Integrated Modeling of Metabolism and Transcriptional Regulation in *Rhodobacter sphaeroides*

By

Saheed Rotimi Imam

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Cellular and Molecular Biology)

at the

UNIVERSITY OF WISCONSIN-MADISON

2014

Date of final oral examination: 04/23/2014

The dissertation is approved by the following members of the Final Oral Committee:

Timothy J. Donohue, Professor, Bacteriology

Daniel R. Noguera, Professor, Civil and Environmental Engineering

Nicole T. Perna, Professor, Genetics

Jennifer L. Reed, Associate Professor, Chemical and Biological Engineering

Robert C. Landick, Professor, Biochemistry

Abstract

Obtaining a systems-level understanding of how microbes transduce extracellular stimuli into concerted regulatory and metabolic responses is pivotal to our understanding of cellular physiology and to the development of rational approaches for strain design. Due to its metabolic diversity and complex regulatory network, I conducted a systems-level analysis of metabolism and transcriptional regulation in the photosynthetic bacterium, *Rhodobacter sphaeroides*. Using genome information and previous experimental evidence, I constructed a genome-scale metabolic model for this organism, iRsp1095, which I validated both qualitatively and quantitatively. I subsequently refined and extended iRsp1095 using data from high-throughput substrate utilization analysis and chemostats operated at varying light intensities. The refined metabolic model, iRsp1140, consists of 1416 reactions, 878 metabolites and 1140 genes. iRsp1140 showed significantly improved predictive power over iRsp1095, with its growth rate predictions showing a high correlation to experimental observations ($R > 0.92$). By integrating ~200 global gene expression datasets with genome sequence information from 8 closely related α-Proteobacteria, I also constructed a large-scale transcriptional regulatory network (TRN) for *R. sphaeroides* consisting of 120 gene clusters, 1211 genes (including 93 transcription factors (TFs)), 1858 regulatory interactions and 76 DNA sequence motifs. In addition to being consistent with known regulatory modules in *R. sphaeroides* and other bacteria, the TRN showed a high level of functional coherence and predicted several novel interactions and links between regulatory sub-networks. I validated predictions for this TRN by assessing genome-wide protein-DNA interactions for 9 TFs (PpsR, FnrL, PrrA, CrpK, RSP_2888, RSP_3341, RSP_0489, CcmR and AkgR), 6 of which I characterized for the first time. Overall the TRN predictions had a precision of ~60% and recall of ~50% for these 9 TFs. Finally, using 2 different approaches, I combined the metabolic and TRN models to generate integrated models for metabolism and transcriptional regulation for *R. sphaeroides*, which can enable the prediction of several condition-dependent and gene deletion phenotypes not achievable with the metabolic model

alone. I propose these models and future iterations of them will provide novel systems-level insights into regulatory and metabolic systems that are central to activities of *R. sphaeroides* and related bacteria.

Acknowledgments

My journey through graduate school has been immensely fulfilling and for that I have a several very important people to thank. First off, mom and dad! I wouldn't have made it here without all your support – moral, spiritual, financial and much more. For that I would like to express my deepest thanks and gratitude. This body of work is dedicated to you both!

I would also like to thank my three siblings (Abe, Bisi and Ola) for all their encouragement throughout the course of my graduate education.

I am very thankful that I decided to conduct my graduate studies under the guidance of Dr. Timothy Donohue. By giving me the freedom to develop my thesis research from the ground up, you have allowed me to significantly broaden my scientific horizons. So I would like to say a big thank you! I also have to thank all the other members of my thesis committee who have also guided me through this process with great enthusiasm.

I would also like to take this opportunity to thank Dr. Mecky Pohlschroder, who played an instrumental role in helping me take my first real steps as a scientist and also in my decision to continue my graduate education beyond my master's degree.

Finally, the long road to a PhD can be full of many challenging moments and so it's important to have like-minded people with whom you can share those difficulties. Thus, I would like to thank all the members of the Donohue lab and associated labs for all their support and contributions to making my time here in Madison some of the most enjoyable and fulfilling of my life… Thanks guys!!!

Table of Contents

List of figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 7

Chapter 8

List of tables

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

Chapter 8

Chapter 1: Introduction

Obtaining a systems-level understanding of cellular complexity

Cells are multifaceted biological systems consisting of numerous interacting components which include macromolecules – DNA, RNA, proteins, lipids and carbohydrates – and a wide variety of cofactors, metabolites and inorganic compounds. Over the years, our understanding of how cells function has been driven by detailed analyses of how its individual components work, via genetic analysis of interactions or biochemical characterization of macromolecular structure and function. As this knowledgebase increased, the need arose to gain a more integrative understanding of how these individual components function in context of the multitude of other components within the cell. Furthermore, with technological advances enabling system-wide and dynamic quantification of the states, activities and interactions of these cellular components, systems-level approaches to harnessing these data and obtaining new, holistic perspectives of biological systems has come to the forefront of scientific research.

Cellular processes such as signal transduction, transcription, translation and metabolism are central to life and inter-dependent on one another. Computational modeling offers an avenue for understanding the dynamics of these biological processes from a global perspective. With the availability of techniques for obtaining system-wide data sets, modeling of these processes has become feasible, through the application of some simplifying assumptions that reduce the complexity of the system. Such models of living systems, while not complete, can generate new hypotheses and make predictions that guide scientific discovery.

My thesis research focused on building, validating and utilizing systems-level models of metabolism and transcriptional regulation to better understand microbial activities. To achieve this, I analyzed the well-studied photosynthetic bacterium, *Rhodobacter sphaeroides*. In this introductory chapter, I discuss some of the features of *R. sphaeroides* that make it a useful organism for systems-level analyses. I then provide brief introductions into the computational approaches available for constructing large-scale systems

biology models of metabolism and transcriptional regulation, as well as approaches for integrating these large-scale models of living systems.

Rhodobacter sphaeroides as a model system for studying cellular activity

While numerous cell types can be used to gain a systems-level understanding of biology, bacteria represent ideal organisms for this analysis because of their relative simplicity, ease of growth, genetic tractability, amenability to genomic tools, large number of sequenced genomes, significant amounts of large-scale datasets and prior knowledge available for many organisms [1]. For such reasons, *Escherichia coli* has served as a paradigm for systems-level analysis of prokaryotic biology [2]. However, a large amount of diversity exists among prokaryotes, with many groups of bacteria exhibiting unique lifestyles or functions. Thus, to study these specialized lifestyles, some of which are crucial to function of many ecosystems (for example photosynthesis, carbon dioxide or nitrogen fixation etc), other prokaryotes must be studied.

Rhodobacter sphaeroides is a member of the alpha sub-group of Proteobacteria. It is a rod-shaped, uni-flagellated purple non-sulfur bacterium that has been studied for decades as a model system for bacterial photosynthesis [3]. Another important feature of *R. sphaeroides* that makes it an excellent candidate for systems analysis is its metabolic diversity. *R. sphaeroides* is a facultative bacterium that grows by aerobic respiration, anaerobic respiration and anoxygenic photosynthesis [3] (Figure 1-1). In addition, *R. sphaeroides* is capable of fixing both atmospheric carbon dioxide (CO_2) and nitrogen (N_2) under anaerobic conditions (Figure 1-1). Thus, from a fundamental knowledge perspective, understanding the complexities of maintaining, regulating and coordinating these various lifestyles represents an important scientific endeavor. I aimed to contribute to this endeavor by a systems-level analysis of this organism. From an applied perspective, *R. sphaeroides* is naturally capable of producing significant amounts of a variety of compounds, which could be of socio-economic value. For instance, under anaerobic conditions, *R. sphaeroides* produces large amounts hydrogen gas (H_2) through activities of its nitrogenase complex

[3-5], as well as increased amounts of membrane components to house its photosynthetic machinery. Given its ability to increase membrane synthesis, *R. sphaeroides* also produces more ubiquinone than many other bacteria , which can serve as a pharmaceutical or food supplement [6]. It also has the native ability to accumulate polyhydroxybutyrate (PHB), which can be used in the manufacture of biodegradable plastics [7] (Figure 1-1). *R. sphaeroides* has also been applied for the purposes of bioremediation of radioactive contamination [8]. I aimed to use systems approaches to better understand these metabolic processes and their regulation, with the aim of optimizing their production.

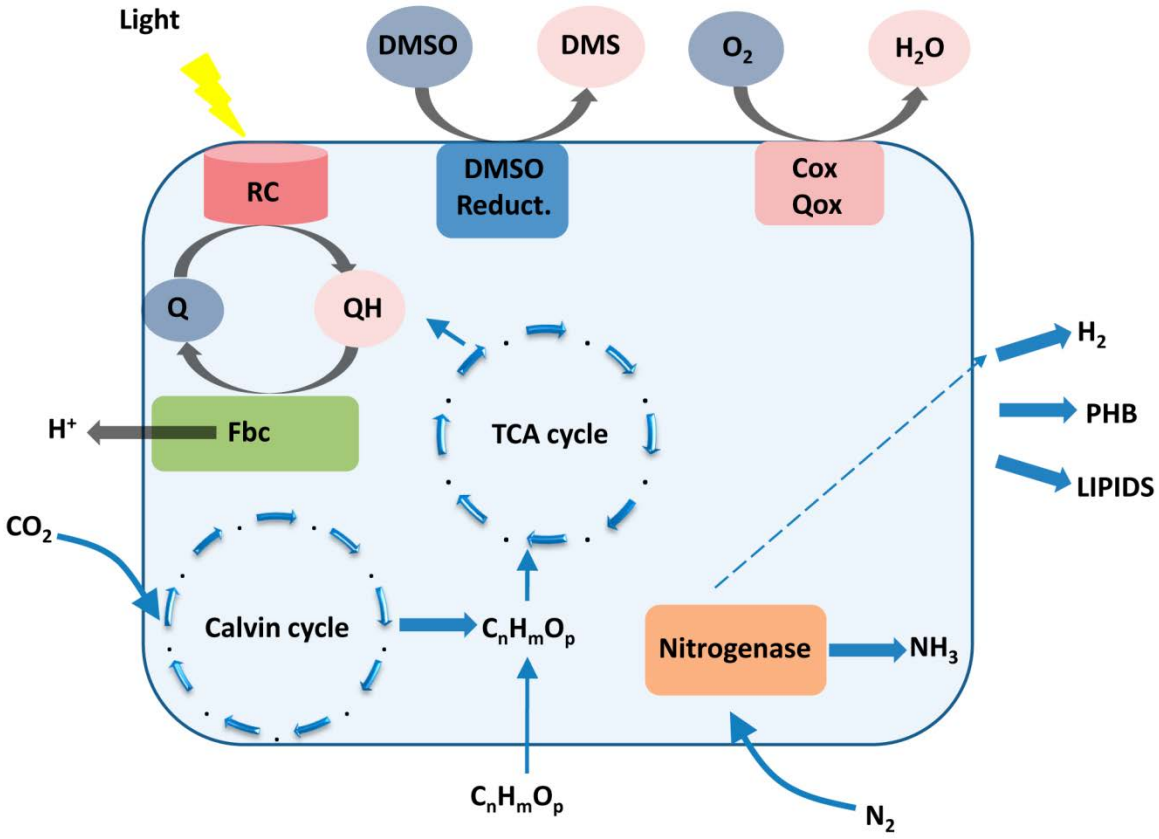

Figure 1-1. Metabolic capabilities of *R. sphaeroides*. Schematic highlights the metabolic versatility of *R. sphaeroides*, showing its major growth modes (anoxygenic photosynthetic, aerobic respiratory and anaerobic respiratory using dimethyl sulfoxide (DMSO) as a terminal electron acceptor), as well as its ability to produce a variety of interesting metabolites like H_2, lipids and PHB. RC – reaction center, DMS – dimethyl sulfide, Cox – cyctochrome oxidase, Qox – quinol oxidase, Q – ubiquinone, QH – ubiquinol, fbc – cytochrome bc_1 complex, $C_nH_mO_p$ – electron rich carbon source.

R. sphaeroides is also a genetically tractable microbe, with tools available for gene deletion, over-expression, site directed mutagenesis and promoter functional analyses [9-11]. A variety of genomic tools have also been developed for this system including an Affymetrix genechip for global transcript level analysis [12] and a Nimblegen tiling array for genome-wide analysis of protein-DNA interactions [13]. In addition, next generation sequencing based technologies are also being implemented for use in *R. sphaeroides*. These approaches have been used to generate a variety of new datasets informing us on the functions of a variety of important proteins, as well as changes in cell behavior in response to genetic or environmental perturbations. For instance, close to 200 microarray datasets are publicly available for *R. sphaeroides* studying effects of light intensity, oxygen tension, salt concentration, gene deletions etc, on transcription in *R. sphaeroides* [14-19]. In addition, the genome-wide binding locations for several transcriptional regulators have also been analyzed including those for σ-factors (RpoD, RpoE, $RpoH_I$, $RpoH_{II}$) and the transcription factor (TF) FnrL [15, 20].

Thus, *R. sphaeroides* provides a unique opportunity to study important aspects of cellular physiology and complexity, with a wealth of prior biochemical or genetic knowledge that can be leveraged for building detailed computational models and an array of experimental tools to study predictions from such models or provide insight into previously uncharacterized aspects of this or related organisms. Given that ~30% of the *R. sphaeroides* genome is still annotated as "hypothetical" [21], with an additional ~12% of the open reading frames having nondescript annotations, much still remains to be learned about the various function encoded in its genome.

Anoxygenic Photosynthesis: an ancient metabolic lifestyle of R. sphaeroides

Photosynthesis is arguably the most important biological process for the sustainability of life on our planet. Photosynthetic organisms ranging from higher plants to microbes, harness solar energy and fix atmospheric CO_2 making them integral to the global food chain, carbon cycling and energy conservation. Oxygenic photosynthetic organisms such as higher plants and cyanobacteria, carry out photosynthesis by

a combination of cyclic- and non-cyclic photophosphorylation, absorbing photons and extracting electrons from water to ultimately generate ATP and NADPH used by the Calvin cycle for CO_2 fixation, with O_2 produced as a byproduct of photochemistry by one of two photosystems [22] (Figure 1-2).

In anoxygenic phototrophs such as green sulfur and purple bacteria like *R. sphaeroides*, photosynthesis occurs only via cyclic photophosphorylation, with no requirement for water splitting or production of O_2 [23] (Figure 1-2). Indeed, each of the two photosystems in oxygenic phototrophs are proposed to be derived from ancestral membrane-bound complexes found in modern day anoxygenic phototrophs [24, 25].

The *R. sphaeroides* photosynthetic apparatus is housed in the intracytoplasmic membrane (ICM), invaginations of the cell membrane which perform an analogous function to the thylakoid membrane in chloroplasts [26]. Light harvesting complexes in the *R. sphaeroides* ICM absorb light energy at characteristic wavelengths in the infrared range (800 – 875 nm) and use this energy to excite chlorophyll pigments in the photosynthetic reaction center to a higher energy state [27]. Electrons are released from light-excited reaction center pigments, transferred to ubiquinone (reducing it to ubiquinol) and then to the lower energy electron carrier cytochrome c_2, with the energy of this electron transfer conserved by the pumping of protons across the ICM by ubiquinol-cytochrome *c* reductase (cytochrome bc_1 complex) (Figure 1-2). This proton gradient eventually drives ATP synthesis [23]. Electrons for CO_2 fixation via the Calvin cycle can be derived either from H_2 oxidation (during photoautotrophic growth) or from electron rich-carbon sources (during photoheterotrophic growth).

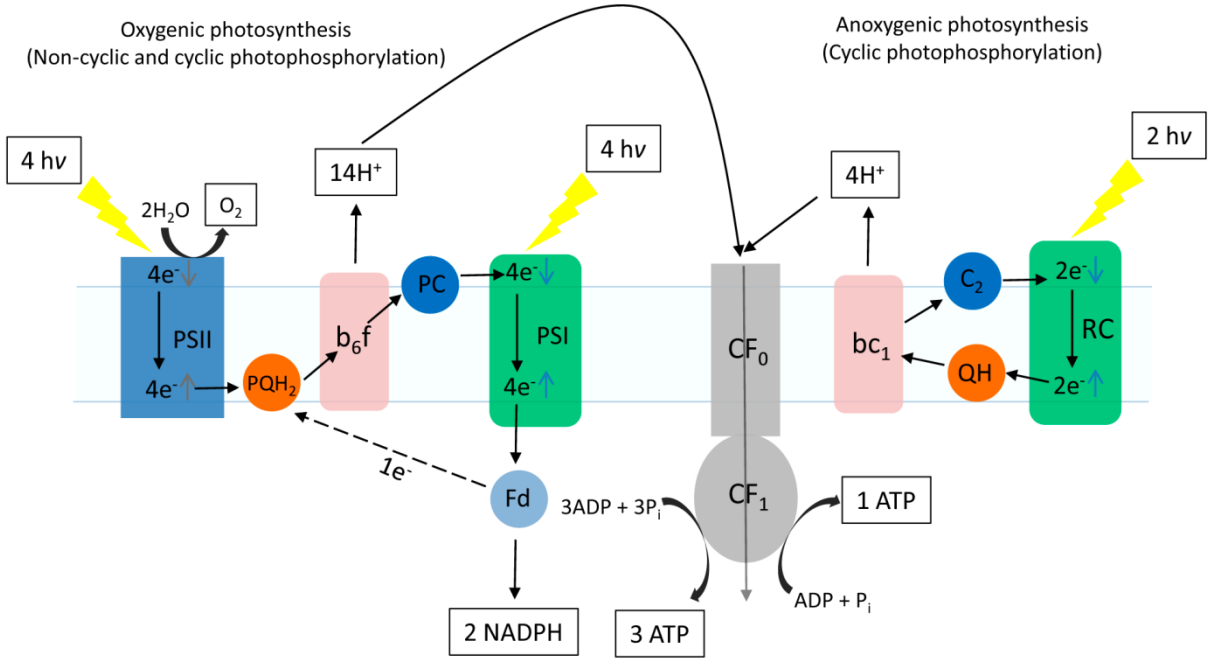

Figure 1-2. Cyclic and non-cyclic photophosphorylation. A comparison of anoxygenic photosynthesis by cyclic photophosphorylation, as it occurs in *R. sphaeroides*, to oxygenic photosynthesis that uses both non-cyclic and cyclic photophosphorylation to generate ATP and NADPH for carbon dioxide fixation. $e^-\uparrow$ indicates high energy electrons while $e^-\downarrow$ indicates low energy electrons. Some intermediate electron carriers are omitted for brevity. h*ν* – photon, RC – reaction center, QH – ubiquinol, C2 – cytochrome c_2, bc_1 – cytochrome bc_1 complex, PSI – photosystem I, PSII – photosystem II, PQH_2 – plastoquinol, PC – plastocyanin, b_6f – cytochrome b_6f complex, Fd – ferredoxin, CF_0F_1 – ATP sythase.

Regulation of Photosynthesis in R. sphaeroides

Previous studies of anoxygenic photosynthesis in *R. sphaeroides* revealed that this lifestyle is controlled at the transcriptional level by 3 TFs: PpsR, FnrL and PrrA (Figure 1-3). PpsR is proposed to function as a transcriptional repressor, reducing expression of photosynthesis related genes under aerobic conditions [28]. The repression of photopigment biosynthesis under aerobic conditions is important to prevent the production of reactive oxygen species by photopigments in the presence of light, which can damage cellular components. PpsR activity is regulated by its cognate O_2-sensitive anti-repressor, AppA, which binds to PpsR under anaerobic conditions preventing it from repressing its target genes [29-32]. AppA has also been proposed to sense blue light [32]. Chromatin immunoprecipitation (ChIP) analysis showed that PpsR directly binds the promoter region for the bacteriochlorophyll biosynthetic operon *bchFNBHLM-puhA* (RSP_0284-91) and *ppaA* (RSP_0283) [33]. In addition, some genes not directly involved in photosynthesis were proposed to be as direct PpsR targets, including *ccoO* (RSP_0695), RSP_2122 and RSP_3241 [33].

R. sphaeroides FnrL is a conserved, oxygen-sensitive, iron-sulfur (Fe-S) cluster-binding TF belonging to the Crp/Fnr family. FnrL has been shown to be required for both anaerobic respiratory and photosynthetic growth in *R. sphaeroides* [34, 35]. Furthermore, ChIP followed by hybridization to a tiling array (ChIP-chip) analysis showed that FnrL is a global regulator directly involved in controlling transcription of genes that are required for photosynthesis and anaerobic respiration in *R. sphaeroides* and related α-proteobacteria [20]. Under anaerobic conditions, FnrL activates the expression of its target genes, which include the bacteriochlorophyll biosynthetic operon *bchEJGP* (RSP_0281-76) and the anaerobic respiration regulatory histidine kinase *dorS* (RSP_3044).

PrrA is the response regulator of the PrrAB to component system, which is required for photosynthetic growth in *R. sphaeroides* [36]. Global gene expression analysis conducted on a PrrA deletion mutant under anaerobic respiratory growth conditions, indicates that PrrA is a global regulator that could directly

or indirectly regulate over 25% of the *R. sphaeroides* genome [16]. PrrA is proposed to activate the expression of photosynthesis related genes at low oxygen tensions [16, 36]. *In vitro* analysis of PrrA has also shown that this TF directly binds to the promoters of *hemA* and Calvin cycle operons in *R. sphaeroides* [37, 38].

Recently, a small non-coding RNA, PcrZ, has also been implicated in regulating the expression of photosynthesis related genes [39]. While much is known about the transcriptional regulation of photosynthesis in *R. sphaeroides*, much of the focus of this prior research has been directed toward the control of genes known to be required for anaerobic photosynthetic growth. Thus, there are likely still uncharacterized TFs and other systems that are important for the anaerobic or photosynthetic lifestyles of this facultative bacterium. I aim to use large-scale models of transcriptional regulation to gain new insights into the regulation of photosynthesis and other aspects of the physiology of *R. sphaeroides* and related bacteria.

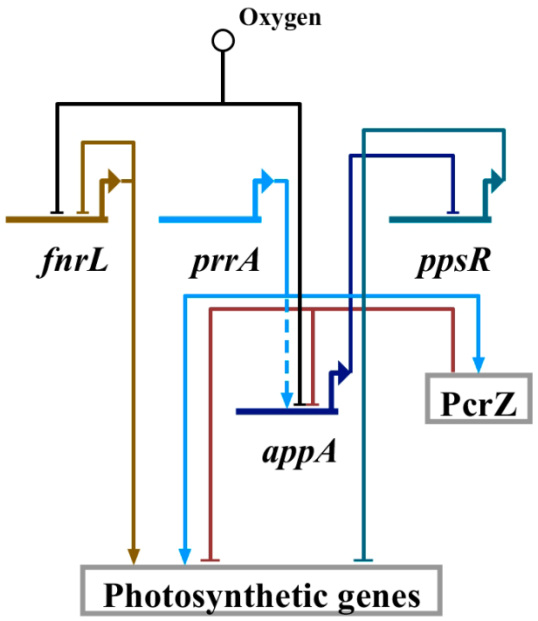

Figure 1-3. Photosynthetic gene regulatory network. The known gene regulatory network controlling photosynthesis in *R. sphaeroides*.

Other well-studied aspects of R. sphaeroides physiology

In addition to photosynthesis, *R. sphaeroides* has been used to study other key cellular processes including:

Nitrogen fixation and hydrogen production – Like most other purple non-sulfur bacteria, *R. sphaeroides* is a diazotroph capable of fixing atmospheric nitrogen (N_2) into ammonia [40, 41]. N_2 fixation is accomplished through the use of the conserved, molybdenum-dependent, multi-subunit nitrogenase enzyme, which catalyzes the energy-dependent reduction of N_2 to ammonia via the following reaction [40]:

$$N_2 + 16\ ATP + 4\ Reduced\ ferredoxin + 16\ H_2O \rightarrow 2\ NH_3 + H_2 + 16\ ADP + 16\ P_i + 4\ Oxidized\ ferredoxin$$

Ammonium (NH_4^+) is the preferred nitrogen source for *R. sphaeroides* and many bacteria. Thus, given the high energetic cost of the nitrogenase reaction, this reaction is prevented in the presence of NH_3 both transcriptionally (via inactivation of NifA – the nitrogenase operon transcriptional regulator) [42] and post-transcriptionally (via ADP-ribosylation of the enzyme) [43, 44]. In addition, both nitrogenase and NifA are oxygen-sensitive [41], thus expression of the nitrogenase structural genes occurs only under anaerobic growth conditions in *R. sphaeroides*. In α-Proteobacteria, the nitrogenase structural genes *nifHDK* and proteins required for assembly of the nitrogenase complex such as *nifXNE*, are transcriptionally regulated by NifA, which recruits RNA polymerase containing the RpoN σ-factor, which is generally involved in transcription of nitrogen metabolism genes [40].

The production of H_2 in *R. sphaeroides* is closely associated with the nitrogen status of the cell, as the nitrogenase enzyme is the major source of H_2 in this bacterium [5, 19, 40]. During photoheterotrophic growth on reduced carbon sources under nitrogen-limiting conditions, the nitrogenase operon is actively transcribed and nitrogenase activity is high [19]. Under these conditions, when alternative nitrogen

sources like glutamate are provided to the cells in the absence of N_2, *R. sphaeroides* produces large amounts of H_2 via the nitrogenase activity [5, 19].

Stress responses – The response of *R. sphaeroides* to photo-oxidative and heat stresses have also been extensively studied [15, 45-47]. Photo-oxidative stress is caused by the reactive oxygen species (ROS) singlet O_2, which is formed by the transfer of energy from light-excited bacteriochlorophyll molecules to O_2 [47, 48]. Singlet O_2 can damage cells and cellular components, so to survive, cells need to respond to the presence of this ROS by induction of sets of genes required to quench singlet O_2 and repair any cellular damage caused [47]. In *R. sphaeroides* and many other bacteria, the master regulators of the response to photo-oxidative stress are the alternative σ-factor σ^E and its cognate anti-σ factor ChrR [14, 15, 49]. σ^E-ChrR sense singlet O_2 via an as yet undetermined mechanism, but in the presence of this ROS, σ^E activates the transcription of a set genes required to amount an appropriate stress response [14, 15, 50]. One of the direct target of σ^E is $RpoH_{II}$, a σ^{32} family σ-factor that regulates an additional ~144 genes required to respond to singlet O_2 stress.

R. sphaeroides has two σ^{32} paralogs, $RpoH_I$ and $RpoH_{II}$. While $RpoH_{II}$ is part of the regulatory cascade required for responding to photo-oxidative stress, $RpoH_I$ is directly regulates genes required for mounting the heat shock stress response, activating ~175 target genes [45]. Global analysis of the regulons of $RpoH_I$ and $RpoH_{II}$ showed that these two σ-factors regulate overlapping but distinct targets required for responding to both photo-oxidative and heat shock stresses, suggesting an evolutionary convergence of the transcriptional responses to these stresses [45].

Motility and chemotaxis – *R. sphaeroides* has a single uni-directionally rotating flagellum used for swimming and chemotaxis [51]. While chemotaxis and motility has been studied extensively in *E. coli*, the chemotaxis operons are more numerous, and the sensory mechanism more complex, in *R. sphaeroides* [52]. Hence, it has been used for several decades as an alternative model system for studying bacterial chemotaxis [52]. The *R. sphaeroides* genome encodes 3 chemotaxis (*che*Op_1, *che*Op_2 and *che*Op_3) and 2

flagella (RSP_0052-66 and RSP_0083-71) operons [53-55]. *che*Op$_2$ and *che*Op$_3$ are required for normal chemotatic responses [53, 56], while the precise functional role of *che*Op$_1$ is yet to be determined [53]. Regulation of flagella biosynthesis and cell motility, as well as chemotaxis, is jointly controlled by 3 TFs: FliA, RpoN2 and FleQ [56, 57].

Understanding biological systems through computational modeling

Mathematical models of biological phenomena (biomodels) have played an important role in scientific inquiry for decades [58]. Biomodels are mathematical representations of biological systems or processes that try to capture their core or essential features, while excluding other, presumably less important, components of the system [59]. Such simplifications are necessary due to the multifaceted nature of living systems, so biomodels are inherently incomplete. However, if the essence of the biological process being modeled is captured, biomodels can be very useful in providing new insight and guiding scientific discovery [58, 59]. Such biomodels can be used to make quantitative predictions about the system, test a large number of hypothesis *in silico*, and guide experimental analysis. They can serve as databases for cataloguing components of a living system, providing a framework for better understanding the organism and identifying its most important features. In addition, they can also aid in building tools that allow for intuitive visualizations of the system [1, 58, 59].

With recent technological advances permitting high-throughput genomics, transcriptomics, proteomics, lipidomics, glycomics and metabolomics analyses, systems biology – the study of the whole of a system, not just its individual parts – has become an integral component of biological scientific research, with its main challenge being the conversion of these system-wide measurements into systems-level insight [1]. A major goal of systems biology is the creation of predictive mathematical models of important biological processes, entire organisms or ecosystems. Consequently, models for processes such as protein folding [60], metabolism [61], gene regulation [62], protein translation [63] and signal transduction [64, 65] have been built in different organisms, as well as integrated models of some of these processes [66-69].

Furthermore, cellular models have also been constructed [70, 71], with projects underway to model mammalian organs (such as the human brain [72, 73]) and multi-cellular organisms (e.g., *Caenorhabditis elegans* [74]).

Computational modeling of metabolism

An application of mathematical or computational modeling in biology that has been highly successful in recent years, is the analysis of cellular metabolism [61, 75]. By considering the genes, proteins, reactions and metabolites involved in specific pathways or the entire metabolic network of an organism, models have been used to make qualitative and quantitative predictions about outcomes of growth conditions and genetic perturbations, which have led to novel discoveries and strategies for strain improvements [61, 75]. Many approaches are utilized for modeling of cellular metabolism, including kinetic [76], ensemble [77, 78] and constraint-based [79, 80] modeling. Each of these approaches has their strengths and limitations. For instance, kinetic metabolic models can provide quantitative prediction of fluxes through pathways, however these are limited by lack of experimental information on kinetic parameters for individual reactions. Similarly, ensemble modeling, which also permits quantitative predictions, is often limited both by availability of reference steady state fluxes required for simulation and by computing time [78]. Thus, both these approaches are generally limited to modeling of relatively small sets of reactions. On the other hand, constraint-based models (CBMs) bypass the need for kinetic parameters and are less computationally demanding, allowing of modeling of larger-scale metabolic networks. The drawback of CBMs is a relatively large solution space, potentially making them less quantitative. However, some of the limitations of CBMs can be reduced by application of relevant constraints such as data from system-wide measurements.

Modeling metabolism using constraint-based techniques

Constraint-based modeling is founded on the assumption that the flux of metabolites through a metabolic network at steady state is dependent on enzyme capacity, the stoichiometry of the reactions that make up

the network and their inherent thermodynamic constraints [79, 80]. As a consequence, this approach bypasses the need for kinetic parameters. In addition, many constraint-based problems can be formulated as linear programming problems allowing for computationally efficient and accurate solution strategies, even on a genome-scale. As with other approaches, constraint-based modeling begins with the assembly of all the genes, proteins, metabolites and reactions known or predicted to occur in the relevant biological system. This process of metabolic network reconstruction can be achieved using the annotated genome of the target organism, along with available evidence from literature or databases about its ability to carry out specific reactions [81] (Figure 1-4). This initial draft metabolic network reconstruction is converted into a mathematical representation, generally referred to as the stoichiometric (or S) matrix, and used for initial simulations. Similar to the process of building most models, this process of metabolic network reconstruction is an iterative one [81]. Predictions from each iteration of the CBM are compared with known phenotypes of the target organism, with discrepancies serving as indications of missing or aberrant components in the CBM. These components are added, removed or modified in subsequent iterations and this process is continued until a model is generated that is in good agreement with known phenotypes of the target organism (Figure 1-4). This iterative process of network reconstruction continues as more data becomes available for the target organism, as exemplified by the *E. coli* metabolic reconstruction which has gone through numerous refinements and extension of over the last 20 odd years [61]. I used this approach to build and iteratively refine a genome-scale CBM for *R. sphaeroides* (Chapters 2 and 3).

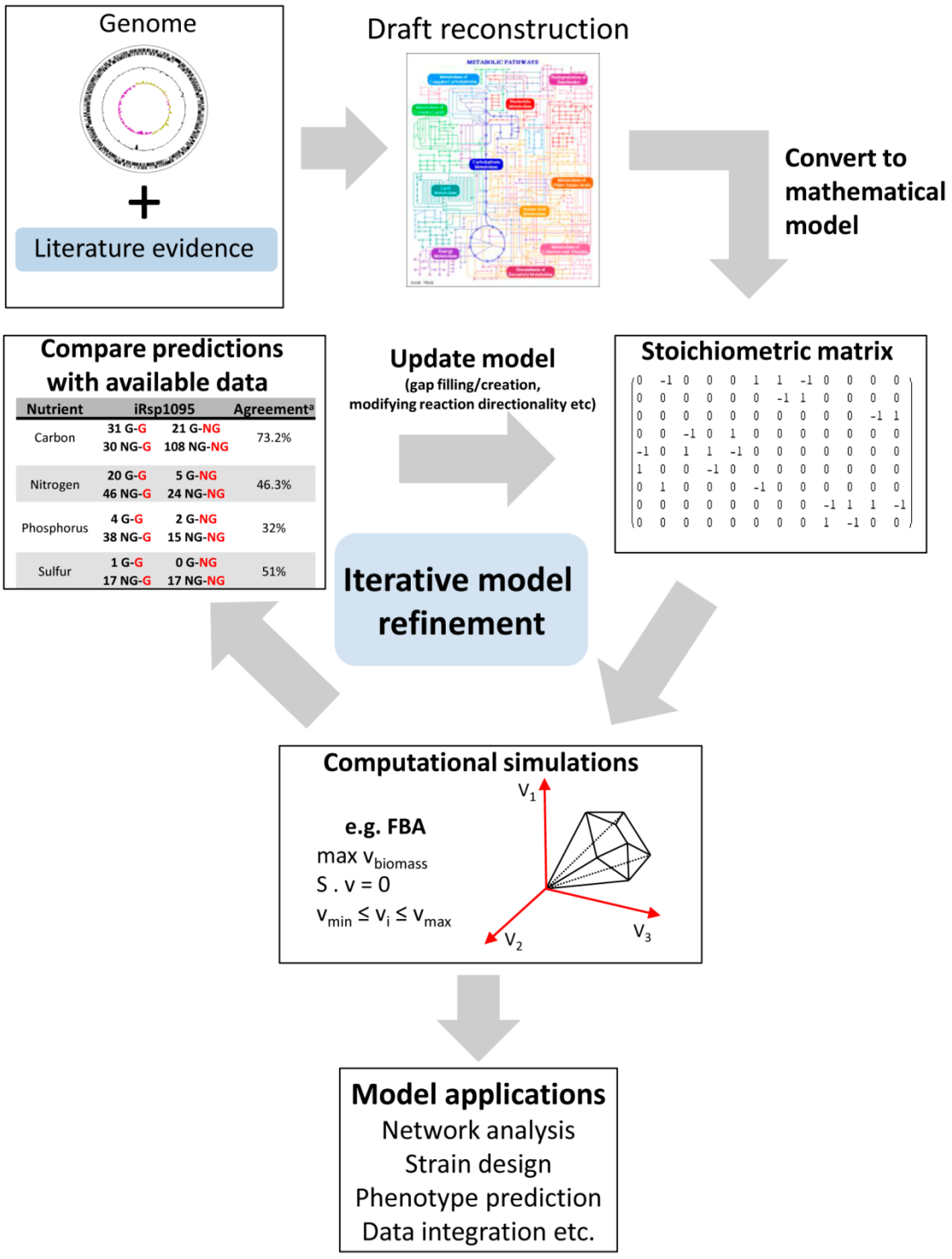

Figure 1-4. Reconstruction and iterative refinement of genome-scale metabolic models.

Simulations are carried out with CBMs using techniques such as flux balance analysis (FBA) [80], which can be used to make experimentally testable qualitative or quantitative predictions about growth rates, phenotypes, flux distributions and production rates of specific metabolites. FBA is based on three main constraints: (i) that cells are at steady state; (ii) that the direction of flux through a reaction is determined by thermodynamic constrains; and (iii) that the amount of flux through any reaction are within specified bounds (measured or arbitrary). These constraints, in addition to the stoichiometric constraints of the metabolic model, allow for the computationally efficient simulation of the steady state flux distributions through an entire network. To conduct these simulations, a relevant objective function is optimized. Typically, for bacterial systems, this objective function is production of cell biomass, which is often a relevant physiological objective of microbial cells [81, 82].

In addition to FBA, numerous other constraint-based modeling approaches have been developed over the last 2 decades, which incorporate modifications that enable alternative forms of network analysis such as flux variability analysis (FVA) [83], alternative optima analysis [84, 85] and extreme pathway analysis [86], amongst others. In addition to approaches that analyze the properties of the network and predict steady state fluxes, methods have also been developed to enable integration of a variety of high-throughput datasets from transcriptomics, proteomics and metabolomics [87-92] into CBMs by using them as additional constraints during simulations. Inclusion of these and other datasets can significantly reduce the optimal solution space of predicted fluxes, thus improving prediction accuracy.

Strain optimization and predicting gene deletion phenotypes

Another practical application of CBMs is the development of genetic strategies for improving the performance of bacterial systems. These strategies could involve genetic changes to remove or add genes or reactions to the metabolic network. Approaches like FBA, minimization of metabolic adjustment (MOMA) [93] and regulatory on/off minimization of metabolic flux changes (ROOM) [94] have been used to make predictions of the effects of genetic perturbations to metabolic flux distributions. MOMA

has also been applied in strain design for improving the production of lycopene and L-valine in *E. coli* through the sequential analysis of effect of deleting genes and/or reactions from the metabolic network [95, 96]. Several other optimization approaches have been developed for improving metabolite production such as optknock [97], optGene [98], optORF [99], optStrain [100] etc, which have met with varying levels of success. I aim to implement some of these approaches with the genome-scale metabolic model for *R. sphaeroides* to develop strategies for the improving the production of compounds like PHB, H_2 and lipids (Chapter 8).

Reverse engineering of transcriptional regulatory networks (TRNs)

Regulation of cellular processes occurs at many levels including at the transcriptional, post-transcriptional, translational and post-translational levels, enabling the cell to make rapid and robust responses to changes in its internal or external milieu. Our ability to reverse engineer, on a systems-level, the regulatory networks at any of these stages depends on our ability to accurately monitor these regulatory processes on a global scale. Currently, the most developed technologies for high-throughput measurement of regulatory processes are for measurement of transcriptional processes, in form of global transcript level analysis using microarray or high-throughput sequencing (RNA-seq) or protein-DNA interaction analysis via ChIP-chip or ChIP followed by high-throughput sequencing (ChIP-seq). This has made reverse engineering of TRNs a focus of statistical and computational analysis in recent years.

Many approaches to reconstructing TRNs utilize only data obtained from global gene expression analysis and try to predict the underlying regulatory rules based on the assumption that the expression profile of a TF across a set or subset of experiments is related to that of its target genes [62, 101-103]. These methods could either be direct TRN inference approaches, wherein the relationship between every possible TF-target pair is assessed to determine the most likely targets of a given TF (e.g., GENIE3) [104-106], or module-based TRN inference approaches wherein genes are first grouped into modules or clusters (i.e.,

groups of genes with related expression profiles) before attempting to predict the TF that is most likely to regulate a given cluster (e.g., LeMoNe) [62, 102, 103] (Figure 1-5).

To assess the performance of the myriad of approaches available for TRN assembly from gene expression data, the "Dialogue for Reverse Engineering Assessments and Methods" (DREAM) project was initiated in 2006 to provide unbiased assessments of available TRN inference approaches via an annual challenge providing both simulated and blinded real world datasets [107]. These challenges have highlighted many of the advantages and limitations of the available expression-based inference approaches, with many published approaches scoring relatively poorly in these challenges. Among the most informative observations thus far from these challenges are: (i) the most useful expression datasets from inferring TRNs are TF deletion and time series experiments, as these often provide the strongest data supporting causality; (ii) the predictions from the best performing inference approaches appear to be complementary, as their underlying assumptions allow them to capture different aspects of the information embedded in the expression data sets; and (iii) current state of the art expression-based inference approaches have almost no predictive power for eukaryotic systems [108]. These observations will help guide future experimentation and algorithm development to improve TRN inference from expression data.

To complement information from gene expression data, some TRN inference approaches integrate sequence information from the upstream-regulatory regions of genes to aid gene clustering (Figure 1-5). For instance, the bi-clustering (i.e., clustering over a genes and conditions) algorithm cMonkey [109] uses an iterative process of first clustering based on gene expression, then looking for a shared motif in the cluster, where possible, and refining members the cluster based on the possession of the shared motif. Thus, the clusters created are more likely to consist of co-regulated genes as they share both a motif(s) and similar expression profiles. The TFs most likely to regulate these clusters are then determined based on the relationship of their expression profiles to those of the members of a cluster [110]. In addition to these, other TRN inference pipelines, which have been applied to well studied bacteria, also try to incorporate data from protein-DNA interaction assays [111, 112].

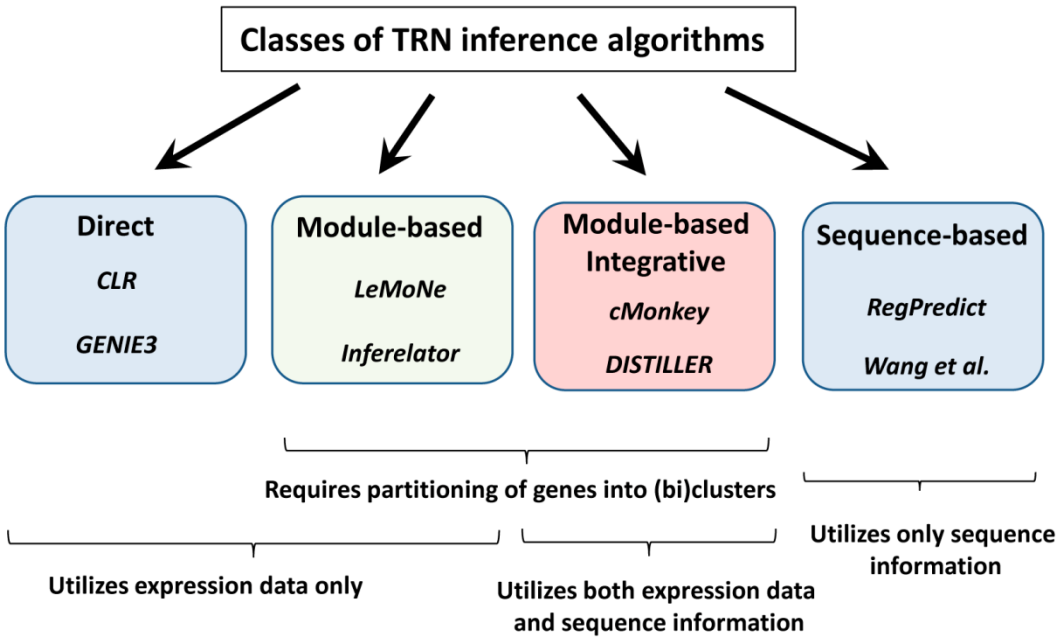

Figure 1-5. Classification of TRN inference algorithms.

Alternative approaches to TRN inference, which do not consider gene expression data, include the comparative genomics-based approaches that attempt to reconstruct TRNs using information of closely-related organisms [113-115] (Figure 1-5). These approaches are based on the assumption that transcriptional regulatory programs including the relevant TF, targets and motifs, are conserved across closely-related organisms. Thus, by comparing the upstream-regulatory sequences of orthologous genes (i.e., evolutionarily related genes), one might be able to identify evolutionarily conserved regulatory motifs, with all genes sharing a particular conserved sequence considered as being co-regulated.

Each of these inference approaches has its strengths and limitations, proving very effective in capturing the regulons (i.e., complement of genes under direct regulation of a given TF) of some TFs with particular characteristics, while being less effective for others [101, 108]. For instance, expression-based inference approaches generally perform very well for TFs whose expression profiles share some relationship with that of their target genes. However, many TFs in bacteria, and even more so in eukaryotes, are post-transcriptionally regulated, thus their gene expression profiles have very little relationship, if any, to that of their target genes, making expression-based approaches of limited predictive value for these types of TFs. Furthermore, indirect control of gene transcription through the activity of regulatory cascades can also lead to additional prediction errors [116]. On the other hand comparative genomics based approaches, which do not rely on expression data, may fare better in these cases.

Another approach to assembling large-scale TRNs is the use of integrative module-based inference approaches that use both expression and sequence information for clustering. These integrative approaches are particularly useful for regulators with sizeable regulons, but often fail to identify interactions for regulators with only a few target genes due to limited sequence information for motif detection and filtering steps used during clustering. On the other hand, direct TRN inference approaches are better at identifying such interactions, but less effective with large regulons.

Yet, another example involves TRNs built using comparative genomics approaches. While these approaches are powerful, they are highly dependent on the *de novo* motif detection algorithms used and thresholds applied during detection and comparison of motifs during clustering. Furthermore, some TFs and their targets genes may only occur within a very narrow range of species or even strains of an organism, making them inaccessible to comparative genomics approaches. Thus, a systematic approach to combining these approaches, which can be complementary, could result in generation of TRN models with fuller coverage and improved predictive power. I used such a systematic approach to reconstruct a large-scale TRN for *R. sphaeroides* (Chapter 5).

Validation of inferred TRNs

Large-scale TRN models inferred from high-throughput data can provide significant new insight into the regulatory circuits present in a target organism by identifying novel interactions, functions and links between apparently independent sub-networks (Figure 1-6). However, given that the TRN inference problem is generally considered an underdetermined one (i.e., the available data is insufficient to converge on a unique solution or network [116, 117]), TRN models can be excepted to be incomplete and contain erroneous predictions. For instance, the *R. sphaeroides* genome contains 4300 genes, while only ~200 non-unique experiments are available to infer its TRN, making this problem underdetermined (more genes than samples). Furthermore, as mentioned above, some experiments are apparently more informative than others for the purposes of TRN inference. Thus, once assembled, the quality of the resulting TRN must be assessed using available or newly generated experimental data (Figure 1-6). For organisms whose TRNs have been extensively studied such as *E. coli* and *Bacillus subtilis*, databases of manually curated, experimentally verified interactions (such as RegulonDB [118] and DBTBS [119] respectively) exist consisting of thousands of known interactions, which can be used for assessing the quality of the inferred TRN. However, for organisms like *R. sphaeroides*, with less well characterized TRNs, network validation will require extensive experimental verification, requiring the use of high-throughput assays like ChIP-seq and global expression profiling on mutant strains, as well as lower

throughput experiments such as gene-specific ChIP-qPCR, reporter gene fusion assays, DNase footprinting and electrophoretic mobility shift assays (ESMA) to test individual predictions. Using these approaches, the quality of predictions in a TRN model can be assessed, novel interactions identified, and the TRN refined. I validated predictions from the large-scale TRN I constructed for *R. sphaeroides* by analyzing the regulons of several known and previously uncharacterized TFs (Chapters 5 – 7).

As with the creation of any biomodel, the process of constructing TRNs has to be iterative, with new experimental data being used to refine and update the initial TRN (Figure 1-6). In addition to the use of experimental data to refine TRNs, other computational approaches can be used to extend specific sub-networks via supervised, semi-supervise or unsupervised approaches [101]. For instance, the supervised inference algorithm SEREND [111] takes known information about the properties of a TF (including its binding motif, target genes and their expression profiles across previous and newly generated dataset) as training data. SEREND then uses the features obtained from the training set to predict novel members of the regulon sharing those features, thereby potentially extending the TRN.

In general, the process of TRN refinement and extension should be a community-based effort integrating findings from different groups. To make this possible, it is important that all regulatory interactions predicted and/or verified are catalogued and stored in databases to allow efficient updates. Thus, I also aim to create a publicly accessible database archiving all known and predicted regulatory interactions in *R. sphaeroides*, which can be updated as new information becomes available.

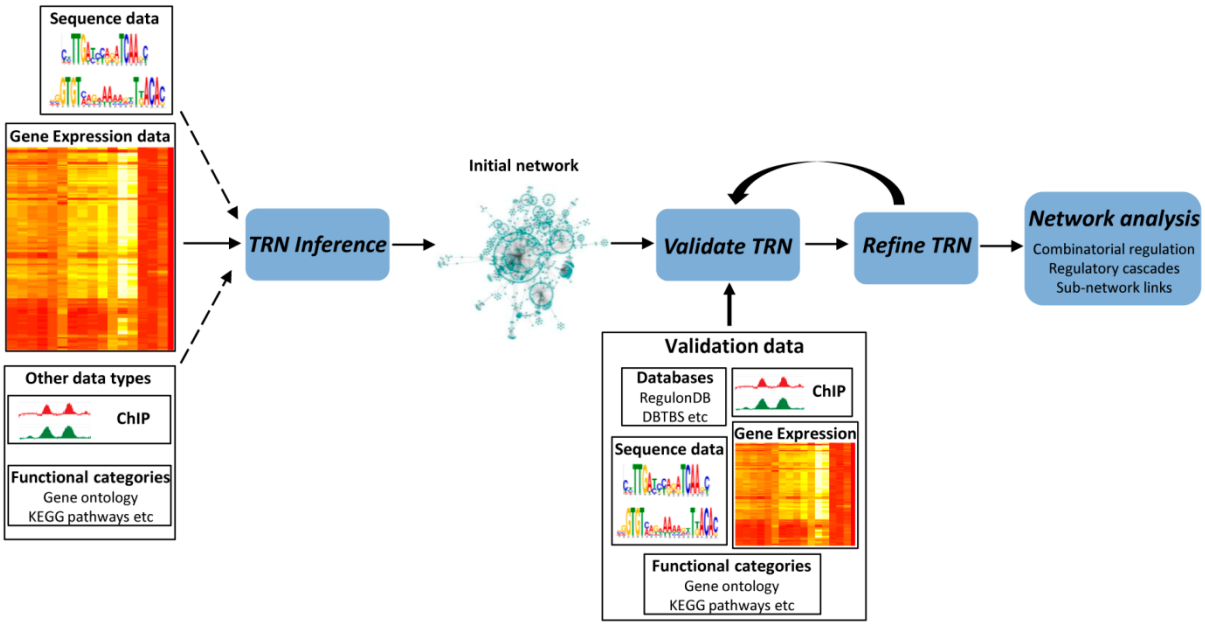

Figure 1-6. Summary of TRN inference workflow. Schematic showing the various stages of TRN reconstruction including inference, validation, refinement and network analysis. TRN inference can involve the use of gene expression data only or the integration of other data types, while a variety of data types can be used for validation.

Dynamic regulatory models

While reconstruction of TRN topology can allow prediction of a large number of novel interactions, potentially providing a significant amount of new scientific insight, this model is static. That is, it does not allow quantitative predictions of the outcomes of changes to that topology. One of the goals of systems biology is to create dynamic models that provide quantitative predictions about how the system would react to changes in one or more of its components [1, 58, 59]. To this end, some TRN inference approaches enable the generation of models based on multiple regression or ordinary differential equations, which permit quantitative predictions of how perturbations to the system might affect expression profiles of the genes included in the model [62]. For instance, TRN models built using the cMonkey-inferelator workflow [62, 109, 110], allow predictions of a new state of the TRN in response to genetic or environmental perturbations. In this approach, the expression state of a gene can be thought of as a multiple regression equation consisting of several predictors (e.g., TFs). If one of these predictors is removed from the system, the expression of the gene can be determined from the input of the other remaining predictors, thus allowing a quantitative prediction of the new state of the system.

An alternative approach to making quantitative predictions with TRN models, is by integrating them with other models that allow dynamic predictions, wherein perturbations to either the TRN or the model to which it is combined, affect the final predicted output of the integrated system. To this end, several approaches have been developed to integrate TRNs with biomodels like CBMs, wherein the final predicted output of the integrated model is dependent on the states of both the TRN and CBM [66, 67, 69].

Integrated modeling of metabolism and transcriptional regulation

A systems-level understanding of connections between transcriptional regulation and metabolism is key to understanding cellular growth and for the rational strain design. While useful information and a wide range hypothesis can be generated from individual metabolic or TRN models independently, there is often

significant interplay between these systems, as TFs control expression of genes encoding metabolic enzymes and some enzymes produce metabolites that alter the activity of TFs [69]. Thus, to generate improved models of cellular lifestyles, integrating individual network models is a necessary and important goal of systems biology.

Approaches have been developed that enable integration of TRNs with CBMs (Figure 1-7). One of the earliest of these, regulatory FBA (rFBA) [67], uses the rules derived from a TRN to determine the activity state of the genes in the CBM depending on growth conditions, prior to growth simulation by FBA. Thus, genes are turned "on or off" depending on the predictions of the TRN and simulated growth conditions. As a result, reactions catalyzed by enzymes encoded by genes included in the TRN are given on or off activity states, prior to simulating growth, while those not included in the TRN are left unconstrained. During an rFBA simulation, only a subset of reactions will be capable of carrying fluxes, thereby reducing the CBM's solution space based on TRN interactions.

An alternative to rFBA, steady state regulatory FBA (SR-FBA) [69], also assigns on or off states to genes in the metabolic network based on activity of their corresponding regulators under a given condition. However, unlike rFBA, gene, protein and reaction states are determined during steady state simulations in SR-FBA. The relationships between TFs, genes, proteins and reactions are formulated as linear equations, to be solved as part of the simulation of steady state flux distributions with FBA. While both rFBA and SR-FBA models are based on Boolean logic and thus relatively rigid, as genes are only allowed to exist in one of two states (i.e. on or off), they have been successfully applied in organisms like *E. coli* [67, 69].

An alternative to these Boolean-based approaches was developed more recently. To circumvent this binary logic, probabilistic regulation of metabolism (PROM) uses gene expression data to estimate the probability that a TF regulates a given metabolic gene under a particular condition [66] (Figure 1-7). Here, the relationship between a TF and a given gene is derived from the TRN, then PROM tries to extend this relationship by assessing how often this relationship holds true, based on observations from

gene expression datasets. For instance, if TF A is known or predicted to activate gene B from the TRN, one assesses the percentage of times gene B is on when the gene encoding A is on from gene expression datasets. If this is 50% of the time, then the probability A activates B would be 0.5. This calculated probability is then used to constrain the maximum flux through the reactions catalyzed by the gene product of B if A is turned off either due to the current experimental condition or the deletion of A from the integrated model. When PROM was applied to the *E. coli* and *Mycobacterium tuberculosis* networks, it resulted in improved accuracy in prediction of growth rate and gene deletion phenotypes [66].

As might be expected, each of these approaches also has strengths and weaknesses. For instance, to generate reasonable estimates of the probabilities of TF-target interactions for PROM, a substantial amount of gene expression data is required. This might make PROM inaccessible to most organisms. Furthermore, PROM assumes some relationship exists between the expression profile of a TF and its target genes. For post-transcriptionally regulated TFs, whose expression profiles might be constant, probabilities calculated via PROM might not capture the true relationship between the TF and its targets, negatively impacting predictions. On the other hand, the Boolean logic based approaches rely entirely on the information contained within the regulatory rules of the TRN. As the TRN will be incomplete and, if inferred, may contain errors, predictions via these approaches would be more sensitive to this lack of information. Nevertheless, these approaches provide veritable means for integrating TRNs and CBMs and have been validated in models systems. I built integrated models of metabolism and transcriptional regulation for *R. sphaeroides* using rFBA and PROM, and assessed the performance of these integrated models using available and newly generated datasets (Chapter 8).

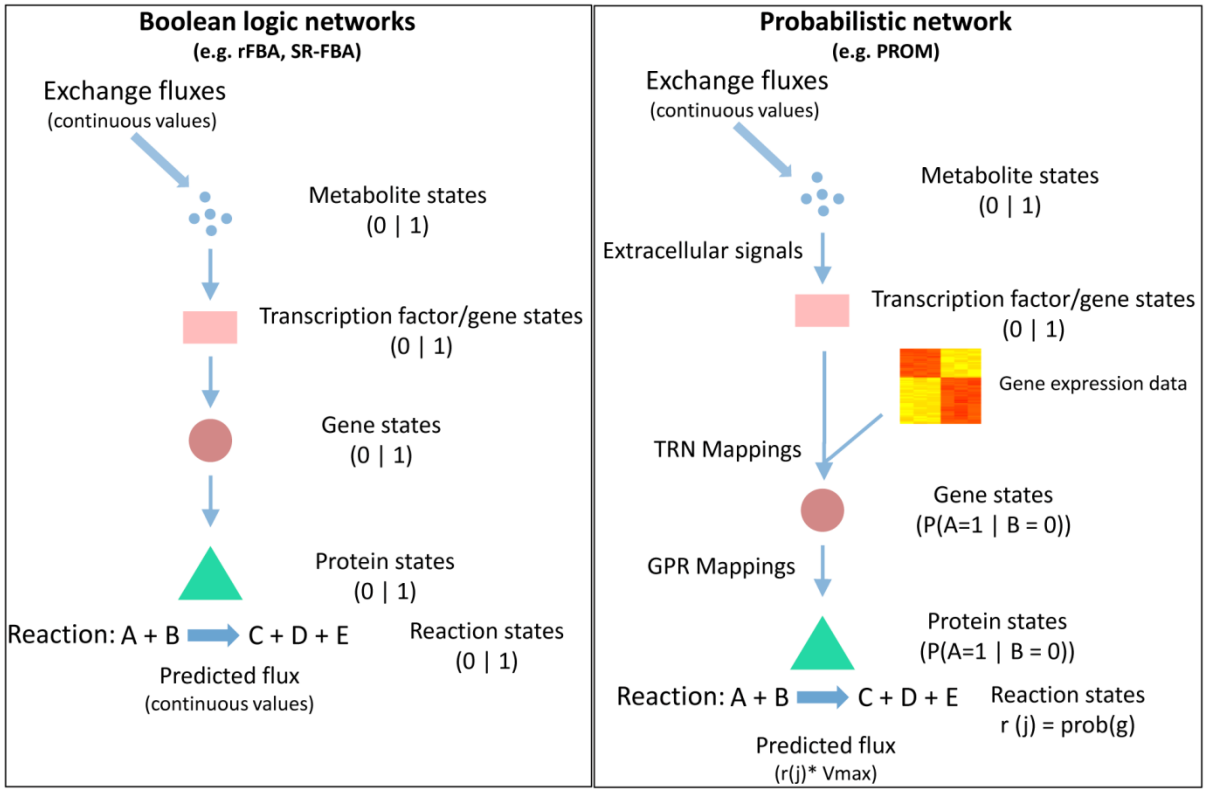

Figure 1-7. Approaches from integrating TRNs and CBMs. Comparison of Boolean logic based and probabilistic approaches to integrating TRNs with CBMs.

Thesis outline

The focus of my thesis research was the construction, validation and refinement of large-scale models of metabolism and transcriptional regulation for the photosynthetic bacterium, *R. sphaeroides*. By employing an approach that combined computational predictions with experimental analysis, I aimed to gain new insights into the physiology *R. sphaeroides* and related microbes. At the commencement of my project, there were no large-scale models of metabolism or transcriptional regulation for *R. sphaeroides* or any closely related organism. Furthermore, knowledge of the regulatory networks employed by *R. sphaeroides* was limited, focused mostly on those involved in regulation of photosynthesis and stress responses.

Thus, I began my research by building a manually curated genome-scale metabolic model for *R. sphaeroides*, which I named iRsp1095, consisting of 1,095 genes, 796 metabolites and 1158 reactions (Chapter 2). iRsp1095 also included experimentally determined *R. sphaeroides*-specific biomass reactions for simulating aerobic and anaerobic growth. In addition to serving as a knowledgebase for all known and predicted metabolic capabilities of *R. sphaeroides*, iRsp1095 allowed genome-scale predictions of metabolic fluxes during aerobic respiratory, anaerobic respiratory, photoheterotrophic and photoautotrophic growth modes. iRsp1095 also allowed accurate predictions of growth rate under these various conditions while providing useful predictions for the production of important metabolites such as H_2 and PHB. Using data I generated from high-throughput phenotype analysis of *R. sphaeroides* cells, I refined and extended iRsp1095, generating an updated metabolic model, which I named iRsp1140, with significantly improved predictive capabilities and coverage of *R. sphaeroides* metabolic repertoire (Chapter 3). iRsp1140 consists of 1416 reactions, 878 metabolites and accounts for 1140 genes. To further improve the predictive capability of these models, the photosynthetic maintenance energy requirements of *R. sphaeroides* at different light intensities were determined and inclusion of these parameters, as well as the measured light uptake rates, significantly improved iRsp1140's predictions,

while considerably shrinking optimal solution space (Chapter 4). iRsp1140 growth rate predictions currently show a high correlation to experimental observations ($R > 0.92$).

I also constructed a large-scale TRN for *R. sphaeroides*, implementing a newly developed reconstruction workflow that leveraged available gene expression data from *R. sphaeroides*, sequence information from closely related bacteria and intrinsic properties of bacterial TFs (Chapter 5). The resulting large-scale TRN model included 120 clusters consisting of 1211 genes (including 93 TFs), 1858 regulatory interactions and 76 regulatory motifs. About 67% of the identified gene clusters were significantly enriched for cellular functions ranging from photosynthesis and central carbon metabolism to environmental stress responses. Furthermore, the features of many of the identified clusters were consistent with known regulatory interactions in *R. sphaeroides* and/or other bacteria. I validated and refined predictions from this large-scale TRN by conducting genome-wide ChIP-seq and gene expression analysis for 9 transcription factors, which included 6 involved in photosynthesis and maintenance of iron homeostasis (PpsR, FnrL, PrrA, CrpK, RSP_2888 and RSP_3341) and 3 involved in regulation of carbon metabolism (RSP_0489, CcmR and AkgR). Of these, 6 TFs (CrpK, RSP_2888, RSP_3341, RSP_0489, CcmR and AkgR) were entirely novel predictions from the TRN. Thus, my experimental analysis represent the first characterizations of these proteins. In addition, my predictions and genomic analyses resulted in the significant expansion of the PpsR, FnrL and PrrA regulons (Chapters 5 – 7). Data from these experimental analyses were used to refine and extend the TRN.

To improve our ability to model the lifestyles of *R.* sphaeroides, I used 2 approaches to integrate iRsp1140 with the large-scale TRN, rFBA and PROM. I assessed the performance of these integrated models of metabolism and transcriptional regulation using a limited dataset of growth and gene deletion phenotypes available for *R. sphaeroides* and compared these to predictions made from iRsp1140 alone. Predictions from the integrated models showed good agreement with known growth and gene deletion phenotypes, while providing increased scope compared to iRsp1140 alone (Chapter 8). However, differences existed between predictions from each of the integrated models and further analysis of these

models with a more extensive dataset may be needed to determine the best approach to use for integration given the amount of data currently available.

Overall, I accomplished the major goals which I envisaged at the start of my research, developing useful tools for computational modeling of metabolism and transcriptional regulation, while generating a wealth of new experimental data. I believe all of these analyses have given the scientific community a significant amount of new, systems-level insight into the complexity and diversity of the physiology of *R. sphaeroides* and related bacteria.

References

1. Stelling J: **Mathematical models in microbial systems biology**. *Curr Opin Microbiol* 2004, **7**(5):513-518.
2. Lee SY: **Systems biology and biotechnology of *Escherichia coli***: Springer; 2009.
3. Mackenzie C, Eraso JM, Choudhary M, Roh JH, Zeng X, Bruscella P, Puskas A, Kaplan S: **Postgenomic adventures with *Rhodobacter sphaeroides***. *Annu Rev Microbiol* 2007, **61**:283-307.
4. Imam S, Yilmaz S, Sohmen U, Gorzalski AS, Reed JL, Noguera DR, Donohue TJ: **iRsp1095: a genome-scale reconstruction of the *Rhodobacter sphaeroides* metabolic network**. *BMC Syst Biol* 2011, **5**:116.
5. Yilmaz LS, Kontur WS, Sanders AP, Sohmen U, Donohue TJ, Noguera DR: **Electron partitioning during light- and nutrient-powered hydrogen production by *Rhodobacter sphaeroides***. *Bioenerg Res* 2010, **Volume**(1):55 - 66.
6. Kien NB, Kong IS, Lee MG, Kim JK: **Coenzyme Q10 production in a 150-l reactor by a mutant strain of *Rhodobacter sphaeroides***. *J Ind Microbiol Biotechnol* 2010, **37**(5):521-529.
7. Khatipov E, Miyake, M., Miyake J. and Y. Asada: **Polyhydroxybutyrate accumulation and hydrogen evolution by *Rhodobacter sphaeroides* as a function of nitrogen availability**. *Biohydrogen* 1999, **III**:157 - 161.
8. Sasaki K, Morikawa H, Kishibe T, Mikami A, Harada T, Ohta M: **Practical removal of radioactivity from sediment mud in a swimming pool in Fukushima, Japan by immobilized photosynthetic bacteria**. *Biosci Biotechnol Biochem* 2012, **76**(4):859-862.
9. Ind AC, Porter SL, Brown MT, Byles ED, de Beyer JA, Godfrey SA, Armitage JP: **Inducible-expression plasmid for *Rhodobacter sphaeroides* and *Paracoccus denitrificans***. *Appl Environ Microbiol* 2009, **75**(20):6613-6615.
10. Schafer A, Tauch A, Jager W, Kalinowski J, Thierbach G, Puhler A: **Small mobilizable multi-purpose cloning vectors derived from the *Escherichia coli* plasmids pK18 and pK19: selection of defined deletions in the chromosome of *Corynebacterium glutamicum***. *Gene* 1994, **145**(1):69-73.
11. Simon R, Priefer U, Puhler A: **A broad host range mobilization system for *in vivo* genetic engineering: transposon mutagenesis in Gram-negative bacteria**. *Nat Biotechnol* 1983, **1**:784 - 791.
12. Pappas CT, Sram J, Moskvin OV, Ivanov PS, Mackenzie RC, Choudhary M, Land ML, Larimer FW, Kaplan S, Gomelsky M: **Construction and validation of the *Rhodobacter sphaeroides* 2.4.1 DNA microarray: transcriptome flexibility at diverse growth modes**. *J Bacteriol* 2004, **186**(14):4748-4758.
13. Dufour YS, Wesenberg GE, Tritt AJ, Glasner JD, Perna NT, Mitchell JC, Donohue TJ: **chipD: a web tool to design oligonucleotide probes for high-density tiling arrays**. *Nucleic Acids Res* 2010, **38**(Web Server issue):W321-325.
14. Anthony JR, Warczak KL, Donohue TJ: **A transcriptional response to singlet oxygen, a toxic byproduct of photosynthesis**. *Proc Natl Acad Sci U S A* 2005, **102**(18):6502-6507.
15. Dufour YS, Landick R, Donohue TJ: **Organization and evolution of the biological response to singlet oxygen stress**. *J Mol Biol* 2008, **383**(3):713-730.
16. Eraso JM, Roh JH, Zeng X, Callister SJ, Lipton MS, Kaplan S: **Role of the global transcriptional regulator PrrA in *Rhodobacter sphaeroides* 2.4.1: combined transcriptome and proteome analysis**. *J Bacteriol* 2008, **190**(14):4831-4848.
17. Gomelsky L, Moskvin OV, Stenzel RA, Jones DF, Donohue TJ, Gomelsky M: **Hierarchical regulation of photosynthesis gene expression by the oxygen-responsive PrrBA and AppA-PpsR systems of *Rhodobacter sphaeroides***. *J Bacteriol* 2008, **190**(24):8106-8114.

18. Arai H, Roh JH, Kaplan S: **Transcriptome dynamics during the transition from anaerobic photosynthesis to aerobic respiration in *Rhodobacter sphaeroides* 2.4.1**. *J Bacteriol* 2008, **190**(1):286-299.
19. Kontur WS, Ziegelhoffer EC, Spero MA, Imam S, Noguera DR, Donohue TJ: **Pathways involved in reductant distribution during photobiological H_2 production by *Rhodobacter sphaeroides***. *Appl Environ Microbiol* 2011, **77**(20):7425-7429.
20. Dufour YS, Kiley PJ, Donohue TJ: **Reconstruction of the core and extended regulons of global transcription factors**. *PLoS Genet* 2010, **6**(7):e1001027.
21. Kontur WS, Schackwitz WS, Ivanova N, Martin J, Labutti K, Deshpande S, Tice HN, Pennacchio C, Sodergren E, Weinstock GM *et al*: **Revised sequence and annotation of the *Rhodobacter sphaeroides* 2.4.1 genome**. *J Bacteriol* 2012, **194**(24):7016-7017.
22. Allen JF: **Cyclic, pseudocyclic and noncyclic photophosphorylation: new links in the chain**. *Trends Plant Sci* 2003, **8**(1):15-19.
23. McEwan AG: **Photosynthetic electron transport and anaerobic metabolism in purple non-sulfur phototrophic bacteria**. *Antonie Van Leeuwenhoek* 1994, **66**(1-3):151-164.
24. Blankenship RE, Madigan MT, Bauer CE: **Anoxygenic photosynthetic bacteria**, vol. XLVIII; 1995.
25. Blankenship RE, Hartman H: **The origin and evolution of oxygenic photosynthesis**. *Trends Biochem Sci* 1998, **23**(3):94-97.
26. Chory J, Donohue TJ, Varga AR, Staehelin LA, Kaplan S: **Induction of the photosynthetic membranes of *Rhodopseudomonas sphaeroides*: biochemical and morphological studies**. *J Bacteriol* 1984, **159**(2):540-554.
27. Kiley PJ, Kaplan S: **Molecular genetics of photosynthetic membrane biosynthesis in *Rhodobacter sphaeroides***. *Microbiol Rev* 1988, **52**(1):50-69.
28. Gomelsky M, Kaplan S: **Genetic evidence that PpsR from *Rhodobacter sphaeroides* 2.4.1 functions as a repressor of *puc* and *bchF* expression**. *J Bacteriol* 1995, **177**(6):1634-1637.
29. Gomelsky M, Kaplan S: ***appA*, a novel gene encoding a trans-acting factor involved in the regulation of photosynthesis gene expression in *Rhodobacter sphaeroides* 2.4.1**. *J Bacteriol* 1995, **177**(16):4609-4618.
30. Gomelsky M, Kaplan S: **AppA, a redox regulator of photosystem formation in *Rhodobacter sphaeroides* 2.4.1, is a flavoprotein. Identification of a novel fad binding domain**. *J Biol Chem* 1998, **273**(52):35319-35325.
31. Gomelsky M, Kaplan S: **Molecular genetic analysis suggesting interactions between AppA and PpsR in regulation of photosynthesis gene expression in *Rhodobacter sphaeroides* 2.4.1**. *J Bacteriol* 1997, **179**(1):128-134.
32. Masuda S, Bauer CE: **AppA is a blue light photoreceptor that antirepresses photosynthesis gene expression in *Rhodobacter sphaeroides***. *Cell* 2002, **110**(5):613-623.
33. Bruscella P, Eraso JM, Roh JH, Kaplan S: **The use of chromatin immunoprecipitation to define PpsR binding activity in *Rhodobacter sphaeroides* 2.4.1**. *J Bacteriol* 2008, **190**(20):6817-6828.
34. Zeilstra-Ryalls JH, Kaplan S: **Aerobic and anaerobic regulation in *Rhodobacter sphaeroides* 2.4.1: the role of the fnrL gene**. *J Bacteriol* 1995, **177**(22):6422-6431.
35. Zeilstra-Ryalls JH, Kaplan S: **Role of the *fnrL* gene in photosystem gene expression and photosynthetic growth of *Rhodobacter sphaeroides* 2.4.1**. *J Bacteriol* 1998, **180**(6):1496-1503.
36. Eraso JM, Kaplan S: ***prrA*, a putative response regulator involved in oxygen regulation of photosynthesis gene expression in *Rhodobacter sphaeroides***. *J Bacteriol* 1994, **176**(1):32-43.
37. Dangel AW, Tabita FR: **Protein-protein interactions between CbbR and RegA (PrrA), transcriptional regulators of the cbb operons of *Rhodobacter sphaeroides***. *Mol Microbiol* 2009, **71**(3):717-729.

38. Ranson-Olson B, Jones DF, Donohue TJ, Zeilstra-Ryalls JH: ***In vitro* and *in vivo* analysis of the role of PrrA in *Rhodobacter sphaeroides* 2.4.1 *hemA* gene expression**. *J Bacteriol* 2006, **188**(9):3208-3218.
39. Mank NN, Berghoff BA, Hermanns YN, Klug G: **Regulation of bacterial photosynthesis genes by the small noncoding RNA PcrZ**. *Proc Natl Acad Sci U S A* 2012, **109**(40):16306-16311.
40. Masepohl B, Hallenbeck PC: **Nitrogen and molybdenum control of nitrogen fixation in the phototrophic bacterium *Rhodobacter capsulatus***. *Adv Exp Med Biol* 2010, **675**:49-70.
41. Hallenbeck PC: **Recent advances in phototrophic prokaryotes**. New York: Springer; 2010.
42. Drepper T, Gross S, Yakunin AF, Hallenbeck PC, Masepohl B, Klipp W: **Role of GlnB and GlnK in ammonium control of both nitrogenase systems in the phototrophic bacterium *Rhodobacter capsulatus***. *Microbiology* 2003, **149**(Pt 8):2203-2212.
43. Kim E, Lee M, Kim M, Lee JK: **Molecular hydrogen production by nitrogenase of *Rhodobacter sphaeroides* and by Fe-only hydrogenase of *Rhodospirillum rubrum***. *International Journal of Hydrogen Energy* 2008, **33**(5):1516-1521.
44. Yakunin AF, Hallenbeck PC: **Short-term regulation of nitrogenase activity by NH_4^+ in *Rhodobacter capsulatus*: multiple *in vivo* nitrogenase responses to NH_4^+ addition**. *J Bacteriol* 1998, **180**(23):6392-6395.
45. Dufour YS, Imam S, Koo BM, Green HA, Donohue TJ: **Convergence of the transcriptional responses to heat shock and singlet oxygen stresses**. *PLoS Genet* 2012, **8**(9):e1002929.
46. Nuss AM, Glaeser J, Berghoff BA, Klug G: **Overlapping alternative sigma factor regulons in the response to singlet oxygen in *Rhodobacter sphaeroides***. *J Bacteriol* 2010, **192**(10):2613-2623.
47. Ziegelhoffer EC, Donohue TJ: **Bacterial responses to photo-oxidative stress**. *Nat Rev Microbiol* 2009, **7**(12):856-863.
48. Borland CF, McGarvey DJ, Truscott TG, Cogdell RJ, Land EJ: **Photophysical studies of bacteriochlorophyll a and bacteriopheophytin a – singlet oxygen generation**. *J Photochem Photobiol* 1987, **1**:93–101.
49. Anthony JR, Newman JD, Donohue TJ: **Interactions between the *Rhodobacter sphaeroides* ECF sigma factor, sigma(E), and its anti-sigma factor, ChrR**. *J Mol Biol* 2004, **341**(2):345-360.
50. Greenwell R, Nam TW, Donohue TJ: **Features of *Rhodobacter sphaeroides* ChrR required for stimuli to promote the dissociation of sigma(E)/ChrR complexes**. *J Mol Biol* 2011, **407**(4):477-491.
51. Armitage JP, Macnab RM: **Unidirectional, intermittent rotation of the flagellum of *Rhodobacter sphaeroides***. *J Bacteriol* 1987, **169**(2):514-518.
52. Porter SL, Wadhams GH, Armitage JP: **Signal processing in complex chemotaxis pathways**. *Nat Rev Microbiol* 2011, **9**(3):153-165.
53. Porter SL, Warren AV, Martin AC, Armitage JP: **The third chemotaxis locus of *Rhodobacter sphaeroides* is essential for chemotaxis**. *Mol Microbiol* 2002, **46**(4):1081-1094.
54. Ward MJ, Bell AW, Hamblin PA, Packer HL, Armitage JP: **Identification of a chemotaxis operon with two *cheY* genes in *Rhodobacter sphaeroides***. *Mol Microbiol* 1995, **17**(2):357-366.
55. Hamblin PA, Maguire BA, Grishanin RN, Armitage JP: **Evidence for two chemosensory pathways in *Rhodobacter sphaeroides***. *Mol Microbiol* 1997, **26**(5):1083-1096.
56. Martin AC, Gould M, Byles E, Roberts MA, Armitage JP: **Two chemosensory operons of *Rhodobacter sphaeroides* are regulated independently by sigma 28 and sigma 54**. *J Bacteriol* 2006, **188**(22):7932-7940.
57. Wilkinson DA, Chacko SJ, Venien-Bryan C, Wadhams GH, Armitage JP: **Regulation of flagellum number by FliA and FlgM and role in biofilm formation by *Rhodobacter sphaeroides***. *J Bacteriol* 2011, **193**(15):4010-4014.
58. Wooley JC, Herbert SL: **Catalyzing inquiry at the interface of computing and biology**. Washington, DC: The National Academies Press; 2005.

59. Ingalls BP: **Mathematical modeling in systems biology**. Cambridge, MA: The MIT Press; 2013.
60. Rizzuti B, Daggett V: **Using simulations to provide the framework for experimental protein folding studies**. *Arch Biochem Biophys* 2013, **531**(1-2):128-135.
61. McCloskey D, Palsson BO, Feist AM: **Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli***. *Mol Syst Biol* 2013, **9**:661.
62. Bonneau R, Facciotti MT, Reiss DJ, Schmid AK, Pan M, Kaur A, Thorsson V, Shannon P, Johnson MH, Bare JC *et al*: **A predictive model for transcriptional control of physiology in a free living cell**. *Cell* 2007, **131**(7):1354-1365.
63. Thiele I, Jamshidi N, Fleming RM, Palsson BO: **Genome-scale reconstruction of *Escherichia coli*'s transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization**. *PLoS Comput Biol* 2009, **5**(3):e1000312.
64. Janes KA, Yaffe MB: **Data-driven modelling of signal-transduction networks**. *Nat Rev Mol Cell Biol* 2006, **7**(11):820-828.
65. Kestler HA, Wawra C, Kracher B, Kuhl M: **Network modeling of signal transduction: establishing the global view**. *Bioessays* 2008, **30**(11-12):1110-1125.
66. Chandrasekaran S, Price ND: **Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia coli* and *Mycobacterium tuberculosis***. *Proc Natl Acad Sci U S A* 2011, **107**(41):17845-17850.
67. Covert MW, Palsson BO: **Transcriptional regulation in constraints-based metabolic models of *Escherichia coli***. *J Biol Chem* 2002, **277**(31):28058-28064.
68. Lerman JA, Hyduke DR, Latif H, Portnoy VA, Lewis NE, Orth JD, Schrimpe-Rutledge AC, Smith RD, Adkins JN, Zengler K *et al*: **In silico method for modelling metabolism and gene product expression at genome scale**. *Nat Commun* 2012, **3**:929.
69. Shlomi T, Eisenberg Y, Sharan R, Ruppin E: **A genome-scale computational study of the interplay between transcriptional regulation and metabolism**. *Mol Syst Biol* 2007, **3**:101.
70. Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B, Jr., Assad-Garcia N, Glass JI, Covert MW: **A whole-cell computational model predicts phenotype from genotype**. *Cell* 2012, **150**(2):389-401.
71. Ishii N, Robert M, Nakayama Y, Kanai A, Tomita M: **Toward large-scale modeling of the microbial cell for computer simulation**. *J Biotechnol* 2004, **113**(1-3):281-294.
72. Markram H: **Human Brain Project**. In.: https://www.humanbrainproject.eu.
73. Waldrop MM: **Computer modelling: Brain in a box**. *Nature* 2012, **482**(7386):456-458.
74. Palyanov A, Khayrulin S, Larson SD, Dibert A: **Towards a virtual *C. elegans*: a framework for simulation and visualization of the neuromuscular system in a 3D physical environment**. *In Silico Biol* 2012, **11**(3-4):137-147.
75. Oberhardt MA, Palsson BO, Papin JA: **Applications of genome-scale metabolic reconstructions**. *Mol Syst Biol* 2009, **5**:320.
76. Schallau K, Junker BH: **Simulating plant metabolic pathways with enzyme-kinetic models**. *Plant Physiol* 2010, **152**(4):1763-1771.
77. Tan Y, Liao JC: **Metabolic ensemble modeling for strain engineers**. *Biotechnol J* 2012, **7**(3):343-353.
78. Tran LM, Rizk ML, Liao JC: **Ensemble modeling of metabolic networks**. *Biophys J* 2008, **95**(12):5606-5617.
79. Palsson B: **The challenges of in silico biology**. *Nat Biotechnol* 2000, **18**(11):1147-1150.
80. Varma A, Palsson BO: **Metabolic flux balancing: basic concepts, scientific and practical use.** *Nature Biotechnology* 1994, **12**:994 - 998.
81. Thiele I, Palsson BO: **A protocol for generating a high-quality genome-scale metabolic reconstruction**. *Nat Protoc* 2010, **5**(1):93-121.
82. Feist AM, Palsson BO: **The biomass objective function**. *Curr Opin Microbiol* 2010, **13**(3):344-349.

83. Mahadevan R, Schilling CH: **The effects of alternate optimal solutions in constraint-based genome-scale metabolic models**. *Metab Eng* 2003, **5**(4):264-276.
84. Lee S, Phalakornkule, C., Domach, M.M., and Grossmann, I.E.: **Recursive MILP model for finding all the alternate optima in LP models for metabolic networks**. *Computers & Chemical Engineering* 2000, **24**(2 - 7):711 -716.
85. Reed JL, Palsson BO: **Genome-scale in silico models of *E. coli* have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states**. *Genome Res* 2004, **14**(9):1797-1805.
86. Papin JA, Price ND, Palsson BO: **Extreme pathway lengths and reaction participation in genome-scale metabolic networks**. *Genome Res* 2002, **12**(12):1889-1900.
87. Colijn C, Brandes A, Zucker J, Lun DS, Weiner B, Farhat MR, Cheng TY, Moody DB, Murray M, Galagan JE: **Interpreting expression data with metabolic flux models: predicting *Mycobacterium tuberculosis* mycolic acid production**. *PLoS Comput Biol* 2009, **5**(8):e1000489.
88. Shlomi T, Cabili MN, Herrgard MJ, Palsson BO, Ruppin E: **Network-based prediction of human tissue-specific metabolism**. *Nat Biotechnol* 2008, **26**(9):1003-1010.
89. Kim J, Reed JL: **RELATCH: relative optimality in metabolic networks explains robust metabolic and regulatory responses to perturbations**. *Genome Biol* 2012, **13**(9):R78.
90. Becker SA, Palsson BO: **Context-specific metabolic networks are consistent with experiments**. *PLoS Comput Biol* 2008, **4**(5):e1000082.
91. Yizhak K, Benyamini T, Liebermeister W, Ruppin E, Shlomi T: **Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model**. *Bioinformatics* 2010, **26**(12):i255-260.
92. Jensen PA, Papin JA: **Functional integration of a metabolic network model and expression data without arbitrary thresholding**. *Bioinformatics* 2011, **27**(4):541-547.
93. Segre D, Vitkup D, Church GM: **Analysis of optimality in natural and perturbed metabolic networks**. *Proc Natl Acad Sci U S A* 2002, **99**(23):15112-15117.
94. Shlomi T, Berkman O, Ruppin E: **Regulatory on/off minimization of metabolic flux changes after genetic perturbations**. *Proc Natl Acad Sci U S A* 2005, **102**(21):7695-7700.
95. Alper H, Jin YS, Moxley JF, Stephanopoulos G: **Identifying gene targets for the metabolic engineering of lycopene biosynthesis in *Escherichia coli***. *Metab Eng* 2005, **7**(3):155-164.
96. Park JH, Lee KH, Kim TY, Lee SY: **Metabolic engineering of *Escherichia coli* for the production of L-valine based on transcriptome analysis and in silico gene knockout simulation**. *Proc Natl Acad Sci U S A* 2007, **104**(19):7797-7802.
97. Burgard AP, Pharkya P, Maranas CD: **Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization**. *Biotechnol Bioeng* 2003, **84**(6):647-657.
98. Patil KR, Rocha I, Forster J, Nielsen J: **Evolutionary programming as a platform for *in silico* metabolic engineering**. *BMC Bioinformatics* 2005, **6**:308.
99. Kim J, Reed JL: **OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains**. *BMC Syst Biol* 2010, **4**:53.
100. Pharkya P, Burgard AP, Maranas CD: **OptStrain: a computational framework for redesign of microbial production systems**. *Genome Res* 2004, **14**(11):2367-2376.
101. De Smet R, Marchal K: **Advantages and limitations of current network inference methods**. *Nat Rev Microbiol* 2010, **8**(10):717-729.
102. Joshi A, De Smet R, Marchal K, Van de Peer Y, Michoel T: **Module networks revisited: computational assessment and prioritization of model predictions**. *Bioinformatics* 2009, **25**(4):490-496.
103. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data**. *Nat Genet* 2003, **34**(2):166-176.

104. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS: **Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles**. *PLoS Biol* 2007, **5**(1):e8.
105. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P: **Inferring regulatory networks from expression data using tree-based methods**. *PLoS One* 2010, **5**(9).
106. Kuffner R, Petri T, Tavakkolkhah P, Windhager L, Zimmer R: **Inferring gene regulatory networks by ANOVA**. *Bioinformatics* 2012, **28**(10):1376-1382.
107. Stolovitzky G, Monroe D, Califano A: **Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference**. *Ann N Y Acad Sci* 2007, **1115**:1-22.
108. Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Stolovitzky G: **Wisdom of crowds for robust gene network inference**. *Nat Methods* 2012, **9**(8):796-804.
109. Reiss DJ, Baliga NS, Bonneau R: **Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks**. *BMC Bioinformatics* 2006, **7**:280.
110. Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, Thorsson V: **The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets *de novo***. *Genome Biol* 2006, **7**(5):R36.
111. Ernst J, Beg QK, Kay KA, Balazsi G, Oltvai ZN, Bar-Joseph Z: **A semi-supervised method for predicting transcription factor-gene interactions in *Escherichia coli***. *PLoS Comput Biol* 2008, **4**(3):e1000044.
112. Lemmens K, De Bie T, Dhollander T, De Keersmaecker SC, Thijs IM, Schoofs G, De Weerdt A, De Moor B, Vanderleyden J, Collado-Vides J *et al*: **DISTILLER: a data integration framework to reveal condition dependency of complex regulons in *Escherichia coli***. *Genome Biol* 2009, **10**(3):R27.
113. Novichkov PS, Rodionov DA, Stavrovskaya ED, Novichkova ES, Kazakov AE, Gelfand MS, Arkin AP, Mironov AA, Dubchak I: **RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach**. *Nucleic Acids Res* 2010, **38**(Web Server issue):W299-307.
114. Rodionov DA: **Comparative genomic reconstruction of transcriptional regulatory networks in bacteria**. *Chem Rev* 2007, **107**(8):3467-3497.
115. Wang T, Stormo GD: **Identifying the conserved network of cis-regulatory sites of a eukaryotic genome**. *Proc Natl Acad Sci U S A* 2005, **102**(48):17400-17405.
116. Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, Stolovitzky G: **Revealing strengths and weaknesses of methods for gene network inference**. *Proc Natl Acad Sci U S A* 2010, **107**(14):6286-6291.
117. Siegenthaler C, Gunawan R: **Assessment of network inference methods: how to cope with an underdetermined problem**. *PLoS One* 2014, **DOI: 10.1371/journal.pone.0090481**.
118. Gama-Castro S, Salgado H, Peralta-Gil M, Santos-Zavaleta A, Muniz-Rascado L, Solano-Lira H, Jimenez-Jacinto V, Weiss V, Garcia-Sotelo JS, Lopez-Fuentes A *et al*: **RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units)**. *Nucleic Acids Res* 2011, **39**(Database issue):D98-105.
119. Sierro N, Makita Y, de Hoon M, Nakai K: **DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information**. *Nucleic Acids Res* 2008, **36**(Database issue):D93-96.

Chapter 2

iRsp1095: A genome-scale reconstruction of the *Rhodobacter sphaeroides* metabolic network.

This chapter is published under the same title:

Imam S, Yilmaz S, Sohmen U, Gorzalski AS, Reed JL, Noguera DR, Donohue TJ. BMC Syst Biol. 2011 Jul 21;5:116. doi: 10.1186/1752-0509-5-116. PMID: 21777427.

SI constructed and curated iRsp1095. SI and LSY participated in assessment of iRsp1095. US, ASG and LSY set up *R. sphaeroides* continuous cultures and obtained analytical data. SI and LSY participated in the determination of the *R. sphaeroides* biomass composition.

Abstract

Background

Rhodobacter sphaeroides is one of the best studied purple non-sulfur photosynthetic bacteria and serves as an excellent model for the study of photosynthesis and the metabolic capabilities of this and related facultative organisms. The ability of *R. sphaeroides* to produce hydrogen (H_2), polyhydroxybutyrate (PHB) or other hydrocarbons, as well as its ability to utilize atmospheric carbon dioxide (CO_2) as a carbon source under defined conditions, make it an excellent candidate for use in a wide variety of biotechnological applications. A genome-level understanding of its metabolic capabilities should help realize this biotechnological potential.

Results

Here we present a genome-scale metabolic network model for *R. sphaeroides* strain 2.4.1, designated iRsp1095, consisting of 1,095 genes, 796 metabolites and 1158 reactions, including *R. sphaeroides*-specific biomass reactions developed in this study. Constraint-based analysis showed that iRsp1095 agreed well with experimental observations when modeling growth under respiratory and phototrophic conditions. Genes essential for phototrophic growth were predicted by single gene deletion analysis. During pathway-level analyses of *R. sphaeroides* metabolism, an alternative route for CO_2 assimilation was identified. Evaluation of photoheterotrophic H_2 production using iRsp1095 indicated that maximal yield would be obtained from growing cells, with this predicted maximum ~50% higher than that observed experimentally from wild type cells. Competing pathways that might prevent the achievement of this theoretical maximum were identified to guide future genetic studies.

Conclusions

iRsp1095 provides a robust framework for future metabolic engineering efforts to optimize the solar- and nutrient-powered production of biofuels and other valuable products by *R. sphaeroides* and closely related organisms.

Introduction

Photosynthetic organisms perform many functions of significance to the planet and society. Plants and photosynthetic microbes are responsible for harvesting solar energy, evolving oxygen and sequestering atmospheric carbon dioxide [1]. In addition, algae, cyanobacteria and photosynthetic bacteria are either naturally able to or have been modified to evolve hydrogen (H_2), accumulate oils and hydrocarbons, or produce alcohols or other compounds that can reduce society's dependence on fossil fuels [2, 3]. The ability to understand, capitalize on or improve these activities is limited by our knowledge of the metabolic blueprint of photosynthetic organisms. To fill this knowledge gap, we are modeling the flow of carbon and reducing power in the well-studied photosynthetic bacterium *Rhodobacter sphaeroides*. This facultative bacterium is capable of either aerobic or anaerobic respiration, depending on the availability of oxygen (O_2) or alternative electron acceptors. When O_2 is absent or limiting, light energy can be harnessed by a photosynthetic electron transport chain that has features similar to those used by plants and other oxygen-evolving phototrophs [1]. During photosynthetic growth, *R. sphaeroides* is capable of autotrophic or heterotrophic growth using either carbon dioxide (CO_2) or organic carbon sources [4, 5]. Thus, it provides an ideal system for studying the details of each lifestyle and the mechanisms of transition between these various metabolic states.

R. sphaeroides has also received significant attention due to its biotechnological potential, with its ability to produce large amounts of carotenoids or isoprenoids as a source of biocommodities, H_2 as a potential biofuel, or polyhydroxybutyrate (PHB) as raw material for biodegradable plastics [6]. Furthermore, the autotrophic metabolism of *R. sphaeroides* makes it a potential organism for use in the synthesis of chemicals or polymers that can serve as raw materials in the production of biofuels, or as a means of sequestering atmospheric or industrially-produced CO_2 [2]. To understand and tap into the activities or products of this photosynthetic bacterium, detailed knowledge of its metabolic pathways is necessary. To provide this knowledge, we are generating computational models of the metabolic network of *R.*

sphaeroides that are based on genomic information, which can be informed and integrated with laboratory analysis of wild type and mutant strains [3, 7].

Over the last decade the field of constraint-based metabolic modeling has witnessed significant progress, which has led to major advances in the modeling, understanding and engineering of different biological systems [8-11]. As a consequence, high quality genome-scale metabolic reconstructions have been generated for many organisms [9]. These reconstructions serve both as structured databases of all the known and/or predicted metabolic functions of an organism and as the basis for the construction of mathematical models used in constraint-based analysis. The ability of constraint-based analyses to provide new biological insights has the potential to increase with the influx of high-throughput biological data sets [8, 9]. Thus far, genome-scale reconstructions have been published for only one photosynthetic microbe, the oxygenic cyanobacterium *Synechocystis sp.* PCC 6803 [12-14]. Models of photosynthetic electron transport [15] and small scale *R. sphaeroides* metabolic networks that use flux balance analysis (FBA) [16] and ensemble modeling [17] have also been published.

Here we present iRsp1095, a manually curated genome-scale metabolic reconstruction for *R. sphaeroides* strain 2.4.1 consisting of 796 metabolites, 858 transformation reactions and 300 transport reactions. The reconstruction includes 1,095 genes, covering about 25% of the recognized *R. sphaeroides* open reading frames. To facilitate improved predictions, the biomass composition of *R. sphaeroides* was determined under a variety of growth conditions and used in generating biomass objective functions suitable for developing predictive models. FBA [18-20], flux variability analysis (FVA) [21] and alternate optima analysis [22, 23] were used to predict metabolic fluxes under chemoheterotrophic (aerobic respiration), photoheterotropic and photoautotrophic (anaerobic) growth conditions. The predictive ability of iRsp1095 was validated by comparison with experimentally determined growth rate and fluxes of key metabolic products from continuous cultures. iRsp1095 was also used to predict metabolic flux distributions through key pathways including CO_2 fixation and the electron transport chain. Overall, iRsp1095 shows good qualitative and quantitative agreement with experimental observations. Thus, iRsp1095 provides concepts

and a basis for extensive future studies of this bacterium, other related bacteria and photosynthetic organisms in general.

Results

Model Reconstruction

The initial *R. sphaeroides* metabolic network was constructed by extracting genomic and metabolic information from KEGG [24], and combining this with results from metaSHARK [25] analysis (see Additional File 1 for details). We assigned directions to reactions in the network via a combination of thermodynamic and heuristic calculations/assumptions, which have been used previously [26] (see Additional File 1). The *R. sphaeroide*s model was further analyzed for stoichiometrically balanced cycles (SBCs) – internal network loops that carry flux in a closed system (i.e., when all exchange reactions are closed) with no net production or consumption of metabolites [20, 27]. SBCs were manually eliminated from the network leading to the assignment of directionality to an additional 29 reactions in the network (see Additional File 1, Additional File 2 – Table S4). The remaining 150 (13%) reactions for which there was insufficient thermodynamic information were assigned as reversible. The directionality assignments in iRsp1095 are summarized in Table 2-1.

Table 2-1. Summary of the reaction directionality assignments in the model

	Number in each group	% of total Reactions
Total Irreversible	**401**	**35**
Thermodynamics only	109	9
Heuristics + Thermodynamics	125	11
ABC Transporter/tRNA charging	93	8
Spontaneous	8	1
Others*	66	6
Total Reversible	**757**	**65**
Thermodynamics/Heuristics	607	52
Unknown	150	13
Total no. of reactions	**1158**	

* Others includes groups of reactions assigned as irreversible based on SBC analysis or literature (e.g. other databases)

Gaps in the initial reconstruction, representing limitations in our current understanding of *R. sphaeroides* metabolism, were identified and filled (see Additional File 1). This process led to the addition of 30 transformation and 65 transport reactions to the network (see Additional File 2 – Table S11) and produced a model capable of predicting the production of biomass under defined conditions. FVA analysis with a completely open system (i.e., all exchange reactions allowed to carry flux) showed 140 blocked reactions remained at this stage, but these generally involved reactions (or pathways) required for the biosynthesis of low abundance end products (minor carotenoids and phospholipids) that are not considered as part of our biomass objective function. Thus, these 140 reactions are related to dead ends in iRsp1095.

Formulation of biomass objective function

To obtain qualitative and quantitative outputs from constraint-based modeling using genome-scale models, the use of a meaningful objective function is critical [28]. Currently, the most widely used objective function in constraint-based modeling is the biomass objective function (BOF), as it represents a meaningful, though not necessarily accurate, ultimate goal of a microbial cell. While *R. sphaeroides* is a gram-negative bacterium, and in many respects similar to *E. coli* during aerobic growth, photosynthetic growth requires significant changes in metabolic machinery, and thus biomass composition, most notably in the pigment and lipid composition, as large amounts of chlorophyll or carotenoid pigments and phospholipids are contained in intracytoplasmic membrane (ICM) that houses the photosynthetic apparatus [29]. Thus, to generate representative BOFs for *R. sphaeroides*, we experimentally determined the major macromolecular constituents of aerobically and photosynthetically grown cells (Material and Methods). Based on these experimentally determined macromolecular components (Table 2-2), available genome sequence data [30] and published compositions of fatty acids and lipids [31-37], the BOFs were formulated as weighted combinations of precursors, with coefficients directly related to their percent composition of the biomass [20, 38]. Details of the biomass calculations are contained in Additional File 3. The growth associated maintenance (GAM) energy requirement was estimated as previously described [20].

Table 2-2. Percent composition of cellular biomass of *R. sphaeroides* during photoheterotrophic and aerobic growth*

Components	% Composition of biomass (Photo)	% Composition of biomass (Aero)
DNA	1.9	2.8
RNA	5.1	7.1
Protein	53.6	49.3
Lipids[a]	17.1	12.8
PHB	10.4	17.6
Bacteriochlorophyll	0.4	0
Carotenoids	0.1	0
Glycogen	1.4	0.4
Lipopolysaccharides[b]	3	3
Cell Wall[b]	2	2

* Inorganic fraction estimated to be about 5% of biomass [66].

[a] Lipid composition of *R. sphaeroides* consisting of phospholipids and sulfolipids.

[b] Estimated from *E. coli* cell wall and lipopolysaccharides percent contribution.

Overview of iRsp1095

iRsp1095 consists of 796 unique metabolites, 858 transformation reactions, 300 transport reactions and 148 exchange reactions (Table 2-3). The list of reactions, metabolites, thermodynamic calculations, genes and references used are in Additional File 2. The network is divided into 3 compartments (extracellular, periplasmic and cytoplasmic), with appropriate transport reactions across the outer and inner membranes. Individual metabolites, including cytoplasmic, periplasmic or extracellular instances of a given metabolite, were given reconstruction-specific unique identifiers for internal use, which were mapped to other database identifiers (PubChem, Cas, KEGG and BiGG). The iRsp1095 reconstruction accounts for 1,095 genes representing ~25% of the annotated *R. sphaeroides* open reading frames. Of the 1158 reactions in iRsp1095, 1,049 (90.6%) have gene-protein-reaction (GPR) assignments, with 203 of these having associated experimental data, while 95 (8.2%) of the reactions without GPR assignments correspond to place holder reactions for which a putative gene could not be assigned. The remaining 14 reactions correspond to known spontaneous or diffusion reactions (Table 2-3, see Additional File 2 – Table S1). The breakdown of the sub-system distribution of the reactions is shown in Figure 2-1A. Analysis of the distribution of the gene products in iRsp1095 using cluster of orthologous groups (COGs) classification [39], shows that 13 of the 22 COG categories are significantly enriched for the proteins present in the model (p-value <0.01, hypergeometric test), with amino acid metabolism having the highest number and nucleotide metabolism showing the greatest coverage (Figure 2-1B). The genome-scale reconstruction was converted into a stoichiometric matrix consisting of 796 rows and 1309 columns, including exchange reactions to allow metabolites to be taken up or secreted in to the extracellular space, as well as 3 demand reactions for key metabolites not included in the biomass reaction (PHB, glycogen and minor carotenoids) (Table 2-3). The equivalent SBML format of the model was generated for distribution and potential use in other modeling environments (see Additional File 4). This file has been deposited in the BioModels database [40] (accession: MODEL1106220000).

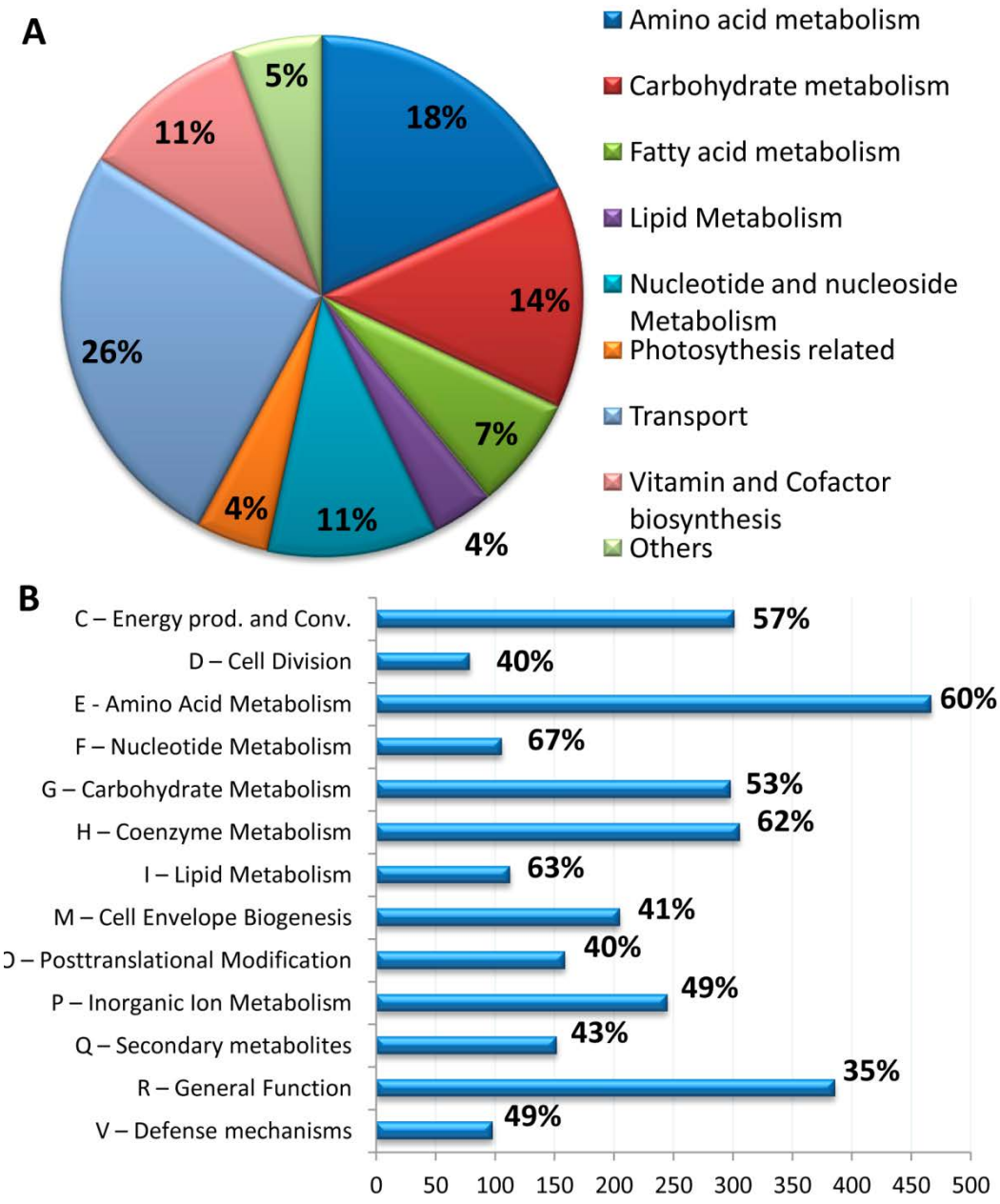

Figure 2-1. Distribution of reactions and gene products in iRsp1095. (A) The pie chart depicts the subsystem distribution of the reactions in iRsp1095, with the percent contribution of each subsystem of reactions indicated in the corresponding section of the chart. It can be seen that amino acid, carbohydrate and nucleotide metabolism dominate the enzymatic reactions present in iRsp1095, while photosynthesis related reactions represent a smaller but significant fraction. (B) The bar chart depicts the distribution of gene products in iRsp1095 based on COG classification with the percent coverage shown for each class. Only COG classes significantly enriched for proteins present in the model ($p < 0.01$, hypergeometric distribution) are shown.

Table 2-3. Overview of iRsp1095

Categories	No. in the reconstruction	
Genes	**1095**	
ORFs	1049	96.0%
tRNA genes	46	4.0%
Metabolites	**1096**	
Unique metabolites	796	
Cytoplasmic	795	
Periplasmic	151	
Extracellular	150	
Reactions	**1158**	
Enzymatic Reactions	858	74.1%
Transport reactions	300	25.9%
Reactions Associated with genes	1049	90.6%
Reactions based on experimental evidence	203	17.5%
Reactions inferred based on gene homology	846	73.1%
Spontaneous/Diffusion reactions	14	1.2%
Reactions without gene association	95	8.2%
Reactions associated with multi-protein complexes	130	11.2%
Reactions associated with isozymes	262	22.6%
Reversible Reactions	757	65.4%
Irreversible Reactions	401	34.6%
Exchange Reactions	148	
Demand Reactions	3	

Model validation

We used FBA and other constraint-based approaches to interrogate the properties of the iRsp1095, with simulations conducted for aerobic respiration, dark anaerobic respiration in the presence of the electron acceptor dimethyl sulfoxide (DMSO) and photoheterotrophic growth (anaerobic growth in the presence of light and an electron-rich carbon source) using Sistrom's minimal media (SIS) [41] containing one of a variety of carbon sources (see Additional File 2 – Table S9). Photoautotrophic growth with CO_2 as the sole source of carbon and H_2 or hydrogen sulfide (H_2S) as the electron donor was also simulated.

Qualitative Assessment of Metabolic model

As a first step in assessing the performance and breadth of iRsp1095, we used FBA to test for the ability of the model to predict the production of biomass and H_2 while supplied with SIS minimal media. The model was capable of predicting growth in the dark in the presence of O_2 or DMSO as known electron acceptors, under photoautotrophic conditions using CO_2 as the sole carbon source and either H_2 or H_2S as electron donor, and photoheterotropically with a variety of organic carbon sources (Table 2-4). In addition, when the ability to utilize the various carbon, nitrogen, phosphorus and sulfur sources present in iRsp1095 was tested, it predicted photosynthetic growth on 129 potential carbon sources, 72 potential nitrogen sources, 46 potential phosphorus sources and 9 potential sulfur sources. While no high throughput phenotypic screens have been conducted for *R. sphaeroides*, growth on 25 of the carbon sources predicted by iRsp1095 to support net biomass formation (~20%) have previously been reported [6, 42, 43] (see Additional File 2 – Table S6), while those carbon sources not yet tested as growth substrates in the literature provide candidates for future validation and correction of the model.

Table 2-4. Growth phenotypes predicted by the model under a variety of routinely utilized laboratory conditions*

	Light	Dark		
		Electron Acceptor		
		O_2[a]	DMSO[a]	None
Succinate + NH_3	+/+[b]	+/-	+/-	-/-
Succinate + Glutamate	+/+[b]	+/-	+/-	-/-
Lactate + NH_3	+/+[b]	+/-	+/-	-/-
Glutamate only	+/-	+/-	+/-	-/-
CO_2 + H_2 + NH_3	+/-	-/-	-/-	-/-
CO_2 + H_2 + N_2	+/-	-/-	-/-	-/-

* +/+ Growth and H_2 production predicted; +/- Growth but no H_2 production predicted; -/- No growth

[a] Oxygen (O_2) or DMSO was used as the sole electron acceptors in simulations.

[b] Succinate and lactate uptake rates were set to 3 mmol/g DW h, while the NH_3 and glutamate uptake rates were set to 1 mmol/g DW h**,** as these are within the rate of experimentally observed uptake rates for these substrates. CO_2, H_2 and N_2 uptake rates were set to 1 mmol/g DW h for photoautotrophic growth simulations.

An extensive set of *R. sphaeroides* mutants does not currently exist for validation of gene knock-out simulations using iRsp1095. However, gene essentiality analysis still allows us to generate hypotheses about genes and reactions that are potentially essential under one or more growth conditions. We used FBA to conduct single reaction and gene deletion analyses during simulations of photoheterotrophic growth using succinate as a carbon source and ammonia as the nitrogen source (with light uptake left unconstrained). Under these conditions, iRsp1095 predicts that a core set of 293 reactions (25% of the network) are essential for growth (Figure 2-2A). Seventy of these "essential" reactions are associated with isozymes and thus would potentially require multiple gene deletions to inactive the cognate pathway. FVA analysis at optimal growth rate, predicts that 415 (36%) of the reactions in the network are capable of carrying flux, but are not essential for growth on minimal media containing succinate as a carbon source. An additional 310 (27%) of the reactions are predicted to be incapable of carrying flux during photoheterotrophic growth on succinate and ammonia and correspond to transport and transformation steps not required under these conditions but could potentially be essential under alternative growth conditions. The remaining 140 (12%) of the reactions in the network cannot carry flux under any of the conditions tested (i.e. blocked reactions). Furthermore, single gene deletion analysis showed that 217 (20%) of the 1095 genes present in iRsp1095 were essential for growth under these conditions (see Additional File 2 – Table S10). The distribution of these gene products based on COG classification is shown in Figure 2-2B.

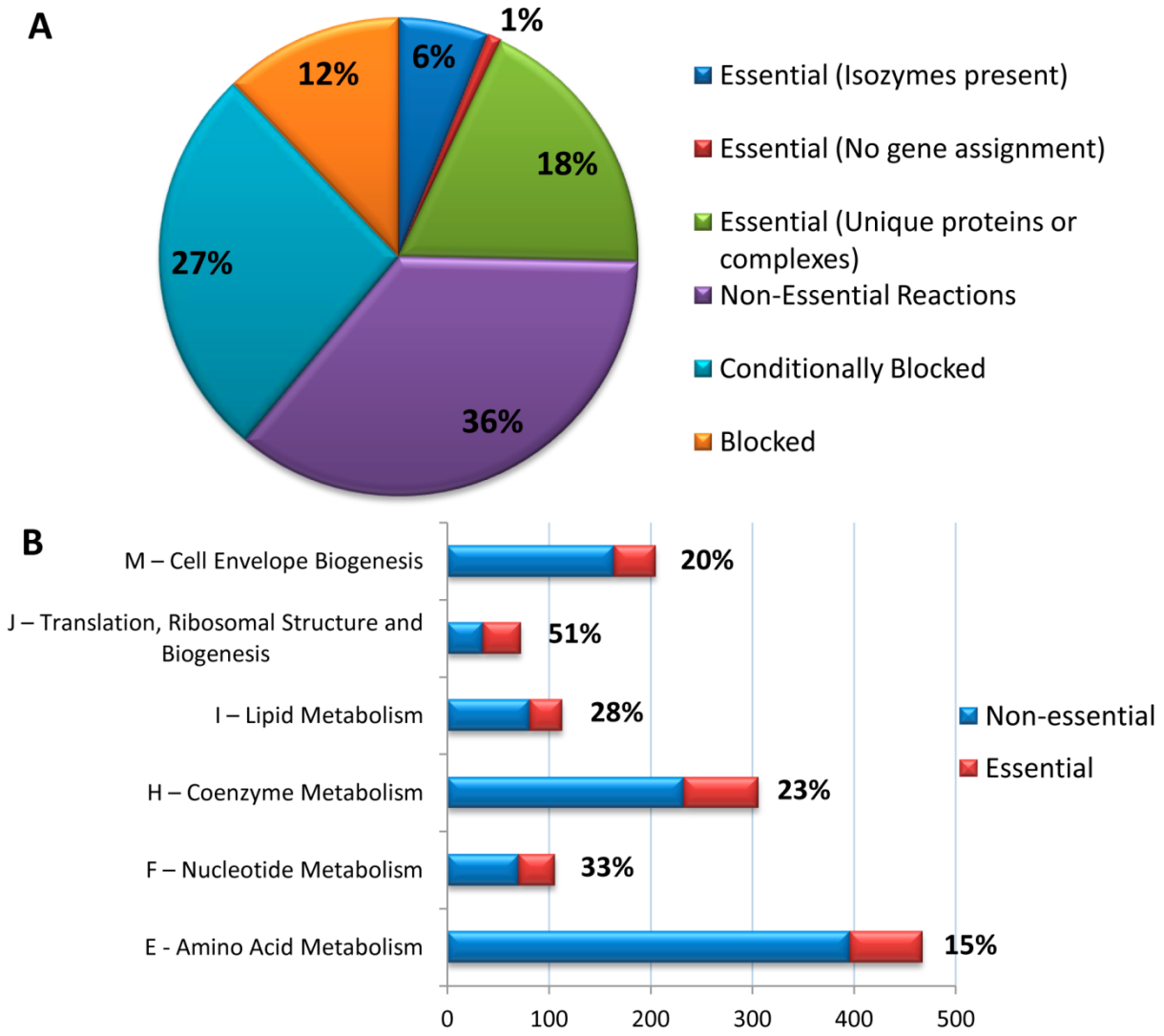

Figure 2-2. Summary of gene and reaction essentiality analysis. (A) Categorization of the reactions present in the model based on their requirement for growth in SIS minimal media supplemented with succinate. About 25% of the reactions in the model are known or predicted to be required for growth under these conditions. (B) Distribution of essential gene products by COG functional classes. The total number of gene products in each class (essential + non-essential) represent only those present in the current model. The percentage of essential proteins in each class is indicated. COG classes depicted are those significantly enriched for essential gene products ($p < 0.01$, hypergeometric distribution).

Quantitative Assessment of iRsp1095

To assess iRsp1095 quantitatively, we used FBA and alternate optima analysis to sample the feasible solution space and make predictions about specific growth rate and the rate of production of key metabolic products during photoheterotrophic growth on a variety of carbon and nitrogen sources (succinate + ammonia, succinate + glutamate, glucose + glutamate and glutamate only), as well as during aerobic growth on succinate and ammonia. We compared the predicted fluxes to experimentally determined growth rate and production rates for these key metabolites during *R. sphaeroides* growth in continuous culture. The model was constrained with experimentally determined uptake rates for the various carbon and nitrogen sources, while being freely allowed to take up all other media components, as well as absorb light. We found that iRsp1095 was capable of accurately predicting cellular growth rate, with predictions generally within 0.4 - 25% of the experimentally observed growth rate (Figure 2-3A), with an overall correlation of 0.75 ($P = 0.012$) across the conditions tested. The FBA predicted growth rate is generally slightly higher than that observed experimentally (especially during growth on succinate + NH_3). These observed differences could be the result of several factors, including stress and feedback inhibition, which cannot be captured in stoichiometric models. Furthermore, many laboratory strains are not necessarily evolved for maximization of growth and thus do not meet the FBA predicted growth rate prior to adaptive evolution experiments [44]. Nevertheless, the predicted growth rates are closer to experimental observations than results previously seen in some other organisms [45, 46], suggesting *R. sphaeroides* strain 2.4.1 is not as far from optimal growth under the conditions we analyzed.

Solutions to linear programming problems are not always unique [21], thus several distinct flux distributions could potentially result in the predicted optimal growth rate. To search the feasible solution space for the possible optimal solutions achievable by iRsp1095 given the constraints on substrate uptake rates, we used a mixed integer linear programming (MILP)-based alternate optima algorithm [22, 23]. A small subset of the reactions in iRsp1095 predicted to function as sinks for excess reducing power were used in sampling the optimal subspace (Materials and Methods). This analysis led to the identification of

some 2 – 17 equivalent optimal solutions, across the various conditions tested, that differed in their pattern of flux distributions. The optimal solution with fluxes for H_2, PHB and CO_2 presented in Figure 2-3B, C and D respectively, represents one where non-zero fluxes for all 3 metabolites were observed in the same solution and which most closely matched the observed experimental data. In addition, the FVA predicted maximum and minimum production rates of these metabolites were assessed. Overall, the predicted amounts of H_2, PHB and CO_2 generally ranged from within 4% to 200% of the experimentally measured fluxes (Figure 2-3B, C and D). Furthermore, constraining the model with the experimentally observed fluxes for PHB, H_2 and CO_2 did not result in decreases in the maximum predicted growth rate for most cultures, suggesting that these experimentally determined flux distributions are also within the optimal subspace. However, applying these constraints to simulations of photoheterotrophic growth on succinate + glutamate and aerobic growth on succinate + NH_3 decreases the predicted growth rate, suggesting the organism is growing sub-optimally under these conditions (Figure 2-3E). Overall, the experimentally measured fluxes generally fell within the optimal solution space of our simulations.

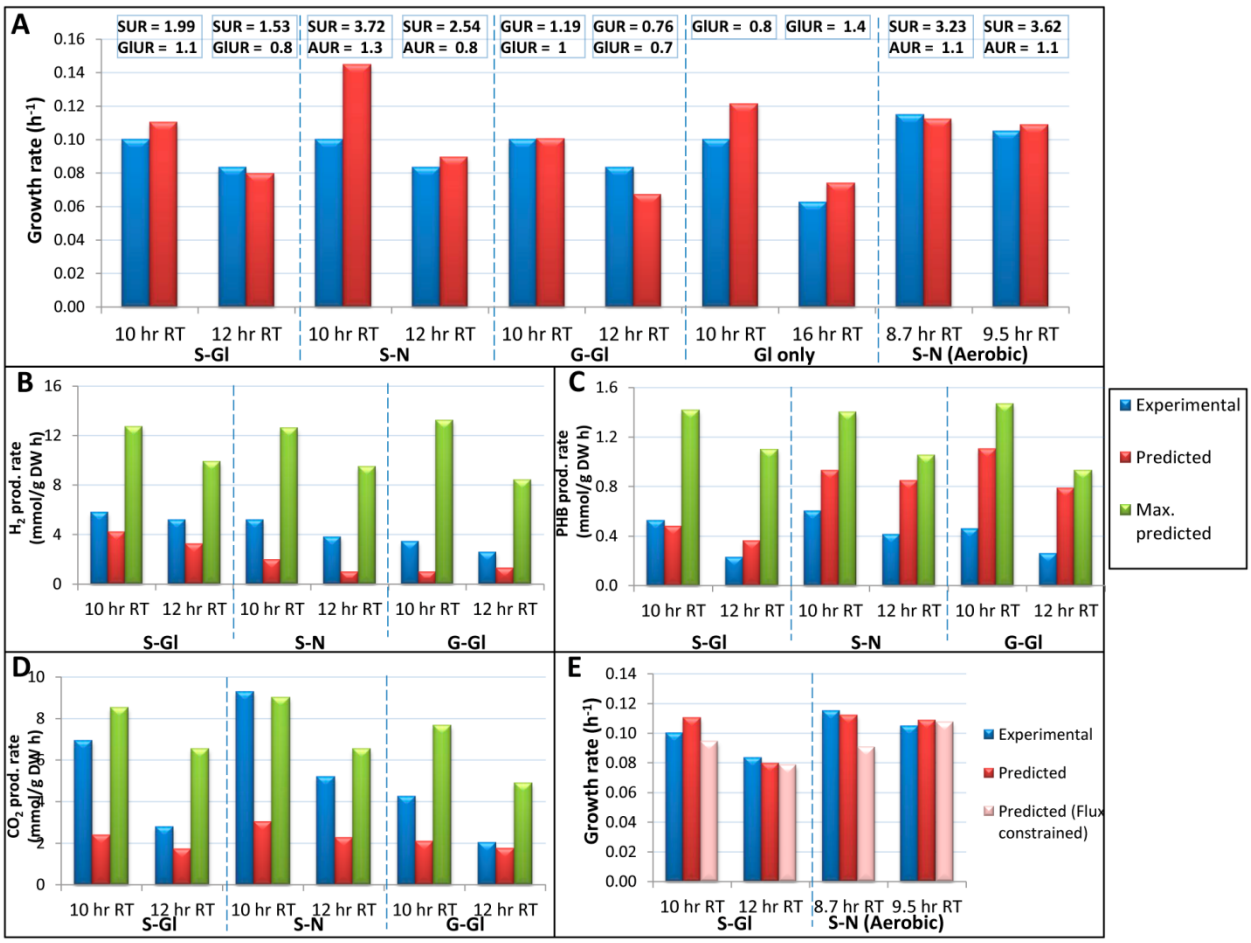

Figure 2-3. Quantitative assessments. The bar graphs compare experimentally measured (A) growth rate, (B) H_2 production, (C) PHB production and (D) CO_2 production with predictions made by iRsp1095. Growth conditions included photoheterotrophic growth on the following carbon and nitrogen sources: succinate + glutamate (S-Gl), succinate + NH_3 (S-N), glucose + glutamate (G-Gl), glutamate as a sole source of carbon and nitrogen (Gl only) and aerobic growth on succinate + ammonia (S-N (Aerobic)). The experimentally determined succinate uptake rate (SUR), ammonia uptake rate (AUR), glucose uptake rate (GUR) and glutamate uptake rate (GlUR) (all in mmol/g DW h) used in all simulations are provided in the boxes above the bars in Figure 2-3A. The FVA predicted maximum production rate of each of the metabolites during optimal growth (Max. predicted) is also provided. In each case the FVA predicted minimum production rate was zero. It should be noted that while ammonia can inhibit nitrogenase activity and thus H_2 production, the ammonia uptake rates in S-N chemostats are not high enough to inhibit nitrogenase activity, hence the production of H_2 under these conditions. No detectable amounts of PHB, H_2 or CO_2 were produced during growth on glutamate only, which is in line with model predictions. Figure 2-3E shows the decreases in predicted growth rate for S-Gl and S-N (Aerobic) conditions upon constraining the model with experimentally observed fluxes for H_2, PHB and/or CO_2 during simulations (Predicted (flux constrained)). The predicted optimal growth rates for other conditions shown in Figure 2-3A were unaffected by imposing these fluxes.

Sensitivity analysis

Further analyses were conducted to evaluate the effects of BOF composition, light uptake and P/O ratio on growth and metabolite production rates in iRsp1095 (see Additional File 5). These analyses showed that: (i) growth rate predictions are not significantly affected by changes in BOF composition, however the production rate of certain metabolites (e.g., H_2) can be affected (see Additional File 5 – Figure S1); (ii) the predicted growth rate and production rates for PHB and H_2 increased with increasing light until they reached a plateau, while the predicted CO_2 production decreased with light uptake, presumably reflecting improved carbon assimilation as biomass increased (see Additional File 5 – Figure S2); and (iii) the P/O ratio can have a significant impact on growth rate, as seen in other metabolic models [38] (see Additional File 5 – Figure S3).

Evaluation of H_2 Production by R. sphaeroides

H_2 serves as a major electron sink for the dissipation of excess substrate reducing power during anoxic phototrophic growth in *R. sphaeroides* [3]. H_2 production in *R. sphaeroides* mainly results from nitrogenase activity, through the coupling of N_2 fixation with H_2 production [47]. However, nitrogenase will also reduce protons, producing H_2 when N_2 is absent [48]. Since high levels of ammonium inhibit nitrogenase activity, H_2 production can be stimulated by supplying the culture with an alternative nitrogen source, such as glutamate [49]. While there is no evidence of H_2 production by the hydrogenase of *R. sphaeroides*, H_2 accumulation in *R. sphaeroides* cultures can also be affected by the presence of this enzyme if H_2 is reutilized by the cells [47].

As a specific application of iRsp1095 we evaluated H_2 production when *R. sphaeroides* is grown on one of several carbon sources with glutamate used as the only nitrogen source, under anoxic photosynthetic conditions. Figure 2-4A shows sensitivity plots of the relationship between growth rate and H_2 production capacity. The theoretical maximum H_2 production while maximizing growth is achieved at the optimal growth rate of 0.076 h^{-1} (e.g., 10.2 mmol/g DW h for succinate). However, for all carbon

sources tested the theoretical H_2 production maxima were reached under suboptimal growth conditions with biomass fluxes around 0.055-0.060 h^{-1}. For comparison to experimental production rates, a reference chemostat yielded a H_2 flux that was about two thirds of the theoretical maximum of 11.5 mmol/g DW h (Figure 2-4A). For cells using glutamate as the sole source of carbon and nitrogen (Figure 2-4A), iRsp1095 predicts little to no H_2 production near maximum growth, consistent with our experimental observations with glutamate only cultures, which produced no detectable H_2.

In Figure 2-4A, the maximal value derived by iRsp1095 was predicted to be larger for more reduced compounds (lactate and glucose) and smaller for less reduced carbon sources (fumarate and pyruvate). Thus, an important question is whether maximum H_2 production is a function of the substrate reducing power only or is also affected by substrate-specific pathways. To address this question, we converted substrate uptake and H_2 production rates to electron fluxes using the stoichiometry of half reactions for electron donation and acceptance [50]. We found that the maximum H_2 production potential for growing cells was linearly related to the available electrons from the substrates (carbon source and glutamate) as shown in Figure 2-4B (see below for the no-growth condition). The linear trend indicates that H_2 producing capacity is proportional to substrate reducing power, irrespective of the carbon source. The intercept of this relationship, where no electrons are available to support H_2 production, shows the reducing power that supports growth alone. The derived slope, which equals 1, indicates that maximizing H_2 production can theoretically be achieved by directing all electrons in excess of that required for growth to H_2 production. This is a significant finding since there are multiple competing pathways that can dissipate substrate reducing power, so this result suggests that H_2 production can be increased from experimental values to theoretical maxima if these other pathways are silenced.

An interesting prediction from these data (Figure 2-4) is that growing cells can support a larger H_2 production potential than resting cells, since, in all cases studied, metabolism with no flux in the biomass reaction yielded the lowest maxima of H_2 flux. Therefore, the breakdown of substrates in biomass synthesis pathways seems necessary to provide maximal reducing power for H_2 production. The

relationship of theoretical maxima at the no-growth condition to the reducing power of the substrates was similar to those with growing cells (Figure 2-4B). That is, the slope of the no-growth curve was also equal to 1, indicating that H_2 can be theoretically maximized when all excess electrons are converted to H_2. However, the model also predicted a baseline of reducing power not converted to H_2, which is represented in Figure 2-4B by the intercept of the no-growth line with the horizontal axis. The flux distribution output from iRsp1095 suggests H_2S as the product accumulating this reducing power baseline.

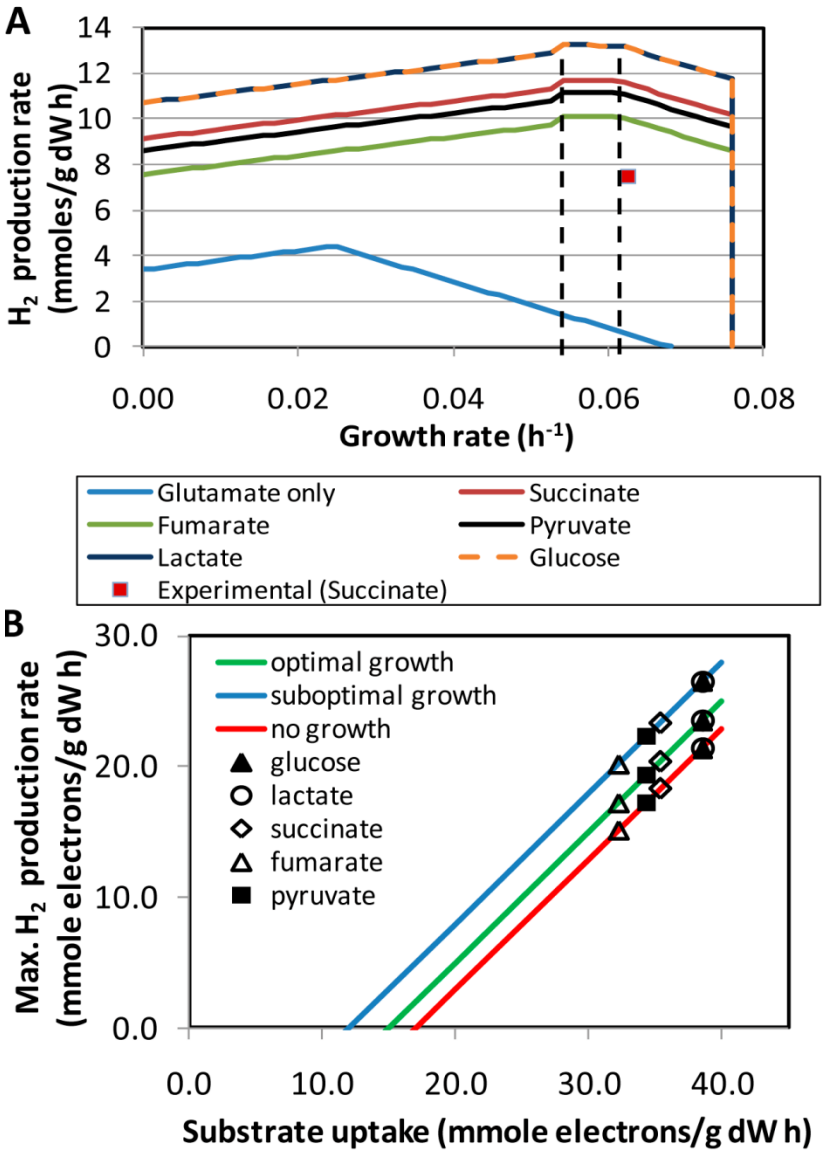

Figure 2-4. H_2 production potential of iRsp1095 with different carbon sources and glutamate as the nitrogen source. (A) Relationship between growth rate and H_2 production. Uptake rates were determined based on a reference chemostat (16h retention time) fed with succinate (SUR = 1.57 mmol/g DW h) and glutamate (GlUR=0.75 mmol/g DW h), such that, GlUR was not varied and any other carbon source was supplied by keeping the total number of carbon atoms constant. Experimentally observed H_2 flux from the reference chemostat is shown as a data point. Dashed lines indicate the region that represents the suboptimal growth rates at which the theoretical maximum amount of H_2 is produced for each carbon source. (B) Relationship between substrate reducing power and electrons used in H_2 production (glutamate only case not included). Best fitting lines correspond to three phases of growth in (A) (R^2 = 1 in all cases).

Metabolic flux distributions

We used FBA to predict metabolic flux distributions during aerobic, photoheterotrophic and photoautotrophic growth.

Photoautotrophic growth

As expected, during photoautotrophic growth iRsp1095 predicts there is a high flux through ribulose-1,5-bisphosphate carboxylase/oxygenase (RubisCO) and the Calvin cycle, as it represents a major pathway for CO_2 assimilation [17, 51]. However, previous analysis of *R. sphaeroides* has shown that a RubisCO mutant (in which form I and form II RubisCO have been deleted) is still capable of photoautotrophic growth, when using less reduced electron donors than H_2 (e.g., thiosulfate or sulfide) [51], suggesting that alternative CO_2 assimilation pathways can support growth under these conditions. Pyruvate carboxylase has previously been shown not be one of these alternative routes [7]. iRsp1095 predicts that the ethylmalonyl pathway, involved in acetyl-coA assimilation, is a candidate for CO_2 sequestration under these conditions. The first enzyme in this pathway, crotonyl-CoA carboxylase/reductase, catalyzes the reductive carboxylation of crotonyl-CoA to ethylmalonyl-CoA [52, 53] and iRsp1095 predicts this pathway can carry sufficient flux for photoautotrophic growth in the absence of RubisCO. Only when the flux through both the RubisCO and crotonyl-CoA carboxylase/reductase reactions are set to zero in the model, does photoautotrophic growth with thiosulfate or sulfide cease to be predicted by iRsp1095, suggesting it is potentially an alternative route of CO_2 fixation in *R. sphaeroides,* and the only one currently incorporated in the model that is capable of supporting photoautotrophic growth in the absence of RubisCO.

Photoheterotrophic growth

FBA simulation of photoheterotrophic growth on succinate and ammonia predicts metabolic flux through reactions involved in the TCA cycle, as might be expected, with significant amounts of H_2 being produced as the rate of ammonia uptake used in simulation (1 mmol/g DW h) results in nitrogen limiting

conditions, allowing excess succinate supplied to the model to be converted to H_2. iRsp1095 does not predict flux through RubisCO to be essential for photoheterotrophic growth; however, it is known that RubisCO is essential for photoheterotrophic growth of wild-type *R. sphaeroides* on carbon sources like succinate and malate, where there is reductive assimilation of CO_2 [7]. Alternate optima analysis [22, 23] predicts a few FBA optima exist wherein RubisCO is used as a major electron sink, however other FBA optima predict the extensive utilization of one or more alternative pathways to recycle excess reducing power including: (i) nitrogenase activity resulting in the production of large amounts of H_2; (ii) the sulfite reductase reaction resulting in the production of H_2S; (iii) PHB synthesis; or (iv) the use of the ethylmalonyl pathway (Figure 2-5). Previous, analyses of *R. sphaeroides* RubisCO mutants have shown that cells are capable of reprogramming their regulatory network to restore photoheterotrophic growth on electron-rich carbon sources [7]. The alternative reactions known to be utilized to restore photoheterotrophic growth under these conditions include nitrogenase reaction yielding H_2 and sulfate reduction to H_2S [7, 17]. Thus the observed alternate optima predicted in iRsp1095 likely represent distinct functional states, all achievable by *R. sphaeroides* based on its metabolic capabilities, but the wild type organism is largely restricted to only a limited number of these as a result of its complex and highly evolved regulatory network, which keeps most of these other functional states silent in the absence of perturbation. Given that these regulatory constraints are not present in iRsp1095, the majority of these functional states are thus achievable, allowing for the prediction of growth in the absence of RubisCO.

Analysis of electron transport chain activity during photoheterotrophic growth shows significant flux through ubiquinol-cytochrome c reductase (Fbc complex) and NADH dehydrogenase, with both enzymes predicted as being essential during growth on succinate and ammonia. The essentiality of the Fbc complex might be expected as it serves as the only means of providing reduced cytochromes required for the photosynthetic light reaction [1, 54]. In contrast, the requirement for NADH dehydrogenase activity during photoheterotrophic growth on succinate is proposed to reflect the need to oxidize ubiquinol and generate NADH for anabolic reactions [55]. Indeed, during anaerobic growth, iRsp1095 predicts that

NADH dehydrogenase uses the transmembrane electron potential to drive the oxidation of ubiquinol to ubiquinone and the concomitant reduction of NAD+ to NADH (Figure 2-5), thus freeing up ubiquinone for use in the cell, while providing NADH for biosynthetic reactions. Furthermore, iRsp1095 predicts that addition of DMSO would restore photoheterotrophic growth in the absence NADH dehydrogenase, as might be expected if cells lacking this enzyme were unable to balance electron flux. It should be noted however, that the predicted essentiality of NADH dehydrogenase during photoheterotrophic growth appears to be conditional, as iRsp1095 predicts that growth occurs with other carbon sources which apparently have less of a requirement for NADH dehydrogenase activity.

Aerobic growth

FBA simulations of aerobic respiratory growth on succinate and ammonia predict significant flux through the TCA cycle and reactions specific to succinate metabolism with the concomitant production of large amounts of CO_2 and trace amounts of urea. iRsp1095 also predicts that cyctochrome c oxidase (Cox) activity is sufficient and required for optimal aerobic respiratory growth. In the absence of Cox activity, quinol oxidase (Qox), which is capable of ubiquinol oxidation to ubiquinone coupled to direct O_2 reduction, is predicted to support aerobic respiratory growth, but the predicted growth rate in this mutant is only 60% of the predicted optimum. A similar reduced growth rate is also predicted in the absence of the Fbc complex, as this also results in flux being directed through Qox in order to oxidize ubiquinol (Figure 2-5). This observed reduction in growth rate might be expected as flux through the Fbc and the Cox complexes pumps 8 protons across the membrane, while flux through Qox, which bypasses both enzymes, results in only 2 protons being pumped across the membrane, thus providing much less energy for the cell (Figure 2-5). Interestingly, NADH dehydrogenase, which is predicted by iRsp1095 to be essential during photoheterotrophic growth on succinate and ammonia, is not predicted to be essential during aerobic respiration. Indeed, only an ~8% decrease in growth rate is predicted during aerobic respiration in the absence of NADH dehydrogenase activity. Since ubiquinol can be oxidized either via the Fbc-Cox pathway or Qox, NADH dehydrogenase activity is no longer required these conditions.

Thus, iRsp1095 predicts that NADH dehydrogenase functions in NADH oxidation during aerobic respiration and contributes to formation of a proton gradient across the membrane (Figure 2-5).

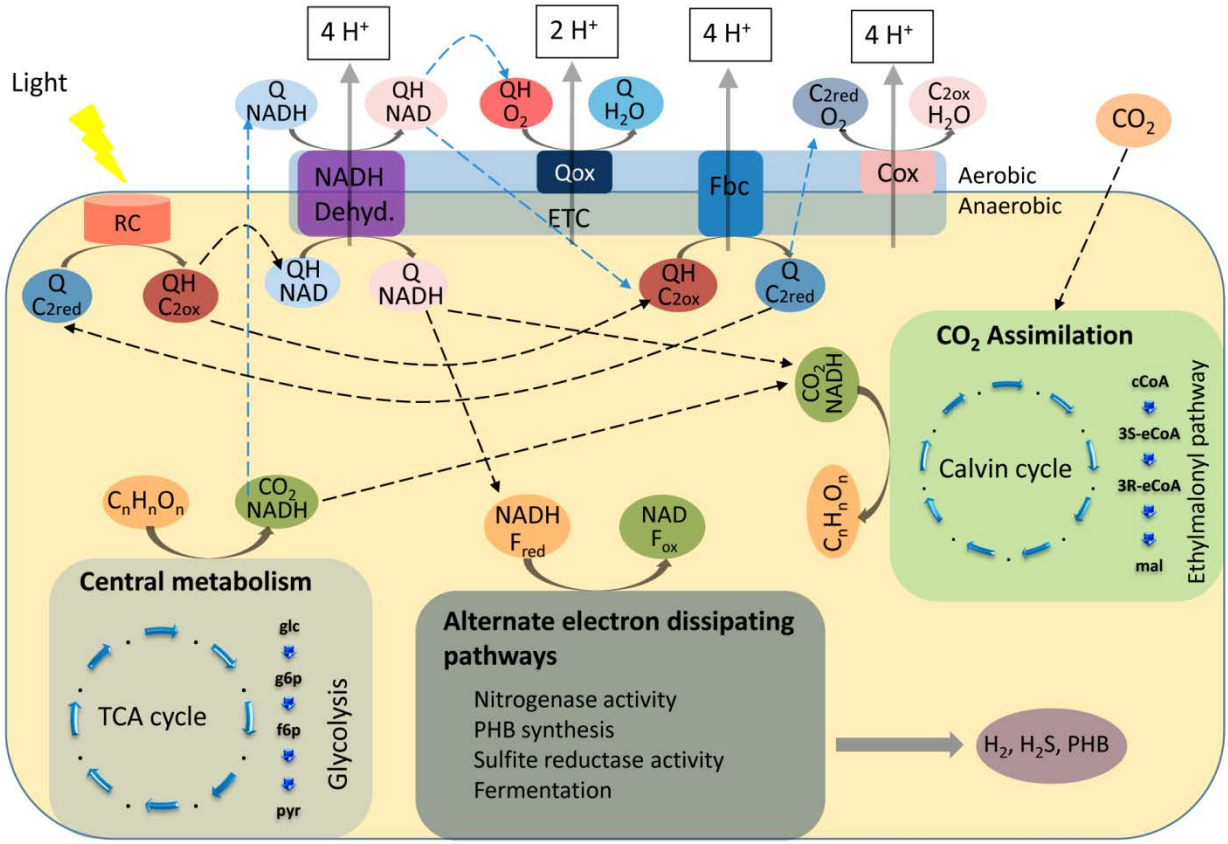

Figure 2-5. Overview of the flux distributions under various growth conditions. Figure shows the flow of electrons during aerobic (light blue dashed arrows) and photosynthetic (black dashed arrows) growth. During anaerobic growth electrons flow from the photosynthetic reaction center (RC), along the electron transport chain (ETC) and back to the RC via reduced cytcochrome ($C2_{red}$) in a cyclic photosynthetic system [54], with protons pumped into the periplasm. NADH dehydrogenase is predicted to function in the reverse direction, reducing NAD to NADH while oxidizing ubiquinol (QH) to ubiquinone (Q). Under these conditions, excess electrons obtained from the oxidation of electron rich carbon sources ($C_nH_mO_p$) would be dissipated via the alternative electron consuming pathways or via carbon fixation. During aerobic growth either cytochrome oxidase (Cox) or quinol oxidase (Qox) can be used in the oxidation of QH to Q, with the Cox reaction favored as more protons are pumped across the cytoplasmic membrane. Under these conditions, NADH dehydrogenase functions in the forward direction oxidizing NADH to NAD. ATP synthase is omitted from the ETC for simplicity. C_{2ox} – oxidized cytochrome; F_{red} – reduced ferredoxin; F_{ox} – oxidized ferredoxin.

Discussion

Previous research has shown the potential of constraint-based analysis for understanding metabolic networks [9]. Given the well-studied photosynthetic lifestyle and biotechnological potential of *R. sphaeroides*, iRsp1095 provides an enabling framework that should increase our understanding of and ability to improve its metabolic machinery. One of the major challenges faced by photosynthetic and many other bacteria is the need to balance the generation of reducing equivalents obtained from light or carbon sources with pathways that consume these electrons. Previous analysis has shown that *R. sphaeroides* partitions significant proportions of reducing equivalents into cellular biomass, PHB, excreted organic acids or H_2 [3]. Furthermore, genetic analysis suggests that CO_2 fixation via RubisCO is also essential for recycling excess reductant during photoheterotrophic growth. Analysis of the flow of reducing equivalents in iRsp1095 reveals that *R. sphaeroides* has several alternate means to potentially recycle reducing equivalents, but not all of these are functional in wild type cells. In addition to known processes like CO_2 fixation via RubisCO, PHB synthesis and H_2 production [3, 7, 17], iRsp1095 also predicts H_2S production by sulfite reductase activity, reductive carbon assimilation via the ethylmalonyl pathway and secretion of metabolites (e.g., lactate and formate) as alternative routes for dissipating excess reducing power. While the role of some of these processes have been experimentally verified [17], others represent novel predictions. The dependence of wild type *R. sphaeroides* on RubisCO for photoheterotrophic growth, a phenotype not corroborated by iRsp1095, suggests that these alternative routes for dissipating excess reducing power could either represent silent functional states or are insufficiently active to support growth in the absence of the Calvin cycle.

Our evaluation of H_2 production potential of *R. sphaeroides* with iRsp1095 showed that continuous culture performance reached two-thirds of the predicted maximum H_2 production (Figure 2-4). To harness the remaining potential predicted by iRsp1095, pathways that contribute to and compete with H_2 production needed to be determined. Interestingly, biomass synthesis is predicted by iRsp1095 to be a contributor to H_2 production. Therefore, we analyzed the pathways that divert electrons from H_2

production in growing cells, when succinate and glutamate were the substrates. This analysis predicts a set of reactions (Table 2-5) whose collective elimination would yield a H_2 production rate of 11.3 mmol/g DW h, very close to the theoretical maximum of 11.5 mmol/g DW h that was predicted by iRsp1095. Five of the products in Table 2-5 are intermediates in cell synthesis pathways and cannot compete with H_2 production under optimal growth conditions (i.e., when biomass flux is maximized). Hence, these reactions provide predictions on the pool of electrons that can be diverted from biomass synthesis to H_2 production.

Table 2-5. Key electron sinks that compete with H_2 production**

Electron Sink	Pathway	Reaction	Responsible Genes
PHB	Butanoate Metabolism	RXN0589	RSP0382, RSP1257
H_2S	Sulfur Metabolism	RXN0866	RSP1942
Glycogen	Starch & Sucrose Metabolism	RXN0849	RSP2887
Formate	One Carbon Pool by Folate	RXN0323	RSP0944
Glycerate	Glycerolipid Metabolism	RXN0030	RSP1292, RSP1507, RSP3740, RSP4003, RSP2372
4-Coumarate	Nitrogen Metabolism	RXN0495	RSP3574
Methanethiol	Cysteine & Methionine Metabolism	RXN1096	RSP1851
4-Aminobutyraldehyde*	Glycolysis / Gluconeogenesis	RXN0031	RSP4003, RSP3740, RSP2372, RSP1507, RSP1292
Chitobiose*	Amino Sugar and Nucleotide Sugar Metabolism	RXN0040	RSP2941
D-1-Aminopropan-2-ol O-phosphate*	Cobalamin Metabolism	RXN0653	RSP0430
Heme*	Heme Metabolism	RXN0632	RSP1197
Ethanolamine*	Glycerophospholipid Metabolism	RXN0378	RSP0113

*Competitors for sub-optimal growth conditions only.

** Secretion of central metabolism intermediates pyruvate, fumarate, and malate were blocked, assuming that cells are programmed to reuse these compounds as carbon sources even after secretion. This was based on observations with batch cultures [3], but might not hold true for continuous cultures.

FBA has also enabled us to model the flow of electrons through the aerobic respiratory chain. *R. sphaeroides* possesses two cytochrome oxidases (Cox): aa3-type cytochrome c oxidase and cbb3-type cytochrome c oxidase, which carry out the same reaction but have different oxygen affinities. *R. sphaeroides* also possesses two quinol oxidases (Qox) – QoxBA and QxtAB - that provide a less energetically efficient means for recycling reduced electron carriers [56]. While it is possible that both Cox and Qox could be used simultaneously, maximization for biomass during FBA simulations results in only the more efficient Fbc-Cox portion of ETC being utilized. Mutational analysis has shown that deletion of either Qox or Qxt has no effect on aerobic growth rate [56], which is in agreement with the predictions of iRsp1095. In addition, mutation of the Fbc complex, which is predicted to redirect flux through the Qox pathway, results in a two fold increase in doubling time experimentally, which is almost identical to the predictions of iRsp1095 (see results and [56]). Furthermore, the loss of both Cox and Qox activity is also correctly predicted by iRsp1095 to be lethal under aerobic conditions. Thus, iRsp1095 accurately models the flux distribution through the aerobic respiratory chain. The reversibility of the NADH dehydrogenase reaction predicted by iRsp1095 and its essentiality during photosynthetic growth has previously been observed in the closely related photosynthetic bacterium *Rhodobacter capsulatus* [55, 57]. Furthermore, conclusions on the essential role of NADH dehydrogenase in synthesizing NADH for anabolic processes under photosynthetic conditions are in agreement with predictions of iRsp1095. Unlike *R. capsulatus*, *R. sphaeriodes* is predicted to contain two isozymes of the NADH dehydrogenase complex, with genes encoding both enzymes being expressed during photoheterotrophic growth [58]. Experimental analysis of the role of each NADH dehydrogenase isozymes during anaerobic growth in *R. sphaeroides* is required to compare with the predictions of iRsp1095.

Finally our simulations predict that several alternative optimal solutions are often possible under any given condition, reinforcing the need to analyze the space of alternate optima [21-23]. The diverse metabolic capabilities of *R. sphaeroides* reinforces the challenge of making accurate predictions about condition-dependent metabolic fluxes as not all feasible functional states are relevant to wild type cells.

Thus, to obtain improved predictions of the flux distributions through the network of wild-type *R. sphaeroides*, additional constraints on iRsp1095 will be required.

Conclusions

iRsp1095 represents the first comprehensive genome-scale metabolic reconstruction for a facultative photosynthetic bacterium. This genome-scale reconstruction has enabled us to examine the metabolic capabilities of this purple non-sulfur bacterium. Our modeling results predict that *R. sphaeroides* possesses multiple pathways that could be exploited as electron sinks during photoheterotrophic growth, though experimental results suggest many of these are silent in wild type cells. Other results predict that additional gains in H_2 production are possible as the production capacity of wild type cells is only about two-thirds of the theoretical maximum, with pathways and reactions that could increase production predicted using iRsp1095. An alternative route for CO_2 fixation, the ethylmalonyl pathway, was predicted using iRsp1095. This prediction could potentially resolve the question of how *R. sphaeroides* assimilates CO_2 in the absence of RubisCO. iRsp1095 also predicts the reversibility of the NADH dehydrogenase complex and its essentiality during photoheterotrophic growth on succinate, where it plays a key role in oxidation of ubiquinol. Further experimental work is needed to confirm these predictions and improve our understanding of the metabolic network of this and possibly other related bacteria. Finally, quantitative predictions made using iRsp1095 showed good agreement with experimental observations, verifying the utility of the model and highlighting the potential for its use in quantitative analysis of *R. sphaeroides* metabolism.

Methods

Constraint-based simulations

A stoichiometric matrix, $S_{m \times n}$, was generated from the reconstruction with the rows (m) representing the metabolites, the columns (n) representing the reactions and the entries in the matrix representing the stoichiometric coefficients for metabolites involved in each reaction. Flux balance analysis (FBA) [19] was used to simulate *in silico* growth by solving the linear programming problem:

$$\max v_{Biomass} \quad (1)$$

s.t

$$\mathbf{S} \bullet \mathbf{v} = \mathbf{0} \quad (2)$$

$$\mathbf{v_{min}} \leq \mathbf{v} \leq \mathbf{v_{max}} \quad (3)$$

where $v_{Biomass}$ is the flux through biomass objective function (BOF); $\mathbf{v}$ is the vector of steady state reaction fluxes; and $\mathbf{v}_{min}$ and $\mathbf{v}_{max}$ are the minimum and maximum allowable fluxes. The values in $\mathbf{v}_{min}$ and $\mathbf{v}_{max}$ were set to -1000 and 1000 mmol/g DW h for reversible reactions, 0 and 1000 mmol/g DW h for forward only reactions, and -1000 and 0 mmol/g DW h for backward only reactions, respectively. During simulation all exchange reactions were assigned as being forward only (allowing metabolites to be secreted into the medium but no taken up), except the exchange reactions for media components required by the cell for growth, which were set to measured values for limiting substrates – carbon and nitrogen sources, or allowed to be freely exchanged with the extracellular space, i.e., $-1000 \leq v \leq 1000$. In addition, the non-growth associated ATP maintenance limit was set to 8.39 mmol/gDW h [38].

Flux variability analysis (FVA) was carried out as described in [21] by first determining the flux through the BOF using FBA, then determining the maximum and minimum possible fluxes through each of the reactions in the network, while the BOF is fixed at the FBA optimum, using equations (4) and (5) below.

$$\max v_i \text{ s.t } \mathbf{S \bullet v = 0}, v_{Biomass} = Z, \mathbf{v_{min} \leq v \leq v_{max}} \quad \text{for i = 1... n (4)}$$

$$\min v_i \text{ s.t } \mathbf{S \bullet v = 0}, v_{Biomass} = Z, \mathbf{v_{min} \leq v \leq v_{max}} \quad \text{for i = 1... n (5)}$$

where Z is the optimal flux through the BOF predetermined using FBA.

Alternate optima analysis was conducted as described in [23], using a mixed integer linear programming algorithm that is a modification of that previously used in [22], which prevents revisiting of already identified optimal solutions. In addition to the FBA constraints outlined above (i.e., Equations 1, 2 and 3), the alternate optima algorithm requires the implementation of the following additional constraints:

$$\sum_{i \in NZ^{j-1}} y_i \geq 1 \qquad (6)$$

$$\sum_{i \in NZ^{k}} w_i \leq \left| NZ^{k} \right| - 1 \qquad k = 1, 2, \ldots, j-1 \qquad (7)$$

$$y_i + w_i \leq 1 \qquad \text{for all i} \qquad (8)$$

$$\mathbf{v_{min}} \cdot w_i \leq \mathbf{v_i} \leq \mathbf{v_{max}} \cdot w_i \quad \text{for all i} \qquad (9)$$

where y_i, $w_i \in \{0, 1\}$, NZ is a set of indices that keeps track of non-zero fluxes of interest for each iteration j. During each iteration through j, at least one of these fluxes of interest v_i with a non-zero value must be set to zero and thus y_i for the corresponding flux is set to 1 (Equation 6). When y_i takes the value 1, w_i is forced to 0 (Equation 8), forcing the upper and lower bounds of v_i to zero (Equation 9). Equation 7 ensures that previously identified optima are not revisited by forcing at least one non-zero flux to have a zero value for the next iteration. Fluxes of interest used in our alternate optima analysis were restricted to those observed to be involved in redox balancing or for which we had experimental measurements for comparison (RXN1205, RXN0222, RXN0109, RXN1427, RXN1308, RXN1425, RXN1441, RXN1121,

and RXN0681 – see Additional File 2 – Table S1 contains reaction details). The use of this set of reactions proved more efficient at sampling the optimal solution space for desired solutions, than using all the reactions in iRsp1095, as it identified an equivalent number or more optimal solutions in which all 3 measured metabolites (i.e., CO_2, H_2 and PHB) had non-zero fluxes.

Deletion analysis was initially carried out at the reaction level by sequentially setting the flux of each reaction to zero, then using FBA to compute the optimal growth rate. Reactions which led to the production of no biomass were considered essential. At the gene level, the fluxes of all reactions associated with a particular gene were set to zero and FBA used to compute the optimal growth rate. Genes encoding proteins whose reactions were required for the formation of biomass, and for which there existed no isozymes in the model, were considered essential.

For analysis of potential carbon, nitrogen, phosphorus and sulfur sources utilized in iRsp1095, simulations were conducted using SIS as the baseline media, which contains succinate, ammonium, phosphate, and sulfate as the only the sources of carbon, nitrogen, phosphorus, and sulfur, respectively. To test a different source, the original metabolite was removed and replaced with the metabolite to be tested. When needed, temporary sink reactions [20] were added for each metabolite to be tested and these reactions were removed at the completion of the analysis. Metabolites which resulted in the predicted growth rate greater than 0 were considered as potential growth substrates. All simulations were conducted under the GAMS programming environment (GAMS Development Corporation, Cologne, Germany) using the CPLEX solver.

Continuous Cultures

To obtain steady state growth data for FBA, wild type *R. sphaeroides* 2.4.1 was cultured in 20mL chemostats at ~30^{o}C, either continuously illuminated by an incandescent light source for photosynthetic growth (~10 W/m^2, as measured with a Yellow–Springs–Kettering model 6·5-A radiometer through a Corning 7–69 filter), or continuously aerated (4mL/min from a compressed air cylinder) in dark

conditions for aerobic growth. The turbidity of photosynthetic cells was monitored using a Klett-Summerson photoelectric colorimeter (Klett MFG Co., NY), while that of aerobic cultures was measured spectrophotometrically at 600nm wavelength with a UV-1601 Spectrophotometer (Shimadzu Scientific, Columbia, MD). Reactors were started in batch mode [3] until cells reached >100 Klett units or >300 O.D. at 600nm, and were then continuously fed with medium using Masterflex peristaltic pumps (Cole-Palmer Instrument Co., Vernon Hills, IL). To reach the desired retention time, an appropriate amount of medium was replaced by 5-min continuous pumping every hour. Reactors were checked approximately every 12 hours, and when necessary, pumping was manually adjusted to correct small changes in reactor volume due to marginal imbalances of inflow and outflow. Cultures were grown for at least 5 retention times and stopped when steady state was established as evidenced from constant turbidity measures or –in case of some photosynthetic cultures- constant gas rates. All reactors were fed with Sistrom's minimal medium [41] containing one of the following pairs of carbon and nitrogen sources in the respective order: 33.9 mM succinate and 7.5mM ammonia, 33.9 mM succinate and 8.1 mM glutamate, 19.8 mM glucose and 8.1 mM glutamate, and 26.6mM glutamate as both carbon and nitrogen sources.

Biomass composition analysis

Cultures were centrifuged (6,000 rpm, 12 mins, 4^{o}C) to obtain cell pellets for biomass analysis. Cell pellets pooled from several chemostats were resuspended in 1X SIS medium, mixed, and distributed into different subsamples for measuring individual biomass components. The major cellular components measured were protein, DNA, RNA, cell wall, lipids, bacteriochlorophyll, carotenoids, glycogen and PHB.

Total cellular protein was quantified via the Lowry assay [59, 60], while total DNA and RNA were determined spectrophotometerically after phenol/chloroform and perchloric acid extraction respectively [61]. Total cellular lipid content was estimated using the sulfo-phospho-vanillin assay on crude lysates [62], while the phospholipid component was determined by total phosphorus assay on lipids extracted via

standard chloroform/methanol extraction [63]. Total cellular bacteriochlorophyll was determined spectrophotometrically at 770nm following acetone/methanol extraction. Bacteriochlorophyll levels were used in estimating cellular carotenoid content based on the previously determined 2:1 ratio of bacteriochlorophyll to carotenoids in the B800-850 complex of *R. sphaeroides* [64].

The PHB content of cells was determined by GC-MS (GC-2010 gas chromatograph coupled to a QP-2010S mass spectrometer detector; Shimadzu Scientific) [3, 65]. Cellular glycogen content was determined by digestion of glycogen in cellular extracts to glucose using amyloglucosidase (Sigma-Aldrich) and quantification of glucose, using a glucose (HK) assay kit (Sigma-Aldrich). Identically treated dilutions of glycogen (Sigma-Aldrich) were used as standard for quantification. Cell wall composition of biomass was assumed to be similar to that of *E. coli* [38]. Finally, the fraction of inorganic material was based on ash content in previous biomass analyses of closely related species [66].

Biomass reaction and net cell dry weight (dW)

The biomass reaction of the metabolic model was formulated using major biomass components as detailed in Additional File 3. Since PHB and glycogen varied significantly based on carbon source and growth conditions, they were not included in the biomass reaction. Instead, they were modeled via the addition of demand reactions to allow their accumulation during simulation. Accordingly, the normalization of all fluxes was done using a dry weight (dW) calculation that excludes PHB and glycogen from the cell mass estimate. For this, we calculated dW using a chemical oxygen demand (COD) mass balance approach [3] as shown in Equation 10, where $COD_{biomass}$ represents the overall measurement of COD in cells, COD_{PHB} and $COD_{glycogen}$ represent COD of PHB and glycogen obtained from experimental measurements and theoretical COD/weight ratios (1.67 mgCOD/mgPHB and 1.18 mgCOD/mg glycogen), and θ is the COD/weight ratio for the cell mass according to the biomass reactions established in this study (θ is 1.62 for photosynthetic and 1.56 for aerobic cultures).

$$dW = \frac{COD_{biomass} - (COD_{PHB} + COD_{glycogen})}{\theta} \quad (10)$$

Other Analytical Measurements

Substrate uptake rates of all cultures and gas composition of the headspace in phototrophic chemostats were measured using previously described protocols [3] (see Additional File 1). No gas evolution measurements were taken for aerobic cultures as they were continuously aerated. However, the amount of O_2 uptake, an important parameter in the FBA of aerobic growth, was indirectly measured by COD, which was previously used for analysis of electron fate in photosynthetic cultures of *R. sphaeroides* [3]. Briefly, the COD of the medium (inflow) and of the reactor effluent were measured using HACH High Range (0-1500 mg/L) COD kits (HACH Company, Loveland, CO), and the difference between medium and effluent (including cells) gave the estimated O_2 utilization by the cells due to the mass balance of electrons [3].

References

1. Hunter CN, Daldal, F., Thurnauer, M. C. and Beatty, J. T.: **The purple phototrophic bacteria**, vol. 28: Springer; 2009.

2. Atsumi S, Higashide W, Liao JC: **Direct photosynthetic recycling of carbon dioxide to isobutyraldehyde**. *Nat Biotechnol* 2009, **27**(12):1177-1180.

3. Yilmaz LS, Kontur WS, Sanders AP, Sohmen U, Donohue TJ, Noguera DR: **Electron partitioning during light- and nutrient-powered hydrogen production by *Rhodobacter sphaeroides***. *Bioenerg Res* 2010, **Volume**(1):55 - 66.

4. Eraso JM, Kaplan S: **Regulation of gene expression by PrrA in *Rhodobacter sphaeroides* 2.4.1: role of polyamines and DNA topology**. *J Bacteriol* 2009, **191**(13):4341-4352.

5. Mackenzie C, Eraso JM, Choudhary M, Roh JH, Zeng X, Bruscella P, Puskas A, Kaplan S: **Postgenomic adventures with *Rhodobacter sphaeroides***. *Annu Rev Microbiol* 2007, **61**:283-307.

6. Khatipov E, Miyake, M., Miyake J. and Y. Asada: **Polyhydroxybutyrate accumulation and hydrogen evolution by *Rhodobacter sphaeroides* as a function of nitrogen availability**. *Biohydrogen* 1999, **III**:157 - 161.

7. Wang X, Falcone DL, Tabita FR: **Reductive pentose phosphate-independent CO_2 fixation in *Rhodobacter sphaeroides* and evidence that ribulose bisphosphate carboxylase/oxygenase activity serves to maintain the redox balance of the cell**. *J Bacteriol* 1993, **175**(11):3372-3379.

8. Feist AM, Palsson BO: **The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli***. *Nat Biotechnol* 2008, **26**(6):659-667.

9. Oberhardt MA, Palsson BO, Papin JA: **Applications of genome-scale metabolic reconstructions**. *Mol Syst Biol* 2009, **5**:320.

10. Price ND, Papin JA, Schilling CH, Palsson BO: **Genome-scale microbial *in silico* models: the constraints-based approach**. *Trends Biotechnol* 2003, **21**(4):162-169.

11. Segre D, Vitkup D, Church GM: **Analysis of optimality in natural and perturbed metabolic networks**. *Proc Natl Acad Sci U S A* 2002, **99**(23):15112-15117.

12. Knoop H, Zilliges Y, Lockau W, Steuer R: **The metabolic network of *Synechocystis sp.* PCC 6803: systemic properties of autotrophic growth**. *Plant Physiol* 2010, **154**(1):410-422.

13. Montagud A, Navarro E, Fernandez de Cordoba P, Urchueguia JF, Patil KR: **Reconstruction and analysis of genome-scale metabolic model of a photosynthetic bacterium**. *BMC Syst Biol* 2010, **4**:156.

14. Shastri AA, Morgan JA: **Flux balance analysis of photoautotrophic metabolism**. *Biotechnol Prog* 2005, **21**(6):1617-1626.

15. Klamt S, Schuster S, Gilles ED: **Calculability analysis in underdetermined metabolic networks illustrated by a model of the central metabolism in purple nonsulfur bacteria**. *Biotechnol Bioeng* 2002, **77**(7):734-751.

16. Golomysova A, Gomelsky, M. and Ivanov P. S.: **Flux balance analysis of photoheterotrophic growth of purple nonsulfur bacteria relevant to biohydrogen production**. *International Journal of Hydrogen Energy* 2010, **35**(23):12751 - 12760.

17. Rizk ML, Laguna R, Smith KM, Tabita FR, Liao JC: **Redox homeostasis phenotypes in RubisCO-deficient *Rhodobacter sphaeroides* via ensemble modeling**. *Biotechnol Prog* 2010, **27**(1):15-22.

18. Price ND, Thiele I, Palsson BO: **Candidate states of *Helicobacter pylori*'s genome-scale metabolic network upon application of "loop law" thermodynamic constraints**. *Biophys J* 2006, **90**(11):3919-3928.

19. Varma A, Palsson BO: **Metabolic flux balancing: basic concepts, scientific and practical use.** *Nature Biotechnology* 1994, **12**:994 - 998.

20. Thiele I, Palsson BO: **A protocol for generating a high-quality genome-scale metabolic reconstruction**. *Nat Protoc* 2010, **5**(1):93-121.

21. Mahadevan R, Schilling CH: **The effects of alternate optimal solutions in constraint-based genome-scale metabolic models**. *Metab Eng* 2003, **5**(4):264-276.

22. Lee S, Phalakornkule, C., Domach, M.M., and Grossmann, I.E.: **Recursive MILP model for finding all the alternate optima in LP models for metabolic networks**. *Computers & Chemical Engineering* 2000, **24**(2 - 7):711 -716.

23. Reed JL, Palsson BO: **Genome-scale in silico models of *E. coli* have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states**. *Genome Res* 2004, **14**(9):1797-1805.

24. Kanehisa M, Goto S, Kawashima S, Nakaya A: **The KEGG databases at GenomeNet**. *Nucleic Acids Res* 2002, **30**(1):42-46.

25. Pinney JW, Shirley MW, McConkey GA, Westhead DR: **metaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of *Plasmodium falciparum* and *Eimeria tenella***. *Nucleic Acids Res* 2005, **33**(4):1399-1409.

26. Henry CS, Zinner JF, Cohoon MP, Stevens RL: **iBsu1103: a new genome-scale metabolic model of *Bacillus subtilis* based on SEED annotations**. *Genome Biol* 2009, **10**(6):R69.

27. Price ND, Famili I, Beard DA, Palsson BO: **Extreme pathways and Kirchhoff's second law**. *Biophys J* 2002, **83**(5):2879-2882.

28. Feist AM, Palsson BO: **The biomass objective function**. *Curr Opin Microbiol* 2010, **13**(3):344-349.

29. Kiley PJ, Kaplan S: **Molecular genetics of photosynthetic membrane biosynthesis in *Rhodobacter sphaeroides***. *Microbiol Rev* 1988, **52**(1):50-69.

30. Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O: **The comprehensive microbial resource**. *Nucleic Acids Res* 2001, **29**(1):123-125.

31. Benning C: **Biosynthesis and function of the sulfolipid sulfoquinovosyl diacylglycerol**. *Annu Rev Plant Physiol Plant Mol Biol* 1998, **49**:53-75.

32. Benning C, Somerville CR: **Identification of an operon involved in sulfolipid biosynthesis in *Rhodobacter sphaeroides***. *J Bacteriol* 1992, **174**(20):6479-6487.

33. Catucci L, Depalo N, Lattanzio VM, Agostiano A, Corcelli A: **Neosynthesis of cardiolipin in *Rhodobacter sphaeroides* under osmotic stress**. *Biochemistry* 2004, **43**(47):15066-15072.

34. De Leo V, Catucci L, Ventrella A, Milano F, Agostiano A, Corcelli A: **Cardiolipin increases in chromatophores isolated from *Rhodobacter sphaeroides* after osmotic stress: structural and functional roles**. *J Lipid Res* 2009, **50**(2):256-264.

35. Donohue TJ, Cain BD, Kaplan S: **Purification and characterization of an N-acylphosphatidylserine from *Rhodopseudomonas sphaeroides***. *Biochemistry* 1982, **21**(11):2765-2773.

36. Gage DA, Huang ZH, Benning C: **Comparison of sulfoquinovosyl diacylglycerol from spinach and the purple bacterium *Rhodobacter spaeroides* by fast atom bombardment tandem mass spectrometry**. *Lipids* 1992, **27**(8):632-636.

37. Marinetti GV, Cattieu K: **Lipid analysis of cells and chromatophores of *Rhodopseudomonas sphaeroides***. *Chemistry and Physics of Lipids* 1981, **28**(3):241 - 251.

38. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO: **A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information**. *Mol Syst Biol* 2007, **3**:121.

39. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN *et al*: **The COG database: an updated version includes eukaryotes**. *BMC Bioinformatics* 2003, **4**:41.

40. Li C, Donizelli M, Rodriguez N, Dharuri H, Endler L, Chelliah V, Li L, He E, Henry A, Stefan MI *et al*: **BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models**. *BMC Syst Biol* 2010, **4**:92.

41. Sistrom WR: **The kinetics of the synthesis of photopigments in *Rhodopseudomonas spheroides***. *J Gen Microbiol* 1962, **28**:607-616.

42. Garrity GM, Brenner, D. J., Krieg, N. R., Staley, J. T. and Krieg, N. R.: **The proeobacteria: part C the alpha-, beta-, delta-, and epsilon-proteobacteria**: Springer; 2005.

43. Novak RT, Gritzer RF, Leadbetter ER, Godchaux W: **Phototrophic utilization of taurine by the purple nonsulfur bacteria *Rhodopseudomonas palustris* and *Rhodobacter sphaeroides***. *Microbiology* 2004, **150**(Pt 6):1881-1891.

44. Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, Adkins JN, Schramm G, Purvine SO, Lopez-Ferrer D *et al*: **Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models**. *Mol Syst Biol* 2010, **6**:390.

45. Gowen CM, Fong SS: **Genome-scale metabolic model integrated with RNAseq data to identify metabolic states of *Clostridium thermocellum***. *Biotechnol J* 2010, **5**(7):759-767.

46. Ibarra RU, Edwards JS, Palsson BO: ***Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth**. *Nature* 2002, **420**(6912):186-189.

47. Kim E, Lee M, Kim M, Lee JK: **Molecular hydrogen production by nitrogenase of *Rhodobacter sphaeroides* and by Fe-only hydrogenase of *Rhodospirillum rubrum***. *International Journal of Hydrogen Energy* 2008, **33**(5):1516-1521.

48. Rivera-Ortiz JM, Burris RH: **Interactions among substrates and inhibitors of nitrogenase**. *J Bacteriol* 1975, **123**(2):537-545.

49. Gabrielyan L, Torgomyan H, Trchounian A: **Growth characteristics and hydrogen production by *Rhodobacter sphaeroides* using various amino acids as nitrogen sources and their combinations with carbon sources**. *International Journal of Hydrogen Energy* 2010, **35**(22):12201-12207.

50. Rittman B, McCarty PL: **Environmental biotechnology: principles and applications**: McGraw-Hill Science Engineering; 2000.

51. Wang X, Modak HV, Tabita FR: **Photolithoautotrophic growth and control of CO_2 fixation in *Rhodobacter sphaeroides* and *Rhodospirillum rubrum* in the absence of ribulose bisphosphate carboxylase-oxygenase**. *J Bacteriol* 1993, **175**(21):7109-7114.

52. Erb TJ, Berg IA, Brecht V, Muller M, Fuchs G, Alber BE: **Synthesis of C5-dicarboxylic acids from C2-units involving crotonyl-CoA carboxylase/reductase: the ethylmalonyl-CoA pathway**. *Proc Natl Acad Sci U S A* 2007, **104**(25):10631-10636.

53. Erb TJ, Frerichs-Revermann L, Fuchs G, Alber BE: **The apparent malate synthase activity of *Rhodobacter sphaeroides* is due to two paralogous enzymes, (3S)-Malyl-coenzyme A (CoA)/{beta}-methylmalyl-CoA lyase and (3S)- Malyl-CoA thioesterase**. *J Bacteriol* 2010, **192**(5):1249-1258.

54. McEwan AG: **Photosynthetic electron transport and anaerobic metabolism in purple non-sulfur phototrophic bacteria**. *Antonie Van Leeuwenhoek* 1994, **66**(1-3):151-164.

55. Herter SM, Kortluke CM, Drews G: **Complex I of *Rhodobacter capsulatus* and its role in reverted electron transport**. *Arch Microbiol* 1998, **169**(2):98-105.

56. Mouncey NJ, Gak E, Choudhary M, Oh J, Kaplan S: **Respiratory pathways of *Rhodobacter sphaeroides* 2.4.1(T): identification and characterization of genes encoding quinol oxidases**. *FEMS Microbiol Lett* 2000, **192**(2):205-210.

57. Dupuis A, Peinnequin, A., Darrouzzet, E. and Lunardi J.: **Genetic disruption of the respiratory NADH-ubiquinone reductase of *Rhodobacter capsulatus* leads to an unexpected photosynthesis-negative phenotype**. *FEMS Microbiology Letters* 1997, **148**(1):107 - 113.

58. Arai H, Roh JH, Kaplan S: **Transcriptome dynamics during the transition from anaerobic photosynthesis to aerobic respiration in *Rhodobacter sphaeroides 2.4.1***. *J Bacteriol* 2008, **190**(1):286-299.

59. Hartree EF: **Determination of protein: a modification of the Lowry method that gives a linear photometric response**. *Anal Biochem* 1972, **48**(2):422-427.

60. Lowry OH, Rosebrough NJ, Farr AL, Randall RJ: **Protein measurement with the Folin phenol reagent**. *J Biol Chem* 1951, **193**(1):265-275.

61. Benthin S. NJaJV: **A simple and reliable method for the determination of cellular RNA content**. *Biotechnology Techniques* 1991, **5**(1):39 - 42.

62. Izard J, Limberger RJ: **Rapid screening method for quantitation of bacterial cell lipids from whole cells**. *J Microbiol Methods* 2003, **55**(2):411-418.

63. Rouser G, Fkeischer S, Yamamoto A: **Two dimensional then layer chromatographic separation of polar lipids and determination of phospholipids by phosphorus analysis of spots**. *Lipids* 1970, **5**(5):494-496.

64. Evans MB, Cogdell, R. J. and G. Britton: **Determination of the bacteriochlorophyll:Carotenoid ratios of the B890 antenna complex of *Rhodospirillum rubrum* and the B800–850 complex of *Rhodobacter sphaeroides*.** *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 1988, **935**(3):292 - 298.

65. Villas-Boas SG, Delicado DG, Akesson M, Nielsen J: **Simultaneous analysis of amino and nonamino organic acids as methyl chloroformate derivatives using gas chromatography-mass spectrometry**. *Anal Biochem* 2003, **322**(1):134-138.

66. Kobayashi M, Kobayashi M: **Waste remediation and treatment using anoxygenic phototrophic bacteria.**: Kluwer academic publishers; 1995.

Chapter 3

Global insights into energetic and metabolic networks in *Rhodobacter sphaeroides*

This chapter is published under the same title:

Imam S, Noguera DR, Donohue TJ. BMC Syst Biol. 2013 Sep 13;7:89. doi: 10.1186/1752-0509-7-89. PubMed PMID: 24034347.

I performed all the experiments and analyses in this chapter.

Abstract

Background

Improving our understanding of processes at the core of cellular lifestyles can be aided by combining information from genetic analyses, high-throughput experiments and computational predictions.

Results

We combined data and predictions derived from phenotypic, physiological, genetic and computational analyses to dissect the metabolic and energetic networks of the facultative photosynthetic bacterium *Rhodobacter sphaeroides*. We focused our analysis on pathways crucial to the production and recycling of pyridine nucleotides during aerobic respiratory and anaerobic photosynthetic growth in the presence of an organic electron donor. In particular, we assessed the requirement for NADH/NADPH transhydrogenase enzyme, PntAB during respiratory and photosynthetic growth. Using high-throughput phenotype microarrays (PMs), we found that PntAB is essential for photosynthetic growth in the presence of many organic electron donors, particularly those predicted to require its activity to produce NADPH. Utilizing the genome-scale metabolic model iRsp1095, we predicted alternative routes of NADPH synthesis and used gene expression analyses to show that transcripts from a subset of the corresponding genes were conditionally increased in a Δ*pntAB* mutant. We then used a combination of metabolic flux predictions and mutational analysis to identify flux redistribution patterns utilized in the Δ*pntAB* mutant to compensate for the loss of this enzyme. Data generated from metabolic and phenotypic analyses of wild type and mutant cells were used to develop iRsp1140, an expanded genome-scale metabolic reconstruction for *R. sphaeroides* with improved ability to analyze and predict pathways associated with photosynthesis and other metabolic processes.

Conclusion

These analyses increased our understanding of key aspects of the photosynthetic lifestyle, highlighting the added importance of NADPH production under these conditions. It also led to a significant improvement in the predictive capabilities of a metabolic model for the different energetic lifestyles of a facultative organism.

Introduction

Information about an organism's capabilities can be derived from a variety of sources. When genomic information is combined with biochemical, phenotypic or genetic data, functional roles and interrelationships of components within metabolic or regulatory networks become better defined [1-5]. Thus, to obtain a global view of an organism's capabilities, it is often beneficial to develop models that integrate data from different types of experiments. In obtaining such integrated views, genome-scale metabolic network models can serve both as databases for storage and organization and as tools for the combination and analysis of heterogeneous data sets [6]. A particular interest of our laboratory is developing an integrated understanding of metabolic networks in photosynthetic microbes, because of their abundance in nature, the unique aspects of a solar-driven lifestyle and their growing importance in biotechnological applications [7-9].

We study purple non-sulfur bacteria, a group of photosynthetic microbes that display great metabolic and energetic diversity [10]. The purple non-sulfur bacterium *Rhodobacter sphaeroides* represents one of the best studied photosynthetic organisms, and has been used to develop models of photon capture, light-driven energy metabolism and other aspects of its diverse lifestyles [11, 12]. This facultative microbe is capable of anoxygenic photosynthetic growth, aerobic respiration and anaerobic respiration [11, 12]. Furthermore, *R. sphaeroides* has been studied for potential biotechnological applications including the ability to produce H_2 [13-15] and ubiquinone [16], production of polyhydroxybutyrate, which can be used as a source of biodegradable plastics [17], remediation of radioactive contamination [18], and its ability to fix CO_2 and N_2 [7, 19, 20]. The available genetic, genomic and physiological tools [12] also make *R. sphaeroides* an excellent system in which to improve our understanding of solar energy capture, metabolic and energetic aspects of photosynthesis and other energetic pathways, and the networks which regulate processes of societal and biotechnological interest. To obtain an integrated understanding of photosynthesis or other aspects of *R. sphaeroides*' lifestyles requires the use of high-throughput data to develop better predictive models of its metabolic network.

In this work, we take a systematic approach to expand our knowledge of the metabolic and energetic networks of *R. sphaeroides* by combining data from genetic, phenotypic and transcriptional analyses with constraint-based modeling. We use high-through phenotypic microarrays to show that wild type *R. sphaeroides* grows on a diverse array of substrates and that this nutrient utilization profile varies significantly between photosynthetic and non-photosynthetic growth conditions. Using the conserved bioenergetic enzyme pyridine nucleotide transhydrogenase (PntAB) as an example, we identify carbon sources where recycling of pyridine nucleotides by this enzyme is essential for photosynthetic or non-photosynthetic growth. We use a genome-scale metabolic model to predict flux distributions and identify alternative NADPH producing reactions that can compensate for the loss of PntAB and thereby explain the conditional growth of ΔPntAB cells on selected carbon sources. Transcriptional and phenotypic analyses of defined single and double mutants were used to verify the potential use of some of these alternative NADPH producing reactions under defined conditions. The new data derived from analyzing the growth of wild type and mutant cells were used to develop iRsp1140, a significant update to the existing genome-scale reconstruction of the *R. sphaeroides* metabolic network [11], with increased coverage of metabolic pathways and improved predictive ability. iRsp1140 accounts for 1140 genes, 878 metabolites and 1416 reactions. This work illustrates the new insights into important cellular processes that can be acquired by integrating data from genetic, genomic and other complementary experiments into predictive models of biological systems.

Results and Discussion

Global analysis of substrate utilization by *R. sphaeroides*

One important step in acquiring a global understanding of cellular processes in an organism is to develop a broad perspective of its metabolic repertoire. An assessment of the literature reveals that *R. sphaeroides* has been reported to grow on 27 carbon, 3 nitrogen, 1 phosphorus and 4 sulfur sources [11]. In contrast, the existing genome-scale model of *R. sphaeroides* metabolic network, iRsp1095, predicted an ability to grow on a significantly larger number of substrates (Table 3-1) [11]. Thus, to improve our knowledge of the metabolic capabilities of *R. sphaeroides*, and aid subsequent analyses of genetic or physiological perturbations, we used phenotype microarrays (PM) [21, 22] to assess the ability of wild type (WT) cells to utilize carbon sources under anoxygenic photoheterotrophic conditions (using a single carbon source as an external electron donor; this is hereafter referred to as photosynthetic growth), aerobic and anaerobic respiratory conditions (see Methods). We also determined the suite of, nitrogen (N), phosphorus (P) and sulfur (S) sources that do or do not support photosynthetic growth.

R. sphaeroides utilizes different arrays of nutrients across growth conditions

The results from analysis of substrate utilization by WT *R. sphaeroides* cells (Table 3-1, Additional File 1 – Tables S1-S4), significantly expands the array of compounds that support growth of this organism. While the carbon utilization profiles were largely similar during photosynthetic and aerobic respiratory growth, several important differences were observed. Eight carbon sources appeared to support growth photosynthetically but not aerobically, while 15 supported growth aerobically but not photosynthetically (Additional File 2 – Figure S1A, Additional File 1 – Table S1). Potential causes for the observed differences might include: (i) longer lag times under individual conditions (Additional File 2 – Figure S1B), which may result in an apparent inability to utilize the carbon source under one experimental condition; (ii) insurmountable metabolic, bioenergetic or regulatory bottlenecks (Additional File 2 –

Figure S1C); or (iii) potential differences between the data derived from the photosynthetic PM assay (which measures an increase in optical density) and the aerobic PM assay (that measures respiration) [23].

Of the 53 carbon sources that were used both photosynthetically and aerobically, 41 were tested for their ability to support growth under anaerobic respiratory conditions using dimethyl sulfoxide (DMSO) as the terminal electron acceptor (Additional File 1 – Table S1). Only 16 of these carbon sources were capable of supporting anaerobic respiratory growth (as measured by an increase in optical density) after 10 days of incubation. We propose that the inability of WT *R. sphaeroides* to grow in the presence of several carbon substrates during anaerobic respiratory growth is likely due to regulatory and/or bioenergetic constraints, as pathways required for their catabolism are either known or predicted to be present in the genome.

PM assays also revealed that 66 nitrogen, 42 phosphorus and 18 sulfur sources supported growth photosynthetically in WT *R. sphaeroides* (Table 3-1, Additional File 1 – Tables S2-S4). This is a number of nitrogen, phosphorous and sulfur substrates which is similar to those shown to support growth of other well-studied facultative bacteria like *Escherichia coli* [24] and *Bacillus subtilis* [25].

The ability of *R. sphaeroides* to grow on a wide variety of carbon, nitrogen, phosphorus and sulfur sources (Additional File 1 – Table S8) is a further demonstration of its metabolic versatility. Of particular interest for future studies is the pattern of substrate utilization observed under different growth conditions, which we propose likely reflects regulatory differences, since enzymes needed to carry out the required reactions are predicted to be encoded in the genome. Below we show that these PM analyses of WT cells provide important reference points for studying the effects of mutations on the metabolic, energetic and regulatory pathways that are potentially used during various modes of growth.

Table 3-1. Substrate utilization profile of *R. sphaeroides* under different growth conditions.

Nutrient source	Previously known*	Predicted by iRsp1095**	Based on PM assay[a]		
			Photo	**Aero**	**Anaerobic[b]**
Carbon	27	64	61 (190)	68 (190)	16 (41)
Nitrogen	3	31	66 (95)		
Phosphorus	1	6	42 (59)		
Sulfur	4	4	18 (35)		
Total	35	105	187 (379)		

* Previously known growth substrates under photosynthetic conditions.

** Includes 28 false positive predictions and predictions for 20 substrates not included on the various Biolog plates. The simulations for substrate utilization in iRsp1095 were done under aerobic and photosynthetic conditions.

[a] The values in parenthesis represent the total number of carbon, nitrogen, phosphorus or sulfur sources tested using PM and related approaches.

[b] Photo – photosynthetic growth; Aero – aerobic respiratory growth; Anaerobic – anaerobic respiratory growth using DMSO as the terminal electron acceptor.

Analyzing the role of PntAB under defined growth conditions

To illustrate how knowledge of the substrate utilization profile of WT *R. sphaeroides* can be used to assess the effects of genetic perturbation on the metabolic network, we describe insights gained from analyzing an important and widely conserved energetic enzyme, pyridine nucleotide transhydrogenase (PntAB). PntAB is a heterotetrameric membrane-bound enzyme consisting of α and β subunits that catalyzes the reversible exchange of reducing equivalents between pyridine nucleotides based on the magnitude of the proton gradient across the cytoplasmic membrane [26, 27].

$$H^+_{periplasm} + NADP^+ + NADH \leftrightarrow NADPH + NAD^+ + H^+_{cytosol}$$

Thus PntAB plays a major role in maintaining the balance of cellular pyridine nucleotides (NADH/NADPH). NADPH is a source of reducing equivalents in a large number of crucial anabolic pathways such as the Calvin cycle in autotrophic cells, fatty acid biosynthesis and tetrapyrrole or pigment biosynthesis in photosynthetic organisms [28].

Extensive studies have shown that *E. coli* PntAB expression is induced when there is a demand for NADPH [29]. In addition, *E. coli* PntAB is required for optimal growth on carbon sources whose metabolism does not directly generate NADPH, such as glycerol [29]. *E. coli* also possesses an energy independent soluble transhydrogenase, UdhA, which is induced when there is an excess of NADPH (e.g., growth on acetate) and mediates conversion of NADPH to $NADP^+$ [29, 30]. In addition to PntAB and UdhA, glucose-6-phosphate dehydrogenase (Zwf) and isocitrate dehydrogenase (Icd) can help maintain bacterial NADPH pools under specific conditions [29, 30]. We compared growth of *R. sphaeroides* wild type and Δ*pntAB* cells (PntA1 [31]) using PMs to identify conditions with an increased need for NADPH in this bacterium.

PntAB is conditionally essential for photosynthetic and aerobic respiratory growth

Under photosynthetic conditions, only 25 carbon sources supported growth of PntA1 compared to 61 substrates that were used by WT cells. Importantly, only PntA1 cells using D-glucose achieved a final optical density that was equivalent to that of its parent, while PntA1 cells using D-aspartate grew well, whereas the WT parent did not grow on this substrate (Table 3-2, Additional File 1 – Table S5), suggesting these were the only tested carbon sources that supported normal photosynthetic growth in PntA1. To independently confirm the differences observed in the PM assays, we compared photosynthetic growth between PntA1, its WT parent and a PntA1 cells expressing PntAB from an IPTG inducible plasmid (Figure 3-1). The combined results of these analyses can place the carbon sources that support photosynthetic growth in *R. sphaeroides* into 3 groups (Figure 3-1A-C, Additional File 1 – Table S7). Group I carbon sources such as D-glucose, result in net production of NADPH during their metabolism, via enzymes like the glucose-6-phosphate dehydrogenase. These Group I carbon sources support comparable photosynthetic growth in WT and PntA1 cells (Figure 3-1A, Additional File 2 – Figure S2). Group II carbon sources (such as acetate) are incapable of supporting photosynthetic growth in the absence of PntAB. Metabolism of these Group II carbon sources require net consumption of NADPH, in addition to that required for anabolic processes (Figure 3-1B, Additional File 2 – Figure S2). Growth of PntA1 using Group III carbon sources (such as succinate and many others, Additional File 1 – Table S7), was significantly impaired compared to that of WT cells, and exhibited a long lag before growth commenced (Figure 3-1C). It should be noted however, that several Group III carbon sources such as succinate, failed to support photosynthetic growth on agar plates (Additional File 2 – Figure S2), reinforcing the need for PntAB activity when using these carbon sources.

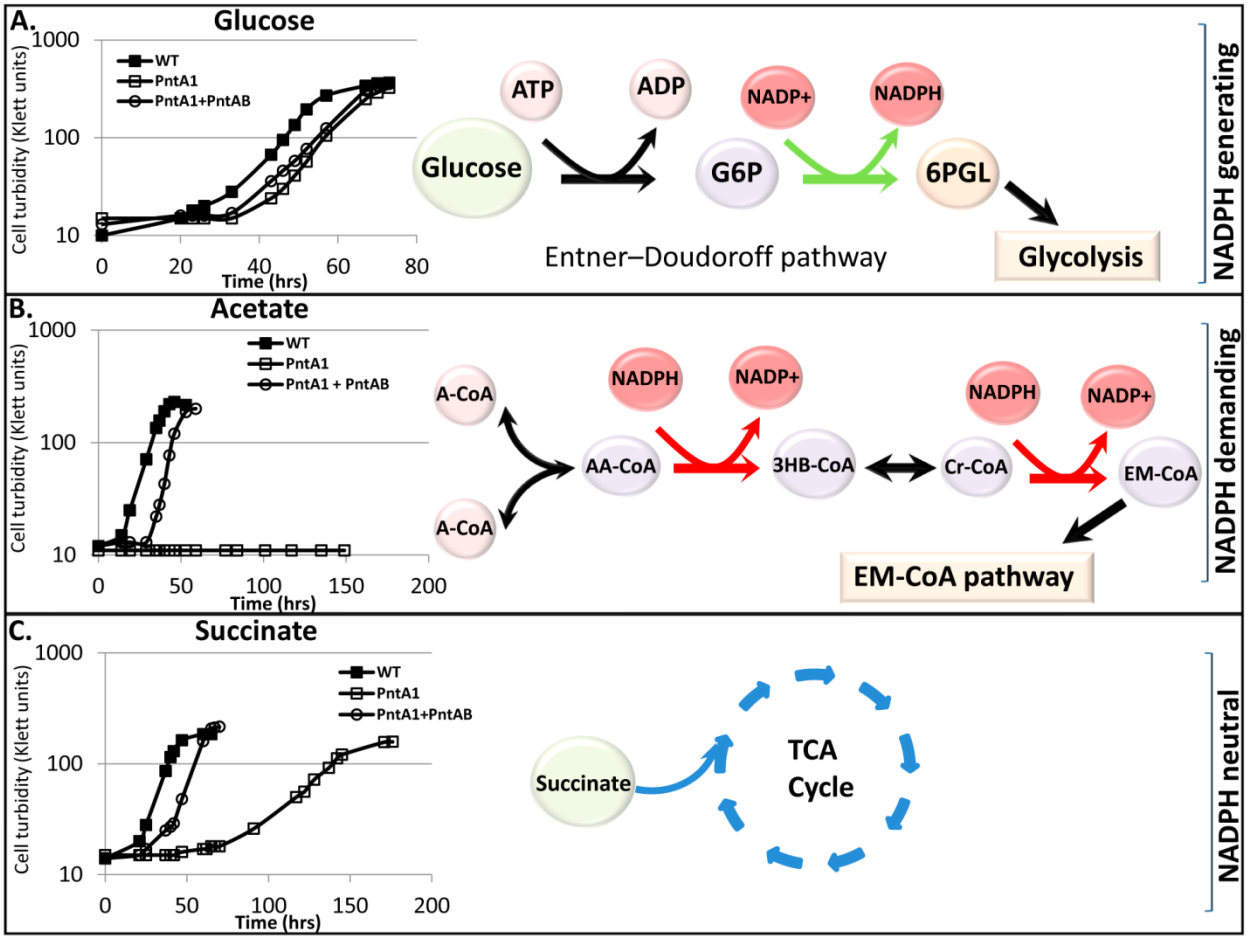

Figure 3-1. Categorization of growth substrates based on NADPH demand and requirement for PntAB under photosynthetic conditions. (A) *R. sphaeroides* growth on glucose via the Entner-Doudoroff pathway includes the NADPH-generating glucose-6-phosphate (G6P) dehydrogenase enzyme (Zwf), which supports normal photosynthetic growth even in the absence of PntAB. (B) *R. sphaeroides* growth on acetate is predicted to occur via the ethylmalonyl-CoA (EM-CoA) pathway, involving oxidation of 2 molecules of NADPH per 3 molecules of acetyl-CoA (A-CoA) consumed [59]. Note that a third molecule of A-CoA used up in later steps of the EM-CoA pathway is not depicted in this illustration. (C) During photosynthetic growth on substrates that do not directly produce NADPH (e.g. succinate), PntAB is required. In some cases, after a long lag period, cells adapt and grow photosynthetically albeit at a slower growth rate (17.2 ± 0.56 hrs compared to 7.53 ± 0.56 hrs for wild type cells). Abbreviations: AA-CoA – acetoacetyl-CoA; 3HB-CoA – 3-hydroxy butyryl-CoA; Cr-CoA – crotonyl-CoA; 6PGL – phosphoglucono-δ-lactone.

Combined, these data suggest that PntAB is the major source of NADPH for photosynthetic growth on carbon sources except glucose and D-aspartate. Furthermore, the fact that PntA1, but not its WT parent grows on D-aspartate suggests either that this strain contains unlinked mutations that allow it to metabolize this substrate or that the metabolism of the substrate is induced when the NADPH pool is reduced. This observation could also indicate that *R. sphaeroides* possesses an NADPH-linked pathway for aspartate catabolism [32, 33], even though its genome does not contain an open reading frame with significant amino acid sequence similarity to known enzymes with such an activity. In the case of Group III carbon sources (succinate and many others), other NADPH producing pathways could be activated to support growth albeit at a slower rate. For example, the delayed photosynthetic growth that is seen with some Group III carbon sources might be the result of metabolic or genetic alterations that are needed to provide PntAB-independent routes for NADPH synthesis (see below). In addition, we predict that for Group II substrates (such as acetate), the date predict that the combined NADPH demand for metabolism and anabolic processes is too high to be provided by such alternative pathways.

In contrast, PM assays show that under aerobic conditions PntA1 grows similarly to its WT parent on the many of carbon sources assessed (Figure 3-2, Table 3-2, Additional File 1 – Table S6). Of the 73 carbon sources tested, 40 allowed essentially equivalent growth (as measured by the respiratory output of the PM assays) between PntA1 and its WT parent. In addition, 23 carbon sources that allowed aerobic growth of the WT parent did not support aerobic respiration/growth of PntA1 after 96 hrs of incubation, while another 9 showed significantly reduced aerobic respiration/growth in cells lacking PntAB. Both L- and D-aspartate, supported improved aerobic respiration/growth in PntA1 cells compared to its WT parent (Figure 3-2). These results underscore potential differences in the relative need for PntAB activity under distinct metabolic states and provide a further indication that an, as yet unidentified NADPH-linked pathway for metabolism of aspartate and potentially other carbon sources exists in *R. sphaeroides*.

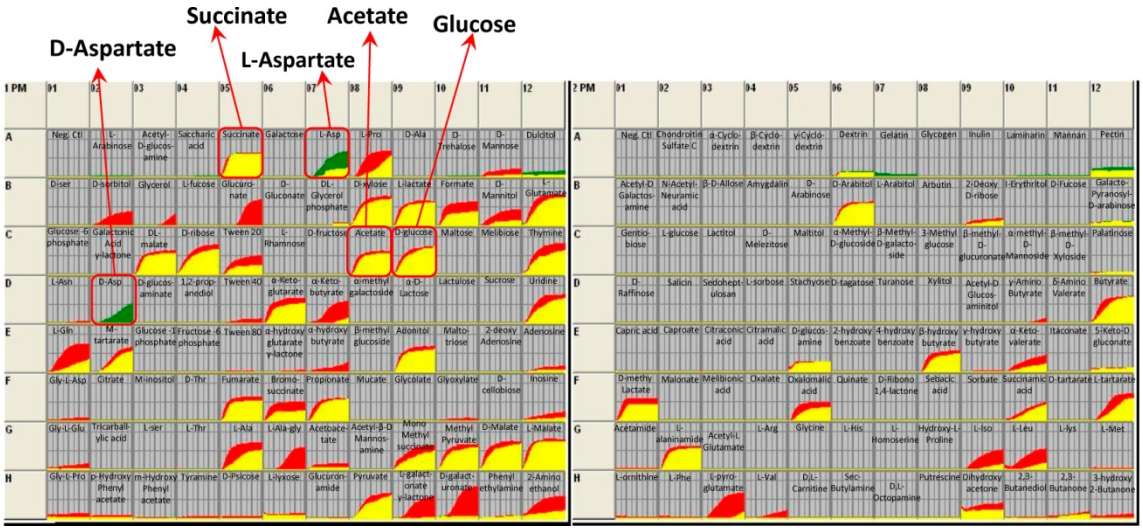

Figure 3-2. Growth data from Biolog PM1 and PM2A for WT and PntA1 under aerobic respiratory conditions. The plots compare the respiration rates across 190 carbon sources between wild type and PntA1 cells. Regions of the graph in red indicate better growth by WT cells than PntA1, while portions in green indicate better growth in PntA1. Regions in yellow represent the overlap between the kinetic data from the 2 strains.

Table 3-2. Summary of carbon utilization in WT and PntA1 cells under aerobic conditions.

Total number of Carbon sources utilized	**Photo**[a]	**Aero**
PntA1	25	51
WT	61	73
Differences in Carbon sources utilized		
Equivalent growth in PntA1 and WT	1	40
PntA1 only	1	1
Increased growth in PntA1	0	1
Reduced growth in PntA1	23	9
Growth in WT only	37	23

[a] Photo – photosynthetic growth; Aero – aerobic respiratory growth.

Alternative NADPH-generating pathways can be utilized under different growth conditions

For Group II and III carbon sources (such as acetate and succinate respectively), which require PntAB for photosynthetic growth, growth of PntA1 was indistinguishable from its WT parent under aerobic respiratory conditions (Figure 3-2, Additional File 1 – Table S6, Additional File 2 – Figure S2). One explanation for this observation is that alternative NADPH-generating reactions can function to replace PntAB under aerobic respiratory conditions. Indeed, iRsp1095, predicts 6 other routes that could generate NADPH under defined conditions in *R. sphaeroides* (Table 3-3). The relative expression of these candidate alternative NADPH-generating pathways was assayed using quantitative reverse transcriptase PCR (qRT-PCR) to determine if transcript levels for the respective genes are increased in the absence of PntAB when compared to a parent strain, when cells were grown by aerobic respiration in cultures containing one of the 3 groups of carbon sources defined above (i.e., glucose, acetate and succinate) (Figure 3-3A-H). This analysis revealed carbon source-dependent differential expression of potential alternative NADPH generation pathways. During aerobic respiratory growth on a carbon source that requires NADPH consumption (acetate), the loss of PntAB was accompanied by an ~8 fold increase in expression of a putative NADPH:ferredoxin reductase gene (RSP_1939). When using glucose, a carbon source that generates net NADPH, we found that *zwf* (RSP_2734, glucose-6-phosphate dehydrogeanse) transcript levels are increased >5 fold during aerobic respiratory growth on glucose compared to acetate or succinate, in both the WT parental strain and PntA1. Furthermore, transcripts for genes encoding putative malic enzyme (RSP_1217), isocitrate dehydrogenase (RSP_0446) and methylenetetrahydrofolate dehydrogenase (RSP_0661) enzymes were each increased >2 fold in PntA1 during aerobic respiratory growth compared to its WT parent (Figure 3-3A-H). We recognize that changes in gene expression may not necessarily result in equivalent changes in flux through the corresponding reactions in the direction of NADPH synthesis (see below). However, these data illustrate the potential of alternative NADPH-generating pathways during aerobic respiratory growth to contribute to the growth of PntA1 when using many carbon sources.

Table 3-3. Reactions predicted by iRsp1095 to be involved in NADPH generation

Gene Identifier	Enzyme name	Reaction catalyzed*	EC number
RSP_0239 & RSP_0240	PntAB	$NADP^+ + NADH + 2\ H^+[p] = > NADPH + NAD^+ + 2\ H^+$	1.6.1.2
RSP_1939	NADPH-ferredoxin reductase	Reduced ferredoxin + $NADP^+$ + H + <= > Oxidized ferredoxin + NADPH	1.18.1.2
RSP_2734	Zwf	D-Glucose 6-phosphate + $NADP^+$ < = > 6PGL + NADPH	1.1.1.49
RSP_1217 RSP_1593	Malic enzyme	(S)-Malate + $NADP^+$ = > Pyruvate + CO_2 + NADPH	1.1.1.40
RSP_0661	THF dehydrogenase	mlthf + $NADP^+$ < = > methf + NADPH	1.5.1.5
RSP_0446 RSP_1559	Icd	Isocitrate + $NADP^+$ < = > 2-Oxoglutarate + CO_2 + NADPH	1.1.1.42
RSP_1146 & RSP_1149	Glutamate synthase	2 L-Glutamate + $NADP^+$ < = > L-Glutamine + 2-Oxoglutarate + NADPH + H^+	1.4.1.13

* 6PGL – D-Glucono-1,5-lactone 6-phosphate; mlthf – 5,10-Methylenetetrahydrofolate; methf – 5,10-Methenyltetrahydrofolate; THF dehydrogenase – 5,10-methylene-tetrahydrofolate dehydrogenase.

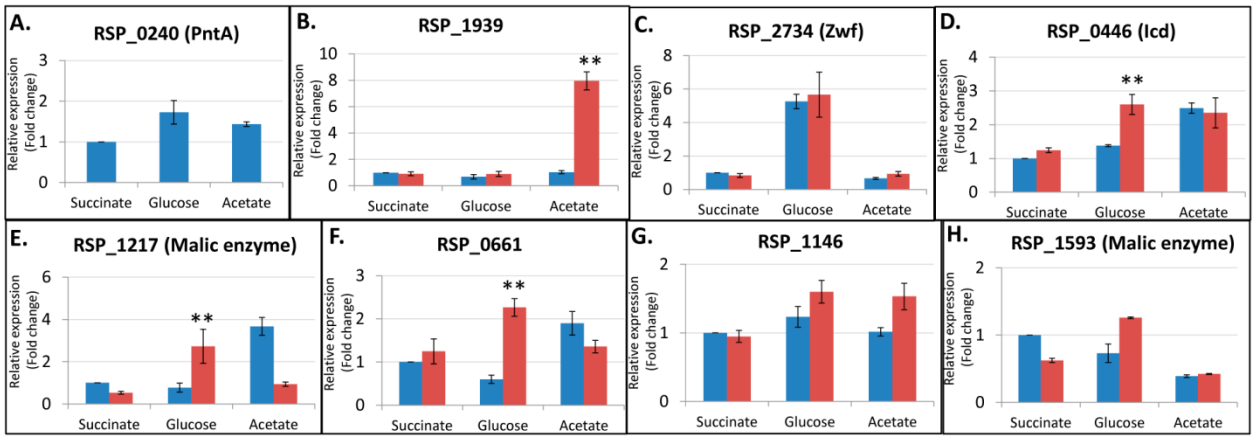

Figure 3-3. Transcript levels of genes encoding putative NADPH-generating enzymes in *R. sphaeroides* under aerobic respiratory conditions. Transcript levels for genes encoding enzymes predicted by iRsp1095 to be involved in NADPH generation (Table 3-3), were assayed via qRT-PCR in the presence or absence of PntAB. Transcript levels in WT cells is shown in blue bars (■), while those for PntA1 is in red bars (■). All fold change values represent expression relative to WT succinate-grown cells, whose relative expression was set to 1. ** Significantly increased expression in PntA1 relative to WT.

Interestingly, none of the transcript levels for the tested genes were significantly increased in the succinate-grown PntA1 cells under aerobic respiratory conditions. This could indicate that the contribution of PntAB is minor under this condition and that alternative NADPH-generating pathways provide sufficient NADPH to support growth. Alternatively, other as yet unidentified NADPH-generating pathways or post-transcriptional events might exist in both wild type and *ΔpntAB* cells.

The metabolic flux predictions from iRsp1095 predict that gluconeogenesis is one potential route which could be utilized during growth on succinate in the absence of PntAB, to produce glucose-6-phosphate, which would then be metabolized via the Entner-Doudoroff pathway to produce NADPH (Figure 3-4). To test this hypothesis, we experimentally analyzed the impact of loss of Zwf under different growth and nutrient conditions. We found that loss of Zwf alone (strain Zwf1) did not significantly impair aerobic respiratory growth on succinate, consistent with the prediction that this enzyme is not required for growth under these conditions. However, loss of Zwf in PntA1 (PntA1-Zwf1) resulted in a significant decrease in aerobic respiratory growth on succinate compared to the PntA1 mutant (Figure 3-5A), suggesting that Zwf makes a significant contribution to NADPH production, but only in the absence of PntAB, even though *zwf* transcript levels are not increased in PntA1 cells under this condition. Additionally, the observed growth difference between the Zwf1 and PntA1-Zwf1 mutants indicates that PntAB makes even a more significant contribution to NADPH production when cells are grown aerobically with succinate as a carbon source (Figure 3-5A). These data confirm the model's prediction and suggest that Zwf is a major NADPH generating enzyme utilized by the cell in the absence of PntAB during aerobic respiratory growth on succinate.

Under photosynthetic conditions, the PntA1-Zwf1 double mutant was incapable of growth on either glucose, succinate or acetate (representatives of each of the 3 major classes of carbon source we defined previously). However, growth is partially restored for glucose and succinate when the PntA-Zwf1 mutant is complemented with a plasmid containing the *zwf* gene (Figure 3-5B and D). These findings indicate that Zwf also serves as a major route for NADPH generation during photosynthetic growth with these

carbon sources, an observation that was predicted by metabolic flux analysis in iRsp1095 (Additional File 2 – Figures S3 and S4). Additionally, growth of Zwf1 cells is impaired when using glucose either photosynthetically or aerobically (Figures 3-5C and D, Additional File 2 – Figure S2), suggesting that the Entner-Doudoroff pathway is the major glycolytic pathway in *R. sphaeroides*, despite significant expression of genes encoding enzymes of the Embden-Meyerhof-Parnas pathway in these cultures. This conclusion is also consistent with both experimental analysis of ^{13}C-glucose metabolism in *R. sphaeroides* under aerobic respiratory conditions [34] and with the metabolic flux predictions made by iRsp1095 (Additional File 2 – Figure S4).

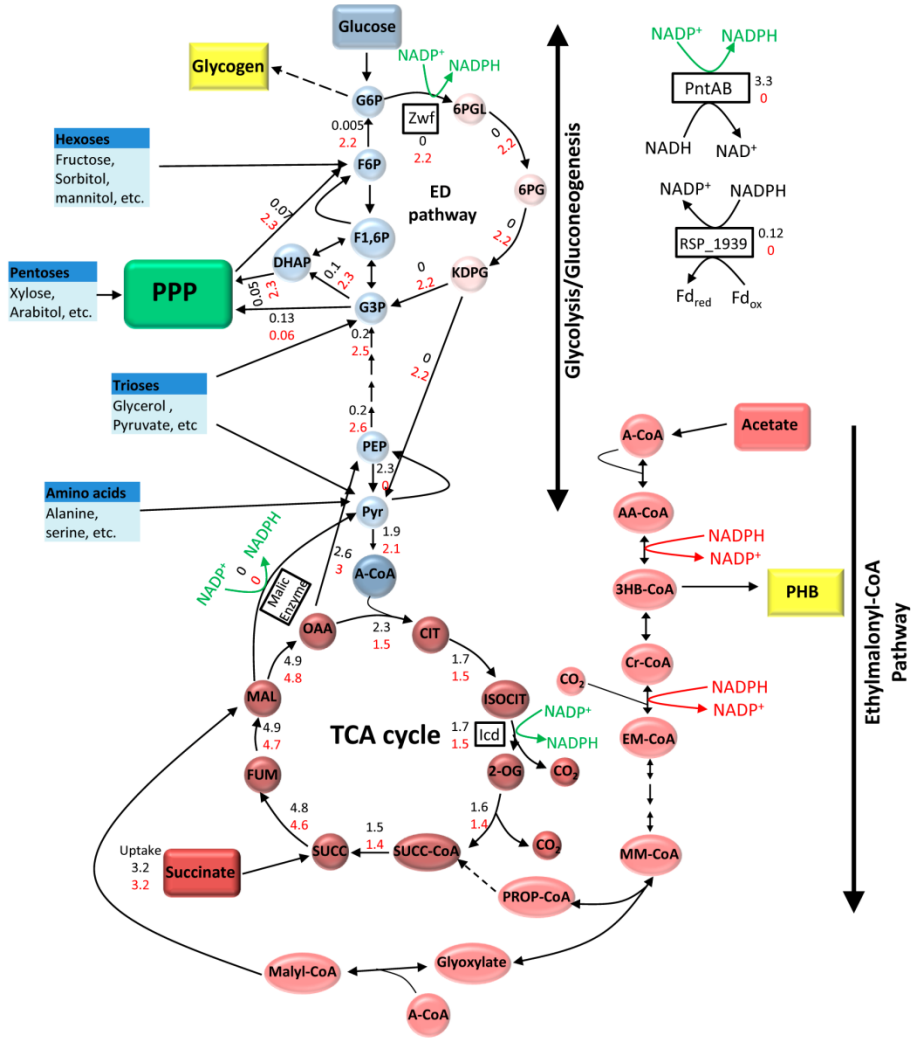

Figure 3-4. Predicted flux distributions during aerobic respiratory growth with succinate. Map of *R. sphaeroides* central carbon metabolism showing flux distributions predicted by iRsp1095 during aerobic growth with succinate. All fluxes are in mmol/gDW h. The fluxes in black are those predicted for the wild type cells, while those in red are predictions made for the Δ*pntAB* deletion strain. Reactions with no flux values are predicted to have a zero flux. To improve accuracy of predictions, fluxes were constrained using publicly available genome-wide expression data from wild type cells grown aerobically on succinate (see Methods). Green arrows indicate the predicted NADPH generating reactions under these conditions. The entry point of some other carbon sources utilized by *R. sphaeroides* are also shown (blue boxes). It should be noted that the predicted fluxes shown represent only one of many optimal solutions from the FBA solution space. Also note that *R. sphaeroides* does not possess a homolog of 6-Phosphogluconate dehydrogenase, which links the pentose phosphate pathway (PPP) with the Entner-Doudoroff (ED) pathway in some other organisms. TCA – tricarboxylic acid; G6P – Glucose 6-phosphate; F6P – Fructose 6-Phosphate; F1,6P – Fructose 1,6-bisphosphate; G3P – Glyceraldehyde 3-phosphate; DHAP – Dihydroxyacetone phosphate; 6PGL – phosphoglucono-δ-lactone; 6PG – 6-Phosphogluconate; KDPG – 2-Keto-3-deoxy-6-phosphogluconate; PEP – Phosphoenolpyruvate; Pyr – Pyruvate; A-CoA – Acetyl-CoA; CIT – Citrate; ISOCIT – Isocitrate; 2-OG – 2-oxoglutarate; SUCC-CoA – Succinyl CoA; SUCC – Succinate; FUM – Fumarate; MAL – Malate; OAA – Oxaloacetate; AA-CoA – acetoacetyl-CoA; 3HB-CoA – 3-hydroxy butyryl-CoA; Cr-CoA – crotonyl-CoA; EM-CoA – Ethylmalonyl-CoA; MM-CoA – β-methylmalyl-CoA; PROP-CoA – Propionyl-CoA; PHB – Polyhydroxybutyrate.

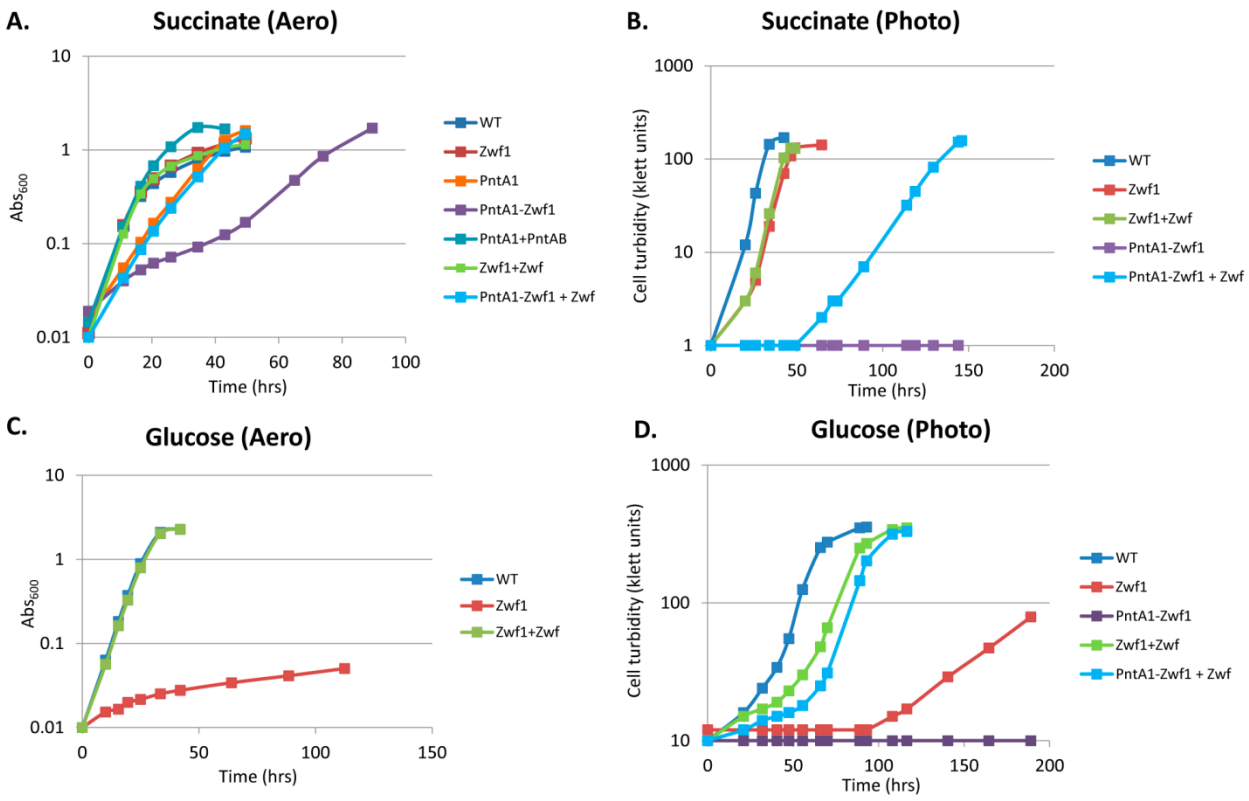

Figure 3-5. The role of Zwf during growth with succinate. Comparison of the growth rates of wild type (WT), PntA1 and PntA1+PntAB cells, to *zwf* deletion strains Zwf1 and PntA1-Zwf1 and complemented strains Zwf1+Zwf and PntA1-Zwf+Zwf during aerobic respiratory (A and C) and photosynthetic (B and D) growth using succinate or glucose, respectively.

The conditional ability of alternative pathways to compensate for the loss of PntAB could reflect a higher demand for NADPH during photosynthetic or other energetic conditions. For example, under photosynthetic conditions, *R. sphaeroides* produces significantly larger amounts of fatty acids, photopigments and other components of the photosynthetic apparatus [35] that are each synthesized via NADPH-dependent pathways. Indeed, iRsp1095 predicts there is a 2–4 fold increase in the demand for NADPH under photosynthetic conditions in cells using succinate or glucose as a major carbon source (Figure 3-6). If this prediction is accurate, our data suggests that this significantly greater demand for NADPH during photosynthetic growth cannot be met by using alternative NADPH-generating pathways that are sufficient under aerobic respiratory conditions. However, iRsp1095 also predicts that the need for NADPH during aerobic respiratory growth in the presence of acetate is greater than that required for photosynthetic growth on either succinate or glucose, indicating that the cell might have the metabolic capacity to generate sufficient amount NADPH to support photosynthetic growth via these alternative pathways. Thus, additional studies are needed to determine if there are additional constraints under photosynthetic conditions which make alternative NADPH-generating pathways that function under aerobic respiratory conditions insufficient to support growth under other energetic states. Combined, these data illustrate the centrality, and previously unrecognized importance, of the routes for NADPH production in the photosynthetic lifestyle of *R. sphaeroides*. Given the ubiquitous nature of PntAB across biology, it is possible that this enzyme plays a major energetic role in other photosynthetic and non-photosynthetic organisms.

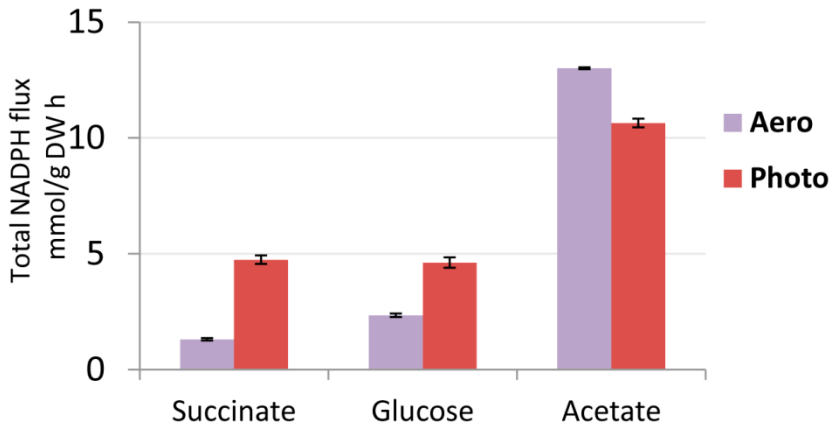

Figure 3-6. Comparison of the predicted NADPH flux during photosynthetic or aerobic respiratory growth. Predictions were made for the total NADPH flux during growth with succinate, glucose and acetate under photosynthetic and aerobic respiratory conditions. The error bars represent the standard error of mean of 1000 fluxes through this reaction obtained from 1000 alternative optimal solutions (see Methods). It should be noted that the predicted larger NADPH demand under aerobic conditions compared to photosynthetic conditions during growth with acetate is due to a larger predicted uptake rate of acetate under aerobic conditions.

iRsp1140: a revised experimentally-validated genome-scale metabolic reconstruction for *R. sphaeroides*

The above results provide new information about the metabolic capabilities of *R. sphaeroides* that can be utilized to refine biochemically, genetically and genomically structured databases employed in constraint-based analysis [36]. We previously constructed and validated a genome-scale metabolic reconstruction for *R. sphaeroides*, called iRsp1095, using its annotated genome, published organism-specific data and information from continuous cultures of WT cells [11].

However, the substrate utilization predictions of iRsp1095 could not explain a large amount of the data obtained during PM analysis of wild type and mutant cells (Figure 3-7A). Consequently, we performed a 2-step refinement and extension of iRsp1095 that involved addition and removal of appropriate reactions, metabolites and genes to increase its agreement with the PM data, while also incorporating newly available metabolic data for *R. sphaeroides* since release of iRsp1095 (see Methods). An initial refinement process resulted in addition of 280 reactions, 81 metabolites and 27 genes to iRsp1095, while removing 23 reactions and 3 genes (Additional File 3 – Tables S1–3). Ninety-six of these 280 additional reactions consisted of newly added enzymatic activities, for which no apparent genes were present in the annotated *R. sphaeroides* genome. To identify candidate enzymes for these reactions, we conducted new BLAST analyses utilizing protein sequence of experimentally verified enzymes from other organisms. These searches resulted in the identification and annotation of 6 new candidate enzymes. An additional 11 enzymes already included in iRsp1095, with previously predicted functions, were identified as candidates to catalyze new reactions in the model (Additional File 3 – Tables S4).

Of the 280 reactions added to iRsp1095, 92 are predicted membrane transport reactions. Organisms utilize a variety of transporters to import nutrients [37, 38]. However, *R. sphaeroides* appears to favor the use of ATP-dependent ABC transporters (~7.84% of its genome is dedicated to ABC transport functions compared to 4.9% in *E. coli* [39]). An analysis of the *R. sphaeroides* genome reveals that of the genes encoding ABC transporters with no known or predicted functions, 11 are organized in operons encoding proteins with an ATPase domain, an integral membrane permease domain and a substrate-specific substrate binding domain (typical of importers [40]) (Table 3-4). To test the function of these putative transport operons, we deleted the gene for each of these 11 predicted substrate binding proteins and assessed growth of the resulting mutants using PM. This analysis, plus subsequent liquid growth studies led to identification of substrate-specific growth defects in mutants lacking individual genes for candidate transporters (Additional File 2 – Figure S6). Based on this data set we have provisionally identified substrates for 5 previously uncharacterized transporters (Table 3-4). These transport reactions were incorporated into iRsp1095 by the addition of 15 genes, 2 reactions and one metabolite (Additional File 3 – Tables S4).

A.

Nutrient	iRsp1095		Agreement[a]
Carbon	31 G-G 30 NG-G	21 G-NG 108 NG-NG	73.2%
Nitrogen	20 G-G 46 NG-G	5 G-NG 24 NG-NG	46.3%
Phosphorus	4 G-G 38 NG-G	2 G-NG 15 NG-NG	32%
Sulfur	1 G-G 17 NG-G	0 G-NG 17 NG-NG	51%

B.

Nutrient	iRsp1140		Agreement[a]
Carbon	53 G-G 8 NG-G	5 G-NG 124 NG-NG	93.2%
Nitrogen	53 G-G 13 NG-G	2 G-NG 27 NG-NG	84.2%
Phosphorus	33 G-G 9 NG-G	0 G-NG 17 NG-NG	84.7%
Sulfur	13 G-G 5 NG-G	0 G-NG 17 NG-NG	85.7%

Figure 3-7. *R. sphaeroides* substrate utilization and transport. A detailed comparison between the predicted substrate utilization in iRsp1095 (A) and iRsp1140 (B) with data obtained from PM analysis. G-G – **G**rowth (predicted by model) and **G**rowth (observed in PM); G-NG – **G**rowth (predicted by model) and **N**o **G**rowth (observed in PM); NG-G – **N**o **G**rowth (predicted by model) and **G**rowth (observed in PM); NG-NG – **N**o **G**rowth (predicted by model) and **N**o **G**rowth (observed in PM). [a]Percentage represents agreement between predictions and PM data from photosynthetic analysis only.

Table 3-4. Predicted *R. sphaeroides* ABC transporter operons tested for substrate specificity using Biolog PM

Operons	Identified substrate[a]
RSP_0200-1,RSP_6101	Ribitol
RSP_0342-5	Methyl D-lactate
RSP_0346-9	Asparagine
RSP_1442-5	D-serine
RSP_2208-11	Uridine
RSP_0644-6	ND
RSP_2596-8	ND
RSP_2809-11	ND
RSP_3166-8	ND
RSP_3500-3	ND
RSP_3557-60	ND

[a] ND - Not Determined. Specific substrates for these transporters could not be determined from PM assay and the analysis of mutants lacking single genes in each of these predicted operons (see text).

These updates resulted in a refined metabolic model for *R. sphaeroides*, iRsp1140, consisting of 1416 reactions, 878 metabolites and accounting for 1140 genes within its genome (other properties of iRsp1140 are summarized in Table 3-5). Overall, iRsp1140 includes a larger number of reactions, metabolites and genes, and it is supported by increased physiological, genetic and genomic evidence for the inclusion of many components in the network. The predictions of iRsp1140 are in better agreement with the PM data compared with those of iRsp1095 (Figure 3-7), verifying the improved predictive ability of iRsp1140. In addition, predictions made by iRsp1140 for growth rate or production of H_2, polyhydroxybutyrate and CO_2 evolution were in excellent agreement with published data [11] (Additional File 3 – Table S5). We expect that iRsp1140 will enable improved modeling of metabolic behavior under a larger number of conditions. In addition, the increased number of genes in iRsp1140 provides a larger set of targets for assessing the effects of genetic perturbations on growth or metabolic activity under both photosynthetic and non-photosynthetic conditions.

Table 3-5: Comparison of the properties of iRsp1095 and iRsp1140

Categories	iRsp1095	iRsp1140
Genes	**1095**	**1140**
Genes based on experimental evidence	334 (30.5%)	354 (31%)
Genes inferred based on gene homology	761 (69.5%)	786 (69%)
Metabolites	**796**	**878**
Reactions	**1158**	**1416**
Enzymatic Reactions	858	953
Transport reactions	300	463
Reactions Associated with genes	1049	1204
Spontaneous/Diffusion reactions	14	17
Reactions without gene association	95	195
Reversible Reactions	757	942
Irreversible Reactions	401	474
Exchange Reactions	148	231
Demand Reactions	3	3

Conclusions

While the development of high-throughput or global approaches can provide large amounts of data, the task of extracting meaningful biological insight from this information is still challenging [4]. To gain new biological insights, constraint-based and other modeling approaches can be used to integrate various data sets [1-3, 41].

In this study, we took an integrated approach to gain new insight into the metabolic, energetic and photosynthetic lifestyles of *R. sphaeroides*. We extended the number of nutrients that can support growth of WT cells. We also showed that a conserved bioenergetic enzyme (PntAB) which can provide reduced pyridine nucleotides is essential for photosynthetic growth on many carbon sources. We used a genome-scale model for *R. sphaeroides* to make flux predictions, as well as generate and test hypotheses on alterative NADPH-generating pathways that allow growth in the absence of PntAB. The products of various anabolic pathways require NADPH that is derived mainly from PntAB activity, so exploiting these and other alternative NADPH generating pathways we identified, could improve growth and metabolic end products derived from photosynthetic and non-photosynthetic wild type cells or those engineered to produce compounds of biotechnological interest.

Previous studies have shown the utility of high-throughput data sets in refining and validating genome-scale metabolic models [42-44]. We used our data to produce a 2^{nd} generation genome-scale reconstruction for *R. sphaeroides*, iRsp1140 with significantly improved coverage of metabolic functions and predictions that are in better agreement with experimental data. iRsp1140, provides an improved depiction of the *R. sphaeroides* metabolic network, so it will be useful in studying photosynthesis, as well as a wider range of metabolic processes in this and related organisms.

Methods

Bacterial strains and growth conditions

R. sphaeroides strains 2.4.1 and Ga were used as parental strains. All mutants were made in strain 2.4.1 except PntA1 [31] and Zwf1, which were constructed in Ga, and PntA1-Zwf1, which was constructed in PntA1 (Additional File 4 – Table S1). *E. coli* DH5α was used as a plasmid host, and *E. coli* S17-1 was used to conjugate DNA into *R. sphaeroides*.

R. sphaeroides cultures were incubated at 30 ^{o}C in Sistrom's Minimal Medium (SMM) [45] lacking glutamate and aspartate, with succinate (33.9 mM), or an alternative sole carbon source. The molar concentration of carbon atoms of the carbon source was kept constant at 135.5 mM. Photosynthetic cultures were incubated in screw capped tubes at a light intensity of ~10 W/m^2, while anaerobic respiratory cultures were incubated in screw capped tubes in the dark with the media supplemented with 0.9% DMSO. Aerobic cultures were shaken in flasks. Optical density of photosynthetic cells in screw capped tubes was measured using a Klett-Summerson photometer and is expressed in Klett units (1 Klett unit equals approximately 10^7 cells/mL). Other optical density measurements were made by measuring optical density at 600 or 650 nm in a spectrophotometer. When required, the media was supplemented with 100 μM IPTG, 25 μg/mL kanamycin or 25 μg/mL spectinomycin. *E. coli* cells were grown in Luria Bertani medium at 37 ^{o}C, supplemented with 25 μg/mL kanamycin where needed.

Phenotype microarray analysis

To determine the substrate utilization profiles of *R. sphaeroides*, phenotype microarrays (PMs) were used with a few modifications. To assay aerobic respiratory growth on different carbon sources (Biolog PM1 and PM2A), cells were grown on SMM agar plates aerobically for 3 days. Cells were swabbed from the agar plates and suspended in 4 mL inoculation fluid (Biolog inc.) to an OD_{600} of 0.38. Two mL of this mixture was placed in 10 mL of inoculation fluid (IF) containing 24 μL of tetrazolium-based dye A

(Biolog inc.), resulting in a final OD_{600} of ~0.07. Then 1.2 µL of vitamin solution (1 g nicotinic acid, 0.5 g thiamine-HCl and 0.01 g biotin in 100 mL of water) was added and 100 µL was dispensed into each well of a 96 well plate. Cultures were incubated at 30˚ C for 72 to 96 hrs in an OmniLog plate reader (Biolog inc).

To assay photosynthetic growth, 10mM $NaHCO_3$, 0.4mM sodium thioglycolate and 1 µM methylene green were added to the IF and this was kept in an anaerobic chamber for ~7 days with periodic shaking to facilitate it becoming anaerobic (the methylene green in the IF turns colorless once oxygen is removed). *R. sphaeroides* cells were grown photosynthetically on SMM agar plates and PM plates set up as described above, except that the tetrazolium dye was omitted, as thioglycolate reduces the tetrazolium-based dye turning it purple independent of cellular respiration. The PM plates were put in an anaerobic chamber, inoculated, placed in heat sealed anaerobic bags (Biolog inc.) [46] and incubated under constant illumination (light intensity of ~10 W/m^2) at 30˚ C for 72 to 96 hrs, after which OD_{650} readings were taken. Anaerobic indicator strips and ageless oxygen absorbers (MITSUBISHI Gas Chemical America, Inc.) were placed in the sealed bags to report on and maintain an anaerobic environment throughout the experiment.

Any growth in the negative control well (A1) was subtracted from the measured optical density for both aerobic and photosynthetic PM. A threshold OD_{650} of 0.05 (after background correction) was used as a baseline for scoring photosynthetic growth, as this was the highest value obtained from any well in control experiments where cells were kept in the dark. A threshold of 5 Omnilog units (after background correction) was used as a baseline for aerobic respiratory growth. Only carbon sources that resulted in reproducible growth above the baseline across all replicates were considered to be growth substrates.

To assay photosynthetic growth on different nitrogen, phosphorus or sulfur sources (Biolog PM3B and PM4A), *R. sphaeroides* cells were grown aerobically for 5 days on a modified R2A agar [22, 47] (0.25 g of yeast extract, 0.25 g of Proteose Peptone, 0.25 g of Casamino Acids, 0.12 g of K_2HPO_4, 0.025 g of

$MgSO_4.7H_2O$, 0.5 g of sodium pyruvate and 15 g of agar per liter of water). Plates were set up as described for photosynthetic growth with the addition of 20mM sodium succinate and 2μM ferric citrate. Cell cultures were grown under constant illumination (10 W/m^2) at 30˚ C for 48 hrs, after which OD_{650} readings were taken. Aerobic growth on the different nitrogen, phosphorus and sulfur sources is not reported due to significant background growth in the negative control wells, an issue that has been observed previously [25].

Due to comparatively slow growth rates of *R. sphaeroides* under anaerobic respiratory conditions, evaluation of these growth modes could not be reliably conducted with Biolog PM plates due to media evaporation. Thus, to assay anaerobic respiratory growth on different carbon sources, 96 well microwell plates (Fisher Scientific) were set up to analyze 41 of the carbon sources identified as *R. sphaeroides* growth substrates from PM (see Additional File 1 – Table S1 for a list of these substrates). The carbon sources were normalized for the total number of carbon atoms in each compound (135.5 mM). *R. sphaeroides* cells were grown aerobically in SMM and centrifuged. Cells were then washed with SMM media lacking a carbon source (SMM no carbon), suspended in anaerobic SMM no carbon (which had been kept in an anaerobic chamber for at least 4 days) to an OD_{600} of ~0.1 and DMSO was added to a final concentration of 0.9%. Then, 300 μL of a cell suspension was dispensed into wells containing a different carbon source in an anaerobic chamber. Plates were incubated at 30˚ C for 10 days with continuous shaking in a Tecan M200 plate reader located within the anaerobic chamber, with OD_{650} readings taken every 6 minutes. Alternatively, plates were sealed in anaerobic bags, wrapped in foil and incubated at 30˚ C for 10 days with periodic shaking.

Construction of mutants

All *R. sphaeroides* mutants we constructed contained in-frame markerless deletions. Briefly, regions spanning ~1500 bp upstream and downstream of the target gene were amplified using sequence specific primers containing restriction sites for EcoRI, XbaI or HindIII. These fragments were digested with the

appropriate restriction enzymes and ligated into pK18mobsacB plasmid [48], digested with EcoRI and HindIII, by three-way ligation to generate the various gene deletion constructs, which were confirmed by sequencing (Additional File 4 – Table S1 and S2). The pK18mobsacB-based plasmids were separately mobilized from *E. coli* S17-1 into *R. sphaeroides* strains. Cells in which the plasmid had successfully integrated into the genome via homologous recombination were identified by selection on SMM plates supplemented with kanamycin. These cells were then grown overnight in SMM without kanamycin. Cells that had lost the deletion plasmid (and thus the *sacB* gene) via a second recombination event were identified by growth on SMM plates supplemented with 10% sucrose. Individual gene deletions were confirmed by PCR and sequencing with specific primers.

Ectopic expression plasmids were made by amplifying the target genes from the genome using sequence specific primers containing restriction sites for NdeI and BglII, HindIII or BamHI. These DNA fragments were digested with the appropriate enzymes and cloned into pIND5 digested with the same enzymes. These plasmids were conjugated from *E. coli* S17-1 into the relevant *R. sphaeroides* mutant. Cells which harbor the desired plasmid were identified by selection on SMM plates supplemented with kanamycin.

RNA extraction, qRT-PCR and microarray analyses

RNA was isolated from exponential phase cultures of *R. sphaeroides* strains that were grown either photosynthetically in 16 mL screw cap tubes or aerobically in 500 mL conical flasks. RNA isolation and subsequent cDNA synthesis were performed as previously described [49]. qRT-PCR experiments were conducted in triplicate for each biological replicate using SYBR Green JumpStart Taq ReadyMix (Sigma-Aldrich). Relative expression was determined via the $2^{-\Delta\Delta C}{}_{T}$ method with efficiency correction [50]. *R. sphaeroides rpoZ* was used as a housekeeping gene for normalization. Primers used in this analysis are provided in Additional File 4 – Table S2.

Constraint based analysis and model refinement

Separate stoichiometric matrices, $S_{m \times n}$, were generated from the reconstructions (i.e., iRsp1095 and iRsp1140) with the rows (m) representing metabolites, the columns (n) representing reactions and entries in the matrices representing stoichiometric coefficients for metabolites involved in each reaction. Flux balance analysis (FBA) [51] was used to simulate *in silico* growth by solving the linear programming problem:

$$\max v_{Biomass}$$

$$\text{s.t}$$

$$\mathbf{S \bullet v = 0}$$

$$\mathbf{v_{min} \leq v \leq v_{max}}$$

where $v_{Biomass}$ is the flux through biomass objective function; $\mathbf{v}$ is the vector of steady state reaction fluxes; and $\mathbf{v}_{min}$ and $\mathbf{v}_{max}$ are the minimum and maximum allowable fluxes. The values in $\mathbf{v}_{min}$ and $\mathbf{v}_{max}$ were set to -100 and 100 mmol/g DW h for reversible reactions and 0 and 100 mmol/g DW h for forward only reactions. During simulation, all exchange reactions were assigned as being forward only (allowing metabolites to be secreted into the medium but not taken up), except the exchange reactions for media components required for cell growth, which were set to measured values for limiting substrates such as ammonia, or allowed to be freely exchanged with the extracellular space, i.e., $-100 \leq v \leq 100$. The non-growth associated ATP maintenance limit was set to 8.39 mmol/gDW h [24]. Flux variability analysis [52] was also used to determine minimum and maximum allowable flux through reactions in the network.

Initial simulations with iRsp1095 in which the transhydrogenase reaction was deleted resulted in the prediction of optimal growth, suggesting alternative NADPH generating reactions existed in the metabolic network. Analysis of iRsp1095 revealed that it includes at total of 61 NADPH requiring reactions, of which only 29 were independently non-essential and capable of functioning in the direction of NADPH synthesis. To identify all reactions within iRsp1095 capable of producing NADPH to support

growth, all 29 non-essential NADPH-requiring reactions within iRsp1095, capable carrying flux the direction of NADPH synthesis, were turned off. This resulted in a predicted growth rate of 0. Optimal growth was restored solely by turning on the transhydrogenase reaction, consistent with transhydrogenase being sufficient for generating NADPH required for growth. All other reactions capable of independently restoring growth, while the other NADPH-requiring reactions were still off, were considered as a candidate NADPH producing reaction (Table 3-3).

To assess the predicted NADPH demand during aerobic or photosynthetic growth across growth substrates (i.e., succinate, glucose and acetate), all predicted NADPH generating reactions (Table 3-3) set to have a zero flux, except the transhydrogenase reaction. Using a previously described mixed integer linear programming approach [53, 54], 1000 alternative optimal FBA solutions were identified under each condition. The flux through the transhydrogenase reaction, and thus the amount of NADPH predicted to be required, under each condition was averaged over the 1000 alternative optimal solutions and this average was used as an estimate of NADPH demand under those conditions.

To predict fluxes through central metabolism, we used an extension of FBA called E-flux [55]. E-flux limits the maximum and minimum fluxes ($\mathbf{v}_{max}$ and $\mathbf{v}_{min}$ respectively) through the reactions in the network based on genome-wide gene expression measurements. To achieve this publicly available microarray data obtained from cells grown on succinate and glucose (GEO platform GPL162), as well as from cells grown acetate (this study), were normalized and used to constrain the fluxes through each reaction in the network as previously described [55]. For reactions without gene-to-protein-to-reaction (GPR) assignments, the fluxes through these reactions were allowed to have a $\mathbf{v}_{max}$ of 100 mmol/g DW h and a $\mathbf{v}_{min}$ of 0 or -100 mmol/g DW h, if the reactions were forward only or reversible respectively. For reactions catalyzed by isozymes, the expression value of gene for the isozyme with the highest expression was used to constrain the reaction, while for multi-subunit enzymes the gene for the subunit with the lowest expression was used to constrain the reaction. After setting the upper and lower bounds, subsequent simulations were conducted with FBA as described above.

The previously published genome-scale model for *R. sphaeroides* iRsp1095 [11] was used as the starting point for a 2-step model refinement. In the first step, PM data was used to guide model refinement, which involved the manual addition and removal of reactions to bring it into better alignment with the PM data. PM data for carbon (C), nitrogen (N), phosphorus (N) and sulfur (S) utilization were compared to model predictions from FBA simulations in which equivalent compounds were provided as the sole sources of these nutrients. The uptake rate of the tested carbon source was set to -4 mmol/g DW h, while that of N, P or S sources was set to -1 mmol/g DW h, as these are in the range of uptake rates typically observed in *R. sphaeroides* [11]. Succinate was used as the C source when testing for N, P and S utilization, while ammonium, inorganic phosphate and inorganic sulfate served as N, P and S sources when accessing C utilization (consistent with the PM analysis). For these simulations, nutrients which resulted in a FBA predicted growth rate greater than zero were considered growth substrates. No Growth-Growth (NGG – no growth predicted by model but growth observed in PM) inconsistencies were manually rectified by addition of appropriate transport and/or enzymatic reactions from the multi-organism databases KEGG [56] and BRENDA [57, 58]. The required enzymatic reaction(s) were added to the model based on one of the following 2 criteria: (i) the presence genes in the *R. sphaeroides* genome encoding the proteins potentially capable of catalyzing the new reaction(s) to be added; and (ii) if no putative enzymes were identified, the metabolic route that required the addition of the fewest reactions to iRsp1095 was selected. Growth-No Growth (GNG – growth predicted by model but no growth observed in PM) inconsistencies in iRsp1095 were resolved by removal of transport reactions for the substrate in question, when this did not introduce and new inconsistencies with the PM data.

In a second step of model refinement, putative enzymes capable of catalyzing reactions added to iRsp1095 where identified by BLAST searches using the protein sequences from other organisms previously verified to carry out the reaction in question. A BLAST E-value cutoff of $10e\text{-}^{20}$ was selected as a threshold for significance. Enzymatic functions not previously included in iRsp1095 and which were encoded by genes without any previously defined specific function were considered as newly annotated

genes, while those with previously defined putative functions were considered as having additional functionality (Additional File 3 – Table S4). Updated information from KEGG [56] database, new information obtained from mutant analysis in this study, and data from recent literature searches were incorporated to generate iRsp1140. iRsp1140 in SBML format is provided in Additional File 5 and can be accessed in the BioModels database with ID MODEL1304240000.

References

1. Cavill R, Kamburov A, Ellis JK, Athersuch TJ, Blagrove MS, Herwig R, Ebbels TM, Keun HC: **Consensus-phenotype integration of transcriptomic and metabolomic data implies a role for metabolism in the chemosensitivity of tumour cells**. *PLoS Comput Biol* 2011, **7**(3):e1001113.
2. Hirai MY, Yano M, Goodenowe DB, Kanaya S, Kimura T, Awazuhara M, Arita M, Fujiwara T, Saito K: **Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana***. *Proc Natl Acad Sci U S A* 2004, **101**(27):10205-10210.
3. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L: **Integrated genomic and proteomic analyses of a systematically perturbed metabolic network**. *Science* 2001, **292**(5518):929-934.
4. Joyce AR, Palsson BO: **The model organism as a system: integrating 'omics' data sets**. *Nat Rev Mol Cell Biol* 2006, **7**(3):198-210.
5. Ratnakumar S, Hesketh A, Gkargkas K, Wilson M, Rash BM, Hayes A, Tunnacliffe A, Oliver SG: **Phenomic and transcriptomic analyses reveal that autophagy plays a major role in desiccation tolerance in *Saccharomyces cerevisiae***. *Mol Biosyst* 2011, **7**(1):139-149.
6. Oberhardt MA, Palsson BO, Papin JA: **Applications of genome-scale metabolic reconstructions**. *Mol Syst Biol* 2009, **5**:320.
7. Atsumi S, Higashide W, Liao JC: **Direct photosynthetic recycling of carbon dioxide to isobutyraldehyde**. *Nat Biotechnol* 2009, **27**(12):1177-1180.
8. Gronenberg LS, Marcheschi RJ, Liao JC: **Next generation biofuel engineering in prokaryotes**. *Curr Opin Chem Biol* 2013, **17**(3):462-471.
9. Jaschke PR, Saer RG, Noll S, Beatty JT: **Modification of the genome of *Rhodobacter sphaeroides* and construction of synthetic operons**. *Methods Enzymol* 2011, **497**:519-538.
10. Tabita FR: **The biochemistry and metabolic regulation of carbon metabolism and CO_2-fixation in purple bacteria**. In: *Anoxygenic photosynthetic bacteria.* Edited by Blankenship RE, Madigan MT, Bauer CE. The Netherlands: Kluwer Academic Publishers; 1995: 885–914.
11. Imam S, Yilmaz S, Sohmen U, Gorzalski AS, Reed JL, Noguera DR, Donohue TJ: **iRsp1095: a genome-scale reconstruction of the *Rhodobacter sphaeroides* metabolic network**. *BMC Syst Biol* 2011, **5**:116.
12. Mackenzie C, Eraso JM, Choudhary M, Roh JH, Zeng X, Bruscella P, Puskas A, Kaplan S: **Postgenomic adventures with *Rhodobacter sphaeroides***. *Annu Rev Microbiol* 2007, **61**:283-307.
13. Yilmaz LS, Kontur WS, Sanders AP, Sohmen U, Donohue TJ, Noguera DR: **Electron partitioning during light- and nutrient-powered hydrogen production by *Rhodobacter sphaeroides***. *Bioenerg Res* 2010, **Volume**(1):55 - 66.
14. Kim E, Lee M, Kim M, Lee JK: **Molecular hydrogen production by nitrogenase of *Rhodobacter sphaeroides* and by Fe-only hydrogenase of *Rhodospirillum rubrum***. *International Journal of Hydrogen Energy* 2008, **33**(5):1516-1521.
15. Kontur WS, Ziegelhoffer EC, Spero MA, Imam S, Noguera DR, Donohue TJ: **Pathways involved in reductant distribution during photobiological H_2 production by *Rhodobacter sphaeroides***. *Appl Environ Microbiol* 2011, **77**(20):7425-7429.
16. Kien NB, Kong IS, Lee MG, Kim JK: **Coenzyme Q10 production in a 150-l reactor by a mutant strain of *Rhodobacter sphaeroides***. *J Ind Microbiol Biotechnol* 2010, **37**(5):521-529.
17. Khatipov E, Miyake, M., Miyake J. and Y. Asada: **Polyhydroxybutyrate accumulation and hydrogen evolution by *Rhodobacter sphaeroides* as a function of nitrogen availability**. *Biohydrogen* 1999, **III**:157 - 161.

18. Sasaki K, Morikawa H, Kishibe T, Mikami A, Harada T, Ohta M: **Practical removal of radioactivity from sediment mud in a swimming pool in Fukushima, Japan by immobilized photosynthetic bacteria**. *Biosci Biotechnol Biochem* 2012, **76**(4):859-862.
19. Connor MR, Atsumi S: **Synthetic biology guides biofuel production**. *J Biomed Biotechnol* 2010.
20. Wahlund TM, Conway T, Tabita FR: **Bioconversion of CO_2 to ethanol and other compounds**. *American Chemical Society Division of Fuel Chemistry* 1996, **3**:1403-1405.
21. Bochner BR: **New technologies to assess genotype-phenotype relationships**. *Nat Rev Genet* 2003, **4**(4):309-314.
22. Bochner BR, Gadzinski P, Panomitros E: **Phenotype microarrays for high-throughput phenotypic testing and assay of gene function**. *Genome Res* 2001, **11**(7):1246-1255.
23. Borglin S, Joyner D, DeAngelis KM, Khudyakov J, D'Haeseleer P, Joachimiak MP, Hazen T: **Application of phenotypic microarrays to environmental microbiology**. *Curr Opin Biotechnol* 2012, **23**(1):41-48.
24. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO: **A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information**. *Mol Syst Biol* 2007, **3**:121.
25. Oh YK, Palsson BO, Park SM, Schilling CH, Mahadevan R: **Genome-scale reconstruction of metabolic network in *Bacillus subtilis* based on high-throughput phenotyping and gene essentiality data**. *J Biol Chem* 2007, **282**(39):28791-28799.
26. Bragg PD: **Site-directed mutagenesis of the proton-pumping pyridine nucleotide transhydrogenase of *Escherichia coli***. *Biochim Biophys Acta* 1998, **1365**(1-2):98-104.
27. Bragg PD, Davies PL, Hou C: **Function of energy-dependent transhydrogenase in *Escherichia coli***. *Biochem Biophys Res Commun* 1972, **47**(5):1248-1255.
28. Nelson DL, Cox MM: **Lehninger: Principles of biochemistry**. New York: W. H. Freeman and Company; 2005.
29. Sauer U, Canonaco F, Heri S, Perrenoud A, Fischer E: **The soluble and membrane-bound transhydrogenases UdhA and PntAB have divergent functions in NADPH metabolism of *Escherichia coli***. *J Biol Chem* 2004, **279**(8):6613-6619.
30. Lee HC, Kim JS, Jang W, Kim SY: **High NADPH/NADP+ ratio improves thymidine production by a metabolically engineered *Escherichia coli* strain**. *J Biotechnol* 2010, **149**(1-2):24-32.
31. Hickman JW, Barber RD, Skaar EP, Donohue TJ: **Link between the membrane-bound pyridine nucleotide transhydrogenase and glutathione-dependent processes in *Rhodobacter sphaeroides***. *J Bacteriol* 2002, **184**(2):400-409.
32. Yang Z, Savchenko A, Yakunin A, Zhang R, Edwards A, Arrowsmith C, Tong L: **Aspartate dehydrogenase, a novel enzyme identified from structural and functional studies of TM1643**. *J Biol Chem* 2003, **278**(10):8804-8808.
33. Yoneda K, Kawakami R, Tagashira Y, Sakuraba H, Goda S, Ohshima T: **The first archaeal L-aspartate dehydrogenase from the hyperthermophile *Archaeoglobus fulgidus*: gene cloning and enzymological characterization**. *Biochim Biophys Acta* 2006, **1764**(6):1087-1093.
34. Fuhrer T, Fischer E, Sauer U: **Experimental identification and quantification of glucose metabolism in seven bacterial species**. *J Bacteriol* 2005, **187**(5):1581-1590.
35. Kiley PJ, Kaplan S: **Molecular genetics of photosynthetic membrane biosynthesis in *Rhodobacter sphaeroides***. *Microbiol Rev* 1988, **52**(1):50-69.
36. Schellenberger J, Park JO, Conrad TM, Palsson BO: **BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions**. *BMC Bioinformatics* 2010, **11**:213.
37. Ames GF: **Bacterial periplasmic transport systems: structure, mechanism, and evolution**. *Annu Rev Biochem* 1986, **55**:397-425.

38. Higgins CF: **ABC transporters: physiology, structure and mechanism--an overview**. *Res Microbiol* 2001, **152**(3-4):205-210.
39. Linton KJ, Higgins CF: **The *Escherichia coli* ATP-binding cassette (ABC) proteins**. *Mol Microbiol* 1998, **28**(1):5-13.
40. Rees DC, Johnson E, Lewinson O: **ABC transporters: the power to change**. *Nat Rev Mol Cell Biol* 2009, **10**(3):218-227.
41. Yoon SH, Han MJ, Jeong H, Lee CH, Xia XX, Lee DH, Shim JH, Lee SY, Oh TK, Kim JF: **Comparative multi-omics systems analysis of *Escherichia coli* strains B and K-12**. *Genome Biol* 2012, **13**(5):R37.
42. Barua D, Kim J, Reed JL: **An automated phenotype-driven approach (GeneForce) for refining metabolic and regulatory models**. *PLoS Comput Biol* 2010, **6**(10):e1000970.
43. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, Palsson BO: **A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism--2011**. *Mol Syst Biol* 2011, **7**:535.
44. Zomorrodi AR, Maranas CD: **Improving the iMM904 *S. cerevisiae* metabolic model using essentiality and synthetic lethality data**. *BMC Syst Biol* 2010, **4**:178.
45. Sistrom WR: **A requirement for sodium in the growth of *Rhodopseudomonas spheroides***. *J Gen Microbiol* 1960, **22**:778-785.
46. Borglin S, Joyner D, Jacobsen J, Mukhopadhyay A, Hazen TC: **Overcoming the anaerobic hurdle in phenotypic microarrays: generation and visualization of growth curve data for *Desulfovibrio vulgaris* Hildenborough**. *J Microbiol Methods* 2009, **76**(2):159-168.
47. Reasoner DJ, Geldreich EE: **A new medium for the enumeration and subculture of bacteria from potable water**. *Appl Environ Microbiol* 1985, **49**(1):1-7.
48. Schafer A, Tauch A, Jager W, Kalinowski J, Thierbach G, Puhler A: **Small mobilizable multi-purpose cloning vectors derived from the *Escherichia coli* plasmids pK18 and pK19: selection of defined deletions in the chromosome of *Corynebacterium glutamicum***. *Gene* 1994, **145**(1):69-73.
49. Tavano CL, Podevels AM, Donohue TJ: **Identification of genes required for recycling reducing power during photosynthetic growth**. *J Bacteriol* 2005, **187**(15):5249-5258.
50. Livak KJ, Schmittgen TD: **Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method**. *Methods* 2001, **25**(4):402-408.
51. Varma A, Palsson BO: **Metabolic flux balancing: basic concepts, scientific and practical use.** *Nature Biotechnology* 1994, **12**:994 - 998.
52. Mahadevan R, Schilling CH: **The effects of alternate optimal solutions in constraint-based genome-scale metabolic models**. *Metab Eng* 2003, **5**(4):264-276.
53. Lee S, Phalakornkule, C., Domach, M.M., and Grossmann, I.E.: **Recursive MILP model for finding all the alternate optima in LP models for metabolic networks**. *Computers & Chemical Engineering* 2000, **24**(2 - 7):711 -716.
54. Reed JL, Palsson BO: **Genome-scale in silico models of *E. coli* have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states**. *Genome Res* 2004, **14**(9):1797-1805.
55. Colijn C, Brandes A, Zucker J, Lun DS, Weiner B, Farhat MR, Cheng TY, Moody DB, Murray M, Galagan JE: **Interpreting expression data with metabolic flux models: predicting *Mycobacterium tuberculosis* mycolic acid production**. *PLoS Comput Biol* 2009, **5**(8):e1000489.
56. Kanehisa M, Goto S, Kawashima S, Nakaya A: **The KEGG databases at GenomeNet**. *Nucleic Acids Res* 2002, **30**(1):42-46.
57. Scheer M, Grote A, Chang A, Schomburg I, Munaretto C, Rother M, Sohngen C, Stelzer M, Thiele J, Schomburg D: **BRENDA, the enzyme information system in 2011**. *Nucleic Acids Res* 2011, **39**(Database issue):D670-676.
58. Schomburg I, Chang A, Schomburg D: **BRENDA, enzyme data and metabolic information**. *Nucleic Acids Res* 2002, **30**(1):47-49.

59. Alber BE: **Biotechnological potential of the ethylmalonyl-CoA pathway**. *Appl Microbiol Biotechnol* 2011, **89**(1):17-25.

Chapter 4

Quantifying the effects of light intensity on bioproduction and maintenance energy during photosynthesis

This chapter is formatted as a manuscript and has been submitted for publication:

Imam S, Fitzgerald CM, Cook EM, Donohue TJ and Noguera DR.

SI performed all computational modeling analysis. CMF and EMC set up *R. sphaeroides* bioreactors and obtained analytical data. SI, CMF and EMC participated in the determination of the *R. sphaeroides* biomass composition. SI, CMF and DRN analyzed data obtained from bioreactors.

Abstract

Obtaining a better understanding of the physiology and bioenergetics of photosynthetic microbes is an important step towards optimizing these systems for light energy capture or production of valuable commodities. In this work, we analyzed the effect of light intensity on bioproduction, biomass formation and maintenance energy during photoheterotrophic growth in the bacterium, *Rhodobacter sphaeroides*. Using data obtained from steady state bioreactors operated at varying dilution rates and light intensities, we found that irradiance had a significant impact on biomass yield and composition, with significant changes in photopigment, phospholipid and biopolymer storage contents. We also observed a linear relationship between incident light intensity and H_2 production rate between 3 and 10 W m^{-2}, with saturation observed at 100 W m^{-2}. The light conversion efficiency to H_2 was also higher at lower light intensities. Photosynthetic maintenance energy requirements were also significantly affected by light intensity, with links to differences in biomass composition and the need to maintain redox homeostasis. Inclusion of the measured condition-dependent biomass and maintenance energy parameters and the measured photon uptake rate into a genome-scale metabolic model for *R. sphaeroides* (iRsp1140) significantly improved its predictive performance. We discuss how our analyses provide new insights into the light-dependent changes in bioenergetic requirements and physiology during photosynthetic growth of *R. sphaeroides* and potentially other photosynthetic organisms.

Introduction

Microbes hold great potential as industrial systems for economical production of high value commodities through their vast array of metabolic processes [1, 2]. To optimize these biological systems for efficient bioproduction, it is imperative we gain better knowledge of their physiological and bioenergetic processes, so improvements can be made in culture conditions or strain design. Photosynthetic microbes, with their ability to harness light energy, are the major contributor to global carbon cycling and can be instrumental in development of bioprocesses for sustainable energy and bioremediation [1-3]. We are interested in gaining an improved systems-level understanding of the effect of light availability on cellular physiology and bioenergetics in order to take advantage of the activities of photosynthetic microbes.

Rhodobacter sphaeroides is perhaps the most well studied photosynthetic bacterium and has been used as a model system for investigating light energy capture and conversion, as well as for studying various other aspects of this lifestyle [4, 5]. In addition to its ability to grow under a variety of conditions (aerobic respiratory, anaerobic respiratory and anoxygenic photosynthetic) and utilize a diverse array of growth substrates [5, 6], *R. sphaeroides* is also capable of producing significant quantities of a suite of valuable commodities including H_2 [7-9], polyhydroxybutyrate (PHB) [10] and ubiquinone [11], as well as being able to fix CO_2 and N_2 [3, 12-14]. While several aspects of *R. sphaeroides* metabolic processes have been studied in detail, no detailed systems-level analysis of light absorption and its effects on bioenergetics and physiology has been conducted.

In recent years, constraint-based modeling has proved to be a very useful tool for the systems-level analysis of cellular metabolism, yielding new insights into condition-dependent metabolic flux distributions and cellular physiology. Furthermore, constraint-based analysis can guide metabolic engineering efforts to optimize the production of cell biomass and metabolites of interest, amongst other applications [15]. Previous work led to the construction and refinement of high-quality genome-scale metabolic models for *R. sphaeroides*, iRsp1095 [4] and iRsp1140 [6], which have been used for analysis

of its metabolic processes [4, 6, 16]. However, previous application of these and many other models of photosynthetic organisms have, with a few exceptions [17, 18], not considered measured light uptake as an important modeling constraint, generally relying on arbitrary thresholds and estimates [19-22].

An important factor in determining the efficiency of a cell for bioprocessing is the amount of metabolic energy (ATP) utilized for biomass production and maintenance of cellular homeostasis [23, 24]. In general, an ideal microbial cell factory would have low maintenance energy demands, allowing for the direction of more of its metabolic energy towards production of commodities of interest [23, 25]. There are several studies of how changes in light intensity impact growth of *R. sphaeroides* and other phototrophic microbes in batch or continuous culture [26-31]. However, the steady state maintenance energy requirements for *R. sphaeroides* at different light intensities have not been determined. Knowledge of this parameter can be crucial for the accurate prediction of biomass and metabolite production using constraint-based models. Thus, an in depth understanding of the maintenance energy requirements and bioenergetic challenges that are unique to the photosynthetic lifestyle will be important for downstream modeling and strain design efforts.

In this study, we assessed the impact of light intensity on biomass formation, biomass composition, biomolecule production and cellular maintenance energy requirements under photoheterotrophic growth conditions in *R. sphaeroides*. Using a series of chemostats run at varying dilution rates and irradiated at low, moderate and high light intensities, we observed that biomass composition, as well as production of several target biomolecules (lipid, PHB and H_2) changed with variation in light intensity. We also found that the photosynthetic maintenance energy requirements were also significantly affected by changes in light intensity. The *R. sphaeroides* metabolic model, iRsp1140, updated with the experimentally determined biomass and maintenance energy parameters, was used to study the impact of light intensity on growth and metabolic flux distributions. The incorporation of biomass and maintenance energy parameters significantly improved the predictive capability of iRsp1140. Overall, these analyses provide

new insights into the light-dependent changes in bioenergetic requirements and cellular behavior during photosynthetic growth of *R. sphaeroides*.

Results and Discussion

Chemostat measurements

To assess the impact of light intensity on *R. sphaeroides* physiology and bioenergetics, we set up bioreactors irradiated at low (3 and 5 W m^{-2}), moderate (10 W m^{-2}) and high (100 W m^{-2}) light intensities (corresponding to ~20, 33.5, 66.9 and 669 μmol photons m^{-2} s^{-1} respectively, at 800 nm). At each light intensity, multiple reactors were run at retention times ranging between 8 and 39 hours (i.e., dilution rates between 0.026 h^{-1} and 0.125 h^{-1}). Cells were harvested from reactors at steady state and the biomass and PHB composition determined. The uptake rates for the components in the media with highest concentration, succinate, glutamate, aspartate and sulfate were empirically determined, while H_2 production was measured at steady state prior to harvesting cells. The specific light supply rate (SLSR) in mmol photons gDW^{-1} h^{-1} was also calculated from the measured biomass and light intensities (Table S1).

Biomass yield on light energy

To determine how varying light intensities affected growth yields, we assessed the relationship between light supply and dilution rate. From a comparison of SLSR and the dilution rate (*D*) (i.e., the specific growth rate), we observed a linear relationship for reactors illuminated at 3, 5 and 10 W m^{-2} with R^2 of 0.9, 0.95 and 0.8 (p-values of 0.0016, 0.024 and 0.0018), respectively (Figure 4-1a), indicating that the supplied light has yet to reach saturation and is a limiting factor on steady state growth at these intensities and reactor configurations. However, there was no correlation between the SLSR and *D* at 100 W m^{-2} (R^2 = 0.007) (Figure S1a), indicating that light is saturating at this intensity and suggesting that the majority of the supplied light was not being used to support growth.

The biomass yields on light energy ($Y_{x,E}$) were calculated from the inverse of the slopes in Figure 4-1a [32]) to be 0.0016, 0.0015 and 0.0012 gDW mmol photons^{-1} at 3, 5 and 10 W m^{-2} respectively. These represent reasonably high biomass yields that compare favorably to those reported for other phototrophs [18, 32], though the growth mode used here is photoheterotrophic as opposed to photoautotrophic in the other studies. The lower yield at 10 W m^{-2} could reflect the nitrogen source becoming limiting at this light intensity, which was not the case at the lower intensities (Table S1). In addition, at these sub-saturating light intensities, the light-derived maintenance energy requirements (i.e., the y intercepts of the regression plots in Figure 4-1a [32]) were 5.4, 0.83 and 16.7 mmol photons gDW^{-1} h^{-1} at 3, 5 and 10 W m^{-2}, respectively.

We also observed a high correlation between the measured cell dry weight and *D* for reactors run at sub-saturating light intensities, with R^2 values ranging from 0.8 to 0.92 (Figure 4-1b). The observed negative correlation between cell mass and *D* is consistent with cells growing under limiting conditions, as higher dilution rates result in less dense cultures. In these cultures, organic substrates in the supernatant were abundant (Table S1), indicating that light intensity was the main limiting factor under these conditions. At the saturating light intensity of 100 W m^{-2}, we did not observe any meaningful correlation between biomass and *D* (Figure S1b), an observation that suggests light is not limiting under these growth conditions.

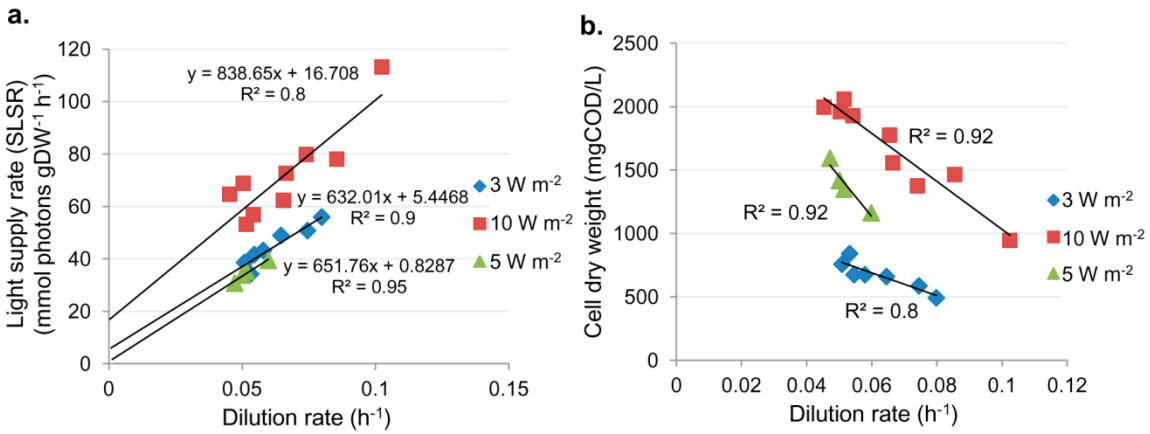

Figure 4-1. Relationship of dilution rate to SLSR and biomass production Regression of (a) specific light supply rate (SLSR) against dilution rate *D* and (b) cell dry weight against *D*, at 3 (♦), 5 (▲) and 10 (■) W m^{-2}. There was no statistically significant correlation between these measurements at 100 W m^{-2} (Figure S1).

Biomass yield on nutrients

In contrast to the relationships between SLSR and *D*, the correlation between nutrient uptake and *D* was poor at sub-saturating light intensities (Figure S2). There was very little correlation between uptake of succinate (the major carbon source) or sulfate and *D* at these light intensities, indicating that neither of these were limiting nutrients. However, glutamate (the main nitrogen source) uptake showed a better correlation to *D* ($R^2 > 0.6$) at these light intensities (Figure S2). On the other hand, there were high correlations between uptake of all nutrients measured and *D* at an incident light intensity of 100 W m^{-2} ($R^2 > 0.93$) (Figure S2). The observed biomass yield from carbon uptake ($Y_{x,C}$), obtained from the regression of cell dry weight on the overall organic substrate uptake, produced good fits ($R^2 > 0.88$) at 5, 10 and 100 W m^{-2} with $Y_{x,C}$ equal to 0.64, 0.61 and 0.57 $gCOD_x$ $gCOD_c^{-1}$ respectively (Figure S3). However, the fit was much poorer at 3 W m^{-2}, since organic substrates were not limiting. Overall, these observations are consistent with a model in which incident light intensity is the main constraint on growth at sub-saturating light intensities of 10 W m^{-2} and below. Under saturating light conditions (100 W m^{-2}), the major limitation to growth becomes the rate of organic substrate utilization, hence the strong correlation between nutrient uptake and biomass observed under these conditions.

Steady-state biomass composition is significantly altered by changing light intensities

Previous analysis of batch cultures has shown that light intensity can significantly impact aspects of *R. sphaeroides* physiology and biomass composition. In particular, the amount of light capturing pigments produced by *R. sphaeroides* increases significantly with decreases in incident light intensities [33, 34]. We assessed the impact of light intensity on *R. sphaeroides* biomass components at steady state.

Incident light intensity had only a small effect on bulk cellular protein content but it was slightly higher at low light intensities (3 and 5 W m^{-2}) making up about 71 to 73% of total cell dry weight (DW), compared to moderate (10 W m^{-2}) and high light intensities (100 W m^{-2}), where protein made up about 60 to 64% of the DW (Figure 4-2a). On the other hand, light intensity had a significant impact on PHB production in *R.*

sphaeroides. At low light intensities (3 and 5 W m^{-2}), PHB made up only ~3% of the DW, which was more than 4-fold less than that observed at moderate (10 W m^{-2}) and high (100 W m^{-2}) light intensities, where PHB accounted for 13 to 14.8% of the DW respectively (Figure 4-2b). Consistent with previous observations in batch cultures, the total bacteriochlorophyll (Bchl) content of the cell was highly dependent on incident light intensity. The cellular Bchl content was highest at low light intensities, constituting about 0.71% and 0.91% of DW at 3 and 5 W m^{-2}, respectively. In contrast a significant drop in Bchl content was observed at moderate (0.53%) and high (0.17%) light intensities, representing about a 5 fold difference between high and low light conditions (Figure 4-2c). The observations are consistent with the fact that at lower light intensities more photopigments are required to absorb the limited amount of incident light.

During photosynthetic growth, *R. sphaeroides* significantly increases its membrane surface area by forming intracytoplasmic membranes to house it photosynthetic apparatus [35]. Phospholipids are a major constituent of *R. sphaeroides* cell membranes [36, 37], and are also expected to change as a function of light intensity. However, while changes in light intensity resulted in significant changes in Bchl content, the relative magnitude of the change in cellular phospholipid levels with respect to light intensity was smaller. Similar to Bchl content, the highest amount of phospholipids was obtained for cells grown at 5 W m^{-2} (~6% of DW), while the lowest was observed for cell grown at 100 W m^{-2} (~4% of DW), only a 1.5 fold difference (Figure 4-2d). In addition, phospholipid content was higher in the 10 W m^{-2} cultures than 3 W m^{-2} in contrast to changes in Bchl level, indicating that the impact of light intensity on phospholipid formation is small and not tied to that of photopigment production.

Overall, we did not observe any significant changes in biomass composition with respect to chemostat dilution rate at any of the light intensities tested.

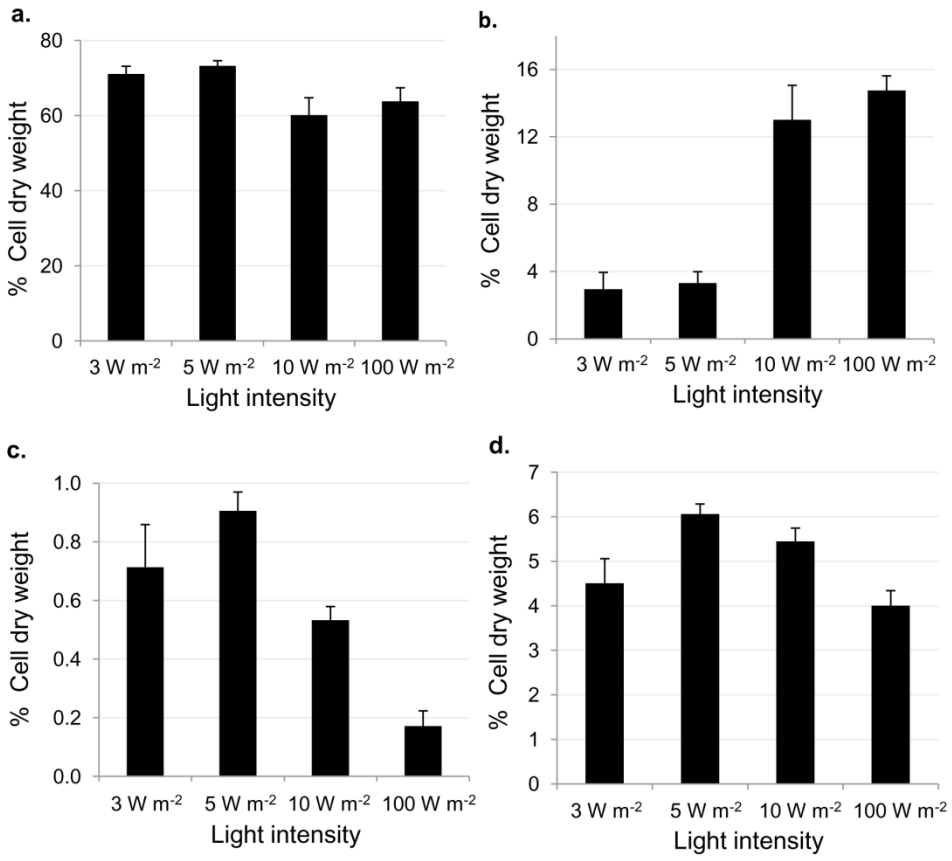

Figure 4-2. Biomass composition is significantly affected by light intensity Assessment of the contribution of (a) Protein (b) PHB (c) Bacteriochlorophyll and (d) Phospholipids to cell dry weight at 3, 5, 10 and 100 W m^{-2} light intensities.

Assessing the effect of light intensity on R. sphaeroides metabolite secretion

In addition to determining the effect of light intensity on *R. sphaeroides* biomass composition, we also analyzed its effects on selected metabolic end products at steady state.

H_2 production is dependent on light intensity

We assessed the impact of light intensity on H_2 production and observed a strong linear relationship between intensity of light supplied and the amount of H_2 produced, with increasing amounts of H_2 produced as light intensity increased. At low (3 to 5 W m^{-2}) to moderate (10 W m^{-2}) light intensities, the increase in H_2 production appeared to be directly proportional to the increase in light intensity, resulting in an increase of ~0.25 mmol gDW^{-1} h^{-1} of H_2 production (~0.1 mL h^{-1} with the experimental conditions used) ($R^2 > 0.99$) per unit of irradiance increased (Figure 4-3a). This observation indicates that light is a limiting factor for H_2 production, consistent with previous observations in batch culture [38], and it predicts that cells might be using a proportional amount of the incident light for H_2 production within this range. While there was significant increase in the amount of H_2 produced when comparing moderate (10 W m^{-2}) to saturating or high (100 W m^{-2}) light, an ~1.7 fold increase, the magnitude of this increase was not proportional to the amount of additional light, providing another indication that the incident light is saturating at steady state in the later condition (Figure 4-3b), and suggesting that H_2 production reached its maximum rate (an average of ~4.9 mmol gDW^{-1} h^{-1} (or ~1.28 mL h^{-1})).

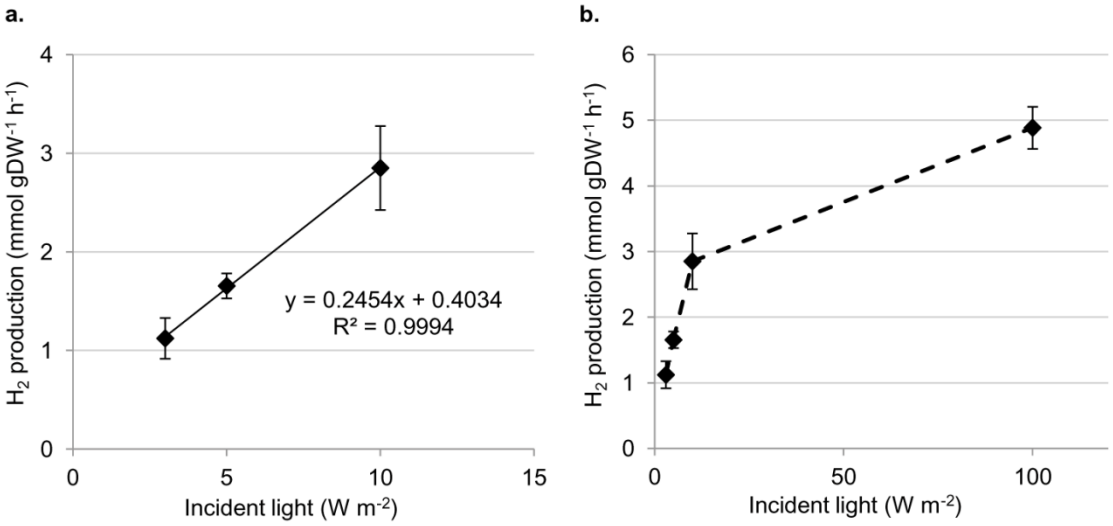

Figure 4-3. Relationship of H_2 production rate to light intensity (a) Regression of average H_2 production rate against light intensity between 3 and 10 W m^{-2}. (b) Plot of average H_2 production against light intensity extended to 100 W m^{-2}, showing loss of linear relationship due to light saturation.

The efficiency of light energy conversion to an end product like H_2 is a useful parameter for assessing the effectiveness of organisms for H_2 production [38, 39]. We found that the light energy conversion efficiency of these steady state cultures was relatively high at low and moderate light intensities with average efficiencies of 4.75, 6.35 and 5.77 % at 3, 5 and 10W m^{-2} respectively. The energy conversion efficiencies obtained at these light intensities are in general agreement with previously reported values for cells grown in batch cultures [38, 39]. On the other hand, at 100 W m^{-2} the light conversion efficiency dropped down to just 0.77%, consistent with light being provided in excess at this intensity. Similar decreases in energy conversion efficiency at high light intensities have previously been observed [38]. In addition, to the light conversion efficiency, another important parameter is the substrate conversion efficiency (i.e., ratio of observed H_2 produced to the theoretical maximum obtainable for hydrogen atoms in the main growth substrate, succinate in our chemostats) [39, 40]. At 3W m^{-2} the substrate conversion efficiency was ~21%, while the efficiencies significantly increased at higher light intensities – 40 %, 34 % and 34 % at 5, 10 and 100 W m^{-2} respectively. Pathways that compete for the reducing equivalents provided by succinate include PHB production, biomass growth, and the production of soluble microbial products (SMPs) [9].

Light intensity alters the end product profiles in continuous culture

Previous analysis of *R. sphaeroides* growing in batch cultures indicated the production of substantial amounts of SMPs, such as TCA cycle intermediates, lactate and pyruvate [9]. Thus, to determine if significant amounts of SMPs were produced during steady state growth, and to see if this was affected by light intensity, we compared the measured chemical oxygen demand (COD) of the spent media to its calculated COD (based on measured extracellular concentrations of succinate, glutamate and aspartate; see Material and Methods). For chemostats operated at low light intensities (3 W m^{-2} and 5 W m^{-2}), the measured and estimated COD were very similar, with the difference generally less than 10%. Conversely, at moderate to high light intensities (10 W m^{-2} and 100 W m^{-2}), the difference in measured and estimated COD was ~20%. We analyzed the spent reactor media for soluble organics (glucose, xylose, pyruvate,

xylitol, lactate, glycerol, formate, acetate, malate and ethanol) at varying light intensities in order to potentially identify SMPs. In general, only small amounts of pyruvate, lactate, glycerol, xylitol and xylose were observed in the spent media but the levels of these metabolites did not show any trends across the reactors. Thus, the chemical identity of most of the SMPs in steady state cultures remains unresolved. The increased spent media COD observed at higher light intensities could reflect altered metabolic flexibility in these cultures, where conserving metabolic energy by nutrient oxidation is less of a constraint on the cells due to the increased availability of light energy. Alternatively, it could be an indication of increased cellular stress or photo-damage at higher light intensities, with the increased extracellular COD coming from secretion of cellular constituents into the media.

Maintenance energy requirements change with light intensity

The energy a cell expends for growth and maintaining cellular homeostasis is an important factor in determining its performance in the laboratory or in industrial settings [23]. Thus, we assessed the maintenance energy requirements of *R. sphaeroides* under photosynthetic conditions at varying light intensities to gain a better understanding of its bioenergetic requirements and limitations. In genome-scale metabolic models the maintenance energy required by a cell is divided into two major components: the growth associated maintenance (GAM) and non-growth associated maintenance (NGAM) energy requirements, used for biomass formation and maintenance of cellular homeostasis respectively [41]. GAM and NGAM are typically estimated from the slope and y-intercept respectively, of the regression of the predicted maximum ATP hydrolysis/production on the experimentally determined specific growth rates from steady state cultures [41]. In our simulations, GAM represents the ATP required for polymerization and formation of biomass components, while NGAM represents the ATP used by the cell for processes that are not directly linked to growth.

Efficiency of light utilization

For photosynthetic organisms, the total amount of ATP that can be predicted to be produced will depend on the efficiency with which the incident light is utilized by the cells. This efficiency, which is dependent on the light absorbing capacity of the photosynthetic apparatus, can also be significantly impacted by cell shading in dense cultures, back scatter of light by cells and the reactor configuration (depth, etc). Thus, prior to determining maintenance energy requirements at various light intensities, we used iRsp1140 to computationally assess the impact of the efficiency of light utilization (Φ) on maintenance energy estimates.

Constraining iRsp1140 using the measured uptake rates for succinate, glutamate, aspartate and sulfate, as well as the observed growth rates at steady state, we predicted maximum ATP hydrolysis rates at Φs of 10, 25, 50, 75 and 100%, at incident light intensities of 3, 5 and 10 W m^{-2}, and compared this to the observed growth rate. From these analyses, we observed that the best fit between the maximum ATP hydrolysis rate and growth rates was achieved at a Φ of 75% for all sub-saturating light intensities (Figure 4-4a), with R^2 values of 0.86, 0.93 and 0.74 at 3, 5 and 10 W m^{-2} respectively. Light utilization efficiencies of 10 and 25, resulted in significantly poorer fits between the predicted ATP hydrolysis rates and μ, while 75% efficiency provided marginally better fits than 50% and 100% (Figure 4-4a). Based on this, we estimated the Φ of *R. sphaeroides* to be ~75%, so this value was used for all subsequent maintenance energy estimates. Interestingly, this value is similar to the maximum light utilization efficiency of 80% estimated for algal cells [18].

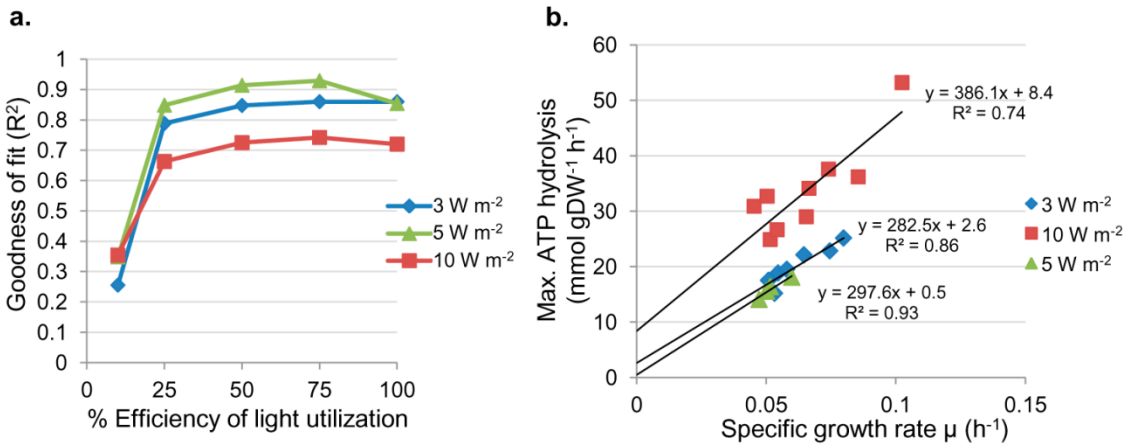

Figure 4-4. Assessing photosynthetic maintenance energy in *R. sphaeroides* (a) Comparison of the goodness of fit (i.e., R^2 values) of the regression of maximum ATP hydrolysis on specific growth rate μ considering a range of efficiencies of light utilization (i.e., 10, 25, 50, 75 and 100 % of measured light intensity) in each simulation. The analysis was conducted with data obtained from chemostats operated at 3 (♦), 5 (▲) and 10 (■) W m^{-2} light intensity. The best fit obtained across all data was at 75% efficiency. (b) Regression of maximum ATP hydrolysis against specific growth rate μ at 3 (♦), 5 (▲) and 10 (■) W m^{-2} using a Φ of 75%.

Maintenance energy requirements vary with light intensity

At light intensities of 3 and 5 W m^{-2}, we estimated the GAM to be 282.5 and 297.6 mmol ATP gDW^{-1} respectively, while NGAM was estimated to be 2.6 and 0.5 mmol ATP gDW^{-1} h^{-1} respectively (Figure 4-4b). Overall, there did not appear to be significant differences between these parameters at these two light intensities, which might be expected given the small difference in light intensities and the general similarities in the biomass compositions of steady state cells under these conditions (Figure 4-2). Given the larger number of data points and wider range of growth rates over which data was collected for cultures grown at 3 W m^{-2}, these data possibly provide a better estimate of GAM and NGAM at lower light intensity.

At a light intensity of 10 W m^{-2}, the calculated GAM was 386.1 mmol ATP gDW^{-1}, which is significantly higher than that obtained at lower light intensities, likely reflecting the significant differences in biomass composition between cells grown under these conditions (Figure 4-4b). Furthermore, the estimated NGAM was 8.4 mmol ATP gDW^{-1} h^{-1}, significantly higher than that observed at the two lower light intensities. This larger NGAM value could be the result of increased stress (heat, photochemical or others) at higher light intensities. Alternatively, this difference could reflect the increased ATP required to maintain homeostasis, for instance via ATP-dependent H_2 production by nitrogenase [8], which is significantly higher at 10 W m^{-2}. Incorporating ATP-dependent H_2 production constraints into iRsp1140 with the calculated GAM and NGAM parameters, showed that requiring additional ATP for this process significantly deteriorates the model's predictions as light (and thus ATP) is limiting under these conditions (Figure S4). This suggests that ATP utilized for H_2 production is already accounted for in the estimated maintenance energy parameters.

Interestingly, the iRsp1140 derived maintenance energy values are very similar to the previously estimated light derived maintenance energy values obtained from regressing SLSR on *D* (Figure 4-1a). Given that 2 photons leads to generation of ~1 ATP during cyclic photophosphorylation [42], the GAM

(NGAM) from that analysis is equivalent to 316 (2.72), 325.9 (0.41) and 419.3 (8.4) mmol ATP g DW^{-1} (mmol ATP g DW^{-1} h^{-1}) at 3, 5 and 10 W m^{-2} respectively, very similar to that predicted by iRsp1140. The small differences between these values likely result from aspects of ATP utilization accounted for elsewhere in iRsp1140 (such as charging of tRNAs with amino acids), as well as the inclusion of nutrient derived ATP in the model's predictions and the rotational fold symmetry of *R. sphaeroides* ATP synthase under these conditions. Nevertheless, this indicates that the vast majority of energy used for growth and maintenance of cellular homeostasis during photoheterotrophic growth is derived from photon absorption, leaving energy in organic substrates to be utilized for bioproduct formation.

We could not use similar methods to accurately calculate maintenance energy requirements at 100 W m^{-2} as the plot of maximum ATP hydrolysis versus growth rate obtained under these conditions did not show any significant correlation (Figure S5). This observation might reflect the fact that light is saturating at this intensity so only a small fraction of incident light is being used by cells under these conditions. However, utilizing more conservative estimates of light absorption efficiency did not improve these results (Figure S5). An alternative explanation might lie in increasing cellular damage from light, heat, photochemical stress and other factors at high light intensities, which would complicate attempts to estimate maintenance energy under these conditions.

Use of light uptake rates and maintenance energy to improve modeling predictions

The above data indicated that, at low and moderate light intensities, photons are a significant growth limiting factor, while this is not the case at high light intensities. Consistent with those observations, iRsp1140 predicts that at low to moderate light intensities (3 to 10 W m^{-2}), the measured SLSR and the specific light uptake rate (LUR) required to achieve optimal growth (optimal LUR), given the substrate uptake rates, are comparable (Figure 4-5a). In most cases, the measured SLSR is less than the predicted optimal light uptake rate, consistent with photons being a growth limiting factor. Conversely, the

measured SLSR was about 10 times the predicted optimal LUR at 100 W m^{-2}, consistent with photons being supplied in excess under these conditions (Figure 4-5a).

The above analysis also suggests that LUR can have a significant impact on the observed and predicted behavior of photosynthetic cells and, if ignored or inaccurately measured, could have a negative effect on the validity of computational predictions and vice versa. Indeed, plotting the predicted optimal growth rates against observed specific growth rates (i.e., dilution rate), when LUR and all measured uptake and production rates are taken into account (Figure 4-5b), and comparing this to a plot of predictions generated by omitting LUR (thus predicting optimal growth based on substrate uptake and assuming that light is not limiting), showed that inclusion of the LUR significantly improves the predictions of iRsp1140 (Figure 4-5c). The fit between predicted and observed growth rates improves from an R^2 of 0.63 ($p = 2.9 \times 10^{-5}$) to 0.85 ($p = 7.2 \times 10^{-9}$) with the inclusion of the measured LUR. Only cultures grown at low to moderate light intensities (3 to 10 W m^{-2}) were considered for this analysis since at saturating light (100 W m^{-2}) the SLSR is not an adequate estimation of specific LUR. It should also be noted that inclusion of the experimentally determined parameters for GAM and NGAM, also significantly improved the predictions of iRsp1140 compared to using the previously estimated maintenance energy requirements [4] (Figure 4-5d).

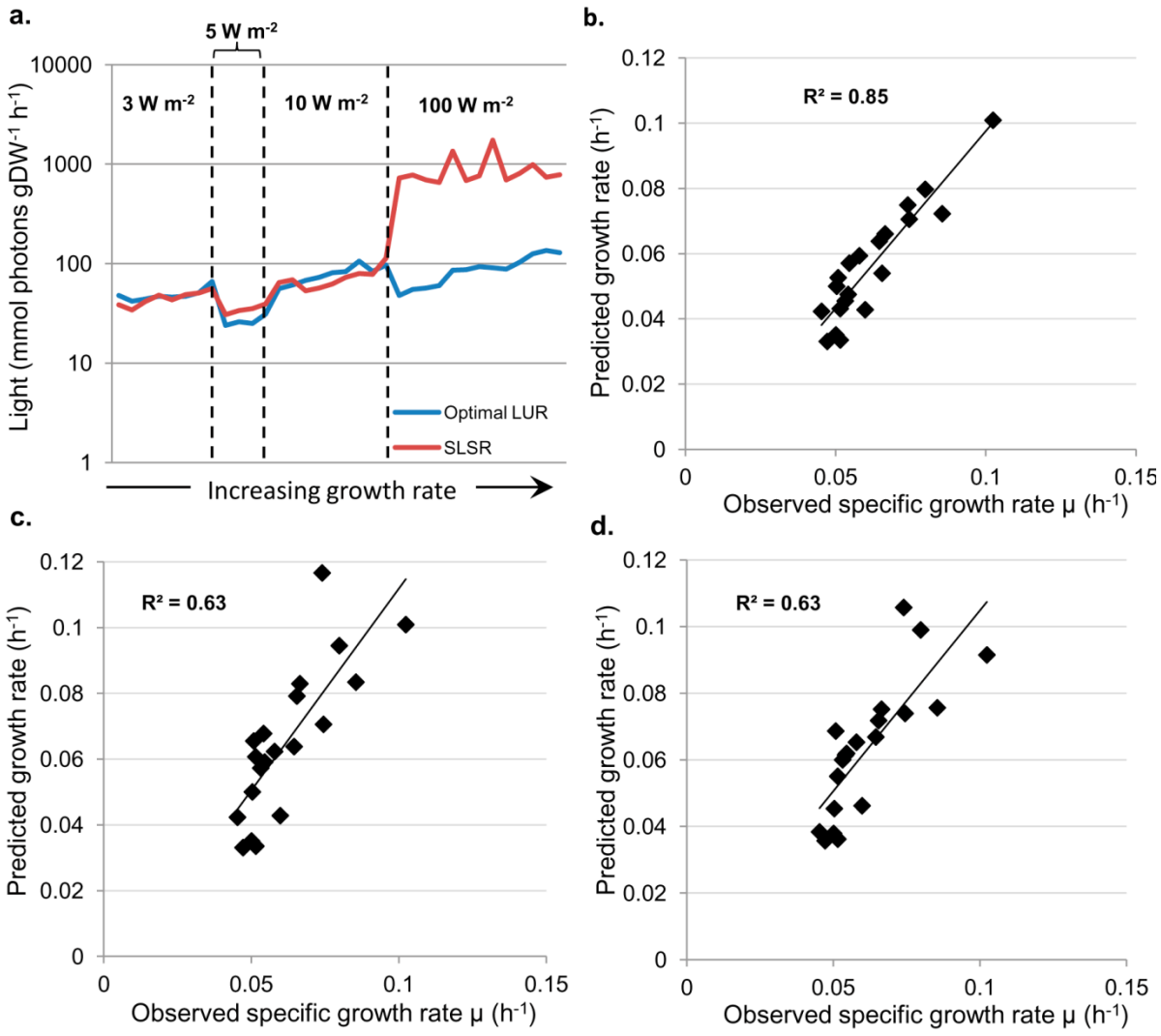

Figure 4-5. Assessing the impact of SLSR, LUR and maintenance energy on modeling predictions (a) Plot comparing the measured SLSR and the specific LUR predicted by iRsp1140 to be required for optimal growth via FBA (optimal LUR), given the measured substrate uptake rates. Comparison of the observed and iRsp1140 predicted specific growth rates using all measured parameters including LUR (b) and excluding LUR (c) as a constraint for FBA predictions. (d) Comparison of iRsp1140 growth rate predictions to the observed growth rates using previously estimated maintenance energy parameters [4].

The impact of light limitation on the solution space of metabolic fluxes

While constraint-based analysis can be a very useful tool for studying metabolism on a genome-scale, one of its limitations lies in the large solution space of feasible flux distributions that can result in the predicted optimal objective function [43]. However, with the application of additional relevant constraints, one can significantly reduce the size of this solution space. We observed that LUR, in particular, has a significant impact on the solution space of iRsp1140 predictions for this photosynthetic bacterium. Using flux variability analysis (FVA) [43], we assessed the range of predicted flux values for each reaction at optimal growth rate under light limiting and saturating conditions (i.e., with and without use of the specific LUR as an additional constraint, see above). Using all available nutrient uptake data from a representative chemostat operated at a light intensity of 10 W m^{-2}, a total of 795 reactions were able to carry flux under light saturating conditions, compared to just 448 when the LUR constraint were imposed. Of these 448 reactions, which carried flux under both conditions, at total of 342 reactions (76.3%) showed no variability in flux under light limiting conditions, compared to only 123 reactions (27.5%) with invariable fluxes under light saturating conditions. These results indicate that there is a significant reduction in the solution space when the LUR constraint is included. Examination of the fluxes through central metabolism illustrates this difference in flux distribution (Figure 4-6), as the fluxes through most TCA cycle reactions show no change over the optimal solution space under light constrained conditions, whereas significant variability in these fluxes is predicted with unconstrained light. In addition, the range of values predicted for variable reactions was also significantly larger without including LUR as a modeling constraint. This reduced solution space likely reflects the energy limitations under light limiting conditions. The implications of this is that flux through energy demanding reactions will be avoided or minimized by the cells where possible when light is limiting.

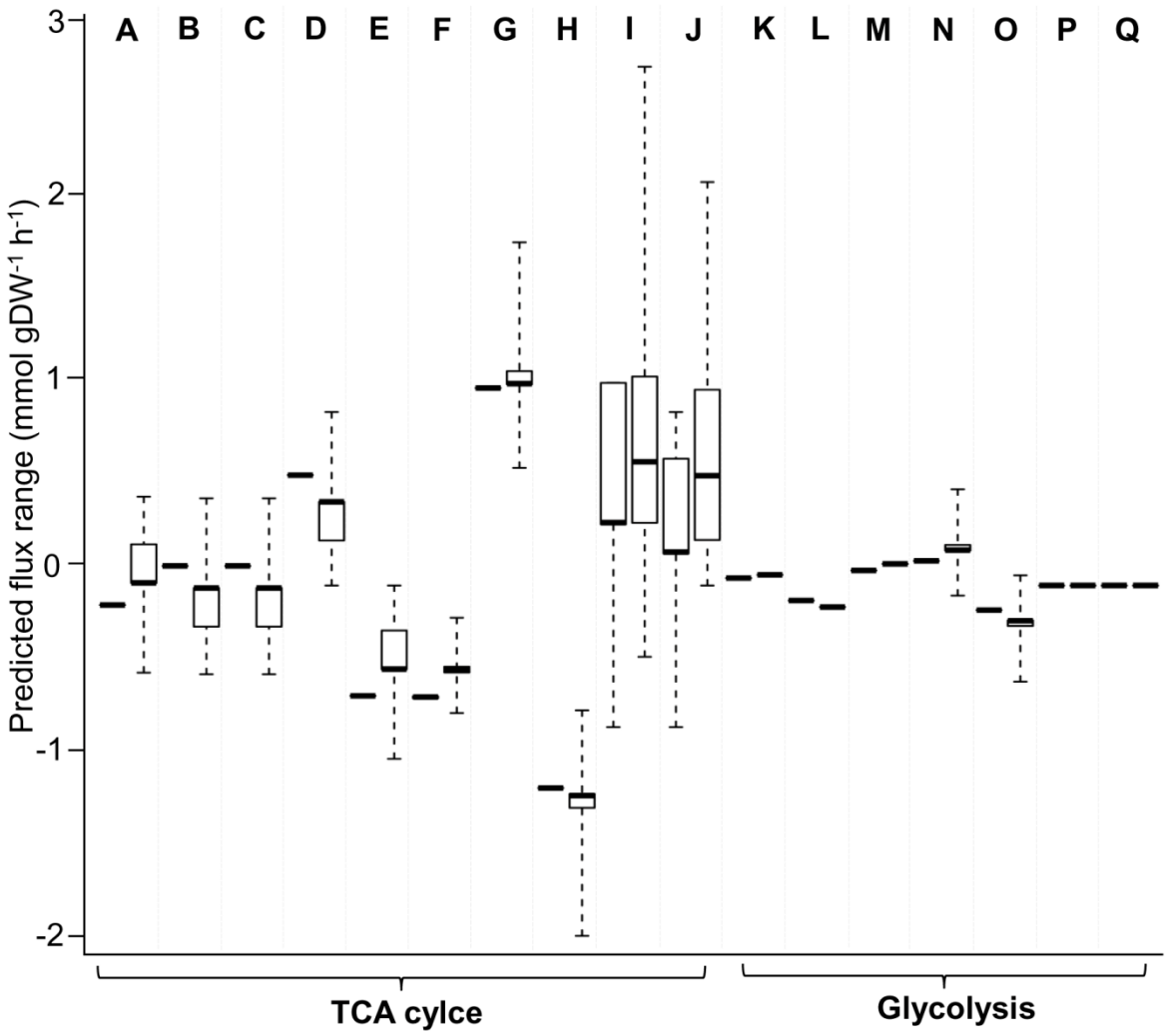

Figure 4-6. Variability of predicted flux distributions under light limiting and saturating conditions Box plots of flux distribution for reactions of the TCA cycle (A-J) and glycolytic pathway (K-Q) in *R. sphaeroides* under photoheterotrophic conditions. The first box plot in each group represents the flux distribution under light constrained conditions, while the second box represents flux distributions with unconstrained light. Distributions were obtained by sampling 1000 alternative optimal solutions. It should be noted that outliers were omitted from the box plots for the unconstrained light analysis to allow better visualization of the differences in predicted flux ranges. Data used in the simulations corresponded to a chemostat run at 10W m^{-2}. In both sets of simulations, the model was constrained with the uptake rates for succinate, glutamate, aspartate, sulfate, as well as production rates of PHB and H_2. For the light-constrained simulation, LUR was set to the measured SLSR. A – citrate synthase; B – aconitase; C – isocitrate dehydrogenase; D and E – α-ketoglutarate dehydrogenase; F – Succinyl-CoA synthetase; G – Succinate dehydrogenase; H – Fumarate reductase; I – Malate dehydrogenase; J - phosphoenolpyruvate carboxykinase; K - triosephosphate isomerase; L – glyceraldehydes-3-phosphate dehydrogenase; M – phosphoglycerate kinase; N – phosphoglyceromutase; O – phosphopyruvate hydratase; P – fructose-bisphosphate aldolase; Q - D-fructose 1,6-bisphosphatase.

Conclusions

Photosynthetic organisms play a key role in global carbon cycling and the planetary food chain. They have the potential to make significant biotechnological contributions, as their ability to efficiently harness solar energy and produce value-added commodities can help develop a more sustainable society. Well-studied photosynthetic microbes such as *R. sphaeroides* can provide high value products and guide future research given its well-studied photochemical cycle and the available genetic, genomic and computational tools available to study this organism. To take advantage of photosynthetic systems, it is imperative to gain an improved understanding of their physiology, energetics and contribution of light intensity changes in these processes.

In this study, we gained a better quantitative understanding of the impact of light on a photosynthetic organism by examining its effects on the physiology and bioenergetics of *R. sphaeroides*. We found significant light-dependent changes in biomass composition as well as PHB and H_2 production in steady state cells. Furthermore, our analysis showed that the metabolic energy required for cellular maintenance under photosynthetic conditions is significantly affected by light intensity, but it is also linked to resultant differences in biomass composition and the need to maintain redox homeostasis. At low light intensities, *R. sphaeroides* maintenance energy requirements were relatively low, suggesting that it could serve as an excellent microbial factory for bioproduction under these conditions. Maintenance energy increased significantly at higher light intensities, but with a tradeoff of producing significantly larger amounts of potentially useful end products like PHB and H_2.

We also showed that use of the condition-dependent parameters for maintenance energy we derived can significantly improve predictions of a genome-scale metabolic model of photosynthetic metabolism in *R. sphaeroides*. By incorporating photon uptake for the first time in modeling photosynthetic growth in *R. sphaeroides*, we also significantly improved the predictions of an existing constraint-based metabolic model [6]. Furthermore, we showed that the solution space of modeling predictions is significantly

reduced when light limiting conditions are considered, with only a handful of reactions in the network having variable fluxes under these conditions. Overall, these analyses have improved our understanding of the effects of light on *R. sphaeroides* photosynthetic metabolism, its bioenergetics and physiology. In addition, it has provided several new parameters (Table 4-1) that can be used to improve the modeling and performance of this and related photosynthetic organisms for industrial uses.

Table 4-1. Properties of *R. sphaeroides* during photoheterotrophic growth

	3 W m^{-2}	**5 W m^{-2}**	**10 W m^{-2}**	**100 W m^{-2}**
Substrates				
Carbon	Succinate	Succinate	Succinate	Succinate
Nitrogen	Glutamate	Glutamate	Glutamate	Glutamate
Yields				
Biomass yield on light, $Y_{x,E}$ (g DWmmol photons^{-1})	0.0016 ± 0.0002	0.0015 ± 0.0002	0.0012 ± 0.0002	ND
Biomass yield on substrate, $Y_{x,C}$ ($gCOD_x$ $gCOD_c^{-1}$)	ND	0.64 ± 0.17	0.61 ± 0.07	0.57 ± 0.04
Maintenance energy				
Light maintenance (mmol photons g DW^{-1} h^{-1})	5.4 ± 6.4	0.8 ± 5.4	16.7 ± 11.7	ND
GAM (mmol ATP gDW^{-1})	282 ± 51	298 ± 58	386 ± 86	ND
NGAM (mmol ATP gDW^{-1} h^{-1})	2.6 ± 3.2	0.5 ± 3	8.4 ± 5.9	ND
Hydrogen				
Average H_2 production (mmol g DW^{-1} h^{-1})	1.1 ± 0.2	1.7 ± 0.1	2.9 ± 0.4	4.9 ± 0.3
Substrate conversion efficiency (%)	21 ± 5	40 ± 7	34 ± 3	34 ± 4
Light conversion efficiency (%)	4.8 ± 0.7	6.4 ± 0.4	5.8 ± 0.2	0.8 ± 0.05
Biomass (% DW)				
Protein	71.1 ± 2.0	73.3 ± 1.3	60.1 ± 4.6	63.8 ± 3.6
PHB	3.0 ± 1.0	3.3 ± 0.7	13.01 ± 2.0	14.8 ± 0.9
Pigment	0.7 ± 0.2	0.9 ± 0.06	0.5 ± 0.05	0.2 ± 0.05
Lipid	4.5 ± 0.6	6.1 ± 0.2	5.5 ± 0.3	4.0 ± 0.3

Materials and methods

Chemostat set up

R. sphaeroides 2.4.1 cells were grown on Sistrom's Minimal Medium (SMM) [44] with 8.1 mM glutamate used to replace ammonia as the main nitrogen source. Under these conditions, H_2 production via nitrogenase activity is promoted [8, 9]. A total volume of 16.5 mL of cells was cultured in continuously stirred, sealed 20 mL chemostats at 30°C, which were illuminated by an incandescent light box at 3, 5 or 10 W m^{-2}, or with an incandescent flood light source at 100 W m^{-2}. Reported light intensities were measured with a Yellow-Springs-Kettering model 6·5-A radiometer through a Corning 7-69 filter (transmission 750 – 900 nm, i.e., the photosynthetically active radiation (PAR) for *R. sphaeroides*). Culture turbidity was monitored using a Klett-Summerson photoelectric colorimeter (Klett MFG Co., NY).

The chemostats were operated in batch-fed mode, in which a portion of spent media was wasted for 1 minute immediately followed by a 3 minute feeding cycle of an equivalent volume of new media. This cycle was repeated every 20 minutes. The feeding and wasting rates were varied to attain a range of retention times varying from 8 to 39 hours. Reactors were started in batch mode until cells reached an optical density of >100 Klett units (1 Klett unit equals ~10^7 cells mL^{-1}) [4]. Cells were continuously fed with medium using Watson Marlow model 120U peristaltic pumps (Watson Marlow Inc., Wilmington, MA) and media was wasted using Masterflex model C/L peristaltic pumps (Cole-Palmer Instrument Co., Vernon Hills, IL). Reactors were checked approximately every 12 hours, and when necessary, pumping was manually adjusted to correct for any changes in reactor volume due to imbalances of liquid inflow and outflow. Cultures were grown for at least 5 retention times and analyzed when steady state was established (verified from constant turbidity measures and/or constant gas production rates).

Light supply rates

The measured incident light intensity at the PAR measured in W m^{-2} was converted into standard units of photon flux (PF) of μmol photon m^{-2} s^{-1} at a wavelength of 800 nm using the equation:

$$\text{PF} = \frac{I * \lambda * 10^{-9}}{h * c * N_A}$$

where I is the irradiance in W m^{-2}; λ is the wavelength of light in nm; h is Planck's constant (6.626068 x 10^{-34} m^2 kg s^{-1}); c is the speed of light (2.997925 x 10^8 m s^{-1}) and N_A is Avogadro's number (6.02 x 10^{17} μmol^{-1}).

The PF was then converted into a specific light supply rate (SLSR) in mmol g DW^{-1} h^{-1} for model simulations with the equation:

$$\text{SLSR} = \frac{PF * 3600 * A_{reactor} * \Phi}{DW * V_{reactor}}$$

where PF is the photon flux in μmol photon m^{-2} s^{-1}; $A_{reactor}$ is the surface area of the bioreactor in m^2; DW is the cell dry weight in g L^{-1}; $V_{reactor}$ is the volume of the bioreactor in mL; and Φ is the efficiency of light utilization taken to be 0.75 based on best fits from regression of max. ATP production on μ (see Results).

Light conversion to H_2 efficiency was calculated [39] with the equation:

$$\%\ \text{Efficiency} = \frac{33.61 * \rho_{H2} * V_{H2}}{I * A}$$

where ρ_{H2} is the density of H_2 (0.08988 g/L); V_{H2} is the volume of H_2 produced per hour in Lhr^{-1}; I is the irradiance in W m^{-2}; and A is the total surface area of the bioreactor in m^2. Substrate conversion to H_2 efficiency was calculated using succinate as the substrate [39, 40].

Quantification of biomass and cell dry weight

Samples from individual chemostats were centrifuged at 5,500 rcf at 4°C to obtain cell pellets for biomass analysis. Total cellular protein was quantified via the Lowry assay [45, 46], while total cellular bacteriochlorophyll was determined spectrophotometrically [47]. The phospholipid component was determined by total phosphorus assay on lipids extracted via standard chloroform/methanol extraction [48]. The PHB content of cells was determined by GC-MS (GC-2010 gas chromatograph coupled to a QP-2010S mass spectrometer detector; Shimadzu Scientific) [9]. Other biomass components were assumed to remain constant based on previous measurements [4]. Cell dry weight was calculated using the chemical oxygen demand (COD) mass balance approach [4, 9].

Quantification of nutrient uptake and biomolecule secretion

To analyze the presence of media nutrients and products, culture supernatants were filtered through a 0.22 μm Nylon Filter (Fisher Scientific, Pittsburg, PA) and frozen at -80 °C until analysis. The uptake rates for succinate, glutamate and aspartate were measured as previously described [9] by comparing their concentrations in the filtered supernatant from the chemostat cultures to that in the initial media. Sulfate was quantified using Dionex 2100 Series Ion Chromatogram (Thermo Scientific) using EPA method 300.1. Ammonia concentrations were analyzed using the salicylate method (Method 10031, Hach Company, Loveland, CO). Chemical oxygen demand (COD) was analyzed using the reactor digestion method (Method 8000, Hach Company, Loveland, CO).

Lactate, pyruvate and malate were measured using GC-MS [9]. Glucose, xylose, pyruvate, xylitol, glycerol, formate, acetate and ethanol were quantifying using HPLC (High Performance Liquid Chromatography) with a C18 column and a Refractive Index Detector (RID-10A) on an Agilent system (Agilent Technologies, Santa Clara, CA), in collaboration with the Metabolomics Facility of the Great Lakes Bioenergy Research Center. Total gas production and the gas composition of the headspace were determined as previously described [9].

Assessment of media for secretion of untargeted SMPs

To test if other SMPs not targeted in our analysis might be present in the media, the measured COD of the culture supernatants was compared to the estimated COD of the supernatant. The estimated COD of the supernatant was calculated based upon the measured concentrations of succinate, glutamate and aspartate in the culture supernatant. The estimated COD contribution for each compound was calculated using the following formula:

$$ThCOD_i = (\sum \mathrm{C_i} + 0.25 * \sum \mathrm{H}_i + 0.5 * \sum \mathrm{N}_i - 0.75 * \sum \mathrm{O}_i) * (32 mg\ \mathrm{O_2}/mmol) * [\mathrm{Conc}_i]$$

where $\mathbf{\Sigma}C_i$, $\mathbf{\Sigma}H_i$, $\mathbf{\Sigma}N_i$ and $\mathbf{\Sigma}O_i$ are the number of carbon, hydrogen, nitrogen and oxygen molecules, respectively present in compound *i*; $[Conc_i]$ is the concentration of compound *i* in units of mmol L^{-1}. $ThCOD_i$ is the estimated COD_i in units of mgCOD L^{-1} [49].

Constraint based analysis

The genome-scale metabolic model for *R. sphaeroides*, iRsp1140 [6], was used for all simulations. Flux balance analysis (FBA) [50] was used to simulate *in silico* growth by solving the linear programming problem:

$$\begin{aligned} &\max\ v_{biomass} \\ &s.t. \\ &S \bullet v = 0 \\ &v_{\min} \le v \le v_{\max} \end{aligned}$$

where $v_{biomass}$ is the flux through biomass objective function; v is the vector of steady state reaction fluxes; and v_{min} and v_{max} are the minimum and maximum allowable fluxes set to -1000 and 1000 mmol g DW^{-1} h^{-1} for reversible reactions and 0 and 1000 mmol g DW^{-1} h^{-1} for forward only reactions. Measured uptake rates for light, succinate, glutamate, aspartate and sulfate, as well as the measured growth rates and the production rates of PHB and H_2 were used as initial input constraints for modeling when required. Other

media components were allowed to be freely exchanged with the extracellular space. FVA [43] and alternative optima analysis [51, 52] was conducted as previously described [4]. For assessing differences in flux distribution between light limiting and saturating conditions, 1000 alternative solutions were sampled from the optimal solution space and summary statistics obtained from these were used generate the box plots in Figure 4-6. All simulations were conducted under the GAMS programming environment (GAMS Development Corporation, Cologne, Germany) using the CPLEX solver.

Determination maintenance energy requirements

To determine the cellular maintenance energy requirements at a given light intensity, the maximum ATP hydrolysis rate for each reactor was calculated by constraining the model with uptake rates for light, succinate, glutamate, aspartate and sulfate and the experimentally determined growth rate. At each light intensity, the photosynthetic biomass reaction of iRsp1140 (RXN1306) was replaced with a new biomass reaction corrected for condition-specific differences in biomass composition and removing the previously estimated GAM value of 53.65 mmol ATP g DW^{-1} from the reaction. Simulations were conducted by maximizing the ATP hydrolysis reaction (RXN0765) under these constraints. For each light intensity, the maximum ATP hydrolysis rate ($q_{ATP\text{-}total}$) for each reactor was plotted against the specific growth rate (μ). The ATP balance for anaerobic photoheterotrophic growth can be considered as:

$$q_{ATP-carbon} + q_{ATP-light} = q_{ATP-total} = GAM_{ATP} * \mu + NGAM_{ATP}$$

where $q_{ATP\text{-}carbon}$ is the ATP production rate from carbon sources taken up from the media (succinate, glutamate and aspartate), $q_{ATP\text{-}light}$ is the ATP production rate from light and $q_{ATP\text{-}total}$ is the ATP production/hydrolysis rate all in mmol g DW^{-1} h^{-1}. GAM_{ATP} is the growth associated maintenance energy requirement (in mmol g DW^{-1}) and $NGAM_{ATP}$ is the non-growth associated maintenance energy requirement (in mmol g DW^{-1} h^{-1}). GAM_{ATP} and $NGAM_{ATP}$ were determined from the slope and y-intercept, respectively, of the regression of $q_{ATP\text{-}total}$ on μ.

References

1. Gronenberg LS, Marcheschi RJ, Liao JC: **Next generation biofuel engineering in prokaryotes**. *Curr Opin Chem Biol* 2013, **17**(3):462-471.
2. Peralta-Yahya PP, Zhang F, del Cardayre SB, Keasling JD: **Microbial engineering for the production of advanced biofuels**. *Nature* 2012, **488**(7411):320-328.
3. Atsumi S, Higashide W, Liao JC: **Direct photosynthetic recycling of carbon dioxide to isobutyraldehyde**. *Nat Biotechnol* 2009, **27**(12):1177-1180.
4. Imam S, Yilmaz S, Sohmen U, Gorzalski AS, Reed JL, Noguera DR, Donohue TJ: **iRsp1095: a genome-scale reconstruction of the *Rhodobacter sphaeroides* metabolic network**. *BMC Syst Biol* 2011, **5**:116.
5. Mackenzie C, Eraso JM, Choudhary M, Roh JH, Zeng X, Bruscella P, Puskas A, Kaplan S: **Postgenomic adventures with *Rhodobacter sphaeroides***. *Annu Rev Microbiol* 2007, **61**:283-307.
6. Imam S, Noguera DR, Donohue TJ: **Global insights into energetic and metabolic networks in *Rhodobacter sphaeroides***. *BMC Syst Biol* 2013, **7**(1):89.
7. Kim E, Lee M, Kim M, Lee JK: **Molecular hydrogen production by nitrogenase of *Rhodobacter sphaeroides* and by Fe-only hydrogenase of *Rhodospirillum rubrum***. *International Journal of Hydrogen Energy* 2008, **33**(5):1516-1521.
8. Kontur WS, Ziegelhoffer EC, Spero MA, Imam S, Noguera DR, Donohue TJ: **Pathways involved in reductant distribution during photobiological H_2 production by *Rhodobacter sphaeroides***. *Appl Environ Microbiol* 2011, **77**(20):7425-7429.
9. Yilmaz LS, Kontur WS, Sanders AP, Sohmen U, Donohue TJ, Noguera DR: **Electron partitioning during light- and nutrient-powered hydrogen production by *Rhodobacter sphaeroides***. *Bioenerg Res* 2010, **Volume**(1):55 - 66.
10. Khatipov E, Miyake, M., Miyake J. and Y. Asada: **Polyhydroxybutyrate accumulation and hydrogen evolution by *Rhodobacter sphaeroides* as a function of nitrogen availability**. *Biohydrogen* 1999, **III**:157 - 161.
11. Kien NB, Kong IS, Lee MG, Kim JK: **Coenzyme Q10 production in a 150-l reactor by a mutant strain of *Rhodobacter sphaeroides***. *J Ind Microbiol Biotechnol* 2010, **37**(5):521-529.
12. Connor MR, Atsumi S: **Synthetic biology guides biofuel production**. *J Biomed Biotechnol* 2010.
13. Masepohl B, Hallenbeck PC: **Nitrogen and molybdenum control of nitrogen fixation in the phototrophic bacterium *Rhodobacter capsulatus***. *Adv Exp Med Biol* 2010, **675**:49-70.
14. Wahlund TM, Conway T, Tabita FR: **Bioconversion of CO_2 to ethanol and other compounds**. *American Chemical Society Division of Fuel Chemistry* 1996, **3**:1403-1405.
15. Oberhardt MA, Palsson BO, Papin JA: **Applications of genome-scale metabolic reconstructions**. *Mol Syst Biol* 2009, **5**:320.
16. Carapezza G, Umeton R, Costanza J, Angione C, Stracquadanio G, Papini A, Lio P, Nicosia G: **Efficient behavior of photosynthetic organelles via Pareto optimality, identifiability, and sensitivity analysis**. *ACS Synth Biol* 2013, **2**(5):274-288.
17. Vu TT, Stolyar SM, Pinchuk GE, Hill EA, Kucek LA, Brown RN, Lipton MS, Osterman A, Fredrickson JK, Konopka AE *et al*: **Genome-scale modeling of light-driven reductant partitioning and carbon fluxes in diazotrophic unicellular**

cyanobacterium *Cyanothece* sp. ATCC 51142. *PLoS Comput Biol* 2012, **8**(4):e1002460.
18. Kliphuis AM, Klok AJ, Martens DE, Lamers PP, Janssen M, Wijffels RH: **Metabolic modeling of *Chlamydomonas reinhardtii*: energy requirements for photoautotrophic growth and maintenance**. *J Appl Phycol* 2011, **24**(2):253-266.
19. Montagud A, Navarro E, Fernandez de Cordoba P, Urchueguia JF, Patil KR: **Reconstruction and analysis of genome-scale metabolic model of a photosynthetic bacterium**. *BMC Syst Biol* 2010, **4**:156.
20. Rex R, Bill N, Schmidt-Hohagen K, Schomburg D: **Swimming in light: a large-scale computational analysis of the metabolism of *Dinoroseobacter shibae***. *PLoS Comput Biol* 2013, **9**(10):e1003224.
21. Dal'Molin CG, Quek LE, Palfreyman RW, Nielsen LK: **AlgaGEM--a genome-scale metabolic reconstruction of algae based on the *Chlamydomonas reinhardtii* genome**. *BMC Genomics* 2011, **12 Suppl 4**:S5.
22. Nogales J, Gudmundsson S, Knight EM, Palsson BO, Thiele I: **Detailing the optimality of photosynthesis in cyanobacteria through systems biology analysis**. *Proc Natl Acad Sci U S A* 2012, **109**(7):2678-2683.
23. Stouthamer AH, van Verseveld HW: **Microbial energetics should be considered in manipulating metabolism for biotechnological purposes.** *Trends in biotechnology* 1987, **5**(5):149-155.
24. Russell JB, Cook GM: **Energetics of bacterial growth: balance of anabolic and catabolic reactions**. *Microbiol Rev* 1995, **59**(1):48-62.
25. Tannler S, Decasper S, Sauer U: **Maintenance metabolism and carbon fluxes in *Bacillus* species**. *Microb Cell Fact* 2008, **7**:19.
26. Aiking H, Sojka G: **Response of *Rhodopseudomonas capsulata* to illumination and growth rate in a light-limited continuous culture**. *J Bacteriol* 1979, **139**(2):530-536.
27. Schumacher A, Drews G: **Effects of light intensity on membrane differentiation in *Rhodopseudomonas capsulata***. *Biochim Biophys Acta* 1979, **547**(3):417-428.
28. Golecki JR, Schumacher A, Drews G: **The differentiation of the photosynthetic apparatus and the intracytoplasmic membrane in cells of *Rhodopseudomonas capsulata* upon variation of light intensity**. *Eur J Cell Biol* 1980, **23**(1):1-5.
29. Oelze J: **Regulation of tetrapyrrole synthesis by light in chemostat cultures of *Rhodobacter sphaeroides***. *J Bacteriol* 1988, **170**(10):4652-4657.
30. Biel AJ: **Control of bacteriochlorophyll accumulation by light in *Rhodobacter capsulatus***. *J Bacteriol* 1986, **168**(2):655-659.
31. Campbell TB, Lueking DR: **Light-mediated regulation of phospholipid synthesis in *Rhodopseudomonas sphaeroides***. *J Bacteriol* 1983, **155**(2):806-816.
32. Zijffers JW, Schippers KJ, Zheng K, Janssen M, Tramper J, Wijffels RH: **Maximum photosynthetic yield of green microalgae in photobioreactors**. *Mar Biotechnol (NY)* 2010, **12**(6):708-718.
33. Aagaard J, Sistrom WR: **Control of synthesis of reaction centre bacteriochlorophyll in photosynthetic bacteria**. *Photochem Photobiol* 1972, **15**(2):209-225.
34. Chory J, Kaplan S: **Light-dependent regulation of the synthesis of soluble and intracytoplasmic membrane proteins of *Rhodopseudomonas sphaeroides***. *J Bacteriol* 1983, **153**(1):465-474.

35. Kiley PJ, Kaplan S: **Molecular genetics of photosynthetic membrane biosynthesis in *Rhodobacter sphaeroides***. *Microbiol Rev* 1988, **52**(1):50-69.
36. Donohue TJ, Cain BD, Kaplan S: **Purification and characterization of an N-acylphosphatidylserine from *Rhodopseudomonas sphaeroides***. *Biochemistry* 1982, **21**(11):2765-2773.
37. Marinetti GV, Cattieu K: **Lipid analysis of cells and chromatophores of *Rhodopseudomonas sphaeroides***. *Chemistry and Physics of Lipids* 1981, **28**(3):241 - 251.
38. Miyake J, Kawamur S: **Efficiency of light energy conversion to hydrogen by the photosynthetic bacterium *Rhodobacter sphaeroides***. *Int J Hydrogen Energy* 1987, **12**(3):147-149.
39. Koku H, Eroglu I, Gunduz U, Yucel M, Turker L: **Aspects of the metabolism of hydrogen production by *Rhodobacter sphaeroides***. *Int J Hydrogen Energy* 2002, **27**(11-12):1315-1329.
40. Sasikala K, Ramana VC, Rao RP, Kovacs KL: **Anoxygenic phototrophic bacteria: physiology and advances in hydrogen production technology**. *Adv Appl Microbiol* 1993, **38**:211-295.
41. Thiele I, Palsson BO: **A protocol for generating a high-quality genome-scale metabolic reconstruction**. *Nat Protoc* 2010, **5**(1):93-121.
42. Allen JF: **Cyclic, pseudocyclic and noncyclic photophosphorylation: new links in the chain**. *Trends Plant Sci* 2003, **8**(1):15-19.
43. Mahadevan R, Schilling CH: **The effects of alternate optimal solutions in constraint-based genome-scale metabolic models**. *Metab Eng* 2003, **5**(4):264-276.
44. Sistrom WR: **A requirement for sodium in the growth of *Rhodopseudomonas spheroides***. *J Gen Microbiol* 1960, **22**:778-785.
45. Hartree EF: **Determination of protein: a modification of the Lowry method that gives a linear photometric response**. *Anal Biochem* 1972, **48**(2):422-427.
46. Lowry OH, Rosebrough NJ, Farr AL, Randall RJ: **Protein measurement with the Folin phenol reagent**. *J Biol Chem* 1951, **193**(1):265-275.
47. Cohen-Bazire G, Sistrom WR, Stanier RY: **Kinetic studies of pigment synthesis by non-sulfur purple bacteria**. *J Cell Physiol* 1957, **49**(1):25-68.
48. Rouser G, Fkeischer S, Yamamoto A: **Two dimensional then layer chromatographic separation of polar lipids and determination of phospholipids by phosphorus analysis of spots**. *Lipids* 1970, **5**(5):494-496.
49. Rittmann B, McCarty PL: **Environmental biotechnology: principles and applications**. New York: McGraw-Hill Science Engineering; 2001.
50. Varma A, Palsson BO: **Metabolic flux balancing: basic concepts, scientific and practical use.** *Nature Biotechnology* 1994, **12**:994 - 998.
51. Lee S, Phalakornkule, C., Domach, M.M., and Grossmann, I.E.: **Recursive MILP model for finding all the alternate optima in LP models for metabolic networks**. *Computers & Chemical Engineering* 2000, **24**(2 - 7):711 -716.
52. Reed JL, Palsson BO: **Genome-scale in silico models of *E. coli* have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states**. *Genome Res* 2004, **14**(9):1797-1805.

Chapter 5

An integrated approach to reconstructing genome-scale transcriptional regulatory networks

This chapter is formatted as a manuscript and has been submitted for publication:

Imam S, Noguera DR and Donohue TJ.

I performed all the experiments and analyses in this chapter.

Abstract

Transcriptional regulatory networks (TRNs) program cells to dynamically alter their gene expression in response to changing internal or environmental conditions. In this study, we employ a novel workflow to generate a large-scale TRN that integrates comparative genomics data, global gene expression analyses, and intrinsic properties of transcription factors (TFs). We used this workflow to build a large-scale TRN for the α-Proteobacterium *Rhodobacter sphaeroides* that includes 120 gene clusters, 1211 genes (including 93 TFs), 1858 predicted protein-DNA interactions and 76 DNA binding motifs. We found that ~67% of the predicted gene clusters in this TRN are enriched for specific functions ranging from photosynthesis or central carbon metabolism to environmental stress responses. We also found that members of many of the predicted gene clusters were consistent with prior knowledge in *R. sphaeroides* and/or other bacteria. We experimentally validated some TRN predictions by genomic analyses of TFs involved in photosynthesis (PpsR), carbon metabolism (RSP_0489) and iron homeostasis (RSP_3341), respectively. These analyses showed that the proposed workflow generates a TRN with precision and recall of up to 90% for individual TFs. In addition, we illustrate how this integrative approach enables generation of TRNs with increased information content relative to TRNs built via other approaches. We also show how this approach simultaneously produces a TRN for each related organism used in the comparative genomics analysis. Our results highlight the advantages of integrating comparative genomics of closely related organisms with gene expression data to assemble large-scale TRNs that can make high-quality predictions.

Introduction

Coordinating cellular behavior in response to internal or external signals requires dynamic regulation at several levels [1, 2]. Our ability to understand cellular dynamics requires detailed knowledge of each regulatory network and will, in part, depend on our ability to reconstruct models that integrate the datasets that report on these processes. Of the various levels at which cellular activities are regulated, transcriptional regulatory networks (TRNs) represent a particularly active area for modeling, as high-throughput techniques to monitor RNA levels and protein-DNA interactions can be applied in a wide range of organisms [2, 3]. Using such datasets, one can analyze, model and engineer TRNs [3, 4]. We are interested in reconstructing the TRNs of microbes because of the roles they play in health, agriculture, and biotechnology (such as nitrogen fixation, carbon cycling, food, chemical or fuel production, etc.).

Many published approaches to TRN inference depend on gene expression datasets to make predictions about direct interactions between transcription factors (TFs) and their target genes, assuming that the expression profile of a gene or cluster of genes, is directly related to that of a cognate TF(s) [5-11]. However, predictions based on this premise alone can be compromised by well-known indirect effects (e.g., co-expressed but not co-regulated genes) and post-transcriptionally regulated TFs, whose cellular levels remain relatively constant under conditions where their activity is significantly altered. In attempts to improve the TRN inference process, sequence analysis of the promoter regions of target genes has been used to inform models on the likelihood of a TF directly regulating a set of target genes [5, 6, 12-16]. In addition, the apparent conservation of TFs and regulatory interactions across species has been leveraged to build TRNs across related species [13-16]. While these individual approaches to TRN inference have their strengths and limitations, these approaches can be complementary and could potentially be combined to construct TRNs of greater coverage and better predictive power [3, 6]. In this work, we developed a workflow to construct TRNs which integrates comparative genomics data, global gene expression analyses, and intrinsic properties of transcription factors (TFs).

Purple non-sulfur bacteria (PNB) are a group of photosynthetic microbes displaying great breadth in their metabolic, bioenergetic and regulatory diversity [17-19]. *Rhodobacter sphaeroides* represents arguably the best studied PNB, serving for decades as a model system for photosynthetic growth, being used to understand photon capture, light-driven energy metabolism and other aspects of the photosynthetic lifestyle [19, 20]. In addition to anoxygenic photosynthetic growth, this facultative bacterium is capable of aerobic and anaerobic respiration [19]. *R. sphaeroides* can also fix CO_2 and N_2, and produce H_2, polyhydroxybutyrate or other compounds of industrial importance [17-26]. Coordinating the activities that form the basis for this metabolic and bioenergetic diversity is likely to require a robust regulatory system as internal or environmental conditions change. Thus, gaining a detailed understanding of its TRN will be pivotal in extending our knowledge of how these various lifestyles and metabolic processes are regulated in this and possibly other related bacteria.

In this work, we use an integrated approach to develop a large-scale reconstruction of the *R. sphaeroides* TRN. Exploiting evolutionarily conserved aspects of transcriptional regulation in closely related α-Proteobaceteria, we combined a genome-wide comparative genomics-based approach with a compendium of *R. sphaeroides* gene expression data sets to identify clusters of co-regulated genes. Using known properties of TFs, we made predictions on DNA binding proteins that are likely to regulate individual gene clusters. By focusing on sub-networks predicted to be pertinent to major cellular processes, we show that predictions of our TRN are consistent with prior knowledge in this and related bacteria. In addition, experimental analysis of select TFs using chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) and global gene expression analyses, provided direct validation of the predictive power of this large-scale TRN model. Our analyses illustrate the utility of this approach to assemble TRNs, highlighting the workflow and the new insights into important biological processes in this and other organisms that result from the inferred large-scale TRN.

Results and Discussion

TRN Inference

To reconstruct the large-scale *R. sphaeroides* TRN, we developed an approach that uses both sequence information from closely related bacteria and gene expression data (Figure 5-1). Our network inference approach began with the selection of 7 other closely related α-Proteobacteria (Additional File 1 – Figure S1, Materials and Methods). Using a comparative genomics-based approach, we conducted a genome-wide analysis of these bacteria to identify evolutionarily conserved DNA sequence motifs upstream of orthologous genes in the 8 selected genomes. This analysis led to the identification of 914 evolutionarily conserved DNA sequence motifs. *R. sphaeroides* genes possessing conserved DNA sequence motifs of high similarity were then grouped together to generate 76 gene clusters with shared evolutionarily conserved DNA sequence motifs. These comparative genomics-based clusters were then integrated with the several hundred publically-available *R. sphaeroides* gene expression datasets to identify conditions (or sub-conditions) under which members of each gene cluster are co-expressed [27]. This process generated clusters of co-regulated genes with both shared DNA sequence motifs and gene expression patterns.

We subsequently made predictions on which of the 216 known or predicted TFs in the *R. sphaeroides* genome most likely regulate these identified clusters by exploiting 4 known characteristics of bacterial TFs, which included: (i) *correlation* in expression profiles between a TF and its target genes [3, 6-8]; (ii) *proximity* of a TF to the location of its closest binding site [12, 14, 28, 29]; (iii) similarity in DNA motifs bound by TFs having similar *DNA binding domains* [29, 30]; and (iv) *phylogenetic correlation* of the occurrence of a TF and occurrence of a DNA sequence motif across species [29]. This analysis allowed assignment of high scoring TFs to 60 of the 76 identified gene clusters. The resulting large-scale TRN consisted of 76 gene clusters and DNA sequence motifs, 49 TFs and 1217 TF (or motif)-target interactions (Additional File 2 – Table S1).

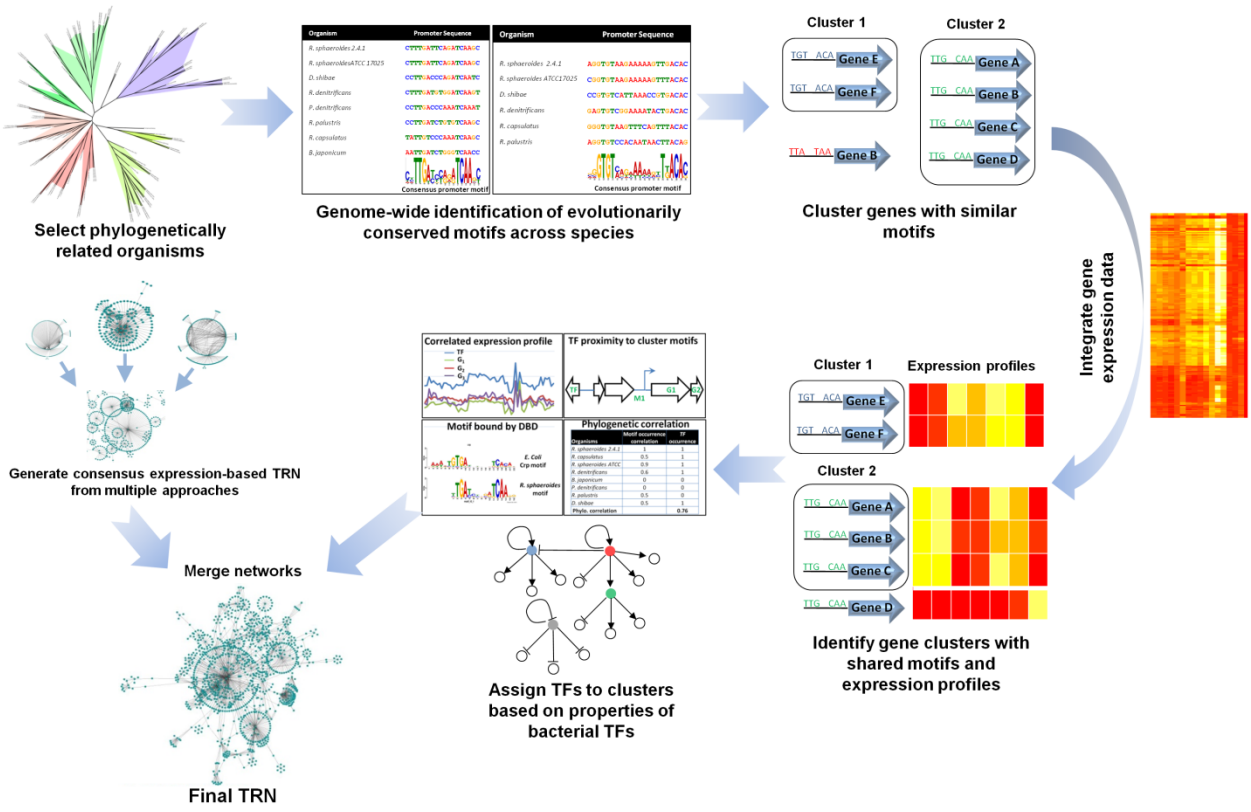

Figure 5-1. Overview of TRN reconstruction approach. A summary of the various steps involved in our TRN reconstruction workflow.

This approach enabled us to identify several conserved clusters of putatively co-regulated genes, but its utility can be limited by the evolutionary distance and the degree of conservation of the individual regulatory modules across the organisms used to generate the TRN. For example, it may be difficult to identify conserved regulatory sequences across closely related species if these sequences or regulatory mechanisms have undergone significant evolution. Furthermore, individual sub-networks that are specific to a lifestyle or response of an individual species, genus and/or clade might not be captured via a comparative genomics-based approach. For instance, of the 216 known/predicted TFs in *R. sphaeroides*, 42 were unique to *R. sphaeroides* 2.4.1 strain (Additional File 2 – Table S5). Thus, a comparative genomics-based approach might not be able to make predictions for these TFs.

To complement predictions from the comparative genomics-based analysis, we also utilized the consensus predictions of multiple high performing direct expression-based inference approaches [3, 9-11], to make predictions for TFs not included in our initial TRN. Using this complementary approach, we made predictions for an additional 44 *R. sphaeroides* TFs, corresponding to a total of 641 TF- target interactions.

In the final step, we fused the 2 predicted networks to generate a combined TRN consisting of 120 clusters, 93 TFs, 76 distinct evolutionarily conserved DNA sequence motifs and 1858 TF (or motif)-target interactions (Figure 5-2, Additional File 2 – Table S1). This large-scale TRN encompasses 1211 *R. sphaeroides* genes (about 28% of the open reading frames predicted in its genome [31, 32]). Details of the implementation of the steps used in reconstruction of the TRN are provided elsewhere (Materials and Methods, Additional File 2 –Tables S1-9).

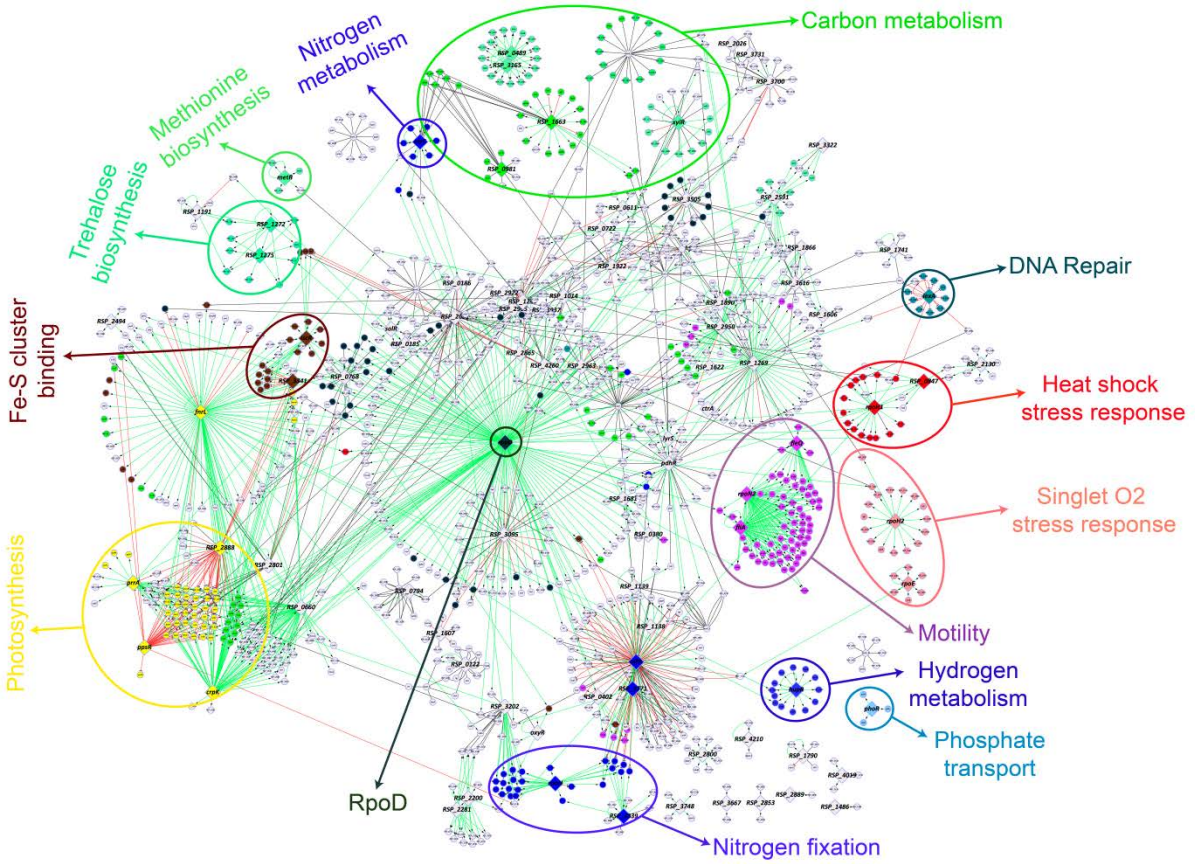

Figure 5-2. Overview of the reconstructed TRN for *R. sphaeroides*. A high-level visualization of the TRN constructed for *R. sphaeroides* consisting of 1221 nodes and 1858 edges. Some sub-networks consisting of genes and their regulating TFs enriched for different GO functional categories are highlighted. Green edges represent activation, red edges represent repression, while back edges indicate undetermined regulatory control. Cytoscape 3.0.2 [101] was used for network visualization.

Figure 5-3. Overview of functional categories captured in the TRN. Heat map showing the most significantly enriched GO terms for 48 of the 120 clusters identified in our analysis. The predicted regulators for each cluster is shown on the right hand side of the map, while the GO categories are at the bottom. Darker shades of blue indicated greater significance.

Overview of the Inferred TRN

Reconstructed TRN encompasses a wide variety of functions

This reconstructed TRN encompasses a wide variety of cellular functions ranging from central carbon metabolism and global stress responses, to processes more specific to *R. sphaeroides* like nitrogen fixation and photosynthesis (Figure 5-2, Figure 5-3). Of the 120 identified gene clusters, 80 were significantly enriched for at least one gene ontology (GO) [33] category (Additional File 2 – Table S1, Figure 5-3), indicating our TRN captures a high degree of functional information even though this type of functional data was not used in the network inference workflow. Below, we provide an overview of some of the predicted sub-networks in the TRN.

Photosynthesis

Previous analyses of the photosynthetic lifestyle of *R. sphaeroides* have implicated 3 TFs in this process: PpsR [34, 35], FnrL (a homolog of FNR) [36-38] and PrrA (the response regulator of the PrrAB two component system) [39-43] (Figure 5-4). More recently a small non-coding RNA, PcrZ has been implicated in the regulation of photosynthesis in *R. sphaeroides* [44]. Despite extensive prior analysis, our TRN predicts at least 2 additional regulators of photosynthesis: CrpK (RSP_2572) and RSP_2888 (Figure 5-4). To illustrate the predictive ability of our TRN, below we provide details about the known or predicted TFs in the *R. sphaeroides* photosynthetic lifestyle.

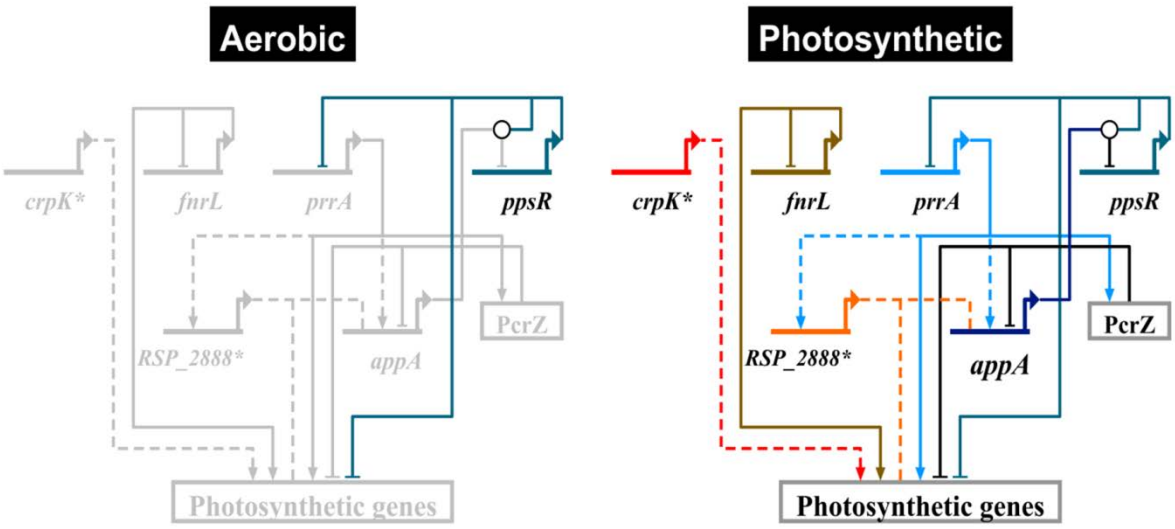

Figure 5-4. Photosynthetic gene regulatory network. An overview of the *R. sphaeroides* photosynthetic gene regulatory network, showing all the known/predicted transcriptional regulators. Solid lines indicate experimentally verified interactions, while dashed lines indicate predicted but as yet unverified interactions. Under aerobic conditions, AppA – the anti-repressor of PpsR – is inactivate, allowing PpsR to repress photosynthetic genes (grey nodes and edges indicate inactivity). Under photosynthetic conditions, AppA becomes active and interacts with PpsR via protein-protein interactions (depicted with a white circle), thereby inhibiting PpsR repression. The transcriptional activators, PrrA and FnrL become active under these conditions and drive the expression of photosynthetic genes. CrpK and RSP_2888 are also predicted to be involved in this process. PcrZ is a sRNA shown to negatively impact photopigment gene expression under photosynthetic conditions. Biotapestry was used for network visualization [102]. * indicates newly added components of the photosynthetic gene regulatory network identified in this study.

Previous analysis of **PpsR (RSP_0282)** in the regulation of photosynthesis identified this TF as a repressor of photopigment production under aerobic conditions [34, 35, 45, 46]. The activity of PpsR is regulated by its cognate anti-repressor, AppA, which is reported to respond to both oxygen and blue light [47-50]. To gain a more complete picture of the PpsR regulon, as well as assess the predictive performance of our inferred TRN for this TF, we determined the genome-wide binding of PpsR to its target sites by ChIP-seq using a 3X-myc tagged PpsR protein that complements a defined *ΔppsR* mutant. We identified a total of 19 PpsR binding sites in the genome that were located upstream of 15 operons, only 2 of which had been previously verified as direct targets for this TF [34] (Table 5-1, Figure 5-5A). Consistent with its role in regulation of photopigment formation, the majority of PpsR target operons had known or predicted photosynthesis-related functions (Table 5-1). Interestingly, PpsR was bound upstream of the *prrA* gene, which encodes another transcriptional regulator of photosynthesis in *R. sphaeroides* [39-43], suggesting a previously unknown genetic interaction between these TFs.

In addition to photosynthesis-related targets, PpsR was bound upstream of RSP_2095 and RSP_3000, which encode proteins of unknown function. However, these genes were not found to be significantly differentially expressed (DE) in a pair-wise comparison of RNA levels between a *ΔppsR* mutant and its parental strain [34], nor did their expression profiles show significant correlation to other members of the PpsR regulon across the available microarray dataset compendium (Figure 5-5B), suggesting these might represent non-functional binding sites in the genome, despite possessing strong PpsR motifs (Table 5-1). Consistent with the known role of PpsR as a transcriptional repressor, all DE PpsR targets we identified were predicted to be repressed by PpsR as RNA levels were increased in cells lacking this TF (Table 5-1).

Our TRN predicted a total of 13 PpsR target operons, 12 of which were verified via ChIP-seq analysis (Additional File 2 – Table S1 (cluster 60), Table 5-1), corresponding to a recall of 80% and a precision of 92.3%. The only predicted PpsR target site not verified by ChIP-seq analysis (RSP_4172 – a hypothetical protein) was classified as a false-positive since enrichment for PpsR binding was not detected by subsequent ChIP-qPCR analysis under the growth conditions tested (Additional File 1 – Figure S2). On

the other hand, 3 PpsR sites identified in our ChIP-seq assay were not predicted in our TRN (RSP_2095, RSP_3000 and *hemE*). However, given that putative targets such as RSP_2095 and RSP_3000 were not DE in the absence of PpsR (Table 5-1, Figure 5-5B), these might represent non-functional or false positive binding events. Independent ChIP-qPCR validation of ChIP-seq identified sites suggest that RSP_2095 and RSP_3000 are likely bound by PpsR but not DE under the conditions tested (Additional File 1 – Figure S2). Overall, our inferred TRN provided an accurate and expanded picture of PpsR binding sites across the genome with a large coverage of true binding sites. Thus, the consensus DNA sequence motifs obtained for PpsR from ChIP-seq and phylogenetic footprinting analysis are very similar (Figure 5-5C).

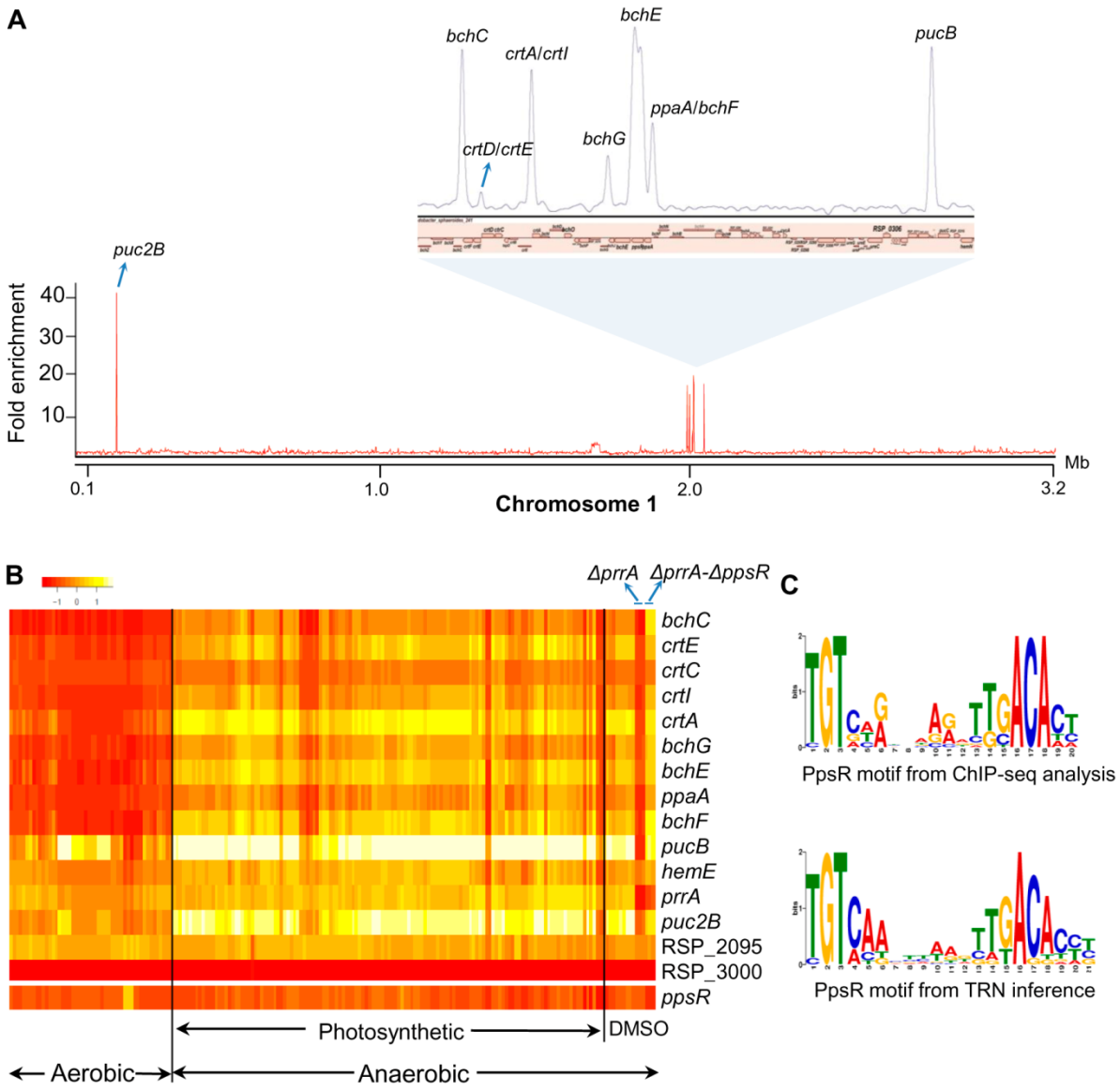

Figure 5-5. Analysis of the PpsR regulon in *R. sphaeroides*. (A) Using ChIP-seq, we identified the binding sites for PpsR across the *R. sphaeroides* genome, with several binding sites across chromosome 1 highlighted. MochiView [103] was used for visualization of binding profile. (B) Heat map depicts the expression profiles of the first members of PpsR targets operons across our microarray compendium of 198 experiments conducted under aerobic respiratory (Aerobic), anoxygenic photosynthetic (Photosynthesis) and anaerobic respiratory conditions (DMSO). Expression profiles for experiments conducted on the *ΔprrA* and *ΔprrA-ΔppsR* strains are highlighted. Deletion of PpsR from *ΔprrA* results in derepession of PpsR target genes. (C) Position weight matrix logo generated for PpsR using targets identified by ChIP-seq compared to logo generated from our TRN inference analysis.

Table 5-1. PpsR binding sites across the *R. sphaeroides* genome identifed by ChIP-seq

	ID	Annotation†	chrID	peakStart	peakStop	FC[a]	Motif	Expr[b]
1	RSP_0263-59[c]	*bchCXYZ-pufQ*	chr1	1987800	1988799	24.3	**TGT**CCAATAAAGTTG**ACA**CT	-36.61
2	RSP_0265-4[c]	*crtEF*	chr1	1990400	1990799	3.3	**TGT**AAGAAAAAGTTG**ACA**CC	-8.98
	RSP_0266-7[c]	*crtCD*						-2.65
3	RSP_0271-69[c]	*crtIB-tspO*	chr1	1996200	1996799	20.2	**TGT**CTAGTCAGGTTT**ACA**AT	-11.75
	RSP_0272-5[c]	*crtA-bchIDO*						-20.05
4	RSP_0279-6[c]	*bchG-pucC-bchP*	chr1	2005000	2005599	7.8	**TGT**AAGGATAGATTG**ACA**CT	-8.03
5	RSP_0281-80[c]	*bchEJ*	chr1	2007600	2009599	21.9	**TGT**CAACTGAAATGG**ACA**CA	-9.60
6							**TGT**CCAGTGCGTCTG**ACA**CT	
7	RSP_0283*[,c]	*ppaA*	chr1	2010000	2010799	12.6	**TGT**CAAAGAAAATTG**ACA**CC	-7.48
8	RSP_0284-91*[,c]	*bchFNBHLM-puhA*	chr1				**TGT**AAGTCAGAATTG**ACA**CT	-36.33
9	RSP_0314-RSP_6256[c]	*pucBA*	chr1	2042200	2042799	23.5	**TGT**CAGCGCAATGTG**ACA**CC	-112.17
10							**TGT**CAGCCAACACTG**ACA**TT	
11	RSP_0680	*hemE*	chr1	2424000	2424400	1.7	**TGT**CCATTTGCCCTG**ACA**AC	-2.23
12	RSP_1518[c]	*prrA*	chr1	105181	105204	2.1	**CGT**CAAAGGAAGTTG**ACA**CA	NA
13	RSP_1556-RSP_6158[c]	*puc2B2A*	chr1	146000	146599	58.4	**TGT**CTGCATGGCATG**ACA**TA	-8.99
14	RSP_2095	hypothetical protein	chr1	694600	694999	2.5	**TGT**GTGCGCAGTTGG**ACA**CC	-1.09
15	RSP_3000	hypothetical protein	chr1	1697500	1697700	3	**TGT**CCATATGGGTTG**ACA**TT	-1.21
16			chr1	4000	4200	3.5	**TGT**GTGTCAAGATGC**ACA**CT	ND
17			chr1	1680000	1680599	3.2	**TGT**CTATGACATTTC**ACA**AT	ND
18			chr2	4000	4200	3.4	**TGT**GTGTCAAGATGC**ACA**CT	ND
19			chr2	33000	39599	5	**TGT**GTGTCAAGATGC**ACA**CT	ND

* Previously experimentally verified as direct PpsR target

[a] Fold enrichment of PpsR-myc ChIP over control myc antibody ChIP in WT.

[b] Fold change in gene expression from microarray analysis of ΔPpsR and its parental strain. NA - Not applicable (*prrA* is deleted from both strains used for expression analysis). ND - Not determined (binding sites not located upstream of any annotated gene(s)).

[c] PpsR targets predicted in the TRN

† See Additional File 2 – Table S2 for descriptive gene annotations.

FnrL (RSP_0698) is an iron-sulfur cluster-containing Crp-family TF which previous studies have reported to be essential for anaerobic growth in *R. sphaeroides* [37, 38]. Previous ChIP-chip of genome-wide FnrL binding sites *in vivo* indicated the direct involvement of this TF in a host of processes including photosynthetic and anaerobic respiratory growth [36]. Our inferred TRN captured a significant portion of the known FnrL regulon, predicting a total of 59 FnrL target operons (Additional File 2 – Table S10, Additional File 2 – Table S1 (cluster 11)) that included 24 of the 25 previously identified FnrL target operons, a recall of 96%. The only previously verified FnrL target operon not identified in our analysis was RSP_6116, which is not represented on the *R. sphaeroides* gene chip, and thus dropped out during the integration of gene expression data. In addition to previously identified sites, our large-scale TRN predicted an additional 35 FnrL target operons not previously known or predicted to be under the control of FnrL (Additional File 2 – Table S10). Each of these new FnrL target operons have putative binding sites with strong similarity to the FnrL consensus and sharing a co-expression profile with other members of the FnrL cluster (Additional File 2 – Table S3). Several of these newly predicted FnrL targets encode functions for this TF has been previously implicated, including the regulation of Fe-S cluster biogenesis (e.g., RSP_1949) and Fe-S binding proteins (e.g., RSP_0692_89 - *rdxBHIS*). However, several new functions for FnrL that are predicted in this data set need to be tested experimentally. If these predictions are correct, it would significantly broaden the functional role of FnrL in this species.

In addition to PpsR and FnrL, whose regulons were globally characterized in this or previous studies, our TRN also made predictions for direct targets of less-well characterized TFs. For instance, our TRN made several new predictions for target of the photosynthesis regulator **PrrA (RSP_1518)**. PrrA has previously been proposed to be major global regulator in *R. sphaeroides* and other bacteria [40]. PrrA is essential for photosynthetic growth in *R. sphaeroides* and direct control of photosynthesis related operons, tetrapyrolle biosynthesis (*hemA*) and the Calvin–Benson–Bassham (CBB) cycle genes has be shown *in vitro* [42, 51]. Our TRN predicts that a total of 17 operons are directly regulated by PrrA (Additional File 2 – Table S1 (cluster 96)). Of these, 7 predicted PrrA target operons have a photosynthesis related role, including

pufLMX (RSP_0255-7), *pufA* (RSP_0258), *ppaA* (RSP_0283), *bchFNBHLM-puhA* (RSP_0284-91), *hemC* (RSP_0679), *hemA* (RSP_2984) and *appA* (RSP_1565). However, only two of these operons (*bchF* and *hemA*) have previously been experimentally verified as PrrA-dependent in *R. sphaeroides* [51], so direct analysis of PrrA binding to these newly proposed targets is required.

CrpK (RSP_2572) is a Crp/Fnr-family TF, which possesses predicted cyclic nucleotide-binding and Crp-like helix-turn-helix domains. However, unlike FnrL, CrpK does not possess N-terminal cysteine residues required for coordination of iron-sulfur clusters, suggesting CrpK might not directly sense oxygen. Our TRN predicts that CrpK regulates overlapping targets to FnrL, including several photosynthesis related operons such as *bchEJGP* (RSP_0281-76) and *hemA* (RSP_2984) (Additional File 2 – Table S1 (cluster 105)), as well as several other known FnrL target genes including *nuoA-N* (RSP_0100-12) and *ccoNOQP* (RSP_0696-3), amongst others. These predictions suggest CrpK could substitute for FnrL under some conditions, providing added, previously unappreciated, robustness to the photosynthetic TRN of this bacterium and possibly others containing homologs of both FnrL and CrpK.

RSP_2888 is a BadM/Rrf2 family TF predicted by our TRN to control photosynthesis gene expression in *R. sphaeroides*. Predictions from our TRN suggest a direct role of RSP_2888 in the regulation of a bacteriochlorophyll biosynthesis operon *bchFNBHLM* (RSP_0284-91), in addition to key photosynthesis related genes, such as *appA* (RSP_1565) (Additional File 2 – Table S1 (cluster 110)). RSP_2888 mRNA levels are increased under photosynthetic conditions in our expression datasets and this gene is predicted in our TRN to be under the control of PrrA. These observations are consistent with a proposed role for RSP_2888 in the photosynthesis sub-network of the TRN.

Overall our TRN captures a significant portion of the known regulatory interactions in the photosynthesis sub-network (Figure 5-4), while making a large number of novel predictions that should provide new insights into the complex combinatorial regulation of this lifestyle in PNB. Further experimental

validation will be required to fully understand the specific role played by RSP_2888, as well as CrpK, in the predicted regulation of photosynthesis functions.

Central and Alternative Carbon Metabolism

For cells to survive in nature, they must adapt to the types and quantities of nutrients present in their environment. For instance, *E. coli* uses the cAMP receptor protein (CRP), in part, to preferentially utilize glucose over other nutrient sources, if present in its environment [52]. On the other hand, the ArcAB two-component global regulator represses portions of *E. coli*'s central metabolic pathways under anaerobic respiratory conditions [53, 54]. In addition to these global regulators, the Cra/FruR regulator specifically regulates carbon and energy metabolism in enteric bacteria [55].

R. sphaeroides is not predicted to possess proteins analogous to CRP or ArcAB. However, our TRN predicts that the regulation of central carbon metabolism in *R. sphaeroides* is controlled by a LacI family transcriptional regulator, **RSP_1663**. RSP_1663 is predicted to regulate transcription of genes encoding the central carbon metabolism enzymes Mdh (RSP_0968), PckA (RSP_1680), malic enzyme (RSP_1593), PdhAB (RSP_2968-RSP_4047-RSP_4050), succinate dehydrogenase (RSP_0974-6), as well as glycolytic enzymes Zwf (RSP_2734), Pgl (RSP_2735), Pgi (RSP_2736) and FbaB (RSP_4045), potentially making this TF a major regulator of carbon metabolism under many conditions (Figure 5-6). This predicted RSP_1663 regulon might make it functionally analogous to the Cra/FruR regulator in enteric bacteria [55] and the RpiR family TF HexR in β- and γ-proteobacteria [56]. RSP_1663 is predicted to bind to an inverted repeat DNA motif with the sequence [A/G/T]GTT N_{6-8} AAC[A/C/T] (where N is any nucleotide) (Figure 5-6). In addition, differences in spacer between the inverted repeats divides the genes predicted to be regulated by this TF into 2 clusters (Additional File 2 – Table S1 (clusters 15 and 36)). Further experimental analysis is needed to understand the functional role of RSP_1663.

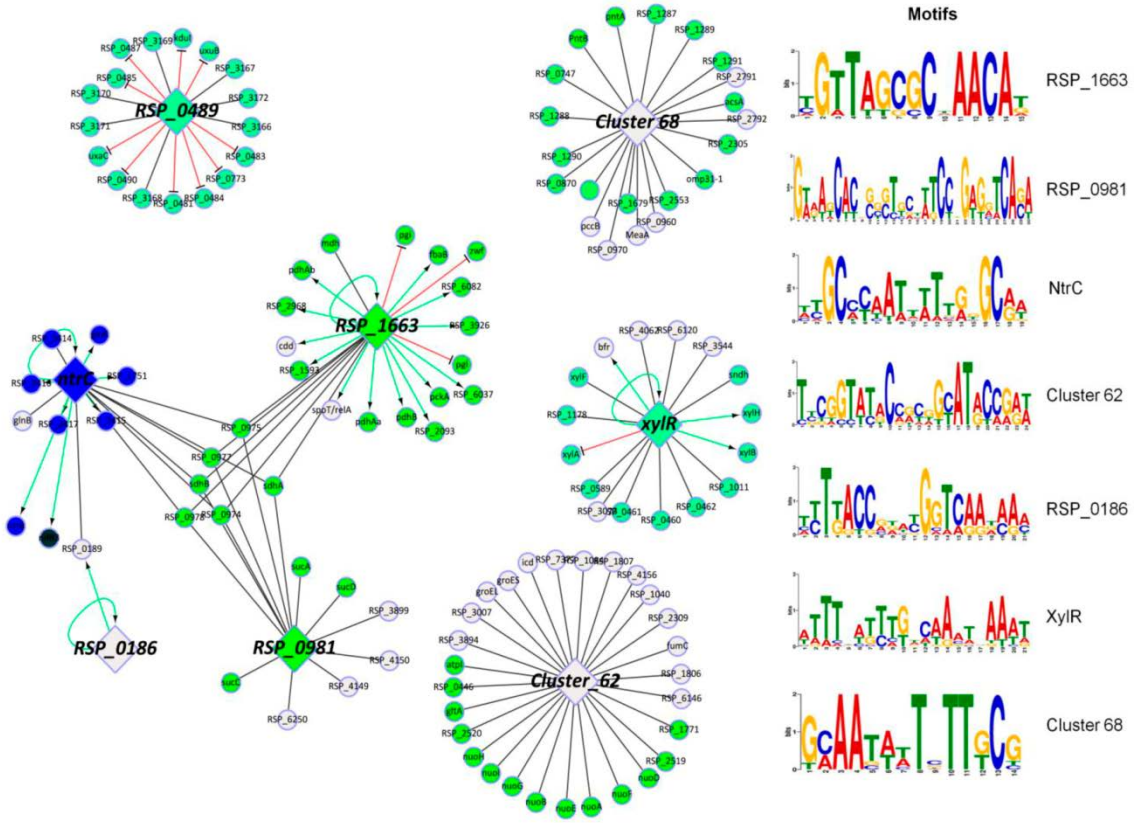

Figure 5-6. Predicted gene regulatory network controlling central and alternative carbon metabolism in *R. sphaeroides*. Sub-network highlighting the regulons of the major TFs predicted to be involved in the regulation of carbon metabolism in *R. sphaeroides*. TFs are represented by diamond shaped nodes while non-TF target genes are presented as circular nodes. Green edges represent activation, red edges represent repression, while black edges represent undetermined regulation. Green nodes indicate genes known or predicted to be involved in carbon metabolism, while blue nodes are related to nitrogen metabolism. Motifs predicted to be bound by the various TFs in this sub-network are shown on the right.

In addition to RSP_1663, **RSP_0981** – a GntR family transcriptional regulator, is predicted to regulate transcription of genes encoding the succinyl-CoA synthetase (RSP_0967-6), succinate dehydrogenase (RSP_0974-6) and α-ketoglutarate dehydrogenase (RSP_0965-62) complexes of the tricarboxylic acid cycle (Figure 5-6, Additional File 2 – Table S1 (cluster 48)), while **NtrC (RSP_2838)** is also predicted to be involved in the regulation of the succinate dehydrogenase complex (Figure 5-6, Additional File 2 – Table S1 (cluster 1)). **Cluster 62** in our TRN (Figure 5-6, Additional File 2 – Table S1) also contains a number of genes encoding enzymes involved in central carbon metabolism including Icd (RSP_0446 and RSP_1559), L-malyl-CoA lyase (RSP_1771), citrate synthase (RSP_1994) and NuoA-N (RSP_2512-23). The members of cluster 62 share the inverted repeat motif (Figure 5-6), indicating that these central metabolism genes are under the joint control of an as yet unidentified TF.

Our TRN also made predictions about regulation of metabolism of several other carbon sources. For instance, **RSP_0489** – a GntR family transcriptional regulator, is predicted to regulate transcription of genes encoding enzymes that are involved in the metabolism of carboxylic acids including UxuA (RSP_0773), UxaC (RSP_0488), KduID-UxuB (RSP_0482-80) and carbohydrate kinase (RSP_0490), as well as substrate transport (RSP_0487-3 and RSP_3168-5) (Figure 5-6, Additional File 2 – Table S1 (cluster 83)), making it functionally analogous to UxaR [57]. We tested these predictions by comparing RNA levels between wild type (WT) and *ΔRSP_0489* cells, and ChIP-seq with a myc-tagged version of RSP_0489 (Figure 5-7). A total of 55 genes were DE (1.5 fold change, p – 0.05) between WT and *ΔRSP_0489* cells, including predicted targets *uxuA*, *kduID-uxuB*, *uxaC* and RSP_0487-3, which were repressed in the presence of RSP_0489 by as much as 36-fold (Figure 5-7A, Additional File 2 – Table S11). Several other genes involved in substrate transport and metabolism were also DE in this data set (Figure 5-7A, Additional File 2 – Table S11). ChIP-seq analysis with a 3X myc tagged variant of RSP_0489 revealed that RSP_0489 binds at the promoters for *uxuA*, the *uxaC* operon (RSP_0488-0), RSP_0489, RSP_0490 and within the coding regions of substrate transporter (RSP_3372-70 and RSP_2667-3) (Figure 5-7B, Table 5-2), verifying several predictions from our TRN. Overall 4 out of

these 6 RSP_0489 target operons (~67%) were correctly predicted in our TRN. The conserved DNA sequence motif derived from sites bound by RSP_0489 also showed similarities to that obtained from phylogenetic footprinting analysis of the RSP_0489 promoter (Figure 5-7C). Other genomic locations enriched for RSP_0489 but with no corresponding DE genes are listed in Additional File 2 – Table S12.

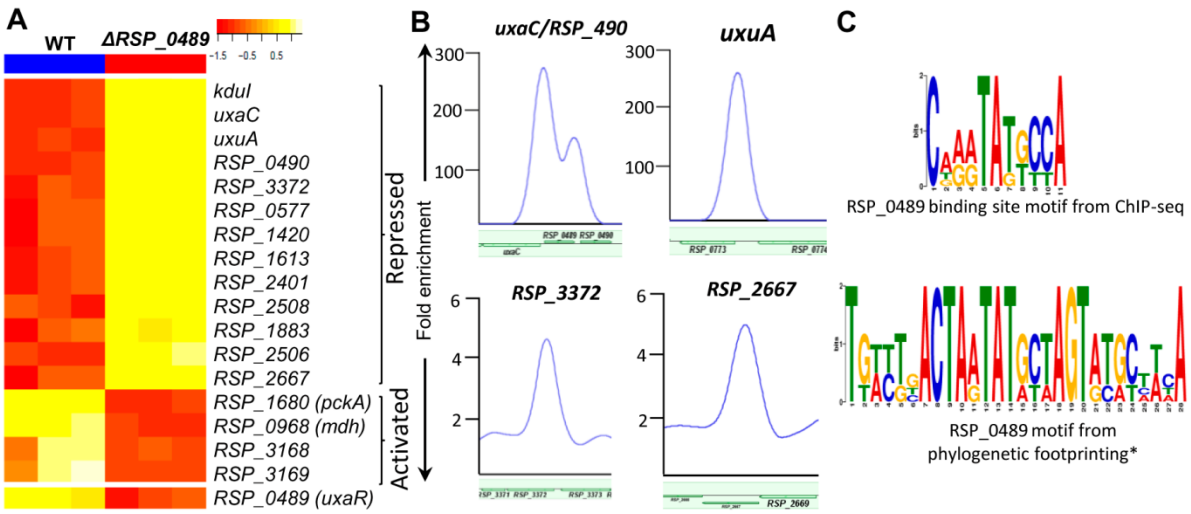

Figure 5-7. The RSP_0489 regulon. (A) Heat map of metabolic genes DE between wild-type (WT) and *ΔRSP_0489* mutant cells from global gene expression analysis. Only the first members of DE operons are depicted in the heat map for brevity. RSP_0490 (carbohydrate kinase), RSP_3372 (TRAP-T family transporter), RSP_0577 (hypothetical protein), RSP_1420 (TRAP-T family transporter), RSP_1613 (TRAP-T family transporter), RSP_2401 (putative 6-aminohexanoate-cyclic-dimer hydrolase), RSP_2508 (Methylcrotonyl-CoA carboxylase beta chain), RSP_1883 (ABC polyamine/opine transporter), RSP_2506 (Isovaleryl-CoA dehydrogenase), RSP_3168 (ABC transporter), RSP_3169 (FAA-hydrolase-family protein). (B) Direct binding of RSP_0489 to the *uxaC*, RSP_0490, *uxuA*, RSP_3372 and RSP_2667 promoters identified by ChIP-seq. (C) RSP_0489 binding site motif obtained from ChIP-seq analysis compared to that obtained from phylogenetic footprinting analysis of the RSP_0489 promoter.

Table 5-2. RSP_0489 direct targets identified by ChIP-seq and expression profiling

	ID	Annotation†	chrID	peakStart	peakStop	FC[a]	Motif	Expr[b]
1	RSP_0488-80*	*uxaC-kduID-uxuB*	chr1	2222600	2223600	248	TGTCTGA**CTAATATGCTA**GTATGC	-36
	RSP_0489*	GntR family TF	chr1					
2	RSP_0490*	Carbohydrate kinase	chr1	2223800	2224200	135	CGGCGGT**CAGATAGTCCA**CCTCCG	-2.4
3	RSP_0773*	*uxuA*	chr1	2515000	2515799	225	TAATATG**CAAGTATGCCA**GTTTGC	-26
4	RSP_3372-70	TRAP-T transporter	chr2	437000	437600	6	TCGCCCG**CGAATATGTCA**CGCGGG	-2.4
5	RSP_2667-3	ABC transporter	chr1	1310200	1310600	5	CATCGCG**CAGGTATTCCA**GTTTCC	-1.5

* RSP_0489 targets also predicted in TRN

[a] Fold enrichment of RSP_0489-myc ChIP over control myc antibody ChIP in WT.

[b] Fold change in gene expression in WT w.r.t ΔRSP_0489.

† See Additional File 2 – Table S2 for descriptive gene names

Another TF predicted to be involved in regulation of *R. sphaeroides* carbon metabolism is **RSP_1181**, which is annotated as a xylose operon repressor, *xylR*. Consistent with this annotation, RSP_1181 is predicted to regulate a putative xylose ABC transporter (RSP_1178-80), *xylB* (RSP_1177), *xylA* (RSP_1176), as well as its own expression in our TRN. RSP_1181 is also predicted to regulate transcription of L-sorbosone dehydrogenase (RSP_3072), as well as other metabolic and transport proteins, suggesting a potentially broader role for this TF beyond xylose utilization (Figure 5-6, Additional File 2 – Table S1 (cluster 37)). A phylogenetically conserved inverted repeat binding motif (TTTN[A/T]TTTG N CAAA[T/A]NAAA) was identified for this TF (Figure 5-6).

RSP_0185 and **RSP_0186** were predicted to regulate genes involved in pyrimidine utilization in *R. sphaeroides* such as allantoate amidohydrolase (RSP_0184), dihydropyrimidine dehydrogenase/glutamate synthase (RSP_0189), as well as their own expression. Both of these TetR family TFs are homologs of *E. coli* RutR, which is known to regulate pyrimidine and purine utilization [58]. Furthermore, the predicted phylogenetically conserved inverted repeat motif for these TFs (T[T/G]ACC N_4 CCT[C/A]AA) is similar to one previously identified for *E. coli* RutR [58] (Figure 5-6), indicating that our TRN might accurately capture the regulons of these TFs.

R. sphaeroides lacks enzymes required for the metabolism of acetate via the glyoxylate shunt, but instead utilizes the recently elucidated ethylmalonyl-CoA pathway for this purpose [59]. A large number of the genes encoding enzymes known or predicted to function in the ethylmalonyl-CoA pathway are found in **Cluster 68** (Additional File 2 – Table S1, Figure 5-6). These include (R)-3-hydroxybutyryl-CoA dehydrogenase (RSP_0747), crotonyl-CoA carboxylase/reductase (RSP_0960), (2R)-ethylmalonyl-CoA mutase (RSP_0961), Malyl-CoA thioesterase (RSP_0970), (2S)-methylsuccinyl-CoA dehydrogenase (RSP_1679) and propionyl-CoA carboxylase (RSP_2189), crotonase (RSP_2305). In addition, the NADH/NADPH transhydrogenase enzyme (RSP_0239-40) is also a member of cluster 68, which might be expected given the large requirement of NADPH for function of the ethylmalonyl-CoA pathway [18]. The members of this cluster are predicted to share a common conserved promoter motif of [G/T][C/A]AA

N_5 TT[G/T]C and have a shared co-expression pattern suggesting they are under joint control of a common TF. However, we were unable to link any high scoring TF to cluster 68, so additional experiments are needed to resolve this.

Overall, the predictions of our TRN suggest *R. sphaeroides* uses a robust TRN to control carbon metabolism. This is not surprising, given its ability to utilize at least 68 different carbon sources for growth [18]. Regulation of central metabolism is well studied in *E. coli* [55, 60] and *Bacillus subtilis* [61, 62], but not *R. sphaeroides* or other α-Proteobacteria. To fully utilize the properties of *R. sphaeroides* and other bacteria for biotechnological purposes, it is imperative that the gene regulatory networks controlling these crucial aspects of its metabolic network are better understood. Our TRN provides a blueprint to gaining additional understanding of carbon metabolism in *R. sphaeroides* and related bacteria.

Nitrogen Metabolism, Nitrogen Fixation and Hydrogen Production

Many α-Proteobacteria and most PNB are diazotrophs capable of fixing atmospheric nitrogen into ammonia [63]. Consistent with previous observations about the regulation of nitrogen fixation genes in *R. capsulatus* [63], the *R. sphaeroides* Mo-nitrogenase operon, *nifHDK* (RSP_0539-41) and accessory nitrogen fixation genes, *nifXNE* (RSP_0535-38), are predicted in our TRN to be under the joint control of **RpoN (RSP_0527)** and **NifA (RSP_0547)**. In other α-Proteobacteria, NifA is a known transcriptional activator of the nitrogenase structural genes, recruiting RpoN to the promoter of the *nif* operon to stimulate transcription under nitrogen-limiting conditions [63]. A similar positive control mechanism for NifA in nitrogen fixation is predicted in our TRN. In addition to RpoN and NifA, another TF predicted by our TRN to participate in the regulation of the *nif* operons is **RSP_3339** – a GntR family TF. In addition to the *nif* operons, RSP_3339 is also predicted to regulate spermidine/putrescine (RSP_3337-8) and di-/oligopeptide transporters (RSP_3892) (Additional File 2 – Table S1 (cluster 45)). Thus, RSP_3339 might control nitrogen metabolism in the presence of specific nitrogen sources or under certain nitrogen-limiting conditions. Finally, our TRN predicts that **RSP_3771** – a XRE-family TF, negatively regulates the

expression of the *nif* genes functioning antagonistically to NifA and potentially preventing the unnecessary expression of proteins involved in the energetically demanding process of nitrogen fixation.

In addition to the *nif* genes, RpoN is predicted to regulate a small number of other operons including those encoding proteins proposed to transfer electrons to the nitrogenase [24, 63] including *rnfABCDGEH* (RSP_3192-9) and the genes in the RSP_3191-88 operon (Additional File 2 – Table S1 (cluster 5)). The features of the predicted RpoN regulon in our TRN model could reflect either a highly specialized role for this σ factor in *R. sphaeroides* or a limitation of our approach in identifying targets for this transcriptional regulator.

NtrC is predicted in our TRN model to bind a phylogenetically conserved inverted repeat motif of GC N_{11} GC (Additional File 2 – Table S1 (cluster 1)). NtrC is predicted to regulate transcription of genes encoding proteins involved in maintaining cellular nitrogen homeostasis including GlnB (RSP_0146), GlnK (RSP_0889), AmtB (RSP_0888), dihydropyrimidine dehydrogenase/glutamate synthase (RSP_0189), as well as its own operon that includes genes like NifR3 (RSP_2836) and NtrB (RSP_2837). These predictions are consistent with the known role of NtrC in regulating genes and pathways in response to environmental nitrogen level [63].

The production of H_2 in *R. sphaeroides* is intimately linked to the nitrogen status of the cell, as the nitrogenase enzyme is the major source of H_2 in this bacterium [24, 26, 63]. Transcription of the uptake hydrogenase enzyme, which splits H_2 into protons and electrons, is increased under H_2 producing conditions [64]. H_2 is sensed via HupTUV, while the hydrogenase genes *hupSL* are directly bound by the response regulator HupR in *R. capsulatus* [65]. Consistent with these observations, our TRN predicts that genes encoding subunits of the uptake hydrogenase *hupSL* (RSP_0495-6) and several accessory proteins (RSP_0497-RSP_0509) are under the control of the response regulator, **HupR (RSP_0507)** in *R. sphaeroides* (Additional File 2 – Table S1 (cluster 84)).

Fe-S cluster biogenesis and iron homeostasis

Genes of the Fe-S biogenesis pathway (*iscSUA-hscBA-fdx*) are regulated by the Rrf2-family TF IscR, in *E. coli* and several other bacteria [66, 67]. In *E. coli*, IscR is a global regulator that is able bind to two different DNA target sequences depending on whether it is ligated to a 2Fe-2S cluster [66, 67]. The *R. sphaeroides* homolog of IscR, **RSP_0443**, differs from *E. coli* IscR as it does not possess cysteine residues required for the ligation to a 2Fe-2S cluster, suggesting that this protein is unable to ligate a Fe-S cluster. If this is true, then the upstream signaling pathway utilized and target genes regulated by RSP_0443 is likely to differ from that of *E. coli* IscR.

Consistent with observations in *E. coli*, RSP_0443 is predicted in our TRN to regulate transcription of its own operon (RSP_0443-31). However, the RSP_0443 operon encodes homologs of the Suf Fe-S biogenesis pathway (*sufABCDSE*), which is also a direct IscR target in *E. coli* [68]. In addition, RSP_0443 is predicted in our TRN to regulate transcription of catalase (RSP_2779), bacterioferritin-associated ferredoxin (RSP_1547), imelysin (RSP_1548), biopolymer transport protein TonB-ExbBD (RSP_0920-2), *napEFDABC* (RSP_4112-8), all gene products with Fe-S cluster or heme-binding domains or predicted to be involved in iron uptake. Thus, members of the predicted RSP_0443 regulon could play a significant role in maintaining cellular iron homeostasis, possibly to provide the metal needed for Fe-S centers. There is also a strong positive correlation between RSP_0443 RNA levels and transcription of its predicted target genes in *R. sphaeroides*, suggesting this TF functions as an activator.

In addition to RSP_0443, FnrL is directly involved in regulating transcription of genes encoding iron transporters such as *feoABC*, as well as a number of Fe-S and heme containing proteins in *R. sphaeroides*. Thus, our TRN predicts that RSP_0443 and FnrL both play an important role in regulation of cellular iron homeostasis. Furthermore, FnrL is also predicted in our TRN to directly activate **RSP_3341**, a putative iron binding RirA-like [69] protein in *R. sphaeroides*, which in turn is predicted to negatively regulate the putative 4Fe-4S binding nitrate reductase (*napEFDABC*). We tested this prediction by comparing

RNA abundance levels between wild type (WT) and *ΔRSP_3341* cells, and via ChIP-seq analysis using a myc-tagged version of RSP_3341 (Figure 5-8A). We found a total of 69 genes were DE (2 fold change, p <0.05) between WT and *ΔRSP_3341* cells including several members of the nitrate reductase operon (*napEFDABC*), which were all repressed by RSP_3341 (Figure 5-8A, Additional File 2 – Table S13). In addition, transcription of genes encoding other iron dependent proteins (such as cytochromes and ferredoxins), were also repressed by RSP_3341 (Figure 5-8A, Additional File 2 – Table S13). The mRNA level RSP_0443 was 2 fold higher in WT relative to *ΔRSP_3341* cells, suggesting there might be some cross talk between these TFs. We conducted ChIP-seq analysis with a 3X myc tagged version of RSP_3341 and confirmed the direct regulation of *napEFDABC* by this protein, consistent with our gene expression data and TRN predictions (Figure 5-8B). In addition, RSP_3341 binding was found near Hsp70 DnaK (RSP_1173) and *cycJ* (RSP_2945) (Figure 5-8B, Table 5-3). These genes were also DE in our gene expression dataset, thus were considered as additional direct RSP_3341 targets (Table 5-3). Twenty-two other sites showing significant enrichment for RSP_3341 but for which no genes in those genomic locations were DE are provided in Additional File 2 – Table S14. These data verify the prediction of our TRN of the involvement of RSP_3341 in the direct and indirect regulation of iron-dependent genes in *R. sphaeroides*.

Another RirA-like protein in *R. sphaeroides*, **RSP_2888** predicted to be important in regulation of photosynthesis is also involved in the regulation of iron containing proteins such as AppA (RSP_1565) and those involved in bacteriochlorophyll biosynthesis. Thus, the maintenance of iron homeostasis and the transcriptional regulation of genes ecoding iron-dependent enzymes appears to involve a complex gene regulatory network in *R. sphaeroides* (Figure 5-8C).

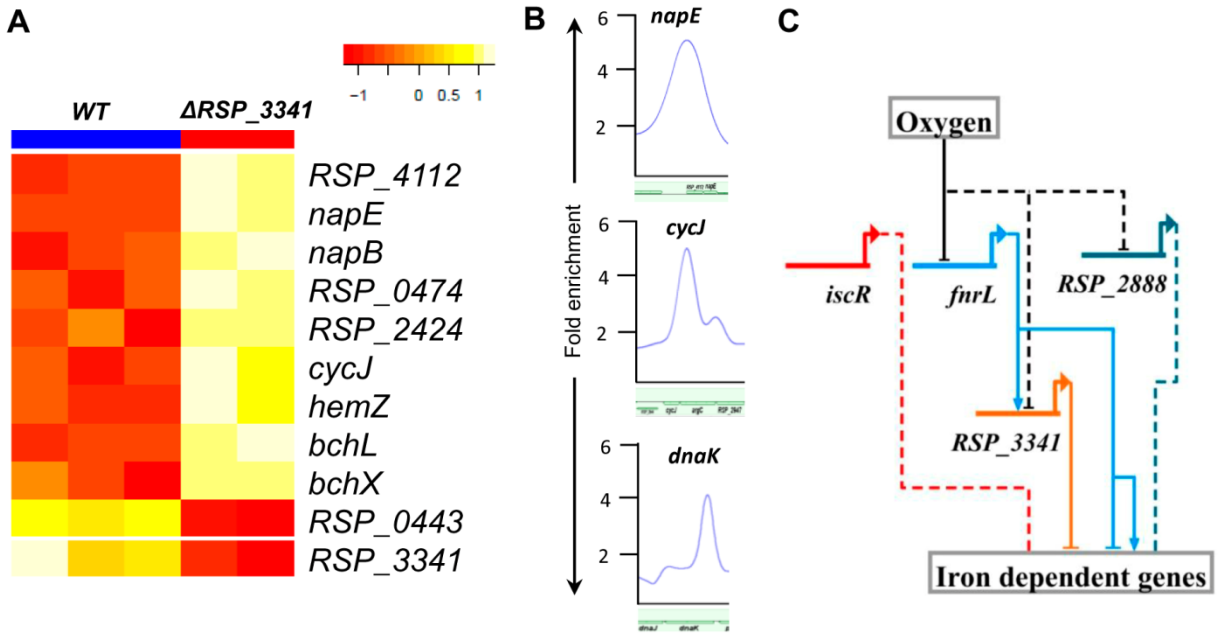

Figure 5-8. Regulation of iron-dependent genes in *R. sphaeroides*. (A) Heat map of iron-dependent DE genes between wild-type (WT) and *ΔRSP_3341* mutant cells from global gene expression analysis. RSP_4112 (hypothetical protein), RSP_0474 (Cytochrome c'), RSP_2424 (ferredoxin II), RSP_2945 (cytochrome c-type biogenesis protein CcmE). (B) Direct binding of RSP_3341 to the *napEFGABC*, *cycJ* and *dnaK* promoters identified by ChIP-seq. (C) Predicted gene regulatory network controlling iron-homeostasis in *R. sphaeroides*. Both RSP_2888 and RSP_3341 are RirA like proteins with C-terminal cysteine residues potentially capable of binding Fe-S clusters and sensing oxygen. Solid lines indicate experimentally verified interactions, while dashed lines indicated predicted but as yet unverified interactions.

Table 5-3. RSP_3341 direct targets identified by ChIP-seq and expression profiling

	ID	Annotation	chrID	peakStart	peakStop	FC[a]	Expr[b]
1	RSP_1173	Heat shock protein Hsp70 (*dnaK*)	chr1	2941200	2941599	4.4	-2.7
2	RSP_2945	cytochrome c-type biogenesis protein CcmE (*cycJ*)	chr1	1626000	1626599	5.0	-2.5
3	RSP_4112-8*	Nitrate reductase (*napEFDABC*)	plasmidC	79400	79799	5.1	-6.3

* RSP_3341 targets also predicted in TRN

[a] Fold enrichment of RSP_RSP_3341-myc ChIP over control myc antibody ChIP in WT.

[b] Gene expression in WT w.r.t ΔRSP_3341.

DNA Repair

Similar to other well studied bacteria such as *E. coli* [70, 71], the DNA repair machinery in *R. sphaeroides* is predicted in our TRN to be controlled by the **LexA (RSP_1997)** repressor (Additional File 2 – Table S1 (cluster 59)). Our TRN predicts LexA regulates transcription of *recA* (RSP_0452), putative DNA repair enzyme (RSP_1458), *lexA* (RSP_1997), *uvrD* (RSP_2092), DNA-binding protein (RSP_2234) and *uvrA* (RSP_2966), all genes known or predicted to be involved in DNA repair and which are verified LexA targets either in *R. sphaeroides* [71, 72] or other bacteria [70]. Our TRN also predicts that LexA binds to the direct repeat motif GTTC N_7 GTTC, consistent with a previously identified SOS box for *R. sphaeroides* [71].

In addition, our TRN predicts that LexA regulates genes involved in other processes such as class I diheme cytochrome c (RSP_0306), ferredoxin (RSP_0352), aldo/keto reductase (RSP_0423), UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase (RSP_2115), and several hypothetical proteins (RSP_0424, RSP_1656, RSP_2624, RSP_6187, and RSP_6249). Most of these predicted target genes are likely involved in various aspects of metabolism, suggesting that the functional role of LexA in *R. sphaeroides* might extend beyond DNA repair, as has been reported in some other bacteria [73, 74].

Flagella biosynthesis, cell motility and chemotaxis

R. sphaeroides has a single uni-directionally rotating flagellum used for swimming and chemotaxis [75]. The *R. sphaeroides* genome encodes 3 chemotaxis (*che*Op_1 (RSP_2443-33), *che*Op_2 (RSP_1583-89) and *che*Op_3 (RSP_0049-2)) and 2 flagella (RSP_0052-66 and RSP_0083-71) operons [76-78]. Consistent with previous observations [79, 80], our TRN predicts that the regulation of flagella biosynthesis and cell motility, as well as chemotaxis, is jointly controlled by 3 TFs: **FliA (RSP_0032), RpoN2 (RSP_0068) and FleQ (RSP_0071)** (Additional File 2 – Table S1 (clusters 77, 78 and 79)).

FliA and RpoN2 are predicted in our TRN to jointly regulate *che*Op$_2$ and *che*Op$_3$, as well as both flagella biosynthesis operons. In addition, several genes located outside of the chemotaxis and flagella operons but likely important in motility and chemotaxis were also predicted to be part of this shared regulon, including RSP_3083 (methyl accepting chemotaxis protein), RSP_3302 (chemotaxis response regulator CheY4), RSP_3303 (putative methyl accepting chemotaxis protein McpG) and others (Additional File 2 – Table S1). On the other hand, FleQ is predicted to regulate only *che*Op$_2$ as well as the flagella operon RSP_0052-66, consistent with the previous implication of FleQ in the regulation of these flagella genes [80]. FleQ is also predicted to regulate genes outside the chemotaxis/flagella operons including RSP_0513 (a putative glycoside hydrolase) and RSP_3432 (methyl-accepting chemotaxis protein).

Interestingly none of the 3 TFs in this motility sub-network was predicted to regulate *che*Op$_1$. Previous mutational analysis of the *che*Op$_1$ operon showed its loss had little impact on *R. sphaeroides* motility, unlike *che*Op$_2$ and *che*Op$_3$ [75, 77],which are required for normal chemotatic responses [76-78]. These results might mean that *che*Op$_1$ is not expressed under experimental conditions analyzed thus far and might be under the control of a completely different, as a yet unidentified, regulatory program from the other chemotaxis and flagella genes.

Heat Shock and Oxidative Stress Responses

The ability of an organism to survive in nature is also highly dependent on its ability to mount effective responses to various environmental stresses. Previous analysis of two *R. sphaeroides* σ^{32} paralogs, **RpoH$_I$ (RSP_2410)** and **RpoH$_{II}$ (RSP_0601)**, revealed the overlapping but distinct regulons of these σ factors, indicating a convergence in the transcriptional responses to heat shock and singlet oxygen [81, 82] (Figure 5-9). RpoH$_I$, primarily involved in the cellular response to heat shock, directly regulates a large regulon consisting of ~175 target genes involved in a wide variety of functions ranging from protein folding to fatty acid biosynthesis [81]. Our TRN accurately captures a portion of this regulon, identifying several genes directly involved in heat shock response such as HtpX (RSP_0554), a zinc dependent

protease HtpX (RSP_2649), RSP_0559, Hsp20 (RSP_1572) and a small heat shock protein (RSP_1016), in addition to several other proteins of varying function previously identified as part of the $RpoH_I$ regulon (Additional File 2 – Table S1 (cluster 4)) [81]. In total, 15 experimentally verified $RpoH_I$ target operons were identified in our TRN model. In addition, new putative targets such as *recA* (RSP_0452) and *lolA* (RSP_1497) were also predicted as direct targets for $RpoH_I$.

On the other hand, $RpoH_{II}$ expression is induced in response to the presence of the reactive oxygen species singlet oxygen in *R. sphaeroides* [81]. Previous ChIP-chip and expression analysis identified 144 direct targets of $RpoH_{II}$, 45 of which overlap with $RpoH_I$ targets. Consistent with the previously verified regulon for this σ factor, our TRN predicts $RpoH_{II}$ regulates the expression of genes involved in maintenance of the glutathione pool (glutathione peroxidase (RSP_2389), hydroxyacylglutathione hydrolase (RSP_2294), glutathione S-transferase (RSP_1591)) and oxidative DNA damage repair (oxidoreductases (RSP_2314, RSP_3537), RSP_2388), amongst others (Additional File 2 – Table S1 (cluster 85)). In total 17 verified direct $RpoH_{II}$ targets were identified in our TRN with no new predicted members of this regulon.

While $RpoH_{II}$ regulates transcription of the genes required to amount a stress response to singlet oxygen, **σ^E (RSP_1092)** is the master regulator this response in *R. sphaeroides* [83]. Previous analysis of the σ^E using ChIP-chip and expression analysis, identified the direct targets of σ^E [84]. Our TRN captured a significant portion of the known σ^E regulon including RSP_1088-91, RSP_1092-3, *phrB* (RSP_2143) and RSP_1852. However, other target genes like $RpoH_{II}$, RSP_1409 and *cycA* (RSP_0296) were not predicted in our analysis (Additional File 2 – Table S1 (cluster 90)).

The cellular response to oxidative stress caused by H_2O_2 is modulated by OxyR in many bacteria including *E. coli* and *R. sphaeroides* [85]. Our TRN predicts that *R. sphaeroides* **OxyR (RSP_2780)** regulates its own expression, as well as that of catalase (RSP_2779), a verified direct OxyR target in *R. sphaeroides* [85]. In addition, a ferretin-like protein (RSP_0850) and a TetR family TF (RSP_6137) are

predicted to be *R. sphaeroides* OxyR targets (Additional File 2 – Table S1 (cluster 8)), suggesting that this TF may initiate a transcriptional cascade in this bacterium. OxyR regulation of ferritin has previously been reported in other bacteria [86], where it is proposed to serve as an iron reservoir to prevent Fenton chemistry. While motifs for LysR TFs like OxyR are often difficult to identify, our phylogenetic footprinting analysis allowed us to identify several LysR-type motifs, including a highly conserved motif for OxyR in *R. sphaeroides* (Figure 5-9), highlighting another advantage of utilizing a comparative genomics approach for TRN reconstruction.

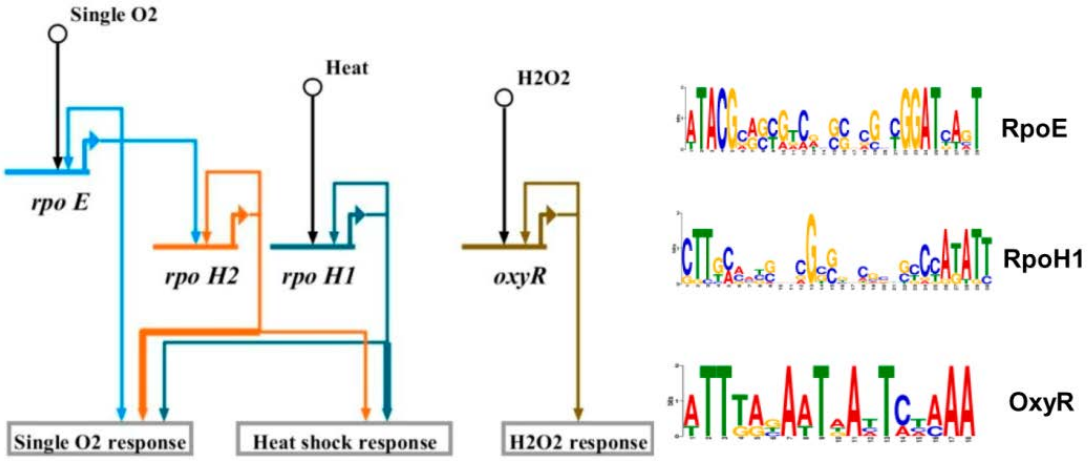

Figure 5-9. Stress response gene regulatory network. Summary of the gene regulatory network controlling singlet O_2, heat and H_2O_2 stress responses in *R. sphaeroides*. RpoH1 and RpoH2 both regulate genes involved in singlet O_2 and heat shock stress response with RpoH1 primarily controlling heat shock related genes and RpoH2 primarily controlling singlet O_2 response related genes (thicker arrows). RpoE is the master regulator of singlet O_2 stress response, while OxyR controls the response to H_2O_2 stress. Motifs identified for RpoE, RpoH1 and OxyR are shown on the right.

Other major cellular sub-networks

Consistent with its role as the housekeeping σ factor, **RpoD (RSP_0395)** is predicted to control a large number of genes within our TRN. Seven different clusters were predicted to be under the control of RpoD, with these clusters being enriched in functions ranging from electron transport to protein synthesis (Figure 5-2, clusters 12, 26, 39, 43, 52, 54, 58 – Table S1), consistent with known global role of RpoD in bacteria. The predicted DNA sequence motifs in all clusters resembled typical -35 -10 motifs bound by RpoD with variations in each motif sufficient to cause them to be split into different clusters. Another σ factor predicted to be involved in the regulation of general processes is **RSP_2681**, a Group IV or ECF family alternative sigma factor. RSP_2681 is predicted to be involved in the regulation of 2 gene clusters that encode proteins involved in protein synthesis and fatty acid biosynthesis (Additional File 2 – Table S1 (clusters 32 and 57)).

In addition to the above sub-networks, several other predictions of our TRN are either consistent with previous knowledge or represent entirely novel predictions which remain to be verified. For instance, **PhoR (RSP_2599)** is predicted regulate the expression of RSP_2601-3 (phosphate ABC transporter), while **RSP_2800** is predicted to regulate genes encoding proteins involved in pyrroloquinoline quinone biosynthesis PqqEDCB (RSP_0790-3)**. RSP_2801** is predicted to regulate transcription of carotenoid biosynthesis genes *crtEF* (RSP_0265-4) and *crtDE* (RSP_0266-7). On the other hand, **RSP_1243** and **RSP_2963** are predicted to be involved in the regulation of genes involved in fatty acid/PHB biosynthesis, while **RSP_2171 (MetR)** is predicted to regulate methionine biosynthesis (Additional File 2 – Table S1).

Links between sub-networks

In addition to the depth and variety of networks captured in our TRN, we also identified several new and interesting links between these sub-networks. For instance, our TRN predicts a previously unrecognized connection between photosynthesis and iron homeostasis in *R. sphaeroides*. The predicted photosynthesis

regulators RSP_2888 and CrpK, as well as the known regulator FnrL, are predicted to regulate several iron/heme-dependent and iron transport genes. Furthermore, FnrL is also predicted to regulate RSP_3341, which we have shown in this work to be directly involved in regulation of other iron-dependent genes. These data suggest that regulation of photosynthesis, which employs several iron-dependent proteins, and iron homeostasis need to be coordinated in *R. sphaeroides* to achieve optimal growth under anaerobic photosynthetic conditions.

NtrC, which is predicted in our TRN to be involved in regulation of nitrogen metabolism, is also predicted to control transcription of genes for central carbon metabolism (Figure 5-2), suggesting a possible previously unrecognized link between carbon and nitrogen metabolism in *R. sphaeroides*. Similar links between carbon and nitrogen metabolism have been identified in *B. subtilis* via the global regulator of carbon metabolism CcpA [87]. Our TRN also captures previously known links between sub-networks controlling the response to heat shock, singlet oxygen stress and DNA repair.

While this description of sub-networks is by no means exhaustive, it provides a useful overview of the various functionalities and connections captured in the TRN. Overall our TRN captures a significant amount of known transcriptional regulatory interactions in *R. sphaeroides*, while including a large number of new novel predictions for this bacterium, which are consistent with analyses in other organisms. Furthermore, our TRN also makes a large number of novel predictions unique to *R. sphaeroides*, which represent high-quality targets for future experimental verification. In sum, given the high predictive ability of our TRN for known TFs, we propose that it provides an excellent roadmap for future analysis of the *R. sphaeroides* TRN and those of related bacteria.

Using an integrative approach for TRN inference

While there are numerous approaches to inferring TRNs, many of which work well in bacteria [3, 6], our approach has the added advantage of leveraging both expression data and sequence information of closely related bacteria. This approach produced a TRN with high predictive capability and information content,

especially given the somewhat limited availability of expression data and prior knowledge about genome-wide transcriptional control in *R. sphaeroides* compared to other organisms for which large-scale TRNs reconstruction has been reported. Below we highlight a few of the crucial features of this inference approach.

An integrated TRN inference approach provides significant improvement in information content

To assess whether our integrated approach provided improved predictive capability over other previously published TRN inference approaches, we compared the our TRN to others built with our gene expression dataset using the direct inference approaches CLR [9] and GENIE3 [10] and a module-base inference approach LeMoNe [88]. Selecting networks of similar size (i.e., the top ~1900 predicted TF-target predictions from each approach), we found that our integrated approach generated a TRN with significantly improved information content (Figure 5-10). Of the 120 clusters identified in our TRN, 80 (~67%) were enriched for at least one GO functional category compared to 34, 35 and 53% for network built with CLR, GENIE and LeMoNe respectively. This comparison suggests our approach captures more functional information. Furthermore, the number of *de novo* detected DNA sequence motifs obtained in our TRN (88 motifs corresponding to ~73% of the clusters), significantly supersedes that obtained by searching the intergenic regions of predict TF targets obtained from CLR, GENIE and LeMoNe analyses (7, 13 and 11 motifs corresponding to 4, 10 and 17% of the clusters respectively) (Figure 5-10). This suggests that these other approaches can group potentially functionally related and co-expressed genes. However, it also suggests that the resulting clusters likely do not include a sufficiently high percentage of co-regulated genes, so the ability to detect conserved promoter motifs from these predicted clusters/regulons is very low. Thus, it appears that initiating our TRN inference with motif detection prior to incorporating expression data significantly improved its information content and allowed us to overcome some of the limitations in our gene expression datasets.

While the regulons of only a handful of TFs have been studied on a genome-scale in *R. sphaeroides*, assessing predictions made for some of these TFs highlights other advantages of our approach. For instance, none of CLR, GENIE or LeMoNe were able to predict targets for PpsR or FnrL, likely due to almost invariant expression profiles of these TFs (Figure 5-5B), as their activities are regulated post-transcriptionally. However, by taking other features of bacterial TFs into consideration, we were able to accurately link PpsR and FnrL to their respective regulons, while making predictions across our network for other similarly regulated TFs. On the other hand, for the alternative sigma factor σ^E whose binding elements are separated by a variable length spacer region and whose regulon might differ considerably across the species used in our comparative genomics analysis, the expression based approaches performed better at identifying members of this regulon. Thus, incorporating consensus predictions for expression-based inference approaches allowed us to capture such predictions in our final TRN.

While the use of comparative genomics significantly improved the predictive capability of our TRN, *de novo* motif detection alone can lead to the identification of an unknown level of false positive binding sites, depending on thresholds, background distributions and the characteristics of the target sequence. Thus, the use of gene expression data to refine these sequence-based clusters significantly improved the predictive accuracy of our TRN model. For instance, we were able to exclude 4 falsely predicted PpsR targets obtained from our sequence analysis by integrating expression data into our analysis (Additional File 1 – Figure S5).

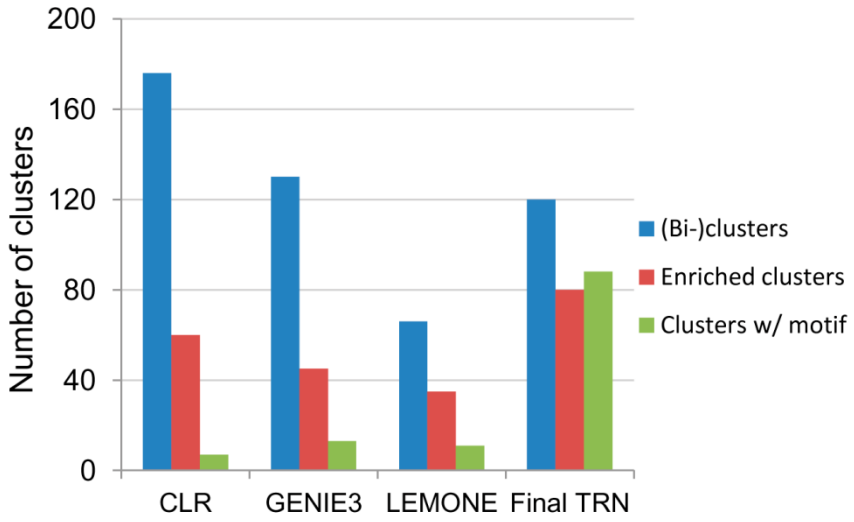

Figure 5-10. Comparison of predictions from our workflow to those from other inference approaches. Bar chart depicts the number of clusters (or regulons) predicted by CLR, GENIE3, LeMoNe and our approach (Final TRN). It also shows the number of these clusters that are significantly enriched for at least one GO functional category ($p < 0.00001$) and the number of these clusters where we could identify a shared conserved motif using the same *de novo* motif detection approach.

Targets of some TFs remain difficult to identify

Though our approach performed relatively well for many TFs, a large number of verified target genes were not identified for some σ^{32} type regulators (i.e., $RpoH_I$ and $RpoH_{II}$). This could possibly be due to difficulties in discriminating DNA binding motifs for closely related σ-factors as well as limitations in available gene expression data. Alternatively, it could be the result of constraints used in *de novo* motif detection. While these constraints performed well at identifying likely binding sites of many traditional TFs, they might be too prohibitive for identification of σ-type motifs.

Preliminary TRNs for other α-Proteobacteria

In addition to inferring a large-scale high information content TRN for *R. sphaeroides*, our comparative genomics analysis enabled us to build and provide preliminary sequence-based TRNs for the community for the 7 other bacteria utilized in our analysis: *R. sphaeroides* ATCC 17025, *Rhodobacter capsulatus* SB 1003, *Roseobacter denitrificans* Och 114, *Dinoroseobacter shibae* DFL 12, *Rhodopseudomonas palustris* CGA009, *Bradyrhizobium japonicum* USDA 110 and *Paracoccus denitrificans* PD1222 (Additional File 2 – Table S15). We expect that these first-generation TRNs will provide insights into the peculiarities of the TRNs of these α-Proteobacteria. They can also serve as starting points for construction of more detailed global TRNs for these and other related bacteria.

Conclusions

In this study, we developed a new workflow to generate genome-scale TRNs, which integrates information in genome sequence and gene expression data, as well as taking into consideration properties of bacterial TFs. We show the utility of this workflow by building a large-scale TRN for the facultative bacterium *R. sphaeroides*. This TRN consists of 120 gene clusters and 1858 regulatory interactions encompassing ~28% of the genes for this organism. Several observations indicated that this approach generates a large-scale TRN with high predictive power. The majority of the predicted gene clusters were enriched for specific functions and the genes found in many of these clusters were consistent with prior knowledge in *R. sphaeroides* or other bacteria. Experimental validation of select *R. sphaeroides* TFs showed that this approach could result in a precision and recall of up to 90% for individual proteins. We provide information that this approach enabled generation of large-scale TRNs with increased information content relative to those built via other approaches. An additional benefit of our approach is that it simultaneously reconstructed TRNs for each of the 7 other related organisms that were used in the comparative genomics analysis. Overall, the workflow presented here represents a powerful approach by which to reconstruct TRNs for bacteria for which similar data types are available. It has also provided a large amount of new insight into transcriptional regulation in a phototroph, correctly capturing many aspects of the diverse lifestyles of *R. sphaeroides*, while providing novel predictions into regulatory networks that await experimental validation. Thus, this large-scale TRN should serve as an indispensible data source for those interested in *R. sphaeroides* and related bacteria.

Materials and Methods

TRN reconstruction

To build a large-scale TRN for *R. sphaeroides*, we utilized an approach that combined comparative genomics, gene expression data bi-clustering and intrinsic properties of bacterial TFs. The workflow used for our reconstruction is detailed below in a stepwise fashion (Additional File 1 – Figure S3).

Selecting Genomes for Phylogenetic Footprinting

Our TRN reconstruction workflow began with exploiting the sequence information from closely related bacteria [13-15]. In order to identify evolutionarily conserved sequences upstream of homologous genes across multiple species (i.e., phylogenetic footprinting), it is important that relatively closely related species are used, as regulatory mechanisms are more likely to be conserved across these organisms [89]. However, if species are too closely related analysis of upstream sequences becomes uninformative, as large stretches of identical or highly similar sequences prevent the identification of relevant regulatory sequences. Thus, species selected for phylogenetic footprinting analysis were carefully chosen to increase the utility of this approach [12]. To select organisms for our analysis, we used a combination of orthology, phylogeny and physiological information. We considered 3 factors in organism selection: (i) the number of orthologs shared between a given organism and our target organism, *R. sphaeroides* (as a larger number of shared orthologs would enable identification of a potentially larger set of regulatory motifs); (ii) phylogenetic distance (as more closely related species would be more likely to have conserved regulatory mechanisms); and (iii) metabolic diversity (in addition to general cellular processes, we considered the regulation of processes peculiar to this group of metabolically diverse organisms, such as photosynthesis). Based on these criteria, we restricted the organisms selected for phylogenetic footprinting to those belonging to the orders Rhodobacterales and Rhizobiales, as these organisms share a larger number of orthologs with *R. sphaeroides* (Additional File 1 – Figure S1), are close phylogenetic relatives to *R. sphaeroides* (Additional File 1 – Figure S1) and are more metabolically diverse than many

members of other α-Proteobacterial orders. From these two orders we selected 8 organisms for our phylogenetic footprinting analysis: *R. sphaeroides* 2.4.1, *R. sphaeroides* ATCC 17025, *R. capsulatus* SB 1003, *R. denitrificans* Och 114, *D. shibae* DFL 12, *R. palustris* CGA009, *B. japonicum* USDA 110 and *P. denitrificans* PD1222. The criteria used for limiting our analysis to 8 organisms is discussed in the section "Identifying phylogenetically conserved motifs". Sequence information for the selected organisms were downloaded from NCBI at the following URLs: *R. sphaeroides* 2.4.1 (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Rhodobacter_sphaeroides_2_4_1_uid57653/), *R. sphaeroides* ATCC 17025 (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Rhodobacter_sphaeroides_ATCC_17025_uid58451/), *R. capsulatus* SB 1003 (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Rhodobacter_capsulatus_SB_1003_uid47509/), *R. denitrificans* Och 114 (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Roseobacter_denitrificans_OCh_114_uid58597/), *D. shibae* DFL 12 (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Dinoroseobacter_shibae_DFL_12_uid58707/), *R. palustris* CGA009 (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Rhodopseudomonas_palustris_CGA009_uid62901/), *B. japonicum* USDA 110 (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Bradyrhizobium_japonicum_USDA_110_uid57599/) and *P. denitrificans* PD1222 (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Paracoccus_denitrificans_PD1222_uid58187/).

Identifying Orthologous Genes between Species

To identify orthologs shared between the selected organisms, we used orthoMCL version 2.0.2 [90]. The blastall function was run with the following parameters: -v 100000 -b 100000 -F 'm S' -m 8 -e $1e^{-5}$. All other functions were run with their default settings. Identified orthologous groups (i.e., all orthologs for a given *R. sphaeroides* 2.4.1 gene across all 8 species) containing paralogs in a subset of species were either

resolved manually by comparing the genomic context of the paralogs to that of the other unique orthologs within the group or, in instances where this could not be resolved by genomic context, all paralogs were removed from the orthologous group. In addition, as our target organism was *R. sphaeroides* 2.4.1, each of the identified orthologous groups was required to have an ortholog in this species. The use of these criteria resulted in the identification of 3387 orthologous groups.

Intergenic regions (IGRs) greater than 40bp in length, upstream of every gene in the genome, were obtained for each of the 8 organisms in our analysis using the regulatory sequence analysis tools website [91]. For each of the 3387 orthologous groups, the IGRs upstream of each gene in the group were then extracted from the appropriate organism, if they existed (genes within operons would generally not contain IGRs of sufficient length). As subsequent motif finding steps would require a sufficient number of sequences to identify meaningful motifs shared by the orthologs, we restricted the orthologous groups carried over to the motif finding step to those having at least 4 IGR sequences (from the possible maximum of 8). A total of 1326 groups of sequences met these criteria and were used for subsequent *de novo* motif detection.

Identifying Phylogenetically Conserved Motifs

These 1326 groups of intergenic sequences upstream of orthologous genes were used as input for *de novo* motif detection. Motif detection was conducted using MEME [92] with the following parameters: -dna -mod zoops -evt 0.01 -nmotifs 3 -maxw 30. A third order background distribution file was generated using all the intergenic sequences from all the organisms selected from this analysis and was used to aid subsequent motif detection. A total of 914 phylogenetically conserved (PC) *de novo* motifs were detected from these sequences, which were represented by position weight matrices (PWMs) (Additional File 1 – Figure S4A). It should be noted that increasing the number of organism used in our phylogenetic footprinting analysis did not significantly increase the number of identified PC motifs (Additional File 1 – Figure S4B). For instance, by the number of identified motifs in relation to the number of species

included in the analysis showed to only an additional 150 motifs were identified by including up to 12 additional closely α-proteobacteria to the analysis. Importantly, these additional motifs represented only new instances of motifs previously identified using the initial set of 8 organisms, which were all recovered in a subsequent genome-wide motif finding step. This analysis also indicated as few as 6 organisms could be sufficient to carry out this analysis, if they possess the appropriate characteristics with respect to the target organism.

Clustering of Identified Motifs

The set of 914 PC motifs identified in the phylogenetic footprinting step will necessarily contain a significant amount of redundancy, as multiple instances of essentially the same motif, corresponding to multiple binding sites of a specific transcription factor (TF), exist in this set. To eliminate this redundancy from the data set, we grouped identical or very similar motifs into clusters based on their similarity. To achieve this, we first conducted a pair-wise comparison of all 914 identified PC motifs using TOMTOM [92, 93], generating q-values as measures of the similarity of these motifs to one another. Only motif pairs with q-values <0.01 were considered as potentially identical motifs and retained for subsequent clustering analysis. We then used MAST [92] to identify all the instances of each of the 914 PC motifs (represented as PWMs) across the *R. sphaeroides* genome. The set of instances identified for a given motif were called "motif groups". We then conducted a pair-wise comparison of all these motif groups to one another. Motif group pairs showing a high degree of overlap (based on identification of the same motif instances across the genome – threshold set to 33%) and for which the parent motif pairs had a q-value <0.01 were considered "identical" and clustered into one group. This approach is based on the idea that multiple variants of the same motif would look similar and thus have a high pair-wise similarity score (i.e., small q-value) and would also be able to identify other instances of the motif across the genome and this shared identity could be used group similar motifs together. These clustered motif groups, thus theoretically contain all the targets for a putative TF within the *R. sphaeroides* genome. The identified target sequences in the *R. sphaeroides* genome were then used to generate *R. sphaeroides* specific PWMs based

on all instances of each motif identified (Additional File 2 – Table S3). Based on these analyses, the 914 PC motifs were clustered into 76 unique motifs based on their similarity. PC motifs which occurred only once in the *R. sphaeroides* genome (i.e., no additional instances of these motifs were found during MAST searches), did not have enough information for generating organism-specific motifs and thus were not included in downstream processing.

Processing of Gene Expression Data

R. sphaeroides represents one of the best studied α-proteobacteria and has served as model system for the study of photosynthesis and several processes. As a result of this prior research, a relatively large number of global gene expression datasets generated using the *R. sphaeroides* Affymetrix geneChip [94] are publically available for data mining. Collecting all of the publicly available microarray datasets from *R. sphaeroides* from the gene expression omnibus (GEO platform GPL162) (totaling 174 microarrays) and combining these with unpublished microarray experiments conducted in our lab (totaling 24 microarrays), we generated a compendium of 198 microarrays encompassing experiments conducted under a variety of conditions (summarized in Additional File 2 – Table S4). All microarray analysis were conducted on the same Affymetrix platform and were normalized together using Robust Multichip Average (RMA) to $\log_2$ scale with background adjustment and quantile normalization [95]. The RMA normalized data were standardized by row normalization.

Identifying Clusters of Co-regulated Genes

Based on the phylogenetic footprinting analysis above, we identified a total 76 clusters of putatively co-regulated genes that shared conserved motifs. While this sequence based network is rich in information content about co-regulated genes and their putative shared cis-acting regulatory sequences, the information content of such networks could be improved by integration of gene expression data, as genes regulated by the same TFs are likely to have similar transcriptional profiles at least under a subset of conditions [8, 12, 27]. Thus, a gene which has a predicted shared motif with the other genes in a cluster

but to does not share a similar transcriptional profile with any other genes in the cluster, over at least a subset of experimental conditions represented in our compendium, could potentially be a false positive prediction and thus potentially be filtered out by using expression data as additional information in the reconstruction process. Furthermore, by utilizing bi-clustering algorithms that allow identification of subsets of conditions under which genes are co-expressed, one can potentially determine under what experimental conditions the genes of different clusters are active, potentially providing an indication of their functional roles and/or signals to which they are responsive [12, 27]. To integrate the data generated from phylogenetic footprinting with our expression datasets, we utilized the data integration frame work DISTILLER [27]. DISTILLER takes in motif information as a binary file indicating whether a particular *de novo* detected motif is present or not. It also takes an expression matrix of normalized expression data across conditions. It then uses an itemset data mining approach to predict what conditions genes sharing a common motif, show correlated expression patterns. We ran DISTILLER on our data set using the following parameters: binary supports: 1, box supports: 30, box pvalues: 0.001, number of randomizations: 100000, size of random modules: 4, minimal module size: 4, number of greedy modules: 200. This analysis resulted in the identification of bi-clusters for 27 of the 76 sequences based clusters (Additional File 2 – Table S1 – Seq + Expr clusters). Genes within these clusters share both a common motif and tight co-expression pattern under a least a subset of conditions in our microarray compendium.

The integration of expression data into our predictions resulted in the removal of a subset of genes from the original sequence-based clusters due to an inability to identify sub-conditions under which they are co-expressed with other members of the cluster. For instance, in the case of target genes predicted for cluster 60, eighteen genes were predicted to be members of this cluster based on our phylogenetic footprinting analysis (Additional File 1 – Figure S5), while only 13 of these genes showed strong co-expression with other members of the operon, under at least a subset of conditions. The genes not showing strong co-expression were thus removed from the cluster. Subsequent experimental analysis of the predicted transcriptional regulator for this bi-cluster (PpsR, see Results) verified that these excluded

genes were likely false positive predictions from our phylogenetic footprinting step. Thus, at least in subset of instances, integration of our gene expression data sets using DISTILLER appeared to improve the overall accuracy of our TRN. While this approach allowed us to refine the predictions of several of our sequence-based clusters, limitations in the diversity of our gene expression compendium meant only 27 of these cluster (~36% of all identified clusters) could be refined via this approach. The remaining sequence-based clusters were retained in our final network, as they were rich in information about putative regulation interactions.

Operon Extension

While phylogenetic footprinting analysis enabled us to identify putative binding sites for TFs within the *R. sphaeroides* genome and thus identify the closest gene(s) to the binding site, other genes within close proximity of this target gene, and potentially in an operon with it, were not captured in the initial analysis. To incorporate *R. sphaeroides* operon structure into our predictions, we combined distance-based operon predictions for *R. sphaeroides* from microbes online [96], with correlation data from our microarray compendium. Genes predicted to be in an operon based on distance and had a Pearson's correlation coefficient of at least 0.8 across the entirety of our microarray compendium, were considered to be in an operon. This information was used to extend to predictions in our TRN to take into account genes that might be in an operon with targets identified via our sequence-based analysis.

Prediction of Transcriptional Regulators for Clusters

Based on information garnered from *R. sphaeroides* genome annotation [31, 32] and complementary Pfam [97] analysis of its proteins sequences, we identified a total of 216 known or predicted TFs (including sigma factors) in the genome (Additional File 2 – Table S5). Haven identified and refined our clusters of co-regulated genes using sequence and gene expression information, we then predicted the most likely of these 216 TFs to regulate each of these clusters. To achieve this we used a combination of

4 criteria based on known properties of bacterial TRNs (Additional File 1 – Figure S6). They consisted of:

1. **Correlation** between a TF and its target genes [3, 6-8]
2. **Proximity** of a TF to the location of its closest binding site in the genome [12, 14, 28, 29].
3. Similarity in DNA motifs bounds by TFs having similar **DNA binding domains** (DBD) [29, 30].
4. **Phylogenetic correlation** of the occurrence of a TF and occurrence of a motif across species [29].

Below we provide more details as to how these features were used and combined into a likelihood score for predicting TFs that regulate identified clusters.

Correlation

Based on previous observations, the expression of many bacterial TFs is known to exhibit some level of correlation to many of their target genes [3, 6, 8, 9]. To use this feature to discriminate between potential transcriptional regulators of each cluster, the average Pearson's correlation coefficient was determined for each TF by determining the correlation between the TF (x) and each gene (g_i) within the cluster (y) containing n genes, across all (or subsets of) conditions from our gene expression data set and averaging the absolute values of these correlations (eqn. 1). This was carried out for all 216 known/predicted TFs in the *R. sphaeroides* genome to determine the average correlation of each TF in relation to each cluster. These average correlation scores were then converted into p-values ($\mathbf{P}_{corr}$) by random permutation. Briefly, 1000 TF-cluster average correlation scores were randomly generated, then each of the previously calculated TF-cluster average correlation scores was compared to the set of randomly generated values. The total number of randomly generated scores greater than or equal to a given TF-cluster average correlation score divided by 1000, was used as an estimate of the p-value (eqn. 2).

$$\text{Ave. corr. } (TF_x, \text{Cluster}_y) = \frac{1}{n}\sum_{i=1}^{n} |\, corr\,(TF_x, Cluster_y(g_i))\,| \quad \textbf{(eqn. 1)}$$

$$\mathbf{P}_{corr}(TF_x, Cluster_y) = \frac{\text{Total no. of random scores} \geq \text{Ave. corr.}(TF_x, Cluster_y)}{1000} \textbf{ (eqn. 2)}$$

Proximity

Many bacterial TFs are known or predicted to bind at their own promoters [28]. Furthermore, a large number of bacterial TFs are also known to bind to targets within close proximity of the gene encoding the TF [12, 28, 29]. To exploit this binding site proximity feature, we determined the closest distance (in number of genes) between each TF's location in the genome and the genes present in a given cluster. Thus, this proximity score would have a value of 0 (if the TF is a member of a cluster for a given TF-cluster pair) or larger. This distance score was determined for every TF-cluster pair where at least one member of the cluster was located on the same replicon (chromosome or plasmid – *R. sphaeroides* 2.4.1 has 2 chromosomes and 5 plasmids) as the TF. These minimum distance scores were also converted into p-values ($\mathbf{P}_{prox}$) by random permutation similar to the procedure described above (eqn. 3).

$$\mathbf{P}_{prox}(TF_x, Cluster_y) = \frac{\text{Total no. of random min. dist.} \leq \text{Min dist.}(TF_x, Cluster_y)}{1000} \textbf{ (eqn. 3)}$$

DNA Binding Domain

Previous analysis of TF binding sites has shown that TFs with similar DBD structures tend to bind similar types of motifs [29, 30]. To exploit this feature in our TRN predictions for *R. sphaeroides*, we began by determining the DBD family to which the 216 TFs in *R. sphaeroides* belong to using Pfam analysis [97]. This allowed us to group 212 of these TFs into 35 families based on the predicted structures of their DBDs (Additional File 2 – Table S6). We subsequently retrieved all the *E. coli* TFs from RegulonDB [98], which had binding motif information (85 in total at the time of this analysis). We also retrieved their position specific scoring matrices (PSSMs). These 85 TFs were then analyzed via Pfam and grouped into 33 families based on the predicted structures of their DBDs (Additional File 2 – Table S7). Only 23 of the 35 families *R. sphaeroides* TFs (190 TFs) had representatives of the same family in the *E. coli* TF data

set. Next we generated PSSMs for each of our 76 *de novo* detected motifs, using the sequences from the predicted binding sites (Additional File 2 – Table S3).

For each TF-cluster pair to be assessed, the PSSM for the *de novo* detected motif of the cluster under consideration was compared to the PSSM(s) from *E. coli* whose associated TF(s) belongs to the same DBD family as the *R. sphaeroides* TF under consideration. The *R. sphaeroides* TF was then assigned the most significant (smallest) q-value from these set of comparisons, which were $-\log_{10}$ transformed to generate the DBD score for that TF-cluster pair. PSSM comparisons were made using TOMTOM [92, 93] and all possible TF-cluster pairs were assessed similarly. These DBD scores were also converted into p-values ($\mathbf{P}_{dbd}$) by random permutation as previously described (eqn. 4).

$$P_{dbd}(TF_x, Cluster_y) = \frac{\text{Total no. of random DBD scores} \geq \text{DBD score}(TF_x, Cluster_y)}{1000} \quad \text{(eqn. 4)}$$

Phylogenetic Correlation

The occurrence of binding sites for a TF across species typically shows a strong correlation with the presence of that TF across the species being studied, as TFs and their target genes generally co-evolve [29, 99]. We used this feature of gene regulatory networks first by determining the occurrence of a given motif across all the genomes used in our analysis. For each *de novo* detected motif we used MAST [92] to search for all instances of that motif in the IGR of each of the 8 species used in our analysis. The p-values for each MAST hit located upstream of a given gene for each genome were stored in vector. The Pearson's correlation coefficient was then calculated between the MAST hits p-value vector for *R. sphaeroides* and that for all the species used in our analysis (*R. sphaeroides* 2.4.1 inclusive). These correlations were referred to as "motif occurrence correlations" (Additional File 1 – Figure S6D shows the motif occurrence correlation for cluster 60). We then determined the occurrence of each TF across the 8 genomes from the orthoMCL ortholog analysis (Additional File 2 – Table S5) . Finally, the Pearson's correlation coefficient between the motif occurrence correlation and TF occurrence was calculated to

determine the phylogenetic correlation. These phylogenetic correlation scores were converted to p-values by random permutation as described above.

$$\mathbf{P}_{pc}(TF_x, Cluster_y) = \frac{\text{Total no. of random Phylo.corr.scores} \geq \text{Phylo.corr.score}\,(TF_x, Cluster_y)}{1000} \textbf{ (eqn. 5)}$$

To provide an estimate the likelihood that a particular TF (x) regulates a given cluster (y), the $-\log_{10}$ of the p-values calculated for the 4 difference criteria were summed together to generate a final score $\mathbf{R}_{score}$ (eqn. 6).

$$\mathbf{R}_{score}(TF_x, Cluster_y) = -\log_{10}(\mathbf{P}_{corr}(TF_x, Cluster_y) * \mathbf{P}_{prox}(TF_x, Cluster_y) * \mathbf{P}_{dbd}(TF_x, Cluster_y) * \mathbf{P}_{pc}(TF_x, Cluster_y)) \text{ (eqn. 6)}$$

Thus the highest possible $\mathbf{R}_{score}$ a TF-cluster pair could attain would be 12. The TF having the highest $\mathbf{R}_{score}$ for each cluster is provided in Additional File 2 – Table S1, while the TFs with the 3 highest $\mathbf{R}_{score}$ values for each identified cluster are provided in Additional File 2 – Table S8. Clusters where the highest scoring TFs had $\mathbf{R}_{score}$s below 4, do not have any TFs assigned to them as these scores were considered to be of low predictive value.

Inferring regulatory interactions solely from expression data

Recent analysis has shown that combining the predictions from a small number of high performing expression-based TRN inference approaches can result in significantly improved accuracy of the predicted network [3]. Thus, to make predictions for TFs not captured in the comparative genomics-based TRN, we employed a combination expression-based TRN inference approaches to try to identify regulatory interactions using only our microarray datasets. For this analysis, we combined the predictions from 3 well-established, high performing direct inference approaches: context likelihood of relatedness (CLR) [9], GENIE3 [10] and a two-way analysis of variance (ANOVA) based approach [11]. As these

approaches have previously been described [3], thus we do not provide any details of implementation or assumptions peculiar to each approach.

Our RMA normalized and row standardized gene expression data from 198 microarray experiments were used as input data for these 3 inference approaches. The list of 216 *R. sphaeroides* TFs was also provided as potential transcriptional regulators. In addition, information on specific deleted or over-expressed genes was provided as additional input for ANOVA. The top 50,000 predicted TF-target interactions from each approach were selected. For each inference approach, the scores of TF-target predictions were converted to p-values by random permutation, generating 10000 random TF-target scores for each approach and comparing actual TF-target scores to this set (as described above). To determine the likelihood that TF i regulates target gene j, the predictions from each of the 3 approaches for that specific interaction were then combined by averaging the –log10 of the p-values from each approach (eqn. 7). In instances, where the no prediction was made for a particular TF-target interaction in any one approach, but was predicted by at least one of the other 2 approaches, a score of 0 was assigned. Potential TF-target interactions not in the top 50,000 of any of the 3 prediction lists were not considered.

$$\mathbf{R}_{exp}(TF_i, target_j) = \frac{1}{3} -\log_{10}(\mathbf{P}_{CLR}(TF_i, target_j) * \mathbf{P}_{GENEI3}(TF_i, target_j) * \mathbf{P}_{ANOVA}(TFi, Targetj) \quad \textbf{(eqn. 7)}.$$

Predicted targets for each TF were then extended to include all genes in a potential operon, as described above, to generate the expression based TRN.

To determine a score threshold to use as a cut-off for interactions to be retained in our expression-based network, we collated all previously generated genome-wide protein-DNA interaction (ChIP) datasets for *R. sphaeroides* and used this to determine the precision and recall of the network relative to an increasing number of interactions included in the network. Genome-wide protein-DNA interaction for FnrL [36], σ^E [84], $RpoH_I$ and $RpoH_{II}$ [81], corresponding to a total of 467 TF-target interactions, were used for this analysis. We used these interactions as our set of true positives (TP). Precision-recall curves were

generated for ranked lists of predictions from CLR, ANOVA, GENEI3 and the combined network (Additional File 1 – Figure S7), with precision and recall calculated at intervals 100 predictions. Typically precision is calculated as:

$$\frac{TPP}{TPP + FPP} = \frac{True\ positive\ predictions}{True\ positive\ predictions + False\ positive\ predictions} = \frac{True\ positive\ predictions}{All\ predictions\ made}$$

(eqn. 8)

where TPP and FPP are assessed based on the number of interactions considered and a gold standard of true positives interactions (TP) [100]. However, due to the small number of TP available for assessment, for each interval of 100 predictions from our TRN, we only considered predictions for TFs for which we had ChIP data and thus we redefined precision as follows:

$$\frac{TPP}{TPP + FPP} = \frac{TPP\ \ for\ TFs\ with\ ChIP\ data}{TPP\ \ for\ TFs\ with\ ChIP\ data + FPP\ \ for\ TFs\ with\ ChIP\ data}$$

(eqn. 9)

$$= \frac{TPP\ \ for\ TFs\ with\ ChIP\ data}{All\ \ predictions\ made\ \ for\ TFs\ with\ ChIP\ data}$$

Recall was calculated as previously described [100]:

$$\frac{TPP}{TP} = \frac{True\ positive\ predictions}{All\ known\ true\ positives}$$ **(eqn. 10)**

We selected a precision cut off of 95%, which corresponded to a recall of 8% and a score cut-off of 1.3 (Additional File 1 – Figure S7). Using this cut off for the entire set of predicted interactions resulted in a total of 1100 predicted TF-target interactions. In this analysis the best performing of the individual approach selected was GENEI3, whose performance was very close to the final composite TRN, though the predictions retained the final composite network differed significantly from the predictions of any one of the individual networks (Additional File 1 – Figure S8), as predictions supported by at least 2 of the 3 approaches received higher scores.

Combining sequence-based and expression-based networks

While the various TRN prediction approaches have their limitations, they could potentially also be complementary as the rely on different data, assumption and algorithms, thus potentially capture different aspects of the regulatory network being studied [6]. To leverage the potential complementarity between our reconstructed sequence- and expression-based networks, we merged predictions from the 2 networks giving precedence to predictions made using the comparative genomics-based approach as many of these predictions were supported by both sequence and expression data. Thus, for TFs for which predictions were made in both our comparative genomics- and expression-based networks, only the predictions from the comparative genomics-base network were retained in our final combined network. Based on this a total of 641 TF-target interactions from our expression-based analysis were retained in the final combined network. This included a total of 44 TFs. The final set of interactions predicted using expression-based approaches is provided in Additional File 2 – Table S9.

Experimental analysis

Details of growth conditions, construction of mutants, microarray and ChIP-seq analyses are provided in Additional file 3.

References

1. Pardee AB: **Multiple molecular levels of cell cycle regulation**. *J Cell Biochem* 1994, **54**(4):375-378.
2. Herrgard MJ, Covert MW, Palsson BO: **Reconstruction of microbial transcriptional regulatory networks**. *Curr Opin Biotechnol* 2004, **15**(1):70-77.
3. Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Stolovitzky G: **Wisdom of crowds for robust gene network inference**. *Nat Methods* 2012, **9**(8):796-804.
4. Stolovitzky G, Monroe D, Califano A: **Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference**. *Ann N Y Acad Sci* 2007, **1115**:1-22.
5. Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, Thorsson V: **The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo**. *Genome Biol* 2006, **7**(5):R36.
6. De Smet R, Marchal K: **Advantages and limitations of current network inference methods**. *Nat Rev Microbiol* 2010, **8**(10):717-729.
7. Joshi A, De Smet R, Marchal K, Van de Peer Y, Michoel T: **Module networks revisited: computational assessment and prioritization of model predictions**. *Bioinformatics* 2009, **25**(4):490-496.
8. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data**. *Nat Genet* 2003, **34**(2):166-176.
9. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS: **Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles**. *PLoS Biol* 2007, **5**(1):e8.
10. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P: **Inferring regulatory networks from expression data using tree-based methods**. *PLoS One* 2010, **5**(9).
11. Kuffner R, Petri T, Tavakkolkhah P, Windhager L, Zimmer R: **Inferring gene regulatory networks by ANOVA**. *Bioinformatics* 2012, **28**(10):1376-1382.
12. Fadda A, Fierro AC, Lemmens K, Monsieurs P, Engelen K, Marchal K: **Inferring the transcriptional network of *Bacillus subtilis***. *Mol Biosyst* 2009, **5**(12):1840-1852.
13. Novichkov PS, Rodionov DA, Stavrovskaya ED, Novichkova ES, Kazakov AE, Gelfand MS, Arkin AP, Mironov AA, Dubchak I: **RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach**. *Nucleic Acids Res* 2010, **38**(Web Server issue):W299-307.
14. Rodionov DA: **Comparative genomic reconstruction of transcriptional regulatory networks in bacteria**. *Chem Rev* 2007, **107**(8):3467-3497.
15. Wang T, Stormo GD: **Identifying the conserved network of cis-regulatory sites of a eukaryotic genome**. *Proc Natl Acad Sci U S A* 2005, **102**(48):17400-17405.
16. Stormo GD, Tan K: **Mining genome databases to identify and understand new gene regulatory systems**. *Curr Opin Microbiol* 2002, **5**(2):149-153.
17. Tabita FR: **The biochemistry and metabolic regulation of carbon metabolism and CO_2-fixation in purple bacteria**. In: *Anoxygenic Photosynthetic Bacteria.* Edited by Blankenship RE, Madigan MT, Bauer CE. The Netherlands: Kluwer Academic Publishers; 1995: 885–914.
18. Imam S, Noguera DR, Donohue TJ: **Global insights into energetic and metabolic networks in *Rhodobacter sphaeroides***. *BMC Syst Biol* 2013, **7**(1):89.
19. Mackenzie C, Eraso JM, Choudhary M, Roh JH, Zeng X, Bruscella P, Puskas A, Kaplan S: **Postgenomic adventures with *Rhodobacter sphaeroides***. *Annu Rev Microbiol* 2007, **61**:283-307.

20. Imam S, Yilmaz S, Sohmen U, Gorzalski AS, Reed JL, Noguera DR, Donohue TJ: **iRsp1095: a genome-scale reconstruction of the *Rhodobacter sphaeroides* metabolic network**. *BMC Syst Biol* 2011, **5**:116.
21. Atsumi S, Higashide W, Liao JC: **Direct photosynthetic recycling of carbon dioxide to isobutyraldehyde**. *Nat Biotechnol* 2009, **27**(12):1177-1180.
22. Khatipov E, Miyake, M., Miyake J. and Y. Asada: **Polyhydroxybutyrate accumulation and hydrogen evolution by *Rhodobacter sphaeroides* as a function of nitrogen availability**. *Biohydrogen* 1999, **III**:157 - 161.
23. Kien NB, Kong IS, Lee MG, Kim JK: **Coenzyme Q10 production in a 150-l reactor by a mutant strain of *Rhodobacter sphaeroides***. *J Ind Microbiol Biotechnol* 2010, **37**(5):521-529.
24. Kontur WS, Ziegelhoffer EC, Spero MA, Imam S, Noguera DR, Donohue TJ: **Pathways involved in reductant distribution during photobiological H(2) production by *Rhodobacter sphaeroides***. *Appl Environ Microbiol* 2011, **77**(20):7425-7429.
25. Wahlund TM, Conway T, Tabita FR: **Bioconversion of CO_2 to ethanol and other compounds**. *American Chemical Society Division of Fuel Chemistry* 1996, **3**:1403-1405.
26. Yilmaz LS, Kontur WS, Sanders AP, Sohmen U, Donohue TJ, Noguera DR: **Electron partitioning during light- and nutrient-powered hydrogen production by *Rhodobacter sphaeroides***. *Bioenerg Res* 2010, **Volume**(1):55 - 66.
27. Lemmens K, De Bie T, Dhollander T, De Keersmaecker SC, Thijs IM, Schoofs G, De Weerdt A, De Moor B, Vanderleyden J, Collado-Vides J *et al*: **DISTILLER: a data integration framework to reveal condition dependency of complex regulons in *Escherichia coli***. *Genome Biol* 2009, **10**(3):R27.
28. Martinez-Antonio A, Collado-Vides J: **Identifying global regulators in transcriptional regulatory networks in bacteria**. *Curr Opin Microbiol* 2003, **6**(5):482-489.
29. Tan K, McCue LA, Stormo GD: **Making connections between novel transcription factors and their DNA motifs**. *Genome Res* 2005, **15**(2):312-320.
30. Sandelin A, Wasserman WW: **Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics**. *J Mol Biol* 2004, **338**(2):207-215.
31. Kontur WS, Schackwitz WS, Ivanova N, Martin J, Labutti K, Deshpande S, Tice HN, Pennacchio C, Sodergren E, Weinstock GM *et al*: **Revised sequence and annotation of the *Rhodobacter sphaeroides* 2.4.1 genome**. *J Bacteriol* 2012, **194**(24):7016-7017.
32. Mackenzie C, Choudhary M, Larimer FW, Predki PF, Stilwagen S, Armitage JP, Barber RD, Donohue TJ, Hosler JP, Newman JE *et al*: **The home stretch, a first analysis of the nearly completed genome of *Rhodobacter sphaeroides* 2.4.1**. *Photosynth Res* 2001, **70**(1):19-41.
33. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.
34. Bruscella P, Eraso JM, Roh JH, Kaplan S: **The use of chromatin immunoprecipitation to define PpsR binding activity in *Rhodobacter sphaeroides* 2.4.1**. *J Bacteriol* 2008, **190**(20):6817-6828.
35. Gomelsky M, Kaplan S: **Genetic evidence that PpsR from *Rhodobacter sphaeroides* 2.4.1 functions as a repressor of puc and bchF expression**. *J Bacteriol* 1995, **177**(6):1634-1637.
36. Dufour YS, Kiley PJ, Donohue TJ: **Reconstruction of the core and extended regulons of global transcription factors**. *PLoS Genet* 2010, **6**(7):e1001027.
37. Zeilstra-Ryalls JH, Kaplan S: **Role of the fnrL gene in photosystem gene expression and photosynthetic growth of *Rhodobacter sphaeroides* 2.4.1**. *J Bacteriol* 1998, **180**(6):1496-1503.
38. Zeilstra-Ryalls JH, Kaplan S: **Aerobic and anaerobic regulation in *Rhodobacter sphaeroides* 2.4.1: the role of the fnrL gene**. *J Bacteriol* 1995, **177**(22):6422-6431.
39. Eraso JM, Kaplan S: **prrA, a putative response regulator involved in oxygen regulation of photosynthesis gene expression in *Rhodobacter sphaeroides***. *J Bacteriol* 1994, **176**(1):32-43.

40. Eraso JM, Roh JH, Zeng X, Callister SJ, Lipton MS, Kaplan S: **Role of the global transcriptional regulator PrrA in *Rhodobacter sphaeroides* 2.4.1: combined transcriptome and proteome analysis**. *J Bacteriol* 2008, **190**(14):4831-4848.
41. Eraso JM, Kaplan S: **Oxygen-insensitive synthesis of the photosynthetic membranes of *Rhodobacter sphaeroides*: a mutant histidine kinase**. *J Bacteriol* 1995, **177**(10):2695-2706.
42. Dangel AW, Tabita FR: **Protein-protein interactions between CbbR and RegA (PrrA), transcriptional regulators of the cbb operons of *Rhodobacter sphaeroides***. *Mol Microbiol* 2009, **71**(3):717-729.
43. Laguri C, Phillips-Jones MK, Williamson MP: **Solution structure and DNA binding of the effector domain from the global regulator PrrA (RegA) from *Rhodobacter sphaeroides*: insights into DNA binding specificity**. *Nucleic Acids Res* 2003, **31**(23):6778-6787.
44. Mank NN, Berghoff BA, Hermanns YN, Klug G: **Regulation of bacterial photosynthesis genes by the small noncoding RNA PcrZ**. *Proc Natl Acad Sci U S A* 2012, **109**(40):16306-16311.
45. Zeilstra-Ryalls J, Gomelsky M, Eraso JM, Yeliseev A, O'Gara J, Kaplan S: **Control of photosystem formation in *Rhodobacter sphaeroides***. *J Bacteriol* 1998, **180**(11):2801-2809.
46. Penfold RJ, Pemberton JM: **Sequencing, chromosomal inactivation, and functional expression in Escherichia coli of ppsR, a gene which represses carotenoid and bacteriochlorophyll synthesis in *Rhodobacter sphaeroides***. *J Bacteriol* 1994, **176**(10):2869-2876.
47. Gomelsky M, Kaplan S: **appA, a novel gene encoding a trans-acting factor involved in the regulation of photosynthesis gene expression in *Rhodobacter sphaeroides* 2.4.1**. *J Bacteriol* 1995, **177**(16):4609-4618.
48. Gomelsky M, Kaplan S: **AppA, a redox regulator of photosystem formation in *Rhodobacter sphaeroides* 2.4.1, is a flavoprotein. Identification of a novel fad binding domain**. *J Biol Chem* 1998, **273**(52):35319-35325.
49. Gomelsky M, Kaplan S: **Molecular genetic analysis suggesting interactions between AppA and PpsR in regulation of photosynthesis gene expression in *Rhodobacter sphaeroides* 2.4.1**. *J Bacteriol* 1997, **179**(1):128-134.
50. Masuda S, Bauer CE: **AppA is a blue light photoreceptor that antirepresses photosynthesis gene expression in *Rhodobacter sphaeroides***. *Cell* 2002, **110**(5):613-623.
51. Ranson-Olson B, Jones DF, Donohue TJ, Zeilstra-Ryalls JH: **In vitro and in vivo analysis of the role of PrrA in *Rhodobacter sphaeroides* 2.4.1 hemA gene expression**. *J Bacteriol* 2006, **188**(9):3208-3218.
52. Botsford JL, Harman JG: **Cyclic AMP in prokaryotes**. *Microbiol Rev* 1992, **56**(1):100-122.
53. Park DM, Akhtar MS, Ansari AZ, Landick R, Kiley PJ: **The bacterial response regulator ArcA uses a diverse binding site architecture to regulate carbon oxidation globally**. *PLoS Genet* 2013, **9**(10):e1003839.
54. Cunningham L, Georgellis D, Green J, Guest JR: **Co-regulation of lipoamide dehydrogenase and 2-oxoglutarate dehydrogenase synthesis in *Escherichia coli*: characterisation of an ArcA binding site in the lpd promoter**. *FEMS Microbiol Lett* 1998, **169**(2):403-408.
55. Saier MH, Jr., Ramseier TM: **The catabolite repressor/activator (Cra) protein of enteric bacteria**. *J Bacteriol* 1996, **178**(12):3411-3417.
56. Leyn SA, Li X, Zheng Q, Novichkov PS, Reed S, Romine MF, Fredrickson JK, Yang C, Osterman AL, Rodionov DA: **Control of proteobacterial central carbon metabolism by the HexR transcriptional regulator: a case study in *Shewanella oneidensis***. *J Biol Chem* 2011, **286**(41):35782-35794.
57. Rodionova IA, Scott DA, Grishin NV, Osterman AL, Rodionov DA: **Tagaturonate-fructuronate epimerase UxaE, a novel enzyme in the hexuronate catabolic network in *Thermotoga maritima***. *Environ Microbiol* 2012, **14**(11):2920-2934.
58. Shimada T, Ishihama A, Busby SJ, Grainger DC: **The *Escherichia coli* RutR transcription factor binds at targets within genes as well as intergenic regions**. *Nucleic Acids Res* 2008, **36**(12):3950-3955.

59. Alber BE: **Biotechnological potential of the ethylmalonyl-CoA pathway**. *Appl Microbiol Biotechnol* 2011, **89**(1):17-25.
60. Perrenoud A, Sauer U: **Impact of global transcriptional regulation by ArcA, ArcB, Cra, Crp, Cya, Fnr, and Mlc on glucose catabolism in *Escherichia coli***. *J Bacteriol* 2005, **187**(9):3171-3179.
61. Blencke HM, Homuth G, Ludwig H, Mader U, Hecker M, Stulke J: **Transcriptional profiling of gene expression in response to glucose in *Bacillus subtilis*: regulation of the central metabolic pathways**. *Metab Eng* 2003, **5**(2):133-149.
62. Goelzer A, Bekkal Brikci F, Martin-Verstraete I, Noirot P, Bessieres P, Aymerich S, Fromion V: **Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of *Bacillus subtilis***. *BMC Syst Biol* 2008, **2**:20.
63. Masepohl B, Hallenbeck PC: **Nitrogen and molybdenum control of nitrogen fixation in the phototrophic bacterium *Rhodobacter capsulatus***. *Adv Exp Med Biol* 2010, **675**:49-70.
64. Colbeau A, Vignais PM: **Use of hupS:lacZ gene fusion to study regulation of hydrogenase expression in *Rhodobacter capsulatus*: stimulation by H_2**. *J Bacteriol* 1992, **174**(13):4258-4264.
65. Dischert W, Vignais PM, Colbeau A: **The synthesis of *Rhodobacter capsulatus* HupSL hydrogenase is regulated by the two-component HupT/HupR system**. *Mol Microbiol* 1999, **34**(5):995-1006.
66. Schwartz CJ, Giel JL, Patschkowski T, Luther C, Ruzicka FJ, Beinert H, Kiley PJ: **IscR, an Fe-S cluster-containing transcription factor, represses expression of *Escherichia coli* genes encoding Fe-S cluster assembly proteins**. *Proc Natl Acad Sci U S A* 2001, **98**(26):14895-14900.
67. Rajagopalan S, Teter SJ, Zwart PH, Brennan RG, Phillips KJ, Kiley PJ: **Studies of IscR reveal a unique mechanism for metal-dependent regulation of DNA binding specificity**. *Nat Struct Mol Biol* 2013, **20**(6):740-747.
68. Yeo WS, Lee JH, Lee KC, Roe JH: **IscR acts as an activator in response to oxidative stress for the suf operon encoding Fe-S assembly proteins**. *Mol Microbiol* 2006, **61**(1):206-218.
69. Todd JD, Wexler M, Sawers G, Yeoman KH, Poole PS, Johnston AW: **RirA, an iron-responsive regulator in the symbiotic bacterium *Rhizobium leguminosarum***. *Microbiology* 2002, **148**(Pt 12):4059-4071.
70. Butala M, Zgur-Bertok D, Busby SJ: **The bacterial LexA transcriptional repressor**. *Cell Mol Life Sci* 2009, **66**(1):82-93.
71. Fernandez de Henestrosa AR, Rivera E, Tapias A, Barbe J: **Identification of the *Rhodobacter sphaeroides* SOS box**. *Mol Microbiol* 1998, **28**(5):991-1003.
72. Tapias A, Campoy S, Barbe J: **Analysis of the expression of the *Rhodobacter sphaeroides* lexA gene**. *Mol Gen Genet* 2000, **263**(6):957-965.
73. Oliveira P, Lindblad P: **LexA, a transcription regulator binding in the promoter region of the bidirectional hydrogenase in the cyanobacterium *Synechocystis sp.* PCC 6803**. *FEMS Microbiol Lett* 2005, **251**(1):59-66.
74. Oliveira P, Lindblad P: **Novel insights into the regulation of LexA in the cyanobacterium *Synechocystis sp.* Strain PCC 6803**. *J Bacteriol* 2011, **193**(15):3804-3814.
75. Armitage JP, Macnab RM: **Unidirectional, intermittent rotation of the flagellum of *Rhodobacter sphaeroides***. *J Bacteriol* 1987, **169**(2):514-518.
76. Porter SL, Warren AV, Martin AC, Armitage JP: **The third chemotaxis locus of *Rhodobacter sphaeroides* is essential for chemotaxis**. *Mol Microbiol* 2002, **46**(4):1081-1094.
77. Ward MJ, Bell AW, Hamblin PA, Packer HL, Armitage JP: **Identification of a chemotaxis operon with two cheY genes in *Rhodobacter sphaeroides***. *Mol Microbiol* 1995, **17**(2):357-366.
78. Hamblin PA, Maguire BA, Grishanin RN, Armitage JP: **Evidence for two chemosensory pathways in *Rhodobacter sphaeroides***. *Mol Microbiol* 1997, **26**(5):1083-1096.

79. Martin AC, Gould M, Byles E, Roberts MA, Armitage JP: **Two chemosensory operons of *Rhodobacter sphaeroides* are regulated independently by sigma 28 and sigma 54**. *J Bacteriol* 2006, **188**(22):7932-7940.
80. Wilkinson DA, Chacko SJ, Venien-Bryan C, Wadhams GH, Armitage JP: **Regulation of flagellum number by FliA and FlgM and role in biofilm formation by *Rhodobacter sphaeroides***. *J Bacteriol* 2011, **193**(15):4010-4014.
81. Dufour YS, Imam S, Koo BM, Green HA, Donohue TJ: **Convergence of the transcriptional responses to heat shock and singlet oxygen stresses**. *PLoS Genet* 2012, **8**(9):e1002929.
82. Nuss AM, Glaeser J, Berghoff BA, Klug G: **Overlapping alternative sigma factor regulons in the response to singlet oxygen in *Rhodobacter sphaeroides***. *J Bacteriol* 2010, **192**(10):2613-2623.
83. Anthony JR, Warczak KL, Donohue TJ: **A transcriptional response to singlet oxygen, a toxic byproduct of photosynthesis**. *Proc Natl Acad Sci U S A* 2005, **102**(18):6502-6507.
84. Dufour YS, Landick R, Donohue TJ: **Organization and evolution of the biological response to singlet oxygen stress**. *J Mol Biol* 2008, **383**(3):713-730.
85. Zeller T, Mraheil MA, Moskvin OV, Li K, Gomelsky M, Klug G: **Regulation of hydrogen peroxide-dependent gene expression in *Rhodobacter sphaeroides*: regulatory functions of OxyR**. *J Bacteriol* 2007, **189**(10):3784-3792.
86. Rocha ER, Smith CJ: **Transcriptional regulation of the *Bacteroides fragilis* ferritin gene (ftnA) by redox stress**. *Microbiology* 2004, **150**(Pt 7):2125-2134.
87. Wacker I, Ludwig H, Reif I, Blencke HM, Detsch C, Stulke J: **The regulatory link between carbon and nitrogen metabolism in *Bacillus subtilis*: regulation of the gltAB operon by the catabolite control protein CcpA**. *Microbiology* 2003, **149**(Pt 10):3001-3009.
88. Michoel T, Maere S, Bonnet E, Joshi A, Saeys Y, Van den Bulcke T, Van Leemput K, van Remortel P, Kuiper M, Marchal K *et al*: **Validating module network learning algorithms using simulated data**. *BMC Bioinformatics* 2007, **8 Suppl 2**:S5.
89. Dermitzakis ET, Clark AG: **Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover**. *Mol Biol Evol* 2002, **19**(7):1114-1121.
90. Li L, Stoeckert CJ, Jr., Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes**. *Genome Res* 2003, **13**(9):2178-2189.
91. Thomas-Chollier M, Defrance M, Medina-Rivera A, Sand O, Herrmann C, Thieffry D, van Helden J: **RSAT 2011: regulatory sequence analysis tools**. *Nucleic Acids Res* 2011, **39**(Web Server issue):W86-91.
92. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS: **MEME SUITE: tools for motif discovery and searching**. *Nucleic Acids Res* 2009, **37**(Web Server issue):W202-208.
93. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS: **Quantifying similarity between motifs**. *Genome Biol* 2007, **8**(2):R24.
94. Pappas CT, Sram J, Moskvin OV, Ivanov PS, Mackenzie RC, Choudhary M, Land ML, Larimer FW, Kaplan S, Gomelsky M: **Construction and validation of the *Rhodobacter sphaeroides* 2.4.1 DNA microarray: transcriptome flexibility at diverse growth modes**. *J Bacteriol* 2004, **186**(14):4748-4758.
95. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185-193.
96. Price MN, Huang KH, Alm EJ, Arkin AP: **A novel method for accurate operon predictions in all sequenced prokaryotes**. *Nucleic Acids Res* 2005, **33**(3):880-892.
97. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL *et al*: **The Pfam protein families database**. *Nucleic Acids Res* 2004, **32**(Database issue):D138-141.

98. Gama-Castro S, Salgado H, Peralta-Gil M, Santos-Zavaleta A, Muniz-Rascado L, Solano-Lira H, Jimenez-Jacinto V, Weiss V, Garcia-Sotelo JS, Lopez-Fuentes A *et al*: **RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units)**. *Nucleic Acids Res* 2011, **39**(Database issue):D98-105.
99. Tan K, Moreno-Hagelsieb G, Collado-Vides J, Stormo GD: **A comparative genomics approach to prediction of new members of regulons**. *Genome Res* 2001, **11**(4):566-584.
100. Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, Xue X, Clarke ND, Altan-Bonnet G, Stolovitzky G: **Towards a rigorous assessment of systems biology models: the DREAM3 challenges**. *PLoS One* 2010, **5**(2):e9202.
101. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome Res* 2003, **13**(11):2498-2504.
102. Longabaugh WJ, Davidson EH, Bolouri H: **Visualization, documentation, analysis, and communication of large-scale gene regulatory networks**. *Biochim Biophys Acta* 2009, **1789**(4):363-374.
103. Homann OR, Johnson AD: **MochiView: versatile software for genome browsing and DNA motif analysis**. *BMC Biol* 2010, **8**:49.

Chapter 6

Global analysis of photosynthesis transcriptional regulatory networks

This chapter is formatted as a manuscript and has been submitted for publication:

Imam S, Noguera DR and Donohue TJ.

I performed all the experiments and analyses in this chapter.

Abstract

Photosynthesis is a crucial biological process that depends on the interplay of many components. This work analyzed the gene targets for 4 transcription factors, FnrL, PrrA, CrpK and RSP_2888, which are known or predicted to control photosynthesis in *Rhodobacter sphaeroides*. Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) identified 52 operons under direct control of FnrL, illustrating its regulatory role in photosynthesis, iron homeostasis, nitrogen metabolism and regulation of sRNA synthesis. Using global gene expression analysis combined with ChIP-seq, we mapped, for the first time, the regulons of PrrA, CrpK and RSP_2888. PrrA regulates ~34 operons encoding mainly photosynthesis and electron transport functions, while CrpK, a previously uncharacterized Crp-family protein, regulates genes involved in photosynthesis and maintenance of iron homeostasis. Furthermore, CrpK and FnrL share similar DNA binding determinants, possibly explaining our observation of the ability of CrpK to partially compensate for the growth defects of a ΔFnrL mutant. We show that the Rrf2 family protein, RSP_2888, plays an important role in photopigment biosynthesis, as part of an incoherent feed-forward loop with PrrA. Our results reveal a previously unrealized, high degree of combinatorial regulation of photosynthetic genes and significant cross-talk between their transcriptional regulators, while illustrating previously unidentified links between photosynthesis and the maintenance of iron homeostasis.

Author Summary

Photosynthetic organisms are the among the most abundant life forms on earth. Their unique ability to harvest solar energy and use it to fix atmospheric carbon dioxide is at the foundation of the global food chain. This paper reports the first comprehensive analysis of networks that control expression of photosynthesis genes using *Rhodobacter sphaeroides*, a microbe that has been studied for decades as a model of solar energy capture and other aspects of the photosynthetic lifestyle. We find a previously unappreciated complexity in the level of control of photosynthetic genes, while identifying new links between photosynthesis and central processes like iron availability. This organism is an ancestor of modern day plants, so our data can inform studies in other photosynthetic organisms and improve our ability to harness solar energy for food and industrial processes.

Introduction

Photosynthetic organisms are central to life on the planet. Their ability to harness solar energy and fix atmospheric carbon dioxide makes them integral parts of most ecosystems. Furthermore, many photosynthetic microbes, either naturally or via modifications, are capable of producing a variety of valuable commodities such as grain for food, hydrocarbons, hydrogen gas and valuable chemicals [1-4]. These properties will likely make them important in efforts to develop more sustainable societies. We are interested in obtaining new knowledge about the transcriptional networks of photosynthetic cells that underlie these important activities.

Anoxygenic photosynthetic bacteria have and continue to provide significant insight into the networks that govern photosynthetic activities because of their ease of growth, genetic tractability, and prior knowledge about solar energy capture and other aspects of this lifestyle [3,5]. The advent of genomic approaches has allowed development of metabolic and transcriptional regulatory network (TRN) models for bacterial photosynthesis, the latter of which has led to predictions about regulatory networks in photosynthetic cells that extend beyond prior knowledge (Imam et al. submitted) [6-9]. Thus, there is likely still much more to be learned about photosynthesis through testing the predictions of metabolic and TRN models in well-studied photosynthetic organisms.

To obtain this new knowledge, we analyze *Rhodobacter sphaeroides*, the best studied member of the purple non-sulfur bacteria – a group of photosynthetic microbes displaying great metabolic versatility and having significant biotechnological potential [1,7,10-17]. *R. sphaeroides* is capable of growing by aerobic respiration, anaerobic respiration and anaerobic anoxygenic photosynthesis. Prior analysis indicates that transitions between aerobic respiratory and anaerobic photosynthetic growth is achieved, in part, via a TRN involving 3 global transcription factors (TFs) – PrrA, FnrL and PpsR – that act to activate or repress relevant operons depending on the presence of oxygen or other signals. For instance, PrrA (the response regulator of the PrrAB two component system) and FnrL (the *R. sphaeroides* homolog of FNR) directly

activate transcription of photosynthesis related genes at low oxygen tensions [9,18-24]. On the other hand, PpsR represses the expression of photosynthesis related genes at high oxygen tensions [8,25,26]. In addition to these TFs, a small non-coding RNA, PcrZ has recently been implicated in the regulation of photosynthesis gene expression in *R. sphaeroides* [27]. While there is considerable information on how these regulators impact some photosynthesis genes, global information on their targets and how they act together to impact this lifestyle are lacking. Furthermore, a large-scale reconstruction of the *R. sphaeroides* TRN (Imam et al. unpublished data) predicts that at least 2 previously uncharacterized TFs, CrpK and RSP_2888, regulate transcription of operons that encode key functions involved in photosynthesis, suggesting that the photosynthetic TRN of this organism is more complex than previously thought.

In this work, we use a combination of genetic, genomic and physiological analysis to dissect the roles of 4 TFs known or predicted to be involved in the regulation of the photosynthetic lifestyle of *R. sphaeroides*. The regulons of the previously characterized TFs, PrrA and FnrL, were refined and extended, while those of CrpK and RSP_2888 were characterized for the first time. Our analysis confirmed many predictions of the large-scale *R. sphaeroides* TRN, revealed the existence of significant overlap in direct targets for these TFs, as well as the high degree of combinational regulation of key operons. We also identified how components in this photosynthetic TRN provide robustness and fine tuned expression of target genes. Overall, this study provides a large amount of new insight into the photosynthetic TRN of *R. sphaeroides* that is likely to be conserved in other related photosynthetic bacteria.

Results

Genome-wide analysis of known regulators of photosynthesis in *R. sphaeroides*

Based on previous analysis in *R. sphaeroides* and related purple non-sulfur bacteria, FnrL, PrrA and PpsR have been identified as key regulators of the photosynthetic lifestyle [8,9,19,23,26,28]. We have previously characterized the genome-wide binding sites of PpsR via chromatin immunoprecipitation followed by sequencing (ChIP-seq) (Imam et al. unpublished data). Here, we analyze the regulons of FnrL and PrrA using both ChIP-seq and global gene expression analysis.

FnrL – a global regulator of anaerobic growth in R. sphaeroides

FnrL is an iron-sulfur cluster-containing Crp-family TF that has been reported to be essential for both photosynthetic and anaerobic respiratory growth in *R. sphaeroides* [9,24]. ChIP-chip analysis has previously been used to map genome-wide FnrL binding sites *in vivo*, identifying targets that indicate the direct involvement of this protein in a host of processes including those required for photosynthetic and anaerobic respiratory growth [23]. However, a large-scale reconstruction of *R. sphaeroides* TRN predicted that the FnrL regulon is significantly larger than previous analyses suggested. Thus, we re-examined the FnrL regulon using new complementary genomic approaches.

Analyzing the FnrL regulon using ChIP-seq

We determined the genome-wide FnrL binding sites using ChIP-seq with wild type (WT) cells grown under anoxygenic photosynthetic conditions. We reproducibly identified a total of 62 FnrL binding sites across 3 independent ChIP-seq experiments, corresponding to 52 known or predicted operons (Table 6-1). These included several sites immediately upstream of genes involved in bacteriochlorophyll synthesis (*bchEJGP*), early steps in tetrapyrrole biosynthesis (*hemN*, *hemZ* and *hemA*), as well as genes that regulate anaerobic respiration using DMSO as a terminal electron acceptor (*dorS*) (Figure 6-1A, Table 6-1). FnrL binding sites were also found upstream of genes encoding functions for iron transporter

(*feoABC*) and iron sulfur cluster assembly (RSP_1949). When we compared this set of ChIP-seq identified FnrL binding sites to data from ChIP-chip analysis [23], 24 of the 27 FnrL binding sites identified previously were also detected in our analysis (Table 6-1). The three previously identified FnrL binding sites not identified in our ChIP-seq analysis do not appear to contain a significant FnrL motif and likely represent false binding events. Furthermore, we found an additional 38 FnrL binding sites in the ChIP-seq dataset, implicating this TF as a direct regulator of a wide variety of new functions, ranging from protein synthesis and substrate transport to polyphosphate dependent phosphorylation and nitrogen metabolism (Table 6-1).

In addition to the 49 FnrL binding sites that were identified upstream of operons, 13 binding sites were outside upstream regulatory regions of any annotated genes. These sites could represent non-functional sites or binding sites in the upstream regulatory regions of other unannotated genomic elements. For instance, 2 of these 13 additional sites were in putative promoter regions of recently identified sRNA in *R. sphaeroides* - RSs0019 and RSs2461 [29]. Thus, it is conceivable that these other unassigned binding sites are located in the upstream regulatory regions of other as of yet unidentified genomic elements. It is also worth noting that 41 of 59 (69.5%) FnrL target operons predicted in the large-scale TRN reconstruction (Imam et al. unpublished data) were verified via this ChIP-seq analysis (Figure 6-1C), including 17 operons that were novel predictions in that TRN inference study (Table 6-1).

To independently assess the functional role of FnrL in the regulation of target genes identified in our ChIP-seq analysis, we conducted microarray analysis to compare the gene expression of WT cells to a *ΔfnrL* deletion strain [9,24] during growth on acetate as the sole carbon source, a condition we found that allows photosynthetic growth of both WT and *ΔfnrL* strains (Figure S1). Consistent with FnrL being a global regulator, a total of 303 genes were DE between the 2 strains (cutoff – 1.5 fold change (FC), $p < 0.01$), with 166 and 137 genes showing increased and decreased transcript abundance, respectively, between WT and *ΔfnrL* cells (Table S1). Of the 48 operons in which we found FnrL binding via ChIP-seq (and for which probes exist on the *R. sphaeroides* gene chip), 24 were DE between WT and *ΔfnrL* cells

(Figure 6-1B), indicating that a least under the conditions assayed, the expression of these target genes is significantly influenced by FnrL. While the change in expression of some of the FnrL targets did not meet the significance cut-off used for this analysis (Table S1), changes in their expression in response to the loss of FnrL was sufficient to allow a tentative assignment of their control by FnrL (Table 6-1, Figure 6-1B). It should also be noted that transcripts for *hemA* and *nuoA-N* (RSP_0100-12) did not show a significant difference in levels between WT and *ΔfnrL*, while transcripts encoding *bchEJGP* were elevated in *ΔfnrL* cells, despite the proposed positive role of FnrL in transcription of these operons [23]. These results suggest that transcription of some FnrL target operons might be under control of other TFs, wherein loss of FnrL is partially or fully compensated for by the activity of other transcriptional regulators in the *ΔfnrL* strain. Indeed, *hemA* transcription is known to also be directly activated by the response regulator, PrrA under anaerobic conditions [30,31].

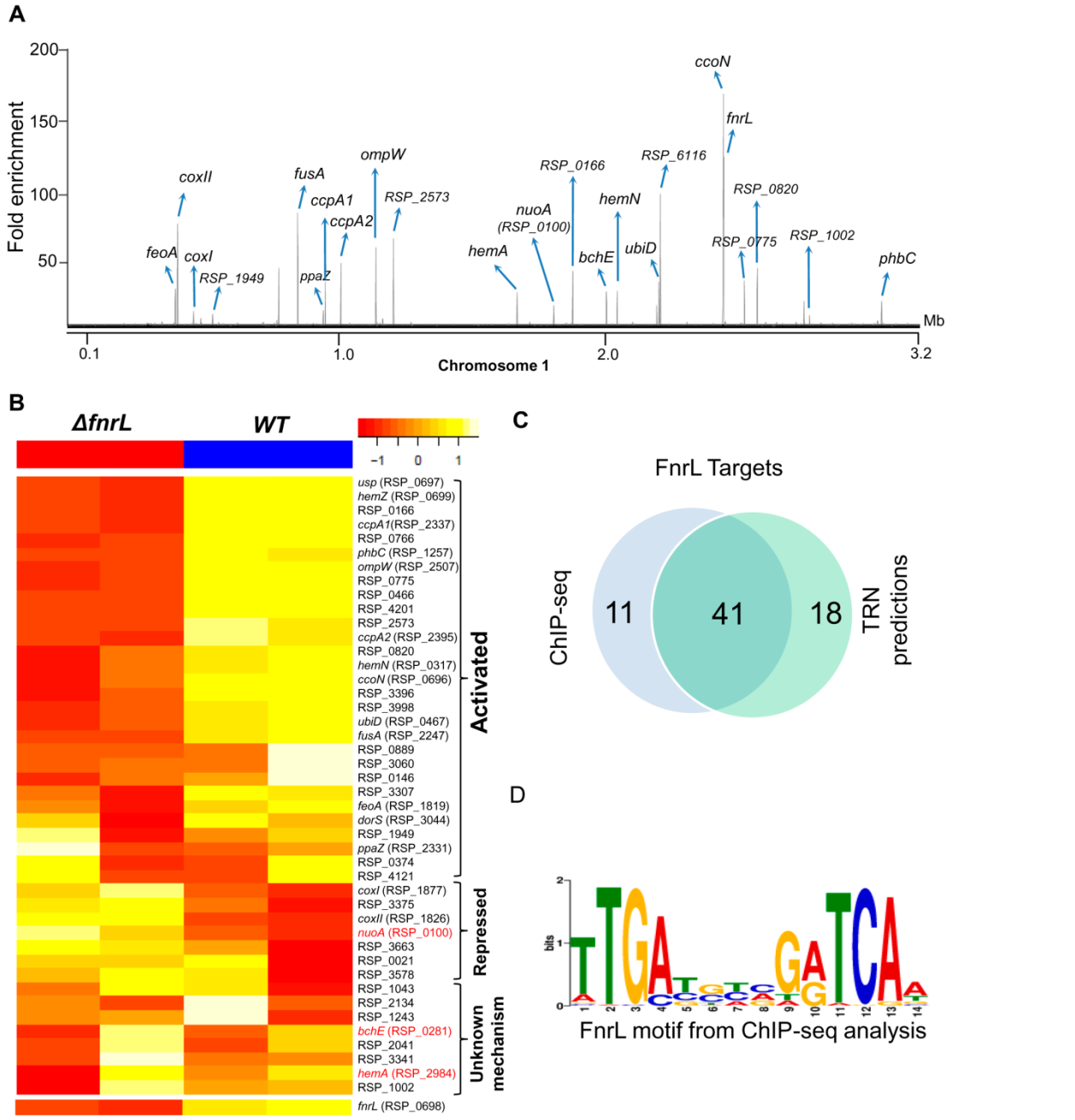

Figure 6-1. Analysis of the FnrL regulon in *R. sphaeroides*. (A) A total of 62 FnrL binding sites were identified by ChIP-seq across the *R. sphaeroides* genome. Binding sites across chromosome 1 are highlighted. MochiView [58] was used for visualization of binding profile. (B) Heat map depicts the differential expression FnrL target operons between wild type (WT) and *ΔfnrL* cells grown on acetate-based media. For brevity only the first members of the target operons are presented. The relative expression of *nuoA*, *bchE* and *hemA* (highlighted in red), which are known to be positively regulated by FnrL, are either not DE or DE in the opposite direction. (C) Venn diagram depicting the overlap between our FnrL ChIP-seq analysis and prediction from the large-scale reconstruction of *R. sphaeroides* transcriptional network. (D) Position weight matrix logo generated for FnrL using targets identified by ChIP-seq.

Table 6-1. FnrL target genes identified by ChIP-seq analysis of *R. sphaeroides* cells grown photosynthetically

S/N†	Gene[a]	Annotation	ChrID	Peak start	Peak stop	FC[b]	Motif start	Reg[c]
1	RSP_0021[d]	30S ribosomal protein S9	chr1	1722800	1723199	6.4	1722853 1722997 1723162	-
2	RSP_0100-12*	*nuoA-N*	chr1	1811600	1812599	34.1	1812207	+
3	RSP_0146-7[d]	*glnB-glnA1*	chr1	1859600	1859999	3.0	1859730	+
4	RSP_0166*	TraR-like protein	chr1	1881800	1882599	78.1	1882413	+
5	RSP_0281-76*	*bchEJGP*	chr1	2007400	2008199	49.3	2007816	+
6	RSP_0317*	*hemN*	chr1	2047000	2047599	51.0	2047244	+
7	RSP_0374-8[d]	ABC basic amino acid transporter	chr1	2103800	2104399	4.6	2104074	+
8	RSP_0466-4*	Hypothetical protein	chr1	2201200	2201999	61.2	2201632	+
	RSP_0467-8	*ubiD*						+
9	RSP_0695-3	*ccoOQP*	chr1	2438800	2439399	41.7	2439129	+
10	RSP_0696-3*	*ccoNOQP*	chr1	2440000	2440799	410.9	2440417	+
	RSP_0697*	Universal stress protein (Usp)					2440385	+
11	RSP_0698*	*fnrL*	chr1	2441600	2442599	453.4	2442182	-
	RSP_0699	*hemZ*						+
12	RSP_0766-5[d]	polyphosphate kinase 2	chr1	2509000	2509399	3.4	2509195	+
13	RSP_0775*	Class I monoheme cytochrome c	chr1	2518400	2518999	61.3	2518696	+
14	RSP_0820-RSP_6134*	cytochrome B561, hypothetical protein	chr1	2565800	2566399	85.7	2566094	+
15	RSP_0889-8[d]	*glnK-amtB*	chr1	2637800	2638399	5.1	2638117	+
16	RSP_1002-3[d]	Aspartate carbamoyltransferase	chr1	2760000	2760599	14.3	2760402	
17	RSP_1043[d]	50S ribosomal protein L31	chr1	2804600	2804799	2.8	2804780	
18	RSP_1243-2	Transcriptional regulator LacI family	chr1	3012800	3013199	6.3	3013043	+
19	RSP_1257-4*	Polyhydroxyalkanoic synthase	chr1	3026600	3027399	31.5	3027028	+
20	RSP_1266-9	Potassium-transporting P-type ATPase	chr1	3032600	3033199	3.0	3032852	
21	RSP_1775	Hypothetical protein	chr1	359600	360399	5.2	360109	
22	RSP_1819-7*	*feoABC*	chr1	408800	409399	50.3	409223	+
23	RSP_1826-9*	*coxII-X-XI-III*	chr1	417600	418199	146.8	417975	-
24	RSP_1877-6*	*coxI*	chr1	476800	477399	17.7	477119	-
25	RSP_1949[d]	FeS assembly SUF system protein	chr1	547000	547399	15.3	547144	+
26	RSP_2041[d]	Hypothetical protein	chr1	635400	635799	2.4	635560	
27	RSP_2134-6[d]	Cytochrome c-type biogenesis protein	chr1	735400	735799	5.1	735526 735575	
28	RSP_2247*	*fusA*	chr1	862400	863199	176.2	862812 862579	+
29	RSP_2331[d]	*ppaZ*	chr1	957000	957599	22.2	957181	+
30	RSP_2337*	*ccpA1*	chr1	964200	964999	69.8	964492	+

31	RSP_2395*	*ccpA2*	chr1	1022200	1022799	96.2	1022541	+
32	RSP_2507*	*ompW*	chr1	1152400	1152999	135.3	1152640	+
33	RSP_2573*	Hypothetical protein	chr1	1217400	1218399	127.0	1217769	+
34	RSP_2984*	*hemA*	chr1	1675800	1676399	51.5	1676046	+
35	RSP_3044*	*dorS*	chr2	77800	78399	63.5	78069	+
36	RSP_3060-3[d]	Possible O-acetylserine synthase	chr2	100800	101399	22.2	101055	-
37	RSP_3307[d]	Hypothetical protein	chr2	367600	367999	13.5	367810	+
38	RSP_3341[d]	BadM/Rrf2 family transcription factor	chr2	403800	404199	6.8	403914	+
39	RSP_3375-3[d]	ABC efflux pump	chr2	440800	441199	5.3	440985	-
40	RSP_3396-3	ABC opine/polyamine transporter	chr2	461000	461399	3.8	461144	+
41	RSP_3578	Putative gas vesicle synthesis protein	chr2	674000	674399	5.5	674158	-
42	RSP_3663-1[d]	TRAP-T family transporter	chr2	771600	771999	2.7	771762	-
43	RSP_3998[d]	Glycine betaine transporter	plasmidB	49600	50199	7.9	49888	+
44	RSP_4121 RSP_4122	Hypothetical proteins	plasmidC	87800	88399	4.2	88138	+
45	RSP_4201-4*	Transcriptional regulator ArsR family	plasmidD	51400	51999	100.3	51739	+
46	RSP_6024	Hypothetical protein	chr1	486800	486999	2.9	486945	NA
47	RSP_6116*	Hypothetical protein	chr1	2206400	2207199	201.5	2206759	NA
48	RSP_6142	Hypothetical protein	chr1	2739400	2739999	35.1	2739722	NA
49	RSP_7211	RepB partitioning protein	plasmidC	71800	72399	20.4	72065	NA
50	RSs2461	small RNA RSs2461	chr1	2598000	2598399	4.5	2598237	NA
51	RSs0019	small RNA RSs0019	chr2	28600	28999	2.6	28843	NA
52			chr1	1177800	1178199	9.0	1178055	NA
53			chr1	1282200	1282599	4.0	1282408	NA
54	*		chr1	2193200	2193799	29.6	2193494	NA
55			chr1	403800	404199	4.1	403983	NA
56	*		chr1	792200	792799	84.4	792528	NA
57			chr2	12800	13199	14.2	12987	NA
58			chr1	2306000	2306399	3.9	2306212	NA
59			chr1	2710800	2711199	5.0	2710949	NA
60			chr1	731200	731599	3.6	731382	NA
61			chr1	785400	785999	6.0	785643	NA
62			chr2	95200	95599	4.7	95451	NA

* Previously identified by ChIP-chip analysis (Dufour et al. 2010).
[a] Instances where the FnrL binding site could potentially regulate two operons is separated by a forward slash
[b] Fold enrichment of the wild type ChIP sample over the ΔfnrL ChIP control.
[c] Likely regulatory role of FnrL on these target operons based on change in gene expression between WT and ΔfnrL cells. + = positively regulated by FnrL. - = negatively regulated by FnrL. NA - Not applicable (i.e., these genes are not represented on the R. sphaeroides) Affymetrix gene chip).
[d] New FnrL targets predicted in TRN and verified via ChIP-seq analysis. All ChIP-chip identified sites except RSP_6116 were also identified in the TRN
† Some binding sites correspond to more than one operon.

PrrA – a global regulator of photosynthetic growth in R. sphaeroides

PrrA is the response regulator of the two component PrrAB system that has previously been proposed to be a major global TF in *R. sphaeroides* and related purple non-sulfur bacteria [19]. PrrA is essential for photosynthetic growth in *R. sphaeroides* and direct control of photosynthesis-related operons by PrrA has be shown via the use of *in vitro* experiments [21,30]. To obtain a better understanding of the functional role of PrrA, we assessed PrrA activity using genome-wide gene expression data and ChIP-seq.

Redefining the set of PrrA target genes

Previous gene expression profiling experiments comparing mRNA abundance in a *ΔprrA* strain to WT cells under anaerobic respiratory conditions, showed that over 1000 genes were DE in the absence of PrrA [19]. However, a large percentage of these genes encoded functions related to protein synthesis and cell growth [19], suggesting that these might also reflect differences in growth rates between the 2 strains, possibly resulting from unlinked mutations in the *ΔprrA* strain. Consistent with this, we found that the *ΔprrA* strain used in the previous analysis (PrrA2) grew significantly faster than our WT strain under anaerobic respiratory conditions (Figure S2A). In contrast, an independently constructed markerless *ΔprrA* strain made for this study (PrrA3) grew similarly to WT under anaerobic respiratory conditions (Figure S2) and showed similar pigmentation phenotypes to the original *ΔprrA* mutant. Consequently, we reassessed differences in gene expression between WT and *ΔprrA* using the PrrA3 mutant strain. We found a total of 255 genes were DE between WT and *ΔprrA* (2 FC, $p<0.01$) (Table S2), significantly less than the 1058 previously reported at a similar cut off [19]. In addition, this set of 255 DE genes did not include any protein synthesis genes. We believe this set of DE genes, which are essentially a subset of those previously identified [19], provide a better picture of potential PrrA target genes.

Consistent with previous knowledge on PrrA, the 255 DE genes that we identified were enriched for genes known or predicted to be involved in photosynthetic processes (Table 6-2). In addition to photosynthesis related functions, other GO terms significantly enriched for DE genes in this data set

include categories such as the TCA cycle, electron transport chain and iron binding (Table 6-2). Overall, of the 255 DE genes (corresponding to 182 operons), mRNA levels from 148 were increased in the presence of PrrA, while 107 were decreased, supporting previous suggestions that PrrA functions as both a transcriptional activator and repressor [19].

Table 6-2. GO functional categories significantly enriched for genes regulated by PrrA

GO ID	GO description	Number of genes in set	DE genes[a]	P-value[b]	Reg[c]
GO:0015979	Photosynthesis	34	20	0	+
GO:0022900	Electron transport chain	33	16	3.51E-13	+, -
GO:0006099	Tricarboxylic acid (TCA) cycle	11	6	6.01E-07	-
GO:0008299	Isoprenoid biosynthetic process	12	6	1.37E-06	+
GO:0018189	Pyrroloquinoline quinone biosynthetic process	4	3	1.17E-05	-
GO:0004129	Cytochrome-c oxidase activity	8	4	3.26E-05	-
GO:0033014	Tetrapyrrole biosynthetic process	5	3	5.56E-05	+
GO:0005506	Iron ion binding	52	10	1.62E-04	+, -
GO:0004497	Monooxygenase activity	17	4	2.32E-03	+, -

[a] The 255 DE genes obtained using a 2 FC cut off were utilized for this analysis.
[b] p-value based on the hypergeometric distribution
[c] Regulatory role PrrA on DE genes within the gene sets. + = positive regulation; - = negative regulation; +,- = some genes upregulated while others are downregulate in the gene set.

PrrA regulates transcription from only a subset of its binding sites

To determine which DE genes are directly regulated by PrrA, we conducted ChIP-seq analysis with a 3X myc-tagged PrrA protein that complements the photosynthetic growth defect of PrrA3 (Figure S2B). We observed significant enrichment for PrrA at ~140 sites across the *R. sphaeroides* genome (Figure 6-2A). Analysis of the sequences under all of these peaks did not reveal any strong consensus sequence shared by a significant number of these sites, suggesting, as has recently been found for TFs such as ArcA and IscR [32,33], that some of these peaks may not reflect sequence-specific DNA binding by PrrA. Thus, to help determine the transcriptionally regulated direct targets of PrrA, only operons with both a significant peak and which were DE in PrrA3, were considered as candidate direct targets of this TF. A total of 34 operons met these criteria, including 18 photosynthesis related operons (Table 6-3, Figures 6-2B and C). In addition to photosynthesis-related genes to which PrrA had previously been linked, these analyses indicate that PrrA is also a direct regulator of electron transport (regulating operons encoding *fbcFBC*, *fbcQ-soxDA* and RSP_0820 (cytochrome B561)), tetrapyrrole synthesis (*hemA*, *hemC* and *hemE*) and terpenoid backbone biosynthesis (*dxr*). Our data predicts that all but one of these operons are positively regulated by PrrA since RNA levels from these genes were lower in PrrA3 (Figure 6-2B, Table 6-3). Other enriched sites in the genome not included in this set of transcriptionally regulated direct PrrA targets are provided in Table S3.

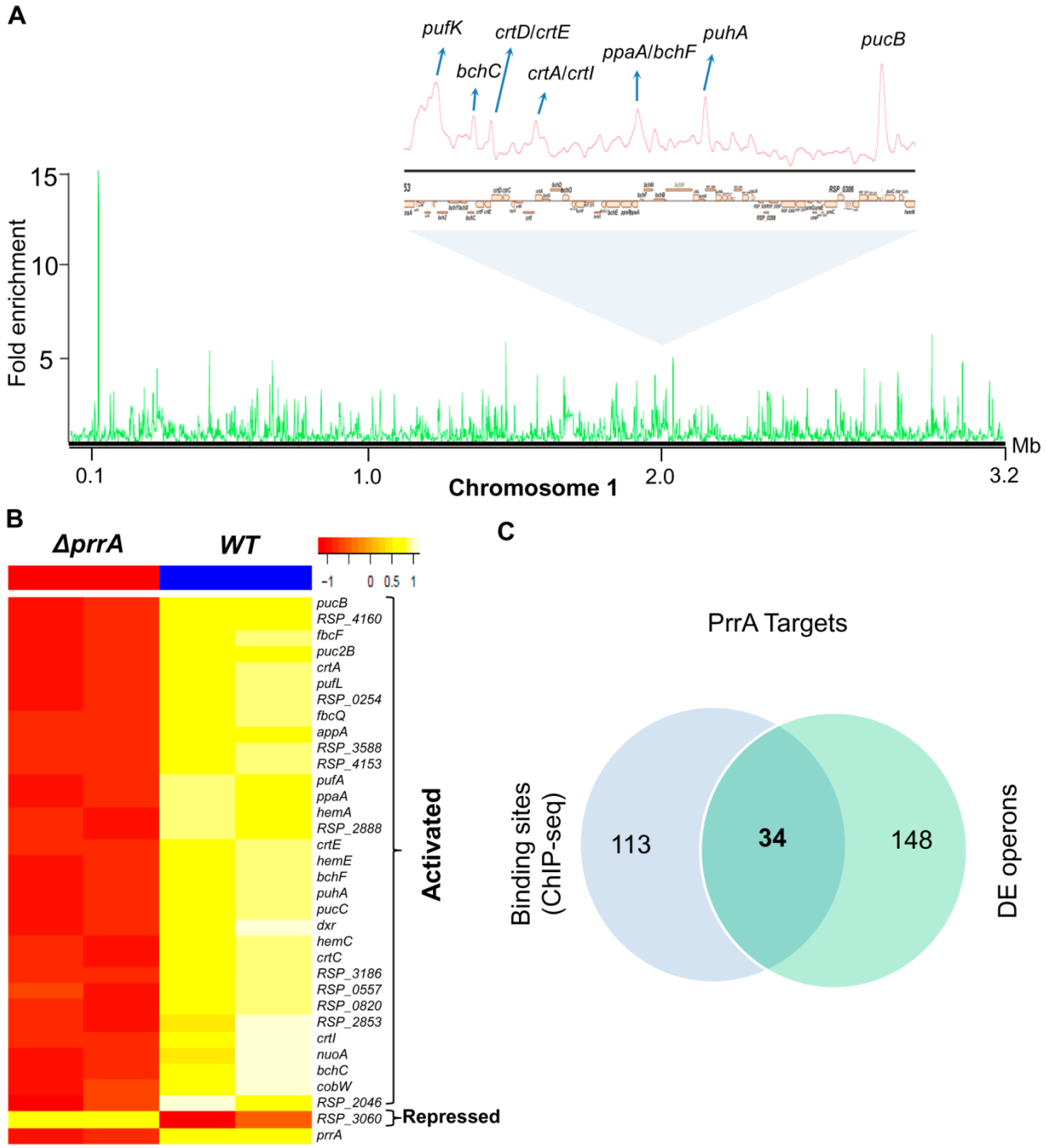

Figure 6-2. Analysis of the PrrA regulon in *R. sphaeroides*. (A) PrrA binding sites across chromosome 1. Binding sites within the photosynthetic gene cluster are enlarged. (B) Heat map depicting the PrrA targets genes from a pair-wise comparison of transcript levels from WT and *ΔprrA* cells grown under anaerobic respiratory conditions. (C) Venn diagram showing the overlap between identified ChIP-seq binding sites for PrrA and DE operons from microarray analysis. It should be noted that some binding sites are located between two divergently transcribed operons, which were both DE. In such cases, both operons were considered as direct PrrA targets.

Table 6-3. PrrA target genes identified by ChIP-seq and gene expression analysis of *R. sphaeroides* cells.

S/N†	GeneID	Gene Annotation*	chrID	Start[a]	Stop[a]	FC[b]	Reg[c]
1	RSP_0100-12	*nuoA-N* (NADH dehydrogenase)	chr1	1811600	1812599	2	+
	RSP_0254	*dxsA*	chr1	1980200	1983799	2.9	+
2	RSP_0257-5	*pufLMX (*Photosynthetic reaction center proteins)					+
	RSP_0258-RSP_6109	*pufKAB* (LHI alpha, Light-harvesting B875 protein)					+
3	RSP_0263-59	*bchCXYZ-pufQ* (Chlorophyll synthesis)	chr1	1988000	1988599	3	+
4	RSP_0265-4	*crtEF* (Carotenoid biosynthesis)	chr1	1990400	1990799	2.87	+
	RSP_0266-7	*crtCD* (Carotenoid biosynthesis)					+
5	RSP_0271-69	*crtIB-tspO* (Carotenoid biosynthesis)	chr1	1996400	1996799	2.57	+
	RSP_0272-5	*crtA-bchIDO* (Carotenoid biosynthesis)					+
6	RSP_0283	*ppaA* (Regulatory protein)	chr1	2010000	2010799	2.96	+
	RSP_0284-91	*bchFNBHLM-puhA* (Bacteriohlorophyll synthesis)					+
7	RSP_0290-1	*puhA* (Light-harvesting 1 (B870) complex assembly)	chr1	2019000	2019599	3.7	+
8	RSP_0314-RSP_6256	*pucBA* (LHII beta, light-harvesting B800/850 protein)	chr1	2042400	2043399	5.35	+
	RSP_0315	*pucC* (Light-harvesting 1 (B870) complex assembly)					+
9	RSP_0557	Hypothetical protein	chr1	2294400	2294599	2.2	+
10	RSP_0679	*hemC* (Hydroxymethylbilane synthase)	chr1	2424000	2424399	2.1	+
	RSP_0680	*hemE* (Uroporphyrinogen decarboxylase)					+
11	RSP_0820-RSP_6134	Cytochrome B561	chr1	2565800	2566199	2	+
12	RSP_1396-4	*fbcFBC* (Ubiquinol-cytochrome C reductase)	chr1	3174800	3175199	2.85	+
13	RSP_1518	*prrA* (Response regulator)	chr1	104400	105399	15.9	NA
14	RSP_1556-RSP_6158	*puc2B2A* (Light-harvesting complex)	chr1	146000	146799	3.4	+
15	RSP_1565	*appA* (Antirepressor of PpsR)	chr1	156200	156999	3.9	+
16	RSP_2046	Hypothetical protein	chr1	640600	642399	3.26	+
17	RSP_2687-90	*fbcQ-soxDA* (Cytochrome b-c1 subunit IV)	chr1	1332200	1332999	3.1	+
18	RSP_2709-10	*dxr*	chr1	1357000	1357599	2.38	+
19	RSP_2829-24	*cobWNHIJK* (Cobalamin synthesis proteins)	chr1	1481600	1482199	2.7	+
20	RSP_2853-55	Transcriptional regulator	chr1	1451200	1451599	2.74	+
21	RSP_2888	Transcriptional regulator	chr1	1565000	1565199	1.7	+
22	RSP_2984	*hemA* (5-aminolevulinate synthase)	chr1	1674200	1676199	4.49	+
23	RSP_3060-3	Possible O-acetylserine synthase	chr2	101000	101399	2.27	-
24	RSP_3186	Hypothetical protein	chr2	239200	239999	3.49	+
25	RSP_3588-5	Hypothetical protein	chr2	680600	681599	4.74	+
26	RSP_4153-RSP_7385	Hypothetical protein	plasmidD	13600	14799	5	+
27	RSP_4160-2	Hypothetical protein	plasmidD	19600	20199	3.67	+

[a] Chromosomal locations of start and stop of ChIP-seq peaks.
[b] Fold enrichment of PrrA-myc ChIP over control myc antibody ChIP in WT control.
[c] Regulatory role of PrrA on target operons based on change in gene expression between WT and ΔprrA cells. + = positively regulated by PrrA. - = negatively regulated by PrrA. NA - Not applicable.
† Number of binding sites. Some binding sites correspond to more than one operon.
* *dxsA* - 1-deoxy-D-xylulose-5-phosphate synthase; *dxr* - 1-deoxy-D-xylulose 5-phosphate reductoisomerase.

Analysis of newly predicted regulators of photosynthesis in *R. sphaeroides*

A recent large-scale reconstruction of the TRN of *R. sphaeroides* predicted that there are additional regulators of photosynthesis (Imam et al. unpublished data). Among the highest scoring TFs that fell into this category were: (i) CrpK (RSP_2572), a Crp/Fnr-family regulator, and (ii) RSP_2888, a BadM/Rrf2-family protein. Using a combination of physiological, genetic and genomic analysis, we investigated the contributions made by these proteins to regulation of photosynthesis in *R. sphaeroides*.

CrpK – a member of the Crp/FnrL family that controls many photosynthesis genes

CrpK is a Crp/Fnr-family TF, which based on Pfam analysis [34], shares similar cyclic nucleotide-binding and Crp-like helix-turn-helix domains as FnrL. However, unlike FnrL, CrpK does not possess the 4 cysteine residues at its N-terminus required for coordination with iron-sulfur clusters, suggesting CrpK might not directly sense oxygen. Nevertheless, ectopic expression of CrpK in an *ΔfnrL* mutant from an IPTG-inducible plasmid restores photosynthetic growth on succinate (Figure S3A), indicating CrpK might directly regulate a similar set of genes as FnrL. However, a markerless *crpK* deletion mutant is capable of photosynthetic growth on succinate (Figure S3B), indicating that FnrL and CrpK might also have distinct targets.

CrpK and FnrL regulons have overlapping but distinct members

ChIP-seq analysis using a 3X myc tagged variant of CrpK (Figure S3A) identified a total of 38 binding sites for CrpK in the *R. sphaeroides* genome (Table 6-4, Figure 6-3A). Consistent with its predicted involvement in regulation of photosynthetic genes, CrpK was found to bind to the upstream regulatory regions of *bchEJGP* and *hemA*, both known to be involved the biosynthesis of photopigments or their tetrapyrrole precursors [35-37]. CrpK binding sites were also found upstream of genes encoding iron transporters (*feoABC, ccmD*) and iron-sulfur cluster binding proteins (*rdxBHIS*). In addition, 23 (60.5%) of the identified CrpK sites were also identified as FnrL target sites (Table 6-4, Table 6-1), possibly providing an explanation for the ability of CrpK to at least partially compensate for the loss of FnrL

(Figure S3A). The remaining 15 CrpK binding sites were not occupied by FnrL under the conditions we tested (Table 6-4). On the other hand, FnrL was found bound to 39 sites that were not recognized by CrpK. These observations of TF occupancy for a subset of these overlapping and distinct sites were also verified via independent ChIP-qPCR analysis (Figure S3C), but it should be noted that many of the "CrpK unique sites" (i.e., those not also bound by FnrL) are relatively low enrichment sites in ChIP-seq assays. However, this set of "CrpK unique sites" all possessed a shared motif to other identified CrpK sites, so we consider it likely that these are actually direct targets for control by this TF. The ChIP-seq peaks for CrpK and FnrL at sites bound by both TFs were centered at the same location for both TFs and consequently the predicted binding motifs for both TFs bear strong DNA sequence similarity at both shared and unique sites (Figure 6-3B). This observation is consistent with general motif type recognized by Crp/Fnr-family TFs and the relatively high degree of amino acid sequence similarity in the predicted DNA binding motifs of CrpK and FnrL [23]. However, subtle differences between the motifs that are assembled by analysis of the "FnrL and CrpK unique sites" could be discerned, which might allow for future computational or experimental discrimination between target sites for each TF (Figure 6-3B).

CrpK controls expression of its predicted target genes

To independently test the role of CrpK on its predicted target genes, we conducted microarray analysis comparing the expression of a *ΔfnrL* strain to a *ΔfnrL* strain expressing *crpK* from an IPTG-inducible plasmid, to bypass the regulatory effect of FnrL in these target operons (unfortunately, a *ΔfnrLΔcrpK* strain was not viable under any anaerobic or photosynthetic conditions that we tested). Of the 28 CrpK target operons identified by ChIP-seq (and for which probes exist on the *R. sphaeroides* gene chip), 14 of these were DE (1.5 FC, $p<0.01$) under these conditions (Figure 6-3C, Table 6-4), indicating that CrpK controls expression from these target promoters in this reporter strain. Transcripts from 13 of these DE operons were increased including those encoding RSP_0166 (a TraR-like protein), UbiD, RPS_0697 (universal stress protein, Usp) and iron transporter FeoABC, whereas 1 operon was down regulated in the absence of CrpK (Figure 6-3C). Just as in the case of FnrL, regulatory control at some of the CrpK target

operons such as *hemA, nuoA-N* and *bchEJGP* might be obscured by reprogramming of the transcriptional network in the *ΔfnrL* strain. Thus, the number of functional CrpK targets is likely larger. In general, the predicted direction of regulation by CrpK in all of the DE operons were that same as that observed with FnrL (Table 6-4, Figure 6-3C).

As an added test of the ability of CrpK to control expression of its target promoters, we analyzed expression of promoter-*lacZ* fusions, using a few CrpK and FnrL target genes, in *ΔfnrLΔcrpK* double mutant reporter strains. Consistent with the predictions of our genome-wide analysis, we observed that CrpK was able to increase β-galactosidase activity from the *bchE,* universal stress protein (RSP_0697) and *ccoN* promoters, similar to FnrL though albeit with lower promoter activity (Figure 6-3D). This result also indicates a positive regulatory role for both FnrL and CrpK have on *bchE,* a fact which was not possible to infer from our global gene expression data. No significant β-galactosidase activity was observed with the RSP_3341 promoter upon ectopic expression of CrpK, whereas significant β-galactosidase activity was obtained when FnrL synthesis was induced. This data is also consistent with the ChIP-seq data, which indicates that FnrL, but not CrpK, binds to the RSP_3341 promoter. For the one unique CrpK site tested in this analysis, RSP_2349, we did not observe any increase in β-galactosidase activity from this promoter when either CrpK or FnrL was ectopically expressed, suggesting that this CrpK binding site might not be functional under the growth conditions tested. Nevertheless, the ability of CrpK to bind the upstream regulatory regions of several FnrL target genes such as *bchEJGP* and *hemA*, as well as control their expression, provides a direct explanation for the ability of increased CrpK expression to restore photosynthetic growth to an FnrL mutant of *R. sphaeorides*.

CrpK and FnrL share similar DNA binding determinants

To test the above prediction that CrpK and FnrL can recognize related DNA sequences, we made 2 separate base substitutions in the predicted FnrL/CrpK consensus sequence of the *bchE* promoter, substituting thymine of the TCAA (at position -57 relative to the start codon) with either a guanine or

cytosine. When these base substitutions were introduced into the *bchE* promoter fused to a promoterless *lacZ* gene, we found that they caused a significant decrease in β-galactosidase activity relative to the WT promoter when either FnrL or CrpK synthesis was induced (Figure 6-3E). An ~90% reduction in β-galactosidase activity was observed with both promoter mutations when FnrL synthesis was induced, while ~70 and 80% decreases were observed with the individual guanine and cytosine mutations, respectively, when CrpK synthesis was induced. These data indicate that both TFs recognize similar sequences, consistent with predictions from the motif finding analysis and the fact that they belong to the same TF family.

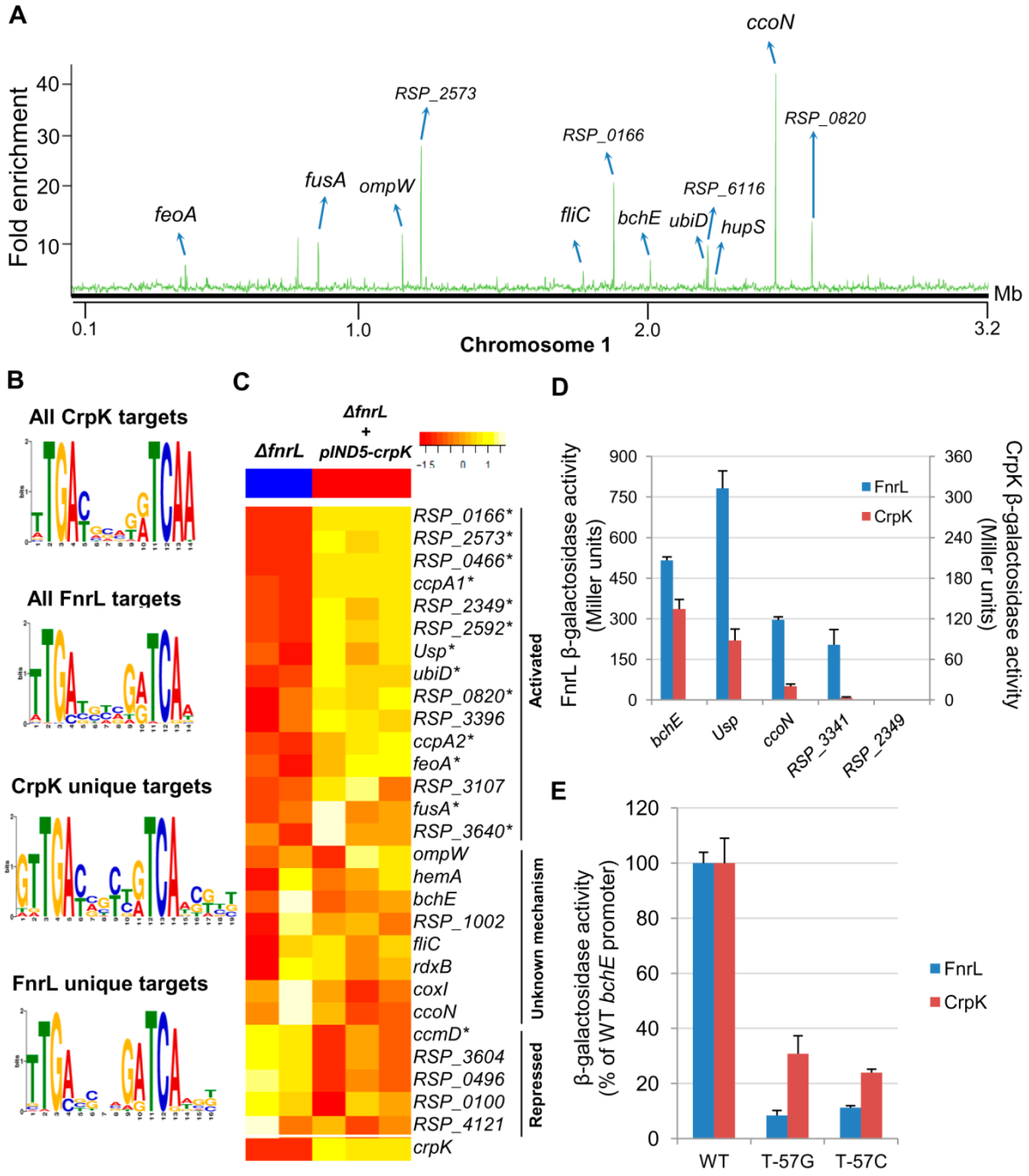

Figure 6-3. Analysis of the CrpK regulon in *R. sphaeroides*. (A) CrpK binding sites across chromosome 1. (B) Comparison of motifs generated for all CrpK and FnrL ChIP-seq identified targets, as well as those for targets exclusive to the CrpK and FnrL ChIP-seq data sets. (C) Pair-wise comparison of the global transcript level data between a *ΔfnrL* strain and a *ΔfnrL* strain over-expression CrpK (*ΔfnrL*+pIND5-*crpk*) for all ChIP-seq identified CrpK target operons. Only the first members of the operons are shown. Significantly DE genes are indicated with *. (D) β-galactosidase activity observed after inducing either CrpK or FnrL synthesis using promoter-*lacZ* fusions of the *bchE*, RSP_0697 (Usp), *ccoN*, RSP_3341 and RSP_2349 promoters, integrated into the chromosome of a *ΔfnrL-ΔcrpK* reporter strain. (E) Percentage β-galactosidase at mutant *bchE* promoters relative to the WT promoter.

Table 6-4. CrpK binding sites across the *R. sphaeroides* genome identified by ChIP-seq.

S/N†	ID	Annotation	chrID	Start[a]	Stop[a]	FC[b]	Reg[c]
1	RSP_0069*	*fliC* (Flagellar filament protein)	chr1	1776800	1777199	3.84	
2	RSP_0100-12	*nuoA-N* (NADH dehydrogenase)	chr1	1812000	1812399	2.7	-
3	RSP_0166	TraR-like protein	chr1	1882200	1882599	22.6	+
4	RSP_0281-76	*bchEJGP* (Bacteriochlorophyll biosynthesis)	chr1	2007600	2007999	6.5	+
5	RSP_0466-4	Hypothetical protein	chr1	2201400	2201999	5.3	+
	RSP_0467-8	*ubiD*					+
6	RSP_0496*	*hupL* (Hydrogenase protein large subunit)	chr1	2232000	2232399	3.15	-
7	RSP_0692-89*	*rdxBHIS* (Iron-sulfur cluster-binding protein)	chr1	2436600	2436799	2.33	
8	RSP_0696-3	*ccoNOQP* (Cbb3-type cytochrome c oxidase)	chr1	2440200	2440799	41.4	+
	RSP_0697	Universal stress protein (Usp)					+
9	RSP_0820	Cytochrome B561	chr1	2565800	2566399	13	+
10	RSP_1002-3	Aspartate carbamoyltransferase	chr1	2760200	2760599	3.7	
11	RSP_1804*	*ccmD* (Heme exporter protein D)	chr1	388000	388399	3.42	-
12	RSP_1819-17	*feoABC* (Ferrous iron transport protein)	chr1	409200	409399	3.46	+
13	RSP_1877-6	Cytochrome c oxidase aa3 type (*coxI*)	chr1	477000	477399	3.61	-
14	RSP_2247	*fusA* (Elongation factor G)	chr1	862600	862999	9.77	+
15	RSP_2337	*ccpA1*	chr1	964200	964799	2.7	+
16	RSP_2349*	Hypothetical protein	chr1	975000	975399	3.2	
17	RSP_2395	*ccpA2* (BCCP, cytochrome c peroxidase)	chr1	1022400	1022799	3.13	+
18	RSP_2507	*ompW* (Outer membrane protein)	chr1	1152400	1152999	13.2	
19	RSP_2573	Hypothetical protein	chr1	1216800	1217999	21.26	+
20	RSP_2592*	Hypothetical protein	chr1	1234000	1234399	3.25	+
21	RSP_2984	*hemA* (5-aminolevulinate synthase)	chr1	1675800	1676199	2.83	
22	RSP_3107*	Hypothetical protein	chr2	152400	152999	3.27	+
23	RSP_3396-3	ABC opine/polyamine transporter	chr2	461000	461399	2.96	+
24	RSP_3604*	tRNA 2-selenouridine synthase	chr2	698600	698999	4.6	-
25	RSP_3640*	Hypothetical protein	chr2	748400	748599	3	+
26	RSP_4121	Hypothetical protein	plasmidC	88000	88399	2.94	-
27	RSP_6116	Hypothetical protein	chr1	2206600	2206999	8.7	NA
28	RSs0019	small RNA RSs0019	chr2	28600	28999	3.5	NA
29			chr1	403800	404199	5.26	NA
30			chr1	785400	785799	3.27	NA
31			chr1	792200	792799	10.18	NA
32			chr1	2193200	2193599	2.97	NA
33	*		chr1	2338000	2338399	3.1	NA
34	*		chr2	449800	450399	11.28	NA
35	*		chr2	856800	857399	7.1	NA
36	*		plasmidB	24000	24599	6.2	NA
			plasmidB				NA
37	*		plasmidD	17600	18199	5.56	NA
38	*		plasmidD	90000	90399	4.7	NA

* ChIP-seq binding sites bound by CrpK but not FnrL.

[a] Chromosomal locations of start and stop of ChIP-seq peaks.

[b] Fold enrichment of CrpK-myc ChIP over control myc antibody ChIP in WT control.

[c] Regulatory role of CrpK on target operons based on change in gene expression between WT and ΔcrpK cells. + = positively regulated by CrpK. - = negatively regulated by CrpK. NA - Not applicable (i.e., these genes are not represented on the R. sphaeroides)

† Number of binding sites. Some binding sites correspond to more than one operon.

RSP_2888 – a newly identified negative regulator of photopigment biosynthesis

Another TF predicted by the large-scale TRN to play a role in the control of *R. sphaeroides* photosynthesis genes is the BadM/Rrf2 family TF, RSP_2888. RSP_2888 transcript levels are increased under photosynthetic conditions in WT cells and RSP_2888 is predicted to be a direct target of PrrA (Table 6-3, Figure 6-2B). Consistent with this, RSP_2888 transcript levels are more than 5 fold higher in WT cells relative to *ΔprrA*, being the most DE TF in that data set (Table S2). To test the role of RSP_2888 in regulation of photosynthesis, we conducted a combination of physiological, gene expression and protein-DNA binding assays for this TF.

RSP_2888 negatively modulates photopigment synthesis

To assess the physiological role of RSP_2888, we constructed and analyzed the properties of a RSP_2888 deletion mutant (ΔRSP_2888). Furthermore, ΔRSP_2888 was complemented with RSP_2888 from an IPTG-inducible plasmid (ΔRSP_2888+pIND5-*RSP_2888*). The WT and ΔRSP_2888 strains both exhibited similar growth rates (Figure 6-4A), while the complemented strains also grew at similar rates up to 10μM IPTG, beyond which photosynthetic growth, but not aerobic growth, was severely negatively impacted (Figure 6-4A, Figure S4A). This suggested a role for RSP_2888 in one or more aspects of photosynthesis.

It appeared that RSP_2888 had some impact on photopigment biosynthesis as there was a reduction of colony pigmentation when this protein was expressed from an IPTG-inducible plasmid. To test this hypothesis, we assessed the total amount of bacteriochlorophyll in cells containing or lacking RSP_2888. These experiments showed that the ΔRSP_2888 mutant strain produces >50% more bacteriochlorophyll than its WT parent (Figure 6-4B). Furthermore, ectopic expression of RSP_2888 lowered the amount of cellular bacteriochlorophyll (a 2-fold decrease from WT levels at 10uM IPTG), providing additional support for the role of RSP_2888 as a regulator of photopigment synthesis (Figure 6-4B). These data also indicate that RSP_2888 functions to negatively modulate photopigment synthesis.

RSP_2888 is a transcriptional repressor of photopigment and other photosynthetic genes

Given the observed negative role of RSP_2888 on photopigment synthesis, we conducted global gene expression analysis on the WT, ΔRSP_2888 and ΔRSP_2888 + pIND5-*RSP_2888* strains under photosynthetic conditions. Consistent with the above observations, a variety of photosynthesis related genes were significantly DE between these strains. A total of 17 genes other than RSP_2888 were DE (cut-off – 1.5 fold, $p < 0.05$) between the WT and *ΔRSP_2888* (Figure 6-4C, Table S4), with 12 of these genes having photosynthesis related functions. Most of these DE genes include genes from 7 operons (*appA*, *rdxBHIS*, *bchCXYZ*, *hemF*, *bchF*, *crtIB*, RSP_0278). Furthermore, transcript levels from all but one of the DE genes were all down regulated in the presence of RSP_2888, indicating that it functions as a transcriptional repressor. This predicted negative regulatory function for RSP_2888 is consistent with the decreased photopigment phenotype seen when this TF is over-expressed. Given that RSP_2888 transcript levels are increased under photosynthetic growth conditions relative to aerobic conditions in WT cells, its function is likely to fine-tune photopigment synthesis, similar to the predicted role for the sRNA, PcrZ [27].

By comparing the global gene expression profile of the ΔRSP_2888 strain with that of cells ectopically expressing this TF, we found a total of 36 genes DE (1.5 FC, $p < 0.05$), including 14 of the 17 genes that were DE between WT and ΔRSP_2888 (Table S5). The other 22 DE genes that were DE when RSP_2888 was ectopically expressed included additional photosynthesis related genes like those encoding light-harvesting proteins (*pucC* and *puc2B*), as well as genes involved in functions ranging from iron and heme transport to fatty acid biosynthesis (Figure 6-4C, Table S5). Twenty six of these 36 DE genes were also down regulated when RSP_2888 was ectopically expressed, consistent with this TF functioning as a repressor of photosynthesis and other functions.

RSP_2888 directly regulates genes encoding photosynthesis related proteins

To further test the predicted direct role of RSP_2888 in regulation of photosynthesis, we performed 2 independent ChIP-seq analyses using a *ΔRSP_2888* strain containing a 3X myc-tagged variant of RSP_2888 that complements the phenotype of the parent strain (Figure S4B). This ChIP-seq analysis identified a total of 52 RSP_2888 binding sites across the genome. We found that most of the genes downstream of these binding sites were not DE in the presence or absence of RSP_2888 in any of global gene expression data sets. In addition, we were unable to identify any conserved DNA motif shared by a significant number of these target sites. Thus, to identify potential direct targets of RSP_2888, only operons that were both DE in either of our global gene expression data sets and had a significant ChIP-seq peak were considered. When these criteria were applied, we identified 9 potential RSP_2888 target operons: *bchCXYZ*, *bchFNBH*, *rdxBHIS*, *appA*, RSP_1257-4, RSP_2692, RSP_2888, RSP_2961 and RSP_3718 (Table 6-5, Figure 6-4D). Transcript levels from all of these operons were lower in the presence of RSP_2888 and they represent the most high confidence direct RSP_2888 targets in our dataset. The 43 other sites which showed significant enrichment for RSP_2888 are provided in Table S6.

Table 6-5. RSP_2888 target genes identified by ChIP-seq and gene expression analysis

	ID	Annotation	chrID	Start[a]	Stop[a]	FC[b]	Reg[c]
1	RSP_0263-59	*bchCXYZ-pufQ*	chr1	1988000	1988199	2	-
2	RSP_0284-91	*bchFNBH-chlL-bchM-puhA*	chr1	2010200	2010799	3.23	-
3	RSP_0692-89	*rdxBHIS*	chr1	2436600	2436999	3.35	-
4	RSP_1257-4	Polyhydroxyalkanoic synthase	chr1	3026800	3027199	4.3	-
5	RSP_1565	*appA*	chr1	156400	156599	2.75	-
6	RSP_2692	Acyltransferase domain (LPS)	chr1	1338400	1338799	2.61	-
7	RSP_2888	Transcriptional regulator, BadM/Rrf2 family	chr1	1564600	1565599	17.4	NA
8	RSP_2961	Protein containing a CBS domain	chr1	1643600	1643799	2.45	-
9	RSP_3718	Hypothetical protein	chr2	841800	842199	2.24	-

[a] Chromosomal locations of start and stop of ChIP-seq peaks.
[b] Fold enrichment of RSP_2888-myc ChIP over control myc antibody ChIP in WT control.
[c] Regulatory role of RSP_2888 on target operons based on change in gene expression between ΔRSP_2888 and WT or ΔRSP_2888+pIND5-RSP_2888 cells. - = negatively regulated by RSP_2888. NA - Not applicable.
† Chromosomal locations of start and stop of ChIP-seq peaks.

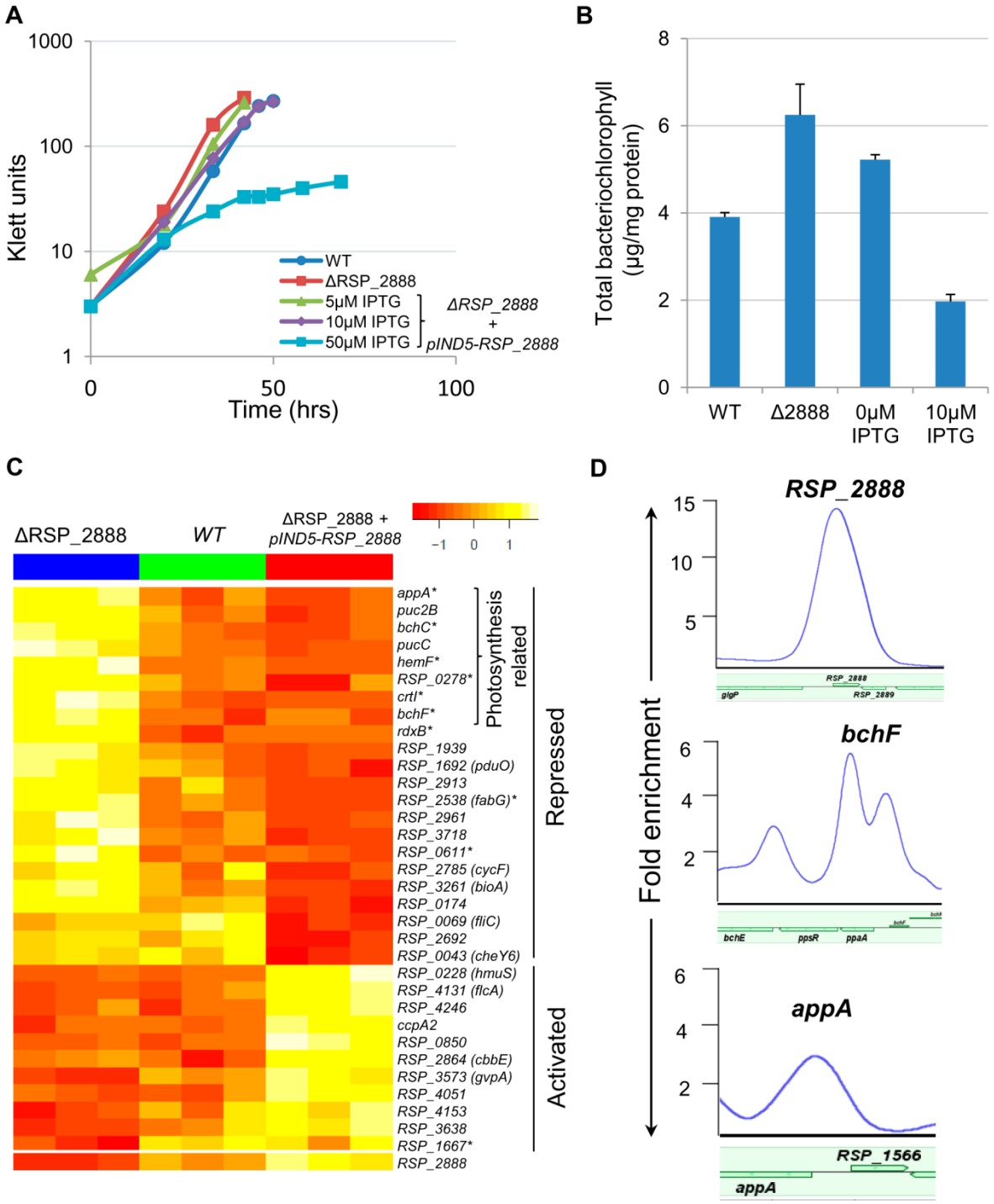

Figure 6-4. Physiological and genomic analysis of RSP_2888 regulation. (A) Growth of WT, *ΔRSP_2888* and *ΔRSP_2888*+pIND5-*RSP_2888* with increasing IPTG concentrations under photosynthetic conditions. (B) Amounts of bacteriochlorophyll produced in WT, *ΔRSP_2888* and *ΔRSP_2888*+pIND5-*RSP_2888*. (C) Expression profiles of genes differentially expressed in response to the loss of RSP_2888 (*ΔRSP_2888*) or over-expression of RSP_2888 (*ΔRSP_2888*+pIND5-*RSP_2888*) strains. Genes DE in the *ΔRSP_2888* only are indicated with an asterisk (*). (D) ChIP-seq binding profile of RSP_2888 at the RSP_2888, *bchF* and *appA* promoters.

Discussion

Our analyses have provided new information on the TRN controlling bacterial photosynthesis in *R. sphaeroides*. We confirmed the predicted involvement of two previously uncharacterized TFs, CrpK and RSP_2888, in the regulation of photosynthesis related genes. We also extended the regulons of PrrA and FnrL, which had previously been implicated in regulation of the photosynthetic lifestyle of *R. sphaeroides*. Our analyses, combined with previous analyses of PrrA, FnrL and PpsR, illustrate the depth, complexity and robustness of the photosynthetic TRN, highlighting significant combinatorial regulation of target genes, cross-talk between regulators and redundancy in the use of TFs within this network (Figure 6-5).

New insights into regulation by PrrA and FnrL

Previous analysis of photosynthetic gene expression in *R. sphaeroides* had established the importance of 3 global TFs, PpsR, PrrA and FnrL, in the regulation of this lifestyle [8,9,18,19,23,25,26]. Our analysis extends previous observations by comprehensively identifying the direct targets for PrrA and FnrL, complementing our genome wide analysis of the PpsR regulon (Imam et al. unpublished data). As part of this study, a new global gene expression analysis for PrrA provided an improved view of the targets controlled by PrrA in *R. sphaeroides*. Our analysis indicated the total number of genes, directly or indirectly, controlled by PrrA is ~4 times smaller than previously reported [19], providing a picture of the PrrA regulon not influenced by apparent growth-rate differences between wild type and the mutant used in the original study. In addition, data obtained from ChIP-seq analysis provided the first genome-scale view of PrrA interactions *in vivo* and verified the major direct role played by this TF in photosynthesis gene expression. Our data shows that PrrA, in addition to controlling genes required for light energy capture, also regulates a large number of genes involved in electron transport both directly (e.g., *fbcFBC* complex, cytochrome B561 and *cycA* [38]) and indirectly (Table S2).

Although we identified several new PrrA direct targets, we were unable to identify a strong consensus binding motif for this TF. While PrrA, and its analog in *Rhodobacter capsulatus* RegA, have been proposed to bind a degenerate GCG inverted repeat with a varying length spacer region, previous analyses of PrrA binding in *R. sphaeroides* have suggested that both DNA curvature and sequence specificity might contribute to target site recognition [22,30]. These potential features, together with the GC-rich nature of the *R. sphaeroides* genome and the Fis-like nature of the PrrA-binding domain [22], possibly made it difficult to identify a shared motif among target genes from our analysis. Thus, the determinants of sequence specific DNA binding by PrrA in *R. sphaeroides* remains an open question.

A large-scale reconstruction of the *R. sphaeroides* TRN predicted the FnrL regulon in *R. sphaeroides* was larger than previously described [23]. Our studies verified several of these additional direct FnrL targets, significantly extending the size and function of genes in its regulon. Some of the newly identified targets include nitrogen regulatory proteins, iron sulfur assembly proteins, ABC transporters, additional TFs and recently identified sRNAs, all of which significantly increase the scope of genes and functions that are controlled by FnrL.

Redundancy of the photosynthetic TRN

One of the previously uncharacterized TFs we tested for a role in the photosynthetic lifestyle in this study was CrpK. Genome-wide analysis of CrpK targets revealed an overlapping but distinct regulon to that of FnrL, providing an explanation for the ability of CrpK to rescue the photosynthesis defect of an FnrL deletion strain. While bacterial TRNs are often tightly controlled, they also need to be robust to allow cells to adapt to potentially deleterious changes to these networks. Given the central role of FnrL in regulating photosynthesis and a large number of anaerobic processes, the redundancy observed between this TF and the CrpK regulon might function to provide robustness to the *R. sphaeroides* photosynthetic TRN. Alternatively, CrpK might have a broader function under different conditions from FnrL. For instance, the absence of an oxygen sensitive iron-sulfur cluster in CrpK might allow this protein to

function at oxygen tensions that would inactivate FnrL, facilitating photosynthesis or other metabolic functions during microaerophilic or semi-aerobic growth in nature. Interestingly, while FnrL binds at the promoter of *dorS* (the histidine kinase of the DorSR two-component system involved in regulation of anaerobic respiratory growth on DMSO [39]), CrpK binding was not observed at this promoter. This suggests that CrpK's ability to functionally replace FnrL might not extend to FnrL's role in regulation of anaerobic respiration.

Although the predicted consensus motifs derived from the FnrL and CrpK binding sites were similar, the observation that both proteins can recognize unique, as well as overlapping sites, indicates there must be some subtle but functionally significant differences in DNA recognition by these TFs. Closer inspection of these DNA sequence motifs, suggest there may be a greater tolerance by FnrL for deviations from the TGA N_6 TCA consensus, while bases within the spacer region or outside the core target site might play a role. Under the anaerobic photosynthetic growth conditions typically used in the lab, the CrpK transcript is present at a significantly lower level than that of FnrL. Thus, while it is possible that both proteins might compete for some shared binding sites under these conditions, our analysis of a few shared or unique promoters suggested this was likely not a major factor under the conditions we tested (Figure S3D). However, reproducibly greater FnrL enrichment was observed at the *ccoN*/RSP_0697 promoter in the absence of CrpK (Figure S3C), so the possibility that both proteins compete for binding at selected sites cannot be ruled out without additional genetic and biochemical studies.

Incoherent feed forward regulation of the photosynthesis TRN

The second photosynthesis-related TF characterized for the first time in this study was RSP_2888. Our data showed that RSP_2888 functions as a direct transcriptional repressor of photopigment biosynthesis, including the *bchCXYZ* and *bchFNBHM* operons, with high cellular levels of this protein inhibiting photosynthetic growth. In addition, transcripts from several other operons that encode photosynthesis-related functions were indirectly repressed by RSP_2888. Our data predict that much of this indirect

regulation of photosynthesis function is achieved through the direct regulation of the gene that encodes the anti-repressor, AppA, by RSP_2888. Reduced cellular levels of AppA caused by the presence of RSP_2888, would in principle cause accumulation of free PpsR under photosynthetic conditions, which would lead to repression of the photosynthesis-related genes that are PpsR targets (Figure 6-5). Given that RSP_2888 transcript levels are significantly elevated during photosynthetic growth, its function in repressing photopigment synthesis would appear to be counterintuitive, similar to the observation for the sRNA, PcrZ [27]. Since no significant difference in photosynthetic growth was observed between WT and ΔRSP_2888 cells, the additional pigment produced in the RSP_2888 mutant strain did provide increased fitness, potentially equating to a waste of cellular resources in the production of this extra pigment. In addition, the presence of excess photopigment could be a source of metabolic stress, especially since they can result in production of reactive oxygen species if light is present under microaerophilic conditions in the lab or in nature [40]. Thus, RSP_2888 may function as a negative modulator of pigment synthesis to ensure the optimal expression and tight coordination between expression of photopigment biosynthetic pathway genes and those for other components of the photosynthetic apparatus. Based on observed function of RSP_2888, we propose naming this TF, MppG (modulator of photopigment genes).

Newly-identified links between photosynthetic and iron homeostasis gene regulatory networks

In addition to its role in photosynthesis, RSP_2888 also regulates, either directly or indirectly, a variety genes encoding iron/heme dependent proteins (AppA, RdxBHIS, BchX, BchL, RSP_2785) and iron/heme transporters (RSP_2913, HmuS). RSP_2888 shares a high degree of amino acid sequence similarity to RirA, which was previously shown to regulate iron-responsive genes in *Rhizobium leguminosarum* [41,42]. Thus, in addition to its role in regulation of photopigment synthesis, RSP_2888 appears to have a previously unidentified role in maintaining iron homeostasis during photosynthetic growth in *R. sphaeroides*. Furthermore, like RirA, RSP_2888 possesses a set of cysteine residues in its C-terminal

region, which could coordinate an iron-sulfur cluster or some other metal, potentially allowing it to directly sense signals such as oxygen or metal availability.

Our data also provide new evidence that both FnrL and CrpK directly regulate genes encoding iron-dependent, iron transport and iron-sulfur biogenesis proteins, as well as several proteins involved in tetrapyrrole biosynthesis. In addition, we showed that FnrL directly activates expression of another RirA-like protein, RSP_3341, which has also been shown to directly regulate other iron dependent genes in *R. sphaeroides* (Imam et al. unpublished data). Thus, we have provided new evidence that the TRNs and TFs controlling photosynthesis and iron homeostasis are tightly linked in *R. sphaeroides*. This link is likely, at least in part, due to the anaerobic anoxygenic mode of photosynthesis in *R. sphaeroides*, the sensitivity of Fe-S clusters to oxygen, and the involvement of a variety of iron-dependent proteins in light energy capture or other aspects of photosynthesis.

Cross-talk between TFs that regulate photosynthetic gene expression

The TRN controlling bacterial photosynthesis appears to be remarkably complex when compared to other characterized bacterial TRNs. For complex TRNs to function effectively the components of the network often need to communicate with one another, and this is the case with the photosynthetic TRN. For example, our previous analysis of the PpsR regulon identified a PpsR binding site upstream of *prrA*, in an intra-operonic promoter shown to be occupied by σ^{70} under photosynthetic conditions [23]. If this PpsR binding site upstream of *prrA* is functional, it would provide an additional, previously unrecognized, mechanism to prevent aerobic expression of photosynthetic genes. We also found that PrrA directly activates both AppA and RSP_2888, which in turn represses AppA, forming an incoherent feed-forward loop to control photosynthetic genes (Figure 6-5) similar to the situation proposed for PcrZ [27]. Thus, our data provides new support for the previous hypotheses that the control of *appA* transcription serves as a major point of integration of regulatory signals, integrating opposing regulatory inputs from PrrA, RSP_2888, PcrZ, oxygen and possibly other as of yet unidentified factors. We predict that this network

architecture likely results in a rapid response of cells to small environmental perturbations and allows optimal expression of photosynthetic genes under anaerobic conditions.

Concluding remarks

Using a combination of genetic, genomic and physiological approaches, guided in large part by computational predictions from a large-scale reconstruction of the *R. sphaeroides* TRN, we obtained a significant amount of new knowledge about regulation of photosynthesis in *R. sphaeroides*. Our analyses highlight the important role computational predictions can play in guiding biological discovery, as novel components of the photosynthetic TRN, not previously identified using traditional approaches, were identified computationally, with those predictions serving as the basis for this work. We expect that predictions from this large-scale TRN will continue to provide new insights into other aspects of *R. sphaeroides* diverse metabolic and energetic lifestyles, including those involved in production of high-value commodities such as biofuel precursors. In addition, given the ancestral relationship of *R. sphaeroides* to plants and other oxygenic phototrophs, we predict that knowledge of this photosynthetic TRN will help inform parallel or future studies in other photosynthetic organisms. Integration of the available large-scale networks of metabolism and transcriptional regulation for *R. sphaeroides*, will broaden the predictive capabilities of these models and further guide future experimental efforts.

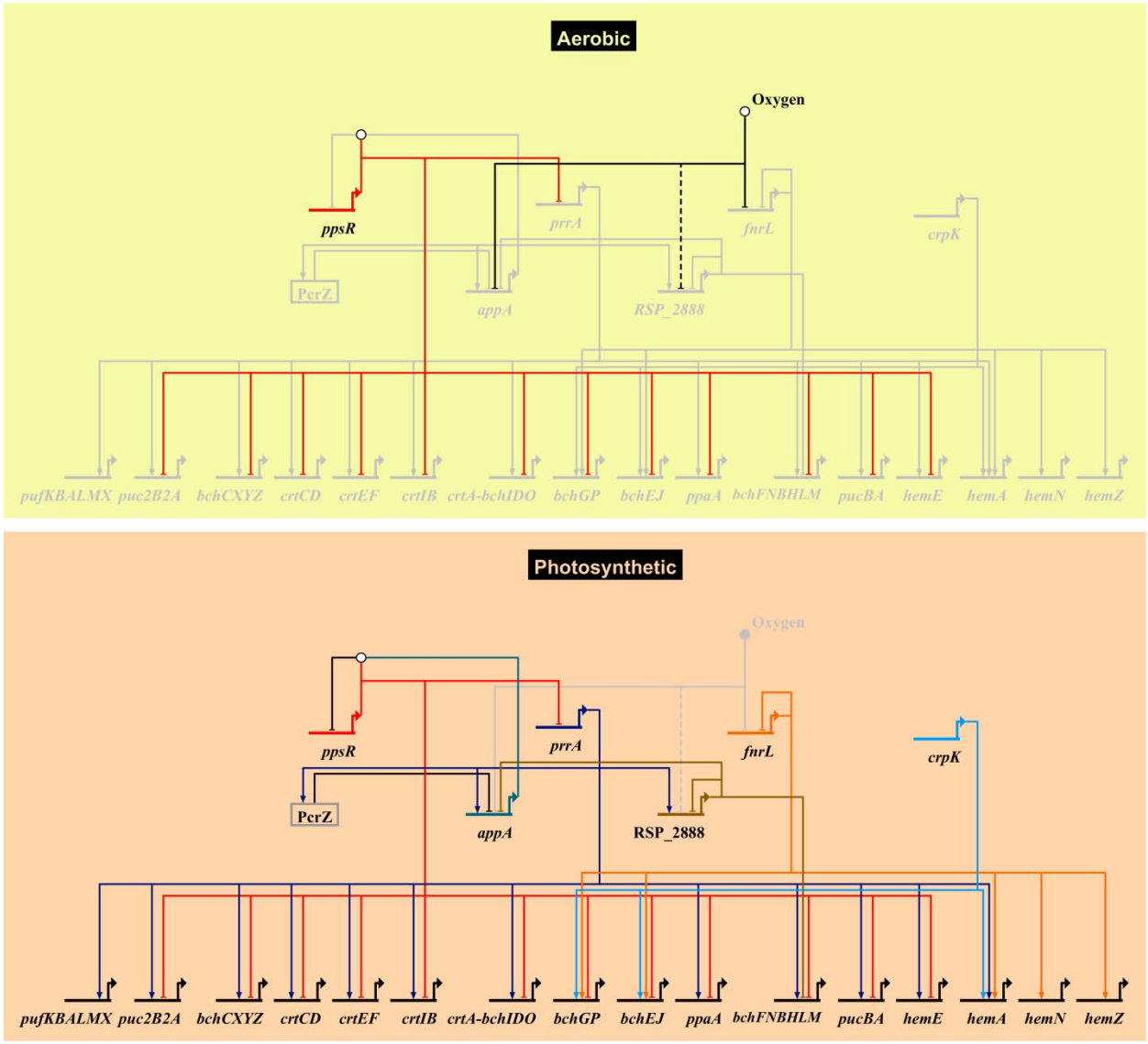

Figure 6-5. Photosynthetic gene regulatory network. An overview of the *R. sphaeroides* photosynthetic gene regulatory network, showing known transcriptional regulators and their photosynthesis related direct target genes. Under aerobic conditions (top panel), the oxygen sensitive anti-repressor protein AppA is inactivate, allowing PpsR to repress the expression of photosynthesis related genes including the *prrA*, thereby preventing the production of photopigments. In addition, oxygen inhibits the activity of FnrL and potentially RSP_2888 (the uncertain nature of effect is indicated by a dashed edge). Under anaerobic conditions, AppA becomes active and inhibits PpsR activity via protein-protein interactions (depicted by the white circle above *ppsR*). In addition, the activators of photosynthesis, PrrA, FnrL and CrpK, become active under these conditions and induce the expression of photosynthetic genes. Under these conditions, the photopigment gene repressors RSP_2888 and PcrZ are also active, negatively modulating photopigment gene expression. The expression of *appA* and the activity of its gene product is dependent on regulatory inputs from PrrA, RSP_2888, PcrZ and oxygen. Biotapestry was used for network visualization [59].

Materials and Methods

Bacterial strains and growth conditions

R. sphaeroides 2.4.1 was used as the parental (wild type) strain and all mutants were constructed in this background except the *ΔfnrLΔcrpK* double deletion strain which was constructed in an existing *ΔfnrL* mutant background [9] (Table S7). *E. coli* DH5α was used as a plasmid host, and *E. coli* S17-1 was used to conjugate DNA into *R. sphaeroides*. *R. sphaeroides* cultures were incubated at 30 ^{o}C in Sistrom's minimal medium (SMM) [43]. When required, the media was supplemented with 100 μM IPTG, 25 μg/mL kanamycin, 25 μg/mL spectinomycin, or 1 μg/mL tetracycline. *E. coli* cells were grown in Luria Bertani medium at 37 ^{o}C, supplemented with 50 μg/mL kanamycin, 25 μg/mL spectinomycin, or 20 μg/mL tetracycline where needed.

Quantification of bacteriochlorophyll

Photosynthetic pigments were quantified in *R. sphaeroides* strains grown photosynthetically in screw cap tubes at a light intensity of ~10 W/m^2 as previously described [44]. Briefly, 5 mL of culture was centrifuged and supernatant discarded. Cells were resuspended in 100 μL of water, transferred to 15 mL glass corex centrifuge tubes held in centrifuge adaptors and covered with rubber stoppers to prevent exposure to light. 4.9 mL of a 7:2 mixture of acetone and methanol was added to the cell suspension and vortexed thoroughly in the dark. Samples were centrifuged for 10 at 10000g. Absorbance of the supernatant was measured at 775 nm and total bacteriochlorophyll was determined as follows: Abs_{775} * total volume of sample (5mL) * (bacteriochlorophyll molecular weight (914 g/mol)/ bacteriochlorophyll millimolar extinction coefficient (75 mM^{-1} cm^{-1})). Total bacteriochlorophyll in each sample was normalized to total protein content of samples determined using the Lowry assay [45].

Construction of mutants and expression plasmids

All mutants constructed for this study contained in-frame markerless deletions. Regions spanning ~1500 bp upstream and downstream of the target gene were amplified using sequence-specific primers containing restriction sites for EcoRI, XbaI or HindIII. These fragments were digested with the appropriate restriction enzymes and ligated into pK18mobsacB [46], which had been digested with EcoRI and HindIII, by three-way ligation to generate the various gene deletion constructs, which were confirmed by sequencing. The pK18mobsacB-based plasmids were separately mobilized from *E. coli* S17-1 into relevant *R. sphaeroides* strains. Cells in which the plasmid had successfully integrated into the genome via homologous recombination were identified by selection on SMM plates supplemented with kanamycin. These cells were then grown overnight in SMM without kanamycin. Cells that had lost the *sacB* gene via a second recombination event were identified by growth on SMM plates supplemented with 10% sucrose. Gene deletions were confirmed by PCR and sequencing with specific primers (Table S7).

To construct plasmids for the ectopic expression of 3x Myc tagged proteins, we modified pIND5 [47] to include 3 copies of a codon optimized Myc tag (EQKLISEEDL – GAGCAGAAGCTGATCTC**G**GAGGAGGACCTG) within the plasmid's multiple cloning site. New multiple cloning sites were added to allowing tagging of proteins either C-terminally (NdeI-PstI-NcoI) or N-terminally (BamHI-SalI-BglII-HindIII). Individual expression plasmids were made by amplifying the target genes from the genome using sequence specific primers (Table S7) containing restriction sites for NdeI and BglII, HindIII or BamHI for cloning into pIND5 and NdeI/NcoI or BamHI/HindIII for cloning into pIND5-3xMyc. These DNA fragments were digested with the appropriate enzymes and cloned into pIND5 or pIND5-3xMyc digested with the same enzymes. These plasmids were conjugated from *E. coli* S17-1 into the relevant *R. sphaeroides* strains. Cells which harbor the desired plasmid were identified by selection on SMM plates supplemented with kanamycin.

Construction of *lacZ* reporter promoter fusions and β-galactosidase assays

To assay the activity of FnrL and CrpK *in vivo*, β-galactosidase assays were conducted in *ΔfnrLΔcrpK* deletion strains containing different promoter-*lacZ* fusions integrated into the genome. To construct these reporter strains, ~200 – 300 bp regions upstream of putative target genes (RSP_0281 (*bchE*), RSP_0696 (*ccoN*), RSP_0697 (*usp*), RSP_2346 and RSP_3341), were amplified from genomic DNA using specific primers having NcoI and XbaI restriction sites at their ends (Table S7). The amplified DNA fragments, as well as a pSUP202 suicide vector containing a promoterless *lacZ* gene [48], were digested with NcoI-XbaI. DNA fragments containing the upstream regulatory sequences were cloned into pSUP202. These promoter-*lacZ* fusion plasmids were then individually conjugated into the *ΔfnrLΔcrpK* strain, generating single copy promoter-*lacZ* fusions integrated in the genome after selecting for the plasmid-encoded tetracycline resistance activity. The *fnrL* and *crpK* genes cloned into pIND5 were conjugated into individual reporter strains and cells harboring the reporter construct and the ectopic expression plasmid were identified by selection with tetracycline and kanamycin. These strains were grown aerobically by shaking 10 mL of culture in 125 mL conical flasks until exponential phase, then were treated with 100 μM IPTG for 3 hrs to increase expression of the indicated TF before measuring β-galactosidase activity as previously described [49].

To assess the contribution of specific bases to FnrL and CrpK activity, β-galactosidase assays were conducted in *ΔfnrLΔcrpK* double deletion strains containing reporter gene fusions of the RSP_0281 (*bchE*) upstream regulatory region with individual point mutations (see Results). These reporter strains were constructed as described above, with individual point mutations being generated by overlap extension PCR. β-galactosidase assays were conducted as described above.

RNA extraction, qRT-PCR and microarray analyses

RNA was isolated from exponential phase cultures of *R. sphaeroides* strains that were grown photosynthetically in 16 mL screw cap tubes or 500 ml cultures in roux bottles with bubbling (95% N_2,

5% CO_2). RNA isolation and subsequent cDNA synthesis, labeling and hybridization to *R. sphaeroides* GeneChip microarrays (Affymetrix, Santa Clara, CA) were performed as previously described [50]. Microarray datasets were normalized by Robust Multichip Average (RMA) to $\log_2$ scale with background adjustment and quantile normalization [51]. Statistical analysis of normalized data to identify differentially expressed (DE) genes was done using the limma package [52]. Correction for multiple testing was done using Benjamini-Hochberg correction [53]. All analyses were conducted in the R statistical programming environment (http://www.R-project.org).

Chromatin immunoprecipitation analysis (ChIP-qPCR and ChIP-seq analysis)

R. sphaeroides cells were grown photosynthetically in 500 ml cultures (see above). For FnrL studies, 3 independent ChIP-seq experiments were conducted for WT cells grown photosynthetically with succinate (2 replicates) or acetate as sole carbon source. For tagged TFs, plasmids expressing the tagged variant of the gene from an IPTG inducible promoter, were cloned into the appropriate mutants (Table S7). Cells were harvested at mid-exponential phase and chromatin immunoprecipication was conducted as previously described [54], using polyclonal antibody against FnrL [23] or against the Myc epitope tag (ab9132, Abcam plc) for all other TFs analyzed. Immunoprecipitated DNA samples were PCR-amplified, gel purified (size selection ~200bp) and sequenced at the UW Biotechnoloy Center using the HiSeq 2000 sequencing system (Illumina, Inc). The initial 100bp sequence tags were trimmed to 70bp, to remove less reliable DNA sequences, and mapped to the *R. sphaeroides* strain 2.4.1 genome (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Rhodobacter_sphaeroides_2_4_1_uid57653/) using SOAP version 2.21 [55], allowing a maximum of 3 mismatches and no gaps. Peaks that represent potential TF binding sites were identified using MOSAiCS [56] at a false discovery rate of 0.05. The MOSAiCS analysis was conducted as a two-sample analysis, with control ChIP-seq data generated from Δ*fnrL* grown on acetate (for FnrL analysis), myc antibody ChIP in WT cells (for myc-tagged proteins) or input DNA. Only peaks that were called as significant using both input DNA and an appropriate ChIP control were considered as true peaks. Motifs were identified from within peak regions using MEME [57].

References

1. Atsumi S, Higashide W, Liao JC (2009) Direct photosynthetic recycling of carbon dioxide to isobutyraldehyde. Nat Biotechnol 27: 1177-1180.
2. Gronenberg LS, Marcheschi RJ, Liao JC (2013) Next generation biofuel engineering in prokaryotes. Curr Opin Chem Biol 17: 462-471.
3. Hunter CN, Daldal, F., Thurnauer, M. C. and Beatty, J. T. (2009) The purple phototrophic bacteria: Springer.
4. Peralta-Yahya PP, Zhang F, del Cardayre SB, Keasling JD (2012) Microbial engineering for the production of advanced biofuels. Nature 488: 320-328.
5. Blankenship RE, Madigan MT, Bauer CE (1995) Anoxygenic photosynthetic bacteria; Blankenship RE, editor. 1330 p.
6. Gomelsky L, Moskvin OV, Stenzel RA, Jones DF, Donohue TJ, et al. (2008) Hierarchical regulation of photosynthesis gene expression by the oxygen-responsive PrrBA and AppA-PpsR systems of *Rhodobacter sphaeroides*. J Bacteriol 190: 8106-8114.
7. Mackenzie C, Eraso JM, Choudhary M, Roh JH, Zeng X, et al. (2007) Postgenomic adventures with *Rhodobacter sphaeroides*. Annu Rev Microbiol 61: 283-307.
8. Zeilstra-Ryalls J, Gomelsky M, Eraso JM, Yeliseev A, O'Gara J, et al. (1998) Control of photosystem formation in *Rhodobacter sphaeroides*. J Bacteriol 180: 2801-2809.
9. Zeilstra-Ryalls JH, Kaplan S (1995) Aerobic and anaerobic regulation in *Rhodobacter sphaeroides* 2.4.1: the role of the fnrL gene. J Bacteriol 177: 6422-6431.
10. Imam S, Noguera DR, Donohue TJ (2013) Global insights into energetic and metabolic networks in *Rhodobacter sphaeroides*. BMC Syst Biol 7: 89.
11. Imam S, Yilmaz S, Sohmen U, Gorzalski AS, Reed JL, et al. (2011) iRsp1095: a genome-scale reconstruction of the *Rhodobacter sphaeroides* metabolic network. BMC Syst Biol 5: 116.
12. Khatipov E, Miyake, M., Miyake J. and Y. Asada (1999) Polyhydroxybutyrate accumulation and hydrogen evolution by *Rhodobacter sphaeroides* as a function of nitrogen availability. Biohydrogen III: 157 - 161.
13. Kien NB, Kong IS, Lee MG, Kim JK (2010) Coenzyme Q10 production in a 150-l reactor by a mutant strain of *Rhodobacter sphaeroides*. J Ind Microbiol Biotechnol 37: 521-529.
14. Kontur WS, Ziegelhoffer EC, Spero MA, Imam S, Noguera DR, et al. (2011) Pathways involved in reductant distribution during photobiological H_2 production by *Rhodobacter sphaeroides*. Appl Environ Microbiol 77: 7425-7429.
15. Tabita FR (1995) The biochemistry and metabolic regulation of carbon metabolism and CO_2-fixation in purple bacteria. In: Blankenship RE, Madigan MT, Bauer CE, editors. Anoxygenic photosynthetic bacteria. The Netherlands: Kluwer Academic Publishers. pp. 885–914.
16. Wahlund TM, Conway T, Tabita FR (1996) Bioconversion of CO_2 to ethanol and other compounds. American Chemical Society Division of Fuel Chemistry 3: 1403-1405.
17. Yilmaz LS, Kontur WS, Sanders AP, Sohmen U, Donohue TJ, et al. (2010) Electron partitioning during light- and nutrient-powered hydrogen production by *Rhodobacter sphaeroides*. Bioenerg Res Volume: 55 - 66.
18. Eraso JM, Kaplan S (1994) *prrA*, a putative response regulator involved in oxygen regulation of photosynthesis gene expression in *Rhodobacter sphaeroides*. J Bacteriol 176: 32-43.
19. Eraso JM, Roh JH, Zeng X, Callister SJ, Lipton MS, et al. (2008) Role of the global transcriptional regulator PrrA in *Rhodobacter sphaeroides* 2.4.1: combined transcriptome and proteome analysis. J Bacteriol 190: 4831-4848.
20. Eraso JM, Kaplan S (1995) Oxygen-insensitive synthesis of the photosynthetic membranes of *Rhodobacter sphaeroides*: a mutant histidine kinase. J Bacteriol 177: 2695-2706.

21. Dangel AW, Tabita FR (2009) Protein-protein interactions between CbbR and RegA (PrrA), transcriptional regulators of the cbb operons of *Rhodobacter sphaeroides*. Mol Microbiol 71: 717-729.
22. Laguri C, Phillips-Jones MK, Williamson MP (2003) Solution structure and DNA binding of the effector domain from the global regulator PrrA (RegA) from *Rhodobacter sphaeroides*: insights into DNA binding specificity. Nucleic Acids Res 31: 6778-6787.
23. Dufour YS, Kiley PJ, Donohue TJ (2010) Reconstruction of the core and extended regulons of global transcription factors. PLoS Genet 6: e1001027.
24. Zeilstra-Ryalls JH, Kaplan S (1998) Role of the *fnrL* gene in photosystem gene expression and photosynthetic growth of *Rhodobacter sphaeroides* 2.4.1. J Bacteriol 180: 1496-1503.
25. Bruscella P, Eraso JM, Roh JH, Kaplan S (2008) The use of chromatin immunoprecipitation to define PpsR binding activity in *Rhodobacter sphaeroides* 2.4.1. J Bacteriol 190: 6817-6828.
26. Gomelsky M, Kaplan S (1995) Genetic evidence that PpsR from *Rhodobacter sphaeroides* 2.4.1 functions as a repressor of *puc* and *bchF* expression. J Bacteriol 177: 1634-1637.
27. Mank NN, Berghoff BA, Hermanns YN, Klug G (2012) Regulation of bacterial photosynthesis genes by the small noncoding RNA PcrZ. Proc Natl Acad Sci U S A 109: 16306-16311.
28. Gomelsky M, Kaplan S (1995) *appA*, a novel gene encoding a trans-acting factor involved in the regulation of photosynthesis gene expression in *Rhodobacter sphaeroides* 2.4.1. J Bacteriol 177: 4609-4618.
29. Berghoff BA, Glaeser J, Sharma CM, Vogel J, Klug G (2009) Photooxidative stress-induced and abundant small RNAs in *Rhodobacter sphaeroides*. Mol Microbiol 74: 1497-1512.
30. Ranson-Olson B, Jones DF, Donohue TJ, Zeilstra-Ryalls JH (2006) *In vitro* and *in vivo* analysis of the role of PrrA in *Rhodobacter sphaeroides* 2.4.1 *hemA* gene expression. J Bacteriol 188: 3208-3218.
31. Willett J, Smart JL, Bauer CE (2007) RegA control of bacteriochlorophyll and carotenoid synthesis in *Rhodobacter capsulatus*. J Bacteriol 189: 7765-7773.
32. Myers KS, Yan H, Ong IM, Chung D, Liang K, et al. (2013) Genome-scale analysis of *Escherichia coli* FNR reveals complex features of transcription factor binding. PLoS Genet 9: e1003565.
33. Park DM, Akhtar MS, Ansari AZ, Landick R, Kiley PJ (2013) The bacterial response regulator ArcA uses a diverse binding site architecture to regulate carbon oxidation globally. PLoS Genet 9: e1003839.
34. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, et al. (2004) The Pfam protein families database. Nucleic Acids Res 32: D138-141.
35. Addlesee HA, Fiedor L, Hunter CN (2000) Physical mapping of *bchG*, orf427, and orf177 in the photosynthesis gene cluster of *Rhodobacter sphaeroides*: functional assignment of the bacteriochlorophyll synthetase gene. J Bacteriol 182: 3175-3182.
36. Addlesee HA, Hunter CN (1999) Physical mapping and functional assignment of the geranylgeranyl-bacteriochlorophyll reductase gene, *bchP*, of *Rhodobacter sphaeroides*. J Bacteriol 181: 7248-7255.
37. Oh JI, Eraso JM, Kaplan S (2000) Interacting regulatory circuits involved in orderly control of photosynthesis gene expression in *Rhodobacter sphaeroides* 2.4.1. J Bacteriol 182: 3081-3087.
38. Comolli JC, Carl AJ, Hall C, Donohue T (2002) Transcriptional activation of the *Rhodobacter sphaeroides* cytochrome c(2) gene P2 promoter by the response regulator PrrA. J Bacteriol 184: 390-399.
39. Mouncey NJ, Kaplan S (1998) Cascade regulation of dimethyl sulfoxide reductase (*dor*) gene expression in the facultative phototroph *Rhodobacter sphaeroides* 2.4.1T. J Bacteriol 180: 2924-2930.
40. Ziegelhoffer EC, Donohue TJ (2009) Bacterial responses to photo-oxidative stress. Nat Rev Microbiol 7: 856-863.
41. Todd JD, Wexler M, Sawers G, Yeoman KH, Poole PS, et al. (2002) RirA, an iron-responsive regulator in the symbiotic bacterium *Rhizobium leguminosarum*. Microbiology 148: 4059-4071.

42. Yeoman KH, Curson AR, Todd JD, Sawers G, Johnston AW (2004) Evidence that the *Rhizobium* regulatory protein RirA binds to cis-acting iron-responsive operators (IROs) at promoters of some Fe-regulated genes. Microbiology 150: 4065-4074.
43. Sistrom WR (1960) A requirement for sodium in the growth of *Rhodopseudomonas spheroides*. J Gen Microbiol 22: 778-785.
44. Cohen-Bazire G, Sistrom WR, Stanier RY (1957) Kinetic studies of pigment synthesis by non-sulfur purple bacteria. J Cell Physiol 49: 25-68.
45. Lowry OH, Rosebrough NJ, Farr AL, Randall RJ (1951) Protein measurement with the Folin phenol reagent. J Biol Chem 193: 265-275.
46. Schafer A, Tauch A, Jager W, Kalinowski J, Thierbach G, et al. (1994) Small mobilizable multi-purpose cloning vectors derived from the *Escherichia coli* plasmids pK18 and pK19: selection of defined deletions in the chromosome of *Corynebacterium glutamicum*. Gene 145: 69-73.
47. Ind AC, Porter SL, Brown MT, Byles ED, de Beyer JA, et al. (2009) Inducible-expression plasmid for *Rhodobacter sphaeroides* and *Paracoccus denitrificans*. Appl Environ Microbiol 75: 6613-6615.
48. Dufour YS, Imam S, Koo BM, Green HA, Donohue TJ (2012) Convergence of the transcriptional responses to heat shock and singlet oxygen stresses. PLoS Genet 8: e1002929.
49. Schilke BA, Donohue TJ (1995) ChrR positively regulates transcription of the *Rhodobacter sphaeroides* cytochrome c2 gene. J Bacteriol 177: 1929-1937.
50. Tavano CL, Podevels AM, Donohue TJ (2005) Identification of genes required for recycling reducing power during photosynthetic growth. J Bacteriol 187: 5249-5258.
51. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185-193.
52. Smyth G (2004) Applications in genetics and molecular biology 3: Berkeley Electronic Press.
53. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate - a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B-Methodological 57: 289-300.
54. Dufour YS, Landick R, Donohue TJ (2008) Organization and evolution of the biological response to singlet oxygen stress. J Mol Biol 383: 713-730.
55. Li R, Yu C, Li Y, Lam TW, Yiu SM, et al. (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25: 1966-1967.
56. Kuan PF, Chung D, Pan G, Thomson JA, Stewart R, et al. (2011) A statistical framework for the analysis of ChIP-seq data. Journal of the American Statistical Association 106: 891-903.
57. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, et al. (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37: W202-208.
58. Homann OR, Johnson AD (2010) MochiView: versatile software for genome browsing and DNA motif analysis. BMC Biol 8: 49.
59. Longabaugh WJ, Davidson EH, Bolouri H (2009) Visualization, documentation, analysis, and communication of large-scale gene regulatory networks. Biochim Biophys Acta 1789: 363-374.

Chapter 7

Global analysis of the regulation of central carbon and energy metabolism in α-Proteobacteria

This chapter is formatted as a manuscript for publication:

Imam S and Donohue TJ.

I performed all the experiments and analyses in this chapter.

Abstract

Many pathways of central carbon and energy metabolism are conserved across the phylogeny but the systems that regulate their expression and activity vary considerably among organisms. In this work, we analyzed two previously uncharacterized transcription factors (TFs) that are predicted to regulate genes encoding enzymes of central carbon and energy metabolism in the α-Proteobacterium *Rhodobacter sphaeroides*. Using a combination of *in vivo* and *in vitro* assays, we show that the LacI family TF CcmR (RSP_1663) is required for normal growth on many carbon sources. We found that CcmR is a global regulator that directly controls the transcription of genes encoding enzymes in central carbon and energy metabolism and that DNA binding by this TF is inhibited by the Entner-Doudoroff pathway intermediate 6-phosphogluconate. Mutational analysis of a predicted CcmR DNA binding site identified conserved bases that are important of activity of this TF. In addition, comparative genomics analysis predicts that orthologs of CcmR are found in several other α-Proteobacteria, where they appear to have a conserved function in controlling the expression of enzymes in central carbon and energy metabolism. We also showed that the GntR family TF, AkgR (RSP_0981), is a direct transcriptional regulator of genes encoding tricarboxylic acid (TCA) cycle enzymes. We identified growth conditions under which AkgR activity is increased and required for normal growth. AkgR is also conserved in several α-Proteobacteria, where it appears to directly regulate transcription of genes encoding TCA cycle enzymes. The properties of CcmR and AkgR mutants illustrate that these two TFs act co-operatively to regulate the flux through central carbon and energy metabolism. These data represent the first experimental characterization of TFs controlling central carbon and energy metabolism in α-Proteobacteria. Given the metabolic versatility and biotechnological potential of this group of organisms, these findings should aid future efforts to engineer improved strains of these bacteria.

Introduction

To survive in different environments, cells need to be able to adapt to a changing spectrum of accessible nutrients. Many bacteria respond to changes in nutrient availability through the use of networks of signaling and transcriptional regulatory components that sense metabolites and orchestrate changes in gene expression or protein activity states, to allow effective utilization of these nutrients [1, 2]. Thus, gaining new insights into these networks is required for an improved understanding of cellular physiology and can be crucial in efforts to develop improved bacterial hosts for production of foods, chemicals or other products.

The transcriptional regulatory networks (TRNs) controlling carbon utilization have been studied extensively in bacteria such as *Escherichia coli* and *Bacillus subtilis*. *E. coli* preferentially utilizes glucose over other carbon sources, if present in its environment, and the regulation of this process of catabolite repression is partly controlled by cAMP receptor protein (CRP), which is a global regulator of several processes in addition to central carbon metabolism pathways [3]. Furthermore, the ArcAB two-component system represses portions of *E. coli*'s central metabolic pathways under anaerobic respiratory conditions [4, 5]. In addition to these global regulators, Cra/FruR – a LacI-family regulator, regulates carbon and energy metabolism in *E. coli* and related enteric bacteria [6-8]. In *B. subtilis* the transcription factor (TF) CcpA functions as a global regulator of carbon metabolism, mediating catabolite repression in the presence of glucose, activating the expression of glycolytic genes while repressing transcription of many genes encoding proteins involved in central metabolism or the use of alternative carbon sources [9, 10]. In some β- and γ-proteobacteria, regulation of central carbon metabolism is controlled by the RpiR family TF HexR, which is often co- or divergently transcribed with a glucose metabolism operon [11-13]. However, to date no transcriptional regulator of central carbon metabolism has been characterized in other bacteria.

α-Proteobacteria are a group of metabolically versatile and well-studied bacteria which have numerous activities or pathways not found in many well-studied systems (e.g., photosynthesis, carbon dioxide fixation, etc.) [14-16]. *Rhodobacter sphaeroides* represents one of the best studied α-Proteobacteria from biochemical, genetic and genomic perspectives. This facultative bacterium is capable of growing by aerobic respiration, anaerobic respiration and anoxygenic photosynthesis [15, 17, 18]. Furthermore, *R. sphaeroides* can utilize a wide array of nutrients for growth, including at least 68 different carbon sources [19]. Despite this metabolic versatility, transcriptional regulators of carbon metabolism have not been studied in *R. sphaeroides* and other α-proteobacteria.

A recent large-scale reconstruction of the *R. sphaeroides* TRN provided predictions of TFs that regulate carbon metabolism. In particular, this TRN identified several co-regulated clusters of genes that are enriched for functions involved in central carbon metabolism (Imam et al. submitted). Of these clusters, two were predicted to be regulated by previously uncharacterized TFs, RSP_1663 (a LacI family TF hereafter referred to as *c*entral *c*arbon *m*etabolism *r*egulator, CcmR) and RSP_0981 (a GntR family TF hereafter refer to as *a*lpha-*k*eto*g*lutarate regulator, AkgR), orthologs of which are present in many other α-Proteobacteria (Imam et al. submitted). This large-scale TRN predicts that CcmR regulates genes involved in glycolysis, the tricarboxylic acid (TCA) cycle and other aspects of central carbon or energy metabolism, while AkgR was predicted to regulate some TCA cycle genes.

In this work, we investigated the role of CcmR and AkgR using a combination of physiological, biochemical, genetic and genomic assays. We show that CcmR directly and indirectly regulates transcription of genes involved in carbon and energy metabolism, including those encoding enzymes of the Entner-Doudoroff (ED) pathway, TCA cycle, ATPase and others. We find that CcmR is required for normal growth on many carbon sources. We predicted a DNA sequence motif recognized by RSP_1663 and verified the role of bases in this region and the ED pathway intermediate, 6-phosphogluconate, in DNA binding activity of this TF. We also show that AkgR directly regulates transcription of genes encoding proteins that catalyze 3 reactions of the TCA cycle and that this TF is important for normal

growth on carbon sources which require high metabolic flux through these reactions. Using comparative genomics, we predict conservation and differences in members of the CcmR and AkgR regulons across other α-Proteobacteria. Overall, our data provide the first genome-level view on the regulation of central carbon and energy metabolism in *R. sphaeroides* and related α-Proteobacteria.

Results

CcmR is a regulator of central carbon and energy metabolism in *R. sphaeroides*

A reconstruction of the *R. sphaeroides* TRN predicted that the LacI family CcmR is a regulator of central carbon metabolism in this and several other α-Proteobacteria (Imam et al. submitted). Below, we use a combination of genetic, physiological and genomic approaches to study the role of CcmR in *R. sphaeroides*.

CcmR is required for normal growth on a large number of carbon sources

If CcmR is a regulator of central carbon and energy metabolism, deletion of this gene might be expected to have a significant impact on *R. sphaeroides* growth with different carbon sources. To test this hypothesis, we constructed an in-frame markerless deletion of *ccmR* (ΔCcmR) and assessed its growth phenotypes. Under aerobic conditions in Sistrom's minimal medium (SMM) [20], in which succinate is the main carbon source, ΔCcmR grew about 1.5 times slower than its wild type (WT) parent (Figure 7-1A), while complementation of this mutant with *ccmR* from an inducible plasmid restored its growth rate to a level similar to that of the WT. Under photosynthetic conditions, the ΔCcmR strain had a doubling time of more than 3 times slower than WT cells, with a normal growth rate restored in the complemented strain (Figure 7-1B). These data indicated that CcmR had an important role, particularly under photosynthetic conditions, when using succinate as the main carbon source.

When we compared growth of the ΔCcmR mutant to WT cells across 25 carbon sources under aerobic conditions, we observed significantly slower growth of this mutant on 15 of these compounds (Figure 7-1C). This indicates that CcmR may regulate genes required for metabolism of several carbon sources. The severest impact on aerobic growth of the ΔCcmR mutant was observed on carboxylic and amino acids, with up to a 6-fold difference in growth rate between ΔCcmR and WT cells when using pyruvate as a carbon source. The only carboxylic acids that permitted normal growth of the ΔCcmR strain were acetate

and L-tartarate (Figure 7-1C). On the other hand the observed growth rates of ΔCcmR and WT cells were generally similar on all the sugars we tested. These data indicate that while CcmR regulates genes important for the metabolism of many carbon sources, other TFs might control metabolism of some select carbon sources such as acetate, consistent with predictions of the large-scale TRN (Imam et al. submitted).

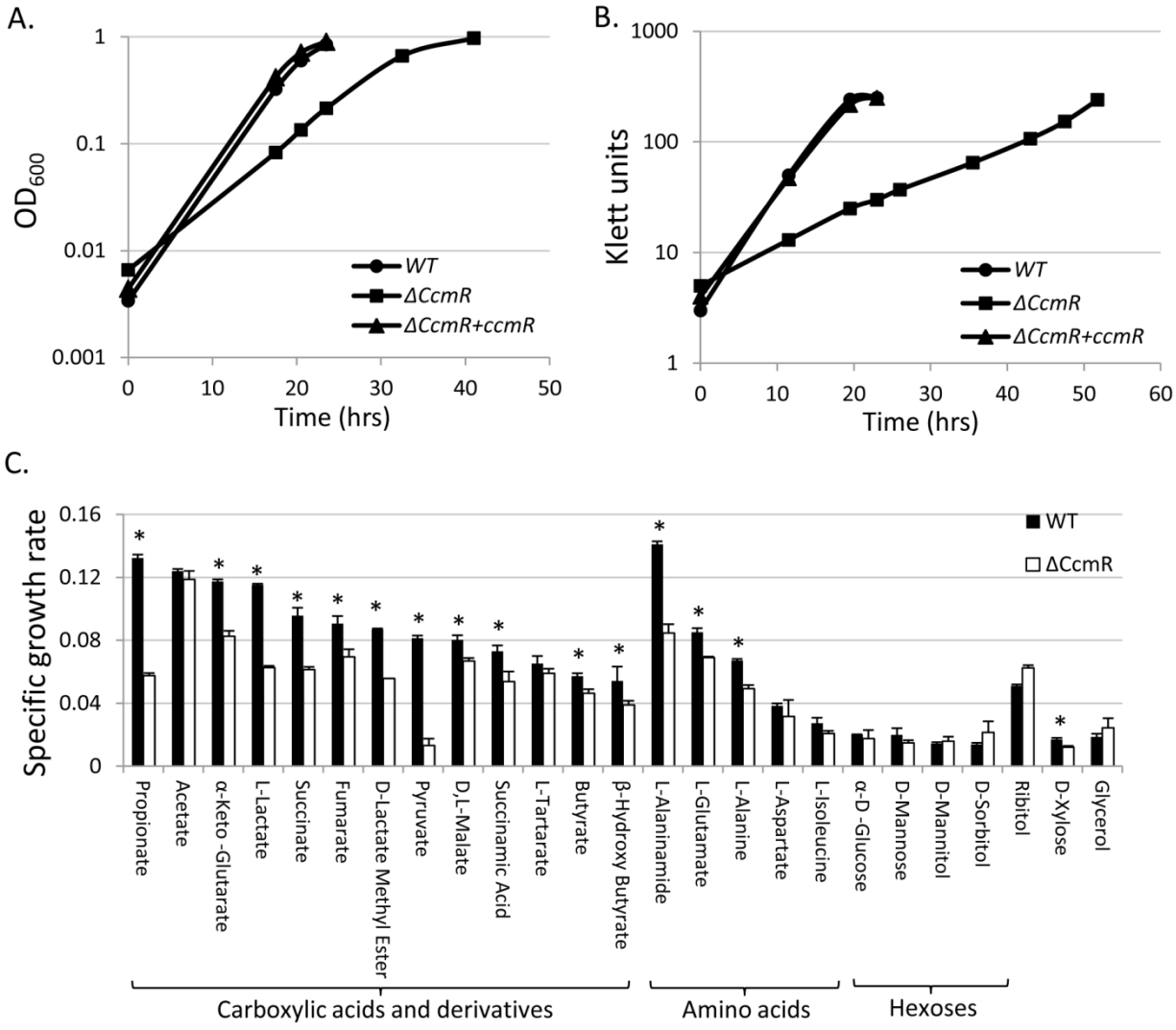

Figure 7-1. Growth phenotypes of ΔCcmR. Growth of WT, ΔCcmR and ΔCcmR cells complemented with *ccmR* from an IPTG-inducible plasmid (ΔCcmR + *ccmR*) on SMM under aerobic (A) and photosynthetic (B) conditions. Doubling times of WT, ΔCcmR and ΔCcmR + *ccmR* were 3.24 (3.22), 4.74 (10.35) and 2.95 (3.14) hours respectively under aerobic (photosynthetic) conditions (C) Specific growth rates of WT and ΔCcmR strains on 25 individual carbon sources under aerobic respiratory conditions. Error bars represent standard error from 3 independent replicates. * Indicate carbon sources on which WT grew significantly faster than the ΔCcmR mutant.

CcmR regulates the expression of a wide array of metabolic pathways

Given the role of CcmR in utilization of multiple carbon sources, we used global gene expression analyses to identify genes, directly or indirectly, regulated by CcmR during aerobic growth. We found that transcript levels from a total of 225 genes were differentially expressed (DE) between the WT and ΔCcmR (cutoff 1.5 fold change (FC), $p < 0.01$) (Supplemental Table S1) during aerobic growth with SMM, with transcripts from 125 genes present at higher levels and 100 decreased in WT cells relative to the mutant. About 54 of these DE genes are annotated as being involved in carbon metabolism, transport, electron transport and energy metabolism. These included ~31 genes involved in central carbon and energy metabolism (Figure 7-2A) including genes encoding enzymes of the Entner-Doudoroff (ED) pathway (*zwf*, *pgl* and *eda*), the TCA cycle (*sdhDA*, *mdh*, *fumC*, *sucCD* and *sucB*), glycolysis/gluconeogenesis (*pgi*, *fbaB*, fructose1,6 - bisphosphatase, *pdhAB* and *pckA*) and other aspects of carbon or energy metabolism (*atpBEF*, *atpHAC*, *nuoL*, cytochrome c-554 and malic enzyme) (Figure 7-2A). RNA levels of the vast majority of these genes were more abundant in the WT strain than ΔCcmR (i.e., suggesting they are activated by CcmR), with only those encoding the ED pathway enzymes (*zwf*, *pgl* and *eda*) and RSP_2785 (cytochrome c-554) less abundant in WT cells (suggesting they are repressed by CcmR) (Figure 7-2A). These observations are consistent with the predicted role of CcmR in regulating genes that encode enzymes involved in carbon and energy metabolism.

Most of the other genes that were DE in a CcmR-dependent manner encoded functions either relating to protein or photopigment biosynthesis (Supplemental Table S1). The differential expression of protein synthesis related genes, such as those encoding tRNAs and ribosomal proteins, is likely a reflection of the difference in growth rates between the WT and ΔCcmR strains under aerobic growth conditions (Figure 7-1A). On the other hand, the photopigment related DE genes may reflect an altered oxidation-reduction state of the cell or electron transport chain in the absence of the *ccmR* gene. For instance, the altered expression of genes encoding electron transport proteins in ΔCcmR cells could change the redox state of the cell and activate the PrrAB two component system, which has previously been proposed to activate

expression of photosynthesis related genes in response to changes in the cellular redox state [21]. The increased aerobic pigmentation of the ΔCcmR strain was also rescued in complemented strain, indicating that it was also a result of the loss of CcmR function.

CcmR binds to the promoters of key central carbon metabolism genes

While the above data predict a role for CcmR in some aspect of central carbon metabolism, we wanted to determine direct targets for this TF. To do this, we assessed CcmR DNA binding across the genome using chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) with a 3X myc tagged version of CcmR that complements the phenotype of the ΔCcmR mutant. This analysis identified a total of 19 CcmR binding sites across the genome (Table 7-1). These CcmR binding sites were located in the upstream regions of 16 operons that encode a total of 32 genes. Consistent with the observations from the global gene expression analysis, CcmR was bound upstream of a variety of genes involved in carbon and energy metabolism including the glycolytic genes (*zwf*, *pgl*, *pgi*, *eda* and *fbaB*), the ATP synthase operons (*atpHAGDC* and *atpBEXF*), genes encoding subunits of the succinate dehydroganse complex (*sdhAB*), the phosphoenolpyruvate carboxykinase gene (*pckA*), as well as its own promoter (Figure 7-2B, Table 7-1). Of the 16 operons bound by CcmR, transcript levels from 5 were significantly increased while transcripts from 2 were decreased in the presence of CcmR (Table 7-1), indicating that CcmR functions both as an activator and a repressor. The operons predicted to be directly repressed by CcmR encoded genes in the ED pathway (*zwf*, *pgl* and *eda*) as well as *pgi*. All other central metabolic genes directly regulated by CcmR were apparently activated by this TF, since transcript levels for these genes were reduced in the ΔCcmR mutant. In addition to these known metabolic genes, CcmR was bound upstream of genes (RSP_6037, RSP_6082 and RSP_7376) encoding proteins of unknown functions, which could indicate these proteins might have unknown roles in carbon metabolism.

CcmR was also bound upstream of *rpoH1* (the σ-factor known to mediate heat shock response in *R. sphaeroides* [22, 23]) and *flgM* (the anti-sigma factor to *fliA* [24]) (Table 7-1, Figure 7-2B). This might

indicate that the regulatory influence of CcmR extends into chemotaxis and stress responses, however neither of these genes were DE in a CcmR-dependent manner in our global gene expression datasets. Thus, conditions under which these CcmR binding sites might be functional, if any, remain unresolved.

Conducting *de novo* motif detection analysis on the sequences bound by CcmR in ChIP-seq assays revealed the presence of a shared inverted repeat sequence of [T/C]GTT N_6 AAC[A/T] (Figure 7-2C) that is very similar to the CcmR binding site predicted from the large-scale *R. sphaeroides* TRN (Imam et al. submitted). This putative CcmR DNA binding site bears some sequence similarity to one that is proposed to be recognized by HexR [11], but shares little similarity to the predicted Cra-box [8].

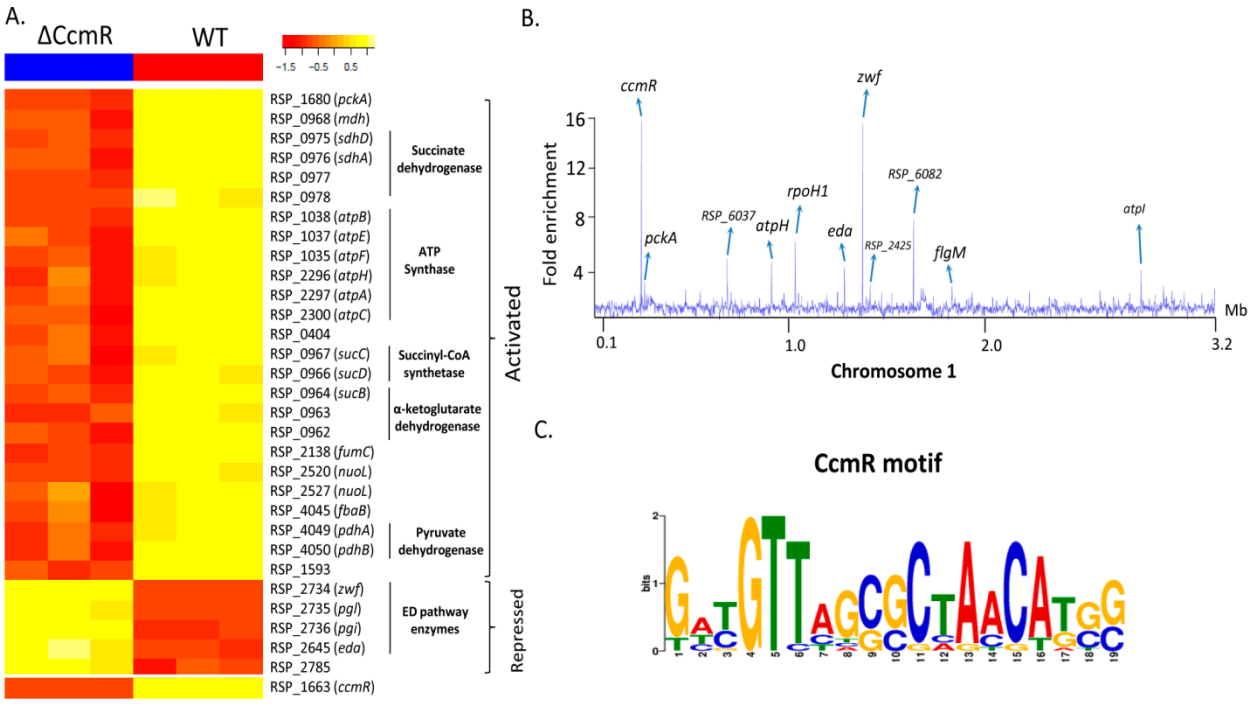

Figure 7-2. Genome-wide analysis of the role of CcmR. (A) Heatmap highlighting DE metabolic genes between WT and ΔCcmR cells under aerobic conditions. (B) Binding profile of CcmR on *R. sphaeroides* chromosome 1. The gene names of only the first members of each operon are indicated. (C) Predicted CcmR binding motif shared by sites identified in the ChIP-seq analysis.

Table 7-1. CcmR binding sites across the R. sphaeroides genome identifed by ChIP-seq

s/n	Gene ID	Annotation	ChrID	Start[a]	Stop[a]	FC[b]	Predicted sequence	Reg[c]
1	RSP_0037	*flgM*	chr1	1742400	1742799	2.4	GA**GGTT**CGGCCT**AACA**TGC	
2	RSP_0974-9	Succinate dehydrogenase complex (*sdh*)	chr1	2734600	2735000	1.7	CA**TGAT**TGCGCA**AACA**TGT	+
3	RSP_1039-35	F0F1 ATP synthase (*atpIBEXF*)	chr1	2800800	2801799	5.3	GA**TGTT**CGTGGT**GCAT**TCG	+
4	RSP_1663	*ccmR*	chr1	255200	256799	18.7	GC**CGTT**ATCGCT**AACA**TGG	NA
5	RSP_1680	*pckA*	chr1	272000	272599	5.4	GA**CGTT**AGCGCA**ACCA**GCG	+
6	RSP_2296-2300	F0F1 ATP synthase (*atpHAGDC*)	chr1	918600	919399	7.3	GA**TGTT**AGCGGC**ACCA**TCG	+
7	RSP_2410	RNA polymerase factor sigma-32 (*rpoH1*)	chr1	1039000	1039399	8	GA**CGTC**AGCGCT**ACCA**TGG	
8	RSP_2425	putative CarD-like transcriptional regulator	chr1	1056000	1056400	2.4	GA**TTTT**CTAGCG**ACCA**TTC	
9	RSP_2646-5	edd, eda	chr1	1289400	1290199	6.7	TT**TGTT**AGCGCT**AACT**AGC	-
10	RSP_2734-6	*zwf, pgl, pgi*	chr1	1382800	1383399	16.9	GC**CGTT**AGCGCT**AACA**GGC	-
11	RSP_3896	*repC*	plasmidA	89600	90200	7.6	GC**AGTT**TCTGCG**AACA**AGG	
12	RSP_3926	UDP-glucuronate 5'-epimerase	plasmidA	100200	100600	2.5	CA**TGAT**AGCGCA**AACA**TCG	
13	RSP_4045	fructose-bisphosphate aldolase (*fbaB*)	chr1	1127000	1127400	2.3	GT**TGTT**CGCGCT**AACT**TGG	+
14	RSP_6037	hypothetical protein	chr1	691800	692999	7.1	GC**TGTT**AGGGCA**AACA**TGG	ND
15	RSP_6082	hypothetical protein	chr1	1642200	1642599	8.6	GT**TGTT**AGCGCC**AACA**TTC	ND
16	RSP_7376	hypothetical protein	plasmidC	52600	52799	2.9	GA**TGTT**CCGCCT**AACA**GCG	ND
17			chr2	704600	704999	2.6	GA**GGTC**CGTGGC**AACA**TGG	ND
18			plasmidD	29800	30200	3.9	TA**CGTT**TAGCCT**AACA**GCG	ND
19			plasmidC	54000	54400	2.7	GT**TGTT**AGCCCT**GAGA**TCC	ND

[a] Chromosomal locations of start and stop of ChIP-seq peaks.
[b] Fold enrichment of CcmR-myc ChIP over control myc antibody ChIP in WT control.
[c] Regulatory role of CcmR on target operons based on change in gene expression between WT and ΔCcmR cells. + = positively regulated by CcmR. - = negatively regulated by CcmR. NA - Not applicable. ND - Not determined (i.e., these genes are not represented on the *R. sphaeroides*).

Bases in the predicted consensus motif are required for CcmR binding

To test if the predicted CcmR binding motif contains sequence elements recognized by this TF, we analyzed CcmR binding at target promoters using an electrophoretic mobility shift assay (EMSA) with WT and mutated DNA sequences. When we assessed the ability of purified CcmR to bind specifically to target sequences identified from our ChIP-seq analysis, we found it was able to bind with high affinity to DNA fragments containing the predicted binding motif obtained from the *R. sphaeroides ccmR* and *zwf* promoters, with increasing concentrations of CcmR protein increasing the amounts of bound DNA (Figure 7-3A and B). Based on the assumption that CcmR binds DNA as a tetramer, as is the case for other LacI family members [25, 26], these data indicate that CcmR binds the *ccmR* and *zwf* promoters with an apparent micromolar affinity. On the other hand, no detectable CcmR binding was observed with a DNA fragment from the *bchF* promoter (Figure 7-3C), which was not identified as a CcmR target in the ChIP-seq data. This indicates that the high-affinity CcmR binding observed at the *ccmR* and *zwf* promoter sequences were due to sequence specific protein-DNA interactions.

We then made specific point mutations in the predicted CcmR binding site of the *ccmR* promoter to assess the impact of base substitutions on DNA binding using this assay. When we substituted the conserved thymidine at position 6 and adenines at position 13 and 14 of the predicted CcmR motif (Figure 7-2C, corresponding to positions -92, -85 and -84 relative to the *ccmR* start codon) with guanines, each of these mutations had a detrimental effect on CcmR binding (Figure 7-3D-F). The T6G and A13G mutations resulted in virtually complete loss of CcmR binding activity at all protein concentrations tested, indicating that these bases are critical for sequence recognition by CcmR at the *ccmR* promoter (Figure 7-3D and E). While the A14G mutation also resulted in significant impairment of CcmR binding, at higher protein concentrations, DNA binding was observed to this mutant DNA template (Figure 7-3F), suggesting that CcmR DNA binding activity is not completely nullified by this mutation. Based on the conservation of bases in the predicted consensus CcmR motif (Figure 7-2C), the adenine at position 14 has less information content than either T6 or A13, so the reduced impact on CcmR binding of the A14G

compared to the T6G or A13G mutations is in agreement with the predicted properties of the CcmR binding site. In sum, these data indicate that the conserved sequences predicted from our ChIP-seq analysis are important for CcmR DNA binding, so they likely represent a consensus binding site for this TF.

6-phosphogluconate inhibits CcmR DNA binding activity

LacI family TFs are generally composed of an N-terminal helix-turn-helix DNA binding domain and a C-terminal sugar or effector binding domain (EBD) that is capable of binding specific ligands [26]. In well-studied members of this family, interactions between a ligand(s) and the EBD results in a conformational change of the TF, modulating its DNA binding activity in response to changing cellular metabolic states [26]. To test for ligands potentially recognized by the CcmR EBD, we assayed the effects of central carbon metabolism intermediates 6-phosphogluconate (6-PG), 2-Keto-3-deoxy-6-phosphogluconate (KDPG), phosphoenolpyruvate (PEP), glucose, glucose-6-phosphate (G6P) and fructose-1,6-bisphosphate (F16P), on CcmR binding to the *ccmR* promoter fragment by EMSA (Figure 7-3G and H). These assays showed that, of these metabolites, only 6-PG significantly inhibits CcmR DNA binding activity at the concentrations used (0.5mM to 10mM) (Figure 7-3G). We then assessed the impact of lower concentrations of 6-PG on CcmR DNA binding and observed inhibitory effects at 6-PG concentrations as low as 10 μM (Figure 7-3I). These data indicate that 6-PG has the ability to inhibit CcmR DNA binding at micromolar concentrations. Given that glucose is metabolized almost exclusively via the ED pathway in *R. sphaeroides* [19, 27], the observed inhibition of CcmR (a regulator of the genes in this pathway) DNA binding at what are likely to be physiologically relevant concentrations of a known intermediate (i.e., 6-PG), links activity of this TF to flux through the ED pathway.

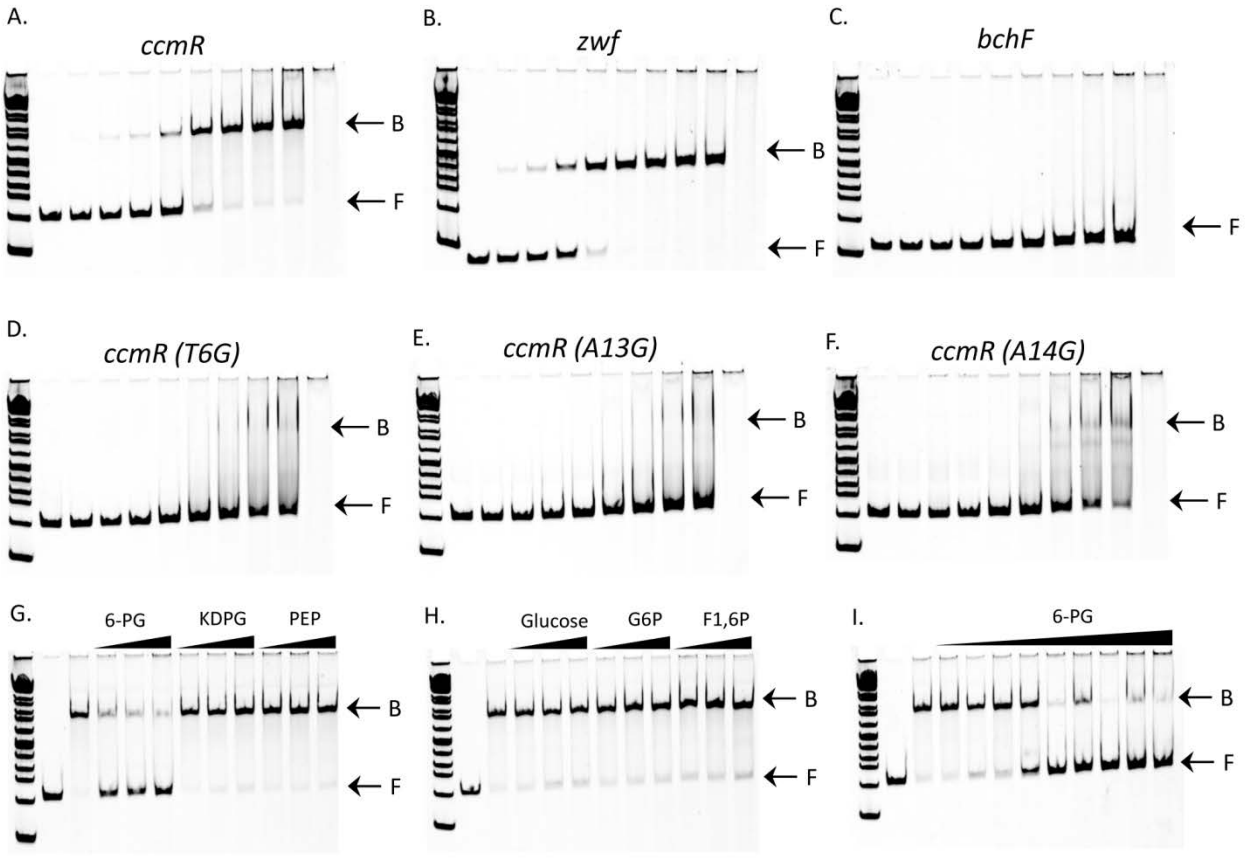

Figure 7-3. CcmR binding specificity. Purified CcmR was used for EMSA analysis with DNA fragments from the *ccmR* (A), *zwf* (B)and *bchF* (C) promoters. Approximately 0.05, 0.1 or 0.085 μM of *ccmR*, *zwf* or *bchF* DNA fragment respectively, was incubated with increasing amounts of CcmR in each experiment. (D-F) EMSA analysis using templates with indicated point mutations, i.e., T6G, A13G and A14G (or T-92G, A-85G and A-84G relative to the *ccmR* start codon respectively), in the predicted CcmR binding site of the *ccmR* promoter. For experiments in panels A-F samples were loaded as follows: lane 1 – 1kb DNA ladder; lane 2 – DNA only; lanes 3-10 – DNA + CcmR (0.12, 0.24, 0.48, 0.97, 1.45, 1.94, 2.4 and 2.9 μM respectively); lane 11 – CcmR only. B – bound DNA, F – free DNA. (G, H) Assessment of the effects of 6-PG, KDPG, PEP, glucose, G6P and F1,6P on CcmR binding to DNA fragment from the *ccmR* upstream regulatory region. Lane 1 – 1kb ladder; lane 2 – DNA only; lane 3 – DNA + CcmR. Three concentrations (0.5 mM, 2 mM and 10 mM) were tested for each metabolite, with the increasing concentration gradient highlight with the black triangle. Approximately 0.03 μM DNA and 1.94 μM CcmR were used in each reaction. (I) Assessment of the effect of a range of 6-PG concentrations on CcmR binding activity. Samples were loaded as follows: lanes 1-3 1kb ladder, DNA only and DNA + CcmR only respectively; lanes 4 – 10 DNA + CcmR + 6-PG (0.01μM, 0.1μM, 1μM, 10 μM, 50 μM, 100 μM, 0.5 mM, 2 mM and 10 mM respectively). Approximately 0.03 μM DNA and 1.94 μM CcmR were used in each reaction.

AkgR activates transcription of selected TCA cycle genes in *R. sphaeroides*

Another TF predicted to regulate transcription of genes encoding proteins involved in central carbon metabolism is the GntR family TF AkgR. We used a combination of physiological, transcriptomics and ChIP-seq analyses to assess the function of AkgR in *R. sphaeroides*.

AkgR is required for normal growth on small number of carbon sources

To assess the role of AkgR, we made an in-frame markerless deletion mutant of *akgR* (ΔAkgR) and compared its growth phenotypes to that of WT cells. The growth of ΔAkgR cells on SMM was equivalent to that of WT cells under both aerobic respiratory and anaerobic photosynthetic conditions (Figure S1), suggesting the regulatory role of AkgR under these conditions is limited. When we assessed the growth of ΔAkgR on 25 carbon sources under aerobic conditions, we found that it only exhibited a growth defect compared to WT cells on 4 carbon sources: α-ketoglutarate, propionate, pyruvate and acetate (Figure 7-4A), suggesting a more specific role for this TF compared to CcmR. Additional growth experiments with α-ketoglutarate as the sole carbon source, showed that under aerobic conditions ΔAkgR cells grew ~1.5 times slower than WT, while the photosynthetic doubling time of this mutant was ~2.5 times slower than its WT parent on this carbon source (Figure 7-4B).

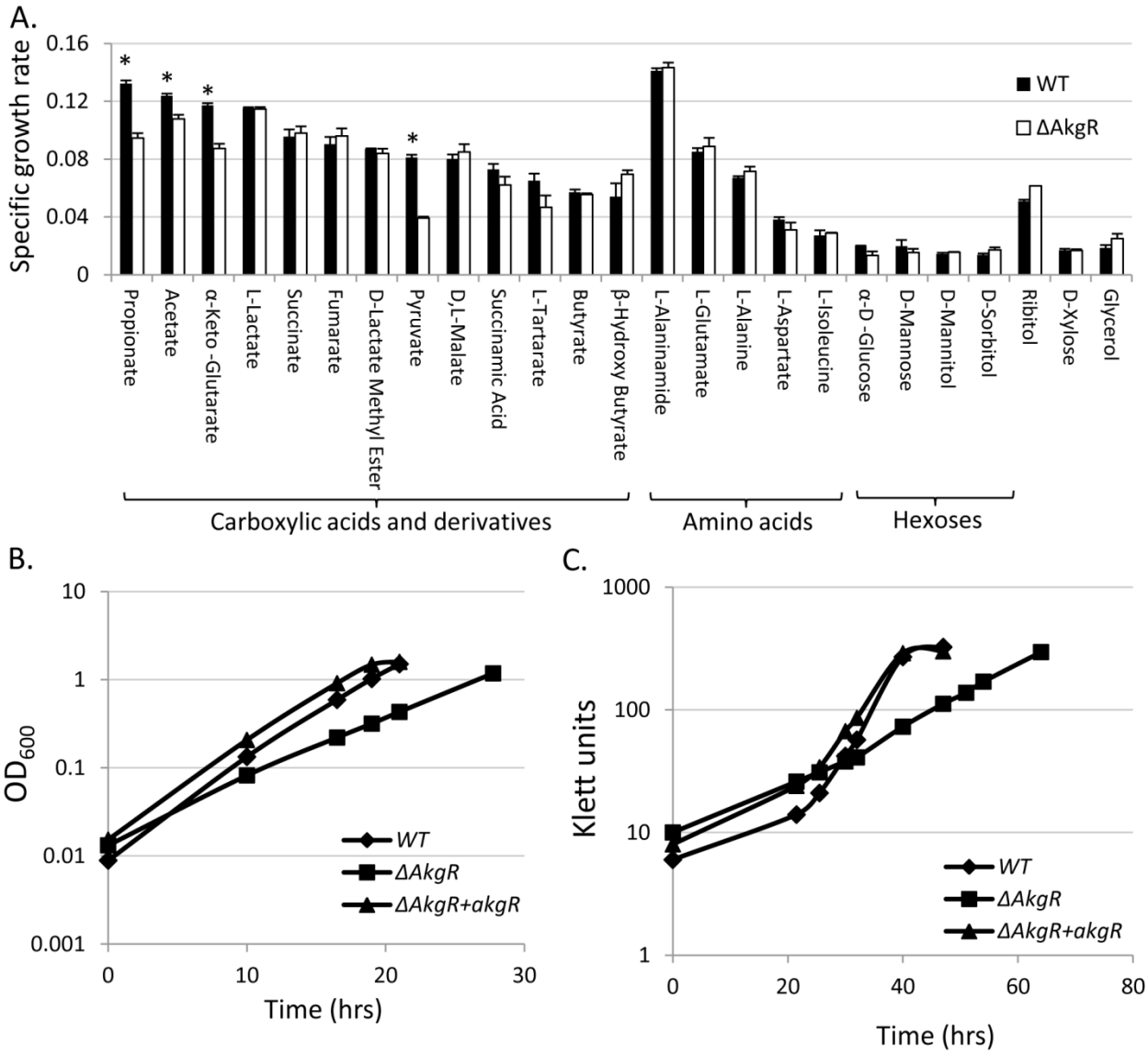

Figure 7-4. Growth of WT and ΔAkgR cells. (A) Specific growth rates of WT and ΔAkgR strains on 25 different carbon sources under aerobic conditions. Error bars represent standard error from 3 independent replicates. * Indicate carbon sources on which WT grew significantly faster than ΔAkgR cells. (B) Growth curves of WT, ΔAkgR and ΔAkgR complemented with *akgR* from an IPTG-inducible plasmid (ΔAkgR + *akgR*) on α-ketoglutarate under aerobic conditions. The doubling times of WT, ΔAkgR and ΔAkgR + *akgR* were 3. 03, 4.61 and 3.04 hrs respectively (C) Growth curves of WT, ΔAkgR and ΔAkgR + *akgR* on α-ketoglutarate under photosynthetic conditions. The doubling times of WT, ΔAkgR and ΔAkgR + *akgR* were 4.5, 10.6 and 4.6 hrs respectively.

AkgR activates the expression of TCA cycle genes and its expression is induced during growth on α-ketoglutarate

We investigated the transcriptional regulatory role of AkgR by conducting global gene expression analysis on WT and ΔAkgR grown aerobically on SMM, which contains succinate as the main carbon source. The only DE metabolic genes (cutoff 1.5 FC, $p < 0.01$) identified under this condition were genes in the RSP_0967-62 operon which include *sucD* (RSP_0966; succinyl-CoA synthetase α subunit) and *sucB* (RSP_0964; α-ketoglutarate dehydrogenase E2 component), each of which had ~1.8 fold higher transcript levels in the WT strain relative to ΔAkgR (Figure 7-5A). The transcript levels of *akgR* in the WT were also relatively low under this growth condition, being only ~1.3 fold above background. These modest changes in global transcript levels are consistent with the lack of a growth phenotype observed for ΔAkgR grown on SMM.

Ectopic expression of *akgR* from an IPTG-inducible plasmid revealed that AkgR can have a more significant impact on expression of TCA cycle genes. A comparison of the transcript levels between the AkgR ectopic expression strain (ΔAkgR + *akgR*) and the ΔAkgR mutant, showed that transcripts encoding the subunits of succinyl-CoA synthetase (RSP_0967-6), subunits of the succinate dehydrogenase complex (RSP_0974-9), and those in the α-ketoglutarate dehydrogenase complex (RSP_0965-2) were 2- to 3-fold higher in the ΔAkgR + *akgR* strain than ΔAkgR (Figure 7-5A). This indicates that AkgR is capable of directly or indirectly controlling the expression of genes encoding these 3 enzymes of the TCA cycle.

Given the impaired growth phenotype of ΔAkgR during aerobic growth on α-ketoglutarate (Figure 7-4A and B), we used qRT-PCR to assess the transcript levels of *akgR*, *sucA* and *sucC* in WT and ΔAkgR cells during growth on media containing either succinate (as a control) or α-ketoglutarate as the main carbon source (Figure 7-5B). We found that transcript levels of *sucA* and *sucC* were ~ 2.5 and 2-fold higher respectively in WT cells grown on α-ketoglutarate relative to those grown on succinate. These

observations are consistent with an increased need for succinyl-CoA synthetase (SucCD) and α-ketoglutarate dehydrogenase (SucAB) activity during growth on α-ketoglutarate. In addition, we observed that there was an ~1.5-fold increase in the *akgR* RNA levels during growth of WT cells on α-ketoglutarate relative to growth on succinate, suggesting that AkgR may be involved in activating transcription of these TCA cycle genes. Consistent with this, the transcript levels of both *sucA* and *sucC* were ~2 fold lower in ΔAkgR cells relative to WT cells grown on α-ketoglutarate (Figure 7-5B).

AkgR binds to the promoters of genes encoding three TCA cycle enzymes

To determine the genome-wide DNA binding locations of AkgR, we conducted ChIP-seq with a 3X myc tagged AkgR protein. We identified 9 genomic locations significantly enriched for AkgR binding (Table 7-2). These included high enrichment binding sites located upstream of the succinyl-CoA synthetase and α-ketoglutarate dehydrogenase complex operon (*sucCDAB*; RSP_0967-2) and the succinate dehydrogenase complex operon (*sdh*; RSP_0967-2) (Figure 7-5C). In addition to these regions, 7 other sites were found to be enriched for AkgR across the genome, but the physiological importance of these interactions is unknown, as no genes in the vicinity of these putative binding sites were DE in our gene expression datasets (Table 7-2).

De novo motif detection analysis of sequences under the AkgR binding sites revealed the presence of a shared DNA sequence motif, GTGATCAC (Table 7-2). Interestingly, the upstream regulatory regions of the succinyl-CoA synthetase/α-ketoglutarate dehydrogenase complex and the succinate dehydrogenase complex operons possessed 2 copies of this GTGATCAC motif with a 13 bp spacer region (Figure 7-5D). Given that these were the only 2 operons found to be DE in our global gene expression analyses, we propose that the motif recognized by AkgR at these operons is the inverted repeat sequence GTGATCAC N_{13} GTGATCAC. If this is true, then the 7 other binding locations identified in the genome-wide ChIP-seq analysis could represent non-functional AkgR half-sites or binding sites that require an additional TF that is not active under the conditions tested to activate transcription.

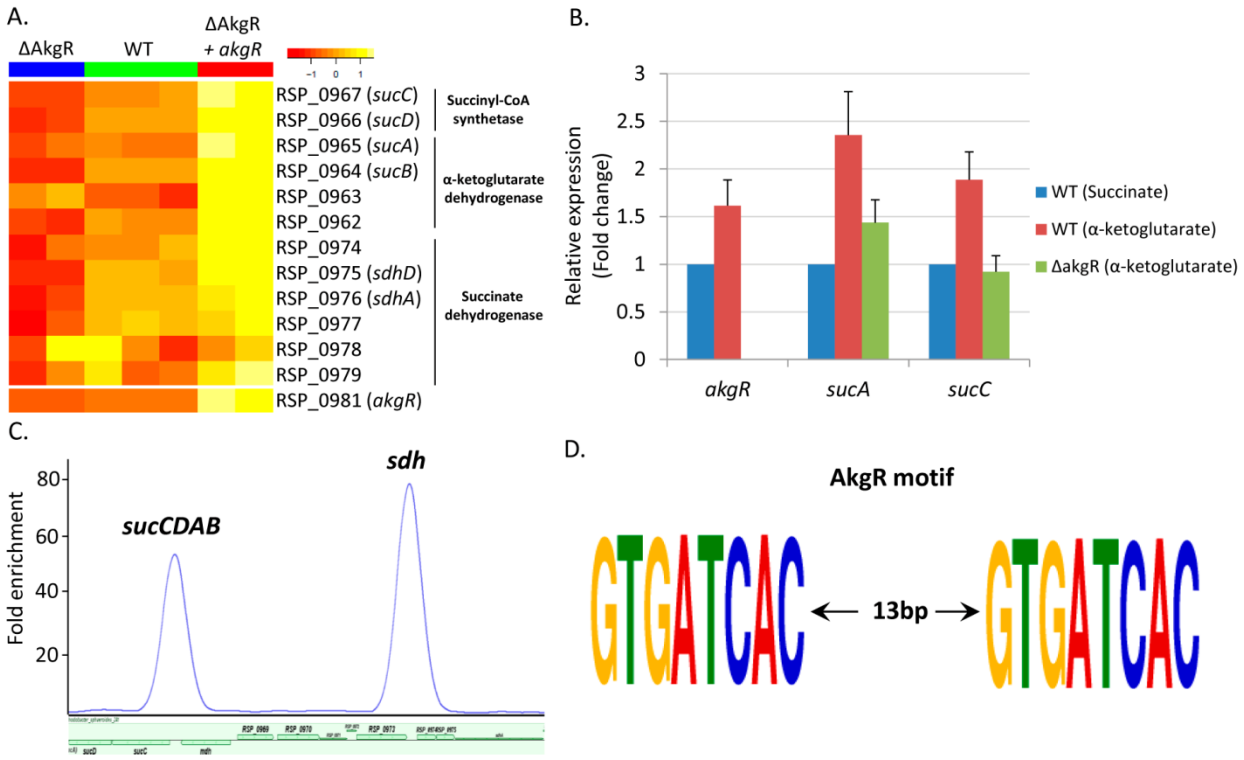

Figure 7-5. Genomic analysis of AkgR targets. (A) Heatmap highlighting DE metabolic genes between WT, ΔAkgR and ΔAkgR + *akgR* during aerobic growth on SMM. (B) qRT-PCR analysis of the transcript levels of *akgR*, *sucC* and *sucA* in WT and ΔAkgR cells during aerobic growth on succinate or α-ketoglutarate. Error bars represent standard error from 2 independent biological and 2 independent technical replicates. (C) AkgR ChIP-seq binding profile at *sucCDAB* and succinate dehydrogenase (*sdh*) promoter regions. (D) Predicted AkgR binding motif shared by ChIP-seq identified target sites.

Table 7-2. AkgR binding sites across the R. sphaeroides genome identifed by ChIP-seq

s/n	Gene ID	Annotation	ChrID	Start[a]	Stop[a]	FC[b]	Predicted sequence	Reg[c]
1	RSP_0967-2	*sucCDAB*	chr1	2729800	2730399	57.8	**GTGATCAC**GGGCTCGAAGGGT**GTGATCAC**	+
2	RSP_0974-9	*sdh*	chr1	2734600	2735199	84.4	**GTGATCAC**AGGCCCGCATCTT**GTGATCAC**	+
	Other binding sites							
3	RSP_1718	50S ribosomal protein L23	chr1	309200	309599	6.7	**GTGATCAC**	
4	RSP_3523	ABC peptide transporter	chr2	603000	603399	3.8	**GTGATCAC**	
5	RSP_2434	Putative MCP methyltransferase CheR1	chr1	1067600	1068199	4	**GTGATCAC**	
6	RSP_0386	Cold-shock DNA-binding domain protein	chr1	2119200	2119599	2.8	**GTTATCAC**	
7			chr1	950000	950599	7.8	**GTGATCAC**	
8			chr2	588000	588599	3.1	**GTGATCAC**	
9			chr1	3136400	3136799	2.9	**GTGATCAC**	

[a] Chromosomal locations of start and stop of ChIP-seq peaks.
[b] Fold enrichment of AkgR-myc ChIP over control myc antibody ChIP in WT control.
[c] Regulatory role of AkgR on target operons based on change in gene expression between WT and ΔAkgR cells. + = positively regulated by AkgR.

CcmR and AkgR co-operatively regulate central metabolism

The data provided above show that both CcmR and AkgR are involved in regulation of parts of central carbon metabolism in *R. sphaeroides*. While CcmR has a larger regulon than AkgR, their regulons overlap, as both TFs regulate expression of genes that encode subunits of the succinate dehydrogenase complex and thus potentially have complementary roles in the cell. When growing on carbon sources whose metabolism requires significant flux through large portions of the TCA cycle such as α-ketoglutarate and pyruvate, growth is significantly impaired in both the ΔAkgR and ΔCcmR mutants (Figure 7-6A and B), consistent with their predicted roles in regulating expression of genes encoding TCA cycle enzymes. Furthermore, a ΔCcmRΔAkgR double deletion strain shows an even more severe growth defect on these carbon sources (Figure 7-6), indicating that these 2 TFs co-operatively regulate the expression of TCA cycle enzymes. Thus, under these conditions CcmR and AkgR perform non-redundant roles in regulating the flux through the TCA cycle and both are required for normal growth on these carbon sources.

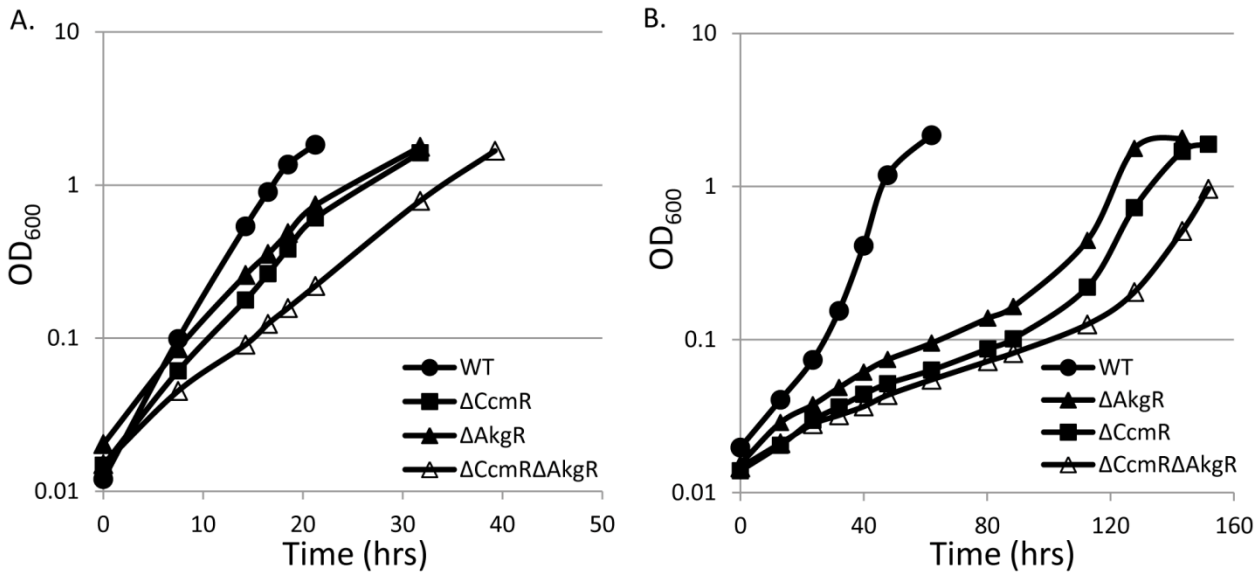

Figure 7-6. Growth of the ΔCcmRΔAkg*R* double deletion mutant on pyruvate and α-ketoglutarate. A comparison of the growth of WT, ΔCcmR, ΔAkgR and ΔCcmRΔAkgR with α-ketoglutarate (A) or pyruvate (B) as the main carbon source under aerobic conditions.

Discussion

Our results provide important new insights into the transcriptional control of central carbon and energy metabolism genes in the α-Proteobacterium *R. sphaeroides*. In particular, we show that two previously uncharacterized TFs, CcmR and AkgR, are direct transcriptional regulators of carbon and energy metabolism. Our analyses revealed that CcmR is a global regulator controlling key aspects of central carbon and energy metabolism and is required for optimal growth under many conditions. In addition, we find that AkgR is a direct transcriptional regulator of selected TCA cycle enzymes. Together, these 2 TFs regulate activities of a significant portion of the *R. sphaeroides* central metabolic pathways (Figure 7-7). Below we summarize the major new knowledge on transcriptional control of carbon and energy metabolism genes in this and related bacteria.

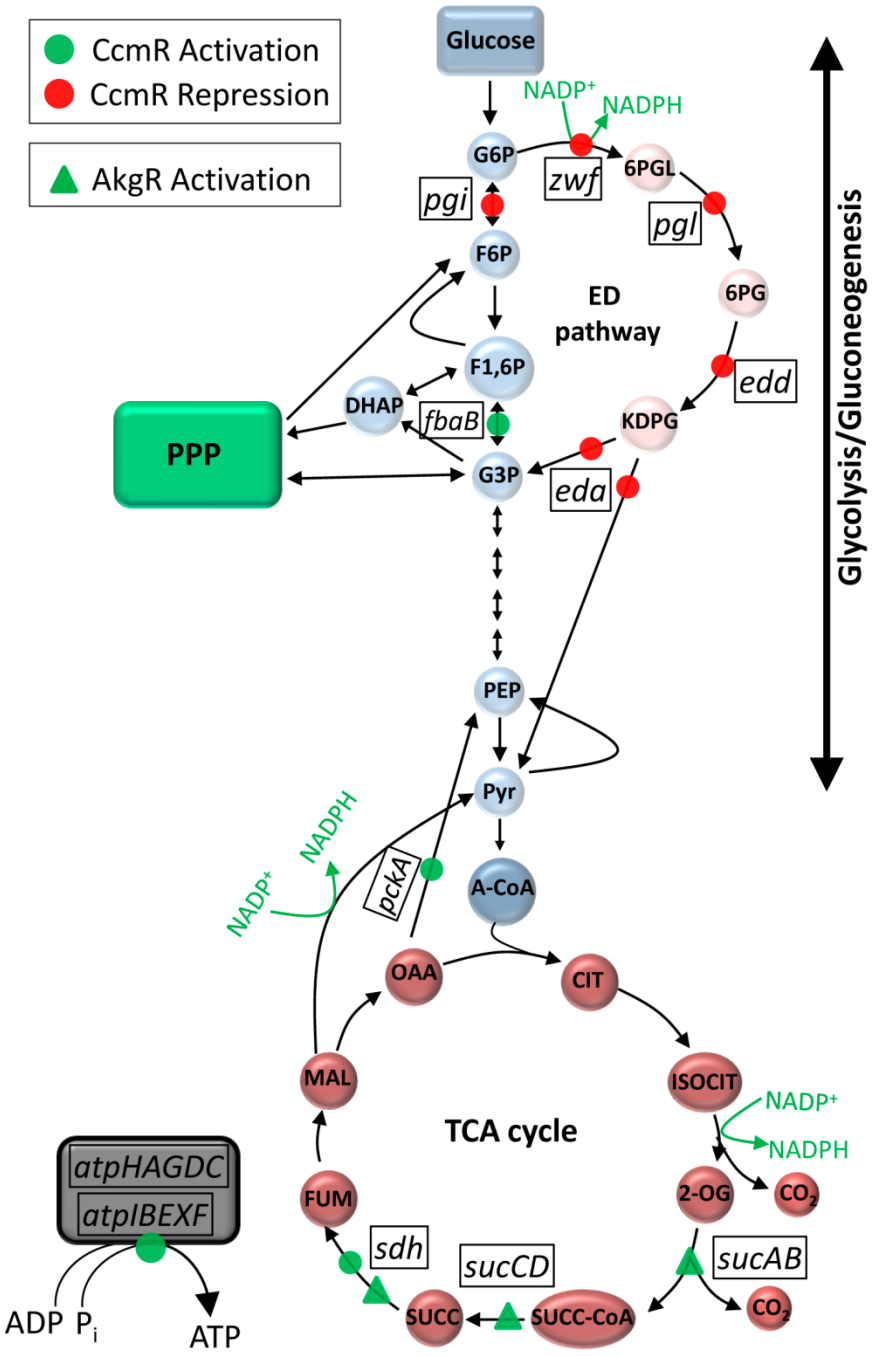

Figure 7-7. Map of central carbon metabolism highlighting CcmR and AkgR targets. Metabolic map highlighting genes regulated by CcmR and AkgR in *R. sphaeroides*. Figure modified from Figure 4 in [19]. TCA – tricarboxylic acid; G6P – glucose 6-phosphate; F6P – fructose 6-phosphate; F1,6P – fructose 1,6-bisphosphate; G3P – glyceraldehyde 3-phosphate; DHAP – dihydroxyacetone phosphate; 6PGL – phosphoglucono-δ-lactone; 6PG – 6-phosphogluconate; KDPG – 2-keto-3-deoxy-6-phosphogluconate; PEP – phosphoenolpyruvate; Pyr – pyruvate; A-CoA – acetyl-CoA; CIT – citrate; ISOCIT – isocitrate; 2-OG – 2-oxoglutarate (α-ketoglutarate); SUCC-CoA – succinyl CoA; SUCC – succinate; FUM – fumarate; MAL – malate; OAA – oxaloacetate; PPP – pentose phosphate pathway.

CcmR is functionally analogous to, but distinct from, other known global regulators of central metabolism

Previous analysis of the regulation of central metabolic genes in *B. subtilis*, *E. coli* and *Pseudomonas putida* led to identification of CcpA [10], Cra [6] and HexR [12] respectively, as transcriptional regulators of central carbon metabolism in these and related organisms [11, 28]. While regulation of central carbon metabolism had not previously been studied in α-Proteobacteria, computational predictions suggested that the LacI family TF, herein named CcmR, was a likely regulator of this process. Our data shows that CcmR is indeed a direct transcriptional regulator of genes that encode enzymes of central carbon metabolism in *R. sphaeroides*. Similar to Cra and HexR, CcmR is a global regulator that functions as both an activator and repressor of central carbon and energy metabolism and, as such, it is required for normal growth under many conditions. The genes directly repressed by CcmR encode proteins of ED glycolytic pathway, similar to genes that are repressed by Cra and HexR [6, 12]. However, more diversity exists in the genes activated by Cra, HexR and CcmR. For instance, CcmR directly regulates transcription of the ATP synthase operons in *R. sphaeroides*, which is not a known or predicted target for Cra or HexR. In addition, Cra and HexR each regulate transcription of genes encoding enzymes that function in the glyoxylate shunt, a pathway that is missing or incomplete in *R. sphaeroides*. Overall, the complement of genes regulated by these TFs and the resulting phenotypes of mutant strains make them seem functionally analogous.

However, Cra, HexR and CcmR have major differences. For example, HexR is a RpiR family TF that is often found in close association with glucose utilization operons in some β- and γ-proteobacteria [11]. On the other hand, Cra and CcmR belong to the LacI family of TFs and their structural genes are not located proximal to glucose utilization operons. LacI homologs are a large class of TFs mainly involved in regulating the uptake and utilization of carbon sources [26]. However, other members of this family such as Cra and CcmR can have broader regulatory functions. A phlyogenetic analysis of 159 LacI family TFs encoded in 22 α-, β- and γ-Proteobacteria indicates that Cra and CcmR belong to different clades or sub-

families of LacI TFs (Figure S3). Based on this analysis, CcmR is most closely related *E. coli* gluconate metabolism regulatory proteins, GntR and IdnR, while the uncharacterized LacI family TF RSP_3700 in *R. sphaeroides* is the most closely related TF to Cra. Consistent with the predicted evolutionary distance between Cra and CcmR, the inverted repeat DNA sequence recognized by Cra (GCTGAAnCGnTTCA) [8] is quite different from the CcmR binding site identified in this study ([T/C]GTT N_6 AAC[A/T]) (Figure 7-2C). In addition, differences also appear to exist in effector molecules recognized by Cra and CcmR. Previous analyses have shown that frucoste-1-phosphate and fructose-1,6-bisphosphate affect Cra binding *in vitro*, reducing its ability to bind target sequences [29, 30]. However, addition of fructose-1,6-bisphosphate did not have any effect on CcmR binding *in vitro* (Figure 7-3H). On the other hand, 6-PG significantly reduced CcmR's ability to bind DNA (Figure 7-3G), indicating Cra and CcmR respond to a different set of effector molecules and thus likely have significant differences in their EBDs. The effector molecules used by CcmR and Cra to control DNA binding likely reflect differences the preferred routes of glucose metabolism between *R. sphaeroides* and *E. coli*. *E. coli* metabolizes glucose primarily via the Embden-Meyerhof pathway [27] of which fructose-1-phosphate and fructose-1,6-bisphosphate are intermediates. On the other hand, *R. sphaeroides* metabolizes glucose primarily via the ED pathway of which 6-PG is an intermediate [19, 27]. Thus, the ability to bind these intermediates provides a mechanism for derepression of these pathways in the presence of glucose.

Our data also show that CcmR also controls transcription of genes that encode electron carriers (cytochromes, NADH dehydrogenase subunits) and integral membrane bioenergetics proteins such as the F_1F_0 ATPase. To date, there is little information about the physiological signals or TFs that control transcription of these genes, either in *R. sphaeroides* or other bacteria. Indeed, published analyses of *E. coli atp* operon expression indicate that it is constitutively expressed [31]. In contrast, our data predicts that CcmR activates the transcription of two *atp* operons, so it appears that regulation of these operons likely to respond to 6-PG levels that control the activity of this TF. Thus, we propose that the role of

CcmR in activating *atp* operon transcription reflects a heretofore unrecognized link between where central carbon and energy metabolism in this and possibly other bacteria.

CcmR and AkgR are conserved across several α-Proteobacteria species

While this study analyzed the roles of CcmR and AkgR in *R. sphaeroides*, these 2 TFs are also conserved in several other α-Proteobacteria. To gain additional insight into the conservation and diversity of the CcmR and AkgR regulons in other α-Proteobacteria, we used comparative genomics to predict their regulons in other species. Utilizing the derived position weight matrices for CcmR and AkgR binding sites, we searched for related sequences across 21 representative α-Proteobacteria species. We found that these CcmR and AkgR binding sites are most highly conserved among the Rhodobacterales (Figure 7-8, Figure S2), with only a few instances of the motifs found in other orders of α-Proteobacteria. Within the Rhodobacterales, the CcmR binding site was often found upstream of genes involved in central carbon metabolism, suggesting that it serves a similar function in these species. However, we found that the predicted CcmR target genes can vary considerably in individual Rhodobacterales species, encompassing other central metabolism genes such as glyceraldehydes-3-phosphate dehydrogenase (*gapA1 and gapA2*), enolase (*eno*), pyruvate kinase (*pykA*), gluconate 5-dehydrogenase (*idnO*), glycogen phosphorylase, fructose-6-phosphate aldolase and malate dehydrogenase (*mdh*), which were not identified as direct CcmR targets in *R. sphaeroides* (Figure 7-8). This analysis also predicts that in *Azospirillum* sp. B510, CcmR has a more specific role in pyruvate (through regulation of pyruvate dehydrogenase regulator, *pdhR*) and lactate metabolism. These predicted differences in regulon composition might reflect differences in the substrate utilization profiles and metabolic lifestyles of these diverse groups of bacteria.

A similar comparative genomics analysis predicts that the proposed AkgR binding motif of GTGATCAC N_{13} GTGATCAC is also conserved across α-Proteobacteria species. However, in contrast to CcmR, the predicted AkgR binding sites were restricted the *sucCDAB* and/or succinate dehydrogenase operon among the Rhodobacterales (Figure S2), with only a few other putative members of an extended regulon such as

tripartite tricarboxylate transporter (*tctA*), shikimate kinase and quinoprotein ethanol dehydrogenase (*exaA2*) in some species (Figure S2). Thus, it appears that the AkgR regulon is also highly conserved across α-Proteobacteria species.

Regulation of central carbon and energy metabolism in R. sphaeroides involves a set of previously uncharacterized TFs

Leveraging computational predictions of transcriptional regulation in *R. sphaeroides*, we used a combination of biochemical genetic, genomic and physiological analyses to determine the role of previously uncharacterized TFs that regulate central carbon metabolism in *R. sphaeroides* and likely other related bacteria. These findings highlight the utility of large-scale TRN inference in guiding scientific discovery.

However, the computational predictions of this TRN suggest there are likely still other, as yet unidentified, TFs that might play important roles in regulation of central carbon metabolism in *R. sphaeroides* and related bacteria. For example, genes encoding enzymes of the TCA cycle and electron transport chain such as citrate synthase, isocitrate dehydrogenase and NADH dehydrogenase, which are not members of either the CcmR or AkgR regulons, share conserved DNA sequence motifs in their upstream regulatory regions suggesting a common mode of transcriptional regulation. However, at this point specific TF(s) are yet to be associated with regulation of these genes. Furthermore, genes encoding enzymes required for acetate utilization via the ethylmalonyl-CoA pathway in *R. sphaeroides* [32] are also predicted to be jointly regulated via as yet unknown TF(s). Thus, the regulation of central carbon metabolism in *R. sphaeroides* and possibly many other α-Proteobacteria, likely involves the interplay of several TFs, some of which are still to be discovered. Our analysis represents a first step in experimental characterization of this complex regulatory network.

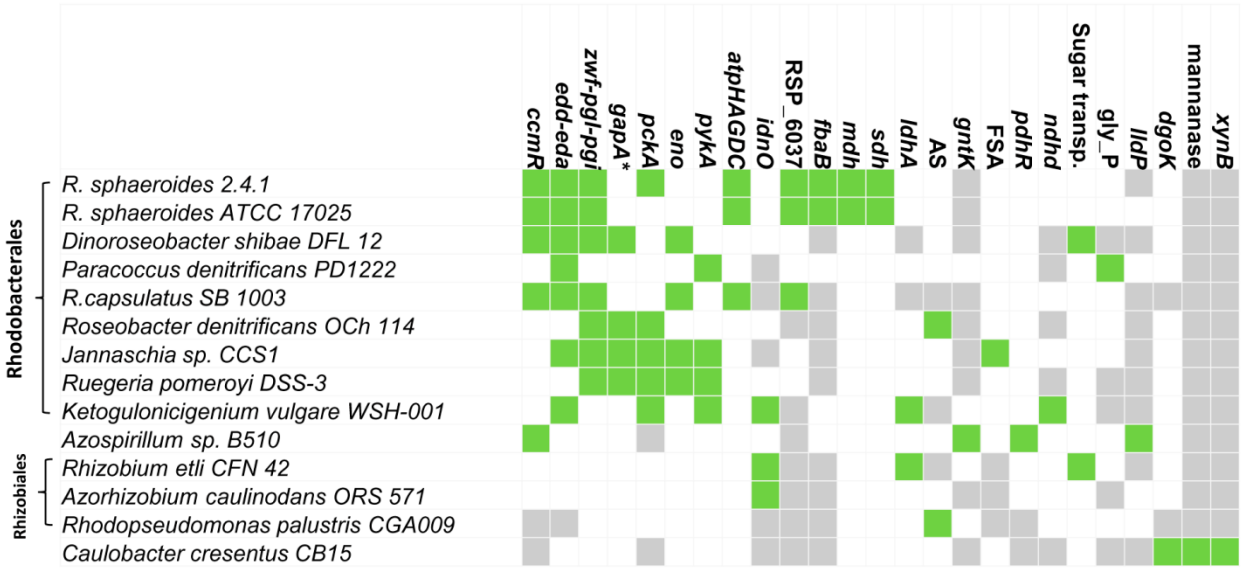

Figure 7-8. Conservation of CcmR regulon across α-Proteobacteria. Predicted conservation of the CcmR regulon across 14 α-Proteobacterial species based on the presence of a predicted CcmR binding site upstream of the indicated genes. Green boxes indicated a candidate CcmR binding site was located upstream of the gene, gray boxes indicate no ortholog for the specific protein was identified in that genome, while white boxes indicate presence of orthologous protein but no candidate CcmR binding site upstream of its encoding gene. CcmR targets specific to *R. sphaeroides* 2.4.1 (i.e., *rpoH1*, *flgM*, *aptI*, RSP_2425, RSP_3896, RSP_3926 and RSP_7376) were omitted for brevity. *gapA* – glyceraldehydes-3-phosphate dehydrogenase; *eno* – enolase; *pykA* – pyruvate kinase; *idnO* – gluconate 5-dehydrogenase; *fbaB* – fructose-bisphosphate aldolase; *mdh* – malate dehydrogenase; *ldhA* – lactate dehydrogenase related; AS – acyl-CoA synthetase; *gntK* – gluconokinase; FSA – fructose-6-phosphate aldolase; *pdhR* – pyruvate dehydrogenase regulator; *ndhd* – NADH dehydrogenase; Sugar transp. – ABC sugar transporter; gly_P – glycogen phosphorylase; *lldP* – lactate transporter; *dgoK* – 2-dehydro-3-deoxygalactonokinase; *xynB* – xylan beta-1,4-xylosidase. * *gapA* includes both *gapA1* and *gapA2* in all species with predicted CcmR binding sites upstream of these genes.

Materials and Methods

Bacterial strains and growth conditions

R. sphaeroides 2.4.1 was used as the parental strain in this study. The individual and double ΔCcmR and ΔAkgR mutants were constructed in this background. *E. coli* DH5α was used as a plasmid host and *E. coli* S17-1 was used to conjugate DNA into *R. sphaeroides* (Table S2). *R. sphaeroides* cultures were incubated at 30°C in Sistrom's Minimal Medium (SMM) [20] or in SMM lacking glutamate and aspartate, and with succinate replaced with an alternative carbon source. The molar concentration of carbon atoms of the carbon source was kept constant at 135.5 mM, equivalent to that of succinate in SMM. When required, the media was supplemented with 3-5 μM IPTG and 25 μg/mL kanamycin. *E. coli* cells were grown in Luria Bertani (LB) medium at 37°C, supplemented with 50 μg/mL kanamycin where needed.

Construction of mutants and expression plasmids

ΔCcmR, ΔAkgR and ΔCcmRΔAkgR strains were constructed as in-frame markerless deletions [19] and deletions were confirmed by PCR and sequencing with specific primers (Table S2). Plasmid constructs for the ectopic expression of 3X Myc tagged CcmR (C-terminal tag) and AkgR (N-terminal tag), were made using sequence specific primers (Table S2) and conjugated into the relevant *R. sphaeroides* strains selecting for plasmid-encoded kanamycin resistance.

Protein purification

To purify CcmR, a His tagged protein was made by cloning *ccmR* into pIND5[33] using specific primers (Table S2) lacking the native stop codon, thus allowing inclusion of the pIND5 encoded C-terminal 6X His tag**.** The pIND5-*ccmR*-6XHis plasmid was transferred to *E. coli* BL21 (DE3) (Novagen). 5 mL of *E. coli* cells harboring pIND5-*ccmR*-6X His were grown overnight in LB supplemented with 50 μg/mL kanamycin, then inoculated into 250mL of fresh LB supplemented with 50 μg/mL kanamycin and 500

μM IPTG, and incubated at 30°C for ~5 hrs until cells reached an OD_{600} ~ 0.6. 100 mL of cells was harvested by centrifugation and lysed by sonication in buffer containing 50 mM NaH_2PO_4, 100 mM NaCl, 10 mM imidazole, 10% glycerol and 0.25m/mL lysozyme. The suspension was centrifuged and CcmR-6X His was purified from the supernatant by Ni^{2+} affinity chromatography (Qiagen). The slurry was washed 3 times and then CcmR-6X His was eluted in buffer containing 50 mM NaH_2PO_4, 300 mM NaCl, 250 mM imidazole and 10% glycerol. Protein samples were stored in ~50% glycerol at -80°C until use. Protein was quantified using Lowry assay [34].

Electrophoretic mobility shift assay (EMSA)

To assess *in vitro* binding of CcmR to target promoters, assays were conducted with the EMSA kit (E33075, Life technologies). Target DNA fragments were amplified by PCR from genomic DNA using specific primers (Table S2) and the DNA fragments gel purified. Increasing amounts of purified CcmR (final concentration of monomers ranging from ~0.12 – 2.9 μM) was incubated with ~0.05 μM of purified target DNA for 45 minutes at room temperature (RT) in 1X EMSA binding buffer in 10 μL reactions. Samples were run on 6% polyacrylamide retardation gel with pre-chilled TBE running buffer at 200 V for 35 minutes at 4°C. Gels were stained with SYBR green DNA stain for 25 minutes with continuous shaking in the dark at RT, washed twice with distilled water and visualized on the omega lum C imager (Aplegen, Inc.). To assess the impact of metabolites on CcmR binding 0.5, 2 and 10 mM 6-phosphogluconate (6-PG), 2-Keto-3-deoxy-6-phosphogluconate, phosphoenolpyruvate, glucose, glucose 6-phosphate and fructose-1,6-bisphosphate were added to binding assays containing ~0.03 μM DNA and ~ 1.94 μM purified CcmR. Lower concentrations ranging from 0.01μM to 10mM of 6-PG were used to further assess the effect of 6-PG on CcmR binding.

To test the role of individual bases in the predicted CcmR motif on DNA binding by this CcmR, a 194bp DNA fragment upstream of *ccmR*, starting from the base before the *ccmR* start codon, was amplified from genomic DNA by PCR using specific primers (Table S2). Three point mutations at positions A-84G, A-

85G and T-92G relative to the *ccmR* start codon, were made in the predicted CcmR binding site by overlap extension PCR. The WT and mutated DNA fragments were incubated with purified CcmR and EMSA analysis conducted as described above.

RNA extraction, qRT-PCR and microarray analyses

RNA was isolated from exponential phase *R. sphaeroides* cultures that were grown aerobically in 500 ml roux bottles bubbled with 69% N_2, 30% O_2, 1% CO_2. RNA isolation, cDNA synthesis, labeling, and hybridization to *R. sphaeroides* GeneChip microarrays (Affymetrix, Santa Clara, CA) were performed as previously described [35]. Microarray datasets were normalized by Robust Multichip Average (RMA) to $\log_2$ scale with background adjustment and quantile normalization [36]. Statistical analysis of normalized data to identify differentially expressed genes was done using the Limma package [37]. Correction for multiple testing was done using Benjamini-Hochberg correction [38]. All analyses were conducted in the R statistical programming environment (http://www.R-project.org). qRT-PCR experiments were conducted in duplicates for each biological replicate using SYBR Green JumpStart Taq ReadyMix (Sigma-Aldrich). Relative expression was determined via the $2^{-\Delta\Delta CT}$ method with efficiency correction [39]. The *R. sphaeroides rpoZ* gene was used for normalization. Primers used in this analysis are provided in Table S2.

Chromatin immunoprecipitation analysis

R. sphaeroides cells (*ΔccmR*+pIND5-*ccmR*-3XMyc and *ΔakgR*+pIND5-*akgR*-3XMyc) were grown aerobically in 500 ml cultures with bubbling, as described above. Cells were treated with 3 to 5 μM IPTG at inoculation (i.e., the lowest IPTG concentration required to restore normal aerobic growth with the tagged protein) and harvested at an OD_{600} of ~ 0.35. Chromatin immunoprecipication was conducted [40] using polyclonal antibodies against the Myc epitope tag (ab9132, Abcam plc). Immunoprecipitated DNA samples were PCR-amplified, gel purified (size selection ~200bp) and sequenced at the UW Biotechnology Center sequencing facility, using the HiSeq 2500 sequencing system (Illumina, Inc). The

50bp sequence tags were mapped to the *R. sphaeroides* 2.4.1 genome (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Rhodobacter_sphaeroides_2_4_1_uid57653/) using SOAP version 2.21 [41], allowing a maximum of 2 mismatches and no gaps.

Peaks were identified using MOSAiCS [42] at a false discovery rate of 0.05. The MOSAiCS analysis was conducted as a two-sample analysis, with ChIP-seq data from either input DNA or ChIP conducted with anti-Myc antibody in the WT strain without a Myc-tagged protein used as a control. Only peaks that were called as significant using both controls (i.e., the intersect of the 2 analyses) were considered as true peaks. Motifs were identified from sequences under the peak regions using MEME [43]. Genomic locations with both a significant ChIP-seq peak and shared motifs were considered true binding sites.

Phylogenetic tree construction

To construct a phylogenetic tree for LacI family TFs, 159 known or predicted LacI proteins from 22 species (7 γ-, 3 β- and 12 α-proteobacteria) (Figure S3) were aligned using ClustalX [44] and a phylogenetic tree constructed via the neighbor-joining phylogenetic method with the ProtDist program of the PHYLIP [45], using 100 bootstrap pseudo-replicates to construct the consensus tree. The resulting tree was visualized with display tools from the interactive tree of life (iTOL) website [46].

Identification of putative CcmR and AkgR regulons across α-Proteobacteria

To determine the conservation of the CcmR and AkgR regulons across α-Proteobacteria, we used position weight matrices (PWMs) built from verified *R. sphaeroides* CcmR and AkgR targets to search all intergenic sequences from 21 representative α-Proteobacteria species: *R. sphaeroides* 2.4.1, *R. sphaeroides* ATCC 17025, *R. capsulatus* SB 1003, *Roseobacter denitrificans* Och 114, *Dinoroseobacter shibae* DFL 12, *Paracoccus denitrificans* PD1222, *Rhodopseudomonas palustris* CGA009, *Bradyrhizobium japonicum* USDA 110, *Sinorhizobium meliloti*, *Ruegeria pomeroyi*, *Jannaschia sp. CCS1*, *Mesorhizobium ciceri*, *Azospirillum* sp. B510, *Rhizobium etli*, *Starkeya novella*, *Azorhizobium caulinodans*, *Xanthobacter autotrophicus*, *Methylobacterium chloromethanicum*, *Rhodospirillum rubrum*,

Ketogulonicigenium vulgare, *Caulobacter crescentus CB15*. Searches were carried out using MAST [43]. Only PWM hits with a p-value of $< 10^{-6}$ were considered as putative CcmR/AkgR targets. Orthologous proteins across species were determined via orthoMCL analysis [47].

References

1. Hueck CJ, Hillen W: **Catabolite repression in *Bacillus subtilis*: a global regulatory mechanism for the gram-positive bacteria?** *Mol Microbiol* 1995, **15**(3):395-401.
2. Stulke J, Hillen W: **Coupling physiology and gene regulation in bacteria: the phosphotransferase sugar uptake system delivers the signals**. *Naturwissenschaften* 1998, **85**(12):583-592.
3. Botsford JL, Harman JG: **Cyclic AMP in prokaryotes**. *Microbiol Rev* 1992, **56**(1):100-122.
4. Park DM, Akhtar MS, Ansari AZ, Landick R, Kiley PJ: **The bacterial response regulator ArcA uses a diverse binding site architecture to regulate carbon oxidation globally**. *PLoS Genet* 2013, **9**(10):e1003839.
5. Cunningham L, Georgellis D, Green J, Guest JR: **Co-regulation of lipoamide dehydrogenase and 2-oxoglutarate dehydrogenase synthesis in *Escherichia coli*: characterisation of an ArcA binding site in the lpd promoter**. *FEMS Microbiol Lett* 1998, **169**(2):403-408.
6. Saier MH, Jr., Ramseier TM: **The catabolite repressor/activator (Cra) protein of enteric bacteria**. *J Bacteriol* 1996, **178**(12):3411-3417.
7. Ramseier TM, Bledig S, Michotey V, Feghali R, Saier MH, Jr.: **The global regulatory protein FruR modulates the direction of carbon flow in *Escherichia coli***. *Mol Microbiol* 1995, **16**(6):1157-1169.
8. Shimada T, Yamamoto K, Ishihama A: **Novel members of the Cra regulon involved in carbon metabolism in *Escherichia coli***. *J Bacteriol* 2011, **193**(3):649-659.
9. Moreno MS, Schneider BL, Maile RR, Weyler W, Saier MH, Jr.: **Catabolite repression mediated by the CcpA protein in *Bacillus subtilis*: novel modes of regulation revealed by whole-genome analyses**. *Mol Microbiol* 2001, **39**(5):1366-1381.
10. Tobisch S, Zuhlke D, Bernhardt J, Stulke J, Hecker M: **Role of CcpA in regulation of the central pathways of carbon catabolism in *Bacillus subtilis***. *J Bacteriol* 1999, **181**(22):6996-7004.
11. Leyn SA, Li X, Zheng Q, Novichkov PS, Reed S, Romine MF, Fredrickson JK, Yang C, Osterman AL, Rodionov DA: **Control of proteobacterial central carbon metabolism by the HexR transcriptional regulator: a case study in *Shewanella oneidensis***. *J Biol Chem* 2011, **286**(41):35782-35794.
12. del Castillo T, Duque E, Ramos JL: **A set of activators and repressors control peripheral glucose pathways in *Pseudomonas putida* to yield a common central intermediate**. *J Bacteriol* 2008, **190**(7):2331-2339.
13. Daddaoua A, Krell T, Ramos JL: **Regulation of glucose metabolism in *Pseudomonas*: the phosphorylative branch and entner-doudoroff enzymes are regulated by a repressor containing a sugar isomerase domain**. *J Biol Chem* 2009, **284**(32):21360-21368.
14. Blankenship RE, Madigan MT, Bauer CE: **Anoxygenic photosynthetic bacteria**, vol. XLVIII; 1995.
15. Hunter CN, Daldal, F., Thurnauer, M. C. and Beatty, J. T.: **The purple phototrophic bacteria**, vol. 28: Springer; 2009.

16. Garrity GM, Brenner, D. J., Krieg, N. R., Staley, J. T. and Krieg, N. R.: **The proeobacteria: part C the alpha-, beta-, delta-, and epsilon-proteobacteria**: Springer; 2005.
17. Imam S, Yilmaz S, Sohmen U, Gorzalski AS, Reed JL, Noguera DR, Donohue TJ: **iRsp1095: a genome-scale reconstruction of the *Rhodobacter sphaeroides* metabolic network**. *BMC Syst Biol* 2011, **5**:116.
18. Mackenzie C, Eraso JM, Choudhary M, Roh JH, Zeng X, Bruscella P, Puskas A, Kaplan S: **Postgenomic adventures with *Rhodobacter sphaeroides***. *Annu Rev Microbiol* 2007, **61**:283-307.
19. Imam S, Noguera DR, Donohue TJ: **Global insights into energetic and metabolic networks in *Rhodobacter sphaeroides***. *BMC Syst Biol* 2013, **7**(1):89.
20. Sistrom WR: **A requirement for sodium in the growth of *Rhodopseudomonas spheroides***. *J Gen Microbiol* 1960, **22**:778-785.
21. Swem LR, Gong X, Yu CA, Bauer CE: **Identification of a ubiquinone-binding site that affects autophosphorylation of the sensor kinase RegB**. *J Biol Chem* 2006, **281**(10):6768-6775.
22. Dufour YS, Imam S, Koo BM, Green HA, Donohue TJ: **Convergence of the transcriptional responses to heat shock and singlet oxygen stresses**. *PLoS Genet* 2012, **8**(9):e1002929.
23. Karls RK, Brooks J, Rossmeissl P, Luedke J, Donohue TJ: **Metabolic roles of a *Rhodobacter sphaeroides* member of the sigma32 family**. *J Bacteriol* 1998, **180**(1):10-19.
24. Poggio S, Osorio A, Dreyfus G, Camarena L: **The flagellar hierarchy of *Rhodobacter sphaeroides* is controlled by the concerted action of two enhancer-binding proteins**. *Mol Microbiol* 2005, **58**(4):969-983.
25. Lewis M, Chang G, Horton NC, Kercher MA, Pace HC, Schumacher MA, Brennan RG, Lu P: **Crystal structure of the lactose operon repressor and its complexes with DNA and inducer**. *Science* 1996, **271**(5253):1247-1254.
26. Weickert MJ, Adhya S: **A family of bacterial regulators homologous to Gal and Lac repressors**. *J Biol Chem* 1992, **267**(22):15869-15874.
27. Fuhrer T, Fischer E, Sauer U: **Experimental identification and quantification of glucose metabolism in seven bacterial species**. *J Bacteriol* 2005, **187**(5):1581-1590.
28. Antunes A, Camiade E, Monot M, Courtois E, Barbut F, Sernova NV, Rodionov DA, Martin-Verstraete I, Dupuy B: **Global transcriptional control by glucose and carbon regulator CcpA in *Clostridium difficile***. *Nucleic Acids Res* 2012, **40**(21):10701-10718.
29. Ramseier TM, Negre D, Cortay JC, Scarabel M, Cozzone AJ, Saier MH, Jr.: **In vitro binding of the pleiotropic transcriptional regulatory protein, FruR, to the *fru, pps, ace, pts* and *icd* operons of *Escherichia coli* and *Salmonella typhimurium***. *J Mol Biol* 1993, **234**(1):28-44.
30. Shimada T, Fujita N, Maeda M, Ishihama A: **Systematic search for the Cra-binding promoters using genomic SELEX system**. *Genes Cells* 2005, **10**(9):907-918.
31. Kasimoglu E, Park SJ, Malek J, Tseng CP, Gunsalus RP: **Transcriptional regulation of the proton-translocating ATPase (*atpIBEFHAGDC*) operon of *Escherichia coli*: control by cell growth rate**. *J Bacteriol* 1996, **178**(19):5563-5567.

32. Erb TJ, Berg IA, Brecht V, Muller M, Fuchs G, Alber BE: **Synthesis of C5-dicarboxylic acids from C2-units involving crotonyl-CoA carboxylase/reductase: the ethylmalonyl-CoA pathway**. *Proc Natl Acad Sci U S A* 2007, **104**(25):10631-10636.
33. Ind AC, Porter SL, Brown MT, Byles ED, de Beyer JA, Godfrey SA, Armitage JP: **Inducible-expression plasmid for *Rhodobacter sphaeroides* and *Paracoccus denitrificans***. *Appl Environ Microbiol* 2009, **75**(20):6613-6615.
34. Lowry OH, Rosebrough NJ, Farr AL, Randall RJ: **Protein measurement with the Folin phenol reagent**. *J Biol Chem* 1951, **193**(1):265-275.
35. Tavano CL, Podevels AM, Donohue TJ: **Identification of genes required for recycling reducing power during photosynthetic growth**. *J Bacteriol* 2005, **187**(15):5249-5258.
36. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185-193.
37. Smyth G: **Applications in genetics and molecular biology 3**: Berkeley Electronic Press; 2004.
38. Benjamini Y, Hochberg Y: **Controlling the false discovery rate - a practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society Series B-Methodological* 1995, **57**:289-300.
39. Livak KJ, Schmittgen TD: **Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method**. *Methods* 2001, **25**(4):402-408.
40. Dufour YS, Landick R, Donohue TJ: **Organization and evolution of the biological response to singlet oxygen stress**. *J Mol Biol* 2008, **383**(3):713-730.
41. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J: **SOAP2: an improved ultrafast tool for short read alignment**. *Bioinformatics* 2009, **25**(15):1966-1967.
42. Kuan PF, Chung D, Pan G, Thomson JA, Stewart R, Keleş S: **A statistical framework for the analysis of ChIP-seq data**. *Journal of the American Statistical Association* 2011, **106**(495):891-903.
43. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS: **MEME SUITE: tools for motif discovery and searching**. *Nucleic Acids Res* 2009, **37**(Web Server issue):W202-208.
44. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R *et al*: **Clustal W and Clustal X version 2.0**. *Bioinformatics* 2007, **23**(21):2947-2948.
45. Felsenstein J: **PHYLIP - Phylogeny interference package (Version 3.2)**. *Cladistics* 1989, **5**:164–166.
46. Letunic I, Bork P: **Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy**. *Nucleic Acids Res* 2011, **39**(Web Server issue):W475-478.
47. Li L, Stoeckert CJ, Jr., Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes**. *Genome Res* 2003, **13**(9):2178-2189.

Chapter 8: Integrating large-scale network models, summary and future directions

Integrated modeling of metabolism and transcriptional regulation in *R. sphaeroides*

While the construction, validation and refinement of individual networks of metabolism and transcriptional regulation in *R. sphaeroides* resulted in a significant amount of new biological insight into various aspects of its physiology (Chapters 2 – 7), integrating these network models could potentially yield additional insights not obtainable with either model alone. As described in Chapter 1, there are currently 3 established approaches for integrating constraint-based models (CBMs) and transcriptional regulatory networks (TRNs), namely regulatory flux balance analysis (rFBA), steady state regulatory FBA (SR-FBA) and probabilistic regulation of metabolism (PROM) [1-3]. In this section, I'll describe the integrated regulatory-metabolic models I constructed for *R. sphaeroides* using the rFBA and PROM approaches. In addition, I will provide a summary of the predictive performances of the models in comparison to iRsp1140 using data I compiled for substrate utilization and phenotypes of defined mutants.

Building an integrated model for R. sphaeroides using rFBA

Combining CBMs and TRNs using rFBA involves using regulatory rules from the TRN to predict the activity states (i.e., on or off) of reactions within the CBM prior to FBA simulations. To achieve this, the activity states of metabolites, transcription factors (TFs), metabolic proteins and reactions were set based on the growth condition used for simulation. These were achieved as follows:

Exchange fluxes determine the activity state of metabolites – Depending on the growth conditions being simulated, the set of exchange reactions (i.e., reactions that allow simulation of exchange of metabolites

between the cell and its environment) that supply metabolites to the cell will differ. For instance, during simulation of aerobic growth the oxygen exchange reaction would be allowed to deliver oxygen to the cells, while this would not be permitted under anaerobic conditions. To keep track of which exchange reactions carried flux under a given condition, metabolites' activity states were linked to their exchange reactions. Thus, if an exchange reaction carried flux of metabolite A into the cell, A would be turned on. The amount of flux required to turn A on could also be specified.

Metabolite activity states determine TF and enzyme activity states – Based on information garnered from literature, certain TFs and other proteins are predicted to be able to directly or indirectly sense the presence of specific metabolites. For instance, in *R. sphaeroides* FnrL and NifA are known to be oxygen sensitive [4, 5]. Thus, the activity states of these proteins can be determined based on the activity states of the metabolites they sense (e.g., if oxygen is on, FnrL would be off). Consequently, based on the growth conditions being simulated the activity states of some TFs in the TRN, as well as some enzymes in the CBM can be determined.

TF activity states determine protein activity states – Having determined the activity states of a subset of TFs based on the experimental conditions (or genetic perturbations, if for example a TF gene is deleted), the activity states of other proteins within the TRN (encompassing a subset of proteins within the CBM) are set based on TF activity states. To do this, TF-TF interactions are resolved prior to other TF-target interactions, as the states of all TFs need to be set first in order to factor in all regulatory influences on downstream targets. Protein activity states could also be determined by genetic perturbations, if the gene encoding a given protein is deleted.

Protein activity states determine reaction activity states – With protein activity states determined, the reaction activity states are set using the genes-to-proteins-to-reactions (GPR) associations from the CBM. A reaction catalyzed by a protein(s) with an off activity state would also have an off activity state and *vice versa*. For reactions catalyzed by multiple isozymes, all isozymes would need be in off states for the

reaction to be off, while for reactions catalyzed by multi-subunit enzymes, if one of the subunits is off, then the reaction would be off. Reactions with an off activity state are not allowed to carry flux, while those with an on activity state are allowed to carry flux in the FBA simulation.

The final result of these associations is an integrated network in which the environmental conditions or genetic perturbations determine the set of active reactions (i.e., reactions allowed to carry flux) in the CBM. Initially, the genes, proteins and reactions are assumed to be on, with specific genes, proteins or reactions turned off based on mappings determined by the growth conditions or genetic perturbations. On the other hand, all metabolites are initially considered off and turned on based on exchange reaction fluxes. Using this approach, I implemented rFBA with iRsp1140 and the large-scale TRN model. A total of 591 (51.8%) of the 1140 genes included in iRsp1140 have some level of regulatory control from the TRN. This rFBA model will hereafter be referred to as iRsp1140-rFBA.

Building an integrated model for R. sphaeroides using PROM

To generate an integrated regulatory-metabolic model for *R. sphaeroides* using PROM, similar mappings to those implemented for rFBA were used to link exchange reactions to metabolites and metabolites to TFs. Other links between the components of the integrated model were achieved using PROM-specific formulations [1] as described below:

TF activity states determine protein activities – To calculate activity of a protein under a given condition, its activity was estimated using information in the existing 198 microarray datasets for *R. sphaeroides* (Chapter 5). This was achieved by considering all possible permutations of on/off states of the TF(s) regulating a given protein, then estimating how active a protein is for each permutation by the determining fraction of times the gene encoding the protein is on in the microarray dataset for that particular permutation of TF states. For instance, if gene A is regulated by TFs B, C and D, 2^3 possible TF states exist (e.g., B is on, C is on, D is off or B is off, C is on, D is on etc). For each of these permutations,

the probability that A is on is determined by counting the total number of times A is on when a particular permutation of TFs states occurs and dividing by the total number times that permutation occurs. Thus,

$$P(A=1 \mid B=1, C=1, D=0) = \frac{N(A=1 \mid B=1, C=1, D=0)}{N(B=1, C=1, D=0)}$$

The threshold used for calling a gene on or off was determined from control probes on the *R. sphaeroides* Affymetrix chip and was set to between 7.5 and 8 for $\log_2$-transformed normalized array signal intensities.

Protein activities determine reaction activities – Using the GPR associations of the CBM, the activities of the reactions were estimated from the activities of their corresponding proteins. For reactions catalyzed by a single protein, its activity was set to be equal to protein activity. When cells have multiple isozymes to catalyze a given reaction, the reaction activity was set to be equal to the isozyme with the highest activity (i.e., the highest probability estimated from the microarray datasets). In cases of a reaction catalyzed by a multi-subunit enzyme, its activity was set to be equal to the subunit with the lowest activity. The allowable fluxes through the reactions are determined by multiplying the reaction bounds by the reaction activities.

The result of these associations is an integrated network in which the environmental conditions or genetic perturbations determine the allowable amount of flux through reactions in the CBM. This PROM model will hereafter be referred to as iRsp1140-PROM.

PROM simulations were run as previously described [1] using the following constraints:

$$\max \left(v_{biomass} - \sum_j (\kappa * \alpha_j + \kappa * \beta_j)\right)$$

s.t.

$$S \bullet v = 0$$

$$P_j * v_{j,\min} - \alpha_j \le v_j \le P_j * v_{j,\max} + \beta_j$$

$$\alpha_j, \beta_j \ge 0$$

where $v_{biomass}$ is the flux through the biomass reaction; $P_j * v_{j,min}$ is the minimum allowable flux through reaction j; $P_j * v_{j,max}$ is the maximum allowable flux through reaction j; α_j and β_j are positive variables that allow some deviation from the TRN imposed flux bounds; and $\Sigma (\kappa * \alpha_j + \kappa * \beta_j)$ is a penalty applied for deviation from the TRN imposed flux bounds. κ was set to 1 for iRsp1140-PROM simulations as it yielded the best results.

Comparing predictions from iRsp1140, iRsp1140-rFBA and iRsp1140-PROM

To assess predictions from iRsp1140, iRsp1140-rFBA and iRsp1140-PROM, I compiled a list of known growth and gene deletion phenotypes for *R. sphaeroides* from literature, the analyses conducted as part of this dissertation, and other unpublished sources. These included 51 TF gene deletion phenotypes, 71 metabolic gene deletion phenotypes, 162 carbon source utilization growth phenotypes and 6 other condition-dependent growth phenotypes (Tables 8-1, 8-2, 8-3 and 8-4). These data were used to assess the performance of individual models in predicting aspects of the *R. sphaeroides* lifestyle.

TF deletion phenotypes – iRsp1140 does not include TFs, so no predictions for the impact of TF deletions can be made using this model. iRsp1140-PROM appeared to perform better than iRsp1140-rFBA at predicting the outcomes of TF deletions, accurately predicting 80.4% of the observed growth phenotypes compared to 66.7% by iRsp1140-rFBA (Tables 8-1). While iRsp1140-PROM appeared to outperform iRsp1140-rFBA in this assessment, some known growth phenotypes were captured by iRsp1140-rFBA but not iRsp1140-PROM. For instance, the essentiality of the photosynthesis regulators PrrA and FnrL under anaerobic conditions was captured by iRsp1140-rFBA but not iRsp1140-PROM, suggesting that certain phenotypes might be better captured by Boolean logic. However, it should be noted that when the κ parameter (i.e., penalty for deviation from flux bounds) was increased from 1 to 1.5 in the PROM simulations, the phenotypes of these two mutants were also captured by iRsp1140-PROM, but its overall predictive performance dropped to 70.6% agreement with experimental data. Intermediate values of κ might result in improved prediction by iRsp1140-PROM, but these were not tested.

Metabolic gene deletion phenotypes - iRsp1140, iRsp1140-rFBA and iRsp1140-PROM performed similarly at predicting the outcomes of metabolic gene deletions, correctly predicting 85.9, 83.1 and 83.1% of the growth phenotypes respectively (Table 8-2). This indicates that the constraints of the metabolic model were likely still the most important factors for predicting the outcomes of these genetic perturbations.

Substrate utilization phenotypes - iRsp1140 and iRsp1140-rFBA performed significantly better that iRsp1140-PROM for predicting the ability of *R. sphaeroides* to grow with various carbon sources under aerobic respiratory and anaerobic photosynthetic conditions, with agreements of 91.7 and 85% respectively aerobically (compared to 66.7% by iRsp1140-PROM), and 85 and 83.3% respectively under photosynthetic conditions (compared to 75% by iRsp1140-PROM) (Table 8-3). Conversely, iRsp1140-PROM performed better at predicting substrate utilization under anaerobic respiratory conditions with an agreement of 66.7% with experimental data compared to 45 and 47.6% for iRsp1140 and iRsp1140-rFBA (Table 8-3).

Condition-dependent growth phenotypes – A few other condition-dependent growth phenotypes such as the inhibition of H_2 production by NH_3 and O_2, as well as the inhibition of photosynthetic growth by O_2 were assessed, with the integrated models being better equipped to predict these phenotypes than iRsp1140 alone (Table 8-4).

Concluding remarks

Overall, the integrated models enable analysis of a wider range of growth phenotypes than iRsp1140, significantly increasing its scope and potential utility. However, incorporation of these TRN constraints led to a reduction of some of the predictive accuracy of iRsp1140, particularly in substrate utilization analysis. It should be noted that both iRsp1140-rFBA and iRsp1140-PROM are first generation integrated models, which underwent only minimal refinement before assessment of their predictive capabilities (compared to iRsp1140, which is the result of over 2 years of refinement). Given the fact that the TRN

used to build the integrated models is inferred mostly from high-throughput datasets, it might contain errors (see Table 8-5) that may negatively impact the performance of the integrated models. Furthermore, if the full complement of TFs controlling the expression of a given metabolic gene has not been determined, predictions for TF deletion phenotypes might suffer, as genes might be erroneously turned off due to a TF deletion, because its alternative regulators are not yet known. An additional factor that limits the performance of these integrated models is the very limited amount of data available about the effectors of the different TFs in *R. sphaeroides*. If for instance, a transcriptional repressor is permitted to be on under conditions when it should be off (given the presence/absence of its effector), all reactions regulated by this TF may erroneously be turned off, due to lack of information, negatively impacting predictions. These limitations would have significantly more impact on Boolean based approaches like rFBA, but they would also alter the performance of PROM-based integrated models. Thus, to improve the predictive performance of iRsp1140-rFBA and iRsp1140-PROM, iterative refinements of both the TRN and the integrated models will be required to bring them into better agreement with observed growth and gene deletion phenotypes. These refinements could simply involve the omission of a subset of the regulatory rules from the TRN, to bring the integrated models' predictions into better agreement with available data. Alternatively, as more information becomes available about regulatory interactions and TF effectors, these could be incorporated into the models to potentially improve their predictive capabilities. In addition, more extensive datasets will likely need to be generated, which would allow for a more rigorous assessment of the predictive performance of these models and serve as a data source for further refinements. Nevertheless, the performance of iRsp1140, iRsp1140-rFBA and iRsp1140-PROM under different conditions indicate that they and any subsequent refinements should serve as useful tools for studying *R. sphaeroides* physiology.

Table 8-1. TF deletion phenotypes under different energetic conditions

Deleted Genes[a]	Experimental[b]			iRsp1140			iRsp1140-rFBA			iRsp1140-PROM			Ref
	Aero	Photo	DMSO	Aero	Photo	DMSO	Aero	Photo	DMSO	Aero	Photo	DMSO	
prrA	NE	E	NE	NA	NA	NA	✓	✓	×	✓	×	✓	[6]
fnrL	NE	E	E	NA	NA	NA	✓	✓	✓	✓	×	×	[7]
RSP_2888	NE	NE	NA	NA	NA	NA	✓	✓	NA	✓	✓	NA	This study
RSP_3341	NE	NE	NA	NA	NA	NA	✓	✓	NA	✓	✓	NA	This study
RSP_0489	NE	NE	NA	NA	NA	NA	✓	✓	NA	✓	✓	NA	This study
ccmR	NE*	NE*	NA	NA	NA	NA	×	×	NA	×	×	NA	This study
akgR	NE	NE	NA	NA	NA	NA	✓	✓	NA	✓	✓	NA	This study
akgR (α-kg)	NE*	NE*	NA	NA	NA	NA	✓	×	NA	✓	×	NA	This study
crpK	NE	NE	NA	NA	NA	NA	✓	✓	NA	✓	✓	NA	This study
ppsR	NE	NE	NE	NA	NA	NA	✓	✓	✓	✓	✓	✓	[6]
appA	NE	NE*	NA	NA	NA	NA	✓	✓	NA	✓	✓	NA	[6]
prrA-ppsR	NE	NE	NE	NA	NA	NA	✓	×	×	✓	✓	✓	[6]
cbbR	NE	E	NA	NA	NA	NA	×	×	NA	✓	×	NA	Burger BT†
ccmR-akgR	NE*	NE*	NA	NA	NA	NA	×	×	NA	×	×	NA	This study
nifA	NE	NE	NA	NA	NA	NA	✓	✓	NA	✓	✓	NA	†
nifA (S + G)	NE	E	NA	NA	NA	NA	✓	×	NA	✓	×	NA	†
RSP_0443	NE*	NE*	NA	NA	NA	NA	✓	×	NA	✓	✓	NA	This study
rpoE	NE	NE	NA	NA	NA	NA	✓	✓	NA	✓	✓	NA	[8]
rpoH1	NE	NE	NA	NA	NA	NA	×	×	NA	✓	✓	NA	[9]
rpoH2	NE	NE	NA	NA	NA	NA	✓	×	NA	✓	✓	NA	[9]
rpoH1-rpoH2	NE	NE	NA	NA	NA	NA	×	×	NA	✓	✓	NA	[9, 10]
oxyR	NE	NE	NA	NA	NA	NA	✓	✓	NA	✓	✓	NA	[11]
rpoD	E	E	E	NA	NA	NA	✓	✓	✓	✓	✓	✓	
Totals	**51 Phenotypes**			0	0	0	18	13	3	21	16	4	
	Agreement			0			**34** (66.7%)			**41** (80.4%)			

[a] Growth media for experiments and simulations was Sistrom's minimal medium (SMM) [12], except where indicated (α-kg = α-ketoglutarate, S+G = succinate+glutamate).

[b] E – Essential; NE – Non-Essential; NA – Not Applicable (i.e., growth condition not tested experimentally and/or computationally). Aero – aerobic respiration; Photo – photosynthesis; DMSO – anaerobic respiratory with DMSO as terminal electron acceptor. If predicted growth rate was ≤ 5% of that of wild type cells, the gene was considered essential.

* Impaired growth, either observed or predicted.

† Personal communication (in the case of the *nifA* deletion mutant, this is a personal observation from a mutant I made but not discussed anywhere else in this thesis)

✓Prediction agrees with experimental observation

×Prediction disagrees with experimental observation

Table 8-2. Metabolic gene deletion phenotypes under different energetic conditions

Deleted Genes[a]	Experimental[b]			iRsp1140			iRsp1140-rFBA			iRsp1140-PROM			Ref
	Aero	Photo	DMSO	Aero	Photo	DMSO	Aero	Photo	DMSO	Aero	Photo	DMSO	
cbbI Operon	NE	NE	NA	✓	✓	NA	✓	✓	NA	✓	✓	NA	[13]
cbbII Operon	NE	NE	NA	✓	✓	NA	×	×	NA	✓	✓	NA	[13]
cbbI-II	NE	E	NA	✓	×	NA	×	✓	NA	✓	×	NA	[13]
nifHDK	NE	NE	NA	✓	✓	NA	✓	✓	NA	✓	✓	NA	[14]
nifHDK (S +G)	NE	E	NA	✓	×	NA	✓	×	NA	✓	×	NA	[14]
qox	NE	NE	NA	✓	✓	NA	✓	✓	NA	×	✓	NA	[15]
fbcFBC	NE*	E	NA	✓	✓	NA	✓	✓	NA	✓	✓	NA	[15]
fbcFBC-qox	E	E	NA	×*	✓	NA	×*	✓	NA	✓	✓	NA	[15]
dorC	NE	NE	E	✓	✓	✓	✓	✓	✓	✓	✓	✓	[16]
nuo (RSP2512)	NE*	E	NA	✓	×	NA	✓	×	NA	×	×	NA	Spero MA†
nuo (RSP0100)	NE	NE	NA	✓	✓	NA	✓	✓	NA	✓	✓	NA	Spero MA†
nuo (Double)	NE*	E	NA	✓	✓	NA	✓	✓	NA	×	✓	NA	Spero MA†
pntAB	NE	NE*	NA	✓	✓	NA	✓	✓	NA	✓	✓	NA	This study
pntAB (A)	NE*	E	NA	✓	×	NA	✓	×	NA	✓	×	NA	This study
RSP_0745 (A)	NE*	NE*	NA	✓	✓	NA	✓	✓	NA	✓	✓	NA	[17]
RSP_0745	NE	NE	NA	✓	✓	NA	✓	✓	NA	✓	✓	NA	[17]
RSP_0960 (A)	NE*	NE*	NA	✓	×	NA	✓	×	NA	×	×	NA	[17]
RSP_0960	NE	NE	NA	✓	✓	NA	✓	✓	NA	✓	✓	NA	[17]
RSP_0961 (A)	E	E	NA	×*	✓	NA	×*	✓	NA	✓	✓	NA	[18]
RSP_0961	NE	NE	NA	✓	✓	NA	✓	✓	NA	✓	✓	NA	[18]
RSP_1679 (A)	E	E	NA	×*	✓	NA	×*	✓	NA	✓	✓	NA	[19]
RSP_1679	NE	NE	NA	✓	✓	NA	✓	✓	NA	✓	✓	NA	[19]
RSP_1771 (A)	E	E	NA	✓	✓	NA	✓	✓	NA	✓	✓	NA	[17, 20]
RSP_1771	NE	NE	NA	✓	✓	NA	✓	✓	NA	✓	✓	NA	[17, 20]
RSP_0970 (A)	E	E	NA	✓	✓	NA	✓	✓	NA	✓	✓	NA	[17, 20]
RSP_0970	NE	NE	NA	✓	✓	NA	✓	✓	NA	✓	✓	NA	[17, 20]
RSP_2191 (A)	NE*	NE*	NA	✓	✓	NA	✓	✓	NA	✓	✓	NA	Alber BE†
RSP_2191	NE	NE	NA	✓	✓	NA	✓	✓	NA	✓	✓	NA	Alber BE†
RSP_2189 (A)	NE*	NE*	NA	✓	✓	NA	✓	✓	NA	✓	✓	NA	Alber BE†
RSP_2189	NE	NE	NA	✓	✓	NA	✓	✓	NA	✓	✓	NA	Alber BE†
RSP_2192 (A)	NE*	NE*	NA	✓	✓	NA	✓	✓	NA	×	✓	NA	Alber BE†
RSP_2192	NE	NE	NA	✓	✓	NA	✓	✓	NA	✓	✓	NA	Alber BE†
RSP_0653 (A)	E	E	NA	×	×	NA	×	×	NA	×	×	NA	Alber BE†
RSP_0653	NE	NE	NA	✓	✓	NA	✓	✓	NA	✓	✓	NA	Alber BE†
hupSL	NE	NE	NA	✓	✓	NA	✓	✓	NA	✓	✓	NA	
Totals	**71 Phenotypes**			31	29	1	29	29	1	29	29	1	
	Agreement			**61 (85.9%)**			**59 (83.1%)**			**59 (83.1%)**			

[a] Growth media for experiments and simulations was Sistrom's minimal medium (SMM) [12], except where indicated (A = Acetate, S+G = succinate+glutamate).

[b] E – Essential; NE – Non-Essential; NA – Not Applicable (i.e., growth condition not tested experimentally and/or computationally). Aero – aerobic respiration; Photo – photosynthesis; DMSO – anaerobic respiratory with DMSO as

terminal electron acceptor. If predicted growth rate was ≤ 5% of that of wild type cells, the gene was considered essential.
* Impaired growth, either observed or predicted.
† Personal communication
✓Prediction agrees with experimental observation
×Prediction disagrees with experimental observation

Table 8-3. Substrate utilization phenotypes under varying growth conditions

	Aero[a]				Photo				DMSO			
Nutrient	**Biolog**	**PROM**	**rFBA**	**iRsp1140**	**Biolog**	**PROM**	**rFBA**	**iRsp1140**	**Biolog**	**PROM**	**rFBA**	**iRsp1140**
L-Lysine	G	×	✓	✓	G	×	✓	✓	NG	✓	×	×
L-Glutamine	G	✓	✓	✓	G	✓	✓	✓	NG	✓	×	×
L-L eucine	G	×	✓	✓	G	✓	✓	✓	G	×	✓	✓
L-Isoleucine	G	×	✓	✓	NG	×	×	×	NG	✓	×	×
L-Valine	G	✓	✓	✓	G	✓	✓	✓	NG	×	×	×
L-Proline	G	✓	✓	✓	G	✓	✓	✓	NG	×	×	×
D-Sorbitol	G	✓	✓	✓	G	✓	✓	✓	G	✓	✓	✓
D-Mannitol	G	✓	✓	✓	G	✓	✓	✓	G	✓	✓	✓
Glycerol	G	✓	✓	✓	G	✓	✓	✓	NG	×	×	×
D-Ribose	G	✓	✓	✓	G	✓	✓	✓	NA	NA	NA	NA
D-Xylose	G	✓	×	✓	G	✓	×	✓	G	✓	×	✓
D-Fructose	G	✓	✓	✓	G	✓	✓	✓	G	✓	✓	✓
L-Glutamic Acid	G	✓	✓	✓	G	✓	✓	✓	G	✓	✓	✓
L-Asparagine	G	✓	✓	✓	G	✓	✓	✓	NG	✓	×	×
Succinic Acid	G	✓	✓	✓	G	✓	✓	✓	G	✓	✓	✓
D,L-Malic Acid	G	✓	✓	✓	G	✓	✓	✓	G	✓	✓	✓
Fumaric Acid	G	✓	✓	✓	G	✓	✓	✓	G	✓	✓	✓
Pyruvic Acid	G	✓	✓	✓	G	✓	✓	✓	G	✓	✓	✓
α-Ketovalerate	G	×	✓	✓	NG	×*	×	×	NA	NA	NA	NA
L-Alanine	G	✓	✓	✓	G	✓	✓	✓	G	✓	✓	✓
D-Mannose	G	✓	✓	✓	G	✓	✓	✓	NG	×	×	×
α-D –Glucose	G	✓	✓	✓	G	✓	✓	✓	NG	×	×	×
Acetic Acid	G	×	✓	✓	G	✓	✓	✓	G	×	✓	✓
L-Serine	G	✓	✓	✓	G	✓	✓	✓	NG	✓	✓	✓
Glyoxylic Acid	G	✓	✓	✓	NG	×	×	×	NG	✓	✓	✓
Formic Acid	G	×	×	✓	G	×	✓	✓	NG	✓	✓	✓
L-Lactic Acid	G	✓	✓	✓	G	✓	✓	✓	NG	×	×	×
Thymidine	G	×	✓	✓	G	×	✓	✓	NG	✓	×	×
Propionic Acid	G	×	✓	✓	G	×	✓	✓	NG	✓	×	×
D-Arabitol	G	✓	✓	✓	G	✓	✓	✓	NA	NA	NA	NA
Adonitol (Ribitol)	G	✓	✓	✓	G	✓	✓	✓	G	✓	✓	✓
Glycolic Acid	NG	✓	×	×	G	✓	✓	✓	NG	✓	×	×
α-Ketoglutarate	G	✓	✓	✓	G	✓	✓	✓	G	✓	✓	✓
Methanol	G	×	✓	✓	G	✓	✓	✓	NA	NA	NA	NA

2-Aminoethanol	G	×	✓	✓	G	✓	✓	✓	NG	✓	✓	✓
α-Hydroxybutyrate	G	×	✓	✓	NG	✓	×	×	NA	NA	NA	NA
L-Tartaric Acid	G	✓	✓	✓	G	✓	✓	✓	NG	×	×	×
Oxalo malic Acid	G	✓	✓	✓	G	✓	✓	✓	NA	NA	NA	NA
Palatinose	G	✓	✓	✓	NG	×	×	×	NA	NA	NA	NA
D-Tartaric Acid	NG	×	×	×	G	✓	✓	✓	NG	×	×	×
Citramalic Acid	NG	×	×	×	G	✓	✓	✓	NA	NA	NA	NA
Dextrin	G	✓	✓	✓	NG	×	×	×	NA	NA	NA	NA
Acetoin	G	×	✓	✓	NG	×	×	×	NA	NA	NA	NA
δ-Aminovalerate	G	✓	✓	✓	G	×*	✓	✓	NG	✓	×	×
γ-Aminobutyrate	G	×	✓	✓	G	×	✓	✓	NG	✓	×	×
D-Galacturonate	G	✓	×	✓	G	✓	×	✓	NA	NA	NA	NA
Uridine	G	×	✓	✓	G	×	✓	✓	NG	✓	×	×
Inosine	NG	×	×	×	G	✓	✓	✓	NG	×	×	×
Butyric Acid	G	✓	✓	✓	G	✓	✓	✓	G	×	✓	✓
Acetoacetic Acid	G	×	✓	✓	NG	✓	×	×	NA	NA	NA	NA
β-Hydroxybutyrate	G	×	✓	✓	G	×	✓	✓	NG	✓	×	×
5-Keto-D-Gluconate	G	✓	×	✓	NG	×	✓	×	NA	NA	NA	NA
L-Galactonic Acid-γ-	G	✓	✓	✓	G	✓	✓	✓	NA	NA	NA	NA
Glycyl-L-Proline	G	✓	✓	✓	G	✓	✓	✓	NA	NA	NA	NA
L-Alanyl-Glycine	G	✓	✓	✓	G	✓	✓	✓	NA	NA	NA	NA
Glycyl-L-Aspartate	G	✓	✓	✓	G	✓	✓	✓	NG	×	×	×
Succinamic Acid	G	✓	✓	✓	G	✓	✓	✓	NG	×	×	×
L-Alaninamide	G	✓	✓	✓	G	✓	✓	✓	NG	✓	✓	✓
Glycyl-L-Glutamate	G	✓	✓	✓	G	✓	✓	✓	NA	NA	NA	NA
Mucic Acid	NG	×	×	×	G	✓	✓	✓	NA	NA	NA	NA
Totals	**60**	**40** **66.7%**	**51** **85%**	**55** **91.7%**	**60**	**45** **75%**	**50** **83.3%**	**51** **85%**	**42**	**28** **66.7%**	**19** **45%**	**20** **47.6%**

[a] G – Growth; NG – No growth; NA – Not Applicable (i.e., growth condition not tested experimentally and/or computationally). Aero – aerobic respiration; Photo – photosynthesis; DMSO – anaerobic respiratory with DMSO as terminal electron acceptor. Data obtained from Biolog analysis for *R. sphaeroides* [21].
* Impaired growth, either observed or predicted.
✓Prediction agrees with experimental observation
×Prediction disagrees with experimental observation

Table 8-4. Other predicted condition-dependent growth phenotypes

Phenotypes	iRsp1140	iRsp1140-rFBA	iRsp1140-PROM
Photoautotrophic growth on CO_2, H_2 & NH_3	✓	✓	✓
hupSL essentiality for photoautotrophic growth	✓	✓	✓
Inhibition of photosynthetic growth by O_2	×	✓	×*
Inhibition of H_2 production by ammonia	×	✓	✓
Inhibition of H_2 production by oxygen	×	✓	✓
No H_2 production on glutamate only	✓	✓	✓
6 Phenotypes	3	6	5

* Impaired growth.

Summary

At the start of my thesis research, I set 3 main objectives to gain a greater systems-level understanding of the metabolically versatile microbe, *R. sphaeroides*. These objectives were: (i) construction, validation and refinement of a genome-scale metabolic model for *R. sphaeroides*; (ii) inference, validation and refinement of a large-scale transcriptional regulatory network (TRN) for *R. sphaeroides*; and (iii) construction of an integrated model of metabolism and transcriptional regulation for *R. sphaeroides*. Over the course of my studies, I have successfully applied a variety of approaches, both experimental and computational, to achieve each of these objectives.

I. Metabolic modeling of R. sphaeroides

To gain a systems-level understanding of the metabolic capabilities of *R. sphaeroides*, as well as to enable computational modeling of its metabolism, I built, validated and refined genome-scale models of the *R. sphaeroides* metabolic network (Chapters 2 – 4) [21, 22]. Using information gleaned from its genome, literature and databases, as well as the inclusion of a minimal number of reactions to fill in gaps in our knowledge of *R. sphaeroides* metabolic capabilities, I built iRsp1095, the first manually curated genome-scale model for *R. sphaeroides* and the first for an anoxygenic photosynthetic bacterium [22]. iRsp1095's predictions of growth rate and production of metabolites like PHB, H_2 and CO_2 were validated experimentally with data obtained from cells grown in continuous culture, with predictions being in good agreement with observed data.

I then conducted a series of high-throughput phenotype microarray experiments to determine substrate utilization profile of wild type and mutant *R. sphaeroides* strains [21]. In addition, to broadening our knowledge of the *R. sphaeroides* metabolic repertoire, these data were used to significantly refine and extend iRsp1095, leading to the generation of iRsp1140 [21]. iRsp1140, which consists of 1416 reactions, 878 metabolites and 1140 genes, captures all of *R. sphaeroides'* known and inferred metabolic capabilities and includes experimentally determined, condition-dependent biomass and maintenance

energy parameters that significantly improve its predictions over iRsp1095. Growth rate predictions from iRsp1140 are in better agreement with data obtained from chemostat grown cells than iRsp1095 ($R \approx 0.92$, compared to ≈ 0.75 for iRsp1095). Predictions from iRsp1140 also show significantly better agreement to the results from phenotype microarray screens. iRsp1140 also makes predictions for PHB, H_2 and lipid production and the accuracy for these and other predicted metabolic fluxes improves significantly with inclusion of a measured light uptake rate constraint.

Overall iRsp1140 represents a useful systems-level model of the *R. sphaeroides* metabolic network that can be used for qualitative and quantitative assessment of growth rate, growth phenotypes, metabolic flux distributions and metabolite production rates. iRsp1140 can also be used to guide strain design to improve the production of value-added commodities and iRsp1095 has already been used to generate genetic strategies for improving the production of H_2, PHB and phospholipids through the use of sequential gene deletion analysis (see Figure 8-1), though these prediction are yet to be investigated experimentally. Like other genome-scale models, iRsp1140 can be integrated with high-throughput genomic data sets to better understand metabolic activities under specific growth conditions [21]. Thus, iRsp1140 has and should continue to serve as a useful tool for analyzing metabolism in *R. sphaeroides* and related bacteria. To make iRsp1140 accessible to the public in a user friendly format I, in collaboration with Yury Bukhman (GLBRC IT) and Christopher Henry (DOE Kbase) have uploaded iRsp1140 to the DOE Kbase server, which has a user friendly interface and a wide variety of constraint-based modeling tools for analysis of genome-scale metabolic models. I expect that as new data become available on *R. sphaeroides* metabolic capabilities, iRsp1140 will continue to be refined and updated to improve its predictive value.

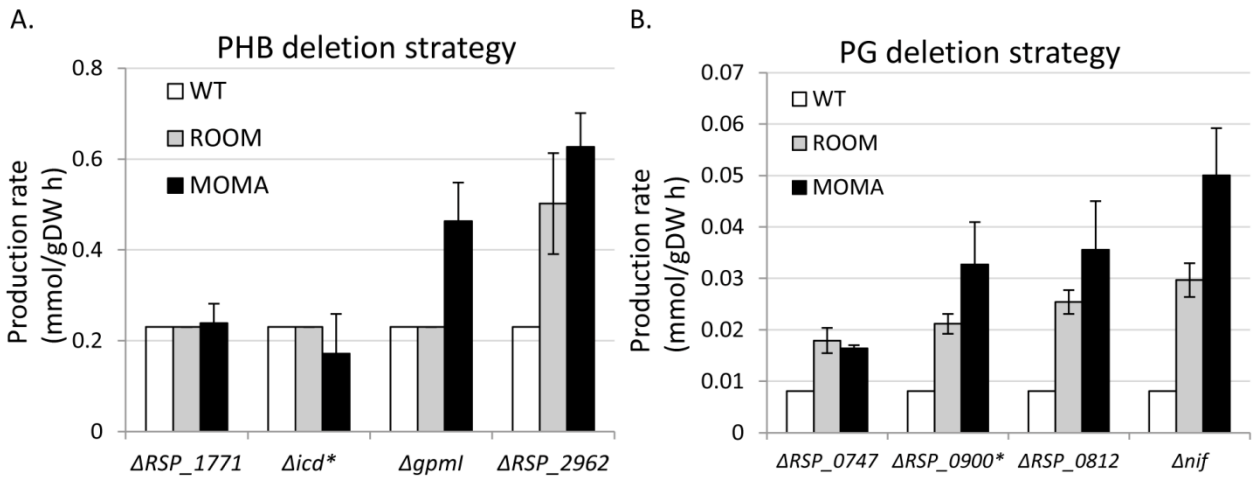

Figure 8-1. Deletion strategies for improving PHB and phospholipid production. iRsp1095 predicted gene deletion strategy for improving PHB (A) and phosphatidylglycerol (PG) (B) production involving the sequential deletion of 4 genes. * Two isozymes of Icd (RSP_1556 and RSP_0446) and another isozymes for RSP_0900 (i.e., RSP_2793) exist in *R. sphaeroides*. Error bars represent standard error from 1000 alternative optimal solutions. RSP_0747 – (R)-3-Hydroxybutanoyl-CoA:NADP+ oxidoreductase ; RSP_0900 – Ferredoxin:NAD oxidoreductase ; RSP_0812 – methylmalonyl-CoA epimerase; *nif* – nitrogenase; RSP_1771 – methylmalonyl-CoA carboxy-lyase; icd – Isocitrate dehydrogenase; RSP_0934 – phosphoglyceromutase; RSP_2962 – (S)-Methylmalonate semialdehyde:NAD+ oxidoreductase (CoA-propanoylating).

II. Reconstruction of a large-scale R. sphaeroides transcriptional regulatory network (TRN)

Due to *R. sphaeroides'* metabolic and bioenergetic versatility, coordination of its activities relies on a robust TRN, which facilitates transitions between metabolic states and lifestyles in response to changing internal or external conditions. To obtain a global view of this TRN, I developed a new integrated workflow that combined available gene expression datasets, sequence information from closely related bacteria and intrinsic properties of bacterial transcription factors (TFs) (Chapter 5). This workflow enabled me to generate a large-scale TRN for *R. sphaeroides* consisting of 120 gene clusters, 1211 genes (including 93 TFs), 1858 regulatory interactions and 76 regulatory motifs. This TRN showed a high degree of functional coherence and captured several modules known to occur in *R. sphaeroides* and/or other bacteria. Given that the vast majority of predictions in the reconstructed TRN represented novel interactions, I began validating these predictions by testing the function of TFs predicted in the TRN to be involved in regulation of key cellular processes, in particular photosynthesis and carbon metabolism, as these is are integral parts of *R. sphaeroides* physiology.

Re-evaluation of the R. sphaeroides photosynthetic TRN

From previous analysis it was known that PpsR, FnrL and PrrA were important transcriptional regulators of photosynthesis in *R. sphaeroides* [7, 23-29]. Consistent with this, my reconstructed TRN predicted regulons for these 3 TFs consisting of a significant number of photosynthetic genes but also included several previously unknown targets for these TFs. Using transcriptomics and ChIP-seq analyses, I experimentally verified several of these new targets. I identified a total of 101 operons (240 genes) as direct targets of these 3 TFs, significantly broadening their regulons (Chapters 5 and 6). In addition to these previously studied TFs, predictions from my TRN indicated that at least 2 previously uncharacterized proteins, CrpK and RSP_2888, were involved in regulation of photosynthesis. I used a combination of physiological, genetic and genomic approaches to determine the regulons of these TFs and their DNA binding determinants (in the case of CrpK), while proposing roles for these 2 TFs in

photosynthesis in *R. sphaeroides*. These analyses significantly extended prior knowledge of the transcriptional regulation of photosynthesis, while identifying mechanisms for robustness in the photosynthetic TRN. These data also provided information on new links between the regulation of photosynthesis and other central metabolic functions such as the maintenance of iron homeostasis in *R. sphaeroides*.

Studying transcriptional regulation of central carbon and energy metabolism

My analysis of *R. sphaeroides'* metabolic activities, both predicted (iRsp1140) and observed (using phenotypic microarrays), showed that it is able to use a larger number of carbon sources than previously reported. While the transcriptional control of central carbon metabolism had previously been studied in γ- and β-Proteobacteria, no similar analysis had been conducted in α-Proteobacteria. Furthermore, no proteins analogous to TFs that control central carbon metabolism in other bacteria had been identified in α-Proteobacteria. Predictions from the large-scale TRN indicated 2 previously uncharacterized proteins played important roles in regulation of central carbon metabolism in *R. sphaeroides* and other α-Proteobacteria: CcmR and AkgR. I used a combination of physiological, genetic and genomic approaches to study the functional roles of these proteins and map their regulons. These analyses verified the important role played by these 2 TFs in regulating transcription of genes encoding key enzymes involved in glycolysis/gluconeogenesis, the TCA cycle, ATP synthesis and other aspects of central metabolism. These TFs could serve as useful targets for maximizing or controlling the flux through central carbon metabolism pathways in *R. sphaeroides* and related bacteria. In addition, homologues of these proteins exist in other α-Proteobacteria, where they are predicted from my comparative genomics analysis to also be involved in regulation central carbon metabolism. Thus, this TRN provided new insights into central carbon metabolism and has the potential to provide similar insights into other important pathways in *R. sphaeroides* and other bacteria.

Overall performance of reconstructed TRN

In addition to the above mentioned TFs, I used genomic analysis to map the regulons of RSP_0489 (an UxaR analog involved in regulating the metabolism of carboxylic acids) and RSP_3341 (a RirA-like protein involved in regulating iron-dependent genes), verifying several predictions from the TRN (Chapter 5). Thus, I studied a total of 9 TFs to validate the reconstructed TRN.

Given the limited amount of gene expression data used in construction of the large-scale *R. sphaeroides* TRN, this network can likely be improved. To gain an idea of the accuracy level of a TRN's predictions, known regulatory interactions can be compared to predictions to calculate the precision (percentage of predictions made that are true positives based on a reference dataset) and recall (percentage of all known interactions in the reference dataset predicted in the TRN) of the TRN predictions. The precision and recall obtained for each of the 9 TFs and cumulatively were calculated (Table 8-5). Based on this analysis, the precision of the *R. sphaeroides* TRN for each TF ranges from 30 to 100% while the recall ranges from 19 to 87%. Overall, predictions for the TRN had a precision of ~60% and a recall of ~50% (i.e., ~60% of the TRNs predictions were correct and it captures ~50% of target regulons). It should however be noted that the genomic analyses used for validation here are subject to their own limitations (i.e., false positives and negatives). For instance, several TF binding sites predicted from the ChIP-seq analyses were not considered as direct targets because of a lack of differential expression upon deleting the target TF or due to the absence of a predicted shared motif. Thus, the calculated precision and recall of this TRN (Table 8-5) should only be considered as approximations. Nevertheless, these numbers indicate the TRN provides reasonably high accuracy and coverage of *R. sphaeroides* regulatory networks.

Table 8-5. Precision and recall of the large-scale *R. sphaeroides* TRN[a]

TF	TP	Predicted	TPP	FPP	Precision	Recall
PpsR	39	35	34	1	0.97	0.87
FnrL	110	91	65	26	0.71	0.59
PrrA	91	29	17	12	0.59	0.19
CrpK	62	69	27	42	0.39	0.44
RSP_2888	26	30	9	21	0.30	0.35
RSP_3341	9	7	7	0	1.00	0.78
RSP_0489	19	19	11	8	0.58	0.58
CcmR	32	24	15	9	0.63	0.47
AkgR	12	13	9	4	0.69	0.75
Overall	**400**	**317**	**194**	**123**	**0.61**	**0.49**

[a] TP – True positives (i.e., experimentally verified interactions), Predicted – Total number of TRN predicted interactions, TPP – True positive predictions (i.e., total number of TRN predictions that are TP), FPP – False positive predictions (i.e., total number of TRN predictions that are not in the TP dataset), Precision – TPP/Predicted, Recall – TPP/TP

III. Integrated modeling of metabolism and transcriptional regulation

To study the connections between metabolism and transcriptional regulation in *R. sphaeroides*, I integrated iRsp1140 with regulatory interactions predicted in the large-scale TRN to generate integrated regulatory-metabolic models for *R. sphaeroides*. I implemented this using 2 previously published approaches – rFBA [2] and PROM [1]. To assess the predictive performance of the integrated models relative to iRsp1140, I compared the predictions of all 3 models to known gene deletion and condition-dependent growth phenotypes, which I assembled from literature and my own analysis. Preliminary analyses of these integrated models show that they perform well at predicting the phenotypes of mutants lacking individual TFs and metabolic proteins. However, the addition of TRN information into integrated models negatively impacted some of the substrate utilization predictions of iRsp1140, indicating additional refinements will be required to improve their predictive accuracy.

Overall these models of metabolism and transcriptional regulation, as well as the integrated models, have proved to useful representations of the respective *R. sphaeroides* networks, capable of making both qualitative and quantitative predictions about various aspects of its physiology and will hopefully prove useful in current and future attempts to study and engineer this organism for production of valuable commodities. It also seems likely that the approaches I have taken will be useful in modeling or experimentally analyzing the lifestyles of other organisms.

Future directions

My thesis research was focused on building systems-level models for studying the physiology of *R. sphaeroides*. Overall, it resulted in generation of useful models for studying metabolism and transcriptional regulation in this well-studied bacterium. As with all models of living systems, they capture what we currently know or can predict about *R. sphaeroides* metabolic and transcriptional regulatory processes, but they are inherently incomplete due to limitations in our current knowledge. Thus, as new information becomes available about *R. sphaeroides* metabolic and regulatory capabilities, it can be used to iteratively refine and extend the existing models. For instance, very little information is available on the impact of metabolites as allosteric regulators of enzyme and TF behavior in *R. sphaeroides*, limiting the scope and predictive performance of iRsp1140 and the integrated models. As these data become available they can be incorporated into these models and potentially improve their predictive performance and/or ability to inform our understanding of this and related bacteria.

Utilizing models for strain design

A widely used application of CBMs is in the development of genetic strategies to improve production of desirable commodities in bacteria. This could involve identification of gene deletions required to divert metabolic flux towards the production of a commodity of interest or the addition of genes to the metabolic network encoding functions that facilitate production of a desired substance or a combination of these and other approaches. I previously used iRsp1095 for predicting such genetic strategies aimed at improving the production of PHB, H_2 and the phospholipid phosphatidylglycerol (as a surrogate for fatty acid production), with some promising results (Figure 8-1). These predictions were made through the analysis of sequential reaction deletions from iRsp1095 using MOMA or ROOM [30, 31], in combination with sampling the optimal solution space. Overall, these predictions highlight the utility of iRsp1095 for guiding metabolic engineering in *R. sphaeroides*. However, similar analyses have yet to be conducted with iRsp1140 or any of the integrated regulatory-metabolic models, nor have any of these predictions

been tested experimentally. Furthermore, a wide variety of alternative algorithms have been developed and successfully implemented for strain design [32-34], which have also not used with any of the *R. sphaeroides* models. Thus, future efforts aimed at improving the yield of substances like PHB, H_2 and phospholipids, or even compounds not natively produced by *R. sphaeroides*, should take advantage of these models and algorithms to develop strategies that will guide these experimental efforts.

Assessment of TRN predictions

As part of my analyses of the *R. sphaeroides* regulatory processes, I built a large-scale model of its TRN, which I began to validate via the use genomic analyses (Chapters 5-7). However, I only assessed the predictions for 9 TFs, while the TRN includes predictions for over 90 TFs, the vast majority of which have never been studied in *R. sphaeroides* or any other bacteria. Thus, future efforts at understanding uncharacterized aspects of transcriptional regulation in *R. sphaeroides* and related bacteria could use these predictions as starting points in their analysis. Just like the CBMs constructed as part of my thesis, this large-scale TRN will also need to go through iterative steps of validation and refinement to improve its quality. Thus, future efforts should be directed at studying specific TFs and regulatory modules of interest.

Validating and refining regulatory-metabolic models

While I conducted an initial assessment of the performance of the integrated regulatory-metabolic models, iRsp1140-rFBA and iRsp1140-PROM, this was based on a limited set of available mutants and phenotypes. To conduct a more rigorous assessment of the predictive capabilities of these models, larger scale datasets may be required. Work is underway to develop genome-scale, condition-dependent gene essentiality data for *R. sphaeroides* using high-throughput methods (Burger BT, personal communication). Data from these and other analyses can be used to obtain a more global assessment of the predictions generated from these models and potentially enable significant refinements to improve their predictive capabilities.

Alternative approaches to integrating CBMs and TRNs

Given the fact that cellular processes such as signal transduction, transcription, translation and metabolism are inter-dependent, meaningful ways of integrating data that monitor these processes will be important to increasing the scientific insight gained from computational models of living systems. In this chapter, I described the construction of integrated regulatory-metabolic models for *R. sphaeroides* using previously published approaches [1, 2], with models capturing several condition-dependent and TF deletion phenotypes not currently obtainable with the standalone metabolic model. My studies documented an increase in the predictive power of these integrated models, but they also confirm some of the limitations of integrating CBMs and TRNs with rFBA and PROM (as discussed in Chapter 1). Thus, alternative approaches to integrating CBMs and TRNs would be very useful to the scientific community. Such approaches could involve the use of gene and/or protein expression levels as constraints, within context of the regulatory rules of the TRN and the GPR associations of the CBM. In such a formulation, the allowable flux bounds of a given reaction could be dependent on some factor derived from an integration of the expression levels of the regulating TFs and catalyzing reactions. Here, instead of TFs having binary states as used in the rFBA and PROM implementations, they could potentially be represented continuous values derived from expression data. This could potentially allow for convenient simulation of gene deletion and/or over-expression phenotypes.

Concluding remarks

As technological advancements have improved our ability to quantitatively monitor cellular processes, studying living systems from a global perspective is both feasible and gaining increased importance for the analysis of cells, multi-cellular organisms and entire ecosystems. These data can now allow us to build mathematical models of how biological systems work, enabling computational simulations that my help guide future scientific research or provide insight into biological phenomena which are not testable by available experimental approaches. My thesis work involved a systems-level analysis of microbial life, focusing on the key processes of metabolism and transcriptional regulation in a model phototrophic bacterium. By harnessing different available datasets for *R. sphaeroides*, leveraging well-established computational approaches and conducting complementary experiments, I constructed, validated, extended, refined and integrated large-scale models of metabolism and transcriptional regulation for this microbe. These models, as well as experiments conducted in validation of hypotheses generated from them, have provided new global insights into both well-studied and previously uncharacterized aspects of *R. sphaeroides* physiology. I expect that these models and their predictions will continue to serve as useful resources to scientific researchers, particularly those interested in microbial physiology.

References

1. Chandrasekaran S, Price ND: **Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia coli* and *Mycobacterium tuberculosis***. *Proc Natl Acad Sci U S A* 2011, **107**(41):17845-17850.
2. Covert MW, Palsson BO: **Transcriptional regulation in constraints-based metabolic models of *Escherichia coli***. *J Biol Chem* 2002, **277**(31):28058-28064.
3. Shlomi T, Eisenberg Y, Sharan R, Ruppin E: **A genome-scale computational study of the interplay between transcriptional regulation and metabolism**. *Mol Syst Biol* 2007, **3**:101.
4. Masepohl B, Hallenbeck PC: **Nitrogen and molybdenum control of nitrogen fixation in the phototrophic bacterium *Rhodobacter capsulatus***. *Adv Exp Med Biol* 2010, **675**:49-70.
5. Hunter CN, Daldal, F., Thurnauer, M. C. and Beatty, J. T.: **The purple phototrophic bacteria**, vol. 28: Springer; 2009.
6. Gomelsky L, Moskvin OV, Stenzel RA, Jones DF, Donohue TJ, Gomelsky M: **Hierarchical regulation of photosynthesis gene expression by the oxygen-responsive PrrBA and AppA-PpsR systems of *Rhodobacter sphaeroides***. *J Bacteriol* 2008, **190**(24):8106-8114.
7. Zeilstra-Ryalls JH, Kaplan S: **Aerobic and anaerobic regulation in *Rhodobacter sphaeroides* 2.4.1: the role of the fnrL gene**. *J Bacteriol* 1995, **177**(22):6422-6431.
8. Anthony JR, Warczak KL, Donohue TJ: **A transcriptional response to singlet oxygen, a toxic byproduct of photosynthesis**. *Proc Natl Acad Sci U S A* 2005, **102**(18):6502-6507.
9. Green HA, Donohue TJ: **Activity of *Rhodobacter sphaeroides* RpoHII, a second member of the heat shock sigma factor family**. *J Bacteriol* 2006, **188**(16):5712-5721.
10. Dufour YS, Imam S, Koo BM, Green HA, Donohue TJ: **Convergence of the transcriptional responses to heat shock and singlet oxygen stresses**. *PLoS Genet* 2012, **8**(9):e1002929.
11. Zeller T, Klug G: **Detoxification of hydrogen peroxide and expression of catalase genes in *Rhodobacter***. *Microbiology* 2004, **150**(Pt 10):3451-3462.
12. Sistrom WR: **A requirement for sodium in the growth of *Rhodopseudomonas spheroides***. *J Gen Microbiol* 1960, **22**:778-785.
13. Hallenbeck PL, Lerchen R, Hessler P, Kaplan S: **Roles of CfxA, CfxB, and external electron acceptors in regulation of ribulose 1,5-bisphosphate carboxylase/oxygenase expression in *Rhodobacter sphaeroides***. *J Bacteriol* 1990, **172**(4):1736-1748.
14. Kontur WS, Ziegelhoffer EC, Spero MA, Imam S, Noguera DR, Donohue TJ: **Pathways involved in reductant distribution during photobiological H_2 production by *Rhodobacter sphaeroides***. *Appl Environ Microbiol* 2011, **77**(20):7425-7429.
15. Mouncey NJ, Gak E, Choudhary M, Oh J, Kaplan S: **Respiratory pathways of *Rhodobacter sphaeroides* 2.4.1(T): identification and characterization of genes encoding quinol oxidases**. *FEMS Microbiol Lett* 2000, **192**(2):205-210.
16. Mouncey NJ, Choudhary M, Kaplan S: **Characterization of genes encoding dimethyl sulfoxide reductase of *Rhodobacter sphaeroides* 2.4.1T: an essential metabolic gene function encoded on chromosome II**. *J Bacteriol* 1997, **179**(24):7617-7624.
17. Alber BE, Spanheimer R, Ebenau-Jehle C, Fuchs G: **Study of an alternate glyoxylate cycle for acetate assimilation by *Rhodobacter sphaeroides***. *Mol Microbiol* 2006, **61**(2):297-309.
18. Erb TJ, Retey J, Fuchs G, Alber BE: **Ethylmalonyl-CoA mutase from *Rhodobacter sphaeroides* defines a new subclade of coenzyme B12-dependent acyl-CoA mutases**. *J Biol Chem* 2008, **283**(47):32283-32293.
19. Erb TJ, Fuchs G, Alber BE: **(2S)-Methylsuccinyl-CoA dehydrogenase closes the ethylmalonyl-CoA pathway for acetyl-CoA assimilation**. *Mol Microbiol* 2009, **73**(6):992-1008.
20. Erb TJ, Frerichs-Revermann L, Fuchs G, Alber BE: **The apparent malate synthase activity of *Rhodobacter sphaeroides* is due to two paralogous enzymes, (3S)-Malyl-coenzyme A**

(CoA)/{beta}-methylmalyl-CoA lyase and (3S)- Malyl-CoA thioesterase. *J Bacteriol* 2010, **192**(5):1249-1258.

21. Imam S, Noguera DR, Donohue TJ: **Global insights into energetic and metabolic networks in *Rhodobacter sphaeroides***. *BMC Syst Biol* 2013, **7**(1):89.
22. Imam S, Yilmaz S, Sohmen U, Gorzalski AS, Reed JL, Noguera DR, Donohue TJ: **iRsp1095: a genome-scale reconstruction of the *Rhodobacter sphaeroides* metabolic network**. *BMC Syst Biol* 2011, **5**:116.
23. Dufour YS, Kiley PJ, Donohue TJ: **Reconstruction of the core and extended regulons of global transcription factors**. *PLoS Genet* 2010, **6**(7):e1001027.
24. Eraso JM, Kaplan S: ***prrA*, a putative response regulator involved in oxygen regulation of photosynthesis gene expression in *Rhodobacter sphaeroides***. *J Bacteriol* 1994, **176**(1):32-43.
25. Eraso JM, Roh JH, Zeng X, Callister SJ, Lipton MS, Kaplan S: **Role of the global transcriptional regulator PrrA in *Rhodobacter sphaeroides* 2.4.1: combined transcriptome and proteome analysis**. *J Bacteriol* 2008, **190**(14):4831-4848.
26. Gomelsky M, Kaplan S: ***appA*, a novel gene encoding a trans-acting factor involved in the regulation of photosynthesis gene expression in *Rhodobacter sphaeroides* 2.4.1**. *J Bacteriol* 1995, **177**(16):4609-4618.
27. Gomelsky M, Kaplan S: **AppA, a redox regulator of photosystem formation in *Rhodobacter sphaeroides* 2.4.1, is a flavoprotein. Identification of a novel fad binding domain**. *J Biol Chem* 1998, **273**(52):35319-35325.
28. Gomelsky M, Kaplan S: **Genetic evidence that PpsR from *Rhodobacter sphaeroides* 2.4.1 functions as a repressor of *puc* and *bchF* expression**. *J Bacteriol* 1995, **177**(6):1634-1637.
29. Zeilstra-Ryalls JH, Kaplan S: **Role of the *fnrL* gene in photosystem gene expression and photosynthetic growth of *Rhodobacter sphaeroides* 2.4.1**. *J Bacteriol* 1998, **180**(6):1496-1503.
30. Shlomi T, Berkman O, Ruppin E: **Regulatory on/off minimization of metabolic flux changes after genetic perturbations**. *Proc Natl Acad Sci U S A* 2005, **102**(21):7695-7700.
31. Segre D, Vitkup D, Church GM: **Analysis of optimality in natural and perturbed metabolic networks**. *Proc Natl Acad Sci U S A* 2002, **99**(23):15112-15117.
32. Kim J, Reed JL: **OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains**. *BMC Syst Biol* 2010, **4**:53.
33. Pharkya P, Burgard AP, Maranas CD: **OptStrain: a computational framework for redesign of microbial production systems**. *Genome Res* 2004, **14**(11):2367-2376.
34. Burgard AP, Pharkya P, Maranas CD: **Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization**. *Biotechnol Bioeng* 2003, **84**(6):647-657.