Model-based Analysis Methods in Statistical Genomics

by

Qiuling He

A submitted in partial fulfillment of the

requirements for the degree of

Doctor of philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN – MADISON

2012

Date of final oral examination: 09/21/2012

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

Michael A. Newton, Professor, Statistics

Mark Craven, Professor, Biostatistics & Medical Informatics

Bret Hanlon, Assistant Professor, Statistics

Christina Kendziorski, Associate Professor, Biostatistics & Medical Informatics

Sijian Wang, Assistant Professor, Statistics

Abstract

This thesis aims to solve two problems in statistical genomics: (1) how to model agreement among genome-wide RNA interference (RNAi) studies; and (2) how to integrate experimentally derived genomic data with functional annotations. The problems are distinct in their specific elements but share two important features: (1) solutions could have significant implications for the practice of statistical genomics, and (2) our approaches to solve them use common model-based tools and techniques.

The RNAi analysis concerns four recent genome-wide studies of influenza virus replication. All studies identified genes whose inactivation alters a cell's ability to produce virus, and they all had a similar experimental design. In total 614 human genes were confirmed to have an affect on viral replication, however there were very limited agreement between the studies. For instance, only one gene was confirmed by all four studies.

Under the guidance of Professor Michael A. Newton.

The apparent lack of agreement raises questions about the rate of false positives and false negatives in genome-wide RNAi. We develop a generative sampling model to describe the RNAi data, and with likelihood methods we use this model to assess the relative magnitude of false positive and false negative effects. The model accommodates many aspects of RNAi, but it is sufficiently simple that closed form inference summaries are available. Evidence points to a relatively high false negative rate.

In the second part of the thesis (Chapter 3), we investigate the problem of genomic data integration, specifically, the problem of integrating experimentally derived data with data on the known functional profiles of the annotated genes. Such functional category analysis is important to data reduction and for weak-signal identification, though state-of-the-art methodology does not adequately handle the complexity of growing systems of functional categories. We show that a leading model-based empirical Bayesian approach suffers inconsistency and inefficiency, and we propose a new approach to connect these problems.

To my parents and family

Acknowledgments

Here I would like to express my gratitude to the people who have helped me make this thesis possible. I am grateful to my advisor Professor Michael Newton for his guidance and support throughout the five years of my graduate study. He has always been a role model and and an instrumental piece to my path in becoming a successful statistician in the future. I would like to sincerely thank my committee members Profs. Sijian Wang, Bret Hanlon, Christina Kendziorski and Mark Craven for generously sharing their ideas and advice, and encouraging me to complete my thesis research.

A special thanks goes to my collaborators Dr. Linhui Hao and Prof. Paul Ahlquist who gave me opportunities to participate in their interesting and significant research projects. I am also grateful to my funding source during the past two years, Morgridge Institute for Research and Computation & Informatics in Biology and Medicine (CIBM) at UW-Madison. This cross-disciplinary experience inspired me to continue working in the field of biomedical research.

I am indebted to all my dear peer students and friends Ning Leng, Darlene Lu, Meng Song, Yi Chai, Drs. Yang Zhao, Xu He, Yuan Jiang, Wei Zheng and Zhenyu Liu, and

many others I can not include here, without whose help and support I would not have been here today.

Last but not least I want to thank my parents and my family for their unconditional love and belief in me. I would like to express my greatest gratitude by dedicating this thesis to them.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis aims to solve two problems in statistical genomics, one a specific inference task in the analysis of RNA interference (RNAi), and a second, more general inference problem that arises when integrating experimentally derived data with functional annotations. Although the problems are distinct, they are linked through shared aspects of the statistical modeling approaches that are developed to solve them. It is anticipated that the solutions to both will have relevance beyond the specific case studies explored in this thesis.

The RNAi problem emerged from a collaborative project with Drs. P. Ahlquist, L. Hao, M. Craven, and M. Newton to understand genes involved with influenza-virus replication. The results of this collaborative effort are in a manuscript in preparation (Hao et al. (2012)). Chapter 2 of this thesis fully develops one aspect of that project concerning a model for agreements among replicated genome-wide RNAi studies. There remains relatively little work on statistical analysis of genome-wide RNAi, though the technology represents a powerful approach to understand gene function. Many sources of variation affect the RNAi data and our effort was to estimate these sources in order

to assess the relative size of false-positive and false-negative errors.

The second problem considered a generic data analysis task in genomics, namely the integration of experimentally derived data with functional annotations on the measured genes. We use the term *functional category analysis* to refer to all such methods of analysis. Various useful methods have already been developed, but limitations still remain to be addressed. Chapter 3 introduces our model-based method called Simultaneous Functional Category Analysis (SFCA). It is developed to solve the central model called *the role model*, originally proposed by Bauer et al. (2010) and further developed in Newton et al. (2012). SFCA develops approximate solutions using MCMC methods but its operating characteristics show advantages in comparison to Bauer et al.'s method.

This chapter is organized to briefly introduce each problem and related methodologies. We start with background of genome-wide RNA interference in Section 1.1. A review on the existing methods of functional category analysis and the issues involved is given In Section 1.2. In the end of this chapter is a summary our major contributions. Model-based methods are proposed to deal with the two central problems respectively in Chapter 2 and Chapter 3.

1.1 Genome-wide RNA interference

RNA interference (RNAi) is a gene-specific silencing process directed by short double stranded RNAs or small interfering RNAs (siRNAs) that can knock down expression of

a selected gene by inducing messenger RNA (mRNA) degradation in a sequence-specific manner (Mohr et al. (2010)). This technique has been widely used to selectively and robustly induce suppression and inhibit expression of targeted genes. By applying this technique to large-scale screens, high-throughput RNAi analysis becomes a powerful approach to study gene functions that support or modulate any biological process of interest. Genome-wide RNAi analyses have been used to study many important biological processes, for example, identifying host genes that are important for replication of a certain virus, i.e. HIV and Influenza virus.

An emerging challenge with RNAi studies is the limited agreement in the lists of identified genes from studies of the same cellular phenotype. This limited agreement must be due to false positive factors, false negative factors or both. If it is primarily due to false positive factors, then the majority of findings from RNAi studies would be erroneous. On the other hand, if false negative factors are dominant, then either the phenotype is extremely complex or the genetic causes extremely difficult to measure. In fact, there is good evidence that both false positive and false negative factors contribute to variations among studies. What is less well characterized is the relative magnitude of these effects. It is an important task to figure out what causes the limited agreement for better understanding of RNAi screening and its results.

False positive and false negative factors arise for various reasons, primarily from technical and biological sources of variation that affect data generation. Major factors causing false positives are (1) off-target effects, in which an siRNA leads to silencing one

or more genes besides the targeted one, owing to incomplete sensitivity and specificity of siRNAs; and (2) false positive errors that are intrinsic to the complex phenotypic readouts used to measure the cells. Similarly, experimental issues contribute to false negatives: (1) the cells under study may have redundancies that limit the accessibility of certain functions to phenotypic manipulation by knocking down a single gene; (2) genes with undetectable expression or whose knock down results in cytotoxicity can not be interrogated with RNAi in certain cell type; and (3) inefficiencies in knocking down targeted genes can also generate false negatives.

To better understand RNAi screening and assess effects of false positive and false negative factors, useful statistical tools are needed to model RNAi data. By this work we aim to contribute to this relatively new research area.

1.2 Functional category analysis

We define functional category analysis as the integration of experimental genomic data with functional information. It is demanded in routine data analysis applications. Various forms of experimental data are suited to this data analysis, including lists of genes identified by some genome-wide assay or quantitative gene-level scores on gene expression or differential expression. For our purposes the exogenous functional information refers to all that has been recorded in relevant databases regarding biological properties of the genes under study. A substantial amount of functional content is recorded in

GO (The Gene Ontology Consortium, Ashburner et al. (2000)) and KEGG (The Kyoto Encyclopedia of Genes and Genomes, Ogata et al. (1999)), and other systems hold other important information. In this thesis we focus on functional content represented by sets: specifically, where *biological property* p = *set of genes having property* p. The phrase *functional category* refers to a set of genes having a specific biological property. Properties involving an ordered series of events lose something when forced into this system (*e.g.*, signaling or metabolic pathways), however substantial statistical challenges remain to be addressed even within this class of set-defined functional properties.

Sometimes the analytical purpose is description and data reduction. The data analyst is faced with interpreting experimental data from across the genome; this interpretation is facilitated by summarizing data at the level of functional categories. For example, a list of 100 lead genes may represent just a few known biological functions. In other applications the functional information boosts the signal-to-noise ratio. For example, the signal representing the difference of two experimental conditions at the gene level might be very subtle and hard to detect, but it gets boosted at the category level when contributions from genes in the category are consistent.

A number of useful statistical methods are available for integrating experimental and functional data (Goeman and Bühlmann (2007), provides an early review). The methods aim to either calibrate category-level differential expression or test the over-representation (i.e. enrichment) of categories in a short list of genes. Most enrichment methods adapt Fisher's exact test (e.g. Draghici and Krawetz (2003), Beißbarth and

Speed (2004), Grossmann et al. (2007), Jiang and Gentleman (2007), Newton et al. (2007)). Other methods treat the gene-level data in a functional category as a multivariate observation, and then aim to assess affects of various covariates on the associated joint distribution (e.g., Barry et al. (2005), Subramanian et al. (2005), Efron and Tibshirani (2007)). Methods differ in terms of how much experimental data needs to be incorporated, what hypotheses are being tested, and how inferences are computed.

Almost all available methods develop inference for individual functional categories, treating the multiplicity of categories as an afterthought. For instance GSEA, by Subramanian et al. (2005), computes set-level statistics for each category and calibrates afterwards to target a false discovery rate. Other methods are similar in that respect. One-at-a-time methods are inefficient at prioritizing categories because they fail to incorporate the complexity of the functional record, especially the large number of categories, their varying size, and their extensive overlap. Variation in category size causes a power imbalance across categories. Power is related to the size of both category and its effect. Ranking categories by p-values tends to favor large categories, while ranking by a set average statistics tends to favor small categories, since sample variation is higher in this case. Although size of a category is not important to the scientific relevance, it has an undue influence in the summary statistics. Overlap patterns in GO or KEGG also complicate inference. For one-at-a-time methods, overlapping categories have positively correlated summary statistics. One practical problem caused by the overlap is that a summary list of significant functional categories can be very large, thus not facilitating

a simple interpretation of functional content of experimental data.

GO is comprised of three directed acyclic graphs where nodes are categories and directed edges link parent nodes to child nodes representing proper subsets of the parent category. This hierarchical structure is used by some statistical methods (*e.g.* sequential testing methods, Liang and Nettleton (2010)). Sequential testing methods are often difficult to interpret; further, the directed graphs do not express all of the overlap relations among categories.

Statistical methods that address category differential expression assert that the category on test is non-null if and only if any of the genes in that category is non-null. (note on Barry's method III, which is not this but essentially so). The trouble with this assertion is that many genes are multi-functional, and this multi-functionality is expressed as category overlap in the functional record. A gene has different *roles* depending on what it is doing in the cell. For example, the NXF1 gene (nuclear RNA export factor 1) is annotated to 20 different GO categories including nucleocytoplasmic transporter activity and mRNA export from nucleus. When this gene is differentially expressed under different experimental conditions, it might be due to one of its roles but not necessarily all of them. Implicating all of NXF1's roles leads to spurious inference (false positives).

Model-based gene-set analysis methods are promising because they can be constructed to incorporate the functional record simultaneously over all categories and thereby can handle difficulties to do with category size and overlap (Lu et al. (2008), Bauer et al. (2010), Newton et al. (2012)). In Bauer *et al.* (2010), the generative model

has non-null behavior starting with the functional category rather than the genes; each gene inherits non-null behavior from the non-null categories to which it is annotated. This transformation means that the inference on a given category does not only depend on activities of genes inside this category but is also related to behaviors of other overlapping categories. The form of *role model* was originally proposed in Bauer *et al.* (2010) and inference computations were made available in R package MGSA (Bauer et al. (2011)). The work was investigated and developed further in Newton et al. (2012), and techniques from probabilistic graphical modeling were invoked to approximate inference computations. However, role-model inference computations via probabilistic graphical models remain complicated. Analysis of the structure of role mode computations deployed in MGSA shows a critical problem with inference, owing to a failure to respect specific constraints of the parameter space. It is our goal to develop statistical methods that overcome the limitations of existing procedures.

1.3 Summary of main contributions

Our contributions in the RNAi project include proposing a novel sampling model that deals with multi-record observational data from RNA interference studies, and developing sophisticated computational schemes for likelihood inference. The design and structure of RNAi experiments forces us to propose a specification for multi-study, two-stage

(detection/confirmation), genome-wide RNAi data. To deal with important experimental factors that lack observational data, hidden variables are introduced and carefully chosen statistical distributions are assigned to them. The challenge of computation comes from marginalizing these hidden variables in a non-linear model. We overcome this challenge and precisely measure the contribution of false positives and false negatives in Influenza RNAi.

For functional category analysis, we contribute to improve both modeling and computations of the role model. An important condition called the *activation hypothesis* is developed to establish equivalence between gene- and category-level activities. By clarifying intrinsic constraints among role model parameters, we more accurately infer categories that contribute to gene-level signals. For computations, we have considered three directions but focus on one most promising one in this thesis. We had investigated methods for exact computations and inferences via probabilistic graphical models. However, this remains challenging due to complexity of graphs for large-scale genome-wide problems. A second approach was to relax the role model by embedding it in a larger set of probability distributions, and then to fit the larger model and connect the results back to the original model. In preliminary work, we proposed two relaxations of the role model that have different generalized-linear-model (GLM) representations of gene-level data. Regularized regression and quadratic programming were developed to fit the relaxed models and provide for the selection of the most significant functional categories. The major difficulties in optimization of non-convex objective function and

efficiently accommodating activation hypothesis by linear constraints did not allow us to succeed in this direction. The third method, presented here in Chapter 3 applies MCMC methods to approximate posterior inferences. With activation hypothesis the chain is sampled on a highly restricted space that requires sophisticated updating rules. Altogether we have established superiority of a method for category analysis in terms of posterior consistency and efficiency compared to existing methods and also we have developed effective algorithm to implement posterior computations. Although unresolved, the first two approaches remain to be valuable future work.

Chapter 2

Sampling model for a meta analysis of genome-wide RNAi studies

2.1 Overview

We are interested in four recent RNAi studies that all aimed to identify host genes involved in influenza virus replication (Karlas et al. (2010), Hao et al. (2008), Brass et al. (2009), König et al. (2009)). They followed similar experimental procedures but showed limited overlap in their final gene findings. To provide reasoned inferences about factors affecting among-study gene-level agreement, we developed a statistical model and corresponding likelihood-based analysis methods. The model formulates relationships among: (1) system-level parameters that affect sensitivity and various error rates, (2) gene-level and study-level latent variables that transduce information about the system to information at the gene-level, and (3) gene-level, multi-study data on both detection and confirmation by RNAi screening. In its generative form, the model specifies the probability of observing any particular multi-study data set. In its inferential form, it

indicates the likelihood assigned to any particular parameter setting in light of observed data. In general, this work involves model development, mathematical analysis, likelihood and Bayesian computation inference and diagnostic checking. Our sampling model sufficiently fits the data and the results suggest that false negative factors contribute more to the limited agreement than false positive factors. This chapter is organized to introduce in turn the sampling model, computational methods, inferences, model checks and an extended application.

2.2 Background

RNAi is a gene-specific silencing process that knocks down expression level of targeted genes using dsRNA or siRNA. Genome-wide RNAi has become a powerful tool to study gene functions in regulating biological processes. In our meta analysis, four studies applied this technique to identify genes whose inactivation affects the cell's ability to reproduce Influenza virus.

Basic information of four studies is presented in Table 1, including RNAi libraries and cell lines. Despite differences in detailed deployment, all four studies used a similar two-stage experimental design. All studies started with a high-throughput *primary screen* to target each gene across the whole genome small interference RNAs (siRNA) from an RNAi library and candidate genes are selected as detected genes by the primary

screen.These genes then were re-tested for function in virus replication in repeated secondary validation assays with individual siRNAs. We call genes that are selected in the end confirmed by the secondary screen. Numbers of genes detected and confirmed from primary and secondary screens are collected from the 4 studies and listed in the last two rows of Table 1. They serve as data of our sampling model.

There are in total 984 genes detected by the 4 RNAi studies jointly, and 641 of these detected genes are further confirmed. Figure 1 gives a basic summary of these confirmed genes in terms of their distribution. The bar plot in panel (A) is divided into 4 sections by solid lines. That four high vertical bars are located in the first section means that most of genes are confirmed by only one study but not the rest. There is only one gene confirmed by all 4 studies. Panel (B) gives the number of overlapping genes confirmed by any pair of studies and percentages relative to the total number of genes confirmed by each study. These evidence all suggest that the four RNAi studies of interest present limited agreement on their confirmed gene lists. It motivates us to develop a statistical approach to model RNAi data and explain the low agreement.

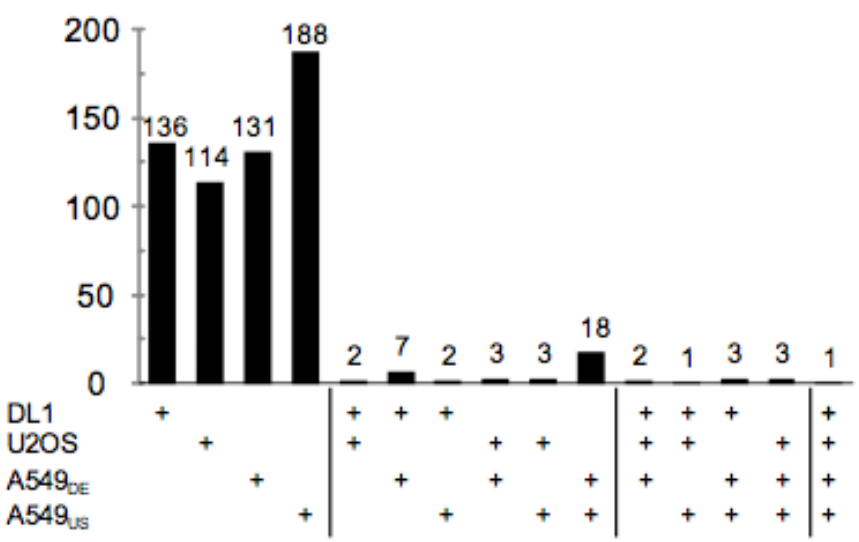
2.3 Modeling approach

The assessment of agreements and disagreements among studies has long been a focus of model-based statistical analysis, from seminal work by R.A. Fisher and colleagues on species abundance estimation in ecology (Fisher et al. (1943)) to more recent and relevant

Table 1: Basic information on the 4 RNAi studies of interest. Acronyms are assigned by cell lines they used.

Studies	DL-1	U2OS	A549DE	A549US
Reference	Hao et al. (2008)	Brass et al. (2009)	Karlas et al. (2010)	König et al. (2009)
Cell line	Drosophila cell line	Human osteosarcoma cell line	Human lung adenocarcinoma epithelial cell line	Human lung adenocarcinoma epithelial cell line
siRNA Library	Ambion	Dharmacon siARRAY	Qiagen Hu genome 1.0	Qiagen druggable set V2
Gene targets	7096 Drosophila genes ≈ 15,000 human homologs	17,877 human genes	22,843 human genes	19,628 human genes
# Detected in primary screen	237	250	287	294
# Confirmed in secondary screen	154	129	168	219

A



B

	DL-1	U2OS	A549DE	A549US
DL-1	96	6 (6.3%)	13 (13.5%)	7 (7.3%)
U2OS	6 (4.7%)	129	9 (7.0%)	8 (6.2%)
A549DE	13 (7.7%)	9 (5.4%)	168	25 (14.9%)
A549US	7 (3.2%)	8 (3.7%)	25 (11.4%)	219

Figure 1: Basic summary on 641 confirmed genes. (A) Distribution of genes confirmed by the four RNAi studies designated as DL-1, U2OS, A549US and A549DE. Vertical bars show numbers of genes confirmed by studies indicated by +. Note that most genes were confirmed by only one study. (B) Pairwise overlap of confirmed genes. Number of genes implicated by each study are on the diagonal line. In flanking cells are pairwise overlaps between pairs of indicated studies. A percentage is calculated against the number of genes confirmed in the study of the relevant column.

precursors to our own calculations, including Raftery (1988), Craig et al. (1997), and Basu and Ebrahimi (2001). The rationale for this general approach is that the specific findings of any study are affected by numerous factors, some of which are systematic and shared in some predictable way among studies, and some of which are idiosyncratic. To capture the systematic effects we treat them as parameters in a stochastic process presumed to have generated the observed data, and we infer the parameter values by calculating the probability of observed data (the likelihood). The design and structure of RNAi experiments forces us to go beyond previously described probability models and propose a specification for multi-study, two-stage (detection/confirmation), genome-wide RNAi data. In this section, we will in turn introduce data, latent variables, parameter, the likelihood model and how to analytically develop likelihood calculation.

2.3.1 Data

For each study s in the set of four studies, and each gene g in the human genome, we introduce $D_{g,s}$ to indicate whether or not (1 or 0) gene g was detected in the primary screen of study s, and similarly $C_{g,s}$ to indicate whether or not g was confirmed in the corresponding secondary screen.

$$\begin{aligned} D_{g,s} &= 1\,[\text{gene } g \text{ is detected by the primary screen of study } s\,], \\ C_{g,s} &= 1\,[\text{gene } g \text{ is confirmed by the secondary screen of study } s\,] \end{aligned}$$

for genes $g = 1, 2, ..., G$ and studies $s = 1, 2, 3, 4$. We consider the genome to be the union of genes that are covered by the siRNA libraries used in the four studies, and assume the genome size $G = 22000$. According to the experimental design, we observe detection indicators $\{D_{g,s}\}$ for all g and s and confirmation indicators $\{C_{g,s}\}$ only for cases g and s where $D_{g,s} = 1$. That is, the primary screen of each study is viewed as scanning the full genome; the secondary screen aims to confirm those primary findings. It is technically convenient to allow $C_{g,s}$ to be defined even when $D_{g,s} = 0$, though by the study designs such $C_{g,s}$ is unobserved and does not enter our computations. The four studies differ in details of their secondary screens. To simplify our analysis we model studies similarly, in terms of results $C_{g,s,k}$ from further assays $k = 1, 2, 3, 4$, wherein the kth assay entails the application of the kth siRNA from the pool of (typically) 4 siRNAs that targets gene g. Then we have confirmation $C_{g,s} = 1$ if (and only if) at least two of these assays indicates a phenotype; i.e. if $\sum_k C_{g,s,k} \geq 2$. Several studies (e.g., U2OS) have this precise structure, although we have not used any assay-level data $C_{g,k,s}$ in subsequent computations (these data are not complete; we use the summary calls $C_{g,s}$). Not all studies have this structure (e.g., DL-1); the key consequence of modeling this way is that the secondary validation is more stringent for filtering non-involved genes.

There are three observable states of $(D_{g,s}, C_{g,s})$ for a given gene g in a given study s:

$$\{(0, 0), (1, 0), (1, 1)\}.$$

The proposed model entails gene-specific latent random effects, and thus marginally

the $(D_{g,s}, C_{g,s})$ is not independent from $(D_{g,s'}, C_{g,s'})$ for any two studies s and s'. Our model does entail independence among genes, and therefore the likelihood (probability of observed data) can be expressed as the probability of the multinomial count vector $\{N_\pi\}$ over the $3^4 = 81$ possible multi-study observation states, or *patterns*, $\{\pi\}$ (see Table 3), where

$$N_\pi = \sum_{g=1}^{G} 1\left[\{(D_{g,s}, C_{g,s})\} \text{ has pattern } \pi\right].$$

(On the independence among genes assumption, this is conditional on involvement (see below) and expresses the fact that separate cells and assays are used for different genes within a given study.) In these terms, the log-likelihood is

$$\mathcal{L} = \log \text{Prob(data)} = \sum_{\pi} N_\pi \log P_\pi, \tag{2.1}$$

where pattern probabilities $\{P_\pi\}$ are defined by a smaller number of parameters through a stochastic model of genome-wide RNAi.

2.3.2 Latent variables and parameters

To calculate pattern probabilities, various experimental factors that contribute to false positives and false negatives need to be modeled including (1) involvement, (2) accessibility, (3) off target, (4) knockdown efficiency, (5) measurement errors including both false positive and false negative errors. The first three factors are specified using latent random effects, and the last two are calibrated with system-level parameters.

Involvement

Whether or not the gene g is truly involved in influenza-virus replication is unknown a priori, and this fact is expressed by the latent binary variable I_g.

$$I_g = 1[g \text{ is } \textit{involved} \text{ in influenza virus replication}]$$

In some cell type, an error-free measurement of a true knockdown, in the absence of off-target effects, would show a phenotype if and only if $I_g = 1$. Parameter θ is used as the genome-wide rate of true involvement. Fixing the genome size at G, the number of truly involved genes is $N = \sum_{g=1}^{G} I_g$, which has expected value $G\theta$. The distribution of gene-level data depends on I_g through additional factors expressing sources of variation that affect knockdown and phenotype. One could alternatively classify the $\{I_g\}$ as a high-dimensional parameter, but in doing Bayesian inference we would immediately cover it with a prior, so we treat it as a vector of latent factors in the notation.

Accessibility

A variety of factors could block either the knockdown of a gene or the phenotype of a knocked-down gene, for example, gene may not be expressed in the particular cell line, siRNAs may induce cytotoxicity, or due to phenotypical masking. We introduce latent, binary accessibility variables $A_{g,s}$ to accommodate this general effect that contributes to false negatives,

$$A_{g,s} = 1[g \text{ is } \textit{accessible} \text{ in study } s\ ].$$

If $A_{g,s} = 1$ we say that gene g is accessible to study s. if g is also involved and fully knocked down, a phenotypic effect would show. In the absence of more specific knowledge we treat the $A_{g,s}$s as independent Bernoulli-distributed variables. Analysis supports allowing the accessibility rate to vary among studies, and we allow this flexibility to better accommodate study-study heterogeneity.

Off targets

The pool of siRNAs that target gene g in study s may not be fully specific, and thus may inadvertently knock down some number of influenza-involved off targets. These off targets are a subset of the involved off-targets associated with all siRNAs used for gene g across all studies, not accounting for inaccessible genes in any given study. By modeling $T_{g,s}$ as a subset of a total T_g, we allow potential dependencies between studies attributable to use of the same siRNA in different studies. Further in k^{th} individual assay of the secondary screen, $V_{g,s,k}$ is used as a subset of $T_{g,s}$. The three layers of latent variables used to model numbers of involved off targets are:

T_g = number of involved off-*targets* for gene g, relative to a pool of siRNAs that might be used to target gene g

$T_{g,s}$ = size of the accessible subset of T_g in study s

$V_{g,s,k}$ = size of the accessible subset of $T_{g,s}$ in assay k of secondary screen in study s .

The number T_g counts involved off-targets from all siRNAs in play for a given gene:

we consider it to have mean value $K\theta\nu$, where K is the average number of distinct siRNAs used per gene across all four studies, θ is the involvement rate, and ν is the mean number of off-targets per siRNA. Evidence indicates that rates of phenotypic response increase with T_g, but there remain little data on the distribution of it beyond computational predictions based on sequence homology (Kulkarni et al. (2006)). From first principles, we treat T_g as Poisson distributed though we investigate over-dispersed alternatives in model diagnostics (Section 2.5). In study s, four (typically) siRNAs are used and these carry a subset of $T_{g,s}$ involved and accessible off-targets, having a Binomial distribution on t trials with success probability $4\gamma_s/K$ given that $T_g = t$. (An involved off-target that is not accessible in a given study cannot affect the phenotype in that study.) Similarly, with given $T_{g,s} = u$, $V_{g,s,k}$ is simplified to be a Binomial distribution with size u and success probability 1/4. Numbers of off targets from the 4 individual assays are actually not independent but rather negatively correlated as their sum $\sum_{k=1}^{4} V_{g,s,k}$ equals $T_{g,s}$. Sensitivity analysis showed that ignoring this negative correlation did not affect likelihood computaions.

mRNA Knockdown

We developed a model to have the following three basic features. First, the larger the number of either on-target or off-target events, the higher the probability of a phenotypic effect. Second, if there are multiple off-target events from a pool of siRNAs, then distinct off-targeted genes are affected. Third, we suppose that multiple on-target hits

(i.e., from multiple siRNAs targeting the same gene) deliver a higher probability of phenotypic effect than do the same number of off-target hits. A mathematical device to achieve this structure imagines that every targeting or off-targeting event (i.e. every potential knock down of an involved gene) is associated with a uniform (0,1) random variable representing the fraction of mRNA remaining after knock down by that event. An error-free measurement then would show a phenotypic effect if any of the involved, accessible genes had mRNA levels reduced below a threshold, parameterized by $\omega \in (0,1)$. By assumption, off-target effects work in parallel. If $T_{g,s} = t$, the probability that any of the off-targeted genes has mRNA knocked down below ω is $1-(1-\omega)^t$. The assumptions similarly form the on-target model as a series circuit: the probability that the targeted mRNA is knocked down below ω after hits from a pool of 4 siRNAs becomes $1 - G_{4,1}(-\log(\omega))$, where $G_{4,1}$ is the cumulative distribution function of a gamma distribution with shape 4. (Details provided in Section 2.3.3).

Measurement errors

Our meta-analysis analyzes summary gene-level data from four two-stage genome-wide studies. Whether or not a gene is detected or confirmed in any study depends on details of the quantitative assays used to assess the phenotypic effect, as well as on all the intrinsic factors indicated above. These assays are subject to various sources of measurement error that may create both false negative and false positive recordings. We allow both types, and have found improved model fits by allowing the false negative

rate to be study specific. Parameter α is for type I (false positives) and $\{\beta_j\}_{j=1}^4$ are for type II (false negatives).

Here we summarize the system-level parameters that are used to specify the probability structure of latent variables and observed data; they describe the basic system in terms of rates governing the latent variables as well as quantities affecting false-positive and false-negative detections and confirmations:

$$
\begin{aligned}
\theta &= \text{proportion of genome involved in influenza virus replication} \\
\alpha &= \text{false positive measurement error} \\
\beta_s &= \text{false negative measurement error of study } s \\
\gamma_s &= \text{rate at which genes are accessible in study } s \\
\omega &= \text{expression threshold in knockdown model} \\
\nu &= \text{average number of off-target genes per siRNA .}
\end{aligned}
$$

2.3.3 Model specification

Based on previous descriptions, probability distributions are assigned to the latent variables.

$$\begin{aligned} I_g &\sim \text{Bernoulli}(\theta) \\ A_{g,s} &\sim \text{Bernoulli}(\gamma_s) \\ T_g &\sim \text{Poisson}(K\theta\nu) \\ T_{g,s}\,|[T_g = t] &\sim \text{Binomial}\left(t, \frac{4\gamma_s}{K}\right) \\ V_{g,s,k}\,|[T_{g,s} = u] &\sim \text{Binomial}\left(u, \frac{1}{4}\right). \end{aligned} \tag{2.2}$$

Finally, we have a model for observations $D_{g,s}$ and $C_{g,s}$.

$$\begin{aligned} {}_{D_{g,s}|\,[I_g = i, A_{g,s} = a, T_{g,s} = t]} &\sim \text{Bernoulli}\left[1 - \beta_s + (\alpha + \beta_s - 1)\left[G_{4,1}(-\log\omega)\right]^{ai}(1-\omega)^t\right] \\ {}_{C_{g,s,k}|\,[I_g = i, A_{g,s} = a, V_{g,s,k} = v]} &\sim \text{Bernoulli}\left[1 - \beta_s + (\alpha + \beta_s - 1)(1-\omega)^{ai+v}\right] \end{aligned} \tag{2.3}$$

where $G_{4,1}(.)$ is the c.d.f. of a gamma distribution with shape parameter 4 and scale parameter 1. The model allows heterogeneity across genes and studies. Targeted genes that are involved ($i = 1$) need to be accessible ($a = 1$), otherwise they are detected at the lower rate of non-involved genes. The constant 4 enters here because we have modeled a typical study that targets a gene by pooling four different siRNAs (with each additional siRNA improving the detection rate).

Figure 2 presents a probability model for detection $D_{g,s}$ and confirmation $C_{g,s}$ conditional upon accessibility, involvement, and off-target count. For detections, each edge in the circuit has a probability, and the fate of cells considered prior to experimentation

(left) is a path through the circuit to some end state (right). For example, a phenotypic effect (+) is possible if either (1) there is a successful knockdown of some involved gene (either on or off target) and there is no (type II) measurement error, or (2) there is neither on nor off target knockdown and there is a (type I) measurement error. Note that confirmations are modeled similarly to detections, but we consider a typical study in which the four individual siRNAs that had been pooled in the primary screen were applied separately in four assays. Confirmation on assay k is indicated by $C_{g,s,k}$, and we have $C_{g,s} = 1$ if and only if $\sum_{k=1}^{4} C_{g,s,k} \geq 2$; that is, if at least two of the single siRNA assays also yielded a positive phenotype. Figure 2(B) breaks down contributions to the conditional distribution of $C_{g,s,k}$. Detection probability and conditional confirmation rate are assembled by adding along paths in this circuit.

Another system-level parameter we fix a priori and do not estimate from the data is

$$K = \text{average number of siRNAs that target a gene.}$$

Ideally if all 4 studies use the same siRNA library, we would expect $K = 4$ i.e. there are exactly 4 siRNAs targeting every gene as designed; if each study uses a distinct library and there is no chance of genes being targeted by siRNAs other than the designed ones, then $K = 16$. K controls the correlation in numbers of off-targets between studies, i.e. $\text{cor}(\text{T}_{\text{g,s}}, \text{T}_{\text{g,s}'}) \to 0$, as $K \to \infty$. In our case, we fix $K = 12$ since the 4 studies of interest use 3 different libraries. Diagnostic computations showed little sensitivity to this setting.

A full specification of conditional independence assumptions is in Figure 3. The

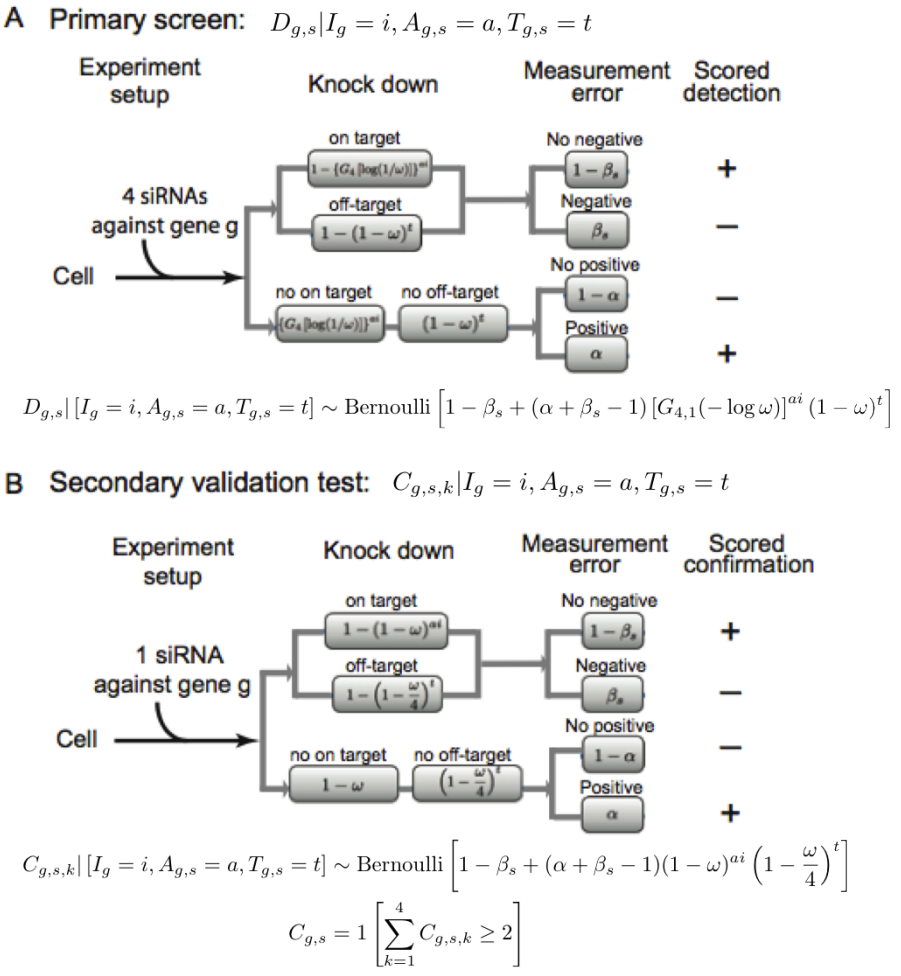


Figure 2: Circuit diagram providing a schematic for gene-specific outcomes and probabilities for (A) primary screen and (B) secondary screen. (A) Cells are treated with a pool of 4 siRNA's and can traverse one of two main branches during an experiment: the top branch involves some kind of knock down event, either on or off target (or both), and the bottom branch involves neither on nor off target knock down. In either case, measurement error could affect the final measured phenotype. An effect is either observed (+), or unobserved (-). Along each path through the circuit are shown probabilities associated with cells traversing that path. Similarly, (B) illustrates one of the four individual assays where cells are treated with only 1 siRNA.

various modeling elements have been introduced to address known features of genome-wide siRNA screening. For example, every additional siRNA applied to an involved gene increases the chance that the harboring cells exhibit a phenotype. The higher the rate of involved genes, the higher the rate of an off-target phenotype. There is heterogeneity among genes, owing to whether or not they are involved, and owing to varying amounts of off-targets associated with their targeting pools of siRNAs, but there is among-gene independence in terms of siRNA detection/confirmation. The studies are heterogeneous, because they may entail different sets of accessible genes and these accessibility rates (γ_s) are study specific, and also there may be different false negative measurement errors (β_s) involved in each study due to individual experimental environment. (We had considered a single parameter γ and β, but saw substantial improvements when we allow the extra flexibility.) From study to study the data are not independent, owing to genetic factor and common targets among studies in RNAi libraries (i.e. I_g and T_g, which get marginalized in our likelihood computation).

To further explain the model structure, the curious $G_{4,1}$ term and related terms enter because of our knockdown model. We suppose that each targeting or off-targeting event is associated with a uniformly distributed variable on $(0, 1)$. For on-targets, we suppose that four hits reduce the expression of the target such that the amount left over equals the product of the four uniforms, and if this amount is less than ω, we would see an effect, in the absence of measurement error. This happens with probability $G_{4,1}(-\log\omega)$. We further assume that off-target events hit separate genes, and act in a parallel fashion,

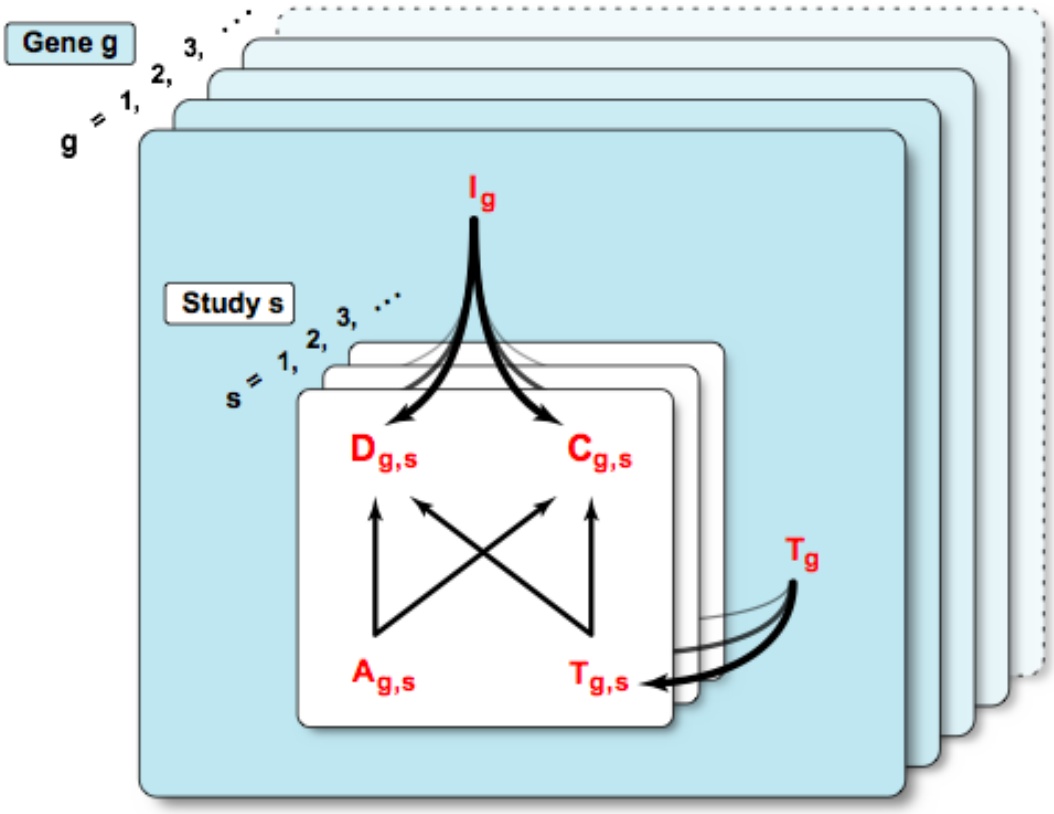


Figure 3: Plate diagram for dependence structure of the sampling model. Latent factors I_g, $A_{g,s}$ and $T_{g,s}$ affect the distribution of observable detections $D_{g,s}$ and confirmations $C_{g,s}$. For example, the probability that a gene is detected in a given study depends on whether it is truly involved in flu, whether it is accessible in the system used, and the number of involved and accessible off-targets (3 arrows coming towards $D_{g,s}$.) Conveniently, we assume that $C_{g,s}$ exists independently of detection, and is latent unless $D_{g,s} = 1$.

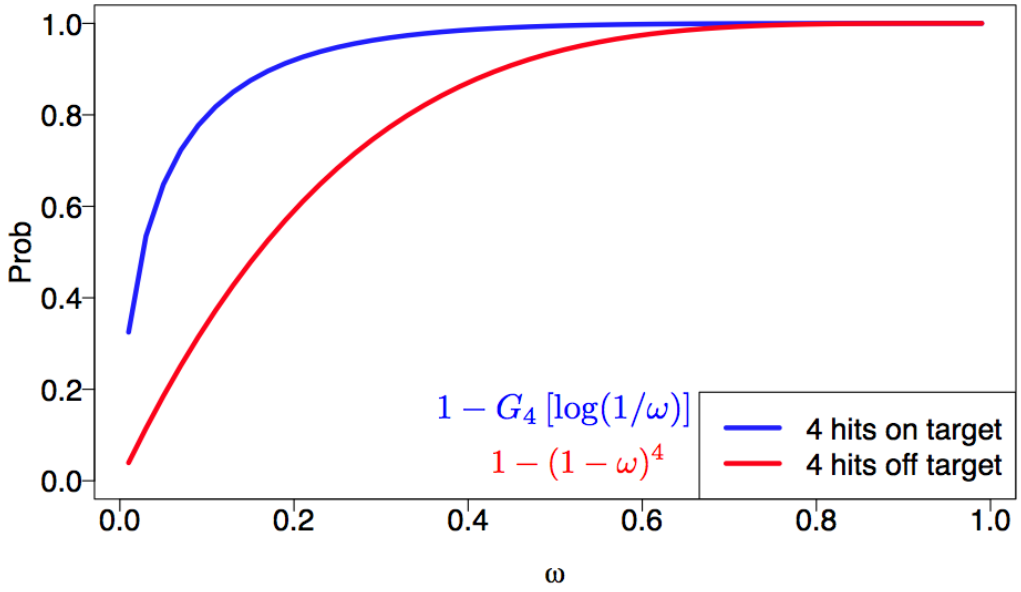


Figure 4: Knockdown model properties. Four on target hits are more effective at producing an observed effect than four off-target hits. Shown is the case of no measurement error, as a function of the threshold for an effect.

so that a phenotypic effect occurs if any of the mRNA levels is reduced below ω; this happens with probability $1-(1-\omega)^t$ when there are t off-targeting events. By modeling this way we allow that multiple on-target hits are more effective than the same number of dispersed off-target hits (Figure 4).

The log-likelihood $\mathcal{L}$ in (2.1) is a function of these 12 parameters, which we collect in a vector $\psi = (\theta, \alpha, \beta, \gamma, \omega, \nu)$, where $\beta = \{\beta_s\}_{s=1}^4$, $\gamma = \{\gamma_s\}_{s=1}^4$. Thus $\mathcal{L} = \mathcal{L}(\psi)$.

2.3.4 Calculation of pattern probabilities

The 81 multi-study pattern probabilities $\{P_\pi\}$ (and thus the log-likelihood $\mathcal{L}(\psi)$) in (2.1) are obtained as a function of the 12 system-level parameters ψ by summing out the discrete-valued latent variables. Considering among-gene independence, we focus on a single gene, and sum out values of the involvement indicator I_g, the four accessibility indicators $A_{g,s}$, and the off-target counts T_g, the four $T_{g,s}$, and the $\{V_{g,s,k}\}_{k=1}^4$ for each study. (We model $T_{g,s}$'s as subsets of a common T_g to reflect the possibility that different studies share siRNAs.) All but the target counts are binary sums; more complicated is the elimination of the off-target counts. To investigate this calculation, write the vector $a = \{a_s\}$ and the conditional probability of data pattern π as,

$$P_\pi(i, a) = P\left(\pi \middle| I_g = i, \{A_{g,s}\}_{s=1}^4 = a\right).$$

Each multi-study pattern probability P_π is computed as a summation of these $P_\pi(i, a)$ over the 2^5 values of its arguments. The trickier computation is the evaluation of each $P_\pi(i, a)$, which requires marginalization of the off-target counts.

To marginalize the off-target counts, first recognize that each pattern π is an intersection of four study-specific patterns $\pi = \bigcap_s \pi_s$. For example $\pi = 3111$ indicates that the gene is confirmed and detected in the first study and neither detected nor confirmed

in any of the remaining three studies. The modeling assumptions give

$$\begin{aligned} P_\pi(i,a) &= \sum_{t=0}^{\infty} P(T_g = t)\ P\left(\pi | I_g = i, \{A_{g,s}\}_{s=1}^4 = a, T_g = t\right) \\ &= \sum_{t=0}^{\infty} \text{Pois}(t)\ \prod_{s=1}^{4} P\left(\pi_s | I_g = i, A_{g,s} = a_s, T_g = t\right) \\ &= \sum_{t=0}^{\infty} \text{Pois}(t)\ \prod_{s=1}^{4} \sum_{u=0}^{t} B_s(t,u)\ P\left(\pi_s | I_g = i, A_{g,s} = a_s, T_{g,s} = u\right) \\ &= \sum_{t=0}^{\infty} \text{Pois}(t)\ \prod_{s=1}^{4} \sum_{u=0}^{t} B_s(t,u)\ Q_{s,i,a_s,u} \end{aligned} \tag{2.4}$$

where $\text{Pois}(t) = P(T_g = t) = \exp\{-K\theta\nu\}(K\theta\nu)^t/t!$ by the Poisson assumption, $B_s(t,u)$ is the Binomial mass function at u with t trials and success probability $\frac{4\gamma_s}{K}$, and where each contribution $Q_{s,i,a_s,u} = P\left(\pi_s | I_g = i, A_{g,s} = a_s, T_{g,s} = u\right)$ is computed from the stochastic model (2.3). Coming back to pattern $\pi = 3111$ for example, the four sub-pattern probabilities are:

$$\begin{aligned} Q_{1,i,a_1,u} &= P\left(D_{g,1} = 1 | I_g = i, A_{g,1} = a_1, T_{g,1} = u\right)\ P\left(C_{g,1} = 1 | I_g = i, A_{g,1} = a_1, T_{g,1} = u\right) \\ Q_{2,i,a_2,u} &= P\left(D_{g,2} = 0 | I_g = i, A_{g,2} = a_2, T_{g,2} = u\right)\ P\left(C_{g,2} = 0 | I_g = i, A_{g,2} = a_2, T_{g,2} = u\right) \\ Q_{3,i,a_3,u} &= P\left(D_{g,3} = 0 | I_g = i, A_{g,3} = a_3, T_{g,3} = u\right)\ P\left(C_{g,3} = 0 | I_g = i, A_{g,3} = a_3, T_{g,3} = u\right) \\ Q_{4,i,a_4,u} &= P\left(D_{g,4} = 0 | I_g = i, A_{g,4} = a_4, T_{g,4} = u\right)\ P\left(C_{g,4} = 0 | I_g = i, A_{g,4} = a_4, T_{g,4} = u\right). \end{aligned}$$

where

$$P\left(C_{g,s} = 1 | I_g = i, A_{g,s} = a_s, T_{g,s} = u\right) = P\left(\sum_{k=1}^{4} C_{g,s,k} \geq 2 \,\middle|\, I_g = i, A_{g,s} = a_s, T_{g,s} = u\right)$$

We make the simplifying approximation that $C_{g,s,k}$ are conditionally independent (and thus $C_{g,s}$ is governed by Binomial masses), though in fact they have some negative

dependence attributable to the divvying up the off-target count $T_{g,s}$ among the four separate assays.

A key to simplifying the computation further is to recognize that with respect to the count variable u, each $P\left(D_{g,s}|I_g, A_{g,s}, T_{g,s}\right)$ is a polynomial of $\xi_1 = 1-\omega$, and each $P\left(C_{g,s}|I_g, A_{g,s}, T_{g,s}\right)$ is a polynomial of $\xi_2 = 1-\frac{\omega}{4}$. Thus, each $Q_{s,i,a_s,u}$ is a bivariate polynomial in ξ_1 and ξ_2, of degree at most u and $4u$ respectively. By careful book-keeping, we identify coefficients $\{b_{s,p,q}\}$ (depending on system parameters ψ and the pattern π) such that

$$Q_{s,i,a_s,u} = \sum_{p=0}^{1}\sum_{q=0}^{4} b_{s,p,q}\ \left(\xi_1^p\,\xi_2^q\right)^u$$

Thus the inner factor of (2.4)

$$\begin{aligned}
\sum_{u=0}^{t} B_s(t,u)\, Q_{s,i,a_s,u} &= \sum_{u=0}^{t} B_s(t,u)\ \sum_{j=0}^{8} b_{s,j}\,\xi^{uj} \\
&= \sum_{p=0}^{1}\sum_{q=0}^{4} b_{s,p,q} \sum_{u=0}^{t} \left(\xi_1^p\,\xi_2^q\right)^u B_s(t,u) \\
&= \sum_{p=0}^{1}\sum_{q=0}^{4} b_{s,p,q} \left(1 - \frac{4\gamma_s}{K} + \frac{4\gamma_s}{K}\xi_1^p\,\xi_2^q\right)^t \\
&= \sum_{p=0}^{1}\sum_{q=0}^{4} b_{s,p,q} {e_{s,p,q}}^t,
\end{aligned}$$

with the second-last line obtained from the moment generating function of a Binomial variable, and with $e_{s,p,q} = 1-\frac{4\gamma_s}{K}+\frac{4\gamma_s}{K}\xi_1^p\,\xi_2^q$. Incorporating this back into (2.4), we obtain

for the conditional probability of a pattern given accessibility and involvement:

$$\begin{aligned}
P_\pi(i,a) &= \sum_{t=0}^{\infty} \text{Pois}(t) \prod_{s=1}^{4} \sum_{p=0}^{1} \sum_{q=0}^{4} b_{s,p,q} {e_{s,p,q}}^t \\
&= \sum_{t=0}^{\infty} \text{Pois}(t) \sum_{p_1=0}^{1} \sum_{p_2=0}^{1} \sum_{p_3=0}^{1} \sum_{p_4=0}^{1} \sum_{q_1=0}^{4} \sum_{q_2=0}^{4} \sum_{q_3=0}^{4} \sum_{q_4=0}^{4} \left(\prod_{s=1}^{4} b_{s,p_s,q_s} \right) \left(\prod_{s=1}^{4} e_{s,p_s,q_s} \right)^t \\
&= \sum_{p_1=0}^{1} \sum_{p_2=0}^{1} \sum_{p_3=0}^{1} \sum_{p_4=0}^{1} \sum_{q_1=0}^{4} \sum_{q_2=0}^{4} \sum_{q_3=0}^{4} \sum_{q_4=0}^{4} \left(\prod_{s=1}^{4} b_{s,p_s,q_s} \right) \sum_{t=0}^{\infty} \text{Pois}(t) \left(\prod_{s=1}^{4} e_{s,p_s,q_s} \right)^t \\
&= \sum_{p_1=0}^{1} \sum_{p_2=0}^{1} \sum_{p_3=0}^{1} \sum_{p_4=0}^{1} \sum_{q_1=0}^{4} \sum_{q_2=0}^{4} \sum_{q_3=0}^{4} \sum_{q_4=0}^{4} \left(\prod_{s=1}^{4} b_{s,p_s,q_s} \right) \exp \left\{ K\theta\nu \left[\left(\prod_{s=1}^{4} e_{s,p_s,q_s} \right) - 1 \right] \right\}.
\end{aligned}$$

where the last line comes from the moment generating function of a Poisson distribution. Finally, the pattern probability P_π is obtained by summing over the 2^5 states of i and a, as indicated previously. This provides a route to computing all 81 multi-study pattern probabilities required for likelihood evaluation. R program is developed to implement this calculation.

2.4 Inference computations

Here we describe our computational approach to likelihood-based inference via numerical optimization and Markov chain Monte Carlo (MCMC). The methods were implemented in code and tested extensively to assure the fidelity of reported numerical findings. Some computational tests are discussed in Section 2.5.

2.4.1 Likelihood evaluation and maximization

Based on formulas for pattern probabilities $\{P_\pi\}$ the log-likelihood (2.1) was available numerically. We used some convenient facilities in the R system to organize the rather complex sums (R Core Development Team, 2011, version 2.13.1). To maximize the log-likelihood, we used the R function nlminb. Working on the log scale for ν and the logit scale for all probabilities, the optimization code was initiated at the zero vector to compute maximum likelihood estimates (MLEs). Numerical experiments showed insensitivity to a range of starting configurations. To go beyond point estimates, we develop Bayesian inference under a flat prior for the system parameters in ψ.

2.4.2 Posterior computation

The Metropolis-Hastings method was used to construct a Markov chain to simulate the joint posterior density

$$P(\psi|\text{data}) \propto \exp \mathcal{L}(\psi)$$

(i.e., flat prior). Importantly, we did not run MCMC over the high-dimensional space including latent variables, because we were able to solve these analytically. Chains were initiated at the MLE values, run for length 250000 scans, and subsampled every 100 scans for final output. Our sampler produced a sequence $\psi^1, \psi^2, \ldots, \psi^B$ of parameter vectors according to standard Metropolis-Hastings updating rules (e.g., Robert and Cassella, 1999, page 231.) We systematically scanned the 12 parameter values, and used

a base set of proposal distributions that modified one parameter at a time. The base proposal distribution for ν was exponential, with mean at the fixed value 1/50. All other parameters resided in $(0, 1)$, and for each we used a uniform window proposal; window length 0.025 gave acceptance rates in the range 28% to 70%. Concerned about possible poor mixing caused by posterior correlation between θ and ν, we included a joint update involving $(\theta, \nu) \longrightarrow (\theta c, \nu/c)$ for a Gamma-distributed multiplier c (shape, rate both 50, so mean 1). Starting positions of parameters were the MLE's from the numerical optimization. Trace plots (Figure 5) and autocorrelation plots (Figure 6) indicate good mixing properties.

We used marginal posterior means to estimate the parameters and the equi-tail percentile method to obtain confidence (equivalently *credible*) intervals. Marginal posterior means and MLE's gave comparable results; the sampling approach lead more directly to confidence intervals, and so we report only results from the MCMC output.

2.4.3 Posterior distribution of N

The total number of influenza-involved genes is $N = \sum I_g$. By our approach we have marginalized the involvement indicators, and so the posterior of N needs to be obtained through further post-processing of the MCMC output. We estimate N by $G \cdot \hat{\theta}$ where $\hat{\theta}$ is the mean of posterior distribution of θ from MCMC. The posterior distribution of N

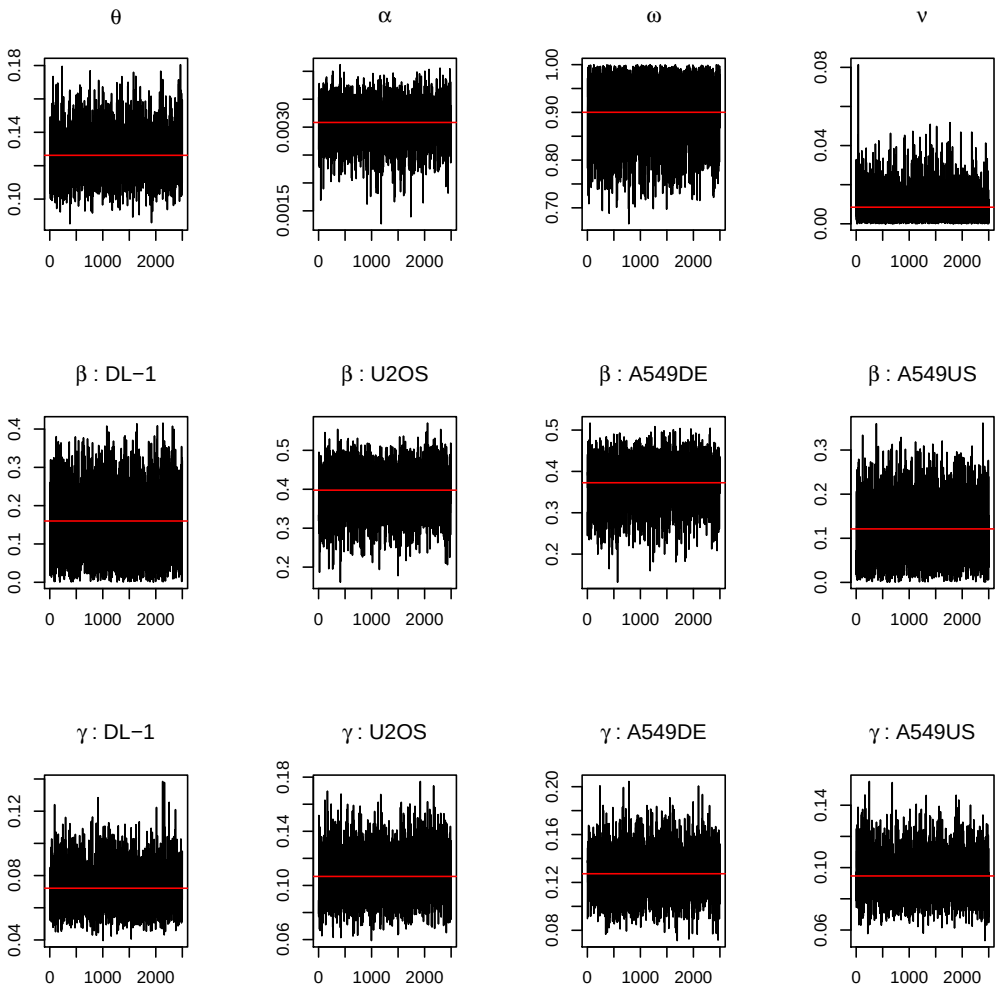


Figure 5: Trace plots parameter values from MCMC output.

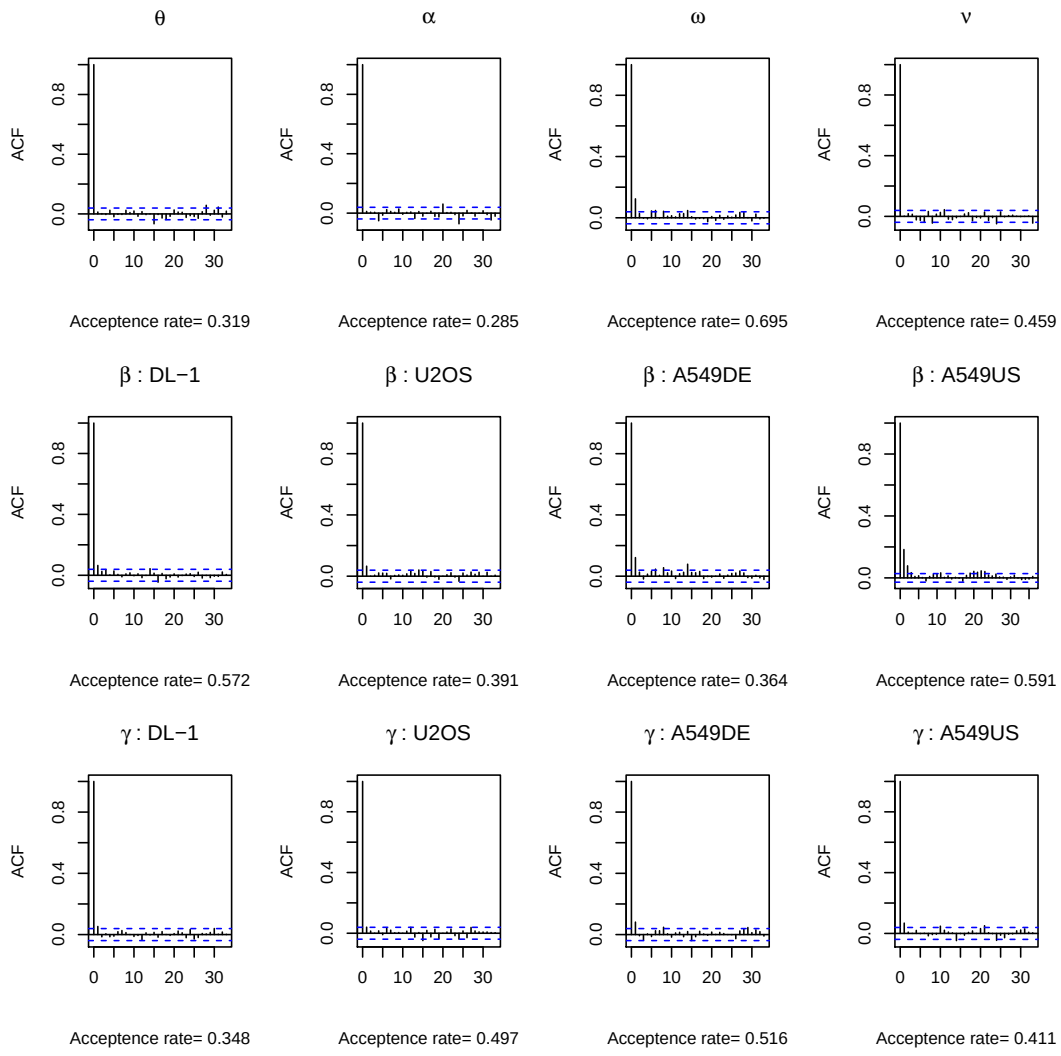


Figure 6: Autocorrelation plots of MCMC output.

is approximated as follows. For $n = 0, 1, \ldots, G$,

$$\begin{aligned} P(N = n|\text{data}) &= \int P(N = n|\psi, \text{data}) p(\psi|\text{data})\, d\psi \\ &\approx \frac{1}{B} \sum_{b=1}^{B} P(N = n|\psi^b, \text{data}) \end{aligned} \tag{2.5}$$

where $\{\psi^b\}$ constitute the MCMC output. A priori, N given ψ is distributed Binomial(G, θ), but in conditioning on the data we have a different distribution for N, even with ψ in hand. Being a sum of independent but differently-distributed Bernoulli trials, N has a Poisson Binomial distribution (Thomas and Taub, 1982). For example, the one gene that is confirmed by all 4 studies is more likely to be truly involved than a gene confirmed just once. Thomas and Taub's recursion method is applied to evaluate the probability mass of N at each sampled parameter setting ψ^b. Estimated distribution of N is illustrated in Figure 7.

2.4.4 Error rate inference

Depending on the reference set of genes, there are different ways to measure false positive and false negative error rate. In any case, we are thinking of errors in a single study, s, and define four rates

$$\begin{aligned} \text{FDR}(\psi) &= P\left(I_g = 0 \middle| D_{g,s} = C_{g,s} = 1\right) \\ \text{FNDR}(\psi) &= P\left(I_g = 1 \middle| D_{g,s} \times C_{g,s} = 0\right) \\ \text{FP}(\psi) &= P\left(D_{g,s} = C_{g,s} = 1 \middle| I_g = 0\right) \\ \text{FN}(\psi) &= P\left(D_{g,s} \times C_{g,s} = 0 \middle| I_g = 1\right). \end{aligned}$$

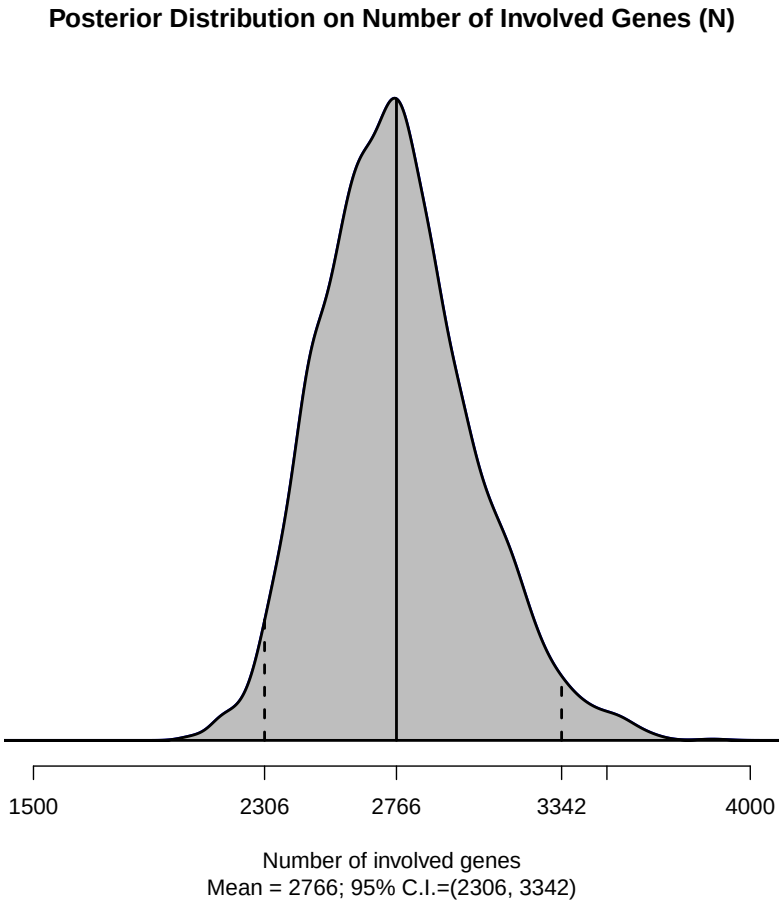


Figure 7: Posterior distribution of number N of involved genes

Respectively, these are rates of false discovery (FDR), false nondiscovery (FDNR), false positive (FP), and false negative (FN), and they all depend on the vector of system-level parameters ψ. These rates depend on probabilities in the proposed models, and are marginal to latent variables recording accessibility and off-target counts. Specifically, $P(C_{g,s} = c, D_{g,1} = d | I_g = i)$ is a summation of $Q_{s,i,a_s,u}$ over values of $A_{g,s}$ and $T_{g,s}$, as presented in Section 2.3.4 of this supplement. That covers FP and FN; for FDR and FNDR, observe that

$$\begin{aligned}
\mathrm{FDR}(\psi) &= \frac{P(D_{g,s} = C_{g,s} = 1 | I_g = 0)\, P(I_g = 0)}{P(D_{g,s} = C_{g,s} = 1 | I_g = 0)\, P(I_g = 0) + P(D_{g,s} = C_{g,s} = 1 | I_g = 1)\, P(I_g = 1)} \\
&= \frac{\mathrm{FP}(\psi)\,(1-\theta)}{\mathrm{FP}(\psi)\,(1-\theta) + (1 - \mathrm{FN}(\psi))\;\theta} \\
\mathrm{FNDR}(\psi) &= \frac{[1 - P(D_{g,s} = C_{g,s} = 1 | I_g = 1)]\, P(I_g = 1)}{[1 - P(D_{g,s} = C_{g,s} = 1 | I_g = 1)]\, P(I_g = 1) + [1 - P(D_{g,s} = C_{g,s} = 1 | I_g = 0)]\, P(I_g = 0)} \\
&= \frac{\mathrm{FN}(\psi)\,\theta}{\mathrm{FN}(\psi)\,\theta + (1 - \mathrm{FP}(\psi))\;(1-\theta)}.
\end{aligned}$$

Point estimates of error rates were obtained by plugging in an estimate $\hat{\psi}$ of system parameters, using the DL-1 study as a reference. Bayesian confidence sets were obtained by percentile error rate values computed across MCMC samples $\{\psi^b\}$. Density plots of error rates (Figure 8) show that false negative errors are higher their false positive counter parts. Point estimates and credible intervals are summarized in Table 2.

2.5 Diagnostics

This section provides details on model validation. We employed a variety of computer experiments for related purposes: (1) to test that our code was calculating what we intended it to calculate, (2) to assess goodness-of-fit of the proposed model, (3) to obtain

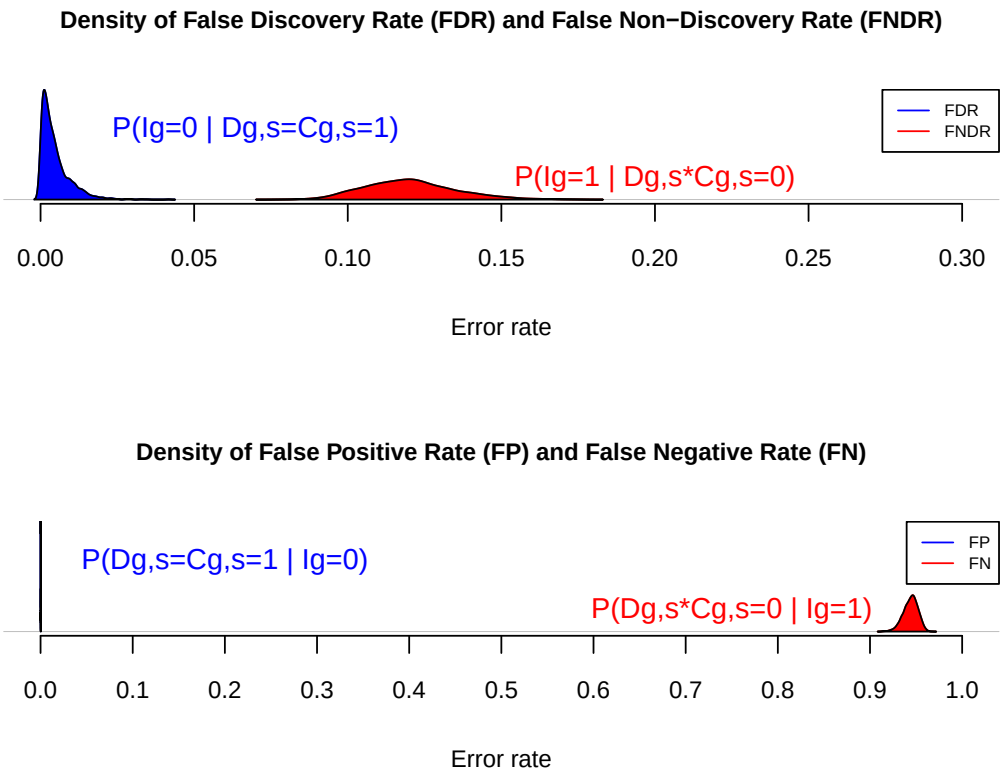


Figure 8: Posterior density of error rates. Upper: false discovery rate (FDR) versus false non-discovery rate (FNDR). Lower: false positive (FP) versus false negative (FN). Estimations are based on posterior samples. The false negative rates are higher than false positive rates.

Table 2: Point estimates of parameters, number of involved genes and error rates and their 95% credible intervals, multi-study influenza data.

Parameter	Point Estimate		95% C.I.
	MLE	Posterior Mean	
$\hat{\theta}$	0.128	0.126	(0.101, 0.158)
$\hat{\alpha}$	0.003	0.003	(0.002, 0.004)
$\hat{\beta} : DL-1$	0.112	0.159	(0.011, 0.338)
$\hat{\beta} : U2OS$	0.360	0.398	(0.267, 0.502)
$\hat{\beta} : A549DE$	0.333	0.372	(0.246, 0.470)
$\hat{\beta} : A549US$	0.067	0.122	(0.010, 0.264)
$\hat{\gamma} : DL-1$	0.065	0.072	(0.050, 0.103)
$\hat{\gamma} : U2OS$	0.097	0.107	(0.075, 0.145)
$\hat{\gamma} : A549DE$	0.116	0.127	(0.091, 0.169)
$\hat{\gamma} : A549US$	0.086	0.095	(0.069, 0.126)
$\hat{\omega}$	0.834	0.900	(0.754, 0.996)
$\hat{\nu}$	< 0.001	0.009	$(< 0.001, 0.032)$
Number of Involved Genes	MLE	Posterior Mean	95% C.I.
N	2821	2766	(2306, 3342)
Error Rate of DL-1 Study	MLE	Posterior Mean	95% C.I.
FDR	< 0.001	0.005	(0.000, 0.017)
$FNDR$	0.122	0.120	(0.095, 0.152)
FP	< 0.001	< 0.001	$(0, < 0.001)$
FN	0.945	0.944	(0.927, 0.958)

and evaluate model-based predictions, and to (4) check the robustness of conclusions to various model assumptions,

2.5.1 Consistency checks

The code base was relatively complex and required substantial testing. Among the basic checks was a useful consistency check. In parametric models, the MLE is known to be consistent. Hence, if we simulated sufficiently many draws from the 81-cell multinomial (i.e., sufficiently many genes), the computed MLE would need to be close to the generating parameter vector. In one test, we increased the genome size from 22000 to 10^6, generated the 81-pattern counts from various parameter settings, and ran the optimization code to estimate the underlying parameters. Parameters values were accurately recovered in all cases.

2.5.2 Predictive checks

Model development was characterized by a series of tests of the model's ability to recapitulate features in the data, as well as to represent presumed structures in RNAi data. (i.e. we did not start with a model as complex as the one finally presented here!) We employed forward simulation to generate synthetic multi-study data: i.e., we repeatedly simulated latent involvement indicators, accessibility indicators, and off-target counts, followed by detection and confirmation indicators, after fixing the system-level parameters at certain values. based on various predictive checks. Table 3 shows observed counts

compared to estimates from 1000 multi-study simulations at the fitted parameter values. Discrepancies in the generally good fit are attributable to Monte Carlo error and also the approximation error originating in our treatment of the confirmation-screen data. Figure 9 is a marginal histogram, from this same simulation, showing the number of confirmed genes from across the multiple studies (we observed 614 and the fitted predictive distribution covers this value well.) Figure 10 reveals another characteristic of the observed data that is well approximated by the fitted model; namely, the overall numbers of detected and confirmed genes per study. These three basic checks indicate a good model fit for the statistics considered. We note that a version of the model which did not allow parameter heterogeneity among studies showed lack of fit in the detection/confirmation plot.

2.5.3 Leave-one-study-out diagnostics

As further validation of our model-based approach, we checked how well it estimated parameters when data from three studies were used to fit the model. This cross-validation exercise provides some assessment of the stability of inference. With four studies there are four leave-one-out cases; for each we developed inference computations for model fitting. Some care was required to reduce from the table of 81 (3^4) four-study patterns to tables of 27 (3^3) three-study detection/confirmation patterns.

Following the posterior prediction strategy described below (Section 4), we simulated counts of how many novel genes would be confirmed by a fourth study given data from

Table 3: Multi-study data in count format. Column "observation" shows the number of genes N_π having detection and confirmation pattern π. For each study, code 0 means not detected in the primary screen, 1 means detected in the primary screen but not confirmed in the secondary screen, and 2 means detected and confirmed in both screens. Assume a full genome is of size G=22000. For the observed 81-pattern counts, the last two columns list their estimated values from the fitted model and the empirical estimates from 1000 simulations.

DL-1	U2OS	A549DE	A549US	Pattern π	Observation N_π	Model Fit	Simulation Mean	Std
0	0	0	0	0000	21016	20999.02	20799.02	34.35
0	0	0	1	0001	71	75.8	98.84	9.87
0	0	0	2	0002	179	180.96	165.38	12.27
0	0	1	0	0010	106	109.12	185.22	13.8
0	0	1	1	0011	0	0.61	2.66	1.63
0	0	1	2	0012	6	4.06	10.62	3.2
0	0	2	0	0020	126	138.24	130.39	11.51
0	0	2	1	0021	2	0.86	2.44	1.6
0	0	2	2	0022	18	12.16	11.77	3.53
0	1	0	0	0100	113	104.34	168.22	13.07
0	1	0	1	0101	0	0.56	2.32	1.53
0	1	0	2	0102	1	3.71	9.4	3.13
0	1	1	0	0110	0	1.24	7.29	2.78
0	1	1	1	0111	0	0.01	0.13	0.36
0	1	1	2	0112	0	0.08	0.62	0.78
0	1	2	0	0120	2	2.83	7.37	2.68
0	1	2	1	0121	0	0.02	0.13	0.36
0	1	2	2	0122	0	0.25	0.67	0.82
0	2	0	0	0200	111	105.26	99.19	9.88
0	2	0	1	0201	1	0.65	1.96	1.39
0	2	0	2	0202	3	9.26	8.93	3.03
0	2	1	0	0210	2	2.36	6.41	2.56
0	2	1	1	0211	0	0.02	0.1	0.32
0	2	1	2	0212	0	0.21	0.6	0.8
0	2	2	0	0220	3	7.09	6.91	2.5
0	2	2	1	0221	0	0.04	0.13	0.36
0	2	2	2	0222	3	0.63	0.62	0.84
1	0	0	0	1000	80	74.85	96.79	10.16
1	0	0	1	1001	0	0.37	0.98	1.01
1	0	0	2	1002	2	1.15	2.97	1.68
1	0	1	0	1010	0	0.58	2.47	1.61
1	0	1	1	1011	0	0.01	0.04	0.2
1	0	1	2	1012	0	0.03	0.18	0.44
1	0	2	0	1020	1	0.88	2.42	1.59
1	0	2	1	1021	0	0.01	0.04	0.2
1	0	2	2	1022	0	0.08	0.21	0.47
1	1	0	0	1100	0	0.54	2.15	1.49
1	1	0	1	1101	0	0	0.04	0.18
1	1	0	2	1102	0	0.02	0.17	0.44
1	1	1	0	1110	0	0.01	0.12	0.33
1	1	1	1	1111	0	0	0	0.04
1	1	1	2	1112	0	0	0.01	0.1
1	1	2	0	1120	0	0.02	0.14	0.36
1	1	2	1	1121	0	0	0.01	0.09
1	1	2	2	1122	0	0	0.01	0.11
1	2	0	0	1200	0	0.67	1.83	1.34
1	2	0	1	1201	0	0	0.04	0.18
1	2	0	2	1202	0	0.06	0.17	0.42
1	2	1	0	1210	0	0.02	0.1	0.32
1	2	1	1	1211	0	0	0.01	0.08
1	2	1	2	1212	0	0	0.01	0.09
1	2	2	0	1220	0	0.04	0.12	0.35
1	2	2	1	1221	0	0	0	0.07
1	2	2	2	1222	0	0	0.01	0.09
2	0	0	0	2000	127	126.22	116	10.51
2	0	0	1	2001	1	0.79	2.1	1.45
2	0	0	2	2002	2	11.09	10.34	3.16
2	0	1	0	2010	4	2.83	7.45	2.69
2	0	1	1	2011	0	0.02	0.13	0.37
2	0	1	2	2012	0	0.25	0.64	0.79
2	0	2	0	2020	6	8.49	8.12	2.88
2	0	2	1	2021	0	0.05	0.14	0.37
2	0	2	2	2022	3	0.75	0.72	0.83
2	1	0	0	2100	4	2.59	6.66	2.57
2	1	0	1	2101	0	0.02	0.12	0.35
2	1	0	2	2102	0	0.23	0.55	0.73
2	1	1	0	2110	0	0.06	0.42	0.64
2	1	1	1	2111	0	0	0.01	0.08
2	1	1	2	2112	0	0.01	0.03	0.17
2	1	2	0	2120	1	0.17	0.46	0.69
2	1	2	1	2121	0	0	0.01	0.08
2	1	2	2	2122	0	0.02	0.04	0.2
2	2	0	0	2200	2	6.46	6.16	2.56
2	2	0	1	2201	0	0.04	0.12	0.34
2	2	0	2	2202	0	0.57	0.56	0.75
2	2	1	0	2210	0	0.14	0.34	0.56
2	2	1	1	2211	0	0	0.01	0.09
2	2	1	2	2212	1	0.01	0.04	0.19
2	2	2	0	2220	2	0.44	0.43	0.67
2	2	2	1	2221	0	0	0.01	0.08
2	2	2	2	2222	1	0.04	0.05	0.22

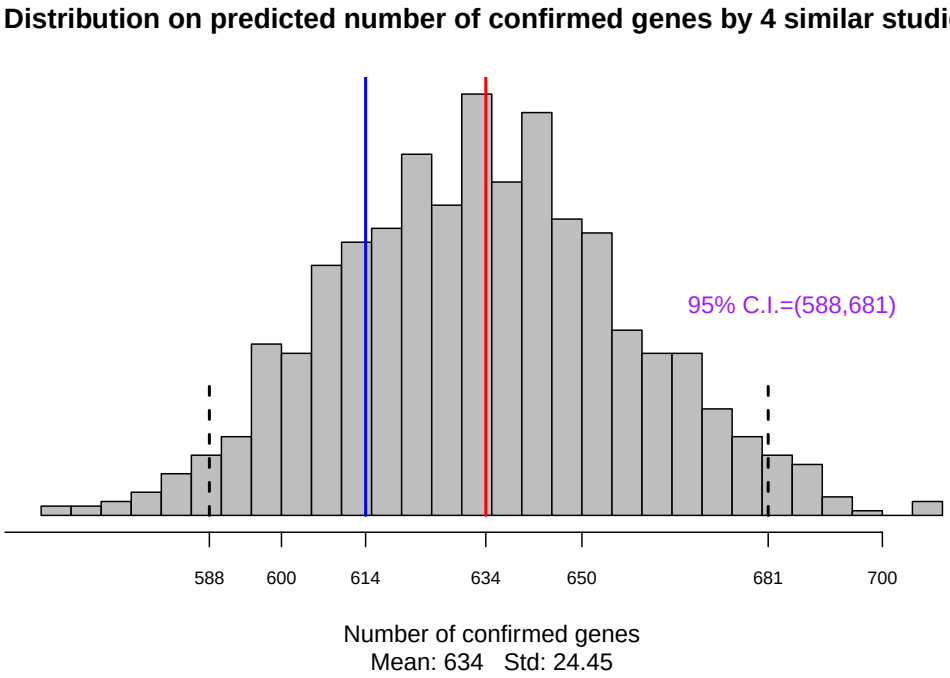


Figure 9: Goodness-of-fit simulations. Histogram of number of genes confirmed jointly by all 4 studies from 1000 simulations based on the fitted model.

three studies, and we compared these predictions to available data (Table 4). In all four cases, the count predicted from triple-study training data matched well to the observed test data that had been left out.

Parameter estimates from the four leave-one-out cases are shown in Table 5. Reflecting inferential stability, these estimates are very similar to results based on all four studies. Sizes of detection and confirmation patterns influence the results. For example, $A549DE$ has the most overlap with other studies in confirmed genes, when it is excluded, the estimated θ value is most affected.

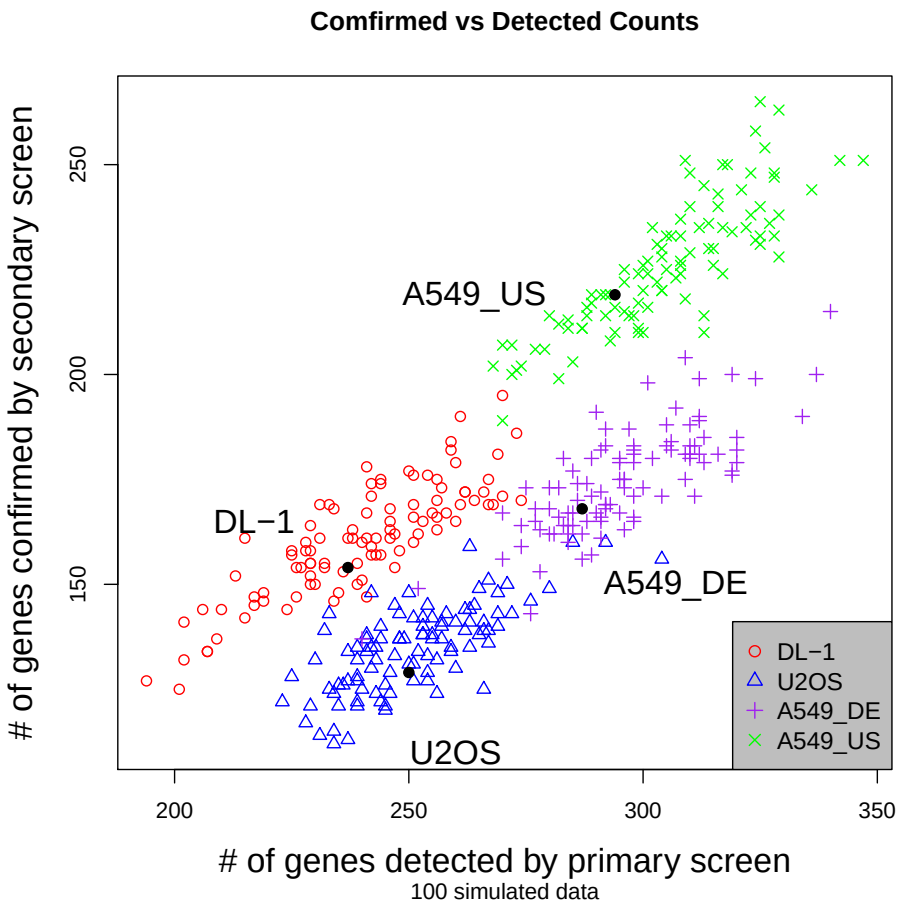


Figure 10: Observed numbers of detections/confirmations over four studies (black dots) compared to simulated values (colored symbols) from fitted model.

Table 4: Predicted number of extra genes confirmed by a 4^{th} study based on modeling the other three studies.

Leave Out	Predicted Additional	95% Prediction Interval	Observed Additional
DL-1	139	(56, 253)	136
U2OS	143	(67, 253)	114
A549DE	156	(57, 330)	131
A549US	131	(55, 240)	188

Table 5: Estimated parameters by four ways of leaving out one study.

Leave out	$\hat{\theta}$	$\hat{\omega}$	$\hat{\gamma}$			
			DL-1	U2OS	A549DE	A549US
DL-1	0.106	0.896	-	0.367	0.350	0.111
U2OS	0.102	0.885	0.190	-	0.368	0.127
A549DE	0.192	0.892	0.203	0.422	-	0.129
A549US	0.115	0.890	0.121	0.380	0.337	-
Leave out	$\hat{\alpha}$	$\hat{\nu}$	$\hat{\beta}$			
			DL-1	U2OS	A549DE	A549US
DL-1	0.003	0.010	-	0.118	0.144	0.112
U2OS	0.003	0.010	0.096	-	0.159	0.120
A549DE	0.003	0.013	0.054	0.079	-	0.066
A549US	0.003	0.013	0.075	0.113	0.129	-

2.5.4 Robustness checks

In developing model-based inference for factors affecting multi-study RNAi data, we had formulated a range of models prior to the final model presented here. We settled on the final model because it exhibited a goodness of fit, it made plausible predictions, and it captured what we could formulate about the key systematic sources of variation. Earlier models (not shown) failed on one or more of these criteria. We report here one additional test of the final model assumptions.

Our main computations treated the number T_g of influenza-involved off-targets of each first-round siRNA pool as Poisson distributed (with mean of $K\theta\nu$ to account for the pool size, the involvement rate, and the overall rate of off targeting). A first-principles argument supports this assumption, and experience suggests that the impact of violations in this assumption on other inferences is probably minimal. However, the

limited data on off-target rates suggests variation in T_g that is more extensive than the Poisson (Kulkarni, *et al.* 2006). To check the robustness of our Poisson-based approach, we investigated replacing the Poisson distribution with the Negative Binomial distribution to allow potential overdispersion. Specifically, for a Gamma distributed random variable C, with both shape and rate parameters equal to κ (and thus mean 1), we considered:

$$T_g|[C=c] \;\sim\; \text{Poisson}(K\theta\nu c),$$

which implies

$$T_g \;\sim\; \text{Negative Binomial}\left(\kappa, \frac{K\theta\nu}{K\theta\nu+\kappa}\right)$$

and parameterized so the mean continues to be $K\theta\nu$. Small $\kappa > 0$ corresponds to substantial overdispersion, while $\kappa \longrightarrow \infty$ recapitulates the Poisson model. Complexity of the multi-study pattern probabilities put a full analysis of the Negative Binomial model beyond our reach, though we were able to obtain pattern probabilities in several boundary cases. As the siRNA pool size K gets large, off-target counts $T_{g,s}$ from different studies become independent, and thus data from the separate studies become conditionally independent given the involvement indicators. In this limiting case,

$$T_{g,s} \sim \text{Negative Binomial}\left(\kappa, \frac{4\theta\nu\gamma_s}{4\theta\nu\gamma_s+\kappa}\right)$$

and the simplifications arising from independent studies enabled us to compute all pattern probabilities for likelihood analysis. In this independent-study case, we recomputed

the maximum likelihood parameter values over a grid of κ values in $(0, 1000)$. We found very little dependendence of estimates on the value of κ. To link back to the actual case (among-study dependence, and small K), we retained the Poisson model but varied K over the range $(4, 1000)$. Again we saw very little dependence of the MLEs on the value of K. This lack of sensitivity to K and κ may be due to the data favoring very small mean off-target rate ν; with small ν, the likelihood surface is relatively flat over the domains of K and κ.

As a further investigation of the off-target rate, we considered a range of values ν (on a grid) and at each one profiled the remaining parameters by maximum (profile) likelihood. Results shown in the main text show that increasing ν does not explain the data well (decreasing likelihood fit), and further that a reason for this is the constrained model's inability to explain the relatively high confirmation rate. Figure 11 presents another view of this lack-of-fit, in the spirit of the goodness-of-fit plot in Figure 10.

2.6 Predicting outcomes in future siRNA studies

The model-based approach provides a mechanism for predicting outcomes of further siRNA studies. We pursued posterior predictive simulation in which parameter draws from the MCMC output were used to seed forward simulation of further siRNA studies (we mixed over the four posteriors for study-specific error rates to incorporate parameter settings for these hypothetical future studies.) Specifically, for each of 2000 posterior

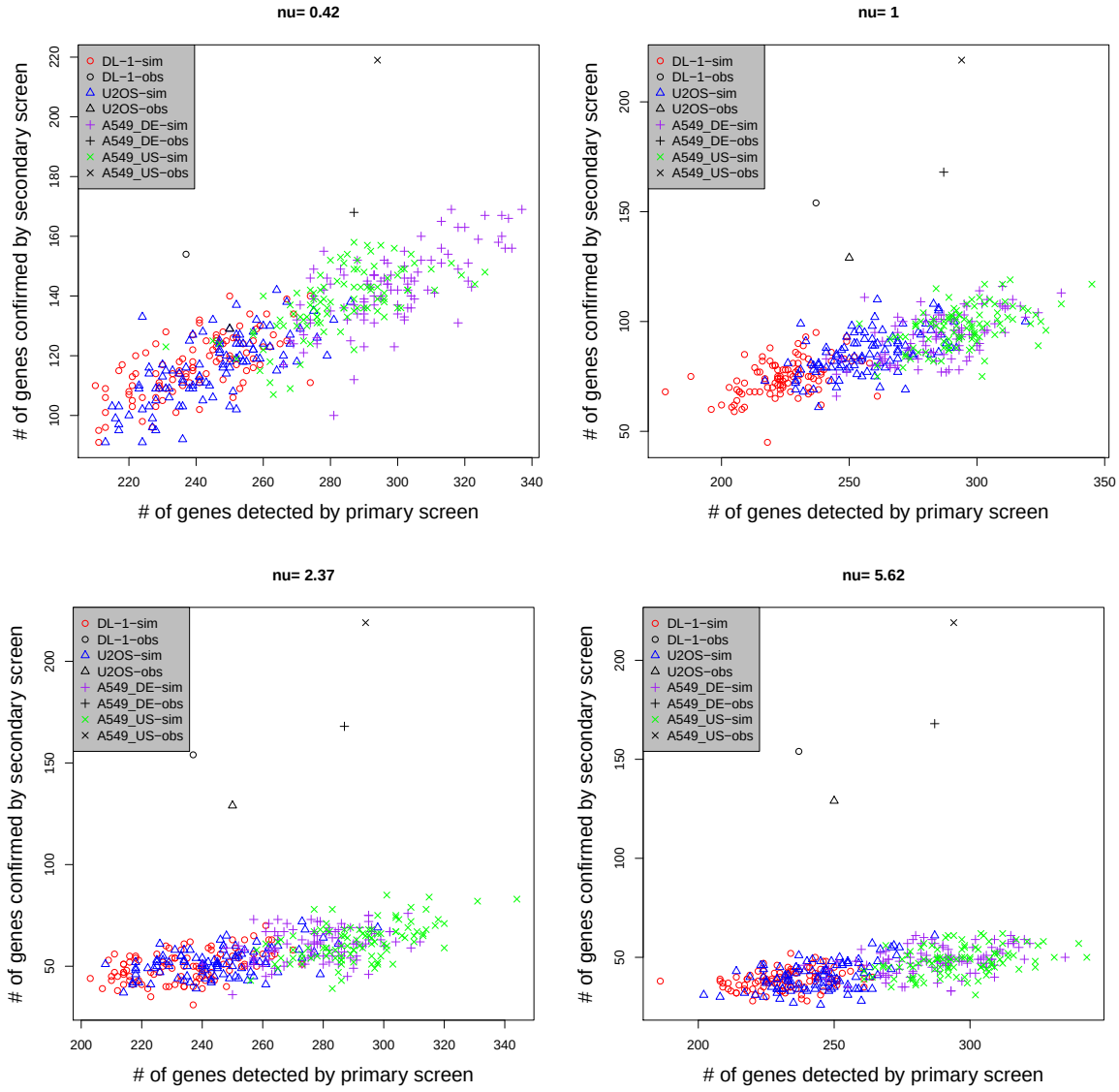


Figure 11: Lack-of-fit consequence of raising off-target rate ν. In profile computations, we fixed ν at a moderately large value, and estimated other parameters by maximum likelihood. Shown is a scatterplot revealing the constrained model's inability to explain the high confirmation rate. Simulation data points go astray further from observations as ν increases. (Compare to Figure 10.)

draws, we simulated a future trajectory of up to 50 future studies. Each trajectory represented a state of nature, and so corresponded to a single draw of involvement indicators $\{I_g\}$ and off-target numbers $\{T_g\}$. Along each trajectory, we sampled accessibilities and study-specific off-target numbers $T_{g,s}$ at each step, and we generated detections and confirmations. We kept track of how many novel genes were confirmed along the way. A subtlety of the computation was making it *posterior* predictive. There were up to 81 different kinds of trajectories, depending on the four-study data on a given gene from the existing data; and each kind corresponded to different involvement and off-target inferences. Predictions are shown in Figure 12.

2.7 Application to HIV studies

As a further validation exercise we checked how the model-based approach worked on an independent collection of three RNAi experiments from the study of HIV (Brass *et al.* 2008; Zhou *et al.* 2008; Konig *et al.* 2008). These studies used similar two-stage designs and experimental procedures to the four influenza studies, and so we organized the mult-HIV-study data into a table holding patterns of detection/confirmation across the three studies and we fit the proposed model to these data.

Because only one of the three HIV studies reported data on both primary and secondary screens, we did not have access to all $3^3 = 27$ counts, and were forced to use a reduced set of 12 pattern counts (see Table 6). Recall that in original patterns, digits

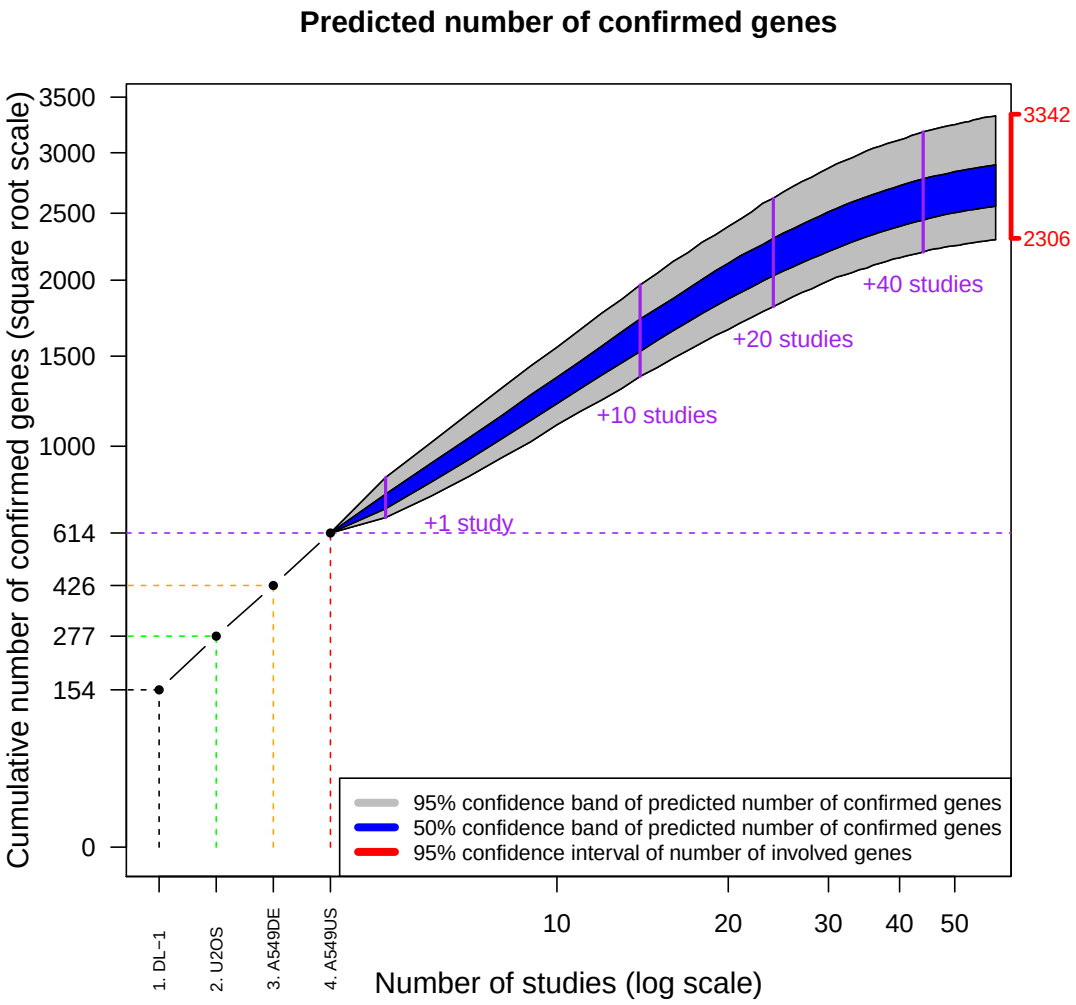


Figure 12: Predictions from 2000 simulated study sequences, with each sequence determined by a parameter setting obtained by Markov chain Monte Carlo and subsequently with future-study counts simulated prospectively from the specified multinomial model. The number of confirmed genes increases and stabilizes after 40 studies to a range consistent with the inferred number of influenza-involved genes (indicated in red number, as CI 95%). Grey and blue bands express different levels of confidence.

$0, 1, 2$ refer to respectively detection and confirmation status $\{D_{g,s} = C_{g,s} = 0\}, \{D_{g,s} = 1, C_{g,s} = 0\}, \{D_{g,s} = C_{g,s} = 1\}$. For instance, that a gene has pattern 201 means that it is detected and confirmed by study 1, not detected nor confirmed by study 2, and detected but not confirmed by study 3. Suppose we only have detection and confirmation data from study 3 in HIV meta analysis, then we are able to identify detection and confirmation status 0, 1, or 2 for only study 3. For the other 2 studies, we are only able to identity if the status is 2 or not, but not able to differentiate 0 from 1. Therefore, what have previously been patterns 201 and 211 need to be collapsed into one single pattern which is collapsed pattern 9 in Table 6. Because of the limited available pattern information, we used a common false negative error β instead of 4 study specific ones for a better model fit.

Table 6: HIV analysis: relation between collapsed patterns and original patterns.

Collapsed Patterns	Original Patterns			
1	222			
2	221			
3	220			
4	202	212		
5	022	122		
6	200	210		
7	020	120		
8	021	121		
9	201	211		
10	002	012	102	112
11	000	010	100	110
12	001	011	101	111

Estimated parameters, number of involved genes, error rates and their 95% credible

intervals are summarized in Table 7. We use acronyms to refer to the 3 studies (SCI: Brass *et al.* 2008; CHM: Zhou *et al.* 2008; CEL: Konig *et al.* 2008). Error rates are calculated based on accessibility rate estimated for CHM study.

Table 7: Estimated parameters, number of involved genes and error rates and their 95% credible intervals in HIV studies.

Parameter	Point Estimate	95% C.I.
$\hat{\theta}$	0.285	(0.209, 0.390)
$\hat{\alpha}$	0.002	(0.000, 0.003)
$\hat{\beta}$	0.078	(0.009, 0.154)
$\hat{\gamma} : SCI$	0.051	(0.036, 0.070)
$\hat{\gamma} : CEL$	0.055	(0.038, 0.074)
$\hat{\gamma} : CHM$	0.043	(0.030, 0.058)
$\hat{\omega}$	0.833	(0.651, 0.990)
$\hat{\nu}$	0.017	(0.000, 0.061)
Number of Involved Genes	Point Estimate	95% C.I.
N	6277	(4591, 8620)
Error Rate of CHM Study	Point Estimate	95% C.I.
FDR	0.008	(0.000, 0.032)
$FNDR$	0.278	(0.201, 0.384)
TP	0.037	(0.026, 0.051)
TN	1.000	(1.000, 1.000)

2.8 Concluding remarks

This part of thesis concerns modeling agreement among replicated genome-wide RNAi studies, more specifically, estimating sources that cause variations in data and assessing the relative size of false-positive and false-negative errors.

Our contributions in the RNAi project include proposing a novel sampling model

that deals with multi-record observational data from RNAi studies, and developing sophisticated computational schemes for likelihood inference. In its generative form, the model specifies the probability of observing any particular multi-study data set. In its inferential form, it indicates the likelihood assigned to any particular parameter setting in light of observed data. We generated likelihood-based inference via both numerical optimization and Markov chain Monte Carlo (MCMC), and confirmed that the point estimates are consistent. The posterior inference of error rates point to false negative factors to account for more of the limited agreement.

To be more confident of the whole modeling approach, we conducted various model diagnostics including (1) consistency checks to test whether our code was calculating what we intended it to calculate, (2) predictive checks to evaluate goodness-of-fit, (3) leave-one-out studies to test stability of the inference and (4) robustness checks of conclusions to various model assumptions. The results show that our sampling model passed all the tests.

The numbers of genes that would be confirmed by future studies of the similar system were predicted from posterior predictive simulation. The trend suggests that the total number of confirmed genes will increase notably and stabilizes after 40 studies to a range consist with the inferred number of influenza-involved genes. Our method was also applied to model agreement from outcomes of 3 RNAi studies for HIV.

Chapter 3

Simultaneous functional category analysis (SFCA)

3.1 Overview

Integrating experimental genomic data with exogenous functional information is important in statistical genomics for the purposes of effective data reduction and boosting weak gene-level signals. Most available functional category analysis methods, introduced in Section 1.2, can be categorized into three classes: (1) one-at-a-time methods (e.g. Subramanian et al. (2005), Newton et al. (2007)) which ignore complexity of the functional record; (2) sequential methods (e.g. Liang and Nettleton (2010)) that fail to incorporate the overlapping structure of categories and are hard to interpret; (3) model-based methods which either oversimplify the model assumption (e.g. Bauer et al. (2010)) or are unduly challenging in computations (e.g. Newton et al. (2012)). Our goal is to develop a methodology that addresses these limits. In particular, we are most interested in answering the following questions: (1) Can we propose a model that incorporates

the overlapping and hierarchical structure of functional categories and generates inferences that respect this structure? (2) Compared to existing model-based methods, does our model perform better in detecting subtle signals? (3) Can we develop an efficient algorithm to apply the proposed method to large-scale category analysis problems?

This chapter is organized to provide evidence in response to these questions. Our method, called SFCA, is described in Section 3.3. It relies on a model originally proposed in model-based gene set analysis (MGSA) by Bauer et al. (2010), followed by an important model assumption called *activation hypothesis* based on which our model is developed to assure identifiability of the model. In Section 3.4 we demonstrate advantages of SFCA over MGSA in terms of consistency and efficiency based on analytically developed posterior summaries in different scenarios. Section 3.5 introduces Markov Chain Monte Carlo algorithm we have developed for computations. SFCA is applied to genome-wide data from the meta analysis of Influenza virus replication in the previous chapter with its finding being compared with MGSA in Section 3.6 . Two other approaches to address role model posterior inferences that we have investigated are presented at the end of this chapter.

3.2 Data structure

First, let us introduce the data structure and some notation that will be used throughout this chapter. Two forms of data are being integrated in the proposed analysis:

genome-wide experimental gene-level data and functional information about the genes as recorded in a bioinformatics resource such as GO or KEGG. With respect to functional information, each gene has annotation profile, which is a vector holding binary indicators of whether or not this gene is annotated to each functional category. Often genes have distinct profiles, but there can be ties. An *atom* is defined to be a maximal set of genes sharing a common profile, following Boca et al. (2010). Thus each atom i corresponds to a profile x_i, and these profiles are usefully arranged as the rows of an incidence matrix X of dimension $N \times C$, where N is the number of atoms (collapsed from G genes, and $N \leq G$) and C is the number of categories recorded in the resource. Element $x_{i,j} = 1$ if and only if genes in atom i are in category j, otherwise $x_{i,j} = 0$. For example, consider a simple system made of the first 4 KEGG pathways by ID order, as summarized below. The gene level incidence matrix to present the functional profiles has dimensions 145×4, where 145 is the total number of genes and 4 is the number of categories.

ID	Functional Category	Number of genes
00010	Glycolysis/Gluconeogenesis	62
00020	Citrate cycle (TCA cycle)	32
00030	Pentose phosphate pathway	26
00040	Pentose and glucuronate interconversions	25

By collapsing the identical rows the new incidence matrix is of dimension 7×4, where 7 is the number of atoms. Each atom is a distinctive row representing an annotation profile.

For example, the first 4 atoms contain genes that are annotated to only 1 pathway; there are 7 genes involved in both *Glycolysis/Gluconeogenesis* and *Citrate cycle (TCA cycle)* (the first two pathways) and not involved in the other two.

Atom i	Number of genes (n_i)
1000	44
0100	25
0010	14
0001	24
1100	7
1010	11
0011	1

Observed genomic data can take various forms, depending on the nature of the experimental system. We focus here on the simplest case in which gene-level data are binary indicators, for example representing which genes show significant differential expression in a microarray study. The initial methods development is for binary gene-level data. The binary case covers a large number of applications where gene lists are reported from experimental data. We later extend the methods to multinomial outcomes. For the binary case, we collapse gene-level data to the atom level by simple counting, and thus denote Y_i to be the number of *positive* genes at atom i, from among the n_i genes at that atom.

For modeling purposes we treat genes sharing the same annotation profile as producing independent and identical distributed data. Also notice that atoms are mutually exclusive and any category can be decomposed into a certain number of atoms. Thus, this meaningful parameterization does not only reduce dimensionality of the incidence matrix but also assists computations and inferences of gene- and category level activities, as we will see in later sections. Boca et al. (2010) introduced atoms in a decision-theoretic analysis of the same basic data-integration problem. They used atoms differently from us, in that they sought a subset of atoms (rather than functional categories) whose activation could explain gene-level data.

3.3 Modeling approach

3.3.1 The role model

Our study is based on the following observation model. For all atoms i, $i \in \{1, 2, ..., N\}$

$$\begin{aligned} Y_i|p_i &\sim \text{Binomial}\{n_i, p_i\} \\ p_i &= p(x_i) \end{aligned} \tag{3.1}$$

where Y_i records atom level data, x_i is the annotation profile of atom i ad p_i is a success probability. Simply, atoms deliver binomial data where the success probability depends in some way on the annotation profile x_i. We further assume that the Y_i are mutually

conditionally independent given $\{p_i\}$. Thus the joint distribution of observations is:

$$P\left(Y_i = y_i, \forall i | \{p_i\}\right) = \prod_{i=1}^{N} p_i^{y_i} (1-p_i)^{n_i - y_i} \tag{3.2}$$

In particular, the role model asserts that p_i depends on profile x_i through latent binary activation variable A_i:

$$p_i = p^{\mathrm{RM}}(x_i) \quad = \quad \begin{cases} \alpha & \text{if } A_i = 0 \\ \gamma & \text{if } A_i = 1 \end{cases} \tag{3.3}$$

where

$$\alpha \quad = \quad \text{false positive measurement error}$$

$$\gamma \quad = \quad 1 - \text{false negative measurement error.}$$

System-wide parameters α and $\gamma > \alpha$ are both in $(0, 1)$. The latent binary variable $A_i \in \{0, 1\}$ indicates atom activity, i.e. $A_i = 1$ means that atom i is *active*, or all genes sharing the same annotation profile are *active*, and $A_i = 0$ otherwise. In this paper, "on/off" are also used to describe states of atoms or categories, as equivalence to "active/inactive". The model says that Y_i has rate of γ if atom i is active, otherwise it has rate α. Simply, active atoms have a higher success probability for observations than do inactive atoms. We also call it "Hot-cold" model to refer to the dichotomous success rates delivered (Newton et al. (2012)).

Bauer et al. (2010) introduced this model for functional category analysis, and developed a method called model-based gene set analysis (MSGA). A key contribution was

to relate latent atom activities $\{A_i\}$ to activities $\{Z_j\}$ which are associated with the functional categories in view. Specifically, the model asserts that an atom is active if and only if at least one of the categories it is annotated to is active, or equivalently

$$A_i = \max_{j:x_{i,j}=1} Z_j = 1 - \prod_{j:x_{i,j}=1} Z_j. \tag{3.4}$$

To each category j latent binary variable $Z_j \in \{0, 1\}$ indicates category's activity, i.e. $Z_j = 1$ means that category j is *active*, and $Z_j = 0$ otherwise. Active is just another way to say non-null, and our inference seeks to identify non-null categories.

Bauer et al. (2010) used a Bayesian network to model gene response with category activities. An important aspect of the model is that it starts with category activities. Atom activities and observed data follow then, as opposed to almost all the other approaches. In Figure 13, from left to right there are different layers of dependency. First of all, category level activities follow *i.i.d.* Bernoulli prior distributions with success rate of π, where π means the proportion of active categories. It is another system parameter that needs to be specified or estimated. Secondly, atom level activities are determined by categories to which they are annotated. For instance, A_1 is decided by Z_1 and Z_2 together. Finally, observations depend on both systematic factors which are atom activities and experimental factors which are false positive and false negative errors involved in the process. For inferences, they proposed to rank categories by MCMC-approximated marginal posterior probability $P(Z_j = 1|Y)$.

Compared to one-at-a-time and sequential methods, MGSA is compelling in that (1)

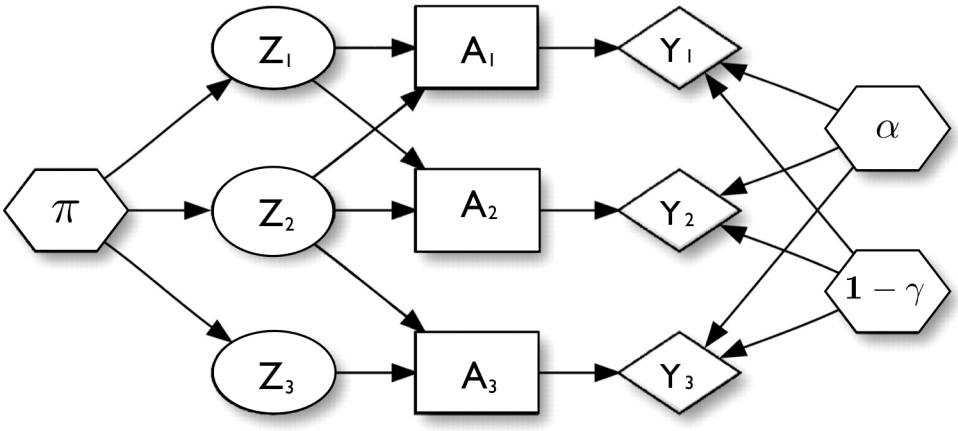


Figure 13: A Bayesian network to model gene response with category activities originally proposed by Bauer et al. (2010). Here is a simple case with 3 categories and 3 atoms.

the role model conveniently incorporates complexity of the category structure and (2) it utilizes structural information for posterior inference of category activities. MGSA is made available in R package (Bauer et al. (2011)).

3.3.2 Activation hypothesis

A curious aspect of MGSA is its use of an *i.i.d.* Bernoulli prior for the category activities $\{Z_j\}$. Considering the possibly extensive overlap among categories, it seems plausible that the activity variables ought to be related. Consider a category c that is fully obtained in another category c'. We call c' a parent set of c and c a child set of c'. This parent-child relationship means that the child set has more specific biological function than the parent set while the parent set has more general function. It routinely occurs in GO. For category activities to have an observationally verifiable meaning, they ought

to respect some basic logical constraints.

In category analysis, to say that "biological property p is activated" is equivalent to say that "genes having property p are activated". If parent category c' is active, then all genes with property c' are active. It means that any subset of c' is active, thus all genes with property c are active, which leads to the activation of child category c. Notice that the implication is not symmetric. If a subset is active it does not follow that a containing set is active. On the other hand, if the child category c is inactive, meaning not all the genes in c are active, it can be inferred that the parent c' must also be inactive, as it contains the inactive genes as well. This basic idea is conveyed in Bauer's model, in that a gene is active if and only if any of the categories to which it is assigned is active. However, the constraint is not respected in the *i.i.d.* prior or the sampled posteriors used in MGSA. For example, in a system with only two categories c and c', MGSA's prior would falsely assign positive probability to the joint outcome $(Z_c, Z_{c'}) = (0, 1)$. To make progress, we require a clear definition of activation. The following assumption is key.

Activation hypothesis: A category is active if and only if all atoms(genes) in the category are active.

The activation hypothesis is equivalent to asserting that any subset of an active category is itself active. It also implies that a set is active only if all its subsets are

active. Since categories could be decomposed into a collection of atoms, the activation hypothesis conveniently applies to atom activities by replacing 'genes' with 'atoms'. Also, the hypothesis is related to the true path rule used in GO, which conveys logical constraints on collections of related categories. One might object to the activation hypothesis for being too strict because it does not allow categories to be activated by a subset of their genes. However, a rich collection of categories, for example in an extreme case all categories constituting a power set of all genes in the system, the active categories we ought to detect are the one with only all active genes and its child sets, rather than any larger category. Furthermore, our language could get unduly complicated if we allow active categories to contain inactive genes.

The collection of functional categories, through the incidence matrix X, thus imposes a possibly large number of constrains on category activation states $\{Z_j\}$ under the activation hypothesis. To proceed, denote space containing all conceivable combinations of category activation states by $\mathcal{Z}_0$, i.e. $\mathcal{Z}_0 = \{Z = (Z_1, Z_2, ..., Z_C) \in \{0, 1\}\}$, where C is the number of categories. Then $|\mathcal{Z}_0| = 2^C$. We define its subspace $\mathcal{Z}$ to include only valid joint states, i.e.

$$\mathcal{Z} = \{Z \in \mathcal{Z}_0 : Z \text{ satisfies the activation hypothesis.}\}$$

Constraints from activation hypothesis are explicitly described below.

- Denote by $c \subset c'$ that category c is a child of category c', then

$$Z_c = 0 \rightarrow Z_{c'} = 0$$

$$Z_{c'} = 1 \rightarrow Z_c = 1;$$

- Let $\{c_1, c_2, ..., c_m\}$ be a group of active categories, i.e. $Z_{c_j} = 1, j = 1, ..., m$, and $\cup_{j=1}^{m} c_j$ the union of all genes annotated to these categories, then

$$Z_c = 1, \text{ for } \forall c \subset \cup_{j=1}^{m} c_j.$$

Depending on category structure and true activations, sometimes $\mathcal{Z} = \mathcal{Z}_0$. For example, in an extreme case where all categories are mutually exclusive, each joint state $Z \in \mathcal{Z}_0$ activates a unique subset of atoms. Thus, all states on $\mathcal{Z}_0$ respect the activation hypothesis, i.e. $\mathcal{Z} = \mathcal{Z}_0$. More generally, a sufficient condition for $\mathcal{Z} = \mathcal{Z}_0$ is that for every $j \in \{1, 2, ..., C\}$ there exists an 'singleton' atom which is annotated to only category j. This condition is obviously satisfied when there is no overlap between categories. Compared to $\mathcal{Z}_0$, $\mathcal{Z}$ is typically smaller and its magnitude can be greatly reduced when categories are heavily overlapping and form hierarchies (examples shown later).

The original role model (Bauer et al. (2010)) gives mapping only from category level to atom level activation states (3.4). The activation hypothesis is helpful because it allows us to invert this mapping. First, consider the range of mapping (3.4)

$$\mathcal{A} = \{a = (a_1, a_2, ..., a_N) : a = a(z), z \in \mathcal{Z}.\}$$

where N is the number of atoms. We proved in Newton et al. (2012),

Proposition 3.1. *Under the activation hypothesis, atom and category activations are in one-to-one correspondence. For* $\forall Z \in \mathcal{Z}$, $\exists A \in \mathcal{A}$, *s.t.*

$$\begin{aligned} Z_j &= \min_{i:x_{i,j}=1} A_i = \prod_{i:x_{i,j}=1} A_i \\ A_i &= \max_{j:x_{i,j}=1} Z_j = 1 - \prod_{j:x_{i,j}=1} Z_j \end{aligned} \tag{3.5}$$

and vice versa.

This property allows identifiability of category activations from atom activations, hence consistency in a scenario when information of atoms goes up (Section 3.4.2). It is also important to develop our computation methods (Section 3.5).

3.3.3 Priors over activation states

Denote by P_0 the *i.i.d.* Bernoulli prior used in Bauer et al. (2010), i.e.

$$P_0(Z_j = 1) = \pi, \forall j$$

where $\pi \in (0, 1)$ is the success probability.

Let $Z = (Z_1, Z_2, ..., Z_C)$. The first prior we consider is, for $z \in \mathcal{Z}$,

$$P_1(Z = z) = \frac{P_0(Z = z)}{P_0(Z \in \mathcal{Z})}$$

A similar approach is possible from the perspective of atom activations. Let $A = (A_1, A_2, ..., A_N)$, for $a \in \mathcal{A}$,

$$P_2(A = a) = \frac{P_{2'}(A = a)}{P_{2'}(A \in \mathcal{A})}$$

where $P_{2'}$ is the *i.i.d.* Bernoulli prior over $\mathcal{A}$, i.e. $P_{2'}(A_i = 1) = \pi, \forall i$. We still use P_2 for the corresponding marginalized distribution associated with categories.

Both priors P_1 and P_2 are conditional priors supported on activities that satisfy the activation hypothesis. They are usually non-uniform and different from each other. When $\pi = 0.5$, all joint states on discrete spaces $\mathcal{Z}$ and $\mathcal{A}$ are uniformly distributed and their marginalized distribution associated with categories P_1 and P_2 are equivalent, because of the one-to-one mapping between $\mathcal{Z}$ and $\mathcal{A}$. In practice, choice of π is discussed in Section 3.4.1.

With a prior distribution $P(Z)$, the posterior distribution of category level activation states given binary observational data in the role model is:

$$\begin{aligned} P(Z|Y) &\propto P(Z)P(Y|Z) \\ &= P(Z)\prod_{i=1}^{N} P(Y_i|\max_{j:x_{i,j}=1} Z_j) \\ &= P(Z)\prod_{i=1}^{N} \left[\alpha^{y_i}(1-\alpha)^{1-y_i}\right]^{1-\max_{j:x_{i,j}=1} Z_j} \left[\gamma^{y_i}(1-\gamma)^{1-y_i}\right]^{\max_{j:x_{i,j}=1} Z_j} \end{aligned} \tag{3.6}$$

where $P(Y|Z)$ is from (3.2). In numerical examples, we consider both priors P_1 and P_2 and denote by SFCA1 and SFCA2 posterior inference using role model and these two priors respectively.

3.3.4 Extending the role model

The basic setting of the role model may be limited by its restriction to binary gene-level data and by an assumed homogeneity of responses within the activated and inactivated

classes. Though the focus of this thesis is still dealing with binomial data, we show that it could be conveniently extended in two different ways. The first extension is to allow extra-binomial variations in the observation component and the second is to address multinomial gene-level data.

Beta-binomial model

Currently in the role model, all inactivated states deliver conditionally independent responses with success probability $\gamma > \beta$. This constrains the atom level counts Y_i to be Binomially distributed given the activation states. A more flexible and simple extension within the general framework allows each gene to have its own Beta distributed success probability, then atom level counts y_i are more broadly distributed as Beta-binomial counts. For this extension, we only need to add one more parameter $c \in (0, +\infty)$ to control the variation of Beta distribution while remaining its mean at either α or γ give the true category activation states.

$$p_i = p^{\mathrm{BB}}(x_i) \sim \begin{cases} \mathrm{Beta}\left(c\alpha, c(1-\alpha)\right) & \text{if } \max_{j:x_{i,j}=1} Z_j = 0 \\ \mathrm{Beta}\left(c\gamma, c(1-\gamma)\right) & \text{if } \max_{j:x_{i,j}=1} Z_j = 1 \end{cases} \tag{3.7}$$

When $c = +\infty$, the Beta distribution reduces to constant and (3.7) is equivalent to (3.3); when $c \to 0$, $\mathrm{var}(Y_i)$ increases and reaches its maximum at $c = 0$. Posterior computations may benefit from flattening out of the posterior distribution over activation states.

Fitting multinomial data

When binary gene-level data from multiple studies are combined, we demonstrate how the role model is extended to utilize gene-level multinomial record for category level inferences. Take the meta-analysis data analyzed in Chapter 2 as an example, there are in total 81 patterns in the observational Multinomial distribution. For each atom i, record Y_i becomes a vector of length 81 counting the number of genes falling into each pattern. For each pattern m, there is a pair of parameter γ_m and α_m to calibrate the hot and cold delivery of success probabilities given the category activation states.

Let $Y_i = \{Y_{im}\}$, $\gamma = \{\gamma_m\}$, $\alpha = \{\alpha_m\}$, $m = 1, 2, ..., M$.

$$\begin{aligned} Y_i &\sim \text{Multinomial}(n_i, p_{i1}, p_{i2}, ..., p_{iM}) \\ p_{im} = p_m^{\text{Mul}}(x_i) &= \begin{cases} \alpha_m & \text{if } \max_{j:x_{i,j}=1} Z_j = 0 \\ \gamma_m & \text{if } \max_{j:x_{i,j}=1} Z_j = 1 \end{cases} \end{aligned} \tag{3.8}$$

3.4 Operating characteristics of posterior inference

3.4.1 Setup

Compared to MGSA, our proposed SFCA restricts computations to a smaller but highly constrained space. It requires more sophisticated computational methods to deploy posterior inference. An important question to ask is what would be the benefits from taking this extra effort? In this section, operational characteristics of both SFCA and MGSA

are compared via simulation studies in relatively simple systems enabling exact posterior computations. We find that for overlapping and hierarchical category structures (1) SFCA generates consistent posterior inference while MGSA fails in some situations; (2) SFCA is more efficient in detecting weak signals in the data; (3) SFCA provides useful inferences on joint category activation states.

First we consider two artificial examples that display properties relevant to more realistic scenarios. Example I presents a hierarchical structure where all categories have parent-child relationship with others. It is used to show that MGSA can be inconsistent in the sense that posterior distributions do not converge to true states. Example II resembles a system with highly overlapping categories, where SFCA detects weak signal more efficiently than MGSA. The third system is constructed with GO categories and represents overlapping and hierarchical categories more typical of practice. Similar operating characteristics are seen in this more realistic case.

In each example, true activation states $\{Z_{true}, A_{true}\}$ and parameters $\{\alpha, \gamma\}$ are fixed. Size of each atom n_i is gradually increased to a large number in each scenario. By the *Law of large numbers*, when $n_i \to +\infty, \forall i$,

$$\frac{Y_i}{n_i} \quad \to \quad \begin{cases} \alpha & \text{if } A_{true,i} = 0 \\ \gamma & \text{if } A_{true,i} = 1 \end{cases}$$

It means that the level of signal in the data increases accordingly and truly on/off atoms are to be identified. We will go over this point and provide proof at the end of this section. We set parameters $\alpha = 0.45$, and $\gamma = 0.55$ in most cases to represent the

situation where both false positive and false negative errors are as high as 0.45 and renders the true signal weak at small atom size. In the first two examples, atom sizes are equal and changed from 10 to 1000 in different scenarios. In the third example, true sizes of GO terms are amplified proportionally from 1 to 1000 times.

Data sets are simulated under each scenario by following the role model (3.3). First, activation state A_i of atom i is decided by the fixed category level states and the incidence matrix. Then, data point Y_i is sampled from Binomial distribution with size being size of the atom n_i and success probability being either γ if the atom is on or α if the atom is off. As the examples are stylized, we are able to calculate all prior and posterior probabilities numerically. We denote by SFCA1, SFCA2 and MGSA analysis under prior P_1, P_2 or P_0 respectively.

On choice of parameter π

Under P_0 prior probability for each category being active is π, while under P_1 and P_2 the prior probabilities are not constant over categories. In order to make MGSA and SFCA comparable, π needs to be carefully chosen for each prior. Denote by π_k the parameter value chosen for prior $P_k, k = 0, 1, 2$. In MGSA, π_0 is chosen to be the proportion of on categories in the true joint state Z_{true}, which is also the MLE of π knowing the true states. Following this idea, we could simply let $\pi_1 = \pi_0$ and π_2 equal the proportion of on atoms in A_{true}. However, they are not controlled under some common criteria to be comparable. Instead, the following adjustment is adopted. Note that given incidence

matrix X the mean of prior probabilities $P_k(Z_j = 1)$ is a function of only π_k. Let $F_k(\pi) = \frac{1}{C}\sum_{j=1}^{C} P_k(Z_j = 1), k = 0, 1, 2$. It is easy to see that F_0 of MGSA is an identity function, i.e. $F_0(\pi_0) = \pi_0$. For SFCA, π_k is chosen so that the mean of prior probabilities is also π_0, i.e.

$$\pi_k = F_k^{-1}(\pi_0) \tag{3.9}$$

Note that this is not the only way to choose π for priors, but it is useful to help compare the operating characteristics of posterior inference generated by them. We apply this method in all following examples.

Maximum active set

One direct conclusion from activation hypothesis is that if a category j is on, all its children i.e. categories that are proper subsets of j are on. If there is a large category being active in a joint state, presenting this category and all its children being on is just equivalent to knowing this large category is on. We call a category a *maximum active set* in a joint state if it is on and does not have any active parent. Any valid joint state with at least one category on has at least one maximum active set. Maximum active sets can be overlapping and altogether present information that can not be contained by any single category. If all the active categories are mutually exclusive then all of them are maximum active sets. Marginal probability of each category being a maximum active set is useful in addition to marginal probability of being active.

For small-scale examples, we can derive prior and posterior probability of each category being a maximum active set analytically. Let

$$\begin{aligned} W_j &= I(\text{category } j \text{ is a maximum active set}) \\ &= I(Z_j = 1 \text{ and category } j \text{ does not have any active parent category}) \end{aligned}$$

Then prior and posterior probability of each category being a maximum active set are:

$$\begin{aligned} P(W_j = 1) &= \sum_{s=1}^{S} W_j^s P(Z^s) \\ P(W_j = 1|Y) &= \sum_{s=1}^{S} W_j^s P(Z^s|Y) \end{aligned}$$

where W_j^s indicates if category j is a maximum active sets in joint state Z^s, and $S = |\mathcal{Z}|$.

Bayes factor

Due to overlapping and hierarchical structure, P_1 and P_2 in most cases are non-flat. Therefore, it is important to consider statistics that take into account both prior and posterior information when prioritizing categories, for example, the Bayes factor. Bayes factor of category j being active is as follows.

$$BF_{Z_j} = \frac{P(Z_j = 1|Y)}{P(Z_j = 1)} \tag{3.10}$$

P_0 is flat regardless of the category structure, so ranking by Bayes factor is the same as ranking by marginal posteriors. Similarly, Bayes factor of category j being a maximum active set is

$$BF_{W_j} = \frac{P(W_j = 1|Y)}{P(W_j = 1)}.$$

Performances of different priors are compared in terms of the statistics introduced above.

3.4.2 Posterior consistency

Example I (Table 8) shows an incidence matrix that represents a hierarchical structure. There are 15 categories each containing a subset of the 4 atoms. Category 15 is the largest and also a parent of the rest. Due of the hierarchical structure, only 16 out of 2^{15} combinations of category activation states respect activation hypothesis, i.e. $S = |\mathcal{Z}| = |\mathcal{A}| = 16$ while $|\mathcal{Z}_0| = 2^{15}$. Consider true category level joint state be $Z_{true} = (\underbrace{1,...,1}_{7},\underbrace{0,...,0}_{8})$ and corresponding atom level $A_{true} = (1,1,1,0)$. Essentially, the first 7 categories activate the first 3 atoms.

Table 8: Incidence matrix representing a hierarchical category structure in Example I.

atom/category	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
2	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
3	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
4	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1

Figure 14 shows functions $F_k, k = 0, 1, 2$ with settings of Example I. Given Z_{true}, π_0 is estimated as $7/15 \approx 0.47$. π_1 and π_2 are solved from equations $F_k(\pi) = \pi_0, k = 1, 2$. In order to keep the mean prior probabilities at the same level, π_1 is slightly boosted from 0.47 to 0.54; and π_2 is pressed from the proportion of on atoms which is $3/4 = 0.75$ to 0.68.

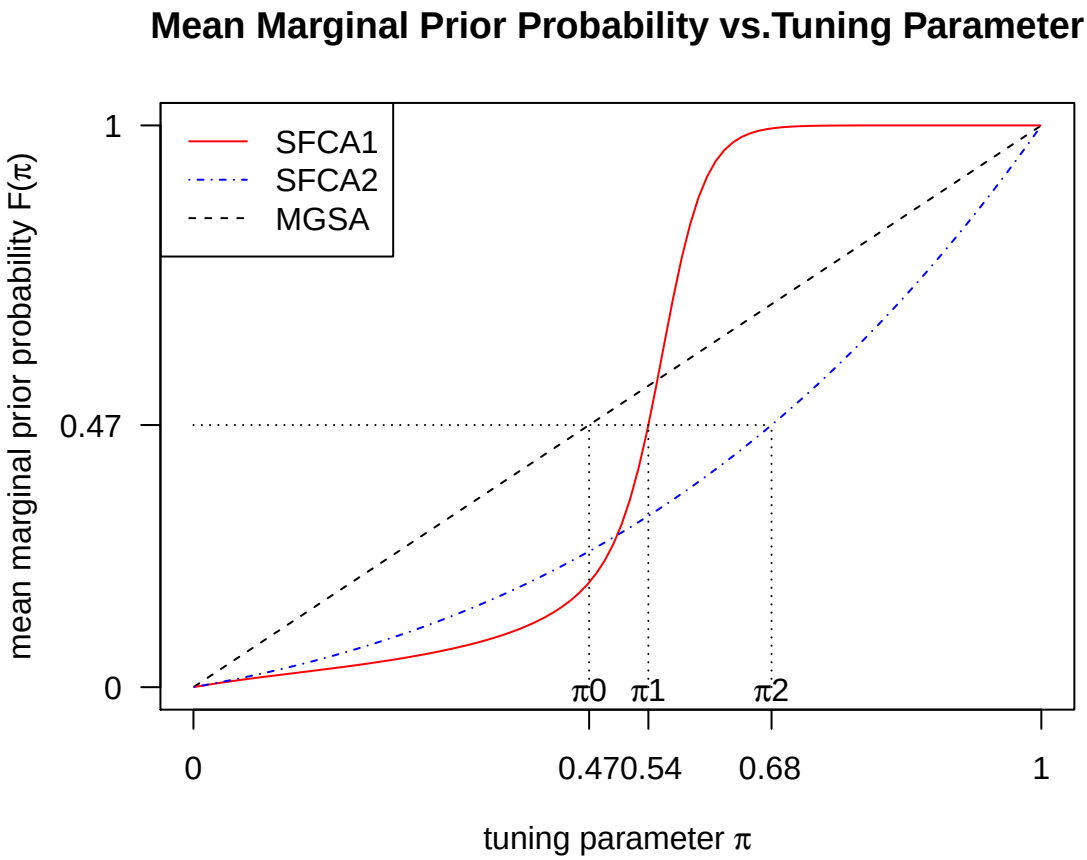


Figure 14: Example I: Mean prior probability $F_k(\pi)$ as a function of π. π_1 and π_2 solved to achieve mean prior probability at π_0.

Consistency in posterior probability of each category being active

With π taking the chosen value as described above, the other two parameters are fixed at $\alpha = 0.45$, and $\gamma = 0.55$ to represent high level of noise. We change the size of atom n_i from 10, 100 to 1000 to increase the level of signal in the data and compare performances of methods at each signal level. At each n_i, 1000 data sets are simulated. Marginal posterior probability $P(Z_j = 1|Y)$ are calculated under all 3 priors.

Marginal posterior calculations for all 3 methods are illustrated in Figure 15. Atom size increases from 10 to 1000 from top to bottom panel. Red blocks represent 'signal', posterior probabilities of the 7 truly on categories being active, and green blocks represent 'noise', posterior probabilities that falsely call the 8 truly off categories to be active. We see that when atom size is 10 and 100, SFCA1 and SFCA2 always perform better than MGSA to separate signal and noise. When atom size becomes 1000, noise converges to 0 in all methods but signal converges to 1 only in SFCA methods but not MGSA. It suggests that MGSA estimator for posterior probability $P(Z_j = 1|Y)$ is not consistent. Table 9 lists mean posterior probability for the 7 truly on categories from 1000 simulations and atom size is 1000. The inconsistency is due to the hierarchical structure of these categories as illustrated in Figure 16. Recall that $A_{true} = (1, 1, 1, 0)$. Without activation hypothesis there are multiple invalid configurations of joint states on $\mathcal{Z}_0$ that activate the first 3 atoms, for example, any joint state with category 7 and one of its children on. With all these invalid states sparing probabilities, MGSA fails in identifying truly on categories. We find that with truly active categories forming

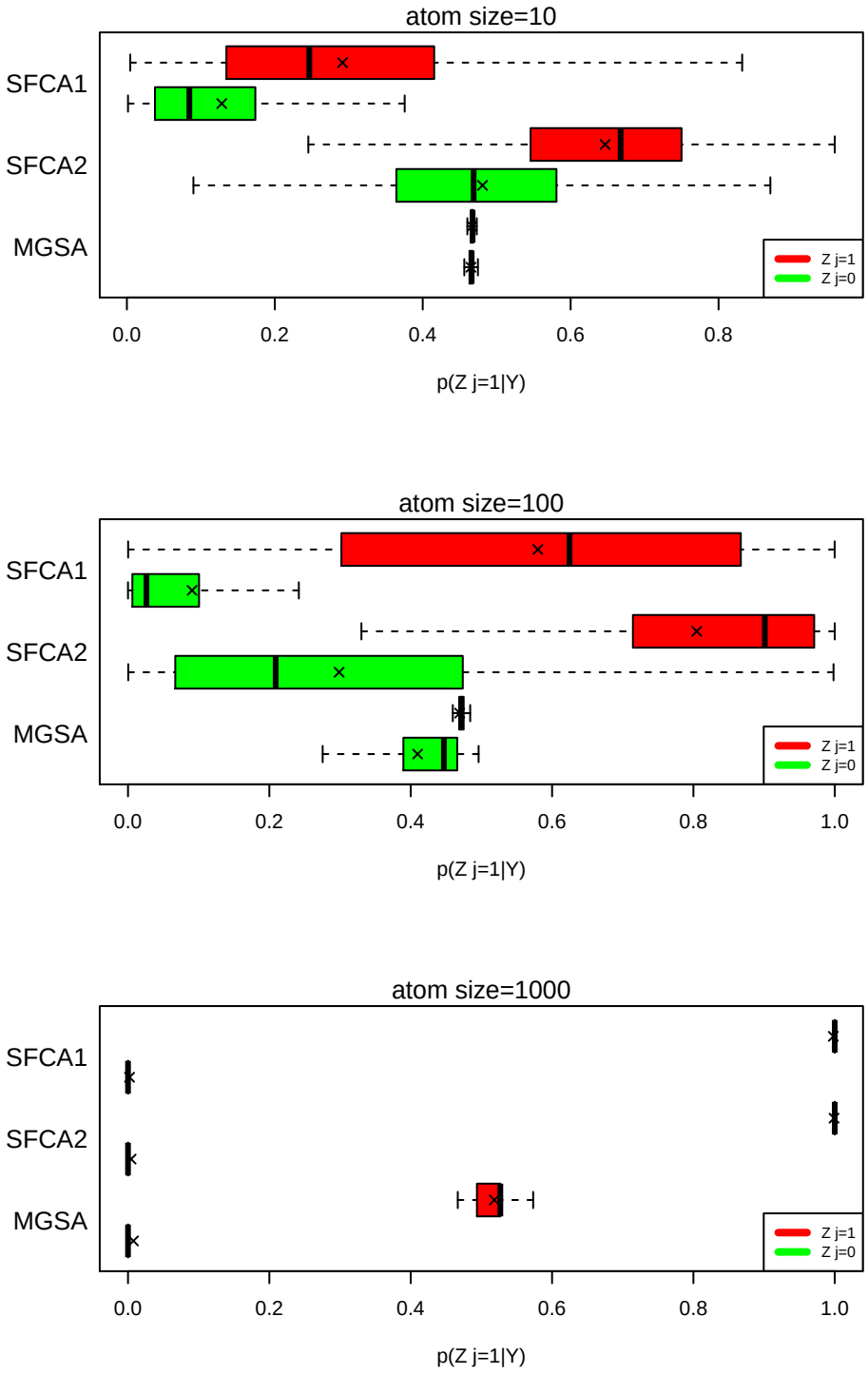


Figure 15: Example I: Box plot of posterior probability of each category being active by atom size and method. Red blocks represent truly on categories and green blocks represent off categories. "x" marks mean of the distribution.

parent-child relationship with other sets, MGSA is not consistent in estimating posterior probability $P(Z_j = 1|Y)$. Our demonstration is based on numerical evaluation of posterior inference on simulated data in a very simple system. However, further analysis supports a general claim.

Table 9: Mean posterior probability of the 7 truly on categories being active from 1000 simulations of Example I. Each atom is of size 1000.

category	# of atoms	mean posterior		
		SFCA1	SFCA2	MGSA
1	1	1.00	1.00	0.49
2	1	1.00	1.00	0.49
3	2	1.00	1.00	0.53
4	1	1.00	1.00	0.49
5	2	1.00	1.00	0.53
6	2	1.00	1.00	0.53
7	3	1.00	1.00	0.57

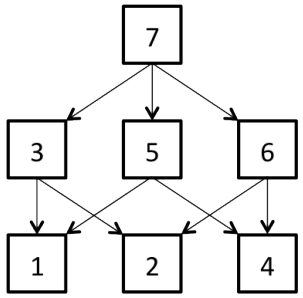


Figure 16: Hierarchical structure of category 1-7 of Example I. Arrows start from parent categories to direct children categories (no children in between).

Consistency in posterior probability of each category being a maximum active set

According to definition of maximum active set, in Z_{true} of Example I, category 7 is the only maximum active set since its only parent category 16 is inactive. Figure 17 shows box plot of posterior of each category being a maximum active set from 1000 simulations.

The simulation results unsurprisingly show that MGSA fails to identify category 7 as the only maximum active set with posterior of 1 when atom size is big enough. In Figure 17, $P(W_7 = 1|Y) = 1$ in the top two panels representing SFCA methods, while in the bottom penal $P(W_7 = 1|Y) = 0.57$ and stabilizes at this value when atom size is increased further (not shown when atom size increases). Also, $P(W_j = 1|Y)$ is supposed to be 0 for all other categories, which is violated by MGSA. Thus, with existence of parent-child relationship between categories MGSA might not able to provide consistent inference to posterior probability of each category being a maximum active set.

Proof of consistency

In the end, we provide a simple proof of consistency for SFCA and explain why it fails in MGSA. Given true category activation states $Z_{true} \in \mathcal{Z}$, it maps to a unique true atom activation states $A_{true} \in \mathcal{A}$ by Proposition 3.1. Suppose parameters $\{\alpha, \gamma\}$ are known. The role model (3.3) can be written as

$$p_i = \alpha(1 - A_{true,i}) + \gamma A_{true,i}.$$

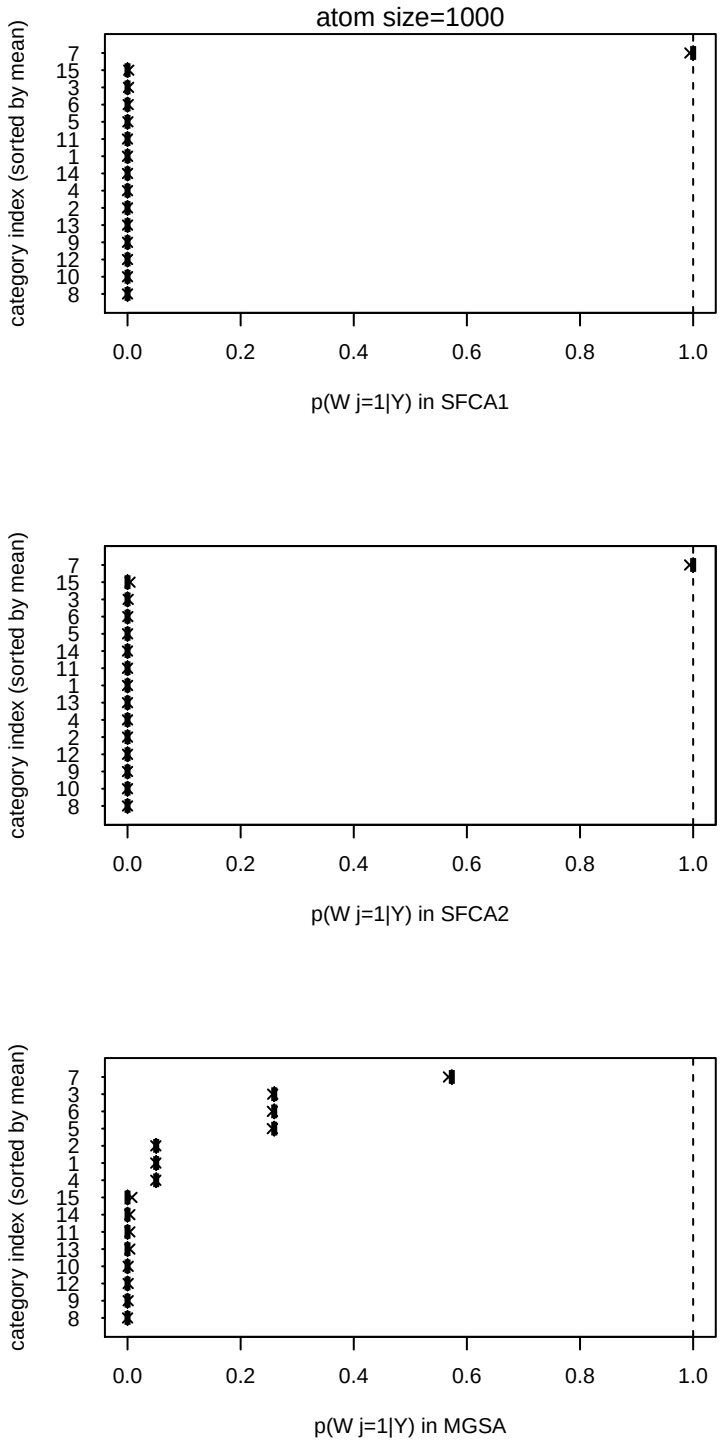


Figure 17: Box plot of posterior of each category being a maximum active set from 1000 simulations in Example I. Each atom is of size 1000. Each panel represents a method. "x" marks the mean of the distribution.

Since for $\forall i, Y_i/n_i$ is the MLE of p_i, $\frac{Y_i/n_i - \alpha}{\gamma - \alpha}$ is the MLE and also an consistent estimator of $A_{true,i}$. By finiteness of the system of atoms, we become confident in the true atom level activations under all priors $P_k, k = 0, 1, 2$,

$$P_k\left(A = A_{true}|Y\right) \to 1, \text{ as } \min(n_i) \to +\infty$$

where $Y = \{Y_i\}$, by posterior consistency in finite spaces (Schervish (1997)). For SFCA, because of the one-to-one mapping between category and atom level activation states, we also have $k = 1, 2$,

$$P_k\left(Z = Z_{true}|Y\right) \to 1, \text{ as } \min(n_i) \to +\infty$$

However, for MGSA there maybe no unique inverse of A_{true}, depending on the particular overlapping structure of categories and the true state. In such cases, denote by $\mathcal{Z}^*$ the set of joint category level joint states that A_{true} maps to,

$$\mathcal{Z}^* = \left\{ Z = (Z_1, Z_2, ..., Z_C) \in \mathcal{Z}_0 : Z_j = \min_{i:x_{i,j}=1} A_{true,i} \right\} \subset \mathcal{Z}_0$$

then $Z_{true} \in \mathcal{Z}^*$, but there are also other $Z \in \mathcal{Z}^*$ and $Z \neq Z_{true}$. For MGSA,

$$P_0\left(Z \in \mathcal{Z}^*|Y\right) \to 1, \text{ as } \min(n_i) \to +\infty.$$

but different states in $\mathcal{Z}^*$ cannot be distinguished. Curiously, this carries over to marginal posterior summaries on activated categories $\{j\}$ in such a way that for some sets j,

$$P_0\left(Z_j = 1|Y\right) \to \phi, \text{ as } \min(n_i) \to +\infty$$

where $\phi \in (0, 1)$.

3.4.3 Efficiency

Other than consistency, SFCA is more efficient in detecting subtle signals in the data compared to MGSA. We demonstrate this point in the case when consistency issue does not occur. Thus, at a fairly small atom size, method that performs better in separating signals from noise is claimed to be more efficient.

Table 10 presents incidence matrix for Example II. There are 11 atoms constituting 7 highly overlapping categories. The relatively high rate of pairwise overlap among categories is intended to model redundancies in GO. Only 30 out of $2^7 = 128$ category-level joint states are valid, i.e. $S = |\mathcal{Z}| = 30$. Category 1 and 7 are set to be truly on and activate all atoms except for atom 6, i.e. $Z_{true} = (1, \underbrace{0, ..., 0}_{5}, 1)$ and $A_{true} = (\underbrace{1, ..., 1}_{5}, 0, \underbrace{1, ..., 1}_{5})$.

Table 10: Incidence matrix of a highly overlapping category structure of Example II.

atom/category	1	2	3	4	5	6	7
1	0	0	0	0	0	0	1
2	0	0	0	0	0	1	1
3	0	0	0	0	1	1	1
4	0	0	0	1	1	1	1
5	0	0	1	1	1	1	1
6	0	1	1	1	1	1	0
7	1	1	1	1	1	0	0
8	1	1	1	1	0	0	0
9	1	1	1	0	0	0	0
10	1	1	0	0	0	0	0
11	1	0	0	0	0	0	0

Prior probability of each category being active with π at chosen values are illustrated

in Figure 18. Due to the structure, category 4 which has the most overlap with other categories has the highest prior probability, while active category 1 and 7 have the smallest priors. We know that prior is most influential to posterior inference when true signal in the data is vague compared to noise. This effect of prior becomes weak when signal gets stronger. Thus, it is useful to compare Bayes factor which accounts for both prior and posterior.

To provide a thorough comparison, we show box plots of both posterior and Bayes factor at different atom size in Figure 19. The left column illustrates posterior of each category being active by method and at three different atom sizes. On the right are corresponding plots for Bayes factors. If a method provides consistent inference of marginal posteriors, when the atom size is big enough, we should observe (1) on the left the red and green bar converges to 1 and 0 respectively, (2) on the right the blue bar converges to 0 and the orange bar converges to some positive value ($1/P(Z_1 = 1)$). We see that it is confirmed by the two plots at bottom, which means all three priors compared in this setting are consistent.

The top two plots represent the case where signal is weakest. As expected, in both SFCA methods truly on categories have smaller mean posterior probabilities than truly off categories due to the structure, but Bayes factor is able to recover the true ordering. Mean of all Bayes factors are above 1, meaning averaged posteriors are boosted from corresponding priors. In general all priors perform similarly, and SFCA1 shows a slightly wider gap between mean Bayes factors of on and off categories.

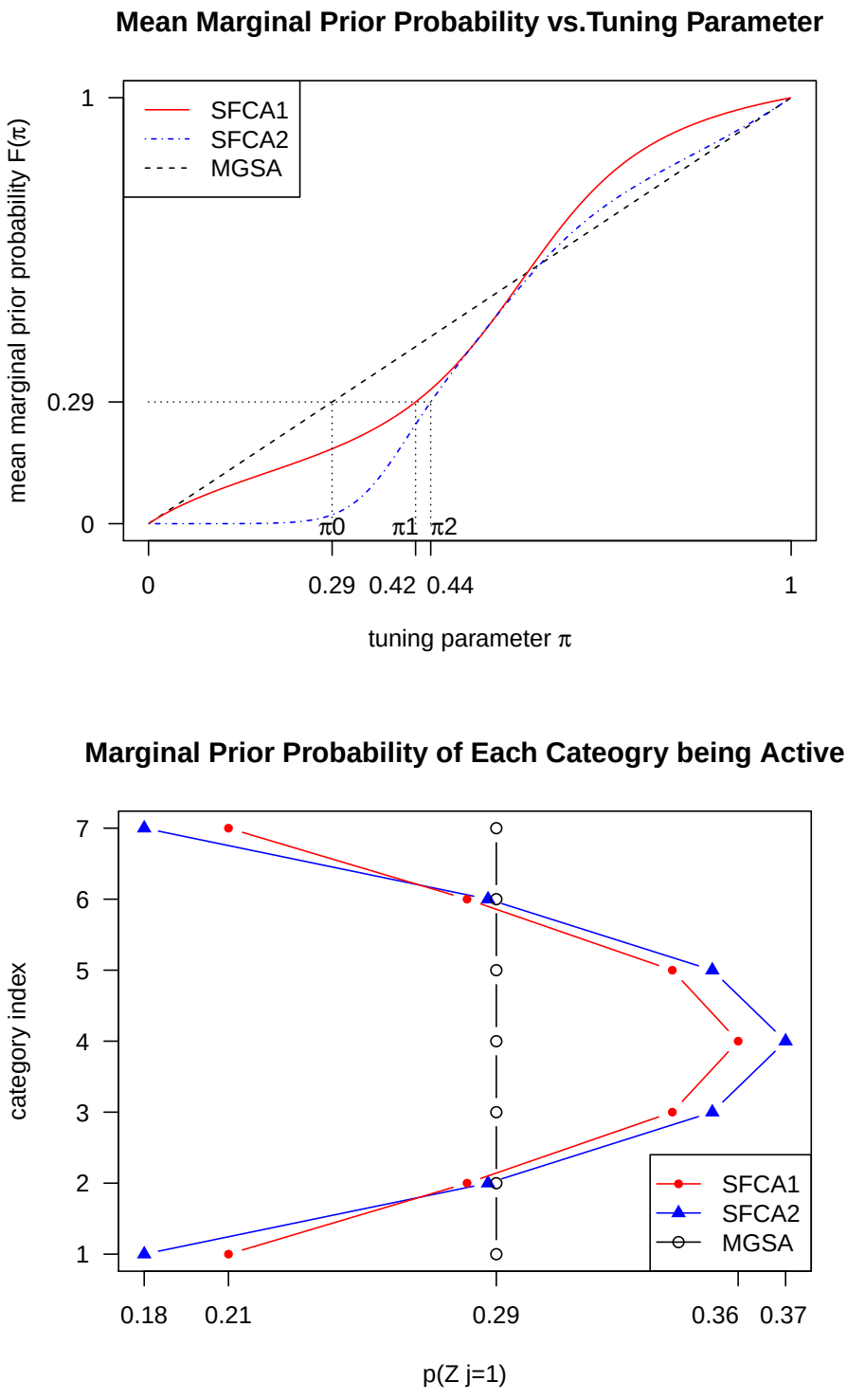


Figure 18: Example II. Upper: mean prior probability $F_k(\pi)$ as a function of π. π_1 and π_2 solved to achieve mean prior probability at π_0. Lower: prior probability of each category being active with π at chosen values.

The efficiency gain of SFCA1 compared to MGSA is demonstrated in the middle panels where atom size increases. Note that despite the small prior probabilities of the truly on categories, SFCA1 performs best in separating mean posterior probabilities and Bayes factors of the on and off groups. Also, both mean and median Bayes factor of truly off categories start to drop below 1 only in SFCA1. When the atom size is 100, SFCA1 shows absolute advantages in differentiating signal from noise and SFCA2 starts to perform better than MGSA.

3.4.4 Inference on joint activation states

In Bauer et al. (2010), it is suggested that 0.5 be used as a cutoff value. Categories with marginal posterior probability higher than 0.5 are selected as final results, meaning they have more chance to be on than off. A question to ask is that whether a list of categories with highest marginal posteriors is the best combination to explain the data. We will show in this section that relying on only marginal posterior inference has limitations. For this reason, inference on joint activation states is very useful in addition to marginal inference.

Incidence matrix of Example II is used again here with artificial data in the last two columns of Table 11. n_i is the size of atom i, and data point y_i is the number of genes observed as active. The bottom two rows give a category level summary on size and number of active genes. Suppose we don't know the true activation states and study

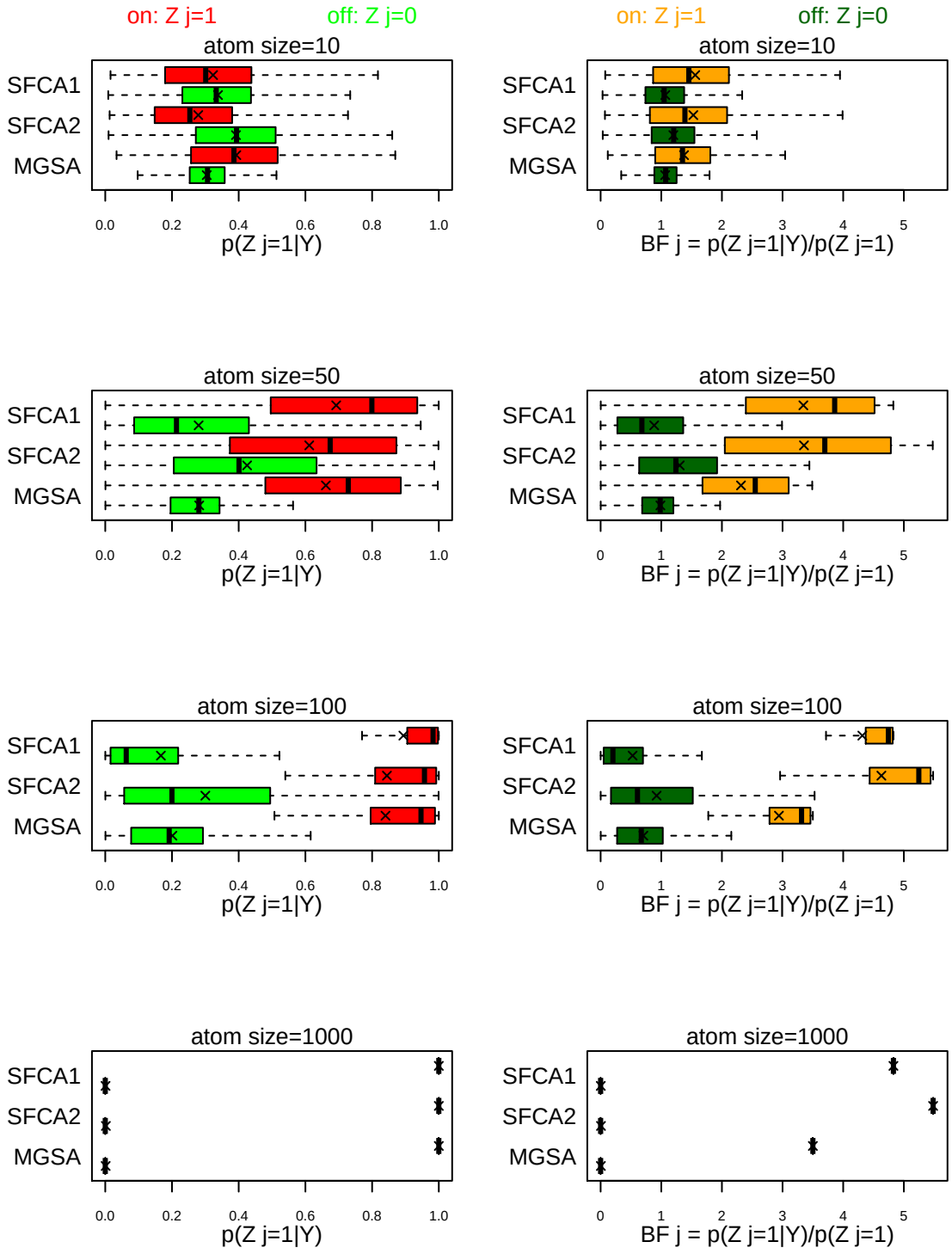


Figure 19: Example II. Left column: box plots of posterior of each category being active by method, at different atom sizes. Right column: box plots of marginal Bayes factor by method, at different atom sizes. Red and orange bars represent categories 1 and 7 which are truly active.

this example with the following parameters to roughly match the observations.

$$\alpha = 0.5, \gamma = 0.75;$$

$$\pi_0 = \pi_1 = \pi_2 = 0.5.$$

Table 11: Example II: incidence matrix and basic data.

atom/category	1	2	3	4	5	6	7	y_i	n_i
1	0	0	0	0	0	0	1	1	2
2	0	0	0	0	0	1	1	1	2
3	0	0	0	0	1	1	1	1	2
4	0	0	0	1	1	1	1	2	2
5	0	0	1	1	1	1	1	2	2
6	0	1	1	1	1	1	0	2	2
7	1	1	1	1	1	0	0	2	2
8	1	1	1	1	0	0	0	2	2
9	1	1	1	0	0	0	0	1	2
10	1	1	0	0	0	0	0	1	2
11	1	0	0	0	0	0	0	1	2
# active genes	7	8	9	10	9	8	7		
size	10	10	10	10	10	10	10		

Note that when $\pi_1 = \pi_2 = 0.5$, SFCA1 and SFCA2 are equivalent. Marginal prior and posterior probabilities are summarized in Table 12. From posterior estimations we see that three categories {3,4,5} would be named on a short list of categories targeting no more than 50% posterior false discovery rate. In fact, the gene-level activity data are well explained using only the activation of category 4. Actually the joint state with only $Z_4 = 1$ is the maximum a posteriori (MAP) estimate of the joint state as shown in Table 13. Each joint state is presented as a vector of individual category states. Out of 30 valid joint states, we list the top 10 with highest marginal posterior probabilities (as

well as Bayes factors) in SFCA. We know that the MAP estimate is the Bayes estimate under 0 -1 loss, while a Hamming-loss delivers the estimate {3,4,5} (e.g., Carvalho and Lawrence (2008)). It is not a major issue to decide which one is better, but having access to all sorts of posterior summaries will surely give us a better understanding of the high-dimensional parameter space. With one-to-one correspondence between atom and category level joint states, SFCA is able to generate meaningful inferences on category level joint states from atom level input, which is not available from MGSA.

Table 12: Example II: Marginal prior and posterior probability of each category being active at fixed parameters $\alpha = 0.5, \gamma = 0.75, \pi_0 = \pi_1 = 0.5$.

		MGSA		SFCA	
category	# on genes/size	$P_0(Z_j = 1)$	$P_0(Z_j = 1\|Y)$	$P_1(Z_j = 1)$	$P_1(Z_j = 1\|Y)$
1	0.7	0.5	0.39	0.27	0.16
2	0.8	0.5	0.46	0.40	0.35
3	0.9	0.5	0.52	0.50	0.58
4	1.0	0.5	0.58	0.53	0.78
5	0.9	0.5	0.52	0.50	0.58
6	0.8	0.5	0.46	0.40	0.35
7	0.7	0.5	0.39	0.27	0.16

3.4.5 An example in GO

Last but not least let us look at a more realistic example. GO presents a complex structure that makes it difficult to detect subtle signals. In this part, Example III is constructed by GO categories of interest and represents a typical overlapping and hierarchical GO structure. MGSA and both SFCA priors are applied to analyze simulation

Table 13: Example II: prior and posterior probability of joint states with highest posteriors in SFCA at fixed parameters $\alpha = 0.5, \gamma = 0.75, \pi_0 = \pi_1 = 0.5$.

rank	joint state	prior	posterior
1	0001000	0.033	0.104
2	0011000	0.033	0.078
3	0001100	0.033	0.078
4	0111000	0.033	0.059
5	0011100	0.033	0.059
6	0001110	0.033	0.059
7	1111000	0.033	0.044
8	0111100	0.033	0.044
9	0011110	0.033	0.044
10	0001111	0.033	0.044

data sets and their performances are compared in terms of consistency, efficiency and sufficiency in recovering the truth.

Category structure

In the meta-analysis presented in Chapter 2, there are 614 genes jointly confirmed by 4 studies of interest. By applying one-at-a-time gene set enrichment analysis method *allez* (Newton et al. (2007)), 19 GO terms are reported (Hao et al. (2012)) as most enriched (p value $< 10^{-6}$) with the confirmed genes. Annotation profiles of these 19 GO terms are extracted from Bioconductor database org.Hs.eg.db (version 2.7.1 up to Sep 9, 2012) and collapsed to create an atom level incidence matrix to be used by our analysis. Identical columns indicating the same atom content are collapsed to one. In the end, there are 17 categories left in the system containing 28 atoms. When dealing with overlapping structure, one-at-a-time category analysis methods tend to select a list of

correlated categories with related functions, which is the case in this example. GO terms listed in Table 14 show that they are involved in majorly two functions: intracellular transportation and endocytosis. The GO are accordingly separated into two mutually exclusive groups A and B, each relates to only one of the two functions. As illustrated in Figure 20, hierarchical structures are formed within each group with arrows pointing from parent categories to their direct children categories. Category 11 does not have any children or parent in this system but overlaps with every other category in group A. It is connected to the bottom category in each hierarchy by a dashed line.

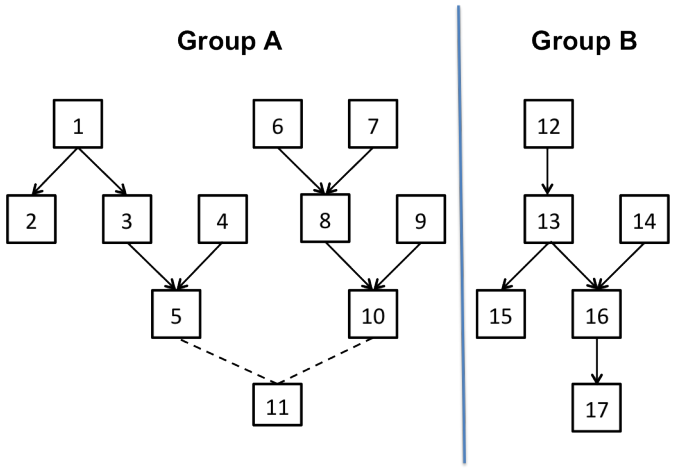


Figure 20: Hierarchical and overlapping category structure formed by 17 GO terms (labeled by indices in Table 14) in Example III. Arrows point from parent categories to direct children categories. Dashed lines connect overlapping categories. Categories of group A all are involved in intracellular transportation and group B is related to endocytosis. They are mutually exclusive.

Table 14: Information on 17 GO terms studied in Example III.

Index	ID	term	# atoms	# genes	group
1	GO:0006900	membrane budding	11	32	A
2	GO:0006901	vesicle coating	10	31	A
3	GO:0048194	Golgi vesicle budding	7	14	
4	GO:0048199	vesicle targeting, to, from or within Golgi	9	25	A
5	GO:0048200	Golgi transport vesicle coating	6	13	A
6	GO:0030137	COPI-coated vesicle	7	20	A
7	GO:0030660	Golgi-associated vesicle membrane	8	35	A
8	GO:0030663	COPI coated vesicle membrane	5	16	A
9	GO:0030120	vesicle coat	7	42	A
10	GO:0030126	COPI vesicle coat	3	14	A
11	GO:0006890	retrograde vesicle-mediated transport, Golgi to ER	6	24	A
12	GO:0006818	hydrogen transport	7	79	B
13	GO:0015992	proton transport	6	77	B
14	GO:0016469	proton-transporting two-sector ATPase complex	5	45	B
15	GO:0015985	energy coupled proton transport, down electrochemical gradient	2	18	B
16	GO:0033176	proton-transporting V-type ATPase complex	2	21	B
17	GO:0033179	proton-transporting V-type ATPase, V0 domain	1	6	B

Simulation studies

Similarly to previous examples we fix $\alpha = 0.45$ and $\gamma = 0.55$ to set high false positive and false negative errors in the system. An amplification index a is adopted to proportionally increase atom size and accordingly the level of signal in the data. For example when $a = 1$ atoms are at their original sizes; when $a = 1000$, all atom sizes are increased by 1000 times. At each index a, 500 data sets are simulated and analyzed by SFCA and MGSA. Two scenarios are designed. The first one demonstrates MGSA's inconsistency in posterior inference. The second case is to show SFCA's efficiency in detecting subtle signals compared to MGSA. Instead of analytically developed calculations we use MGSA's R package (Bauer et al. (2011)) to generate its results. By default MGSA samples 5 chains of length 10^6, and reports the average marginal posterior probabilities over 5 chains as the final. We find that results from MGSA's R program are always very close to analytical calculations (marginal deviance in this example is controlled under 10^{-3}).

Categories 6 and 7 are set to be active in the first scenario. According to activation hypothesis, their children sets 8 and 10 are also active. For MGSA, $\pi_0 = 4/17 = 0.24$. By setting mean prior probability of categories being active to 0.24, we solve that $\pi_1 = 0.33$ and $\pi_2 = 0.41$. Both categories 6 and 7 are maximum active sets of the true joint state. Figure 21 illustrates marginal posterior calculations when atom size is amplified by 1000 times. We see that MGSA is able to identify only categories 6 and 7 as active with marginal posterior probability 1, but misses the two active children sets 8 and 10. It

confirms that with truly active categories involved in parent-child relationships, MGSA suffers from identifiability issue and generates inconsistent posterior inference.

Another interesting finding is that when signal is weak MGSA might point to irrelevant category as being important. Let atoms stay at their original sizes to represent the case with weak signal in the data. Figure 22 shows box plots of Bayes factors for categories being a maximum active set from 500 simulations. Each panel represents results from one prior. Because the prior for each category to be a maximum active set is not flat due to the structure, we rank categories by averaged Bayes factor $P(W_j = 1|Y)/P(W_j = 1)$ over simulated data sets. We see that both SFCA1 and SFCA2 are able to rank categories 6 and 7 at the top with barely any other category's Bayes factor exceeding 1. For MGSA, category 7 is at the second place. Categories 15, 16 and 17 which belong to group B are ranked higher than category 6. Since categories in group A and B are mutually exclusive and each gourd is related to a different function, prioritizing irrelevant categories as important might be even worse.

The second scenario only has one truly active category which is category 11, and it is also the only maximum active set. Parameters π for each prior are $\pi_0 = 0.06$, $\pi_1 = 0.14$, and $\pi_2 = 0.31$. Although parent-child relationship exists in the system, under this structure MGSA is able to identify category 11 being active and also the only maximum active set as shown in the bottom panel of Figure 23. With consistency of posterior inference guaranteed, we are able to compare efficiency of methods in detecting signal from noise. Starting from the top panel, SFCA1 converges fastest to the truth (bottom

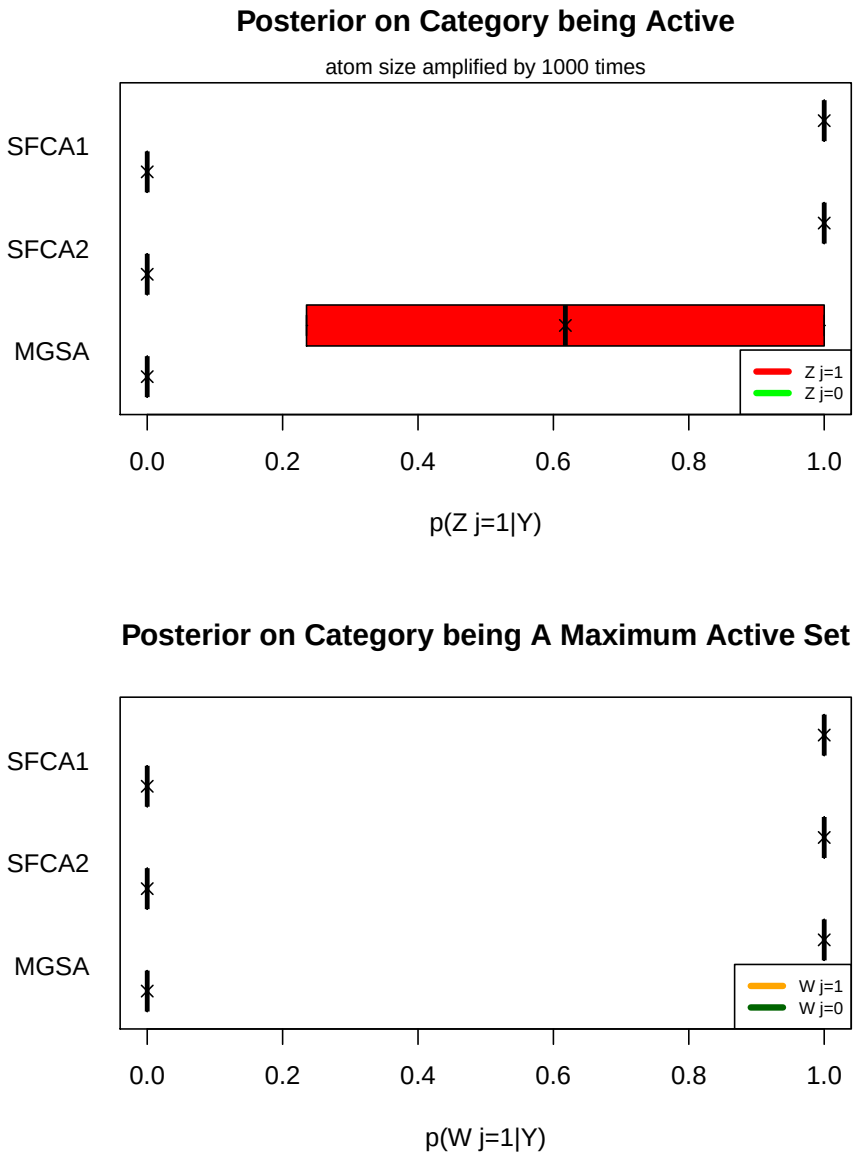


Figure 21: Marginal posterior probabilities by true state and method. Upper: probability on each category being active. Lower: probability on each category being a maximum active set. Red and orange bars all represent categories 6 and 7.

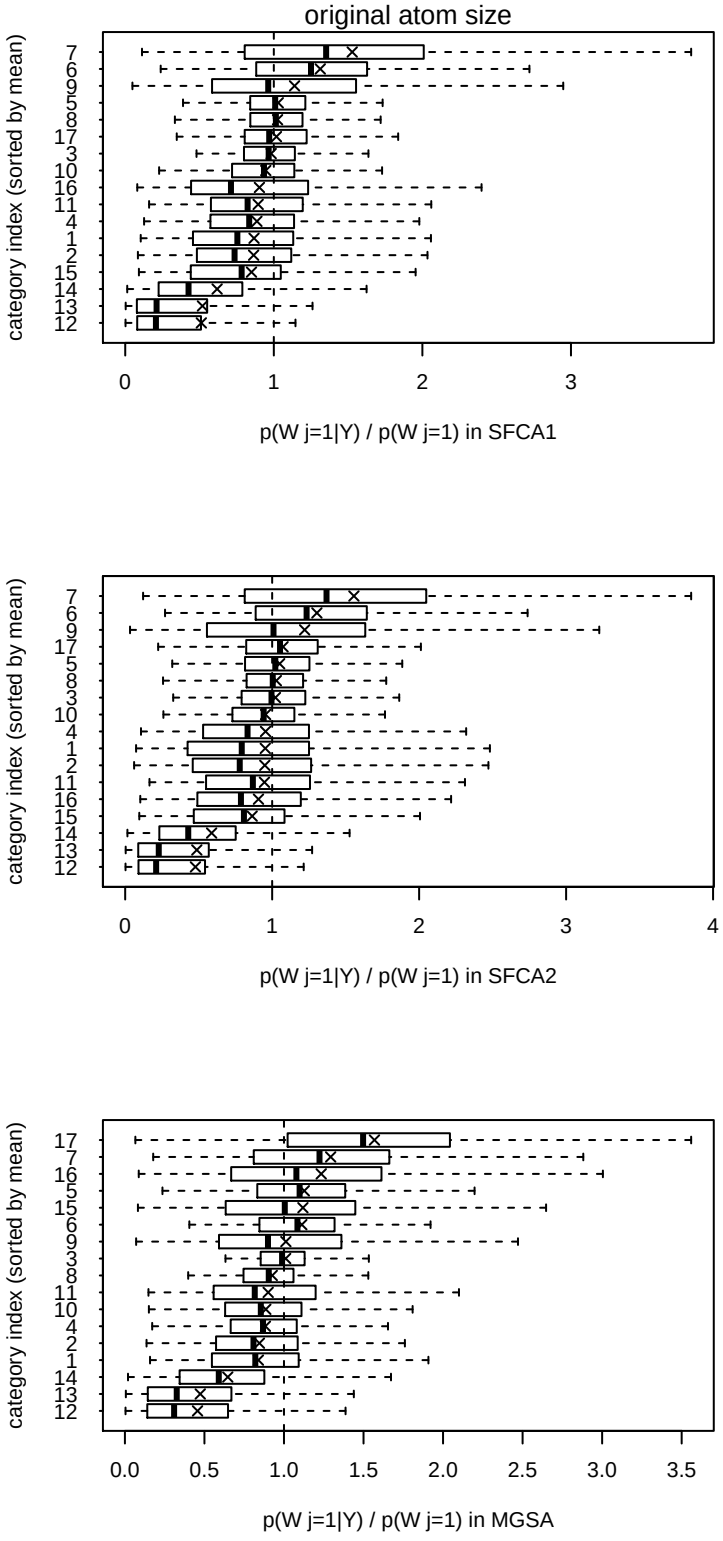


Figure 22: Bayes factors of categories being a maximum active set from 500 simulations when signal of data is weak. Each atom is of its original size.

panel) in terms of both marginal posterior probability and Bayes factor. SFCA2 does not perform well in this case. Posteriors of both truly on and off groups converge slower to their limiting probabilities compared to SFCA1 and MGSA.

In the end let us compare characteristics of prior calibrations of SFCA1 and SFCA2 indicated by this example. Figure 24 illustrates prior probabilities of the 17 categories in each method. Compared to MGSA's flat distribution, both SFCA methods separate priors in two distinct groups. The average prior probabilities of all three methods equal 0.06 here. For SFCA1, categories $\{5, 10, 11, 15, 17\}$ have priors above average. Figure 20 shows that they are the ones at bottom of each hierarchy (categories 5, 10, 15, 17) or does not belong to any hierarchy but overlaps most of its content with bottom categories of hierarchies (category 11). Sitting at the bottom of a hierarchy means this category certainly has fewer atoms than its parents but not necessarily compared to others. For instance, category 5 has 6 atoms which ranks in the middle of the 17 categories. Since one direct conclusion of activation hypothesis is that if a category is on then all its children are on, the lower level a category is located in a hierarchy the more likely it is active *a priori*.

SFCA2 assigns higher prior probabilities to categories $\{10, 15, 16, 17\}$ which have the smallest numbers of atoms (Table 14). Each atom is a group of genes sharing a common annotation profile, the more categories an atom is annotated to the less specific this atom is in terms of functions. Activation hypothesis can also be interpreted in terms of atoms: if an atom is on and denote by $\mathcal{C}$ all categories it is annotated to, then less

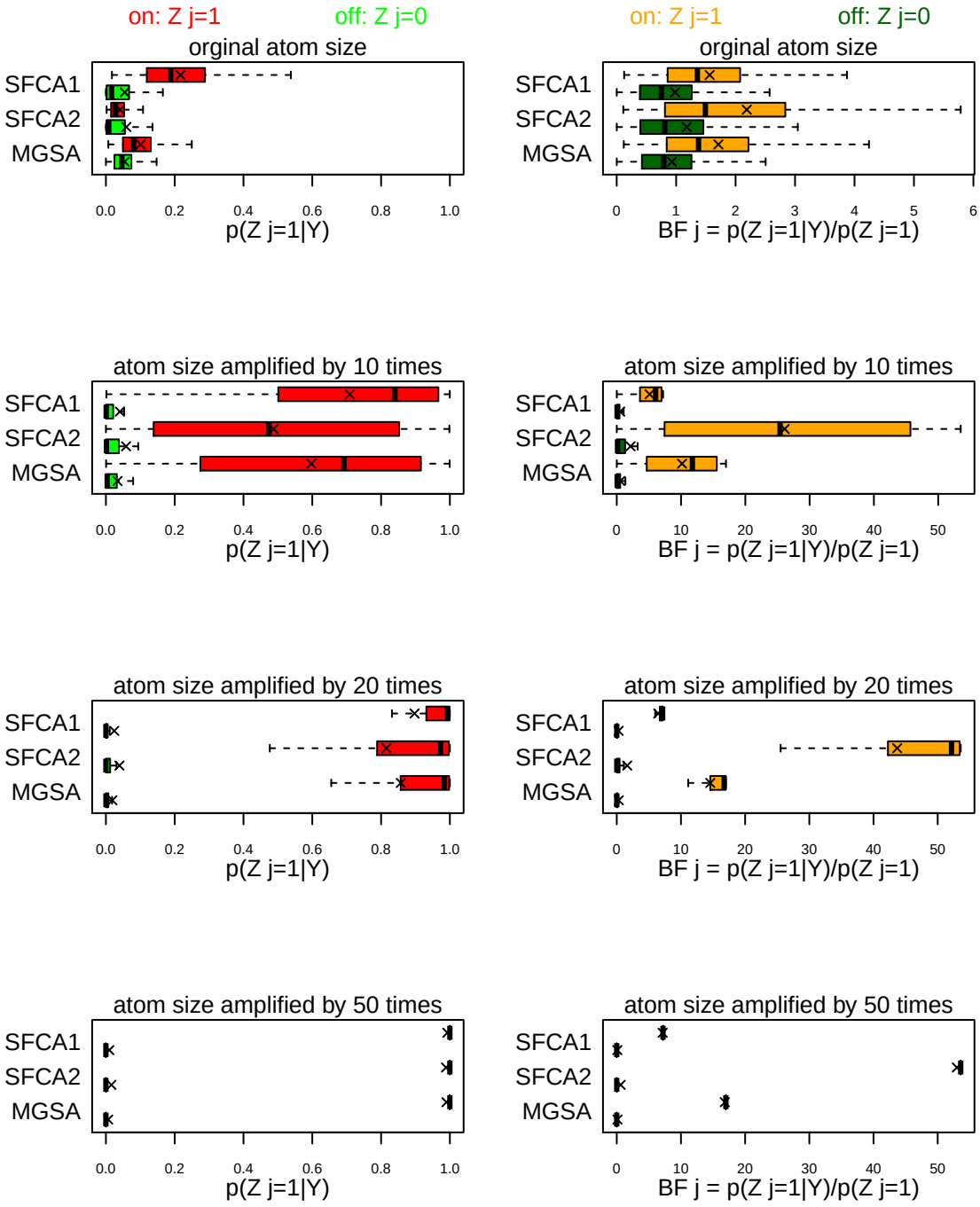


Figure 23: Example III. Left column: box plots of posterior of each category being active by method. Right column: box plots of marginal Bayes factor by method. Red and orange bars represent category 11 which is the only truly active set. Atom size is proportionally amplified by up to 50 times.

specific atoms that are also annotated to $\mathcal{C}$ are on. It suggests that another hierarchical system can be constructed to illustrate atom activities. In such a hierarchical graph, each node is an atom and a *parent atom* is annotated to more categories and less specific in functions than its children. This perspective is discussed with more details in Newton et al. (2012) and skipped here.

The purpose of listing both interpretations of activation hypothesis is that they relate to respectively SFCA1 and SFCA2 . We are not at a position to conclude which prior is better, because they are literally equivalent and the differences come from structure of specific examples and the choice of parameter π. Since P_1 used for SFCA1 has more straightforward definition and is similar to MGSA's prior P_0, in further demonstrations we present results from only SFCA1 to compare with MGSA.

3.5 Computation via Markov Chain Monte Carlo method

3.5.1 General description

Category analysis is developed eventually to deal with large scale genome-wide problems where posterior inferences can not be derived analytically. MGSA (Bauer et al. (2010)) has developed a Metropolis-Hasting algorithm for marginal posterior inferences. However, this method is not able deal with the activation hypothesis and generates only

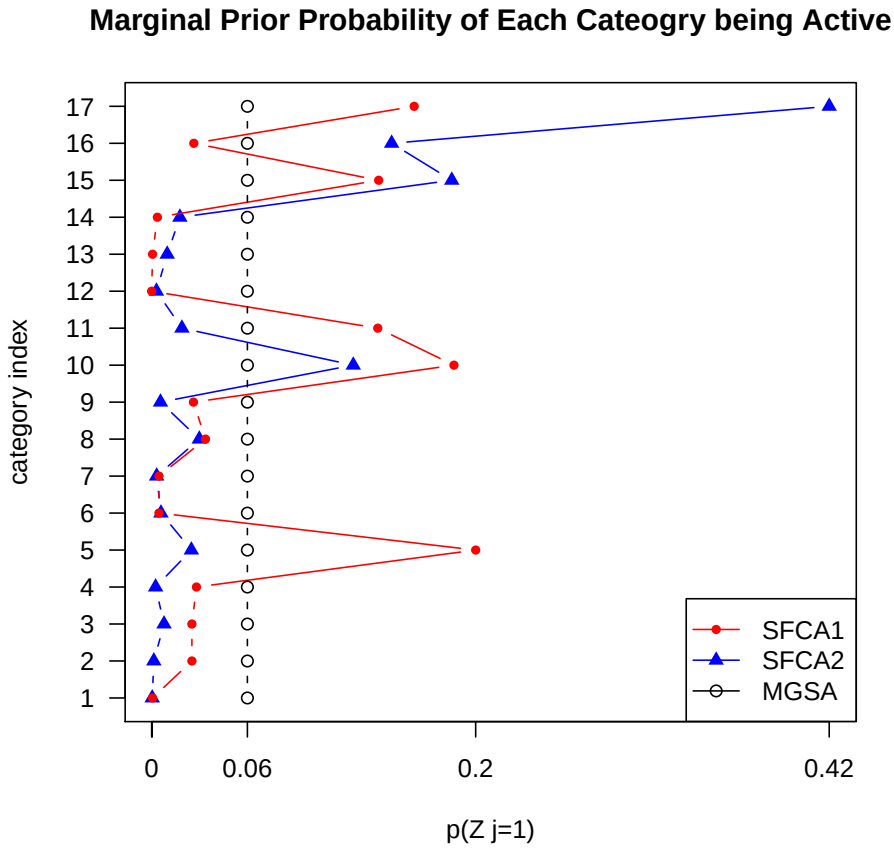


Figure 24: Marginal prior probability of each category being active with $\pi_0 = 0.06$, $\pi_1 = 0.14$, and $\pi_2 = 0.31$. Category 11 is set to be the only active set.

marginal posterior summaries. Thus, it is our task to develop a computational scheme to account for activation hypothesis and generate different kinds of posteriors summaries.

We also adopt Metropolis-Hasting (M-H) algorithm to draw samples from the target distribution $P(Z|Y)$. MGSA derives posterior inferences for parameters $\zeta = (\pi, \alpha, \gamma)$ along with the random walk over configurations of joint states. For SFCA, we fix ζ first and generate posterior inference of joint states. We think that these parameters should not be conjectured along with category analysis, but estimated from external sources of information and assist category analysis. If the observational data is a list of gene labeled as interesting, α means the false positive rate existing in the gene lists and $1-\gamma$ is the false negative error measurement. They should be controlled at a certain level when the gene list is generated. The choice of π is more arbitrary but it should reflect how concise we want the final category list that represents the genomic data to be.

Given the current joint state $Z = \{Z_j\}_{j=1}^{C}$, the M-H algorithm proposes a neighbor state Z^* according to a proposal distribution $Q(\cdot|Z)$. Accept the proposal with probability P_{accept} which is defined as:

$$\begin{aligned} P_{accept} &= \min(1, r) \\ \text{where } r &= \frac{P(Z^*|Y)Q(Z|Z^*)}{P(Z|Y)Q(Z^*|Z)} \\ &= \frac{P(Z^*)P(Y|Z^*)Q(Z|Z^*)}{P(Z)P(Y|Z)Q(Z^*|Z)}, \end{aligned} \tag{3.11}$$

where $P(Z)$ is the prior, and $P(Y|Z)$ is the probability of data given category activities. Formula (3.6) gives an explicit expression of $P(Y|Z)$ with data being binary records. It

could be conveniently extended to beta-binomial model with (3.7) or to fit in multinomial data with (3.8). In practice, randomly generate $u \sim \text{Unif}(0, 1)$, and accept Z^* if $r > u$. This procedure is iteratively applied to collect samples. A burn-in period consisting a certain number of iterations is used to initialize the chain.

If we replace $P(Y|\cdot)$ by 1 in (3.11), the ratio r becomes:

$$r = \frac{P(Z^*)Q(Z|Z^*)}{P(Z)Q(Z^*|Z)} \tag{3.12}$$

hence M-H algorithm can be applied to sample from $P(Z)$ as a target distribution. It is trivial for prior is P_0, but it is useful to estimate P_1 and P_2 which are not easy to derive analytically, especially when dimensionality is large and even the magnitude of $\mathcal{Z}$ is not accessible. We will use this method to study different priors in simulations and real data analysis.

3.5.2 Proposal

MGSA proposes a new joint state by either toggling the "on/off" state of one category or switching a pair of states with one on and one off. This proposal is straightforward and easy to apply, however, with constraints imposed by the activation hypothesis, it generates invalid states. Possible ways to fix it include following this proposal and (1) reject all invalid states and keep only valid ones, (2) convert invalid states to valid ones, and (3) develop new proposal to directly generate valid states. We claim that the first two solutions are inefficient or unrealistic given the dimensionality of category analysis.

When C is big and there are a large number of overlapping categories, most combinations of binary states are invalid. Thus, it would be extremely inefficient for method (1) to move to valid states. Method (2) requires the mapping between 2^C possible joint states and all valid states in $\mathcal{Z}$ to calculate the jumping probabilities, which is also intractable when C is big. Thus, we resort to method (3) to directly propose valid states that is local to the current state.

Updating rules

Before describing the proposal, we need to introduce an operation that corrects an invalid joint state to valid. Any given configuration of a joint activation state has two equivalent presentations: category level joint state and atom level joint state denoted by Z and A respectively. If this state is invalid, i.e. $Z \notin \mathcal{Z}$ or $A \notin \mathcal{A}$, correction can be operated in two ways. First we could start with Z, map it to $\mathcal{A}$ by (3.4) to obtain a joint atom level activation state A^*, then map A^* back to $\mathcal{Z}$ by (3.5) to get a new category level joint state Z^*. Finally, update Z to Z^* and A to A^*. We call this operation a *max-min* rule as we obtain atom level joint state by performing maximization followed by a minimization to get an updated category level joint state. Similarly, we could define a *min-max* rule by starting with A: map it to $\mathcal{Z}$ by (3.5) to get Z^*, then map Z^* to $\mathcal{A}$ by (3.4) to obtain A^*. Since the mapping between $\mathcal{Z}$ and $\mathcal{A}$ is one-to-one, either *max-min* or *min-max* rule guarantees to generate valid states. Whether new states generated by these two rules are the same or not is decided by incidence matrix X which presents the structural

information. These operations can be mathematically expressed as follows.

Max-min rule:

1. Start with $Z = \{Z_1, Z_2, ..., Z_C\}$ and $A = \{A_1, A_2, ..., A_N\}$;
2. *max* step: $A_i^* = \max_{j:x_{i,j}=1} Z_j, i = 1, 2, ..., N$.
3. *min* step: $Z_j^* = \min_{i:x_{i,j}=1} A_i^*, j = 1, 2, ..., C$.
4. Update joint states to $Z^* = \{Z_1^*, Z_2^*, ..., Z_C^*\}$ and $A^* = \{A_1^*, A_2^*, ..., A_N^*\}$.

Min-max rule:

1. Start with $Z = \{Z_1, Z_2, ..., Z_C\}$ and $A = \{A_1, A_2, ..., A_N\}$;
2. *min* step: $Z_j^* = \min_{i:x_{i,j}=1} A_i, j = 1, 2, ..., C$.
3. *max* step: $A_i^* = \max_{j:x_{i,j}=1} Z_j^*, i = 1, 2, ..., N$.
4. Update joint states to $Z^* = \{Z_1^*, Z_2^*, ..., Z_C^*\}$ and $A^* = \{A_1^*, A_2^*, ..., A_N^*\}$.

Detailed proposal

First let us introduce some notation. Combine Z and its corresponding A and denote by $S = (Z, A)$ category and atom activities. Given incidence matrix X, let K be the number of 1's in X. Then there are K index pairs where each consists of a category index j and one of its atoms' index $i(j)$. If the k^{th} pair is $(j, i(j))$, define $S_k = \left(Z_j, A_{i(j)}\right)$ to be the current states of category j and one of its atom, and S/S_k to be current states

of all other categories and atoms. We see that S_k is uniquely decided by the current state S, $\forall k = 1, 2, ..., K$.

Now consider a mixture of individual proposals where each is associated with a particular pair of category and atom. For example, the k^{th} proposal is designed to operate on S_k. It is easy to see that S_k can only take value from $\{(0,0),(0,1),(1,1)\}$. It can not be $(1,0)$ because according to activation hypothesis, when $Z_j = 1$ all atoms annotated to category j are on, hence $A_{i(j)} = 1$. Define σ_k to be the set of possible values for the pair other than its current value and denote by

$$\begin{aligned} \sigma_k &= \{(0,0),(0,1),(1,1)\}/\left(S_k\right) \\ &= \{s_1, s_2\}. \end{aligned}$$

The proposal associated with index pair k is described as follows.

1. Replace S_k with s_1, and remain Z/Z_k to make a temporary joint state $S' = (Z', A')$ where $S'_k = s_1$.

2. Apply one of the updating rules (either *min-max* or *max-min*) to S' to generate a valid joint state $S'' = (Z'', A'')$ and extract S''_k.

3. If $S''_k = s_1$, let $S^{*1} = S''$. Otherwise, $S^{*1} = S$.

4. Apply steps (1)-(3) to s_2. Denote by S^{*2} the resulting joint state.

5. New state S^* is sampled from proposal distribution $Q_k(\cdot|S)$ given the current state S.

- If $S^{*1} \neq S$ and $S^{*2} \neq S$, (in this case, $S^{*1} \neq S^{*2}$ since $s_1 \neq s_2$)

$$Q_k(S^*|S) = \begin{cases} \frac{1}{2} & \text{if } S^* = S^{*1} \\ \frac{1}{2} & \text{if } S^* = S^{*2} \\ 0 & \text{otherwise.} \end{cases}$$

- If $S^{*1} \neq S$, $S^{*2} = S$ (or $S^{*2} \neq S$, $S^{*1} = S$),

$$Q_k(S^*|S) = \begin{cases} 1 & \text{if } S^* = S^{*1} \text{ (or } S^{*2}) \\ 0 & \text{otherwise.} \end{cases}$$

- If $S^{*1} = S^{*2} = S$,

$$Q_k(S^*|S) = \begin{cases} 1 & \text{if } S^* = S \\ 0 & \text{otherwise.} \end{cases}$$

Similarly we can obtain $Q_k(\cdot|S^*)$ by repeating these steps and calculating jumping probability from S^* to exactly the original state S. At each iteration of this algorithm, an index k is sampled from $\{1, 2, ..., K\}$ with probability $1/K$, and then a new state S^* is proposed from $Q_k(\cdot|S)$. Calculate P_{accept} by replacing $Q(\cdot|\cdot)$ with $Q_k(\cdot|\cdot)$ in (3.11) and (3.12).

The rules regarding whether to apply *min-max* or *max-min* to generate new valid joint state are shown in Figure 25. It depends on both current S_k and proposed value s_1 (or s_2). For example, if $(0, 1)$ is proposed to update $S_k = (0, 0)$. Since Z_j remains at 0, applying *max-min* rule will not move the current state. Instead we should replace $A_{i(j)}$ by 1 and apply *min-max*. When transition is between $(0, 0)$ and $(1, 1)$, and both

Z_j and $A_{i(j)}$ are subject to change, either *min-max* or *max-min* has potential to update current state. We decide to adopt the rules as shown in Figure 25 because by doing so each rule has to be applied once to update S_k to s_1 and s_2, hence every current state is to be updated in two different ways.

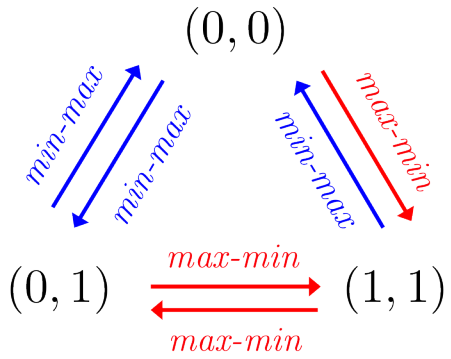


Figure 25: Updating rules regarding usage of *min-max* and *max-min*.

3.5.3 Presentation via factor graph

We use a bipartite graph called factor graph to demonstrate how this proposal moves the current state locally. For simplicity, Example IV only has 4 atoms and 3 categories. Category 2 is overlapping with the other two categories. The incidence matrix X is of dimension 4×3.

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

There are 7 valid states with respect to activation hypothesis as listed in Table 15. Space of valid joint states has magnitude: $|\mathcal{Z}| = |\mathcal{A}| = 7$.

Table 15: Example IV: 7 valid joint states presented in category level (middle column) and atom level (right column).

No.	$\mathcal{Z}$	$\mathcal{A}$
1	000	0000
2	100	1100
3	010	0110
4	001	0011
5	110	1110
6	011	0111
7	111	1111

The annotation profiles presented by X and current states of categories and atoms can be illustrated uniquely by a bipartite factor graph as shown in Figure 26. Each square represents a category and each circle represents an atom. An edge connecting a category and an atom indicates the atom is annotated to the category. The graph is bipartite in that any pair of squares or circles can not be connected directly by an edge. The number of edges in the graph equals the number of 1's in the incidence matrix X. We use shading to indicate the activation state is on. For example, in Figure 26 category 2 is on and it activates atoms 2 and 3 that are annotated to it. The current joint state represented by the graph is state 3 in Table 15.

Suppose we are currently at state 3 with $Z = (0, 1, 0)$, $A = (0, 1, 1, 0)$, and propose to update $S_1 = (Z_1, A_2)$. Since $S_1 = (0, 1)$, the set of proposal values is $\sigma_1 = \{(0, 0), (1, 1)\}$. Figure 27 shows the procedure of proposing new valid states following our description. In

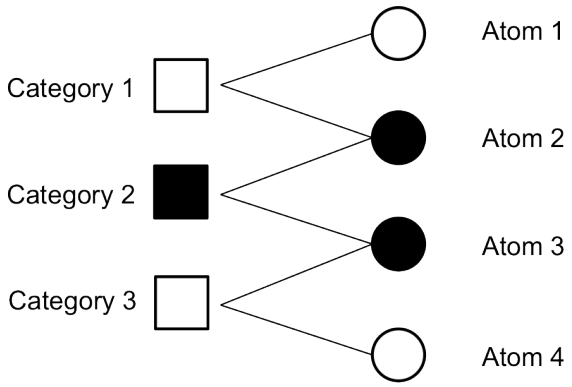


Figure 26: A bipartite factor graph presenting annotation profiles and category/atom level joint states.

both upper and lower panels, factor graphs on the left show temporary states generated by directly replacing S_1 with proposed values. These states whether being valid or not will be corrected by either *min-max* or *max-min* rule to generate valid states S^*. Both panels show that after correction, updated S_1 match the proposed values, i.e. in the upper panel $S_1^* = (0, 0)$ and S^* correspond to state 1; in the lower panel $S_1^* = (1, 1)$ and S^* correspond to state 6. Either state 1 or state 6 will be proposed next with equal probabilities.

3.5.4 Posterior summaries

As we see from earlier examples, it is useful to access all sorts of posterior summaries to characterize the high dimensional space of activation states. Here we show how to calculate 4 major summary statistics from samples collected by the MCMC algorithm. Let $Z^t = \{Z_1^t, Z_2^t, ..., Z_C^t\}$ be the joint state sampled at t^{th} iteration and T be the total

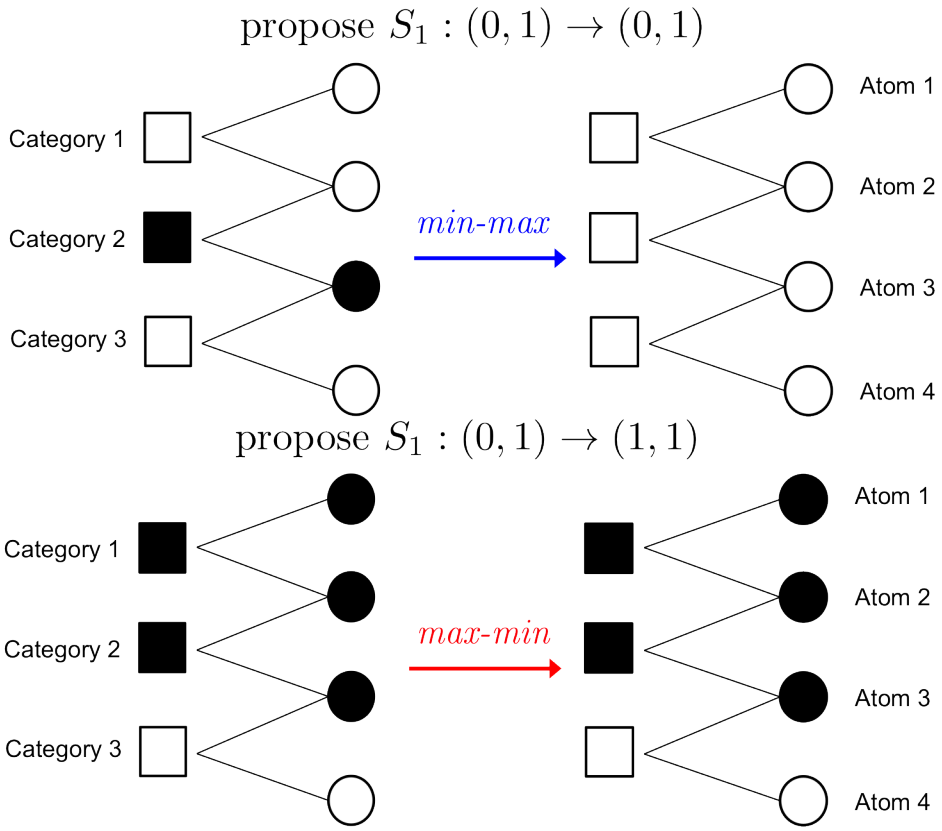


Figure 27: Upper: replace current value of $S_1 = (Z_1, A_2)$ with $(0, 0)$ and apply *min-max* rule to generate state 1; lower: replace current value of S_1 with $(1, 1)$ and apply *max-min* rule to reach state 6.

number of samples collected excluding those from burn-in period.

- Marginal posterior probability of each category being active $P(Z_j = 1|Y), j = 1, 2, ..., C$ is the most direct result from MCMC to show how likely a category is active given observational data, category structure and system parameters.

$$P(Z_j = 1|Y) \approx \frac{\sum_{t=1}^{T} Z_j^t}{T} \tag{3.13}$$

- Marginal posterior probability of each category being a maximum active set gives concise presentation of active categories. It is useful in addition to the previous summary.

$$\begin{aligned} \text{Let } W_j^t &= I(Z_j^t = 1 \text{ and category } j \text{ is a maximum active set}) \\ P(W_j = 1|Y) &\approx \frac{\sum_{t=1}^{T} W_j^t}{T}. \end{aligned} \tag{3.14}$$

- Bayes factors for both Z_j and W_j:

$$\begin{aligned} BF_{Z_j} &= \frac{P(Z_j = 1|Y)}{P(Z_j = 1)} \\ BF_{W_j} &= \frac{P(W_j = 1|Y)}{P(W_j = 1)} \end{aligned}$$

- We know that it is ideal to have posterior distribution on joint states across $\mathcal{Z}$, however, it is not realistic for large scaled problems. Instead of getting a distribution of joint states, we could calculate the probability of data given the joint state at different MCMC samples $P(Y|Z^t)$ and find which state yields the maximum results, i.e.

$$\text{Find } Z^{max} \text{ such that } P(Y|Z^{max}) = \max_{t=1,2,...,T} \left\{ P(Y|Z^t) \right\} \tag{3.15}$$

3.5.5 Convergence issues

In this section we discuss issues regarding convergence of the proposed Metropolis-Hasting algorithm. Most importantly, we need to show that the sequence of MCMC samples generated by the algorithm converges to the target distribution, i.e. the posterior distribution of joint state.

To prove the validity of our algorithm we need two conditions: (1) the simulated sequence is a Markov chain with a unique stationary distribution; (2) the stationary distribution equals the target distribution. Condition (2) is guaranteed by the design of M-H algorithm if the stationary distribution exists. Regarding condition (1), it is sufficient to show that the chain is ergodic and irreducible. First, the chain is aperiodic in that given any current state, it is always possible to reject the proposal and stay at the same state. Since there are finite states ($|\mathcal{Z}| \leq 2^C$), the chain is ergodic. A markov chain is irreducible means that any pair of states can transit to each other, or they communicate. Irreducibility is obvious when all categories are mutually exclusive. Since switching on/off any individual category does not affect states of others, transition between any pair of states is to simply switch individual categories that make the joint states different one at a time. It can be realized by *max-min* or *min-max* rule at each step. For overlapping and hierarchical cases, we have not developed theoretical proof for irreducibility yet. However, from various artificial examples we have checked we find that in a plot where nodes are valid joint states and edges connect communicating states, transition paths are formed to connect all-zero state to all-one state via the other joint

states, and all states belong to at least one of these paths. That is to say, any pair of joint states can communicate by following certain paths via either all-zero state or all-one state. Thus, the chain is irreducible. Figure 28 shows transition paths of Example IV representing overlapping structure.

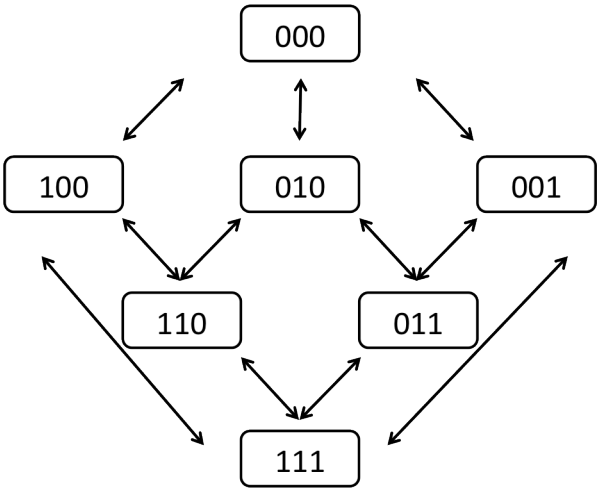


Figure 28: Transition paths of Example IV. Nodes represent valid joint states (presented by category level states). Edges connect pairs of states that communicate with each other.

We have developed R program to implement the proposed M-H algorithm and applied it to various scenarios to assess convergence of the generated chains. All the examples tested are of small scales with fixed parameters so that we could analytically develop the posterior inferences and compare with results from MCMC. Depending on complexity of the data structure, different chain lengths are required to control the deviance of prior and posterior estimation from the truth which is developed analytically. For instance, for Example II a chain length of 10^6 is required to control the deviance of prior and

posterior probability of any joint state under 10^{-3}. With 30 joint states in total Monte Carlo error is estimated as $\sqrt{\frac{1}{30}\frac{29}{30}\frac{1}{10^6}} \approx 2\times10^{-4}$ which is slightly smaller. Since marginal probabilities are calculated by collapsing involved joint states probabilities, deviances are cumulated. For this chain, deviances of all prior and posterior probabilities from analytical results are below 5×10^{-3}. In terms of computational speed, the current program could be improved in the future by implementing major calculations in fundamental languages like C. In practice, instead of depending on one extremely long chain we could generate multiple chains from dispersed starting states and develop inferences based on all samples collected.

3.5.6 Algorithm

Input:

- incidence matrix $X_{N\times C}$, atom size $\{n_i\}_{i=1}^N$, number of observed on genes per atom $Y = \{y_i\}_{i=1}^N$;
- parameters $\zeta = (\pi, \alpha, \gamma)$;
- chain length T.

Result:

- prior of each category being active: $P(Z_j = 1) = \frac{\sum_{t=1}^T Z_j^t}{T}, j = 1, 2, ..., C$ by specifying $P(Y|\cdot) = 1$ in ratio r;

Algorithm 1 A Metropolis-Hasting algorithm to generate posterior inferences

$Z^t \Leftarrow \{0, 0, ..., 0\}_{1\times C}$, $A^t \Leftarrow \{0, 0, ..., 0\}_{1\times N}$, and $S^t = (Z^t, A^t)$
for $t = 1 \rightarrow T$ **do**
  Sample $k \in \{1, 2, ..., K\}$
  Generate $S^* \sim Q_k(\cdot|S^t)$

$$r \leftarrow \frac{P(Z^*)P(Y|Z^*)Q_k(S^t|S^*)}{P(Z^t)P(Y|Z^t)Q_k(S^*|S^t)}$$

  $u \sim \text{Unif}(0, 1)$

  if $u < r$ **then**
    $S^t \leftarrow S^*$
  end if

  $W^t \leftarrow \{0, 0, ..., 0\}$
  $W_j^t = I(Z_j^t = 1$ and category j is a maximum active set)

end for
return $\{S^1, S^2, ..., S^T\}$ and $\{W^1, W^2, ..., W^T\}$

- marginal posterior of each category being active: $P(Z_j = 1|Y) = \frac{\sum_{t=1}^T Z_j^t}{T}$;
- Bayes factor: $BF_{Z_j} = \frac{P(Z_j=1|Y)}{P(Z_j=1)}$;
- prior/posterior/Bayes factor of each category being a maximum active set: $P(W_j = 1) = \frac{\sum_{t=1}^T W_j^t}{T}$, $P(W_j) = \frac{\sum_{t=1}^T W_j^t}{T}$, and $BF_{W_j} = \frac{P(W_j=1|Y)}{P(W_j=1)}$;
- joint state Z^{max} such that $P(Y|Z^{max}) = \max_{t=1,2,...,T}\{P(Y|Z^t)\}$.

3.6 Case study

3.6.1 Data and method

In Chapter 2 we present a meta analysis of RNAi studies in influenza virus replication. The 4 studies of interest have confirmed a total number of 614 genes after two screens. A gene set enrichment analysis has been conducted in GO by applying *allez* (Newton et al. (2007)) and finds 19 categories most significantly enriched with the confirmed genes. With 2 redundant ones being removed, 17 unique GO categories are listed in Table 14). As expected they are overlapping and present similar biological functions, since *allez* belongs to one-at-a-time functional category analysis methods that ignore GO structure and treat categories independently. In this section we apply two model-based category analysis methods MGSA (mgsa version 1.4.0) and SFCA in R (version 2.15.1) to this confirmed gene list ("flu data") in both KEGG and GO system and compare their findings.

For convenience of calculation, KEGG and GO system are first trimmed to include only categories that have no more than 50 genes and overlap with the list of 641 genes. Then atom level incidence matrices are created by collapsing common annotation profiles and keeping only unique columns. The most updated Bioconductor database KEGG.db 2.7.1 and org.Hs.eg.db 2.7.1 are used to extract annotation information. Information of data used for analysis are listed in Table 16. The numbers of categories and atoms of trimmed data are dimensions of corresponding incidence matrix X. The last column

refers to the total number of confirmed genes in each study set. They are reported to be the active genes but subject to the false negative error. The numbers are always below 614 as some genes are not annotated by KEGG/GO yet or they are annotated only to big categories that are trimmed.

Table 16: Basic information of GO and KEGG system used to analyze flu data. For each system, original data on annotation profiles are trimmed to include only categories that have no more than 50 genes and overlap with the list of 614 confirmed genes in flu data.

		# categories	# genes	# atoms	# confirmed genes
KEGG	original	3152	75100	-	390
	trimmed	61	1460	172	130
GO	original	15492	14572	-	535
	trimmed	2682	8284	5770	442

3.6.2 Data analysis in KEGG

After trimming there are 61 categories left in KEGG. Every category has overlap with some others but they do not form any parent-child relationship. It means that marginal inferences from SFCA and MGSA should be very similar if not the same. For both SFCA and MGSA, parameters are fixed at $\{\alpha = 0.05, \gamma = 0.5, \pi_0 = 0.1\}$. The choice of parameter values are reasonable for the following reasons: (1) A list of important genes from the first step analysis, i.e. differential expression analysis are usually selected with false discovery rate (FDR) controlled under 5%; (2) a marginal enrichment rate of 50% is considered as fairly high for a category; (3) $\pi_0 = 0.1$ for MGSA means that we expect

about 10% of categories to be truly active. For SFCA the parameter π will be adjusted if the estimated mean prior from MCMC suggests that it deviates far from input value 0.1. Five chains are generated where each has a length of 2×10^6 with no data input to generate the prior distribution of SFCA, and the same procedure is followed with input of the flu data for posterior inference. For each chain, the first 10^4 samples are excluded as burning period. The mean acceptance rate for prior and posterior chains are respectively 18.4% and 2.2%. Chains of the same lengths are also collected from MGSA, but the acceptance rate is not accessible.

Table 17 reports information and marginal inferences on categories that have posterior probabilities above 0.01 when rounded to 2 decimal places. We see that MCMC results from MGSA and SFCA are very close, as expected. The estimated mean prior from SFCA is 0.100 if rounded to 3 decimal places, so there is no need to adjust π. Prior and posterior inferences on categories being maximum actives sets are skipped as W_j is equivalent to Z_j with no hierarchy in the structure.

By MGSA's criteria, the top 3 categories will be reported since their marginal posterior probabilities exceed 0.5. Since the prior is flat over all categories, ranking by marginal posterior is the same as by Bayes factor. In addition to marginal inferences, SFCA's MCMC results suggests that joint state with categories {hsa03050, hsa04966, hsa05219, hsa03060, hsa04977} being active has the largest $P(Y|Z)$, which means that this joint state is most likely to generate the observational data.

Table 17: Inferences on categories with top marginal posterior probability $P(Z_j|Y)$ from SFCA and MGSA. Categories labelled with $*$ are most likely to jointly generate the observational data.

ID	term	# genes	# reported active genes	SFCA $P_1(Z_j)$	SFCA $P_1(Z_j\|Y)$	SFCA BF_{Z_j}	MGSA $P_0(Z_j\|Y)$
hsa03050*	Proteasome	45	14	0.101	1.000	9.83	1.000
hsa04966*	Collecting duct acid secretion	27	7	0.100	0.744	7.41	0.747
hsa05219*	Bladder cancer	43	10	0.100	0.563	5.64	0.560
hsa03060*	Protein export	23	5	0.100	0.096	0.97	0.096
hsa00750	Vitamin B6 metabolism	6	1	0.100	0.043	0.43	0.043
hsa05216	Thyroid cancer	29	6	0.099	0.018	0.18	0.018
hsa00603	Glycosphingolipid biosynthesis-globo series	14	2	0.100	0.005	0.05	0.005
hsa04122	Sulfur relay system	10	1	0.100	0.003	0.03	0.003
hsa04977*	Vitamin digestion and absorptions	24	4	0.100	0.003	0.03	0.003
hsa00770	Pantothenate and CoA biosynthesis	16	2	0.100	0.001	0.01	0.001

3.6.3 Data analysis in GO

Data structure for GO remains complicated with overlaps and hierarchies after trimming. The incidence matrix X is of dimension 5770×2682. There are 42922 1's in X which means there are a large number of individual proposals to sample from when moving the chain forward. With possibly low acceptance rate, the chain length of MCMC should be fairly large to achieve stable results. For a system of this complexity, the Monte Carlo error has not been well controlled. Experiments on more restricted cases need to be conducted.

3.7 Computation via graphical probabilistic models

Our major solution for role model posterior inferences is MCMC. So far this approach can only access marginal posterior probabilities and develop very limited summary on posteriors of joint states, which may not be sufficiently informative due to the high dimension of the parameter space. Also in any event, MCMC error is very difficult to assess, with the real prospect of poorly mixing chains. In Newton et al. (2012), we have considered possible non-MCMC computations via message-passing algorithms and techniques from probabilistic graphical models. In particular, 2 kinds of graphs are introduced to facilitate posterior calculation.

The first graph is called *intersection graph* where nodes are categories, and edges connect categories that share common genes. Figure 3.7 gives a simple example of the

category intersection graph.

ID	functional category
00010	Glycolysis/Gluconeogenesis
00020	Citrate cycle (TCA cycle)
00030	Pentose phosphate pathway
00040	Pentose and glucuronate interconversions
00051	Fructose and mannose metabolism

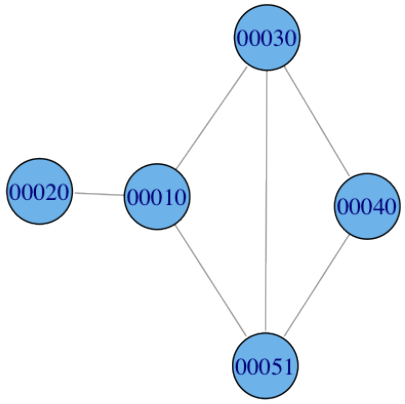


Figure 29: Category intersection graph for 5 KEGG pathways. Each node represents a category and every edge connects a pair of overlapping categories.

In Newton et al. (2012), we prove the following proposition and show that the joint posterior of category activation states factorizes into local functions over the intersection graph, and category intersection s graph can be used in principle to support various inference computations implied by the role model.

Proposition 3.2. *The role model posterior (3.6) satisfies:* $P(Z|Y) \propto \prod_{j=1}^{C} \Psi_j \left[Z_j, Z_{nb(j)}\right]$, *where* Ψ_j *is a data-dependent function of both* Z_j *and neighboring states* $Z_{nb(j)} = \{Z_{j'} : j \cap j' \neq \emptyset\}$.

Ideally, one would like to utilize the entirety of GO or KEGG. However the associated intersection graphs are highly complex and prohibit exact numerical methods as illustrated by Figure 3.7.

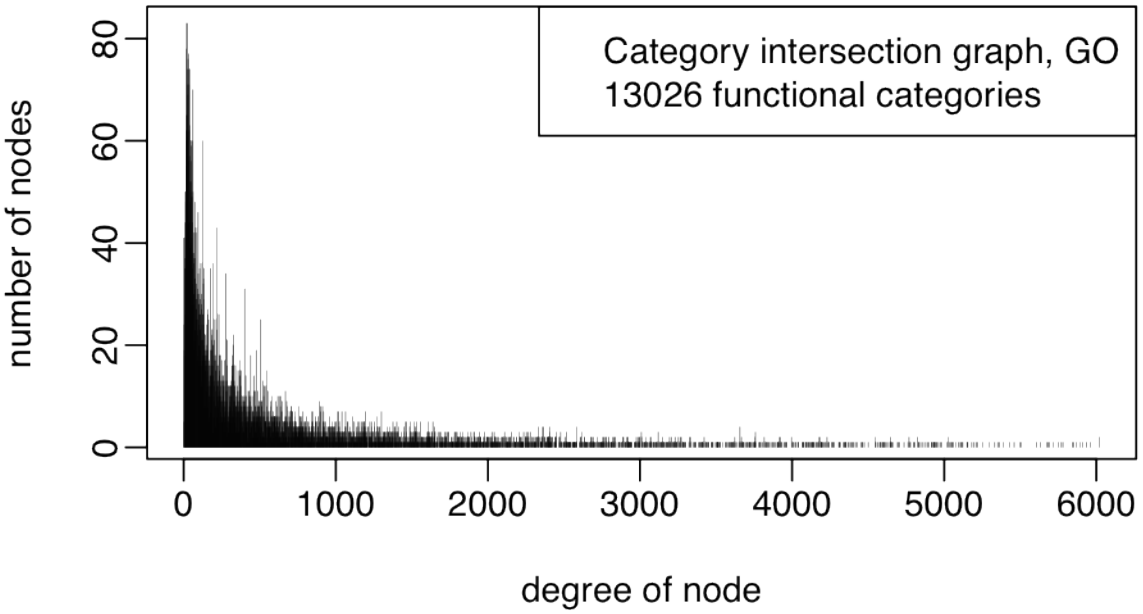


Figure 30: Degree distribution of intersection graph of GO (categories holding between 1 and 500 human genes) from Bioconductor database org.Hs.eg.db 2.6). It is somewhat remarkable that so many overlaps are possible. The most extreme case is the category cell motility (GO:0048870), which annotates 495 human genes and shares genes with 6160 other categories among the 13026 GO categories that annotate between 1 and 500 human genes. These 13026 categories annotate 14047 genes. The median number of other cate- gory assignments per cell-motility gene is 64, and one gene happens to be in 631 other categories.

With one-to-one mapping between spaces of joint category- and atom level activation states (3.1), we consider reparameterization of joint posterior from category to atom level to approximate inference and reduce computational complexity. With proper prior distribution $P(A)$ on valid space of atom level activation states $\mathcal{A}$ chosen, joint posterior 3.6 can be expressed as $P(A|Y) \propto P(A) \prod_{i=1}^{N} P(y_i|A_i)$. Just as the intersection graph of the categories is the data structure supporting posterior inference in the original parameterization, we develop another graph called *function profile graph* that supports atom level computations. Its nodes are the atoms. There is a directed edge we from ν and ν' if: (1) the assignments at ν' are a proper subset of the assignments at ν, and also (2) there is no other atom ν'' with assignments that are a subset of assignments at ν and a superset of assignments at ν'. We say ν is a parent of ν' and ν' is a child of ν. An example is shown in Figure 3.7.

To support inference we need an undirected version of the function profile graph, which we obtain by a form of moralization used in graphical models analysis. Specifically, we include an undirected edge between any two nodes ν and ν' that are both parents of a common child. We also include an undirected edge between any two nodes ν and ν' that are children of a common parent. This two-way moralization comes from the fact that information flows both ways along a given directed edge. Finally we make all remaining directed edges undirected.The resulting graph is the undirected function profile graph. An example is given in Figure 3.7.

Similarly we prove the following proposition dealing with posterior on atom level

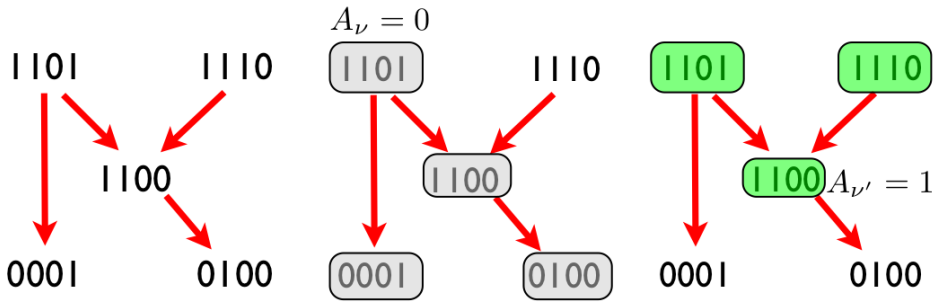


Figure 31: Reparameterizing the role model with a function profile graph: The nodes in each panel represent 5 atoms. Each atom shows a profile of assignments (1) or not (0) to 4 categories. A directed edge goes from ν to ν' if the assignments at ν include those at ν' (except we omit redundant edges e.g., no edge from 1110 to 0100.) The middle and right panels show logical dependencies on activity variables. E.g., in the middle panel, knowing $A_\nu = 0$ implies $A_{\nu'} = 00$ for all downstream atoms, and knowing $A_{\nu'} = 1$ on the right panel implies $A_\nu = 1$ for all upstream atoms.

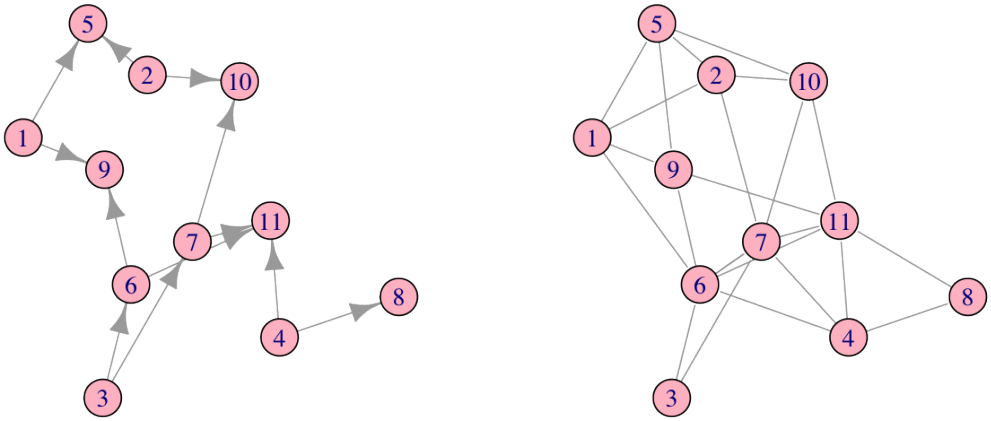


Figure 32: Function profile graphs for the small KEGG example shown in Figure 3.7, with 11 atoms as listed.

activation states.

Proposition 3.3. *For suitable prior $P(A)$ over $\mathcal{A}$, the posterior distribution on atom level activation states is the product of functions $\tilde{\Psi}_i$ that are local in the undirected function profile graph: $P(A|Y) \propto \prod_{i=1}^{N} \tilde{\Psi}_i \left[A_i, A_{nb(i)}\right]$, where $\tilde{\Psi}_i$ is a data-dependent function of both A_i and neighboring states $A_{nb(i)}$.*

Coupled with mapping between $\mathcal{A}$ and $\mathcal{Z}$ (Proposition 3.1), the above result indicates that we can perform inference computations on the function profile graph, and then transform back as needed to get inference on category level activation states.

In GO, for example, the transformation provides a much simpler graph (Figure 33). Unfortunately even this simpler graph is still too complicated for exact numerical methods. Approximation methods remain under investigation.

3.8 Relaxation of the role model

In addition to exact calculations via graphical model and approximation by MCMC methods, we have also considered a third option for role model computations. We proposed two relaxations of the role model that give different generalized-linear-model (GLM) representations of gene-level data. Regularized regression and quadratic programming were developed to fit the relaxed models and provide selection of the most significant functional categories. In this section, we introduced this two relaxed models and the major difficulties we have encountered. With more sophisticated optimization

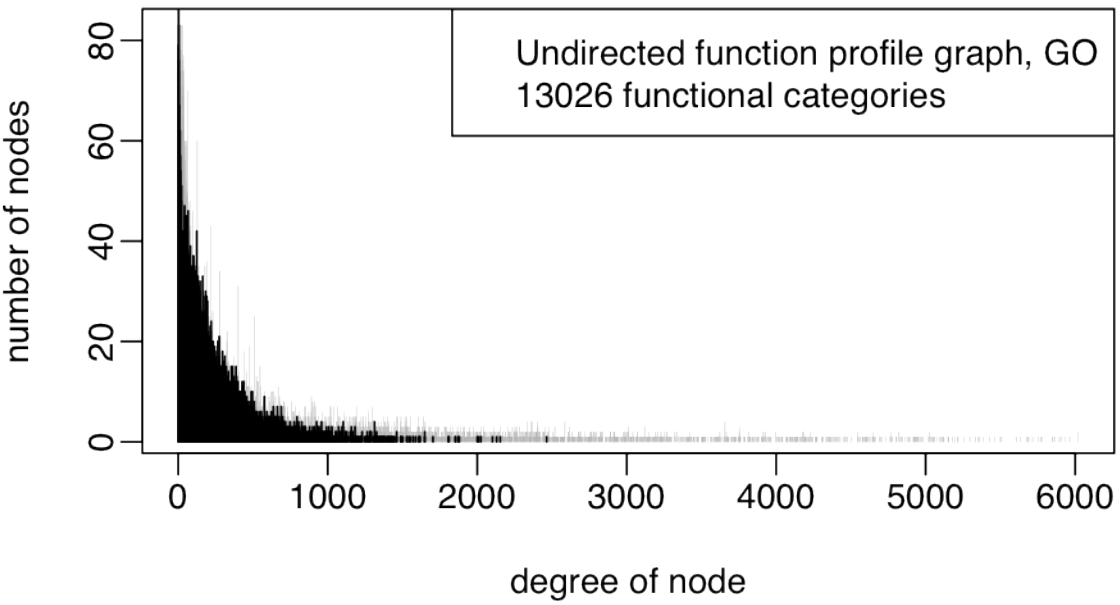


Figure 33: Degree distribution of the undirected function profile graph of GO (categories holding between 1 and 500 human genes). The maximal degree is 2464; the graph itself has 10366 nodes (atoms). The corresponding results for the category intersection graph (from Figure 3.7) are repeated here in grey. Not shown are results for the directed function profile graph, which is much simpler, having maximal degree 268.

and computational techniques, we think it valuable for future work.

We are guided by work on generalized linear models (GLMs), and express the success rate p_i in the central model (3.1) as a function of a linear predictor. Specifically, assume we have parameters $\alpha, \gamma \in (0, 1)$, with $\alpha < \gamma$ as before, and also that we have extended-real-valued parameters β_j, for $j = 1, 2, \ldots, C$, such that $\beta_j \geq 0$. Let β denote the column vector $(\beta_1, \beta_2, \ldots, \beta_C)^T$ and consider the linear predictor $\eta_i = x_i^T \beta$. In model GLM I, we assume

$$p_i = p^{\mathrm{I}}(x_i) = \gamma - (\gamma - \alpha)e^{-\eta_i}$$

Note that $p^{\mathrm{I}}(x_i) \in [\alpha, \gamma]$ for any parameter settings. Further, we can prove that the role model is a sub-model of GLM I, i.e. if $\beta_j = -\log(1 - Z_j)$ for all j and for Z_j's in $\{0, 1\}$, then $p^I(x_i) = p^{\mathrm{RM}}(x_i)$, for all annotation profiles x_i.

$$\begin{aligned} p_i &= p^I(x_i) = \gamma - (\gamma - \alpha)e^{-\eta_i} \\ \text{where } \eta_i &= x_i^T \beta, \beta \geq 0, \gamma > \alpha \end{aligned}$$

In the GLMs described above, the coefficients in β are contributions of categories on the gene-level data. Instead of evaluating contributions from each and every category, we are interested in selecting a list of most representative ones that jointly explain the observed data. Thus, we need to apply effective model selection schemes to generate sparse solutions. An immediate thought is to adopt the LASSO penalty due to its nice property of forcing some coefficients to exactly 0. If without taking into account activation hypothesis, to fit L_1-regularized GLM I at fixed α and γ is to minimize the

negative log likelihood function $l_{\mathrm{I}}(\beta)$ with the LASSO penalty:

$$\min_{\beta \geq 0} -l_{\mathrm{I}}(\beta) + \|\beta\|_1$$

where $l_{\mathrm{I}}(\beta) = \sum_{i=1}^{N} y_i \log\left[\gamma - (\gamma - \alpha)e^{-\eta_i}\right] + (n_i - y_i)\log\left[1 - \gamma + (\gamma - \alpha)e^{-\eta_i}\right]$.

We find that the objective function is not convex as the Heissian matrix $\frac{d^2 l_{\mathrm{I}}(\beta)}{d\beta d\beta^t}$ is not positive semidefinite. That is to say it is not guaranteed to have a global optimal, and local minimum solutions are intractable. Not to mention adding constraints for activation hypothesis. For this reason, we consider model GLM II whose link function is the logit function used for logistic regression to ensure convexity. The success probability is expressed as:

$$\begin{aligned} p_i &= p^{\mathrm{II}}(x_i) = \frac{\exp^{\eta_i}}{1 + \exp^{\eta_i}} \\ \eta_i &\in \left[\log\left(\frac{\alpha}{1-\alpha}\right), \log\left(\frac{\gamma}{1-\gamma}\right)\right] \end{aligned} \tag{3.16}$$

Linear constraint (3.16) ensures $p_i \in [\alpha, \gamma]$. By adding an intercept $\beta_0 = \log\left(\frac{\alpha}{1-\alpha}\right)$ to η_i the constraint becomes

$$\eta_i \in \left[0, \log\left(\frac{\gamma(1-\alpha)}{\alpha(1-\gamma)}\right)\right].$$

Now we can add non-negative constraints on β_j. The objective function is:

$$\begin{aligned} &\min_{\beta} -l_{\mathrm{II}}(\beta) \\ &\text{s.t.} \sum(\beta_j) < t \\ &\quad \eta_i \in \left[0, \log\left(\frac{\gamma(1-\alpha)}{\alpha(1-\gamma)}\right)\right] \end{aligned}$$

To take into account activation hypothesis, more constraints on β are needed. For example, if category 1 is a parent of category 2 meaning the latter contains a subset of gene in the former, according to activation hypothesis we must have $\beta_1 \leq \beta_2$ to guarantee that if category 1 appears in the final selection all its children must be selected as well. This condition adds a system of linear constrains whose size equals the number of parent-child relationships, or the number of 1's in the incidence matrix. Other than the large number of constraints to deal with, another computational challenge is that the objective function needs to be quadratic approximated to fit in quadratic programming. Even if a global minimum at each iteration is reached, it is not guaranteed that they converge to the optimal solution of the original problem.

Last but not least we recognize that expressing activity of atoms as linear combinations of category level coefficients is inconsistent with the role model. In the role model, atom is activated if any one of the categories it is annotated to is active. Thus more categories being on does not increase the chance that this atom is on. We see that categories' joint effect is not cumulative as in a linear combination. Thus, regression approach is very limited in providing the role model solution.

3.9 Concluding remarks

Functional category analysis deals with integration of experimentally derived data with functional annotation data. It is very important in statistical genomics as it serves to

describe observational data more concisely and detects weak gene-level data more effectively. Most available methods derive category-level inference based on only activities of genes annotated to them. Model-based category analysis methods take a different approach that non-null behavior start with functional categories and gene activities are modeled by taking into account activities of all categories involved. The complex category structure is therefore conveniently incorporated into the model and inferences on category activities are derived simultaneously.

Our model-based category analysis method SFCA is developed based on an existing method MGSA (Bauer et al. (2010)) and our previous investigation(Newton et al. (2012)). Our first contribution is that an important condition called the activation hypothesis is developed to establish one-to-one correspondence between gene- and category-level activities. By clarifying intrinsic constraints among role model parameters, we are able to infer categories more accurately to explain gene-level signals. For computations, we developed MCMC methods to approximate posterior inference, especially a sophisticated algorithm which is able to sample the chain on a highly restricted space.

We have shown that MGSA suffers inconsistency in posterior inference from our empirical simulation studies, and also less efficient compared to SFCA in dealing with overlapping and hierarchical category structures, which is often the case in reality. In addition to marginal posterior inference on categories being active, SFCA is able to generate other useful posterior summaries including inference on joint activation states and on maximum active sets.

Other than observations being Binomial distributed, SFCA is also flexible enough to accommodate different types of variations and data. Two other approaches for role model posterior computations including exact calculations via probabilistic graphical models and relaxed modeling have also been investigated.

Bibliography

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J. et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, **25** 25.

Barry, W., Nobel, A. and Wright, F. (2005). Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, **21** 1943–1949.

Basu, S. and Ebrahimi, N. (2001). Bayesian capture-recapture methods for error detection and estimation of population size: Heterogeneity and dependence. *Biometrika*, **88** 269–279.

Bauer, S., Gagneur, J. and Robinson, P. (2010). Going bayesian: model-based gene set analysis of genome-scale data. *Nucleic acids research*, **38** 3523–3532.

Bauer, S., Robinson, P. and Gagneur, J. (2011). Model-based gene set analysis for bioconductor. *Bioinformatics*, **27** 1882–1883.

Beissbarth, T. and Speed, T. (2004). Gostat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, **20** 1464–1465.

Boca, S., Bravo, L. J. L., H.C. and Parmigiani, G. (2010). A decision-theory approach to interpretable set analysis for high-dimensional data.

Brass, A., Huang, I., Benita, Y., John, S., Krishnan, M., Feeley, E., Ryan, B., Weyer, J., Van Der Weyden, L., Fikrig, E. et al. (2009). The ifitm proteins mediate cellular resistance to influenza a h1n1 virus, west nile virus, and dengue virus. *Cell*, **139** 1243–1254.

Carvalho, L. and Lawrence, C. (2008). Centroid estimation in discrete high-dimensional spaces with applications in biology. *Proceedings of the National Academy of Sciences of the United States of America*, **105** 3209.

Craig, B., Newton, M., Garrott, R., Reynolds III, J. and Wilcox, J. (1997). Analysis of aerial survey data on florida manatee using markov chain monte carlo. *Biometrics* 524–541.

Draghici, S. and Krawetz, S. (2003). Global functional profiling of gene expression data. *A practical approach to microarray data analysis* 306–325.

Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *The Annals of Applied Statistics* 107–129.

Fisher, R., Corbet, A. and Williams, C. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology* 42–58.

Goeman, J. and Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23** 980.

Grossmann, S., Bauer, S., Robinson, P. and Vingron, M. (2007). Improved detection of overrepresentation of gene-ontology annotations with parent–child analysis. *Bioinformatics*, **23** 3024–3031.

Hao, L., He, Q., Craven, M., Newton, M. and Ahlquist, P. (2012). Influenza-virus rnai screens exhibit limited gene-level agreement owing more to false-negative than false-positive factors (in submission).

Hao, L., Sakurai, A., Watanabe, T., Sorensen, E., Nidom, C., Newton, M., Ahlquist, P. and Kawaoka, Y. (2008). Drosophila rnai screen identifies host genes important for influenza virus replication. *Nature*, **454** 890–893.

Jiang, Z. and Gentleman, R. (2007). Extensions to gene set enrichment. *Bioinformatics*, **23** 306.

Karlas, A., Machuy, N., Shin, Y., Pleissner, K., Artarini, A., Heuer, D., Becker, D., Khalil, H., Ogilvie, L., Hess, S. et al. (2010). Genome-wide rnai

screen identifies human host factors crucial for influenza virus replication. *Nature*, **463** 818–822.

König, R., Stertz, S., Zhou, Y., Inoue, A., Hoffmann, H., Bhattacharyya, S., Alamares, J., Tscherne, D., Ortigoza, M., Liang, Y. et al. (2009). Human host factors required for influenza virus replication. *Nature*, **463** 813–817.

Kulkarni, M., Booker, M., Silver, S., Friedman, A., Hong, P., Perrimon, N. and Mathey-Prevot, B. (2006). Evidence of off-target effects associated with long dsrnas in drosophila melanogaster cell-based assays. *Nature methods*, **3** 833–838.

Liang, K. and Nettleton, D. (2010). A hidden markov model approach to testing multiple hypotheses on a tree-transformed gene ontology graph. *Journal of the American Statistical Association*, **105** 1444–1454.

Lu, Y., Rosenfeld, R., Simon, I., Nau, G. and Bar-Joseph, Z. (2008). A probabilistic generative model for go enrichment analysis. *Nucleic acids research*, **36** e109–e109.

Mohr, S., Bakal, C. and Perrimon, N. (2010). Genomic screening with rnai: results and challenges. *Annual review of biochemistry*, **79** 37–64.

Newton, M., He, Q. and Kendziorski, C. (2012). A model-based analysis to infer the functional content of a gene list. *Statistical Applications in Genetics and Molecular Biology*, **11**.

Newton, M., Quintana, F., den Boon, J., Sengupta, S. and Ahlquist, P. (2007). Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *The Annals of Applied Statistics* 85–106.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **27** 29.

Raftery, A. (1988). Inference for the binomial n parameter: A hierarchical bayes approach. *Biometrika*, **75** 223–228.

Schervish, M. (1997). *Theory of statistics*. Springer.

Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E. et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **102** 15545.