LEVERAGING RNA-SEQ TO DETECT NOVEL PROTEIN VARIATIONS VIA MASS SPECTROMETRY

by

Gloria M. Sheynkman

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Chemistry)

at the

UNIVERSITY OF WISCONSIN–MADISON

2014

Date of final oral examination: 5/12/2014

The dissertation is approved by the following members of the Final Oral Committee:
Lloyd Smith, Professor, Chemistry
Colin Dewey, Professor, Biostatistics & Medical Informatics and Computer Sciences
Lingjun Li, Professor, Pharmaceutical Sciences and Chemistry
Tony Stretton, Professor, Zoology
Michael Sussman, Professor, Biochemistry

To Leon.

ACKNOWLEDGMENTS

I want to thank my advisor, Lloyd Smith, for teaching me so many valuable lessons including that above all I need to have fun in science! He gave me so much encouragement (and guidance when I needed it) and so many great opportunities throughout these past five years. I could not have asked for a better advisor and scientific role model.

I gratefully acknowledge the staff scientists of the group for helping me grow as a scientist. I thank Michael Shortreed for, early on, he believed in me and that encouragement gave me the confidence to listen to my own research compass. Later, he became someone who I could turn to talk about research, brainstorm new ideas, and discuss science philosophy. Through example, he also taught me the importance of balance: balance between creative and analytical thought during a research project and balance between work and play in life. I also thank Brian Frey for always patiently welcoming discussions about chemistry, grants, papers, projects, and any of a dozen things that I needed to talk about. Brian's keen ability to distill complex ideas into their essential parts helped me clarify my own thinking. He also taught me about working well with others and the oft-ignored human elements of research. I thank Mark Scalf for all the support regarding all things mass spec., and showing me the importance of intuition in research.

I want to thank Timothy Griffin, Pratik Jagtap, and Jim Johnson from the University of Minnesota and Minnesota Supercomputing Institute for welcoming a collaboration to work on Galaxy-P workflows. They were a pleasure to work with and they made me realize that science not only occurs within the walls of one research group, but can happen across any boundary.

I also want to thank my committee for all of their input and advice regarding my graduate education. Colin Dewey and his student, Bo Li, introduced me to the world of RNA sequencing. I also thank Lingjun Li, Tony Stretton, and Michael Sussman for taking time to be on my committee.

I was fortunate to cross paths with so many wonderful graduate students, scientists, and support staff throughout the years. I thank Jason Russell for initially

training me in proteomics methods. I enjoyed working with Ashlan Musante, Yuan Yuan, and Cheng-Hsien within the Wisconsin CEGS project. I appreciated coffee breaks with Julia Kennedy-Darling, Rachel Knoener, and Ranran Liu during which we had wonderful discussions about science and life. I want to also thank Gergana Heinrichs and Will Horvat, two excellent undergraduate students who aided me in my research projects. I was also so lucky to work in proximity to supportive departmental staff: Cheri Stephens, Sue Martin-Zernicke, and Kristi Heming.

I also thank those who encouraged me to enter a life of research in the first place. Mayland Chang and Shahriar Mobashery from the University of Notre Dame were the first to encourage me to go to graduate school; they placed that idea in my head. Victor Rucker, my first boss from Gilead Sciences also urged me to pursue research. Encouragements from these scientists and a few others strongly influenced my own perception of what I could achieve.

And last, but not least, I thank my family, but especially my husband, Leon Sheynkman. He has believed in me from day one and has fully supported me in chasing my dreams. I owe all the good things that happen in my life to him. For that, I am deeply thankful.

AUTHOR CONTRIBUTIONS

Below are descriptions of the intellectual and scientific contributions of G.M.S. towards each paper.

Chapter 2 was published in *Molecular & Cellular Proteomics*:

Sheynkman, G. M., Shortreed, M. R., Frey, B. L., and Smith, L. M. (2013) Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. *Mol Cell Proteomics* 12, 2341-2353

G.M.S. conceived of the study, designed the experiments, prepared the samples (cell culture, RNA extraction, proteomic sample preparation/fractionation), collected the mass spectrometry (MS) data, performed the RNA-Seq and MS data analysis, wrote the computational algorithms, and wrote the manuscript.

Chapter 3 was published in *Journal of Proteome Research*

Sheynkman, G. M., Shortreed, M. R., Frey, B. L., Scalf, M., and Smith, L. M. (2014) Large-scale mass spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences. *J Proteome Res* 13, 228-240

G.M.S. conceived of the study, designed the experiments, prepared the samples (cell culture, RNA extraction, proteomic sample preparation/fractionation), performed the RNA-Seq and MS data analysis, wrote the computational algorithms, and wrote the manuscript.

Chapter 4 was submitted for publication.

Sheynkman, G., Johnson, J. E., Jagtap, P. D., Shortreed, M. R., Onsongo, G., Frey, B. L., Griffin, T. J., and Smith, L. M. (2014) Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. *Genome Biol*

G.M.S. conceived of the study along with J.P.D., T.J.G, and L.M.S. G.M.S. designed and built the Galaxy-P workflows and wrote the manuscript. With the guidance of G.M.S., J.E.J. wrote the programming code underlying the customized Galaxy-P programs.

CONTENTS

Contents v

List of Tables vii

List of Figures viii

Abstract x

1 Introduction 1
- *1.1 Proteomic variation* 1
- *1.2 MS-based proteomics* 1
- *1.3 MS database searching* 4
- *1.4 Target-decoy searching* 5
- *1.5 RNA-Seq* 7
- *1.6 RNA-Seq bioinformatic analysis* 9
- *1.7 The human reference proteome* 10
- *1.8 Augmenting the reference proteome with customized databases* 11
- *1.9 RNA-Seq-derived protein databases* 13

References 14

2 Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq 18
- *2.1 Abstract* 18
- *2.2 Introduction* 19
- *2.3 Experimental Procedures* 22
- *2.4 Results* 27
- *2.5 Discussion* 42
- *2.6 Acknowledgements* 45

References 45

3 Large-scale mass spectrometric detection of variant peptides resulting from non-synonymous nucleotide differences 52
3.1 Abstract 52
3.2 Introduction 53
3.3 Experimental Procedures 56
3.4 Results 62
3.5 Discussion 78
3.6 Acknowledgements 82
References 82

4 Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations 89
4.1 Abstract 89
4.2 Introduction 90
4.3 Experimental Procedures 92
4.4 Results 94
*4.5 Conclusions*102
*4.6 Acknowledgements*104
*References*104

LIST OF TABLES

2.1 Frequency of splicing events represented by the 57 junction peptides passing 1% local FDR. 41

4.1 Results from creating SAP databases and using them for searching proteomic datasets. 98
4.2 Results from creating splice junction databases and using them for searching proteomic datasets. 100
4.3 Results from MS searching with the original Ensembl protein database and the reduced database. 102

LIST OF FIGURES

1.1 Sequence and structural variations found in genomes and transcriptomes. 2
1.2 The central dogma of molecular biology. 3
1.3 Mass spectrometry-based proteomics workflow. 4
1.4 The MS database search method use target-decoy validation. 6
1.5 Overview of the RNA-Seq workflow 8
1.6 Evolution of DNA sequencing technologies and protein databases. . . 12

2.1 Results overview for the bioinformatic pipeline. 28
2.2 Transcriptomic and proteomic data collection. 30
2.3 Bioinformatics workflow to convert raw RNA-Seq reads into junction peptide sequences. 32
2.4 Read depth frequency distribution for Tophat-detected annotated and unannotated junctions. 33
2.5 Junction sequence processing before translation into peptide sequences. 35
2.6 Distribution of the Tophat exon lengths for unannotated junctions. . . . 36
2.7 Comparison of peptide score distributions for canonical and junction peptides. 38

3.1 Overview of sample-specific SAP peptide detection from custom databases. 63
3.2 Bioinformatic workflow numbers for customized SAP database construction and subsequent MS search results. 64
3.3 Plot of RNA-Seq read depth versus quality score for each called SNV. . 65
3.4 Comparison of average XCorr scores for peptides matching the RefSeq protein, dbSNP-SAP, or custom (RNA-Seq) SAP database. 68
3.5 Bioinformatic workflow numbers for the dbSNP-derived SAP database construction and MS search. 69
3.6 Bioinformatic workflow numbers for the dbSNP-derived SAP database construction and MS search. 72

3.7 Cumulative number of identified SAP peptide and nsSNV sites with consolidated protease digest data. 74
3.8 Distribution of transcript abundances for transcripts encoding detected proteins and transcripts encoding detected SAP peptides. 76
3.9 Comparing SIFT and PolyPhen-2 functional effect prediction scores between all nsSNVs and nsSNVS with a SAP peptide ID. 77
3.10 RNA and protein-level allele-specific expression. 79

4.1 Experimental overview. 95
4.2 Overview of workflows implemented in Galaxy-P that utilize RNA-Seq data for improved proteomics. 96
4.3 Improved peptide confidence scores from reduced database searches. . 102
4.4 Transcript versus protein abundance expression. 103

ABSTRACT

Current practice in mass spectrometry-based proteomics is to identify peptides by comparison of experimental spectra with theoretical spectra derived from a reference protein database. However, this strategy necessarily fails to detect peptides and proteins whose amino acid sequence differs from the reference sequence, such as when there is a genetic difference between the sample and reference genome. Fortunately, next generation sequencing (NGS), specifically RNA-Seq, enables comprehensive determination of the coding transcript sequences present in a given sample. These transcript sequences can then be translated *in silico* to the corresponding proteins and used to build a customized proteomic database that captures all sample-specific (i.e. specific to an individual) protein variations including those resulting from alternative splicing, single amino acid polymorphisms (SAPs), insertions, deletions, translational frameshifts, fusion genes, and RNA editing events. In this dissertation, I show how customized proteomic databases derived from RNA-Seq data can be employed during MS-searching to both enhance proteomic analysis and discover novel peptides. Chapter 2 describes the discovery of novel splice-junction peptides. Chapter 3 describes the large-scale detection of SAP-containing peptides. Finally, Chapter 4 combines the bioinformatic pipelines from Chapter 2 and 3 and implements them within Galaxy-P, a web-based platform for the flexible construction of NGS and proteomic workflows.

1 INTRODUCTION

1.1 Proteomic variation

The foremost challenge in the field of genomics is understanding the connection between genotype and phenotype. Since the sequencing of the human genome and advances in DNA sequencing technologies, an incredible number of genomic and transcriptomic variations have been discovered in diverse species, tissues, and cell-types. These variations include single nucleotide polymorphisms, insertions, deletions, inversions, gene fusions, alternatively spliced mRNA, and changes in nucleotide sequence from RNA editing (Figure 1.1). If the variation resides within or overlaps a protein-coding region, it can change the sequence of the encoded protein (Figure 1.2), which could have profounds effect on the phenotype or disease state of an organism. Therefore, it has become increasingly important to not only detect and characterize variations at the level of the genome and transcriptome but to also directly detect these variations at the level of proteins, where the effect of the variation typically plays out within the cellular system. In other words, characterization of protein variations will be crucial to help understand the link between genotype and phenotype.

1.2 MS-based proteomics

In 1984, John Fenn discovered that large biomolecules could be introduced into the gas phase through the process of electrospray ionization, allowing peptides and proteins to be readily analyzed by a mass spectrometer [1]. Over the next few decades, mass spectrometry (MS) instrumentation and sample preparation methods steadily advanced in throughput and utility [2]. Owing to these advances, MS-based proteomics has become the preeminent method for the identification and quantification of proteins in a sample. One of the most popular MS-based proteomics methods is the bottom-up or shotgun proteomics strategy, in which peptides are detected in a high throughput manner. In bottom-up proteomics, proteins

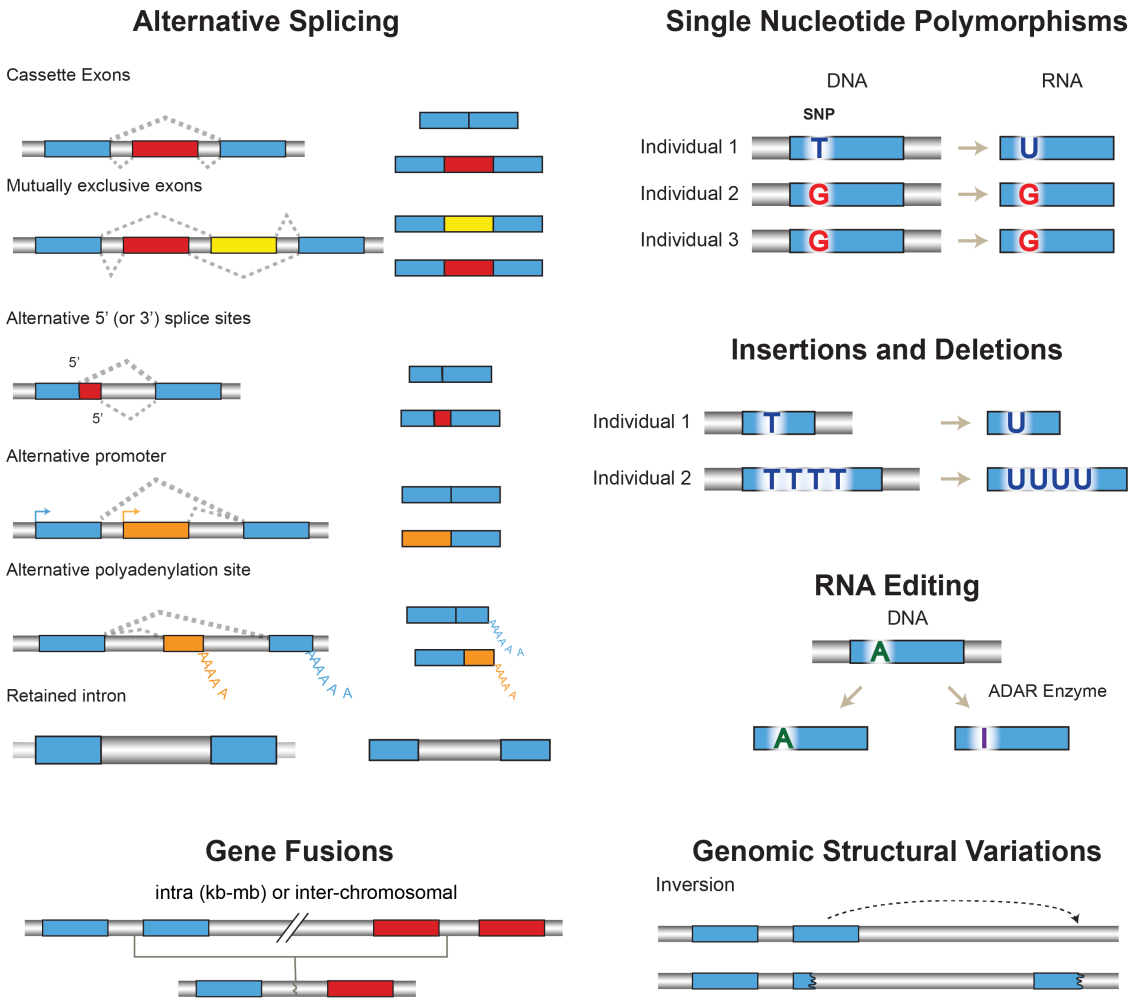

Figure 1.1. Examples of sequence and structural variations found in genomes and transcriptomes. Nucleotide sequence variations within the genome include single nucleotide polymorphisms, insertions, and deletions. Large chromosomal structural rearrangements that contain breakpoints within protein-coding regions, such as with gene fusion events or inversions, can dramatically change protein sequence. Other sources of variations occur at the RNA-level. Alternative splicing of exons produces several distinct protein products from one genetic locus. RNA editing changes nucleotide sequences post-transcriptionally. Given the astounding amount of variation expressed in higher eukaryotes, it will become increasingly important to understand their role and functional importance in various biological contexts.

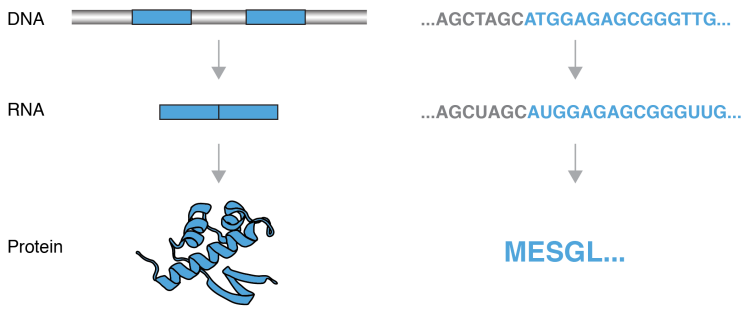

Figure 1.2. The central dogma of molecular biology. Variations that occur at the level of DNA or RNA can influence the coding potential of proteins.

are extracted by lysing cells or tissues with a detergent or chaotrope-containing lysis buffer. Next, an enzyme such as trypsin, which hydrolyzes amide bonds C-terminal to lysines and arginines, is used to digest the protein into peptides. Note that peptides are more amenable to LC-MS analysis as compared to intact proteins due to their favorable physicochemical properties (solubility, chromatographic separability, gas-phase charge state, etc.). The resultant peptides typically comprise a complex mixture of millions of distinct peptide sequences thus peptides are chromatographically fractionated or separated to reduce sample complexity before MS analysis. Peptides are introduced into the mass spectrometer through electrospray ionization, where the peptides are driven from the liquid to gas phase through application of a voltage gradient. Charged peptides in the gas-phase are then directed into the mass spectrometer and their mass-to-charge (m/z) ratio is measured with a mass analyzer, such as with ion-cyclotron resonance or an Orbitrap[3, 4]. For the purposes of high-throughput sequencing of peptide mixtures, mass spectrometers are frequently operated in data-dependent mode, where peptide precursor mass-to-charges are first detected in a full-scan (i.e. MS^1 scan) and then iteratively selected for fragmentation via collisionally induced dissociation. During dissociation, peptide ions break along the amide backbone generating b and y ions [5]. The mass-to-charges of these fragments are measured in a tandem mass spectral scan (i.e. MS^2 scan). After collection of the set of full and tandem

mass spectra, peptides may be identified through database search methods. Figure 1.3 depicts the MS-based proteomics workflow just described.

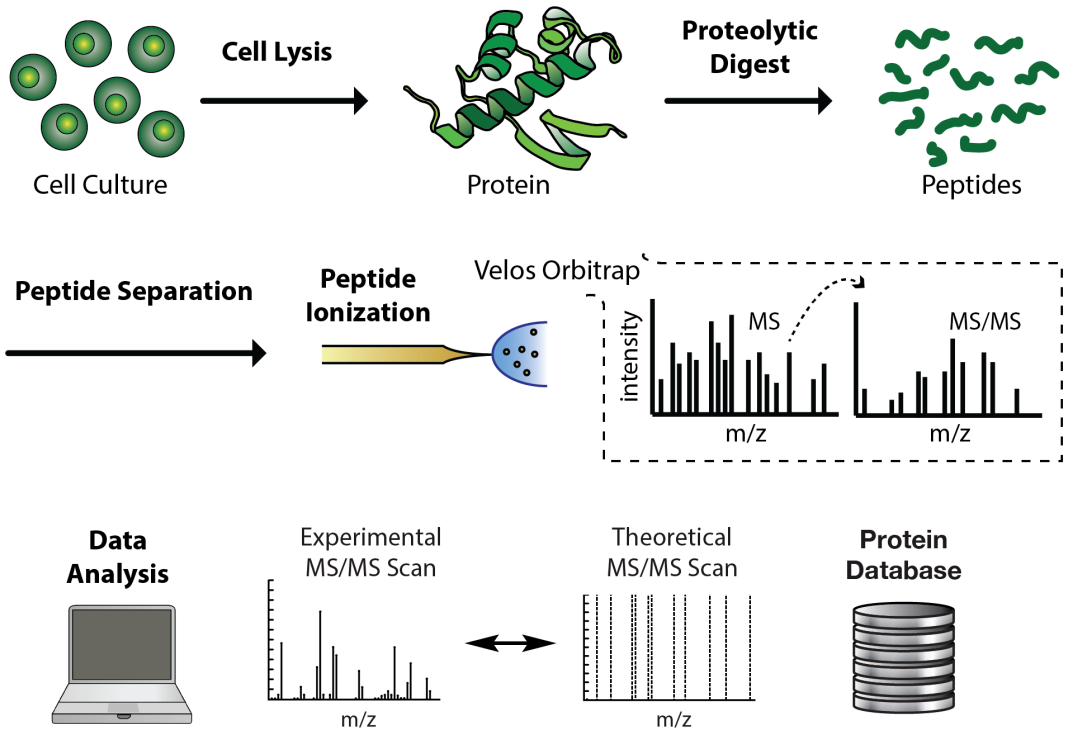

Figure 1.3. Mass spectrometry-based proteomics workflow. A typical experiment involves several steps: extraction of proteins from cells or tissues, enzymatic digestion of the proteins into peptides, chromatographic separation of peptides, introduction of peptides into the MS instrument through electrospray ionization, and sequencing of peptides through iterative isolation and fragmentation of peptide ions. Post-analysis database searching identifies peptides, which can be used to support protein identifications.

1.3 MS DATABASE SEARCHING

MS-based proteomics experiments can produce millions of peptide mass spectra. This sheer volume of data precludes manual analysis; therefore, computational methods have been crucial for matching peptides to their corresponding tandem

mass spectra. The most widely employed computational method for the identification of peptides is the MS database search strategy [6]. Here, a reference proteome for the species under study is obtained and the proteins are digested *in silico* to generate a list of all candidate peptides that could have been present in the sample. For each candidate peptide, theoretical tandem mass spectra containing all possible peptide fragment mass-to-charges is generated. Following this routine, two sets of tandem mass spectra are produced: the set of experimental spectra that was collected on the mass spectrometer and the set of theoretical spectra that was derived from the candidate peptides. Each experimental spectrum is compared with every other theoretical spectrum and the highest scoring experimental-theoretical spectral pair is considered a peptide spectrum match (PSM). Thus, it is the scoring of the degree of match between each spectrum pair that is central to all database search methods. For instance, the first such algorithm reported, SEQUEST, scores the degree of match using a cross-correlation function [7]. Once the experimental-theoretical spectrum comparisons are completed for every experimental spectrum in the whole set (e.g. group of LC-MS runs) and PSMs are generated, statistical validation may be done to decide which PSMs are counted as a peptide identification (Figure 1.4).

1.4 Target-decoy searching

The target-decoy strategy is a widely employed method for estimation of the false discovery rate in a group of PSMs [8]. MS searching suffers from the problem of high-dimensionality —a single MS analysis generates large sets of tandem mass spectra that need to be compared with even larger sets of theoretical tandem mass spectra. Because there are multiple-hypotheses being tested, it is difficult to estimate the number of false positives that occur. The target-decoy method is an elegant solution to this problem. In this strategy, experimental mass spectra are compared with not only the target proteome (i.e. reference protein database) but also a decoy proteome in which the protein sequences are reversed or shuffled to represent spurious sequences *not* present in the sample (Figure 1.4). The decoy PSMs provide

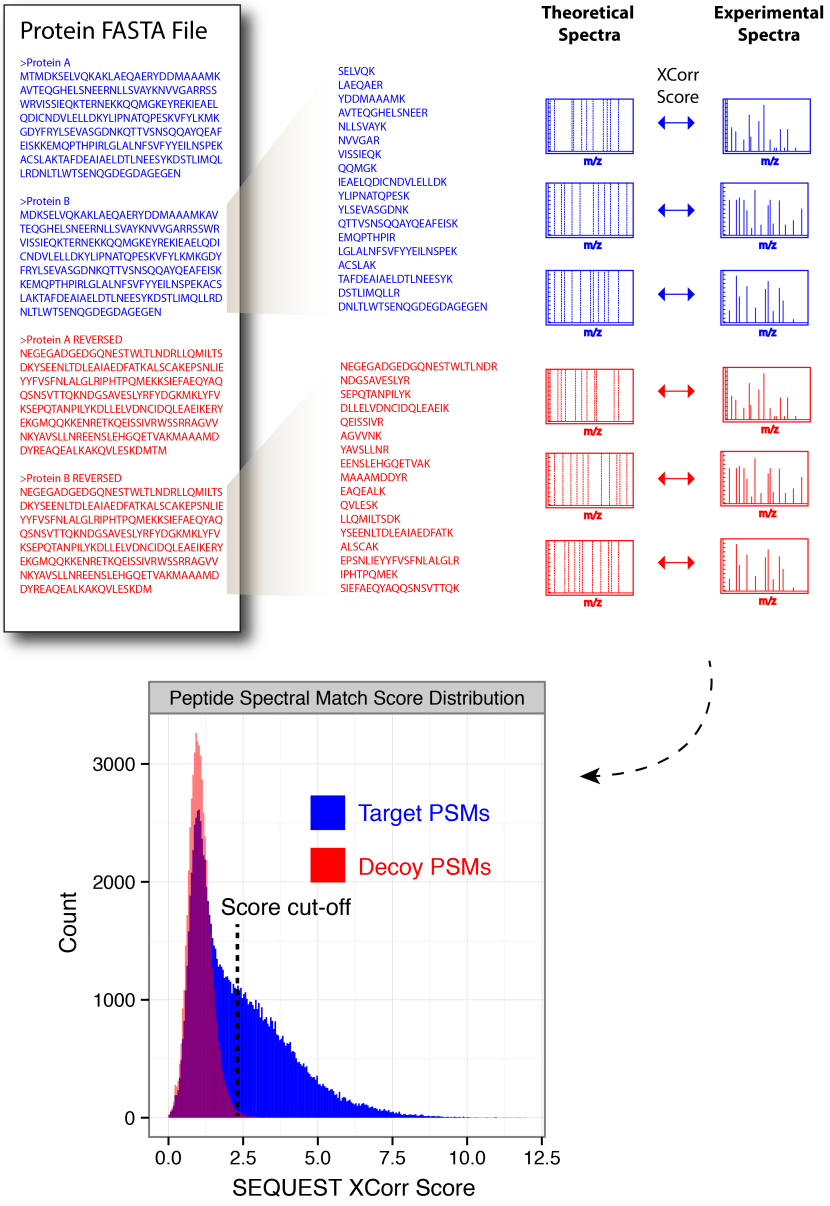

Figure 1.4. MS database searching combined with target-decoy validation. In the target-decoy search method, the set of theoretical tandem mass spectra derived from peptides generated from *in silico* digestion of both the forward (target, shown in blue) and reverse (decoy, shown in red) protein sequences are compared to the experimental spectra and scored. Once the scoring is complete, a score distribution for the target (containing false positives and true positives) and decoy (assumed to contain only false positives) can be utilized to set appropriate peptide score cut-offs.

a key piece of information: the score distribution of false positives distributed among the target PSMs. The score distribution for decoy PSMs (in red, Figure 1.4) are plotted along with the score distribution for target PSMs (in blue, Figure 1.4) and the false discovery rate for groups of target PSMs passing certain score thresholds may be estimated (Figure 1.4) [8]. The success of this method can be attributed to its versatility as it has been successfully applied to data collected from various proteomics workflows using different instruments.

The target-decoy method makes the assumption that the target proteome appropriately represents the sample proteome both in size and composition. This requirement shall become increasingly important in understanding the interplay between database size and peptide identification quality.

1.5 RNA-Seq

Introduced in 2009, RNA sequencing (i.e. RNA-Seq) has dramatically increased the ability to characterize the transcriptome [9]. RNA-Seq experiments start with total RNA isolated from cells or tissues using either phenol-chloroform or column-based extraction. Next, poly(dT) beads are used to isolate mRNAs that contain poly(A) tails as these mRNAs are likely protein-coding. The mRNAs are then fragmented by addition of a divalent cation and heat to catalyze strand breaks randomly along each mRNA. Using random DNA hexamer primers, a reverse transcriptase is added to the mRNA fragments to synthesize short cDNA sequences. These cDNAs are amplified via PCR and then further processed using vendor-specific sample preparation protocols. For example, Illumina sequencing protocols require the ligation of Illumina-specific adapters which are used to immobilize cDNAs onto a flowcell. Most next generation sequencing instruments operate on the principle of template-based strand synthesis, in which a DNA polymerase incorporates, one at a time, a fluorescently labeled nucleotide and the emitted signal is recorded on a CCD camera. A short (30-200 bp) stretch of cDNA is sequenced and each sequence fragment is called an RNA-Seq read. Figure 1.5 depicts the Illumina-based workflow.

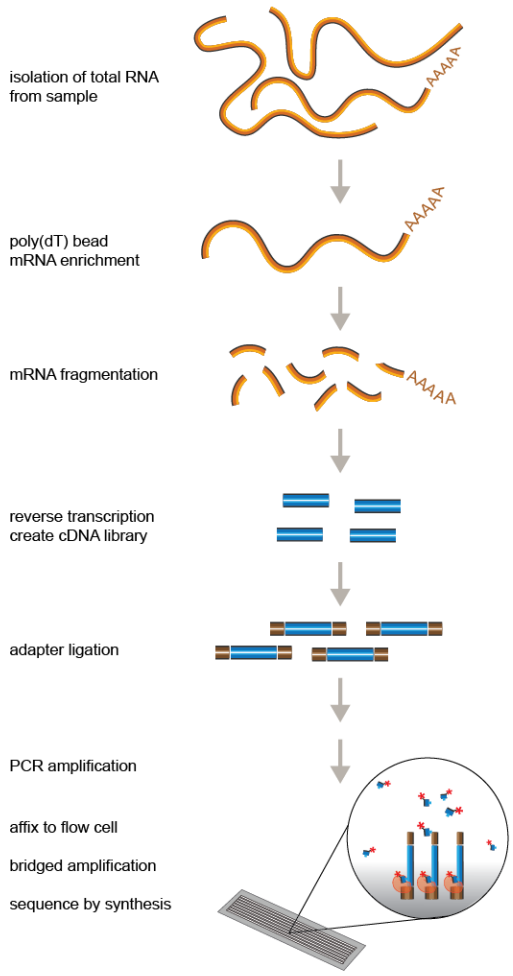

Figure 1.5. Overview of the RNA-Seq workflow. Total RNA is extracted from the sample of interest and poly(A)+ mRNA are enriched using poly(dT) beads. mRNA is fragmented and reverse transcribed to create a cDNA library (shown in blue). In the case of Illumina-based sequencing, specialized adapters are ligated to each cDNA and PCR amplified (brown). The cDNAs are randomly attached to a flowcell through DNA hybridization. Each cDNA undergoes bridged amplification, creating distinct clusters of cDNAs each with identical sequence to the original cDNA. The four bases (A, T, C, G) are flowed over the cell one at a time and DNA polymerases incorporate a single complementary nucleotide during each A-T-C-G cycle. In this sequencing-by-synthesis method, the pattern and order of cluster signal emission corresponds to cDNA sequence.

1.6 RNA-Seq bioinformatic analysis

A typical RNA-Seq experiment can generate several million to a billion short RNA-Seq reads, representing the RNA fragments sampled from a cell's transcriptome [10]. Today's sequencers typically produce RNA-Seq reads with lengths of up to a few hundred nucleotides, and because of the short read lengths, transcripts must be fragmented before sequencing in order to gain sequence coverage evenly across the whole transcript. Because RNA-Seq datasets consist of random subsequences, not full-length transcripts, computational tools play a crucial role in their analyses.

There are myriad computational and bioinformatic programs and they can generally be categorized under a few core tasks: read alignment, transcript reconstruction, variant calling (e.g. single nucleotide polymorphisms, indels), and transcript quantification [11].

In *read alignment*, each RNA-Seq read is aligned to a reference genome, where the optimal alignment is the position for which there is highest correspondence between the nucleotide sequences of the RNA-Seq read and reference genome. The reference genome, in effect, serves as a scaffold through which related RNA-Seq reads may be grouped together. Reads that overlap or are aligned close to each other on the genome may be derived from the same genomic region or from the same transcript. Not all reads will align perfectly to the genome because they may span multiple exons. Spliced-aware aligners will account for the alternative splicing of transcripts by allowing for "splits" between reads during alignment. *Transcript reconstruction* is the process of inferring full-length transcript sequences from aligned RNA-Seq reads. Similar to transcript reconstruction is *de novo* assembly, which still infer transcript sequences but without the help of a reference genome. Once the transcript sequences have been assembled, *variant calling* methods may be used to detect small sequence differences like single nucleotide polymorphisms (SNPs). Finally, transcriptional abundance may be measured using *transcript quantification* tools that count the number of RNA-Seq reads mapping up to each transcript and, employing various normalization routines, estimate the concentrations of each transcript.

1.7 The human reference proteome

MS database searching relies on the completeness and accuracy of the reference proteome for the identification of peptides. Consequently, a peptide sequence being absent from the reference proteome precludes detection of that peptide using MS database search methods, even if that peptide (protein) is expressed in the sample. The tandem mass spectrum of the missing peptide may even be incorrectly assigned to the wrong peptide sequence, creating a false positive peptide identification.

Following the sequencing of the human genome, various organizations have worked to annotate genes and their protein products [12]. For instance, the Ensembl team developed sensitive *ab initio* gene prediction algorithms for the computational prediction of protein coding regions across the genome and within expressed sequence tags (ESTs) listed in sequencing repositories [13]. The consensus coding sequence (CCDS) project, on the other hand, employed comparative genomics for protein annotation, specifically examining protein-coding sequences that were conserved between human and mouse genomes [14]. The Swiss Institute for Bioinformatics (SIB) and the UniProt Swiss-Prot/Trembl group specifically focuses on protein curation, designating a full-length protein sequence as their fundamental annotation unit instead of a genomic locus [15]. Being protein-centric curators, the Swiss-Prot team has amassed the most popular and reliable human reference proteome with each protein entry linked to highly relevant functional annotations.

Regardless of the organization providing the human reference proteome, most curators of protein entries strive to build a complete and accurate proteome. Interestingly, they are also guided by two principles: assembling an "average" proteome and minimizing sequence redundancy. First, the principle of assembling an "average" proteome arises from the fact that innumerable protein variations exist in different individuals, tissue-types, and cell-types. There is no one reference proteome, but, in fact, millions of distinct proteomes. Therefore, protein curators must choose which protein form to include in the database. Usually the best set of proteins to include for a gene represents a composite of all the possible proteins. For example, UniProt incorporates only the most relevant protein forms, defined as

those most frequently detected in sequencing projects or supporting literature [16]. Second, the principle of minimizing sequence redundancy arises from the need to create a reference proteome that is compact, with minimal redundancy or overlap in protein sequences. Maintaining non-redundancy was important in the construction of early protein databases where the same or similar protein sequences were listed multiple times. Today, this principle is reflected through UniProt's curation process where the decision to include a newly discovered protein sequence takes into account the degree to which a novel protein form diverges from the "canonical" protein sequence . For example, subtle splicing events that cause a protein to differ by a few amino acids are seldom listed in the protein reference database. Overall, the choices of assembling an "average" human proteome with minimal sequence redundancy helps maintain a consistent, stable reference proteome. However, this means that reference proteomes do not necessarily reflect the actual proteomes expressed in certain cells or individuals.

1.8 Augmenting the reference proteome with customized databases

The initial candidate protein sequences that protein curators and bioinformatic programs must start with to build a reference proteome must be derived from translation of existing DNA and RNA sequence data, which is typically obtained from large-scale sequencing projects (A few protein sequences have been determined through Edman degradation experiments, but this is true for only a small number of protein entries.) [12]. For that reason, protein sequence annotation depends on the availability of nucleotide sequence data. This dependence is reflected in the parallel evolution of protein databases and nucleotide sequencing technologies (Figure 1.6).

In 1995, databases of expressed sequence tags (ESTs) —sequences corresponding to partial or full-length mRNAs —were used by Yates and colleagues to automatically match tandem mass spectra to peptide sequences [17]. By the time the draft

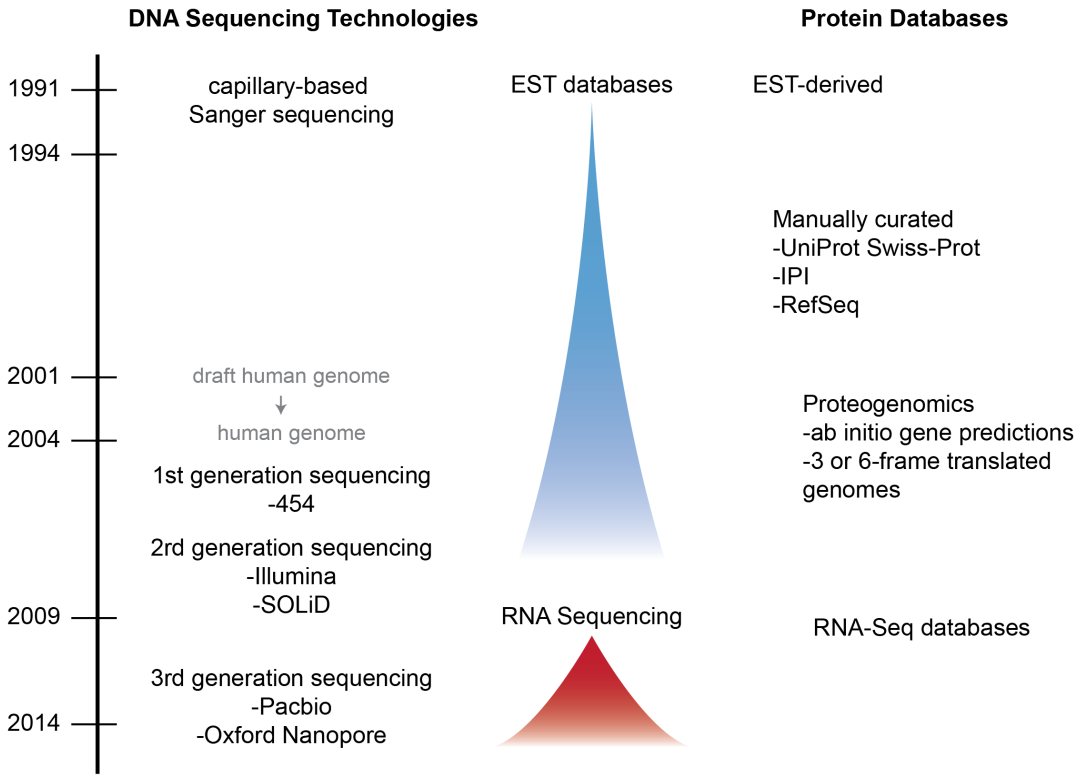

Figure 1.6. Evolution of DNA sequencing technologies and protein databases. Protein reference sequences are typically derived from translation of nucleotide sequencing databases, thus the protein reference databases largely rely on available nucleotide sequence repositories.

human reference genome was finished in 2001, protein curators had available to them large human EST databases, genomic sequence, and a wealth of genomic sequence data from small-scale biological studies reported in the literature [12]. Using these various sources of data along with improved bioinformatic algorithms, several organizations were working to annotate the human proteome. During this time a few researchers employed auxiliary nucleotide sequence datasets to increase the number of novel peptides detected by MS. Edwards created a computational approach that compresses the entire human EST database into a compact, MS-

searchable database [18]. Schandorff et. al. utilized SNP databases derived from EST data to allow for MS-detection of single amino acid polymorphisms (SAPs) [19]. Tanner et. al. proposed utilizing the sequence of full genomes and introduced the concept of searching MS data against 6-frame translated reference genomes, which gave birth to the field of proteogenomics [20, 21].

1.9 RNA-Seq-derived protein databases

RNA-Seq has revolutionized transcriptomic analysis. It enables high-throughput and deep sequencing of an entire transcriptome from a single sample, including each transcripts' sequence and structure (i.e. transcriptional start/end sites; exon-exon connectivites) [10]. Not surprisingly, various research groups have utilized RNA-Seq data to create customized proteomic databases [22–27]. Here, both RNA-Seq and MS-based proteomics data are collected from the same or similar groups of samples. The detected mRNA sequences provided through RNA-Seq are translated into protein and compiled into customized protein databases. The RNA-Seq-derived sequences may include putative splice variants, SAPs, and transcripts corresponding to novel genes.

Ning et. al. reported the earliest example of incorporating RNA-Seq data for the creation of a protein database. They used RNA-Seq data to detect novel exon-exon junctions that were expressed at the transcript level. These junctions were translated into protein and compiled into a database for searching against a mitochondrial MS dataset; seven novel peptides were discovered [22]. This preliminary study hinted that there were indeed sample-specific peptides not represented in the reference proteome and that RNA-Seq could allow for direct, empirical measurement of what peptide sequences may be expressed. **Chapter 2** describes a bioinformatic method which uses a spliced-aware aligner, Tophat, to detect novel alternative splice junctions without prior knowledge of existing exon locations [28]. This method enabled the detection of 57 novel splice junction peptides and showed that many alternative splicing patterns have not been fully characterized in humans.

Wang et. al. showed that RNA-Seq datasets could be mined for single nucleotide

polymorphisms and thus used to create a database of possible amino acid polymorphisms (SAPs) [24]. This resulted in the identification of peptides containing SAPs, an average of 38 per cell line. **Chapter 3**, extends this work by conducting an in-depth analysis of SAPs for a single human cell line. An unprecedented 695 SAP-containing peptides were identified, which allowed for investigation of various aspects of SAP detection including allele-specific expression.

For species without a reference genome, RNA-Seq reads may also be converted into full-length transcripts using *de novo* assemblers. In recent years, researchers have showed that *de novo* assembled transcripts can provide putative protein sequences that may be expressed in non-model organisms [25, 26]. With the advent of third generation sequencing technologies, such as the single-molecule sequencing platforms provided by Pacific Biosystems, the next years may usher in a new era where the concept of "reference proteome" may be supplanted by the concept of proteomes built for each cell-type, tissue-type, or even individual.

Despite the many benefits of using RNA-Seq to create customized databases, the main limitation has been in the complex computational workflows that are needed for RNA-Seq analysis and MS database creation. **Chapter 4** describes the implementation of the bioinformatic workflows described in **Chapter 2** and **Chapter 3** within Galaxy-P, so that the proteomics community may build upon these methods as new sequencing and proteomics technologies arise. Galaxy-P, or Galaxy for Proteomics, is an extension of Galaxy, a popular web-based bioinformatic platform that allows for streamlined analysis of sequencing data [29].

References

[1] J. B. Fenn et al. Electrospray ionization for mass-spectrometry of large biomolecules. *Science* 246.4926 (1989), pp. 64–71.

[2] R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature* 422.6928 (2003), pp. 198–207.

[3] A. G. Marshall, C. L. Hendrickson, and G. S. Jackson. Fourier transform ion cyclotron resonance mass spectrometry: A primer. *Mass Spectrometry Reviews* 17.1 (1998), pp. 1–35.

[4] Q. Z. Hu et al. The Orbitrap: a new mass spectrometer. *Journal of Mass Spectrometry* 40.4 (2005), pp. 430–443.

[5] H. Steen and M. Mann. The ABC's (and XYZ's) of peptide sequencing. *Nature Reviews Molecular Cell Biology* 5.9 (2004), pp. 699–711.

[6] A. I. Nesvizhskii. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteomics* 73.11 (2010), pp. 2092–2123.

[7] J. K. Eng, A. L. McCormack, and J. R. Yates. An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *Journal of the American Society for Mass Spectrometry* 5.11 (1994), pp. 976–989.

[8] J. E. Elias and S. P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* 4.3 (2007), pp. 207–214.

[9] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10.1 (2009), pp. 57–63.

[10] F. Ozsolak and P. M. Milos. RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics* 12.2 (2011), pp. 87–98.

[11] Manuel Garber et al. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Meth* 8.6 (2011), pp. 469–477.

[12] M. Yandell and D. Ence. A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics* 13.5 (2012), pp. 329–342.

[13] T. Hubbard et al. The Ensembl genome database project. *Nucleic Acids Research* 30.1 (2002), pp. 38–41.

[14] K. D. Pruitt et al. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Research* 19.7 (2009), pp. 1316–1323.

[15] R. Apweiler et al. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research* 32 (2004), pp. D115–D119.

[16] C. O'Donovan and R. Apweiler. "A Guide to UniProt for Protein Scientists". *Bioinformatics for Comparative Proteomics*. Ed. by C. H. Wu and C. Chen. Vol. 694. Methods in Molecular Biology. Totowa: Humana Press Inc, 2011, pp. 25–35.

[17] John R. Yates, Jimmy K. Eng, and Ashley L. McCormack. Mining Genomes: Correlating Tandem Mass Spectra of Modified and Unmodified Peptides to Sequences in Nucleotide Databases. *Analytical Chemistry* 67.18 (1995), pp. 3202–3210.

[18] N. J. Edwards. Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. *Molecular Systems Biology* 3 (2007).

[19] S. Schandorff et al. A mass spectrometry-friendly database for cSNP identification. *Nature Methods* 4.6 (2007), pp. 465–466.

[20] Stephen Tanner et al. Improving gene annotation using peptide mass spectrometry. *Genome Research* 17.2 (2007), pp. 231–239.

[21] N. Castellana and V. Bafna. Proteogenomics to discover the full coding content of genomes: A computational perspective. *Journal of Proteomics* 73.11 (2010), pp. 2124–2135.

[22] K. Ning and A. I. Nesvizhskii. The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment. *Bmc Bioinformatics* 11 Suppl 11 (2010), S14.

[23] C. Adamidi et al. De novo assembly and validation of planaria transcriptome by massive parallel sequencing and shotgun proteomics. *Genome Research* 21.7 (2011), pp. 1193–1200.

[24] Xiaojing Wang et al. Protein Identification Using Customized Protein Sequence Databases Derived from RNA-Seq Data. *Journal of Proteome Research* (2011).

[25] Vanessa C. Evans et al. De novo derivation of proteomes from transcriptomes for transcript and protein identification. *Nat Meth* advance online publication (2012).

[26] G. Lopez-Casado et al. Enabling proteomic studies with RNA-Seq: The proteome of tomato pollen as a test case. *Proteomics* 12.6 (2012), pp. 761–774.

[27] Sunghee Woo et al. Proteogenomic database construction driven from large scale RNA-seq data. *Journal of Proteome Research* (2013).

[28] C. Trapnell, L. Pachter, and S. L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25.9 (2009), pp. 1105–1111.

[29] Daniel Blankenberg et al. "Galaxy: A Web-Based Genome Analysis Tool for Experimentalists". *Current Protocols in Molecular Biology*. John Wiley & Sons, Inc., 2010.

2 DISCOVERY AND MASS SPECTROMETRIC ANALYSIS OF NOVEL SPLICE-JUNCTION PEPTIDES USING RNA-SEQ

This chapter has been published in *Molecular & Cellular Proteomics*:

Sheynkman, G. M., Shortreed, M. R., Frey, B. L., and Smith, L. M. (2013) Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. *Mol Cell Proteomics* 12, 2341-2353

2.1 Abstract

Human proteomic databases required for MS peptide identification are frequently updated and carefully curated, yet are still incomplete because it has been challenging to acquire every protein sequence from the diverse assemblage of proteoforms expressed in every tissue and cell type. In particular, alternative splicing has been shown to be a major source of this cell-specific proteomic variation. Many new alternative splice forms have been detected at the transcript level using next generation sequencing (NGS) methods, especially RNA-Seq, but it is not known how many of these transcripts are being translated.

Leveraging the unprecedented capabilities of NGS, we collected RNA-Seq and proteomics data from the same cell population (Jurkat cells) and created a bioinformatics pipeline that builds customized databases for the discovery of novel splice-junction peptides. Eighty million paired-end Illumina reads and ~500,000 tandem mass spectra were used to identify 12,873 transcripts (19,320 including isoforms) and 6,810 proteins. We developed a bioinformatics workflow to retrieve high-confidence, novel splice junction sequences from the RNA data, translate these sequences into the analogous polypeptide sequence, and create a customized splice junction database for MS searching. Based on the RefSeq gene models,

we detected 136,123 annotated and 144,818 unannotated transcript junctions. Of those, 24,834 unannotated junctions passed various quality filters (e.g. minimum read depth) and these entries were translated into 33,589 polypeptide sequences and used for database searching. We discovered 57 splice junction peptides not present in the Uniprot-Trembl proteomic database comprising an array of different splicing events, including skipped exons, alternative donors and acceptors, and non-canonical transcriptional start sites. To our knowledge this is the first example of using sample-specific RNA-Seq data to create a splice-junction database and discover new peptides resulting from alternative splicing.

2.2 Introduction

Mass spectrometry-based proteomics relies on accurate databases to identify and quantify proteins, including those derived from splice variants, indels, and single nucleotide variants (SNVs) [1]. Most computational search algorithms detect peptides by scoring the degree of similarity between in silico derived and experimental peptide spectra, and thus can only identify peptides that are present in the proteomic database. If the polypeptide sequence is not present in the database used for searching, even if the peptide is present in the sample, it will fail to be detected.

Human proteomic databases used for mass spectrometric peptide identification are frequently updated and carefully curated, yet are still incomplete. Despite efforts to comprehensively annotate every gene product, there are still many undiscovered proteoforms [2] because the complete human proteome—the aggregate of all protein products expressed in every tissue, cell, and cellular state—turns out to be vastly more complex than was predicted [3–5]. Furthermore, each cell or tissue-type may express a unique subset of all possible proteoforms, many of which may not be represented in existing proteomic databases. These databases are assembled from multiple datasets originating from an assortment of different human tissue and cell samples [6–11].

In recent years, alternative splicing has been shown to be a major source of cell-specific proteomic variation in humans [3, 4, 12]. Human genes are composed

of introns and protein-coding exons; a protein machine, the spliceosome, removes introns from pre-mRNAs, joining exons to form a mature transcript ready for translation. Since exons can be joined in various configurations, one gene typically produces a "canonical" protein (defined as the most abundant form of the protein) as well as one or more alternatively spliced protein products, which are often thought to have modulated or altered biological function [13–16]. Many alternative splice variants have been detected at the transcript level using next generation sequencing methods, especially RNA-Seq. However, it is not known exactly how many of these newly discovered alternatively spliced transcripts are being translated and if these translated products are functional.

Several approaches have been employed in the last decade to expand detection of alternatively spliced proteins using mass spectrometry. Initial approaches searched proteomic data against databases containing splice variant sequences and then confirmed the translation of a spliced sequence by detecting a peptide unique to that form [17–26]. Other approaches expanded the number of alternatively spliced sequences beyond entries present in databases by constructing exon-exon databases. In this approach, exon coordinates are first determined by obtaining exon sequences from databases such as Ensembl or by using ab initio computational algorithms to predict the location of putative exon boundaries. Next, these exon sequences are assembled into all theoretical exon-exon (and exon-intron) combinations, and then the sequences are translated into polypeptide sequences and used for MS-based searching to discover novel splice variant peptides [27–30]. To extend this approach, several research groups have restricted their exon-exon database to include only those sequences corroborated with transcript expression data [1, 31, 32], thereby eliminating spurious sequences. Two other approaches developed include a method that directly translates RNA sequence from expressed sequence tag (EST) contigs [33–36] and a proteogenomics strategy that uses the genome as a template for peptide sequence alignment [37, 38].

Several of the above methods expand proteomic databases to include entries for putative or experimentally confirmed splice variants; however, unbounded addition of more and more splice variants compiled from thousands of human cell-

types is not the preferred solution. MS searching with inordinately large databases containing many more proteins than actually present in the sample causes decreased peptide identification sensitivity (as the probability of spurious spectral matches to in silico peptide spectra is greater) [39, 40], complications in protein parsimony (from including many redundant sequences) [41, 42], and longer analysis and search times.

Given the unprecedented advances of next generation sequencing and the maturation of RNA-Seq—longer read length, improved accuracy, increased affordability, better software—the whole transcriptome of a single sample can now be sequenced in a matter of days. As a result, all of the alternative splice junctions expressed in a single cell-type can be determined empirically. Many of the aforementioned splice detection methods rely on gene prediction programs, where reliable detection of splice forms is a challenge, or the use of data from public repositories, an amalgamation of data from multiple samples that may not reflect the splicing patterns in a given cell-type. Because RNA-Seq methods are increasingly accessible to proteomicists and these methods can empirically determine the full spectrum of alternative splicing in a sample, there is a need for bioinformatic methods that provide sample-specific, splice junction proteomic sequences from RNA-Seq data for mass spectrometry database searching.

Though the focus of this paper is the study of alternative splice junctions, other bioinformatics strategies to extract information from RNA-Seq data have been employed to create customized mass spectrometry databases. These include reducing a database to only include sequences with transcript expression evidence [39], including fusion or chimeric sequences (44), incorporating non-synonymous single nucleotide polymorphism (SNP) or single nucleotide variant (SNV) sequences [39], and, for non-model systems, building a proteomic database from de novo assembled transcripts [43, 44]. The advent of next generation proteomics will most certainly arrive when all these sources of transcriptomic variation can be seamlessly incorporated into sample-specific proteomic databases.

We have developed a method to create a sample-specific splice junction database from RNA-Seq data and used it to discover novel splice junction peptides. We col-

lected both RNA-Seq and proteomic data from the same cell population (Jurkat cells) and identified 12,873 transcripts and 6,810 proteins. We developed a bioinformatics pipeline to retrieve high-confidence, novel splice junction sequences, translate these sequences into the analogous polypeptide sequences, and then create customized splice-sequence databases that allow for novel splice junction discovery. We discovered 57 splice junction peptides not present in the Uniprot-Trembl proteomic database using appropriately stringent MS search parameters and post-processing steps, including the use of a conservative 1% local false discovery rate and manual validation of junction peptide MS2 spectra. To our knowledge this is the first example of using sample-specific RNA-Seq data to discover new peptides resulting from alternative splicing.

2.3 Experimental Procedures

Cell Culture

The Jurkat cell line (TIB-152) was obtained from the American Type Culture Collection (ATCC, Manassas, VA). Jurkat cell culture was grown in 10% Fetal Bovine Serum and 90% RPMI-1640 buffer (ATCC, Manassas, VA) at 37°C. Cell concentration was measured using the TC10 Automated Cell Counter system (BioRad, Hercules, CA), which was validated via hemocytometer counting. Before harvesting, cells were grown to approximately 1.3×10^6 cells/mL and had 95%+ viability as measured with the trypan blue assay.

Proteomic Sample Preparation and Analysis

Approximately 25 mL of Jurkat cell suspension was centrifuged at 180g at 4°C for 10 minutes. After removal of the supernatant, cells were resuspended in an equivalent volume of ice-cold PBS buffer (Invitrogen, Grand Island, NY) and centrifuged again. This step was repeated twice and the final pellet was stored at -80°C. For cell lysis, pellets were thawed on ice and a volume of SDT lysis buffer

equaling 2/3 the volume of the cell pellet was added. The pellet was pipetted up and down to assist in its solubilization, followed by incubation of the solution at 95°C for 5 minutes. The SDT lysis buffer consisted of 4% SDS, 500 mM Tris-HCl (pH 7.4), and 180 mM dithreothreitol (DTT) (all reagents from Sigma-Aldrich, St. Louis, MO). The resulting lysate was sonicated (power level between 2 and 3) on ice—alternating between 30 seconds on and 30 seconds off—for 3-5 minutes until the viscous chromatin was solubilized and lysate had an aqueous consistency for improved sample pipetting during later steps (Misonix Sonicator XL2015, Misonix microtip PN/418, Farmingdale, NY). Protein content was measured using the 660 nm Protein Assay and the Ionic Detergent Compatibility Reagent (Pierce, Rockford, IL) to allow for accurate protein quantification in the presence of SDS.

Detergents and salts in the sample were removed and the protein was subjected to tryptic digestion by following the Filter-Aided Sample Preparation or FASP protocol developed by Wisniewski et. al. [45]. Five aliquots of lysate containing approximately 150 μg of protein were each added to a 100K MW Amicon Ultra filter (Millipore, Billerica, MA). After multiple FASP wash steps, reduction, and alkylation, trypsin was added directly to the filters (50:1 protein:trypsin w/w) and digested overnight at 37°C. The next morning, filters were centrifuged at 14,000 g for 15 minutes and the amount of peptide recovered was assessed via the Nanodrop UV-Vis spectrometer (Thermo Fisher Scientific, Wilmington, DE).

Approximately 500 μg of tryptic peptide digest was fractionated using high pH reverse-phase chromatography on a Shimadzu HPLC system (LC-10AD, SCL-10A VP, SPD-10A VP, Shimadzu, Columbia, MD) and a Phenomenex C18 Gemini 3μ, 110Å, 3.0×150mm column (Phenomenex, Torrance, CA). The high pH method was adopted from Gillar et. al. [46]. Mobile phase A (MPA) was 20 mM ammonium formate, pH 10, and B (MPB) was 20 mM ammonium formate, pH 10, in 70% acetonitrile. The HPLC flow was 0.5 mL/min and the gradient is as follows: 0% MPB isocratic for 15 minutes (trapping step), linear ramp to 100% MPB over 60 minutes, hold at 100% MPB for 5 minutes, to 0% MPB over 2 minutes, and equilibration at 0% MPB for 20 minutes. Fractions were collected every minute using a Gilson 203 fraction collector (Gilson, Middleton, WI) for a total of 27 fractions collected within

the range of peptide elution as discernable from the UV-Vis trace. Fractions were dried down using vacuum centrifugal concentration (Savant SpeedVac, Thermo, Pittsburgh, PA) and stored at -80°C.

Each of the lyophilized fractions generated from the high pH LC separation was reconstituted in sample solution consisting of 2% acetonitrile and 0.2% formic acid in water and then chromatographically separated on a nanoAquity LC system (Waters, Milford, MA) using a 20 cm reverse phase capillary column (100 μm i.d.) packed with 3 μm MAGIC aqC18 beads (Bruker-Michrom, Auburn, CA). Mobile phase A was 0.2% formic acid in water and B was 0.2% formic acid in acetonitrile. The full HPLC method was 180 minutes long and included a 90 minute gradient. The mass spectrometric analysis was conducted on a Velos-Orbitrap mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) operating in data-dependent mode. A full scan (300-1500 m/z) was collected at a resolution of 30,000 followed by fragmentation of the top ten precursor peptides, with +2 charge or higher, in HCD mode (collision energy=40) and analysis of the tandem mass spectra in the Orbitrap at a resolution of 7,500. Precursor fragmentation repeat count was set to two and the dynamic exclusion was set to 60 seconds. XCalibur software version #2.1.0 was used for data collection.

RNA-Seq Analysis

RNA was extracted from Jurkat cells using Trizol Reagent (Life Technologies, Grand Island, NY). 2 mL of Jurkat culture (~2.6×10^6 cells) was centrifuged at 110 g and 4°C for 5 minutes. After removal of the supernatant, 1 mL of Trizol reagent was added to the pellet and solution was incubated for 15 minutes at room temperature. The subsequent steps are described in the Trizol Reagent RNA isolation procedure. The final total RNA pellet was solubilized in 20μL water. The amount of RNA extracted was quantified using the Nanodrop UV-Vis spectrometer (Thermo, Rockford, IL) and mRNA integrity (RIN≈10) was assessed using a 2100 Agilent Bioanalyzer (Agilent, Santa Clara, CA).

RNA-Seq paired end libraries were prepared using the Illumina TruSeq RNA

Sample Prep Rev. A (kit lot #6849988, Illumina, San Diego, CA). First, mRNA was purified from total RNA using poly dT bead isolation and fragmented by heating in the presence of a divalent cation. The fragmented RNA was then converted to cDNA with reverse transcriptase using random hexamer priming and the resultant double stranded cDNA was purified. cDNA ends were repaired, adenylated at the 3‘ ends, and then ligated to Illumina adapter sequences. Primers matching the adapter sequences were then used to PCR amplify the cDNA sequences. These sequences were run on an Invitrogen 2% Size Select Gel (Lot# R19090-01) and a band corresponding to ~350 base pairs was excised and used for paired end (2×200bp) sequencing on an Illumina HiSeq 2000. Raw cluster station data was post-processed and a total of 80 million RNA-Seq reads in fastq format were used for splice junction discovery. All fastq files used in this study can be accessed at NCBI‘s Gene Expression Omnibus (GEO) repository [47] by using the following link: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45428.

Construction of the Splice Junction Database

Splice Junction Discovery with Bowtie-Tophat Software.

Annotated and unannotated junctions were detected using the Bowtie (v0.12.7) and Tophat (v1.4.0) splice-junction discovery programs [48, 49]. All default Bowtie parameters were used. In Tophat, the mate inner distance was set to 150. Two rounds of Bowtie-Tophat processing were conducted with a supplied set of RefSeq gene model annotations in GTF format [7]: the first round detected junctions only matching the gene annotation file (option −−no-novel-junctions) and the second round detected all junctions, both aligning to the GTF file and novel (option -G). All data processing was conducted on the Phoenix cluster at the University of Wisconsin-Madison Chemistry department. The set of novel junctions not matching the RefSeq gene annotation was extracted from sets of output .bed files by in-house Perl scripts.

Translation of the Junction Nucleotide Sequences.

The set of unannotated splice junction coordinates containing six or more supporting RNA-Seq reads were translated into putative peptide splice junctions. The exon coordinates were extended by 66 nucleotides on both of the flanking ends of the junction. Junctions frequently overlapped with known genes, therefore the transcriptional strand (i.e. forward or reverse) of the junction was inferred from this association. The sequences resulting from a three frame translation, either on the forward or reverse strand, were extracted from the reference genome (hg19), translated to amino acid sequence, and trimmed to the first arginine or lysine (MS data was from a tryptic digest). Sequences less than 5 amino acids or containing a stop codon near the splice site were removed. All splice junction sequences were appended to the canonical Uniprot proteomic (release-2012_10; 20,225 entries) and GPM CrAP database (version 2012.01.01, 115 sequences). Two additional customized databases were built by appending junction sequences to the Uniprot/Trembl (release-2012_10; 86,881 entries) and to the Ensembl (release GRCh37.70.pep.all) protein databases and searches were conducted as described below.

Mass Spectrometry Junction Database Searching.

Raw mass spectrometry files were searched against the customized UniProt+CrAP+ Junction (53,476 entries total) database using the Percolator search node within Proteome Discoverer (v1.3.0.339, Thermo Fisher Scientific, San Jose, CA). Percolator is a machine-learning supplement to the SEQUEST search algorithm that increases the sensitivity and specificity of peptide identifications [50]. Default peaklist-generating parameters were used. Precursor m/z tolerance was set to 10 ppm and product m/z tolerance was set to 0.05 Da. Peptides with up to two missed cleavages (trypsin) were permitted. Variable methionine oxidation and static carbamidomethylation were used. Using reversed sequences as a decoy database, peptides passing both 1% and 5% global FDR and 1% and 5% local FDR (splice junction hit group) were used for downstream analysis. Validation was based on q-values generated by Percolator.

For identification of a protein using Proteome Discoverer, protein grouping and strict parsimony principle was enabled, leucine and isoleucine were considered equal, and only peptides passing 1% FDR and having a delta Cn higher than 0.15 were used. A minimum of two peptides per protein was required for identification.

All mass spectrometric raw files associated with this study may be downloaded via FTP from the PeptideAtlas data repository [51] by accessing the following link: http://www.peptideatlas.org/PASS/PASS00215.

2.4 Results

Overview

RNA-Seq and MS-based proteomics data was collected; 19,873 transcripts, which map to 12,873 genes, and 6,810 proteins were identified. RNA-Seq reads were used to discover 144,818 unannotated splice junctions using Bowtie and Tophat software. 24,834 Tophat junctions passing an expression cut-off were translated into polypeptide sequences. Either three frames or the one frame inferred from comparison to gene models was translated, resulting in 33,136 polypeptide entries. The splice junction sequences were appended to the Uniprot canonical database (~20,000 entries) and searched against the mass spectrometric data. 210 splice junction peptides that were absent in the complete Uniprot/Trembl database (~87,000 entries) but present in RNA-Seq derived junctions passed 5% global FDR. A local FDR was applied to the splice junction peptides and 72 (5% local FDR) and 57 (1% local FDR) peptides were identified. An overview of these results are depicted in Figure 2.1.

Terminology employed to define the types of peptides identified in this study.

"Uniprot peptide" are all peptides identified by searching proteomics data against the full Uniprot/Trembl database that includes isoforms (86,766 entries). "Splice

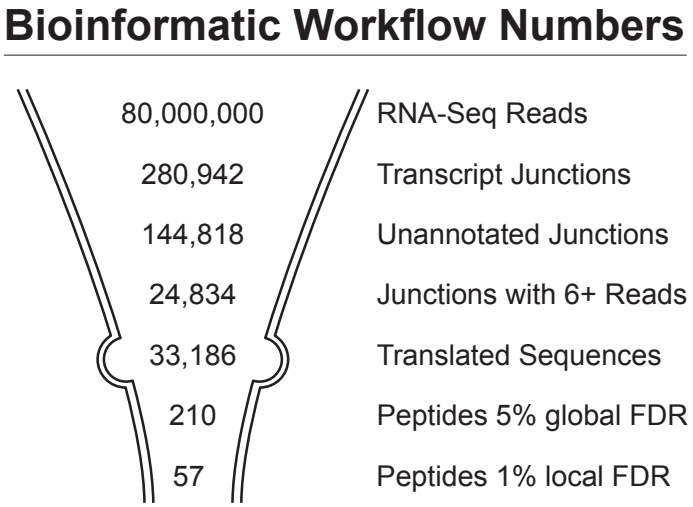

Figure 2.1. Results overview for the bioinformatic pipeline.

junction peptides" are all the peptides identified from RNA-Seq data (translated splice junctions) in this study that were not present in the full Uniprot/Trembl database.

mRNA and Protein Data Collection.

The transcriptomic and proteomic data collection workflow was designed to allow for accurate splice peptide detection (Figure 2.2). The protein and mRNA samples were extracted from the same Jurkat cell population to build a sample-specific junction database, one with minimal intra and inter-laboratory variation. Protein was extracted from cells using an SDS and DTT-based buffer (SDT) and the FASP protocol [45]. This protocol allows unbiased extraction and digestion of all protein groups (including hard-to-solubilize transmembrane proteins), an important factor when seeking to identify a proteoform [2]. Wisniewski et al. demonstrated that the composition of proteins identified using FASP corresponded to the expected abundances of Gene Ontology groups, with all protein groups evenly represented

[45]. In addition, when we compared SDT-based and urea-based extractions, we found that approximately 20% more protein (BCA assay) was extracted with SDT and that membrane proteins were more represented (results not shown). Total RNA was extracted from cells using a standard Trizol protocol. To provide a comprehensive RNA-Seq dataset for the sensitive discovery of alternative splice forms, 80 million reads of the longest RNA-Seq read type available on the Illumina platform were analyzed: libraries were derived from 350 bp cDNA sequences and 100 bp paired-ends were sequenced. In summary, the RNA and protein wet laboratory experiments were designed so that transcript-level junctions are sensitively detected and included in a comprehensive splice-junction database and the maximum number of discoverable splice-junction peptides using bottom-up proteomics are detected.

We measured the number of transcripts and proteins detected from both the RNA-Seq and peptide MS data, respectively, in order to compare the transcriptomic and proteomic datasets. RNA-Seq reads were processed by RSEM (RNA-Seq by Expectation-Maximization) to estimate transcript abundances [52]. Reads were aligned to a synthetic transcriptome and the number of reads associated with a given transcript was used to estimate that transcript's abundance in TPM (transcripts per million). RSEM processing of 80 million RNA-Seq reads resulted in 19,320 transcripts that mapped to 12,873 genes (TPM>1). Tandem mass spectra were processed by Proteome Discoverer (SEQUEST + Percolator algorithm) to infer protein identities. Experimental peptide MS spectra were processed with SEQUEST, followed by rounds of semi-supervised machine learning with Percolator, a target-decoy search using a 1% FDR, and grouping of proteins using maximum parsimony. Proteome Discoverer processing of 488,149 MS2 HCD spectra resulted in 77,733 Uniprot peptides and 6,810 proteins. Full results are in the supplemental table.

We also searched the mass spectrometric data against the UniProt/Trembl database (~87,000 entries) in order to measure the number of isoforms. We were able to detect two or more protein isoforms for 86 genes, where each isoform required at least one unique peptide that passed a 1% FDR cut-off. However, this number is likely to be artificially low since the detection of isoforms using bottom-up

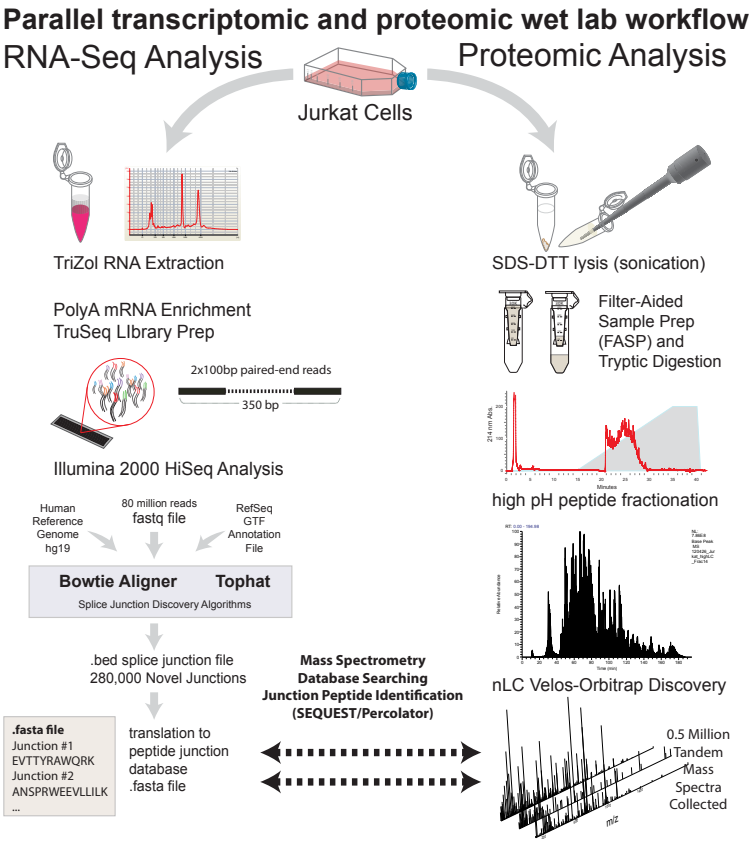

Figure 2.2. Transcriptomic and proteomic data collection. Both RNA-Seq and MS-based proteomics data were collected from the same Jurkat cell population. RNA-Seq data was processed using Bowtie/Tophat to discover new junctions that were then converted into polypeptide sequences. Protein was extracted and tryptically digested using the FASP protocol. Peptides were fractionated and subsequently analyzed on a nanoLC-Velos Orbitrap operating in data-dependent mode. A customized junction database derived from the RNA-Seq data was used for MS database searching.

proteomics requires a tryptic peptide unique to each isoform and protein sequence coverage is typically low (<25% coverage). The actual number of genes expressing more than one protein isoform is believed to be much higher [9].

Discovering Alternative Splice-Junctions from RNA-Seq Data.

Bowtie and Tophat software were used to discover splice junctions from 80 million RNA-Seq reads, and from these junctions, a peptide junction database was created for use in mass spectrometric data searching. Part of the procedure described in this section is illustrated in Figure 2.3.

Bowtie software efficiently aligns short RNA-Seq reads to a reference sequence (human reference genome, synthetic transcriptome, etc.) and Tophat discovers junctions not represented in the gene models. Both methods work together to discover novel junctions. Tophat discovers novel junctions primarily by finding RNA-Seq reads that span an exon-exon boundary, the most direct evidence of transcript splicing. It does this by segmenting the reads into subsequences and aligning the subsequences to the genome. When a read is "split"—one half of the read aligns upstream of an intron and the other half of the read aligns downstream of the intron—this is evidence for a novel splice junction. Tophat utilizes Bowtie for the alignment process and since both programs efficiently process RNA-Seq reads, the software can be run on desktop computers or local computer clusters accessible to most labs.

Processing of 80 million paired-end reads by Tophat/Bowtie resulted in a total of 280,942 junctions before filtering: 136,123 junctions present in RefSeq annotations (NM accession entries, representing RefSeq mRNA sequences) and 144,818 unannotated junctions. The list of annotated splice junctions were derived from NCBI RefSeq gene annotations because RefSeq has high quality, conservative annotations with minimum redundancy. Of the 144,818 unannotated junctions, 19,942, 1,185, and 22 junctions had over 10X, 100X, and 1000X read coverage (depth), respectively.

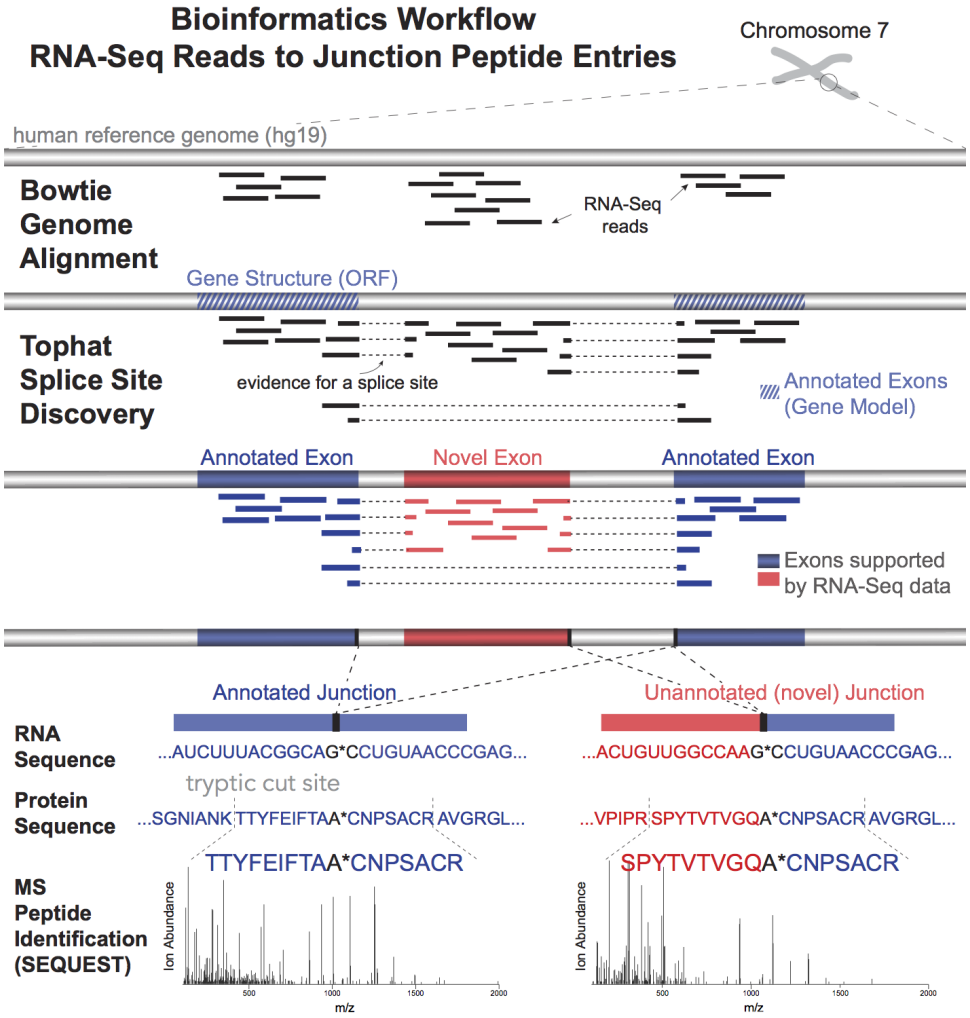

Figure 2.3. Bioinformatics workflow to convert raw RNA-Seq reads into junction peptide sequences. Bowtie and Tophat are used to align reads to the genome and annotated gene structure sequences (RefSeq). During Tophat splice junction alignment, reads are segmented and can align across exon-exon boundaries. When many reads support the presence of a novel splicing event, the junction is reported in .bed format. The list of unannotated junctions are converted to peptide sequence and searched against tandem mass spectra. Here, we show an example of a canonical and an alternative splice site identification from the detection of two tryptic peptides, where A* represents the amino acid residing, alanine (A) in this case, at the junction.

Selection of Minimum RNA-Seq Read Depth for Junction Inclusion in Database.

The RNA-Seq read depth, the number of RNA-Seq reads supporting the existence of a novel junction, was examined: a majority of the unannotated junctions were lower in abundance (read depth), and a significant number of junctions had just one supporting RNA-Seq read (Figure 2.4). We hypothesize that junctions containing a small number of supporting reads are either expressed at low levels and represent stochastic transcription [53, 54] or are the result of errors in the sequencing reads or Bowtie alignment step [55, 56]. We reasoned that many of these low coverage junction sequences are unlikely to result in a peptide identification, either because they are false positives or expressed at an extremely low-level (below 1 copy/cell).

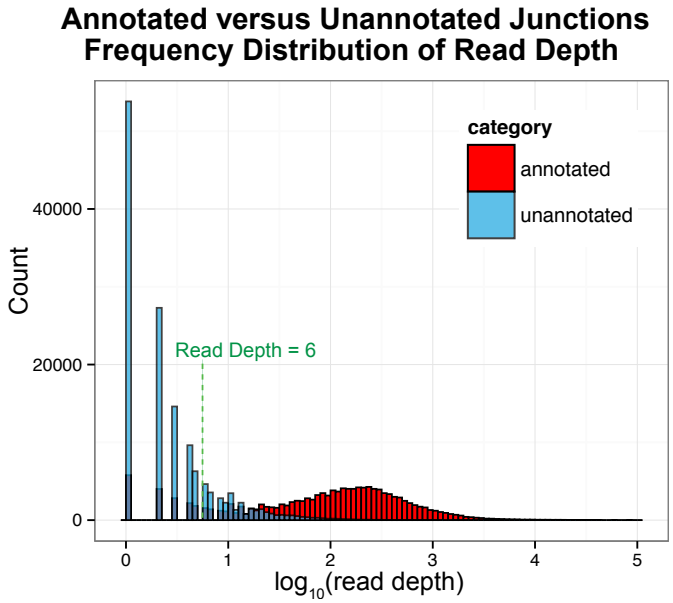

Figure 2.4. Read depth frequency distribution for Tophat-detected annotated and unannotated junctions. A majority of the reads aligned to RefSeq annotated junctions, as shown in the red histogram, while a lower number of reads, on average, aligned to unannotated junctions. Unannotated junctions with fewer than 6 supporting reads were removed prior to downstream analysis (green dotted line).

We elected to use junctions with six supporting reads or higher in the customized database to strike a balance between inclusion of novel junction sequences to promote peptide discovery and exclusion of junction sequences to minimize false positives. Two observations support using this cut-off. First, the transcript expression levels (RSEM output) were plotted against the protein expression levels (spectral counting), and the minimum transcriptional abundance required to detect a protein corresponded to ~6 RNA-Seq reads per junction. Second, multiple proteomic searches were performed, each differing by the minimum RNA-Seq read depth required for a junction sequence to be included in the database. For example, one search was against a database that included junctions having 1X RNA-Seq read depth or higher while another search was against a database that included junctions having 10X RNA-Seq read depth or higher. Uniprot peptide and splice junction peptide score distributions (see nomenclature section above) were compared to determine the incidence of false positives in the group of splice junction peptide identifications. After taking into account the above observations, a lower read depth cut-off of six was selected for database construction.

Construction of a Customized Junction Database from RNA-Seq Data.

A pipeline was developed to convert unannotated junction sequences into putative polypeptide entries for mass spectrometry searching. 24,834 junction sequences with 6 or more reads were translated into 33,186 amino acid sequences. To accomplish this, junction ends were extended, translation frame was inferred (when possible), and improbable sequences were trimmed or removed.

Transcript sequences were extended upstream and downstream of the Tophat junction. Each Tophat junction is represented by four coordinates: the start and end nucleotides of both the upstream and downstream Tophat "exon" (coordinates 1,2,3, and 4 in Figure 2.5). In humans, the average exon size is 148 nucleotides in length (7), but the reported Tophat "exons" ranged from 8 to 100 nucleotides and are an average of 64 nucleotides (Figure 2.6). This exon size distribution results

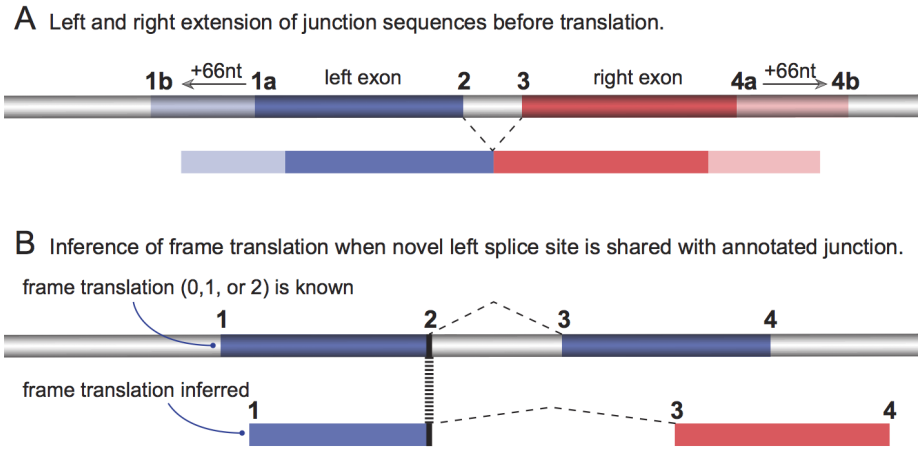

Figure 2.5. Junction sequence processing before translation into peptide sequences. 4A) Each Tophat junction consists of four coordinates: 1a, 2, 3, and 4a. Sequences were extended by 66 nucleotides before translation to increase the probability of detecting peptides that may partially protrude past 1a and 4a. 4B) The frame translation was inferred for the subset of cases in which the left splice site, 2, of the unannotated junction corresponded exactly to the left splice site of an annotated junction.

from the Tophat Software. Tophat reports only the stretch of sequence —upstream and downstream of the splice site —that has evidence: aligned 100 bp RNA-Seq reads that overlap the junction. In order to increase the probability of detecting peptides that extend past Tophat junction ends (coordinates 1 and 4 in Figure 2.5a), additional sequence was appended to both sides of the junction. Before translation, the sequence coordinates of each Tophat junction were thus lengthened at flanking exon ends (5′ end of upstream exon, 3′ end of downstream exon) by 66 nucleotides.

The frame translation was inferred for a subset of the Tophat junction sequences. In the case that a novel junction's left splice site (coordinate 2 in Figure 2.5b) corresponded to the left splice site of a known gene structure, the frame translation was inferred. This is reasonable because the upstream exon is part of a known gene model and will most likely be translated in the same frame as the canonical splice form. For all other junctions where the frame could not be inferred, such as when there were novel left and right (coordinate 2 and 3 in Figure 2.5b) splice sites or a

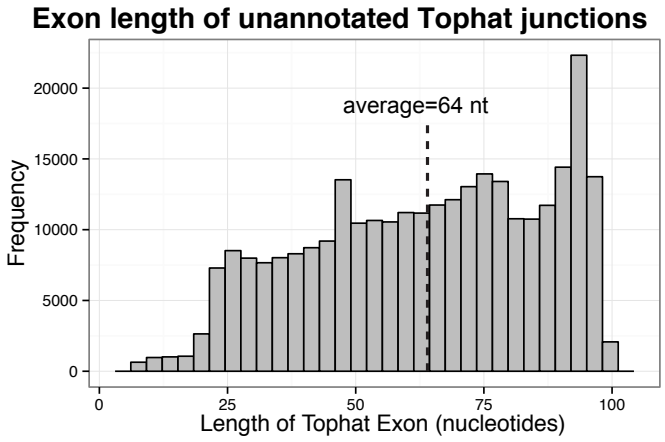

Figure 2.6. Distribution of the Tophat exon lengths for unannotated junctions. The Tophat exon lengths are a consequence of the de novo splice discovery program, where only RNA-Seq reads spanning a splice site can be used as direct evidence for the existence of a junction.

novel left splice site (coordinate 2), all three frames were translated.

Improbable sequences were either removed or trimmed. First, short peptides (<5 a.a.), peptides with an abundance of stop codons, and peptides that did not include the splice site amino acid were removed. Second, polypeptide start and end sites each were trimmed to the first occurrence of a lysine (K) or arginine (R), preventing the inclusion of non-tryptic fragment sequences. Sequences were trimmed because the proteomics data for this study was based on detection of tryptic peptides, all of which begin after the C-terminus of a lysine (K) or arginine (R) and likewise end with a K or R.

After subjecting the 33,186 unannotated transcript-level junctions to the aforementioned processing steps, 24,834 remained and these sequences were translated into 33,589 junction peptide entries (the higher number of peptide entries resulted from requisite 3-frame translations) and were integrated into a customized junction database (see supplemental table for full list). Most of the 8,352 junctions filtered out were due to high frequency stop codons, or possibly to out-of-frame translation. To create customized junction databases, the junction sequence entries were appended

to the following protein databases: canonical UniProt reference (20,225 entries), UniProt/Trembl (86,881 entries), or Ensembl (104,785 entries, version 70). The addition of junction peptide entries increased the size of the Uniprot, UniProt/Trembl, and Ensembl databases by 13.1% (1,474,776 aa were added to 11,291,209 aa), 4.1% (1,474,776 aa were added to 36,164,128 aa), and 3.7% (1,474,776 aa were added to 39,786,499 aa), respectively. The raw MS files were searched against each of these three combination databases to identify the subset of splice junction peptides. The lists of splice junction peptides among the three searches were very similar (see supplemental table); we have chosen here to focus on MS results from the UniProt reference + junction sequence database. Junction peptide sequences identified from the Uniprot reference + junction sequence database were BLAST searched against the full UniProt/Trembl database (~87,000 entries) and peptides not present in UniProt/Trembl (hence new splice junction peptides) were retrieved.

Balancing Splice Peptide Discovery and False Positives.

It has previously been demonstrated in multiple settings that when expanding a proteomic database to include possible proteoform sequences or when searching MS data against six-frame translated reference genomes, the false positive rate increases and the sensitivity of peptide identification decreases [40, 57, 58]. Therefore, proper statistical methods and scoring thresholds must be employed for accurate identification of new variants.

A conservative local FDR based on Posterior Error Probability (PEP) values was used for identified splice junction peptides to reduce the number of false positives [59]. MS data was processed using Percolator, a machine-learning adaptation to SEQUEST [50], and a reverse target-decoy database. The search yields were 77,733 and 83,385 identified Uniprot peptides at a 1% and 5% false discovery rate, respectively. To ascertain any peptide scoring biases for the Uniprot and splice junction peptides, the delta precursor ppm and XCorr SEQUEST scores were plotted for different subsets of peptides (Figure 2.7). Figure 2.7b shows a comparison of splice junction peptide score distributions to the score distributions from the same number of

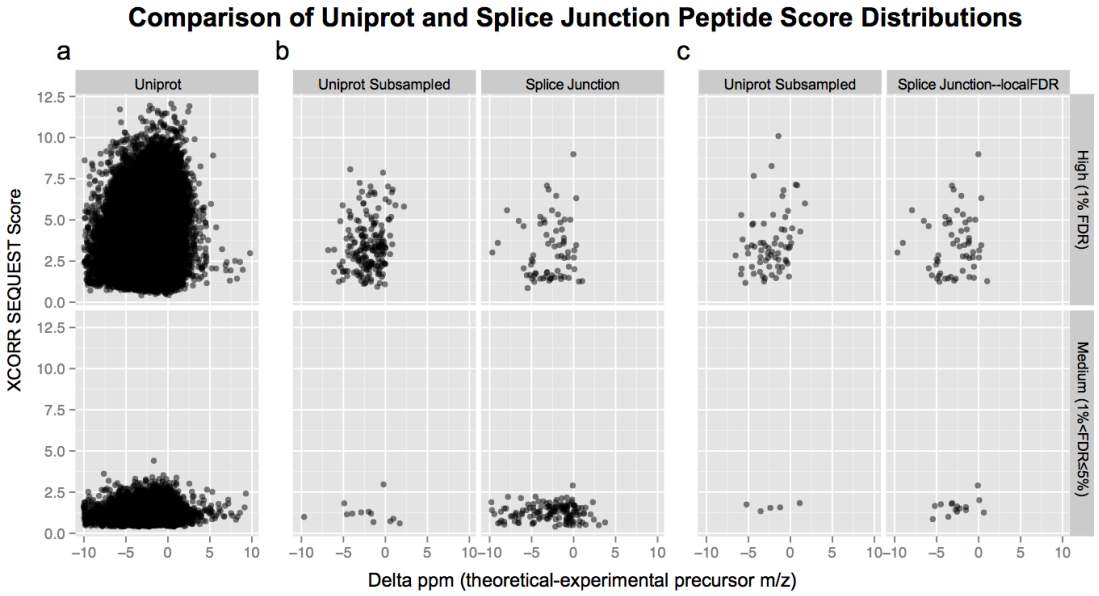

Figure 2.7. Comparison of peptide score distributions for canonical and junction peptides. For the comparison of peptide scores, delta ppm (difference, in parts per million, between measured and experimental precursor m/z) versus XCorr (cross-correlation value from MS search) SEQUEST score was plotted. 7A) Score distributions for peptides matching the UniProt/Trembl human proteomic database. 7B) Junction peptide score distribution (n=210, 5% FDR) compared to the score distribution of 210 peptides randomly subsampled from panel 7A. 7C) Local FDR junction peptide score distribution (n=72, 5% local FDR) compared to score distributions for 72 peptides randomly subsampled from panel 7A. The plotted points in Figure 7B illustrate that junction peptides tend to have lower scores than the canonical ones when employing the global FDR thresholds. This mismatch indicates the 210 junction peptides will have greater than 5% false positives. Figure 7C shows the remedy to this situation, namely calculation of a more strict local FDR (based on the Percolator posterior expectation probability score), which then makes the canonical and junction distributions quite similar.

sub-sampled Uniprot peptides (n=210). The population of splice junction peptides contained a disproportionately higher number of lower scoring peptides (passing 5% FDR, but not 1% FDR). This is probably due to the large number of specious sequences resulting from the three-frame translated or even non-coding junctions. To resolve this issue, we elected to apply a 1% local FDR threshold to the splice junction peptides based on the posterior error probability, or PEP, values. The PEP for a peptide identification is the probability that the experimental spectra actually originated from the sequence reported. The local FDR for a subgroup of peptides is calculated by dividing the expected number of false positives (the sum of PEP values for all peptides within the group) by the total number of peptides [60]. Figure 2.7c shows that applying a local FDR threshold to the splice junction peptides achieves a similar score distribution to the sub-sampled Uniprot peptides (n=57, 1% FDR; n= 72, 5% FDR). Thus, 57 novel junction peptides have been discovered at the local false discovery rate of 1%.

Presence of Splice Junction Peptides in Various Databases.

The UniProt reference protein set is a popular database used in proteomics; however, many other databases and sequence repositories are also available for researchers to use. Therefore, we checked to see how many of the 72 splice junction peptide sequences that were not in the Uniprot/Trembl database were present in other publically available databases. We did this by BLAST searching each of the 72 splice junction peptide sequences against the human datasets found within NCBI's dbEST, INSDC (the International Nucleotide Sequence Database Collection, which includes Genbank and the DNA Data Bank of Japan), Ensembl, Genscan, and the NIST peptide mass spectral library. We determined how many sequences were already present in these databases and found 22, 22, 7, 12, and 5 peptides, respectively. A table showing each peptide sequence and the database(s) it was found in is available in the supplementary information. All in all, 39 of the sequences corresponding to the 72 splice junction peptides were found in one or more of the nucleotide sequence or proteomic repositories; however, most of these had limited or no evidence of

protein expression. It may be noted that although BLAST analysis easily determines if a particular sequence is present in a database, that does not mean that the splice junction peptide would be identified with statistical significance in a mass spectral search against that same database.

Discovered Alternative Splice Junction Peptides.

We designed a bioinformatic workflow that leverages RNA-Seq data to create a customized splice-junction database. Despite the comprehensiveness of the UniProt/Trembl human proteomic database—86,766 discrete protein entries ranging from manually validated to computationally predicted sequences (entries without evidence for the expression of the protein)—we still discovered 57 novel splice junction peptide sequences that were absent in the UniProt/Trembl database. The RNA-Seq customized splice junction database provided a promising mechanism for discovery of these peptides.

The discovered peptides represented many different types of splicing including exon skipping events, alternative donors and acceptors, novel exons, alternative transcriptional start sites and novel exon-exon junctions (Table 2.1). A full table of each splice junction peptide that includes information such as the observed canonical peptide, a description of the splicing event (e.g. exon skipping), and transcript level alternative/canonical splicing frequencies, may be found in the supplemental table. The most frequent splicing types exhibited by the splice-junction peptides were alternative acceptor and donor sites and skipped exons.

The most common splicing events were small insertions and deletions (indels) occurring at the 3' acceptor exons, frequently characterized by the NAGNAG motifs where two AG dinucleotide splice site acceptors sit in close proximity to each other: this agrees with recent gene validation efforts of the GENCODE gene annotation project in which mass spectrometry data retrieved from the Global Proteome Machine (GPM) and PeptideAtlas were aligned to GENCODE gene models to assess the number of translated products [17]. NAGNAG tandem splicing may cause subtle changes in the protein sequences, just the insertion or deletion of one amino

Events represented by 57 discovered splice-junction peptides

Splicing Event		Frequency
Alternative Acceptor	+3nt	13
	-3nt	2
	35nt	1
	77nt	1
Skipped Exon	1 Exon	9
	2 Exon	2
	3 Exon	1
Novel Exon	Left	3
	Right	4
Completely Unannotated		7
Alternative Donor (-21,-12, +12, +23, +24, +58)		5
Alternative Transcriptional Start Site (TSS)		2
Within Intron		2
Cross Gene		1

Table 2.1. Frequency of splicing events represented by the 57 junction peptides passing 1% local FDR. A variety of different splicing events were detected from RNA-Seq specific splice junction entries.

acid, yet there is evidence that these alternative forms are not merely the result of stochastic noise from splicing machinery. Recently, evidence has been mounting that NAGNAG splicing plays a functional role. These splicing sequences have been shown to be evolutionarily conserved across species and the ratio of canonical to alternative splicing has been shown to be tissue-specific—facts that suggest NAGNAG splicing is important to protein function [61]. The PSI (ψ) or "Percentage Spliced In" [63] was calculated for all fifteen peptides exhibiting alternative acceptor splicing. "Percentage Spliced In" is the fraction of minor and major isoforms, expressed as a percentage. The PSI ranged from a low of 0.2% to a high of 27.1% and the average was 5.6%.

2.5 Discussion

A peptide or protein sequence must be listed in the database to be identified by mass spectrometry; hence, proteomics relies on databases to discover new proteoforms. Despite the large strides that groups curating databases such as Swiss-Prot/Trembl and GENCODE have made in completing gene models, including improving pipelines to better discriminate between putative and actual protein sequences by incorporating the latest high-throughput MS data, not all proteins are listed. The diversity of human proteoforms is immense and proteoforms expressed in thousands of human cell types have yet to be catalogued. Furthermore, the list of protein entries in the human reference proteome is consolidated from the human cells studied to date and may not reflect variants present in any particular sample. One of the major sources of cell-type specific proteomic variation is alternative splicing, where the protein coding exons of a gene are stitched together in various combinations to create multiple splice forms. While there have been efforts to create expanded databases that capture all alternative splicing variants, we suggest that the solution should not be unbounded expansion of a central database, but rather the customization of databases for specific cell-types. Due to recent unprecedented advances in next generation sequencing and RNA-Seq, this proteomics strategy is now within reach.

We describe here a novel strategy to use a sample-specific RNA-Seq dataset to characterize new cell-type specific splicing events not yet captured in proteomic databases. We collected RNA-Seq and proteomic data from a single cell population (Jurkat cells), constructed an empirically derived splice-junction database from RNA-Seq data, searched the accompanying mass spectrometry data against the customized splice-junction database, and discovered new splice-junction peptides that were absent from the UniProt/Trembl proteomic database, which includes all putative gene annotations predicted from the Ensembl pipeline. To our knowledge, this is the first report of using RNA-Seq data to discover mRNA splice junctions de novo from direct alignment of RNA-Seq reads with the reference genome (exon boundaries not supplied) and construction of a customized splice junction database

from the splicing events that were detected.

We found that an important element in creating such customized databases is achieving a balance between the inclusion of all putative proteoform sequences (for which there is transcript-level evidence) to maximize discovery of new forms, and the reduction of database size to control for sequence redundancy and false positives. Unbounded expansion of databases by including additional protein sequences, such as those derived from proteogenomics (6 frame translation), ab initio gene predictors, and transcriptomics data (3 or 6 frame translation), is problematic because it increases false positives, redundancy, and MS search times. The false positive rate is increased when many spurious protein sequences, corresponding to proteins not expressed in the sample, are added to the database, because the presence of these sequences increases the probability that an experimental spectrum matches that sequence by random chance [57]. Note that some of the junction peptide sequences described in this paper were found in expanded databases (e.g. GenBank), but mass spectrometric searching against these large, all-inclusive databases is problematic for the reasons stated above. Redundancy is also increased by adding many closely related proteoforms, and this confounds protein parsimony, the inference of protein from peptides [41, 42]. Conversely, in the case of our experimentally determined splice-junctions, strict reduction of the database to include only those sequences with the highest expression levels (>30 transcripts per million, TPM) was inappropriate: there are plenty of examples of low transcript abundance but high protein abundance and vice versa [64, 65]. Therefore, to strike a balance between discovering novel alternative splice junctions and minimizing the number of spurious sequences, we included junction sequences with six or more supporting RNA-Seq reads and used a local 1% FDR for splice junction peptides.

Another important issue in the discovery of alternative splice forms at the protein level is the low number of splice-specific peptides actually identified, an issue that has been revealed by work reported in the literature [17, 19, 22, 24, 25, 30, 66]. Part of the reason for the low number of alternative splice variants detected are the technical differences between RNA-Seq and bottom-up proteomics, namely sequence coverage and detection sensitivity. RNA-Seq reads are obtained by, first,

randomly fragmenting mRNA molecules with a divalent cation and heat, and second, reverse transcribing these RNA fragments into cDNA and using PCR to amplify this initial cDNA library. These steps allow for the detection of reads spanning the whole transcript (100% coverage) and corresponding to transcripts expressed at a low-level [67]. Peptide spectra, on the other hand, are obtained by, first, employing a proteolytic enzyme to cleave the protein at prescribed sites, and second, directly electrospraying the peptide into a mass spectrometer and collecting spectral scans. These steps allow the detection of only those peptides amenable to LC-MS/MS (~5-25% coverage) and corresponding to proteins expressed at a high enough level for detection (attomoles-femtomoles). The consequence of these RNA and protein measurement differences is that it is much more difficult to detect alternative splice variants at the protein level than the RNA level. Transcripts can be sensitively (<1 transcript/cell) and completely (100% sequence coverage) characterized, but for proteins, only moderately or highly expressed (>1 protein molecules/cell) proteins are usually detected and amino acid sequence coverage is typically low (~5-25%). Alternatively spliced proteins are difficult to detect because 1) they have lower cellular abundances than the canonical forms, 2) require at least one splice form-specific peptide for unambiguous detection, likely one spanning a junction or residing in a splice form-specific exon [24], and, 3) the alternative splice variant sequence is sometimes not yet in the database.

The number of alternative splice forms expected to be detected in a bottom-up proteomics experiment has been estimated using computational approaches [19, 22, 24]. Some authors reported that they identified the expected number of splice-specific peptides while other authors identified far fewer peptides than predicted. These discrepancies were attributed to the underlying assumptions of their statistical models. In any case, this paper shows that new splice junction peptides can be detected directly from customized databases built from RNA-Seq data. It is likely that these peptides represent the tip of the iceberg, and that there are many more splice-specific peptides that are currently undetected. Extensions of the strategy employed in this paper may be employed to increase the ability to detect splice junction peptides. For example, utilizing multiple proteolytic enzymes (LysC,

GluC, etc.) will increase the odds of creating a splice-specific peptide detectable by LC-MS/MS, or targeted proteomics strategies such as selected reaction monitoring (SRM) analysis could be employed to decrease detection limits for splice junction peptides of interest that have low abundances.

Next generation sequencing and RNA-Seq has developed rapidly and its cost has decreased greatly making it accessible to most research organizations. Because of this technological revolution, there is a great opportunity for next generation proteomics to utilize sample-specific, customized databases built from RNA-Seq data. The present work on discovery of novel splice junctions is one important aspect of proteomic variation, but there are many other variations (e.g. SNVs, RNA fusion products) that may also be captured in custom databases. As RNA-seq technologies continue to become increasingly affordable, accessible, and sensitive, the power and utility of this new strategy for the discovery of proteomic variation will continue to expand.

2.6 Acknowledgements

This work was supported by NIH grants 1P01GM081629 and 1P50HG004952. GMS was supported by the NIH Genomic Sciences Training Program 5T32HG002760. The Phoenix Computing Cluster at the University of Wisconsin-Madison Chemistry Department is supported by the National Science Foundation Grant CHE-0840494. We would like to thank Dr. Mark Scalf for assistance with the mass spectrometric data collection. We would like to thank Dr. Victor Ruotti and Dr. Colin Dewey for helpful discussions regarding the transcriptomics pipeline. RNA-Sequencing work was performed at the University of Wisconsin –Madison Biotechnology Center.

References

[1] K. Ning, D. Fermin, and A. I. Nesvizhskii. Computational analysis of unassigned high-quality MS/MS spectra in proteomic data sets. *Proteomics* 10.14 (2010), pp. 2712–2718.

[2] Lloyd M. Smith and Neil L. Kelleher. Proteoform: a single term describing protein complexity. *Nat Meth* 10.3 (2013), pp. 186–187.

[3] Q. Pan et al. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* 40.12 (2008), pp. 1413–1415.

[4] E. T. Wang et al. Alternative isoform regulation in human tissue transcriptomes. *Nature* 456.7221 (2008), pp. 470–476.

[5] E. S. Lander et al. Initial sequencing and analysis of the human genome. *Nature* 409.6822 (2001), pp. 860–921.

[6] W. J. Kent et al. The human genome browser at UCSC. *Genome Research* 12.6 (2002), pp. 996–1006.

[7] K. D. Pruitt, T. Tatusova, and D. R. Maglott. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* 35 (2007), pp. D61–D65.

[8] A. Bairoch et al. The universal protein resource (UniProt). *Nucleic Acids Research* 33 (2005), pp. D154–D159.

[9] J. Harrow et al. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research* 22.9 (2012), pp. 1760–1774.

[10] P. J. Kersey et al. The International Protein Index: An integrated database for proteomics experiments. *Proteomics* 4.7 (2004), pp. 1985–1988.

[11] T. Hubbard et al. The Ensembl genome database project. *Nucleic Acids Research* 30.1 (2002), pp. 38–41.

[12] T. Kwan et al. Genome-wide analysis of transcript isoform variation in humans. *Nature Genetics* 40.2 (2008), pp. 225–231.

[13] S. Stamm et al. Function of alternative splicing. *Gene* 344 (2005), pp. 1–20.

[14] B. J. Blencowe. Alternative splicing: New insights from global analyses. *Cell* 126.1 (2006), pp. 37–47.

[15] P. R. Romero et al. Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc Natl Acad Sci U S A* 103.22 (2006), pp. 8390–5.

[16] G. S. Wang and T. A. Cooper. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature Reviews Genetics* 8.10 (2007), pp. 749–761.

[17] I. Ezkurdia et al. Comparative Proteomics Reveals a Significant Bias Toward Alternative Protein Isoforms with Conserved Structure and Function. *Molecular Biology and Evolution* 29.9 (2012), pp. 2265–2283.

[18] Rajasree Menon et al. Identification of Novel Alternative Splice Isoforms of Circulating Proteins in a Mouse Model of Human Pancreatic Cancer. *Cancer Research* 69.1 (2009), pp. 300–309.

[19] Michael Tress et al. Proteomics studies confirm the presence of alternative protein isoforms on a large scale. *Genome Biology* 9.11 (2008), R162.

[20] R. Menon and G. S. Omenn. Proteomic Characterization of Novel Alternative Splice Variant Proteins in Human Epidermal Growth Factor Receptor 2/neu-Induced Breast Cancers. *Cancer Research* 70.9 (2010), pp. 3440–3449.

[21] M. L. Tress et al. The implications of alternative splicing in the ENCODE protein complement. *Proceedings of the National Academy of Sciences of the United States of America* 104.13 (2007), pp. 5495–5500.

[22] E. I. Severing, A. D. J. van Dijk, and Rchj van Ham. Assessing the contribution of alternative splicing to proteome diversity in Arabidopsis thaliana using proteomics data. *Bmc Plant Biology* 11 (2011).

[23] Leoni G. et al. Coding potential of the products of alternative splicing in human. *Genome Biology* 12.1 (2011).

[24] Paul Blakeley et al. Investigating protein isoforms via proteomics: A feasibility study. *Proteomics* 10.6 (2010), pp. 1127–1140.

[25] T. Hubbard et al. Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome. *Genome Research* 21.5 (2011), pp. 756–767.

[26] Danny A. Bitton et al. An Integrated Mass-Spectrometry Pipeline Identifies Novel Protein Coding-Regions in the Human Genome. *Plos One* 5.1 (2010), e8949.

[27] B. Y. Lin et al. A compatible exon-exon junction database for the identification of exon skipping events using tandem mass spectrum data. *Bmc Bioinformatics* 9 (2008).

[28] X. B. Xing et al. The discovery of novel protein-coding features in mouse genome based on mass spectrometry data. *Genomics* 98.5 (2011), pp. 343–351.

[29] A. Zhou, F. Zhang, and J. Y. Chen. PEPPI: a peptidomic database of human protein isoforms for proteomics experiments. *Bmc Bioinformatics* 11 Suppl 6 (2010), S7.

[30] K. Y. Chang et al. Detection of Alternative Splice Variants at the Proteome Level in Aspergillus flavus. *Journal of Proteome Research* 9.3 (2010), pp. 1209–1217.

[31] G. Lopez-Casado et al. Enabling proteomic studies with RNA-Seq: The proteome of tomato pollen as a test case. *Proteomics* 12.6 (2012), pp. 761–774.

[32] Stephen Tanner et al. Improving gene annotation using peptide mass spectrometry. *Genome Research* 17.2 (2007), pp. 231–239.

[33] J. H. Chen et al. Improved protein identification using a species- specific protein/peptide database derived from expressed sequence tags. *Plant Omics* 4.5 (2011), pp. 257–263.

[34] K. A. Power et al. High-Throughput Proteomics Detection of Novel Splice Isoforms in Human Platelets. *Plos One* 4.3 (2009).

[35] N. J. Edwards. Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. *Molecular Systems Biology* 3 (2007).

[36] John R. Yates, Jimmy K. Eng, and Ashley L. McCormack. Mining Genomes: Correlating Tandem Mass Spectra of Modified and Unmodified Peptides to Sequences in Nucleotide Databases. *Analytical Chemistry* 67.18 (1995), pp. 3202–3210.

[37] N. E. Castellana et al. Discovery and revision of Arabidopsis genes by proteogenomics. *Proceedings of the National Academy of Sciences of the United States of America* 105.52 (2008), pp. 21034–21038.

[38] N. E. Castellana et al. Template Proteogenomics: Sequencing Whole Proteins Using an Imperfect Database. *Molecular & Cellular Proteomics* 9.6 (2010), pp. 1260–1270.

[39] Xiaojing Wang et al. Protein Identification Using Customized Protein Sequence Databases Derived from RNA-Seq Data. *Journal of Proteome Research* (2011).

[40] N. Castellana and V. Bafna. Proteogenomics to discover the full coding content of genomes: A computational perspective. *Journal of Proteomics* 73.11 (2010), pp. 2124–2135.

[41] A. I. Nesvizhskii and R. Aebersold. Interpretation of shotgun proteomic data - The protein inference problem. *Molecular & Cellular Proteomics* 4.10 (2005), pp. 1419–1440.

[42] K. Meyer-Arendt et al. IsoformResolver: A Peptide-Centric Algorithm for Protein Inference. *Journal of Proteome Research* 10.7 (2011), pp. 3060–3075.

[43] C. Adamidi et al. De novo assembly and validation of planaria transcriptome by massive parallel sequencing and shotgun proteomics. *Genome Research* 21.7 (2011), pp. 1193–1200.

[44] V. C. Evans et al. De novo derivation of proteomes from transcriptomes for transcript and protein identification. *Nat Methods* 9.12 (2012), pp. 1207–11.

[45] J. R. Wisniewski et al. Universal sample preparation method for proteome analysis. *Nature Methods* 6.5 (2009), 359–U60.

[46] Martin Gilar et al. Orthogonality of Separation in Two-Dimensional Liquid Chromatography. *Analytical Chemistry* 77.19 (2005), pp. 6426–6434.

[47] T. Barrett et al. NCBI GEO: archive for functional genomics data sets-10 years on. *Nucleic Acids Research* 39 (2011), pp. D1005–D1010.

[48] B. Langmead et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 10.3 (2009).

[49] C. Trapnell, L. Pachter, and S. L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25.9 (2009), pp. 1105–1111.

[50] Lukas Käll et al. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Meth* 4.11 (2007), pp. 923–925.

[51] F. Desiere et al. The PeptideAtlas project. *Nucleic Acids Research* 34 (2006), pp. D655–D658.

[52] B. Li and C. N. Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *Bmc Bioinformatics* 12 (2011), p. 323.

[53] E. Melamud and J. Moult. Stochastic noise in splicing machinery. *Nucleic Acids Research* 37.14 (2009), pp. 4873–4886.

[54] D. Hebenstreit et al. RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Molecular Systems Biology* 7 (2011).

[55] Manuel Garber et al. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Meth* 8.6 (2011). 10.1038/nmeth.1613, pp. 469–477.

[56] F. Ozsolak and P. M. Milos. RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics* 12.2 (2011), pp. 87–98.

[57] A. I. Nesvizhskii. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteomics* 73.11 (2010), pp. 2092–2123.

[58] Paul Blakeley, Ian M. Overton, and Simon J. Hubbard. Addressing Statistical Biases in Nucleotide-Derived Protein Databases for Proteogenomic Search Strategies. *Journal of Proteome Research* (2012).

[59] Lukas Käll et al. Posterior Error Probabilities and False Discovery Rates: Two Sides of the Same Coin. *Journal of Proteome Research* 7.1 (2007), pp. 40–44.

[60] H. Choi, D. Ghosh, and A. I. Nesvizhskii. Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. *Journal of Proteome Research* 7.1 (2008), pp. 286–292.

[61] R. K. Bradley et al. Alternative Splicing of RNA Triplets Is Often Regulated and Accelerates Proteome Evolution. *Plos Biology* 10.1 (2012).

[62] M. Hiller et al. Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *Nature Genetics* 36.12 (2004), pp. 1255–1257.

[63] Y. Katz et al. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods* 7.12 (2010), 1009–U101.

[64] Martin Beck et al. The quantitative proteome of a human cell line. *Mol Syst Biol* 7 (2011). 10.1038/msb.2011.82.

[65] Nagarjuna Nagaraj et al. Deep proteome and transcriptome mapping of a human cancer cell line. *Mol Syst Biol* 7 (2011). 10.1038/msb.2011.81.

[66] K. Ning and A. I. Nesvizhskii. The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment. *Bmc Bioinformatics* 11 Suppl 11 (2010), S14.

[67] L. C. Jiang et al. Synthetic spike-in standards for RNA-seq experiments. *Genome Research* 21.9 (2011), pp. 1543–1551.

3 LARGE-SCALE MASS SPECTROMETRIC DETECTION OF VARIANT PEPTIDES RESULTING FROM NON-SYNONYMOUS NUCLEOTIDE DIFFERENCES

This chapter has been published in *Journal of Proteome Research*

Sheynkman, G. M., Shortreed, M. R., Frey, B. L., Scalf, M., and Smith, L. M. (2014) Large-scale mass spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences. *J Proteome Res* 13, 228-240

3.1 Abstract

Each individual carries thousands of non-synonymous single nucleotide variants (nsSNVs) in their genome, each corresponding to a single amino acid polymorphism (SAP) in the encoded proteins. It is important to be able to directly detect and quantify these variations at the protein level in order to study post-transcriptional regulation, differential allelic expression, and other important biological processes. However, such variant peptides are not generally detected in standard proteomic analyses, due to their absence from the generic databases that are employed for mass spectrometry searching. Here, we extend previous work that demonstrated the use of customized SAP databases constructed from sample-matched RNA-Seq data. We collected deep coverage RNA-Seq data from the Jurkat cell line, compiled the set of nsSNVs that are expressed, used this information to construct a customized SAP database, and searched it against deep coverage shotgun MS data obtained from the same sample. This approach enabled detection of 421 SAP peptides mapping to 395 nsSNVs. We compared these peptides to peptides identified from a large generic search database containing all known nsSNVs (dbSNP) and found that more

than 70% of the SAP peptides from this dbSNP-derived search were not supported by the RNA-Seq data, and thus are likely false positives. Next, we increased the SAP coverage from the RNA-Seq derived database by utilizing multiple protease digestions, thereby increasing variant detection to 695 SAP peptides mapping to 504 nsSNV sites. These detected SAP peptides corresponded to moderate to high abundance transcripts (30+ transcripts per million, TPM). The SAP peptides included 192 allelic pairs; the relative expression levels of the two alleles were evaluated for 51 of those pairs, and found to be comparable in all cases.

3.2 Introduction

DNA sequencing technologies have allowed researchers to uncover an astounding amount of genetic variation in humans, including a multitude of single nucleotide variations, insertions, deletions, tandem repeats, inversions, translocations, and duplications [1]. Among these variations, single nucleotide variants (SNVs), the single nucleotide differences between two genomes that occur on average about once every 860 base pairs, have been the most intensely researched, mainly through genome-wide association studies that seek to uncover the sets of causative SNVs that are responsible for a disease or trait [1–3]. Advances in sample preparation, sequencing instrumentation, and computational data analysis have made it easier for researchers to rapidly sequence and discover the millions of SNVs found within a genome, and thus the challenge today is not how to discover these variations but how to sift through them to find those with functional significance [4].

One way to simplify the study of SNVs is to focus on those SNVs that lie within coding regions, because these SNVs can cause a change in the protein amino acid sequence and are thus most likely to modify the function of a protein. Coding SNVs can be classified into three types: (1) synonymous, which does not change the corresponding amino acid, (2) nonsense, which introduces a premature stop codon, and (3) non-synonymous, also called missense, which changes the corresponding amino acid. While it is well accepted that synonymous SNVs do not affect the protein function, and nonsense SNVs usually cause a loss of function (because the

protein is truncated)[5], it is harder to determine what effect a non-synonymous SNV (nsSNV) has on a given protein's function [6].

Current strategies employed to study the functional effects of nsSNVs include determining statistical associations between well phenotyped populations (i.e. genome-wide association studies), computationally predicting the functional effect of an SNV using programs like SIFT and PolyPhen-2 [7, 8], and, most recently, evaluating the nsSNV within the biological system, such as in a protein-protein interaction or regulatory network [9]. These approaches guide the prioritization of nsSNVs for subsequent validation and hypothesis testing using in vitro and in vivo functional assays. Though these statistical and bioinformatic strategies have aided the study of nsSNVs, another valuable piece of information is the direct measurement of the variant-containing protein.

The direct detection of proteins containing single amino acid polymorphisms (SAPs) encoded by an nsSNV can aid researchers in studying the functional significance of these variants. Directly measuring these SAP-containing proteoforms [10] is essential to understanding how an SNV influences a variety of processes at the protein-level such as post-translational regulation of protein expression (e.g. protein degradation and stability), localization of the protein, modulation of protein-protein interactions, and influence of the SAP on patterns of post-translational modifications (PTMs). Furthermore, understanding the influence of SAPs across various cell states would be very difficult without technologies to measure these protein variations. Fortunately, mass spectrometry-based proteomics has undergone remarkable development in the past decade and can now be used to comprehensively identify and quantify large portions of the proteome [11–13]. MS-based proteomics has tremendous potential to detect SAPs on a large scale, providing researchers with valuable information regarding the relationship between genomic variations and the ultimate protein products they encode.

The main impediment to the wide-spread adoption of variant peptide detection using mass spectrometry has been the lack of proteomic databases that include sample-specific variant sequences. The current practice in proteomics to identify peptides or proteins is to search the mass spectra against the sequences contained

in a reference proteomic database, which is derived from either the human reference genome or cDNA sequence repositories [14–17]. Since the reference protein sequences do not contain the amino acid variations specific to a sample, a mass spectrum produced from a variant-containing peptide will not correctly match to a sequence and, therefore, will fail to be detected.

Several researchers have addressed this problem by constructing proteomic databases that include SAPs and then searching these databases against tandem mass spectra to detect SAP peptides. One approach relies on the construction of an exhaustive SAP database which includes amino acid changes resulting from every hypothetical nucleotide change in the genome [18–20]. Another approach relies on the construction of a database that includes every SAP found within SNV or cancer mutation repositories, such as dbSNP or COSMIC [21–34]. Both of these approaches successfully allowed the detection of SAP peptides that are absent from the reference proteome and thus show the potential of proteomics to characterize variant peptides. However, the databases are greatly increased in size by tens of thousands of SAP-containing sequences, many of which are not expressed in the sample. This results in a concomitant increase in the false positive rate and a decrease in peptide identification sensitivity [18, 21, 35]. These problems were overcome in two studies that used RNA-Seq data to build SAP databases customized for a sample, enabling the detection of dozens of SAP peptides, including peptides containing novel variants resulting from either rare SNVs or de novo mutations [36, 37]. These studies showed how rapid advances in next generation sequencing technologies and the ease with which scientists can empirically measure all the coding SNVs in a sample can be harnessed to expand the detection of SAPs on a proteome-wide scale.

Here, we build upon those studies by comprehensively investigating SAP peptide detection in the Jurkat human cell line. This study follows from previous work in which we used RNA-Seq data to detect novel splice-junction peptides [38]. We collected deep coverage RNA-Seq data from the Jurkat cell line, compiled the set of nsSNVs that are expressed, used this information to construct a customized SAP database, and searched it against deep coverage shotgun MS data obtained from

the same sample. The SAP peptides identified from this customized database workflow were of much higher quality as compared to those identified using a larger aggregate database that incorporates all known nsSNVs (dbSNP). We employed multiple protease digestions to increase proteomic coverage and, thus, the number of SAP peptide identifications. These detected SAP peptides represent the most comprehensive study to date. Using this dataset, we describe various characteristics of the detected SAP peptides, including their corresponding transcriptional abundance, SNV functional effect scores, and degree of allele-specific expression.

3.3 Experimental Procedures

Mammalian cell culture

Jurkat cells (TIB-152) were grown in 10% Fetal Bovine Serum and 90% RPMI-1640 buffer at 37°C to a concentration of ~1.3×10^6 cells/mL (cell line and media were purchased from ATCC, Manassas, VA). In total, there were 12 flasks each containing 25 mL of Jurkat cell suspension. Upon harvesting, cell viability for each flask was determined with the trypan blue assay and cells were counted on a TC10 Automated Cell Counter system (BioRad, Hercules, CA). All cell cultures had 95%+ viability.

Mass spectrometry sample preparation and data collection

The proteomic sample preparation has been described previously in detail [38]. Briefly, Jurkat cell suspension was pelleted and rinsed twice in cold PBS buffer before storage at -80°C. Cell lysis was performed by following the FASP protocol [39]. Pellets were solubilized in SDT lysis buffer (4% w/v SDS, 100 mM DTT, 50 mM Tris-HCl), heated, sonicated, and 150 μg aliquots of protein were transferred to a 100K MW Amicon Ultra filter (Millipore, Billerica, MA). For this study, the FASP protocol was slightly modified to allow for multiple enzymatic digestions. The FASP method was followed for initial wash steps, alkylation, and the last three wash steps, which employed 50 mM ammonium bicarbonate. Then, each filter

was washed with two additional rounds of buffer compatible with a protease and the enzyme was added directly to the filter as listed here: 3 μg of trypsin (50:1 protein to enzyme ratio) in 50 mM ammonium bicarbonate at 37°C for 16 hours (Promega, Madison, WI); 1.5 μg of rLysC (100:1) in 25 mM Tris-HCl pH 8.5, 1 mM EDTA, 4 M urea at 37°C for 16 hours (Promega, Madison, WI); 1.5 μg of ArgC (100:1) in 270 μL of 50 mM Tris-HCl pH 7.6, 5 mM $CaCl_2$, and 2 mM EDTA and 30 μL of 50 mM Tris-HCl pH 7.6, 50 mM DTT, and 2 mM EDTA at 37°C for 16 hours (Promega, Madison, WI); 1.5 μg of AspN (100:1) in 50 mM sodium phosphate pH 8.0 at 25°C for 16 hours (Roche, Indianapolis, IN); 1.5 μg of GluC (100:1) in 25 mM ammonium bicarbonate at 25°C for 16 hours (Roche, Indianapolis, IN); and 1.5 μg of chymotrypsin (100:1) in 100 mM Tris-HCl pH 8.0 and 10 mM $CaCl_2$ at 25°C for 4 hours (Promega, Madison, WI). The final volume for each digestion was approximately 400 μl. At the end of the incubation time, each filter was centrifuged at 14,000 g for 15 minutes and the amount of peptide recovered was quantified via the Nanodrop UV-Vis spectrometer (Thermo Fisher Scientific, Wilmington, DE).

At least 100 μg of peptide digest was fractionated on a Shimadzu HPLC system (LC-10AD, SCL-10A VP, SPD-10A VP, Shimadzu, Columbia, MD) using a Phenomenex C18 Gemini 3μ, 110Å, 3.0×150mm column (Phenomenex, Torrance, CA) and high pH mobile phases. Mobile phase A (MPA) was aqueous 20 mM ammonium formate pH 10, and B (MPB) was 20 mM ammonium formate pH 10, in 70% acetonitrile. The HPLC flow was 0.5 mL/min and the gradient was as follows: 0% MPB isocratic for 15 minutes (trapping step), linear ramp to 100% MPB over 60 minutes, hold at 100% MPB for 5 minutes, to 0% MPB over 2 minutes, and equilibration at 0% MPB for 20 minutes. A Gilson 203 fraction collector (Gilson, Middleton, WI) was used to collect 28 fractions for the tryptic digest and 11 fractions for each of the LysC, ArgC, AspN, GluC, and chymotrypsin digests during detected (214 nm UV absorbance) peptide elution. Fractions were dried down using vacuum centrifugal concentration (Savant SpeedVac, Thermo, Pittsburgh, PA) and stored at -80°C.

Each of the dried down fractions were reconstituted in 2% acetonitrile and 0.2% formic acid in water and then chromatographically separated on a nanoAquity

LC system (Waters, Milford, MA) using a 20 cm reverse phase capillary column (100 μm i.d.) packed with 3 μm MAGIC aqC18 beads (Bruker-Michrom, Auburn, CA). Mobile phase A was 0.2% formic acid in water and B was 0.2% formic acid in acetonitrile. The full HPLC method was 180 minutes long and included online trapping, a 90 minute gradient, and re-equilibration time. A Velos-Orbitrap mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) was programmed to collect a full scan (300-1500 m/z) at a resolution of 60,000 followed by the top ten precursor HCD fragmentation spectra at a resolution of 7,500. Precursor fragmentation repeat count was set to two and the dynamic exclusion was set to 60 seconds. XCalibur software version #2.1.0 was used for data collection.

RNA Sequencing

The RNA-Seq data collection has previously been described in detail [38]. Briefly, total RNA was extracted from a 2 mL aliquot of each Jurkat culture (~2.6×10^6 cells) using the TRIzol®Reagent (Life Technologies, Grand Island, NY) and the RNA integrity was evaluated on a 2100 Agilent Bioanalyzer (Agilent, Santa Clara, CA). Illumina paired-end libraries were prepared for each of 12 samples using the TruSeq RNA Sample Prep Rev. A (kit lot #6849988, Illumina, San Diego, CA). Briefly, mRNA was isolated with poly dT beads, fragmented, reverse transcribed to cDNA, and then cDNA ends were repaired, adenylated, and ligated to Illumina adapters. The cDNA library was run on an Invitrogen 2% Size Select Gel (Lot# R19090-01) and a ~350 base pair band was excised and sequenced on an Illumina HiSeq 2000 in paired-end mode (2×100bp). An average of 12 million reads were generated per sample, and some samples were run multiple times, resulting in a total of ~300 million reads.

RNA-Seq data analysis

Bowtie/Tophat RNA-Seq read alignment

RNA-Seq reads were aligned to the human reference genome (hg19) using Bowtie (v0.12.7) and Tophat (v1.4.0) [40, 41]. Alignments were performed within Tophat, which uses Bowtie. The Tophat mate inner distance was set to 150. All other parameters were default. RefSeq gene models were supplied in GTF format and reads were aligned to both RefSeq genes and novel genes (option -G). RefSeq is NCBI's curated, non-redundant reference sequence database and includes DNA, RNA, and protein sequences and annotations [42]. The binary alignment or BAM file was used for subsequent SNV calling.

SAMtools SNV calling

SAMtools (v0.1.18) was used to call SNVs, nucleotide differences between the aligned RNA-Seq reads and the human reference genome. The mpileup command was used with the -u and -D options. Bcftools was then used (-bvcg options) to format the binary call format or BCF file. Finally, the SAMtools vcfutils.pl script was used to create a variant call format or VCF file. Only SNVs with a read depth (DP) higher than 10 and a quality score (QUAL) higher than 10 were used for subsequent analysis. QUAL is a phred-scaled score that reflects the confidence of the SNV call.

All RNA-Seq data processing was performed on the Phoenix cluster at the University of Wisconsin-Madison Chemistry department.

Retrieval of amino acid polymorphisms

The variant_effect_predictor.pl Perl script (version 2.7) downloaded from Ensembl along with the human annotation file (Ensembl v72) was used to convert the SNVs to amino acid coordinates and retrieve the calculated SIFT and PolyPhen-2 scores [43]. Only SNVs passing the DP and QUAL filters were used. Each SNV coordinate contained the chromosome, chromosome position, forward strand reference nucleotide, and forward strand alternative nucleotide. After analysis, the program output a

variant effect predictor (VEP) formatted file containing all the non-synonymous SNVs, and each entry included the corresponding amino acid change, the amino acid index within a RefSeq protein sequence, and the associated SIFT and PolyPhen-2 score.

Construction of a customized SAP database

SAP coordinate information was converted into a customized SAP FASTA database. Within the VEP file output from the previous step, SNVs that resided within RefSeq protein coding regions were retrieved. The RefSeq protein FASTA file was downloaded from NCBI's FTP site (ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/mRNA_Prot/human.protein.faa.gz, release 59) [42]. For each coding SNV, the reference and alternative nucleotide and its position within the genome was listed, as well as the reference and alternative amino acid and position within a RefSeq protein entry (NP accession). An in-house perl script was used to extract an 80 aa substring containing the SAP and change the reference aa to the variant aa. A FASTA header including the amino acid change and position within the RefSeq NP entry was linked to each SAP-containing sequence and all these sequences were appended to the RefSeq protein and cRAP FASTA file. cRAP or the common Repository of Adventitious Proteins is a database of protein sequences that are found as contaminants in proteomics experiments (http://www.thegpm.org/crap/).

Construction of a SAP database from the dbSNP repository

For comparison purposes, a FASTA file containing SAPs derived from NCBI dbSNP repository was constructed. The ASN-1 flat file containing all 53,233,155 dbSNP rs entries for human was downloaded from NCBI's ftp site (/snp/organisms/human_9606, build 137) and the 691,356 rs entries representing missense mutations (fxnclass = missense) were retrieved. Each rs entry lists the reference and alternative amino acid position within a RefSeq protein entry. An in-house perl script was used to extract an 80 aa substring containing the SAP and change the reference aa to the variant aa. A FASTA header including the amino acid change and its position

within the RefSeq NP entry was added to each dbSNP-SAP-containing sequence and all these sequences were appended to the RefSeq protein and cRAP FASTA file.

Mass spectrometry searching

Raw mass spectrometry files were searched against the customized SAP+RefSeq+cRAP and the dbSNP-SAP+RefSeq+cRAP FASTA files using the SEQUEST/Percolator search algorithm within ProteomeDiscoverer (v1.3.0.339, Thermo Fisher Scientific, San Jose, CA). Default peaklist-generating parameters were used. Precursor m/z tolerance was set to 10 ppm and product m/z tolerance was set to 0.05 Da. Peptides with up to two missed cleavages (proteolytic) were permitted. Variable methionine oxidation and static carbamidomethylation were used. Using reversed sequences as a decoy database, peptides passing both a 1% and 5% global FDR were used for downstream analysis. Validation was based on q-values generated by Percolator. For identification of a protein using ProteomeDiscoverer, protein grouping and strict parsimony principle was enabled, leucine and isoleucine were considered equal, and only peptides passing a 1% FDR and having a delta Cn higher than 0.15 were used. Each peptide identification counted only if that peptide had a unique primary sequence. A minimum of two peptides per protein was required for identification. MS data collected from alternative enzymatic digests were separately searched against the customized SAP+RefSeq+cRAP FASTA file with identical parameters to the trypsin search except with the relevant enzyme specificity.

Estimation of allele-specific protein expression

Using Skyline software (v1.4)[44], MS1 extracted ion chromatograms were integrated for heterozygous peptide pairs that had a high degree of structural similarity (same length, only one amino acid difference). Only peaks that overlapped a target peptide MS^2 identification, contained minimal background interference, and had an appropriate chromatographic peak shape were accepted. Default Skyline parameters for peak integration were used.

3.4 Results

Overview

Each human cell line or tissue sample contains thousands of non-synonymous SNVs (nsSNVs) that give rise to single amino acid polymorphisms (SAPs); however, these variations are typically absent from generic proteomic databases. Therefore, sample-specific peptides containing these SAPs fail to be identified during mass spectrometry searching. Fortunately, RNA-Seq can be used to experimentally detect the nsSNVs in a sample, which allows for the creation of a customized SAP database, thereby enabling identification of SAP peptides [37].

Here, we describe the comprehensive detection and evaluation of SAP peptides from a human cell line. We created a customized SAP database using RNA-Seq data collected from Jurkat cells that enabled the detection of 421 SAP peptides mapping to 395 nsSNV sites. For comparison purposes, we constructed an all-inclusive SAP database derived from all known human nsSNVs (NCBI's dbSNP) leading to the identification of 891 SAP peptides. Though there were a higher number of SAP peptides passing a 1% FDR using this all-inclusive database, we show that the peptide spectral matches (PSMs) were of much lower quality, indicating a false positive issue. After this finding, we proceeded to determine the extent of SAP peptide detection using the customized database. We employed multiple protease digestions to increase proteomic coverage and thus identified 695 SAP peptides mapping to 504 nsSNV sites (9% of total nsSNVs, 504/5755). These SAP peptides corresponded to transcripts with a median of 44 transcripts per million, indicating that they are derived from moderate to high abundance transcripts. For all the SAP peptides, we report the computationally predicted functional effect scores (SIFT, PolyPhen-2). And last, the detected SAP peptides included 192 allelic pairs, in which the reference and SAP peptide were both detected; we measured the relative allele-specific expression for 51 of these pairs.

Construction and use of the customized RNA-Seq database

RNA-Seq data was collected from Jurkat cell culture and used to create a customized SAP database used for MS searching. The detection of variant peptides from SAP databases is shown in Figure 1 and the bioinformatic workflow numbers are shown in Figure 2.

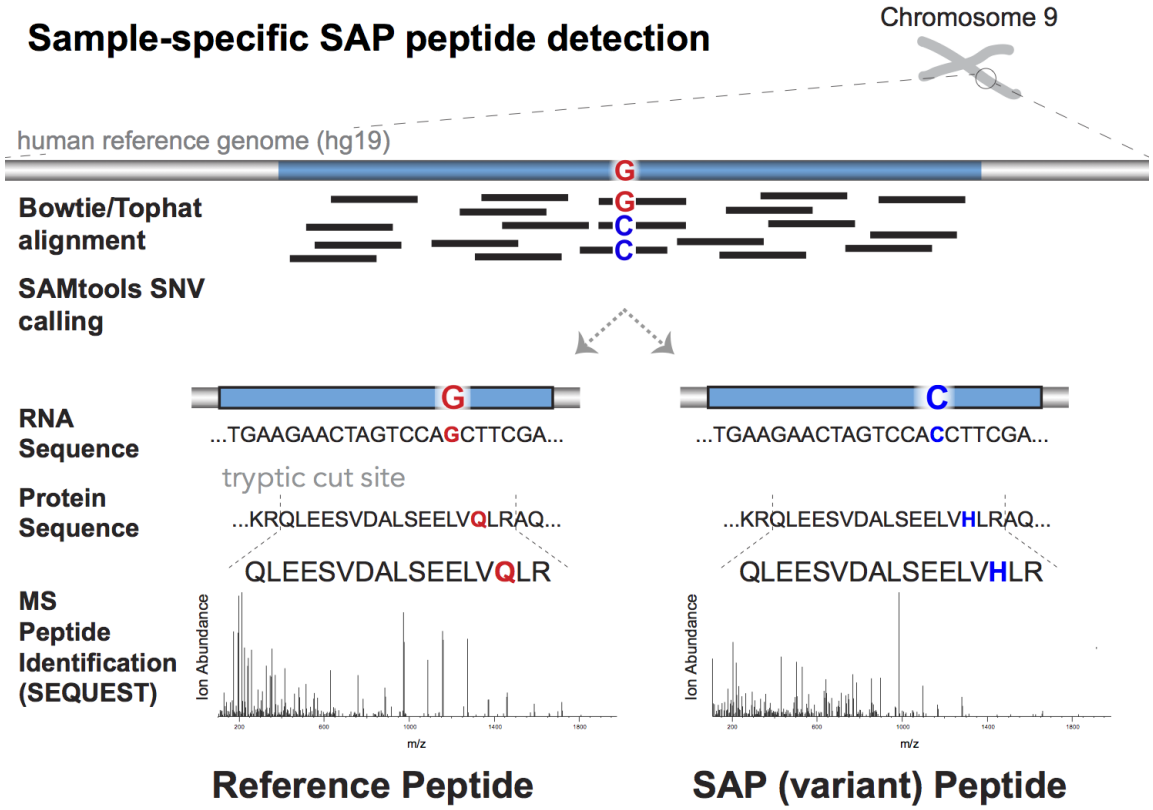

Figure 3.1. Overview of sample-specific SAP peptide detection from custom databases. Single nucleotide variants (SNVs) are detected directly from RNA-Seq reads by finding differences between the transcript and human reference genome nucleotide sequences. The set of non-synonymous SNVs are converted into amino acid sequences that are consolidated into a customized protein database that is used for MS searching. Here, both the reference and variant (SAP) peptides are detected, demonstrating that both allelic forms are expressed at the protein level.

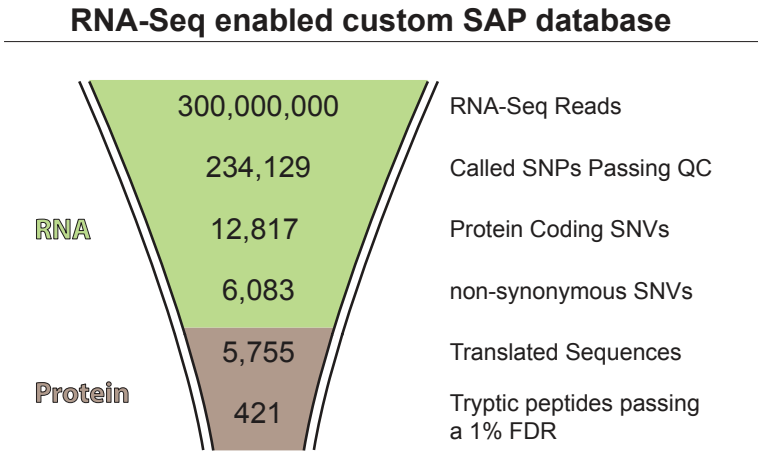

Figure 3.2. Bioinformatic workflow numbers for customized SAP database construction and subsequent MS search results.

First, RNA-Seq and MS data was collected from Jurkat human cell culture. Total RNA was extracted from several Jurkat cultures (95%+ viability, trypan blue) using the TRIzol ®method and each sample was used to create a barcoded Illumina cDNA library using the TruSeq protocol. Each library was sequenced at least once on an Illumina HiSeq 2000, resulting in a total of ~300 million paired end reads (350bp, 2×100bp). Protein was extracted and digested from the Jurkat cultures using the FASP method and the resulting peptides were fractionated via a high pH HPLC and run on a nanoLC-Velos Orbitrap operating in data-dependent mode. Approximately 500,000 mass spectra were collected.

The RNA-Seq data were analyzed to find Jurkat cell-specific SNVs. Bowtie and Tophat were used to align the RNA-Seq reads to the human reference genome (hg19). RefSeq gene models were used to guide alignment, but reads that aligned to novel genes were also allowed. 82.8% of the singletons (one member of the read pair) and 67.7% of the full read pair were successfully aligned. All read alignments were stored within a binary alignment (BAM) file (61.41 GB). Next, SAMtools (mpileup command) was used to call SNVs. Here, the genome is traversed one nucleotide at a time and for each nucleotide position, the reads overlapping a nucleotide is

examined. If there is evidence that the nucleotide sequence within the RNA-Seq reads differ from the nucleotide in the reference genome with statistical significance, an SNV is "called" or reported. After SNV calling, several quality metrics are used to filter SNVs, including the quality of the nucleotides at the SNV site, the score of the read alignment, and the depth (i.e. coverage) of the reads. From the mapped reads in this study, a total of 473,868 SNVs were called while 234,129 SNVs passed quality filtersâĽ"read depth (DP) of 10 or higher and quality score (QUAL) of 10 or higher. Figure 3 shows the distribution of read depth versus quality score for all the SNVs called, with filtered out SNVs shaded in gray.

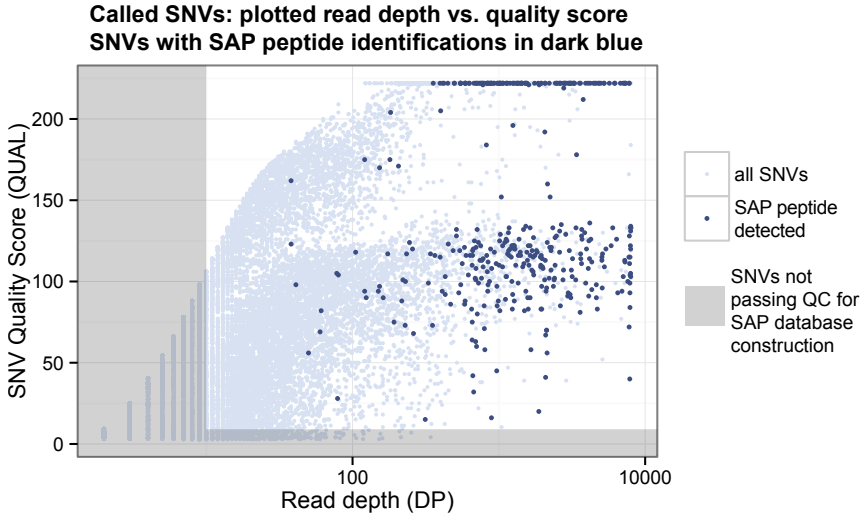

Figure 3.3. Plot of RNA-Seq read depth versus quality score for each called SNV. This graph shows the distribution of depth and quality scores for the SNVs called using SAMtools, with discarded SNVs highlighted in gray. The bimodal shape is due to the presence of homozygous (top portion) and heterozygous (bottom portion) alleles. The nsSNVs that resulted in a SAP peptide identification are dark blue. These nsSNVs tend to be of higher read depth and quality.

Of the 234,129 SNVs that passed quality filters, 12,817 SNVs were found to reside within RefSeq protein coding regions. 6,535 (52%) were synonymous SNVs and 6,083 (47%) were non-synonymous SNVs (nsSNVs). These percentages are similar to percentages reported by the 1000 Genomes Project (55% synonymous, 45%

non-synonymous; average values from 1,092 individuals) [1]. The high proportion (94.5%) of SNVs that did not reside in coding regions were predominantly located within UTRs or introns, and this was especially true for nucleotides near the 3′ end of the transcript. This suggests that many untranslated SNVs are detected from incompletely spliced mRNAs that were isolated during the polydT bead enrichment step of the Illumina library preparation protocol.

The set of nsSNVs found in the RNA-Seq data was used to derive all SAP-containing polypeptide sequences in RefSeq. To accomplish this, each amino acid position and index within the RefSeq protein sequence (NP accession number) was retrieved. For each SAP, a custom Perl script was used to extract an 80 aa subsequence containing the SAP position, and the amino acid at that position was changed to the variant form. In a few cases (5%) the RefSeq protein sequence corresponded to the SAP encoded by the nsSNV. This is because of minor discrepancies between the RefSeq and hg19 sequence data, due to their different origins —hg19 is the product of genome sequencing efforts, whereas RefSeq is derived from cDNA sequencing data. 5,755 SAP-containing sequences mapping to 3,837 distinct NP accessions were extracted and appended to the RefSeq protein (35,930 entries) and cRAP (155 entries) databases to create a customized SAP database. The SAP entries marginally increased the size of the database by 2.2% (442,740 aa added to 19,899,407 aa). 38% (2,162 entries out of 5,755) of the SAPs were not present in dbSNP and are likely to represent undocumented variations, including somatic mutations, rare variants, and variations exclusively in the RNA from RNA editing or RNA polymerase nucleotide misincorporations.

The RefSeq+cRAP+SAP database was searched against the MS data using the Percolator/SEQUEST algorithm. 73,552 peptides (each with unique sequences) were identified at a 1% FDR. From these, there were a total of 421 SAP peptides mapping to 395 unique SNVs, corresponding to 0.6% of all peptides. This percentage, representing the proportion of SAP peptides detected in a shotgun proteomics experiment, is similar to previous findings [37]; however, the present study identified over ten times the number of SAP peptides. The significantly higher number of SAP peptides identified is likely due to the deep proteomic sampling achieved

in this study. This suggests that even more SAP peptides could be discovered by the collection of deeper-coverage proteomics data. A list of the SAP peptide identifications may be found in Supplementary Table S1 in the Supplementary Information (SI).

The relative quality of peptide spectral matches (PSMs) was compared between RefSeq and SAP peptides. When MS searches are performed against proteomic databases that are augmented with putative sequences (e.g. splice junction sequences), there is an increased chance of false positives [38]. A typical indication that there are false positive issues is when peptides matching the non-canonical database (e.g. SAP peptide) have lower than expected MS search scores. Therefore, the average MS search scores —in this case, the SEQUEST XCorr score that represents the degree of match (via the cross-correlation function) between the theoretical and experimental MS^2 spectra —were compared between RefSeq and SAP peptides. Surprisingly, the SAP peptide XCorr scores, on average, were actually higher than the RefSeq peptide scores, indicating that the SAP peptide identifications are of high quality. Figure 4 shows these comparisons.

Construction and use of the dbSNP database

The nsSNVs listed in dbSNP were used to create an exhaustive SAP database, which was then used for MS searching. Key bioinformatic workflow numbers describing this process are shown in Figure 5.

NCBI's dbSNP is one of the largest repositories of known SNVs consolidated from various sources of data such as sequence tagged sites, Genbank, and the 1000 genomes project [33]. dbSNP was used to create an exhaustive SAP database for proteomic searching. A human dbSNP ANS-1 flat file containing all 53,555,486 entries was downloaded from NCBI's FTP site (May 3rd, 2013). Of those entries, 679,490 were classified as non-synonymous SNVs (fxn-class=missense) and 378,986 as synonymous (fxn-class=synonymous). The 679,490 non-synonymous SNVs mapped to 33,557 distinct RefSeq NP sequences and, therefore, the dbSNP nsSNVs covered nearly all RefSeq protein sequences.

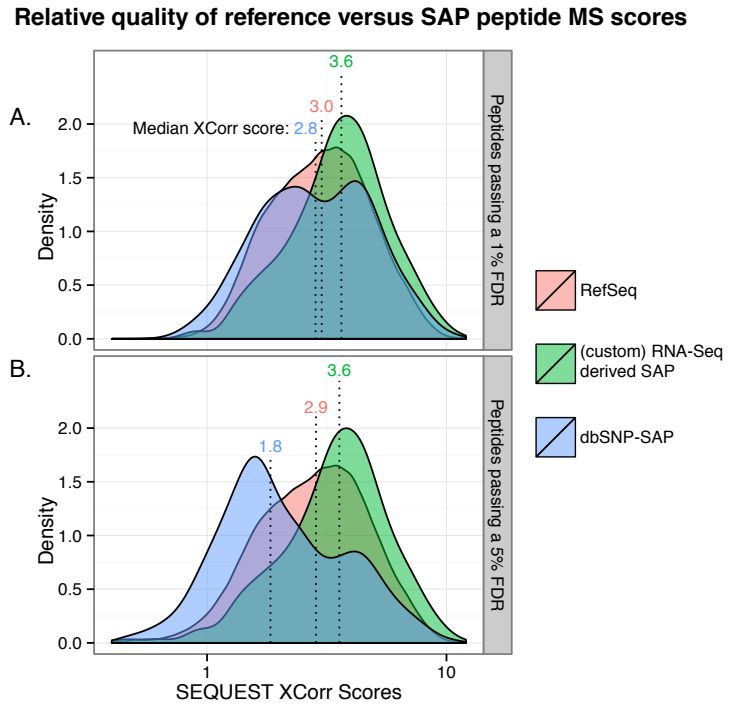

Figure 3.4. Comparison of average XCorr scores for peptides matching the RefSeq protein, dbSNP-SAP, or custom (RNA-Seq) SAP database. SAP peptides identified from the custom database tended to have higher XCorr scores than those identified from the dbSNP database. Score distributions for peptides passing a 1% FDR (A) and 5% FDR (B) are shown.

A SAP-containing polypeptide sequence was created from the SNV coordinate information listed in each dbSNP entry. Using the dbSNP nsSNV coordinate information, a custom Perl script was used to extract, from the RefSeq protein entry, the 80 amino acid stretch of protein sequence containing the SAP and to change the amino acid to reflect the variant form. Each entry was created in FASTA format and the header included the chromosome and protein position of the nucleotide and amino acid change, respectively. In total, 691,356 dbSNP-SAP entries were created. Some dbSNP entries contained two or more alternative alleles, thereby generating multiple SAP entries from a single dbSNP. The dbSNP-SAP entries were appended to the RefSeq protein (35,930 entries) and cRAP (155 entries) databases to create the

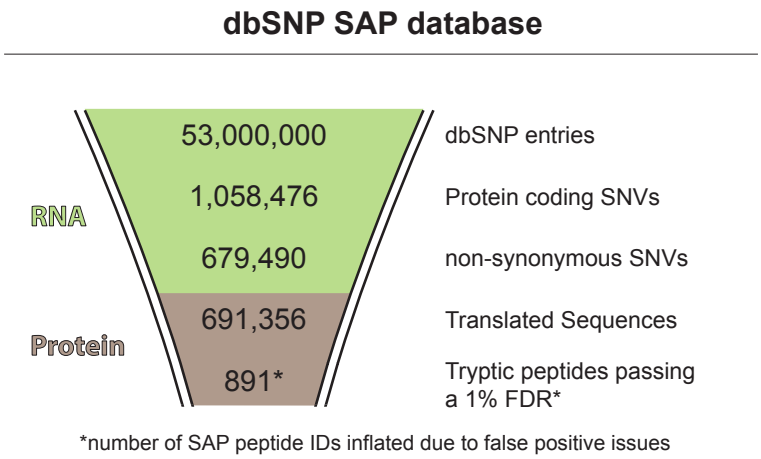

Figure 3.5. Bioinformatic workflow numbers for the dbSNP-derived SAP database construction and MS search. Though more SAP peptides were detected using the dbSNP database, the peptide identifications had low peptide spectral match (PSM) scores, indicating a false positive issue.

dbSNP-SAP database. The dbSNP-SAP entries drastically increased the size of the database by 268% (53,233,115 aa added to 19,899,407 aa).

The RefSeq+cRAP+dbSNP-SAP database was searched against the MS data using the Percolator/SEQUEST algorithm. 72,250 RefSeq peptides (each with unique sequences) were identified at a 1% FDR. A total of 891 dbSNP-SAP peptides were identified. An additional 652 dbSNP-SAP peptides were identified at a 5% FDR threshold. A list of the dbSNP-SAP peptide identifications may be found in Supplementary Table S2 in the SI. Though at first glance it may seem that more SAP peptides were identified with the dbSNP-SAP database, there were false positive issues that bring into question the quality of these peptide identifications. This topic is discussed in the next section.

Comparing RNA-Seq and dbSNP-derived SAP peptides

The dbSNP-SAP database represents all the nsSNVs found in any number of different human cell and tissue types, whereas the custom SAP database derived herein is from a single sample-matched RNA-Seq dataset and represents the set of nsSNVs that exist in this particular single cell-line. Although use of an aggregate database, such as the set of dbSNP-derived SAPs, obviates the need to collect sample-specific RNA-Seq data, these databases contain an extremely large number of polypeptide sequences that do not exist in the sample. Inclusion of a large number of extraneous sequences in proteomics databases increases the probability that a theoretical mass spectrum derived from an extraneous peptide sequence falsely matches to an experimental mass spectrum by mere chance, a well-known phenomenon [45].

A strong disadvantage of using an aggregate database, like the dbSNP-derived SAP database, is that there are many false positives in the set of SAP peptides identified. Evidence for this phenomenon can be seen in the comparison of MS search score distributions of the RefSeq and SAP peptides. Figure 4A shows that for peptides passing a 1% FDR, the median XCorr score for RefSeq (canonical) peptides was 3.0: The custom SAP peptides had a median value of 3.6, which was even better than the RefSeq median, but, notably, the dbSNP-SAP peptides had lower XCorr scores, a median of 2.8. These trends for RefSeq, custom SAP, and dbSNP-SAP were even more pronounced when comparing median XCorr scores for peptides passing a 5% FDR, that is, 2.9, 3.6, and 1.8, respectively (Figure 4B), underscoring both the high quality of RNA-Seq derived custom SAP peptide identifications, and the low quality and higher number of false positives within the dbSNP-SAP peptide identifications. Note that the peptide posterior error probabilities (PEP) and q-values for the peptide groups also showed similar trends (Figures S1 and S2).

We examined the extent of overlap in peptide identifications between RNA-Seq versus dbSNP-derived SAP peptides. Venn diagrams are shown in Figure 6. A large fraction of the RNA-Seq SAP peptides (42% of peptides passing a 1% FDR) were not present in the dbSNP database, showing that despite dbSNP's large size, it still does not include every SNV in this particular human cell line. Moreover,

it is reasonable to assume that aggregate databases, as they stand today, would fail to detect a number of variants in other cell or tissue types, as many SNVs are yet to be documented. Conversely, a large fraction of dbSNP-SAP peptides (73% of peptides passing a 1% FDR, and 84% passing a 5% FDR) lacked evidence of expression in the deep coverage RNA-Seq data and, hence, are most likely false positives. This would suggest that the nominal false discovery rates for 1% and 5% FDR passing dbSNP-SAP peptides are actually 73% and 84%, respectively. While the total number of dbSNP-SAP peptides identified is greater than the number of RNA-Seq SAP peptides identified, the exceedingly high actual false positive rate compromises their utility.

Next, we asked if the dbSNP-SAP peptide false positive issue could be remedied by applying more stringent peptide identification thresholds. It is well known that MS searches against extremely large databases tend to produce many false positive peptide identifications, and various strategies have been developed to reduce the incidence of false positives, including sequential (multi-tiered) MS searches and calculation of local FDRs [45, 47]. We calculated a local FDR for the dbSNP-SAP peptides by utilizing posterior error probability (PEP) values (see Supplemental Table S2)[38, 48]. We found that even with the application of a local FDR threshold, the dbSNP-SAP peptide score distributions were still slightly shifted to lower values (Figure S3). And, more importantly, applying the local FDR cut-off did not eliminate many false positive dbSNP-SAP peptides, as shown in the Venn diagrams in Figure 6B, where more than 70% of dbSNP-SAP peptides were not present in the RNA-Seq data and are therefore likely to be false positives.

The coverage and accuracy of the SAP peptide identifications must be high to be of use in biological applications such as the confirmation of nsSNV translation. These results show that utilizing sample-matched RNA-Seq data to identify SAP peptides offers significant advantages in these respects.

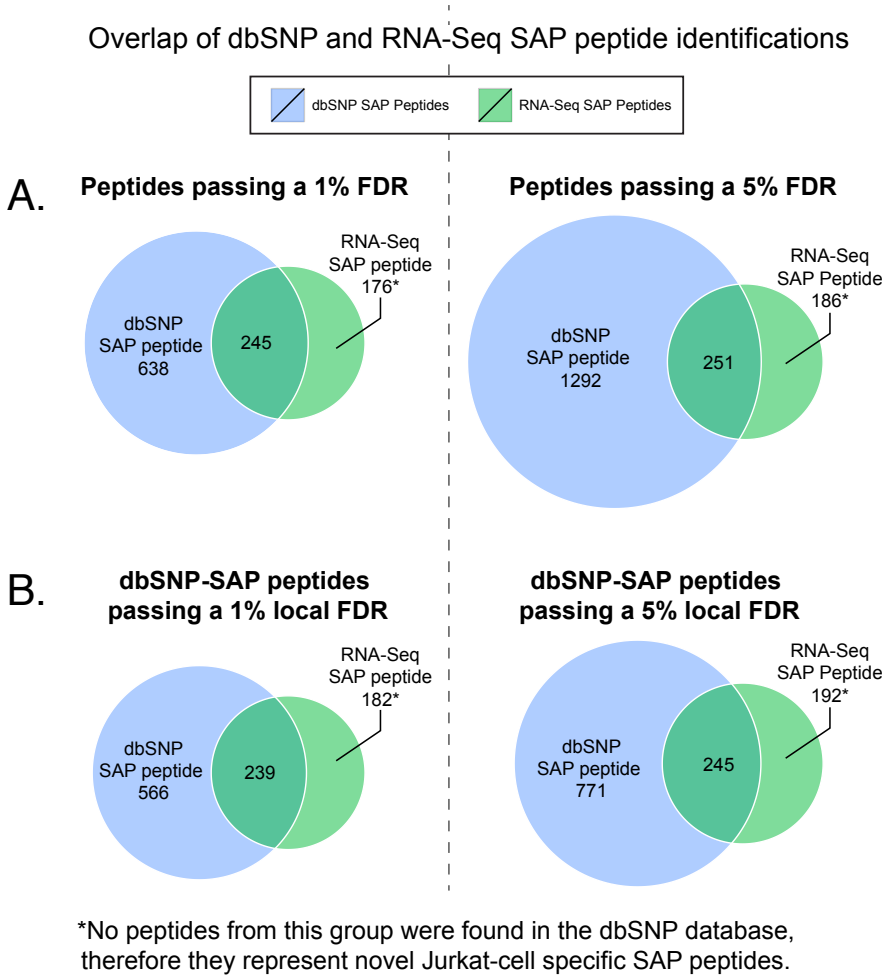

Figure 3.6. Comparison of dbSNP versus RNA-Seq derived SAP peptide identifications. Venn diagrams show the overlap of SAP peptides identified from MS searching. For example, 245 SAP peptides passing a 1% FDR were identified in both the dbSNP and RNA-Seq SAP database searches. (A) dbSNP-SAP and RNA-Seq SAP peptides passing global FDRs, (B) dbSNP-SAP peptides re-analyzed to pass a local FDR and then compared to the same RNA-Seq SAP peptides. The terms "local" and "global" FDR are explained by Käll, et al. [46]

Multiple protease digests to expand SAP peptide detection

It was shown above that 395 SNV sites were detected at the protein level from searching the custom (RNA-Seq derived) SAP database against MS data collected on tryptically-digested lysate. As far as we know, this is the largest number of SAP peptides detected for a single human cell line. However, these SAP peptides represent only 6.9% (395/5755) of all possible translated nsSNVs. Of the 5755 total SAP sequences, 4325 contain SAP peptides that are between 6 and 39 amino acids, the typical range of peptide lengths that are identified in shotgun proteomics studies. Using this reduced number, a larger fraction of length-filtered SAP peptides were identified, specifically 9.7% (395/4325). Assuming that the nsSNVs detected at the RNA level are indeed translated into protein, these results provide a good estimate of the proportion of nsSNVs corresponding to detectable SAPs.

We asked what fraction of nsSNVs could be detected at the protein-level with shotgun proteomics. To explore this question, we collected high coverage proteomics data by employing multiple protease digestions. Jurkat cell lysate was separated into five aliquots and was digested with either LysC, ArgC, AspN, GluC, or chymotrypsin. Each of the five peptide digests were fractionated on a high pH HPLC and analyzed on a Velos-Orbitrap mass spectrometer in data dependent mode, and each dataset was searched against the RefSeq+cRAP+SAP database. Similarly to the trypsin-derived SAP peptides, the SAP peptides had higher XCorr distributions than RefSeq peptides on average (Figure S4). Figure 7 shows the peptide and SNV site identification results. Note that the trypsin dataset was based on 28 high pH HPLC fractions whereas the datasets for the other enzymes were based on 11. The number of SAP peptides with unique sequences was calculated for cumulative combinations of proteolytic search results. For example, 508 unique SAP sequences were found with combined trypsin and LysC data and 547 unique SAP sequences were found with combined trypsin, LysC, and ArgC data. When the multiple protease data was compared with the original tryptic dataset, the number of unique SAP peptides increased by 65% while the number of unique nsSNV sites for which there was direct peptide evidence increased by 28%. In

other words, while data from all six enzymes detected 695 unique SAP peptide sequences, these peptides corresponded to only 504 unique nsSNV sites. These results suggest that higher coverage shotgun proteomics data increases the number of identified SAP peptides with unique sequences, but that many of these SAP peptides are repeatedly sampling the same set of SNVs. All multiple protease SAP peptide search results may be found in Supplementary Table S3 in the SI.

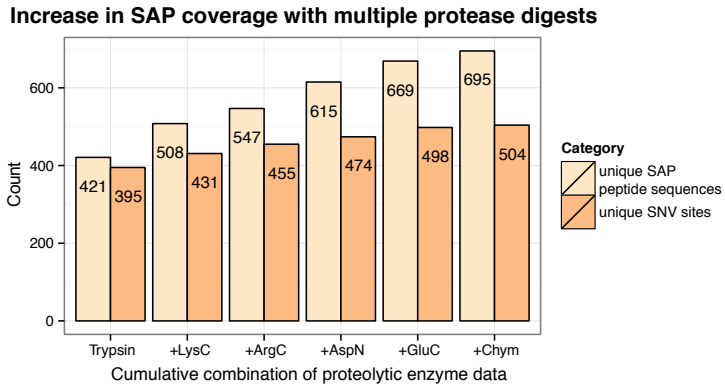

Figure 3.7. Cumulative number of identified SAP peptide and nsSNV sites with consolidated protease digest data. The enhanced protein coverage afforded by multiple protease digestions increased the number of translated nsSNVs detected by 28%.

Transcript abundances for detected SAP peptides

With high coverage proteomic data, 8.8% (504/5755) of the total number of possible nsSNVs were identified at the protein level. This represents a much higher fraction of detected SAP peptides as compared to previous studies [21, 24, 27, 35, 37], but it lags in comparison to the SNV detection sensitivity afforded by next generation sequencing technologies. MS-based proteomics can only detect a small fraction of all possible protein-level variants within a sample. To understand why, the abundance distribution, in transcripts per million (TPM), was plotted for all transcripts and for transcripts in which the corresponding protein was identified (Figure 8). The median TPM for transcripts with a protein identification was much

higher than the median TPM for all transcripts. Two reasons for this are: first, some lower abundance transcripts are not translated, especially for transcripts that are stochastically expressed, and, second, mass spectrometry is not as sensitive as RNA-Seq and the sampling depth of peptides is limited by many factors such as peptide ionization efficiency, sample complexity, and the MS duty cycle. The abundance distribution for transcripts in which there was a detected SNV was also plotted and compared to the abundances of transcripts for which there was a detected SAP peptide (Figure 8B). This plot shows that SAP peptides are primarily detected from highly expressed transcripts and suggests that as MS sensitivity and sampling depth increases, the number of SAP peptides detected will also increase.

Computationally predicted functional effect scores

The functional consequence of a given SNV can be computationally predicted using a variety of tools such as SIFT and PolyPhen-2 [7, 8]. SIFT examines the degree of evolutionary conservation of the nucleotide polymorphism and depends on the assumption that an SNV found in a highly conserved genomic region is more likely to affect the function of the protein. PolyPhen-2 examines the physicochemical properties of the amino acid change and how much this change affects conserved protein domains. Because the number of discovered SNVs far exceeds the number of SNVs that can be biologically validated, both SIFT and PolyPhen-2 are ubiquitously used to analyze and rank SNVs discovered in genome research.

We were interested in evaluating the functional predictive scores for both the RNA and protein-level SNVs. We used Ensembl's Variant Effect Predictor (VEP) program to retrieve the SIFT and PolyPhen-2 scores for each nsSNV (see Supplementary Table S4 in SI). The distribution of SIFT and PolyPhen-2 scores for nsSNVs detected at the RNA level and the subset of nsSNVs that was detected at the protein level, as evidenced by a SAP peptide ID, were similar. Figure 9 shows histograms of both SIFT and PolyPhen-2 score distributions. 27% of all nsSNVs and 29% of nsSNVs with peptide evidence had a SIFT score less than 0.05, which is categorized as "deleterious". 16% nsSNVs and 14% of nsSNVs with peptide evidence had a

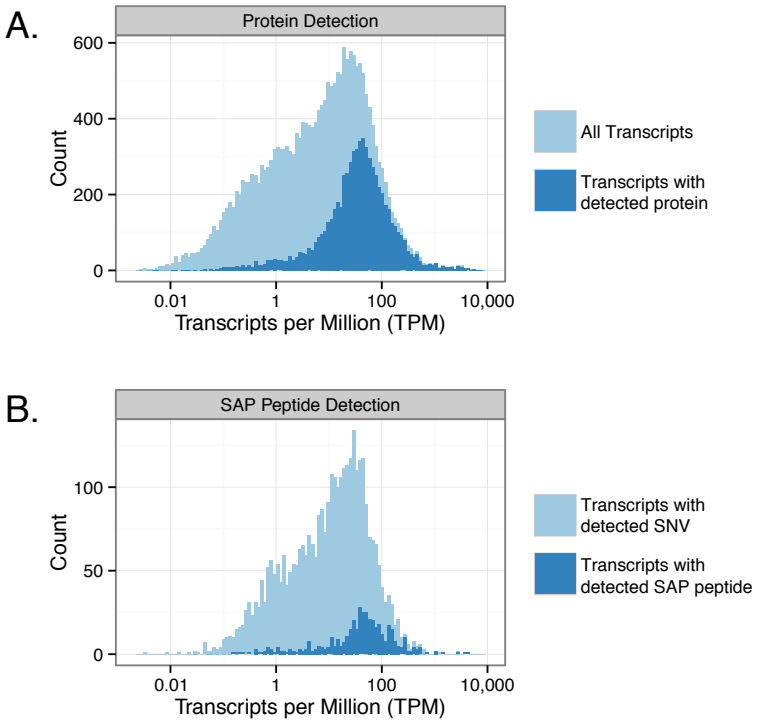

Figure 3.8. Distribution of transcript abundances for transcripts encoding detected proteins and transcripts encoding detected SAP peptides. (A) The abundance distribution for all transcripts (light blue) versus just those transcripts with a protein identification (dark blue). (B) The abundance distribution for transcripts with an nsSNV (light blue) versus just those transcripts with a detected SAP peptide (dark blue).

PolyPhen-2 scores greater than 0.903, which is categorized as "probably damaging".

RNA and protein allele-specific expression

In diploid organisms such as human, there are two copies of each chromosome, and thus each RNA or protein is derived from one of two alleles. When the gene is homozygous, the sequence of the allelic pair is identical and there is no way to distinguish which chromosomes the gene products come from. But when the gene is heterozygous, the sequences of the allelic pair are different and it is possible to

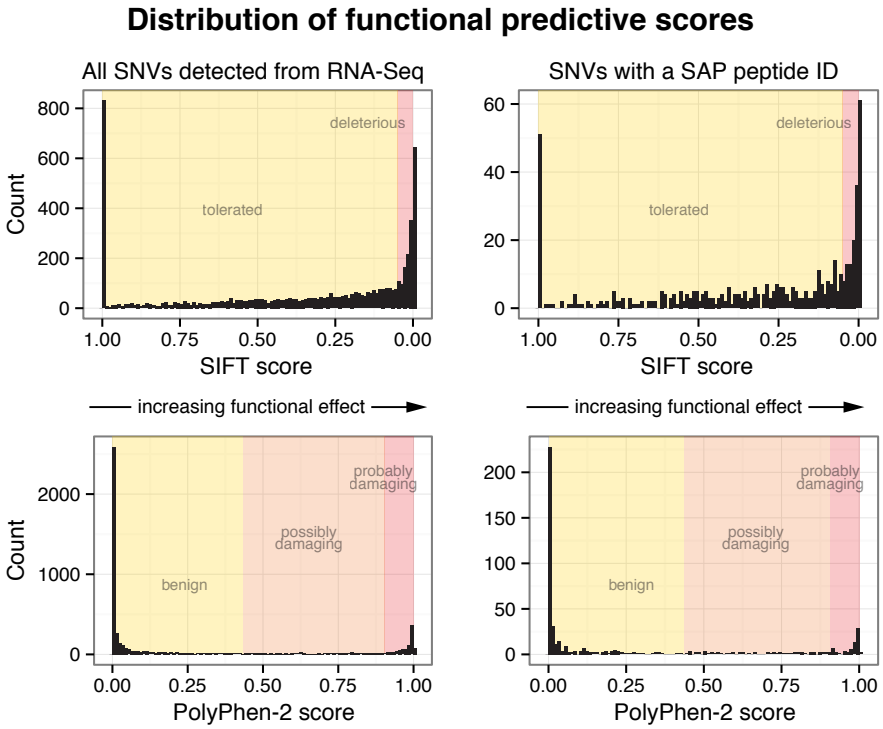

Figure 3.9. Comparing SIFT and PolyPhen-2 functional effect prediction scores between all nsSNVs and nsSNVS with a SAP peptide ID. The distributions were similar between the two groups.

track which gene an RNA or protein arose from by detecting the RNA-Seq read or SAP peptide containing the SNV or SAP, respectively. Additionally, it is possible to quantify allele-specific expression (ASE). The ASE at the RNA-level can be estimated by comparing the depth of reads mapping to the reference and alternative SNVs [49]. Analogously, the ASE at the protein-level can be estimated by quantifying the amount of reference and SAP peptide [50]. Previously, a SILAC-based approach was developed that allowed global quantification of ASE in yeast [51].

We examined the RNA-Seq and mass spectrometry datasets to identify, at the protein-level, the number of detected allelic pairs and to measure ASE. At least one SAP peptide was detected for each of 504 nsSNV sites, as shown in an earlier section

of this paper. Both the reference and SAP peptides were detected for 38% (192 out of 504) of those nsSNV sites showing that a significant number of heterozygous peptide pairs are readily detected by shotgun proteomics. The amino acid sequences of the heterozygous peptide pairs were either significantly different (e.g. the SAP introduces a lysine causing the SAP peptide to be much shorter than the reference peptide) or highly similar (e.g. the SAP is a single amino acid change in the middle of the peptide sequence). 74 heterozygous peptide pairs were found in the latter category. The peptides in these pairs have highly similar sequences (i.e. a difference of only one amino acid). They could be considered structural analogues of each other; the predicted HPLC retention time using SSRCalc [52] and the predicted ionization efficiency using ESPPredictor [53] between these pairs were found to be near-identical. We estimated the relative SAP to reference peptide concentrations by integrating the area of MS^1 extracted ion chromatograms using the Skyline program [44].

Figure 10 displays a plot of the estimated allelic expression for peptide and RNA-level heterozygous pairs. The reference to alternative peptide ratio was distributed around 1:1, for both the nsSNVs (RNA-Seq reads) or SAPs (peptide) measured. As expected, allele-specific peptide expression shows greater variability than allele-specific RNA expression due to MS variables such as electrospray current and complexity of the sample matrix (i.e. co-eluting pepties). Future work could utilize heavy-labeled internal standards and employ more precise methods of quantification to further explore allele-specific expression. All ASE results can be found in Supplementary Table S5 in the SI.

3.5 Discussion

The full repertoire of SNVs expressed in RNA can be detected using the latest sequencing technologies but the power to detect the corresponding SAPs at the protein level has been lagging. Direct detection of the SAP within a peptide (or protein) is important for understanding how variants influence biological phenomena such as post-transcriptional regulation and differential allelic expression. Little

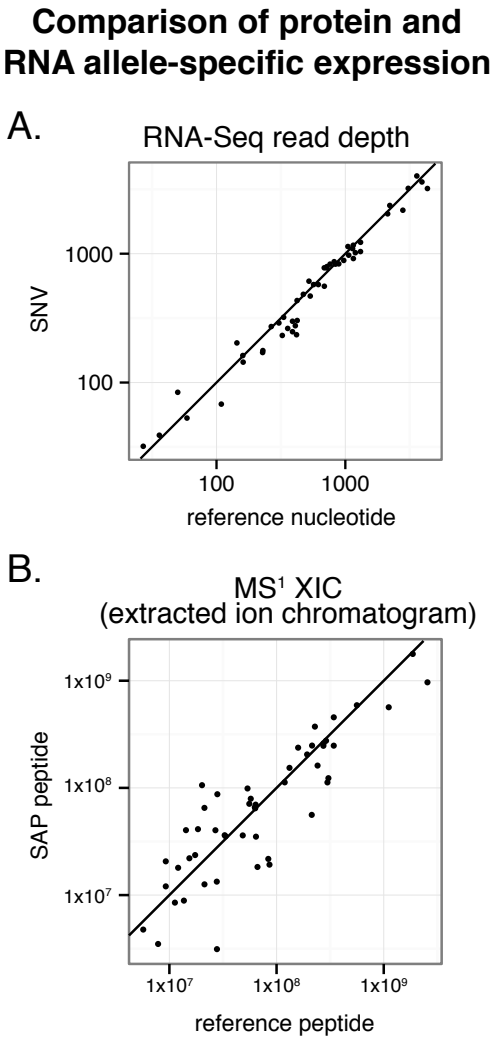

Figure 3.10. RNA and protein-level allele-specific expression. A line corresponding to 1:1 allelic expression has been overlaid. For both RNA (A) and protein (B), the expression levels for allelic pairs are roughly the same. Protein-level expression had higher variability.

work has been done to date to measure SAP peptides on a large-scale using mass spectrometry because the conventional strategy for identifying peptides is through database searches against a generic proteomic database that does not include the variant sequences.

We have described the large-scale detection of SAP peptides made possible through the construction of a customized SAP database from sample-matched RNA-Seq data. With the customized database, we confirmed the translation of hundreds of non-synonymous SNVs that were specific to the Jurkat cell line, representing the most comprehensive set of SAP peptide identifications to date. To determine how many SAP peptides are detectable by shotgun proteomics, we employed multiple protease digestions and collected even higher coverage proteomics data, allowing us to detect 695 sequence-unique SAP peptides corresponding to 504 unique nsSNV sites, or ~10% of all RNA-level nsSNVs (504/5755). These results illustrate that a significant number of SAP peptides are detectable through shotgun proteomics, but also indicate that further improvements in proteomics technologies are needed for them to equal the coverage of variants that can be obtained at the RNA level with next generation sequencing technologies.

The unusually high number of SAP peptides identified in this work along with the sample-matched RNA-Seq data provided us with the opportunity to analyze properties of nsSNVs and the SAP peptides identified via mass spectrometry. The SAP peptides, similarly to all peptides identified, corresponded to moderate to high abundance transcripts (30+ transcripts per million, TPM). The distribution of these detected SAP peptides' computationally predicted functional effects (e.g. SIFT, PolyPhen-2) was similar to the distribution for the complete set of all possible SAPs, indicating no selection of particular SAP types. Finally, for 192 out of the 504 SNVs, we detected both the reference and SAP peptides, confirming that a significant fraction of heterozygous alleles are expressed at the protein level. Related to this finding, we also investigated the feasibility of quantifying differential allelic expression on a large scale. Previously, SRM methods employing stable isotope labeled peptide standards were developed to quantify three allelic peptide pairs [54] and a small number of related mutant peptides [50, 55]. Here, we presented

preliminary label-free quantification of allele-specific expression based on the integrated MS1 extracted ion chromatograms from 51 allelic peptide pairs.

We compared the number and quality of SAP peptide identifications resulting from MS searches against (1) an aggregate SAP database derived from NCBI's dbSNP repository and (2) a customized SAP database derived from sample-specific RNA-Seq data —which contained only those nsSNVs detected in the human cell line of study (Jurkat cells). There were many clear advantages to using a customized database, including its smaller size (reducing the incidence of false positive peptide IDs), inclusion of nsSNVs not yet in public SNV repositories, and the ability to compare RNA and protein nsSNV expression. The aggregate database may be an option in the case that NGS data cannot be collected, but we found that the large database size (over 100 times larger than the customized database) caused the identification of many false positive SAP peptides, a problem not remedied by application of stringent MS search cut-offs (e.g. local FDR). In light of these findings, it is recommended to use some strategy for condensing or customizing proteomic databases when searching for novel protein variations.

An issue that will become important as methodology for the detection of sample-specific SAP peptides is adopted is that the various genetic, transcriptomic, and proteomic databases have discrepancies in sequence. These sequence discrepancies make it difficult to assess the incidence and extent of protein variations in samples. The genomics community has solved this problem by calling an SNV when there is a nucleotide that is different from the human reference genome that is maintained by the Genome Reference Consortium [56]. No such convention has yet been implemented in the area of proteomics. For example, many proteomics researchers use protein databases containing sequences that are not derived from the human reference genome, such as UniProt, so the set of SAPs called from the reference genome will be different from those called from UniProt.

As outlined in the introduction, it would be beneficial if MS-based proteomics could detect and quantify all the translated nsSNVs in a human sample. In this study, we show that up to ~10% of nsSNVs identified in RNA were also detected at the protein level, meaning that there are many SAP peptides that are not presently

detected. Two factors could improve SAP detection: higher proteomic coverage and increased MS sampling sensitivity. With high proteomic coverage, there is a better chance of detecting a peptide corresponding to an nsSNV. In this study, we used multiple proteases and increased the number of detected SNV sites by ~25%. With increasing sensitivity, there are improved chances of detecting SAPs that are expressed at lower levels. Whereas MS instrument sensitivity is an inherent feature of each MS platform, another factor affecting sensitivity that we can control is the sampling depth, that is, the ability for the instrument to choose precursor peptides of low ion intensity (within a complex matrix) for subsequent MS^2 fragmentation. For example, one solution to increase the number of SAP peptides detected would be to employ a targeted approach by using selected reaction monitoring (SRM) assays [57], SAP peptide inclusion lists during data dependent acquisition, or even intelligent data acquisition (IDA) strategies [58]. These SAP peptide targeting approaches could be employed in future work to detect a larger fraction of translated nsSNV sites.

3.6 Acknowledgements

We thank Gergana Hinrichs and William Horvat for technical assistance with the cell culture and proteomics sample preparation. This work was supported by NIH grants 1P01GM081629 and 1P50HG004952. This research was also supported in part by National Science Foundation Grant CHE-0840494 through use of the University of Wisconsin-Madison chemistry computing resources. RNA-Sequencing work was performed at the University of Wisconsin-Madison Biotechnology Center. GMS was supported by the NIH Genomic Sciences Training Program 5T32HG002760.

References

[1] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491.7422 (2012), pp. 56–65.

[2] Mark I. McCarthy et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 9.5 (2008), pp. 356–369.

[3] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447.7145 (2007), pp. 661–78.

[4] T. A. Manolio et al. Finding the missing heritability of complex diseases. *Nature* 461.7265 (2009), pp. 747–753.

[5] D. G. MacArthur et al. A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science* 335.6070 (2012), pp. 823–828.

[6] S. R. Sunyaev. Inferring causality and functional significance of human coding DNA variants. *Human Molecular Genetics* 21 (2012), R10–R17.

[7] I. A. Adzhubei et al. A method and server for predicting damaging missense mutations. *Nature Methods* 7.4 (2010), pp. 248–249.

[8] Pauline C. Ng and Steven Henikoff. Predicting Deleterious Amino Acid Substitutions. *Genome Research* 11.5 (2001), pp. 863–874.

[9] E. Khurana et al. Interpretation of Genomic Variants Using a Unified Biological Network Approach. *Plos Computational Biology* 9.3 (2013).

[10] Lloyd M. Smith and Neil L. Kelleher. Proteoform: a single term describing protein complexity. *Nat Meth* 10.3 (2013), pp. 186–187.

[11] Martin Beck et al. The quantitative proteome of a human cell line. *Mol Syst Biol* 7 (2011).

[12] E. Lundberg et al. Defining the transcriptome and proteome in three functionally different human cell lines. *Molecular Systems Biology* 6 (2010).

[13] Nagarjuna Nagaraj et al. Deep proteome and transcriptome mapping of a human cancer cell line. *Mol Syst Biol* 7 (2011).

[14] A. Bairoch et al. The universal protein resource (UniProt). *Nucleic Acids Research* 33 (2005), pp. D154–D159.

[15] D. A. Benson et al. GenBank. *Nucleic Acids Research* 40.D1 (2012), pp. D48–D53.

[16] T. Hubbard et al. The Ensembl genome database project. *Nucleic Acids Research* 30.1 (2002), pp. 38–41.

[17] J. Harrow et al. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research* 22.9 (2012), pp. 1760–1774.

[18] D. M. Creasy and J. S. Cottrell. Error tolerant searching of uninterpreted tandem mass spectrometry data. *Proteomics* 2.10 (2002), pp. 1426–1434.

[19] D. Hyatt and C. L. Pan. Exhaustive database searching for amino acid mutations in proteomes. *Bioinformatics* 28.14 (2012), pp. 1895–1901.

[20] C. L. Gatlin et al. Automated identification of amino acid sequence variations in proteins by HPLC/microspray tandem mass spectrometry. *Analytical Chemistry* 72.4 (2000), pp. 757–763.

[21] M. K. Bunger et al. Detection and validation of non-synonymous coding SNPs from orthogonal analysis of shotgun proteomics data. *Journal of Proteome Research* 6.6 (2007), pp. 2331–2340.

[22] Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* (2011).

[23] S. Schandorff et al. A mass spectrometry-friendly database for cSNP identification. *Nature Methods* 4.6 (2007), pp. 465–466.

[24] M. Chen et al. Annotation of Non-Synonymous Single Polymorphisms in Human Liver Proteome by Mass Spectrometry. *Protein and Peptide Letters* 17.3 (2010), pp. 277–286.

[25] A. L. Chernobrovkin et al. Identification of Single Amino Acid Polymorphisms in MS/MS Spectra of Peptides. *Doklady Biochemistry and Biophysics* 437.1 (2011), pp. 90–93.

[26] G. Alves, A. Y. Ogurtsov, and Y. K. Yu. RAId_DbS: mass-spectrometry based peptide identification web server with knowledge integration. *Bmc Genomics* 9 (2008).

[27] S. Mathivanan et al. Identifying mutated proteins secreted by colon cancer cell lines using mass spectrometry. *Journal of Proteomics* 76 (2012), pp. 141–149.

[28] H. Xi et al. SysPIMP: the web-based systematical platform for identifying human disease-related mutated sequences from mass spectrometry. *Nucleic Acids Research* 37 (2009), pp. D913–D920.

[29] H. Nijveen et al. HSPVdb-the Human Short Peptide Variation Database for improved mass spectrometry-based detection of polymorphic HLA-ligands. *Immunogenetics* 63.3 (2011), pp. 143–153.

[30] T. Kawabata, M. Ota, and K. Nishikawa. The protein mutant database. *Nucleic Acids Research* 27.1 (1999), pp. 355–357.

[31] S. A. Forbes et al. "The Catalogue of Somatic Mutations in Cancer (COSMIC)". *Current Protocols in Human Genetics*. John Wiley & Sons, Inc., 2001.

[32] J. Li, D. T. Duncan, and B. Zhang. CanProVar: A Human Cancer Proteome Variation Database. *Human Mutation* 31.3 (2010), pp. 219–228.

[33] S. T. Sherry et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* 29.1 (2001), pp. 308–311.

[34] Yum L. Yip et al. Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. *Human Mutation* 29.3 (2008), pp. 361–366.

[35] J. Li et al. A Bioinformatics Workflow for Variant Peptide Detection in Shotgun Proteomics. *Molecular & Cellular Proteomics* 10.5 (2011).

[36] Vanessa C. Evans et al. De novo derivation of proteomes from transcriptomes for transcript and protein identification. *Nature Methods* 9.12 (2012), pp. 1207–11.

[37] Xiaojing Wang et al. Protein Identification Using Customized Protein Sequence Databases Derived from RNA-Seq Data. *Journal of Proteome Research* (2011).

[38] Gloria M. Sheynkman et al. Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. *Molecular & Cellular Proteomics* (2013).

[39] J. R. Wisniewski et al. Universal sample preparation method for proteome analysis. *Nature Methods* 6.5 (2009), 359–U60.

[40] B. Langmead et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 10.3 (2009).

[41] C. Trapnell, L. Pachter, and S. L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25.9 (2009), pp. 1105–1111.

[42] K. D. Pruitt, T. Tatusova, and D. R. Maglott. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* 35 (2007), pp. D61–D65.

[43] William McLaren et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 26.16 (2010), pp. 2069–2070.

[44] B. Schilling et al. Platform-independent and Label-free Quantitation of Proteomic Data Using MS1 Extracted Ion Chromatograms in Skyline APPLICATION TO PROTEIN ACETYLATION AND PHOSPHORYLATION. *Molecular & Cellular Proteomics* 11.5 (2012), pp. 202–214.

[45] N. Castellana and V. Bafna. Proteogenomics to discover the full coding content of genomes: A computational perspective. *Journal of Proteomics* 73.11 (2010), pp. 2124–2135.

[46] Lukas Käll et al. Posterior Error Probabilities and False Discovery Rates: Two Sides of the Same Coin. *Journal of Proteome Research* 7.1 (2007), pp. 40–44.

[47] Pratik Jagtap et al. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *PROTEOMICS* 13.8 (2013), pp. 1352–1357.

[48] H. Choi, D. Ghosh, and A. I. Nesvizhskii. Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. *Journal of Proteome Research* 7.1 (2008), pp. 286–292.

[49] H. Yan et al. Allelic variation in human gene expression. *Science* 297.5584 (2002), pp. 1143–1143.

[50] Q. Wang et al. Mutant proteins as cancer-specific biomarkers. *Proceedings of the National Academy of Sciences of the United States of America* 108.6 (2011), pp. 2444–2449.

[51] Z. Khan et al. Quantitative measurement of allele-specific protein expression in a diploid yeast hybrid by LC-MS. *Molecular Systems Biology* 8 (2012).

[52] O. V. Krokhin. Sequence-specific retention calculator. Algorithm for peptide retention prediction in ion-pair RP-HPLC: Application to 300-and 100-angstrom pore size C18 sorbents. *Analytical Chemistry* 78.22 (2006), pp. 7785–7795.

[53] V. A. Fusaro et al. Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nature Biotechnology* 27.2 (2009), pp. 190–198.

[54] Zhi-Duan Su et al. Quantitative detection of single amino acid polymorphisms by targeted proteomics. *Journal of Molecular Cell Biology* 3.5 (2011), pp. 309–315.

[55] Isabel Ruppen-Cañás et al. An improved quantitative mass spectrometry analysis of tumor specific mutant proteins at high sensitivity. *PROTEOMICS* 12.9 (2012), pp. 1319–1327.

[56] R. Nielsen et al. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics* 12.6 (2011), pp. 443–451.

[57] P. Picotti and R. Aebersold. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nature Methods* 9.6 (2012), pp. 555–566.

[58] D. J. Bailey et al. Instant spectral assignment for advanced decision tree-driven mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America* 109.22 (2012), pp. 8411–8416.

4 USING GALAXY-P TO LEVERAGE RNA-SEQ FOR THE DISCOVERY OF NOVEL PROTEIN VARIATIONS

This chapter has been submitted for publication.

Sheynkman, G., Johnson, J. E., Jagtap, P. D., Shortreed, M. R., Onsongo, G., Frey, B. L., Griffin, T. J., and Smith, L. M. (2014) Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. *In Review*

4.1 Abstract

Current practice in mass spectrometry (MS)-based proteomics is to identify peptides by comparison of experimental mass spectra with theoretical mass spectra derived from a reference protein database; however, this strategy necessarily fails to detect peptide and protein sequences that are absent from the database. We have recently shown that customized proteomic databases derived from RNA-Seq data can be employed for MS-searching to both improve MS analysis and identify novel peptides. While this general strategy constitutes a significant advance for the discovery of novel protein variations, it has not been readily transferable to other laboratories due to the need for many specialized software tools. To address this problem, we have implemented readily accessible, modifiable, and extensible workflows within Galaxy-P, short for Galaxy for Proteomics, a web-based bioinformatic extension of the Galaxy framework for the analysis of multi-omics (e.g genomics, transcriptomics, proteomics) data. These workflows allow the user to upload raw RNA sequencing reads and convert the data into high-quality customized proteomic databases suitable for MS searching. We show the utility of these workflows on human and mouse samples, identifying 544 peptides containing single amino acid polymorphisms (SAPs) and 187 peptides corresponding to unannotated splice

junction peptides, increasing the number of overall peptide identifications through database reduction, and correlating protein and transcript expression levels.

4.2 Introduction

Mass spectrometry-based proteomics is widely employed to characterize proteins in myriad organisms, ranging from E. coli to human. Fundamental to almost all proteomics analyses is the database search step, where experimental peptide mass spectra are matched with theoretical peptide mass spectra derived from a protein reference database [1]. This MS database searching strategy relies on the completeness and quality of the protein reference database, meaning that peptides and proteins are only identified if their correct sequence is present in the protein reference file. But individual organisms often possess genetic variations that differ from the canonical sequences present in the database. These variations are often not represented in the reference database causing the corresponding peptides to be invisible to MS-based analyses.

In recent years, high-throughput RNA sequencing has been used to empirically determine the transcript sequences expressed in a given sample, strain, cell line, or tissue, and has become accessible to many researchers [2, 3]. Taking advantage of this powerful new capability, we and others have developed novel strategies to leverage RNA-Seq for the detection of sample-specific protein variations [4–11]. In this strategy parallel RNA-Seq and proteomics data are collected from the same or related samples. Novel sequences discovered from RNA-Seq data are translated into proteins and added to the MS search database, which can then be employed to detect the corresponding protein variations.

RNA-Seq derived databases tailored for a given sample can improve proteomics in two main ways. First, and most importantly, RNA-Seq can be used to reveal novel single nucleotide polymorphisms (SNPs), indels, alternative splice forms, and gene fusions at the transcript level that, when translated, yield protein sequences that are not in the reference protein database. These novel protein sequences are then appended to the reference database and employed for MS-searching, enabling

the detection of novel peptides. Second, RNA-Seq can be used to increase the sensitivity and quality of peptide detection. To accomplish this, RNA-Seq is used to quantify transcript levels, and all protein entries in the database that fall below a threshold expression level for the corresponding transcript are removed. This is useful because, in general, proteins that are not associated with reasonable transcript abundances are unlikely to be expressed at detectable levels. This strategy results in fewer "false" sequences in the database, which increases the number and quality of overall peptide identifications [10, 12, 13].

The greatest bottleneck in harnessing RNA-Seq data for the discovery of protein variations is not data generation —deep coverage RNA-Seq data is readily and inexpensively produced —but rather in creating accessible and flexible bioinformatic pipelines to process the data. Given that sequencing platforms and software tools are rapidly evolving, researchers need an environment where it is easy to quickly integrate new transcriptomic and proteomic tools and readily modify workflows to suit their system of study. There is a dire need for transparency and sharing of workflows so that other labs can build upon prior work. These problems are magnified when considering the troves of next generation sequencing (NGS) data that are currently underutilized in the field of proteomics.

Here we address the bioinformatic bottleneck in RNA-Seq-based protein database construction by introducing flexible, extensible, and sharable workflows within usegalaxyp.org, the public version of Galaxy-P. Galaxy-P is an extension of the original web-based Galaxy framework [14–16], with a focus on proteomic and multi- omic data analysis applications. We present three workflows that can be used for RNA-Seq-derived proteomic database construction. These workflows are transparent, easily shared, and flexible, so researchers, especially those without expertise in computer science and bioinformatics, can quickly extend and evolve the workflows for their needs. We describe the workflows and show their utility in discovering novel peptides in both human (Jurkat cells) and mouse (pancreatic islet) samples. The implementation of these workflows in Galaxy-P will help researchers utilize NGS data for the detection and discovery of protein variations via mass spectrometry.

4.3 Experimental Procedures

Jurkat cell RNA-Seq

Jurkat cells were grown in 90% RPMI and 10% FBS (ATCC, Manassas, VA) to 1.3×10^6 cells/mL. Total RNA was extracted using the TriZol and its protocol (Invitrogen). RNA libraries were prepared using the Illumina TruSeq protocol, which included a dT bead enrichment of polyadenylated mRNAs and size selection of 350bp cDNA fragments. ~80 million paired end reads (350bp, 2×100bp) were sequenced on an Illumina HiSeq2000. More information about this dataset may be found in [5]. Jurkat cell MS-based proteomics MS-based proteomics data collection has been previously described [5]. Briefly, protein was extracted and digested using the FASP protocol [17]. Peptides were fractionated on a high-pH HPLC and 28 fractions were analyzed on a nanoflow HPLC integrated with a Velos-Orbitrap mass spectrometer. The MS raw files for the Jurkat cell lysate samples are available via FTP from the PeptideAtlas data repository [18] by accessing the following link: http://www.peptideatlas.org/PASS/PASS00215.

B6 and Cast Mouse Islet RNA-Seq

Pancreatic islets were isolated from two B6 mice and two CAST mice. Total RNA was extracted from ~250 islets of each mouse strain using the Qiagen RNeasy Mini Kit (Qiagen, Hilden, Germany). RNA-Seq data was collected as described for the human sample.

B6 and Cast Mouse Islet proteomics

Protein was extracted from ~400 B6 islets (~470 CAST islets), and then proteomics data was collected in the same manner as for the human sample, except that 9 fractions were collected during the high-pH HPLC fractionation. MS raw files for the mouse samples are available via FTP from the PeptideAtlas data repository [18] by accessing the following link http://www.peptideatlas.org/PASS/PASS00470.

Workflows

Three workflows were created within Galaxy-P that allow for the conversion of RNA-Seq data into customized protein databases. Full details of these workflows can be found in the following links:

SAP database workflows

Link to HTML: Human_SAP_DB_Workflow.html
Link to HTML: Mouse_SAP_DB_Workflow.html
URL to workflow within Galaxy Toolshed:
http://toolshed.g2.bx.psu.edu/view/galaxyp/proteomics_rnaseq_sap_db_workflow

Splice database workflows

Link to HTML: Human_Splice_DB_Workflow.html
Link to HTML: Mouse_Splice_DB_Workflow.html
URL to workflow within the Galaxy Toolshed:
http://toolshed.g2.bx.psu.edu/view/galaxyp/proteomics_rnaseq_splice_db_workflow

Reduced database workflows

Link to HTML: Human_Reduced_DB_Workflow.html
Link to HTML: Mouse_Reduced_DB_Workflow.html
URL to workflow within the Galaxy Toolshed:
http://toolshed.g2.bx.psu.edu/view/galaxyp/proteomics_rnaseq_reduced_db_-workflow

Database searching of mass spectrometry data

For each of the three sample types described above (human, mouse B6, mouse CAST), Galaxy-P workflows generated a SAP, splice, and reduced database which

was concatenated with the cRAP database of common MS contaminants. The resultant reduced+SAP+splice+cRAP databases, one created for each of the three samples, were searched against the matched raw mass spectra data using the Percolator search node within Proteome Discoverer (v1.4, Thermo Fisher Scientific, San Jose, CA). Default peaklist-generating parameters were used. Precursor m/z tolerance was set to 10 ppm and product m/z tolerance was set to 0.05 Da. Peptides with up to two missed cleavages (trypsin) were permitted. Variable methionine oxidation and static carbamidomethylation were used. Using reversed sequences as a decoy database, peptides passing a 1% global FDR were accepted as identified (except in cases where a more stringent 1% local FDR was mentioned in the text).

Post-search peptide filtering and annotation

Peptide identifications were filtered using the "Filter In Reference" tool we developed within Galaxy-P, which finds and annotates the novel peptides not listed in the reference proteome. An example workflow may be found in the following links: Example_Novel_Peptide_Filter.html URL to workflow within the Galaxy Toolshed: http://toolshed.g2.bx.psu.edu/view/galaxyp/proteomics_novel_peptide_filter_workflow

4.4 Results

Galaxy workflows

We have developed workflows in Galaxy-P that convert RNA-Seq data into three types of readily usable proteomic databases. These are databases containing novel single amino acid polymorphisms; databases containing novel splice junction sequences; and a reduced database, which only contains protein sequences with corresponding transcripts that are expressed over a threshold level of abundance.

We demonstrated the utility of these workflows on parallel RNA-Seq and proteomics datasets collected from the same sample. Figure 1 shows an overview of

the experimental design employed to collect RNA-Seq and proteomic data from human Jurkat cells and mouse pancreatic islets from B6 and CAST mice. From each sample, paired-end RNA-Seq reads (350bp, 2×100bp) from polyadenylated mRNAs were sequenced on an Illumina HiSeq2000 and tandem mass spectra of tryptically digested peptides were collected on a Velos-Orbitrap mass spectrometer. Figure 2 gives an overview of the three bioinformatic workflows, which are described below. These workflows can serve as the starting point for more complex bioinformatic pipelines and are designed to be readily edited, extended, and evolved.

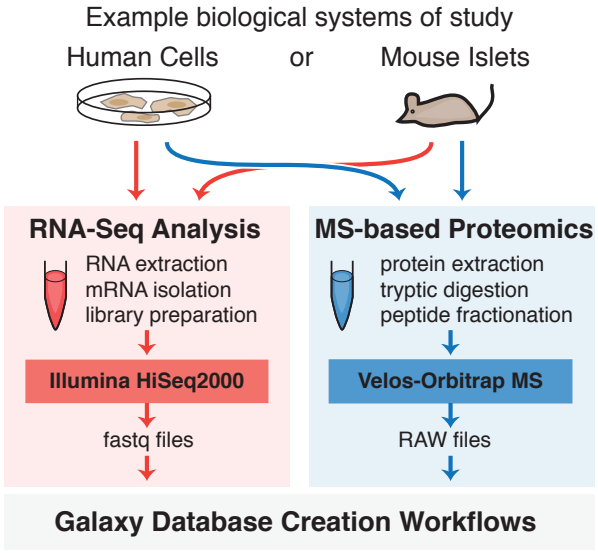

Figure 4.1. Experimental overview. The Galaxy-P workflows take as input sample-specific RNA-Seq data and create sample-specific protein databases. These protein databases are then employed for MS-based proteomics database searching. The workflows were developed on datasets generated from human (Jurkat cells) and mouse (B6 and CAST islets) samples.

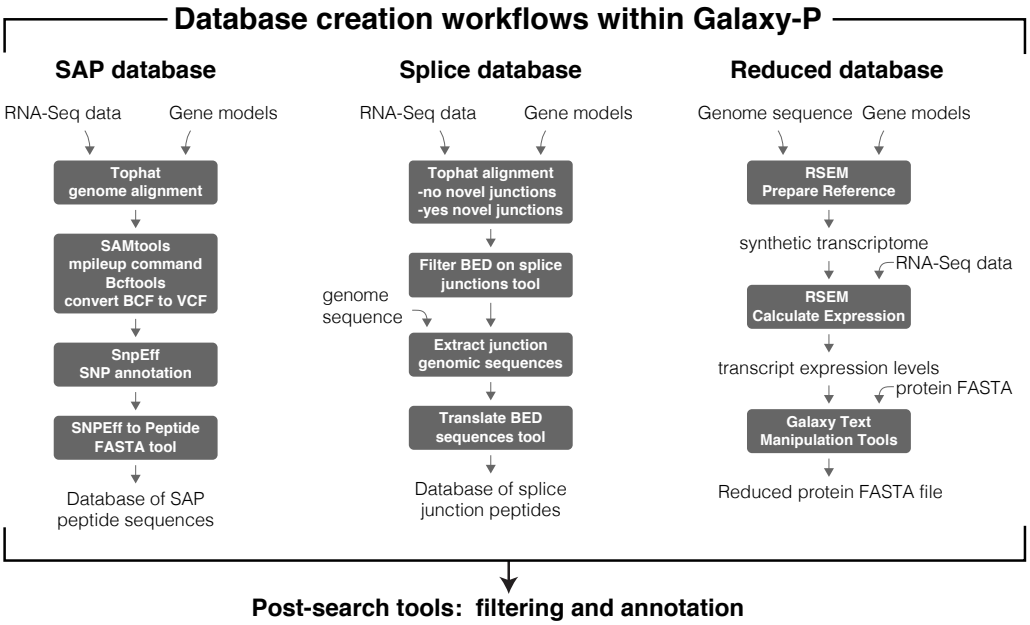

Figure 4.2. Overview of workflows implemented in Galaxy-P that utilize RNA-Seq data for improved proteomics. The single amino acid polymorphism (SAP) database workflow detects non-synonymous SNPs that yield SAPs. The splice database workflow detects alternatively spliced transcripts and the corresponding novel splice junction polypeptide sequences. The reduced database workflow quantifies the sample's transcriptome, optionally removes likely unexpressed protein sequences, and allows determination of RNA-protein correlations. Post-search tools are used to filter and annotate novel peptides.

Galaxy workflows for RNA-Seq-derived database creation

SAP database

SNPs are single nucleotide differences between genomes of different individuals and are one of the most common types of genetic variation [19]. SNPs that reside within a protein coding region and change the coding amino acid are termed non-synonymous SNPs (nsSNPs) and the corresponding amino acid is then called a single amino acid polymorphism (SAP). Since a change in protein coding sequence can potentially alter a protein's function, it is important to directly measure SAP-

containing proteins by mass spectrometry. This would allow the evaluation of the post-translational consequences of a given variant. In addition, detection of proteins from heterozygous genes would allow for measurement of differential allelic expression at the protein level to complement measurement of differential allelic expression at the RNA level, which is already possible [20].

Most reference protein databases contain only those amino acid sequences that are translated from the reference genome, which typically represent nucleotide sequences derived from one or more representative individuals or strains [21]. Therefore, SAPs present in a particular experimental sample will be missed unless they are explicitly added to the database. To solve this problem, we and others have shown that customized SAP polypeptide databases can be constructed from RNA-Seq data. The set of nsSNPs encoded in a sample's transcriptome can be detected by RNA-Seq and the stretches of RNA sequences containing nsSNPs can be translated into SAP-containing protein sequences for database searching [4, 10].

The SAP database workflow in Galaxy-P inputs raw RNA-Seq data and outputs a database of SAP polypeptide entries that can be used for MS searching. The workflow aligns RNA-Seq reads to the reference genome using Tophat [22], calls SNPs using SAMtools [23], and annotates the SNPs that reside within protein-coding regions using SNPeff [24]. To convert the annotated SNPs into a SAP-containing polypeptide database, the workflow uses a tool we developed within Galaxy-P called "SNPeff to Peptide Fasta". Within this tool, the user specifies the number of amino acids to the left and right of each detected SAP to include in the final SAP database. Each entry in the database contains an informative header specifying the location of both the SNP and SAP on the transcript and protein, respectively. Additionally, if the user would like to employ an alternative SNP calling tool, like GATK, they can modify the workflow to include it [25].

We used the Galaxy-P SAP database workflow to create and employ custom SAP databases for the human and mouse samples. Using the human RNA-Seq dataset, this workflow produced a SAP database comprising 6,168 SAP polypeptide entries, which was combined with the Ensembl reference proteome. After MS database searching, 522 SAP peptides that mapped up to 491 unique SNP sites on

the genome were identified. These SAP peptides would not have been detected if only the canonical Ensembl protein sequences were used for database searching. When comparing the SAP peptides detected in the present study (522) with SAP peptides detected using our previously published SAP workflow (491) [4], which used different gene models (RefSeq instead of Ensembl), there was an 89% overlap in peptide identifications.

We also demonstrated the utility of this SAP database workflow on the two mouse strains, B6 and CAST. For B6, the workflow produced, as expected, only 1 SAP entry, a likely false positive or recent mutation since the mouse reference genome is based on B6 [26]. For CAST, however, the workflow output a database with 476 SAPs, which was concatenated with the Ensembl reference proteome and subsequently used for MS searching. 22 SAP peptides mapping to 19 unique SNP sites were identified. The difference between B6 and CAST SAP databases illustrates that the number of SAPs detected is dependent on the relationship between the sample and the reference genome. B6, which is in fact the strain from which the reference genome is based, did not have detected variants while CAST, a less well characterized disease model system for Type II diabetes, had many. This illustrates the importance of utilizing RNA-Seq data for proteomics analysis, especially for organisms, strains, and disease models that have not been thoroughly characterized or contain sparsely annotated reference proteomes.

Results for both human and mouse data are summarized in Table 1.

	SAP database		Proteomic Identifications	
Sample	SAPs	SNP sites	SAP Peptide IDs*	SNPs ID'd
Jurkat human cells	9,168	6,924	522	491
B6 mouse islets	1	1	N/A	N/A
CAST mouse islets	476	249	22	19

*peptide passing a 1% FDR

Table 4.1. Results from creating SAP databases and using them for searching proteomic datasets.

Splice database

A majority of genes in higher eukaryotes are alternatively spliced resulting in the production of multiple mRNA forms from the same gene. The spliceosome processes pre-mRNAs by excising introns and combining specific exons to produce a mature RNA. The ubiquity of splicing, especially in humans, has been revealed by next generation sequencing methods that allow unbiased, global characterization of splicing in many cell and tissue types [27, 28].

Despite the high number of novel splice forms detected at the transcript level, proteomic databases for MS searching are far from complete in terms of splicing. There are still novel splice events in certain cell types or disease models that are not yet annotated. Consequently, the polypeptide sequences corresponding to these novel splice sites are not in the protein reference database and are thus missed during standard MS-based proteomic analyses.

Within Galaxy-P, we have created a workflow for the detection and subsequent incorporation of novel splice sequences into custom splice-junction databases. The splice database workflow first aligns RNA-Seq data to the genome twice, first to only those splice junctions found in the Ensembl gene models and second to both the Ensembl gene models and reference genome. The output BED files, which contain the coordinates of all detected junctions, are compared to each other and only those coordinates corresponding to splice junctions not present in the gene models are retrieved. Next, the genomic sequences for each splice junction is retrieved. We developed a program within Galaxy-P, "Translate BED sequences", which translates the splice junctions and compiles all splice-junction polypeptide sequences into a database. The user may choose to filter out splice junction entries that contain stop codons, are less than a certain length, or are below a certain expression level as measured by the RNA-Seq read depth at each splice junction.

We used the splice database workflow to create and employ custom splice-junction databases for the human and mouse samples. Using the human RNA-Seq dataset, this workflow produced a splice-junction database comprising approximately 33,000 splice-junction polypeptide entries. Previously, we have found it

was important to use a stringent score cut-off for peptide spectral matches corresponding to splice junction peptides [5]. Therefore, we required the same 1% local FDR for splice-junction peptide identifications in the present study. After MS searching against the splice-junction database, 67 novel splice junction peptides, defined as those peptides not present in the Ensembl reference proteome, were identified. There was a 57% overlap of splice-junction peptides identified in this and a previous study, which used a similar though not identical workflow (e.g. RefSeq gene models) [5].

Application of the workflow for analysis of the mouse islet RNA-Seq data resulted in a splice junction database containing approximately 32,000 (B6) and 20,000 (CAST) splice junction polypeptides. After MS searching, 58 (B6) and 72 (CAST) novel splice junction peptides were identified at a 1% local FDR.

Results for human and mouse data are summarized in Table 2. These results show that many sample-specific peptides derived from novel alternative splice events are missed when using only the reference protein database for MS searching.

	splice database		
Sample	size	min. depth	peptide IDs*
Jurkat human cells	33,372	6	67
B6 mouse islets	57,587	4	64
CAST mouse islets	43,244	4	66

*peptide passing a 1% local FDR

Table 4.2. Results from creating splice junction databases and using them for searching proteomic datasets.

Reduced database

Target decoy search strategies are widely used in mass spectrometry-based proteomics to permit a determination of the false discovery rate (FDR) for peptide identifications [29]. The underlying assumption in this approach is that the target database, which comprises the sequences of the protein reference database, reflects the protein sequences actually present in the sample. However, this is rarely the

case; for example, human cells have been found to express fewer than 50% of the proteins encoded in their genome at any given time [30–32]. RNA-Seq data can be employed to quantify transcripts and then remove those protein sequences from the reference database that have minimal or undetected mRNA expression levels [33]. This produces a smaller, sample-specific "reduced" database that improves the number and quality of peptide identifications [10, 12, 13].

In the reduced database Galaxy-P workflow, the sample-matched raw RNA-Seq data serves as input and RSEM [34] is used to quantify transcripts based on Ensembl gene models (e.g. GTF file), and the output is a list of each transcripts' abundance in Transcripts Per Million (TPM). Next, Galaxy Text Manipulation tools are used to link each protein entry in the protein FASTA file to its corresponding transcript and the transcript's abundance in TPM. The user selects the minimum transcriptional abundance a protein must have to be included in the reduced database (e.g. >1TPM).

We used the human and mouse datasets to test the reduced database workflow by creating reduced databases comprised of only those proteins with transcript abundances above 1 TPM. For human, the Ensembl protein database was reduced from approximately 104,000 to 83,000 entries. The MS search against this reduced database yielded 313 more peptide identifications as compared to the original database search. For mouse, the Ensembl protein database was reduced from approximately 52,000 to 18,000 (B6) or 17,000 (CAST) entries, increasing the number of peptide identifications for each strain by 166 (B6) and 146 (CAST). Though these increases in peptide identifications are modest, another benefit of reduced databases is that the overall quality of peptide identifications improves, as shown in Figure 3. Full results for the reduced databases are listed in Table 3.

The reduced database workflow is easily modified to accommodate different datasets and can also enable measurement of RNA-Protein expression correlations. For example, one can easily change the TPM cut-off employed for various proteomic datasets that have different depths of coverage. If available, alternative gene models besides Ensembl can be used, as can different transcript quantification programs. Since the reduced databases contain TPM values for each protein, the user can easily determine RNA-Protein expression correlations such as those shown in Figure 4

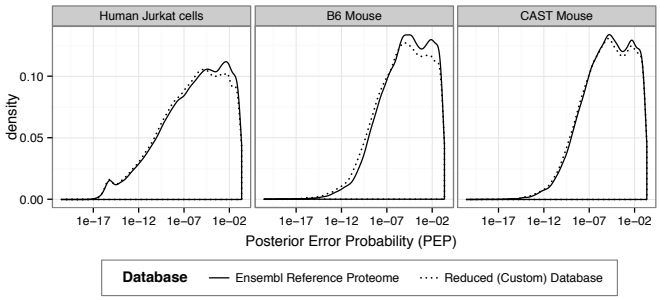

Figure 4.3. Improved peptide confidence scores from reduced database searches. The comparison of distributions of posterior error probability (PEP) values for peptides passing a 1% FDR cut-off for MS searches of the reduced database versus the (canonical) Ensembl reference proteome. A lower PEP value reflects a higher degree of confidence for the peptide identifications. In all cases, the reduced database searches improves PEP scores, as shown by the shift of PEP distributions to the left.

	Rna-Seq	Mass	original database		reduced database		
Sample	reads	spectra	# entries	peptide IDs*	# entries	peptide IDs*	% increase
Jurkat human cells	80M	500K	104,310	73,123	82,101	73,436	0.4
B6 mouse islets	94M	250K	52,165	30,212	18,052	30,220	0.3
CAST mouse islets	126M	250K	52,165	28,756	16,940	28,823	0.2

*peptide passing a 1% FDR

Table 4.3. Results from MS searching with the original Ensembl protein database and the reduced database.

for the human and B6 mouse samples in the present study.

4.5 Conclusions

Using RNA-Seq data to enhance MS analysis is a promising strategy to discover novel peptides specific to a sample and, more generally, to improve proteomics results. The main bottleneck for widespread adoption of this strategy has been the lack of easily used and modifiable computational tools. We provide a solution to this problem by introducing a set of workflows within Galaxy-P that easily convert raw RNA-Seq data into proteomic databases. Development within Galaxy-P brings

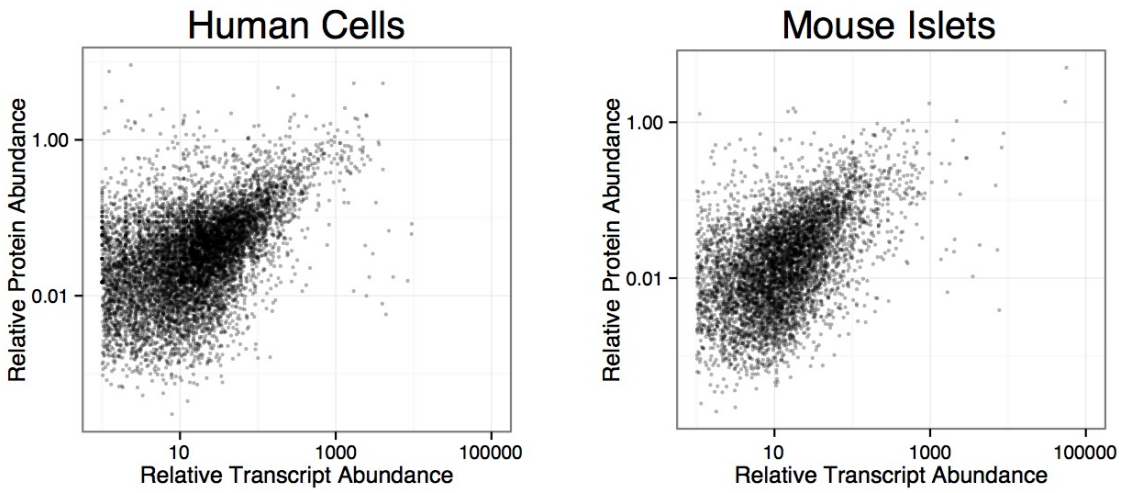

Figure 4.4. Transcript versus protein abundance expression. The reduced database workflow outputs for each protein the transcript abundance level in transcripts per million (TPM). This allows researchers to easily determine RNA-Protein correlations. For this particular plot, individual protein abundances were estimated by counting the number of peptide spectral matches and normalizing by the protein's length (e.g. 400 amino acids).

unique benefits due to the inherent characteristics of the Galaxy-framework [14–16], such as easy publication and sharing of complete workflows with other users. Flexibility is a key benefit, as users can easily customize workflows to account for sample- or experiment-specific parameters, and also incorporate emerging new tools as desired. Although the complete workflows are available for use on the public Galaxy-P instance (i.e. implementation), the tools used and developed here are either already a part of the main Galaxy build or have been deposited in the Galaxy Tool Shed (http://toolshed.g2.bx.psu.edu/) under the "Proteomics" link. Thus these workflows should be usable on local Galaxy instances as well.

These workflows were tested on RNA and protein datasets that were collected in parallel from human and mouse samples. The results show that incorporating RNA-Seq data into proteomic analyses enables discovery of novel peptides arising from genetic variation and alternative splice forms, improves the number and quality of peptide identifications, and enables measurement of RNA-Protein expression

correlations. These workflows and the benefits of the Galaxy framework provide a sound basis upon which to build newer and more sophisticated methods of RNA-Seq analysis for the continued advancement of proteomics, as newer tools and technologies arise.

4.6 Acknowledgements

This work was supported by NIH grants 1P01GM081629, U54DK093467, 1P50HG004952 to LMS and NSF grant 1147079 to TJG. RNA-Sequencing work was performed at the University of Wisconsin-Madison Biotechnology Center. We thank Donnie Stapleton, Mark Keller, and Alan Attie for supplying the mouse islet samples (National Institute of Diabetes and Digestive Kidney Diseases grants 58037 and 66369). Galaxy-P is maintained by the Minnesota Supercomputing Center at the University of Minnesota. GMS was supported by the NIH Genomic Sciences Training Program 5T32HG002760.

References

[1] J. K. Eng, A. L. McCormack, and J. R. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 5.11 (1994), pp. 976–89.

[2] Manuel Garber et al. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Meth* 8.6 (2011), pp. 469–477.

[3] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10.1 (2009), pp. 57–63.

[4] G. M. Sheynkman et al. Large-scale mass spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences. *J Proteome Res* 13.1 (2014), pp. 228–40.

[5] G. M. Sheynkman et al. Discovery and Mass Spectrometric Analysis of Novel Splice-junction Peptides Using RNA-Seq. *Molecular & Cellular Proteomics* 12.8 (2013), pp. 2341–2353.

[6] S. Woo et al. Proteogenomic database construction driven from large scale RNA-seq data. *J Proteome Res* 13.1 (2014), pp. 21–8.

[7] G. Lopez-Casado et al. Enabling proteomic studies with RNA-Seq: The proteome of tomato pollen as a test case. *Proteomics* 12.6 (2012), pp. 761–74.

[8] G. Menschaert et al. Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Mol Cell Proteomics* 12.7 (2013), pp. 1780–90.

[9] V. C. Evans et al. De novo derivation of proteomes from transcriptomes for transcript and protein identification. *Nat Methods* 9.12 (2012), pp. 1207–11.

[10] X. Wang et al. Protein identification using customized protein sequence databases derived from RNA-Seq data. *J Proteome Res* 11.2 (2012), pp. 1009–17.

[11] M. Frenkel-Morgenstern et al. Chimeras taking shape: potential functions of proteins encoded by chimeric RNA transcripts. *Genome Res* 22.7 (2012), pp. 1231–42.

[12] P. Jagtap et al. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *Proteomics* 13.8 (2013), pp. 1352–7.

[13] P. Blakeley, I. M. Overton, and S. J. Hubbard. Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *J Proteome Res* 11.11 (2012), pp. 5221–34.

[14] Daniel Blankenberg et al. "Galaxy: A Web-Based Genome Analysis Tool for Experimentalists". *Current Protocols in Molecular Biology*. John Wiley & Sons, Inc., 2010.

[15] B. Giardine et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* 15.10 (2005), pp. 1451–5.

[16] J. Goecks et al. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 11.8 (2010), R86.

[17] J. R. Wisniewski et al. Universal sample preparation method for proteome analysis. *Nat Methods* 6.5 (2009), pp. 359–62.

[18] F. Desiere et al. The PeptideAtlas project. *Nucleic Acids Research* 34 (2006), pp. D655–D658.

[19] D. M. Altshuler et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491.7422 (2012), pp. 56–65.

[20] T. Pastinen. Genome-wide allele-specific analysis: insights into regulatory variation. *Nat Rev Genet* 11.8 (2010), pp. 533–8.

[21] T. Hubbard et al. The Ensembl genome database project. *Nucleic Acids Res* 30.1 (2002), pp. 38–41.

[22] C. Trapnell, L. Pachter, and S. L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25.9 (2009), pp. 1105–11.

[23] H. Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27.21 (2011), pp. 2987–93.

[24] P. Cingolani et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* 6.2 (2012), pp. 80–92.

[25] M. A. DePristo et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43.5 (2011), pp. 491–8.

[26] Consortium Mouse Genome Sequencing et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420.6915 (2002), pp. 520–62.

[27] E. T. Wang et al. Alternative isoform regulation in human tissue transcriptomes. *Nature* 456.7221 (2008), pp. 470–6.

[28] Q. Pan et al. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 40.12 (2008), pp. 1413–5.

[29] J. E. Elias and S. P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* 4.3 (2007), pp. 207–14.

[30] T. Geiger et al. Comparative Proteomic Analysis of Eleven Common Cell Lines Reveals Ubiquitous but Varying Expression of Most Proteins. *Molecular & Cellular Proteomics* 11.3 (2012), pp. M111.014050–M111.014050.

[31] M. Beck et al. The quantitative proteome of a human cell line. *Mol Syst Biol* 7 (2011), p. 549.

[32] N. Nagaraj et al. Deep proteome and transcriptome mapping of a human cancer cell line. *Mol Syst Biol* 7 (2011), p. 548.

[33] A. Mortazavi et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5.7 (2008), pp. 621–8.

[34] B. Li and C. N. Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12 (2011), p. 323.