Title
Proteogenomic method to identify mutated peptides and immunoglobulin rearrangements using NGS data, and it's application to cancer data

Permalink
https://escholarship.org/uc/item/83j7w87r

Author
Woo, Sunghee

Publication Date
2015

Peer reviewed|Thesis/dissertation
UNIVERSITY OF CALIFORNIA, SAN DIEGO

Proteogenomic method to identify mutated peptides and immunoglobulin rearrangements using NGS data, and its application to cancer data

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy

in

Electrical Engineering (Computer Engineering)

by

Sunghee Woo

Committee in charge:

Professor Vineet Bafna, Chair
Professor Nuno Banderia
Professor Chung Kuan Cheng
Professor Clark Guest
Professor William Hodgkiss

2015
The Dissertation of Sunghee Woo is approved and is acceptable in quality and form for publication on microfilm and electronically:

________________________________________

________________________________________

________________________________________

________________________________________

Chair

University of California, San Diego
2015
TABLE OF CONTENTS

Signature Page ........................................................................................................ iii

Table of Contents ................................................................................................... iv

List of Figures .......................................................................................................... vi

List of Tables ........................................................................................................... x

Acknowledgements ................................................................................................. xi

Vita ............................................................................................................................ xii

Abstract of the Dissertation ..................................................................................... xiii

Chapter 1  Proteogenomic database construction driven from large scale RNA-seq data ................................................................. 1
1.1 Introduction ....................................................................................................... 1
1.2 Method .............................................................................................................. 4
  1.2.1 Converting splice graph structure to a FASTA file format ................... 7
  1.2.2 Datasets and experimental procedure .................................................. 11
1.3 Results ............................................................................................................ 13
1.4 Conclusions .................................................................................................... 20

Acknowledgements ................................................................................................ 22

Chapter 2  Proteogenomic strategies for identification of aberrant cancer peptides using large-scale Next Generation Sequencing data .................................................. 23
2.1 Introduction ..................................................................................................... 23
2.2 Method ............................................................................................................ 27
  2.2.1 Database creation from RNA-seq data ................................................. 27
  2.2.2 Database Search Details ...................................................................... 30
  2.2.3 FDR based error control strategies ..................................................... 31
  2.2.4 Sample preparation and LC-MS/MS analysis ..................................... 33
2.3 Results ............................................................................................................ 34
2.4 Discussion ........................................................................................................ 39

Acknowledgements ................................................................................................ 43

Chapter 3  Integrative proteogenomic pipeline for identification of mutated peptides and immunoglobulin gene rearrangements, and its application to colon cancer ................................................................. 50
3.1 Introduction ..................................................................................................... 50
3.2 Results ................................................................. 53
3.3 Discussion ............................................................. 62

Acknowledgements ......................................................... 64

Appendix A Proteogenomic database construction driven from large scale RNA-seq data .................. 70
A.1 Comparison with other gene prediction methods ......................... 71
A.2 Calculation of split mapped coordinates from CIGAR string in SAM file format ........................................... 72
A.3 Detailed RNA-seq methods ............................................... 76
A.4 Comparison of spectra dataset used in this study with Merrihew et al. (2008) [45] ........................... 77
A.5 Proof of correctness and completeness in applying Rule1, Rule2, and Rule3 .............................................. 79
A.6 Proof of correctness and completeness in DFS algorithm implementation of Rule1, Rule2, and Rule3 . 82
A.6.1 Rule 1: ................................................................. 82
A.6.2 Rule 2: ................................................................. 83
A.6.3 Rule 3: ................................................................. 83

Appendix B Proteogenomic strategies for identification of aberrant cancer peptides using large-scale Next Generation Sequencing data .................. 85

Appendix C Integrative proteogenomic pipeline for identification of mutated peptides and immunoglobulin gene rearrangements, and its application to colon cancer ......................................... 89

Bibliography ................................................................. 104
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.1</td>
<td>(a) Given RNA-seq read, find overlapping regions with the existing splice graph. (b) Split and add nodes. (c) Assign edges for each spliced-read. (d) Revisit each pair of contiguous nodes. The nodes are merged if there is no edge at the boundaries.</td>
</tr>
<tr>
<td>Figure 1.2</td>
<td>(a) Depth first search (DFS) traversing of the graph. (b) In traversing, we only maintain a length $L - 1$ suffix. (c) In encountering an incoming visited edge, only maintain a length $L - 1$ prefix. (d) For a pair of sequences, combine two sequences.</td>
</tr>
<tr>
<td>Figure 1.3</td>
<td>(a) Growth of the database file size while incorporating more RNA-seq data. (b) Increase in the percentage of covered splice junctions compared to RefSeq. (c) Increase in the number of splice junctions expressed in splice graph.</td>
</tr>
<tr>
<td>Figure 1.4</td>
<td>(a) Shows a novel gene area where two peptides are identified in a non-genomic region. (b) Two peptides with alternative splice junctions. (c) Peptide matches in a different frame compared to the gene eef-2 with Two neighboring peptides.</td>
</tr>
<tr>
<td>Figure 2.1</td>
<td>Number of peptide identifications in 439,858 spectra collected from a single sample (sample id: TCGA-24-1467) using different FDR based error control strategies.</td>
</tr>
<tr>
<td>Figure 2.2</td>
<td>Overlap between novel identifications from unified and single sample database.</td>
</tr>
<tr>
<td>Figure 2.3</td>
<td>Alignment of identified spectra of mutated peptides.</td>
</tr>
<tr>
<td>Figure 2.4</td>
<td>Alignment of identified spectra of novel junction peptides.</td>
</tr>
<tr>
<td>Figure 2.5</td>
<td>Alignment of identified spectra of mutated peptides.</td>
</tr>
<tr>
<td>Figure 2.6</td>
<td>Insertions and substitutions are represented as additional node and edges having negative coordinate values. Deletions are represented same as splice junctions with actual DNA coordinates.</td>
</tr>
<tr>
<td>Figure 3.1</td>
<td>Illustration of proteogenomic database construction for IG peptide identifications.</td>
</tr>
</tbody>
</table>
Figure 3.2. (a) Comparison of aberrant peptide identifications against previous findings using multi-stage FDR (b) Comparison of overlapping aberrant peptide identifications using combined FDR. .................. 67

Figure 3.3. (a) Genes containing most frequent somatic mutations reported by the TCGA study. (b) Most frequently mutated genes in DNA level are under expressed in protein level. (c) Percentage of samples containing identified protein mutations. ............................ 68

Figure 3.4. Percentage of IG gene peptide identifications in each sample normalized by the number of known peptide identifications across sample subtypes. .................................................. 69

Figure A.1. In filtering stage, RNA-seq reads that have identical splice junctions are merged, and extended in both ends ....................... 71

Figure A.2. Combining pairs of sequences that share a prefix and suffix string. (a) If \( \ell < L \), the generated sequences cannot share an identical prefix and suffix. (b) If \( l \geq 2L \), the prefix and suffix of generated sequences will not overlap .............................................. 72

Figure A.3. Illustration of hashing. (a) In first visited node path, generated sequence is the full path. (b) In traversing, we store all the split nodes present in a candidate list. (c) When a merge node is encountered, we traverse the subsequent path, querying the hash table. ........... 73

Figure A.4. Description of parameter \( W \) .................................................. 74

Figure A.5. Alignment result of novel gene example. The highlighted region corresponds to the alignment of identified peptide against the sequence of hypothetical protein CRE_09558 [Caenorhabditis remanei]. ............................................. 74

Figure A.6. Translated UTR spectral counts throughout different developmental stages .......................................................... 75

Figure A.7. RNA vs peptide transcription level ................................. 75

Figure B.1. Diagram describing different FDR based error control strategies applied in this study. .................................................. 86

Figure B.2. Structure of hash table for accessing the original RNA-seq meta information ...................................................... 87
Figure B.3. UCSC genome browser plot of our novel peptide identifications within complex immunoglobulin region rearrangements included in our peptide identification result. ................................. 87
Figure B.4. UCSC genome browser plot of peptide identification in pseudo gene area. ................................................................. 88
Figure B.5. UCSC genome browser plot of peptide identification in a possible novel gene area where a gene prediction method also reported as a possible gene. .................................................. 88
Figure C.1. (a) Potentially missed reads from a somatically recombined heavy chain transcript as greyed out, while mapping reads as darker. (b) Example de Bruijn graph showing how differences in sequence manifest as differences in topology. ........................................ 90
Figure C.2. Multistage-FDR strategy. .................................................. 90
Figure C.3. (a) Number of known peptide identifications obtained by applying combined FDR versus multi-stage FDR. (b) Number of novel peptide identifications applying combined versus multi-stage FDR. 91
Figure C.4. Comparison between results obtained using MSGF+ and Comet MS/MS search tools. ...................................................... 92
Figure C.5. (a) Example of peptide identifications in immunoglobulin rearrangements. (b) Diagram illustrating the peptide identifications of V(D)J recombination junctions. ................................. 94
Figure C.6. (a) Plot of RNA-seq read counts from IG variable region versus IG constant region. (b) Spectra counts of peptide identifications from IG constant versus variable region. (c) Plot of spectra counts covering IgG constant region. ................................. 95
Figure C.7. Percentage of peptide identifications with somatic mutations in each sample normalized by the number of known peptide identifications across sample subtypes. ........................................ 96
Figure C.8. Identification of somatic mutation in gene SMAD4. This mutation had 1 spectra count with unique genomic location and 15 RNA-seq read depth. This mutation is also reported as somatic mutation in TCGA study [47], and COSMIC [25]. ................................. 97
Figure C.9. Identification of somatic mutation in gene KRAS. TCGA colon cancer study [47] reported this mutation as ‘somatic’ in 25 different colon cancer samples and also reported by COSMIC [25] and dbSNP [58]. ......................................................... 98

Figure C.10. Identification of somatic mutation in gene FGA. 3 overlapping peptide sequences had total 4 spectra counts and unique genomic locations. This SNV location is reported by both COSMIC [25] and dbSNP [58]. ......................................................... 99

Figure C.11. Identification of somatic mutation in gene PIGR. Total spectra count of both peptide was 137 and RNA-seq read depth of this mutation was 11005. Matching mutation of this region were found in both COSMIC [25] and dbSNP [58]. ................................. 100

Figure C.12. Identified alternative splice junction peptide. This peptide had 11 spectra counts (with unique genomic location) and total 386 RNA-seq reads were mapped to this alternative splice junction. . . 101

Figure C.13. Identified deletion and two neighboring SNP mutated peptides. This peptide with deletion had 7 spectra counts (across 6 different tumor protein samples) with unique genomic location and 996 RNA-seq read depth (across 10 different tumor DNA samples). . . 102

Figure C.14. Identified fusion gene peptides. This shows a possible gene fusion region where two junctional peptides are identified across two different genes. Two fusion peptide shown in this region had 15 spectra counts. HBA1 and HBA2 are Hemoglobin related genes. . 103
## LIST OF TABLES

| Table 1.1. | The statistics of novel events identified | 20 |
| Table 2.1. | Statistics of created cancer databases | 44 |
| Table 2.2. | List of novel findings (alternative splice junctions indicate novel junctions that shares identical splice site with a RefSeq gene in one side.) | 44 |
| Table 3.1. | Characterization of aberrant events. 61 samples out of 90 had blood (normal) samples available as a matched reference. | 65 |
| Table A.1. | Overall statistics of splice graph data structure | 71 |
| Table A.2. | Number of overlapping sequences between identified novel peptides using our proteogenomics pipeline versus protein sequences generated from other gene prediction methods. | 72 |
| Table C.1. | Database statistics. Total 90 RNA-seq BAM files which matches with the tumor samples used in the study of Zhang et al. [78] were used in creating proteogenomic database. | 89 |
| Table C.2. | Statistics of identified novel events using combined FDR 1% cut-off. This statistics was generated by applying conventional combined FDR strategy. We obtained large number of novel peptide identifications compared to the result from multi-stage FDR strategy. | 93 |
ACKNOWLEDGEMENTS

I would like to acknowledge Professor “Vineet Bafna” for his guidance and support to this research. I would also like to acknowledge the “Seong Won Cha” who worked on multiple proteogenomics projects together and who shares equal contributions to all research developments written in this thesis.

Chapter 1, in full, is a reprint of the material as it appears in Journal of Proteomic Research. S. Woo., S. Cha, G. Merrihew, Y. He, N. Castellana, C Guest, M. MacCoss, and V. Bafna, 2013. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is a reprint of the material as it appears in Journal of Proteomic Research. S. Woo., S. Cha, S. Na, C. Guest, T. Liu, R. Smith, K. Rodland, S. Payne, and V. Bafna, 2013. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in part is currently being prepared for submission for publication of the material. S. Woo., S. Cha, S. Bonisso, S. Na, D. Tabb, P. Pevzner, V. Bafna. The dissertation author was the primary investigator and author of this material.
VITA

2009 Bachelor of Science, Korea University, Seoul
2009–2015 Research Assistant, University of California, San Diego
2015 Doctor of Philosophy, University of California, San Diego
ABSTRACT OF THE DISSERTATION

Proteogenomic method to identify mutated peptides and immunoglobulin rearrangements using NGS data, and its application to cancer data

by

Sunghee Woo

Doctor of Philosophy in Electrical Engineering (Computer Engineering)

University of California, San Diego, 2015

Professor Vineet Bafna, Chair

Cancer is driven by the acquisition of somatic DNA lesions. Distinguishing the early driver mutations from subsequent passenger mutations is key to molecular sub-typing of cancers, understanding cancer progression, and the discovery of novel biomarkers. The advances of genomics technologies (whole-genome exome, and transcript sequencing, collectively referred to as NGS(Next Generation Sequencing)) have fueled recent studies on somatic mutation discovery. However, the vision is challenged by the complexity, redundancy, and errors in genomic data, and the difficulty of investigating the proteome translated portion of aberrant genes using only genomic approaches.
Combination of proteomic and genomic technologies are increasingly being employed.

This thesis provides a discussion of applying different strategies relating to large database search, and FDR(False Discovery Rate) based error control, and their implication to cancer proteogenomics. Moreover, it extends and develops the idea of a unified genomic variant database that can be searched by any mass spectrometry sample. Furthermore, we introduce a novel database creation method targeted for immunoglobulin peptide search.

Finally, by applying our integrative proteogenomics pipeline, we have identified various types of mutated peptides and immunoglobulin gene rearrangements. Overall statistics and important examples of our proteogenomic discoveries will be shown throughout this study.
Chapter 1

Proteogenomic database construction driven from large scale RNA-seq data

1.1 Introduction

With the advent of inexpensive DNA sequencing technologies, researchers finally have the opportunity to sequence thousands of individuals in a population. This presents the scenario that every individual will be sequenced, perhaps multiple times in their lifetimes, providing a comprehensive and unbiased look at genomic variability in the population. A few large scale studies have explored this genomic variability [59, 70], and have shown that the genomes are surprisingly plastic, diverging not only with single nucleotide variations, but include large structural changes involving deletions, inversions, translocations, and duplications of large portions of the genome. It is only to be expected that these genomic changes also modify the structure, splicing patterns, and the primary sequence of the expressed transcripts and proteins.

Historically, gene finding has been solely the province of the genomics community. In addition to de novo signals for coding regions and splicing, gene finding tools also make use of transcript information to identify genic regions, splicing, and other information. The availability of RNA-Seq and other deep sequencing technologies for RNA has the promise to radically improve genomic annotation. ENCODE and other, similar projects
have made effective use of RNA-Seq, ChIP-seq, and other technologies to improve the functional annotation of the genome [64].

Nevertheless, challenges remain, even with simple gene finding. Although RNA-seq provides a deep sampling of expressed genes within the sample, not all genes are expressed at one time. Therefore, RNA-seq data generated from multiple experiments must be used in a cumulative manner. The transcribed portion of the genome appears to greatly exceed the translated portion, and everything that is transcribed may not be translated. Transcriptomes do not provide information on the reading frame, and large amounts of pre-spliced and un-spliced RNA mask true splicing events.

The emerging field of proteogenomics attempts to remedy this by using proteomic information derived using tandem mass-spectrometry to augment the transcript information. For example, we can search MS spectra against a translation of RNA-seq reads, but this is both inefficient and redundant. Typical RNA-seq database sizes match the size of the genome, while only sampling a small fraction (∼3%) of it. An improvement is to assemble RNA-seq fragments into longer transcripts, and search these reduced databases [67, 27]. However, this approach also has many shortcomings. First, information is lost during the assembly, and indeed a wrong call might be made among competing splicing events. A peptide might match multiple isoforms derived from the same set of reads. Information on mutations is often discarded during assembly.

Further, the best sensitivity is obtained by accumulating, and searching RNA data across multiple conditions and cell-types. However, it is technically difficult to assemble multiple RNA-seq data-sets given the huge numbers of experiments. As an extreme example from humans a single project (The Cancer Genome Atlas or TCGA) project lists over 240Tb of RNA-Seq data across multiple cancer sub-types [69]. It is not clear that there is an effective way to search all of these data-sets, even when limited to a specific sub-type. Previous studies such as Wang et al. (2012) [72], focused on creating
customized proteomic database by reducing the search space using RNA-seq in order to increase the sensitivity of the peptide identification proteomic database.

Our study has different goals, focused on finding novel gene events. For this reason, we explicitly search both RNA derived, and 6-frame translated data-sets. Therefore, we focus on maximizing the sensitivity of the database itself with respect to gene features such as splicing and translated regions. In addition, rather than using matched RNA and proteomic samples, we work on aggregated RNA datasets from multiple experiments to maximize sensitivity, and remove the constraint of proteomic and RNA data being from the same sample. To reduce the search space, we construct a non-redundant compact database that contains useful splicing information expressed in RNA-seq reads, along with enough information that any MS search tool can identify peptides. We have developed a tool to construct a splice graph database using RNA-seq fragment mappings. The database encodes a graph $G$ where genomic intervals (exonic regions) correspond to nodes, and edges correspond to pairs of exons that are putatively spliced together. RNA-seq read mappings that split across splice-junctions are used to determine edges. The huge compression in database size comes from the tremendous redundancy of transcript generation and mapping. Mutations, including small insertions and deletions are repeatedly sampled both within a data-set, and also across many data-sets. Unlike transcript assembly, the splice-graph does not have to select between specific splicing paths. Thus, there is no loss of sensitivity, even with the large compression. We have previously used a splice-graph encoding of cDNA sequences (typically from EST projects) for proteogenomic studies [11, 12]. Here we modified the tool to deal with very large RNA-seq data-sets and release it for general use. As shown in the results, we can compress large (> 490Gb) of RNA-seq mapping data to a compact database of 0.4Gb.

While some MS2 identification software can search splice-graphs directly [61, 12, 11], most tools require a FASTA formatted sequence database. To generate a universal
database, we also present a tool to convert splice graph data structure into a multi-FASTA formatted file. The naive approach would enumerate all paths in the graph, leading to a large expansion in size. Instead, we exploit the short length of typical ‘bottom-up’ peptides. Using an adjustable parameter \( L \) as the maximum length of a peptide, our tool generates a highly compressed FASTA file that encodes all splice-graph peptides of length \( \leq L \). Applying this method reduced 496.2GB of aligned RNA-seq SAM files to a 410MB splice graph database written in FASTA format. This corresponds to 1000\( \times \) compression of data size, without loss of sensitivity.

We performed a proteogenomics study using our pipeline and identified a total of 4044 novel events (as compared to \emph{C. elegans} gene set version WBcel215.68 [24]). The identified events included 215 novel genes, 808 novel exons, 12 alternative splicings, 618 gene-boundary corrections, 245 exon-boundary changes, 938 frame-shifts, 1166 reverse-strands, and 42 translated UTR. Our results highlight the usefulness of transcript+proteomic integration for improved genome annotations.

### 1.2 Method

The pipeline has two major parts: splice-graph construction from mapped reads, and splice-graph to FASTA conversion.

**Input preparation:** RNA-seq data used in our study was generated by the Waterston lab. RNA-seq reads were mapped using the method described in previous studies [29, 26] as part of the modENCODE project. RNA-seq read alignments are in SAM format [37], a column based encoding of the alignment of each read. The alignment itself is stored in a compact CIGAR string of the SAM file. We parse the CIGAR string to obtain the split information in spliced reads. Resulting coordinates of the mappings are converted into GFF, a simple interval-based flexible format for representing genomic intervals.
Calculation of the split mapped coordinates from the CIGAR string is described in the Supplementary Material section. While RNA-Seq reads can span multiple splice junctions, we did not find such instances in our data set which had 76bp sequences. However, the software automatically, through its parse of the CIGAR string, identifies, and creates (multiple) splice junctions per read. The functionality of encoding genomic variants such as insertions, deletions, and mutations, is not applied in this study due to the lack of known genomic variant information for this data-set, but it was used in other data not presented here.

Filtering RNA-seq reads: We employed a number of filtering steps to reduce the size of the database. To start, note that the final proteogenomic study involves searching MS/MS spectra against three different databases. These are a database of known proteins, a 6-frame translation of the entire genome, and the splice graph database. In the past, a search of the 6-frame database has been eschewed when transcript data is available. However, typical RNA-seq databases are often the size of the genome (or larger), and when multiple RNA-seq databases are being searched, the burden of searching the 6-frame translation becomes less dominant. Moreover, all non-spliced, non-mutated, peptides can be identified in the search. Therefore, as a first filtering step, we discarded all RNA-seq fragments that are not split-reads, i.e., they map to the genome without splicing. This simple step removed 71.79% of the total RNA-seq reads.

In a second filtering step, we considered mappings of the split-reads to the genome and compress all the reads sharing the same intron boundary into a single read. As in Figure A.1, if a pair of reads share the same intron coordinates, we merged them into a single read preserving the boundaries of the intron, and extend both ends of the read. In this process, we maintained hash table using intron coordinates as a key value which represents a unique splice junction (intron coordinate pair). Entries of the hash table
contain information of expanding exon boundaries on each side, RNA-seq read counts, and original RNA-seq file name. Once all files were processed, we filtered out all putative introns that are not covered by at least $c$ reads, where $c$ is a user-defined parameter with default value $c = 2$. This second filtering stage removed 98.16% of the reads that survived from the first filtering step. After applying the above two steps of filtering, a total of 4,669,116,388 RNA-seq reads were reduced to 517,326 merged RNA-seq components, and the average exon length on each side of the component was 83.42bp. Note that the goal is not to cover the entire exon, but to cover split peptides of maximum length $L$. In our study, we used 90bp (30 amino-acids) for the value of parameter $L$.

To further reduce the computational burden, we partitioned the merged data into multiple files, based on mapping coordinates. The construction ensures that splice junctions do not exist between multiple files, and therefore the true splice graph is simply a concatenation of the splice-graphs from each file. The splice-graph construction was done in parallel for each file.

**Constructing the splice-graph:** In the splice graph data structure, nodes represent exons, and edges represent splice junctions. The construction is schematically illustrated in Figure 1.1. Starting with the empty graph, the splice-graph is augmented/updated read by read. (here, a read represents a single merged component of RNA-seq reads which is the output of the previously described filtering stage.)

See Figure 1.1 for an example. Given RNA-seq read $r_1$, node $s_1$ is split into nodes $u_1$ and $u_2$, and node $u_3$ is added. Next, we assign edges for each spliced-read. In Figure 1.1(c), edge $e_4$ is added to the current set ($e_1-e_3$). Finally, we revisit each pair of contiguous nodes, where contiguous means that there is no coordinate gap between the previous and next node. In Figure 1.1, $u_1$ and $u_2$ are contiguous, while $u_5$ and $u_6$ are not contiguous since there exists a gap in between. The contiguous nodes are merged if there
is no assigned edge between the corresponding pair; otherwise, they are connected by an additional edge. For example in Figure 1.1(d), $u_1$ and $u_2$ are merged since there is no edge between. On the other hand, $u_2$ and $u_3$ cannot be merged due to the existence of $e_1$, and the additional edge $e_5$ is assigned.

### 1.2.1 Converting splice graph structure to a FASTA file format

While the splice-graph database is a compact encoding of splice patterns, it cannot be searched directly by standard MS/MS search tools. To overcome this limitation, we developed a tool that generates a FASTA formatted database from the splice-graph.

The generated FASTA database must have certain properties that relate it to the splice-graph database. Following Edwards and Lippert [18], we say that a FASTA database $F$ is $L$-Complete w.r.t a splice-graph database $G$ if every length $L$ sequence in $G$ is a substring in $F$. In addition, $F$ is correct w.r.t $G$ if every string in $F$ is also a substring in $G$. Given a splice-graph $G$, and a user-defined parameter $L$, our objective is to generate a minimum size (number of amino acids in database) FASTA database $F$ that is correct and $L$-complete w.r.t $G$. A naive approach for splice graph to FASTA conversion is to retrieve all possible paths within $G$ and generate a new FASTA sequence. Such a database is complete (for all $L$), and correct, but will be greatly increased in size, growing exponentially in the average node-degree of $G$. Similar to Edwards and Lippert [18], we also describe a novel method which finds a greedy but effective solution of this problem. However, unlike Edwards and Lippert [18], our method uses a genomic-coordinate based data structure (represented in base pairs) rather than minimizing the amino acid sequence overlap. We claim that for proteogenomic analysis, the coordinate based approach is more appropriate since it can easily reconstruct the original genomic coordinate of the identified peptide.
Figure 1.1. (a) Given RNA-seq read, find overlapping regions with the existing splice graph. (b) Split and add nodes. (c) Assign edges for each spliced-read. (d) Revisit each pair of contiguous nodes. The nodes are merged if there is no edge at the boundaries.
FASTA conversion strategies: We used three rules to eliminate shared sub-paths.

1. For a pair of paths, \( xz \) and \( yz \) with a shared string \( z \), we generate two FASTA strings: \( xz \), and \( y \cdot \text{pref}_L(z) \), where \( \text{pref}_L(z) \) denotes a length \( L - 1 \) prefix of string \( z \).
2. For a pair of paths, \( xz \) and \( xy \) with a shared prefix \( x \), we generate two FASTA strings: \( xz \), and \( \text{suff}_L(x) \cdot y \), where \( \text{suff}_L(x) \) denotes a length \( L - 1 \) suffix of string \( x \).
3. For paths \( xy \) and \( yz \), which have a prefix-suffix match with \( y \geq L \), generate the FASTA string: \( xyz \).

Rules 1 and 2 can be implemented in an enumerating procedure during a depth first search (DFS) traversal of the splice graph. Recall that in a standard DFS search, a node is marked the first time it is visited. Thus if a previously-visited node \( v \) is revisited, we keep only the length \( L - 1 \) path from outgoing edges to \( v \). Likewise if a traversal touches a node with multiple outgoing edges, we need to only maintain a length \( L - 1 \) suffix to attach to each of the outgoing paths. (See Figure 1.2)

Rule 3 allows us to combine pairs of sequences that share a prefix and suffix string. First, we identify overlap-node-pairs as pairs of merge nodes (out degree > 1) and split nodes (in-degree > 1) with length \( \ell (L \leq \ell < 2L) \) sequence between the two. If \( \ell < L \) (seq1 and seq2 in Figure S. A.2(a)), the generated sequences cannot share an identical prefix and suffix. If \( \ell \geq 2L \), the prefix and suffix of generated sequences will not overlap (Figure S. A.2(b)).

For implementation of rule 3, we used a hashing technique to rapidly identify overlap-node-pairs. Traversing the graph in a depth first fashion, we store all the split nodes present in a candidate list. For each split node \( u \), we consider the sequence of nodes encompassing the length \( L \) prefix of \( u \), and hash the prefix string using the first 3 nodes as key (Figure S. A.3(b)), so that each key contains the list of the paths such that prefix of the paths is the same as the corresponding key. Every time a merge node
Figure 1.2. (a) Depth first search (DFS) traversing of the graph. (b) In traversing, we only maintain a length \( L - 1 \) suffix. (c) In encountering an incoming visited edge, only maintain a length \( L - 1 \) prefix. (d) For a pair of sequences, combine two sequences.

is encountered in the DFS, we traverse the subsequent path, querying the hash table continuously using 3 node triplets. For example in Figure S. A.3(c), key2, key3, and key4 are used to query the hash table. When a match is found (e.g., between key4 and key1), the hash table returns a list of sequences that corresponding paths starting with the appropriate key. (e.g., \( \text{‘TCG’+‘CG’+‘GG’+‘AAC’+‘CCTA’+‘AATATG’} \)). We search each sequence within the returned sequences, using the remaining suffix of the queried sequence. In our example, the remaining sequence is ‘A’ which appears right after the key4. We merge the matched sequence with queried sequence, and translate it into three different frames. Finally, we output the 3-frame translated sequences to a FASTA file.

**Heuristic constraints to prevent exponential growth**

Growth of the final FASTA database is dependant on the complexity of the splice graph. Splice junctions expressed in RNA-seq reads are expressed as edges in the graph. Multiple edges assigned within a small region (less than \( L \)) will increase the complexity of
the splice graph structure. Therefore, the final FASTA file will grow exponentially in the case where many splice junctions are found within a small region. To prevent exponential growth in very complex regions, we added an additional constraint to our conversion strategy. Based on the RefSeq known protein database, we set a proper length parameter $W$ as the minimum distance between adjacent splice-junctions. In our implementation, if two splice junctions appear within $W$ bp of each other, the FASTA sequence generation selects each splicing independently, but not in combination. We used $W = 20$bp in this study (See Figure A.4 for the description of $W$ parameter). Note that the average length of exons in RefSeq C.elegans database was 207.62bp (s.d. 262bp). Only 1.01% of known exons were shorter than 20bp. The proofs of correctness and completeness in applying all methods above are illustrated in Supplementary material.

1.2.2 Datasets and experimental procedure

RNA-seq data was generated by the Waterston lab as part of the modENCODE project. The dataset used was 111 experiments from multiple Caenorhabditis species and developmental stages. RNA-seq reads were mapped as described in the studies [29, 26, 45]. Detailed RNA-seq methods are illustrated in Supplementary material.

For the mass spectrometry data, eleven developmental stages of C. elegans were analyzed - N2 embryo, N2 L1, N2 L2, N2 L3, N2 L4, N2 YA, N2 dauer, spe-9 L4, spe-9 YA, spe-9 adult and him-8. Each developmental stage was grown on agar plates at 20°C, seeded with the NA22 strain of E. coli. [9], sucrose floated, lysed in the presence of protease inhibitors (Roche) and centrifuged to separate insoluble and soluble fractions. A 200µg soluble lysate of each developmental stage was reduced with DTT(Sigma) and separated into 15 molecular weight fractions ranging from 3.5 to 500 kDa using the GelFree 8100TM fractionation system (Protein Discovery/Expedeon) [65]. Each fraction was alkylated with IAA (Sigma) and trypsin (Promega) digested. SDS was removed with
SDS removal columns (Pierce) and salts were removed with MCX columns (Waters). The peptides from each fraction were analyzed using a 35 cm fused silica 75 µg column and a 4 cm fused silica Kasil1 (PQ Corporation) frit trap loaded with Jupiter C12 reverse phase resin (Phenomenex) with a 120-minute LC-MS/MS run on a Thermo LTQ-Orbitrap Velos mass spectrometer coupled with an Eksigent nanoLC 2D. A biological and analytical replicate was performed for each sample.

Using the constructed splice graph database, we launched a proteogenomics search of C.elegans MS/MS spectra dataset. C.elegans MS/MS spectra is a total of 81 GB in size, consisting of 11,123,595 spectra. The spectra dataset was produced by the MacCoss lab, and comparison to Merrihew et al. (2008) [45] is illustrated in Supplementary material. We used MSGFDB (version 20120106) [35] for the database using the following parameters: 30ppm for parent mass tolerance, semi-tryptic search, Carbamidomethylation of C as fixed modification, and Oxidation of M as optional modification. For each spectrum, we selected PSMs with the lowest SpecProb reported by MSGFDB across all database search results (known proteins, 6-frame, and splice-graph-fasta). The reversed decoy database was also searched for all databases to apply the target-decoy approach. The database search resulted in 65,874 peptides better than 1% spectral level FDR cut-off. Among identified peptides, 52,292 corresponded to known peptides, but 13,582 peptides were novel. The novel PSMs were mapped back to their genomic coordinates using automated scripts. The 13,582 novel peptides mapped to 15,205 different genomic locations, giving on average of 1.12 locations per peptide. Among 15,205 locations, 3,484 were identified from the splice graph database, and 11,721 were from the 6-frame database.

Using previously developed tools [11, 12], identified peptides were grouped together into a single event in a pairwise fashion if they were located within ≤ 1000bp apart. The novel events were called automatically along with an event probability. We
filtered out low quality results by setting the event probability cut-off as 0.998. For further validation, the novel events were plotted using the UCSC genome browser [31] and verified using comparative genomics (protein level BLAST [6]).

A second part of the software not used here, tracks how the novel findings were supported by specific RNA-seq data-sets, allowing for a more accurate correlation between protein and RNA evidence. In future work, we plan to apply this pipeline to compare MS and RNA data acquired on identical biological samples. Our software for splice graph construction and FASTA conversion is available at CCMS webpage (http://proteomics.ucsd.edu/Software.html).

1.3 Results

A splice graph was created from 496.2GB of aligned RNA-seq SAM files to 410MB of a splice graph database written in FASTA format. Overall statistics of our splice graph data structure is illustrated in Table A.1. The 6-frame translation database was created from the reference genome and also written in 102MB of FASTA formatted data.

Data compression

Figure 1.3(a) shows the overall increase in database size as a function of accumulating RNA-seq data. On the x-axis, the number of data-sets are progressively incremented up to 149 (496GB). The y-axis describes (on a log-scale) the growth of the corresponding splice graph and FASTA sequence database.

The 496.2GB RNA-seq data was compressed into a 410MB FASTA database, a 1000× compression in terms of file size. Most of the gains are due to the filtering of spliced reads. Within the filtering stage, 71.79% reduction was achieved from filtering split mapped reads, and the remainder was from merging identical splice junctions and
discarding ambiguously mapped reads. Since most of the size reduction was achieved in
the filtering stage, this indicates the strong advantage of aligning RNA-seq reads before
database construction, unlike other methods [18] where no coordinate information is
used in database creation. Additionally, we observed that the rate of growth of the splice
graph decreases after using 45 data-sets due to a saturation in the splicing information.

Note that our design choice of filtering out the non-spliced reads works because we
also search a 6-frame translation, which is 102MB in size. Thus, the 6-frame translation
acts as a compressed version of unspliced RNA reads, which in combination with the
splice graph reduces the 496.2GB file to (410 + 102)MB total. Moreover, due to the
large variation in transcript abundance, we observe that even the large set of RNA data
includes only 91% of known splice junctions, and we expect a similar ratio for known
exons. Therefore, the addition of 6-frame translation also improves the sensitivity of the
search by capturing all possible non-spliced translations.

The total computation time required for the database creation was 12 CPU-hours
for filtering, 2.5 CPU-hours for graph construction, and 300 CPU-seconds for FASTA
conversion. This database creation computation was performed on a Desktop PC with
Intel Core i7 2.67GHz processor and 9.0 GB of RAM.

**Database validation**

To validate the splicing information, we compared our Splice Graph database
with RefSeq [55] Accession NC_003279. The ideal Splice Graph database should cover
all known splicing junctions. Define the **RefSeq-splice-coverage (RSC)** as the fraction
of all RefSeq splicing events covered by the SpliceGraph. We observed that the RSC
value saturated after about 45 data-sets (Figure 1.3(b)), with most (though not all) RefSeq
splicings incorporated.

Additional growth in the Splice Graph was due to the incorporation of novel
splicings in the RNASeq data, but not in RefSeq. Figure 1.3(c) plots the number of novel splice junctions in the Splice Graph. Comparing the growths in Figure 1.3(b) and Figure 1.3(c), we observed a similar growth curve, with the observed rates being $1.2 \times (75.0\% \text{ to } 91.3\%)$ in coverage and $3.37 \times (123,670 \text{ to } 416,176)$ in number of novel splice junctions. The tremendous growth in novel splicing events, which might not be translated, highlights the ambiguity in locating gene events using RNA data alone, and underscores the importance of protein level validation via proteogenomics. Our proteogenomic search identified 2,126 novel spliced peptide locations from the total 416,176 novel putative splicings encoded in our splice graph database.

**Proteogenomics discoveries**

The compacted FASTA representation of splicings was used in conjunction with a known protein database in a proteogenomic analysis of a bottom up LC-MS/MS LTQ-Orbitrap data-set generated by the MacCoss lab. We used the existing proteogenomic pipeline, ENOsi [12, 11], which automatically searches all of the spectra against the custom databases, accumulates all results, employs FDR calculations to identify Peptide Spectra Matches (PSMs), clusters novel peptides, and calls events automatically. *Event Probability Score* for each novel event was calculated as, $1 - \left( \prod_{i \in S} \left( 1 - \frac{(1-FDR) \text{LocationCount}}{LocationCount} \right) \right)$, where $S$ is the set of peptides assigned in current event, *LocationCount* is the number of genomic locations of the identified peptide, and *FDR* is the calculated FDR value of the corresponding PSM. We only reported novel events with *Event Probability Score* larger than 0.998. The proteogenomic search of 11,123,595 spectra was performed using a cluster server with 125 cores in parallel. 6-frame database search was done in 34.96 wall time hours, splice-graph-fasta database search took 15.31 wall time hours, and known protein database search took 6.54 wall time hours.

We note that the splice-graph FASTA has extensive header information that
Figure 1.3. (a) Growth of the database file size while incorporating more RNA-seq data. (b) Increase in the percentage of covered splice junctions compared to RefSeq. (c) Increase in the number of splice junctions expressed in splice graph.
describes the mapped coordinates of reads and how they are split. The sequence part
of the splice graph is 114MB in size, and it still contains some redundant sequence
that can be efficiently handled by MSGFDB [35], which indexes the database using
suffix-tree techniques. Therefore, the search time of the splice graph FASTA is less than
the six-frame translation.

The search revealed 4,044 events, as shown in Table 1.1. Figure 1.4 shows a few
examples of the novel findings taken from our result which were further plotted using
UCSC genome browser [31]. Red blocks represent identified peptides, and sky-blue
blocks represent split mapped RNA-seq reads used in creating splice graph database.

We note that each event must contain at least one uniquely located peptide
sequence match (PSM). However, a peptide group contains multiple peptides, a single
group can represent multiple events. For example, there could be a group of peptides
that extend the end of a gene by jumping past the stop codon, and adding a new terminal
exon. In this case, the peptides support both ‘alternative splicing’ and ‘novel exon’. In
another way to parse the solution, we identified 5463 unique novel peptide locations, and
3979 novel clusters.

**Novel genes:** We identified 215 novel gene events Table 1.1 where a collection of
novel peptides were located ≥ 3Kbp from any annotated gene, again underscoring the
impact of proteomic data on the discovery of new genes. Moreover, even with the ex-
tensive RNA information, we identify new genes from the 6-frame search as well. An
example is shown in Figure 1.4(a), with 2 peptides, R.SRKSLPRTSQSPSSNFSGFY.V
and R.CYRYIIVSDIEKAFHQVRLQKA FR.N, both from 6-frame database search.
We looked for comparative evidence using Blastx [6]. A query sequence was ex-
tected from the DNA region ‘chrV:19272600-19274335’ and searched against the
nr protein database [43]. The top blastx hit was ‘hypothetical protein CRE-09558
[Caenorhabditis remanei] with e-value 0.0. As shown in Figure A.5, peptide R.CYRYIIVSDIEKAFHQVRLQKAHR.N was aligned with the protein sequence of Caenorhabditis remanei indicating that a similar CDS region exists in Caenorhabditis remanei which is a positive evidence of protein translation. Furthermore, we also found a supporting evidence from a predicted protein sequence generated from the GeneFinder [45] in the same region containing the peptide sequence ‘R.CYRYIIVSDIEKAFHQVRLQKAHR.N’.

**Gene corrections:** The majority of the events in Table 1.1 are corrections to existing gene structures, including novel exons, extensions to UTRs, alternative splicing, frame correction, and even reverse strand events. We identified 12 alternative splicing events, with junctions that differed from RefSeq. Figure 1.4(b) shows 2 novel spliced peptides, ‘T.LNVNGQE:IVYSMENEK.L’, and ‘R.EIKK:QHTSFQVSGPKEEIVYSMENEK.L’ in their genomic context. The notation ‘:’ indicates where the splice junctions are located. In peptide ‘T.LNVNGQE:IVYSMENEK.L’, the splice junction spans the amino acid ‘I’. The peptides are well represented on either side, and located uniquely in the genome. Splice junction of peptide ‘T.LNVNGQE:IVYSMENEK.L’ was supported by 13 split mapped RNA-seq reads, and peptide ‘R.EIKK:QHTSFQVSGPKEEIVYSMENEK.L’ was supported by 40 reads. The peptides identify a novel splicing in the gene vit-5, part of a 5-member family of vitellogenin genes involved in maternal yolk production [16].

We identified 938 ‘frame-shift events’, where peptides match to known genes but in a different frame. In Figure 1.4(c), we identified the peptide ‘TIVFTVPLSQCMVSP-MISK.E’ (in the gene eef-2), which matches in a different frame. Two neighboring peptides, ‘R.FIEPIEDIPSGNIAGLVGVDYL:S.R’, and ‘G.HVFEESQVTGTPMVFV:R.L’ were identified with 1 bp deletion, that allow for the frame-shift to occur. This region has complex RNA-seq mapping containing many small deletions, implying DNA assembly error, or a high degree of polymorphism in the region.
**Figure 1.4.** (a) Shows a novel gene area where two peptides are identified in a non-genomic region. (b) Two peptides with alternative splice junctions. (c) Peptide matches in a different frame compared to the gene eef-2 with two neighboring peptides.

We measured the distribution of the novel peptides across different developmental stages. Figure A.6 shows the spectral counts of novel peptide spectra related to translated UTR events across different developmental stages. There is a small bias toward early developmental stages relating to translated UTR events. The translated UTR events suggest new transcription start sites (and alternative regulation) of genes in early developmental (‘N2 L1’) stage.

To compare peptide versus RNA abundance, we computed a scatter-plot (Supplemental Figure A.7) of RNA-seq read counts mapped to a known gene (x-axis) versus the spectral count of peptides (y) falling within the region. The correlation between two values was calculated as 0.31 which implies that only a weak correlation is observed. However, since we are looking at the statistics of the accumulative data collected from various studies, we may need more detailed information on time specificity and sample consistency in order to study this correlation.
Table 1.1. The statistics of novel events identified

<table>
<thead>
<tr>
<th>Novel Events</th>
<th># of events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternative Splice</td>
<td>12</td>
</tr>
<tr>
<td>Novel Exon</td>
<td>808</td>
</tr>
<tr>
<td>Novel Gene</td>
<td>215</td>
</tr>
<tr>
<td>Exon Boundary</td>
<td>245</td>
</tr>
<tr>
<td>Gene Boundary</td>
<td>618</td>
</tr>
<tr>
<td>Reverse Strand</td>
<td>1166</td>
</tr>
<tr>
<td>Frame Shift</td>
<td>938</td>
</tr>
<tr>
<td>Translated UTR</td>
<td>42</td>
</tr>
</tbody>
</table>

1.4 Conclusions

Our manuscript makes two points. First, cumulative mass spectrometry information (acquired in multiple studies) is a useful data resource for improving genome annotation, and should be applied as a standard part of continuing annotation efforts. Second, incorporating RNA-seq toward genome-annotation remains non-trivial due to high redundancy, large data sizes, but also, the difficulty of assuming translation given transcription information.

At the same time, judicious use of RNA-seq databases can be made by compacting and saving the information non-redundantly and using it to gather proteomic (translation level) information as an aide to genome annotation efforts. On the well-annotated C. elegans data, we still succeeded in identifying over seven thousand novel events. In developing our methods, we made many design choices, including ‘mapping raw RNA reads’ versus transcript assembly, and maintaining a single comprehensive database of all RNA. Our results suggest that this is a better, and more inclusive approach to combining RNA and protein data, and can be reused for all organisms. Our splice graph database construction pipeline produces a conventional FASTA database that can be applied to any kind of proteomics study, while achieving large scale data compression with no loss of useful information.
Utilizing RNA-seq information in proteogenomics database construction has many other benefits, including the computation of sample specific expression, and genomic variation identification. Future improvements of our pipeline will extend to mapping all variations in addition to splicing events, and further the use of proteomic data in genetic studies.
ACKNOWLEDGEMENTS

This chapter, in full, is a reprint of the material as it appears in Journal of Proteomic Research. S. Woo., S. Cha, G. Merrihew, Y. He, N. Castellana, C Guest, M. MacCoss, and V. Bafna, 2013. The dissertation author was the primary investigator and author of this paper.
Chapter 2

Proteogenomic strategies for identification of aberrant cancer peptides using large-scale Next Generation Sequencing data

2.1 Introduction

Cancer is driven by the acquisition of somatic DNA lesions. Understanding of the progression of the lesions, distinguishing the early driver mutations from subsequent passenger mutations, deciphering the role of somatic mutations in regulating protein expression are all under active investigation. The availability of genomics technologies (mainly whole-genome and exome sequencing, and transcript sampling via RNA-seq, collectively referred to as NGS) have fueled recent studies on these topics [36, 8]. It is very likely that the some of the discovered mutations will aid in molecular sub-typing of cancers, and act as diagnostic and prognostic bio-markers.

A challenge to this vision comes from the complexity, redundancy, and errors in genomic data, and the difficulty of investigating the proteome translated portion of aberrant genes using only genomic approaches. In comparative studies, while protein and RNA expression matched for the most abundant molecules, the correlation for lower
abundance molecules was much worse ($\sim 0.4$) [28]. Others found that as many as 20% of transcripts do not have a matching protein identification, often due to a different frame of translation [53]. The high variability between protein and genomic expression in these studies suggests that a combination of proteomic and genomic technologies are the best bet for identifying coding variants and their use as biological markers of cancer, and such searches are increasingly employed [39, 38, 73]. Moreover, one cannot rely on comparison of RNA and protein data from the same sample.

The problem of searching all protein samples and all RNA samples becomes a significant challenge for proteogenomics, especially for bottom up mass spectrometric protocols, where a short peptide spectrum is matched against theoretical databases of spectra derived from genomic sequences. The chance of a false identification grows with increasing database sizes. A typical RNA-seq alignment file is around 10 GB, and is different for each sample. The TCGA resources [1] alone lists around 5Tb of RNA-seq data for Ovarian Carcinoma. In order to utilize large-scale NGS data in proteomics search, efficient methods for managing the large data-size are essential. This paper provides an efficient method to search the large search space of NSG data and discussion of applying more accurate FDR based error control strategies, and their implication to cancer proteogenomics.

**Large Database Search.** As our goal is to discover aberrant peptides in cancer, including fusion genes, splicing variants, and possibly even novel expressed genes, we cannot rely on the human proteome, hence large databases. First, a six-frame translation of the human genome is already $\sim 6Gb$, but that pales in comparison to the available transcript data that encodes many of the variants. For example, the TCGA resources [1] lists around 5Tb of RNA-seq data for a single tumor type (Ovarian Carcinoma).

Some approaches have been suggested to handle the big-data overload. These
include, sample specific search to reduce the search space by generating a curated individual database for RNA-seq obtained from each sample [39, 38, 73], and direct translation of the outputs from available genomics assembly tools [66, 33, 32, 44, 17]. Alternatively, our method favors a graph structured accumulative approach [75] that combines multiple sample NGS data into a unified database. A graph based approach enables us to efficiently encode cumulative/large information from multiple RNA-seq data-sets into a compact unified database. Moreover, unified database approach also enables us to maintain a single FDR threshold throughout the entire analysis. Finally, our approach also enables peptide identifications with combinatorial multiple splice junctions or variants. Based on our proposed method, we released a JAVA and python based tool called SpliceDB (https://bix-lab.ucsd.edu/display/CCMSwebsite) which generates FASTA formatted splice graph database from multiple RNA-seq alignments.

**False discovery rate based error control strategies.** One of the challenges with proteogenomic studies is the aggressive and variable choices of False Discovery Rates (FDR) strategies, all designed to maximize the discovery of aberrant peptides. In most conventional proteomic studies, a global peptide level FDR with 1% FDR cut-off is used. However it has been shown from other studies that FDR threshold can be biased in a larger database search space such as PTM and SNP [39] tolerant searches. In this study, in order to discuss the effect of applying different FDR strategies, we performed a benchmark study under identical condition applying three different FDR based peptide error control strategies which are Combined FDR (Supplemental Figure B.1(a)), Separate FDR (Supplemental Figure B.1(b)), and Two-stage FDR (Supplemental Figure B.1(c)). While more sophisticated FDR approaches can be further applied in combination of our FDR strategies, here we calculated standard global FDR threshold in order to mainly focus on the effect of separate, multi-stage calculation strategies. As shown in our result,
in applying conservative FDR error control, our results are robust to the choice of FDR.

**Calling peptides versus events.** An important part of proteogenomics search for discovering aberrant events is that we are looking for events (alternative splicing, gene-fusions, etc.), not peptides. The SpliceDB tool described here can be used in stand-alone fashion just for FASTA database creation, but also can be paired with our integrative proteogenomics pipeline Enosi [61, 12, 11, 13]. To focus on the effect of choosing different strategies in ‘peptide identification results’, we will not describe cancer specific event calling here, but will present some results describing events we could identify in a proteogenomic search of a single primary ovarian carcinoma sample.

To summarize, the manuscript makes the following contributions. First, we extend the SpliceDB database construction to scale to human cancer data-sets, and include all different types of variation. We build and present a generic ovarian cancer database that can be searched with any proteomic data-sets. We utilized a total of 879 BAM files downloaded from TCGA [8, 36] repository and created total 4.34 GB ($10^3 \times$ compression) of unified FASTA database which contained 2,787,062 novel splice junctions, 38,464 deletions, 1105 insertions, and 182,302 substitutions.

Next, we systematically test the impact of applying different strategies regarding to database construction and FDR based error control on the identification of aberrant peptides in cancer. Total 439,858 spectra collected from a single ovarian cancer sample were searched against the both the created FASTA database as well as a sample specific database. By applying most conservative FDR measure, we could identified 524 novel peptides and 65,578 known peptides at 1% FDR threshold. Moreover, selected detailed examples of doubly mutated peptide and different-sample-recruited mutation identifications were shown to emphasize the strength of our method, and the large number of identifications from a single sample underscore the value of proteogenomic searches in
identifying aberrant peptides in cancer.

## 2.2 Method

Following our previous study [75], we extended the splice graph database construction method to encode a more extended list of genomic variants. Splice graph [75] is a data structure which represents exons as nodes, and splice junctions as edges. The graph is constructed from the junction information extracted from the RNA-seq alignments and all types of mutations reported from VCF (Variant Call Format) files. For variant calling from RNA-seq alignments (in BAM format), we used GATK [44, 17] tool with parameters ‘−stand_call_conf 30.0 –stand_emit_conf 10.0’. Detailed descriptions on initial RNA-seq information handling, graph algorithms, and FASTA conversion algorithms, are described in our previous study [75]. The graph construction is done in an accumulative fashion, and the last FASTA conversion step must be performed each time when additional information is incorporated into the graph. Our graph construction approach also conserves the property of compactness [18, 75] and completeness [18, 75] of the original search space (state of proof shown in our previous study [75]). In this study, we introduce a concept of variant graph which enables additional nodes and edges representing arbitrary length deletions, insertions, and substitutions.

### 2.2.1 Database creation from RNA-seq data

RNA-seq data is downloaded from TCGA data repository [1] in BAM formatted files. Total size of the downloaded BAM files are shown in Figure C.1. Our first step in database creation is to extract useful information from RNA-seq alignment/coordinate files (BAM/SAM [37], GFF, BED) and variant calls (VCF). Details of BAM file processing can be found in our previous study [75].
Storing genomic information for post-processing usage. RNA-seq level information (sample ID, read counts, junctions, variants, and so on) is not only used in database creation but also heavily utilized in further proteogenomics analysis. In order to efficiently maintain and retrieve various types of RNA-seq level information, we implemented a multiple depth hash table structure which enables fast access to the source information.

As described in Supplemental Figure B.2, SpliceDB[4] extracts information from input files (BAM/SAM, GFF, BED, VCF) and generates a hash table using three key-data pairs. Three key values used in this hashing stage are, (1) category of a variant call (splice, deletion, insertion, and substitution), (2) chromosome name, and (3) beginning coordinate of a junction/variant. For example, in case when VCF file calls an ‘AT’ insertion in chromosome 1 at 30000th base pair, an entry containing RNA-seq level information (sample ID, read counts, junctions, variants, and so on.) is created using three key pairs of (insertion, chr1, 30000). Information maintained in the hash table is written to an intermediate file (.spl) for future usages such as cumulative database concatenation and validation of proteogenomic discoveries [13].

Variant graph construction. Our next step in database creation is to construct a graph data structure using information collected from the previous stage. The underlying method in graph construction and FASTA conversion for variant peptide database is shared from our previous study [75]. In this study, we extended our method to a population based study where individual genomes differ from the standard genomic reference due to the presence of mutations. These mutations may be a germ-line, somatic mutations, or even polymorphic, i.e. Somatic mutations can be distinguished from germ-line mutations by comparison with DNA from the same individual. Since protein level data used in this study does not include non-tumor samples, we treated both types of mutations equally during the MS/MS search but only differentiate it in the post-processing
stage while retrieving the originated genomic level information. In order to encode all types of variants into the graph structure, we added additional types of nodes and edges. While deletions can be expressed similar to splice junctions within the graph, insertions and substitutions cannot be incorporated using the same concept. Since insertions and substitutions cannot share the coordinate system of the reference DNA, we introduced insertion and substitution nodes having artificial coordinates which can be inserted to the existing graph. As described in Figure 2.6, negative numbers in different ranges are used to distinguish between inserted and substituted nodes.

The variant graph can be written in FASTA format by applying the conversion strategy introduced in our previous study [75]. In the FASTA conversion stage, the coordinate information of each entry is written in the FASTA header in order to reconstruct the original genomic coordinates of identified peptides.

**Restoring genomic information.** In proteogenomic analysis, genomic information of identified peptides such as original coordinates, and RNA-seq meta-data must be restored after the MS/MS search. We restore this information by using FASTA file headers and intermediate (.spl) files created during the database creation process. First, original coordinates of identified peptide sequence are calculated according to the corresponding FASTA header entry.

After having the peptide identifications, we can reconstruct the original coordinate of each peptide. For example, if we have a variant graph shown in the Figure 2.6, the graph can traverse the path ‘n1-n2-n3-n4-n6-n8’, and its corresponding nucleotide and amino acid sequence will be ‘GCTGCGCCAGAACCTACAATCGGA’, and ‘AAPEPTIG’. Next let’s assume that we have an identified peptide ‘PEPTI’, then we can find the coordinate of the peptide from its FASTA header. In this example, ‘PEPTI’ begins at the third amino acid of ‘AAPEPTIG’, so the beginning coordinate of the peptide will be ‘10006’ and the
ending coordinate will be located after traversing 15 nucleotide starting from ‘10006’. In this case, ‘chr1: [10006:10009] [-2:-1] [10009:10016] [-1003:-1001] [10018:10020]’ is the actual coordinate of ‘PEPTI’. Moreover, these restored genomic coordinates are next used in retrieving the hashed RNA-seq level data. In the above example, a set of hash keys indicating the first insertion and the second substitution will be generated as (insertion, chr1, 10009) and (substitution, chr1, 10016).

2.2.2 Database Search Details

MS/MS data used in this study was generated from PNNL (Pacific Northwest National Laboratory) as part of the CPTAC [2] (Clinical Proteomic Tumor Analysis Consortium) project. From the total MS/MS data generated by our collaborators, in this study we used a single sample. Additionally, iTRAQ quantification information is not utilized in this study since the goal of this study is focused on aberrant peptide identification. The 439,858 spectra acquired from a single ovarian cancer sample (sample id: TCGA-24-1467, see Methods) were used in this study, and searched against all proteogenomic databases (Table C.1). We used MSGF+ [35] for MS/MS database search with following parameters: parent mass tolerance 20ppm, semi-tryptic, Fixed Carbamidomethyl C, optional Oxidized methionine, and fixed iTRAQ related modifications. Known protein database was downloaded from Ensembl [24, 3](version GRCh37.70) which contained 104,785 sequences. We attempt to use this comparably richer set of known protein database in order to be more conservative in our novel sequence calling. By categorizing any previously known genomic variants included in the Ensembl known protein database as ‘known peptide sequences’, we tempt to focus more on identifying possible ‘cancer related’ mutations. The reversed decoy database of the same size was created for each database and also searched for all databases to apply the target-decoy approach. Using 100 CPU nodes of the CCMS cluster server in parallel, the total search
took 28.63 wall clock hours. For each spectrum, we selected PSMs with the lowest SpecProb reported by MSGFDB across all database search results (known proteins, 6-frame, and proteogenomics FASTA).

### 2.2.3 FDR based error control strategies

In this study, we applied three different FDR based error control strategies approaches for testing their effect on novel peptide identifications. In order to design accurate benchmark comparisons and to highlight the effect of combined, separation, and multi-stage FDR strategies, we calculated the global level FDR without applying further more sophisticated FDR calculations. (for example, further sub-classifying PSMs into different charge states, or utilizing peptide length and modification rates)

Indeed, the challenges with current proteogenomic studies is the aggressive and variable choice of False Discovery Rates (FDR) based error control strategies, all designed to maximize the discovery of aberrant peptides. In most conventional proteomic studies, a global peptide level FDR calculation with 1% FDR cut-off is used. However it has been shown from other studies that FDR based error control can be biased in a larger database search space such as PTM and SNP [39] tolerant searches. Here, we discuss based on a single conservative choice of FDR based error control strategies, how different ways to execute a search (Supplemental Figure B.1) may lead to different results.

**Combined FDR (Supplemental Figure B.1(a)).** (1) Merge MS/MS search results from every target and decoy database according to the best scored PSM per each spectra. (2) Calculate FDR using the combined PSM result.

This FDR is identical to the conventional peptide level FDR based error control used in most proteomics studies. The term ‘combined’ refers to combining reference protein databases and translated RNA databases into a single database, to be used for
proteogenomic searches.

**Separate FDR (Supplemental Figure B.1(b)).**  (1) Merge MS/MS search results from every target and decoy database according to the best scored PSM per each spectra. (2) Iterate all merged PSM and check the origin of the matched database entry. (3) If a PSM is matched to the known protein and it’s decoy database, put the corresponding PSM to the known sequence PSM set. (4) If a PSM is matched to any proteogenomic database and their decoy, put this into the novel sequence PSM set. (5) Calculate separate FDR in each known and novel PSM set.

Following the FDR approach suggested in the study of Jing et al. (2011) [39], the search uses a combined database to score and rank peptides, but separates known and novel PSMs prior to FDR calculation. This results in conservative novel peptide identifications [39]. The separation step is done strictly by iterating over every PSM to extract peptides that have string matches within the known and known-decoy database. Note that this procedure is different from simply launching the MS/MS search using only the proteogenomics database (excluding the known protein database) and calculating conventional FDR. Since most proteogenomics databases partially overlap with the known protein database, the proteogenomic database search might contain known peptide hits.

**Two-stage FDR (Supplemental Figure B.1(c)).**  1. Search only known protein DB and it’s decoy. 2. Calculate FDR in the known database only PSM result. 3. Search all proteogenomic databases and calculate FDR using only the spectra that are not identified through the previous known protein database search.

The two stage FDR is very similar to separate FDR but differ in some aspect. For example, it is possible that a search algorithm may assign higher score to a spectrum in
a novel peptide sequence compared to a similar (homologous) known peptide sequence. This can happen in the case when the spectrum contains high noise or missing peaks. This suggested multi-stage process guarantees that every known peptide that can be identified by conventional MS/MS search is not misinterpreted as novel peptide (this is important especially in the case of SNV mutated peptides due to the sequence similarity).

### 2.2.4 Sample preparation and LC-MS/MS analysis

TCGA ovarian tumor tissues were cryo-pulverized and homogenized in lysis buffer (8M urea, 100 mM NH4HCO3, pH 7.8, 0.1% NP - 40, 0.5% sodium deoxycholate, and protease inhibitors), after which the extracted proteins were reduced, alkylated and tryptically digested (Promega, Madison, WI) overnight. The resulting tryptic peptides were then cleaned up using strong cation exchange SPE and reversed phase C18 SPE columns (Supelco, Bellefonte, PA), dried and labeled with 4-plex iTRAQ reagents according to the manufacturer’s instructions (AB Sciex, Foster City, CA). The 4-plex iTRAQ labeled sample was separated on a XBridge C18 column (Waters, Milford, MA) using a LC gradient starting with a linear increase of solvent A (10 mM triethylammonium bicarbonate, pH 7.5) to 10% B (10 mM triethylammonium bicarbonate, pH 7.5, 90% acetonitrile) in 6 min, then 86 min to 30%B, 10 min to 42.5%B, 5 min to 55%B and another 5 min to 100%B. The flow rate was 0.5 mL/min. A total of 96 fractions were collected and concatenated into 24 fractions by combining 4 fractions that are 24 fractions apart. The concatenated fractions were dried down and re-suspended in 0.1% trifluoroacetic acid to a peptide concentration of 0.15 µg/µL for LC-MS/MS analysis.

The LC system was custom built using Agilent 1200 nanoflow pumps (Agilent Technologies, Santa Clara, CA). A 35 cm x 360µm.d. x 75µm i.d reversed-phase column was slurry packed with 3µm Jupiter C18 (Phenomenex, Torrence, CA). Mobile phase flow rate was 300nL/min and consisted of 0.1% formic acid in water (A) and 0.1%
formic acid acetonitrile (B) with a gradient profile as follows (min:%B); 0:5, 1:10, 85:28, 93:60, 98:75, 100:75. MS analysis was performed using a LTQ Orbitrap Velos mass spectrometer (Thermo Scientific, San Jose, CA) outfitted with a custom electrospray ionization interface. The ion transfer tube temperature and spray voltage were 300°C and 1.8 kV, respectively. Orbitrap spectra (AGC 3x106) were collected from 300 – 1800m/z at a resolution of 30K followed by data-dependent HCD MS/MS (centroid mode, at a resolution of 7500, collision energy 45%, activation time 0.1 ms, AGC 5x104) of the ten most abundant ions using an isolation width of 2.5 Da. Charge state screening was enabled to reject unassigned and singly charged ions. A dynamic exclusion time of 30 sec was used to discriminate against previously selected ions (within 0.55 Da to 2.55 Da).

2.3 Results

Database statistics. RNA-seq data used in this study was downloaded from TCGA [8, 36] repository.

The statistics of RNA-seq data-set and corresponding constructed FASTA DB are illustrated in Table C.1. We used ovarian and breast cancer sample data with a total of 6.27 TB of RNA-seq alignments in 879 BAM formatted files. In order to predict small nucleotide variants expressed the protein samples, we only used the subset of 67 files among 879 files that TCGA-sample-id that match with the PNNL selected samples for CPTAC [2] study. Thus, a total of 879 files are used in junction prediction, and 67 BAM files matching samples selected by PNNL for the ovarian cancer CPTAC [2] study. These 67 selected BAM files were plugged into the GATK [44, 17] tool for variant calling analysis. Using our SpliceDB workflow, a total of 879 BAM files were used in creating the splice graph, and separately, 67 VCF formatted GATK [44, 17] output files were used in the variant graph construction. Our final FASTA database size was 4.34 GB in
total, and contained 1,466,449 novel junctions (which includes 1,180,071 canonical GT-AG splice junction sites, and 24,433 small deletions less than 10bp), 38,464 deletions, 1,105 insertions, and 182,302 substitutions. The database will be made available for the usage researchers working on cancer proteogenomics in a way that agrees with the TCGA [36, 8] data usage guide lines.

Moreover, for the comparison experiment performed in the following section, we additionally created a genomic database from a single RNA-seq sample (sample id: TCGA-24-1467). From this sample, using our SpliceDB workflow applying same parameters, we created a 187 MB splice graph and variant graph database in FASTA format. This single sample variant graph encoded 168,289 novel splice junctions (which includes 161,935 canonical GT-AG splice junction sites, and 3,322 small introns less than 10bp), 62 insertions, 3,150 deletions, and 7,109 substitutions.

**Comparison between different FDR based error control strategies.** In order to test the effect of different FDR calculation strategies in proteogenomic searches, we applied three different FDR approaches to our PSM results. Figure 2.1 shows the number of known and novel peptide identifications using different target-decoy based FDR strategies. The diagram showing the descriptions of each FDR strategy is shown in Supplemental Figure B.1.

With combined FDR, we identified 60,877 known peptides and 1238 novel peptides. In contrast, the two stage FDR resulted 65,578 known and 524 novel peptide identifications. However, in combined FDR, we note that the number of peptides hitting the decoy database under a certain FDR threshold is very different in novel database versus known database. After applying combined FDR approach, we explicitly separated the PSMs from known, known decoy, novel, novel decoy, database and calculated the FDR in both novel and known peptide hits. As shown in Figure 2.1, we get 36% FDR in
novel peptides, and 0.03% FDR in known peptides while combined FDR was calculated as 1%. This indicates that combining the two PSM distributions raises the FDR cut-off for known peptides (lower identifications) and lowers it for novel peptides.

We choose a two-stage FDR approach in recognition of the differences in database sizes for the two searches. While results from the separate FDR is very similar to the two-stage FDR, two-stage FDR shows more conservative threshold on ‘novel identifications’.

**Comparison between single-sample-matched and unified database search.** In order to explore the trade-off between creating a single sample targeted database versus multiple sample unified database, we performed a computational experiment using 439,858 spectra collected from an identical sample (sample id: TCGA-24-1467). Figure 2.2 shows the comparison of MS/MS search results between the single sample database (187 MB in FASTA) and all the sample unified database (total 4.34 GB in FASTA). As shown in Figure 2.2, the unified database shows a higher number of novel peptide identifications for every FDR estimation strategy, even with much a larger (x20) search space. Moreover, we observed that the overlapping portion of peptide identifications between the unified and sample matched database increases while applying more accurate novel peptide FDR calculations.

**MS-MS search results.** The MS/MS search identified 524 novel peptides and 65,578 known peptides at 1% FDR threshold. (using two-stage FDR strategy B.1(c)) By applying our integrative proteogenomics pipeline [75, 13], 470 novel findings were called (Table 3.1) from 524 identified novel peptides. In assigning proteogenomic events, we removed all peptides that can be mapped to more than 3 genomic locations, and multiply located peptides are used only as supportive evidences of uniquely located peptides to prevent overestimation of our findings [61, 12, 11, 13]. Details in handling multiply located
peptides and event level error control will be discussed in our further study. Peptides identified as ‘novel sequence’ from our pipeline carry mutations which are not part of the wildtype proteome. We analyzed the novel peptide events, and selected examples that showcase the strength of our method. The other results can be found in Supplementary data ‘Supp_1_novel_identifications.xlsx’. (list of known peptide identifications can be found in ‘Supp_2_known_identifications.xlsx’)

**Examples of identified mutated peptides.** As mentioned, a key advantage of our method lies in the capability of identifying combinatorial multiple variants and the possibility of utilizing large scale information from multiple sample data. Note that the CPTAC [2] project which provided the proteomic data, did not include matched tumor-normal controls which could help in identifying somatic versus genetic variants. It is possible to mine the ‘genomic’ data to distinguish between genetic and somatic variants in post-processing analysis that can be performed after the peptide identification. However, in this study we didn’t perform in depth diseases related analysis in order to strictly focus on providing the solid peptide identification results along with benchmark reports while applying different approaches.

Our novel peptides include 13 multiply mutated peptide identifications. Figure 2.3(a) shows a selected example with two substitutions within a single peptide (‘S(F)TFVQAGQDLEENMDED(V)SEK’, spectra count:2). Both substitutions are supported by significant read-depth across multiple samples. Note that both substitutions are reported by dbSNP [58] which also supports the validity of our finding.

The next example in Figure 2.3(b) is a case where we identified a SNV mutated peptide (‘TQTHATL(C)STSAK’, spectra count:2) using distinct sample RNA evidence (selected out of a total of 285 similar substitution events). Interestingly, this mutation was not found within the GATK [44, 17] variant call result the matched sample RNA-seq
but heavily reported by 58 different samples. In order to explore the possibility that genomics alignment or variant calling tools might have filtered out this mutation, we went back and examined the original BAM file of this particular sample. As shown in Figure 2.3(c), we found a RNA-seq read alignment in this region that carries this exact mutation (note that this RNA-seq read also spans known splice junctions of RefSeq gene DPDY). This indicates that GATK filtered out this mutation while processing the single sample RNA-seq file due to the presence of splice junction, low quality score, or insufficient read depth.

Junction peptides are particularly difficult because the span on one end is often too small to make a definitive call. In Figure 2.4, we show an example of multiple junction peptides that confirm a single alternative splice junction event. Two identified peptides ‘SPPDSPT:DALMQLAK’ (spectral count:3) and ‘QNLLQAAAGNVGQASGELLQQIGESDTDPHFQ:ICASR’ (spectral count:1) both indicate alternative splice junctions which share one junction each with the refseq gene ‘TLN1’. Exon in the middle of each peptide also shares the same translation frame indicating a possible novel exon region. Moreover junctions in both peptides had strong RNA-seq level coverage evidence of 22,559 read depth in ‘SPPDSPT:DALMQLAK’ and 17,749 read depth in ‘QNLLQAAAGNVGQASGELLQQIGESDTDPHFQ:ICASR’, across multiple samples.

Figure 2.5(a) is an example of deletion peptide identification (selected from a total of 3 such cases). In this peptide identification, amino acid ’S’ was deleted from the original peptide sequence of ’F(S)SPTLELQGEFSPLQSSLPCDIHLVNLR’ (spectra count:1). This deletion site was expressed across 16 different RNA-seq samples. Together, these examples illustrate the power of proteogenomics searches in confirming translation of DNA lesions.
**Novel peptides identified from outside of general protein coding regions.** The bulk of our novel peptides are mutations on known proteins (275). One of the suggested strategies to reduce false identifications is to limit identification of novel peptides to genes where unmodified peptides have already been discovered. To test the usefulness of a more general method, we investigated peptides not from known gene regions. We do see 60 peptides within immunoglobulin regions (Supplemental Figure B.3) which (because of their high variability) are detected by our pipeline as various types of novel sequences such as alternative/novel splice junctions, fusion genes, and substitutions. A detailed description of proteogenomic event handling and observations from global multiple protein sample results are beyond the scope of this study and will be addressed in our future work. We also find peptides in annotated pseudo-genes (Supplemental Figure B.4), and yet another novel peptide in an unannotated region (Supplemental Figure B.5) which had previously been marked as a gene by a computational tool. Our pipeline automatically identifies many peptides outside of general protein coding regions. However, more detailed study will be necessary to investigate the underlying biology of these regions.

### 2.4 Discussion

Few approaches have been suggested in the community to search large scale genomic data using conventional MS/MS search algorithms. Here, we briefly revisit other approaches and compare with our suggested method. First, a sample specific search can be done to reduce the genomic data size and MS/MS search space. To apply this, one can generate a curated individual database for RNA-seq obtained from each sample [39, 38, 73], and search it against proteomic data from the same sample. This targeted database approach has advantages in increasing the peptide identifications by assigning lower FDR threshold due to reduced database search space. However, as shown
in Table 2.2, we claim that multiple sample driven database can improve the search results by incorporating shared information from many samples. Moreover, it requires coordination between genomic and proteomic laboratories to ensure that both sets of data are created for the same individual. Finally, each sample will use different FDR error control strategy which results in having multiple FDR thresholds throughout the whole process.

A second alternative approach is to assemble the transcript and DNA evidence into compact isoforms and variant calls using available tools [66, 33, 32, 44, 17], and then search translated versions of the isoforms and variants with proteomic data. The advantage of this approach is that it doesn’t require a genomic reference [21] and is very easy to implement. However, the assembly of transcripts from multiple samples is unsolved, and the confidence with which we can assemble and compact all RNA data is limited. There is no study to date which has integrated a large multi-sample RNA-seq data into a single compacted transcript database, and used it for proteomic searches.

In comparison, our method implements an approach that merges and compress large-scale RNA-seq data (all) into a single database by applying graph-based algorithm. We achieve the large scale incorporation of RNA-seq data (from multiple samples) with no loss of information [75], while maintaining the reasonable database size for conventional MS/MS search engines. This unified database approach also enables us to maintain a single and more conservative FDR threshold throughout the entire analysis. Moreover, a graph based approach enables us to efficiently encode cumulative/large information from multiple RNA-seq data-sets into a compact unified database. Our approach also enables peptide identifications with combinatorial multiple splice junctions or variants. Based on our proposed method, we released a JAVA and python based tool called SpliceDB [4] which generates FASTA formatted splice graph database from multiple RNA-seq alignments.
FDR based error control strategies in large database search. Due to the larger database search space, major criticisms of large-scale proteogenomics studies have been focused on the possibilities of false positive novel peptide identifications included in the result, which naturally emphasizes the need of stringent FDR estimation strategy [68]. We agree that different approaches can dramatically change the number of novel peptide identifications. In this study we have shown that combined FDR strategy might boost the total novel identification results by introducing a global bias towards high-scored known peptide PSMs. Even the separate FDR approach still bears the possibility of identifying homologous peptides as novel sequences. Therefore, without introducing any unverified-novel FDR calculations, we applied a simple multi-stage approach where it can remove the FDR calculation bias in combined approach and additionally remove ambiguities in high scored false matches. As expected, two-stage FDR approach shows most conservative FDR measure (Figure 2.1) in novel identifications compared to other approaches. Therefore, we claim that two-stage FDR approach is most suitable for large database searches.

Further study. In this study, we have shown that multiple RNA-seq data-sets can be efficiently incorporated and utilized by the unified MS/MS search database. However, due to the limited MS/MS data access (obtained from a single sample), further integrative analysis is not shown in this study. Note that the results shown in this study are based on peptide level identifications without introducing advanced proteogenomic events. Detailed strategies of proteogenomic events to handle multiple genomic locations properly and advanced event level analysis will be shown in our future work.

Interestingly, novel sequence identifications shown in this study already includes results from highly variable regions such as Immunoglobulin genes and other novel sequence identifications (such as translated pseudo genes, transcript genes, novel gene
areas, and so on) where known peptide identifications may not exist near by. Approaches on properly handling these identification in proteogenomic event level will be discussed in our future work.

While our approach focus mostly on ‘novel sequence identifications’ and follow conventional/standard approaches in ‘known sequence identifications’, other proteogenomic approaches [39, 73, 74, 38] focus heavily on improving ‘known sequence identifications’. Since we used relatively richer set of known proetins (Ensembl), we reason that by utilizing genomic level information to create a targeted and curated known protein database following other studies [39, 73, 74, 38] will improve our known identification results.

In conclusion, we reason that applying different philosophies in proteomic database creation will show various trade-offs. As we can see in the test experiment of this study (Figure 2.2), we claim that incorporation of large transcriptome data can increase the chance of novel peptide identification.
ACKNOWLEDGEMENTS

This chapter, in full, is a reprint of the material as it appears in Journal of Proteomic Research. S. Woo, S. Cha, S. Na, C. Guest, T. Liu, R. Smith, K. Rodland, S. Payne, and V. Bafna, 2013. The dissertation author was the primary investigator and author of this paper.
Table 2.1. Statistics of created cancer databases

<table>
<thead>
<tr>
<th></th>
<th>Single OV sample</th>
<th>OV(PNNL)</th>
<th>OV(splice)</th>
<th>BRCA(splice)</th>
</tr>
</thead>
<tbody>
<tr>
<td># samples</td>
<td>1</td>
<td>67</td>
<td>228</td>
<td>484</td>
</tr>
<tr>
<td>BAM size</td>
<td>7.2 GB</td>
<td>750 GB</td>
<td>2.0 TB</td>
<td>3.2 TB</td>
</tr>
<tr>
<td>FASTA size</td>
<td>187 MB</td>
<td>607 MB</td>
<td>395 MB</td>
<td>814 MB</td>
</tr>
<tr>
<td>Novel splice</td>
<td>168,289</td>
<td>321,587</td>
<td>498,233</td>
<td>646,629</td>
</tr>
<tr>
<td>Deletions</td>
<td>3150</td>
<td>38,464</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Insertions</td>
<td>62</td>
<td>1,105</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>substitutions</td>
<td>7109</td>
<td>182,302</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2.2. List of novel findings (alternative splice junctions indicate novel junctions that shares identical splice site with a RefSeq gene in one side.)

<table>
<thead>
<tr>
<th>Type of novel findings</th>
<th># of novel findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substitution</td>
<td>236</td>
</tr>
<tr>
<td>Deletion</td>
<td>5</td>
</tr>
<tr>
<td>Novel splice junctions</td>
<td>90</td>
</tr>
<tr>
<td>Alternative splice junctions</td>
<td>74</td>
</tr>
<tr>
<td>Novel gene</td>
<td>49</td>
</tr>
<tr>
<td>TranslatedUTR</td>
<td>2</td>
</tr>
<tr>
<td>Exon boundary</td>
<td>4</td>
</tr>
<tr>
<td>Novel exon</td>
<td>6</td>
</tr>
<tr>
<td>Reverse strand</td>
<td>4</td>
</tr>
</tbody>
</table>
Figure 2.1. Number of peptide identifications in 439,858 spectra collected from a single sample (sample id: TCGA-24-1467) using different FDR based error control strategies.
Figure 2.2. Overlap between novel identifications from unified and single sample database.
(a) Identified spectra with 2 substitutions

GATK didn’t report this mutation using single sample RNA-seq (TCGA-24-1467, matched with spectra). However, it is reported in other 58 different TCGA sample GATK results with total 1101 RNA-seq read depth.

(b) Identified spectra with 1 mutation

(c) UCSC Genome Browser plot of identified SNV mutated peptide

**Figure 2.3.** Alignment of identified spectra of mutated peptides.
(a) Identified spectra with alternative splice junctions: ':' is inserted within the peptide sequence to indicate where the splice junction appears.

(b) UCSC Genome Browser plot of alternative splice junctions.

**Figure 2.4.** Alignment of identified spectra of novel junction peptides.
Figure 2.5. Alignment of identified spectra of mutated peptides.

(a) Identified spectra with deletion

(b) UCSC Genome Browser plot of identified deleted peptide

Figure 2.6. Insertions and substitutions are represented as additional node and edges having negative coordinate values. Deletions are represented same as splice junctions with actual DNA coordinates.
Chapter 3

Integrative proteogenomic pipeline for identification of mutated peptides and immunoglobulin gene rearrangements, and its application to colon cancer

3.1 Introduction

Cancer is marked by a progression of somatically acquired genomic lesions. Recent availability of advanced genomic technologies has led to deep insights into the molecular basis of the disease and a better understanding of the mutations that drive the progression of these diseases [36, 8, 47]. The impact of mutations at the protein level, however, is not as well understood.

To close this gap in understanding, recent studies, including recent publications from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) [2], are focusing on analyzing cancer tissue using proteomic (mainly mass spectrometry based) technologies and workflows, with large-scale direct comparisons between transcript and proteomic expression patterns [78]. The results confirm large differences between protein and transcript expression and underscore the need for robust proteomic technologies, particularly in the identification of ‘variant’ peptides as translational evidence for genomic events.
such as mutations, splicing, structural variation, and others. As peptides are typically identified by comparing acquired spectra against theoretical spectra from candidate peptides, a customized database of candidate peptides must be created to include variants observed in genomic tumor samples. The term proteogenomics often refers to the search of mass spectra against these specialized databases [75, 76, 13, 11].

Despite many proteogenomic methods having been recently proposed [38, 39, 73, 74, 23], serious methodological challenges remain. While the initial goal of the CPTAC [2] colon cancer study has been delivered [78], the use of more sophisticated approaches can enable additional discoveries from this existing cancer dataset. Most proteogenomic methodologies focus on identifying single amino-acid polymorphisms (SAP) by adding peptides that capture the alternative allele [78, 38, 39, 73, 74, 23]. However, a large portion of mutational variants, such as insertions, deletions, substitutions, fusion genes, and immunoglobulin genes, are not captured systematically by such an approach. In some cases, transcript evidence is used as a means of reducing the reference database size, while ignoring their potential of identifying novel mutation forms [78]. In other cases, small transcript data-sets are used to mine junction peptides, without a robust framework for handling available big data-sets of Next Generation Sequencing (NGS) data. For colorectal cancer, the single TCGA [47] (The Cancer Genome Atlas) project alone lists more than 1300 RNA-seq data-sets (5.31 TB).

In our approach, we attempt to address the limitations of previous proteogenomic methods namely, computational scaleability, false discovery controls, and novel variant detection. We started by building a comprehensive and compact database that non-redundantly stores variant peptide information through a proteogenomic compaction of multiple RNA-seq datasets. To achieve this compression without loss of sensitivity, we use a graph based approach to model junction and variant peptides. From this representation, we derive a compact linear database [75, 76]. This approach results in
a considerable reduction in database size; from 348 GB of RNA-seq alignments, to a compact proteomic database of 888 MB.

In addition to reducing database size, a crucial step is controlling the number of false positive identifications. We demonstrate how the ‘richness’ (defined below) of the database determines the false discovery rate, and extend our own previous approaches [12, 11, 13, 76] to develop a conservative strategy for proteogenomic event handling and multi-stage false discovery control. We observe that the use of improper false discovery rate (FDR) strategies, such as traditional combined methods, leads to the overestimation of novel peptide identifications. These can result in over 47.44% of actual FDR when calculated separately. The proposed multi-stage FDR strategy strictly maintains FDR to the desired rate (1%). Moreover, the proteogenomic event handling method eliminates the multiple counting of identifications of identical mutational variants. This removes ambiguity in reporting novel findings through the downstream proteogenomic analysis. From 2,367 novel peptide identifications, we reported 1,884 proteogenomic events by grouping compatible peptides and utilizing peptides with ambiguous genomic locations only as supporting evidence. These are in addition to the 130,640 known peptides that were also identified.

In addition to improving the identification of proteogenomic events, we also introduce a novel approach to identify rearranged immunoglobulin genes, a task that has been infeasible in proteogenomic studies to date. While the role of T-lymphocytes in tumor immunology is well understood [50, 48], recent reports have also highlighted the role of B-cells, which also aggregate in tumors. Once there, they form germinal centers, undergo class switching, and differentiate into plasma cells [51]; producing multiple antibodies which are part of the proteome extracts. However, they remain unexplored because standard databases are unable to represent the highly divergent sequences induced by the B-cell differentiation. We developed a customized RNA-seq antibody database,
using a combination of mapped RNA-seq reads, and partial assemblies using de Bruijn graphs [52, 77, 14]. These customized databases permit the identifications of tumor antibodies, and explore their potential role in the molecular characterization of colorectal cancers. Surprisingly, our result showed that 56.37% of our novel proteogenomic event identifications were from immunoglobulin gene database search. This result underscores the importance of our proposed immunoglobulin peptide search when analyzing cancer samples, adding a new host-immune dimension to our analysis.

The value of proteomic evidence over transcript, or genomic evidence has been debated, with recent reports supporting the complementary information available from proteomic data. Our proteogenomic pipeline maintains summary level information in the transcript derived databases that allow for seamless querying of the relative frequencies of specific variants in DNA/transcript data. Through the reanalysis of 90 distinct colorectal tumors from the CPTAC project, we also addressed questions regarding recurrent somatic mutations in tumor genomes. As the result, we have identified $2.3 \times$ more variations compared to the initial CPTAC study [78]. Moreover, it should be noted that by applying conventional FDR strategy, $15.3 \times$ additional novel identifications were found which includes $96.25\%$ of RNA expressed mutations from the initial CPTAC colon cancer study [78].

3.2 Results

The CPTAC colorectal cancer data-set. MS/MS spectra of Adenocarcinoma (COAD) and Rectum Adenocarcinoma (READ) were downloaded from CPTAC data portal [2], for a total of 12,827,616 spectra collected from 90 distinct tumor samples.

Proteogenomic database creation for splice junction and mutation search. We acquired RNA-seq data where available for the CPTAC samples, from the TCGA [8, 36]
repository (90 overlapping samples, 151.08 GB of sequence data), and used it to create specialized splice junction databases. We separated junction variants and mutational variants into separate databases. In the case of junction variants, mapped reads were used to identify recurrent junctions and mutations, and specialized FASTA formatted databases were developed encoding all coding region and junction information, while ignoring mutational data to create a compact database (1.43 GB) encoding 1,245,069 novel splice junctions, and 85.29% of all known splicing events. In case of mutational variants, single nucleotide variant (SNV) and short substitution/insertion/deletion information from the RNA-seq alignments (from TCGA project), encoded in VCF files, were also used to construct an 1.14 GB MutationDB FASTA database encoding putative variant peptides. The compact databases, critical to maintaining a low FDR, can be attributed to (a) building a splice-graph to encode junctions in a non-redundant fashion, and (b) creating a specialized FASTA database derived from the splice-graph to enable efficient database search (see online methods, and our previous approaches [75, 76]).

**Extended proteogenomic database for immunoglobulin peptide search.** The DB construction for immunoglobulin genes is more challenging, as the antibodies are the result of genomic recombination, splicing, and non-templated DNA insertion, making it difficult to map them to the standard reference sequence. As illustrated in Figure 3.1, we developed a customized proteogenomic database targeted to Ig gene peptide identifications.

First, in order to select immunoglobulin related RNA-seq reads, we employed a two-step procedure for selecting reads for the IgH locus. For the first pass, we filtered and retained all reads mapping to the IgH locus. The majority of these reads mapped to the constant (C) and variable (V) gene-segments. Additionally, we retained unmapped reads with 10-mers that matched to the V, D, J, or 5’ end of C reference gene-segments,
and matched additional filters (online methods). This set of remaining reads was referred to as the putative IgH read set. While not very stringent, the filtering eliminated most non-IgH originating reads. Further pruning was performed on the de Bruijn graph data structure.

We constructed the de Bruijn graph using k-mers from these reads in the following manner: Nodes in this graph represent all \((k - 1)\)-mers over the putative IgH read set. Nodes \(u, v\) in set \(V\) are connected by a directed edge (arc) \((u, v) \in E\) if \(u\) is a prefix, and \(v\) is a suffix of some \(k\)-mer in a read. This graph \(G = (V, E)\) is termed the repertoire graph, as it is built over the putative IgH read set. Figure C.1(b) displays an example of the de Bruijn graph built on 6-mers from the two sequences shown, while a value of \(k = 21\) is used to construct the repertoire graph. More detailed explanations of de Bruijn graphs as a data-structure for assembly can be found elsewhere [14].

Next, paths in the repertoire graph were converted to a compact FASTA formatted protein database using a specialized algorithm that guarantees sensitivity while keeping the database as compact as possible [75]. The specialized IG gene database derived from the larger corpus of 150Gb RNA-seq reads was only 467Mbp. Table C.1 shows the overall statistics of the database sizes and number of genomic variations encoded in our final proteogenomic database. The complete search also used a database of “known-proteins” from Ensembl [24] (version GRCh37.70).

**MS-MS search results.** A ‘target-decoy’ based FDR strategy is commonly deployed to control the false discovery rate of peptide identifications. The traditional approach to FDR calculation [19] creates a single, combined target database, and a similar-sized reversed (or permuted), decoy database to estimate the false-discovery rate. This leads to a distortion in proteogenomic searches, where the novel variant databases can be very large, but have a smaller fraction of identifications with a higher false positive rate.
To understand the behavior of FDR controls on databases of different sizes, consider a database of a specific size, and richness $\alpha$, where the richness corresponds to the fraction of peptide spectrum matches (PSMs) that are correctly mapped to the peptide. Thus, the value of $\alpha$ is high for known proteins, but low for many of the variant encoding databases. Let $C, I, T, D$ be randomly chosen peptides spectrum match scores from correct, incorrect, target-database, and decoy-database PSMs. These random variables are distributed according to $f_C, f_I, f_T, f_D$ respectively. Further, let $F_C(x) = \int_{u=x}^{\infty} f_C(u)du$ denote the cumulative tail probability. To control the FDR, we would like to identify the minimum threshold $\tau$ such that

$$\frac{F_D(\tau)}{F_T(\tau)} \leq 0.01$$

where 0.01 (1%) is the desired FDR. We assume that $f_D(x) = f_I(x)$ for all $x$, and note that

$$f_T(x) = \alpha f_C(x) + (1 - \alpha) f_I(x)$$

Integrating and substituting, the goal is to find a minimum threshold $\tau$ s.t.

$$\frac{F_D(\tau)}{F_T(\tau)} = \frac{F_I(\tau)}{\alpha \cdot F_C(\tau) + (1 - \alpha) \cdot F_I(\tau)} \leq 0.01$$

Denominator of known-protein DB is larger than that of proteogenomic DB, and vice versa for the numerator. Therefore, if the proteogenomic DB has larger size and poor quality, then the FDR of known-protein DB is always smaller than FDR of proteogenomic DB, so the same cut-off cannot be applied to the different databases (see online methods and data). We employed a conservative, multi-stage strategy [76] with a 1% FDR cut-off at each stage. The databases were searched in a specific order, starting with a ‘known protein’ database first, followed by Ig Database, SpliceDB, MutationDB, and six-frame
in order. Spectra that passed the FDR threshold in an earlier database were not considered for subsequent searches (see online methods). Figure C.3 shows a comparison of the two strategies, where the combined strategy results in more identifications, but with a higher false-discovery rate for the novel (variant) peptides.

The 12,827,616 Adenocarcinoma and Rectum Adenocarcinoma tumor MS/MS spectra were searched against the known protein and specialized proteogenomic database using MSGF+ [35]. The multi-stage search resulted in 130,640 known peptide identification (5,673,517 PSMs) and 1,416 aberrant peptides (14,484 PSMs) at 1% PSM level FDR cut-off. The extended immunoglobulin database search for IG peptides yielded 439 distinct peptides (58,778 PSMs) from the constant region, and 951 peptides (7,091 PSMs) from the variable regions.

**Comparisons with different MS/MS database search approaches.** We have benchmarked our search against previous searches of the same MS data, including Zhang et al. [78] who used their own databases (CanProVar), and against a second search-tool using Comet [20] on our specialized databases as a control. The Comet results [20] showed 357 novel peptide identifications while 70.86% of the peptide overlapped with MSGF+ [35] results (Figure C.4). In generating novel events reported in this paper, we used the union set of both MSGF+ [35] and Comet [20] peptide identification results, adding an additional 104 peptides. In general, our tools are agnostic to the choice of a specific search-tools.

In comparing against CanProVar results (Figure 3.2), we note that in both the multi-stage and combined-search, we predicted an excess of junction peptides and IG peptides. These were ignored by previous approaches due to the challenge in identification. The number of mutations were comparable in both studies, with 276 overlapping mutations. Among the mutated peptides predicted by CanProVar alone, 290 were not
represented in our database as their databases included public sources encoding variation [25, 58, 47], while our customized databases study were created directly from matching sample RNA-seq data-set. The remaining missed identifications were mainly due to FDR controls (211 of 230) and could have been discovered via the ‘combined’ FDR search, but at the cost of a higher false-discovery rate (Figure 3.2b).

**Peptide identifications to proteogenomic events.** Novel variations were grouped by locations and automatically classified into distinct events (see methods). Peptides mapping to two locations were used only to support other events, ensuring that each event had at least one uniquely mapping peptide. Grouped novel peptide identifications with reading frame compatibility are shown in the following section with specific proteogenomic examples.

**Comparisons in protein versus genomic level mutation analysis.** Initial comparisons between the expression and recurrence of variant peptides suggested significant differences [78]. As we did not have matched proteomic data from normal samples, we used an earlier study from TCGA [47] to call somatic variations. The TCGA study paired 224 of 243 tumor samples with matched blood samples, while the MS data had 90 samples that overlapped with TCGA, and 61 also had matched blood. We identified 108 SNV mutations and 1 insertion that were called somatic in the TCGA study, and compared their recurrence with versus genomic mutations.

Figure 3.3(a) shows the top 30 most frequently mutated genes reported by the TCGA study [47]. However, these genes have extremely low protein expression (as measured by spectral counts) even for non-mutated peptides (Figure 3.3(b)). In contrast, the most recurrent proteins with somatic mutations have variable recurrence using RNA expression (Figure 3.3(c)), but the list identifies many genes of interest. However, genes
such as TNC [60], HSPG2 [30], PML [71], GBP-1 [10], TF [57], NES [63], have all been implicated in colorectal tumor angiogenesis.

**Peptide identifications from immunoglobulin rearrangements.** Our search also resulted in a large number of IG peptides, including 439 peptides (58,778 PSMs) mapping to the IG constant region, and 1,094 peptides (8,701 PSMs) IG variable regions (see online methods). Figure C.5 shows an example of peptides supporting specific V(D)J recombinations. The complexity of these peptides suggests that there could be bias in their discovery patterns. To test for bias, we compared the IG peptide spectral counts to RNA-seq data, and observed a strong correlation (Figure C.6(a)). The high correlation extended to spectral counts between heavy and constant regions in each sample (Figure C.6(a); $\rho = 0.77$). Finally, while there is variation in the location of IG constant region peptides, all regions with tryptic digestion sites are well sampled (Figure C.6(c)). As there is no specific bias, we used the data to investigate IG peptide concentrations within cancer-subtypes. As mature antibodies are expressed only in differentiated lymphocytes, the excess of IG peptides is indicative of an immune response mediated by B-lymphocyte infiltration in the tumor cells. While the role of T-lymphocytes in tumor immunology is well understood [50], the role of B-cells is still being elucidated, although some reports suggest that B-cells aggregate in tumors [41, 49], where they form germinal centers, undergo class switching, and differentiate into plasma cells [48, 51].

**Distribution of IG peptides across colorectal subtypes.** The CPTAC study classified the 90 samples into five subtypes, marked ‘A’ through ‘E’ [78] based on expression patterns. Figure 3.4 shows the plot of IG gene peptide spectra counts (normalized by the total number of known spectra identifications in each group) between each sample subtypes. In addition, MS/MS data-set from normal colon/rectal [34] tissue, and colon
cell-line [22] were used as controls. We observed that IG peptide identification rate in all sub-types were similar to the normal sample compared to the majority of cancer samples (except samples within group ‘C’) while cell-line sample showed a markedly smaller number of IG peptide identifications. The one exception was the significant over-expression of IG peptides in subtype ‘C’ ($p$-value $< 0.0001, \chi^2 = 2927.71$), comprising of samples that are hypermutated, and microsatellite instability (MSI) high. Moreover, samples in subtype ‘C’ also show high overlap with the both ‘stem-like’ and ‘colon cancer subtype3’ group defined by the Sadanandam et al. [56] and De et al. [15]. Our results suggest that a strong immune response could be a molecular marker of CRC sub-types.

We also tested the distribution of somatic mutations across samples sub-types (Figure C.7) and observed slightly higher frequency of somatic mutation identifications in subtype ‘B’ ($p$-value $< 0.0001, \chi^2 = 40.39$). In the initial TCGA [47] and CPTAC [78] colon cancer study, samples in both ‘B’ and ‘C’ subtypes are reported as hypermutated while group ‘C’ is characterized as showing both MSI-high within hypermutated samples. Our results support this partitioning based on differential distribution of variant peptides in the two sub-types.

**Identifying mutated peptides for follow-up.** The TCGA transcript analysis largely identified somatic mutations with low recurrence except for a few key genes. Moreover, the recurrently mutated genes (e.g., APC) are tumor suppressors, and had reduced protein expression; the mutations are therefore not seen in the proteome. Thus, we focused here on identifying SNV mutations and other events that were not highly recurrent, but together, could be part of targeted proteomic studies characterizing colorectal cancer sub-types.

Our study revealed 679 identified substitutions, of which 108 SNV mutation
overlapped with the TCGA reported somatic mutations in colon cancer samples and 424 SNV mutations are reported in dbSNP.

**Exemplars of Somatic SNV mutations.** The tumor suppressor *SMAD4* mediates the TGFbeta signaling pathway suppressing epithelial cell growth, and inactivation of the Smad4 gene through an intragenic mutation occurs frequently in association with malignant progression [46, 42]. We identified a single PSM ‘VPSSCPIVTVGDVGTVDPSSGD; H;FCLGQLSNVHR’ (*R361H*, Figure C.8) supporting a known mutation in colorectal cancer [25], appearing with low frequency in transcript data (7 of 243 TCGA samples).

The wildtype *KRAS* gene is required for drug efficacy in metastatic colorectal cancer [7]. We identified a known, low-frequency, mutated peptide, ‘LVVVGAG:D:VGK’ (*G12D*, Figure C.9) in 4 of 90 proteome samples, matching the low transcript frequency (25 of 243 transcript samples).

Expression of the polymeric immunoglobulin receptor (pIgR), a transporter of polymeric IgA and IgM, is commonly increased in response to viral or bacterial infections, linking innate and adaptive immunity. Abnormal expression of pIgR in cancer was also observed [5]. We identified a mutation (Figure C.11) with strong overlapping peptide identification. We also identified overlapping peptides in FGA gene, which has been proposed as a marker for other cancers [62].

**Alternative splice junctions.** We categorized the identified splice junctions as ‘novel’ when both splice sites does not overlap with any known splice junctions, while ‘alternative’ junctions indicate that at least one splice site is shared with any existing known junctions. We identified 97 novel splice junctions and 11 alternative splice junctions. Figure C.12 shows an example of alternative splice junction peptide ‘VKEENPE:G:PPNANE DYR’ in STK39 (a cellular stress response pathway gene [54]) along with its spectral
alignment.

**Deletions.** Figure C.13 shows an example of mutated peptide identified with the presence of deletion (from 4 deleted peptides in total) in the Ladinin-1 gene across 6 samples. A related SNV mutation of the peptide ‘K.NLPSLA:E:QGASDPPTVASR.L’ (K−→E) was also reported by TCGA [47] colon cancer somatic mutation calls with 10,711 read depth.

**Fusion genes.** Figure C.14 shows a possible gene fusion region (selected from total 8 possible gene fusion peptide identifications) where two junctional peptides are identified across two different genes (HBA1 and HBA2). Two fusion peptide shown in this region had unique genomic location and total 15 spectra counts from two protein samples. These hemoglobin related genes act as anti-oxidants, attenuating oxidative stress-induced damage in cervical cancer cells [40].

### 3.3 Discussion

We present here a systematic pipeline for identifying mutated peptides in cancer focusing on many challenging issues such as a compact, integrated transcript derived database for searching, FDR controls, and event calling. In addition, we also developed customized databases for searching for IG peptides allowing us to quantify the antibody response to cancer.

Our results follow other results in suggesting a significant gap between genomic versus protein level mutation identifications, mediated on the fact that recurrent mutations in transcripts may not be observed in the proteome due to reduced protein expression of the mutated gene. Thus, the development of protein based biomarkers must be prefaced by proteome related studies. The mutations observed during transcription and
translation have different characteristics. However, a pipeline such as ours which searches a comprehensive database of transcript derived mutations against spectra allows for a joint exploration of the proteogenomic space.

The significant number of peptide identifications in immunoglobulin regions point to active immunoglobulin responses within certain sub-types of cancer, and provide a new direction towards molecular sub-typing of cancer. Implicitly, our results can also lead to an analysis of antibody sub-type switching, and predicting the host response to infections. These will be fully investigated in future studies.

Finally, the proteogenomic analysis leads to the identification of a number of aberration peptide identifications that will serve as candidates for targeted studies of tumor subtyping and tumor progression.
ACKNOWLEDGEMENTS

This chapter, in part is currently being prepared for submission for publication of the material. S. Woo., S. Cha, S. Bonisso, S. Na, D. Tabb, P. Pevzner, V. Bafna. The dissertation author was the primary investigator and author of this material.
Table 3.1. Characterization of aberrant events. 61 samples out of 90 had blood (normal) samples available as a matched reference.

<table>
<thead>
<tr>
<th>Type of novel findings</th>
<th># of novel findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Somatic substitution</td>
<td>106</td>
</tr>
<tr>
<td>Germline substitution</td>
<td>298</td>
</tr>
<tr>
<td>Uncategorized substitution</td>
<td>246</td>
</tr>
<tr>
<td>Somatic insertion</td>
<td>1</td>
</tr>
<tr>
<td>Uncategorized insertion</td>
<td>3</td>
</tr>
<tr>
<td>Deletion</td>
<td>4</td>
</tr>
<tr>
<td>Transcript gene</td>
<td>10</td>
</tr>
<tr>
<td>Fusion gene</td>
<td>5</td>
</tr>
<tr>
<td>Translated-UTR</td>
<td>16</td>
</tr>
<tr>
<td>Alternative splice</td>
<td>11</td>
</tr>
<tr>
<td>Novel splice</td>
<td>91</td>
</tr>
<tr>
<td>Exon boundary</td>
<td>6</td>
</tr>
<tr>
<td>Frame shift</td>
<td>4</td>
</tr>
<tr>
<td>Novel exon</td>
<td>2</td>
</tr>
<tr>
<td>Novel gene</td>
<td>4</td>
</tr>
<tr>
<td>Reverse strand</td>
<td>1</td>
</tr>
<tr>
<td>Pseudo gene</td>
<td>14</td>
</tr>
<tr>
<td>IG gene variable region</td>
<td>1,062</td>
</tr>
</tbody>
</table>
Figure 3.1. Illustration of proteogenomic database construction for IG peptide identifications.
Figure 3.2. (a) Comparison of aberrant peptide identifications against previous findings using multi-stage FDR (b) Comparison of overlapping aberrant peptide identifications using combined FDR.
Figure 3.3. (a) Genes containing most frequent somatic mutations reported by the TCGA study. (b) Most frequently mutated genes in DNA level are under expressed in protein level. (c)Percentage of samples containing identified protein mutations.
**Figure 3.4.** Percentage of IG gene peptide identifications in each sample normalized by the number of known peptide identifications across sample subtypes.
Appendix A

Proteogenomic database construction driven from large scale RNA-seq data
Figure A.1. In filtering stage, RNA-seq reads that have identical splice junctions are merged, and extended in both ends.

Table A.1. Overall statistics of splice graph data structure

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of components $(G)$</td>
<td>116355</td>
</tr>
<tr>
<td>Number of nodes</td>
<td>652936</td>
</tr>
<tr>
<td>Number of edges</td>
<td>337648</td>
</tr>
<tr>
<td>Average node length</td>
<td>57.00 bp</td>
</tr>
<tr>
<td>Average number of edges per node</td>
<td>0.44</td>
</tr>
</tbody>
</table>

A.1 Comparison with other gene prediction methods

We compared the list of identified novel peptides using our proteogenomics pipeline against various gene prediction results provided from other groups using identical RNA-seq data-sets. The file of gene predictions is available [45], and includes GeneFinder [45], single exon gene predictions, predictions based on RNA-seq, and predictions from conserved ORFs (against *C. briggsae*). A total of 688 novel peptides were matched to 1194 different predicted gene sequences (Table A.2).
Figure A.2. Combining pairs of sequences that share a prefix and suffix string. (a) If $\ell < L$, the generated sequences cannot share an identical prefix and suffix. (b) If $l \geq 2L$, the prefix and suffix of generated sequences will not overlap.

Table A.2. Number of overlapping sequences between identified novel peptides using our proteogenomics pipeline versus protein sequences generated from other gene prediction methods.

<table>
<thead>
<tr>
<th>Prediction methods [45]</th>
<th># of overlapped sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>GeneFinder</td>
<td>286</td>
</tr>
<tr>
<td>Conserved C. briggsae ORFs</td>
<td>364</td>
</tr>
<tr>
<td>RNA-seq data</td>
<td>543</td>
</tr>
<tr>
<td>Single exon gene predictions</td>
<td>1</td>
</tr>
</tbody>
</table>

A.2 Calculation of split mapped coordinates from CIGAR string in SAM file format

CIGAR string of the SAM format file is used to determine splice junctions. For example, consider a match, starting at coordinate $x$, with the accompanying CIGAR string given by ‘35M1000N35M’ which is translated to “match 35bp”, “skip 1000bp”, followed by “match 35bp”. We convert this string to two GFF lines, denoting the intervals $[x, x + 35]$ and $[x + 1035, x + 1070]$. Here, $[x + 35, x + 1035]$ represents a junction.
Figure A.3. Illustration of hashing. (a) In first visited node path, generated sequence is the full path. (b) In traversing, we store all the split nodes present in a candidate list. (c) When a merge node is encountered, we traverse the subsequent path, querying the hash table.
Figure A.4. Description of parameter $W$.

Figure A.5. Alignment result of novel gene example. The highlighted region corresponds to the alignment of identified peptide against the sequence of hypothetical protein CRE_09558 [Caenorhabditis remanei].
**Figure A.6.** Translated UTR spectral counts throughout different developmental stages

**Figure A.7.** RNA vs peptide transcription level
A.3 Detailed RNA-seq methods

RNA-seq Alignment methods:

Step 1:

1. identify all reads beginning with at least 4 TTs
2. identify all reads that begin with at least 6 bases of SL on the front
3. identify adaptor sequence on the 5’ and 3’ ends of the reads
4. align the reads against the WS220 genome using cross-match
5. align the reads against the AG1003 aggregate genelet transcriptome (transformed into WS220 coordinates) using cross-match

From combining the information from the output of these steps, if \( z \leq 5 \) bases on either end of the read are unclassified/unaligned, then the read is considered to be mapped.

The non-C. elegans worms, C. briggsae and C. remanei, were searched against the WS225 database for each of those genomes. C. japonica and C. brenneri, were searched against the WS227 database.

Step 2:

Next, all reads that had at least 30 bases of match to the genome but were not yet successfully placed, are aligned to the WS220 genome using splice-aware cross-match. Those results are integrated with the alignments to step 1 to again decide which reads are now considered to be fully mapped.

Step 3:

Reads still not placed but with at least 30 bases of match in the genome, are aligned against a splice junction database using cross-match. The splice junction database
contains all confirmed and predicted splice junctions (wormbase and RNAseq etc.) plus all possible novel combinations of those junctions (within 4kb of one another) with 75 bases appended on either side of the junction. Combine these alignments with the information from step 1 to determine if the read is fully accounted for (\(\leq 5\) bases unaligned from either end).

Step 4:
For reads still not placed, look for multi-segment alignments from bwasm that suggest multiple as yet unpredicted exon pairs and identify splice junctions to join those multi-segment alignments. Combine these alignments with the information from step 1 to determine if the read is fully accounted for (\(\leq 5\) bases unaligned from either end).

A.4 Comparison of spectra dataset used in this study with Merrihew et al. (2008) [45]

The Merrihew et al. (2008) [45] paper used different fractionation methods, samples and data analysis than the current submission to discover novelty.

The 2008 paper used biochemical fractionation of all stages of C. elegans to improve identifications while the new dataset uses molecular weight fractionation of different stages of C. elegans. The molecular weight fractionation gives us information about the protein before digestion and giving us the potential to map the peptides back to different isoforms. Sampling the different stages of C. elegans improves the identifications and provides information about when proteins are expressed.

The 2008 data analysis relied on a search database made up of the following components: Wormbase (version WS150) protein-coding genes, less conservative predictions from a version of Genefinder, and intergenic ORFs from Wormbase (version WS130) greater than 30 codons with homology between C. elegans and C. briggsae. All
of the above components are outdated and most, if not all of the novel findings from the 2008 paper have been confirmed by other experimental methods provided by the modENCODE project. Wormbase is currently on version WS236 which incorporates all these modENCODE findings. This paper describes a different method for assessing novelty using a non-redundant compact database of information from RNA-seq reads to identify novel events in mass spectrometry data. The 2008 paper used some RNA-seq data from the Green lab but it only used the data to confirm the novel events found based on our database search. Also when the 2008 paper was written the Green lab had only RNA-seq data for part of the C. elegans genome.

Additionally our chromatography conditions and mass spectrometers have improved tremendously since 2008. The new data was collected using nano-flow liquid chromatography and using a mass spectrometer with higher resolution, increased sensitivity and faster scanning. The 2008 data used standard flow liquid chromatography and a standard mass spectrometer.
A.5 Proof of correctness and completeness in applying Rule1, Rule2, and Rule3

We use three rules to eliminate shared sub-paths.

1. For a pair of paths, $xz$ and $yz$ with a shared string $z$, we generate two FASTA strings $xz$, and $y \cdot \text{pref}_L(z)$, where $\text{pref}_L(z)$ denotes a length $L - 1$ prefix of string $z$.

2. For a pair of paths, $xz$ and $xy$ with a shared prefix $x$, we generate two FASTA strings $xz$, and $\text{suff}_L(x) \cdot y$, where $\text{suff}_L(x)$ denotes a length $L - 1$ suffix of string $x$.

3. For paths $xy$ and $yz$, which have a prefix-suffix match with $y \geq L$, generate the FASTA string $xyz$.

Claim: Applying rules 1, 2, and 3, doesn’t violate completeness and correctness

Proof: Let $G$ be a splice graph with nodes and edges. Each node represents exons containing sequence of nucleotides, and edges represents the possible event of splicing.

First we’ll define the followings,

- $S$ is a set of every sequences from graph $G$.
- $S_1$ is a set of sequences from $S$ with applying rule 1.
- $S_2$ is a set of sequences from $S_1$ with applying rule 2.
- $S_3$ is a set of sequences from $S_2$ with applying rule 3.
- $S_\alpha(l)$ is a set of length $l$ sequences from $S_\alpha$.

It is clear that $S(l)$ contains every length $l$ sequences that can be generated from $G$, and doesn’t have any sequences that cannot be generated from $G$. So, the claim will be satisfied if $S(l) = S_3(l)$
To show \( S(l) = S_3(l) \), we need to show \( S(l) \supset S_1(l) \supset S_2(l) \supset S_3(l) \) and \( S(l) \subset S_1(l) \subset S_2(l) \subset S_3(l) \)

Because of rule 1, 2 eliminating the sequence or subsequence of elements in \( S_1 \) and \( S_2 \), \( S(l) \supset S_1(l) \supset S_2(l) \) is clear. Also, rule 3 may produce extra length \( l \) path during the combining procedure, \( S_2(l) \subset S_3(l) \) is also clear.

What we need to show are the following,

1. \( S(l) \subset S_1(l) \)
2. \( S_1(l) \subset S_2(l) \)
3. \( S_3(l) \subset S_2(l) \)

1. \( S(l) \subset S_1(l) \)

It is clear that \( x \) is an element of \( S_1(l) \) if \( S_1 \) has at least one element \( y \) which has \( x \) as a subsequence. So, it is enough to show that for \( \forall x \in S(l) \), \( \exists y \) such that \( y \in S_1 \) and \( x \) is a subsequence of \( y \). Recall that \( S_1 \) is generated from \( S \) with applying rule 1 which preserve one sequence for the shared suffix. This means that rule 1 eliminates the suffix only when there is at least one sequence which contains the same suffix. Hence \( y \) exists in \( S_1 \).

2. \( S_1(l) \subset S_2(l) \)

Recall that every element in \( S_1 \) has distinct length \( l \) suffix and \( S_2 \) is generated from \( S_1 \) with rule 2. So \( |S_2| = |S_1| \). Set a bijection between \( S_2 \) and \( S_1 \) which has same length \( l \) suffix. Then the only difference between these sets is part of the prefix that is eliminated by rule 2. But rule 2 also never eliminates the prefix until one of the other elements in \( S_2 \) has same prefix. Therefore every length \( l \) sequence in \( S_1(l) \) is also in \( S_2(l) \).
3. $S_3(l) \subset S_2(l)$

Rule 3 is applied only when two elements in $S_2$ shares the node as their suffix and prefix. So, newly generated length $l$ sequences have their corresponding location in graph $G$. By 1 and 2, $S_2(l)$ have every length $l$ sequence in $G$. So there is no element $x$ such that $x \in S_3(l), x \not\in S_2(l)$.
A.6 Proof of correctness and completeness in DFS algorithm implementation of Rule1, Rule2, and Rule3

Claim: Our algorithm doesn’t violate the completeness and correctness.

Proof: From above, we have shown that application of Rule 1, 2, and 3 doesn’t violate the constraints. Here, we want to show our algorithm correctly apply rule 1, 2, and 3.

We begin with the two functions used in our implementation which follows conventional DFS algorithm. DFS: An algorithm same as DFS, but stop when it enumerate visited node

DFSFiniteLength : An algorithm same sa DFS, but stop when it enumerate length $L$ further from desired node.

A.6.1 Rule 1:

For Rule 1, we will show this in a two step.

Step 1: For a certain merge node $n$, if every input edges are retrieved, then keep search DFS for one edge and call DFSFiniteLength for others. We want to show this is same as application of Rule 1 for the sequence set generated from $n$ by DFS.

Step 2: For every merge node in $G$, DFS retrieve every input edges of all merge nodes.

If Step 1 and Step 2 are true, DFS and DFSFiniteLength correctly apply rule 1 for every merge nodes. Step 1 and Step 2 are shown as the following.

1. Assume we have merge node $n$ and multiple input edges $e_1, e_2, \ldots, e_k$.

Define:

- path $p_i$: a path that passes through $e_i$ and stops at node $n$. 
• set $E_i$: set of paths such that all elements in the set have $p_i$ as a prefix and are a result of enumeration from $n$ by DFS.

$|E_i|$ are all identical because they are generated at same node $n$ by DFS. Therefore, we can set a bijection for any element from $E_i$ to $E_j$ which shares the suffix. Keep $E_1$, and for $E_2, \ldots, E_k$, eliminate the suffix of all elements except $(L - 1)$ from corresponding sequence of $n$.

This is the same as application of Rule 1 for same merge node $n$.

2. Define: DFS´ is an algorithm the same as DFS, but stopping when it enumerates a visited node.

If we color the edges enumerated by DFS´, then we can easily see that all edges in $G$ will be colored. This means that every merge node in $G$ will be visited by DFS from all its incoming edges.

By 1. and 2., Rule 1 is implemented except for some path set that shares the suffix but does not share the path.

A.6.2 Rule 2:

Whenever DFS´ visit splitting node during the enumeration, every sequence set from that node will share the prefix. Therefore keep one and eliminate the prefix except $(L - 1)$ from the node is same as application if Rule 2 for that node.

DFS´and DFSFiniteLength apply this for every splitting node, hence Rule 2 is implemented for all splitting nodes.

A.6.3 Rule 3:

Candidate pair set contains every possible prefix-suffix ‘coordinate’ overlap which shares the sequence, so combining prefix-suffix pair in candidate pair set implements Rule 3.
Note that our algorithm applies Rule 1, Rule 2, and Rule 3 only when sequences have ‘coordinate’ overlap (having only sequence level overlap does not satisfy this condition, and having coordinate overlap guarantees that the sequences have sequence level overlap within the same reference DNA system). This is different than Edwards and Lippert [18] where they merge all overlapping sequences (less than a certain parameter), and do not consider coordinate information. Again, unlike Edwards and Lippert [18], our method uses a genomic coordinate-based data structure (represented in base pairs) rather than minimizing the amino acid sequence overlap. We claim that for proteogenomic analysis the coordinate based approach is more appropriate since it can easily reconstruct the original genomic coordinate of the identified peptide.
Appendix B

Proteogenomic strategies for identification of aberrant cancer peptides using large-scale Next Generation Sequencing data
Figure B.1. Diagram describing different FDR based error control strategies applied in this study.
Figure B.2. Structure of hash table for accessing the original RNA-seq meta information

Figure B.3. UCSC genome browser plot of our novel peptide identifications within complex immunoglobulin region rearrangements included in our peptide identification result.
Figure B.4. UCSC genome browser plot of peptide identification in pseudo gene area.

Figure B.5. UCSC genome browser plot of peptide identification in a possible novel gene area where a gene prediction method also reported as a possible gene.
Appendix C

Integrative proteogenomic pipeline for identification of mutated peptides and immunoglobulin gene rearrangements, and its application to colon cancer

Table C.1. Database statistics. Total 90 RNA-seq BAM files which matches with the tumor samples used in the study of Zhang et al. [78] were used in creating proteogenomic database.

<table>
<thead>
<tr>
<th>DB attributes</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA-seq input files</td>
<td># of samples</td>
</tr>
<tr>
<td></td>
<td>BAM size</td>
</tr>
<tr>
<td>Protein DB</td>
<td>MutationDB FASTA size</td>
</tr>
<tr>
<td></td>
<td>IG DB FASTA size</td>
</tr>
<tr>
<td></td>
<td>Total AA searched</td>
</tr>
<tr>
<td># of mutations encoded in DB</td>
<td>Novel splice</td>
</tr>
<tr>
<td></td>
<td>Deletions</td>
</tr>
<tr>
<td></td>
<td>Insertions</td>
</tr>
<tr>
<td></td>
<td>Substitutions</td>
</tr>
</tbody>
</table>
Figure C.1. (a) Potentially missed reads from a somatically recombined heavy chain transcript as greyed out, while mapping reads as darker. (b) Example de Bruijn graph showing how differences in sequence manifest as differences in topology.

Figure C.2. Multistage-FDR strategy.
Figure C.3. (a) Number of known peptide identifications obtained by applying combined FDR versus multi-stage FDR. (b) Number of novel peptide identifications applying combined versus multi-stage FDR.
Figure C.4. Comparison between results obtained using MSGF+ and Comet MS/MS search tools.
Table C.2. Statistics of identified novel events using combined FDR 1% cut-off. This statistics was generated by applying conventional combined FDR strategy. We obtained large number of novel peptide identifications compared to the result from multi-stage FDR strategy.

<table>
<thead>
<tr>
<th>Type of novel findings</th>
<th># of novel findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substitutions</td>
<td>2,090</td>
</tr>
<tr>
<td>Insertion</td>
<td>9</td>
</tr>
<tr>
<td>Deletion</td>
<td>7</td>
</tr>
<tr>
<td>IG gene</td>
<td>428</td>
</tr>
<tr>
<td>Transcript gene</td>
<td>587</td>
</tr>
<tr>
<td>Fusiton gene</td>
<td>99</td>
</tr>
<tr>
<td>TranslatedUTR</td>
<td>376</td>
</tr>
<tr>
<td>Alternative splice</td>
<td>239</td>
</tr>
<tr>
<td>Novel splice</td>
<td>2,355</td>
</tr>
<tr>
<td>Exon boundary</td>
<td>76</td>
</tr>
<tr>
<td>Frame shift</td>
<td>1,516</td>
</tr>
<tr>
<td>Novel exon</td>
<td>523</td>
</tr>
<tr>
<td>Novel gene</td>
<td>733</td>
</tr>
<tr>
<td>Reverse strand</td>
<td>1,578</td>
</tr>
<tr>
<td>Pseudo gene</td>
<td>197</td>
</tr>
</tbody>
</table>
Figure C.5. (a) Example of peptide identifications in immunoglobulin rearrangements. (b) Diagram illustrating the peptide identifications of V(D)J recombination junctions.
Figure C.6. (a) Plot of RNA-seq read counts from IG variable region versus IG constant region. (b) Spectra counts of peptide identifications from IG constant versus variable region. (c) Plot of spectra counts covering IgG constant region.
Figure C.7. Percentage of peptide identifications with somatic mutations in each sample normalized by the number of known peptide identifications across sample subtypes.
Figure C.8. Identification of somatic mutation in gene SMAD4. This mutation had 1 spectra count with unique genomic location and 15 RNA-seq read depth. This mutation is also reported as somatic mutation in TCGA study [47], and COSMIC [25].
(a) Alignment of identified spectra with somatic mutation in gene KRAS

(b) UCSC Genome Browser plot of identified somatic mutation in gene KRAS.

**Figure C.9.** Identification of somatic mutation in gene KRAS. TCGA colon cancer study [47] reported this mutation as ‘somatic’ in 25 different colon cancer samples and also reported by COSMIC [25] and dbSNP [58].
(a) Three identified spectra alignments indicating an identical SNV mutation in gene FGA.

(b) UCSC Genome Browser plot of identified somatic mutation in gene FGA.

**Figure C.10.** Identification of somatic mutation in gene FGA. 3 overlapping peptide sequences had total 4 spectra counts and unique genomic locations. This SNV location is reported by both COSMIC [25] and dbSNP [58].
Figure C.11. Identification of somatic mutation in gene PIGR. Total spectra count of both peptide was 137 and RNA-seq read depth of this mutation was 11005. Matching mutation of this region were found in both COSMIC [25] and dbSNP [58].
Figure C.12. Identified alternative splice junction peptide. This peptide had 11 spectra counts (with unique genomic location) and total 386 RNA-seq reads were mapped to this alternative splice junction.
(a) Identified spectra with deletion

(b) UCSC Genome Browser plot of identified deletion.

(c) Two different adjacent SNV mutations in LAD1 gene located within the same exon with identified deleted peptide.

**Figure C.13.** Identified deletion and two neighboring SNP mutated peptides. This peptide with deletion had 7 spectra counts (across 6 different tumor protein samples) with unique genomic location and 996 RNA-seq read depth (across 10 different tumor DNA samples).
(a) Identified spectra indicating possible gene fusion

(b) UCSC Genome Browser plot of possible fusion gene identifications

**Figure C.14.** Identified fusion gene peptides. This shows a possible gene fusion region where two junctional peptides are identified across two different genes. Two fusion peptide shown in this region had 15 spectra counts. HBA1 and HBA2 are Hemoglobin related genes.


[64] The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Ele-

[65] J. C. Tran and A. A. Doucette. Multiplexed size separation of intact proteins in


[67] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren,
S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification
by RNA-Seq reveals unannotated transcripts and isoform switching during cell


Miller, L. Ding, T. Golub, J. P. Mesirov, G. Alexe, M. Lawrence, M. O’Kelly,
P. Tamayo, B. A. Weir, S. Gabriel, W. Winckler, S. Gupta, L. Jakkula, H. S. Feiler,
J. G. Hodgson, C. D. James, J. N. Sarkaria, C. Brennan, A. Kahn, P. T. Spellman,
R. K. Wilson, T. P. Speed, J. W. Gray, M. Meyerson, G. Getz, C. M. Perou, and
D. N. Hayes. Integrated genomic analysis identifies clinically relevant subtypes of
glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1.

[70] M. Via, C. Gignoux, and E. G. Burchard. The 1000 Genomes Project: new

Frezza, A. M. Ruzzo, S. Rizzo, P. Crucitti, S. Galluzzo, A. Zoccoli, C. Rabitti, A. O.
Muda, A. Russo, A. Falcone, and G. Tonini. PML as a potential predictive factor
of oxaliplatin/fluoropyrimidine-based first line chemotherapy efficacy in colorectal

B. Zhang. Protein identification using customized protein sequence databases

[73] X. Wang, R. J. Slebos, D. Wang, P. J. Halvey, D. L. Tabb, D. C. Liebler, and
B. Zhang. Protein identification using customized protein sequence databases

[74] X. Wang and B. Zhang. customProDB: an R package to generate customized
protein databases from RNA-Seq data for proteomics search. *Bioinformatics*,

