UNIVERSITY OF CALIFORNIA, SAN DIEGO

Proteogenomic approach to discover cancer aberrant peptides and antibody peptides using large-scale next-generation sequencing data

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

in

Electrical Engineering (Communication Theory and Systems)

by

Seong Won Cha

Committee in charge:

Professor Vineet Bafna, Chair
Professor Drew Hall, Co-Chair
Professor CK Cheng
Professor William Hodgkiss
Professor Pavel A Pevzner

2017

The dissertation of Seong Won Cha is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California, San Diego

2017

DEDICATION

I dedicate this dissertation to my beloved parents, Myungsik Cha and Jinjune Kim, for all their love, patience, kindness, and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Many thanks to Professor Vineet Bafna for being such an incredible advisor to me. Through his guidance, I have been able to pursue the work in this dissertation. I am thankful to my committee members, Professor Drew Hall, Professor Pavel A Pevzner, Professor William Hodgkiss, and Professor CK Cheng, for their valuable comments and suggestions.

Chapter 1, in full, is a reformatted reprint of the material as it appears in Journal of Proteome Research, Jan; 13(1):21-28, 2014. "Proteogenomic database construction driven from large scale RNA-seq data." Sunghee Woo; Seong Won Cha, Gennifer Merrihew; Yupeng He; Natalie Castellana; Clark Guest; Michael MacCoss; and Vineet Bafna. The thesis author was a author of this paper.

Chapter 2, in full, is a reformatted reprint of the material as it appears in Proteomics, Nov; 14(0):2719-2730, 2014. "Proteogenomic strategies for identification of aberrant cancer peptides using large-scale Next Generation Sequencing data." Sunghee Woo; Seong Won Cha; Seungjin Na; Clark Guest; Tao Liu; Richard D Smith; Karin D Rodland; Samuel Payne; Vineet Bafna. The thesis author was a primary investigator and author of this paper.

Chapter 3, in full, is a reformatted reprint of the material as it appears in Journal of Proteome Research, Sep; 14(9):3555-3567, 2015. "Advanced Proteogenomic Analysis Reveals Multiple Peptide Mutations and Complex Immunoglobulin Peptides in Colon Cancer." Sunghee Woo; Seong Won Cha; Stefano Bonissone; Seungjin Na; David L. Tabb; Pavel A. Pevzner; and Vineet Bafna. The thesis author was a primary investigator and author of this paper.

Chapter 4, in full, is a reformatted reprint of the material as it appears in Molecular and Cellular Proteomics, Dec; 16(12):2111-2124, 2017. "The Antibody Repertoire of Colorectal Cancer." Seong Won Cha; Stefano Bonissone; Seungjin Na; Pavel A. Pevzner; and Vineet Bafna. The thesis author was a primary investigator and author of this paper.

VITA

2010	B. S. in Electrical Engineering, Brigham Young University
2013	M. S. in Electrical Engineering (Communication Theory and Systems), University of California, San Diego
2017	Ph. D. in Electrical Engineering (Communication Theory and Systems), University of California, San Diego

PUBLICATIONS

"Proteogenomic Database Construction Driven from Large Scale RNA-seq Data.", Journal of Proteome Research, 13(1):21-28, January 2014.

"Proteogenomic Strategies for Identification of Aberrant Cancer Peptides Using Large-scale Next-generation Sequencing Data.", Proteomics, 14(0):2719-2730, November 2014.

"Advanced Proteogenomic Analysis Reveals Multiple Peptide Mutations and Complex Immunoglobulin Peptides in Colon Cancer.", Journal of Proteome Research, 14(9):3555-3567, September 2015.

"Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer", Cell, 166(3):755-765, July 2016.

"The antibody repertoire of colorectal cancer". Molecular and Cellular Proteomics, 16(12):2111-2124, December 2017.

ABSTRACT OF THE DISSERTATION

Proteogenomic approach to discover cancer aberrant peptides and antibody peptides using large-scale next-generation sequencing data

by

Seong Won Cha

Doctor of Philosophy in Electrical Engineering (Communication Theory and Systems)

University of California, San Diego, 2017

Professor Vineet Bafna, Chair
Professor Drew Hall, Co-Chair

Next Generation Sequencing (NGS) and other deep sequencing technologies provide information on transcribed regions, splicing events, and single nucleotide variants in a variety of cellular conditions. The advances of this genomic technologies lead us to have better understanding about the cancer including molecular subtype, cancer progression, and biomarker discovery, but the complexity, redundancy, and errors in genomic data make it difficult to investigate aberrant genes using only genomic approaches. Combination of proteomic and genomic technologies called proteogenomics are increasingly being employed. Various strategies have been employed to allow the usage of large-scale NGS data, however, serious methodological challenges remain, especially in the identification of multiple mutational variants, structural variations, or immune system genes even

if they play an important role in cancer. This dissertation introduce the integrative proteogenomic method that extends the limit of proteogenomic searches to identify multiple variant peptides as well as immunoglobulin gene variations using advanced RNA-seq read assembly method. The result provide thousands of aberrant peptides observed in colorectal cancer and extensive characterization of tumor immune response. The result demonstrate the presence of co-expressed pair of antibody and aberrant peptides, and show that these pairs are correlated with survival time of the individuals.

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 [108, 123], 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 [115].

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 ($\sim 3\%$) of it. An improvement is to assemble RNA-seq fragments into longer transcripts, and search these reduced databases [118, 39]. 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 [122]. 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) [127], 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. Li *et al*. (2011) [65] focused on encoding SNPs (Single Nucleotide Polymorphisms) to the 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 [16, 17]. 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 [112, 17, 16], 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× 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 *C. elegans* gene set version WBcel215.68 [34]). 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 [42, 38] as part of the modENCODE project. RNA-seq read alignments are in SAM format [62], 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 [27], 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

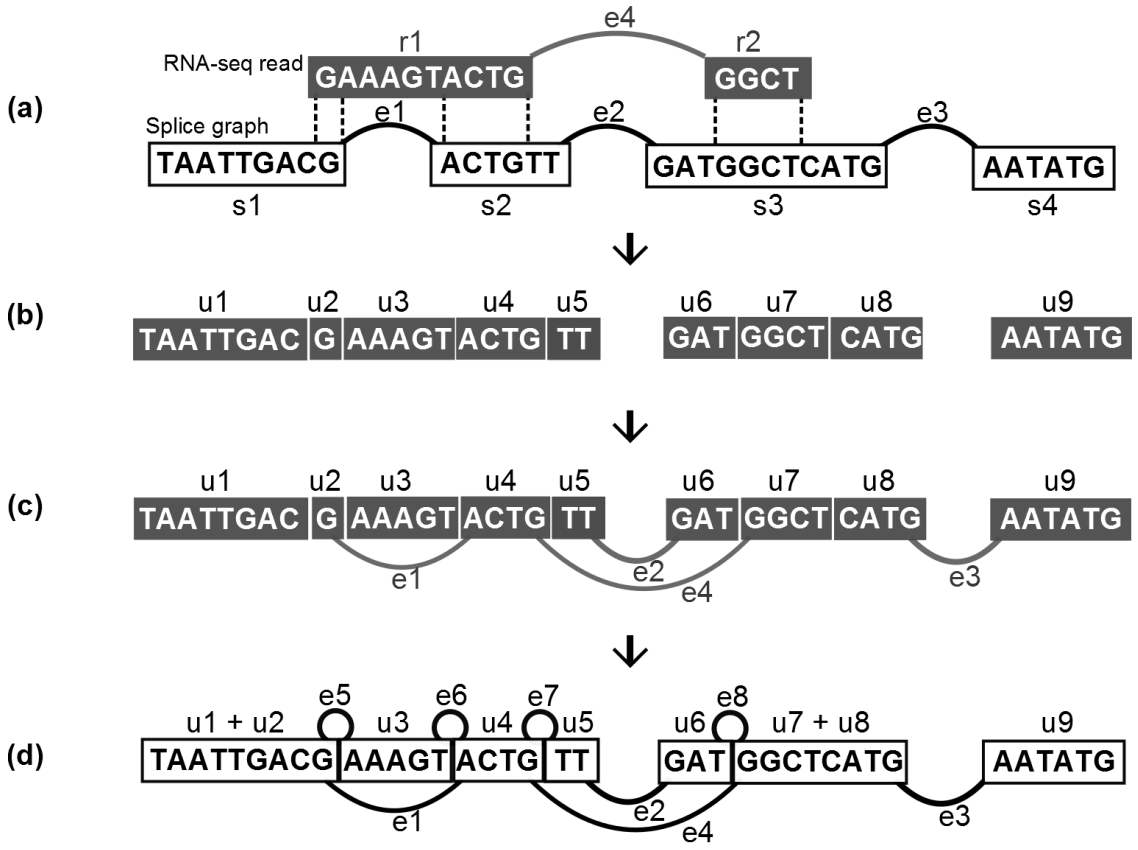


Figure 1.1: (a) Given RNA-seq read, find overlapping regions with the existing splice graph. (b) Split and add nodes. (r_1, node s_1 is split into nodes u_1 and u_2, and node u_3 is added.) (c) Assign edges for each spliced-read. (d) Revisit each pair of contiguous nodes. The nodes are merged if there is no edge at the boundaries. (Nodes u_1 and u_2 are merged, while e_5 is added between u_2 and u_3.)

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 [27], we also describe a novel method which finds a greedy but effective solution of this problem. However, unlike Edwards and Lippert [27], 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.

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 ℓ ($L \leq \ell < 2L$) 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 $l \geq 2L$, 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 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., ('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

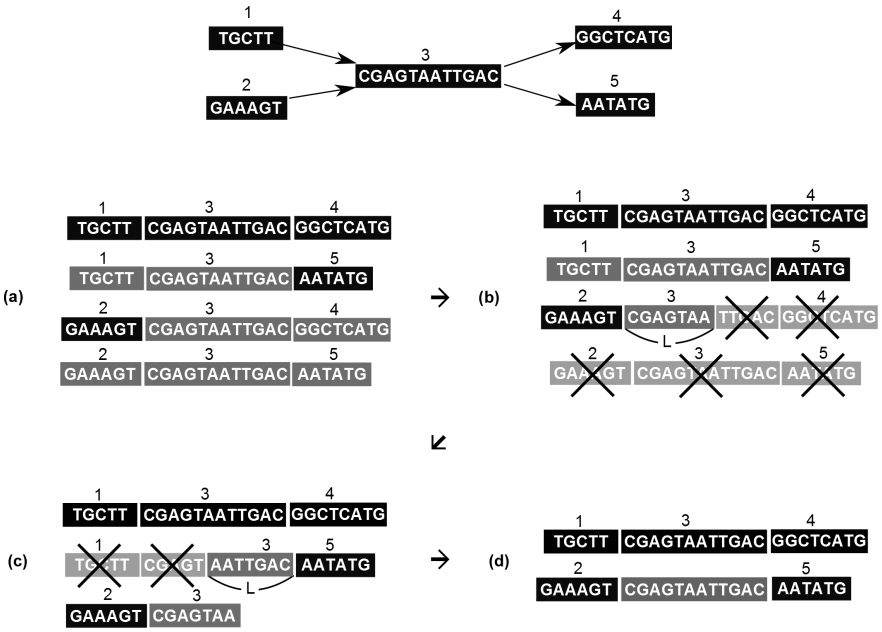


Figure 1.2: (a) By traversing the graph using a depth first search(DFS), we generate a sequence from the first visited start to end node path. (b) While traversing in DFS, when we encounter an outgoing edge that is already visited, only maintain a length $L-1$ suffix. (c) While traversing in DFS, when we encounter an incoming edge that is already visited, only maintain a length $L-1$ prefix. (d) For a pair of sequences(paths) with a prefix-suffix match, combine two sequences.

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 [42, 38, 73]. 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. [12], sucrose floated, lysed in the presence of protease inhibitors (Roche) and centrifuged to separate insoluble and soluble fractions. A $200\mu g$ 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) [116]. 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 $75\mu g$ 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) [73] is illustrated in Supplementary material. We used MSGFDB (version

20120106) [54] 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 [16, 17], 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 [48] 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 web page(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 $1000\times$ 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 [27] 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.

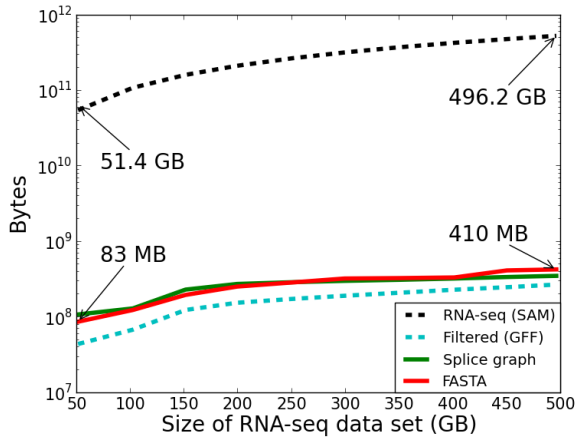
Database validation

To validate the splicing information, we compared our Splice Graph database with RefSeq [92] 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.

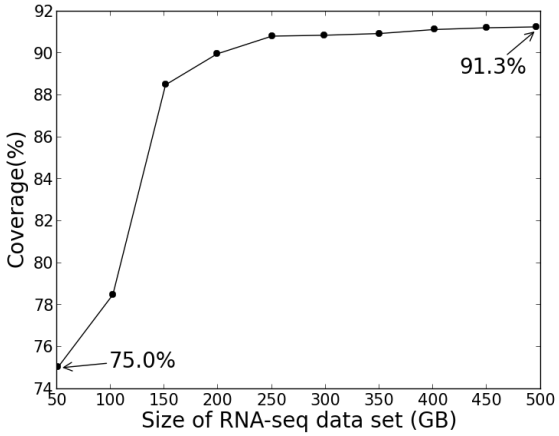
Additional growth in the Splice Graph was due to the incorporation of novel splicings in the RNASeq data, but not in RefSeq. Figure 1.3(c) plots the number of novel splice junctions in the Splice Graph. Comparing the growths in Figure 1.3(b) and Figure 1.3(c), we observed a similar growth curve, with 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

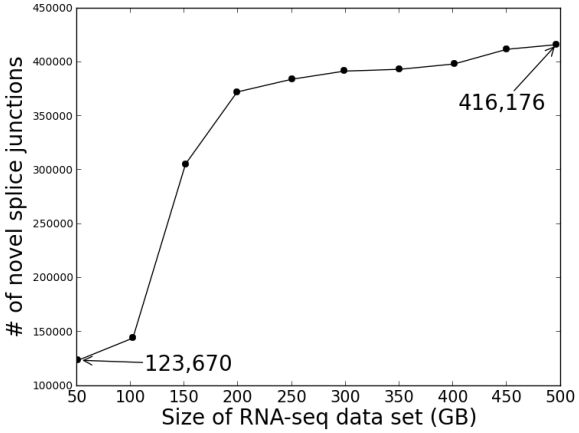
The compacted FASTA representation of splicings was used in conjunction with a known protein database in a proteogenomic analysis of a bottom up LC-MS/MS LTQ-Orbitrap data-set generated by the MacCoss lab. We used the existing proteogenomic pipeline, ENOsi [17, 16], which automatically searches all of 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, $1 - (\prod_{i \in S}(1 - \frac{(1-FDR)}{LocationCount}))$, 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



(a) Database file size growth



(b) Coverage plot



(c) Number of novel splice junction

Figure 1.3: (a) Growth of the database file size(Bytes) 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 database which does not exist in RefSeq.

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 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 [54], 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 [48]. 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.

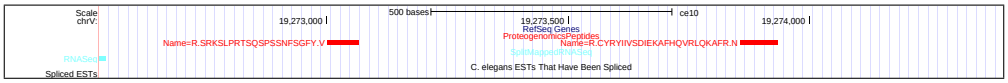
Novel genes: We identified 215 novel gene events Table 1.1 where a collection of novel peptides were located $\geq$ 3Kbp from any annotated gene, again underscoring the impact of proteomic data on the discovery of new genes. Moreover, even with the extensive RNA information, we identify new genes from the 6-frame search as well. An example is shown in Figure 1.4(a), with 2 peptides, 'R.SRKSLPRTSQSPSSNFSGFY.V' and 'R.CYRYIIVSDIEKAFHQVR LQKAFR.N', both from 6-frame database search. We looked for comparative evidence using Blastx [6]. A query sequence was extracted from the DNA region 'chrV:19272600-19274335' and searched against the nr protein database [70]. The top blastx hit was 'hypothetical protein CRE-09558 [Caenorhabditis remanei]' with e-value 0.0. As shown in Figure A.5, peptide

'R.CYRYIIVSDIEKAFHQVRLQKAFR.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 [73] in the same region containing the peptide sequence 'R.CYRYIIVSDIEKAFHQVRLQKAFR.N'.

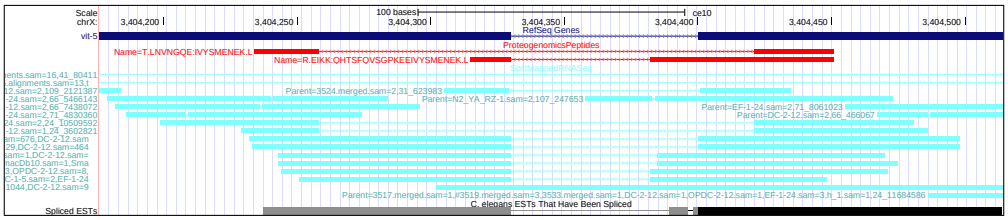
Gene corrections: The majority of the events in Table 1.1 are corrections to existing gene structures, including novel exons, extensions to UTRs, alternative splicing, frame correction, and even reverse strand events. We identified 12 alternative splicing events, with junctions that differed from RefSeq. 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 [24].

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 'TIVFTVPLSQCQMVSPMISK.E' (in the gene eef-2), which matches in a different frame. Two neighboring peptides, 'R.FIEPIEDIPSGNIAGLVG VDQYL:S.R', and 'G.HVFEESQVTGTPMFVV: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.

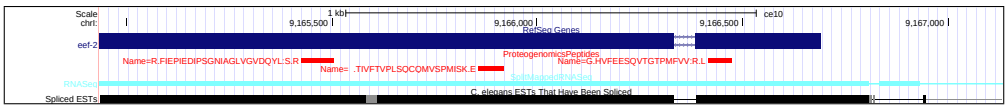
We measured the distribution of the novel peptides across different developmental stages. Figure A.6 shows the spectral counts of novel peptide spectra related to translated UTR events across different developmental stages. There is a small bias toward early developmental stages relating to translated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



(a) NovelGene



(b) NovelSplice



(c) FrameShift

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. Peptide T.LNVNGQE:IVYSMENEK.L is supported by 13 split mapped RNA-seq reads, and R.EIKK:QHTSFQVSGPKEEIVYSMENEK.L is supported by 40 reads. (c) Peptide 'TIVFTVPLSQCQMVSPMISK.E' matches in a different frame compared to the gene eef-2. Two neighboring peptides, 'R.FIEPIEDIPSG NIAGLVGVDQYL:S.R', and 'G.HVFEESQVTGTPMFVV:R.L' are identified with 1 bp deletion, that allow for the frame-shift to occur.

Table 1.1: The statistics of novel events identified

Novel Events	# of events
Alternative Splice	12
Novel Exon	808
Novel Gene	215
Exon Boundary	245
Gene Boundary	618
Reverse Strand	1166
Frame Shift	938
Translated UTR	42

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.

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 maintaing 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

Chapter 1, in full, is a reformatted reprint of the material as it appears in Journal of Proteome Research, Jan; 13(1):21-28, 2014. “Proteogenomic database construction driven from large scale RNA-seq data.” Sunghee Woo; Seong Won Cha, Gennifer Merrihew; Yupeng He; Natalie Castellana; Clark Guest; Michael MacCoss; and Vineet Bafna. The thesis author was a 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 [56, 10]. 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 (~ 0.4) [41]. Others found that as many as 20% of transcripts do not have a matching protein identification, often due to a different frame of translation [89]. 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 [64, 63, 126]. 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 [64, 63, 126], and direct translation of the outputs from

available genomics assembly tools [117, 52, 51, 72, 25]. Alternatively, our method favors a graph structured accumulative approach [133] 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 [64] 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 [112, 17, 16, 18].

To focus on the effect of choosing different strategies in 'peptide identificatoin 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 [10, 56] 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 [133], we extended the splice graph database construction method to encode a more extended list of genomic variants. Splice graph [133] 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 [72, 25] 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 [133]. 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 [27, 133] and completeness [27, 133] of the original search space (state of proof shown in our previous study [133]). 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 2.1. Our first step in database creation is to extract useful information from RNA-seq alignment/coordinate files (BAM/SAM [62], GFF, BED) and variant calls (VCF). Details of BAM file processing can be found in our previous study [133].

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 [18].

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 [133]. 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 [133]. 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 2.1). We used MSGF+ [54] 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 [34, 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 [64] 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) [64], 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 [64]. 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 database 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, $100mM$ 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 ($10mM$ triethylammonium bicarbonate, pH 7.5) to 10% B($10mM$ 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 $0.5mL/min$. 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 $0.15\mu g/\mu L$ for LC-MS/MS analysis.

The LC system was custom built using Agilent 1200 nanoflow pumps (Agilent Technologies, Santa Clara, CA). A $35cm$ x $360\mu mo.d.$ x $75\mu m$ i.dreversed-phase column wasslurry packed with $3\mu m$ 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 $0.1ms$, 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 [10, 56] repository.

The statistics of RNA-seq data-set and corresponding constructed FASTA databases are illustrated in Table 2.1. 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 nucleotidevariants 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 totalof 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 [72, 25] 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 [72, 25] 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 [56, 10] 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 [133, 18], 470 novel findings were called (Table 2.2) 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 [112, 17, 16, 18]. 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 [105] 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 [72, 25] 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 'QNLLQAAGNVGQASGELLQQIGESDTDPHFQ: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 'QNLLQAAGNVGQASGELLQQIGESDTDPHFQ: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 [64, 63, 126], 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 [117, 52, 51, 72, 25], 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 [30] 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 [133], 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 [121]. 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 [64, 126, 128, 63] 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 [64, 126, 128, 63] 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.

Spectra used in this study can be found at ftp://MSV000078631@massive.ucsd.edu.

ACKNOWLEDGEMENTS

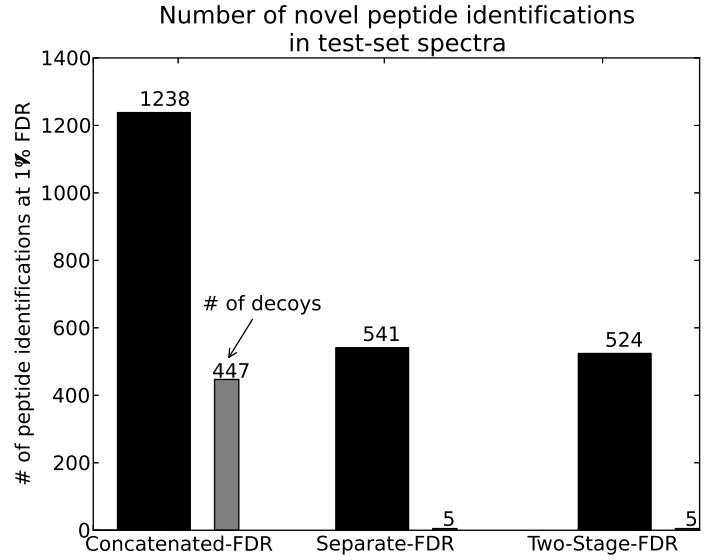
Chapter 2, in full, is a reformatted reprint of the material as it appears in Proteomics, Nov; 14(0):2719-2730, 2014. “Proteogenomic strategies for identification of aberrant cancer peptides using large-scale Next Generation Sequencing data.” Sunghee Woo; Seong Won Cha; Seungjin Na; Clark Guest; Tao Liu; Richard D Smith; Karin D Rodland; Samuel Payne; Vineet Bafna. The thesis author was a primary investigator and author of this paper.

Table 2.1: Statistics of created cancer databases

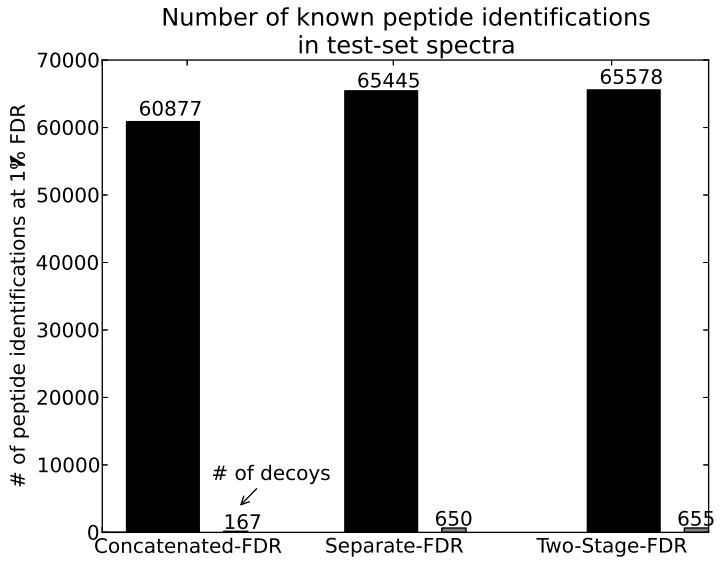
	Single OV sample	OV (PNNL samples)	OV (splice only)	BRCA (splice only)
# samples	1	67	228	484
BAM size	7.2 GB	750 GB	2.0 TB	3.2 TB
FASTA size	187 MB	607 MB	395 MB	814 MB
Novel splice	168,289	321,587	498,233	646,629
Deletions	3150	38,464	-	-
Insertions	62	1,105	-	-
substitutions	7109	182,302	-	-

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.)

Type of novel findings	# of novel findings
Substitution	236
Deletion	5
Novel splice junctions	90
Alternative splice junctions	74
Novel gene	49
TranslatedUTR	2
Exon boundary	4
Novel exon	6
Reverse strand	4

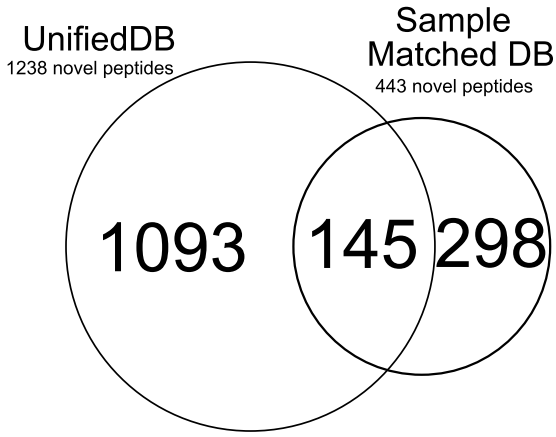


(a) Novel identifications

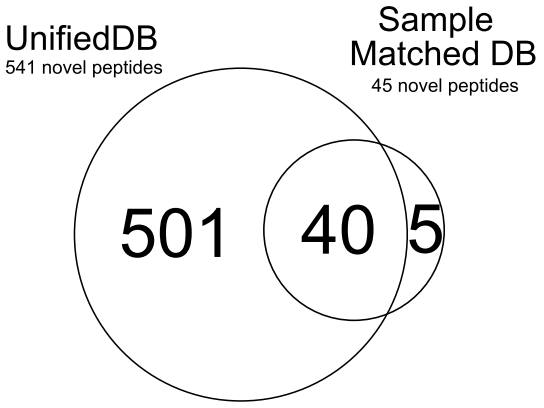


(b) Known identifications

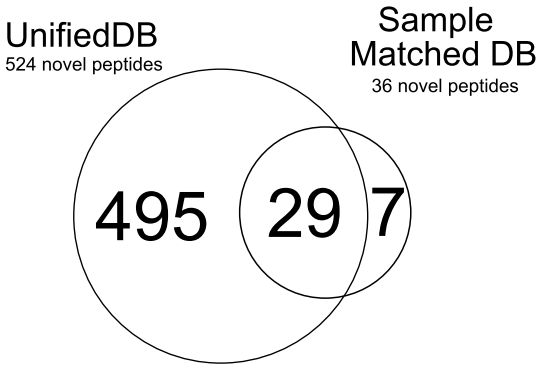
Figure 2.1: Number of peptide identifications in 439,858 spectra collected from a single sample (sample id: TCGA-24-1467) using different FDR based error control strategies.



(a) Combined-FDR

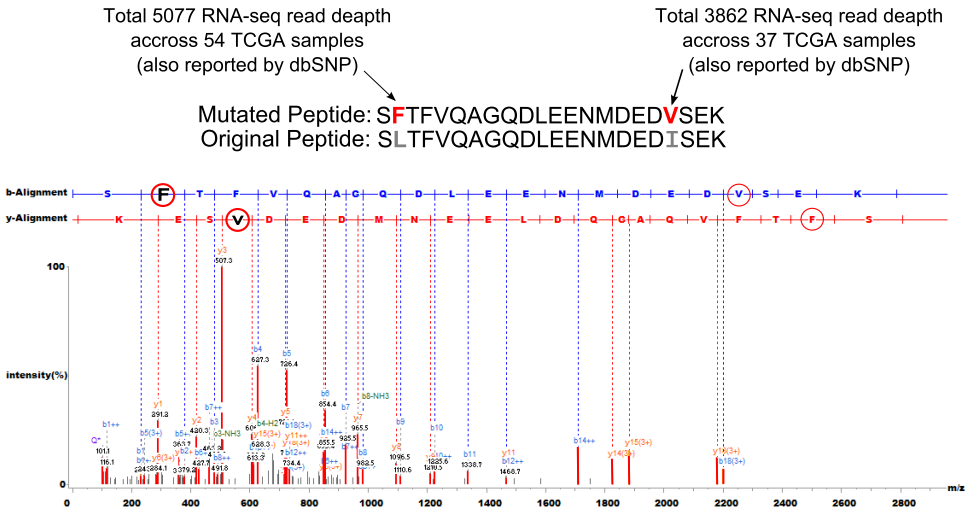


(b) Separate-FDR

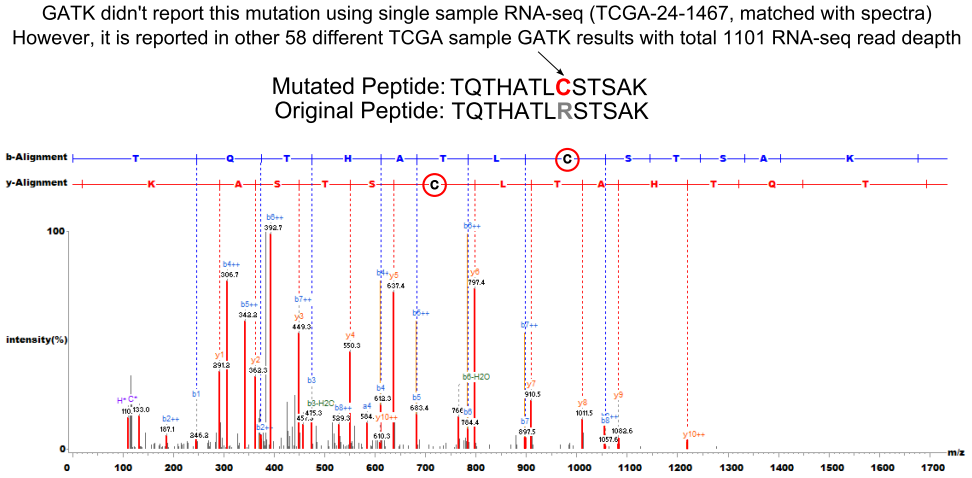


(c) Two-stage-FDR

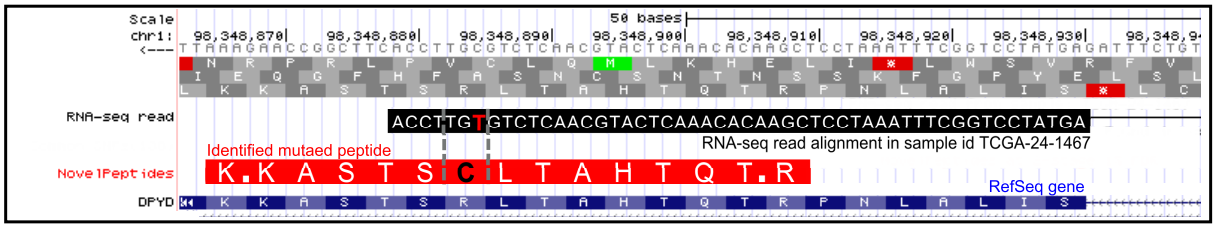
Figure 2.2: Overlap between novel identifications from unified and single sample database.



(a) Identified spectra with 2 substitutions

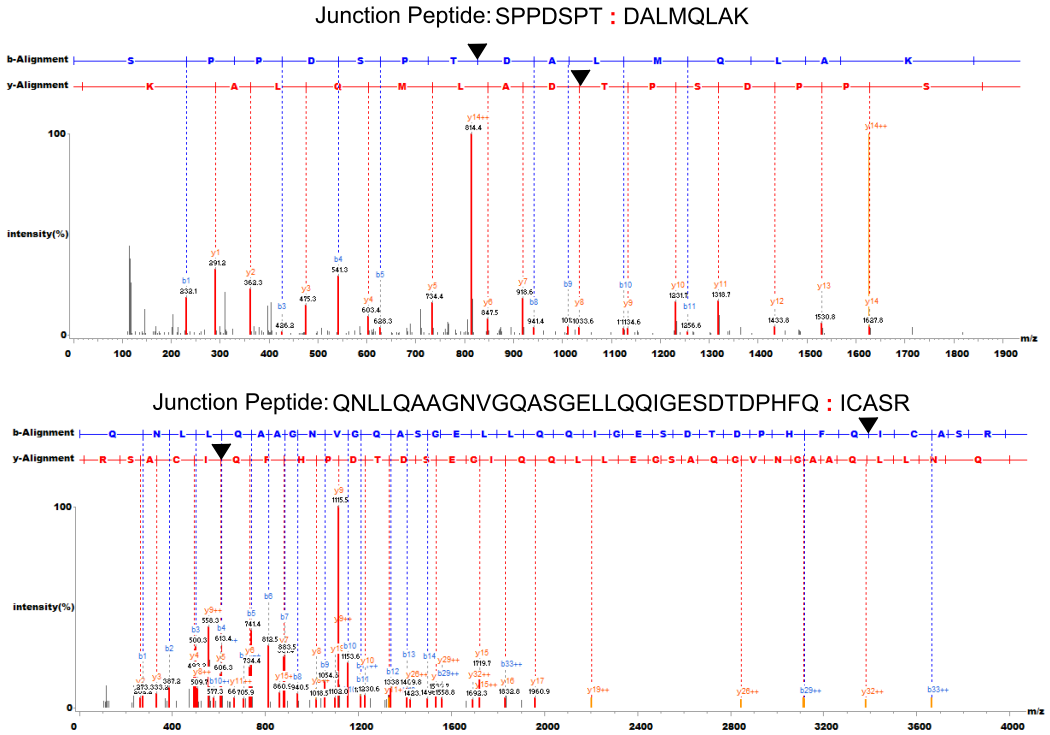


(b) Identified spectra with 1 mutation

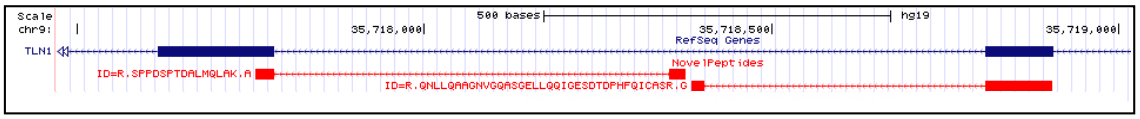


(c) UCSC Genome Browser plot of identified SNV mutated peptide

Figure 2.3: Alignment of identified spectra of mutated peptides.

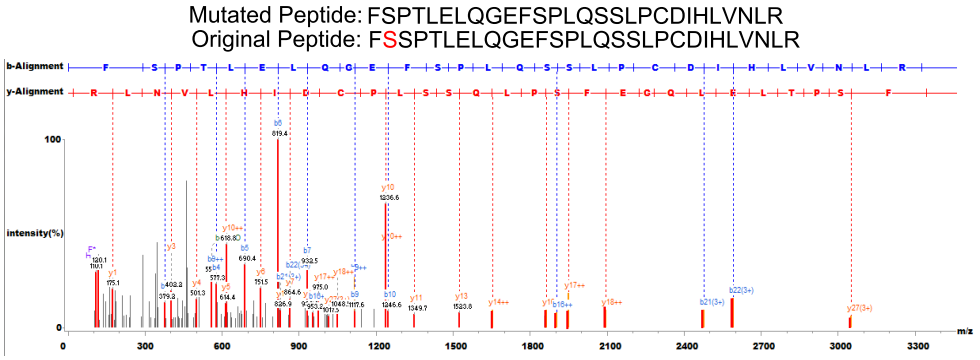


(a) Identified spectra with alternative splice junctions: ':' is inserted within the peptide sequence to indicate where the splice junction appears

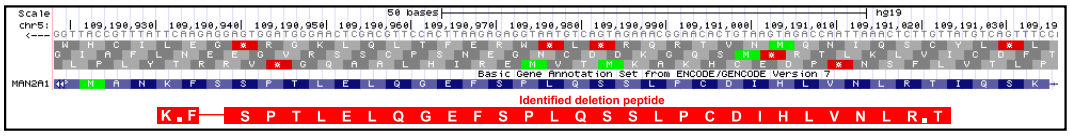


(b) UCSC Genome Browser plot of alternative splice junctions

Figure 2.4: Alignment of identified spectra of novel junction peptides.



(a) Identified spectra with deletion



(b) UCSC Genome Browser plot of identified deleted peptide

Figure 2.5: Alignment of identified spectra of mutated peptides.

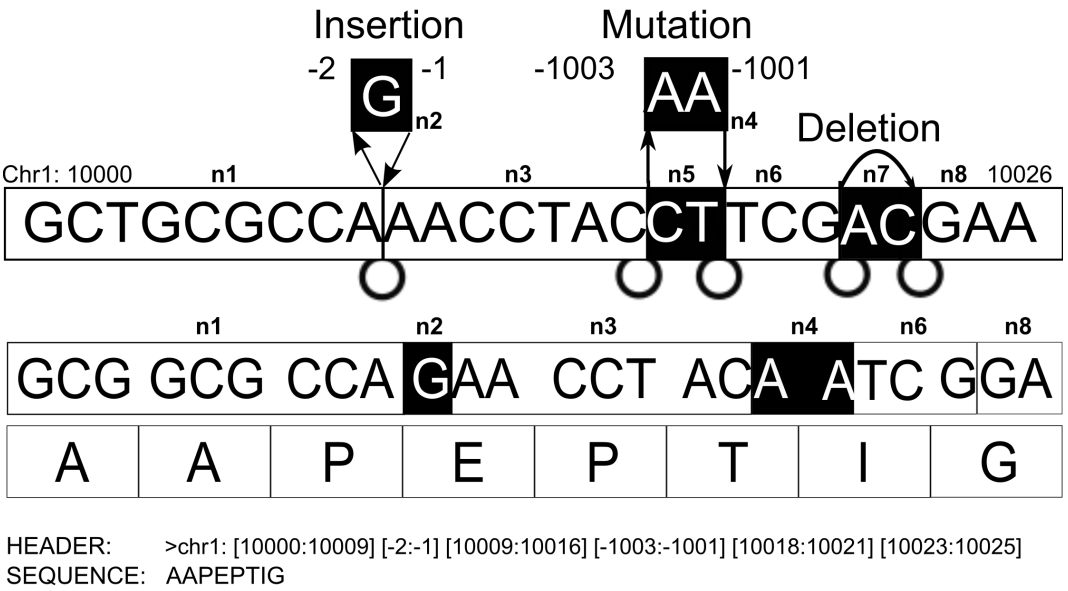


Figure 2.6: Insertions and substitutions are represented as additional node and edges having negative coordinate values. Deletions are represented same as splice junctions with actual DNA coordinates.

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 [56, 10, 75]. 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 [138]. 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[133, 131, 18, 16].

Despite many proteogenomic methods having been recently proposed [63, 64, 126, 128, 33], serious methodological challenges remain. While the initial goal of the CPTAC [2] colon cancer study has been delivered [138], 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 [138, 63, 64, 126, 128, 33]. 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 [138]. 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 [75] (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 [133, 131]. 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 rat, and extend our own previous approaches [17, 16, 18, 131] 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 [80, 77], 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 [81]; 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 [87, 137, 19]. 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× more variations compared to the initial CPTAC study [138]. Moreover, it should be noted that by applying conventional FDR strategy, 15.3× additional novel identifications were found which includes 96.25% of RNA expressed mutations from the initial CPTAC colon cancer study [138].

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 [10, 56] 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 [133, 131]).

Extended proteogenomic database for immunoglobulin peptide search. The database 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 [19].

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 [133]. 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 [34] (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 [28] 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* α, where the richness corresponds to the fraction of peptide spectrum matches (PSMs) that are correctly mapped to the peptide. Thus, the value of α 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, and 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 τ such that

$$\frac{F_D(\tau)}{F_T(\tau)} \leq 0.01 \tag{3.1}$$

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

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

Integrating and substituting, the goal is to find a minimum threshold τ s.t.

$$\frac{F_D(\tau)}{F_T(\tau)} = \frac{F_I(\tau)}{\alpha \cdot F_C(\tau) + (1-\alpha) \cdot F_I(\tau)} \leq 0.01 \tag{3.2}$$

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 [131] 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 (COAD) and Rectum Adenocarcinoma (READ) tumor MS/MS spectra were searched against the known protein and specialized proteogenomic database using MSGF+ [54]. 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 benchmarked our search against previous searches of the same MS data, including Zhang et al. [138] who used their own databases (CanProVar), and against a second search-tool using Comet [29] on our specialized databases as a control. The Comet results [29] showed 357 novel peptide identifications while 70.86% of the peptide overlapped with MSGF+ [54] results (Figure C.4). In generating novel events reported in this paper, we used the union set of both MSGF+ [54] and Comet [29] 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 [35, 105, 75], 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 [138]. As we did not have matched proteomic data from normal samples, we used an earlier study from TCGA [75] 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 [75]. 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 [110], HSPG2 [46], PML [125],

GBP-1 [13], TF [104], NES [114], 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 [80], the role of B-cells is still being elucidated, although some reports suggest that B-cells aggregate in tumors [67, 78], where they form germinal centers, undergo class switching, and differentiate into plasma cells [77, 81].

Distribution of IG peptides across colorectal subtypes. The CPTAC study classified the 90 samples into five subtypes, marked 'A' through 'E' [138] 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 [53] tissue, and colon cell-line [31] 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' ($p\text{-}value < 0.0001, \chi^2 = 2927.71$), 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. [99] and De et al. [22]. 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 [75] and CPTAC [138] 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 [74, 68]. We identified a single PSM 'VPSSCPIVTVDGYVDPSGGD;H;FCLGQLSNVHR' ($R361H$, Figure C.8) supporting a known mutation in colorectal cancer [35], 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 propsoed as a marker for other cancers [113].

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:PPNANEDYR' in STK39 (a cellular stress response pathway gene [91]) 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 [75] 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 accross 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 [66].

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 n 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

Chapter 3, in full, is a reformatted reprint of the material as it appears in Journal of Proteome Research, Sep; 14(9):3555-3567, 2015. “Advanced Proteogenomic Analysis Reveals Multiple Peptide Mutations and Complex Immunoglobulin Peptides in Colon Cancer.” Sunghee Woo; Seong Won Cha; Stefano Bonissone; Seungjin Na; David L. Tabb; Pavel A. Pevzner; and Vineet Bafna. The thesis author was a primary investigator and author of this paper.

Table 3.1: **Enosi characterization of aberrant events**. 61 samples out of 90 had blood (normal) samples available as a matched reference. Using DNA level normal) sample mutation calls, we were able to distinguish 106 somatic and 298 germline mutations among 650 substitutions. (246 substitutions remained uncategorized due to the absence of normal reference samples)

Type of novel findings	# of novel findings
Somatic substitution	106
Germline substitution	298
Uncategorized substitution	246
Somatic insertion	1
Uncategorized insertion	3
Deletion	4
Transcript gene	10
Fusion gene	5
Translated-UTR	16
Alternative splice	11
Novel splice	91
Exon boundary	6
Frame shift	4
Novel exon	2
Novel gene	4
Reverse strand	1
Pseudo gene	14
IG gene variable region	1,062

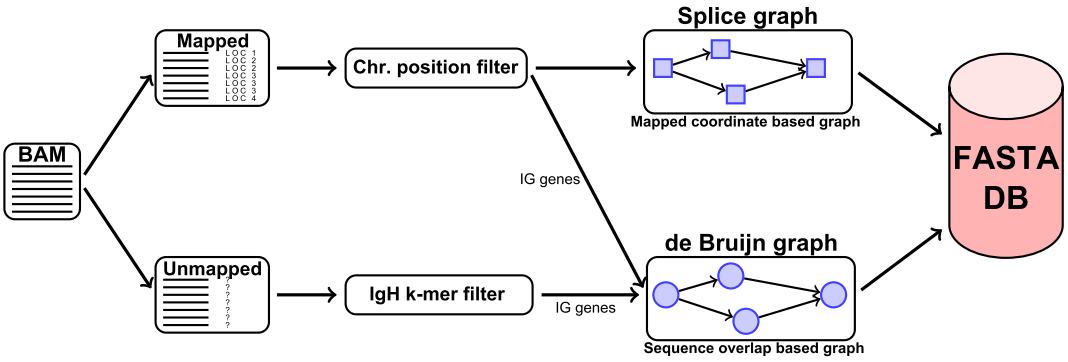


Figure 3.1: Illustration of proteogenomic database construction for immunoglobulin peptide identifications.

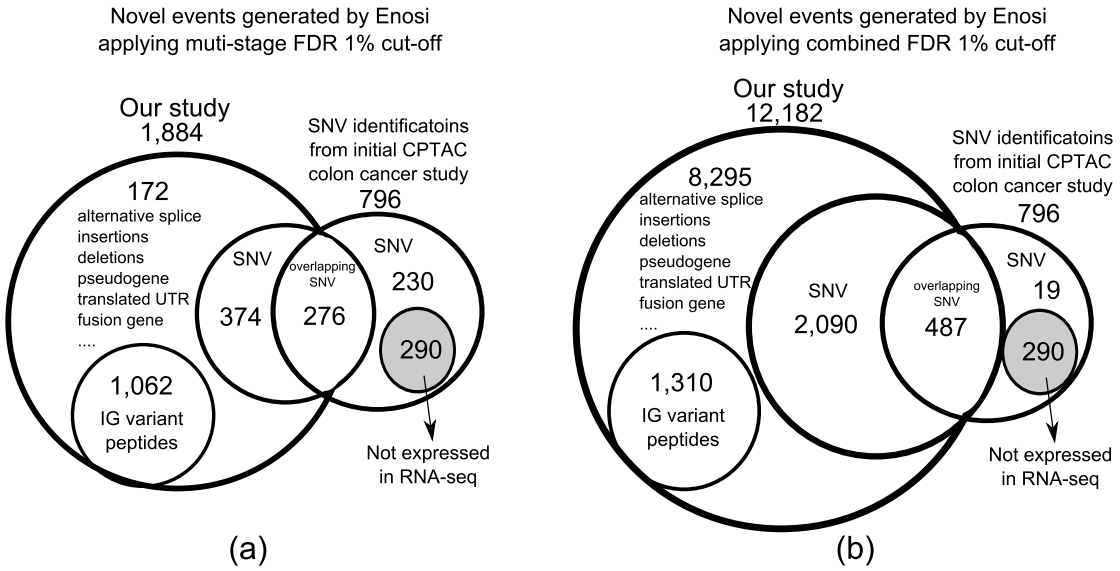
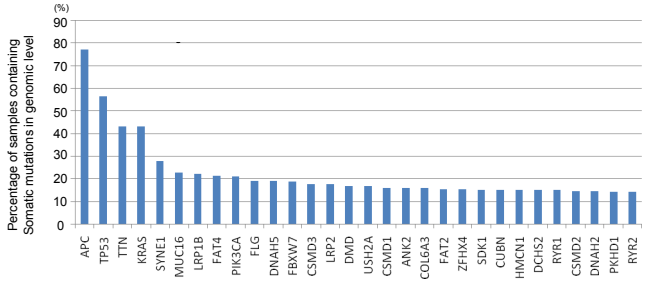
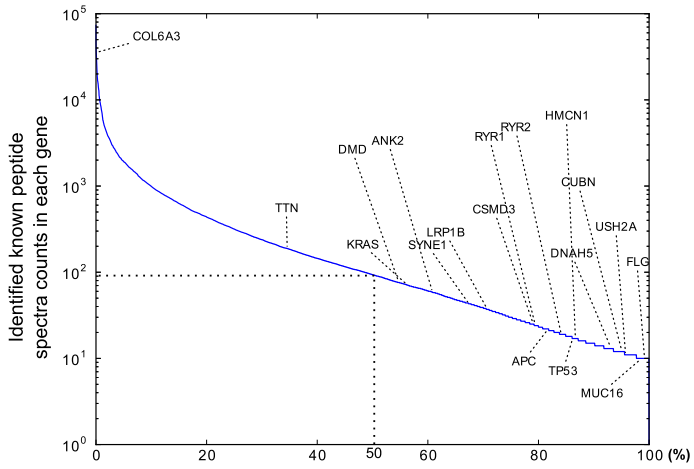


Figure 3.2: (a) Comparison of aberrant peptide identifications against previous findings using multi-stage FDR (b)Comparison of overlapping aberrant peptide identifications using combined FDR. Our proteogenomic database was created from raw RNA-seq alignments from TCGA repository and database used in Zhang et al. [138] is created from SNV informations reported by dbSNP [105], COSMIC [35], and TCGA somatic mutation calls [75].

(a)



(b)



(c)

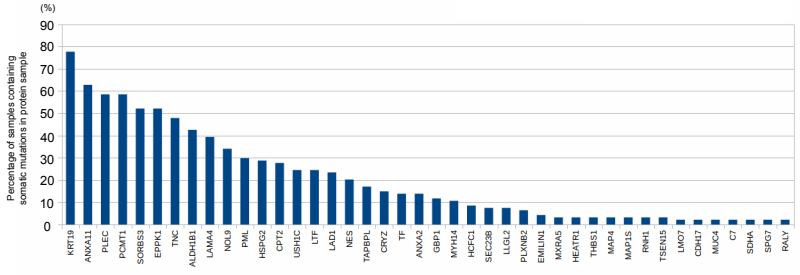


Figure 3.3: (a) Genes containing most frequent somatic mutations reported by the TCGA study. (b) RefSeq identified spectra per gene in 10 based log scale. Most frequently mutated genes in DNA level are under expressed in protein level. COL6A3 had 35463 spectra counts, TTN (188), KRAS (71), DMD (76), SYNE1 (43), LRP1B (37), ANK2 (59), and rest of the DNA level highly mutated genes had less than 25 spectra counts. (c)Percentage of samples containing identified protein mutations in TCGA reported most frequently genes. While most of the DNA level top frequently mutated genes were under expressed in protein level, we observed that some genes showed even higher mutation frequencies across samples in protein level.

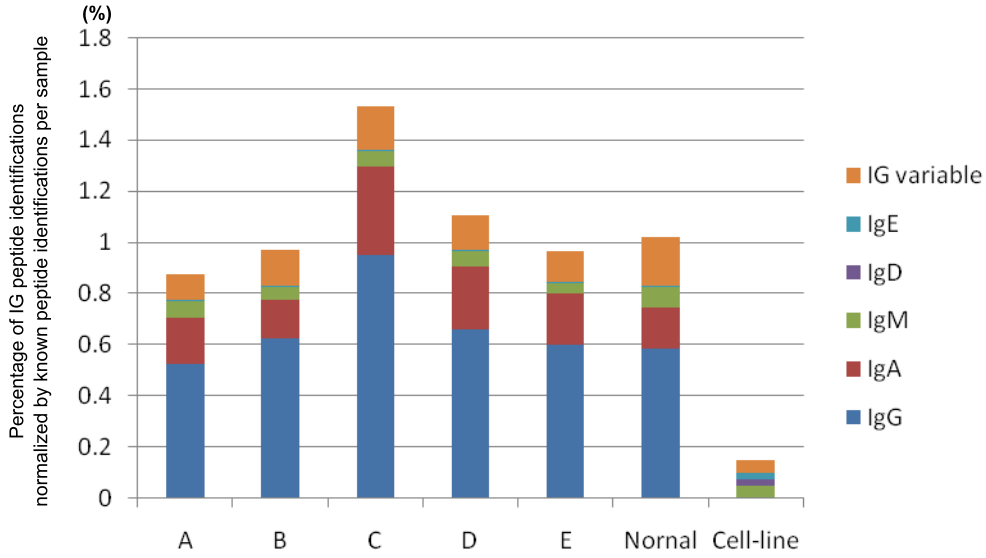


Figure 3.4: Percentage of IG gene peptide identifications in each sample normalized by the number of known peptide identifications across sample subtypes. This percentile ratio is calculated by dividing the number of known peptide identifications from the total number of IG peptide identifications within each sample. (ratio = (# of IG peptides) / (# of known peptides) * 100) Different kinds of IG gene segments are colored. Subtype C (sample groups showing both hypermutation and CIMP characteristics) showed comparably high number of IG gene peptide identification compared to other sample subtypes. Chi-squared test of this plot showed $p\text{-}value < 0.0001, \chi^2 = 2927.71$.

Chapter 4

The antibody repertoire of colorectal cancer

4.1 Abbreviations page

Abbrev.	Definition
TCGA	The Cancer Genome Atlas Project
CPTAC	Clinical Proteomic Tumor Analysis Consortium
TIL	tumor infiltrating lymphocyte
Ig	immunoglobulin
SdB	'split' de Bruijn
dB	de Bruijn
FR	framework region
CDR	complementarity determining region
IMGT	the international ImMunoGeneTics information system
COAD	colon adenocarcinoma
SAAV	Single Amino-Acid Variant

4.2 Introduction

Cancer immunotherapy, which attempts to tackle cancer using the body's own immune response, has been very successful in boosting the survival rates of patients with leukemia and other

blood cancers [21, 93, 124]. This field of research is expanding rapidly, and has been extended to include other cancer sub-types, including solid tumors [47, 58, 23].

Immunotherapy is more specific than generic typical cancer treatments targeting fast-growing cells directly. It can take the form of cancer vaccines (neoantigens that stimulate an immune response) [69, 97], monoclonal antibodies, which target cancer cells expressing specific (neoantigenic) proteins [102] or immune checkpoint inhibitors that activate suppressed immune cells [11, 15, 43]. The development of new forms of cancer immunotherapy could be greatly helped by knowledge of the cancer specific immune response, especially in understanding the antibodies and neoantigens specific to cancer.

This is a challenge because of the millions of distinct antibodies that are circulating in the blood. We still have only limited knowledge of the antibody responses that target individual disease-related antigens and epitopes. There are only a few known examples in infectious disease [57] and autoimmune disease [45]. On top of that, recent methods that characterize the antibody repertoire use serum or plasma samples as their source for antibody analysis. However, the antibodies in these samples include the pool of all antibodies binding to multiple antigens, as well as the antibodies produced by numerous previous immune responses [95, 61, 90, 88, 60]. Screening the antibodies based on their binding to pre-selected antigens may also not work, as all possible neoantigens existing in a sample cannot be known, and some important antigens may be post-translationally modified [111, 49] or cleaved [119].

Another approach to understanding the antibody repertoire is by isolating the B-cells that respond to a target immunogenic antigen. Plasmablasts [135, 111], memory B cells [71, 36, 26], and tissue infiltrating B cells [8, 109, 107] have been used to characterize the functional antibody repertoire [94, 37]. The method works, but it requires a dedicated workflow to isolate the B-cells and sequence the antibody clones. Here, we propose a more direct method for discovering antibody peptides in tumor samples.

Recently, we and others have developed pipelines for identifying mutated peptides expressed specifically in cancer [132, 134, 130]. In our approach, we mine a general transcript resource (such

as The Cancer Genome Atlas Project) to extract transcript sequences, identify novel mutations, and junctions, then encode them into a complex database. This database is then searched via a proteogenomics approach, to identify peptides that are seen only in tumor proteome samples. Interestingly, our initial search of the CPTAC colorectal tumor samples identified a number of antibody peptide sequences [132]. Supplemental Fig. D.1 shows the example of some antibody peptides identified in the search. At the time, there were questions regarding the provenance of the discovery, as we did not expect to find antibody peptides in colon tissue. They could be antibodies from tumor infiltrating lymphocytes (TIL), circulating antibodies from blood contamination, encoding general proteome variation, or even mis-identifications. Moreover, our databases were not specifically designed to capture Ig regions, so we were only identifying peptides from some of the annotated Ig genes on the human reference.

AbScan is a new tool for identifying all antibody (Ig) peptides in a sample by searching mass spectral datasets against RNA-seq datasets. AbScan is a proteogenomic tool that scans transcript and genomic data, preferably, but not exclusively from the same samples as the proteomic data; it creates specialized antibody sequence databases that can search tandem mass spectra. As the antibody sequences are hypervariable, identifying and characterizing transcripts encoding Ig genes is a challenging endeavor. We devised a special construct called the 'split' de Bruijn (SdB) graph to encode all Ig transcripts in a compact fashion, then show the power of this approach compare to other methods. AbScan also uses a customized pipeline to search these antibody databases and identify expressed antibody peptides, while controlling for false discoveries. We evaluated sensitivity and specificity of AbScan by benchmarking it on simulated datasets, pure antibody mixtures, normal colon tissues, and colorectal cell-lines. We further applied AbScan to 90 colorectal samples from the CPTAC project and demonstrated that the antibody repertoire was characterized by significant co-occurrence pattern in 163 pairs of antibody peptides and Single Amino-Acid Variant (SAAV) pairs, and the co-occurring pairs were correlated with patient survival.

4.3 Experimental Procedures

Experimental MS data sets, sequence databases, and search parameters. We analyzed four spectral datasets, which have been described in previous work.

- 90 colorectal tumor samples [139]
- 30 normal colon biopsies [139]
- Colon cancer cell-lines LIM1215, LIM1899, and LIM2405 [32], and
- Purified polyclonal antibody mixture [100].

We searched each tandem mass spectra against three different databases. These included:

- Ensembl database version GRCh38 [20]
- The 'split' de Bruijn (SdB) graph based database driven by the method described in this paper, and
- A de Bruijn (dB) graph based database [130].

We used MS-GF+ (version 1.1.0) [55] with the following parameters: parent mass tolerance of 20 ppm, and allowed post-translational modifications of fixed carbamidomethyl C and optional oxidized Methionine. Common contaminants were excluded. ProteoWizard (v3.0.3827) [50], and ReAdW (v1.1 and v4.3.1) [84] were used for the peaklist-generating software. Number of missed and non-specific cleavages permitted was 1. Trypsin was used to generate peptides for three colorectal data sets, and Trypsin, Asp-N, Chymotrypsin, Elastase were used for the purified polyclonal antibody mixture data set. A multi-stage FDR (See Experimental Procedures–'Multi-stage search') was applied to identify the PSMs from the SdB and dB driven databases (See Supplemental Table 1).

Database construction using 'split' de Bruijn (SdB) and de Bruijn (dB) graphs. AbScan constructs the 'split' de Bruijn (SdB) graphs for multiple RNA-seq datasets from tumors. A de Bruijn (dB) is constructed for a fair comparison. Followings are the steps to generate a custom MS searchable database:

1. **Read filtering**. Filter out all RNA-seq reads not sampling an Ig gene.

2. **SdB graph construction**. Create a SdB graph based database from filtered reads.
3. **Error correction**. Identify and eliminate sequencing errors.
4. **FASTA database construction**. The SdB graph is used to generate an MS searchable FASTA formatted database, as well as scripts to identify the context of the peptide on the antibody sequence.

For comparing the performance of SdB graphs to dB graphs, we used an implementation of dB graphs customized for the discovery of antibody peptides [130].

Read filter. All antibodies are a combination of relatively fixed framework (FR), and hyper-variable complementarity determining regions (CDR), with the order given by 'FR1, CDR1, FR2, CDR2, FR3, CDR3, FR4'. The typical lengths of CDRs in human are 15 to 30 nt for CDR1 and CDR2, and 24 to 36 nt for CDR3 [98]. On the other hand, lengths of RNA-seq reads in our datasets varied from 76 to 100 nt. Therefore, we expect most RNA-seq reads to cover some part of a framework region, and could use this to filter RNA-seq reads from Ig genes. In addition, we employed keyword matching to recover non-mapped Ig gene encoding reads by creating a list of k-mer sequences from all Ig genes in the IMGT reference [59], and selecting all reads that matched one of the k-mers.

An appropriate value of parameter k was determined by comparison with decoy data obtained by reversing the IMGT reference sequences. As k is made smaller, we can quantify the false matches by the number of reads that match decoy k-mers. For any value of k, the false discovery rate is given by

$$\mathrm{FDR}(k) = \frac{\#\text{number of reads matching decoy } k\text{-mers}}{\#\text{number of reads matching target } k\text{-mers}}$$

We selected the smallest value of k that resulted in a FDR below 1%, $k = 19$ was used for filtering (Supplemental Fig. D.2). Quality filtering was applied additionally to trim the part of the poor quality reads. We trimmed the $3'$ end of the reads if their quality threshold were less than the threshold value (10). We excluded the read if the trimmed part was longer than $\frac{2}{3}$ of the read length or the overall quality of the reads were below than the threshold value (25).

SdB graph construction. Typical de Bruijn (dB) graph construction is as follows: Given a set of reads, the de Bruijn (dB) graph for this set is defined as follows: each k-mer from reads is a node in the graph. Nodes u and v are connected by an edge, if there exists a $(k+1)$-mer in reads whose k-suffix is u and whose k-prefix is v. dB graphs are a powerful construct because they help remove redundancy in read coverage, and can be efficiently constructed without the need to compare all pairs of reads to test for overlap [86, 14, 96]. In the ideal case, each of the Ig genes is a path in the graph, and each path in the graph corresponds to a putative Ig gene. Errors can arise in dB graph construction if two unrelated reads share the same k-mer (we denote these as 'false-positive' overlaps), or if reads from the same molecule do not share a k-mer due to sequencing errors (false-negatives).

False edges in the dB graph can also arise due to repeated k-mers. Specifically, a repeated substring of size greater than or equal to k will lead to false edges in a k-mer based dB graph. The error could be controlled by selecting larger values of k, but that would result in a higher false-negative rate. We reasoned that the exact match requirement in k-mer dB graph is restrictive. For example, consider two reads that overlap over 40 bp. The probability that this overlap contain $k = 30$ consecutive nucleotides with no error in both reads is 65.5%. On the other hand, the probability that this overlap contain a 30 consecutive nucleotides with at most one error is 93.0%. Therefore allowing for an approximate match improves sensitivity from 65.5% to 93.0%. See Supplemental Method – 'Analytical comparison of SdB and dB graphs' for a rigorous analysis.

An alternative aproach is to do an error-correction prior to matching. BayesHammer uses a Bayesian approach on 1-neighborhoods of k-mers to correct reads, before constructing a k-mer dB graph [79]. In Ig genes, however, we use RNA data to identify variation, and the variable coverage makes it difficult to distinguish sequencing errors from true genetic variation. The SdB graph handles this problem of non-uniform coverage through correction on local nodes similar to the IGdb, Trinity and IDBA-tran [40, 85, 130].

On top of that, SdB graph applies a binning technique to solve the approximate matching problem efficiently. To obtain 1-neighborhoods of k-mers, we need the pairwise distance of every

existing k-mers observed from the reads, which may increase the computation time for large value of k. We divided the k-mers into two parts: one (r-mer) for binning, and the other (ℓ-mer) for 1-neighborhood testing. The size of the bin decides the average number of nodes required the pairwise distance computation. Note that the SdB graph is a generalization of prior approaches with bin size of k (respectively, 0) corresponding to a standard dB graph (respectively, BayesHammer like graph). In our tests, we did some empirical tests to choose r and ℓ and found the performance to be robust to difference choices. Therefore, we worked with $r = 10, \ell = 20$, leaving the optimization of parameters to future work. However, to allow for fair comparisons, we tested the SdB graphs against dB graphs using a range of values of k.

Given r, ℓ, we build a SdB graph as follows:

1. Each node initially corresponds to a distinct $(r+\ell)$-mer from the read. Node $u = (x, y)$, where x is a length r of prefix of the node, and y is a length ℓ of suffix of the node.
2. Consider nodes $u = (x, y)$, and $v = (x', y')$. We connect u and v by an edge, if the $r + \ell - 1$ suffix of u matches the prefix of v, and a read matches the combined sequence. The weight of an edge (u, v) is the number of reads that contain the combined sequence. The weight of node u is the maximum of the sum of incoming or outgoing edge-weights. This operation mimics a standard dB graph construction.
3. Consider nodes in order of decreasing weights, and repeat the following until no node is left:
   (a) Pick node $u = (x, y)$. For all nodes $u' = (x', y')$, merge u' with u (and remove from further consideration) if $d_h(x, x') = 0$, $d_h(y, y') \leq 1$ and u is the heaviest, where $d_h(x, x')$ is hamming distance between x and x'. Merge any multi-edges into a single edge of weight equal to the sum of the weights of the merged edges. Note that the actual implementation speeds this computation by hashing on the prefix strings.

The construction of a SdB is illustrated in Fig. D.3. A $(3, 3)$ SdB graph successfully compacted the data with no false-positives or false-negatives except due to sequencing error. As Fig. D.3 shows, there are two distinct paths, corresponding to the two genes. However, due to

sequencing error, we see a small branching in gene 1. This can be controlled by an error correction procedure, described in the next section. In contrast, Supplemental Fig. D.4 shows examples of dB graphs with the choice of $k = 4$ and $k = 5$ using same reads. A 4-mer dB graph connected false edges at node '*GAAT*', producing false paths combining gene 1 and 2. On the other hand, a 5-mer dB graph failed to connect edges in both genes, and neither gene could be represented by a single path. In the Results section, we systematically compare the performance of SdB graphs and dB graphs.

Error correction. Sequencing errors also result in false overlaps. An error towards the end of the read (within k nucleotides) leads to a 'tip' in the dB graph, while an error in the middle of the read leads to a 'bulge.' Subsequent to its construction, the SdB graph can be viewed as a regular $(r+\ell)$-mer graph; graph simplification methods, such as tip clipping and bulge removing can be applied. For transcript assembly, uniform coverage pruning may delete some true sequences, so we use a proportional approach to rescue lower-abundance transcripts similar to the one used by IGdb, Trinity and IDBA-tran [40, 85, 130].

Assuming for simplicity that sequencing errors are independent and identically distributed with ε_s denoting the nucleotide error probability. The number of reads matching a specific k-mer is proportional to $(1-\varepsilon_s)^k$. On the other hand, the number of reads matching a k-mer with a mismatch at a specific position is proportional to $\frac{1}{3}\cdot\varepsilon_s\cdot(1-\varepsilon_s)^{k-1}$. The expected ratio of read depths of the true edge to any false edge is given by

$$\frac{(1-\varepsilon_s)}{\frac{1}{3}\cdot\varepsilon_s},$$

The expression is usually $\gg 1$, for typical values of $\varepsilon_s \sim 1\%$. Therefore, sequencing errors can be overcome as long as the sequence coverage is high enough.

AbScan differentiates true mutations from sequencing errors using the same idea. In ideal case, any genes conveying mutations are regarded as separate genes and the graph maintains separate paths. However, if two genes are separated only by a few polymorphisms, then the graph may merge some nodes in paths. For SdB graphs, an exact match requirement for the r-mer would result

in a bulge where one collection of r-mers carry the mutation, and the other collection contains r-mers carrying the reference nucleotide. In the case of a true mutation, these bulge would be well-supported by reads, and not removed during error correction.

FASTA conversion. To use the SdB graph to construct a FASTA database, we associated a sequence with each node. The sequence of the source is the r-mer; the sequence associated with the sink is the last nucleotide of its r-mer concatenated with its ℓ-mer. For all non-source, non-sink nodes, the associated sequence is simply the last nucleotide of the r-mer. The sequence of a path in the graph is the concatenation of sequences associated with nodes on the path. A compact FASTA database is constructed from the SdB graph by enumerating the paths as described. The sequences in the path were converted to the amino-acid FASTA format to generate a database for the MS/MS database search tools, using the SpliceDB tool [132] for this conversion. 69.3MB of FASTA form amino acid database was created, concentrating on the antibody sequence generated from 162.7GB of RNA-seq bam files.

Multi-stage search. The antibody database adds some noise to the search and it is possible that a PSM to a known peptide has a better score against an antibody peptide, leading to false identification. As a conservative strategy to avoid false identifications, we use a modified multi-stage search [134]. We first searched all sepctra using MS-GF+ against a known protein database (Ensembl version GRCh38) [20]. All PSMs identified as a non-Ig known peptide from the spectrum level 1% FDR search of the Ensembl database were excluded from the second search. Spectra that could not be matched were searched using MS-GF+ against the antibody database, using a target-decoy strategy with 1% spectrum level FDR.

Comparison to rnaSPAdes. As transcriptome assembly is a well established research area, we built database using a popular transcriptome assembly tool, rnaSPAdes[9] to compare with AbScan. To make a fair comparison, we applied the identical read set for assembly. SPAdes version 3.9.0 was used with options "–only-assembler" and "–rna". The output nucleotide sequence translated to FASTA form amino-acid sequence for MS/MS search.

Identifying Antibody peptide location. For all identified antibody PSMs, we found the most likely position in the antibody structure. To do this, we recovered the nucleotide sequence of the peptide from the SdB graph, then compared it to sequences with the IMGT sequence to find the best matched position of each PSM to IMGT reference sequences. Finally, we incorporated gaps to the position using IMGT multiple sequence alignments to get a normalized position. Fig. 4.1 shows the expected position of the peptides we identified from the colorectal tumor MS/MS data and polyclonal antibody MS/MS data. Each horizontal black line represents the distinct peptide sequence. Peptides that do not map to IMGT reference sequences are not displayed.

Statistical test for antibody enrichment. The two stage search resulted in PSM identifications with spectra matching 'known peptides' and antibody peptides (SdB database). If in some sample, the MS/MS data was known to not contain any antibody peptide (e.g., cell-line), then any PSM in the SdB database corresponds to a false identification. The number of false identifications is expected to grow linearly with the number of known peptide identifications. Therefore, we considered the fraction

$$\frac{\text{\# PSMs in SdB database}}{\text{\# PSMs in known peptide database}}$$

for all MS/MS data, and considered the Null hypothesis that this fraction was constant in all cases, colorectal tumor, colorectal normal, and colorectal cell-lines. To calculate the p-values, we applied Pearson's χ^2 test in a 2×2 contingency table.

$$\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$$

where

$$\chi^2 = \text{Pearson's cumulative test statistic,}$$

$$O_i = \text{Number of observation of type i}$$

$$E_i = \text{Theoretical frequency of type i}$$

The p-value was calculated from the χ^2 distribution table.

Antibody and SAAV peptides correlation test. Consider a table, where columns correspond to samples (each column is a different sample), and rows correspond either to SAAV peptides (possible antigens) or to antibody peptides. The cells mark the presence or absence of the peptides in the specific sample. For any pair of antibody and SAAV peptides, we used the Fisher's exact test to measure correlation of occurrence. As a large number of pairs were used, we used a target-decoy based approach to compute the false discovery rate for any p-value cut-off.

For each row, the columns were permuted independently so that any correlation between two rows (an antibody peptide, SAAV peptide pair) was just by chance, and a Fisher exact test was used to compute the correlation between all pairs. Highly correlated pairs of antibody and SAAV peptides were identified by applying a 5% FDR threshold.

Measuring immune response. Measurement of the immune response for each individual was accomplished by counting the number of antibody PSMs. As the total number of spectra and their quality were not identical for every sample, the antibody PSMs were normalized by the total number of PSMs to the 'known database' search. Fig. 4.2(c) shows the distribution of the normalized immune response of each individual in both tumor and normal samples. We simply took the top 45% and bottom 45% group of individuals in terms of their normalized immune response.

Survival rate comparison. We designed a method that takes a collection of peptides, and samples, groups samples based on co-occurrence of peptides, and tests if the individuals groups have different survival times. Specifically, we used the following strategy:

1. Represent each peptide p as a binary vector $\mathbf{p}$ over all samples with $\mathbf{p}_i = 1$ (respectively, $\mathbf{p}_i = 0$) indicating the presence (respectively, absence) of peptide p in sample i.
2. Cluster the peptide vectors into two groups (arbitrarily labeled $+$, $-$) using 2-means clustering.
3. Assign score S_i to each sample i using:

   $S_i = (\text{Number of '+' assigned peptides in sample } i)$

– (Number of '−' assigned peptides in sample i)

4. Pick two sets of samples: Bottom (45%) of all samples with the lowest score, and top (45%) with the highest score.
5. Perform the Kaplan Meier log-rank test on the two groups of samples to test for correlation with clinical outcome.

We performed this test using all antibody and mutated peptides that significantly co-occurred in the samples exceeding a 5% FDR threshold of correlation test. The test statistic could include some unknown bias, and it wasn't clear if they followed the χ^2 distribution used to compute a p-value. To test this, we set two groups of patients where each group included 45% of random samples without replacement, and then we calculated the test statistics of two groups of random patients by log-rank test. We repeated the process $10,000$ times to create the distribution of test statistics (Supplemental Fig. D.5).

4.4 Results

Analytical comparison of SdB and dB graphs. We compared the performance of SdB graphs versus dB graphs using both analytical methods as well as empirical data from simulations. Let p_s denote the probability that a randomly chosen pair of nucleotides are identical. Thus, the probability of a false k-mer match is p_s^k. To allow for fair comparisons, parameters r, ℓ, k were selected so that the probability of an (r, ℓ) match between unrelated reads in a SdB graph is the same as the probability of a k-mer match in dB graph. Specifically (Supplementary Methods–'Analytical comparison of SdB and dB graphs'),

$$p_s^k \geq p_s^{r+\ell} \cdot \left(1 + \ell \frac{1-p_s}{p_s}\right), \tag{4.1}$$

$$k \leq r + \ell + \log_{p_s}\left(1 + \ell \frac{1-p_s}{p_s}\right). \tag{4.2}$$

For any r, ℓ, we chose k to be the largest value satisfying constraint 4.2. We also computed the probability of false overlaps. (See Supplemental Methods–'Comparison between the SdB and dB graph mathematically'). Using these calculations, we can show that SdB graphs have significantly lower false negative rates compared to dB graphs. For example, let $p_s = \frac{1}{4}$. When $r = 10$, $\ell = 20$, the choice of $k = 27$ equated the false overlap rates for both methods at $5.55 \cdot 10^{-17}$. However, for an overlap of 40 bp, $\varepsilon = 0.01$, we computed a false negative rate of 13.9% for the dB graphs versus a rate of 2.2% for the SdB graphs.

To test these theoretical results, data was simulated by generating the $100,000$ overlapping regions, of length from 30 to 100, with uniform sequencing error rate ε. Reads were connected by a path in the dB graph, if there was at least one k-mer consecutive sequence without an error. Similarly, they were connected by a path in a SdB graph, if there was an $(r+\ell)$-mer in which the first r nucleotides had no error and the following ℓ-mer had at most one mismatch. Supplemental Fig. D.6 showed a complete concordance between theoretical and simulated results. The sensitivity for all methods increases with length of overlap and decreases with higher ε. SdB graphs consistently outperform dB graphs.

Comparison on simulated Antibody reads. To provide a more direct comparison of the performance of SdB graphs and dB graphs on Ig sequences, we employed a second simulation, starting with a single IMGT reference antibody sequence denoted by $\mathcal{A}$. Note that an antibody (Supplemental Fig. D.7) is a 'Y' shape protein consist of a variable region and constant region. The variable region is formed by selecting a gene from each of 3 sets V, D, and J which are brought together by recombination and splicing. The combined variable region itself can be divided into a framework (FR) which is relatively constant, and three hypervariable complementarity determining regions (CDRs; Supplemental Fig D.7) [98]. In the simulation, $\mathcal{A}$ was created by joining known V, D, and J regions (IGHV1-18*01, IGHD1-1*01, IGHJ1*01; [103]; Supplemental Fig. D.8). We generated a collection D of decoy sequences in which each nucleotide was chosen uniformly at random, except for the insertion (at a random position) of a single sub-string of $\mathcal{A}$. The insertions were of varying lengths ranging from 20 to 26. The antibody reference $\mathcal{A}$ and decoy gene sequence collection D

were used as a template to simulate reads, using the tool *wgsim* (https://github.com/lh3/wgsim), with sequence error rate set at $\varepsilon = 1\%$. A dB graph and an SdB graph was built using these reads to measure the false positive and false negative results from these graphs.

Let $G = (V, E)$ denote the dB or SdB graph, depending on context, while the graph $G_A = (V_{\mathcal{A}}, E_{\mathcal{A}})$ is constructed solely using $\mathcal{A}$. In the ideal scenario, G and $G_{\mathcal{A}}$ should be identical. Therefore (Supplementary Methods), the false negative rate (denoted by $\mathcal{F}$) can be estimated using

$$\mathcal{F} = \frac{|E - E_{\mathcal{A}}|}{|E_{\mathcal{A}}|}$$

The false positive rate was measured indirectly, using divergence (Supplementary Methods):

$$\mathcal{D} = \frac{\Sigma_{n \in V_{\mathcal{A}}}((n_i - 1) + (n_o - 1))}{|E_{\mathcal{A}}|},$$

where n_i is the in-degree and n_o is the out-degree of node $n \in V_{\mathcal{A}}$. The divergence provides a measure of false connections for the antibody sequence $\mathcal{A}$.

Note that false positive edges can also arise due to sequencing errors. However, most dB construction corrects for such errors by choosing an appropriate threshold for coverage, along with other methods [40, 85, 130]. However, the appropriate threshold is different for each value of coverage. We chose a principled method for choosing coverage to remove false positive edges due to sequencing errors for both dB, and SdB. Subsequent to coverage filtering, the false negative rate and divergence was measured as a function of increased coverage (Fig. D.9).

Fig. D.9 shows an explicit tradeoff between false negatives, and divergence (false positives) for dB graph methods. At any specific fixed coverage parameter(e.g. $10\times$), the false negative rate of the dB graph increases with increasing values of k, even as divergence decreases, making it difficult to simultaneously improve both metrics. In contrast, SdB graphs show consistently lower divergence and false negatives for all coverage values.

Read filtering. Before we construct the split de Bruijn graph, we need to collect the reads that encode Ig gene transcripts (See Experimental Procedures – 'Read filter'). We tested the quality of read filtering by a partial alignment of filtered reads to the reference antibody sequences. A virtual antibody reference was set to represent the variable regions of all antibodies. We adjusted the gap between this virtual antibody and each individual antibody using the IMGT antibody reference with gap. The matching k-mer was used to anchor the alignment, and the extent of the alignment was determined simply by the length of read on each side of the k-mer. The anchored position of the read was transferred to virtual antibody position and used to estimate the overall coverage. We counted the number of reads passing through each unique position of the virtual antibody. Supplemental Fig. D.10 describes a coverage due to the partial alignment of all filtered reads, and shows that the reads are filtered without apparent bias except at the very end of the sequence.

MS-MS based discovery of antibody peptides. We used four mass spectrometry data-sets. To test the algorithms, a data-set of spectra acquired from a purified polyclonal antibody mixture (*antibody purified*) was used [100]. To test for antibody peptides in tumor samples, we used a collection of MS/MS spectra from 90 distinct colorectal tumor samples from the CPTAC project [139] (*colorectal tumor*). As negative control, we used spectra acquired from 30 normal colon biopsies [139] (*colorectal normal*). As a second control, we used spectra from colon cancer cell-lines LIM1215, LIM1899, and LIM2405 (denoted as *colon cell-lines*). [32].

SdB and dB graphs were designed and implemented, utilizing 162.7GB RNA-seq reads of 90 individuals downloaded from The Cancer Genome Atlas (TCGA) repository[76]. The two approaches resulted in a 69.3MB and 107.8MB FASTA-formatted amino-acid database. A multi-stage search (See Experimental Procedures –'Multi-stage search') using known proteins and SdB graphs (respectively, known proteins and dB graphs) was conducted to identify peptide spectrum matches (PSMs). A summary of the results of those searches are presented in Table 4.1. The list of identified spectra and other details are presented in Supplemental Table 1 – 'Link to the list of PSM and spectrum image'.

Note that the antibody-purified data-set presents an interesting challenge as the SdB graph

was constructed from RNA of completely different individuals. Even so, our search identified 16,404 antibody PSMs (3,167 peptides) out of 116,018 total spectra (PSM identification rate 14%). Fig. 4.1 (a) shows that the identified peptides cover the entire space of the antibody. Table 4.1 also allows for a comparison of the SdB graph and dB graph databases, as both use the same read set as their input. At identical FDR cut-off (1%), SdB graphs identify 3.3× as many PSMs as the dB for the colorectal tumor, and 1.7× as many PSMs for antibody purified data-set, consistent with simulation results. On the other hand, the number of PSMs identified in the colorectal normal, and cell-line colorectal samples are similar, validating the proposition that SdB graphs can filter out erroneous PSMs at the same rate as dB graphs. Therefore, SdB graphs reduce both false positives and false negatives in the real data, identifying more true PSMs without increasing false PSMs.

In the sample matched colorectal tumor spectra, 54,909 PSMs (1,940 peptides) were identified. We asked if these large numbers of antibody peptides originated from tumor infiltrating lymphocytes, or from other sources. For example, these immunoglobulin identifications could simply correspond to floating antibodies from blood contamination, or they could be mis-identified (modified) peptides. In the first case, we would expect to see similar numbers of antibody peptides in colorectal tumor and colorectal normal data set. In the second case, we would expect to see similar numbers of antibody peptides in colorectal cancer, and colon cell-lines (Fig. 4.2(a)).

We normalized PSM counts to the number of PSMs in the Known DB before comparing across samples (Fig. 4.2(b)). The normalized PSM count in the colorectal normal data-set was only 4.69% of the colorectal tumor counts (p-value < 0.0001; See Experimental Procedures –'Statistical test for antibody enrichment'). The normalized PSM count in colon cell-lines was 0 consistent with the observation that TILs were the source of the antibody peptides observed in the colorectal tumor. While it is likely that the actual numbers would depend upon experimental handling of tumor interstitial fluid (TIF), the tumor and normal cells were processed in an identical fashon, and would have similar biases in terms of TIF handling. We additionally tested the samples for presence of biomarkers, and identified 113 PSMs matching CD38 in tumour samples compared to 1 PSM in normal samples. Similarly, we observed CD74 predominantly in tumour samples (684 PSMs versus

34 PSMs). These represent significant enrichment even after accounting for the 3x larger number of tumor samples. CD38 is a glycoprotein found on the surface of many immune cells including CD4+, CD8+, B lymphocytes and natural killer cells [82, 83], while CD74 has been reported to possibly reflect an intratumoural immune response with TIL association [129].

The assembly of RNA-seq reads is a well established research area of genomics [86, 40, 85]. However, genome assembly tools are designed to be general, and may not do a good job of assembling Ig genes. As the reconstruction of Ig genes from the RNA-seq reads was a key part of our pipeline, we asked if the use of the RNA assembly tools could provide better results. To test this, a popular transcriptome assembly tool, rnaSPAdes[9], was used to assemble RNA-seq reads from one colorectal tumor sample. We searched the MS/MS data from the same sample against databases constructed using rnaSPAdes and SdB graphs. The number of spectra identified using the SdB graph method was $2,450$, compared to 528 using rnaSPAdes, suggesting that general purpose transcript assembly tools were not suitable for studying the antibody repertoire at the protein level.

SAAV discovery. We used the SAAV peptides from the results of previous studies [134, 130], but with additional filtering. We remove all peptides where the mutation has a mass difference of one. We also enumerate all reference peptides with common modifications that shared some sequence tag with the mutated peptide and scored them to see if a reference peptide could better explain the data. The final list of mutated peptides is presented in Supplemental Table 3, and the annotated mass spectra are in MassIVE, and link is provided in Supplemental Table 1. The filtered list contains 677 SAAV peptides.

Antibody peptide-SAAV peptide correlation. We asked if the antibody peptides discovered in the colorectal tumor dataset could be targeting specific neo-antigens. The neo-antigens are possibly mutated peptides that are recognized by TILs and antibodies. Many somatically mutated peptides had been detected in the colorectal tumor data in the original seminal study[139] and our own groups re-analysis [130]. Supplemental Table 4 shows the occurrence of mutated non-reference peptides, and all antibody peptides in each of the 90 samples. As peptides that are polymorphic

in the population could still be somatic in individuals, and some polymorphisms are known to be functionally deleterious, we used all mutated non-reference peptides.

For every antibody peptide-SAAV peptide pair in this table, a calculation was made to determine the significance of co-occurrence using the Fisher exact test. Since many pairs were to be tested, a target-decoy approach was used to compute the false discovery rate for significant pairs. The decoy statistics were computed by permuting the occurrence of each peptide in the sample. Fig. 4.3 shows the distribution of p-values computed from the target and the decoy table. At a nominal p-value threshold of 0.00025, we see 163 pairs that exceed this threshold, versus 5 decoy pairs, suggesting a small false discovery rate of $\leq 5\%$.

One example of these co-occurring pairs is the the antibody peptide NTLYLQMDSLR, and SAAV peptide AAQAQGQSCEYSLMVGYQCGQVF(Q→R). The antibody peptide NTLYLQMDSLR belongs to variable region of IGHV3-64D*06 and the mutated peptide *pep*= AAQAQGQSCEYSLMVGYQCGQVF(Q→R) belongs to the gene FBLN1 reported to be down-regulated in colorectal cancer cells [136]. Among 90 samples, both peptides are expressed in 26 samples, and neither is found in 42 samples, giving a Fisher exact test p-value of 2.59×10^{-6}. Supplemental Fig. D.11 shows examples of peptide spectrum matches of these peptides. The mutation in peptide *pep* a known polymorphism (dbSNP rsID136730). However, the mutation is very low frequency in normal population surveys 0.14% in ExAC, and 0.04% in 1000 Genomes project[106] compared to its occurrence in 34 out of 90 samples. It is also known that non-somatic, self peptides can elicit an immune response against tumor cells [44]. Therefore, the functional relevance of *pep* cannot be rejected based solely on its classification as (non-)somatic. Finally, it is important to assert that co-occurence does not indicate co-occurence only between the specific antibody peptide and the SAAV, but rather between the antibody carrying NTLYLQMDSLR and some peptide in the mutated version of FBLN1 product. In fact we see another antibody peptide LSCAASGFSFR in the FR1/CDR1 region that also co-occurs with *pep* (p-value:9.93×10^{-5}). Therefore, we did not filter antibody peptides by their location (CDR/FR) before testing for correlation.

We also used both co-occurence and co-absence to test correlation. While 'absence' of

a peptide may be due to experimental protocol, co-absence is also indicative of a correlation. As an extreme example, the pair of antibody peptides NGPSVFPLAPSSK and mutated peptide AGRPVICATQMLESMIK were observed in 29 samples and 28 samples, respectively. If they had no correlation, then over the 90 samples, we would expect 9 samples to carry them both, just by chance. Instead we see zero (p-value $1.44 \cdot 10^{-6}$). This suggests that the existence of this mutated peptide (perhaps indirectly) reduced the affinity to this specific antibody leading to a negative correlation, and the effect was independent of the event that we missed identifying the peptides.

Correlation between antibody expression and survival status. The antibody peptide repertoire might provide a snapshot of the immune response to cancer. We anticipated that the patients with higher immune response could have a different clinical outcomes than those with lower immune response because of the role of TILs in mediating response to cancer[120, 140, 101].

We first measured the immune response of an individual as the fraction of identified peptides that came from the antibody repertoire, and identified a subset of individuals as high-responders and low-responders (See Experimental Procedures –'Measuring immune response', and Fig. 4.2(c)). We used the days-to-death values to get the Kaplan-Meier survival estimator for the two groups. Next, we used a log-rank test to compute a p-value for the difference between the two curves. The p-value was 0.75, indicating that we could not reject the Null hypothesis (Supplemental Fig. D.12).

We also considered the possibility that some, but not all peptides mediate a positive clinical outcome. Further, these peptides would be expressed in multiple individuals with similar outcomes. To test this hypothesis, we designed a method that takes any group of peptides, and clusters samples based on co-expression, but without knowledge of the clinical outcome in the individuals (See Experimental Procedures –'Survival rate comparison'). For a given collection of peptides, we tested the null hypothesis that there is no correlation between sample grouping and the clinical outcome.

We computed an empirical null distribution by choosing random subsets of individuals, and performing the log-rank test against clinical outcome. Supplemental Fig. D.5 shows that the test statistic under null hypothesis closely follows the theoretical χ^2 distribution.

In contrast, when we tested sample grouping using the correlated antibody, SAAV peptide

pairs (See Experimental Procedures –'Antibody and SAAV peptides correlation test'), we observed a significant differential response with p-value: 0.032 (Fig.4.4 (a)). We also tested this method using two other groups of peptides. When we used all antibody peptides we also obtained a differential response with p-value 0.040. (Fig.4.4 (b)). However, testing with all mutated peptides, we did not observe significant differential response, obtaining a p-value of 0.522 (Fig.4.4 (c)). The small number of samples implies that our study is not fully powered and the results need to be replicated in larger cohorts. Nevertheless, they do show that antibody expression could be correlated with the clinical outcomes.

4.5 Discussion and future study

Understanding the immune response to cancer is key to cancer immunotherapy. Current approaches use serum or plasma samples and specifically focus on isolating differentiated B-cells for analyzing antibodies. However, the serum antibody repertoire may contain a larger pool of antibody sequences, not just the ones responding to tumor neo-antigens. In this paper, we mined spectra acquired from isolated (colorectal) tumor cells, and identified a large number of antibody peptides. Our results suggest that infiltrating lymphocytes in the tumors generate antibodies in response to the tumor. They also suggest that somatic coding mutations in the tumor genome act as neoantigens triggering antibody generation. We observed recurrence of antibody and mutated peptide sequences that cannot be explained as chance events, and showed a positive association between clinical outcome (survival time), and the antibody response. Together, the results underscore the need for systematic analysis of the tumor antibody repertoire.

The identification of antibody peptides using tandem mass spectrometry is technically challenging. In the ideal case, the spectra should be searched against transcript data from differentiated B-cells from the same individual. However, that data may not always be available. Moreover, it is not known if circulating B-cells have the same antibody repertoire as the tumor infiltrating lymphocytes. In this paper, we used RNA-seq data, not from isolated B-cells, but from the same

tissue that the proteome was extracted. Nevertheless, we managed to get significant coverage of antibody peptides. We identified a large number of peptides even when we used MS data from unmatched samples. Future research will focus on the differences between different sequencing approaches, such as IG-seq, and RNA-seq.

The hyper-variability of antibody sequences makes it challenging to construct databases that can be searched with MS spectra. We proposed a new structure, called the SdB graph, and showed improved performance in compressing and creating MS-searchable databases relative the dB graphs. The SdB graphs are later converted into Fasta formatted databases that can be used for search with any tool. The software for developing SdB graph should be generally applicable for any hypervariable region, and is available for download. These techniques described here can be further improved and those will be the focus of future research.

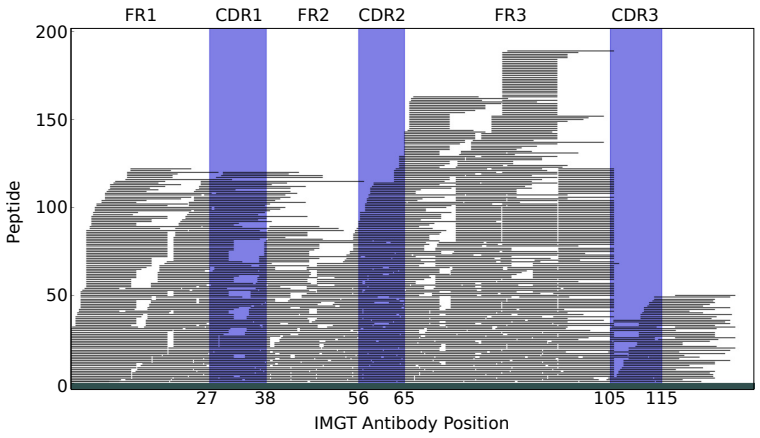
We found that the SdB graph database generated from RNA-seq of TCGA tumor samples was also helpful in identifying antibodies from completely different samples. This raises the possibility that multiple RNA-seq samples from a specific tumor type could be used as a universal database, reducing the need for matched RNA and protein samples for decoding the immune repertoire. This will be explored in future work. At the end, we also hope that our preliminary results spurs a further investigation of the clinical outcome based on immune system response, and the development of diagnostic tools and therapies that can emerge from an analysis of the tumor immune repertoire.

ACKNOWLEDGEMENTS

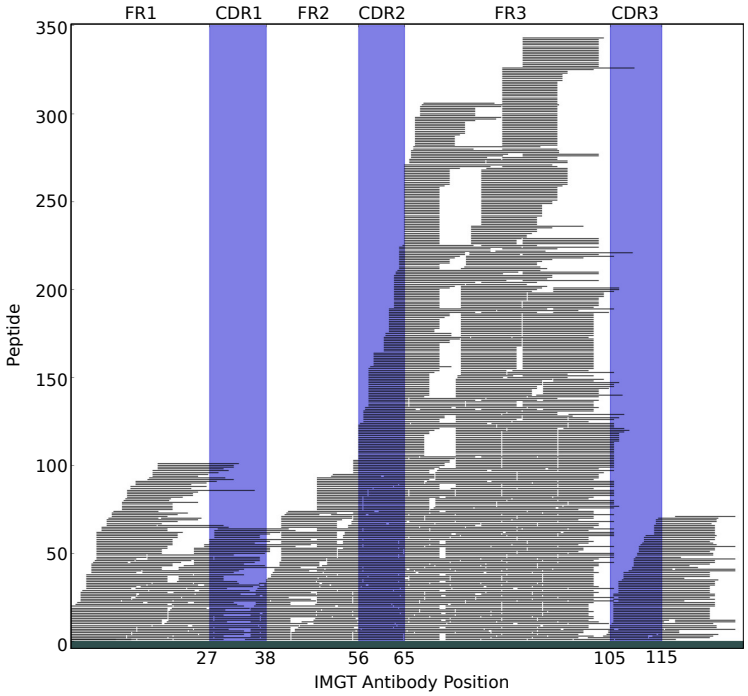
Chapter 4, in full, is a reformatted reprint of the material as it appears in Molecular and Cellular Proteomics, Dec; 16(12):2111-2124, 2017. “The Antibody Repertoire of Colorectal Cancer.” Seong Won Cha; Stefano Bonissone; Seungjin Na; Pavel A. Pevzner; and Vineet Bafna. The thesis author was a primary investigator and author of this paper.

Table 4.1: Number of identified PSM(peptides)

Data set	SdB graph DB	dB graph DB	Known DB
Cell line	0(0)	0(0)	117,679(14,527)
Normal	711(113)	700(96)	1,705,785(85,956)
Tumor	54,909(1,940)	16,364(1,088)	5,573,094(129,886)
IG purified	16,404(3,029)	9,576(2,338)	989(246)

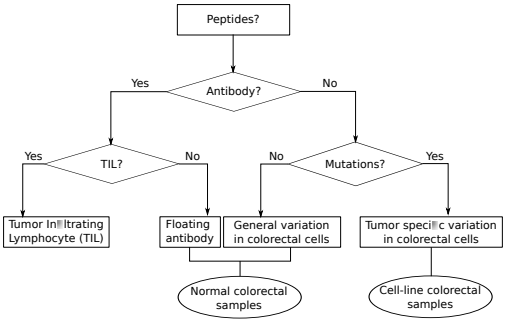


(a)

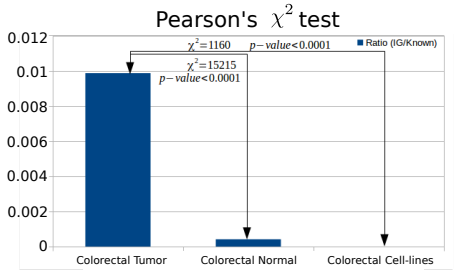


(b)

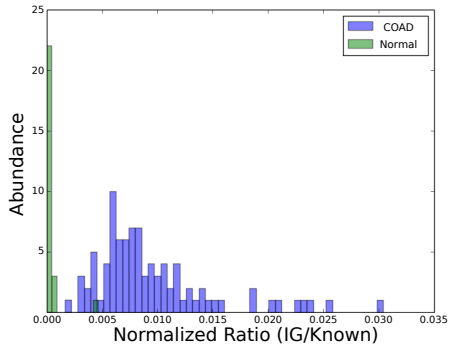
Figure 4.1: **Relative locations of identified antibody peptides.** Each horizontal black line represents a distinct peptide sequence. Trypsin was applied for the colorectal tumor MS/MS spectra assessment, and four different enzymes were applied for polyclonal antibody MS/MS spectra assessment. Both spectra sets were searched against the same antibody database constructed using tumor RNA-seq reads driven by TCGA. **(a)** Antibody PSMs from colorectal tumor MS/MS data. **(b)** Antibody PSMs from polyclonal antibody MS/MS data.



(a)



(b)



(c)

Figure 4.2: **Comparison of identified antibody PSMs per experiment and sample** (a) The source of antibody peptides in different samples. PSMs that match non-reference peptides are either mutations or antibody peptides. Antibody peptides should not be observed in cell-lines. However, floating antibodies could be observed in normal colorectal samples. Antibodies from Tumor infiltrating lymphocytes should only be observed in tumor samples. (b) Occurrence of antibody peptides in tumor, normal, and tumor derived cell-lines are significantly different for MS/MS spectra of tumor, normal, and cell-line colorectal samples. Each spectra set were searched against the Ensembl GRCh38 protein database[20] and a custom antibody database. The number of PSMs identified as antibody peptides were $54K$ (*colorectal tumor*), 711 (*colorectal normal*), and 0 (*Cell-lines*). The PSM counts were normalized against the number of PSMs to known peptides. $5.5M$ in *colorectal tumor*, $1.7M$ in *colorectal normal*, and $0.1M$ in *Cell-lines*. The normalized ratios suggest that a significantly larger fraction of the colorectal tumor PSMs are antibody peptides, compared to the other two data-sets (Pearson's χ^2 p-val $< 10^{-4}$). (c) The distribution of the number of samples carrying a normalized fraction of antibody peptides. COAD samples carry a higher fraction of antibody peptides.

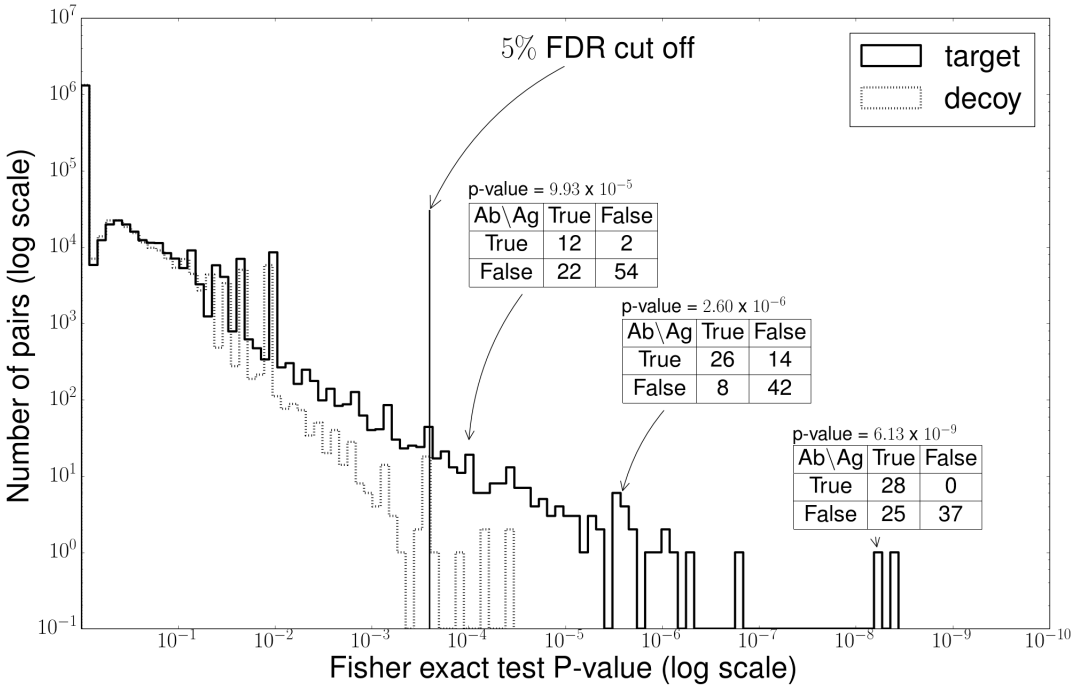
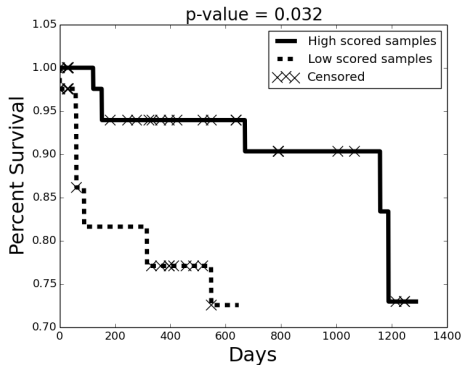
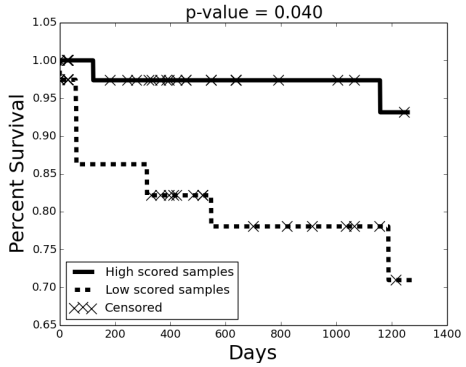


Figure 4.3

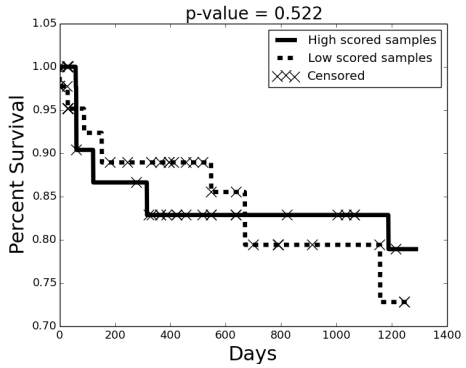
Figure 4.3: **Peptide correlation test.** We tested the correlation between the antibody peptides and mutated peptides. For every pair of peptides, we counted the number of samples co-occurring with these peptides and then we applied Fisher exact test to calculate the p-value. For example, the peptide pairs of *NTLYLQMDSLR* (antibody) and *AAQAQGQSCEYSLMVGYQCGQVF*$(Q \rightarrow R)$ (SAAV peptide) co-occurred in 26 samples, and there was a co-absence in 42 samples. It was revealed that 68 of the 90 samples shared the co-occurrence of this pair with a p-value of 2.60×10^{-6}. We drew the histogram of p-values of all pairs in Supplemental Table 4. We also drew the histogram of the p-values from the decoy table generated by the random permutation of values. A 5% FDR threshold was applied to collect the high correlated pairs.



(a)



(b)



(c)

Figure 4.4: **Kaplan-Meier survival estimator.** For any subset of peptides, we bi-partioned peptides based on co-expression in samples. Next, we scored each sample based on the homogeneity of peptides from a single partition in that sample (Methods). The highest and lowest scoring samples (one-third each) were grouped, and were tested to determine the clinical outcome. The Kaplan-Meier survival estimator and log-rank test were applied to test the difference of the clinical outcome of two groups. When testing with co-occurring mutated peptide/antibody peptide pairs, we observed a significant correlation with survival (Plot (a): p-value $= 0.032$). In contrast, the correlation was reduced when testing with only antibody peptides (Plot (b): p-value $= 0.040$), and there was no-correlation when testing with mutated peptides. (Plot (c): p-value $= 0.522$).

Appendix A

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

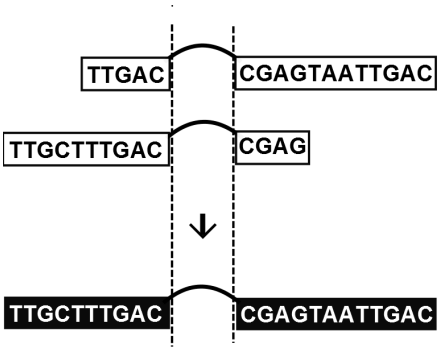


Figure A.1: In filtering stage, RNA-seq reads that have identical splice junctions are merged, and extended in both ends

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

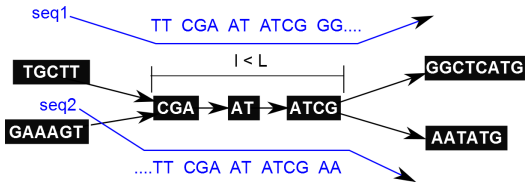
Number of components(G)	116355
Number of nodes	652936
Number of edges	337648
Average node length	57.00 bp
Average number of edges per node	0.44

A.1 Comparison with other gene prediction methods

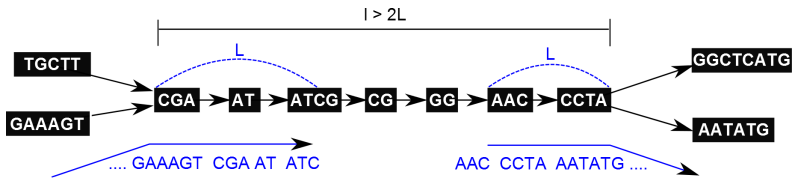
We compared the list of identified novel peptides using our proteogenomics pipeline against various gene prediction results provided from other groups using identical RNA-seq data-sets. The file of gene predictions is available [73], and includes GeneFinder [73], single exon gene predictions, predictions 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).

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

Prediction methods [73]	# of overlapped sequence
GeneFinder	286
Conserved *C. briggsae* ORFs	364
RNA-seq data	543
Single exon gene predictions	1



(a) $\ell < L$



(b) $l \geq 2L$

Figure A.2: Combining 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 ℓ ($L \leq \ell < 2L$) sequence in between the two. (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

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

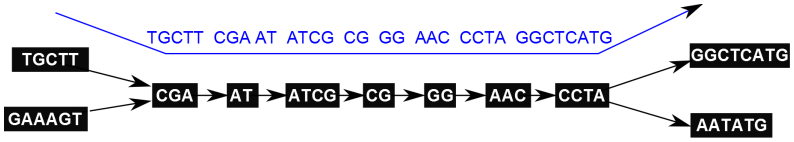
The CIGAR string of the SAM format file is used to determine splice junctions. For example, consider a match, starting at coordinate x, with the accompanying CIGAR string given by '35M1000N35M' which is translated to "match 35bp", "skip 1000bp", followed by "match 35bp". We convert this string to two GFF lines, denoting the intervals $[x, x+35]$ and $[x+1035, x+1070]$. In this example, $[x+35, x+1035]$ represents a splice junction.

A.3 Detailed RNA-seq methods

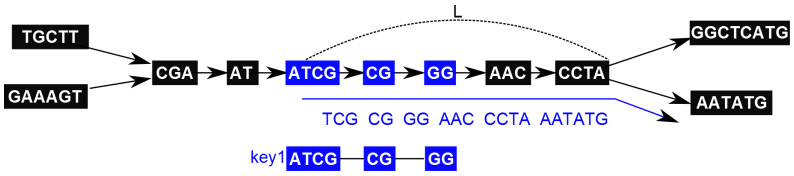
RNA-seq Alignment methods:

Step 1:

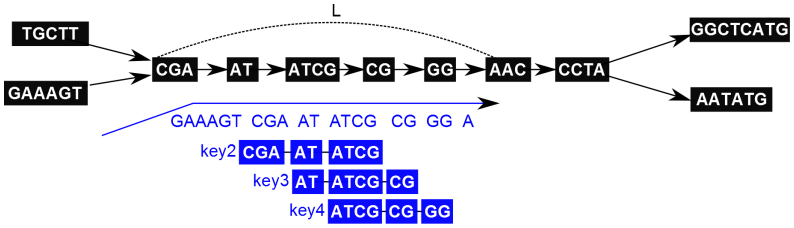
1. identify all reads beginning with at least 4 TTs



(a) Sequence generated from the first visited nodes



(b) Sequence generated from split nodes



(c) Sequence generated from merge nodes

Figure A.3: Illustration of hashing technique to rapidly identify overlap-node-pairs. (a) For the first visited node path from a start to an end node, the generated sequence is the full path from the corresponding start to end node. This full path cannot be merged with others. (b) In 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 hash the prefix string using the first 3 nodes as key(key1), so that each key contains the list of the paths such that prefix of the paths same as the corresponding key. (c) Every time a merge node is encountered in the DFS, we traverse the subsequent path, querying the hash table continuously using 3 node triplets(key2, key3, and key4) to query the hash table. When a match is found (key4 and key1), the hash table returns a list of sequences that corresponding paths starting with the appropriate key. ('TCG'+'CG'+'GG'+'AAC'+'CCTA'+'AATATG'). We search each sequence within the returned sequences, using remaining suffix of the queried sequence. In our example, the remaining sequence is 'A' which appears right after key4. We merge the matched sequence with queried sequence and output to a FASTA file.

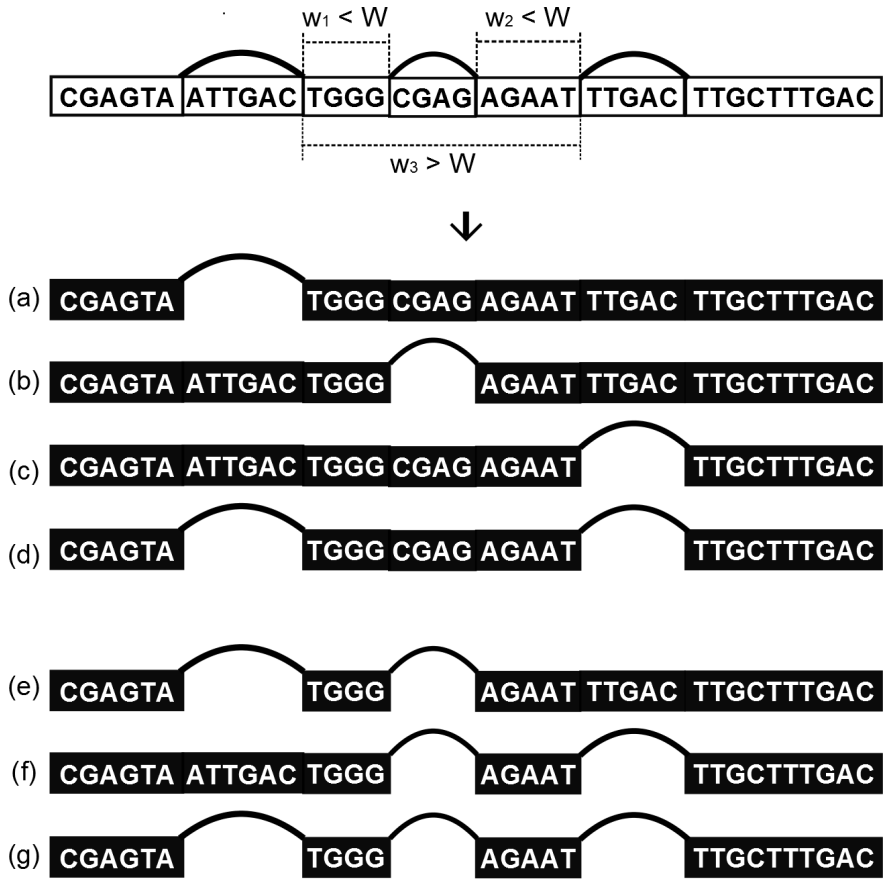


Figure A.4: Description of parameter W. In this example, W is set to 10bp. From the splice graph all possible combinations of the resulting sequences, considering all splicings, total 7 as shown above ((a) through (g)). If multiple spice edges exist within Wbp, and only when the corresponding node has a following consecutive node, then the splice path will be ignored. As a result, (a), (b), (c), and (d), are converted and expressed to the FASTA file. On the other hand, (e), (f), and (g) are discarded.

```
Query  1375  NLVTPLFGILIRARCYRYIIVSDIEKAFHQVRLQKAFRNVTQFLWIQDPSKPTVEDNLCR
             N++TP+FGIL+R R    I+V+DIEKAFHQVRLQ  FRNVT FLW++D + P   DN+
Sbjct  1383  NMITPIFGILVRVRFPPIIVVADIEKAFHQVRLQPEFRNVTMFLWLKDVTAPATADNIQV
```

Figure A.5: Alignment result of novel gene example. The highlighted region corresponds to the alignment of identified peptide 'R.CYRYIIVSDIEKAFHQVRLQKAFR.N' against the sequence of hypothetical protein CRE_ 09558 [Caenorhabditis remanei].

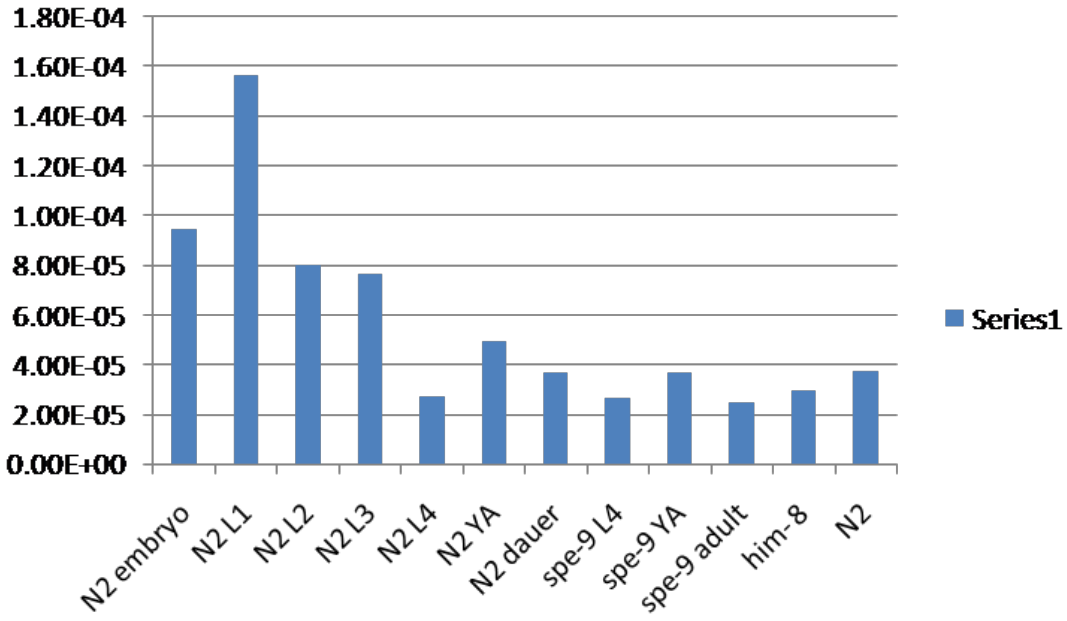


Figure A.6: Translated UTR spectral counts throughout different developmental stages

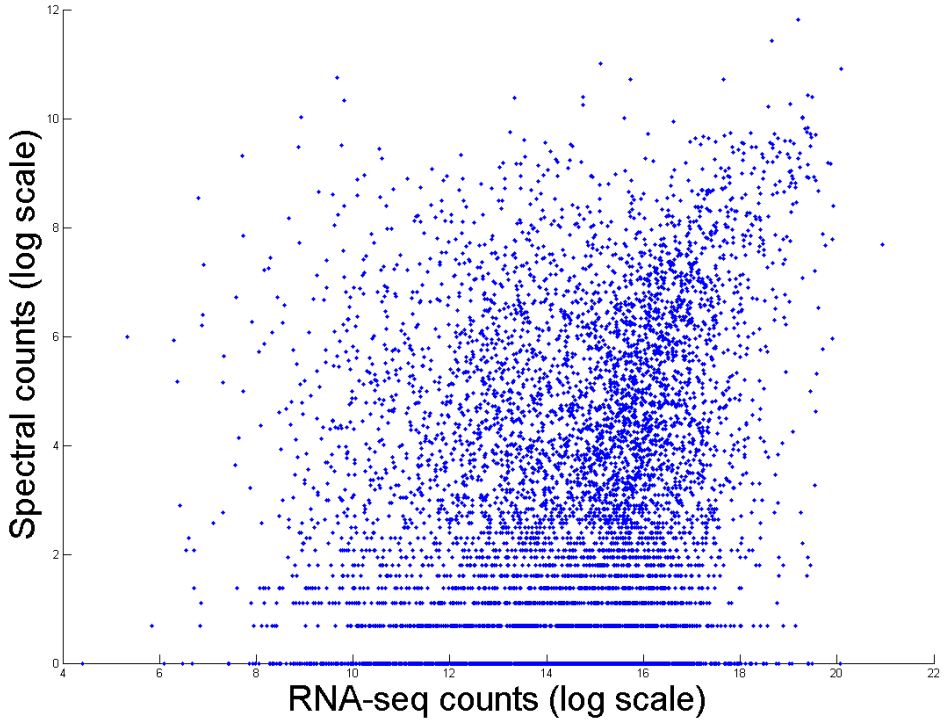


Figure A.7: RNA vs peptide transcription level

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<=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 (<=5 bases unaligned from either end).

Step 4:

For reads still not placed, look for multi-segment alignments from bwasw 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 (¡=5 bases unaligned from either end).

A.4 Comparison of spectra dataset used in this study with Merrihew *et al*. (2008) [73]

The Merrihew *et al*. (2008) [73] 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_α.

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 $\mid S_2 \mid = \mid S_1 \mid$. 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 sequnce in G. So there is no element x such that $x \in S_3(l), x \notin S_2(l)$.

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. DFŚ: An algorithm same as DFS, but stop when it evumerate 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 retrived, 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, DFSśfetrieve every input edges of all merge nodes.

If Step 1 and Step 2 are true, DFSánd 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 $e_1, e_2, ..., e_k$.

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 bijenction for any element from E_i to E_j which shares the suffix. Keep E_1, and for $E_2, ..., E_k$, 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 DFSfrom 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.

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

A.6.3 Rule 3:

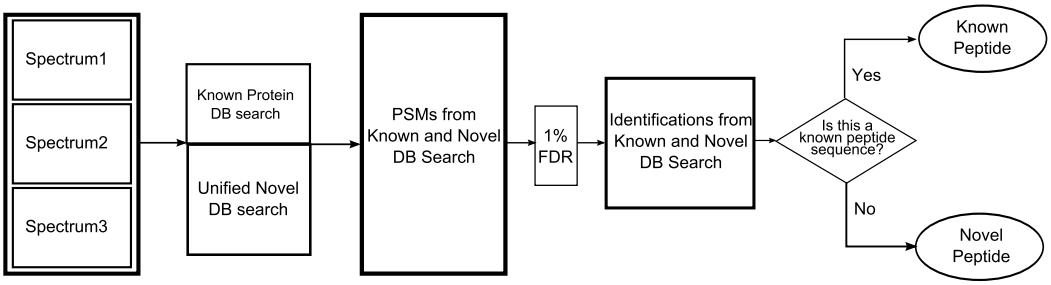
Candidate pair set contains every possible prefix-suffix 'coordinate' overlap which shares the sequence, so combining prefix-suffix pair in candidate pair set implements Rule 3.

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 then Edwards and Lippert [27] where they merge all

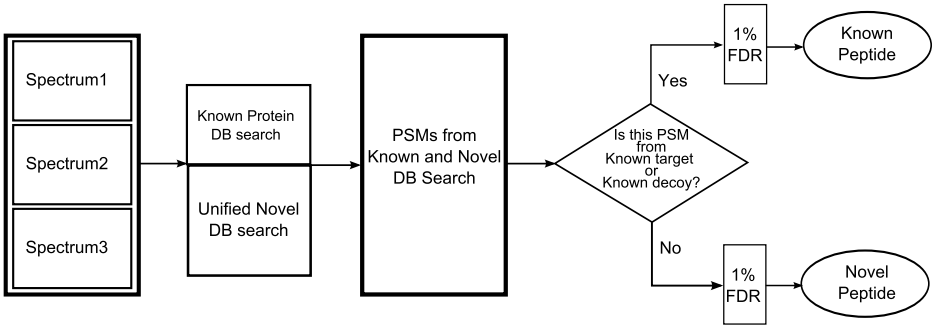
overlapping sequences (less than a certain parameter), and do not consider coordinate information. Again, unlike Edwards and Lippert [27], our method uses a genomic coordinate-based data structure(represented in base pairs) rather than minimizing the amino acid sequence overlap. We claim that for proteogenomic analysis the coordinate based approach is more appropriate since it can easily reconstruct the original genomic coordinate of the identified peptide.

Appendix B

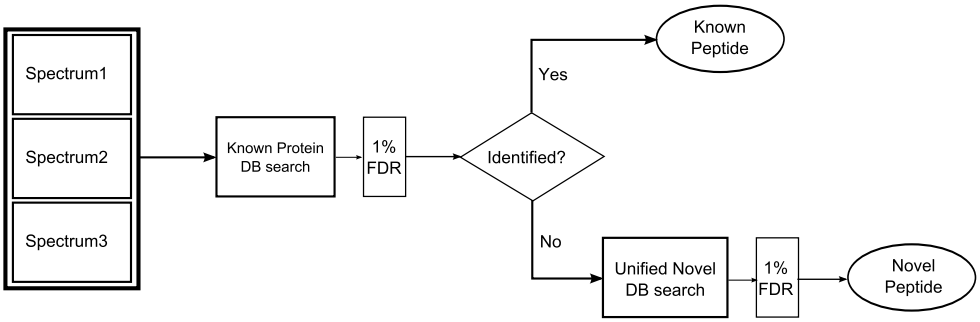
Appendix: Proteogenomic strategies for identification of aberrant cancer peptides using large-scale Next Generation Sequencing data



(a) Combined-FDR



(b) Separate-FDR



(c) Two-Stage-FDR

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

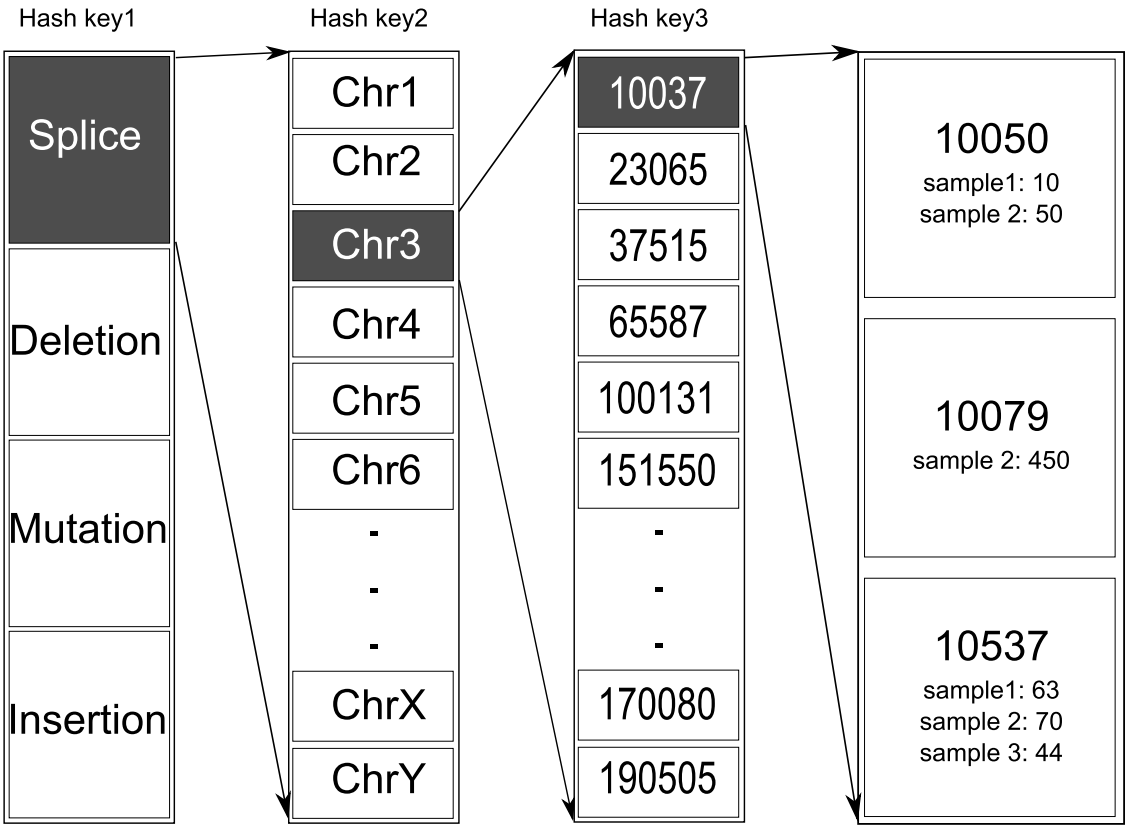


Figure B.2: Structure of hash table for accessing the original RNA-seq meta information

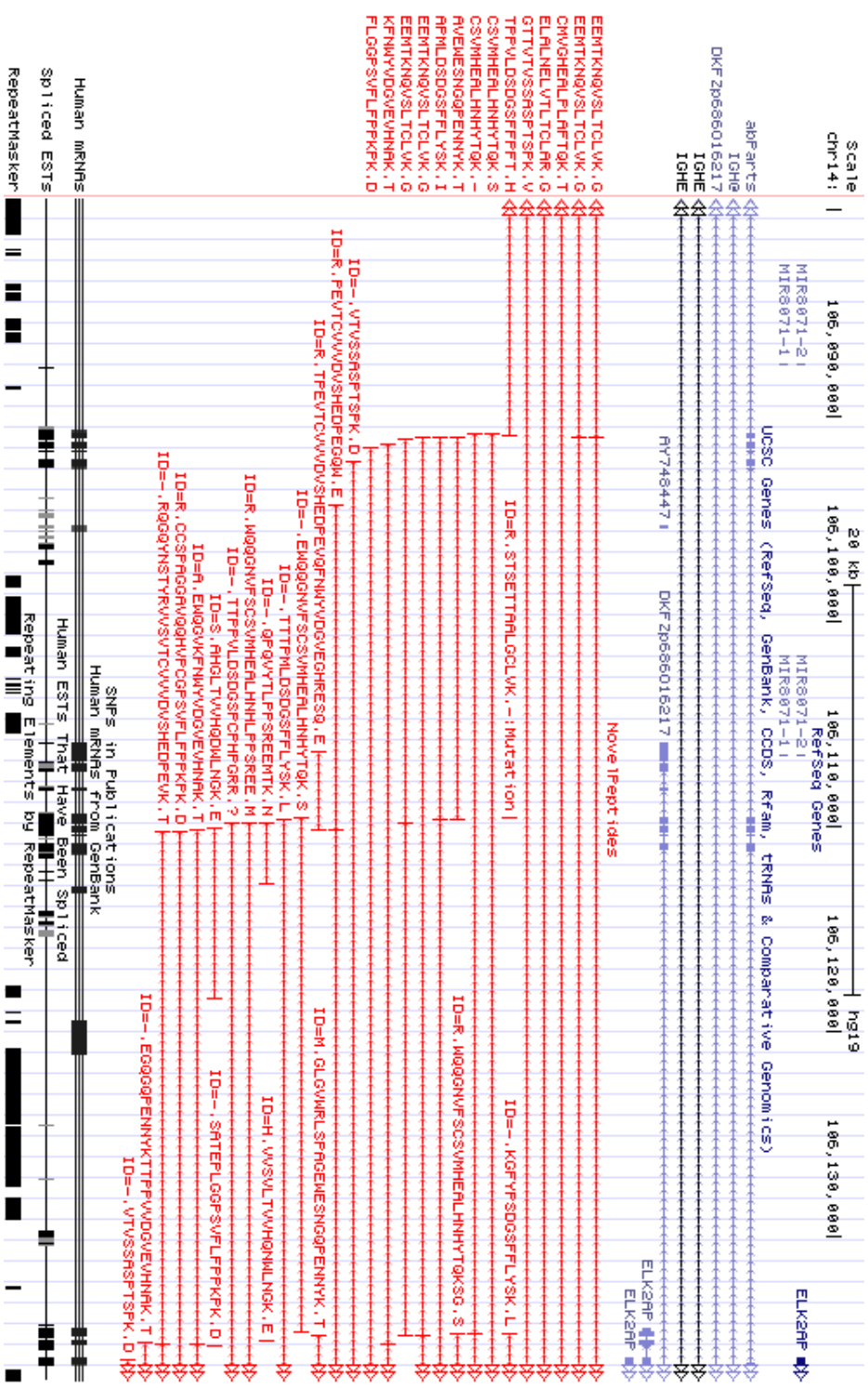


Figure B.3: UCSC genome browser plot of our novel peptide identifications within complex immunoglobulin region rearrangements included in our peptide identification result.

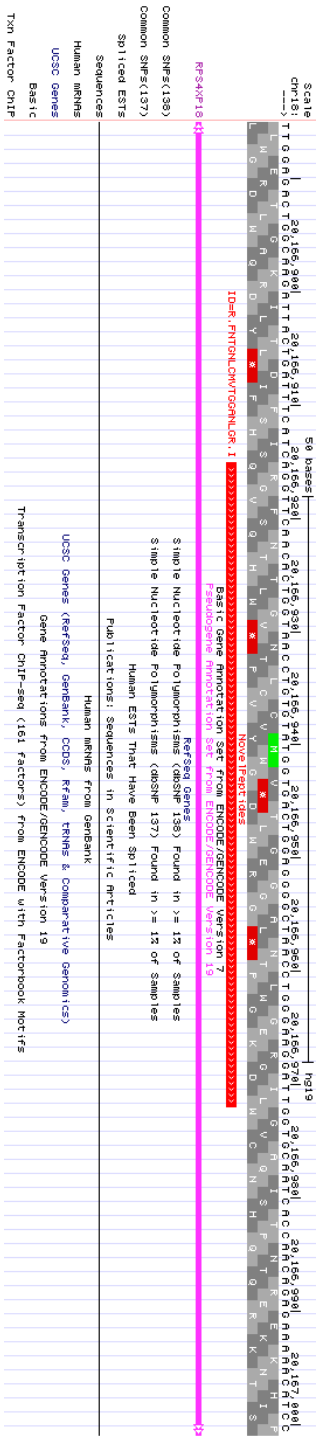


Figure B.4: UCSC genome browser plot of peptide identification in pseudo gene area.

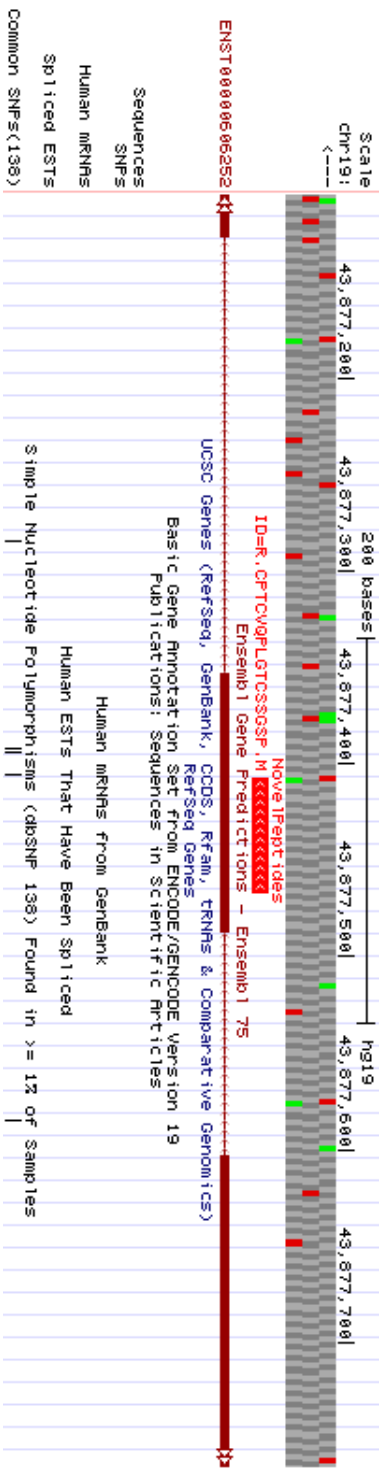


Figure B.5: UCSC genome browser plot of peptide identification in a possible novel gene area where a gene prediction method also reported as a possible gene.

Appendix C

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

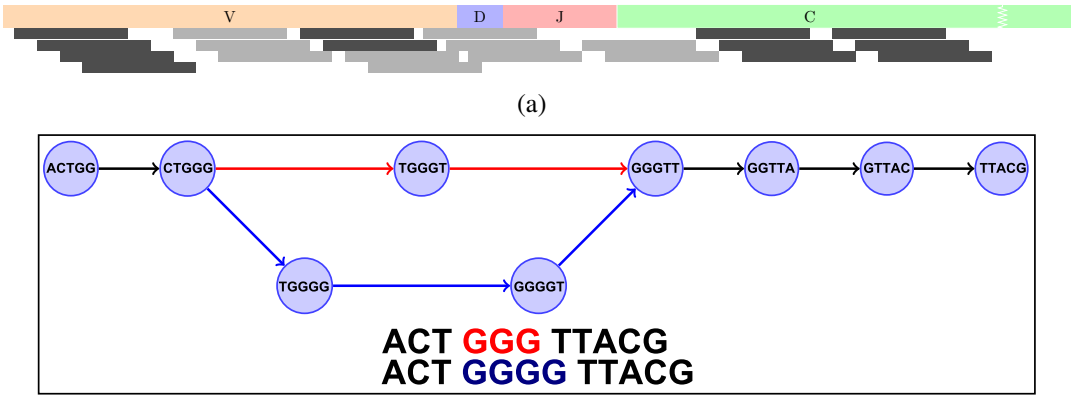


Figure C.1: (a)Potentially missed (in RNA-seq read alignment) reads from a somatically recombined heavy chain transcript as greyed out, while mapping reads as darker. (b)Example de Bruijn graph showing how differences in sequence manifest as differences in topology. In this example $k = 6$, and a single homopolymer difference is shown.

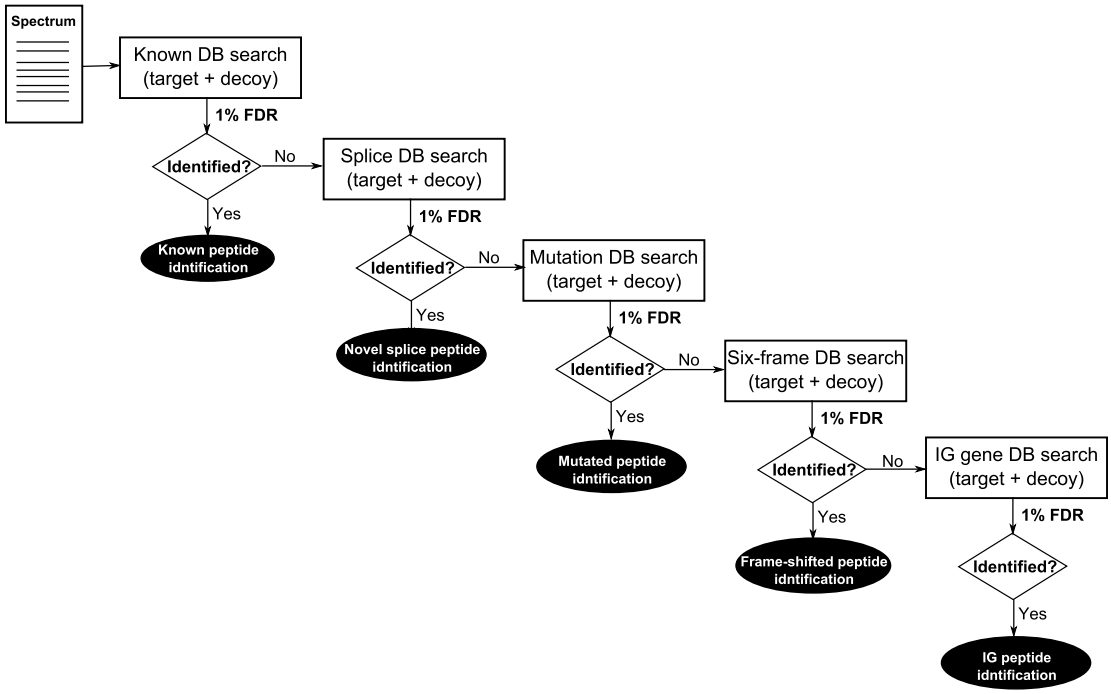
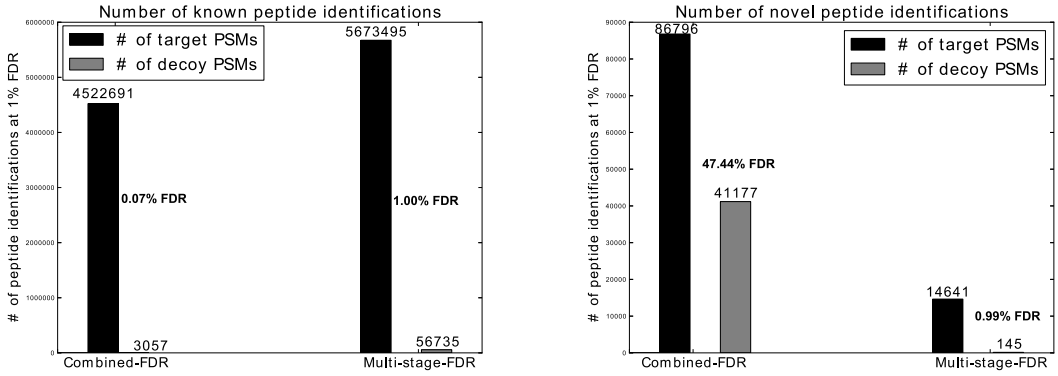


Figure C.2: Multistage-FDR strategy. Every spectrum will be searched against the known peptide database first, and are reported as a known peptides. In following stages, only the unidentified portion of the spectra are searched and assigned a new FDR threshold. Similar procedure is applied in the following order. Splice DB $\rightarrow$ Mutation DB $\rightarrow$ Sixframe DB $\rightarrow$ Immunoglobulin DB.

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. [138] were used in creating proteogenomic database. Using total 348 GB of BAM files, 2.57 GB of FASTA formatted protein database were created. By removing all FASTA headers (containing sample and genomic coordinate information), we searched total 888 MB of amino acid sequence characters. Final proteogenomic database contained, 605,171 substitutions, 20,263 deletions, 1,130 insertions, and 1,245,069 novel splice junctions.

	DB attributes	Statistics
RNA-seq input files	# of samples	90
	BAM size	348 GB
Protein DB	MutationDB FASTA size	2.57 GB
	IG DB FASTA size	467 MB
	Total AA searched	888 MB
# of mutations encoded in DB	Novel splice	1,245,069
	Deletions	20,263
	Insertions	1,130
	Substitutions	605,171



(a) (b)

Figure C.3: In order to calculate the accurate FDR separately in known and novel peptide identifications in combined FDR strategy, we explicitly distinguished and parsed out the PSMs resulted from known target and known decoy versus novel target and novel decoy database from the concatenated PSM list. (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. We observed that the actual FDR threshold has been distorted significantly in both novel and known peptide identifications when combined FDR strategy is applied.

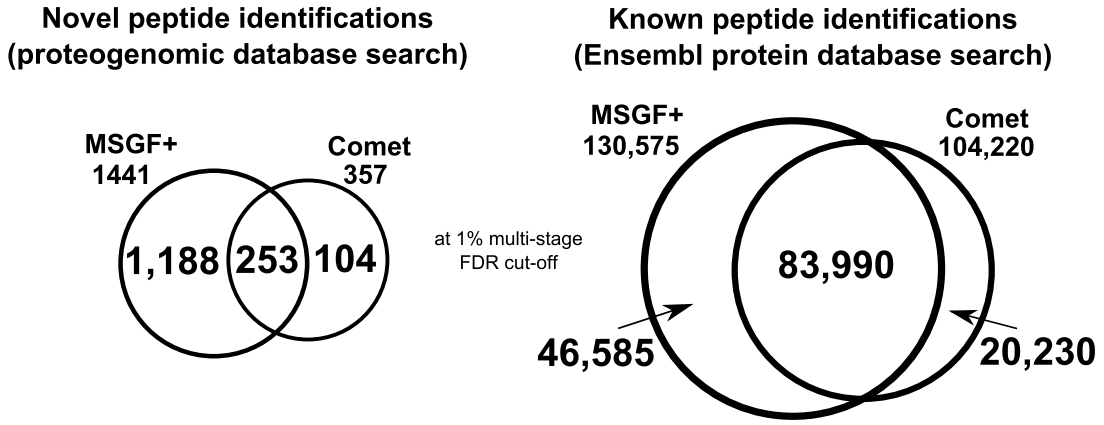


Figure C.4: Comparison between results obtained using MSGF+ and Comet MS/MS search tools. MSGF+ [54] showed more peptide identification results in both known (Ensembl [34] protein database) and novel (proteogenomic database) protein search with significant overlap.

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. However, as stated earlier, we reason that traditional combined FDR strategy could distort the FDR threshold significantly especially in novel peptide identifications.

Type of novel findings	# of novel findings
Substitutions	2,090
Insertion	9
Deletion	7
IG gene	428
Transcript gene	587
Fusiton gene	99
TranslatedUTR	376
Alternative splice	239
Novel splice	2,355
Exon boundary	76
Frame shift	1,516
Novel exon	523
Novel gene	733
Reverse strand	1,578
Pseudo gene	197

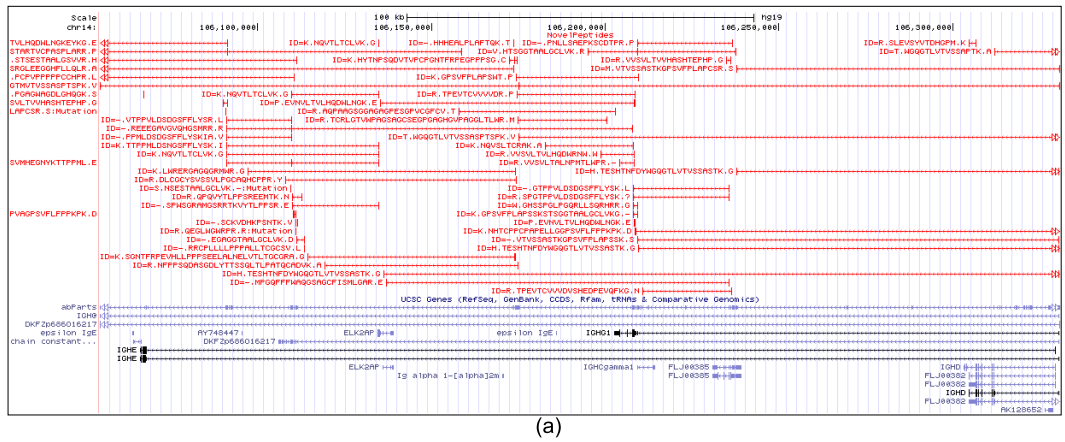


Figure C.5: (a) Example of peptide identifications resulted from immunoglobulin rearrangements. We have identified clusters of peptides spanning junctions of V(D)J recombinations. (b) Diagram illustrating the peptide identifications of V(D)J recombination junctions. We identified clusters of peptides in IG region which connects various V(D)J segments.

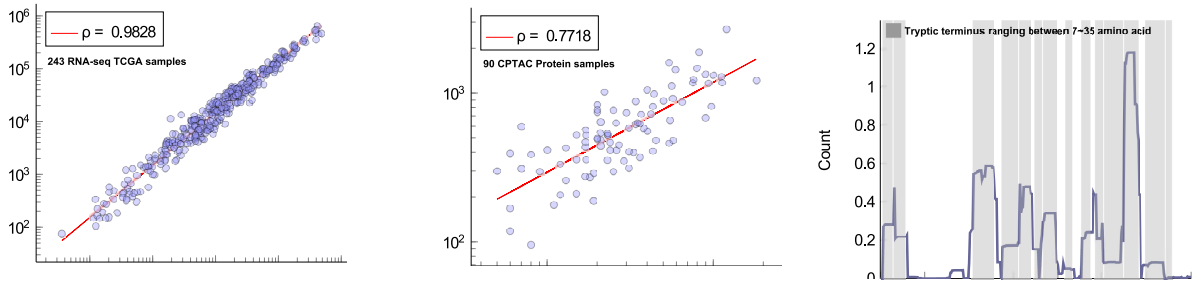


Figure C.6: (a) Plot of RNA-seq read counts from IG variable region versus IG constant region. We observed high correlation between RNA-seq reads that mapped to IG constant region versus variable region (filtered out using IG filter used in this study). (b) Spectra counts of peptide identifications from IG constant versus variable region. 90 protein samples overlapping with TCGA samples are plotted. We also observed a high correlation in peptide spectra counts in IG variable versus constant regions. (c) Plot of spectra counts covering IgG constant region. All possible tryptic terminus ranging between 7-35 amino acids are greyed out. We identified a large number of spectra covering all possible tryptic terminus in this region.

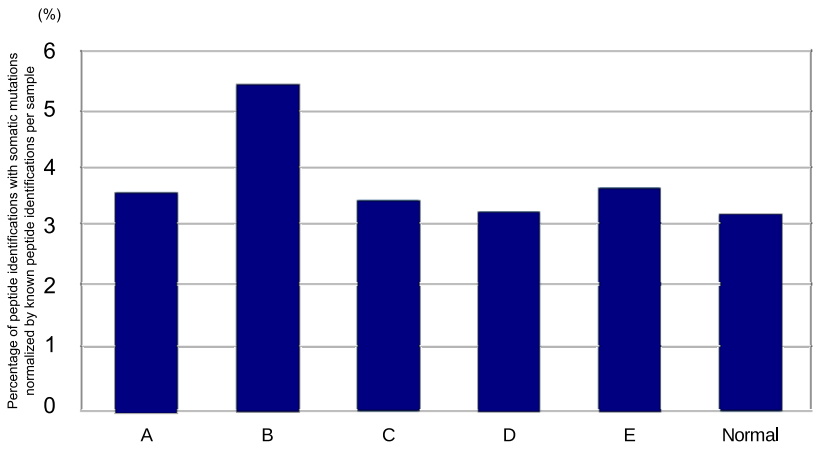
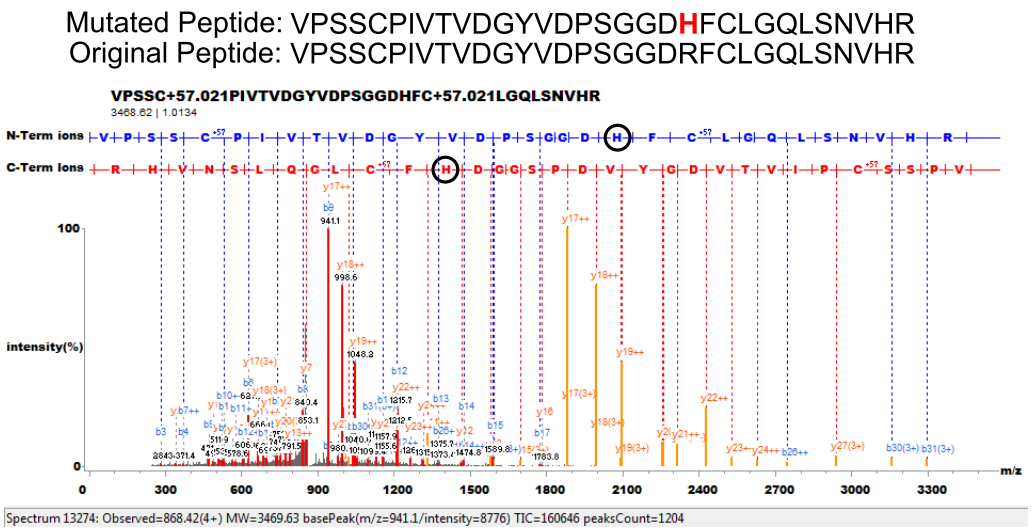
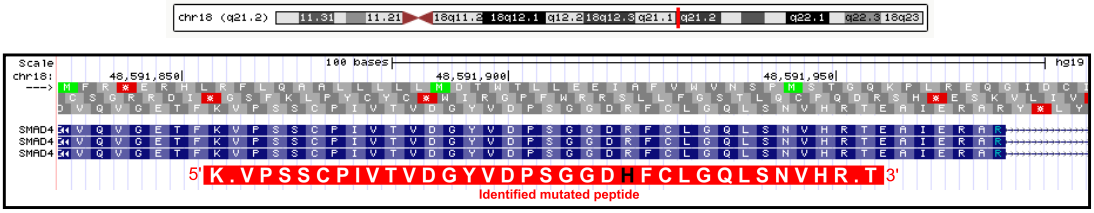


Figure C.7: Percentage of peptide identifications with somatic mutations in each sample normalized by the number of known peptide identifications across sample subtypes. This percentile ratio is calculated by dividing the number of known peptide identifications from the total number of IG peptide identifications within each sample. (ratio = (# of IG peptides) / (# of known peptides) * 100). Subtype B (sample groups showing hypermutation and non-CIMP characteristics) showed comparably high number of somatic mutations identified through peptide compared to other sample subtypes. Chi-squared test of this plot showed $p\text{-}value < 0.0001, \chi^2 = 40.39$.

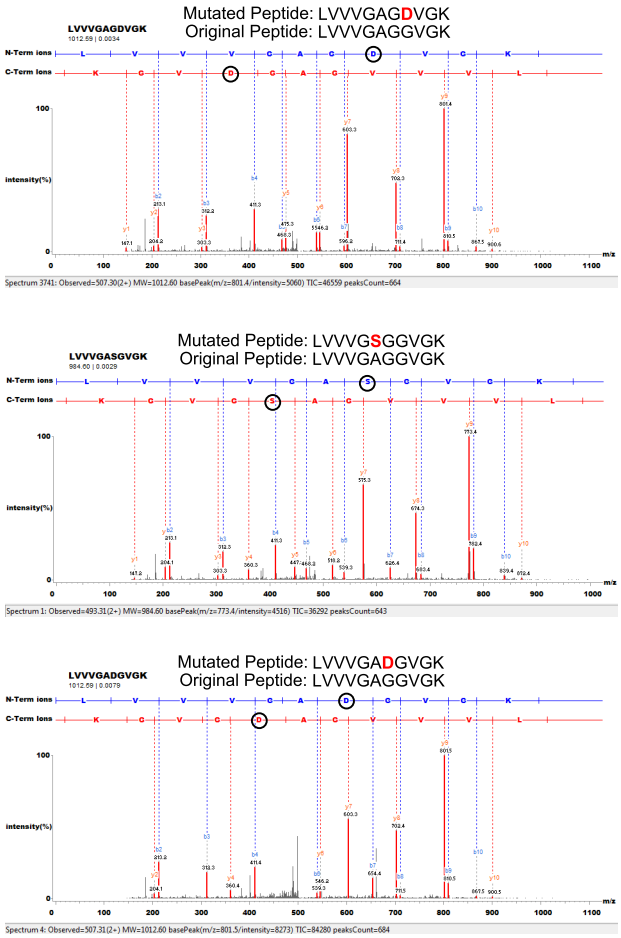


(a) Alignment of identified spectra with somatic mutation in gene SMAD4.

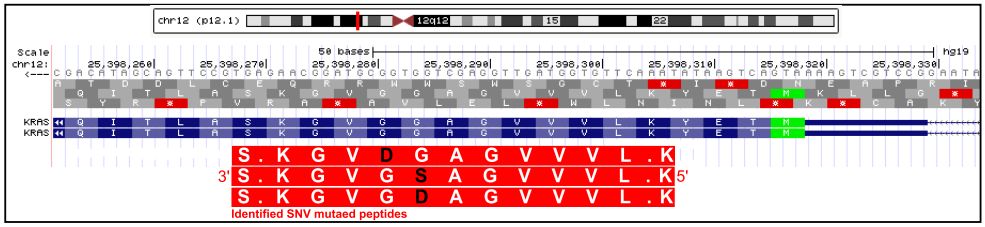


(b) UCSC Genome Browser plot of identified somatic mutation.

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 7 different samples from TCGA colon cancer study [75], and overlapping mutation existed in COSMIC [35] database.

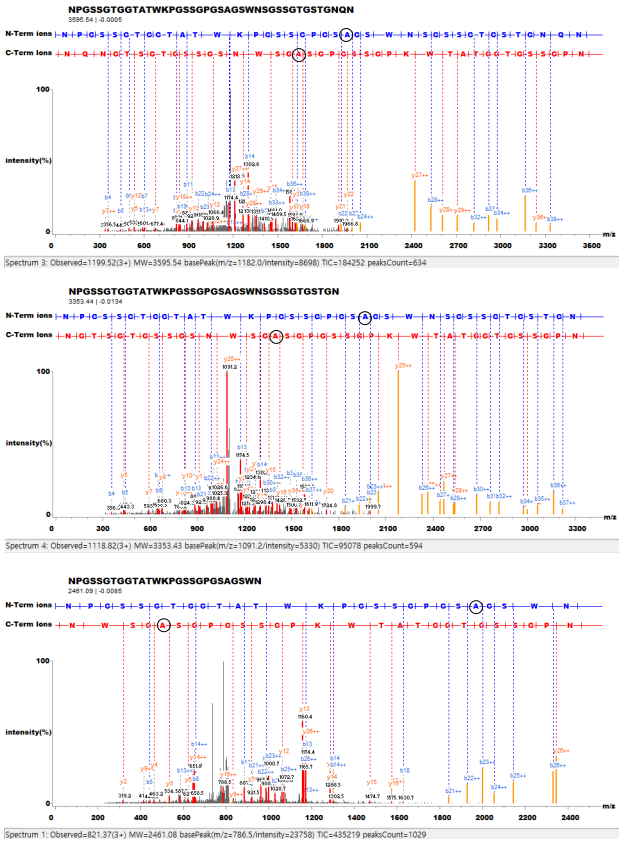


(a) Alignment of identified spectra with somatic mutation in gene KRAS

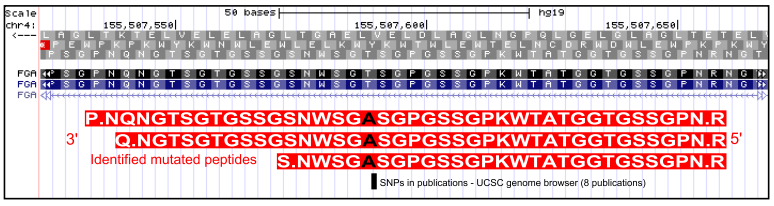


(b) UCSC Genome Browser plot of identified somatic mutation in gene KRAS.

Figure C.9: Identification of somatic mutation in gene KRAS. TCGA colon cancer study [75] reported this mutation as 'somatic' in 25 different colon cancer samples and also reported by COSMIC [35] and dbSNP [105]. Peptide 'LVVVGAG:D:VGK' (G− >D) had 1 spctra count and unique genomic location.

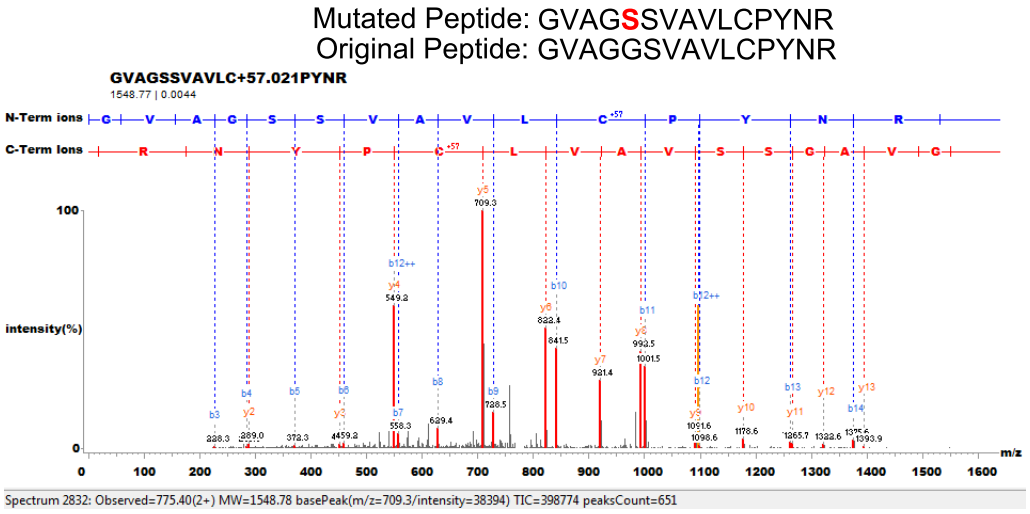


(a) Three identified spectra alignments indicating an identical SNV mutation in gene FGA

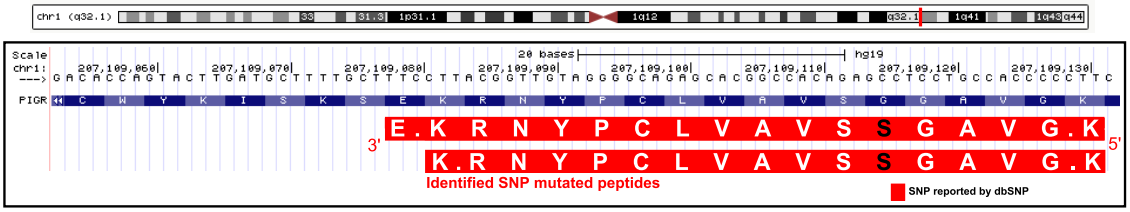


(b) UCSC Genome Browser plot of identified somatic mutation in gene FGA.

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 [35] and dbSNP [105].

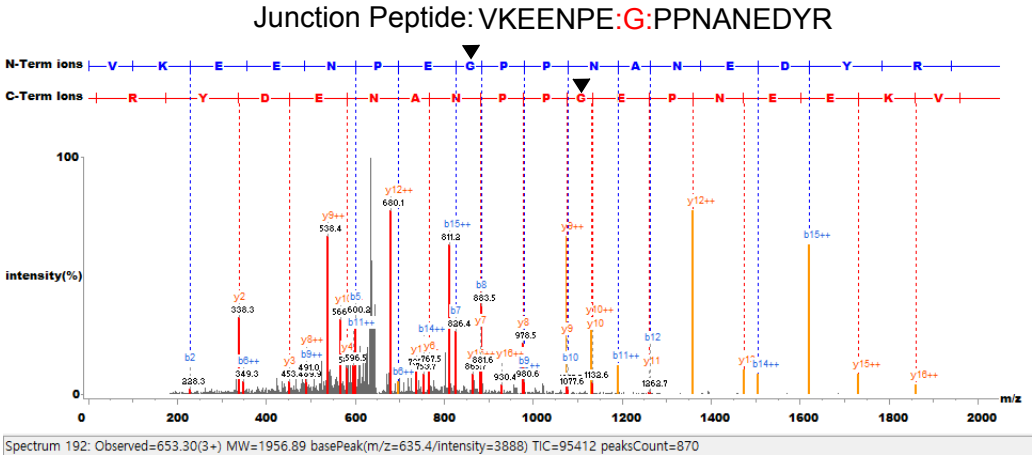


(a) Two identified spectra indicating an identical SNV mutation in gene PIGR

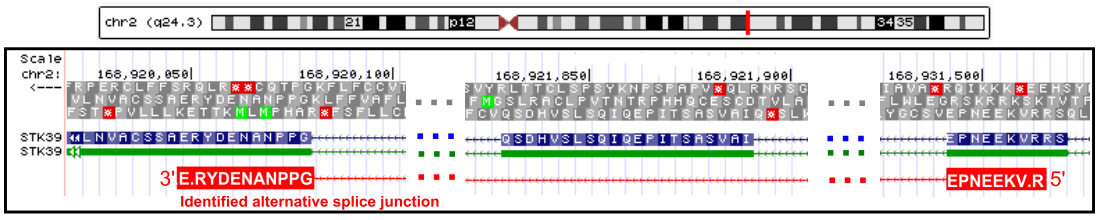


(b) UCSC Genome Browser plot of identified SNV mutation in gene PIGR.

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. We found these two mutated peptides in a single protein sample that was categorized as subtype 'C' (subtype with high-IG peptide identification rate). Matching mutation of this region were found in both COSMIC [35] and dbSNP [105].

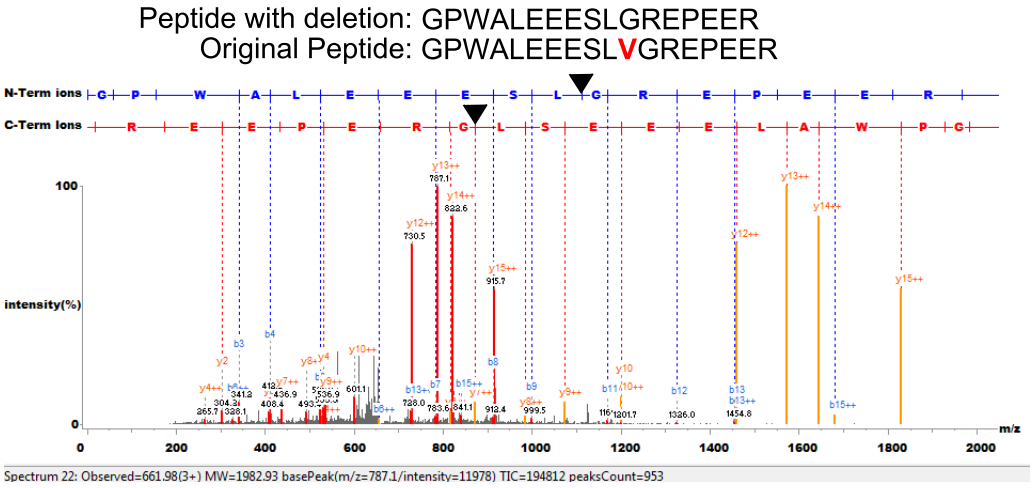


(a) Identified spectra with alternative splice junction

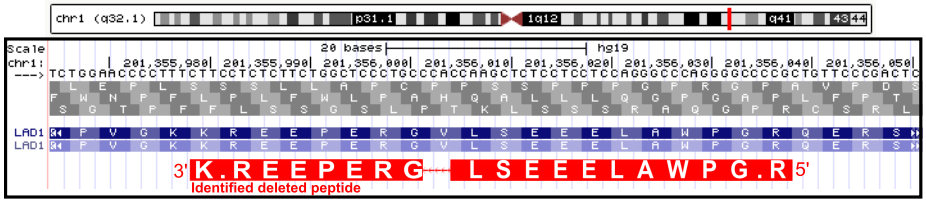


(b) UCSC Genome Browser plot of identified alternative splice junction

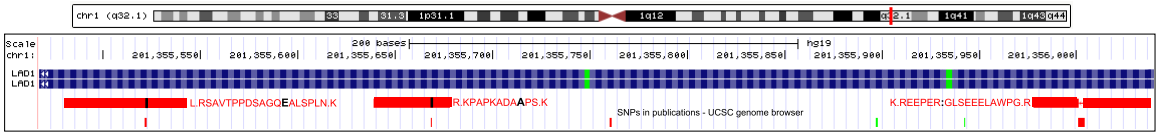
Figure C.12: Identified alternative splice junction peptide. Peptide 'VKEENPE:G:PPNANED YR' (junction existing in the middle of amino acid 'G') had 11 spectra counts (with unique genomic location) and total 386 RNA-seq reads were mapped to this alternative splice junction.



(a) Identified spectra with deletion

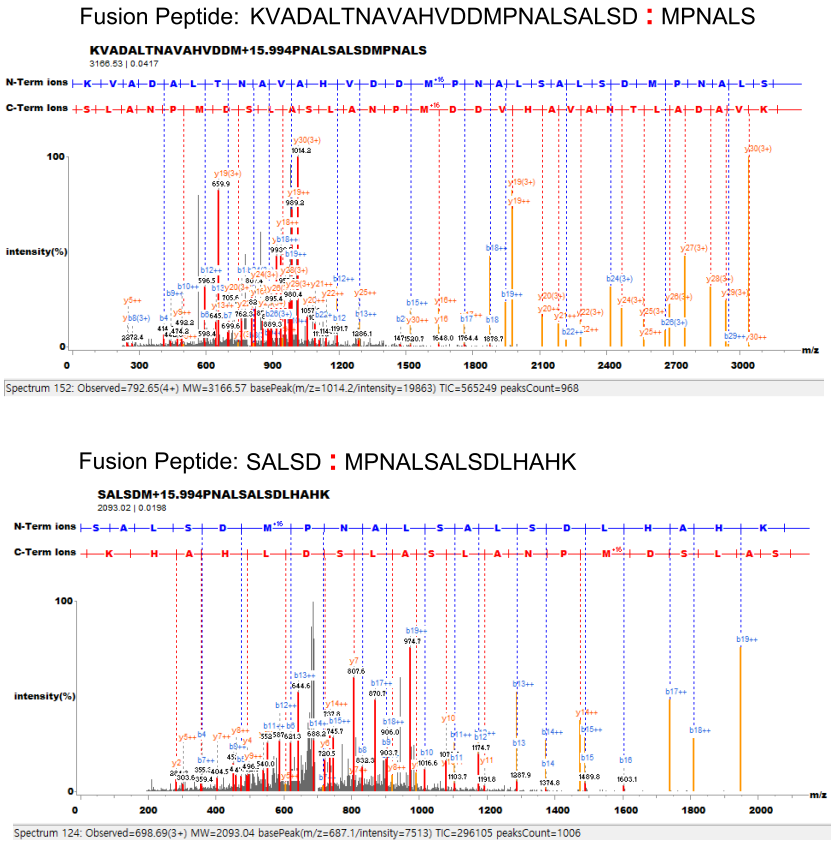


(b) UCSC Genome Browser plot of identified deletion.

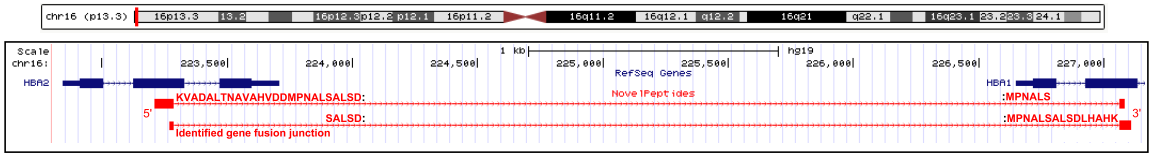


(c) Two different adjacent SNV mutations in LAD1 gene located within the same exon with identified deleted peptide.

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). Additionally, two SNV mutations were further identified within the same exon. All mutations found in this exon had external supporting evidences from dbSNP. SNV mutation of the peptide 'K.NLPSLA:E:QGASDPPTVASR.L' (K− >E) was also reported by TCGA [75] colon cancer somatic mutation calls with 10,711 read depth.



(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 accross two different genes (HBA1 and HBA2). Two fusion peptide shown in this region had unique genomic location and total 15 spectra counts. HBA1 and HBA2 are Hemoglobin related genes.

Appendix D

Appendix: The antibody repertoire of colorectal cancer

D.1 Supplemental method

D.1.1 Antibody structure and IMGT reference

Antibody is a 'Y' shape protein that recognizes an antigen. As illustrated in Fig D.7, antibodies are composed of four polypeptides: two identical copies of a heavy chain, and two identical copies of a light chain. In mammals, there are five types of heavy chain denoted by α, δ, ε, γ, and μ, and two types of light chain denoted by λ, and κ. Five types of heavy chain define the five different isotype of antibody (respectively IgA, IgD, IgE, IgG, and IgM), and each isotype may be composed of either λ or κ type of a light chain.

Both heavy and light chains are consist of variable region and constant region denoted by *V* and *C* respectively. Variable region include complementarity determining regions (CDR), which are "hotspots" in DNA containing high variation in sequence and length. Between the CDRs, there are framework regions (FR) whose sequence and length are relatively stable. The length of the CDRs are varies mainly between $15 \backsim 30$ nt for CDR1 and CDR2, and $24 \backsim 36$ nt for CDR3 [98].

We downloaded the reference antibody including all framework region sequences from

ImMunoGeneTics (IMGT) web site (www.imgt.org). IMGT is the global reference in immunogenetics and immunoinformatics created by Marie-Paule Lefranc. IMGT provide the resource for immunoglobulins (IG), T cell receptors (TR), and major histocompatibility (MH) and related proteins of the immune system (RPI). In this paper, we accessed the resource for immunoglobulin heavy chain variable and constant region (IGHV and IGHC) whose genomic locus is $14q32.33$ and immunoglobulin light chain λ (IGL) and κ (IGK) whose genomic locus are $22q11.2$ and $2p11.2$ respectively.

D.1.2 Node grouping method for SdB graph

Recall that each node in a SdB graph is a pair (x,y) representing r and ℓ letters of the sequence. A pair of nodes $u=(x,y)$ and $v=(x',y')$ are connected by a directed edge if the $\ell+r-1$ suffix of u matches the corresponding prefix of v, and there is a read that spans the combined sequence. Unlike dB graphs, however, we glue nodes in SdB graph if the first r letters/nucleotides are identical and following ℓ letters differ by at most one nucleotide mismatch (no indels).

To glue the nodes, we apply the following procedure:

1. Weigh each node u by the number of reads that support it.
2. Arrange nodes into stacks using the following criterion: Nodes $u=(x,y), v=(x',y')$ belong to the same stack if $x=x'$. Furthermore, the nodes in stack S are ordered by decreasing weights.
3. While S is not empty:
   (a) Let representative $r=(x,y)$= POP(S). In other words, the node with the highest weight is the current representative.
   (b) Glue all unglued nodes $v=(x,y')$ in S to r if y,y' differ by at most one mismatch, and remove those nodes from S.

In the implementation, we speed up this procedure by considering the stacks in a breadth-first search manner based on the directed SdB graph. For each stack, we consider the predecessor nodes of the

nodes, and glue them if their predecessor nodes are glued as a first step. The gluing algorithm is then applied only to the representatives of each glued set in the stack.

D.1.3 Mathematical comparison between the SdB and dB graph

The purpose of the SdB and dB graph is reconstructing the original genes from the fragmented reads. In the ideal scenario, the graph connects the edges if and only if two reads are from same gene. However, the graph often connects the two reads from the different genes (false positive) or fails to connect the reads from the same gene (false negative). To show the performance of the SdB and dB graph, we calculated the false positive and false negative mathematically. In the following equations, we use the following notation:

$p_s \equiv \Pr(\text{A pair of randomly chosen nucleotides are identical})$

$\varepsilon \equiv$ Probability that a nucleotide is sequenced incorrectly

$D(x) \equiv$ The event that two reads originating from the same gene and overlapping by x nucleotides, share at least one k-mer

$S(x) \equiv$ The event that two reads originating from the same gene and overlapping by x nucleotides have an approximate match as follows: there is at least one $(r+\ell)$-mer where the first r nucleotides match exactly, and the remaining ℓ nucleotides have at most one mismatch.

$L =$ Read length

Computing false positives. The probability of two reads from the different genes would share a identical sequence with a probability of p_s. Hence, the false positive of dB and SdB graph can be

calculated as following,

False positive of dB graph over k-mer

$$\mathrm{FP}_{dB_k} = p_s^k$$

False positive of SdB graph over $\ell + r$-mer

$$\begin{aligned}
\mathrm{FP}_{SdB_{\ell+r}} &= \Pr(\ell + r\text{-mer sequences have no error}) \\
&\quad + Pr(\text{first } r\text{-mer sequences have no error and following } \ell\text{-mer sequence have} \\
&\quad 1 \text{ error} \\
&= p_s^{r+\ell} + p_s^r \cdot \binom{\ell}{1}(1-p_s)p_s^{\ell-1} \\
&= p_s^{r+\ell} + \ell(1-p_s)p_s^{r+\ell-1} \\
&= p_s^{r+\ell} \cdot (1 + \ell(\frac{1-p_s}{p_s}))
\end{aligned}$$

Enforcing the condition of the false-overlap rate $\mathrm{FP}_{SdB_{\ell+r}} \leq \mathrm{FP}_{dB_k}$ to set $\mathrm{FP}_{SdB_{\ell+r}} \leq \mathrm{FP}_{dB_{\ell+r}}$ while $\mathrm{FP}_{dB_k} \leq \mathrm{FP}_{dB_{\ell+r}}$ for all $k \leq \ell + r$.

$$\begin{aligned}
p_s^k &\geq p_s^{r+\ell} \cdot (1 + \ell(\frac{1-p_s}{p_s})) \\
k &\leq r + \ell + \log_{p_s}(1 + \ell(\frac{1-p_s}{p_s}))
\end{aligned} \tag{D.1}$$

We chose the largest k satisfying Eqn. D.1 to keep that the false positive of SdB graph is always smaller than or equal to that of dB graph.

Computing false negatives. By definition, it is sufficient to show that $\Pr(D(x)) > Pr(S(x))$ for all $x \leq L$ to prove that SdB graphs have fewer false negative edges than dB graphs. Let $\mathcal{D}_x = \Pr(D(x))$. We first show that:

$$\mathcal{D}_x = \begin{cases} 0 & \text{if } x < k \\ (1-\varepsilon)^k & \text{if } x = k \\ \mathcal{D}_{x-1} + \varepsilon(1-\varepsilon)^k (1 - \mathcal{D}_{x-k-1}) & \text{if } x > k \end{cases}$$

Note that the base cases ($x \leq k$) are trivial. For reads overlapping by x bp, and $0 \leq a \leq x$, let $D(x,a)$ denote the event that the first a nucleotides in the overlap contain a matching k-mer, and $D(x,-a)$ denote the event that the last a nucleotides in the overlap contain a matching k-mer. Note that $\Pr(D(x,a)) = \Pr(D(x,-a)) = \mathcal{D}_a$ for all $0 \leq a \leq x$. By definition,

$$\begin{aligned} D(x) &= D(x,x-1) \cup D(x,-k) \\ \mathcal{D}_x &= \Pr(D(x,x-1) \cup D(x,-k)) \\ &= \Pr(D(x,x-1)) + \Pr(D(x,-k)) - \Pr(D(x,-k)) \cdot \Pr(D(x,x-1)|D(x,-k)) \\ &= \mathcal{D}_{x-1} + \mathcal{D}_k - \mathcal{D}_k \cdot \Pr(D(x,x-1)|D(x,-k)) \end{aligned}$$

If last k nucleotides match, then so must the last $k-1$ nucleotides over the first $x-1$ nucleotides. Therefore, $D(x,x-1)|D(x,-k)$ is true if the $(k)_{\text{th}}$ nucleotide from the end matches, or the $(k)_{\text{th}}$ mismatches, but $D(x,x-k-1)$ is true. In other words,

$$\Pr(D(x,x-1)|D(x,-k)) = (1-\varepsilon) + \varepsilon \Pr(D(x,x-k-1))$$

Therefore,

$$\mathcal{D}_x = \mathcal{D}_{x-1} + \varepsilon(1-\varepsilon)^k (1 - \mathcal{D}_{x-k-1})$$

For SdB graphs, let $\mathcal{S}_x = \Pr(S(x))$. Applying a similar approach, we can show that

$$\mathcal{S}_x = \begin{cases} 0 & \text{if } x < r+\ell \\ (1-\varepsilon)^r \left((1-\varepsilon)^\ell + \ell\varepsilon(1-\varepsilon)^{(\ell-1)}\right) & \text{if } x = r+\ell \\ \mathcal{S}_{x-1} + (1-\varepsilon)^r \left((1-\varepsilon)^\ell + \ell\varepsilon(1-\varepsilon)^{\ell-1}\right) & \\ -\{((1-\varepsilon)^{r+\ell} + (\ell-1)\varepsilon(1-\varepsilon)^{r+\ell})u_{r+\ell}(x) & \\ +\varepsilon^2(1-\varepsilon)^{r+l-1} \cdot \Sigma_{k=1}^{min(x-2r-\ell-1,\ell-r-1)}(1-\varepsilon)^{k+r}u_{2r+\ell}(x) & \\ +\varepsilon(1-\varepsilon)^{r+\ell-1}\left((1-\varepsilon)^{r+1} + (r+1)\varepsilon(1-\varepsilon)^r\right)u_{2r+\ell}(x) & \\ +re^2(1-\varepsilon)^{r+\ell-1}\mathcal{S}_{x-\ell-r-1}u_{2(r+\ell)}(x) & \\ +\varepsilon^2(1-\varepsilon)^{r+\ell-1}\Sigma_{i=1}^{\ell-r-1}\Sigma_{y=1}^{r+i}\varepsilon(1-\varepsilon)^{y-1}\mathcal{S}_{x-(y+r+\ell+1)}u_{2(r+\ell)}(x) & \\ +\varepsilon(1-\varepsilon)^{r+\ell-1}\Sigma_{z=2}^{r+1}(z-1)\varepsilon(1-\varepsilon)^{z-2}\mathcal{S}_{x-(z+r+\ell)}u_{2(r+\ell)}(x)\} & \text{otherwise} \end{cases}$$

Note that the base case ($x \leq r+\ell$) is trivial. We applied similar approaches to calculate $x > r+\ell$.

We denote $\mathcal{S}_x, S(x,a), S(x,-a), \mathcal{E}_a$, and $u_c(x)$ as follow:

$\mathcal{S}_x = \Pr(S(x))$

$S(x,a) =$ first a nucleotides contain a matching condition over the x nucleotide overlap

$S(x,-a) =$ last a nucleotides contain a matching condition over the x nucleotide overlap

$\mathcal{E}_a =$ set of all positions of the errors over the first a nucleotide matches,

(Ex, $\mathcal{E}_a = \{i, j\}$ errors in i_{th} and j_{th} position up to a)

$$u_c(x) = \begin{cases} 1 & \text{if } x > c \\ 0 & \text{if } x \leq c \end{cases}$$

We calculate the probability satisfying both $S(x, r+\ell)$ and $S(x, -(x-1))$ true simultaneously

$(\Pr(S(x, r+\ell) \cap S(x, -(x-1))))$ to finalize the calculation of $\mathcal{S}_x$,

$$\mathcal{S}_x = \Pr(S(x, r+\ell)) + \Pr(S(x, -(x-1))) - \Pr(S(x, r+\ell) \cap S(x, -(x-1)))$$

Note that $\Pr(S(x,a)) = \Pr(S(x,-a)) = \mathcal{S}_a$ for all $a \leq x$.

Recall that the matching condition for SdB graph is that first the r nucleotides match exactly and following ℓ nucleotides match at most one error. Under the condition of first $r+\ell$ nucleotides satisfying the matching condition, we explicitly consider all possible conditions. Let $\mathcal{E}_a$ denote the positions of mismatches in the first 'a' nucleotides in the overlapping part . For example, if there is no error over the first r nucleotides, then $\mathcal{E}_r = \{\}$. Similarly, if there is only one error at position $r+\ell-1$ in the first $r+\ell$ nucleotides, then $\mathcal{E}_{r+\ell} = \{r+\ell-1\}$. We enumerate possible positions of the first error conditioned on the fact that $S(x, r+\ell) \cap S(x, -(x-1))$ is true. (Supplemental Fig. D.13)

1. $\mathcal{E}_{r+\ell} = \{\}$. In other words, no mismatch in the first $r+\ell$ positions.
2. $\mathcal{E}_{r+\ell} = \{r+1\}$. In other words, a mismatch at $r+1$-st position in the first $r+\ell$ positions. The first mismatch after $r+1$ must occur after $r+\ell$ because $S(x, r+\ell)$ is true. This leads to different sub-conditions:
   (a) $|\mathcal{E}_{2r+\ell+1}| \leq 2$. In other words, there is at most one mismatch between $r+\ell+1$ and $2r+\ell+1$.
   (b) $|\mathcal{E}_{2r+\ell+1}| > 2$ and $S(x, -(x-a))$ is true, where a is the position of the third mismatch. In other words, the number of errors in the first $2r+\ell+1$ is larger than 2 and there is a true $(r+\ell)$-mer in the last $(x-a)$ positions.
3. $\mathcal{E}_{r+\ell+1} = \{a\}$, where $r+2 \leq a \leq r+\ell$. In other words, the single mismatch in the first $r+l$ nucleotides is after $r+1$, and there is a match at the $r+\ell+1$-st position.
4. $\mathcal{E}_{r+\ell+1} = \{a, r+\ell+1\}$, where $r+2 \leq a \leq r+\ell$. In other words, the single mismatch in the first $r+l$ nucleotides is after $r+1$, and there is a mis-match at the $r+\ell+1$-st position. This itself has more specific cases:

(a) $\mathcal{E}_{r+\ell+1} = \{a', r+\ell+1\}$ and $S(x, -(x-(r+\ell+1)))$ is true, where $\ell < a' \leq \ell + r$. In other words, first mismatch between ℓ and $r+\ell$ and there is a true $(r+\ell)$-mer in the last $x-(r+\ell+1)$ positions.

(b) $\mathcal{E}_{r+\ell+1} = \{a'', r+\ell+1\}$, where $r+2 \leq a'' \leq \ell$. In other words, first mismatch lies between $r+2$ and ℓ, then there are more specific cases:

i. $\mathcal{E}_{min(r+\ell+a'', x)} = \{a'', r+\ell+1\}$, where a'' is the position of the first mismatch. In other words, no mismatch between $r+\ell+2$ and minimum of ($r+\ell+a''$ and x).

ii. $\mathcal{E}_{r+\ell+b} = \{a'', b, r+\ell+1\}$ and $S(x, -(x-d))$ is true, where a'' is the position of the first mismatch and b is the position of third mismatch. In other words, mismatch exists between $r+\ell+2$ and $r+2\ell+a''+1$ and there is a true $(r+\ell)$-mer in the last $x-b$ position.

All possible combinations can be summarized in 7 cases ($1, 2a, 2b, 3, 4a, 4b(i)$, and $4b(ii)$). Note that the overlapping length x should be larger than the minimum length requirement of each cases. It will be added in the equation using the step function, $u_c(x)$. The probabilities of each condition is calculated as follows:

1	$(1-\varepsilon)^{r+\ell} u_{r+\ell}(x)$
$2a$	$\varepsilon(1-\varepsilon)^{r+\ell-1}\left((1-\varepsilon)^{r+1} + (r+1)\varepsilon(1-\varepsilon)^{r}\right) u_{2r+\ell}(x)$
$2b$	$\varepsilon(1-\varepsilon)^{r+\ell-1}\Sigma_{z=2}^{r+1}(z-1)\varepsilon^2(1-\varepsilon)^{z-2}\mathcal{S}_{x-(z+r+\ell)} u_{2(r+\ell)}(x)$
3	$(\ell-1)\varepsilon(1-\varepsilon)^{r+\ell} u_{r+\ell}(x)$
$4a$	$re^2(1-\varepsilon)^{r+\ell-1}\mathcal{S}_{x-\ell-r-1} u_{2(r+\ell)}(x)$
$4b(i)$	$\varepsilon^2(1-\varepsilon)^{r+l-1} \cdot \Sigma_{k=1}^{min(x-2r-\ell-1, \ell-r-1)}(1-\varepsilon)^{k+r} u_{2r+\ell}(x)$
$4b(ii)$	$\varepsilon^2(1-\varepsilon)^{r+\ell-1}\Sigma_{i=1}^{\ell-r-1}\Sigma_{y=1}^{r+i}\varepsilon(1-\varepsilon)^{y-1}\mathcal{S}_{x-(y+r+\ell+1)} u_{2(r+\ell)}(x)$

Supplemental Fig. D.6 shows that the probability of connection for SdB graphs is higher than dB graphs for all overlapping length x.

Supplemental Figures

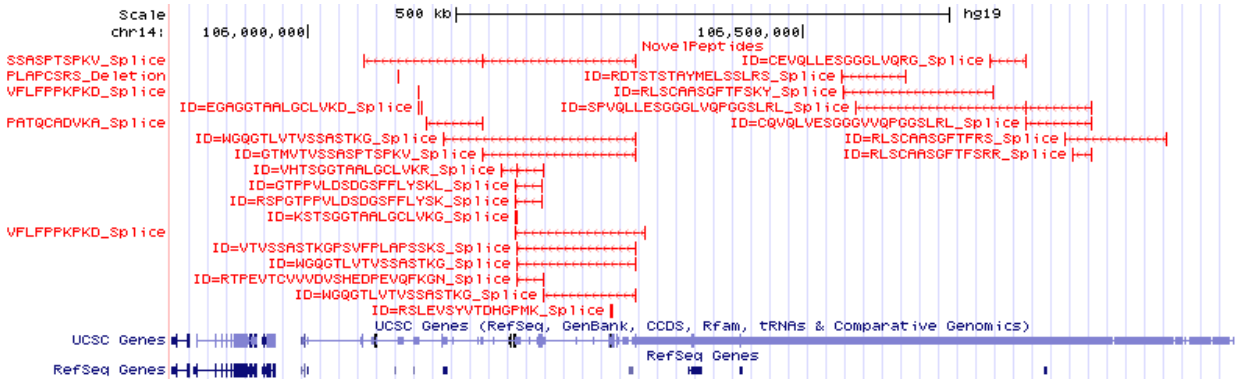


Figure D.1: **Examples of antibody peptides identified by ENOSI.** A UCSC genome browser view depicting locations of antibody peptides identified using ENOSI. The discoveries suggest that a more careful search is needed to identify all antibody peptides in tumor samples.

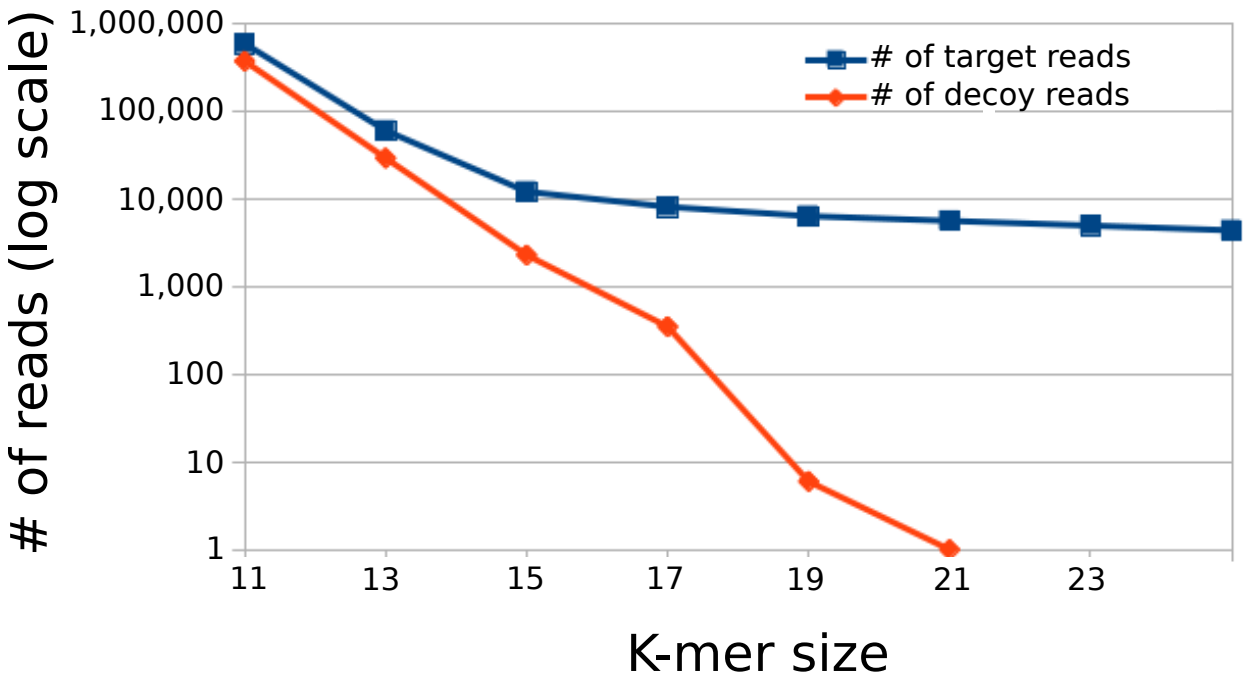


Figure D.2: **Read filter.** We filtered all unmapped reads containing k-mers that were found in the IMGT reference [59] of antibody sequences. The plot shows the number of unmapped reads with matching k-mers, as a function of k. As negative control, we reversed sequences in the IMGT database, and used them as decoy. The decoy matches only a small number of reads once $k \geq 19$. Therefore, we used 19-mers to filter for antibody sequences.

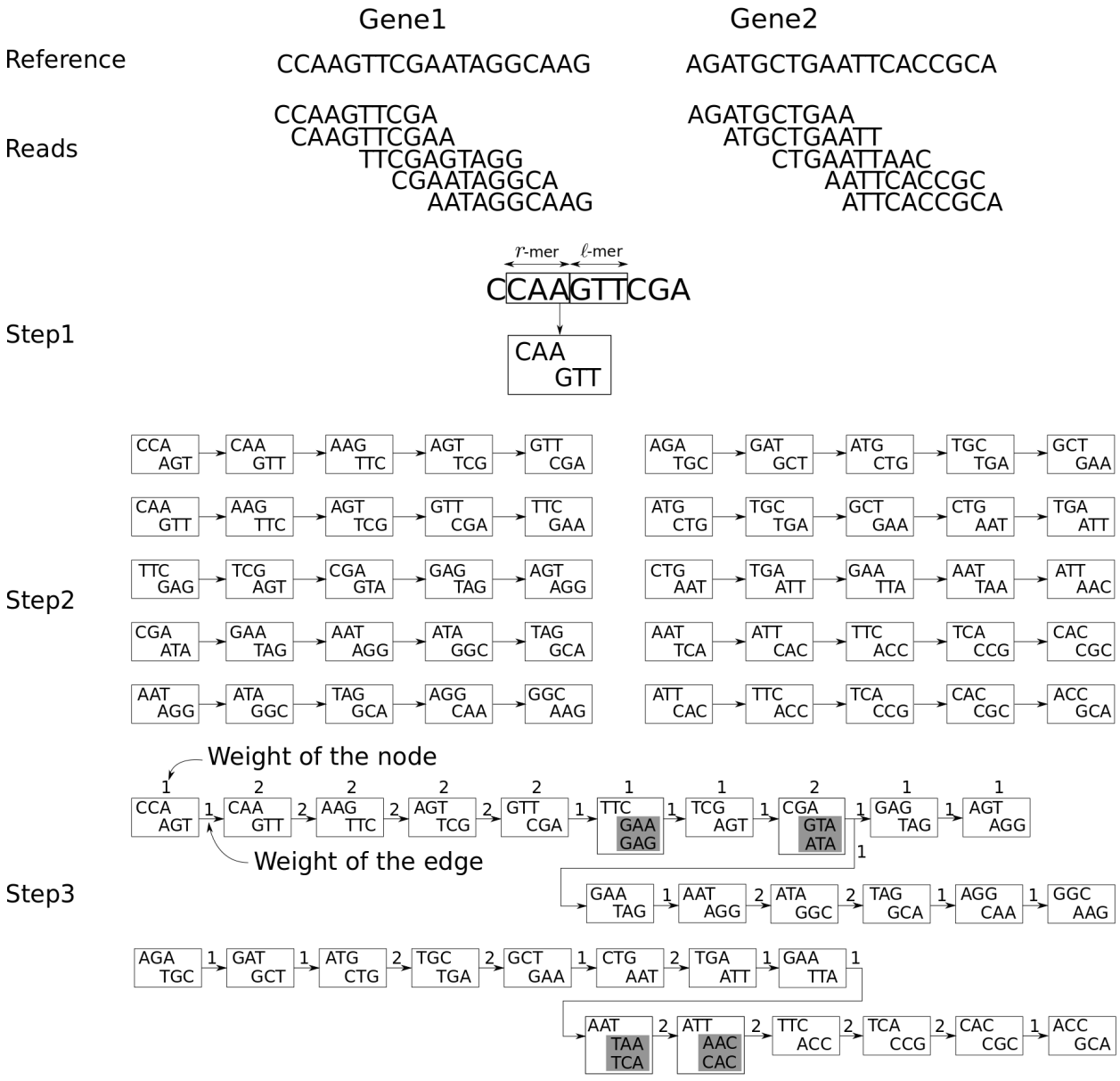


Figure D.3: **Schematic illustration of SdB graph construction.** The figure shows 5 reads each from *gene 1* and *gene 2*. Step 1: Each node $u = (x, y)$ initially corresponds to a distinct $(r+\ell)$-mer from the read, where x is a length r prefix and y is a length ℓ suffix sequence. Step 2: Edges are added, connecting nodes corresponding to adjacent $(r+\ell)$-mers in a sliding window in each read. Step 3: Pairs of nodes $u = (x, y)$ and $v = (x', y')$ are merged if $d_h(x, x') = 0$ and $d_h(y, 'y) \leq 1$, where $d_h(x, x')$ is a hamming distance between x and x'. Weights on edges corresponds to the number of reads supporting the edge. Weights on nodes correspond to the maximum of the sum of incoming and outgoing edge weights (See Methods).

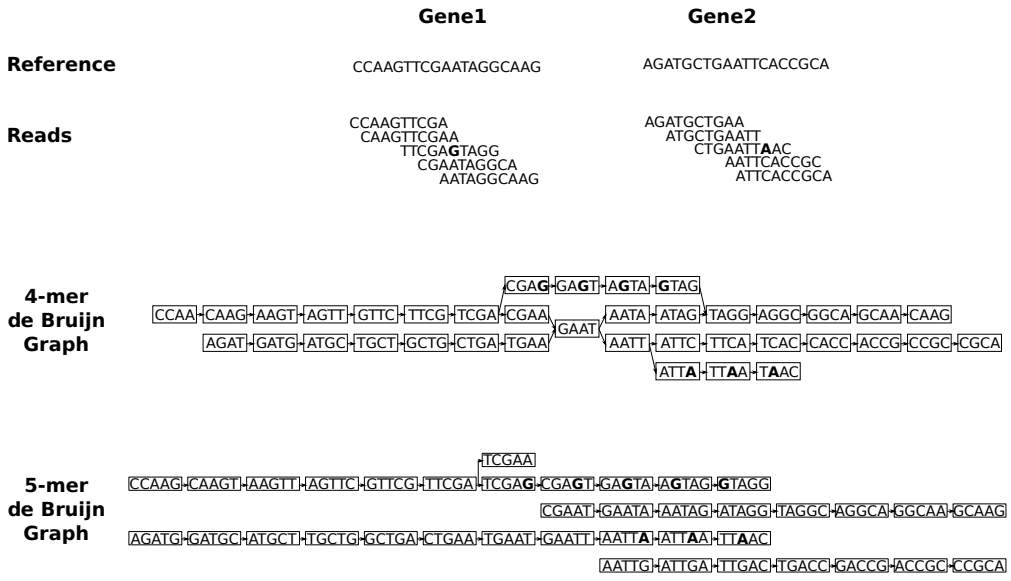


Figure D.4: **dB graph construction example with a parameter $k = 4$ and $k = 5$.** The genes and reads applied in the example of SdB graph (Fig. D.3) used to construct the dB graph with parameter $k = 4$, and $k = 5$. Note that 4-mer graph failed to differentiate two genes, and 5-mer graph failed to connect true edges within the genes.

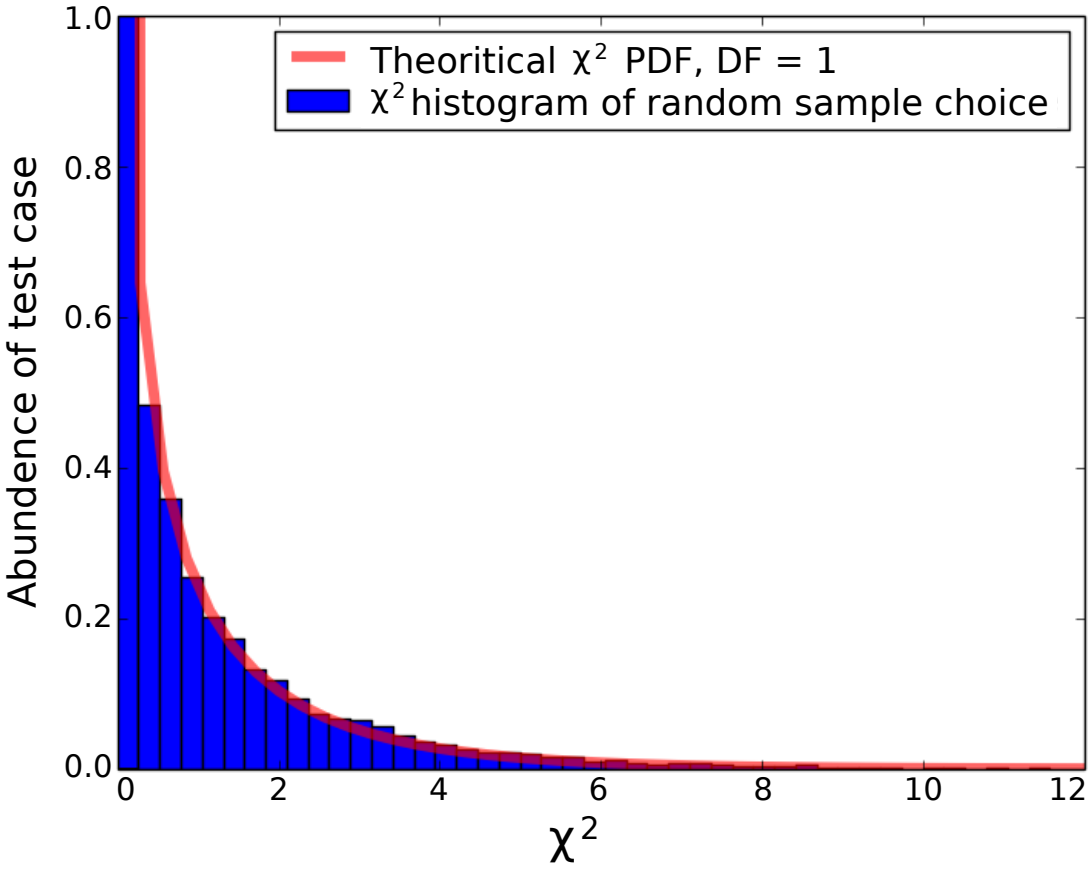
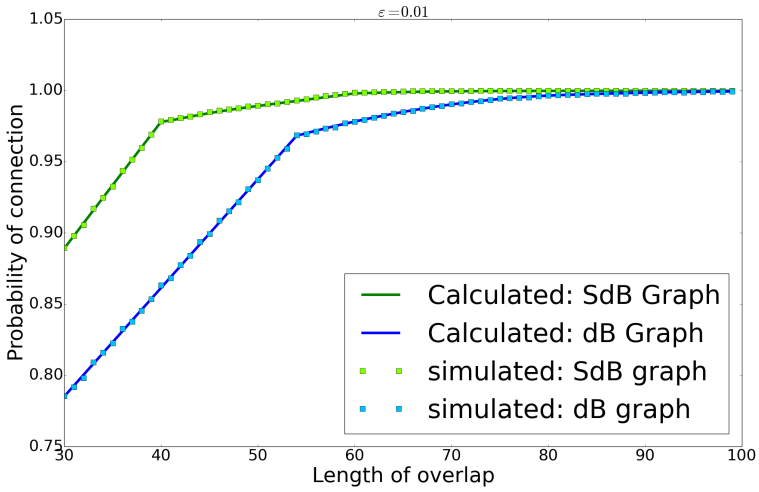
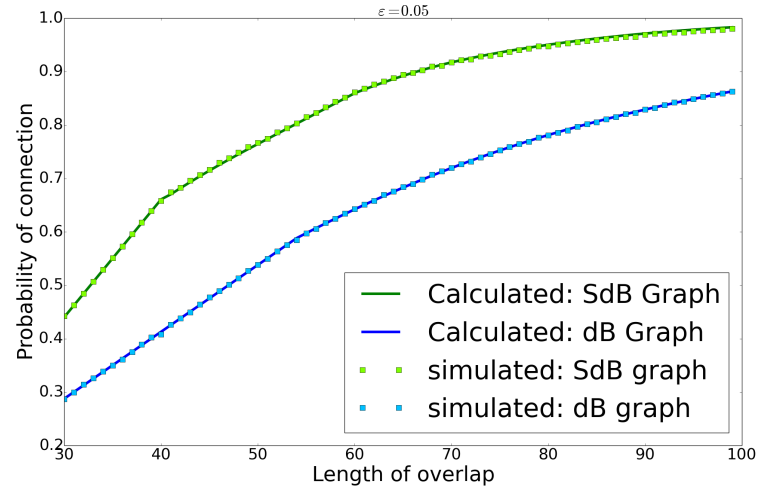


Figure D.5: χ^2 **distribution of log-rank test.** 10,000 iterations of random sample choice were performed to estimate the null distribution of the log-rank test. Blue bar shows the distrubution of simulated result and solid red line shows the theoritical χ^2 distribution.



(a) $\varepsilon = 0.01$



(b) $\varepsilon = 0.05$

Figure D.6: **Performance gap between SdB and dB graph.** Sensitivity of SdB and dB graphs are compared as a function of the length of overlap, and the error rate ε. Solid lines show the analytically computed sensitivity values (See Supplemental method–'Comparison between the SdB and dB graph mathematically'), while the dots represent the means of 100,000 simulation experiments. We observe concordance between analytical and simulation results. The sensitivity increases with length of overlap and decreases with higher ε. SdB graphs consistently outperform dB graphs.

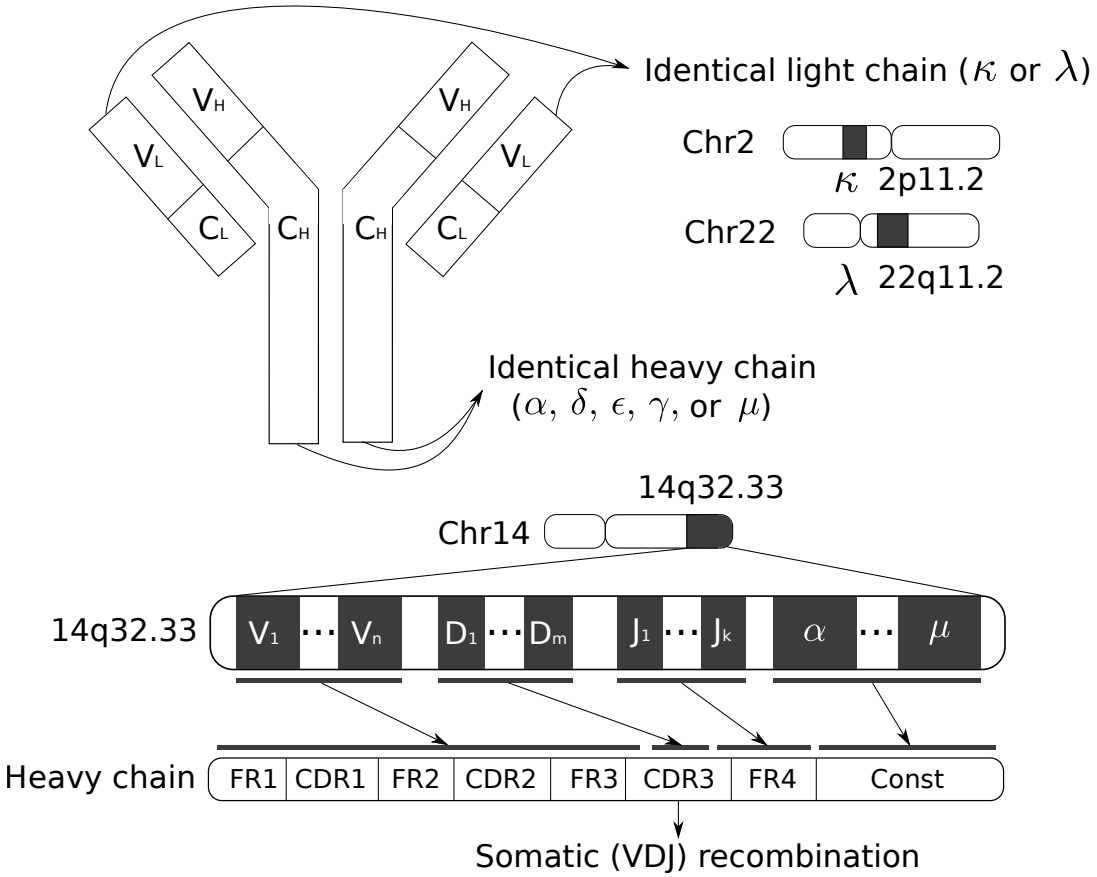
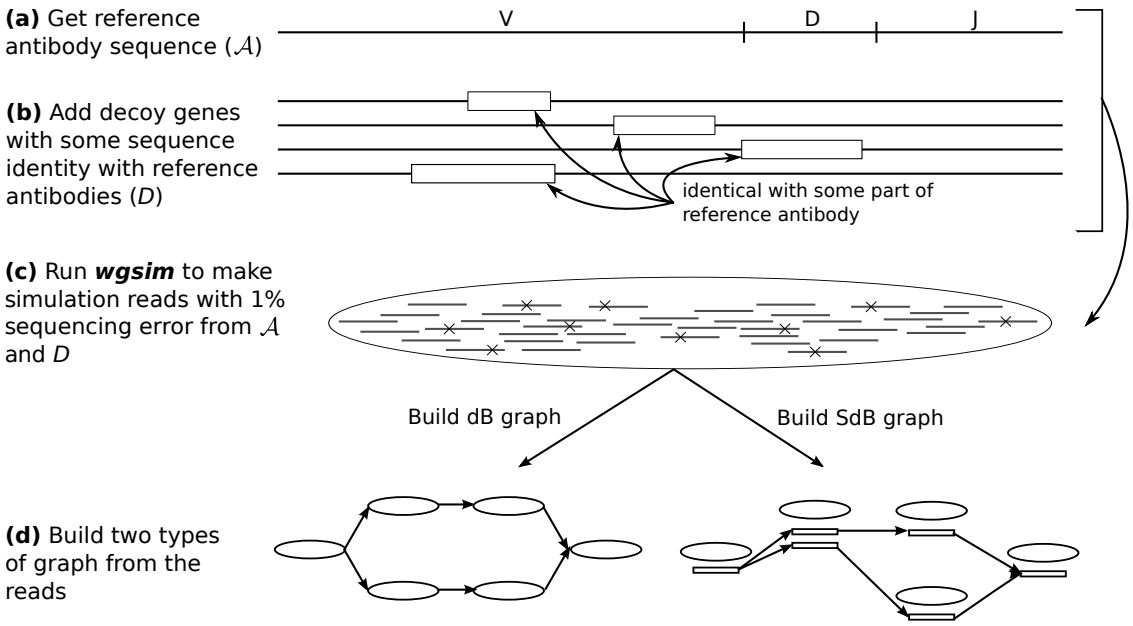


Figure D.7: **Antibody structure.** Antibodies are composed of four polypeptides: two identical copies of each of a light chain, and a heavy chain. The two light chain loci λ and κ are located at $22q11.2$ and $2p11.2$, respectively. There are 5 heavy chain sub-types, α, δ, ε, γ, and μ, all located at $14q32.33$. Both light and heavy chains contain variable and constant regions. The heavy chain genomic locus contains of V,D,J, and C segments. Recombination of V,D, and J segments forms a variable region. The recombined variable region sequence consists of stable, framework, regions (FR), as well as three hyper-variable complementarity determining regions (CDR). The constant region is formed by one out of five sub-types, and determines the type of the antibody (IgA,IgD,IgE,IgG, or IgM).



(e) Estimate the false negative rate ($\mathcal{F}$) and false positive rate ($\mathcal{D}$) of both graph

$$\mathcal{F} = \frac{|E \cap E_A|}{|E_A|}$$

$$\mathcal{D} = \frac{\Sigma_{n \in V_A}((n_i - 1) + (n_o - 1))}{|E_A|}$$

Figure D.8: **Simulation method.** (***a***) The reference VDJ combination of antibody sequence, $\mathcal{A}$, is obtained from IMGT database ($V : M99641, D : X97051, J : J00256$). **(*b*)** Random genes, D, are added with a part of the identical sequence being a different length than the reference. **(*c*)** The ***wgsim*** tool generated the simulation reads with a 1% error rate from the genes generated in Part ***a*** and ***b***. **(*d*)**. Build the dB and SdB graph, using the reads from Part ***c***. **(*e*)** False negative rate and divergence result can be estimated from the graph. The (V, E) value denote the set of nodes and edges of dB or SdB graph, and (V_A, E_A) denote the sets of nodes and edges of each graph using only $\mathcal{A}$. The n_i/n_o denote a number of incoming/outgoing edges of a node.

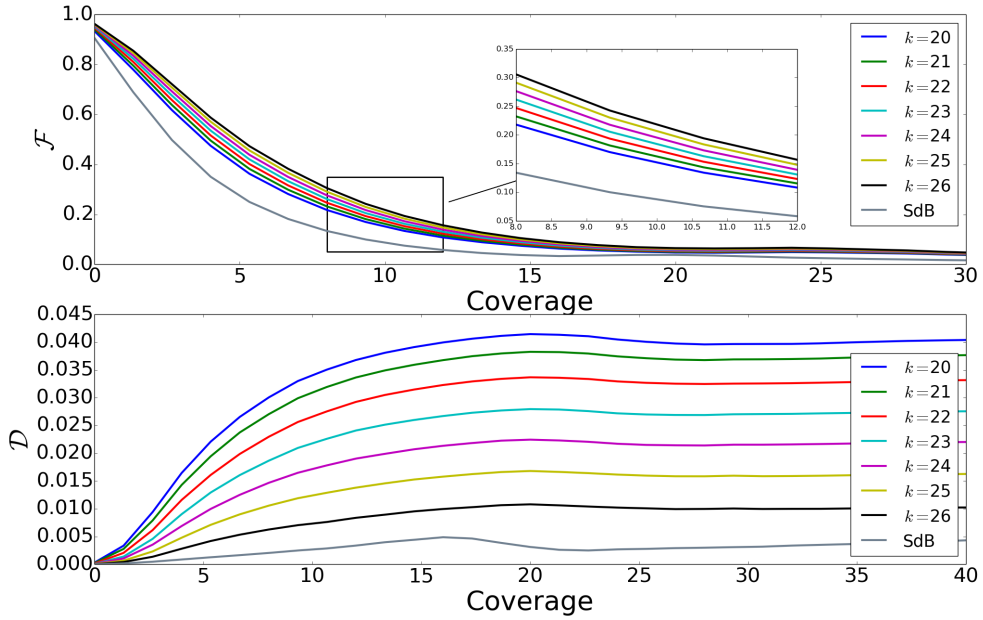


Figure D.9: **Comparison of SdB, and dB graphs on simulated data. (a)** The false negative rate $\mathcal{F}$ is the fraction of missing true edges. As k was increased from 20 to 26, the false negative rate increased monotonically. **(b)** The divergence $\mathcal{D}$ denotes the number of false edges normalized by the length of the true path. A coverage based filtering (filter out the edges with a low read depth) is applied to remove most of the sequencing errors in the reads before measuring $\mathcal{D}$. The divergence monotonically decreases with increase in k, while $\mathcal{F}$ increases. For all coverage values, SdB graphs performed better on both $\mathcal{F}$ and $\mathcal{D}$ measures.

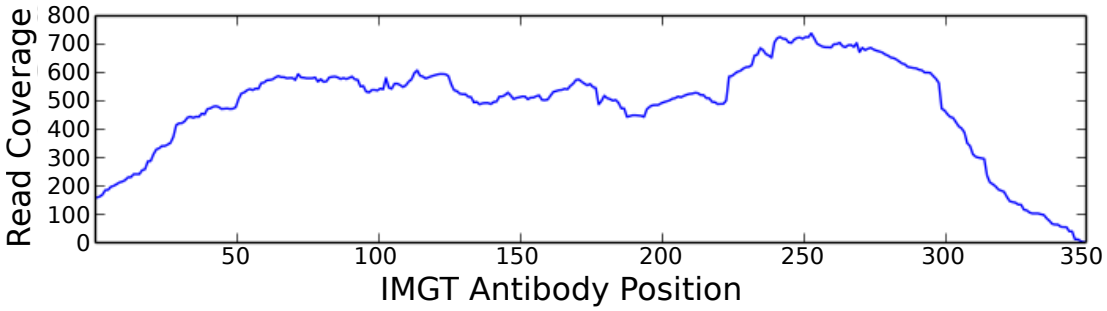
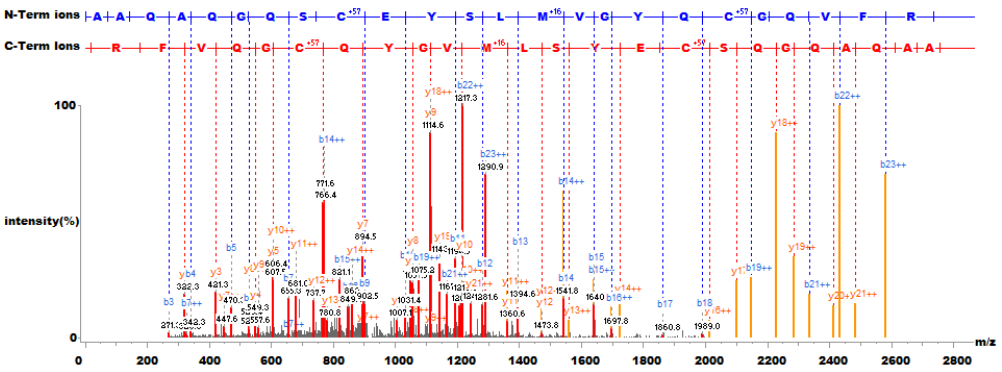
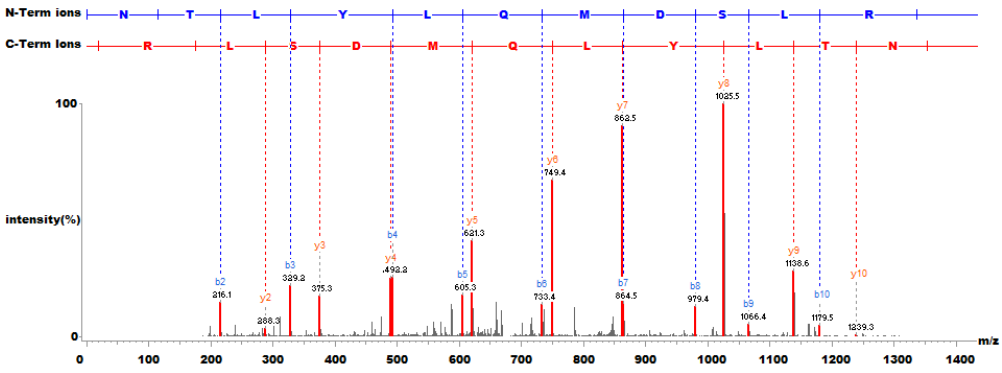


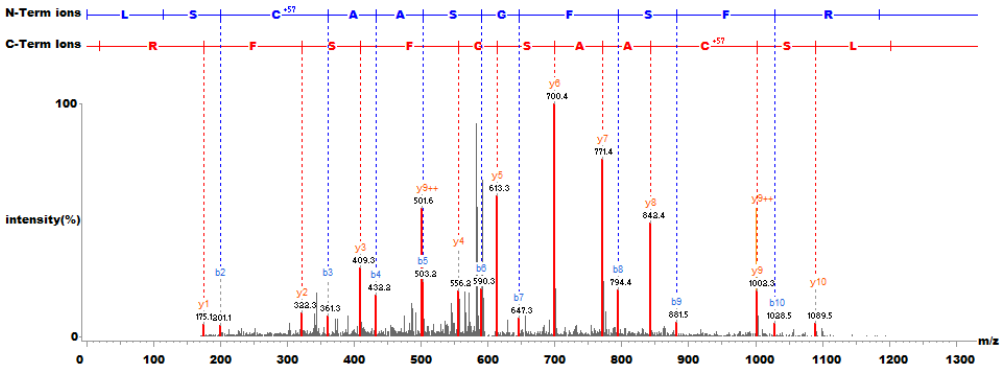
Figure D.10: **Number of reads mapping to the specific Ig gene position after the filtration:** For every read pass the filtration, we found the expected position of the read in the reference, and counted the number of reads pass each position of the reference.



(a)



(b)



(c)

Figure D.11: **Example PSMs of highly co-occurred peptides.** SAAV peptide *AAQAQGQ SCEYSLMVGYQCGQVF*$(Q \rightarrow R)$, Antibody peptide *NTLYLQMDSLR* and *LSCAASGFSFR* were an example of highly co-occurred peptides. Each figure shows the PSMs of these peptides. Note that the Fisher exact test p-value between SAAV peptide and *NTLYLQMDSLR* is 2.59×10^{-6} and *LSCAASGFSFR* is 9.93×10^{-5}

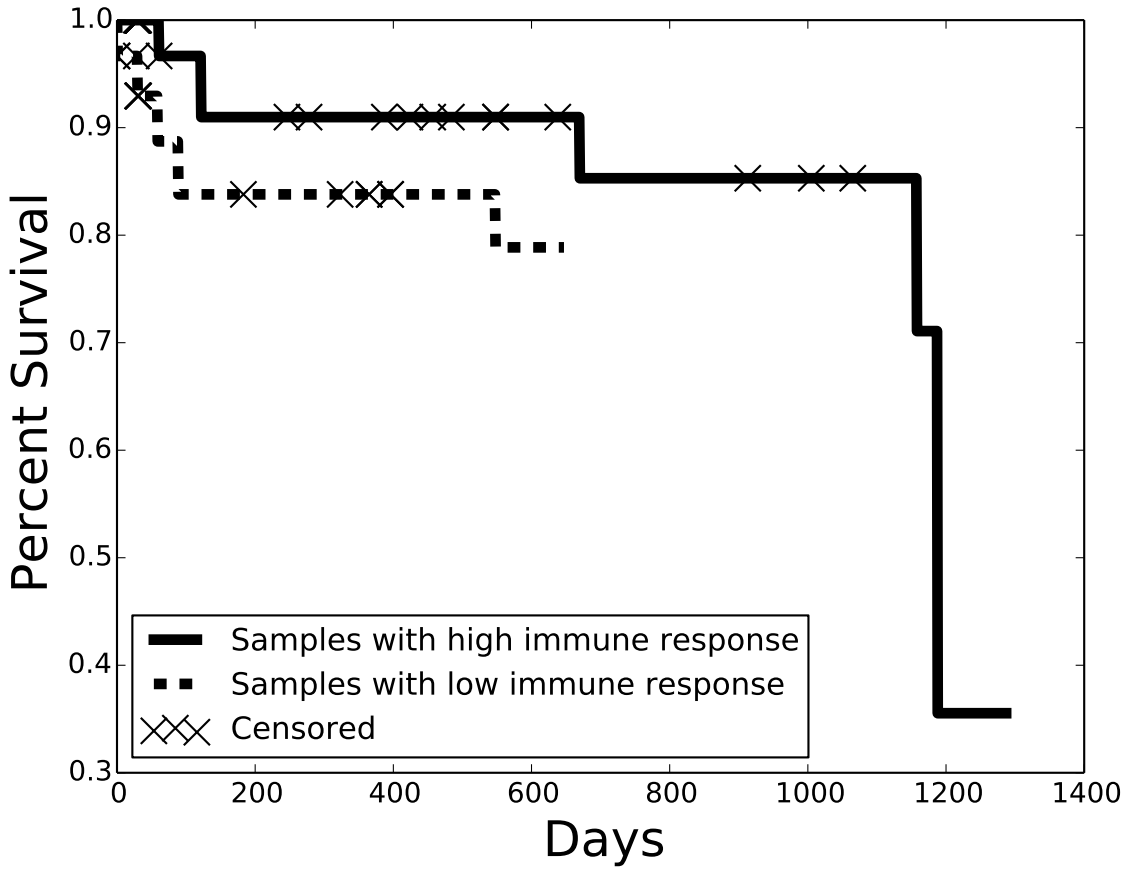


Figure D.12: **Kaplan-Meier survival estimator for patient with high vs low immune response.** The samples were divided into two groups based on the immune responses measured in Fig. 4.2(c). The clinical outcomes of each group were estimated by the Kaplan-Meier survival estimator (p-value $= 0.75$).

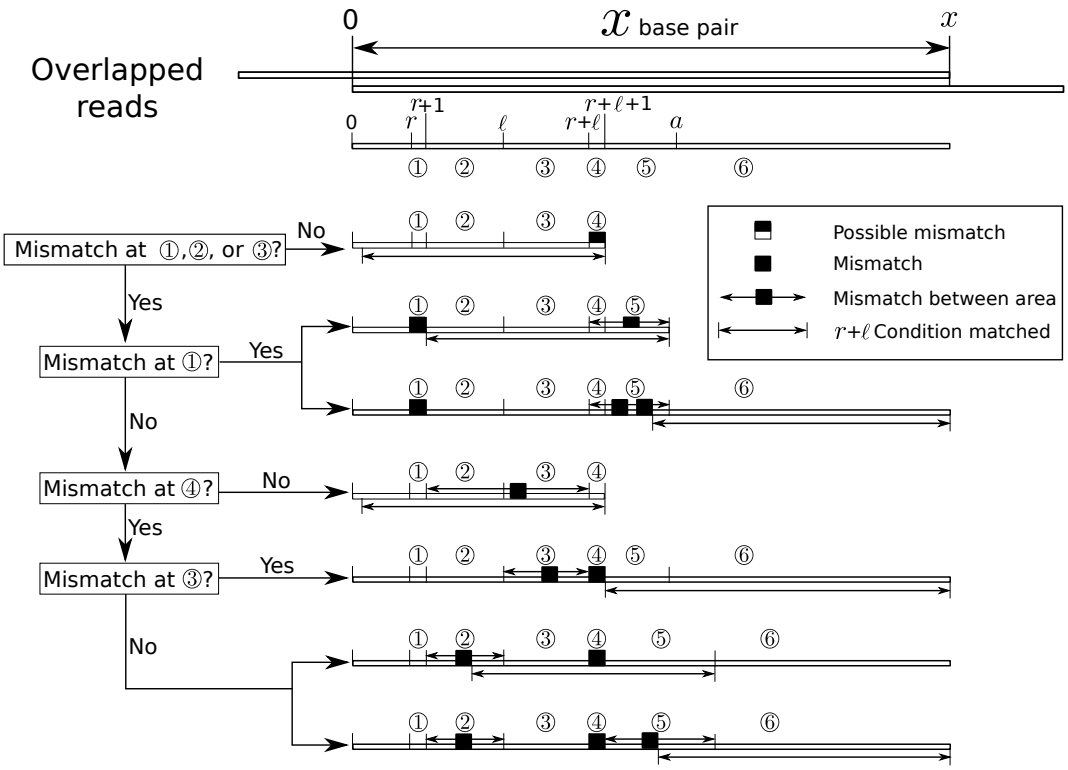
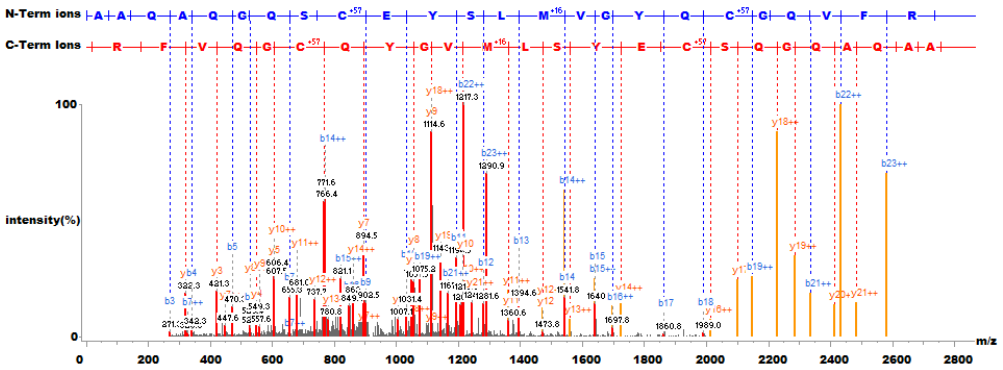
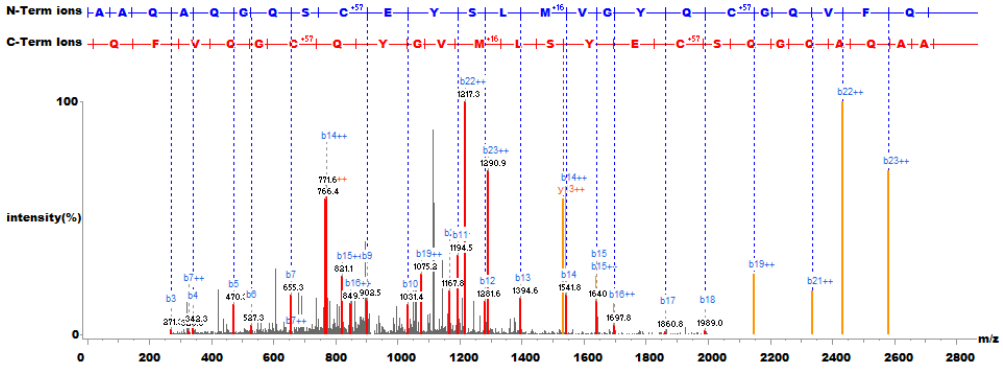


Figure D.13: **All possible conditions satisfying** $\mathcal{S}(x, r+\ell) \cap \mathcal{S}(x, -(x-1))$. Each line illustrated the possible cases described in the Supplemental Methods–'Computing false negative' of SdB graph in the order of $1, 2a, 2b, 3, 4a, 4b(i), 4b(ii)$.



(a)



(b)

Figure D.14: **Comparison between non-modified and modified version of SAAV example PSM.** SAAV peptide $AAQAQGQSCEYSLMVGYQCGQVF(Q \rightarrow R)$ shows poor matches $AAQAQGQSCEYSLMVGYQCGQVFQ$ on C-term ion sides suggesting that $(Q \rightarrow R)$ is a valid mutation.

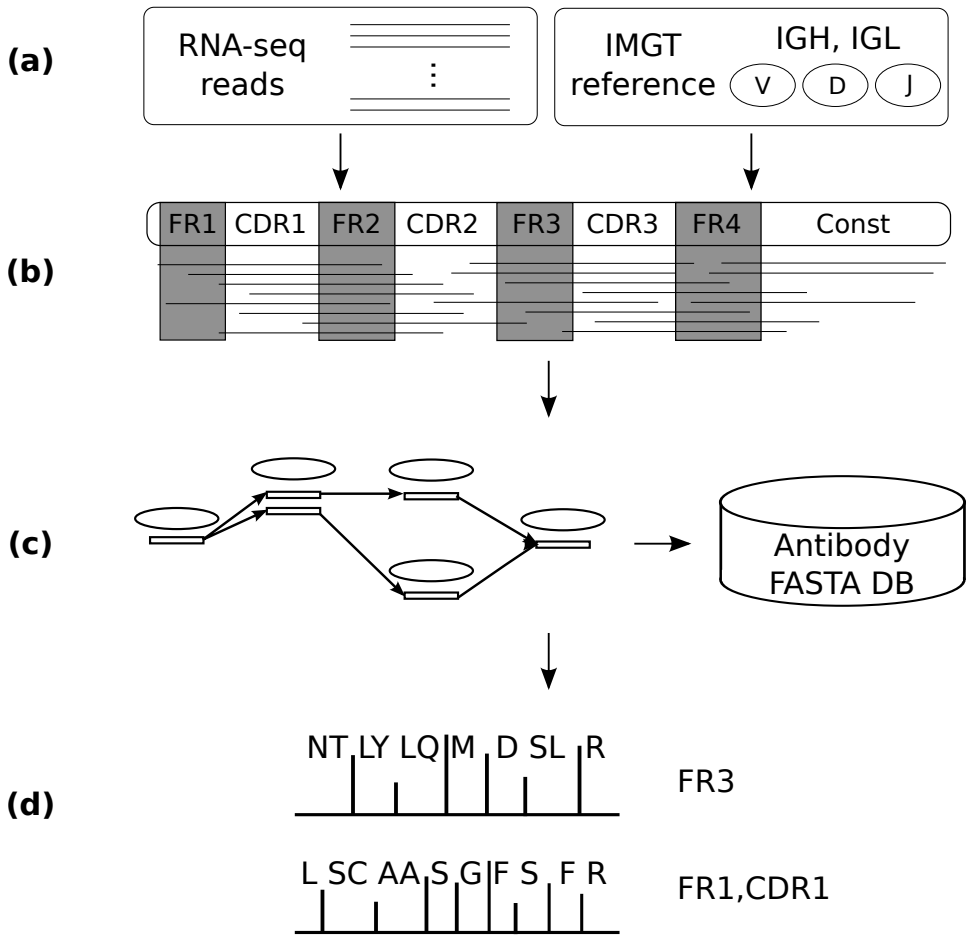


Figure D.15: **AbScan pipeline.** (a) AbScan takes two input files: the RNA-seq reads file and a reference antibody file downloaded from IMGT. (b) RNA-seq are filtered based on similarity to the IMGT reference. (c)An SdB graph is used for the read assembly, and the graph is used to construct a FASTA formatted amino-acid database for MS/MS search. (d) AbScan identify the PSMs and annotates peptides according to their derived location on a reference antibody.

Bibliography

[1] *Cancer Genomics Hub Repository*. UC Santa Cruz, https://cghub.ucsc.edu.

[2] *Clinical Proteomic Tumor Analysis Consortium*. Clinical Proteomic Tumor Analysis Consortium, http://proteomics.cancer.gov.

[3] *Ensembl*. http://uswest.ensembl.org/info/data/ftp/index.html.

[4] *SpliceDB*. UC San Diego, CCMS, http://bix.ucsd.edu/tmp/SpliceDB/SpliceDB.zip.

[5] J. Ai, Q. Tang, Y. Wu, Y. Xu, T. Feng, R. Zhou, Y. Chen, X. Gao, Q. Zhu, X. Yue, Q. Pan, S. Xu, J. Li, M. Huang, J. Daugherty-Holtrop, Y. He, H. E. Xu, J. Fan, J. Ding, and M. Geng. The role of polymeric immunoglobulin receptor in inflammation-induced tumor metastasis of human hepatocellular carcinoma. *J. Natl. Cancer Inst.*, 103(22):1696–1712, Nov 2011.

[6] Altschul, Stephen F and Gish, Warren and Miller, Webb and Myers, Eugene W and Lipman, David J and others. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[7] R. G. Amado, M. Wolf, M. Peeters, E. Van Cutsem, S. Siena, D. J. Freeman, T. Juan, R. Sikorski, S. Suggs, R. Radinsky, S. D. Patterson, and D. D. Chang. Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. *J. Clin. Oncol.*, 26(10):1626–1634, Apr 2008.

[8] K. Amara, J. Steen, F. Murray, H. Morbach, B. M. Fernandez-Rodriguez, V. Joshua, M. Engstrom, O. Snir, L. Israelsson, A. I. Catrina, H. Wardemann, D. Corti, E. Meffre, L. Klareskog, and V. Malmstrom. Monoclonal IgG antibodies generated from joint-derived B cells of RA patients have a strong bias toward citrullinated autoantigen recognition. *J. Exp. Med.*, 210(3):445–455, Mar 2013.

[9] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, and P. A. Pevzner. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, 19(5):455–477, May 2012.

[10] D. Bell, A. Berchuck, M. Birrer, J. Chien, and et al. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, Jun 2011.

[11] C. Boutros, A. Tarhini, E. Routier, O. Lambotte, F. L. Ladurie, F. Carbonnel, H. Izzeddine, A. Marabelle, S. Champiat, A. Berdelou, E. Lanoy, M. Texier, C. Libenciuc, A. M. Eggermont, J. C. Soria, C. Mateus, and C. Robert. Safety profiles of anti-CTLA-4 and anti-PD-1 antibodies alone and in combination. *Nat Rev Clin Oncol*, 13(8):473–486, Aug 2016.

[12] Sydney Brenner. The genetics of caenorhabditis elegans. *Genetics*, 77(1):71–94, 1974.

[13] N. Britzen-Laurent, K. Lipnik, M. Ocker, E. Naschberger, V. S. Schellerer, R. S. Croner, M. Vieth, M. Waldner, P. Steinberg, C. Hohenadl, and M. Sturzl. GBP-1 acts as a tumor suppressor in colorectal cancer cells. *Carcinogenesis*, 34(1):153–162, Jan 2013.

[14] J. Butler, I. MacCallum, M. Kleber, I. A. Shlyakhter, M. K. Belmonte, E. S. Lander, C. Nusbaum, and D. B. Jaffe. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res.*, 18(5):810–820, May 2008.

[15] L. H. Camacho, S. Antonia, J. Sosman, J. M. Kirkwood, T. F. Gajewski, B. Redman, D. Pavlov, C. Bulanhagui, V. A. Bozon, J. Gomez-Navarro, and A. Ribas. Phase I/II trial of tremelimumab in patients with metastatic melanoma. *J. Clin. Oncol.*, 27(7):1075–1081, Mar 2009.

[16] N. Castellana and V. Bafna. Proteogenomics to discover the full coding content of genomes: a computational perspective. *J Proteomics*, 73(11):2124–2135, Oct 2010.

[17] N. E. Castellana, S. H. Payne, Z. Shen, M. Stanke, V. Bafna, and S. P. Briggs. Discovery and revision of Arabidopsis genes by proteogenomics. *Proc. Natl. Acad. Sci. U.S.A.*, 105(52):21034–21038, Dec 2008.

[18] N. E. Castellana, Z. Shen, Y. He, J. W. Walley, and et al. An automated proteogenomic method uses mass spectrometry to reveal novel genes in Zea mays. *Mol. Cell Proteomics*, 13(1):157–167, Jan 2014.

[19] Phillip EC Compeau, Pavel A Pevzner, and Glenn Tesler. How to apply de bruijn graphs to genome assembly. *Nature biotechnology*, 29(11):987–991, 2011.

[20] F. Cunningham, M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. G. Giron, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, N. Johnson, T. Juettemann, A. K. Kahari, S. Keenan, F. J. Martin, T. Maurel, W. McLaren, D. N. Murphy, R. Nag, B. Overduin, A. Parker, M. Patricio, E. Perry, M. Pignatelli, H. S. Riat, D. Sheppard, K. Taylor, A. Thormann, A. Vullo, S. P. Wilder, A. Zadissa, B. L. Aken, E. Birney, J. Harrow, R. Kinsella, M. Muffato, M. Ruffier, S. M. Searle, G. Spudich, S. J. Trevanion, A. Yates, D. R. Zerbino, and P. Flicek. Ensembl 2015. *Nucleic Acids Res.*, 43(Database issue):D662–669, Jan 2015.

[21] E. Curran, L. Corrales, and J. Kline. Targeting the innate immune system as immunotherapy for acute myeloid leukemia. *Front Oncol*, 5:83, 2015.

[22] F. De Sousa E Melo, X. Wang, M. Jansen, E. Fessler, A. Trinh, L. P. de Rooij, J. H. de Jong, O. J. de Boer, R. van Leersum, M. F. Bijlsma, H. Rodermond, M. van der Heijden, C. J.

van Noesel, J. B. Tuynman, E. Dekker, F. Markowetz, J. P. Medema, and L. Vermeulen. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nat. Med.*, 19(5):614–618, May 2013.

[23] J. De Vries and C. Figdor. Immunotherapy: Cancer vaccine triggers antiviral-type defences. *Nature*, 534(7607):329–331, 06 2016.

[24] A. S. DePina, W. B. Iser, S. S. Park, S. Maudsley, M. A. Wilson, and C. A. Wolkow. Regulation of Caenorhabditis elegans vitellogenesis by DAF-2/IIS through separable transcriptional and posttranscriptional mechanisms. *BMC Physiol.*, 11:11, 2011.

[25] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, and et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, 43(5):491–498, May 2011.

[26] N. A. Doria-Rose, R. M. Klein, M. M. Manion, S. O'Dell, A. Phogat, B. Chakrabarti, C. W. Hallahan, S. A. Migueles, J. Wrammert, R. Ahmed, M. Nason, R. T. Wyatt, J. R. Mascola, and M. Connors. Frequency and phenotype of human immunodeficiency virus envelope-specific B cells from patients with broadly cross-neutralizing antibodies. *J. Virol.*, 83(1):188–199, Jan 2009.

[27] N. J. Edwards. Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. *Mol. Syst. Biol.*, 3:102, 2007.

[28] J. E. Elias and S. P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods*, 4(3):207–214, Mar 2007.

[29] J. K. Eng, T. A. Jahan, and M. R. Hoopmann. Comet: an open-source MS/MS sequence database search tool. *Proteomics*, 13(1):22–24, Jan 2013.

[30] V. C. Evans, G. Barker, K. J. Heesom, J. Fan, and et al. De novo derivation of proteomes from transcriptomes for transcript and protein identification. *Nat. Methods*, 9(12):1207–1211, Dec 2012.

[31] S. Fanayan, J. T. Smith, L. Y. Lee, F. Yan, M. Snyder, W. S. Hancock, and E. Nice. Proteogenomic analysis of human colon carcinoma cell lines LIM1215, LIM1899, and LIM2405. *J. Proteome Res.*, 12(4):1732–1742, Apr 2013.

[32] S. Fanayan, J. T. Smith, L. Y. Lee, F. Yan, M. Snyder, W. S. Hancock, and E. Nice. Proteogenomic analysis of human colon carcinoma cell lines LIM1215, LIM1899, and LIM2405. *J. Proteome Res.*, 12(4):1732–1742, Apr 2013.

[33] E. R. Fearon. Molecular genetics of colorectal cancer. *Annu Rev Pathol*, 6:479–507, 2011.

[34] P. Flicek, I. Ahmed, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, C. Garcia-Giron, L. Gordon, T. Hourlier, S. Hunt, T. Juettemann, A. K. Kahari, S. Keenan, M. Komorowska, E. Kulesha, I. Longden, T. Maurel, W. M. McLaren, M. Muffato, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, G. R. Ritchie, M. Ruffier, M. Schuster, D. Sheppard,

D. Sobral, K. Taylor, A. Thormann, S. Trevanion, S. White, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, J. Harrow, J. Herrero, T. J. Hubbard, N. Johnson, R. Kinsella, A. Parker, G. Spudich, A. Yates, A. Zadissa, and S. M. Searle. Ensembl 2013. *Nucleic Acids Res.*, 41(Database issue):48–55, Jan 2013.

[35] S. A. Forbes, N. Bindal, S. Bamford, C. Cole, and et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.*, 39(Database issue):D945–950, Jan 2011.

[36] Bettina. Franz, Kenneth F. May, Glenn. Dranoff, and Kai. Wucherpfennig. Ex vivo characterization and isolation of rare memory b cells with antigen tetramers. *Blood*, 2011.

[37] G. Georgiou, G. C. Ippolito, J. Beausang, C. E. Busse, H. Wardemann, and S. R. Quake. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.*, 32(2):158–168, Feb 2014.

[38] M. B. Gerstein, Z. J. Lu, E. L. Van Nostrand, C. Cheng, B. I. Arshinoff, T. Liu, K. Y. Yip, R. Robilotto, A. Rechtsteiner, K. Ikegami, P. Alves, A. Chateigner, M. Perry, M. Morris, R. K. Auerbach, X. Feng, J. Leng, A. Vielle, W. Niu, K. Rhrissorrakrai, A. Agarwal, R. P. Alexander, G. Barber, C. M. Brdlik, J. Brennan, J. J. Brouillet, A. Carr, M. S. Cheung, H. Clawson, S. Contrino, L. O. Dannenberg, A. F. Dernburg, A. Desai, L. Dick, A. C. Dose, J. Du, T. Egelhofer, S. Ercan, G. Euskirchen, B. Ewing, E. A. Feingold, R. Gassmann, P. J. Good, P. Green, F. Gullier, M. Gutwein, M. S. Guyer, L. Habegger, T. Han, J. G. Henikoff, S. R. Henz, A. Hinrichs, H. Holster, T. Hyman, A. L. Iniguez, J. Janette, M. Jensen, M. Kato, W. J. Kent, E. Kephart, V. Khivansara, E. Khurana, J. K. Kim, P. Kolasinska-Zwierz, E. C. Lai, I. Latorre, A. Leahey, S. Lewis, P. Lloyd, L. Lochovsky, R. F. Lowdon, Y. Lubling, R. Lyne, M. MacCoss, S. D. Mackowiak, M. Mangone, S. McKay, D. Mecenas, G. Merrihew, D. M. Miller, A. Muroyama, J. I. Murray, S. L. Ooi, H. Pham, T. Phippen, E. A. Preston, N. Rajewsky, G. Ratsch, H. Rosenbaum, J. Rozowsky, K. Rutherford, P. Ruzanov, M. Sarov, R. Sasidharan, A. Sboner, P. Scheid, E. Segal, H. Shin, C. Shou, F. J. Slack, C. Slightam, R. Smith, W. C. Spencer, E. O. Stinson, S. Taing, T. Takasaki, D. Vafeados, K. Voronina, G. Wang, N. L. Washington, C. M. Whittle, B. Wu, K. K. Yan, G. Zeller, Z. Zha, M. Zhong, X. Zhou, J. Ahringer, S. Strome, K. C. Gunsalus, G. Micklem, X. S. Liu, V. Reinke, S. K. Kim, L. W. Hillier, S. Henikoff, F. Piano, M. Snyder, L. Stein, J. D. Lieb, and R. H. Waterston. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science*, 330(6012):1775–1787, Dec 2010.

[39] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, 29(7):644–652, Jul 2011.

[40] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev.

Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, 29(7):644–652, May 2011.

[41] S. P. Gygi, Y. Rochon, B. R. Franza, and R. Aebersold. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.*, 19(3):1720–1730, Mar 1999.

[42] L. W. Hillier, V. Reinke, P. Green, M. Hirst, M. A. Marra, and R. H. Waterston. Massively parallel sequencing of the polyadenylated transcriptome of C. elegans. *Genome Res.*, 19(4):657–666, Apr 2009.

[43] F. S. Hodi, S. J. O'Day, D. F. McDermott, R. W. Weber, J. A. Sosman, J. B. Haanen, R. Gonzalez, C. Robert, D. Schadendorf, J. C. Hassel, W. Akerley, A. J. van den Eertwegh, J. Lutzky, P. Lorigan, J. M. Vaubel, G. P. Linette, D. Hogg, C. H. Ottensmeier, C. Lebbe, C. Peschel, I. Quirt, J. I. Clark, J. D. Wolchok, J. S. Weber, J. Tian, M. J. Yellin, G. M. Nichol, A. Hoos, and W. J. Urba. Improved survival with ipilimumab in patients with metastatic melanoma. *N. Engl. J. Med.*, 363(8):711–723, Aug 2010.

[44] A. N. Houghton and J. A. Guevara-Patino. Immune recognition of self in immunity against cancer. *J. Clin. Invest.*, 114(4):468–471, Aug 2004.

[45] W. Hueber and W. H. Robinson. Proteomic biomarkers for autoimmune disease. *Proteomics*, 6(14):4100–4105, Jul 2006.

[46] X. Jiang and J. R. Couchman. Perlecan and tumor angiogenesis. *J. Histochem. Cytochem.*, 51(11):1393–1410, Nov 2003.

[47] C. Jimenez-Luna, J. Prados, R. Ortiz, C. Melguizo, C. Torres, and O. Caba. Current Status of Immunotherapy Treatments for Pancreatic Cancer. *J. Clin. Gastroenterol.*, 50(10):836–848, 2016.

[48] Kent, W. J. and Sugnet, C. W. and Furey, T. S. and Roskin, K. M. and Pringle, T. H. and Zahler, A. M. and Haussler, D. . The human genome browser at UCSC. *Genome Res.*, 12(6):996–1006, Jun 2002.

[49] P. F. Kerkman, Y. Rombouts, E. I. van der Voort, L. A. Trouw, T. W. Huizinga, R. E. Toes, and H. U. Scherer. Circulating plasmablasts/plasmacells as a source of anticitrullinated protein antibodies in patients with rheumatoid arthritis. *Ann. Rheum. Dis.*, 72(7):1259–1263, Jul 2013.

[50] D. Kessner, M. Chambers, R. Burke, D. Agus, and P. Mallick. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics*, 24(21):2534–2536, Nov 2008.

[51] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, and et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, 14(4):R36, Apr 2013.

[52] D. Kim and S. L. Salzberg. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.*, 12(8):R72, 2011.

[53] M. S. Kim, S. M. Pinto, D. Getnet, R. S. Nirujogi, S. S. Manda, R. Chaerkady, A. K. Madugundu, D. S. Kelkar, R. Isserlin, S. Jain, J. K. Thomas, B. Muthusamy, P. Leal-Rojas, P. Kumar, N. A. Sahasrabuddhe, L. Balakrishnan, J. Advani, B. George, S. Renuse, L. D. Selvan, A. H. Patil, V. Nanjappa, A. Radhakrishnan, S. Prasad, T. Subbannayya, R. Raju, M. Kumar, S. K. Sreenivasamurthy, A. Marimuthu, G. J. Sathe, S. Chavan, K. K. Datta, Y. Subbannayya, A. Sahu, S. D. Yelamanchi, S. Jayaram, P. Rajagopalan, J. Sharma, K. R. Murthy, N. Syed, R. Goel, A. A. Khan, S. Ahmad, G. Dey, K. Mudgal, A. Chatterjee, T. C. Huang, J. Zhong, X. Wu, P. G. Shaw, D. Freed, M. S. Zahari, K. K. Mukherjee, S. Shankar, A. Mahadevan, H. Lam, C. J. Mitchell, S. K. Shankar, P. Satishchandra, J. T. Schroeder, R. Sirdeshmukh, A. Maitra, S. D. Leach, C. G. Drake, M. K. Halushka, T. S. Prasad, R. H. Hruban, C. L. Kerr, G. D. Bader, C. A. Iacobuzio-Donahue, H. Gowda, and A. Pandey. A draft map of the human proteome. *Nature*, 509(7502):575–581, May 2014.

[54] S. Kim, N. Mischerikow, N. Bandeira, J. D. Navarro, L. Wich, S. Mohammed, A. J. Heck, and P. A. Pevzner. The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. *Mol. Cell Proteomics*, 9(12):2840–2852, Dec 2010.

[55] S. Kim and P. A. Pevzner. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun*, 5:5277, Oct 2014.

[56] D. C. Koboldt, R. S. Fulton, M. D. McLellan, and H. et al. Schmidt. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, Oct 2012.

[57] W. C. Koff, D. R. Burton, P. R. Johnson, B. D. Walker, C. R. King, G. J. Nabel, R. Ahmed, M. K. Bhan, and S. A. Plotkin. Accelerating next-generation vaccine development for global disease prevention. *Science*, 340(6136):1232910, May 2013.

[58] L. Kottschade, A. Brys, T. Peikert, M. Ryder, L. Raffals, J. Brewer, P. Mosca, and S. Markovic. A multidisciplinary approach to toxicity management of modern immune checkpoint inhibitors in cancer therapy. *Melanoma Res.*, 26(5):469–480, Oct 2016.

[59] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M.

Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, and J. Szustakowki. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

[60] H. B. Larman, Z. Zhao, U. Laserson, M. Z. Li, A. Ciccia, M. A. Gakidis, G. M. Church, S. Kesari, E. M. Leproust, N. L. Solimini, and S. J. Elledge. Autoantigen discovery with a synthetic human peptidome. *Nat. Biotechnol.*, 29(6):535–541, May 2011.

[61] J. B. Legutki, Z. G. Zhao, M. Greving, N. Woodbury, S. A. Johnston, and P. Stafford. Scalable high-density peptide arrays for comprehensive health monitoring. *Nat Commun*, 5:4785, Sep 2014.

[62] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug 2009.

[63] J. Li, D. T. Duncan, and B. Zhang. CanProVar: a human cancer proteome variation database. *Hum. Mutat.*, 31(3):219–228, Mar 2010.

[64] J. Li, Z. Su, Z. Q. Ma, R. J. Slebos, and et al. A bioinformatics workflow for variant peptide detection in shotgun proteomics. *Mol. Cell Proteomics*, 10(5):M110.006536, May 2011.

[65] J. Li, Z. Su, Z. Q. Ma, R. J. Slebos, P. Halvey, D. L. Tabb, D. C. Liebler, W. Pao, and B. Zhang. A bioinformatics workflow for variant peptide detection in shotgun proteomics. *Mol. Cell Proteomics*, 10(5):M110.006536, May 2011.

[66] X. Li, Z. Wu, Y. Wang, Q. Mei, X. Fu, and W. Han. Characterization of adult - and -globin elevated by hydrogen peroxide in cervical cancer cells that play a cytoprotective role against oxidative insults. *PLoS ONE*, 8(1):e54342, 2013.

[67] Michael Linnebacher. Tumor-infiltrating b cells come into vogue. *World journal of gastroenterology: WJG*, 19(1):8, 2013.

[68] F. Liu. SMAD4/DPC4 and pancreatic cancer survival. Commentary re: M. Tascilar et al., The SMAD4 protein and prognosis of pancreatic ductal adenocarcinoma. Clin. Cancer Res., 7: 4115-4121, 2001. *Clin. Cancer Res.*, 7(12):3853–3856, Dec 2001.

[69] P. L. Lollini, F. Cavallo, P. Nanni, and G. Forni. Vaccines for tumour prevention. *Nat. Rev. Cancer*, 6(3):204–216, Mar 2006.

[70] editors. McEntyre J, Ostell J, editor. *The NCBI Handbook [Internet]*. National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/books/NBK21101/, 2002-.

[71] M. McHeyzer-Williams, S. Okitsu, N. Wang, and L. McHeyzer-Williams. Molecular programming of B cell memory. *Nat. Rev. Immunol.*, 12(1):24–34, Dec 2011.

[72] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, and et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, 20(9):1297–1303, Sep 2010.

[73] G. E. Merrihew, C. Davis, B. Ewing, G. Williams, L. Kall, B. E. Frewen, W. S. Noble, P. Green, J. H. Thomas, and M. J. MacCoss. Use of shotgun proteomics for the identification, confirmation, and correction of C. elegans gene annotations. *Genome Res.*, 18(10):1660–1669, Oct 2008.

[74] M. Miyaki and T. Kuroki. Role of Smad4 (DPC4) inactivation in human cancer. *Biochem. Biophys. Res. Commun.*, 306(4):799–804, Jul 2003.

[75] D. M. Muzny, M. N. Bainbridge, K. Chang, H. H. Dinh, J. A. Drummond, G. Fowler, C. L. Kovar, L. R. Lewis, M. B. Morgan, I. F. Newsham, J. G. Reid, J. Santibanez, E. Shinbrot, L. R. Trevino, Y. Q. Wu, M. Wang, P. Gunaratne, L. A. Donehower, C. J. Creighton, D. A. Wheeler, R. A. Gibbs, M. S. Lawrence, D. Voet, R. Jing, K. Cibulskis, A. Sivachenko, P. Stojanov, A. McKenna, E. S. Lander, S. Gabriel, G. Getz, L. Ding, R. S. Fulton, D. C. Koboldt, T. Wylie, J. Walker, D. J. Dooling, L. Fulton, K. D. Delehaunty, C. C. Fronick, R. Demeter, E. R. Mardis, R. K. Wilson, A. Chu, H. J. Chun, A. J. Mungall, E. Pleasance, A. Robertson, D. Stoll, M. Balasundaram, I. Birol, Y. S. Butterfield, E. Chuah, R. J. Coope, N. Dhalla, R. Guin, C. Hirst, M. Hirst, R. A. Holt, D. Lee, H. I. Li, M. Mayo, R. A. Moore, J. E. Schein, J. R. Slobodan, A. Tam, N. Thiessen, R. Varhol, T. Zeng, Y. Zhao, S. J. Jones, M. A. Marra, A. J. Bass, A. H. Ramos, G. Saksena, A. D. Cherniack, S. E. Schumacher, B. Tabak, S. L. Carter, N. H. Pho, H. Nguyen, R. C. Onofrio, A. Crenshaw, K. Ardlie, R. Beroukhim, W. Winckler, G. Getz, M. Meyerson, A. Protopopov, J. Zhang, A. Hadjipanayis, E. Lee, R. Xi, L. Yang, X. Ren, H. Zhang, N. Sathiamoorthy, S. Shukla, P. C. Chen, P. Haseley, Y. Xiao, S. Lee, J. Seidman, L. Chin, P. J. Park, R. Kucherlapati, J. T. Auman, K. A. Hoadley, Y. Du, M. D. Wilkerson, Y. Shi, C. Liquori, S. Meng, L. Li, Y. J. Turman, M. D. Topal, D. Tan,

S. Waring, E. Buda, J. Walsh, C. D. Jones, P. A. Mieczkowski, D. Singh, J. Wu, A. Gulabani, P. Dolina, T. Bodenheimer, A. P. Hoyle, J. V. Simons, M. Soloway, L. E. Mose, S. R. Jefferys, S. Balu, B. D. O'Connor, J. F. Prins, D. Y. Chiang, D. Hayes, C. M. Perou, T. Hinoue, D. J. Weisenberger, D. T. Maglinte, F. Pan, B. P. Berman, D. J. Van Den Berg, H. Shen, T. Triche, S. B. Baylin, P. W. Laird, G. Getz, M. Noble, D. Voet, G. Saksena, N. Gehlenborg, D. DiCara, J. Zhang, H. Zhang, C. J. Wu, S. Y. Liu, S. Shukla, M. S. Lawrence, L. Zhou, A. Sivachenko, P. Lin, P. Stojanov, R. Jing, R. W. Park, M. D. Nazaire, J. Robinson, H. Thorvaldsdottir, J. Mesirov, P. J. Park, L. Chin, V. Thorsson, S. M. Reynolds, B. Bernard, R. Kreisberg, J. Lin, L. Iype, R. Bressler, T. Erkkila, M. Gundapuneni, Y. Liu, A. Norberg, T. Robinson, D. Yang, W. Zhang, I. Shmulevich, J. J. de Ronde, N. Schultz, E. Cerami, G. Ciriello, A. P. Goldberg, B. Gross, A. Jacobsen, J. Gao, B. Kaczkowski, R. Sinha, B. Aksoy, Y. Antipin, B. Reva, R. Shen, B. S. Taylor, T. A. Chan, M. Ladanyi, C. Sander, R. Akbani, N. Zhang, B. M. Broom, T. Casasent, A. Unruh, C. Wakefield, S. R. Hamilton, R. Cason, K. A. Baggerly, J. N. Weinstein, D. Haussler, C. C. Benz, J. M. Stuart, S. C. Benz, J. Sanborn, C. J. Vaske, J. Zhu, C. Szeto, G. K. Scott, C. Yau, S. Ng, T. Goldstein, K. Ellrott, E. Collisson, A. E. Cozen, D. Zerbino, C. Wilks, B. Craft, P. Spellman, R. Penny, T. Shelton, M. Hatfield, S. Morris, P. Yena, C. Shelton, M. Sherman, J. Paulauskis, J. M. Gastier-Foster, J. Bowen, N. C. Ramirez, A. Black, R. Pyatt, L. Wise, P. White, M. Bertagnolli, J. Brown, T. A. Chan, G. C. Chu, C. Czerwinski, F. Denstman, R. Dhir, A. Dorner, C. S. Fuchs, J. G. Guillem, M. Iacocca, H. Juhl, A. Kaufman, B. Kohl, X. Van Le, M. C. Mariano, E. N. Medina, M. Meyers, G. M. Nash, P. B. Paty, N. Petrelli, B. Rabeno, W. G. Richards, D. Solit, P. Swanson, L. Temple, J. E. Tepper, R. Thorp, E. Vakiani, M. R. Weiser, J. E. Willis, G. Witkin, Z. Zeng, M. J. Zinner, C. Zornig, M. A. Jensen, R. Sfeir, A. B. Kahn, A. L. Chu, P. Kothiyal, Z. Wang, E. E. Snyder, J. Pontius, T. D. Pihl, B. Ayala, M. Backus, J. Walton, J. Whitmore, J. Baboud, D. L. Berton, M. C. Nicholls, D. Srinivasan, R. Raman, S. Girshik, P. A. Kigonya, S. Alonso, R. N. Sanbhadti, S. P. Barletta, J. M. Greene, D. A. Pot, K. R. Shaw, L. A. Dillon, K. Buetow, T. Davidsen, J. A. Demchok, G. Eley, M. Ferguson, P. Fielding, C. Schaefer, M. Sheth, L. Yang, M. S. Guyer, B. A. Ozenberger, J. D. Palchik, J. Peterson, H. J. Sofia, and E. Thomson. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, Jul 2012.

[76] D. M. Muzny, M. N. Bainbridge, K. Chang, H. H. Dinh, J. A. Drummond, G. Fowler, C. L. Kovar, L. R. Lewis, M. B. Morgan, I. F. Newsham, J. G. Reid, J. Santibanez, E. Shinbrot, L. R. Trevino, Y. Q. Wu, M. Wang, P. Gunaratne, L. A. Donehower, C. J. Creighton, D. A. Wheeler, R. A. Gibbs, M. S. Lawrence, D. Voet, R. Jing, K. Cibulskis, A. Sivachenko, P. Stojanov, A. McKenna, E. S. Lander, S. Gabriel, G. Getz, L. Ding, R. S. Fulton, D. C. Koboldt, T. Wylie, J. Walker, D. J. Dooling, L. Fulton, K. D. Delehaunty, C. C. Fronick, R. Demeter, E. R. Mardis, R. K. Wilson, A. Chu, H. J. Chun, A. J. Mungall, E. Pleasance, A. Robertson, D. Stoll, M. Balasundaram, I. Birol, Y. S. Butterfield, E. Chuah, R. J. Coope, N. Dhalla, R. Guin, C. Hirst, M. Hirst, R. A. Holt, D. Lee, H. I. Li, M. Mayo, R. A. Moore, J. E. Schein, J. R. Slobodan, A. Tam, N. Thiessen, R. Varhol, T. Zeng, Y. Zhao, S. J. Jones, M. A. Marra, A. J. Bass, A. H. Ramos, G. Saksena, A. D. Cherniack, S. E. Schumacher, B. Tabak, S. L. Carter, N. H. Pho, H. Nguyen, R. C. Onofrio, A. Crenshaw, K. Ardlie, R. Beroukhim, W. Winckler, G. Getz, M. Meyerson, A. Protopopov, J. Zhang, A. Hadjipanayis, E. Lee, R. Xi, L. Yang, X. Ren, H. Zhang, N. Sathiamoorthy, S. Shukla, P. C. Chen, P. Haseley, Y. Xiao,

S. Lee, J. Seidman, L. Chin, P. J. Park, R. Kucherlapati, J. T. Auman, K. A. Hoadley, Y. Du, M. D. Wilkerson, Y. Shi, C. Liquori, S. Meng, L. Li, Y. J. Turman, M. D. Topal, D. Tan, S. Waring, E. Buda, J. Walsh, C. D. Jones, P. A. Mieczkowski, D. Singh, J. Wu, A. Gulabani, P. Dolina, T. Bodenheimer, A. P. Hoyle, J. V. Simons, M. Soloway, L. E. Mose, S. R. Jefferys, S. Balu, B. D. O'Connor, J. F. Prins, D. Y. Chiang, D. Hayes, C. M. Perou, T. Hinoue, D. J. Weisenberger, D. T. Maglinte, F. Pan, B. P. Berman, D. J. Van Den Berg, H. Shen, T. Triche, S. B. Baylin, P. W. Laird, G. Getz, M. Noble, D. Voet, G. Saksena, N. Gehlenborg, D. DiCara, J. Zhang, H. Zhang, C. J. Wu, S. Y. Liu, S. Shukla, M. S. Lawrence, L. Zhou, A. Sivachenko, P. Lin, P. Stojanov, R. Jing, R. W. Park, M. D. Nazaire, J. Robinson, H. Thorvaldsdottir, J. Mesirov, P. J. Park, L. Chin, V. Thorsson, S. M. Reynolds, B. Bernard, R. Kreisberg, J. Lin, L. Iype, R. Bressler, T. Erkkila, M. Gundapuneni, Y. Liu, A. Norberg, T. Robinson, D. Yang, W. Zhang, I. Shmulevich, J. J. de Ronde, N. Schultz, E. Cerami, G. Ciriello, A. P. Goldberg, B. Gross, A. Jacobsen, J. Gao, B. Kaczkowski, R. Sinha, B. Aksoy, Y. Antipin, B. Reva, R. Shen, B. S. Taylor, T. A. Chan, M. Ladanyi, C. Sander, R. Akbani, N. Zhang, B. M. Broom, T. Casasent, A. Unruh, C. Wakefield, S. R. Hamilton, R. Cason, K. A. Baggerly, J. N. Weinstein, D. Haussler, C. C. Benz, J. M. Stuart, S. C. Benz, J. Sanborn, C. J. Vaske, J. Zhu, C. Szeto, G. K. Scott, C. Yau, S. Ng, T. Goldstein, K. Ellrott, E. Collisson, A. E. Cozen, D. Zerbino, C. Wilks, B. Craft, P. Spellman, R. Penny, T. Shelton, M. Hatfield, S. Morris, P. Yena, C. Shelton, M. Sherman, J. Paulauskis, J. M. Gastier-Foster, J. Bowen, N. C. Ramirez, A. Black, R. Pyatt, L. Wise, P. White, M. Bertagnolli, J. Brown, T. A. Chan, G. C. Chu, C. Czerwinski, F. Denstman, R. Dhir, A. Dorner, C. S. Fuchs, J. G. Guillem, M. Iacocca, H. Juhl, A. Kaufman, B. Kohl, X. Van Le, M. C. Mariano, E. N. Medina, M. Meyers, G. M. Nash, P. B. Paty, N. Petrelli, B. Rabeno, W. G. Richards, D. Solit, P. Swanson, L. Temple, J. E. Tepper, R. Thorp, E. Vakiani, M. R. Weiser, J. E. Willis, G. Witkin, Z. Zeng, M. J. Zinner, C. Zornig, M. A. Jensen, R. Sfeir, A. B. Kahn, A. L. Chu, P. Kothiyal, Z. Wang, E. E. Snyder, J. Pontius, T. D. Pihl, B. Ayala, M. Backus, J. Walton, J. Whitmore, J. Baboud, D. L. Berton, M. C. Nicholls, D. Srinivasan, R. Raman, S. Girshik, P. A. Kigonya, S. Alonso, R. N. Sanbhadti, S. P. Barletta, J. M. Greene, D. A. Pot, K. R. Shaw, L. A. Dillon, K. Buetow, T. Davidsen, J. A. Demchok, G. Eley, M. Ferguson, P. Fielding, C. Schaefer, M. Sheth, L. Yang, M. S. Guyer, B. A. Ozenberger, J. D. Palchik, J. Peterson, H. J. Sofia, and E. Thomson. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, Jul 2012.

[77] B. H. Nelson. CD20+ B cells: the other tumor-infiltrating lymphocytes. *J. Immunol.*, 185(9):4977–4982, Nov 2010.

[78] Julie S Nielsen, Rob A Sahota, Katy Milne, Sara E Kost, Nancy J Nesslinger, Peter H Watson, and Brad H Nelson. Cd20+ tumor-infiltrating lymphocytes have an atypical cd27- memory phenotype and together with cd8+ t cells promote favorable prognosis in ovarian cancer. *Clinical Cancer Research*, 18(12):3281–3292, 2012.

[79] S. I. Nikolenko, A. I. Korobeynikov, and M. A. Alekseyev. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics*, 14 Suppl 1:S7, 2013.

[80] S. Nzula, J. J. Going, and D. I. Stott. Antigen-driven clonal proliferation, somatic hypermuta-

tion, and selection of B lymphocytes infiltrating human ductal breast carcinomas. *Cancer Res.*, 63(12):3275–3280, Jun 2003.

[81] S. Ogino, K. Nosho, N. Irahara, J. A. Meyerhardt, Y. Baba, K. Shima, J. N. Glickman, C. R. Ferrone, M. Mino-Kenudson, N. Tanaka, G. Dranoff, E. L. Giovannucci, and C. S. Fuchs. Lymphocytic reaction to colorectal cancer is associated with longer survival, independent of lymph node count, microsatellite instability, and CpG island methylator phenotype. *Clin. Cancer Res.*, 15(20):6412–6420, Oct 2009.

[82] M. Orciani, O. Trubiani, S. Guarnieri, E. Ferrero, and R. Di Primio. CD38 is constitutively expressed in the nucleus of human hematopoietic cells. *J. Cell. Biochem.*, 105(3):905–912, Oct 2008.

[83] S. Partida-Sanchez, L. Rivero-Nava, G. Shi, and F. E. Lund. CD38: an ecto-enzyme at the crossroads of innate and adaptive immune responses. *Adv. Exp. Med. Biol.*, 590:171–183, 2007.

[84] P. G. Pedrioli, J. K. Eng, R. Hubley, M. Vogelzang, E. W. Deutsch, B. Raught, B. Pratt, E. Nilsson, R. H. Angeletti, R. Apweiler, K. Cheung, C. E. Costello, H. Hermjakob, S. Huang, R. K. Julian, E. Kapp, M. E. McComb, S. G. Oliver, G. Omenn, N. W. Paton, R. Simpson, R. Smith, C. F. Taylor, W. Zhu, and R. Aebersold. A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.*, 22(11):1459–1466, Nov 2004.

[85] Yu. Peng, Henry C. M. Leung, Siu-Ming. Yiu, Ming-Ju. Lv, Xin-Guang. Zhu, and Francis Y. L. Chin. IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. *Bioinformatics (Oxford, England)*, 2013.

[86] P. A. Pevzner, H. Tang, and M. S. Waterman. An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. U.S.A.*, 98(17):9748–9753, Aug 2001.

[87] Pavel A Pevzner, Haixu Tang, and Michael S Waterman. An eulerian path approach to dna fragment assembly. *Proceedings of the National Academy of Sciences*, 98(17):9748–9753, 2001.

[88] E. Phizicky, P. I. Bastiaens, H. Zhu, M. Snyder, and S. Fields. Protein analysis on a proteomic scale. *Nature*, 422(6928):208–215, Mar 2003.

[89] L. Ponnala, Y. Wang, Q. Sun, and K. J. van Wijk. Correlation of mRNA and protein abundance in the developing maize leaf. *Plant J.*, Feb 2014.

[90] J. V. Price, S. Tangsombatvisit, G. Xu, J. Yu, D. Levy, E. C. Baechler, O. Gozani, M. Varma, P. J. Utz, and C. L. Liu. On silico peptide microarrays for high-resolution mapping of antibody epitopes and diverse protein-protein interactions. *Nat. Med.*, 18(9):1434–1440, Sep 2012.

[91] K. D. Pruitt, G. R. Brown, S. M. Hiatt, F. Thibaud-Nissen, A. Astashyn, O. Ermolaeva, C. M. Farrell, J. Hart, M. J. Landrum, K. M. McGarvey, M. R. Murphy, N. A. O'Leary,

S. Pujar, B. Rajput, S. H. Rangwala, L. D. Riddick, A. Shkeda, H. Sun, P. Tamez, R. E. Tully, C. Wallin, D. Webb, J. Weber, W. Wu, M. DiCuccio, P. Kitts, D. R. Maglott, T. D. Murphy, and J. M. Ostell. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, 42(Database issue):D756–763, Jan 2014.

[92] K. D. Pruitt, T. Tatusova, W. Klimke, and D. R. Maglott. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.*, 37(Database issue):D32–36, Jan 2009.

[93] J. C. Riches and J. G. Gribben. Immunomodulation and immune reconstitution in chronic lymphocytic leukemia. *Semin. Hematol.*, 51(3):228–234, Jul 2014.

[94] W. H. Robinson. Sequencing the functional antibody repertoire–diagnostic and therapeutic discovery. *Nat Rev Rheumatol*, 11(3):171–182, Mar 2015.

[95] W. H. Robinson, C. DiGennaro, W. Hueber, B. B. Haab, M. Kamachi, E. J. Dean, S. Fournel, D. Fong, M. C. Genovese, H. E. de Vegvar, K. Skriner, D. L. Hirschberg, R. I. Morris, S. Muller, G. J. Pruijn, W. J. van Venrooij, J. S. Smolen, P. O. Brown, L. Steinman, and P. J. Utz. Autoantigen microarrays for multiplex characterization of autoantibody responses. *Nat. Med.*, 8(3):295–301, Mar 2002.

[96] R. Ronen, C. Boucher, H. Chitsaz, and P. Pevzner. SEQuel: improving the accuracy of genome assemblies. *Bioinformatics*, 28(12):i188–196, Jun 2012.

[97] S. A. Rosenberg, J. C. Yang, and N. P. Restifo. Cancer immunotherapy: moving beyond current vaccines. *Nat. Med.*, 10(9):909–915, Sep 2004.

[98] M. Ruiz and M. P. Lefranc. IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. *Immunogenetics*, 53(10-11):857–883, Feb 2002.

[99] A. Sadanandam, C. A. Lyssiotis, K. Homicsko, E. A. Collisson, W. J. Gibb, S. Wullschleger, L. C. Ostos, W. A. Lannon, C. Grotzinger, M. Del Rio, B. Lhermitte, A. B. Olshen, B. Wiedenmann, L. C. Cantley, J. W. Gray, and D. Hanahan. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat. Med.*, 19(5):619–625, May 2013.

[100] Y. Safonova, S. Bonissone, E. Kurpilyansky, E. Starostina, A. Lapidus, J. Stinson, L. DePalatis, W. Sandoval, J. Lill, and P. A. Pevzner. IgRepertoireConstructor: a novel algorithm for antibody repertoire construction and immunoproteogenomics analysis. *Bioinformatics*, 31(12):53–61, Jun 2015.

[101] R. Salgado, C. Denkert, S. Demaria, N. Sirtaine, F. Klauschen, G. Pruneri, S. Wienert, G. Van den Eynden, F. L. Baehner, F. Penault-Llorca, E. A. Perez, E. A. Thompson, W. F. Symmans, A. L. Richardson, J. Brock, C. Criscitiello, H. Bailey, M. Ignatiadis, G. Floris, J. Sparano, Z. Kos, T. Nielsen, D. L. Rimm, K. H. Allison, J. S. Reis-Filho, S. Loibl, C. Sotiriou, G. Viale, S. Badve, S. Adams, K. Willard-Gallo, and S. Loi. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. *Ann. Oncol.*, 26(2):259–271, Feb 2015.

[102] J. H. Sampson, G. E. Archer, D. A. Mitchell, A. B. Heimberger, and D. D. Bigner. Tumor-specific immunotherapy targeting the EGFRvIII mutation in patients with malignant glioma. *Semin. Immunol.*, 20(5):267–275, Oct 2008.

[103] D. Scaviner, V. Barbie, M. Ruiz, and M. P. Lefranc. Protein displays of the human immunoglobulin heavy, kappa and lambda variable and joining regions. *Exp. Clin. Immunogenet.*, 16(4):234–240, 1999.

[104] J. Q. Sheng, S. R. Li, Z. T. Wu, C. H. Xia, X. Wu, J. Chen, and J. Rao. Transferrin dipstick as a potential novel test for colon cancer screening: a comparative study with immuno fecal occult blood test. *Cancer Epidemiol. Biomarkers Prev.*, 18(8):2182–2185, Aug 2009.

[105] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, and et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, 29(1):308–311, Jan 2001.

[106] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, 29(1):308–311, Jan 2001.

[107] Y. Shigematsu, T. Hanagiri, K. Kuroda, T. Baba, M. Mizukami, Y. Ichiki, M. Yasuda, M. Takenoyama, K. Sugio, and K. Yasumoto. Malignant mesothelioma-associated antigens recognized by tumor-infiltrating B cells and the clinical significance of the antibody titers. *Cancer Sci.*, 100(7):1326–1334, Jul 2009.

[108] N. Siva. 1000 genomes project. *Nature biotechnology*, 26(3):256–256, 2008.

[109] J. N. Stern, G. Yaari, J. A. Vander Heiden, G. Church, W. F. Donahue, R. Q. Hintzen, A. J. Huttner, J. D. Laman, R. M. Nagra, A. Nylander, D. Pitt, S. Ramanan, B. A. Siddiqui, F. Vigneault, S. H. Kleinstein, D. A. Hafler, and K. C. O'Connor. B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Sci Transl Med*, 6(248):248ra107, Aug 2014.

[110] Y. Takahashi, G. Sawada, J. Kurashige, T. Matsumura, R. Uchi, H. Ueo, M. Ishibashi, Y. Takano, S. Akiyoshi, T. Iwaya, H. Eguchi, T. Sudo, K. Sugimachi, H. Yamamoto, Y. Doki, M. Mori, and K. Mimori. Tumor-derived tenascin-C promotes the epithelial-mesenchymal transition in colorectal cancer cells. *Anticancer Res.*, 33(5):1927–1934, May 2013.

[111] Y. C. Tan, S. Kongpachith, L. K. Blum, C. H. Ju, L. J. Lahey, D. R. Lu, X. Cai, C. A. Wagner, T. M. Lindstrom, J. Sokolove, and W. H. Robinson. Barcode-enabled sequencing of plasmablast antibody repertoires in rheumatoid arthritis. *Arthritis Rheumatol*, 66(10):2706–2715, Oct 2014.

[112] S. Tanner, Z. Shen, J. Ng, L. Florea, R. Guigo, S. P. Briggs, and V. Bafna. Improving gene annotation using peptide mass spectrometry. *Genome Res.*, 17(2):231–239, Feb 2007.

[113] Y. L. Tao, Y. Li, J. Gao, Z. G. Liu, Z. W. Tu, G. Li, B. Q. Xu, D. L. Niu, C. B. Jiang, W. Yi, Z. Q. Li, J. Li, Y. M. Wang, Z. B. Cheng, Q. D. Liu, L. Bai, C. Zhang, J. Y. Zhang, M. S. Zeng, and Y. F. Xia. Identifying FGA peptides as nasopharyngeal carcinoma-associated biomarkers by magnetic beads. *J. Cell. Biochem.*, 113(7):2268–2278, Jul 2012.

[114] N. Teranishi, Z. Naito, T. Ishiwata, N. Tanaka, K. Furukawa, T. Seya, S. Shinji, and T. Tajiri. Identification of neovasculature using nestin in colorectal cancer. *Int. J. Oncol.*, 30(3):593–603, Mar 2007.

[115] The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306(5696):636–640, Oct 2004.

[116] J. C. Tran and A. A. Doucette. Multiplexed size separation of intact proteins in solution phase for mass spectrometry. *Anal. Chem.*, 81(15):6201–6209, Aug 2009.

[117] C. Trapnell, L. Pachter, and S. L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, May 2009.

[118] 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 differentiation. *Nat. Biotechnol.*, 28(5):511–515, May 2010.

[119] P. J. Utz and P. Anderson. Posttranslational protein modifications, apoptosis, and the bypass of tolerance to autoantigens. *Arthritis Rheum.*, 41(7):1152–1160, Jul 1998.

[120] F. Vanky, E. Klein, J. Willems, K. Book, T. Ivert, A. Peterffy, U. Nilsonne, A. Kreicbergs, and T. Aparisi. Lysis of autologous tumor cells by blood lymphocytes tested at the time of surgery. Correlation with the postsurgical clinical course. *Cancer Immunol. Immunother.*, 21(1):69–76, 1986.

[121] E. Venter, R. D. Smith, and S. H. Payne. Proteogenomic analysis of bacteria and archaea: a 46 organism case study. *PLoS ONE*, 6(11):e27587, 2011.

[122] R. G. Verhaak, K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, C. R. 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. *Cancer Cell*, 17(1):98–110, Jan 2010.

[123] M. Via, C. Gignoux, and E. G. Burchard. The 1000 Genomes Project: new opportunities for research and social challenges. *Genome Med*, 2(1):3, 2010.

[124] K. Vincent, D. C. Roy, and C. Perreault. Next-generation leukemia immunotherapy. *Blood*, 118(11):2951–2959, Sep 2011.

[125] B. Vincenzi, D. Santini, G. Perrone, F. Graziano, F. Loupakis, G. Schiavon, A. M. 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 cancer patients. *J. Cell. Physiol.*, 227(3):927–933, Mar 2012.

[126] X. Wang, R. J. Slebos, D. Wang, P. J. Halvey, and et al. Protein identification using customized protein sequence databases derived from RNA-Seq data. *J. Proteome Res.*, 11(2):1009–1017, Feb 2012.

[127] 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 derived from RNA-Seq data. *J. Proteome Res.*, 11(2):1009–1017, Feb 2012.

[128] X. Wang and B. Zhang. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics*, 29(24):3235–3237, Dec 2013.

[129] Z. Q. Wang, K. Milne, J. R. Webb, and P. H. Watson. CD74 and intratumoral immune response in breast cancer. *Oncotarget*, 8(8):12664–12674, Feb 2017.

[130] S. Woo, S. W. Cha, S. Bonissone, S. Na, D. L. Tabb, P. A. Pevzner, and V. Bafna. Advanced Proteogenomic Analysis Reveals Multiple Peptide Mutations and Complex Immunoglobulin Peptides in Colon Cancer. *J. Proteome Res.*, 14(9):3555–3567, Sep 2015.

[131] S. Woo, S. W. Cha, and et al. Proteogenomic strategies for identification of aberrant cancer peptides using large-scale Next Generation Sequencing data. *PROTEOMICS*, in press.

[132] S. Woo, S. W. Cha, G. Merrihew, Y. He, N. Castellana, C. Guest, M. MacCoss, and V. Bafna. Proteogenomic database construction driven from large scale RNA-seq data. *J. Proteome Res.*, 13(1):21–28, Jan 2014.

[133] S. Woo, S. W. Cha, G. Merrihew, Y. He, and et al. Proteogenomic database construction driven from large scale RNA-seq data. *J. Proteome Res.*, 13(1):21–28, Jan 2014.

[134] S. Woo, S. W. Cha, S. Na, C. Guest, T. Liu, R. D. Smith, K. D. Rodland, S. Payne, and V. Bafna. Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data. *Proteomics*, 14(23-24):2719–2730, Dec 2014.

[135] J. Wrammert, K. Smith, J. Miller, W. A. Langley, K. Kokko, C. Larsen, N. Y. Zheng, I. Mays, L. Garman, C. Helms, J. James, G. M. Air, J. D. Capra, R. Ahmed, and P. C. Wilson. Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. *Nature*, 453(7195):667–671, May 2008.

[136] Z. Xu, H. Chen, D. Liu, and J. Huo. Fibulin-1 is downregulated through promoter hypermethylation in colorectal cancer: a CONSORT study. *Medicine (Baltimore)*, 94(13):e663, Apr 2015.

[137] Daniel R Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome research*, 18(5):821–829, 2008.

[138] B. Zhang, J. Wang, X. Wang, J. Zhu, Q. Liu, Z. Shi, M. C. Chambers, L. J. Zimmerman, K. F. Shaddox, S. Kim, S. R. Davies, S. Wang, P. Wang, C. R. Kinsinger, R. C. Rivers, H. Rodriguez, R. R. Townsend, M. J. Ellis, S. A. Carr, D. L. Tabb, R. J. Coffey, R. J. Slebos, D. C. Liebler, S. A. Carr, M. A. Gillette, K. R. Klauser, E. Kuhn, D. R. Mani,

P. Mertins, K. A. Ketchum, A. G. Paulovich, J. R. Whiteaker, N. J. Edwards, P. B. McGarvey, S. Madhavan, P. Wang, D. Chan, A. Pandey, I. e. M. Shih, H. Zhang, Z. Zhang, H. Zhu, G. A. Whiteley, S. J. Skates, F. M. White, D. A. Levine, E. S. Boja, C. R. Kinsinger, T. Hiltke, M. Mesri, R. C. Rivers, H. Rodriguez, K. M. Shaw, S. E. Stein, D. Fenyo, T. Liu, J. E. McDermott, S. H. Payne, K. D. Rodland, R. D. Smith, P. Rudnick, M. Snyder, Y. Zhao, X. Chen, D. F. Ransohoff, A. N. Hoofnagle, D. C. Liebler, M. E. Sanders, Z. Shi, R. J. Slebos, D. L. Tabb, B. Zhang, L. J. Zimmerman, Y. Wang, S. R. Davies, L. Ding, M. J. Ellis, and R. R. Townsend. Proteogenomic characterization of human colon and rectal cancer. *Nature*, 513(7518):382–387, Sep 2014.

[139] B. Zhang, J. Wang, X. Wang, J. Zhu, Q. Liu, Z. Shi, M. C. Chambers, L. J. Zimmerman, K. F. Shaddox, S. Kim, S. R. Davies, S. Wang, P. Wang, C. R. Kinsinger, R. C. Rivers, H. Rodriguez, R. R. Townsend, M. J. Ellis, S. A. Carr, D. L. Tabb, R. J. Coffey, R. J. Slebos, D. C. Liebler, S. A. Carr, M. A. Gillette, K. R. Klauser, E. Kuhn, D. R. Mani, P. Mertins, K. A. Ketchum, A. G. Paulovich, J. R. Whiteaker, N. J. Edwards, P. B. McGarvey, S. Madhavan, P. Wang, D. Chan, A. Pandey, I. e. M. Shih, H. Zhang, Z. Zhang, H. Zhu, G. A. Whiteley, S. J. Skates, F. M. White, D. A. Levine, E. S. Boja, C. R. Kinsinger, T. Hiltke, M. Mesri, R. C. Rivers, H. Rodriguez, K. M. Shaw, S. E. Stein, D. Fenyo, T. Liu, J. E. McDermott, S. H. Payne, K. D. Rodland, R. D. Smith, P. Rudnick, M. Snyder, Y. Zhao, X. Chen, D. F. Ransohoff, A. N. Hoofnagle, D. C. Liebler, M. E. Sanders, Z. Shi, R. J. Slebos, D. L. Tabb, B. Zhang, L. J. Zimmerman, Y. Wang, S. R. Davies, L. Ding, M. J. Ellis, and R. R. Townsend. Proteogenomic characterization of human colon and rectal cancer. *Nature*, 513(7518):382–387, Sep 2014.

[140] L. Zhang, J. R. Conejo-Garcia, D. Katsaros, P. A. Gimotty, M. Massobrio, G. Regnani, A. Makrigiannakis, H. Gray, K. Schlienger, M. N. Liebman, S. C. Rubin, and G. Coukos. Intratumoral T cells, recurrence, and survival in epithelial ovarian cancer. *N. Engl. J. Med.*, 348(3):203–213, Jan 2003.