UNIVERSITY OF CALIFORNIA, SAN DIEGO

Deconvolution of immunome and degradome repertoires using computational proteomics

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

in

Bioinformatics and Systems Biology

by

Stefano R. Bonissone

Committee in charge:

Professor Pavel A. Pevzner, Chair
Professor Steven Briggs, Co-Chair
Professor Vineet Bafna
Professor Nuno Bandeira
Professor Sergei Kosakovski Pond

2015

The dissertation of Stefano R. Bonissone is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California, San Diego

2015

DEDICATION

To my family, whose love, support, and encouragement helped me through the journey that is graduate school.

To Jocelyne, for making the journey that much easier and enjoyable.

EPIGRAPH

Reality is frequently inaccurate.

—Douglas Adams,

The Restaurant at the End of the Universe

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I have had the good fortune of working with many talented individuals over the course of my graduate studies. During the course of my first project in bioinformatics I worked with Nitin Gupta, who acted as an important student mentor. Additionally, Ralph A. Bradshaw provided great insight into the biology of N-termini.

In subsequent projects I worked with several members of Dr. Pevzner and Dr. Bafna's labs, in particular Sunghee Woo, Seong W. Cha, and Yana Safonova.

I have also been lucky to share lab space with many great colleagues, providing insight, support, and friendship. Including: Anand Patel, Eric Scott, June Snedecor, Kyowon Jeong, Marcus Kinsella, Nitin Gupta, Nitin Udpa, Roy Ronen, Seong W. Cha, Son K. Pham, Sunghee Woo, Viraj Deshpande.

I have also been fortunate to have a mentor such as Pavel, who introduced me to bioinformatics and showed me how wonderful it can be.

Chapter 2 is adapted from **S. Bonissone**, N. Gupta, M. Romine, R.A. Bradshaw, and P.A. Pevzner. N-terminal protein processing: A comparative proteogenomic analysis. Molecular & Cellular Proteomics, 12(1):1428, 2013. The dissertation author was the primary author of this paper.

Chapter 3 is adapted from **S.R. Bonissone** and P.A. Pevzner. Immunoglobulin classification using the colored antibody graph. In Research in Computational Molecular Biology, volume 9029 of Lecture Notes in Computer Science, pages 44-59. Springer International Publishing, 2015. The dissertation author was the primary author of this paper.

Chapter 4 is adapted from 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):i53i61, 2015. The dissertation author was one of the primary authors of this paper.

Chapter 5 is adapted from 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. Journal of Proteome Research, 2015. The dissertation author was one of the primary authors of this paper.

VITA

2006	B. S. in Computer Science, Rensselaer Polytechnic Institute, Troy, NY May
2010	M. S. in Computer Science, University of California, San Diego
2015	Ph. D. in Bioinformatics and Systems Biology, University of California, San Diego

PUBLICATIONS

S. Bonissone, N. Gupta, M. Romine, R.A. Bradshaw, and P.A. Pevzner. N-terminal protein processing: A comparative proteogenomic analysis. Molecular & Cellular Proteomics, 12(1):14-28, 2013

S.H. Payne, **S. Bonissone**, S. Wu, R.N. Brown, D.N. Ivankov, D. Frishman, L. Pasa-Toli c, R.D. Smith, and P.A. Pevzner. Unexpected diversity of signal peptides in prokaryotes. mBio, 3(6), November/December 2012

D. N. Ivankov, S. H. Payne, M. Y. Galperin, **S. Bonissone**, P. A. Pevzner, and D. Frishman. How many signal peptides are there in bacteria? Environmental microbiology, 15(4):983-990, 2013

S.R. Bonissone and P.A. Pevzner. Immunoglobulin classification using the colored antibody graph. In Research in Computational Molecular Biology, volume 9029 of Lecture Notes in Computer Science, pages 44-59. Springer International Publishing, 2015

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):i53-i61, 2015

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. Journal of Proteome Research, 2015

ABSTRACT OF THE DISSERTATION

Deconvolution of immunome and degradome repertoires using computational proteomics

by

Stefano R. Bonissone

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2015

Professor Pavel A. Pevzner, Chair
Professor Steven Briggs, Co-Chair

In recent years, we have seen the advent of high-throughput sequencing technologies, providing us with the unprecedented ability to query DNA, RNA, and proteins. We have witnessed the vast variability of molecules at the level of sequences. Disentangling this molecular variability, and describing the array of changes to molecules, is critical to understanding certain biological processes. This dissertation asks questions about certain functions, e.g., immune response and N-terminal events, and attempts to ascertain any patterns in the data.

While the N-terminus may seem like an innocuous portion of translated proteins, it in fact, harbors great variability. Despite all proteins starting with a methionine (Met), many lose this first residue, to N-terminal methionine excision (NME), exposing the second residue.

Additionally, some proteins gain an acetyl group to the N-terminus. These N-terminal post-translational modifications (PTMs) make up the N-terminome.

Another biological process that introduces variability is that of the immune repertoire. B-cells, responding to an antigen, produce heavy and light chain proteins to form an antibody. These antibodies are comprised of three gene-segments, termed V, D, and J. Unlike typical exonic splicing events, VDJ somatic recombination occurs at the genomic level within B-cells. Additionally, somatic hypermutation (SHM) introduces a very large number of mutations. Thus, the population of B-cells will produce a large variety of antibody heavy and light chains, in an attempt to target an antigen. This immunoglobulin repertoire forms part of the immunome that we attempt to characterize in different projects.

Chapter 1

Introduction

Bioinformatics is driven by the data we are able to generate. High-throughput technologies have enabled the creation of numerous experiments probing DNA, RNA, and proteins. New assays modifying these underlying technologies enable exploring new hypothesis, and frequently, require new computational tools. This environment provides a wealth of existing datasets, and an ever changing arsenal of new experimental approaches with which to ask questions.

With these new high-throughput technologies, we are now able to characterized and quantify the vast variability at the molecular level. These molecular repertoires can be captured and analyzed using different approaches. In this dissertation, repertoires from the N-terminome and immunome are examined.

1.1 N-terminome

The N-terminus of proteins harbors a great deal of variability, despite all translated proteins begin with a methionine. N-terminal methionine excision (NME) is the process of removing this initiating methionine, exposing the second residue to the N-terminus. Additional post-translational modifications (PTMs) occur on the N-terminus, in particular N-terminal acetylation, which will be discussed in the dissertation.

Chapter 2 explores N-terminal methionine excision by employing comparative protegenomics approaches to analyze millions of tandem mass-spectra (MS/MS), across 57 different organisms. The comparative genomics postulate, that function suggests conservation, is applied to proteomic searches.

In this project, whole cell lysates from many different bacterial and eukarotyes were analyzed using tandem mass-spectrometry. These experiments did not enrich for the N-terminus, but instead relied on the whole cell capture of proteins. The lack of enrichment obviously only allows us to examine those proteins which are captured by the mass-spectrometer, favoring those medium and high-abundance proteins. While this bias may be problematic for some hypothesis, for exploring NME specificity it did not pose a challenge.

1.2 Immunome

The immunoglobulin gene is of particular interest since it can target a multitude of pathogens, and provides a critical line in our body's immune system. As a result, a recombinant version can be engineered to a specific target, providing potential therapeutic uses. However, identifying a candidate antibody from which to start is a difficult task.

Immunoglobulin sequencing (Ig-seq) is an assay targeting immunoglobulin transcripts, and allows for the querying of an antibody repertoire; the overall process is depicted in Figure 1.1. The heavy or light chain transcripts from B-cells are targeted, amplified, and sequenced, to provide a measure of the resulting antibody repertoire.

The resulting Ig-seq reads are remarkably diverse due to the extreme variability of immunoglobulins themselves. Both tools described in Chapters 3 and 4 address this variability. Chapter 3 provides a means of identifying the original, germ-line, composition of each immunoglobulin read. While Chapter 4 provides a way of organizing Ig-seq data, and better searching for peptides in MS/MS data of immunoglobulins.

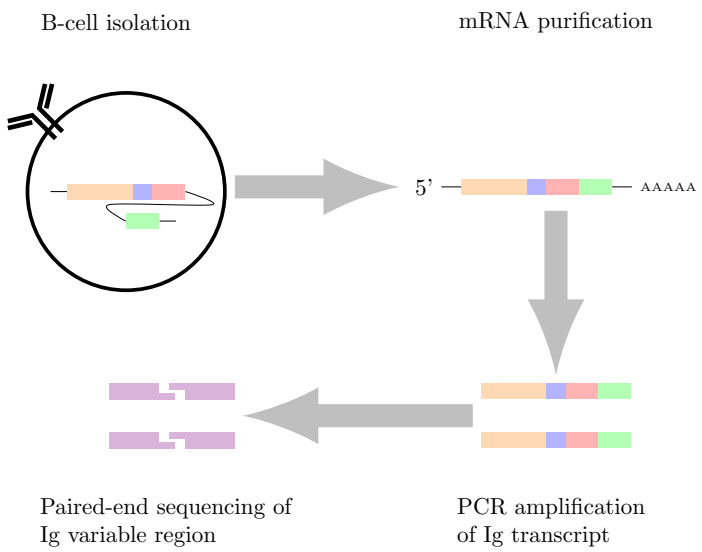


Figure 1.1: Immunoglobulin sequencing diagram. B-cells are isolated, transcripts purified, amplified, and ultimately sequenced. The heavy or light chains can be targeted using this approach.

1.3 Outline

Chapter 2 focuses on the N-terminome and uses a comparative proteogenomic approach to ask the the question about the function of N-terminal methionine excision (NME), and the role it plays in protein degradation.

Chapter 3 describes the problem of identifying germline V, D, J gene-segments from antibody transcript sequencing. This operation is a critical first step to analyzing the antibody response as queried by immunoglobulin sequencing (Ig-seq).

Chapter 4 continues to address the problem of antibody repertoire analysis, by introducing a tool to perform clustering and error correction of Ig-seq reads. Further, database search of mass-spectra from antigen pulldowns is performed to reveal the ability to identify antibody repertoires at the proteomic level.

Chapter 5 examines tumor specific features: tumor mutations and tumor infiltrating lymphocytes (TILs) by measuring immunoglobulin peptides from infiltrating B-cells. Here mutations and immunoglobulin peptides are compared to find differences across colorectal tumor subtypes. Additionally, TILs are also considered in the context of transcripts. These

antibody transcripts present at the tumor site are assembled, and isotypes are analyzed for similarities across tumor samples and differences between normal samples when available. Colon and breast cancers are investigated.

Chapter 2

N-terminal post-translational modifications

2.1 Comparative proteogenomics

2.1.1 Introduction

Although methionine is used to initiate protein synthesis for essentially all proteins, it is subsequently removed in a large percentage of cases, either by cleavage of an N-terminal signal peptide (as part of cellular translocation mechanisms or precursor activations) or by the action of specific methionine aminopeptidases (MetAPs). Approximately two-thirds of the proteins in any proteome are potential substrates for the latter N-terminal methionine excision (NME), and MetAPs appear in all organisms from bacteria to eukaryotes [GBM04]. The second, or P2, amino acid in protein substrates is crucially important for NME since MetAP specificity mainly depends on the nature of this residue, a selectivity that is conserved across all species [GBM04, AB88, FMP$^+$06, MSG06, BBW98]. These enzymes generally excise the N-terminal Met when the second residue is Gly, Ala, Ser, Thr, Cys, Pro, or Val [FMP$^+$06, HEL$^+$87, MG08], which are the amino acids smallest in size (based on radius of gyration of the side chain [Lev76]). NME is a necessary process for proper cell functioning; it is included in the minimal genome

set of eubacteria [HPG+99]. Eukaryotes contain two MetAPs derived from a version in bacteria (MetAP1), and another found in archea (MetAP2) [AKH+95]. Just as the deletion of MetAP eubacteria is lethal, the deletion of both MetAPs in yeast is also lethal [LC95].

In 1988, Arfin and Bradshaw [AB88] observed that the specificity of NME coincided with that of the N-end rule (NER) [BFV86, Var96], a ubiquitin-dependent protein degradation process that is based on the recognition of N-terminal residues. The stabilizing residues for the NER include Gly, Ala, Ser, Cys, Thr, Pro, and Val and, with the exception of Met, the destabilizing residues are all found to be in the class of P2-residues that are not substrates for the MetAPs. This suggested that NME acts to release Met from proteins whose stability is unaffected by the NER creating at the same time a second class of proteins, who have the potential for regulated turnover downstream of the co-translational processing, when, and if, the N-terminal Met is subsequently removed by a mechanism other than the co-translational action of the MetAPs. However, despite extensive studies, this type of programmed protein turnover (requiring downstream removal of Met) has not been demonstrated to occur. An implication of this correlation is that exposing of the stabilizing residues may also contribute to increasing their lifetime.

The stabilizing residues exposed by the action of the MetAPs can be further modified. The most extensive of these reactions is N-terminal acetylation (NTA), which can occur on as much as 70-80% of the mass of the soluble protein in eukaryotes. Although the specificity of the N-acetyltransferase (NAT) responsible is not as rigid as the MetAPs, the principal substrates in the stabilizing class are usually the four smallest residues (Gly, Ala, Ser, and Thr) [HEL+87, PFHJ85]. A second class of NATs can also modify the retained Met when the adjacent residues are Asp, Glu or Asn [PNT+99]. The functional importance of this modification (in either case) is not known although it has been suggested that it may exert a protective effect against spurious aminopeptidase cleavages. Recently, Hwang et al., [HSV10] have extended the NER to include Nα-acetylated termini as also destabilizing thus providing another possible function for this modification. In contrast, to date, very few instances of

Nα-acetylation have been observed in bacteria. Other modifications can also occur in both eukaryotes and prokaryotes although they are generally much more limited in scope.

The specificity of the MetAPs suggest an apparent connection between NME and protein degradation. However, this connection has never been examined using high-throughput mass spectrometric data or a comparative genomics approach; thus it remains unclear whether exposing these stabilizing residues contributes to increasing protein half-life and thus represents a primary purpose of NME . (The connection between NME and NER in bacteria, which has an NER with a somewhat different profile [TSRV91], is even more obscure.) Recent studies provide some examples where disruption of NME via a single-residue substitution in the P2 position causes protein degradation [PS02, WB99, GVM03]; however, some of these experimental results are in conflict with the NER [Var96]. Giglione et al., [GVM03] have shown that NME triggers degradation of D2 protein in *Caenorhabditis reinhardtii* in the PSII complex after replacing the second (stabilizing) Thr residue by another amino acid to prevent NME. This replacement results in early degradation of D2 and instability of the PSII complex. From this, Giglione et al. [GVM03] postulated that NME determines protein life-span via currently unknown machinery. However, since Bachmair et al. [BFV86] classified Met as a stabilizing residue, it is not entirely clear why substituting one stabilizing residue (Met) by another one (Gly, Ala, Ser, Cys, Thr, Pro, or Val) should affect protein stability and the substitution may have other deleterious effects that are manifested in different ways.

The logic for analyzing NME and NER is shown in Figure 2.1 NME exposes 7 different residues as new N-termini of proteins. The natural conclusion that has become a dogma of NME is that these 7 residues are exposed for a functional reason. The broad scope of NME suggests a universal reason that surpasses any particular protein's role. In turn the comparative genomics postulate (function suggests conservation) leads to the conclusion that the 7 residues should be evolutionarily conserved at position P2 of proteins. However, since only 2 out of the 7 residues are conserved, we argue that one of the two assumptions in Figure 2.1(a) must be incorrect and put forth the alternative logic depicted in Figure 2.1(b), which matches our

analysis across dozens of species. According to this logic, NME accomplishes the goal of exposing Ala and Ser by exposing all residues with side chains smaller or comparable in size to Ala and Ser (G, T, V, P, C). These residues are thus inconsequential players that are not functionally important (and are not evolutionarily conserved) at P2.

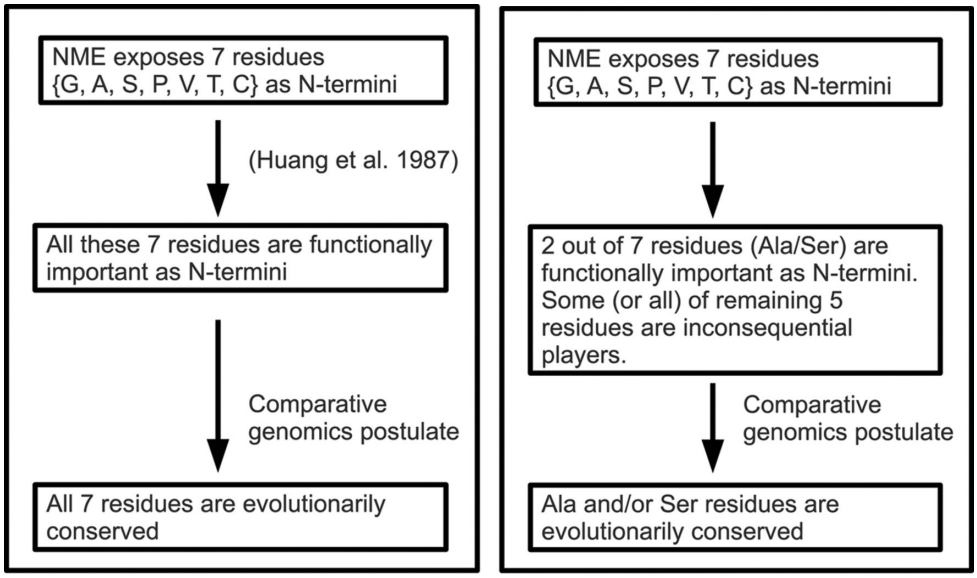


Figure 2.1: Two alternative cases for NME function. (a) NME exposes 7 residues to be new N-termini of proteins. Since this is presumably for some functional reason, the conventional assumption is that all 7 residues must have functional importance as N-termini. By the comparative genomics postulate (as defined in the text), evolutionary conservation of all 7 at P2 should be observed. If all of these residues are not conserved, one of the two assumptions must be incorrect; either not all 7 residues are important or the comparative genomics postulate is invalid. (b) Given that the comparative genomics postulate holds, and only 2 of the 7 residues are of functional importance as N-termini, then the other 5 residues are inconsequential players and only these two residues should be evolutionarily conserved.

In this report, we examine the connection between the specificity of NME and stabilizing residues of NER. In doing so, datasets from bacteria (including 112 million mass spectrometric spectra from 57 species), yeast, and mammals, were analyzed for N-terminal peptides both with respect to the excision (or not) of initiator Met residues and the distribution of P2-residues. The results reveal a strong preference of Ala and Ser as P2-residues. However, this process does not appear to be linked to the NER other than being generally compatible with it. These

studies also demonstrate a much greater than expected number of Nα-acetylation events in some bacteria.

2.1.2 Methods

Spectral Datasets

A tandem mass spectrometry (MS/MS) dataset from *Saccharomyces cerevisiae* was utilized in the analysis of NME and other post-translational modifications (PTMs). Six million spectra were analyzed, searched against an *S.cerevisiae* database, yielding 410 protein identifications with N-terminal evidence.

In addition, 112 million spectra from 57 different bacterial organisms were searched to provide a large dataset containing 16511 classifiable proteins. InsPecT [TSF+05] was used for searching the spectra followed by the MS-generating function approach (MSGF), as detailed in Kim et al. [KGP08] to filter for significant identifications at a false discovery rate (FDR) of 1%. Further description of this dataset is presented in Text S1.

N-terminal PTM search

Of the available 57 datasets from bacterial organisms, 47 were searched for all N-terminal modifications (including unexpected ones) across 68 million spectra. In order to identify PTMs occurring on the N-terminus of proteins, MSGF [KGP08] was utilized. Given a protein p in proteome P, the first tryptic cut of p was taken as the peptide to search. Therefore, the N-terminus of the peptide was the N-terminus of p, while the C-terminus was either K or R. In addition to searching for this peptide, the initial Met was removed and the Met-less version was also searched. So should $|P|$ denote the number of proteins in the proteome P, there are $M = 2 * |P|$ peptides to search.

Each of the M peptides was paired with each available spectrum for the organism in question. Therefore, with N different spectra, one obtained $M * N$ potential peptide-spectra matches (PSM) to test. For each PSM, there will likely be a difference in parent mass of

the spectrum and of the peptide. This difference, or offset, was concatenated to the peptide annotation as an N-terminal modification to be searched. As a result our search represents a blind (or universal) search for PTM [TTZ$^+$05] that allows us to find unexpected PTMs.

For most bacteria, M is between 1,000 and 8,000, while N varies greatly. However, there are frequently 40 million to 400 million PSMs with accompanying spectral probabilities. Only those PSMs with spectral probabilities less than 1.0×10^{-10} were retained as statistically significant. This greatly reduced the number of PSMs, and potential PTMs, to consider.

A permutation test was used to determine over-representation of Nα-acetylated Ser-proteins from the 248 proteins found by MS/MS data in yeast containing evidence of N-terminal acetylation. This test was performed by considering all proteins in *S.cerevisiae* with Ala, Ser, or Thr at position 2. Of all proteins satisfying that constraint, 248 proteins were randomly selected. Counts for the number of proteins with Ala, Ser, and Thr at P2 were determined for each iteration. This process of randomized sampling was iterated 10,000 times to obtain a null distribution. Thresholds for 5% and 1% were then determined using the conservative Bonferroni correction for the tests of the three residues.

Sequelogs

In order to test if NME was critical for a function of a specific protein, its sequelogs from related species must be analyzed (in this report the term sequelog from [Var04] is used instead of the commonly used term ortholog). The protein sequence data used for comparing sequelogs were compiled into three sets: bacteria, yeast, and mammal. The bacterial dataset comprises 19 Shewanella species (1860 proteins had sequelogs in all species). The yeast dataset comprised seven Saccharomyces species (1502 proteins had sequelogs in all species) obtained from the Saccharomyces Genome Database (http://downloads.yeastgenome.org/). The mammalian dataset was composed of six species: human, chimpanzee, macaque, cow, opossum, and rat, obtained from the MSOAR project (http://msoar.cs.ucr.edu/). There were 7934 proteins that had sequelogs in all species. Sequelog mappings for each dataset were obtained from each

respective project. A protein was included in a particular dataset if there existed a sequelog in all species considered in the dataset in question. No further annotations or processing were performed to determine or modify the provided sequelog mapping.

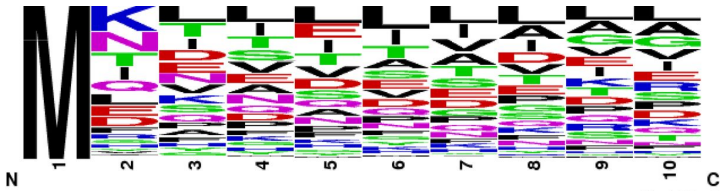
2.1.3 Results

NME Specificity

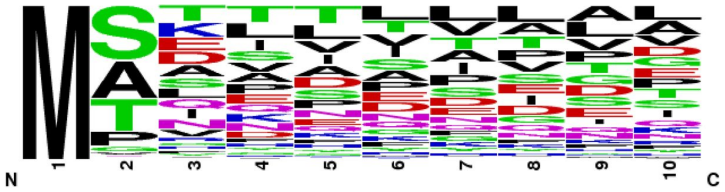
Large-scale mass spectrometric studies [GTJ$^+$07] have confirmed that NME occurs with nearly 100% efficiency from proteins containing Gly, Ala, Ser, Thr, Cys, Pro, and Val as the P2-residues, which provides the simplest rule for predicting NME. More detailed studies [FMP$^+$06, MG08] have devised rules for NME prediction that utilize the amino acids in positions 2 and 3. The simple rule (that classifies a protein as NME-positive if P2 is in X where $X = \{G,A,P,S,T,V,C\}$) results in an 8% error rate, most of which represents false positives (96 of 107 errors are false positives). The accuracy of this simple prediction rule on the *Shewanella* MS/MS dataset is acceptable, given that some of the errors may reflect false positive peptide identifications rather than a shortcoming of the rule.

As part of these analyses, an MS/MS dataset consisting of $\approx$112 million spectra from 57 different bacteria generated at Pacific Northwest National Laboratory were examined (see [FMS$^+$10] for a detailed description of this dataset). 6811 N-terminal peptides representing Met retention and 9700 representing Met removal (16511 peptides in total) were identified. This is the largest dataset of N-terminal peptides analyzed to date. Figure 2.2(a) and 2.2(b) represent the logos for peptides with retained and cleaved Met, respectively. Table S1, located in Text S1, summarizes the organisms searched and the results obtained.

Interestingly, these analyses also revealed previously unknown variability in NME specificity across different species. An example of a species with a less specific NME (with 8 rather than 7 residues in the second position that trigger NME) is shown in Figure 2.3, which compares NME specificities in *Brachybacterium faecium*, *Escherichia coli*, and *Deinococcus radiodurans*. Asn is exposed in 20 of 22 cases, in contrast to the standard NME specificity that



(a)



(b)

Figure 2.2: Sequence motifs for the first ten residues of each protein determined by MS/MS. (a) 6811 samples which retained Met and (b) 9700 samples where Met was cleaved.

conventionally includes this residue in the group of non-MetAP substrates. However, it is consistent with the observations of Walker and Bradshaw [WB99] that yeast MetAP1 could readily cleave a synthetic peptide with a Met-Asn N-terminus. *Deinococcus radiodurans* and various *Geobacter* species show a different variance where, contrary to the usual MetAP substrate profile, NME retains rather than removes Met before Thr. This is unlikely to be an artifact since the same effect is observed in all Geobacter species in this dataset, which includes *Geobacter metallireducens*, *Geobacter sulfurreducens*, and *Geobacter uraniireducens*. Another smaller but significant variation in NME specificity relates to Val P2-residues where Met is mainly retained rather than excised in ≈30% of the species considered (cleavage specificity is less than 50%). Therefore, while the simple MetAP specificity is a good overall predictor for most species, there is some flexibility observed in some species. This poses an opportunity for NME predictor software, such as the TERMINATOR web server [MTV$^+$08], to improve and reflect these variations in NME specificity (currently all eubacteria are assumed to have the same NME specificity).

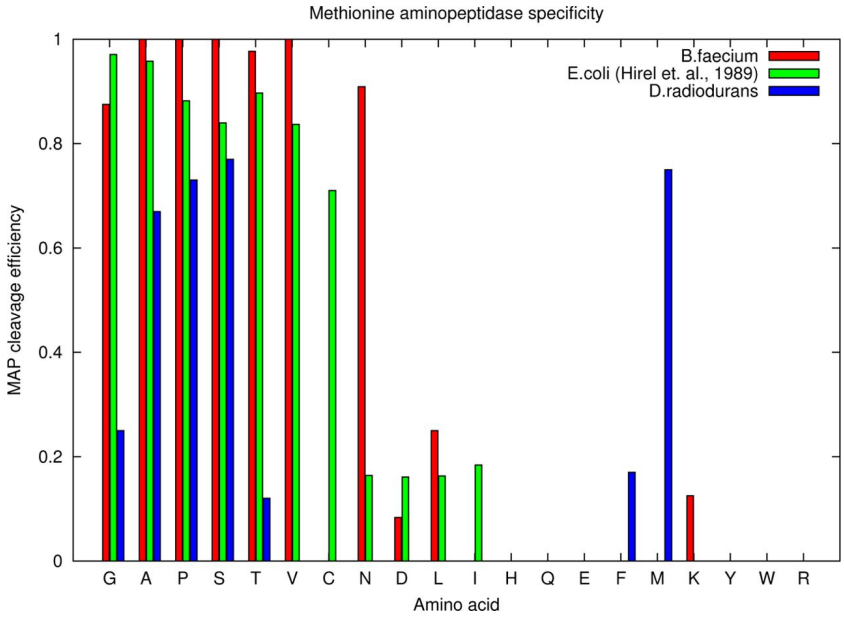


Figure 2.3: Comparison of MetAP specificities in *Brachybacterium faecium*, *E.coli*, and *Deinococcus radiodurans*. *B. faecium*, *E.coli*, and *D. radiodurans* significantly differ in NME specificity (Met occurring before Asn was released in 90% of cases in *B. faecium*). The specificity of *D. radiodurans* differs from expected with very few Thr residues exposed (only 10 of 82 samples).

N-terminal modifications

A typical MS/MS experiment identifies only a fraction of all tryptic peptides (and hence only a fraction of those derived from the N-terminus) and the accuracy of MS/MS-based peptide identification tools further deteriorates while detecting modified (e.g., acetylated) peptides (see [HGR$^+$10] for complications in identification of N-acetylated peptides). Therefore, the acetylation status of most proteins cannot be inferred from a typical MS/MS experiment.

Mass spectrometry data for *S. cerevisiae* were searched for PTMs (Figure 2.4). Nα-acetylation is a common PTM in yeast proteins, with Ser, Ala, Gly, and Thr being the acetylated residues when the initiating Met is cleaved [Wal06]. Indeed, of the 270 Nα-acetylated sites found in *S. cerevisiae* by mass spectrometry, 197 are on Ser, 31 on Ala, and 20 on Thr. Permutation tests revealed that only Ser is significantly over-represented in these PTM observations, the only residue with a p-value less than 0.01 (see Methods). In addition, Figure 2.4(d) shows that Nα-acetylation prefers Ser-proteins since Thr- and Ala-proteins increase in frequency

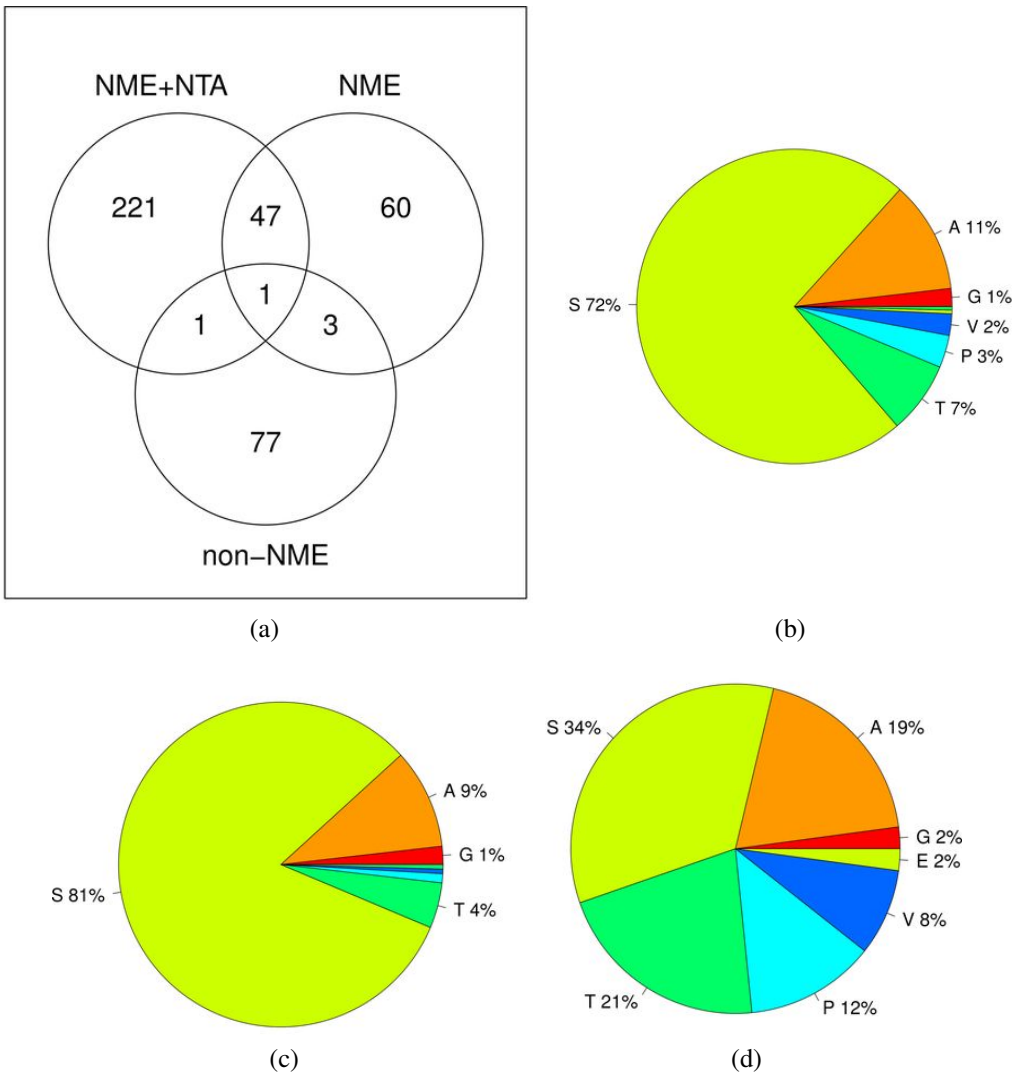


Figure 2.4: Results from PTM searches in *Saccharomyces cerevisiae*. (a) Breakdown of NME, non-NME, and NME+NTA (NME and Nα-acetylation) events. Approximately 65% of N-terminal events in yeast were found to be N-acetylations. (b) Residue breakdown of the P2 position of proteins that underwent NME+NTA. Ser is strongly favored in yeast, much as Ser-conservation is also pronounced in yeast. (c) 221 identifications of only NME+NTA proteins show that Ser dominates the composition. (d) The intersection between NME+NTA and NME-only show a wide distribution among NME residues.

when considering the intersection of Nα-acetylation and NME-only events. While proteins with N-terminal Ser, Ala, Thr, and Gly are all Nα-acetylated in yeast, N-temrinal Ser-proteins appear to be heavily favored.

Serine is favored in *S.cerevisiae*, however, it is not universally favored for Nα-acetylation events. Helbig et al. [HGR$^+$10] provide mass spectrometry evidence for Nα-acetylation in human cells, showing a strong preference for Ala. In the same study, the Nα-acetylation profile for *Drosophila melanogaster* was also shown to differ, with Met, Ala, and Ser nearly equally represented. The bacterial Nα-acetylation data in Table 2.1 and Figure 2.5 also show a slightly different profile, still with Ala and Ser as prominent residues.

Widespread N-terminal acetylation in bacteria

Acetylation is currently viewed as a rare modification in bacteria. In previous studies with in *E. coli* five different Nα-acetylated proteins were identified: ribosomal subunits S18 and S5 [YISI87], L12 [TMYI89], along with SecB [SSRS96], and elongation factor Tu [ACD$^+$80]. Utilizing the dataset of bacterial organisms, blind PTM searches detected many Nα-acetylated proteins in multiple organisms, although certainly not all.

Figure 2.6 shows the potential PTMs as the number of occurrences of a particular mass shift. Mass shifts of -131 correspond to a loss of Met (i.e. an NME event) while a shift of 0 signifies the retention of the initial Met residue; these are expected events. The surprising number of -89 mass shifts, corresponding to a loss of Met (-131) along with the addition of an acetyl group (+42), was considered indicative of Nα-acetylation. A protein was considered to be Nα-acetylated if it was observed with this offset within 1Da.

Figure 2.5(a) diagrams the findings of NME-only, non-NME, and NME proteins that also undergo Nα-acetylation (NME+NTA), using this un-targeted PTM search method. There is considerable overlap between NME-only and NME+NTA sets, suggesting that many proteins undergoing NTA do not require this modification, or require both (as with ribosomal protein L7/L12 in E.coli). Figure 2.5 summarizes the overall P2 residue distribution of NME events

Table 2.1: Nα-acetylation P2 residue breakdown using a 1.0×10^{-10} spectral probability threshold and described filtering procedures, using the definition as: -89 ± 1Da neighborhood.

orgnum	Organism	G	A	S	C	T	P	V	D	N	L	I	Q	E	H	M	F	K	Y	W	R	Σ
005	Anaeromyxobacter dehalogenans	0	1	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	2
007	Actinosynnema mirum DSM	3	8	33	-	60	3	2	2	-	-	-	-	-	-	-	-	-	-	-	-	111
013	Anabaena variabilis ATCC 29413	1	23	39	-	71	5	26	-	4	3	3	-	1	-	-	-	1	1	-	1	179
014	Arthrobacter FB24	1	14	95	-	124	3	4	-	-	1	-	1	-	-	1	-	-	-	-	-	244
017	Bacillus anthracis	2	12	3	-	1	1	4	-	1	1	-	-	-	-	-	-	-	-	-	-	25
019	Brachybacterium faecium DSM4810	1	9	47	-	90	4	-	-	1	1	-	-	-	-	-	-	-	-	-	-	153
021	Chloroflexus aurantiacus	1	18	40	-	51	-	7	-	6	-	2	-	-	-	1	-	-	-	-	-	126
028	Cyanobacterium synechocystis	2	9	10	-	11	1	8	-	-	-	1	-	-	-	-	-	-	-	-	-	42
033	Cyanothece PCC7424	0	4	4	-	13	-	14	-	-	-	1	-	-	-	-	-	-	-	-	-	36
034	Cyanothece PCC7425	0	3	4	-	12	-	2	-	-	1	2	-	-	-	-	-	-	-	-	-	24
036	Cyanothece PCC8801	1	6	5	-	18	-	9	1	-	-	6	-	-	-	-	-	-	-	-	-	46
039	Dethiosulfovibrio peptidovorans DSM11002	0	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
041	Deinococcus radiodurans	0	5	19	-	122	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	147
042	Desulfovibrio vulgaris	2	22	35	-	16	13	3	1	-	-	-	-	-	1	-	-	1	-	-	2	96
044	Ehrlichia chaffeensis	1	2	1	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	7
046	Escherichia coli BL21	0	1	9	-	2	-	-	2	-	-	1	-	-	-	-	-	1	-	-	1	17
050	Geobacter metallireducens GS15	4	4	11	-	1	1	-	1	-	-	-	-	-	-	-	-	-	1	-	-	23
051	Geobacter sulfurreducens	1	2	3	-	-	2	-	1	-	-	-	1	1	-	-	-	-	-	-	-	11
052	Geobacter uraniireducens RF4	1	5	6	-	-	2	-	1	1	-	-	1	1	-	-	2	-	-	-	-	20
055	Heliobacterium modesticaldum Ice1	1	4	8	-	3	2	2	-	-	-	1	-	1	-	-	-	-	-	-	-	22
061	Kineococcus radiotolerans SRS30216	1	13	115	-	75	-	4	1	2	1	1	-	-	-	2	-	-	-	-	1	216
075	Nakamurella multipartita DSM44233	0	3	26	-	61	-	-	1	2	1	-	-	1	-	-	-	-	-	-	-	95
080	Pelobacter carbinolicus	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
087	Pelagibacter ubique HTCC1002	0	-	11	-	-	1	-	-	-	1	-	-	-	-	-	-	-	-	-	-	13
092	Roseiflexus castenholzii DSM13941	1	22	32	-	24	-	-	-	-	-	1	-	-	-	1	-	-	-	-	-	81
094	Rhodopseudomonas palustris TIE1	1	14	22	-	7	8	1	-	4	-	-	1	-	-	2	-	-	-	-	-	60
099	Salmonella typhi	2	22	33	-	9	5	1	2	4	-	1	-	-	-	1	-	-	-	-	-	80
100	Shewanella amazonesis SB2B	0	3	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5
101	Shewanella baltica OS155	0	8	7	-	2	2	-	3	1	-	-	1	-	-	-	1	-	-	-	-	25
102	Shewanella baltica OS185	0	11	10	1	-	1	-	-	-	1	-	-	-	-	-	1	1	-	-	1	27
103	Shewanella baltica OS195	0	7	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	8
104	Shewanella baltica OS223	0	2	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5
106	Shewanella denitrificans OS217	1	1	2	-	-	-	-	-	1	-	-	-	-	-	-	1	-	-	-	-	6
107	Shewanella frigidimarina	0	10	8	-	1	1	-	-	2	-	-	-	-	-	-	-	-	-	-	-	22
108	Syntrophobacter fumaroxidans MPOB	0	5	2	-	-	1	-	1	-	-	-	-	-	-	-	-	-	-	-	-	9
109	Slackia heliotrinireducens DSM20476	0	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
110	Sanguibacter keddieii	0	9	55	-	83	2	1	-	-	-	-	-	-	-	-	-	-	-	-	-	150
111	Stackebrandtia nassauensis DSM44728	1	2	8	-	13	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	25
114	Shewanella putrefaciens CN32	0	5	2	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8
117	Shewanella ANA3	0	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5
121	Shewanella MR4	0	5	2	-	-	1	-	-	-	-	-	-	-	-	-	-	-	1	-	-	9
122	Shewanella MR7	0	4	2	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	7
124	Shewanella W3 18 1	0	4	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7
126	Synechococcus PCC7002	1	7	13	-	19	2	14	-	-	-	1	-	-	-	-	1	-	-	-	-	58
135	Xylanimonas cellulosilytica DSM15894	2	29	99	-	135	4	5	2	2	1	-	-	-	-	-	-	-	-	-	-	279
136	Yersinia enterocolitica	0	3	14	-	1	1	1	-	1	-	1	1	-	-	-	-	-	-	-	-	23
138	Yersinia pestis CO92	0	4	5	-	1	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	12
139	Yersinia pseudotuberculosis	0	6	29	1	4	1	-	-	-	-	2	-	-	-	1	-	1	-	-	-	45
	Σ	32	357	878	2	1033	70	109	19	32	12	25	6	6	1	10	6	5	3	0	7	2613

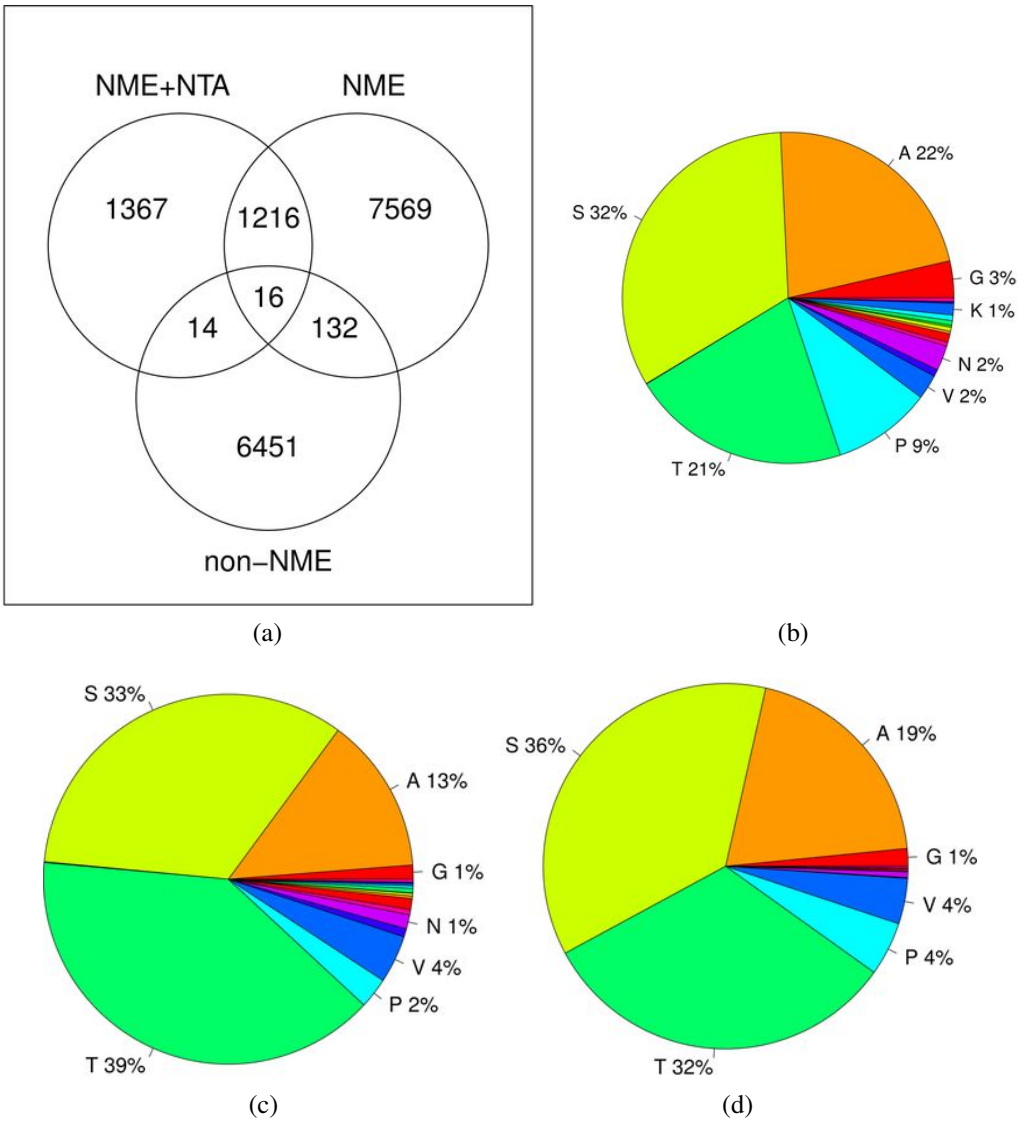


Figure 2.5: Distribution of identifications in 45 different bacterial organisms. Any proteins appearing in both sets are only counted once. Residues with greater than or equal to 1% representation are labeled in each pie chart. (a) Aggregation of Venn diagrams of N-terminal PTM runs on 68 million bacterial spectra across 45 organisms. Of the 2613 NME+NTA, 8933 NME, and 6613 non-NME identifications, the only considerable overlap is across NME+NTA and NME, which is expected. A Venn diagram for each organism was constructed, and the counts for each set and overlapping sets were aggregated into this larger diagram. Sequelog relationships were not taken into account in the diagrams creation. 16% of bacterial N-terminal events were found to be N-acetylations. (b) 8933 NME events with Ser, Ala, and Thr well represented. (c) 2613 Nα-acetylation identifications (NME+NTA) where Ser, Ala, and Thr are again the major residues. (d) The intersection between NME+NTA and NME yields 1216 identifications, comprising mostly of Ser and Thr.

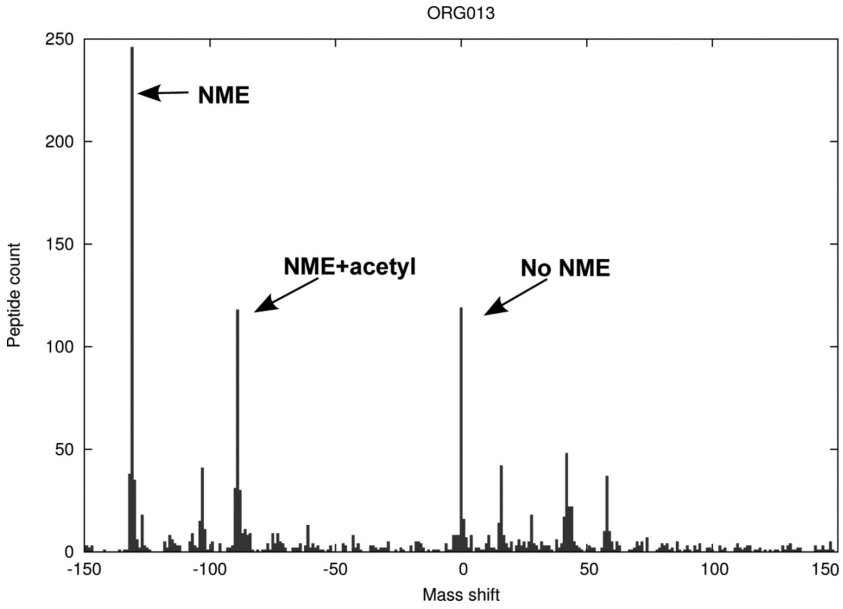


Figure 2.6: Offset counts for the range of -150 to +150 in *A.variabilis*. NME events are represented as peaks located at -131, NME and Nα-acetylation at -89. Retention of Met (no NME event) is represented by a peak at 0.

and NME+NTA events from Table 2.2 and Table 2.1, respectively. It was observed that Ser is well represented in both groups, showing that it is an important component of NME. Table 2.1 shows the breakdown of amino acids at position P2 of each protein with observed NME and NTA. Ala, Ser, and Thr showed a strong tendency for losing the initial Met and for being acetylated. Table 2.2 shows the same amino acid breakdown for P2 of identified NME events using this same MSGF approach.

While Thr is highly represented in these N-acetylated identifications, seen in Figure 2.5(c), no such elevated Thr-conservation at position P2 was observed in the *Shewanella* dataset. This could be due to a more complicated role of Thr in Nα-acetylated proteins, or a reduced role of NTA in the *Shewanella* organisms compared to other bacteria (data not shown). Unfortunately, sequelogs are difficult to determine for distantly related bacterial organisms, preventing further analysis of conservation in bacteria.

Table 2.2: NME P2 residue breakdown using a 1.0×10^{-10} spectral probability threshold and described filtering procedures, using the definition as: -131 ± 1Da neighborhood.

orgnum	Organism	G	A	S	C	T	P	V	D	N	L	I	Q	E	H	M	F	K	Y	W	R	Σ
005	Anaeromyxobacter dehalogenans	4	29	19	-	1	15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	68
007	Actinosynnema mirum DSM	7	34	52	-	35	20	5	-	3	-	-	1	-	-	-	-	2	1	-	-	160
013	Anabaena variabilis ATCC 29413	10	70	76	-	69	38	20	1	6	1	7	1	1	-	2	1	-	-	-	-	303
014	Arthrobacter FB24	15	107	149	-	150	50	14	9	16	2	3	1	4	1	2	-	2	-	-	-	525
017	Bacillus anthracis	17	23	29	-	23	5	4	2	3	2	2	1	2	-	1	1	4	-	-	-	119
019	Brachybacterium faecium DSM4810	8	63	57	-	64	37	3	2	19	2	1	-	-	1	-	-	1	1	-	-	259
021	Chloroflexus aurantiacus	1	4	5	-	2	6	1	1	3	1	1	-	1	-	-	-	3	-	-	1	30
028	Cyanobacterium synechocystis	7	47	55	-	51	17	17	-	7	1	-	-	-	-	-	-	1	-	-	-	203
033	Cyanothece PCC7424	5	20	49	-	38	21	14	1	4	1	-	-	-	-	1	1	2	-	-	-	157
034	Cyanothece PCC7425	7	27	36	-	34	15	1	1	3	-	-	-	-	-	-	-	1	-	-	-	125
036	Cyanothece PCC8801	4	24	40	-	26	10	5	-	1	-	-	-	-	-	-	-	1	-	-	-	111
039	Dethiosulfovibrio peptidovorans DSM11002	7	13	32	-	18	10	2	1	-	-	1	-	-	-	1	-	4	1	-	1	91
041	Deinococcus radiodurans	4	24	51	-	41	15	1	-	-	1	-	-	-	-	3	-	3	-	-	3	146
042	Desulfovibrio vulgaris	6	46	77	-	27	26	4	4	-	-	-	-	-	-	1	-	1	-	1	-	193
044	Ehrlichia chaffeensis	1	4	14	-	3	2	3	1	3	-	-	-	-	-	-	1	1	-	-	-	33
046	Escherichia coli BL21	3	39	66	-	21	8	2	-	1	-	-	-	-	-	1	1	1	-	-	1	144
050	Geobacter metallireducens GS15	9	42	55	-	13	15	-	1	3	1	-	-	-	-	2	-	-	-	-	-	141
051	Geobacter sulfurreducens	13	50	69	1	14	20	3	1	1	-	3	1	4	-	-	4	3	-	-	1	188
052	Geobacter uraniireducens RF4	25	109	126	1	36	46	9	1	3	1	4	1	5	-	1	5	11	-	-	1	385
055	Heliobacterium modesticaldum Ice1	9	17	21	-	14	8	3	-	-	-	-	-	-	-	-	1	-	-	-	-	73
061	Kineococcus radiotolerans SRS30216	11	75	142	1	105	61	14	3	9	2	1	1	-	-	-	-	-	-	1	5	431
075	Nakamurella multipartita DSM44233	10	36	45	-	68	19	5	2	1	2	-	-	-	1	-	1	1	-	-	1	192
080	Pelobacter carbinolicus	1	12	13	-	1	3	-	1	-	-	1	-	-	-	-	-	-	-	-	-	32
087	Pelagibacter ubique HTCC1002	2	19	58	-	13	10	1	2	4	-	-	-	1	-	-	1	2	-	-	-	113
092	Roseiflexus castenholzii DSM13941	2	29	30	-	18	23	-	2	1	1	2	1	1	1	2	-	1	-	-	-	114
094	Rhodopseudomonas palustris TIE1	2	61	68	-	79	16	4	3	12	-	-	-	1	-	1	-	-	-	-	-	247
099	Salmonella typhi	6	83	137	1	83	22	9	1	12	1	4	1	4	3	6	5	12	3	-	3	396
100	Shewanella amazonesis SB2B	8	37	68	-	24	9	2	1	1	1	1	1	-	1	-	2	3	-	-	-	159
101	Shewanella baltica OS155	6	64	107	-	58	17	3	1	12	-	2	-	-	2	1	1	5	1	-	1	281
102	Shewanella baltica OS185	7	58	109	-	57	18	4	1	7	-	-	-	-	-	2	2	3	-	-	1	269
103	Shewanella baltica OS195	5	32	64	-	32	11	-	1	4	-	2	-	-	-	-	2	4	-	-	-	157
104	Shewanella baltica OS223	3	34	77	-	42	9	-	-	3	-	-	1	-	-	1	1	1	-	-	-	172
106	Shewanella denitrificans OS217	3	33	83	-	37	10	1	-	3	-	-	-	-	-	2	1	-	1	-	1	175
107	Shewanella frigidimarina	4	33	79	-	36	10	1	2	4	2	2	-	-	2	-	2	6	1	-	-	184
108	Syntrophobacter fumaroxidans MPOB	4	32	21	-	6	10	2	-	2	-	-	-	-	-	2	-	1	-	-	-	80
109	Slackia heliotrinireducens DSM20476	8	58	27	-	34	22	3	1	-	-	-	-	1	-	-	-	-	-	-	-	154
110	Sanguibacter keddieii	12	52	82	-	100	31	8	2	5	3	2	-	2	-	-	-	-	-	-	-	299
111	Stackebrandtia nassauensis DSM44728	12	32	63	-	67	23	3	1	26	2	-	-	-	-	-	-	-	-	-	-	229
114	Shewanella putrefaciens CN32	5	42	72	-	42	11	2	-	2	-	1	-	-	-	1	1	2	-	-	-	181
117	Shewanella ANA3	6	42	64	-	33	11	-	-	2	-	1	-	-	-	-	1	2	-	-	-	162
121	Shewanella MR4	4	42	69	-	33	8	-	-	-	-	1	-	-	-	-	1	3	-	-	-	161
122	Shewanella MR7	3	41	62	-	31	13	1	-	2	-	2	-	-	-	-	1	2	-	-	1	159
124	Shewanella W3 18 1	4	46	69	-	36	8	2	-	-	-	1	-	-	-	2	1	1	-	-	-	170
126	Synechococcus PCC7002	6	44	56	-	38	17	14	1	4	-	3	2	-	1	1	1	1	-	-	1	190
135	Xylanimonas cellulosilytica DSM15894	19	76	81	-	120	65	15	5	10	6	4	1	-	-	-	-	3	-	-	5	410
136	Yersinia enterocolitica	2	25	28	-	10	3	4	1	5	1	8	3	3	1	1	3	2	2	-	2	104
138	Yersinia pestis CO92	1	13	27	-	7	5	2	-	1	-	2	-	-	-	3	-	-	-	-	-	61
139	Yersinia pseudotuberculosis	4	34	62	-	22	12	4	5	8	3	12	5	6	3	-	6	5	3	-	3	197
	Σ	322	1977	2941	4	1912	861	215	62	216	37	74	22	36	17	40	48	101	14	2	32	8933

NME-conserved Proteins

In order to observe NME importance using comparative genomic approaches, we made certain assumptions about the sequence data.

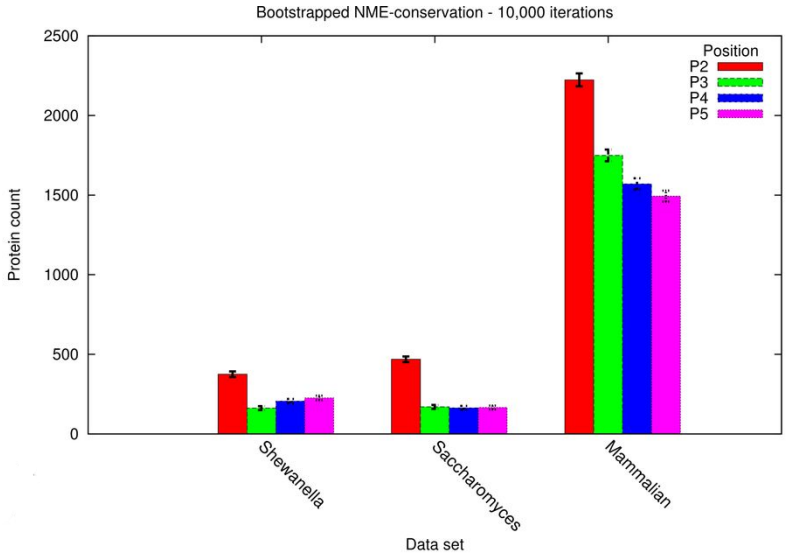
Postulate 1 (Comparative genomics postulate): If a protein undergoes NME in one species and it is essential for its function, then related proteins in other species are likely to undergo NME as well. A protein is called (strongly) NME-conserved if its related proteins (*sequelogs*) across all studied species also undergo NME. Alternatively, a protein can be called *mostly* NME-conserved if *most* of its sequelogs (statistically significant increase as compared to null hypothesis) also undergo NME. This fuzzier definition would also capture functional importance regarding NME, but would add additional complexity to our analysis. We opt for a strict definition of (strongly) NME-conserved through the remainder of the text.

Postulate 2: Finding X-residues in all sequelogs of a protein is statistically surprising unless NME is truly important for the function of the protein. While similar to the first axiom, it is distinct from it since the first one specifies a general relation among sequelogs, and this one relates X-residues to NME and describes the type of signal we should observe from proteins that require NME.

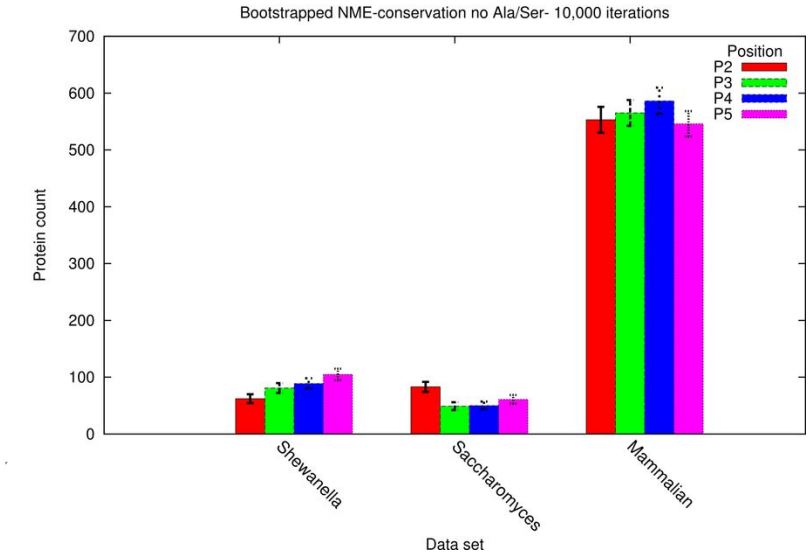
Postulate 3: The baseline compares X-residues at P2 with X-residues at P3, P4, and P5. Since X-residues at P2 are related to NME, and presumably *only* NME, those at P3, P4, and P5 are not related to NME but are instead related to each individual protein's function. These three positions can be used as a baseline for the background of expected X-residue statistics.

Our statistic for X-residues is now defined as follows: For a protein P (having sequelogs in all species) and a position i, define $X(P,i)$ as the number of species in which the amino acid at position i is an X-residue. $X(P,i)$ can vary from 0 to the number of species, n. The number of proteins with $X(P,i) = t$ defines $count(t,i)$ ($count(t,i)$ varies from 0 to the number of proteins, k). Since we care about proteins with X-residues at P2 across all n species, as per postulates 1 and 2, the statistics we compute are $count(n,2)$, $count(n,3)$, $count(n,4)$, and $count(n,5)$. In order to obtain an estimate on each statistic's variance, bootstrap analysis is applied to each

data set using 10,000 iterations.



(a)



(b)

Figure 2.7: Mean NME-conserved protein counts are shown for P2 through P5 for bacterial, yeast, and mammalian datasets. An NME-conserved protein is one containing the residue at the position in question (2 through 5) in X across all sequelogs of that protein. All values are obtained from bootstrap estimated mean and standard deviation of each statistic. (a) The set of residues considered is defined as $X = \{G,A,S,P,V,T,C\}$. It is evident that P2 contains a larger number of NME-conserved than P3-P5. (b) NME-conserved protein counts, without Ala and Ser, are shown for P2 through P5 for the three datasets. Here the X set does not contain Ala and Ser, and is defined as $X = \{G,P,V,T,C\}$. Not including Ala and Ser in X produces no elevated counts at P2 compared to P3, P4, and P5. This suggests that the boost in NME-conserved counts seen in (a) can be attributed to the Ala and Ser residues.

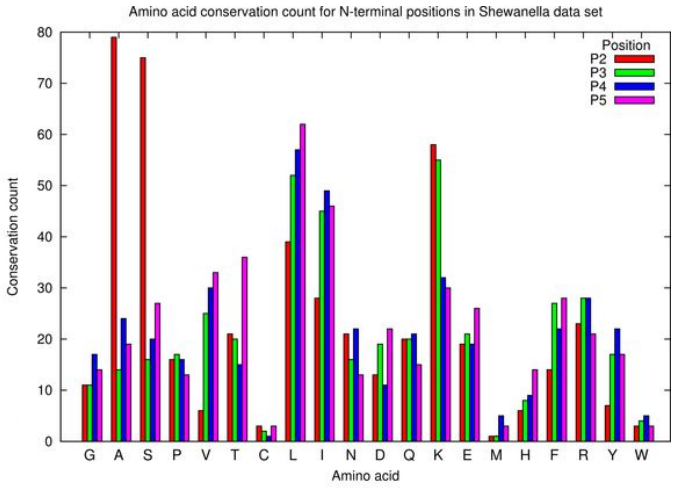
The mean counts for each dataset are shown in Figure 2.7(a), for the *Shewanella* dataset there are 375 such proteins for position 2, in contrast to only 162 for position 3. This is a significant increase in conservation of X-residues at the position directly affecting NME. Similar trends are observed in yeast and mammalian species, also seen in Figure 2.7(a). Yeast shows a similar pattern with ≈300 more proteins with perfectly conserved X-residues at position 2 as compared to other positions. The mammalian species also show a large difference (≈500 proteins) between position 2 and other positions. Furthermore, the differences in X-residue counts between positions 2 compared to 3, 4, and 5 cannot be attributed to variance in the statistic used.

The higher conservation of X-residues at the second position, seen in all datasets in Figure 2.7(a), compared to Figure 2.7(b), indicates that NME may be functionally critical for some proteins and provides a candidate list of such proteins. Proteins whose sequelogs in all species have an X-residue in the second position are designated as NME-conserved proteins.

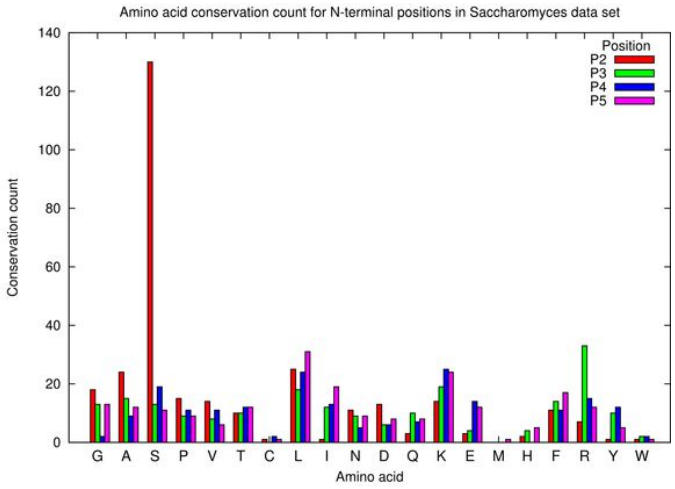
Conservation of Ala- and Ser- as the P2-residue in proteins

After analyzing for group conservation of all seven X-residues at initial positions of the proteins (checking whether the second residue in each sequelog is an X-residue), analyses of individual amino acids (instead of the set X) at the second position and comparison to the conservation at the third position (where conservation is not expected) as a control were performed. Figure 2.8 and Table 2.3 show that Ala is conserved among all *Shewanella* species in 79 proteins at position 2 but only in 14 proteins at position 3. Ser is conserved in 75 proteins at position 2 but only in 16 proteins at position 3. Conservation levels of other amino acids are rather similar between positions 2 and 3, suggesting that proteins with Ala and Ser in the P2 position may be the important targets for NME in *Shewanella*. Since Ala and/or Ser are exceptionally well-conserved, it is reasonable to assume that NME affects the function of these proteins (if one accepts the comparative genomics postulate).

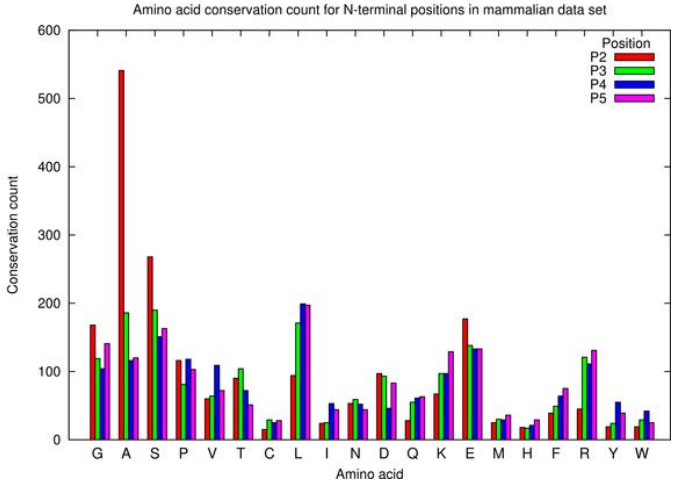
Figure 2.8: Conservation of residues are shown for P2 through P5 for datasets from (a) *Shewanella* (b) *Saccharomyces* (c) mammalian. Ala and Ser are by far the most conserved in P2 across all three datasets. The number of proteins with Ala in the second position in *Shewanella*, *Saccharomyces*, and the mammalian datasets are 79, 24, and 541, respectively. The number of proteins with Ser in P2 in *Shewanella*, *Saccharomyces*, and mammalian datasets are 75, 130, and 268, respectively.


Amino acid conservation count for N-terminal positions in Shewanella data set
Position
P2
P3
P4
P5
Conservation count
Amino acid
G A S P V T C L I N D Q K E M H F R Y W

(a)


Amino acid conservation count for N-terminal positions in Saccharomyces data set
Position
P2
P3
P4
P5
Conservation count
Amino acid
G A S P V T C L I N D Q K E M H F R Y W

(b)


Amino acid conservation count for N-terminal positions in mammalian data set
Position
P2
P3
P4
P5
Conservation count
Amino acid
G A S P V T C L I N D Q K E M H F R Y W

(c)

Table 2.3: Conservation of single amino acids for *Shewanella* sequence data set (positions 2 and 3). Ala and Ser (shown in bold) are by far the most conserved residues in the 2nd position.

AminoAcid	P2	P3
G	11	11
A	**79**	**14**
S	**75**	**16**
P	16	17
V	6	25
T	21	20
C	3	2
L	39	52
I	28	45
N	21	16
D	13	19
Q	20	20
K	58	55
E	19	21
M	1	1
H	6	8
F	14	27
R	23	28
Y	7	17
W	3	4

Table 2.4: Conservation of single amino acids for *Saccharomyces* sequence data set (positions 2 and 3). Ser (shown in bold) is by far the most conserved residue in the 2nd position.

AminoAcid	P2	P3
G	18	13
A	24	15
S	**130**	**13**
P	15	9
V	14	8
T	10	10
C	1	0
L	25	18
I	1	12
N	11	9
D	13	6
Q	3	10
K	14	19
E	3	4
M	0	0
H	2	4
F	11	14
R	7	33
Y	1	10
W	1	2

Ala and Ser also exhibit much higher conservation at the second position as compared

Table 2.5: Conservation of single amino acids for mammalian sequence data set (positions 2 and 3). Ala (shown in bold) is by far the most conserved residue in the 2nd position.

AminoAcid	P2	P3
G	168	119
A	**541**	**186**
S	268	190
P	116	81
V	60	64
T	90	104
C	15	29
L	94	171
I	24	25
N	53	59
D	97	93
Q	28	55
K	67	97
E	177	138
M	25	30
H	18	17
F	39	49
R	45	121
Y	19	24
W	19	29

to other N-terminal positions indicating that they are specifically required here (rather than in a few initial positions) and suggests that this elevated conservation is relevant to NME. *Shewanella*, yeast, and mammalian species all show much higher levels of conservation for Ala and/or Ser in the second position compared to other residues. The one exception to this pattern is seen for Lys in *Shewanella*, Figure 2.8(a), where Lys is elevated at P2 and P3 compared to P4 and P5. This violates our third postulate, and is therefore considered to not be an NME relevant pattern.

This trend of Ala/Ser importance appears again in Figure 2.7(b), which displays X-residue conservation across bacterial, yeast, and mammalian datasets. For the data in this figure, Ala and Ser are removed from the X set, i.e. $X = \{G, P, T, C, V\}$ and there is little change in the level of conservation from P2 to P3, P4 or P5 for sequelogs within each dataset. Most of the change from P2 to P3, P4 or P5, of which there is little, can be accounted for by variability of the statistic. Comparing Figure 2.7(a) to 2.7(b) shows that Ala and Ser are responsible for the elevated levels of conservation at position 2 for these sequelogs.

2.1.4 Conclusion

Although it has been known for some time that protein synthesis is essentially universally initiated with Met [JH70], it has only become appreciated over the intervening years that this residue is subsequently removed from a considerable portion of the proteins actively synthesized in any given organism. Much of this NME for cytoplasmic and nuclear proteins occurs as a co-translational event catalyzed by specific MetAPs [BBW98]. There are two classes of these enzymes but both have quite similar substrate profiles [AKH$^+$95]. In a eukaryote such as yeast, the number of proteins predicted to lose or retain the initiator Met (on an individual gene basis) is nearly equal; however, based on protein mass, a great preponderance of soluble proteins undergo this modification and are, moreover, then Nα-acetylated.

The functional purposes of NME are poorly understood and there are several possible reasons that NME is such an essential process. In the first place, it has been argued that NME allows the rapid recycling of Met to keep the levels of the intracellular pools at an adequate level for the many roles that this amino acid plays in addition to its involvement in protein synthesis (both as an initiator and as a constituent residue). Met is generally relatively scarce and the recovery of most (on a mass basis) of the Met expended in initiating protein synthesis is certainly likely to be beneficial to the cell. However, this ascribed role for NME begs the question of why Met is recovered from only a subset and not all proteins. The answer, while likely complex, may be that having two classes of protein (those that lose their Met and those that do not) is desirable for other reasons. Indeed, it is possible that primordial MetAPs did not make the distinction between these two substrate types and Met was recovered from all synthesized proteins. However, if this was the case, evolution altered these enzymes to their present day specificity very early on and these changes were adopted in all life forms that are presently known. There are few biological processes that show this level of conservation in all living organisms. Walker and Bradshaw [WB99] demonstrated that NME specificity in vitro can be changed (by site-directed mutagenesis of yeast MetAP1) to expand the number of P2-residues that allowed Met removal, so even within the present MetAPs there is no structural

reason to prevent the development of an enzyme (through mutation and natural selection) that has a broader and possibly comprehensive substrate profile. The fact that this did not happen suggests additional roles for NME and the need for the two classes of protein that it generates.

By the same token, being able to expand the specificity does not ensure that it would be possible to further close down the repertoire of substrates. It may well be that the substrate specificity that the present day MetAPs display is basically at the lower limit (in terms of the size of the P2-residue) that nature can engineer with this scaffold. This may explain the dominance of Ala/Ser substrates (controlled by evolutionary pressure on the second position) while still allowing the five or so additional substrates that are also cleaved.

A second proposed function of NME is linked to protein stability/degradation. Thus, the profile of NME produced by MetAP action fits, at least in eukaryotes, to that dictated by the NER. This suggests, given the very early development of MetAP specificity, that the recognition component of the NER that binds certain N-terminal residues to allow the polyubiquitinylation and subsequent proteasomal degradation evolved to be compatible with it, rather than the other way around. As such, the removal of Met (from the 7 amino acids with the smallest side chains) is not so much a stabilizing event as it is a permissive one, i.e. these N-termini are simply ignored by the NER machinery. Since there is a singular lack of evidence that the proteins in the non-NME class are ever subsequently degraded via the NER (after exposure downstream of the destabilizing residues), it suggests that the role of NER is to capture and destroy mis-folded and/or damaged proteins, spuriously formed fragments and any other peptidic detritus that may arise in the course of other cellular activities rather than as part of programmed protein degradation. Recently, Hwang et al. [HSV10] have revised the original NER by showing that Nα-acetylation can produce N-termini that are also recognized by the germane ubiquitin ligase and thus arguing now that NER is really a quality control mechanism for protein stoichiometry in the cell. They also conclude that it mainly functions as a scavenger for improperly folded proteins.

A third raison d'etre for NME is the preparation of the N-terminus for further post-

translational modification. For purposes of this discussion, no strong distinction is drawn between co- and post-translational events. In eukaryotes, Nα-acetylation is a prominent modification of four of the seven residues exposed by NME but is certainly not the only one. In addition, the initiator Met is acetylated when the P2-residue is Asp, Glu or Asn [PNT+99]. Hwang et al. [HSV10] have recently reported that 5 residues (Ala, Val, Ser, Thr, Cys) are the primary sites of NTA after NME and argue that they should now be considered destabilizing residues in the NER. However, in an extensive study of Nα-acetylation by mass spectrometry, it was shown that Ala and Ser account for 89% of all the proteins modified by acetylation that are NME targets in Homo sapiens, implying that Nα-acetylation of Val, Cys, and even Thr is also rare in eukaryotes [HGR+10]. The NATs responsible for NTA (the NatA, NatB, and NatC complexes) have been studied in a variety of eukaryotic organisms [PAS09]. NatA modifies Ala, Ser, Gly, and Thr so it requires NME to prepare its substrates. Knockdown of Naa10 (previously Ard1, standardization introduced [PAS09]) and Naa15 (previously Nat1), which comprise NatA, in *Caenorhabditis elegans* resulted in embryonic lethality [SKW+05]. The human NatA complex has also been shown to be important in apoptosis as shown in Naa10 (previously hARD1) and Naa15 (previously NATH) knockdowns in human HeLa cells [MSEE02]. The necessity of NatA in yeast is less clear-cut. Polevoda et al. [PNT+99] produced deletion mutants of Naa10/Naa15 (previously ard1 and nat1, of the NatA complex), Naa20 (previously Nat3, of the NatB complex), and Naa30 (previously Mak3, of the NatC complex), all of which are viable; however, all display negative phenotypes.

The extensive occurrence of this reaction in eukaryotes suggests that it is an important modification. However, there is little evidence for any direct role of this group in any function. It may be that it is involved in, as yet unidentified, protein-protein interactions perhaps through a specific binding motif such as found for other PTMs. Alternatively it may contribute generally to the cellular environment by blocking a significant number of α-amino groups to prevent non-enzymatic glycation with reducing sugar metabolites (which would make these energy rich compounds unavailable for catabolic breakdown). It seems not to be significantly involved, as

was once proposed, in stabilizing proteins against spurious aminopeptidase attack.

In contrast, modifications such as N-myristoylation clearly have functional consequences. This modification has been shown to be important for a small subset of proteins, and it is estimated that 0.5% of eukaryotic proteins undergo this modification [MSEE02]. The functional role of NME is unlikely to be strongly connected to N-myristoylation as only a small percentage of proteins undergo this addition and even fewer require it for proper protein function. Also, N-myristoylation is not as universal as NME since it does not occur in bacteria. The analyses herein (Table 2.5) reveal some over-conservation of Gly in mammals (168 Gly-conserved proteins for the second position as compared to 119 Gly-conserved proteins for the third position). However, this 40% increase provides a somewhat weaker support for the functional role of Gly as compared to 190% increase for Ala. Therefore, while Gly-conserved proteins undoubtedly exist, their modest amount of over-conservation does not allow one to conclude that exposing Gly is an important role of NME.

The inclusion of the commonly acetylated N-termini in the NER [HSV10] significantly changes the original model. The N-termini now considered to be degrons compose 18 of 20 residues, including 5 of 7 X-residues: A, V, S, T, and C, leaving out only X-residues G and P due to the fact that they are rarely Nα-acetylated (in yeast). We argue that if G and P are not considered degrons, then V, T, and C should also be removed from the degron list due to their rare acetylation state, (seen in Figure 2.4). This leaves Ala and Ser as degrons that are heavily Nα-acetylated in yeast, suggesting that the predominant removal of Met and subsequent acetylation of Ala and Ser combine to provide a principal role for NME and the connection to the NER degradation pathway.

One explanation for the conservation of Ala and Ser might be the influence of the underlying nucleotide sequences of the Met and second codon. The Kozak consensus RccATGG is a regulatory motif playing an important role in the initiation of translation in vertebrate and some other eukaryotes [Koz89]. Since the Kozak consensus (where ATG represents the first codon in the protein) suggests that the second codon starts with G, it might be expected that

there would be higher levels of Ala at P2 in vertebrates due to the frequency of G at the first position of Ala codons. However, by using much larger datasets than available to Kozak at the time the rule was formulated, Xia [Xia07] showed that, contrary to the Kozak rule, there is no link between G at the fourth position and translation initiation efficiency. Xia further suggested that it is the presence of specific amino acids at P2 that led Kozak to (erroneously) conclude that G at position 4 affects the translation efficiency. The observations reported herein that Ala and/or Ser are universally over-represented in the second position in nearly all species cannot be explained by the Kozak consensus rules since it is not universal, e.g., it is not applicable to bacteria and does not extend beyond the first codon in fruit flies. Yeast consensus codons UC[U/C] as reported in Hamilton et al. [HWdB87] also appear not to be in support of translation initiation efficiency (shown in Text S2).

It is appropriate to emphasize that the goal of this study was to detect whatever statistically significant differences might occur between the second and third positions that could offer insight into NME function. Indeed, the fact that Ser/Ala have elevated levels of conservation at P2 in species as diverse as bacteria and humans illustrates that there exists an enormous evolutionary pressure for retaining them in the second position. Since this analysis reveals no over-conservation for 18 out of 20 amino acids, it can be argued that exposure of Ala and/or Ser by NME defines its key function.

While this study reveals only two over-conserved residues in the second position in the species studied, it cannot be ruled out that additional residues (e.g., Thr) may be over-conserved in other species. However, the 7 seemingly equivalent residues (with respect to NME specificity) clearly have vastly different evolutionary patterns suggesting that some of them (e.g., Ala and Ser) play an important functional role while others may represent inconsequential players. Thus, NME appears to be a mechanism evolved to resolve the "conflict" between the uniformity of the translation initiation mechanism and the variety of functions affecting the N-termini of proteins.

2.2 Acknowledgements

Chapter 2 is published in full as **S. Bonissone**, N. Gupta, M. Romine, R.A. Bradshaw, and P.A. Pevzner. N-terminal protein processing: A comparative proteogenomic analysis. Molecular & Cellular Proteomics, 12(1):1428, 2013 . The dissertation author was the primary author of this paper.

Chapter 3

Immunoglobulin classification

3.1 Introduction

The antibody molecule is comprised of two pairs of two distinct proteins: the *heavy* and *light* chains. In humans, there exist a single heavy chain locus, and two light chain loci. These heavy and light chains pair with one another to form a 'Y' shaped protein structure. The tips of this immunoglobulin (Ig) molecule interact and bind to different antigens within one's body, signaling an immune response. Unlike typical transcripts within eukaryotic cells, the heavy and light chain transcripts are not directly taken from exonic segments of the individual's genome. Instead, there are three distinct classes of exon-esque gene-segments, termed the variable (V), diversity (D), and joining (J) gene-segments. Each of these classes of gene-segments contains many different variants encoded in an individual's genome. The light chain transcript contains only V and J gene-segments, while the heavy chain transcript contains V, D, and J gene-segments. Both heavy and light chains also contain a constant (C) gene-segment that does not contribute to combinatorial diversity.

Unlike typical exonic splicing, that is precise, somatic recombination of antibody gene-segments is inexact with the exonuclease removing several base-pairs from each end of the gene-segments. Ligation of D to J, and subsequently DJ to the V gene-segment, is also

imprecise with deoxynucleotidyl transferase (TdT) incorporating non-templated base pairs into the resulting gene [DYP$^+$84]; a process known as V(D)J recombination. In addition to the variability induced by somatic recombination, somatic hypermutation (SHM) events introduce additional deviations from germline gene-segments. The end result of this process is a B-cell that produces a single type of antibody, a monoclonal antibody (mAb). This increased variability allows for a larger search space of antibody configurations to be explored for specificity to a particular antigen. While this is advantageous from the perspective of our immune system's adaptability to foreign substances, analysis of these highly variable immunoglobulin genes becomes difficult.

Repertoire construction forms the basis for the analysis of antibodies; characterizing the pool of gene-segments that were selected for a particular antigen. A prerequisite step for repertoire analysis is the labeling of V, D, and J gene-segments for the read of each heavy and light chain. This VDJ labeling problem can be described as the following: given reference gene-segment sets $\mathcal{V}$, $\mathcal{D}$, $\mathcal{J}$, and a read, return the "most likely" labels $v \in \mathcal{V}$, $d \in \mathcal{D}$, and $j \in \mathcal{J}$ for this read. Despite this problem being easily described, it remains unclear how to design an adequate and easy-to-compute likelihood estimator for VDJ classification. As a result, this classification can be difficult and error prone, particularly for the heavy chain. While all described approaches also operate on the light chain, we focus on the heavy chain due to its difficulty in correctly identifying composite gene-segments.

Existing tools for repertoire characterization rely on aligning reads against the reference sequences of V, D, and J gene-segments from the organism in question [WJW$^+$09, ALC$^+$11, CPZ$^+$12, JHW$^+$13]. This strategy is exemplified by IMGT-VQUEST [BLG08] (the most widely used VDJ classification tool) and other tools [VCK06, GMJ$^+$07, WWZ$^+$08, YMMO13, SCLR$^+$04, OLNLB06]. Most of these tools rely on an iterative approach where first the best matching V gene-segment is identified, then J, and finally D. This specific order of alignments (from V to J to D gene-segment) is appealing because it starts from the longest (and thus resulting in the most confident alignment) gene-segment and ends with the shortest

(and thus resulting in the least confident alignment) gene-segment. However, it also suffers from uncertainties in alignment (there are usually multiple optimal alignments) and sequential dependencies in the iterative alignment (at each step, previously matched nucleotides are removed from future alignments).

To address this sequential dependency bias shortcoming, we describe a colored de Bruijn graph based approach, which leverages the current understanding of V(D)J structuring of antibody transcripts. Similarly to recent attempts to remove biases of previous alignment-based approaches in genomics applications, we now introduce the concept of de Bruijn graph to immunoinformatics. Iqbal et al., 2012 [ICT$^+$12] introduced the colored de Bruijn graph for identifying variants across genomes, we repurpose this approach for use with antibodies. The resulting algorithm IgGraph does away with the sequential nature of iterative alignment, and provides accurate labeling of reads. IgGraph is shown to perform well on both real immunoglobulin sequencing (Ig-seq) datasets, and simulated datasets with varying levels of deviations from reference gene-segments. At the same time, we show that the problem of VDJ classification is far from being resolved as the leading tools produce remarkably different results when applied to large Ig-seq datasets.

3.2 Colored antibody graph for antibody classification

3.2.1 Methods

Antibody sequencing and the CDRs

The transcripts of the heavy/light chains can be sequenced using reverse primers located in the constant regions, and forward primers located at different positions of the different V gene-segments. Sequencing of these transcripts can then be performed after PCR amplification. The VDJ region of heavy chains is approximately 110 amino acids (330bp), which is why the previous literature favored the Roche 454 platform due to its larger read lengths

of approximately 450bp. However, with Illumina's increasing read length and throughput, recent and future studies face the challenge of analyzing large repertoires with millions of reads [SBK+15].

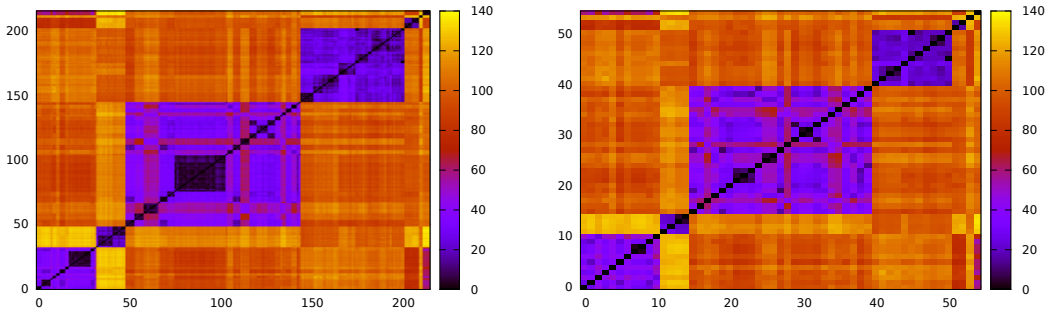
The heavy and light chains have three subsequences, termed *complementary determining regions* (CDRs) due to the role they play in defining a particular antibody's antigen binding specificity. These CDRs, denoted CDR1, CDR2, and CDR3, while located along the length of each immunoglobulin chain, are in close spacial proximity at the physical 'tips' of the antibody structure. The location at the junction, along with exonucleotide chewback and non-templated nucleotide addition, all contribute to the larger variability in CDR3 length. Since CDR1 and CDR2 are located entirely within the V gene-segment, they are only subjected to somatic hypermutation.

V/D/J antibody segments

213 V, 30 D, and 13 J gene-segments are annotated as functional (and complete) in the international ImMunoGeneTics (IMGT) database [RHM+13]. Of these 213 V gene-segments, many are allelic variants of one another; differing in a few nucleotides from another allelic variant of the same gene. High similarity between these allelic variants adds complexity to the problem of VDJ classification. Even after collapsing allelic variants to their consensus sequences (that results in only 55 *consensus* V gene-segments), there are still many similar fragments between these consensus sequences. Figure 3.1(a) visualizes similarities between 213 V gene-segments and Figure 3.1(b) visualizes similarities between 55 consensus V gene-segments.

Simulating antibodies

To generate simulated data that properly represent the challenge of VDJ labeling from reads, we needed to simulate the VDJ somatic recombination events that drive the diversity of the CDR3 region of antibodies. Unfortunately, there are no publicly available



(a) (b)

Figure 3.1: Edit distances between (a) 213 human V gene-segments (alleles) and (b) 55 consensus V gene-segments. The consensus V gene-segments illustrate that, even after collapsing highly similar allelic variants into consensus V gene-segments, many of the 55 consensus V gene-segments remain similar to each other.

antibody simulators, despite many existing tools having used simulated antibody sequences to demonstrate performance [ALC+11, VCK06, WWZ+08]. To this end, a simulated monoclonal antibody (smAb) is generated by the process detailed in Supplementary Figure A.1: selecting a V, D, and J gene-segment to comprise our smAb; exonuclease chewback on the 3' V, 5' and 3' D, and 5' J segments; and finally non-templated nucleotide addition to these same regions. To simulate these biological processes, empirical distributions for exonuclease chewback length [JGSC04], as well as composition and length distributions for non-templated nucleotide additions [BHM83], were used. Using this process to create a smAb, we are able to generate datasets with labeled V/D/J segments with simulated *biological* diversity.

These smAbs can then be sampled using a read simulator, to further introduce *sequencing* errors. The Grinder [AWR+12] read simulator can be used to generate Illumina and 454 reads. Additionally, we also want to generate datasets of smAbs with a fixed number of deviations from the germline sequence, i.e., mutations. To this end, positions along the V gene-segment were selected from a distribution of mutations created from 23,051 annotated IMGT sequences. These positions were selected without replacement to ensure a fixed divergence from germline references. The V, D, and J gene-segments from human were collected from the IMGT database [RHM+13] as the basis for the simulation of each smAb.

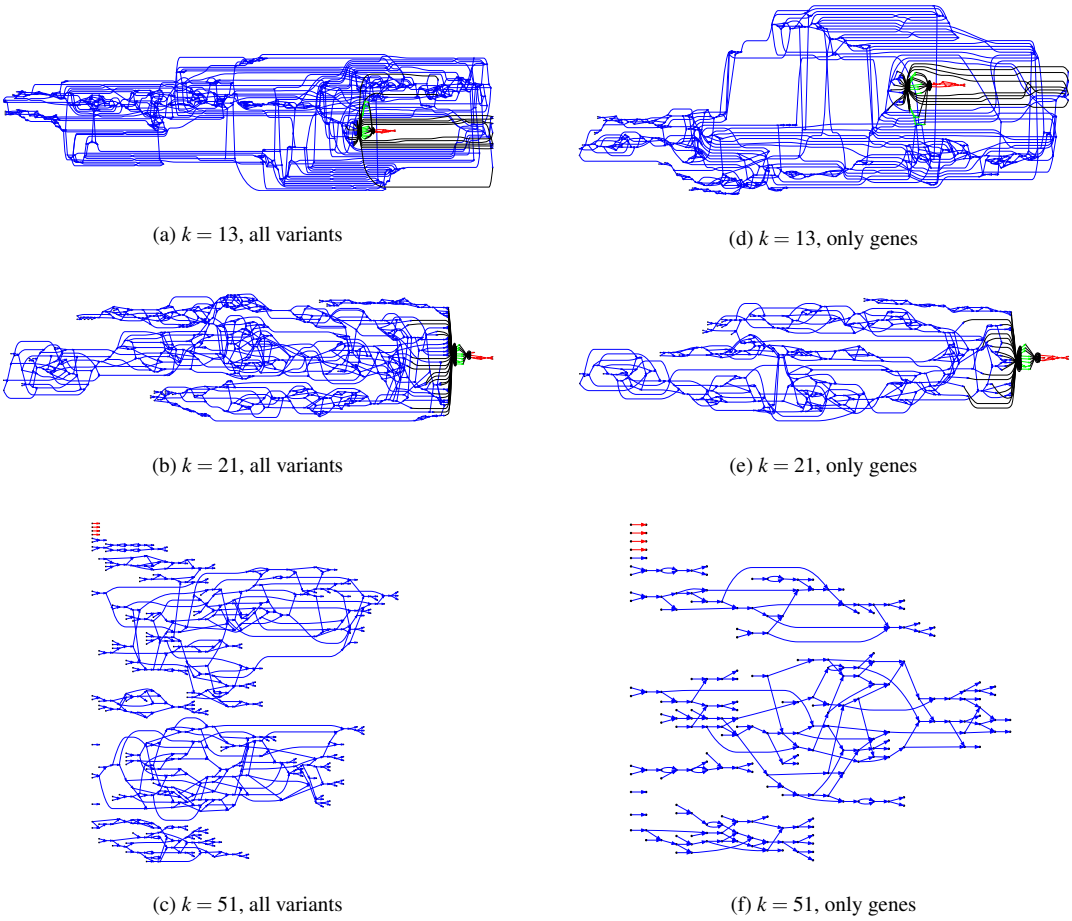
Canonical antibody graph

The canonical antibody graph is created by constructing a de Bruijn graph of each set of V, D, and J gene-segments, and creating an artificial joining of nodes at the V/D and D/J segments. Figure 3.2 shows multiple versions of this graph for different parameters k. The differences when creating the canonical antibody graph with either all alleles (left), or all consensus gene-segments (right), is shown. The arcs in this graph are colored blue for V gene-segments, green for D gene-segments, and red for J gene-segments. The arcs artificially joining gene-segments are colored black. This canonical antibody graph was created for three values of k to show the connectivity between the different sets of reference gene-segments. The graph constructed with $k = 13$ shows sharing of k-mers between V and D gene-segments, as well as amongst different V genes. The parameterization of $k = 13$ results in a very complicated graph, this complexity of the visual representation is partially exacerbated by the graph visualization layout algorithm. It is the relative comparison of complexity between the graphs in Figure 3.2 that is meaningful.

Antibody graph

Given a set of reads $\mathcal{R}$ from mAbs, we construct the de Bruijn graph (termed *antibody graph*) over the k-mers of these reads in the following manner. Nodes in this graph represent all $(k-1)$-mers over the set of reads $\mathcal{R}$. Nodes u, v are connected by a directed edge (arc) (u, v) if u is a prefix, and v is a suffix of some k-mer in a read from $\mathcal{R}$. More on applications of the de Bruijn graphs for assembly can be found in [CPT11].

We can also incorporate IMGT reference gene-segments into the antibody graph. Reference gene-segments $\mathcal{C}$ can be added to the antibody graph, and considered as 'colored' reads. For example, the human antibody graph has 213 + 30 + 13 = 256 colors (corresponding to 213 V, 30 D, and 13 J gene-segments). In comparison, the mouse antibody graph has 242 V, 27 D, and 8 J gene-segments, for a total of 277 colors. A total of $|\mathcal{C}|$ reference gene-segments are added to the antibody graph in a similar manner as the (virtual) reads, with an additional data



(a) $k = 13$, all variants

(d) $k = 13$, only genes

(b) $k = 21$, all variants

(e) $k = 21$, only genes

(c) $k = 51$, all variants

(f) $k = 51$, only genes

Figure 3.2: The canonical antibody graph for different values of k (arcs corresponding to the V, D, and J gene-segments are colored blue, green, and red, respectively) constructed for all alleles (left) and all consensus gene-segments (right). All non-branching paths are collapsed to a single arc, and at each junction, a dummy node is created to connect V gene-segments to D gene-segments, and D gene-segments to J gene-segments, these arcs are colored black. These graphs are constructed with $k = 13$ ((a) and (d)), $k = 21$ ((b) and (e)), and $k = 51$ ((c) and (f)). (b) shows V, D, and J gene-segments completely separated, while (a) shows considerably more sharing of arcs in the V segments, and some shared in the D gene-segments. Increasing the value of k (c) greatly simplifies the relationship among V gene-segments. This is not a feasible parameter for our purposes (as no D segments are captured) but does show the complexity of V gene-segments. In the case of $k = 51$, the graph becomes disconnected (and green edges disappear) since it exceeds the length of the longest D gene-segment.

structure. Each arc along a reference read path i is assigned the color $c_i \in \mathcal{C}$. A hash of arcs to a set of colors, $\mathcal{H}_C$, is maintained as each reference sequence is added to the graph. The hash can then be queried given an arc e, e.g., $\mathcal{H}_C[e] = \{c_1, c_3, c_4\}$, to retrieve all the colors present on that arc. Edges from reads are assigned a special, "non-colored" symbol representing their lack of color (shown as black edges in subsequent examples).

The antibody graph incorporating reference gene-segments is termed the *colored antibody graph*. This graph represents the sequenced mAb repertoire and their similarity to reference gene-segments, an idealized depiction of this graph is shown in Figure 3.3.

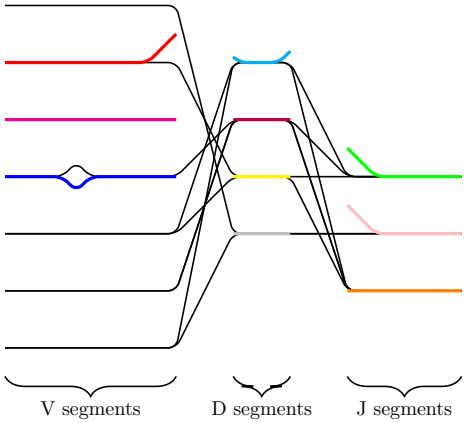


Figure 3.3: Colored antibody graph. An idealized colored antibody graph built over the reads, with reference gene-segments represented as distinct colors. Imperfect overlay of reference gene-segments at V/D and D/J segments is common. Also detectable is the divergence of V-segments from their references, helpful in determining differences in CDR1 and CDR2 regions.

Color profile

The coloring of the antibody graph relies on common structures in de Bruijn graphs referred to as *bulges* and *tips* [CPT11] that will help us to repaint black edges into colors corresponding to reference gene-segments. Given a set of reference gene-segments $\mathcal{C}$, a $|\mathcal{C}| \times n$ *color profile* matrix C can be constructed for a read of length n, representing the associations of each color to each position of the read. At first glance, it is unclear how to assign new colors to arcs on the black path TCC $\rightarrow$ CCG $\rightarrow$ CGC $\rightarrow$ GCA $\rightarrow$ CAG in Figure 3.4. However, one can

note that this path forms a bulge with the colored path TCC $\rightarrow$ CCA $\rightarrow$ CAC $\rightarrow$ ACA $\rightarrow$ CAG that we will use for coloring the black path as described below. A similar approach is applied to tips, such as ATA $\rightarrow$ TAT. Construction of the color profile matrix is accomplished by considering each color $c_i \in \mathcal{C}$, and traversing each arc e from read r, noting when $c_i \in \mathcal{H}_{\mathcal{C}}[e]$. This condition determines the value at $C[c_i][e]$, the cell in the color profile matrix for the color and position, which is updated to note the match/mismatch with color c_i at the position of arc e. Figure 3.4 shows an example graph with $\mathcal{C} = \{\text{red}, \text{blue}, \text{green}\}$, and a single read depicted with black arcs. In this example, read arcs (in black) TAT $\rightarrow$ ATC, ATC $\rightarrow$ TCC, and CAG $\rightarrow$ AGG are shared with different reference segments, the contents of $\mathcal{H}_{\mathcal{C}}$ for these arcs are shown in the Figure. It is worthy to note in this example that reference segments share arcs, e.g., red and green sharing three arcs, something that is common for allelic variants of V gene-segments. This color profile represents an abstraction for scoring the reference gene-segments to a read r.

Color propagation

Deviations from reference gene-segments create bulges and tips [PTT04, ZB08]. A bulge is created when a read deviates from a reference gene-segment and is not near either end. A tip is created when this deviation occurs near either end of a reference gene-segment or read. The assignment of each color to the read can be greatly affected by bulges and tips between a read and a colored reference sequence. This particularly effects V gene-segments due to somatic hypermutations, as such, we must ensure the color propagates through these arcs such that small differences between a read and a gene segment do not result in a "loss of color". Bulges arising from mutations in the V gene-segments are traversed, and the color profile is adjusted accordingly. Figure 3.5a shows color propagation for a de Bruijn graph constructed with $k = 5$ when a single reference segment (red arc color), and a single read (black arcs), have a single nucleotide variation between them. Above each red arc is the arc marginal (last nucleotide of the corresponding k-mer) for the reference, similarly, below each black arc is the marginal for the read.

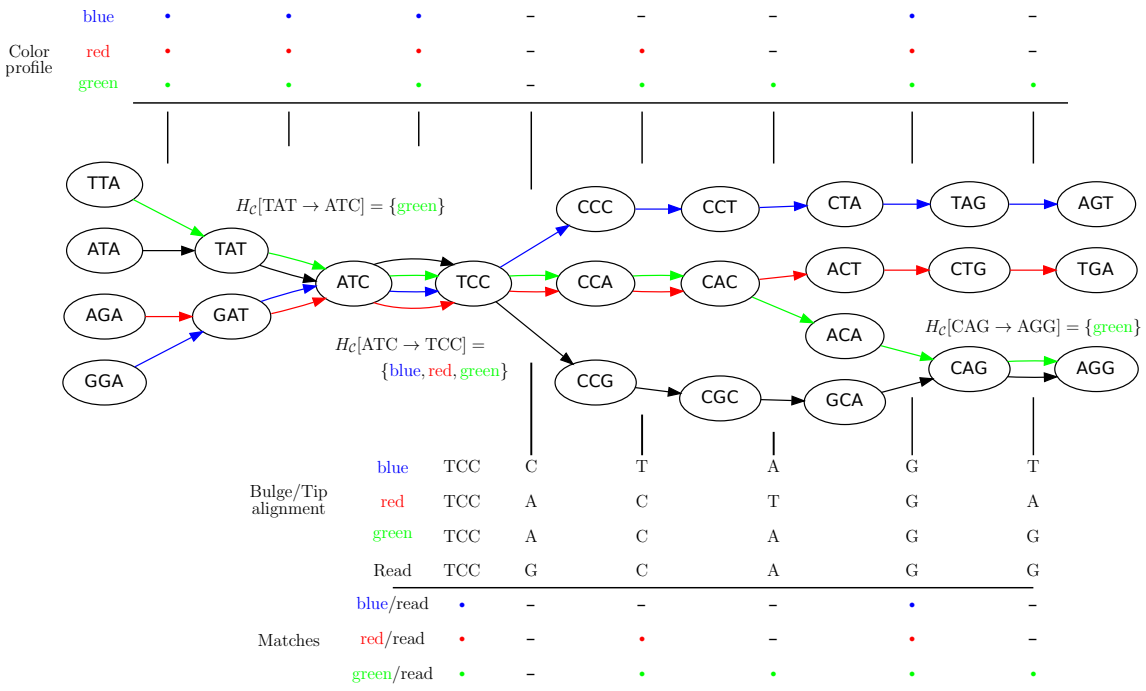


Figure 3.4: An example antibody graph with three reference segments, colored by red, blue, and green arcs. A single read is shown here with black arcs. The color hash $\mathcal{H}_C$ is shown for the three arcs from the read that are shared with reference gene-segments, TAT → ATC, ATC → TCC, and CAG → AGG. Bulge/tip traversal and color assignment is shown below the graph. E.g., to obtain the matching for the green reference, the green/black bulge is traversed, and marginals are aligned. Tips are also traversed, shown here with red and blue references. Matching/mismatching nucleotides are noted for each colored reference to the read at the bottom of the figure. Matches are noted with a •, and mismatches with a -.

The information contained in the arc marginal aids us in creating the color profile of a read. In our example (Figure 3.5a), this matrix is of dimension 1×10, since we have only a single color in our set of colors $\mathcal{C} = \{\text{red}\}$, and a fragment represented by 10 arcs. Two different color profiles are shown in Figure 3.5a, a 'Raw' and a 'Propagated'. These color profiles are shown with red/black rectangles denoting matches/mismatches over each position i.e., each arc marginal. If we merely traverse the arcs of the read, we would obtain the color profile 'Raw' showing five mismatches, colored black, in C. If we instead traverse the bulge, i.e., traverse both the read and reference paths, we obtain the subsequences of the read and reference over the bulge. These subsequences can then be aligned to fill in the color profile and only report a single mismatch, shown as 'Propagated' in Figure 3.5a. A similar propagation is performed for tips from the read/reference that could be caused by mutations in the first and/or last k base pairs. Figure 3.4 depicts the traversal of bulges and tips, and their subsequent color propagation. The full color profile, after color propagation, is shown at the top of the figure, aligned with the arcs of the read.

Using the colored antibody graph, we label each read's V, D, and J gene-segments for repertoire analysis. Figure 3.5(b) depicts a single read, shown in black, with multiple colored reference gene-segments sharing some subsequences. A single read, can be traversed to create the color profile for that read. This profile consists of all the colors that paint the path of the read, i.e., all the reference gene-segments that share some k-mer with the read. Figure 3.5(c) shows the $9 \times n$ color profile matrix C for the example represented by the nine (3 V, 3 D, and 3 J) reference gene-segments, and n positions. From this color profile matrix, we can select the top m scoring gene-segments for each V, D, and J gene-segment set. Scoring each row of this matrix, by a variety of scoring schemes described below, will allow us to select the top gene-segments.

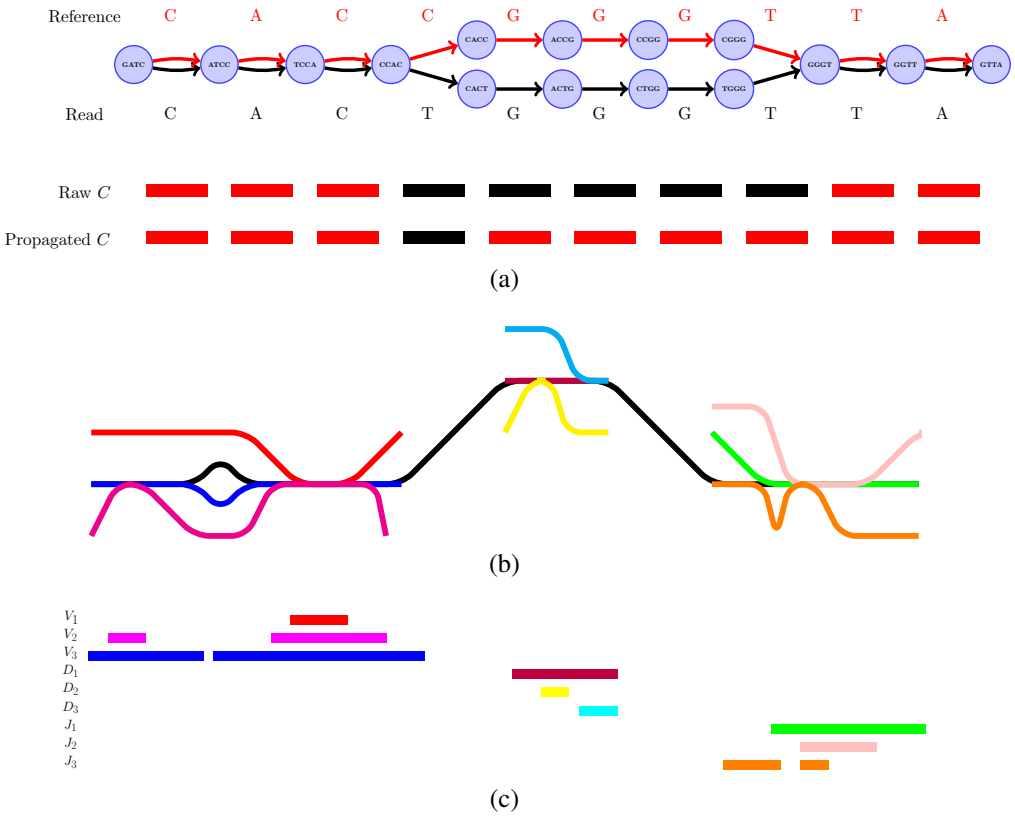


Figure 3.5: Color propagation and colored antibody graph with single read. (a) Color propagation example. Two sequences with a single nucleotide difference between them; GATCCAC**T**GGGTTA (read shown by black edges) and GATCCAC**C**GGGTTA (reference shown by red edges). The de Bruijn graph in this example is created with $k = 5$. Edges shared between the two sequences are colored red and black. A single nucleotide difference creates 5 mismatches in the color profile of this read, shown as the 'Raw' *C*. IgGraph traverses this bulge and propagates the color to reduce the number of mismatches to the single nucleotide difference, shown as 'Propagated' *C*. (b) A single read (shown in black) along with V, D, and J gene-segments shown as different colors. Shared *k*-mers between the read and different gene-segments are shown as merged paths, while divergences are shown as bulges and tips. (c) The $9 \times n$ color profile matrix for the example is shown. Each row represents one of nine gene-segments, and each column is a different position in the read. From this matrix, we can score each row to select the V, D, and J labels for the read.

Scoring the color profile

To utilize the color profile C, a scoring scheme must be defined. A simple scoring scheme with match and mismatch values can be used for the D and J gene-segments, as they exhibit far fewer mutations. In this simple case, the most popular color can be selected as the reference label. The V gene-segments frequently contain many mutations, some having known associated motifs [RK92, DFFL98, CGPvV06]. Rogozin and Kolchanov, 1992 [RK92] first exposed the RGYW motif, and Doerner et. al.,1998 [DFFL98] showed the inverse motif, WRCY also promotes mutations. As a result, the simple scoring does not leverage this additional information and thus does not perform well on V gene-segments. However, this information can be easily incorporated into the model to improve gene-segment labeling. Mutations in the V gene-segments are known to be positionally dependent [CGPvV06], with fewer occurring in framework regions, and more in CDR regions. This is incorporated with discovered 4-mer motifs into a probabilistic score. At each position in the scoring matrix, i, there is an event of either a mutation or a match. There is an associated l-mer b_i and a read position p_i. From these, the probability of an event $m \in \{\text{match}, \text{mutation}\}$ is $P(m|b_i, p_i)$. We compute the probability of the read r, given each reference $R \in \mathcal{V}$, with each reference being equally likely as $P(r|R) = \Pi_{\text{all positions } i \text{ in the reference}} P(m|b_i, p_i)$.

The computation of $P(r|R)$ can be performed over a row R of color profile C, $C[R]$. Each column i of $C[R][i]$ provides us with positional information, p_i, and its surrounding sequence context. In the uncommon cases when bulge/tip color propagation is unable to resolve differences in the sequences, we must assume that all differences arise from mutations without any reference sequence context. This is computed for all references in the V gene-segment set $\mathcal{V}$.

The probabilities for mutation and matching events are computed from 23,051 human IMGT annotated sequences, resulting in 67,108 mutation events and 1,487,059 matching events. Any events that include an indel from the alignment of read to reference are discarded. Once probabilities for each reference (i.e., color) are computed, a rank score is associated with each

color. The top ranked colors, cumulatively comprising a certainty cutoff, are all awarded a tie for top rank. Each other color is assigned the rank of its probability; only the top ranked colors are returned..

3.2.2 Results

Datasets

In order to test the labeling performance of the IgGraph, two approaches were utilized: simulating datasets of smAbs with varying levels of divergence, and testing on three Ig-seq datasets. Comparison on simulated datasets is deemed as supervised since ground truth labels are known. Comparison on Ig-seq datasets is computed on similarity of predictions by different tools since ground truth cannot be known; i.e., unsupervised evaluation.

Obtaining true labels for real data, (like the Stanford_S22 dataset) is difficult and error prone. We thus include Stanford_S22 dataset as an example of real Ig-seq data, all of which are shown in Table 3.1, and compare predictions on it in an unsupervised manner.

While the datasets of real Ig-seq data are invaluable, they are likely to be biased in favor of certain V/D/J gene-segments selected for by the immune system (Supplementary Figure A.5). This bias is not a desirable property when benchmarking a tool. Rather, we wish to test performance on all combinations of gene-segments, so an ideal dataset will have a uniform distribution of VDJ usage (Supplementary Figure A.7). The simulated dataset was generated by using V, D, and J gene-segments from human reference gene-segments, using the method described in Figure A.1. The distributions of exonuclease chewback, nucleotide additions, and V(D)J combinations, are represented across the datasets. Furthermore, each dataset included a fixed number of mutations per smAb, testing the ability to perform VDJ classification at varying degrees of divergence from the reference gene-segment. Considering a single read, it can be labeled by one, or more, reference gene-segments. Ideally, only a single segment should be returned. However, there are occasions when exonuclease chewback makes unique

identification infeasible. We select the maximum number of gene-segments to return, above which we return no label (Supplementary Figures A.8 and A.9).

Table 3.1: Table of datasets used for benchmarking. Simulated datasets are evaluated in a supervised manner, and real datasets are compared in an unsupervised manner.

Dataset	sequencing	# unique entries	size (MB)
Simulated Ig	simulated	2 000	1.1MB
Stanford_S22 [JBGC10]	Roche 454	13 153	3.4MB
Mouse Ig-seq [HGH$^+$14]	Illumina MiSeq	204 462	80.0MB
Human Ig-seq [SBK$^+$15]	Illumina MiSeq	3 099 967	1 173.0MB

We attempted to benchmark as many tools as possible, and while many exist [GMJ$^+$07, BLG08, YMMO13, WWZ$^+$08, SCLR$^+$04, VCK06, OLNLB06], few are available for download, and only IgBlast is able to be run on the large number of sequences produced by current Ig-seq experiments. This is likely the cause for why so many analyses of Ig-seq experiments produce their own approaches to VDJ classification [WJW$^+$09, JWP$^+$11, ALC$^+$11, WBL$^+$13, HGH$^+$14]. Even for IgBlast, while scaling well to process millions of Ig-seq reads, its output was not immediately usable and a wrapper parser had to be written to convert its output to a more concise format. Other tools only provide a web based interface [BLG08, OLNLB06, SCLR$^+$04] which have varying limitations on the number of sequences; none of which could handle the mouse or human Ig-seq datasets, listed in Table 3.1. The lack of usable, efficient, and standardized tools suggest the potential usefulness of IgGraph for this increasingly used analysis of Ig-seq datasets.

Performance on Ig-seq datasets

In order to compare the predicted classes of various tools, we separate comparisons on labeled data (i.e., simulated data) and unlabeled data (i.e., Ig-seq datasets). Comparisons on labeled data are supervised, and reported as accuracy. Unlabeled data is compared in an unsupervised manner and reported as the Jaccard index over two sets A and B, computed as $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$. Both alleles and genes are compared using the Jaccard index [Jac08],

as are partitions. Clusters based on the junction sequences, as reported by the tools, are compared using the Fowlkes-Mallows index [FM83]. Further explanation of the tool used for this comparison is described in the supplement.

Table 3.2 shows the runtimes of different tools over the datasets. iHMMune was not run on the Stanford_S22 dataset since predictions were provided [JBGC10]. The unsupervised comparison of all pairs of tools is shown in Table 3.3. The full pairwise comparisons for the Stanford_S22 dataset are shown in Figure A.6. Predictions between IgGraph and IgBlast are very similar for J alleles/genes, while slightly less similar for V alleles/genes. The predictions for D alleles/genes are where the two tools diverge the most, but the difference in clone partitioning is very divergent for the Stanford_S22 dataset, but similar for the mouse Ig-seq dataset.

Table 3.2: Table of runtimes for each tool on the datasets tested. iHMMune-align was not run on the Stanford_S22 dataset, but was analyzed using the published predictions. iHMMune-align was not run on the Mouse Ig-seq and Human Ig-seq datasets due to its high estimated run time from its time per entry on the smAb_2k dataset.

Dataset	tool	CPU time (sec)	time per entry (sec)
Simulated Ig	IgGraph	54	0.027
	IgBlast	151	0.075
	iHMMune	3724	1.862
Stanford_S22	IgGraph	191	0.014
	IgBlast	641	0.048
	iHMMune	NA	NA
Mouse Ig-seq	IgGraph	10311	0.050
	IgBlast	20114	0.098
Human Ig-seq	IgGraph	99813	0.032
	IgBlast	367545	0.118

Performance on simulated datasets

To evaluate the performance of IgGraph in the case when somatic hypermutations (SHM) are prevalent, we generated simulated datasets of smAbs with increasing numbers of mutations, ranging from zero up to 30. In these datasets, a mutation is a change to a non-

Table 3.3: Table of Ig-seq datasets showing pairwise comparision using unsupervised evaluation criteria. Criteria for allele and gene levels is Jaccard index, while Fowlkes-Mallows is used to compare the clone clusterings.

			Alleles					Genes			
Dataset	Tools		IGHV	IGHD	IGHJ	Total	Clone cluster	IGHV	IGHD	IGHJ	Total
	IgBlast - IgGraph		0.944	0.824	0.983	0.774	0.153	0.960	0.824	0.983	0.787
Stanford_S22	IgBlast - iHMMune		0.739	0.878	0.921	0.696	-	0.903	0.889	0.923	0.862
	iHMMune - IgGraph		0.814	0.771	0.921	0.674	-	0.913	0.781	0.923	0.766
Mouse Ig-seq	IgBlast - IgGraph		0.948	0.426	0.947	0.426	0.997	0.948	0.426	0.947	0.426
Human Ig-seq	IgBlast - IgGraph		0.936	0.583	0.951	0.526	0.563	0.945	0.594	0.954	0.541

germline nucleotide with uniform probability, it is not meant to simulate true motifs found within antibodies. Mutations were selected only along the V gene-segment, as sampled from our mutation distribution obtained from IMGT data. Figure 3.6(a) shows the V gene-segment performances of each mutation dataset with an even number of mutations (datasets with an odd number are used for parameter selection, Supplementary Figure A.9) when the divergence from the germline increases. The difference in performance between these parameterizations with varying k-mer sizes correlates with the complexity of the canonical antibody graph shown in Figure 3.2. With more nodes shared with using the smaller k, reconstruction is more difficult. The green curve shows the performance of IgBlast on these same datasets when run with default parameters. While IgGraph with $k = 21$ and $k = 15$ outperform IgBlast, the $k = 11$ parameterization does underperform when reference divergence is increased.

One option is to provide multiple values of k for the different V, D, and J gene-segments; since larger values for k perform better for V and J, while smaller values of k are required for recovering many D gene-segments. This can be done by creating the graph for each gene-segment type, and one or more reads, as described previously. The resulting accuracies, for pairs of values of k for V/J and D, on a simulated dataset are shown in Figure A.12.

VDJ partitioning comparison

Partitioning an input read into the germline gene-segments is a useful output for VDJ classification. To adequately compare the similarity in partitioning between the tools, a dataset of 7,532 antibody sequences was downloaded from the IMGT database. This approach of using

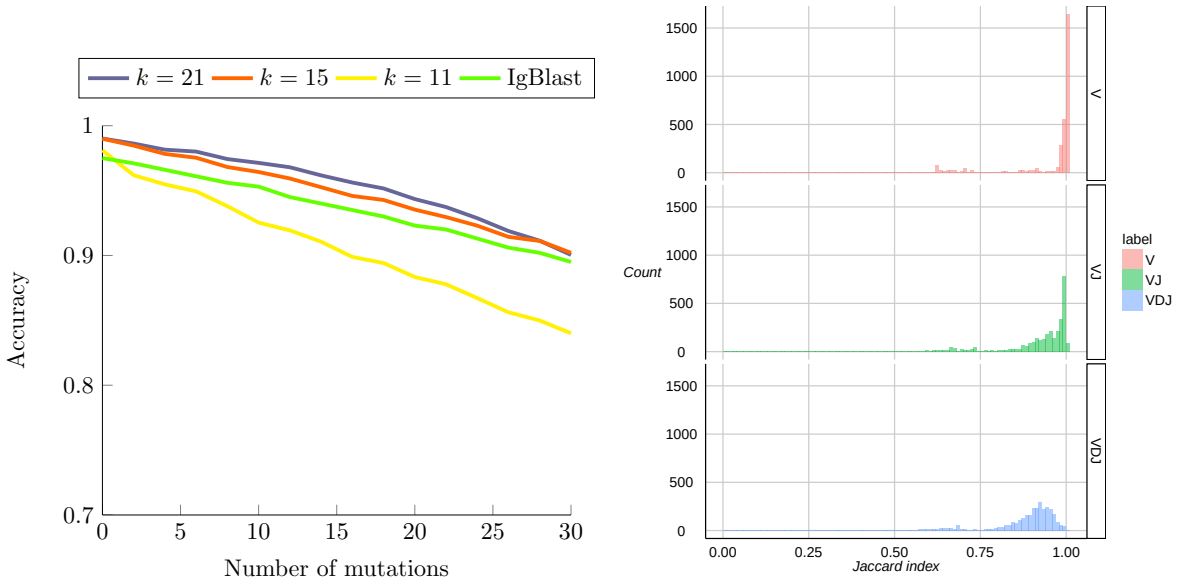


Figure 3.6: Labeling and partitioning comparison. (a) shows the accuracy of IgGraph for V gene-segments when a fixed number of mutations are inserted in each smAb V gene-segment. Only datasets with an even number of mutations are plotted. The blue, orange, and yellow curves represent IgGraph results with parameterizations of $k = 21$, $k = 15$, and $k = 11$, respectively. The green curve represents the IgBlast tool run with default parameters. (b) Jaccard index over partitions. The similarity of the partitioning for range sets of V, VJ, and VDJ gene-segments are measured by computing the Jaccard index for predictions from IgGraph and IgBlast, for each sequence.

a collection of unlabeled, experimentally derived, sequences, for comparison was employed in previous approaches [GMJ+07, YMMO13]. This set was selected by collecting all fully annotated, human heavy chain antibody sequences in the IMGT database whose length ranged from 350bp to 500bp.

Figure 3.6(b) shows the similarities in partitioning between IgBlast and IgGraph as the Jaccard index between the partitioning ranges considered. For each tool, each range of positions for V, D, and J is considered a set, and the Jaccard index over two sets A and B is computed, $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$. While there are differences between the two tools where the partitions are drawn, they are largely similar. As the difficulty in labeling gene-segments increases, so too do the differences between the reported partitions.

3.2.3 Conclusion

We presented a new IgGraph approach to VDJ gene-segment labeling for immunoglobulin transcripts. Our colored antibody graph departs from the alignment based methods (IMGT, SoDA, IgBlast, and others) and HMM-based methods (iHMMune-align). Recently, colored de Bruijn graphs have been used to identify genomic variants [ICT$^+$12], we repurpose and extend this idea to identify immunoglobulin gene-segments. Further, our approach utilizes a scoring model for V gene-segments that considers mutation motifs and position dependence, something that many other tools do not model. iHMMune-align is one of the few that explicitly model known mutation motifs, however they do so in a static fashion. Our scoring is based on probabilities learned from IMGT data, discovering known, and potentially novel, mutation motifs.

We have shown that our approach performs well on simulated datasets and on real Ig-seq datasets. While this approach performs well, it does have its limitations, namely the reliance on sufficiently large k-mers. This is of concern particularly for small D gene-segments, as there must be some k-mers that match on these segments that have been shortened by exonuclease chewback. However, selecting too small a value for k to ensure coverage on D gene-segments can create an overly complicated graph, potentially connecting k-mers in V gene-segments to those in J gene-segments. While we do not observe any significant reductions in performance in either our simulated datasets or real ones due to this, this can limit the potential applications, namely to T-cell receptors (TCR). TCRs share the same V/D/J structure as immunoglobulins, but in human have only two D gene-segments, 12bp and 16bp long. While some approaches may claim to recover these D gene-segments, our colored antibody graph will likely be unable to as long as exonuclease chewback sufficiently reduces its length.

3.3 Acknowledgements

Chapter 3 is published in full as **S.R. Bonissone** and P.A. Pevzner. Immunoglobulin classification using the colored antibody graph. In Research in Computational Molecular Biology, volume 9029 of Lecture Notes in Computer Science, pages 44-59. Springer International Publishing, 2015 . The dissertation author was the primary author of this paper.

Chapter 4

Immunoproteogenomics

4.1 Antibody repertoire construction and immunoproteogenomics analysis

4.1.1 Methods

Antibody repertoires

If we view an antibody as a center of a cluster formed by reads derived from this antibody, then construction of a repertoire corresponds to a difficult clustering problem with many closely located centers so that the radius of a cluster may exceed the distance from one cluster to another one. Since the standard clustering techniques (like k-means clustering) are not applicable to such problems ([PEP04]), we have designed IGREPERTOIRECONSTRUCTOR, a novel algorithm for constructing antibody repertoires.

Each antibody in an antibody repertoire is characterized by its sequence and abundance; estimated by the number of reads derived from this antibody (Figure 4.1b). The complexity of the antibody repertoire mirrors the complexity of the immune system, e.g., clonal selection leads to a highly uneven distribution of abundances of antibodies ([Bur76]). Abundant antibodies mutate and yield new antibodies that share the same VDJ recombination pattern, but differ only

by somatic hypermutations. As a result, the antibody repertoire contains a mixture of closely related antibodies with differing abundances. The abundances of ≈ 2.3 million antibodies in our dataset vary from 1 to $\approx 33,000$ with the most abundant antibody representing $\approx 1\%$ of all reads.

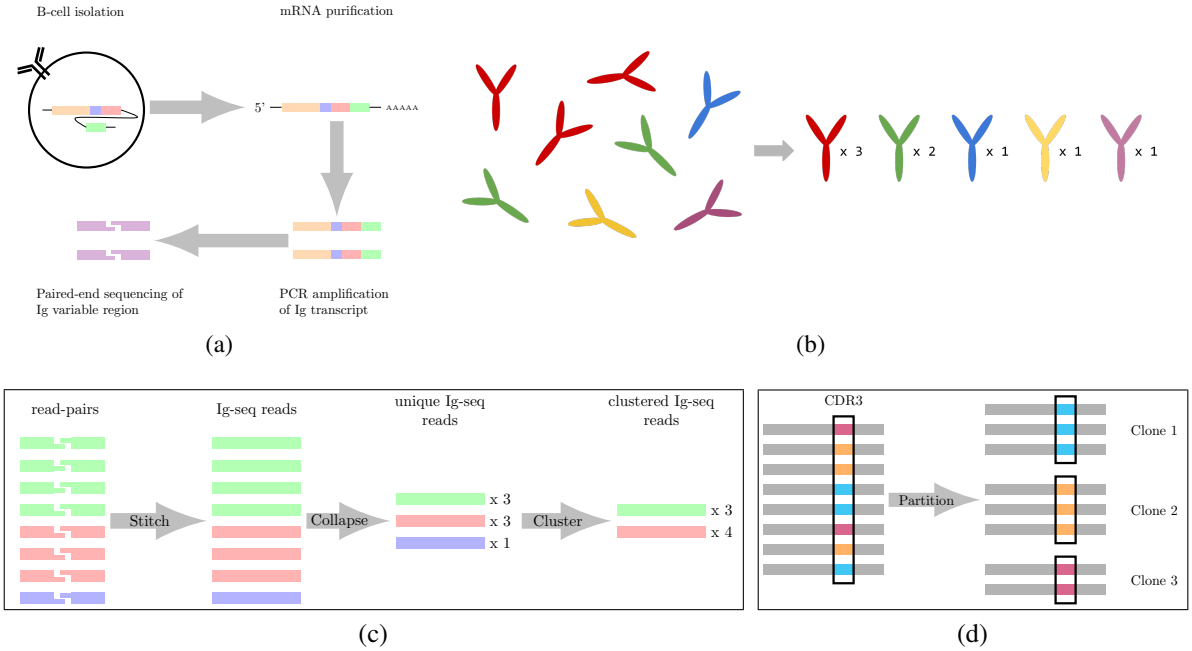


Figure 4.1: (a) An overview of immunoglobulin (Ig-seq) sequencing. Briefly, B-cells are isolated; transcripts are purified; antibody heavy (or light) chain are amplified by PCR; and finally paired-end sequencing of the Ig variable region is performed on the amplified Ig transcript molecules. (b) An antibody repertoire containing five different antibodies (shown on the left) is characterized by a set of pairs <sequence, abundance> (shown on the right). For example, the abundance of the "red" antibody is 3. (c) The varying levels of sequence information. First, the paired reads are stitched together to form contiguous reads. These reads are then compressed to unique reads with count information, and finally clustered reads. E.g., the red and blue unique reads (with counts 3 and 1) are clustered into a single cluster with count 4 because they represent reads (with errors) derived from the same antibody. (d) Reads are partitioned according to identical CDR3 sequences (shown in the black rectangles). Each resulting cluster of antibodies is referred to as a *clone*.

The major challenge in constructing antibody repertoires is the identification of all reads that are derived from a single antibody. If reads were error-free, we would simply group together reads that are identical (up to small shifts) into *unique reads* to generate an antibody repertoire. In reality, reads are error-prone necessitating *error-correction* of reads prior to any

analysis. IGREPERTOIRECONSTRUCTOR error-corrects reads; partitions them into clusters; and computes the consensus sequence and abundance of each antibody. Figure 4.1c depicts the different levels of clustering performed by IGREPERTOIRECONSTRUCTOR. First, we computationally stitch paired-end reads of Ig molecules to derive the contiguous *Ig-seq reads*. Subsequently, the Ig-seq reads are grouped together to provide *unique Ig-seq reads*. Finally, the unique reads are clustered to obtain *clustered* Ig-seq reads (antibodies). In addition, we can represent antibodies according to the somatically recombined B-cell from which they originate, i.e., their *clonality*. We define an *antibody clone* as the set of all antibodies in the repertoire with the same CDR3 sequence (as determined by IgBlast ([YMMO13]). Figure 4.1d diagrams this clone identification process. A clone is trivial if it consists of a single cluster and non-trivial otherwise. The sharp distribution of clone sizes (Figure A1) can be attributed to B-cell response to an antigen, i.e., clonal selection ([WJW$^+$09]).

Limitations of existing error correction tools

At first glance, it appears that the problem of error-correction in immunosequencing is not unlike the problem of error-correction in genome assembly ([PTW01]). However, popular error-correction tools (e.g., Quake ([KSS10]) or BayesHammer ([NKA13]), that were optimized for genome assembly, are not suited for immunosequencing data. Indeed, Ig-seq data contain a large number of sequences differing by very few mismatches or indels, and feature the extremely uneven coverage of various antibodies by reads (since abundances of antibodies differ by orders of magnitudes). Both Quake and BayesHammer start by identifying *solid* k-mers in reads (that are likely to be present in the genome) and use them to correct reads. However, to find solid k-mers, Quake uses the read coverage, an approach that is not applicable in the case of immunosequencing with highly variable abundances. In contrast, BayesHammer (part of the SPAdes assembler, [BNA$^+$12]) was designed to assemble single cell sequencing data with uneven coverage. However, it is also not applicable to immunosequencing since an antibody repertoire yields numerous similar, correct, k-mers from different antibodies

(Figure A2). BayesHammer is unable to distinguish these correct k-mers from similar incorrect k-mers derived from the same antibody.

Hamming graph for analyzing immunosequencing data

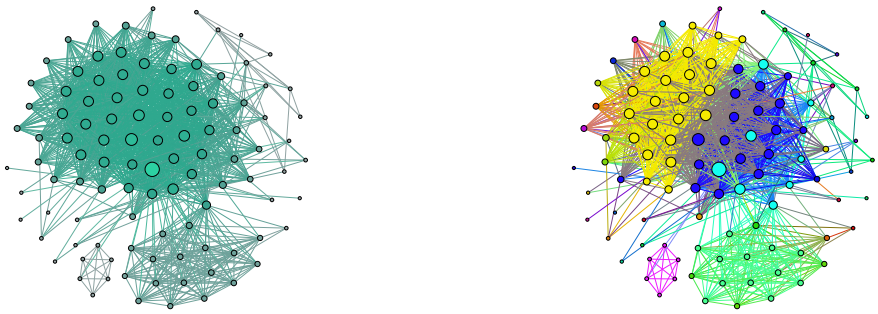
IGREPERTOIRECONSTRUCTOR uses the idea of the *Hamming graph* for error correction ([MSKP11, NKA13]) to correct Illumina reads. With the emergence of longer (250-nt) Illumina reads in 2013 (until 2013, Illumina technology generated shorter reads that did not fully cover the variable region of antibodies), it is now possible to interrogate repertoires using accurate high-throughput Illumina technology. Our focus in this paper is repertoire reconstruction using Illumina technology rather than the less accurate and lower throughput 454 technology that dominated previous immunoproteogenomics studies. The *Hamming distance* $d(s_1, s_2)$ between sequences s_1 and s_2 of equal length is defined as the number of positions where the symbol in s_1 differs from a symbol in s_2 (Figure A3a). We extend the concept of Hamming distance to any two sequences (including sequences with different lengths) by considering all sufficiently long overlaps between sequences s_1 and s_2 (longer than the default value δ), and computing the Hamming distance between the overlapping parts. We define $\tilde{d}(s_1, s_2)$ as the minimum of such distances (Figure A3b). We define the *Hamming Graph* $HG(Strings)$ as the complete weighted graph whose vertices correspond to a collection of sequences *Strings* and the weight of the edge (s_1, s_2) is equal to $\tilde{HD}(s_1, s_2)$. The *Bounded Hamming Graph*, denoted $HG(Strings, \tau)$, is a subgraph of the Hamming Graph where edge (s_1, s_2) exists iff $\tilde{d}(s_1, s_2) \leq \tau$. The time- and space-efficient construction of large Hamming Graphs is a challenging problem that was addressed in [NKA13] and adapted in IGREPERTOIRECONSTRUCTOR. Note that compared to Hammer ([MSKP11]) and BayesHammer ([NKA13]), we construct the Bounded Hamming Graph on the entire reads (rather than on k-mers) and use the generalized Hamming distance.

Repertoire construction and search for dense subgraphs

We construct an antibody repertoire by partitioning reads into clusters that correspond to the same antibody. Our goal is to place reads differing by sequencing errors into the same cluster, while placing reads corresponding to different antibodies into different clusters. This becomes difficult since the number of errors in a read from a given cluster may be larger than the number of differences between antibodies from different clusters. We define the antibody sequence as the consensus of reads in a cluster, and its *abundance* as the number of reads in a cluster.

Since Illumina reads have a small indel rate, the generalized Hamming distance between reads from the same cluster should be low. We thus construct the Bounded Hamming Graph $HG(Reads, \tau)$ from all reads. Our analysis revealed that the generalized Hamming distance for most Ig-seq reads, originating from the same antibody, does not exceed 3 and that many antibodies form complete, or nearly complete, subgraphs of the Bounded Hamming Graph with $\tau = 3$ (see Appendix A). Ideally, we would like to choose τ in such a way that $HG(Reads, \tau)$ is a *clique graph*, i.e., a graph where each connected component is a complete subgraph (*clique*).

In reality, the large connected components of the Bounded Hamming Graph often have a more complex structure. Given a connected component with m edges and n vertices, we define its *edge fill-in* as the ratio of the number its edges (m) to the maximal possible number of edges in the graph on n vertices ($n \cdot (n-1)/2$). Fig. 4.2a presents a connected component of the Bounded Hamming graph with edge fill-in 0.25 ($\tau = 3$). The lion's share of large connected components in the Bounded Hamming Graph (i.e., components with more than 100 vertices) have similar structures characterized by small edge fill-ins; the average edge fill-in for large components is 0.32 (Fig. A4). Additional analysis of the connected components reveals that the nearly all of them (98.6%) consist of *dense* (complete or nearly complete) subgraphs connected by very few edges. Most vertices in these dense subgraphs correspond to error-prone reads derived from a single antibody or from highly similar antibodies differing from each other by a small number of somatic hypermutations.



(a) (b)

Figure 4.2: (a) A connected component with 107 vertices and 1426 edges in the Bounded Hamming graph with $\tau = 3$ (fill-in is 0.25). The sizes of vertices are proportional to their degrees. (b) Clusters constructed as result of vertex decomposition of the Bounded Hamming Graph. Vertices of the same colors define the dense subgraphs in the decomposition (the colors are coordinated with Figure 4.3 (bottom right)). IGREPERTOIRECONSTRUCTOR constructs 42 clusters but 35 of them are trivial, i.e., are induced by a single read. Sizes and edge fill-ins (in brackets) of the remaining 7 non-trivial clusters are: 2 (1.0), 3 (1.0), 6 (1.0), 8 (1.0), 12 (1.0), 18 (0.9), and 23 (0.9).

Thus, the first step in constructing an antibody repertoire is solving a very large instance of the Corrupted Cliques Problem: finding the smallest number of additions and removals of edges that transform the Bounded Hamming Graph into a clique graph. While there exist a number of algorithms for solving the Corrupted Cliques Problem (such as CAST [BDSY99]), they are too slow for the Bounded Hamming Graphs with many vertices. We thus developed a different approach for analyzing the Bounded Hamming Graph that is based on transforming it into a *triangulated graph* (i.e., a graph where every cycle of length longer than three has a *chord*) rather than a clique graph using the Minimum Fill-in Problem ([GJ79]).

Repertoire construction and Minimum Fill-in Problem

The *minimum fill-in edge-set* for a graph is the edge-set of minimal size whose addition turns this graph into a triangulated graph. We are interested in triangulated graphs because maximal cliques in these graphs can be generated in polynomial time ([GHP95]) and because maximal cliques in the triangulated Bounded Hamming Graph help to reveal dense subgraphs of the original Bounded Hamming Graph (see Appendix B for details).

While the Minimum Fill-in Problem is NP-complete ([Yan81]), there exist efficient approximation algorithms for solving it, e.g., METIS algorithm ([KK99]). The METIS algorithm is based on the equivalence between triangulated graphs and *perfect elimination orderings*. A perfect elimination ordering in a graph is such an ordering of its vertices that, for each vertex v, v and the vertices following it in the order, form a clique, an example shown in Figure A11d. A graph is triangulated if and only if it has a perfect elimination ordering ([RTL76]). METIS finds an ordering that generates an approximation of a minimum fill-in edge set and this ordering can be used for finding cliques in the triangulated graph ([GHP95]). As we mentioned above, these cliques correspond to dense subgraphs in the original graph. To construct maximal dense subgraphs, we additionally merge subgraph connected by many edges.

IGREPERTOIRECONSTRUCTOR solves the Minimum Fill-in Problem in the Bounded Hamming Graph using METIS and converts its solution into a list of dense subgraphs in the original Bounded Hamming Graph. Note that some of the resulting dense subgraphs may share vertices forcing us to assign these shared vertices to one of the dense subgraphs. To assign a vertex v to a single dense subgraph, we select a subgraph with maximum number of vertices adjacent to v. Thus, dense subgraphs generated by METIS provide us with a vertex decomposition of the Bounded Hamming Graph. A vertex decomposition of the graph in Fig. 4.2a is shown in Fig. 4.3 (top right). Analysis of all found subgraphs in the decomposition of the Bounded Hamming Graph reveals that the lion's share of them have high edge fill-ins (the average edge fill-in is 0.94), thus confirming that IGREPERTOIRECONSTRUCTOR indeed finds dense subgraphs of the Bounded Hamming Graph. The histogram of edge fill-in for all subgraphs in this decomposition is shown in Fig. A5.

Dense subgraphs correspond to clusters of Ig-seq reads representing either identical or very similar antibodies (i.e., antibodies differing by very few substitutions). However, to construct the antibody repertoire, we need to further partition some of the dense subgraphs (that correspond to multiple antibodies) into subgraphs corresponding to single antibodies. To illustrate this challenge, consider the *SHM-triggering patterns* RGYM/WRCY ([RK92]) and

define an edge in the Bounded Hamming Graph as an *SHM-edge* if at least one mismatch on this edge conforms to the RGYM/WRCY motif. Fig. 4.3 (bottom left) shows the Bounded Hamming Graph where the SHM-edges are highlighted in orange. This coloring reveals that the yellow dense subgraph in Fig. 4.3 (upper right)) corresponds to two similar antibodies rather than to a single one (Fig. 4.3 (bottom right)). Indeed, the multiple alignment of reads corresponding to the yellow subgraph shows a mismatch in the CDR1 region which separates reads into two groups (right panel of Fig. 4.3). Thus, we need to develop an algorithm for splitting constructed dense subgraphs by detection of SHMs. The final solution is shown in Fig. 4.3 (bottom right) and illustrated in Fig. 4.2b. See Appendix C for more details on splitting dense subgraphs, and Appendices D and E on benchmarking of IGREPERTOIRECONSTRUCTOR on real and simulated antibody Ig-seq datasets.

Immunoproteogenomics search

The previous immunoproteogenomics studies ([CBZ$^+$12, SBP$^+$12, BHW$^+$14]) conducted searches on a database of unique Ig-seq reads that we refer to as *unique reads* database. We argue that a better option is the antibody repertoire database (formed by centers of clusters constructed by IGREPERTOIRECONSTRUCTOR) as it eliminates many sequencing errors. Furthermore, to assess the divergence from reference gene-segments, a dataset of canonical V, D, and J gene-segments was searched, this database is termed the *canonical VDJ* database. In order to obtain peptide identifications from the constant region, we also included all 41 intact variants of the constant region, obtained from the IMGT repository ([LGG$^+$09]). These sequences are concatenated to the VDJ database.

Proteomic searches were conducted using MS-GF+ ([KGP08, KP14]) on partially digested peptides (e.g., for trypsin, semi-tryptic peptides were considered). The false discovery rate (FDR) was controlled by selecting a MS-GF+ threshold of spectral probabilities such that we maintained a 1% FDR. Figure A6 shows the distribution of spectral probabilities for the target and decoy datasets.

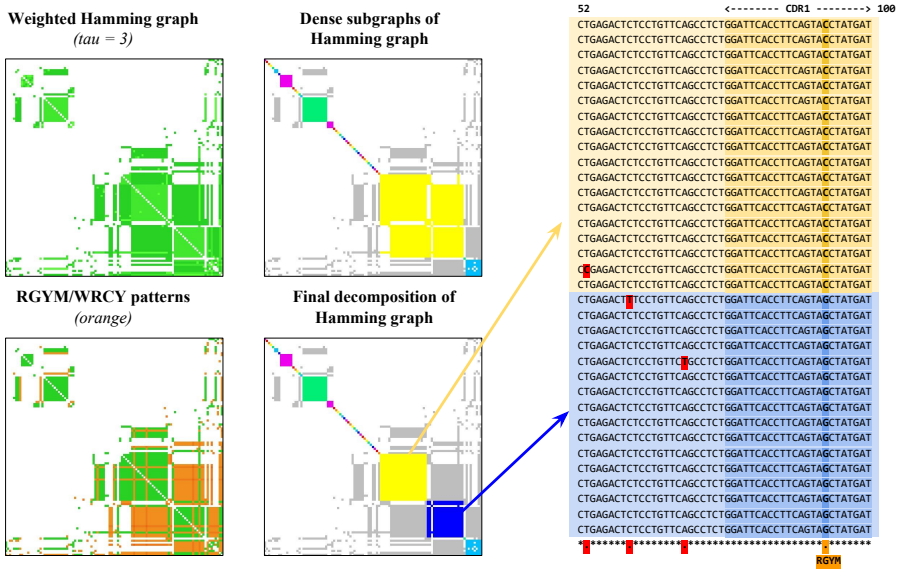


Figure 4.3: Construction of the antibody repertoire based on the decomposition of the Bounded Hamming Graph into dense subgraphs. (Top left) The adjacency matrix of the Bounded Hamming Graph shown in Fig. 4.2a. Each element in the matrix corresponds to a pair of vertices x and y and is colored green if the edge (x,y) is presented in the graph. (Top right) Decomposition of the Bounded Hamming Graph into dense subgraphs (highlighted by different colors). Edges connecting vertices from different dense subgraph are colored in grey. (Bottom left) The adjacency matrix with edges corresponding to SHM-triggering patterns RGYM/WRCY highlighted in orange. This coloring reveals that that the big yellow subgraph in the upper right matrix consists of two subgraphs that differ from each other by SHMs. (Bottom right) The final decomposition of the Bounded Hamming Graph takes into account the multiple alignment of reads corresponding to the same subgraph in the decomposition and breaks the large yellow subgraph (top right sub-figure) into two smaller subgraphs highlighted in yellow and blue. The multiple alignment of "yellow" and "blue" reads from these smaller subgraphs is shown on the right (limited to positions 52–100). Note that all "yellow" reads are similar to each other and all "blue" reads are similar to each other (the differences are highlighted in red and likely represent sequencing errors). However, there exists a systematic difference (C/G mismatch within RGYM pattern in CDR1 region) between "yellow" and "blue" reads that allows IgRepertoireConstructor to split the large yellow subgraph in top right subfigure into two smaller ones.

Blind modification searches were performed using MODa ([NBP12]), allowing for a single modification with mass between -200Da and 200Da. Peptides with at least one enzymatic end were considered and a 1% FDR was enforced.

As discussed in [BHW$^+$14], immunoproteogenomics searches require new algorithmic and statistical approaches since the standard peptide identification algorithms were not designed

for searches in large and highly repetitive immunoproteogenomics databases. We argue that yet another key difference between the standard and immunoproteogenomics searches is that, in the latter case, after constructing the antibody repertoire, we have information about antibody abundances. Since higher-abundance antibodies are better candidates for spectral searches than lower-abundance antibodies (despite limited correlation between genomics- and proteomics-derived abundances), we partition all antibodies into *layers* according to their abundances. The rationale for such partitioning is that higher-abundance antibodies form much smaller protein databases than lower-abundance antibodies. For example, there are 1564, $10,782$, and $48,564$ antibodies with abundances in the intervals from 100 to $30,000$, from 10 to 99, and from 2 to 9, respectively. This contrasts with $2,267,863$ singleton antibodies with abundance 1. Thus, since E-values of PSMs rapidly deteriorate with the increase in the database size ([GBKP11]), we partition all antibodies into 4 layers (according the abundance intervals specified above) and employ a separate 1% FDR control, for each layer, based on selecting a spectral probability threshold in MS-GF+. Note that our *multi-layer* approach is very different from the *two-stage* MS/MS search approach with logical dependencies between two stages. Since there are no such dependencies in the multi-layer approach, the controversy about the statistical foundations of the two-stage approach ([GBKP11]) does not extend to our multi-layer approach.

4.1.2 Results

Datasets

We have benchmarked IGREPERTOIRECONSTRUCTOR on multiple Mi-Seq and Orbitrap datasets generated at Genentech. Below we only describe the results for a single heavy chain dataset. Similarly to BayesHammer, the running time of IGREPERTOIRECONSTRUCTOR is dominated by the construction of the Bounded Hamming graph ($\approx$ 5 hours for the heavy chain dataset). All further steps (finding and splitting dense subgraphs, etc) take less than 30 minutes on a single thread. Mass-spectrometry searches and subsequent cluster/clone peptide

assignment take ≈ 8 hours to complete.

Ig-seq dataset. The Ig-seq library contains overlapping paired-end reads that cover the variable region of heavy chain (3.83 million 250-nt long reads with average insert size 366 nucleotides). We pre-process the immunosequencing library by merging overlapping paired-end reads, and removing contaminants as described in Appendix F. After pre-processing, IGREPERTOIRECONSTRUCTOR generated $2,925,095$ unique reads, $2,406,121$ clustered reads, and $586,341$ clones. See Appendix G for the analysis of contaminants.

Spectral dataset. We analyzed CID tandem mass-spectra generated using the following digestive enzymes; AspN ($21,385$ spectra), chymotrypsin ($24,956$ spectra), trypsin ($26,740$ spectra), and elastase ($20,604$ spectra). Enzymes with differing cleavage specificity improve coverage over the length of the antibody sequence. We searched spectral datasets against the protein databases derived from the antibody repertoire. 3-frame translations were created for each antibody in the repertoire, and any frames containing a stop codon were discarded; $165,675$ antibodies ($\approx 7\%$) had a stop codon in all frames.

Analysis of antibody repertoires

Below we compare the repertoires formed by unique reads and by IGREPERTOIRECONSTRUCTOR. To compare the repertoires, we used various metrics measuring cluster sizes (*# clusters*, *# singletons* (single-element clusters), *max cluster size*, *# clusters of size exceeding X* (where *X* is a parameter)) as well as metrics based on CDR3 analysis. Table 4.1 illustrates that IGREPERTOIRECONSTRUCTOR generates a rather different (more compact) representation of antibodies than the set of unique reads used in previous immunoproteogenomics studies.

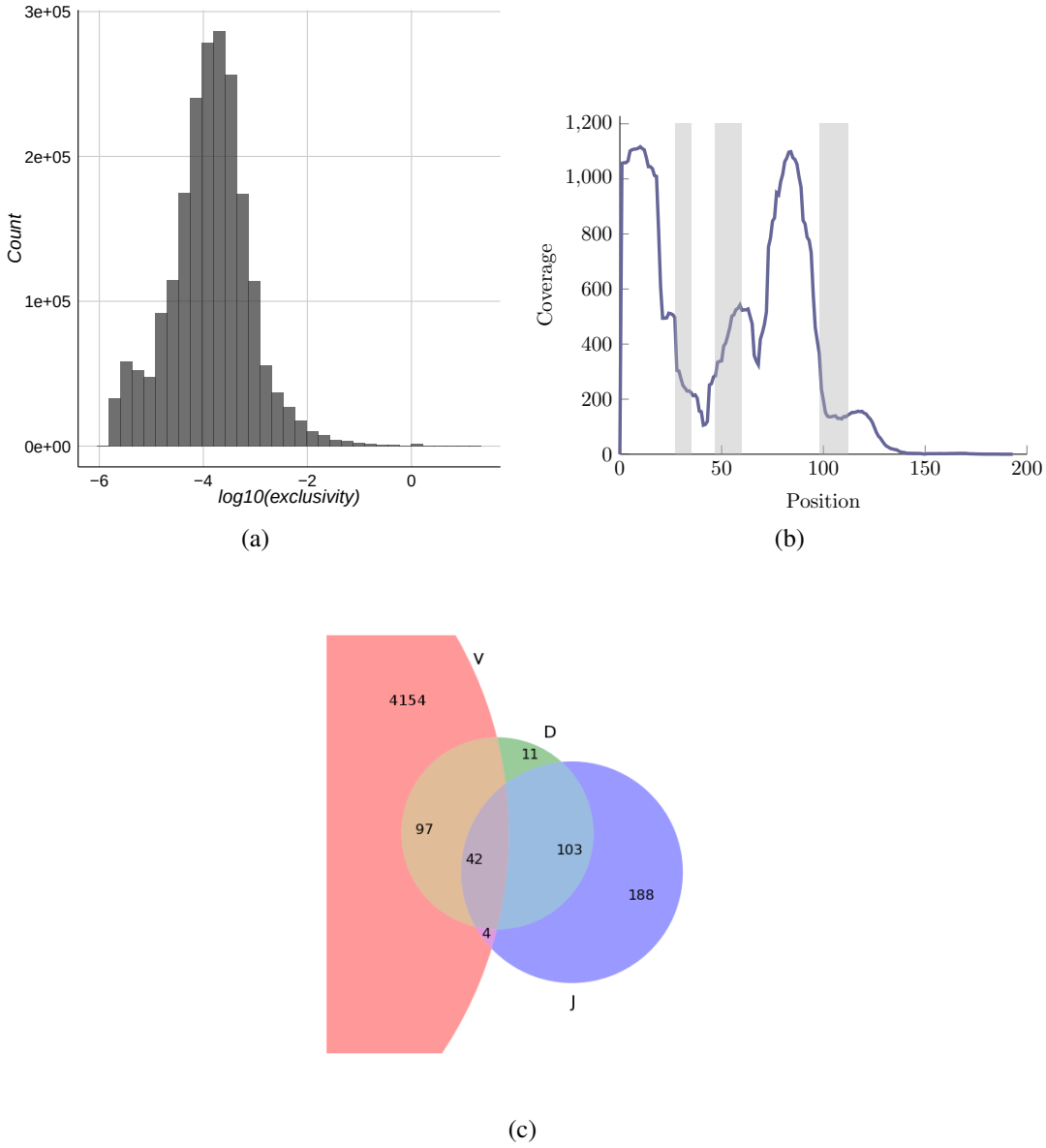


Figure 4.4: (a) Distribution of exclusivity scores of antibodies. (b) PSM coverage along positions of each cluster. Positions of CDR1, CDR2, and CDR3 shown in gray as determined for a single cluster. Coverage is normalized for shared peptides using their exclusivity scores. (c) Origin of identified peptides. For each identified peptide, a representative cluster sequence was used to determine from which reference segment it originated; V, D, or J. Each peptide is classified as V-, D-, or J-peptide depending on whether it overlaps with segments marked as V, D, or J regions for the heavy chain sequence (peptides spanning more than one region, e.g., V and J, are classified as both V-peptides and J-peptides).

Table 4.1: Comparison of the antibody repertoire generated by IGREPERTOIRECONSTRUCTOR with the set of unique reads (heavy chain immunosequencing data). The *avg clone divergence* metric is computed as the fraction of the number of columns in the multiple alignment of all antibodies in a clone that have mutations or indels. The *avg non-trivial clone divergence* shows the average clone divergence computed over all non-trivial clones.

	Unique reads	**IGREPERTOIRECONSTRUCTOR**
# clusters	3,099,967	2,328,773
# singletons	3,027,123	2,267,863
max cluster size	2203	33,021
# clusters (> 10)	5532	12,346
# clusters (> 50)	377	3571
# clusters (> 500)	7	206
# clones	602,536	538,928
# non-trivial clones	151,612	132,431
avg non-trivial clone size	15.64	12.90
max clone size	30,571	15,977
avg non-trivial clone divergence	0.21	0.23

Peptide identifications

Table 4.2 shows the number of identified peptides and PSMs. Modified peptides are considered identical to those without modifications should their sequences be the same; and hence are not counted when considering unique peptides. Surprisingly, many peptides from antibodies are captured only in modified form (e.g., we do not identify the unmodified variants of these peptides) implying that previous immunoproteogenomics studies likely missed many peptides. Note the large number of peptides identified only with modifications (i.e., unmodified versions of these peptides were not identified) suggesting that future immunoproteogenomics searches should include search for PTMs. Overall, we identify nearly 13% of all spectra when performing restrictive PTM searches at 1% FDR. The number of identified peptides is further boosted when employing a multi-layer strategy, noted by the "layer" column in the table. Blind modification search was performed on the trypsin dataset only (since MODa is not designed for spectral datasets generated with other digestive enzymes). MODa identified 3334 PSMs with modifications, corresponding to 970 peptide IDs; 815 of which were identified only by the blind modification search. It brings the total percentage of identified spectra to $\approx 22.6\%$ at 1% FDR ([CBZ$^+$12] identified 6% of spectra at 2% FDR). See Appendix H on specific modifications found by our blind search. Figure 4.4c shows the breakdown of the origin of

each identified peptide.

Table 4.2: Peptides and PSMs identified by MS-GF+. The number of peptide identifications at an 1% FDR cutoff for each spectral dataset. For example, a 1% FDR cutoff corresponds to a spectral probability cutoff of 1.4E-08 for AspN, 2.3E-10 for chymotrypsin, 7.5E-10 for trypsin, and 3.8E-10 for elastase datasets, when searching antibodies with the constant region appended, and restrictive PTM search. The total column shows the number of total peptides, or PSMs, across the four different mass-spectrometry datasets. The total % column shows the percentage of identified spectra, among all spectra. The "layer" column denotes the type of search; single layer (s), or multi-layer (m). The restrictive MS-GF+ searches for post-translational modifications (PTMs) were conducted by searching for carbamidomethyl (C+57) as a fixed modification, oxidation of methionine, oxidation (single and double) of tryptophan, and N-terminal pyroglutamate (Q-17, E-18) as optional modifications.

			Peptides					PSMs					
Database	Layer	PTM	AspN	chymo	trypsin	elastase	total	AspN	chymo	trypsin	elastase	total	total %
repertoire	m	✓	814	1706	2505	776	5801	2665	4989	10021	1786	19461	20.77%
repertoire	s	✓	832	1441	2291	628	5192	1881	2365	6675	878	11799	12.59%
repertoire	s		61	636	1756	357	2810	89	896	3753	377	5115	5.46%
constant region	s		279	205	109	107	700	865	583	933	286	2667	2.85%
canonical VDJ	s		25	122	122	69	338	115	441	618	173	1347	1.44%

Assigning peptides to multiple antibodies

Interestingly, only $67,124$ antibodies (2.6% of all antibodies) did not encode any identified peptide. Moreover, as expected, these are mainly antibodies with minimal abundance 1 (total abundance of these $67,124$ antibodies is $73,764$ and maximal abundance is 300). This (surprisingly) low number of antibodies with no peptide evidence is due to the fact that many identified peptides map to multiple antibodies. As a result, the number of antibodies A in a repertoire R encoding a peptide P is often large. We define *exclusivity score of a peptide* as $exclusivity(P,R) = 1/\text{number of antibodies in } R \text{ encoding } P$ and *exclusivity score of an antibody* as $exclusivity(A,R) = \sum_{\text{all peptides } P \text{ mapping to } A} exclusivity(P,R)$. The exclusivity score distribution of the antibodies, seen in Figure 4.4a, shows few antibodies having peptides exclusive to them alone (only 1472 antibodies with $exclusivity(A,R) > 1.0$). Figure 4.4b shows the peptide coverage over the position of each clone. Figure A8 illustrates the peptide coverage of a single clone (the grayed out ranges show CDR3).

Correlation between Ig-seq and MS/MS abundances

To compare the relation between peptides and their Ig-seq counterparts, we introduce the notions of *total Ig-abundance* and *maximal Ig-abundance* of a peptide. Total (maximal) Ig-abundance of a peptide is the total (maximal) abundance of antibodies that encode this peptide. Figure A9a shows the relation of the total Ig-abundance for each peptide to its spectral count (number of PSMs). Figure A9b shows a histogram of spectral peptide counts binned over maximal Ig-abundance for each peptide. A strikingly large number of peptides, 2702, can be attributed to singleton antibodies. The remarkable lack of correlations between genomics-based and proteomics-based abundances further amplifies the concern first expressed in [CBZ$^+$12]. Figure A9c shows the correlations between clone abundances measured by MS/MS and Ig-seq (compare with Figure A7 that measures antibody, rather than clone, abundance). These plots show the difference when considering the unit of a repertoire (antibody), and the unit of antibody evolution (the clone) raising the concern that Ig-seq data do not adequately represent antibody abundances. When considering only the antibodies, there is no correlation with the mass-spectrometry evidence, as previously reported by [CBZ$^+$12]. However, when considering the amalgam of antibodies forming each clone, a moderate correlation emerges ($\rho = 0.5687614$). One possible explanation is that certain antibodies, within highly expressed clones, are not captured by mass-spectrometry.

4.1.3 Conclusion

Our study is the first to validate the constructed antibody repertoires (by using complementary proteomics data) that confirmed that IgRepertoireConstructor generates accurate repertoires. With an accurate tool for constructing antibody repertoires, we can move to studies of evolution of antibody repertoires, the analysis that has not been possible in the past. Since analysis of antibody repertoires is not unlike analysis of repeat subfamilies, the existing algorithms for analyzing repeat evolution ([PEP04, CB09]) can be applied to study evolution of antibodies. We also addressed the problem of peptide identification in large and highly

repetitive databases by designing multi-layer immunoproteogenomics search algorithm. Finally, we revealed an alarming lack of correlation between NGS-based and MS-based quantitation of antibodies (consistent with [CBZ+12]) and proposed a way to partially restore this correlation by considering clone abundances rather than individual antibody abundances.

4.2 Acknowledgements

Chapter 4 is published in full as 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):i53i61, 2015 . The dissertation author was one of the primary authors of this paper.

Chapter 5

Cancer immunoglobulin proteogenomics

5.1 Tumor infiltrating lymphocytes

Cancer is a complex disease, embodied by the heterogeneity of clones at the tumor site. It is further complicated by the fact that the tumor sample is additionally comprised of non-tumor host cells. This tumor microenvironment also harbors a variety of immune cell infiltrates, including: T-cells, macrophages, dendritic cells, and B-cells. Such infiltrates where shown to be found in pan-cancer analysis [RSW$^+$15]. These tumor infiltrating lymphocytes (TILs) represent the adaptive immune systems attempt at targeting and eliminating tumors. Obviously, this system breaks down and can allow for cancers to progress and proliferate.

The following chapter will investigate TILs from tumor samples, specifically the immunoglobulins produced by B-cells. These infiltrating immunoglobulins can provide us with a snapshot of if/when the host can create clonally expanded antibodies, and their potential prognostic value. Additionally, understanding similarities in repertoire response and isotype switching can provide new insights into the immune systems role in this tumor microenvironment.

5.2 Proteogenomic analysis of colorectal cancer reveals mutations and immunoglobulin peptides

5.2.1 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 [KFMS12, BBB$^+$11, MBC$^+$12]. 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) [cpt], 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 [ZWW$^+$14]. The results confirm large differences between protein and transcript expression and underscore the need for robust proteomic technologies, particularly in the identification of 'variant' peptides as translational evidence for genomic events such as mutations, splicing, structural variation, and others. As peptides are typically identified by comparing acquired spectra against theoretical spectra from candidate peptides, a customized database of candidate peptides must be created to include variants observed in genomic tumor samples. The term proteogenomics often refers to the search of mass spectra against these specialized databases [WCM$^+$14, WCN$^+$14, CSH$^+$14, CB10].

Despite many proteogenomic methods having been recently proposed [LDZ10, LSM$^+$11, WSW$^+$12, WZ13, Fea11], serious methodological challenges remain. While the initial goal of the CPTAC [cpt] colon cancer study has been delivered [ZWW$^+$14], the use of more sophisticated approaches can enable additional discoveries from this existing cancer dataset. Most proteogenomic methodologies focus on identifying single amino-acid polymorphisms (SAP) by adding peptides that capture the alternative allele [ZWW$^+$14, LDZ10, LSM$^+$11, WSW$^+$12,

WZ13, Fea11]. 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 [ZWW$^+$14]. In other cases, small transcript data-sets are used to mine junction peptides, without a robust framework for handling available big data-sets of Next Generation Sequencing (NGS) data. For colorectal cancer, the single TCGA [MBC$^+$12] (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 [WCM$^+$14, WCN$^+$14]. 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 [CPS$^+$08, CB10, CSH$^+$14, WCN$^+$14] 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 [NGS03, Nel10], 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 [ONI$^+$09]; 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 [PTW01, ZB08, CPT11]. These customized databases permit the identifications of tumor antibodies, and explore their potential role in the molecular characterization of colorectal cancers. Surprisingly, our result showed that 56.37% of our novel proteogenomic event identifications were from immunoglobulin gene database search. This result underscores the importance of our proposed immunoglobulin peptide search when analyzing cancer samples, adding a new host-immune dimension to our analysis.

The value of proteomic evidence over transcript, or genomic evidence has been debated, with recent reports supporting the complementary information available from proteomic data. Our proteogenomic pipeline maintains summary level information in the transcript derived databases that allow for seamless querying of the relative frequencies of specific variants in DNA/transcript data. Through the reanalysis of 90 distinct colorectal tumors from the CPTAC project, we also addressed questions regarding recurrent somatic mutations in tumor genomes. As the result, we have identified $2.3\times$ more variations compared to the initial CPTAC study [ZWW$^+$14]. Moreover, it should be noted that by applying conventional FDR strategy, $15.3\times$ additional novel identifications were found which includes 96.25% of RNA expressed

mutations from the initial CPTAC colon cancer study [ZWW+14].

5.2.2 Results

The CPTAC colorectal cancer data-set. MS/MS spectra of Adenocarcinoma (COAD) and Rectum Adenocarcinoma (READ) were downloaded from CPTAC data portal [cpt], 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 [BBB+11, KFMS12] 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 [WCM+14, WCN+14]).

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 5.1, we developed a customized proteogenomic database targeted to IG gene peptide identifications.

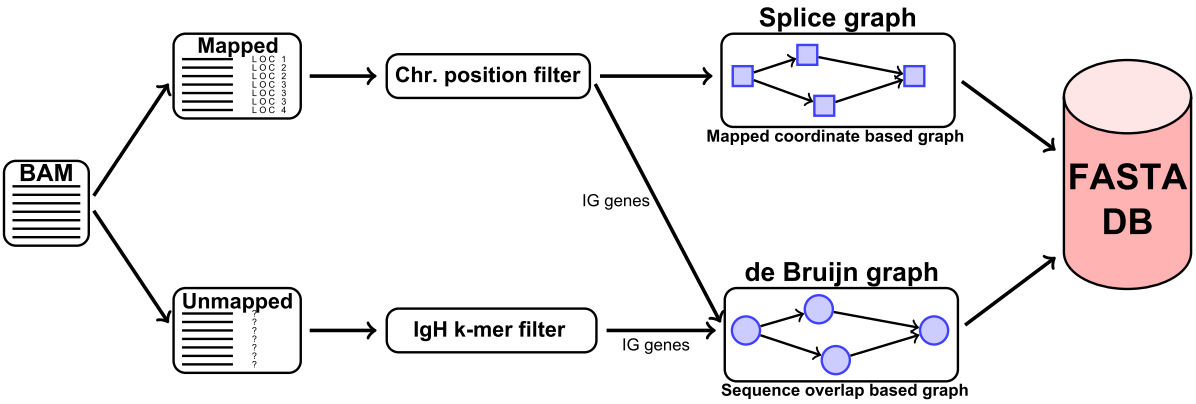


Figure 5.1: Illustration of proteogenomic database construction for immunoglobulin peptide identifications.

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 [CPT11].

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 [WCM+14]. 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 [FAA+13] (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 [EG07] 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

$$\frac{F_D(\tau)}{F_T(\tau)} \leq 0.01 \tag{5.1}$$

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.

$$\frac{F_D(\tau)}{F_T(\tau)} = \frac{F_I(\tau)}{\alpha \cdot F_C(\tau) + (1-\alpha) \cdot F_I(\tau)} \leq 0.01 \tag{5.2}$$

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 [WCN+14] 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+ [KMB+10]. 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. [ZWW+14] who used their own databases (CanProVar), and against a second search-tool using Comet [EJH13] on our specialized databases as a control. The Comet results [EJH13] showed 357 novel peptide identifications while 70.86% of the peptide overlapped with MSGF+[KMB+10] results

(Figure C.4). In generating novel events reported in this paper, we used the union set of both MSGF+ [KMB+10] and Comet [EJH13] 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 5.2), 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 [FBB+11, SWK+01, MBC+12], 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 5.2b).

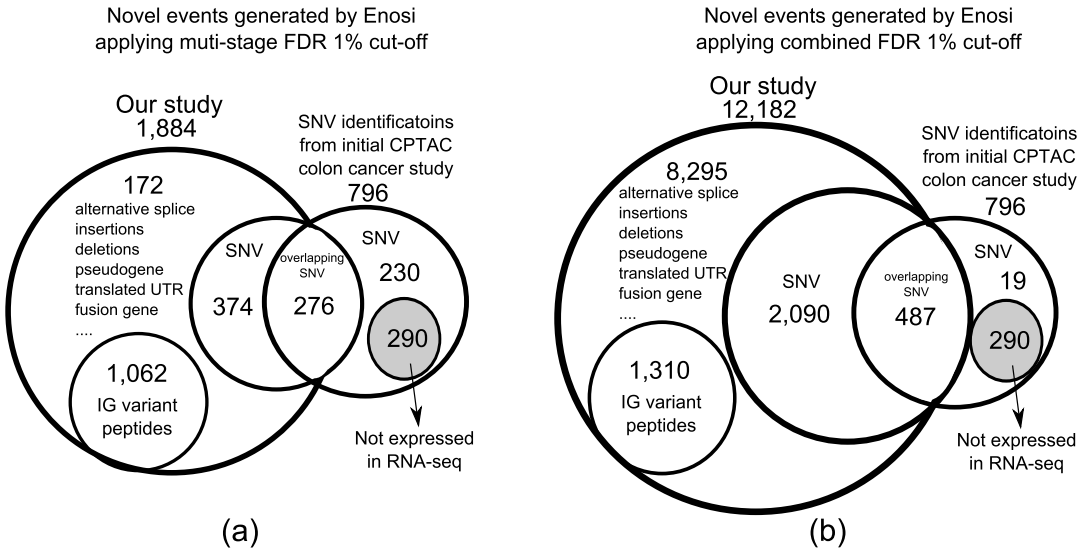


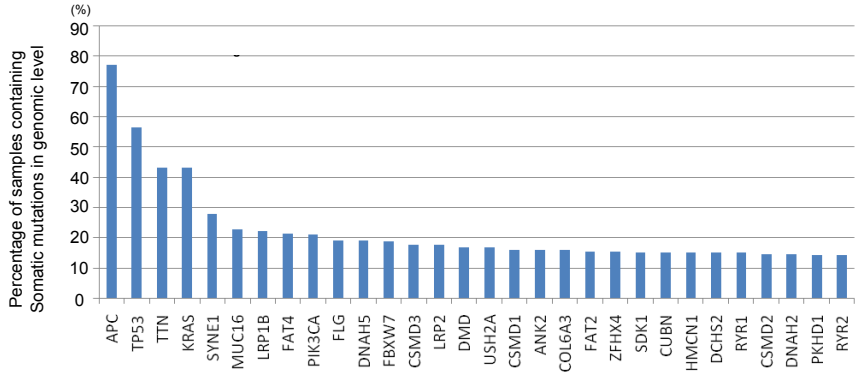
Figure 5.2: (a) Comparison of aberrant peptide identifications against previous findings using multi-stage FDR (b) Comparison of overlapping aberrant peptide identifications using combined FDR. Our proteogenomic database was created from raw RNA-seq alignments from TCGA repository and database used in Zhang et al. [ZWW+14] is created from SNV informations reported by dbSNP [SWK+01], COSMIC [FBB+11], and TCGA somatic mutation calls [MBC+12].

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.

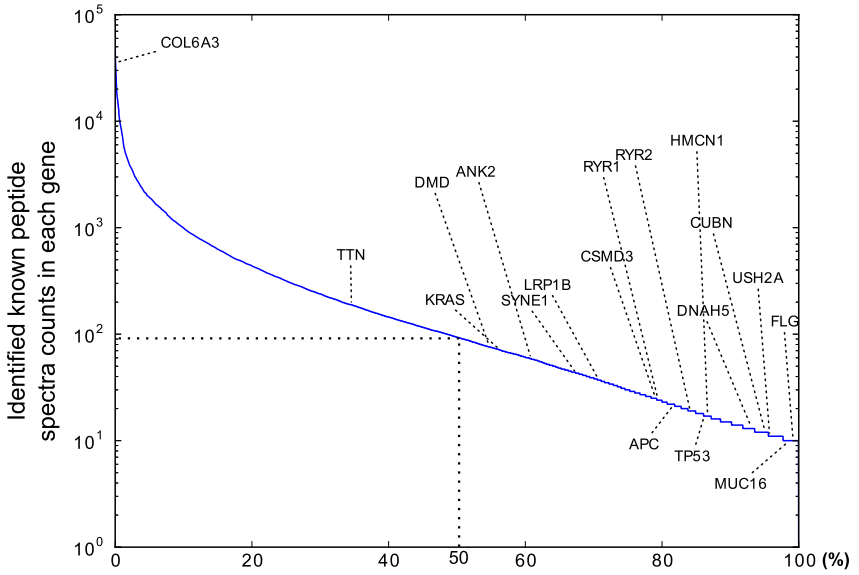
Comparisons in protein versus genomic level mutation analysis. Initial comparisons between the expression and recurrence of variant peptides suggested significant differences [ZWW+14]. As we did not have matched proteomic data from normal samples, we used an earlier study from TCGA [MBC+12] to call somatic 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 5.3(a) shows the top 30 most frequently mutated genes reported by the TCGA study [MBC+12]. However, these genes have extremely low protein expression (as measured by spectral counts) even for non-mutated peptides (Figure 5.3(b)). In contrast, the most recurrent proteins with somatic mutations have variable recurrence using RNA expression (Figure 5.3(c)), but the list identifies many genes of interest. However, genes such as TNC [TSK+13], HSPG2 [JC03], PML [VSP+12], GBP-1 [BLLO+13], TF [SLW+09], NES [TNI+07], have all been implicated in colorectal tumor angiogenesis.

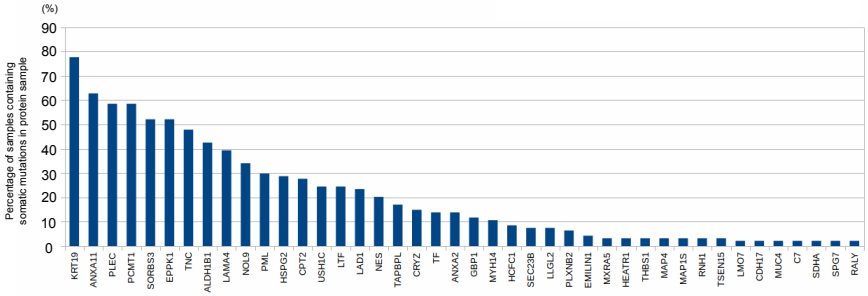
Figure 5.3: (a) Genes containing most frequent somatic mutations reported by the TCGA study. (b) RefSeq identified spectra per gene in 10 based log scale. Most frequently mutated genes in DNA level are under expressed in protein level. COL6A3 had 35463 spectra counts, TTN (188), KRAS (71), DMD (76), 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.


(%)
Percentage of samples containing Somatic mutations in genomic level
90
80
70
60
50
40
30
20
10
0
APC
TP53
TTN
KRAS
SYNE1
MUC16
LRP1B
FAT4
PIK3CA
FLG
DNAH5
FBXW7
CSMD3
LRP2
DMD
USH2A
CSMD1
ANK2
COL6A3
FAT2
ZFHX4
SDK1
CUBN
HMCN1
DCHS2
RYR1
CSMD2
DNAH2
PKHD1
RYR2

(a)


Identified known peptide spectra counts in each gene
10^5
10^4
10^3
10^2
10^1
10^0
0
20
40
50
60
80
100 (%)
COL6A3
TTN
DMD
ANK2
KRAS
SYNE1
LRP1B
CSMD3
RYR1
RYR2
HMCN1
CUBN
USH2A
DNAH5
FLG
APC
TP53
MUC16

(b)


(%)
Percentage of samples containing somatic mutations in protein sample
90
80
70
60
50
40
30
20
10
0
KRT19
ANXA11
PLEC
PCMT1
SORBS3
EPPK1
TNC
ALDH1B1
LAMA4
NOL9
PML
HSPG2
CPT2
USH1C
LTF
LAD1
NES
TAPBPL
CRY2
TF
ANXA2
GBP1
MYH14
HCFC1
SEC23B
LLGL2
PLXNB2
EMILIN1
MXRA5
HEATR1
THBS1
MAP4
MAP1S
RNH1
TSEN15
LMO7
CDH17
MUC4
C7
SDHA
SPG7
RALY

(c)

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 [NGS03], the role of B-cells is still being elucidated, although some reports suggest that B-cells aggregate in tumors [Lin13, NSM$^+$12], where they form germinal centers, undergo class switching, and differentiate into plasma cells [Nel10, ONI$^+$09].

Distribution of IG peptides across colorectal subtypes. The CPTAC study classified the 90 samples into five subtypes, marked 'A' through 'E' [ZWW$^+$14] based on expression patterns. Figure 5.3 shows the plot of IG gene peptide spectra counts (normalized by the total number of known spectra identifications in each group) between each sample subtypes. In addition, MS/MS data-set from normal colon/rectal [KPG$^+$14] tissue, and colon cell-line [FSL$^+$13] were used as controls. We observed that IG peptide identification rate in all sub-types were similar to the normal sample compared to the majority of cancer samples (except samples within group 'C') while cell-line sample showed a markedly smaller number of IG peptide identifications. The one exception was the significant over-expression of IG peptides in subtype 'C' ($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. [SLH+13] and De et al. [DSEMWJ+13]. Our results suggest that a strong immune response could be a molecular marker of CRC sub-types.

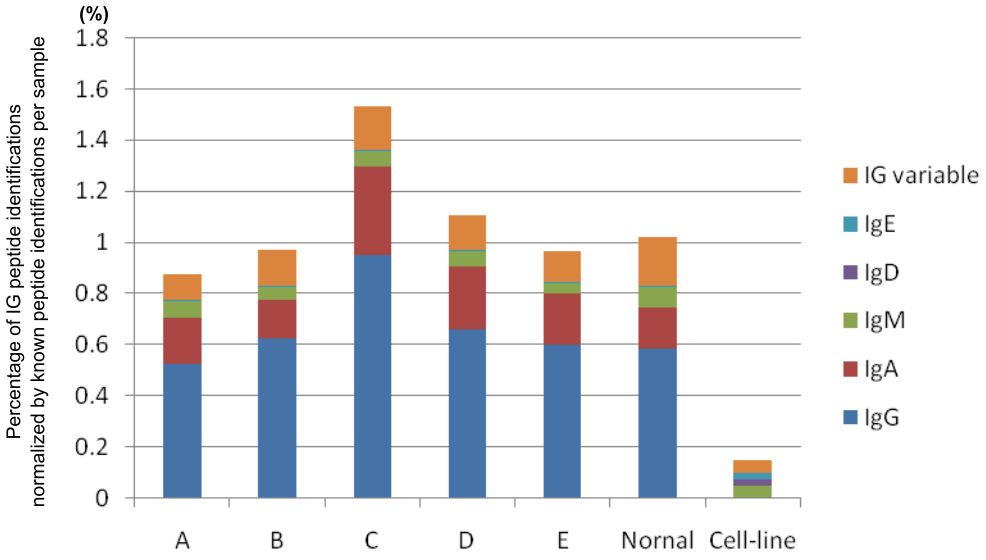


Figure 5.3: Percentage of IG gene peptide identifications in each sample normalized by the number of known peptide identifications across sample subtypes. This percentile ratio is calculated by dividing the number of known peptide identifications from the total number of IG peptide identifications within 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 $p\text{-}value < 0.0001, \chi^2 = 2927.71$.

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' ($p\text{-}value < 0.0001, \chi^2 = 40.39$). In the initial TCGA [MBC+12] and CPTAC [ZWW+14] 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 TGF-beta signaling pathway suppressing epithelial cell growth, and inactivation of the Smad4 gene through an intragenic mutation occurs frequently in association with malignant progression [MK03, Liu01]. We identified a single PSM 'VPSSCPIVTVDGYVDPSGGD;H;FCLGQLSNVHR' ($R361H$, Figure C.8) supporting a known mutation in colorectal cancer [FBB$^+$11], 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 [AWP$^+$08]. We identified a known, low-frequency, mutated peptide, 'LVVVGAG:D:VGK' ($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 [ATW$^+$11]. We identified a mutation (Figure C.11) with strong overlapping peptide identification. We also identified overlapping peptides in FGA gene, which has been propsoed as a marker for other cancers [TLG$^+$12].

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 [PBH$^+$14]) 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 [MBC$^+$12] 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 [LWW$^+$13].

5.2.3 Conclusion

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.

5.3 Immunoglobulin assemblies from TILs

5.3.1 Introduction

Tumor infiltrating lymphocytes

Immune cell infiltrates into tumors is a well studied phenomenon due to their potential for identifying tumor cells, as well as to understand how the immune system fails to eradicate the tumor. While tumors evade the immune system, these infiltrating cells have been used as a prognostic indicator for some cancers [BHAH$^+$13, NSM$^+$12]. These tumor infiltrating lymphocytes (TILs) have largely been studied in the context of T-cells, however, recent studies suggest that infiltrating B-cells may provide insights as well [Nel10, LM12, Lin13, WLM$^+$14, MRTM14, KGJ$^+$13].

While the role of B-cell infiltrates is being investigated [LM12, WLM$^+$14, KGJ$^+$13], it is not being done so at the level of molecular sequences. There are few examples of studies

Table 5.1: 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

performing sequence analysis of infiltrating B-cells [CTK$^+$01], as such, this relm remains largely uncharacterized. High throughput sequencing of immunoglobulins has now become feasible, therefore representing the T-cell and B-cell repertoires at the nucleotide level is now a reality. Unfortunately, no tools exist for the purpose of utilizing general RNA-Seq data for the analysis of T-cell receptors (TCR) or immunoglobulins (Ig). General tools for short read RNA-Seq data will be unable to identify many reads originating from these regions due to the high divergence from germline references.

Immunoglobulin sequencing

High throughput sequencing of immunoglobulin transcripts from B-cells was pioneered in model organisms [WJW$^+$09, JWP$^+$11]. Immunoglobulin sequencing (Ig-seq) has now

become a viable approach to studying human immunoglobulin repertoires at the molecular level [ALC+11, JHW+13, VSW+13, LVGM+14]. Earlier Ig-seq studies relied on Roche 454 sequencing to provide sufficiently long reads to cover the variable region of the antibody; this came at the cost of lower throughput and sparser coverage of the repertoire. Illumina read lengths are now sufficiently long to cover the entire variable region when overlapping paired-end reads are used [SBK+15].

Ig-seq has been employed in humans to study the antibody response to vaccine antigens [JHW+13, VSW+13, LVGM+14]. This is a very powerful way to view the changing immune system, and has the potential for identifying antibody candidates. However, Ig-seq has not been applied to studying tumor infiltrates; perhaps because their potential has not been fully described, or due to the variable nature with which they are observed.

While no such Ig-seq data has been created, many RNA-Seq experiments have been performed on primary tumor samples to examine the expression patterns, or identify fusion transcripts. These RNA-Seq experiments seek to understand the molecular basis of cancers by probing the transcriptome. However, since conducted on primary tumors these samples contain tumor cells, host cells, as well as immune infiltrates. A recent study extensively characterized T-cell cytolytic activity across TCGA cancer samples [RSW+15]; showing that molecular signatures for T-/B-cells can be identified from RNA-Seq. We will use this general approach of using primary tumor RNA-Seq data to reconstruct and analyze a targeting component of the immune system: the antibody molecule.

Ig-mining

While characterizing tumor infiltrates from general RNA-Seq has been shown possible [RSW+15, GNL+15]; obtaining immunoglobulin variable regions from short read data is not the same task. In order to extract B-cell immunoglobulin transcripts from such general experiments using short read data, we cannot rely on standard tools. Such immunoglobulin mining (Ig-mining) from short read data, will require a different approach.

Standard read mapping tools will miss the majority of reads from antibody transcripts due to the high divergence from reference; yet it is this divergence that is of particular interest! To overcome this shortcoming, we have developed TILAPIA which works by: 1) recruiting unmapped reads potentially originating from the immunoglobulin heavy chain (IGH) locus, together with those already mapping; 2) the de Bruijn graph over these recruited reads is created and pruned; 3) antibody variable regions are assembled from the de Bruijn graph. These assembled antibody variable regions can then be analyzed in the same manner as data obtained from Ig-seq.

Here we describe the tool TILAPIA: Tumor Infiltrating Lymphocyte Analysis Performed by Immunoglobulin Assembly. In addition to describing TILAPIA, we will justify the specialized antibody assembler on simulated data, and present results on 397 colon cancer samples. We further show that B-cell infiltrate response differs from colon to breast cancer with regards to isotype switching.

5.3.2 Methods

Datasets

Colon cancer dataset

Sequencing data of mRNA from 397 human colon cancer tumor tissues were downloaded from The Cancer Genome Atlas (TCGA) [N$^+$12]. Clinical, pathology, and slide metadata files associated with these samples were also downloaded and used in the analysis. 344 samples are retained for further analysis of the constant region, based on the number of mapping reads (Supplemental Figure D.1).

Breast cancer dataset

Sequencing data of mRNA from human breast tissue, previously analyzed [VGR$^+$14], were downloaded from the NCBI short read archive (SRP042620). BAM files were gener-

ated from the downloaded FASTQ files that were aligned to human reference HG19 using the STAR aligner [DDS+13]. For expression analysis, a count matrix was computed using HTSeq [APH14], and differential expression computed using DESeq2 [LHA14]. 141 samples were used in total: 42 triple negative breast cancer primary tumors (TNBC); 42 Estrogen receptor positive and HER2 negative primary tumors (ER+); 30 uninvolved breast tissue samples adjacent to ER+ tumors, 21 uninvolved breast tissue samples adjacent to TNBC tumors; and 5 breast tissue samples from reduction mammoplasty procedures on patients with no known cancer.

Mass-spectrometery dataset

The data used for mass-spectrometry analysis consists of mRNA sequencing data from 73 human colon and rectal tumor tissues were downloaded from The Cancer Genome Atlas (TCGA), of which 44 are colon cancer, and 29 are rectal cancer. Mass-spectrometry datasets from these same samples were obtained from The Clinical Proteomic Tumor Analysis Consortium (CPTAC). These samples were analyzed at a proteome-level scale in a previous study [ZWW+14].

Filtering IGH locus from RNA-Seq data

Since the majority of mRNA sequenced are from tumor cells of the colon, we must filter out only those reads mapping to the immunoglobulin heavy chain (IGH). Any transcript mapping to the IGH locus would suggest that it originated from a B-cell, specifically a tumor infiltrating B-cell (TIB). Typical mapping procedures, e.g., bowtie, will miss many transcripts from the IGH variable region due to the imprecise junction of V/D and D/J gene-segments, and high somatic mutation rates of the IGHV gene-segments. Figure 5.4(a) shows the potentially missed reads from a somatically recombined heavy chain transcript as grayed out, while mapping reads as darker. This shows that we can obtain mapping reads for some parts of the V gene-segment, and the constant region, but will need a different approach for the junction

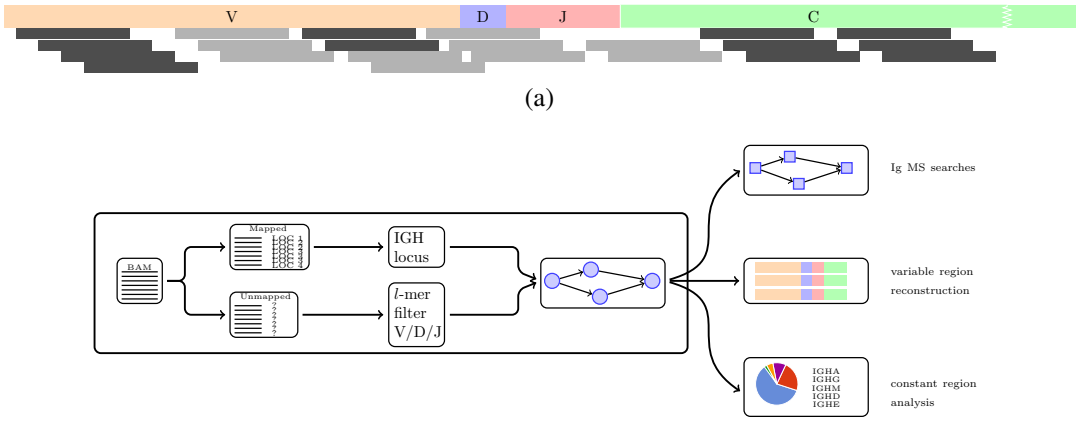


Figure 5.4: Immunoglobulin mining overview. (a) Read mapping example. Standard reference mapping approaches will fail to properly recruit many reads from antibodies due to the highly variable nature of these transcripts. (b) IGH reconstruction. Outlines the approach used by TILAPIA for reconstructing antibody transcripts from short read data, in a sample that is a mixture of tumor, normal, and immune cells.

region and highly mutated areas of the V gene-segment.

Figure 5.4(b) shows the approaches used for reconstructing the IGH locus. Starting with a BAM file for the RNA-Seq experiment for an individual sample, a series of l-mer filters are applied to retain reads only pertaining to the IGH locus. These reads are used to construct the de Bruijn graph of the expressed IGH transcripts, termed the *repertoire graph*. Finally, the heavy chain transcripts are assembled from the graph.

Variable region assembly and analysis

To obtain IGH transcripts, we must assemble them from the read data. All unmapped reads are filtered for matches to l-mers in the V, D, J, and C gene-segment references. Those with matches are retained, along with mapping reads, to construct the de Bruijn graph. Pruning of this graph is performed to remove errors. The resulting repertoire graph is then assembled to obtain near full length transcripts.

Filtering for variable region

Starting from an RNA-Seq sample, reads that are likely to pertain to the IGH locus are filtered out. The full filtering procedure is outlined in Algorithm 4 in the supplement. All reads mapping to the IGH locus (chr14:105566277-106879844) are retained, and combined with all unmapped reads. This subset of reads is then filtered for quality by: removing any reads containing an 'N'; removing any reads with a mean quality value below 25; remove any reads without any l-mers matching V, D, J, and C reference l-mers. Any retained reads are then trimmed based on quality values from both the 5' and 3' ends. Bases with quality values below 10 are trimmed from both ends of the read and if more than 66% of the read is trimmed in this manner, then the read is discarded. All retained reads are *putative IGH reads*, and retained in set R for assembly. Figure D.1 shows the number of variable region and constant region reads retained over the colon cancer data sets.

Selection of l for the l-mer filter is an important step. The l parameter presents a tradeoff between sensitivity (selecting reads from true V, D, J, and C gene-segments), and specificity (missing reads due to high variability from somatic hypermutation and recombination). One way to approach this tradeoff is to select the smallest l that has an acceptable false positive rate of including reads. This can be represented by spurious reference l-mer hits to each read (described in the Figure D.7 and Supplemental section: Modeling spurious l-mer matches).

Repertoire graph construction

The putative IGH read set R is used as input to construct the de Bruijn graph $G_{Ig} = (V,E)$, termed the *repertoire graph*, over k-mers of these reads. Nodes in this graph represent all $(k-1)$-mers over the set of reads R. Nodes $u,v \in V$ are connected by a directed edge $(u,v) \in E$ if there exists some k-mer in a read, contained in R, whose prefix is u and suffix is v.

Algorithms using de Bruijn graphs have been applied to genomic assembly [PTW01, ZB08], transcriptomic assembly [GHY$^+$11, PLY$^+$13], as well as meta-genomic assembly [PLYC11]. Transcript assembly algorithms such as Trinity [GHY$^+$11] or IDBA-tran [PLY$^+$13] could be

used to assemble IGH transcripts. These assemblers make certain simplifying assumptions, that work on the majority of loci. Unfortunately, the IGH locus violates many of these simplifications. For example, Trinity [GHY$^+$11] exhaustively enumerates branching paths to a predefined depth cutoff. This is to ensure that the correct transcript is assembled. This works when branches are uncommon; repertoire graphs have many branches, so this strategy will miss many antibody contigs (Figure D.2).

Once the repertoire graph G_{Ig} is constructed, error correction can be employed. Similar approaches to transcript assemblers [GHY$^+$11, PLY$^+$13] are employed. Tip clipping/bubble removal is performed using a proportional coverage heuristic. Additionally, a uniform coverage rule is applied after tip/bulge simplification to remove any very low coverage tips that will not be able to be incorporated into contigs.

The repertoire graph is further simplified by retaining only the largest connected component within the graph. This largest component is assumed to represent the IGH locus due to the pruning/recruiting of reads mostly pertaining to this locus. Retaining the largest connected component removes many small, spurious, components likely created by reads that erroneously passed the filters.

Repertoire graph assembly

The main difference between the assembly approach using the repertoire graph and other transcript assemblers rests in two simple aspects. Since a relatively low number of reads, compared to the total, are considered, we can retain all read information on edges in the graph. Additionally, the heuristics used to determine which which branches to be considered differs. It is common for a branch to be considered only when a significant number of nucleotides overlap on each side of the branching node. Otherwise, the approach is similar to that used for IDBA-tran.

When considering the heaviest path between source and sink nodes, the reads are considered for determining coverage over the edges on the path. Edges continuing from

previous heaviest paths are considered as candidates, and those with the most overlap in continuing reads are retained. Reads starting from the candidate arc are given a lower weighted score. At each node, the n heaviest paths are retained. Nodes are processed in topological order. Full details can be found in Supplement Algorithms 3 and 4.

VJ labeling and analysis

Labeling of assembled antibodies was performed using IgBlast [YMMO13]. An e-value threshold of 10^{-10} was employed to reduce poor labeling. Mismatches in the V gene-segment were noted and are used for mutation analysis.

For each sample, all $52 * 6 = 312$ VJ pairs are quantified by aggregating the effective read count from eXpress [RP13], for each VJ labeled transcript, into a vector v. For a sample i, we obtain a vector v_i of counts for each VJ pair. Each v_i is unity normalized $\hat{v}_i = \frac{v_i - \min(v_i)}{\max(v_i) - \min(v_i)}$ to facilitate comparisons across samples.

Clustering is performed on unity normalized VJ vectors, comparing two samples of V gene expressions using cosine distance and Ward clustering metric. VJ clusters are plotted by summing each $\hat{v}_i$ for all samples i, and representing each V-J pair using the unity normalized expression.

VJ permutation test

A permutation test was carried out to determine if the most abundant VJ pairs were observed significantly more than random. Each sample i considered has a representing vector of VJ pair abundances, v_i. To determine the null distribution of VJ clusters represented as a set of unity transformed VJ abundances, each $\hat{v}_i$ was randomly shuffled, effectively relabeling the abundance of each VJ pair for sample i. This was performed for all samples, and the corresponding VJ clusters were computed as described above.

The maximum valued VJ cluster from each iteration is noted in the cumulative count vector c. E.g., if the maximum valued VJ cluster is k for an iteration, $c_{0:k}$ are incremented by

one. After 10,000 iterations, c contains a distribution for the maximum VJ value expected by chance. We can then select the cutoff at a 0.01 level by selecting the largest index j whose cumulative sum from j to 312 is below this threshold. In other words, select the largest j such that satisfies $\sum_{i=j}^{312} c_i \leq 100$.

Constant region analysis

Constant region mapping

Reads that map to the IGH constant region genes, e.g., IgG1, IgG2, IgA1, IgA2, IgM, etc, are used for the analysis of isotype abundance, and analysis of mutations along the constant region. These reads are determined by mapping of the read to the reference segments using bowtie, and selecting the reference with the fewest number of mismatches. If there is a tie, both (or multiple) reference are assigned the read, but with a fractional value of $1/n$, for n matching references. The constant region mapping for breast cancer samples was performed differently, described previously, to better accommodate a differential expression analysis.

Mutations in constant region

The analysis of mutations on the constant region is carried out by identifying which mutation positions are observed within each sample. A position is deemed as containing a mutation if it cannot be explained by the error rate of sequencing. This is done by calculating the probability of it coming from the distribution of noise, modeled by a binomial distribution, seen in equation 5.3.

$$P(x|n,p) = \binom{n}{x} p^x (1-p)^{(n-x)} \tag{5.3}$$

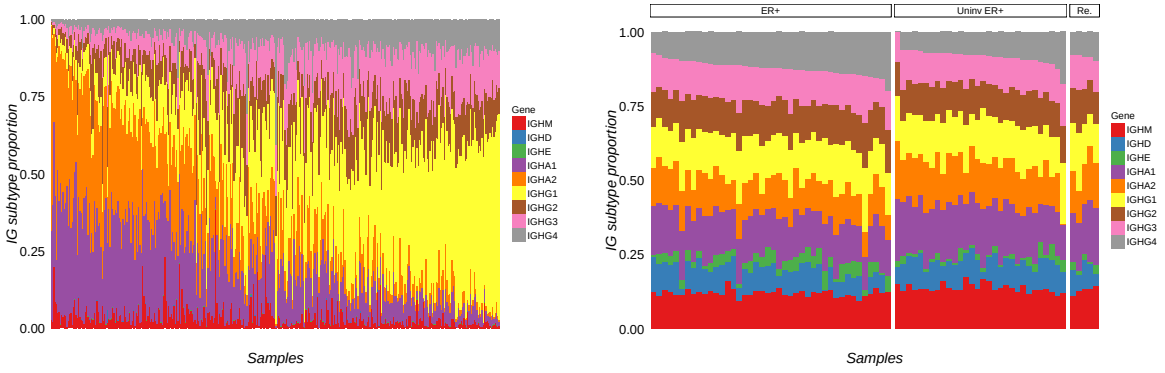
Here, n is the coverage at the position in question, x are the number of mutated sequences, and p is the probability of an error, set at 0.01. Equation 5.3 will provide the probability of observing x substitutions, from a coverage of n, from errors. Any positions below

a 0.05 cutoff, are deemed significant. Bonferroni correction is employed, applied to tests for all positions across all constant gene-segments.

5.3.3 Results

Constant region

Isotype expression

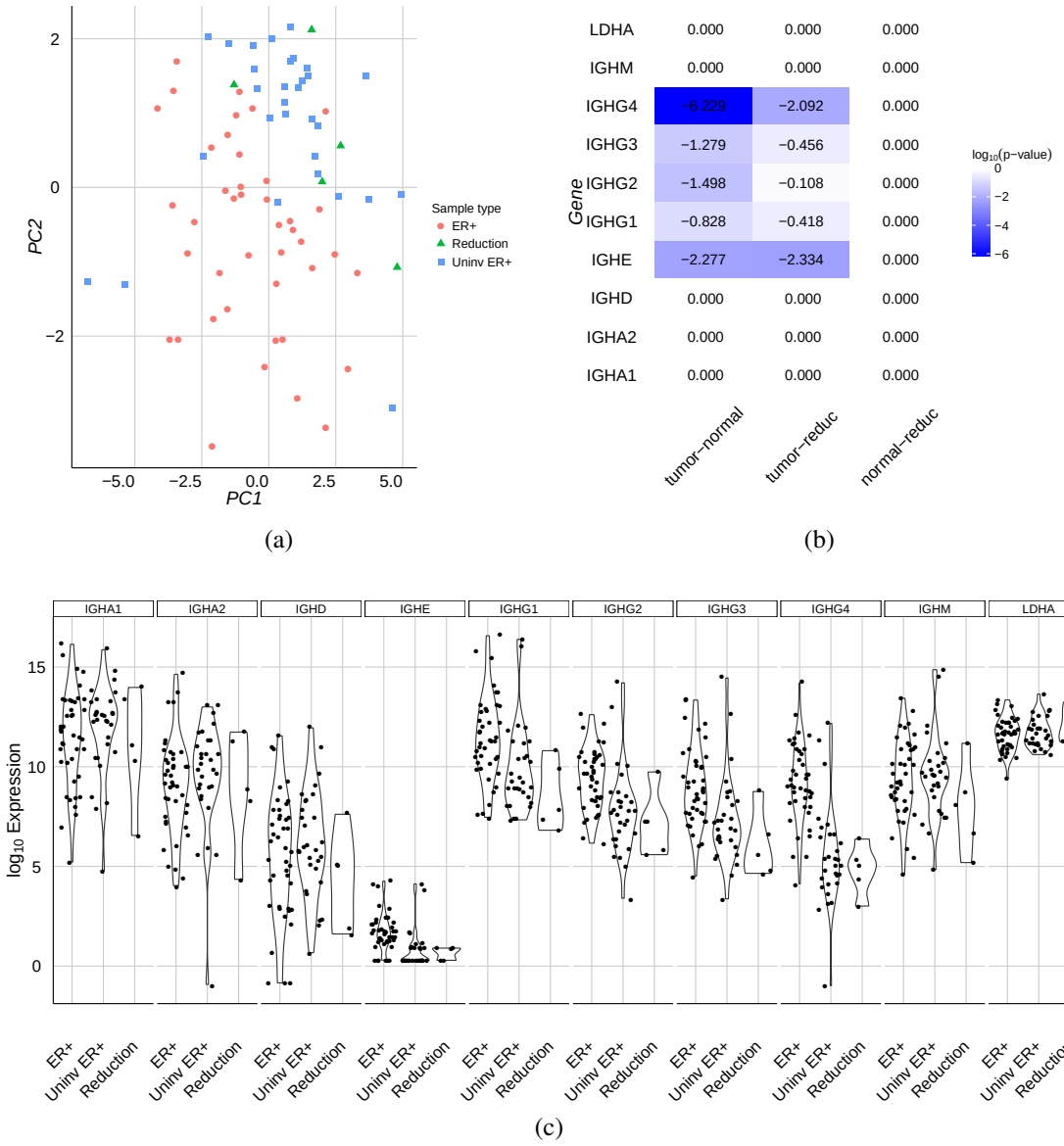


(a) (b)

Figure 5.5: (a) Isotype distribution in colon cancer. The relative abundances among 344 samples is shown, sorted according to IgG1. (b) Isotype distribution in breast cancer, split according to ER+, Uninvolved ER+, and reduction (Red.) samples.

The constant region of antibody chains are far less variable than their V, D, and J gene-segments. As such, standard mapping of reads to references can be performed. Additionally, analysis of mutations from mapped reads is performed as described in Methods:IGH constant region. The relative proportion of each isotype in colon cancer is shown in Figure 5.5(a), depicting a clear split between IgA's and IgG's (Supplemental Figure D.4(a) and D.4(b)). Furthermore, IgG1 is shown to be dominant among the IgG isotypes. The breast cancer dataset, originally from Varley et. al., 2014 [VGR+14], contains two tumor subtypes (ER+ and TNBC); along with uninvolved tissues for each, intended as a control. An additional 5 breast reduction mammoplasties are included, intended as an additional negative control. The proportion of Ig constant regions in breast cancer is shown in Figure 5.5(b), sorted according to IgG4. The

subtype proportions in breast cancer are less variable, however there is an increase in relative IgG4 expression in ER+ compared to uninvolved and reduction samples.



(a) (b)

(c)

Figure 5.6: (a) Principle component analysis of breast cancer samples using only IGH constant region genes. Samples shown are ER+, Uninvolved ER+, and reduction mammoplasty. (b) Bonferroni corrected p-values for the T-test of each gene, for each pair of sample type. (c) Violin plot of IGH constant region expression (log10) for ER+, Uninvolved ER+, and reduction mammoplasty samples.

The experimental setup from Varley et. al., 2014 allows us to verify if the adjacent uninvolved tumor samples can be considered a reasonable control for measuring infiltrate

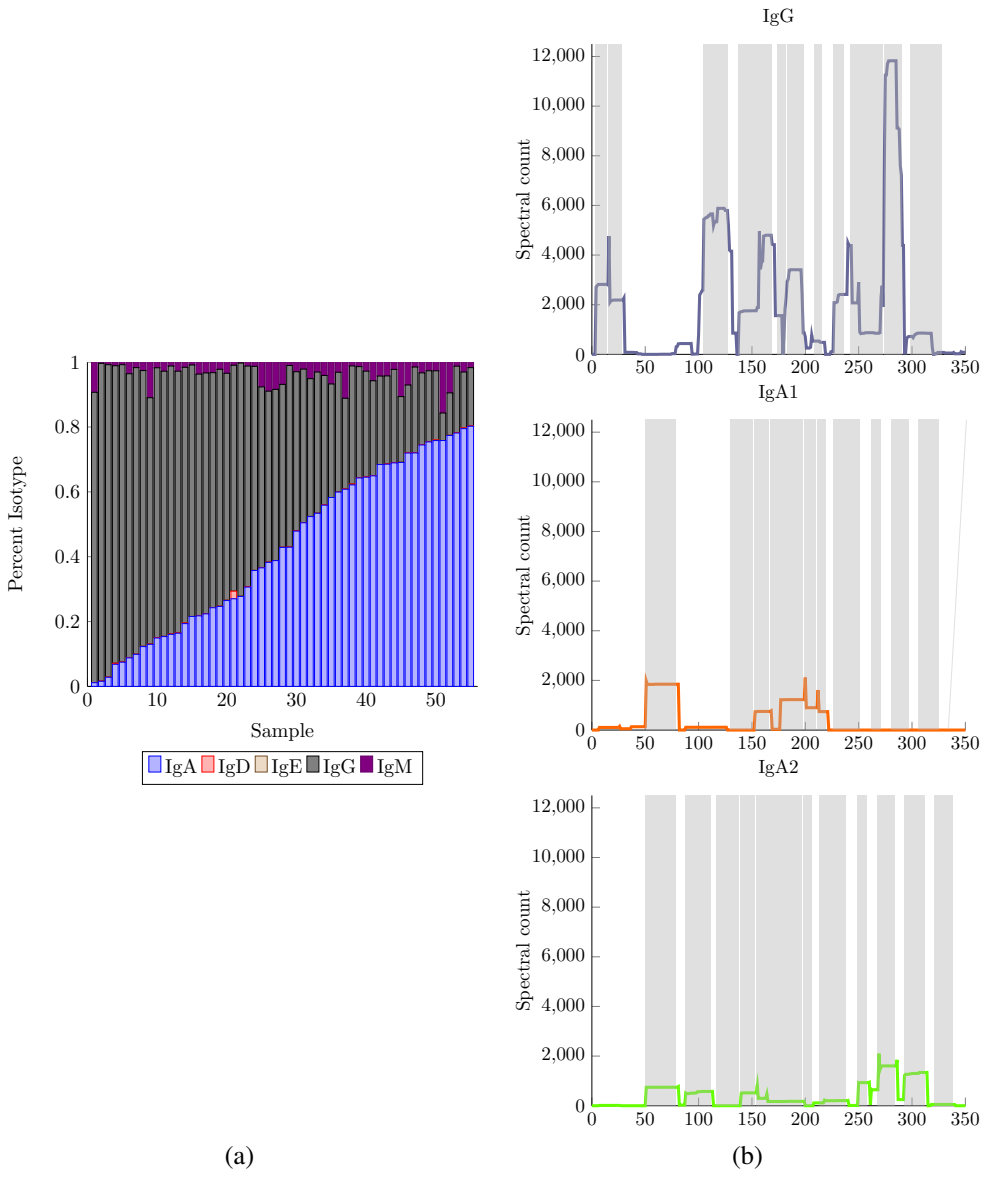
response. It also allows us to compare the immune response to two, clearly defined, tumor subtypes in breast cancer. After normalization of RNA-Seq expression across samples was performed, IGH constant region genes were extracted, and used for analysis. Additionally, a single housekeeping gene, LDHA, is shown as a negative control. Figure 5.6(a) shows the first two principle components (representing 50% and 23% of the variance, respectively) of IGH constant region genes for ER+, uninvolved ER+, and reduction mammoplasty samples. Most of the ER+ samples group with one another, separate from the uninvolved ER+ and reduction samples. While this grouping is not as clear as that with all normalized genes (Supplementary Figure D.5), it does suggest that there is an IGH specific response to this tumor subtype. Further, there appears to be little difference between uninvolved ER+ and reduction samples, for both IGH genes (Figure 5.6(a)) and all genes (Supplementary Figure D.5), suggesting that uninvolved tissues can be considered a proper control when measuring IGH infiltrate response.

The p-values of pairwise comparisons of each gene to different samples, using a T-test with Bonferroni multiple testing correction, is shown in Figure 5.6(b). The same comparison is visualized using a violin plot in Figure 5.6(c). These show that for the ER+ subtype, IgG4 exhibits a significantly higher response for the tumor when compared to the adjacent tissue, and reduction mammoplasty. No such distinction can be drawn from the TNBC samples (Supplemental Figure D.6).

This observed increase of IgG4 levels is present in breast cancer, but absent in our colon cancer samples. However, increases in relative IgG4 levels have been reported for melanoma [KGJ$^+$13] and shown to be a potential immune escape mechanism of the tumor. This effect is described by Karagiannis et. al., 2013 showing that IgG4 competes with IgG1 for FcγRI activation, reducing tumoricidal activity.

Deviations in constant regions and mass-spectrometry evidence

From all reads mapping to an IGH constant gene-segment in the colon cancer samples, Figure 5.7(a) shows the relative abundance of each isotype, ordered by IgA. This plot shows



(a) (b)

Figure 5.7: (a) IGH constant region read counts. (b) Spectral count of IGH constant region tryptic peptides. The IgG plot is reproduced from Woo et. al., 2015 [WCB+15].

a clear delineation between IgG and IgA. The spectral counts, for all samples, along the positions of IgG, IgA1, and IgA2 are shown in Figure 5.7(b). Tryptic peptides of lengths 7 to 35 are grayed out. The absolute abundance of IgG genes eclipses that of IgA, however, closer inspection shows that the majority of IgA1 is never sampled by mass-spectrometry, particularly the C-terminal end. This poor sampling of IgA1 is in contrast to to all tryptic peptides highly covered in the IgG subfamily, and nearly all tryptic peptides covered in IgA2, at much lower abundances.

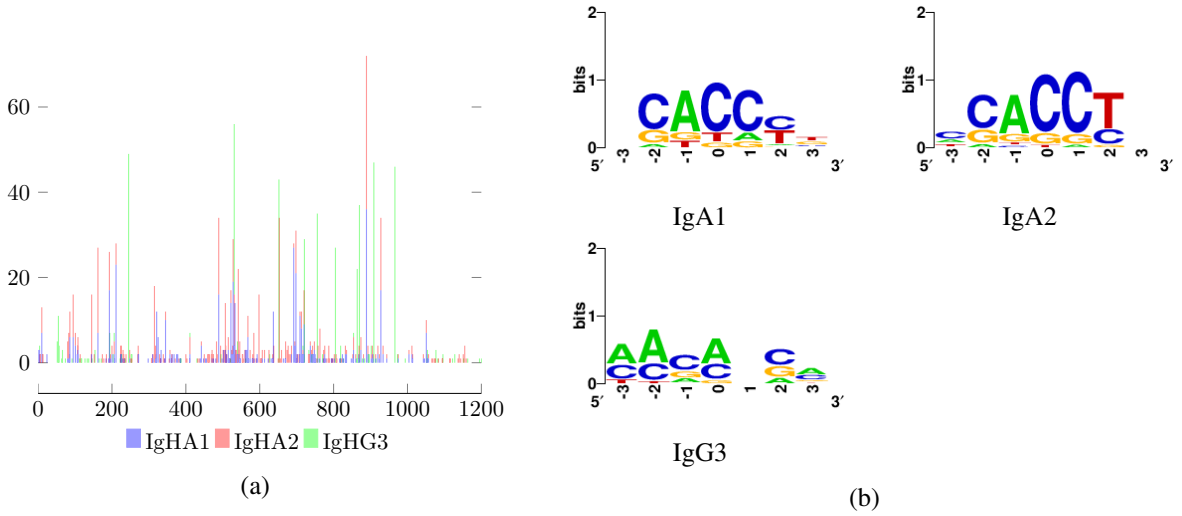


Figure 5.8: (a) Stacked bar plot of positions with mutations, and the number of samples, out of 74, that are significant. (b) Motif of mutation positions in IgA1, IgA2, and IgG3.

While the variable region of the IGH transcript is highly mutated by somatic hypermutation events, we observe many substitutions along this constant region as well. Differentiating read errors from true substitutions is important, and we take a statistical approach, described in the methods section. Figure 5.8(a) shows, for each position (bp), the stacked bar plot of the number of samples with significant substitution at that position, for each isoform. Figure 5.8(b) shows the sequence logos centered around positions with significant substitutions, that occur in more than 10 samples. IgA1 and IgA2 show nearly identical motifs, and fall within the known motif for somatic hypermutation, WRCY. IgG3, was the only IgG gene-segment containing a motif, however, it does not correspond to any known pattern and is a weaker signal than those

from IgA1 and IgA2, seen in Figure 5.8(b).

Immunoglobulin tumor repertoire

Once assembling, quantifying, and labeling is performed, the resulting variable regions can be clustered. Figure 5.11 shows the heatmap and clustering of the V gene-segments, and the VJ pairings as a dot plot, across 160 samples. Samples with few reads recruited for the variable region were unable to produce assemblies, another example of the heterogeneous nature of immune response to tumors.

Despite typical heterogeneity in strength of response, measured by recruited reads, Figures 5.9, 5.10, and 5.11 show that across samples, there are commonalities in B-cell responses. Figure 5.9 shows distinct clusters around V gene-segments IGHV3-11, IGHV3-23, and IGHV3-30. No discernible pattern with cancer subtype emerges (subtype derived from cluster labels from Broad firehose), shown below dendrogram.

Figure 5.10 shows the VJ pairs across all colon samples. The pairwise VJ plot is created by summing the normalized vectors of VJ expression, across all samples (see Methods). The aggregate suggests that certain germline preferences may exist in the colon; IGHV3-11, IGHV3-23, and IGHV3-30 all pair with IGHJ4, and are common across many samples. All three of these pairs are significantly elevated compared to chance when performing a permutation test (see Methods). No iteration of the permutation test yielded VJ pairs with such high counts. This germline preference may not be related to tumor specificity, but could merely be a preference seen at a population level.

Mutations along antibody V gene-segments can improve specificity and determine how long the clone has been selected. Figure 5.11(a) shows the distribution of mutations along each V gene-segment, partitioned among samples with the most prevalent VJ pairs. This distribution shows that most transcripts contain approximately ten mutations, while many contain nearly 30. The density of mutations for each partition is shown in Figure 5.11(b). Two of the prominent partitioning are bimodal in nature.

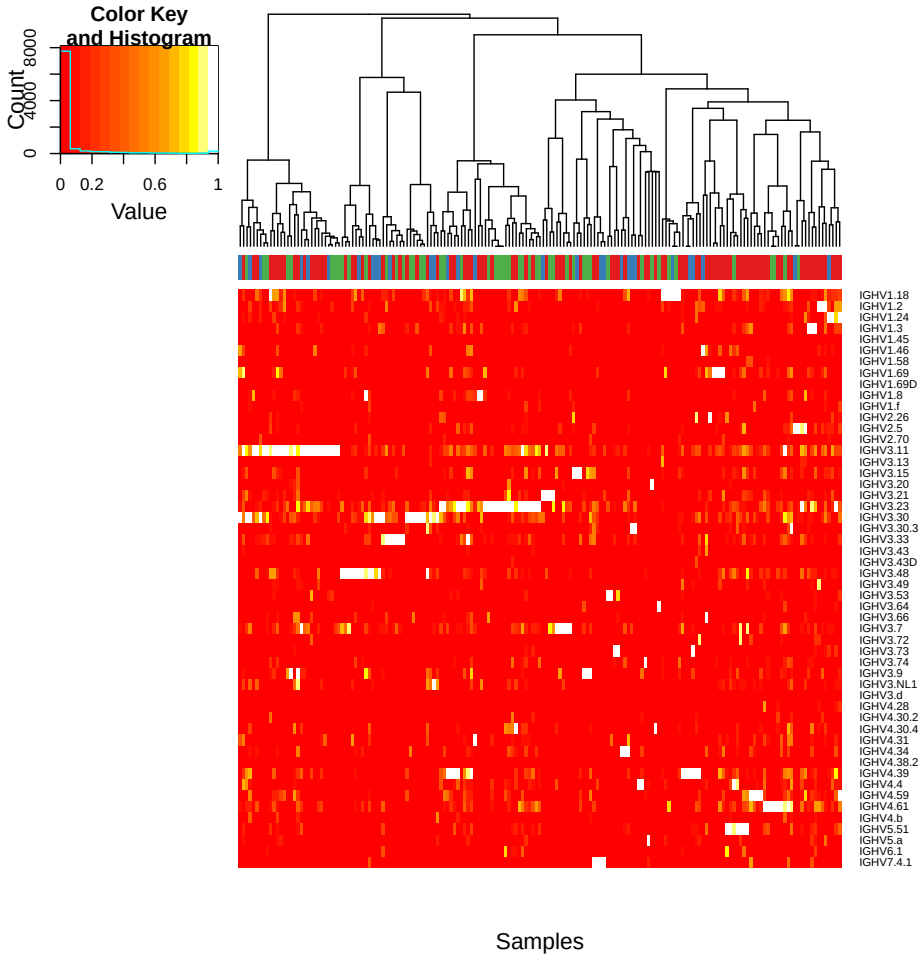


Figure 5.9: V gene clustering. Clustering of unity normalized V gene expression across samples using cosine distance and Ward clustering metric.

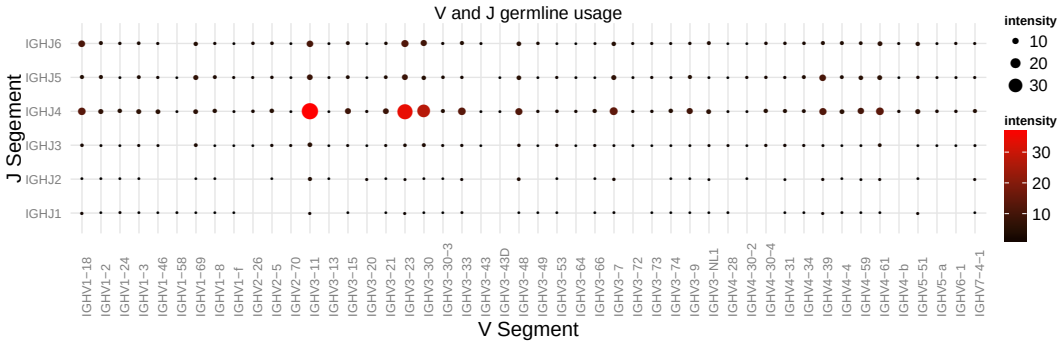
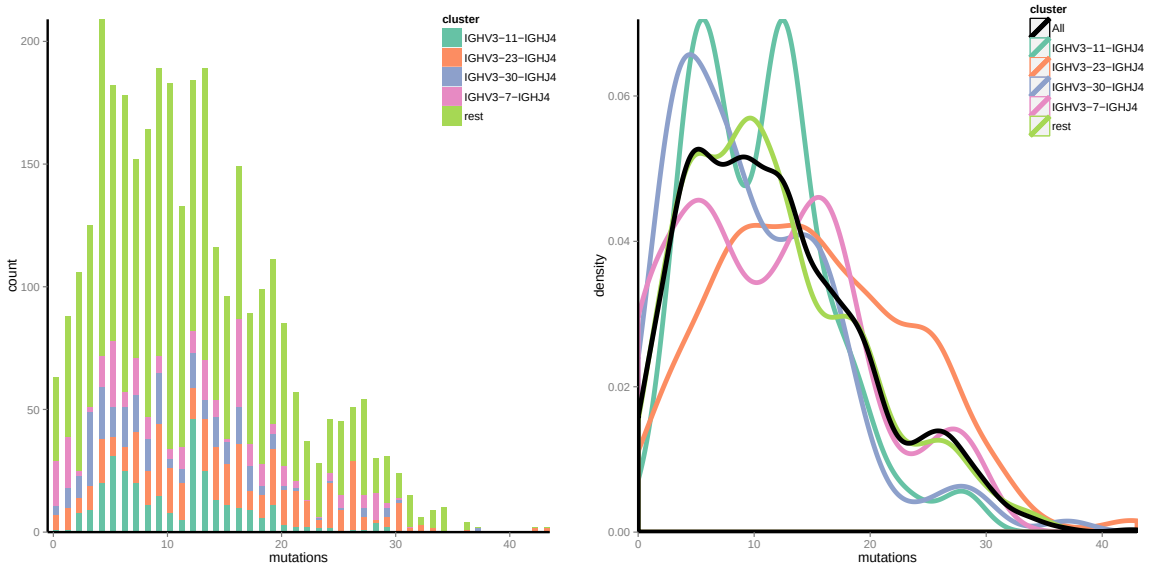


Figure 5.10: VJ pairs across samples. Aggregate of normalized expression of VJ pairs across colon cancer samples, showing that some VJ pairs are over-represented.



(a) (b)

Figure 5.11: (a) Mutation distribution. Distribution of mutations along V gene-segments with samples partitioned according to the four largest VJ pairs. (b) Density of mutations. The density of mutations for each partition of VJ pairs.

5.3.4 Conclusion

The precise role immune infiltrates play in the tumor microenvironment remains unclear [LM12, KGJ+13], this is especially true for B-cells. While some studies suggest a potential benefit [BHAH+13], others hint at a negative correlation with outcome [KGJ+13]. We focused on immunoglobulin heavy chains as a marker for infiltrating B-cells, and found differences in isotype switching of the constant region between colon and breast cancers. The relative abundances of isotypes also differed dramatically between colon and breast cancer. In addition to differences in isotype abundances, we found considerably higher expression of IgG's over IgA's at the proteomic level. Additionally, we identified several common hotspots for mutations in constant regions, across many samples, that do not correspond to allelic variants. These mutation hotspots, found in the IgA isotype, correspond to a known mutation motif for somatic hypermutation [RK92, DFFL98]. While this evidence does not prove that somatic hypermutation occurs in the constant region, further investigation is warranted. Prior work shows that the constant region can affect antigen binding affinity [PMD+00, MTE+02], so

mutations to this region could be of interest.

While we consider immune infiltrates from existing datasets, similar analysis have been performed recently [RSW+15, GNL+15], although their focus had been on other immune cell types. In addition to considering antibody heavy chain constant regions, we additionally show how to reconstruct the variable regions of infiltrating B-cell secretions. An informatic tool, TILAPIA, is presented and benchmarked to show sensitivity in reconstructing immunoglobulin transcripts. Additionally we perform analysis of reconstructed repertoires of heavy chains across colon cancer.

One novelty to this approach, beside the re-purposing RNA-Seq experiments, is the ability to leverage the large cohorts of sequenced tumor samples from projects such as the TCGA. This allows us to compare reconstructed repertoires from many individuals, a population-level comparison that is not feasible for targeted immunosequencing experiments. We observed significant similarities in germline usage across multiple individuals. It is, however, difficult to determine if these similarities are due to convergence to a tumor antigen, or preferences at a population level. While we did not observe any correlations between germline usage and tumor subtype, we were able to characterize the mutation profiles of these reconstructed immunolgobulins. Additional studies performing Ig-seq on these tumor infiltrates could yield interesting results, and provide a more complete view of the tumor response repertoire.

5.4 Acknowledgements

Chapter 5 is published partially as 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. Journal of Proteome Research, 2015 . The dissertation author was one of the primary authors of this paper.

Chapter 6

Conclusion

Over the course of this dissertation, the repertoires of the N-terminus and immunoglobulin population where characterized. The proteogenomic approaches employed allowed us to identify many events occurring post-transcriptionally or somatically. In Chapter 2, proteogenomics was used to elucidate the post-translational N-terminal events, specifically N-terminal methionine excision (NME) and N-terminal acetylation. These events introduce variability into the otherwise homogenous start of a protein. This induced N-terminal repertoire was then analyzed and used to argue for NME's true function: to reveal Ala and Ser residues to the N-terminus, possibly for N-terminal acetylation. Additionally, this approach showed that the N-terminal repertoire was more diverse in bacteria that previously thought, since N-terminal acetylation should no longer be considered a rare event for bacteria.

In Chapter 3 a tool for characterizing the immunoglobulin repertoire for analysis was presented. This colored de Bruijn graph approach allowed for identification of germline gene-segment use, despite the high level of divergence. Additionally, it also provided a faster alternative to current tools.

Chapter 4 presented a project in the proteogenomics of immunoglobulins. Here, the immunoglobulin repertoires were characterized at the level of transcripts (as is most commonly performed), and at the proteomic level. The proteomic information of the repertoire showed

even more variety in the form of post-translational modifications; more than was shown in a previous study in rabbit [BHW$^+$14].

Chapter 5 also presented the immunoglobulin repertoire, but at a different scale and in response to a tumor. While the somatic mutations from the tumor were also described in detail, and could also be considered as a repertoire of neoantigens, I focus on the immunoglobulin peptides and transcripts that were identified.

The chapter showed that B-cell infiltrates can exhibit an observable immunoglobulin repertoire, and show changes to isotype switching relating to cancer types. This approach to mining existing datasets for immunoglobulins (Ig-mining), can be contrasted with the deeper querying of the repertoire from Ig-seq experiments in Table 6.1.

Table 6.1: Comparing Ig-seq to Ig-mining approaches.

	IgSeq	Ig-mining
Procedure	Amplify/purify	Whole cell RNA-seq/MS
Depth	High	Low
# samples	Low	High
Antigen	known vaccine	unknown tumor
Examples	Cheung et al., 2012, Sato et al., 2012	Woo et al., 2015

While Ig-mining was approached by developing new methods for existing data, the result suggests that using targeted experiments, e.g., Ig-seq, to study this phenomenon could be beneficial. Indeed, the same targeted approaches to studying the immune repertoire, as described in Chapter 3 and 4, could be used to study the repertoire in a tumor microenvironment. Such a study would provide deeper coverage to the local repertoire at the tumor microenvironment. This would allow for a deeper analysis of repertoire at the tumor-immune boundary, and could be further extended into a proteogenomic framework by attempting to purify immunoglobulin proteins using anti-IgG antibodies (e.g., Protein A, Protein G). An identical framework as that presented in Chapter 4 could then be used to identify, and characterize, potential tumor specific antibodies.

Appendix A

Supplement to Immunoglobulin classification

Algorithm description

Algorithm 1 Overview of the IgGraph algorithm

Inputs:
$\mathcal{R}$: set of mAb reads
k: k-mer length
Output: L: $|\mathcal{R}| \times 3$ matrix of most likely V/D/J labelings of each read

```
1:  procedure IGGRAPH(R, k)
2:      G ← ANTIBODY-GRAPH(R, k)                ▷ create de Bruijn graph over mAbs
3:      G, H_C ← ADD-REFERENCES(G, V, D, J)     ▷ add V, D, and J
4:      for all r ∈ R do
5:          for all X ∈ {V, D, J} do
6:              C^X ← COLOR-PROFILE(G, r, X, H_C)
7:              L[r][X] ← max_{x∈X} Σ_i C^X[x][i]
8:          end for
9:      end for
10:     return L
11: end procedure
```

The functions ANTIBODY-GRAPH() and ADD-REFERENCES() create an antibody graph, and add colored reference sequences, as are described in the Methods section of the main text. The selection of labels is shown in line 7 as taking simple maximum of the sum of columns

of a specific row j of a color profile matrix C. This is the simplest form, and is improved by selecting thresholds for V gene-segments, as described in the main text and Supplemental Figures A.8 and A.9.

Algorithm 2 Get color profile for a read r, given a set of reference gene-segments $\mathcal{C}$, and colored antibody graph G

Inputs:
r read
$\mathcal{C}$: set of reference gene-segments
G: colored antibody graph
Output: C: $|\mathcal{C}| \times n$ color profile matrix, where n is the maximum length of a read

procedure COLOR-PROFILE($G, r, \mathcal{C}$)
  $P \leftarrow a_1 a_2 \ldots a_n$ ▷ get path of read r
  for all $c \in \mathcal{C}$ **do**
    for $i \leftarrow 1 \ldots n$ **do**
      if $c \in \mathcal{H}_C[a_i]$ **then**
        $C[c][a_i] \leftarrow score$[match]
      else if $c \notin \mathcal{H}_C[a_i]$ **then**
        $C[c][a_i] \leftarrow score$[mismatch]
      end if
    end for
    $C[c] \leftarrow$PROPAGATE-COLOR($P, c, r, C[c]$) ▷ Traverse bulges for color
  end for
  return C
end procedure

The PROPAGATE-COLOR() function traverses bulges, performs alignment, and propagates color; as described in Figure 5 of the main text.

Simulating Antibody Sequences

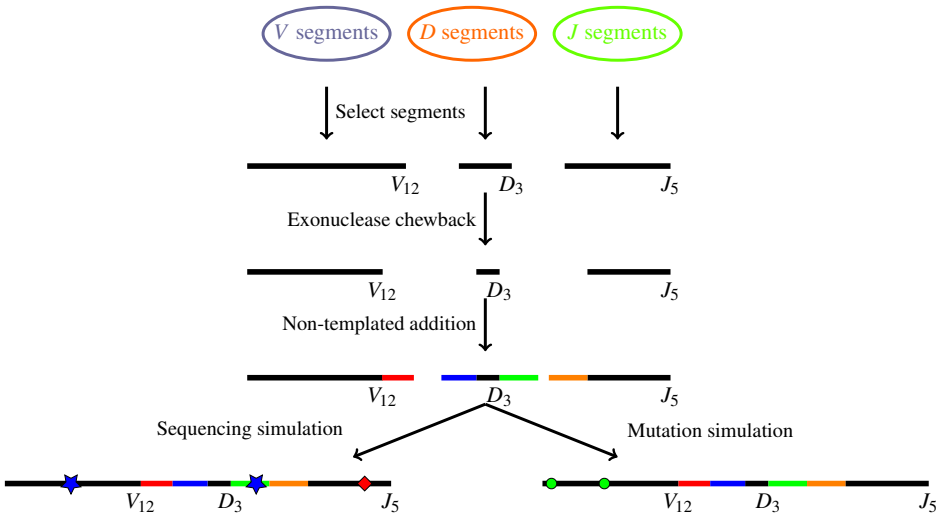


Figure A.1: Diagram of antibody simulation procedure. From pools of V, D, and J gene-segments, one of each is selected for a single smAb. Exonuclease chewback, and non-templated nucleotide additions, are performed based on empirical distributions. Finally, the read data can be generated from the smAb using a sequencing simulator for Roche 454, Illumina, or bypassing sequencing simulator altogether. The Roche 454 simulator can introduce homopolymer insertions and deletions, represented here by blue stars and red diamonds, respectively. Alternatively, the simulated antibodies can obtain mutations along the V gene-segment, shown as green circles.

Mutation distribution

Figure A.2 shows the probability of mutating the first 275 positions of a V gene-segment. This distribution was computed from the 23,051 IMGT annotated sequences; as described in the main text. It was used for creating datasets with a given number of mutations, sampled from this distribution, without replacement.

Scoring V segments

Somatic hypermutation occurs primarily in V gene-segments, and can have an effect on their labeling. A simple scoring approach can be taken for D and J gene-segments, with

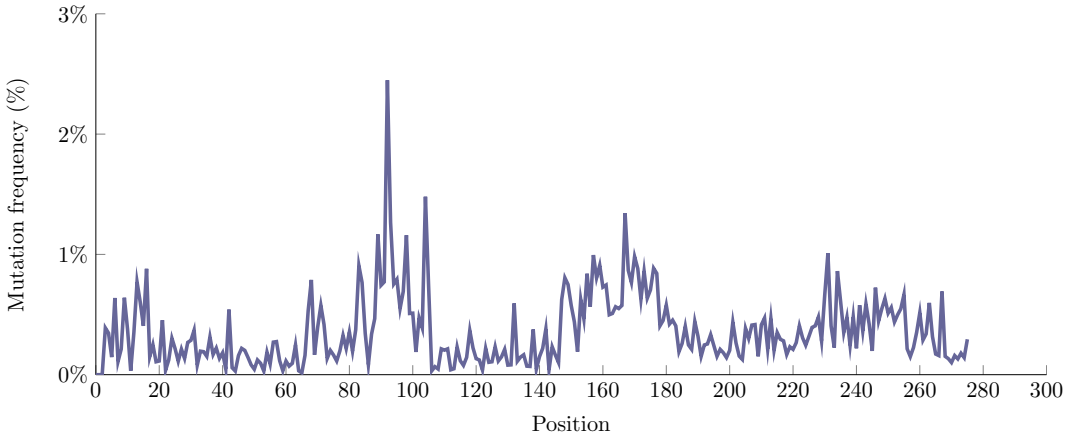


Figure A.2: Distribution of mutations along V gene-segments (CDR3 region is not included). Positions from 0 to 275 are shown on the x-axis, and the y-axis shows the mutation frequency at a given position on the human dataset described in the main text. This mutation distribution shows peaks representing CDR1 and CDR2 regions, and lower probability of mutations in framework regions. The positions are purposefully truncated since deviations from the reference are difficult to attribute to somatic hypermutation events, or to recombination events near the 3' end of the V gene-segment.

match/mismatch penalties, but when considering somatic hypermutation events, sequence properties and positional information must be taken into account. Considering only sequence properties, e.g., 4-mer motifs, ignores a strong signal that mutations occur primarily in CDR1 and CDR2. While only considering position can obviously over-call mutations as well.

We consider both 4-mer motifs and position in our scoring method. For the sake of simplicity, below we assume that the references R and read r are aligned to the same position. In reality, these can differ, an we can tolerate a shift in the read or reference. We wish to compute the probability of observing read r, given reference R, written as: $P(r|R) = \Pi_i P(a_i|b_i, i)$, for a nucleotide in the read a_i and an l-mer in the reference b_i, at position i.

If we instead consider mutations and not nucleotides a_i, then we can write the probability of a read r given a reference R as: $P(r|R) = \Pi_i P(m|b_i, i)$, for $m = \text{mut}$ representing a mutation (i.e., mismatch) and $m = \neg\text{mut}$ representing a match to reference. This turns into:

$$P(m|b_i,i) = \frac{P(b_i,i|m)P(m)}{\sum_{n\in\{\text{mut},\neg\text{mut}\}} P(b_i,i|n)P(n)} \quad \text{(A.1)}$$

where the probability of a mutation can be the aggregate of non-matching identities at the jth position of 4-mer b_i, denoted b_i^j.

Unfortunately, the data used to compute the conditional joint probability $P(b_i,i|m = \text{mut})$ is sparse when using 67,108 mutation events. To remedy this, we simplify the joint probability to $P(b_i,i|m) = P(b_i|m)P(i|m)$, resulting in the form:

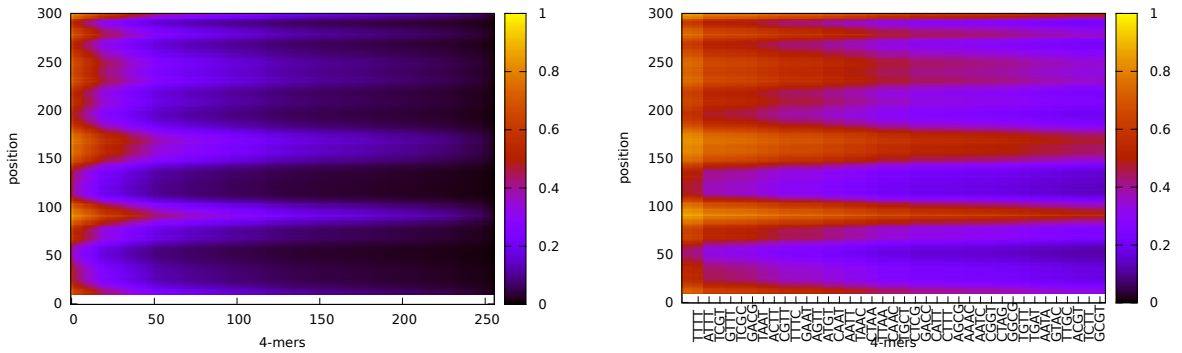
$$P(m|b_i,i) = \frac{P(b_i|m)P(i|m)P(m)}{\sum_{n\in\{\text{mut},\neg\text{mut}\}} P(b_i|n)P(i|n)P(n)} \quad \text{(A.2)}$$

whose probabilities can be easily computed from the data. This is the same relaxation that is performed for Naive Bayes classifiers to avoid the curse of dimensionality. Further, counts from data used to compute the conditional probability, $P(i|m = \text{mut})$, are modified to smooth the counts. A window of 5 downstream, and 5 upstream positions is used to compute the local average count. This smoothing is to overcome any effects caused by indels in the alignment of the IMGT database.

Plotting $P(m = \text{mut}|b_i,i)$, sorting each 4-mer according to the sum across their positions, provides Figure A.3(a). The top 40 4-mers are shown in Figure A.3(b). These plots clearly show strong favoring of some 4-mers in CDR1 and CDR2.

Alternate model

The probabilistic model used in IgGraph is detailed in the Supplemental section: Scoring V segments. This is ostensibly a Naive Bayes model, trained on the dataset described in the main text for predicting matching or mutated base pairs in the V gene-segment. For completeness,



(a) All 4-mers

(b) Top 40 4-mers

Figure A.3: Probabilities of mutation for a given 4-mer and position. Shown are all 4-mers, across positions from 10 to 300, sorted according to the sum of mutation values across all positions for each 4-mer. The mutation value in each square of the heatmap represents the probability of a mutation at that position, for that 4-mer. (a) All sorted 4-mers reveal CDR1, CDR2, and CDR3 regions. (b) The top 40 4-mers show which motifs are favored in each region.

this model's performance is shown using 5-fold cross-validation on the described dataset in Figure A.4. Performance of 4-mers, 5-mers, and 6-mers are shown. Due to a large class imbalance, $\approx 4\%$ samples are mutations, SMOTE [CBHK02] is employed to over-sample the minority class, while down-sampling the majority class. This class redistribution is performed only on each fold's training set, while the fold's test set remain unchanged.

Validating and evaluating predicted classifications

In order to assess the similarity of predicted VDJ classifications of antibody reads, we separate this task into two components: supervised and unsupervised comparisons. Supervised comparisons consist of comparing the predictions to ground truth labels. This supervised validation uses simulated reads since there is no guaranteed way of determining the true antibody references from immunoglobulin sequencing data.

There is also value in assessing the similarity of tools prediction on real data, despite the uncertainty of true labels. In this unsupervised comparison, the similarity of predictions between different tools is assessed.

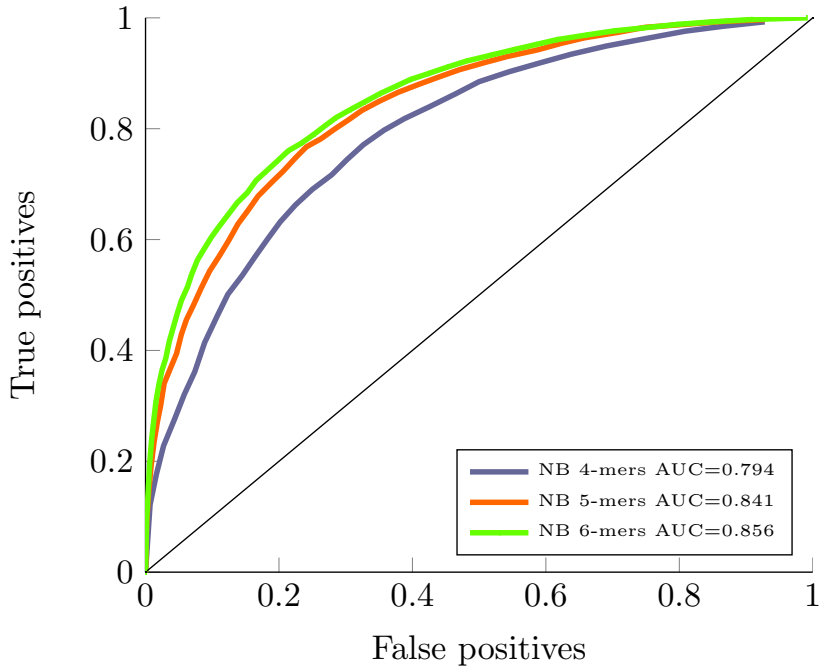


Figure A.4: Receiver operating characteristic (ROC) curve for different models. The ROC curve on 5-fold cross-validated dataset for Naive Bayes using different size k-mers as input. The training datasets are re-sampled using SMOTE [CBHK02], while the testing sets are unmodified.

Comparing the classifications of V, D, J, and total segments at the gene and allele level are performed using the Jaccard index. The junction sequence is compared for absolute equality, since this is often used to characterize the distribution of sizes across repertoires.

Additionally, comparing the clone partitioning is important to determine if different tools would cluster reads into the same, or different, clones which could drastically impact downstream analysis of clone evolution. Clone partitioning can be compared using indices for comparing clustering algorithms, specifically the Rand index, Jaccard index [Jac08], and Fowlkes-Mallows index [FM83]. These three indices can be computed by comparing pairs of points (i.e., reads), and determining if partition A has clustered them together (1) or separate (0). Similarly, a 1 or 0 can be assigned to the same pair of points for partitioning B. Thus, each pair of points can be in one of the four categories: n_{11} for A and B both placing the pair in the same cluster; n_{00} for both A and B placing the pair in separate clusters; n_{10} for A placing them together while B separates them; and n_{01} for A separating them while B places them together.

When A is viewed as the predicted clustering and B as the ground truth, n_{11}, n_{00}, n_{10}, n_{01} can be viewed as true positives, true negatives, false positives, and false negatives, respectively.

With this formulation, we can compute any of the Rand, Jaccard, or Fowlkes-Mallows indices. However, due to the nature of clone abundances in Ig-seq datasets containing many clones, this results in many clusters. The Rand index will be skewed due to the high number of different clusters. As such, the Fowlkes-Mallows (FM) index [FM83] provides us with an alternative. The FM index is computed as:

$$FM = \frac{n_{11}}{\sqrt{(n_{11}+n_{01})*(n_{11}+n_{10})}} \tag{A.3}$$

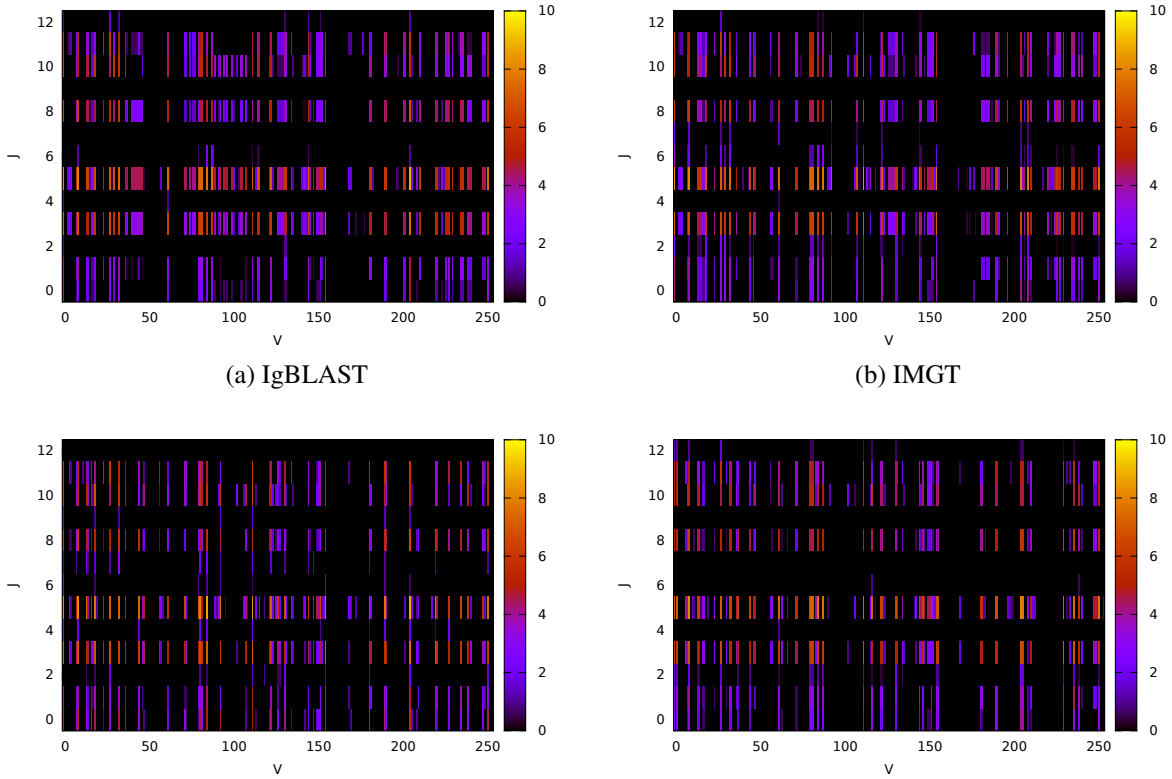
and can be interpreted as the geometric mean of the precision in recall in a supervised setting. In our setting the geometric mean can normalize differences in scale between the two partitionings, while still providing a similarity comparison. A tool for simplified comparison of predicted labels, partition ranges, and clonal clusterings is provided as `IgVALVE` (Ig Vdj Antibody Labeling Validation and Evaluation). `IgVALVE` will compute accuracy in a supervised setting, i.e., when using simulated reads; or can compare predictions using the Jaccard index and Fowlkes-Mallows index, as described above, in an unsupervised setting when labels are not available. The validation and evaluation tool `IgVALVE` is available for use at: https://bitbucket.org/sbonisso/ig_valve

Stanford_S22 VJ analysis

Analyzing the VJ pairings of labeled heavy/light chains can provide an idea of the distribution for the selection of which V and J are favored within a population of B-cells. When comparing different V/D/J labeling tools, this can help show the differences in labeling. The Stanford_S22 dataset from [JBGC10] is used to highlight differences between tools. The predictions from IgBlast, IMGT, and SoDA are taken from the online resource [1], and IgGraph was run with $k = 21$ since only V and J gene-segments were sought. This returns slightly

[1] http://www.emi.unsw.edu.au/~ihmmune/IGHUtilityEval/eval_help_datasets.php

better results for V/J gene-segments as a longer k is more robust for the longer segments, but obviously, misses D segments that are smaller than this parameter. The values in Figure A.5 below are computed by counting VJ pairs for which a prediction for both exists, and if more than one gene-segment is predicted, each predicted label is given an equal proportion of the read. Meaning, if three V segments are predicted for a single read, each obtains 0.33 points. This seeks to normalize for predicting many gene segments. The values for each VJ pair are then log2 transformed for easier comparison.



(a) IgBLAST

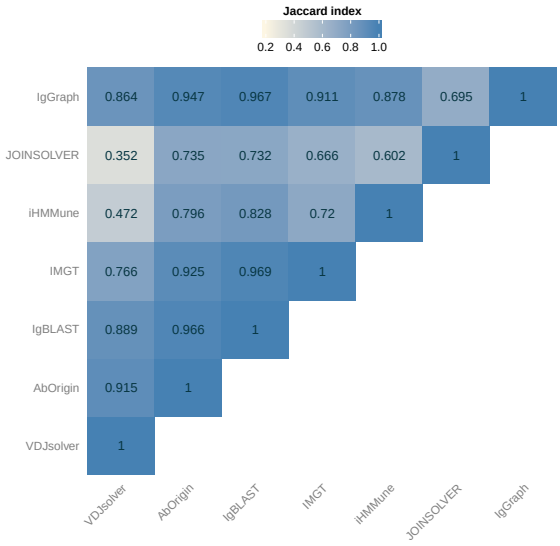
(b) IMGT

(c) SoDA

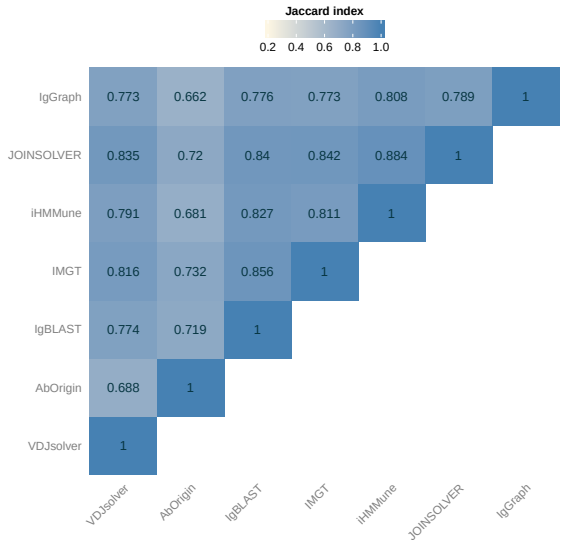
(d) IgGraph

Figure A.5: VJ pairs for IgBlast, IMGT, SoDA, and IgGraph. The V and J gene-segments are index on the x-axis and y-axis, respectively. Each cell represents a V-J pair, and shows the log transformed count indicated by the colorbar.

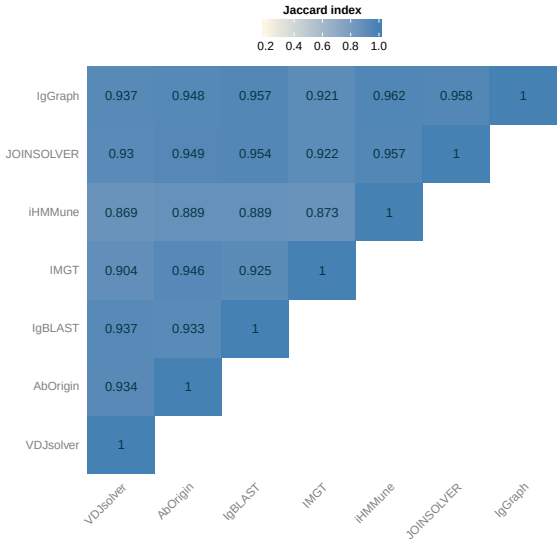
Figure A.5 shows the distributions of VJ pairings for various tools. One element to note is how IgBLAST, seen in Figure A.5(a), dilutes its predictions across multiple allelic variants.



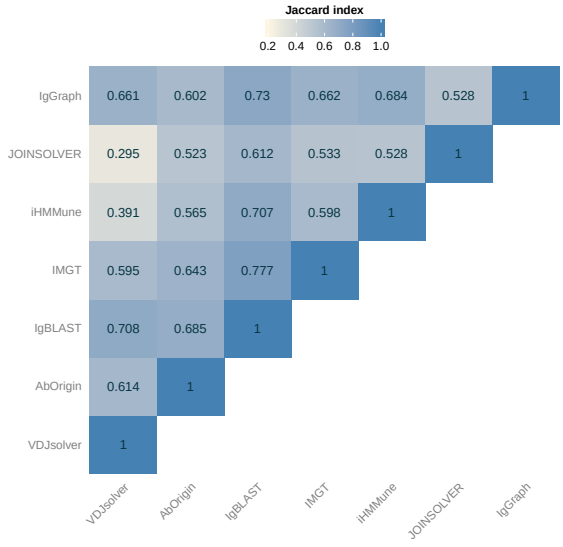
(a) V



(b) D



(c) J



(d) total

Figure A.6: Comparison of predictions from different tools on the Stanford_S22 dataset. Similarity is measured by the Jaccard index in predictions of (a) V, (b) D, (c) J, and (d) total.

Benchmarking

Table A.1 compares performance of various tools for VDJ classification and illustrates that IgGraph performs well for all gene-segments. While the error percentage is higher for V gene-segments, this could potentially be further improved with a more sophisticated scoring model than the one we employed. One detail to note, is that, like for most other VDJ classification tools, the majority of errors are mispredictions of allelic variants. These types of errors particularly difficult to distinguish, but our approach (along with most others) is able to correctly identify the correct genotype. An example of a typical error is shown in Supplementary Figure A.10.

Table A.1: Table of error percentages on Stanford_S22 dataset reproduced from [JBGC10] with the colored antibody graph (IgGraph) appended. The errors shown are the percentage of incorrect allelic variant reported, and the percentage of incorrect gene reported; a rarer event than incorrect allelic variant. The total column represents the percentage of sequences that include an incorrect gene or allele for either the V, D, or J gene-segments. The results for IgGraph shown are with $k = 11$ and $m = 2$.

		Alleles				Genes		
Citation	Utility	IGHV	IGHD	IGHJ	Total	IGHV	IGHD	IGHJ
[GMJ+07]	iHMMune-align	3.21	2.21	1.95	7.11	0.21	1.27	0.0
[BLG08]	IMGT	4.90	5.09	1.55	10.87	0.22	2.81	0.0
[YMMO13]	IgBLAST	3.84	3.96	0.85	8.39	0.75	2.16	0.0
[WWZ+08]	Ab-origin	4.06	7.94	2.53	13.74	0.22	5.53	0.0
[SCLR+04]	JOINSOLVER	6.17	6.93	1.24	7.89	0.86	4.92	0.0
[VCK06]	SoDA	2.68	6.82	1.50	10.37	0.29	6.63	0.0
[OLNLB06]	VDJSolver	6.87	1.96	0.71	9.09	0.48	0.79	0.0
	IgGraph	5.47	0.93	0.65	6.07	0.15	0.82	0.0

Table A.2 benchmarks performance of IgGraph for different values of k-mer sizes (k) along with different sizes of l-mers used for scoring. While increasing k improves performance labeling V gene-segments, the l-mer used influences the performance. Similarly to Figure 3.1, the left part of the table referes to 213 alleles, while the right part refers to 55 genes.

Simulated dataset VJ distribution

The simulated dataset with no mutations, described in the main text, is visualized in Figure A.7 by showing the distribution of counts for all VJ pairs.

Table A.2: Table of error percentages for V gene-segments on Stanford_S22 dataset for different parameterizations of IgGraph, and different sized l-mers for scoring. Errors involving an incorrect gene, and errors for an incorrect allelic variant, are shown.

	Alleles			Genes		
Parameters	4-mers	5-mers	6-mers	4-mers	5-mers	6-mers
$k = 11$	7.65	5.47	8.90	0.12	0.15	0.74
$k = 15$	6.32	4.29	7.67	0.09	0.10	0.68
$k = 21$	5.94	3.95	7.51	0.09	0.08	0.68

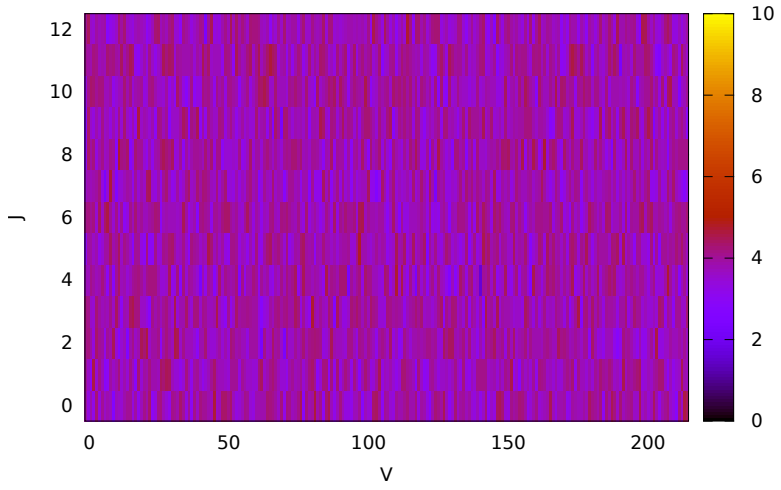


Figure A.7: VJ pairs represented in the simulated antibody dataset containing no mutations. The V and J gene-segments are index on the x-axis and y-axis, respectively. Each cell represents a V-J pair, and shows the log transformed count indicated by the colorbar.

Threshold determination

To select the maximal number of labels to return for each gene segment, the mean accuracy of labelings, varying the threshold for maximal labels, m, is plotted. Providing an additional option allows for improved accuracy, without providing a long candidate list. Manual inspection showed that these improvements are typically in cases where the gene segment is indistinguishable from a different allelic variant. Based on this plot, we select a threshold of 3 for human.

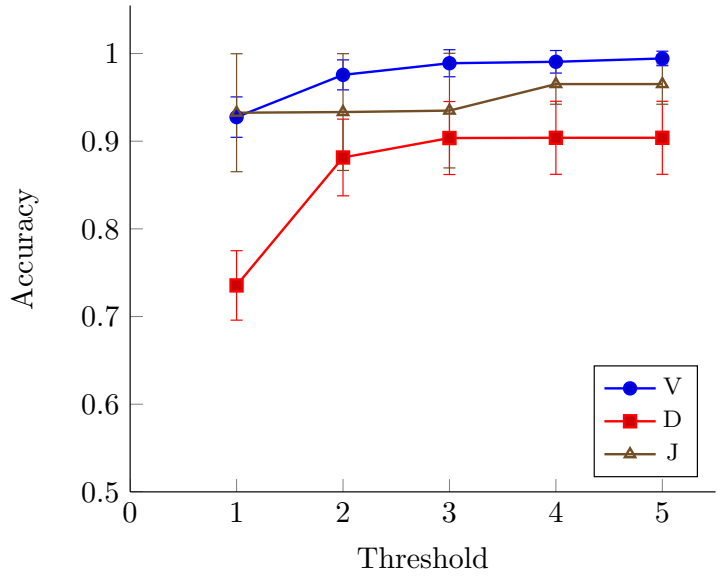
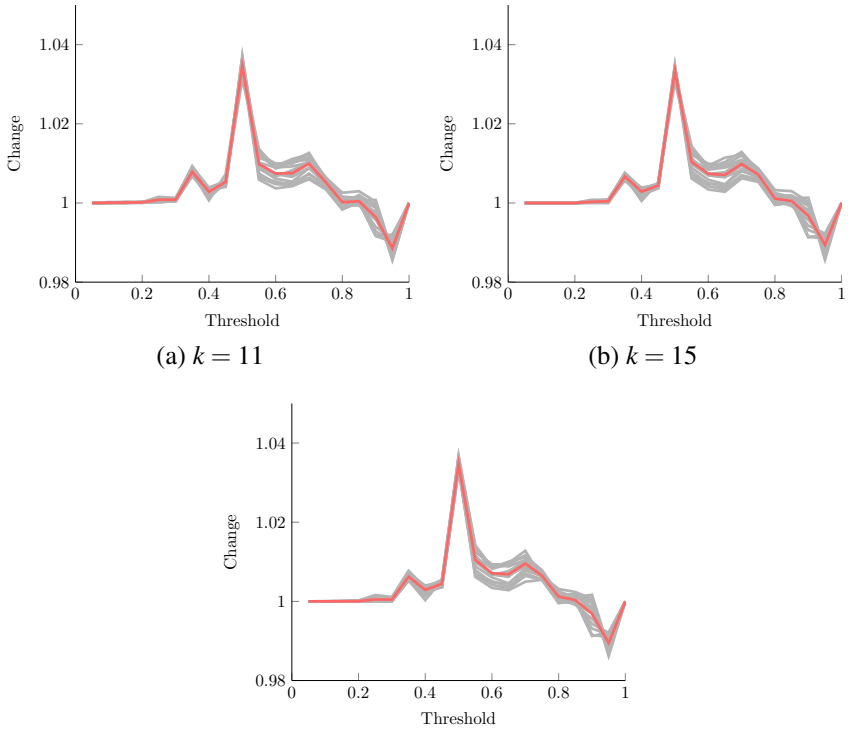


Figure A.8: Mean accuracy of labeling each class of gene segments on human dataset using colored antibody graph with $k = 11$. Providing even a single alternative explanation, i.e. a threshold of 2, improves the accuracy greatly, particularly for D gene segments, while there are diminishing returns for additional options.

The threshold m determines the maximal number of gene-segment labels to report, but we allow for 0 to m labels to be reported for each read. This is based on the probabilistic score of each label. Rather than simply selecting the top m scoring labels, we select the top m scoring labels that cumulatively exceed a threshold t. This means that if there are $m+1$ labels that cumulatively do not exceed t, no label is reported. Conversely, if a single label exceeds t, despite $m = 2$, only that single label is reported. This threshold t is not a user tunable parameter, and was selected by identifying the threshold at which there were diminishing

returns correctly identified labels. Figure A.9 shows, for each threshold, the ratio between the number of correctly identified gene-segments at that threshold compared to the previous threshold. The threshold when the delta plateaus and/or drops below 1.0 is how a $t = 0.8$ is selected for use in all datasets; all mutation datasets with an even number of mutations and the Stanford_S22 dataset.. The curves are generated over the mutation datasets with an odd number of mutations, each curve in gray, and their average shown in red.



(a) $k = 11$ (b) $k = 15$

(c) $k = 21$

Figure A.9: Change in correctly labeled samples when varying the threshold. Each gray curve represents a single dataset with a fixed number of odd mutations, and the red curve is the mean across all used datasets. Thresholds are varied and the change in number of correctly labeled samples is shown on the y-axis. The threshold at which the change plateaus or drops below 1.0 is selected.

Analysis of common errors

One common error IgGraph commits on the Stanford_S22 dataset is calling IGHV3-33*01 as IGHV3-33*06, an example of this error is shown in Figure A.10 below. The beginning of the read matches perfectly with IGHV3-33*01, at position 166 of the read, where the SNP that distinguishes IGHV3-33*01 from IGHV3-33*06 occurs, the nucleotide matches IGHV3-33*06. This is labeled as IGHV3-33*01, however, the 4-mer surrounding the mismatch is not among the most prevalent 4-mers from the training dataset, and as such, receives a low probability. Disambiguating such a labeling is difficult to perform, with significant confidence, in favor of either gene-segment.

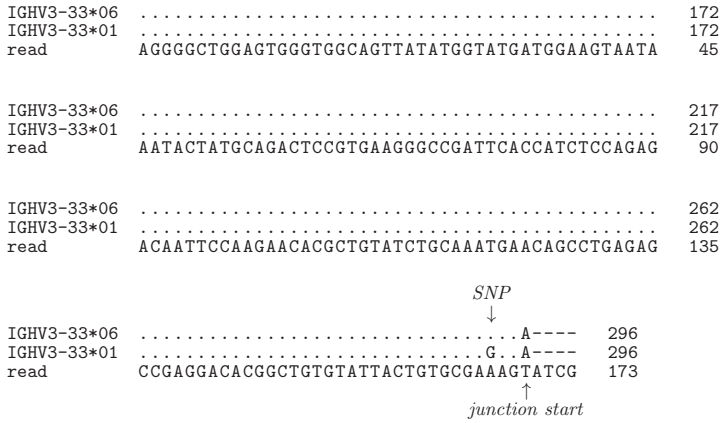


Figure A.10: Alignment of read and reference sequences. This alignment shows a read from the Stanford_S22 dataset and reference gene-segments IGHV3-33*01 and IGHV3-33*06. This read has ground truth labeling as IGHV3-33*01, and IgGraph assigns it as IGHV3-33*06. Distinguishing between these two is difficult due to the SNP at position 293 of IGHV3-33*01.

Selectng the *k*-mer size

The selection of *k* can have consequences on performance, and must be chosen carefully. Here, we show how *k* can negatively impact the antibody graph by creating cycles, or loops, bridging two segments of the antibody that are far away from each other. For example, if a (*k*-1)-mer is shared between a V segment and a J segment, this can bridge the two gene-

segments and cause difficulties in assigning a label. Figure A.11 shows the mean number of shared k-mers across all 99,450 human, 4,290 rabbit, and 52,272 mouse combinations of VDJ gene-segments.

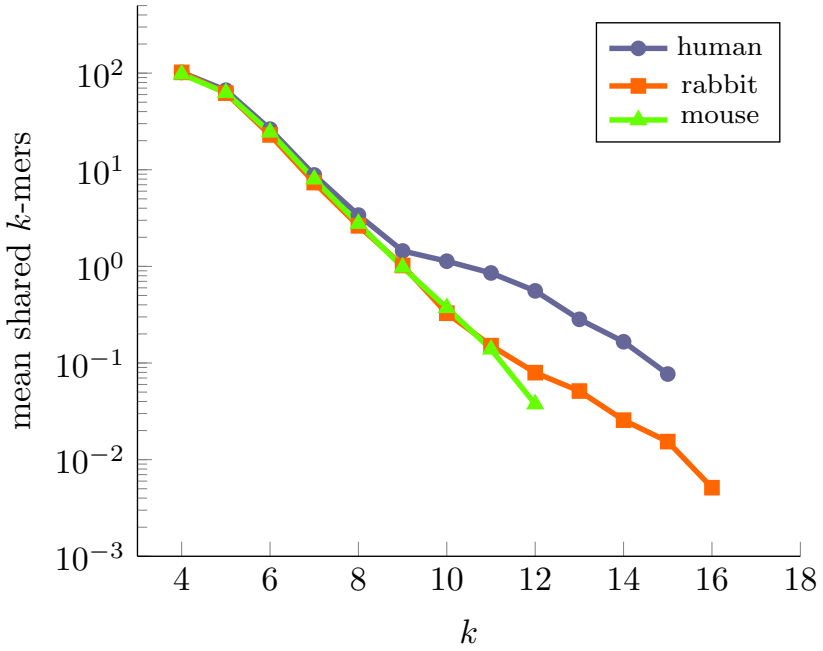


Figure A.11: The effect of varying k over the average number of shared k-mers per antibody transcript. All values of k between 4 and 25 where computed, and each curve is truncated at the value of k where all k-mers are unique within each transcript.

Each curve ends where there no longer exist any shared k-mers among any possible re-arrangement of VDJ gene-segments. This shows that a selection of k=18 would ensure no sharing for any dataset, this selection for k will miss some D gene-segments in human and mouse. For this reason, we are willing to tolerate some redundancy and can see that smaller selections for k can be tolerated. This plot is a simplification of the scenario since it does not take into account non-templated nucleotides and how they may affect k-mer sharing. This is particularly possible since these nucleotides are GC rich.

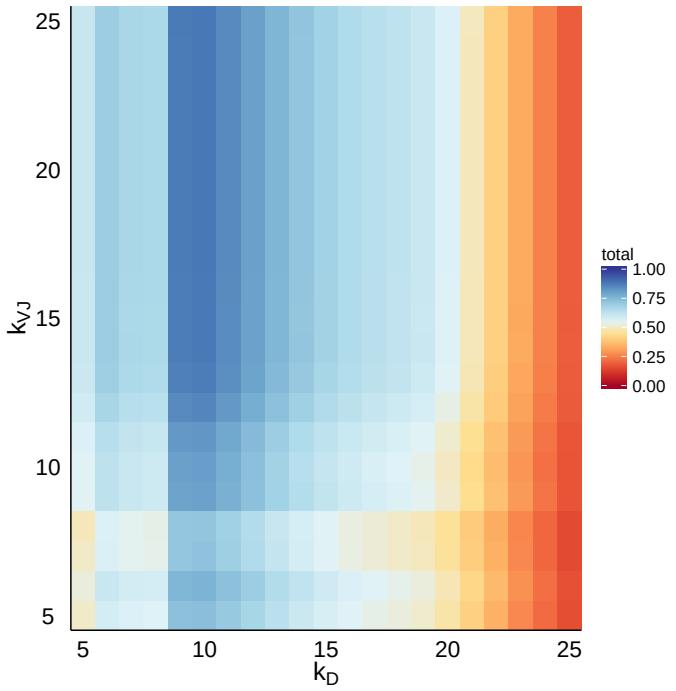


Figure A.12: Accuracy over alleles for different values of k for V/J (k_{VJ}) and D (k_D) gene-segments. Accuracy for total V, D, and J segments over a dataset of 2000 simulated antibody reads is shown. Standard scoring was used.

Appendix B

Supplement to Immunoproteogenomics

Supplemental figures

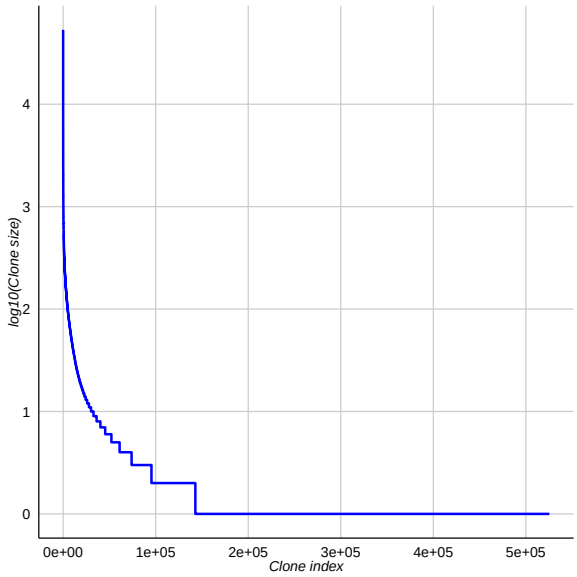


Figure B.1: Clone size distribution of Ig-seq.

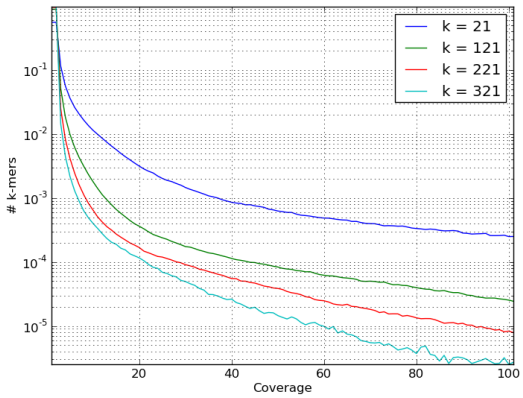


Figure B.2: Histograms of k-mer coverage distribution for $k = 21, 121, 221$, and 321 illustrate that the coverage of the antibody repertoire by short k-mers is orders of magnitude higher compared to the coverage by long k-mers. Note that the y-axis is given in logarithmic scale.

```
s1 CAGGTCTGATGCAGTCTGGGACTTAGCTGGGGCGCT
s2 CAGGTCTGGTGCAATCTGGGACTGAGCTGGGTCGCT

d(s1, s2) = 3
```

(a)

```
s1 CTCAGGTCTGATGCAGTCTGGGACTTAGCTGGGGCGCT
s2   CAGGTCTGGTGCAATCAGGGACTGAGCTGGGTCGCTA

d̃(s1, s2) = 4
```

(b)

Figure B.3: Hamming distance $d(s_1, s_2)$ (a) and generalized Hamming distance $\tilde{d}(s_1, s_2)$ (b). Note that $d(s_1, s_2)$ and $\tilde{d}(s_1, s_2)$ for sequences s_1 and s_2 of equal length are not necessarily the same.

The Hamming graph bound selection

The important question is how to select the bound τ while constructing the Bounded Hamming Graph $HG(Strings, \tau)$. The input to IGREPERTOIRECONSTRUCTOR is overlapping paired-end reads that are merged into single reads covering the variable region of the antibody (about 400 nt). If sequencing errors in reads (the error rate in Illumina reads is $\approx 1\%$) are not corrected by merging, the merged read are expected to contain ≈ 4 errors on average. Thus, unless the rate of sequencing errors is reduced by the merging procedure, two merged reads are

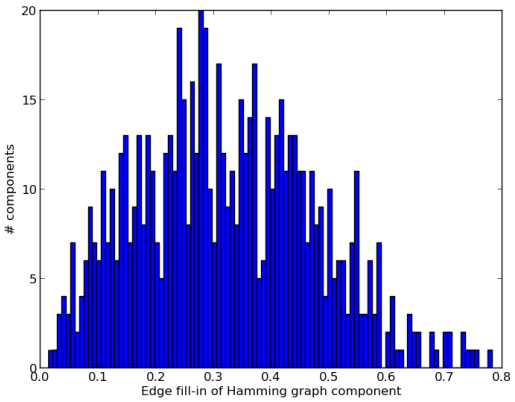


Figure B.4: Histogram of the distribution of edge fill-ins computed for 721 large ($\geq$ 100 vertices) connected components of the Bounded Hamming graph with $\tau = 3$. The average size of components with edge fill-ins > 0.7 is 151.

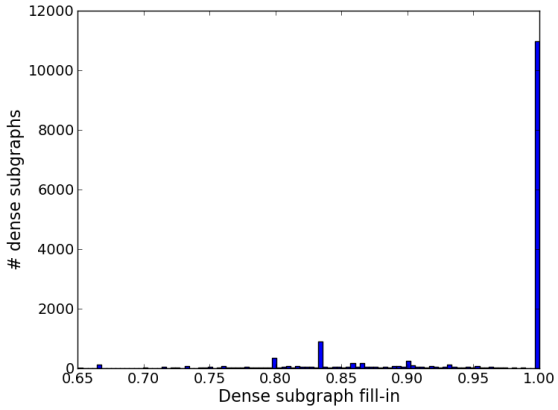


Figure B.5: Histogram of edge fill-in distribution for non-trivial dense subgraphs constructed from 721 large ($\geq$ 100 vertices) connected components. The total number of the constructed non-trivial dense subgraphs is $15,996$.

expected to have $\approx 4+4 = 8$ mismatches on average. Unfortunately, the threshold $\tau = 8$ will not work for error-correction in immunosequencing since different antibodies often differ by less than 8 mismatches (Figure B.10).

However, it turns out that our algorithm benefits from the fact that most errors are concentrated at the ends of reads resulting in merged reads with significantly higher accuracy than the accuracy of the original paired-end reads (see Appendix B). To estimate the average

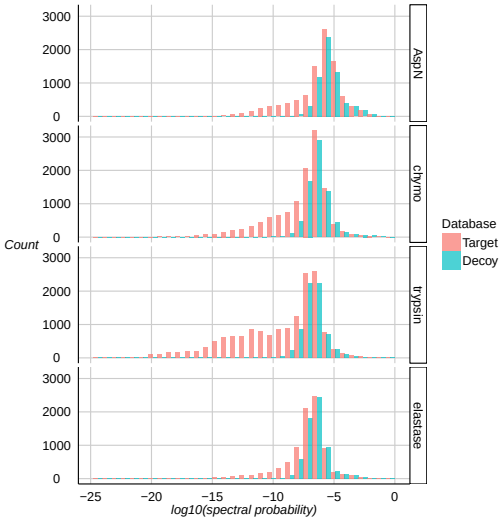


Figure B.6: Spectral probability distributions of target/decoy identifications for each spectral dataset.

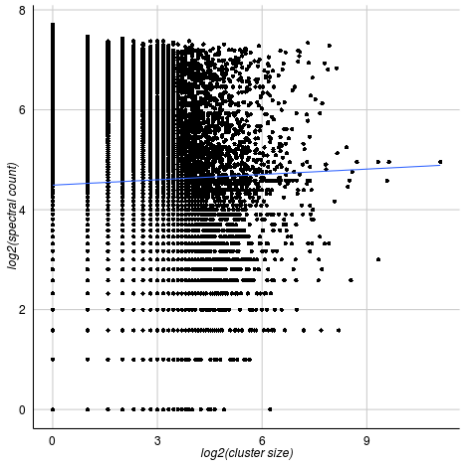


Figure B.7: Scatterplot of NGS-based and MS-based abundances of antibodies (cluster size vs. spectral count). Pearson correlation $\rho = 0.007544031$.

number of errors in the merged reads, we extracted reads corresponding to the known contaminant (*Streptococcus pneumoniae*) in our Ig-seq dataset and aligned them to the *Streptococcus pneumoniae* genome. It turned out that 96% of merged reads differ from the reference genome by at most 1 mismatch (98% of merged reads differ from the reference genome by at most 2 mismatches). Thus, we have selected the bound $\tau = 3$ for constructing the Bounded Hamming Graph.

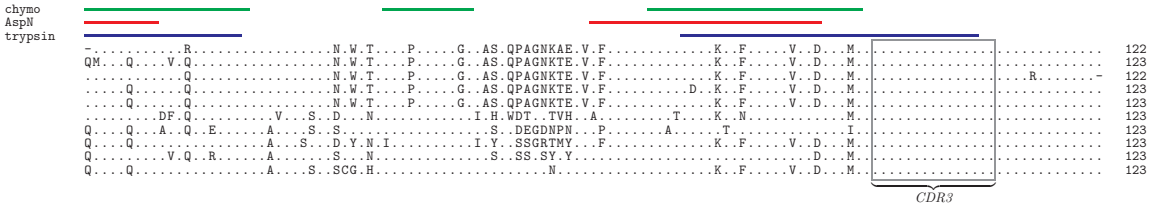
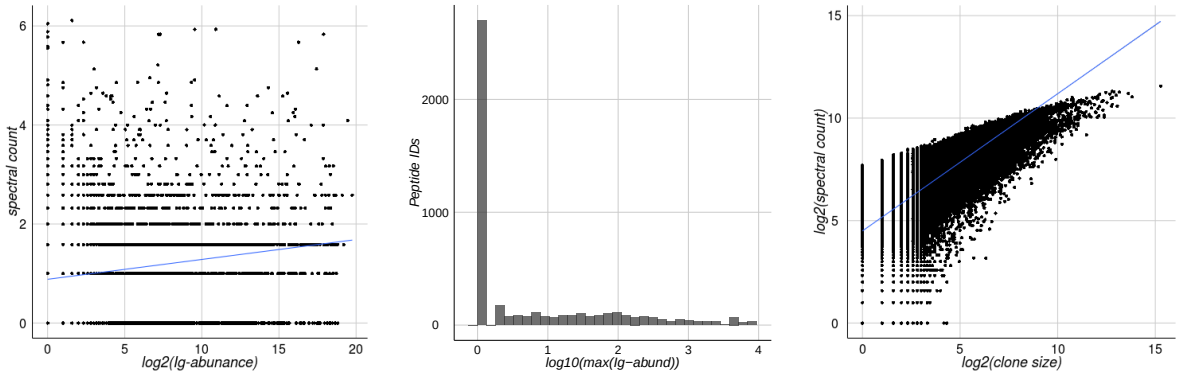


Figure B.8: Alignment of sequences of a single clone with peptide evidence. The 11 antibodies with most peptide evidence within this clone are aligned to one another to show the sequence diversity, while the antibody with most peptide evidence is omitted. Identified peptides for the omitted sequence are shown above the alignment. The CDR3 region is noted, and shown in gray box. Positions differing from the sequence with peptide evidence shown are noted, positions agreeing with the omitted sequence are shown as a dot.

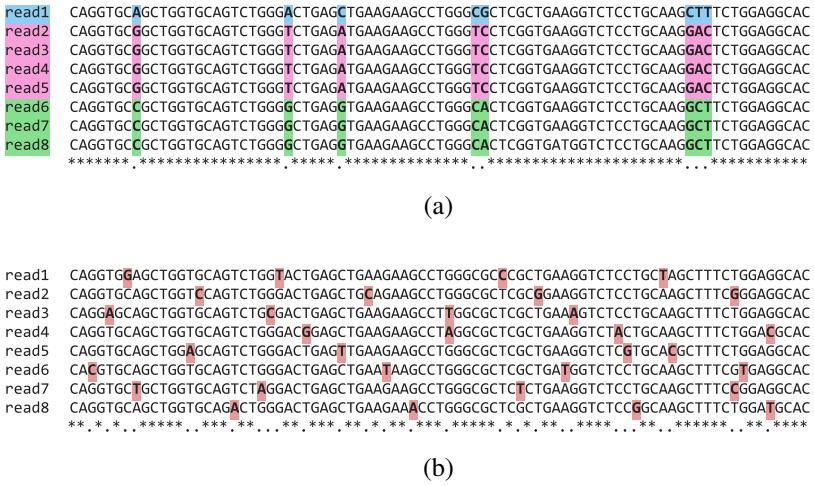


(a) (b) (c)

Figure B.9: Scatterplot of genomics-based and proteomics-based abundances. Cluster size compared to spectral counts of each cluster; Pearson correlation $\rho = 0.007544031$. (a) The total Ig-abundance compared against the spectral count for all peptide. Pearson correlation $\rho = 0.1724002$. (b) Histogram of peptide IDs by the maximal Ig-abundance of each peptide. (c) Spectral count of each clone, related to clone size; Pearson correlation $\rho = 0.5687614$. The spectral count of a clone is the total number of PSMs originating from all antibodies within that clone. Normalization of spectra for shared peptides is not performed in these plots.

Dense subgraphs and cliques in triangulated graph

Fig. B.11a shows a triangulated graph G containing three dense subgraphs, yellow, green, and violet. Fig. B.11b shows a *clique overlap graph* of G where vertices correspond to maximal cliques in G and edges connect cliques that share vertices. The weight of an edge



(a)

(b)

Figure B.10: (a) An alignment of reads corresponding to three similar antibodies (highlighted in blue, violet, and green). For the sake of simplicity, we show error-free reads. For comparison, (b) shows the alignment of reads originating from the same antibody, that contain randomly located errors in reads.

is the number of vertices shared between two cliques. A *clique tree* is defined as a maximum spanning tree of the clique overlap graph (Fig. B.11c). Fig. B.11d show a perfect elimination order for the graph from Fig. B.11a. Fig. B.11e shows the vertex elimination process for the same graph.

Dense subgraphs are formed by mutliple cliques in the clique tree. For example, vertices from the dense yellow subgraph can be found among four nodes from the the clique tree. To construct dense subgraphs, we merge maximal cliques connected by many edges. E.g., we merge two yellow cliques resulting in a subgraph on 5 vertices with a high edge fill-in.

Note that some of the resulting dense subgraphs may share vertices forcing us to assign these shared vertices to one of the dense subgraphs. To assign each vertex v to a single dense subgraph, we select a the subgraph that has the maximum number of vertices adjacent to v among all subgraphs containing v.

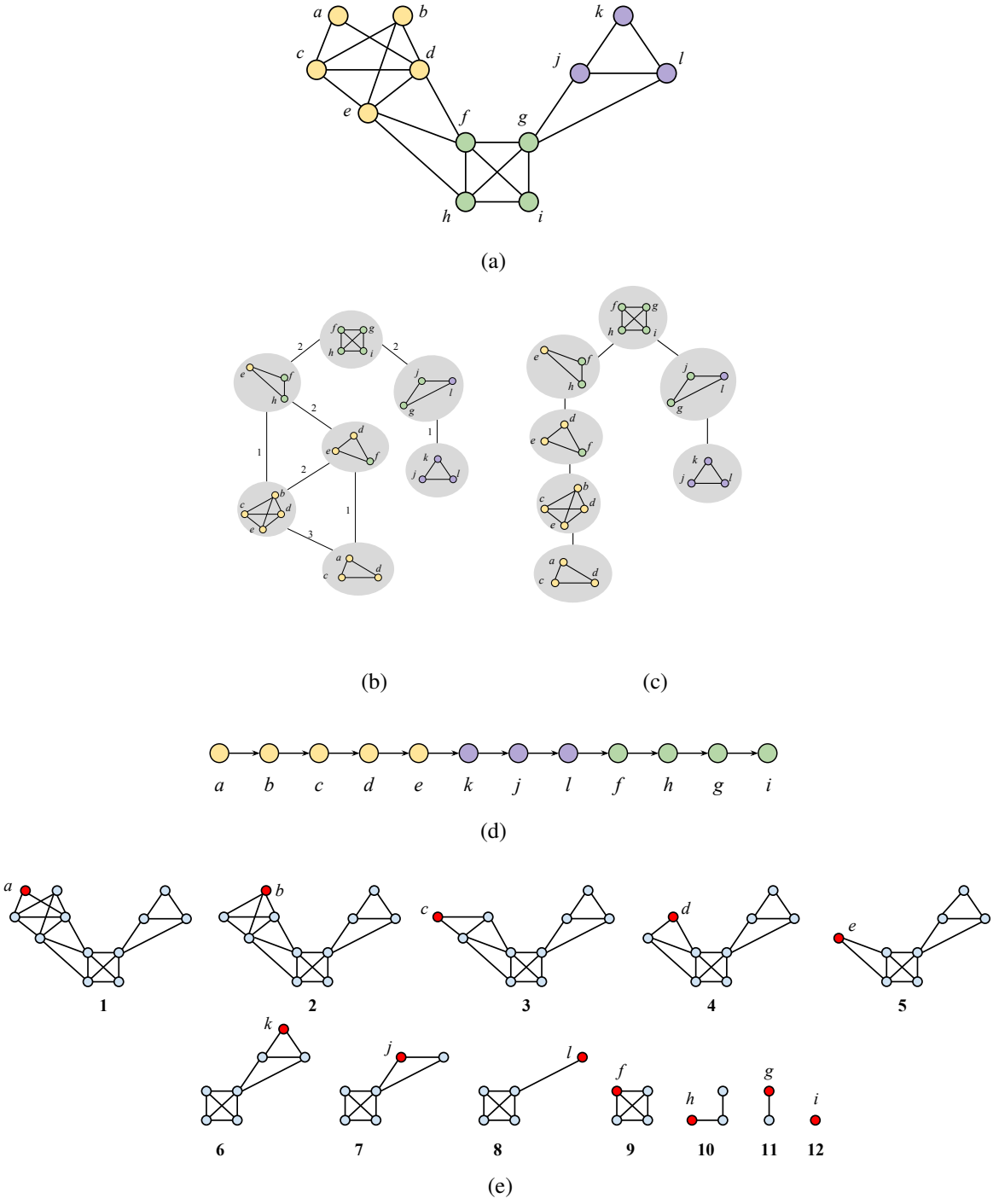


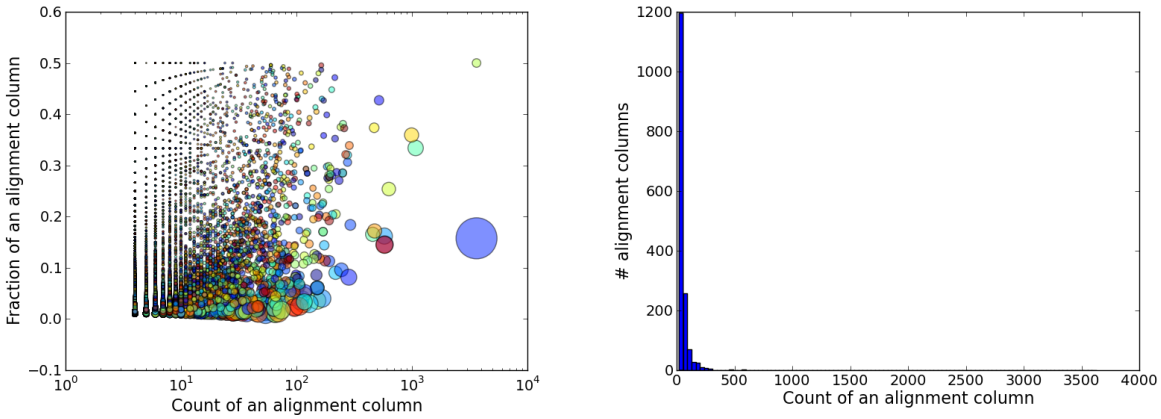
Figure B.11: A triangulated graph (a), its clique graph (b) and its clique tree (c). (d) perfect elimination order for the triangulated graph in (a). (e) the vertex elimination process for the triangulated graph in (a).

Splitting dense subgraphs using SHM detection

In practice, SHMs associate with various patterns ([DBB$^+$97]) making it difficult to apply the approach for breaking dense subgraphs described in the main text. To bypass this complication, we define the notion of a *mutation-edge-set* as the set of all edges corresponding to a given mutation. We further define an SHM as a mutation whose mutation-edge-set splits the subgraph (cluster of reads) into relatively large sub-clusters. Thus, IGREPERTOIRECONSTRUCTOR attempts to split each constructed dense subgraph by identifying SHMs. The split subgraphs revealed by this final step define the reads contributing to each antibody in the antibody repertoire. Below we explain how IGREPERTOIRECONSTRUCTOR splits dense subgraphs by identifying SHMs.

To design our splitting rule, we aligned reads from each dense subgraph. For each column i in the alignment, we define $count_i$ and $fraction_i$ and the count and fraction of *second* most frequent nucleotide in the i-th column. We define thresholds $count_{min}$ (the default value is 4) and $fraction_{min}$ (the default value 0.01) and limit our attention to all columns with surprisingly fractions of the second most frequent nucleotides ($fraction_i > fraction_{min}$) among columns where there is substantial number of this nucleotides ($count_i > count_{min}$). We further refer to such columns as *SHM columns* and split all dense subgraph that have SHM columns. While it may appear that the default value $count_{min} = 0.01$ is too small to distinguish SHMs from sequencing errors, we note that it applies to *stitched* Ig-seq reads that feature rather small error rates (0.0022 on average).

Overall, we detected $18,097$ such columns distributed over 4126 dense subgraphs constructed by IGREPERTOIRECONSTRUCTOR (25.79% percent of all dense subgraphs). Figure B.12a presents the scatter plot of $(count_i, fraction_i)$. Figure B.12b shows the histogram of the distribution of the $count_i$ values.



(a) (b)

Figure B.12: (a) the scatter plot of pairs $(count_i, fraction_i)$ for filtered alignment columns using thresholds $count_{min} = 4$ and $fraction_{min} = 0.01$. The area of a circle is proportional to the value of $count_i$. Colors of circles are individual for each alignment column. (b) the histogram of the distribution of the $count_i$ values.

Evaluating the constructed repertoires

We evaluate the constructed repertoire by checking whether the Ig-Seq reads from the same cluster exhibit variations typical for errors in reads (as expected from correctly constructed clusters) or variations typical for incorrectly constructed clusters formed by multiple antibodies. In order to analyze the pattern of variations, we align reads from each cluster and compute the distribution of positions of mismatches along the length of the reads. If this distribution is roughly uniform (as expected from sequencing errors), we conclude that the cluster is constructed correctly. However, if this distribution reveals some peaks (e.g., peaks in CDR regions), we conclude that two different antibodies were merged into a single cluster.

Fig. B.13 shows a histogram of mismatch positions averaged over all clusters. Since this distribution is rather uniform (except for peaks in the beginning and end of reads typical for the error profile of Illumina reads), we conclude that most clusters correspond to a single antibody. In contrast, the distribution of variations for antibody clones (groups of multiple antibodies with the same CDR3) shows pronounced peaks in CDR1 and CDR2 regions indicating that the

clones are formed by multiple antibodies (Fig. B.17).

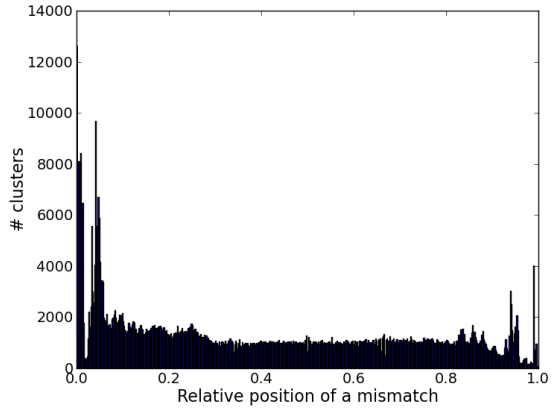


Figure B.13: Histogram of distribution of the relative mismatch positions for the constructed antibody clusters.

Benchmarking IGREPERTOIRECONSTRUCTOR on simulated immunosequencing data

In order to check accuracy of IGREPERTOIRECONSTRUCTOR, we generated small simulated immunosequencing data set using IGSIMULATOR ([SLL15]) with the following parameters: *# base sequences* $= 10,000$, *# mutated sequences* $= 100,000$ and *expected repertoire size* $= 1,000,000$. The simulated repertoire contains $105,438$ clusters (size of the maximal cluster is 112, number of singletons is $10,025$). Experiments showed that IGREPERTOIRECONSTRUCTOR accurately recovers clusters in the simulated repertoire except for several small clusters broken into singletons in the constructed repertoire.

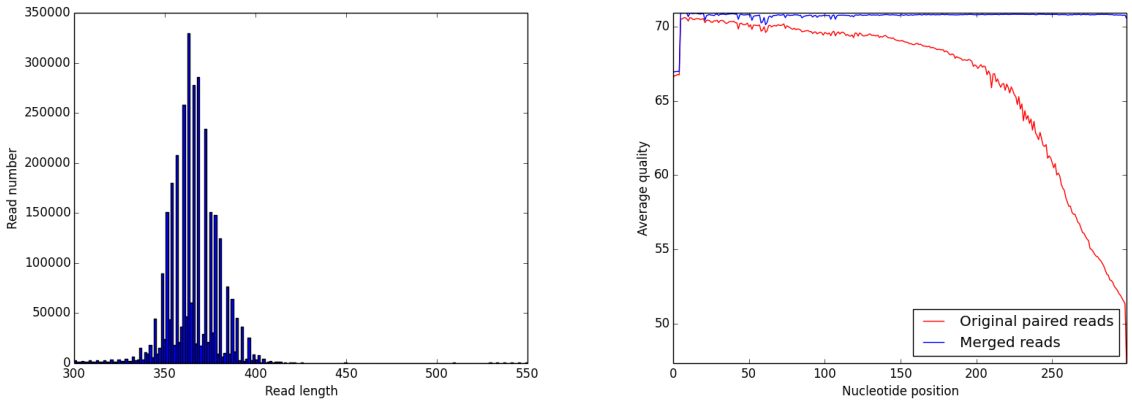
Ig-seq data preprocessing

Read merging. IGREPERTOIRECONSTRUCTOR works with single reads that cover the entire variable regions of antibodies. These reads are generated by merging the paired-end Illumina

reads.

Since paired reads in our dataset have average insert size 366 nt (Figure B.14a), they are expected to overlap by $\approx 250 + 250 - 366 = 134$ nt. After finding the overlap, we merge the two reads within a read-pair into a single merged read that is expected to cover the entire variable region of antibody. This procedure results in a significant reduction of error rates. Since the accuracy of Illumina reads drops towards the end of reads, we take advantage of the overlapping region and form its consensus by selecting the nucleotide with maximal quality value at each position in the overlap. Figure B.14b shows the difference in error rates before and after merging.

Contamination clean-up. We used IgBlast ([YMMO13]) to align merged reads to Ig germline database and removed reads that have alignment with E-value exceeding 0.001. We further filtered reads and assembled them with SPAdes assembler ([BNA$^+$12]) resulting in 10 contigs (Table B.1). Blast alignments of constructed contigs show that filtered reads were correctly classified as contamination and can be safely removed from the Ig-seq library.



(a) (b)

Figure B.14: (a) shows histogram of merged read length distribution. (b) shows the average quality of reads before (red) and after (blue) merging, and illustrates that merging of overlapping reads significantly improves their quality.

Table B.1: Contigs assembled from reads filtered as contaminants. The table shows length, coverage and the best Blast alignment for each contig from assembly.

ID	*Length (nt)*	*Coverage*	*Blast alignment*
1	1512	1.2	Escherichia coli genome assembly FHI92
2	1195	25.4	Homo sapiens major histocompatibility complex
3	959	1.9	Escherichia coli genome assembly FHI89
4	929	1.0	Homo sapiens protein tyrosine phosphatase
5	827	29.0	Homo sapiens O-sialoglycoprotein endopeptidase
6	868	1.3	Escherichia coli genome assembly FHI89
7	780	1.3	Homo sapiens immunoglobulin heavy locus (IGH)
8	734	1.0	Homo sapiens B lymphoid tyrosine kinase (BLK)
9	722	31.0	Homo sapiens uncharacterized LOC102725417
10	240	14.7	Homo sapiens long non-coding RNA

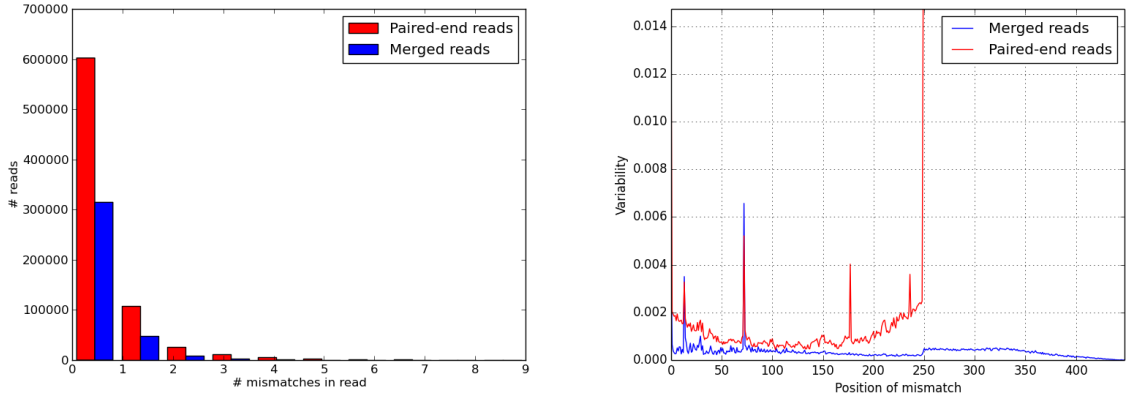
Contaminated reads analysis

We used the fact that some our Ig-seq libraries contain contaminants to estimate the average error rate of the paired-end and merged Ig-seq reads. We identified reads corresponding to the genome of *Streptococcus sp. VT 162* and aligned both paired-end, and merged, reads to the reference genome. The average number of mismatches per read is 0.85 and 0.22 for the paired-end and merged reads, respectively. Figure B.15 shows that the overlapping parts of the merged reads contain fewer errors as compared to paired-end reads.

Blind modification search results

Figure B.16 shows the prevalence of modifications over different peptide sets; those observed only with a modification Figure B.16a and B.16b, and those observed both with and without the modification Figure B.16c and B.16d.

MODa identified many PTMs with offsets -17/-18, +16, and +42Da. The 17Da and 18Da losses can be explained as pyroglutamic acid, occurring on Q and E, particularly on the N-terminus. The 16Da gains are nearly all located on methionine and tryptophan, consistent with oxidation. While some of the 42Da gains occur on serines consistent with acetylation, the majority occur on cystines. These are likely the result of N-isopropyl iodoacetamide



(a) (b)

Figure B.15: Analysis of the error rate of Ig-seq libraries using reads from contaminants. (a) shows the histogram of mismatches number per read distribution. (b) shows the histogram of mismatches position distribution. Since the length of merged reads is not fixed, we compute the position of the mismatch as the distance to the nearest start from the overlapping reads. Thus, positions of mismatches are normalized from 1 to 250.

(NIPCAM), since cystines are searched with a fixed +57Da offset.

The PTMs with offset +1Da identified by MODa are largely attributed to asparagine (N), which could signify a mutation to isoleucene/leucine (I/L). MODa found many modifications on tryptophan (W) centered around offsets of +32Da, +16Da, and +5Da, all of which can be attributed to oxidations of tryptophan. Additional prominent offsets were +12Da gain on glycine (G), and -9Da loss on arginine (R), both seen in Figure B.16b. A 9Da loss on R can be explained as a mutation to phenylalanine (F). However, the addition to G cannot be explained with common modifications or mutations.

Validating antibody repertoires

The effect from IGREPERTOIRECONSTRUCTOR is evident when comparing the peptides identified from the unique reads database and the antibody repertoire. Only 0.6% of peptides identified from the unique reads database do not appear in the antibody repertoire. This demonstrates that IGREPERTOIRECONSTRUCTOR rarely over-corrects reads implying that

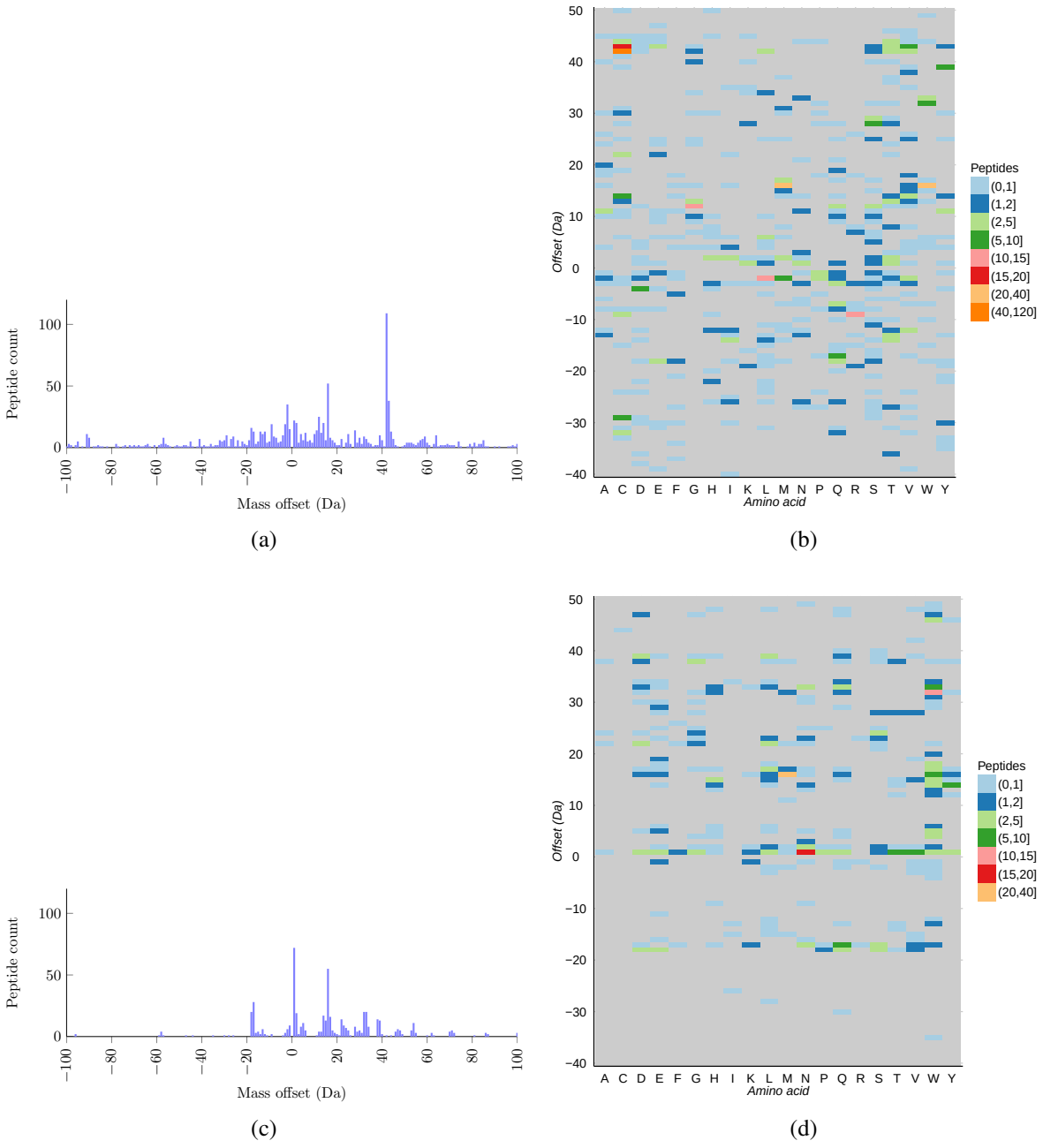


Figure B.16: Modifications of peptides identified only with that modification, or with and without the modification. (a) Histogram of offsets over 1099 peptides with only modifications, (b) and their breakdown by residue. (c). 1051 out of these 1497 peptides were not identified in restrictive MS/MS searches. Computed on 544 peptides with observed non-modified and modified versions, (d) along with residue breakdown.

hardly any information is lost as the result of error correction and that the antibody repertoire is a better option for immunoproteogenomics searches than the previously used the unique reads

database.

To further evaluate the constructed repertoires, we performed additional analysis of CDR3 regions. Differing antibody clusters with shared CDR3 sequence (and V region labeling) partition all antibodies into clones. We refer to the *capacity of the clone* as the number of antibodies composing it.

Since coincidence of CDR3 region of two unrelated antibodies is an unlikely event, there are two possible explanations for non-trivial clones: (i) erroneous partitioning of reads from the same antibody into multiple clusters due to insufficient error correction, and (ii) correct clustering of multiple differing antibodies into multiple clusters with the same CDR3. In the latter case, since these multiple antibodies were not exposed to diversity mechanisms (such as SHM) in their CDR3 region, we expect that variations in these antibodies dominate in CDR1 and CDR2 regions.

To test whether the case (ii) holds, we used CLUSTAL W, version 2.0 ([LBB+07]) to align all antibodies within a clone, and to identify the variable positions in all non-trivial clones. Figure B.17 shows the histogram of the variable positions for the constructed heavy chain repertoire. Since the peaks in the histograms are located at approximate positions of CDR1 and CDR2 regions, we conclude that most non-trivial clones indeed were formed by related and diversified antibodies (rather than by errors in clustering). See Appendix B for the peptide coverage over the CDR3 region.

Coverage of CDR3 region by peptides

Peptide coverage of antibodies and clones is of interest since it can provide us with direct proteomic evidence of which antibodies/clones are specific to the introduced antigen. Of particular interest is the peptide coverage over the region which defines clonality; the CDR3 region. Figure B.18 shows the coverage distribution for each clone over the CDR3 region and reveals that often very few residues are being covered at the junction of CDR3 region. However,

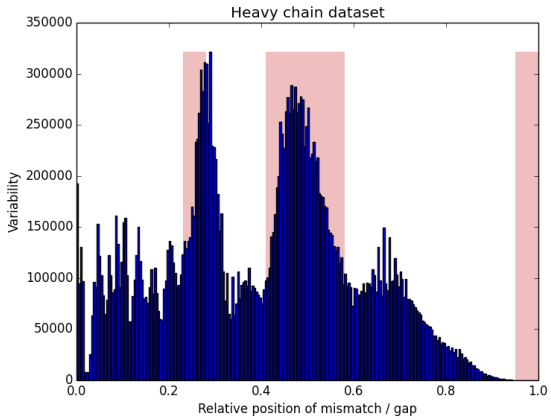


Figure B.17: Histogram of the mutated positions among non-trivial clones. Mutated positions are computed as relative positions of columns in multiple alignment of antibodies from each clone corresponding to a mismatch or an indel. The histogram was cut off at the right border of the CDR3 region. Red vertical bars correspond to positions of CDRs as specified in [Mur12].

few clones have high coverage over the entire region (99 clones with 90% or more coverage). Additional representations of clone coverage are shown in Figure B.19.

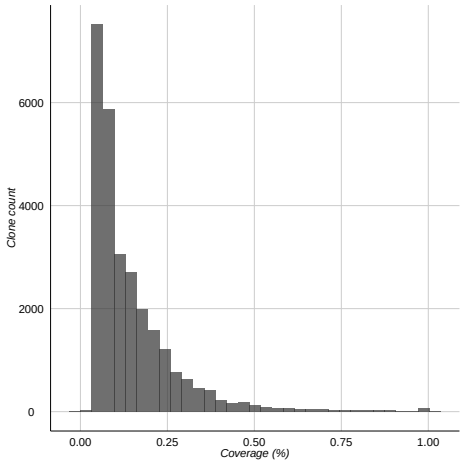
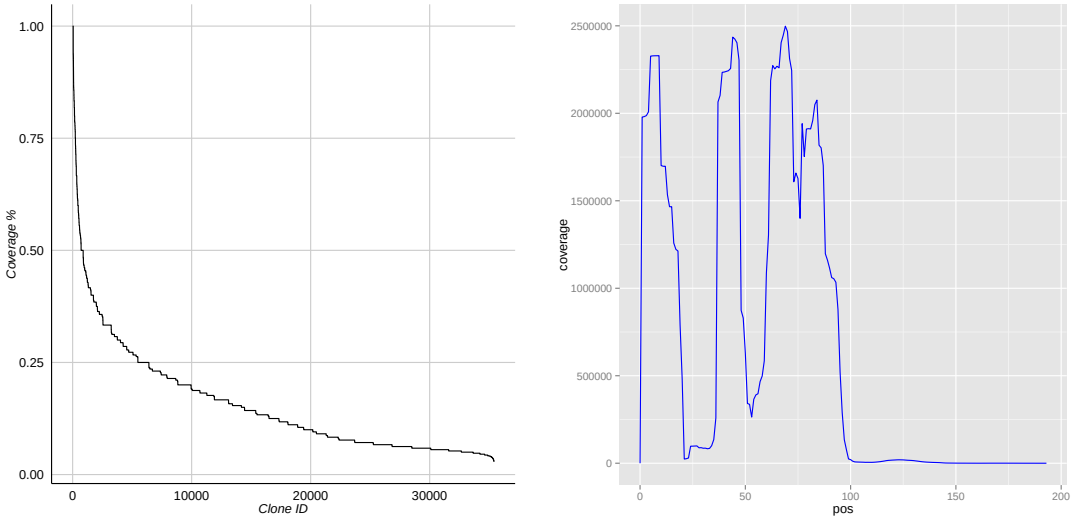


Figure B.18: Peptide coverage distribution of CDR3.



(a) (b)

Figure B.19: (a) Clones sorted by percent coverage of the CDR3 region by peptides. (b) Peptide coverage over positions of each antibody. No normalization of coverage is performed.

Appendix C

Supplement to Proteogenomic analysis of colorectal cancer

Table C.1: Database statistics. Total 90 RNA-seq BAM files which matches with the tumor samples used in the study of Zhang et al. [ZWW+14] 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

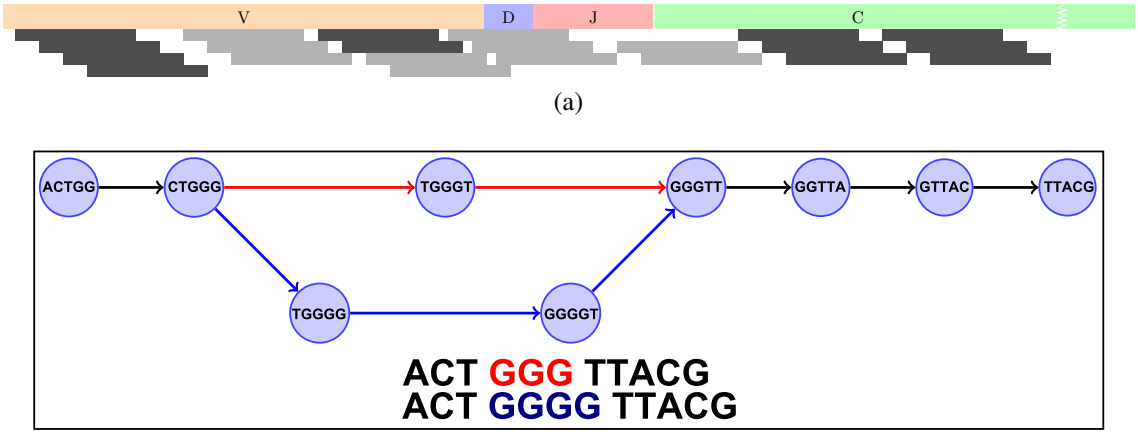


Figure C.1: (a) Potentially missed (in RNA-seq read alignment) reads from a somatically recombined heavy chain transcript as greyed out, while mapping reads as darker. (b) Example de Bruijn graph showing how differences in sequence manifest as differences in topology. In this example $k = 6$, and a single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

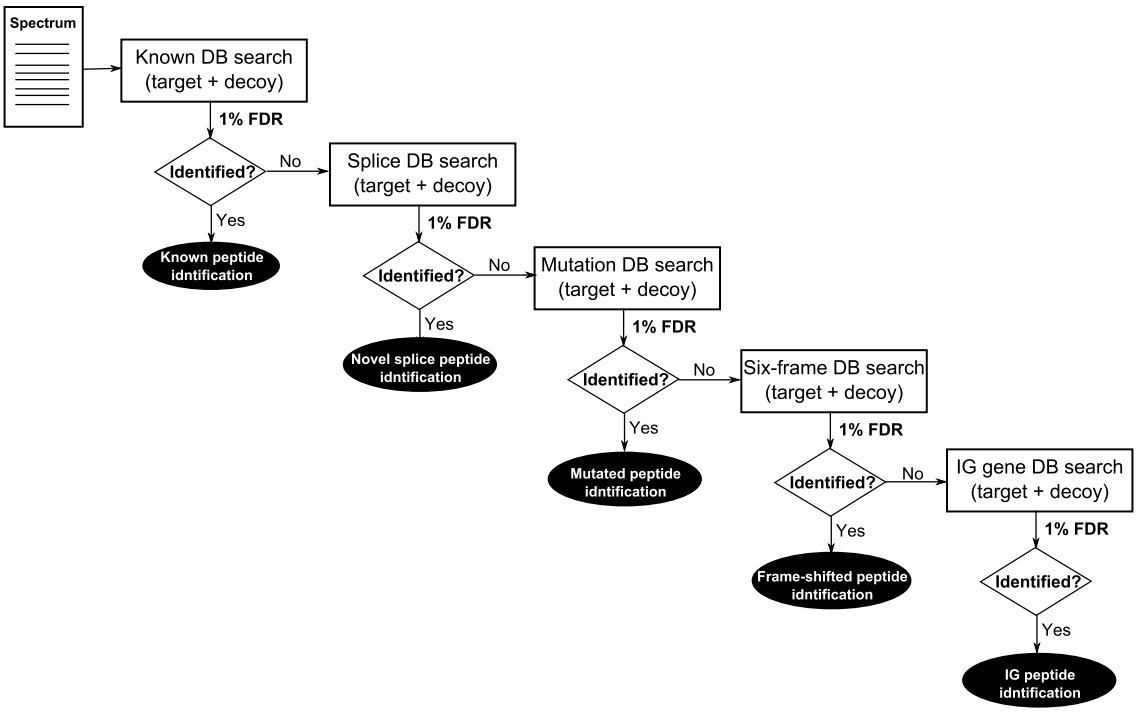
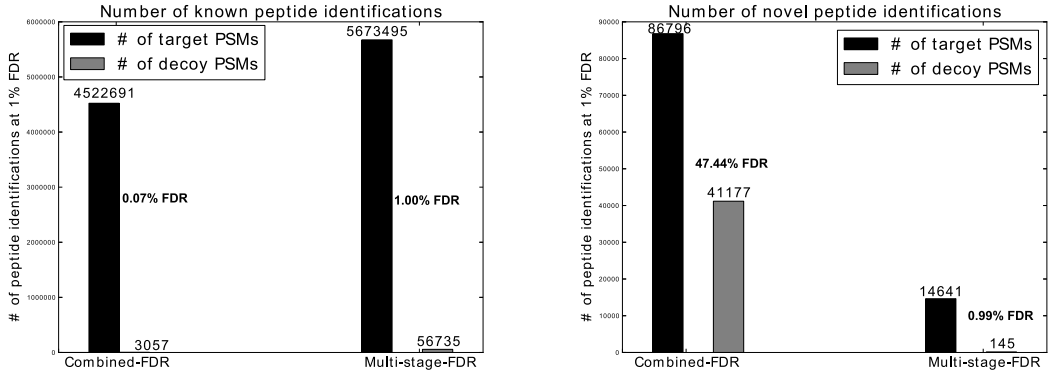


Figure C.2: Multistage-FDR strategy. Every spectrum will be searched against the known peptide database first, and are reported as a known peptides. In following stages, only the unidentified portion of the spectra are searched and assigned a new FDR threshold. Similar procedure is applied in the following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



(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.

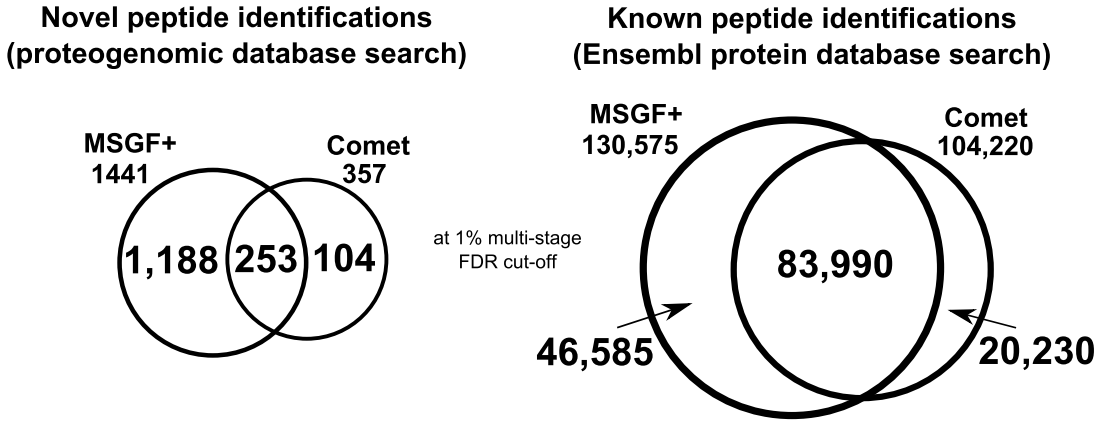
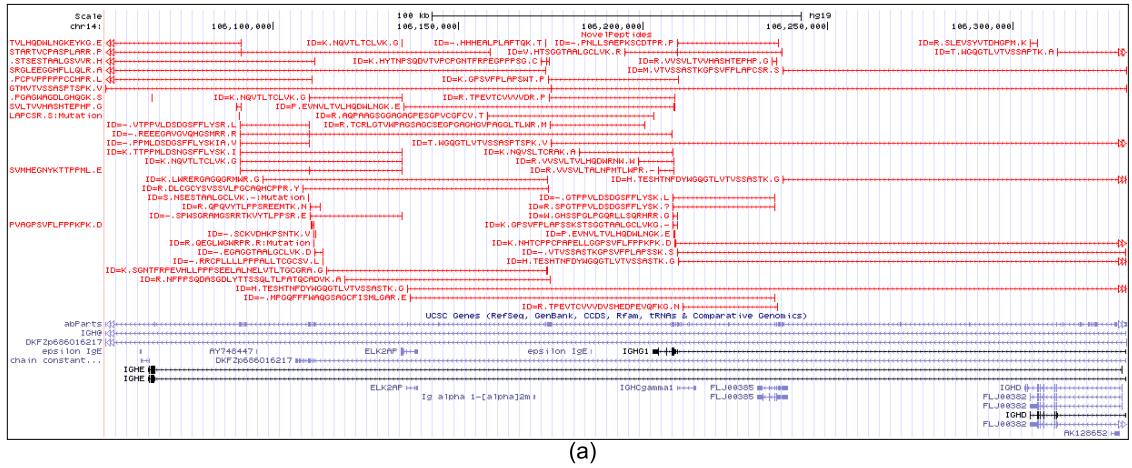


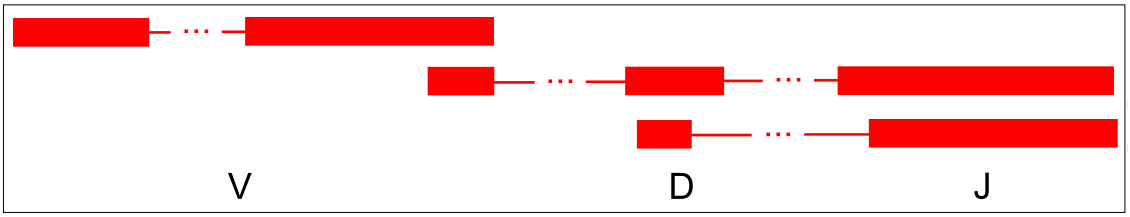
Figure C.4: Comparison between results obtained using MSGF+ and Comet MS/MS search tools. MSGF+ [KMB$^+$10] showed more peptide identification results in both known (Ensembl [FAA$^+$13] protein database) and novel (proteogenomic database) protein search with significant overlap.

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



(a)



(b)

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.

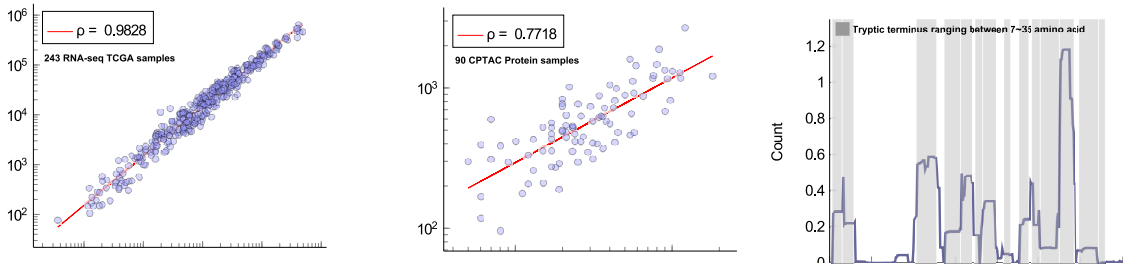


Figure C.6: (a) Plot of RNA-seq read counts from IG variable region versus IG constant region. We observed high correlation between RNA-seq reads that mapped to IG constant region versus variable region (filtered out using IG filter used in this study). (b) Spectra counts of peptide identifications from 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.

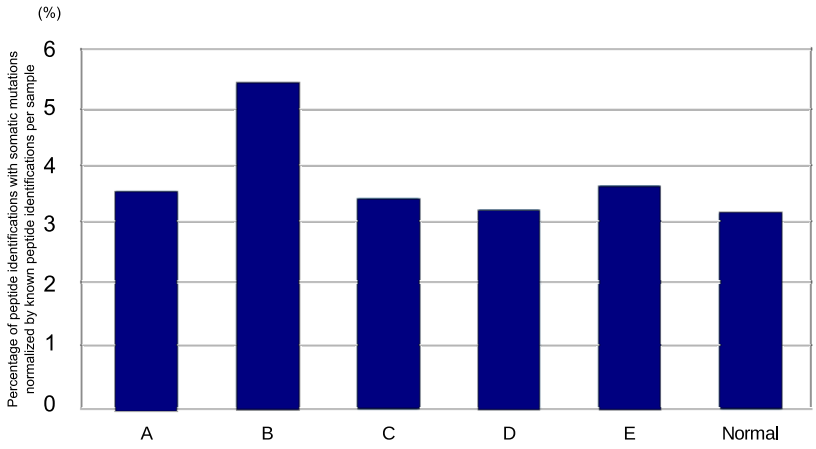
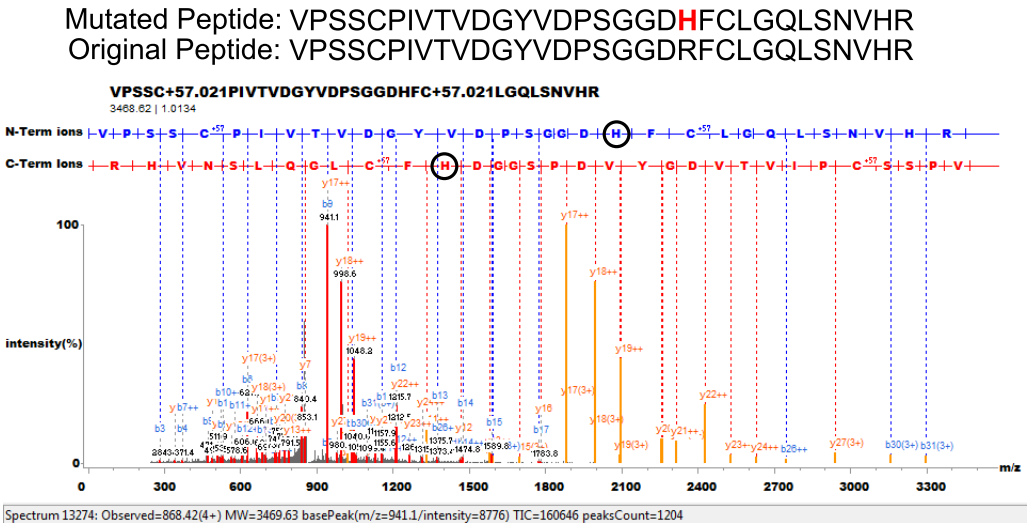
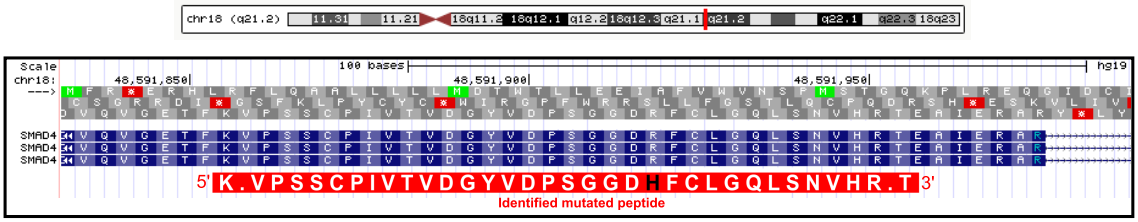


Figure C.7: Percentage of peptide identifications with somatic mutations in each sample normalized by the number of known peptide identifications across sample subtypes. This percentile ratio is calculated by dividing the number of known peptide identifications from the total number of IG peptide identifications 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 $p\text{-}value < 0.0001, \chi^2 = 40.39$.

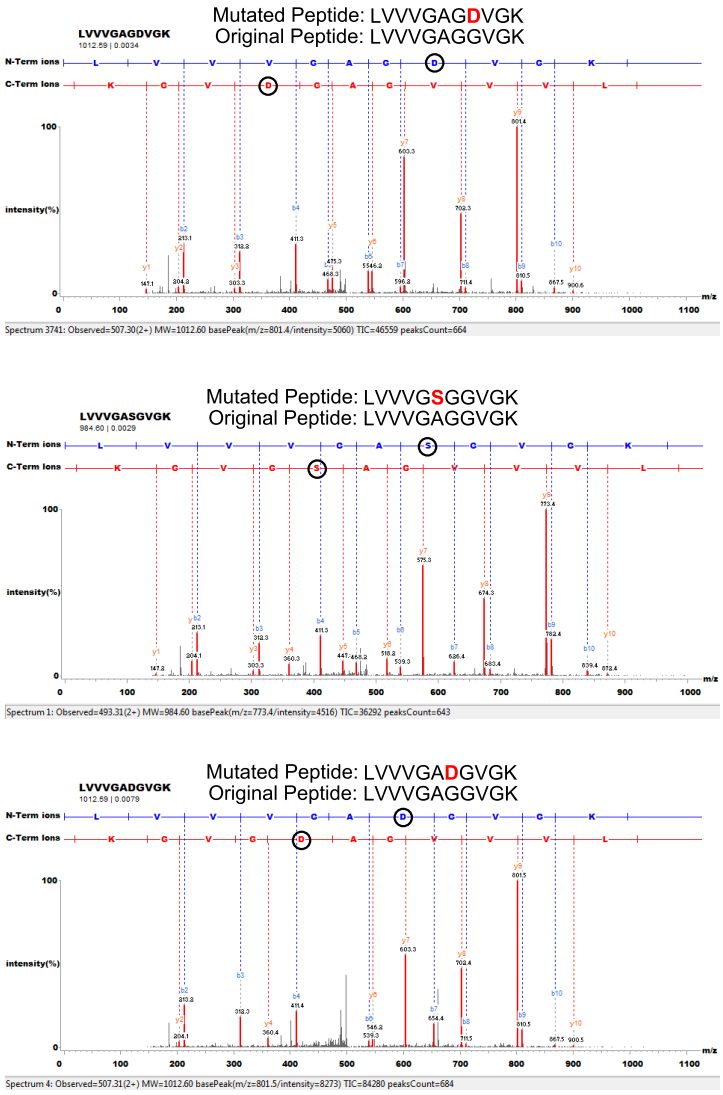


(a) Alignment of identified spectra with somatic mutation in gene SMAD4.

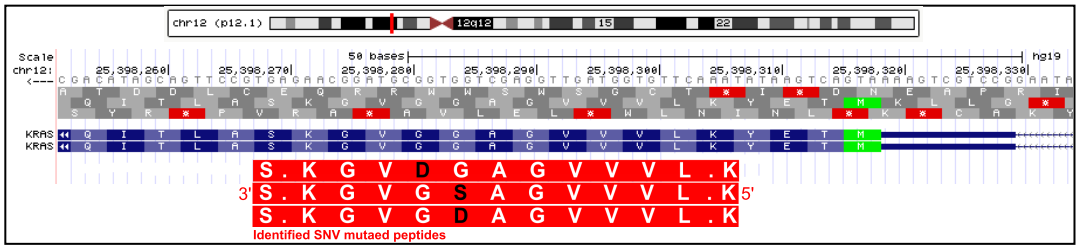


(b) UCSC Genome Browser plot of identified somatic mutation.

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 [MBC^{+}12], and overlapping mutation existed in COSMIC [FBB^{+}11] database.

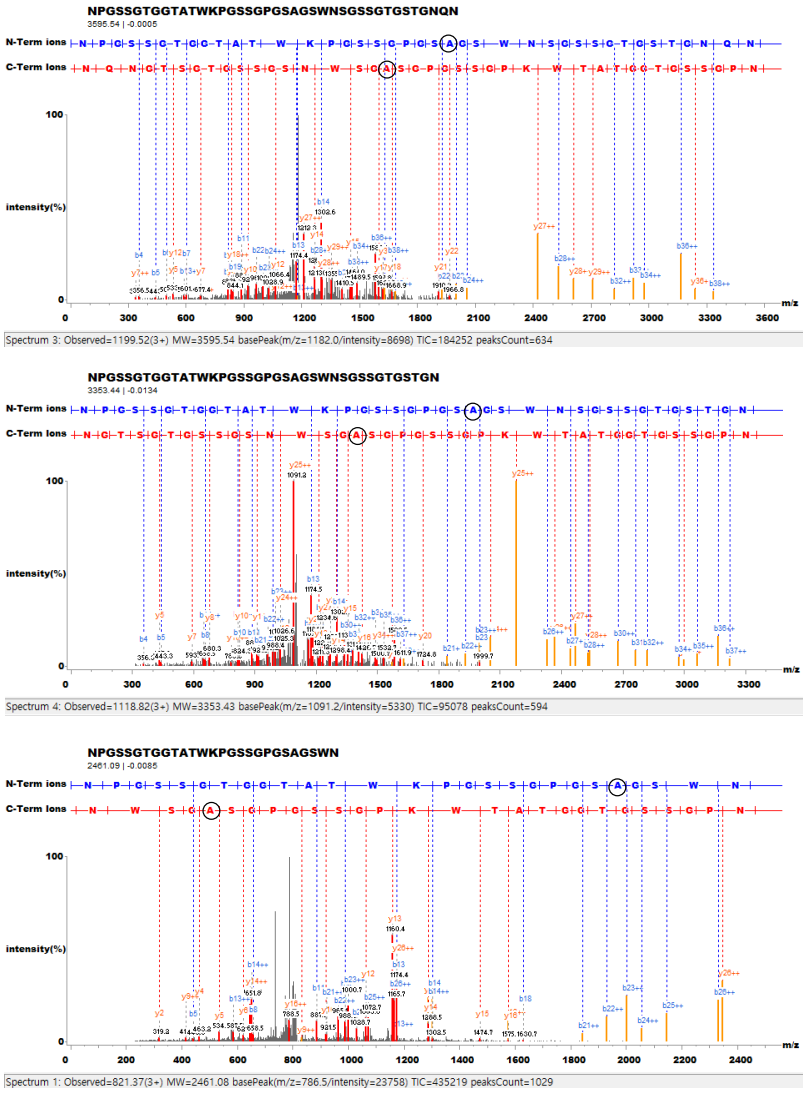


(a) Alignment of identified spectra with somatic mutation in gene KRAS

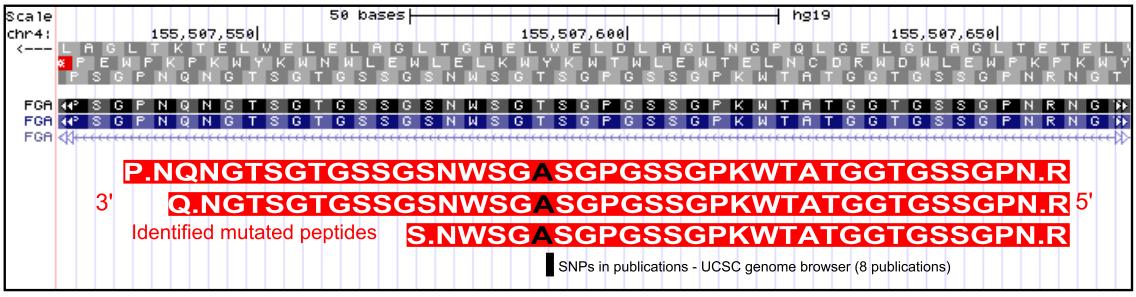


(b) UCSC Genome Browser plot of identified somatic mutation in gene KRAS.

Figure C.9: Identification of somatic mutation in gene KRAS. TCGA colon cancer study [MBC^{+}12] reported this mutation as 'somatic' in 25 different colon cancer samples and also reported by COSMIC [FBB^{+}11] and dbSNP [SWK^{+}01]. Peptide 'LVVVGAG:D:VGK' (G$-$ $>$D) had 1 spctra count and unique genomic location.

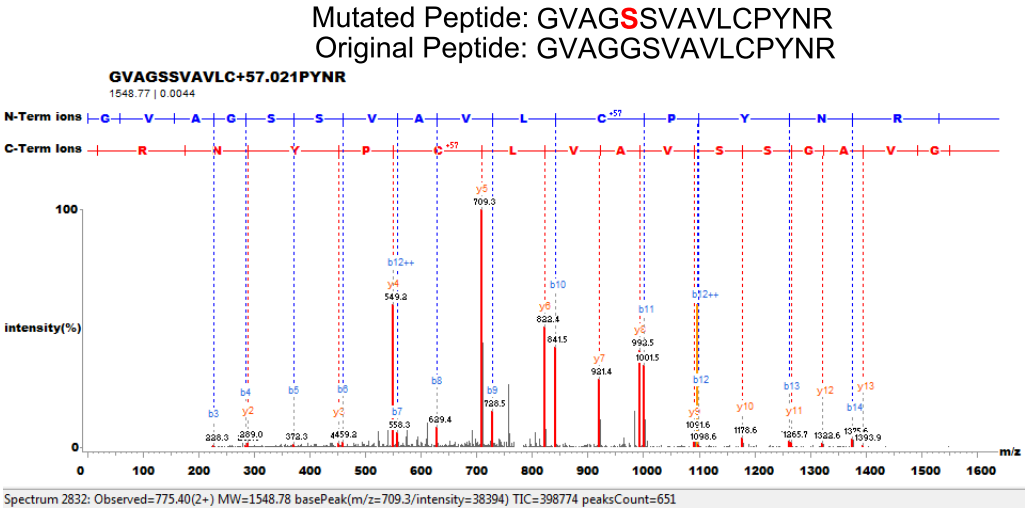


(a) Three identified spectra alignments indicating an identical SNV mutation in gene FGA

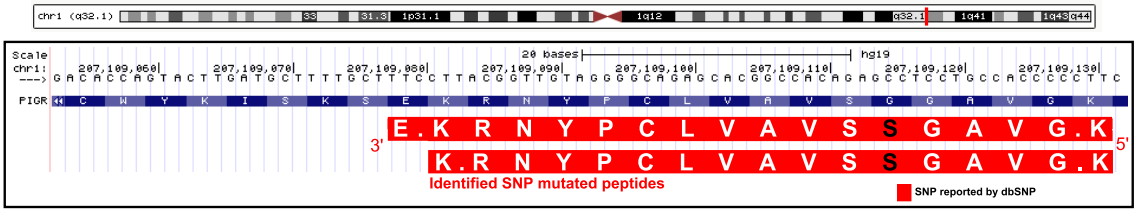


(b) UCSC Genome Browser plot of identified somatic mutation in gene FGA.

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 [FBB$^+$11] and dbSNP [SWK$^+$01].

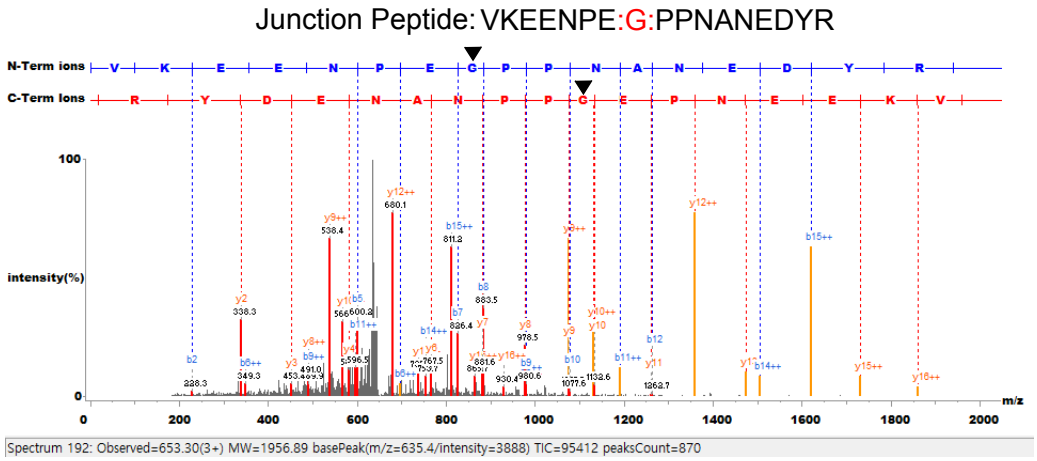


(a) Two identified spectra indicating an identical SNV mutation in gene PIGR

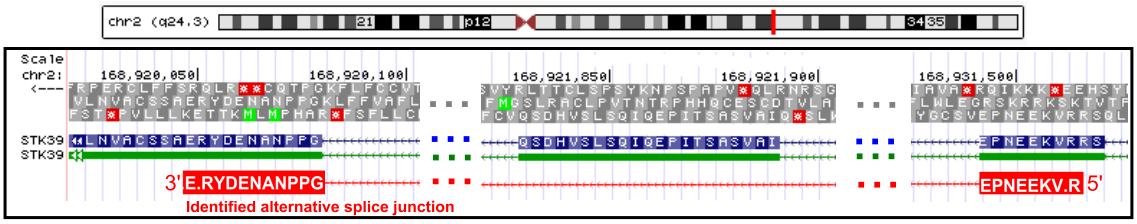


(b) UCSC Genome Browser plot of identified SNV mutation in gene PIGR.

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 [FBB+11] and dbSNP [SWK+01].

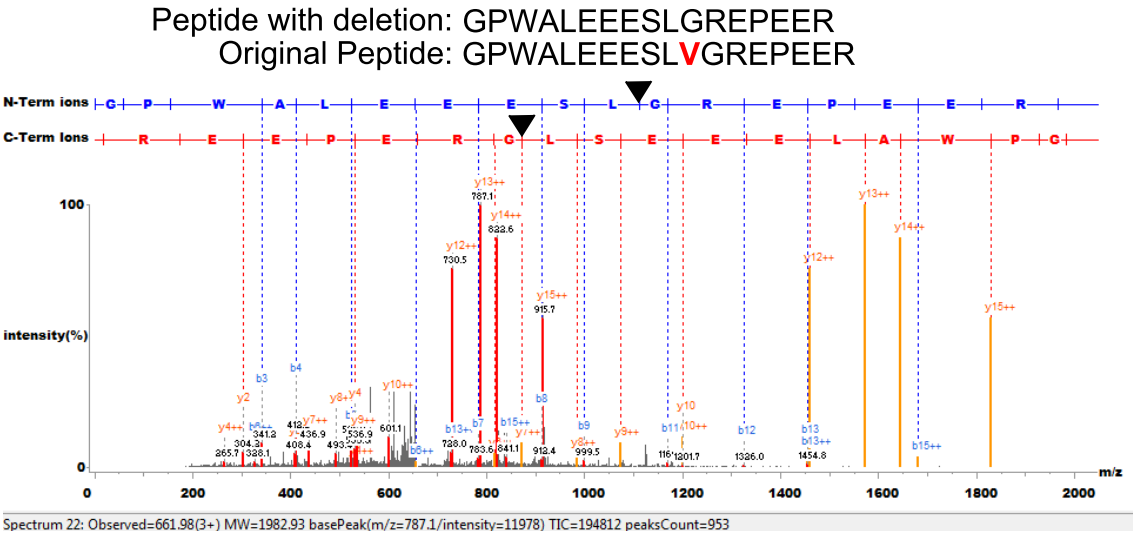


(a) Identified spectra with alternative splice junction

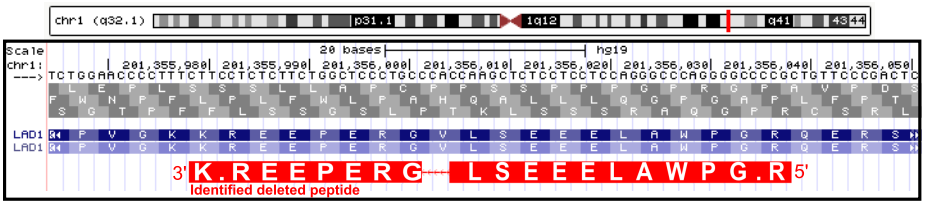


(b) UCSC Genome Browser plot of identified alternative splice junction

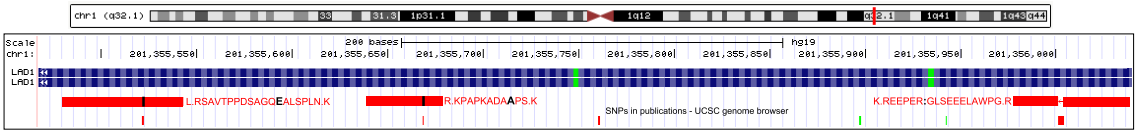
Figure C.12: Identified alternative splice junction peptide. Peptide 'VKEENPE:G:PPNANEDYR' (junction existing in the middle of amino acid 'G') had 11 spectra counts (with unique genomic location) and total 386 RNA-seq reads were mapped to this alternative splice junction.



(a) Identified spectra with deletion

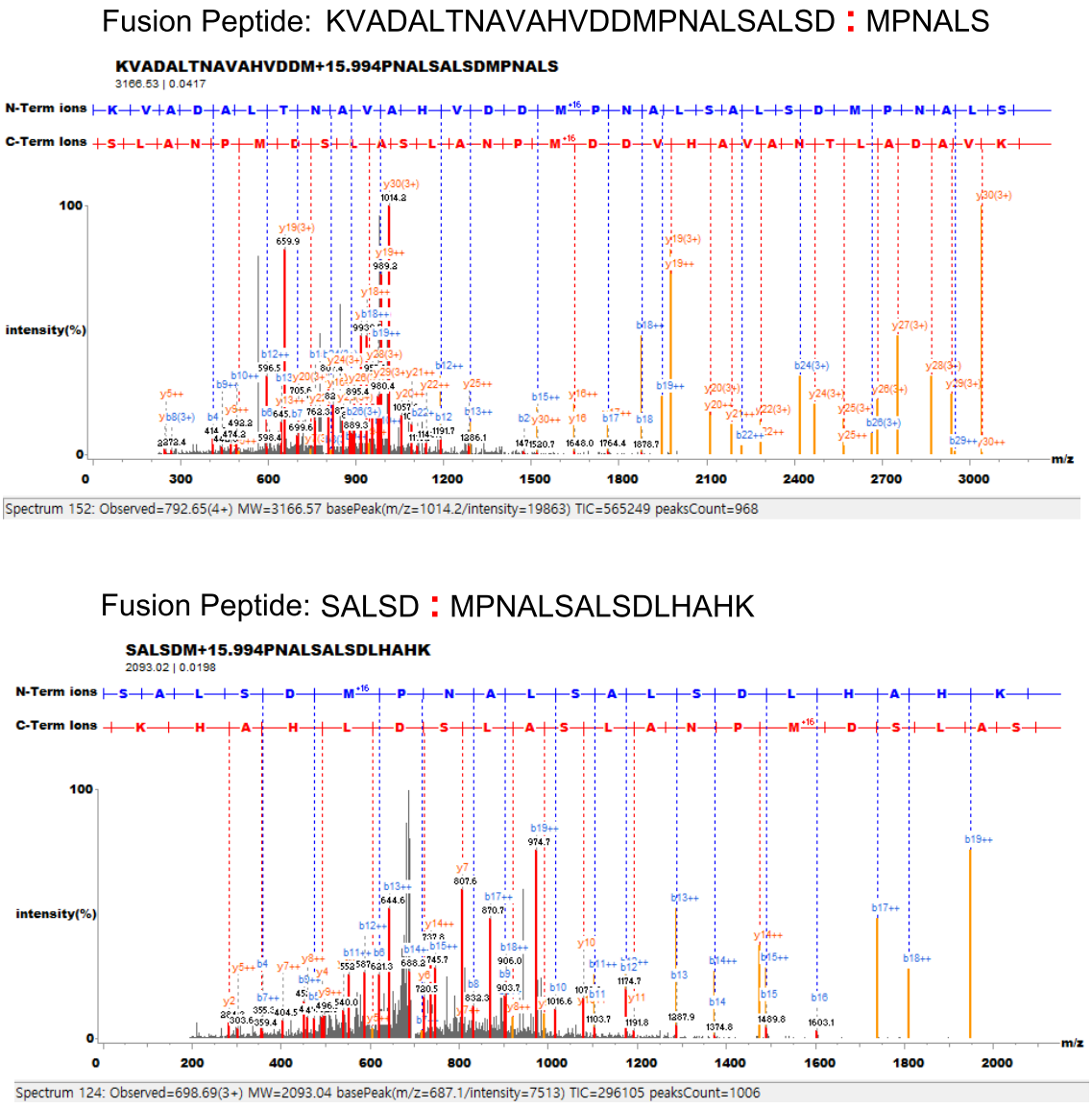


(b) UCSC Genome Browser plot of identified deletion.

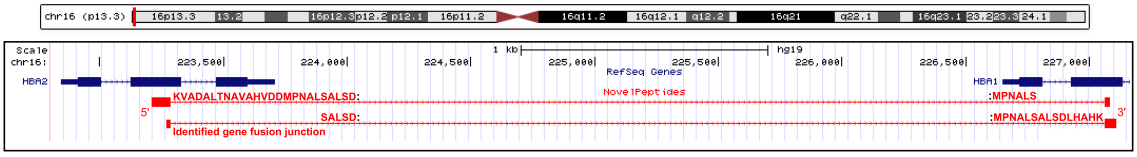


(c) Two different adjacent SNV mutations in LAD1 gene located within the same exon with identified deleted peptide.

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 [MBC^{+}12] 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

Supplement to Immunoglobulin assemblies from TILs

Algorithm description

Algorithm 3 Overview of TILAPIA Ig assembly

b **INPUTS:**
B: BAM file of RNA-seq reads
l: l-mer length
k: k-mer length
OUTPUT:
A: Set of assembled Ig sequences

1: **procedure** IG-ASSEMBLY(B, l, k)
2: $R_m \leftarrow$GET-IGH-MAPPED-READS(B)
3: $R_u \leftarrow$GET-UNMAPPED-READS(B)
4: $R' \leftarrow R_m \cup R_u$ ▷ Join mapped and unmapped reads
5: $R \leftarrow$FILTER-IG-READS(R', l)
6: $G_{Ig} \leftarrow$CREATE-IG-GRAPH(R,k)
7: $A \leftarrow$ASSEMBLE-IG(G_{Ig})
8: **end procedure**

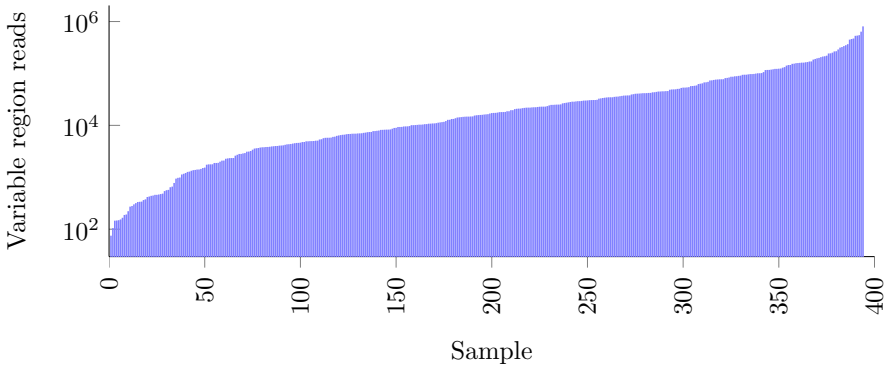
The methods GET-IGH-MAPPED-READS(B) and GET-UNMAPPED-READS(B) are simple commands to the `samtools` package for extracting reads from the BAM file B. $R' \leftarrow R_m \cup R_u$ joins the mapped and unmapped reads, while the internal workings of the call to method

CREATE-IG-GRAPH(R,k) is described in the main text in the Methods section. The methods FILTER-IG-READS(R', l) and ASSEMBLE-IG(G_{Ig}) are described below.

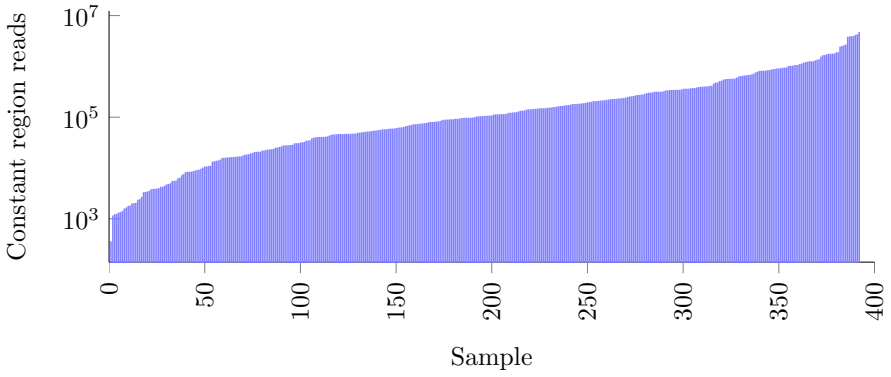
Algorithm 4 Filtering procedure

INPUTS:
R': subset of RNA-seq reads
l: l-mer length
EXTERNAL DATA STRUCTURES:
$\mathcal{R}$: set of concatenated V, D, J, and first 75bp of C, reference sequences
OUTPUT:
R: subset of high-quality, potential, IgH reads

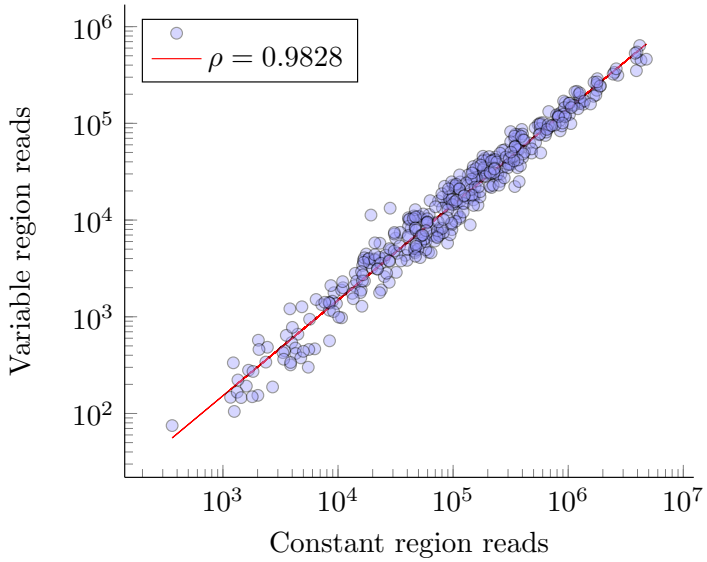
procedure FILTER-IG-READS(R', l)
    $R \leftarrow \{\}$
    $\mathcal{R}_l \leftarrow$ALL-l-MERS($\mathcal{R}$)
    for all $r \in R'$ **do**
        if N $\in r$ **then** SKIP **end if**
        if mean-qual(r) ≤ 25 **then** SKIP **end if**
        $C_l \leftarrow$ALL-l-MERS(r)
        if $C_l \cap \mathcal{R}_l = \emptyset$ **then** SKIP **end if**
        $r' \leftarrow$ TRIM-READ(r)
        $R \leftarrow R \cup \{r'\}$
    end for
return R
end procedure



(a)



(b)



(c)

Figure D.1: IgH read counts. (a) Reads mapping to the IgH constant region. (b) Filtered reads with l-mer matches to either the V, D, J, or first 75bp of the constant region. (c) Scatterplot of constant region reads to those of the variable region for each sample, with a fitted line.

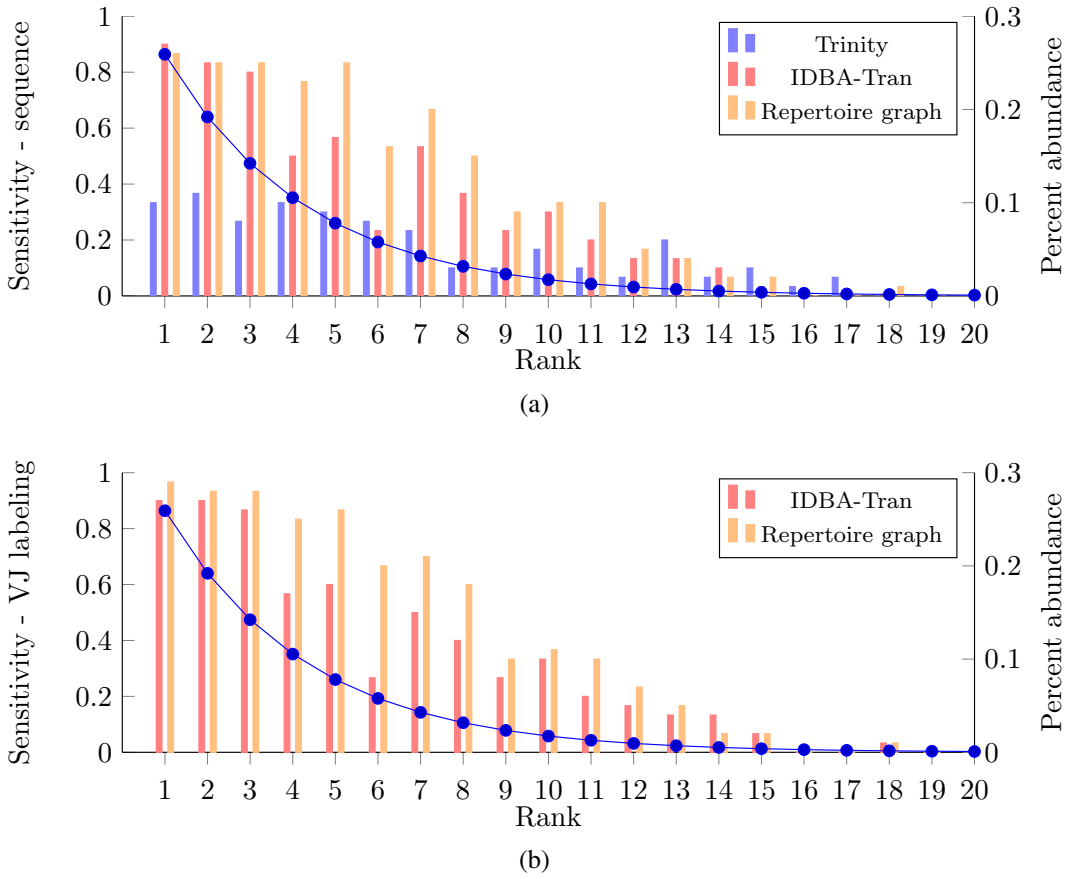


Figure D.2: Sensitivity on simulated data. Shown are the sensitivities on simulated data using different assemblers for a repertoire of 30 simulated antibodies, distributed according to the exponential distribution shown. The sensitivities are over 30 independent datasets, with each monoclonal rank shown independently. Sensitivity is measured by (a) sequence identity, and (b) VJ labeling.

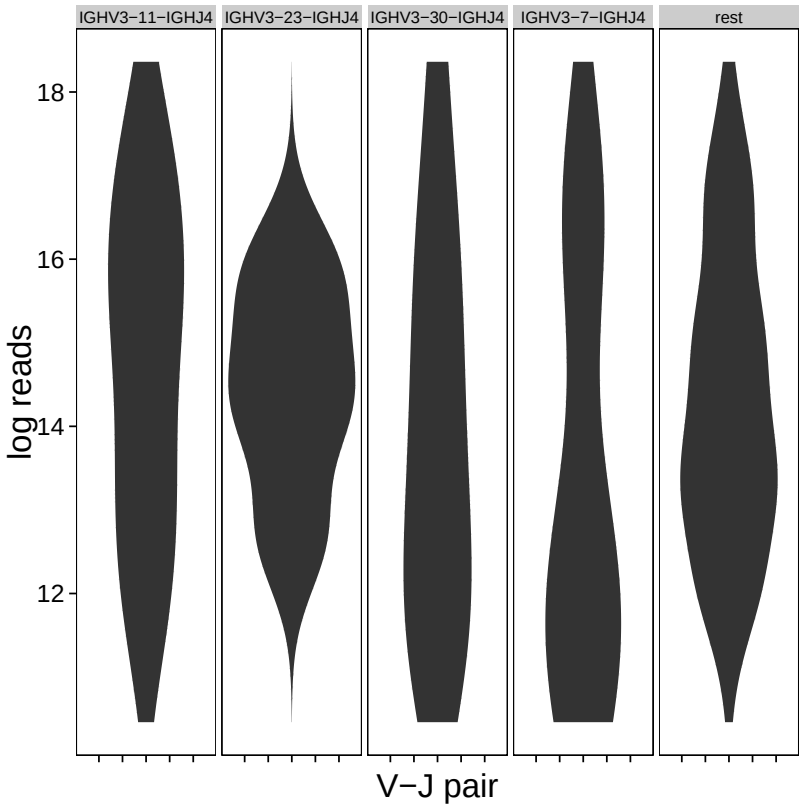
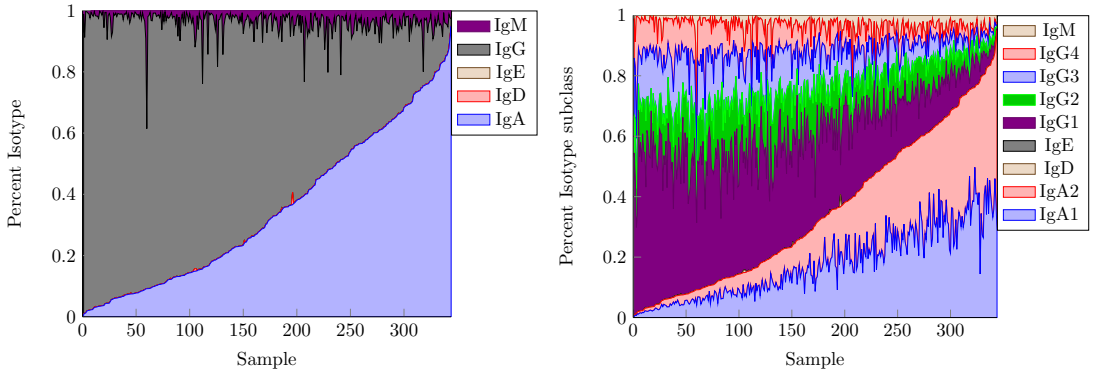


Figure D.3: Variable read count by VJ pair. Shown are the three most prominent VJ pairings, and the distribution of their read count in the variable region. The remaining samples are shown as the *rest* group. No bias is seen in these partitioning of samples.



(a) (b)

Figure D.4: (a) Isotype distribution in colon cancer. The relative abundance of isotypes among 344 samples is shown, sorted according to IgA. (b) Isotype subclass distribution. The same samples, and sorting, in (a), showing each isotype subclass.

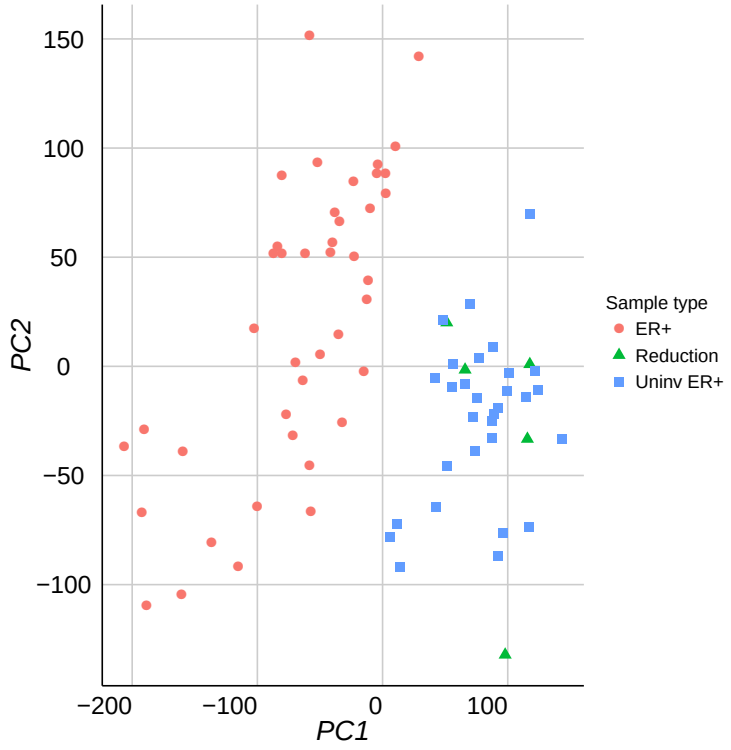


Figure D.5: Principle components analysis of ER+ samples using all genes with non-zero variance.

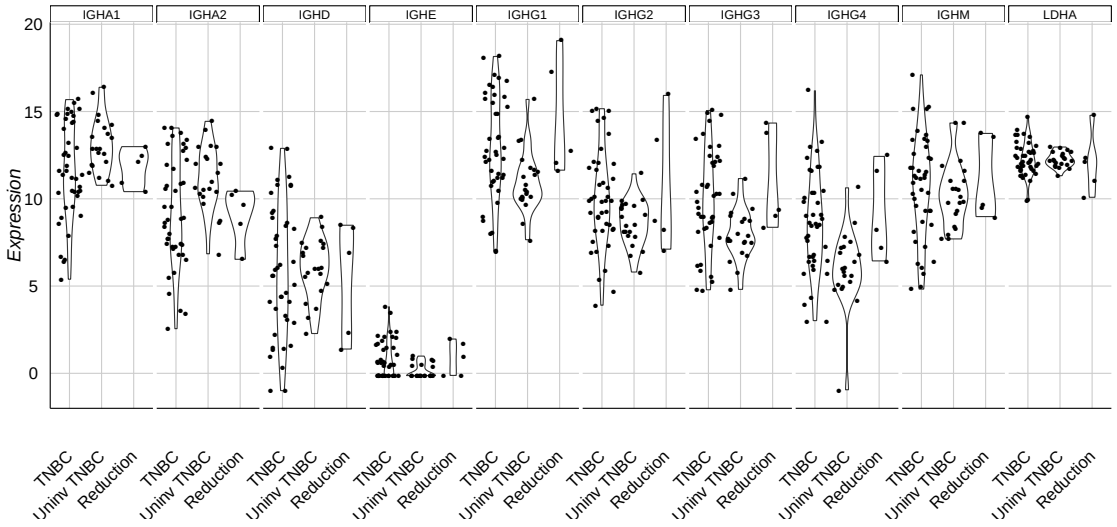


Figure D.6: Triple negative breast cancer (TNBC) IGH isotype responses.

Modeling spurious l-mer matches

Spurious l-mers in the filtering procedure can be modeled by a binomial distribution. The set of l-mers from the concatenation of V, D, J, and C reference gene-segments $\mathcal{R}$, is $\mathcal{R}_l \leftarrow$All-l-mers$(\mathcal{R})$. The probability of an l-mer matching by random chance is $p_l = \frac{|\mathcal{R}_l|}{4^l}$. So the probability of any l-mer match within a read of size n is:

$$\sum_{i=1}^{n-l} \binom{n-l}{i} p_l^i (1-p_l)^{n-l-i} \tag{D.1}$$

The values of equation D.1 for various values of l is shown in Figure D.7(a) for read lengths of 35 and 75. The largest l that has an acceptable value of false positives (less than 0.01) is $l = 13$ for both read lengths. Figure D.7(b) shows the number of reads retained on a single dataset for various values of l.

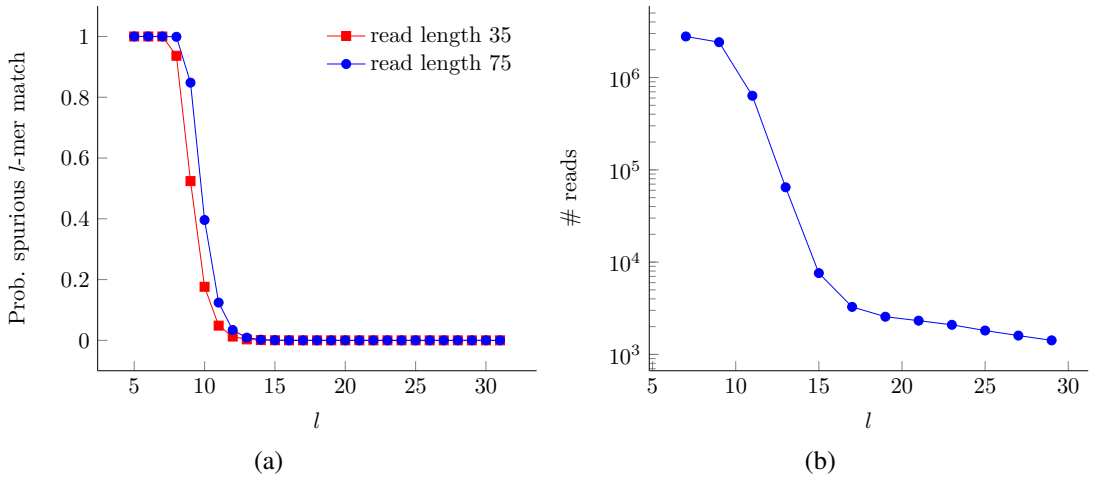


Figure D.7: (a) Probability of a spurious l-mer match within a read, for different read lengths. (b) Number of reads retained for different odd values of l on a single dataset.

Bibliography

[AB88] Stuart M Arfin and Ralph A Bradshaw. Cotranslational processing and protein turnover in eukaryotic cells. *Biochemistry*, 27(21):7979–7984, 1988.

[ACD+80] K Arai, B F Clark, L Duffy, M D Jones, Y Kaziro, R A Laursen, J L'Italien, D L Miller, S Nagarkatti, S Nakamura, K M Nielsen, T E Petersen, K Takahashi, and M Wade. Primary structure of elongation factor tu from escherichia coli. *Proceedings of the National Academy of Sciences*, 77(3):1326–1330, 1980.

[AKH+95] Stuart M Arfin, Richard L Kendall, Linda Hall, Larry H Weaver, Albert E Stewart, Brian W Matthews, and Ralph A Bradshaw. Eukaryotic methionyl aminopeptidases: two classes of cobalt-dependent enzymes. *Proceedings of the National Academy of Sciences*, 92(17):7714–7718, 1995.

[ALC+11] R. Arnaout, W. Lee, P. Cahill, T. Honan, T. Sparrow, M. Weiand, C. Nusbaum, K. Rajewsky, and S.B. Koralov. High-resolution description of antibody heavy-chain repertoires in humans. *PloS One*, 6(8):e22365, 2011.

[APH14] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. Htseq–a python framework to work with high-throughput sequencing data. *Bioinformatics*, page btu638, 2014.

[ATW+11] 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.

[AWP+08] 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.

[AWR+12] Florent E Angly, Dana Willner, Forest Rohwer, Philip Hugenholtz, and Gene W Tyson. Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Research*, 40(12):e94–e94, 2012.

[BBB+11] D. Bell, A. Berchuck, M. Birrer, J. Chien, and et al. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, Jun 2011.

[BBW98] Ralph A Bradshaw, William W Brickey, and Kenneth W Walker. N-terminal processing: the methionine aminopeptidase and n α-acetyl transferase families. *Trends in biochemical sciences*, 23(7):263–267, 1998.

[BDSY99] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *J. Comp. Bio*, 6(3-4):281–97, 1999.

[BFV86] Andreas Bachmair, Daniel Finley, and Alexander Varshavsky. In vivo half-life of a protein is a function of its amino-terminal residue. *Science*, 234(4773):179–186, 1986.

[BHAH+13] Anna Bachmayr-Heyda, Stefanie Aust, Georg Heinze, Stephan Polterauer, Christoph Grimm, Elena I Braicu, Jalid Sehouli, Sandrina Lambrechts, Ignace Vergote, Sven Mahner, et al. Prognostic impact of tumor infiltrating cd8+ t cells in association with cell proliferation in ovarian cancer patients-a study of the ovcad consortium. *BMC cancer*, 13(1):422, 2013.

[BHM83] Mitali Basu, Mahabaleshwar V Hegde, and Mukund J Modak. Synthesis of compositionally unique dna by terminal deoxynucleotidyl transferase. *Biochemical and biophysical research communications*, 111(3):1105–1112, 1983.

[BHW+14] Daniel R Boutz, Andrew P Horton, Yariv Wine, Jason J Lavinder, George Georgiou, and Edward M Marcotte. Proteomic identification of monoclonal antibodies from serum. *Analytical chemistry*, 86(10):4758–4766, 2014.

[BLG08] X. Brochet, M.P. Lefranc, and V. Giudicelli. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized VJ and VDJ sequence analysis. *Nucleic Acids Research*, 36(suppl 2):W503–W508, 2008.

[BLLO+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.

[BNA+12] 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. *Journal of Computational Biology*, 19:455–477, 2012.

[Bur76] F. M. Burnet. A modification of Jerne's theory of antibody production using the concept of clonal selection. *CA Cancer J Clin*, pages 119–21, 1976.

[CB09] Richard Cordaux and Mark A Batzer. The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, 10(10):691–703, 2009.

[CB10] 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.

[CBHK02] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[CBZ+12] W.C. Cheung, S.A. Beausoleil, X. Zhang, S. Sato, S.M. Schieferl, J.S. Wieler, J.G. Beaudet, R.K. Ramenani, L. Popova, M.J. Comb, et al. A proteomics approach for the identification and cloning of monoclonal antibodies from serum. *Nature biotechnology*, 2012.

[CGPvV06] Louis A Clark, Skanth Ganesan, Sarah Papp, and Herman WT van Vlijmen. Trends in antibody sequence changes during the somatic hypermutation process. *The Journal of Immunology*, 177(1):333–340, 2006.

[CPS+08] 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.

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

[CPT11] 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.

[CPZ+12] W. Chen, P. Prabakaran, Z. Zhu, Y. Feng, E.D. Streaker, and D.S. Dimitrov. Identification of cross-reactive IgG antibodies from an acute HIV-1-infected patient using phage display and high-throughput sequencing technologies. *Experimental and Molecular Pathology*, 2012.

[CSH+14] 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.

[CTK+01] Julia A Coronella, P Telleman, GA Kingsbury, TD Truong, S Hays, and RP Junghans. Evidence for an antigen-driven humoral immune response in medullary ductal breast cancer. *Cancer research*, 61(21):7889–7899, 2001.

[DBB+97] T. Dorner, H. P. Brezinschek, R. I. Brezinschek, S. J. Foster, R. Domiati-Saad, and P. E. Lipsky. Analysis of the frequency and pattern of somatic mutations within nonproductively rearranged human variable heavy chain genes. *J Immunol.*, 6(158):2779–89, 1997.

[DDS+13] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[DFFL98] Thomas Dörner, Sandra J Foster, Nancy L Farner, and Peter E Lipsky. Somatic hypermutation of human immunoglobulin heavy chain genes: targeting of RGYW motifs on both DNA strands. *European Journal of Immunology*, 28(10):3384–3396, 1998.

[DSEMWJ+13] 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.

[DYP+84] Stephen V Desiderio, George D Yancopoulos, Michael Paskind, Elise Thomas, Michael A Boss, Nathaniel Landau, Frederick W Alt, and David Baltimore. Insertion of N regions into heavy-chain genes is correlated with expression of terminal deoxytransferase in B cells. *Nature*, 311:752–755, 1984.

[EG07] 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.

[EJH13] 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.

[FAA+13] 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.

[FBB+11] 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.

[Fea11] E. R. Fearon. Molecular genetics of colorectal cancer. *Annu Rev Pathol*, 6:479–507, 2011.

[FM83] Edward B Fowlkes and Colin L Mallows. A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383):553–569, 1983.

[FMP+06] Frédéric Frottin, Aude Martinez, Philippe Peynot, Sanghamitra Mitra, Richard C Holz, Carmela Giglione, and Thierry Meinnel. The proteomics of n-terminal methionine cleavage. *Molecular & Cellular Proteomics*, 5(12):2336–2349, 2006.

[FMS+10] AM Frank, ME Monroe, AR Shah, RJ Moore, GR Anderson, RD Smith, and PA Pevzner. Spectral archives: A novel approach to analyzing tandem mass spectra. *Submitted*, 2010.

[FSL+13] 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.

[GBKP11] Nitin Gupta, Nuno Bandeira, Uri Keich, and PavelA. Pevzner. Target-decoy approach and false discovery rate: When things may go wrong. *Journal of The American Society for Mass Spectrometry*, 22(7):1111–1120, 2011.

[GBM04] C Giglione, A Boularot, and T Meinnel. Protein n-terminal methionine excision. *Cellular and Molecular Life Sciences CMLS*, 61(12):1455–1474, 2004.

[GHP95] P. Galinier, M. Habib, and C. Paul. Chordal graphs and their clique graphs. In *In WG 95*, pages 358–371. Springer-Verlag, 1995.

[GHY+11] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, et al. Full-length transcriptome assembly from rna-seq data without a reference genome. *Nature Biotechnology*, 29(7):644–652, 2011.

[GJ79] M. R. Garey and D. S. Johnson. *Computers and Intractability, A Guide to the Theory of NP-Completeness*. W.H. Freeman and Company, New York, 1979.

[GMJ+07] Bruno A Gaëta, Harald R Malming, Katherine JL Jackson, Michael E Bain, Patrick Wilson, and Andrew M Collins. iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. *Bioinformatics*, 23(13):1580–1587, 2007.

[GNL+15] Andrew J Gentles, Aaron M Newman, Chih Long Liu, Scott V Bratman, Weiguo Feng, Dongkyoon Kim, Viswam S Nair, Yue Xu, Amanda Khuong, Chuong D Hoang, et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nature medicine*, 2015.

[GTJ+07] Nitin Gupta, Stephen Tanner, Navdeep Jaitly, Joshua N. Adkins, Mary Lipton, Robert Edwards, Margaret Romine, Andrei Osterman, Vineet Bafna, Richard D. Smith, and Pavel A. Pevzner. Whole proteome analysis of post-translational modifications: Applications of mass-spectrometry for proteogenomic annotation. *Genome Research*, 17(9):1362–1377, 2007.

[GVM03] Carmela Giglione, Olivier Vallon, and Thierry Meinnel. Control of protein life-span by n-terminal methionine excision. *The EMBO journal*, 22(1):13–23, 2003.

[HEL+87] S Huang, RC Elliott, PS Liu, RK Koduri, LC Blair, KM Bryan, P Ghosh-Dastidar, B Einarson, and RL Kendall. Specificity of cotranslational amino-terminal processing of proteins in yeast. *Biochemistry*, 26(25):8242–8246, 1987.

[HGH+14] Kalani Halemano, Kejun Guo, Karl J Heilman, Bradley S Barrett, Diana S Smith, Kim J Hasenkrug, and Mario L Santiago. Immunoglobulin somatic hypermutation by apobec3/rfv3 during retroviral infection. *Proceedings of the National Academy of Sciences*, 111(21):7759–7764, 2014.

[HGR+10] Andreas O Helbig, Sharon Gauci, Reinout Raijmakers, Bas van Breukelen, Monique Slijper, Shabaz Mohammed, and Albert JR Heck. Profiling of n-acetylated protein termini provides in-depth insights into the n-terminal nature of the proteome. *Molecular & Cellular Proteomics*, 9(5):928–939, 2010.

[HPG+99] Clyde A Hutchison, Scott N Peterson, Steven R Gill, Robin T Cline, Owen White, Claire M Fraser, Hamilton O Smith, and J Craig Venter. Global transposon mutagenesis and a minimal mycoplasma genome. *Science*, 286(5447):2165–2169, 1999.

[HSV10] Cheol-Sang Hwang, Anna Shemorry, and Alexander Varshavsky. N-terminal acetylation of cellular proteins creates specific degradation signals. *Science*, 327(5968):973–977, 2010.

[HWdB87] Robert Hamilton, Colin K Watanabe, and Herman A de Boer. Compilation and comparison of the sequence context around the aug startcodons in saccharomyces cerevisiae mrnas. *Nucleic acids research*, 15(8):3581–3593, 1987.

[ICT+12] Zamin Iqbal, Mario Caccamo, Isaac Turner, Paul Flicek, and Gil McVean. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics*, 44(2):226–232, 2012.

[Jac08] Paul Jaccard. *Nouvelles recherches sur la distribution florale*. 1908.

[JBGC10] Katherine JL Jackson, Scott Boyd, Bruno A Gaëta, and Andrew M Collins. Benchmarking the performance of human antibody gene alignment utilities using a 454 sequence dataset. *Bioinformatics*, 26(24):3129–3130, 2010.

[JC03] X. Jiang and J. R. Couchman. Perlecan and tumor angiogenesis. *J. Histochem. Cytochem.*, 51(11):1393–1410, Nov 2003.

[JGSC04] Katherine JL Jackson, Bruno Gaeta, William Sewell, and Andrew M Collins. Exonuclease activity and P nucleotide addition in the generation of the expressed immunoglobulin repertoire. *BMC Immunology*, 5(1):19, 2004.

[JH70] Richard Jackson and Tony Hunter. Role of methionine in the initiation of haemoglobin synthesis. *Nature*, 227:672–676, 1970.

[JHW+13] Ning Jiang, Jiankui He, Joshua A. Weinstein, Lolita Penland, Sanae Sasaki, Xiao-Song He, Cornelia L. Dekker, Nai-Ying Zheng, Min Huang, Meghan Sullivan, Patrick C. Wilson, Harry B. Greenberg, Mark M. Davis, Daniel S. Fisher, and Stephen R. Quake. Lineage structure of the human antibody repertoire in response to influenza vaccination. *Science Translational Medicine*, 5(171):171ra19, 2013.

[JWP+11] N. Jiang, J.A. Weinstein, L. Penland, R.A. White III, D.S. Fisher, and S.R. Quake. Determinism and stochasticity during maturation of the zebrafish antibody repertoire. *Proceedings of the National Academy of Sciences*, 108(13):5348–5353, 2011.

[KFMS12] 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.

[KGJ+13] Panagiotis Karagiannis, Amy E Gilbert, Debra H Josephs, Niwa Ali, Tihomir Dodev, Louise Saul, Isabel Correa, Luke Roberts, Emma Beddowes, Alexander Koers, et al. Igg4 subclass antibodies impair antitumor immunity in melanoma. *The Journal of clinical investigation*, 123(4):1457, 2013.

[KGP08] Sangtae Kim, Nitin Gupta, and Pavel A Pevzner. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *Journal of proteome research*, 7(8):3354–3363, 2008.

[KK99] G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20(1):35992, 1999.

[KMB+10] 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.

[Koz89] Marilyn Kozak. The scanning model for translation: an update. *The Journal of cell biology*, 108(2):229–241, 1989.

[KP14] Sangtae Kim and Pavel A Pevzner. Ms-gf+ makes progress towards a universal database search tool for proteomics. *Nature communications*, 5:5277, 2014.

[KPG+14] 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.

[KSS10] D.R. Kelley, M.C. Schatz, and S.L. Salzberg. Quake: quality-aware detection and correction of sequencing errors. *Genome Biol*, 11(11):R116, 2010.

[LBB+07] M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, and D.G. Higgins. Clustal w and clustal x version 2.0. *Bioinformatics*, pages 2947–8, 2007.

[LC95] Xuan Li and Yie-Hwa Chang. Amino-terminal protein processing in saccharomyces cerevisiae is an essential function that requires two distinct methionine aminopeptidases. *Proceedings of the National Academy of Sciences*, 92(26):12357–12361, 1995.

[LDZ10] J. Li, D. T. Duncan, and B. Zhang. CanProVar: a human cancer proteome variation database. *Hum. Mutat.*, 31(3):219–228, Mar 2010.

[Lev76] Michael Levitt. A simplified representation of protein conformations for rapid simulation of protein folding. *Journal of molecular biology*, 104(1):59–107, 1976.

[LGG+09] Marie-Paule Lefranc, Veronique Giudicelli, Chantal Ginestoux, Joumana Jabado-Michaloud, Geraldine Folch, Fatena Bellahcene, Yan Wu, Elodie Gemrot, Xavier Brochet, Jerome Lane, et al. Imgt®, the international immunogenetics information system®. *Nucleic acids research*, 37(suppl 1):D1006–D1012, 2009.

[LHA14] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol*, 15(12):550, 2014.

[Lin13] Michael Linnebacher. Tumor-infiltrating b cells come into vogue. *World journal of gastroenterology: WJG*, 19(1):8, 2013.

[Liu01] 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.

[LM12] Michael Linnebacher and Claudia Maletzki. Tumor-infiltrating b cells: The ignored players in tumor immunology. *Oncoimmunology*, 1(7):1186–1188, 2012.

[LSM+11] 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.

[LVGM+14] Uri Laserson, Francois Vigneault, Daniel Gadala-Maria, Gur Yaari, Mohamed Uduman, Jason A Vander Heiden, William Kelton, Sang Taek Jung, Yi Liu, Jonathan Laserson, et al. High-resolution antibody dynamics of vaccine-induced immune responses. *Proceedings of the National Academy of Sciences*, 111(13):4928–4933, 2014.

[LWW+13] 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.

[MBC+12] 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.

[MG08] Thierry Meinnel and Carmela Giglione. Tools for analyzing and predicting n-terminal protein modifications. *Proteomics*, 8(4):626–649, 2008.

[MK03] M. Miyaki and T. Kuroki. Role of Smad4 (DPC4) inactivation in human cancer. *Biochem. Biophys. Res. Commun.*, 306(4):799–804, Jul 2003.

[MRTM14] Miguel Martinez-Rodriguez, Alec K Thompson, and Carlos Monteagudo. A significant percentage of cd20-positive tils correlates with poor prognosis in patients with primary cutaneous malignant melanoma. *Histopathology*, 65(5):726–728, 2014.

[MSEE02] Sebastian Maurer-Stroh, Birgit Eisenhaber, and Frank Eisenhaber. N-terminal n-myristoylation of proteins: prediction of substrate proteins from amino acid sequence. *Journal of molecular biology*, 317(4):541–557, 2002.

[MSG06] Thierry Meinnel, Alexandre Serero, and Carmela Giglione. Impact of the n-terminal amino acid on targeted protein degradation. *Biological chemistry*, 387(7):839–851, 2006.

[MSKP11] P. Medvedev, E. Scott, B. Kakaradov, and P. Pevzner. Error correction of high-throughput sequencing datasets with non-uniform coverage. *Bioinformatics*, 27(13):i137–41, 2011.

[MTE+02] G. R. McLean, M. Torres, N. Elguezabal, A. Nakouzi, and A. Casadevall. Isotype Can Affect the Fine Specificity of an Antibody for a Polysaccharide Antigen. *The Journal of Immunology*, 169(3):1379–1386, 2002.

[MTV+08] Aude Martinez, José A Traverso, Benoît Valot, Myriam Ferro, Christelle Espagne, Geneviève Ephritikhine, Michel Zivy, Carmela Giglione, and Thierry Meinnel. Extent of n-terminal modifications in cytosolic proteins from eukaryotes. *Proteomics*, 8(14):2809–2831, 2008.

[Mur12] K. P. Murphy. *Janeway's immunobiology*, chapter Antigen Recognition by B-cell and T-cell Receptors. Garland Science, 8th edition, 2012.

[N+12] Cancer Genome Atlas Network et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, 2012.

[NBP12] Seungjin Na, Nuno Bandeira, and Eunok Paek. Fast multi-blind modification search through tandem mass spectrometry. *Molecular & Cellular Proteomics*, 11(4):M111–010199, 2012.

[Nel10] Brad H Nelson. CD20+ B cells: the other tumor-infiltrating lymphocytes. *The Journal of Immunology*, 185(9):4977–4982, 2010.

[NGS03] S. Nzula, J. J. Going, and D. I. Stott. Antigen-driven clonal proliferation, somatic hypermutation, and selection of B lymphocytes infiltrating human ductal breast carcinomas. *Cancer Res.*, 63(12):3275–3280, Jun 2003.

[NKA13] S.I. Nikolenko, A. Korobeynikov, and M.A. Alekseyev. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics*, 14:S7, 2013.

[NSM+12] 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.

[OLNLB06] Line Ohm-Laursen, Morten Nielsen, Stine R Larsen, and Torben Barington. No evidence for the use of DIR, D–D fusions, chromosome 15 open reading

frames or VHreplacement in the peripheral repertoire was found on application of an improved algorithm, JointML, to 6329 human immunoglobulin H rearrangements. *Immunology*, 119(2):265–277, 2006.

[ONI+09] 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.

[PAS09] Bogdan Polevoda, Thomas Arnesen, and Fred Sherman. A synopsis of eukaryotic nα-terminal acetyltransferases: nomenclature, subunits and substrates. In *BMC proceedings*, volume 3, page S2. BioMed Central Ltd, 2009.

[PBH+14] 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.

[PEP04] Alkes L Price, Eleazar Eskin, and Pavel A Pevzner. Whole-genome analysis of alu repeat elements reveals complex evolutionary history. *Genome research*, 14(11):2245–2252, 2004.

[PFHJ85] Bengt Persson, Christofer FLINTA, Gunnar HEIJNE, and Hans JÖRNVALL. Structures of n-terminally acetylated proteins. *European Journal of Biochemistry*, 152(3):523–527, 1985.

[PLY+13] Yu Peng, Henry CM Leung, Siu-Ming Yiu, Ming-Ju Lv, Xin-Guang Zhu, and Francis YL Chin. Idba-tran: a more robust de novo de bruijn graph assembler for transcriptomes with uneven expression levels. *Bioinformatics*, 29(13):i326–i334, 2013.

[PLYC11] Yu Peng, Henry CM Leung, Siu-Ming Yiu, and Francis YL Chin. Meta-idba: a de novo assembler for metagenomic data. *Bioinformatics*, 27(13):i94–i101, 2011.

[PMD+00] O Pritsch, C Magnac, G Dumas, J P Bouvet, P Alzari, and G Dighiero. Can isotype switch modulate antigen-binding affinity and influence clonal selection? 30(12):3387–95, 2000.

[PNT+99] Bogdan Polevoda, Joakim Norbeck, Hikaru Takakura, Anders Blomberg, and Fred Sherman. Identification and specificities of n-terminal acetyltransferases from saccharomyces cerevisiae. *The EMBO journal*, 18(21):6155–6168, 1999.

[PS02] Bogdan Polevoda and Fred Sherman. The diversity of acetylated proteins. *Genome Biol*, 3(5):0006–1, 2002.

[PTT04] Paul A Pevzner, Haixu Tang, and Glenn Tesler. De novo repeat classification and fragment assembly. *Genome Research*, 14(9):1786–1796, 2004.

[PTW01] 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.

[RHM+13] James Robinson, Jason A Halliwell, Hamish McWilliam, Rodrigo Lopez, Peter Parham, and Steven GE Marsh. The IMGT/HLA database. *Nucleic Acids Research*, 41(D1):D1222–D1227, 2013.

[RK92] Igor B Rogozin and Nikolai A Kolchanov. Somatic hypermutagenesis in immunoglobulin genes: II. influence of neighbouring base sequences on mutagenesis. *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression*, 1171(1):11–18, 1992.

[RP13] Adam Roberts and Lior Pachter. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature methods*, 10(1):71–73, 2013.

[RSW+15] Michael S Rooney, Sachet A Shukla, Catherine J Wu, Gad Getz, and Nir Hacohen. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell*, 160(1):48–61, 2015.

[RTL76] D. J. Rose, R. E. Tarjan, and G. S. Lueker. Algorithmic aspects of vertex elimination on graphs. *SIAM J. Comput.*, 2(5):26683, 1976.

[SBK+15] Yana Safonova, Stefano Bonissone, Eugene Kurpilyansky, Ekaterina Starostina, Alla Lapidus, Jeremy Stinson, Laura DePalatis, Wendy Sandoval, Jennie Lill, and Pavel A Pevzner. IgRepertoireConstructor: a novel algorithm for antibody repertoire construction and immunoproteogenomics analysis. *Bioinformatics*, 31(12):i53–i61, 2015.

[SBP+12] S. Sato, S.A. Beausoleil, L. Popova, J.G. Beaudet, R.K. Ramenani, X. Zhang, J.S. Wieler, S.M. Schieferl, W.C. Cheung, and R.D. Polakiewicz. Proteomics-directed cloning of circulating antiviral human monoclonal antibodies. *Nature Biotechnology*, 30(11):1039–1043, 2012.

[SCLR+04] M Margarida Souto-Carneiro, Nancy S Longo, Daniel E Russ, Hong-wei Sun, and Peter E Lipsky. Characterization of the human Ig heavy chain antigen binding complementarity determining region 3 using a newly developed software algorithm, JOINSOLVER. *The Journal of Immunology*, 172(11):6790–6802, 2004.

[SKW+05] B Sönnichsen, L B Koski, A Walsh, P Marschall, B Neumann, M Brehm, A-M Alleaume, J Artelt, P Bettencourt, E Cassin, M Hewitson, C Holz, M Khan,

S Lazik, C Martin, B Nitzsche, M Ruer, J Stamford, M Winzi, R Heinkel, M Röder, J Finell, H Häntsch, S J M Jones, M Jones, F Piano, K C Gunsalus, K Oegema, P Gönczy, A Coulson, A A Hyman, and C J Echeverri. Full-genome RNAi profiling of early embryogenesis in Caenorhabditis elegans. *Nature*, 434(7032):462–9, 2005.

[SLH+13] 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.

[SLL15] Y. Safonova, A. Lapidus, and J. Lill. IgSimulator: a versatile immunosequencing simulator. *Submitted*, 2015.

[SLW+09] 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.

[SSRS96] Virginia F Smith, Brenda L Schwartz, Linda L Randall, and Richard D Smith. Electrospray mass spectrometric investigation of the chaperone secb. *Protein science: a publication of the Protein Society*, 5(3):488, 1996.

[SWK+01] 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.

[TLG+12] 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.

[TMYI89] Seiji Tanaka, Yasuhiko Matsushita, Akikazu Yoshikawa, and Katsumi Isono. Cloning and molecular characterization of the gene riml which encodes an enzyme acetylating ribosomal protein l 12 of escherichia coli k 12. *Molecular and General Genetics*, 217(2):289–293, 1989.

[TNI+07] 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.

[TSF+05] Stephen Tanner, Hongjun Shu, Ari Frank, Ling-Chi Wang, Ebrahim Zandi, Marc Mumby, Pavel A Pevzner, and Vineet Bafna. Inspect: identification of posttranslationally modified peptides from tandem mass spectra. *Analytical chemistry*, 77(14):4626–4639, 2005.

[TSK+13] 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.

[TSRV91] John W Tobias, Thomas E Shrader, Gabrielle Rocap, and Alexander Varshavsky. The n-end rule in bacteria. *Science*, 254(5036):1374–1377, 1991.

[TTZ+05] D. Tsur, S. Tanner, E. Zandi, V. Bafna, and P.A. Pevzner. Identification of post-translational modifications via blind search of mass-spectra. In *Computational Systems Bioinformatics Conference, 2005. Proceedings. 2005 IEEE*, pages 157–166, Aug 2005.

[Var96] Alexander Varshavsky. The n-end rule: functions, mysteries, uses. *Proceedings of the National Academy of Sciences*, 93(22):12142–12149, 1996.

[Var04] Alexander Varshavsky. spalogand sequelog: neutral terms for spatial and sequence similarity. *Current Biology*, 14(5):R181–R183, 2004.

[VCK06] Joseph M Volpe, Lindsay G Cowell, and Thomas B Kepler. SoDA: implementation of a 3D alignment algorithm for inference of antigen receptor recombinations. *Bioinformatics*, 22(4):438–444, 2006.

[VGR+14] Katherine E Varley, Jason Gertz, Brian S Roberts, Nicholas S Davis, Kevin M Bowling, Marie K Kirby, Amy S Nesmith, Patsy G Oliver, William E Grizzle, Andres Forero, et al. Recurrent read-through fusion transcripts in breast cancer. *Breast cancer research and treatment*, 146(2):287–297, 2014.

[VSP+12] 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.

[VSW+13] Christopher Vollmers, Rene V Sit, Joshua A Weinstein, Cornelia L Dekker, and Stephen R Quake. Genetic measurement of memory b-cell recall using antibody repertoire sequencing. *Proceedings of the National Academy of Sciences*, 110(33):13463–13468, 2013.

[Wal06] Christopher Walsh. *Posttranslational modification of proteins: expanding nature's inventory*. Roberts and Company Publishers, 2006.

[WB99] Kenneth W Walker and Ralph A Bradshaw. Yeast methionine aminopeptidase i alteration of substrate specificity by site-directed mutagenesis. *Journal of Biological Chemistry*, 274(19):13403–13409, 1999.

[WBL+13] Yariv Wine, Daniel R Boutz, Jason J Lavinder, Aleksandr E Miklos, Randall A Hughes, Kam Hon Hoi, Sang Taek Jung, Andrew P Horton, Ellen M Murrin, Andrew D Ellington, et al. Molecular deconvolution of the monoclonal antibodies that comprise the polyclonal serum response. *Proceedings of the National Academy of Sciences*, 110(8):2993–2998, 2013.

[WCB+15] Sunghee Woo, Seong Won Cha, Stefano Bonissone, Seungjin Na, David Lee Tabb, Pavel A Pevzner, and Vineet Bafna. Advanced proteogenomic analysis reveals multiple peptide mutations and complex immunoglobulin peptides in colon cancer. *Journal of proteome research*, 2015.

[WCM+14] 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.

[WCN+14] Sunghee Woo, Seong Won Cha, Seungjin Na, Clark Guest, Tao Liu, Richard D Smith, Karin D Rodland, Samuel Payne, and Vineet Bafna. Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data. *Proteomics*, 14(23-24):2719–2730, 2014.

[WJW+09] J.A. Weinstein, N. Jiang, R.A. White, D.S. Fisher, and S.R. Quake. High-throughput sequencing of the zebrafish antibody repertoire. *Science*, 324(5928):807–810, 2009.

[WLM+14] Jason R Woo, Michael A Liss, Michelle T Muldong, Kerrin Palazzi, Amy Strasner, Massimo Ammirante, Nissi Varki, Ahmed Shabaik, Stephen Howell, Christopher J Kane, et al. Tumor infiltrating b-cells are increased in prostate cancer tissue. *J Transl Med*, 12:30, 2014.

[WSW+12] 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.

[WWZ+08] Xiaojing Wang, Di Wu, Siyuan Zheng, Jing Sun, Lin Tao, Yixue Li, and Zhiwei Cao. Ab-origin: an enhanced tool to identify the sourcing gene segments in germline for rearranged antibodies. *BMC Bioinformatics*, 9(Suppl 12):S20, 2008.

[WZ13] 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.

[Xia07] Xuhua Xia. The+ 4g site in kozak consensus is not related to the efficiency of translation initiation. *PLoS One*, 2(2):e188–e188, 2007.

[Yan81] M. Yannakakis. Computing the minimum fill-in is np-complete. *SIAM J. Alg. Disc. Meth.*, 2:77–79, 1981.

[YISI87] Akikazu Yoshikawa, Setsuko Isono, Abraham Sheback, and Katsumi Isono. Cloning and nucleotide sequencing of the genes rimi and rimj which encode enzymes acetylating ribosomal proteins s18 and s5 of escherichia coli k12. *Molecular and General Genetics MGG*, 209(3):481–488, 1987.

[YMMO13] Jian Ye, Ning Ma, Thomas L. Madden, and James M. Ostell. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Research*, 41(W1):W34–W40, 2013.

[ZB08] 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.

[ZWW+14] Bing Zhang, Jing Wang, Xiaojing Wang, Jing Zhu, Qi Liu, Zhiao Shi, Matthew C Chambers, Lisa J Zimmerman, Kent F Shaddox, Sangtae Kim, et al. Proteogenomic characterization of human colon and rectal cancer. *Nature*, 2014.