The Dissertation Committee for Andrew Pitchford Horton Certifies that this is the approved version of the following dissertation:

Methods for Proteomic Characterization of Antibody Repertoires and *De Novo* Peptide Sequencing

Committee:

Edward Marcotte, Supervisor

George Georgiou

Jennifer Brodbelt

Ning Jenny Jiang

Gregory Ippolito

Methods for Proteomic Characterization of Antibody Repertoires and *De Novo* Peptide Sequencing

by

Andrew Pitchford Horton, B.S.Biomed.E.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2016

To my mom.

Acknowledgements

I have been surrounded by so many wonderful people throughout this long journey, and this dissertation is in a large part due to them. To my colleagues, my family, and my friends inside and outside of science, thank you so much. I want to first acknowledge my advisor Edward Marcotte. Thank you, Edward, for your guidance, the freedom to follow my interests, and for engendering in your lab the ideal environment for scientific inquiry and growth. You are milling the keys to unlock a whole new systems-wide understanding of biology. May you and your lab one day open that door.

To all my friends and colleagues in the Marcotte lab, past and present, thank you for helping the time go by too fast. I can't imagine a place with a higher concentration of brilliant and fun people performing fascinating research at the intersection of so many different domains. I am incredibly lucky to have been able to work with each of you. I am also grateful to my friend, Jeff Plaisance, for teaching me so much from his wealth of computer science knowledge and programming experience. I am further indebted to Daniel Boutz, master of mass spectrometry, for showing me the ways of proteomics. Dan, you are a model scientist, and I owe so much to our work together through the years. To George Georgiou and Jenny Brodbelt, thank you for providing me the research opportunities in your fields of interest. I have enjoyed immensely the wonderful and unique puzzles your data present and look forward to continued collaboration. In addition to the aforementioned, I wish to thank the many others that contributed over the years, including Costas Chrysostomou, Victoria Cotham, Kam Hon Hoi, Gregory Ippolito, Jason Lavinder, Jiwon Lee, Bill Press, Scott Robotham, Sebastian Schätzle, Christine Vogel and Yariv Wine. To all others I have had the pleasure of working with, thank you, too. You are not forgotten.

Finally, I must thank my family, the most important people in my life. To my parents, Eileen and Bill, thank you for everything. Your unwavering support and encouragement made all of this possible, and I could not have asked for more. Thank you, Mom. You gave me the curiosity necessary for research, the perfectionistic tendency that kept me here so long, and the drive to keep going. I wish you were still here and miss you every day. Dad, thank you for skipping the bedtime stories and going straight to discussion of power plant scrubber systems (and other topics). Rather than put me to sleep, it helped awaken my engineering mindset. Thank you, too, for being my role model of an honest and good person. And to my sister, Leah, thank you for your friendship and for being the person I can talk to about anything, regardless of how serious of silly it may be.

Methods for Proteomic Characterization of Antibody Repertoires and *De Novo* Peptide Sequencing

Andrew Pitchford Horton, Ph.D.

The University of Texas at Austin, 2016

Supervisor: Edward M. Marcotte

Driven by the increased performance and availability of protein mass spectrometry and next generation sequencing technologies, research in proteomics and systems biology has expanded far beyond the study of model organisms. This heralds a deeper understanding of biology, the world, and human health. However, it also brings significant new challenges to the interpretation of sequencing and mass spectrometry data, the current software tools ill-suited for many modern studies. The first half of this dissertation explores some of these challenges and solutions in the context of a particularly demanding domain – that of serological antibody proteomics. Our team has developed a combined sequencing and proteomics approach for profiling the human serum antibody repertoire. This opens an unprecedented view into the nature of the adaptive immune system and provides insight on antibody repertoire dynamics in both health and disease. The platform also provides effective means to evaluate vaccine efficacy and identify potential antibody therapeutics. Chapter 1 reviews recent advances in and results from such molecular level characterization of the serum antibody repertoire. Detailed in the second chapter, challenges specific to antibody repertoire proteomics preclude the use of standard analysis methods and motivated our development of novel tools and approaches for interpreting serum repertoire proteomic data. I will shift focus in chapters 3 and 4 to present an

experimental and computational workflow for accurate and full-length *de novo* peptide sequencing. We applied 351 nm ultraviolet photodissociation (UVPD) on chromophore-tagged peptides and developed software for sequencing the resultant UVPD mass spectra. Improvements described in chapter 4 enable the software to automatically learn from and interpret new types and combinations of spectra from the same precursor peptide. We demonstrate the effectiveness of this machine learning framework on CID/UVPD spectral pairs and obtain results, from low resolution spectra, comparable to current state of the art. Continued development of these *de novo* interpretation and sequencing methods, in part or in whole, may sidestep many of the remaining challenges facing repertoire proteomics, and successful application of these efforts promises further advancement in antibody repertoire characterization and understanding.

Table of Contents

List of Tables

List of Figures

Chapter 1: Next-generation sequencing and protein mass spectrometry for the comprehensive analysis of human cellular and protein antibody repertoires

Recent developments of high-throughput technologies are enabling the molecular-level analysis and bioinformatic mining of antibody-mediated (humoral) immunity in humans at an unprecedented level.* These approaches explore either the sequence space of B-cell receptor repertoires using next-generation deep sequencing (BCR-seq), or the amino acid identities of antibody in blood using protein mass spectrometry (Ig-seq), or both. Generalizable principles about the molecular composition of the protective humoral immune response are being defined, and as such, the field could supersede traditional methods for development of diagnostics, vaccines, and antibody therapeutics. Three key challenges have driven recent advances: (1) incorporation of innovative techniques for paired BCR-seq to ascertain the complete antibody variable-domain VH:VL clonotype, (2) integration of proteomic Ig-seq with BCR-seq to reveal how the serum antibody repertoire compares with the antibody repertoire encoded by circulating B cells, and (3) a demand to link antibody sequence data to functional meaning (binding and protection).

INTRODUCTION

Since the landmark discovery of antibody (or immunoglobulin) in blood serum more than 100 years ago, we now know conclusively that serum is composed of a complex spectrum of distinct antibodies (is polyclonal) which are generated by individual B-cell clones through extraordinary modes of genetic recombination, diversification, and selection by antigen (*anti*body *gen*erator) according to rules outlined in the paradigmatic

* This chapter draws heavily from [Lavinder, J. J.; Horton, A. P.; Georgiou, G.; Ippolito, G. C. *Current Opinion in Chemical Biology* **2015**, *24*, 112–120.] and less so from [Wine, Y.; Horton, A. P.; Ippolito, G. C.; Georgiou, G. *Current Opinion in Immunology* **2015**, *35*, 89–97.]. I contributed writing, research, editing, and the illustrations.

"clonal selection theory". Remarkably, however, there had been no way to identify, and determine the relative concentrations, of the monoclonal antibodies that compose the serum polyclonal pool elicited in response to vaccination or natural infection, until recently.[1–3] Understanding the composition of the antigen-specific serum antibody *protein* repertoire, the properties (e.g. affinities, epitopes recognized) of the respective immunoglobulins, and finally, the relationship between circulating immunoglobulin and the presence of clonally expanded B cells is profoundly important for the comprehensive understanding of humoral antibody responses.

The current era of modern genomics and proteomics is providing extraordinary new tools for examining antibody repertoires. Next Generation Sequencing (NGS) allows millions of B cell immunoglobulin sequences to be obtained in a single experiment, and NGS approaches to studying the human antibody repertoire[4] not only aim to aid in the discovery of elite antibodies potentially useful as therapeutics, but also to comprehensively catalogue the antibody sequences that are elicited during an adaptive immune response.[5] Previously a limitation with NGS, the ability to obtain the endogenous variable heavy and light chain (VH:VL) pairs within NGS datasets is now feasible.[1,6] This paired VH:VL sequencing represents a major breakthrough in repertoire analysis, obviating the need for multiplexed screening to identify functionally paired VH and VL. NGS has also provided a stepping stone to the direct characterization of serum antibodies using NGS database-driven high resolution mass spectrometry,[1–3,7] providing a direct means to comprehensive delineation of the antibody repertoire.

TWO ANTIBODY REPERTOIRES: THE CELLULAR AND THE SEROLOGICAL

B cells, serum immunoglobulin, and antibody repertoire persistence

Antibody molecules are composed of two heavy (H) and two light (L) chains, and antigen binding specificity is determined by the variable region of the antibody (or B cell antigen receptor, BCR) gene. This region is not coded for in the germline genome but arises through V(D)J recombination, a process through which one each of V, (D) and J germline gene segments are selected from a combinatorially enormous pool and joined together (Figure 1.1). The heavy chain variable region (VH) contains a VDJ junction, and nucleotides are added or removed in the V-D and D-J joining sites at random. The CDRH3 (complementary determining region 3, heavy) spans the D segment and flanking junctions and is the primary determinant of antibody specificity as well as the most variable part of an antibody molecule. The light chain VL region arises through a similar process of recombination of V and J segments. Interspersed between the CDRs (of which there are three in both the VH and VL chains) are the highly conserved framework regions (FR). Further antibody diversity develops through somatic hypermutation (SHM) of the CDRs during affinity maturation, whereby a single B cell lineage may branch out into a multitude of clonally related but distinct B cells.[8,9] Antibodies secreted from clonally related plasma B cells will have largely the same antigen specificity but may differ in affinity, or binding strength.

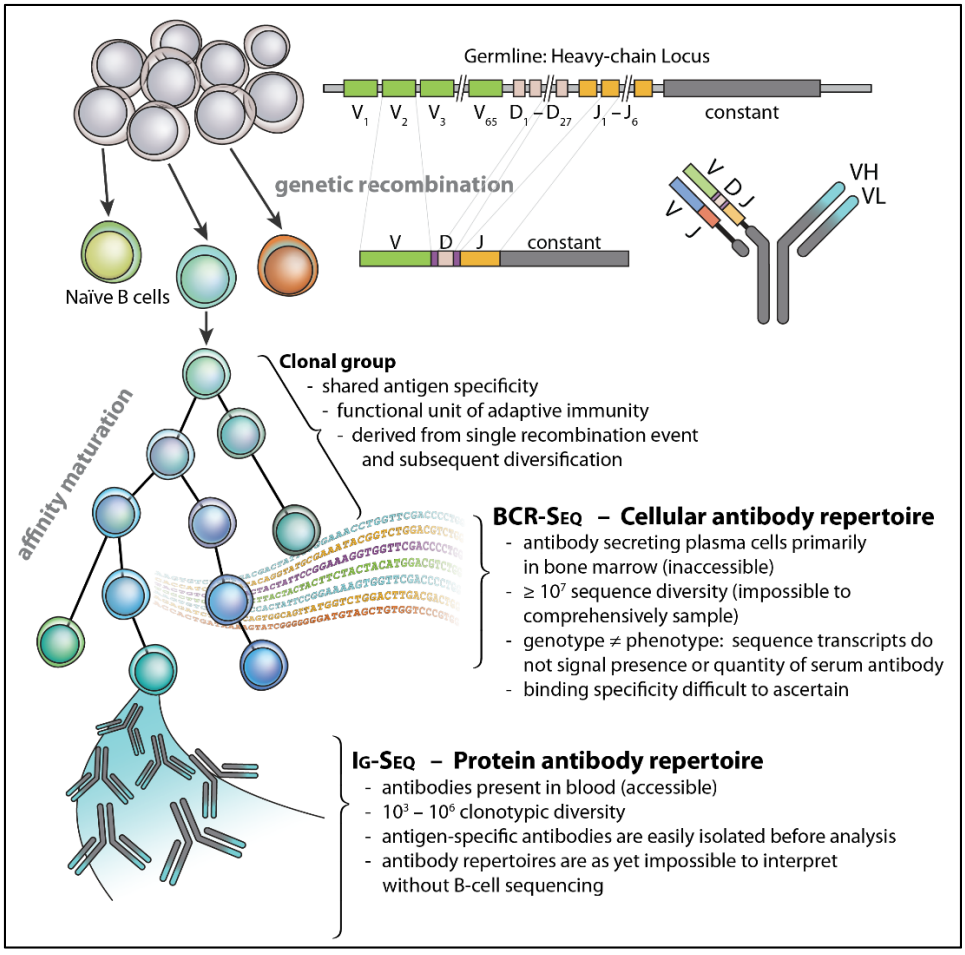

Figure 1.1: Generation and composition of the antibody and B cell repertoires.

Whereas all newly formed B cells express antibody on their surface as BCR, and subsequently emigrate from their generative site in bone marrow to seed the periphery, it is only a small minority that might ultimately differentiate during the course of an immune response to become memory B cells (mBCs) and an even smaller fraction that secrete their BCR as soluble antibody. In this regard, we can conceive of the functional antibody repertoire as consisting of two major components: (1) the set of BCRs expressed on the surface of B lymphocytes, and (2) the collection of soluble Ig found in blood and secretions, produced predominantly (>95%) by terminally differentiated plasma cells in the bone

marrow (PCs, BMPCs) (Figure 1.1).[10] Humoral immunity against pathogens can be sustained for greater than a half century, requiring steady-state expression of serum antibodies that are believed to be maintained by long-lived BMPCs.[10] However, bone marrow specimens are often impractical to obtain in humans, and the vast majority of studies that examine the human antibody repertoire interrogate peripheral plasmablasts (PBs), circulating mBCs, or all peripheral blood mononuclear cells (PBMCs).

In summary, the two major components of the antibody repertoire—the quiescent mBC cellular and the plasma cell-secreted serological—are both generated during a primary response to antigen and persist for sustained yet indeterminate life spans. It is therefore reasonable to ask what degree of overlap is present in the molecular composition of these compartments. If they are not congruent, as there is reason to think,[9] then what might be the selective mechanisms that govern their differential recruitment, and what might be the consequences to protective humoral immunity?

Antibody Serology

The essence of serological immunity is predicated on the existence of a diverse repertoire of antibodies, elicited over the life of the host and representing the integrated response to numerous antigenic stimuli. Due to the complexity and temporally dynamic nature of the antibody repertoire, the identification of its component immunoglobulins represents a formidable challenge. Since antigen can trigger B cells to proliferate, mutate, and expand, it is a useful metric to enumerate the clonotypes. In the classical sense, a clonotype is defined as a repertoire of unique B-cell specificities,[11] and serological studies to date have relied on the detection of an *ensemble* of antibodies that either could be resolved by a certain analytical technique or bound to a specified antigen (Figure 1.2). Among the most useful metrics for assessing humoral immunity, the presence of

neutralizing antibodies in the serum following vaccination or infection represents the best correlate for vaccine efficacy and for protection during invasive infections.[12,13] The limitations imposed by the inability to resolve complex serum antibody mixtures into their constituent clonal representatives and the need to have pre-established the identity of antigens of potential interest have obscured central questions of profound basic and clinical significance.

First and foremost, there is nearly no information on the number of sequences (clonal diversity), functions and relative concentrations of the individual antibodies in serum. Considering that the BCR repertoire diversity in the memory and plasmablast compartments is orders of magnitude greater than that of the serological repertoire[1] it follows that the overwhelming majority of peripheral B cell-encoded antibodies are unlikely to be present in detectable amounts as soluble proteins in blood or secretions and thus do not contribute to humoral immunity. Second, while it is well established that a significant fraction of antibodies display polyreactivity and that these antibodies have important physiological functions in processes such as the clearance of cell debris and in pathogen recognition,[14] there is a paucity of methods for quantifying and characterizing the polyreactive fraction of the serological response. There is a clear need to understand the mechanisms that drive polyreactivity and its implications in health and disease. One possible explanation is that polyreactivity originates from B cells that were not removed from the repertoire during B-cell development. For some pathogens, notably HIV, polyreactivity may confer a selective advantage to pathogen–specific antibodies.[15] Third, in many instances the antigens that are recognized by serum antibodies are not known *a priori*. The significance of identifying antigens, antigen surrogates (i.e. antigen-mimics distinct from the antigens that elicited an antibody response) and immunosignatures for disease diagnosis is being increasingly recognized.[16] Additionally, mapping the serum

antigen reactivity profile in a comprehensive manner is key to understanding which environmental exposures play a more dominant role in shaping humoral immunity.

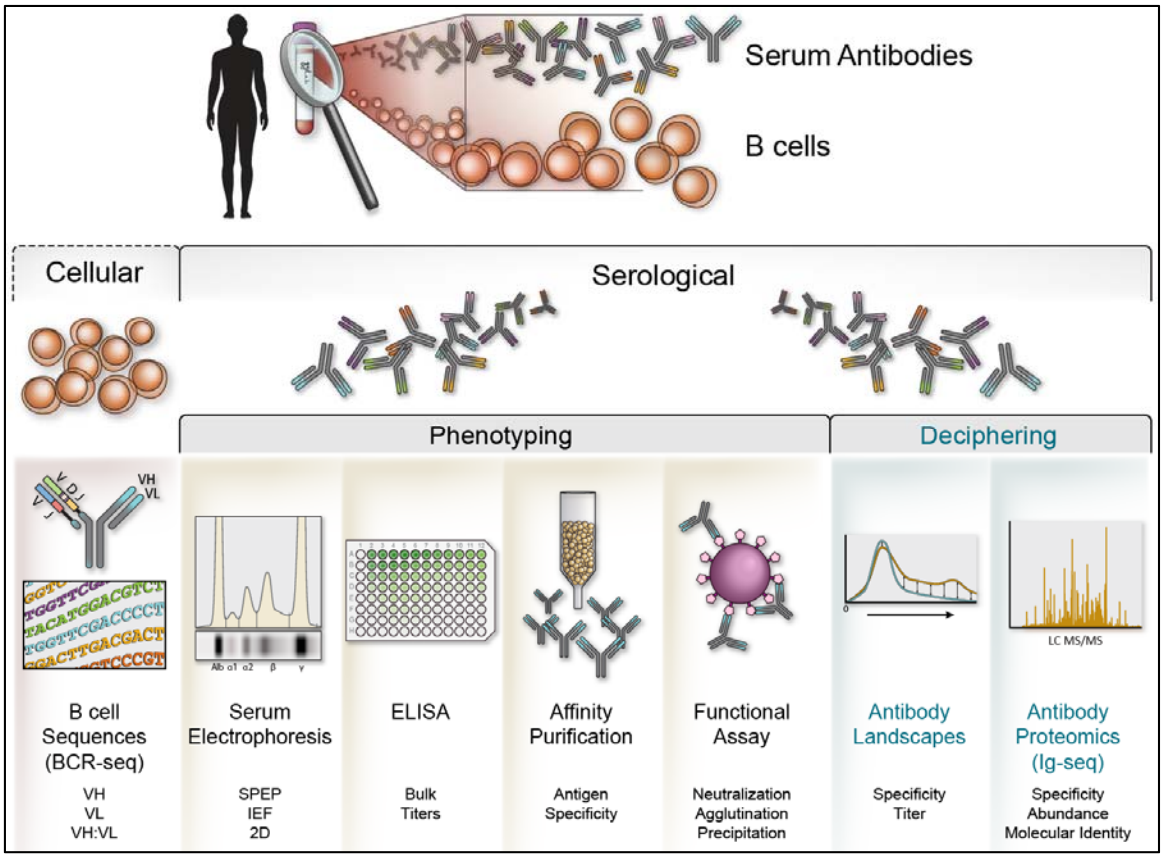

Figure 1.2: Approaches for the analysis of human antibody repertoires.

B cell high-throughput sequencing generates the antibody repertoire encoded by B cells (cellular repertoire, left side of the figure). The corresponding serum immunoglobulins are isolated from the samples and can be analyzed by various methods including well established technologies such as 2D gels or by recently established methodologies such as high resolution shotgun proteomics (serological repertoire, right side of the figure). The methodologies for serological immunoglobulin analysis can be broadly based upon the phenotype of an antibody subpopulation (e.g., ELISA titer of antigen-specific fraction) or upon decipherment of the molecular identity and sequence determination of an antibody subpopulation (e.g., LC–MS/MS immunoglobulin sequencing, Ig-seq).

Defining the Cellular and Serological Antibody Repertoires

The earliest and still most common NGS-enabled studies of the repertoire have focused on the CDR-H3 region—its length, peptide sequence, and IGHV and IGHJ gene usage patterns—to define heavy chain clonotype dynamics. CDR-H3 clonotypes, defined in this regard, typically share the same inferred germline IGHV and IGHJ, CDR-H3 length, and have ≥90% peptide sequence homology. A recent and significant innovation in NGS is the development of methods to maintain the correct pairing of the VH and VL in the B-cell repertoire.[6,17–19] This is achieved through single-cell sorting, VH:VL linkage PCR performed in an emulsion or single-cell wells, and NGS. With regard to clonotyping methods, this technological advance will allow more accurate assignment of the complete VH:VL antibody clonotype and can additionally take account of important features in VL domains.

Until recently determining the sequence and relative concentration of the antibodies in the serum repertoire was considered a nearly impossible task: biological fluids contain many thousands of different antibodies all of which are chemically very similar, having an overall high degree of sequence identity, and whose concentrations can vary by several orders of magnitude in a dynamic fashion. 2008 saw the first use of LC-MS/MS for serum immunoglobulin peptide detection.[20,21] However, the Ig-derived peptides detected in these earlier studies, were overwhelmingly derived from the framework regions and did not provide sufficient information to piece together complete antibody sequences. By restricting the diversity of the antigen-specific antibody pool from the serum of immunized rabbits using antigen-affinity chromatography under stringent elution conditions, Polakiewicz and coworkers succeeded in using LC-MS/MS with NGS to identify complete V genes. Combinatorial pairing of separate identified VH and VL sequences was then used

to produce several antibodies displaying high affinity for antigen, first from rabbits and subsequently from humans.[3,22]

In an alternate approach, our lab independently invented a technology for determining the serological repertoire to individual antigens (Figure 1.3). This combined: (i) V gene sequencing from peripheral memory B cells and plasmablasts; (ii) enrichment of the pool of antigen-specific antibodies by affinity chromatography; (iii) identification of immunoglobulin peptides using bottom up LC-MS/MS and searching against the V gene sequences; and (iv) the application of stringent informatics filters and *in silico* clonotype estimation to identify antigen-specific VH genes from the peptide assignments.[2,7,23] Comparison to the natively paired VH:VL sequence repertoire could then reveal the complete antibody sequence, which then could be produced and studied *in vitro*. Thanks to the exquisite sensitivity of modern MS instrumentation, individual serum antibodies can be detected semi-quantitatively at levels as low as 0.4 ng/ml.[1] For an antibody to bind to antigen it has to be present at a concentration at or above its equilibrium dissociation constant, which is estimated to have a floor of around 0.1 nM.[24] Thus, the approach outlined above has more than adequate sensitivity for the detection of the repertoire of physiologically relevant antibodies in a sample.

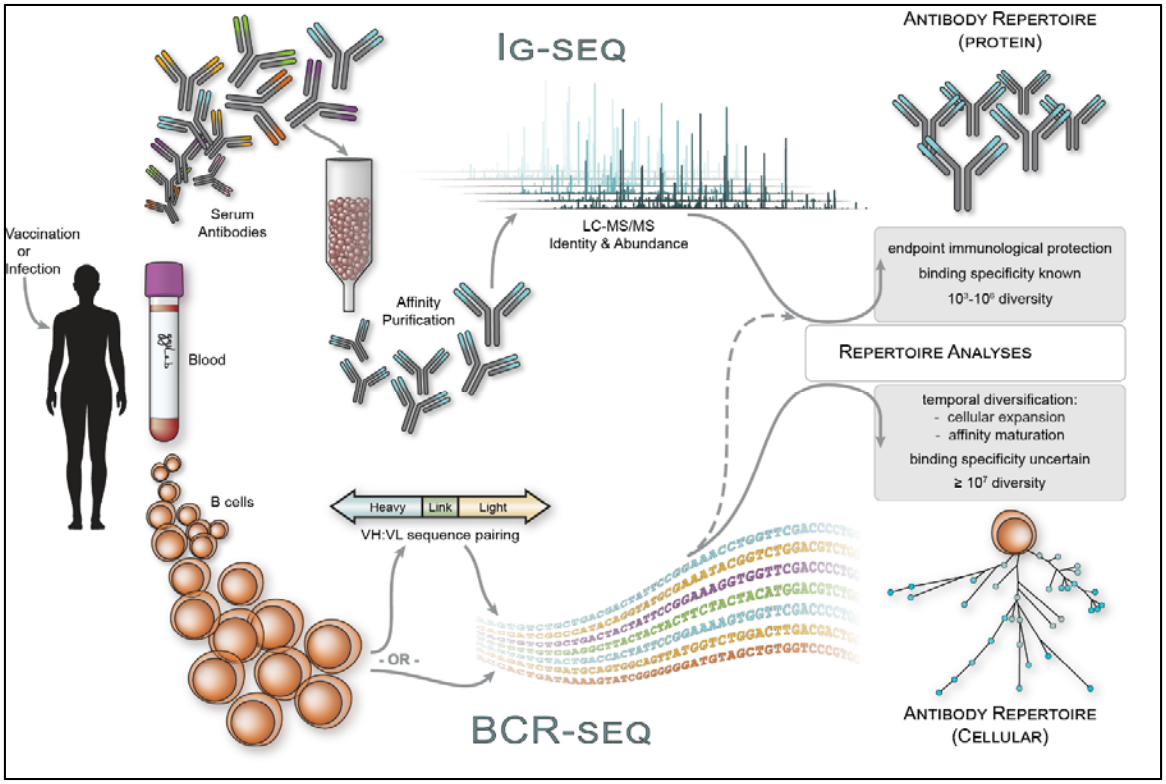

Figure 1.3: NGS and MS analysis of human antibody repertoires from peripheral blood.

The functional antibody repertoire consist of two major components: (bottom) the total set of BCRs expressed on the surface of peripheral blood B cells, and (top) the collection of soluble serum antibody circulating in the blood. The ability to compare and functionally characterize these two types of antibody repertoires provides a new paradigm in the study of the humoral response. This involves the isolation and proteomic analysis of affinity purified serum antibody (lg-seq, top) in parallel with VH:VL pairing and/or NGS of peripheral B cell V gene repertoires (BCR-seq, bottom). The bioinformatic analyses of both the diversified cellular humoral immune response and the endpoint serological antibody response provides an avenue for effective antibody discovery, exhaustive antibody repertoire characterization, and an improved understanding of humoral immunity.

Generalizable Principles from Normal Antibody Repertoires

Natural variation within the primary repertoire

Because of the sheer size of the human antibody repertoire, estimates of breadth and diversity become limited by sampling and only recently have such statistical

approaches become feasible. In one study, the total human peripheral B cell repertoire size was estimated to contain ~10^6-10^7 unique CDR-H3.[25] However, the theoretical sequence space of the repertoire far exceeds the antibody diversity found within an individual at any one point in time, and shared antibody sequences are extremely uncommon.[26–28] Yet, at a broader level, the overall usage of germline IGHV, IGHD, and IGHJ segments is observed to be unequally distributed, yet at a very consistent ratio among individuals, indicating a measured amount of determinism in the generation of primary antibody diversity.[25] It was subsequently shown that this determinism likely arises from genetic factors intrinsic to human B cells.[29–33]

Unlike the antigen-specific mBC repertoire, which is very diverse and in the case of chronic or repeated infections can comprise millions of distinct clones and a much larger space of unique VH sequences, the serological repertoire is orders of magnitude more restricted. There is indirect evidence that these compartment also differ in humans.[9] Likewise, comparison of the antigen-specific VH gene repertoire of transient PB cells with the steady state serological repertoire (in other words, the serological memory) has revealed that a very small fraction of the CDR-H3 clonotypes encoded by peripheral B cells are observed in the polyclonal serum response (<5% of the CDR-H3 clonotypes in the peak response PB repertoire and <0.1% in the steady-state peripheral mBC repertoire).[1]

The dramatic discordance between humoral immunity and the VH gene repertoire in antigen-stimulated peripheral B cells is not widely appreciated but it can be readily illustrated with a simple quantitative analysis of humoral immunity. For circulating antibodies to be physiologically relevant, they have to be present in serum at concentrations exceeding their equilibrium binding constant, K_D. Assuming an average K_D of IgGs to persistent antigen exposure or re-stimulation of 1-5 nM (or approx. 0.2-1.0 μg/ml) and given that serum titers to pathogens rarely exceed 100 mg/ml it follows that the diversity

of physiologically relevant antibodies present in the serological repertoire must be of the order of 10^2- 10^3 (100 μg/ml divided by 0.2-1 μg/ml), or 3 or more orders of magnitude smaller than the typical antigen-specific peripheral mBC repertoire. With regard to an *upper* bound, the total serological antibody repertoire could be as large as 10^6 distinct binding specificities (clonotypes).[34] This argument assumes that (1) the lowest Ab concentration required for the elimination of antigen ~ 10 ng/mL, (2) there is 10 mg/mL total IgG in serum, and (3) each distinct Ab is present at threshold concentration. Even at 10^6, this still places the serological repertoire at least 10^1 smaller, perhaps 10^3 smaller, than the cellular repertoire by lower-bound NGS estimates.[25] Proteomic analyses from our lab estimate IgG clonotypic diversity of $\geq 10^4$. The discrepancy between the peripheral mBC (and also the antigen-specific peak response PB repertoire) and the serum antibody repertoire argue strongly that determination of the serological repertoire is critical for a comprehensive understanding of antibody- mediated protection mechanisms.

GENERALIZABLE PRINCIPLES FROM VACCINE-INDUCED ANTIBODY REPERTOIRES

BCR-seq of vaccine-specific VH antibody repertoires

Almost all vaccines confer immunity through the induction of antibodies in serum or in mucosal tissues.[12] Systematic analysis of the antibody-mediated humoral immune response to vaccination at high-throughput requires experimental distinction between the vaccine-specific and the total antibody repertoire in an individual. One approach to inferring antigen specificity is to examine the dynamics in the peripheral B cell repertoire before and after vaccination. Jackson et al. showed a direct correspondence between the number of clonally expanded peripheral B cell lineages at day 7 post-vaccination and an increase in serum titer.[35] Jiang et al. similarly used antibody dynamics before and after vaccination to identify expanded vaccine-specific antibody lineages within the peripheral

B cell repertoire.[36] They found that 11 of 16 confirmed influenza-specific VH sequences mapped to the expanded lineages. Further, lineage structure analysis revealed that elderly patients have fewer, more highly mutated IgG lineages as compared to younger patients. Wu et al. also detected higher IgG mutation in the elderly after influenza vaccination, as well as significantly longer CDR-H3 in the IgM and IgA lineages that were expanded at day 7 post-vaccination.[37] Clearly, these studies and others indicate the significance of antibody diversity and CDR-H3 characteristics to the humoral immune response and support the use of such metrics for studying vaccine efficacy.[38,39] A refinement of these metrics will include a transition from VH-only BCR-seq to BCR-seq of complete VH:VL clonotypes, as well as a quantitative exploration of their absence or presence in the serological repertoire using Ig-seq.

BCR-seq of vaccine-specific paired VH:VL antibody repertoires

Other, more conventional methods for distinguishing antigen specificity include the labeling of antigen-specific mBCs or the isolation of bulk PBs using flow cytometry.[40] For a variety of viral infections and most immunizations, the appearance of vaccine-specific PBs is strikingly consistent in that they peak in number at approximately one week post-vaccination, or day 10 for primary vaccinations.[41] This "plasmablast signature" and its predictive capacity for the *magnitude* of antibody production has been observed by several research groups;[42,43] how this might relate to the functional quality or exact molecular nature of the endpoint serological antibody response is unknown and has yet to be comprehensively examined, but current investigations are *very* tantalizing.[35]

As but one example, DeKosky et al. isolated tetanus toxoid (TT) specific PBs on day 7 post-vaccination and, using emulsion linkage PCR and NGS from 200 sorted cells, identified 86 TT-specific antibody VH:VL pairs in a single experiment.[6] A significant

improvement in throughput utilizes a newly-developed, low-cost single-cell emulsion-based technology which flows B cells through a vibrating nozzle to encapsulate individual B cells in lysis/PCR reaction droplets that contain magnetic beads for mRNA capture.[17] VH:VL amplicons generated through subsequent emulsion RT-PCR are used for 2 x 300 Illumina MiSeq NGS. This method increases the B-cell VH:VL yields 100X, to >$2x10^6$ B cells per experiment with demonstrated pairing precision >97%. This orders-of-magnitude increase in B-cell throughput and VH:VL sequencing depth potentially obviates an explicit need for antigen-specific cell sorting because literally millions of B-cells can be processed and interrogated in a single experiment by a single experimentalist. Thus, an advantage of this method is that it allows complete sequencing of *all* antigen-specific B-cells within a finite collected pool (e.g., day 7 PBs or total mBCs at 14 days post-booster vaccination).[40]

Serum antibody proteomics (Ig-seq) of vaccine-specific antibody repertoires

It is now also possible to identify affinity-purified serum antibodies using high-resolution proteomics.[1–3,7] The goal of serum antibody proteomics, or Ig-seq, is to systematically identify the distinct antibodies present in a serum sample, as assayed using protein tandem mass spectrometry (Figure 1.3). In a study of the immune response to tetanus toxoid (TT), Lavinder et al. exhaustively characterized the constituent serum antibodies elicited by a vaccine and discovered that the steady state anti-TT serum IgG repertoire is composed of a limited number of antibody clonotypes (80-100), with three clonotypes accounting for >40% of the response.[1] Only a small fraction (<5%) of TT-specific, vaccine-responsive PBs at day 7 were found to encode antibodies that could be detected in the serological memory response 9 months post-vaccination. This suggests that only a minority of the antigen-specific, transient PBs give rise to bone marrow long-lived plasma cells (BMPCs). This result is not altogether unexpected since huge variability in

both the number, kinetics, and the antigen-specificity of transient PBs has been repeatedly observed for a variety of vaccination and natural infection contexts.[44] This is also in agreement with previous data demonstrating that only a fraction (5-10%) of responding PBs migrate to the bone marrow after vaccination.[45] These differences in the peak responding PBs and the effective levels of serum antibodies is a significant finding in that the antigen-specific repertoire in vaccinated humans is typically assessed by DNA sequencing of these responding PBs, a large number of which do not constitute the post-boost steady-state antigen-specific serum IgG repertoire.

Molecular convergence of antibody responses

Lastly, in the course of the TT study summarized above, we discovered a VH clonotype shared between two donors (a stereotype) and also with a third donor analyzed independently in a distinct laboratory,[46] providing intriguing evidence for the existence of "public" clones, or "convergent immune signatures," emerging after antigen challenge. NGS deep sequencing of VH repertoires has discovered the emergence of stereotyped serological clones in other vaccinations, such as seasonal influenza H1N1 vaccination[35] and Dengue viral infection,[47] and would in principle be detected by the methodology developed here. The convergent detection of stereotyped or "public" serum clonotypes detected by Ig-seq might correlate strongly with vaccine efficacy, seroconversion, or the production of neutralizing antibodies.

With certain antigens, convergence has been shown to be quite prevalent, producing CDR-H3 lineages that are universally identified across individual antibody repertoires. Vollmers et al. developed a NGS barcoding technique that uniquely identified each starting VH or VL transcript.[26] This allowed consensus-based filtering of sequencing error to allow an accurate measurement of the memory recall response to vaccination. It also provided an

accurate measurement of antibody repertoire convergence, revealing that 25 of ~100,000 VH sequences were shared between vaccinated individuals. However, all of these were of low abundance and had very low amounts of mutation and short CDR-H3, indicating stochastic overlap within the naïve B cell repertoire. Jackson et al., however, identified the molecular convergence of an antibody response to influenza H1N1 vaccination when comparing data with H1N1-vaccinated donors from two additional studies, revealing a stereotypic VH lineage that utilized the same V segment, J segment, and highly identical CDR-H3s;[35] although a striking result, it remains to be determined if this stereotyped rearrangement exists not merely as a VH-only but also as a VH:VL clonotype, and whether this clonotype exists in the serum antibody repertoire or can be correlated with seroconversion and viral neutralization. It is not entirely known how common stereotypic B cell responses are in vaccines. However, as detailed below, shared antibody sequences have enormous significance as potential biomarkers in both infectious disease and autoimmunity.[16]

THE ANTIBODY REPERTOIRE IN THE DISEASE STATE

Infectious disease

Broadly neutralizing antibodies (bNAbs) directed against HIV and influenza viruses have been identified via the cloning of antibody V genes from peripheral B cells isolated from infected patients that displayed neutralizing serum titers.[48–50] Sequence analyses of bNAbs and homologous V genes of antigen-specific cells from the individual from which the bNAbs had been isolated, are providing insights on the evolution of broadly protective B cell immunity, on preferential usage of certain germline V genes, on somatic hypermutation patterns, and other features of neutralizing immune responses.

The significance of the information gained from these studies notwithstanding, it is not clear whether peripheral B cell-encoded bNAbs actually play a dominant role in the serological protection against infection *in vivo*. As discussed above, serological immunity is overwhelmingly contributed by BMPCs, which are often experimentally inaccessible in humans. Nonetheless, it should be appreciated that it is the cells of the BMPC compartment and not peripheral B cells, which actually secrete the Ig that maintains long-term serological memory. So far, it has not been ascertained whether bNAbs isolated from peripheral mBCs are actually present in the serum at all, let alone at physiologically relevant concentrations (i.e. above K_D) as is required in order for these antibodies to play a role in virus elimination and protection *in vivo*.

The most compelling examples of bNAb functionality is from convergent humoral responses that occur within chronic HIV infection. It is known that up to 25% of individuals with advanced HIV can develop bNAbs against the virus. NGS is now being used to track antibody and viral co-evolution, and the identification of elite bNAbs is now central to the study of HIV.[51] One of these bNAbs, VRC01, is specific to the CD4 binding site of gp120 and can cross-neutralize ~90% of HIV-1 isolates. Like many other identified HIV-specific antibodies, it shows striking amounts of affinity maturation (70 amino acid differences from germline). Wu et al. isolated VRC01-like antibodies from separate HIV-1 infected donors using FACS sorting against a CD4-binding site probe.[52] NGS and phylogenetic analysis of both donors revealed similarities between the affinity maturation pathways for the VRC01-like antibodies and demonstrated how NGS data can be used to identify large clades of antibody sequences based upon selective criteria, such as V(D)J usage, amount of mutation, and sequence identity to known monoclonal antibodies.

Autoimmunity and cancer

Autoantibody repertoires likely contain a wealth of information both in regards to the early diagnosis of immunopathology, as well as providing an increased understanding of disease progression. Unfortunately, very little is known regarding these potentially significant sources of biomarkers. In a pair of recent studies on patients with multiple sclerosis (MS), expanded B cell clonotypes in the periphery (peripheral blood in one study and cervical lymph nodes in the other) were overlapping with B cell sequences found in the CNS of the patient.[53,54] It was shown that the founding members of these overlapping clonotypes were prevalent in the cervical lymph nodes and that overlapping members in the peripheral blood were primarily class-switched B cells. In addition to cross-tissue overlap, convergence has also been detected in the autoantibody response. Doorenspleet et al. used NGS of B cells from peripheral blood and joint synovial fluid from patients with early and establish rheumatoid arthritis (RA), demonstrating potential convergence in the dominant B cell lineages within synovial fluid of early RA patients.[55] These dominant lineages heavily utilized the V segment IGHV4-34 and had significantly longer CDR-H3. A few recent studies have also used high-resolution mass spectrometry to proteomically identify molecular signatures in the autoreactive antibody response to the Ro/La ribonucleoprotein complex in patients with Sjögren's syndrome (SS).[56–58] Although these studies only identified a handful of public (shared) antibody lineages and V gene mutations across SS patients, it highlights the great potential of serum antibody proteomics in autoimmune biomarker discovery.

Similarly, the antibody response to tumor-associated antigens can provide early diagnostic cues in detecting malignancy.[59] A series of studies have utilized NGS, as well as serum antibody proteomics, to facilitate the early detection and monitoring of Non-Hodgkin's lymphoma,[60] leukemias,[61,62] and multiple myeloma.[63,64] Typically, these

lymphocyte malignancies are detected and monitored via PCR specific to the malignant lineage(s), which requires patient-specific primers to examine values of minimal residual disease (MRD). There is great potential for NGS and serum antibody proteomics in developing further metrics for diagnostic and prognostic applications in both autoimmunity and cancer.

CONCLUSIONS

NGS has revolutionized the manner in which we study adaptive immunity, providing millions of sequence reads from an enormously complex repertoire of lymphocytes. However, these gigantic data sets have an inherent interpretability problem in that the sequences specific to "your favorite antigen" are metaphorical needles in a haystack. Determining the functionality of the Ig reads obtained in BCR-seq is an enormous challenge, and it has now become a common appeal in the field to link function with sequence.[65,66] As discussed above, in certain cases, convergence or dynamics within BCR repertoires can often lead the discovery effort for antigen-specific clonotypes in response to vaccination or infection. However, this is not a generalizable strategy and such obvious levels of dynamics or determinism may be restricted to certain antigens or only evident in cases where the humoral response is robust or ongoing.

The recent development of paired VH:VL BCR-seq and Ig-seq represent a new paradigm of antibody discovery in which functionality (binding) is directly linked to NGS of natively-paired antibody gene sequences. This ability to quickly link paired NGS data and antibody functionality is a key step forward in antibody discovery and repertoire analysis. Not only is this applicable to the vaccine-elicited antibody repertoire, but it also enables the link between antibody sequence and function to be ascertained in infectious disease, as well as autoimmunity and cancer.

Chapter 2: Proteomic Identification of Monoclonal Antibodies from Serum

Characterizing the *in vivo* dynamics of the polyclonal antibody repertoire in serum, such as might arise in response to stimulation with an antigen, is difficult due to the presence of many highly similar immunoglobulin proteins, each specified by distinct B lymphocytes.* These challenges have precluded the use of conventional mass spectrometry for antibody identification based on peptide mass spectral matches to a genomic reference database. Recently, progress has been made using bottom-up analysis of serum antibodies by nanoflow liquid chromatography/high-resolution tandem mass spectrometry combined with a sample-specific antibody sequence database generated by high-throughput sequencing of individual B cell immunoglobulin variable domains (V genes). Here, we describe how intrinsic features of antibody primary structure, most notably the interspersed segments of variable and conserved amino acid sequences, generate recurring patterns in the corresponding peptide mass spectra of V gene peptides, greatly complicating the assignment of correct sequences to mass spectral data. We show that the standard method of decoy-based error modeling fails to account for the error introduced by these highly similar sequences, leading to a significant underestimation of the false discovery rate. Because of these effects, antibody-derived peptide mass spectra require increased stringency in their interpretation. The use of filters based on the mean precursor ion mass accuracy of peptide-spectrum matches is shown to be particularly effective in distinguishing between "true" and "false" identifications. These findings highlight

* This chapter has been previously published in: Boutz, D. R.; Horton, A. P.; Wine, Y.; Lavinder, J. J.; Georgiou, G.; Marcotte, E. M. *Analytical Chemistry* **2014**, *86* (10), 4758–4766.
D.R.B, A.P.H., and Y.W. contributed equally.

important caveats associated with the use of standard database search and error-modeling methods with non-standard datasets and custom sequence databases.

INTRODUCTION

The ability of the humoral immune system to provide broad protection against a diverse and constantly changing population of invasive pathogens stems largely from the antigen-binding capabilities of the antibody (immunoglobulin, Ig) repertoire. Antibodies recognize foreign molecules (antigens) through epitope-binding sites in the variable domains of the antigen binding fragment (Fab), and alert immune cells to putative threats through interaction sites in the constant domain of the tail region. Individual antibodies will preferentially bind a particular antigenic epitope, with specificity largely determined by the antigen-binding site sequences in the variable domains of immunoglobulin heavy chain (V_H) and light chain (V_L) genes. In order to provide coverage against a large variety of potential antigens, the B cell-encoded antibody repertoire is incredibly diverse, estimated to comprise $>10^8$ immunoglobulins with distinct variable domain sequences in human serum,[67,68] resulting in an antibody population capable of binding a broad range of antigens with high affinity and specificity.

This massive diversification of sequence is the product of two processes: V(D)J recombination during B cell maturation, and somatic hypermutation during B cell affinity maturation.[69] In the heavy chain specifically, the variable domain is generated by recombination of V, D, and J gene segments, with a single subgene of each segment selected from multiple variants encoded in the germline genome (Figure 2.1). Two of the three hypervariable loops responsible for antigen-binding (CDR-H1 and CDR-H2) are encoded within the V gene segment, while the third (CDR-H3) is largely non-templated and is constructed by the addition of random nucleotides (N-nucleotides) between the

recombination joints of the V, D, and J segments.[69,70] V(D)J recombination generates a single pair of V_H and V_L genes per B cell, such that every B cell expresses only one antibody variant. Somatic hypermutation during humoral immune response fine-tunes affinity for antigen by introducing additional mutations in the variable domain, further increasing the sequence variation and in turn expanding the sequence diversity within a clonotype.[9] Consequently, antibodies that originate from the same B cell precursor lineage are designated as belonging to the same clonotype and generally exhibit specificity for the same antigen.[71]

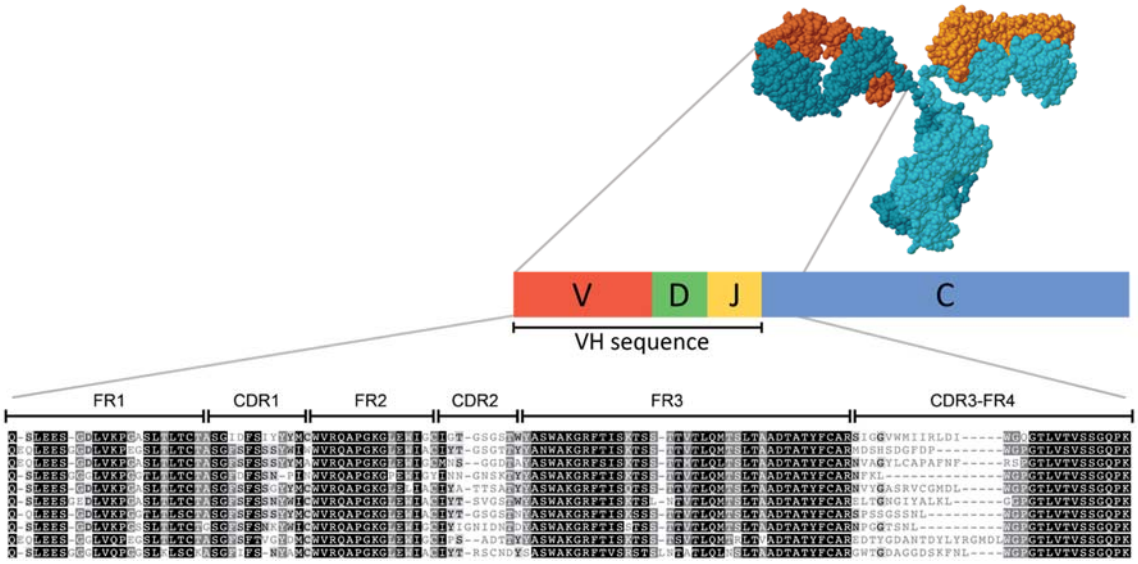

Figure 2.1: Schematic of the structure and representative sequences of the immunoglobulin (Ig) heavy chain variable domain.

The heavy chain variable domain (VH) sequence is created by recombination of V, D, and J subgenes and encodes epitope binding sites for antigen-recognition. Complementarity determining regions (CDRs) represent uniquely non-degenerate fingerprints, interspersed between constant framework sequences (FRs), and manifest as hypervariable and conserved sequences, respectively, in the multiple sequence alignment. Antigen binding specificity is primarily dictated by the CDR-H3 region. Hence, the challenge of antibody repertoire proteomics can be largely reduced to the problem of successfully identifying CDR-H3-containing peptides.

The process of Ig diversification has been elucidated, and methods for the identification and expression of monoclonal antibodies, including creation of hybridomas, immortalization of B lymphocytes, and cloning of antibody genes from primary lymphocytes, have revolutionized diagnostics and expanded our understanding of how immune responses induce the production of circulating antibodies that help clear a pathogen. Recently, next-generation (NextGen) sequencing has made possible investigations of the scope and sequence composition of the antibody repertoire, as represented in the population of B cells sequenced.[72,73] With technical and financial barriers to personalized sequencing substantially dropping with advances in NextGen technologies, immune-related repertoire sequencing is becoming more commonplace.[26,74] However, the B cell repertoire includes many sequences which are not represented in the circulating pool of serum immunoglobulins. Characterization of the polyclonal serum response thus requires direct observation of the constituent monoclonal antibodies present at functionally relevant concentrations.

Unfortunately, the proteomic analysis of serum immunoglobulins by mass spectrometry (MS) presents several challenges. One such challenge arises from the fact that antibody genes are not encoded in the germline but are assembled *via* DNA recombination and diversified within individual B cells. As a result, the typical strategy of constructing a reference database from the genome sequence is not useful for interpreting antibody-derived mass spectra.[21,75] The use of *de novo* peptide sequencing for mass spectral interpretation does not require a reference database,[76,77] thus offering a promising solution to this problem. Current methods are not yet capable of handling the complexity of peptide sequence diversity present in serum. With further development, the *de novo* workflow we present in chapter 4 could potentially suffice. However, a reference database would still be necessary to identify full-length V_H sequences.

A strategy has recently emerged which largely overcomes these barriers by utilizing high-throughput sequencing of the immunoglobulin variable domain (V gene) from an individual's B cell population to construct a sample-specific antibody sequence database for the interpretation of antibody-derived mass spectral data.[3,7,78] With the ability to generate a personalized reference database it is now possible to apply shotgun-style MS proteomics to the analysis of serum antibodies, as demonstrated by recent studies identifying antigen-specific monoclonal antibodies directly from serum.[1,3,7,22] Yet even with the availability of such a database, confident identification of monoclonal antibodies is not trivial. The high degree of sequence identity shared across antibodies introduces additional complications in sequence-to-spectrum assignments and protein inference, making proteomic analysis of the repertoire particularly challenging.

The complexity of the V gene repertoire can best be understood as a massively expanded set of homologous proteins, each sharing regions of highly conserved (or identical) sequences with short intervening hypervariable sequences. From a proteomics perspective, this creates a large pool of potential peptide sequences with at least partial sequence identity. Proteolytic digestion of antibodies for shotgun proteomics yields many peptides that map to multiple clonotypes and are therefore non-informative for monoclonal antibody identification, or that share partial sequence identity with many other candidate peptides, resulting in highly similar mass spectra that are difficult to interpret unambiguously, even with the high resolution and mass accuracy of current mass spectrometers.

In this paper, we detail how these interspersed segments of variable and conserved amino acid sequences create unusual features in the corresponding antibody peptide mass spectra. We demonstrate the importance of using high mass accuracy liquid chromatography mass spectrometry (LC-MS/MS) and describe how antibody proteomics

requires a particularly high stringency in the interpretation of the peptide mass spectra for reasons that are intrinsic to antibody gene structure. Finally, we offer specific guidelines for the interpretation of antibody peptide mass spectra focusing on correctly distinguishing CDR-H3 peptides with shared subsequences.

EXPERIMENTAL METHODS

Materials and Reagents

Concholepas concholepas hemocyanin (CCH), Protein A agarose, Protein G Plus agarose, N-hydroxysuccinimide (NHS)-activated agarose, immobilized pepsin resin, and Zeba spin columns were acquired from Pierce (Thermo Fisher Scientific, Rockford, IL). Incomplete Freund's Adjuvant (IFA), TRIS hydrochloride (Tris-HCl), ammonium bicarbonate (NH_4HCO_3), 2,2,2-trifluoroethanol (TFE), dithiothrietol (DTT), triethylphosphine (TEP), iodoacetamide (IAM), and iodoethanol (IE) were obtained from Sigma-Aldrich (St. Louis, MO). Urea and AG-5Ol-X8 resin were purchased from Bio-Rad (Hercules, CA). Microcon 10 kDa MWCO (Microcon-10) centrifugal filter columns from Millipore (Bedford, MA) and Hypersep SpinTip C18 columns (C18-SpinTips) from Thermo Scientific (Rockford, IL) were used in LC-MS/MS sample preparation along with LC-MS Grade water, acetonitrile (ACN), and formic acid from EMD (Billerica, MA).

Rabbit immunization, V gene sequencing, and preparation of serum antibodies

Methods for immunization, V gene sequencing, and preparation of antibodies for this study were previously described in *Wine, et al*.[7] Briefly, a New Zealand white rabbit was immunized with 100 µg CCH protein. Booster immunization with antigen in IFA was administered at days 14 and 28. The animal was sacrificed at day 35. Total RNA was isolated from femoral bone marrow cells (BM), peripheral B cells (PBCs), and CD138+ bone marrow plasma cells (BM-PCs) and cDNA libraries were generated from poly(A)+

RNA. V gene cDNA was amplified by 5'RACE with primers complementary to rabbit IgG CH1 and sequenced using the Roche 454 GS FLX Titanium platform (Roche Diagnostics GmbH, Mannheim, Germany). Sequencing data was processed using sequence quality and signal filters in the 454 Roche analysis pipeline, followed by identification of conserved framework regions and V germline gene identification using the IMGT/HighV-Quest Tool. Additional filters were applied to remove truncations (sequence length <70 amino acids, misalignment of framework regions FR1 and FR4) and sequences containing stop codons or ambiguous reads. In total, $>1.5x10^5$ reads were obtained, resulting in 107,672 unique full-length, in-frame V_H genes. For reference sequence database construction, single read sequences were excluded to reduce the impact of sequencing errors (18,593 V_H genes $\geq$2 reads).

Serum IgG was purified by protein A agarose affinity chromatography, and $F(ab')_2$ fragments generated by digestion with immobilized pepsin. Antigen-specific IgG-derived $F(ab')_2$ was isolated by affinity chromatography against CCH protein coupled to NHS-activated agarose and eluted in 100mM glycine pH 2.7. Immediately following elution, the pH was neutralized with 1M Tris-HCl, pH 8.5. Protein concentrations were measured using an ND-1000 spectrophotometer (Nanodrop, DE, USA).

Alternative cysteine alkylation and trypsin digestion

Protein samples were concentrated on Microcon-10 columns and split into aliquots for alternative cysteine modification. For IAM alkylation, aliquots were resuspended in 50% (v/v) TFE, 50 mM NH_4HCO_3 and 2.5 mM DTT, and incubated at 37°C for 60 min. Reduced samples were then alkylated with 32 mM IAM at room temperature, in the dark, for 60 min. Alkylation was quenched by addition of 7.7 mM DTT. Samples were diluted

to 5% TFE and digested with trypsin at a ratio of 1:75 trypsin:protein at 37 °C for 5 hours. Digestion was halted by addition of formic acid to 1% (v/v) concentration.

For IE alkylation, trypsin digestion in the presence of urea was carried out as previously described[79] with the following modifications: Samples were resuspended in 8 M urea, then diluted to a final reaction solution consisting of 2.4 M urea, 200 mM NH_4HCO_3 pH 11.0, 49% (v/v) ACN, 8.5 mM TEP, and 65 mM IE. pH was adjusted to 10 and samples incubated at 37 °C for 60 min. Samples were concentrated by SpeedVac (Eppendorf, NY, USA) and resuspended in 100 mM Tris-HCl, pH 8.5 to reach a final urea concentration of 1.6 M prior to trypsin digestion. Trypsin was added at a ratio of 1:75 trypsin:protein at 37 °C for 5 hours. The digestion was quenched with 1% formic acid.

Human raw spectral data and V_H sequence database

All human data used in this study corresponds to the donor HD1 dataset previously described in Lavinder, *et al.*[1] In summary, a healthy human subject (HD1) was administered the tetanus toxoid/diphtheria toxoid vaccine (Sanofi Pasteur MSD GmbH, Leimen, Germany) for booster immunization 7 years after previous booster. V_H and V_L gene sequences from plasmablasts and memory B cells isolated at 7 days and 3 months post-boost were determined by Roche 454 sequencing. Sequence data was processed and filtered as described for rabbit sequencing. In total, 70,326 V_H gene sequences were used in construction of the human HD1 reference sequence database.

IgG was purified by affinity chromatography with Protein G Plus agarose from serum samples collected at pre-vaccination (day 0), 7 days, 3 months, and 9 months post-vaccination, and digested with immobilized pepsin resin to generate $F(ab')_2$ fragments. Antigen-specific $F(ab')_2$ was isolated by affinity chromatography against vaccine-grade tetanus toxoid protein (Statens Serum Institut, Copenhagen, Denmark) coupled to NHS-

activated agarose and eluted with 20 mM HCl (pH 1.7). Eluted samples were neutralized with 1 M NaOH, 10 mM Tris-HCl and desalted on a 2 ml Zeba spin column prior to denaturation with 50% TFE, reduction with 10 mM DTT, and alkylation with 32 mM IAM. Samples were diluted 10-fold with 50 mM NH_4HCO_3 and digested with trypsin (1:35 trypsin:protein) overnight at 37°C. Digestion was quenched with 1% formic acid.

Sample preparation for LC-MS/MS

Digested IAM (human, rabbit) and IE (rabbit) samples were concentrated by SpeedVac, resuspended in Buffer C (5% ACN, 0.1% formic acid), and loaded and washed on C18-SpinTips according to the manufacturer's protocol. Bound peptides were eluted in 60% ACN, 0.1% formic acid, concentrated by SpeedVac, resuspended in Buffer C and filtered through Microcon-10 columns prior to LC-MS/MS analysis.

Construction of target and decoy databases

Sample-specific target protein sequence databases were constructed for SEQUEST searches of rabbit and human mass spectral data. The CCH rabbit database consisted of V_H and V_L gene sequences (≥2 reads), Ensembl rabbit protein-coding sequences (OryCun2.0), and common contaminants (from MaxQuant website, http://maxquant.org/contaminants.zip). The human HD1 database included V_H and V_L gene sequences, Ensembl human protein-coding sequences (release 64, longest sequence variant/gene), and MaxQuant common contaminants.

Decoy databases were constructed for rabbit and human analyses to evaluate the effects of decoy variants on error modeling of V-gene peptides. Reversed and shuffled databases were generated for each database at the protein level. Additionally, conserved-J region shuffled decoys were generated by preserving the conserved J-segment sequence (which directly follows the CDR-H3) of V_H gene sequences. For the remaining V gene

sequence, amino acids between arginine and lysine residues were shuffled, with Arg/Lys residues fixed to preserve peptide length and precursor mass distributions.

Computational interpretation of peptide mass spectra

Spectra were searched against the various protein sequence and decoy databases described above using SEQUEST (Proteome Discoverer 1.3, Thermo Scientific). Fully-tryptic peptides with up to 2 missed cleavages were considered. Mass tolerance filters of 5 ppm (MS1) and 0.5 Da (MS2) were applied. Static cysteine modifications of either carbamidomethylation (IAM-alkylation, +57.0215 Da) or ethanolyl (IE-alkylation, +44.0262 Da) were included based on which modifying reagent was used. Oxidation of methionine (+15.9949 Da) was allowed as a dynamic modification. PSMs were filtered using Percolator (implemented in Proteome Discoverer) to control false discovery rates (FDR) to <1% as determined using a reverse-sequence decoy database.[80] All observed precursor masses were recalibrated according to the methods of Cox, *et al*.,[81] and the average mass deviation (AMD) was calculated for all high-confidence PSMs (Percolator FDR <1%) matching the same reference peptide, as the mean difference between the observed precursor masses and the expected mass of that reference peptide in units of ppm. Due to the high frequency of isobaric peptides with isoleucine-leucine substitutions in V-gene sequences, we considered all Iso/Leu sequence variants as a single group, and mapped the group to all CDR-H3 peptides associated with any of the group members. For other isobaric pairings (e.g. Asp/Gly-Gly, Gln/Gly-Ala) and ambiguous identifications where MS/MS spectral differences can distinguish between pairings, we considered only the top-ranked PSM determined by the SEQUEST-Percolator pipeline.

Survey of covalent peptide modifications

In order to confirm the specificity of cysteine modifications and to assess the general overall presence of covalent post-translational modifications (PTMs) among antibody peptides, raw peptide mass spectra from the rabbit samples were computationally searched for the dominant, differentially observed PTMs as follows: Tandem mass spectral sets were first reduced in size and complexity through spectral clustering, in which merged spectra were represented by a single consensus spectrum. For each sample, spectra were initially grouped based on precursor mass so that all the members within a group were within 25 ppm of at least 1 other member. Hierarchical clustering was performed on the tandem mass spectra of each weight group using a fuzzy cosine similarity metric and weighted linkage criteria with a distance cutoff of 0.25. The fuzzy cosine similarity, or correlation, between two spectra A and B is defined as

$$similarity = Cos(A,B) = \frac{A_c \cdot B}{\|A\|\|B\|}$$

where A_c is the convolution of spectrum A with a Gaussian 1 Da in width. This serves to influence the correlation by both the intensity of each peak pair and the closeness of the peaks in m/z. Spectra composing each cluster were then reduced into a single consensus spectrum. An average parent ion mass was then assigned to each cluster.

All pairs of spectral clusters between IAM- and IE-labeled samples were compiled with the constraint that the parent ion mass difference between pair members fell within ±60.5 Da. Similarity measures were calculated for each pair, the sum of which was a composite metric for judging spectral correlation. Pairs were then binned in 2D arrays by mass offset and composite correlation score. Because clusters had varying numbers of members, all cluster pairs were not equal and were therefore weighted by 0.5 plus the log of the product of the two membership counts. The sum of these weights gave a single

summary statistic for each bin, and the data was visualized as a stacked bar graph consisting of 121 offset bins of width 0.02 Da that are centered at an integer value.

Differential analysis of cysteine modifications

PTM analysis (described above) was used to identify pairs of spectral clusters exhibiting an observed parent mass difference of 12.995 +/-0.005 Da (or 25.99 +/-0.005 Da for two Cys) between IAM- and IE-treated samples. Paired clusters with similar elution times and fragmentation patterns were flagged as originating from cysteine-containing peptides. The top-ranked SEQUEST peptide identification for each cluster was then considered. If the same sequence was identified in both treatments (inherently requiring the presence of cysteine to match), the peptide sequence was flagged as a likely correct, or "true positive", identification. If the peptide identification differed between treatment sets (precluding the presence of cysteine in the sequence), the corresponding peptide sequences were flagged as definitely incorrect, or "false positive", identifications.

RESULTS AND DISCUSSION

The goal of serum antibody proteomics is to systematically identify the distinct antibodies present in a serum sample, as assayed through the use of shotgun proteomics mass spectrometry. To achieve this, our approach relies on the integration of two main experimental pipelines:

1. High-throughput sequencing of B lymphocyte cDNAs to generate a database of class-switched antibody variable domain sequences in a particular individual.
2. A protein biochemistry and mass spectrometry-based proteomics pipeline for the identification of peptides derived from antigen-specific antibodies.

A personalized reference sequence database generated by the high-throughput sequencing pipeline is used in the interpretation of antibody-derived peptide mass spectra

obtained through the proteomics pipeline. Identified peptides can be mapped back onto the antibody sequence database to determine the distribution of specific clonotypes comprising the antigen-specific repertoire. However, the frequency of degenerate peptides mapping to multiple clonotypes complicates this analysis. Given that the CDR-H3 is the most hypervariable region in immunoglobulins and is overwhelmingly responsible for antigen specificity, as well as being the primary determinant of clonality, this problem can be largely simplified to that of the quantitation and sequence determination of CDR-H3 peptides. The remaining sequence of each antibody can then be retrieved from the V gene reference database.

For this study, we largely focused on analysis of serum samples from a New Zealand white rabbit (*Oryctolagus cuniculus*) immunized with *Concholepas concholepas* hemocyanin (CCH). Sequencing data for this rabbit was previously described,[7] and is summarized in *Materials and Methods.* We focus here only on the V_H sequences; while the partner V_L chain contributes to antibody stability and binding characteristics, native VH-VL pairing information cannot be determined by proteomic analysis, but can be derived by other methods once V_H chains are known.[6,7]

From this rabbit we prepared antigen-specific F(ab')$_2$ fragments, proteolytically digested them with trypsin and analyzed the resulting peptides by quantitative shotgun proteomics, employing nanoflow LC-MS/MS (see *Materials and Methods*). A conventional analysis of the peptide mass spectra would involve comparing the spectra against the rabbit's V_H gene database in order to identify those antibodies actually present in the serum. However, as we next discuss, the conventional proteomics database search process is insufficient for the analysis of antibody peptide mass spectra due to intrinsic properties of the antibody sequences.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium[82] via the PRIDE partner repository, with dataset identifiers PXD000916 (Rabbit) and PXD000917 (Human).

Limitations of standard peptide-spectrum assignments and decoy-based error modeling

While the general process of identifying the best peptide-spectrum match (PSM) is well established for conventional datasets searched against normal proteomic sequence databases,[83,84] V-gene databases contain unique sequence characteristics which pose challenges to this standard method of data interpretation.

Under the standard target-decoy approach, candidate peptides within a specified mass range of the parent ion are initially scored based on cross-correlation to the observed fragmentation spectrum (XCorr), subjected to additional quality filters, and ultimately assigned confidence scores by reference to the score distributions of decoy sequences. For a conventional proteome, the occurrence of multiple peptides sharing partial sequence identity and mass is extremely rare, as can be seen for proteins sampled from the human proteome (Figure 2.2a). Thus, while multiple theoretical peptides may fall close in mass to a given precursor ion, the correct peptide sequence will almost always match the MS2 spectrum with a significantly higher score than competing, incorrect peptides. This is reflected by the positive correlation between XCorr and the normalized difference in XCorr between the top two PSMs of a given spectrum (ΔCN) (Figure 2.3a).

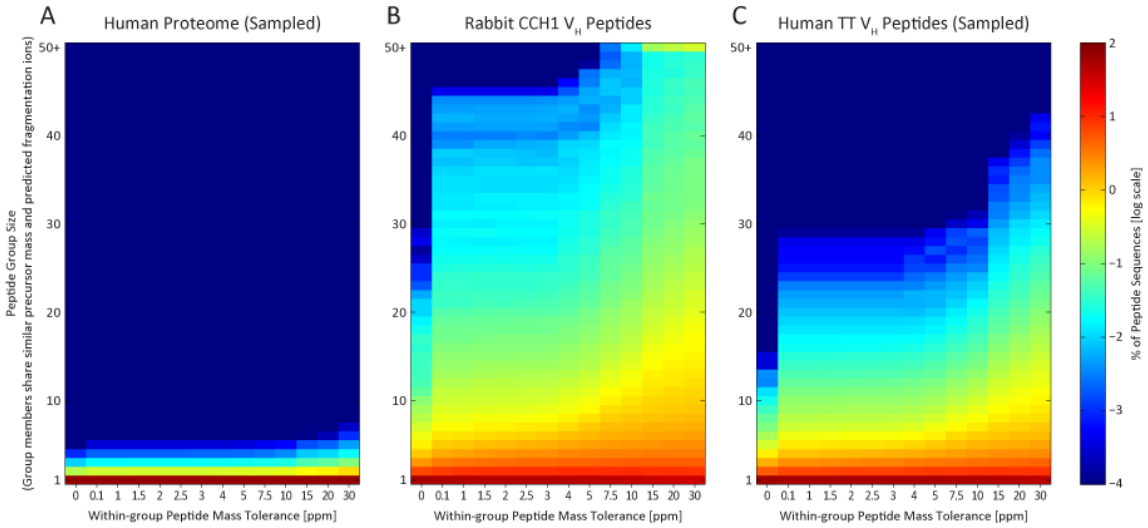

Figure 2.2: Theoretical extent of peptide-spectral match ambiguity for human proteome and antibody peptides.

In contrast to the proteome in general, antibody peptide sequences resemble each other in both mass and expected fragmentation patterns. The peptide sequence search space is thus strongly dependent on mass accuracy, as seen by plotting the extent of theoretical peptide-spectral match ambiguity, for (A) human proteome peptide sequences, (B) rabbit CCH antibody V_H peptides , and (C) human tetanus toxoid antibody V_H peptides. Reducing precursor mass tolerance thus more strongly affects the potential for false identifications in V_H peptides than for a typical proteome. Here, an *in silico* digest of the rabbit CCH V_H antibody sequences generated 505,790 unique peptide sequences (constrained to fully tryptic peptides of ≥ 8 amino acids, ≤ 6000 Da theoretical mass, and ≤ 2 missed cleavages). Each peptide sequence contributes to a y-axis bin defined by the self-inclusive count of all theoretical peptides within a specified mass tolerance (x-axis) and sharing at least 60% predicted fragmentation ion similarity. For comparison, the human proteome (A) and human TT V_H (C) sequence databases were processed likewise and subsampled to include the same number of peptide sequences as (B). The inter-sequence similarity evident in the antibody sets is negligible in this size-matched human proteome control.

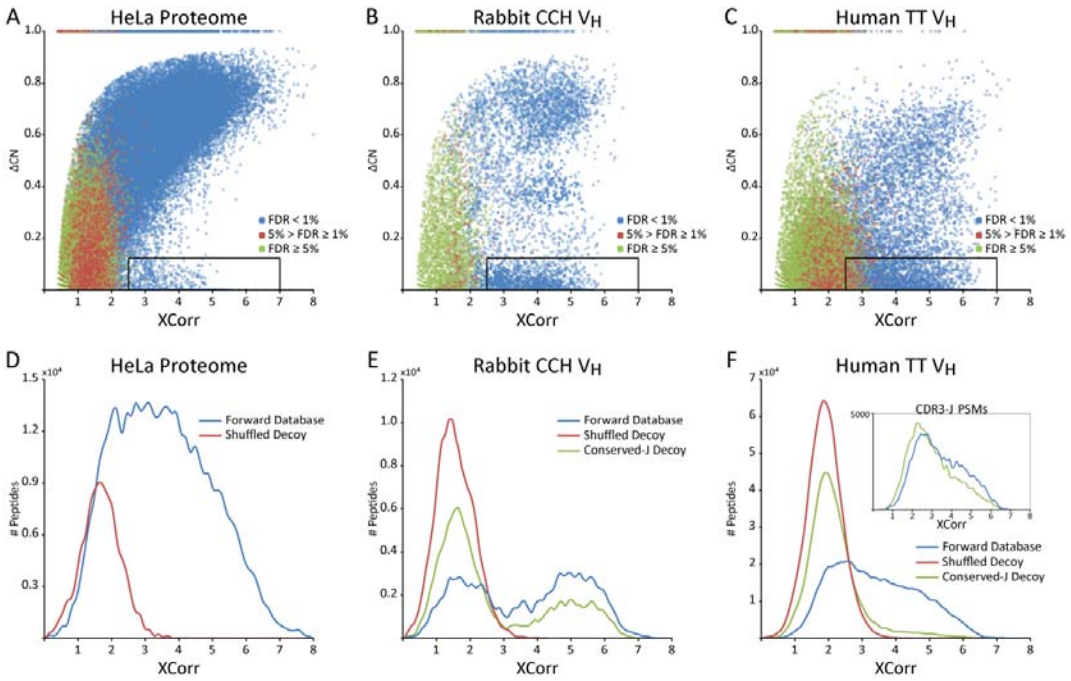

Figure 2.3: Antibody V_H spectra often show multiple high-scoring PSMs, problematic for standard target-decoy FDR methods.

Confidently-identified spectra from most proteomics samples generally score well against only one database sequence. The interspersal of conserved (framework) and variable regions in antibody F(ab')2 sequences often leads to multiple high-scoring PSMs for a single IgG-V_H peptide spectrum. Plotting the primary PSM score (XCorr) vs. the normalized difference in XCorr scores between the two top-scoring matches (ΔCN) from proteomic analysis of (A) human HeLa cell lysate compared to (B) rabbit and (C) human IgG-V_H peptide spectra reveals a substantial proportion of high XCorr/low ΔCN PSMs (denoted by black boxes) in the IgG-V_H datasets. Standard false discovery rate (FDR) calculations fail for these PSMs, as illustrated by high (blue), medium (green), and low (red) Percolator confidence scores: many high XCorr/low ΔCN PSMs are erroneously assigned high confidence in spite of high-scoring second hits implicit in the low ΔCN values. Filtering out low ΔCN PSMs inadvertently removes many true hits. Comparison of PSM XCorr distributions between target (blue) and decoy (red) databases reveals that standard decoys do not adequately model the non-random structure of IgG-V_H peptides [(D) human proteome, (E) rabbit IgG-V_H, (F) human IgG-V_H]. This is attributable to high-scoring, incorrect matches to IgG framework region-derived sequences. By constructing an alternate decoy database that preserved J-region framework regions ("Conserved-J Decoy"), ambiguity of CDR-H3,J peptide assignment can be modeled (green). These peptides account for the majority of high-XCorr PSMs in rabbit (E), while additional framework-derived peptides add to the complexity of the human IG-V_H sample (F, inset).

For the case of immunoglobulin variable genes, however, large numbers of peptide sequences overlap in both mass and partial sequence identity (as plotted for our VH datasets in Figure 2.2b,c), yielding sets of highly-similar theoretical MS2 spectra. This confounds proteomics analysis and often results in, for a single spectrum, multiple ambiguous peptide-spectral matches sharing similarly high PSM correlation scores (observed as high-XCorr/low-ΔCN, i.e. high scoring-second rank hits) (Figure 2.3b,c). In some cases, incorrect sequences out-score the correct PSMs. Even when applying an extremely strict mass accuracy filter— requiring a peptide mass to fall within 5 ppm of the observed precursor ion mass to be considered—false identifications are still prevalent.

V-gene sequence similarity also effects decoy-based error-modeling. Standard errors in PSM assignment normally arise from poor quality spectra, which contain significant noise and/or additional peaks due to unaccounted for contaminating peptide fragments following ion isolation. In order to assign PSM confidence and calculate a false identification rate, a decoy reference database of either reversed or shuffled protein sequences is generally used to model this standard error, allowing for confidence-filtering based on discernible differences in the distribution of true and false positive PSMs (Figure 2.3d).[84,85] Software programs such as Percolator[80] analyze multiple parameters of target and decoy results (including XCorr, ΔCN, and others) in order to determine a set of high-confidence PSMs at a given FDR (Figures 2.3a-c). For the case of Ig V genes, reversing or shuffling sequences did not replicate the high incidence of high scoring-second rank hits observed in the forward search, demonstrating that a standard decoy database fails to model this aspect of IgG sequences (Figures 2.3e,f).

Immunoglobulin PSM ambiguity arises from Ig peptides containing highly immutable framework regions

To further investigate this trend, we focused on the partial sequence identity of CDR-H3-containing peptides. Most such peptides also contained the entirety of the J-region subsequence in both the rabbit and human samples, generally a series of 12 or more residues sharing exceptional self-similarity within each species. Hence, peptides containing the J-region shared a significant fraction of identical peaks within their fragmentation spectra, in addition to peaks contributed by the variable CDR-H3 sequence (Figure 2.4).

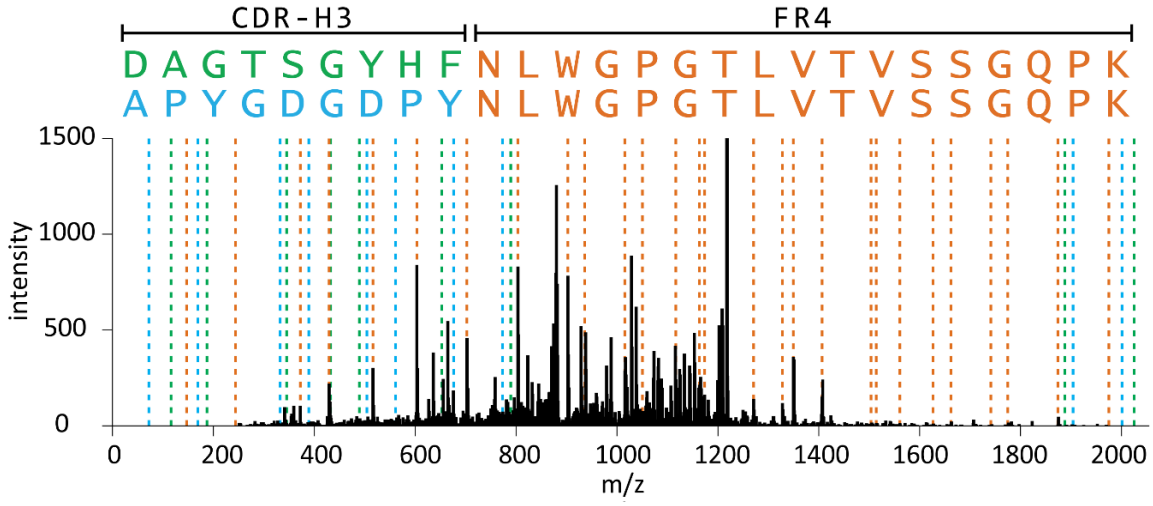

Figure 2.4: CDR-H3 sequence differences between high-confidence PSMs are not reflected in the MS/MS spectrum.

High-scoring PSMs for antibody CDR-H3 peptide mass spectra are dominated by matches to peptides sharing identical C-terminal J region FR4 framework sequences. This is illustrated by two top-scoring peptide sequences mapped to a single observed rabbit spectrum, with shared (orange) and unique *in silico* predicted MS2 fragmentation peaks associated with *APYGDGDPY*NLWGPGTLVTVSSGQPK (blue) and *DAGTSGYHF*NLWGPGTLVTVSSGQPK (green). Both sequences exhibit PSMs with XCorr > 4.7 with a normalized difference in XCorr scores (ΔCN) of 0.006. A similar trend accounts for a large proportion of the high-scoring matches in Figs. 2.3b and d.

In order to assess the magnitude of this effect on the resulting PSM scores, we generated sample-specific shuffled decoy databases in which the J-region residues were explicitly preserved ("Conserved-J Decoy"). Importantly, the Conserved-J Decoy database reproduced the incidence of high scoring-second rank hits observed in the J-region peptides and evident in the V_H forward peptide database (Figure 2.3e,f[inset]). A significant portion of high scoring-second rank hits can therefore be attributed to CDR-H3-containing peptides partially matching other CDR-H3-containing peptides by their conserved J region sequences. More generally, Ig peptides containing an antibody framework region at one terminus are subject to this kind of ambiguous PSM assignment. Consequently, standard decoy-based error modeling significantly underestimates false identifications for this class of peptides.

Construction of a high-confidence set of rabbit V_H identifications

In order to determine the prevalence of incorrect identifications and find characteristics on which to discriminate between true and false matches, we employed differential labeling of cysteine residues to create a set of higher confidence identifications consistent with the cysteine labeling data and to flag a subset of definitively incorrect identifications as high-scoring false positives (Figure 2.5a). Rabbit $F(ab')_2$ fragments were divided into two aliquots. One aliquot was alkylated with iodoacetamide (IAM), while the second was alkylated with iodoethanol (IE). This created equivalent samples with the exception of a 13 Da mass difference between modified cysteine residues in the two samples.

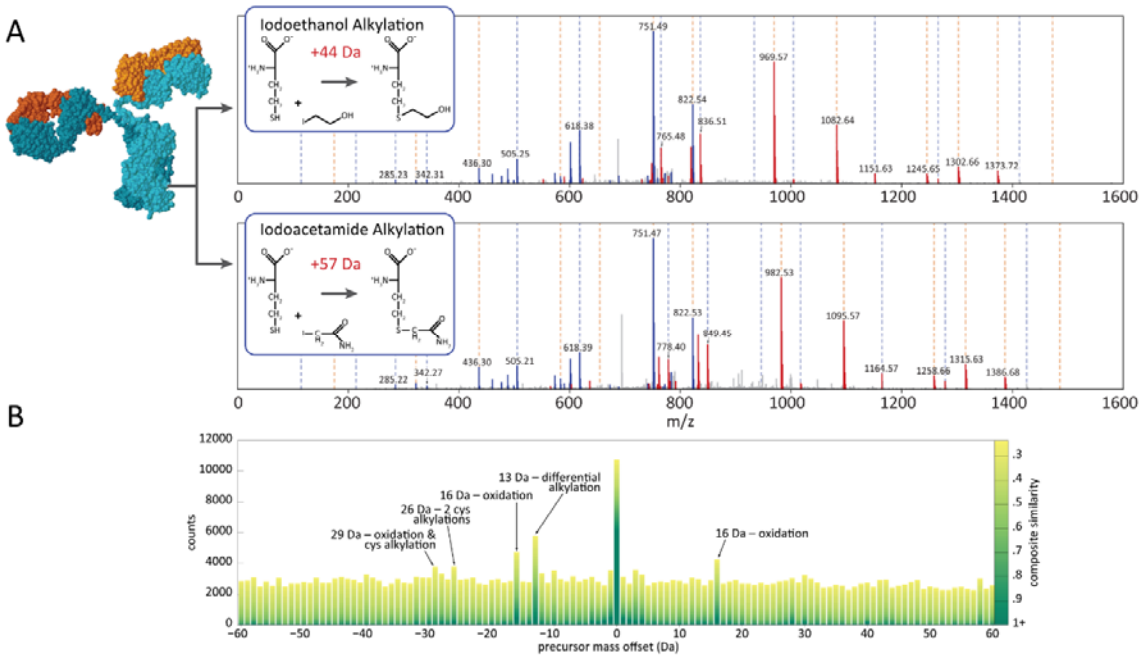

Figure 2.5: Differential cysteine modification and naïve identification of common PTMs.

A limited set of higher-confidence identifications can be created using differential covalent modification to flag cysteine-containing peptides. (A) Comparison of rabbit CCH spectra from samples treated with iodoacetamide (Cys +57Da) vs. iodoethanol (Cys +44Da) results in a 13Da mass difference per cysteine. PSMs for paired spectra exhibiting a mass shift but no cysteine residues in the corresponding matched sequences can be flagged as false identifications. (B) Comparison of precursor mass offsets between differentially labeled rabbit CCH samples confirms alkylation and oxidation account for the most abundant modifications.

Following LC-MS/MS analysis, datasets corresponding to IAM- and IE-treated samples were compared to identify parent ion pairs across the two datasets exhibiting the signature 13 Da mass difference, similar chromatographic elution times, and correlated MS2 fragmentation spectra. Qualifying ion pairs were considered cysteine-containing; upon peptide-spectrum sequence assignment, ion pairs with identical sequences containing cysteine residues and displaying the 13 Da difference in the two aliquots were flagged as more likely to be correct, and considered for these purposes to be "true positive" identifications. In contrast, those spectra shifted by 13 Da but lacking a cysteine residue in

their assigned sequences were considered definitely incorrect, or “false positive”. By flagging peptides in this manner, we defined a set of 53 “true positive” and 40 “false positive” peptide identifications comprising 11,077 and 425 PSMs, respectively. This set was used both to diagnose PSM assignment error and to define filtering criteria appropriate for more general application across all PSMs, not just those containing cysteine residues.

To further assess these samples, we examined the frequency of all potential precursor ion mass offsets between the differentially treated samples so as to survey the most common covalent modifications, thus confirming the cysteine modifications and testing for other potential modifications (Figure 2.5b). Besides modified cysteine, only one other prevalent modification was found, occurring in both samples at a mass offset of 15.99 Da and consistent with oxidation. Detailed manual analysis of fragmentation spectra confirmed oxidized methionine as the main contributor to this offset peak.

A stringent average mass accuracy filter successfully removes false identifications

Using the high confidence true and false identification sets, we searched for mass spectral properties that distinguished these cases. We observed a robust difference in mass accuracy distributions (defined as the difference between observed precursor mass and expected peptide mass, in units of parts per million (ppm)), with the “true positive” PSMs centered around 0.127 ppm with a standard deviation of 0.637 ppm (following mass recalibration), while “false positive” PSMs were more evenly distributed throughout the mass range. This signal, while clear, was not suitable for direct use as a mass accuracy filter at the level of PSMs, since many individual “true positive” PSMs still deviated from expected mass by several ppm. Application` of a strict mass accuracy filter to remove false PSMs would therefore inevitably remove many true PSMs as well (Figure 2.6a).

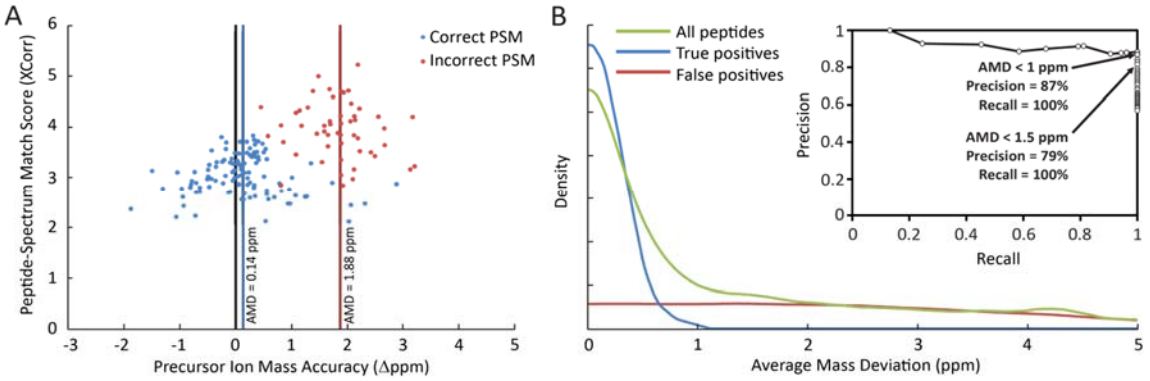

Figure 2.6 A large average mass deviation indicates peptide misidentification.

Correctly matched PSMs exhibit a systematically smaller average mass deviation (AMD) compared to incorrect identifications. (A) Plotting the difference in precursor ion mass from expected peptide mass (Precursor Mass Accuracy) vs. XCorr scores of individual rabbit CCH PSMs reveals overlapping mass accuracy distributions for PSMs matched to the same peptide sequence for correct (blue) and incorrect (red) identifications. While individual incorrect PSMs may achieve higher XCorr scores than correct matches, the average precursor mass accuracy across all PSMs for a given peptide (AMD) discriminates well between correct and incorrect identifications. (B) For the set of high-confidence rabbit CCH PSMs derived from cysteine-labeling, true identifications exhibit low AMD scores while false identifications are more uniformly distributed. Thus, filtering by AMD strongly controls misidentifications. Here, controlling AMD to within 1.5 ppm provides 100% recall of true identifications and increases precision from near 50% (background rate) to 79%. Requiring AMD < 1 ppm further increases precision to 87% with no loss of recall.

However, the average mass deviation (AMD) of a peptide identification, calculated as the average mass accuracy of all high-confidence PSMs associated with a given peptide, showed an extremely narrow distribution for the "true positive" set (mean 0.141 ppm, stdev 0.238 ppm). In contrast, the "false positive" set exhibited a roughly uniform AMD distribution across the mass range. Consequently, filtering hits by applying a strict AMD filter was feasible without substantial loss of true identifications. Requiring AMD < 1.5 ppm in this dataset improved the precision from a prior rate of approximately 50% to 79%, with no loss of true identifications. Applying an even stricter AMD threshold of 1 ppm further improved the precision to 87%, again with no loss of true identifications (Figure

2.6b). High mass accuracy LC-MS/MS is therefore sufficient to identify antibody CDR-H3 peptides from serum at relatively high precision when combined with a stringent AMD filter beyond the conventional proteomics analytical pipeline.

CONCLUSIONS

Proteomic analysis of serum immunoglobulins has only recently become feasible with the ability to generate appropriate mass spectrometry reference databases *via* next-generation sequencing of individual B cell antibody repertoires. Even with an appropriate custom database in hand, however, antibody sequences still present significant challenges for mass spectral interpretation due to the frequency of interspersed variable and conserved amino acid sequences within the same peptides. We've shown how these sequence properties lead to certain systematic trends in the fragmentation spectra of antibody-derived peptides, which introduce additional errors in peptide-spectrum correlation scoring not accounted for by standard decoy-based error modeling. The observation of similar sequence properties in rabbit and human datasets indicates that these are intrinsic features of immunoglobulin primary structure which should be accounted for in any proteomic analysis of antibody repertoire, regardless of species. To this end, we have demonstrated a strategy to reduce false discovery and improve the accuracy of antibody identification by shotgun proteomics through the use of high mass accuracy LC-MS/MS and high stringency filters applied to groups of peptide-spectral matches, rather than individual PSMs.

These findings highlight the importance of evaluating methods of data analysis when applied to non-standard datasets. While we specifically addressed complications encountered in the analysis of antibodies, we would expect similar trends for any protein samples where many close variant sequences might be present, such as in samples assaying human genetic variants or large protein families with related sequences.

Chapter 3: UVnovo: A *De Novo* Sequencing Algorithm Using Single Series of Fragment Ions via Chromophore Tagging and 351 nm Ultraviolet Photodissociation Mass Spectrometry

De novo peptide sequencing by mass spectrometry represents an important strategy for characterizing novel peptides and proteins, in which a peptide's amino acid sequence is inferred directly from the precursor peptide mass and tandem mass spectrum (MS/MS or MS^3) fragment ions, without comparison to a reference proteome.* This method is ideal for organisms or samples lacking a complete or well-annotated reference sequence set. One of the major barriers to *de novo* spectral interpretation arises from confusion of N- and C-terminal ion series due to the symmetry between *b* and *y* ion pairs created by collisional activation methods (or *c*, *z* ions for electron-based activation methods). This is known as the 'antisymmetric path problem' and leads to inverted amino acid subsequences within a *de novo* reconstruction. Here, we combine several key strategies for *de novo* peptide sequencing into a single high-throughput pipeline: high efficiency carbamylation blocks lysine side chains, and subsequent tryptic digestion and N-terminal peptide derivatization with the ultraviolet chromophore AMCA yields peptides susceptible to 351 nm ultraviolet photodissociation (UVPD). UVPD-MS/MS of the AMCA-modified peptides then predominantly produces *y* ions in the MS/MS spectra, specifically addressing the antisymmetric path problem. Finally, the program UVnovo applies a random forest algorithm to automatically learn from and then interpret UVPD mass spectra, passing results to a hidden Markov model for *de novo* sequence prediction and scoring. We show this combined strategy provides high performance *de novo* peptide sequencing, enabling the *de novo* sequencing of thousands of peptides from an *E. coli* lysate at high confidence.

* Chapter 3 has been previously published in: Robotham, S. A.; Horton, A. P.; Cannon, J. R.; Cotham, V. C.; Marcotte, E. M.; Brodbelt, J. S. *Anal. Chem.* **2016**, *88* (7), 3990–3997.
S.A.R. and A.P.H. contributed equally.

INTRODUCTION

The breadth of proteomic studies has never been greater, as a growing trend in proteomics research pushes mass spectrometry experiments beyond the study of model organisms, proteotypic peptides, and common post-translational modifications. This strains the limits of traditional spectral interpretation using sequence databases, and it has driven development of more flexible search methods and proteogenomic pipelines. De novo peptide and protein sequencing is one potential strategy for characterizing novel peptides.[86] Rather than comparing a peptide spectrum to theoretical candidate spectra from a reference protein sequence database, *de novo* analysis directly infers a peptide sequence from the precursor peptide mass and tandem mass spectrum (MS/MS or MS^3) fragment ions.[87] This method is ideal for organisms or samples lacking a complete or well-annotated reference sequence set. In the event that gene sequences are available, *de novo* approaches are well suited for interpreting unidentified spectra and discovering unknown splice variants, intergenic peptides, sequence polymorphisms, and other novel peptides.

Given an ideal MS/MS spectrum, *de novo* peptide sequence assignment is a trivial exercise. Such a spectrum would exhibit a complete series of ions, all of a single fragment type (N-terminal *a*/*b*/*c* or C-terminal *x*/*y*/*z* ions) and known charge state, that span an entire precursor peptide. The sequence could then be read directly from the spectrum by matching the mass difference between each consecutive ion pair to its corresponding amino acid. Technological developments, notably high-resolution MS/MS acquisition and concurrent collection of complementary fragmentation spectra (e.g. paired collision-induced dissociation (CID)/Electron transfer dissociation (ETD) mass spectra), have greatly improved the potential of *de novo* peptide sequencing, but spectra still suffer from incomplete peptide fragmentation, complex fragmentation patterns and neutral losses, and uninterpretable noise. CID,[88,89] HCD,[90] ETD,[91,92] and dual fragmentation (EThcD,

ETciD),[93] have all been applied for *de novo* sequencing. Infrared multiphoton dissociation (IRMPD) and ultraviolet photodissociation (UVPD)[94–97] are also emerging as viable alternatives for tandem mass spectrometry of peptides.

One of the major barriers to *de novo* spectral interpretation arises from confusion of N- and C-terminal ion series due to the symmetry between *b* and *y* ion pairs created by collisional activation methods (or *c*, *z* ions for electron-based activation methods). This is known as the 'antisymmetric path problem' and leads to inverted amino acid subsequences within a *de novo* reconstruction.[98] A related difficulty arises when fragment ions with similar *m/z* values cannot be independently resolved.[99] Biased peptide backbone fragmentation, the most serious problem, leads to spectral regions without fragment ion evidence and precludes definition of a complete amino acid sequence. These issues have made it unrealistic in practice to assign full and accurate peptide sequences in an automated *de novo* fashion. Therefore, database searches still greatly outperform *de novo* in any complex bottom-up shotgun proteomics experiment for which representative sequence data is available. Many modern *de novo* algorithms compensate by reporting tens or hundreds of putative sequences for a single peptide spectrum or only partial peptide sequences containing gaps where amino acids cannot be derived.[100,101] The results are most useful after manual curation or homology-based database comparisons, where such hybrid sequence tag-based homology searching combines the flexibility of *de novo* sequencing with the identification power provided through database comparison.

Among the many *de novo* programs available today, a few of the more popular established and emerging options include PEAKS, PepNovo, NovoHMM, pNovo, DirecTag and Novor for bottom-up proteomics and Twister for top-down analysis.[99,102–107] Most such tools use statistical models of peptide fragmentation for spectral interpretation prior to sequence generation, or for scoring candidate *de novo* sequences constructed from

simple initial assumptions and rules. These fragmentation models are rooted in the idea of the offset frequency function (OFF), introduced by Dančík et al. in 1999.[108] Fundamentally, the OFF treats fragmentation as a stochastic process whereby specific ions (ex. b^+, y^+-*NH3*) have a certain chance for being observed from each peptide residue position. These models are highly dependent on the type of spectra used during construction, limiting the application of existing software for new spectral paradigms.

In parallel to the continued development of *de novo* interpretation software, considerable effort has focused on creating "ideal" spectra for *de novo* sequencing through novel sample preparation and instrumentation methods.[87] Most of these methods have been implemented to overcome the antisymmetric path problem or more generally, the issue of discerning product ion type. Differential labeling between two samples, via isotopic or chemical modification of peptide N- or C-termini, is applied to evoke a mass difference between product ions pairs and allows MS^2 ion type annotation.[109,110] Alternatively, spectral simplification through chemical derivatization and charge sequestration can either enhance or eliminate a particular fragment ion series. In particular, peptide termini may be modified to increase the relative intensity of either the N- or C-terminal ion series.[111,112] Changing the basicity or charge of a peptide terminus influences the charge distribution along an ionized peptide and, consequently, produces a more prominent series of fragmentation ions from the end where charge is concentrated.

We recently demonstrated marked spectral simplification through a combination of chromophore derivatization and UVPD-MS.[113,114] The simplification mechanism, fundamentally different from those described above, destroys rather than neutralizes redundant fragment ions. By attaching the chromophore 7-amino-4-methylcoumarin 3-acetic acid (AMCA) to a peptide N-terminus, the peptide becomes susceptible to 351 nm photoactivation. The selectivity of 351 nm UVPD ensures that only AMCA-derivatized

peptides undergo photodissociation, and successive laser pulses effectively eliminate N-terminal chromophore-containing ions. C-terminal product ions remain unaffected by the UVPD, and the process yields a clean series of *y* ions uniformly distributed along the entire peptide length.

In this paper, we combine three key strategies for *de novo* peptide sequencing into a single high-throughput pipeline: (i) covalent modification of peptides and (ii) 351 nm UVPD fragmentation to favor N-terminal fragment ions with (iii) a dedicated software platform, UVnovo, to interpret these data. We introduce an improved strategy for selective peptide N-terminal AMCA derivatization. This is accomplished through highly efficient carbamylation of lysine side-chain amines[115] prior to tryptic digestion and AMCA labeling. LC-UVPD-MS/MS of the AMCA-modified peptides then predominantly produces *y* ions in the MS/MS spectra, specifically addressing the antisymmetric path problem. Finally, the program UVnovo applies a random forest (RF) algorithm[116] to automatically learn from and then interpret UVPD spectra, passing results to a hidden Markov model (HMM) for *de novo* sequence prediction and scoring. We show this combined strategy provides high performance *de novo* peptide sequencing.

MATERIALS AND METHODS

Materials

Trypsin Gold, Mass Spectrometry Grade was purchased from Promega (Madison, WI, USA). LC-MS grade acetonitrile and water were purchased from EMD Millipore (Darmstadt, Germany). Phosphate buffered saline (PBS) and dimethyl sulfoxide (DMSO) were purchased from Thermo Fisher Scientific Inc. (San Jose, CA, USA). Sulfosuccinimydyl-7-amino-4-methyl-coumarin-3-acetic acid (Sulfo-NHS-AMCA) was

purchased from Pierce Biotechnology (Rockford, IL, USA). *E. coli* lysate was graciously donated by Dr. M. Stephen Trent's research group at the University of Texas at Austin.

Modification of *E. coli* lysate

Figure 3.1 shows the process for N-terminal AMCA peptide derivatization. 50 µg of *E. coli* lysate in 100 µL of 50 mM sodium carbonate and 8 M urea was heated at 80 °C for 4 hours to carbamylate lysine side chains (ε-amines) and N-termini primary amines, blocking subsequent reaction with AMCA. Urea was removed through PBS buffer exchange, and proteins were then digested using trypsin at 37 °C overnight. After digestion, 25 µL of 20 mM sulfo-NHS-AMCA in DMSO was added to approximately 270 µL of the digest to label the peptides' primary N-terminal amines, and the solution was kept in the dark overnight at room temperature. Samples were cleaned using a C18 SPE cartridge to facilitate removal of unreacted AMCA, evaporated to dryness, and resuspended for LC-MS/MS (98% water/2% acetonitrile with 0.1% formic acid).

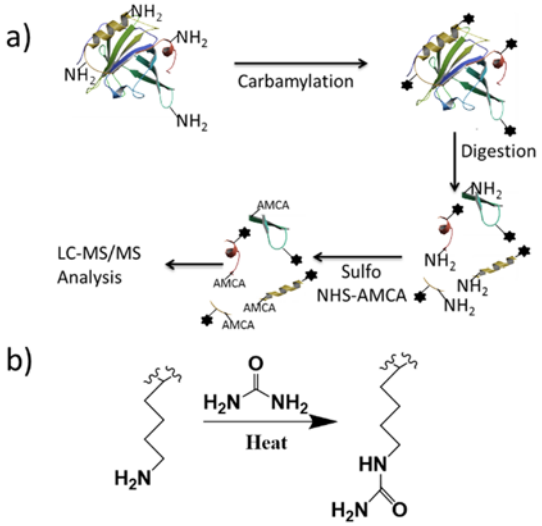

Figure 3.1: Workflow for peptide N-terminal AMCA derivatization.

(a) Carbamylation blocks primary amines (Lys and protein N-terminus) from subsequent AMCA derivatization and prevents lysine from tryptic cleavage. After digestion, AMCA labeling occurs at the new peptide N-terminal amines. (b) The carbamylation reaction.

LC-MS/MS analysis of *E. coli* lysate

All mass spectra were acquired using a Thermo Scientific Velos Pro dual linear ion trap mass spectrometer (Thermo Scientific; San Jose, CA) modified for UVPD by addition of a Coherent 351 nm excimer laser (Coherent; Santa Clara, CA, USA) to allow 351 nm UV excitation of ions present in the ion trap.[117] The laser was set to 3 mJ per pulse at 500 Hz, with 15 pulses per scan. Peptides were separated by reverse phase chromatography and eluted into the mass spectrometer using a Dionex NSLC 3000 nanoLC system (Thermo Scientific; Waltham, MA, USA). We used a 15 cm capillary column (75 μm ID) packed with 3.5 μm particles (C18 stationary phase) with a pore size of 140 Å, loading 5 μg of peptide mixture (via 1 μL injection). Sample elution followed a 360 minute gradient starting at 3% B and increasing to 50% B with a flow rate of 300 nL/min; solvent A was water with 0.1 % formic acid (v/v), and solvent B was acetonitrile with 0.1% formic acid (v/v).

SEQUEST

In order to obtain a list of high confidence peptide spectral matches, raw spectra were analyzed using the SEQUEST and Percolator nodes of Proteome Discoverer v. 1.4 (Thermo Fisher Scientific, San Jose, CA). AMCA was required as a fixed N-terminal modification, and optional oxidized methionine in any position was allowed. The precursor mass tolerance was set at ±1.6 Da due to the low resolution of ion trap spectra. Because trypsin does not cleave at carbamylated lysines, SEQUEST protease specificity was set to trypsin(R) and included the proline rule. We considered only *y* ion fragments for the UVPD data sets, searching spectra against the UniProt *E. coli* strain K12 reference proteome.

UVnovo *de novo* sequencing

We implemented UVnovo, a *de novo* sequencing program for analysis of UVPD spectra, in the MATLAB programming language. All top-ranked high confidence SEQUEST peptide spectrum matches (PSMs) from charge 2+ precursor ions were used to train and validate UVnovo using three-fold cross validation as follows:

Spectral partitioning and preprocessing

Spectra were randomly partitioned into three sets. All spectra from a given peptide, collapsing PTM variants, were allocated to the same set, preventing their use for both training and validation. During each of the three cross-validation rounds, a different partition was treated as an 'unknown' test set, and the 'known' spectra in the remaining two partitions were used for model training. We repeated this three times, withholding a different test partition each time, to evaluate the performance of UVnovo against the high confidence SEQUEST PSMs.

Thermo *.raw files were converted to the mzXML format using MSConvert with peak picking, and peaks with an intensity < 5 were removed. Through an unexplained artifact of UVPD spectral generation, all fragment ions in the MS^2 spectra from all precursors less than m/z 817.2 were systematically shifted up 0.16 m/z by the instrument, whereas peaks of the remaining spectra displayed no such systematic mass error. This was corrected in preprocessing by subtracting 0.16 m/z from all peaks of the affected spectra. Additionally, because our goal was to evaluate the potential of UVPD-MS/MS for automated *de novo* peptide sequencing, we chose to marginalize the effect of incorrect precursor mass on *de novo* sequence assignment. We set the precursor mass for each spectrum to the integer mass nearest its respective SEQUEST PSM. Thus, our results should be understood as being contingent on an accurate definition of precursor mass, to the nearest 1 Dalton, well within the capacity of modern high-resolution instruments.

UVnovo Overview

Figure 3.2 presents the overall software workflow. Following data import and preprocessing, spectral interpretation follows four main steps:

1. Transform each MS/MS spectrum into a spectral representation of peptide cleavage site probability at each possible mass position. This applies a random forest model for peptide fragmentation pattern deconvolution.
2. Refine the backbone cleavage site predictions using a hidden Markov model.
3. Identify amino acid sequences that best fit the predictions.
4. Score and rank the *de novo* sequence reconstructions.

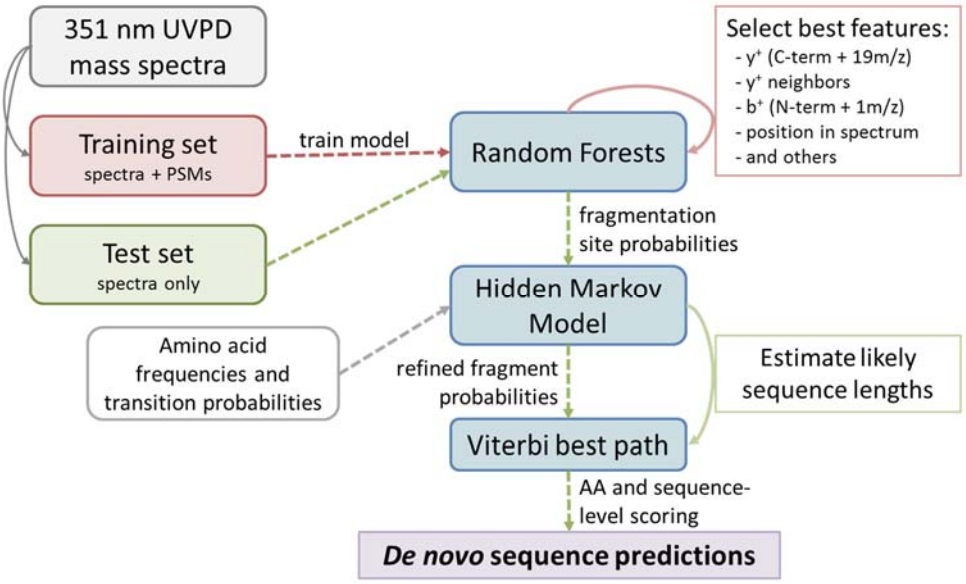

Figure 3.2: UVnovo workflow for *de novo* sequencing.

Spectra are divided into training and test sets. A random forest, trained on known spectra, transforms an unknown spectrum into a simplified representation of peptide cleavage site probabilities. At each position in this 'simplified spectrum', a hidden Markov model (HMM) refines the probability, also incorporating amino acid frequencies and requiring valid mass transitions. The best valid path through the HMM yields the *de novo* sequence prediction, and the individual fragmentation site probabilities provide a means to score each sequence.

Spectral interpretation using machine learning with random forests

Terms and notation

For the purpose of simplicity, all subsequent references to peptide and fragment mass will be understood as the nominal mass, or integer count of protons and neutrons, composing the species under consideration. This reframing simplifies computational treatment as well, and we divide real and observed spectral masses by a mass defect normalization factor, 1.000468, and round to the nearest integer (e.g. as for Kendrick mass calculations).[118,119] We retain the highest intensity peak in the case of coincident peaks during spectrum normalization.

We define bare peptide mass M as the total mass of amino acids composing a peptide, without the water (18 Da) that is also present in the precursor ion. Each interior position is referred to by its prefix (N-terminal) or suffix (C-terminal) mass, m_{pre} and m_{suff} respectively, and $m_{pre} + m_{suff} = M$. An interior position m_{pre} represents a fragmentation site when the sum of amino acid masses preceding that position equal m_{pre}. In other words, the mass of each N-terminal amino acid subseries is a *fragmentation site*. Therefore, $m_{pre} + 1$ denotes a *b* ion and $m_{suff} + 19$ a *y* ion.

Fragmentation pattern deconvolution

Similar to charge deconvolution, where a predictable series of distinct peaks can be mapped to single mass, fragmentation pattern deconvolution combines the evidence from multiple related peaks into support for a single base peak. For example, the presence of associated *b*, *y*, and *y*-H_2O ion peaks is strongly indicative of a specific fragmentation site along a peptide's backbone. The deconvolution process simplifies the original spectrum and assigns a probability to each mass m_{pre} that it corresponds to a true precursor peptide backbone fragmentation site (the nominal mass of the N-terminal amino acid series

preceding that fragmentation site). Such interpretation of fragmentation ion patterns has long been the domain of descriptive count-based statistical models,[108] but it can likewise be approached from a machine learning perspective.[106] We introduce here the application of random forests to transform MS/MS spectra into predictions of peptide fragmentation sites.

Random forests

Random forest (RF) classifiers are an ensemble machine learning method that has been successfully applied across a wide range of fields.[116] This includes in proteomics where Degroeve and Martens used RF in their tool MS^2PIP to predict MS/MS peak intensities given a peptide sequence.[120] An ensemble, in machine learning, aggregates the results from many individual predictive models into a single output prediction. In particular, a random forest ensemble grows a 'forest' of independent decision trees. Individual trees are simple classifiers often prone to over-fitting, a problem that must be carefully guarded against. When combined into a RF or related ensemble, however, they turn into very powerful models.

Decision trees, and therefore RF, can handle an arbitrary number of feature (or predictor) variables, either continuous or categorical. Each tree is fit on a random resampling with replacement (bootstrapping) from a training data set, using a random subset of the predictors at each node in the tree. Enforcing randomness in the construction of the individual trees ensures that the trees are uncorrelated, and this reduces the problem of over-fitting for the final random forest ensemble. When applied to an unknown observation, each tree in the RF returns a binary value, 0 or 1, of class membership. The mode of these is taken as the RF prediction and the mean can be treated as a probability of

class membership. However, these probability estimates tend toward the extreme, 0 and 1, and must be treated cautiously.[121,122]

Training random forests to decipher MS/MS fragmentation patterns

We used the MATLAB implementation of random forests to create an ensemble of 400 decision trees. A supervised-learning algorithm, RF requires both positive and negative training examples (observations), each comprising a potentially large set of feature variables (a predictor vector) particular to that observation. Examples are labeled with their respective class during RF construction. Here, the positive examples included, for every spectrum in the training data, a predictor vector for each fragmentation site. An equal number of negative examples were likewise created after applying random shifts (-50 to +50 Da) to the true fragmentation sites.

An observation, representing a specific spectrum and mass position m_{pre}, was described by its predictor vector. This feature set primarily included MS/MS peak scores for all peak bins (width 1 m/z) spanning the ±50 m/z windows around both m_{pre} and the reciprocal C-terminal position at $m_{suff} = M - m_{pre}$. Doubly charged ions were also included, adjusting the window locations and widths accordingly. Each of the resulting 404 features took a normalized peak score rather than a raw intensity, calculated as follows: Peak intensities were ranked from highest to lowest in a spectrum, and the rank was divided by the bare peptide mass M. Local fluctuation in peak intensity was reduced by subtracting the minimum rank within a sliding window of ±50 Da.

Additionally, each predictor vector contained three derived predictors: the fractional position within the peptide (m_{pre}/M), and the mass positions relative to the N- and C-termini discretized into 100 Da regions ($\lceil m_{pre}/100 \rceil$ and $\lceil m_{suff}/100 \rceil$). The final

two indicated which 100 Da regions the fragment site falls into from the N and C-terminal directions. This gave a total of 407 predictor variables at each fragmentation site.

Random forests can return importance measures of training features as a side effect of their construction,[116] and we used this to perform feature selection, successively identifying and retaining a subset of the most useful predictors through four consecutive RF training stages. This iterative RF creation and feature selection reduced the original set of 407 predictors down to the 30 most important for m_{pre} classification. These 30 features directed construction of the final RF model.

By including as features each integer mass offset within ±50 Da of m_{pre} and m_{suff}, this set of predictors indirectly included all of the common fragment ion types a, b, c, x, y, and z. We let the machine learning process determine which of these if any should be used for the final RF. Additionally, the training automatically identified and utilized interactions between predictors. This obviated the need for human oversight and explicit model definition seen in most other *de novo* sequencing packages.

RF interpretation of unknown spectra

RF, trained as above, could then be used to interpret unknown spectra. We created predictor vectors spanning all potential fragmentation sites m_{pre} along a spectrum, as described above, but only included the 30 features relevant to the RF. The RF converted each vector into a probability that its associated m_{pre} represented a true fragmentation site.

Therefore, from an unknown spectrum, the RF generates a new 'spectrum' of peak mass versus fragmentation site probability. This deconvolution process effectively reduces a raw spectrum – its various ion peaks, spectral artifacts, noise, and intensity inconsistencies – into a much cleaner representation of peptide composition.

HMM for refinement of fragmentation site predictions

We implemented a hidden Markov model (HMM) to refine the RF predictions and re-compute probabilities across all potential peptide fragmentation sites.[123] Each node, or state, in the HMM is a possible fragmentation site and is described by both its position along the spectrum (m_{pre}) and its position in a sequence (n, the number of residues preceding it). These are the hidden states $s_{m_{pre},n}$ of the HMM, and the set of unique paths through the states represent all possible amino acid sequences with mass equal to the precursor.

The transition between two states in the HMM is equivalent to an amino acid mass. Starting from $s_{0,0}$ ($m_{pre} = 0$ Da and $n = 0$), the HMM propagates belief across all possible paths using the forward-backward dynamic programming algorithm.[124] It only records a combined (posterior) probability for each state $s_{m_{pre},n}$ it reaches, regardless of how many paths (equal-length, isobaric amino acid sequences) coincide at the state. The likelihood of a state at $n > 0$ and m_{pre} depends on the RF estimate at m_{pre} and the likelihoods of neighboring states at $n - 1$ and $n + 1$. Additionally, it is influenced by state transition probabilities representing single amino acid frequencies and dipeptide probabilities in the forward and backward directions. The amino acid state transition probabilities vary positionally within the peptide, and we incorporated into the HMM the expected frequencies of observing an amino acid at either of the N or C-termini or as an interior residue.

The state transition probability matrices were generated based on amino acid frequencies in an *in silico* digested *E. coli* reference proteome. They could, in practice, be based on any sequence database similar to the organism or sample under study. The probability of observing methionine was divided between its oxidized and un-oxidized forms at a rate (25%) matching that observed in a proteomically identified *E. coli* peptide

data set. (Note that incorporation of an expectation-maximization algorithm for HMM parameter estimation would remove the dependence on preexisting sequence or proteomic references.) Frequency and transition probabilities were marginalized against nominal residue mass, meaning indistinguishable residue pairs I/L, and F/M-oxidation were combined. Being carbamylated, lysine was distinct from glutamine. We removed cysteine from consideration as the peptides were not alkylated, and disulfide bonding would make observation of cysteine-containing peptides exceedingly rare.

As the length of each true peptide sequence was not known, HMMs were constructed for the set of most likely candidate lengths of a particular spectrum. Generally, between one and seven models were created for each spectrum. The fragment site probabilities were different in each model because the set of possible sequences composing each were necessarily disjoint.

Sequence assignment

The forward-backward algorithm and HMM construction described above can identify the most likely fragment mass states at each position along the spectrum. These states only represent which masses likely correspond to a fragmentation site, and they may not all derive from the same peptide. There is often not a viable amino acid path through all of the most probable states. We use the Viterbi algorithm[123,125] on the HMM results to obtain the most likely amino acid sequence reconstruction. This moves through the fragmentation site HMM posterior probabilities and identifies the single most likely path through the set of nodes. Our algorithmic approach is currently limited in that we find, for a given spectrum, only the single best path for each putative amino acid sequence length. Extending this to the provably correct set of top k paths becomes a more challenging problem.

De novo sequence scoring and filtering

Sequence reconstructions were scored, ranked, and filtered to create a final set of *de novo* sequences. The probability that a *de novo* sequence is correct is equivalent to the joint probability that all predicted nodes are correct and all true nodes were predicted. As the nodes are not independent, the true probability cannot be easily derived. A simple upper bound is given by the Fréchet inequality, equivalent to the probability of the single lowest-scoring node in a sequence:[126]

$$P(sequence) \leq P(n_1 \& n_2 \& \ \ldots \ \& n_k) \leq min\big(P(n_1), P(n_2), \ldots, P(n_k)\big)$$

This was often biased in favor of the shorter of two similar sequences when ranking reconstructions from a single spectrum. We applied it instead to filter unlikely sequencing results, removing *all* results for a given spectrum if *none* of the reconstructions surpassed a threshold value. In the same manner, we removed spectra without at least one high scoring reconstruction, using as a metric the average sequence node score. As detailed in the main text, these two filters greatly enriched the fraction of correctly sequenced spectra, removing over two-thirds of those that were not sequenced successfully.

Under the HMM framework discussed above, a spectrum may have a pair of probable HMM nodes at the same mass and consecutive sequence positions, $S_{m_{pre},n-1}$ and $S_{m_{pre},n}$. When, for example, the node at n falls on the true peptide sequence path, any potentially correct node at n-1 will have a decreased likelihood due to the incorrect $S_{m_{pre},n-1}$. This effect explains to a large degree the bias toward shorter peptides noted above, as the nodes of shorter peptides from a spectrum will have better positional resolution. We implemented a node rescoring scheme to address this problem. After defining a putative sequence, we removed any incorrect node that shared a mass with an on-sequence node. Taking the average node score following node renormalization at each

sequence position yielded a new sequence score metric. Candidate sequences for each spectrum were ranked by this score metric. Empirically, this performed significantly better than the two probability-based metrics above, but it was not as effective for filtering.

Amino acid precision and recall

We also assigned scores to the individual amino acids of the sequence predictions and used these for the construction of the precision-recall curves below. A residue spans the gap between two nodes and does not inherently have a score. Following the inequality defined above, we set an amino acid confidence score as the smaller likelihood from the two nodes defining that amino acid. Amino acids are ranked and sorted by decreasing confidence, and we calculate precision and recall values cumulatively for each increasingly large set of the top n predictions:

$$precision = \frac{number\ of\ correct\ amino\ acids}{number\ of\ predicted\ amino\ acids} = \frac{\sum_{i=1}^{n} \mathrm{iscorrect}(i)}{n}$$

$$recall = \frac{number\ of\ correct\ amino\ acids}{total\ number\ of\ amino\ acids\ in\ test\ set} = \frac{\sum_{i=1}^{n} \mathrm{iscorrect}(i)}{T}$$

T is the total number of amino acids across all SEQUEST PSMs in the test set. We plot precision versus recall for all values for n, where $1 \leq n \leq N$ and N is the number of all amino acid predictions. A correct amino acid is one that matches both the mass position and identity of a residue in the known peptide sequence, allowing for I/L and F/M^{Oxy} ambiguity. The tradeoff between precision and recall indicates our residue confidence score is well-calibrated for ranking the pooled amino acids.

RESULTS AND DISCUSSION

We based our strategy to enhance *de novo* peptide spectrum interpretation on the ability of UVPD to efficiently generate C-terminal fragment ion (*y* ion) series while

eliminating N-terminal ions (*a, b* ions). This strategy required efficient attachment of a UV chromophore to the N-terminus of each peptide in order to target them by 351 nm UVPD, while avoiding labeling of lysine side-chains that would result in indiscriminant chromophore attachment. We describe a sample processing scheme that accomplishes these goals and enables UVPD-based *de novo* peptide sequencing. We also introduce the *de novo* sequencing program UVnovo, as to date there is no *de novo* sequencing program suitable for analysis of 351 nm UVPD mass spectra.

Lysine capping with carbamylation

In order to confine AMCA modification to the N-termini of peptides, the *epsilon* amino group on lysine side chains must first be blocked. We have previously employed lysine guanidination for this purpose, converting lysines into homoarginines via reaction with *O*-methylisourea in the presence of 7 N ammonium hydroxide.[93,114] Here, we improve on this strategy and instead convert lysine to homocitrulline via carbamylation. This provides a quick and efficient alternative to guanidination for blocking the reactive *epsilon*-amino group on lysine side chains. Heating samples at 80 °C for four hours in an 8 M urea solution resulted in complete reaction of reactive primary amines on model proteins, including the N-termini and lysine side-chains.[115] As a proof of concept we evaluated carbamylation efficiency on intact myoglobin molecules before and after the carbamylation reaction, using direct injections of the intact protein into a high resolution Thermo Orbitrap Elite. With nineteen lysine resides and a free N-terminus, myoglobin has 20 amine reactive sites (Figure 3.3a). We observed a mass shift of 860.09 Da between the modified and unmodified forms, very close to the 860.116 Da expected from complete carbamylation (20*43.0058 Da). We estimated to be nearly complete based on the ESI mass spectrum

shown in Figure 3.3b and c. A similar analysis of intact ubiquitin (data not shown) also revealed complete lysine carbamylation.

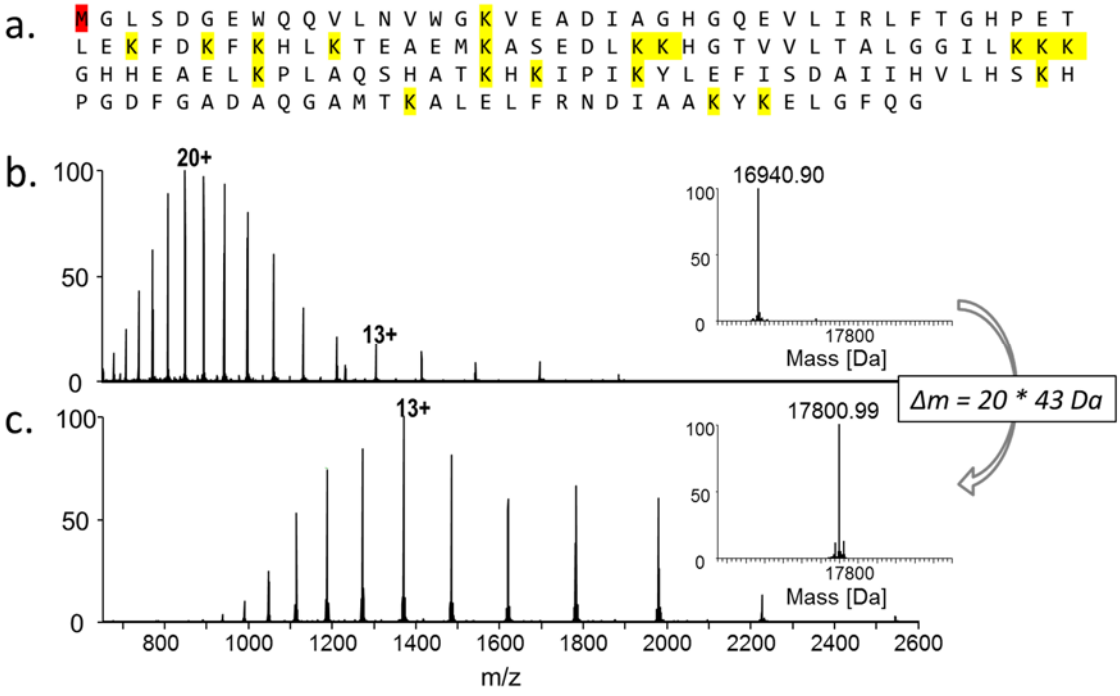

Figure 3.3: Demonstration of virtually complete Myoglobin lysine carbamylation.

Lysine carbamylation is observed after protein incubation in 8 M urea (4 h at 80 °C). (a) Myoglobin has 20 possible carbamylation sites: the protein N-terminus and 19 Lys residues. The mass difference of 860.09 Da between (b) unmodified and (c) carbamylated myoglobin indicates modification at all 20 sites. High accuracy ESI mass spectra were collected on an Orbitrap Elite.

351 nm UVPD spectra

Figure 3.4 presents a representative UVPD mass spectrum for a peptide from *E. coli* elongation factor G protein. The clean series of *y* ions is consistent with 351 nm UVPD and demonstrates the effective annihilation of *b* ions during the activation period (i.e. 15 laser pulses). The *b* ions (which contain the N-terminus) retain the AMCA chromophore and are susceptible to photoabsorption and dissociation during successive laser pulses.

Very few internal fragment ions are observed. While fragment ions are often diminished C-terminal to proline, peptide cleavage otherwise produces a comprehensive series of observable *y* ions.

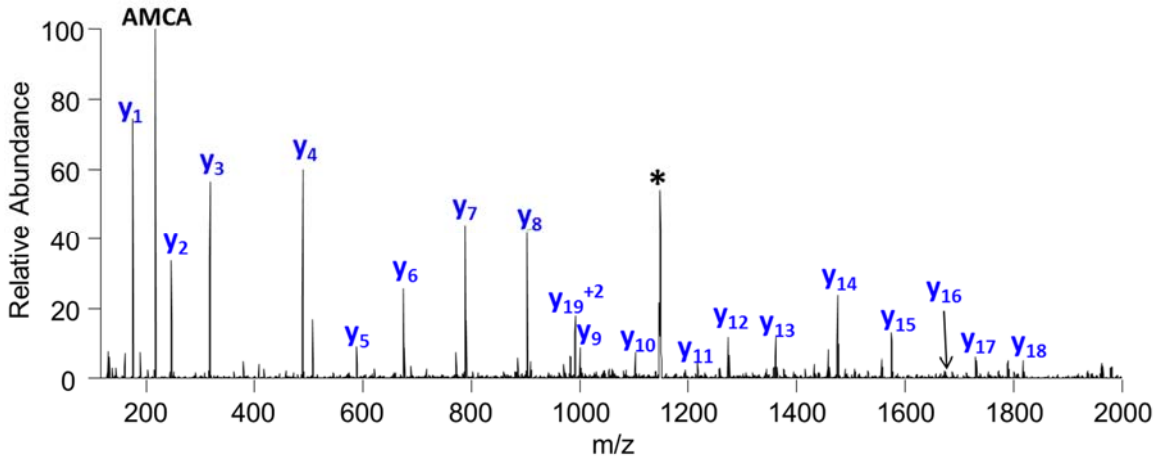

Figure 3.4: UVPD spectrum of peptide V[AMCA]YSGVVNSGDTVLNSVK[carbamyl]AAR.

UVPD (3 mJ per pulse, 15 pulses) mass spectrum of charge 2+ Elongation factor G peptide V[AMCA]YSGVVNSGDTVLNSVK[carbamyl]AAR from a trypsin-digested *E. coli* lysate. The precursor is labeled with an asterisk.

In one regard, spectral symmetry is beneficial for low resolution data because *b* and *y* ion pairs provide the most effective means for correct *de novo* precursor mass assignment.[104,108,119,127] The lack of complementary ion pairs and other telltale MS/MS signatures of precursor mass in our data precluded effective mass error correction. After a baseline correction of systematic error, only 63% of the *E. coli* lysate spectra we used for benchmarking (described below) had a mass within ±0.5 Da of the SEQUEST PSM. In all results below, the precursor mass was therefore assigned as the integer nearest the PSM mass.

However, the benefits of the UVPD method for *de novo* sequencing are twofold, and they cannot be overstated. First, with a complete *y* ion ladder, full-length, gapless *de*

novo reconstructions are frequently attainable for non-trivial peptides. Second, the spectra display an ion ladder from only the C-terminus, eliminating the computationally intractable antisymmetric path problem (where mirror-image sequences propagate along both N-terminal and C-terminal ion ladders). *De novo* algorithms commonly address this problem by making imprecise assumptions, such as requiring that *b* and *y* ions not share the same mass node. Such assumptions are unnecessary with 351 nm UVPD mass spectra.

UVnovo

We developed UVnovo to de novo interpret AMCA-treated UVPD spectra. As illustrated in Figure 3.2, the UVnovo spectral processing pipeline progresses through four main steps for each MS/MS spectrum**.** Briefly, the spectrum is simplified using a random forest (RF) classifier.[116] At each integer mass position along the spectrum, the RF merges evidence from 30 spectral features to predict whether that position falls at a peptide bond of the precursor peptide backbone. Next, a Hidden Markov Model is used to estimate the probability that each site corresponds to a true fragment ion.[124] Each spectrum is then assigned one or more potential sequences using the Viterbi algorithm, with a single best sequence generated for each likely spectrum peptide length.[125] The candidate *de novo* sequences are scored and ranked using the HMM fragment node probabilities.

Validation of UVnovo using *E. coli* lysate

In order to measure performance on a complex protein sample, we applied the AMCA-UVPD strategy on a full *E. coli* lysate. The lysate was carbamylated, digested, derivatized with AMCA, and analyzed via LC-MS/MS with 351 nm UVPD. Spectra from triplicate *E. coli* runs were processed with Proteome Discoverer SEQUEST using the Percolator node and allowing a ±1.6 Da precursor mass tolerance. Limiting the results to doubly charged precursors and top-ranked matches, 7911 high confidence identifications

matching 2983 unique peptide were obtained from the 106,870 spectra collected across all three samples. We benchmark UVnovo against these high confidence PSMs using three-fold cross validation (CV) to maintain independence between training and test sets.

Each CV repetition was trained independently for UVPD spectral interpretation. During random forest generation, 30 predictor variables were automatically identified as the most important out of a total space of 407 potential features. Feature scoring and selection was largely consistent between the 3 CV repetitions, and 27 of the selected features were the same between each of the CV repetitions. These primarily represented spectral peaks at specific mass offsets relative to the base position, and as expected, the most important feature corresponded with y^+ ions. Many of the features have not been used in prior de novo software, although the fact that they independently emerged among the most important from each of the CV rounds shows their value and the power of an open machine learning approach to spectral analysis.

UVnovo generated a list of sequences for each spectrum, typically with no more than seven candidates, and sequences were ranked based on descending confidence score. We required a 'correct' sequence reconstruction to exactly match its corresponding SEQUEST PSM, after allowing for indistinguishable residue pairs I/L and $F/M^{Oxidation}$. No sequence gaps or truncations were permitted.

UVnovo produced correct top-ranked sequences for 47.4% of the *E. coli* mass spectra, and when considering the top three *de novo* sequences for each spectrum, 59.8% had a match to the corresponding SEQUEST PSM (Figure 3.5a). The number of correct reconstructions drops substantially with decreasing *de novo* sequence rank (Figure 3.5b). Peptides with correct sequences ranged in size between 6 and 24 amino acids and had an average length of 11.0 residues. This compares to an average peptide length of 11.8 from the total set of SEQUEST PSMs, only two of which were longer than 24 residues.

Exclusion of spectra without high scoring *de novo* reconstructions dramatically improved sequencing precision. This filtered out two thirds of the false positive predictions while retaining 85.5% of true predictions, boosting the precision to 66.4% and 80.4% for top one and top three *de novo* sequence sets, respectively (Figure 3.5c,d). Our ability to identify correct full length sequences from the majority of the test set demonstrates the benefits of AMCA-UVPD for comprehensive and interpretable peptide fragmentation.

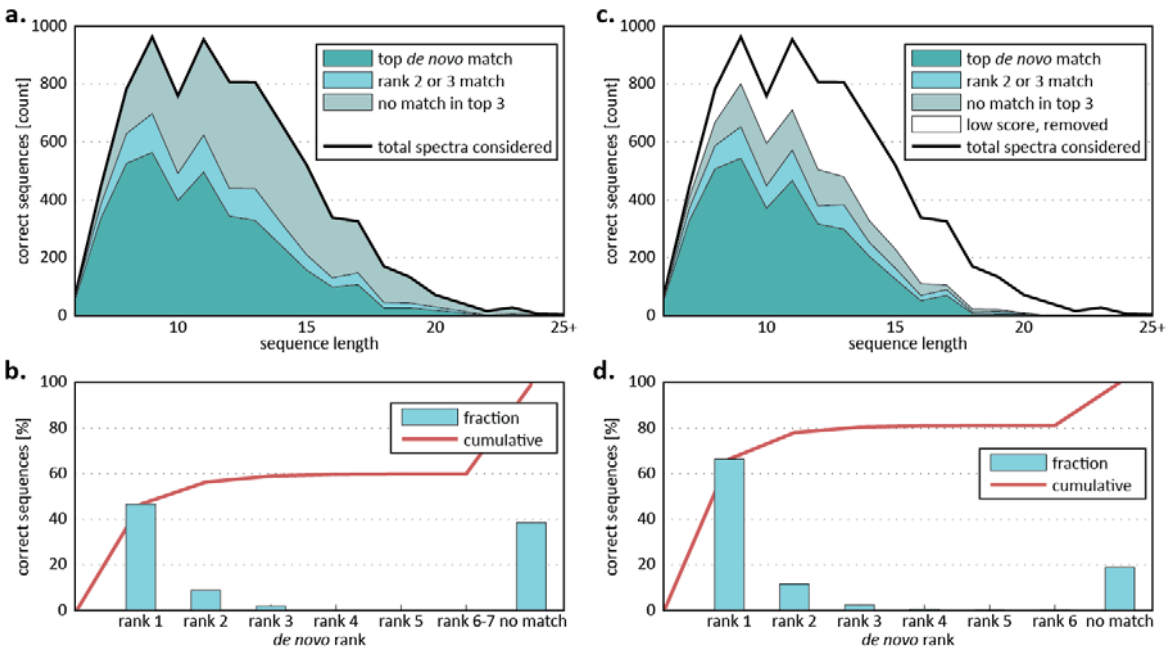

Figure 3.5: UVnovo *de novo* results for the *E. coli* lysate UVPD spectra.

A correct sequence matches the SEQUEST PSM exactly with no gaps. UVnovo scores each sequence reconstruction and ranks it relative to others from the same spectrum. (a) Number of correct sequences versus peptide length for the top-ranked *de novo* result and for the top three *de novo* results. (b) Fraction of correct sequences versus *de novo* rank. (c,d) Filtering of low scoring *de novo* predictions improves sequence-level precision. 5062 of the original 7911 spectra remain, and over 75% of those removed had no correct match.

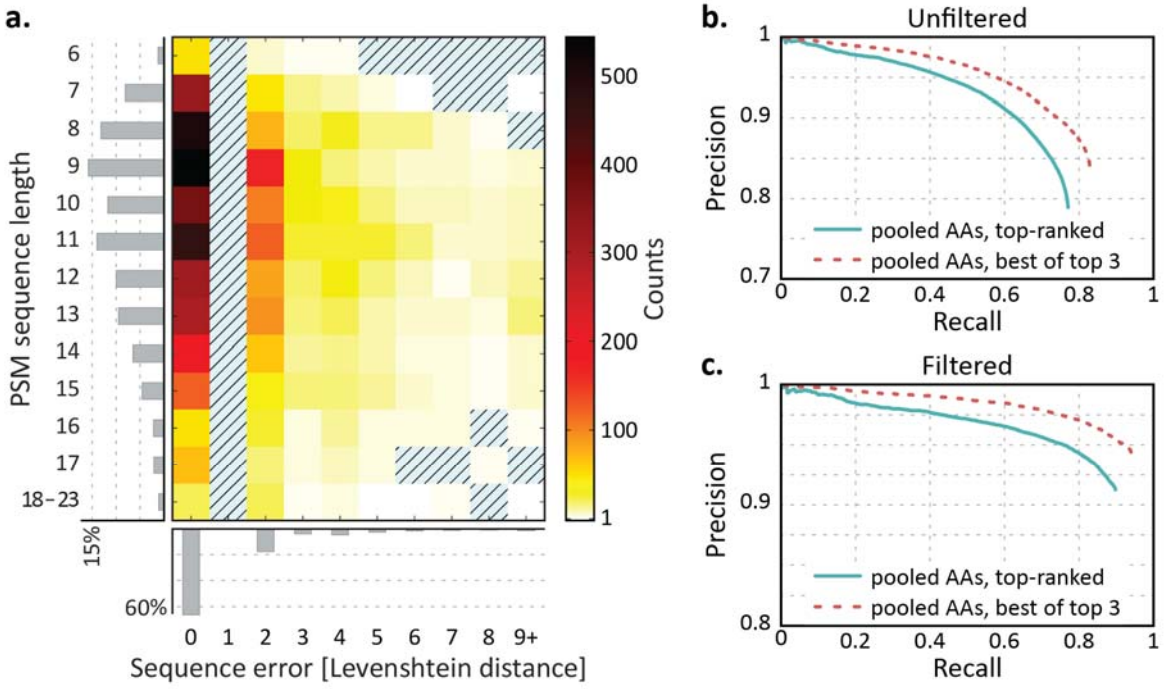

Figure 3.6: UVnovo sequencing error and precision recall of residue predictions.

(a) Amino acid error versus peptide length for top-ranked *de novo* sequences from the filtered set of higher-confidence predictions. Most sequences are correct with no insertions or deletions. Incorrect sequences tend to diverge from SEQUEST PSMs by only 2 residues (a single fragmentation site misprediction). Histograms show fractional counts in each dimension. (b,c) Amino acid precision-recall for the complete and filtered *de novo* results. AAs are pooled and sorted by residue-level score from (blue) the top-ranked *de novo* predictions for each spectrum or (dashed red) the best match among the top 3 predictions for each spectrum.

For those spectra without a correct full-length identification, the highest scoring prediction often differed from its matching PSM at only a single fragmentation site, corresponding to a difference of two amino acids. Figure 3.6a displays the frequency and extent of amino acid sequencing errors versus peptide length in the filtered set. Over half (52.0%) of the misidentified sequences differ from the SEQUEST PSM by only two amino acids, meaning that only one fragmentation site per peptide is recognized incorrectly. Furthermore, each amino acid in a sequence reconstruction has an associated score. Pooling

all residue predictions and sorting by descending score allowed us to plot the amino acid-level precision and recall of residue assignments, a common metric for *de novo* algorithm performance.[104,106,128] In brief, precision is measured as the fraction of correct predictions out of all amino acids predictions, and recall is the fraction of all amino acids in the test set that are correctly identified. Correct counts must match both the residue assignment and mass position along the spectrum. Shown for the total set in Figure 3.6b and the filtered set in Figure 3.6c, the UVnovo precision-recall curves confirm high sequence coverage and low error at the amino acid-level.

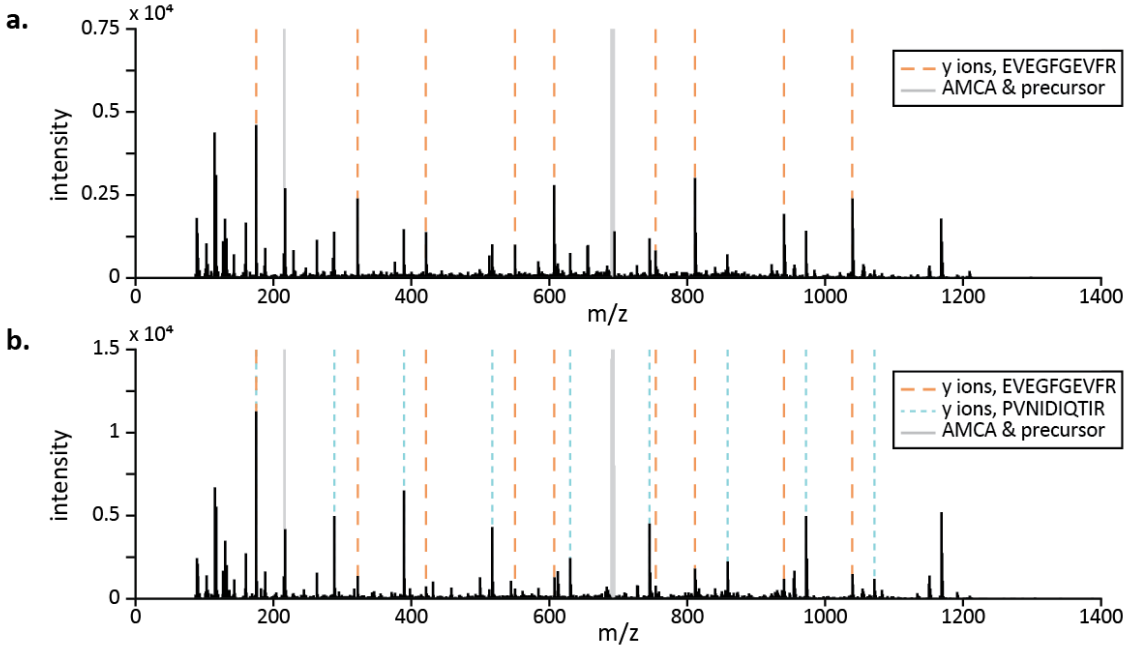

Figure 3.7: UVnovo and SEQUEST each identify different peptides from co-eluting pair.

a) UVnovo and SEQUEST both assign the sequence EVEGFGEVFR. b) Spectrum is acquired 49 seconds after (a). Here, UVnovo assigns PVNIDIQTIR, conflicting with the SEQUEST identification, EVEGFGEVFR. Both sequences are present within the *E. coli* reference database.

Co-eluting peptides in our data sometimes manifest as differences between the *de novo* sequence and SEQUEST PSM for a spectrum. In some cases, this resulted in a hybrid

de novo sequence blended from the two precursor peptides. Ideally, however, the differing *de novo* results complement the SEQUEST identification, and both are correct. As an example, Figure 3.7 presents a pair of co-eluting peptides observed across two spectra. Both SEQUEST and UVnovo identified the first spectrum as EVEGFGEVFR (1383.62 Da). The second spectrum, acquired 49 seconds later in the same injection, took the same SEQUEST PSM, while UVnovo assigned the sequence PVNLDLQTIR (1383.73 Da). Both are present within the *E. coli* sequence database, though the latter was not included in the SEQUEST search due to the presence of proline C-terminal to the tryptic arginine residue. We also observed other 'incorrect' *de novo* identifications with exact matches to semi-tryptic *E. coli* peptides. Such examples indicate inflated error rate estimates in our results and point to the power of *de novo* methods in general for identifying unanticipated peptide variants.

Finally, we note that our results compare favorably to the performance of leading *de novo* sequencing algorithms on high-resolution datasets in general, although specific comparisons on this dataset were not feasible due to the nature of the modifications and ion series employed here. For example, while UniNovo was designed to interpret novel fragmentation spectra, it does not permit user-defined peptide modifications or custom protease specificities.[100] More generally, most available *de novo* software is designed to recognize peptide fragmentation patterns generated through HCD, CID, or ExD, very different from the single *y* ion series we observe, and many of these programs address the antisymmetric path problem with assumptions that would negatively affect results for spectra with unambiguous directionality. Nonetheless, by employing stringent benchmarking criteria (e.g., requiring complete peptide sequence predictions that exactly match corresponding database PSMs), our data show that UVPD/UVnovo accurately

identifies peptide sequences in complex samples and cell lysate contexts through a fully *de novo* sequencing approach.

CONCLUSIONS

We describe new experimental methods and the UVnovo software package for de novo peptide sequencing by UVPD. High efficiency carbamylation blocks lysine side chains, and subsequent tryptic digestion and N-terminal peptide derivatization with the UV chromophore AMCA yields peptides susceptible to 351 nm ultraviolet activation. The UVPD mass spectra, primarily composed of *y* ions, are particularly well suited for *de novo* sequencing. As illustrated in the present study, 351 nm UVPD alleviates two of the fundamental limitations for *de novo* sequencing of standard spectra: incomplete or biased peptide sequence coverage, and spectral symmetry due to observation of both N- and C-terminal ions. Because of the proclivity to generate abundant *y* ions, the spectral peaks are easier to interpret, and the antisymmetric path problem is nonexistent. Additionally, the comprehensive peptide backbone cleavage of UVPD provides the means to reconstruct full or nearly full sequences for most high-quality peptide spectra.

Development of UVnovo was motivated by a lack of appropriate tools for analysis of 351 nm UVPD peptide mass spectra. UVnovo combines random forests and Hidden Markov models to simplify and interpret UVPD fragmentation spectra, enabling the *de novo* sequencing of thousands of peptides from an *E. coli* lysate at high confidence. UVnovo performance, seen here for low-resolution ion trap spectra, broadly matches that of leading *de novo* programs on high-resolution MS/MS spectra. Due to the full sequence coverage provided through UVPD, our workflow offers unprecedented capability for full-length peptide *de novo* sequencing. Further refinement of the UVnovo algorithm is underway and will capitalize on integrating CID and UVPD paired spectra.

Chapter 4: UVnovo *de novo* sequencing of paired CID/UVPD spectra

We describe collection and *de novo* peptide sequencing of matched CID and 351 nm UVPD spectral pairs.* Each precursor ion is isolated twice, the instrument switching between CID and UVPD activation to obtain a complementary MS/MS pair. We generalize UVnovo to concurrently synthesize information from both spectra into a single fusion spectrum, from which it generates *de novo* sequence predictions. Through machine learning, UVnovo shifts the burden of fragmentation model definition from the programmer to the machine, and opens up the model parameter space for inclusion of nonobvious features and interactions. Applied to a benchmark *E. coli* lysate CID/UVPD dataset, the program reconstructed correct full-length *de novo* sequences for 83% of the spectral pairs. For 70% of the pairs, this was the top-ranked *de novo* prediction. This chapter presents the CID/UVPD workflow and demonstrates the capacity of UVnovo for accurate and high-throughput *de novo* peptide sequencing.

INTRODUCTION

The adoption of high throughput bottom-up mass spectrometry for proteomics has accelerated rapidly in the past decade. This is largely driven by improvements in both instrumentation and software interpretation of protein mass spectral data, and more spectra of higher quality may be identified from an experiment than ever before. However *de novo* peptide sequencing, where spectra are identified without use of a reference sequence database, has not seen a commensurate advancement. Limitations inherent in most spectra prevent accurate and full-length *de novo* sequence assignment for bottom-up proteomics

* This chapter represents a manuscript in preparation with contributions from A. P. Horton, S. A. Robotham, J. R. Cannon, D. D. Holden, E. M. Marcotte, and J. S. Brodbelt. S.A.R. and J.S.B. designed the experimental workflow. S.A.R. performed the sample preparation and data collection. D.D.H. modified the instrument control software. A.P.H. designed and developed the UVnovo analysis software and wrote the manuscript.

workflows, and the gap is being filled with ever more elaborate database search workflows and custom processing pipelines.

The previous chapter illustrates how 351nm UVPD can alleviate two major obstacles to successful de novo sequencing: incomplete or biased peptide sequence coverage, and spectral symmetry due to observation of both N and C-terminal ions. We now describe modifications to a Thermo Velos Pro ion trap mass spectrometer that allow collection of CID/UVPD spectral pairs and demonstrate high-throughput collection of CID/UVPD pairs. The greater fragment ion diversity in CID spectra can be a valuable complement to the simple *y* ion series generated through UVPD, and the combination of UVPD and CID spectra offers superior *de novo* sequencing performance compared to use of UVPD alone.

The use of paired spectra has over a decade of precedence for *de novo* sequencing. Matched spectra, generally pairs or triplets produced from different precursor activation methods, contain complementary information that can substantially improve sequencing performance. Savitski et al., in 2005, described "proteomics-grade" *de novo* sequencing from high resolution CID/ECD spectral pairs, and other groups followed.[129] Many software tools now support sequencing of paired or triplet spectra, PEAKS and pNovo+ being the most popular.[99,100,102,130–134] These programs are each limited to one of a few combinations from CID, HCD, ETD, and ECD spectra. Here we report on UVnovo improvements that provide the capability to merge information from multiple spectra of the same peptide.

We initially developed the UVnovo software for *de novo* interpretation and sequencing of 351 nm UVPD mass spectra. UVnovo differs from most other *de novo* sequencing programs in how it models and interprets fragmentation spectra. Most are developed around the offset frequency function (OFF), which represents a descriptive statistical model for understanding and interpreting peptide fragmentation.[108] These

programs may use a directed graph structure to capture simple dependencies between the expected fragmentation peaks or features (ex. *b* ions, neutral losses).[103] The models are typically hand-tuned or, at least, provided a concrete set of features and dependencies, and for well-characterized type of spectra they can perform quite well.

In a departure from the frequency-based statistical models, UVnovo employs a machine learning approach and uses a random forest (RF) algorithm to automatically learn from and interpret mass spectra. RF is a popular and powerful algorithm that combines the predictions of many individual decision trees into a single ensemble.[116] Decision trees and random forest ensembles can exploit a much larger space of features and feature interactions when compared to classical statistical models.[121] The *de novo* sequencing program Novor employs this advantage through use of two large decision trees for spectral interpretation and scoring.[106] UVnovo, too, combines and utilizes spectral features at a scale that is combinatorially impractical with OFF-based models.

While Novor uses the same set of spectral features regardless of activation type, however, UVnovo selects automatically those it finds most important from a much larger space of potential features. In this regard UVnovo follows the work of Datta and Bern, whose spectrum fusion algorithm used the OFF to learn features important for paired CID and ETD spectra interpretation.[119] It then modeled simple feature dependencies and constructed effective tree-augmented networks for making predictions. Unfortunately, their algorithm was only presented as a demonstration of the automated supervised learning technique and was not released for general use. UniNovo also applies the OFF in a similar scheme for automated fragment ion learning.[100] It does not permit user-defined peptide modifications or custom protease specificities, therefore precluding its use for UVPD spectra of AMCA-derivatized peptides.

The generalized UVnovo framework for learning novel fragmentation patterns removes a great burden from the programmer and obviates the need for explicit definition of the fragment ions, their importance and correlation structure. We apply UVnovo to complementary pairs of UVPD and CID spectra and show that *de novo* sequencing performance is greatly improved relative to that for individual spectra.

METHODS

Materials

We purchased Trypsin Gold, mass spectrometry grade, from Promega (Madison, WI, USA), LC-MS grade acetonitrile and water from EMD Millipore (Darmstadt, Germany), and phosphate buffered saline (PBS) and dimethyl sulfoxide (DMSO) from Thermo Fisher Scientific Inc. (San Jose, CA, USA). Sulfosuccinimydyl-7-amino-4-methyl-coumarin-3-acetic acid (Sulfo-NHS-AMCA) was purchased from Pierce Biotechnology (Rockford, IL, USA). *E. coli* lysate was graciously donated by Dr. M. Stephen Trent's research group at the University of Texas at Austin.

Instrumentation for paired CID/UVPD collection

We used a Thermo Velos Pro dual linear ion trap mass spectrometer (Thermo Scientific, San Jose, CA, USA*) coupled to a Coherent 351 nm excimer laser (Coherent, Santa Clara, CA, USA) for UVPD ion activation in the ion trap.[117] Collection of paired CID/UVPD data required modification to the standard instrument runtime procedures, accomplished with custom scripts in Thermo Fisher Scientific's proprietary ion trap control language (ITCL).[135]

Briefly, each selected precursor ion was isolated twice in succession, first for activation by CID and again for UVPD. By setting the NCE parameter to a value above 0, precursors for CID scans were retained in the high-pressure cell (HPC), and the laser was

not triggered. The CID fragment ions were transferred to the low-pressure cell (LPC) for detection. Setting the NCE value to 0 directed precursor ions to the low-pressure cell (LPC), and the laser was pulsed for UVPD activation.

Sample preparation for UVPD analysis

We used whole cell *E. coli* lysate for development and testing of the paired UVPD/CID acquisition and sequencing workflow presented here. Sample preparation and N-terminal chromophore peptide modification were performed as described previously.[77] *E. coli* lysate was carbamylated to block the reactive primary amines of the lysine side-chains by mixing 50 µg of lysate in 50 mM sodium carbonate with 8 M urea and heating for 4 hours at 80 °C. The resulting carbamylated proteins were then buffer exchanged into PBS to remove urea and subsequently digested using trypsin at 37 °C overnight. After digestion, 25 µL of 20 mM AMCA in DMSO was added to the solution and kept in the dark overnight at room temperature. A C18 SPE cartridge was used to clean the samples and remove residual AMCA. Finally, the samples were dried and reconstituted for LC-MS/MS in 98% water / 2% acetonitrile with 0.1% formic acid.

LC-MS/MS analysis and acquisition of a CID/UVPD dataset for benchmarking

Peptides were separated by reverse phase chromatography using a Dionex NSLC 3000 nanoLC (Thermo Scientific; Waltham, MA, USA) interfaced to the UVPD-enabled Thermo Velos Pro described above. Samples eluted over a 360 min gradient, starting with 3% B and increasing to 50% B and using a flow rate of 300 nL/min. Mobile phase A was water with 0.1 % formic acid (v/v), and mobile phase B was acetonitrile with 0.1% formic acid (v/v). Approximately 5 µg of peptide mixture was loaded on a 15 cm column, packed in-house with C18 stationary phase 3.5 µm particles of 140 Å pore size. Five precursor ions were selected following each MS1 scan. Fragmentation switched between CID (NCE

35, 10 ms) and UVPD (15 pulses at 500 Hz and 3 mJ per pulse), and a complementary pair of MS/MS were generated for each selected precursor ion.

UVnovo

UVnovo is implemented in MATLAB (version R2013a), and *de novo* sequence prediction proceeds through four main steps:

Spectrum fusion: Information from all spectra of a precursor ion is synthesized into a single fusion spectrum, where peak 'intensity' scores reflect the predictions of a random forest model trained to recognize peptide fragmentation sites.

HMM prediction refinement: Fusion spectrum predictions are put into a hidden Markov model (HMM) framework and integrated with knowledge of amino acid relationships and valid mass transitions. This model assigns a probability distribution across the valid prefix masses for each successive peptide prefix fragment. A separate HMM is created for each likely sequence length of a precursor peptide.

De novo sequence prediction: A greedy best-path algorithm identifies the most likely peptide sequence spanning each HMM.

Sequence scoring: Candidate peptide sequences are scored and ranked relative to others from the same precursor.

Terms and notation

Subsequent use of the term *mass* indicates a value with nominal mass units, equivalent to the integer count of protons and neutrons composing the referred to species. The conversion from Daltons is imprecise, and we take the product, to the nearest integer, of Dalton mass and a mass defect normalization factor (1/1.000468) to obtain a nominal mass.

The *peptide bare mass*, M, is the mass of residues composing a peptide, without the water and proton (19 Da) of the conventional peptide mass. Each interior position of a peptide is addressed by its N-terminal *prefix mass*, m_{pre}, or its *suffix mass*, m_{suff}, where $m_{pre} + m_{suff} = M$. Here, a *fragmentation site* is the position of a cleavage between two consecutive amino acids, its value given as the integer mass sum of residues N-terminal to that cleavage site.

Spectrum fusion

The spectrum fusion stage applies a random forest predictive model for interpretation of spectra deriving from a common precursor, in effect deconvoluting the fragmentation patterns of multiple MS/MS into a single spectral representation of likely peptide fragmentation sites. While the following description assumes CID/UVPD spectral pairs, the same implementation is used regardless of spectral type or number of different activation methods applied.

The random forest is first trained against matched CID and UVPD spectral pairs to recognize MS/MS ion signatures indicative of peptide fragmentation sites. During spectrum fusion, the model then predicts at each interior peptide location m_{pre} whether that mass represents a true fragmentation site. Each prediction is based on a small set of important features drawn from across the spectral pair, this set learned during the aforementioned RF training procedure.

Before RF training or prediction, spectra are normalized so that values of locally-prominent peaks are more consistent, both across a spectrum and between spectra from the same activation method. Peaks are ranked by descending intensity and then assigned to unit-width mass bins, the lowest rank (most intense) taking precedence in case of collision. Broad differences in peak intensity along a spectrum are reduced by subtracting the

minimum peak rank within a ± 50 Da sliding window, with values then divided by the precursor mass. This neighborhood-adjusted rank score replaces peak intensities, and spectral fusion is performed using these normalized spectra.

The random forest takes as input a feature vector for each mass position m_{pre} where a prediction is to be made. The initial feature space for paired spectra modeling includes 811 variables, though only a subset are used for prediction. Through an iterative feature selection process, UVnovo identifies this smaller set as those most valuable for constructing an effective RF. Trained models use 30 features in the present study. RF training and feature selection is discussed below and includes a description of the 811 initial features. In short, the feature vector for an RF prediction at m_{pre} primarily comprises peaks at specific mass offsets relative to the m_{pre} and m_{suff}, from both of the normalized CID and UVPD spectra.

The fusion spectrum is generated as follows. A feature vector is constructed for each m_{pre}in a peptide and passed to the RF. The 600 decision trees in the ensemble individually predict whether the m_{pre} does (1) or does not (0) represent a fragmentation site, and the mean of the 600 binary predictions becomes the fusion spectrum score at that mass position. Repeating this for each integer mass within a peptide produces the fusion spectrum for that spectral pair.

De novo sequencing of fusion spectra

Details of the hidden Markov modeling, *de novo* sequence assignment and scoring are described in full in the methods section of Chapter 3. We provide an overview here.

A hidden Markov model framework integrates the fusion spectrum predictions with knowledge of amino acid relationships and valid mass transitions.[136] An HMM is created for each likely sequence length l of the precursor and describes all possible length l

sequences with mass equal to the precursor. Every possible fragmentation site is represented by a node in the HMM, and nodes are indexed by m_{pre} and the count of preceding residues. A path through the HMM defines a viable peptide sequence, and the nodes (or hidden states) along the path are each separated by the mass of an amino acid. As co-localized theoretical cleavage sites share the same node (e.g. PEA|S and APE|S), the HMM only provides a prefix mass probability distribution for each sequence fragment site, not the probability for every path through the model.

The single most likely path through an HMM is identified using the Viterbi algorithm.[124] Due to the mass resolution limitations of ion trap spectra, UVnovo does not distinguish between the residue pairs I/L and F/M^{oxidation}. *De novo* sequence predictions were scored as described in Chapter 3.

Random forest training and parameterization

The 811 initial features comprise 404 peak features from each normalized spectrum and 3 features derived only from M and m. The first group includes peak features corresponding to b- and y- ions. However, these common fragmentation products are not explicitly encoded and are instead recognized automatically through the machine learning process.

The peak features are extracted from four ranges in each normalized spectrum, one centered at each of the charge 1^+ and 2^+ prefix mass m_{pre} and suffix mass m_{suff} positions. Each range spans the -50 to +50 Da window around its center, and includes as a feature the normalized rank score for each of the 101 included peaks. For example, the charge 2^+ prefix set contains a feature for each position $({m_{pre}}^{2+} - 25\ \mathrm{m/z})$ to $({m_{pre}}^{2+} + 25\ \mathrm{m/z})$ with a 0.5 m/z step size. The three derived features are calculated as: m_{pre}/M; $\lceil m_{pre}/100 \rceil$; and

$\lceil m_{suff}/100 \rceil$. Respectively, these represent the position within the peptide, the 100 Da region containing the prefix mass, and the 100 Da region containing the suffix mass.

UVnovo applies a backward variable elimination method for feature selection. It trains an initial RF using all 811 features, which are then ranked by their predictive importance in the trained model. The RF is discarded and trained anew using only higher-ranked features from the previous round. This process is repeated thrice more, yielding a final RF trained on 30 features and containing 600 decision trees.

We chose to use $n = 30$ features for constructing the CID/UVPD random forest models after completing a sweep across n from 10 to 55 (step size 5). RF parameter choice can be evaluated easily during model training. Each decision tree in an RF ensemble omits roughly a third of training examples during its construction. Evaluating for each tree the examples that tree left out during training (out-of-bag, OOB) provides an estimate of the RF classification error. With this OOB error metric, we observed how RF performance changes with different feature sets. RF were constructed for each set of the top n features, as ranked by a preliminary RF of 60 features. We repeated this experiment five times, using a different random number generation seed for each of the replicates. This ensured reproducibility and limited some of the randomness of within-replicate comparisons for the different values n.

UVnovo is trained using examples drawn from a representative set of CID and UVPD pairs. Defined by the corresponding PSMs, the collection of true fragmentation sites composes the positive RF training examples. A set of negative cases was created by randomly shifting each of the real fragmentation sites. Therefore, the number of positive and negative examples is balanced 1:1 during the UVnovo training regime, and the OOB error approximates the misclassification rate on these examples. Therefore, this metric does not directly reflect RF classification error on actual spectra. During spectral fusion, the RF

is applied at every potential precursor fragmentation site, where negative predictions are expected to outnumber true fragmentation sites by about 100:1. However, weighting the positive and negative examples to better match expected rates of observation degraded the subsequent HMM performance and *de novo* accuracy (data not shown). By constructing RF with equal representation between positive and negative training examples, the RF is skewed toward more false positive prediction error. This is desired because the HMM *de novo* sequence assignment is much more sensitive to 'missing' peptide fragmentation sites in the RF fusion spectrum.

Benchmarking

SEQUEST

We processed the *E. coli* CID/UVPD spectra, acquired over three injections, using the SEQUEST and Percolator nodes in Proteome Discoverer v. 1.4 (Thermo Fisher Scientific, San Jose, Ca, USA). Both N-terminal AMCA and lysine carbamylation were required as fixed modifications, and optional methionine oxidation was allowed. Spectra were searched against a sequence database that included the UniProt *E. coli* strain K12 reference proteome and common contaminants (from MaxQuant website, http://maxquant.org/contaminants.zip). From the resultant PSMs, we established a high confidence dataset for UVnovo training and validation. We limited this set to spectral pairs from doubly charged precursors with high confidence PSMs (Percolator false discovery rate <%1) with theoretical precursor mass within ± 1.5 Da of the observed value. Additionally, only PSMs for the CID spectra were considered.

UVnovo

We evaluated UVnovo *de novo* sequencing performance on the CID/UVPD spectral pairs as well as on the isolated sets of CID and UVPD spectra. Using a 3-fold cross

validation (CV) regime, we were able to test performance against all PSMs while maintaining independence between the data used for training and testing. Spectra were divided, based on their assigned peptide sequence, into three partitions. In each of the three CV rounds, spectra and PSMs from two of the partitions were used for training an RF, with which UVnovo used for sequencing 'unknown' spectra from the remaining partition.

Due to the low resolution afforded through ion trap analysis, the instrument-defined isolation windows and precursor masses often vary greatly from that of the monoisotopic parent peptide. We therefore performed the RF interpretation and *de novo* sequencing three times for each spectral pair, once using the observed mass and again at -1 Da and +1 Da from the observed mass.

The UVnovo sequence predictions for a given precursor were ranked by an HMM-derived score metric and compared against the corresponding SEQUEST PSM. To be considered correct, each residue in the full-length PSM corresponded exactly in position and mass to a residue in the *de novo* sequence. In other words, no sequence gaps were allowed, and ambiguous residue pairs I/L and $F/M^{oxidation}$ were treated as equivalent in the comparison.

Results and Discussion

E. coli lysate was carbamylated at all primary amines and digested with trypsin. With lysines blocked through carbamylation, subsequent peptide modification installed AMCA at each newly generated peptide N-terminal primary amine. Derivatization with this UV chromophore rendered peptides susceptible to 351 nm UV photoactivation.

ESI-MS/MS was performed using a dual cell linear ion trap mass spectrometer equipped with a 351 nm laser. Following each MS1 survey scan, CID/UVPD spectral pairs were acquired for each of five selected precursors. The generation of UVPD/CID spectra

required modification of the ion trap control software. This instructed the instrument to switch between CID in the high-pressure cell and UVPD in the low-pressure cell for alternating scans.

UVPD of AMCA-modified peptides produces spectra with a strong *y* ion series. 351 nm photoactivation induces cleavage of the C–N peptide bonds and generates b- and y-type fragment ions. The *b* ions retain the N-terminal chromophore and are consequently annihilated with repeated laser pulses during UVPD, while the *y* ions were not activated further and survived in the trap. CID spectra from the same peptides display conventional *b* and *y* ions, the *b* ions shifted by +215 Da, the mass of AMCA. Figure 4.1 shows matched CID and UVPD spectra for peptide $E^{[AMCA]}LVTAAK^{[carbamyl]}LGGGDPDANPR$ identified from *E. coli* lysate.

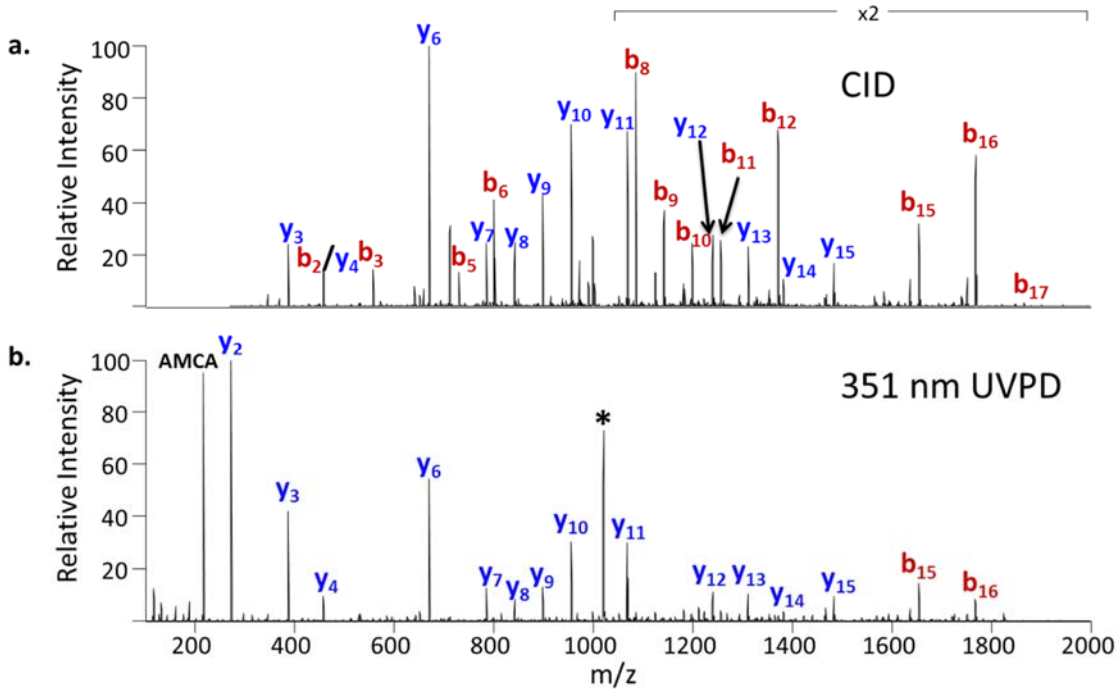

Figure 4.1: CID and UVPD spectra for *E. coli* peptide ELVTAAKLGGGDPDANPR.

(a) CID (NCE 35) and (b) UVPD (3 mJ per pulse, 15 pulses). The precursor is labeled with an asterisk.

UVnovo overview

Spectra, individual or paired, were transformed using a random forest classifier into a single vector of prediction scores for each potential N-terminal fragmentation site. High-scoring positions in the resultant fusion spectrum ideally manifested as a complete and clean ‘sequence ladder’ traversal of the parent peptide in the N- to C-terminal direction. Constructed from this fusion spectrum, a hidden Markov model put the precursor into a framework from which a peptide sequence most descriptive of the data was identified and scored.

Multiple sequence predictions for a precursor were generated, one for each likely peptide length. This was repeated for each mass within ±1 Da of the observed precursor. All sequence predictions for a given precursor were then ranked by the HMM-derived score.

Random forests for CID and UVPD spectral fusion

Random forests are generally robust to small changes in parameterization and have a paucity of tunable parameters relative to most other machine learning algorithms. These strengths, combined with their generally exceptional performance, make random forests among the most popular machine learning algorithms. As with most, however, feature selection still presents a challenge. When presented a surfeit of unimportant features, use of an optimal subset will improve RF time and space efficiency and can reduce prediction error.

UVnovo applies an iterative 'backwards variable elimination' approach for feature selection, by which it winnows the total set of 811 down to a final set of 30 used in production, on RF ensembles grown to 600 trees. The choice of 30 features reliably produced among the most performant CID/UVPD random forests, as measured by OOB error (Figure 4.2).

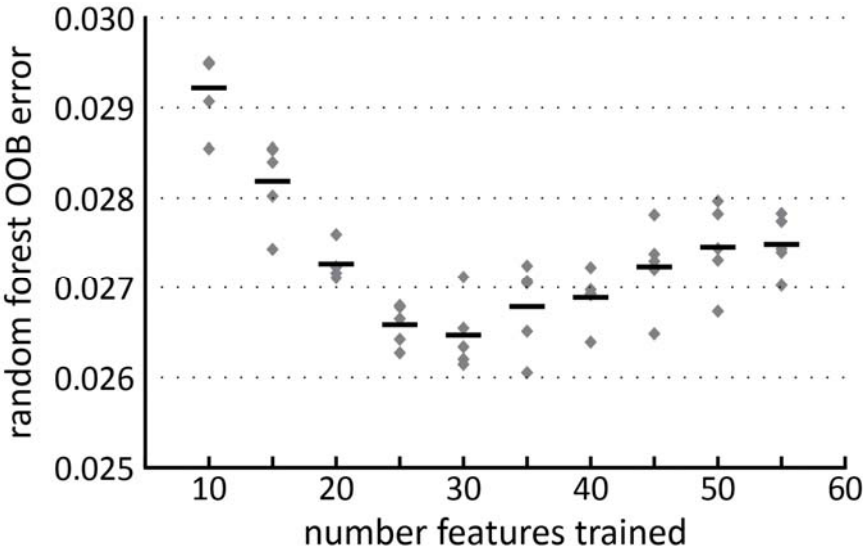

Figure 4.2: Random forest OOB error versus number of features used for training.

Random forests for CID/UVPD spectral pairs were trained varying the number of features used during construction. Black bars show the average of 5 replicates (diamonds). Performance drops with fewer than 30 features as useful predictors are removed.

UVnovo benchmarking on *E. coli* lysate

51,525 UVPD/CID spectral pairs (103,050 MS2 spectra) were collected across three replicate injections. We processed the spectra using Proteome Discoverer SEQUEST with the Percolator node and identified, for charge 2+ CID/UVPD pairs, a set of 4616 high-confidence PSMs covering 1842 unique peptides. These 4616 pairs were applied for UVnovo testing, using 3-fold cross validation (CV) to maintain independence between training and testing examples. UVnovo generated predictions for each at its observed precursor mass and additionally at -1 Da and +1 Da from observed. This was required, as 43% of pairs diverged from the assigned PSMs by ±1 Da.

Comparison of de novo predictions to SEQUEST PSMs

We benchmarked UVnovo predictions from the CID/UVPD spectral pairs against the corresponding SEQUEST PSMs, counting a *de novo* sequence as correct if it matched

exactly the PSM, with no gaps allowed. This is a more stringent criterion than commonly used for *de novo* benchmarks. We additionally compared UVnovo performance on paired spectra to that using only the CID or UVPD scan subsets. Results are presented in Table 4.1 and Figure 4.3.

		rank 1	rank 2	rank 3	rank 4	rank 5	rank 6+	no match
CID/UVPD	correct predictions	3258	449	103	33	13	10	750
	frequency [%]	70.58%	9.73%	2.23%	0.71%	0.28%	0.22%	16.25%
	cumulative	70.58%	80.31%	82.54%	83.25%	83.54%	83.75%	100.00%
CID	correct predictions	1822	463	146	69	34	35	2047
	frequency [%]	39.47%	10.03%	3.16%	1.49%	0.74%	0.76%	44.35%
	cumulative	39.47%	49.50%	52.66%	54.16%	54.90%	55.65%	100.00%
UVPD	correct predictions	2395	618	205	103	46	73	1176
	frequency [%]	51.88%	13.39%	4.44%	2.23%	1.00%	1.58%	25.48%
	cumulative	51.88%	65.27%	69.71%	71.95%	72.94%	74.52%	100.00%

Table 4.1: Count and frequency of correct *de novo* sequences by descending UVnovo rank. Results show performance of UVnovo using CID/UVPD spectral pairs or the individual CID or UVPD spectrum of each pair. The total set contains 4616 charge 2+ paired spectra examples from E. coli lysate with corresponding high confidence SEQUEST PSMs. Frequency describes the fraction of the 4616 with a correct prediction at each UVnovo rank. Correct sequence predictions match the full-length PSM with no gaps allowed. I/L and F/M^{oxidation} residue assignments are treated as equivalent.

Each of the 4616 CID/UVPD examples in the *E. coli* benchmark was provided potentially several putative *de novo* sequence assignments, and the predictions were scored during the sequencing process and ranked once complete. The top-ranked UVnovo sequence correctly matched the corresponding PSM for 3258 (70.6%) of the paired CID/UVPD spectra. In contrast, UVnovo sequencing on the individual spectral types produced only 1822 (39.5%) and 2395 (51.9%) correct top-ranked predictions, for CID and UVPD respectively. When including the three best scoring *de novo* predictions for each precursor, UVnovo correctly sequences 82.5% (CID/UVPD), 52.7% (CID) and 69.7%

(UVPD) of the *E. coli* examples. There is substantial overlap in these correct assignments (Figure 4.3e).

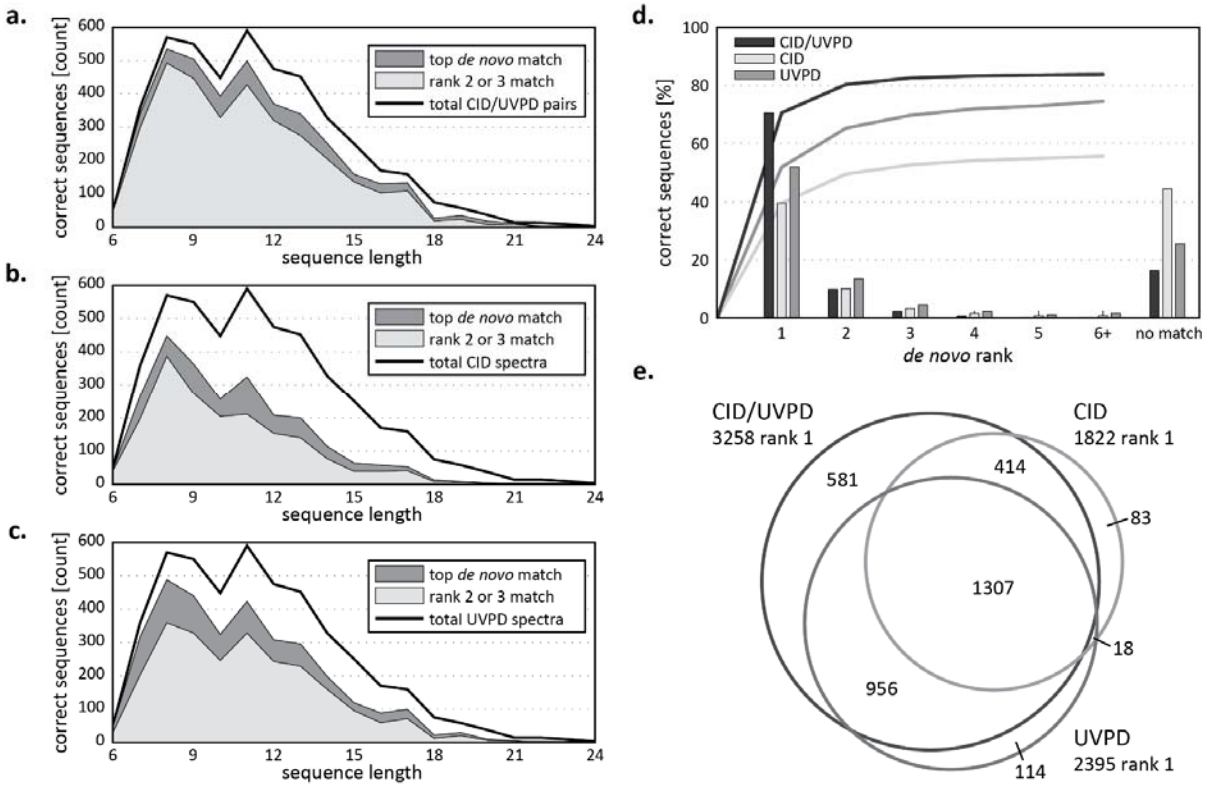

Figure 4.3: UVnovo results for paired and individual *E.coli* lysate spectra.

(a-c) Count of correct sequence reconstructions versus peptide length from the benchmark set of 4616 spectral pairs. Correct *de novo* sequences match the corresponding high confidence SEQUEST PSM, with no gaps allowed. Sequencing of (a) paired CID/UVPD outperforms that using only the (b) CID or (c) UVPD subset of spectra. (d) Fraction and cumulative fraction of correct sequences recovered by descending UVnovo prediction rank. Rank correlates well with prediction accuracy. (e) Overlap [counts] of correct top-ranked sequences between the paired, CID, and UVPD predictions.

The length of SEQUEST PSMs averaged 11.4 residues across the dataset, and correct top-ranked *de novo* reconstructions from the CID/UVPD and UVPD-only results averaged 10.9 residues (Figure 4.3a,c). The CID were shorter on average, at 10.2 residues (Figure 4.3b). UVnovo correctly sequences peptides up to 24 residues in length, the largest

in our dataset, and four of the six peptides with at least 23 residues were identified by their top-ranked CID/UVPD result. To compare, none were identified in the CID analysis and only two in the UVPD.

These results show the advantage of UVPD for *de novo* sequencing, compared to CID, and the largest benefits are realized through synthesis of both CID and UVPD spectra, whereby UVnovo can harness the best properties of each activation method.

UVPD provides comprehensive fragmentation and sequence directionality

UVPD fragmentation occurs consistently across a whole peptide, and the resulting spectra provide better sequence coverage than seen from CID. This characteristic is essential for successful full peptide sequencing and accounts for much of the difference in CID and UVPD performance.

Additionally, the absence of strong N-terminal ions in UVPD spectra eliminates one of the central problems in standard *de novo* analysis. Known as the 'asymmetric path problem', confusion of ion series directionality can lead to inversions in the assigned *de novo* sequence. This is a factor for all fragmentation methods which generate symmetric pairs of N and C-terminal ions,[98] and it is a particular issue for interpretation of CID spectra, due to the presence of symmetric *b* and *y* ion series. No commonly used proteomics workflows for *de novo* sequencing avoid this problem, though various specialized labeling and instrumentation methods have been developed to address the issue.[87,137] As presented here, our AMCA-derivatization and 351nm UVPD workflow offers another way to overcome this problem.

CID complexity complements UVPD simplicity

The complexity of CID spectra can be beneficial in combination with the interpretability of UVPD spectra, and the additional *b* ions and neutral losses substantiate

evidence for true fragmentation peaks. The symmetry between *b* and *y* ions can, detrimental on its own, is effective for inference of the MS/MS precursor mass in UVnovo. UVPD spectra, lacking N- and C-terminal ion symmetries, do not provide such a means. The symmetries in CID spectra are modeled automatically during RF construction as interactions between important features, for example those representing *b* and *y* ion peaks. The symmetries then influence spectral fusion fragment site predictions. When initialized with an incorrect precursor mass, the symmetric features will be misaligned. True fragmentation sites usually score lower, as will the subsequent sequence predictions. Therefore, sequence predictions made using the correct precursor mass will typically rank higher for interpretations involving CID spectra.

Similarly, discrimination between ions of co-eluting peptides is difficult or impossible when using only UVPD spectra, and this can lead to chimeric sequence predictions spanning the fragmentation sites of two or more peptide species. CID symmetries can again be useful and here provide a means to separate peptide ions from species of different mass, sometimes enabling *de novo* sequence generation for both. For example, UVnovo recovers from a single CID/UVPD pair the sequences 'TENLYILPASQTR' and 'VYDALEVQNGNER'. Both appear as tryptic peptides in the *E. coli* sequence database and differ by one unit mass (0.92 Da). The *b* and *y* ion symmetries manifest differently for each and enable correct sequencing for both. Neither was identified when either of the CID or UVPD where processed alone.

Future improvements

These results for paired spectra are comparable to the current state of the art in *de novo* peptide sequencing. When considering that the data was collected on a low resolution ion trap mass spectrometer, this becomes particularly impressive and illustrates the benefit

of 351nm UVPD, stand-alone or complemented with a second activation method, for *de novo* analysis. High resolution mass spectrometry is considered by many to be "exceedingly important" for accurate *de novo* sequencing,[86] and we expect translation of our methods to high resolution CID/UVPD acquisition and analysis would further improve de novo sequencing performance. As a concrete example, the use of accurate precursor mass provides a simple filter for database search results. UVnovo could apply a similar mass filter and flag for removal or resequencing any prediction more than 5 ppm different from the precursor mass. In the current *E. coli* experiment, this mass filter would catch 74.5% of the incorrect CID/UVPD results.

Accurate precursor mass also allows amino acid compositional analysis techniques that improve whole sequence accuracy,[138] and can help fill short sequence gaps. Collection of high resolution MS/MS spectra can provide even more benefit to de novo sequencing platforms. The Vonode software, for example, uses fragment ion mass accuracy for scoring short sequence tags within a peptide.[139]

In our data, nearly half of the incorrect top-ranked CID/UVPD sequence assignments differ at only one fragment site from the corresponding PSM. Both residues flanking this misprediction are therefore incorrect. High resolution MS/MS spectra, or even an accurate precursor mass, would help substantially in correcting these point errors. It would also greatly aid local sequence confidence scores and our ability to fill gaps with the correct residues.

CONCLUSIONS

We have generalized UVnovo to work with matched MS/MS produced through any combination of precursor activation methods, demonstrating here its application on complementary CID/UVPD spectral pairs. Provided a set of training examples, UVnovo

effectively learns from and then utilizes the best properties of each activation method for spectral interpretation. We are not aware of any other software that provides this capability.

The initial state of UVnovo is agnostic to all fragment ion types and any expected patterns or correlations between ions (e.g. neutral losses). Instead, it identifies these features automatically during construction of a random forest classifier. This model is then used for interpretation of unknown spectra.

In the case of paired CID/UVPD spectra, the UVnovo random forest model synthesizes evidence from both to derive stronger predictions of peptide bond location (fragmentation site) than either spectrum could provide on its own. UVPD provides comprehensive fragmentation coverage and a clear directionality for ion series, while the symmetries and redundancies in CID spectra are necessary for precursor mass assignment and improve fragmentation site.

UVnovo generated correct full-length *de novo* sequences for 83% of CID/UVPD spectral pairs, in an *E. coli* lysate dataset with charge 2+ high-confidence PSMs. These results, obtained from low-resolution ion trap mass spectra, demonstrate the effectiveness of a CID/UVPD paired spectra workflow for *de novo* peptide sequencing. Furthermore, with the software improvements presented herein, UVnovo now provides a means for *de novo* interpretation of matched MS/MS generated through any combination of ion activation methods. Continued development of UVPD workflows, on high resolution instrumentation and perhaps using alternatives to CID, will offer exciting prospects for the future of *de novo* proteome analysis.

Conclusions

The theory of a humoral immunity mediated by serum antibodies was proposed over a century ago, though only now can we characterize the individual antibodies composing the serological response to vaccination or infection. Our methods for serum antibody repertoire profiling and VH:VL sequence pairing open an unprecedented view into the nature of the adaptive immune system and provide fresh insight on repertoire organization and diversification, immune system development and memory, and serological repertoire dynamics in both health and disease. Applied to medicine, these techniques will offer powerful tools for disease diagnostics and monitoring, vaccine development and efficacy studies, and therapeutic antibody discovery.

Antibodies are the primary effectors of the B cell adaptive immune response, and as such, direct observation of the serum repertoire is paramount for delineating the response. This has proven difficult because an individual's antibody repertoire encompasses a large diversity of non-germline encoded proteins. We were instrumental in developing methods for the high throughput sequencing of B cells and mass spectrometry of affinity purified serum antibodies, using the individualized antibody sequence database for identification of the antibody spectra.

Challenges specific to antibody repertoire proteomics preclude the use of standard analysis methods and motivated our development of novel tools and approaches for interpretation of human polyclonal antibody repertoires. In particular the process of antibody sequence generation and diversification, through which conserved gene framework regions are interspersed with variable complementarity determining regions (CDRs) and mutated, leads to an enormous expansion of highly similar but distinct antibody proteins. The very similar nature of the resultant peptides, where thousands may

share the majority of their sequence, brings unique difficulties to proteomic interpretation. In particular, standard peptide mass spectra often lack sufficient fragment ion coverage to unambiguously identify CDR-H3 peptides.

I implemented the UVnovo package for *de novo* sequencing of UVPD spectra and later generalized it to work with any combination of various spectral types. Constructed around a random forest machine learning framework, UVnovo automatically learns from annotated training examples how it can best interpret future unknown instances of the same spectra types. This provides a flexible and powerful means for synthesizing information from any number of complementary spectra, and as far as I am aware, UVnovo is unique in this capability. So generated, the UVnovo fusion spectrum is currently used for *de novo* sequence predictions, though it could easily be transformed into a spectral representation optimized for use by standard database search algorithms such as SEQUEST and MaxQuant.

Our methods for collection of paired CID/UVPD spectra and subsequent *de novo* interpretation offer potential improvements to serum antibody characterization. UVPD generates comprehensive precursor peptide fragmentation that manifests as a clean series of *y* fragment ions, and these characteristics make it ideal for *de novo* sequencing. This uniform fragmentation could improve CDR-H3 peptide spectra, especially at the N-terminus where CID coverage is persistently poor. Such application could substantially reduce ambiguous PSM assignments and consequently improve our detection of CDR-H3 peptides. This promises to improve our capabilities for sequence-guided repertoire proteomic studies, and with continued development, our methods may provide a means to identify at the very least, CDR-H3 peptide sequences *de novo*, and more ambitiously, serum monoclonal antibodies through proteomics alone.

References

(1) Lavinder, J. J.; Wine, Y.; Giesecke, C.; Ippolito, G. C.; Horton, A. P.; Lungu, O. I.; Hoi, K. H.; DeKosky, B. J.; Murrin, E. M.; Wirth, M. M.; Ellington, A. D.; Dörner, T.; Marcotte, E. M.; Boutz, D. R.; Georgiou, G. *Proc. Natl. Acad. Sci.* **2014**, *111* (6), 2259–2264.

(2) Boutz, D. R.; Horton, A. P.; Wine, Y.; Lavinder, J. J.; Georgiou, G.; Marcotte, E. M. *Anal. Chem.* **2014**, *86* (10), 4758–4766.

(3) Cheung, W. C.; Beausoleil, S. A.; Zhang, X.; Sato, S.; Schieferl, S. M.; Wieler, J. S.; Beaudet, J. G.; Ramenani, R. K.; Popova, L.; Comb, M. J.; Rush, J.; Polakiewicz, R. D. *Nat. Biotechnol.* **2012**, *30* (5), 447–452.

(4) Georgiou, G.; Ippolito, G. C.; Beausang, J.; Busse, C. E.; Wardemann, H.; Quake, S. R. *Nat. Biotechnol.* **2014**, *32* (2), 158–168.

(5) Mathonet, P.; Ullman, C. G. *Front. Immunol.* **2013**, *4*.

(6) DeKosky, B. J.; Ippolito, G. C.; Deschner, R. P.; Lavinder, J. J.; Wine, Y.; Rawlings, B. M.; Varadarajan, N.; Giesecke, C.; Dörner, T.; Andrews, S. F.; Wilson, P. C.; Hunicke-Smith, S. P.; Willson, C. G.; Ellington, A. D.; Georgiou, G. *Nat. Biotechnol.* **2013**, *31* (2), 166–169.

(7) Wine, Y.; Boutz, D. R.; Lavinder, J. J.; Miklos, A. E.; Hughes, R. A.; Hoi, K. H.; Jung, S. T.; Horton, A. P.; Murrin, E. M.; Ellington, A. D.; Marcotte, E. M.; Georgiou, G. *Proc. Natl. Acad. Sci.* **2013**, *110* (8), 2993–2998.

(8) Rajewsky, K. *Nature* **1996**, *381* (6585), 751–758.

(9) Tarlinton, D.; Good-Jacobson, K. *Science* **2013**, *341* (6151), 1205–1211.

(10) Slifka, M.; Ahmed, R. *Trends Microbiol.* **1996**, *4* (10), 394–400.

(11) Metcalf, E. S.; Klinman, N. R. *J. Exp. Med.* **1976**, *143* (6), 1327–1340.

(12) Plotkin, S. A. *Clin. Infect. Dis.* **2008**, *47* (3), 401–409.

(13) Rappuoli, R.; Bottomley, M. J.; D'Oro, U.; Finco, O.; De Gregorio, E. *J. Exp. Med.* **2016**, *213* (4), 469–481.

(14) Baumgarth, N. *Immunol. Rev.* **2013**, *255* (1), 82–94.

(15) Scheid, J. F.; Mouquet, H.; Ueberheide, B.; Diskin, R.; Klein, F.; Oliveira, T. Y. K.; Pietzsch, J.; Fenyo, D.; Abadir, A.; Velinzon, K.; Hurley, A.; Myung, S.; Boulad, F.; Poignard, P.; Burton, D. R.; Pereyra, F.; Ho, D. D.; Walker, B. D.; Seaman, M. S.; Bjorkman, P. J.; Chait, B. T.; Nussenzweig, M. C. *Science* **2011**, *333* (6049), 1633–1637.

(16) Kodadek, T. *Chem. Biol.* **2014**, *21* (9), 1066–1074.

(17) DeKosky, B. J.; Kojima, T.; Rodin, A.; Charab, W.; Ippolito, G. C.; Ellington, A. D.; Georgiou, G. *Nat. Med.* **2015**, *21* (1), 86–91.

(18) Tan, Y.-C.; Blum, L. K.; Kongpachith, S.; Ju, C.-H.; Cai, X.; Lindstrom, T. M.; Sokolove, J.; Robinson, W. H. *Clin. Immunol.* **2014**, *151* (1), 55–65.

(19) Lu, D. R.; Tan, Y.-C.; Kongpachith, S.; Cai, X.; Stein, E. A.; Lindstrom, T. M.; Sokolove, J.; Robinson, W. H. *Clin. Immunol.* **2014**, *152* (1-2), 77–89.

(20) Obermeier, B.; Mentele, R.; Malotka, J.; Kellermann, J.; Kümpfel, T.; Wekerle, H.; Lottspeich, F.; Hohlfeld, R.; Dornmair, K. *Nat. Med.* **2008**, *14* (6), 688–693.
(21) de Costa, D.; Broodman, I.; VanDuijn, M. M.; Stingl, C.; Dekker, L. J. M.; Burgers, P. C.; Hoogsteden, H. C.; Sillevis Smitt, P. A. E.; van Klaveren, R. J.; Luider, T. M. *J. Proteome Res.* **2010**, *9* (6), 2937–2945.
(22) Sato, S.; Beausoleil, S. A.; Popova, L.; Beaudet, J. G.; Ramenani, R. K.; Zhang, X.; Wieler, J. S.; Schieferl, S. M.; Cheung, W. C.; Polakiewicz, R. D. *Nat. Biotechnol.* **2012**, *30* (11), 1039–1043.
(23) Reddy, S.; Ge, X.; Lavinder, J.; Boutz, D.; Ellington, A. D.; Marcotte, E. M.; Georgiou, G. Rapid Isolation of Monoclonal Antibodies from Animals. US20110312505 A1, December 22, 2011.
(24) Foote, J.; Eisen, H. N. *Proc. Natl. Acad. Sci.* **2000**, *97* (20), 10679–10681.
(25) Arnaout, R.; Lee, W.; Cahill, P.; Honan, T.; Sparrow, T.; Weiand, M.; Nusbaum, C.; Rajewsky, K.; Koralov, S. B. *PLoS ONE* **2011**, *6* (8), e22365.
(26) Vollmers, C.; Sit, R. V.; Weinstein, J. A.; Dekker, C. L.; Quake, S. R. *Proc. Natl. Acad. Sci.* **2013**, *110* (33), 13463–13468.
(27) Glanville, J.; Kuo, T. C.; Büdingen, H.-C. von; Guey, L.; Berka, J.; Sundar, P. D.; Huerta, G.; Mehta, G. R.; Oksenberg, J. R.; Hauser, S. L.; Cox, D. R.; Rajpal, A.; Pons, J. *Proc. Natl. Acad. Sci.* **2011**, *108* (50), 20066–20071.
(28) Boyd, S. D.; Marshall, E. L.; Merker, J. D.; Maniar, J. M.; Zhang, L. N.; Sahaf, B.; Jones, C. D.; Simen, B. B.; Hanczaruk, B.; Nguyen, K. D.; Nadeau, K. C.; Egholm, M.; Miklos, D. B.; Zehnder, J. L.; Fire, A. Z. *Sci. Transl. Med.* **2009**, *1* (12), 12ra23–12ra23.
(29) Ippolito, G. C.; Hoi, K. H.; Reddy, S. T.; Carroll, S. M.; Ge, X.; Rogosch, T.; Zemlin, M.; Shultz, L. D.; Ellington, A. D.; VanDenBerg, C. L.; Georgiou, G. *PLoS ONE* **2012**, *7* (4), e35497.
(30) Larimore, K.; McCormick, M. W.; Robins, H. S.; Greenberg, P. D. *J. Immunol.* **2012**, *189* (6), 3221–3230.
(31) Benichou, J.; Glanville, J.; Prak, E. T. L.; Azran, R.; Kuo, T. C.; Pons, J.; Desmarais, C.; Tsaban, L.; Louzoun, Y. *J. Immunol.* **2013**, *190* (11), 5567–5577.
(32) Ippolito, G. C.; Schelonka, R. L.; Zemlin, M.; Ivanov, I. I.; Kobayashi, R.; Zemlin, C.; Gartland, G. L.; Nitschke, L.; Pelkonen, J.; Fujihashi, K.; Rajewsky, K.; Schroeder, H. W. *J. Exp. Med.* **2006**, *203* (6), 1567–1578.
(33) Trad, A.; Tanasa, R. I.; Lange, H.; Zemlin, M.; Schroeder, H. W.; Lemke, H. *Front. Immunol.* **2014**, *5*.
(34) Langman, R. E.; Cohn, M. *Mol. Immunol.* **1987**, *24* (7), 675–697.
(35) Jackson, K. J. L.; Liu, Y.; Roskin, K. M.; Glanville, J.; Hoh, R. A.; Seo, K.; Marshall, E. L.; Gurley, T. C.; Moody, M. A.; Haynes, B. F.; Walter, E. B.; Liao, H.-X.; Albrecht, R. A.; García-Sastre, A.; Chaparro-Riggers, J.; Rajpal, A.; Pons, J.; Simen, B. B.; Hanczaruk, B.; Dekker, C. L.; Laserson, J.; Koller, D.; Davis, M. M.; Fire, A. Z.; Boyd, S. D. *Cell Host Microbe* **2014**, *16* (1), 105–114.

(36) Jiang, N.; He, J.; Weinstein, J. A.; Penland, L.; Sasaki, S.; He, X.-S.; Dekker, C. L.; Zheng, N.-Y.; Huang, M.; Sullivan, M.; Wilson, P. C.; Greenberg, H. B.; Davis, M. M.; Fisher, D. S.; Quake, S. R. *Sci. Transl. Med.* **2013**, *5* (171), 171ra19–ra171ra19.
(37) Wu, Y.-C. B.; Kipling, D.; Dunn-Walters, D. K. *Front. Immunol.* **2012**, *3*.
(38) Martin, V.; Wu, Y.-C. (Bryan); Kipling, D.; Dunn-Walters, D. *Phil Trans R Soc B* **2015**, *370* (1676), 20140237.
(39) Tabibian-Keissar, H.; Hazanov, L.; Schiby, G.; Rosenthal, N.; Rakovsky, A.; Michaeli, M.; Shahaf, G. L.; Pickman, Y.; Rosenblatt, K.; Melamed, D.; Dunn-Walters, D.; Mehr, R.; Barshack, I. *Eur. J. Immunol.* **2015**.
(40) Frolich, D.; Giesecke, C.; Mei, H. E.; Reiter, K.; Daridon, C.; Lipsky, P. E.; Dorner, T. *J. Immunol.* **2010**, *185* (5), 3103–3110.
(41) Giesecke, C.; Frolich, D.; Reiter, K.; Mei, H. E.; Wirries, I.; Kuhly, R.; Killig, M.; Glatzer, T.; Stolzel, K.; Perka, C.; Lipsky, P. E.; Dorner, T. *J. Immunol.* **2014**, *192* (7), 3091–3100.
(42) Tsang, J. S.; Schwartzberg, P. L.; Kotliarov, Y.; Biancotto, A.; Xie, Z.; Germain, R. N.; Wang, E.; Olnes, M. J.; Narayanan, M.; Golding, H.; Moir, S.; Dickler, H. B.; Perl, S.; Cheung, F. *Cell* **2014**, *157* (2), 499–513.
(43) Pulendran, B. *Proc. Natl. Acad. Sci.* **2014**, *111* (34), 12300–12306.
(44) Fink, K. *Front. Immunol.* **2012**, *3*.
(45) Hofer, T.; Muehlinghaus, G.; Moser, K.; Yoshida, T.; E. Mei, H.; Hebel, K.; Hauser, A.; Hoyer, B.; O. Luger, E.; Dorner, T.; Manz, R. A.; Hiepe, F.; Radbruch, A. *Immunol. Rev.* **2006**, *211* (1), 295–302.
(46) Franz, B.; May, K. F.; Dranoff, G.; Wucherpfennig, K. *Blood* **2011**, *118* (2), 348–357.
(47) Parameswaran, P.; Liu, Y.; Roskin, K. M.; Jackson, K. K. L.; Dixit, V. P.; Lee, J.-Y.; Artiles, K. L.; Zompi, S.; Vargas, M. J.; Simen, B. B.; Hanczaruk, B.; McGowan, K. R.; Tariq, M. A.; Pourmand, N.; Koller, D.; Balmaseda, A.; Boyd, S. D.; Harris, E.; Fire, A. Z. *Cell Host Microbe* **2013**, *13* (6), 691–700.
(48) Wrammert, J.; Smith, K.; Miller, J.; Langley, W. A.; Kokko, K.; Larsen, C.; Zheng, N.-Y.; Mays, I.; Garman, L.; Helms, C.; James, J.; Air, G. M.; Capra, J. D.; Ahmed, R.; Wilson, P. C. *Nature* **2008**, *453* (7195), 667–671.
(49) Kwong, P. D.; Mascola, J. R. *Immunity* **2012**, *37* (3), 412–425.
(50) Corti, D.; Lanzavecchia, A. *Annu. Rev. Immunol.* **2013**, *31* (1), 705–742.
(51) Liao, H.-X.; Lynch, R.; Zhou, T.; Gao, F.; Alam, S. M.; Boyd, S. D.; Fire, A. Z.; Roskin, K. M.; Schramm, C. A.; Zhang, Z.; Zhu, J.; Shapiro, L.; Becker, J.; Benjamin, B.; Blakesley, R.; Bouffard, G.; Brooks, S.; Coleman, H.; Dekhtyar, M.; Gregory, M.; Guan, X.; Gupta, J.; Han, J.; Hargrove, A.; Ho, S.; Johnson, T.; Legaspi, R.; Lovett, S.; Maduro, Q.; Masiello, C.; Maskeri, B.; McDowell, J.; Montemayor, C.; Mullikin, J.; Park, M.; Riebow, N.; Schandler, K.; Schmidt, B.; Sison, C.; Stantripop, M.; Thomas, J.; Thomas, P.; Vemulapalli, M.; Young, A.; Mullikin, J. C.; Gnanakaran, S.; Hraber, P.; Wiehe, K.; Kelsoe, G.; Yang, G.; Xia, S.-M.; Montefiori, D. C.; Parks, R.; Lloyd, K. E.; Scearce, R. M.; Soderberg, K. A.; Cohen, M.; Kamanga, G.; Louder, M. K.; Tran, L. M.; Chen, Y.; Cai, F.; Chen, S.;

Moquin, S.; Du, X.; Joyce, M. G.; Srivatsan, S.; Zhang, B.; Zheng, A.; Shaw, G. M.; Hahn, B. H.; Kepler, T. B.; Korber, B. T. M.; Kwong, P. D.; Mascola, J. R.; Haynes, B. F. *Nature* **2013**.

(52) Wu, X.; Zhang, Z.; Schramm, C. A.; Joyce, M. G.; Do Kwon, Y.; Zhou, T.; Sheng, Z.; Zhang, B.; O'Dell, S.; McKee, K.; Georgiev, I. S.; Chuang, G.-Y.; Longo, N. S.; Lynch, R. M.; Saunders, K. O.; Soto, C.; Srivatsan, S.; Yang, Y.; Bailer, R. T.; Louder, M. K.; Mullikin, J. C.; Connors, M.; Kwong, P. D.; Mascola, J. R.; Shapiro, L. *Cell* **2015**, *161* (3), 470–485.

(53) Stern, J. N. H.; Yaari, G.; Vander Heiden, J. A.; Church, G.; Donahue, W. F.; Hintzen, R. Q.; Huttner, A. J.; Laman, J. D.; Nagra, R. M.; Nylander, A.; Pitt, D.; Ramanan, S.; Siddiqui, B. A.; Vigneault, F.; Kleinstein, S. H.; Hafler, D. A.; O'Connor, K. C. *Sci. Transl. Med.* **2014**, *6* (248), 248ra107–ra248ra107.

(54) Palanichamy, A.; Apeltsin, L.; Kuo, T. C.; Sirota, M.; Wang, S.; Pitts, S. J.; Sundar, P. D.; Telman, D.; Zhao, L. Z.; Derstine, M.; Abounasr, A.; Hauser, S. L.; von Budingen, H.-C. *Sci. Transl. Med.* **2014**, *6* (248), 248ra106–ra248ra106.

(55) Doorenspleet, M. E.; Klarenbeek, P. L.; de Hair, M. J. H.; van Schaik, B. D. C.; Esveldt, R. E. E.; van Kampen, A. H. C.; Gerlag, D. M.; Musters, A.; Baas, F.; Tak, P. P.; de Vries, N. *Ann. Rheum. Dis.* **2014**, *73* (4), 756–762.

(56) Thurgood, L. A.; Arentz, G.; Lindop, R.; Jackson, M. W.; Whyte, A. F.; Colella, A. D.; Chataway, T. K.; Gordon, T. P. *Clin. Exp. Immunol.* **2013**, *174* (2), 237–244.

(57) Lindop, R.; Arentz, G.; Bastian, I.; Whyte, A. F.; Thurgood, L. A.; Chataway, T. K.; Jackson, M. W.; Gordon, T. P. *Clin. Immunol.* **2013**, *148* (1), 27–34.

(58) Arentz, G.; Thurgood, L. A.; Lindop, R.; Chataway, T. K.; Gordon, T. P. *J. Autoimmun.* **2012**, *39* (4), 466–470.

(59) Murphy, M. A.; O'Leary, J. J.; Cahill, D. J. *J. Proteomics* **2012**, *75* (15), 4573–4579.

(60) He, J.; Wu, J.; Jiao, Y.; Wagner-Johnston, N.; Ambinder, R. F.; Diaz, L. A.; Kinzler, K. W.; Vogelstein, B.; Papadopoulos, N. *Oncotarget* **2011**, *2* (3), 178–185.

(61) Faham, M.; Zheng, J.; Moorhead, M.; Carlton, V. E. H.; Stow, P.; Coustan-Smith, E.; Pui, C.-H.; Campana, D. *Blood* **2012**, *120* (26), 5173–5180.

(62) Gawad, C.; Pepin, F.; Carlton, V. E. H.; Klinger, M.; Logan, A. C.; Miklos, D. B.; Faham, M.; Dahl, G.; Lacayo, N. *Blood* **2012**, *120* (22), 4407–4417.

(63) Barnidge, D. R.; Tschumper, R. C.; Theis, J. D.; Snyder, M. R.; Jelinek, D. F.; Katzmann, J. A.; Dispenzieri, A.; Murray, D. L. *J. Proteome Res.* **2014**, *13* (4), 1905–1910.

(64) Barnidge, D. R.; Dasari, S.; Botz, C. M.; Murray, D. H.; Snyder, M. R.; Katzmann, J. A.; Dispenzieri, A.; Murray, D. L. *J. Proteome Res.* **2014**, *13* (3), 1419–1427.

(65) Galson, J. D.; Pollard, A. J.; Trück, J.; Kelly, D. F. *Trends Immunol.* **2014**, *35* (7), 319–331.

(66) Calis, J. J. A.; Rosenberg, B. R. *Trends Immunol.* **2014**, *35* (12), 581–590.

(67) Poulsen, T. R.; Meijer, P.-J.; Jensen, A.; Nielsen, L. S.; Andersen, P. S. *J. Immunol.* **2007**, *179* (6), 3841–3850.

(68) Glanville, J.; Zhai, W.; Berka, J.; Telman, D.; Huerta, G.; Mehta, G. R.; Ni, I.; Mei, L.; Sundar, P. D.; Day, G. M. R.; Cox, D.; Rajpal, A.; Pons, J. *Proc. Natl. Acad. Sci.* **2009**, *106* (48), 20216–20221.
(69) Briney, B. S.; Crowe Jr, J. E. *Front. Immunol.* **2013**, *4*.
(70) Murphy, K.; Travers, P.; Walport, M.; Janeway, C. *Janeway's immunobiology*, 8th ed.; Garland Science: New York, 2012.
(71) Lavinder, J. J.; Horton, A. P.; Georgiou, G.; Ippolito, G. C. *Curr. Opin. Chem. Biol.* **2015**, *24*, 112–120.
(72) Weinstein, J. A.; Jiang, N.; White, R. A.; Fisher, D. S.; Quake, S. R. *Science* **2009**, *324* (5928), 807–810.
(73) Reddy, S. T.; Ge, X.; Miklos, A. E.; Hughes, R. A.; Kang, S. H.; Hoi, K. H.; Chrysostomou, C.; Hunicke-Smith, S. P.; Iverson, B. L.; Tucker, P. W.; Ellington, A. D.; Georgiou, G. *Nat. Biotechnol.* **2010**, *28* (9), 965–969.
(74) Britanova, O. V.; Putintseva, E. V.; Shugay, M.; Merzlyak, E. M.; Turchaninova, M. A.; Staroverov, D. B.; Bolotin, D. A.; Lukyanov, S.; Bogdanova, E. A.; Mamedov, I. Z.; Lebedev, Y. B.; Chudakov, D. M. *J. Immunol.* **2014**, 1302064.
(75) Dekker, L. J. M.; Zeneyedpour, L.; Brouwer, E.; Duijn, M. M. van; Smitt, P. A. E. S.; Luider, T. M. *Anal. Bioanal. Chem.* **2011**, *399* (3), 1081–1091.
(76) Bandeira, N.; Pham, V.; Pevzner, P.; Arnott, D.; Lill, J. R. *Nat. Biotechnol.* **2008**, *26* (12), 1336–1338.
(77) Robotham, S. A.; Horton, A. P.; Cannon, J. R.; Cotham, V. C.; Marcotte, E. M.; Brodbelt, J. S. *Anal. Chem.* **2016**, *88* (7), 3990–3997.
(78) Wine, Y.; Horton, A. P.; Ippolito, G. C.; Georgiou, G. *Curr. Opin. Immunol.* **2015**, *35*, 89–97.
(79) Hale, J. E.; Butler, J. P.; Gelfanova, V.; You, J.-S.; Knierman, M. D. *Anal. Biochem.* **2004**, *333* (1), 174–181.
(80) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. *Nat. Methods* **2007**, *4* (11), 923–925.
(81) Cox, J.; Michalski, A.; Mann, M. *J. Am. Soc. Mass Spectrom.* **2011**, *22* (8), 1373–1380.
(82) Vizcaíno, J. A.; Deutsch, E. W.; Wang, R.; Csordas, A.; Reisinger, F.; Ríos, D.; Dianes, J. A.; Sun, Z.; Farrah, T.; Bandeira, N.; Binz, P.-A.; Xenarios, I.; Eisenacher, M.; Mayer, G.; Gatto, L.; Campos, A.; Chalkley, R. J.; Kraus, H.-J.; Albar, J. P.; Martinez-Bartolomé, S.; Apweiler, R.; Omenn, G. S.; Martens, L.; Jones, A. R.; Hermjakob, H. *Nat. Biotechnol.* **2014**, *32* (3), 223–226.
(83) Marcotte, E. M. *Nat. Biotechnol.* **2007**, *25* (7), 755–757.
(84) Nesvizhskii, A. I. *J. Proteomics* **2010**, *73* (11), 2092–2123.
(85) Elias, J. E.; Gygi, S. P. *Nat. Methods* **2007**, *4* (3), 207–214.
(86) Ma, B.; Johnson, R. *Mol. Cell. Proteomics* **2012**, *11* (2).
(87) Seidler, J.; Zinn, N.; Boehm, M. E.; Lehmann, W. D. *PROTEOMICS* **2010**, *10* (4), 634–649.
(88) Mitchell Wells, J.; McLuckey, S. A. In *Methods in Enzymology*; A. L. Burlingame, Ed.; Academic Press, 2005; Vol. Volume 402, pp 148–185.

(89) Laskin, J.; Futrell, J. H. *Mass Spectrom. Rev.* **2003**, *22* (3), 158–181.
(90) Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M. *Nat Meth* **2007**, *4* (9), 709–712.
(91) Mikesh, L. M.; Ueberheide, B.; Chi, A.; Coon, J. J.; Syka, J. E. P.; Shabanowitz, J.; Hunt, D. F. *Biochim. Biophys. Acta BBA - Proteins Proteomics* **2006**, *1764* (12), 1811–1822.
(92) Wiesner, J.; Premsler, T.; Sickmann, A. *PROTEOMICS* **2008**, *8* (21), 4466–4483.
(93) Madsen, J. A.; Brodbelt, J. S. *Anal. Chem.* **2009**, *81* (9), 3645–3653.
(94) Brodbelt, J. *J. Am. Soc. Mass Spectrom.* **2011**, *22* (2), 197–206.
(95) Reilly, J. P. *Mass Spectrom. Rev.* **2009**, *28* (3), 425–447.
(96) Ly, T.; Julian, R. R. *Angew. Chem. Int. Ed.* **2009**, *48* (39), 7130–7137.
(97) Madsen, J. A.; Xu, H.; Robinson, M. R.; Horton, A. P.; Shaw, J. B.; Giles, D. K.; Kaoud, T. S.; Dalby, K. N.; Trent, M. S.; Brodbelt, J. S. *Mol. Cell. Proteomics* **2013**, *12* (9), 2604–2614.
(98) Song, Y.; Yu, M. *Inf. Process. Lett.* **2015**, *115* (2), 377–381.
(99) Chi, H.; Chen, H.; He, K.; Wu, L.; Yang, B.; Sun, R.-X.; Liu, J.; Zeng, W.-F.; Song, C.-Q.; He, S.-M.; Dong, M.-Q. *J. Proteome Res.* **2012**.
(100) Jeong, K.; Kim, S.; Pevzner, P. A. *Bioinformatics* **2013**, *29* (16), 1953–1962.
(101) Kim, S.; Bandeira, N.; Pevzner, P. A. *Mol. Cell. Proteomics* **2009**, *8* (6), 1391–1400.
(102) Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G. *Rapid Commun. Mass Spectrom.* **2003**, *17* (20), 2337–2342.
(103) Frank, A.; Pevzner, P. *Anal. Chem.* **2005**, *77* (4), 964–973.
(104) Fischer, B.; Roth, V.; Roos, F.; Grossmann, J.; Baginsky, S.; Widmayer, P.; Gruissem, W.; Buhmann, J. M. *Anal. Chem.* **2005**, *77* (22), 7265–7273.
(105) Tabb, D. L.; Ma, Z.-Q.; Martin, D. B.; Ham, A.-J. L.; Chambers, M. C. *J. Proteome Res.* **2008**, *7* (9), 3838–3846.
(106) Ma, B. *J. Am. Soc. Mass Spectrom.* **2015**, 1–10.
(107) Vyatkina, K.; Wu, S.; Dekker, L. J. M.; VanDuijn, M. M.; Liu, X.; Tolić, N.; Dvorkin, M.; Alexandrova, S.; Luider, T. M.; Paša-Tolić, L.; Pevzner, P. A. *J. Proteome Res.* **2015**.
(108) Dančík, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. *J. Comput. Biol.* **1999**, *6* (3-4), 327–342.
(109) Richards, A. L.; Vincent, C. E.; Guthals, A.; Rose, C. M.; Westphall, M. S.; Bandeira, N.; Coon, J. J. *Mol. Cell. Proteomics* **2013**, *12* (12), 3812–3823.
(110) Devabhaktuni, A.; Elias, J. E. *J. Proteome Res.* **2016**.
(111) Keough, T.; Youngquist, R. S.; Lacey, M. P. *Proc. Natl. Acad. Sci.* **1999**, *96* (13), 7131–7136.
(112) Robinson, M. R.; Madsen, J. A.; Brodbelt, J. S. *Anal Chem* **2012**.
(113) Wilson, J. J.; Brodbelt, J. S. *Anal. Chem.* **2007**, *79* (20), 7883–7892.
(114) Robotham, S. A.; Kluwe, C.; Cannon, J. R.; Ellington, A.; Brodbelt, J. S. *Anal. Chem.* **2013**, *85* (20), 9832–9838.
(115) Angel, P. M.; Orlando, R. *Rapid Commun. Mass Spectrom.* **2007**, *21* (10), 1623–1634.

(116) Breiman, L. *Mach. Learn.* **2001**, *45* (1), 5–32.
(117) Gardner, M. W.; Smith, S. I.; Ledvina, A. R.; Madsen, J. A.; Coon, J. J.; Schwartz, J. C.; Stafford, G. C.; Brodbelt, J. S. *Anal. Chem.* **2009**, *81* (19), 8109–8118.
(118) Bajrami, B.; Shi, Y.; Lapierre, P.; Yao, X. *J. Am. Soc. Mass Spectrom.* **2011**, *20* (11), 2124–2134.
(119) Datta, R.; Bern, M. *J. Comput. Biol.* **2009**, *16* (8), 1169–1182.
(120) Degroeve, S.; Martens, L. *Bioinformatics* **2013**, btt544.
(121) Malley, J. D.; Kruppa, J.; Dasgupta, A.; Malley, K. G.; Ziegler, A. *Methods Inf. Med.* **2011**, *51* (1), 74–81.
(122) Niculescu-Mizil, A.; Caruana, R. In *Proceedings of the 22nd International Conference on Machine Learning*; ACM: Bonn, Germany, 2005; pp 625–632.
(123) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes: The Art of Scientific Computing*, 3rd ed.; Cambridge University Press: Cambridge, UK, 2007.
(124) Rabiner, L. R. *Proc. IEEE* **1989**, *77* (2), 257–286.
(125) Forney, G. D. *Proc. IEEE* **1973**, *61* (3), 268–278.
(126) Nelsen, R. B.; Molina, J. J. Q.; Lallena, J. A. R.; Flores, M. Ú. *J. Multivar. Anal.* **2004**, *90* (2), 348–358.
(127) Grossmann, J.; Roos, F. F.; Cieliebak, M.; Lipták, Z.; Mathis, L. K.; Müller, M.; Gruissem, W.; Baginsky, S. *J. Proteome Res.* **2005**, *4* (5), 1768–1774.
(128) Mo, L.; Dutta, D.; Wan, Y.; Chen, T. *Anal. Chem.* **2007**, *79* (13), 4870–4878.
(129) Savitski, M. M.; Nielsen, M. L.; Kjeldsen, F.; Zubarev, R. A. *J. Proteome Res.* **2005**, *4* (6), 2348–2354.
(130) Bertsch, A.; Leinenbach, A.; Pervukhin, A.; Lubeck, M.; Hartmer, R.; Baessmann, C.; Elnakady, Y. A.; Müller, R.; Böcker, S.; Huber, C. G.; Kohlbacher, O. *ELECTROPHORESIS* **2009**, *30* (21), 3736–3747.
(131) He, L.; Ma, B. *J. Bioinform. Comput. Biol.* **2010**, *08* (06), 981–994.
(132) Yan, Y.; Kusalik, A. J.; Wu, F.-X. In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*; 2014; pp 150–155.
(133) An, M.; Zou, X.; Wang, Q.; Zhao, X.; Wu, J.; Xu, L.-M.; Shen, H.-Y.; Xiao, X.; He, D.; Ji, J. *Anal. Chem.* **2013**, *85* (9), 4530–4537.
(134) Guthals, A.; Clauser, K. R.; Frank, A. M.; Bandeira, N. *J. Proteome Res.* **2013**, *12* (6), 2846–2857.
(135) Perkel, J. M. *BioTechniques* **2012**, *53* (6), 339–343.
(136) Blei, D. M. *Annu. Rev. Stat. Its Appl.* **2014**, *1* (1), 203–232.
(137) Brownstein, N. C.; Guan, X.; Mao, Y.; Zhang, Q.; DiMaggio, P. A.; Xia, Q.; Zhang, L.; Marshall, A. G.; Young, N. L. *Rapid Commun. Mass Spectrom.* **2015**, *29* (7), 659–666.
(138) Spengler, B. *J. Am. Soc. Mass Spectrom.* **2004**, *15* (5), 703–714.
(139) Pan, C.; Park, B. H.; McDonald, W. H.; Carey, P. A.; Banfield, J. F.; VerBerkmoes, N. C.; Hettich, R. L.; Samatova, N. F. *BMC Bioinformatics* **2010**, *11* (1), 118.