Prospecting the MHC

Final Summary

Bioinformatics is taking a more prominent role in analysis of HLA and its role in scientific advancement. Every improvement in the tools for molecular analysis, and the techniques that employ them brings about new methods of approaching scientific queries. Improvements in bioinformatic methods enable us to create new scientific queries, which creates a self-perpetuating cycle. This thesis has discussed my use of bioinformatics to make scientific contributions to our understanding of immunogenetics, HLA and the MHC, and how these ideas apply to clinical diagnostics. HLA diagnostics has been evolving, from serological typing, to exon-based molecular genotyping, to identification of full-length gene sequence, to elucidating entire MHC haplotypes. It is critical that our ways of thinking match the evolving technology, in order to fully realize our potential in understanding HLA and immunogenetics.

This thesis describes a two-fold approach to scientific developments. In the first section (**Rules and Tools of HLA Analysis**) I have described techniques that enable molecular analysis, and how we can use these techniques to answer scientific questions. **Chapter 2** describes efforts in collecting and curating sequences in standard databases, which provides more high-quality data to be used in sequencing analysis and diagnostics. Collecting and submitting allele sequences is a difficult and time-consuming project, but the availability of full-length reference sequence is critical to identifying HLA polymorphism, and efforts to simplify this process are of a great value to the community. Community-based collaboration is clearly critical for successful scientific endeavors, which is reflected in **Chapter 3**. Several laboratories contributed samples and sequencing data to a collaborative workshop component, and resulted in a valuable expansion of the IPD-IMGT/HLA database. As our understanding of the MHC advances, the content of IPD-IMGT/HLA will advance as well, and the community will still need to have reference data to understand HLA polymorphism and what it represents.

Targeting and collecting high quality HLA sequences commonly begins with PCR, and one successful technique is described in **Chapter 5.** This assay was analyzed critically, and proven that the resulting amplification produces product that is accurate as well as reliable. We also showed that this product is suitable for use on applicable sequencing platforms, and subsequent use in diagnostics and high-throughput typing. **Chapter 4** described our contribution in bringing a new analysis technique to clinical diagnostics. In a diagnostic setting, it is valuable to have a tool that can quickly and inexpensively type HLA, but there is no room for ambiguities and errors. Applying a new technique (*e.g.* MinION) to a clinical setting brings challenges, and these challenges were met by a focused effort to identify analysis techniques that overcome them. Our techniques were validated, demonstrating that this sequencing platform can be used as a diagnostic tool in a clinical setting.

In the second section (**What's in a Haplotype?**) I have described our contributions to our understanding of MHC haplotypes and immunogenetics. The importance of full-gene polymorphism was discussed in **Chapter 6**, where we indicated how extended polymorphism of HLA-DRA indicates haplotype patterns. Limiting analysis to exons can give an incomplete picture of an HLA gene, and indicates that this gene plays a very limited role. Expanding analysis to intron and UTR sequence gives important insights into the evolutionary history of HLA-DRA, and how it reshapes our understanding of DR~DQ haplotypes. **Chapter 7** expanded on the importance of UTR polymorphism by demonstrating its role in dividing DRB1*13 allele groups and showing that it represents multiple haplotypes. Allele groups are often intended to correlate to serological subtypes, and demonstrating that alleles in this group do not possess characteristic epitopes suggests that another look should be taken at the definition of the DRB1*13 group.

Defining allele groups based on functional distinctions is especially challenging in the HLA-DP region, as we have not yet reached an understanding of what polymorphism is relevant in transplantation. This was explored in **Chapter 8**, where DPA1~Promoter~DPB1 haplotypes were defined and clustered by the intergenic promoter sequences. Correlating the promoter patterns with hypervariable regions demonstrated that clustering the DP haplotypes by this method may be a more effective way to think about this unique HLA region, and may suggest a more functional nomenclature. Allele groups and their correlations with nomenclature and immunogenetics were also a major focus in **Chapter 9**, where we predicted serological specificity of HLA alleles based on amino acid patterns. Efforts to identify serological typing can help to identify permissive mismatches in transplantation, and our efforts to correlate sequence to serology improve the understanding of their relationship, hopefully contributing to general understanding of the nature of HLA.

Together, these projects have created a story about the role of bioinformatics in HLA and immunogenetics. Bioinformatics' multi-faceted and wide reaching approaches makes it perfectly suitable to apply to these fields. I have applied these approaches to identify patterns in HLA sequences, and to observe how these sequences contribute to MHC haplotypes and identifying immunogenic epitopes. Bioinformatics has given me the tools and methodology that I need to approach scientific questions using effective techniques, and it provides new ways to think about problems and develop scientific vision for the future. Bioinformatics techniques will continue to develop, and I am excited to be a part of bringing it into the future.

Valorisation

Rules and Tools of HLA Analysis

Typing and matching of HLA alleles is clearly beneficial in Stem Cell Transplantations, and matching of HLA reduces the effects of Graft-vs-Host disease. Sequencing and matching the full length of HLA at high resolution has also been correlated with improved outcomes, and matching of phased HLA haplotypes improves outcomes even further. High resolution HLA matching is also a strong consideration for Solid Organ Transplantations. The presence of anti-HLA antibodies is the main contraindication for SOT, and high resolution sequencing defines the epitopes that are recognized by the antibodies. Advancements in the platforms and techniques used in HLA sequencing improve the speed and cost-effectiveness of HLA typing, and allow the characterization of full-length HLA polymorphism.

A PCR approach that reliably amplifies 11 HLA loci in four reactions to help in library preparation for sequencing is described in **Chapter 5**. PCR is one of the greatest costs involved in sequence-based typing of HLA alleles, and reduction of these costs enables more HLA laboratories to easily and accurately type these alleles. The availability of a standard primer set allows a more reliable sequencing assay and more consistent analysis and comparison of sequencing data. **Chapter 4** describes the validation of an HLA typing approach which uses nanopore sequencing. MinION sequencing is portable, requires less up-front costs, and requires only minimal laboratory equipment. It generates full-length single-molecule reads, which allows phasing of relatively distant polymorphism and reduces the inherent difficulties of phasing cis/trans polymorphism in heterozygous sequencing by short-read technology. The smaller form factor and relatively short time required for sequencing makes it an attractive target as an on-call typing device. The benefits of MinION are however balanced by challenges in implementation. Basecalling models can struggle with regions of low sequence variation, especially homopolymer sequences, and bioinformatics approaches are necessary to correctly interpret the data. Interpretations of sequencing data, especially from novel platforms, must be validated for accuracy and reliability. MinION sequencing, combined with a validated analysis technique, enables a wider variety of laboratories to sequence and type HLA, to the benefit of the HLA and transplantation communities.

The HLA genes are hyperpolymorphic, which is apparent in the number of unique allele sequences in IPD-IMGT/HLA. The sequence data in this repository is freely available, and is often used in commercial software packages for HLA analysis. The availability of a standard HLA database with official names from the WHO nomenclature committee is of great value to the community. It allows standardization and unambiguous typing and comparison of HLA alleles which can be communicated between any HLA laboratory. Many of the alleles

have only partial sequences available; just 27.3% of the available HLA alleles have full-length (5′ UTR to 3′ UTR) sequences available (release 3.39.0), a significant improvement over the <8% reported by Dr. Steven J. Mack in 2015. This improvement is thanks to local and international efforts to fill the gaps in available full-length sequences. **Chapter 3** describes the results of an international collaboration at the 17th HLA workshop where 34 HLA alleles were extended with complete full-length sequences. Matching of full-length HLA sequences allows matching of a greater amount of polymorphism, compared with matching only the antigen presentation domains, and the availability of full-length sequences allows more specific studies that compare polymorphism between groups.

Sequencing of an individual's HLA genes, especially individuals from under-represented populations, regularly produces novel allele sequences. The submission and naming of these sequences in IPD-IMGT/HLA provides a continuous increase in the known HLA polymorphism, to the benefit of HLA researchers and transplantation clinicians. However, the submission process can be cumbersome, and highly-curated databases often have higher requirements for submission. Gathering the necessary metadata and documentation requires some human effort, **Chapter 2** describes an effort to ease that process. Saddlebags is a freely available tool designed to simplify the process of submission to EMBL/ENA, an important step in submission to IPD-IMGT/HLA, allowing laboratories to more easily participate in submission of novel HLA alleles. Saddlebags has been used by laboratories around the world for submission of HLA class I sequences, and development is continuing to support HLA class II and bulk sequence submission.

What's in a Haplotype?

The HLA genes do not exist in isolation, they are part of a complex and variable MHC region. The second part of this thesis is entitled "What's in a Haplotype?" which reflects a major theme of this thesis. Outside of the HLA field, a haplotype may represent only two linked SNPs, but for HLA researchers a haplotype represents polymorphism in multiple genes, and possibly all polymorphism across an entire chromosome. Regardless, haplotypes are a critical concept in HLA studies. Determining haplotype patterns is an important step in identifying patterns in linkage disequilibrium between SNPs within a gene, or polymorphism at completely different loci. Haplotype studies allow researchers to identify polymorphism that is conserved through evolution, or polymorphism that is commonly inherited together. We can find the relationship between polymorphism of alleles at two loci that encode a protein heterodimer, and clarify how it affects the behavior of the resulting molecule. Haplotypes help us to find new patterns in the organization of genes, and sheds light on the nature of the MHC.

In addition to applications in answering research questions, haplotypes have an important role in transplantation. Matching of phased HLA haplotypes in addition to the unphased

genotypes provides further benefits in stem cell transplantations, perhaps due to implicit matching of unsequenced polymorphism. Haplotypes provide an advantageous effect in the context of haploidentical transplants, that seems to overcome the effects of mismatched HLA alleles. Sequencing haplotypes may help to clarify the linkage disequilibrium patterns and poorly understood mechanisms that provide these beneficial effects. It is clear that identifying patterns in haplotypes increases our understanding of the mechanisms within the MHC, to the benefit of both scientific and clinical applications.

This thesis has expanded our understanding of HLA haplotypes, especially in the class II region. In **Chapter 6**, we explored the role of HLA-DRA polymorphism in DR~DQ haplotypes. Previous literature has described DRA as monomorphic, with a consistent locus within well-defined haplotype patterns. The exon sequences were found to have minimal polymorphism compared to other HLA genes, but we described 20 novel SNPs in the introns and UTR sequences. Haplotype analysis revealed that patterns of polymorphism are correlated with specific HLA-DRB and DQB1 alleles, suggesting that although the DR alpha subunit is evolutionarily conserved, the non-coding polymorphism of HLA-DRA suggests distinct evolutionary lineages and plays an important role in defining DR-DQ haplotypes.

Although previous studies have categorized haplotypes into one of just a few patterns, **Chapter 7** expands our understanding of HLA-DRB1*13 haplotypes and explores the theory of a flexible MHC. We have suggested that the MHC is a flexible and dynamic system which is subject to continued evolution, and that existing haplotypes may not always fall within the definitions of known patterns. This model is presented, not as a conclusive and final definition, but as an idea that can be expanded in further studies by others in the community. As more individuals are sequenced, and more research projects to determine haplotype patterns are carried out by researchers worldwide, the community will further understand how the HLA and non-HLA genes fit together, and how evolutionary pressures affect differentiation between individuals and ethnic groups.

Our understanding of haplotypes was further extended in our studies of the HLA-DP region (**Chapter 8**). DPA1 and DPB1 have an interesting head-to-head orientation with a shared overlapping promoter region. Unlike other HLA loci, HLA-DP nomenclature is not based on allele groups defined by serology. Sequencing the entire region identified common promoter patterns, and haplotype analysis indicates that sequence clusters based on these patterns form strong correlations with the hypervariable regions in DPB1. This suggests a relationship between the promoter region, which likely affects HLA-DP expression levels, and the hypervariable regions in the antigen presentation domain, which affect the HLA-DP immunogenicity. The allele clusters defined by promoter sequences were defined with the goal that future studies and collaborations, such as the

International HLA & Immunogenetics Workshop, can expand on the patterns and clarify their clinical consequences.

The relationship between polymorphism and immunogenicity was further explored in **Chapter 9**. The use of serological HLA typing in a clinical setting is generally decreasing, and the serological typing for many allele sequences is unknown. Serological subtypes of specific HLA-B alleles are not known, and can be difficult to assign due to scarcity of available sera. The serotyping is critical in determining if patient donor-specific antibodies (DSAs) are specific to the transplanted tissue, and models have been proposed to predict serology based on sequence polymorphism. We have proposed one technique for using patterns in specific amino acid polymorphism, compared with alleles with known serotypes, to predict the potential serological subtype of an unknown HLA-B*15 sequence. This method is proposed as an alternative model to existing models that use machine learning-based serology prediction, and it's accuracy and efficacy are free to explore by the community.

The HLA Community

For many of the projects in this thesis, specific software tools were developed for analysis, and software that we created for analysis of HLA sequences is provided as open-source software whenever possible. This includes the code for Saddlebags, as well as Nanopore Prospector, the collection of code and scripts that has provided some capability to analyze MinION reads and HLA allele sequences. The code is available on Github, a widely-used repository for open-source software, and is provided under the GNU GPL 3.0 license, which means that it can be freely downloaded and modified and repurposed for different applications. Providing open-source software has remained a high priority during these studies, since it increases the clarity of how the analysis was performed, and benefits the community by helping other researchers to formulate techniques for analyzing sequencing data or HLA alleles.

The International HLA & Immunogenetics Workshop is a worldwide gathering of researchers and clinicians who work to standardize methodologies, definitions, nomenclatures, and concepts and collaborate on community-focused well-defined projects related to HLA and immunogenetics. The workshop occurs once every 2-5 years, and workshop projects have been a recurring theme in several chapters in this thesis. **Chapter 3** is the direct result of a 17th workshop project where labs collaborated to sequence and submit (**Chapter 2**) full-length HLA allele sequences. This project will be expanded and continued at the 18th workshop. The 18th workshop also features a project focused on DPA1-promoter-DPB1 haplotypes, which will expand on the results identified in our HLA-DP project (**Chapter 8**). We explored the ideas of polymorphic epitopes in **Chapters 7 & 9**, which are related to the planned projects of identifying immunogenetic epitopes and an update of the

HLA dictionary. The 18th workshop will also feature projects focused on bioinformatics, including analysis of recombinations in inherited haplotypes, population genetics, and a community-focused DASH data standards hackathon, all of which relate to projects in this thesis.

All studies in this thesis have been performed with a goal of improving our understanding of HLA for the benefit of patients, clinicians, and researchers in the HLA community. We have put priority on sharing of our results and data whenever feasible, and on active participation in the collaborative congresses and hackathons. This thesis has been focused on the creation and use of software tools which clarify our knowledge of the MHC and which can be applied to many HLA research questions. The projects represented by this thesis are a snapshot in a continuing timeline; it expands on the discoveries of earlier HLA researchers, and the results have the goal of extending the capabilities of future researchers to continue to advance the field of HLA and immunogenetics.

List of Publications

Matern BM, Olieslagers TI, Voorter CEM, Groeneweg M, Tilanus MGJ: Insights into the polymorphism in HLA-DRA and its evolutionary relationship with HLA haplotypes. HLA 2020 Feb;95(2):117-127. doi: 10.1111/tan.13730.

Truong L, Matern BM, D'Orsogna L, Martinez P, Tilanus MGJ, De Santis D: A novel multiplexed 11 locus HLA full gene amplification assay using next generation sequencing. HLA 2020 Feb;95(2):104-116. doi: 10.1111/tan.13729.

Voorter CEM, Matern BM, Tran TH, Fink A, Vidan-Jeras B, Montanic S et al.: Full-length extension of HLA allele sequences by HLA allele-specific hemizygous Sanger sequencing (SSBT). Human Immunology 2018 Nov;79(11):763-772. doi: 10.1016/j.humimm.2018.08.004.

Matern BM, Groeneweg M, Voorter CEM, Tilanus MGJ: Saddlebags: A software interface for submitting full-length HLA allele sequences to the EMBL-ENA nucleotide database. HLA 2018 Jan;91(1):29-35. doi: 10.1111/tan.13179.

Duygu B, Matern BM, Groeneweg M, Voorter CEM, Tilanus MGJ: Polymorphism at residue 156 of the new HLA-A*02:683 allele suggests immunological relevance. HLA 2017 Aug;90(2):107-109. doi: 10.1111/tan.13059.

Matern BM, Olieslagers TI, Groeneweg M, Tilanus MGJ: Division of HLA-DRB1*13 haplotypes by extended HLA-DRA 3'UTR polymorphism refines HLA-DRB1*13~HLA-DRB3~HLA-DQB1 haplotypes and gives clues to HLA-DR13 immunogenicity. In Preparation

Duygu B, Matern BM, Wieten L, Voorter CEM, Tilanus MGJ: Specific amino acid patterns define split specificities of HLA-B15 antigens enabling conversion from DNA based typing to serological equivalents. Submitted (HLA)

Truong L, Matern BM, Groeneweg M, D'Orsogna L, Martinez P, Tilanus MGJ, De Santis D: Polymorphism clustering of the 21.5kb DPA-Promoter-DPB region reveals novel extended full length haplotypes. Submitted (HLA)

Matern BM, Olieslagers TI, Groeneweg M, Duygu B, Wieten L, Tilanus MGJ, Voorter CEM: Long-read nanopore sequencing validated for HLA typing in routine diagnostics. Submitted (Journal of Molecular Diagnostics)