HHS Public Access

Author manuscript

Cell Immunol. Author manuscript; available in PMC 2017 January 01.

Published in final edited form as:

Cell Immunol. 2016 January ; 299: 6–13. doi:10.1016/j.cellimm.2015.10.012.

iWAS- A Novel Approach to Analyzing Next Generation Sequence Data for Immunology

Benjamin Vincent[1,9], **Adam Buntzman**[2,3,9], **Benjamin Hopson**[4,5], **Chris McEwen**[5], **Lindsay Cowell**[6], **Ali Akoglu**[7], **Helen Zhang**[8], and **Jeffrey Frelinger**[2]

Benjamin Vincent: benjamin.g.vincent@gmail.com; Chris McEwen: chris.mcewan@cambridgeconsultants.com; Lindsay Cowell: Lindsay.Cowell@utsouthwestern.edu; Ali Akoglu: akoglu@email.arizona.edu; Helen Zhang: hzhang@math.arizona.edu; Jeffrey Frelinger: jfrelin@email.arizona.edu

[1]Department of Medicine, University of North Carolina, Chapel Hill NC 2714

[2]Department of Immunobiology, University of Arizona, Tucson AZ 85724

[5]Cambridge Consultants, Science Park, Milton Rd, Cambridge, CB4 0DW, UK

[6]Division of Biomedical Informatics, Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, TX 75390

[7]Department of Electrical and Computer Engineering, University of Arizona,Tucson, AZ 85721

[8]Department of Mathematics, University of Arizona,Tucson, AZ 85721

Abstract

In this communication we describe a novel way to use Next Generation Sequence from the receptors expressed on T and B cells. This informatics methodology is named iWAS, for immunonome Wide Association Study, where we use the immune receptor sequences derived from T and B cells and the features of those receptors (sequences themselves, V/J gene usage, length and character each of the CDR3 sub-regions) to define biomarkers of health and disease, as well as responses to therapies. Unlike GWAS, which do not provide immediate access to mechanism, the associations with immune receptors immediately suggest possible and plausible entrée's into disease pathogenesis and treatment.

Keywords

Immune Receptors; Repertoire; V(D)J Recombination; Association Study

[3]Present Address, Department of Medicine, University of Arizona, Tucson AZ 85724, buntzman@deptofmed.arizona.edu
[4]School of Engineering, The University of Edinburgh, Edinburgh, EH9 3JL, UK, Present address: Cambridge Consultants ben.hopson@cambridgeconsultants.com
[9]These authors contributed equally to this work.

1. Introduction

Immunological tolerance is one of the key characteristics of adaptive immunity. The mammalian immune system has the ability to differentiate between self and foreign agents, by tolerating the presence of antigens from self and commensal microbes, while responding robustly to foreign antigens. It has been 70 years since Ray Owen's seminal observation in dizygotic cows that led to the development of the concept of tolerance [1]. The subsequent 7 decades of work on lymphocyte development, antigen processing/presentation, thymic regulation and peripheral tolerance mechanisms have shed light on the machinery responsible for establishing tolerance [1,2]. It is clear that lymphocyte deletion, lymphocyte anergy, and lymphocyte repression are central to the development of self tolerance by establishment of a tolerant adaptive immune receptor repertoire.

While we know much about WHERE and HOW lymphocytes are selected and enter the circulating repertoire, we know precious little about WHICH lymphocytes are selected in the process of developing tolerance. The challenge is identifying the characteristics of adaptive immune receptor sequences that increase or decrease the likelihood of selection. The main limitation in identifying which T cells and B cells are modulated (deleted, anergized or repressed) during tolerance is the sheer magnitude of the potential repertoire of variation that is possible in the species of T cell receptors (TCR) and B cell receptor (BCR=antibodies) as well as their partially non-germline encoded character. Initially, the sequencing of immune receptors was only possible in small numbers that were insufficient to find patterns in the tolerance process. A clear example of the under-sampling of the pre- and post-selection thymic repertoire utilizing classical sequencing techniques is highlighted in research from the Mathis lab [3]. In this work, over 600 TCR sequences were analyzed in an intensive sequencing effort to characterize the repertoire changes that occur during central tolerance in the thymus. In spite of this effort, it is clear that the thymic repertoire was under-sampled and the main conclusions from the manuscript were restricted to observations of skewed J-beta gene usage for selected V-gene bearing TCRs. Ten years after that study, a novel strategy to overcome this sampling conundrum was developed by the same group, wherein a TCR α minilocus transgenic mouse with limited diversity in the alpha chain locus was utilized to assess the changes in TCR repertoire during central tolerance [4]. In spite sequencing many hundreds of TCRs, the authors came to the conclusion that class I and II restricted clones differ but “no overriding rule emerges, and minute sequence differences can switch MHC class preference”. However, with the advent of high throughput sequencing technologies, this sampling limitation has been essentially circumvented. Pre-selection and post-selection repertoires can now be readily assessed at sufficient coverage to allow the interrogation of the selection process to, in essence, identify the cellular clonotypes that are deleted, anergized or repressed upon tolerance induction.

A remaining challenge is the ability to informatically mine such datasets for those critical immune receptors that are involved in the tolerance process. Inspired by the computations utilized in Genome Wide Association Studies (GWAS), we propose an algorithm that we now term an immunome Wide Association Study (iWAS) to mine immune repertoire for immune receptor sequences that correlate with immunological phenotypes (e.g. including but not limited to the development of tolerance).

1.1 History of Immune Tolerance Genetics

The idea of using genetic approaches to interrogate the status of the immune system in health and disease is not new. In the 1960s, data in mice showed that genetics could control immune responses. Biozzi successfully selected two strains of mice from an outbred colony for high and low antibody responses to sheep erythrocytes, demonstrating that differences in adaptive immune responses were characteristic of a particular genotype and thus under the control of genetic regulation [5]. Benacerraf and colleagues were then able to identify guinea pigs that were high and low responders to synthetic polypeptides, and that the Major Histocompatibility Complex (MHC) controlled this response [6,7]. This research was quickly extended to mice where more refined genetic tools solidified the result. It became clear that genetic regulation of the adaptive immune response was generalizable and genetic regulation could be mapped to specific gene loci. These seminal studies, along with the need for histocompatibility matching in transplantation, drove the characterization of the MHC. Skin transplantation in humans and other outbred species demonstrated that nearly all unrelated individuals rejected skin grafts. This was the first evidence that the MHC was highly polymorphic. The following years demonstrated that the MHC was perhaps the most polymorphic region of the mouse and human genome [8], with more than 12,000 alleles of human class I and II genes identified to date [9]. This extensive polymorphism created a tool for examining the association of the immune response and a disease susceptibility or resistance phenotype. Thus, the MHC locus proved useful as a marker for association studies, to allow for the examination of the correlation of a particular allele or alleles with a particular disease state [10-12]. The best known of these include the striking association of HLA-B27 family alleles with ankylosing spondylitis (AS). The relative risk for developing AS is nearly 100 fold higher in HLA-B27+ individuals compared to the general population. Important susceptibility and protective associations have also been found. In Type 1 Diabetes there are both disease associated and protective alleles. DR3 and DR4 individuals have increased risk for developing Type I Diabetes and in DR3/DR4 heterozygotes the risk is further increased [13]. In contrast, DR2 individuals are protected from developing Type 1 Diabetes [14]. Thus, the concept of using immune markers in disease association is well established. While there are many studies examining the genetics of host resistance, the allele phenotypes tend to be simple and the polymorphism not extensive.

The most heavily investigated alleles in the MHC locus, Class I and Class II, are associated with antigen processing and presentation of peptides to CD8+ and CD4+ T cells respectively. Variations in these alleles clearly impact the immune response directly. However, there are many other immune related genes that map within or close to the Class I and Class II genes. These include complement components C4, C2 and Factor B, TNFa, Lymphotoxin and B as well as the transporter of peptides, TAP. All of these genes show some polymorphism and murine gene knockouts show robust phenotypes raising the possibility that they could contribute to the involvement of the MHC and immune phenotypes.

1.2 Genome-Wide Associations with Immune Tolerance

The discovery that single nucleotide polymorphisms (SNP) are common throughout the genome, coupled with the ability to perform detailed interrogation of many individuals using

microarray based technology opened a new frontier. Many polymorphisms could be tested simultaneously and correlated with disease phenotype. Over the ensuing years, many studies utilizing this approach have identified non-MHC candidate genes that modulate immunological phenotypes. Generally referred to as Genome Wide Association studies (GWAS), it was widely applied to large cohorts of patients and controls. Literally hundreds of disease phenotypes have been examined in GWAS studies ranging from autoimmune diseases like Type 1 Diabetes, Multiple Sclerosis, and psoriasis, to immunologically related birth defects [15]. A major limitation to optimal utilization of GWAS data is statistical and computational. When many traits (tens of thousands) are interrogated, false positive associations are common and statistical correction becomes challenging. A second limitation in this approach is that most SNPs do not occur in protein coding regions, and thus it difficult to directly infer function. SNPs in noncoding regions potentially regulate transcription by altering enhancers or promoters, but direct linkage to function is difficult given sequence variant data alone. Thus SNPs may allow associations of a phenotype with a region of the genome, but not with a specific gene. This is formally equivalent to the statistical analysis of Quantitative trait loci, where well-developed mathematical approaches were both applied directly and further developed. Additional limitations of GWAS include its weakness at rare variant discovery [16], continued issues with “missing heritability” [17], and the presences of ascertainment biases that lead to overestimation of the effect size during the initial discovery of an association in under-powered GWA studies, termed the “winner' curse” [18]. These and other characteristics of GWAS have led the field to increase the size of the cohorts to enormous numbers. Since the initial GWAS in 2005 the samples sizes have grown from 146 to over 339,000 individuals [19,20]. An interesting analysis of GWAS, focusing on the immune system [21], catalogs many of the discussions about the utility of GWAS and its limitations.

Nonetheless, GWAS has become a powerful tool for the discovery of genes that contribute to heritable disease phenotypes in monogenic and polygenic diseases. Rheumatoid Arthritis, Type I Diabetes, Lupus, celiac disease, Chrohn's disease, Multiple Sclerosis, Behcet's disease, Vitiligo, ankylosing spondylitis, psoriasis, Graves disease, eczma and other atopic allergic diseases, Wegener's, Sjogren's, and ANCA-associated vasculitis have all been investigated in large productive GWAS studies [22]. Association studies have also identified immune system contributions to unexpected diseases, most notably schizophrenia [23]. These studies have extended our understanding of the immune system, however GWAS studies have been less successful in elucidating genetic contributions of the adaptive immune system genes that undergo V(D)J recombination, class switch recombination and somatic hyper-mutation (TCR alpha, beta, gamma, delta, Ig Heavy and light chains) due their extreme non-germline modifications. The extended Major Histocompatibility Complex (MHC) locus has been identified as a disease-contributing locus in a large number (over 200) of GWAS studies [15], whereas only a limited number of studies have mapped the genetic contributions of the adaptive immune loci (TCR and Ig) to disease phenotypes. The sleep disorder Narcolepsy, Trypanosoma cruzi cardiomyopathy, poor Renal transplant outcome (TRAV19/20), and Intraoccular pressure (TRAJ17) have been associated with the TCR alpha extended locus [24-27]. The Ig heavy chain locus has been genetically associated with Kawasaki's disease, Alzheimer's, and Multiple Myeloma [28-30]. There is an

imbalance in the number of associations between the TCR and the MHC loci, the very genes whose products present peptide antigens to those of the TCR loci. This imbalance suggests that GWAS fails to identify the contributions of TCR genes to relevant disease processes.

1.3 iWAS: immunome Wide Association Study

The non-germline nature of V(D)J recombination, at the TCR/Ig loci, necessitates a novel experimental and statistical approach. Using high throughput immune receptor sequencing to identify “biomarkers” of the adaptive immune system we developed the immunome Wide Association Study (iWAS) to compensate for the weakness of GWAS studies when investigating adaptive immune receptor repertoire characteristics. The iWAS tool should be able to identify the differential characteristic of the Ig and T cell receptor repertoire that correlate with stratified immunological outcomes, like characterizing the repertoire differences between unsuccessful and protective vaccination outcomes [31].

However, most relevant to Ray Owen's work on tolerance, iWAS could be used for identifying the repertoire differences between pre-selection and post-selection thymic TCR repertoire (DN1 vs single positive thymic populations). While measurement of SNPs allowed the development of GWAS analysis, until recently there was no comparable method to interrogate the adaptive immune system receptor sequences. The great challenge for developing an immunome Wide Association Study (iWAS) is that the targeting receptors (T cell Receptors and Antibodies) used by lymphocytes to confer antigen specificity of the immune response are only partially germline encoded. During lymphocyte development, each lymphocyte has the potential to generate a unique immune receptor.through enzymatic rearrangement of Variable (V), Diversity (D), and Joining (J) germline gene segments (TCRβ, TCRδ, and Ig heavy chain) or V and J germline gene segments (TCRα, TCRγ, Ig light chain). A variable number of germline nucleotides may be removed and non-germline encoded nucleotides added at each of the recombination junctions. These non-templated additions are created stochastically by both nucleases and polymerases in each lymphocyte and as such can not be mapped using the SNP analysis performed in classical GWAS. Thus the extent of variation of the immune receptors that the adaptive immune system depends on for its function cannot be captured, which represents the most significant limitation to using GWAS analysis to study the role of the adaptive immune system in immunological phenotypes. The magnitude of diversity of the immune receptor repertoire also poses a significant impediment to analysis, in that the potential repertoire is estimated to be between 10^{15} and 10^{18} unique receptors. While there are only about 10^{11} immune cells in a human, this still represents a significant increase in data complexity over classical SNP analysis which typically relies on 10^{5}-10^{6} SNPs [32].

Before Next Generation Sequencing (NGS), it was impossible to determine the impact of the adaptive immune receptor repertoire on the subsequent immune response because the prior techniques provided insufficient sampling of the repertoire. DNA sequencing of immune receptors TCR and antibodies previously required cloning and sequencing each immune cell receptor one by one. It was a great effort and cost to obtain even a few hundred sequences. This was far too few to effectively interrogate the breadth of an immune repertoire. Thus, sequence analysis was largely confined to examining the antigen specific response with the

idea of better understanding the biochemical interaction of immune receptors and their targets. There have been many past efforts to define sequence motifs that correlate with the ability of a specific T cell to respond to a particular epitope. These were traditionally defined by examining the TCR sequences of a set of antigen specific T cell clones and defining motifs among the TCR expressed by the distinct clones. This analysis was done by visual inspection and was performed by first asking if the Vb or Jb gene usage itself was skewed, then picking the most frequent Vb and aligning the CDR3 regions for the highest sequence similarity and determining visually if a consensus is created from the set of alignments. The establishment of a consensus motif allowed the examination of datasets derived from unselected cells for the presence of those sequences and their relative clone frequencies. This type of analysis was narrow in scope and, by necessity, ignored the contribution of the remaining repertoire of immune receptors. In addition, the requirement that sequences with a similar phenotype have a similar primary sequence motif is unreasonable in polyclonal responses in the setting of high polymorphism at the MHC and other loci. The advent of Next Generation Sequencing (NGS) has both increased the throughput of sequencing and decreased the cost more than five orders of magnitude in the last ten years. This comes purely as a result of technical improvements and the ability to parallelize the biochemistry and data acquisition. NGS techniques have been adapted from genomics and RNAseq applications to immunomics by focusing the sequencing efforts on the immune receptor's VDJ recombination junction. A number of basic science and commercial groups have developed assays for sequencing TCR/BCR repertoire utilizing NGS instruments (e.g. Illumina, Roche 454, Ion Torrent) and have coined multiple names for the technique including Repertoire-Seq, Rep-Seq, innunoSEQ, iR-Profile, ClonoSeq, T-cell metagenomics, LymphoSIGHT, etc [33-37]. The field has progressed sufficiently so that we are capable of sequencing tens of millions of immune receptors from hundreds of individuals; hence we can now comprehensively evaluate the extent and breadth of the adaptive immune system.

The throughput of sequencing immune receptor is now sufficient to replace the markers used in the classical association studies (SNP's) with markers that will reflect the actual diversity of the adaptive immune response, the immune receptor sequences (TCR and antibodies). While the field of informatics has also greatly improved, it lags significantly behind the biochemistry of high throughput sequencing. Current analyses have described the extent of clonal diversity in healthy repertoires and in some cases correlated repertoire diversity measures to clinical outcomes [31,38], however detailed analyses of population sequence-level characteristics to immunologic or clinical phenotypes have not been done. Informatics to assess the contribution of the adaptive immune system to immune phenotypes is virtually non-existent, necessitating the development of the iWAS pipeline.

2. Material and Methods

In principle, the concept of the GWAS and the iWAS is the same. There are an array of features or markers (TCR/Ig sequences and/or TCR/Ig sequence characteristics; analogous to the SNPs used in a GWAS), which are measured in each immune receptor for each individual within two populations (e.g. control and effected). These populations could represent patients who are either well (outcome 1) or unwell (not outcome 1), or the

populations could represent those who exhibit differential biological characteristics (e.g. good vaccine responders vs. poor vaccine responders), or the repertoire sequences themselves can be grouped by intrinsic classifiers (e.g. sequences shared by multiple individuals vs. those unique to one individual). iWAS provides a method by which adaptive immune receptor repertoire features can be associated with outcome/biological group with the opportunity to perform robust statistical significance testing.

While there are many statistical methods to test for associations between quantitative predictive markers and a trait or a biological characteristic of interest, the nature of the marker variables measured often impacts the choice of statistical methods chosen for determining an association. The choice of statistical tool often depends on whether the outcome trait variable is continuous or categorical and it also depends on the complexity and size of the datasets (e.g. if there are a large number of parameters relative to the number of individuals in the dataset, the so-called large p, small n problem, as in a classical GWAS analysis). In this manifestation of an association study, we tailored the initial embodiment of the iWAS to focus on a binary outcome question, where we desired to identify the predictor variables that most contribute to the variance of the binary outcome (e.g. TCR sharing, described below). We chose to develop an association method that relied upon penalized logistic regression (PLR) where we performed a joint analysis as opposed to a marginal analysis on each predictor variable (as is typically performed in a GWAS). iWAS utilizes shrinkage techniques for culling predictors with minimal to no predictive effects in the model. Briefly, the iWAS analysis minimizes a penalized logistic log-likelihood function to identify a comprehensive set of TCR characteristics (the predictors $\mathbf{X} = (X_1, \ldots, X_p)$) which are associated with the prediction of the response variable Y (e.g. a binary outcome phenotype like shared .vs. unshared). iWAS applies a shrinkage penalty for the selection of a subset of TCR characteristics $\mathbf{X}$ containing the most predictive covariates of Y. Penalties are imposed on regression coefficients in order to simultaneously conduct regression model estimation and variable selection [39-43]. The TCR features with nonzero regression coefficients are those features that are the most predictive of the binary response variable Y. iWAS utilizes the adaptive elastic net which includes the adaptive LASSO as a special case [40,43]. Specifically: Let Y denote the binary phenotype as $Y = 1$ or $Y = 0$ (e.g. shared and unshared in this test case). The data are $(\mathbf{X}_i, Y_i)$ for $i = 1, \ldots, n$, where n is the sample size (number of TCR sequences). Each $\mathbf{X}_i$ is a vector of p features (e.g. V/J gene usage, Artemis activity exonuclease activity at each terminus, palindromic nucleotides, n-nucleotide addition, etc). Thus, the design matrix $\mathbf{X}$ is $n \times p$. We consider the logistic model

$$\mathrm{P}(\mathrm{Y}=1|\mathrm{X}=\mathrm{x})=\frac{\exp(\beta_0+x'\beta)}{1+\exp(\beta_0+x'\beta)}$$

where $\beta = (\beta_1, \ldots \beta_p)'$. We define $p(x; \beta 0, \beta) = P(Y = 1|X = x)$. The log-likelihood function is then

$$l(\beta_0,\beta)=-\sum_{i=1}^{n}\log\{1+\exp(\beta_0+x'\beta)\}+\sum_{i=1}^{n}y_i(\beta_0+x'_i\beta).$$

To identify important features and estimate their regression coefficients, we minimize the penalized log likelihood subject to the adaptive elastic net penalty

$$\min_{\beta_0,\beta} -l(\beta_0,\beta)+\lambda_2\|\beta\|_2^2+\lambda_1\sum_{j=1}^{p}w_j|\beta_j|$$

where $\lambda_1, \lambda_2 > 0$ are the regularization parameters, and the penalty function is employed to shrink small coefficients for “unimportant” (non-predictive) TCR characteristics to exactly zero. To implement PLR with adaptive elastic net penalties, we need to select proper values

for the weights w_j and the regularization parameters λ_1 and λ_2, and to select a single regularization parameter λ_1 for our base algorithm the adaptive LASSO. The weights w_j are data dependent and iWAS will utilize the standard maximum likelihood estimates (MLE) for our application to construct the weights [42]. Following the method of Friedman, Hastie, and Tibshirani [44], we construct a sequence of K values of $\lambda \in [\lambda_{min} + \lambda_{min,}]$ on a log scale. A minimization solution for each value of λ_1 is computed and some tuning criteria are applied to select the best tuning parameter. We utilize two approaches to tuning λ_1, Bayesian Information Criterion (BIC) [45,46] and *k*-fold cross-validation. Minimization of the penalized log likelihood requires iterative numerical approximation, and iWAS will utilize Gradient Descent (GC) or Coordinate Descent (CD) to accomplish this task for large datasets, as a result of the parallelizable nature of those algorithms [47,48]. Minimization proceeds until convergence. The iWAS PLR-based algorithm for repertoire signature identification assesses the contribution of key features in predicting outcome (biological group membership). In an alternative embodiment of the iWAS analysis, the regression can be performed as a marginal analysis by regressing on each TCR/Ig marker individually, as in a classical GWAS.

3. Results: iWAS Test Case: Preparing to Map the Tolerant T Cell Receptor Repertoire

Utilizing iWAS, we have evaluated the TCR sequence determinants of publicity (TCR sequence sharing among multiple individuals), a long-standing conundrum in basic immunology. TCR sequences recovered from the peripheral circulation from multiple individuals, especially those with disparate MHC haplotypes, would constitute a group of TCR's that are more readily synthesized by the thymus and more likely capable of traversing the rigors of positive and negative thymic selection as well as persisting through peripheral tolerance mechanisms. Since the size of the potential TCR repertoire is orders of magnitude larger that the actual TCR repertoire in any individual, researchers initially expected that in V(D)J recombination, a process that was assumed to have a stochastic nature, there would be a small likelihood that any two individuals would produce the same “public” TCR (i.e. share a clonotype). However, it is clear that shared clonotypes are common features of TCR repertoires, not rare exceptions. Multiple theories have been developed as an attempt to explain the appearance of shared TCR clones. These theories include the Convergent Recombination Hypothesis (where the likelihood of sequence sharing is proportional to the number of unique VDJ recombinations that could theoretically create a given sequence), the Conformational Hypothesis (where the shape of the peptide presented to the TCR is postulated to drive the selection of shared TCRs), and the Enzyme Proxy Hypothesis (where the likelihood of sequence sharing is related to the regulation and bias of the V(D)J enzymes Rag, Artemis, TdT, etc; which generate the TCR/Ig sequence) [49]. While the conformational model has lost traction, the other two models (CRH and EPH models) have been hypothesized as the cause of TCR sharing in multiple manuscripts [50,51]. However, the relative contribution of these models to the TCR sharing phenomenon has not been tested systematically. One challenge in testing the relative contributions of these models is the computational complexity of modelling all possible VDJ recombinations possible to make each TCR (recombination “paths”) from a repertoire of biologically relevant size. A

second challenge is the statistical analysis of relative contribution of sequence sharing between the number of recombination paths and other sequence features. We investigated the Convergent Recombination Hypothesis to explain sharing of TCR sequences in C57BL/6 mice. One major prediction of the Convergent Recombination Hypothesis is that the frequency of appearance *in vivo* should be proportional to the number of recombination paths. We evaluated > 100,000 peripheral TCR-beta sequences derived from two un-manipulated C57BL/6J mice by Roche 454 next generation sequencing [52] and > 3.34 million double positive thymic TCR-beta sequences from 3 un-manipulated C57BL/6J mice by Illumina HiSeq next generation sequencing [53] (NCBI SRA Accession # SRA026496). We then devised an *in silico* model to calculate the number of recombination paths, for each TCR observed *in vivo*, using a novel GPU parallel processing approach, which interrogated the theoretical TCR repertoire at greater than 10^{14} different recombination paths [52]. We then compared the observed *in vivo* frequency with the recombination path number. As demonstrated in our prior work (Streimer at al), the recombination paths do not correlate well with high *in vivo* frequency [52]. Thus, one prediction of the CR hypothesis is not supported by this analysis. A second prediction of the Convergent Recombination Hypothesis is that the public sequences that are shared between individuals would be those that are “easier to make”, that is they have more theoretical ways to make them (e.i. higher recombination paths). The Enzyme Proxy Hypothesis however, predicts that the public TCR sequences will differ in the extent of activity of V(D)J enzymes (e.g.Rag, Artemis, TdT), which are quantifiable from the TCR sequences. The recombination path number and the enzyme characteristics are readily representable as TCR features in the iWAS analysis. Using the iWAS method, we analyzed the TCR sequence features and recombination path numbers for enrichment in the shared sequences. As shown in Figure 1, the TCR characteristics that represent the V(D)J recombination enzyme activities yielded non-zero logistic regression coefficients (beta) indicating that those enzyme features contain predictive information about the binary response variable for TCR “publicity” (TCR sharing) in this model. Strikingly, the regression coefficient yielded for the recombination path enumeration was a zero coefficient value indicating that the recombination paths (“the number of ways each TCR can theoretically be made”) for each TCR is not predictive of TCR sharing. The results in Figure #1 A and B, are shown for the two completely independent mouse experiments, performed on different tissue (thymic and peripheral T cells), in different laboratories, and sequenced on different platforms (454 vs Illumina). These results support the Enzyme Proxy Hypothesis, whereby a decreased activity (indicated by the negative beta coefficient values) of the V(D)J enzymes appear to correlate with TCR sharing. This suggests that the enzyme proxy model is perhaps is a better fit to the repertoire data than is the convergent recombination model and the results also suggest that the iWAS algorithm is capable of identifying TCR characteristics that correlate with biological outcomes. When the recombination path counting algorithm was restricted to calculate the recombination paths allowed during the condition of zero n-nucleotide addition, a small regression coefficient was obtained, indicating that the convergent recombination hypothesis may be contributory for cases where no n-nucleotides are added at the V(D)J gene junctions (data not shown). In a traditional GWAS study there is what is termed the “large p, small n” problem where the number of predictor parameter (p) is represented by each SNP and the number of individuals (n) in the study is represented by the

number of patients in the affected group and the number of healthy donors in the control group (i.e. the number of individuals in the response variable). Normally the number (p) predictor parameters (SNPs) is very large and the number of individuals in the binary response variable groups is much smaller (n); leading to false positive errors (Type I) due to the large number of comparisons that are performed. To compensate, GWAS studies often control for this false discovery (Bonferroni) which then leads to false negative errors (Type II) which leads to missed associations [54]. Penalized regression analysis have been used in advanced GWAS studies to balance the Type I and II errors [55], and we have adopted this approach in the iWAS in preparation for clinically based iWAS studies which would also be subject to the “large P, small n” problem. However, in this test case the powering of number of individuals (n) in the two arms of the association study is based on binning the TCR sequences into (1) Shared and (2) Unshared groups not based on the number of individual donors in the trial. While calculations to determine the level of powering for a study of this kind will have to await evaluations for the strength of association of immunological phenotypes, the powering in the 3-mouse dataset exceeds 3 million which is currently 10 fold higher than the largest GWAS performed to date [20]. As such, the test case within this study is likely to be sufficiently powered. Nevertheless, we are currently evaluating both convergent recombination and enzyme proxy models that potentially explain TCR sharing using larger data sets (powered with more individuals and greater sequencing depth per individual) of both murine and human TCR-beta sequences.

4. Discussion: How can iWAS be used?

The major strength of GWAS studies is that many genetic loc can be simultaneously interrogated. Similarly, the major strength of an iWAS is the ability to evaluate very large numbers of immune receptors and meta-data parameters (clinical, biological, or experimental) in parallel While in the past only a small number of parameters could be addressed at once, now with the combination of very large numbers of sequences and sequence features, more power is added to evaluate traits that depend on adaptive immunity. The features that can be used can vary, but include not only the sequences themselves, but the length of the CD3 region, its composition, the length of the nontemplated regions, as well as the V and J gene usage.

4.1 What can be examined using the iWAS approach?

iWAS and Infection/Vaccination—The clear choice for utilizing iWAS will be in the investigation of diseases liked to immunity. These include infectious diseases such as B- and T-cell responses to influenza, West Nile Virus, HIV and Hepatitis C (along with other infectious diseases), as well as vaccination responses. In the example of Hepatitis C, it will be important to stratify the patients by their outcomes; those who clear virus without treatment, those that require treatment and finally those who go on to develop cirrhosis and ultimately hepatocellular carcinoma. Identification of the T cell receptor features associated with effective immunity could allow clinicians to predict the outcomes of disease early in the disease course, and tailor treatments appropriately. In addition, the identification of Ig/TCR repertoire features that associate with protective or unsuccessful vaccination

responses may aid in the development of novel vaccine approaches or adjuvant optimization strategies.

iWAS and Autoimmunity—Autoimmune disorders may also be amenable to iWAS analysis. Candidates with well-defined targets and strong HLA association such as Type 1 diabetes, ankylosing spondylitis, narcolepsy, and psoriasis are optimal analytical targets. In addition, antibody mediated diseases like Hashimoto's thyroiditis and Myasthenia Gravis can have both B and T cell repertoire interrogated. T cell repertoires may be important even for antibody mediated diseases, where T cell help is required for the B cell responses, and in some cases T cells also likely contribute to the disease process. The examination need not be limited to the antigen specific repertoire. iWAS allows users to expand their examinations to evaluate the naïve repertoire of longitudinal samples. Another disease where iWAS would be of value would be in children who contracted severe RSV infections early in life and who have a very high relative risk of later developing asthma. This could potentially stratify the risk of developing severe asthma from those children who will remain non-asthmatic. It will also be important to evaluate if iWAS analysis would be informative for analyzing patients whose disease pathology has a less clear-cut immune etiology in the clinical setting, as in Lyme disease and coccidiomycosis.

iWAS and Transplantation—Transplantation-associated outcomes are particularly relevant, as iWAS may be leveraged to identify differential outcomes both in solid and liquid tumors. The specific clinical applications include, but are not limited to, evaluating the extent of irradiation during immune-ablation, judging post-transplant immunosuppression regimen, monitoring infectious sequela (e.g. CMV reactivation, bacterial sepsis, etc), screening for rejection and graft-vs-host disease, and examining stem cell reconstitution (allogeneic or autologous) [56].

iWAS and Cancer—Cancer represents an important target as well. Many tumors are immunogenic, and alteration of the immune responses by cytokine interventions or immune checkpoint inhibition (modulation of PD-1 and/or CTLA4) have been useful in multiple tumor types [57]. In some of those cancers, responses that are inhibited by regulatory T-cells (Treg) would be important to evaluate both the effector population (to identify potential candidates for T-cell therapies) as well as the regulatory population. iWAS offers the ability to evaluate this area of immuno-cancer surveillance by analyzing the characteristics of the adaptive immune response to tumors that predict successful response or resistance to immunotherapy.

Lympoid tumors (lymphoma, leukemia, and myeloma) are cancers of adaptive immune cells and are an important special case. Not only can lymphoid tumors themselves be immunogenic and thus targeted by the remaining adaptive immune system (hence leveraged for potential immunotherapy just like solid tumors), but lymphoid tumors also harbor the unique V(D)J recombinant receptor that can serve as a truly unique tumor marker. Both T and B cell derived tumors will express a unique immune receptor that can be readily monitored by high throughput sequencing techniques. The sensitivity of detection by sequencing is many times higher than conventional cytopathology and so minimal residual disease can be detected, even in the presence of large numbers of normal lymphocytes.

5. Conclusion: How does this fit with Ray Owen and tolerance?

Ray was always a fan of new technology. He embraced protein sequencing and was a big supporter of mathematical analysis of complex data, no doubt stemming from his own background as an agricultural geneticist. The greatest excitement from the iWAS is not merely that we can identify biomarkers of immunological processes, like the development of tolerance, but that these immune biomarkers immediately suggest a way to test mechanism of biological processes. Thus, at the dawn of the era of immuno-genetics and immuno-informatics of the adaptive immune system, investigators now have new tools to suggest how an important immunological phenotype can be related to function and then how it can be manipulated.

Acknowledgments

Antigen receptor research for an author in this study (LC) was supported by the NIH NIAID (1R01AI097403-01, 5R01AI097403-04).

References

1. Owen RD. Immunogenetic Consequences of Vascular Anastomoses Between Bovine Twins. Science. 1945; 102:400–401.10.1126/science.102.2651.400 [PubMed: 17755278]
2. Paul, WE. Fundamental Immunology. Lippincott Williams & Wilkins; 2012.
3. Candéias S, Waltzinger C, Benoist C, Mathis D. The V beta 17+ T cell repertoire: skewed J beta usage after thymic selection; dissimilar CDR3s in CD4+ versus CD8+ cells. J Exp Med. 1991; 174:989–1000. [PubMed: 1940807]
4. Correia-Neves M, Waltzinger C, Mathis D, Benoist C. The shaping of the T cell repertoire. Immunity. 2001; 14:21–32. [PubMed: 11163227]
5. Biozzi G, Stiffel C, Mouton D, Bouthillier Y. [Genetic regulation of immunoglobulin synthesis during immune response]. - PubMed - NCBI. Annales d' …. 1974
6. O A, B B, Levine Bernard B. Studies on Artificial Antigens : III. The Genetic Control of the Immune Response to Hapten-Poly-L-Lysine Conjugates in Guinea Pigs. J Exp Med. 1963; 118:953. [PubMed: 14112274]
7. Ellman L, Green I, Martin WJ, Benacerraf B. Linkage between the Poly-L-Lysine Gene and the Locus Controlling the Major Histocompatibility Antigens in Strain 2 Guinea Pigs. Proceedings of the …. 1970
8. Sommer S. The importance of immune gene variability (MHC) in evolutionary ecology and conservation. Front Zool. 2005; 2:16.10.1186/1742-9994-2-16 [PubMed: 16242022]
9. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE. The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res. 2015; 43:D423–D431.10.1093/nar/gku1161 [PubMed: 25414341]
10. Simmonds M, Gough S. The HLA Region and Autoimmune Disease: Associations and Mechanisms of Action. Cg. 2007; 8:453–465.10.2174/138920207783591690
11. Snell GD. Histocompatibility Genes of the Mouse II Production and Analysis of Isogenic Resistant Lines2. 1958
12. Snell GD. Histocompatibility Genes of the Mouse I Demonstration of Weak Histocompatibility Differences by Immunization and Controlled Tumor Dosage. 1958
13. Noble JA, Valdes AM. Genetics of the HLA Region in the Prediction of Type 1 Diabetes. Curr Diab Rep. 2011; 11:533–542.10.1007/s11892-011-0223-x [PubMed: 21912932]
14. Noble JA, Valdes AM, Varney MD, Carlson JA, Moonsamy P, Fear AL, et al. HLA Class I and Genetic Susceptibility to Type 1 Diabetes: Results From the Type 1 Diabetes Genetics Consortium. Diabetes. 2010; 59:2972–2979.10.2337/db10-0699 [PubMed: 20798335]

15. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2013; 42:D1001–D1006.10.1093/nar/gkt1229 [PubMed: 24316577]
16. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014; 95:5–23.10.1016/j.ajhg.2014.06.009 [PubMed: 24995866]
17. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010; 467:832–838.10.1038/nature09410 [PubMed: 20881960]
18. Nakaoka H, Inoue I. Meta-analysis of genetic association studies: methodologies, between-study heterogeneity and winner's curse. J Hum Genet. 2009; 54:615–623.10.1038/jhg.2009.95 [PubMed: 19851339]
19. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005; 308:385–389.10.1126/science.1109557 [PubMed: 15761122]
20. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015; 518:197–206.10.1038/nature14177 [PubMed: 25673413]
21. Visscher PM, Brown MA, McCarthy MI. Five Years of GWAS Discovery. J Hum Gene. 2012
22. Ricaño-Ponce I, Wijmenga C. Mapping of Immune-Mediated Disease Genes. Annu Rev Genom Human Genet. 2013; 14:325–353.10.1146/annurev-genom-091212-153450
23. Michel M, Schmidt MJ, Mirnics K. Immune system gene dysregulation in autism and schizophrenia. Devel Neurobio. 2012; 72:1277–1287.10.1002/dneu.22044
24. O'Brien RP, Phelan PJ, Conroy J, O'Kelly P, Green A, Keogan M, et al. A genome-wide association study of recipient genotype and medium-term kidney allograft function. Clin Transplant. 2013; 27:379–387.10.1111/ctr.12093 [PubMed: 23432519]
25. Ozel AB, Moroi SE, Reed DM, Nika M, Schmidt CM, et al. NEIGHBOR Consortium. Genome-wide association study and meta-analysis of intraocular pressure. Hum Genet. 2013; 133:41–57.10.1007/s00439-013-1349-5 [PubMed: 24002674]
26. Han F, Faraco J, Dong XS, Ollila HM, Lin L, Li J, et al. Genome Wide Analysis of Narcolepsy in China Implicates Novel Immune Loci and Reveals Changes in Association Prior to Versus After the 2009 H1N1 Influenza Pandemic. PLoS Genet. 2013; 9:e1003880.10.1371/journal.pgen.1003880 [PubMed: 24204295]
27. Deng X, Sabino EC, Cunha-Neto E, Ribeiro AL, Ianni B, Mady C, et al. Genome Wide Association Study (GWAS) of Chagas Cardiomyopathy in Trypanosoma cruzi Seropositive Subjects. PLoS ONE. 2013; 8:e79629.10.1371/journal.pone.0079629 [PubMed: 24324551]
28. Weinhold N, Johnson DC, Chubb D, Chen B, Försti A, Hosking FJ, et al. The CCND1 c.870G>A polymorphism is a risk factor for t(11;14)(q13;q32) multiple myeloma. Nat Genet. 2013; 45:522–525.10.1038/ng.2583 [PubMed: 23502783]
29. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 2013; 45:1452–1458.10.1038/ng.2802 [PubMed: 24162737]
30. Tsai FJ, Lee YC, Chang JS, Huang LM, Huang FY, Chiu NC, et al. Identification of Novel Susceptibility Loci for Kawasaki Disease in a Han Chinese Population by a Genome-Wide Association Study. PLoS ONE. 2011; 6:e16853.10.1371/journal.pone.0016853 [PubMed: 21326860]
31. Jackson KJL, Liu Y, Roskin KM, Glanville J, Hoh RA, Seo K, et al. Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. Cell Host Microbe. 2014; 16:105–114.10.1016/j.chom.2014.05.013 [PubMed: 24981332]
32. Trepel F. Zahl und Verteilung der Lymphocyten des Menschen. Eine kritische Analyse. Klin Wochenschr. 1974; 52:511–515.10.1007/BF01468720 [PubMed: 4853392]
33. Weinstein JA, Jiang N, White RA, Fisher DS, Quake SR. High-throughput sequencing of the zebrafish antibody repertoire. Science. 2009; 324:807–810.10.1126/science.1170020 [PubMed: 19423829]

34. Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, et al. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. Sci Transl Med. 2009; 1:12ra23.
35. Freeman JD, Warren RL, Webb JR, Nelson BH, Holt RA. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. Genome Res. 2009; 19:1817–1824.10.1101/gr.092924.109 [PubMed: 19541912]
36. Warren RL, Nelson BH, Holt RA. Profiling model T-cell metagenomes with short reads. Bioinformatics. 2009; 25:458–464.10.1093/bioinformatics/btp010 [PubMed: 19136549]
37. Benichou J, Ben-Hamo R, Louzoun Y, Efroni S. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. Immunology. 2012; 135:183–191.10.1111/j.1365-2567.2011.03527.x [PubMed: 22043864]
38. Stern JNH, Yaari G, Vander Heiden JA, Church G, Donahue WF, Hintzen RQ, et al. B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. Sci Transl Med. 2014; 6:248ra107.10.1126/scitranslmed.3008879
39. Tibshirani R. Regression Shrinkage and Selection via the Lasso on JSTOR. … The Royal Statistical Society Series B (Methodological). 1996
40. Zou H, Hastie T. Regularization and variable selection via the elastic net. … Royal Statistical Society: Series B (Statistical …. 2005
41. Zou H. The Adaptive Lasso and Its Oracle Properties. Journal of the American Statistical Association. 2006; 101:1418–1429.10.1198/016214506000000735
42. Zhang HH, Lu W. Adaptive Lasso for Cox's proportional hazards model. Biometrika. 2007; 94:691–703.10.1093/biomet/asm037
43. Zou H, Zhang HH. On the adaptive elastic-net with a diverging number of parameters. Ann Statist. 2009; 37:1733–1751.10.1214/08-AOS625
44. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010; 33:1–22. [PubMed: 20808728]
45. Schwarz G. Estimating the Dimension of a Model. Ann Statist. 1978; 6:461–464.
46. Chen J, Chen Z. Extended Bayesian information criteria for model selection with large model spaces. Biometrika. 2008; 95:759–771.10.1093/biomet/asn034
47. Zhang, T. Icml '04. ACM Press; New York, New York, USA: 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms; p. 116
48. Friedman J, Hastie T, Höfling H, Tibshirani R. Pathwise coordinate optimization. The Annals of Applied Statistics. 2007; 1:302–332.
49. Venturi V, Price DA, Douek DC, Davenport MP. The molecular basis for public T-cell responses? Nat Rev Immunol. 2008; 8:231–238.10.1038/nri2260 [PubMed: 18301425]
50. Gauss GH, Lieber MR. Mechanistic constraints on diversity in human V(D)J recombination. Mol Cell Biol. 1996; 16:258–269. [PubMed: 8524303]
51. Venturi V, Kedzierska K, Price DA, Doherty PC, Douek DC, Turner SJ, et al. Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. Proc Natl Acad Sci USa. 2006; 103:18691–18696.10.1073/pnas.0608907103 [PubMed: 17130450]
52. Striemer G, Krovi H, Akoglu A, Vincent B. IEEE Xplore Abstract - Overcoming the Limitations Posed by TCR-beta Repertoire Modeling through a GPU-Based In-Silico DNA R. … Symposium. 2014
53. Li H, Ye C, Ji G, Wu X, Xiang Z, Li Y, et al. Recombinatorial biases and convergent recombination determine interindividual TCRβ sharing in murine thymocytes. J Immunol. 2012; 189:2404–2413.10.4049/jimmunol.1102087 [PubMed: 22826324]
54. Xie J, Cai TT, Maris J, Li H. False Discovery Rate Control For High Dimensional Dependent Data With an Application to Large-Scale Genetic Association …. 0AD.
55. Cho S, Kim K, Kim YJ, Lee JK, Cho YS, Lee JY, et al. Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis. Ann Hum Genet. 2010; 74:416–428.10.1111/j.1469-1809.2010.00597.x [PubMed: 20642809]

56. Yew PY, Alachkar H, Yamaguchi R, Kiyotani K, Fang H, Yap KL, et al. Quantitative characterization of T-cell repertoire in allogeneic hematopoietic stem cell transplant recipients. Bone Marrow Transplant. 2015; 50:1227–1234.10.1038/bmt.2015.133 [PubMed: 26052909]

57. Topalian SL, Hodi FS, Brahmer JR, Gettinger SN, Smith DC, McDermott DF, et al. Safety, Activity, and Immune Correlates of Anti–PD-1 Antibody in Cancer. N Engl J Med. 2012; 366:2443–2454.10.1056/NEJMoa1200690 [PubMed: 22658127]

Highlights

- Adaptive Immune Receptors (Ig/TCR) Genetics rarely assessed in Genome Wide Association Studies
- immunome Wide Association Studies (iWAS) of Ig/TCR data can reveal immune genetic contributions
- iWAS shows Enzyme Proxy Model correlates with TCR sharing

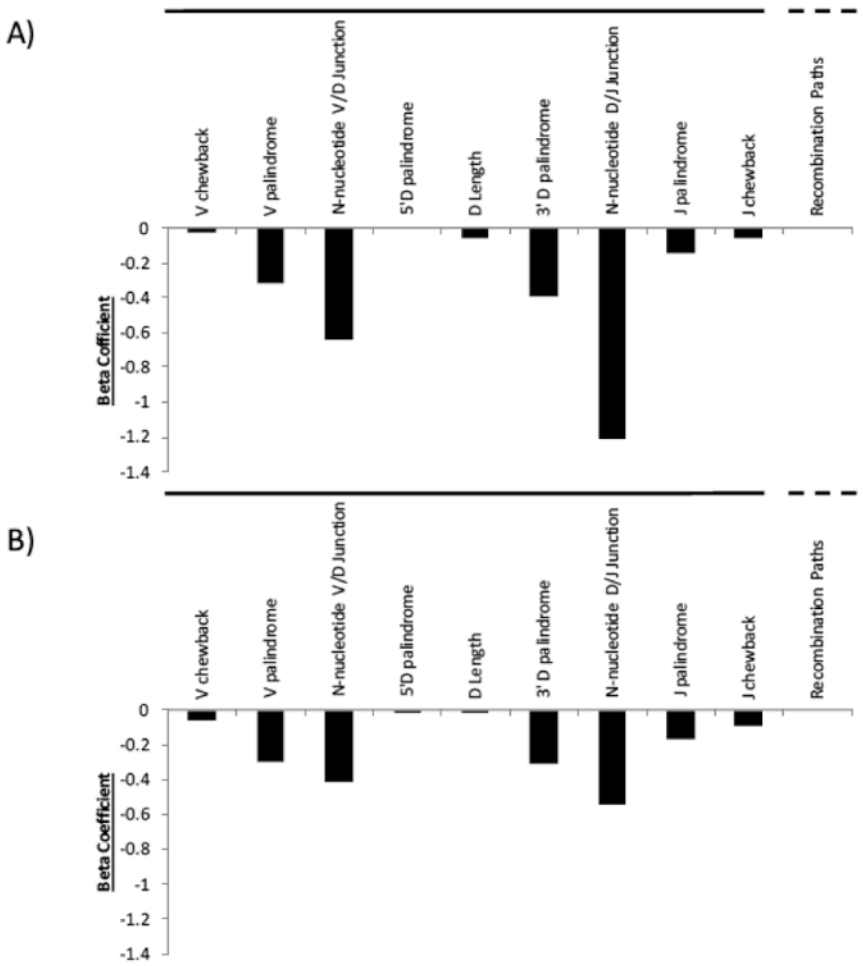

Figure 1. TCR Sharing: iWAS Regression Coefficients

The recombination path numbers resulted in a zero regression coefficient (beta), however the enzyme correlates of the TCR characteristics yielded non-zero coefficients in both datasets. A)2 mouse dataset of CD8+ peripheral T cells sequenced with a Roche 454 instrument, B)3 mouse dataset of double positive (CD4+CD8+) thymocytes sequenced with an Illumina Hiseq2000 instrument (NCBI SRA Accession # SRA026496). The solid line above the graph indicates the predictor variables related to the enzyme correlates, whereas the dotted line indicates the convergent recombination path variable.