Cellular Immunology xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Cellular Immunology

journal homepage: www.elsevier.com/locate/ycimm

Research paper

iWAS – A novel approach to analyzing Next Generation Sequence data for immunology

Benjamin Vincent [a,1], Adam Buntzman [b,*,1], Benjamin Hopson [c,d,2], Chris McEwen [d], Lindsay Cowell [e], Ali Akoglu [f], Helen Zhang [g], Jeffrey Frelinger [b]

[a] Department of Medicine, University of North Carolina, Chapel Hill, NC 2714, United States
[b] Department of Immunobiology, University of Arizona, Tucson, AZ 85724, United States
[c] School of Engineering, The University of Edinburgh, Edinburgh EH9 3JL, UK
[d] Cambridge Consultants, Science Park, Milton Rd, Cambridge CB4 0DW, UK
[e] Division of Biomedical Informatics, Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, TX 75390, United States
[f] Department of Electrical and Computer Engineering, University of Arizona, Tucson, AZ 85721, United States
[g] Department of Mathematics, University of Arizona, Tucson, AZ 85721, United States

ARTICLE INFO

Article history:
Received 1 October 2015
Revised 28 October 2015
Accepted 28 October 2015
Available online xxxx

Keywords:
Immune receptors
Repertoire
V(D)J recombination
Association study

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.

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

* Corresponding author at: Department of Medicine, University of Arizona, Tucson, AZ 85724, United States.
E-mail addresses: benjamin.g.vincent@gmail.com (B. Vincent), buntzman@deptofmed.arizona.edu (A. Buntzman), ben.hopson@cambridgeconsultants.com (B. Hopson), chris.mcewan@cambridgeconsultants.com (C. McEwen), Lindsay.Cowell@utsouthwestern.edu (L. Cowell), akoglu@email.arizona.edu (A. Akoglu), hzhang@math.arizona.edu (H. Zhang), jfrelin@email.arizona.edu (J. Frelinger).
[1] These authors contributed equally to this work.
[2] Present address: Cambridge Consultants.

http://dx.doi.org/10.1016/j.cellimm.2015.10.012
0008-8749/

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, Crohn's disease, Multiple Sclerosis, Behcet's disease, Vitiligo, ankylosing spondylitis, psoriasis, Graves disease, eczema 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)

Please cite this article in press as: B. Vincent et al., iWAS – A novel approach to analyzing Next Generation Sequence data for immunology, Cell. Immunol. (2015), http://dx.doi.org/10.1016/j.cellimm.2015.10.012