Genome-wide Analysis of Immune System Genes by Expressed Sequence Tag Profiling

Cosmas C. Giallourakis, Yair Benita, Benoit Molinie, Zhifang Cao, Orion Despo, Henry E. Pratt, Lawrence R. Zukerberg, Mark J. Daly, John D. Rioux and Ramnik J. Xavier

*J Immunol* published online 24 April 2013
http://www.jimmunol.org/content/early/2013/04/23/jimmunol.1203471

Supplementary Material
http://www.jimmunol.org/content/suppl/2013/04/24/jimmunol.1203471.DC1

Subscription
Information about subscribing to *The Journal of Immunology* is online at:
http://jimmunol.org/subscription

Permissions
Submit copyright permission requests at:
http://www.aai.org/About/Publications/JI/copyright.html

Email Alerts
Receive free email-alerts when new articles cite this article. Sign up at:
http://jimmunol.org/alerts
Genome-wide Analysis of Immune System Genes by Expressed Sequence Tag Profiling

Cosmas C. Giallourakis, Yair Benita, Benoit Molinie, Zhifang Cao, Orion Despo, Henry E. Pratt, Lawrence R. Zukerberg, Mark J. Daly, John D. Rioux, and Rammik J. Xavier

Profiling studies of mRNA and microRNA, particularly microarray-based studies, have been extensively used to create compendia of genes that are preferentially expressed in the immune system. In some instances, functional studies have been subsequently pursued. Recent efforts such as the Encyclopedia of DNA Elements have demonstrated the benefit of coupling RNA sequencing analysis with information from expressed sequence tags (ESTs) for transcriptomic analysis. However, the full characterization and identification of transcripts that function as modulators of human immune responses remains incomplete. In this study, we demonstrate that an integrated analysis of human ESTs provides a robust platform to identify the immune transcriptome. Beyond recovering a reference set of immune-enriched genes and providing large-scale cross-validation of previous microarray studies, we discovered hundreds of novel genes preferentially expressed in the immune system, including noncoding RNAs. As a result, we have established the Immunogene database, representing an integrated EST road map of gene expression in human immune cells, which can be used to further investigate the function of coding and noncoding genes in the immune system. Using this approach, we have uncovered a unique metabolic gene signature of human macrophages and identified PRDM15 as a novel overexpressed gene in human lymphomas. Thus, we demonstrate the utility of EST profiling as a basis for further deconstruction of physiologic and pathologic immune processes. The Journal of Immunology, 2013, 190: 000–000.

The immune system is our central defense against bacterial and viral pathogens. As such, misregulation of certain genes that are critical for normal immune function can result in immunodeficiency and autoimmune disorders. Furthermore, disorders such as diabetes, obesity, and cancer are increasingly viewed as having roots in immunological dysfunction (1). Despite this recognition of the immune system as a central player in health and disease, the full spectrum of genes and pathways capable of modulating immune responses is not known.

Analyzing the function of genes that are highly enriched or specific for immune cell/tissue expression may aid in capturing important regulators of immune function. It is important to note that this premise does not require that ubiquitously expressed genes cannot be critically important in immune function and development. However, considering examples such as BCR/TCR signaling cascade components, it is striking that genes that are expressed specifically in immune cells/tissues are often central players in immune physiology. In addition, genes responsible for Mendelian immunodeficiencies or autoimmunity often show enriched expression patterns in the immune system. Examples include the causes of Omenn syndrome (RAG1/2; Online Mendelian Inheritance in Man [OMIM] identification number 603554), hyper-IgM syndrome (AICDA; OMIM identification number 308230), and X-linked lymphoproliferative disorder (SH2D1A; OMIM identification number 308240).

Historically, gene discovery in the immune system has relied on unbiased searches for genes exhibiting differential expression in cells/tissues of the immune system as compared with other cells/tissues. These approaches have used techniques such as differential display, microarray profiling, serial analysis of gene expression, and, most recently, RNA sequencing (RNA-Seq) (2–11). In the case of microarray profiling, experiments often reveal a large number of differentially expressed genes, making cross-validation by alternative approaches a challenge. In addition, microarrays suffer from difficulty in detecting low-abundance transcripts and may not contain valid probes for mRNAs, noncoding RNAs (ncRNAs), or microRNAs, because their design is based on human annotation. Recent studies have used RNA-Seq data to examine expression profiles of immune cells; however, this has not diminished the utility of expressed sequence tag (EST) data, as recently demonstrated by the Encyclopedia of DNA Elements (ENCODE).
To identify potential divergent ncRNAs emanating near the TSSs of immune-enriched protein-coding genes, we downloaded EST data from the University of California Santa Cruz genome browser. Using custom PERL scripts, we analyzed the data to identify sequences that met the following requirements: 1) at least one EST is spliced and transcribed on the opposite strand of a coding gene and has a start site within 5 kb of the TSS of the RefSeq Immunegene; 2) the spliced EST does not show evidence of splicing to a neighboring RefSeq gene; and 3) no open reading frame of >100 aa could be identified in the longest EST associated with the coding RefSeq, if there was more than one EST that met our criteria. We used the data feature “intronOrientation” downloaded from the University of California Santa Cruz Genome Browser to confirm EST strand orientation.

**Gene ontology and Kyoto Encyclopedia of Genes and Genomes pathway functional enrichment analysis**

Enrichment analyses for gene ontology (GO) biological processes and Kyoto Encyclopedia of Genes and Genomes (KEOG) pathways were conducted for various gene list categories including the EST-defined Immuno- nogenes with Entrez IDs (n = 2,232) and the reference set of genes (n = 570) documented in the literature as being preferentially expressed in the immune system. These gene lists were submitted to the Database for Annotation, Visualization and Integrated Discovery (http://david.abcc.ncifcrf.gov) for enrichment analyses of GO biological processes and KEGG pathways, with the number of genes in each category and Bonferroni-corrected p values displayed (19).

**Microarray analysis**

GC robust multirarray average—normalized gene expression profiles across 79 tissues were obtained from the Genomics Institute of the Novartis Research Foundation (GNF) consortium (20). Expression profiles from the Affymetrix U133A platform (Affymetrix) and GNF custom probes were used. In accordance with Affymetrix guidelines, probes with highest expression value below log2(100) across all tissues were removed. Expression profiles were clustered using Cluster 3 (21) and visualized using JavaTreeView (22). Immune enrichment was calculated with the R program (version 2.5.1) using the Wilcoxon rank-sum test for each probe, and p values were corrected using the FDR method (18). In cases in which multiple probes were mapped to the same gene, the most significant p value was accepted. The following tissues in the GNF dataset were classified as immune and tested versus all other tissues: bone marrow, CD19 B cells, tonsils, lymph nodes, thymus, CD4 T cells, CD8 T cells, CD56 T cells, whole blood, CD33 myeloid cells, CD14 monocytes, dendritic cells, fetal liver, CD105 endothelial cells, leukemia cell lines, lymphoma cell lines, and erythroid cells. For Fig. 5E, we used the normalized log2 median-centered expression of PRDM15 and several coregulated genes obtained from published microarray expression profiling (23).

**Human-mouse homolog genes**

For each human protein, the homolog mouse protein was identified using NCBI HomoloGene (24). In some cases, multiple mouse proteins matched a single human protein and vice versa. For the purpose of immune enrichment, at least one mouse homolog was required to be immune enriched.

**Cells, plasmids, and Abs**

Human PRDM15 (GenBank identification number BC067102) was obtained from Open Biosystems and subcloned into N-terminal and C-terminal FLAG-tagged pCMV-3XFLAG vectors using 5' EcoRI and 3' HindIII; vectors were generated by modifying ClonTech pCMV-Myc vector by introduction of appropriate epitope tags and modification of the multiple cloning site. Primary Abs used were anti-PRDM15 (Santa Cruz Biotechnology; sc-83314), anti-FLAG M2 (Sigma-Aldrich; F1804), and anti-actin (Sigma-Aldrich; F2066). Secondary Abs for ECL development were HRP-conjugated anti-mouse (Amersham Biosciences; NA931) and anti-rabbit (Amersham Biosciences; NA934).

**Western blotting, immunofluorescence, and immunohistochemistry**

Human Cell Line Blot II (ProSci; 1502) was obtained containing 15 μg protein from various human cell lines. Anti-PRDM15 Western blotting was performed under standard procedures with 1:1000 primary Ab and 1:3000 secondary Ab. For immunofluorescence, HEK293T cells were transfected with Lipofectamine (Invitrogen) with PRDM15 FLAG-tagged constructs and stained with either anti-FLAG (1:400) or anti-PRDM15 (1:200) Abs. Immunohistochemistry on human tonsil or follicular lymphomas was performed from the Pathology Core at Massachusetts General Hospital was performed as described using anti-PRDM15 Ab (1:100) (25).
RNA-Seq analysis

Reads per kilobase of exon model per million mapped reads calls from an RNA-Seq atlas were downloaded from Medicalgenomics (http://medicalgenomics.org/ma_seq_atlas), which encompases ribodepleted RNA-Seq data from 11 different tissues (colon, heart, hypothalamus, kidney, liver, lung, skeletal muscle, spleen, testes, ovary, and adipose tissue). Immune enrichment was calculated for each gene using the Wilcoxon signed-rank test. In the RNA-Seq atlas, spleen was classified as an immune tissue and tested versus all other tissues. In this dataset, 21,725 RefSeq isoforms for mRNAs or ncRNAs were designated by NM or NR accessions, encompassing 14,713 unique gene symbols (10). Of the 26,552 UniGene clusters analyzed for their WIP score, there were RNA-Seq data for 11,449 corresponding genes. Furthermore, of the 3352 Immunogene clusters identified in the UniGene database (see above analysis), there were RNA-Seq data for 1336 corresponding genes. The ρ-, W-, and Z-score statistics were calculated for all 11,449 genes versus the 1,336 Immunogenes to test for the presence of a trend toward immune enrichment based on RNA-Seq data.

Human disease single nucleotide polymorphism analysis

The genome-wide association study (GWAS) single nucleotide polymorphism (SNP) catalog was downloaded from http://www.genome.gov/gwastudies/. SNPs were considered immune disease/trait-associated if they associated with the following diseases/trait: atopic dermatitis, systemic lupus erythematosus, Behcet’s disease, primary biliary cirrhosis, rheumatoid arthritis, psoriasis, psoriatic arthritis, vitiligo, sarcoidosis, inflammatory bowel disease, Crohn’s disease, ulcerative colitis, primary sclerosing cholangitis, type 1 diabetes, type 1 diabetes autoantibodies, multiple sclerosis (MS), ankylosing spondylitis, juvenile idiopathic arthritis, IgE levels, IgA levels, allergic rhinitis, leprosy, IgE grass sensitization, CMV Ab response, Graves disease, or celiac disease. These diseases/trait-associations were identified with 1102 unique entries in the catalog, encompassing 808 unique immune disease/trait-associated SNPs giving overlapping identification of SNPs among diseases and studies. The remaining diseases/trait-associations were classified as nonimmune and were associated with 8795 entries encompassing 7355 unique SNPs. We next calculated the distance from each immune disease/trait-associated SNP, as well as each nonimmune disease/trait-associated SNP, to the nearest TSS of a coding RefSeq. We then calculated the frequency of immune disease/trait-associated SNPs within binned distances from EST-defined Immunogenes versus the frequency of nonimmune disease/trait-associated SNPs within the same binned distances from EST-defined Immunogenes (Fig. 4C). We computed a p value by permutation testing, selecting 808 SNPs at random from the non-immune disease/trait-associated SNP pool 10,000 times and calculating the percentage of SNPs with a nearest immune or nonimmune tissues (Supplemental Fig. 1A).

Gold-standard immune-enriched database annotation

We compiled a list of genes documented in the literature to be preferentially expressed in the immune system. We defined evidence for immune enrichment based on RT-PCR or Northern blots. To remove selection biases as much as possible, we cataloged the expression pattern of sets of genes that would be generally regarded as preferentially expressed in the immune system (e.g., cluster of differentiation Ags, cytokine genes and receptors, and TLR signaling components) using various Web-based resources that catalog these genes. We consider this very stringent approach necessary, because not all TLR genes (e.g., TLR5) or CD Ags (e.g., CD56) are enriched in the immune system.

RT-PCR analysis

First-strand cDNA from human spleen was obtained from OriGene. Reactions were performed in 25 μl volumes with 40 ng first-strand cDNA and 640 nM concentrations of both forward and reverse primers. PCR was performed under the following conditions: initial denaturation of 95°C for 5 min followed by 34 cycles of 95°C for 30 s and 60°C for 30 s. To control for specificity of PCR reactions and primer-dimers, the same PCR reaction was also conducted with no input cDNA. All primers were designed to be exon spanning. PCR products were cloned and sequenced to verify amplification of expected target long ncRNA (IncRNA). A complete list of UniGene clusters and associated PCR primers used to amplify spliced ncRNAs is shown in Table II. For quantitative RT-PCR experiments, total RNA was extracted from the indicated cell lines; the RT step was performed with Superscript, and the quantitative PCR (qPCR) step was performed with SYBR Green (Invitrogen).

Results

EST profiling is a robust method to measure mRNA immune enrichment

To identify immune-enriched genes using EST datasets, we developed a quantitative schema outlined in Fig. 1. We use the term “immune-enriched” to indicate higher mRNA transcript levels in immune tissues compared with nonimmune tissues.

The NCBI Human UniGene database (build 202) contained 6,731,038 ESTs derived from 8,648 human cDNA libraries. These ESTs were clustered by sequence alignment with no genomic scaffold into 171,154 clusters. Of the resulting clusters, we focused on the 26,552 that had at least 10 ESTs, encompassing a total of 6,049,575 ESTs. The UniGene tissue library descriptors were annotated and standardized, resulting in 469 tissues/cell types. Of these tissues/cell types, 91 (~20%) were classified as immune tissues, with the remaining defined as nonimmune tissues. Because each EST in a particular UniGene cluster has an associated library identification number, we were able to create a binary classification scheme, categorizing each EST as to whether it derived from immune or nonimmune tissues (Supplemental Fig. 1A).

Table II. Complete list of UniGene clusters and associated PCR primers used to amplify spliced ncRNAs

<table>
<thead>
<tr>
<th>UniGene Cluster (Primer Name)</th>
<th>Sequence</th>
<th>Product (bp)</th>
<th>EST Accession No.</th>
<th>NR Accession No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hs.662020 (ncRNA4)</td>
<td>F: 5'-GAGGAGATTATGCTCAAGACA-3'</td>
<td>346</td>
<td>AK097922</td>
<td>NR_045486.1</td>
</tr>
<tr>
<td></td>
<td>R: 5'-TGCGGCAACAGGATGCTAAT-3'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.690944 (ncRNA6)</td>
<td>F: 5'-CACCTCATTACGTGCTTCT-3'</td>
<td>353</td>
<td>DA953527</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R: 5'-GCGAAATGTGCGGAAAGA-3'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.439791 (ncRNA12)</td>
<td>F: 5'-TGCTCCACCACCTTTGTCAG-3'</td>
<td>823</td>
<td>CR598049</td>
<td>NR_024464.1</td>
</tr>
<tr>
<td></td>
<td>R: 5'-CGAGTAGACTGCTGAAAGA-3'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.667175 (ncRNA16)</td>
<td>F: 5'-CCCCAAGCAGGAAATGGTPT-3'</td>
<td>275</td>
<td>AL832284</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R: 5'-CGTTCGCCCGCCGAGCTTCT-3'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.657722 (ncRNA17)</td>
<td>F: 5'-CCTCTTCACCTCCCTTAT-3'</td>
<td>399</td>
<td>CD367998</td>
<td>NR_026544.1</td>
</tr>
<tr>
<td></td>
<td>R: 5'-TGTTAGGTTATGCGGAAAGA-3'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.130180 (ncRNA18)</td>
<td>F: 5'-GCCATCTGATCAGTCCGAA-3'</td>
<td>332</td>
<td>DA937565</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R: 5'-AGGATGTGTGCGGCCATGACA-3'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.672896 (ncRNA20)</td>
<td>F: 5'-AGGATGGGACACCTGCTCAAA-3'</td>
<td>366</td>
<td>DA468624</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R: 5'-GGTCTCCACCCATCTTTGAC-3'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(AS-LRRK2)</td>
<td>F: 5'-GATCGGAGACCTGCTCAAA-3'</td>
<td>158</td>
<td>DA673065</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R: 5'-ATACGGCTCCCTCCGAGAC-3'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(LRRK2)</td>
<td>F: 5'-TGGAGAAGTGATGCGGAAAGA-3'</td>
<td>128</td>
<td>NM_198578.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R: 5'-TGGAAGATGATGCGGAAAGA-3'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(PRDM15)</td>
<td>F: 5'-GCCCTAAGGAGACGTCTGGC-3'</td>
<td>186</td>
<td>NM_022115.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R: 5'-AGCTGTAATGGCGGCTAGG-3'</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

F: Forward; R, reverse.
Next, we introduced a scoring system to quantify the level of immune expression for each of the 26,552 clusters. Specifically, we calculated a WIP score for each UniGene cluster which reflects the normalized percent of ESTs derived from immune tissues within each cluster. To assess whether the WIP score for each UniGene cluster was significantly enriched in immune expression, we applied a χ² statistic and corrected for multiple hypothesis testing using the FDR method (see Materials and Methods and Table I). The distribution of the WIP score (mean 0.095) and p values for all 26,552 UniGene clusters are shown in Fig. 2A and Supplemental Table I.

To evaluate whether our EST immune-based expression profiling was sufficiently sensitive and specific to discover immune-related genes, we compiled a reference set of genes (n = 570) documented in the literature as being preferentially expressed in the immune system as defined by RT-PCR or Northern blot analysis (Supplemental Table I). To cross-validate our literature annotation for this immune-expression gold-standard set of genes, we performed GO/KEGG annotation with subsequent experimental follow-up.

We next assessed the strength of EST profiling against microarray profiling to provide: 1) an independent method of cross-validation of each platform; and 2) increased cellular resolution within the immune system of our EST-derived dataset. We quantified immune enrichment by calculating a Wilcoxon rank-sum test and adjusted for multiple hypothesis testing (FDR corrected p < 0.05) using the publicly available BioGPS atlas (9). This atlas is a broad compendium of human tissues, with 22 of 79 profiled tissues categorized as immune tissues. Of the 2232 Immunogenes, 1927 had representative probes on the microarray. Strikingly, 1522 out of 1927 (79%) of the genes found to be immune enriched based on EST profiling WIP score (WIP > 0.15) also exhibited significantly higher immune expression by microarray (p < 0.05; FDR microarray Wilcoxon test) (Fig. 3A). Hierarchical clustering and heat map representation of the 1927 EST Immunogene profiles separated by immune versus nonimmune tissue are shown in Fig. 3B, revealing enrichment of EST-derived Immunogenes in the BioGPS compendia and providing further cross-validation of our methodology.

We also reasoned that if expression is an indicator of potential function, then enrichment would also be found in GO biological process categories and KEGG pathways relevant to the immune system, which was indeed the case (Fig. 3C, Supplemental Fig. 1D). In addition, analysis of publicly available RNA-Seq data comparing spleen to 10 other nonimmune tissues revealed that the list of Immunogenes defined by EST profiling showed a significantly higher tendency to be expressed in spleen versus nonimmune tissues (Z = 7.03 for all genes, Z = 151.42 for Immunogenes, Wilcoxon signed-rank test) (Fig. 3D). KIAA0226L (WIP = 0.37; p = 1.65 × 10⁻³⁸) is an example of a novel immune-enriched gene identified by EST profiling. Interestingly, KIAA0226L has been implicated via proteomics to be involved in autophagy (26), and RNA-Seq has demonstrated expression of KIAA0226L in human spleen and mouse C19+ B cells, but not in mouse CD4+ T cells (3, 10). As the ability of EST analysis to spotlight immune-enriched genes was confirmed by multiple analyses, we designed a browseable Web-based portal of Immunogenes accessible at http://xavierlab2.mgh.harvard.edu/ESTProfiler/index.html.

**EST profiling reveals immune-enriched transmembrane, metabolic, and enzymatic signatures**

To fully characterize the Immunogene dataset, we took advantage of structure-function relationships. For this approach, we analyzed each Immunogene systematically by: 1) determining the presence of potential protein domains using Simple Modular Architecture Research Tool and Pfam database criteria; 2) checking for homology to any known or predicted proteins in multiple species using BLASTP alignments; and 3) identifying predicted or known transmembrane regions using the TMHMM prediction algorithm. Structural annotation/prediction revealed that 20% of Immunogenes (440 of 2232) have known or predicted enzymatic functions using an enzyme classification system or homology modeling. These observations led us to question whether such a large fraction of enzymes in the Immunogene gene set was a reflection of the
unique metabolic pathways and needs of immune cells. We used a similar number of UniGenes (460 of 2611) that were significantly underrepresented in immune tissues (WIP < 0.05; p < 0.05) as a control set of enzymes. Using two independent expression maps, BioGPS (p = 1.4 × 10⁻¹⁵) and NCBI serial analysis of gene expression tags (p = 1.2 × 10⁻²⁰), we observed that our EST immune enzyme inventory was most enriched in monocytes/dendritic cells compared with other cell types. In contrast, the control set of enzymes was enriched in the nervous system (p = 1 × 10⁻¹³). These results strongly support our conclusion that EST profiling has revealed a genome-wide compendium of enzymes that are likely to be particularly relevant to immune system function and development. For example, one novel potential enzyme identified as being immune-enriched was ARL11 (WIP = 0.3; p = 0.028), which is predicted to encode an ADP-ribosylation factor.

In other instances, we identified enzymes that have clear enzymatic function, but which have only been evaluated to a limited degree in immune system function. One such enzyme is fructose 1,6-diphosphatase, which is known to be critical in gluconeogenesis in the liver, but was revealed by EST profiling to be enriched in macrophages. Fructose 1,6-diphosphatase was previously found to be a vitamin D–responsive gene in monocytes (27). Thus, our data highlight how EST profiling can also help focus attention on potential intriguing links between metabolism and immune function (27, 28).

In addition to the enzymatic footprints of immune cells, our structure-function analysis demonstrated highly significant enrichment of genes encoding transmembrane proteins (n = 215; p = 7 × 10⁻¹⁰ “integral to membrane” GO annotation) with an even greater number of genes (n = 285) identified by transmembrane prediction algorithms. This result correlated with our finding that the largest PFAM and InterPro classes of protein domains in the Immunogene set were Ig domains (n = 87; p = 3.9 × 10⁻¹²). These transmembrane proteins included cytokine receptors, ion channels such the calcium channel ORAI2, as well as novel transmembrane proteins (n = 69) and predicted secreted proteins (n = 22). One example of a transmembrane protein that we identified as being immune enriched is the nucleotide transporter SLC29A3, which causes H syndrome, Faisalabad histiocytosis, and Rosai-Dorfman disease (OMIM identification number 602782) when mutated. Patients with H syndrome have overlapping features of hemophagocytic syndromes, suggesting that functional characterization of SCL29A3 is likely to yield insights into immune function (29). The strong representation of known or predicted nutrient and ion transporters is also consistent with the unique metabolic needs of the immune system. Our results provide a guide to further dissect the links between metabolism and immune function. An integrated portrait of immune cell transporters and enzymes generated using Immunogene is shown in Fig. 4A.

**FIGURE 2.** Quantitative EST-based immune expression enrichment analysis utilizing 26,552 UniGene clusters. (A) The frequency distribution of the WIP score (mean 0.095) and p values (χ² metric with FDR correction) for 26,552 UniGene clusters with ≥10 ESTs is shown. Red indicates clusters/gene significantly enriched in immune expression (WIP > 0.095; p < 0.05), whereas blue indicates clusters/gene significantly underexpressed in the immune system (WIP < 0.05 and p < 0.05). (B) The frequency distribution (density) of WIP scores for 26,552 UniGene clusters (gray; mean 0.095) versus the WIP scores of the 570 known immune-enriched reference gene sets (red; mean 0.44). (C) The distribution of WIP scores and p values for the 570 known immune-enriched reference gene sets showing that 85% (n = 388) of the UniGene clusters with WIP > 0.20 and p < 0.05 were not annotated as

Discovery and verification of lncRNAs using EST profiling

Given that our EST profiling is agnostic to the type of transcripts it identifies, we also identified several immune-enriched microRNAs, such as miR-155 (BIC) (p = 2.33 × 10⁻¹⁴). Further inspection of the human EST dataset revealed that 30% (n = 388) of the UniGene clusters with WIP > 0.20 and p < 0.05 were not annotated as
hypothetical or validated reference protein-coding genes at NCBI. Manual curation of these clusters revealed that a large fraction was likely composed of genomic contaminants, alternative exons, or untranslated regions of known protein-coding genes (see Materials and Methods). However, 46 UniGene clusters demonstrated evidence of splicing and did not show any significant similarity to known protein-coding genes using BLASTX, suggesting that these clusters likely represent lncRNAs (>200 bp). Using RT-PCR, we confirmed expression of 10 of 17 (59%) randomly selected clusters in human spleen (see Supplemental Fig. 3A, 3B, and Table II for examples). One such human-specific lncRNA identified as an Immunogene was AK056817 (WIP = 0.34; \( p = 7.0 \times 10^{-9} \)). Although the functions of the ncRNAs are largely unknown, analysis of ENCODE ChIP-Seq data in leukemic K562 cells revealed IFN-inducible binding of STAT1 to the promoter region of AK056817, suggesting that this lncRNA may function in IFN response pathways (30). Thus, we suggest that integrating additional genome-wide datasets can be used to refine potential functional pathways of Immunogenes.

AK056817 represents a likely lncRNA that is transcribed in an intergenic region away from other known genes. Yet many coding genes have recently been found to exist as pairs with lncRNAs that are divergently transcribed on the opposite strand with a TSS within a few kilobases of the known mRNA TSS (31). We therefore asked how many protein-coding Immunogenes harbored a divergently transcribed lncRNA based on the presence of a divergent spliced EST. For this analysis, we relaxed our stringency requirements such that a single EST from any tissue was sufficient to be counted, because most lncRNAs are transcribed at significantly lower levels than mRNAs. Interestingly, we identified 188 Immunogenes with divergently transcribed lncRNAs, including multiple susceptibility genes for immune disorders such as LRRK2, CYLD, and MFHAS1 (Supplemental Fig. 3C). We had previously shown that LRRK2 is an IFN-\( \gamma \)–responsive gene. Strikingly, we show that the antisense LRRK2 lncRNA, which we term antisense (AS)-LRRK2, is coordinately upregulated with LRRK2 upon IFN-\( \gamma \) stimulation (Fig. 4B). Thus, EST profiling can also be used to reveal novel lncRNAs responsive to specific immune-signaling pathways (30, 32, 33).

Immunogenes are enriched in primary immunodeficiency genes and in genomic regions associated with immune diseases/traitsgenetics in GWAS studies.

We found that the Immunogene compendium is significantly enriched in the KEGG category “primary immunodeficiency” (hsa05340; \( n = 28 \) of 34; \( p = 2.03 \times 10^{-14} \)). This KEGG pathway does not include other primary immunodeficiency genes that we identified as Immunogenes such as CARD9, CLEC7A, IL17RA.
familial candidiasis 2, 4, 5; OMIM identification numbers 212050, 613108, and 613953), CXCR4 (WHIM syndrome; OMIM identification number 193670), DOCK8 (hyper-IgE syndrome; OMIM identification number 243700), WAS (Wiskott-Aldrich syndrome; OMIM identification number 301000), and SH2D1A (X-linked lymphoproliferative disease; OMIM identification number 308240) (34) (Supplemental Table I). These results demonstrate the ability of our approach to highlight genes responsible for monogenic immune-related human diseases.

In terms of complex traits, GWAS have identified thousands of genetic loci related to disease phenotypes or traits, although identification of the exact causative polymorphism and mechanism...
(s) of action including relevant target gene(s) remains a challenge. We postulated that EST-defined Immunogenes may be significantly more enriched at loci implicated in immune-related diseases/traits (e.g., inflammatory bowel disease) given their expression in disease/trait-relevant tissue/cells (see *Materials and Methods* for details of immune-related diseases/traits analyzed). To test this possibility, we plotted the frequency of SNPs associated with human immune diseases/trails relative to the distance to the nearest Immunogene. The distance distribution of SNPs associated with nonimmune-related diseases/trails to the nearest Immunogene was used as a comparison. As shown in Fig. 4C, there is a higher frequency of SNPs associated with immune-related diseases/trails near Immunogenes, as compared with SNPs not associated with immune-related SNPs. Strikingly, we found that 254 of 808 (31.44%) immune-related disease/trait-associated SNPs had their closest RefSeq coding gene as an EST-defined Immunogene versus 783 of 7355 (10.65%) nonimmune disease/trait-associated SNPs ($p < 10^{-4}$ based on permutation testing).

A recent GWAS study identified the locus tagging SNP rs140522 as associated with MS. In this study, the authors cited the nearby gene encoding cytochrome oxidase deficient homolog 2 (SCO2) as the MS candidate gene within the locus (35). However, the nearest gene to rs140522 is not SCO2, but rather outer dense fiber of sperm tails 3B (ODF3B), which previously has not been implicated in immune function. Our approach identified ODF3B as novel Immunogene (Hs.531314; WIP = 0.229; $p = 0.005$). Thus, ODF3B is an alternative candidate gene for further testing in the pathophysiology of MS, illustrating the utility of integrating our EST expression profiling results with GWAS studies.

**Identification of PRDM15 as a novel B cell gene overexpressed in human B cell lymphomas**

We identified PRDM15, a zinc finger and SET domain-encoding gene, as one of the most highly immune-enriched (WIP = 0.71; $p = 5.1 \times 10^{-10}$) genes with little known function(s). We chose to focus on PRDM15 given our interest in B cell lymphoma pathogenesis coupled with our observation that a high fraction of the PRDM15-derived ESTs were from either normal germinal center B cells (BGCs) or from B cell lymphoma cell line cDNA libraries (36). BGCs are thought to be the normal counterpart of some human B cell malignancies, including follicular lymphoma, Burkitt lymphoma, and diffuse large B cell lymphoma. To refine and further corroborate our EST data, we therefore analyzed the relative expression pattern of PRDM15 among various types of B cell lymphomas and non-B cell cancer-derived cell lines using published microarray data (23). As shown in Fig. 5A, PRDM15 showed the highest expression in cell lines derived from follicular lymphomas ($n = 9$) and the second highest expression level in Burkitt lymphomas ($n = 17$), compared with germinal B cell– and activated B cell–type non-Hodgkin diffuse large B cell lymphoma ($n = 11$), Hodgkin lymphomas ($n = 9$), lung cancers ($n = 99$), and colon cancers ($n = 52$). These results supported the observation that PRDM15 is specifically overexpressed in human B cell lymphomas. These data are consistent with RT-qPCR data showing that the Burkitt lymphoma cell line, Daudi, has 8–35 times more PRDM15 mRNA expression than several other cell lines/types (Fig. 5B). To examine PRDM15 expression at the protein level, we confirmed the nuclear localization of PRDM15 in HEK293T cells (Fig. 5C) and used Western blot analysis to demonstrate that PRDM15 protein is expressed in multiple Burkitt lymphomas and some monocytic leukemias, but not in T cell leukemias, suggesting that PRDM15 is expressed at higher levels in Burkitt B cell lymphomas not only at the mRNA level, but also at the protein level (Fig. 5D). In terms of follicular lymphoma expression, additional microarray analysis confirmed that PRDM15 was not only overexpressed in follicular lymphoma cell lines, but also in primary follicular lymphomas compared with primary normal B cell populations (Fig. 5E). Strikingly, immunohistochemistry analyses also revealed that PRDM15 was overexpressed in five of seven independent human follicular lymphoma samples examined, as compared with normal tonsillar germinal centers (Fig. 5F). Thus, our results demonstrate how initial insights gained by EST profiling can be used to identify a novel transcription factor overexpressed in human B cell lymphomas.

**Discussion**

In this study, we present the application of a quantitative genome-wide analysis of the human/mouse dbEST database composed of >6 million data points to elucidate a collection of genes enriched in immune expression, which we term *immunogenes*. Using this database, we were able to identify an immune transmembrane and enzymatic signature. Furthermore, we implicated a novel transcription factor in B cell oncogenesis. Our results provide an important complement to other studies aimed at understanding the immune system including the RefDic, ImmGen, and Immune Response In Silico databases, all of which largely used microarray profiling (2–5).

Having a genome-wide set of genes with preferential immune expression, we undertook a systems-level analysis of the Immunogene compendium, finding a unique enzymatic signature of immune cells. The study of the large set of known or predicted enzymes, transporters, and corresponding metabolites is likely to yield critical insights into the metabolism of immunity. Furthermore, the Immunogene database provides an enhanced view of the potential spectrum of potential transmembrane cell-surface markers expressed on various immune cells. These novel transmembrane Immunogenes represent potential candidates for targeted therapeutics or cell purification over the long-term via the development of Abs directed against these Ags. We demonstrated how the Immunogene approach can be used to identify genes relevant to the pathogenesis of lymphomas. Along these lines, we dissected the expression pattern of PRDM15, finding it to be overexpressed at the level of mRNA and protein in human follicular lymphomas. Thus, delineation of PRMD15 function may prove to be important in understanding BGC physiology and B cell malignancies.

The identification of genes involved in both Mendelian and complex traits encompassed within the dataset support the utility of the Immunogene database as an aid in genetic studies. In this regard, we demonstrate how the Immunogene dataset can be used to spotlight candidate genes in which GWAS studies, genetic linkage studies, quantitative trait locus analysis, or mouse mutagenesis screens have identified genomic intervals containing numerous genes. In addition, the Immunogene database can be harnessed to probe the variation of immune system responses through focused analysis of SNPs and copy number variations of the identified Immunogenes. The analysis of genetic variation in the Immunogene dataset, through methods such as coding variant SNP analysis or resequencing, is poised to yield a pool of human genes at the cornerstone of variation in human immune responses, including susceptibility to autoimmune diseases such as lupus, inflammatory bowel disease, and diabetes. In addition, we demonstrated how EST data can be used to identify lncRNAs in autoimmune disease loci, with the identification of IFN-inducible AS-LRRK2 lncRNA in the Crohn’s disease *LRRK2* locus. Our data suggest that Crohn’s disease-associated SNPs such as rs11175593, which lies in an intron of the AS-LRRK2 lncRNA, may modulate the function and/or expression of not only LRRK2, but also its coregulated partner AS-LRRK2. Furthermore, extension of the Immunogene paradigm to
other tissues may aid in deciphering human genes likely to be uniquely associated with other systems, such as the nervous system.

We anticipate that by starting with the Immunogene dataset as a robust input, accompanied by multiple new layers of annotation, this resource will provide a dynamic platform to generate hypotheses to study known and novel genes in a variety of immune pathways.

Acknowledgments

We thank Natalia Nedelsky for substantive editorial assistance.
Disclosures
The authors have no financial conflicts of interest.

References


