Supporting Information

Ergun et al. 10.1073/pnas.1311839110

SI Materials and Methods

ImmGen Consortium. David A. Blair[a], Michael L. Dustin[a], Susan A. Shinton[b], Richard R. Hardy[b], Tal Shay[c], Aviv Regevc[c], Nadia Cohen[d], Patrick Brennan[d], Michael Brenner[d], Francis Kim[e], Tata Nageswara Rao[e], Amy Wagers[e], Tracy Heng[f], Jeffrey Ericson[f], Katherine Rothamel[f], Adriana Ortiz-Lopez[f], Diane Mathis[f], Christophe Benoist[f], Taras Kreslavsky[g], Anne Fletcher[g], Kutlu Elpek[g], Angelique Bellemare-Pelletier[g], Deepali Malhotra[g], Shannon Turley[g], Jennifer Miller[h], Brian Brown[h], Miriam Merad[h], Emmanuel L. Gautier[h,i], Claudia Jakubzick[h], Gwendalyn J. Randolph[h,i], Paul Monach[j], Adam J. Best[k], Jamie Knell[k], Ananda Goldrath[k], Vladimir Jojic[l], Daphne Koller[l], David Laidlaw[m], Jim Collins[n], Roi Gazit[o], Derrick J. Rossi[o], Nidhi Malhotra[p], Katelyn Sylvia[p], Joonsoo Kang[p], Natalie A. Bezman[q], Joseph C. Sun[q], Gundula Min-Oo[q], Charlie C. Kim[q], and Lewis L. Lanier[q].

[a]Skirball Institute of Biomolecular Medicine, New York University School of Medicine, New York, NY 10016; [b]Fox Chase Cancer Center, Philadelphia, PA 19111; [c]Broad Institute and Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142; [d]Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Boston, MA 02115; [e]Section on Islet Cell and Regenerative Biology, Joslin Diabetes Center, Boston, MA 02215; [f]Division of Immunology, Department of Microbiology and Immunobiology, Harvard Medical School, Boston, MA 02115; [g]Department of Cancer Immunology and AIDS, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA 02115; [h]Icahn Medical Institute, Mount Sinai Hospital, New York, NY 10029; [i]Department of Pathology and Immunology, Washington University in St. Louis, St. Louis, MO 63110; [j]Department of Medicine, Boston University, Boston, MA 02118; [k]Division of Biological Sciences, University of California, San Diego, La Jolla, CA 92093; [l]Computer Science Department, Stanford University, Stanford, CA 94305; [m]Computer Science Department, Brown University, Providence, RI 02912; [n]Department of Biomedical Engineering, Howard Hughes Medical Institute, Boston University, Boston, MA 02115; [o]Program in Molecular Medicine, Children's Hospital, Boston, MA 02115; [p]Department of Pathology, University of Massachusetts Medical School, Worcester, MA 01655; and [q]Department of Microbiology and Immunology, University of California, San Francisco, CA 94143.

Alternative Splicing Discovery from RNA-Sequencing Data. RNA from CD4+ T and CD19+ B cells was used for the preparation of nonstrand-specific paired-end cDNA libraries in a format compatible with Illumina sequencing technology, which was modified from the work by Parkhomchuk et al. (1). In brief, an mRNA fraction was enriched from total cellular RNA using oligo(dT) selection, treated with DNase, fragmented, and converted to cDNA with random hexamers. cDNA libraries were subjected to strand-independent Illumina library preparation and sequenced using an Illumina sequencing platform (2 × 76 bp).

We mapped the RNA-sequencing (RNA-seq) data to the mm9 (NCBIM37) genome assembly using TopHat (2), version 1.1.4, which required that reads map uniquely. Using the UCSC known gene annotations, we computed gene expression estimates using Cufflinks (3).

TopHat outputs candidate splice junction predictions and a read count for the number of sequencing reads supporting each prediction. We took these reads to represent RNA splice junctions within our samples. Separately we computed alignment and coverage files to visualize the distribution of sequencing reads in a genome browser. Using the summary read counts at each read position within the genome based on the coverages, we estimated the exon-level expression from the number of reads falling within exon boundaries by computing the mean of the piled-up reads with no subsequent normalization. (In total, we obtained 161,948 and 167,067 splice junctions, where 95,699 and 96,372 of these junctions were observed more than 10 times in CD4+ T and CD19+ B cells, respectively.) We further eliminated junctions ending in multiple genes or located entirely in intronic regions. This process resulted in a total of 92,562 and 91,882 junctions in CD4+ T and CD19+ B cells, respectively.

On the basis of the UCSC mm9 mouse exon annotation boundaries, we further categorized these splice junctions as canonical or alternative. If a junction end perfectly matched to two neighboring exon boundaries, we characterized it as canonical. Any junction with skipped exons or acceptor or donor sites previously not in existing databases was characterized as alternative and binned into one of the categories shown in Fig. 1.

We defined the skipping ratio of an exon as the total read count of junctions skipping that exon divided by the total number of reads skipping it and junctions including it. A junction was considered to include an exon if it fell within ±4 bp around exon-annotated boundaries (Fig. S2*A*).

Alternative Splicing Discovery from Microarray Data. Transcriptional profiling within the Immunological Genome Project (ImmGen) cell populations has been previously described (4) (www.immgen.org). All ImmGen microarray data were generated using Affymetrix MoGene1.0 ST chips, which contain probes across the gene body. With two to three probes falling within an exon on average, we could characterize exon-level expression. To enable cross-comparison between microarray and RNA-seq data, we selected a set of probes with expression that was well-correlated (Pearson coefficient ≥ 0.7) between the microarray and RNA-seq datasets. Our final list was composed of 171,000 probes located in 75,673 exons and 7,261 genes. We estimated exon- and gene-level expression of Affymetrix probes after probe-level robust multiarray average (RMA) normalization (5) and summarized at the probe, exon, or gene level. For analysis of differential use, we selected 172 populations that excluded nonhematopoietic cell types and samples with lesser dynamic range.

Exon-Centric Discovery by Linear Regression. For each of 171,000 probes, we performed a linear regression between the probe's expression and the integrated expression of the corresponding gene across 802 ImmGen samples (three replicates per cell population on average) and computed a residual of each sample to a regression line estimated by a first-degree polynomial fit. We only retained those exons that had a consistent slope for each probe–gene regression for each of the probes and averaged the residual values for each exon and cell type. For each cell type, we selected the top 50 exons with the highest residual value, with exon/gene expression ratio ≤ 1 (Fig. S2*C*) and gene-level expression ≥ 100. We furthered required that at least five cell types have high residual value and that the difference in exon/gene expression ratio was at least 0.4 (75th percentile) among the cell types (Fig. S2*B*).

Population-Centric Analysis of Exon/Gene Ratios. For each cell type, we computed exon–gene ratios and selected exons that had a difference of exon- and gene-level expression > 0.4 (log10 scale) and a gene-level expression ≥ 100. To eliminate exons that consistently showed lower exon expression compared with gene-level

expression, we additionally required that at least five cell types not fall into this range and had a difference in maximum and minimum exon inclusion/exclusion ratios of at least 0.4 (75th percentile) among the cell types. Fig. S2*B* shows the distribution of the difference between maximum and minimum values of the exon/gene expression ratios.

Combining the results from both the exon- and population-centric methods resulted in 4,321 exons residing in 1,842 genes as candidate exons that showed differential splicing among ImmGen cell types. Prediction of exon exclusion patterns from Affymetrix microarrays in CD4+ T and CD19+ B cells for both the exon- and cell type-centric methods showed a similar profile of validation compared with skipped exons obtained from RNA-seq. True positives were 15–20% for up to 100 flagged exons but leveled off to 10% if more exons were flagged (Fig. S2 *D* and *E*).

Clustering of Flagged Exons. For k ranging from 10 to 30, we performed a k-means clustering on the row-normalized exon/gene ratios (a metric of exon inclusion), with 10 random initializations for each value of k. We chose the clusters that had the biggest difference between mean intercluster distances and mean intracluster distances in Euclidean space, which resulted in 28 clusters.

Splicing Regulators. We compiled a list of 229 genes with gene ontology annotation related to splicing and splicing regulatory activity. Among these genes, 165 genes were expressed ≥500 in a cell population. We correlated the expression value of each regulator across ImmGen cell types with the exon inclusion/exclusion metric and computed a P value from a background distribution of correlations of each exon inclusion/exclusion metric vs. a random set of 500 genes. We used a Benjamini–Hochberg procedure to compute false discovery rate values for significance between 165 splicing regulators and 4,321 exons; 7,536 such pairs scored at a corrected false discovery rate < 0.5%. Because of the high number of intercorrelated regulators, we focused on the subset to 27 specific regulators shown in Fig. 5*D*.

1. Parkhomchuk D, et al. (2009) Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res* 37(18):e123.
2. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* 25(9):1105–1111.
3. Trapnell C, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28(5):511–515.
4. Heng TS, Painter MW; Immunological Genome Project Consortium (2008) The Immunological Genome Project: Networks of gene expression in immune cells. *Nat Immunol* 9(10):1091–1094.
5. Irizarry RA, et al. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 31(4):e15.

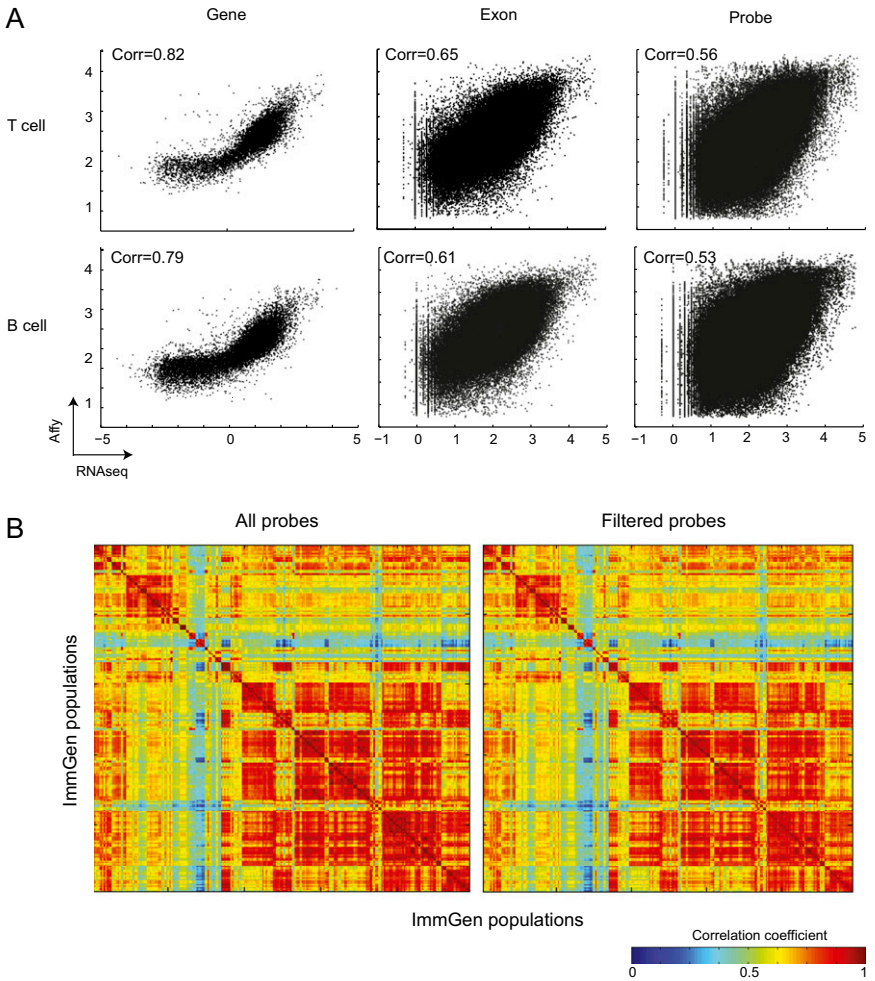

Fig. S1. (*A*) Gene-, exon-, and probe-level expression comparison between microarray and RNA-seq profiles in CD4+ T and CD19+ B cells. (*B*) Pairwise correlation at gene level between ImmGen samples using all probes on the Affymetrix Mouse 1.0 ST chip (*Left*) and the integrated gene-level expression of a selected subset of 171,000 probes (*Right*). Population-level correlations at gene level did not change when we used a filtered subset of probes.

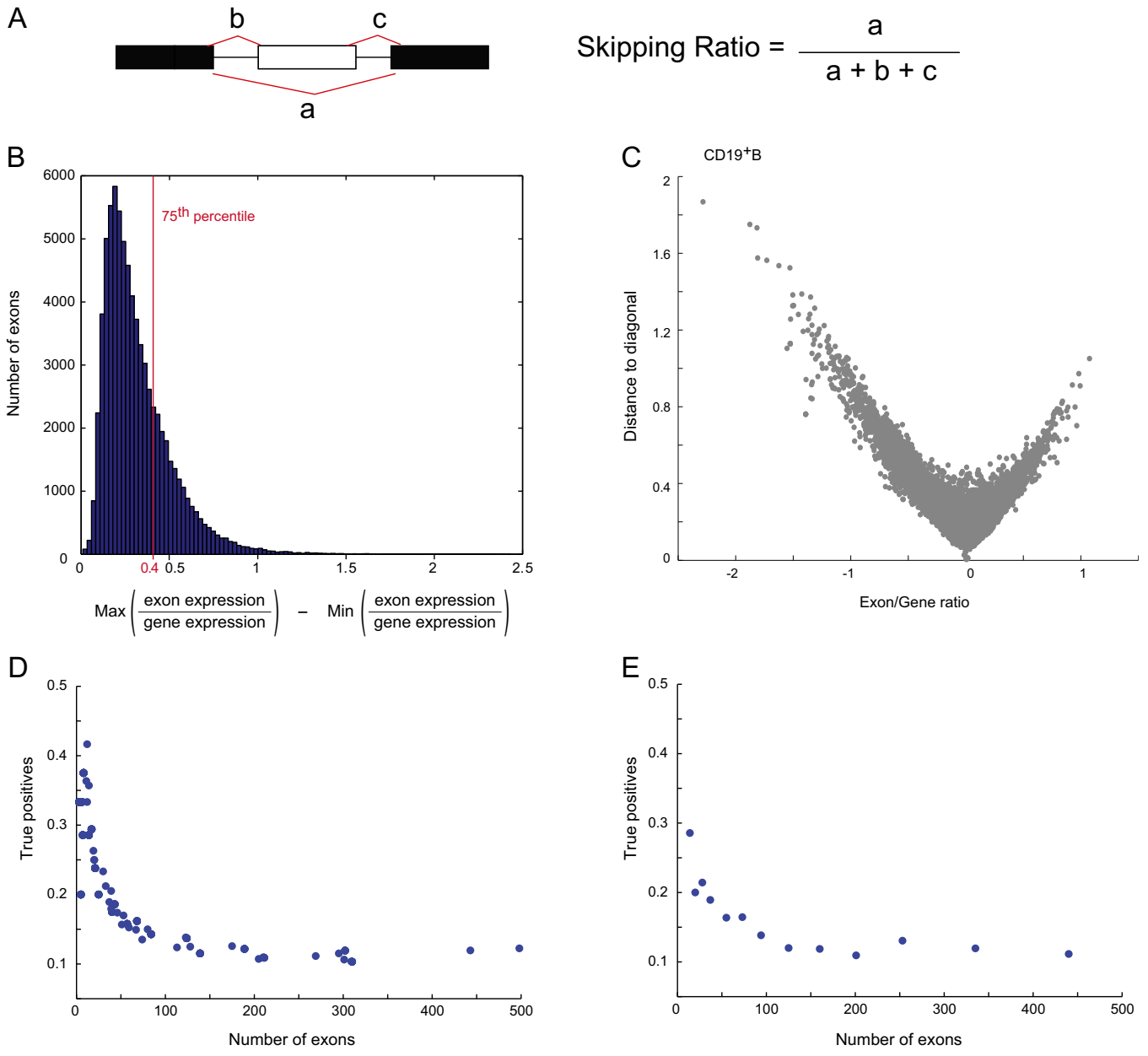

Fig. S2. (*A*) Skipping ratio for an exon computed as the ratio of junctions skipping that exon normalized by the total number of reads skipping it and junctions including it. (*B*) Difference between maximum and minimum exon inclusion/exclusion ratio for a given exon; 75th percentile (0.4) was chosen as a cutoff to ensure variability between inclusion/exclusion in flagged exons for the population-centric method. (*C*) Combined residual values for each population vs. exon/gene ratio from the exon-centric method. (*D* and *E*) Comparison between flagged exons in Affymetrix and skipped exons obtained from exon–exon junctions from RNA-seq data in CD19+ B cells. Percent of true positives flagged by the exon-centric method as a function of exon number and percent of true positives flagged by the population-centric method as a function of exon number, respectively, are shown.

Other Supporting Information Files

Dataset S1 (XLSX)
Dataset S2 (XLSX)
Dataset S3 (XLSX)
Dataset S4 (XLSX)
Dataset S5 (XLSX)
Dataset S6 (XLSX)
Dataset S7 (XLSX)