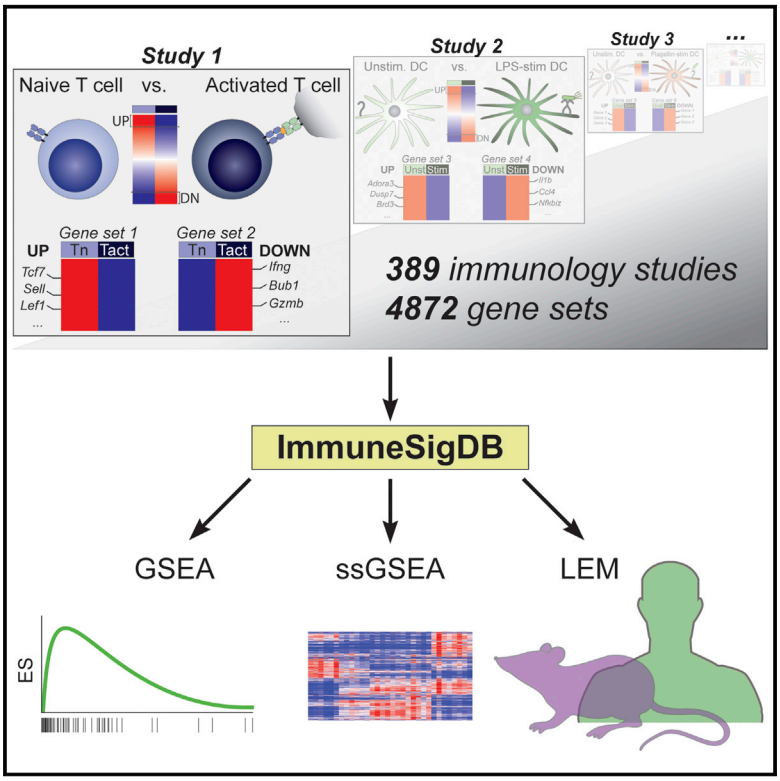
Resource

Immunity

Compendium of Immune Signatures Identifies Conserved and Species-Specific Biology in Response to Inflammation

Graphical Abstract



Highlights

- ImmuneSigDB: Collection of ~5,000 gene sets derived from ~400 immunological studies
- Includes a wide range of mouse and human immune cell states and perturbations
- Designed for use with GSEA and an approach called Leading Edge Metagene analysis
- ImmuneSigDB identified shared and unique biology in human and mouse sepsis response

Authors

Jernej Godec, Yan Tan, Arthur Liberzon, ..., Atul J. Butte, Jill P. Mesirov, W. Nicholas Haining

Correspondence

nicholas_haining@dfci.harvard.edu

In Brief

Meaningful interpretation of gene-expression analyses relies on identifying changes in expression of sets of genes corresponding to specific biological processes and cell states. Haining and colleagues generated a collection of ~5,000 annotated, immunology-specific gene sets and uncovered shared and species-specific biology in mouse and human transcriptional responses to sepsis.

Godec et al., 2016, Immunity *44*, 194–206
January 19, 2016
http://dx.doi.org/10.1016/j.immuni.2015.12.006

CellPress

CellPress

Immunity
Resource

Compendium of Immune Signatures Identifies Conserved and Species-Specific Biology in Response to Inflammation

Jernej Godec,[1,2] Yan Tan,[3,4] Arthur Liberzon,[3] Pablo Tamayo,[3] Sanchita Bhattacharya,[5] Atul J. Butte,[5] Jill P. Mesirov,[3,4] and W. Nicholas Haining[1,3,6,*]

[1]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, USA
[2]Department of Microbiology and Immunobiology, Harvard Medical School, Boston, MA 02115, USA
[3]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[4]Bioinformatics Program, Boston University, Boston, MA 02215, USA
[5]Institute for Computational Health Science, University of California, San Francisco, San Francisco, CA 94158, USA
[6]Division of Hematology/Oncology, Children's Hospital, Harvard Medical School, Boston, MA 02115, USA
*Correspondence: nicholas_haining@dfci.harvard.edu
http://dx.doi.org/10.1016/j.immuni.2015.12.006

SUMMARY

Gene-expression profiling has become a mainstay in immunology, but subtle changes in gene networks related to biological processes are hard to discern when comparing various datasets. For instance, conservation of the transcriptional response to sepsis in mouse models and human disease remains controversial. To improve transcriptional analysis in immunology, we created ImmuneSigDB: a manually annotated compendium of ~5,000 gene-sets from diverse cell states, experimental manipulations, and genetic perturbations in immunology. Analysis using ImmuneSigDB identified signatures induced in activated myeloid cells and differentiating lymphocytes that were highly conserved between humans and mice. Sepsis triggered conserved patterns of gene expression in humans and mouse models. However, we also identified species-specific biological processes in the sepsis transcriptional response: although both species upregulated phagocytosis-related genes, a mitosis signature was specific to humans. ImmuneSigDB enables granular analysis of transcriptomic data to improve biological understanding of immune processes of the human and mouse immune systems.

INTRODUCTION

Experiments in both human cells and mouse models have been used to discover many of the mechanisms by which the immune system functions. Identifying aspects of immunobiology that are evolutionarily conserved between humans and mouse models is useful because it can reveal mechanisms of fundamental importance to both species. Moreover, it can provide reassurance that information gleaned from mouse models will be applicable to the human condition. This is crucial, because much of immunobiology cannot be examined physiologically in humans due to inaccessibility of certain tissues or cell types or the difficulty in recapitulating complex biological milieu in vitro. However, considerable controversy exists as to the degree to which mouse models can recapitulate events occurring in immunologic disease states in humans (Davis, 2008; Hackam and Redelmeier, 2006; Rice, 2012; Shay et al., 2014; van der Worp et al., 2010; Warren et al., 2014). These concerns have extended to the analysis of genome-wide analysis of mRNA levels where analyses of the same datasets from mouse and human sepsis reached opposite conclusions regarding the degree of cross-species similarity (Seok et al., 2013; Takao and Miyakawa, 2014). Contradictory findings have also been reported in the comparison of gene expression across a range of human and mouse tissues (Gilad Y, 2015; Lin et al., 2014).

One of the challenges in identifying similarities between gene-expression datasets is that major changes in the cell state can be associated with relatively small alterations in the expression level of a relatively large numbers of genes. Analysis of co-regulated changes in sets of functionally related genes, rather than individual genes, has therefore become an important strategy to identify subtle, but biologically meaningful, differences in gene expression (Haining and Wherry, 2010; Mootha et al., 2003; Subramanian et al., 2005). This is a particularly useful approach when analyzing samples in which experimental variability (such as those collected from heterogeneous human subjects) or evolutionary divergence (such as comparisons between species) add experimental "noise" to gene-expression profiles. Several approaches for testing for the enrichment of gene sets have been developed, including gene set enrichment analysis (GSEA) (Subramanian et al., 2005). GSEA has been made more powerful by the availability of curated collections of gene-expression signatures extracted from a variety of sources including published experimental datasets. The largest of these collections, the Molecular Signatures Database (MSigDB), contains more than 8,000 signatures (Liberzon et al., 2011). However, only a small fraction of these gene sets pertain to immune processes and cell types.

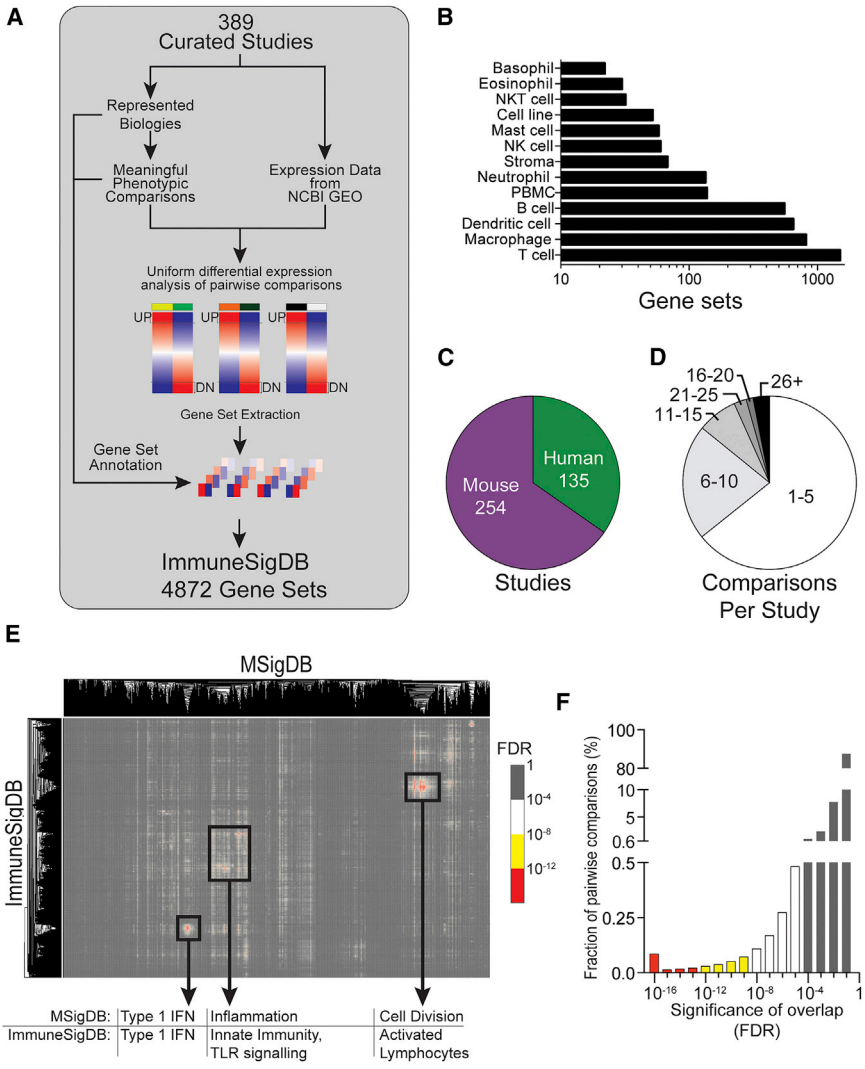


Figure 1. ImmuneSigDB Collection Is Derived from Re-Analysis of Published Data

(A) A schematic of the ImmuneSigDB pipeline.

(B) Number gene sets corresponding to major immune lineages or cell lines and (C) species of origin contained in ImmuneSigDB.

(D) Number of pairwise comparisons made per each study used in ImmuneSigDB.

(E) Overlap in gene set membership in ImmuneSigDB with MSigDB gene sets. Heatmap indicates False Discovery Rate (FDR) values of each pairwise comparison between gene sets. Highlighted are representative biological processes in each of the significantly overlapping clusters of gene sets.

(F) Distribution of the FDR ranges of significance across all pair-wise comparisons of gene set membership. See also Figures S1 and S2.

RESULTS

Generating a Compendium of Gene Signatures Curated from Immune-Expression Profiles

We generated a comprehensive compendium of gene sets pertaining to immune biology. The term "gene-set" in this study refers to groups of genes identified by selecting either up- or downregulated genes in comparisons of gene-expression profiles of interest. We identified and uniformly analyzed 389 published studies in the immunology literature that included genome-wide expression profiling data (outlined in Figure 1A). We selected studies to analyze based on immunological key words in the title or abstract followed by additional manual review. We prioritized studies published in immunology journals of broad interest (Table S1). We identified the corresponding publicly available datasets in the NCBI Gene Expression Omnibus (GEO) and, for uniformity, focused on studies performed on Affymetrix platforms (Table S1) that included three or more biological replicates. Each study was reviewed to identify and annotate the biology represented and to define meaningful pairwise comparisons that would create biologically useful gene sets. For example, an individual study might include a single comparison, such as stimulated versus unstimulated cells, or might have multiple comparisons, as is the case where several cell types were subjected to different culture conditions and analyzed at several time points. In such cases, only meaningful pairwise comparisons, rather than all possible comparisons, were made (Figure S1).

The raw expression data obtained from each GEO study was pre-processed uniformly (see Experimental Procedures). We identified and extracted differentially expressed genes (see Experimental Procedures and Figure 1A), which comprised the gene sets for the ImmuneSigDB collection. These sets represented genes coordinately up- or downregulated in many major

We now report the creation of ImmuneSigDB (http://software.broadinstitute.org/gsea/msigdb/collections.jsp#C7), a compendium of ~5,000 well-annotated signatures generated by analysis of 389 published studies of cell states and perturbations in the mouse and human immune systems. Using this collection of signatures, we demonstrated that signatures of cell differentiation in lymphoid cells and endotoxin stimulation in myeloid cells are highly conserved between humans and mouse models. Moreover, analysis of the transcriptional response to sepsis in human samples and mouse models showed that there was highly significant conservation of gene expression between the species when measured at the gene set level. However, in addition to the conserved transcriptional programs, we also identify species-specific differences in the biological processes associated with sepsis. These findings help interpret contradictory observations regarding the extent of evolutionary conservation in the transcriptional response to sepsis. ImmuneSigDB will enable the detailed analysis of cross-species gene expression that is critical to establishing which biological processes are conserved and which are not, thus allowing mouse models to better inform our understanding of human disease.

immune cell types (Figure 1B) either in their baseline state or following a range of genetic or chemical perturbations. ImmuneSigDB included data from healthy human subjects, patients with immune or non-immune diseases, and mouse models. Mapping orthologous genes to a common identifier allowed us to include both human (n = 135) and mouse (n = 254) studies (Figure 1C). From these studies, we identified 2,436 meaningful comparisons and extracted 4,872 gene sets of up- or downregulated genes comprising the ImmuneSigDB (see Experimental Procedures). The number of gene sets identified per published study ranged from one comparison, (e.g., representing an activated versus unperturbed state or knockout versus wild-type cell) to over 50 (e.g., often representing several cell types cultured in different conditions for varying amounts of time) (Figure 1D). Particular biological conditions over-represented in the literature, such as those related to T cell biology, are correspondingly over-represented in ImmuneSigDB, with slightly fewer gene sets from myeloid cells and B cells (Figure 1B). ImmuneSigDB is publicly available at http://www.msigdb.org.

ImmuneSigDB Expands the Biological Coverage of the MSigDB

We compared the gene sets generated from immune cells (ImmuneSigDB) with those in gene sets in the MSigDB collection. MSigDB is a curated collection of gene sets generated from published gene-expression studies that are generally not from the immunology literature (Liberzon et al., 2011). We measured overlap in constituent genes between each gene set in the ImmuneSigDB and all the other MSigDB collections and found that only a small minority of gene sets significantly overlapped (Figures 1E and 1F), suggesting that ImmunSigDB added a large amount of new transcriptional information. A small subset of gene sets in ImmuneSigDB and MSigDB were highly similar (0.64% of gene sets with $p < 10^{-8}$) and these could be clustered into three groups related to proliferation, inflammation, or type 1 interferon response (Figure 1E). This suggested that with the exception of these core biological processes, gene sets derived from immune cell expression profiles contained genes distinct from non-immune-related gene-expression profiles that previously predominated the MSigDB.

We performed an analogous analysis of pairwise overlaps in gene membership between gene sets within ImmuneSigDB. While most were unique, we found a larger number of gene sets with significant overlap (1.46% with $p < 10^{-8}$) within ImmuneSigDB than between ImmuneSigDB and MSigDB (Figure S2A). These gene sets largely related to lineage-specific signatures shared between datasets generated from similar types of cells. Therefore, ImmuneSigDB has minimal overlap with MSigDB and provides new gene sets describing immune biology.

ImmuneSigDB Provides a Complementary Resource to Existing Immune Module Collections

Several groups have previously created collections of gene modules in the immune system. In studies by Chaussabel et al. (2008) and Li et al. (2014), existing studies of gene-expression profiles in human peripheral blood mononuclear cell (PBMC) or whole blood were analyzed to identify modules of co-regulated genes to aid in the analysis of gene-expression profiles from immune cells. Several features distinguish ImmuneSigDB from either of these collections (summarized in Table S2). First, ImmuneSigDB was generated by direct comparison of the genes that were up- or downregulated in two known sample classes from each study. This allowed the published study to serve as a source of comprehensive annotation of each gene set, in contrast to either of the module collections that were generated by analysis of aggregated pools of samples, limiting the direct experimental annotation of each module. Second, ImmuneSigDB was considerably larger than either module collection (Table S2). Third, ImmuneSigDB included data from both mouse models and humans, and from 13 cell or tissue types, rather than solely from human PBMC and whole blood profiles.

To compare directly the gene-sets in ImmuneSigDB with the module collections of Chaussabel and Li, we measured overlap in constituent genes between each gene set in the ImmuneSigDB and all modules in either the Chaussabel or Li collections (Figures S2B and S2C). We found that only a small fraction of ImmuneSigDB gene sets significantly overlapped with either collection (0.06% and 0.18% with FDR of $< 10^{-8}$ for Chaussabel and Li, respectively), suggesting that the gene-sets within ImmuneSigDB and the modules in the Chaussabel and Li collections were largely distinct. The small number of significantly overlapping gene-sets and/or modules contained genes predominantly related to immune cell lineages (e.g., T cell or myeloid) or to the response to interferon-α (IFN-α) or Toll-like receptor (TLR) ligands. Interestingly, the overlap between modules contained in the Chaussabel and Li collections was similarly limited (Figure S2D), suggesting analysis of immune-expression profiles using each of the three collections could provide complementary information.

Finally, we performed GSEA using four published datasets in human immune cells (LPS stimulated DC, Tregs, plasma B cells, and memory B cells) to compare the results using ImmuneSigDB with the module collections by Chaussabel and Li (Figure S2E). A larger number of ImmuneSigDB gene sets were significantly enriched (even after correcting for multiple hypothesis testing) in each of the four datasets than with either the Chaussabel or Li collections. Moreover, inspection of the top 20 most enriched gene sets from ImmuneSigDB and modules from the Chaussabel or Li collections illustrates the extensive biological annotations (including links to the original studies) available for each ImmuneSigDB gene set (Table S3). Thus analysis with ImmuneSigDB provides a resource for the analysis of gene-expression data in the immune system that is complementary to existing collections.

Enrichment of ImmuneSigDB Gene Sets Recapitulates Known Lineage-Specific Differences in Mouse and Human Hematopoietic Cell Lineages

We next tested whether enrichment analysis of gene expression using ImmuneSigDB could recapitulate known differences in lineage-specific gene expression within the immune system. We analyzed a large, publicly available dataset of gene-expression profiles from the Immunological Genome Project (ImmGen) consisting of immune cell types and cell states in mice (Heng et al., 2008) using a single sample version of GSEA (ssGSEA) (Barbie et al., 2009). In this approach, gene sets are tested for enrichment in the list of genes in a single sample ranked by absolute expression rather than by comparison with another sample. The resulting ssGSEA scores provided an estimate of the

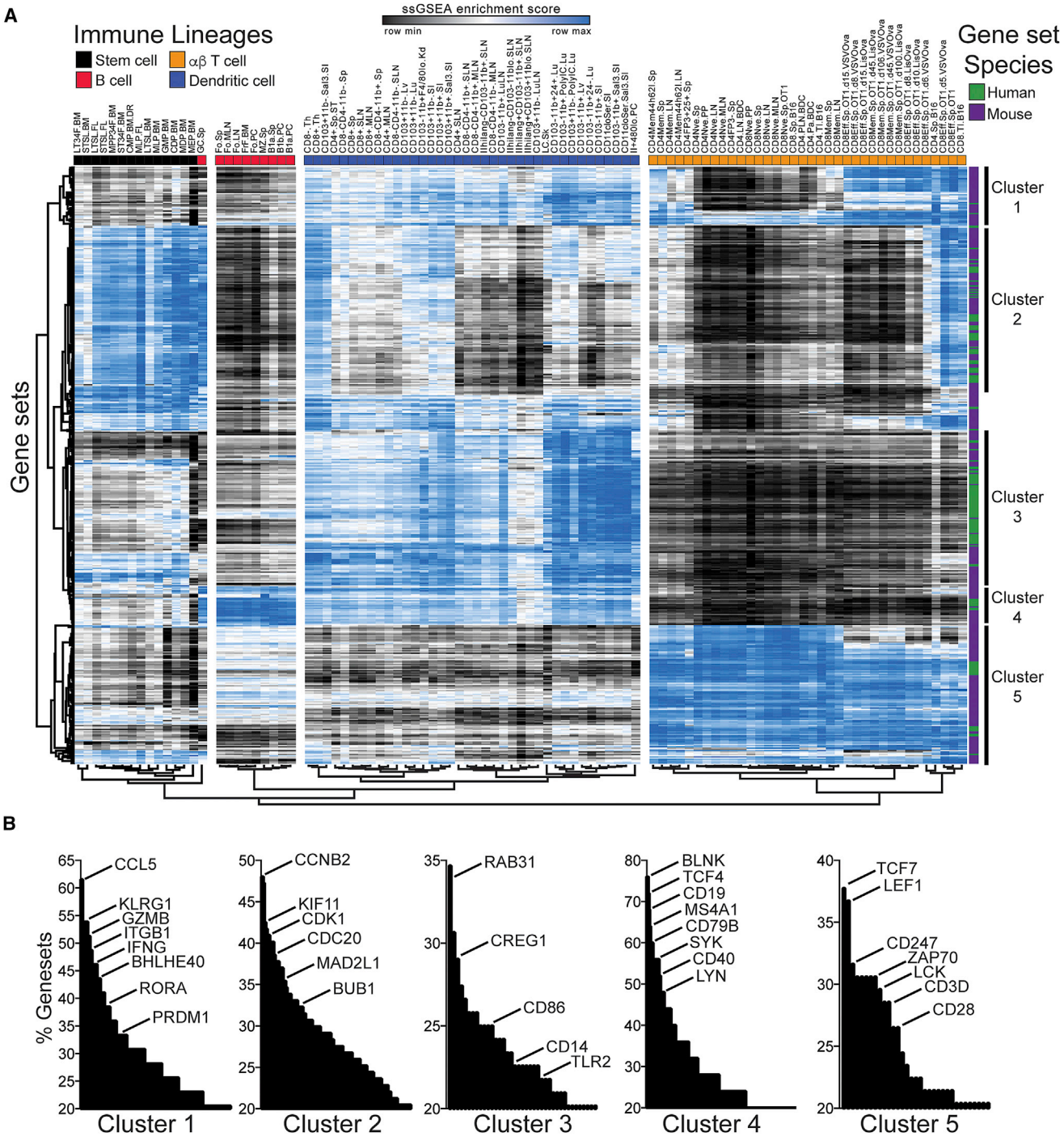


Figure 2. Mouse Immune Lineages Are Accurately Clustered using ImmuneSigDB Enrichments

(A) Unsupervised bi-clustering of ssGSEA values using ImmuneSigDB in samples of four representative mouse immune lineages. Hierarchical clustering of the 10% of gene sets with highest mean absolute deviation is shown. Species of origin of gene sets indicated by green (human) and purple (mouse) bars on the right. Major branches of the gene set dendrogram clusters are indicated by the numbered black bars on the right.

(B) Distribution of genes contained in gene sets in the same gene set dendrogram clusters as indicated in (A). See also Figure S3.

degree of enrichment of each ImmuneSigDB gene set in each individual sample in the dataset. In this way, we generated a dataset containing as rows the profiles of enrichment of each ImmuneSigDB gene set and as columns the samples. Unsupervised hierarchical clustering of samples from four distinct immune cell types—dendritic cells, B cells, αβ T cells, and stem cells—in the space of gene set enrichment scores revealed near-perfect clusters of the respective cell types (Figure 2A).

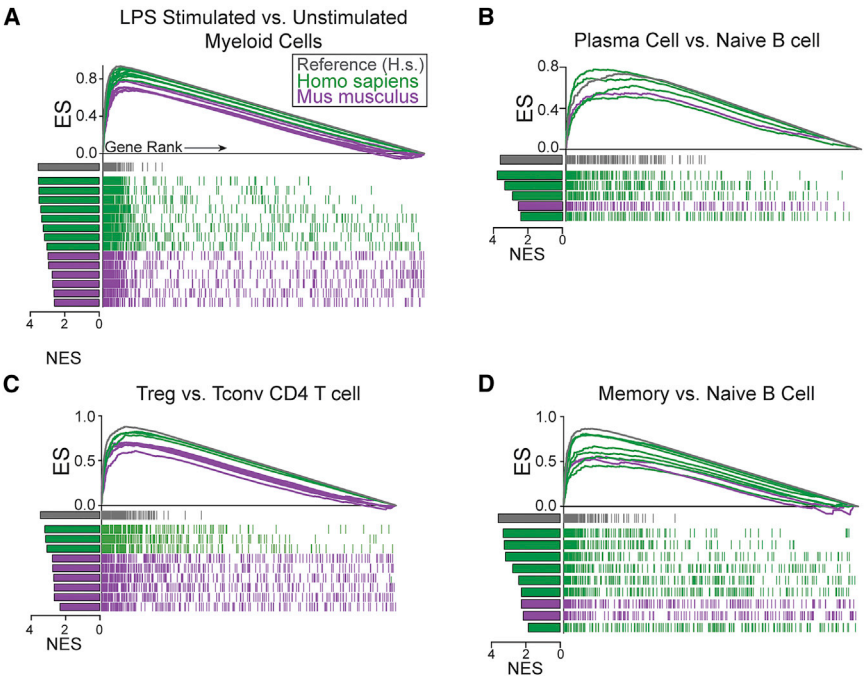


Figure 3. Transcriptional Programs Are Conserved across Mouse and Human Immune Lineages

(A) GSEA of a randomly selected human study comparing LPS-stimulated and unstimulated dendritic cells using ImmuneSigDB gene sets derived from the study itself (gray) or gene sets from other mouse (purple) or human (green) datasets of LPS-stimulated myeloid cells. Mountain plots show all genes ranked by differential expression in sepsis versus control conditions on the x axis, and the curves indicate cumulative enrichment (measured by enrichment score on the y axis). The ticks below the line correspond to the position of genes in each gene set.

(B-D) Analysis as in (A) for three additional cell differentiation states: plasma cells (B), Tregs (C), and memory B cells (D). All gene sets shown are significantly enriched (FDR < 0.001).

Within each lineage, subgroups, such as naive T cells or memory T cells also were clustered accurately together. Similarly, accurate clustering of different lineages was observed when we analyzed human-derived cells in the Differentiation Map (DMAP) (Figure S3) (Novershtern et al., 2011). Hematopoietic stem cells (HSCs) were accurately distinguished from other lineages despite the fact that very few (1.62%) gene sets derived from stem cells were included in the ImmuneSigDB. This suggests that HSCs are characterized by differential expression of gene sets related to biological processes shared with immune cells (Luckey et al., 2006).

We noted that distinct clusters of gene sets showed differential enrichment in specific cell lineages (Figure 2A, clusters 1–5). We characterized these gene set clusters by determining the relative frequencies of genes shared by the gene sets in these clusters (Figure 2B). For example, gene sets in cluster 1, which predominantly distinguished effector and memory T cells from naive T cells, most commonly included genes encoding effector molecules such as granzyme B (*GZMB*), IFN-γ (*IFNG*), as well as Blimp1 (*PRDM1*) and integrin beta 1 (*ITGB1*), and were predominantly derived from expression profiles of effector and memory $CD8^+$ T cells in the context of viral infection and anti-tumoral responses (Table S4, top). Cluster 5, which predominantly distinguished T cells from other cell lineages, included T cell genes such as transcription factors TCF7 and LEF1 as well as components of T cell receptor signaling, CD3ζ (*CD247*), CD3δ (*CD3D*), ZAP70, and Lck and included most gene sets derived from comparing T cells to other immune cell types (Figure 2B and Table S4, bottom). Stem cells showed strong enrichment of gene sets in cluster 2 whose predominant genes play a dominant role in regulating cell cycle (Figure 2B and Table S4, bottom). B cells and dendritic cells were distinguished by a separate cluster of gene sets that included known genes representing those lineages.

We noted that each gene set cluster that distinguished different lineages included both human and mouse gene sets (Figure 2, purple and green bars), suggesting that similar patterns of enrichment were observed using gene sets derived from expression profiles from tissues of both species. Thus, ImmuneSigDB robustly clustered human and mouse immune lineages based on whole transcriptome enrichments of both mouse and human-derived gene sets.

Analogous Cell Types and Contexts in Mice and Humans Show Common Patterns of Gene Expression

We, and others, have previously used GSEA to show that the transcriptional profiles from memory and exhausted $CD8^+$ T cells are highly concordant between mouse and human datasets (Baitsch et al., 2011; Haining et al., 2008; Quigley et al., 2010). To test whether this similarity in gene expression between species is also observed for other cell states we extended this analysis to other cell types and perturbations included in ImmuneSigDB. We focused on four separate biological comparisons (Figure 3) where analysis of gene expression had been made in analogous cell types or perturbations in both human and mouse immune cells. This allowed us to test whether sets of genes differentially expressed in mouse immune cells showed enrichment in profiles from the analogous comparisons in humans and vice versa (Sweet-Cordero et al., 2005).

We first identified 15 studies (6 mouse; 9 human) in which the transcriptional response to lipopolysaccharide (LPS) stimulation had been profiled in myeloid cells; each study had been used to generate a gene set in ImmuneSigDB. We selected one human dataset and generated a ranked list of genes differentially expressed following LPS stimulation. We then performed GSEA using gene sets from the other 14 mouse and human datasets. We found that both human- and mouse-derived gene sets showed highly significant enrichment (FDR < 0.001), suggesting a strong conservation in the transcriptional response to LPS between the two species (Figure 3A). Gene sets derived from studies on

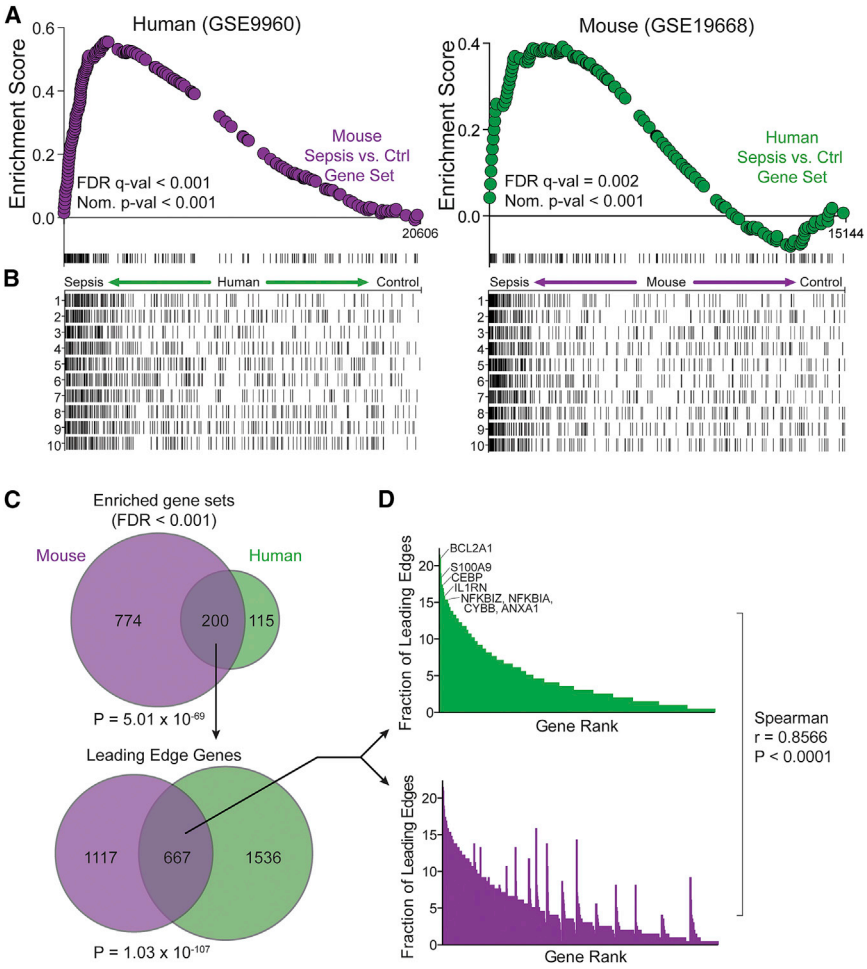


Figure 4. The Transcriptional Response to Sepsis Is Conserved in Humans and Mouse Models

(A and B) GSEA of the set of genes upregulated in mouse sepsis (GSE19668, C57BL/6) in the ranked list of genes upregulated in human sepsis (GSE9960, Gram negative infection) (A, left); and of the corresponding human sepsis gene set enriched in rank ordered list of genes upregulated in mouse sepsis (A, right). Mountain plots indicate cumulative enrichment, and (B) ticks below the line correspond to the position of genes in the ten most enriched gene sets from ImmuneSigDB in the rank order of genes upregulated in sepsis versus control conditions (x axis).

(C) Venn diagram showing overlap in the identity of significantly enriched ImmuneSigDB gene sets in mouse (purple) or human (green) sepsis dataset (top) and the number of shared leading edge genes in the gene sets enriched in both species (bottom). Statistical significance calculated by the hypergeometric test.

(D) Frequency of the genes occurring in the leading edge of a gene sets enriched in human (green) and mouse (purple) sepsis datasets. Statistical significance of the similarity in gene rank is calculated by the Spearman test. See also Figure S4.

human cells tended to show slightly higher enrichment scores than those generated from mouse cells.

We then selected additional cross-species comparisons that were represented by multiple datasets within the ImmuneSigDB. We found cross-species similarity in the gene-expression profiles of comparisons of plasma versus naive B cells, memory B cells versus naive B cells, and regulatory T (Treg) versus conventional T cells (Tconv) (FDR < 0.001, Figure 3B–3D and Table S5). Furthermore, we observed that the biology was not just conserved to the same extent but in some cases mouse-derived gene sets were more strongly enriched in human datasets than other human gene sets, as depicted by the peak height of the respective graphs of their GSEA enrichment scores. These findings indicate that components of the transcriptional signatures of LPS stimulation and some T and B cell differentiation programs are similar in humans and mouse models.

Blood Cells from Human and Mouse Sepsis Share Conserved Biology Reflected in Their Transcriptomes

As we observed common patterns of gene expression in these cross-species comparisons, we next studied more a complex transcriptional dataset from human sepsis and the corresponding mouse models to test whether ImmuneSigDB could resolve similarities or differences between human and mouse transcriptional profiles. Recent studies have analyzed the transcriptional response to sepsis in multiple datasets of gene-expression profiles obtained from human PBMC samples or from mouse models (Seok et al., 2013; Takao and Miyakawa, 2014). However, these studies have differed in their conclusions regarding the degree of similarity between species. We reasoned that analysis with ImmuneSigDB might allow a more detailed analysis of immune signatures elicited by the sepsis response in both species.

We began by using ImmuneSigDB to compare the similarity in gene expression in human and mouse datasets included in the previous studies of the transcriptional responses to sepsis. We selected, at random, a pair of human and mouse studies in which peripheral blood cell gene expression was measured in sepsis versus control conditions (PBMC following sepsis in human [GSE9960] and mouse [GSE19668]) (Ahn et al., 2010; Tang et al., 2009). We first identified genes upregulated following sepsis in each study and tested whether that signature was enriched in the corresponding profile of sepsis versus control in the other species.

We observed strong enrichment of the set of sepsis-induced genes derived from the mouse study in the human dataset (FDR < 0.001, Figure 4A, left). Similarly, we found that a gene set comprising genes upregulated in human PBMC samples in sepsis versus control was strongly enriched in the mouse sepsis gene-expression profile (FDR = 0.002, Figure 4A, right). This internal comparison suggests that there was marked similarity between the genes upregulated by sepsis in humans and in a mouse model.

Next, we identified similarity in gene expression in the sepsis response by testing for enrichment of all gene sets in ImmuneSigDB gene sets in the same pair of human and mouse studies. We compared the ImmuneSigDB gene sets that were significantly enriched in the gene-expression profiles of the human and mouse gram negative and/or positive sepsis response (Figure 4B and Figures S4A–S4D). We observed marked similarity in ImmuneSigDB gene sets that were enriched the sepsis-induced signatures in each species ($p = 5.01 \times 10^{-69}$, Figures 4B and 4C, and Table S6).

To identify which genes in the gene sets that were enriched in both species were "driving" the enrichment of the shared gene sets, we focused on the "leading edge" of enrichment. Leading edge genes in a gene set enrichment analysis are those that contribute most to the enrichment of a particular gene set and include the most significantly upregulated genes in a given gene set. We found that the leading edges of gene sets that were enriched in both species were similar (Spearman $r = 0.857$, $p < 0.0001$; Figure 4D, S4D, and Table S7) indicating that the strong enrichment of shared gene sets is due to the upregulation of similar genes. We found the same results when we performed the same set of analyses using a pair of human and mouse datasets where both were from gram-positive sepsis or when we analyzed gene sets enriched in downregulated genes in sepsis compared to control (Figure S4, Table S7). These data demonstrate a high degree of concordance in gene sets that are enriched following sepsis in humans and mouse models.

Identifying Species-Specific Components of Transcriptional Responses Induced by Sepsis in Human and Mouse

We noted that while many gene sets in ImmuneSigDB were enriched in both species, there were also many gene sets enriched in one species but not the other (Figures 4C and S6). This suggested that in addition to similarities in the sepsis response, there might be species-specific differences in the transcriptional signatures of sepsis. In order to identify the biological basis for these species-specific differences, we devised an analytic approach, termed Leading Edge Metagene (LEM) analysis, to identify main biological "themes" in groups of ImmuneSigDB gene sets enriched in the sepsis datasets. We introduce LEM here and describe it in more detail elsewhere (see Experimental Procedures). LEM analysis is a novel method to identify the groups of co-regulated genes—which we term metagenes—that are highly enriched in multiple gene sets in a comparison of interest (such as sepsis versus control).

For LEM analysis, we first considered all gene sets that were significantly enriched in each dataset of sepsis versus control comparison group (FDR < 0.001). We then filtered the genes in these enriched gene sets to include only leading edge genes (Figure 5A, top and middle). These leading edge genes represented the subset of genes in the group of enriched gene-sets that drive the enrichment score with respect to upregulation in the sepsis phenotype. We then used non-negative matrix factorization (NMF) (Brunet et al., 2004; Lee and Seung, 1999; Lee, 2000; Tamayo et al., 2007) to identify groups of genes that are members of multiple gene sets (Figures S5A and S5B). NMF analysis therefore identifies groups of genes—which we term metagenes—that are members of the leading edge of multiple gene sets that are enriched in the transcriptional response to sepsis (Figure 5A, bottom).

LEM analysis of the gene sets enriched in human sepsis (316 gene sets) and mouse sepsis (974 gene sets) studied in Figure 4 identified three metagenes that were correlated with the sepsis response in each study. Individual metagenes were strongly overrepresented for genes related to distinctive biological processes as annotated by GO terms and Reactome (Ashburner et al., 2000; Croft et al., 2011) (Figure 5B). For instance, in the human sepsis response, we identified a metagene with an overrepresentation of genes involved in mitosis ($p = 4.9 \times 10^{-22}$) such as *CCNA2*, *BUB1*, and *KIF11.* A second metagene was enriched for genes related to phagocytosis ($p = 2.02 \times 10^{-13}$; *LAMP2*, *NCF4*, and *ATPV0B*) and a third metagene was enriched for genes related to inflammation ($p = 3.7 \times 10^{-4}$; *IL1A*, *NFKB1*, and *CCL20*) (Figure 5B). Overlap between the metagene gene memberships and specific GO terms revealed one predominant biological process in each (Figure 5B). However, the while each metagene was significantly enriched for one predominant GO term, only 5%–15% of genes contained in each metagene overlapped with genes in the predominant GO term (Figures S5C–S5F). This suggests that the genes contained in each LEM are related to recognizable biological processes, but that the metagenes represent discrete modules of genes that overlap with but are distinct from GO term categories.

We reasoned that metagenes would provide a more "refined" list of functionally related genes than their parental gene-sets. We therefore tested whether leading edge metagenes were more highly enriched for genes related to biological processes (again as annotated by overlap with GO terms) than their parental gene sets (Figure 5C). We tested the set of three leading edge metagenes for overlap with the collection of GO annotated gene lists, and determined the significance of each GO term's overlap. We compared the p values generated by GO term overlap with the set of genes comprising each metagene with an equivalent number of genes randomly sampled from the original pool of leading edge genes, or from all genes in the genome. We found that the significance of GO term overlap was much higher in the leading edge metagenes than in the original leading edge genes or in a random set of genes. LEM analysis therefore is an effective strategy to both identify major biological processes active in a phenotype of interest and simplify the list of 315 and 974 enriched gene sets in human and mouse, respectively, to a core set of 3 metagenes in each organism that correspond to major biological themes.

We next compared the similarity between metagenes identified in the sepsis response in humans with those in mouse sepsis models. We visualized the pairwise overlap in genes in each metagene using a Circos plot (Figure 5D) and determined the significance of the overlap for each pairwise comparison of mouse and human metagenes (Figures 5E and S6) (Krzywinski et al., 2009). We found striking cross-species similarities for some but not all metagenes. For example, a metagene annotated as "Phagocytic Vesicle" correlated with both the human and mouse sepsis response and contained a very similar set of genes (hypergeometric test $p = 1.09 \times 10^{-31}$, dark blue ribbon, Figure 5D). Similarly there was a highly significant overlap in the metagene annotated as "inflammatory response" in the human dataset and "TRIF-mediated TLR signaling" in the mouse model ($p = 2.79 \times 10^{-23}$).

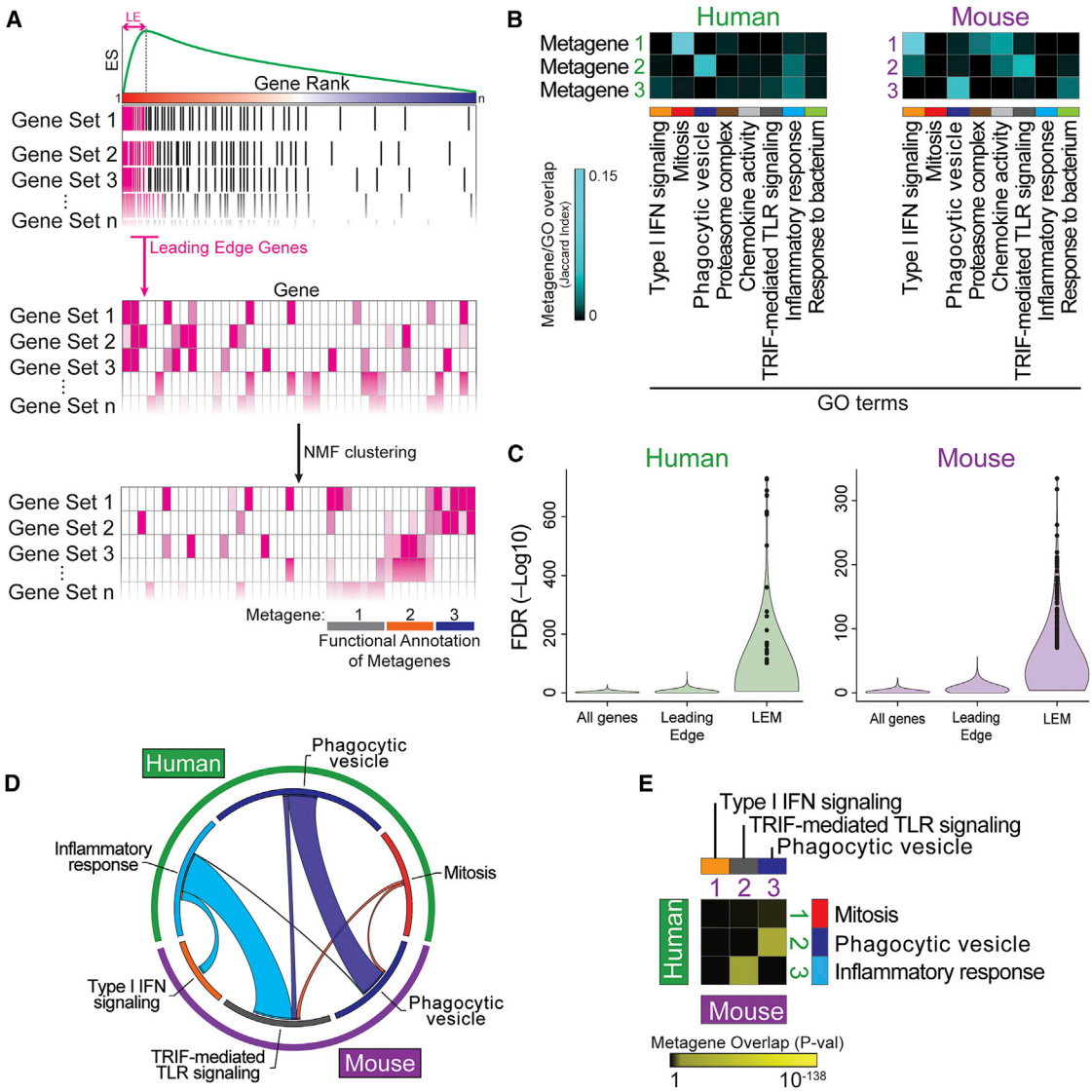


Figure 5. Leading Edge Clustering using Non-Negative Matrix Factorization Identifies Metagenes Representing Distinct Biological Processes

(A) A schematic of the process by which leading edge metagenes are identified.

(B) Biological annotation of metagenes identified in the studies analyzed in Figure 4 generated using GO terms.

(C) Violin plots showing p values of significance of GO Term overlaps with human (left) and mouse (right) sepsis metagenes (LEM), or equivalent-size samples of leading edge genes, or randomly selected genes.

(D) Circos plot of the relative size and overlap of metagenes in mouse (purple, outer segment) and human (green, outer segment) sepsis datasets. Relative number of genes in metagenes is indicated by segment length of the inner circle. Thickness of the ribbon corresponds to the relative number of genes shared between metagenes in the two species.

(E) Heatmap of p values corresponding to significance of overlap in pairwise comparison of metagene gene membership (yellow, highly significant; black, not significant). Statistical significance of the overlap was calculated by hypergeometric test. See also Figures S5 and S6.

However, we also identified metagenes that were not conserved between humans with sepsis and the mouse model. For example, a metagene enriched for genes pertaining to cell cycle ("mitosis" GO term) in humans did not share a corresponding metagene in the mouse model. In the mouse, a type 1 interferon signaling metagene overlapped with very few human metagenes. This analysis approach using ImmuneSigDB suggests that while some biological processes are strongly conserved between these two human and mouse datasets (e.g., phagocytosis, TLR mediated inflammatory response), other biological components are not (e.g., mitosis).

Global Shared and Species-Specific Biological Processes Can Be Identified using ImmuneSigDB and NMF Clustering

We next extended this approach to six datasets of sepsis versus control conditions from three independent studies in humans and from four comparisons in two mouse studies. We identified

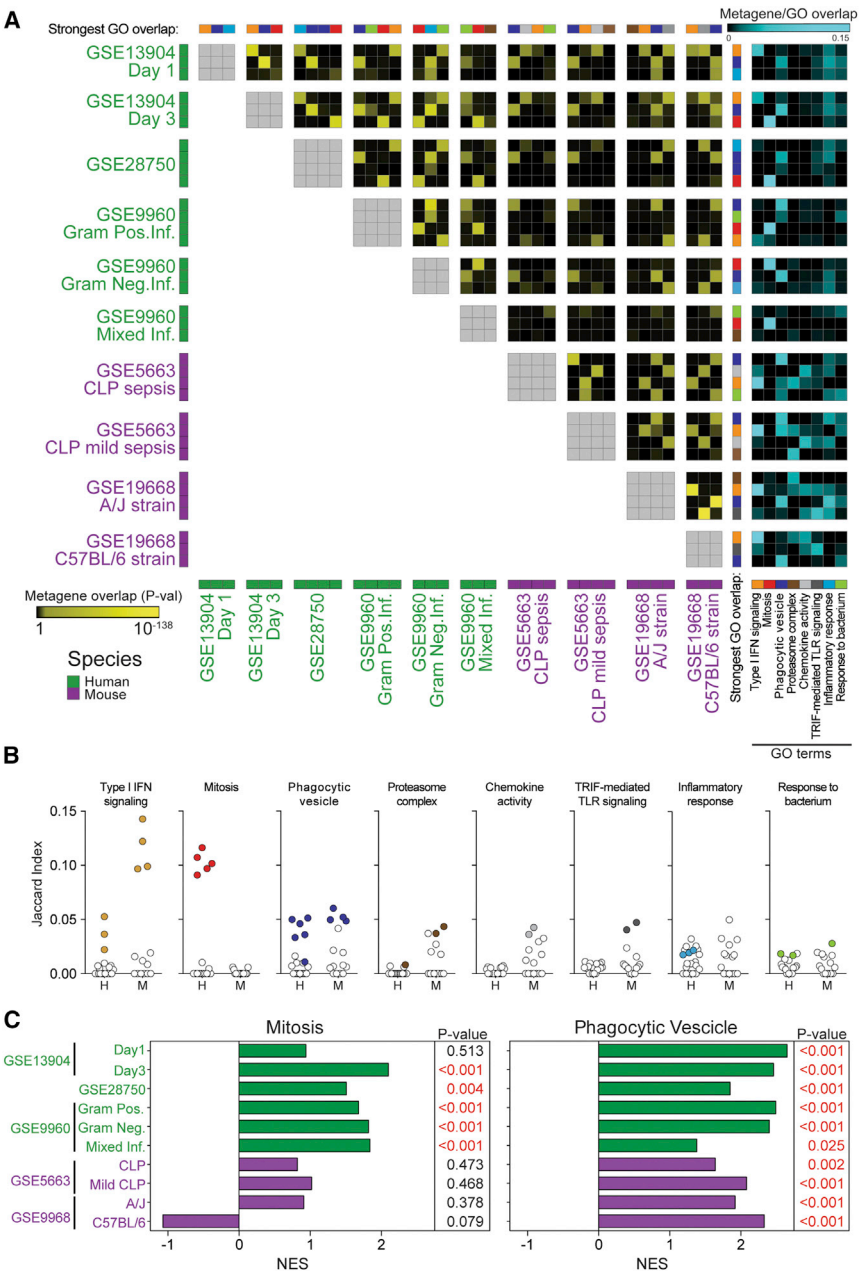


Figure 6. ImmuneSigDB Identifies Shared and Unique Biology in Mouse and Human Sepsis Studies

(A) Pairwise overlaps of all metagenes from mouse (purple bars) and human (green bars) sepsis studies. Heatmap indicates p values corresponding to significance of overlap between each metagene (small squares) in each study (larger squares; yellow, highly significant; black, not significant). The biological annotation of each metagene is based on the significance of enrichment of the GO term indicated (blue, large overlap; black, no overlap) (right). The most significantly enriched GO term annotating each metagene is indicated by the key in lower right.

(B) Jaccard index representing the extent of overlap of metagenes from human (H) and mouse (M) studies. Colored are metagenes that are annotated with the respective biological process as in (A).

(C) Enrichment scores of biological processes that are species-specific (e.g., mitosis, left) or shared (e.g., phagocytic vesicle, right) in the human (green bars) and mouse (purple bars) sepsis datasets. Significance of the enrichment of the named biological process in each dataset is indicated by the p values on the right.

between three or four metagenes in each study providing a total of 35 metagenes present in the collected group of sepsis studies (Figure 6A). We annotated each metagene based on enrichment of GO Terms (Figure 6A, right) and evaluated the significance of pair-wise overlap in the genes present in each metagene.

We found that in almost every study, there was at least one metagene induced by sepsis that showed a highly significant overlap (indicated in yellow in the heatmap, Figure 6A) with metagenes from every other sepsis study, regardless of species of origin. One study that proved an exception was the human study GSE9960, which studied a response to mixed infection, and showed relatively little overlap with any mouse study. However the metagenes identified in that study also showed limited overlap with metagenes from other human studies, suggesting that that transcriptional response contained in that study may represent a different type of biological response to the other human and mouse studies.

In addition to these strongly conserved metagenes, we also found that there were metagenes induced by sepsis that had a striking species-specific distribution. For example, the phagocytic vesicle metagene was either present or strongly overlapped with a metagene present in every dataset, both mouse and human. In contrast, the mitosis metagene was much more specific to human datasets with no significant overlap in mouse studies (Figure 6B). To confirm these results, we tested the significance of enrichment of two GO terms—mitosis and phagocytic vesicle. When looking at the whole-genome transcriptional changes, we indeed observed that mitosis was represented exclusively in human cells in sepsis while phagocytic vesicle process was represented in both species, as we predicted based on the LEM analysis (Figure 6C). These data reveal context-specific transcriptional modules induced by sepsis in humans and mice and also highlight the distinct transcriptional components of the response present in one species but not the other.

DISCUSSION

We analyzed expression profiles from 389 published studies of mouse and human immune cells to generate a collection of curated gene signatures corresponding to cell states and

perturbations in the immune system. This collection of almost 5,000 genes sets contains substantial biological information that was not currently contained in existing collections. We used this new compendium to show that transcriptional signatures induced by LPS stimulation in dendritic cells, and transcriptional programs of T cell and B cell differentiation were highly conserved between humans and mouse models. Moreover, we used ImmuneSigDB to analyze expression profiles from patients and mouse models of sepsis and showed highly significant overlap, suggesting that components of the transcriptional response to sepsis were highly conserved between species. However, we also find that there are substantial species-specific differences, both in enriched gene sets and their component metagenes, in sepsis response signatures, suggesting that not all biological processes induced by sepsis evident at the transcriptional level in humans are present in mouse models and vice versa. These findings suggest that ImmuneSigDB provides a useful tool for detecting subtle patterns of similarity and difference in large-scale datasets of gene expression from cells and tissues in the immune system.

Several studies have directly compared the transcriptional programs in the human and mouse immune systems. We, and others, previously identified conserved patterns of gene expression that change during the differentiation of memory T and B cells, and in exhausted to $CD8^+$ T cells (Baitsch et al., 2011; Haining et al., 2008; Quigley et al., 2010). A recent comparison of gene expression in seven immune cell groups from humans and mice also found a highly significant degree of similarity in global patterns of expression and in the putative transcriptional regulators of these genes (Shay et al., 2013). However, in that study, although the majority of genes showed a pattern of expression that was highly correlated between species, 30%–50% of genes did not show significant correlation between species. Two recent studies of the sepsis datasets analyzed in the present study reached opposite conclusions regarding the degree of similarity between the mouse and human response to sepsis (Seok et al., 2013; Takao and Miyakawa, 2014). Thus, the degree of conservation of transcriptional signatures in the mouse and human immune systems remains controversial (Davis, 2008; Gilad and Mizrahi-Man, 2015; Hackam and Redelmeier, 2006; Lin et al., 2014; Rice, 2012; Shay et al., 2014; van der Worp et al., 2010; Warren et al., 2014).

Our analysis using ImmuneSigDB suggests that there are both conserved and species-specific transcriptional programs induced by sepsis in the immune system. Overall, the transcriptional program shows highly significant similarity between sepsis in the human and mouse. Specifically, analysis of the leading edge metagenes across human (six comparisons) and mouse sepsis datasets (four comparisons) found that in many of the datasets from both species there was coordinate upregulation of metagenes involved in interferon-response and phagocytic processes. This suggests that features of the sepsis response such as interferon release and neutrophilia are shared between species.

However, we also show that many gene sets are enriched in only one species, and metagenes related to mitosis were highly enriched in sepsis-induced profiles in humans but were not significantly enriched in the mouse model. Thus, it is likely that although some components of the sepsis response are highly conserved between species, there is also substantial divergence in the biological processes detected by transcriptional profiling each. Detailed analysis of the transcriptional features of the mouse and human immune systems is therefore required to substantiate conclusions regarding the conservation of a particular biology of interest in two datasets. Whether the differences we observe are due to inherent biological differences between the two species remains unclear. For example, it is possible that the mitotic signature is present in human, but not mouse, because the exact timing of the initiation of activation of immune cells in humans with sepsis is not precisely known and might be more variable compared to tightly controlled, narrow window of induction of sepsis in mouse models.

Our compendium adds to a growing list of collections of transcriptionally co-regulated genes in the immune system. In the human immune system, several studies have identified groups of co-regulated gene modules from expression profiles derived from blood samples representing a range of states of health and disease (Chaussabel et al., 2008; Li et al., 2014). This modular approach to the analysis of gene expression can aid interpretation of gene-expression profiles, increase robustness, and facilitate analyses that span multiple datasets. However, ImmuneSigDB is distinct from those previously described in several respects. Studies by Chaussabel and by Li have focused on identifying collections of genes—termed modules—that tend to vary in expression in a coordinate fashion across a reference set of expression profiles (Chaussabel et al., 2008; Li et al., 2014). Defining modules based on network reconstruction across hundreds or thousands of experimental conditions makes it difficult to associate a particular module with a defined cell state or perturbation that usually results in its up- or downregulation. In contrast, the annotations describing each gene set in the ImmuneSigDB include all the experimental details from a published manuscript, allowing a more transparent connection between gene set and biology. Moreover, ImmuneSigDB was designed for use with GSEA, because each gene set contains either up- or downregulated genes only, rather than a combination of both as can appear in Chaussabel or Li modules, which might limit the use of the latter collections in analyses such as GSEA (Figure S2D, Table S3). Finally, each collection of previously-published modules was defined in a single species (humans), making the generalizability of these compendia to other species hard to predict.

The ImmuneSigDB collection differs in another important respect from previous module collections. The studies by Chaussabel et al. and Li et al. were designed to identify non-overlapping modules of gene expression. However, ImmuneSigDB contains gene sets derived from experimental perturbations that are likely to induce multiple biological processes, each of which might be represented by sub-signatures in a given gene set. Moreover, several gene sets might contain the transcriptional correlate of the same biological processes. For some analytic purposes, it might be useful to have a single gene set that includes the multiple biological processes that are initiated by the complex stimulus such as receptor-ligand engagement or cell differentiation. However, for other applications, such as the analysis which we conducted of the sepsis datasets, a more "atomic" approach might be preferred. We have therefore developed an analytic approach to extract

non-redundant leading edge metagenes from the experimentally derived gene-expression profiles.

Analysis with ImmuneSigDB using GSEA or GSEA combined with a leading-edge metagene analysis might therefore provide the systems immunologist with a useful resource for the analysis of gene expression in the immune system.

EXPERIMENTAL PROCEDURES

ImmuneSigDB Generation

We surveyed the immunology literature and identified published studies that included human or mouse microarray Affymetrix gene expression data in NCBI Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/). We downloaded the corresponding datasets from GEO (Barrett et al., 2007). When available, raw microarray data in the form of the CEL files were normalized by the Robust Multichip Average (RMA) (Bolstad et al., 2003) using justRMA function from the R Bioconductor package affy (version 1.40.0) (Gautier et al., 2004). When CEL files were absent, we downloaded processed expression data from GEO by means of R GEO query package (version 2.28.0) (Davis and Meltzer, 2007). We mapped Affymetrix probe set identifiers to human gene symbols using the Collapse Dataset tool (max probe algorithm) of the GSEA program (Subramanian et al., 2007). We used ortholog gene assignments from Mouse Genome Informatics. The specific mappings were retrieved from the MGI web site on 14 April 2012 and contained 17,827 human-mouse ortholog gene pairings. Phenotype classes were assigned manually according to the original sample annotations and based on review of meaningful biological comparisons (Figure S1). We implemented a pipeline in R, which combined processed microarray data with the phenotype annotations and produced standard formatted files (.gct and .cls) for each comparison as needed.

For each two-class comparison, the genes were ranked according to an information-based similarity metric (RNMI) (Abazeed et al., 2013) from top upregulated to bottom downregulated genes in the two groups. Gene sets comprised genes differentially expressed with an FDR < 0.02, and a maximum number of genes was set at 200 (i.e., all gene sets had at most 200 differentially expressed genes). This way we generated two gene sets from each assigned biological comparison of two groups—"Group_A_vs_Group_B_UP" and "Group_A_vs_Group_B_DN," for the top upregulated and bottom downregulated genes, respectively, identified for the genes most different in the samples in group A compared to the samples in group B. The resource is accessible as the C7 collection at http://www.msigdb.org.

Gene Set Enrichment Analysis

Gene set enrichment analysis (GSEA) was performed as described previously (Mootha et al., 2003; Subramanian et al., 2005). To analyze transcriptional data from Immunological Genome Project (ImmGen) (Heng et al., 2008) and Differentiation Map (DMAP) (Novershtern et al., 2011), we used single sample GSEA (ssGSEA) as described previously (Barbie et al., 2009; Reich et al., 2006), to create a matrix in which columns represented individual samples and rows corresponded to gene sets, and the values represented the single sample ssGSEA score of each gene set in each sample. We averaged the biological replicates and filtered this matrix to include only the top 10% of gene sets based on mean absolute deviation (MAD) across sample types and bi-clustered using 1-Pearson Correlation.

Leading Edge Metagene Analysis

We developed an approach to identify groups of genes—termed leading edge metagenes (LEM)—that are both associated with a phenotype of interest and shared between multiple gene sets enriched in that phenotypic comparison (Y.T., unpublished data). We reasoned that groups of genes that are co-regulated in the phenotype of interest and also present in multiple gene sets are likely to represent the core sub-signatures of genes related to distinct biological processes or pathways. Our approach leverages the notion of the leading edge genes in a GSEA analysis, which are the genes whose expression profile is most highly correlated with the phenotype distinction in a comparison of biological states and thus drives the GSEA enrichment statistic. LEM analysis identifies groups of genes (metagenes) that are common to multiple gene sets returned in a GSEA result, and strongly correlated with the phenotype of interest.

First we perform GSEA using the ImmuneSigDB in a two-class comparison of interest (e.g., sepsis versus control). GSEA yields an enrichment score to quantify the overrepresentation of a gene set (e.g., genes coordinately up- or downregulated in previous experiments) at the top or bottom of a ranked list of genes (e.g., generated by differential expression of in a comparison of interest). The leading edge of each enriched gene set is defined as the subset of genes with positive contribution to the enrichment score before it reaches its peak; i.e., those that are most correlated with the phenotype of interest.

We then consolidate the leading-edges of the m top-scoring gene sets into a sparse n by m matrix M, where the number of rows is the cardinality of the union of genes from all the leading-edges in the m top gene sets, and the columns correspond to the genes in the m enriched gene sets. The value of each entry in the matrix is the signal to noise ratio of the corresponding gene between two conditions in comparison (Equation 1) and 0 if the gene is not in the leading edge of that gene set. A large signal to noise ratio indicates a significant difference in expressions of the corresponding genes between the two conditions.

$$s2n = \frac{\mu_A - \mu_B}{\sigma_A - \sigma_B} \quad (1)$$

To identify clusters in this matrix, we use non-matrix factorization (Brunet et al., 2004; Lee and Seung, 1999; Lee and Seung, 2000; Tamayo et al., 2007) to yield two matrices, W and H. W matrix is a low-dimensional representation of the M matrix and each dimension of W is a linear combination of n genes, called a metagene. The entries in the H matrix represent the quantity of each metagene required to reconstruct each of the M gene sets. The coefficient in W matrix can be viewed as the contribution of each gene to the corresponding metagene. Inspection of the W matrix shows that in each metagene, the coefficients of most genes are usually very small, and only a small number of genes have a coefficient significantly larger than 0. As each metagene is a positive linear combination of all the genes, a small coefficient indicates negligible contribution to the metagene. Thus the next step of our algorithm is to filter out genes with small coefficients in each metagene. To do that, we first assume that the background distribution of coefficients fulfills an exponential distribution. We set a filtering threshold at the 95% quantile of the fitted exponential distribution and set all coefficients below this to zero.

Because each gene can contribute to more than one metagene we next need to assign each gene to a single metagene. The assignment of genes to metagenes uses the following rules: (1) if one gene has no contribution to any of the metagenes, then it will be defined as not in any metagene, and (2) each gene with a coefficient above the threshold (defined above) will be assigned to the metagene where it has the largest coefficient. Each metagene is annotated with a biological "theme" based on the Jaccard overlap of its constituent genes with GO categories (Ashburner et al., 2000).

SUPPLEMENTAL INFORMATION

Supplemental Information includes six figures and seven tables and can be found with this article online at http://dx.doi.org/10.1016/j.immuni.2015.12.006.

AUTHOR CONTRIBUTIONS

Conceptualization, J.G., J.P.M., and W.N.H.; Methodology, J.G., Y.T., A.L., P.T., J.P.M., and W.N.H.; Software, Y.T. and A.L.; Formal analysis, J.G., A.L., and P.T.; Investigation, J.G., A.L., and P.T.; Data curation, J.G. and A.L.; Writing of original draft, J.G., J.P.M., and W.N.H.; Review and editing, J.G., A.J.B., S.B., J.P.M., and W.N.H.; Visualization, J.G. and W.N.H.; Funding acquisition, J.P.M., A.J.B., and W.N.H.; Project administration, W.N.H.; Supervision, W.N.H. and J.P.M.

ACKNOWLEDGMENTS

We would like to thank Arlene H. Sharpe, William Kim, and Aravind Subramanian for useful discussions. This work was supported by an Infrastructure and Opportunity Grant from the Human Immunology Project Consortium

(to A.J.B., J.P.M., and W.N.H.), by U19 AI090023 to W.N.H.; NIAID Bioinformatics Support Contract HHSN272201200028C to A.J.B.; R01CA154480, R01CA121941, R01GM074024, and U54CA112962 to P.T. and J.P.M.; and the Cancer Research Institute Predoctoral Emphasis Pathway in Tumor Immunology to J.G.

Received: April 13, 2015
Revised: August 2, 2015
Accepted: September 30, 2015
Published: January 12, 2016

REFERENCES

Abazeed, M.E., Adams, D.J., Hurov, K.E., Tamayo, P., Creighton, C.J., Sonkin, D., Giacomelli, A.O., Du, C., Fries, D.F., Wong, K.K., et al. (2013). Integrative radiogenomic profiling of squamous cell lung cancer. Cancer Res. *73*, 6289–6298.

Ahn, S.H., Deshmukh, H., Johnson, N., Cowell, L.G., Rude, T.H., Scott, W.K., Nelson, C.L., Zaas, A.K., Marchuk, D.A., Keum, S., et al. (2010). Two genes on A/J chromosome 18 are associated with susceptibility to Staphylococcus aureus infection by combined microarray and QTL analyses. PLoS Pathog. *6*, e1001088.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.; The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

Baitsch, L., Baumgaertner, P., Devêvre, E., Raghav, S.K., Legat, A., Barba, L., Wieckowski, S., Bouzourene, H., Deplancke, B., Romero, P., et al. (2011). Exhaustion of tumor-specific $CD8^+$ T cells in metastases from melanoma patients. J. Clin. Invest. *121*, 2350–2360.

Barbie, D.A., Tamayo, P., Boehm, J.S., Kim, S.Y., Moody, S.E., Dunn, I.F., Schinzel, A.C., Sandy, P., Meylan, E., Scholl, C., et al. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature *462*, 108–112.

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., and Edgar, R. (2007). NCBI GEO: mining tens of millions of expression profiles–database and tools update. Nucleic Acids Res. *35*, D760–D765.

Bolstad, B.M., Irizarry, R.A., Astrand, M., and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics *19*, 185–193.

Brunet, J.P., Tamayo, P., Golub, T.R., and Mesirov, J.P. (2004). Metagenes and molecular pattern discovery using matrix factorization. Proc. Natl. Acad. Sci. USA *101*, 4164–4169.

Chaussabel, D., Quinn, C., Shen, J., Patel, P., Glaser, C., Baldwin, N., Stichweh, D., Blankenship, D., Li, L., Munagala, I., et al. (2008). A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. Immunity *29*, 150–164.

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., et al. (2011). Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. *39*, D691–D697.

Davis, M.M. (2008). A prescription for human immunology. Immunity *29*, 835–838.

Davis, S., and Meltzer, P.S. (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics *23*, 1846–1847.

Gautier, L., Cope, L., Bolstad, B.M., and Irizarry, R.A. (2004). affy–analysis of Affymetrix GeneChip data at the probe level. Bioinformatics *20*, 307–315.

Gilad, Y., and Mizrahi-Man, O. (2015). A reanalysis of mouse ENCODE comparative gene expression data. F1000Research *4*, http://dx.doi.org/10.12688/f1000research.6536.1.

Hackam, D.G., and Redelmeier, D.A. (2006). Translation of research evidence from animals to humans. JAMA *296*, 1731–1732.

Haining, W.N., and Wherry, E.J. (2010). Integrating genomic signatures for immunologic discovery. Immunity *32*, 152–161.

Haining, W.N., Ebert, B.L., Subrmanian, A., Wherry, E.J., Eichbaum, Q., Evans, J.W., Mak, R., Rivoli, S., Pretz, J., Angelosanto, J., et al. (2008). Identification of an evolutionarily conserved transcriptional signature of CD8 memory differentiation that is shared by T and B cells. J. Immunol. *181*, 1859–1868.

Heng, T.S., and Painter, M.W.; Immunological Genome Project Consortium (2008). The Immunological Genome Project: networks of gene expression in immune cells. Nat. Immunol. *9*, 1091–1094.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res. *19*, 1639–1645.

Lee, D.D., and Seung, H.S. (2000). Algorithms for Non-negative Matrix Factorization. In Proceedings of Neural Information Processing Systems *13*, 556–562.

Lee, D.D., and Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature *401*, 788–791.

Li, S., Rouphael, N., Duraisingham, S., Romero-Steiner, S., Presnell, S., Davis, C., Schmidt, D.S., Johnson, S.E., Milton, A., Rajam, G., et al. (2014). Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. Nat. Immunol. *15*, 195–204.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739–1740.

Lin, S., Lin, Y., Nery, J.R., Urich, M.A., Breschi, A., Davis, C.A., Dobin, A., Zaleski, C., Beer, M.A., Chapman, W.C., et al. (2014). Comparison of the transcriptional landscapes between human and mouse tissues. Proc. Natl. Acad. Sci. USA *111*, 17224–17229.

Luckey, C.J., Bhattacharya, D., Goldrath, A.W., Weissman, I.L., Benoist, C., and Mathis, D. (2006). Memory T and memory B cells share a transcriptional program of self-renewal with long-term hematopoietic stem cells. Proc. Natl. Acad. Sci. USA *103*, 3304–3309.

Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., et al. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat. Genet. *34*, 267–273.

Novershtern, N., Subramanian, A., Lawton, L.N., Mak, R.H., Haining, W.N., McConkey, M.E., Habib, N., Yosef, N., Chang, C.Y., Shay, T., et al. (2011). Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell *144*, 296–309.

Quigley, M., Pereyra, F., Nilsson, B., Porichis, F., Fonseca, C., Eichbaum, Q., Julg, B., Jesneck, J.L., Brosnahan, K., Imam, S., et al. (2010). Transcriptional analysis of HIV-specific CD8+ T cells shows that PD-1 inhibits T cell function by upregulating BATF. Nat. Med. *16*, 1147–1151.

Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., and Mesirov, J.P. (2006). GenePattern 2.0. Nat. Genet. *38*, 500–501.

Rice, J. (2012). Animal models: Not close enough. Nature *484*, S9.

Seok, J., Warren, H.S., Cuenca, A.G., Mindrinos, M.N., Baker, H.V., Xu, W., Richards, D.R., McDonald-Smith, G.P., Gao, H., Hennessy, L., et al.; Inflammation and Host Response to Injury, Large Scale Collaborative Research Program (2013). Genomic responses in mouse models poorly mimic human inflammatory diseases. Proc. Natl. Acad. Sci. USA *110*, 3507–3512.

Shay, T., Jojic, V., Zuk, O., Rothamel, K., Puyraimond-Zemmour, D., Feng, T., Wakamatsu, E., Benoist, C., Koller, D., and Regev, A.; ImmGen Consortium (2013). Conservation and divergence in the transcriptional programs of the human and mouse immune systems. Proc. Natl. Acad. Sci. USA *110*, 2946–2951.

Shay, T., Lederer, J.A., and Benoist, C. (2014). Genomic responses to inflammation in mouse models mimic humans: We concur, apples to oranges comparisons won't do. Proc. Natl. Acad. Sci. USA *112*, E346.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

Subramanian, A., Kuehn, H., Gould, J., Tamayo, P., and Mesirov, J.P. (2007). GSEA-P: a desktop application for Gene Set Enrichment Analysis. Bioinformatics *23*, 3251–3253.

Sweet-Cordero, A., Mukherjee, S., Subramanian, A., You, H., Roix, J.J., Ladd-Acosta, C., Mesirov, J., Golub, T.R., and Jacks, T. (2005). An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. Nat. Genet. *37*, 48–55.

Takao, K., and Miyakawa, T. (2014). Genomic responses in mouse models greatly mimic human inflammatory diseases. Proc. Natl. Acad. Sci. USA *112*, 1167–1172.

Tamayo, P., Scanfeld, D., Ebert, B.L., Gillette, M.A., Roberts, C.W., and Mesirov, J.P. (2007). Metagene projection for cross-platform, cross-species characterization of global transcriptional states. Proc. Natl. Acad. Sci. USA *104*, 5959–5964.

Tang, B.M., McLean, A.S., Dawes, I.W., Huang, S.J., and Lin, R.C. (2009). Gene-expression profiling of peripheral blood mononuclear cells in sepsis. Crit. Care Med. *37*, 882–888.

van der Worp, H.B., Howells, D.W., Sena, E.S., Porritt, M.J., Rewell, S., O'Collins, V., and Macleod, M.R. (2010). Can animal models of disease reliably inform human studies? PLoS Med. *7*, e1000245.

Warren, H.S., Tompkins, R.G., Moldawer, L.L., Seok, J., Xu, W., Mindrinos, M.N., Maier, R.V., Xiao, W., and Davis, R.W. (2014). Mice are not men. Proc. Natl. Acad. Sci. USA *112*, E345.

Immunity

Supplemental Information

Compendium of Immune Signatures Identifies Conserved and Species-Specific Biology in Response to Inflammation

Jernej Godec, Yan Tan, Arthur Liberzon, Pablo Tamayo, Sanchita Bhattacharya, Atul J. Butte, Jill P. Mesirov, W. Nicholas Haining

Supplemental Figure Legends

Figure S1, related to Figure 1. Selection of biologically meaningful comparisons. Schematic of an example of comparisons that are biologically meaningful (black arrows) or difficult to interpret (red arrows) in one of the studies included in the ImmuneSigDB (GSE17721).

Figure S2, related to Figure 1. ImmuneSigDB largely contains unique gene sets not represented in previous immune signature collections. (A) Left, overlap in gene set membership within ImmuneSigDB. Heatmap indicates significance of overlap indicated by hypergeometric *P*-values. Highlighted lineages or biological process are found in each of the significantly overlapping clusters. Right, summary of hypergeometric *P*-value ranges of all pairwise overlaps on the left. (B) Overlap in gene set membership between ImmuneSigDB and gene signature collection developed by Li et al. Heatmap indicates significance of overlap indicated by FDR values (right). Highlighted are representative biological processes or cell lineages in each of the significantly overlapping clusters of gene sets/modules. Summary of FDR ranges of all pairwise overlaps are shown to the right. (C and D) Analysis as in (B) showing overlapping gene memberships between immune gene modules defined by Chaussabel et al. and ImmuneSigDB (C) and between the Li et al. and Chaussabel et al. collections (D). (E) Significance of enrichment of gene sets contained in ImmuneSigDB or modules in collections by Li et al or Chaussabel et al. Each of the three collections was used to analyze four publically available datasets of gene expression datasets from LPS vs. unstimluated DC, Tregs vs. conventional $CD4^+$ T cells, memory B cells vs. naive B cells, and plasma B cells vs. naive B cells.

Figure S3, related to Figure 2. Human hematopoietic lineages are accurately clustered using ImmuneSigDB enrichments. (A) Unsupervised bi-clustering of ssGSEA values using ImmuneSigDB in cell subsets represented in DMAP. Clustering was performed in gene sets with the top 10% highest mean absolute deviation.

Figure S4, related to Figure 4. Human and mouse blood cells in gram positive sepsis undergo similar transcriptional response. (A and B) GSEA of gene sets of genes up-regulated in Gram positive human (GSE9960) or mouse (GSE19668, C57BL/6) in the opposite species – mouse gene set enriched in the human dataset (A) and the human gene set enriched in mouse dataset (B). Mountain plots indicate cumulative enrichment, and ticks below the line correspond to the position of respective gene set genes up-regulated in sepsis versus control conditions. (C) Number of significantly enriched gene sets mouse (purple) or human (green) dataset. Statistical significance was calculated by the hypergeometric test. (D) Distribution of the number of gene sets including the shared genes in the leading edges of common enriched gene of mouse and human sepsis dataset enrichments. Statistical significance is calculated by the Spearman test. (E - H) Analysis was done as above using GSEA of gene sets of down-regulated in Gram negative bacteria-induced sepsis in human (GSE9960) or gram positive sepsis mouse (GSE19668, C57BL/6) in the opposite species – mouse gene set enriched in the human dataset (left) and the human gene set enriched in mouse dataset (right).

Figure S5, related to Figure 5. LEM analysis identifies commonly represented and co-regulated metagenes in immune biological processes that are unique, yet overlapping with Gene Ontology Terms. Heatmap representing the sparse matrix of

all leading edge genes (FDR<0.001) in GSEA analysis of human (A) and mouse sepsis (B) as in Figure 5. Gene sets were clustered using Pearson correlation and genes ordered based on their LEM membership, annotated above. (C and D) Overlap in the genes contained in LEMs defined for human (C) and mouse (D) sepsis described in Figure 5 and the predominant GO terms. Numbers of genes are indicated in the Venn diagrams, and the statistical significance of each overlap was assessed using hypergeometric test in the space of 20606 (C) and 15183 (D) total genes, based on annotated unique genes or human orthologs. (E and F) Fraction of LEM genes that are contained in the predominant GO term for human (E) and mouse (F) studies.

Figure S6, related to Figure 5. Illustration describing metagene overlap representation. (A) Venn diagrams representing absolute gene overlap of each human metagene with each mouse metagene. Numbers represent the number of unique and overlapping genes in each comparison. Statistical significance was assessed using Hypergeometric text in the space of all shared genes from the two datasets (n=12,634). The relative overlap of each human metagene with each mouse metagene is represented in Circos plot (B) and the significance of each overlap is represented in a heat map form (C).

Supplemental Table Legends

Table S1, related to Figure 1. Left, the list of scientific journals prioritized for selection of immunological studies included in the ImmuneSigDB. Right, Affymetrix gene expression microarray platforms that studies contained in ImmuneSigDB originate from.

Table S2, related to Figure 1 and S2. Comparison of features in ImmuneSigDB collection, Chaussabel et al. modules, and Li et al. gene set collection.

Table S3, related to Figure 1 and S2E. Top 20 gene sets from ImmuneSigDB, Li, and Chaussabel enriched in LPS-stimulated dendritic cells relative to unstimulated cells (GSE14000), in regulatory T cells (Treg) compared to conventional $CD4^+$ T cells (Tconv) (GSE25087), enriched in peripheral blood plasma cells compared to naive B cells (GSE22886), and in peripheral blood memory B cells compared to naive B cells (GSE42724). Represented are the names and associated description, number of genes in a gene set, statistical significance (false discovery rate, FDR), as well as tissue source, species of origin, and PubMed ID (PMID) of the study where the gene set originates.

Table S4, related to Figure 2. Top, ImmuneSigDB gene sets defining each of the clusters in Figure 2 (rows). Bottom, quantification and relative representation of genes across ImmuneSigDB gene set clusters defined in Figure 2.

Table S5, related to Figure 3. Statistical significance of enrichment of human and mouse gene sets derived from studies profiling LPS-stimulated versus unstimulated myeloid cells, plasma cells versus naive B cells, regulatory (Treg) versus conventional

(Tconv) CD4+ T cells, and memory versus naive B cells; and enriched in a randomly chosen human study examining the same phenotype as described in Figure 3A, 3B, 3C, 3D, respectively.

Table S6, related to Figure 4B. Names of ImmuneSigDB gene sets highly enriched in both mouse and human sepsis shown in Figure 4B.

Table S7, related to Figures 4D, S4D, S4H. Lists of genes shown in Figure 4D, S4D, S4H. The shared leading edge genes are ranked based on their relative abundance in human and mouse studies. Listed are the rank in the respective species, the relative abundance, and the number of times a particular gene appears in the leading edge matrix of human and mouse GSEA.

Figure S1

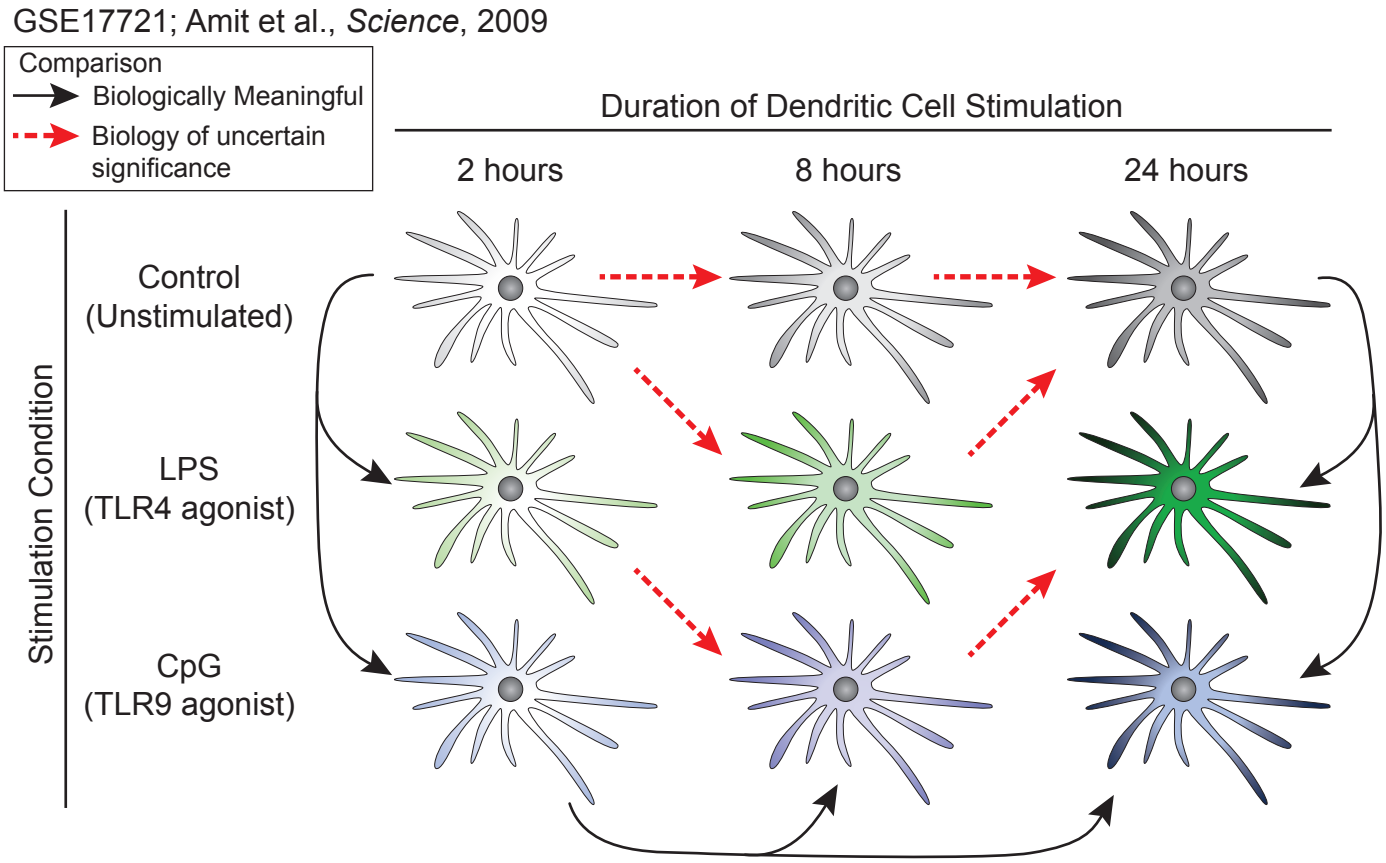

GSE17721; Amit et al., *Science*, 2009
Comparison
Biologically Meaningful
Biology of uncertain significance
Duration of Dendritic Cell Stimulation
2 hours
8 hours
24 hours
Stimulation Condition
Control (Unstimulated)
LPS (TLR4 agonist)
CpG (TLR9 agonist)

Figure S2

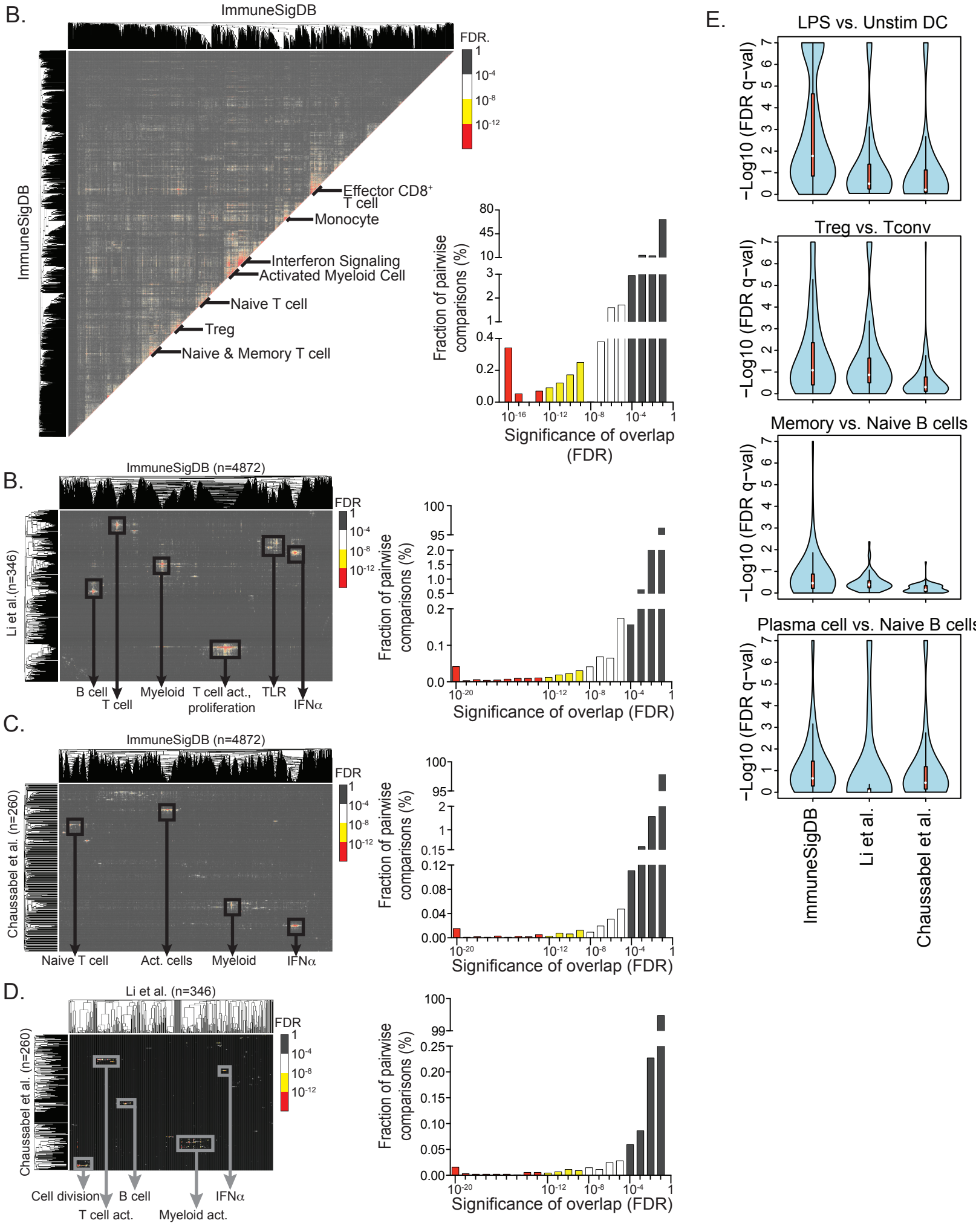

B.
ImmuneSigDB
FDR.
1
10-4
10-8
10-12
ImmuneSigDB
Effector CD8+ T cell
Monocyte
Interferon Signaling
Activated Myeloid Cell
Naive T cell
Treg
Naive & Memory T cell
Fraction of pairwise comparisons (%)
80
45
10
3
2
1
0.4
0.2
0.0
10-16
10-12
10-8
10-4
1
Significance of overlap (FDR)
B.
ImmuneSigDB (n=4872)
FDR
Li et al.(n=346)
B cell
T cell
Myeloid
T cell act., proliferation
TLR
IFNα
Fraction of pairwise comparisons (%)
100
95
2.0
1.5
1.0
0.5
0.2
0.1
0.0
10-20
10-12
10-8
10-4
1
Significance of overlap (FDR)
C.
ImmuneSigDB (n=4872)
FDR
Chaussabel et al. (n=260)
Naive T cell
Act. cells
Myeloid
IFNα
Fraction of pairwise comparisons (%)
100
95
2
1
0.15
0.12
0.08
0.04
0.00
10-20
10-12
10-8
10-4
1
Significance of overlap (FDR)
D.
Li et al. (n=346)
FDR
Chaussabel et al. (n=260)
Cell division
B cell
T cell act.
Myeloid act.
IFNα
Fraction of pairwise comparisons (%)
100
99
0.25
0.20
0.15
0.10
0.05
0.00
10-20
10-12
10-8
10-4
1
Significance of overlap (FDR)
E.
LPS vs. Unstim DC
Treg vs. Tconv
Memory vs. Naive B cells
Plasma cell vs. Naive B cells
−Log10 (FDR q−val)
0 1 2 3 4 5 6 7
ImmuneSigDB
Li et al.
Chaussabel et al.

Figure S3

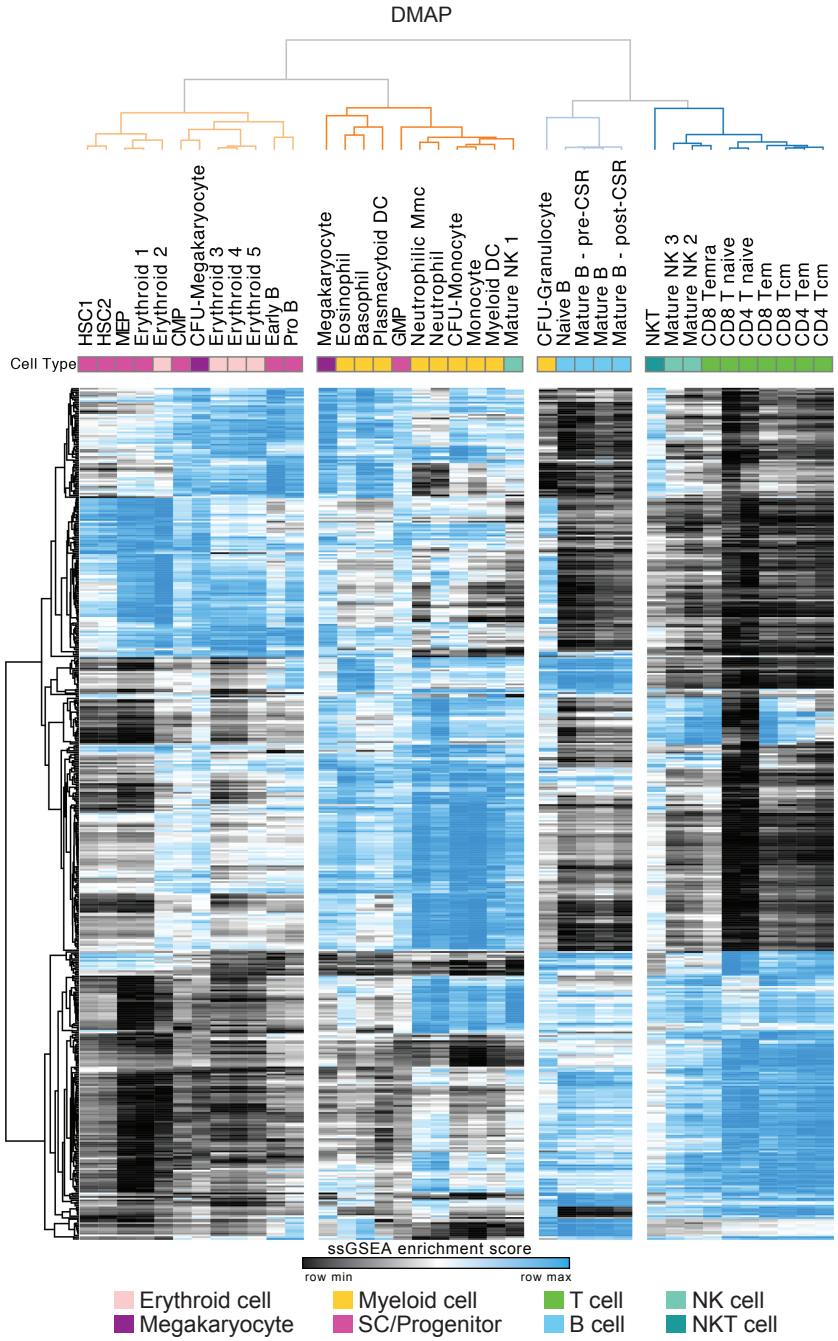

DMAP
HSC1
HSC2
MEP
Erythroid 1
Erythroid 2
CMP
CFU-Megakaryocyte
Erythroid 3
Erythroid 4
Erythroid 5
Early B
Pro B
Megakaryocyte
Eosinophil
Basophil
Plasmacytoid DC
GMP
Neutrophilic Mmc
Neutrophil
CFU-Monocyte
Monocyte
Myeloid DC
Mature NK 1
CFU-Granulocyte
Naive B
Mature B - pre-CSR
Mature B
Mature B - post-CSR
NKT
Mature NK 3
Mature NK 2
CD8 Temra
CD8 T naive
CD4 T naive
CD8 Tem
CD8 Tcm
CD4 Tem
CD4 Tcm
Cell Type
ssGSEA enrichment score
row min
row max
Erythroid cell
Megakaryocyte
Myeloid cell
SC/Progenitor
T cell
B cell
NK cell
NKT cell

Figure S4

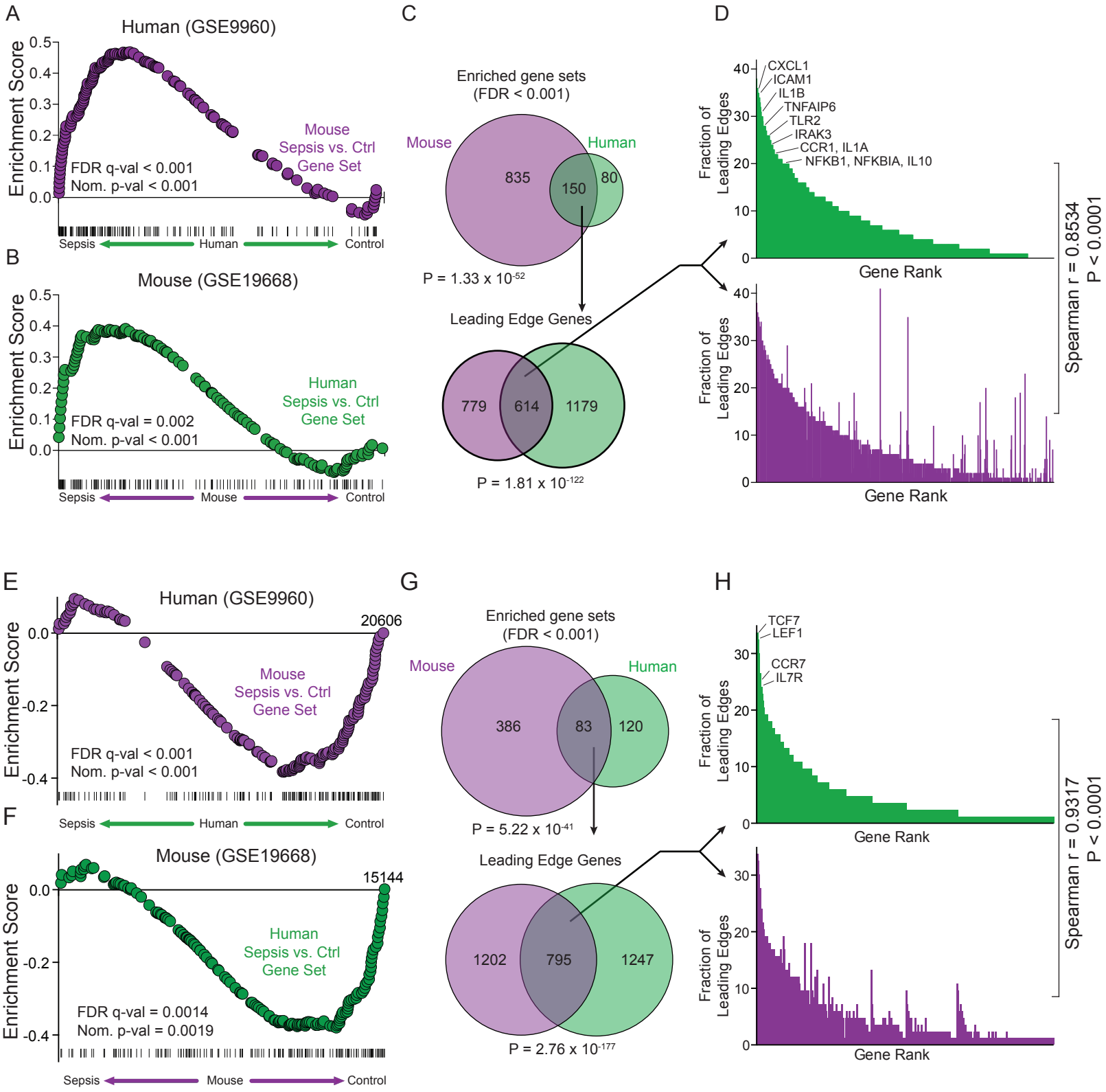

A
Human (GSE9960)
Enrichment Score
FDR q-val < 0.001
Nom. p-val < 0.001
Mouse
Sepsis vs. Ctrl
Gene Set
Sepsis
Human
Control
B
Mouse (GSE19668)
FDR q-val = 0.002
Nom. p-val < 0.001
Human
Sepsis vs. Ctrl
Gene Set
Sepsis
Mouse
Control
C
Enriched gene sets
(FDR < 0.001)
Mouse
Human
835
150
80
P = 1.33 x 10-52
Leading Edge Genes
779
614
1179
P = 1.81 x 10-122
D
CXCL1
ICAM1
IL1B
TNFAIP6
TLR2
IRAK3
CCR1, IL1A
NFKB1, NFKBIA, IL10
Fraction of Leading Edges
Gene Rank
Spearman r = 0.8534
P < 0.0001
E
Human (GSE9960)
20606
Mouse
Sepsis vs. Ctrl
Gene Set
FDR q-val < 0.001
Nom. p-val < 0.001
Sepsis
Human
Control
F
Mouse (GSE19668)
15144
Human
Sepsis vs. Ctrl
Gene Set
FDR q-val = 0.0014
Nom. p-val = 0.0019
Sepsis
Mouse
Control
G
Enriched gene sets
(FDR < 0.001)
Mouse
Human
386
83
120
P = 5.22 x 10-41
Leading Edge Genes
1202
795
1247
P = 2.76 x 10-177
H
TCF7
LEF1
CCR7
IL7R
Fraction of Leading Edges
Gene Rank
Spearman r = 0.9317
P < 0.0001

Figure S5

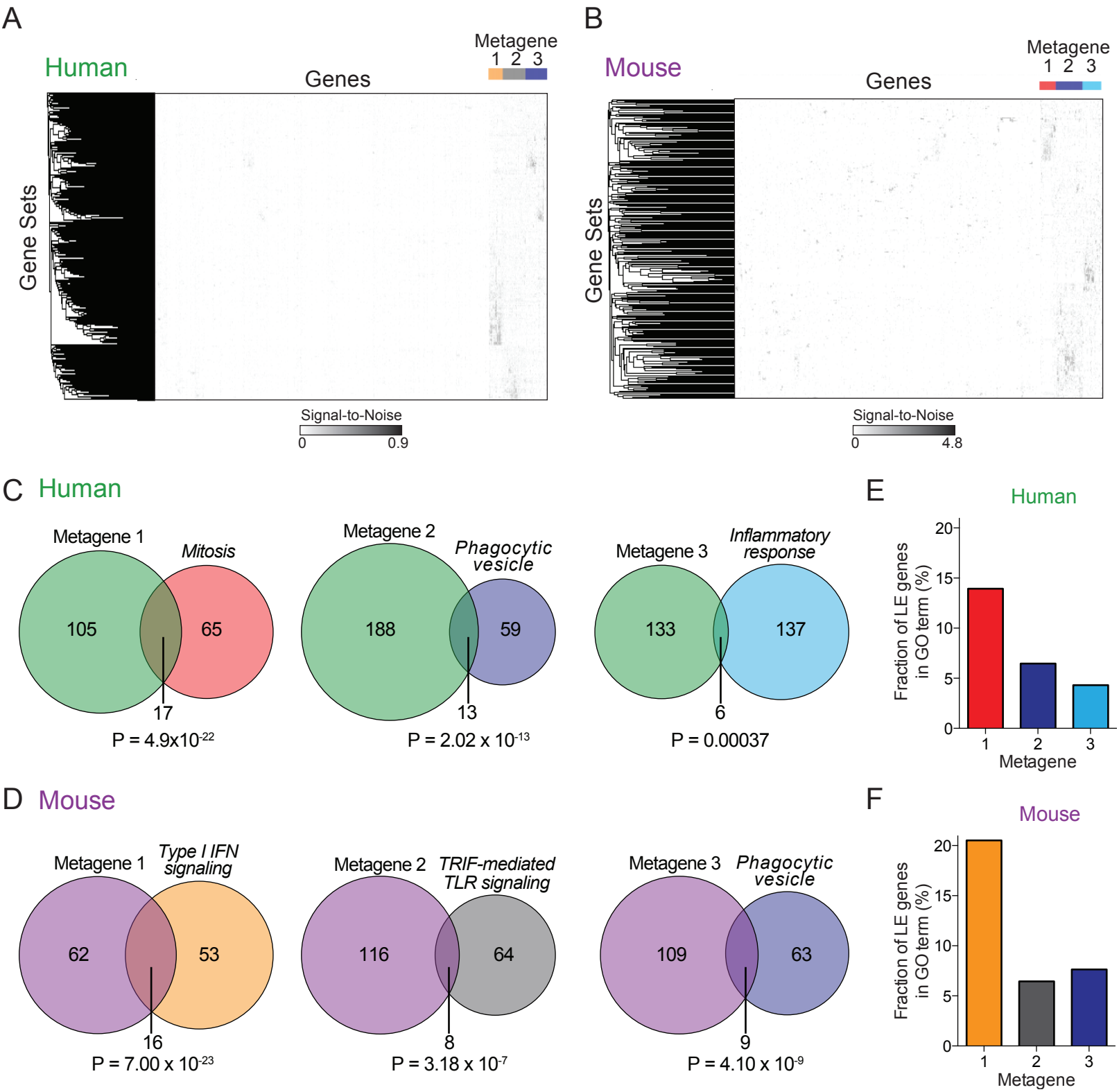

A
Human
Metagene
1 2 3
Genes
Gene Sets
Signal-to-Noise
0
0.9
B
Mouse
Metagene
1 2 3
Genes
Gene Sets
Signal-to-Noise
0
4.8
C Human
Metagene 1
Mitosis
105
65
17
P = 4.9x10-22
Metagene 2
Phagocytic vesicle
188
59
13
P = 2.02 x 10-13
Metagene 3
Inflammatory response
133
137
6
P = 0.00037
D Mouse
Metagene 1
Type I IFN signaling
62
53
16
P = 7.00 x 10-23
Metagene 2
TRIF-mediated TLR signaling
116
64
8
P = 3.18 x 10-7
Metagene 3
Phagocytic vesicle
109
63
9
P = 4.10 x 10-9
E
Human
Fraction of LE genes in GO term (%)
20
15
10
5
0
1
2
3
Metagene
F
Mouse
Fraction of LE genes in GO term (%)
20
15
10
5
0
1
2
3
Metagene

Figure S6

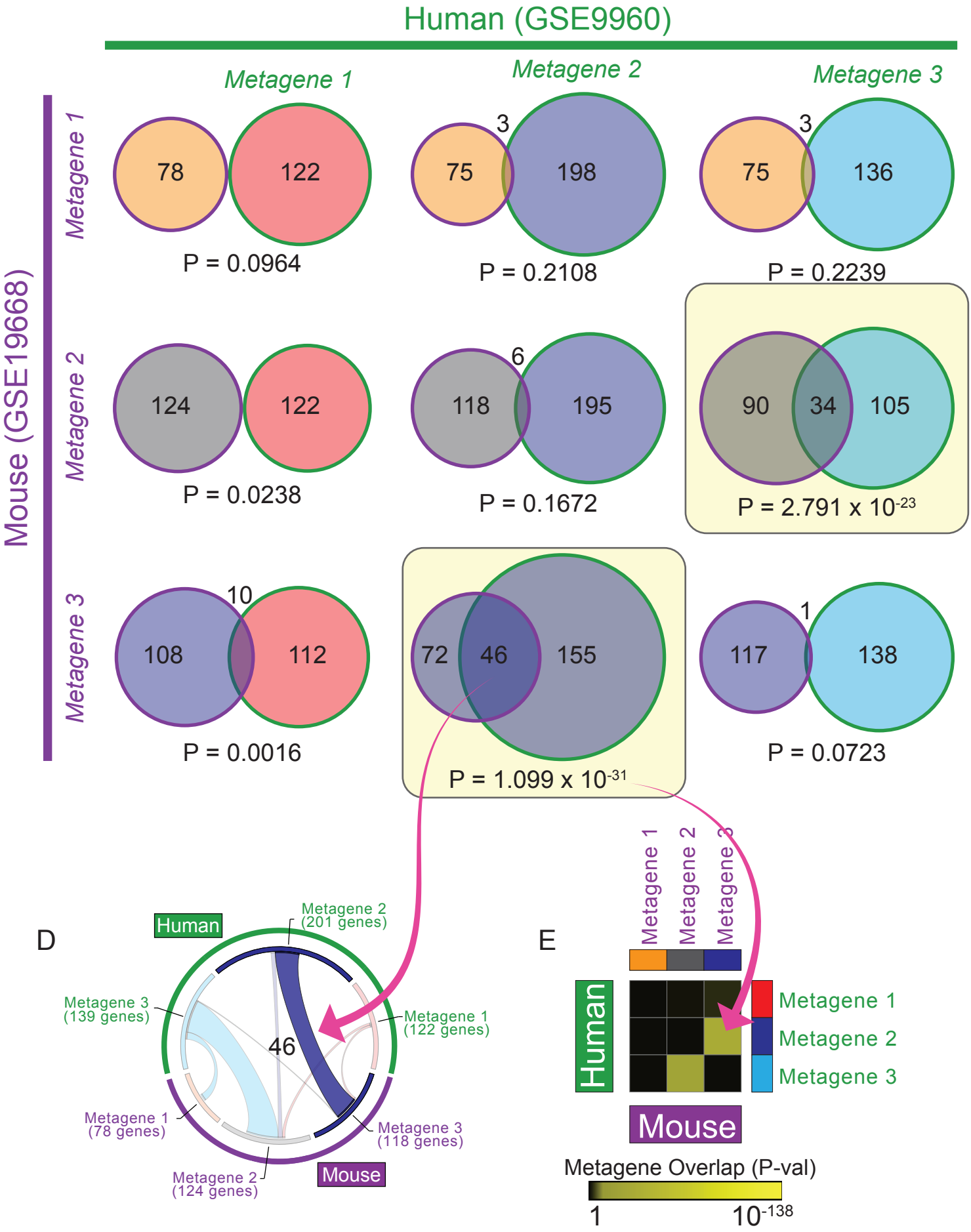

Human (GSE9960)
Metagene 1
Metagene 2
Metagene 3
Mouse (GSE19668)
Metagene 1
Metagene 2
Metagene 3
78
122
P = 0.0964
3
75
198
P = 0.2108
3
75
136
P = 0.2239
124
122
P = 0.0238
6
118
195
P = 0.1672
90
34
105
P = 2.791 x 10-23
10
108
112
P = 0.0016
72
46
155
P = 1.099 x 10-31
1
117
138
P = 0.0723
D
Human
Metagene 2
(201 genes)
Metagene 3
(139 genes)
Metagene 1
(122 genes)
46
Metagene 1
(78 genes)
Metagene 3
(118 genes)
Metagene 2
(124 genes)
Mouse
E
Metagene 1
Metagene 2
Metagene 3
Human
Metagene 1
Metagene 2
Metagene 3
Mouse
Metagene Overlap (P-val)
1
10-138

Table S1

Journal
Nature
Science
Cell
Nature Immunology
Immunity
Journal of Experimental Medicine
Journal of Clinical Investigation
Cell Host and Microbe
Blood
Proc Natl Acad Sci USA
Current Opinion in Immunology
Trends in Immunology
Journal of Allergy and Clinical Immunology
Plos Pathogens
Plos One
Mucosal Immunology
Arthritis and Rheumatism
Seminars in Immunology
Autoimmunity
Journal of Immunology
European Journal of Immunology
Genes and Immunity
Infection and Immunity
Immunology and Cell Biology
Vaccine
Cytokine
Journal of Clinical Immunology
Immunology

Platform ID	Name	Organism
GPL1261	Mouse430_2	*Mus musculus*
GPL339	MOE430A	*Mus musculus*
GPL8321	Mouse430A_2	*Mus musculus*
GPL6246	MoGene-1_0-st	*Mus musculus*
GPL81	MG_U74Av2	*Mus musculus*
GPL570	HG-U133_Plus_2	*Homo sapiens*
GPL96	HG-U133A	*Homo sapiens*
GPL571	HG-U133A_2	*Homo sapiens*
GPL6244	HuGene-1_0-st	*Homo sapiens*
GPL8300	HG_U95Av2	*Homo sapiens*
GPL97	HG-U133B	*Homo sapiens*
GPL91	HG_U95A	*Homo sapiens*
GPL3921	HT_HG-U133A	*Homo sapiens*

Table S2

Parameter	ImmuneSigDB	Li	Chaussabel
Number of gene sets	4872	346	260
Fraction of gene sets annotated	100%	75%	15%
Number of Studies	389	540	9
Number of Samples	6283	32766	410
Species	2	1	1
Cells/tissue types	13	2	1

Table S6

Rank	Gene set
1	GSE22886_NAIVE_BCELL_VS_NEUTROPHIL_DN
2	GSE29618_MONOCYTE_VS_PDC_UP
3	GSE6269_E_COLI_VS_STREP_PNEUMO_INF_PBMC_DN
4	GSE22886_NAIVE_CD4_TCELL_VS_MONOCYTE_DN
5	GSE34156_UNTREATED_VS_24H_NOD2_AND_TLR1_TLR2_LIGAND_TREATED_MONOCYTE_DN
6	GSE34156_UNTREATED_VS_24H_TLR1_TLR2_LIGAND_TREATED_MONOCYTE_DN
7	GSE6269_E_COLI_VS_STREP_AUREUS_INF_PBMC_DN
8	GSE22886_NAIVE_TCELL_VS_MONOCYTE_DN
9	GSE29618_MONOCYTE_VS_MDC_UP
10	GSE22886_NAIVE_CD8_TCELL_VS_MONOCYTE_DN