THE JOURNAL OF IMMUNOLOGY

Variation and Genetic Control of Gene Expression in Primary Immunocytes across Inbred Mouse Strains

Sara Mostafavi, Adriana Ortiz-Lopez, Molly A. Bogue, Kimie Hattori, Cristina Pop, Daphne Koller, Diane Mathis, Christophe Benoist, The Immunological Genome Consortium, David A. Blair, Michael L. Dustin, Susan A. Shinton, Richard R. Hardy, Tal Shay, Aviv Regev, Nadia Cohen, Patrick Brennan, Michael Brenner, Francis Kim, Tata Nageswara Rao, Amy Wagers, Tracy Heng, Jeffrey Ericson, Katherine Rothamel, Adriana Ortiz-Lopez, Diane Mathis, Christophe Benoist, Taras Kreslavsky, Anne Fletcher, Kutlu Elpek, Angelique Bellemare-Pelletier, Deepali Malhotra, Shannon Turley, Jennifer Miller, Brian Brown, Miriam Merad, Emmanuel L. Gautier, Claudia Jakubzick, Gwendalyn J. Randolph, Paul Monach, Adam J. Best, Jamie Knell, Ananda Goldrath, Vladimir Jojic, Daphne Koller, David Laidlaw, Jim Collins, Roi Gazit, Derrick J. Rossi, Nidhi Malhotra, Katelyn Sylvia, Joonsoo Kang, Natalie A. Bezman, Joseph C. Sun, Gundula Min-Oo, Charlie C. Kim and Lewis L. Lanier

J Immunol published online 29 September 2014
http://www.jimmunol.org/content/early/2014/09/28/jimmunol.1401280

The Journal of Immunology is published twice each month by
The American Association of Immunologists, Inc.,
9650 Rockville Pike, Bethesda, MD 20814-3994.

Print ISSN: 0022-1767 Online ISSN: 1550-6606.

Published September 29, 2014, doi:10.4049/jimmunol.1401280

Variation and Genetic Control of Gene Expression in Primary Immunocytes across Inbred Mouse Strains

Sara Mostafavi,* Adriana Ortiz-Lopez,† Molly A. Bogue,‡ Kimie Hattori,† Cristina Pop,* Daphne Koller,* Diane Mathis,† Christophe Benoist,† and The Immunological Genome Consortium[1]

To determine the breadth and underpinning of changes in immunocyte gene expression due to genetic variation in mice, we performed, as part of the Immunological Genome Project, gene expression profiling for CD4$^+$ T cells and neutrophils purified from 39 inbred strains of the Mouse Phenome Database. Considering both cell types, a large number of transcripts showed significant variation across the inbred strains, with 22% of the transcriptome varying by 2-fold or more. These included 119 loci with apparent complete loss of function, where the corresponding transcript was not expressed in some of the strains, representing a useful resource of "natural knockouts." We identified 1222 *cis*-expression quantitative trait loci (*cis*-eQTL) that control some of this variation. Most (60%) *cis*-eQTLs were shared between T cells and neutrophils, but a significant portion uniquely impacted one of the cell types, suggesting cell type–specific regulatory mechanisms. Using a conditional regression algorithm, we predicted regulatory interactions between transcription factors and potential targets, and we demonstrated that these predictions overlap with regulatory interactions inferred from transcriptional changes during immunocyte differentiation. Finally, comparison of these and parallel data from CD4$^+$ T cells of healthy humans demonstrated intriguing similarities in variability of a gene's expression: the most variable genes tended to be the same in both species, and there was an overlap in genes subject to strong *cis*-acting genetic variants. We speculate that this "conservation of variation" reflects a differential constraint on intraspecies variation in expression levels of different genes, either through lower pressure for some genes, or by favoring variability for others. *The Journal of Immunology*, 2014, 193: 000–000.

For more than a century, inbred mice have played a unique role in biomedical research. Their group homogeneity, phenotypic reproducibility, and genetic stability over time have led to key discoveries in essentially every area of biomedical research (1), including the discovery of fundamental concepts of immunology such as histocompatibility, MHC restriction, or genetic susceptibility to autoimmune diseases. The nearly homogeneous nature of an inbred strain's genome underlies the extraordinary power of targeted germline modifications, and it has supported mapping of loci associated with disease or phenotypic traits. The genomes of laboratory strains have been molded by strong selective pressures linked to their domestication by mouse fanciers in China and Europe, then to inbreeding and allele fixation in biomedical research colonies. These genomes incorporate segments from several origins (2), as now clearly established by the decoding of the complete genome of the reference C57BL/6J, followed by a number of other inbred strains (3, 4). Efforts to standardize and integrate phenotypic and genetic information, as exemplified by the Mouse Phenome Database (MPD) project (5), are also helping to exploit the full potential of inbred strains in biomedical research.

The Immunological Genome (ImmGen) project is an international collaboration of laboratories that collectively perform a thorough dissection of gene expression and its regulation in the immune system of the mouse. Genome-wide gene expression data have been collected for ~250 immunological cell types of the mouse, yielding insights into genomic correlates of immunocyte differentiation and lineages (6). The assembled data also enabled predictions about regulatory networks that underlie mouse hematopoiesis (7). The first phase of the ImmGen project mainly used the reference C57BL/6J strain, and it thus focused on identifying changes in gene expression during differentiation and activation in the context of a unique genome. However, there is much value in analyzing the impact of functional genetic variation on gene expression levels. Variants influencing gene expression are pervasive in mammalian species and comprise a large majority of the disease-related variants identified in genome-wide association studies (8). Combined analysis of gene expression and genotype data across a genetically diverse population is a powerful means to understand the impact of genotypic variation on cellular processes, and ultimately to build mechanistic models that link genetic variation to detailed cellular processes in a context-specific manner (8, 9). Several comparative analyses of gene expression have been performed across inbred mouse strains (10–14)

*Department of Computer Science, Stanford University, Stanford, CA 94305; †Division of Immunology, Department of Microbiology and Immunobiology, Harvard Medical School, Boston, MA 02115; and ‡The Jackson Laboratory, Bar Harbor, ME 04609

[1] All authors and their affiliations appear at the end of this article.

Received for publication May 19, 2014. Accepted for publication August 23, 2014.

This work was supported by National Institutes of Health Resource Grant R24 AI072073 to the Immunological Genome Project, as well as by National Institutes of Health Grants DA028420 and AG038070 (to M.A.B.).

Address correspondence and reprint requests to Dr. Diane Mathis and Dr. Christophe Benoist, Division of Immunology, Department of Microbiology and Immunobiology, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115. E-mail address: cbdm@hms.harvard.edu

The online version of this article contains supplemental material.

Abbreviations used in this article: *cis*-eQTL, *cis*-expression quantitative trait locus; eQTL, expression quantitative trait locus; FDR, false discovery rate; GN, granulocyte (polymorphonuclear neutrophil); ImmGen (Project), Immunological Genome (Project); ImmVar (Project), Immune Variation (Project); MAD, mean absolute deviation; MPD, Mouse Phenome Database; PC, principal component; SNP, single nucleotide polymorphism; T4, CD4$^+$ T cell; TF, transcription factor; TSS, transcription start site; TV, true variability.

but were of limited breadth and/or performed in cell types not directly relevant to ImmGen.

In terms of understanding human disease, whereas the mouse models have been invaluable in establishing fundamental paradigms of immunologic function, caution has been suggested in translating findings from the mouse to the human immune system (15). Similarities and differences have been reported in the genomic underpinning of immune lineages of humans and mice, whether at steady-state or after cell activation (16–19). A direct comparison of the genetic underpinning of these differences would also be valuable in ascertaining what mouse models can be usefully applied to understand human diseases and their genetics.

To better understand the effect of genetic variation on the mouse immune system, we generated RNA expression data for 39 of the main inbred strains in the MPD "Priority Strain Panel." Using rigorous ImmGen standard operating procedures, genome-wide expression data were generated for two immunological cell types, CD4$^+$ T cells (T4) and polymorphonuclear neutrophils (granulocytes, GN). These were chosen to represent the main lymphoid and myeloid branches of the immune system, as well as its adaptive and innate facets. This effort paralleled a study of similar design in an ethnically diverse population of healthy humans, the Immune Variation (ImmVar) study, where genotype and gene expression data were collected for T4 and CD14$^+$CD16$^-$ monocytes (20–22). This matching study design allowed us to compare transcriptional variability and its roots in the two species. In the present study, we first report on the impact of genetic background on gene expression levels in mouse T4 and GN, identify *cis* expression quantitative trait loci (*cis*-eQTLs), and chart regulatory interactions that can be inferred from the perturbation of the regulatory network by genetic variation. Second, we compare the impact of functional variation in humans and mice by exploring the overlap between expression variability and its genetics in the two species.

Materials and Methods

Gene expression and genotype data

Inbred mouse strains from the MDP Priority Strain Panel, representing 39 strains, were obtained from The Jackson Laboratory (Bar Harbor, ME) at 5 wk of age. All mice were bred in The Jackson Laboratory under specific pathogen-free conditions. CD3$^+$CD4$^+$CD62L$^+$ naive T splenocytes and CD11b$^+$Ly6G$^+$ bone marrow GN were sorted from pools of two to three mice. Two biological replicates were generated for each strain using the ImmGen standard operating protocol (http://www.immgen.org). Gene expression data were generated for bone marrow GN and T4 using Affymetrix ST1.0 microarrays, the platform used for the main ImmGen compendium, resulting in the quantification of expression levels for 25,134 probes corresponding to 21,951 unique genes. Data were processed and normalized using the ImmGen standard operating protocol (http://www.immgen.org). When indicated, data were filtered to only include genes with $>$0.95 probability of expression (or a mean of $>$120 expression on the intensity scale; see standard operating protocol). This filtering criteria resulted in 11,598 and 11,285 expressed transcripts in T4 and GN, respectively, with 131,85 transcripts expressed in one or the other, and 9,698 transcripts expressed in both cell types. A threshold for absence of expression was also set at $<$0.05 probability of expression (or a $<$42 expression level on intensity scale). Genotype data were obtained from the mouse HapMap genotype resource (http://mouse.cs.ucla.edu/mousehapmap) (23). Only genotyped single nucleotide polymorphisms (SNPs) with minor allele frequency of $>$0.05 and a $\leq$10% missing rate (resulting in a total of 96,779 SNPs) were used in this study.

Defining the true variability metric, bimodality in gene expression, and complete loss of function loci

All analyses were performed in the MATLAB computing environment (R2013a, version 8.1.0.604). At least two biological replicates were available for each mouse and each cell type (for the strains for which there were more than two replicates, we randomly chose two of the replicates for this analysis). For the true variability (TV) metric, two quantities were computed for each gene and each cell type using the log-transformed data: 1) the between-strains mean absolute deviation (MAD), which was divided by the mean gene expression level for that gene; and 2) the average of within-strains MAD, where the MAD for each strain was computed using the two replicates for that strain and then divided by the mean gene expression level for that gene. The TV score for each gene was defined as the difference between the first quantity, representing both meaningful and unwanted variability, and the second quantity, representing the unwanted variability. We note that there are two main differences between the TV metric proposed here and a standard ANOVA approach: first, we chose to quantify variability using MAD as opposed to variance because the latter gives more weight to extreme values. Second, as opposed to an associated F-statistic in ANOVA, where the test statistics (interpreted as the true variability score) is the ratio of two variances, here we use the difference of the two MADs as the score. We chose to use the difference so to emphasize the magnitude of the variability, in addition to the relative variability of the within-strains and between-strains MAD.

Bimodal genes were identified using two criteria: the first criterion was based on the assessment of the fit of a mixture of Gaussian distributions with two components to expression levels across the strains, and the second criterion used a threshold on the fold difference between high- and low-expressing strains. The mixture of Gaussians were fit using MATLAB's *gmdistribution* function (R2013a, version 8.1.0.604). A likelihood ratio test was used to assign a bimodality *p* value to each gene by comparing the likelihood of a mixture of Gaussian distributions with two components with simply the fit of a single Gaussian distribution. Genes with bimodality $p < 10^{-6}$ and at least a 2-fold difference in top two high-expressing and bottom two low-expressing strains were identified as bimodal. Complete loss-of-function loci were identified as those bimodal genes that additionally satisfied a strict threshold on expression levels: an expression of $<$42 (corresponding to $<$0.05 probability of expression) for at least two strains and expression $>$120 (corresponding to $>$0.95 probability of expression) for at least two strains.

eQTL association mapping for mice

It is well appreciated that genetic association studies in inbred strains are impacted by population stratification, which violates the assumptions of standard statistical tests and leads to an abundance of false positive associations (and therefore an inflation of association *p* values) (24). To account for population stratification, we used linear regression, regressing out the effect of the top two genotype PCs from log gene expression data. We chose two PCs by quantifying the inflation of observed *p* values using the λ statistic (25) as we varied the number of removed genotype PCs from one to five. A *cis* window of 1 Mb centered on transcription start site (TSS) was used to identify all *cis* SNPs for each gene.

Joint analysis. To increase statistical power, for the joint analysis, residual expression data (after removing genotype PCs, see above) from both cell types were concatenated (after removing mean expression for each cell type separately), resulting in a dataset with 2 × 39 samples and 13,185 expressed transcripts (expressed in at least one cell type). For each SNP-gene pair, the Wilcoxon rank sum statistic (as implemented in MATLAB R2013a, version 8.1.0.604) was used to test whether the expression of the gene was significantly different between strains with the reference or the alternative allele at the given SNP. Ten thousand permutations were performed for each SNP-gene pair, permuting the assignment of SNP values to strains while keeping intact the correspondence between genotype assigned to the T4 and GN sample for the same strain (thus accounting for "repeated" samples). A gene-level *p* value was assigned that accounted for the number of tested SNPs per gene by using the minimum permutation *p* value across all tested SNPs for that gene as the null distribution (26, 27). The final set of *cis*-eQTLs was defined by setting a 5% false discovery rate (FDR) threshold on the gene-level *p* values.

Cell-specific eQTL analysis. Cell-specific eQTLs were identified by testing the significance of an interaction term between genotype and cell type indicator in a linear regression setting, where the fit of the baseline model (no interaction) with one that additionally included a cell type indicator by genotype interaction term was assessed using an *F* test. In particular, we model the expression level of gene *g* in tissue *t* for strain *i* as $x_{g,t,i} = \alpha_{g,t} + \beta_g s_i + \gamma_{g,t} s_i$ where $\alpha_{g,t}$ is genotype-independent tissue-specific effect for tissue *t* and gene *g*, β_g is the tissue-shared genotype effect, and $\gamma_{g,t}$ represents the cell-specific genotype effect for tissue *t*. As above, gene-level *p* values were computed using 10,000 permutations (permuting the assignment of genotype values to the strains).

Constructing regulatory networks in mouse and validation using Ontogenet links

For constructing regulatory networks, genes expressed in both cell types and identified to have nonnegligible TV scores (as per Supplemental Fig. 1A) were used, which resulted in 3675 analyzed genes. Among these, 164 are transcription factors (TFs; as defined in Ref. 7). Two networks (one for each cell type) were constructed using stepwise regression, where a sparse set of TFs (regulators) was identified for each target gene (set of targets includes both TFs and nonregulatory genes). More specifically, for each target gene, stepwise regression was performed using all regulators (excluding autoregulation), and inferred regulators were identified using a 5% FDR to correct for the number of TFs tested for each target. A "joint network" was also constructed using the same approach but applied on concatenated expression data form both cell types (after removing mean gene expression from each cell type). Networks were constructed on genotype PC–corrected data.

We used the joint network constructed from T4 and GN data to compare the coexpression-based links derived in the present study with those derived from the ImmGen data (using the Ontogenet algorithm; see Ref. 7). We decided to use the joint network, as we observed a high degree of overlap between networks constructed individually from each cell type (see *Results*), and to identify persistent, and thus more likely true positive, relationships. Regulatory interactions and modules defined by Ontogenet were downloaded from the ImmGen Web site (http://www.immgen.org). Note that in Ref. 7, two types of modules were defined: initially 81 larger "coarse-grained modules" were defined, and subsequently some of these modules were refined into smaller modules with more coherent expression, resulting in 334 "fine modules." Coarse modules were constructed to capture the mechanisms that coregulate a larger set of genes in one cell-lineage, whereas fine modules were constructed to capture the distinct regulatory mechanism controlling only a smaller subset of these genes in the sublineage(s). Only "fine modules" and their "top regulators," representing more functionally specific gene groups and links, were used in the present analyses. Based on these data, a list of 4083 testable links connecting the top regulators to all genes in their assigned module was generated. First, the replication rate for this list was computed by assigning a *p* value to each link in the present study based on the coexpression of the corresponding regulator-target pair, and then assessing the proportion of true-positive *p* values using Storey's π_1 (28). To correct for the overall inflation of *p* values between all pairs of genes, as is often observed in coexpression data, we used the distribution of *p* values for coexpression of all gene/gene pairs as the null distribution to assign a *p* value to each of the 4083 links. Second, the links identified in this study were tested for consistency with those identified by the Ontogenet algorithm on the ImmGen data using a hypergeomtric test. This test identified regulators whose inferred targets were also coregulated (i.e., assigned to the same module) according to Ontogenet. Third, we computed the proportion of links identified in the present study that were also reported by Ontogenet and used the hypergeometric test to compute a *p* value for the overlap.

Gene expression, genotype, and eQTL discovery in human

Genotype and gene expression for T4 and neutrophils were obtained from the ImmVar study. As done for the mouse data, *cis*-eQTLs were defined using a 1-Mb window centered on the TSS. Gene expression data were corrected for three genotype PCs and 30 expression PCs (to increase statistical power by removing variability due to environmental or nonlocal genetic factors). The number of removed expression PCs was set by evaluating the improvement in number of *cis*-eQTLs that were detected based on data from one ("training") chromosome (chromosome 18). In particular, to select the number of PCs that are removed, the number of *cis*-eQTL discoveries in raw data was compared to PC-corrected data where we varied the number of removed PCs from 1 to 50. In order to avoid overfitting, we optimized the number of removed PCs based on *cis*-eQTL discovery on just one chromosome (and not the whole dataset). As previously observed (29), the improvement in *cis*-eQTL discovery greatly increased with removal of PCs, and there was a stable plateauing effect when we removed 20–40 PCs (see, for example, Ref. 21). As described for the mouse data above, in the joint eQTL analysis, gene expression data from both cell types were combined and a gene-level *p* value was computed for each gene using permutation analysis (1000 permutations per gene). In this case, the Spearman rank correlation was used as the test statistic.

Constructing regulatory networks for human/mouse comparison

Stepwise regression was used to construct a regulatory network for T4 data. For this analysis, we used the set of genes expressed in both humans and mice (in T4) and were considered to have nonnegligible TV scores for T4 data in mice (as defined by Supplemental Fig. 1A), which resulted in a set of 3407, of which 183 are TFs. For constructing the network, human data were corrected for batch, population structure (three genotype PCs), gender, and age, whereas mouse data were corrected for two genotype PCs (mouse data were done in one batch, and the mice had identical gender and age). Significant links were identified at a 5% FDR.

The replication rate of links identified in one species onto the other was computed using the π_1 statistic to quantify the proportion of true-positives among the coexpression *p* values for the relevant links (links being replicated). As above, coexpression *p* values were adjusted using the distribution of all coexpression *p* values as the null.

The stepwise regression approach above identifies regulatory links in a target-centric manner, identifying "top" regulators for each target. Additionally, in a TF-centric manner, top targets for each TF were identified based on the ranking of their coexpression value (Pearson correlation coefficient) with the given TF. In particular, two analyses were conducted. First, for each TF in mice (humans), the top 10 targets were defined based on coexpression values, and the overlap of these targets was assessed in the top n = [10, 20, 30, 50] targets for the same TF in human (mouse). The significance of the overlaps was determined using the hypergeometric test and corrected for the number of TFs tested. Second, the evidence for conservation of the top n = [10, 20, 30, 50] targets of each TF in mice (humans) was assessed in humans by using the Wilcoxon rank sum test to compare the distribution of the coexpression values for the top n targets compared with the distribution of coexpression values between that TF and all genes.

Results

The mice tested in the present study included 35 classic laboratory inbred strains (*Mus musculus domesticus*) that represent all the major branches of the inbred tree (1) and four "wild-derived" strains (CAST/EiJ, PWD/PhJ, JF1/Ms, and MSM/Ms, which are representative of the *Mus musculus castaenus*, *Mus musculus musculus*, and *Mus musculus molossinus*, respectively). Gene expression data for bone marrow GN and T4 were quantified using Affymetrix ST1.0 microarrays (see *Materials and Methods*). Matching genotype data were obtained from the Mouse HapMap Genotype Imputation Resource (30) and included 132,285 genotyped SNPs (see *Materials and Methods*). Because we did not attempt in the present study to identify causal variants owing to the limitations imposed by the relatively large size of linkage disequilibrium blocks in inbred mice, the analyses only used genotyped SNPs for computational efficiency. All expression data can be browsed or accessed on the ImmGen Web site (http://www.immgen.org).

Extent and distribution of expression variation across strains

We first investigated the nature and extent of the transcript variability across the inbred strain panel. Overall we observed some variability in expression levels for most genes (58% of tested genes, or 8,544 genes in T4–or 39% of genes–and 10,006 genes in GN–46% of genes–at 5% FDR; Supplemental Fig. 1A). Of these, 2508 genes in T4 and 3711 genes in GN had >2-fold difference between the highest two and lowest two expressing strains. Some of the most variable genes correspond to retroviral elements (*Mela*, *EG665955*), and some correspond to loci with known copy number variation (e.g., *Cd244*, *Trim12*, *Glo1*) (31). A TV score was computed for each gene (and per cell type) to identify transcripts whose variance across the strains could be attributed to meaningful differences, by factoring out technical factors and unwanted variability (Fig. 1A, Supplemental Fig. 1A). In practice we computed a TV score for each gene by contrasting a measure of within-strain variability (computed from biological replicates) to between-strain variability (see *Materials and Methods*). We validated the reproducibility of these TV scores 1) by comparing them to TV quantified from a previous gene expression dataset from macrophages for the Hybrid Mouse Diversity Panel, which included 22 of the strains tested here (11); and 2) by assessing the

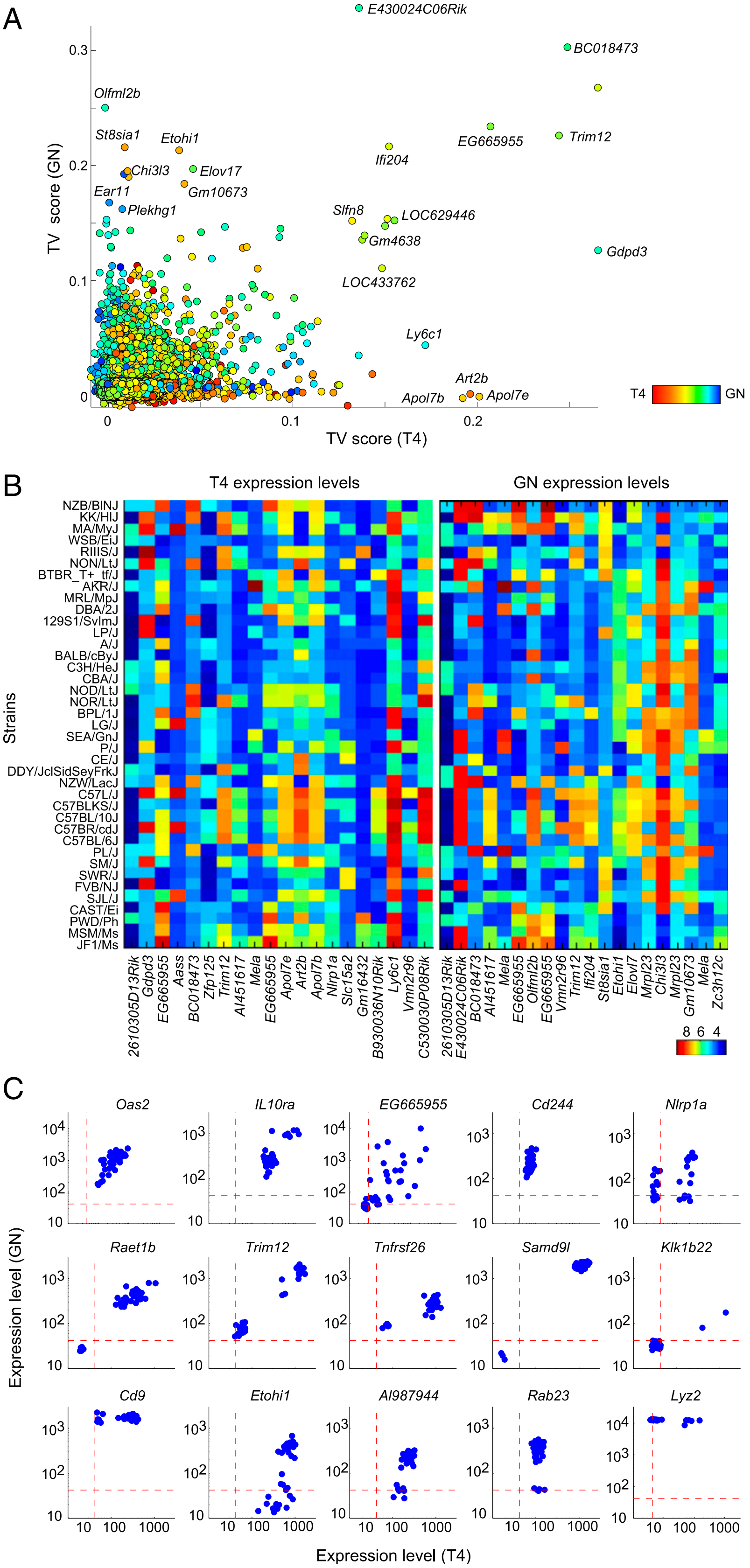

FIGURE 1. The extent and patterns of gene expression variation between inbred mice. A TV score was quantified for each transcript by contrasting a measure of between-strain variability, computed using biological replicates for each strain, with that of within-strain variability. (**A**) TV scores plotted per transcript based on T4 and GN data; each point represents a transcript. Colors depict preferential expression in T4 (red) or GN (blue) as quantified by the difference between mean expression levels. (**B**) Heat map of expression levels for the top 20 most variable transcripts based on T4 and GN data. (**C**) Examples of three types of "variation patterns." For each example transcript, each point represents a (mouse) strain, *x*-axis shows expression in T4, and *y*-axis shows expression in GN.

correspondence with reported variability in DNAase hypersensitivity sites in eight inbred strains (32). Reassuringly, we found a significant correlation between the TV scores in GN and T4 with those computed from macrophage data (Spearman ρ = 0.26 for GN and ρ = 0.2 for T4, $p < 10^{-100}$) (Supplemental Fig. 1B). We also observed significantly higher TV scores for genes previously identified to have variable DNase sites nearby, compared with the background TV scores ($p < 10^{-3}$; Supplemental Fig. 1C).

The distribution of expression across the strains for variable genes covered a wide range with varying patterns (Fig. 1B, 1C). In most cases, a continuous spectrum was observed, hinting at a complex genetic determinism (Fig. 1C, *top row*). In others, bimodal patterns were observed, which we quantified by assessing the fit of a Gaussian mixture model to the expression pattern of each gene (433 and 567 such bimodal genes for T4 and GN expression, respectively, were identified at a Bonferroni-corrected $p < 0.05$; Fig. 1C; see *Materials and Methods*). We also searched for instances of complete loss of function by using a combination of the bimodality test and expression <0.05 probability of expression in at least two strains (see *Materials and Methods*). Overall, we identified 67 and 53 complete loss-of-function loci in T4 and GN, respectively, of which 10 lost expression in both cells (Fig. 1C, *middle row*; a complete list of loss-of-function loci is available from http://www.immgen.org). An example gene displaying such an on/off pattern was *Raet1b*, which encodes an NK cell lectin-like receptor ligand; it was silent in five of the strains but highly expressed in all others. This pattern was consistent for T4 and GN, likely reflecting the variation in composition of the *Rae1α−ε* family, and more generally the multiplicity of targets of NKG2D (33). There were also several instances of "conditional loss-of-function" loci whose expression was sometimes absent in one cell type but present in all strains in the other cell type (Fig. 1C, *bottom row*); for example, *Rab23* transcripts were absent in GNs for some of strains, but present in all T4s. Several of these strains can thus serve as "natural knockouts" or "natural knockdowns" either directly or by backcrossing the segments involved.

We assessed the impact of genetic variation on gene expression at a global level by comparing the relationships between the strains inferred from gene expression data with known genealogies and with genotype-derived relationships (Fig. 2A). Simple examination of the parallel correlation maps of Fig. 2A showed a significant correspondence between strain relationships as derived from the gene expression data and strain genotypes (1, 34). Differences are sharper on the genotype than on the expression matrix, most trivially because the former inherently focuses on differences (SNPs) rather than on transcripts that are largely shared, and/or because most SNPs have no transcriptional consequence. As expected, the wild-derived strains (CAST/EiJ, PWD/PhJ, JF1/Ms, MSM/Ms) were more similar to each other than the classical inbred strains; the CAST/Ei strain, derived from *M. m. castaneus* species, was the most distant outlier, whereas the two *M. m. molossinus*–derived strains (JF1/Ms and MSM/Ms) were more closely related to each other. Other relationships expected from strain histories (35) include the "C57 black" group of strains, the high pairwise similarity between CBA and C3H, or between NOD and NOR, both of which derive from the same stock through selection for susceptibility or resistance to diabetes (36).

For a better handle on the number and identity of differentially expressed transcripts that underlie these relationships, we created a genotype-based dendrogram depicting the relationship between the strains and identified differentially expressed genes that characterized each group (Fig. 2B). The wild-derived group was associated with 2092 differentially expressed genes (5% FDR, of which 204 differ by a fold change >2). These "wild-specific" genes have a range of functionalities, as evidenced by the absence of enrichment for any particular functional category based on gene ontology analysis. Manual exploration of the top associations identified several suggestive differences: the marked underexpression of some TLRs (*Tlr1* and *Tlr7*) in T4 cells from wild-derived strains; several members of the NK family (*Klrd1*, *Klrb1f*) or of the IFN-response pathway (*Ifitm1, Ifitm2*) were uniquely expressed in wild-derived T4; and transcripts encoding cell-surface molecules whose distribution is normally restricted to myeloid cells (*Atp1a3*, *CD163*, *Anxa3*) but were present in T4 from wild-derived strains.

We also noted an intriguing differential expression of *Eps8l1* in the C57 black group. Mutations in *Eps8* family members lead to diverse auditory phenotypes, and the C57 strains are known to develop age-related hearing loss (37). At its inception, this project aimed to find, in the genetic and gene expression data, correlates to the phenotypic traits of these mouse strains, as assembled in the MPD. Unfortunately, a systematic test for association between gene expression levels and an extensive set of behavioral and physiological traits (~1500 traits from the MPD) (38) did not yield significant findings when corrected for random association. Reasons for this may include the limited number of strains for which complete phenotypes were available, buffering of gene expression by regulatory networks, or that the two cell types examined are not relevant to the traits currently in the MPD.

Identifying cis*-eQTLs for neutrophils and T4*

By correlating local genotype and expression data for the mice, we next identified specific *cis* genetic variants that impact gene expression levels in T4 and/or GN (our study did not have the statistical power to detect *trans*-eQTLs). To eliminate broad population-based trends that can result in the inflation of association p values (30), we removed the effect of the top two PCs of the genotype, which represent population structure, from the gene expression data using linear regression. We chose two principal components by assessing the inflation factor λ(24) (see *Materials and Methods*). We performed a *cis*-eQTL analysis with the residuals of this fit, defining *cis* SNPs as mapping in a 1-Mb window from the transcription start site. To increase our power to detect eQTLs that are shared by the two cell types while also detecting cell-specific eQTLs, we performed two analyses: 1) in a "joint analysis," we combined data from the two cell types and evaluated the significance of each SNP-to-gene association using permutation analysis; and 2) in a "cell-specific" analysis, using an ANOVA model, we explicitly tested the significance of a cell-specific SNP effect (see *Materials and Methods*). In both cases, using permutation analysis, we obtained a gene-level p value that took into account the number of tested *cis* variants (26, 27, 39) and defined significant *cis*-eQTLs at 5% FDR based on these gene-level p values.

Using the joint analysis, we identified 1047 genes with *cis*-eQTLs (Fig. 3A, Supplemental Table I; available for browsing on the ImmGen server). The joint analysis increased our discovery power: we identified 262 eQTLs that were not detected in separate analyses of GN and T4 data (774 and 958 eQTLs in separate analysis of T4 and GN, respectively). We observed a significant correlation between *cis*-eQTL association strengths and TV (Spearman $\rho = 0.29$, $p < 10^{-100}$).

Previous studies have identified *cis*-eQTLs for inbred mice in various tissues, including liver (10, 12–14) and immunocytes (10, 11). We compared our set of *cis*-eQTLs with those identified in macrophages (11), which was the most relevant and comparably sized. Orozco et al. (11) identified 1937 genes (corresponding to 4897 SNP-gene pairs) with *cis*-eQTLs controlling transcripts in primary macrophages that were testable in this study. To robustly compare results, we used Storey's π_1 statistic (28) and observed a replication rate of 55% ($p < 0.001$ under permutation testing). This estimate of overlap is similar to those previously reported in the literature for studies involving different cell types or conditions (29, 40–42).

To identify cell-specific *cis*-eQTLs, which should denote genetic impact on cell-specific regulatory pathways, we considered 9698 genes that are expressed in both cell types. We identified 234

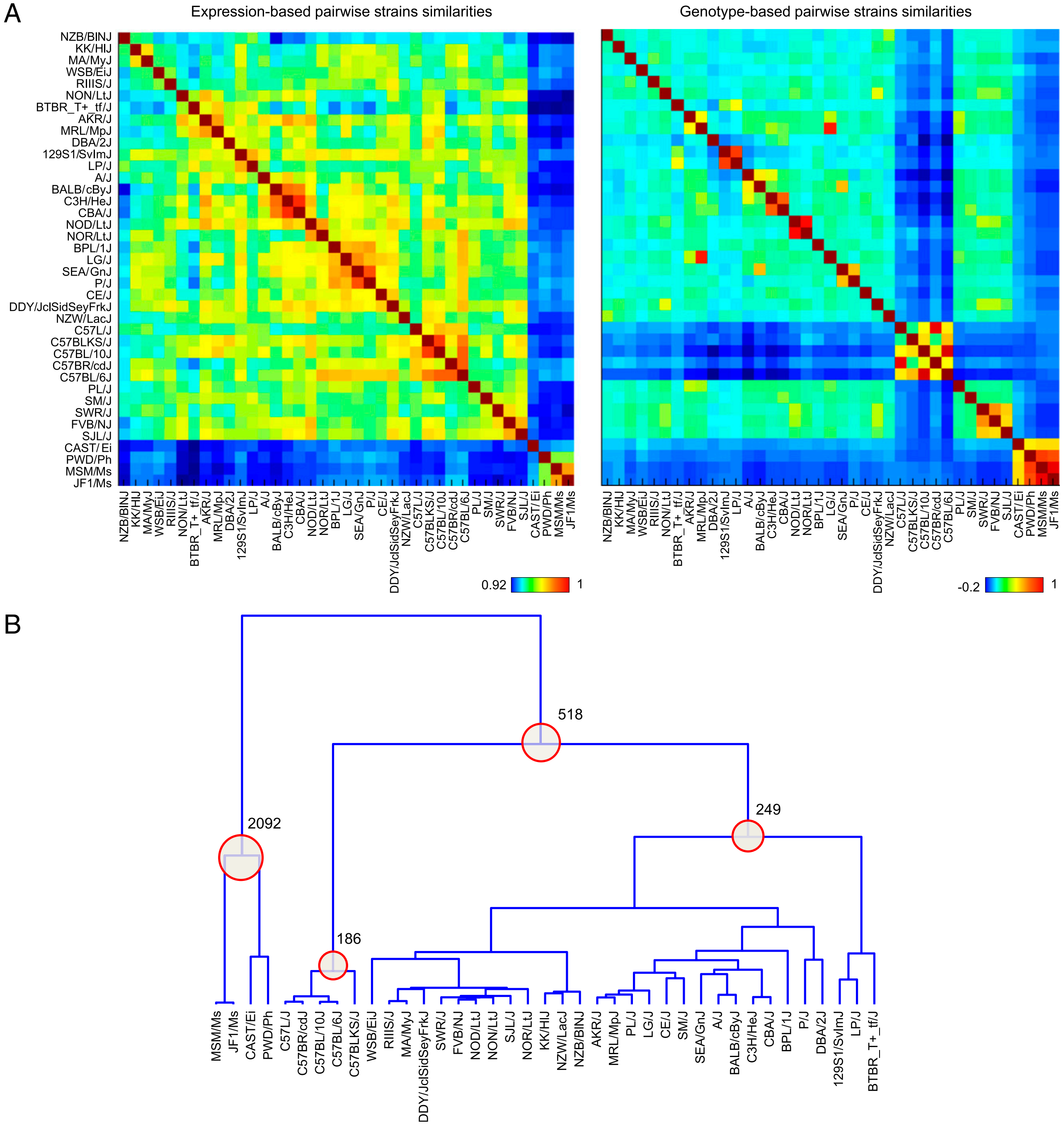

FIGURE 2. Expression-based and genotype-based strain similarities. (**A**) The similarity between each pair of strains was computed from genotype data or gene expression data using Spearman correlation. For expression data, all expressed genes were used in the computation. For genotype data, all variants satisfying the initial criteria (minor allele frequency > 0.05 and missing rate < 10%) were used. Each element in the heat map (matrix) represents the strength of the similarity between a pair of strains. Expression-based and genotype-based similarity heat maps follow the same row and column order. (**B**) The dendrogram was derived from a genotype-based similarity matrix using hierarchical agglomerative clustering. To account for strain-based scale differences between the distribution of similarities, pairwise strain distances for constructing the dendrogram were derived from ranked pairwise similarities for each strain. Each internal node in the dendrogram was used to define two groups (clusters) of strains, which are represented by descending leaves and nondescending leaves. Numbers of differentially expressed genes (defined using *t* test and a 5% FDR threshold) for groupings that yielded >100 differentially expressed genes at 5% FDR are shown.

significant cell-specific *cis*-eQTL, which indicates that ~30% of discovered *cis*-eQTLs are cell-specific (Fig. 3B), an estimate consistent with recent reports of tissue and cell type specificity of eQTLs in human studies (41, 42). For many genes with a cell-specific eQTL signal, we found major differences between effect sizes for the associated SNP in the two cell types (Fig. 3C). This analysis also identified 17 eQTLs where expression values correlate in an opposite manner in the two cell types. For 10 of the 17 genes, the same top SNP was identified from both GN and T4 data. One of the strongest eQTLs with opposite directionality of effect was observed for *Pot1a* (Fig. 3D). The proportion of directional *cis*-eQTLs discovered in the present study is similar to those previously detected using primary immunocytes in humans (21, 43). This divergence may reflect the fact that a factor

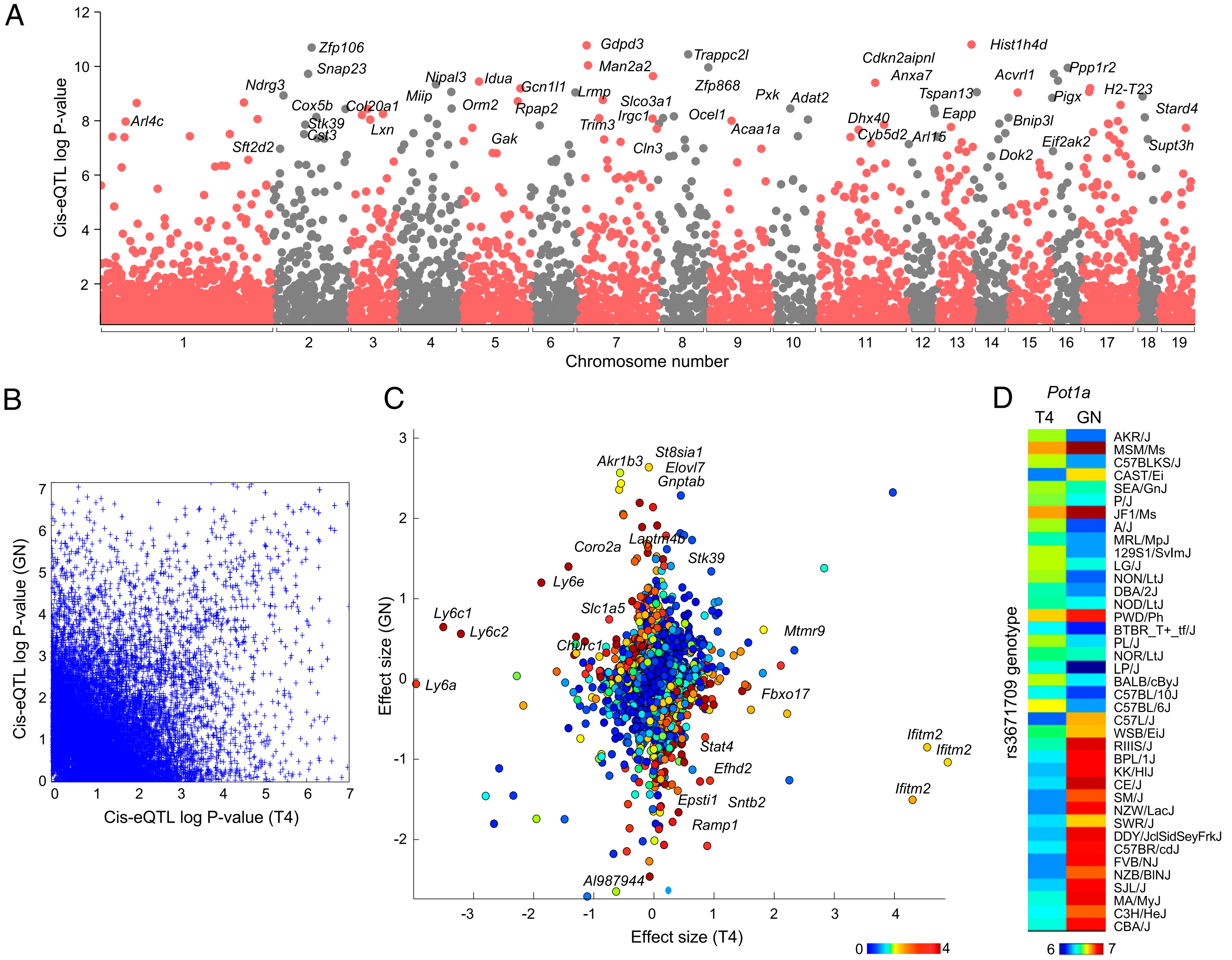

FIGURE 3. GN and T4 joint-discovered and cell-specific *cis*-eQTLs. (**A**) Association *p* values for each transcript and all of its *cis* SNPs (1 Mb from the TSS) were computed using the Wilcoxon rank sum test. Association statistics were computed from both T4 and GN data (joint analysis). Each point represents a transcript; only the best association for each transcript is shown. (**B**) For each cell type, association *p* values for SNP-transcript pairs were computed separately. The association *p* values are shown for the best GN SNP and best T4 SNP for each transcript (i.e., each transcript is represented twice). (**C**) Cell-specific SNP-transcript association *p* values were computed using an ANOVA model. Effect sizes for the best SNP for each transcript and each cell type were computed as the mean difference between strains with alternative and reference alleles. Colors depict the strength (negative $\log_{10}$ *p* value) of cell specificity at gene level. The figure shows effect size and association strengths for the best SNP for all expressed genes. (**D**) Heat map shows the expression levels for the gene *Pot1a* in GN and T4. Strains are sorted based on the genotype of the best SNP for *Pot1a*.

recruited to the same motif acts in an opposite manner in the two cells, but it is also possible that the SNP identified is in linkage disequilibrium with two different causal SNPs, each active in one cell type only.

Identifying regulatory links by coexpression analysis

Gene expression datasets that carry small "perturbations" such as those resulting from genetic variation can be fruitfully exploited to reverse-engineer the structure of genetic regulatory networks (44–46), with the caveat that relationships based solely on baseline coexpression cannot resolve causal from merely correlative associations. We constructed regulatory networks where we inferred interactions (links) between a set of 164 TFs and 3675 candidate downstream targets using stepwise regression. This analysis included only genes that were expressed in both cell types and had a nonnegligible TV score (as per Supplemental Fig. 1A). As above, to avoid artifacts from broad population structure, we used the genotype PC–corrected data. We identified 3462 and 3321 significant (5% FDR) links in T4 data and GN data, respectively, and 4927 links in a joint network constructed using both T4 and GN data. For these networks, few regulatory hubs correlated with expression levels of a large number of targets (>100), and most TFs were linked to ≤15 targets (Fig. 4A, 4B). The major hubs mostly include chromatin modifiers and generic transcriptional activators such as *Smarcd1* and *Smarce1* (SWI/SNF-related chromatin regulators), *Asf1b* (a histone chaperone), *Phf21a* (a histone deacetylase), and the histone deubiquitinase *Mysm1* (Fig. 4B).

We evaluated the overlap between GN- and T4-inferred regulatory links using Storey's π_1 statistic (28). Considering only the interactions passing the statistical significance threshold in the discovery sample (5% FDR), we estimated replication rates of π_1 of 53 and 49% for T4 links in GN and vice versa, respectively, indicating that a large fraction of these associations is shared among the two cell types. Conversely, by directly testing the significance of a cell type–specific effect (see *Materials and Methods*), we estimated that 17% of total interactions are truly cell specific (at 5% FDR). With the interaction test, *Lmo2* was one of

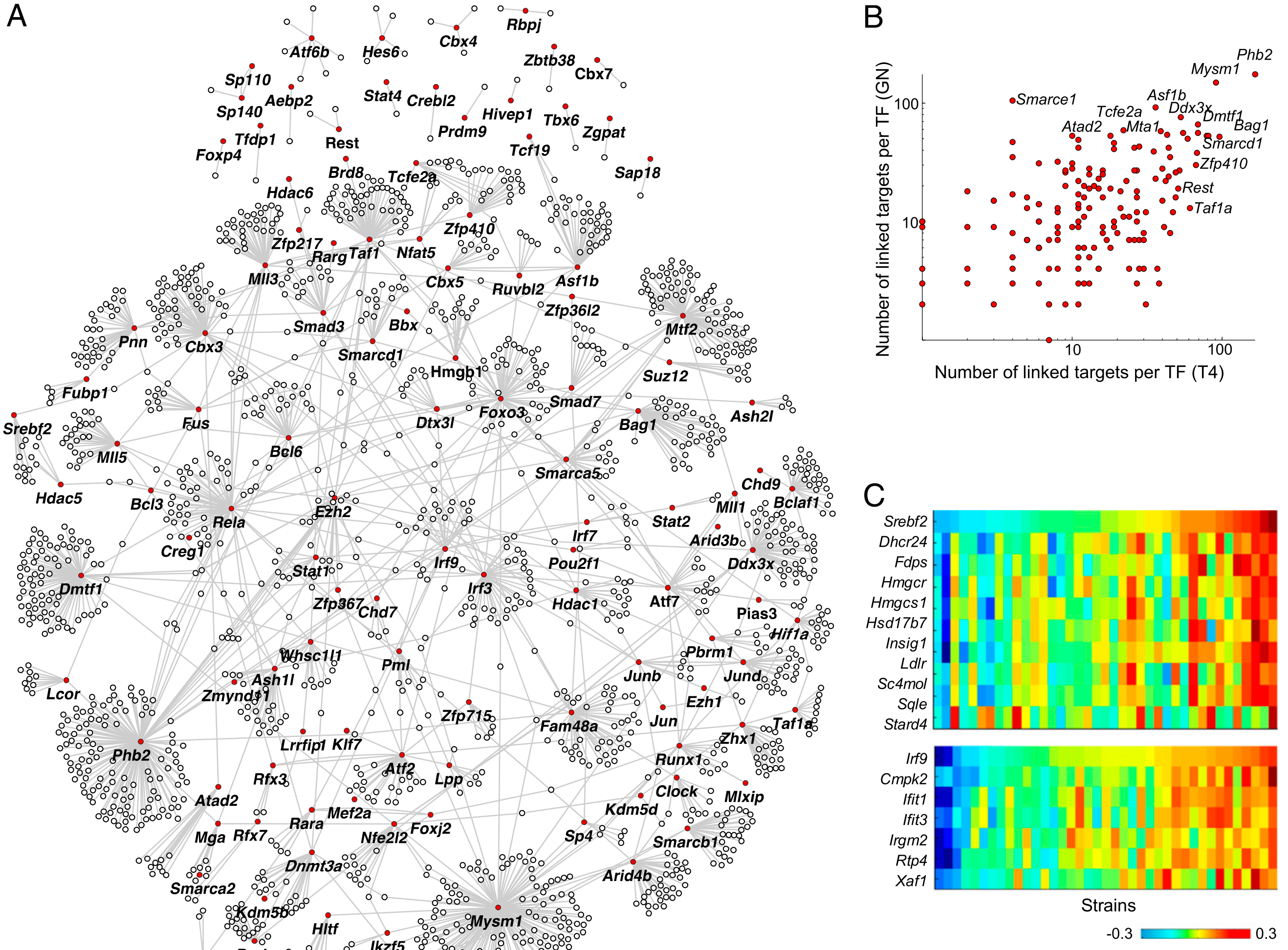

FIGURE 4. Analysis of gene coexpression in mice. (**A**) Overall network showing TF-target links discovered from T4 and GN data (joint network; for visualization purposes, figure only shows a limited set of the strongest links). TFs are marked in red. (**B**) Figure shows the node degree for each TF in T4 (x-axis) and GN (y-axis) networks. (**C**) Expression heat maps for selected regulators and their inferred targets; *top panel* shows data for *Srebf2* and nine of its targets, and the *bottom panel* shows data for *Irf9* and six of its targets (only targets that overlap with the Ontogenet predictions based on ImmGen data are shown).

the most differential regulatory hubs, with 51 inferred links in GN, but only four potential target genes in T4, which likely denotes a very specific role in GN (its targets in GN do not correspond to a distinct functional category in gene ontology analysis).

For an independent validation of coexpression relationships identifiable from this data, we compared a joint set of links identified from analysis of both cell types (joint network; see *Materials and Methods*) with a previous network constructed from the ImmGen compendium using the Ontogenet algorithm. Ontogenet exploits variation in expression through differentiation cascades to identify regulatory relationships (7). We hypothesized that true TF-target pairs identified by Ontogenet would also show evidence of coexpression when natural genetic variation was the network perturbant. First, we evaluated the strength of coexpression between pairs of TFs-targets previously identified by Ontogenet, and, using the π_1 statistic on adjusted p values for coexpression correlation coefficients (see *Materials and Methods*), we found that 27% of these links show evidence of coexpression. Conversely, we checked whether the targets of each TF are also more likely to belong to the same Ontogenet module by testing for significantly enriched Ontogenet modules among the predicted targets of each TF using the hypergeometric test. For 11 of the 127 TFs with at least 10 inferred targets, the targets were significantly enriched in an Ontogenet (fine) module at 5% FDR (Supplemental Table II). For example, *Srebf2*, which encodes a sterol regulatory TF, was associated with 33 genes in this study, 9 of which were part of the same module and predicted by Ontogenet to be regulated by *Srebf2* ($p < 10^{-15}$; Fig. 4C). Another well-known set of replicated links was between *Irf9* and six of its known targets within the IFN response signature ($p < 10^{-8}$; Fig. 4C). Although less robust to differences in inference method and sample sizes, we also directly evaluated the overlap between the inferred regulatory links in this study and those of Ontogenet, where we observed a modest (4%) but significant overlap (hypergeometric $p < 10^{-10}$).

Coexpression relationships that underlie the regulatory links in the present study are not conclusive of directionality. To disentangle causal from simply correlative associations in the present network, we examined the propagated influence of *cis* variants associated with the inferred TFs (47). In practice, we asked whether a *cis*-eQTL SNP for a TF was also correlated with the expression levels of the TF's inferred targets. Within the set of links identified in the joint network (4927 links), 230 links were testable, as they were incident to 1 of the 15 TFs for which a *cis*-eQTL had been identified above; 50 links (22%) were "causally" supported, in that the genotype at the *cis*-eQTL was significantly associated with the expression of the TF's targets at 5% FDR.

Comparison of variation in gene expression in humans and mice

Comparative studies of gene expression patters across species have mainly focused on comparing similarities and differences in expression across tissues, cell types, or responses to triggers. In these studies, conserved cell type specificity or response to similar triggers across species is taken to indicate conserved functionality (19, 48–53). The impact of genetic variation in each species is averaged, smoothed, or factored out in such analyses. Instead, we sought to exploit the diversity of genetic background across inbred mice and across the human population sampled in the ImmVar study (which includes 360 healthy individuals from Asian, African, and European backgrounds with available expression data for CD4 and CD14 cells; the derivation and analysis of ImmVar datasets are detailed elsewhere; see Ref. 21). ImmVar was designed to match the present analysis in several respects (parallel profiling of T4 in both humans and mice). We took advantage of these congruent datasets to explore the similarities and differences in expression variability, the impact of *cis* regulatory variation, and the inferred regulatory interactions in mice and humans. For this analysis, we considered 14,130 genes with one-to-one human/mouse orthology (MGI HMD_Human5 set) and restricted the analysis to 5,964 genes expressed in T4 (we only analyzed the T4 data, because of the exact correspondence of this cell type in our data from the two species).

First, we applied the same TV metric of variation discussed above to compare the variability in genes' expression in humans and mice. The TV scores were calculated for human genes by using replicate samples prepared from the same donor (collected at intervals ranging from 3 to 25 wk) after accounting for batch, age, and gender (using linear regression). The TV metric allowed us to eliminate genewise technical variability and only capture biological variability (responding to environmental and/or genetic cues). Human versus mouse comparison of the TV scores showed interesting patterns (Fig. 5A–C); some genes were variable in one species or the other, but in general there was a correlation between TV in mice and humans (Spearman ρ = 0.16, $p < 10^{-10}$). We categorized genes into five equally sized bins in each species based on TV scores and found significant predictability of TV scores in the second species based on the assigned bin in the first species (Fig. 5C; Wilcoxon rank sum test p values of 10^{-4} to 10^{-8}

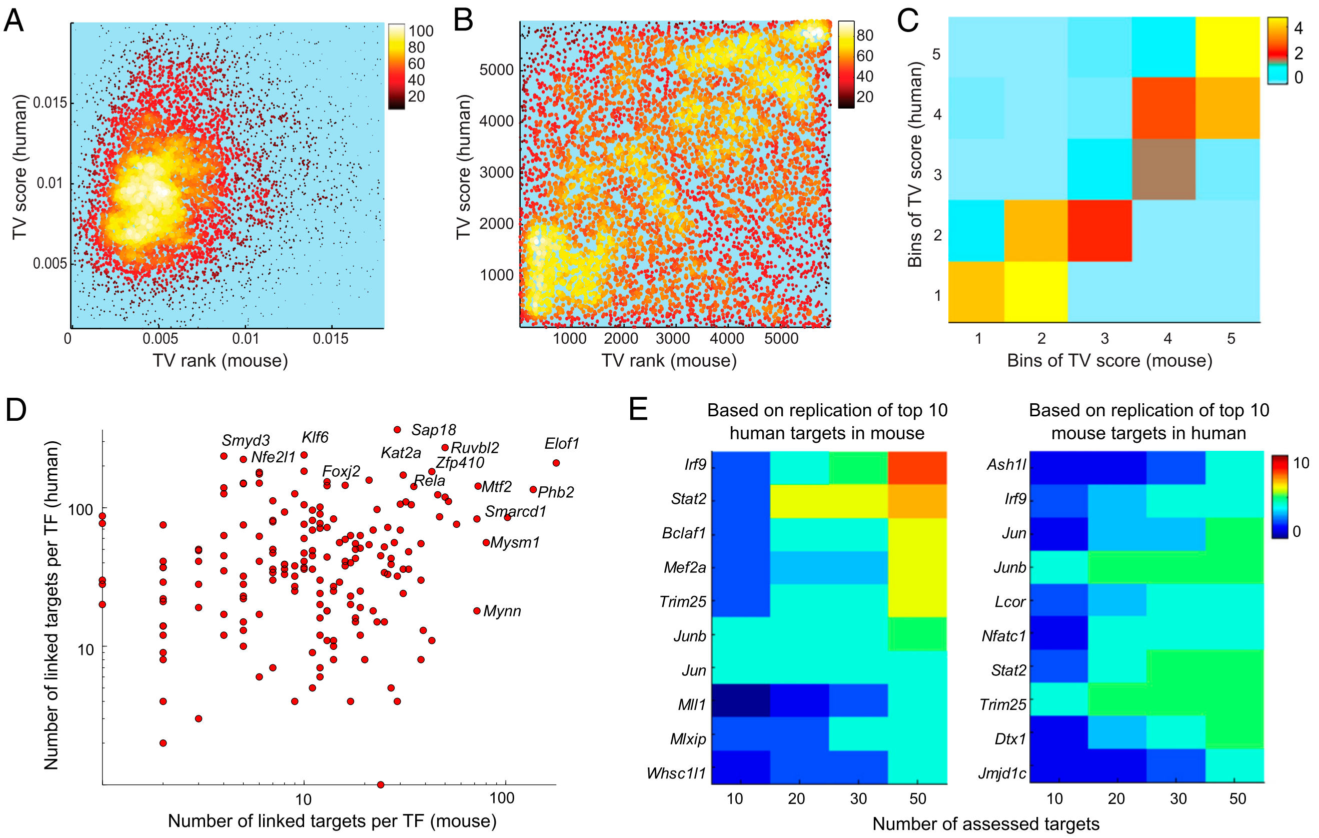

FIGURE 5. Sharing of variability and coexpression in mice and humans. (**A**) Density map of the scatter plot for TV scores (expressed genes only) in T4 of humans and mice (Spearman correlation ρ = 0.16). For visualization purpose, the figure is zoomed in on the high-density region. (**B**) Density map of the scatter plot for TV ranks (TV scores were transformed to ranks; in each species, highest TV score is assigned a rank of 5965). (**C**) To quantify the significance of the overlap of variable genes in the two species, mouse and human TV scores were each binned into five equally sized bins, and for each pair of TV bins (a square), the density of genes observed in that square was compared with density of genes in the same square under random mapping of human genes to mouse genes. Fifty thousand permutations (permuting the mapping between human and mouse genes) were performed to quantify the significance of the observed density in each square. Colors depict the strength of these p values (negative $\log_{10} p$ value). (**D**) Figure shows the network degree (number of inferred links) for each TF in the human and mouse networks. (**E**) Heat map summarizes the number of shared top targets for each TF in humana and mice. The top 10 targets for each TF in humans (mice) were first identified, and then the overlap with the top $n = [10, 20, 30, 40, 50]$ mouse (human) targets for the same TF was computed. Only regulators with at least 4 overlapping targets in the list of top 50 targets ($p < 10^{-6}$) are shown. The second plot shows the result where the top 10 targets for each regulator in mice were evaluated in human data.

for bins created in mice and humans, respectively); for example, 26% of the genes in the top bin (most variable) in one species are also categorized in the top bin the other species.

The variability captured by the TV metric encompasses environmental and other factors beyond the impact of genetic variation. To compare the extent of genetically determined variation in gene expression in both species, we evaluated the overlap of *cis*-eQTLs in humans and mice. Using the same methodology as above, we identified 2285 *cis*-eQTL genes in the human ImmVar datasets among the set of 7098 expressed genes in both species (either cell type) (eQTLs identified using the joint analysis; see *Materials and Methods*). Of the 674 genes associated with an eQTL discovered in mice for this set of expressed genes, 275 were also associated with an eQTL based on human data (hypergeometirc $p < 10^{-6}$), implying that genes that show a significant impact of local genetic variation tend to overlap in mice and humans, even though the variants themselves are certainly unrelated.

Next we compared gene regulatory networks constructed from the T4 dataset for both humans and mice. The motivation was to analyze the evolutionary conservation of these regulatory links, and from a practical standpoint to validate the inferences by confirmation in another species. For each species, we first constructed a network using stepwise regression as above (see *Materials and Methods*). At a global level, we observed a correlation between TFs' out-degree (the number of targets connected to each TF; Fig. 5D), with 38% of the top 20% hubs in one species shared with the second species ($p <$ 0.01). As above, chromatin modifiers tend to be strong regulatory hubs in both species. We used the π_1 statistic to estimate the fraction of TF-to-target links identified in one species that are replicated in the second species. A 47% replication rate was observed for mouse links in the human T4 dataset, and a 19% replication rate was found for human links in the mouse dataset (permutation analysis $p <$ 0.001) (Supplemental Fig. 2A, 2B). Finally, in a regulator-centric analysis, we also assessed the correspondence between top coexpressed links for each TF in the two species. To do so, we assessed the overlap and the distribution of coexpression values (correlation coefficients) for the top n = [10, 20, 30, 50] targets of each TF in the second species (see *Materials and Methods*). Of the 189 TFs that were analyzed, we identified 17 TFs whose top 10 targets were highly conserved (hypergeometic test $p < 10^{-6}$; Fig. 5E), and the top targets of an additional set of 42 TFs showed significant evidence of high coexpression values in the second species (using the Wilcoxon rank sum test; Supplemental Fig. 2C, 2D). Among these highly conserved coexpression links, we identified well-known relationships, including coexpression between *Irf9* and IFN response genes *Dhx58*, *Ifi35*, *Irf1, Pml*, *Trafd1*, and *Stat2* and strong coexpression between *Jun* and *Fos* and known early response genes (*Ier2, Gadd45b*). We did not attempt to interpret the divergent regulatory links within these datasets: these are not conclusive of true differences, because multiple confounding factors can underlie such differences (different environmental influences, much smaller sample size for the mouse data, imperfect mapping of human to mouse probes). Overall, these comparisons show that many of the regulatory connections that can be inferred from the interindividual variation in expression profiles are conserved between these two mammalian species.

Discussion

Our motivation, in the context of the ImmGen and MDP programs, was threefold: to serve as a reference of genomic and genetic information relevant to the immune function in mice, to provide additional material for the dissection of genetic regulatory networks, and to provide a documented basis for comparison of the mouse and human immune systems.

In terms of resources, the present data provide useful information at several levels, and are all available interactively from the ImmGen and MDP Web browsers (http://www.immgen.org, http://phenome.jax.org). We detected a number of genes with a >2-fold change in expression across the strains (the empirical rule of thumb for functional significance). It will be interesting to see how these traits segregate in settings such as the Collaborative Cross strains, where the chromosomal segments can be traced in the recombinant chromosomes, allowing refinement of the genetic control and/or discovery of epistatic modifiers. Variation followed both bimodal and continuous expression patterns across mouse strains, including a few loci with complete loss of expression in some of the strains. As such, these can serve as a resource of natural knockdowns or natural knockouts (some affecting both cell types, others cell-specific). The 1222 *cis*-eQTLs detected in the two immunological cell types are also available through the dedicated ImmGen interface. However, the relatively large sizes of the linkage disequilibrium blocks in these inbred mouse strains, relative to outbred humans or mice (22), make it impossible to pinpoint with precision the causal variant, and the SNPs listed should only be considered as likely proxies of the truly relevant variant. Nevertheless, the patterns of variation and the eQTLs described here, and their conservation across species, may help to interpret differences in susceptibility to infection or autoimmune diseases, in a manner than translates to genetic in risk human populations.

The patterns of interstrain variability followed, as expected, the patterns of genetic distance and genealogical history between the strains. Wild-derived strains were predictably more distant from the classic inbred lines. Some of this genetic distance may be directly related to selective events during mouse domestication or to the input from non-*domesticus* subspecies. We previously reported that a variant at the *Il1b* locus, which leads to a 5- to 10-fold greater IL-1 response to stimulation through innate receptor pathways, is frequent in wild-derived strains but quite rare among classical inbred strains (54), and some of the expression variations uncovered in the present study may be of the same nature (e.g., *Tlr1* and *Tlr7*, although in this instance it is the wild-derived strains that show low or absent expression in T4). Some genes whose expression is normally confined to myeloid cells across the ImmGen compendium were expressed in T4 of the wild-derived strains. Some of these conditionally expressed genes are surprising, such as the expression of CD163, a scavenger receptor on macrophages whose function in T4 is not immediately obvious. We might speculate that this reflects a mode of innate sensing by T4 that was lost during domestication (interestingly, however, human T cells do not express these monocyte genes).

The distribution of *cis*-acting genetic variation was significantly correlated with the variation in expression for the most variable genes, although many of the genes with a high TV score did not show an active *cis*-eQTL. Recent literature indicates a larger impact for local sequence variation, which may have been detectable with larger sample sizes (29, 55), perhaps attainable with a larger study of outbred or Collaborative Cross mice. We note that the number of *cis*-eQTLs detected in the present study is more than what would be expected from an equally sized human dataset (28) where the effect of environment cannot be as effectively controlled.

The coexpression-based network estimated in this study extended the analysis of the regulatory networks of immunocytes initiated in ImmGen (7), and we observed a comforting degree of overlap between the two analyses. Although coexpression cannot formally identify causal directionality in a correlated pair (i.e., who controls whom), the selection of transcriptional regulators provides a functional prior for directionality. Indeed, when we searched for causal chains of associations, by correlating a *cis* variant impacting the expression of a TF with the TF's downstream effects on its inferred targets, a significant portion (22%) of the testable links turned out to be causally driven. Interestingly, connections identified from interstrain variation more frequently involved generic regulators such as chromatin modifiers, which showed up in the present study as major hubs, than classic sequence-specific DNA-binding TFs and lineage determination factors (which were predictably more prevalent in the Ontogenet analysis). This difference is in line with the paucity of *cis*-eQTLs for classic TF regulators involved in differentiation or lineage determination, as previously shown in human cells (29, 56). One might speculate that a degree of "noise" in transcript level resulting from variations in redundant and pleiotropic factors is better tolerated (or even favored) than variation in more specific factors that form the blueprint of cell differentiation and lineage determination. This dominance of broad transcriptional regulators as major coexpression hubs was strikingly reproduced in the human datasets.

Finally, we observed sharing of the patterns of expression variability between humans and mice. Both genetic and nongenetic factors can result in expression variability, and we also observed significantly nonrandom overlaps in genes that are associated with *cis*-eQTLs in both species. From an evolutionary standpoint, this "conservation of variability" can be explained by species-shared strength of selection pressure on gene expression levels (57): variation in more redundant and/or less essential genes is better tolerated, and these characteristics would tend to be conserved. It is also possible that some of this species-shared variability is in genes whose intraspecies variation is favorable. The extraordinary diversification of coding sequence in MHC genes favors heterozygosity in individuals and diversity at the level of the species to best meet variable pathogen challenges (58). Similarly, it may be advantageous to diversify the levels of expression, and hence of response potential, in pathways of the immune system. Genes controlling activating and inhibitory NK receptors would plausibly fall in that category. From a mechanistic standpoint, one might also imagine different scenarios for the roots of this reproducible variability: some regions of the genome may be inherently noisier, a characteristic preserved during the evolutionary shuffling of syntenic chromosomal regions; regulatory feedback loops that control individual genes or sets of genes may be more or less robust; and microRNAs or other noncoding RNAs might make for a variable degree of control. Any of these mechanisms may have been, to an extent, preserved through 200 million years of evolution to conserve immunologically relevant variation.

Acknowledgments

We thank ImmGen participants for comments and suggestions, J. Ericson and S. Davis for help with RNA preparation and profiling, S. Davis for data processing, and J. Flint for helpful discussions.

ImmGen Consortium

David A. Blair,[1] Michael L. Dustin,[1] Susan A. Shinton,[2] Richard R. Hardy,[2] Tal Shay,[3] Aviv Regev,[3] Nadia Cohen,[4] Patrick Brennan,[4] Michael Brenner,[4] Francis Kim,[5] Tata Nageswara Rao,[5] Amy Wagers,[5] Tracy Heng,[6] Jeffrey Ericson,[6] Katherine Rothamel,[6] Adriana Ortiz-Lopez,[6] Diane Mathis,[6] Christophe Benoist,[6] Taras Kreslavsky,[7] Anne Fletcher,[7] Kutlu Elpek,[7] Angelique Bellemare-Pelletier,[7] Deepali Malhotra,[7] Shannon Turley,[7] Jennifer Miller,[8] Brian Brown,[8] Miriam Merad,[8] Emmanuel L. Gautier,[8,9] Claudia Jakubzick,[8] Gwendalyn J. Randolph,[8,9] Paul Monach,[10] Adam J. Best,[11] Jamie Knell,[11] Ananda Goldrath,[11] Vladimir Jojic,[12] Daphne Koller,[12] David Laidlaw,[13] Jim Collins,[14] Roi Gazit,[15] Derrick J. Rossi,[15] Nidhi Malhotra,[16] Katelyn Sylvia,[16] Joonsoo Kang,[16] Natalie A. Bezman,[17] Joseph C. Sun,[17] Gundula Min-Oo,[17] Charlie C. Kim,[17] Lewis L. Lanier[17]

[1]Skirball Institute of Biomolecular Medicine, New York University School of Medicine, New York, NY 10016. [2]Fox Chase Cancer Center, Philadelphia, PA 19111. [3]Broad Institute and Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142. [4]Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Boston, MA 02115. [5]Joslin Diabetes Center, Boston, MA 02215. [6]Division of Immunology, Department of Microbiology and Immunobiology, Harvard Medical School, Boston, MA 02115. [7]Dana Farber Cancer Institute and Harvard Medical School, Boston, MA 02115. [8]Icahn Medical Institute, Mount Sinai Hospital, New York, NY 10029. [9]Department of Pathology and Immunology, Washington University, St. Louis, MO 63110. [10]Department of Medicine, Boston University, Boston, MA 02118. [11]Division of Biological Sciences, University of California San Diego, La Jolla, CA 92093. [12]Computer Science Department, Stanford University, Stanford, CA 94305. [13]Computer Science Department, Brown University, Providence, RI 02912. [14]Department of Biomedical Engineering, Howard Hughes Medical Institute, Boston University, Boston, MA 02115. [15]Program in Molecular Medicine, Children's Hospital, Boston, MA 02115. [16]Department of Pathology, University of Massachusetts Medical School, Worcester, MA 01655. [17]Department of Microbiology and Immunology, University of California San Francisco, San Francisco, CA 94143.

Disclosures

The authors have no financial conflicts of interest.

References

1. Beck, J. A., S. Lloyd, M. Hafezparast, M. Lennon-Pierce, J. T. Eppig, M. F. Festing, and E. M. Fisher. 2000. Genealogies of mouse inbred strains. *Nat. Genet.* 24: 23–25.
2. Wade, C. M., and M. J. Daly. 2005. Genetic variation in laboratory mice. *Nat. Genet.* 37: 1175–1180.
3. Waterston, R. H., K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, et al; Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420: 520–562.
4. Keane, T. M., L. Goodstadt, P. Danecek, M. A. White, K. Wong, B. Yalcin, A. Heger, A. Agam, G. Slater, M. Goodson, et al. 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* 477: 289–294.
5. Bogue, M. A., and S. C. Grubb. 2004. The Mouse Phenome Project. *Genetica* 122: 71–74.
6. Heng, T. S., and M. W. Painter; Immunological Genome Project Consortium. 2008. The Immunological Genome Project: networks of gene expression in immune cells. *Nat. Immunol.* 9: 1091–1094.
7. Jojic, V., T. Shay, K. Sylvia, O. Zuk, X. Sun, J. Kang, A. Regev, D. Koller, A. J. Best, J. Knell, et al; Immunological Genome Project Consortium. 2013. Identification of transcriptional regulators in the mouse immune system. *Nat. Immunol.* 14: 633–643.
8. Montgomery, S. B., and E. T. Dermitzakis. 2011. From expression QTLs to personalized transcriptomics. *Nat. Rev. Genet.* 12: 277–282.
9. Civelek, M., and A. J. Lusis. 2014. Systems genetics approaches to understand complex traits. *Nat. Rev. Genet.* 15: 34–48.
10. Gerrits, A., Y. Li, B. M. Tesson, L. V. Bystrykh, E. Weersing, A. Ausema, B. Dontje, X. Wang, R. Breitling, R. C. Jansen, and G. de Haan. 2009. Expression quantitative trait loci are highly sensitive to cellular differentiation state. *PLoS Genet.* 5: e1000692.
11. Orozco, L. D., B. J. Bennett, C. R. Farber, A. Ghazalpour, C. Pan, N. Che, P. Wen, H. X. Qi, A. Mutukulu, N. Siemers, et al. 2012. Unraveling inflammatory responses using systems genetics and gene-environment interactions in macrophages. *Cell* 151: 658–670.
12. Bennett, B. J., C. R. Farber, L. Orozco, H. M. Kang, A. Ghazalpour, N. Siemers, M. Neubauer, I. Neuhaus, R. Yordanova, B. Guan, et al. 2010. A high-resolution association mapping panel for the dissection of complex traits in mice. *Genome Res.* 20: 281–290.
13. Aylor, D. L., W. Valdar, W. Foulds-Mathes, R. J. Buus, R. A. Verdugo, R. S. Baric, M. T. Ferris, J. A. Frelinger, M. Heise, M. B. Frieman, et al. 2011. Genetic analysis of complex traits in the emerging Collaborative Cross. *Genome Res.* 21: 1213–1222.
14. McClurg, P., J. Janes, C. Wu, D. L. Delano, J. R. Walker, S. Batalov, J. S. Takahashi, K. Shimomura, A. Kohsaka, J. Bass, et al. 2007. Genomewide association analysis in diverse inbred mice: power and population structure. *Genetics* 176: 675–683.
15. Davis, M. M. 2008. A prescription for human immunology. *Immunity* 29: 835–838.
16. Payne, K. J., and G. M. Crooks. 2007. Immune-cell lineage commitment: translation from mice to humans. *Immunity* 26: 674–677.
17. Odom, D. T., R. D. Dowell, E. S. Jacobsen, W. Gordon, T. W. Danford, K. D. MacIsaac, P. A. Rolfe, C. M. Conboy, D. K. Gifford, and E. Fraenkel. 2007. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat. Genet.* 39: 730–732.

18. Ravasi, T., H. Suzuki, C. V. Cannistraci, S. Katayama, V. B. Bajic, K. Tan, A. Akalin, S. Schmeier, M. Kanamori-Katayama, N. Bertin, et al. 2010. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* 140: 744–752.
19. Shay, T., V. Jojic, O. Zuk, K. Rothamel, D. Puyraimond-Zemmour, T. Feng, E. Wakamatsu, C. Benoist, D. Koller, and A. Regev; 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.
20. Lee, M. N., C. Ye, A. C. Villani, T. Raj, W. Li, T. M. Eisenhaure, S. H. Imboywa, P. I. Chipendo, F. A. Ran, K. Slowikowski, et al. 2014. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* 343: 1246980.
21. Raj, T., K. Rothamel, S. Mostafavi, C. Ye, M. N. Lee, J. M. Replogle, T. Feng, M. Lee, N. Asinovski, I. Frohlich, et al. 2014. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science* 344: 519–523.
22. Ye, C. J., T. Feng, H. K. Kwon, T. Raj, M. T. Wilson, N. Asinovski, C. McCabe, M. H. Lee, I. Frohlich, H. I. Paik, et al. 2014. Intersection of population variation and autoimmunity genetics in human T cell activation. *Science* 345: 1254665.
23. Kirby, A., H. M. Kang, C. M. Wade, C. Cotsapas, E. Kostem, B. Han, N. Furlotte, E. Y. Kang, M. Rivas, M. A. Bogue, et al. 2010. Fine mapping in 94 inbred mouse strains using a high-density haplotype resource. *Genetics* 185: 1081–1095.
24. Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin. 2008. Efficient control of population structure in model organism association mapping. *Genetics* 178: 1709–1723.
25. Devlin, B., and K. Roeder. 1999. Genomic control for association studies. *Biometrics* 55: 997–1004.
26. Churchill, G. A., and R. W. Doerge. 1994. Empirical threshold values for quantitative trait mapping. *Genetics* 138: 963–971.
27. Stranger, B. E., S. B. Montgomery, A. S. Dimas, L. Parts, O. Stegle, C. E. Ingle, M. Sekowska, G. D. Smith, D. Evans, M. Gutierrez-Arcelus, et al. 2012. Patterns of *cis* regulatory variation in diverse human populations. *PLoS Genet.* 8: e1002639.
28. Storey, J. D., W. Xiao, J. T. Leek, R. G. Tompkins, and R. W. Davis. 2005. Significance analysis of time course microarray experiments. *Proc. Natl. Acad. Sci. USA* 102: 12837–12842.
29. Battle, A., S. Mostafavi, X. Zhu, J. B. Potash, M. M. Weissman, C. McCormick, C. D. Haudenschild, K. B. Beckman, J. Shi, R. Mei, et al. 2014. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* 24: 14–24.
30. Kang, H. M., C. Ye, and E. Eskin. 2008. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics* 180: 1909–1925.
31. Orozco, L. D., S. J. Cokus, A. Ghazalpour, L. Ingram-Drake, S. Wang, A. van Nas, N. Che, J. A. Araujo, M. Pellegrini, and A. J. Lusis. 2009. Copy number variation influences gene expression and metabolic traits in mice. *Hum. Mol. Genet.* 18: 4118–4129.
32. Hosseini, M., L. Goodstadt, J. R. Hughes, M. S. Kowalczyk, M. de Gobbi, G. W. Otto, R. R. Copley, R. Mott, D. R. Higgs, and J. Flint. 2013. Causes and consequences of chromatin variation between inbred mice. *PLoS Genet.* 9: e1003570.
33. Champsaur, M., and L. L. Lanier. 2010. Effect of NKG2D ligand expression on host immune responses. *Immunol. Rev.* 235: 267–285.
34. Petkov, P. M., Y. Ding, M. A. Cassell, W. Zhang, G. Wagner, E. E. Sargent, S. Asquith, V. Crew, K. A. Johnson, P. Robinson, et al. 2004. An efficient SNP system for mouse genome scanning and elucidating strain relationships. *Genome Res.* 14: 1806–1811.
35. Morse, III, H. C. 1978. Introduction. In *Origins of inbred mice: proceedings of a workshop, Bethesda, Maryland, February 14-16*. H. C. Morse, ed. Academic Press, New York, p. 19–20.
36. Kikutani, H., and S. Makino. 1992. The murine autoimmune diabetes model: NOD and related strains. *Adv. Immunol.* 51: 285–322.
37. Zheng, Q. Y., K. R. Johnson, and L. C. Erway. 1999. Assessment of hearing in 80 inbred strains of mice by ABR threshold analyses. *Hear. Res.* 130: 94–107.
38. Grubb, S. C., C. J. Bult, and M. A. Bogue. 2014. Mouse phenome database. *Nucleic Acids Res.* 42: D825–D834.
39. Montgomery, S. B., M. Sammeth, M. Gutierrez-Arcelus, R. P. Lach, C. Ingle, J. Nisbett, R. Guigo, and E. T. Dermitzakis. 2010. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464: 773–777.
40. Nica, A. C., L. Parts, D. Glass, J. Nisbet, A. Barrett, M. Sekowska, M. Travers, S. Potter, E. Grundberg, K. Small, et al; MuTHER Consortium. 2011. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet.* 7: e1002003.
41. Grundberg, E., K. S. Small, A. K. Hedman, A. C. Nica, A. Buil, S. Keildson, J. T. Bell, T. P. Yang, E. Meduri, A. Barrett, et al; Multiple Tissue Human Expression Resource (MuTHER) Consortium. 2012. Mapping *cis*- and *trans*-regulatory effects across multiple tissues in twins. *Nat. Genet.* 44: 1084–1089.
42. Flutre, T., X. Wen, J. Pritchard, and M. Stephens. 2013. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet.* 9: e1003486.
43. Fairfax, B. P., S. Makino, J. Radhakrishnan, K. Plant, S. Leslie, A. Dilthey, P. Ellis, C. Langford, F. O. Vannberg, and J. C. Knight. 2012. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* 44: 502–510.
44. Segal, E., M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* 34: 166–176.
45. Gardner, T. S., D. di Bernardo, D. Lorenz, and J. J. Collins. 2003. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* 301: 102–105.
46. Bar-Joseph, Z., G. K. Gerber, T. I. Lee, N. J. Rinaldi, J. Y. Yoo, F. Robert, D. B. Gordon, E. Fraenkel, T. S. Jaakkola, R. A. Young, and D. K. Gifford. 2003. Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.* 21: 1337–1342.
47. Zhu, J., B. Zhang, E. N. Smith, B. Drees, R. B. Brem, L. Kruglyak, R. E. Bumgarner, and E. E. Schadt. 2008. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat. Genet.* 40: 854–861.
48. Miller, J. A., S. Horvath, and D. H. Geschwind. 2010. Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proc. Natl. Acad. Sci. USA* 107: 12698–12703.
49. Chan, E. T., G. T. Quon, G. Chua, T. Babak, M. Trochesset, R. A. Zirngibl, J. Aubin, M. J. Ratcliffe, A. Wilde, M. Brudno, et al. 2009. Conservation of core gene expression in vertebrate tissues. *J. Biol.* 8: 33.
50. Zheng-Bradley, X., J. Rung, H. Parkinson, and A. Brazma. 2010. Large scale comparison of global gene expression patterns in human and mouse. *Genome Biol.* 11: R124.
51. Su, A. I., T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* 101: 6062–6067.
52. Strand, A. D., A. K. Aragaki, Z. C. Baquet, A. Hodges, P. Cunningham, P. Holmans, K. R. Jones, L. Jones, C. Kooperberg, and J. M. Olson. 2007. Conservation of regional gene expression in mouse and human brain. *PLoS Genet.* 3: e59.
53. Enard, W., P. Khaitovich, J. Klose, S. Zöllner, F. Heissig, P. Giavalisco, K. Nieselt-Struwe, E. Muchmore, A. Varki, R. Ravid, et al. 2002. Intra- and interspecific variation in primate gene expression patterns. *Science* 296: 340–343.
54. Ohmura, K., A. Johnsen, A. Ortiz-Lopez, P. Desany, M. Roy, W. Besse, J. Rogus, M. Bogue, A. Puech, M. Lathrop, et al. 2005. Variation in IL-1β gene expression is a major determinant of genetic differences in arthritis aggressivity in mice. *Proc. Natl. Acad. Sci. USA* 102: 12489–12494.
55. Lappalainen, T., M. Sammeth, M. R. Friedländer, P. A. 't Hoen, J. Monlong, M. A. Rivas, M. Gonzàlez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, et al; Geuvadis Consortium. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501: 506–511.
56. Ferraro, A., A. M. D'Alise, T. Raj, N. Asinovski, R. Phillips, A. Ergun, J. M. Replogle, A. Bernier, L. Laffel, B. E. Stranger, et al. 2014. Interindividual variation in human T regulatory cells. *Proc. Natl. Acad. Sci. USA* 111: E1111–E1120.
57. Georgi, B., B. F. Voight, and M. Bućan. 2013. From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. *PLoS Genet.* 9: e1003484.
58. Parham, P., and T. Ohta. 1996. Population biology of antigen presentation by MHC class I molecules. *Science* 272: 67–74.