Immune cellular homeostasis in early life is determined by genetic variants of cellular production and turnover

Tania Dubovik[1], Elina Starosvetsky[1], Benjamin LeRoy[2], Rachelly Normand[1], Yasmin Admon[1], Ayelet Alpert[1], Yishai Ofran[1,3], Max G'Sell[2], Shai S. Shen-Orr[1]

[1] Department of Immunology, Faculty of Medicine, Technion - Israel Institute of Technology, Haifa, Israel.

[2] Department of Statistics, Carnegie Mellon University, Pittsburgh, USA.

[3] Department of Hematology, Rambam Health Care Campus, Haifa, Israel.

Abstract

Variation in immune cellular homeostasis exists between individuals, is heritable and due to genes and environment. Environmental influences increase with age, masking genetic determinants and barring human adult studies from informing on the origin of baseline variability. To identify genetic factors affecting individuals' immune profile in early life we profiled the bone marrow of 55 genetically diverse Collaborative Cross mouse strains, using high resolution mass-cytometry. We identified 1,662 genes associated with 11 well defined cell subsets, 862 of which validated in an independent cohort. These genes were strongly enriched for basic housekeeping functions of proliferation and death, predicted cellular abundance and showed a higher mutation rate across multiple human cancers. Thus, akin to mRNA or protein species abundance, cellular variation is determined by genetic variants of production and turnover genes. Given these affect baseline homeostasis, reaching an at-risk immune state would differ for individuals, set by the genetic determinants we identified here.

Introduction

The immune system has a major impact on maintaining our health and preventing the state of disease. Measurements in both humans and mice show a large amount of variation between individuals in cell subset abundance and relative ratios, suggesting a "personalized immune homeostasis". This variation has been exploited for blood based cellular biomarkers yielding important diagnostic health-information; not only of disease (Brodin and Davis, 2017), but also of outcomes among the healthy, such as predicting vaccine response or surgery outcomes (Gaudillière *et al.*, 2014; Tsang *et al.*, 2014).

These cellular immune profiles are highly heritable, suggesting a strong role for environment and genetics in determining these phenotypic traits (Lu *et al.*, 2016). Though data from young twins suggests a high level of genetic determinism (Hall *et al.*, 2000; Pedersen, 2000), the effects of the environment increase with age, eventually dominating over the genetic components (Brodin *et al.*, 2015). In agreement, several large-scale association studies have identified genetic variants of disease genes associated with blood cell subset frequencies and counts (Orrù *et al.*, 2013; Roederer *et al.*, 2015). Yet these appear to be more associated with an individual's life history than to the basic biological processes that drive determination of a cell's abundance, suggesting that the source of variation in immune cellular profile remains unknown. Evidence from basic studies tracking labeled cells have shown that when one changes the rate of a cell subsets' proliferation or death, the total number and relative frequencies of cells would be altered (Mohri *et al.*, 1998; Asquith *et al.*, 2002; Busch *et al.*, 2015) . Yet these findings do not explain the source of individual variability in immune cellular profiles. Thus, identifying genetic determinants of immune cell profiles is at the root of understanding how an individual's immune system state arises and ultimately what biomarkers are 'made of'.

Immune phenotypes are predominantly considered complex traits. Successful mapping of such traits often requires an interplay between power and resolution of mapping: using humans who are genetically diverse as a model reduces the power of mapping, resulting in inability to detect relevant genomic loci, whereas using F2 populations of classical laboratory strains yields a limited resolution due to low sequence diversity (Mott and Flint, 2013). The Collaborative Cross (CC) mice strains form a large panel of inbred mouse strains that are derived from eight inbred founder strains (five classical inbred and three wild-derived strains). The CC mice were outbred in an 8-way cross for two generations and then inbred by inter-sibling mating to achieve

individual genetic homozygosity while maintaining heterogeneity between strains, which is closer to that observed in humans than previously established mice resources (Rogala *et al.*, 2014; Elbahesh and Schughart, 2016). These mice share a similar environment yet are genetically diverse, enabling increased statistical power and focus for gene-trait correlation studies (Srivastava *et al.*, 2017). Moreover, the CC mice have been shown to exhibit large phenotypic variation with respect to immune homeostatic conditions (Graham *et al.*, 2017) and related responses (Rasmussen *et al.*, 2014; Gralinski *et al.*, 2015; Snijders *et al.*, 2016; Brinkmeyer-Langford *et al.*, 2017), such as infection or viral challenge, offering an important venue to study the maintenance and variation of immune system homeostasis and its genetic determinants.

With the aim of identifying the genetic origins of homeostatic variability in early life we coupled the high genetic diversity of the Collaborative Cross mice, which bring statistical power, a clean environment and reproducibility, with the high resolution of mass cytometry (CyTOF), which brings well defined characterization of entire cellular profiles. We identify the involvement of 1662 genes associated with 11 well defined cell subsets and show that baseline variability in immune cell profiles is predominantly determined by genetic variants of cellular proliferation and death genes, information which we validate in a second cohort of mice and show can be used to predict cell abundance and proliferation rates. The majority of these associations are cell subset specific, suggesting unique mechanism of regulation, however as a group they show increased mutation rates across many forms of cancers, suggesting they may be co-opted by tumors to alter production and turnover rates across many cell-types.

Results

A high variability in immune cellular homeostasis in CC mice raised in a clean environment

To assess the non-environmental variation in immune cell subset frequencies we focused on the bone marrow, given it covers major developmental milestones of hematopoiesis. Fitting the need for profiling this complex cellular environment at high dimension, we profiled 30 naive CC mouse strains in duplicate and 8 founder strains in triplicate (See Supp. Table 1) using a large panel of CyTOF phenotypic markers, reflective of hematopoietic populations (See Supp. Table 2 for panel). We noted that for several strains, some markers were unobservable, making it difficult to cluster cell subsets of these mice at high dimension (Supp. Table 3). For the remaining samples, which had a full set of phenotypic markers (15 CC strains and 3 founder strains), we clustered the single cell data using Citrus, a high-dimensional clustering algorithm (Bruggner *et*

al., 2014) (Fig 1A, See Supp. Table 4). Of note, even for these mice expressing the full set of markers, we observed a high variability in marker expression, which could not be fixed via scaling and resulted in immune profile of CC mice clustered by replicate strains, followed by a close similarity to the whole genome genetic kinship distance, suggesting that phenotypic similarity was driven by genetic similarity (Fig 1B). To avoid bias in identification of cluster identity and allow downstream integration of strains which could not be clustered, we leveraged the Scaffold algorithm which maps high dimensional clusters to manually gated populations (Spitzer *et al.*, 2015) (Fig 1C, Supp. Figure 1 for gating scheme, see Methods).

Next, to assess homeostatic variation across mice, we analyzed individual cell subset frequencies across both founder and CC strains. As has previously been noted for other phenotypes (Kelada *et al.*, 2012), CC mice exhibited a continuous range of frequencies whose dynamic range was larger than that observed between the most extreme founder strains, suggesting epistatic interactions (Fig 1C, inset)(Srivastava *et al.*, 2017). An individual's immune cell profile consists of a complex stoichiometry covering multiple cell subsets. As such, to assess the variation between individual mice's homeostatic composition, we performed a Principle Component Analysis (PCA) of bone marrow cell subset composition (Figure 1D). We noted that the first two principle components axes explained 48.3% of total variance, with the first component capturing variation in lymphoid lineage B and T cell subsets and the second component capturing variation in myeloid cell subsets and NK cells. To quantify the extent of variation in homeostatic composition, we measured the pairwise distance between individual mice profiles within replicate mice, founder and CC strains. Here too, CC strains created a continuous scale of phenotypes likely introduced by epistatic events, whose dynamic range was largest and mean distance significantly greater than that observed between replicate strains (Figure 1E, $p<0.001$ by Student's t-test). Interestingly, we noted no significant differences of homeostatic composition distances between bone marrow samples profiled from ten healthy adults and CC strains, supporting the observation that these strains phenotypic variation is close to that observed in humans (Figure 1E, Supp. Figure 2 for human gating scheme). In addition, to understand the extent the observed variability defined an individual 'fingerprint', we profiled spleen of CC and founder mice strains and compared the similarity between individual mice across tissues. To do so, we measured the distance of each strain's immune profile from that of C57BL/6J on the same tissue (see Methods). We observed a positive correlation of homeostatic composition differences between mice across tissues (Figure 1F, Sup. Figure 3, $r=0.41$), suggesting that cellular differentiation, as reflected in the spleen, is a direct function of individual bone marrow composition. Taken together, immune cell subset

variation is high, yet non-random, in genetically diverse mice raised in a shared clean environment and is likely driven by epistatic genetic trait interactions.

Immune cell subset frequencies are complex traits associated with multiple genomic loci

The high phenotypic variation observed in CC mouse strains provides a good platform for detection of quantitative trait loci responsible for immune traits. To identify those genes associated with immune cellular homeostasis, while reducing the number of multiple hypotheses to be tested, we chose to focus on loci harboring genes expressed in immune cells and for which mutations in the exon region were predicted (see Methods, Supp. Table 5). This filtering procedure yielded 6,961 genes covering a broad set of functions, spanning from broad cellular to immune specific functionalities. Next, we mapped per mouse loci, the likelihood of it stemming from a particular founder, using haplotype reconstruction (see Methods) and computed the logs odds ratio (LOD) score of each locus with each of the manually gated cell subset populations (Figure 2A). Of note, as manual gating depends on several representative markers (see gating scheme in Supp. Figure 1) and some mice did not show any pattern of expression for a marker of interest, we could not gate for all populations in all mouse strains and the number of mice used for associations varied between cell subsets (range 44 to 59, Supp. table 3). This procedure yielded a total of 1,619 loci, corresponding to 1,578 genes which matched our filtering criteria, namely below a false discovery rate (FDR) threshold of 15% and passing a leave-one-out procedure (Figure 2B, Supp. Figure 4, Supp. Table 6, see Methods).

While genetic mosaic strains offer high power for statistical association, they also suffer from technical limitations introduced by their non-standard phenotype. Our inability to characterize cell subset populations included a staining loss of 6 markers, CD45, CD43, Sca-1, IA-IE, and Ly6c, one or more of which were undetectable in various combinations across strains. For each marker, we observed at least one founder strain for which the marker expression was undetectable in the stained panel (Figure 2C for CD43 as an example). This suggested that the inability to stain these markers was not due to a new recombination events, but rather was likely a consequence of differences in allelic variants, yielding an inability for epitope detection by the commonly used antibody isoforms (Philbrick *et al.*, 1990). We reasoned that this loss of signal may be overcome using a statistical learning approach that would leverage the large phenotypic panel of CyTOF markers to approximate cell subsets frequencies unmeasurable by manual gating. We focused on CD43, critical for differentiating B-cell subsets in our CyTOF panel (Figure 2C inset) and for which the number of associated genes we identified was low, particularly for early B-cells,

suggesting that we lacked statistical power. Noting that CD19 and CD44 were correlated with CD43, we trained a random forest classifier on those mice whose markers were non-missing (Figure 2C-D, See Methods). The accuracy of this point estimate procedure was 82% (Figure 2E). Next, using this learned classifier we estimated early B-cell proportions in those strains where CD43 expression was un-detectable (Supp. Table 7). Finally, we repeated our genetic association analyses for early B cell subsets, but now adding those strains for which cell subset proportion were estimated. This increased our statistical power and allowed identification of an additional 83 loci, corresponding to 84 genes in early B-cell subsets (Supp. Table 8, see Methods). Though the actual LOD score range for these was lower compared to the rest of associations, newly detected loci remained stable following multiple repeat randomizations in each of which we introduced error, on par with the training set estimation error (see Methods). Thus, the technical difficulties of assaying antibody-based measurements in genetically diverse individuals may be overcome by algorithmic means and a smart antibody panel design, allowing us to reach a total 1,662 genes whose variants were associated with immune cell homeostasis.

Immune cell subset frequencies are regulated by genetic networks that are predominantly cell-specific

The regulation of immune cell subsets homeostasis may be common to all cell subsets, or unique per population or lineage, based on functionality and developmental constraints. We noted that the number of associated hits was not uniformly distributed across immune cell subsets, rather, we detected over 700 genes whose variants were associated with monocytes but only 23 genes whose variants were associated with Pro-B cells (Figure 3A, Supp. Table 9). Of interest, our association picked up 45 genes whose phenotype according to Mammalian Phenotype Ontology is known to affect leukocyte levels, for example genes specific to the immune system: *PAX5, CD28, Ltbp1* and genes with more housekeeping functionality: *Myo3b*, *Hnrnpr* and others (for full list see Supp. Table 10). This imbalance in trait-gene associations may be due to differences in regulatory complexity, yet is most likely is due to multiple complex issues involving both biological and technical limitations in experimental design: differences in allelic composition between genes across CC strains, unequal sample size per cell subset and more. To overcome these issues and increase the strength of detected associations, we repeated our analysis in a second independently collected cohort of 24, non-overlapping, mouse CC strains, whose bone marrow we profiled by CyTOF similar to the first cohort (See Methods, Supp. Table 11-12). This yielded successful validation of 862 genes (52%) from the initially detected set (Figure 3B, Supp.

Table 13 for adjusted p-values). Of note, we observed the highest percent of validated gene-trait associations in the B cell lineage, possibly hinting on a more highly conserved regulatory mechanism. In contrast, for T cells, we observed a lower percentage of validated genes, possibly as their presence in the bone marrow is due to migration and not differentiation processes (Di Rosa and Pabst, 2005; Mazo *et al.*, 2005).

Genetic regulation of a cell's homeostatic condition may be determined by the cell subset itself, or by other cell subsets. For the majority (84%) of genetic variants we detected an association only in a single cell subset. To delineate the relationship between a gene's cellular genetic association and its expression pattern we checked each gene's cellular expression profile across the ImmGen sorted cell gene expression compendium (Heng and Painter, 2008) (see Methods). We detected the overwhelming majority of the genes whose variants were associated with a specific cell subset frequency to be expressed in that same cell subset, that is in *cis* (Figure 3C, mean 86.5%, ranging between 83-100% across subsets). Whereas, 1.5% of genes were expressed in the same lineage, but not in the associated cell subset, and an additional 5% not expressed by any of the tested immune subsets. The remaining 12% were *trans*, that is, genes whose variants were associated with a specific cell subset frequency, but not expressed in that cell or in other subsets in the same lineage. The distribution of genes between these three non-*cis* genes sets was expression threshold dependent yet irrespective of the threshold, the existence of these four groups could not be discounted and robust to false discovery assessments. Taken together this would suggest that a cell's homeostatic condition is determined using distinct regulatory mechanisms per cell subset which are predominantly expressed in that cell.

Genes associated with immune cell subset frequencies are enriched for cell production and turnover functions and highly overlap with mutated genes in various cancers.

To characterize the biological processes of these genetic associations, we performed a pathway enrichment analysis (Figure 4A, Supp. Table 14 for significantly enriched functions, threshold at $p<0.01$ BH adjustment). At this cutoff, the functionally enriched gene set consisted of 310 genes of which 137 (44%) were annotated for one of four main processes: proliferation, cell death, differentiation and cellular movement (*q-value*=10^{-8}, 10^{-5}, 10^{-4}, 10^{-10} respectively by hypergeometric test). Beyond this, we observed an enrichment for other immune functions, including those responsible for homeostatic balance, which highly overlapped in gene membership with that of those four functions (68%). Of note, genes with proliferation and death

functionality that were found were both innate immunity oriented such as *C5AR1*, *MYD88* various *TLR* molecules (2,3 and4) and adaptive derivatives as *IL7r* and *CTLA4.* In addition, we observed a group of genes which are commonly known to affect proliferation processes as *CD69* and apoptotic processes such as *Capase3,8* and *BCL2*, and a considerable group (10%) of cytokines, chemokines and phosphorylated proteins. Analyzing the distribution of these functions between cells, each cell subset included genes with at least one of four functionalities; whereas from a gene perspective, genes in this set were associated with only a single cell subset (Figure 4B). Per cell subset network analysis, identified functionally connected networks in each cell subset, suggesting that genes associated with a cell subset's frequency act synergistically to balance its frequency. For example, analysis of late maturation stage ($CD43^-$) B cell associated genes, identified a genetic network covering B cell differentiation genes, with *Id3* forming a hub protein, as well as more general functionality genes, such as *BCL2*, involved roles in proliferation and apoptosis processes (Figure 4C, Supp. Figure 5 for more cell specific networks). Noting this highly specific gene-cell subset relation for what are generally considered housekeeping type functionalities (i.e. performing similar functionalities in most cell-types); per cell, we next contrasted the LOD score distribution of genes associated to only a single cell subset, that is, below LOD cut-off in other cell subsets, with that a random set of genes of equal size. We observed significantly higher LOD scores for "cell-specific genes associations" across multiple cell-types (Supp. Table 15 for *p-values* by Student's t-test), suggesting that cell-subset-association genes may also play a lesser role in the biological processes involved in other cell subsets, and that homeostatic cell regulation forms a large network with both cell-specific and pan-cell functionalities.

Given these findings, this suggests, that akin to mRNA or protein species abundance, cell subset abundance is determined by genetic variants of genes associated with production or turnover. We thus hypothesized that individuals with similar complex genetic profiles of proliferation genes, would show similar cellular proliferation rates and cell abundances. To test this, we injected IdU to the second cohort of CC mice and assayed them following 48hr (2 injections at 24 hours' interval). IdU incorporates into DNA strands and its level increases as more DNA synthesis occurs such that one can estimate the percent of proliferating cells in each cell subset by measuring number of cells in S phase (Supp. Table 16). Next, we clustered mice based on their genetic distance from one another, defined by their shared genetic variants of the validated (Figure 3B) proliferation genes of a specific cell-type, and clustered the individuals (Figure 5A for late B-cells, Supp. Figure 6 for monocytes). We observed that the mice having PWK/PhJ or

CAST/Eij founder strains contributing to the allelic variants of genes associated with the late B-cell subset, had significantly higher proliferation rates ($p<0.05$) and higher subset abundance ($p<0.01$, Figure 4C). Of interest, genes that showed higher effect on the resulting abundance had basic functionality as *Batf* and *Dicer1* or were previously found to regulate neuro related homeostatic and disease processes *Nf1*, *ADORA2b*, *NOS1*, suggesting that the regulation of the immune cell subset frequency might be affected by neuronal signaling as well.

Given the homeostatic functionality we detected in the associated genes, we reasoned that their mutation may play a role in diseases in which the normal process of proliferation and/or death is disrupted. We thus leveraged information regarding patient somatic mutations in various cancers from TCGA, including both solid and blood tumors. Across the majority of cancers, the number of mutations accumulated in this gene set was at the top percentile compared to equally sized random controls, both when compared to genes sampled from the entire human genome as well as when restricted to genes expressed in immune cell only and excluding all known proliferation and death annotated genes (Figure 5B, see Methods).

Discussion

Here we aimed to identify what are the genetic mechanisms that drive variation in immune cell homeostasis between individuals. To do so we coupled the high variability of Collaborative Cross mice with the high dimensional phenotypic capabilities of mass cytometry to identify genetic determinants that are responsible for between organismal homeostatic differences. Using QTL mapping we identified 1662 genes which we associated with one or more of eleven immune cell subset frequencies of mice bone marrow. These set of genes are highly enriched for functions of cell proliferation, death, migration and differentiation. We validated that these epistatic genetic variants drive cell subset frequencies in a second cohort of CC mice, stratified by their genetic profile of proliferation genes. Transferring these findings to humans, we observed that as a group these genes are highly likely to be mutated across multiple cancer types, even when removing from our hit list all previously annotated proliferation and death genes.

The large variation exhibited by CC mice, coupled with their controlled genetics and clean environment make them an ideal model for System Immunology studies. This is due to the fact that the variation between individual mice is large compared to classical laboratory strains and can thus be used to generate meaningful correlations between measured features (e.g. cells, genes, protein or genetic variants) as well as an improved understanding of the biological processes

studied and at a higher statistical rigor. Moreover, for immunology, where environmental exposure may overcome initial genetic variability, the CC mice offer a means at peeking at a 'dark genetics' that would otherwise be difficult, if not impossible.

Work with the CC strain though is not without difficulties, the use of conventional immunological techniques to characterize the mice may be limited due to reagent standards being largely dictated by C57BL/6J strains. To overcome this, we used statistical learning methodology which leverages the high dimensionality of the data to estimate missing signal, an approach that can be generalized. Beyond this, our findings are dependent on measured relative frequency of immune cell subsets rather than the absolute cell count, which is difficult to obtain in the bone marrow. We validated our model by looking at proliferation rates of different immune cell subsets, however to validate the production/turnover equation suggested by the associated genes in full, one should consider apoptosis and migration functions as well, which would require longitudinal studies to infer rates.

Our analysis revealed an unprecedented number of novel associations of genes with immune cell functionality. These describe a network governing homeostatic balance, whose components are connected and both play a functional role in specific cell-types as well as across multiple cells. The majority of these genes act in *cis*, that is, they are expressed in those cells whose abundance they predominantly regulate. Yet, we also observed lineage specific and *trans* mediated regulation. The latter suggest a complex pattern of immune lineage cross-talk which may be mediated by communication mechanisms, such as exosomes and nanotubes (Mccoy-simandle *et al.*, 2017), to allow maintenance of homeostatic stoichiometry. Beyond this, we focused solely on genes whose expression in the immune system is well established, yet genes expressed in the stromal tissue may drive bone marrow formation as well (Wilson and Trumpp, 2006). For the majority of the genes we identified, no functional association with cellular homeostatic control has been reported to date, these would need to be further studied, especially as we observe their enriched mutation across many cancers, suggesting that tumors may mutate these genes preferentially to control growth.

Over time, environmental exposures (e.g. disease, pregnancy, habitation conditions and other) alter one's immune system. Yet, these changes are always relative to the initial starting conditions which are largely dictated by genetic factors as we show here, and are likely unidirectional towards a non-naïve system. Evidence of immune cellular profiles or immune states being of

clinical relevance for diagnostic purposes has been shown across multiple disease conditions and health outcomes. As these states are reached through a developmental trajectories beginning in early life, the genetic determinants we identify here, affect baseline variability and thus how quickly one may reach an at-risk immune state. This would suggest that by checking the relation of immune responses to genetics, we may be able to predict the immune profile of an individual and their susceptibility to disease.

Methods

Mice samples collection and processing

Founder strain males (n=3 per strain) of the A/J, C57BL/6J, 129S1Sv/ImJ, NOD/ShiLtJ, NZO/H1LtJ, CAST/EiJ, PWK/PhJ, and WSB/EiJ strains were purchased from The Jackson Laboratory and sacrificed at 6-8 weeks. For the Collaborative Cross strains, the Systems Genetics Core Facility (University of North Carolina) provided 129 mice aged 8-14 weeks from 55 different complete lines, at least two mice per strain. Mice genetic information can be found at http://csbio.unc.edu/CCstatus/.

Mice were bred and sacrificed at UNC facility. All procedures involving animals were performed according to the Guide for the Care and Use of Laboratory Animals with prior approval by the Institutional Animal Care and Use Committee within the Association for Assessment and Accreditation of Laboratory Animal Care-accredited program at the UNC at Chapel Hill (Animal Welfare Assurance Number: A-3410-01). Bone marrow and spleen were collected from necropsy following humane euthanasia by CO_2.

Primary conjugates of mass cytometry antibodies were prepared using the MaxPAR antibody conjugation kit (Fluidigm Inc.) according to the manufacturer protocol and optimal concentration was determined by titration.

Bone marrow was flushed with cold CSM (PBS + 0.5% BSA) from the femur and tibia using a 27.5-gauge needle and a 10mL syringe to achieve a single cell suspension. Cells from each mouse were washed twice and a total of 3 million cells were used for staining. Spleen tissue was homogenized and red blood cells were eliminated by lysis buffer. Cells were resuspended in 500ul containing 1:2000 Rh DNA intercelator for 20 min of live/dead cells staining. Samples were washed with CSM buffer and resuspended in total of 100 ul metal-tagged antibody mix for cell surface markers staining for 1 hour. Cells were then fixed in 1.6% PFA (Sigma-Aldrich) in a total volume of 200 ul and stored at 4^0C. DNA intercalator Ir191/193 staining was performed

post-PFA removal for 20 min at 1:2000 concentration in 500 ul volume. Finally, fixed samples were washed 3 times with DIW immediately prior to acquisition.

Human samples

Bone marrow aspirates from AML patients presenting at the Hematology Department in Rambam Medical Health Care Campus were collected. All patients gave informed consent according to the declaration of Helsinki (IRB number: 0573-10). Mononuclear cells were separated by centrifugation over a layer of LymphoprepTM (Axis-Shield PoC AS, Oslo, Norway) and then stored in freezing medium [fetal bovine serum (FBS) with 10% DMSO] in a liquid nitrogen tank.

Samples acquisition and data analysis

Samples were acquired using CyTOF 1 machine at 500 events/sec for a total of 100-200K events. Internal metal isotope bead standards were added for sample normalization as described (Finck *et al.*, 2013) to account for the decline in mean marker intensity over time. Acquired data were uploaded to a Cytobank web server (Cytobank Inc.) for data processing and gating out of dead cells and normalization beads. To account for intra-run declines in mean marker intensity over time, we performed a within-sample-over-time normalization step by using a running window to adjust mean marker intensity throughout each individual run such that the mean expression over time was equal to that measured at the beginning of the run.

We manually gated all the major cell populations in the mouse bone marrow (Supp. Figure 1-3). Resulting phenotypes were than exported and adjusted to the total number of cells in the sample. Cell subset frequencies for each mouse sample are summarized in (Supp. Table 4).

Estimation of cell type proportions

To overcome the loss of signal due to allelic variation in a subset of markers, we devised a statistical learning approach to estimate the proportion of each cell type. Broadly, this consisted of two steps. First, a mixture-based distribution alignment methodology tailored to preserving differences in cell-type proportions between strains, while removing batch differences at the marker level. Second, a supervised estimation of cell-type proportions from the batch-corrected marker distributions using a random forest.

Mixture-based distribution alignment - Single-cell protein expression distributions were aligned across mice using a constrained mixture model to capture sample-to-sample variation not handled by bead-based normalization. This mixture approach aligns the means of individual modes in the

distribution, in contrast to standard alignment approaches that rely on sample-wide averages, quantiles, or regression models. The advantage of this approach is that it prevents variation in cell type proportions from confounding the distribution alignment.

Briefly, for a given marker, we modeled the observed distribution within each mouse by a mixture of Gaussians. We constrained the mean and variance structure of this mixture to be identical across mice, up to a global, linear shift of the distribution in each mouse. However, we allowed the mixture weights to vary to account for shifts in cell type proportion. Finally, we aligned the distributions by removing the estimated constant shift from the distribution of each mouse. The mixture models were fit using an Expectation-Maximization algorithm in the FLEXMIX package in R, initializing the model with the modes computed on the data pooled across all mice.

Estimating cell-type proportions with random forest- We used mice expressing all markers to train a random forest classifier to classify cell subsets. We specifically focused on CD43, critical for differentiating B-cell subsets in our CyTOF panel and selected the 3609 CC strain as the strain used for training as it exhibited reliable classification performance across a range of different training mice. We applied the learned classifier to the mice with missing markers to obtain estimated probabilities of being early or late B-cells. Since correctly calibrating these probabilities on a new mouse actually depends on the unknown proportions of early and late B-cells, these scores cannot directly be interpreted as probabilities. However, their relative values on the log scale are unaffected by this issue. Because the classifier performance is high enough that the estimated log scores are highly bimodal, we were able to fit a mixture model to extract accurate proportion estimates, regardless of the lack of calibration. We fit a mixture model on the logit scale of estimated probabilities to extract the relative weight on each component, and used these mixture weights as estimates of the cell type proportions in that mouse. We validated this approach in held-out mice where all gating markers were present and estimated our error in early B-cell populations. These errors were then used during the genetic association step, to ensure that identified loci were stable. Specifically, for each mouse strain in which a cell subset was estimated, we set the frequency of the cell to be the estimated point value minus the estimation errors where the error is taken from a pooled distribution of errors across all training set strains.

Filtering of loci

To reduce the number of multiple hypotheses tests, we chose to focus on associations with genomic loci that passed the following filter criteria: First, those loci which contained genes for which at least one of the CC founder strain is predicted to have an altered protein structure due to a sequence change compared to C57BL6 in exons of protein coding genes (Keane *et al.*, 2011; Yalcin *et al.*, 2011); Second, those genes determined to be expressed in immune cells based(Ericson *et al.*, 2008). Using these two filtering criteria reduced the number of tested loci from 77,725 to 15,470.

Gene-Phenotype Association

We performed QTL mapping using DOQTL R package and MegaMUGA SNP chip (https://csbio.unc.edu/CCstatus/index.py) given the CC genotype reconstruction and the probability of descent from one out of eight founder strains for each genomic interval. We searched for significant association between immune phenotype and each genomic locus, using an additive linear model. To check for stability of observed association we used a *leave-one-out* approach, calculating the strength of association. We only selected the associations which passed the threshold of significance in a strain independent manner. We set a significant association threshold based on the FDR of same population. We adjusted the threshold such that will allow and FDR of 15% and less. We calculated FDR for each phenotype separately by shuffling each time the mice ancestry in the locus and calculating the strength of the association.

Functional classification of genes and validation of the analysis

We performed pathway enrichment analysis with IPA software (Ingenuity Pathway Analysis Qiagen), using the core analysis module. To do so we estimated expression of genes according to ImmGen., using a threshold of log2(47) (Ericson *et al.*, 2008), as a threshold for genes of intermediate and high probability of expression. We used a second cohort of mice to validate our findings. Each significant gene-phenotype association we identified in the first cohort was tested for significance in the second cohort (*p-value* <0.05).

We injected the mice i.p. with 50mg/kg body weight IdU (5-Iodo-2′-deoxyuridine) twice: 48 hours and 24 hours prior to euthanasia. To test for differences in IdU levels we manually gated all the subsets for IdU positive and negative cells. Mice were clustered, per cell-subset, by hierarchal clustering, based on the loci identified as associated with the cell-subset in the first cohort. One-way ANOVA was performed in order to test for difference in percent of proliferating cells ($p<0.05$).

Cancer mutation enrichment

We obtained somatic mutations of cancer patients TCGA. For each cancer type we computed the overlap between genes mutated in that cancer with a random gene set of equal size to the remaining set of identified genes post filtering for filtering for one-to-one orthologues. We performed this computation 100 times to generate a distribution, and then checked the percentile at which the identified set of genes placed. We considered two sampling options, to sample genes out of the whole human genome or only from genes that are expressed in immune cell subsets. To conduct a more conservative test, we checked whether the results remain stable even when taking out genes that were previously functionally annotated as proliferation and death genes by IPA and/or GO. As such, we analyzed the data in four conditions: every combination of background and gene set options.

References

Asquith, B. *et al.* (2002) 'Lymphocyte kinetics: The interpretation of labelling data', *Trends in Immunology*, 23(12), pp. 596–601. doi: 10.1016/S1471-4906(02)02337-2.

Brinkmeyer-Langford, C. L. *et al.* (2017) 'Host genetic background influences diverse neurological responses to viral infection in mice', *Scientific Reports*. Springer US, 7(1), pp. 1–17. doi: 10.1038/s41598-017-12477-2.

Brodin, P. *et al.* (2015) 'Variation in the human immune system is largely driven by non-heritable influences', *Cell*. Elsevier Inc., 160(1–2), pp. 37–47. doi: 10.1016/j.cell.2014.12.020.

Brodin, P. and Davis, M. M. (2017) 'Human immune system variation', *Nature Reviews Immunology*. Nature Publishing Group, 17(1), pp. 21–29. doi: 10.1038/nri.2016.125.

Bruggner, R. V. *et al.* (2014) 'Automated identification of stratifying signatures in cellular subpopulations', *Proceedings of the National Academy of Sciences*, 111(26), pp. E2770–E2777. doi: 10.1073/pnas.1408792111.

Busch, K. *et al.* (2015) 'Fundamental properties of unperturbed haematopoiesis from stem cells in vivo', *Nature*, 518(7540), pp. 542–546. doi: 10.1038/nature14242.

Elbahesh, H. and Schughart, K. (2016) 'Genetically diverse CC-founder mouse strains replicate the human influenza gene expression signature', *Scientific Reports*. Nature Publishing Group, 6, pp. 1–5. doi: 10.1038/srep26437.

Ericson, J. *et al.* (2008) 'ImmGen microarray gene expression data : Data Generation and Quality Control pipeline .', *http://www.immgen.org/Protocols/ImmGen%20QC%20Documentation_ALL-DataGeneration_0612.pdf*.

Finck, R. *et al.* (2013) 'Normalization of mass cytometry data with bead standards', *Cytometry Part A*. Wiley Online Library, 83A(5), pp. 483–494. doi: 10.1002/cyto.a.22271.

Gaudillière, B. *et al.* (2014) 'Clinical recovery from surgery correlates with single-cell immune

signatures', *Science Translational Medicine*, 6(255). doi: 10.1126/scitranslmed.3009701.

Graham, J. B. *et al.* (2017) 'Extensive Homeostatic T Cell Phenotypic Variation within the Collaborative Cross', *Cell Reports*. ElsevierCompany., 21(8), pp. 2313–2325. doi: 10.1016/j.celrep.2017.10.093.

Gralinski, L. E. *et al.* (2015) 'Genome Wide Identification of SARS-CoV Susceptibility Loci Using the Collaborative Cross', *PLoS Genetics*, 11(10), pp. 1–21. doi: 10.1371/journal.pgen.1005504.

Hall, M. A. *et al.* (2000) 'Genetic influence on peripheral blood T lymphocyte levels', pp. 423–427.

Heng, T. S. P. and Painter, M. W. (2008) 'The Immunological Genome Project: networks of gene expression in immune cells.', *Nature immunology*, 9(10), pp. 1091–4. doi: 10.1038/ni1008-1091.

Keane, T. M. *et al.* (2011) 'Mouse genomic variation and its effect on phenotypes and gene regulation', *Nature*, 477(7364), pp. 289–294. doi: 10.1038/nature10413.

Kelada, S. N. P. *et al.* (2012) 'Genetic analysis of hematological parameters in incipient lines of the collaborative cross.', *G3 (Bethesda, Md.)*, 2(2), pp. 157–65. doi: 10.1534/g3.111.001776.

Lu, Y. *et al.* (2016) 'Systematic Analysis of Cell-to-Cell Expression Variation of T Lymphocytes in a Human Cohort Identifies Aging and Genetic Associations', *Immunity*, 45(5), pp. 1162–1175. doi: https://doi.org/10.1016/j.immuni.2016.10.025.

Mazo, I. B. *et al.* (2005) 'Bone marrow is a major reservoir and site of recruitment for central memory CD8+T cells', *Immunity*, 22(2), pp. 259–270. doi: 10.1016/j.immuni.2005.01.008.

Mccoy-simandle, K. *et al.* (2017) 'HHS Public Access', pp. 44–54. doi: 10.1016/j.biocel.2015.12.006.Exosomes.

Mohri, H. *et al.* (1998) 'Rapid turnover of T lymphocytes in SIV-infected macaques.', *Science*, 279(February), pp. 1223–1227.

Mott, R. and Flint, J. (2013) 'Dissecting quantitative traits in mice', *Annu.Rev Genomics Hum.Genet.*, 14(1545–293X (Electronic)), pp. 421–439. doi: 10.1146/annurev-genom-091212-153419.

Orrù, V. *et al.* (2013) 'Genetic variants regulating immune cell levels in health and disease.', *Cell*, 155(1), pp. 242–56. doi: 10.1016/j.cell.2013.08.041.

Pedersen, N. L. (2000) 'Geneti c and envi r onmental i nfl uences on body fat', (1999), pp. 43–50.

Philbrick, W. M. *et al.* (1990) 'A recombination event in the 5' flanking region of the Ly-6C gene correlates with impaired expression in the NOD, NZB and ST strains of mice.', *The EMBO journal*, 9(8), pp. 2485–2492. Available at: http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med3&NEWS=N&AN=2164472.

Rasmussen, A. L. *et al.* (2014) 'Host genetic diversityenables Ebola hemorrhagic fever pathogenesis and resistance', 346(6212), pp. 987–992.

Roederer, M. *et al.* (2015) 'The genetic architecture of the human immune system: A bioresource

for autoimmunity and disease pathogenesis', *Cell*. Elsevier Inc., 161(2), pp. 387–403. doi: 10.1016/j.cell.2015.02.046.

Rogala, A. R. *et al.* (2014) 'The collaborative cross as a resource for modeling human disease: CC011/Unc, a new mouse model for spontaneous colitis', *Mammalian Genome*, 25(3–4), pp. 95–108. doi: 10.1007/s00335-013-9499-2.

Di Rosa, F. and Pabst, R. (2005) 'The bone marrow: A nest for migratory memory T cells', *Trends in Immunology*, 26(7), pp. 360–366. doi: 10.1016/j.it.2005.04.011.

Snijders, A. M. *et al.* (2016) 'Influence of early life exposure, host genetics and diet on the mouse gut microbiome and metabolome', *Nature Microbiology*. Nature Publishing Group, 2(2), pp. 1–8. doi: 10.1038/nmicrobiol.2016.221.

Spitzer, M. H. *et al.* (2015) 'An interactive reference framework for modeling a dynamic immune system', *Science*, 349(6244). doi: 10.1126/science.1259425.

Srivastava, A. *et al.* (2017) 'Genomes of the mouse collaborative cross', *Genetics*, 206(2), pp. 537–556. doi: 10.1534/genetics.116.198838.

Tsang, J. S. *et al.* (2014) 'Global analyses of human immune variation reveal baseline predictors of postvaccination responses', *Cell*, 157(2), pp. 499–513. doi: 10.1016/j.cell.2014.03.031.

Wilson, A. and Trumpp, A. (2006) 'Bone-marrow haematopoietic-stem-cell niches', *Nature Reviews Immunology*, 6(2), pp. 93–106. doi: 10.1038/nri1779.

Yalcin, B. *et al.* (2011) 'Sequence-based characterization of structural variation in the mouse genome', *Nature*, 477(7364), pp. 326–329. doi: 10.1038/nature10432.

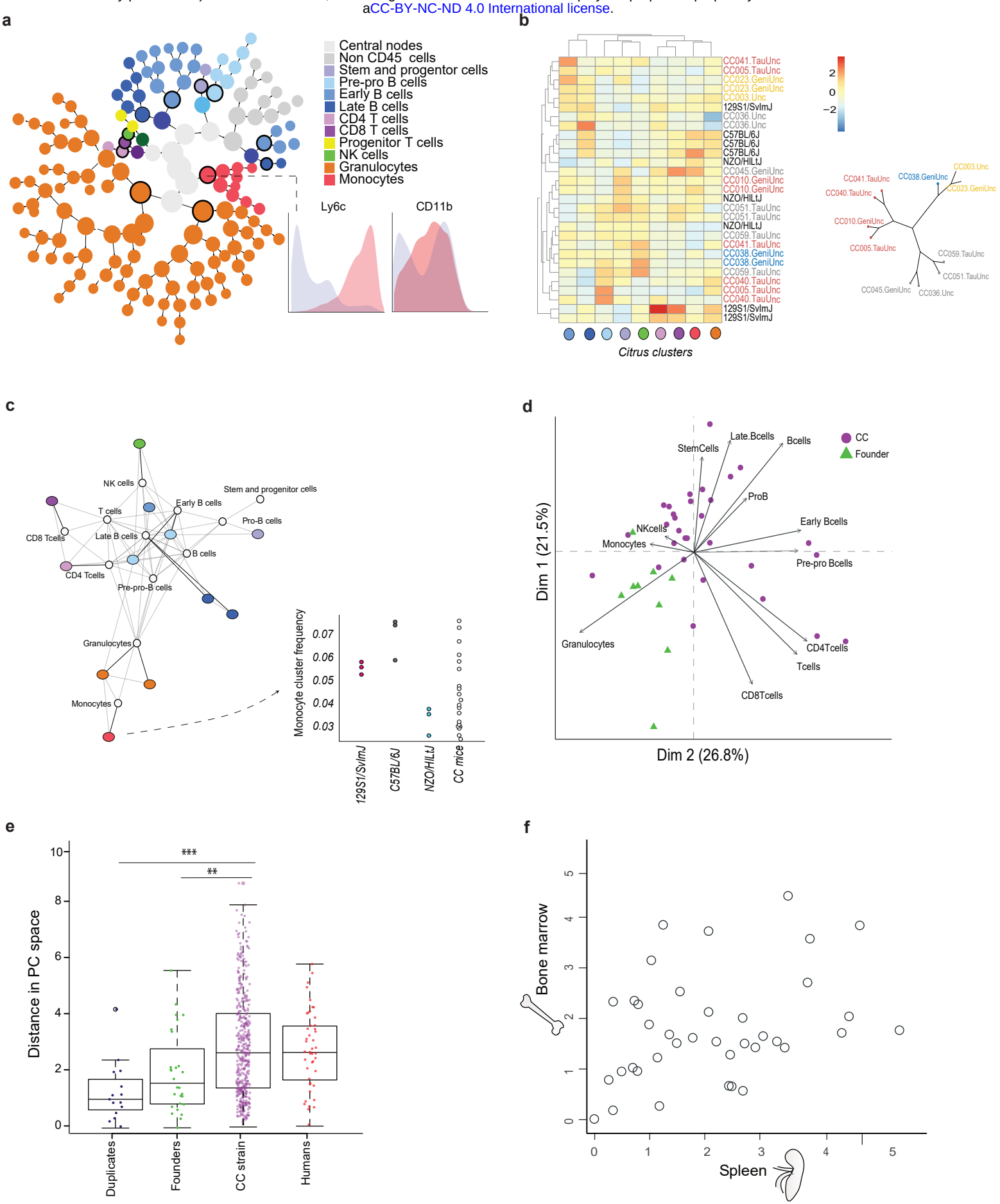
Acknowledgements

We would like to thank Fernando Pardo Manuel de Villena, Darla Miller and Ginger Shaw for providing us the CC mice strains and help with the mouse experiments. Asya Rolls, Tamar Ben-Shannan, Hilla Azulay-Debby, Ben Korin and Maya Schiller for help with mice work and manuscript revision. Amit Ziv-Kennet for help with data preprocessing and Fuad Iraqi for initial setup of the system.

Contributions

T.D., E.S., A.A, Y.O. performed the experiments; T.D, S.S. ,B.L, R.N., Y.A, M.G, performed the analysis; S.S, T.D., E.S, B.L, M.G, A.A, R.N. wrote the manuscript.

Figure 1



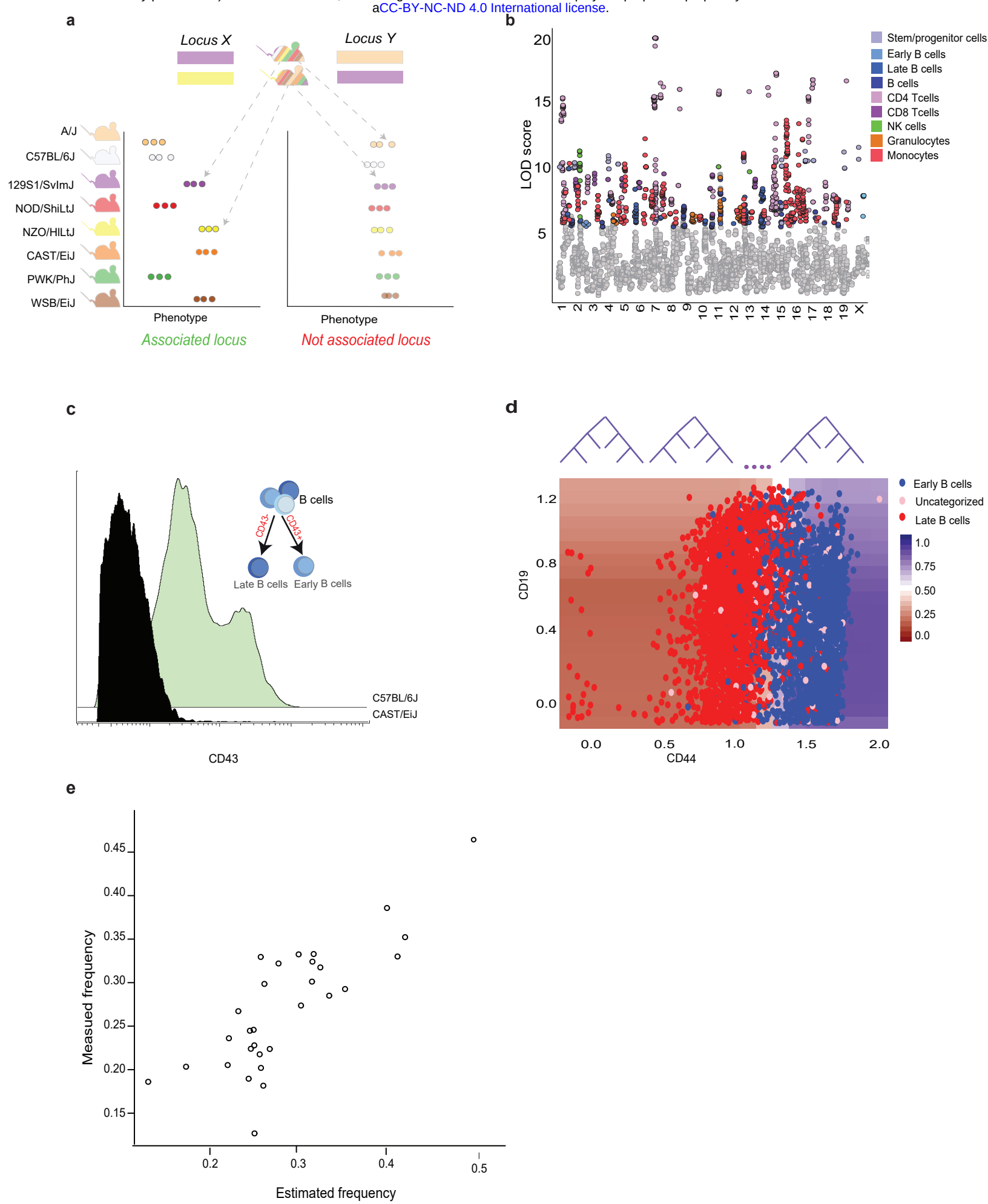


Figure 3

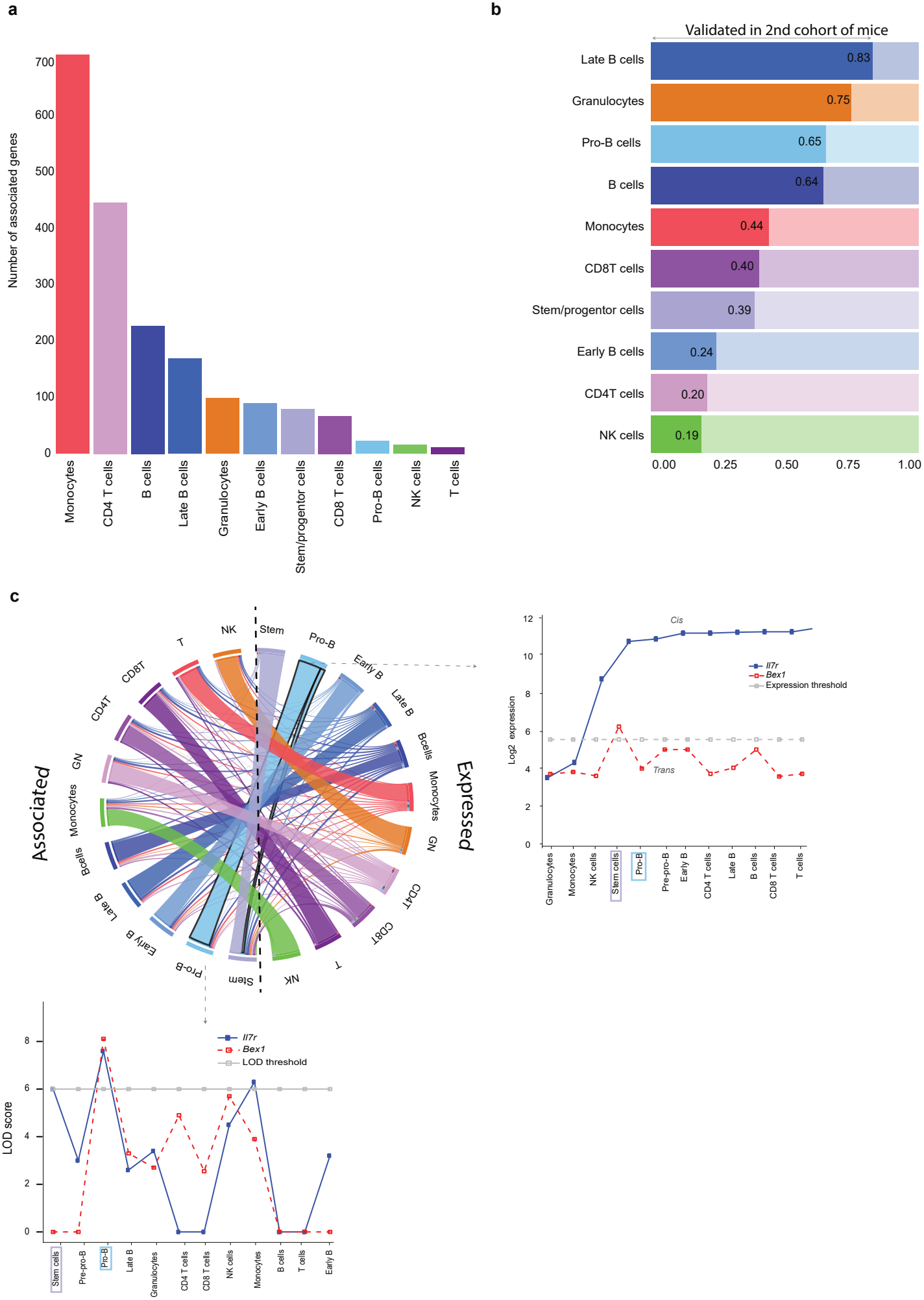


Figure 4

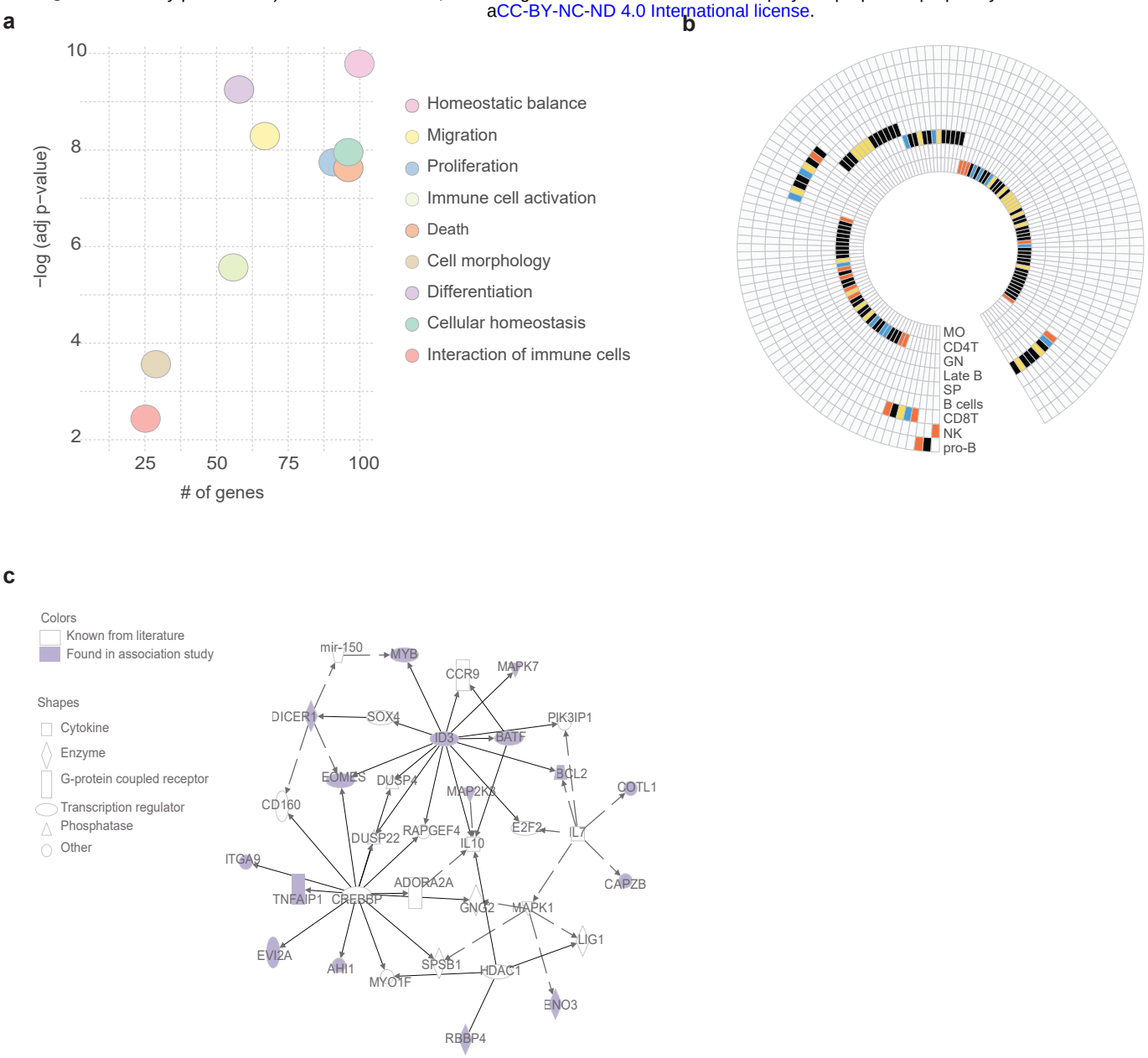


Figure 3

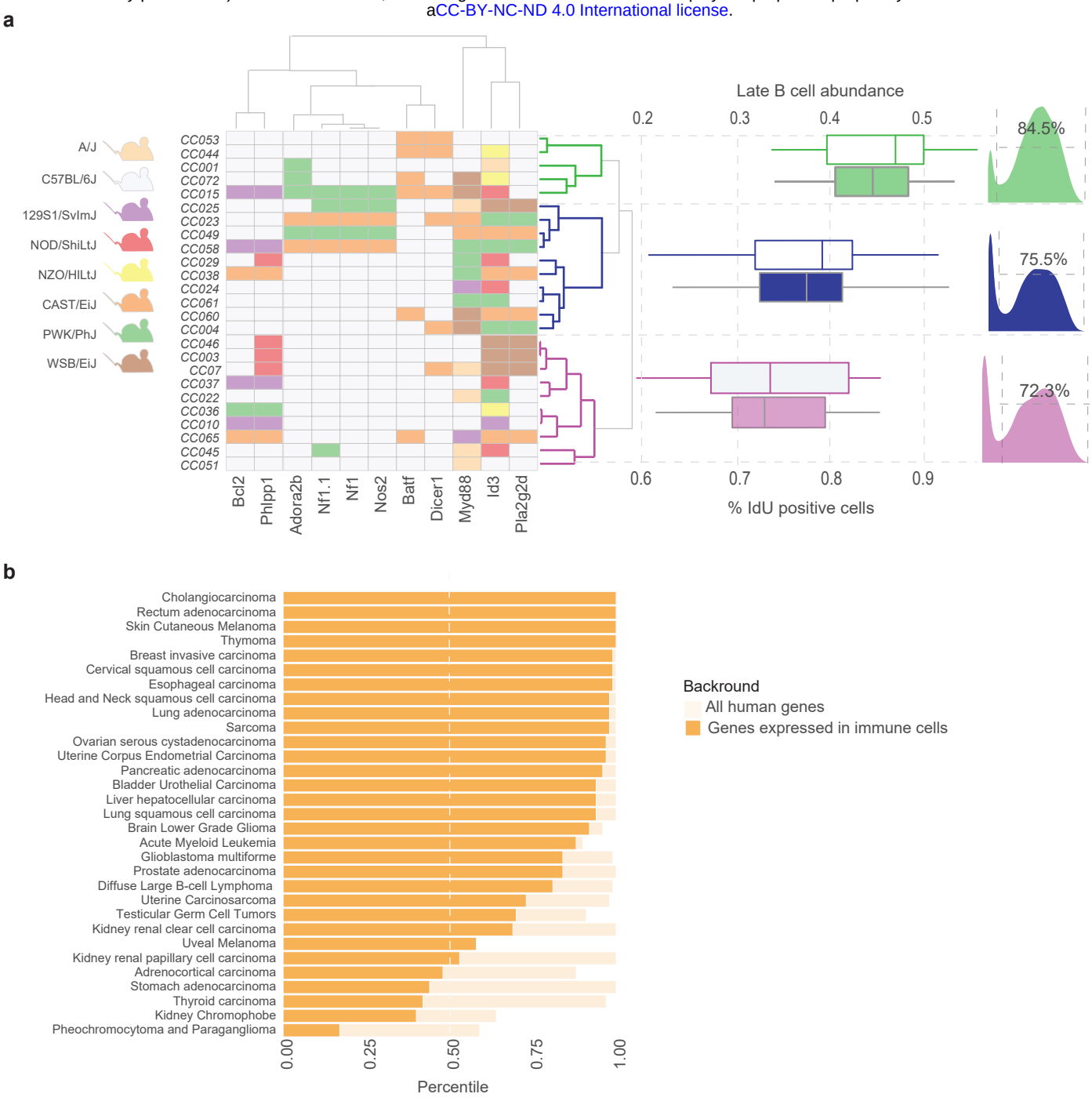


Figure 1. CC mice strains exhibit high variation in immune cell subset frequencies that is comparable to humans and conserved across immune organs. (a) Single cell data of bone marrow populations from CC mice were analyzed by Citrus, a high dimensional clustering analysis algorithm. The tree structure organizes clusters of single cells sharing similar set of 20 markers. Clusters are organized in a hierarchy with leaves of the tree in the periphery being more specific and those inward aggregating multiple leaves. Populations were color-labeled according to phenotypic markers distinctive for major bone marrow populations. Nodes in bold were used for further downstream analysis. Inset shows an example of markers that contributed to cluster characterization of monocyte cluster: Ly6c and CD11b (red and blue representing single cell distribution of marker expression in the cluster versus background respectively). **(b)** Phenotypic profiling of bone marrow populations follows genotypic similarity. Heatmap (left) of scaled Citrus cluster frequency (columns) for CC mice and founder strains (rows). Founders indicated in black and the CC strains are colored according to the phylogenetic tree branch. Phylogenetic tree (right) represents whole genome genetic distance between CC strains. **(c)** Scaffold mapping between manually gated populations and those obtained from Citrus clustering analysis allows to overcome heterogeneity in marker expression. Inset shows monocyte frequency distribution across non-missing CC and founder strains. CC mice exhibited a continuous range of frequencies whose dynamic range was larger than that observed between the most extreme founder strains. **(d)** PCA projection of immune cellular profiles obtained from manual gates of CC (purple) and founder (green) strains. Only strains with complete set of markers were included (Supp. table 3). The two principle components are defined primarily by lymphoid and myeloid lineage subsets respectively. **(e)** CC strains form a continuous phenotypic variation in immune homeostatic profiles whose dynamic range exceeds that of founder strains and is similar to that observed in humans. Boxplots showing pairwise analysis of distances in PC space (the first and second dimensions were used) within groups of replicate, founder or CC strains as well as ten human healthy bone marrow. Student's t-test p-value are: **p<0.01, ***p<0.001. Pairwise distances between replicate strains were only considered in the Duplicate group. **(f)** Individuality of homeostatic immune cellular profiles is conserved between mice across tissues. Distances of each CC strain from C57BL/6J in PC space in bone marrow and spleen tissues.

Figure 2. Immune phenotypes are complex traits that are regulated by multiple genomic loci. (a) Illustration of QTL mapping process using the Collaborative Cross mice: Genetically mosaic CC strains whose genome is reconstructed at each locus to one of eight founder strains are assayed for a quantitative trait, in our case bone marrow immune cell subsets frequencies. For each locus, and each CC strain, mice are split based on the founder strain contributed the locus (e.g. 'Locus X' for the first mouse is contributed by 129S1/SvImJ and for the second mouse by NZO/HlLtJ) and the association with the phenotype, across all founders tested. **(b)** Manhattan plot for all associated loci, the threshold of association is set according

to FDR of each population. Colors denote cell subset gene was mapped to. **(c)** CD43 histogram in two representative founder strains. Loss of staining is seen in CAST/EiJ. The loss causes an inability to distinguish between early and late B-cell subsets (inset). **(d)** Visualization of the two-dimensional joint distribution of CD19, a classical B cell marker and CD44, both of which were found to be highly correlated with CD43 expression in non-missing samples. Shown are multiple overlaid random forest trees. Color range shows the probability of a cell to be an early B-cell based on CD44 (x-axis) and CD19 (y-axis) arcsine transformed expression. **(e)** Scatterplot showing measured versus estimated frequency of early B-cells in non-missing samples. The accuracy was 82%. **(f)** Manhattan plot for all associated loci post-incorporation of estimated proportion for CD43 missing-samples. Loci associated with early B-cells are colored, based on whether they were identified prior to the addition of the estimated early B-cell strains (green) or after (yellow). Associations for other cell subsets are colored black.

Figure 3. Immune cell subset frequencies are regulated by genetic networks that are predominantly cell-specific. (a) An imbalance in genetic associations across immune cell subsets. Bar plot of number of per cell type associated genes, the bar is colored according to the population. **(b)** Validation of gene-trait associations in a second cohort of 24 non-overlapping CC mice strains ($p<0.05$ following BH adjustment). Non-transparent bars show the percent of validation success for each immune cell subset out of the total number of associated loci for that subset. **(c)** Homeostatic condition of immune cells is predominantly determined by *cis* genetic variants expressed in the same cell subset. Split Circos plot, depicting the fraction of genes expressed in each cell subset (right) based on ImmGen sorted cell gene expression patterns and the cells in which we detected an association of these gene's genetic variant with the cell's frequency (left). The color of the ribbon is set according to the cell subset the genes are associated with. The inset on the right side shows log2 expression of *Il7r* and *Bex1* gene. *Il7r* is a *cis*-gene which found to be associated with pro-B cell subset frequency and also expressed in it (blue line). *Bex1* is *trans*-gene, though found to be associated with pro-B cell subset frequency, *Bex1* expressed only in stem and progenitor cell population (red line). Expression threshold indicated by grey line. Bottom inset shows the strength of association for *Il7r* and *Bex1* gene across multiple cell subsets. *Il7r* associated with both pro-B cells and monocytes, while Bex1 associated with pro-B cell subset only. (Short abbreviations used for: T-T cells, NK-NK cells, GN-Granulocytes, Stem- Stem and progenitor cells)

Figure 4. Genes associated with immune cell subset frequencies are enriched for cell production and turnover functions. (a) Bubble plot showing functional enrichment categories. X-axis shows the number of genes in a functional annotation group, y-axis, -log adjusted p-value on enrichment. Repeating related functional terms groups were grouped together (see Supp. Table 14 for full term enrichment listing), and the group with highest adjusted p-value and total number of genes for each functional

enrichment is shown. Functionally enriched genes were annotated for at least one of four main processes: proliferation, cell death, differentiation and cellular movement. **(b)** Cell subset frequencies are regulated by epistatic interactions between proliferation, migration, differentiation and death genes whose genetic variants are associated with the outcome. For every cell, genes associated with each of these four functions are found associated in a cell-specific manner. The subset identity is shown as : MO-monocytes, CD4T- CD4 T cells, GN-Granulocytes, Late B-cells, SP- Stem and progenitor cells, B cells, CD8T- CD8 T cells, NK- NK cells, pro-B cell. The colors of the enriched functions: blue- proliferation, orange- death, yellow migration, black-multi functional gene (all differentiation genes are multi-functional). **(c)** Cell specific network of regulation for late B-cell subset. Gene shapes are annotated based on IPA gene ontology, genes that were found as associated to late B-cell frequency are colored in blue.

Figure 5: Associated genetic variants predict abundance and proliferation rate and are more likely to be mutated across a broad range of cancers.

(a) Validation of the proliferation function. From the left to the right: (i) Clustered genetic profile of genes that are associated with late B-cell subset frequency for each mouse. All possible alleles that differ from the C57bl/6 mouse allele for each gene were extracted, and the resulting profile across all associated genes was clustered by the hamming distance. (ii) For each clustered group, two boxplots are shown: (1) boxplot with colored contour for the abundance and (2) with a color fill for the percent of proliferating cells (top and bottom axes respectively) (iii) Histogram of IdU staining for a representative samples for each clustered group. **(b)** Genes associated with immune cell subset frequency are at the top percentile of mutated genes in cancers, compared to equally sized random controls. Percentile is shown for cancer type, the percentile calculated after removing non-immune genes out of genes expressed in immune cell only (orange), all human genes (light orange).