Meta-analysis of scRNA seq data to define the landscape of human liver immune homeostasis.

Brittany Rocque, MD, MSc[1,2], Arianna Barbetta, MD[1], Pranay Singh[1], Cameron Goldbeck, MS[1], Doumet Georges Helou, PhD[3], Yong-Hwee Eddie Loh, PhD[4], Nolan Ung, PhD[5], Jerry Lee, PhD[5,6], Omid Akbari, PhD[3], and Juliet Emamaullee, MD, PhD[1,3*]

Affiliations:
[1] Division of Abdominal Organ Transplantation and Hepatobiliary Surgery, Department of Surgery, University of Southern California, Los Angeles, CA
[2] Broad Institute, University of Southern California, Los Angeles, CA
[3] Department of Molecular Microbiology and Immunology, University of Southern California, Los Angeles, CA
[4] Norris Medical Library, University of Southern California, Los Angeles, CA
[5] Ellison Institute, University of Southern California, Los Angeles, CA
[6] Department of Medicine and Department of Chemical Engineering and Material Sciences, University of Southern California, Los Angeles, CA

***Corresponding Author:**
Juliet Emamaullee MD PhD FRCSC FACS
1510 San Pablo Street, Suite 412
Los Angeles, CA 90033
Tel: 323.442.5908
Fax: 323.442.6887
Email: Juliet.emamaullee@med.usc.edu

ABSTRACT

Single-cell RNA sequencing (scRNA seq) generates highly dimensional transcriptome data to understand cellular phenotypes. Due to high cost and volume of data obtained from these experiments, most studies analyze few unique samples. Pre-analytical tissue processing and different sequencing techniques can alter quantity and phenotype of cells being examined. This study used existing 'human liver' scRNA seq datasets to determine if heterogeneity exists between datasets and if these data could be integrated to define phenotypes across leukocyte subpopulations in healthy human liver. CD45+ cells from three published studies reporting scRNA seq data from human livers were analyzed and leukocyte subpopulations were examined. All three datasets co-clustered, but with differing cell proportions. Expression correlation demonstrated similarity across all studies. Detailed analysis of differential expression identified meta-signatures for each hepatic immune subpopulation. Herein, we introduce a novel and rigorous framework for meta-analysis of scRNA seq datasets and highlight profiles of liver immune homeostasis.

INTRODUCTION

Single cell RNA sequencing (scRNA seq) has led to rapid advances in our understanding of cellular phenotypes of *in situ* human tissues in recent years.[1,2] Application of scRNA seq to human liver tissue to examine both healthy and pathogenic states at the molecular level has proven useful to identify heterogeneity among various cell types, including immune cell populations and epithelial progenitors.[3–5] Generation of these libraries through high-throughput sequencing techniques such as 10x Genomics or Drop-seq generates highly dimensional data, which is suitable to both answer and also to rapidly generate hypotheses. Given the relatively high cost of scRNA seq, a typical human study only analyzes a small number of unique samples (as few as three unique patients), despite the known genetic heterogeneity across individuals.[6,7] In 2018, the NIH updated the policy for data access of genomic datasets, with the goal of enhancing researchers' ability to perform pooled analyses or meta-analyses on genomics data particularly in human specimens.[8] More recently, there are several NIH Requests for applications specifically targeted at secondary and integrated analytic approaches to harness the power of these existing, funded datasets.[9]

The liver is an immunologically complex organ with an abundance of resident leukocytes which comprise a significant proportion of residing nonparenchymal cells.[10,11] A unique but poorly understood feature of the liver is its immunotolerance, which is thought to be the result of the organ receiving much of its dual blood supply from the portal vein. The portal vein shuttles blood to the liver via the enterohepatic circulation, which carries both nutritionally-derived and bacterial antigens without causing an untoward inflammatory response.[12] This immunotolerance has implications for a vast array of disease processes of the liver including autoimmune hepatitis, cholangiopathy, transplantation and the tumor microenvironment of intrahepatic malignancy, among others.[13]

Combining liver-specific immune cell subsets of scRNA seq datasets yields a higher number of cell libraries across a larger sample of patients, which is crucial when considering that

knowledge generated from these human “atlases” may be used to understand biochemical mechanisms of disease and to develop therapeutics for clinical use. One major drawback of combining unique studies is that differences in technique have been demonstrated to lead to unexpected alterations in sequencing results.[14,15] A study by Bonnycastle et al used scRNA seq in human pancreatic islet cells and compared different tissue processing techniques (including fresh, fixed and cryopreserved tissues). Despite processing tissue samples from the same source, this analysis demonstrated differences among cell type proportions recovered as well as changes in gene expression signatures across different processing techniques.[16] Thus, combining multiple human liver scRNA datasets has the potential to account for more biological variability across different patients, potentially attenuating the effects of pre-analytical variables that are not biologically meaningful.

The aim of this study was to perform a meta-analysis of integrated datasets of normal human liver scRNA seq specimens from high-quality peer-reviewed studies. An extensive comparison was performed across studies in order to establish whether generalizability across datasets would be meaningful. Finally, we characterize expression profiles from the pooled datasets across four major immune subpopulations in order to provide a description of immune homeostasis in the healthy human liver.

RESULTS

Integration of liver immune cell scRNA seq datasets

To begin, immune-specific subsets of human liver tissue were identified and extracted from three unique studies: "LACE^{e}", "L^{nb}" and "LCD45^{e}" (**Figure 1**).[3–5] Techniques used for the RNA sequencing of individual datasets are summarized in **Table 1**. The combined dataset included single-cell RNA sequencing of 17 normal human liver samples with approximately 32,000 hepatic CD45+ cells.

Clustering analysis was performed on all three datasets individually to detect differences in global cellular phenotypes (**Figure 2a-c**). Projection of all three datasets to the same UMAP coordinates revealed a homogeneous interdigitation of each cluster, however the LCD45^{e} subset comprised the majority of cells (~24,000 cells compared to ~4,000 cells in each of the other two studies; **Figure 2d**). Pairwise gene expression correlation analysis was performed which revealed that the LACE^{e} dataset had less concordance with both the L^{nb} (R=0.81; **Figure 2e**), and LCD45^{e} datasets (R=0.79; **Figure 2f**). L^{nb} and LCD45^{e}, both obtained using the 10x platform, demonstrated a more idealized relationship (R=0.95; **Figure 2g**).

Frequency and proportion of immune cell subpopulations across liver scRNA seq datasets

Next, leukocyte subpopulations were identified based on expression of major cell-specific lineage marker genes and compared across the individual datasets (**Supplemental Figure 1**). Clusters were then designated as NK and T cells (combined), myeloid cells, B cells and plasma cells, representing the four major groups of interest for our study (**Figure 3a-b**).

When compared across all three datasets there were differences in proportions of each immune cell subpopulation (X^2 test, p<0.01 for each cell type; **Figure 3d**). Pairwise analysis between datasets also revealed differences in cell-type composition, and only two conditions did not reach statistical significance: comparison of LCD45^{e} and L^{nb} in plasma cells (X^2 test, p=0.16) and comparison of L^{nb} and LACE^{e} in B cells (X^2 test, p=0.08), which may be a consequence of B

cells and plasma cells representing the smallest populations of cells across datasets. This observed scarcity of liver resident B cells and plasma cells is consistent previous reports of liver immune cell composition.[11] Despite some differences in cell-type recovery, the ranking of abundance of each cell type was preserved between datasets, such that NK and T cells were the most numerous (ranging from 51%-69% of cells), followed by myeloid cells (18%-32%), plasma cells (4%-14%), and then B cells (2%-7%; **Figure 3c**).

Phenotypic analysis of leukocyte subpopulations across datasets

A deeper analysis of gene expression profiles within immune cell subpopulation was conducted to identify differences in cellular phenotypes. Initially, analysis of housekeeping genes was conducted, as housekeeping gene expression should provide an internal control of biological noise and thus serve as a marker of technical noise across studies.[17] Expression analysis was performed on an established set of human housekeeping genes across immune cell subpopulations and across datasets using log normalized expression values (**Supplemental Figure 2**). Despite a relatively good agreement between means, comparison with one-way ANOVA yielded significant differences in housekeeping gene expression ($p<0.01$). Pairwise analysis also showed significant differences in expression levels, which is likely the result of the large number of cells included and is thus a highly powered test.

Differences in gene expression profiles across immune cell subpopulations are anticipated due to their distinct biological functions. The number of differentially expressed (DE) genes was quantified between individual immune cell subpopulations (**Supplemental Figure 3**). Volcano plots were used to illustrate meaningfully DE genes, which were plotted as blue dots and indicate a fold-change ratio in expression of either <0.8 or >1.25 and Bonferroni corrected p-value<0.05. Non-DE genes were plotted as red dots. B cells had the fewest DE genes across all studies, whereas cells of the myeloid lineage had the most DE genes. No individual dataset was an outlier with respect to the quantification of differential expression, but the LCD45^{e} dataset had the highest

number of DE genes in B cells and myeloid cells, and the lowest number in plasma cells and NK and T cells.

Due to heterogeneous gene expression between leukocyte subpopulations (**Supplemental Figure 3**) and differences in cell proportions (**Figure 3c-d**), both of which could potentially bias dataset correlation analysis, pairwise expression correlation between datasets was performed and stratified by leukocyte subpopulation (**Figure 4a-d and Supplemental Figure 4**). Comparison of L^{nb} and $LCD45^{e}$ showed a more idealized relationship (*R*=0.93-0.96 depending on leukocyte subpopulation, **Figure 4a.vi.-d.vi**), while $LACE^{e}$ correlated less with the other two studies (**Figure 4a.iv,v-d.iv,v**). The pattern of gene expression correlation was also analyzed using Rank-Rank Hypergeometric Overlap (RRHO) heatmaps (**Figure 4a-d, panels i-iii**).[18] This technique uses heatmaps to indicate the degree of ranked differential expression agreement, where yellow indicates strong statistical overlap and blue represents no statistical overlap between corresponding axes. These panels represent overlap of differential expression in ranked form such that the bottom left of the panel represents the most upregulated genes, and genes in the upper right portion of the panel represent the downregulated genes (thus a higher rank corresponds to downregulation). When comparing NK and T cell expression patterns in $LACE^{e}$ versus L^{nb}, there are two heatmap regions with a high degree of statistical overlap (**Figure 4a**). Examination of the remaining two pairwise comparisons in NK and T cells from different datasets, the overlap among the more highly expressed genes is evident, but genes at lower expression levels have less overlap. Plasma cells and B cells also exhibit RRHO agreement between studies. For myeloid cells, $LACE^{e}$ and L^{nb} have better RRHO for the most highly expressed genes, whereas L^{nb} and $LCD45^{e}$ have better agreement in genes expressed at lower levels. There is a bimodal agreement between $LACE^{e}$ and $LCD45^{e}$ (**Figure 4b**).

Differential gene expression in leukocyte subpopulations across datasets

The RRHO analysis demonstrated that there was relatively strong correlation with genes at the highest levels of expression. To begin to identify the dominant phenotype for each leukocyte subpopulation, the top 100 most abundant transcripts were identified for each cell type in L^{nb} which was the dataset with the most agreement with the other two datasets based on correlation (**Figure 2e-g**). A comparison was performed to determine if the same genes were present as the top 100 most abundant transcripts for each immune cell subpopulation of the other two studies. Depending on leukocyte subpopulation, $LCD45^{e}$ had between 87 and 94 out of 100 'top genes' in agreement. Alternately, $LACE^{e}$ had rather poor agreement with top matching genes for each leukocyte subpopulation (**Supplemental Figure 5**).

Analysis was then performed across datasets in order to quantify how many genes were DE in each immune cell subpopulation (**Figure 5b**). In this analysis, DE genes with a fold change ratio <0.8 or >1.25 and a Bonferroni corrected p-value <0.05 are represented in volcano plots and compare individual gene expression level within an immune cell subpopulation, across datasets.[16] Some comparisons yielded a high number of DE genes between datasets (e.g., 3526 genes when comparing $LCD45^{e}$ and $LACE^{e}$ in myeloid cells). In contrast, when comparing the L^{nb} dataset to the $LCD45^{e}$ dataset among the B cell population, there were only 767 DE genes and the myeloid cell population between these two datasets had 924 DE genes, again demonstrating fewer expression differences between L^{nb} and $LCD45^{e}$ relative to $LACE^{e}$.

To better characterize the heterogeneity in gene expression, the difference of differences was calculated between datasets (pairwise) and stratified by leukocyte subpopulation (particular versus all other). This metric involved calculating a gene's expression difference between the cell subpopulation of interest and all other cell types and then comparing that difference to the difference seen in another dataset (**Figure 5a**). For example, a B cell in the $LACE^{e}$ dataset has a difference of expression of the gene CD25 compared to all other cell types (represented as $\Delta 1$), the difference in CD25 expression between B cells and other cell types in L^{nb} is $\Delta 2$. Subtracting

Δ2 from Δ1 yields the difference of differences (DoD, **Figure 5**) which we used to approximate the proportion of DE genes where expression alteration could be explained by changes relative to other immune cell subpopulations (DoD significance defined as: DE with a fold change ratio <0.8 or >1.25 with $P\text{-adj}_{Bonfcorr}<0.05$ and DoD $P\text{-adj}_{Bonfcorr}<0.05$). We also anticipate that due to comparison across different studies and across differing technical platforms (10x Genomics versus mCel-seq) that there will be differences in the depth of sequencing, a phenomenon which has been known to affect differential expression analysis when comparing scRNA seq datasets.[19] Given this, we propose that our DoD analysis "explains", or at least accounts for, some of the differential expression. Figure 5c lists the number of DE and DoD genes for each analysis as well as a percent total. The last column in the table shows the proportion of DoD genes relative to the number of DE genes. Notably in plasma cells, there are a high number of DE genes, but over 50% of these are also DoD genes. This suggests that approximately half of the observed differences of plasma cells between L^{nb} and $LCD45^{e}$ were accounted for by comparing to the average expression level in the remaining cells of each dataset.

Concordance of dominant gene expression profiles for immune cell subpopulations across datasets

Once DE genes were identified, the most significantly DE genes were chosen among each dataset to determine if an integrated dataset could identify dominant gene expression signatures associated with each immune cell subpopulation within human liver. The top 10 most DE genes for each leukocyte subpopulation in each individual dataset were identified. All three datasets' candidate genes were combined to create a master list of top DE genes for each immune cell subpopulation (**Figure 6d**). Expression levels were plotted as a heatmap for each individual cell and organized by immune cell subpopulation (**Figure 6a-c**). Some heterogeneity was expected within cell type, as the myeloid and NK and T cell populations represent a variety of unique subsets. Despite this, the heatmaps demonstrate that candidate genes listed are indeed DE and

represent reproducible markers for each immune cell subpopulation, regardless of processing technique. Furthermore, the expression patterns outlined in the heatmaps have excellent agreement across studies.

Differential gene expression profiles between datasets reveal signaling pathways affected by scRNA seq pre-analytical tissue processing and techniques

After identification of all the genes that had a statistically significant differential expression (log fold change <0.8 or >1.25 and adjusted $p<0.05$ as shown in volcano plots from Figure 5b) among the three datasets, a canonical pathway analysis was performed using the Ingenuity Pathway Analysis software (IPA, Qiagen). The association study was performed for each leukocyte subpopulation: NK & T cells, B cell, plasma cell and myeloid cells. Input of the DE gene name, average log fold change and *p* values provided an output of canonical pathways which were ranked on their likelihood to be altered between pairwise datasets. The probability that a signaling pathway was affected by differences across datasets was represented as a -log(*P* value) and a cutoff of 7.0 was set in order to establish those canonical pathways which were most different. Disease-specific pathways of no relevance to immune-specific subpopulations were excluded (e.g., Atherosclerosis Signaling, Bladder Cancer Signaling) as they were extraneous to this analysis.

Across immune cell subpopulations, there was general agreement between the canonical pathways affected based on the pairwise dataset comparison and therefore NK & T cell subsets and B cell subsets are shown as examples (**Figure 7a-c** and **Figure 7d-f** respectively) as myeloid and plasma cells demonstrated the same pathways which were affected, albeit to varying degrees.

The analysis of DE genes in NK and T cells for LACE^{e} vs L^{nb} pair identified three pathways: EIF2 signaling (-log(*P* value)=14.7, 20 DE genes involved), oxidative phosphorylation (-log(*P* value)=8.8, 11 DE genes involved), and regulation of eIF4 and 70S6K signaling (-log(*P* value)=7, 11 DE genes involved; **Figure 7a**). B cell DE genes between LACEe vs Lnb, showed a similar result with the addition of mTOR signaling. Of note, many of the genes that led to the presumptive alterations of these pathways were ribosomal proteins with the exception of the oxidative phosphorylation canonical pathway, where genes affected were NADH dehydrogenase subunits, cytochrome c oxidase subunits and ATP synthase subunit F0 subunit 6.

The LCD45^{e} versus LACE^{e} pairwise comparison of canonical pathways was reminiscent of the altered pathways seen in the LACE^{e} versus L^{nb} differential comparison. The expression analysis between L^{nb} and LCD45^{e}, which was the pairwise comparison with the best transcriptome concordance and the fewest DE genes, actually had the highest number of altered canonical pathways (**Figure 7c,f**). Even though many different canonical pathways were identified as having been altered between these two datasets (L^{nb} vs LCD45^{e}) many of the DE genes that made up the pathway differences were the same in each pathway. For instance, ALB or the protein albumin was counted as being a DE molecule in six out of the eight canonical signaling pathways. Likewise, genes such as APOA1, APOA2, ORM1, ORM2 and SERPINA1 contributed to the majority of canonical pathway alterations. In fact, most of the molecules identified were redundant downstream effectors of the pathways and thus do not necessarily reflect true perturbations of the signaling pathway between datasets.

In characterizing the canonical pathways that differed across studies, the major differences were noted among pathways which involved protein synthesis, cell death and apoptosis signaling. Signaling pathways related to immunologic functions, which are the functions that are of relevance to this study were largely preserved. For example, the T cell receptor signaling pathway, CD28 signaling in T helper cells and IL-6 signaling pathways all had 2 or fewer

genes which contributed to differences noted which supports the idea that these pathways were largely comparable between studies and that proceeding with a combined meta-dataset is feasible.

Phenotypic characterization of immune cellular landscape in normal human liver

Given that gene expression correlation was high across datasets and differential gene expression was relatively low, with conservation of phenotypically relevant genes and signaling pathways, we established meta-signatures for each immune cell subpopulation. Genes which were DE from one leukocyte subpopulation versus cells of all other subpopulations, but not DE across studies were identified. This is in contrast with the prior analysis, which used genes DE across datasets. These candidate genes comprising a list of 45 to 146 total genes depending on leukocyte subpopulation were generated and run through the Ingenuity Pathway Analysis Software (Qiagen). The output of this software identified a series of genes with linked cellular functions to determine which canonical pathways were affected.

Pathway analysis output generated characterizations of various cellular biochemical functions. Functions were sub-divided into types including general inflammatory response, cell signaling and interaction, cell death and survival, cell function and maintenance, cell growth and proliferation and cellular movement. Z-score values of whether expression was upregulated or downregulated were used to formulate a heatmap for each function across individual leukocyte subpopulations, establishing a functionality signature for each cell type (**Figure 8**). Most functions were downregulated across all leukocyte subpopulations. There was very mild upregulation of cell functions including apoptosis and colony formation in myeloid cells, development in NK and T cells, and proliferation in B cells and NK and T cells. Cell death of cancer cells and membrane potential of mitochondria were the only functions that were moderately upregulated in B cells. Overall, these results indicate a general low-activity state of liver immune cells.

These pathway analysis data were then transferred for further tertiary analysis in R using the Circlize package to produce meta-signatures as chord plots which relate specific cell functions to individual genes having differential expression. In order to reduce complexity of the figures, redundant cell functions and functions related to disease states noted to be irrelevant to immune subpopulations were excluded. The resulting chord diagrams link genes associated with specific cellular functions, allowing for visualization of the immune phenotype for each leukocyte subpopulation.

NK and T cells represented the largest proportion of cells in our study, but only had the second largest number of DE genes identified by our analysis, with plasma cells having the most genes. For this immune cell subpopulation, 68 genes were identified as being DE, representing 14 major immunologically relevant diseases and functions (**Figure 9a**). Of the DE genes in NK and T cells, only the GTPase GIMAP7 (average log fold change (log_FC)=1.1), perforin 1 (PRF1, log_FC=1.7) and SH2 domain 1A molecule implicated in NK signaling, SH2D1A (log_FC=0.9) were upregulated. The majority of genes were downregulated, which correlates with data in Figure 8. The myeloid lineage was marked by a gene expression signature involving 37 genes across the relevant diseases and functions (**Figure 9b**). Upregulation of the following DE genes was shown: brain protein I3 (BRI3, log_FC=1.0), cystatin c (CST3, log_FC=2.9), the autophagy associated protein cathepsin D (CTSD, log_FC=1.2), dual specificity phosphatase 1 (DUSP1, log_FC=0.4) among others. Again, the majority of genes were downregulated in the myeloid cells, with the most notably downregulated DE genes being CD2 (log_FC= -1.3), CD3E (log_FC= -1.3), CD69 (log_FC= -1.5) and granzyme K (GZMK, log_FC= -1.4; all p values < 0.001).

Plasma cells with 120 DE genes shown in the chord diagram demonstrated upregulation of endoplasmic reticulum lectin 1 (ERLEC1, log_FC=1.0), and immunoglobulins: IGLC7 (log_FC=2.3), IGLV1-40 (log_FC=0.6) and IGLV6-57 (log_FC=1.2; all p values < 0.001). Downregulated genes included cell-cycle regulator BTG anti-proliferation factor 1 (BTG1,

log_FC= -1.7), complement proteins C1QA (log_FC= -1.6) and C1QB (log_FC= -1.6), CST3 (log_FC= -1.8), ID2 (log_FC= -1.6), interleukin 7 receptor (IL7R, log_FC= -1.6) and killer cell lectin like receptor D1 (KLRD1, log_FC= -1.7; **Figure 9c**). Lastly, the B cell meta-signature was comprised of 52 DE genes (**Figure 9d**). Upregulation was seen in Rho GTPase activating protein 24 (ARHGAP24, log_FC=0.8) which is implicated in actin cytoskeleton signaling, baculoviral IAP repeat containing 3 (BIRC3, log_FC=1.3), CD19 (log_FC=0.8), CD22 (log_FC=1.1), CD24 (log_FC=1.1), CD37 (log_FC=1.4), Fc receptor like A (FCRL2, log_FC=0.7), FCRL5 (log_FC=0.5), FCRLA (log_FC=0.9), HLA-DQB1 (log_FC=1.2), TNF receptor superfamily member 13C (TNFRSF13C, log_FC=1.2, all *p* values < 0.001). Genes which were downregulated (log_FC< -1.5) included: complement proteins C1QA (log_FC= -1.5) and C1QB (log_FC= -1.5), C-C motif chemokine ligand 4 (CCL4, log_FC= -1.8) and cystatin c (CST3, log_FC= -1.7), all *p* values < 0.001). A detailed report of genes implicated in leukocyte subpopulation homeostasis is summarized in **Supplemental Table 1**.

DISCUSSION

In this study, meta-analysis of three high-quality scRNA seq studies representing 17 distinct 'normal human liver' samples enabled deep characterization liver immune cell function and features related to immune homeostasis. Focused examination of immune cell types, proportions, and gene expression profiles of revealed that significant differences exist between datasets created from 'normal human liver' samples. Despite these differences, the overall ranking of abundance of immune subsets was preserved, with NK and T cells dominating and plasma and B cells being relatively rare. Further analysis of gene expression profiles involving the most DE genes in the integrated dataset provided the opportunity to minimize the effect of technical 'noise' and establish the gene expression signature of human liver immune homeostasis. These results have the potential to serve as a key reference point for future studies using scRNA seq designed to characterize immune-based pathologies within the liver, such as autoimmune disease, the tumor microenvironment in primary liver malignancies, or liver allograft rejection and tolerance.

To establish validity of performing a meta-analysis of an integrated dataset using scRNA seq data generated in three distinct labs with different tissue handling, processing, and sequencing techniques, visualization of the integrated data with co-clustering of cell populations was performed (**Figure 2**). This approach was based on the work of Bonnycastle et al, who specifically studied the impact of tissue processing and handling on the integrity of scRNASeq data.[16] In our analysis, immune cell proportions and expression profiles differed somewhat across datasets, with highest correlation between the L^{nb} and $LCD45^{e}$ studies (**Figure 4**). This is not particularly surprising, as these two datasets used the same 10x Genomics technique for the scRNA seq. It is important to note that extraction of the CD45+ population and combination with other non-liver derived CD45+ cells did not appear to impact the integrity of the data, as was the case with the $LCD45^{e}$. While various scRNA seq techniques have the potential to introduce noise and bias into the transcriptome data, the Seurat pipeline has been designed to correct for some

of these differences by re-aligning cell clusters based on nearest neighbor correction and allowing for integrated downstream analyses.[20,21] This would suggest that differences we observed in expression analyses could be the result of pre-analytical variables such as tissue processing. We further characterize noise with the quantification of an established set of human housekeeping genes.[17] There were differences noted in expression of human housekeeping genes within an immune cell subpopulation across studies, however, mean expression values were not substantially altered and differences noted here are likely the result of large sample sizes. A potential explanation for differences in cellular proportions may be related to sample size effect as well. The $LACE^{e}$ dataset encompasses the largest sample of human tissues, derived from nine total liver specimens, which likely introduces a higher degree of biological variation than L^{nb} and $LCD45^{e}$ which have five and three liver donors, respectively. The clinical scenario also differs between $LACE^{e}$, which was created using partial hepatectomy specimens from patients with liver malignancy, and the other two studies, which involve specimens derived from adult brain-dead organ donors. Despite attempts at correction, there were differences noted in gene expression analysis, yet immune subpopulations still co-clustered with good interdigitation across studies and DE analysis showed good agreement with heatmaps (**Figures 2 and 8**). As such, we proceeded with deeper immune profiling using the integrated dataset.

Differences in tissue handling and processing has been shown to prompt transcriptional changes within the cell.[15,16,22] A major novelty of our study is the in-depth characterization of differential expression between datasets with identification of relevant canonical pathways affected. Our results indicate that pathways involved in oxidative phosphorylation and cell stress were primarily affected and that immunologically relevant pathways were largely preserved. We suspect that non-immune function related changes may be the consequence of pre-analytical tissue processing. Known pre-analytical variables related to each technique used for the generation of each dataset are summarized in **Table 1**. Unknown variables of potential importance could include the time interval of 'Pringle' clamp time during resection or from the

moment of tissue resection to cell processing and sequencing, which could lead to significant ischemic insults and thus impact transcription profiles. In ischemic injury of rodent and human cortex, insult has been demonstrated to promote pro-inflammatory gene expression alterations.[23,24] The fact that these ischemia-induced pro-inflammatory changes are not demonstrated in this study shows that either 'Pringle' time in liver specimens captured in these datasets were either not significant enough to cause ischemia, or that liver ischemia does not produce as profound an inflammatory response as cortical ischemia. Overall, the observation that immune pathway profiles for each subpopulation were relatively conserved supports the use of this novel meta-analysis approach to examine existing datasets in order to understand immune homeostasis in human livers.

Immunotolerance is a unique feature of hepatic homeostasis and is thought to be a consequence of the constant stream of antigens being delivered to the liver from the gut. This involves regulation of both the adaptive and innate immune systems to prevent an uncontrolled inflammatory cascade in response to foreign antigens in the portal circulation which gain access to the liver.[10,25] One potential theory is that effector T cells are regularly destroyed in the liver. Our profile analysis did not show increased genes involved in apoptosis among NK and T cells and "cell death" as a gene function category was also downregulated. Notably however, this phenomenon would be difficult to capture with scRNA sequencing due to exclusion of dead (but perhaps not dying) cells from liver tissue suspensions as producing individual libraries is expensive and sequencing is ideally performed only on living cells. There was a slight increase in apoptosis genes noted among myeloid cells, but this trend did not extend to any other cell type (**Figure 8**). Another hypothesis surrounding immune tolerance of the liver is a milieu of anti-inflammatory cytokines and inhibitory regulators. In our analysis, the overall cellular functions in the immune system tended to be downregulated indicating that liver-specific immune cells are potentially in a state of globally decreased activity. This provides a potential explanation of the environment of liver immune cell tolerance at the single cell level.

There are limitations to this study. Further differences in scRNA seq technique (comparing 10x Genomics to mCel-seq) likely introduces a differing amount of bias into the sequencing datasets. In addition, because this study explored pre-existing datasets, there was no ability to control for patient-specific parameters in this study design such as inclusion/exclusion based on liver function studies or other potentially relevant clinical factors including duration and storage of samples prior to processing. Despite this, combining scRNA seq and other genomics datasets for the characterization of normal physiology and disease processes will become a more prevalent practice and further efforts should be made to standardize the processes of pre-analytical processing variables as well as to enhance the integration abilities during data analysis.

In conclusion, our results present an integrated dataset of scRNA seq results of the liver immune environment from ~32,000 cells across 17 human livers making it the largest human liver meta-dataset thus far. We generated immune subset expression profiles that describe the landscape of liver immune function and the phenotype of liver immune homeostasis. These results can be incorporated into future RNA sequencing studies and has implications for understanding mechanism of disease and identifying new therapeutic targets in immune-related diseases of the liver.

METHODS

Institutional Review Board approval was not required for this study, as deidentified, publicly available data obtained from human subjects was analyzed. Data is available in the public domain as outlined below.

Systematic Review and Data Acquisition

A COMPREHENSIVE review of the literature for scRNA seq studies involving normal human liver yielded three recent publications: 1. Aizarani *et. al.*, and based on the methods used in the study, it was referred to as "Liver Atlas, Cholangiocyte and Endothelial enriched" ($LACE^e$), 2. Zhao *et. al.*, referred to as "Liver CD45+ enriched" ($LCD45^e$), and 3. MacParland *et. al.*, referred to as "Non-biased Liver" (L^{nb}).[3–5] Each raw dataset was found in the Gene Expression Omnibus, $LACE^e$: GSE124395, $LCD45^e$: GSE125188, L^{nb}: GSE115469.[26] Datasets were imported into R as UMI count matrices using the RStudio interface. The workflow for isolation of our cell populations of interest from each dataset is summarized in **Figure 1**. Briefly, the CD45+ cell compartment was isolated in the $LACE^e$ and L^{nb} studies in order to extract the leukocyte population and exclude hepatocytes and other non-parenchymal cells. For $LCD45^e$, the CD45 population was isolated prior to sequencing, and included cells derived from liver, spleen, and peripheral blood that had been barcoded for source identification. In this case, only the liver cells were extracted for analysis. Authors from each dataset were contacted for clarification on aspects of the datasets as needed in order to ensure the accuracy of our analysis.

Data Analysis and Visualization

The Seurat version 3.0 R toolkit was used for all data analysis due to its ability to integrate across datasets; the data analysis pipeline followed the tutorial outlined by the package developers.[20,21] To normalize the data, UMI counts were scaled by library size and a natural log transformation; gene counts for each cell are divided by the total UMI count of that cell, scaled by a factor of

10,000, and then transformed via a natural log plus 1 function (“NormalizeData”). For downstream analysis, normalized data is additionally scaled so that the mean expression across cells is 0 and the variance is 1 (“ScaleData”). In order to reduce the dimensionality of the data for clustering functions, Principal Component Analysis (PCA) was utilized, and we determined the first 30 principal components explained sufficient observed variance (“RunPCA”). Next, to identify clusters within the reduced dimensional space, cells were embedded in a k-nearest neighborhood-based graph structure (“FindNeighbors”) and were then partitioned into clusters (“FindClusters”). Finally, to aid in visualization, Uniform Manifold Approximation and Projection (UMAP) was run over the reduced dimensional space (‘RunUMAP”) and identified clusters were projected on to the UMAP plot. Accounting for noise and batch differences between the three studies was done by using reference cells from each. These anchors were identified (“FindIntegrationAnchors”), which represent pairs of cells from each dataset identified and scored based on their close proximity using k-nearest neighborhood approach. These anchors are then used to measure the expression difference between studies (“IntegrateData”), which is then removed from the corresponding normalized data. Integration is run between the normalization and scaling steps.[21]

Utilizing the UMAP plot of the integrated data, we were able to manually identify four major immune subpopulations by plotting expression of known biomarkers and grouping previously identified clusters: myeloid cells, NK&T cells, B cells, and plasma cells (**Supplemental Figure 1**). CD3D was used to identify T lymphocytes, and NK cells were identified by expression of KLRF1 and FCGR3A. The myeloid cell lineage was identified using FCGR3A and the specific marker, CD14. The B cell cluster was classified using the CD19 marker, and plasma cells were identified by SDC1 (CD138) expression.[27–31] After clustering, immune cell proportions were quantified and characterized across studies using a Chi-squared test (α=0.05). Gene correlation analysis was

performed across studies as pooled cell types and additionally with stratification by major cell type using both standard linear regression and rank-rank hypergeometric overlap (RRHO).

Differential expression analysis

DE genes were identified across immune cell subpopulations (i.e., between a particular cell subpopulation and all other cell subpopulations) within a dataset. Log normalized gene expression were used, averaging over all cells in a given subpopulation for a specific gene and taking the difference to the same of all other cell subpopulations ("FindMarkers"), using the Wilcoxon rank-sum test for significance. Genes that were DE were examined both within an immune cell subpopulation and across studies. Similarly, DE genes were identified across datasets (i.e., pairwise between datasets) within a particular cell subpopulation using the same methods. To account for some technical variation between dataset, we calculated the difference of differences of a particular subpopulation between datasets and all other cell subpopulations between datasets. Log normalized gene expression was used as described above taking the difference of differences between dataset (pairwise) and cell subpopulation (particular versus all other), using a t-test for significance.

The candidates for immune profile of human liver across our major cell types of interest were identified by applying differential expression analysis across immune cell subpopulations, while identifying genes which were not DE across datasets and pooling expression values.

Pathway analysis

Gene IDs, expression fold changes and p-values generated from differential expression analyses were imported into Ingenuity Pathway Analysis software (IPA, Qiagen). Core expression analysis was performed which generated canonical pathways which were upregulated or downregulated based on these data. Gene function heatmaps were also generated in order to provide a comprehensive expression profile.

Volcano Plots

Pairwise comparisons of genes between data sets or cell types were conducted using Seurat "FindMarkers" function with default Wilcoxon Rank Sum test. Volcano plots were constructed using these results; genes were identified and colored based on $P\text{-}adj_{Bonfcorr}<0.05$ and a ratio of expression (exponentiated log fold change) ≥1.25 or ≤0.8. Additionally, differences of differences were calculated in a similar way comparing the differential expression between datasets and the differential expression between immune cell subpopulations. T-tests were performed to determine significance and genes were colored based on the previous criteria plus $P\text{-}adj_{Bonfcorr}<0.05$ of the difference of differences effect.

Correlation Analysis

Expression levels were averaged over cells and plotted pairwise between datasets. Scatter plots were fit with linear and quadratic regression models to show potential relationships. Additionally, genes were ranked based on their differential expression between cell types and these rankings were compared between datasets. Correlation between the rankings was assessed multiple ways. First via scatter plots and Spearman's correlation coefficient. Second using Rank-Rank Hypergeometric Overlap from the RRHO package in R and constructing heatmaps showing regions of significant overlap between the ranked lists.[32]

Heatmaps

The top 10 DE genes between cell types were identified and combined between each dataset. Raw expression values for these genes were then normalized by subtracting the mean expression and dividing by the standard deviation of a given gene over all cells. Heatmaps were constructed based on these values; identified genes are shown as rows while cells are shown as columns. Cells are grouped based on cell type while genes are grouped based on the cell type differential expression from which they were identified.

Chord Plots

Results from the pathway analysis were used to select cell functions that were relevant to the immune subpopulations. Cell functions and genes were linked to represent the presence of a connection between the two, yes/no. From there, plots were constructed (Circlize "chordDiagram") by connecting selected cell functions to the genes that were identified performing them, grouping by cell function.

DATA AVAILABILITY

Data used in this study are publicly available in the Gene Expression Omnibus, $LACE^{e}$: GSE124395, $LCD45^{e}$: GSE125188, L^{nb}: GSE115469.

CODE AVAILABILITY

R code used for data analysis and creation of graphics is available upon request and can be accessed via github.

REFERENCES

1. Chen, G., Ning, B. & Shi, T. Single-cell RNA-seq technologies and related computational data analysis. *Front. Genet.* **10**, 1–13 (2019).

2. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **50**, (2018).

3. MacParland, S. A. *et al.* Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat. Commun.* (2018) doi:10.1038/s41467-018-06318-7.

4. Zhao, J. *et al.* Single-cell RNA sequencing reveals the heterogeneity of liver-resident immune cells in human. *Cell Discov.* (2020) doi:10.1038/s41421-020-0157-z.

5. Aizarani, N. *et al.* A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature* (2019) doi:10.1038/s41586-019-1373-2.

6. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).

7. Melé, M. *et al.* The human transcriptome across tissues and individuals. *Science (80-.).* **348**, 660–665 (2015).

8. Psaty, B. M., Rich, S. S. & Boerwinkle, E. Innovation in Genomic Data Sharing at the NIH. *N. Engl. J. Med* (2019).

9. RFA-RM-21-007: Pilot Projects Enhancing Utility and Usage of Common Fund Data Sets (R03 Clinical Trial Not Allowed). https://grants.nih.gov/grants/guide/rfa-files/RFA-RM-21-007.html.

10. Freitas-Lopes, M., Mafra, K., David, B., Carvalho-Gontijo, R. & Menezes, G. Differential Location and Distribution of Hepatic Immune Cells. *Cells* **6**, 48 (2017).

11. Racanelli, V. & Rehermann, B. The liver as an immunological organ. *Hepatology* **43**, (2006).

12. Zheng, M. & Tian, Z. Liver-Mediated Adaptive Immune Tolerance. *Front. Immunol.* **10**, (2019).

13. Wang, Y. & Zhang, C. The Roles of Liver-Resident Lymphocytes in Liver Diseases. *Front. Immunol.* **10**, 1582 (2019).

14. Van Den Brink, S. C. *et al.* Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14**, 935–936 (2017).

15. Denisenko, E. *et al.* Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol.* **21**, 1–25 (2020).

16. Bonnycastle, L. L. *et al.* Single-cell transcriptomics from human pancreatic islets: Sample preparation matters. *Biol. Methods Protoc.* (2019) doi:10.1093/biomethods/bpz019.

17. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574 (2013).

18. Plaisier, S. B., Taschereau, R., Wong, J. A. & Graeber, T. G. Rank-rank hypergeometric overlap: Identification of statistically significant overlap between gene-expression

signatures. *Nucleic Acids Res.* **38**, 1–17 (2010).

19. Rizzetto, S. *et al.* Impact of sequencing depth and read length on single cell RNA sequencing data of T cells. *Sci. Rep.* **7**, 1–11 (2017).

20. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).

21. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

22. O'Flanagan, C. H. *et al.* Dissociation of solid tumor tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses. *Genome Biol.* **20**, 1–13 (2019).

23. Barr, T. L. *et al.* Genomic biomarkers and cellular pathways of ischemic stroke by RNA gene expression profiling. *Neurology* **75**, 1009–1014 (2010).

24. Wang, X. *et al.* Concomitant cortical expression of TNF-α and IL-1β mRNAs follows early response gene expression in transient focal ischemia. *Mol. Chem. Neuropathol.* **23**, 103–114 (1994).

25. Heymann, F. & Tacke, F. Immunology in the liver-from homeostasis to disease. *Nature Reviews Gastroenterology and Hepatology* vol. 13 88–110 (2016).

26. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).

27. Roda-Navarro, P. *et al.* Human KLRF1, a novel member of the killer cell lectin-like receptor gene family: Molecular characterization, genomic structure, physical mapping to the NK gene complex and expression analysis. *Eur. J. Immunol.* **30**, 568–576 (2000).

28. Zhang, A. W. *et al.* Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nat. Methods* **16**, 1007–1015 (2019).

29. Shimizu, Y. *et al.* FcγRIIIA-mediated activation of NK cells by IgG heavy chain complexed with MHC class II molecules. *Int. Immunol.* **31**, 303–314 (2019).

30. Wu, Z., Zhang, Z., Lei, Z. & Lei, P. CD14: Biology and role in the pathogenesis of disease. *Cytokine Growth Factor Rev.* **48**, 24–31 (2019).

31. O'Connell, F. P., Pinkus, J. L. & Pinkus, G. S. CD138 (Syndecan-1), a Plasma Cell Marker: Immunohistochemical Profile in Hematopoietic and Nonhematopoietic Neoplasms. *Am. J. Clin. Pathol.* **121**, 254–263 (2004).

32. Cahill, K. M., Huo, Z., Tseng, G. C., Logan, R. W. & Seney, M. L. Improved identification of concordant and discordant gene expression signatures using an updated rank-rank hypergeometric overlap approach. *Sci. Rep.* **8**, 1–11 (2018).

ACKNOWLEDGEMENTS

We thank the Eli and Edythe Broad Center for Regenerative Medicine at the University of Southern California, Keck School of Medicine for providing the Broad Clinical Research Fellowship which was awarded to Dr. Rocque for 2020-2021. We also thank the USC Libraries Bioinformatics core for providing software licensing and resources.

AUTHOR INFORMATION

BR contributed to study design, acquiring data, analyzing data, and writing the manuscript. AB contributed to study design and writing the manuscript. PS was responsible for acquiring initial datasets and performed computational analyses and data integration. CG contributed to study design, analyzing data, statistical analyses and aided with writing the manuscript. DGH helped with study design and writing the manuscript. YEL provided resources for data analysis and contributed to study design. NU provided insights for study design and contributed to verification of data analysis techniques. JL and OA helped with the initial study design. JE conceived the original idea, supervised the project and contributed to study design and data analysis.

COMPETING INTERESTS

The authors declare no competing interests.

TABLES

Table 1. Comparison of methods across three peer-reviewed single cell RNA sequencing studies of the liver

	LACE[e]	L[nb]	LCD45[e]
Samples	9 liver resection pts mCRC ICC	5 caudate lobes of DBD Ltxp donors	3 adult Ltxp donors blood, spleen & liver
Perfusate	HEPES	HTK solution, then HBS + EGTA	Not addressed
Cell fractionation	PHHs and NPCs isolated, mixed then FACS	None	Cell filtration, centrifugation and Ficol
Removal of non-viable cells	Gradient centrifugation	Trypan blue exclusion	eFluor 450 exclusion
Preservation method	Mix of cryopreserved and fresh	Fresh	Fresh
Cell enrichment	Yes - FACS	None	Yes - FACS
Enrichment for Lymphocytes	No	No	Yes - marker CD45
Enrichment for LSECs	Yes - marker CLEC4G	No	No
Enrichment for MaVECs	Yes - marker CD34, PECAM	No	No
Enrichment for Cholangiocytes	Yes - marker EPCAM	No	No
Removal of low-quality cells	Yes - excluded >2% KCNQ1OT1 transcripts	Removed >50% mitochondrial content	No
Number of cells scRNA seq	10,372	8,444	70,706
scRNA seq technique	mCEL-Seq2	10x	10x
Identification of cell types	RaceID3 (mintotal = 1000, minexpr = 2, minnumber = 10, outminc = 2, cln = 15)	Not addressed	Seurat v3
Samples across pts	Co-clustered	Co-clustered	Co-clustered
Different preps within pts	Co-clustered	Not applicable	Not applicable
Cell doublets	Not addressed	Not removed	Removed

LACE[e] *Aizarani et al. A human liver cell atlas reveals heterogeneity and epithelial progenitors*
L[nb] *MacParland et al. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations*
LCD45[e] *Zhao et al. Single-cell RNA sequencing reveals the heterogeneity of liver-resident immune cells in human*

FIGURES AND FIGURE LEGENDS

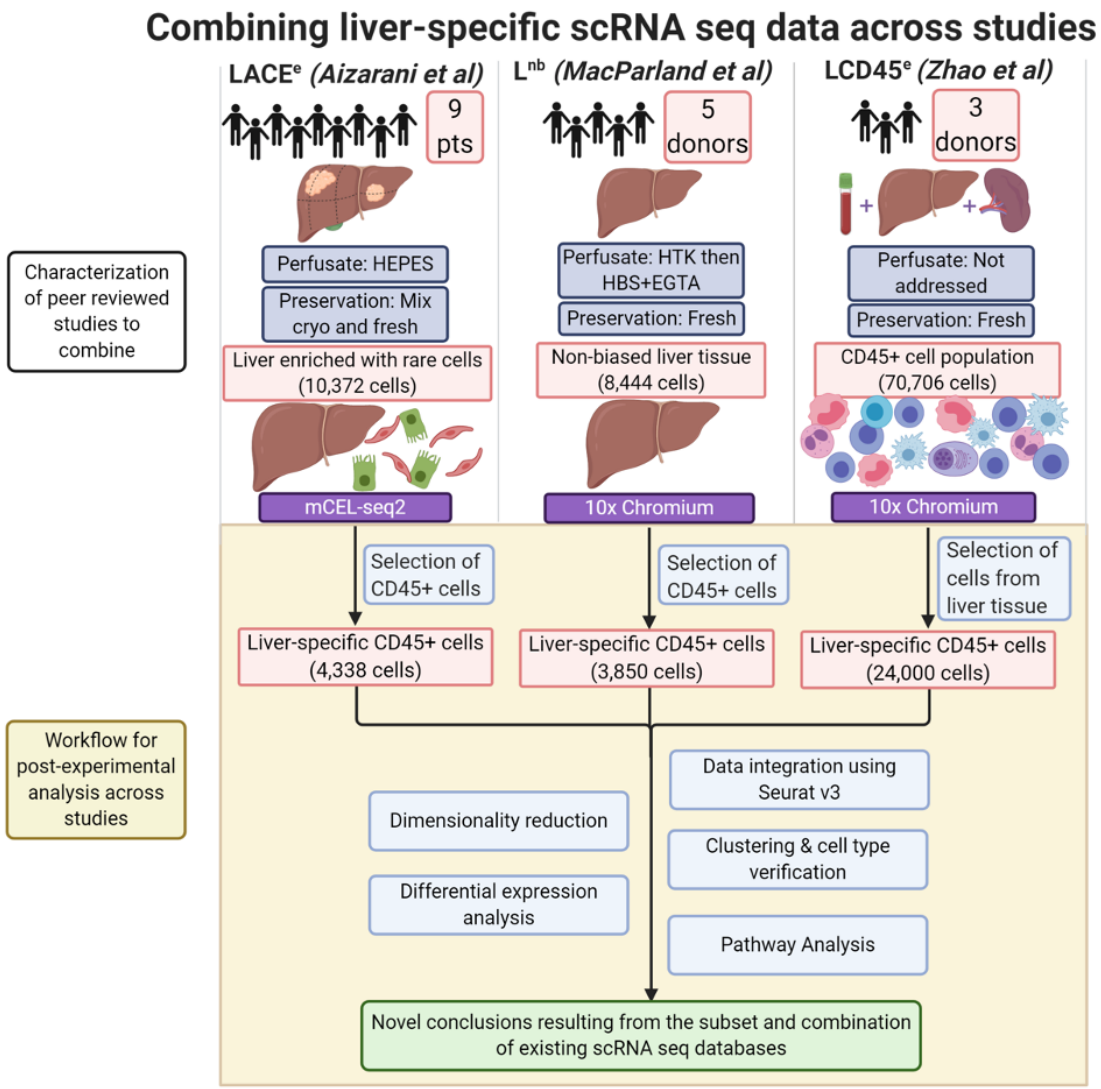


Figure 1: Combining liver-specific scRNA seq data to create a meta-analysis of human liver immune cells.
Schematic diagram summarizing the number of subjects and major pre-analytical variables pertaining to scRNA sequencing of human liver. Post-experimental analysis performed in this study is highlighted in the shaded region. CD45+ cells were selected from the LACE^e and L^{nb} studies. CD45+ cells (which were already pre-selected for in the LCD45e dataset) which were of liver origin were post-experimentally extracted as splenic and peripheral blood cells were also included in the original dataset. The scRNA seq analysis pipeline (Seurat v3 in R) was applied, then clustering and differential expression analysis were used to assess the ability to combine scRNA seq techniques and generate datasets for tertiary pathway analysis of immune function in the normal human liver. Created with BioRender.com.

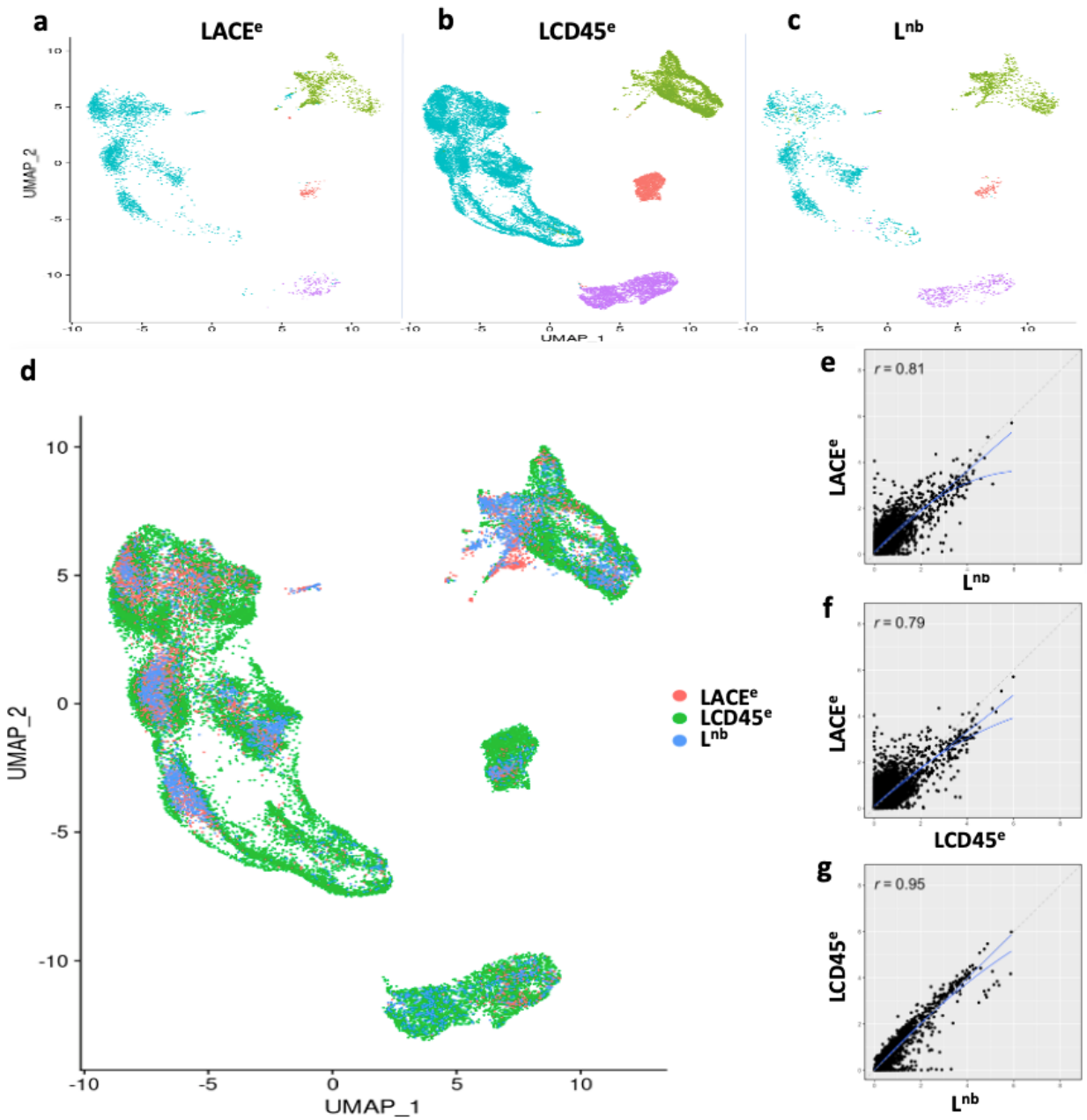
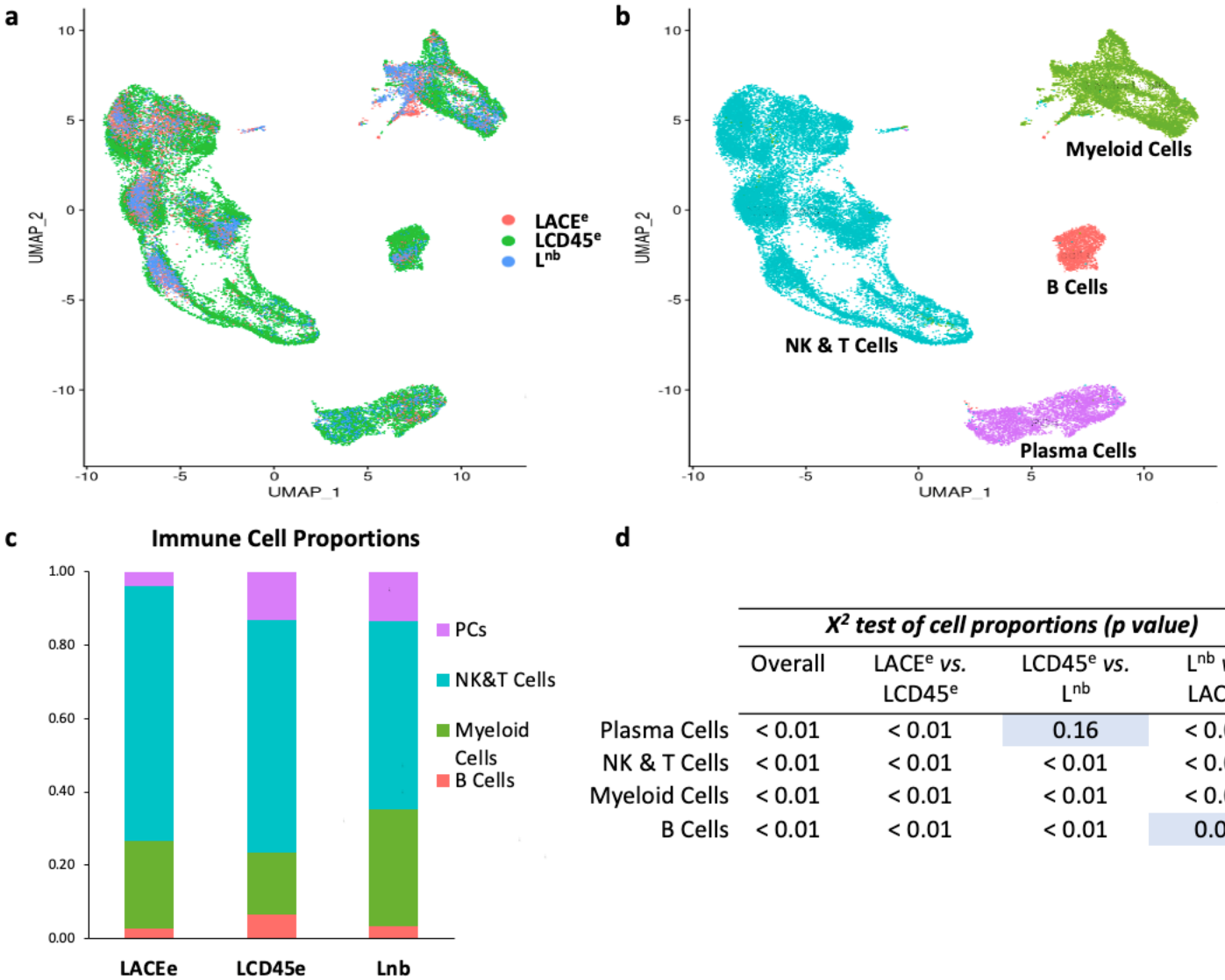


Figure 2: Data visualization with clustering and gene expression correlation across datasets.
a-c. UMAP plots of individual studies. d. Plot showing co-clustering across three single cell datasets. e-g. Gene expression correlation plots with solid lines showing linear and quadratic regressions. Dashed line representing idealized relationship.



	X^2 test of cell proportions (p value)			
	Overall	LACE^e *vs.* LCD45^e	LCD45^e *vs.* L^{nb}	L^{nb} *vs.* LACE^e
Plasma Cells	< 0.01	< 0.01	0.16	< 0.01
NK & T Cells	< 0.01	< 0.01	< 0.01	< 0.01
Myeloid Cells	< 0.01	< 0.01	< 0.01	< 0.01
B Cells	< 0.01	< 0.01	< 0.01	0.08

Figure 3: Cell type composition over CD45-selected liver single cell RNA seq datasets.
a. Cell clustering color-coded by each individual dataset repeated from Figure 2 for side-by-side comparison with panel 3b. b. Cell clustering color-coded by major cell population based on immune cell subpopulation-specific markers. c. Bar graph representation of immune cell proportions across each individual dataset. d. X^2 comparison showing significant differences in proportions of immune cell subpopulations. With pairwise analysis, most chi-squared values reached statistical significance except for plasma cells in LCD45^e versus L^{nb} and for B Cells in L^{nb} versus LACE^e (indicated by blue rectangles).

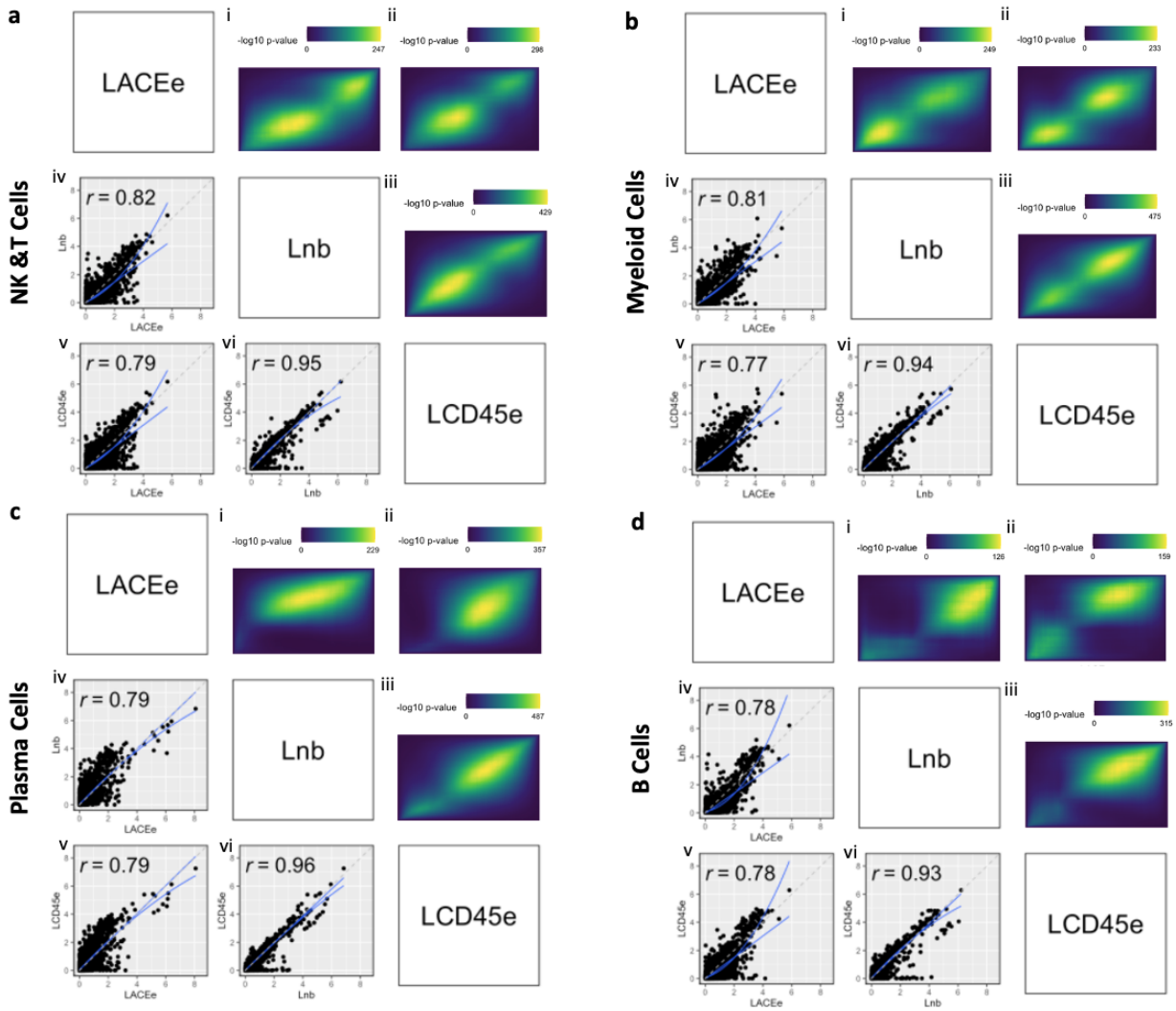
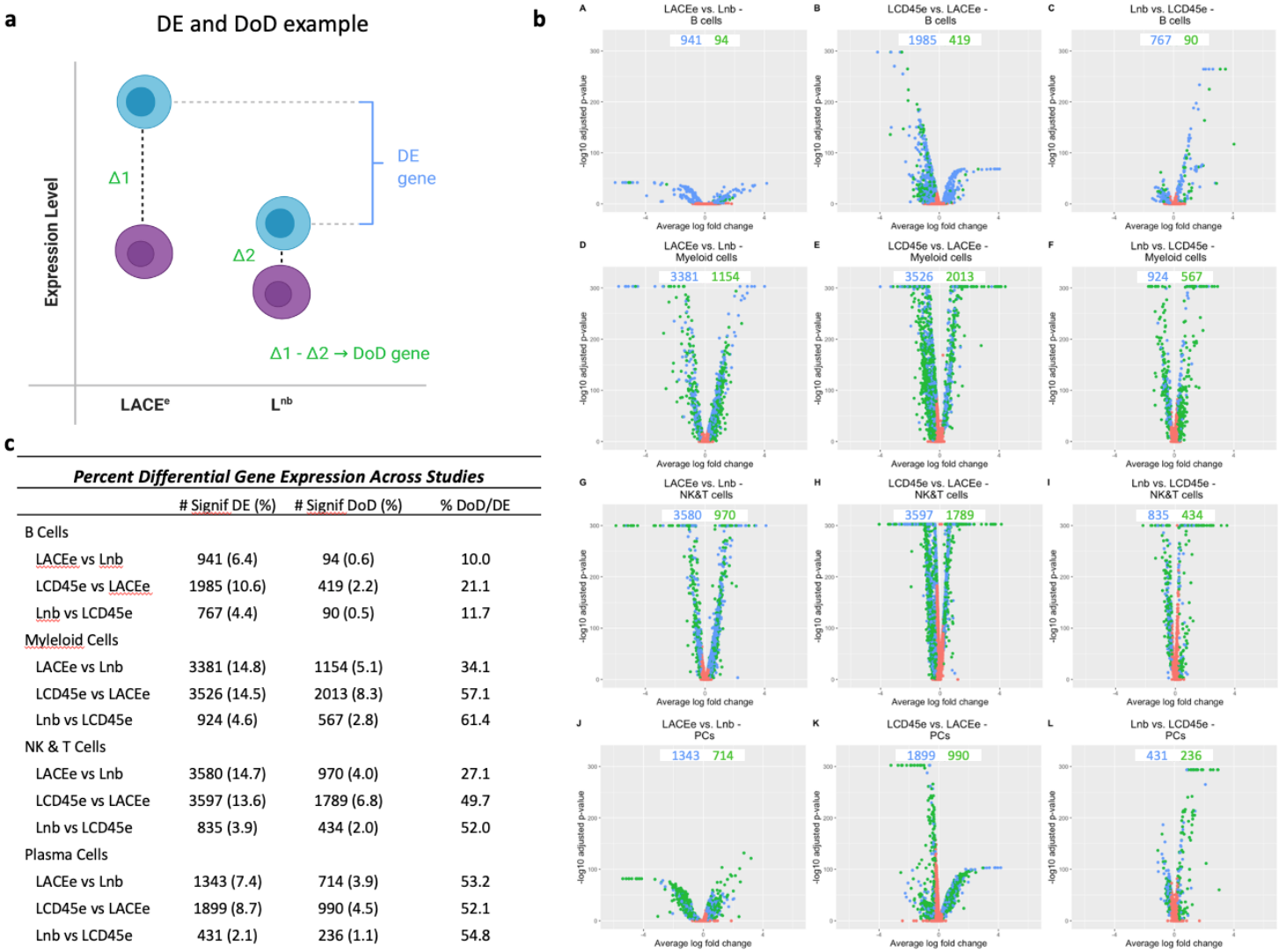


Figure 4: Correlation of gene expression based on immune cell subpopulation.
Pairwise expression correlation analysis was made for each cell type. a. Gene expression among NK and T cells was compared between datasets. The black and gray panels show how well genes correlate between the three datasets and the dashed diagonal line represents an idealized relationship. The straight solid blue line represents the linear best fit and the curved solid blue line shows the quadratic relationship with linear correlation coefficients listed. Colored diagrams show heatmaps of the rank-rank hypergeometric overlap. Yellow regions indicate that for these pairwise comparisons, NK and T cells have a higher amount of agreement or overlap among the more highly expressed genes (lower left quadrant of the graph) and also with a good amount of agreement in genes that are expressed at lower levels (top right quadrant of the graph). b. The same analysis is repeated for myeloid cells. RRHO shows differing amount of overlap depending on the ranked gene expression. c. Correlation analysis for plasma cells. The best correlation is among genes with mid-to-low ranked expression. d. Correlation analysis for B cells. Correlation is maximized in the genes expressed at lower levels.

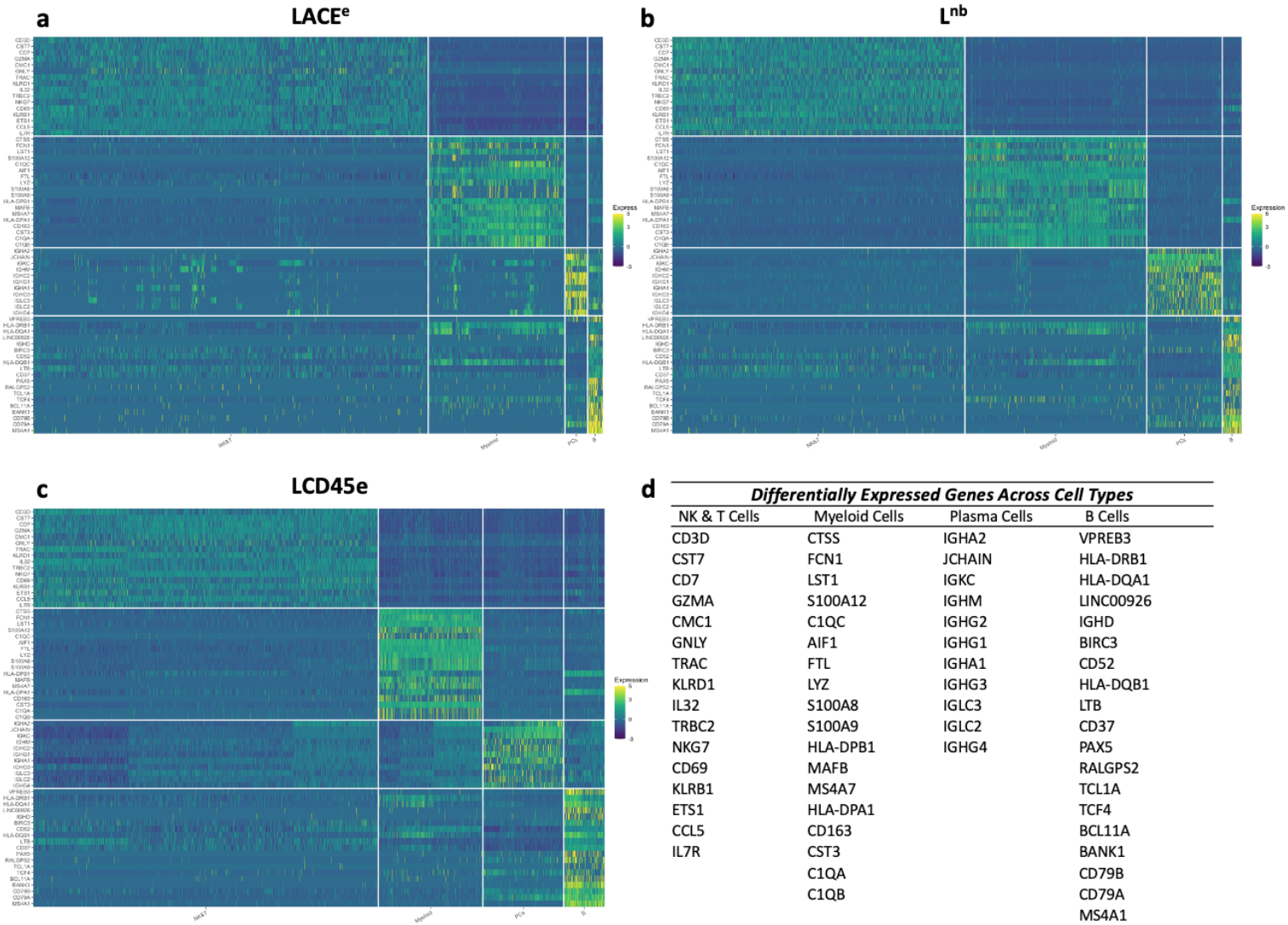


c

Percent Differential Gene Expression Across Studies

	# Signif DE (%)	# Signif DoD (%)	% DoD/DE
B Cells			
LACEe vs Lnb	941 (6.4)	94 (0.6)	10.0
LCD45e vs LACEe	1985 (10.6)	419 (2.2)	21.1
Lnb vs LCD45e	767 (4.4)	90 (0.5)	11.7
Myleloid Cells			
LACEe vs Lnb	3381 (14.8)	1154 (5.1)	34.1
LCD45e vs LACEe	3526 (14.5)	2013 (8.3)	57.1
Lnb vs LCD45e	924 (4.6)	567 (2.8)	61.4
NK & T Cells			
LACEe vs Lnb	3580 (14.7)	970 (4.0)	27.1
LCD45e vs LACEe	3597 (13.6)	1789 (6.8)	49.7
Lnb vs LCD45e	835 (3.9)	434 (2.0)	52.0
Plasma Cells			
LACEe vs Lnb	1343 (7.4)	714 (3.9)	53.2
LCD45e vs LACEe	1899 (8.7)	990 (4.5)	52.1
Lnb vs LCD45e	431 (2.1)	236 (1.1)	54.8

Figure 5: Differential gene expression analysis across datasets.
a. Schematic diagram for identification of a differentially expressed (DE) gene. The blue cells represent different expression levels for a gene in each leukocyte subpopulation (e.g., expression in B cells) and the purple cells represent the expression levels of that same gene, but in all other cell subpopulations. Differential expression only takes levels from the B cell into account. We also examine the Difference of Differences (DoD) which takes the difference in expression between B cells and all other cell types for dataset #1 and subtracts it from the differences in expressed between B cells and all other cell types for dataset #2 which may help categorize the differential expression that is accounted for by technical differences between studies (such as sequencing depth) rather than biologically relevant changes. b. Volcano plots showing the number of DE genes (blue dots) and DoD genes (green dots) for each dataset condition as pairwise comparisons between studies and stratified by cell type. c. Lists of differential gene expression numbers, DoD genes and then the percent of DoD genes out of the number of DE genes.



Differentially Expressed Genes Across Cell Types

NK & T Cells	Myeloid Cells	Plasma Cells	B Cells
CD3D	CTSS	IGHA2	VPREB3
CST7	FCN1	JCHAIN	HLA-DRB1
CD7	LST1	IGKC	HLA-DQA1
GZMA	S100A12	IGHM	LINC00926
CMC1	C1QC	IGHG2	IGHD
GNLY	AIF1	IGHG1	BIRC3
TRAC	FTL	IGHA1	CD52
KLRD1	LYZ	IGHG3	HLA-DQB1
IL32	S100A8	IGLC3	LTB
TRBC2	S100A9	IGLC2	CD37
NKG7	HLA-DPB1	IGHG4	PAX5
CD69	MAFB		RALGPS2
KLRB1	MS4A7		TCL1A
ETS1	HLA-DPA1		TCF4
CCL5	CD163		BCL11A
IL7R	CST3		BANK1
	C1QA		CD79B
	C1QB		CD79A
			MS4A1

Figure 6: Heatmap representation of dominant differentially expressed genes across each immune cell subpopulation.
a. Heatmap of the most DE genes across leukocyte subpopulation plotted using cells from the LACE^{e} dataset. b. Heatmap of the most DE genes across immune cell subpopulation plotted using cells from the L^{nb} dataset. c. Heatmap of the most DE genes across leukocyte subpopulation plotted using cells from the LCD45^{e} dataset. For all heatmaps, z-scores are shown representing standardized, log normalized expression across cells for a given gene. d. Table of the list of genes used for heatmap analysis for each leukocyte subpopulation. Immune cell subpopulations are grouped from left to right: NK&T, myeloid cells, plasma cells then B cells and the gene lists going from top to bottom are ranked in the same order. The same gene lists were used to analyze cells from each of the three datasets.

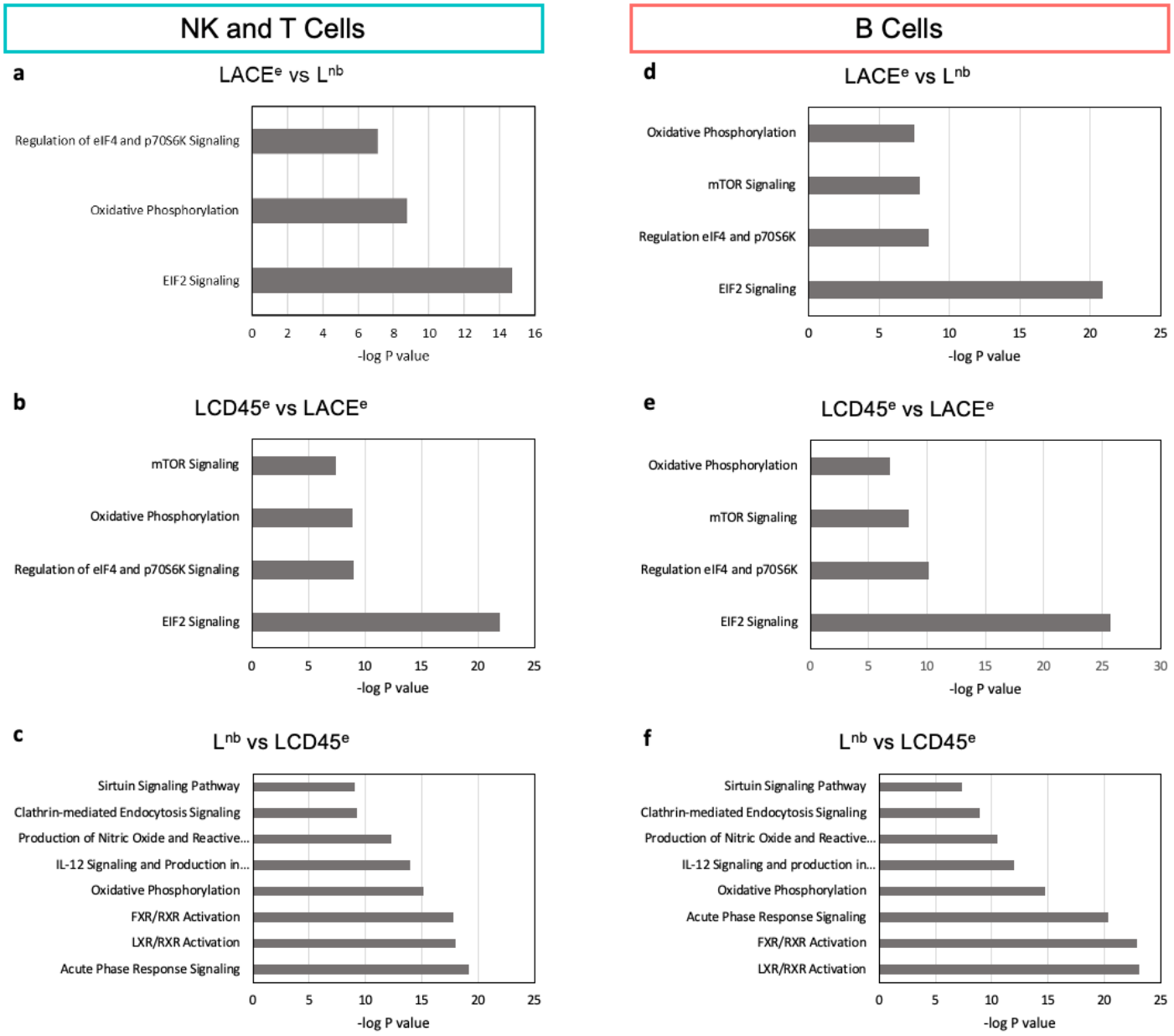


Figure 7: Characterization of differentially expressed genes points to signaling pathways that differ between datasets.
a. Pathway analysis of NK and T cells based on differential expression between the LACE^{e} and L^{nb} datasets. b. Identification of canonical pathways altered between LCD45e and LACEe. c. Pathway analysis of NK and T cells based on differential expression between L^{nb} and LCD45^{e} in NK and T cells. d-f. B cell pathway analysis as in a-c. The same comparisons were made for myeloid cells and plasma cells and were largely similar to what is shown for the immune subpopulations represented in this figure. Bars represent -log(*P* values) of each canonical pathway's likelihood of being altered between comparisons and was obtained with Ingenuity Pathway Analysis Software (Qiagen).

	Function	B Cells	Myeloid	NK&T	PCs
	Inflammatory Response	-1.4	-1.1	-1.3	-1.4
	Cell Death of Cancer Cells	1.9			0.2
Cell Signaling and Interaction	*Activation*	-0.9	-1.4	-1.5	-1.0
	Phagocytosis	-1.8		-2.4	-1.9
	Adhesion	-1.5		-1.4	-1.8
	Binding	-1.7		-1.5	-1.9
	Immune Response	-1.7		-1.5	-2.3
	Interaction			-1.0	-1.5
	Stimulation			-0.9	
	Aggregation			-1.0	-1.3
	Cell-cell Contact			-1.1	-1.6
Cell Death and Survival	*Cell Death*	-1.2	-1.2	0.0	-1.0
	Apoptosis	-0.9	0.4	-0.3	-0.1
	Cytotoxicity	-2.0	-2.2		-2.5
	Cell Viability	-1.1		-2.2	-2.8
	Necrosis	-1.8	-0.5	0.3	-1.4
	Survival	-0.9			
	Fragmentation	-0.3			
	Anoikis			0.5	
	Killing			0.2	
Cell Function and Maintenance	*Engulfment*	-1.9		-2.3	-2.5
	Homeostasis	-0.7	-2.4		-2.2
	Flux	-1.5	-2.0		-2.4
	Internalization	-1.9		-2.0	-1.6
	Endocytosis	-2.0		-1.9	-2.9
	Ion Homeostasis	-1.6	-2.0		-2.4
	Transmembrane Potential	2.0	-1.0		-1.3
	Permeability			-1.1	
	Development			0.4	
	Formation			-2.0	-1.3
Cell Growth and Proliferation	*Proliferation*	0.3	-1.4	0.5	-0.4
	Differentiation	0.1		-1.7	-0.5
	Colony Formation	-0.5	1.0		0.4
	Cell Proliferation	-0.2	-1.7	-0.2	-1.4
	Leukopoiesis	-0.4		-2.1	-0.8
	Myelopoiesis			-1.2	0.3
	Hematopoiesis			-2.1	-0.9
Cellular Movement	*Cell Movement*	-1.3	-2.7	-1.8	-1.9
	Migration	-1.0		-1.5	-1.6
	Chemotaxis	-1.6		-1.9	-1.6
	Invasion	0.3		-1.6	-1.9
	Cell Spreading			-2.2	-0.7

Expression z score
2.5+
1.51-2.5
0.51-1.5
0.5 - -0.49
-0.5 - -1.49
-1.5 - -2.49
-2.5

Figure 8: Phenotypic classification of CD45+ cells in normal human liver points to a low-activity cellular milieu.

Genes that were DE when comparing one leukocyte subpopulation with other subpopulations were identified. From this set of genes, those which were not DE across three datasets were listed as candidate genes of interest. The expression values across datasets were pooled, and pathway analysis was applied to identify upregulation versus downregulation of diseases and functions with 'activation *Z* scores' (calculated from IPA software) for each of the liver immune cell subpopulations: B cells, myeloid cells, NK and T cells and plasma cells. $P<0.05$ was used to establish significance. Gray rectangles denote functions which were not significantly altered within the particular cell type as a result of the DE analysis.

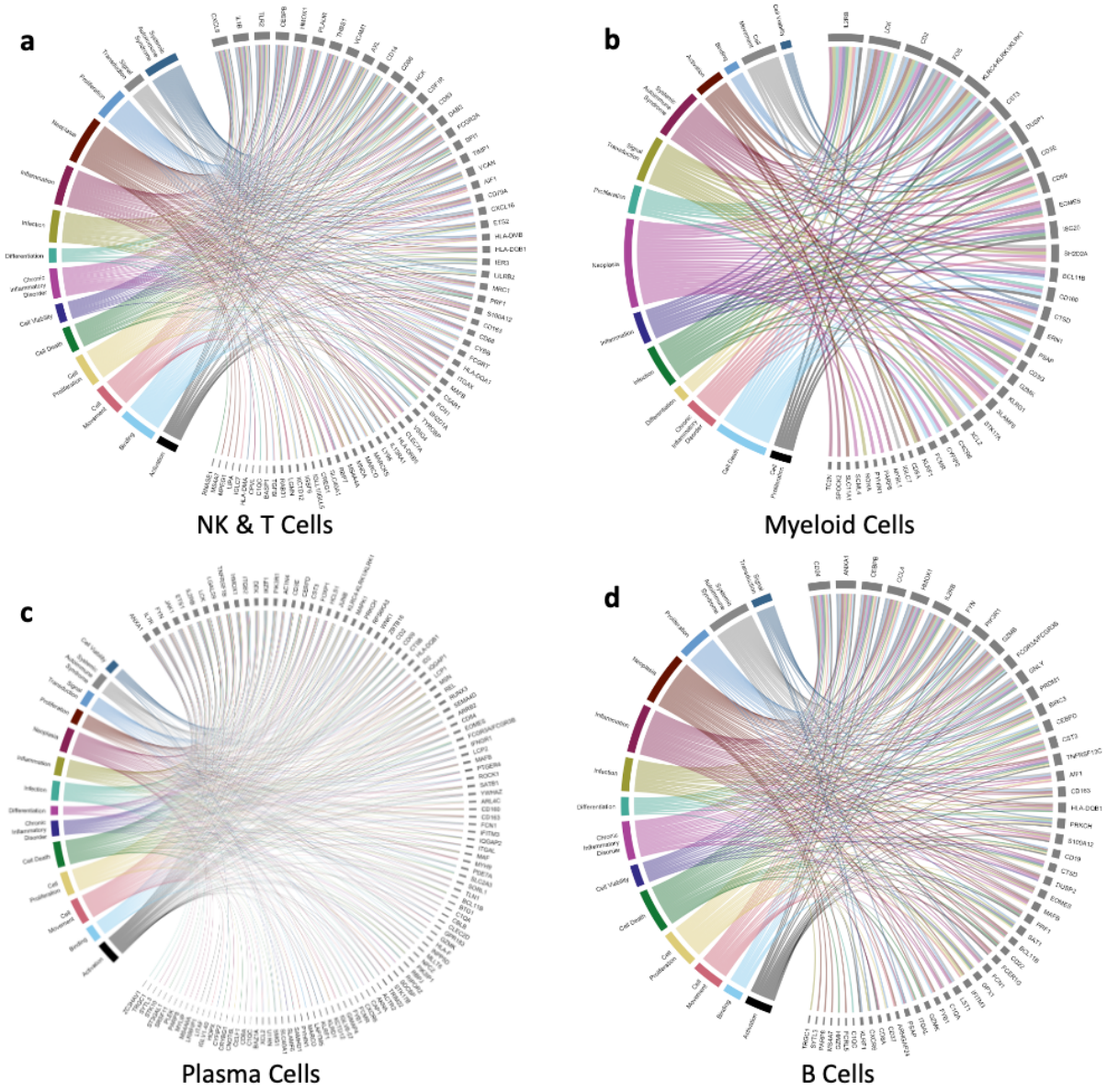


Figure 9: Gene Meta-signatures reveal differential expression signatures among immune cell subpopulations.
a-d. Chord plots representing 14 immune cell-specific diseases or functions (Systemic Autoimmune Syndrome, Signal Transduction, Proliferation, Neoplasia, Inflammation, Infection, Differentiation, Chronic Inflammatory Disorder, Cell Viability, Cell Death, Cell Proliferation, Cell Movement, Binding, Activation) with links to the respective genes which have been identified as DE between the leukocyte subpopulation listed versus all others.