Title

Cellular deconvolution of GTEx tissues powers eQTL studies to discover thousands of novel disease and cell-type associated regulatory variants

Authors

Margaret K. R. Donovan[1,2], Agnieszka D'Antonio-Chronowska[3], Matteo D'Antonio[4], Kelly A. Frazer[3,4#]

[1] Bioinformatics and Systems Biology Graduate Program, University of California, San Diego, La Jolla, CA 92093, USA

[2] Department of Biomedical Informatics, University of California, San Diego, La Jolla, CA 92093, USA

[3] Department of Pediatrics and Rady Children's Hospital, University of California, San Diego, La Jolla, CA 92093, USA

[4] Institute for Genomic Medicine, University of California, San Diego, La Jolla, CA 92093, USA

#Corresponding author

Contact:

Kelly A. Frazer email: kafrazer@ucsd.edu

Abstract

The Genotype-Tissue Expression (GTEx) resource has contributed a wealth of novel insights into the regulatory impact of genetic variation on gene expression across tissues, however thus far has not been utilized to study how variation acts at the resolution of the different cell types composing the tissues. To address this gap, using liver as a proof-of-concept tissue, we show that mouse scRNA-seq can be used as an alternative to human scRNA-seq for the cellular deconvolution of GTEx tissues. Then, using mouse scRNA-seq, we deconvoluted over 6,000 bulk RNA-seq samples corresponding to 28 GTEx tissues and show that we are able to quantify cellular heterogeneity, determining both the different cell types present in each of the tissues and how their proportions vary between samples of the same tissue type. Considering the relative cell type distributions for eQTL analyses in GTEx liver and skin samples, we identified thousands of additional genetic associations that were cell-type-specific and had lower effect sizes. We further show that cell-type-specific eQTLs in skin colocalize with melanoma, malignant neoplasm, and infection signatures, indicating variants that influence gene expression in distinct skin cell types play important roles in skin traits and disease. Overall, our results provide a framework to deconvolute the cellular composition of human bulk RNA-seq using readily available mouse scRNA-seq, which can be implemented immediately for characterizing the functional impact of cell-type-specific genetic variation.

Introduction

Understanding the regulatory impact of genetic variation on complex traits and disease has been a longstanding goal of the field of human genetics. To decipher the mechanistic underpinnings of complex traits, the GTEx Project[1] has generated a large dataset, including over 10,000 bulk RNA-seq samples representing 53 different tissues (corresponding to 30 organs) obtained from 635 genotyped individuals, to link the influence of genetic variants on gene expression levels through expression quantitative trait loci analysis (eQTL). While GTEx has provided important biological insights, it has not yet considered how cellular heterogeneity (i.e. different cell types within a tissue and the relative proportions of each cell type across samples of the same tissue) present in bulk RNA-seq affects genotype-gene expression associations. Because regulation of gene expression varies across cell types, not accounting for cellular composition could result in loss or distortion of signal from relatively rare cell types. It is possible that future studies pursuing cell-type-specific eQTLs may utilize single cell approaches (e.g. single cell RNA-seq; scRNA-seq); however, non-trivial technical challenges, such as hard to dissociate tissues and low capture efficiencies, make the generation of a GTEx-scale single-cell expression dataset a substantial undertaking, which would take years to complete. Thus, as single-cell large-scale scRNA-seq collections progress, our present knowledge of how genetic variation influences cell-type-specific gene expression would greatly benefit from conducting eQTL analyses on bulk GTEx tissue samples whose cellular heterogeneity has been characterized through existing deconvolution methods[2-4].

To characterize the heterogeneity of bulk RNA-seq samples, gene signatures from cell types known to be present in a given tissue can be used to deconvolute the cellular composition (i.e. the proportion of each cell type). The cell-type-specific gene expression signatures needed to deconvolute a heterogenous tissue can be obtained by analyzing scRNA-seq generated from an analogous tissue. However, there are relatively few human scRNA-seq resources currently available[5-9], and thus only a small fraction of GTEx tissues could be deconvoluted using cell-type-specific gene expression signatures derived from existing human single-cell data. While human single-cell data is limited, the Tabula Muris exists[10], which is a powerful resource of scRNA-seq

data from mouse including more than 100,000 cells from 20 tissue types (referred in the Tabula Muris resource as organs and tissues). A recent study showed that similar cell types in humans and mice share sufficient cell-type-specific gene expression signatures to integrate scRNA-seq data between the two species[11], raising the possibility of utilizing the available scRNA-seq from mouse to generate the cell-type-specific gene expression signatures for deconvolution of GTEx tissues.

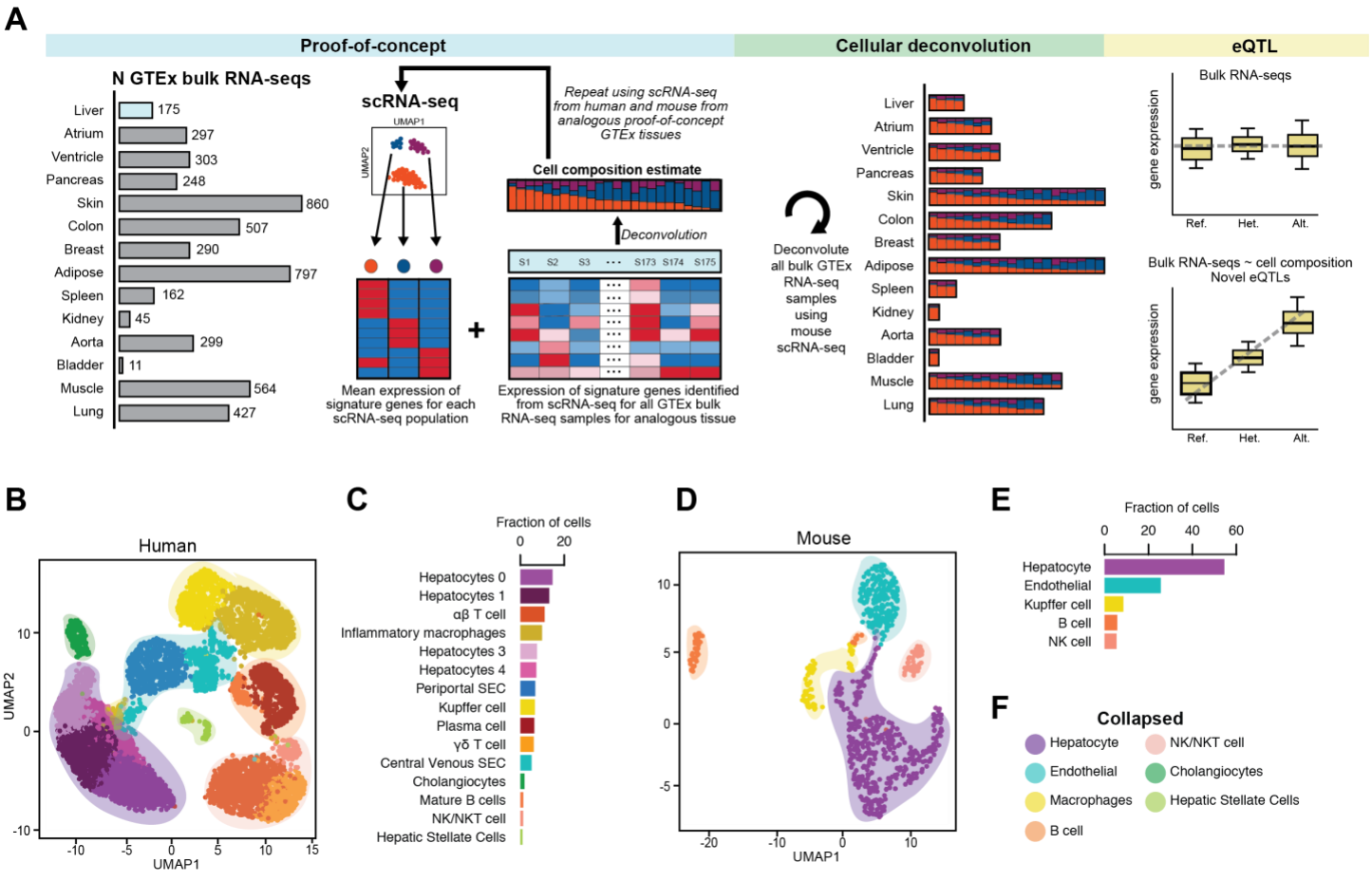
To examine the feasibility of using mouse-derived cell-type-specific gene expression signatures to deconvolute human tissues, we compared cellular composition estimates of GTEx liver samples generated using human scRNA-seq to those generated using the Tabula Muris scRNA-seq resource. We show that the human and mouse single-cell data captured many overlapping cell populations and that using either human-derived or mouse-derived cell-type-specific gene signatures to deconvolute 175 GTEx liver samples resulted in highly correlated estimated cellular compositions. We also show that the number of cells analyzed in a scRNA-seq dataset impacts the ability to both detect less abundant cell types and distinguish between similar cell types (i.e. resolution). We used cell-type-specific gene signatures derived from the Tabula Muris resource to deconvolute over 6,000 additional GTEx samples corresponding to 28 tissues from 14 organs, which enabled us to determine how the fractions of different cell types vary across GTEx samples derived from the same tissue. Using deconvoluted liver and skin GTEx samples for eQTL analyses, we identified thousands of novel (not detected using bulk RNAseq samples) genetic associations that tended to have lower effect sizes, some of which are cell-type-specific. Finally, we show that skin cell-type-specific eQTLs colocalize with GWAS variants for melanoma, malignant neoplasm, and infection signatures, indicating that variants that are functional in limited skin cell types may play major roles in skin traits and disease. Taken together, our study shows the importance of conducting cell-type-specific QTL studies and demonstrates that the estimation of cellular heterogeneity enhances the genetic insights yielded from the GTEx resource.

Results

scRNA-seq from murine and human analogous tissues capture similar cell types

To examine the extent to which scRNA-seq generated from analogous human and mouse tissues captured similar cell types (Table S1), we selected human liver as a proof-of-concept tissue (Figure 1A, "proof-of-concept"). We used previously defined cell types from Tabula Muris mouse liver cells (710 cells; 5 cell types) [10], and to be consistent, we used the Tabula Muris annotation approach to analyze existing human liver scRNA-seq data[5]. In brief, on the 8,119 human liver single-cells, we performed nearest-neighbor graph-based clustering on components computed from principal component analysis (PCA) of variably expressed genes, and then used marker genes to define the cell populations corresponding to each of the 15 previously observed cell types [5]. Human and mouse scRNA-seq from liver captured several shared cell types, including hepatocytes, endothelial cells, and various immune cells (Kuppfer cells, B cells, and natural killer (NK) cells) (Figure 1B-E), however we noted that there were many more distinct cell types for human liver. This was due to the fact that cell type resolution (i.e. the ability to distinguish between similar cell types) increases with the number of cells captured [12]. Some of the 15 cell types identified in the human liver scRNA-seq were highly similar and clustered near each other, for example four hepatocytes populations distinguished by their spatial location (i.e. zonation) and two endothelial cell populations distinguished by zonation (Figure 1B,C). In contrast, for the mouse liver scRNA-seq we only observed one hepatocyte population and one endothelial population (Figure 1D,E). If we collapsed the cell types that were similar to each other in the human scRNA-seq, we obtained 7 distinct cell classes (Figure 1B,F; Table S3), which largely corresponded to the 5 cell types from mouse liver scRNA-seq (cholangiocytes and hepatic stellate cells were absent; Figure D,F). Overall, these results show that scRNA-seq generated from human and mouse liver captured similar cell types and that number of cells analyzed affects the cell type resolution.

Figure 1: Cell composition of liver from human and mouse scRNA-seq



A. Overview of the study design. Our goal was to deconvolute the cellular composition of 28 GTEx tissues from 14 organs using mouse scRNA-seq for the purpose of identifying cell-type-specific eQTLs. We first conducted a **proof-of-concept** analyses, where we compared cellular estimates of each of the proof-of-concept GTEx tissue (liver) after having deconvoluted each using either mouse or human scRNA-seq. We then performed **cellular deconvolution** of the 28 GTEx tissues from 14 organs using CIBERSORT and characterized both the heterogeneity in cellular composition between tissues and the heterogeneity in relative distributions of cell populations between RNA-seq samples from a given tissue. Finally, we used the cell composition estimates as interaction terms for **eQTL analyses** to determine if we could detect novel cell-type-specific genetic associations.

B. UMAP plot of clustered scRNA-seq data from human liver. Each point represents a single cell and color coding of cell type populations (See Methods: Defining the cellular composition of liver) are shown adjacent (Figure 1C). Similar cell types can be collapsed to single cell type classifications and are noted with colored, transparent shading (Figure 1F).

C. Bar plots showing the fraction of each cell type from the scRNA-seq data from human liver. Color-coding of cell types correspond to the colors of the single cells in Figure 1B.

D. UMAP plot of clustered scRNA-seq data from mouse liver. Each point represents a single cell and color coding of cell type populations are shown adjacent (Figure 1E). Each cell type has a corresponding collapsed cell type in human liver and is noted with colored, transparent shading (Figure 1F).

E. Bar plots showing the fraction of each cell type from the scRNA-seq data from mouse liver. Color-coding of cell types correspond to the colors of the single cells in Figure 1D.

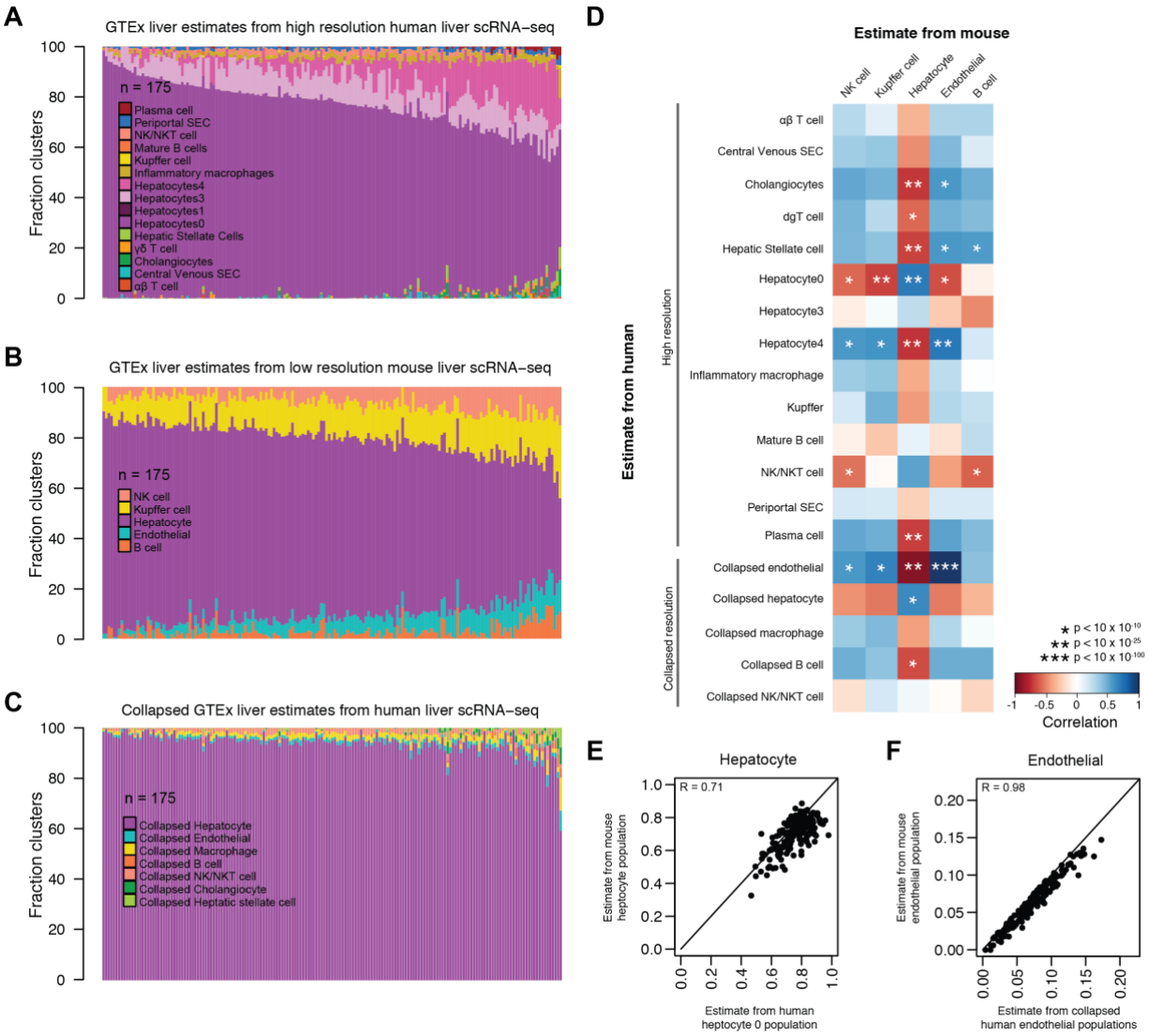
F. Legend showing the colors of collapsed similar cell types from human liver (Shading in UMAP Figures 1B,D; Table S3). All cell types from mouse liver have a corresponding collapsed cell type in human liver (hepatocyte, endothelial, macrophages, B cell, NK/NKT cell) and human liver also contains two additional cell types not present in mouse (cholangiocytes and hepatic stellate cells).

Mouse cell populations can estimate cellular composition of human GTEx samples

To establish the ability to use cell-type-specific gene expression signatures derived from mouse scRNA-seq for the deconvolution of human GTEx tissues, we compared cell composition estimates of bulk RNA-seq deconvoluted using human versus mouse expression profiles (Figure 1A, "proof-of-concept"). To estimate cellular composition, we first obtained signature expression profiles of the top 200 most significantly overexpressed genes for each cell type identified in scRNA-seq from high resolution human liver (i.e. signature genes from 15 cell types) and low resolution mouse liver (i.e. signature genes from 5 cell types) (Table S2). Then, from 175 GTEx bulk liver RNA-seq samples [1], we independently extracted the expression of the signature genes at the two resolutions, and used CIBERSORT[2] to estimate the cellular compositions (i.e. high resolution human liver estimates and low resolution mouse liver estimates) (Figure 2A,B; Table S11,12). To investigate how resolution impacted the correlation between human and mouse signature gene estimates, we also collapsed the high resolution human liver cellular composition estimates for each of the 175 deconvoluted samples by summing the estimates across similar cell types in each of the 7 distinct cell classes (Table S3) (Figures 1B,F and 2C). We then calculated all pairwise-correlations between each of the estimated cell populations in the 175 GTEx liver samples from human (high and collapsed resolution estimates) with the estimated cell populations from mouse (low resolution estimates) (Figure 2D). We found that hepatocyte estimates from mouse liver were positively and highly correlated with the high resolution hepatocyte population

from a single zone (hepatocyte 0) estimate ($r = 0.71$, p-value = 5.4×10^{-28}), less correlated with the collapsed hepatocyte population estimate ($r=0.64$, p-value = 1.015×10^{-21}), but not correlated with any of the other three high resolution hepatocyte populations (Figure 2D,E). This indicates that the low-resolution mouse hepatocyte population corresponds to one of the four human hepatocyte populations/zones. Further, we observed that the endothelial estimates from mouse were not correlated with either high resolution human periportal sinusoidal endothelial cells (SEC) or central venous SECs; however, the collapsed human endothelial population estimates were highly correlated ($r = 0.98$, p-value = 1.2×10^{-115}) (Figure 2F). This indicates that the human endothelial population estimates captured a higher resolution of cell type specificity (i.e. two independent endothelial zones), whereas the mouse endothelial population estimates likely captured a mixture of both cell types (i.e. the two endothelial zones are combined into a single cell population). While in general we observed high correlation in the human and mouse population estimates for most cell types (hepatocytes, endothelial cells, and Kupffer cells), B cells were non-significantly correlated, and NK-like cells were negatively correlated (Figure 2D). This difference in immune cell estimates in GTEx liver is not wholly unexpected, as immune response differences exist between species [13]. Our results show that, while the number of cells captured by scRNA-seq impact the resolution at which cellular composition can be estimated, mouse cell signatures can be used to deconvolute human GTEx bulk RNA-seq samples.

Figure 2: Comparison of GTEx cell estimates using mouse versus human cell signatures



A, B, C. Bar plots showing the fraction of cell types estimated in GTEx liver RNA-seq samples from high resolution human liver scRNA-seq (A), low resolution mouse liver scRNA-seq (B), and collapsed GTEx estimates from high resolution human liver scRNA-seq.

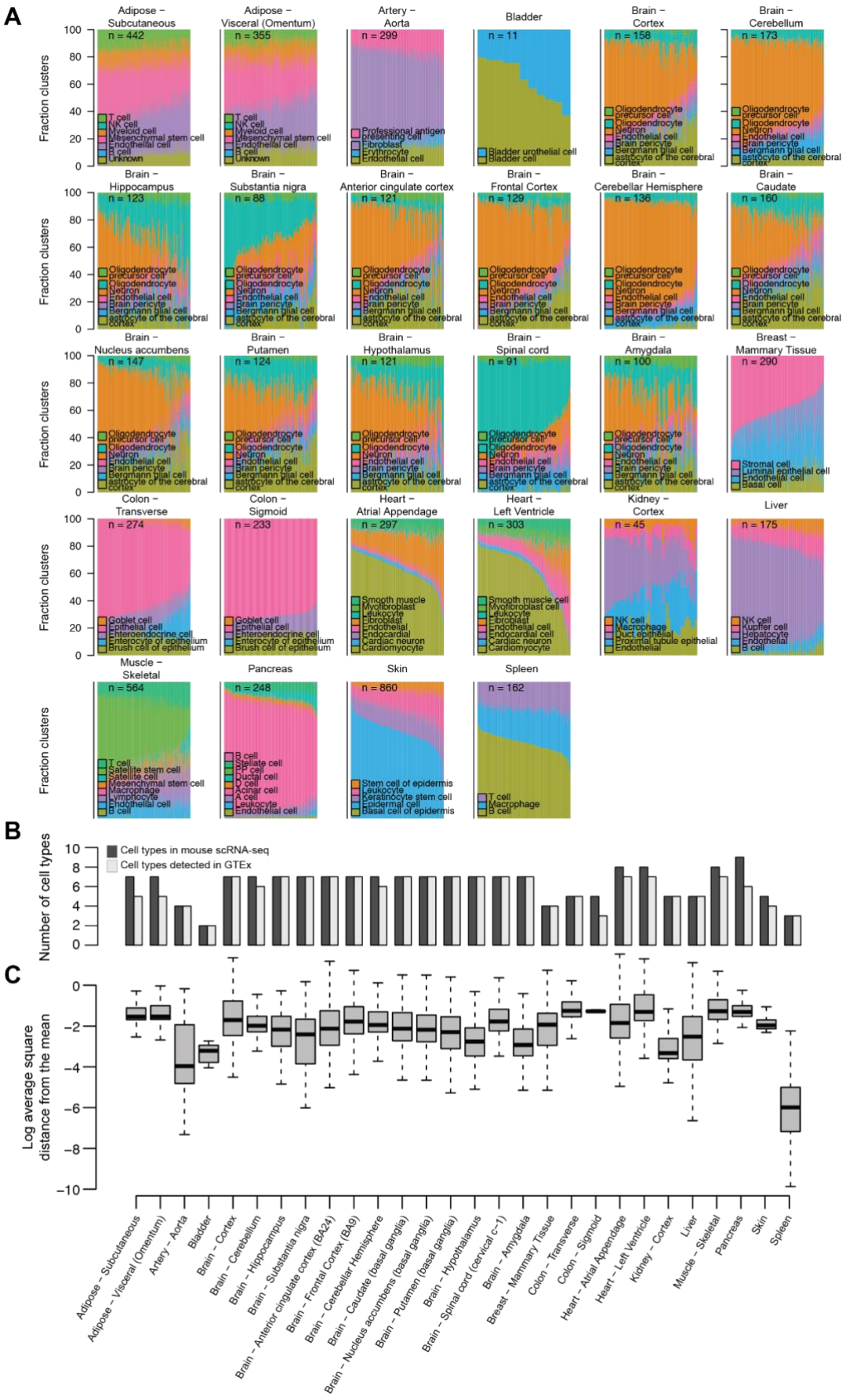
D. Heatmap showing the correlation of GTEx liver cell population estimates from human liver scRNA-seq at high and collapsed resolutions (rows) and mouse liver (columns) at low resolution. Color coding of heatmap scales from red, indicating negative and low correlation in estimates, to blue, indicating positive and high correlation in estimates. Significance is indicated with asterisks.

E, F. Scatter plots of estimated cell compositions across 175 GTEx livers deconvoluted using human scRNA-seq for human hepatocyte 0 population (d) and human collapsed endothelial cells (e) versus estimated cell populations deconvoluted using mouse scRNA-seq.

Cellular deconvolution of GTEx adult tissues show samples are heterogenous

To understand the cellular heterogeneity of GTEx tissues (Figure 1A, "cellular deconvolution"), we used signature genes from 14 mouse tissue types (Table S1,2) to perform cellular deconvolution of 28 GTEx tissues from 14 organs (Figure 3A; Table S1, S4-18), where the number of samples for each GTEx tissue varied from 11 (bladder) to 860 (skin). We found that each deconvoluted GTEx tissue contained a variable number of cell types ranging from two (bladder) to seven (brain and heart) (Figure 3B). Additionally, the relative distribution of the estimated cell types varied between different samples of the same tissue (Figure 3C). Tissues with the least heterogeneous cell population distributions between samples were spleen and aorta (Table S17,4), whereas those with the most heterogeneous cell population distributions between samples were colon, brain (13 tissues), and left ventricle (Table S8,7,18). Examining the tissues corresponding to the same organ, we noted that some had the same cell types estimated at similar distributions (adipose subcutaneous and visceral), some had the same cell types present at variable proportions (heart atrial appendage and left ventricle; 13 brain tissues), and others had variable cell types present/absent (colon transverse and sigmoid). These results reveal a striking heterogeneity in GTEx tissues that has not been previously appreciated and may be contributing noise to eQTL analyses.

Figure 3: Cellular deconvolution of 28 GTEx tissues

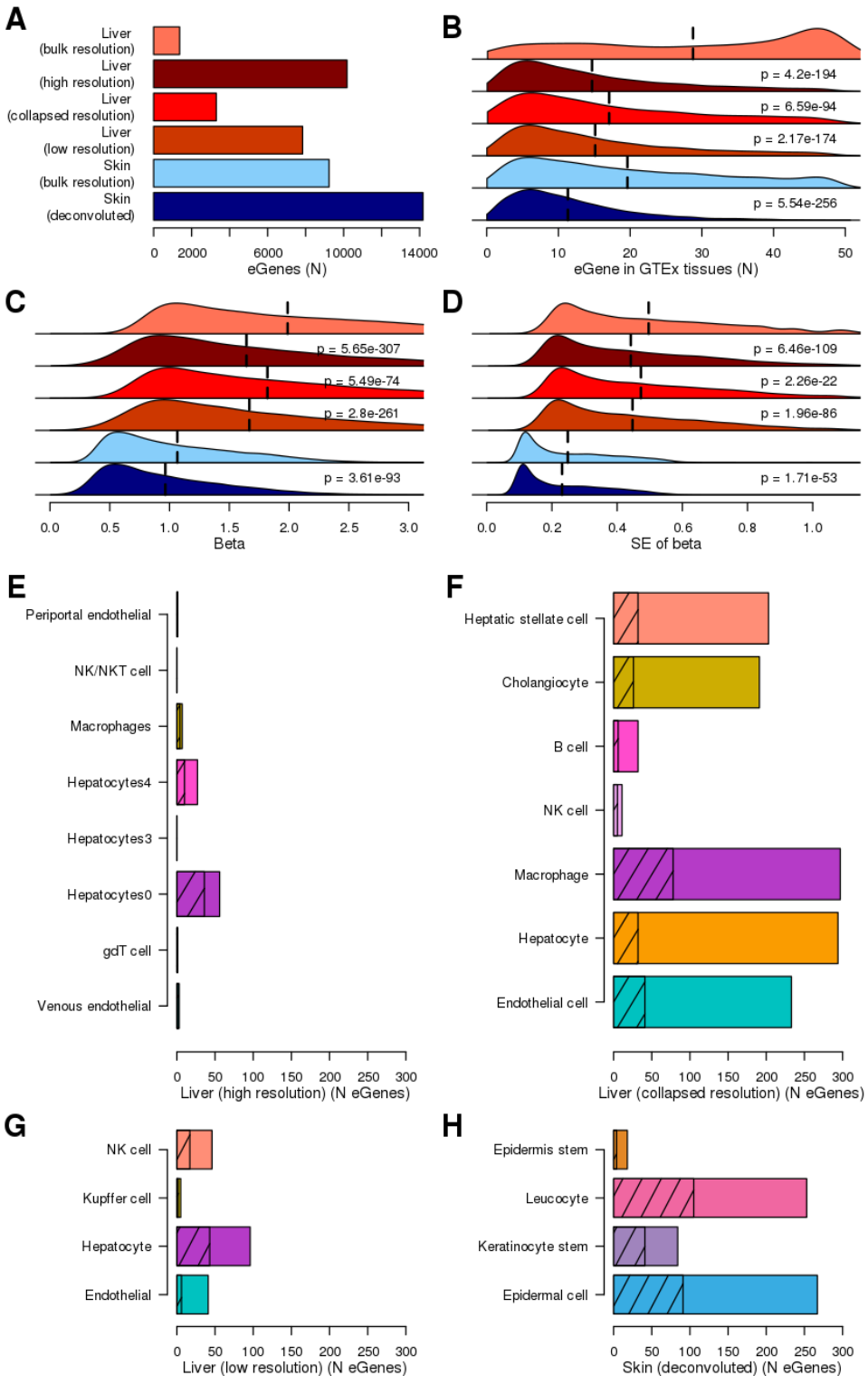


A. Stacked bar plots showing the fraction of cell types estimated in GTEx RNA-seq samples from mouse scRNA-seq.
B. Bar plots comparing the number of cell types discovered in mouse scRNA-seq (light grey) vs. the number of these cell types that were estimable for each GTEx tissue
C. Box plots showing per RNA-seq sample the distribution of the log2 average square distance from the mean estimated cellular compositions for each GTEx tissue

eQTL analyses using deconvoluted tissues increases power

Since we observed heterogeneity in the relative distributions of cell populations across GTEx RNA-seq samples, we hypothesized that considering the cell population distributions of each sample would improve eQTL analysis by increasing our power to detect novel tissue and/or cell-type-specific associations (Figure 1A). We identified 19,621 expressed genes in GTEx liver samples and performed one eQTL analysis not considering cellular heterogeneity (i.e. bulk resolution; Table S19), and three eQTL analysis using cell population estimates as covariates to adjust for cellular heterogeneity (Tables S20-22): 1) considering high resolution human liver estimates (15 cell types; Table S12, 20; Figure 2A); 2) considering collapsed resolution human liver estimates (7 cell types; Table S12,3,21; Figure 2C); and 3) considering low resolution mouse liver estimates (5 cell types; Table S13,22; Figure 2B). Using cell population estimates as covariates we detected many more genes with significant eQTLs (eGenes) than at bulk resolution (Figure 4A). We found that considering high resolution estimates identified the most eGenes (10,117) with 1.3 fold and 3.1 fold more than collapsed and low resolution estimates, respectively. These findings show that conducting eQTL analyses using highly resolved cell population estimates as a covariate significantly increases the power to identify eGenes.

Figure 4: Using cellular deconvolution to discover cell type specific eQTLs



A. Bar plot showing the number of eGenes detected in each eQTL analysis from liver (shades of red) and skin (shades of blue).

B, C, D. Distributions of (b) number of GTEx tissues where each eGene has significant eQTLs, (c) effect size β and (d) standard error of β in liver and skin. Colors are as in panel A. vertical dashed lines represent mean values. P-values were calculated in comparison with the bulk resolution analysis for each tissue using Mann-Whitney U test.

E-H. Bar plots showing the number of eGenes significantly associated with each specific cell population considering cell estimates for: liver high resolution (e), liver collapsed resolution (f), liver low resolution (g), and skin. Total number of eGenes for each cell type indicates the cell type is significantly associated and the hashed number of eGenes for each cell type indicates the association is cell-type-specific (e.g. only significant in that cell type). In cases where a given cell type had no significant association, the bar is not shown.

Given the differences in the number of detected eGenes based on cell-type resolution, we hypothesized that eGenes detected at low powered resolutions (bulk and collapsed resolution) commonly shared eQTLs with other GTEx tissues (i.e. tissue-neutral) and the eGenes detected using higher powered resolutions (high and low resolutions) had more tissue-specific eQTLs (i.e. less frequently in other GTEx tissues). For each resolution, we calculated the number of GTEx tissues in which each eGene has eQTLs. We observed that eGenes identified using cell populations as covariates in general were more tissue-specific (i.e. present in fewer GTEx tissues) than eGenes detected at bulk resolution. Compared to bulk resolution, high resolution eGenes were the most tissue specific ($p = 4.2x10^{-194}$; Mann-Whitney U test), then low resolution eGenes ($p = 2.17x10^{-174}$; Mann-Whitney U test), and collapsed resolution was the least tissue-specific ($p = 6.59x10^{-94}$; Mann-Whitney U test) (Figure 4B), showing that the resolution of cell population estimates used as covariates is correlated with the power of the study to identify tissue-specific eGenes.

Furthermore, using cell populations as covariates resulted in decreased effect size (β) (Figure 4C) and standard error (SE) of β (Figure 4D), where relative to bulk resolution, the higher the resolution of the eQTLs, the smaller the β and SE of β. However, in general the β values for the top hit for each gene were highly correlated between eQTLs detected using cell populations and eQTLs detected without using cell populations ($r > 0.975$,

Figure S1A-C). These results indicate that using cell population distributions as covariates overall reduces the noise, thereby potentially increasing our power to identify eQTLs.

Resolution of deconvoluted tissues impacts the number of identified cell-type-specific regulatory variants

To examine if some of the eQTLs identified using cell population estimates as covariates were cell-type-specific, we used a statistical interaction test[14] to assess if modeling the contribution of a specific cell type significantly improved the observed association between genotype and gene expression. Interaction tests were performed on all independent pairs of eGenes and corresponding lead eQTLs using liver cell type estimates from the high, collapsed, and low resolution as interaction terms. Overall, across the high, collapsed, and low resolutions we respectively detected 74, 528 and 121 cell-type-associated eGenes (i.e. eGene is associated with one or more cell type(s); FDR-corrected p-values < 0.1, χ^2 test, Figure 4E-G) and 54, 220 and 68 cell-type-specific eGenes (i.e. eGene is associated with only one cell type; Figure 4E-G). Notably, using low resolution and collapsed resolution cell populations, we respectively detected 1.6 and 7.1 times more cell-type-associated eQTLs than high resolution cell populations (respectively, $p = 1.9 \times 10^{-7}$ and 7.3×10^{-250}, Fisher's exact test, Figure 4E-G). While initially counter-intuitive to the previous evidence showing higher resolution eGenes are more tissue-specific (Figure 4B) and have decreased noise (Figure 4C,D), it is possible we identify fewer cell-type-associated eGenes using highly resolved cell populations estimates than collapsed estimates, because the eQTL signal in high resolution cell types may be diluted between similar cell populations (i.e. regulatory variants having similar effects across the similar cell types). Overall, these results suggest that accounting for cellular heterogeneity between samples allows for the discovery of novel cell-type-associated (and cell-type-specific) eQTLs.

eQTL analysis of deconvoluted GTEx skin confirms ability to identify cell-type-specific regulatory variants

To further investigate the impact of using cell populations on power to identify novel eGenes and cell-type-specific eQTLs, we conducted eQTL analyses using the GTEx tissue (skin), which includes the largest number of RNA-seq samples (Figure 3A). We deconvoluted 749 skin RNA-seq samples from 510 distinct individuals using signature genes from mouse skin scRNA-seq (5 cell types; Figure 3A). We found that only 4 of the 5 mouse skin populations were detected in GTEx human liver (Figure 3B), which may be due to differences in depth of skin biopsies between GTEx and Tabula Muris (basal cells of the epidermis were absent). We identified 24,029 expressed genes in the 749 skin RNA-seq samples and performed two eQTL analysis: 1) without using cell population distributions (bulk resolution) (Table S23); and 2) using deconvoluted cell populations (skin deconvoluted from mouse; 4 identified cell types) (Table S24). Using cell population distributions as covariates, we detected a 53% increase in genes with significant eQTLs (14,174 compared with 9,232, Figure 4A). We observed that eGenes specific for the eQTL analysis performed using cell populations as covariates had eQTLs in fewer tissues than eGenes detected in at bulk resolution ($p = 5.54 \times 10^{-256}$, Mann Whitney U test; Figure 4B), had a decreased effect size β ($p = 3.61 \times 10^{-93}$, Mann Whitney U test; Figure 4C), and had decreased standard error (SE) of β ($p = 1.71 \times 10^{-53}$, Mann Whitney U test; Figure 4D). We also observed that the β values for the top hit for each eGene were highly correlated between eQTLs detected using cell populations and eQTLs detected without using cell populations ($r = 0.989$, Figure S1D). Further, we detected 417 cell-type-associated eGenes (FDR-corrected p-values < 0.1, χ^2 test, Figure 4H) and 241 cell-type-specific eGenes (FDR-corrected p-values < 0.1, χ^2 test, Figure 4H). The relatively large number of cell-type-associated eGenes compared with the liver could be reflective of sample size differences between skin and liver (749 and 153, respectively) impacting power to detect eGenes. These results show that even in eQTL studies using large sample sizes, accounting for cellular heterogeneity results in the detection of thousands more eGenes, which tend to show cell-type-specific differential regulation.

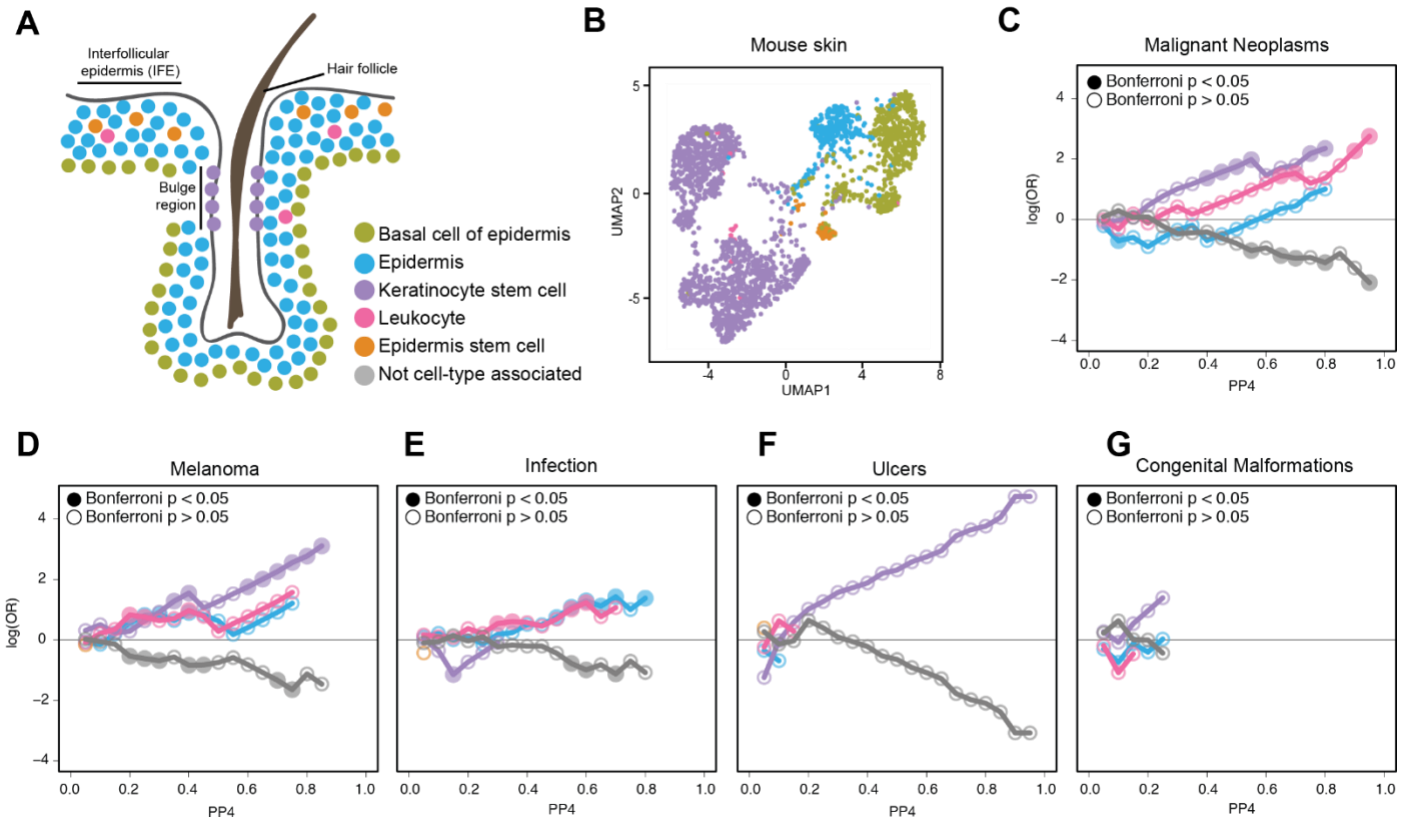
Colocalization identifies cell-type-specific regulatory variants are associated with specific skin diseases

To explore the functional impact of the cell-type-specific eQTLs identified in skin, we examined their overlap with GWAS lead variants for skin traits and disease. From the UK Biobank, we extracted GWAS summary statistics for 23 skin traits where the cell types identified from skin scRNA-seq (Figure 5A,B) likely played a role in the traits (Table 25) and grouped them into seven categories based on trait similarity: 1) malignant neoplasms, 2) melanomas, 3) infections, 4) ulcers, 5) congenital defects, 6) cancer (broad definition, non-malignant neoplasm), and 7) unspecified skin conditions. We performed colocalization to identify skin eQTLs and skin GWAS loci that share common causal variants using coloc[15] and examining instances with PP4 > 0.5 (PP4, posterior probability of the colocalization model having one shared causal variant). We identified 473 variants that showed evidence of colocalization (Table 25). These results show that we could identify hundreds of skin eQTLs that likely share a causal variant with skin GWAS.

We next asked if skin GWAS traits were enriched for eQTLs that are associated with distinct cell types. We tested the enrichment of cell-type-associated eQTLs at multiple PP4 thresholds and found malignant neoplasms were enriched for eQTLs associated with leukocytes ($p = 1.92 \times 10^{-4}$ Fisher's Test; Figure 5C) and keratinocyte stem cells ($p = 7.30 \times 10^{-4}$), melanomas were only enriched for eQTLs associated with keratinocyte stem cells ($p = 4.91 \times 10^{-6}$ Fisher's Test; Figure 5D), and infections were only enriched for eQTLs associated with epidermis cells ($p = 1.3 \times 10^{-3}$ Fisher's Test; Figure 5E). The non-significant but strong signal in ulcers for association with keratinocyte stem cells ($p = 0.25$ Fisher's Test; Figure 5F), was due to a single cell-type-associated eQTL (antisense gene: RP11-524D16__A.3) sharing a signal with one ulcer GWAS locus. We did not observe an enrichment of cell-type-associated eQTLs in congenital malformations (Figure 5G), cancer (broad definition), or unspecified skin conditions. It is unclear if this is to be expected, as it is possible other cell types not estimated may be contributing to the diseases or in the case of congenital malformations, it is possible that expression differences impacting congenital malformations may be functioning during development and not detectable in adult skin. Overall, these results suggest that GWAS lead variants are commonly cell-type-associated regulatory variants, indicating that onset or progression of human disease and traits may be controlled at the cell type level.

As the immune system has been implicated in playing a role in skin cancer[16-18], which includes the various skin cancer types delineated by malignant neoplasms, we next sought to examine the six eGenes with eQTLs significantly associated with leukocytes and colocalized with malignant neoplasms to gain insight into the potential roles they play in disease. Among these eGenes was *TCF19* (PP4 = 0.98), previously implicated in increasing cell proliferation in carcinomas[19], *ATAD3C* (PP4=0.73), previously shown to influence the Fanconi anemia DNA repair pathway and often malfunctioning in human cancers[20], and *SERPINB9* (PP4=0.16), which disruption of the serpinb9 protein in circulating T cells was previously shown to increase risk of skin cancer post-kidney transplantation[21]. Of the remaining eGenes, two (*NT5C2*, and *CD1E*) also have also been found to play a role in cancer progression or immune response[22, 23], supporting our ability to identify cell-type associated eQTLs whose functions are congruent with playing a role in the etiology of malignant neoplasms of the skin. Defects in the final eGene, *ZNF408*, have been linked to diseases of the retina (familial exudative vitreoretinopathy and retinis pimentosa)[24, 25], however it is unclear what role it may play in malignant neoplasms. Together these results show that conducting eQTL studies accounting for cellular heterogeneity can identify the likely causal cell-type associated variants and genes underlying GWAS disease loci.

Figure 5: Colocalization of cell-type-specific skin eQTLs with skin GWAS traits



A. Cartoon of describing approximate organization of cell types identified in scRNA-seq from skin. Colors used for each cell type are used throughout Figure.

B. UMAP plots of clustered scRNA-seq data from mouse skin. Cells are colored following color coding of each cell type from 5A.

C-G. Line plots showing the enrichment of cell-type-associated eQTLs in various GWAS traits: malignant neoplasms (c), melanoma (D), infection (e), ulcers (f), and congenital malformations (g). Enrichment is plotted as the log(OR) (y-axis) over all probabilities of the eQTL signal overlapping (0 = not overlapping – 1 = completely overlapping) with the GWAS signal (x-axis). Lines are colored following color coding of each cell type from 5A.

Discussion

Genetic association studies performed by GTEx have identified a wealth of novel insights into how human genetics function across bulk tissues[1], however these analyses do not consider how cellular heterogeneity can confound these studies through biasing or even masking cell-type-specific signals. Therefore, we sought to deconvolute the cellular composition of GTEx tissues using the mouse scRNA-seq Tabula Muris compendium[10] and to perform eQTL analyses considering cellular heterogeneity as a covariate. Using scRNA-seq from 14 mouse tissue types, we deconvoluted over 6,000 GTEx RNA-seq samples mapping to 28 tissues from 14 organs. We found that considering cellular heterogeneity significantly improved eQTL analyses by increasing power to detect eGenes, as well as by identifying cell-type-specific associations that were masked in an analysis using bulk RNA-seq from the same samples. We further show that cell-type-associated eQTLs colocalize with lead variants from relevant GWAS traits, highlighting a potential path forward for understanding the impact of genetic variation on mechanisms underlying complex traits.

Human scRNA-seq data representative of all tissues in GTEx that can be used to deconvolute more than 10,000 GTEx bulk RNA-seq samples does not yet exist. As the Tabula Muris resource of mouse scRNA-seq from 20 organs was recently released, we sought to determine if mouse scRNA-seq could be used as an alternative for human scRNA-seq for cellular deconvolution by comparing the cellular composition estimates derived from using scRNA-seq from human versus mouse. We established that mouse scRNA-seq was a suitable alternative to human scRNA-seq for estimating cellular heterogeneity in GTEx tissues. In general human and mouse estimates were comparable and discrepancies in cell composition estimates between the species were a result of differences in cell type resolution. As Tabula Muris does not represent all of the GTEx tissues, we were only able to deconvolute a subset of GTEx RNA-seqs (28 of the 53 tissues). However, as we show mouse scRNA-seq can estimate cellular composition comparable to human scRNA-seq, it is possible other scRNA-seq resources from mouse and other mammalian species could be used to deconvolute all of GTEx.

Considering cellular heterogeneity estimated from deconvolution of GTEx liver using high resolution scRNA-seq identifies substantially more eQTLs than from using lower resolutions (low resolution or collapsed resolution); however high resolution cell estimates identify fewer cell-type-associated genetic associations than lower resolutions. It is possible this decrease in associations may be due to dilution of signal between the similar cell types, which indicates varying benefits between resolutions for discovering more eQTLs versus classifying cell-type-specific eQTLs. Overall, this emphasizes that while efforts to generate a resource of scRNA-seq data from human tissues[26] are in progress, studies performing genetic association analyses from human data should utilize already existing scRNA-seq from mouse and other species comparable to human to estimate cellular heterogeneity to optimize power.

Taken together, these data describe a novel approach to obtain cell-type-specific genetic associations by using mouse scRNA-seq to deconvolute bulk human RNA-seq. The framework we propose to deconvolute the cellular composition of bulk RNA-seq from GTEx opens the door to the wealth of publicly available bulk RNA-seq samples that already exist and can be reanalyzed considering their heterogeneity. Our results further emphasize that this straightforward approach has the potential to greatly expand our understanding of the functional impact of genetic variation on complex traits and disease.

Methods

Processing of scRNA-seq from human liver

10X Genomics formatted BAM files from five human liver samples were downloaded (GEO accession: GSE11546) and converted to fastq files using 10X bamtofastq (https://support.10xgenomics.com/docs/bamtofastq). Converted fastq files were then processed using cellranger count utility to generate gene expression count matrices, then the five processed liver samples were merged using cellranger aggr utility.

Mouse single cell transcriptome profiles from 14 mouse organs from Tabula Muris

Single cell transcriptome profiles from 14 organs were used in this study[10]. Briefly, transcriptome profiles were generated from three female and four male mice (C57BL/6JN; 10-15 month-old) from: aorta, atrium, bladder, brain nonmicroglia, colon, fat, kidney, liver, mammary gland, muscle, pancreas, skin, spleen, ventricle (Table S1). Upon extraction of these organs from the mice, single cell transcriptomes were generated by first sorting by fluorescence-activated cell sorting (FACS) (FACS method; SMART-Seq2 RNAseq libraries). We downloaded the normalized gene expression and annotated single-cell clusters from each organ as Seruat[11] Robjects (https://figshare.com/articles/Robject_files_for_tissues_processed_by_Seurat/5821263/1).

Annotation of the cell populations present in human liver scRNA-seq data

Characterization of cell type composition of scRNA-seq from human liver[5] were analyzed following the same approach used to annotate mouse organs[10]. From both tissue types, cells with fewer than 500 detected genes or cells with fewer than 1,000 UMI were filtered from the data, resulting in 8,119 cells analyzed from human liver. Gene expression was then log normalized and variable genes were identified using a threshold of 0.5 for the standardized log dispersion. Principal component analysis (PCA) was performed on the variable genes and significant PCs were identified by visual inspection of the elbow the standard deviations of the PCs observed

through a Scree plot. Clustering was then performed using a shared-nearest-neighbor graph of the significant PCs. Single cells were then visualized using Uniform Manifold Approximation and Projection (UMAP) and cellular subtypes were identified by observing the relative abundance of known liver marker genes[10].

Collapsing liver cell population estimates

Cell population estimates in GTEx liver were merged based on similar cell populations. To collapse similar cell populations, we examined the UMAP from high resolution human liver scRNA-seq (Figure 1B) and compared to the UMAP from low resolution mouse liver scRNA-seq (Figure 1C) to identify broader/lower resolution classifications of cell types present in the liver (Table 3). We identified populations in the human liver scRNA-seq that were similar (e.g. Hepatocyte populations 0, 1, 3, and 4; Figure 1B) with a corresponding population in the mouse liver scRNA-seq (e.g. Hepatocyte; Figure 1C). For populations identified in human not present in mouse, we did not perform any collapsing.

Deconvolution of complex tissues using CIBERSORT

Identification of signature genes from single cell populations: For scRNA-seq from human liver and scRNA-seq from 14 mouse organs, we obtained signature gene profiles for each cell type identified from scRNA-seq (Table S2) as input into CIBERSORT[2] to estimate the cellular composition of GTEx adult tissues (Table S1). To obtain these signature gene profiles, we first identified differentially expression genes from each scRNA-seq cell population within a given tissue using Seurat FindMarkers. Of these differentially expressed genes for each cell type, we extracted the top 200 most significantly overexpressed genes (adjusted p-value < 0.05; average $\log_2$ fold change > 0.25). For signature genes obtained from mouse scRNA-seq, we converted the mouse genes to their human orthologs using the BioMart database[27, 28]. The final gene signature sets only included mouse signature genes that also had a human ortholog. For cases that a mouse gene had more than one human ortholog for a given cells type, only one human ortholog was retained in final signature set. For cases that different

mouse genes corresponded to the same human ortholog for a given cell type, only unique human orthologs were retained in the final signature set.

Cell composition estimation: The mean expression levels of the top 200 genes overexpressed in each of the cell type identified in the scRNA-seq from various human and mouse organs were used as input for CIBERSORT[2] to calculate the relative distribution of the cell populations of 28 GTEx tissues from 14 organs (Table S4-18). CIBERSORT (https://cibersort.stanford.edu/) was run with default parameters using the TPM values for the signature genes identified from scRNA-seq in all RNA-seq samples from the analogous GTEx tissue (https://gtexportal.org/home/datasets) (Table S1). GTEx tissues are defined by the distinct area of the organ where the tissue was taken (variable name SMTSD from sample attributes data table; phv00169241.v7.p2) and organs are defined as the regions where the tissues are sampled from (variable name SMTS from sample attributes data table; phv00169239.v7.p2).

eQTL analysis

To detect eQTLs, we obtained TPM for 153 liver samples and 749 skin samples (sun-exposed and not sun-exposed) from the GTEx V. 7 website (https://gtexportal.org/home/) and downloaded WGS VCF files from dbGaP (525 individuals, phs000424.v7.p2). Only genes with TPM > 0.5 in at least 20% samples were considered (19,621 in liver and 24,029 in skin). Gene expression data was quantile-normalized independently for each tissue type. For all eQTL analyses, we used the following covariates: age, sex and the first five genotype principal components (PCs) calculated using 90,081 SNPs in linkage equilibrium[29]. We fitted different linear mixed models (LMMs) using the lme4 package (https://www.jstatsoft.org/article/view/v067i01/0)) to detect eQTLs in liver and skin. skin, we used the following model[14]:

$$\text{Expression} \sim \text{genotype} + \text{covariates} + (1|\text{subject\_id})$$

Where (1|subject_id) denotes subject-specific random effects. We used the subject-specific random effect for skin because several individuals had two samples. For liver, we used sex as random effect to fit an LMM using a method analogous to skin eQTL analysis:

$$\text{Expression} \sim \text{genotype} + \text{covariates} + (1|\text{sex})$$

Where (1|sex) denotes sex-specific random effects. We calculated associations with all variants (minor allele frequency > 1%) ± 1 Mb around each expressed gene. For each gene, we Bonferroni-corrected p-values and retained the lead variant. To detect eGenes, we used Benjamini-Hochberg FDR at 10% level on all lead variants.

Using cell population distributions to improve eQTL detection

We repeated eQTL detection using LMMs with cellular compositions as covariates. Since several cell types were detected at very low frequency, we only used a subset of the cell types described in Figure 3. Specifically, we detected liver eQTL using human (high resolution and collapsed) and mouse (low resolution) cell populations as covariates. We used the following cell populations: 1) for human high resolution: periportal sinusoidal endothelial cells, central venous endothelial cells, gdT cells, hepatocytes0, hepatocytes3, hepatocytes4, inflammatory macrophages and NK/NKT cells; 2) for human collapsed resolution: endothelial cells, hepatocytes, macrophages, NK cells, B cells, cholangiocytes, and heptatic stellate cells; and 3) for mouse low resolution: endothelial cells of hepatic sinusoid, hepatocytes, Kupffer cells and NK cells. For skin, we used the following cell populations: epidermal cells, epidermal stem cells, keratinocyte stem cells and leucocytes. For each cell population, we compared the following two models:

$$H_0\text{: expression} \sim \text{genotype} + \text{covariates} + \text{cell\_populations} + (1|\text{random})$$

$$H_1\text{: expression} \sim \text{genotype} + \text{covariates} + \text{cell\_populations} + \text{genotype:cell\_population} + (1|\text{random})$$

Where (1|random) denotes each tissue's random effect. We next calculated the difference between the two models using ANOVA and obtained χ^2 p-values using the pbkrtest package (https://www.jstatsoft.org/article/view/v059i09). For each eGene, we compared each cell population to H_0 and

retained only the most significant association. Only eGenes with χ^2 p-values adjusted for Benjamini-Hochberg FDR <0.1 and Akaike's information criterion (AIC) < 0 were labeled as cell population-specific.

Colocalization of UK Biobank GWAS for skin traits and eQTLs identified from skin

For each eGene in the skin eQTL analysis deconvoluted using cell population estimates, we extracted the p-values for all variants that were used to perform the eQTL analysis. From the UK BioBank, we obtained summary statistics for 23 skin-related traits (Table S25). For all the variants genotyped in both GTEx and UK BioBank, we used coloc V. 3.1[15] to test for colocalization between eQTLs and GWAS signal. For each colocalization test, we considered only the posterior probability of a model with one common causal variant (PP4) and tested the enrichment of cell-type-specific associations at multiple thresholds using Fisher's exact test.

Acknowledgements

This work was supported in part by a California Institute for Regenerative Medicine (CIRM) grant GC1R-06673 and NIH grants HG008118-01, HL107442-05, DK105541-03, and DK112155-01. M.K.R.D. was supported by the National Library of Medicine Training Grant T15LM011271.

Author information

K.A.F., M.K.R.D., M.D. conceived the study. M.K.R.D and M.D. performed computational analysis. M.K.R.D. performed scRNA-seq data processing and deconvolution analyses. M.D. performed the eQTL analysis. M.K.R.D and M.D. performed colocalization analysis. K.A.F. oversaw the study. M.K.R.D., M.D., and K.A.F. prepared the manuscript.

References

1. Consortium, G.T. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213 (2017).
2. Newman, A.M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* **12**, 453-457 (2015).
3. Zhong, Y., Wan, Y.W., Pang, K., Chow, L.M. & Liu, Z. Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinformatics* **14**, 89 (2013).
4. Baron, M. et al. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst* **3**, 346-360 e344 (2016).
5. MacParland, S.A. et al. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat Commun* **9**, 4383 (2018).
6. Cheng, J.B. et al. Transcriptional Programming of Normal and Inflamed Human Epidermis at Single-Cell Resolution. *Cell Rep* **25**, 871-883 (2018).
7. Crinier, A. et al. High-Dimensional Single-Cell Analysis Identifies Organ-Specific Signatures and Conserved NK Cell Subsets in Humans and Mice. *Immunity* **49**, 971-986 e975 (2018).
8. Young, M.D. et al. Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science* **361**, 594-599 (2018).
9. Nguyen, Q.H. et al. Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. *Nat Commun* **9**, 2028 (2018).
10. Tabula Muris, C. et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367-372 (2018).
11. 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).
12. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat Rev Genet* **14**, 618-630 (2013).
13. Hagai, T. et al. Gene expression variability across cells and species shapes innate immunity. *Nature* **563**, 197-202 (2018).
14. Alasoo, K. et al. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat Genet* **50**, 424-431 (2018).
15. Giambartolomei, C. et al. A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics* **34**, 2538-2545 (2018).
16. Yu, S.H., Bordeaux, J.S. & Baron, E.D. The immune system and skin cancer. *Adv Exp Med Biol* **810**, 182-191 (2014).
17. Rangwala, S. & Tsai, K.Y. Roles of the immune system in skin cancer. *Br J Dermatol* **165**, 953-965 (2011).
18. Thorsson, V. et al. The Immune Landscape of Cancer. *Immunity* **48**, 812-830 e814 (2018).
19. Zeng, C.X. et al. TCF19 enhances cell proliferation in hepatocellular carcinoma by activating the ATK/FOXO1 signaling pathway. *Neoplasma* **66**, 46-53 (2019).
20. Zhang, T. et al. Fancd2 in vivo interaction network reveals a non-canonical role in mitochondrial function. *Sci Rep* **7**, 45626 (2017).
21. Peters, F.S. et al. Disrupted regulation of serpinB9 in circulating T cells is associated with an increased risk for post-transplant skin cancer. *Clin Exp Immunol* (2019).
22. Tzoneva, G. et al. Activating mutations in the NT5C2 nucleotidase gene drive chemotherapy resistance in relapsed ALL. *Nat Med* **19**, 368-371 (2013).
23. Facciotti, F. et al. Fine tuning by human CD1e of lipid-specific immune responses. *Proc Natl Acad Sci U S A* **108**, 14228-14233 (2011).
24. Habibi, I., Chebil, A., Kort, F., Schorderet, D.F. & El Matri, L. Exome sequencing confirms ZNF408 mutations as a cause of familial retinitis pigmentosa. *Ophthalmic Genet* **38**, 494-497 (2017).

25. Avila-Fernandez, A. et al. Whole-exome sequencing reveals ZNF408 as a new gene associated with autosomal recessive retinitis pigmentosa with vitreal alterations. *Hum Mol Genet* **24**, 4037-4048 (2015).
26. Regev, A. et al. The Human Cell Atlas. *Elife* **6** (2017).
27. Durinck, S., Spellman, P.T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* **4**, 1184-1191 (2009).
28. Durinck, S. et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439-3440 (2005).
29. Panopoulos, A.D. et al. iPSCORE: A Resource of 222 iPSC Lines Enabling Functional Characterization of Genetic Variation across a Variety of Cell Types. *Stem Cell Reports* **8**, 1086-1100 (2017).