UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Understanding the cellular heterogeneity in fetal-like and adult tissues to study cell-type-specific functional genetic variation

Permalink

https://escholarship.org/uc/item/9jj548mp

Author

Donovan, Margaret Kathleen Rose

Publication Date

2019

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA SAN DIEGO

Understanding the cellular heterogeneity of fetal-like and adult tissues to study cell-type-specific functional genetic variation

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Bioinformatics and Systems Biology with a Specialization in Biomedical Informatics

by

Margaret Kathleen Rose Donovan

Committee in charge:

Professor Kelly A. Frazer, Chair
Professor Hannah Carter
Professor Olivier Harismendy
Professor Nathan Lewis
Professor Lucila Ohno-Machado

2019

Copyright

Margaret Kathleen Rose Donovan, 2019

All rights reserved

The Dissertation of Margaret Kathleen Rose Donovan is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2019

DEDICATION

To my parents, Charlene and Richard, my sister, Virginia, my brother, Daniel, and my fiancé, Amin. *1, 2, 3 achieve.*

EPIGRAPH

Great things are done by a series of small things brought together.

Georges Seurat

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my dissertation advisor, Dr. Kelly Frazer, as well as my mentors Dr. Matteo D'Antonio, Dr. Agnienzska Chronowska-D'Antonio, Dr. Paola Benaglio, and Dr. Erin Smith for their support throughout my Ph.D. It was their guidance that brought me through the stages of, as I say, baby-scientist, to scientist-shaped, to scientist presenting my work on the international stage. I also thank my Dissertation committee for their valuable input and guidance in my research. Also, thank you to Dr. Amirali Kia, for without his guidance early in my career, I would not be the scientist I am today (also for telling me to mark my calendar). Further, I would like to acknowledge the entire Frazer Lab, current and past, for their unyielding encouragement and friendship.

Next, I would like to thank my family. Mom and Dad, for their love, support, and igniting my passion in science—beginning over their breakfast chats mentioning sequencing, back when "sequence" to me meant the glittery beads I used for art. Virginia, Daniel, and Zev, for paving the way as big siblings must and creating the stepping-stones I've been lucky enough to follow. Sima, Mojtaba, Neda, and the entire Ronaghi family for reminding me to relax and for cheering me on every step of the way (Merci!). Amin, for inspiring constant curiosity in the world around me (one YouTube video, article, and meme at a time) and for being my pillar, my sounding board, and my partner through this journey.

Chapter 1, in full, is an adapted reprint of the material as it appears in Stem Cell Reports, 2019, Margaret K.R. Donovan, Agnieszka D'Antonio-Chronowska, William W. Greenwald, Jennifer Phuong Nguyen, Kyohei Fujita, Sherin Hashem, Hiroko Matsui, Francesca Soncin, Mana Parast, Michelle C. Ward, Florence Coulet, Erin N. Smith, Eric

Adler, Matteo D'Antonio, Kelly A. Frazer. The dissertation author was one of the primary investigators and authors of this paper.

Chapter 2, in full, has been submitted for publication of the material as it may appear in Nature Communications, 2019, Margaret K.R. Donovan, Agnieszka D'Antonio-Chronowska, Matteo D'Antonio, Kelly A. Frazer. The dissertation author was the primary investigator and author of this paper.

VITA

2009-2013	Bachelor of Science, Bioengineering, Minor, Bioinformatics, University of California, Santa Cruz
2015-2019	Doctor of Philosophy, Bioinformatics and Systems Biology with a Specialization in Biomedical Informatics, University of California San Diego

PUBLICATIONS

First Author

Association of human iPSC gene signatures and X chromosome dosage with two distinct cardiac differentiation trajectories. *Stem Cell Reports.* 2019

Cellular Deconvolution of GTEx tissues powers eQTL studies to discover thousands of novel disease and cell-type associated regulatory variants. *In Review: Nature Communications.* 2019

Pancreatic progenitor differentiation occurs asynchronously and generates a continuum of pancreatic-lineage cell fates. *In Prep: Stem Cell Reports.* 2019

Other Authorship

Allele-specific NKX2-5 binding underlies multiple genetic associations with human EKG traits. *Nature Genetics*. 2019.

Genomic properties of structural variants and short tandem repeats that impact gene expression and complex traits in humans. *In Review: Nature Communications*. 2019

In-depth genetic analysis of 6p21.3 reveals insights into associations between HLA types and complex traits and disease. *eLife*. 2019

Co-regulated paralogs enable deleterious null alleles to exist at high frequencies in humans. *In Review: Nature Communications*. 2019

Insights into the Mutational Burden of Human Induced Pluripotent Stem Cells from an Integrative Multi-Omics Approach. *Cell Reports*. 2018

iPSCORE: A resource of 222 iPSC lines enabling functional characterization of genetic variation across a variety of cell types. *Stem Cell Reports.* 2017

ABSTRACT OF THE DISSERTATION

Understanding the cellular heterogeneity of fetal-like and adult tissues to study cell-type-specific functional genetic variation

by

Margaret Kathleen Rose Donovan

Doctor of Philosophy in Bioinformatics and Systems Biology with a Specialization in Biomedical Informatics

University of California San Diego, 2019

Professor Kelly A. Frazer, Chair

Genome-wide association studies (GWAS) have suggested that the underlying genetic basis of complex traits and disease is driven by large numbers of non-coding variants with modest effects that likely act by modifying gene regulation. Towards understanding the regulatory impact of genetic variation, expression quantitative trait loci

(eQTL) analyses have been performed across dozens of human tissues to link the influence of genetic variants on gene expression levels. While these eQTL studies have provided important biological insights, they are still limited by not considering the contexts in which these variants function, including the stage of development and cell type. Specifically, others have shown increased disease risk in adulthood has links to fetal origins, suggesting that characterizing gene expression in fetal-like cells could identify genetic variants that are associated with adult traits, but function primarily or solely during development. Additionally, as eQTL studies are typically performed across bulk tissues, unaccounted for cellular heterogeneity present in bulk gene expression measurements can affect genotype-gene expression associations. Thus, it is important to identify regulatory variants that alter gene expression in both primitive and adult cell types and to characterize cellular heterogeneity across tissues to comprehensively understand the genetic basis of complex traits and disease.

Here, I present two studies, which utilize gene expression data from fetal-like and adult tissues to characterize cellular heterogeneity at distinct stages of human development. I have examined the cellular heterogeneity in fetal-like induced pluripotent stem cell (iPSC)-derived cardiovascular progenitor cells (CVPCs) using single cell (sc)RNA-seq data to identify cell populations that emerge as a result of the cardiac differentiation. Further, I deconvoluted 180 iPSC-CVPCs and identified factors innate to iPSCs that impacted cardiac fate. Next, I showed that mouse scRNA-seq can be used as an alternative to human scRNA-seq for the deconvolution of adult GTEx bulk tissues and considering cell composition eQTL studies powered the discovery of novel eQTLs, some of which were cell-type-associated and colocalized with GWAS disease loci.

Chapter 1: Association of human iPSC gene signatures and X chromosome dosage with two distinct cardiac differentiation trajectories

1.1 Abstract

Despite the importance of understanding how variability due to non-genetic factors (clone and passage) influences iPSC differentiation outcome, large-scale studies capable of addressing this question have not yet been conducted. Here, we performed 232 directed differentiations of 191 iPSC lines to generate iPSC-derived cardiovascular progenitor cells (iPSC-CVPCs). We observed cellular heterogeneity across the iPSC-CVPC samples due to varying fractions of two cell types: cardiomyocytes (CMs) and epicardium-derived cells (EPDCs). Comparing the transcriptomes of CM-fated and EPDC-fated iPSCs, we discovered that 91 signature genes and X chromosome dosage differences are associated with these two distinct cardiac developmental trajectories. In an independent set of 39 iPSCs differentiated into CMs, we confirmed that sex and transcriptional differences impact cardiac fate outcome. Our study provides novel insights into how iPSC transcriptional and X chromosome gene dosage differences influence their response to differentiation stimuli and hence cardiac cell fate.

1.2 Introduction

Variability in human induced pluripotent stem cell (iPSC) lines compromises their utility for regenerative medicine and as a model system for genetic studies. This variability impacts iPSC differentiation outcome and despite using standardized

differentiation protocols, results in the generation of samples with cellular heterogeneity (i.e. multiple cell types are present within a given sample and the proportions of cell types vary across samples). Previous large-scale quantitative trait loci (QTL) studies in iPSCs [1,2] have shown that genetic variation accounts for the majority of expression differences between iPSC lines, but non-genetic (i.e., clonality and passage) factors also contribute to these differences [3]. Understanding how non-genetic transcriptional differences between iPSC lines impact their differentiation outcome is necessary to improve the ability to generate cell types of interest.

Well-established small molecule protocols for generating iPSC-derived cardiovascular progenitor cells (iPSC-CVPCs) [4] produce fetal-like cardiomyocytes, which can undergo further specification as cells mature in culture into various cardiac subtypes (atrial, ventricular, or nodal) [5]. Based on variable cardiac troponin T (cTnT) staining, the derived samples are known to display cellular heterogeneity [6,7], but the origin of the cTnT-negative non-myocyte cells, and whether the same or different non-myocyte cell types are consistently derived alongside cTnT-positive myocytes across samples has not previously been investigated. The differentiation protocol is dependent on manipulation of WNT signaling, initially through activation of the pathway by GSK3 inhibition, followed by inhibition of the pathway by Porcupine (*PORCN*) inhibition [8,9]. An in-depth analysis of the outcomes of independent differentiations of hundreds of iPSC lines with different genetic backgrounds could provide insights into the origins of the non-myocyte cells, as well as the extent to which non-genetic transcriptional differences between iPSC lines contribute to the iPSC-CVPC cellular heterogeneity.

Here, we used a highly standardized and systematic approach to conduct 232 directed differentiations of 191 iPSC lines into iPSC-CVPCs. We characterized the cellular heterogeneity of the iPSC-CVPC samples and showed that only two distinct cell types were present, cardiomyocytes (CMs) and epicardium-derived cells (EPDCs), which varied in proportion across samples. As differentiation protocols to derive iPSC-CMs and iPSC-EPDCs primarily differ by a step involving WNT inhibition to derive the former, but not the latter [10,11], we hypothesized that the observed cellular heterogeneity could result from suboptimal WNT inhibition in subsets of cells across iPSC lines. To test this hypothesis, we analyzed transcriptional differences between iPSC lines that differentiated into CMs and those that differentiated into EPDCs (e.g. iPSCs with a CM-fate or EPDC-fate) and discovered 91 signature genes associated with these two distinct cardiac differentiation trajectories. These signature genes are involved in differentiation, including the Wnt/β-catenin pathway, muscle differentiation or cardiac-related functions, and the transition of epicardial cells to EPDCs by epithelial-mesenchymal transition (EMT). While the proportion of variance explained by each of the signature genes varied over three orders of magnitude, altogether they captured approximately half of the total variance underlying iPSC fate determination. Additionally, we show variability in X chromosome gene dosage ($X_{active}X_{active}$ vs $X_{active}X_{inactive}$ vs XY) across iPSCs plays a role in cardiac fate determination. The association with X chromosome gene dosage could in part be due to higher expression in CM-fated iPSCs of chrXp11 genes, which encodes *ELK1* and *PORCN*. Transcriptomic analysis of an independent set of 39 iPSCs differentiated to the cardiac lineage using a similar small molecule protocol [12] confirmed our findings.

1.3 Results

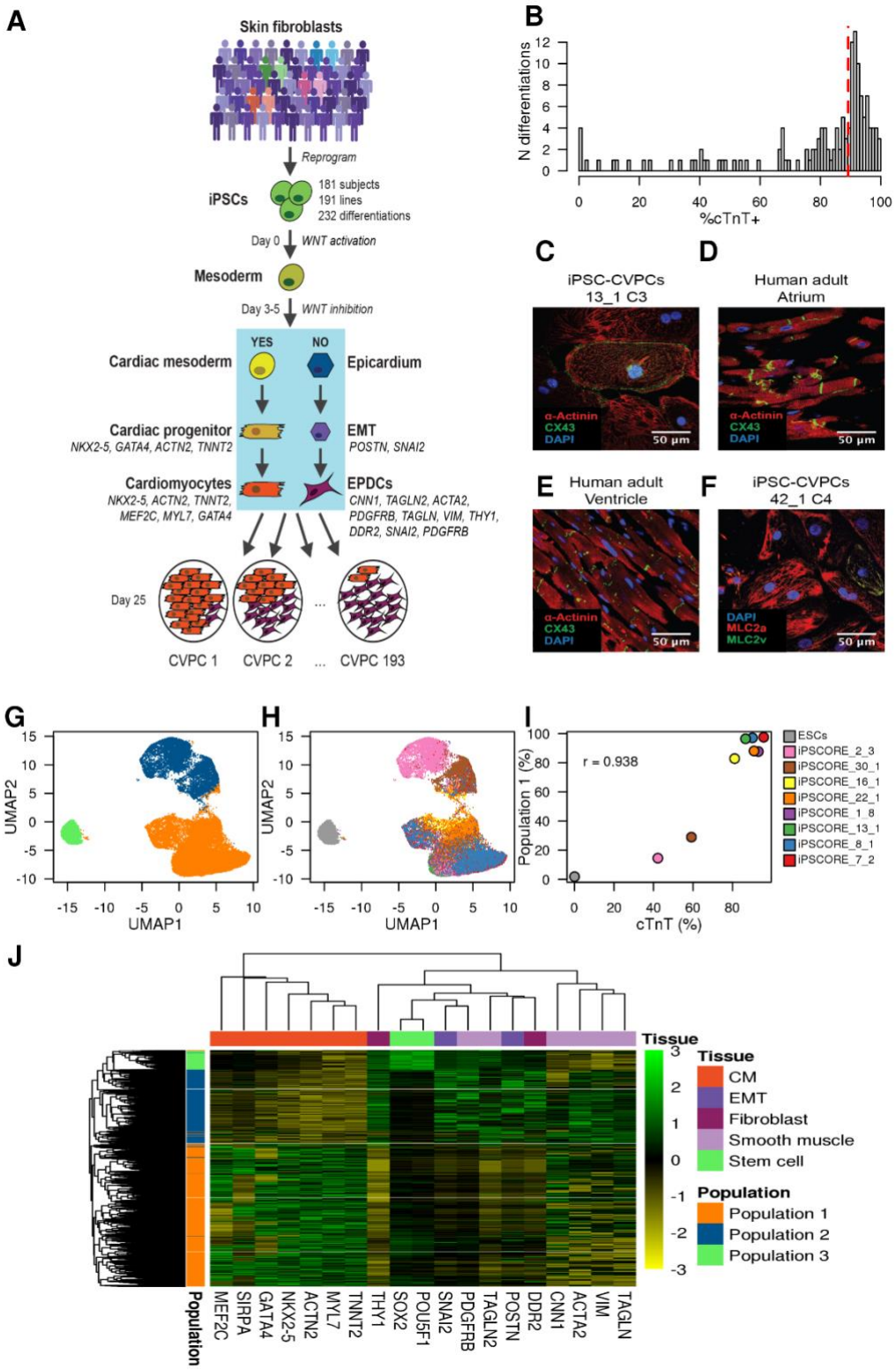
1.3.1 iPSC-CVPCs show cellular heterogeneity across samples

To gain insights into molecular mechanisms that could influence variability in human iPSC differentiation outcome, we employed a highly systematic approach to differentiate 191 pluripotent lines from 181 iPSCORE individuals (Figure 1.1a) into iPSC-derived cardiovascular progenitor cells (iPSC-CVPCs). We used a small molecule cardiac differentiation protocol used to derive cardiomyocytes [13] followed on D15 by lactate selection to obtain pure cardiac cells [14]. In total, we conducted 232 differentiations, of which 193 (83.2%, from 154 lines derived from 144 subjects) were completed, i.e. reached Day 25 of differentiation, while 39 (from 37 lines derived from 37 subjects) were terminated prior to Day 25, because they did not form a syncytial beating monolayer (Table 1.1). The completed iPSC-CVPCs at D25 on average had a high fraction of cells that stained positive for cardiac troponin T (%cTnT, median = 89.2%; Figure 1.1b) and were positive by immunofluorescence (IF) for cardiac markers (Figures 1.1c-f); however, 15 lines had %cTnT < 40%, indicating that despite lactate selection, there was substantial cellular heterogeneity within and across samples.

Table 1.1 Table describing the number of lines and subjects for each attempted differentiation

Cell type	Number of differentiations	Number of lines	Number of subjects
iPSC	232	191	181
Terminated iPSC-CVPC	39	37	37
D25 iPSC-CVPC	193	154	144

Figure 1.1 Characterization of cellular heterogeneity in iPSC-CVPC samples
(A) Overview of the study design. Skin fibroblasts from 181 subjects were reprogrammed to iPSCs and differentiated to iPSC-CVPCs (191 lines, 232 differentiations). After WNT pathway activation at day 0 (D0) and its inactivation by IWP-2 at D3-5, cells differentiate to cardiomyocytes (CMs) if the WNT signaling is successfully inhibited. If WNT signaling is not sufficiently inhibited, cells differentiate to EPDCs. 193 of the 232 differentiations were completed (D25), and we observed that different CVPC samples had different proportions of CMs and EPDCs. (B) Distribution of %cTnT. Dashed redline represents the median value. (C-E) Immunofluorescence staining of (C) iPSC-CVPCs, (D) human atrium, and (E) ventricle with IF markers DAPI (blue), ACTN1 (red), and CX43 (green). (F) Immunofluorescence staining iPSC-CVPCs with IF markers DAPI (blue), MLC2a+ (red) and MLC2v+ (green), and MLC2v+ MLC2a+ (yellow). (G) scRNA-seq UMAP plot showing the presence of three populations: CMs (orange), EPDCs (blue) and ESCs (green). (H) scRNA-seq UMAP plot showing the distribution of the nine analyzed samples (8 iPSC-CVPC lines and one ESC line) across the three clusters. (I) Scatterplot showing the correlation between the %cTnT and the fraction of cells in Population 1 (CMs) for each of the nine samples. (J) Heatmap showing across all 34,905 single cells the expression markers for: 1) stem cells (*POU5F1*; *SOX2*; 2) CMs (*NKX2-5*, *ACTN2*, *TNNT2*, *MEF2C*, *MYL7*, *GATA4*); 3) EMT (*POSTN*, *SNAI2*); 4) fibroblasts (*DDR2*, *THY1*); and 5) smooth muscle (*PDGFRB, TAGLN2*, *CNN1*, *ACTA2*, *VIM*, *TAGLN*).


A
Skin fibroblasts
Reprogram
iPSCs
181 subjects
191 lines
232 differentiations
Day 0
WNT activation
Mesoderm
Day 3-5
WNT inhibition
YES
NO
Cardiac mesoderm
Epicardium
Cardiac progenitor
NKX2-5, GATA4, ACTN2, TNNT2
EMT
POSTN, SNAI2
Cardiomyocytes
NKX2-5, ACTN2, TNNT2, MEF2C, MYL7, GATA4
EPDCs
CNN1, TAGLN2, ACTA2, PDGFRB, TAGLN, VIM, THY1, DDR2, SNAI2, PDGFRB
Day 25
CVPC 1
CVPC 2
...
CVPC 193
B
N differentiations
%cTnT+
C
iPSC-CVPCs 13_1 C3
D
Human adult Atrium
α-Actinin
CX43
DAPI
50 μm
E
Human adult Ventricle
F
iPSC-CVPCs 42_1 C4
DAPI
MLC2a
MLC2v
G
H
UMAP1
UMAP2
I
r = 0.938
Population 1 (%)
cTnT (%)
ESCs
iPSCORE_2_3
iPSCORE_30_1
iPSCORE_16_1
iPSCORE_22_1
iPSCORE_1_8
iPSCORE_13_1
iPSCORE_8_1
iPSCORE_7_2
J
Tissue
CM
EMT
Fibroblast
Smooth muscle
Stem cell
Population
Population 1
Population 2
Population 3
MEF2C
SIRPA
GATA4
NKX2-5
ACTN2
MYL7
TNNT2
THY1
SOX2
POU5F1
SNAI2
PDGFRB
TAGLN2
POSTN
DDR2
CNN1
ACTA2
VIM
TAGLN

1.3.2 Subset of cells show differential response to WNT inhibition during differentiation

To examine the cellular heterogeneity in the iPSC-CVPCs, we performed single-cell RNA-seq (scRNA-seq) on eight samples with varying %cTnT values (42.2 to 95.8%) and combined these data with scRNA-seq from the H9 ESC line (total of 34,905 cells). We detected three distinct cell populations: 1) Population 1, 21,056 cells (60.3%); 2) Population 2, 11,044 cells (31.6%); and 3) Population 3, 2,805 cells (8.1%, Figure 1.1g). While Populations 1 and 2 were comprised of the eight iPSC-derived samples, Population 3 almost exclusively included ESC cells (97.7% of the 2,870 ESC cells, Figures 1.1h). The relative proportions of cells that each of the iPSC-CVPC samples contributed to Population 1 versus Population 2 was strongly correlated with its %cTnT value ($r = 0.938$, $p = 1.89 \times 10^{-4}$, t-test; Figure 1.1i), suggesting that Population 1 was cardiomyocytes.

As cardiomyocytes (CMs) and epicardium lineage cells could both survive lactate purification [14,15], we investigated if the non-myocyte cells composing Population 2 were iPSC-epicardium-derived cells (iPSC-EPDCs). We examined the expression levels of 17 marker genes (Figure 1.1j) specific for either CMs or EPDCs (including smooth muscle, fibroblasts, genes involved in EMT) and two marker genes for stem cells. Consistent with having a high number of cTnT-positive cells, Population 1 expressed high levels of CM-specific genes, while Population 2 expressed high levels of EPDC-specific genes, and Population 3 expressed high levels of the stem cell markers *POU5F1* and *SOX2* (Figure 1.2a,b). Of note, *TNNT2* was expressed in some of the cells in Population 2, which is consistent with the strong, but not absolute correlation between

%cTnT value and fraction of Population 1 (Figure 1.1i), and previous studies showing that some EPDCs express *TNNT2* [7]. These results show that the small molecule differentiation protocol followed by lactate purification resulted in the absence of undifferentiated cells at D25 and in the derivation of two distinct cell populations, one of which expresses high levels of CM markers, including *TNNT2*, *NKX2-5* and *MEF2C* (Population 1), and the other which expresses EPDC markers, including *SNAI2*, *DDR2*, *VIM* and *ACTA2* (Population 2). Of note, the protocols for generating iPSC-derived cardiomyocytes (iPSC-CMs) and iPSC-EPDCs both involve activating the WNT signaling pathway [10,15] and have a shared intermediate mesoderm progenitor, but subsequent WNT inhibition directs differentiating cells to iPSC-CMs and endogenous levels of WNT signaling direct differentiating cells to iPSC-EPDCs [7] (Figure 1.1a). Therefore, our results suggest that iPSC-CVPC cellular heterogeneity results from suboptimal WNT inhibition in a subset of cells during differentiation, which then give rise to EPDCs.

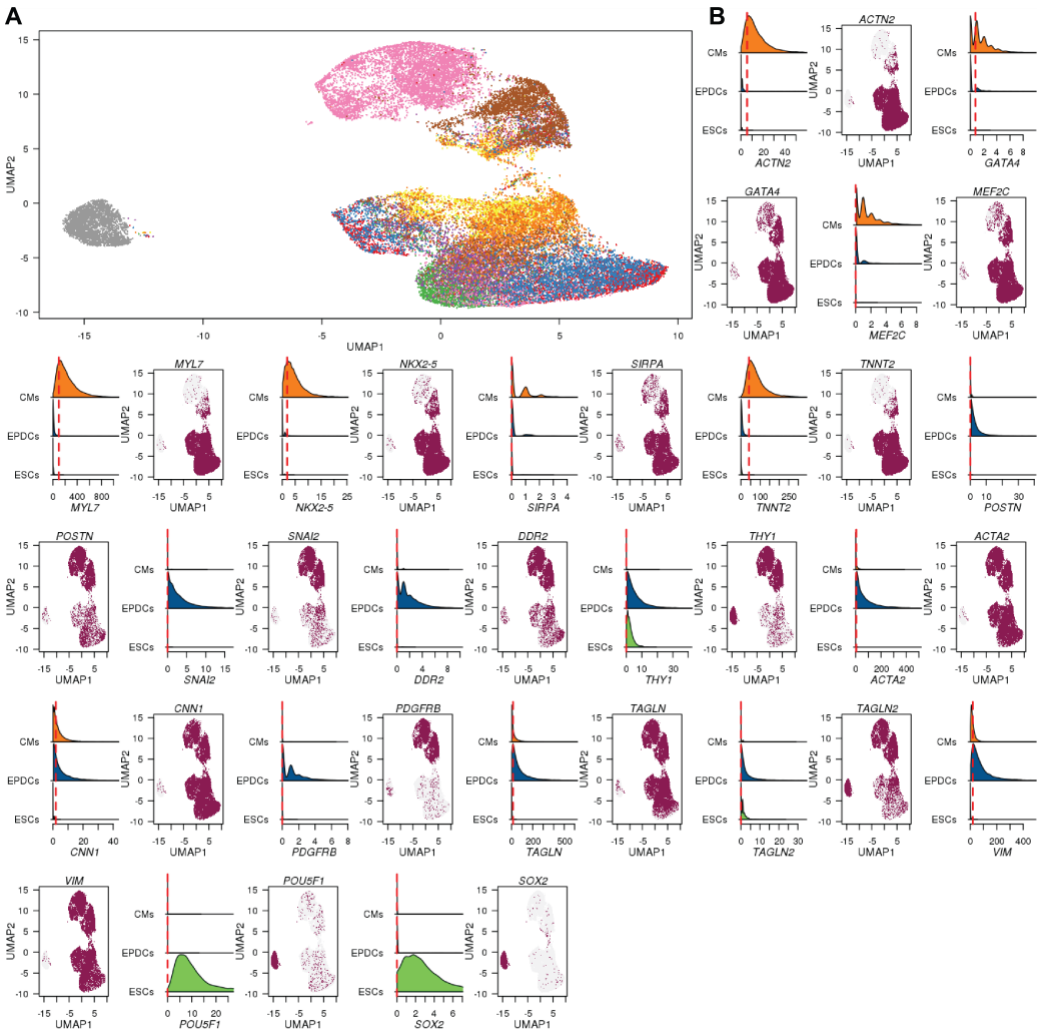


Figure 1.2 Distribution of single cells across three clusters
(a) Distribution of single cells across the three cell populations for the nine analyzed samples: scRNA-seq UMAP plots from 34,905 single cells showing their distributions across the three different clusters for the nine analyzed samples (8 iPSC-CVPCs lines and one ESC line). Each of the nine samples has a different color, as indicated in Figure 1.1i. (b) Expression levels for marker genes: For each gene in Figure 1.1j, density plots show the gene expression distribution across all cells associated with each cell population (Population 1 = orange; Population 2 = blue; Population 3 = green.). Red dashed line represents the median. UMAP plots from 34,905 cells show in maroon all the cells expressing the indicated marker gene higher than its median expression across the three populations.

1.3.3 iPSC-CVPCs are composed of immature CMs and EPDCs

To estimate the relative abundances of CM and EPDC cells across our collection of iPSC-CVPC samples, we selected the top 50 significantly overexpressed genes in each of the three scRNA-seq populations (150 genes in total, $p < 10^{-13}$, edgeR), obtained their expression levels in bulk RNA-seq from 180 iPSC-CVPCs, and inputted these values into CIBERSORT [16]. We observed that the proportions of each cell type varied across the samples, although the iPSC-CVPCs tended to have a greater fraction of CMs (84.8 ± 31.8%, Figure 1.3a) than EPDCs (14.7 ± 32.0%), and essentially no stem cells (0 ± 0.8%). Due to lactate selection, the small number (67) of cells predicted to be ESCs may represent a distinct differentiated cell type that is more similar to stem cells than either CMs or EPDCs. The estimated fraction of CMs and EPDCs in the iPSC-CVPCs was highly correlated with %cTnT values ($r = 0.927$, $p \approx 0$; t-test Figure 1.3b), similar to that observed in the analysis of the scRNA-seq data (Figure 1.1j). Finally, we showed that the iPSC-CVPCs with high estimated CM or EPDC cellular fractions respectively showed higher expression of CM markers (*MEF2C*, *NKX2-5* and *ACTN2*) and EPDC markers (*ACTA2*, *TAGLN*, *DDR2* and *SNAI2*, Figure 1.3c). These results indicate that cellular heterogeneity across iPSC-CVPC samples largely reflects different proportions of CMs and EPDCs.

To characterize the similarities between the iPSC-CVPC transcriptomes and those of adult heart and artery samples, we performed a PCA analysis using the transcriptomes of 184 iPSCORE iPSCs, 180 iPSC-CVPCs, and the 1,072 GTEx samples, including left ventricle, atrial appendage, coronary artery and aorta [17].We found that principal component 1 (PC1) showed that iPSC-CVPCs correspond to an intermediate

state between the iPSCs and adult samples, suggesting that the derived CMs and EPDCs are similar to immature cardiac cells (Figure 1.3d). PC2 divided the samples based on cardiac lineage, namely the myocardium (left ventricles and atrial appendages) and epicardium (coronaries and aorta) [18]. This analysis shows that derived iPSC-CMs and iPSC-EPDCs lie on different cardiac developmental trajectories, with the CMs corresponding to immature myocardium and the EPDCs to immature epicardium.

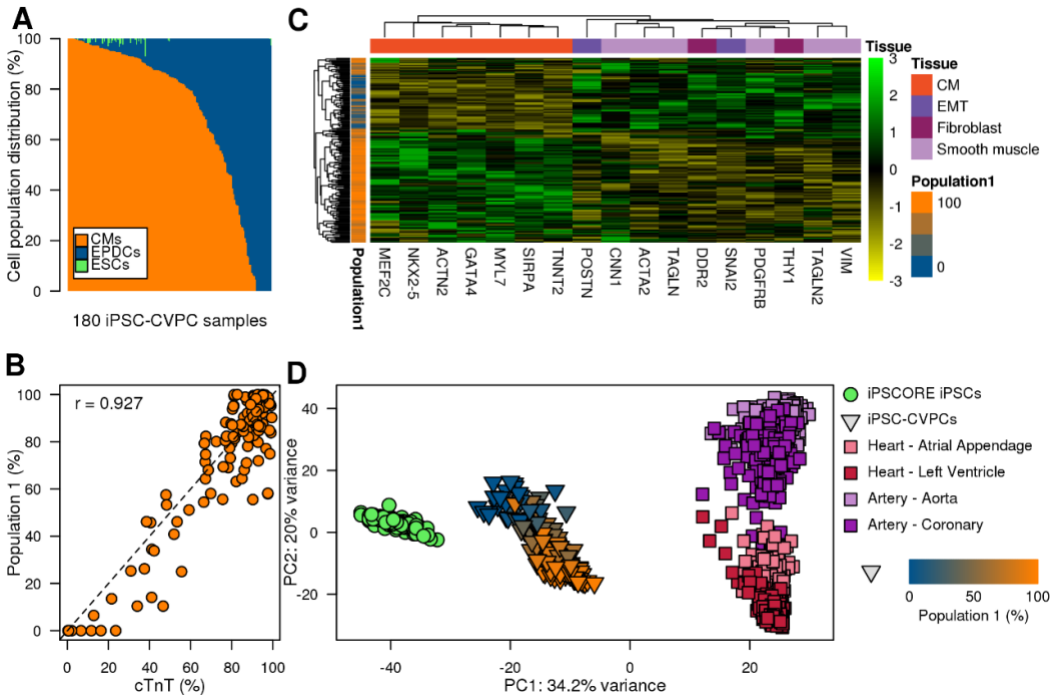
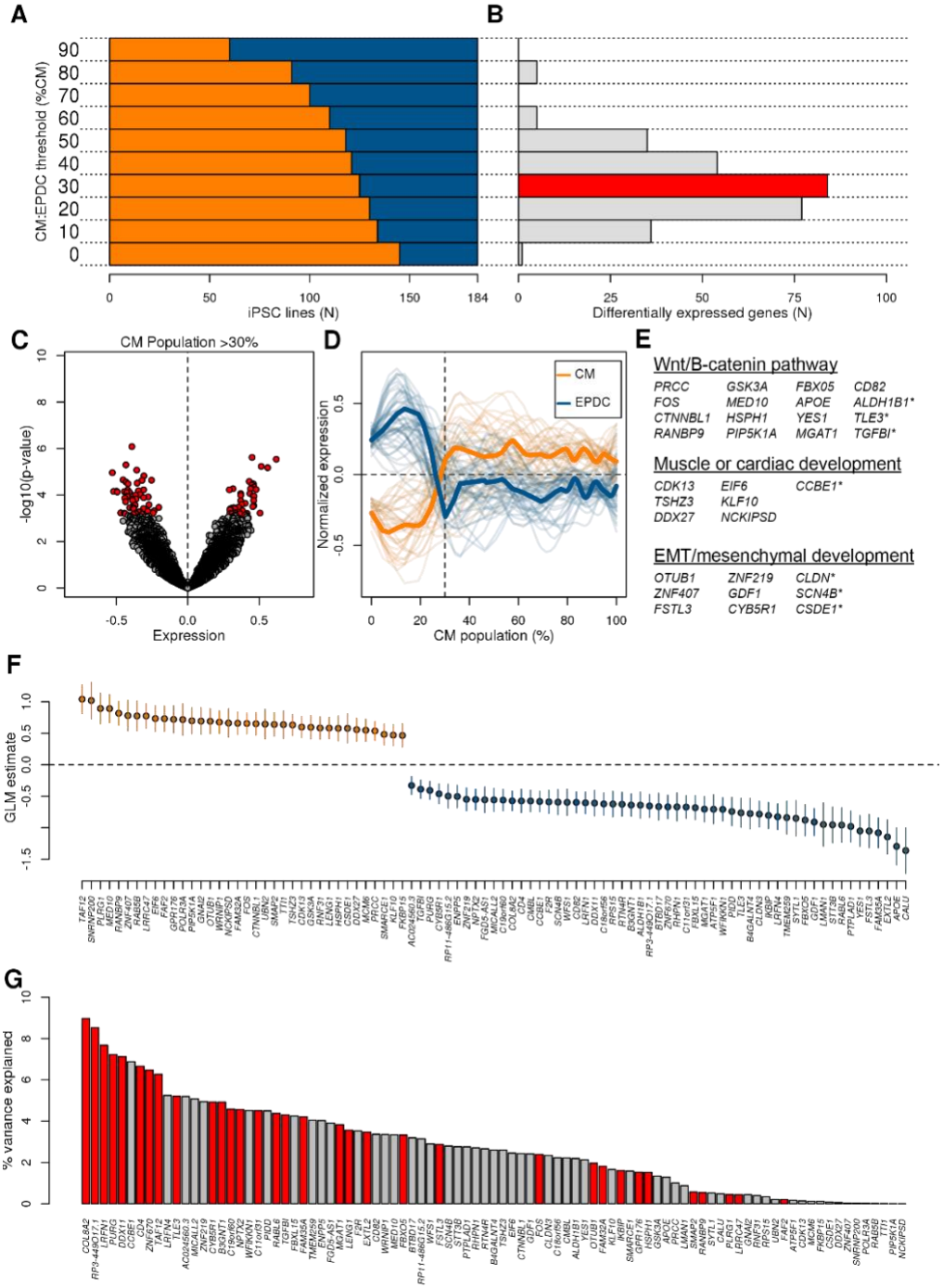


Figure 1.3 Transcriptomic features of 180 iPSC-CVPC samples
(a) Relative distributions of cell populations estimated using CIBERSORT across 180 iPSC-CVPC samples. (b) Scatterplot showing the correlation between %cTnT (X-axis) and the fraction of Population 1 in the iPSC-CVPCs calculated using CIBERSORT (Y-axis). (c) Heatmap showing the expression levels of CM and EPDC marker genes (Figure 1.1j) in 180 iPSC-CVPC samples at D25. Samples are colored based on their fraction of Population 1. (d) PCA on the 1,000 genes with highest variability from 184 iPSC samples, 180 iPSC-CVPC samples (triangles colored according to their % Population 1), and samples from GTEx (squares, left ventricle, right ventricles, coronary artery and aorta).

1.3.4 iPSC expression signatures impact cardiac fate differentiation

Although all iPSCORE iPSCs have previously been shown to be pluripotent [19], we sought to determine if transcriptomic differences existed between the iPSC lines that derived CVPCs containing CMs versus those that gave rise to EPDCs (Figure 1.4a). Given that all 180 iPSC-CVPCs contain both CMs and EPDCs but at different ratios, we initially had to determine the optimal CM:EPDC ratio to group the iPSC lines into those that were CM-fated and those that were EPDC-fated. Thresholds for 193 iPSC-CVPCs that completed differentiation (harvested on D25) were defined by the ratio of CM:EPDC estimates from CIBERSORT (estimated %CM: estimated %EPDC), while the 39 iPSC-CVPC differentiations terminated prior to D25 for not forming a beating syncytium were assigned a CM:EPDC ratio of 0:100 (0% CM:100% EPDC). We tested ten different CM:EPDC ratios and found 116 autosomal genes that were differentially expressed at one or more of these ratios (Storey q-value < 0.1, t-test, Figure 1.4a,b). We observed that the maximum number of the 116 genes (84, 72.5%) were differentially expressed at the 30:70 (CM:EPDC) threshold and 55 of them (47.4%) had their strongest p-value at this ratio. For this reason, we determined that the 30:70 threshold was optimal and grouped the iPSCs into 125 that were CM-fated (produced ≥ 30% CMs), and 59 that were EPDC-fated (produced >70% EPDCs, Figure 1.4b).

Figure 1.4 iPSC gene signatures associated with cardiac differentiation fate
(a) Testing of ten CM:EPDC ratios (0:100 to 90:10, with 10% increments) to determine the optimal threshold for defining an iPSC as CM-fated or EPDC-fated. For each threshold the number of iPSC lines defined as CM-fated (orange) or EPDC-fated (blue) is shown. (B) At the same thresholds indicated in (A), shown are the numbers of differentially expressed autosomal genes between the iPSC lines defined as CM-fated and EPDC-fated. The 30:70 threshold has the maximum number of differentially expressed genes. (C) Volcano plot showing mean difference in expression levels for all autosomal genes between CM-fated iPSC lines (their corresponding derived samples have CM population > 30%) and EPDC-fated iPSC lines (X axis) and p-value (Y axis, t-test). A positive difference indicates over-expression in CM-fated iPSCs, whereas a negative difference indicates over-expression in EPDC-fated iPSCs. Significant genes are indicated in red. (D) Expression levels of the 91 signature genes in iPSCs as a function of the % CM population in their corresponding iPSC-CVPC samples. Thick lines represent the average for 36 genes overexpressed in CM-fated iPSCs (orange) and for 55 genes overexpressed in EPDC-fated iPSCs (blue). (E) WNT/ß-catenin pathway, muscle/cardiac related, or EMT/mesenchymal development signature genes (those differentially expressed with nominal p-values ($p < 0.0015$) indicated with an asterisk). (F) GLM estimate (% CM population ~ expression) calculated for each signature gene. Mean and 95% confidence interval are shown. (G) Bar plot showing the percentage of variability in iPSC fate that is explained by each of the 91 signature genes. Bars highlighted in red show the 35 signature genes identified by L1 normalization that independently contributed to variance. Due to the fact that the 91 genes do not have independent expression, the total sum of the % variance explained is >1.


A
B
CM:EPDC threshold (%CM)
90
80
70
60
50
40
30
20
10
0
0
50
100
150
184
iPSC lines (N)
0
25
50
75
100
Differentially expressed genes (N)
C
CM Population >30%
-log10(p-value)
Expression
D
CM
EPDC
Normalized expression
CM population (%)
E
Wnt/B-catenin pathway
PRCC GSK3A FBX05 CD82
FOS MED10 APOE ALDH1B1*
CTNNBL1 HSPH1 YES1 TLE3*
RANBP9 PIP5K1A MGAT1 TGFBI*
Muscle or cardiac development
CDK13 EIF6 CCBE1*
TSHZ3 KLF10
DDX27 NCKIPSD
EMT/mesenchymal development
OTUB1 ZNF219 CLDN*
ZNF407 GDF1 SCN4B*
FSTL3 CYB5R1 CSDE1*
F
GLM estimate
1.0
0.5
0.0
-0.5
-1.5
G
% vanance explained
0
2
4
6
8
10

Of the 84 autosomal differentially expressed genes at the 30:70 (CM:EPDC) threshold, 35 were overexpressed in the CM-fated iPSC lines and 49 were overexpressed in the EPDC-fated iPSCs (Figure 1.4b,c,d). These genes have functions associated with three differentiation signatures: 1) Wnt/β-catenin pathway (13 genes); 2) muscle and/or cardiac differentiation (six genes); and 3), EMT and/or mesenchymal tissue development (six genes, Figure 1.4e). We noted that seven borderline significant autosomal genes were also involved in one of the three represented signatures, and therefore added them to the final list of differentially expressed genes. We investigated the associations between the expression levels of the final list of 91 signature genes (Table 1.2) in the 184 iPSCs and the fraction of CMs in the resulting iPSC-CVPCs using linear regression, and found significant associations for all genes (Figure 1.4f). These results show that, independently from the 30:70 (CM:EPDC) threshold used in the initial differential expression analysis, the expression levels of these signature genes in the 184 iPSCs were significantly associated with differentiation outcome (e.g. CM- or EPDC-fate).

Table 1.2 All 91 genes significantly differentially expressed between CM-fated and EPDC-fated iPSCs

Gene Name	Expression Difference	p-value (t-test)	Storey q-value
PRCC	0.618529454	2.88E-06	0.01205865
FOS	0.558406525	6.67E-06	0.013943832
WRNIP1	0.513679615	5.87E-06	0.013943832
CDK13	0.463221631	0.000112617	0.034464603
CTNNBL1	0.462866405	4.51E-05	0.02910997
TSHZ3	0.461497119	0.000286561	0.06420694
SMARCE1	0.459118456	0.000110539	0.034464603
RANBP9	0.458788345	3.33E-05	0.024611753
TAF12	0.449478357	2.40E-06	0.01205865
DDX27	0.448880833	3.69E-05	0.025726776
GSK3A	0.444607137	5.14E-05	0.02910997
LRRC47	0.438821079	5.45E-05	0.02910997
FAM32A	0.434495706	8.62E-05	0.031155239
EIF6	0.429786878	2.59E-05	0.024611753
KLF10	0.414685251	0.000616654	0.092558302
LENG1	0.408202719	0.000713826	0.09735479
MCM6	0.401969502	0.000227921	0.054996333
RNF31	0.401277066	0.00073254	0.098832907
FAF2	0.383104624	0.000473391	0.083467392
SMAP2	0.376384072	0.000253781	0.060080777
MED10	0.370823102	0.000102522	0.033852295
HSPH1	0.369954886	0.000175825	0.044122921
OTUB1	0.365142144	0.000641798	0.09281717
FKBP15	0.357407066	0.000555377	0.091691214
GNAI2	0.354845566	0.000334199	0.071073176
PIP5K1A	0.334889726	0.000713791	0.09735479
RAB5B	0.328344869	0.000383614	0.071841008
GPR176	0.321026033	0.000347509	0.071262367
POLR3A	0.316185628	5.36E-05	0.02910997
NCKIPSD	0.30889223	0.000650965	0.09281717
ZNF407	0.306692831	0.000368762	0.071816322
UBN2	0.296411412	0.000394992	0.072205659

Table 1.3 All 91 genes significantly differentially expressed between CM-fated and EPDC-fated iPSCs (Continued)

Gene Name	Expression Difference	p-value (t-test)	Storey q-value
TTI1	0.294890987	0.000619643	0.092558302
PLRG1	0.294560965	0.000432973	0.077609668
SNRNP200	0.290958385	5.93E-05	0.02910997
CSDE1	0.282710822	0.00089626	0.105070434
LMAN1	-0.202369391	0.000339866	0.071073779
STT3B	-0.234785987	0.000478957	0.083467392
SYTL1	-0.242179725	0.000593076	0.092004325
CALU	-0.253299668	2.30E-05	0.024413109
FSTL3	-0.271361948	0.000260981	0.060641254
TMEM259	-0.276598907	0.000120441	0.035144568
RABL6	-0.278855606	0.000499478	0.085851237
TLE3	-0.290777299	0.001171725	0.113969644
EXTL2	-0.306679543	1.49E-05	0.020835271
FBXO5	-0.308185476	5.95E-05	0.02910997
B4GALNT4	-0.30848578	0.000129442	0.036912736
PIDD	-0.309094532	0.000381197	0.071841008
ALDH1B1	-0.317848552	0.001106848	0.112000298
RPS15	-0.318386014	0.000671838	0.093784302
APOE	-0.319862507	2.33E-05	0.024413109
SCN4B	-0.327868094	0.000880088	0.105070434
CLDN3	-0.3318153	0.000834927	0.105070434
ATP5F1	-0.332244288	9.55E-05	0.033278093
WFS1	-0.33560547	0.000517693	0.087779657
FGD5-AS1	-0.340267921	0.000541854	0.090651242
RHPN1	-0.344623087	6.54E-05	0.029583781
PTPLAD1	-0.353806396	3.07E-05	0.024611753
YES1	-0.356807254	8.46E-06	0.015164087
BTBD17	-0.361029531	0.000147429	0.039358532
DDX11	-0.363962468	0.000630168	0.09281717
FBXL15	-0.364558194	0.000601269	0.092004325
IKBIP	-0.365221839	0.000100393	0.033852295
RTN4R	-0.36702925	0.000272395	0.062142609

Table 1.4 All 91 genes significantly differentially expressed between CM-fated and EPDC-fated iPSCs (Continued)

Gene Name	Expression Difference	p-value (t-test)	Storey q-value
CD4	-0.377453722	0.00039707	0.072205659
RP11-486G15.2	-0.377453777	0.00036343	0.071816322
B3GNT1	-0.380958591	0.000672697	0.093784302
ZNF219	-0.383156897	0.000372035	0.071816322
MGAT1	-0.386426098	0.000141543	0.038608449
FAM35A	-0.390447022	8.25E-07	0.010347182
ENPP5	-0.393241942	0.000333879	0.071073176
CMBL	-0.405426981	0.000352127	0.071262367
F2R	-0.407691113	7.54E-05	0.029583781
CCBE1	-0.408620523	0.001349212	0.115952539
LRFN4	-0.417677335	3.27E-05	0.024611753
MICALL2	-0.420069961	6.03E-05	0.02910997
WFIKKN1	-0.426110939	0.000112023	0.034464603
C18orf56	-0.426783944	7.85E-05	0.029840058
ZNF670	-0.432012917	2.76E-05	0.024611753
GDF1	-0.432533942	5.03E-06	0.013943832
C19orf60	-0.434593453	0.00019482	0.047931036
PURG	-0.436080216	0.000643584	0.09281717
LRFN1	-0.440493195	0.000135821	0.037871174
CYB5R1	-0.44313245	0.000600015	0.092004325
C11orf31	-0.443214265	7.12E-05	0.029583781
RP3-449O17.1	-0.452125413	2.22E-05	0.024413109
TGFBI	-0.452163339	0.001308647	0.114534817
AC024560.3	-0.468757978	0.000585005	0.092004325
NPTX2	-0.492895354	7.37E-05	0.029583781
COL8A2	-0.498153533	6.88E-05	0.029583781
CD82	-0.527262952	1.09E-05	0.017147403

1.3.5 Signature genes capture a large fraction of the variance underlying iPSC fate outcome

While the signature genes likely impacted cardiac fate determination, we did not expect each gene to contribute equally. To explore the impact of each gene individually on differentiation outcome, we calculated how much the 91 genes explained the variability underlying iPSC cell fate. To quantify the percent of variance explained by each gene (R^2), we fit a generalized linear regression model with a logit link function to each gene individually. We found that the percent of variance explained by each individual gene varied over three orders of magnitude ($1.73 \times 10^{-3} < R^2 < 8.97\%$; Figure 1.4g).

We next asked how these signature genes altogether captured variability in differentiation fate. As several of the signature genes had correlated expression levels, to reduce overfitting in the regression analysis, we included an L1 norm penalty (i.e. LASSO regression) and used 10-fold cross validation. We identified 35 genes that independently contributed to variance, and whose expression levels collectively explained more than half of the variability in differentiation outcome across iPSC lines (average R^2 from the 10-fold cross validation = 0.512). Together these data show that, while the proportion of variance explained by each of the signature genes varied widely, altogether they captured approximately half of the total variance underlying differential iPSC fate outcome.

1.3.6 Inherited genetic variation does not influence differentiation outcome

We investigated if genetic variation associated with the expression of any of the signature genes contributed to the differentiation outcome of iPSCs. We assessed the genotypes of 8,620,159 variants in each iPSC line and performed a GWAS study to investigate the association between genotype and the fraction of CMs in the corresponding iPSC-CVPCs. We found that none of these variants associated with differentiation outcome at genome-wide significance ($p < 5 \times 10^{-8}$, Figure 1.5a). To further examine the association between genetic background and differentiation outcome, we tested if differentiations of different iPSC clones from the same individual, and from members of the same twin pair, were more likely to yield similar outcomes compared with differentiations of iPSC clones from individuals with different genetic backgrounds, and observed similar distributions Figure 1.5b). While our power to perform a GWAS study was limited, this analysis shows that the genetic background did not contribute to the variance underlying iPSC differentiation outcome, indicating that non-genetic (i.e., clonality and passage) factors played a role in determining whether an iPSC line differentiated to CMs or EPDCs.

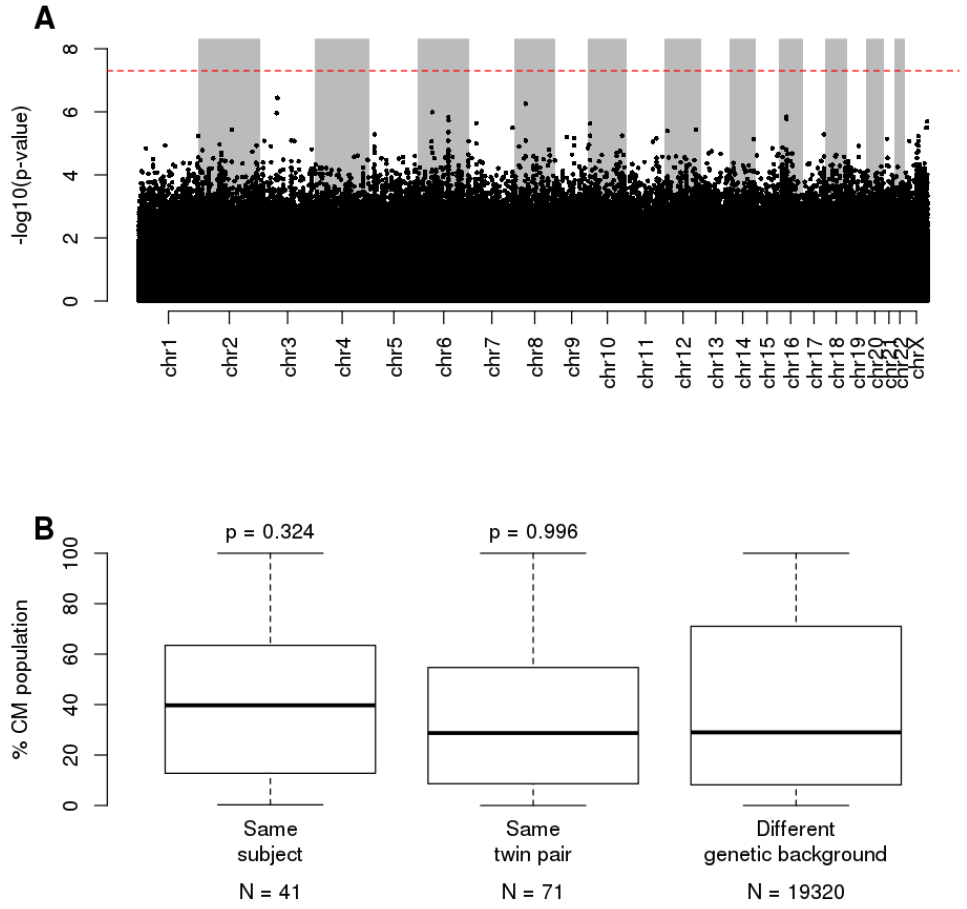
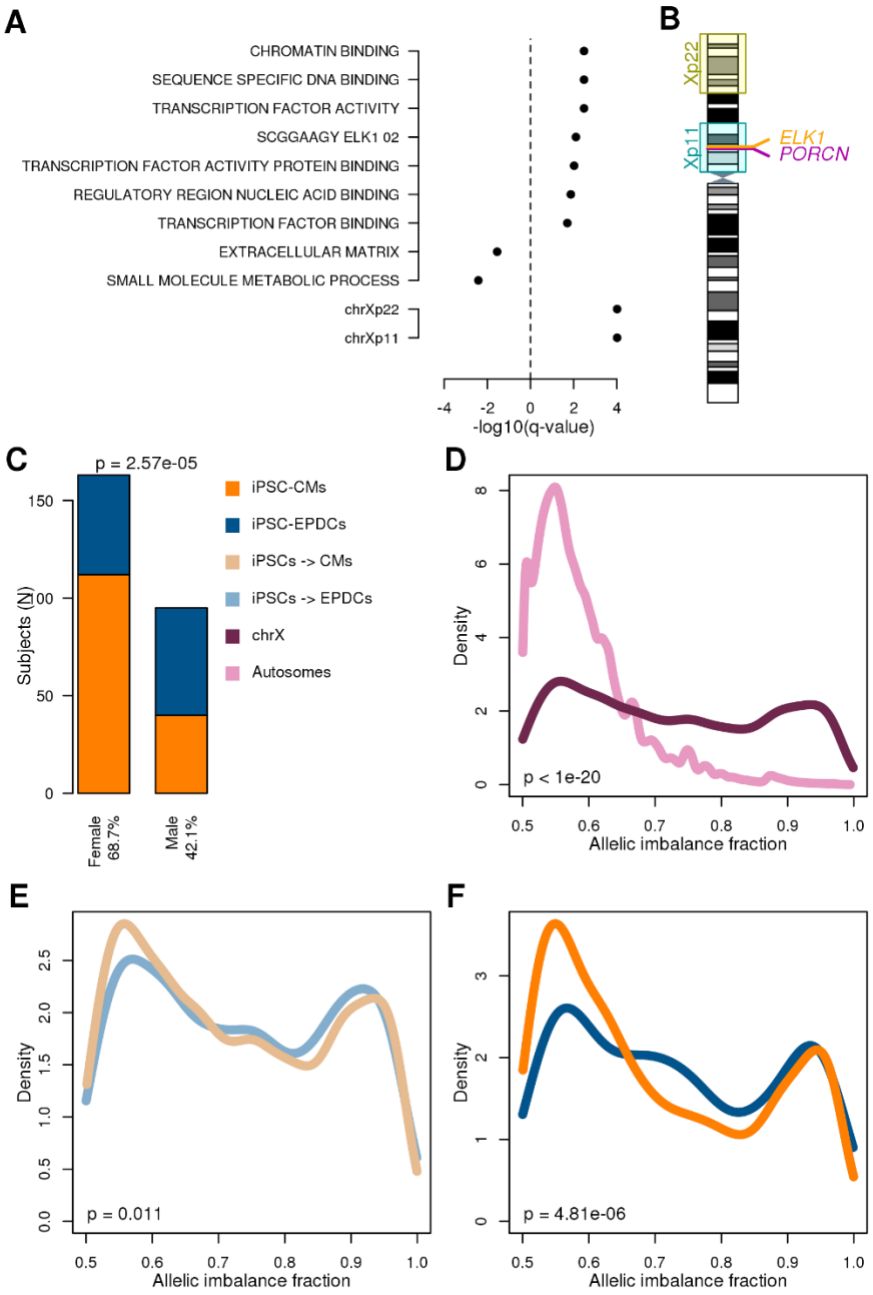


Figure 1.5 Associations between genetic background and differentiation outcome
(a) Manhattan plot showing the association between genetic variation and differentiation outcome (measured as % CM population in iPSC-CVPCs). Red dashed line shows p-value = 0.05 adjusted using Bonferroni's method ($p = 5 \times 10^{-8}$). (b) Boxplots showing distributions of the differences in the %CM population between differentiations of different iPSC clones from the same subject, from the same twin pair, and from individuals with different genetic backgrounds. P-values were calculated using Mann-Whitney U test.

1.3.7 GSEA implicates ELK1 targets and genes on the X chromosome

To understand whether the transcriptomic differences between CM-fated and EPDC-fated iPSCs were associated with alterations in specific pathways or cellular function, we performed a gene set enrichment analysis (GSEA) on 9,808 MSigDB gene sets [20] using the 15,228 expressed autosomal genes in the 184 iPSCs. We identified 22 gene sets that were significantly associated with iPSC cell fate, including enrichment in the 59 EPDC-fated iPSCs for extracellular matrix (Figure 1.6a) and in the 125 CM-fated iPSCs for transcription factor activity and ELK1 targets. To capture gene sets associated with expression differences on the X chromosome, we performed differential expression and GSEA on 113 female iPSC lines (87 CM-fated; 26 EPDC-fated). The two most significant gene sets were loci located within chrXp11 and chrXp22 (Figure 1.6b). Notably, the chrXp11 locus encodes both *ELK1* and *PORCN*, whose protein product (Porcupine) is targeted for WNT inhibition in CM differentiation protocols, but not EPDC differentiation protocols [8,9] (Figure 1.6b). The chrXp22 locus includes the majority of genes (52/99, 52.5%) that are known to escape chromosome X inactivation [21], and thus may potentially have varying X-linked gene dosage across female iPSCs. Overall, GSEA shows that genes differentially expressed between CM-fated and EPDC-fate iPSCs are involved in a variety of pathways including ELK targets and potentially associated with the X chromosome activation status.

Figure 1.6 X chromosome gene dosage plays a role in cardiac differentiation fate
(a) GSEA results: For each gene set, -$\log_{10}$(q-value) is shown. Positive values correspond to gene sets enriched in CM-fated iPSCs, whereas negative values correspond to EPDC-fated iPSCs. For autosomes all iPSCs were included (top), for the chromosome X only the 113 female iPSCs were analyzed (bottom). Storey q-value was used to adjust for multiple testing hypothesis, q-values < 0.05 were considered significant. (B) Cartoon showing the differentially expressed loci on chromosome X and the position of *ELK1* and *PORCN*. (C) Barplot showing the associations between sex and differentiation outcome (orange: iPSC-CVPC samples with CM fraction > 30%; blue: with EPDC fraction > 70%). P-values were calculated using Z-test (glm function in R). (D) Density plot showing the differences in allelic imbalance fraction between autosomal genes (pink) and chrX genes outside of the pseudoautosomal region (maroon) in female iPSCs. (E) Density plot showing the differences in allelic imbalance fraction between chrX genes in female CM-fated (light orange) and EPDC-fated (light blue) iPSCs. (F) Density plot showing the differences in allelic imbalance fraction between chrX genes in female D25 iPSC-CVPC samples with CM fraction > 30% (orange) and EPDC fraction > 70% (blue). P-values in D-F were calculated using Mann Whitney U test.


A
CHROMATIN BINDING
SEQUENCE SPECIFIC DNA BINDING
TRANSCRIPTION FACTOR ACTIVITY
SCGGAAGY ELK1 02
TRANSCRIPTION FACTOR ACTIVITY PROTEIN BINDING
REGULATORY REGION NUCLEIC ACID BINDING
TRANSCRIPTION FACTOR BINDING
EXTRACELLULAR MATRIX
SMALL MOLECULE METABOLIC PROCESS
chrXp22
chrXp11
-4 -2 0 2 4
-log10(q-value)
B
Xp22
Xp11
ELK1
PORCN
C
p = 2.57e-05
iPSC-CMs
iPSC-EPDCs
iPSCs -> CMs
iPSCs -> EPDCs
chrX
Autosomes
Subjects (N)
150
100
50
0
Female
68.7%
Male
42.1%
D
Density
p < 1e-20
0.5 0.6 0.7 0.8 0.9 1.0
Allelic imbalance fraction
E
Density
p = 0.011
0.5 0.6 0.7 0.8 0.9 1.0
Allelic imbalance fraction
F
Density
p = 4.81e-06
0.5 0.6 0.7 0.8 0.9 1.0
Allelic imbalance fraction

1.3.8 Sex is associated with iPSC differentiation outcome

To identify other iPSC factors potentially associated with differentiation outcome, we examined three characteristics of the 181 subjects in our study (sex, ethnicity, and age) and passage of the iPSCs at D0. Analyzing the 125 CM-fated and 59 EPDC-fated iPSC lines with a general linear model, we found no association between differentiation outcome and ethnicity or age ($p > 0.8$; GLM, Z-test; Figure 1.7a,b), but observed a significant association with sex ($p = 2.57 \times 10^{-5}$, GLM, Z-test; Figure 1.6c) and a trend for iPSC passage at D0 ($p = 0.069$, GLM, Z-test; Figure 1.7c). These data suggest that iPSCs derived from female subjects and iPSCs with higher passages at D0 had an increased predisposition for the CM fate. Furthermore, considering only the 191 completed differentiations (D25 iPSC-CVPC samples), we found that iPSC-CVPC samples derived from female subjects compared to those derived from males had significantly higher %cTnT values (mean = 83.0% and 77.7%, respectively for females and males, $p = 6.0 \times 10^{-4}$, Mann-Whitney U test; Figure 1.7d) and a higher fraction of CMs ($p = 6.46 \times 10^{-4}$, Mann-Whitney U test; Figure 1.7e). These results indicate that iPSCs derived from female subjects and, to a lesser extent, iPSCs that have spent more time in cell culture, have a greater inherent predisposition to differentiate towards the CM lineage.

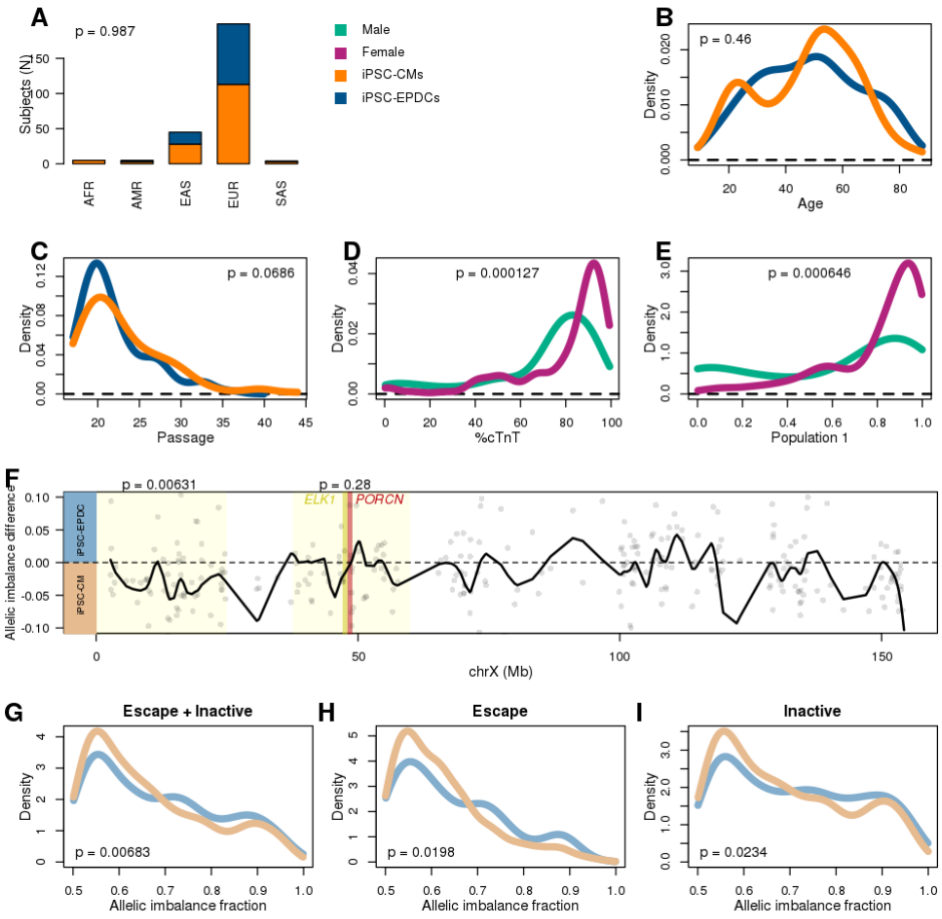


Figure 1.7 X chromosome inactivation in iPSCs
(a,b,c) Associations between differentiation outcome (orange: iPSC-CVPC samples with CM fraction > 30%; blue: with EPDC fraction > 70%) and (a) ethnicity (most similar superpopulation from the 1000 Genomes Project), (b) age at enrollment, and (c) passage at monolayer (D0). (a) is shown as barplots; (b,c) are shown as density plots. P-values were calculated using Z-test (glm function in R). (d,e) Density plots showing the association between sex (teal: males; magenta: females) and (d) %cTnT, and (e) fraction of CM population for 191 iPSC lines. P-values were calculated using Mann-Whitney U test. (f) Allelic imbalance difference between CM-fated and EPDC-fated iPSCs. The dots represent each gene on chrX, while the black solid line corresponds to the smoothed interpolation of differences for all the genes. The locations of the Xp22 and Xp11 loci on the chrX G-banding ideogram are highlighted in yellow, as well as *ELK1* (yellow) and *PORCN* (red). P-values above each locus indicate the difference in allelic imbalance between CM-fated and EPDC-fated iPSCs in each locus (Mann Whitney U test). (g,h,i) Allelic imbalance fraction from inactive and escape genes in Xp22: Density plots showing the allelic imbalance differences in chrX genes on the Xp22 loci in female samples between iPSC lines with CM-fate (light blue) and EPDC-fate (light orange) differentiations. Allelic balances compared in Xp22 are from all genes in the region (g), escape genes (h), and inactive genes (i). P-values were calculated using Mann Whitney U test.

1.3.9 Female iPSCs with X chromosome reactivation associated with CM fate

Given the observation that female iPSCs have a greater potential to differentiate to CMs and that differential expression of chrXp11 genes were associated with differentiation outcome, we asked if variation in X chromosome inactivation (Xi) and activation (Xa) state across female iPSC lines was associated with CM or EPDC-fate. Using RNA-seq data generated from the 113 female iPSCs, we evaluated allele specific effects (ASE) of X chromosome and autosomal genes. We defined the strength of ASE for each gene as the fraction of RNA transcripts that were estimated to originate from the allele with higher expression (hereto referred to as "allelic imbalance fraction", AIF). We observed that AIF in autosomal genes was close to 0.5, indicating that both alleles were equally expressed (Figure 1.6d), while AIF on the X chromosome in iPSCs tended to be bimodal, with some genes showing monoallelic expression (AIF ~1.0; XaXi) and others showing biallelic expression (AIF ~0.5; XaXa). We observed that AIF was less in the 87 CM-fated female iPSCs compared with the 26 EPDC-fated female iPSCs ($p = 0.011$, Mann-Whitney U test, Figure 1.6e) and that this difference in AIF became even more pronounced in the corresponding derived iPSC-CVPC samples ($p = 4.81 \times 10^{-6}$, Mann-Whitney U test, Figures 1.6f, 1.7f). These finding show that differential chromosome XaXi status, as well as altered gene expression in chrXp22 and chrXp11, in iPSCs contribute to differences in cardiac fate differentiation outcome.

Since we observed differences in X chromosome reactivation state between CM-fated and EPDC-fated female iPSCs, we next asked if the two GSEA X chromosome associated intervals (chrXp22 and chrXp11; Figure 1.6a) showed corresponding allelic imbalance trends. We plotted AIF differences, where a positive AIF difference indicates

X chromosome reactivation in the 26 EPDC-fated iPSCs and a negative in the 87 CM fated-iPSCs (Figure 1.7f). We observed that distinct regions across the X chromosome were differentially eroded in the EPDC-fated versus CM-fated iPSCs. In particular, chrXp22 showed X reactivation in CM-fated iPSCs ($p = 6.31 \times 10^{-3}$, Mann Whitney U), with both escape ($p = 0.020$, Mann Whitney U) and non-escape genes ($p = 0.023$, Mann Whitney U) showing evidence of reactivation (Figure 1.7g,h,i). As chrXp22 contains more than half of escape genes on the X chromosome, this observation confirms that increased X reactivation in CM-fated iPSCs results in increased expression of both escape and non-escape genes. As GSEA identified genes on chrXp11 to be overexpressed in CM-fated iPSCs, the lack of X reactivation in this interval ($p = 0.28$, Mann Whitney U) suggests alternative regulatory mechanisms may also alter gene expression levels on the X chromosome. Overall, these results suggest that differential X chromosome reactivation as well as other mechanisms underlying altered regulation of X chromosome and autosomal genes contribute to iPSC cardiac lineage fate determination

1.3.10 Independent iPSC-CM derivation study validates findings

To assess the generalizability of our findings, we examined an independent collection of 39 iPSCs [12] reprogrammed using an episomal plasmid from Yoruba lymphoblastoid cell lines Figure 1.8a). Differentiation of these lines resulted in the successful derivation of 13 iPSC-CMs (%cTnT range at D32: 40 to 96.9), whereas 24 were terminated on or before day 10 due to the fact that they did not form a beating syncytium. To examine if the successfully derived Yoruba iPSC-CMs showed the presence of EPDCs, we used RNA-seq data and CIBERSORT to estimate cellular

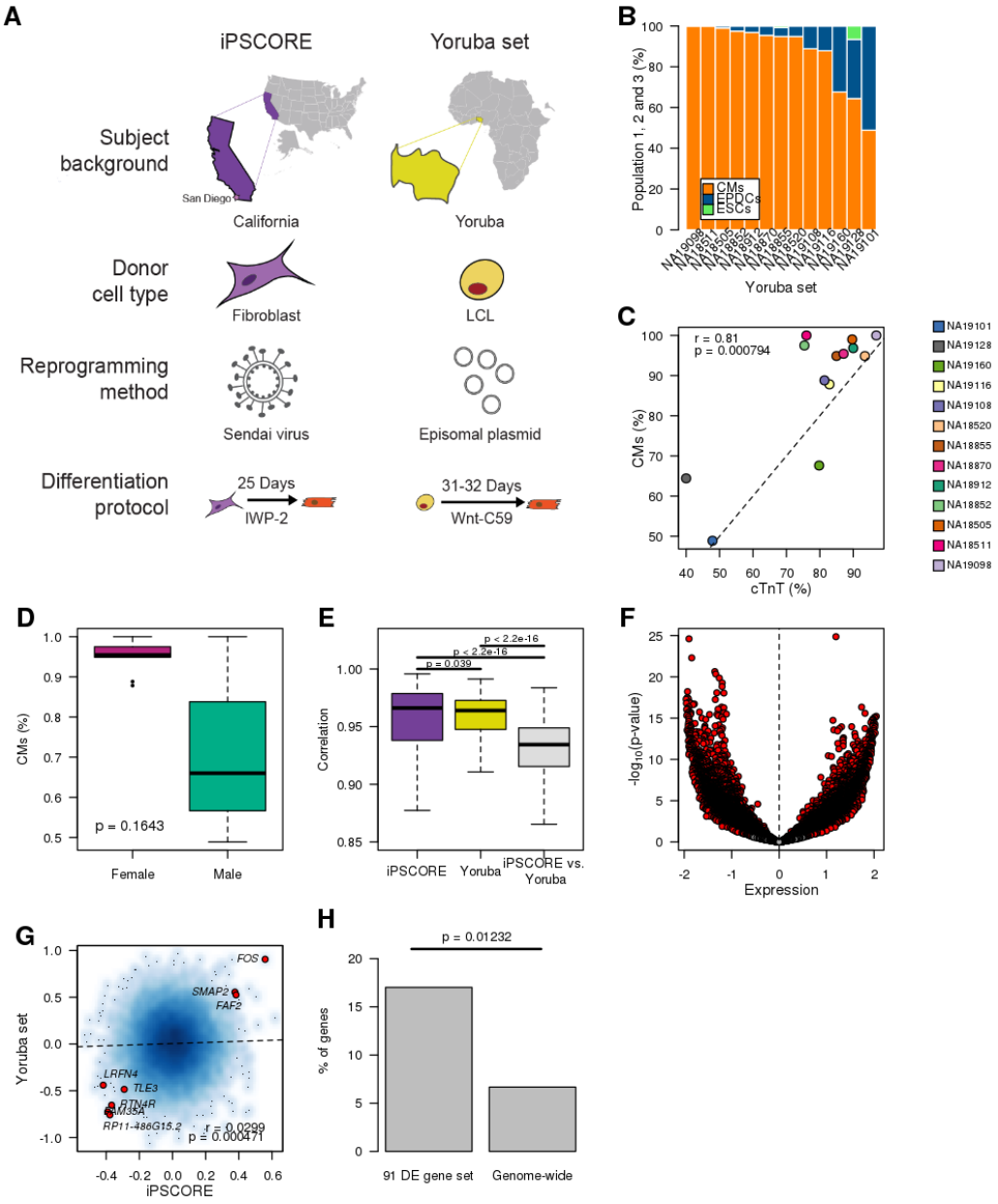
compositions and observed variable relative distributions of CM and EPDC populations (Figure 1.8b). Consistent with our iPSCORE iPSC-CVPC samples, the estimated CM population fractions were significantly correlated with %cTnT values ($r = 0.81$, $p = 7.94 \times 10^{-4}$, t-test; Figure 1.8c). To understand if the CMs and EPDCs appear at the same time during differentiation, we analyzed data generated the from the Yoruba lines a four timepoints [22] and observed that both cardiac lineages are typically present by day 5 and that the ratio of these two cardiac cell types remains relatively stable past day 10. Finally, Yoruba iPSC-CMs derived from females tended to have an increased percentage of CMs compared with those derived from males (Figure 1.8d). These observations show that the Yoruba iPSCs and derived cardiac cells could be used to investigate the generalizability of the associations that we had observed between transcriptomic differences in iPSCs and cardiac fate differentiation outcome.

As several factors (Figure 1.8a) were different between the iPSCORE iPSC and Yoruba iPSC sets (i.e. different reprogramming method, genetic backgrounds, and donor cell types), we expected that there would be significant differences between their transcriptional profiles. We initially analyzed how correlated gene expression was: 1) within iPSCORE iPSCs; 2) within Yoruba iPSCs; and 3) between all pairwise comparisons of the iPSCs in these two different collections (Figure 1.8e). We observed high correlations of gene expression across iPSCs within each collection, however the correlation between samples from different studies was significantly decreased, indicating that the two sets have significant genome-wide gene expression differences. We next examined differential gene expression between the CM-fated iPSCORE iPSC and Yoruba iPSCs that successfully differentiated into iPSC-CMs (Figure 1.8f), and

observed that the majority of genes (69.6% with q-value < .10) were significantly differentially expressed between the two iPSC sets. These results show that there are strong batch effects on gene expression between the iPSCORE and Yoruba iPSC lines.

Figure 1.8 Validation of association between iPSC gene signatures, sex and differentiation outcome

(a) Schematic depicting differences (subject ethnicities, donor cell type, reprogramming method) between the iPSCORE iPSC and Yoruba iPSC as well as differences in the cardiac differentiation protocol. (b) Estimated fractions of CMs and EPDCs for 13 Yoruba iPSC-CM samples from RNA-seq using CIBERSORT (two iPSC-CMs did not have RNA-seq). (c) Scatterplot showing the correlation between %cTnT (X axis) and the fraction of cells in population 1 (% CMs) calculated using CIBERSORT (Y axis) for 13 Yoruba iPSC-CM samples. (d) Boxplots showing the distribution of estimated fraction of cells in population 1 (% CMs) for 9 female Yoruba iPSC-CM and 4 male Yoruba iPSC-CM. P-value was calculated using Mann-Whitney U test. (e) Box plots showing correlation of gene expression in all 184 iPSCORE iPSCs with RNA-seq (purple), 34 Yoruba iPSCs with RNA-seq used for differentiation (yellow; 14 successful iPSC and 20 terminated iPSC, five iPSCs did not have RNA-seq), and the pairwise comparison of the Yoruba iPSC against the iPSCORE iPSC (grey). (f) Volcano plot showing mean difference in expression levels for all autosomal genes between 14 Yoruba iPSC lines that were successfully differentiated and 125 iPSCORE iPSC lines that differentiated to >30% Population 1 (CMs) (X-axis) and p-value (Y-axis, t-test). A positive difference between mean expressions indicate iPSCORE-specific over-expression, whereas a negative difference between mean expressions indicate Yoruba-specific over-expression. Significant genes are indicated in red. (g) Smooth color density scatterplot showing gene expression differences between iPSCs with different fates in 184 iPSCORE iPSC (125 CM-fated vs 59 EPDC-fated) (X-axis) to the expression differences between iPSCs with different outcomes in Yoruba iPSC (14 successful vs 20 terminated) (Y-axis). A positive difference indicates shared over-expression of genes between CM-fated iPSC in iPSCORE and successfully differentiated iPSC in the Yoruba set, whereas a negative difference indicates shared over-expression of genes between EPDC-fated iPSC in iPSCORE and terminated iPSC in the Yoruba set. Of the 91 signature genes that were differentially expressed in the iPSCORE iPSCs based on cell fate, eight had nominally significant expression differences in the same direction in the Yoruba iPSC set (shown in red). (h) Barplot showing that the eight iPSCORE differentially expressed genes (panel g) with nominal significant expression differences in the same direction (e.g. over-expressed or down regulated) in the Yoruba iPSCs is greater than random expectation. Of 13,704 genes expressed both in the iPSCORE and Yoruba iPSCs, we obtained 6,909 for which the average normalized expression differences had either the same positive (CM fate/successful differentiation) or negative (EPDC fate/terminated differentiation) direction. The 6,909 genes included 47 of the 91 iPSCORE signature genes. We found that 466 (6.7%) of the 6,909 genes were nominally significant for being differentially expressed between the 14 successful and 20 terminated differentiations in the Yoruba samples, while 8 of the 47 iPSCORE differentially expressed genes (17.0%) had a nominal $p < 0.05$. This analysis shows that the 91 iPSCORE signature genes are 2.5 times more likely than expected (17.0% vs. 6.7%, $p = 0.012$, Fisher's exact test) to be differentially expressed in the Yoruba samples based on cardiac differentiation fate.


A
iPSCORE
Yoruba set
Subject background
San Diego
California
Yoruba
Donor cell type
Fibroblast
LCL
Reprogramming method
Sendai virus
Episomal plasmid
Differentiation protocol
25 Days
IWP-2
31-32 Days
Wnt-C59
B
Population 1, 2 and 3 (%)
CMs
EPDCs
ESCs
Yoruba set
C
r = 0.81
p = 0.000794
CMs (%)
cTnT (%)
NA19101
NA19128
NA19160
NA19116
NA19108
NA18520
NA18855
NA18870
NA18912
NA18852
NA18505
NA18511
NA19098
D
CMs (%)
p = 0.1643
Female
Male
E
Correlation
p < 2.2e-16
p < 2.2e-16
p = 0.039
iPSCORE
Yoruba
iPSCORE vs. Yoruba
F
-log10(p-value)
Expression
G
Yoruba set
FOS
SMAP2
FAF2
LRFN4
TLE3
PTN4R
SMS5A
RP11-486G15.2
r = 0.0299
p = 0.000471
iPSCORE
H
p = 0.01232
% of genes
91 DE gene set
Genome-wide

We investigated if, despite the strong batch effects on gene expression between iPSCORE and Yoruba iPSCs, we could detect inherent transcriptional differences impacting cardiac fate determination that were shared between the iPSC sets. Given the relatively small size of the Yoruba study there was insufficient power to detect transcriptional differences between the lines with different differentiation outcomes (Successfully completed versus Terminated). Therefore, for each gene, we compared the mean expression differences between iPSCs with different cardiac fate outcomes in iPSCORE (CM-fate – EPDC-fate) to the expression differences between iPSCs with different differentiation outcomes in the Yoruba set (Successfully completed – Terminated, Figure 1.8g). We observed a small, but significant correlation ($r = 0.0299$, $p = 4.71 \times 10^{-4}$, t-test) between genes that were differentially expressed in the iPSCORE iPSCs and those that were differentially expressed in the Yoruba iPSCs. Further, we specifically examined the 91 signature genes significantly associated with iPSCORE iPSC cardiac fate outcome and found eight with nominally significant expression differences in the same direction (e.g. overexpressed or downregulated) in the two sets of iPSCs (Figure 1.8g), which is 2.5 times more than random expectation ($p = 0.012$, Fisher's exact test; Figure 1.8h). These data suggest that the iPSCORE iPSCs and Yoruba iPSCs shared transcriptional differences that impacted cardiac fate differentiation outcome.

1.4 Discussion

While previous directed cardiac differentiation studies have observed the emergence of both cardiomyocytes and a non-contractile cell population, the origin of

these non-contractile cells, and whether the same or different non-myocyte cell types are present across iPSC-CVPC samples has not previously been addressed. We showed that two distinct cell types were present in 154 iPSC-CVPC samples derived from iPSCs in iPSCORE. One of the derived cell types were cardiomyocytes (CMs), characterized by high expression levels of cardiac-specific genes, and the other derived cell type was epicardium-derived cells (EPDCs), characterized by high expression of marker genes for EMT, smooth muscle and fibroblasts. We found the same two cardiac cell types present in iPSC-CMs derived from an independent collection of 39 Yoruba iPSCs; and that both cardiac cell types were typically present by day 5 and their ratios remained relatively stable past day 10 [12,22]. A recent study showed that adding hESC-derived epicardial cells to cardiomyocyte grafts *in vivo* improves transplantation efficacy, as it increases contractility, myofibril structure and calcium handling and decreases tissue stiffness [23]. Our findings suggest that the generation of EPDCs during iPSC-CM differentiation may enhance the structure of the derived CMs; and that to efficiently use iPSC-CVPCs in a clinical setting, future studies may need to optimize the relative proportions of CMs and EPDCs that maximize their transplantation efficiency.

The scale of our study, 232 attempted differentiations of 191 iPSC lines into the cardiac lineage, provided the power to develop a framework to identify non-genetic transcriptional differences in iPSCs that influence their cardiac differentiation outcome. To minimize the factors that might influence differentiation outcome, such as the optimal cell confluency at which to start differentiation, we attempted to standardize all steps in the differentiation protocol, in order to remove subjective decisions and diminish experimental differences between samples. We identified 91 signature genes whose

differential expression was associated with differentiation outcome and showed that many of these genes are involved in cardiac development, including the Wnt/β-catenin pathway, muscle differentiation or cardiac-related functions, and the transition of epicardial cells to EPDCs by EMT (Figure 1.9). Many of the transcriptomic differences between iPSCORE iPSCs with CM-fates versus those with EPDC-fates may be due to aberrant epigenetic landscapes resulting from a combination of the reprogramming method (Sendai virus) and cell of origin (fibroblasts). However, given that the Yoruba iPSCs were reprogrammed using a different method (Episomal plasmid) and cell of origin (LCLs), and yet the iPSCORE and Yoruba iPSCs shared gene expression differences associated with cardiac lineage outcome, it is likely that our findings will likely be generalizable to other collections of iPSCs. We hypothesize that the signature genes associated with cardiac lineage outcome will vary across iPSC collections and depend on the reprogramming method and cell type of origin but will largely be involved in the same pathways identified in this study.

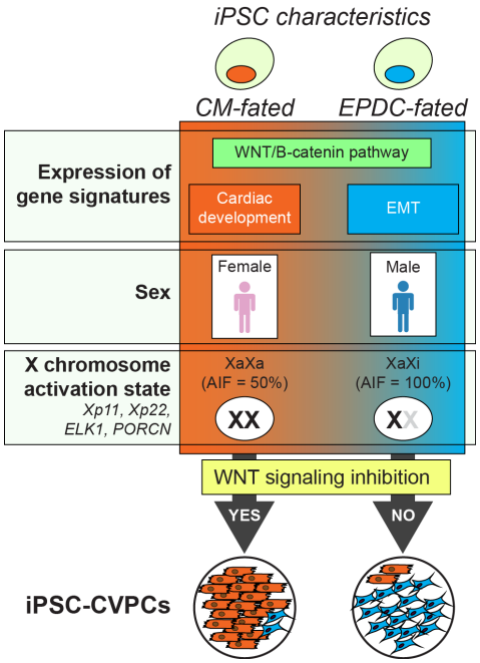


Figure 1.9 iPSC characteristics that influence their cardiac fate determination
(a) Cartoon showing iPSC characteristics that influence their cardiac fate determination, including: 1) the expression levels of 91 genes grouped into three gene signature classes (WNT/B-catenin pathway, cardiac development genes and genes involved in EMT, see Figure 1.4); 2) sex, female iPSCs are more likely to differentiate to CMs than males (see Figure 1.6); and 3) X chromosome activation state, female iPSCs that have activated both X chromosomes (XaXa) are more likely to differentiate to CMs (see Figure 1.6).

We observed that variability across iPSCs on X chromosome gene dosage (XaXa vs XaXi vs XY) played a role in cardiac lineage fate (Figure 1.6). While human iPSCs are known to have only partial XaXa [24,25], we identified two loci (chrXp11 and chrXp22) encoding genes whose expression levels are associated with two the distinct cardiac differentiation trajectories (CMs vs. EPDCs). The higher expression of chrXp11 genes in CM-fated iPSCs may at least in part be due to fact that *ELK1* and *PORCN* are both encoded in this interval, as the protein product of *PORCN* (Porcupine) is inhibited by IWP-2 during CM differentiation [8], but not during EPDC differentiation [7,10,11,15] (some EPDC protocols inhibit Porcupine but then reactivate the WNT pathway at a later time point [26-28]. Furthermore, we found that ELK1 targets are overexpressed in CM-fated iPSCs, which is consistent with previous studies showing that knockdown of *ELK1* in immortalized human bronchial epithelial cells, small airway epithelial cells, and luminal breast cancer cell line (MCF-7) is associated with increased EMT [29,30]. Also consistent with ELK1 playing a role in the association between X chromosome dosage and differentiation outcome is a previous study showing that *ELK1* overexpression or downregulation respectively mimics the phenotypes of XaXa or XaXi pluripotent stem cells [31]. Of note, atrioventricular septal defects (AVSD) occur in ~20% of individuals with Down Syndrome (DS), and has a higher prevalence in female DS patients [32]. Given that EPDCs play an essential role in septal formation [33], our study suggests that future work should investigate the extent to which X chromosome gene expression levels are altered in cardiomyocytes from individuals with DS, and if this is associated with the formation of fewer EPDCs.

Overall, our study suggests that expression differences of 91 signature and X chromosome genes result in the iPSCORE iPSC lines having differential propensities to respond to WNT inhibition during differentiation, and consequently are fated to produce iPSC-CVPC samples with different proportions of CMs and EPDCs. As iPSCs in the iPSCORE collection have passed standard quality checks to confirm their pluripotency and genomic integrity [12,19], these transcriptomic expression differences associated with cardiac lineage outcome are not detected using current quality metrics. In conclusion, our findings suggest that to derive human iPSC lines that respond similarly in differentiation protocols, it may be necessary to improve reprogramming methods such that the transcriptome and X chromosome activation state is fully reset to the naïve state, and incorporate inactivation of one of the X chromosomes in female lines as an early step in differentiation protocols.

1.5 Experimental procedures

Subject information and whole genome sequencing

Individuals (108 female and 73 male) were recruited as part of the iPSCORE project [19] and included 7 MZ twin pairs, members of 32 families (2-10 members/family) and 71 singletons and were of diverse ancestries. Subject descriptions including subject sex, age, family, ethnicity and cardiac diseases were collected during recruitment. As previously described [1], we generated whole genome sequences from the blood or skin fibroblasts of the 181 subjects on the HiSeqX (Illumina; 150 bp-paired end).

iPSC derivation and somatic mutation analysis

As previously described, we reprogrammed fibroblast samples using non-integrative Cytotune Sendai virus (Life Technologies) and the 191 iPSCs (7 subjects had 2 or more clones each) were shown to be pluripotent and to have high genomic integrity with no or low numbers of somatic copy-number variants (CNVs) [19,34].

Large-scale derivation of iPSC-CVPC samples

To generate iPSC-derived cardiovascular progenitors (iPSC-CVPCs) we used a small molecule cardiac differentiation protocol [4]. The 25-day differentiation protocol consisted of five phases, the optimizations for each step are described in detail below: 1) *expansion*: we developed the ccEstimate algorithm to automate the detection of 80% confluency for iPSCs in T150 flasks; 2) *differentiation*: we tested whether increasing the dosage of IWP-2 to induce to inhibit the WNT pathway improved differentiation efficiency and found that 7.5 μM at D3 of the differentiation provided in a single dose for 48 hours results in the most efficient differentiation; 3) *purification*: since fetal cardiomyocytes use lactate as primary energy source and have a higher capacity for lactate uptake than other cell types [35,36], we incorporated lactate metabolic selection for five days to improve iPSC-CVPC purity [14]; 4) *recovery*: after metabolic selection, iPSC-CVPCs were maintained in cell culture for five days; and 5) *harvest*: we collected iPSC-CVPCs at D25 for downstream molecular assays and cryopreserved live cells.

The 232 attempted differentiations of the 191 iPSC lines were performed as follows:

Expansion of iPSC: One vial of each iPSC line was thawed into mTeSR1 medium containing 10 μM ROCK Inhibitor (Sigma) and plated on one well of a 6-well plate coated overnight with matrigel. During the expansion phase, all iPSC passaging was performed in mTeSR1 medium containing 5 μM ROCK inhibitor, when cells were visually estimated to be at 80% confluency. The iPSCs were passaged using Versene (Lonza) from one well into three wells of a 6-well plate. Next, the iPSCs were passaged using Versene onto three 10 cm dishes at $2.54x10^4$ per cm^2 density. The iPSCs molonalyer was plated onto three T150 flasks at the density of 3.66×10^4 per cm^2 using Accutase (Innovative Cell Technologies Inc.). Prior to expansion with Versene, after thaw iPSCs were passaged 1-2 times using Dispase II (20mg/ml; Gibco/Life technologies). iPSCs were at passage 22.7 ± 4.8 (range 17 to 44) at the monolayer stage (i.e., initiation of differentiation).

Differentiation: At 80% iPSC confluency (measured using ccEstimate, see section below "Estimation of optimal time for initiation of iPSC-CVPCs differentiation using ccEstimate") cell lysates were collected from 32 lines for RNA-seq data generation, where these iPSC and subsequent generated molecular data are referred to as D0 iPSC. After reaching 80% confluency (usually within 4-5 days), differentiation was initiated with the addition of the medium containing RPMI 1960 (gibco-life technologies) with Penicillin – Streptomycin (Gibco/Life Technologies) and B-27 Minus Insulin (Gibco/Life Technologies) (hereafter referred to as RPMI Minus supplemented with 12μM CHIR-99021 (D0). After 24h of exposure to CHIR-99021, medium was changed to RPMI Minus (D1). On D3 medium was changed to 1:1 mix of spent and fresh RPMI Minus

supplemented with 7.5μM IWP-2 (Tocris). On D5, after 48h of exposure to IWP-2, the medium was change to RPMI Minus. On D7, medium was changed to RPMI 1960 with Penicillin – Streptomycin (Gibco/Life Technologies) and B-27 Supplement 50X (hereafter referred to as RPMI Plus) (Gibco/Life Technologies). Between D7 and D13, RPMI Plus medium was changed every 48h.

Purification: On D15 the cells were collected from the flask using Accutase and plated onto fresh T150 flasks at confluency 1-1.3 x 10_6 per cm_2. On D16, cells were washed with PBS without Ca^{2+} and Mg^{2+} (Gibco/Life Technologies) and medium was changed for RPMI 1960 no glucose (Gibco/Life Technologies) supplemented with Non-Essential Amino Acids (Gibco/Life Technologies), L-Glutamine (Gibco/Life Technologies), Penicillin-Streptomycin 10,000U (Gibco/Life Technologies) and 4mM Sodium L-Lactate (Sigma) in 1M HEPES (Gibco/Life Technologies). Medium supplemented with lactate was changed on D17 and D19.

Recovery: On D21 cells were washed with PBS and medium was changed for RPMI Plus. On D23 medium was again changed for RPMI Plus. The first beating cells were usually observed between D7 and D9 and as early as D7 (immediately after the media change) and robust beating was usually observed between D8 and D11. During the lactate selection iPSC-CVPC were beating robustly less than 16 hours after reseeding. For all successfully derived iPSC-CVPCs on D25, total-cell lysate material was collected and frozen for downstream RNA-seq assays.

Harvest: On D25 cells were collected using Accutase and processed for the following molecular material for downstream assays: 1) cell lysates (RNA-Seq); 2) permeabilized cells (ATAC-Seq); 3) live frozen cells (scRNA-seq); 4) cross-linked cells (ChIP-Seq, median number of vials/iPSC line = 3; ~1.0 x 10_7 cells/vial), and 5) dry cell pellets (methylation and protein). RNA-seq was generated from 180 iPSC-CVPC differentiations (149 lines from 139 subjects) that successfully reached D25.

Estimation of optimal time for initiation of iPSC-CVPCs differentiation using ccEstimate

Heterogeneity of growth rates across different iPSC lines could result in different confluency at the monolayer stage (i.e., faster growing lines will be more confluent) and hence impact differentiation outcome. To reduce the effects of the iPSC lines having different growth rates, we developed an automatic pipeline that analyzes images of monolayer-grown cells, determines their confluency and predicts when cells reach 80% confluency to initiate the differentiation protocol. Cell confluency estimates (ccEstimate) are performed by first dividing each T150 flask into 10 sections and acquiring images for each section every 24 hours after cells are plated as a monolayer. The final image is acquired immediately after treatment with CHIR, which occurs when their confluence is at least 80% (Day 0). The time required for cells to reach 80% confluence is estimated on the basis of the confluence curve derived for each section in each flask. To digitally measure iPSC confluency, ccEstimate performs image analysis using the EBImage package in R [37]. Images are read using the readImage function. As lighting may be different between the center and the border of an image, only the central part of the image

is retained. To separate cells from the background and calculate confluence (i.e. the fraction of the surface of the flask that is covered by cells) the following operations are performed:

1. The image is transformed to monochromatic by determining the intensity of each pixel as the average of the intensities of the red, green and blue channels.
2. Edges are sharpened using high-pass filter. The matrix used for this filter is 15x15 with values -1 on the diagonals and +28 in the center.
3. Contrasts are enhanced by multiplying the pixel intensities by 2.
4. Mean and standard deviation of the pixel intensities are calculated. The image is transformed from monochromatic to binary by setting all pixels with intensity more than two standard deviations higher than the mean to white (intensity = 1) and all other pixels to black (intensity = 0).
5. The resulting binary image is dilated using a disc-shaped structuring element with diameter 5 pixels.
6. 1,000 50x50 pixels sub-images are randomly selected. For each sub-image, the number of white pixels is calculated. Confluence is estimated as the fraction of the randomly selected sub-images with at least 50% of white pixels.

Confluency measurement data is collected for at least the first three days after plating as monolayer to train a generalized linear model (GLM) using the function glm in R to estimate when cells must be treated with CHIR. Estimation is performed separately for each flask section and CHIR is added to all three flasks associated to a given line when at least 75% of sections have confluence 80%.

Using ccEstimate, we could start differentiation at the same confluency level for each iPSC sample, thereby reducing or neutralizing the effects of different growth rates. On average, each sample required 4.23 ± 1.12 days to reach 80% confluency. The correlation between the number of days required to reach 80% confluency and the %CM population was -0.05, suggesting that iPSC growth rate does not affect differentiation outcome.

Flow cytometry

On D25 of differentiation, iPSC-CVPCs were stained with cTnT antibody, acquired using FACS and analyzed using FlowJo V10.2. Immunofluorescence analysis of iPSC-CVPCs Immunofluorescence was assessed in 5 iPSC-CVPC lines. Live frozen iPSC-CVPC harvested on D25 were thawed, plated for five days, fixed, permeabilized and incubated with antibodies.

Generation of RNA-seq data

For gene expression profiling of iPSCs, we used RNA-seq data from 184 samples (cell lysates collected between passages 12 to 40). For gene expression profiling of iPSC-CVPCs, we generated RNA-seq data from 180 samples at D25 differentiation. All RNA-seq samples were generated and analyzed using the same pipeline to obtain transcript per million bp (TPM) [1].

Generation of scRNA-seq data

To capture the full spectrum of heterogeneity among the iPSC-CVPCs, we selected eight samples with variable %cTnT (42.2 to 95.8%). After removing proliferating cells and doublets we obtained 34,905 cells.

CIBERSORT

Expression levels of the top 50 genes overexpressed in each of the three cell populations (total 150 genes) were used as input for CIBERSORT [16] to calculate the relative distribution of the three cell populations for the 180 iPSC-CVPC samples at D25.

Characterizing transcriptional similarities of iPSCs, iPSC-CVPCs and GTEx adult tissues

We performed principle component analysis on RNA-seq on 184 iPSCs, 180 iPSC-CVPCs and 1,072 RNA-seq samples from GTEx.

Determining optimal CM:EPDC ratio estimates from CIBERSORT to define iPSCs cardiac fates

To obtain the optimal threshold, we conducted a series of differential expression analyses on 15,228 autosomal genes in the 184 iPSC lines (147 completed and 37 terminated) considering the ratio of population frequencies at ten thresholds. The 30:70 (CM:EPDC) ratio resulted in the highest number of differentially expressed genes (84 genes with Storey q-value < 0.1, t-test), which is substantially greater than random expectation. Thus, we grouped the 184 iPSC lines into: 1) those that have CM fates, i.e.

produced iPSC-CVPC with >= 30% Population 1, and 2) those that have EPDC fates, i.e. produced iPSC-CVPC with > 70% Population 2.

Comparing the number of differentially expressed genes with random expectation

To determine if the number of significantly differentially expressed genes was higher than expected by chance, we shuffled the assignments of the 184 iPSC RNA-seq samples to differentiation fate (125 CM and 59 EPDC) 100 times.

Contribution of 91 signature genes in iPSCs to determination of cardiac fate

For each of the 91 signature genes, we built a GLM with the expression of the gene as input and the differentiation outcome (e.g. % Population 1) as output using a logit link function. To understand the cumulative contribution of all 91 signature genes on cardiac differentiation fate, we built a GLM with an L1 norm penalty using the expression of all 91 genes as input and the differentiation outcome as output using. To avoid overfitting the model, we used a 10-fold cross validation.

Detecting associations between genetic background and differentiation outcome

We obtained genotypes for 8,620,159 biallelic SNPs and short indels with allelic frequency >5% in the iPSCORE collection. Genotypes were obtained for each SNP in all individuals using *bcftools view* [38]. Linear regression was used to calculate the associations between the genotype of each variant and differentiation outcome (% CM population in the iPSC-CVPCs), using passage at monolayer and sex as covariates.

Gene set enrichment analysis using the MSigDB collection

We performed GSEA using the R *gage* package [39] on all MSigDB gene sets [20]. FDR correction was performed independently for each collection. The normalized mean expression difference between iPSCs that differentiated to CMs and iPSCs that differentiated to EPDCs was used as input for GSEA.

Associations between iPSC and subject features and differentiation outcome

A GLM was built in R using age, sex, ethnicity, age, and passage of the iPSCs at D0 of differentiation as input and differentiation outcome as output (0 = EPDCs; and 1 = CMs).

Identifying X chromosome inactivation in female iPSCs and iPSC-CVPCs

To analyze X chromosome inactivation, we used 113 female iPSCs, of which 87 where CM-fated and 26 were EPDC-fated. We called ASE in RNA-Seq from iPSC and iPSC-CVPCs as previously described [1]. Genes lying in X chromosome pseudoautosomal (PAR) regions (PAR1: 60001- 2699520, PAR2: 154931044 – 155260560) were removed from analysis. We defined the strength of ASE for each gene as the fraction of RNA transcripts that were estimated to originate from the allele with higher expression (referred to as allelic imbalance fraction, AIF).

Validation of findings in Yoruba iPSC set

The Yoruba iPSCs [12] were generated from LCLs using episomal reprogramming. Differentiation was performed using a small molecular method and iPSC-CMs were

harvested on D31 or D32. 15 lines successfully generated iPSC-CMs and 24 were terminated on or before day 10. We downloaded RNA-seq for 34 of the Yoruba iPSC and 13 iPSC-CM samples from Gene Expression Omnibus (GEO; GSE89895), as well as 297 samples from 19 distinct iPSCs in a timecourse experiment (day 0-15) performed on the same Yoruba iPSC samples [22]. RNA-seq was aligned using STAR, gene expression was quantified using the RSEM package and normalized to TPM. The RNA-seq for the 13 Yoruba iPSC-CMs and from all timecourse time points were analyzed using CIBERSORT similar to the iPSCORE samples.

1.6 Data and software availability

Accession numbers for the RNA-seq data, scRNA-seq, and WGS genotypes are dbGaP: phs00924 and phs001325. The 191 iPSC lines are available through WiCell Research Institute: https://www.wicell.org/; NHLBI Next Gen Collection.

1.7 Acknowledgements

This work was supported by a CIRM grant GC1R-06673-B, NSF-CMMI division award 1728497 and NIH grants HG008118, HL107442, DK105541, DK112155. MKRD and JPN were supported by T15LM011271. WWG was supported by F31HL142151.

1.8 Author information

KAF, ADC, MKRD, and MD conceived the study. ADC and KF performed iPSC-CVPC differentiations. ADC, FC generated molecular data. SH generated IF

images. MCW generated Yoruba iPSC-CMs. FS and ADC generated scRNA-seq data. MKRD, WWG, HM, ENS, JPN and MD performed data processing and computational analyses. MP, EA, and KAF oversaw the study. MKRD, MD, and KAF prepared the manuscript.

Chapter 1, in full, is an adapted reprint of the material as it appears in Stem Cell Reports, 2019, Margaret K.R. Donovan, Agnieszka D'Antonio-Chronowska, William W. Greenwald, Jennifer Phuong Nguyen, Kyohei Fujita, Sherin Hashem, Hiroko Matsui, Francesca Soncin, Mana Parast, Michelle C. Ward, Florence Coulet, Erin N. Smith, Eric Adler, Matteo D'Antonio, Kelly A. Frazer. The dissertation author was one of the primary investigators and authors of this paper.

Chapter 2: Cellular deconvolution of GTEx tissues powers eQTL studies to discover thousands of novel disease and cell-type associated regulatory variants

2.1 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 human 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 and skin as a proof-of-concept tissues, we show that readily available signature genes based on expression profiles of mouse cell types can be used to deconvolute the cellular composition of human GTEx tissues. We then deconvoluted 6,829 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. Conducting eQTL analyses for GTEx liver and skin samples using cell type composition estimates as interaction terms, we identified thousands of novel genetic associations that had lower effect sizes and were cell-type-associated. We further show that cell-type-associated 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 study provides a framework to estimate the relative fractions of different cell types in GTEx tissues using signature genes from mouse cell types and functionally characterize human genetic variation that impacts gene expression in a cell-type-specific manner.

2.2 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[17] 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, unaccounted for 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 can affect genotype-gene expression associations[40]. Since 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, thus characterization of cellular heterogeneity across all GTEx tissues is critical for more comprehensive eQTL studies. It is possible that future studies pursuing cell-type-associated 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-associated gene expression would greatly benefit from conducting eQTL analyses on bulk GTEx tissue samples whose cellular heterogeneity has been characterized through existing deconvolution methods[16,41,42].

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 signature genes needed to deconvolute a heterogeneous tissue can be obtained by analyzing scRNA-seq generated from an analogous tissue. However, there are relatively few human scRNA-seq resources currently available[43-47], and thus only a small fraction of GTEx tissues could be deconvoluted using gene expression signatures derived from existing human single-cell data. While human single-cell data is limited, the Tabula Muris exists[48], 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 gene expression signatures to integrate scRNA-seq data between the two species[49], raising the possibility of utilizing the available scRNA-seq from mouse to generate the gene expression signatures for deconvolution of GTEx tissues.

To examine the feasibility of using mouse-derived gene expression signatures to deconvolute human tissues, we compared cellular composition estimates of GTEx liver and GTEx skin 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 gene signatures to deconvolute the 175 GTEx liver samples and the 860 GTEx skin samples resulted in highly correlated estimated cellular compositions. We show that the main differences between the cell types identified using the human-derived

versus mouse-derived signature genes were due to: 1) subtle biological differences that exist in human and mouse immune cells, and 2) resolution (i.e., the ability to detect less abundant cell types and distinguish between similar cell types) which was impacted by technical differences in the human and mouse scRNA-seq data sets, including the number of cells captured and subjected to scRNA-seq and the spatial location from which the tissue was sampled. We used gene signatures derived from the Tabula Muris resource to deconvolute 6,829 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 (i.e. not detected using bulk RNA-seq samples) genetic associations that tended to have lower effect sizes, some of which are cell-type-associated. Finally, we show that skin cell-type-associated 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 demonstrates two major principles: 1) mouse-derived signature genes can be used to deconvolute the cellular composition of human tissues; and 2) the estimation of cellular heterogeneity by deconvolution enhances the genetic insights yielded from the GTEx resource.

2.3 Results

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

To examine the extent to which scRNA-seq generated from analogous human and mouse tissues (Table 2.1) captured similar cell types, we first examined liver as a proof-of-concept tissue (Figure 2.1a, "proof-of-concept"). We used previously defined cell types from Tabula Muris mouse liver cells (which were purified for viable hepatocyte and non-parenchymal cells followed by FACS sorting; 710 cells; 5 cell types)[48], and to be consistent, we used the Tabula Muris annotation approach to analyze existing human liver scRNA-seq data (total liver homogenate; 8,119 cells; 15 cell types)[43]. 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 [43].

Table 2.1 Mapping of Tabula Muris scRNA-seq tissues/organs used to deconvolute human GTEx tissues

GTEx issue	Tabula Muris organ
Aorta	Heart (aorta subset)
Artium	Heart (left and right atrium subset)
Bladder	Bladder
Amygdala	Brain-nonmicroglia
Anterior cinglulat cortex (BA24)	Brain-nonmicroglia
Caudate (basal ganlia)	Brain-nonmicroglia
Cerebellar Hemisphere	Brain-nonmicroglia
Cerebellum	Brain-nonmicroglia
Cortex	Brain-nonmicroglia
Frontal cortex (BA9)	Brain-nonmicroglia
Hippocampus	Brain-nonmicroglia
Hypothalamus	Brain-nonmicroglia
Nucleus accumbens (basal ganglia)	Brain-nonmicroglia
Putamen (basal ganglia)	Brain-nonmicroglia
Spinal cord (cervical c-1)	Brain-nonmicroglia
Substantia nigra	Brain-nonmicroglia
Colon - Sigmoid	Colon
Colon - Transverse	Colon
Adipose - Subcutaneous	Fat
Adipose - Visceral (Omentum)	Fat
Kidney - Cortex	Kidney
Liver	Liver
Mammary	Mammary
Skeletal muscle	Muscle
Pancreas	Pancreas
Skin - not sun exposed	Skin
Skin - sun exposed	Skin
Spleen	Spleen
Ventricle	Heart (left and right ventricle subset)

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 2.1b-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) can be influenced by 1): the number of cells captured and subjected to scRNA-seq, which may influence the proportion of observed common or rare cell types[50]; and 2) how the tissue was sampled, which may enrich for selected populations or capture how populations are distinguished by spatial location (i.e. zonation). Some of the 15 cell types identified in the human liver scRNA-seq were highly similar and clustered near each other, for example there were four hepatocytes populations and two endothelial cell populations (human periportal sinusoidal endothelial cells (SEC) and central venous SECs) distinguished by their zonation (Figure 2.1b,c). In contrast, for the mouse liver scRNA-seq, which had considerably fewer cells analyzed, we only observed one hepatocyte population and one endothelial population (Figure 2.1d,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 2.1b,f), which largely corresponded to the 5 cell types from mouse liver scRNA-seq (cholangiocytes and hepatic stellate cells were absent due to having been sorted by FACS; Figure 2.1d-f). Overall, these results show that scRNA-seq generated from human and mouse liver captured similar cell types and that technical differences, including the number of cells analyzed and tissue sampling methodology, affects the cell type resolution.

2.3.2 Mouse liver signature genes can estimate cellular composition of human liver samples

To establish the ability to use expression profiles of signature genes derived from mouse scRNA-seq for the deconvolution of human GTEx tissues, we first examined if the similarly annotated cell types identified in the two species (Figure 2.1b-e) clustered together based on their gene expression profiles. We harmonized the human and mouse liver scRNA-seq using canonical correlation analysis (CCA) and visualized using uniform manifold approximation and projection (UMAP) (Figure 2.2a,b). We observed that the corresponding cell types across the two species clustered closely together, indicating that they had highly similar gene expression profiles.

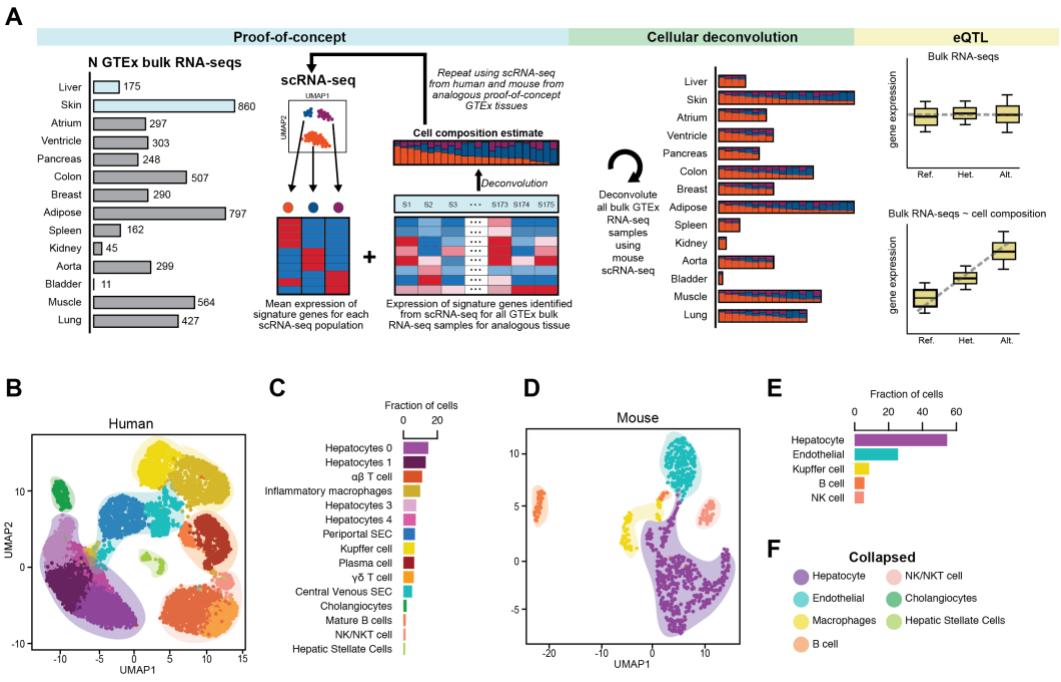
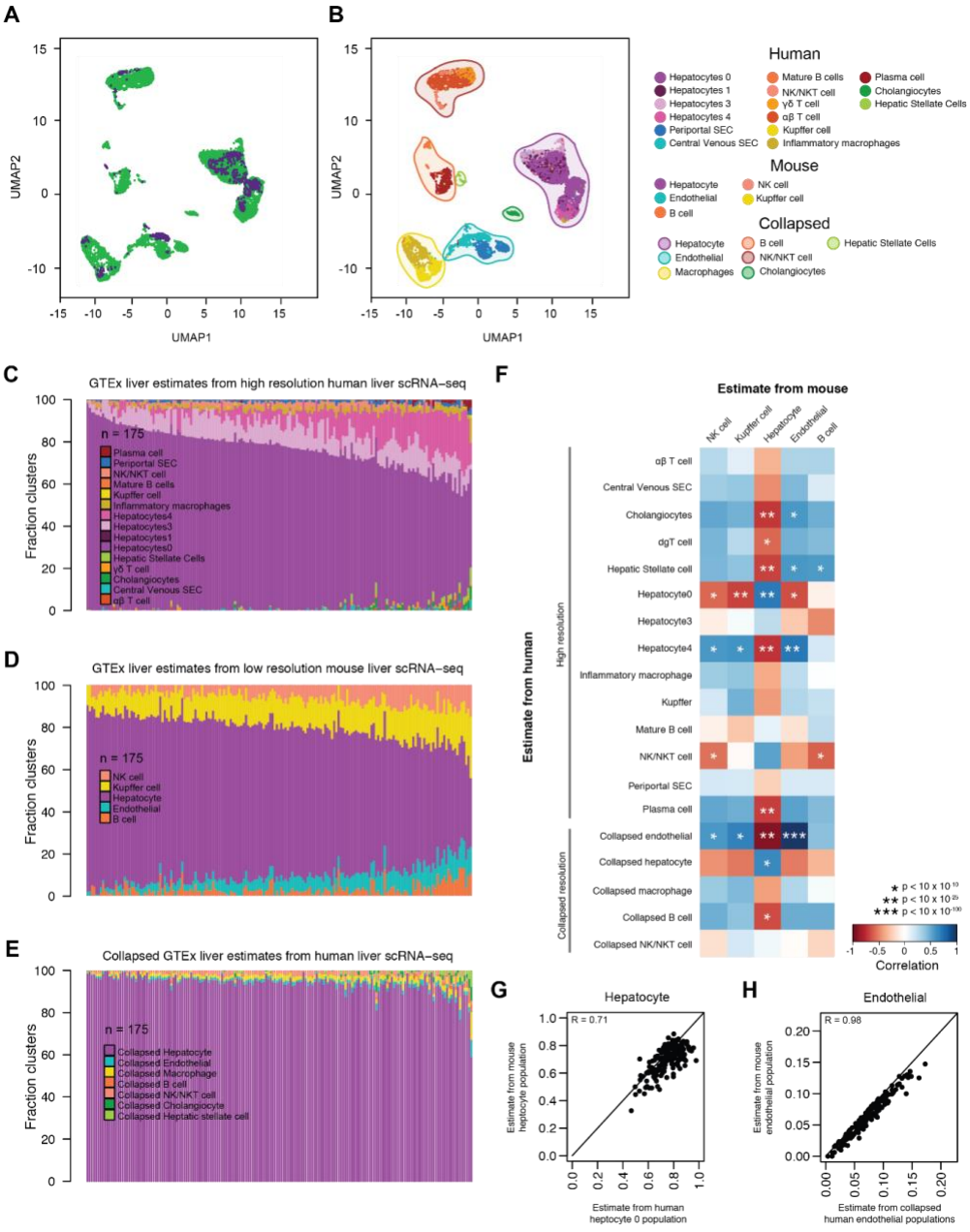


Figure 2.1 Human and mouse liver scRNA-seq contains similar cell types
(a) Overview of the study design. We first conducted proof-of-concept analyses, where we compared cellular estimates of two proof-of-concept GTEx tissues (liver and skin) after having deconvoluted each using both mouse and human signature genes obtained from 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 type composition estimates as interaction terms for eQTL analyses to determine if we could detect novel cell-type-associated 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 2.1c). Similar cell types can be collapsed to single cell type classifications and are noted with colored, transparent shading (Figure 2.1f). (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 2.1f. (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 2.1e). Each cell type has a corresponding collapsed cell type in human liver and is noted with colored, transparent shading (Figure 2.1f). (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 2.1d. (f) Legend showing the colors of collapsed similar cell types from human liver (transparent shading in UMAP Figure 2.1b,d). 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).

Figure 2.2 Comparison of GTEx liver cell estimates using mouse versus human signature gene

(a) UMAP plot of integrated scRNA-seq data from human and mouse liver. Each point represents a single cell and color coding of cells indicates the species the cells were obtained from (human = green; mouse = purple). (b) UMAP plot of integrated scRNA-seq data from human and mouse liver. Each point represents a single cell and color coding of cell type populations are shown in the adjacent legend. The collapsed populations are the same as those shown in Figure 2.1f. (c,d,e) Bar plots showing the fraction of cell types estimated in the 175 GTEx liver RNA-seq samples deconvoluted using gene expression profiles from high resolution human liver scRNA-seq (C), low resolution mouse liver scRNA-seq (D), and GTEx estimates generated by collapsing high resolution human cell types within each of the 7 distinct cell classes (E). (f) 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 correlation in estimates, to blue, indicating positive correlation in estimates. Significance is indicated with asterisks. (g, h) 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.


A
B
UMAP1
UMAP2
Human
Hepatocytes 0
Hepatocytes 1
Hepatocytes 3
Hepatocytes 4
Periportal SEC
Central Venous SEC
Mature B cells
NK/NKT cell
γδ T cell
αβ T cell
Kupffer cell
Inflammatory macrophages
Plasma cell
Cholangiocytes
Hepatic Stellate Cells
Mouse
Hepatocyte
Endothelial
B cell
NK cell
Kupffer cell
Collapsed
Hepatocyte
Endothelial
Macrophages
B cell
NK/NKT cell
Cholangiocytes
Hepatic Stellate Cells
C
GTEx liver estimates from high resolution human liver scRNA−seq
Fraction clusters
n = 175
Plasma cell
Periportal SEC
NK/NKT cell
Mature B cells
Kupffer cell
Inflammatory macrophages
Hepatocytes4
Hepatocytes3
Hepatocytes1
Hepatocytes0
Hepatic Stellate Cells
γδ T cell
Cholangiocytes
Central Venous SEC
αβ T cell
D
GTEx liver estimates from low resolution mouse liver scRNA−seq
n = 175
NK cell
Kupffer cell
Hepatocyte
Endothelial
B cell
E
Collapsed GTEx liver estimates from human liver scRNA−seq
n = 175
Collapsed Hepatocyte
Collapsed Endothelial
Collapsed Macrophage
Collapsed B cell
Collapsed NK/NKT cell
Collapsed Cholangiocyte
Collapsed Heptatic stellate cell
F
Estimate from mouse
NK cell
Kupffer cell
Hepatocyte
Endothelial
B cell
Estimate from human
High resolution
Collapsed resolution
αβ T cell
Central Venous SEC
Cholangiocytes
dgT cell
Hepatic Stellate cell
Hepatocyte0
Hepatocyte3
Hepatocyte4
Inflammatory macrophage
Kupffer
Mature B cell
NK/NKT cell
Periportal SEC
Plasma cell
Collapsed endothelial
Collapsed hepatocyte
Collapsed macrophage
Collapsed B cell
Collapsed NK/NKT cell
★ p < 10 x 10⁻¹⁰
★★ p < 10 x 10⁻²⁵
★★★ p < 10 x 10⁻¹⁰⁰
Correlation
G
Hepatocyte
R = 0.71
Estimate from mouse heptocyte population
Estimate from human heptocyte 0 population
H
Endothelial
R = 0.98
Estimate from mouse endothelial population
Estimate from collapsed human endothelial populations

We next compared the cellular composition estimates of 175 GTEx bulk liver RNA-seq samples[17] obtained by deconvolution using human signature genes to those obtained using mouse signature genes (Figure 2.1a, "proof-of-concept"), which respectively consisted 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). From the 175 GTEx bulk liver RNA-seq samples, we independently extracted the expression of the signature genes at the two resolutions, and used CIBERSORT[16] to estimate the cellular compositions (i.e. high resolution human liver estimates and low resolution mouse liver estimates) (Figure 2.2c,d). 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 (Figures 2.1b,f and 2.2e). 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 2.2f). We found that hepatocyte estimates from mouse liver were positively and highly correlated with the human high resolution hepatocyte 0 population estimate ($r = 0.71$, p-value = 5.4×10^{-28}), but not correlated with any of the other three high resolution hepatocyte populations (1, 3 and 4); and was slightly less correlated with the collapsed hepatocyte population estimate ($r = 0.64$, p-value = 1.015×10^{-21}) (Figure 2.2f,g). This indicates that the low resolution mouse hepatocyte population corresponds to one of the four human hepatocyte

populations/zones potentially due to tissue sampling. Further, we observed that the endothelial estimates from mouse were highly correlated with the collapsed human endothelial population estimates ($r = 0.98$, p-value = 1.2×10^{-115}) but not correlated with either high resolution human periportal SECs or central venous SECs (Figure 2.2h). 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), which is potentially due to the lower number of mouse cells analyzed. 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 2.2f). This difference in immune cell estimates in GTEx liver is not wholly unexpected, as biological differences, including immune response differences, exist between species [51]. To further examine the accuracy of the deconvolution, we conducted simulations to obtain 100 human liver samples with known cell type distributions, and confirmed that the estimated cell population distributions obtained using both human and mouse gene expression signatures were consistent with their expected values (Figure 2.3a,b). Our results show that, while technical differences in scRNA-seq generation and biological differences between humans and mice may impact cell estimation performance, overall mouse signature genes can be used to deconvolute human GTEx bulk RNA-seq samples.

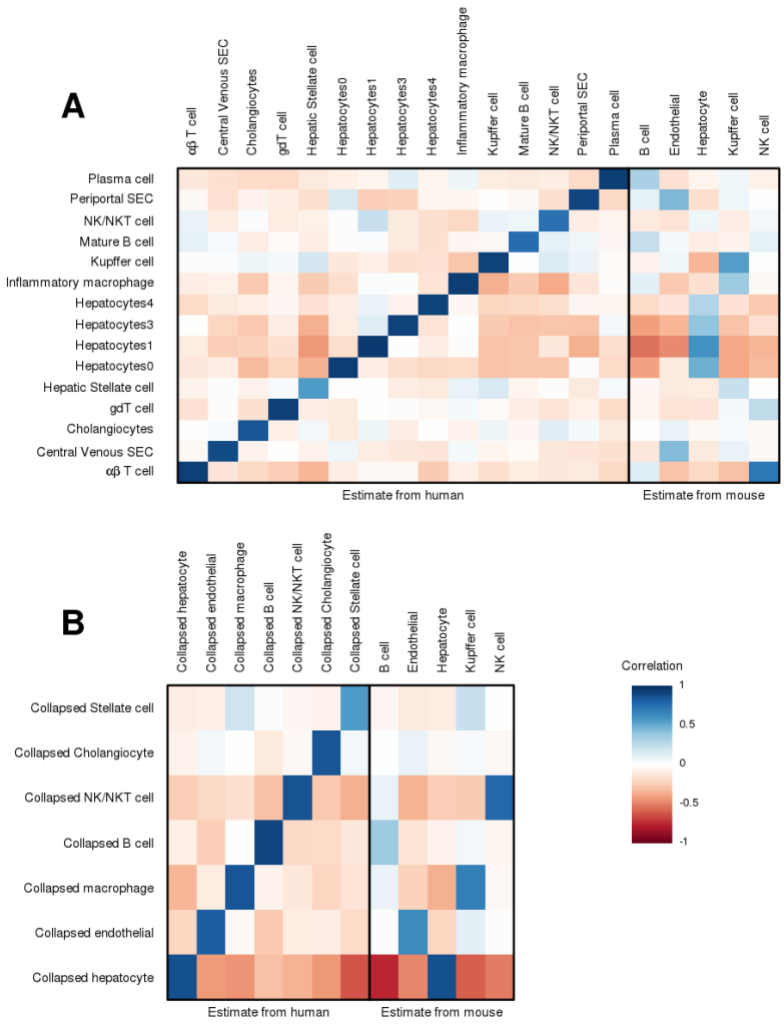


Figure 2.3 Testing the accuracy of deconvolution using simulated samples of known cell type distributions

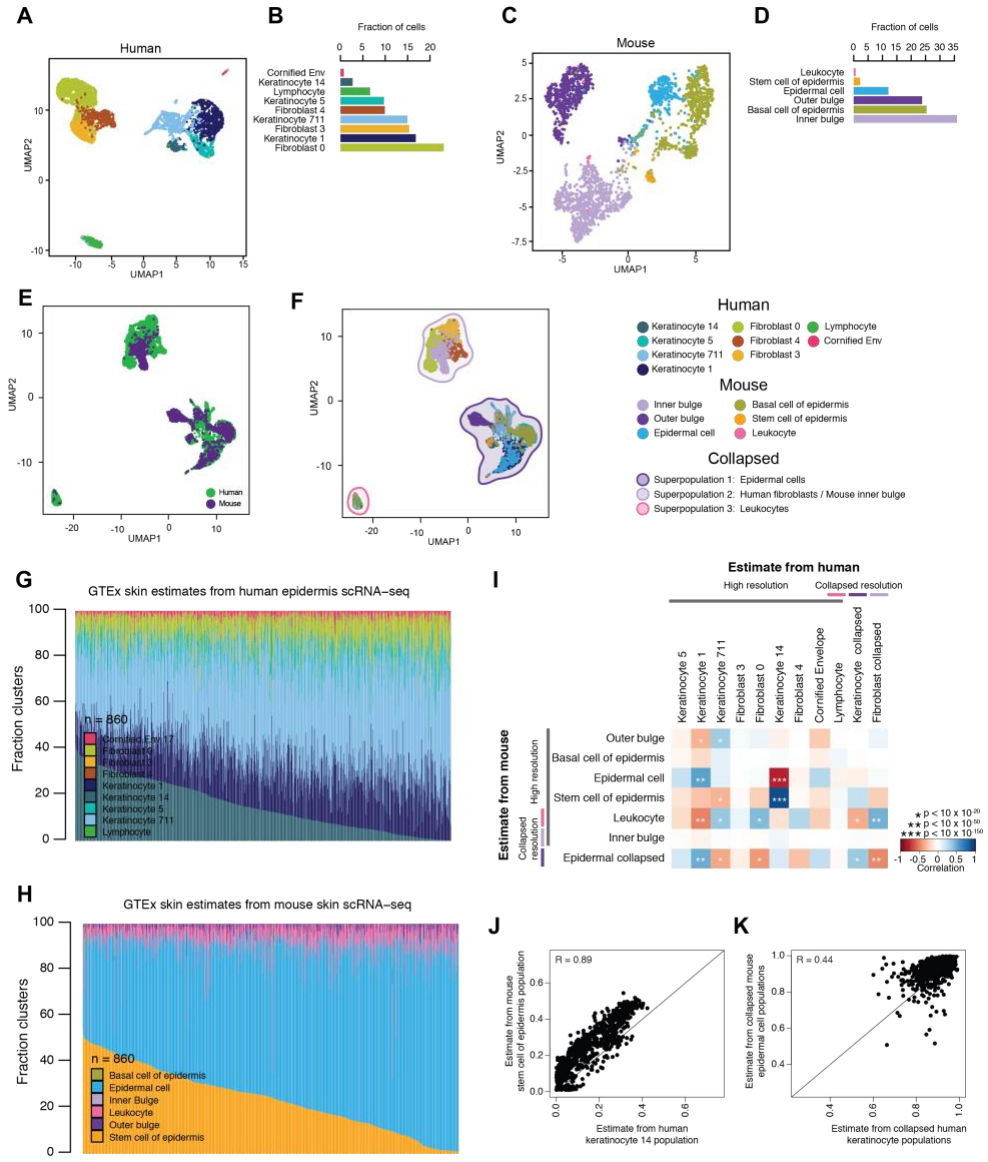
Plots showing the correlation between 100 simulated samples with known cell type distributions cell types (rows) and cell type estimates from human and mouse cell types (columns) for (a) all cell types and (b) collapsed cell types. For most cell types, we observed high correlation ($r > 0.6$) between the known cell type distributions and the population estimates obtained by deconvolution using human and mouse expression signatures. There were two exceptions: 1) human stellate cells ($r = 0.572$ between human estimates and known cell type distributions), likely because their gene expression signatures were derived from only 79 cells; and 2) B-cells ($r = 0.363$ between mouse estimates and known cell type distributions), consistent with between-species immune system differences.

2.3.3 Deconvolution of GTEx skin confirms mouse signature genes can estimate cellular composition

To examine the similarity of cellular estimates across 860 GTEx human skin samples obtained using human-derived versus mouse-derived signature genes, we used scRNA-seq from human epidermal cells[52] (digested dorsal forearm skin biopsies; 5,670 cells; 9 cell types) (Figure 2.4a,b) and Tabula Muris mouse skin cells (FACS sorted epidermal keratinocytes; 2,263 cells; 6 cell types) (Figure 2.4c,d). While the previous human and mouse liver scRNA-seq studies[43,48] used similar naming conventions for the cell type annotations (Figure 2.1b-e), the human and mouse skin scRNA-seq studies[48,52] did not (Figure 2.4a-d), and thus we first needed to identify the corresponding cell types across the two species. To accomplish this, we harmonized the human dermis and mouse skin scRNA-seq using CCA (Figure 2.4e,f) and visualized using UMAP. We observed three distinct superpopulations: 1) superpopulation 1, epidermal cells, consisting of the four human keratinocyte populations (14, 5, 711, and 1) and mouse epidermal cells, basal cells, stem cells of epidermis, and outer bulge cells (keratinocyte stem cells), 2) superpopulation 2, consisting of the three human fibroblast populations (0, 3, and 4) and mouse inner bulge cells (keratinocyte stem cells); and 3) superpopulation 3, leukocytes, consisting of human lymphocytes and mouse leukocytes (Figure 2.4e,f). Further, the different cell types within each of the three clusters expressed corresponding marker genes, confirming that they indeed were similar cell types in the human and mouse skin scRNA-seq. Overall, we found human and mouse skin scRNA-seq captured shared cell types that cluster into three distinct superpopulations.

Figure 2.4 Testing the accuracy of deconvolution using simulated samples of known cell type distributions

(a) UMAP plot of clustered scRNA-seq data from human epidermis. Each point represents a single cell and color-coding of cell type populations are shown adjacent (Figure 2.4b). (b) Bar plots showing the fraction of each cell type from the scRNA-seq data from human epidermis. Color-coding of cell types correspond to the colors of the single cells in Figure 2.4a. (c) UMAP plot of clustered scRNA-seq data from mouse skin. Each point represents a single cell and color-coding of cell type populations are shown adjacent in Figure 2.4d. (d) Bar plots showing the fraction of each cell type from the scRNA-seq data from mouse skin. Color-coding of cell types correspond to the colors of the single cells in Figure 2.4c. (e) UMAP plot of integrated scRNA-seq data from human epidermis and mouse skin. Each point represents a single cell and color-coding of cells indicates the species the cells were obtained from (human = green; mouse = purple). (f) UMAP plot of integrated scRNA-seq data from human epidermis and mouse skin. Each point represents a single cell and color coding of cell type populations and collapsed superpopulations are shown in the adjacent legend. (g,h) Bar plots showing the fraction of cell types estimated in GTEx skin RNA-seq samples from human epidermis scRNA-seq (g) and mouse skin scRNA-seq (h). (i) Heatmap showing the correlation of GTEx skin cell population estimates from mouse skin scRNA-seq at high and collapsed resolutions (rows) and human skin (columns). 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. (j,k) Scatter plots of estimated cell compositions across 860 GTEx skin samples deconvoluted using human scRNA-seq for human keratinocyte 14 population versus mouse stem cell of epidermis population (j) and keratinocyte 1, 5, 14, 711 population versus collapsed mouse epidermal cell populations (k).


A
Human
UMAP2
UMAP1
B
Fraction of cells
0 5 10 15 20
Cornified Env
Keratinocyte 14
Lymphocyte
Keratinocyte 5
Fibroblast 4
Keratinocyte 711
Fibroblast 3
Keratinocyte 1
Fibroblast 0
C
Mouse
UMAP2
UMAP1
D
Fraction of cells
0 5 10 15 20 25 30 35
Leukocyte
Stem cell of epidermis
Epidermal cell
Outer bulge
Basal cell of epidermis
Inner bulge
E
UMAP2
UMAP1
Human
Mouse
F
UMAP2
UMAP1
Human
Keratinocyte 14
Keratinocyte 5
Keratinocyte 711
Keratinocyte 1
Fibroblast 0
Fibroblast 4
Fibroblast 3
Lymphocyte
Cornified Env
Mouse
Inner bulge
Outer bulge
Epidermal cell
Basal cell of epidermis
Stem cell of epidermis
Leukocyte
Collapsed
Superpopulation 1: Epidermal cells
Superpopulation 2: Human fibroblasts / Mouse inner bulge
Superpopulation 3: Leukocytes
G
GTEx skin estimates from human epidermis scRNA-seq
Fraction clusters
n = 860
Lymphocyte
H
GTEx skin estimates from mouse skin scRNA-seq
Fraction clusters
n = 860
Basal cell of epidermis
Epidermal cell
Inner Bulge
Leukocyte
Outer bulge
Stem cell of epidermis
I
Estimate from human
High resolution
Collapsed resolution
Keratinocyte 5
Keratinocyte 1
Keratinocyte 711
Fibroblast 3
Fibroblast 0
Keratinocyte 14
Fibroblast 4
Cornified Envelope
Lymphocyte
Keratinocyte collapsed
Fibroblast collapsed
Estimate from mouse
Outer bulge
Basal cell of epidermis
Epidermal cell
Stem cell of epidermis
Leukocyte
Inner bulge
Epidermal collapsed
Correlation
J
R = 0.89
Estimate from mouse stem cell of epidermis population
Estimate from human keratinocyte 14 population
K
R = 0.44
Estimate from collapsed mouse epidermal cell populations
Estimate from collapsed human keratinocyte populations

We next compared the cellular composition estimates of 860 GTEx bulk skin RNA-seq samples[17] obtained by deconvolution using human gene expression signatures to mouse gene expression signatures (Figure 2.1a, "proof-of-concept"). We obtained signature genes for each cell type identified in scRNA-seq from human skin (i.e. signature genes from each of 9 dermis cell types) and mouse skin (i.e. signature genes from each of 6 skin cell types) and used CIBERSORT to deconvolute the 860 GTEx skin RNA-seqs (Figure 2.4g,h). Given the presence of the three superpopulation clusters observed in the mouse and human scRNA-seq integration analysis (Figure 2.4f), similar to liver, we investigated how resolution impacted the correlation between human and mouse signature gene estimates. We independently collapsed the high resolution human epidermis (9 cell types) and the high resolution mouse skin (6 cell types), by summing the estimates across the cell types in each of the three distinct superpopulations. We then calculated all pairwise-correlations between each of the estimated cell populations in the 860 GTEx skin samples from human estimates (high and collapsed) with the estimated cell populations from mouse (high and collapsed resolution) (Figure 2.4i). Using the integration analysis (Figure 2.4f) as a guide, we examined the similarity of estimates from human and mouse cell populations mapping to each of the three superpopulations. First, we examined the similarity of human cell types in Superpopulation 1 (Keratinocyte 14, Keratinocyte 5, Keratinocyte 711, Keratinocyte 1, cornified envelope, and collapsed estimates of these cell types) and mouse cell types in Superpopulation 1 (epidermal cell, basal cell, stem cell of epidermis, outer bulge, and collapsed estimates of these cell types) (Figure 2.4f; dark purple shading). We observed the human keratinocyte population 14 had a strong positive correlation with the mouse stem cell of the epidermis estimates (R =

0.89; $p = 2.4 \times 10^{-103}$) (Figure 2.4i,j). We also found that collapsed mouse epidermal cell estimates were correlated with collapsed human keratinocyte population estimates ($R = 0.44$, $p = 1.19 \times 10^{-43}$) (Figure 2.4i,k). These results indicate that despite differences in annotations, estimates from mouse and human cell types mapping to the epidermal cell superpopulation are highly correlated. Second, we examined the similarity of human cell types in superpopulation 2 (fibroblast 0, fibroblast 3, fibroblast 4, and collapsed estimates of these cell types) and the single mouse cell type (inner bulge) in this cluster (Figure 2.4f; light purple shading). We found that human fibroblast (high resolution and collapsed) estimates were not correlated with the mouse inner bulge cell population estimates (Figure 2.4i), indicating that, despite similar enough global gene expression patterns for the human fibroblasts and mouse inner bulge cells to cluster together, their signature genes distinguish them as different cell types during deconvolution. Third, we examined the similarity of the human cell type (lymphocyte) and mouse cell type (leukocyte) in superpopulation 3 (Figure 2.4f; pink shading). Similar to the liver estimates, mouse and human leukocyte estimates were not correlated (Figure 2.4i), likely due to known species differences in immune cells. As we observed in liver, we confirmed that technical and biological differences influence cell estimate performance, however overall cell composition estimates derived from human and mouse skin signature genes are correlated, supporting our ability to use mouse scRNA-seq as an alternative to human scRNA-seq for the deconvolution of GTEx tissues.

2.3.4 Cellular deconvolution of GTEx tissues reveals surprising levels of heterogeneity

To understand the extent to which the mouse signature genes obtained from cell types across 14 tissues were able to distinguish between the 28 GTEx tissues, we extracted the expression of the signature genes (Table 2.1) across the 6,829 bulk GTEx RNA-seq samples and visualized how the samples clustered (Figure 2.5a). We observed that the mouse signature genes were able to differentiate between the human GTEx organs, as well as illustrated the existence of organ substructures delineating heterogeneity in tissues belonging to the same organ. For example, tissues from the same organ clustered closely together and distinctly from other organs, including the heart tissues (atrium and ventricle), brain tissues (cortex, frontal cortex, hippocampus, anterior cingulate cortex, amygdala, substantia nigra, spinal cord, putamen, nucleus accumbens, caudate, and hypothalamus), adipose (visceral and subcutaneous), and colon (sigmoid and transverse). Of note, within the brain we also observe clustering according to zonation, including clustering of samples from the cerebellum and cerebellar hemisphere, as well as clustering of samples from the frontal cortex, cortex, and anterior cingulate cortex. These results suggest that signature gene capture both organ differences, as well differences that exist in tissues from the same organ that hint at tissue substructures driven by sample heterogeneity.

To understand the cellular heterogeneity of 28 GTEx tissues (Figure 2.1a, "cellular deconvolution"), we used the signature genes from 14 mouse tissue types (Table 2.1) to perform cellular deconvolution of 28 GTEx tissues from 14 organs (Figure 2.5b; Table 2.1), where the number of samples for each GTEx tissue varied from 11 (bladder)

to 860 (skin). We found that all samples were well-deconvoluted (P-value < 0.001; CIBERSORT, 1,000 permutations) and that each deconvoluted GTEx tissue contained a variable number of cell types ranging from two (bladder) to seven (brain and heart) (Figure 2.5c). In ~30% of the tissues (9 out of 28), we found that not all mouse cell types were estimated, possibly due to the GTEx tissues having been isolated for bulk RNA-seq from a different spatial location than mouse or species differences in cell types. Additionally, the relative distribution of the estimated cell types varied between different samples of the same tissue (Figure 2.5d). Tissues with the least heterogeneous cell population distributions between samples were aorta and spleen, whereas those with the most heterogeneous cell population distributions between samples were brain (13 tissues), colon, and left ventricle. 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.

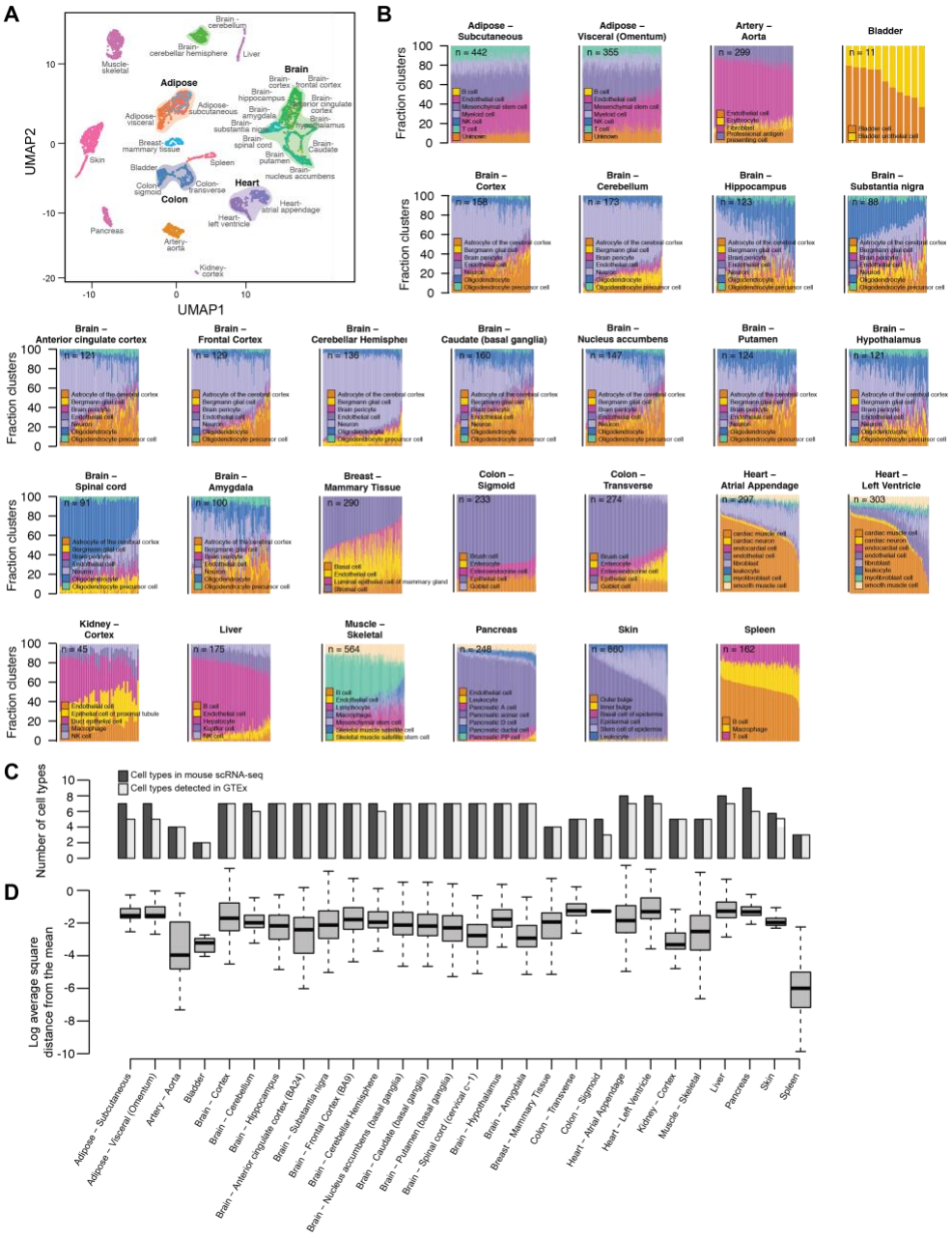


Figure 2.5 Cellular deconvolution of 28 GTEx tissues
(a) UMAP using the expression of all scRNA-seq derived signature genes across the 28 GTEx tissues. (b) Stacked bar plots showing the fraction of cell types estimated in GTEx RNA-seq samples from mouse scRNA-seq. (c) Bar plots comparing the number of cell types discovered in mouse scRNA-seq (light grey) vs. the number of these cell types that were estimable for each GTEx tissue. (d) 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.

2.3.5 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 associations (Figure 2.1a). We identified 19,621 expressed genes in GTEx liver samples and performed one eQTL analysis not considering cellular heterogeneity (i.e. bulk resolution), and three eQTL analysis using cell population estimates as covariates to adjust for cellular heterogeneity: 1) considering high resolution human liver estimates (15 cell types; Figure 2.2a); 2) considering collapsed resolution human liver estimates (7 cell types; Figure 2.2c); and 3) considering low resolution mouse liver estimates (5 cell types; 2.2b). Using cell population estimates as covariates we detected many more genes with significant eQTLs (eGenes) than at bulk resolution (Figure 2.6a). 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.

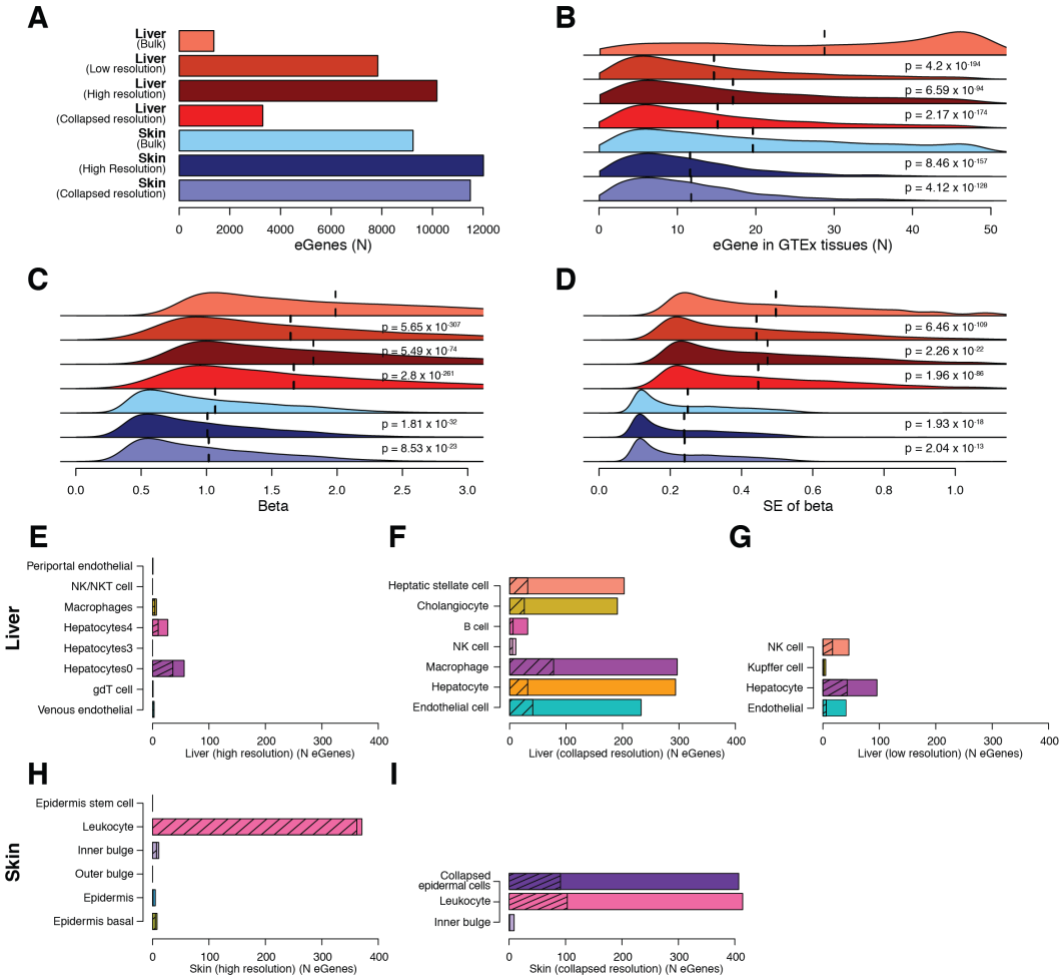


Figure 2.6 Using cellular deconvolution to discover cell-type-associated eQTLs
(a) Bar plot showing the number of eGenes detected in each eQTL analysis from liver (shades of red) and skin (shades of blue). (b,c,d) Distributions of (b) number of GTEx tissues where each eGene has significant eQTLs, (c) 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,f,g,h,i) Bar plots showing the number of eGenes significantly associated with each cell population considering cell estimates for: liver high resolution (e), liver collapsed resolution (f), liver low resolution (g), skin high resolution (h), and skin collapsed resolution (i). 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 had more tissue-associated 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-associated than eGenes detected at bulk resolution. Compared to bulk resolution, high resolution eGenes were the most tissue-associated ($p = 4.2x10^{-194}$; 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 2.6b), showing that the resolution of cell population estimates used as covariates is correlated with the power of the study to identify tissue-associated eGenes.

Furthermore, using cell populations as covariates resulted in decreased effect size (β) (Figure 2.6c) and standard error (SE) of β (Figure 2.6d), 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 ($r > 0.975$). By performing a permutation test, we excluded that the detection of a larger number of eQTLs using cell populations was due to the fact that a larger number of covariates was used. These results indicate that using cell population distributions as covariates overall reduces the noise, thereby potentially increasing our power to identify eQTLs.

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

To examine if some of the eQTLs identified using cell population estimates as covariates were cell-type-associated, we used a statistical interaction test[53-55] to assess if modeling the contribution of a 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, low, and collapsed resolutions we respectively detected 74, 121, and 528 cell-type-associated eGenes (i.e., eGene is more significant considering cell type estimates but associated with one or more cell type(s); FDR-corrected p-values < 0.1, χ_2 test, Figure 2.6e-g) and 54, 68, and 220 cell-type-specific eGenes (i.e. eGene is associated with only one cell type; Figure 2.6e-g). We investigated if relative cell abundance influenced our ability to detect cell-type-associated eGenes (i.e., is there more power for high abundance cells) and determined that it did not play a factor. Further, we noted by 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 2.6e-g). While initially counter-intuitive to the previous evidence showing higher resolution eGenes are more tissue-specific (Figure 2.6b) and have decreased noise (Figure 2.6c,d), it is possible we identify a greater number of cell-type-associated eGenes using low resolution cell population estimates due to prevention of the dilution of eQTL signals between shared cell types, as might occur in cases where a regulatory variant has similar

effects across 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.

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

To further investigate the impact of using cell populations on power to identify novel eGenes and cell-type-associated eQTLs, we conducted eQTL analyses using the GTEx tissue (skin), which includes the largest number of RNA-seq samples (Figure 2.4b). Although we deconvoluted 860 skin RNA-seqs using signature genes from high resolution mouse skin scRNA-seq (6 cell types; Figure 2.4b), only 749 had corresponding genotypes from 510 distinct individuals. We identified 24,029 expressed genes in the 749 skin RNA-seq samples with corresponding genotypes and performed three eQTL analyses: 1) without considering cell population distributions (bulk resolution); 2) considering high resolution mouse skin cell estimates (6 cell types; Figure 2.4c); and 3) considering collapsed resolution mouse skin cell estimates (3 cell types; Figure 2.4c,f). Using cell (high and collapsed) population distributions as covariates, respectively, we detected a 30% and 24% increase in eGenes with significant eQTLs (12,011 and 11,497 compared with 9,232, Figure 2.6a). Similar to our observation in liver, we found that eGenes specific for the eQTL analysis performed using high and collapsed cell populations as covariates, respectively, had eQTLs in fewer tissues than eGenes detected at bulk resolution ($p = 8.46 \times 10^{-157}$; $p = 4.12 \times 10^{-128}$, Mann Whitney U test; Figure 2.6b), had a decreased effect size β ($p = 1.81 \times 10^{-32}$; $p = 8.53 \times 10^{-23}$, Mann Whitney U test,

Figure 2.6c), and had decreased standard error (SE) of β ($p = 1.93 \times 10^{-18}$; $p = 2.04 \times 10^{-13}$, Mann Whitney U test, Mann Whitney U test; Figure 2.6d). We also observed that the β values for the top hit for each eGene were highly correlated between eQTLs detected using high and collapsed cell populations and eQTLs detected without using cell populations ($r = 0.994$; $r = 0.996$). Further, at high resolution we detected 384 cell-type-associated eGenes (FDR-corrected p-values < 0.1, χ^2 test, Figure 2.6h) and 375 cell-type-specific eGenes (FDR-corrected p-values < 0.1, χ^2 test, Figure 2.6h), which were predominantly associated with leukocytes, while at collapsed resolution we detected 511 cell-type-associated eGenes (FDR-corrected p-values < 0.1, χ^2 test, Figure 2.6i) and 220 cell-type-specific eGenes (FDR-corrected p-values < 0.1, χ^2 test, Figure 2.6i), associated with both the collapsed epidermal cell population and leukocytes (Superpopulations 1 and 3; Figure 2.4f). We hypothesize that substantially fewer cell-type-specific associations were observed in the high resolution epidermal cell types (epidermal cell, basal cell, stem cell of epidermis, outer bulge; Figure 2.6h) compared with the collapsed epidermal cells (Figure 2.6i), because of a dilution of signal between similar cell types. The relatively large number of cell-type-associated eGenes in skin compared with the liver could be reflective of sample size differences between the two tissue (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-associated differential regulation.

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

To explore the functional impact of the cell-type-associated eQTLs identified in skin, we examined their overlap with GWAS signals 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 2.7a) likely played a role in the traits 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. As the three collapsed skin superpopulations identified the most cell-type-associated eGenes, we performed colocalization of the eQTLs identified using the collapsed resolution cell estimates and skin GWAS loci to identify shared causal variants using coloc[56] and examining instances with PP4 > 0.5 (PP4, posterior probability of the colocalization model having one shared causal variant). We identified 394 variants that showed evidence of colocalization. These results show that we could identify hundreds of skin eQTLs that likely share a causal variant with skin GWAS traits.

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 and melanomas were enriched for eQTLs associated with keratinocyte stem cells from the inner bulge ($p = 1.13 \times 10^{-3}$, $p = 2.82 \times 10^{-4}$ Fisher's Test; Figure 2.7b,c), and infections were enriched for eQTLs associated with leukocytes ($p = 9.69 \times 10^{-3}$ Fisher's Test; Figure 2.7d). We did not observe a significant enrichment of cell-type-associated eQTLs in ulcers (Figure 2.7e), congenital

malformations (Figure 2.7f), 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.

We next sought to specifically examine the eGenes that most strongly colocalized with malignant neoplasms or melanoma (PP4 $\geq$ 0.8), as bulge stem cells have been implicated in playing a role in cancer[57-62]. We found six eGenes not previously associated with skin cancers with eQTLs significantly associated with inner bulge stem cells, including: 1) *BRIX1*, which has been found to play a role in cancer progression[63]; 2) *RP11-875011*.1, an antisense gene, which has not previously been implicated in cancer, however antisense genes are thought to contribute to the regulation of human cancers[64] ; 3) *MUL1*, which has been associated with the progression of human head and neck cancer[65]; 4) *PMS2P3*, has been implicated in affecting survival in pancreatic cancer[66]; 5) *FTH1*, which has been shown to be involved in regulating tumorigenesis[67,68] and whose increased expression in keratinocytes may be in response to stress[69,70]; and 6) *CNTN2*, which is involved in cell adhesion and has been implicated in tumor development[71,72]. The identification of these disease-associated eGenes supports our ability to identify cell-type-associated eQTLs whose functions are congruent with playing a role in the etiology of cancer. 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.

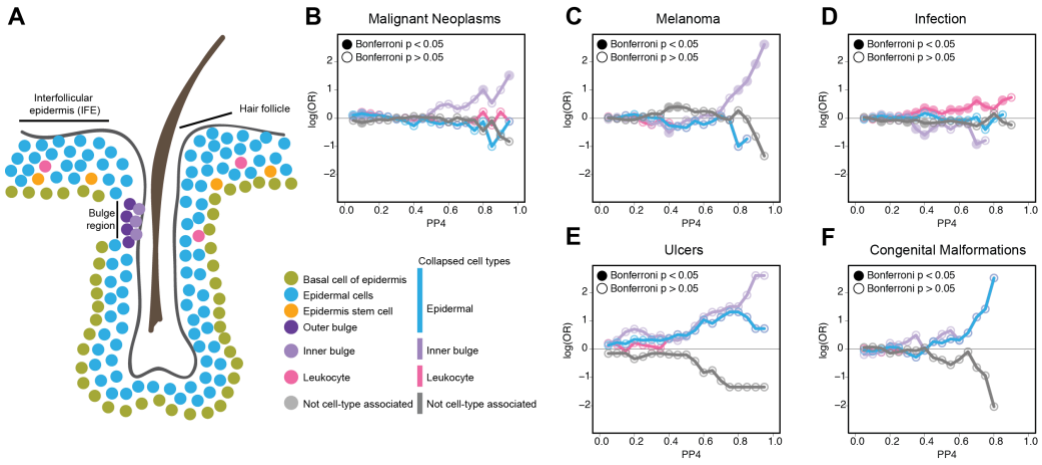


Figure 2.7 Using Colocalization of cell-type-associated skin eQTLs with skin GWAS traits
(a) Cartoon describing the approximate organization of cell types identified in scRNA-seq from skin. Colors used for each cell type are used throughout Figure and described in the adjacent legend. (b,c,d,e,f) Line plots showing the enrichment of cell-type-associated eQTLs in various GWAS traits: malignant neoplasms (b), melanoma (c), infection (d), ulcers (e), and congenital malformations (f). 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 Figure 2.7a.

2.4 Discussion

Human scRNA-seq data representative of all tissues in GTEx that could be used to deconvolute the 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[48], we sought to determine if mouse signature genes obtained from scRNA-seq could be used as an alternative for human signature genes for cellular deconvolution of GTEx RNA-seq samples. Using scRNA-seq from both mouse and human for two proof-of-concept tissues (liver and skin), we derived signature genes and used these expression profiles to deconvolute GTEx liver and skin RNA-seq samples. In general, human and mouse estimates between the two proof-of-concept tissues were comparable, where discrepancies in cell composition estimates between the two species primarily resulted from technical and subtle immunological differences. Specifically, in both liver and skin, technical differences impact the resolution at which cellular composition can be estimated, including: 1) the number of cells captured and subjected to scRNA-seq; and 2) tissue sampling methodology. Further, differences in cell composition estimates for immune cells were observed most likely due to immunological differences between the two species. These differences highlight that high resolution scRNA-seq (more cells/cell types sampled from diverse zones) is key to identifying and estimating the composition of highly specialized and rare cell types. For these reasons, the cell composition estimates we obtained from CIBERSORT using mouse-derived signature genes from proof-of-concept liver and skin scRNA-seq may still be missing cell types not captured in the Tabula Muris resource. An additional challenge we found that influenced our ability to compare cell composition estimates was the scRNA-seq cell annotations in human and

mouse skin did not use consistent naming conventions, thus it was not immediately clear how to compare cell estimates across the studies. We were able to overcome this challenge by integrating the mouse and human scRNA-seq, which allowed us to infer three similar superpopulations of cell types across the two species based on gene expression.

To examine cellular heterogeneity across the GTEx resource, we used the signature genes obtained using scRNA-seq from 14 mouse tissue types to deconvolute 6,829 GTEx RNA-seq samples mapping to 28 tissues from 14 organs. We found that GTEx tissues exhibit substantial cellular heterogeneity, with the number of cell types ranging from two in bladder to seven in brain and heart. Additionally, some of the tissues, including brain, colon, and left ventricle, showed highly variable proportions of estimated cell types between samples, contributing to intra-tissue cellular heterogeneity. Together, these results reveal a source heterogeneity in GTEx tissues that has not been previously considered and may contribute to reduced power to detect eQTLs.

While genetic association studies performed by GTEx have identified a wealth of novel insights into how human genetics function across bulk tissues[17], these analyses have not considered how cellular heterogeneity can confound these studies through biasing or even masking cell-type-specific signals. We found that considering cellular heterogeneity significantly improved eQTL analyses by increasing power to detect lower effect size genetic associations, as well as by identifying cell-type-specific associations that were masked in analyses using bulk RNA-seq data from the same samples. Further, we found resolution of cell heterogeneity influenced eQTL results, where considering high resolution estimates identified substantially more eQTLs than using lower

resolutions (low resolution or collapsed resolution); however, high resolution cell estimates identified fewer cell-type-associated genetic associations than lower resolutions. It is possible this decrease in associations may be due to a dilution of signal between similar cell types. Our observations suggest these two resolutions should both be used to power eQTL analyses in complementary ways: 1) high resolution estimates to power association analyses to discover lower effect size eQTLs; and 2) collapsed resolution estimates to identify cell-type associated eQTLs. 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.

Overall, we demonstrate that while efforts to generate a resource of scRNA-seq data from human tissues[73] are in progress, QTL studies using human bulk RNA-seq data could utilize readily available mouse-derived signature genes to estimate cellular heterogeneity and optimize power to identify cell type-specific genetic associations. As the Tabula Muris resource does not represent all of the human GTEx tissues (28 of 53) it is possible that scRNA-seq resources from other mammalian species could be used to deconvolute the non-represented GTEx tissues. Our study further emphasizes that the straightforward approach of taking tissue heterogeneity into account when conducting genetic association studies has the potential to greatly expand our understanding of the functional impact of genetic variation on molecular and complex human traits.

2.5 Methods

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

Single cell transcriptome profiles from 14 mouse organs were used in this study[48]. 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) for specific populations (FACS method; SMART-Seq2 RNAseq libraries). We downloaded the normalized gene expression and annotated single-cell clusters from each organ as Seruat[49] R objects (https://figshare.com/articles/Robject_files_for_tissues_processed_by_Seurat/5821263/1) .

Processing of scRNA-seq from human liver

10X Genomics formatted BAM files from five human total liver homogenate samples[43] 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.

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

Analysis of scRNA-seq from human liver[43] were conducted following the same approach used to annotate mouse organs[48]. 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. Clustering was performed using a shared-nearest-neighbor graph of the significant PCs and single cells were visualized using Uniform Manifold Approximation and Projection (UMAP). Cell populations were then annotated based on the expression of known liver marker genes[48].

Collapsing liver cell population estimates

To collapse similar cell populations in GTEx liver samples, we examined the UMAP from high resolution human liver scRNA-seq (Figure 2.1b) and compared to the UMAP from low resolution mouse liver scRNA-seq (Figure 2.1c) to identify broader/lower resolution classifications of cell types present in the liver. We identified populations in the human liver scRNA-seq that were similar (e.g. Hepatocyte populations 0, 1, 3, and 4; Figure 2.1b) with a corresponding population in the mouse liver scRNA-seq (e.g. Hepatocyte; Figure 2.1c). For populations identified in human not present in mouse, we did not perform any collapsing.

Annotation of the cell populations present in skin scRNA-seq data

Human: Skin scRNA-seq[49] gene expression data and cell annotations for 8,388 cells were downloaded from http://dom.pitt.edu/rheum/centers-institutes/scleroderma/systemicsclerosiscenter/database/. Cells with fewer than 200 detected genes were filtered from the data. 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. Clustering was performed using a shared-nearest-neighbor graph of the significant PCs and single cells were visualized using Uniform Manifold Approximation and Projection (UMAP). The single cells were then annotated using provided cell annotations and validated using marker gene expression. As the human skin scRNA-seq contained cell types belonging to various layers of the skin, whereas the mouse scRNA-seq was enriched for epidermal cells, we extracted only the 5,670 human cells belonging to the epidermal layer of the skin. We then reanalyzed the subsetted data following the above methods by performing PCA, reclustering, and visualization using UMAP.

Mouse: Tabula Muris cell annotations were confirmed by examining marker gene expression for epidermal cells, basal cells of the epidermis ($Krt1^{High}$), stem cells of the epidermis ($Top2a^{High}$), leukocytes ($Lyz2^{High}$), and keratinocyte stem cells ($Cd34^{High}$). While Tabula Muris annotated a single keratinocyte stem cell population, we reannotated this population by distinguishing between: 1) inner bulge cell population exhibiting $Dkk3^{High}$ and $ITGA6^{Low}$ expression; and 2) outer bulge cell population exhibiting $Fgf18^{High}$ and $ITGA6^{High}$ expression.

Deconvolution of complex tissues using CIBERSORT

Identification of signature genes from single cell populations: For 16 scRNA-seq datasets from human liver, human skin, and scRNA-seq from 14 mouse organs, we obtained gene expression signatures for each annotated cell type and used as input into CIBERSORT[16] to estimate the cellular composition of GTEx adult tissues (Table 2.1). For each tissue, we identified differentially expressed genes using Seurat FindMarkers and then extracted the top 200 most significantly overexpressed genes (adjusted p-value < 0.05; average $\log_2$ fold change > 0.25) for each of the annotated scRNA-seq cell types (gene expression signatures). For signature genes obtained from mouse scRNA-seq, we converted the mouse genes to their human orthologs using the biomaRt database[74,75]. The final gene signature sets only included mouse signature genes that also had a human ortholog. For a given signature gene set: 1) if a mouse gene had more than one human ortholog, only one human ortholog was retained in final signature set; and 2) if different mouse genes corresponded to the same human ortholog, only unique human orthologs were retained in the final signature set.

Cell composition estimation: The mean expression levels of the signature genes were used as input for CIBERSORT to calculate the relative distribution of the cell populations of 28 GTEx tissues from 14 organs. 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 2.1). To determine

the cell types detected in GTEx compared to the cell types modeled from mouse, we classified a given cell type as estimable in GTEx as those with CIBERSORT estimates greater than 0.05% in more than 5% of RNA-seq samples from a given GTEx tissue. To estimate cellular heterogeneity across GTEx RNA-seq samples, heterogeneity was measured as the average square distance from the mean for each GTEx tissue. We further examined how time from death or withdrawal of life-support until each tissue sample was fixed/frozen (i.e. ischemic time) is associated with cellular heterogeneity and we did not observe a consistent trend between ischemic time and cellular heterogeneity. GTEx organs are defined as the regions from which tissues are sampled (variable name SMTS from sample attributes data table; phv00169239.v7.p2) and 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). For example, samples from the GTEx organ, colon, is comprised of two tissues: sigmoid colon and transverse colon.

Harmonization of human and mouse scRNA-seq

To harmonize scRNA-seq from human and mouse liver and skin, mouse genes for each tissue scRNA-seq dataset were first converted to their human orthologs using the BioMart database[74,75]. Mouse and human scRNA-seq were then harmonized by identifying genes that anchor the two datasets using Seurat FindIntegrationAnchors and using these anchors to integrate the datasets using Seurat IntegrateData. Integrated datasets were then visualized using UMAP and corresponding cell types were identified by examining overlap of mouse and human cells.

Detecting eQTLs using a linear mixed model

To detect eQTLs, we obtained gene TPMs for 153 liver bulk RNA-seq samples and 749 skin bulk RNA-seq 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 genes 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[19]. Since some subjects had two skin samples (one sun-exposed and one not sun-exposed), we employed a linear mixed model (LMM) for eQTL detection, using subject ID as random effect (1|subject_id). We fitted LMMs using the lme4 package (https://www.jstatsoft.org/article/view/v067i01/0) to detect eQTLs in skin, described in the following model[55]:

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

For liver, we used sex as random effect (1|sex) to fit an LMM analogous to the skin eQTL analysis method, described in the following model:

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

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.

Improved eQTL detection using cell population distributions as covariates

We repeated eQTL detection adding cellular compositions as covariates to the LMMs described above using the following model:

expression ~ genotype + covariates + cell_populations + (1|random)

The "cell_populations" term denotes the relative cell population distributions in each tissue and (1|random) is each tissue's random effect (1|subject_id in skin samples; 1|sex in liver samples, see: Detecting eQTLs using a linear mixed model).

Specifically, we conducted three eQTL analysis for the liver using human high resolution, human collapsed and mouse low resolution cell populations as covariates. Since several cell types were detected at very low frequency, we only used a subset of the cell types described in Figure 2.5: 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. We conducted two eQTL analysis for skin using mouse high and mouse collapsed resolution cell populations as covariates with the following cell populations: 1) for mouse high resolution: epidermis stem cell, leukocyte, inner bulge, outer bulge, epidermis, and epidermis basal cells; and 2) for mouse collapsed resolution: epidermal cells, leukocyte, and inner bulge cells.

Detecting cell-type-specific and cell-type-associated eQTLs

In order to detect eQTLs associated with one or more cell types, for each cell population we repeated the eQTL analyses described above (see: Improving eQTL detection using cell population distributions as covariates) by adding an interaction term between the genotype and each cell population ($cell_i$) estimate to the model (genotype:$cell_i$). Specifically, for each $cell_i$ estimate, we compared the following two models:

H_0: expression ~ genotype + covariates + cell_populations + (1|random)

H_1: expression ~ genotype + covariates + cell_populations + genotype:$cell_i$ + (1|random)

In both models (H_0 and H_1), the "cell_populations" term denotes the relative cell population distributions in each tissue and (1|random) is each tissue's random effect (1|subject_id in skin samples; 1|sex in liver samples, see: Detecting eQTLs using a linear mixed model). In H_1, "genotype:$cell_i$" is the interaction term between the genotype and $cell_i$ estimate.

To compare the two hypotheses (H_0 and H_1), we calculated the difference between the two models using ANOVA and obtained χ_2 p-values using the pbkrtest package. For each $cell_i$ estimate used as an interaction term in H_1, only eGenes that satisfied two requirements were considered to be associated with $cell_i$: a) Benjamini-Hochberg-adjusted χ_2 p-value <0.1; and b) $\Delta_{AIC} = AIC_{interaction} - AIC_{no\ interaction} < 0$ (i.e. "genotype:$cell_i$" interaction terms that significantly improve the eQTL model). If only one cell population improved the eQTL model, the eQTL was labeled "cell-type specific"; conversely, if more than one cell population improved the eQTL model, the

eQTL was labeled “cell type-associated”. We determined the impact of cell population abundance on power to detect cell-type-associated eGenes by examining the distribution of ß, standard error, and P-value for cell-type-associated-eQTLs from each cell population.

Permutation analysis of liver eQTLs

To test if the detection of more eQTLs using cell populations as covariates was due to improved accuracy of the linear mixed model estimation or was simply associated with an increased number of covariates, for each top hit (defined as the variant with the strongest p-value for each gene), we permuted the cell type distribution across samples, 1,000 times. We obtained the average p-value, beta and standard error of beta across all permutations and compared these values with the measured p-value, beta and standard error of beta for each gene using a paired t-test.

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

For each eGene in the skin eQTL analysis deconvoluted using cell type 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, where the traits were grouped into seven categories based on shared nomenclature in the trait descriptions: 1) malignant neoplasms; 2) melanoma; 3) infection; 4) ulcers; 5) congenital malformations of the skin; 6) other cancer (non-melanoma or malignant neoplasm); and 7) unspecified. For all the variants genotyped in both GTEx and UK BioBank, we used coloc V. 3.1[56] 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). Enrichment of the associations was calculated using a Fisher's Test at multiple PP4 thresholds (0 – 1; by 0.05 bins), where the contingency table consisted of two classifications: 1) *if* the variant was significantly cell-type-associated (FDR < 0.05); and 2) *if* the variant colocalized with the GWAS trait greater than each PP4 threshold.

2.6 Data Availability

Sequence data that support the findings of this study (all Figures) is available for human liver scRNA-seq (GSE11546); for human skin scRNA-seq (http://dom.pitt.edu/rheum/centers-institutes/scleroderma/systemicsclerosiscenter/database/); and for Tabula Muris mouse scRNA-seq (https://figshare.com/articles/Robject_files_for_tissues_processed_by_Seurat/5821263/1). Scripts to process, analyze, and generate Figures from the data is available at https://github.com/mkrdonovan/gtex_deconvolution. The source data underlying all Figures is available in the Source Data available online.

2.7 Acknowledgements

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

2.8 Author information

K.A.F., M.K.R.D., A.D.C., 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. and A.D.C. oversaw the study. M.K.R.D., M.D., and K.A.F. prepared the manuscript.

Chapter 2, in full, has been submitted for publication of the material as it may appear in Nature Communications, 2019, Margaret K.R. Donovan, Agnieszka D'Antonio-Chronowska, Matteo D'Antonio, Kelly A. Frazer. The dissertation author was the primary investigator and author of this paper.

References

1 DeBoever, C., Li, H., Jakubosky, D., Benaglio, P., Reyna, J., Olson, K. M., Huang, H., Biggs, W., Sandoval, E., D'Antonio, M., Jepsen, K., Matsui, H., Arias, A., Ren, B., Nariai, N., Smith, E. N., D'Antonio-Chronowska, A., Farley, E. K. & Frazer, K. A. Large-Scale Profiling Reveals the Influence of Genetic Variation on Gene Expression in Human Induced Pluripotent Stem Cells. *Cell Stem Cell* 20, 533-546 e537, doi:10.1016/j.stem.2017.03.009 (2017).

2 Kilpinen, H., Goncalves, A., Leha, A., Afzal, V., Alasoo, K., Ashford, S., Bala, S., Bensaddek, D., Casale, F. P., Culley, O. J., Danecek, P., Faulconbridge, A., Harrison, P. W., Kathuria, A., McCarthy, D., McCarthy, S. A., Meleckyte, R., Memari, Y., Moens, N., Soares, F., Mann, A., Streeter, I., Agu, C. A., Alderton, A., Nelson, R., Harper, S., Patel, M., White, A., Patel, S. R., Clarke, L., Halai, R., Kirton, C. M., Kolb-Kokocinski, A., Beales, P., Birney, E., Danovi, D., Lamond, A. I., Ouwehand, W. H., Vallier, L., Watt, F. M., Durbin, R., Stegle, O. & Gaffney, D. J. Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* 546, 370-375, doi:10.1038/nature22403 (2017).

3 Panopoulos, A. D., Smith, E. N., Arias, A. D., Shepard, P. J., Hishida, Y., Modesto, V., Diffenderfer, K. E., Conner, C., Biggs, W., Sandoval, E., D'Antonio-Chronowska, A., Berggren, W. T., Izpisua Belmonte, J. C. & Frazer, K. A. Aberrant DNA Methylation in Human iPSCs Associates with MYC-Binding Motifs in a Clone-Specific Manner Independent of Genetics. *Cell Stem Cell* 20, 505-517 e506, doi:10.1016/j.stem.2017.03.010 (2017).

4 Lian, X., Zhang, J., Azarin, S. M., Zhu, K., Hazeltine, L. B., Bao, X., Hsiao, C., Kamp, T. J. & Palecek, S. P. Directed cardiomyocyte differentiation from human pluripotent stem cells by modulating Wnt/beta-catenin signaling under fully defined conditions. *Nat Protoc* 8, 162-175, doi:10.1038/nprot.2012.150 (2013).

5 Burridge, P. W., Matsa, E., Shukla, P., Lin, Z. C., Churko, J. M., Ebert, A. D., Lan, F., Diecke, S., Huber, B., Mordwinkin, N. M., Plews, J. R., Abilez, O. J., Cui, B., Gold, J. D. & Wu, J. C. Chemically defined generation of human cardiomyocytes. *Nat Methods* 11, 855-860, doi:10.1038/nmeth.2999 (2014).

6 Dubois, N. C., Craft, A. M., Sharma, P., Elliott, D. A., Stanley, E. G., Elefanty, A. G., Gramolini, A. & Keller, G. SIRPA is a specific cell-surface marker for isolating cardiomyocytes derived from human pluripotent stem cells. *Nat Biotechnol* 29, 1011-1018, doi:10.1038/nbt.2005 (2011).

7 Witty, A. D., Mihic, A., Tam, R. Y., Fisher, S. A., Mikryukov, A., Shoichet, M. S., Li, R. K., Kattman, S. J. & Keller, G. Generation of the epicardial lineage from human pluripotent stem cells. *Nat Biotechnol* 32, 1026-1035, doi:10.1038/nbt.3002 (2014).

8 Mo, M. L., Li, M. R., Chen, Z., Liu, X. W., Sheng, Q. & Zhou, H. M. Inhibition of the Wnt palmitoyltransferase porcupine suppresses cell growth and downregulates the Wnt/beta-catenin pathway in gastric cancer. *Oncol Lett* 5, 1719-1723, doi:10.3892/ol.2013.1256 (2013).

9 Wang, X., Moon, J., Dodge, M. E., Pan, X., Zhang, L., Hanson, J. M., Tuladhar, R., Ma, Z., Shi, H., Williams, N. S., Amatruda, J. F., Carroll, T. J., Lum, L. & Chen, C. The development of highly potent inhibitors for porcupine. *J Med Chem* 56, 2700-2704, doi:10.1021/jm400159c (2013).

10 Bao, X., Lian, X., Hacker, T. A., Schmuck, E. G., Qian, T., Bhute, V. J., Han, T., Shi, M., Drowley, L., Plowright, A., Wang, Q. D., Goumans, M. J. & Palecek, S. P. Long-term self-renewing human epicardial cells generated from pluripotent stem cells under defined xeno-free conditions. *Nat Biomed Eng* 1, doi:10.1038/s41551-016-0003 (2016).

11 Hartman, M. E., Dai, D. F. & Laflamme, M. A. Human pluripotent stem cells: Prospects and challenges as a source of cardiomyocytes for in vitro modeling and cell-based cardiac repair. *Adv Drug Deliv Rev* 96, 3-17, doi:10.1016/j.addr.2015.05.004 (2016).

12 Banovich, N. E., Li, Y. I., Raj, A., Ward, M. C., Greenside, P., Calderon, D., Tung, P. Y., Burnett, J. E., Myrthil, M., Thomas, S. M., Burrows, C. K., Romero, I. G., Pavlovic, B. J., Kundaje, A., Pritchard, J. K. & Gilad, Y. Impact of regulatory variation across human iPSCs and differentiated cells. *Genome Res* 28, 122-131, doi:10.1101/gr.224436.117 (2018).

13 Lian, X., Bao, X., Zilberter, M., Westman, M., Fisahn, A., Hsiao, C., Hazeltine, L. B., Dunn, K. K., Kamp, T. J. & Palecek, S. P. Chemically defined, albumin-free human cardiomyocyte generation. *Nat Methods* 12, 595-596, doi:10.1038/nmeth.3448 (2015).

14 Tohyama, S., Hattori, F., Sano, M., Hishiki, T., Nagahata, Y., Matsuura, T., Hashimoto, H., Suzuki, T., Yamashita, H., Satoh, Y., Egashira, T., Seki, T.,

Muraoka, N., Yamakawa, H., Ohgino, Y., Tanaka, T., Yoichi, M., Yuasa, S., Murata, M., Suematsu, M. & Fukuda, K. Distinct metabolic flow enables large-scale purification of mouse and human pluripotent stem cell-derived cardiomyocytes. *Cell Stem Cell* 12, 127-137, doi:10.1016/j.stem.2012.09.013 (2013).

15 Iyer, D., Gambardella, L., Bernard, W. G., Serrano, F., Mascetti, V. L., Pedersen, R. A., Talasila, A. & Sinha, S. Robust derivation of epicardium and its differentiated smooth muscle cell progeny from human pluripotent stem cells. *Development* 142, 1528-1541, doi:10.1242/dev.119271 (2015).

16 Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., Hoang, C. D., Diehn, M. & Alizadeh, A. A. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* 12, 453-457, doi:10.1038/nmeth.3337 (2015).

17 Consortium, G. T., Laboratory, D. A., Coordinating Center -Analysis Working, G., Statistical Methods groups-Analysis Working, G., Enhancing, G. g., Fund, N. I. H. C., Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida, Biospecimen Collection Source Site, N., Biospecimen Collection Source Site, R., Biospecimen Core Resource, V., Brain Bank Repository-University of Miami Brain Endowment, B., Leidos Biomedical-Project, M., Study, E., Genome Browser Data, I., Visualization, E. B. I., Genome Browser Data, I., Visualization-Ucsc Genomics Institute, U. o. C. S. C., Lead, a., Laboratory, D. A., Coordinating, C., management, N. I. H. p., Biospecimen, c., Pathology, e, Q. T. L. m. w. g., Battle, A., Brown, C. D., Engelhardt, B. E. & Montgomery, S. B. Genetic effects on gene expression across human tissues. *Nature* 550, 204-213, doi:10.1038/nature24277 (2017).

18 Perez-Pomares, J. M., de la Pompa, J. L., Franco, D., Henderson, D., Ho, S. Y., Houyel, L., Kelly, R. G., Sedmera, D., Sheppard, M., Sperling, S., Thiene, G., van den Hoff, M. & Basso, C. Congenital coronary artery anomalies: a bridge from embryology to anatomy and pathophysiology--a position statement of the development, anatomy, and pathology ESC Working Group. *Cardiovasc Res* 109, 204-216, doi:10.1093/cvr/cvv251 (2016).

19 Panopoulos, A. D., D'Antonio, M., Benaglio, P., Williams, R., Hashem, S. I., Schuldt, B. M., DeBoever, C., Arias, A. D., Garcia, M., Nelson, B. C., Harismendy, O., Jakubosky, D. A., Donovan, M. K. R., Greenwald, W. W., Farnam, K., Cook, M., Borja, V., Miller, C. A., Grinstein, J. D., Drees, F., Okubo, J., Diffenderfer, K. E., Hishida, Y., Modesto, V., Dargitz, C. T., Feiring,

R., Zhao, C., Aguirre, A., McGarry, T. J., Matsui, H., Li, H., Reyna, J., Rao, F., O'Connor, D. T., Yeo, G. W., Evans, S. M., Chi, N. C., Jepsen, K., Nariai, N., Muller, F. J., Goldstein, L. S. B., Izpisua Belmonte, J. C., Adler, E., Loring, J. F., Berggren, W. T., D'Antonio-Chronowska, A., Smith, E. N. & Frazer, K. A. 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, doi:10.1016/j.stemcr.2017.03.012 (2017).

20 Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. & Mesirov, J. P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102, 15545-15550, doi:10.1073/pnas.0506580102 (2005).

21 Tukiainen, T., Villani, A. C., Yen, A., Rivas, M. A., Marshall, J. L., Satija, R., Aguirre, M., Gauthier, L., Fleharty, M., Kirby, A., Cummings, B. B., Castel, S. E., Karczewski, K. J., Aguet, F., Byrnes, A., Consortium, G. T., Laboratory, D. A., Coordinating Center -Analysis Working, G., Statistical Methods groups-Analysis Working, G., Enhancing, G. g., Fund, N. I. H. C., Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida, Biospecimen Collection Source Site, N., Biospecimen Collection Source Site, R., Biospecimen Core Resource, V., Brain Bank Repository-University of Miami Brain Endowment, B., Leidos Biomedical-Project, M., Study, E., Genome Browser Data, I., Visualization, E. B. I., Genome Browser Data, I., Visualization-Ucsc Genomics Institute, U. o. C. S. C., Lappalainen, T., Regev, A., Ardlie, K. G., Hacohen, N. & MacArthur, D. G. Landscape of X chromosome inactivation across human tissues. *Nature* 550, 244-248, doi:10.1038/nature24265 (2017).

22 Strober, B. J., Elorbany, R., Rhodes, K., Krishnan, N., Tayeb, K., Battle, A. & Gilad, Y. Dynamic genetic regulation of gene expression during cellular differentiation. *Science* 364, 1287-1290, doi:10.1126/science.aaw0040 (2019).

23 Bargehr, J., Ong, L. P., Colzani, M., Davaapil, H., Hofsteen, P., Bhandari, S., Gambardella, L., Le Novere, N., Iyer, D., Sampaziotis, F., Weinberger, F., Bertero, A., Leonard, A., Bernard, W. G., Martinson, A., Figg, N., Regnier, M., Bennett, M. R., Murry, C. E. & Sinha, S. Epicardial cells derived from human embryonic stem cells augment cardiomyocyte-driven heart regeneration. *Nat Biotechnol* 37, 895-906, doi:10.1038/s41587-019-0197-9 (2019).

24 Kim, K. Y., Hysolli, E., Tanaka, Y., Wang, B., Jung, Y. W., Pan, X., Weissman, S. M. & Park, I. H. X Chromosome of female cells shows dynamic changes in

status during human somatic cell reprogramming. *Stem Cell Reports* 2, 896-909, doi:10.1016/j.stemcr.2014.04.003 (2014).

25 Barakat, T. S., Ghazvini, M., de Hoon, B., Li, T., Eussen, B., Douben, H., van der Linden, R., van der Stap, N., Boter, M., Laven, J. S., Galjaard, R. J., Grootegoed, J. A., de Klein, A. & Gribnau, J. Stable X chromosome reactivation in female human induced pluripotent stem cells. *Stem Cell Reports* 4, 199-208, doi:10.1016/j.stemcr.2014.12.012 (2015).

26 Zhao, J., Cao, H., Tian, L., Huo, W., Zhai, K., Wang, P., Ji, G. & Ma, Y. Efficient Differentiation of TBX18(+)/WT1(+) Epicardial-Like Cells from Human Pluripotent Stem Cells Using Small Molecular Compounds. *Stem Cells Dev* 26, 528-540, doi:10.1089/scd.2016.0208 (2017).

27 Paik, D. T. & Wu, J. C. Simply derived epicardial cells. *Nat Biomed Eng* 1, doi:10.1038/s41551-016-0015 (2017).

28 Guadix, J. A., Orlova, V. V., Giacomelli, E., Bellin, M., Ribeiro, M. C., Mummery, C. L., Perez-Pomares, J. M. & Passier, R. Human Pluripotent Stem Cell Differentiation into Functional Epicardial Progenitor Cells. *Stem Cell Reports* 9, 1754-1764, doi:10.1016/j.stemcr.2017.10.023 (2017).

29 Tatler, A. L., Habgood, A., Porte, J., John, A. E., Stavrou, A., Hodge, E., Kerama-Likoko, C., Violette, S. M., Weinreb, P. H., Knox, A. J., Laurent, G., Parfrey, H., Wolters, P. J., Wallace, W., Alberti, S., Nordheim, A. & Jenkins, G. Reduced Ets Domain-containing Protein Elk1 Promotes Pulmonary Fibrosis via Increased Integrin alphavbeta6 Expression. *J Biol Chem* 291, 9540-9553, doi:10.1074/jbc.M115.692368 (2016).

30 Desai, K., Aiyappa, R., Prabhu, J. S., Nair, M. G., Lawrence, P. V., Korlimarla, A., Ce, A., Alexander, A., Kaluve, R. S., Manjunath, S., Correa, M., Srinath, B. S., Patil, S., Kalamdani, A., Prasad, M. & Sridhar, T. S. HR+HER2- breast cancers with growth factor receptor-mediated EMT have a poor prognosis and lapatinib downregulates EMT in MCF-7 cells. *Tumour Biol* 39, 1010428317695028, doi:10.1177/1010428317695028 (2017).

31 Bruck, T., Yanuka, O. & Benvenisty, N. Human pluripotent stem cells with distinct X inactivation status show molecular and cellular differences controlled

by the X-Linked ELK-1 gene. *Cell Rep* 4, 262-270, doi:10.1016/j.celrep.2013.06.026 (2013).

32 Diogenes, T. C. P., Mourato, F. A., de Lima Filho, J. L. & Mattos, S. D. S. Gender differences in the prevalence of congenital heart disease in Down's syndrome: a brief meta-analysis. *BMC Med Genet* 18, 111, doi:10.1186/s12881-017-0475-7 (2017).

33 Gittenberger-de Groot, A. C., Vrancken Peeters, M. P., Bergwerff, M., Mentink, M. M. & Poelmann, R. E. Epicardial outgrowth inhibition leads to compensatory mesothelial outflow tract collar and abnormal cardiac septation and coronary formation. *Circ Res* 87, 969-971 (2000).

34 D'Antonio, M., Benaglio, P., Jakubosky, D., Greenwald, W. W., Matsui, H., Donovan, M. K. R., Li, H., Smith, E. N., D'Antonio-Chronowska, A. & Frazer, K. A. Insights into the Mutational Burden of Human Induced Pluripotent Stem Cells from an Integrative Multi-Omics Approach. *Cell Rep* 24, 883-894, doi:10.1016/j.celrep.2018.06.091 (2018).

35 Fisher, D. J., Heymann, M. A. & Rudolph, A. M. Myocardial consumption of oxygen and carbohydrates in newborn sheep. *Pediatr Res* 15, 843-846 (1981).

36 Werner, J. C. & Sicard, R. E. Lactate metabolism of isolated, perfused fetal, and newborn pig hearts. *Pediatr Res* 22, 552-556, doi:10.1203/00006450-198711000-00016 (1987).

37 Pau, G., Fuchs, F., Sklyar, O., Boutros, M. & Huber, W. EBImage--an R package for image processing with applications to cellular phenotypes. *Bioinformatics* 26, 979-981, doi:10.1093/bioinformatics/btq046 (2010).

38 Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987-2993, doi:10.1093/bioinformatics/btr509 (2011).

39 Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D. & Woolf, P. J. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* 10, 161, doi:10.1186/1471-2105-10-161 (2009).

40 Glastonbury, C. A., Couto Alves, A., El-Sayed Moustafa, J. S. & Small, K. S. Cell-Type Heterogeneity in Adipose Tissue Is Associated with Complex Traits and Reveals Disease-Relevant Cell-Specific eQTLs. *Am J Hum Genet* 104, 1013-1024, doi:10.1016/j.ajhg.2019.03.025 (2019).

41 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, doi:10.1186/1471-2105-14-89 (2013).

42 Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., Ryu, J. H., Wagner, B. K., Shen-Orr, S. S., Klein, A. M., Melton, D. A. & Yanai, I. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst* 3, 346-360 e344, doi:10.1016/j.cels.2016.08.011 (2016).

43 MacParland, S. A., Liu, J. C., Ma, X. Z., Innes, B. T., Bartczak, A. M., Gage, B. K., Manuel, J., Khuu, N., Echeverri, J., Linares, I., Gupta, R., Cheng, M. L., Liu, L. Y., Camat, D., Chung, S. W., Seliga, R. K., Shao, Z., Lee, E., Ogawa, S., Ogawa, M., Wilson, M. D., Fish, J. E., Selzner, M., Ghanekar, A., Grant, D., Greig, P., Sapisochin, G., Selzner, N., Winegarden, N., Adeyi, O., Keller, G., Bader, G. D. & McGilvray, I. D. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat Commun* 9, 4383, doi:10.1038/s41467-018-06318-7 (2018).

44 Cheng, J. B., Sedgewick, A. J., Finnegan, A. I., Harirchian, P., Lee, J., Kwon, S., Fassett, M. S., Golovato, J., Gray, M., Ghadially, R., Liao, W., Perez White, B. E., Mauro, T. M., Mully, T., Kim, E. A., Sbitany, H., Neuhaus, I. M., Grekin, R. C., Yu, S. S., Gray, J. W., Purdom, E., Paus, R., Vaske, C. J., Benz, S. C., Song, J. S. & Cho, R. J. Transcriptional Programming of Normal and Inflamed Human Epidermis at Single-Cell Resolution. *Cell Rep* 25, 871-883, doi:10.1016/j.celrep.2018.09.006 (2018).

45 Crinier, A., Milpied, P., Escaliere, B., Piperoglou, C., Galluso, J., Balsamo, A., Spinelli, L., Cervera-Marzal, I., Ebbo, M., Girard-Madoux, M., Jaeger, S., Bollon, E., Hamed, S., Hardwigsen, J., Ugolini, S., Vely, F., Narni-Mancinelli, E. & Vivier, E. High-Dimensional Single-Cell Analysis Identifies Organ-Specific Signatures and Conserved NK Cell Subsets in Humans and Mice. *Immunity* 49, 971-986 e975, doi:10.1016/j.immuni.2018.09.009 (2018).

46 Young, M. D., Mitchell, T. J., Vieira Braga, F. A., Tran, M. G. B., Stewart, B. J., Ferdinand, J. R., Collord, G., Botting, R. A., Popescu, D. M., Loudon, K. W., Vento-Tormo, R., Stephenson, E., Cagan, A., Farndon, S. J., Del Castillo Velasco-Herrera, M., Guzzo, C., Richoz, N., Mamanova, L., Aho, T., Armitage, J. N., Riddick, A. C. P., Mushtaq, I., Farrell, S., Rampling, D., Nicholson, J., Filby, A., Burge, J., Lisgo, S., Maxwell, P. H., Lindsay, S., Warren, A. Y., Stewart, G. D., Sebire, N., Coleman, N., Haniffa, M., Teichmann, S. A., Clatworthy, M. & Behjati, S. Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science* 361, 594-599, doi:10.1126/science.aat1699 (2018).

47 Nguyen, Q. H., Pervolarakis, N., Blake, K., Ma, D., Davis, R. T., James, N., Phung, A. T., Willey, E., Kumar, R., Jabart, E., Driver, I., Rock, J., Goga, A., Khan, S. A., Lawson, D. A., Werb, Z. & Kessenbrock, K. Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. *Nat Commun* 9, 2028, doi:10.1038/s41467-018-04334-1 (2018).

48 Tabula Muris, C., Overall, c., Logistical, c., Organ, c., processing, Library, p., sequencing, Computational data, a., Cell type, a., Writing, g., Supplemental text writing, g. & Principal, i. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* 562, 367-372, doi:10.1038/s41586-018-0590-4 (2018).

49 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, doi:10.1038/nbt.4096 (2018).

50 Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat Rev Genet* 14, 618-630, doi:10.1038/nrg3542 (2013).

51 Hagai, T., Chen, X., Miragaia, R. J., Rostom, R., Gomes, T., Kunowska, N., Henriksson, J., Park, J. E., Proserpio, V., Donati, G., Bossini-Castillo, L., Vieira Braga, F. A., Naamati, G., Fletcher, J., Stephenson, E., Vegh, P., Trynka, G., Kondova, I., Dennis, M., Haniffa, M., Nourmohammad, A., Lassig, M. & Teichmann, S. A. Gene expression variability across cells and species shapes innate immunity. *Nature* 563, 197-202, doi:10.1038/s41586-018-0657-2 (2018).

52 Tabib, T., Morse, C., Wang, T., Chen, W. & Lafyatis, R. SFRP2/DPP4 and FMO1/LSP1 Define Major Fibroblast Populations in Human Skin. *J Invest Dermatol* 138, 802-810, doi:10.1016/j.jid.2017.09.045 (2018).

53 Westra, H. J., Arends, D., Esko, T., Peters, M. J., Schurmann, C., Schramm, K., Kettunen, J., Yaghootkar, H., Fairfax, B. P., Andiappan, A. K., Li, Y., Fu, J., Karjalainen, J., Platteel, M., Visschedijk, M., Weersma, R. K., Kasela, S., Milani, L., Tserel, L., Peterson, P., Reinmaa, E., Hofman, A., Uitterlinden, A. G., Rivadeneira, F., Homuth, G., Petersmann, A., Lorbeer, R., Prokisch, H., Meitinger, T., Herder, C., Roden, M., Grallert, H., Ripatti, S., Perola, M., Wood, A. R., Melzer, D., Ferrucci, L., Singleton, A. B., Hernandez, D. G., Knight, J. C., Melchiotti, R., Lee, B., Poidinger, M., Zolezzi, F., Larbi, A., Wang de, Y., van den Berg, L. H., Veldink, J. H., Rotzschke, O., Makino, S., Salomaa, V., Strauch, K., Volker, U., van Meurs, J. B., Metspalu, A., Wijmenga, C., Jansen, R. C. & Franke, L. Cell Specific eQTL Analysis without Sorting Cells. *PLoS Genet* 11, e1005223, doi:10.1371/journal.pgen.1005223 (2015).

54 Panousis, N. I., Bertsias, G. K., Ongen, H., Gergianaki, I., Tektonidou, M. G., Trachana, M., Romano-Palumbo, L., Bielser, D., Howald, C., Pamfil, C., Fanouriakis, A., Kosmara, D., Repa, A., Sidiropoulos, P., Dermitzakis, E. T. & Boumpas, D. T. Combined genetic and transcriptome analysis of patients with SLE: distinct, targetable signatures for susceptibility and severity. *Ann Rheum Dis* 78, 1079-1089, doi:10.1136/annrheumdis-2018-214379 (2019).

55 Alasoo, K., Rodrigues, J., Mukhopadhyay, S., Knights, A. J., Mann, A. L., Kundu, K., Consortium, H., Hale, C., Dougan, G. & Gaffney, D. J. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat Genet* 50, 424-431, doi:10.1038/s41588-018-0046-7 (2018).

56 Giambartolomei, C., Zhenli Liu, J., Zhang, W., Hauberg, M., Shi, H., Boocock, J., Pickrell, J., Jaffe, A. E., CommonMind, C., Pasaniuc, B. & Roussos, P. A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics* 34, 2538-2545, doi:10.1093/bioinformatics/bty147 (2018).

57 Lapouge, G., Youssef, K. K., Vokaer, B., Achouri, Y., Michaux, C., Sotiropoulou, P. A. & Blanpain, C. Identifying the cellular origin of squamous skin tumors. *Proc Natl Acad Sci U S A* 108, 7431-7436, doi:10.1073/pnas.1012720108 (2011).

58 Song, I. Y. & Balmain, A. Cellular reprogramming in skin cancer. *Semin Cancer Biol* 32, 32-39, doi:10.1016/j.semcancer.2014.03.006 (2015).

59 Jian, Z., Strait, A., Jimeno, A. & Wang, X. J. Cancer Stem Cells in Squamous Cell Carcinoma. *J Invest Dermatol* 137, 31-37, doi:10.1016/j.jid.2016.07.033 (2017).

60 Morris, R. J. A perspective on keratinocyte stem cells as targets for skin carcinogenesis. *Differentiation* 72, 381-386, doi:10.1111/j.1432-0436.2004.07208004.x (2004).

61 Ratushny, V., Gober, M. D., Hick, R., Ridky, T. W. & Seykora, J. T. From keratinocyte to cancer: the pathogenesis and modeling of cutaneous squamous cell carcinoma. *J Clin Invest* 122, 464-472, doi:10.1172/JCI57415 (2012).

62 Kamstrup, M. R., Gniadecki, R. & Skovgaard, G. L. Putative cancer stem cells in cutaneous malignancies. *Exp Dermatol* 16, 297-301, doi:10.1111/j.1600-0625.2007.00547.x (2007).

63 Fan, B., Dachrut, S., Coral, H., Yuen, S. T., Chu, K. M., Law, S., Zhang, L., Ji, J., Leung, S. Y. & Chen, X. Integration of DNA copy number alterations and transcriptional expression analysis in human gastric cancer. *PLoS One* 7, e29824, doi:10.1371/journal.pone.0029824 (2012).

64 Balbin, O. A., Malik, R., Dhanasekaran, S. M., Prensner, J. R., Cao, X., Wu, Y. M., Robinson, D., Wang, R., Chen, G., Beer, D. G., Nesvizhskii, A. I. & Chinnaiyan, A. M. The landscape of antisense gene expression in human cancers. *Genome Res* 25, 1068-1079, doi:10.1101/gr.180596.114 (2015).

65 Kim, S. Y., Kim, H. J., Kang, S. U., Kim, Y. E., Park, J. K., Shin, Y. S., Kim, Y. S., Lee, K. & Kim, C. H. Non-thermal plasma induces AKT degradation through turn-on the MUL1 E3 ligase in head and neck cancer. *Oncotarget* 6, 33382-33396, doi:10.18632/oncotarget.5407 (2015).

66 Dong, X., Li, Y., Hess, K. R., Abbruzzese, J. L. & Li, D. DNA mismatch repair gene polymorphisms affect survival in pancreatic cancer. *Oncologist* 16, 61-70, doi:10.1634/theoncologist.2010-0127 (2011).

67 Chan, J. J., Kwok, Z. H., Chew, X. H., Zhang, B., Liu, C., Soong, T. W., Yang, H. & Tay, Y. A FTH1 gene:pseudogene:microRNA network regulates

tumorigenesis in prostate cancer. *Nucleic Acids Res* 46, 1998-2011, doi:10.1093/nar/gkx1248 (2018).

68 Jiang, X. P. & Elliott, R. L. Decreased Iron in Cancer Cells and Their Microenvironment Improves Cytolysis of Breast Cancer Cells by Natural Killer Cells. *Anticancer Res* 37, 2297-2305, doi:10.21873/anticanres.11567 (2017).

69 Applegate, L. A., Scaletta, C., Panizzon, R. & Frenk, E. Evidence that ferritin is UV inducible in human skin: part of a putative defense mechanism. *J Invest Dermatol* 111, 159-163, doi:10.1046/j.1523-1747.1998.00254.x (1998).

70 Gruber, J. V. & Holtz, R. Examining the impact of skin lighteners in vitro. *Oxid Med Cell Longev* 2013, 702120, doi:10.1155/2013/702120 (2013).

71 Yan, Y. & Jiang, Y. RACK1 affects glioma cell growth and differentiation through the CNTN2-mediated RTK/Ras/MAPK pathway. *Int J Mol Med* 37, 251-257, doi:10.3892/ijmm.2015.2421 (2016).

72 Chen, Y., Wang, L., Xu, H., Liu, X. & Zhao, Y. Exome capture sequencing reveals new insights into hepatitis B virus-induced hepatocellular carcinoma at the early stage of tumorigenesis. *Oncol Rep* 30, 1906-1912, doi:10.3892/or.2013.2652 (2013).

73 Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Gottgens, B., Hacohen, N., Haniffa, M., Hemberg, M., Kim, S., Klenerman, P., Kriegstein, A., Lein, E., Linnarsson, S., Lundberg, E., Lundeberg, J., Majumder, P., Marioni, J. C., Merad, M., Mhlanga, M., Nawijn, M., Netea, M., Nolan, G., Pe'er, D., Phillipakis, A., Ponting, C. P., Quake, S., Reik, W., Rozenblatt-Rosen, O., Sanes, J., Satija, R., Schumacher, T. N., Shalek, A., Shapiro, E., Sharma, P., Shin, J. W., Stegle, O., Stratton, M., Stubbington, M. J. T., Theis, F. J., Uhlen, M., van Oudenaarden, A., Wagner, A., Watt, F., Weissman, J., Wold, B., Xavier, R., Yosef, N. & Human Cell Atlas Meeting, P. The Human Cell Atlas. *Elife* 6, doi:10.7554/eLife.27041 (2017).

74 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, doi:10.1038/nprot.2009.97 (2009).

75 Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A. & Huber, W. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 21, 3439-3440, doi:10.1093/bioinformatics/bti525 (2005).