


METHODOLOGY ARTICLE

Open Access



# scReQTL: an approach to correlate SNVs to gene expression from individual scRNA-seq datasets

Hongyu Liu<sup>1,2†</sup>, N. M. Prashant<sup>1†</sup>, Liam F. Spurr<sup>3,4,5</sup>, Pavlos Bousounis<sup>1</sup>, Nawaf Alomran<sup>1</sup>, Helen Ibeawuchi<sup>1</sup>, Justin Sein<sup>1</sup>, Piotr Słowiński<sup>6,7</sup>, Krasimira Tsaneva-Atanasova<sup>6,7,8,9</sup> and Anelia Horvath<sup>1,10\*</sup> 

## Abstract

**Background:** Recently, pioneering expression quantitative trait loci (eQTL) studies on single cell RNA sequencing (scRNA-seq) data have revealed new and cell-specific regulatory single nucleotide variants (SNVs). Here, we present an alternative QTL-related approach applicable to transcribed SNV loci from scRNA-seq data: scReQTL. ScReQTL uses Variant Allele Fraction (VAF<sub>RNA</sub>) at expressed biallelic loci, and correlates it to gene expression from the corresponding cell.

**Results:** Our approach employs the advantage that, when estimated from multiple cells, VAF<sub>RNA</sub> can be used to assess effects of SNVs in a single sample or individual. In this setting scReQTL operates in the context of identical genotypes, where it is likely to capture RNA-mediated genetic interactions with cell-specific and transient effects. Applying scReQTL on scRNA-seq data generated on the 10 × Genomics Chromium platform using 26,640 mesenchymal cells derived from adipose tissue obtained from three healthy female donors, we identified 1272 unique scReQTLs. ScReQTLs common between individuals or cell types were consistent in terms of the directionality of the relationship and the effect size. Comparative assessment with eQTLs from bulk sequencing data showed that scReQTL analysis identifies a distinct set of SNV-gene correlations, that are substantially enriched in known gene-gene interactions and significant genome-wide association studies (GWAS) loci.

**Conclusion:** ScReQTL is relevant to the rapidly growing source of scRNA-seq data and can be applied to outline SNVs potentially contributing to cell type-specific and/or dynamic genetic interactions from an individual scRNA-seq dataset.

**Availability:** <https://github.com/HorvathLab/NGS/tree/master/scReQTL>

**Keywords:** eQTL, ReQTL, scReQTL, single cell, VAF<sub>RNA</sub>, scVAF<sub>RNA</sub>, scRNA-seq, SNV, Genetic variation, RNA-seq, single cell RNA sequencing, single cell RNA-seq

\* Correspondence: [horvatha@gwu.edu](mailto:horvatha@gwu.edu)

<sup>†</sup>Hongyu Liu and N. M. Prashant contributed equally to this work.

<sup>1</sup>McCormick Genomics and Proteomics Center, School of Medicine and Health Sciences, The George Washington University, Washington, DC 20037, USA

<sup>10</sup>Department of Biochemistry and Molecular Medicine, Department of Biostatistics and Bioinformatics School of Medicine and Health Sciences, George Washington University, Washington, DC 20037, USA

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

In recent years, single cell RNA-seq (scRNA-seq) has become an increasingly accessible platform for genomic studies [1]. By enabling cell-level analyses, scRNA-seq has major advantages for studying gene-regulatory relationships. Among others, the ability to distinguish cell populations and to assess cell-type specific transcriptome features, have shown great potential to identify new regulatory networks [2–4]. Furthermore, scRNA-seq enables the assessment of intracellular molecular relationships, which can reveal cell-specific gene-gene interactions and co-regulated genetic features [2, 5, 6]. These relationships can be reflected in mutually correlated molecular traits, including gene expression (GE) and expression of genetic variants, such as Single Nucleotide Variants (SNVs).

A popular method to study SNVs effects on GE is eQTL (Expressed Quantitative Trait Loci), which is based on testing for a correlation between the number of alleles bearing the variant nucleotide at the position of interest, and the level of local (cis) or distant (trans) GE [7]. eQTLs have been mapped by large-scale efforts such as Genotype-tissue Expression Consortium (GTEx), PsychENCODE, ImmVar BLUEPRINT, and CAGE [8–12].

Recently, pioneering eQTL studies on scRNA-seq data have emerged. By utilizing the advantages of the single cell resolution, these studies have revealed many new regulatory SNVs, including those with cell-specific or transient effects [2, 13–16]. To assess GE, these methods employ approaches specific to single cell transcriptomics, including accounting for drop-outs, classification of cells by type, and assessments of progressive cell stages [2–4, 13–16]. SNV information is traditionally obtained from the genotypes across multiple individuals and encoded as the number of alleles (0, 1 or 2) bearing the variant nucleotide. Accordingly, eQTL analyses are confined to SNVs present in a sufficient number of individuals in the studied group, and frequently exclude variants with low minor allele frequency in the population.

Here, we explore an alternative approach to assess the correlation between GE and expression of SNVs located within transcribed genes from scRNA-seq data. The expression of the SNVs is estimated as the proportion of variant-bearing RNA molecules (Variant Allele Fraction,  $VAF_{RNA}$ ) at biallelic SNV loci. To correlate  $VAF_{RNA}$  to GE from single cells, we estimate  $VAF_{RNA}$  in the individual cell alignments, and correlate  $VAF_{RNA}$  with GE from the individual cells using a linear regression model [17]. To develop the pipeline, we used recent methodologies for calling SNVs and  $VAF_{RNA}$  estimation from RNA-seq data [18–23], as well as scRNA-seq-specific methods to estimate GE [24]. We also adopted a strategy from a method recently developed in our lab to correlate  $VAF_{RNA}$  and GE from bulk RNA-sequencing data –

ReQTL (RNA-eQTL) [25]. We term the application of this technique on single-cell RNA-sequencing data: scReQTL.

We applied scReQTL on publicly available scRNA-seq generated on the 10× Genomics Chromium platform using 3′-based protocol on 26,640 human adipose-derived mesenchymal stem cells (ADSCs), obtained from three healthy donors. This scReQTL analysis includes approximately 4 billion scRNA-seq reads. ScReQTL analysis was performed after classification of the cells by cell type, and only SNVs covered by a minimum of 10 unique sequencing reads per cell were included in the analysis. Across the three samples, we identified 1272 unique scReQTLs. scReQTLs common between individuals or cell types were consistent in terms of the directionality of the relationship and the effect size. In addition, scReQTLs were substantially enriched in known gene-gene interactions and significant genome-wide association studies (GWAS) loci.

## Results

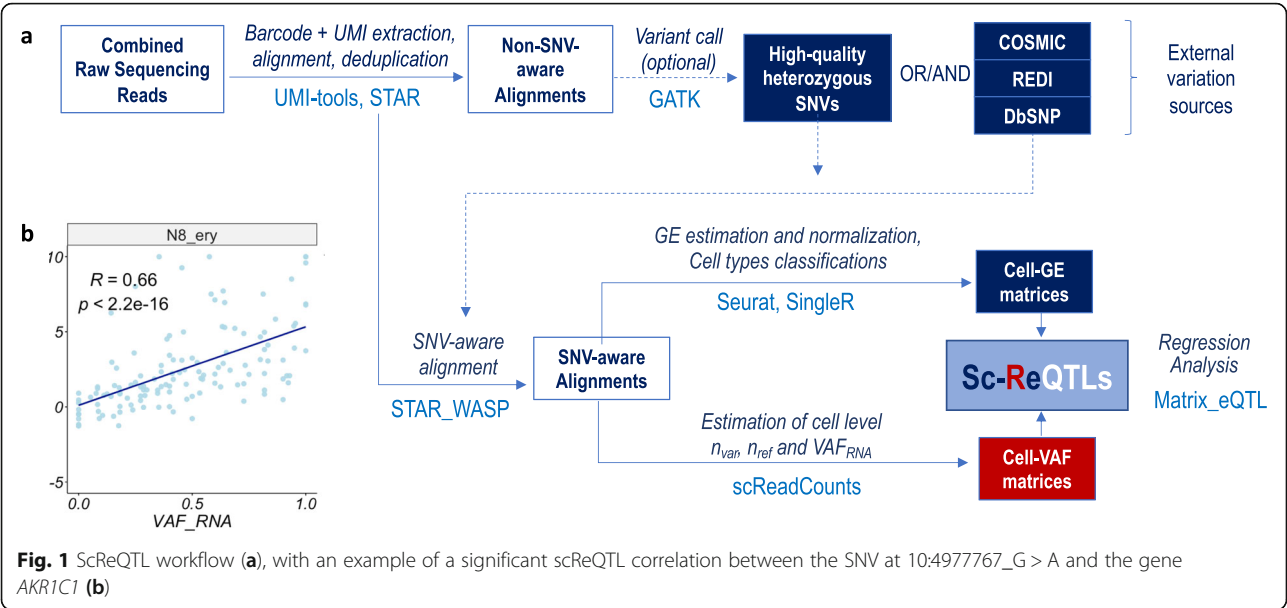
### Overview of scReQTL workflow

An example of scReQTL workflow using publicly available tools is presented in Fig. 1 and outlined in detail in *Methods*. Below, we describe the workflow elements that we identified as specific and essential for the scReQTL analysis.

The scReQTL workflow includes three major components: scRNA-seq data processing,  $VAF_{RNA}$  assessment, and SNV-GE correlation by cell type.

*Processing* includes barcode and UMI modeling, alignment, GE estimation and cell type classification, and can employ a variety of publicly available tools. In the exemplified workflow, we process the barcodes using UMItools, and align using STAR the alignments are then deduplicated based in UMIs [26, 27]. Because  $VAF_{RNA}$  estimations can be sensitive to allele mapping bias, SNV-aware alignment recommended. We perform SNV-aware alignment for the list of positions of interest to be used as input for scReQTL analyses. Here, we perform SNV-aware alignment against the biallelic positions called by GATK in the corresponding pooled alignments applying a two-pass 2-pass STAR-WASP as previously described [18, 28, 29]. Alternatively, scReQTL can be applied on genomic positions of interest from external sources, for example sets of somatic mutations from the COSMIC database, or known RNA-edited loci from the REDI portal [30, 31]; in these cases, the selected sets of loci can be used as input for STAR-WASP alignment for.

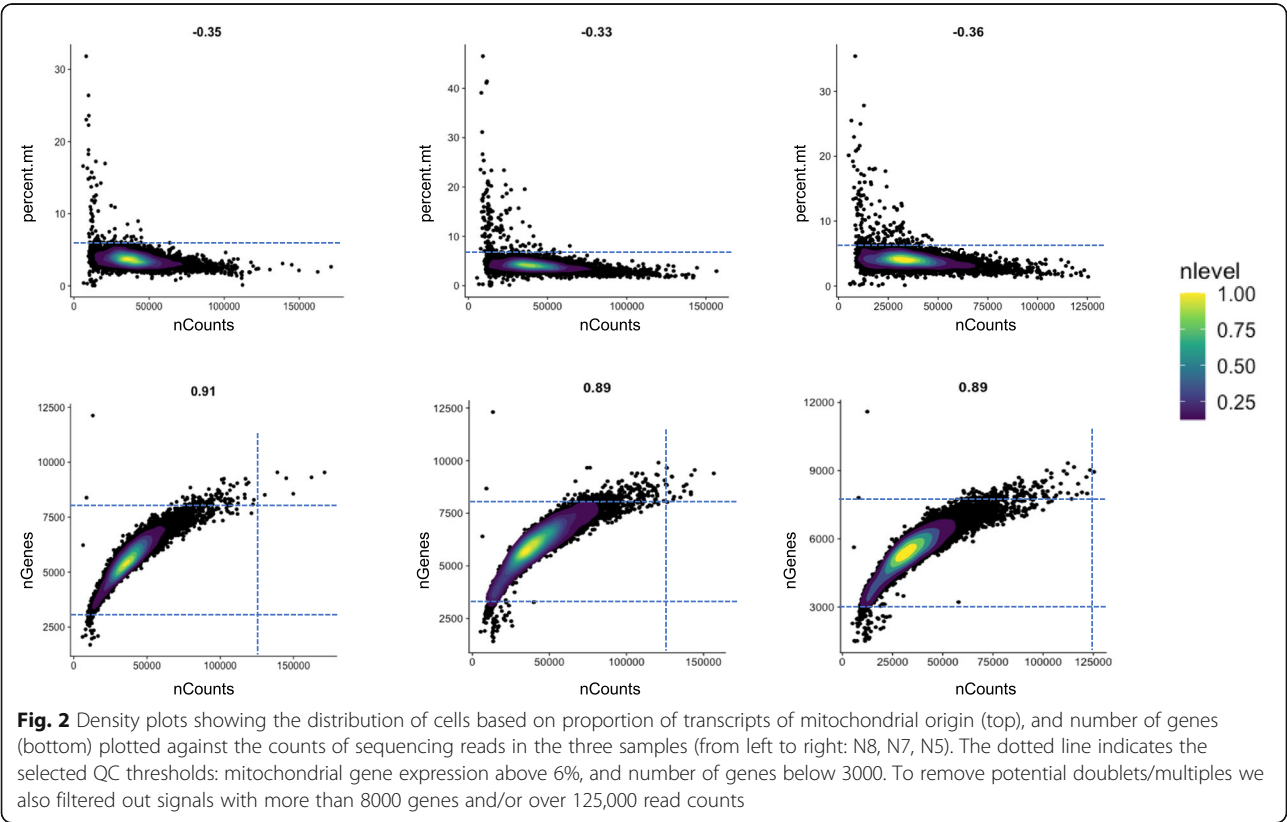
*GE estimation* is performed on the SNV-aware alignments, using FeatureCounts to assess the raw gene counts [32], followed by Seurat for normalization and GE variance stabilization [24, 33]. The generated GE expression values are then used to remove low quality

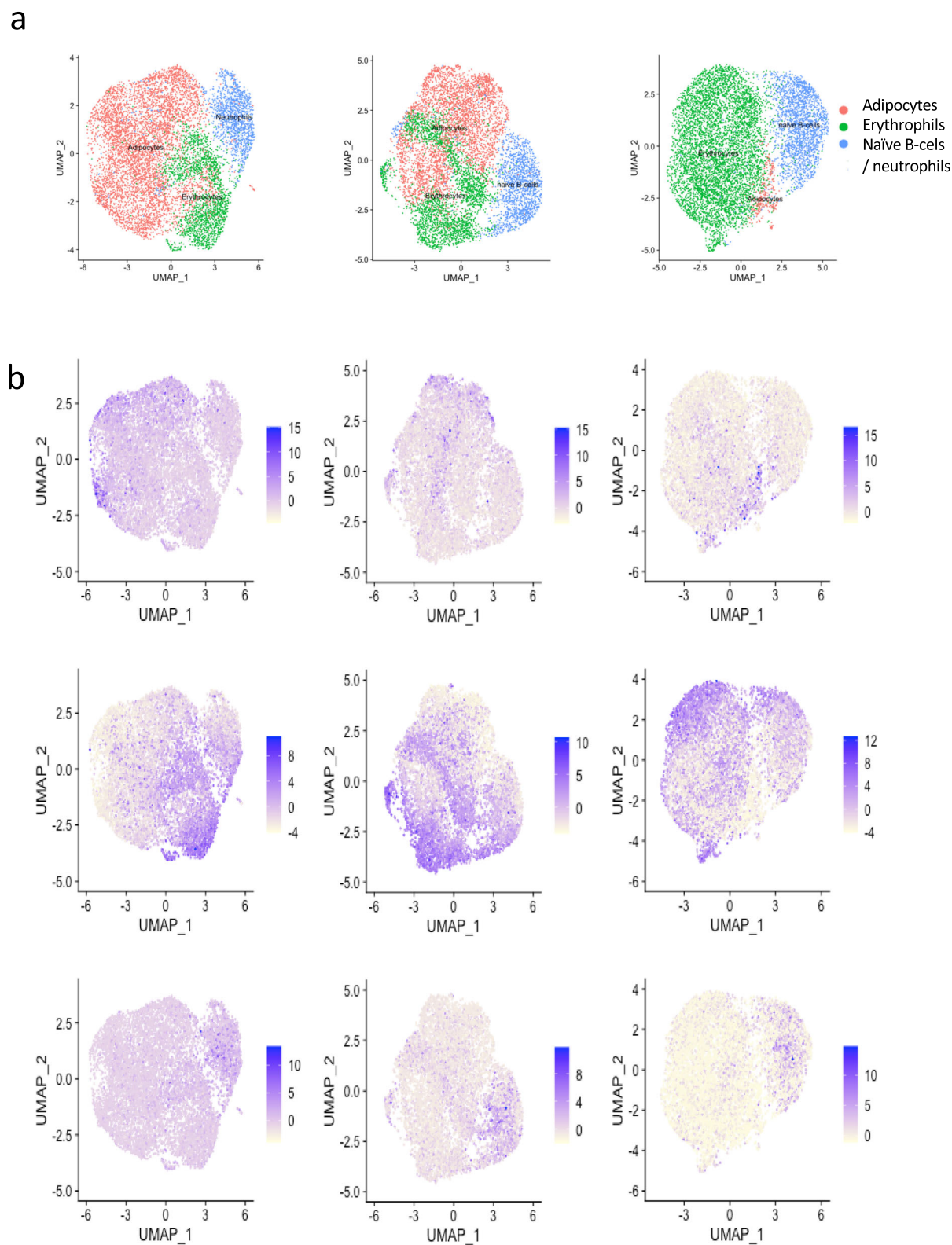


data, batch effects and cell-cycle effects. The distributions of genes and RNA-seq reads, and the selected QC threshold are shown on Fig. 2. The effects of batch-correction and cell-cycle effects removal are shown on Supplementary Figure 1. On the high-quality cell-set we then apply Seurat [33], to normalize gene expression

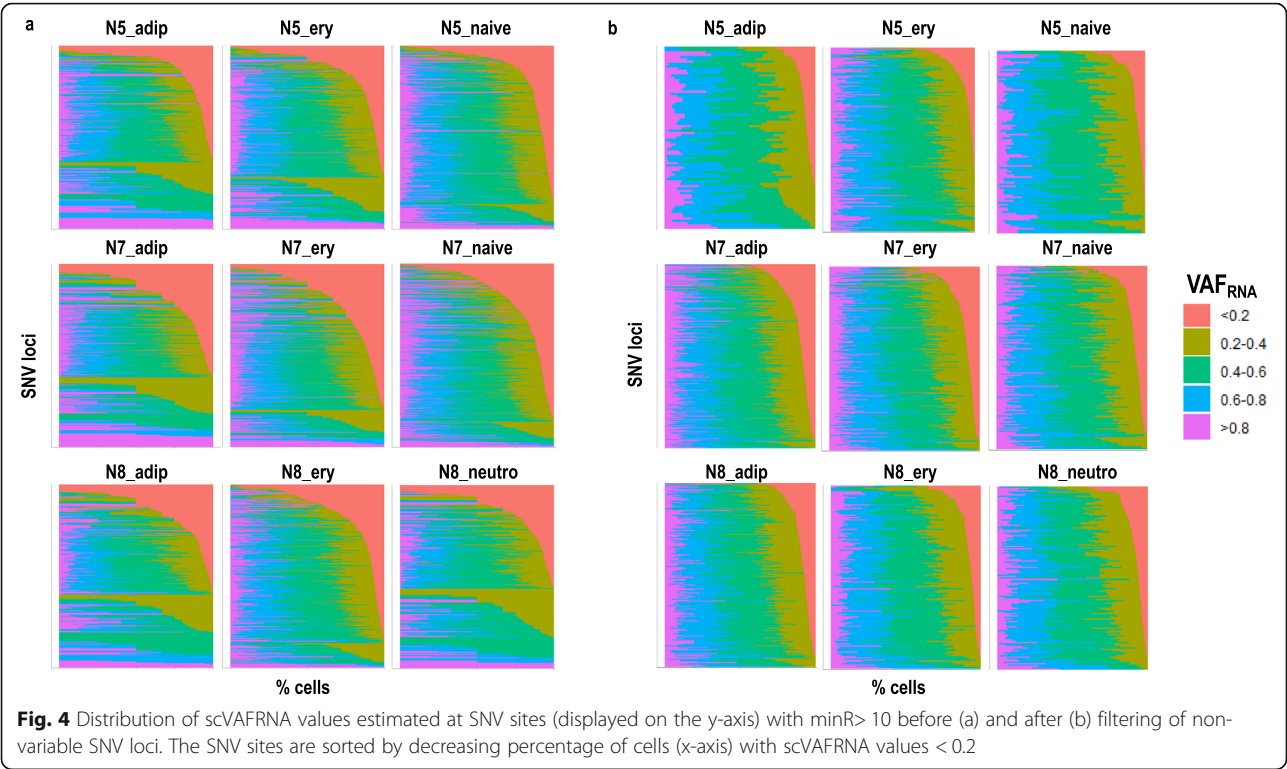
and to identify the most variable genes to be used in the scReQTL analyses (See Methods).

Cell type identification is performed using SingleR [34]. The expression profile of each single cell was correlated to expression data from the Blueprint + ENCODE dataset. Across the three study samples, four major cell





**Fig. 3 a)** Cell types identified in each donor using SingleR. Adipose cells and erythrocytes were found in all three donors, whereas naïve-B-cells were seen in N5 and N7 and neutrophils only in N8. **b)** expression of genes associated with cell types: DCN (adipose cells, top), H2AFZ (erythrocytes, middle), and H1FO (neutrophils and naïve B cells)



types were identified: adipose cells, erythrocytes, neutrophils, and naïve-B cells. Adipose cells and erythrocytes were found in all three samples, whereas naïve-B cells were seen in N5 and N7 and neutrophils – in N8 (Fig. 3 and Supplementary Figure 2).

VAF<sub>RNA</sub> is assessed from the individual cell alignments at the positions of interest using SCReadCounts [35]. For each position, SCReadCounts estimates the number of sequencing reads bearing the variant and the reference nucleotide ( $n_{\text{var}}$  and  $n_{\text{ref}}$  respectively), calculates VAF<sub>RNA</sub> ( $\text{VAF}_{\text{RNA}} = n_{\text{var}} / (n_{\text{var}} + n_{\text{ref}})$ ) and outputs the values in an SNV-barcode matrix. The SNV-barcode matrices are in a format analogous to the GE-barcode matrices and can be directly used in the ReQTL analyses. To address stochasticity of sampling, estimations of VAF<sub>RNA</sub> require a threshold of minimal number of unique sequencing reads (minR). Our previous research shows that current scRNA-seq datasets can contain hundreds of SNV sites covered by minimum of 10 sequencing reads ( $\text{minR} \geq 10$ ) and thousands of SNV sites with  $\text{minR} \geq 5$  [28]. In the herein presented analysis, we used VAF<sub>RNA</sub> estimated at sites with  $\text{minR} \geq 10$ ; from here on, we refer to these loci as informative. We note that for minR we are referring to sequencing reads with unique UMIs which are derived from unique mRNA molecules. The VAF<sub>RNA</sub> distribution of the qualifying SNVs is then examined to identify the most variable VAF<sub>RNA</sub> loci (see Methods). VAF<sub>RNA</sub> distributions

**Table 1** Input parameters for scReQTL analysis, and number of identified scReQTLs per cell type

Sample	N cells	N reads	Mean Reads/Cell	Median Genes/Cell	N cells (per cell type) after filtering		N input SNVs	N input genes	N scReQTLs FDR = 0.05
N5	8,906	1,071,156,174	120,273	5,439	Adipocytes	296	79	2,114	31
					Erythrocytes	5,848	208	2,206	161
					Naïve-B cells	2,033	99	2,138	82
N7	8,478	1,579,342,505	186,287	6,049	Adipocytes	3,819	316	2,442	336
					Erythrocytes	2,788	238	2,395	127
					Naïve-B cells	1,618	167	2,366	102
N8	9,256	1,285,218,728	138,852	5,559	Adipocytes	5,738	230	2,345	299
					Erythrocytes	1,924	157	2,367	72
					Neutrophils	1,433	139	2,340	71
Total/Overall	26,640	3,935,717,407	148,471	5,682	Total/Distinct	25,497	1633 / 644	20,713 / 2,571	1,281 / 1272

before and after filtering of uninformative ( $\text{minR} < 10$ ) and non-variable  $\text{VAF}_{\text{RNA}}$  are shown on Fig. 4a and b, respectively.

**SNV-GE correlations** (scReQTLs) are then computed for each donor, stratified by cell type (see Methods). To qualify for scReQTLs analysis an SNV locus is required to have informative and variable  $\text{VAF}_{\text{RNA}}$  estimations from at least 20 cells per analysis. The variable  $\text{VAF}_{\text{RNA}}$  were correlated to the normalized GE values of the variable genes using linear regression model as implemented in Matrix eQTL [17]; quantile-quantile plots (QQ-plots) are presented on Supplementary Figure 3. Cis- and trans-correlations were annotated as we have previously described for the bulk ReQTLs [25]. Briefly, because scReQTLs are assessed from transcripts, we assign cis-correlation based on the co-location of the SNV locus within the transcribed gene; all the remaining correlations are annotated as trans-scReQTLs).

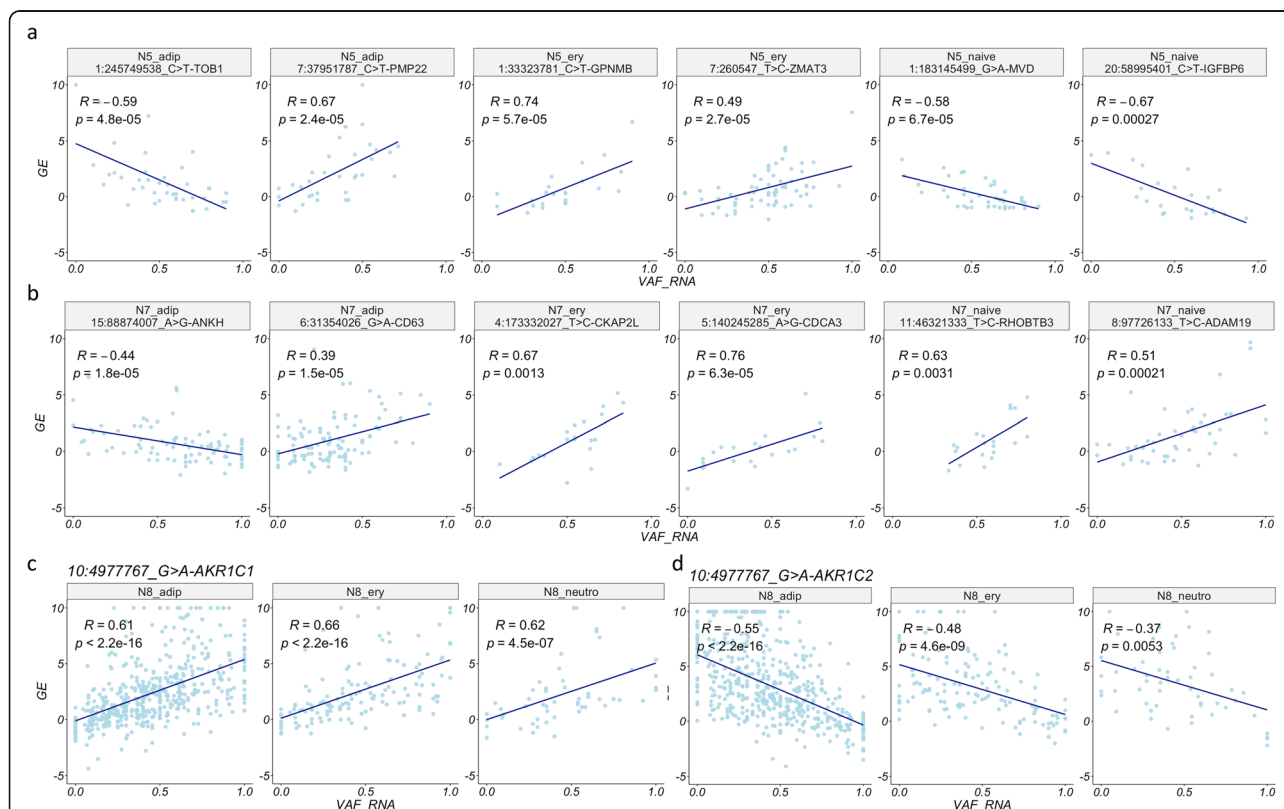
### Overall scReQTL findings

The number of variable genes and  $\text{VAF}_{\text{RNA}}$  loci retained for scReQTL analysis in the three donors (by cell type) is shown in Table 1. We performed scReQTL analysis

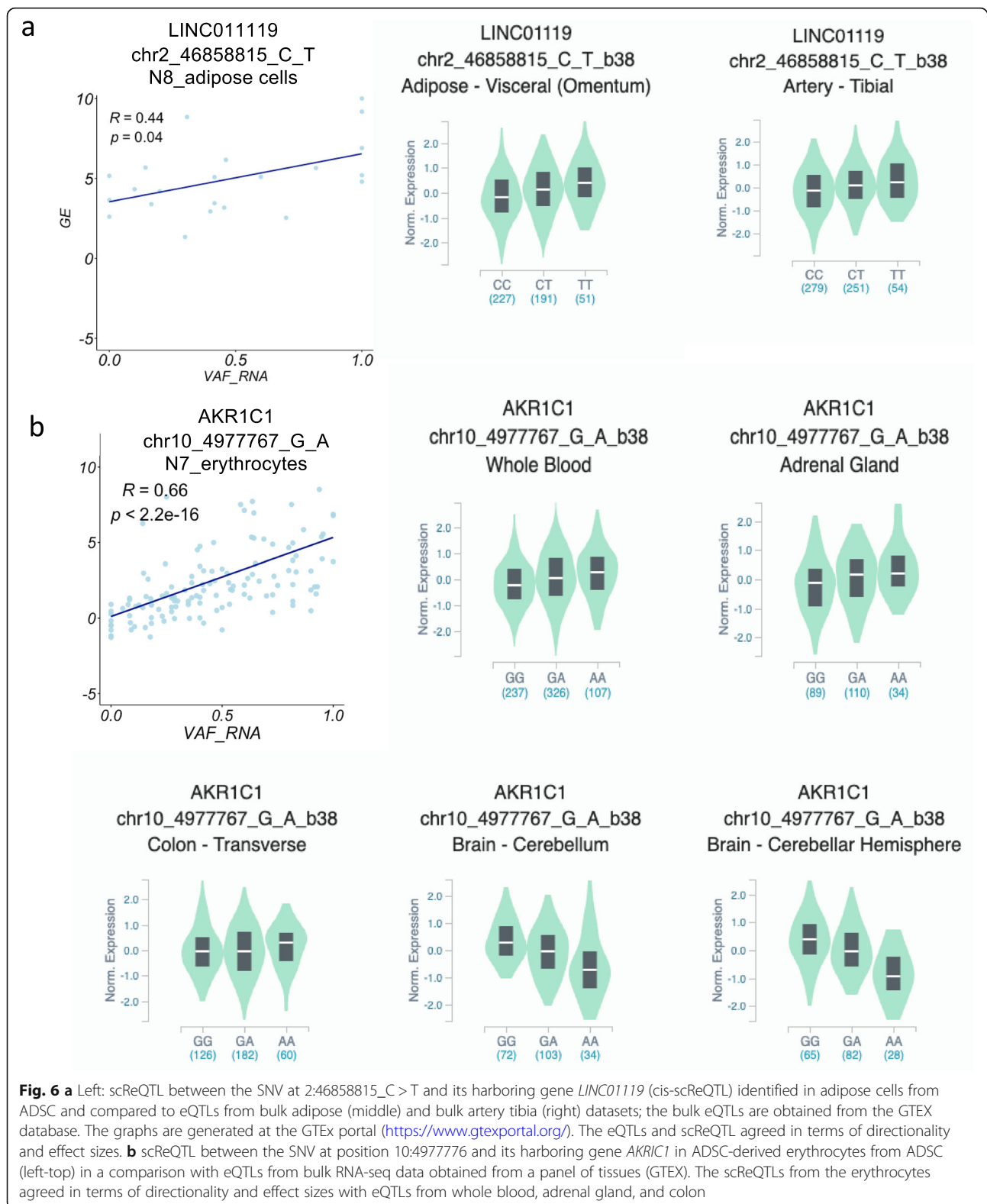
separately for each individual and cell type; accordingly, 9 scReQTL analyses were run. Among the samples and cell types, between 79 and 316 SNV loci, and between 2114 and 2442 genes were used as input for scReQTL analysis. Across the 9 groups, a total of 644 distinct SNVs and 2571 distinct genes were tested. This analysis identified 1272 unique scReQTLs at false discovery rate (FDR) of 0.05. All significant scReQTLs are listed in Supplementary Table 1; examples are shown on Fig. 5.

Among the unique scReQTLs, 7 were identified in more than one cell type or sample (Supplementary Table 2). In all these cases, the correlations were in the same direction, and the effect sizes were similar (See Fig. 5c and d). We note that the number of common input-SNVs across the 3 samples was as low as 20 (numbers of common input SNVs and genes, as well as the common scReQTLs SNVs and genes are shown in Supplementary Figure 4).

Next, we investigated the relationship between cis- and trans-scReQTLs. Of the significant scReQTLs, only 6 represented cis-correlations, representing 4 distinct SNV-gene pairs, (examples shown in Fig. 5c and Fig. 6). This low proportion of cis-scReQTL correlations differs



**Fig. 5** Examples of significant (FDR = 0.05) scReQTL correlations in donor N5 (a), N7 (b) and N8 (c and d). In N8, consistent across the three cell types cis-scReQTL is shown between the SNV at 10:4977767\_G>A and its harboring gene *AKR1C1* (c), and between the same SNV and the nearby positioned gene *AKR1C2* (trans-scReQTL, d). Note that the displayed P-values are calculated based on the input for the plots generated using the R-package ggplot2 and do not represent the FDR—corrected values from the scReQTL analysis performed with Matrix eQTL



from eQTL analyses, which typically identify a high number of significant cis-correlations and is attributed to several factors. First, in contrast to the eQTL

distance-based cis/trans annotation, scReQTL employs gene-based annotation, which results into a cis-to-trans shift for SNVs in nearby genes. Second, cis-scReQTL

estimations require a certain level of expression of the SNV-harboring genes (as defined by minR) thereby confining cis-scReQTL analyses to moderately-to-highly expressed genes in the system. Third, the scReQTL input SNVs are confined to expressed regions, and in the herein employed system (10xGenomics Chromium 3'-protocol) most of them are 3'-UTR-located. Indeed, two of the cis-scReQTL SNVs were located in the 3'UTR of their harboring genes, one was a synonymous substitution (G246G in *TNNT3*) and one was located in an exon of the non-coding RNA *LINC01119*. In contrast, cis-eQTLs are frequently located in the promoter and other regulatory sequences, often transcriptionally silent and therefore not detected by the ReQTL/scReQTL approach.

To further investigate the connection between cis- and trans-scReQTLs, we assessed if some scReQTLs are mediated by cis-effects that do not reach significance at an FDR of 0.05. To do this, we computed the correlation of all SNVs represented in significant trans scReQTLs with their harboring gene. For 26% of the scReQTL SNVs, we detected correlations with their harboring genes with  $0.05 < \text{FDR} < 0.1$  (Supplementary Figure 5). This analysis suggests that a proportion of the SNVs may at least partially exert their trans-effects via weak to moderate regulation of the expression of their harboring gene.

### scReQTL in known genetic networks

To assess to what extent scReQTL findings agree with known SNV-gene, and gene-gene interactions, we intersected the significant scReQTLs with: (a) eQTLs reported in the GTEx database [8], (b) ReQTLs as estimated from bulk adipose sequencing data [25], (c) known gene-gene interaction from the STRING database [36], and (d) significant GWAS loci [37].

### scReQTLs and eQTLs from GTEx

To estimate the overlap between scReQTL and known eQTLs, we used the data from 53 different tissues and cell types from the GTEx database (<https://www.gtexportal.org>). First, we identified the SNVs and genes used as an input for scReQTLs, and participating in known eQTLs: a total of 111 input SNVs and 2024 input genes participated in at least one eQTL reported in GTEx. Across the 49 tissues, scReQTL identified 32 correlations (Supplementary Table 3), comprised of 6 unique SNV-gene pairs (5 SNVs and 6 genes). These pairs included all 4 significant cis-scReQTLs, and two trans-scReQTLs: chr10\_4,977,767\_G > A and *AKRIC2* (see Fig. 5d), and chr1:115337511\_G\_A and *NGF*. For each of the 6 SNV-gene pairs, we compared the scReQTLs and the eQTLs in the different GTEx tissue types. For 3 of the 6 scReQTLs, the corresponding GTEx eQTLs were consistent in terms of directionality and effect size (Fig. 6 and Supplementary Figures 6 and 7).

The other 3 scReQTL were found as both positive and negative eQTLs depending on the tissue type in GTEx. The positive cis-scReQTL, chr6:31354105\_G > A\_HLA-B, was a significant cis-eQTL in 4 GTEx tissues: positive in three, but negative in the testis (Supplementary Figure 8). The last 2 scReQTLs comprised correlations of the SNV at chr10:4977767\_G > A with *AKRIC1* (positive) and *AKRIC2* (negative); these scReQTLs were consistent across cell types (see Fig. 5c and d). In GTEx, the corresponding eQTLs were found in multiple tissues, and in both positive and negative correlations, highlighting tissue-specific effects (Supplementary Figures 9 and 10).

Overall, our analysis on the agreement between significant scReQTLs and eQTLs identified a narrow overlap, within which most observations were consistent, and the remaining were not contradictory. We note that this analysis was limited by the relatively small number of input scReQTL SNVs present in GTEx. Furthermore, the majority of the significant scReQTLs were in trans, which are known to be highly tissue-specific [8]. None of the 4 cell types assessed in our study - adipose cells, erythrocytes, neutrophils, and naïve-B cells obtained from ADSCs - were a direct match to any of the 49 tissues and cell types for which eQTLs were available from the GTEx database. Finally, we expect that the strongest contributor to the low overlap between scReQTL and eQTLs is the detection power of scReQTL. Specifically, depending on the sequencing depth per cell, many cells do not pass the minR requirement for a given SNV (especially at minR > 10, which is the cutoff used here), and are therefore excluded from the analysis. Indeed, while the initial cell counts per scReQTL analysis (except for N5 adipose cells) were over 1000, the majority of the SNV loci were expressed in between 20 (the required minimum) and 100 cells with minR > 10 per cell type (Supplementary Figure 11a). In comparison, the GTEx eQTLs are computed from a minimum of 100, and in most of the tissues, from over 250 individuals (Supplementary Figure 11b).

### scReQTLs and ReQTLs from bulk adipose tissue

Next, we intersected the scReQTL findings with the ReQTLs from bulk RNA-sequencing data. To do this, we performed ReQTL on RNA-seq data from two adipose tissues downloaded from GTEx - adipose subcutaneous (275 samples) and adipose visceral (215 samples) - following the published protocol [25]. Using the same SNVs and genes used as input for the scReQTL, with an FDR = 0.05, ReQTL did not identify significant correlations, whereas with an FDR = 0.1, ReQTL identified 84 (6.6%) and 48 (3.8%) of the significant scReQTLs, in adipose subcutaneous and visceral tissue, respectively. The majority of these ReQTLs had small effect sizes and agreed in the direction with the corresponding scReQTL

in 71% of the cases (Examples shown on Fig. 7a). Of note, the above discussed chr10:4977767\_G > A and *AKR1C1/AKR1C2* did not show any correlation when examined from bulk RNA-seq data (Fig. 7b).

The different sets of SNV-gene pairs identified by scReQTL and ReQTL (as well as eQTL) suggests different regulatory relationships captured by scReQTLs. Bulk ReQTLs and eQTLs show a high overlap between each other and are both based on abundance of variant alleles across multiple individuals with different genotypes. In contrast, scReQTL operates in a setting of identical genotypes, where it is likely to capture RNA-based genetic interactions, possibly with transient and/or cell-specific effects.

#### scReQTLs and known gene-gene interactions

Because the vast majority of the significant scReQTLs were in trans (i.e. representing correlations between two different genes, VAF<sub>RNA</sub> of an SNV located in one of the genes and expression level of the other), we assessed if these gene pairs were enriched in known gene-gene interactions. We downloaded the known gene-gene (human) interactions from the STRING database [36] and intersected these with the scReQTLs. From the 1234 unique gene-gene scReQTLs pairs, 203 (16.4%) were previously annotated in STRING (Supplementary Table 4,  $p < 10e-4$ , permutation test using 10,000

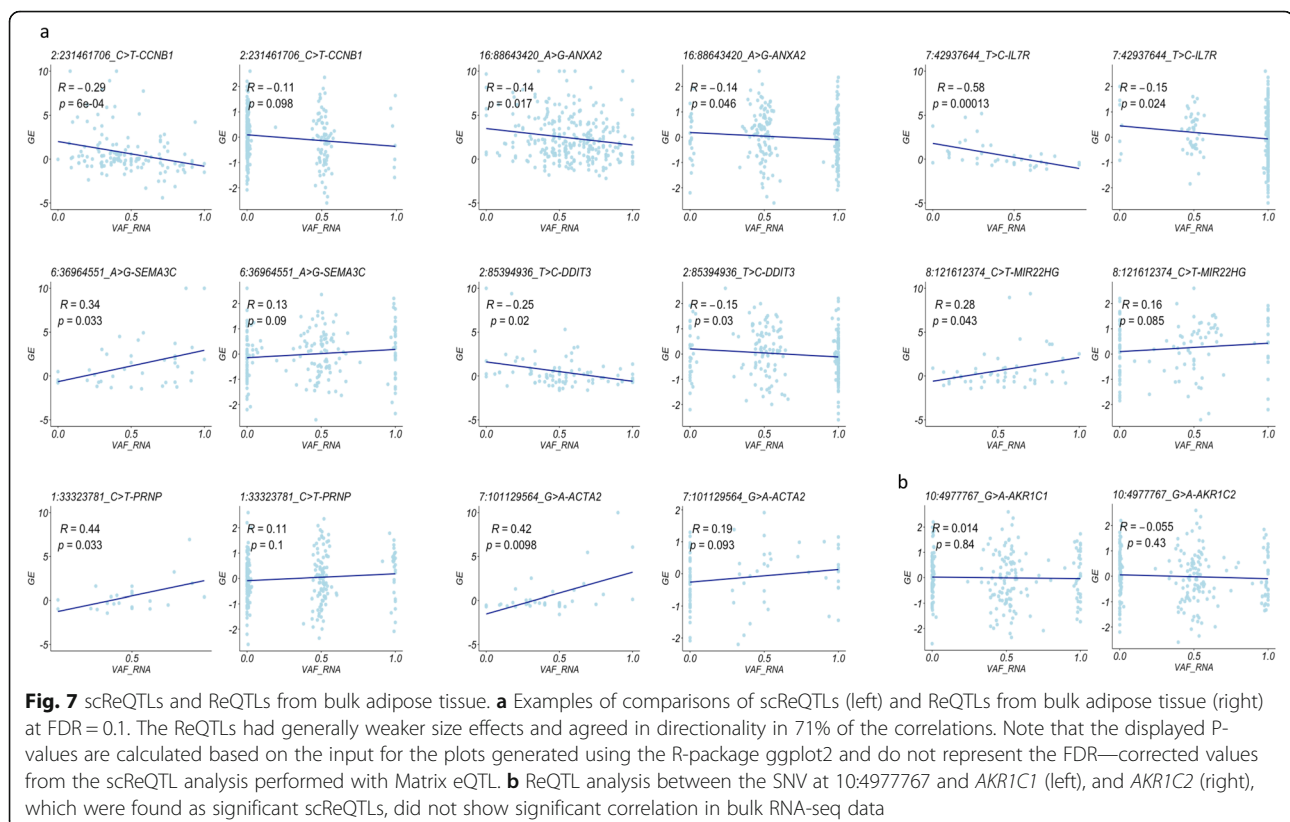
permutations, Fig. 8a). Examples include *IFIT1* and *IFITM2*, *AURKA* and *PLK*, and *CKS2* and *CDC20* (Fig. 8b-c). The strong enrichment of scReQTLs with known genetic networks suggests that scReQTLs may be used to identify allele contributions to gene-gene interactions.

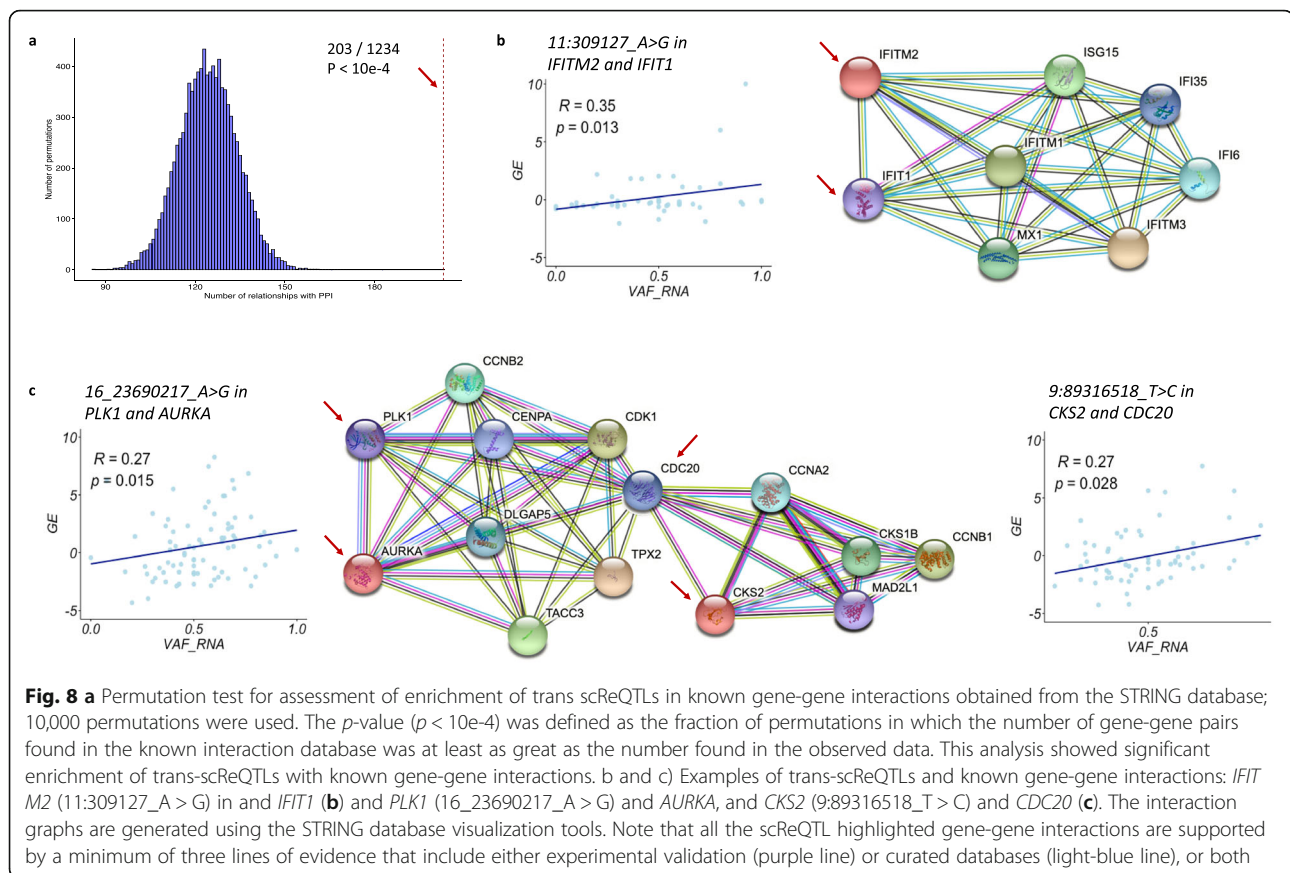
#### scReQTLs and GWAS

Furthermore, we intersected the SNVs participating in scReQTLs with SNVs significantly associated with phenotypes by GWAS [38]. This analysis showed that 18 (out of the 408 unique scReQTL SNVs, 4.4%) were present in GWAS; these 18 SNVs participated in 84 scReQTL correlations (Supplementary Table 5). This percentage is slightly higher than the overlap between GWAS and GTEx eQTLs (3.7 and 3.6% in adipose visceral and adipose subcutaneous tissue respectively), and significantly higher than the overlap with common SNVs from DbSNPv.154, (0.34%,  $p < 10e-6$ ). This analysis shows that scReQTL SNVs are enriched in genetic variants associated with phenotype via large population-based and case-control studies.

#### Functional scReQTLs SNVs annotations

We assessed the SNVs participating in scReQTL in regard to position in the harboring gene and the predicted functional effects. As expected from scRNA-seq data generated using a 3'-based protocol, the majority of the





SNVs resided in the 3'UTR of their harboring gene (70.2%, Supplementary Figure 12); the 3'-UTR SNVs participated in 69.6% of the scReQTLs. 3'-UTR variants are known to strongly affect both GE levels and splicing [39–42]; hence, scReQTLs can be applied to study these aspects of genetic regulation. The second category was exonic SNVs, comprising 16.2% of the unique SNVs and participating in 14.9% of the scReQTLs. Exonic SNVs included missense, nonsense, and near-splice variants, many of which can potentially affect the protein structure and function. Of note, scReQTL captured a substantial number of intronic SNVs – 13%, participating in 11.2% of the scReQTLs. Intronic sequences are reported in 15–25% of the RNA-sequencing reads from both bulk and single-cell RNA-seq [4, 41, 42]. Intron quantitation can be used to estimate the relative abundance of the precursor and mature mRNA, thereby assessing the RNA velocity and dynamic cellular processes [4]. In the allele-specific setting provided by the scReQTLs, the correlations of intronic SNVs with GE can identify SNVs regulating the RNA splicing.

Next, we assessed if the scReQTLs SNVs are enriched in specific clinical phenotypes obtained from the ClinVar database [43]. Fifteen SNVs (3.7% of the total 408 distinct scReQTL SNVs) were associated with known

clinical phenotypes, including circulating phospholipid trans fatty acids, cortisol levels, circadian rhythm, risk for cardiovascular disease, blood pressure, schizophrenia, neuroticism, osteoporosis, anthropometric traits, and asthma (See Supplementary Table 1). This percentage is similar to the overlap between ClinVar and GTEx eQTLs (3.3 and 3.1% of the eQTLs in adipose visceral and adipose subcutaneous tissue respectively), and significantly higher than the overlap with common SNVs from DbSNPv.154, (0.61%,  $p < 10e-6$ ). Finally, we assessed the predicted functional and/or pathogenic scores of the scReQTL SNVs using 17 models including SIFT, Polyphen2, LRT, MutationTaster, MutationAssessor, FATHMM, PROVEAN, VEST3, CADD, DANN, fathmm-MKL, MetaSVM, MetaLR, integratedFit, GERP++, phyloP, and phastCons, as implemented in ANNOVAR [44]; this data is summarized in Supplementary Table 6).

#### scReQTL application

Application of scReQTLs requires consideration of several factors. First, because scReQTLs assess expressed SNV loci, they cannot capture variants in transcriptionally silent genomic regions. In addition, SNV loci with expression levels below the required minimum number

of RNA-seq reads (minR) are not included in the scReQTL analyses. Furthermore, when assessed with the platform used in this study - 10x Genomics Chromium v3 chemistry – the analyzed SNVs are confined to those located within the length of the sequencing read (here, 150 nt) from the 3' end of the transcript. For the above reasons, scReQTLs accessible SNVs represent a subset of the expressed SNVs and are not designed to cover the full set of SNVs in the genome/transcriptome.

Second, when a genetically regulated gene is captured by scReQTL analysis, the scReQTLs may highlight SNVs that are co-allelic to the actual causative SNV(s). This is the case for SNVs positioned outside the transcribed regions or outside the coverage of the sequencing library.

Third, scReQTLs are based on  $VA_{RNA}$  estimation, which can be affected by technical parameters, including allele mapping bias [45, 46]. Therefore, we perform the scReQTL using SNV-aware alignments. Specifically, we apply STAR-alignment with WASP, which removes ambiguously mapped reads after checking for consistency with the reads containing the alternative nucleotide [27, 29].

Another important parameter for  $VA_{RNA}$  estimation is the selection of cutoff for minimal number of reads, minR. When selecting minR for an analysis, a major factor is the balance between the confidence of  $VA_{RNA}$  estimation (high minR) and the inclusivity of SNVs (lower minR values include more loci for scReQTL). Our previous research shows that for current 10x Genomics scRNA-seq datasets,  $\text{minR} \geq 5$  provides a reasonable balance between  $VA_{RNA}$  confidence and SNV inclusivity, while at lower cutoffs (i.e.  $\text{minR} = 3$ ) stochasticity of sampling can affect the  $VA_{RNA}$  estimation [28]. In the present study, we have included SNV loci with  $\text{minR} \geq 10$ . In addition, to assess the scReQTL at lower minR cut-offs, for a subset of the dataset (N7) we performed repeated analyses varying the value of minR in one-step increments between 5 and 9 and analyzed the outputs. The first observation is that, as expected, minR inversely correlates with the number of input SNV loci and the number of significant scReQTLs (Supplementary Figure 13a). Second, the outputs of the scReQTLs show only partial overlap across different minR cut-offs, which is due to the partial overlap between the input SNVs (Supplementary Figure 13b-d). We reason that the partial overlap in the input SNVs loci is largely due to the very stringent filtering criteria applied to retain the loci with the most variable  $VA_{RNA}$ . Specifically, we filter out loci for which over 75% of the  $VA_{RNA}$  values are in the range of  $0.5 \pm 0.1$  (corresponding to stable biallelic expression; this removed more than 50% of all loci), as well as loci with over 75% of the  $VA_{RNA}$  values in the ranges 0–0.25 or 0.75–1 (corresponding to predominantly monoallelic or skewed allelic expression). As we show in our previous research [28], the  $VA_{RNA}$  distribution

changes substantially with different minR, which affects the subset of variants retained after filtering. Importantly, each of the distinct sets of significant SNV-gene pairs obtained at different minR showed very strong enrichment in known gene-gene interactions, which supports confident scReQTL observations with  $VA_{RNA}$  cutoffs of 5 and above (Supplementary Figure 14).

Another threshold to consider is the minimal number of cells to be used for scReQTL, in this study set to 20. Using a minimum of 20 cells for assessing SNV-GE correlations is a result of setting a threshold for filtering loci with non-variable  $VA_{RNA}$ . Specifically, the locus is considered to be variable if in at least 15 (75%) out of the 20 cells  $VA_{RNA}$  is in the range 0.25–0.4 or 0.6–0.75 (i.e. we exclude stable biallelic, monoallelic, and skewed allelic expression). From the binomial distribution we compute that this gives 0.0206 probability of observing variable loci by chance assuming a 50% probability of success on each of at least 15 out of 20 trials. This probability decreases with increasing number of cells. We consider the maximal 0.02 chance threshold for wrongly assigning variability of the loci to be reasonably conservative and acceptable. In addition, we showed in our recent study [23] that 20 VAF values are sufficient to model the characteristics of VAF distributions and that higher numbers will improve the estimates (see Fig. 4 in [23]). For 10 cells, the probability of observing a single variable locus by chance would increase to 0.054 (scReQTLs from 10 to 19 cells are shown on Supplementary Figure 15).

Furthermore, scReQTLs can be affected by the accuracy of the variant call, including assessment of presence or absence of an SNV, and assignment of a biallelic state. The presented pipeline uses scRNA-seq data only, where we call SNVs from pooled scRNA-seq data and select for scReQTL analysis highly confident heterozygous sites based on mapping and Phred quality, genomic position (genic, non-repetitive regions), and previously validated rsIDs. To confidently assign biallelic state, we select SNVs with a minimum of 50 unique reads supporting each allele from the pooled scRNA-seq. By default, this selection excludes heterozygous SNVs with strong non-random monoallelic expression. Therefore, while the above approach is suitable for datasets with no matched DNA, when available, DNA-estimated genotypes can be helpful to interpret the context of the scReQTL findings.

Importantly, scReQTLs do not necessarily require variant calls and can be run on custom pre-defined lists of genomic positions such as a database of somatic mutations or RNA-editing sites. In this case the  $VA_{RNA}$  is estimated for all the input sites, and SNVs not present in the assessed sample (i.e. SNVs with  $VA_{RNA} = 0$  across all cells) are removed during the  $VA_{RNA}$  filtering step (see Fig. 4), while the remaining  $VA_{RNA}$  estimations are used in the scReQTL analysis.

Finally,  $VAF_{RNA}$  varies between different cell types, often due to cell-specific regulatory mechanisms [47]. Due to the dynamic nature of RNA transcription, it is expected that  $VAF_{RNA}$  (similarly to GE) will vary depending on conditions, disease states and stochastic factors. Therefore, scReQTLs are expected to be transient and their interpretation requires consideration of the dynamics of the variables underlying the correlation.

## Discussion

Single-cell RNA-seq eQTL analyses define an emerging research niche that brings major benefits for the understanding of functional genetic variation including the identification of cell-type and condition-specific correlations [2, 13–16, 48]. In this paper, we present a new eQTL-based analysis in a scRNA-seq setting - scReQTL - which uses the  $VAF_{RNA}$  at expressed heterozygous SNVs in place of the genotypes, to correlate allele prevalence to gene-expression levels. By using  $VAF_{RNA}$  across multiple cells of the same sample, scReQTLs introduce several new analytical aspects.

First, and perhaps most importantly, as scReQTL is implemented on multiple single cells from the same sample, it can be applied to assess the effects of SNVs in a single sample or individual. This is particularly applicable for rare SNVs which are challenging to study via population-based approaches. We envision that this scReQTL feature can benefit studies on functionality of infrequent and de novo mutations causing rare phenotypes, as well as somatic mutations in cancer. Second, scReQTLs increase the dynamicity of the SNV-gene correlations, as  $VAF_{RNA}$ , similarly to GE, is both dynamic and cell-type-specific [47]. In particular, in each cell type, scReQTL correlates the most variable  $VAF_{RNA}$  to the most variable genes. Third, as compared to the discrete genotype values (0,1,2),  $VAF_{RNA}$  can obtain continuous values spread along the entire  $VAF_{RNA}$  range ([0,1]), allowing for more precise computation of the proportion of each allele represented in the RNA in a given cell. Fourth, scReQTL operates in the context of (largely) identical genotypes, which narrows the observed effects to RNA-mediated interactions. Finally, scReQTL does not necessarily require matched DNA (although we recommend it for genotyping of heterozygous SNVs, if available), and therefore can be applied on scRNA-seq data alone. Related to that, scReQTL analyses can be performed using pre-defined SNVs of interest, such as RNA-editing sites and sets of dbSNP.

At the same time, compared to single cell and bulk eQTLs, scReQTL analyses have certain limitations. First, the scReQTL accessible SNVs are restricted by depth of coverage per cell (minR) and, in the case of 3'-based scRNA-seq protocols, by the length of the sequencing read. Therefore, scReQTLs can analyze only a

proportion of the transcribed SNVs. This limitation is expected to be gradually reduced with the progress of the sequencing technologies. Additional attenuation of this constraint is possible through reducing the value of minR used in the analysis. Indeed, while in this study we apply  $\text{minR} \geq 10$ , which retained between 308 and 721 input SNVs per sample, in our prior research we show that at  $\text{minR} \geq 5$  the number of SNVs is higher by an order of magnitude [28]. Second, scReQTL appears to have relatively low power to detect cis-acting (on the same gene) SNVs (See Supplementary Figure 3). Specifically, the vast majority of the correlations identified in this study are trans-scReQTLs. Several factors may account for this observation. As mentioned earlier, the definition of "cis"-scReQTLs is based on residing of the SNV within the same gene; hence SNVs that would be classified as "cis" using the eQTL distance-based definition are "trans" for the scReQTLs, increasing the proportion of trans-correlations in the same SNV-gene dataset. Additional possible explanation is that in the explored setting of  $\text{minR} \geq 10$ , cis-acting SNVs are located in genes with high expression, which likely contain a high proportion of stably expressed genes, including with house-keeping functions. Notably, the identified trans-scReQTLs are significantly enriched in known gene-gene correlations (See Fig. 7), therefore we interpret them as indicative of an allelic contribution to these gene-gene interactions. The above factors at least partially account for the narrow overlap between scReQTLs and eQTLs/ReQTLs. At the same time, scReQTLs are able to capture correlations that are masked in the bulk eQTL and ReQTL analyses (See Fig. 8).

Finally, at present, a direct comparison between scReQTLs and single cell eQTLs is limited, to a large extent due to a narrow overlap between the sets of SNVs matching the requirements for scReQTL inputs; these SNVs are located mostly in the expressed 3'-end of the gene, as compared to the genome-wide DNA-genotyped loci used in the sc-eQTLs. Additional contributing factors are the different cell sources as well as the different capturing protocols (i.e. SmartSeq vs 10xGenomics). With the advances of the scRNA-seq technologies, and the extension of single-cell QTL-based approaches to more tissues and cell types, comparisons between sc-eQTLs and scReQTLs are expected to provide meaningful information on underlying mechanisms in a cell-type specific context.

Our scReQTL analysis includes approximately 4 billion RNA-seq reads from 26,640 human ADSCs, obtained from three healthy donors. We chose the 10xGenomics platform due to its growing popularity, high throughput, and the support for unique molecular identifiers (UMI) for the removal of PCR-related sequencing bias. Using stringent cutoff for SNV coverage

( $\text{minR} \geq 10$ ) we identified 1272 distinct scReQTLs. These scReQTLs include a considerable number of correlations which involve SNVs previously highlighted by GWAS and are significantly enriched in known gene-gene interactions. These results demonstrate that scReQTLs can be used to identify novel genetic interactions, including those which are specific to a given cell-type.

## Conclusion

We present a new approach – scReQTL – that correlates SNVs to gene expression from scRNA-seq data. ScReQTL is relevant to the rapidly growing source of scRNA-seq data and can be applied to outline SNVs potentially contributing to cell type-specific and/or dynamic genetic interactions from an individual scRNA-seq dataset.

## Methods

### Data

We used publicly available scRNA-seq data [49] from 26,640 human cells from three healthy donors: N5, N7 and N8. The scRNA-seq data was generated on 10x Genomics Chromium v2 platform; the library preparation and sequencing are described in detail elsewhere [49]. Briefly, cells were partitioned using 10x Genomics Single Cell 3' Chips, and barcodes to index cells (16 bp) and transcripts (10 bp UMI) were incorporated. The constructed libraries were sequenced on an Illumina NovaSeq 6000 System in  $2 \times 150$  bp paired-end mode.

### SNV-aware alignment

The cell barcodes and UMIs were extracted using UMI-tools from the pooled (per donor) raw sequencing reads [26]. The pooled sequencing reads were aligned to the latest version of the human genome reference (GRCh38, Dec 2013) using STAR v.2.7.3.c in 2-pass mode with transcript annotations from the assembly GRCh38.79 [27]. The alignments were deduplicated retaining the reads with the highest alignment scores [26]. SNVs were called in the pooled deduplicated alignments using GATK v.4.1.4.1 [18]. To identify heterozygous SNV positions qualified for  $\text{VAF}_{\text{RNA}}$  analysis, we applied a series of filtering steps. Specifically, heterozygous SNVs were selected based on the presence of minimum of 50 high-quality reads supporting both (reference and alternative) nucleotides in the pooled alignments. SNV loci were annotated using SeattleSeq v.13.00 (dbSNP build 153), and loci positioned in repetitive or intergenic regions were removed. The SNV lists were further filtered based on the following requirements: QUAL (Phred-scaled probability)  $> 100$ , MQ (mapping quality)  $> 60$ , QD (quality by depth)  $> 2$ , and FS (Fisher's exact test estimated strand bias) = 0.000. The filtered SNV lists (per donor)

were then used as an input for a second, SNV-aware alignment using STAR-WASP [29].

### Gene expression estimation

To estimate gene expression, we first apply Feature-Count on the individual alignments to assess the row gene counts per cell [32]. We then normalize and scale the expression data using the *sctransform* function as implemented in Seurat v.3.0 [24, 33], which stabilizes the GE variance using regularized negative binomial regression, and outlines the most variable genes. The *sctransform* function integrates the previous Seurat functions *NormalizeData*, *ScaleData*, and *FindVariableFeatures*. The cell-feature distributions were then plotted to identify and filter out outliers and low-quality cells, which we defined as [1] cells with mitochondrial gene expression over 6%, cells with less than 3000 genes, and 3) cells with more than 8000 detected genes or  $> 12,500$  UMI counts, (to remove potential doublets), as well as cells with mitochondrial genes' expression higher than 6% of the total gene expression, and to correct for batch- and cell-cycle effects (See Fig. 2). We then integrate the datasets and use the function *FindIntegrationAnchors* to identify 'anchors' between pairs of datasets. This analysis resulted in 4099 common genes across the three samples, which we used to correct for batch effects. Next, we split the individual matrices by cell type, and for each cell type, genes which expression in 80% or more of the cells was within 20% or less from the top or bottom of the GE range, were filtered out. This retained between 2114 and 2442 per sample for scReQTL analyses. In addition, after examining the GE distribution across the cells (per cell type), the retained most variable genes were then used for scReQTL analyses (See Table 1).

### Cell type identification

To define individual cell types from the ADSCs, we used SingleR version 1.0.5 [34]. SingleR assigns cellular identity by comparison to reference whole transcriptome expression data sets of pure cell types. SingleR correlates the expression profile of each single cell to whole-transcriptome expression data from established cell types (Blueprint + ENCODE datasets). To select the expression profile most similar to the tested cells, the analysis is rerun iteratively, using only the top cell types from the previous step until only one cell type is retained. Comparing our datasets against 259 bulk RNA-seq profiles representing 24 main cell types and 43 subtypes, SingleR identified four major cell types: adipose cells and erythrocytes were found in all three samples, naïve-B-cells found in N5 and N7, and neutrophils, in N8 (See Fig. 3 and Table 1).

### VAF<sub>RNA</sub> estimation

VAF<sub>RNA</sub> is assessed from the individual alignments as we have previously described [28], using the high quality heterozygous SNV sites as inputs for ReadCounts [22]. At each position of interest, ReadCounts estimates the number of sequencing reads harboring the variant and the reference nucleotide ( $n_{\text{var}}$  and  $n_{\text{ref}}$  respectively), calculates VAF<sub>RNA</sub> ( $\text{VAF}_{\text{RNA}} = n_{\text{var}} / (n_{\text{var}} + n_{\text{ref}})$ ), and filters out positions not covered by the user-defined minimum number of reads (minR); minR is constant across the genome [22]. For the herein presented analysis, we used minR > 10. To qualify for scReQTL, a variant is required to have variable VAF<sub>RNA</sub> from a minimum of 20 cells from the same cell type (per donor). The VAF<sub>RNA</sub> distribution is then examined and loci with non-variable VAF<sub>RNA</sub> are filtered out. Loci were considered non-variable if: (1) over 75% of the VAF<sub>RNA</sub> values are in the range of  $0.5 \pm 0.1$  (corresponding to stable biallelic expression), and (2) over 75% of the VAF<sub>RNA</sub> values are in the ranges 0–0.25 or 0.75–1 (corresponding to predominantly monoallelic or skewed allelic expression).

### ScReQTL computations

*SNV-GE correlations* (scReQTLs) were computed for each donor, across the cells of each type separately. To qualify for scReQTLs analysis, an SNV locus is required to have informative and variable VAF<sub>RNA</sub> estimations (minR ≥ 10) from at least 20 cells per analysis. The variable VAF<sub>RNA</sub> were correlated to the normalized GE values of the most variable genes using a linear regression model as implemented in Matrix eQTL [17]. The top 15 principal components of the GE were used as covariates (Supplementary Figure 16). Cis and trans correlations were annotated as previously described for the bulk ReQTLs [25]. Briefly, because scReQTLs are assessed from transcripts, we assign cis-correlation based on the co-location of the SNV locus within the transcribed gene, using the gene coordinates [50]. All the scReQTLs including SNVs residing in genes different from the expression-correlated genes are annotated as trans-scReQTLs.

### Statistical analyses

Throughout the analysis we used the default statistical tests (with built-in multiple testing corrections) implemented in the used software packages (Seurat, SingleR, Matrix eQTL), where  $p$ -value of 0.05 was considered significant, unless otherwise stated. For estimation of significant scReQTL, we applied FDR as implemented in the Matrix eQTL package. Specifically, once Matrix eQTL discovers a set of significant gene-SNP pairs, it estimates a corresponding  $q$ -value (FDR) for each of them using Benjamini–Hochberg procedure under the assumption that the tests are independent or positively

correlated [17, 51]. For estimation of differences in overlap between scReQTL SNVs, GWAS and ClinVar, chi-square test was used. For assessment of enrichment of scReQTLs in known gene-gene interactions, a permutation test with 10,000 permutations was applied on the findings at minR = 10, and with 1000 permutations for the scReQTL analyses at minR < 10. For each permutation, a random set of gene-gene pairs of the same size as the observed data was selected. The  $p$ -value was defined as the fraction of permutations in which the number of gene-gene pairs found in the known interaction database was at least as great as the number found in the observed data.

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-020-07334-y>.

**Additional file 1.**

**Additional file 2.**

**Additional file 3.**

**Additional file 4.**

**Additional file 5.**

**Additional file 6.**

**Additional file 7.**

### Authors' contributions

AH: Conceptualization, analysis, writing and supervision; HL, NMP, LS, PB, NA, HI, JS, PS, KTA data management and processing, analysis, visualization, and writing—review and editing. The author(s) read and approved the final manuscript.

### Funding

This work was supported by McCormick Genomic and Proteomic Center (MGPC), The George Washington University; [MGPC\_PG2019 to AH].

### Availability of data and materials

All data generated or analyzed during this study are included in this published article and its supplementary information files.

### Ethics approval and consent to participate

Not Applicable.

### Consent for publication

Not Applicable.

### Competing interests

The authors declare that they have no competing interests.

### Author details

<sup>1</sup>McCormick Genomics and Proteomics Center, School of Medicine and Health Sciences, The George Washington University, Washington, DC 20037, USA. <sup>2</sup>Chinese Medicine Toxicological Laboratory, Institute of Traditional Chinese Medicine, Heilongjiang University of Chinese Medicine, Harbin 150040, People's Republic of China. <sup>3</sup>Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA. <sup>4</sup>Cancer Program, The Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. <sup>5</sup>Pritzker School of Medicine, Biological Sciences Division, The University of Chicago, Chicago, IL 60637, USA. <sup>6</sup>Translational Research Exchange at Exeter, University of Exeter, Exeter EX4 4QJ, UK. <sup>7</sup>EPSRC Centre for Predictive Modelling in Healthcare, University of Exeter, Exeter EX4 4QJ, UK. <sup>8</sup>Department of Mathematics & Living Systems Institute, University of Exeter, Stocker Road, Exeter EX4 4QD, UK. <sup>9</sup>Dept. of Bioinformatics and Mathematical Modelling,

Institute of Biophysics and Biomedical Engineering, Bulgarian Academy of Sciences, 105 Acad. G. Bonchev Str, 1113 Sofia, Bulgaria. <sup>10</sup>Department of Biochemistry and Molecular Medicine, Department of Biostatistics and Bioinformatics School of Medicine and Health Sciences, George Washington University, Washington, DC 20037, USA.

Received: 23 August 2020 Accepted: 16 December 2020

Published online: 08 January 2021

## References

- Kulkarni A, Anderson AG, Merullo DP, Konopka G. Beyond bulk: a review of single cell transcriptomics methodologies and applications. *Curr Opin Biotechnol*. 2019.
- Van Der Wijst MGP, Brugge H, De Vries DH, Deelen P, Swertz MA, Franke L. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat Genet*. 2018.
- Villani AC, Satija R, Reynolds G, Sarkizova S, Shekhar K, Fletcher J, et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*. 2017.
- La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, et al. RNA velocity of single cells. *Nature*. 2018.
- Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, et al. From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Res*. 2014.
- Gallivan CP, Ren H, Read EL. Analysis of Single-Cell Gene Pair Coexpression Landscapes by Stochastic Kinetic Modeling Reveals Gene-Pair Interactions in Development. *Front Genet*. 2020.
- Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. *Nat Rev Gene*. 2015.
- Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. *Nature*. 2017.
- Akbayan S, Liu C, Knowles JA, Vaccarino FM, Farnham PJ, Crawford GE, et al. The PsychENCODE project. *Nat Neurosci*. 2015.
- De Jager PL, Hacohen N, Mathis D, Regev A, Stranger BE, Benoist C. ImmVar project: insights and design considerations for future studies of "healthy" immune variation. *Semin Immunol*. 2015.
- Lloyd-Jones LR, Holloway A, McRae A, Yang J, Small K, Zhao J, et al. The genetic architecture of gene expression in peripheral blood. *Am J Hum Genet*. 2017.
- Chen L, Ge B, Casale FP, Vasquez L, Kwan T, Garrido-Martín D, et al. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell*. 2016.
- Cuomo ASE, Seaton DD, McCarthy DJ, Martinez I, Bonder MJ, Garcia-Bernardo J, et al. Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat Commun*. 2020.
- Sarkar AK, Tung PY, Blischak JD, Burnett JE, Li YI, Stephens M, et al. Discovery and characterization of variance QTLs in human induced pluripotent stem cells. *PLoS Genet*. 2019.
- Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol*. 2018.
- Hu Y, Zhang X. SCell: an R package for identifying eQTL from single-cell parallel sequencing data. *bioRxiv*. 2018.
- Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*. 2012.
- Auwerda Mauricio O. GAV der C, Hartl C, Poplin R, Angel G del, Levy-Moonshine A, Jordan T, et al. From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr Protoc Bioinforma*. 2002.
- Deelen P, Zernakova DV, de Haan M, van der Sijde M, Bonder MJ, Karjalainen J, et al. Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels. *Genome Med*. 2015.
- Piskol R, Ramaswami G, Li JB. Reliable identification of genomic variants from RNA-seq data. *Am J Hum Genet*. 2013.
- Horvath A, Pakala SB, Mudvari P, SDN R, Ohshiro K, Casimiro S, et al. Novel insights into breast cancer genetic variance through RNA sequencing. *Sci Rep*. 2013.
- Movassagh M, Alomran N, Mudvari P, Dede M, Dede C, Kowsari K, et al. RNA2DNAAlign: nucleotide resolution allele asymmetries through quantitative assessment of RNA and DNA paired sequencing data. *Nucleic Acids Res*. 2016.
- Slowinski P, Li M, Restrepo P, Alomran N, Spurr LF, Miller C, et al. GeTallele: A Method for Analysis of DNA and RNA Allele Frequency Distributions. *Front Bioeng Biotechnol*. 2020.
- Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*. 2018.
- Spurr L, Alomran N, Bousounis P, Reece-Stremtan D, Prashant NM, Liu H, et al. ReQTL: Identifying correlations between expressed SNVs and gene expression using RNA-sequencing data. *Bioinformatics*. 2019.
- Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res*. 2017.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013.
- Prashant NM, Liu H, Bousounis P, Spurr L, Alomran N, Ibeawuchi H, et al. Estimating the allele-specific expression of snvs from 10x genomics single-cell rna-sequencing data. *Genes (Basel)*. 2020.
- Van De Geijn B, Mcvicker G, Gilad Y, Pritchard JK. WASP: Allele-specific software for robust molecular quantitative trait locus discovery. *Nat Method*. 2015.
- Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res*. 2019.
- Picardi E, D'Erchia AM, Lo GC, Pesole G. REDportal: A comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Res*. 2017.
- Liao Y, Smyth GK, Shi W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014.
- Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol*. 2019.
- D. A, A.P. L, L. E. W, V. F, A. H, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol*. 2019.
- Prashant NM, Alomran N, Chen Y, Liu H, Bousounis P, Movassagh M, et al. SCReadCounts: Estimation of cell-level SNVs from scRNA-seq data. *bioRxiv*. 2020;2020.11.23.394569 Available from: <http://biorxiv.org/content/early/2020/11/23/2020.11.23.394569.abstract>.
- von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, et al. STRING: Known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*. 2005.
- Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 2019.
- Shang L, Smith JA, Zhou X. Leveraging Gene Co-expression Patterns to Infer Trait-Relevant Tissues in Genome-wide Association Studies. *bioRxiv*. 2019.
- Kishore S, Lubner S, Zavolan M. Deciphering the role of RNA-binding proteins in the post-transcriptional control of gene expression. *Brief Funct Genomics*. 2010.
- Hausser J, Zavolan M. Identification and consequences of miRNA-target interactions-beyond repression of gene expression. *Nat Rev Genet*. 2014.
- Chatterjee S, Pal JK. Role of 5'- and 3'-untranslated regions of mRNAs in human diseases. *Biol Cell*. 2009.
- Maiti GP, Ghosh A, Mondal P, Baral A, Datta S, Samadder S, et al. SNP rs1049430 in the 3'-UTR of SH3GL2 regulates its expression: clinical and prognostic implications in head and neck squamous cell carcinoma. *Biochim Biophys Acta - Mol Basis Dis*. 2015.
- Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. 2018.
- Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010.
- Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, et al. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*. 2009.
- Brandt DY, Aguiar VRC, Bitarello BD, Nunes K, Goudet J, Meyer D. Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 genomes project phase I data. *G3 Genes Genomes Genet*. 2015.

47. Savova V, Patsenker J, Vigneau S, Gimelbrant AA. dbMAE: The database of autosomal monoallelic expression. *Nucleic Acids Res.* 2016.
48. van der Wijst MG, de Vries DH, Groot HE, Trynka G, Hon C-C, Bonder M-J, et al. The single-cell eQTLGen consortium. *Elife.* 2020; [cited 2020 Apr 6];9. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/32149610>.
49. X. L, Q. X, F. X, J. H, N. Y, Q. Z, et al. Single-cell RNA-seq of cultured human adipose-derived mesenchymal stem cells. *Sci Data.* 2019.
50. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt. *Nat Protoc.* 2009.
51. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Ser B.* 1995.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

