Genome Biology

METHOD Open Access

DeCAF: a novel method to identify cell-type specific regulatory variants and their role in cancer risk

Cynthia A. Kalita[1] and Alexander Gusev[1,2,3*]

*Correspondence: alexander_gusev@dfci.harvard.edu
[1] Division of Population Sciences, Dana–Farber Cancer Institute & Harvard Medical School, Boston, USA
[2] The Broad Institute, Boston, USA
[3] Division of Genetics, Brigham & Women's Hospital, Boston, USA

Abstract

Here, we propose DeCAF (DEconvoluted cell type Allele specific Function), a new method to identify cell-fraction (cf) QTLs in tumors by leveraging both allelic and total expression information. Applying DeCAF to RNA-seq data from TCGA, we identify 3664 genes with cfQTLs (at 10% FDR) in 14 cell types, a 5.63× increase in discovery over conventional interaction-eQTL mapping. cfQTLs replicated in external cell-type-specific eQTL data are more enriched for cancer risk than conventional eQTLs. Our new method, DeCAF, empowers the discovery of biologically meaningful cfQTLs from bulk RNA-seq data in moderately sized studies.

Background

Genome-wide association studies (GWAS) have been instrumental in identifying a large number of genetic variants associated with risk for many diseases including cancer [1–7]. However, as the majority of GWAS associations are non-coding variants without clear function, the mechanism of action is typically unknown. Expression quantitative trait loci (eQTLs) have been instrumental in linking genetic variation to effects in gene expression [8–11], recently identifying putative susceptibility genes for ovarian cancer [12], prostate cancer [13], and breast cancer [14]. Recently, it has been observed that some eQTL effects can be observed only in specific contexts and often vary across tissue and cell types [15–18]. In the context of cancer, the cell types in the tumor/microenvironment can have substantially different functions [19, 20]: with CD4 and CD8 T cells driving cytotoxic anticancer immunity [19, 21], while regulatory T cells are associated with immune suppression and homeostasis [22, 23]. Identifying and quantifying cell-type-specific eQTLs in tumors is thus a critical step to understanding germline cancer mechanisms and germline-somatic interactions.

To date, cell-type-specific studies have been limited in size due to cost and labor associated with selecting a pure subset (i.e., cell sorting) and therefore have weak power to identify QTLs [24]. While emerging single-cell technologies have the potential to precisely measure expression in specific cell populations, this approach remains too expensive to measure across hundreds of individuals and exhibits very sparse expression, with most genes unexpressed in a given cell. This has led to the development of bulk RNA-seq deconvolution methods which generally work by defining gene sets in reference data from a pure population and ranking the expression of these in the target sample [25, 26], or modeling the target sample as a mixture of signatures [27–30]. These cell fraction estimates can additionally be incorporated into eQTL analyses to identify cell-fraction-specific effects (cfQTLs) [31–35]. However, by testing for an interaction effect on a very noisy outcome, such studies typically require sample sizes in the thousands to achieve adequate power (particularly for cell types present at low frequency) [33].

In addition to conventional eQTLs, genetic effects on expression can be measured by quantifying the ratio of RNA-seq reads at heterozygous variants in exons. A significant departure from a 50% allelic ratio is indicative of a cis-genetic effect on expression, and referred to as allelic imbalance (AI). This approach benefits from being able to control for trans-variation, and, because it measures allelic effects within individuals and not between individuals, can harness power from read depth [36–44]. AI has also been leveraged to identify genes undergoing gene-environment interactions[45, 46], tumor/normal regulatory differences[47] and, recently, cell type specificity [48]. Methods exist and have been proven to work well, which jointly model AI and eQTL effect sizes, including TReCASE [49], RASQUAL [50] and BaseQTL [51]. However, the integration of total expression and AI to detect cfQTLs has been largely unexplored.

Here we propose DeCAF (DEconvoluted cell type Allele specific Function), a method that increases power to identify cfQTLs in bulk data by combining AI and total expression signals. We use DeCAF to identify thousands of allelic cfQTLs in multiple cell types as well as tumor cells in renal cell carcinoma, which we replicate using external cell/tissue type data and link to GWAS risk variants. DeCAF identified many cfQTLs that were not observed in a conventional bulk eQTL analysis, suggesting that this framework may greatly enrich our understanding of cis regulatory mechanisms.

Results

Overview of DeCAF for cfQTL discovery

We sought to identify cfQTL variants whose effect on gene expression varied across bulk samples according to cell type fraction (Fig. 1a) (not to be confused with having a variant that increases proliferation of a certain cell-type in a heterogeneous tissue). As in previous work[52–56], such variants can be identified through a linear interaction model with deconvoluted (individual-level) cell fraction, where a significant interaction between genotype and cell fraction is indicative of an effect size that differs with differing cell-fractions. We introduce an analog of the interaction model for AI data, which we combine with the conventional interaction (ieQTL) model. The ieQTL model is a linear regression of $y \sim \mu + \beta x + f + \beta_f x * f$, where y is total expression, μ is the expression mean, x is the genotype encoding the number of alternative alleles, β is the standard eQTL effect (expected to be zero under the null), f is the individual-level cell fraction, and β_f measures the cell fraction specificity. Our proposed AI model relates the number of reference and alternative allelic reads in heterozygous individuals to their cell fraction as a regression of $REF, ALT \sim \mu_a + \alpha_f f$, where μ_a is the mean allelic fraction across individuals and

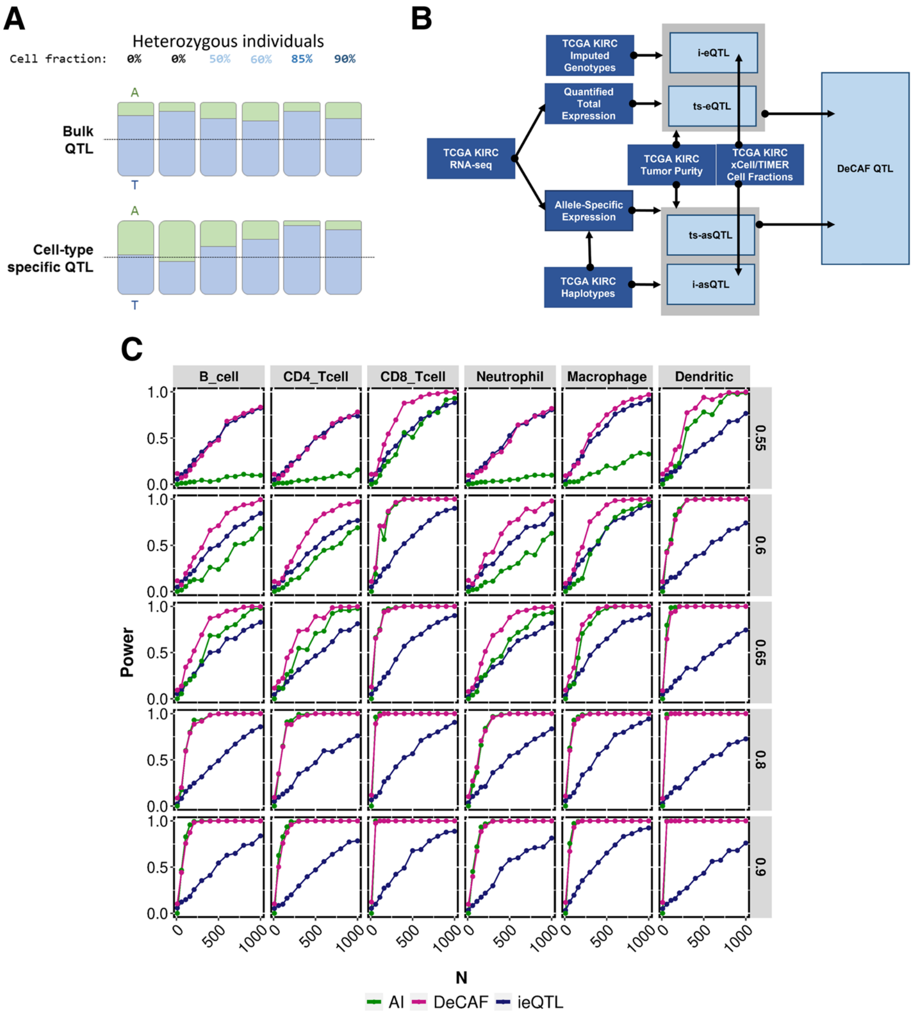

Fig. 1 Identifying cfQTLs. **A** Depiction of AI changing across increasing cell fraction. **B** Flowchart showing the data entering into identifying DeCAF QTLs. Dark blue boxes represent outside data and light blue boxes represent internal calculated data. **C** Comparing cfQTL testing methods for simulated data. Plot depicting the power (*y*-axis) for varying number of individuals (*x*-axis) and either ieQTL (green), iAI (pink), or DeCAF combined method (blue). Each column represents a cell type (denoting simulations using real deconvoluted cell fractions from TIMER) and each row is a different allelic fraction (effect size of AI)

α_f is the modifying effect of the cell-fraction f and our primary effect of interest. Under the null μ_a is 0.5 and deviations from 0.5 indicate an AI effect in the bulk population; likewise, a significantly non-zero α_f indicates the cell fraction modifies the allelic fraction effect and is indicative of a cfQTL. The parameters of the AI model can be estimated by binomial/logistic regression, but we additionally implement beta-binomial regression to accommodate read overdispersion that is common in RNA-seq data [57] (see the "Methods" section). Finally, the β_f and α_f estimates (which are both oriented towards the alternative allele) are combined by meta-analysis to form the complete DeCAF test statistic. In practice, we additionally include a tumor purity term in both models to jointly estimate tumor-specific effects (tsQTLs) [31] and determine whether a variant is acting in the tumor or the microenvironment. We note that, in contrast to some analyses of AI in individual samples, our model always tests for consistent effects in the population at a given variant, polarized to the variant allele. We tested causal variants outside the gene by using haplotype phasing (heterozygous variants were assigned the reads on their corresponding haplotype as in [47]).

Somatic copy number alterations can lead to extra variation and non-genetically driven AI. Indeed, the earliest applications of allelic imbalance in cancer were to identify copy number changes from sequenced DNA. As CNVs have been estimated in TCGA data from tumor/normal genotyping, the local CNV estimate from each individual was included in the model as a fixed effect covariate, to account for an offset in the expression due to carrying a CNV. Accounting for CNV as a covariate was well calibrated and led to only a 1.12× reduction in power (Additional file 2: Fig. S3). In the real data, we additionally conservatively masked out any individuals with deep CNVs (>0.1 segment mean, or Log2 ratios of the tumor copy number to the normal copy number) for a given test.

Evaluation of DeCAF in simulation

We performed wide-ranging simulations reflecting conditions found in real data and evaluated the power of the interaction eQTL, interaction AI, and DeCAF tests (see the "Methods" section). These included allelic effect size, minor allele frequency, sequencing depth, CNV, number of individuals, and read overdispersion. In particular, we sampled cell fractions from the real TIMER-inferred fractions in TCGA across 6 cell-types, to investigate the performance of this approach under realistic cell proportions. Under the null hypothesis where there is no effect, all of the methods were well calibrated and produced null results (Additional file 2: Figs. S1a and S2). Under the alternative hypothesis, DeCAF consistently met or outperformed the power of the conventional interaction eQTL test both under cell fractions generated from a uniform distribution (Additional file 2: S1b) and real cell fractions from TIMER inference in the TCGA KIRC data (Fig. 1c). requiring 0.56× fewer samples on average to reach >75% power (Fig. 1c). In 42% of the simulated models, DeCAF achieved >75% power at <600 individuals (the typical TCGA study size) while the eQTL test did not (Fig. 1c).

Identifying a multitude of QTLs

We applied DeCAF to genotype and RNA-seq data from 503 RCC tumors in TCGA (Fig. 1b) using cell fractions from xCell [26] and TIMER [30] as well as tumor purity (previously estimated by [58]) to account for differences in purity and identify tsQTLs [31] (Fig. 1b; see the "Methods" section). The eQTL and AI effects were highly and significantly correlated in the marginal DeCAF results ($\rho = 0.8$, p value $< 2.2^{-16}$, Additional file 2: Fig. S6). To confirm that DeCAF assumptions (AI and QTL effects are shared and AI effects are not unexpectedly noisy) hold in real data, we tested whether DeCAF would identify more significant effects than the interaction eQTL test. Across all tested cell types, DeCAF identified 5.63× more genes with significant cfQTLs than a conventional interaction QTL analysis. By cell fraction type, DeCAF identified 1753 (xCell), 1220 (TIMER) cfQTL genes, and 691 (purity) tsQTL genes at 10% FDR (defined as unique genes containing a significant corresponding QTL; Additional file 1: Table S1). For comparison, conventional interaction eQTL mapping identified 310 (xCell), 218 (TIMER), and 186 (purity) tsQTL genes (Fig. 2). Applying DeCAF without a cell fraction term identified a total of 3654 marginal eQTL genes, a 1.7× increase over the 2150 eQTL genes identified with the conventional eQTL linear model, suggesting that incorporating AI particularly increases power for interaction analyses where the effective sample sizes are lower. Across the cfQTL genes, we observed 42 genes with significant effects in all cell types while the rest of genes were largely observed in a single cell type (270 genes, Fig. 2b) and were not substantially shared across the xCell and TIMER results. We thus focus on cfQTLs from both methods moving forward.

Of the significant DeCAF cfQTL genes, 56% were also significant DeCAF marginal eQTL genes and 41% were also significant DeCAF tsQTL genes. In contrast, of the significant DeCAF tsQTL genes, 82% were also DeCAF marginal eQTL genes, suggesting that tumor-specific effects often manifest as apparent bulk eQTLs (although they make up a minority of all marginal eQTLs, as previously observed [31]). Remarkably, while we saw a large portion of shared eGenes between cfQTLs, tsQTLs, and marginal eQTLs, the SNPs associated with each eGene were mostly not shared (Fig. 2d). Of the significant DeCAF cfQTLs, 1% were also significant DeCAF marginal eQTLs and 1% were also significant DeCAF tsQTLs. Most strikingly, only 3 SNP:eGene pairs are shared between cfQTLs, tsQTLs, and marginal eQTLs. While differences in the lead eQTL SNP may be due to statistical noise, we further observed that for genes with an eQTL and cfQTL, 92% of SNP pairs had $r^2 < 0.5$ (LD), suggesting that DeCAF identified largely independent genetic signals that were not detectable from conventional bulk analyses.

To provide visual intuition for the DeCAF approach, we identified genes with a significant (Spearman) correlation between cell fraction and individual allelic fraction of their corresponding cfQTL (Fig. 3c). We highlight the cfQTL effect of rs117689555 on the expression of GOLGA6L1 in the context of neutrophils ($\rho = 0.92$, p value = 0.001). A striking linear relationship can be observed between the neutrophil fraction in each heterozygous sample and the allele fraction, shifting from imbalance for the reference allele to imbalance for the alternative allele (i.e., a complete reversal of effect-size direction). We note that this correlation-based approach was statistically much weaker than DeCAF because it did not incorporate read count or read overdispersion, and so is used here only to provide intuition.

In principle, somatic variants will introduce additional noise into the analysis by sporadically knocking out or activating driver genes, but they would not be expected to lead to false positives in the absence of germline-somatic interactions. Analyzing somatic SNV calling data from TCGA whole-exome sequencing, we found that a total of 96 genes had >10 somatic mutation carriers (sufficient to influence population-scale QTL discovery), which overlapped only 9/3654 of the significant marginal DeCAF eGenes we previously identified. For these 9 genes that were significant marginal DeCAF eGenes and had >10 SNV carriers, we tested each lead SNP for a germline-somatic interaction by adding somatic carrier status as a binary interaction term (similar to how purity is included as an interaction term for tsQTL analysis). None of the 9 interaction terms were statistically significant after Bonferroni correction, indicating no significant differences in AI/QTL effect size between carriers and non-carriers. In sum, we find that recurrent SNVs are unlikely to influence the overall DeCAF results due to the generally low number of highly recurrently mutated eGenes, and a formal interaction test of these genes did not modify the associations.

DeCAF can also identify cell-fraction effects in normal tissue. As a demonstration, we applied DeCAF to matched kidney normal RNA-seq data from the same TCGA KIRC cohort ($N = 70$). This much smaller cohort still yielded a total of 287 significant cfQTL eGenes identified by DeCAF, compared to 201 cfQTL eGenes by conventional ieQTL

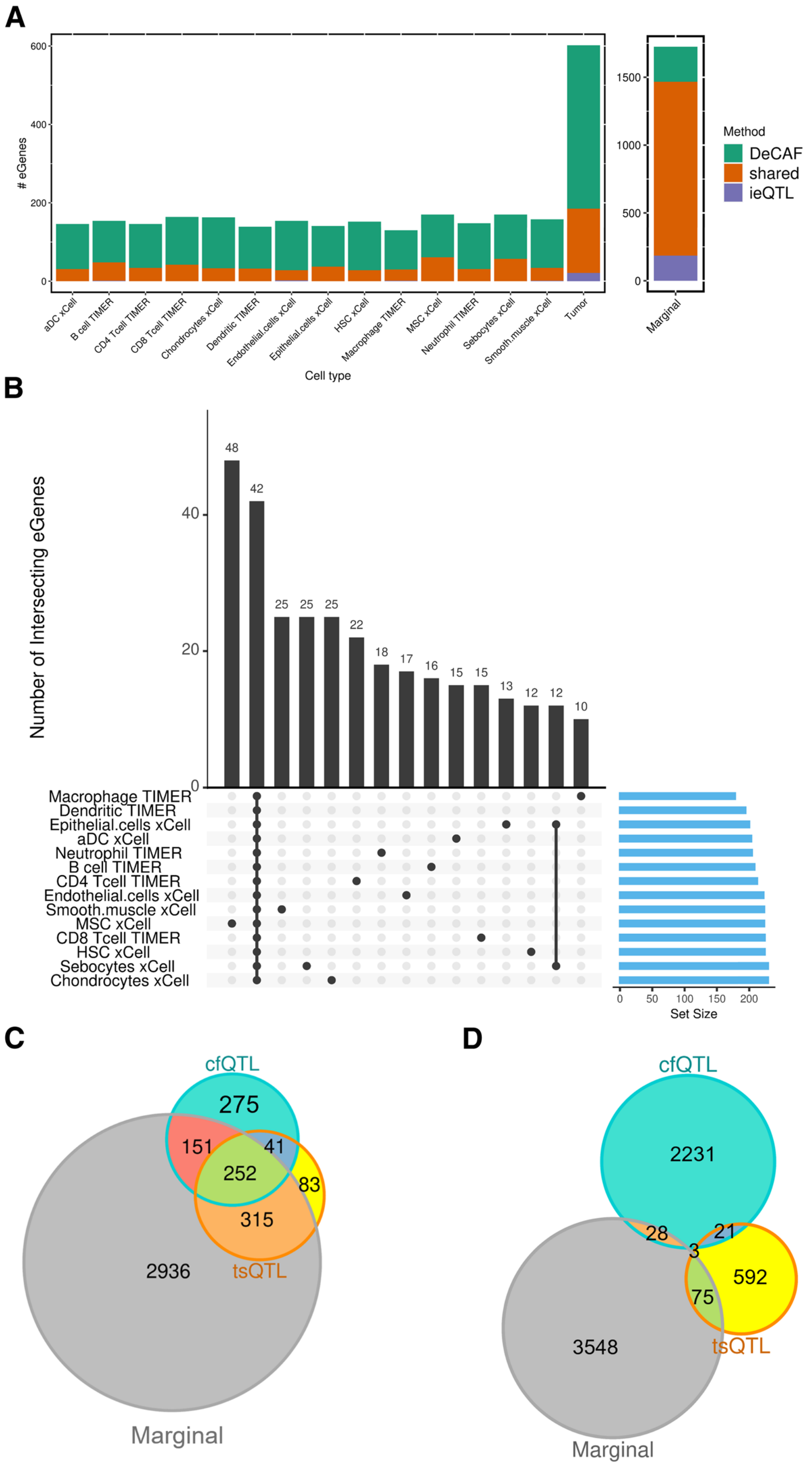

Fig. 2 cfQTL testing methods. **A** Comparing eGenes from DeCAF vs standard interaction QTL (ieQTL) methods. Plot depicting the number of significant cfQTL genes (y-axis) for each cell type (*x*-axis) for DeCAF only (teal), standard interaction eQTL (ieQTL) only (purple), and shared (orange) tests. **B** Intersection of significant eGenes from cancer and cell fraction tests for xCell and TIMER deconvolutions. **C** Overlap of eGenes from cfQTLs, tsQTLs, and marginal eQTLs. **D** Overlap of significant snp:gene pairs from cfQTLs, tsQTLs, and marginal eQTLs

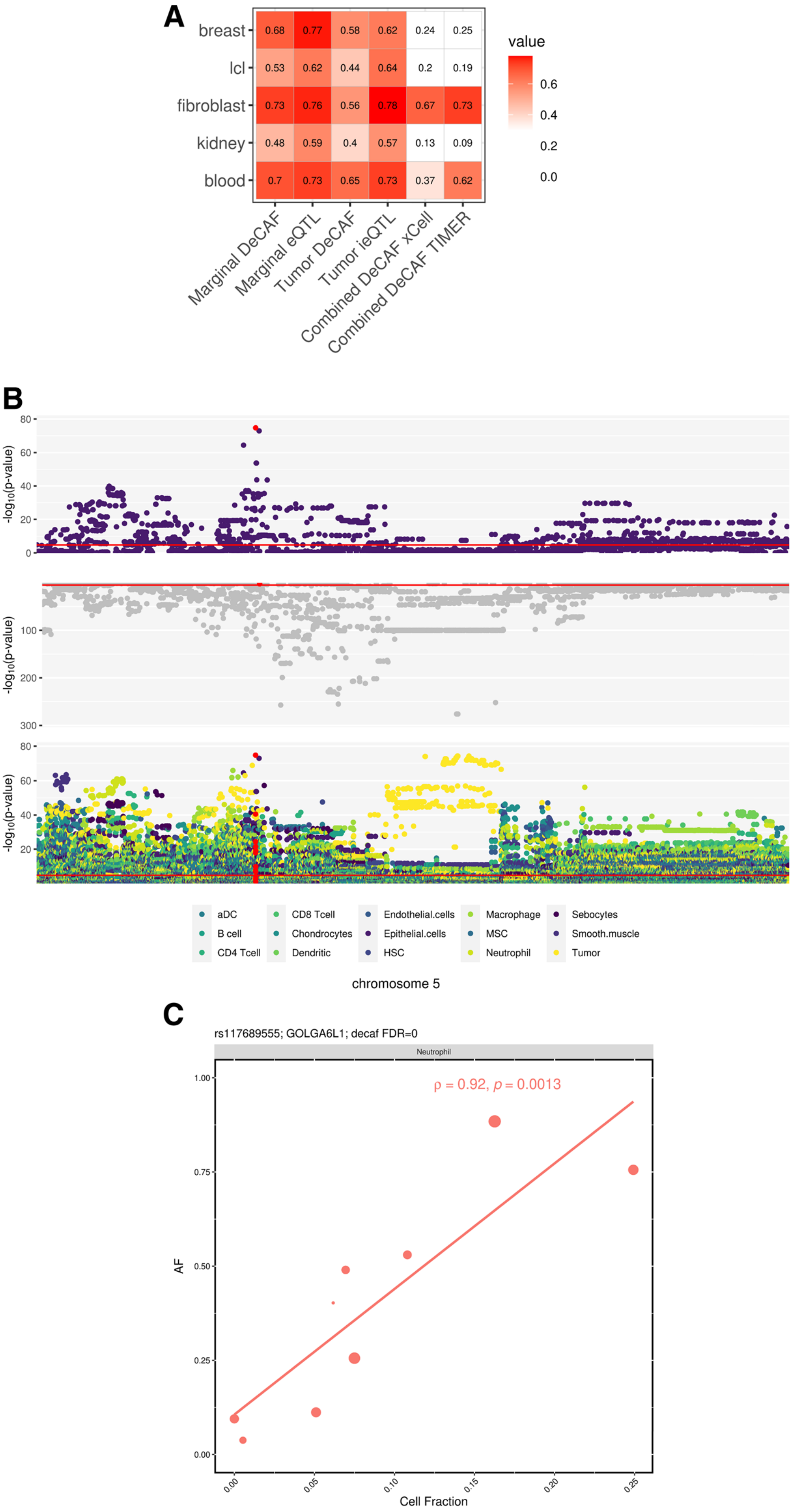

Fig. 3 cfQTL validation. **A** π_1 replication of marginal, cancer, and TIMER and xCell cfQTLs in GTEx. **B** ERAP2 manhattan plot with top: Epithelial cell cfQTL (most significant cell type), middle: Marginal, and bottom: all cfQTL and tsQTL significance levels. Red highlighted SNP is rs26481. **C** Correlation between cell fraction (*x*-axis) and allelic fraction (*y*-axis). Point size indicates number of reads

(FDR 10%). A larger proportion of the identified effects were shared (71% shared cfQTLs for normal versus 32% for tumor; Additional file 2: Fig. S7), likely due to the much lower sample size and lower number of heterozygous individuals for the AI component as can be seen in simulations (Additional file 2: Fig. S1b).

Replication of DeCAF cfQTLs in external studies

Reasoning that the DeCAF cfQTLs would capture components of the tumor immune microenvironment, we sought to replicate them in external non-cancer data. First, we turned to eQTLs in the GTEx (v8 [59]), detected in bulk RNA-seq from whole blood (N = 670), lymphoblastoid cells (LCLs, N = 147), fibroblasts (N = 483), kidney(N = 73), and breast (N = 396) tissues - selected as a sampling of potential cells/tissues in the RCC microenvironment. Across all DeCAF cfQTLs combined, we observed substantial replication across multiple GTEx tissues, with the highest replication rates in fibroblasts: $\pi_1 = 0.67$ and $\pi_1 = 0.73$ for xCell and TIMER-based results respectively (see Methods, Fig. 4). These replication rates were comparable or higher than the replication of marginal eQTLs into GTEx data, where π_1 ranged from 0.59 (kidney) to 0.77 (breast), demonstrating that cfQTLs have similar true positive rates when evaluated in the appropriate target tissue. Moreover, GTEx breast, kidney, and LCLs exhibited the lowest cfQTL replication rates (all $\pi_1 \leq 0.25$), demonstrating that cfQTLs were also specific to a target tissue (in this case fibroblast and blood), in contrast to marginal eQTLs that generally replicated in all target tissues (all $\pi_1 > 0.59$). Taken together, these results suggest that cfQTLs may be capturing genetic effects enriched in cancer-associated fibroblasts and depleted in organ tissues or B-lymphocytes (though we note that replication rates will be biased by statistical power in the target tissue, which is a function of sample size and data quality). Within the individual cell types tested by DeCAF, we found a wide range of replication rates in GTEx tissues with no clear cell-tissue relationships, likely due to the relatively low number of cfQTLs within a given cell type or heterogeneity in the target GTEx samples (Additional file 2: Fig. S10). We note that while GTEx did calculate AI signal [60], it was not used in the identification of these population-level eQTLs, further underscoring the DeCAF replications given the different statistical models employed.

We carried out a second replication using cell-type-specific eQTL data from the BLUEPRINT consortium [24]. Notably, BLUEPRINT data (but not eQTLs) were used to train the xCell deconvolution scores and thus provide an apples-to-apples comparison of pure versus deconvoluted eQTLs in independent data. Across all DeCAF cfQTLs combined, replication rates were notably lower than in GTEx, with π_1 in the 0.30–0.40 range (Additional file 2: Fig. S9). As in GTEx, we again did not observe clear trends between the DeCAF cfQTL cell type and the target eQTL cell type (for example, the strongest replication was from Mesenchymal stem cells (MSCs) to CD4 T cells, with $\pi_1 = 0.60$). The lower replication of BLUEPRINT versus GTEx may be explained by the relatively small sample size of the BLUEPRINT, and thus lower power to detect weak, cell-type-specific effects. Overall, these replications were generally comparable to the replication π_1 range of 0.3–0.67 observed in a recent GTEx ieQTL analysis of similarly sized target studies [35], suggesting that replicating cfQTLs may be generally more challenging than marginal eQTLs.

Turning to tsQTLs, we were surprised to find substantial replication into both external eQTL studies (Fig. 3). The strongest replication in GTEx was in blood ($\pi_1 = 0.65$) and

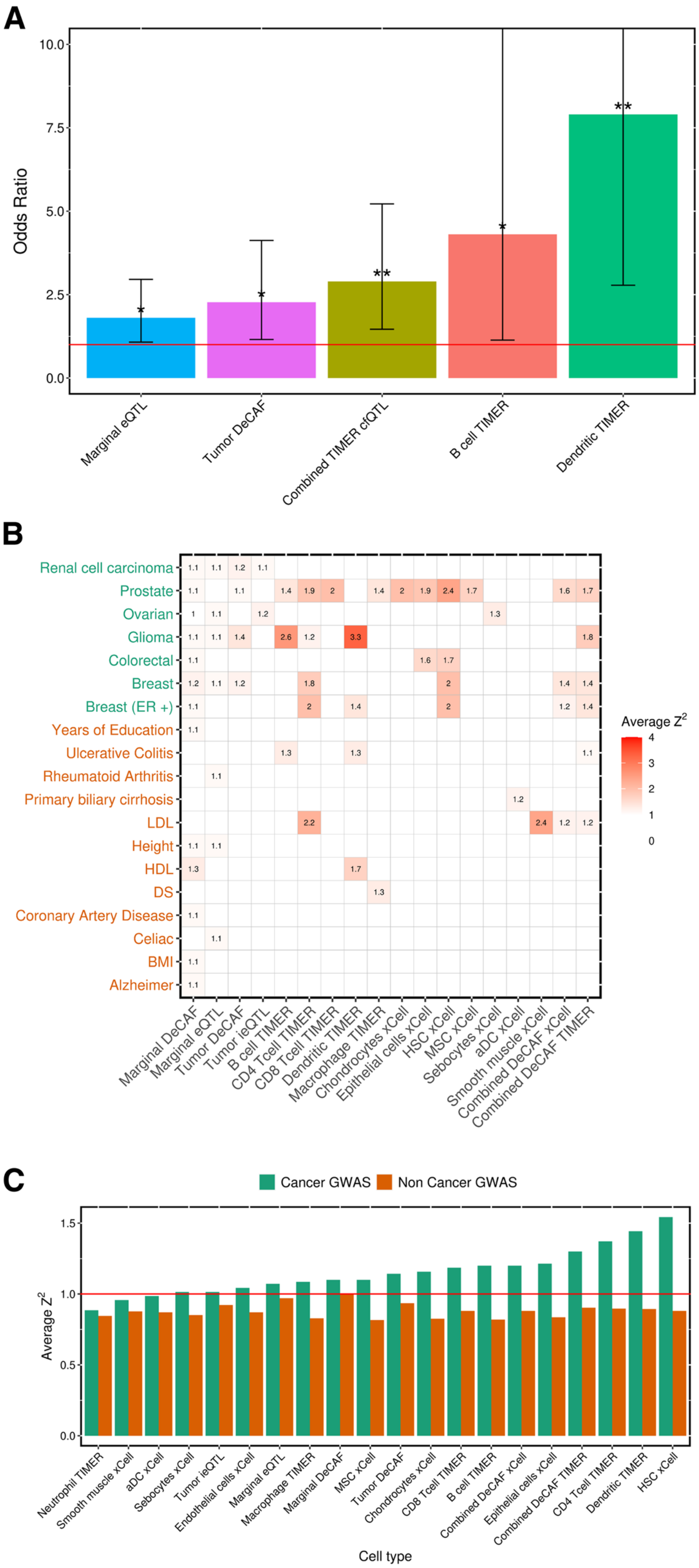

Fig. 4 Enrichment of cfQTLs in GWAS. **A** Significant enrichment of cfQTLs (FDR 20%) in RCC GWAS (*p* value<0.001) from a fisher's test. Two stars above the bar represents significant after Bonferroni correction, one star represents nominal significance (*p* value>0.05). **B** Heatmap showing the average Z^2 enrichment of cfQTLs in GWAS. Insignificant results (no reported value) are based on *p* value>0.05. **C** Comparing average Z^2 (*y*-axis) enrichment from cancer (green) vs non-cancer (orange) GWAS traits for each cell type (*x*-axis)

the strongest replication in BLUEPRINT was in neutrophils and monocytes ($\pi_1 = 0.70$). Strikingly, DeCAF tsQTL replication was higher than DeCAF cfQTL in GTEx breast, LCL, kidney, and blood as well as all BLUEPRINT tissues, and was comparable to that of marginal DeCAF QTLs (Additional file 2: Fig. S9). Given the highest replication was observed in the BLUEPRINT immune cell types, we hypothesize that tsQTLs may be capturing genetic effects from or due to tumor infiltrating lymphocytes (TILs), or a mix of TILs and the tumor cell of origin. These enrichment patterns were broadly similar when testing the smaller number of conventional tumor interaction eQTLs, underscoring the robustness of this effect to detection methodology.

Enrichment of DeCAF cfQTLs in GWAS

We next sought to quantify the relationship between cfQTLs and genome-wide association study (GWAS) variants, which may shed light on the downstream phenotypic mechanisms of cfQTLs. We analyzed the GWAS effect-size (squared Z-score) enrichments across 7 common cancer GWAS [1–7, 61]. The enrichments compared significant cfQTLs to random non-significant cfQTLs as a background to account for the frequency and LD of tested variants, with significance quantified through background sampling (see the "Methods" section). Every cancer GWAS had at least one significant cfQTL enrichment (Fig. 4b) and most cfQTL cell types exhibited enrichment in cancer GWAS (mean 1.16× s.e. 0.04). Notably, cfQTLs consistently showed higher enrichments than marginal DeCAF eQTLs (mean 1.10× s.e. 0.02). Compared to cfQTLs, tsQTLs generally had weaker enrichments, but were still significant in 4/7 cancers. For example, all eight cell types with significant enrichment in prostate cancer GWAS (mean 1.84× s.e. 0.12) had a higher magnitude of enrichment than the corresponding tsQTLs (mean 1.10×). As a comparison set of phenotypes, we expanded our analyses to 12 non-cancer GWAS from recent large-scale biobank studies [61, 62]. We observed broadly lower enrichments in these non-cancer GWAS (mean cfQTL enrichment 0.86× s.e. 0.01), with only 5/12 non-cancer GWAS traits having any significant (p value<0.05) cfQTL enrichments (compared to 7/7 cancer traits), covering a broad range of trait types (Fig. 4b). Overall, cancer GWAS traits had a stronger mean enrichment for every cell type than non-cancer GWAS (Fig. 4c). In sum, cfQTLs are more directly relevant to cancer GWAS mechanisms than eQTLs across a wide range of cancers.

We next focused on enrichments specific to the RCC GWAS [1] where we expected more biologically plausible cell-type-specific cfQTL enrichments. In the previous analysis comparing average effect sizes to non-imbalanced cfQTLs, only tsQTLs were significantly enriched for the RCC GWAS. We thus turned to a more sensitive analysis based on the enrichment of significant versus non-significant effects evaluated by Fisher's Exact Test (see the "Methods" section). After testing multiple DeCAF FDR and GWAS p-value thresholds we found that a more relaxed threshold (see the "Methods" section) produced the most confident results (smallest 95% confidence intervals) while identifying the most commonly seen significant cell types (Additional file 2: Fig. S11a). The core results were relatively stable across thresholds. We identified significant (Bonferoni p value<0.05) enrichment for cfQTLs in (TIMER) dendritic cells (OR=7.90, p value=1.86^{-4}) and all combined TIMER cfQTLs (OR=2.89, p value=1.58^{-3}) (Fig. 5a, Additional file 2: Fig. S11b), with nominal significance for cfQTLs in (TIMER) B cells (OR=4.30,p value=0.02), tsQTLs (OR=2.27, p value=0.01), and marginal eQTLs

(OR=1.80, *p* value=0.02). Again, marginal eQTLs were less enriched than other significant cfQTLs. Interestingly, the RCC GWAS enrichment was generally stronger for cfQTLs in individual cell types than across all cell types, suggesting that RCC GWAS variants may be enriched for specific cell types rather than generic cell-type-specific variants (although the statistical power was too low to identify significant differences between cell types in this analysis).

Biological interpretation of cfQTL examples

Lastly, we highlight two specific examples of cfQTLs and tsQTLs that showcase the utility of this framework.

We identified a highly significant cfQTL at rs26481 and *ERAP2* in epithelial cells (FDR = 5.5^{-69}, Fig. 3b). Intriguingly, rs26481 was clearly an independent signal from the top marginal eQTL for *ERAP2* (rs2927610; FDR = 0), with an r^2 of 0.0008 between the two variants. The marginal eQTL signal also coincided with a highly significant tsQTL, indicative of two independent genetic mechanisms (microenvironment and tumor) operating through *ERAP2*. Interestingly, a previous study found that ERAP2 expression was limited to epithelial cells in normal tissue, and loss of this enzyme was found in 86% of tumors [63]. This loss of *ERAP2* was also significantly linked to lack of HLA class I molecules [63]. In renal cancers, *ERAP2* was found to be frequently downregulated compared to normal

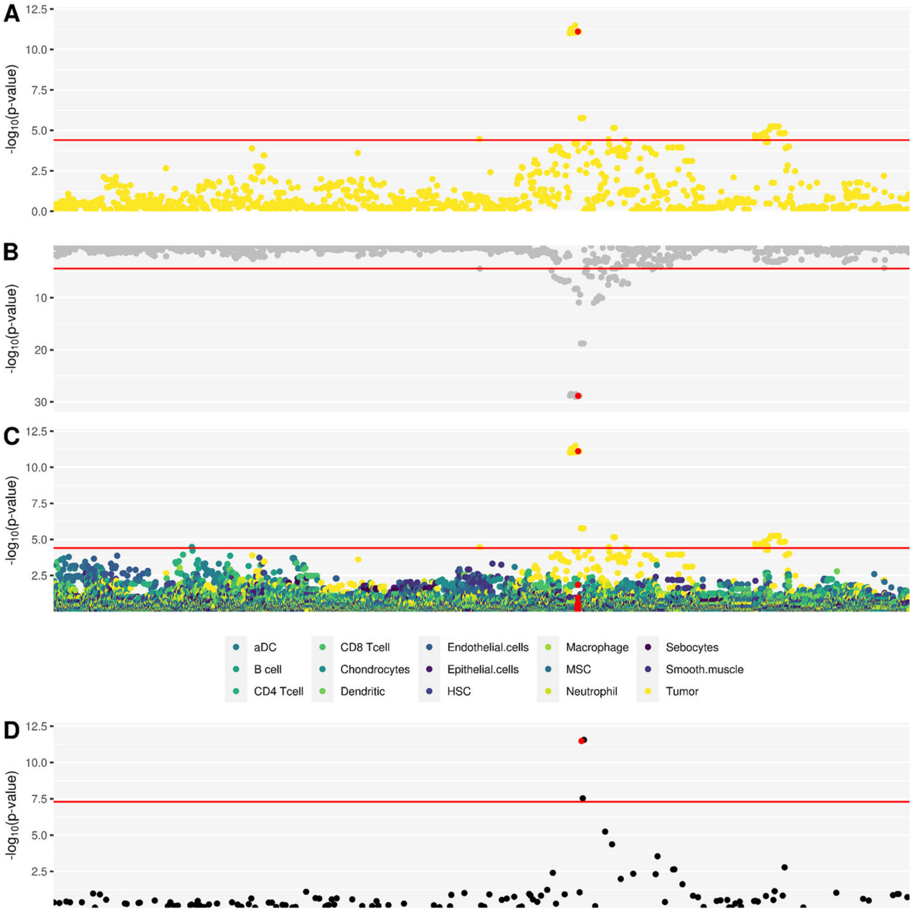

Fig. 5 Enrichment of tsQTLs in GWAS. SCARB1 manhattan plot **A** tsQTL, **B** marginal, **C** ,cfQTL and **D** RCC GWAS significance levels. Red highlighted SNP is rs4765621

tissue, and potentially involved in immune escape mechanisms [64]. The low expression of *ERAP2* has also been associated with poor cancer prognosis [65].

Intersecting the DeCAF results with kidney cancer GWAS data revealed a tsQTL at rs4765621 (FDR = 6.02^{-8}) that colocalized with a genome-wide significant risk variant (*p* value = 3.33^{-12}) for *SCARB1* (Fig. 5a). The same variant was also a marginal eQTL for *SCARB1* but is not an eQTL in the normal [47], strongly supporting its tumor-specific effect (Fig. 5). Previous studies have found that inhibiting *SCARB1* (a HDL cholesterol receptor) could kill and stop proliferation of clear cell renal cell carcinoma (ccRCC) cells [66]. A previous study [47] thoroughly validated the effect of rs4765623 on *SCARB1*, a SNP in perfect LD with the variant identified as a tsQTL in our study. rs4765623/*SCARB1* exhibited tumor-specific enhancer activity and was validated in a 786-O cell line, with the C risk allele enriched for chromatin activity, consistent with its expression increasing effect in our study.

Discussion

We presented DeCAF, a novel method for identifying cell-type-specific QTLs by harnessing signals from both total and allelic expression. No methods currently exist that leverage AI for cell-type-specific discovery, relying solely on QTL signals. By drawing additional power from individual reads, as well as the internal control offered by within-individual variation, DeCAF enables well-powered cfQTL analyses of medium-size QTL studies that would not have sufficient power for conventional tests. DeCAF directly models multiple potential cancer-specific confounders including tumor purity and CNVs, and can distinguish whether a QTL is functional in the tumor, microenvironment cell type, or both.

We applied DeCAF to TCGA data to perform the first cfQTL mapping effort in tumors. DeCAF identified 3664 significant cfQTL genes across all cell types, 5.63× more than was found using conventional ieQTL mapping (consistent with simulations). DeCAF similarly identified 3.72× more tsQTLs than the conventional approach, thus being a powerful method for identifying tumor-specific effects even in the absence of deconvoluted data. Surprisingly, DeCAF tsQTLs had the higher replication levels in independent eQTL data. This high replication, particularly in sorted immune cells, suggests tsQTLs may capture QTL effects in the context of tumor immune infiltration. Though we caution that the tsQTL analysis was substantially better powered than any other individual cfQTL analysis due to the higher variance and abundance of pure tumors, which likely impacts the relative replication performance. As little is known about the relevance of cell types in the tumor microenvironment to cancer risk, we integrated DeCAF data with cancer GWAS variants. We observed significant and cell-type-specific enrichment of cfQTLs with GWAS risk variants across all cancers evaluated; which was substantially higher than in non-cancer GWAS. By looking at the overlap of GWAS variants with cfQTLs and cancer QTLs, we were able to characterize their function and connect them to target genes in a specific cellular context. In sum, we found that DeCAF can identify thousands of cfQTLs from bulk RNA-seq, these cfQTLs replicated significantly in independent eQTL data (particularly tsQTLs), and were more enriched for GWAS risk than conventional eQTLs.

Other methods to detect cell-type-specific eQTL effects from bulk tissue data have recently been applied. GTEx [35] and Westra et al. [33] took the approach of identifying ieQTLs using a standard linear model with a cell fraction - genotype interaction term. The GTEx study evaluated concordance with allelic fractions in the same individuals as a replication, but did not utilize the full power of cell-type AI for discovery. As such, these approaches were generally underpowered for typical, moderately sized eQTL studies. For example, Westra et al. showed very low power below N=1000 (N<1000 represents the vast majority of QTL studies), and the GTEx analysis identified <80 ieQTL genes per cell type-tissue pair on average. Our approach also shares some similarities to recently proposed methods that leverage AI in other contexts. EAGLE [45] is a method to identify gene-environment interactions through AI mapping, which in principle could be extended to quantitative cell-fractions as "environments". However, EAGLE only identifies a gene level effect rather than specific QTLs, and is thus challenging to integrate with external data such as GWAS. The recent method BSCET [48] leveraged AI with deconvoluted cell types, but is similarly intended for analysis of transcribed SNPs rather than cfQTLs, and does not model the count nature of allelic data nor incorporate total expression. DeCAF is thus the first method to integrate total and allelic expression together for powerful cfQTL discovery at small to moderate sample sizes.

While DeCAF has made great strides in identifying cfQTLs, it has some limitations. First, DeCAF has limited power to detect cfQTLs in rare cell types (i.e., low cell fractions or cell fraction variance). Lower power for rare cell types is also a limitation of direct single-cell QTL analyses and motivates larger reference and target studies. Second, DeCAF is inherently dependent on the quality of the deconvolution and cannot test cell types that are not in the reference data. The majority of existing deconvolution methods (including TIMER and xCell) calculate cell fraction scores and not direct percentages. As a consequence, DeCAF cfQTL effect sizes should be interpreted with caution, because they will be influenced by (a) the scale of the deconvoluted score; (b) the differing uncertainty in the deconvolution of different cell types; (c) and the power to detect an effect in rare cell types. While most deconvolution methods have been shown to replicate well in pure bulk cell types [25–30], they continue to be in active development. DeCAF can be applied to any deconvolution framework or score and so will benefit from improved methodologies in the future. Using DeCAF to improve the cell type deconvolution itself (i.e., by maximizing cfQTL discovery) is thus a compelling future direction. Third, we elected to take the conservative approach and remove individuals with high CNV values from the analysis. In principle, DeCAF could be extended to model both cell-type- and CNV-specific interactions using matched allelic data from DNA sequencing, but this remains an open problem. Fourth, AI cannot be used for trans-QTL discovery, as the tested variant must be in phase with the allelic variants, and thus DeCAF's power advantages are limited to the study of cis-eQTLs. Fifth, any other clinical factor that's correlated with the cell fraction could be the causal mediator (for example - tumor stage, site of sampling and BMI), which could be further investigated by incorporating additional relevant interaction terms. Sixth, AI can be the consequence of many mechanisms such as nonsense-mediated decay, imprinting, post-translational effects and alternative splicing [67]. By assuming consistent AI/QTL effects in DeCAF, our approach may lose power when these deviating effects are present. However, prior work has found that, on average, AI and QTL effect sizes are highly concordant [60] and are thus likely capturing the same mechanisms in general. Our real data analyses also show that the combined DeCAF model achieves the largest power (with comparable external replication rate), further supporting these assumptions. Lastly, although DeCAF cfQTLs broadly replicated in external data at comparable rates to prior studies, we generally did not find that a given DeCAF cell type replicated best in the matching target cell type. This lack of cell type consistent replication may be explained by many factors including: differences in cell fractions influencing DeCAF power, differences in the target study sizes influencing replication power, and potential differences between AI signals in the discovery and eQTL signals in the replication. Understanding the relationship between cfQTLs detected in heterogeneous cell populations and cell-type-specific QTLs detected in pure cell types thus continues to be an open question of great interest.

Although DeCAF was applied to primarliy tumor RNA-seq data in this work, it is directly amenable to identifying cis-regulatory effects in other molecular phenotypes and contexts where sequencing read data is available. For example, DeCAF could be applied to identify cfQTLs in larger normal tissue datasets such as in GTEx. By developing a method in tumor tissue, we have addressed a more difficult scenario first. In the case of emerging single-cell QTL studies, DeCAF can be applied to identify cell-type-specific QTLs using the directly estimated cell fractions. Likewise, other measures of cell state, which may span multiple cell types, could be incorporated as interaction terms in the DeCAF framework. Beyond RNA-seq, cfQTLs can be investigated via DeCAF in other molecular measurements such as tumor ATAC-seq data from TCGA [68], which would shed light on cell-type- and tumor-specific mechanisms of DNA accessibility. DeCAF could be extended to estimate effects in multiple cell types jointly, through the inclusion of multiple interaction terms, although the potential for over-fitting and autocorrelation across cell fractions needs to be carefully considered. Beyond cell fraction and tumor purity, any continuous score can be used in DeCAF, including polygenic risk scores capturing germline risk burden [69], integrated clustering of molecular cancer subtypes, and emerging immune subtype definitions [70]. DeCAF thus provides a framework for broad analyses of context-specific QTLs in tumor and non-tumor data.

Conclusions

Rich availlability of bulk RNA-sequencing studies has lead to the development of cell-fraction deconvolution methods for large scale cell-type-specific expression analysis. Here, we present DeCAF as the first to use these analyses to study both cell-type-specific and tumor-specific allelic expression in cancer. By using a combination of AI and eQTL mapping, we gained considerable power over previous studies that considered eQTL mapping alone. Through identifying both microenvironment cell-type and tumor effects we were able to explore effects unique to each and link them back to disease. This high preforming method provides a framework to identify cell-type allelic effects in other cancer types, normal tissue, and a multitude of other continuous traits such as open chromatic from ATAC-seq.

Methods

Statistical model to detect cfQTLs

The conventional QTL-based approach defines y as a per-individual vector of total expression, f as the corresponding cell fraction estimate for a given cell type, and x as the genotype; and the models $y = \mu + \beta x + \beta_f x * f$. The cell fraction ieQTLeffect, β_f, is then estimated by typical linear regression [52–56]. To test cell-type-specific AI, we first restrict to individuals that are heterozygous for a variant in the target gene for which reads have been allelically assigned (we refer to this as the "functional SNP"). For this sub-population, f is again defined as the vector of cell fraction estimates, but instead of x, we introduce π, the population-level allelic fraction for heterozygous carriers of the functional SNP. Unlike conventional allelic imbalance tests that evaluate one individual at a time internally, we test for a consistent trend of cell fraction and imbalance across the population. We then model $\pi = \pi_f f + \pi_0(1 - f)$ where π_f is the allelic fraction in the focal cell type and π_0 is the allelic fraction in the rest of the cell types. While we do not observe the π value, for each individual we see REF, ALT read counts that are sampled from their π. The cell fraction iAI effect is then estimated by binomial regression of $REF, ALT \sim \mu_a + \beta_f f$ where μ_a captures the mean allelic fraction in the population and β_f captures the additional fraction-specific effect. When μ_a is significantly different from 0.5 this variant is exhibiting population-level AI, and when β_f is significantly different from 0, this variant is also exhibiting cell-type-specific AI. Functional (read carrying) SNPs were tested directly. For distal SNPs, the read counts for each allele were computed as the sum of functional SNP reads along the respective haplotype (as shown in previous work [50, 57]). To account for overdispersion that is common in molecular data[57], we leveraged a beta-binomial regression, with overdispersion estimated for each individual across all heterozygous reads (see below for tumors). Finally, we combined the cell fraction ieQTL and iAI tests by Stouffer's method (these tests are independent and so can be combined); we refer to the combined estimate as the DeCAF test statistic. Stouffer's method is the sum of the Z scores (here the test statistics derived from the QTL and AI tests), divided by the square of the number of values input. This combined statistic is equivalent to an inverse-variance weighted meta-analysis between the two associations and thus assumes a shared underlying effect (i.e., effects in opposite directions will become less significant when combined). Our assumption that the total and allelic signals are independent is the same as that made by prior approaches for combining QTL/AI data, including TReCASE [49], RASQUAL [50] and BaseQTL [51]. Even though expression is used from the same individuals, the tests are independent because the tested independent variable is independent: eQTLs use the variance between the 0/1/2 genotypes, whereas AI uses the variance within the 1 genotype (between alleles). We note that this is further supported by our simulations, which simulate consistent allelic effects in the same individuals and show that the test is well controlled under the null (Additional file 2: Fig. S2). In principle, AI and QTL effects could be aggregated with a statistic that ignores sign, such as a Fisher's combined test, but prior work has shown that AI and QTL effects are highly consistent [71].

Statistical model to account for tumor variation

The basic model described above allows for estimation of cfQTLs in normal expression, and we additionally extend this model to account for potential biases due tumor heterogeneity and somatic alterations. First, tumors are a mix of normal and cancer cells which introduces additional variance into the expression and could create a false relationship with a given cell type if it is correlated with the tumor fraction. To account for this, we introduce an additional term corresponding to tumor purity into the AI model, and an interaction term corresponding to the SNP-purity interaction into the eQTL model (as previously proposed in ref.[31]). This extension improves power for cfQTLs by accounting for tumor-specific variance. In addition, tumor-specific QTLs[31] can be inferred by testing for a non-zero effect size on the purity term.

Second, somatic copy number alterations can lead to extra variation and non-genetically driven AI. Indeed, the earliest applications of allelic imbalance in cancer were to identify copy number alterations (CNVs) from sequenced DNA. In the AI component of the DeCAF model, we estimate the beta-binomial overdispersion parameter for each CNV region in an individual separately, to account for CNV-specific variance. We additionally investigated three approaches to account for extra variance due to a significant CNV: (i) excluding variants in a CNV in an individual from the analysis entirely; (ii) including the per-individual CNV estimate in the model as a fixed effect covariate, to account for an offset in the expression due to carrying a CNV; (iii) including a random effect term for all CNV carriers to account for extra variance in expression without a consistent direction. These models make different assumptions about the CNV architecture and we selected the best performing model empirically based on the number of cfQTLs identified and their reproducibility in external data.

Simulation framework

We investigated multiple AI-based approaches and modeling assumptions in simulation. Parameters tested for impact on the performance of these tests included: minor allele frequency (MAF); baseline (0.5) and cell-type-specific effect size (AF); read depth (D, sampled from a Poisson with fixed mean); number of individuals (N); and cell fraction percentages (f, sampled from a uniform distribution or from real data). To generate a single AI individual, a π is defined as $0.5(1-f)+(\pi_f * f)$, ALT reads were drawn from $X \sim \text{BetaBin}(\pi, N)$ with fixed overdispersion parameter, and REF reads were computed as $D - ALT$. Hardy-Weinberg equations were then used to generate the expected number of heterozygous AI individuals based on the MAF. To simulate CNVs, a fraction of individuals were sampled as being carriers and additional variance terms are introduced. In the fixed-effect model, being a CNV carrier adds a constant factor to the π (i.e., one allele is always amplified), whereas in the random effect model, one allele is randomly amplified or deleted for computing the π. Finally, to simulate total expression for the conventional eQTL models, quantitative phenotypes were sampled from a linear model $y = \sigma_{NF} * NF(\sigma_{effect} * SNP * f) + \sigma_{CNV} * CNV + \varepsilon$. The overall simulation was performed 500 times and power for each test was defined as the number of simulations in which that test produces an association with p<0.05/20,000 (where 20,000 is approximately the number of genes to be tested).

Wherever possible, we identified simulation parameters from the real data: mean overdispersion was estimated as 0.0263 from tumor RNA-seq in TCGA; QTL effect sizes are set to 0.04, matching the average eQTL variance explained in real data [72]; normal fraction (NF) effect sizes were set to 0.0269, matching the average variance in expression explained by NF from real data; σ_{CNV} was set to 0.012 to match the variance in CNVs in real data. For cell fraction estimates, as a uniform distribution is an optimistic case (the majority of cell fractions are in the lower range and are variable (Additional file 2: Fig. S4), in addition to running simulations on generated uniform cell fractions we also ran simulations using the real cell fractions identified by TIMER [30].

Deconvoluted cell fractions in TCGA data

TCGA is a rich resource for tumor RNA-seq data which has been deconvoluted into cell types by multiple published methods: xCell [26] (64 cell types), which defines gene sets in a pure population and ranks the expression of these in the sample; TIMER [30] (six tumor-infiltrating immune cell types including B cells, CD4 T cells, CD8 T cells, neutrophils, macrophages, and dendritic cells), which uses a signature gene matrix from bulk expression; and others [73–75]. We downloaded xCell [26] and TIMER [30] cell fraction data for individuals with KIRC data from TCGA. To further improve power for testing in the xCell results, we focus on using cell types with a cell fraction score with an interquartile range >0.1 and cell types expected to be found in kidney tissue (Additional file 2: Fig. S4). Previous estimates of tumor purity [58] were additionally used to account for differences in purity and identify tsQTLs[31] (Fig. 1b).

Allele-specific quantification

We applied DeCAF to genotype and RNA-seq data from 503 TCGA RCC tumors from the KIRC study (Fig. 1b). Germline genotype data and RNA-seq BAMs were downloaded from the NCI Genomic Data Commons. Germline genotype data was imputed and phased through the Michigan Imputation Server. Total quantified and normalized gene expression (upper-quartile normalized count estimates) was accessed as previously described [76]. RNA-seq reads were integrated with the genotype data and run through the WASP allelic mapping pipeline (using BWA alignment) to account for potential mapping biases (Additional file 2: Fig. S5). In brief, for each read containing a polymorphism a new "proxy" read was generated with the alternative allele and mapped to the genome; any reads with proxies that did not map to the same location were then excluded from the analysis. This conservative approach removes reads with potential allele-specific mapping biases. Finally, allelic counts at individual variants were quantified using GATK [77] ASE quantification and assigned to the phased haplotypes. For each individual, allelic reads from all heterozygous variants within the tested gene are aggregated using haplotype information, under the assumption that the regulatory effect is consistent across the haplotype [47, 50, 57]. To process the TCGA data efficiently, we implemented DeCAF in a self-contained and reproducible pipeline on the Cancer Genomics Cloud, creating a workflow that can then be run on any data source (see Web Resources).

cfQTL discovery

For QTL discovery, for each expressed gene and cell type, we tested each common (MAF greater than 1%) germline SNP in the cis-locus of the gene (100kb around the TSS) using each described method. Each SNP is tested individually and separately for eQTL and AI. For the eQTL test, reference/alternate allele genotype dosage is used as the independent variable. For the AI test, heterozygous individuals are assigned the aggregated allelic read counts (see above) corresponding to their reference/alternate haplotypes. For both signals only a single distal variant is evaluated in a given test, but RNA-seq reads are aggregated across the gene. Each SNP then has eQTL and AI effects combined via Stouffer's method (see above) to define the final DeCAF statistic. For the marginal association without cell fraction interactions and tumor variation adjustment (see above), DeCAF estimates the same underlying quantity as the previously proposed WASP and RASQUAL tests[50, 57], with minor differences in how the likelihoods are maximized (regression versus grid search).

Significant cfQTLs in each cell type were determined by identifying the most significant variant for each cell type and gene, applying a genel-level Bonferroni correction. At this point a single SNP per gene and cell-type is then selected (the most significant after bonferroni correction). This set of SNPs then is further BH (Benjamini-Hochberg False Discovery Rate) corrected across genes, per cell-type. The same testing and multiple test correction procedure was applied to identify significant marginal eQTLs in the bulk data. While base results are reported at 10% FDR threshold, using a more stringent threshold did not greatly reduce the number of significant cfQTLs (Additional file 2: Fig. S8).

Testing for variant enrichment

We calculated multiple enrichment and replication estimators for cfQTLs. First, we estimated π_1 replication statistics (i.e., the fraction of associations that are expected to be "non-null") using the qvalue package [78] and (FDR<5%) for the significance threshold for the DeCAF QTL input. Secondly, we conduct a Fisher's exact test for enrichment as follows. A contingency table was constructed where positive-positive was defined as the number of SNPs with significance in both DeCAF (FDR<20%) and GWAS (p-value<0.001). Adjustment was based on the Bonferroni method. Finally, the average Z^2 was calculated as follows. We calculated the average Z^2 (Z score of the SNPs in the category of interest) for significant DeCAF cfQTLs overlapping the category of interest. We then calculated the average Z^2 of a randomly sampled (100×) null and took the ratio of the significant average Z^2 over the null as the measure of enrichment.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-022-02708-9.

Additional file 1. Supplementary Tables. DeCAF sigificant SNPs. Listed for each significant (FDR 10%) SNP is the SNP, gene, cell-type, Z score, *p*-value, and FDR corrected *p*-value.

Additional file 2. Supplementary Figures.

Additional file 3. Review history.

Peer review information

Stephanie McClelland was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

Review history is available as Additional file 3.

Authors' contributions

CK and AG envisaged the study. CK performed the computational analysis for the study, supported and advised by AG. CK and AG wrote the paper, as well as read and approved the manuscript.

Funding

A.G. and C.K. were supported by R01CA227237, R01CA244569, R01MH115676, R01HG006399, R01HG012133, Emerson Collective, and the Doris Duke Charitable Foundation. C.K. was supported by the Louis B. Mayer Foundation and the Claudia Adams Barr Foundation.

The Seven Bridges Cancer Research Data Commons Cloud Resource has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Contract No. HHSN261201400008C and ID/IQ Agreement No. 17X146 under Contract No. HHSN261201500003I and 75N91019D00024.

Availability of data and materials

Software implementing STRATAS DeCAF can be found at zenodo and https://github.com/cakalita/stratAS/tree/DeCAF [79, 80]. Workflow generating the allelic counts can be found on the Cancer Genomics Cloud (CGC) [81] https://cgc.sbgenomics.com/u/cakalita/quickstart/apps/#cakalita/quickstart/stratas_v3.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 11 November 2021 Accepted: 15 June 2022
Published online: 08 July 2022

References

1. Scelo G, Purdue MP, Brown KM, Johansson M, Wang Z, Eckel-Passow JE, et al. Genome-wide association study identifies multiple risk loci for renal cell carcinoma. Nat Commun. 2017;8(1):15724. https://doi.org/10.1038/ncomms15724.
2. Melin BS, Barnholtz-Sloan JS, Wrensch MR, Johansen C, Il'yasova D, Kinnersley B, et al. Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumors. Nat Genet. 2017;49(5):789–94. https://doi.org/10.1038/ng.3823.
3. Huyghe JR, Bien SA, Harrison TA, Kang HM, Chen S, Schmit SL, et al. Discovery of common and rare genetic risk variants for colorectal cancer. Nat Genet. 2019;51(1):76–87. https://doi.org/10.1038/s41588-018-0286-6.
4. McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, Christiani DC, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nat Genet. 2017;49(7):1126–32. https://doi.org/10.1038/ng.3892.
5. Schumacher FR, Al Olama AA, Berndt SI, Benlloch S, Ahmed M, Saunders EJ, et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. Nat Genet. 2018;50(7):928–36. https://doi.org/10.1038/s41588-018-0142-8.
6. Phelan CM, Kuchenbaecker KB, Tyrer JP, Kar SP, Lawrenson K, Winham SJ, et al. Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. Nat Genet. 2017;49(5):680–91. https://doi.org/10.1038/ng.3826.
7. Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. Nature. 2017;551(7678):92–4. https://doi.org/10.1038/nature24284.
8. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat Rev Genet. 2015;16(4):197–212. https://doi.org/10.1038/nrg3891.
9. Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. Nature. 2017;550(7675):204–13. https://doi.org/10.1038/nature24277.
10. Gibbs J, van der Brug M, Hernandez D. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet. 2010;6(5):1–13. https://doi.org/10.1371/journal.pgen.1000952.
11. Melzer D, Perry JRB, Hernandez D, Corsi AM, Stevens K, Rafferty I, et al. A genome-wide association study identifies protein quantitative trait loci (pQTLs). PLoS Genet. 2008;4(5):1–10. https://doi.org/10.1371/journal.pgen.1000072.
12. Gusev A, Lawrenson K, Lin X, Lyra PC, Kar S, Vavra KC, Segato F, Fonseca MAS, Lee JM, Pejovic T, Liu G, Karlan BY, Freedman ML, Noushmehr H, Monteiro AN, Pharoah PDP, Pasaniuc B, Gayther SA. A transcriptome-wide association study of high-grade serous epithelial ovarian cancer identifies new susceptibility genes and splice variants. Nat Genet. 2019;51(5):815–23. https://doi.org/10.1038/s41588-019-0395-x.
13. Mancuso N, Gayther S, Gusev A, Zheng W, Penney KL, Freedman M, Haiman C, Pasaniuc B. Large-scale transcriptome-wide association study identifies new prostate cancer risk regions. Nat Commun. 2018;9(1):4079. https://doi.org/10.1038/s41467-018-06302-1.
14. Wu L, Shi W, Long J, Guo X, Michailidou K, Beesley J, et al. A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. Nat Genet. 2018;50(7):968–78. https://doi.org/10.1038/s41588-018-0132-x.
15. Ardlie KG, DeLuca DS, Segrè AV, Sullivan TJ, Young TR, Gelfand ET, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 2015;348(6235):648–60. https://doi.org/10.1126/science.1262110.
16. Gamazon ER, Segrè AV, van de Bunt M, Wen X, Xi HS, Hormozdiari F, Ongen H, Konkashbaev A, Derks EM, Aguet F, Quan J, GTEx Consortium, Nicolae DL, Eskin E, Kellis M, Getz G, McCarthy MI, Dermitzakis ET, Cox NJ, Ardlie KG. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation,. Nat Genet. 2018;50(7):956–67. https://doi.org/10.1038/s41588-018-0154-4.
17. Zhang T, Choi J, Kovacs MA, Shi J, Xu M, Consortium, Melanoma Meta Analysis, et al. Cell-type-specific eQTL of primary melanocytes facilitates identification of melanoma susceptibility genes. Genome Res. 2018;28(11):1621–35. https://doi.org/10.1101/gr.233304.117.
18. Schmiedel BJ, Singh D, Madrigal A, Valdovino-Gonzalez AG, White BM, Zapardiel-Gonzalo J, Ha B, Altay G, Greenbaum JA, McVicker G, Seumois G, Rao A, Kronenberg M, Peters B, Vijayanand P. Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression,. Cell. 2018;175(6):1701–171516. https://doi.org/10.1016/j.cell.2018.10.022.
19. Chen DS, Mellman I. Elements of cancer immunity and the cancer-immune set point. Nature. 2017;541(7637):321–30. https://doi.org/10.1038/nature21349.
20. Finotello F, Trajanoski Z. Quantifying tumor-infiltrating immune cells from transcriptomics data. Cancer Immunol Immunother. 2018;67(7):1031–40. https://doi.org/10.1007/s00262-018-2150-z.
21. Chen D, Mellman I. Oncology Meets Immunology: The Cancer-Immunity Cycle. Immunity. 2013;39(1):1–10. https://doi.org/10.1016/J.IMMUNI.2013.07.012.

22. Savage PA, Malchow S, Leventhal DS. Basic principles of tumor-associated regulatory T cell biology. Trends Immunol. 2013;34(1):33–40. https://doi.org/10.1016/J.IT.2012.08.005.
23. Finotello F, Trajanoski Z. New strategies for cancer immunotherapy: targeting regulatory T cells. Genome Med. 2017;9(1):1–3. https://doi.org/10.1186/S13073-017-0402-8.
24. 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;167(5):1398–141424. https://doi.org/10.1016/j.cell.2016.10.026.
25. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, Selves J, Laurent-Puig P, Sautès-Fridman C, Fridman WH, de Reyniès A. Erratum to Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. 2016;17(1):218. https://doi.org/10.1186/s13059-016-1113-y.
26. Aran D, Hu Z, Butte AJ. xCell: Digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. 2017;18(1):220. https://doi.org/10.1186/s13059-017-1349-1.
27. Gentles AJ, Newman AM, Liu CL, Bratman SV, Feng W, Kim D, Nair VS, Xu Y, Khuong A, Hoang CD, Diehn M, West RB, Plevritis SK, Alizadeh AA. The prognostic landscape of genes and infiltrating immune cells across human cancers. Nat Med. 2015;21(8):938–45. https://doi.org/10.1038/nm.3909.
28. Finotello F, Mayer C, Plattner C, Laschober G, Rieder D, Hackl H, Krogsdam A, Loncova Z, Posch W, Wilflingseder D, Sopper S, Ijsselsteijn M, Brouwer TP, Johnson D, Xu Y, Wang Y, Sanders ME, Estrada MV, Ericsson-Gonzalez P, Charoentong P, Balko J, De Miranda NFDCC, Trajanoski Z. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. Genome Med. 2019;11(1):34. https://doi.org/10.1186/s13073-019-0638-6.
29. Racle J, de Jonge K, Baumgaertner P, Speiser DE, Gfeller D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. eLife. 2017;6(pii):26476. https://doi.org/10.7554/eLife.26476.
30. Li B, Severson E, Pignon JC, Zhao H, Li T, Novak J, Jiang P, Shen H, Aster JC, Rodig S, Signoretti S, Liu JS, Liu XS. Comprehensive analyses of tumor immunity: Implications for cancer immunotherapy. Genome Biol. 2016;17(1):174. https://doi.org/10.1186/s13059-016-1028-7.
31. Geeleher P, Nath A, Wang F, Zhang Z, Barbeira AN, Fessler J, Grossman RL, Seoighe C, Stephanie Huang R. Cancer expression quantitative trait loci (eQTLs) can be determined from heterogeneous tumor gene expression data by modeling variation in tumor purity. Genome Biol. 2018;19(1):130. https://doi.org/10.1186/s13059-018-1507-0.
32. Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP, et al. Comprehensive functional genomic resource and integrative model for the human brain. Science. 2018;362(6420):8464. https://doi.org/10.1126/science.aat8464.
33. Westra HJ, Arends D, Esko T, Peters MJ, Schurmann C, Schramm K, et al. Cell Specific eQTL Analysis without Sorting Cells. PLoS Genet. 2015;11(5):1005223. https://doi.org/10.1371/journal.pgen.1005223.
34. Zhernakova DV, Deelen P, Vermaat M, Van Iterson M, Van Galen M, Arindrarto W, et al. Identification of context-dependent expression quantitative trait loci in whole blood. Nat Genet. 2017;49(1):139–45. https://doi.org/10.1038/ng.3737.
35. Kim-Hellmuth S, Aguet F, Oliva M, Muñoz-Aguirre M, Kasela S, Wucher V, et al. Cell type–specific genetic regulation of gene expression across human tissues. Science. 2020;369(6509). https://doi.org/10.1126/SCIENCE.AAZ8528.
36. Cowper-Sal lari R, Zhang X, Wright JB, Bailey SD, Cole MD, Eeckhoute J, Moore JH, Lupien M. Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. Nat Genet. 2012;44(11):1191–8. https://doi.org/10.1038/ng.2416.
37. Hasin-Brumshtein Y, Hormozdiari F, Martin L, van Nas A, Eskin E, Lusis AJ, Drake TA. Allele-specific expression and eQTL analysis in mouse adipose tissue. BMC Genomics. 2014;15(1):471. https://doi.org/10.1186/1471-2164-15-471.
38. Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Habegger L, Rozowsky J, Shi M, Urban AE, Hong M-Y, Karczewski KJ, Huber W, Weissman SM, Gerstein MB, Korbel JO, Snyder M. Variation in transcription factor binding among humans. Science. 2010;328(5975):232–5. https://doi.org/10.1126/science.1183621.
39. Kukurba KR, Zhang R, Li X, Smith KS, Knowles DA, How Tan M, Piskol R, Lek M, Snyder M, MacArthur DG, Li JB, Montgomery SB. Allelic Expression of Deleterious Protein-Coding Variants across Human Tissues. PLoS Genet. 2014;10(5):1004304. https://doi.org/10.1371/journal.pgen.1004304.
40. Battenhouse A, Keefe D, Collins FS, Willard HF, Lieb JD, Furey TS, Crawford GE, Iyer VR, Birney E, McDaniell R, Lee BK, Song L, Liu Z, Boyle AP, Erdos MR, Scott LJ, Morken MA, Kucera KS. Heritable Individual-Specific and Allele-Specific Chromatin Signatures in Humans. Science. 2010;328(5975):235–9. https://doi.org/10.1126/science.1184655.
41. McVicker G, Van De Geijn B, Degner JF, Cain CE, Banovich NE, Raj A, Lewellen N, Myrthil M, Gilad Y, Pritchard JK. Identification of genetic variants that affect histone modifications in human cells. Science. 2013;342(6159):747–9. https://doi.org/10.1126/science.1242429.
42. Pastinen T. Genome-wide allele-specific analysis: Insights into regulatory variation. Nat Rev Genet. 2010;11(8):533–8. https://doi.org/10.1038/nrg2815.
43. Timothy E, Reddy TE, Gertz J, Pauli F, Kucera KS, Varley KE, Newberry KM, Marinov GK, Mortazavi A, Williams BA, Song L, Crawford GE, Wold B, Willard HF, Myers RM. The effects of genome sequence on differential allelic transcription factor occupancy and gene expression. Genome Res. 2012;22:860–9. https://doi.org/10.1101/gr.131201.111.22.
44. Skelly DA, Johansson M, Madeoy J, Wakefield J, Akey JM. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. Genome Res. 2011;21(10):1728–37. https://doi.org/10.1101/gr.119784.110.
45. Knowles DA, Davis JR, Edgington H, Raj A, Favé MJ, Zhu X, Potash JB, Weissman MM, Shi J, Levinson DF, Awadalla P, Mostafavi S, Montgomery SB, Battle A. Allele-specific expression reveals interactions between genetic variation and environment. Nat Methods. 2017;14(7):699–702. https://doi.org/10.1038/nmeth.4298.
46. Moyerbrailean GA, Richards AL, Kurtz D, Kalita CA, Davis GO, Harvey CT, Alazizi A, Watza D, Sorokin Y, Hauff N, Zhou X, Wen X, Pique-Regi R, Luca F. High-throughput allele-specific expression across 250 environmental conditions. Genome Res. 2016;26(12):1627–38. https://doi.org/10.1101/gr.209759.116.
47. Gusev A, Spisak S, Fay AP, Carol H, Vavra KC, Signoretti S, Tisza V, Pomerantz M, Abbasi F, Seo J-H, Choueiri TK, Lawrenson K, Freedman ML. Allelic imbalance reveals widespread germline-somatic regulatory differences and prioritizes risk loci in Renal Cell Carcinoma. bioRxiv. 2019;631150. https://doi.org/10.1101/631150.

48. Fan J, Wang X, Xiao R, Li M. Detecting cell-type-specific allelic expression imbalance by integrative analysis of bulk and single-cell RNA sequencing data. PLOS Genet. 2021;17(3):1009080. https://doi.org/10.1371/JOURNAL.PGEN.1009080.
49. Sun W. A Statistical Framework for eQTL Mapping Using RNA-seq Data. Biometrics. 2012;68(1). https://doi.org/10.1111/j.1541-0420.2011.01654.x.
50. Kumasaka N, Knights AJ, Gaffney DJ. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nat Genet. 2016;48(2):206–13. https://doi.org/10.1038/ng.3467.
51. Vigorito E, Lin W-Y, Starr C, Kirk PDW, White SR, Wallace C. Detection of quantitative trait loci from RNA-seq data with or without genotypes using BaseQTL. Nat Comput Sci. 2021;1(6). https://doi.org/10.1038/s43588-021-00087-y.
52. Nédélec Y, Sanz J, Baharian G, Szpiech ZA, Pacis A, Dumaine A, Grenier J-C, Freiman A, Sams AJ, Hebert S, Pagé Sabourin A, Luca F, Blekhman R, Hernandez RD, Pique-Regi R, Tung J, Yotova V, Barreiro LB. Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens,. Cell. 2016;167(3):657–66921. https://doi.org/10.1016/j.cell.2016.09.025.
53. Mangravite LM, Engelhardt BE, Medina MW, Smith JD, Brown CD, Chasman DI, Mecham BH, Howie B, Shim H, Naidoo D, Feng Q, Rieder MJ, Chen Y-DI, Rotter JI, Ridker PM, Hopewell JC, Parish S, Armitage J, Collins R, Wilke RA, Nickerson DA, Stephens M, Krauss RM. A statin-dependent QTL for GATM expression is associated with statin-induced myopathy,. Nature. 2013;502(7471):377–80. https://doi.org/10.1038/nature12508.
54. Maranville J, Luca F, Richards A. Interactions between glucocorticoid treatment and cis-regulatory polymorphisms contribute to cellular response phenotypes. PLoS Genet. 2011;7(7):1–13. https://doi.org/10.1371/journal.pgen.1002162.
55. Maranville JC, Baxter SS, Witonsky DB, Chase MA, Di Rienzo A. Genetic mapping with multiple levels of phenotypic information reveals determinants of lymphocyte glucocorticoid sensitivity. Am J Hum Genet. 2013;93(4):735–43. https://doi.org/10.1016/j.ajhg.2013.08.005.
56. Strober BJ, Elorbany R, Rhodes K, Krishnan N, Tayeb K, Battle A, Gilad Y. Dynamic genetic regulation of gene expression during cellular differentiation. Science. 2019;364(6447):1287–90. https://doi.org/10.1126/science.aaw0040.
57. Van De Geijn B, Mcvicker G, Gilad Y, Pritchard JK. WASP: Allele-specific software for robust molecular quantitative trait locus discovery. Nat Methods. 2015;12(11):1061–3. https://doi.org/10.1038/nmeth.3582.
58. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nat Commun. 2015;6(1):1–12. https://doi.org/10.1038/ncomms9971.
59. Aguet F, Barbeira AN, Bonazzola R, Brown A, Castel SE, Jo B, et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369(6509):1318–30. https://doi.org/10.1126/SCIENCE.AAZ1776.
60. Castel SE, Aguet F, Mohammadi P, Aguet F, Anand S, Ardlie KG, et al. A vast resource of allelic expression data spanning human tissues. Genome Biol. 2020;21(1):1–12. https://doi.org/10.1186/S13059-020-02122-Z/FIGURES/2.
61. Loh PR, Kichaev G, Gazal S, Schoech AP, Price AL. Mixed-model association for biobank-scale datasets. Nat Genet. 2018. https://doi.org/10.1038/s41588-018-0144-6.
62. Gazal S, Finucane HK, Furlotte NA, Loh P-R, Palamara PF, Liu X, Schoech A, Bulik-Sullivan B, Neale BM, Gusev A, Price AL. Linkage disequilibrium dependent architecture of human complex traits shows action of negative selection. Nat Genet. 2017;49(10):1421. https://doi.org/10.1038/NG.3954.
63. Fruci D, Giacomini P, Nicotra MR, Forloni M, Fraioli R, Saveanu L, van Endert P, Natali PG. Altered expression of endoplasmic reticulum aminopeptidases ERAP1 and ERAP2 in transformed non-lymphoid human tissues. J Cell Physiol. 2008;216(3):742–9. https://doi.org/10.1002/JCP.21454.
64. Stoehr CG, Buettner-Herold M, Kamphausen E, Bertz S, Hartmann A, Seliger B. Comparative expression profiling for human endoplasmic reticulum-resident aminopeptidases 1 and 2 in normal kidney versus distinct renal cell carcinoma subtypes. Int J Clin Exp Pathol. 2013;6(6):998.
65. Compagnone M, Cifaldi L, Fruci D. Regulation of ERAP1 and ERAP2 genes and their disfunction in human cancer. Hum Immunol. 2019;80(5):318–24. https://doi.org/10.1016/J.HUMIMM.2019.02.014.
66. Riscal R, Bull CJ, Mesaros C, Finan JM, Carens M, Ho ES, Xu JP, Godfrey J, Brennan P, Johansson M, Purdue MP, Chanock SJ, Mariosa D, Timpson NJ, Vincent EE, Keith B, Blair IA, Skuli N, Simon MC. Cholesterol auxotrophy as a targetable vulnerability in clear cell renal cell carcinoma. Cancer Discov. 2021;0211–2021. https://doi.org/10.1158/2159-8290.CD-21-0211.
67. Cleary S, Seoighe C. Perspectives on Allele-Specific Expression. Ann Rev Biomed Data Sci. 2021;4(1). https://doi.org/10.1146/annurev-biodatasci-021621-122219.
68. Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, Silva TC, Groeneveld C, Wong CK, Cho SW, Satpathy AT, Mumbach MR, Hoadley KA, Robertson AG, Sheffield NC, Felau I, Castro MAA, Berman BP, Staudt LM, Zenklusen JC, Laird PW, Curtis C, Greenleaf WJ, Chang HY. The chromatin accessibility landscape of primary human cancers. Science. 2018;362(6413):1898. https://doi.org/10.1126/science.aav1898.
69. Shi M, O'Brien KM, Weinberg CR. Interactions between a Polygenic Risk Score and Non-genetic Risk Factors in Young-Onset Breast Cancer. Sci Rep. 2020;10(1). https://doi.org/10.1038/s41598-020-60032-3.
70. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, et al. The Immune Landscape of Cancer. Immunity. 2018;48(4):812–83014. https://doi.org/10.1016/j.immuni.2018.03.023.
71. Mohammadi P, Castel SE, Brown AA, Lappalainen T. Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. Genome Res. 2017;27(11):1872–84. https://doi.org/10.1101/gr.216747.116.
72. Ng B, Casazza W, Patrick E, Tasaki S, Novakovsky G, Felsky D, Ma Y, Bennett DA, Gaiteri C, De Jager PL, Mostafavi S. Using Transcriptomic Hidden Variables to Infer Context-Specific Genotype Effects in the Brain. Am J Hum Genet. 2019;105(3):562–72. https://doi.org/10.1016/j.ajhg.2019.07.016.
73. Wang Z, Cao S, Morris JS, Ahn J, Liu R, Tyekucheva S, Gao F, Li B, Lu W, Tang X, Wistuba II, Bowden M, Mucci L, Loda M, Parmigiani G, Holmes CC, Wang W. Transcriptome Deconvolution of Heterogeneous Tumor Samples with Immune Infiltration. Food Sci Hum Wellness. 2018;9(1):451–60. https://doi.org/10.1016/j.isci.2018.10.028.
74. Roman T, Xie L, Schwartz R. Automated deconvolution of structured mixtures from heterogeneous tumor genomic data. PLoS Comput Biol. 2017;13(10):1005815. https://doi.org/10.1371/journal.pcbi.1005815.

75. Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, Hackl H, Trajanoski Z. Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. Cell Rep. 2017;18(1):248–62. https://doi.org/10.1016/j.celrep.2016.12.019.
76. Broad Institute TCGA Genome Data Analysis Center. Broad Institute of MIT and Harvard. 2016. https://doi.org/10.7908/C11G0KM9.
77. Van der Auwera GA, O'Connor BD. Genomics in the Cloud: O'Reilly Media, Inc.; 2020. https://www.oreilly.com/library/view/genomics-in-the/9781491975183/. Accessed 13 Oct 2021.
78. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 2003;100(16):9440–5. https://doi.org/10.1073/pnas.1530509100.
79. Kalita CA. DeCAF Code. 2022. https://doi.org/10.5281/zenodo.6633672.
80. Kalita CA. DeCAF Code. 2022. https://github.com/cakalita/stratAS/tree/DeCAF. Accessed June 2022.
81. Lau JW, Lehnert E, Sethi A, Malhotra R, Kaushik G, Onder Z, Groves-Kirkby N, Mihajlovic A, DiGiovanna J, Srdic M, Bajcic D, Radenkovic J, Mladenovic V, Krstanovic D, Arsenijevic V, Klisic D, Mitrovic M, Bogicevic I, Kural D, Davis-Dusenbery B. The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized—A New Paradigm in Large-Scale Computational Research. Cancer Res. 2017;77(21):3–6. https://doi.org/10.1158/0008-5472.CAN-17-0387.

Publisher's Note

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