Summer Presentation(3)

DeCAF Method

Chenxiao Tian

Date：2022/7/27

OUTLINE

- 1. Background and Motivation
- 2. Methods

  (I)The Case and Model for Normal Tissue

  (II) The Case and Model for Tumor Variation

- 3. Simulations Methods and Real Data Results
- 4. Discussion and Conclusions

Background and Motivation

Background:

- GWAS have been instrumental in identifying a large number of genetic variants associated with risk for many diseases including cancer .
- 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 :

Example： Recently identifying putative susceptibility genes for ovarian cancer, prostate cancer, and breast cancer.

Background and Motivation

- Recently, it has been observed that some eQTL effects can be observed only in specific contexts and often vary across tissue and cell types.

- Example： In the context of cancer, the cell types in the tumor/microenvironment can have substantially different functions：

  (1) With CD4 and CD8 T cells driving cytotoxic anticancer immunity.

  (2)While regulatory T cells are associated with immune suppression and homeostasis.

- Goal: Identifying and quantifying cell-type-specific eQTLs in tumors is thus a critical step to understanding germline cancer mechanisms and germline-somatic interactions.

Background and Motivation

- Problem:
- 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 .
- Example Current Methods for the Problem:
- (1) Emerging single-cell technologies:
- It has the potential to precisely measure expression in specific cell populations,
- but this approach remains too expensive to measure across hundreds of individuals and exhibits very sparse expression.

Background and Motivation

- (2) Bulk RNA-seq deconvolution methods:
- These cell fraction estimates can additionally be incorporated into eQTL analyses to identify cell-fraction specific effects (cfQTLs) .
- 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/fraction) .

- (3) Allelic Imbalance (AI):

A. Genetic effects on expression can be measured by quantifying the ratio of RNA-seq reads at heterozygous variants in exons.

Background and Motivation

- B. A significant departure from a 50% allelic ratio is indicative of a cis-genetic effect on expression, and referred to as allelic imbalance (AI).
- C. Benefits for AI

(I) This approach benefits from being able to control for trans-variation.

(II) Because it measures allelic effects within individuals and not between individuals, can harness power from read depth.

(III) AI has also been used to identify genes undergoing gene-environment interactions, tumor/normal regulatory differences and, recently, cell type specificity.

(IV) Methods exist and have been proven to work well, which jointly model AI and eQTL effect sizes.

D. Problems：

However, the integration of total expression and AI to detect cfQTLs has been largely unexplored.

Background and Motivation

- Goal for this Paper：

- 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.

Methods

Statistical model to detect cfQTLs in Normal Expression

A. Coventional Model

- 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:

$$y = \mu + \beta x + \beta_f x * f.$$

- Then the cell fraction ieQTL effect, β_f, is then estimated by typical linear regression.

Methods

B.The Model for cell-type-specific AI,

Restricted Sub-Population:

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”).

The Population-level Allelic Fraction π :

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.

Methods

C.Different Consistent Trend Test:

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.

D.The Model of The Population-level Allelic Fraction π :

$$\pi = \pi_f f + \pi_0(1 - f)$$

Where π_f is the allelic fraction in the focal cell type and π0 is the allelic fraction in the rest of the cell types.

Methods

- E.Transfer π to REE, ALT:
- While we do not observe the π value, for each individual we see REF,ALT read counts that are sampled from their π.

- F.Estimation of the cell fraction iAI effect:
- The cell fraction iAI effect is then estimated by binomial regression of REF,ALT $\sim \mu_\alpha + \beta_f\ f$:

(I) Where μ_α captures the mean allelic fraction in the population.

(II) β_f captures the additional fraction-specific effect.

(III) When β_f is significantly different from 0, this variant is also exhibiting cell-type-specific AI.

Methods

- G.The Case for distal SNPs:

For distal SNPs, the read counts for each allele were computed as the sum of functional SNP reads along the respective haplotype.

H.The Overdispersion Problem:

To account for overdispersion that is common in molecular data, we leveraged a beta-binomial regression, with overdispersion estimated for each individual across all heterozygous reads.

Methods

- I.The Application of Stouffer’s method for Combining Purpose

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.

- J.Note on the Stouffer’s Method

- (I)Stouffer’s method is the sum of the Z scores (in this paper, the test statistics derived from the QTL and AI tests), divided by the square of the number of values input.

(II)This combined statistic is equivalent to an inverse-variance weighted meta-analysis between the two associations and thus assumes a shared underlying effect that is, the effects in opposite directions will become less significant when combined.

Methods

- K.About the Independent Assumptions:

Independent Assumptions 1: 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, RASQUAL and BaseQTL.

Independent Assumptions 2: Even though expression is used from the same individuals, the tests are independent because the tested independent variable is independent:

i.e. eQTLs use the variance between the 0/1/2 genotypes, whereas AI uses the variance within the 1 genotype (between alleles)

Methods

- 2. The general case and statistical model for tumor variation
- A.Different Biology Background:

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.

B1.Additional Variance and False Relationship Problem:

- 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.

Methods

- Deal: Two Extra Terms into the AI Model:

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:

(I) This extension improves power for cfQTLs by accounting for tumor-specific variance.

(II) In addition, tumor-specific QTLs can be inferred by testing for a non-zero effect size on the purity term.

Methods

- B2.Problem: Extra Variation and Non-genetically Driven AI.

Second, somatic copy number alterations can lead to extra variation and non-genetically driven AI.

Remark: The earliest applications of allelic imbalance in cancer were to identify copy number alterations (CNVs) from sequenced DNA.

Deal with the CNV-specific variance:

To account for CNV-specific variance, in the AI component of the DeCAF model, we estimate the beta-binomial overdispersion parameter for each CNV region in an individual separately.

Methods

- C.We additionally investigated three approaches and models 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.

NOTE: About the Assumptions:

These models make different assumptions about the CNV architecture.

we selected the best performing model empirically based on the number of cfQTLs identified and their reproducibility in external data

Simulations and Results

- A.Date Source Using: 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, TIMER.

- We downloaded xCell and TIMER 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

Simulations and Results

- B.Simulations Framework:
- We investigated multiple AI-based approaches and modeling assumptions in simulation.

Specifically speaking ,parameters tested for impact on the performance of these tests included:

1.minor allele frequency (MAF);

2.baseline (0.5) and cell-type-specific effect size (AF);

3.read depth (D, sampled from a Poisson with fixed mean);

4.number of individuals (N);

5.and cell fraction percentages (f, sampled from a uniform distribution or from real data).

Simulations and Results

- C.Generate a single AI individual:
- To generate a single AI individual, a π is defined as:

$$0.5(1 - f) + (\pi_f * f),$$

- D.The Estimation of the Parameters ALT,REF Expected Number of AI:

(i)ALT reads were drawn from X ~ BetaBin(π, N) with fixed overdispersion parameter,

(ii)REF reads were computed as D − ALT.

(iii) Hardy-Weinberg equations were then used to generate the expected number of heterozygous AI individuals based on the MAF.

Simulations and Results

- E.Simulate CNVs for Different Models:
- To simulate CNVs, a fraction of individuals were sampled as being carriers and additional variance terms are introduced:

(i)In the fixed-effect model, being a CNV carrier adds a constant factor to the π

(ii)In the random effect model, one allele is randomly amplified or deleted for computing the π.

Simulations and Results

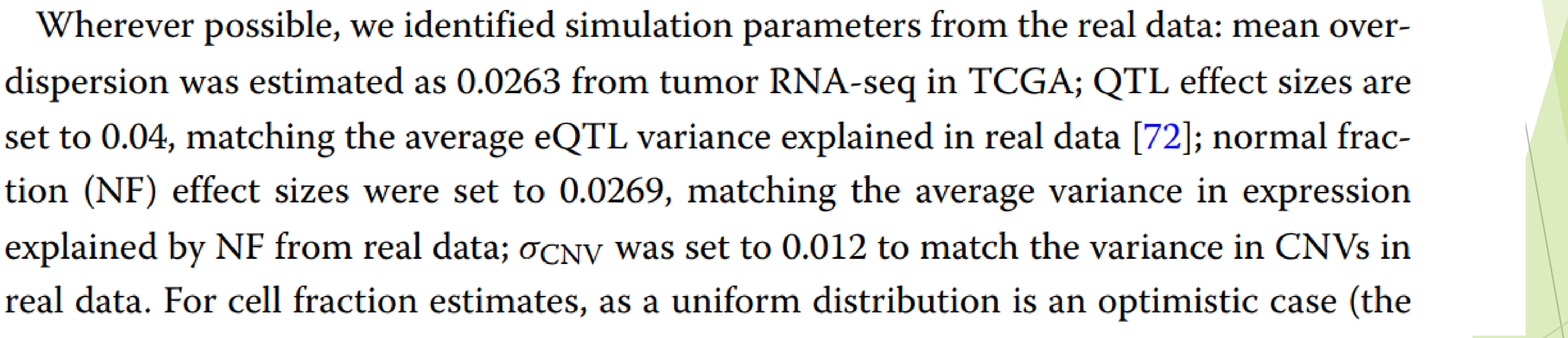
- F.Simulate total expression for the conventional eQTL models:
- To simulate total expression for the conventional eQTL models, quantitative phenotypes were sampled from a linear model：

$$y = \sigma_{NF} * NF(\sigma_{effect} * SNP * f) + \sigma_{CNV} * CNV + \varepsilon.$$

- 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< $0.05/20000$

Simulations and Results

- G.Identified Simulations Parameters from the real data:



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

Simulations and Results

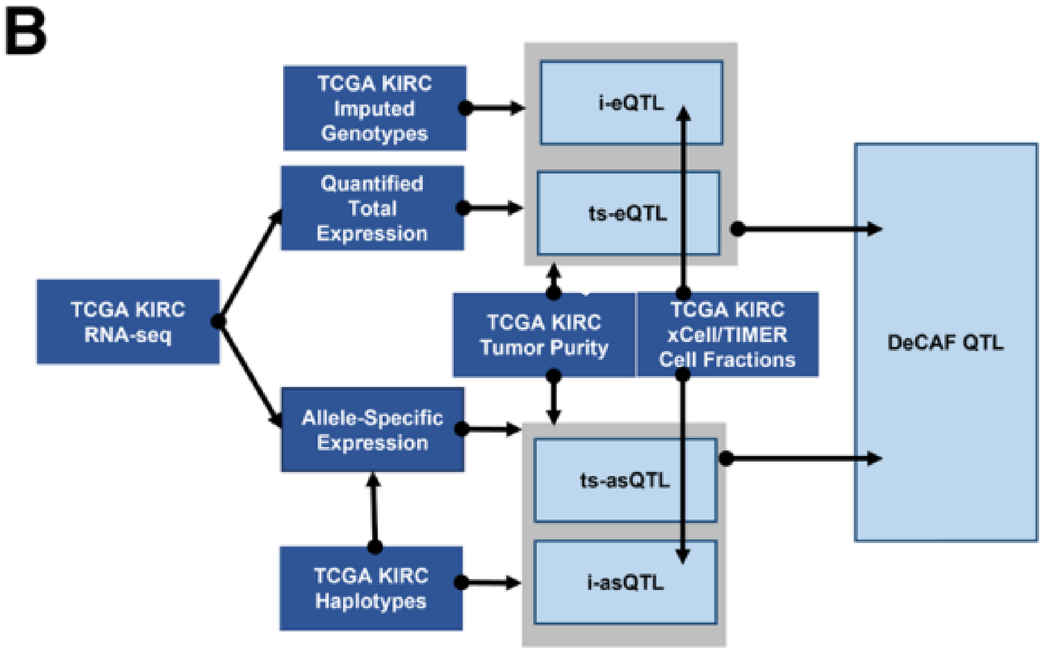
- H.Cell fraction estimates:

(i)Running simulations on generated uniform cell fractions as an optimistic case

(ii)We also ran simulations using the real cell fractions identified by TIMER

Simulations and Results

- I. Real Data and Application
- (1) We applied DeCAF to genotype and RNA-seq data from 503 TCGA RCC tumors from the KIRC study (Fig. 1b), like the following logic procedure graph:



B
TCGA KIRC Imputed Genotypes
Quantified Total Expression
TCGA KIRC RNA-seq
Allele-Specific Expression
TCGA KIRC Haplotypes
i-eQTL
ts-eQTL
TCGA KIRC Tumor Purity
TCGA KIRC xCell/TIMER Cell Fractions
ts-asQTL
i-asQTL
DeCAF QTL

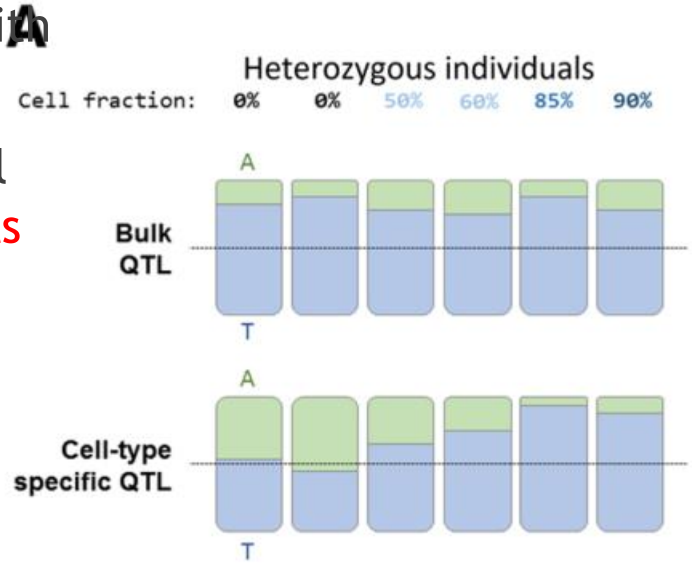
Simulations and Results

J.The Results For Some Real Data:

- We applied DeCAF to TCGA data to perform the first cfQTL mapping effort in tumors(Fig. 1a):

(I) DeCAF identified 3664 significant cfQTL genes across all cell types, 5.63× more than was found using conventional ieQTL mapping (consistent with simulations).

(II) 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

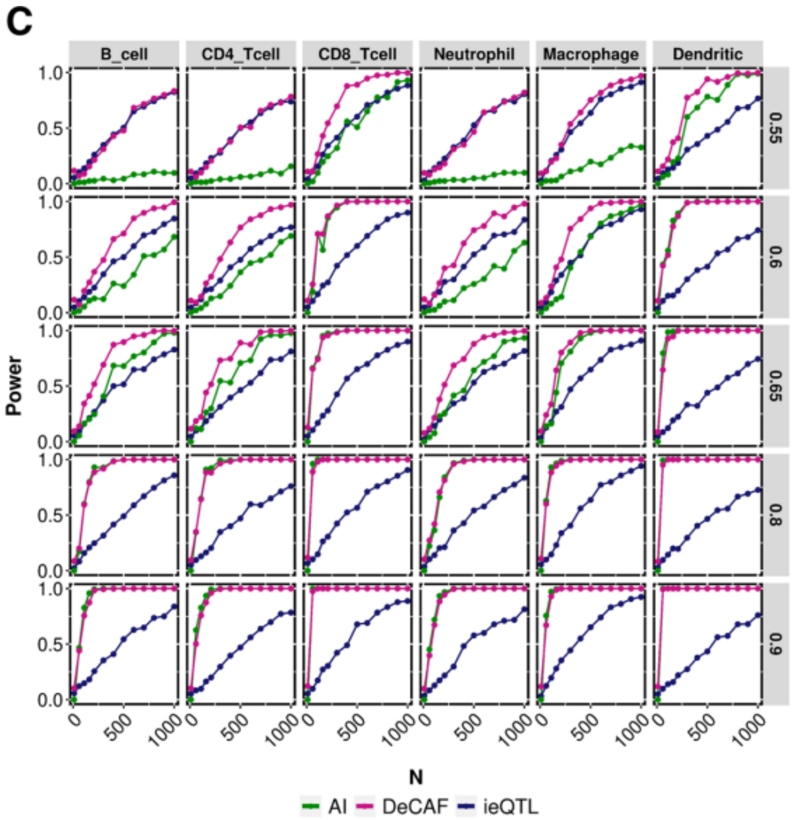


Simulations and Results

- We performed wide-ranging simulations reflecting conditions found in real data and evaluated the power of the interaction eQTL, interaction AI, and DeCAF tests:
- 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).

Simulations and Results

- Real Cell Fractions Case:



Uniform Distribution Case:

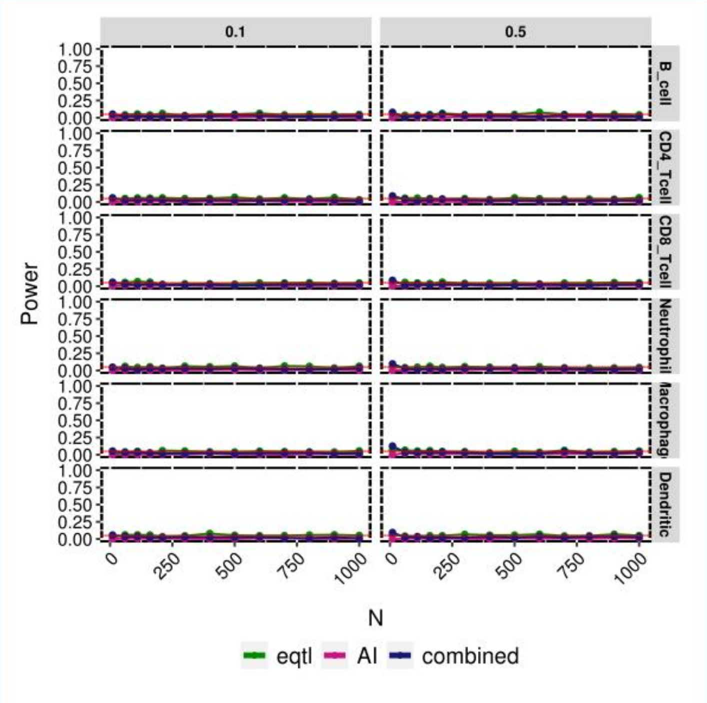
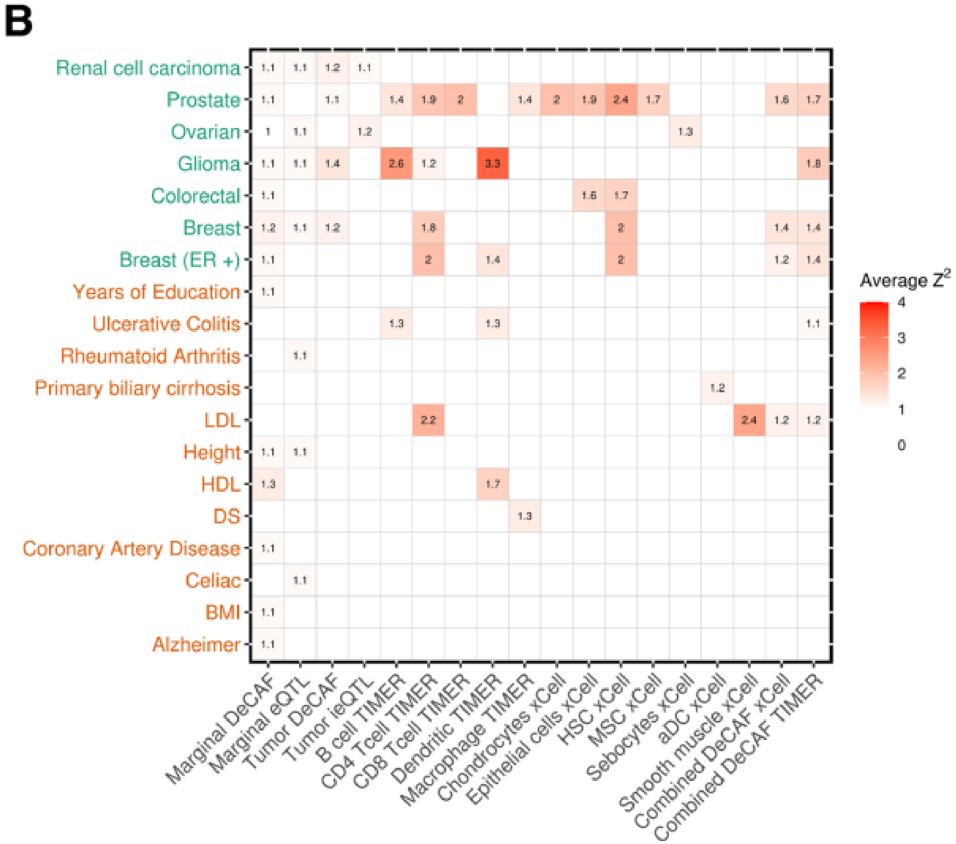


Figure S1 cfQTL models for simulated data from a uniform distribution. A) Plot depicting the power (y-axis) for varying number of individuals (x-axis) and either eQTL (green), AI (pink), or DeCAF combined method (blue) at a null allelic fraction (effect size of AI) and MAF (columns). B) Plot depicting the power (y-axis) for varying number of individuals (x-axis) and either eQTL (green), AI (pink), or DeCAF combined method (blue) at different allelic fractions (rows; effect size of AI) and MAF (columns).

Simulations and Results

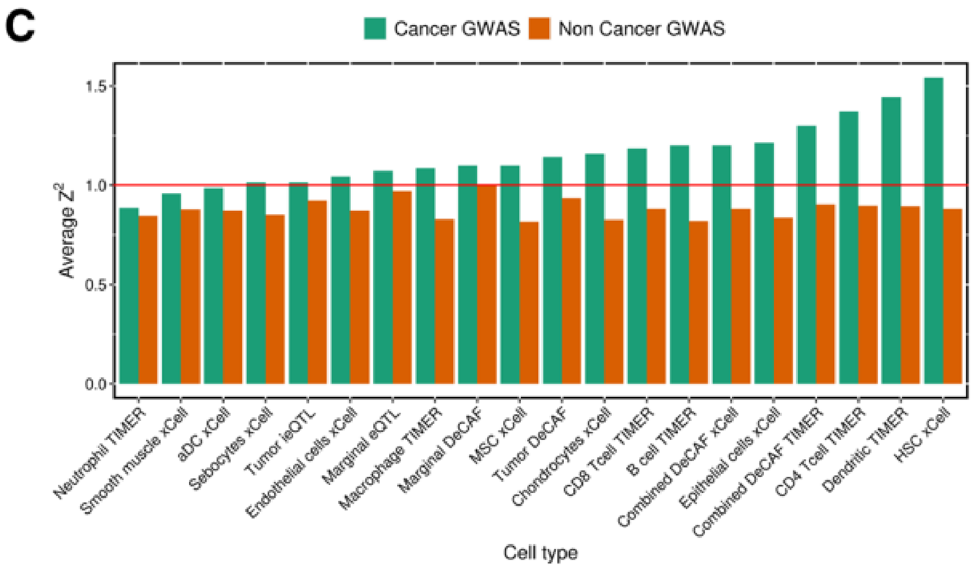
- We observe that every cancer GWAS had at least one significant cfQTL enrichment (Fig. 4b) and most cfQTL cell types exhibited enrichment in cancer GWAS (m



	Marginal DeCAF	Marginal eQTL	Tumor DeCAF	Tumor ieQTL	B cell TIMER	CD4 Tcell TIMER	CD8 Tcell TIMER	Dendritic TIMER	Macrophage TIMER	Chondrocytes xCell	Epithelial cells xCell	HSC xCell	MSC xCell	Sebocytes xCell	aDC xCell	Smooth muscle xCell	Combined DeCAF xCell	Combined DeCAF TIMER
Renal cell carcinoma	1.1	1.1	1.2	1.1														
Prostate	1.1		1.1		1.4	1.9	2		1.4	2	1.9	2.4	1.7				1.6	1.7
Ovarian	1	1.1		1.2										1.3				
Glioma	1.1	1.1	1.4		2.6	1.2		3.3										1.8
Colorectal	1.1										1.6	1.7						
Breast	1.2	1.1	1.2			1.8						2					1.4	1.4
Breast (ER +)	1.1					2		1.4				2					1.2	1.4
Years of Education	1.1																	
Ulcerative Colitis					1.3			1.3										1.1
Rheumatoid Arthritis		1.1																
Primary biliary cirrhosis															1.2			
LDL						2.2										2.4	1.2	1.2
Height	1.1	1.1																
HDL	1.3							1.7										
DS									1.3									
Coronary Artery Disease	1.1																	
Celiac		1.1																
BMI	1.1																	
Alzheimer	1.1																	

Simulations and Results

- Overall, We observe that 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.



Discussion and Conclusions

What We Currently Have Done:

We presented DeCAF, a novel method for identifying cell-type-specific QTLs by harnessing signals from both total and allelic expression.

Compare to Other Methods:

- Although Other methods to detect cell-type-specific eQTL effects from bulk tissue data have recently been applied.
- DeCAF is thus the first method to integrate total and allelic expression together for powerful cfQTL discovery at small to moderate sample sizes
- 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.

Limitations:

- First, DeCAF has limited power to detect cfQTLs in rare cell types (i.e., low cell fractions or cell fraction variance)

Discussion and Conclusions

- Second, DeCAF is inherently dependent on the quality of the deconvolution and cannot test cell types that are not in the reference data.

- Third，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

Discussion and Conclusions

Future Expectations:

- First: DeCAF can be applied to any deconvolution framework or score and so will benefit from improved methodologies in the future:

(Since, most deconvolution methods have been shown to replicate well in pure bulk cell types and they continue to be in active development.)

- Second, 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.

Discussion and Conclusions

- 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.

- Forth, 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

- Finally, 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.

Discussion and Conclusions

- Conclusions:
- (I) 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.
- (II) We present DeCAF as the first to use these analyses to study both cell-type-specific and tumor-specific allelic expression in cancer.
- (III) By using a combination of AI and eQTL mapping, we gained considerable power over previous studies that considered eQTL mapping alone.

Thank you