Functional Genomics of the Pediatric Obese Asthma Phenotype Reveal Enrichment of Rho-GTPase Pathways

Deepa Rastogi, MBBS, MS, Andrew D. Johnston MS, John Nico BS, Lip Nam Loh PhD, Yurydia Jorge MD, Masako Suzuki DVM, PhD, Fernando Macian MD PhD, John M. Greally PhD, DMed

Online Data Supplement

Methods

Study population

One hundred and twenty African American and Hispanic children, ages 7-11 years, with obese asthma (OA, n=59) or normal weight asthma (NwA, n=61) were recruited from clinics at Children's Hospital at Montefiore between 7/2013 to 8/2016. A validation cohort, including 20 obese and 20 normal-weight asthmatic and 15 obese and 15 normal-weight non-asthmatic children, was recruited between 1/2017 and 8/2018. Obesity was defined as Body Mass Index (BMI) >95th percentile for age and sex according to the Center for Disease Control guidelines (1). Asthma was classified based on the clinical diagnosis made by a health care provider that was confirmed from electronic medical records. As previously described (2), all participants underwent anthropometric measurements, skin prick testing for atopic sensitization, and fasting phlebotomy. Pulmonary function testing, including spirometry and lung volume quantification using the nitrogen washout technique, performed according to the American Thoracic Society guidelines, were abstracted from the medical charts (2). Percent predicted values for spirometry indices were calculated using the National Health and Nutrition Examination Survey (NHANES) prediction equations and lung volume indices were calculated using equations developed by the American Thoracic Society workshop (3, 4). The Institutional Review Board at Albert Einstein College of Medicine approved the study.

Study measures

Isolation of CD4+ T (Th) cells and quantification of metabolic measures

Given that we previously observed an association between insulin resistance and Th1 polarization (5), we processed fasting blood for cell and serum separation. Peripheral blood mononuclear cells (PBMCs) were separated using the Ficoll Hypaque method. CD4+ T (Th) cells were isolated from the PBMCs by negative selection using magnetic beads (Easy Sep, Stem Cell Technologies, Tukwila, WA) to avoid any *ex vivo* Th cell stimulation (2). Th cell purity was 95-98% as confirmed by flow cytometry (2). Th cell proportions in PBMCs did not differ between obese (25.1±6.9%) and normal-weight (24.6±6.8%) samples. Fasting serum was used for insulin quantification using radioimmunoassay (Millipore Corporation) on a Wizard2 gamma counter (Perkin Elmer Corporation) and for lipid quantification using an enzymatic immunoassay analyzed on an AU400 chemistry autoanalyzer (Beckman-Coulter Corporation).

Quantification of CD4+ T cell transcriptome using Directional RNA-Seq assay

As previously described, 2.5 µg of RNA extracted from $2x10^6$ unstimulated Th cells underwent directional RNA-seq based library preparation to quantify the obese and normal-weight Th cell transcriptomes (2). All samples underwent quality control testing (2100 Bioanalyzer, Agilent Technologies, Santa Clara, CA) and 112 samples with RNA integrity number of 8 or greater underwent processing for directional RNA-seq library preparation. After removal of ribosomal RNA with the Ribo-Zero rRNA removal kit (Illumina Inc., San Diego, CA), reverse transcription was performed using the SuperScriptIII First-Strand Synthesis system, followed by second strand cDNA synthesis

using dUTP (Thermo Fisher Scientific, Waltham, MA). The double stranded cDNA was fragmented with Covaris (200-300 bp target length), end-repaired, dA tailed, and adaptors added for the Illumina sequencer to allow multiplexing of 8 samples per lane. To maintain directional information, *i.e.* transcribing strand-specific information (6), a combination of dUTP incorporation and uracil-DNA glycosylase were used. All libraries underwent Bioanalyzer testing for quality control and were sequenced on Illumina HiSeq 2500 as 100 bp single-end reads. All bioinformatics analyses, including quality control analysis, and alignment to the Ensembl reference genome were performed on a high performance computing cluster at Albert Einstein College of Medicine, as previously described (2). Picard-tools v 1.119 (7) was used to generate FastQ files, which were trimmed for poor quality bases and adaptor sequences using Trim Galore! v.0.3.7. (8) and aligned to Ensembl release 83 (9) using STAR v.2.5.1b (10) to generate gene counts that were normalized using DESeq (11) on R statistical software, version 3.2.2. Of the initial 120 samples, RNA and RNA-seq libraries from 48 obese and 55 normal-weight asthmatic samples passed QC analysis and were included in the analysis.

Quantification of Th cell subtype proportions using the CD4+ T cell transcriptome

We analyzed the publicly available single-cell RNA-seq reference dataset on naïve (CD4+CD45RA+CD25-), regulatory (CD4+CD25+) and memory (CD4+CD45RO+) Th cells from 10X Genomics (12), using the *Seurat* R package (13) and identified 7 clusters **[Fig. E1a, b]**. After eliminating rRNA, mitochondrial genes, those associated with HLA antigens, and sex chromosome genes, we selected genes which were expressed by at least 30% cells in each cluster, the log-transformed fold change between clusters was greater than 0.32 using the *FindAllMarkers* function of the *Seurat* R package, and then

calculated the median expression value of the genes in each cluster. Cluster 0 overlapped with naïve Th cells while clusters 1, 2 and 3 overlapped with memory Th cells mapped using the 10X Genomics reference dataset. Cluster 4 overlapped with memory and T regulatory cells. Cluster 5, for which several of the top 30 genes overlapped with NK cells, was absent in our samples and Cluster 6 was distinct from the three major Th cell subsets with several of the top 30 genes corresponding to B cells and their precursors. For these reasons, we excluded Clusters 5 and 6 from our further analysis. The signature gene expression of each cluster is summarized in **Table E1**. Using these gene signatures for each cell subtype cluster, the cell subtype proportion estimate for each sample was performed using the *CIBERSORT* function (14) **[Fig. E1c]**.

Quantification of CD4+ T cell DNA methylation using the HELP-tagging assay

Th cell DNA methylation was quantified in 104 samples using the enzyme digestion-based HELP-tagging assay, as previously described (15). One microgram of Th cell genomic DNA, digested by HpaII, was purified and ligated at the cohesive end to the first adaptor including a restriction enzyme site that was then digested by EcoP15I followed by ligation of the second adaptor that was Illumina sequencer compatible. These adaptors served as priming sites for ligation-mediated PCR amplification. To generate longer flanking sequences, a T7 polymerase was added and followed by a reverse transcription step. The libraries were sequenced using Illumina 2500 sequencer at the Einstein Epigenomics Core facility and compared with a reference human MspI library. While MspI digestion is methylation insensitive and cuts the DNA at all CCGG sites, HpaII digestion is methylation sensitive, and cuts DNA only at CCGG sites where the central CG is unmethylated at the cytosine nucleotide. Comparison of HpaII count to MspI count for

each locus thereby allows quantification of DNA methylation. The angle obtained by plotting HpaII count on the y-axis and MspI count on the x-axis provides a quantitative measure of locus-specific methylation. A higher angle formed when HpaII count is high relative to MspI is evidence of relative hypomethylation while a smaller angle is evidence of hypermethylation (16). We quantified percent methylation as the inverse of the methylation angle value. With regards to annotating the CGs, we marked a CG to be located in the gene promoter when it was within 1 kb from the transcription start site, as defined by Ensembl release 83, and in the gene body if it was in the remainder of the gene. The CG was classified to be in a non-promoter cis-regulatory region (enhancer or open chromatin region) as defined in the Ensembl regulatory build (17) or Assay for Transposase Accessible Chromatin followed by high throughput sequencing (ATAC-seq) conducted on Th cells (18). CGs in enhancers were linked to the gene with the closest transcriptional start site, no further than 50,000 bp away. When overlaps of annotation arose, CGs were designated to be only promoter, enhancer, or gene body, in this respective order. In downstream analyses, CGs within promoters and gene bodies were linked to their respective genes. We quantified methylation at $1.4x10^6$ of the $2.4x10^6$ CGs identified by MspI. Of these 135,947 CGs mapped to a gene promoter, 685,377 CGs mapped to a gene body, and 55,924 CGs mapped to an enhancer or open chromatin region. Of the 104 samples, HELP-tagging libraries from 45 obese and 54 normal-weight asthmatic samples passed QC analysis and were included in the analysis.

Identification of expression quantitative trait loci (eQTLs) and methylation quantitative trait loci (meQTLs) in the CD4+ Th study samples

One microgram of genomic DNA isolated from $2x10^6$ PBMCs remaining after Th cell separation was genotyped on the Infinium® Multi-Ethnic Genotyping Array, an array that is enriched for genetic variants associated with complex diseases across diverse ethnicities (19). For all 112 genotyped samples, the signal intensities of the X and Y chromosome were examined for mislabeled samples **[Fig. E3a]**. Array probes perfectly mapping to multiple genomic loci (multi-mapping) and those that do not fully match any locus (missing) were removed. Additionally, for each plate, probes with no calls (NC) in greater than 5% of the samples or whose distribution did not follow the Hardy–Weinberg equilibrium model (p-value < 0.00005) were filtered and removed.

For downstream expression quantitative trait loci (eQTL) and DNA methylation QTL (meQTL) analyses, variants were converted to the Watson (+) strand using the Genome build and Allele definition Conversion Tool (GACT) to correspond to the RNA-seq and HELP-tagging datasets (20). The mbv function of QTLtools v1.1 was implemented to confirm matching between samples with their RNA-seq and HELP-tagging datasets (21). We also examined the relatedness among samples by observing the shared alleles using PLINK v1.90b identity-by-descent **[Fig. E3b]**, and only one half-sibling or closely-related individual out of the pair was retained (22). The genetic diversity and accuracy of self-reported ethnicity was assessed using EIGENSOFT v6.0.1 to perform principle component analysis on samples used for eQTL and meQTL discovery **[Fig. E3c]** (23, 24). Variants were further pruned for the 91 and 76 individuals in eQTL and meQTL analyses, respectively. For both sets, only autosomal variants were retained. All

individuals were missing less than 10% of genotype calls, and variants could only have missing calls for less than 10% of individuals. Finally, variants with a minor allele frequency of less than 0.1 or did not follow the Hardy-Weinberg equilibrium (p-value < 0.000001) were removed.

Validation analyses

To verify RNA-seq findings, we used *PAK3* as a representative biologically relevant gene, since it was differentially expressed between obese and normal-weight samples and was not influenced by Th cell subtype proportions. We also verified *RPS27L* since it was the one e-gene differentially expressed between obese and normal-weight asthmatics. We quantified their expression by quantitative PCR (qPCR) using the TaqMan gene expression assay with commercial qPCR primers (Thermo Fisher Inc, Waltham, MA) and analyzed by the ΔΔCT method. RPLPO was used as the reference gene. Validation was conducted for *PAK3* in a separate cohort of 20 obese and 20 normal-weight asthmatic children, which were compared to 15 obese children and 15 normal-weight children without asthma, to identify the independent contribution of obesity and asthma alone to differential gene expression.

To determine the biological relevance of the differentially expressed and methylated biological pathways in the non-atopic obese asthma phenotype, we conducted functional studies by silencing CDC42, the RhoGTPase most closely linked with PAK3 function (25). Using Amaxa nucleofector, we nucleofected CDC42 and control siRNA (Thermo Fisher Inc, Waltham, MA) in primary human Th cells. Using qPCR, we quantified gene expression of *IFNγ* and *TNF* as pertinent Th1 cytokines, and *IL-4*, as the pertinent Th2

cytokine, before and after CDC42 silencing. *TRAF3* and *HHEX* expression were quantified as measures of off-target effects of nucleofection. The experiment was done in triplicate.

Statistical Analysis

In this multi-omics analysis conducted on R version 3.2.2 **[Fig. 1]**, we quantified differences between the obese and normal-weight asthmatic Th cell transcriptome and Th cell methylome, including the influence of Th cell subtype proportions on these differences. We then investigated the overlap between the obese asthmatic Th cell transcriptome and methylome. To quantify the contribution of genetic variants, we identified the overlap between eQTLs and meQTLs (and their target e-genes and me-genes respectively) with the obese asthmatic Th cell transcriptome and methylome.

Clinical characteristics were compared between obese and normal-weight asthmatics using the Student's T test for continuous variables and χ^2 or Fisher-exact test for categorical variables on STATA version 14. Using R version 3.2.2., we conducted principal component analysis (PCA) to investigate the contribution of biological (age, sex, ethnicity, insulin, and lipid levels) and technical covariates (preparation batch, sequencing batch, percent duplicate reads, total reads and protein coding reads) in the variance of normalized gene expression counts. All except percent duplicate reads, total reads and protein-coding reads were assessed for their contribution to variance of percent DNA methylation. Technical factors, including total number of reads, protein coding reads, percent duplicate reads, and sequencing batch, in addition to insulin, were found to significantly contribute to gene expression variance. The library preparation batch and

sequencing batch, in addition to insulin and LDL, were associated with variance in DNA methylation. We also conducted PCA to quantify the contribution of the Th cell subtypes to variance of gene expression and DNA methylation. In light of the contribution of these biological and technical variables to both gene expression and DNA methylation, we conducted linear regression analysis first without and then with the Th cell subtype proportions to identify their contribution to gene expression and DNA methylation adjusting for the other biologic and technical factors. Age, sex and self-reported ethnicity were included in the model for their demographic relevance. Genes identified by multivariable analysis to be differentially expressed among obese asthmatics with a between-group p-value <0.05 and a false discovery rate (FDR) q-value of <0.05 were retained for further analysis. The CG sites with a between-group methylation difference of 10%, a p-value <0.05, and a q-value of <0.05 were retained as differentially methylated CGs for further analysis. The NetworkAnalyst software (26), with Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database v.10(27) were used to identify relationships and Gene Ontology (GO) pathways enriched in the differentially expressed and differentially methylated genes.

For eQTL discovery, the Reads Per Kilobase Million (rpkm) values were quantified using the QTLtools quan command, and the influence of biological and technical covariates on gene expression principle components (PCs) was investigated **[Fig. E4a]** (28). Based on these covariate contributions to the dataset, we included the first ten PCs of gene expression and the first two PCs of genotype in the *cis*-eQTL linear model. Variants within 1 Mb of a gene's transcription start site were tested for their association to the gene's expression level. Adjusted p-values were calculated by running 10,000 permutations

using QTLtools cis, and the final eQTL set was defined by a conditional pass (28). The discovered eQTLs were then tested for enrichment within genomic regions annotated by an integrative and discriminative epigenome annotation system (IDEAS) for primary T helper naive cells from peripheral blood (29).

A similar strategy was employed to detect meQTLs. Again, DNA methylation PCs were explored for significant covariate influence **[Fig. E4b]**. The first two genotype PCs and the following DNA methylation PCs were incorporated in the meQTL model: 1-4, 6-10, and 12. Using QTLtools permutation and conditional passes, cis-meQTLs were found by 10,000 permutations and testing variants within 1 Mb of a CpG. Both eQTLs and meQTLs are plotted on Manhattan plots.

To investigate the clinical relevance of differential gene expression, a permutation analysis including 1,000 permutations was conducted to determine the mean number of genes, with a non-zero expression value, that were randomly associated with FEV_1/FVC and/or ERV, the two pulmonary function variables most consistently associated with obesity-related asthma **[Fig. E2a-c].** The mean and maximum number of genes randomly associated with FEV_1/FVC and/or ERV were compared with the number of differentially expressed genes associated with these pulmonary function indices, with the latter being marked by the red line in **Figure E2a-c**. Pearson correlation analysis using log-transformed normalized gene count value was then used to quantify the strength of association of a representative set of differentially expressed genes with FEV_1/FVC and ERV among obese and normal-weight children with asthma **[Fig. 6 a-i].**

The gene expression, DNA methylation and genotyping data, and patient characteristics are available at dbGAP study ID 33254.

References

1. Center for Disease Control. https://www.cdc.gov/obesity/. 2018.

2. Rastogi D, Nico J, Johnson AD, Tobias TA, Jorge Y, Macian F, Greally JM. CDC42-related genes are upregulated in T helper cells from obese asthmatic children. *J Allergy Clin Immunol* 2018; 141: 539-548.

3. Hankinson JL, Odencrantz JR, Fedan KB. Spirometric reference values from a sample of the general U.S. population. *Am J Respir Crit Care Med* 1999; 159: 179-187.

4. Stocks J, Quanjer PH. Reference values for residual volume, functional residual capacity and total lung capacity. ATS Workshop and Lung Volume Measurements. Official Statement of The European *Eur Resp J* 1995; 8: 492-506.

5. Rastogi D, Fraser S, Oh J, Huber AM, Schulman Y, Bhagtani RH, Khan ZS, Tesfa L, Hall CB, Macian F. Inflammation, Metabolic Dysregulation and Pulmonary Function Among Obese Asthmatic Urban Adolescents. *Am J Resp Crit Care Med* 2015; 191: 149-160.

6. Podnar J, Deiderick H, Huerta G, Hunicke-Smith S. Next-Generation Sequencing RNA-Seq Library Construction. *Cur Protoc Mol Biol* 2014; 106: 1-19.

7. http://broadinstitute.github.io/picard/.

8. http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

9. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Girón CG, Gordon L, Hourlier T, Hunt S, Johnson N, Juettemann T, Kähäri AK, Keenan S, Kulesha E, Martin FJ, Maurel T, McLaren WM, Murphy DN, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS, Ruffier

M, Sheppard D, Taylor K, Thormann A, Trevanion SJ, Vullo A, Wilder SP, Wilson M, Zadissa A, Aken BL, Birney E, Cunningham F, Harrow J, Herrero J, Hubbard TJ, Kinsella R, Muffato M, Parker A, Spudich G, Yates A, Zerbino DR, Searle SM. Ensembl 2014. *Nucleic Acids Res* 2014: D749-755.

10. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-Seq aligner. *Bioinformatics* 2013; 29: 15-21.

11. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010; 11: R106.

12. https://support.10xgenomics.com/single-cell-gene-expression/datasets.

13. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 2015; 33: 495-502.

14. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* 2015; 12: 453-457.

15. Rastogi D, Suzuki M, Greally JM. Differential epigenome-wide DNA methylation patterns in childhood obesity-associated asthma. *Sci Rep* 2013; 3: 2164.

16. Suzuki M, Jing Q, Lia D, Pascual M, McLellan A, Greally JM. Optimized design and data analysis of tag-based cytosine methylation assays. *Genome Biol* 2010; 11: R36.

17. Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. The ensembl regulatory build. *Genome Biol* 2015; 16: 56.

18. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 2013; 10: 1213-1218.

19. Bien SA, Wojcik GL, Zubair N, Gignoux CR, Martin AR, Kocarnik JM, Martin LW, Buyske S, Haessler J, Walker RW, Cheng I, Graff M, Xia L, Franceschini N, Matise T, James R, Hindorff L, Le Marchand L, North KE, Haiman CA, Peters U, Loos RJ, Kooperberg CL, Bustamante CD, Kenny EE, Carlson CS, Study P. Strategies for Enriching Variant Coverage in Candidate Disease Loci on a Multiethnic Genotyping Array. *PLoS One* 2016; 11: e0167758.

20. Sulovari A, Li D. GACT: a Genome build and Allele definition Conversion Tool for SNP imputation and meta-analysis in genetic association studies. *BMC Genomics* 2014; 15: 610.

21. Fort A, Panousis NI, Garieri M, Antonarakis SE, Lappalainen T, Dermitzakis ET, Delaneau O. MBV: a method to solve sample mislabeling and detect technical bias in large combined genotype and sequencing assay datasets. *Bioinformatics* 2017; 33: 1895-1897.

22. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; 81: 559-575.

23. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet* 2006; 2: e190.

24. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; 38: 904-909.

25. Bagrodia S, Derijard B, Davis RJ, Cerione RA. Cdc42 and PAK-mediated signaling leads to Jun kinase and p38 mitogen-activated protein kinase activation. *J Biol Chem* 1995; 270: 27995-27998.

26. Xia J, Gill EE, Hancock RE. NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. *Nat Protoc* 2015; 10: 823-844.

27. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 2015; 43.

28. Delaneau O, Ongen H, Brown AA, Fort A, Panousis NI, Dermitzakis ET. A complete tool set for molecular QTL discovery and analysis. *Nat Commun* 2017; 8: 15452.

29. Zhang Y, An L, Yue F, Hardison RC. Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic Acids Res* 2016; 44: 6721-6731.

Legends

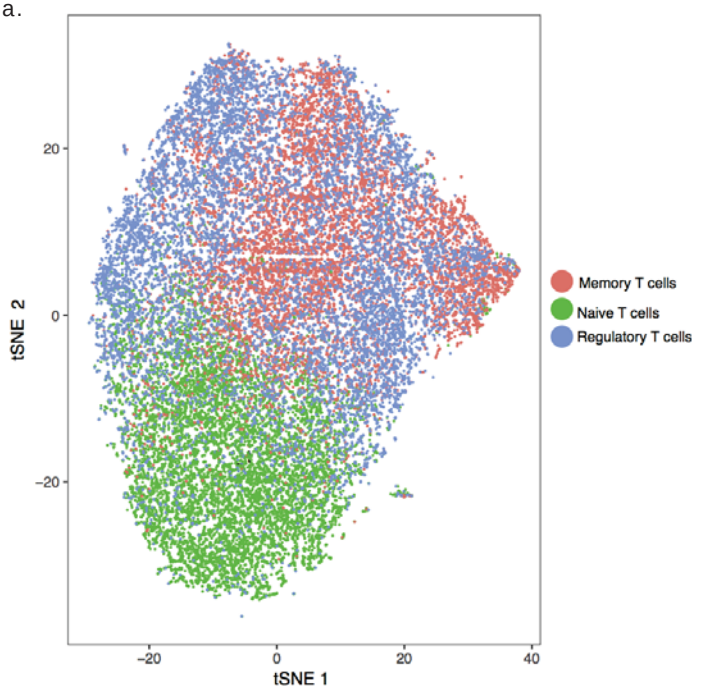
Figure E1. *Quantification of Th cell subtype proportions using the CD4+ T cell transcriptome.* ***a)*** The t-SNE plot summarizes the separation of naïve, memory and regulatory Th cells based on publicly available 10x Genomics single-cell RNA-Seq reference datasets. **b)** Seven Th clusters were identified within the reference naïve, memory and regulatory Th cells, of which 6 were quantified in our samples. **c**) Comparison of the proportion of cells in Th cell clusters between obese and normal-weight asthmatics revealed fewer naïve (cluster 0) and more memory (cluster 1) cells in obese asthmatic samples.

Figure E2. *Permutation analysis to quantify enrichment of the association of differential gene expression with pulmonary function.* A 1000 iteration permutation analysis was conducted to quantify the maximum number of genes whose expression may be randomly associated with pulmonary function indices. As summarized in the bar graphs, where the bars represent the number of times that number of genes were randomly associated with FEV_1/FVC ratio, ERV or both, the maximum number of genes randomly associated **a)** with FEV_1/FVC was 36, **b)** with ERV was 19, **c)** and with both FEV_1/FVC and ERV was 9. Compared to these, as indicated by the red line, of the 157 differentially expressed genes in obese asthmatics **a)** 55 were associated with FEV_1/FVC **b)** 39 were associated with ERV, and **c)** 24 were associated with both FEV_1/FVC and ERV**.** The mean number of genes randomly associated with FEV_1/FVC, ERV or both FEV_1/FVC and ERV was significantly lower ($p<0.001$) as compared to those associated from within the 157 differentially expressed genes suggesting that the differentially expressed genes are enriched in their association with FEV_1/FVC, ERV or both FEV_1/FVC and ERV.

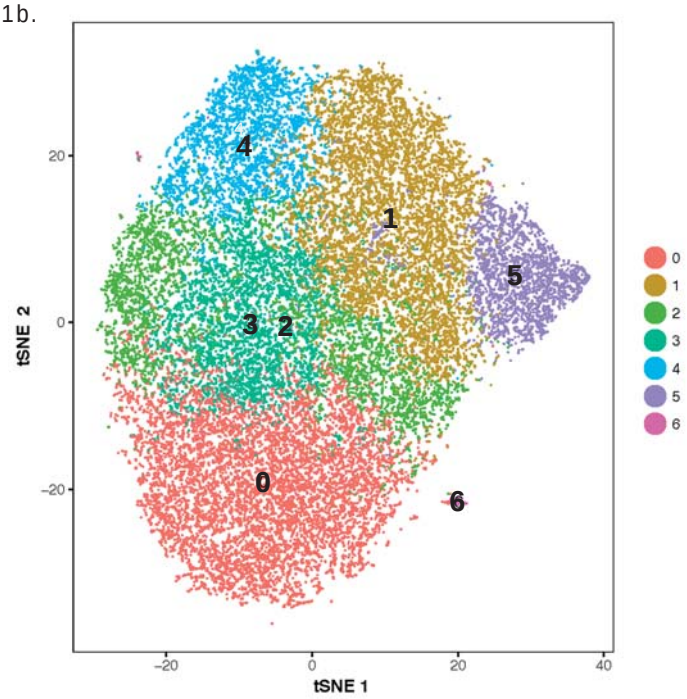
Figure E3. *Genetic array quality control.* **a)** To screen for mislabeled samples, the mean log R ratio (LRR) for the probes on the X and Y chromosomes for each array plate were plotted and colored by the recorded sex on file. One mislabeled sample was identified and removed from downstream analysis. **b)** To avoid genetic similarity confounding our analyses, the probability of 0 (Z0) or 1 (Z1) alleles being identical by descent between a sample pair is plotted to identify closely-related children, potentially not recorded in sample metadata. **c)** The Bronx-based population is primarily of African or Hispanic descent, clustering within the AFR and AMR superpopulation 1000 Genome Project (1000G) distributions when plotting the first two principle components (PC1,PC2) of the genotype data. Additionally, the participant's self-reported ancestry (Study: African American or Hispanic) mostly matches with their ancestry, as determined via principle component analysis. (AFR=African, AMR=Ad Mixed American, EAS=East Asian, EUR= European, SAS=South Asian)

Figure E4. *Contribution of covariates within expression and methylation QTL datasets.* The extent to which technical and biological covariates influence the expression (eQTL) and methylation (mQTL) datasets is captured in this principle component heatmap. The first 20 principle components are separately modeled against each covariate, and significance is denoted by increasingly darker shades of blue (-log of p-value). The percentage of variance that each principle component comprises of the entire dataset is also noted. The identification of principle components, accounting for larger proportion of the overall variance as well as significantly influenced by covariates, informed which components to include in the QTL models.

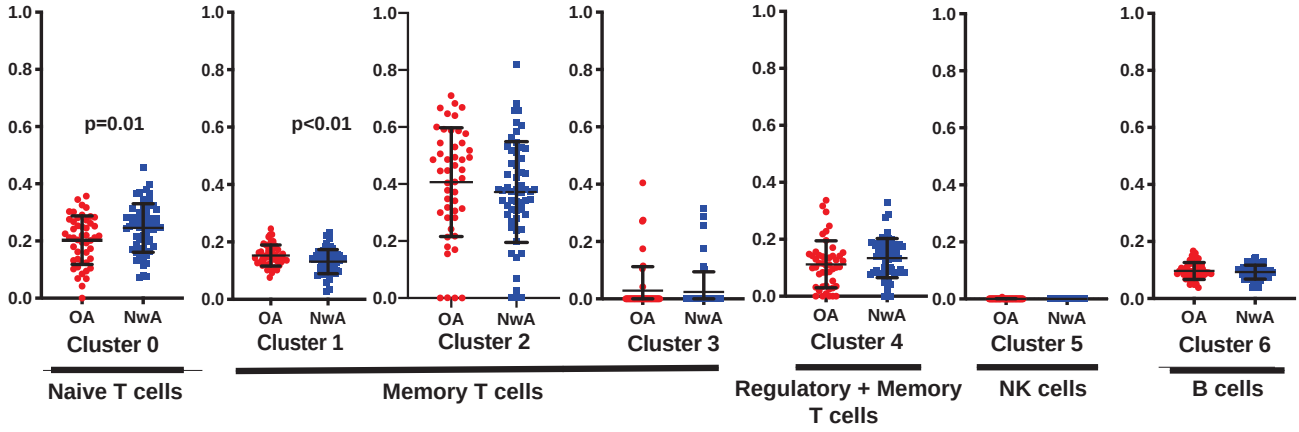
E1a.


tSNE 2
tSNE 1
Memory T cells
Naive T cells
Regulatory T cells

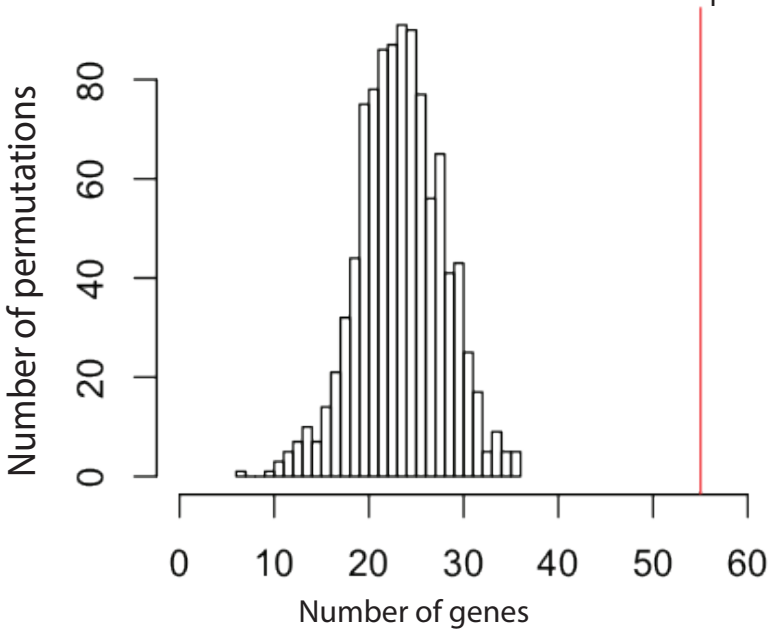
E1b.


tSNE 2
tSNE 1
0
1
2
3
4
5
6

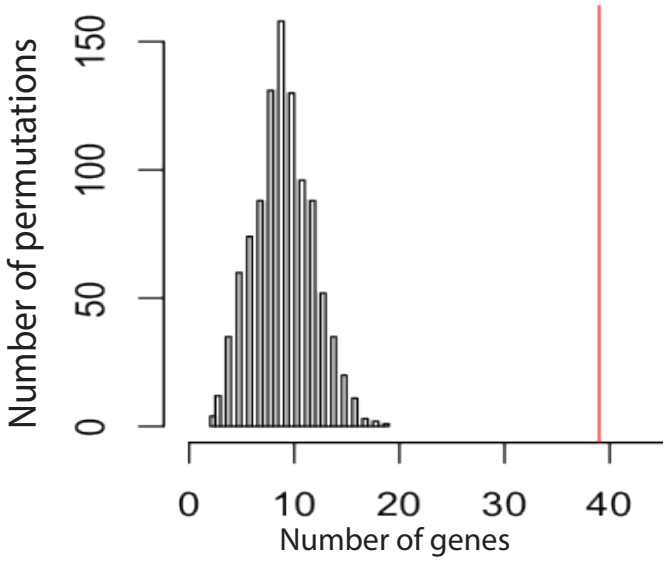
E1c.


p=0.01
p<0.01
OA
NwA
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Naive T cells
Memory T cells
Regulatory + Memory T cells
NK cells
B cells

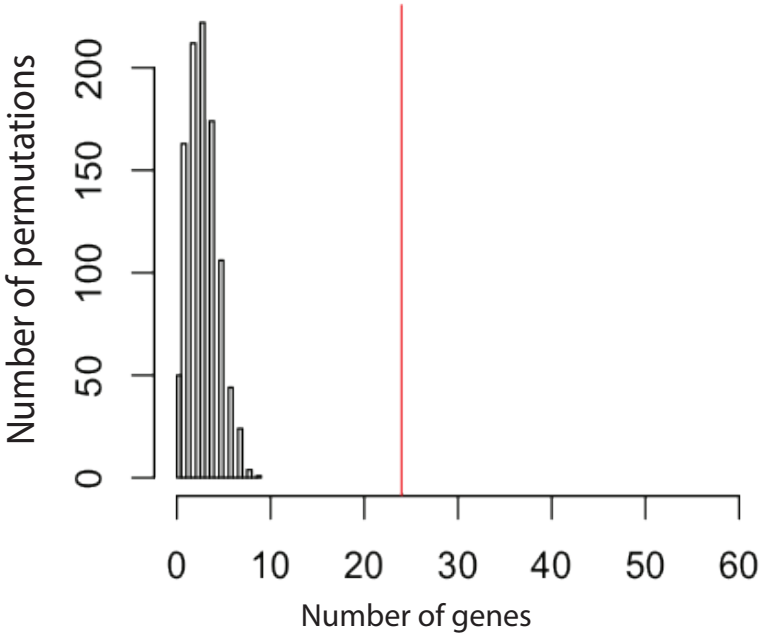
E2a. Enrichment for association with FEV_1/ FVC



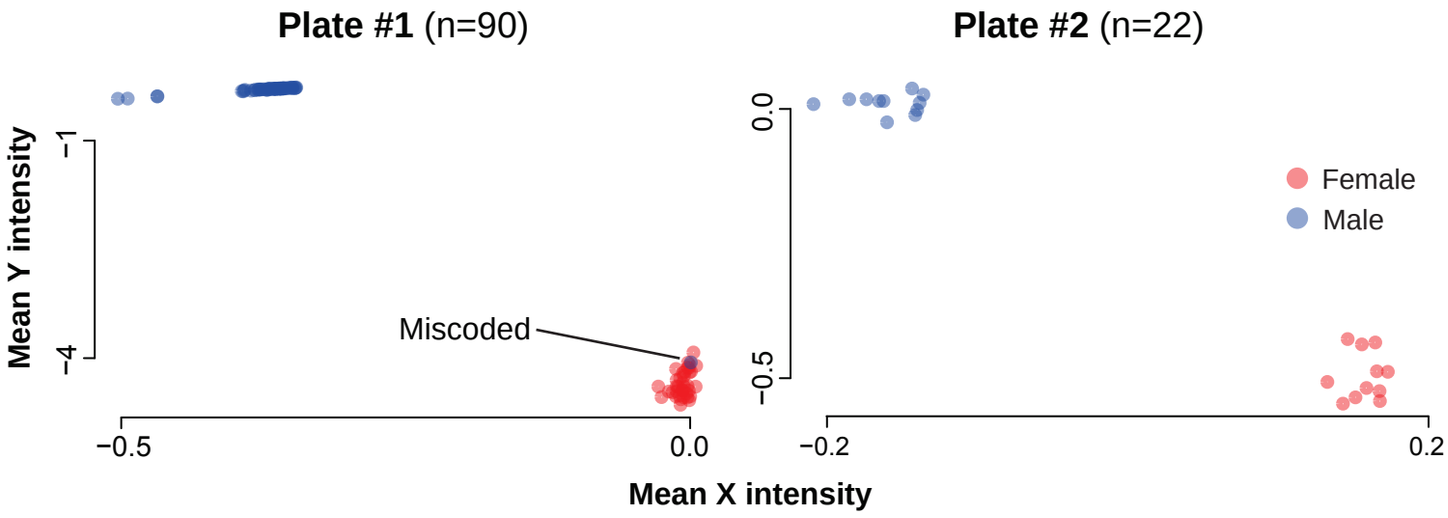
E2b. Enrichment for association with ERV



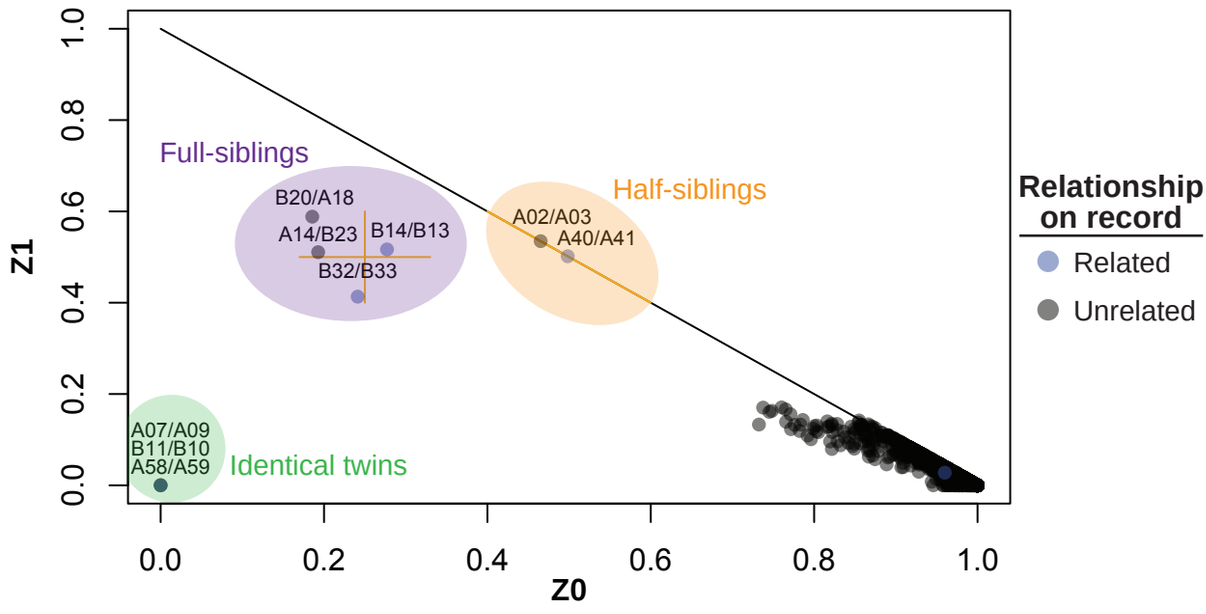
E2c. Enrichment for association with FEV_1/ FVC and ERV



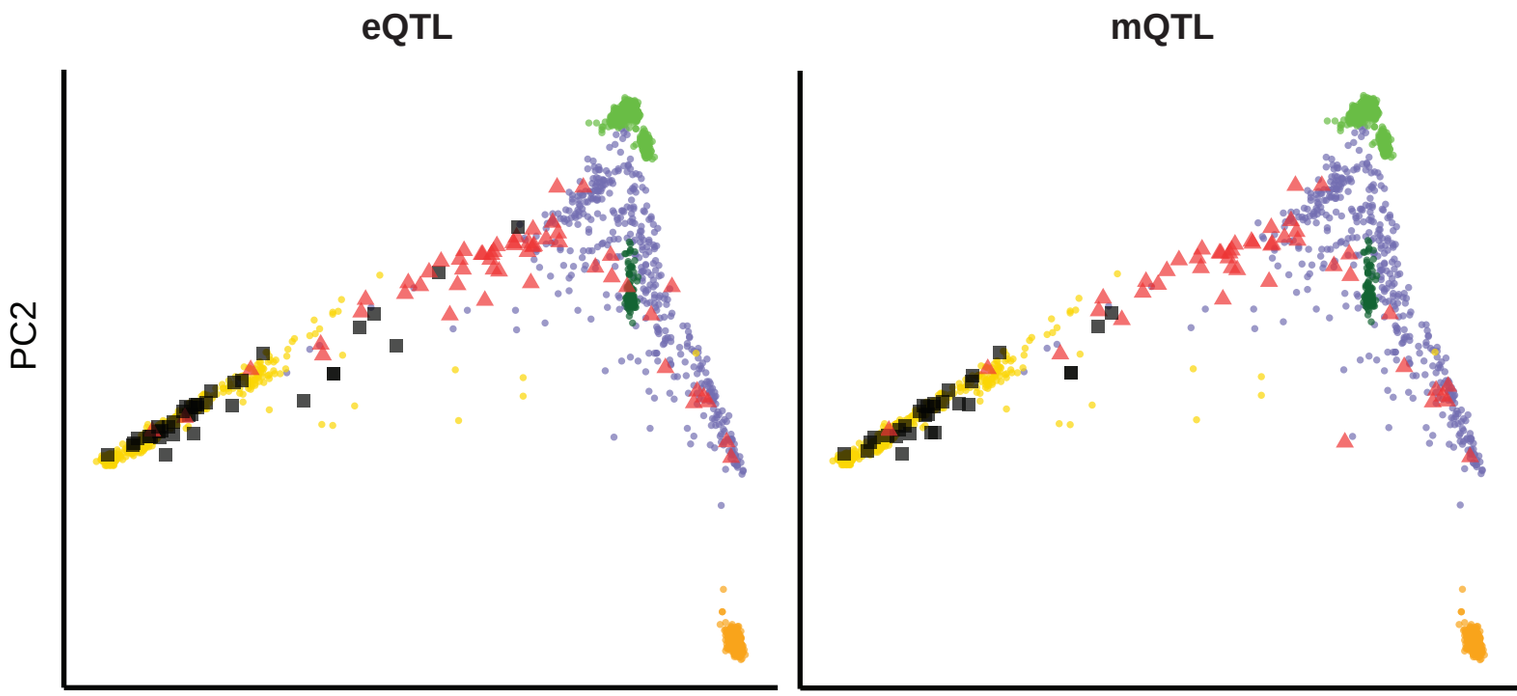
E3a. Assessement of correct sample coding



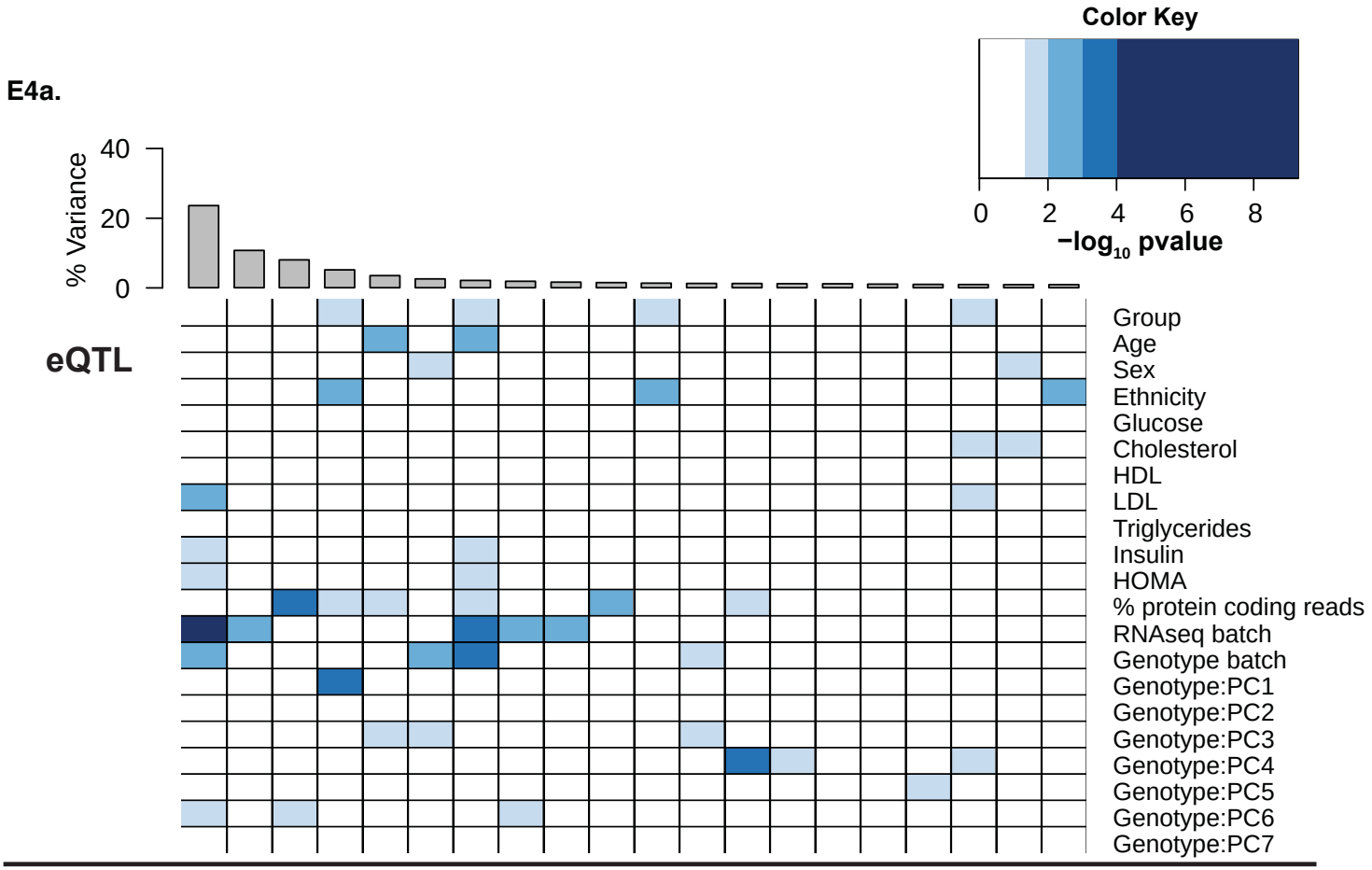
E3b. Identity-by-descent analysis identifies related individuals



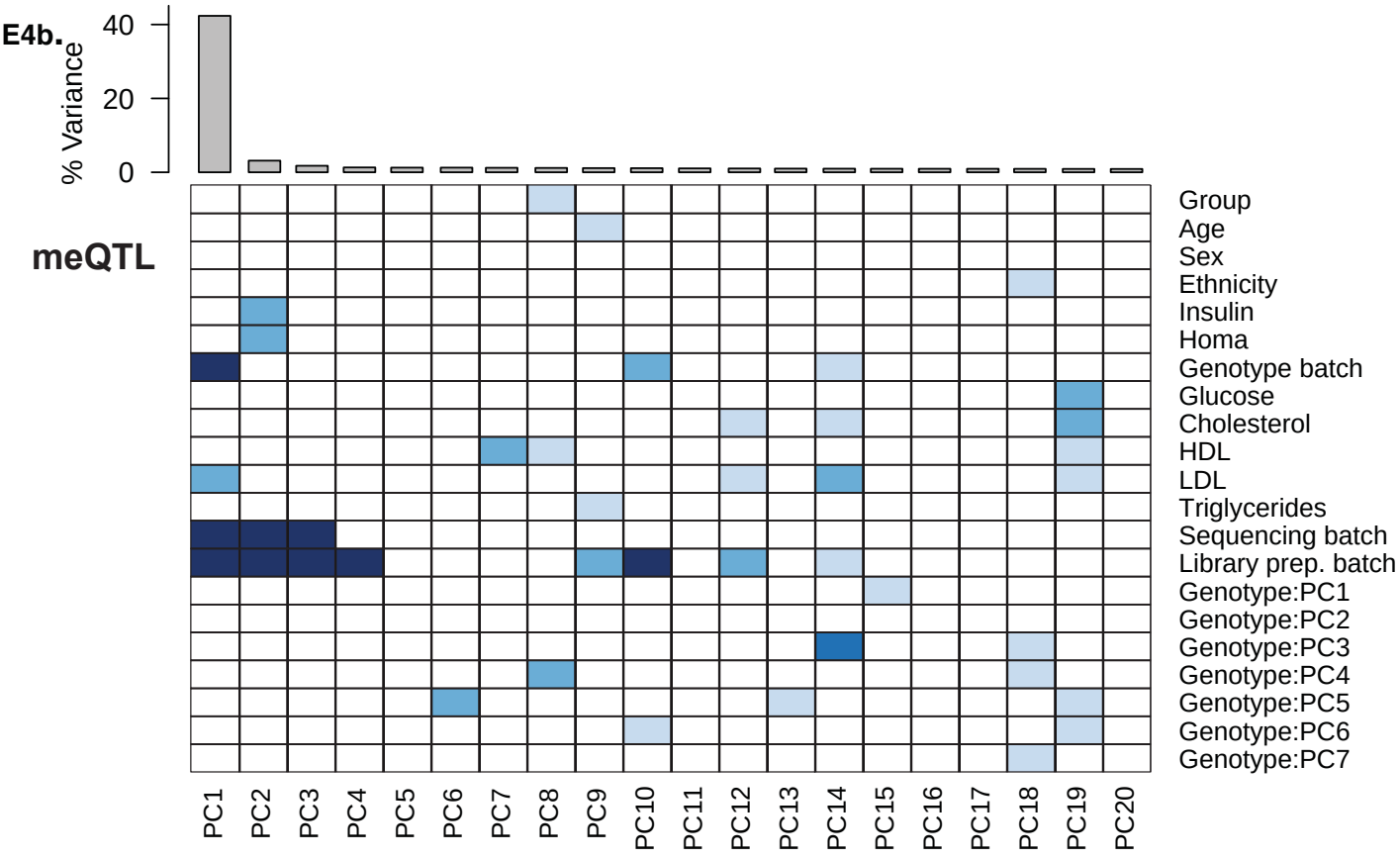
E3c. Genetic distribution of samples used in QTL analyses



E4a.


Color Key
0
2
4
6
8
−log10 pvalue
% Variance
40
20
0
eQTL
Group
Age
Sex
Ethnicity
Glucose
Cholesterol
HDL
LDL
Triglycerides
Insulin
HOMA
% protein coding reads
RNAseq batch
Genotype batch
Genotype:PC1
Genotype:PC2
Genotype:PC3
Genotype:PC4
Genotype:PC5
Genotype:PC6
Genotype:PC7

E4b.


% Variance
40
20
0
meQTL
Group
Age
Sex
Ethnicity
Insulin
Homa
Genotype batch
Glucose
Cholesterol
HDL
LDL
Triglycerides
Sequencing batch
Library prep. batch
Genotype:PC1
Genotype:PC2
Genotype:PC3
Genotype:PC4
Genotype:PC5
Genotype:PC6
Genotype:PC7
PC1
PC2
PC3
PC4
PC5
PC6
PC7
PC8
PC9
PC10
PC11
PC12
PC13
PC14
PC15
PC16
PC17
PC18
PC19
PC20