Integrative genomics analysis identifies ACVR1B as a candidate causal gene of emphysema distribution in non-alpha 1-antitrypsin deficient smokers — Source link ↗

Adel El Boueiz, Robert P. Chase, Andrew Lamb, Sool Lee ...+10 more authors

Institutions: Brigham and Women's Hospital, National Jewish Health

Related papers:

- Genetic control of gene expression at novel and established chronic obstructive pulmonary disease loci

- Expression Quantitative Trait Loci Analysis Identifies Associations Between Genotype and Gene Expression in Human Intestine

- A mega-analysis of expression quantitative trait loci (eQTL) provides insight into the regulatory architecture of gene expression variation in liver.

- Secondary Analyses for Genome-wide Association Studies using Expression Quantitative Trait Loci

- Comparison of quantitative trait loci methods: Total expression and allelic imbalance method in brain RNA-seq.

Share this paper: f ✗ in ✉

View more about this paper here: https://typeset.io/papers/integrative-genomics-analysis-identifies-acvr1b-as-a-2yjh9lo1c0

Integrative genomics analysis identifies *ACVR1B* as a candidate causal gene of emphysema distribution in non-alpha 1-antitrypsin deficient smokers

Adel Boueiz[1,2], Robert Chase[1], Andrew Lamb[1], Sool Lee[1], Zun Zar Chi Naing[1], Michael H. Cho[1,2], Margaret M. Parker[1], Craig P. Hersh[1,2], James D. Crapo[3], Andrew B. Stergachis[4], Ruth Tal-Singer[5], Dawn L. DeMeo[1,2], Edwin K. Silverman[1,2], Xiaobo Zhou[1,2], Peter J. Castaldi[1,6], for the COPDGene investigators.

[1]Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, 02115, USA; [2]Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, 02115, USA; [3]Pulmonary Medicine, National Jewish Health, Denver, CO, 80206, USA; [4]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, 02115, USA; [5]GSK, King of Prussia, 19406, PA; [6]General Medicine and Primary Care, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, 02115, USA.

Corresponding Author: Peter J. Castaldi, MD, MSc, Channing Division of Network Medicine, Brigham and Women's Hospital, 181 Longwood Avenue, Boston, MA, 02115, Email: repjc@channing.harvard.edu

Authors' email addresses: Adel Boueiz (aelboueiz@partners.org), Robert Chase (rerpc@channing.harvard.edu), Andrew Lamb (andrewelamb@gmail.com), Sool Lee (resle@channing.harvard.edu), Zun Zar Chi Naing (rezcn@channing.harvard.edu), Michael H. Cho (remhc@channing.harvard.edu), Margaret M. Parker (rempa@channing.harvard.edu), Craig P. Hersh (craig.hersh@channing.harvard.edu), James D. Crapo (CrapoJ@njhealth.org), Andrew B. Stergachis (astergachis@partners.org), Ruth Tal-Singer (Tal-Singer@gsk.com), Dawn L. DeMeo (redld@channing.harvard.edu), Edwin K. Silverman (reeks@channing.harvard.edu), Xiaobo Zhou (rexiz@channing.harvard.edu), Peter J. Castaldi (repjc@channing.harvard.edu)

ABSTRACT

Background: Several genetic risk loci associated with emphysema apico-basal distribution (EABD) have been identified through genome-wide association studies (GWAS), but the biological functions of these variants are unknown. To characterize gene regulatory functions of EABD-associated variants, we integrated EABD GWAS results with 1) a multi-tissue panel of expression quantitative trait loci (eQTL) from subjects with COPD and the GTEx project and 2) epigenomic marks from 127 cell types in the Roadmap Epigenomics project. Functional validation was performed for a variant near *ACVR1B*.

Results: SNPs from 168 loci with P-values$<5x10^{-5}$ in the largest GWAS meta-analysis of EABD *(Boueiz A. et al, AJRCCM 2017)* were analyzed. 54 loci overlapped eQTL regions from our multi-tissue panel, and 7 of these loci showed a high probability of harboring a single, shared GWAS and eQTL causal variant (colocalization posterior probability≥0.9). 17 cell types exhibited greater than expected overlap between EABD loci and DNase-I hypersensitive peaks, DNaseI hotspots, enhancer marks, or digital DNaseI footprints (permutation P-value < 0.05), with the strongest enrichment observed in $CD4^+$, $CD8^+$, and regulatory T cells. A region near *ACVR1B* demonstrated significant colocalization with a lung eQTL and overlapped DNase-I hypersensitive regions in multiple cell types, and reporter assays in human bronchial epithelial cells confirmed allele-specific regulatory activity for the lead variant, rs7962469.

Conclusions: Integrative analysis highlights candidate causal genes, regulatory variants, and cell types that may contribute to the pathogenesis of emphysema distribution. These findings will enable more accurate functional validation studies and better understanding of emphysema distribution biology.

Keywords: Integrative genomics; Emphysema; Emphysema distribution; Chronic obstructive pulmonary disease.

BACKGROUND

Chronic obstructive pulmonary disease (COPD) is a clinical syndrome with multiple, distinct clinical manifestations including the pathologic loss of lung tissue, i.e. emphysema. The phenotypic variability of COPD has prognostic and therapeutic implications [1-3]. Among patients with emphysema, patterns of lung destruction are often asymmetric [4], and this regional heterogeneity is an important factor in the severity of airflow limitation, disease progression, and the response to lung volume reduction procedures [5-9]. A recent genome-wide association study (GWAS) has identified five genomic loci associated with emphysema apico-basal distribution (EABD) in smokers without alpha-1 antitrypsin deficiency (4q13 near *SOWAHB*, 4q31 near *HHIP*, 8q24 near *TRAPPC9*, 10p12 near *KIAA1462,* and 15q25 near *CHRNA5*) [10]. These loci explain a small proportion of the estimated heritability of EABD, indicating that many more true associations have not yet been identified. In addition, the functional mechanisms of EABD causal variants have not yet been described.

GWAS variants tend to be located in regions of strong linkage disequilibrium (LD), making it difficult to identify the causal variant (or variants) in these regions from genetic association alone. In addition, GWAS-identified regions are frequently located in non-coding DNA and predominantly affect gene expression rather than directly affecting protein structure. This is supported by the observation that GWAS variants are enriched in regulatory domains, including enhancers and regions of open chromatin [11, 12]. We hypothesized that a number of EABD loci regulate gene expression and are located within genomic regions characterized by DNase-I hypersensitivity, enhancer activity, and transcription-factor binding. To test this hypothesis, we performed comprehensive fine-mapping of EABD-associated loci by integrating GWAS results with a large compendium of multi-tissue eQTL and cell-based epigenetic marks. This analysis identifies 7 high-confidence EABD-

3

associated, gene regulatory loci in 42 tissues, and it implicates 17 cell types as likely to participate in EABD pathogenesis, with the strongest enrichment observed for T-cell subsets.

RESULTS

Characteristics of study participants included in the analysis

A previously published GWAS was performed using data from 6,215 non-Hispanic white and 2,955 African-American subjects from COPDGene, 1,538 subjects from ECLIPSE, and 824 subjects from GenKOLS with complete phenotype and genotype data *(Boueiz A. et al, AJRCCM 2017)*. The characteristics of these 11,532 subjects are shown in Table 1S, and the characteristics of the 385 COPDGene NHW subjects with available RNA-seq eQTL data are shown in Table 2S.

Emphysema distribution GWAS signals colocalize with eQTLs

From the whole blood eQTL of the 385 COPDGene subjects, significant associations at 10% FDR were identified for 745,067 unique *cis*-eQTL SNP genotypes associated with the expression level of 17,187 unique transcripts (including protein coding genes, long non-coding RNA, and antisense transcripts) (Table 3S). These eQTLs were analyzed along with the multi-tissue GTEx eQTL data for a total number of unique *cis*-eQTL SNPs per tissue ranging between 129,647 (uterus) and 1,351,125 (thyroid) and unique eQTL transcripts per tissue ranging between 6,452 (uterus) and 27,618 (testis). The workflow of the GWAS-eQTL integration analysis and results are summarized in Figure 1.

Fifty-four genomic loci were associated with both local gene expression and emphysema distribution at FDR 10%, representing 32.1% (54/168) of the total number of candidate EABD-associated loci (GWAS P-value $< 5 \times 10^{-5}$). Given that overlap of GWAS

4

and eQTL signals may occur by chance, a Bayesian test for colocalization was performed to distinguish overlap due to shared causal variants versus chance overlap. 13.1% (22/168) of candidate EABD loci had a reasonable likelihood (>50%) of harboring a shared causal variant for emphysema distribution and eQTL, and 4.2% of loci had a high likelihood (>90%) of harboring a shared causal variant (Table 1), including two regions that harbored genetic variants affecting the expression of *ACVR1B* and *MEI* in lung tissue (Figure 2). The complete set of colocalization results can be viewed interactively at https://cdnm.shinyapps.io/eabd_eqtlcolocalization/.

Emphysema distribution-associated GWAS loci are enriched in T-cell subsets

To quantify the overlap between EABD-associated loci and epigenomic marks in cell types, we investigated whether candidate EABD-associated loci are enriched in gene regulatory regions identified by large scale functional studies performed by the ENCODE and Roadmap Epigenomics consortia [13, 14]. The four types of epigenomic marks that we studied span 39 to 127 diverse cell types and cover on average 0.44% to 2.87% of the genome per cell type (Table 4S). Figure 3 illustrates the workflow for the integrative analysis of GWAS and epigenomic marks.

Forty-five percent of EABD loci (76/168) overlapped at least one of the four studied epigenomic annotations in at least one cell type. Among those, 48 loci overlapped DHS peaks, 54 overlapped DHS hotspots, 60 overlapped enhancer marks, and 26 overlapped DNaseI footprints (28.6%, 32.1%, 35.7%, and 15.5% of EABD loci, respectively).

As with eQTL, some proportion of overlap between EABD loci and regulatory annotations may be due to chance rather than a causal link between the regulatory activity of a locus and its association to emphysema distribution. To better distinguish chance versus causal overlaps, we applied a previously published permutation approach that provides an

estimate of the overall enrichment of GWAS loci in regulatory regions within a given cell type as well as a locus-specific score [15]. As illustrated in Figure 4, a total of 17 different cell types exhibited evidence of enrichment of EABD loci (Permutation p-value <0.05) in at least one set of epigenomic marks, with the most significant enrichment observed for CD4[+], CD8[+], and regulatory T cells. Prioritizing loci by cell type enrichment P-values < 0.05 and overlap locus scores in the lowest 20% of the overall locus score distribution for each epigenomic mark identified, 21 loci (12.5% of the EABD GWAS loci) in 17 different cell types overlapped at least one of the annotations (Table 2). The complete set of results can be viewed at https://cdnm.shinyapps.io/eabd_gwas_roadmap_goshifter/.

Overlaps between EABD GWAS loci, eQTL, and epigenomic marks

Forty-four percent of EABD GWAS-eQTL loci (24/54) also overlapped at least one of the epigenomic marks in at least one cell type (Table 3). Two of these loci (rs12914385 on chromosome 15 *(CHRNA5)* and rs17471079 on chromosome 13 *(STK24)*) had a GWAS-eQTL colocalization P-value ≥ 0.5, cell type epigenomic enrichment P-values < 0.05, and high priority locus scores.

Functional validation of the rs7962469 variant near ACVR1B

Assuming that GWAS regions that colocalize with lung eQTLs may be more likely to play a causal role in emphysema distribution, we focused on the regions near the *ACVR1B* and the *MEI* genes for further functional prioritization. The pattern of association near *MEI* was nearly linear over a broad genomic region of high LD, making the identification of a single causal variant challenging. This, in combination with evidence of colocalization of rs7962469 with DNaseI hypersensitive regions in multiple cell types, led us to prioritize the region near *ACVR1B* for functional characterization. We used the PICS method to identify

6

the 95% credible set of SNPs responsible for the GWAS association near *ACVR1B*, and the causal probability for rs7962469 was estimated to be 96%. Based on these results, we tested for an allelic expression effect of rs7962469 in human bronchial epithelial (16HBE) cells, and we observed that the G variant of rs7962469 has a significantly decreased expression relative to the A variant (Figure 5). The rs7962469 G variant has previously been shown to be associated with COPD susceptibility (OR: 1.13, SE: 1.03; GWAS P-value: 0.002) [16] and with upper lobe emphysema predominance (effect size: 0.06, SE: 0.01, GWAS P-value: 1.7×10^{-5}) [10].

DISCUSSION

Using a multi-cohort GWAS of EABD, multi-tissue eQTL from 45 tissues, and epigenomic marks from 127 cell types, we performed an integrated genetic-epigenomic study to further our functional understanding of common variants associated with this specific manifestation of COPD. eQTL colocalization analysis identified strong evidence of a shared causal variant responsible for observed GWAS and eQTL associations in 7 distinct loci in multiple tissues. GWAS-epigenomic mark enrichment was observed for 17 cell types, with the strongest enrichment in $CD4^+$, $CD8^+$, and regulatory T cells. A region near the promoter of the *ACVR1B* gene demonstrated strong colocalization in lung eQTL data and lies within a DNase-I hypersensitive region that is active in multiple cell types. Reporter assays confirmed allele-specific regulatory activity for the EABD-associated variant, rs7962469, near *ACVR1B* with the G allele associated with decreased reporter gene expression, increased COPD susceptibility, and upper lobe emphysema predominance.

Efforts to understand how genetic variation contributes to common diseases increasingly focus on the regulation of gene expression [17-19]. The enrichment of *cis*-eQTL

7

SNPs has been demonstrated for GWAS-identified loci in COPD [20], and EABD GWAS-eQTL overlap has been previously reported for EABD [21]. With the use of RNA-seq in multiple tissues in GTEx and in blood from subjects with COPD from the COPDGene study, the current study confirmed the previously reported EABD GWAS-eQTL overlap at the 15q25 locus with formal colocalization testing and added evidence for genetic control of gene expression at 12 other emphysema distribution-associated loci.

GWAS-identified loci for some phenotypes are enriched for regulatory marks in tissues that are relevant to the phenotype [21-25]. In the case of emphysema distribution, we observed GWAS-regulatory region overlap in multiple tissue and cell types, highlighting the biological complexity of emphysema and the fact that a significant amount of the regulatory genome is active across multiple tissues and cell types. While some disease loci lie within cell type-specific regulatory regions [26], our results are consistent with Boyle et al.'s work that provided evidence that many complex traits are driven by regulatory variants that tend to be active across many cell types and tissues [27].

Because most GWAS loci discovered to date lie outside of coding regions, it is likely that these loci affect gene regulation. 32.1% of the EABD-associated GWAS loci overlap with eQTL and 45.2% overlap with epigenomic annotations. However, our analysis indicates that a significant proportion of these overlaps may be due to chance, because only 13.1% of candidate EABD regions overlapped an eQTL with a colocalization probability ≥ 0.5, and only 12.5% overlapped an epigenomic mark with a high priority locus score. Furthermore, only 1.2% (2 loci) had a GWAS-eQTL colocalization P-value ≥ 0.5, cell type epigenomic enrichment P-values < 0.05, and high priority locus scores. Our observation that a notable proportion of observed GWAS overlaps with regulatory signals may be due to chance is consistent with previous observations in multiple sclerosis [28]. It is likely that this

observation reflects the high prevalence of regulatory activity in the genome. However, the enrichment methods that we used suffer from important limitations. The Bayesian colocalization method assumes the presence of only one causal variant in a given locus for both GWAS and eQTL signals; this assumption reduces the accuracy of results when the locus contains multiple causal variants [17, 29]. The GoShifter method penalizes regions that are dense in epigenomic annotations and may therefore decrease the power of causal variant identification [15].

It is also interesting to consider the proportion of GWAS signals for which no regulatory overlaps were identified. These instances of non-overlap could be due to false positive GWAS associations at the reduced stringency levels used for this analysis, limited power of the included eQTL analyses [17], limited assessment of the dynamic nature of epigenomic marks in included cell type data, and lack of representation of relevant tissues and cell types for emphysema in GTEx and the Roadmap Epigenomics project, respectively. It is also possible that EABD causal variants affect aspects of gene regulation not observed in our data, such as splicing or post-transcriptional regulation [17].

It has been recognized for many years that emphysema often has a predilection for the upper lobes and subpleural areas in non-alpha-1 antitrypsin deficient smokers [10, 30]. The cause for these regional differences in emphysema distribution is not well understood but has been attributed to regional differences in perfusion, transit time of leukocytes, clearance of deposited dust, mechanical stress and pleural pressure [30-33]. This study provides compelling evidence to support a causal role for the adaptive immune response in EABD, with the strongest enrichment of EABD loci in regulatory activity observed in CD4[+], CD8[+], and regulatory T cells. Furthermore, based on the results of the genomics integrative analysis and the confirmatory reporter assay, our study prioritized the rs7962469 variant, near the *ACVR1B* gene, as a candidate causal variant in emphysema distribution.

9

ACVR1B, also known as *ALK-4*, acts as a transducer of activin-like ligands that are growth and differentiation factors belonging to the transforming growth factor-β (TGF-β) superfamily of signaling proteins. Although *ACVR1B* has not previously been associated with emphysema distribution, prior genetic studies have demonstrated an association of gene polymorphisms of the TGF-β superfamily with COPD [34, 35]. A genome-wide association meta-analysis of 3,497 subjects with severe COPD identified a genome-wide significant association with a locus previously reported to affect the gene expression of *TGFB2* [16]. In addition, *TGFB2* expression levels were reduced in a set of Lung Tissue Research Consortium COPD lung tissue samples compared with controls [16]. A network analysis incorporating COPD GWAS and protein-protein interaction (PPI) data included *ACVR1B* gene in a 10 gene consensus network module associated with COPD case-control status [36]. In addition, differential expression of *ACVR1B* has been found in the epithelial cells of a subset of smokers with lung cancer and in bone marrow micro-metastases from lung cancer patients [37, 38]. More work is warranted to elucidate the role of *ACVR1B* in COPD and emphysema.

The strengths of this study are 1) the use of comprehensive compendia of eQTL and epigenetic marks in multiple tissues and cell types for a novel COPD-related phenotype, 2) application of Bayesian and permutation-based methods to assess the significance of observed overlaps accounting for the genomic abundance of candidate eQTLs and epigenomic annotations, and 3) functional validation and demonstration of allele-specific enhancer activity for a candidate causal variant near *ACVR1B* prioritized by this integrative method.

This study also has important limitations. First, we limited our analysis to *cis*-eQTLs, excluding other classes of gene regulatory variants including *trans*-eQTLs, isoform ratio

10

QTLs, and variants implicated by allele-specific expression analyses. Future studies will be strengthened by the inclusion of these emerging features of eQTL studies [39, 40]. Second, the colocalization and enrichment methods that we used have limitations. The colocalization method does not account for multiple independent signals, and the GoShifter approach is biased against regions that are dense in epigenomic annotations and may therefore decrease the power of causal variant identification [15, 17, 29].

CONCLUSION

This study provides proof of concept for the effectiveness of an approach to leverage multi-tissue compendia of eQTLs and multi-cell compendia of epigenetic marks to refine and characterize functional regulatory GWAS loci. Enrichment analyses implicated a wide range of cells and tissues, emphasizing the importance of having comprehensive compendia of regulatory annotation with respect to tissues, cell types, diseases, and environmental exposures. Based on these integrative analyses, we prioritized and functionally validated a COPD and emphysema-associated variant involved in TGF-β signaling.

MATERIALS AND METHODS

Study populations

We analyzed 11,532 non-alpha-1 antitrypsin deficient current and former smokers with complete genotype and CT densitometry data from four cohorts: The Genetic Epidemiology of COPD study non-Hispanic whites (COPDGene NHW), COPDGene African-Americans (COPDGene AA), the Genetics of Chronic Obstructive Lung Disease (GenKOLS) and the Evaluation of COPD Longitudinally to Identify Predictive Surrogate Endpoints study (ECLIPSE). Detailed descriptions including study populations, genotyping quality control and genotyping imputation have been previously published [10, 41].

CT measurements

Quantitative assessment of emphysema was performed using 3D SLICER density mask analysis (www.chestimagingplatform.org) to determine the percentage of lung voxels with attenuation lower than -950 Hounsfield units (%LAA-950) at maximal inspiration [42]. From these measurements, two correlated but complementary measures of emphysema distribution were constructed: 1) the difference between upper third and lower third emphysema *(diff950)* and 2) the ratio of upper third to lower third emphysema *(ratio950)* [10]. A rank-based inverse normal transformation was applied to both phenotypes to reduce the impact of outliers and deviations from normality [10]. In this current study, given that *ratio950* had a higher heritability and was associated with more genome-wide significant signals compared to *diff950* [10], we performed fine mapping of *ratio950*-associated variants.

Peripheral blood gene expression

385 NHW subjects from the COPDGene study with completed peripheral blood RNA sequencing (RNA-seq) data were analyzed. All samples had RNA integrity number > 7 and RNA concentration $\geq$ 25 µg/µl (COPDGene Phase I dataset, January 6, 2017).

RNA isolation and quality control: Total RNA was extracted from PAXgene Blood RNA tubes using the Qiagen PreAnalytiX PAXgene Blood miRNA Kit. The extraction protocol was performed either manually or with the Qiagen QIAcube extraction robot.

cDNA library preparation and sequencing: Globin reduction and cDNA library preparation for total RNA was performed with the Illumina TruSeq Stranded Total RNA with Ribo-Zero Globin kit. Libraries were pooled and 75bp paired end reads were generated on the Illumina HiSeq2000 platform. Samples were sequenced to an average depth of 20 million reads.

Read alignment, expression quantification, and sequencing quality control: Reads were trimmed of the Truseq adapters using Skewer [43] with default parameters. Trimmed reads were aligned to GRCH38 genome using STAR [44]. Gene level counts were generated using RSubreads with the Ensemble GTF (version 81). Quality control was performed using the Fastqc [45] and RNA-SeQC [46] programs. Samples were included for subsequent analysis if they had >10 million reads, >80% of reads mapped, XIST and Y chromosome expression consistent with reported gender, <10% of R1 reads in the sense orientation, Pearson correlation ≥ 0.9 with other samples in the same library prep batch, and concordant genotype calls between RNA reads and DNA genotyping.

cis-eQTL analysis: Transcript-level expression count data were normalized for library size using the trimmed mean of M values method and then inverse normal transformed [47]. eQTL associations were tested for bi-allelic, autosomal SNPs with minor allele frequency (MAF) > 0.05 and mapping to a dbSNP 142 Reference SNP number. *cis*-eQTL analysis was performed for all SNPs within one megabase of the target gene using Matrix eQTL with a linear model adjusting for age, gender, library prep batch, 3 principal components of genetic ancestry, and 35 PEER factors of gene expression [48]. A total of 5,815,008 SNPs were tested for association with 27,277 transcripts. The threshold for significance was a false discovery rate (FDR) 10%, using the FDR procedure implemented in Matrix eQTL [49].

eQTL-emphysema distribution GWAS colocalization analysis

We integrated emphysema distribution GWAS loci with multi-tissue eQTL data from the Genotype-Tissue Expression (GTEx) project *(version 6)* [10] (Download site: http://www.gtexportal.org/home/datasets; Date of download: November 18, 2015) and whole blood eQTL from 385 COPDGene NHW subjects. GWAS eQTL integration was performed as previously described [21]. Briefly, for each set of eQTL results, GWAS results were

13

filtered to include only SNPs with a significant eQTL association at FDR 10%. In this reduced set of GWAS results, q-values were calculated using the procedure of Storey et al. for the GWAS P-values [50], and SNPs significant in both the eQTL and GWAS analyses at FDR 10% were retained. For each independent association, Bayesian colocalization tests were performed for all SNPs within a 250kb window of the lead GWAS variant at that locus to quantify the probability that the GWAS and eQTL associations were due to a single, shared causal variant [29]. This probability corresponds to the PP_4 number described in the original publication. The workflow of this GWAS-eQTL integration analysis is summarized in Table 2.

Enrichment in tissue-/cell type-specific chromatin states

Genetic variants associated with complex diseases have been shown to overlap regulatory enhancer elements [11, 12]. Based on this finding, we quantified the overlap between emphysema distribution GWAS SNPs (P-value < $5x10^{-5}$) and regulatory elements identified in the Roadmap Epigenomics project *(Release 9)* [51]. For the 127 Roadmap cell types, we downloaded *ChromImpute* imputed annotations that provide the most comprehensive human regulatory region annotation to date for large-scale experimental mapping of epigenomic information [52]. We analyzed *ChromImpute* DNase-I hypersensitive peaks and *ChromImpute* enhancer marks (defined as chromatin states 13 through 18) for all 127 cell types with available data. We also examined DNase-I hypersensitive hotspots that were available for 39 Roadmap cell types. Hotspots are broader regions of DNaseI hypersensitivity encompassing DNase peaks. The hotspot identification algorithm has been previously described [53]. We also analyzed digital DNaseI footprints (DGF) data from 42 uniformly processed cell types in Roadmap. DNaseI footprints are coverage troughs in deeply sequenced DNaseI hypersensitivity data that represent narrow genomic regions shielded from

DNaseI digestion because of transcription factors, and thus represent likely transcription factor binding sites (TFBS).

DHS *ChromImpute* peaks were downloaded from

http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidatedImputed/narrowPeak/ on

July 13, 2016. DHS hotspots were downloaded from

http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/broadPeak/DNase/ on

February 20, 2015. Enhancer marks were downloaded from

http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/imputed12marks/jointModel/final/ on Dec 23, 2015. DGF were downloaded from

http://egg2.wustl.edu/roadmap/data/byDataType/dgfootprints/ on July 13, 2016.

To determine the extent of GWAS-epigenomic annotation overlap, we identified independent emphysema distribution GWAS signals at P-value $< 5 \times 10^{-5}$ within 1MB windows. We then used Genomic Annotation Shifter (GoShifter) to calculate the enrichment for these variants in Roadmap annotations (DHS, enhancer marks, DGF) [15]. This method uses a local permutation strategy to account for the local density of a given epigenomic mark and generates a score for each locus that can be used to prioritize loci where the overlap between a SNP and an annotation is particularly informative. 1,000 permutations were performed using LD information from the 1,000 Genomes EUR population with an r^2 threshold of 0.8. Overlaps in the lowest 20% of locus score distributions and with a cell type enrichment P-value < 0.05 were considered for further analysis (Table 3).

Shiny applications

Searchable tables of the formal metrics, results, and locuszoom plots for each genomic region was compiled and displayed using the R web framework shiny (http://shiny.rstudio.com/) and are available to the public as companion sites for this paper

15

(https://cdnm.shinyapps.io/eabd_eqtlcolocalization/ *(GWAS-eQTL)* and

https://cdnm.shinyapps.io/eabd_gwas_roadmap_goshifter/ *(GWAS-epigenomic annotations))*.

Probabilistic Identification of Causal SNPs - rs7962469

The online Probabilistic Identification of Causal SNPs (PICS) algorithm is a fine mapping algorithm that calculates the probability that an individual SNP is a causal variant given the haplotype structure and observed pattern of association at the locus (https://pubs.broadinstitute.org/pubs/finemapping/pics.php) [54]. We used this algorithm to generate the 95% credible SNP set for the GWAS association identified in the region near the *ACVR1B* gene.

Luciferase reporter assay - rs7962469

Two ~500 base-pair long genomic segments including rs7962469 were obtained from Human Bronchial Epithelial (16HBE) cells heterozygous at rs7962469 and cloned into the sites of XhoI and BglII of a pGL4.23[luc2/minP] vector. Each luciferase construct was co-transfected with TK-Renilla, a luciferase control reporter, in 16HBE cells at ~60-70% confluency by using Lipofectamine 3000 Reagent (Invitrogen), following the manufacturer's protocol. Each luciferase construct was transfected in triplicate at a concentration of 300 ng per well and the TK-Renilla at 15 ng per well. Empty Luciferase vector, pGL4.23[luc2/minP], was also transfected in triplicates as a control. Promoter activity was quantified forty-eight hours post-transfection using the Dual-Luciferase Reporter Assay System (Promega) according to the manufacturer's protocol. Luminescence signals were captured in a Wallac VICTOR3 1420 plate reader (Perkin Elmer) and normalized by the Renilla luciferase readings for each well. The normalized values for each triplicate were then averaged. All plasmids used were confirmed by sequencing. Independent transfection and

16

reporter assays were performed three times. Luciferase activity levels were assessed using Wilcoxon's rank sum test; values were compared with a reference group within an experimental repeat, and results from multiple experiments were included. P-values less than 0.05 were considered significant.

Table 1. Significant colocalizations of emphysema distribution-associated GWAS variants with eQTL from GTEx and the COPDGene study (colocalization probability ≥ 0.9) ordered by the lead SNP GWAS P-values.

Lead SNP	Position	HUGO gene annotation	GWAS P-value	eQTL tissue	eQTL Q-value	Colocalization probability
rs12914385	15:78898723	*CHRNA5*	1.70×10^{-17}	Testis	0.001	0.91
rs2645694	4:77833947	*SOWAHB*	2.40×10^{-8}	Skin: Sun-exposed (Lower leg)	0.007	0.94
				Esophagus: Mucosa	0.003	0.94
rs35500465	19:1158884	*SBNO2*	1.80×10^{-6}	Esophagus: Mucosa	0.001	0.94
rs5758407	22:42076956	*MEI1*	5.90×10^{-6}	Cells: Transformed fibroblasts	4.47×10^{-18}	0.93
				Lung	1.59×10^{-11}	0.91
				Adipose: Subcutaneous	1.53×10^{-9}	0.90
rs17471079	13:99168721	*STK24*	1.10×10^{-5}	Artery: Tibial	1.47×10^{-6}	0.96
rs7962469	12:52348259	*ACVR1B*	1.70×10^{-5}	Lung	0.0005	0.91
rs4468504	14:107122186	*IGHV3-66*	2.70×10^{-5}	Colon: Transverse	6.45×10^{-13}	0.91

Listed are all instances where the Bayesian colocalization results indicate that there is high posterior probability ($PP_4 \geq 0.9$) that the same causal variant is associated with emphysema distribution GWAS and eQTL signals. GWAS: Genome-wide association study; eQTL: Expression quantitative trait loci; Lead SNP: SNP with the lowest P-value for association in a 250-kb window; Colocalization probability: Posterior probability that the same causal variant is associated with emphysema distribution GWAS and eQTL signals.

Table 2. Significant overlaps between emphysema distribution-associated variants and epigenomic annotations (in the lowest 20% of locus score distributions and with a cell type enrichment P-value < 0.05).

A. ChromImpute DNase-I hypersensitive sites

Locus SNP	Position	Nearest Gene	Locus SNP GWAS P-value	Overlap SNP	Overlap SNP GWAS P-value	Cell type	Cell type P-value
rs12914385	15:78862103	*CHRNA3*	1.72×10^{-17}	rs8040868	1.92×10^{-12}	Primary T helper 17 cells PMA-I stimulated;	0.003
						Primary T killer memory cells from peripheral blood	0.01
						Primary T helper memory cells from peripheral blood 1	0.02
						Primary T helper cells PMA-I stimulated	0.02
						Primary T helper naive cells from peripheral blood	0.02
rs141092330	7:17181839	*ANKMY2*	2.44×10^{-5}	rs141092330	2.44×10^{-5}	Primary T helper 17 cells PMA-I stimulated	0.003
						Primary T killer memory cells from peripheral blood	0.01
						Primary T helper memory cells from peripheral blood 1	0.02
						Primary T helper cells PMA-I stimulated	0.02
						Primary T helper naive cells from peripheral blood	0.02
rs142142561	15:100753780	*ADAMTS17*	1.60×10^{-5}	rs142142561	1.60×10^{-5}	Primary T helper 17 cells PMA-I stimulated	0.003
						Primary T killer memory cells from peripheral blood	0.01
						Primary T helper memory cells from peripheral blood 1	0.02
						Primary T helper cells PMA-I stimulated	0.02
						Primary T helper naive cells from peripheral blood	0.02
rs2466200	8:22810931	*PIWIL2*	4.62×10^{-6}	rs2466200	4.62×10^{-6}	Primary T helper 17 cells PMA-I stimulated	0.003

19

Locus SNP	Position	Nearest Gene	Locus SNP GWAS P-value	Overlap SNP	Overlap SNP GWAS P-value	Cell type	Cell type P-value
						Primary T killer memory cells from peripheral blood	0.01
						Primary T helper memory cells from peripheral blood 1	0.02
						Primary T helper cells PMA-I stimulated	0.02
						Primary T helper naive cells from peripheral blood	0.02
rs75992165	11:13631045	*FAR1*	2.27×10^{-6}	rs75992165	2.27×10^{-6}	Primary T helper 17 cells PMA-I stimulated	0.003
						Primary T killer memory cells from peripheral blood	0.01
						Primary T helper memory cells from peripheral blood 1	0.02
						Primary T helper cells PMA-I stimulated	0.02
						Primary T helper naive cells from peripheral blood	0.02

Locus score is probability that an observed instance of overlap between a given genetic variant and set of genomic annotations is due to chance. Cell type enrichment P-value represents the P-value of the degree of overlap with each annotation within each cell type. GoShifter local permutations form the basis of both the locus scores and the cell type enrichment P-values. Annotation of variants to nearest genes was performed using UCSC Genome Browser.

B. DNase hotspots

Locus SNP	Position	Nearest Gene	Locus SNP GWAS P-value	Overlap SNP	Overlap SNP GWAS P-value	Cell type	Cell type P-value
rs12914385	15:78862097	*CHRNA3*	1.72×10^{-17}	rs8040868	1.92×10^{-12}	Primary T cells from peripheral blood	0.02
						Primary Natural Killer cells from peripheral blood	0.04
rs141092330	7:17181846	*ANKMY2*	2.44×10^{-5}	rs141092330	2.44×10^{-5}	Primary T cells from peripheral blood	0.02
						Small Intestine	0.04
						Primary Natural Killer cells from peripheral blood	0.04
rs142142561	15:100753808	*ADAMTS17*	1.60×10^{-5}	rs142142561	1.60×10^{-5}	Primary T cells from peripheral blood	0.02
						Primary Natural Killer cells from peripheral blood	0.04
rs142822120	3:43769809	*ABHD5*	3.28×10^{-5}	rs142822120	3.28×10^{-5}	Small Intestine	0.04
rs182017195	6:17573162	*NUP153*	4.02×10^{-5}	rs77103244	4.72×10^{-5}	Small Intestine	0.04
rs2046399	8:9988262	*MSRA*	3.29×10^{-5}	rs2046399	3.29×10^{-5}	Small Intestine	0.04
rs75992165	11:13631020	*FAR1*	2.27×10^{-6}	rs75992165	2.27×10^{-6}	Primary T cells from peripheral blood	0.02

						Primary Natural Killer cells from peripheral blood	0.04
rs77808082	6:138875121	*NHSL1*	3.32×10^{-5}	rs77808082	3.32×10^{-5}	Small Intestine	0.04

Locus score is probability that an observed instance of overlap between a given genetic variant and set of genomic annotations is due to chance. Cell type enrichment P-value represents the P-value of the degree of overlap with each annotation within each cell type. GoShifter local permutations form the basis of both the locus scores and the cell type enrichment P-values. Annotation of variants to nearest genes was performed using UCSC Genome Browser.

C. ChromImpute enhancer regions (ChromImpute chromatin states 13 through 18)

Locus SNP	Position	Gene annotation	Locus SNP GWAS P-value	Overlap SNP	Overlap SNP GWAS P-value	Cell type	Cell type P-value
rs13141641	4:145435924	*HHIP*	6.34×10^{-18}	rs1813903	1.65×10^{-14}	Thymus	0.03
rs141092330	7:17181789	*ANKMY2*	2.44×10^{-5}	rs141092330	2.44×10^{-5}	Thymus	0.03
rs142822120	3:43769780	*KRBOX1*	3.28×10^{-5}	rs142822120	3.28×10^{-5}	Fetal Heart	0.01
rs12408334	1:201908397	*LMOD1*	3.83×10^{-5}	rs12408334	3.83×10^{-5}	ES-WA7 Cell Line	0.047
rs1452915	7:134486525	*CALD1*	2.12×10^{-5}	rs28485360	2.33×10^{-5}	ES-WA7 Cell Line	0.047
rs17471079	13:99116428	*STK24*	1.07×10^{-5}	rs17574654	3.16×10^{-5}	Thymus	0.03
						Fetal Heart	0.01
rs2046399	8:9988261	*MSRA*	3.29×10^{-5}	rs1484644	3.41×10^{-5}	ES-WA7 Cell Line	0.047
rs2466200	8:22810906	*CPNE3*	4.62×10^{-6}	rs2466200	4.62×10^{-6}	Fetal Heart	0.01
						ES-WA7 Cell Line	0.047
rs56073943	14:94230766	*PRIMA1*	3.11×10^{-5}	rs56073943	3.11×10^{-5}	ES-WA7 Cell Line	0.047
rs5995407	22:37640841	*RAC2*	1.31×10^{-5}	rs5995407	1.31×10^{-5}	ES-WA7 Cell Line	0.047
rs6501394	17:68424096	*KCNJ16*	5.34×10^{-6}	rs6501394	5.34×10^{-6}	ES-WA7 Cell Line	0.047
rs6744412	2:106214065	*FHL2*	2.16×10^{-5}	rs6744412	2.16×10^{-5}	Thymus	0.03
rs72690469	8:87474571	*CPNE3*	7.77×10^{-6}	rs72690446	3.79×10^{-5}	Fetal Heart	0.01
rs73226109	21:42661694	*FAM3B*	1.82×10^{-5}	rs73226109	1.82×10^{-5}	Thymus	0.03
rs966081	13:74864605	*KLF12*	8.45×10^{-6}	rs966081	8.45×10^{-6}	Fetal Heart	0.01

Locus score is probability that an observed instance of overlap between a given genetic variant and set of genomic annotations is due to chance. Cell type enrichment P-value represents the P-value of the degree of overlap with each annotation within each cell type. GoShifter local permutations form the basis of both the locus scores and the cell type enrichment P-values. Annotation of variants to nearest genes was performed using UCSC Genome Browser.

21

D. DNaseI footprints

Locus SNP	Position	Gene annotation	Locus SNP GWAS P-value	Overlap SNP	Overlap SNP GWAS P-value	Cell type	Cell type P-value
rs12914385	15:78862436	*CHRNA3*	1.72×10^{-17}	rs8040868	1.92×10^{-12}	Fetal skin fibroblast	0.03
						Fetal placenta	0.04
rs141092330	7:17182180	*ANKMY2*	2.44×10^{-5}	rs141092330	2.44×10^{-5}	Fetal placenta	0.04
rs1452915	7:134486916	*CALD1*	2.12×10^{-5}	rs28485360	2.33×10^{-5}	iPS 19.11	0.03
rs79958058	3:133597788	*RAB6B*	3.12×10^{-5}	rs79958058	3.12×10^{-5}	Fetal placenta	0.04

Locus score is probability that an observed instance of overlap between a given genetic variant and set of genomic annotations is due to chance. Cell type enrichment P-value represents the P-value of the degree of overlap with each annotation within each cell type. GoShifter local permutations form the basis of both the locus scores and the cell type enrichment P-values. Annotation of variants to nearest genes was performed using UCSC Genome Browser.

Table 3. Number of overlap loci between emphysema apico-basal distribution GWAS, eQTL, and epigenomic marks.

		GoShifter		
		Loci overlapping any annotation (76 loci)	Loci overlapping any annotation with cell type P-value < 0.05 (40 loci)	Loci overlapping any annotation with cell type P-value < 0.05 and in the lowest 20% of locus score distributions (21 loci)
eQTL	All GWAS-eQTL overlap loci (54 loci)	24	15	6
	GWAS-eQTL overlap loci with colocalization probability ≥ 0.5 (22 loci)	11	7	2
	GWAS-eQTL overlap loci with colocalization probability ≥ 0.9 (7 loci)	5	4	2

FIGURE LEGENDS

Figure 1. Workflow and summary of the results of the colocalization analysis of emphysema distribution-associated genetic variants with multi-tissue expression quantitative trait loci (eQTL) data from the Genotype-Tissue Expression (GTEx) project and whole blood eQTL from the COPDGene study.

Figure 2. Genome-wide association study (GWAS) and lung expression quantitative trait loci (eQTL) locus plots for rs5758407 variant near the *MEI1* gene (Panel A) and rs7962469 variant near the *ACVR1B* gene (Panel B).

Figure 3. Workflow and summary of the integrative analysis of emphysema distribution-associated genetic variants with DNase-I hypersensitive peaks, DNaseI hotspots, enhancer marks, and digital DNaseI footprints.

Figure 4. Plot of the cell type enrichment $-\log_{10}$P-value from the Genomic Annotation Shifter (GoShifter) analyses of the overlaps between emphysema distribution-associated genetic variants with DNase-I hypersensitive peaks, DNaseI hotspots, enhancer marks, and digital DNaseI footprints. The cell types with enrichment P-values less than 0.05 are highlighted in red in Figure 4A and are shown in more detail in Figure 4B.

Figure 5. Reporter assays of the rs7962469 variant near the *ACVR1B* gene in Human Bronchial Epithelial (16HBE) cells showing significantly increased expression of the A variant relative to the G variant *(P-value: 0.008)* and the empty Luciferase vector (pGL4.23) *(P-value: 0.00002).* The P-value comparing the G variant to pGL4.23 is 0.02. **: P-value <0.05.

23

REFERENCES

1. Celli BR: **Roger S. Mitchell lecture. Chronic obstructive pulmonary disease phenotypes and their clinical relevance.** *Proc Am Thorac Soc* 2006, **3**(6)**:**461-465.

2. Marsh SE, Travers J, Weatherall M, Williams MV, Aldington S, Shirtcliffe PM, Hansell AL, Nowitz MR, McNaughton AA, Soriano JB, Beasley RW: **Proportional classifications of COPD phenotypes.** *Thorax* 2008, **63**(9)**:**761-767.

3. de Torres JP, Bastarrika G, Zagaceta J, Saiz-Mendiguren R, Alcaide AB, Seijo LM, Montes U, Campo A, Zulueta JJ: **Emphysema presence, severity, and distribution has little impact on the clinical presentation of a cohort of patients with mild to moderate COPD.** *Chest* 2011, **139**(1)**:**36-42.

4. Han MK, Bartholmai B, Liu LX, Murray S, Curtis JL, Sciurba FC, Kazerooni EA, Thompson B, Frederick M, Li D, Schwarz M, Limper A, Freeman C, Landreneau RJ, Wise R, Martinez FJ: **Clinical significance of radiologic characterizations in COPD.** *COPD* 2009, **6**(6)**:**459-467.

5. Venuta F, Anile M, Diso D, Carillo C, De Giacomo T, D'Andrilli A, Fraioli F, Rendina EA, Coloni GF: **Long-term follow-up after bronchoscopic lung volume reduction in patients with emphysema.** *Eur Respir J* 2012, **39**(5)**:**1084-1089.

6. Deslee G, Mal H, Dutau H, Bourdin A, Vergnon JM, Pison C, Kessler R, Jounieaux V, Thiberville L, Leroy S, Marceau A, Laroumagne S, Mallet JP, Dukic S, Barbe C, Bulsei J, Jolly D, Durand-Zaleski I, Marquette CH, REVOLENS Study Group: **Lung Volume Reduction Coil Treatment vs Usual Care in Patients With Severe Emphysema: The REVOLENS Randomized Clinical Trial.** *JAMA* 2016, **315**(2)**:**175-184.

7. Sciurba FC, Chandra D, Bon J: **Bronchoscopic Lung Volume Reduction in COPD: Lessons in Implementing Clinically Based Precision Medicine.** *JAMA* 2016, **315**(2)**:**139-141.

8. Fishman A, Martinez F, Naunheim K, Piantadosi S, Wise R, Ries A, Weinmann G, Wood DE, National Emphysema Treatment Trial Research Group: **A randomized trial comparing lung-volume-reduction surgery with medical therapy for severe emphysema.** *N Engl J Med* 2003, **348**(21)**:**2059-2073.

9. Martinez FJ, Foster G, Curtis JL, Criner G, Weinmann G, Fishman A, DeCamp MM, Benditt J, Sciurba F, Make B, Mohsenifar Z, Diaz P, Hoffman E, Wise R, NETT Research Group: **Predictors of mortality in patients with emphysema and severe airflow obstruction.** *Am J Respir Crit Care Med* 2006, **173**(12)**:**1326-1334.

10. Boueiz A, Lutz SM, Cho MH, Hersh CP, Bowler RP, Washko GR, Halper-Stromberg E, Bakke P, Gulsvik A, Laird NM, Beaty TH, Coxson HO, Crapo JD, Silverman EK, Castaldi PJ, DeMeo DL, COPDGene and ECLIPSE Investigators: **Genome-Wide Association Study of the Genetic Determinants of Emphysema Distribution.** *Am J Respir Crit Care Med* 2017, **195**(6)**:**757-771.

11. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ: **Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS.** *PLoS Genet* 2010, **6**(4)**:**e1000888.

12. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A, Neri F, Lee K, Kutyavin T, Stehling-Sun S, Johnson AK, Canfield TK, Giste E, Diegel M, Bates D, Hansen RS, Neph S, Sabo PJ, Heimfeld S, Raubitschek A, Ziegler S, Cotsapas C, Sotoodehnia N, Glass I, Sunyaev SR, Kaul R,

25

Stamatoyannopoulos JA: **Systematic localization of common disease-associated variation in regulatory DNA.** *Science* 2012, **337**(6099)**:**1190-1195.

13. ENCODE Project Consortium: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**(7414)**:**57-74.

14. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kutyavin T, Lajoie B, Lee BK, Lee K, London D, Lotakis D, Neph S, Neri F, Nguyen ED, Qu H, Reynolds AP, Roach V, Safi A, Sanchez ME, Sanyal A, Shafer A, Simon JM, Song L, Vong S, Weaver M, Yan Y, Zhang Z, Zhang Z, Lenhard B, Tewari M, Dorschner MO, Hansen RS, Navas PA, Stamatoyannopoulos G, Iyer VR, Lieb JD, Sunyaev SR, Akey JM, Sabo PJ, Kaul R, Furey TS, Dekker J, Crawford GE, Stamatoyannopoulos JA: **The accessible chromatin landscape of the human genome.** *Nature* 2012, **489**(7414)**:**75-82.

15. Trynka G, Westra HJ, Slowikowski K, Hu X, Xu H, Stranger BE, Klein RJ, Han B, Raychaudhuri S: **Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci.** *Am J Hum Genet* 2015, **97**(1)**:**139-152.

16. Cho MH, McDonald ML, Zhou X, Mattheisen M, Castaldi PJ, Hersh CP, Demeo DL, Sylvia JS, Ziniti J, Laird NM, Lange C, Litonjua AA, Sparrow D, Casaburi R, Barr RG, Regan EA, Make BJ, Hokanson JE, Lutz S, Dudenkov TM, Farzadegan H, Hetmanski JB, Tal-Singer R, Lomas DA, Bakke P, Gulsvik A, Crapo JD, Silverman EK, Beaty TH, NETT Genetics, ICGN, ECLIPSE and COPDGene Investigators: **Risk loci for chronic obstructive**

pulmonary disease: a genome-wide association study and meta-analysis. *Lancet Respir Med* 2014, **2**(3)**:**214-225.

17. Hormozdiari F, van de Bunt M, Segre AV, Li X, Joo JW, Bilow M, Sul JH, Sankararaman S, Pasaniuc B, Eskin E: **Colocalization of GWAS and eQTL Signals Detects Target Genes.** *Am J Hum Genet* 2016, **99**(6)**:**1245-1260.

18. Mokry M, Harakalova M, Asselbergs FW, de Bakker PI, Nieuwenhuis EE: **Extensive Association of Common Disease Variants with Regulatory Sequence.** *PLoS One* 2016, **11**(11)**:**e0165893.

19. Gusev A, Lee SH, Trynka G, Finucane H, Vilhjalmsson BJ, Xu H, Zang C, Ripke S, Bulik-Sullivan B, Stahl E, Schizophrenia Working Group of the Psychiatric Genomics Consortium, SWE-SCZ Consortium, Kahler AK, Hultman CM, Purcell SM, McCarroll SA, Daly M, Pasaniuc B, Sullivan PF, Neale BM, Wray NR, Raychaudhuri S, Price AL, Schizophrenia Working Group of the Psychiatric Genomics Consortium, SWE-SCZ Consortium: **Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases.** *Am J Hum Genet* 2014, **95**(5)**:**535-552.

20. Castaldi PJ, Cho MH, Zhou X, Qiu W, Mcgeachie M, Celli B, Bakke P, Gulsvik A, Lomas DA, Crapo JD, Beaty TH, Rennard S, Harshfield B, Lange C, Singh D, Tal-Singer R, Riley JH, Quackenbush J, Raby BA, Carey VJ, Silverman EK, Hersh CP: **Genetic control of gene expression at novel and established chronic obstructive pulmonary disease loci.** *Hum Mol Genet* 2015, **24**(4)**:**1200-1210.

21. Castaldi PJ, Cho MH, Zhou X, Qiu W, Mcgeachie M, Celli B, Bakke P, Gulsvik A, Lomas DA, Crapo JD, Beaty TH, Rennard S, Harshfield B, Lange C, Singh D, Tal-Singer R, Riley JH, Quackenbush J, Raby BA, Carey VJ, Silverman EK, Hersh CP: **Genetic control of**

gene expression at novel and established chronic obstructive pulmonary disease loci. *Hum Mol Genet* 2015, **24**(4)**:**1200-1210.

22. Banovich NE, Lan X, McVicker G, van de Geijn B, Degner JF, Blischak JD, Roux J, Pritchard JK, Gilad Y: **Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels.** *PLoS Genet* 2014, **10**(9)**:**e1004663.

23. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE: **Mapping and analysis of chromatin state dynamics in nine human cell types.** *Nature* 2011, **473**(7345)**:**43-49.

24. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A, Neri F, Lee K, Kutyavin T, Stehling-Sun S, Johnson AK, Canfield TK, Giste E, Diegel M, Bates D, Hansen RS, Neph S, Sabo PJ, Heimfeld S, Raubitschek A, Ziegler S, Cotsapas C, Sotoodehnia N, Glass I, Sunyaev SR, Kaul R, Stamatoyannopoulos JA: **Systematic localization of common disease-associated variation in regulatory DNA.** *Science* 2012, **337**(6099)**:**1190-1195.

25. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M: **Linking disease associations with regulatory information in the human genome.** *Genome Res* 2012, **22**(9)**:**1748-1759.

26. Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, Lau E, Jostins L, Plant K, Andrews R, McGee C, Knight JC: **Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression.** *Science* 2014, **343**(6175)**:**1246949.

27. Boyle EA, Li YI, Pritchard JK: **An Expanded View of Complex Traits: From Polygenic to Omnigenic.** *Cell* 2017, **169**(7)**:**1177-1186.

28. Chun S, Casparino A, Patsopoulos NA, Croteau-Chonka DC, Raby BA, De Jager PL, Sunyaev SR, Cotsapas C: **Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types.** *Nat Genet* 2017, **49**(4)**:**600-605.

29. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, Plagnol V: **Bayesian test for colocalisation between pairs of genetic association studies using summary statistics.** *PLoS Genet* 2014, **10**(5)**:**e1004383.

30. Ito I, Nagai S, Handa T, Muro S, Hirai T, Tsukino M, Mishima M: **Matrix metalloproteinase-9 promoter polymorphism associated with upper lung dominant emphysema.** *Am J Respir Crit Care Med* 2005, **172**(11)**:**1378-1382.

31. Gurney JW: **Cross-sectional physiology of the lung.** *Radiology* 1991, **178**(1)**:**1-10.

32. West JB: **Distribution of mechanical stress in the lung, a possible factor in localisation of pulmonary disease.** *Lancet* 1971, **1**(7704)**:**839-841.

33. DeMeo DL, Hersh CP, Hoffman EA, Litonjua AA, Lazarus R, Sparrow D, Benditt JO, Criner G, Make B, Martinez FJ, Scanlon PD, Sciurba FC, Utz JP, Reilly JJ, Silverman EK: **Genetic determinants of emphysema distribution in the national emphysema treatment trial.** *Am J Respir Crit Care Med* 2007, **176**(1)**:**42-48.

34. Warburton D, Shi W, Xu B: **TGF-beta-Smad3 signaling in emphysema and pulmonary fibrosis: an epigenetic aberration of normal development?** *Am J Physiol Lung Cell Mol Physiol* 2013, **304**(2)**:**L83-5.

29

35. Konigshoff M, Kneidinger N, Eickelberg O: **TGF-beta signaling in COPD: deciphering genetic and cellular susceptibilities for future therapeutic regimen.** *Swiss Med Wkly* 2009, **139**(39-40)**:**554-563.

36. McDonald ML, Mattheisen M, Cho MH, Liu YY, Harshfield B, Hersh CP, Bakke P, Gulsvik A, Lange C, Beaty TH, Silverman EK, GenKOLS, COPDGene and ECLIPSE study investigators: **Beyond GWAS in COPD: probing the landscape between gene-set associations, genome-wide associations and protein-protein interaction networks.** *Hum Hered* 2014, **78**(3-4)**:**131-139.

37. Spira A, Beane JE, Shah V, Steiling K, Liu G, Schembri F, Gilman S, Dumas YM, Calner P, Sebastiani P, Sridhar S, Beamis J, Lamb C, Anderson T, Gerry N, Keane J, Lenburg ME, Brody JS: **Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer.** *Nat Med* 2007, **13**(3)**:**361-366.

38. Wrage M, Ruosaari S, Eijk PP, Kaifi JT, Hollmen J, Yekebas EF, Izbicki JR, Brakenhoff RH, Streichert T, Riethdorf S, Glatzel M, Ylstra B, Pantel K, Wikman H: **Genomic profiles associated with early micrometastasis in lung cancer: relevance of 4q deletion.** *Clin Cancer Res* 2009, **15**(5)**:**1566-1574.

39. Sun W, Hu Y: **eQTL Mapping Using RNA-seq Data.** *Stat Biosci* 2013, **5**(1)**:**10.1007/s12561-012-9068-3. Epub 2012 Jun 9 doi:10.1007/s12561-012-9068-3.

40. Joehanes R, Zhang X, Huan T, Yao C, Ying SX, Nguyen QT, Demirkale CY, Feolo ML, Sharopova NR, Sturcke A, Schaffer AA, Heard-Costa N, Chen H, Liu PC, Wang R, Woodhouse KA, Tanriverdi K, Freedman JE, Raghavachari N, Dupuis J, Johnson AD, O'Donnell CJ, Levy D, Munson PJ: **Integrated genome-wide analysis of expression**

quantitative trait loci aids interpretation of genomic association studies. *Genome Biol* 2017, **18**(1)**:**16-016-1142-6.

41. Cho MH, Castaldi PJ, Hersh CP, Hobbs BD, Barr RG, Tal-Singer R, Bakke P, Gulsvik A, San Jose Estepar R, Van Beek EJ, Coxson HO, Lynch DA, Washko GR, Laird NM, Crapo JD, Beaty TH, Silverman EK, NETT Genetics, ECLIPSE, and COPDGene Investigators: **A Genome-wide Association Study of Emphysema and Airway Quantitative Imaging Phenotypes.** *Am J Respir Crit Care Med* 2015, **:**.

42. Coxson HO, Rogers RM, Whittall KP, D'yachkova Y, Pare PD, Sciurba FC, Hogg JC: **A quantification of the lung surface area in emphysema using computed tomography.** *Am J Respir Crit Care Med* 1999, **159**(3)**:**851-856.

43. Jiang H, Lei R, Ding SW, Zhu S: **Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads.** *BMC Bioinformatics* 2014, **15:**182-2105-15-182.

44. 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**(1)**:**15-21.

45. Liao Y, Smyth GK, Shi W: **The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote.** *Nucleic Acids Res* 2013, **41**(10)**:**e108.

46. DeLuca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, Reich M, Winckler W, Getz G: **RNA-SeQC: RNA-seq metrics for quality control and process optimization.** *Bioinformatics* 2012, **28**(11)**:**1530-1532.

47. Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data.** *Genome Biol* 2010, **11**(3)**:**R25-2010-11-3-r25. Epub 2010 Mar 2.

48. Stegle O, Parts L, Durbin R, Winn J: **A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies.** *PLoS Comput Biol* 2010, **6**(5)**:**e1000770.

49. Shabalin AA: **Matrix eQTL: ultra fast eQTL analysis via large matrix operations.** *Bioinformatics* 2012, **28**(10)**:**1353-1358.

50. Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *Proc Natl Acad Sci U S A* 2003, **100**(16)**:**9440-9445.

51. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu YC, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein CB, Gjoneska E, Leung D, Xie W, Hawkins RD, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Hansen RS, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh KH, Feizi S, Karlic R, Kim AR, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong NA, Stevens M, Thurman RE, Wu J, Zhang B, Zhou X, Beaudet AE, Boyer LA, De Jager PL, Farnham PJ, Fisher SJ, Haussler D, Jones SJ, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai LH, Wang W, Waterland RA, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, Kellis M: **Integrative analysis of 111 reference human epigenomes.** *Nature* 2015, **518**(7539)**:**317-330.

52. Ernst J, Kellis M: **Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues.** *Nat Biotechnol* 2015, **33**(4)**:**364-376.

53. John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC, Johnson TA, Hager GL, Stamatoyannopoulos JA: **Chromatin accessibility pre-determines glucocorticoid receptor binding patterns.** *Nat Genet* 2011, **43**(3)**:**264-268.

54. Farh KK, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, Shoresh N, Whitton H, Ryan RJ, Shishkin AA, Hatan M, Carrasco-Alfonso MJ, Mayer D, Luckey CJ, Patsopoulos NA, De Jager PL, Kuchroo VK, Epstein CB, Daly MJ, Hafler DA, Bernstein BE: **Genetic and epigenetic fine mapping of causal autoimmune disease variants.** *Nature* 2015, **518**(7539)**:**337-343.

DECLARATIONS

Ethics approval and consent to participate

Informed consent was obtained from all study subjects and study approval was obtained from institutional review boards for all participating institutions.

Availability of data and material

COPDGene data are available in the NCBI dbGaP database of genotypes and phenotypes under accession number phs000179.v1.p1.

Competing interests

Dr. Hersh reports personal fees from AstraZeneca, grants from Boehringer Ingelheim, personal fees from Mylan, personal fees from Concert Pharmaceuticals which are outside the submitted work. Dr. Cho reports grants from GSK, grants from NIH / NHLBI during the

conduct of the study. Dr. DeMeo reports grants from NIH, personal fees from Novartis outside the submitted work. Dr. Silverman reports grants from NIH during the conduct of the study, personal fees from Novartis, and grant and travel support from GlaxoSmithKline outside the submitted work. Dr. Castaldi reports grants and Advisory Board membership from GSK outside the submitted work. All the other authors have no conflict of interest or financial relationships to disclose. No form of payment was given to anyone to produce the manuscript.

Consent for publication

Not applicable.

Funding

Authors' contributions

Dr. Castaldi had full access to all of the data in the study, takes responsibility for the integrity of the data and the accuracy of the data analysis, had authority over manuscript preparation and the decision to submit the manuscript for publication.

Study concept and design: Boueiz, Castaldi

Acquisition, analysis, or interpretation of data: All authors

Drafting of the manuscript: Boueiz, Castaldi

Critical revision of the manuscript for important intellectual content: All authors

Statistical analysis: Boueiz, Castaldi

Obtained funding: Castaldi, Crapo, Silverman

Study supervision: All authors

All authors gave final approval of the version to be published and have agreed to be accountable for all aspects of the work.

PFT QA Core, Salt Lake City, UT: Robert Jensen, PhD

Data Coordinating Center and Biostatistics, National Jewish Health, Denver, CO: Douglas Everett, PhD, Anna Faino, MS, Matt Strand, PhD, Carla Wilson, MS

Epidemiology Core, University of Colorado Anschutz Medical Campus, Aurora, CO: John E. Hokanson, MPH, PhD, Gregory Kinney, MPH, PhD, Sharon Lutz, PhD, Kendra Young PhD, Katherine Pratte, MSPH, Lindsey Duca, MS.

COPDGene Investigators – Clinical Centers:

Ann Arbor VA: Jeffrey L. Curtis, MD, Carlos H. Martinez, MD, MPH, Perry G. Pernicano, MD

Baylor College of Medicine, Houston, TX: Nicola Hanania, MD, MS, Philip Alapat, MD, Venkata Bandi, MD, Mustafa Atik, MD, Aladin Boriek, PhD, Kalpatha Guntupalli, MD, Elizabeth Guy, MD, Amit Parulekar, MD, Arun Nachiappan, MD

Brigham and Women's Hospital, Boston, MA: Dawn DeMeo, MD, MPH, Craig Hersh, MD, MPH, George Washko, MD, Francine Jacobson, MD, MPH

Columbia University, New York, NY: R. Graham Barr, MD, DrPH, Byron Thomashow, MD, John Austin, MD, Belinda D'Souza, MD, Gregory D.N. Pearson, MD, Anna Rozenshtein, MD, MPH, FACR

Duke University Medical Center, Durham, NC: Neil MacIntyre, Jr., MD, Lacey Washington, MD, H. Page McAdams, MD

Health Partners Research Foundation, Minneapolis, MN: Charlene McEvoy, MD, MPH, Joseph Tashjian, MD

Johns Hopkins University, Baltimore, MD: Robert Wise, MD, Nadia Hansel, MD, MPH, Robert Brown, MD, Karen Horton, MD, Nirupama Putcha, MD, MHS,

Los Angeles Biomedical Research Institute at Harbor UCLA Medical Center, Torrance, CA: Richard Casaburi, PhD, MD, Alessandra Adami, PhD, Janos Porszasz, MD, PhD, Hans Fischer, MD, PhD, Matthew Budoff, MD, Harry Rossiter, PhD

Michael E. DeBakey VAMC, Houston, TX: Amir Sharafkhaneh, MD, PhD, Charlie Lan, DO

Minneapolis VA: Christine Wendt, MD, Brian Bell, MD

Morehouse School of Medicine, Atlanta, GA: Marilyn Foreman, MD, MS, Gloria Westney, MD, MS, Eugene Berkowitz, MD, PhD

National Jewish Health, Denver, CO: Russell Bowler, MD, PhD, David Lynch, MD

Reliant Medical Group, Worcester, MA: Richard Rosiello, MD, David Pace, MD

Temple University, Philadelphia, PA: Gerard Criner, MD, David Ciccolella, MD, Francis Cordova, MD, Chandra Dass, MD, Gilbert D'Alonzo, DO, Parag Desai, MD, Michael Jacobs, PharmD, Steven Kelsen, MD, PhD, Victor Kim, MD, A. James Mamary, MD, Nathaniel Marchetti, DO, Aditi Satti, MD, Kartik Shenoy, MD, Robert M. Steiner, MD, Alex Swift, MD, Irene Swift, MD, Maria Elena Vega-Sanchez, MD

University of Alabama, Birmingham, AL: Mark Dransfield, MD, William Bailey, MD, J. Michael Wells, MD, Surya Bhatt, MD, Hrudaya Nath, MD

University of California, San Diego, CA: Joe Ramsdell, MD, Paul Friedman, MD, Xavier Soler, MD, PhD, Andrew Yen, MD

University of Iowa, Iowa City, IA: Alejandro Cornellas, MD, John Newell, Jr., MD, Brad Thompson, MD

University of Michigan, Ann Arbor, MI: MeiLan Han, MD, Ella Kazerooni, MD, Carlos Martinez, MD

University of Minnesota, Minneapolis, MN: Joanne Billings, MD, Tadashi Allen, MD

University of Pittsburgh, Pittsburgh, PA: Frank Sciurba, MD, Divay Chandra, MD, MSc, Joel Weissfeld, MD, MPH, Carl Fuhrman, MD, Jessica Bon, MD

University of Texas Health Science Center at San Antonio, San Antonio, TX: Antonio

Anzueto, MD, Sandra Adams, MD, Diego Maselli-Caceres, MD, Mario E. Ruiz, MD
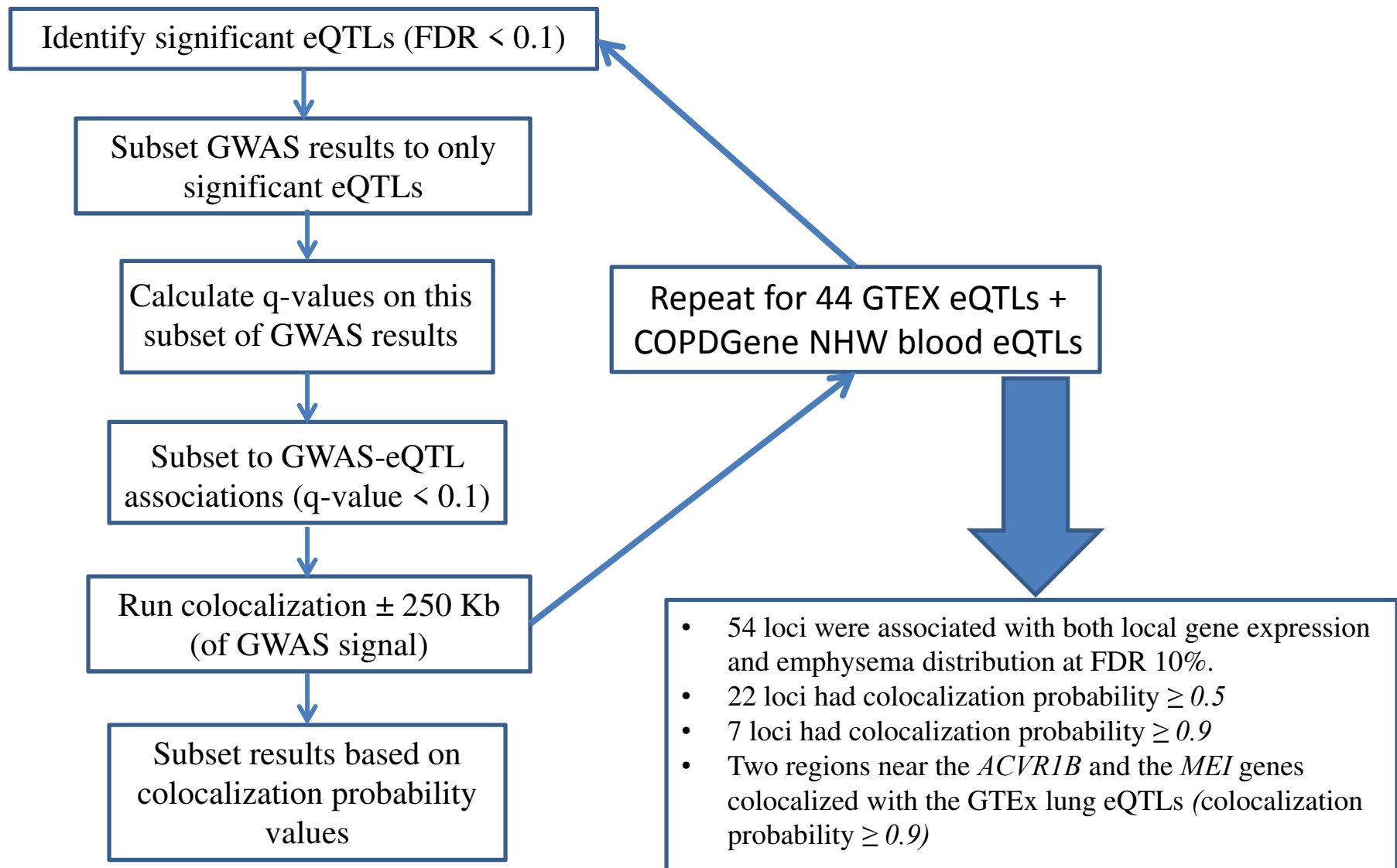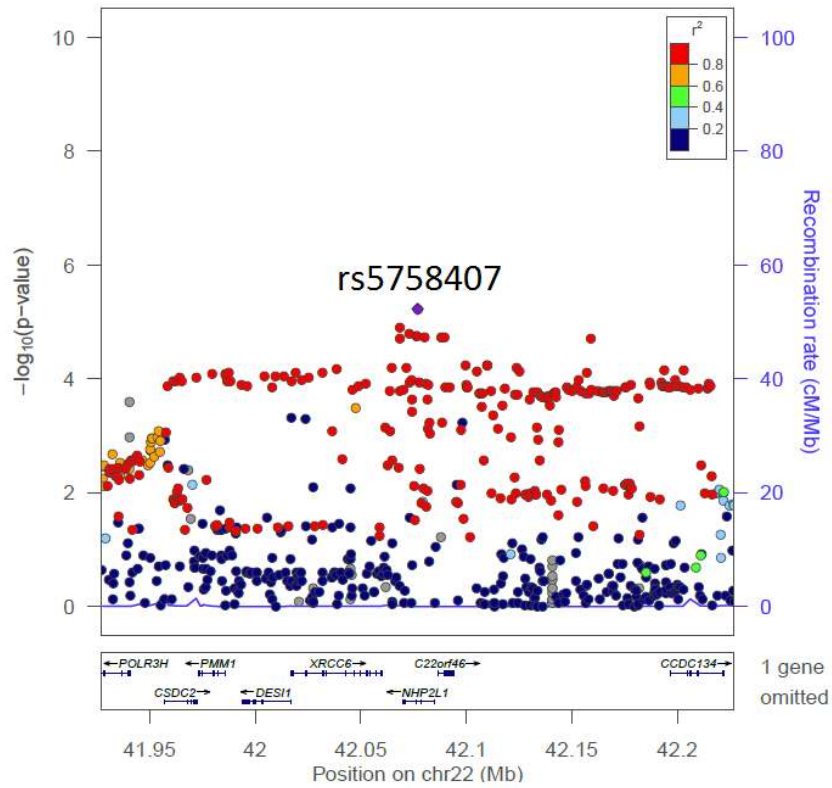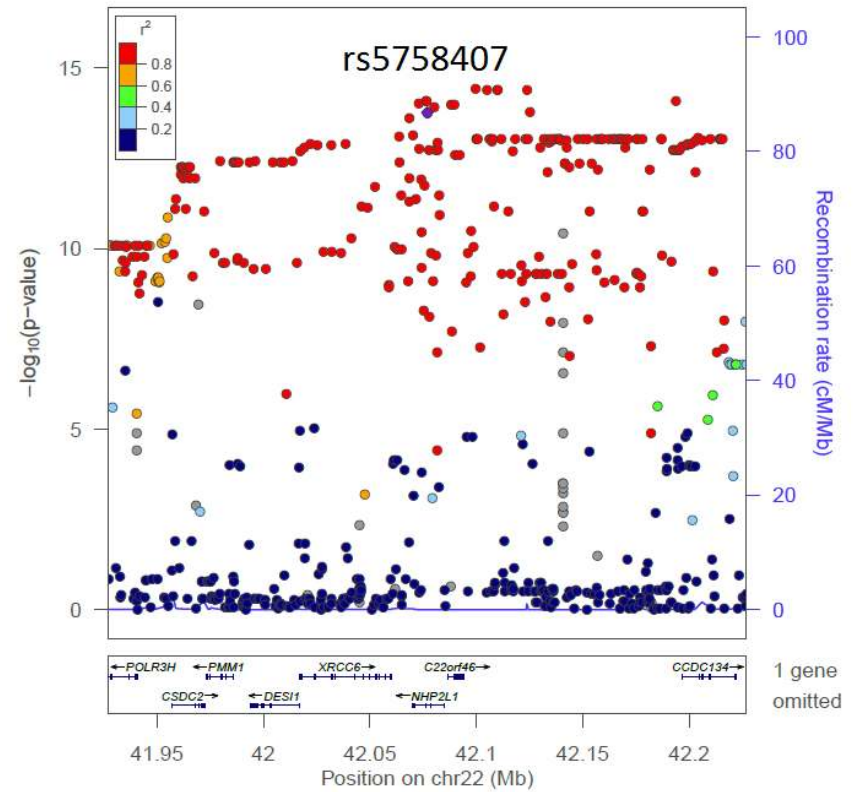
Figure 1

Identify significant eQTLs (FDR < 0.1)

Subset GWAS results to only significant eQTLs

Calculate q-values on this subset of GWAS results

Subset to GWAS-eQTL associations (q-value < 0.1)

Run colocalization ± 250 Kb (of GWAS signal)

Subset results based on colocalization probability values

Repeat for 44 GTEX eQTLs + COPDGene NHW blood eQTLs

- 54 loci were associated with both local gene expression and emphysema distribution at FDR 10%.
- 22 loci had colocalization probability ≥ 0.5
- 7 loci had colocalization probability ≥ 0.9
- Two regions near the *ACVR1B* and the *MEI* genes colocalized with the GTEx lung eQTLs *(colocalization probability ≥ 0.9)*
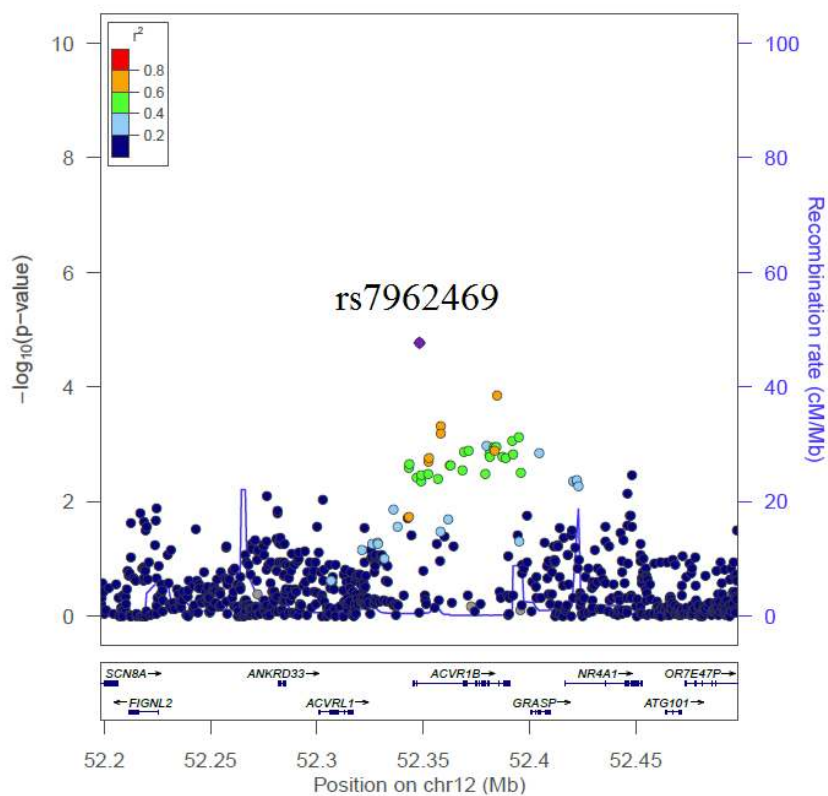
Figure 2

Panel A

rs5758407 variant near the *MEI1* gene

Panel B

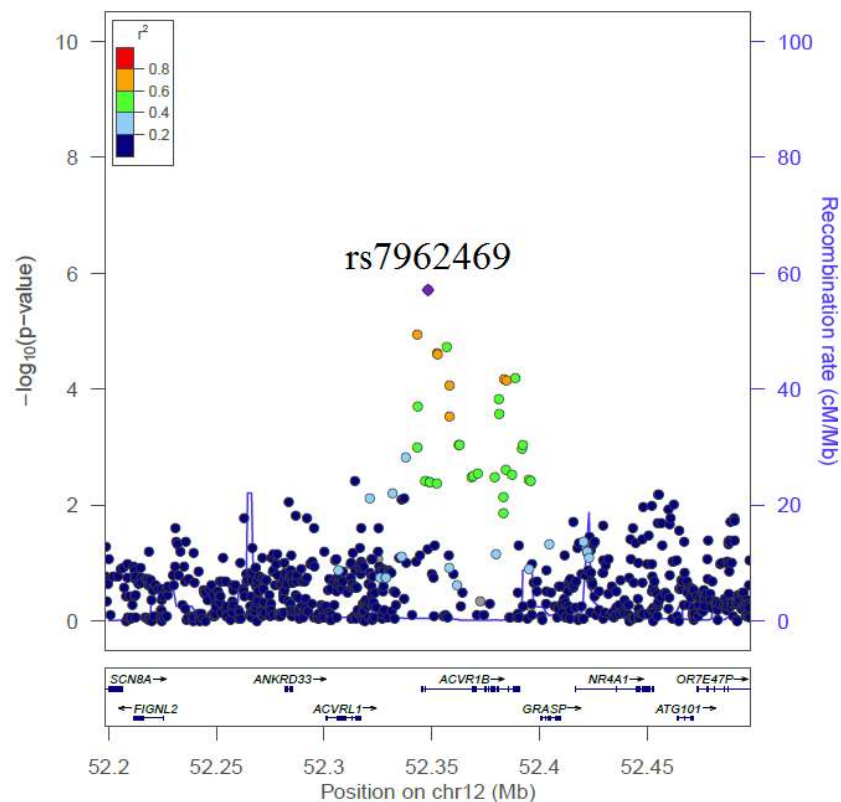# rs7962469 variant near the *ACVR1B* gene

Figure 3

Emphysema distribution associated loci:

- 168 independent loci associated with emphysema distribution at GWAS P-value < 5×10^{-5} *(Boueiz et al, AJRCCM 2017)*

- Identify all variants in LD with each emphysema distribution index SNP ($r^2 > 0.8$ in the 1,000 Genomes Project European samples)
→ Total of 2,033 SNPs

Epigenetic annotations:

- *ChromImpute* DNase-I hypersensitive sites (DHS peaks and hotspots)
- Enhancer marks (*ChromImpute* chromatin states 13 -18)
- Digital DNaseI footprints

Find overlap:

- For each annotation, identify instances of overlap with emphysema distribution-associated SNPs.
    o 48 loci overlapped with DHS peaks
    o 54 loci overlapped with DHS hotspots
    o 60 loci overlapped with enhancer marks
    o 26 loci overlapped with DNaseI footprints
    o 76 loci overlapped with at least one annotation
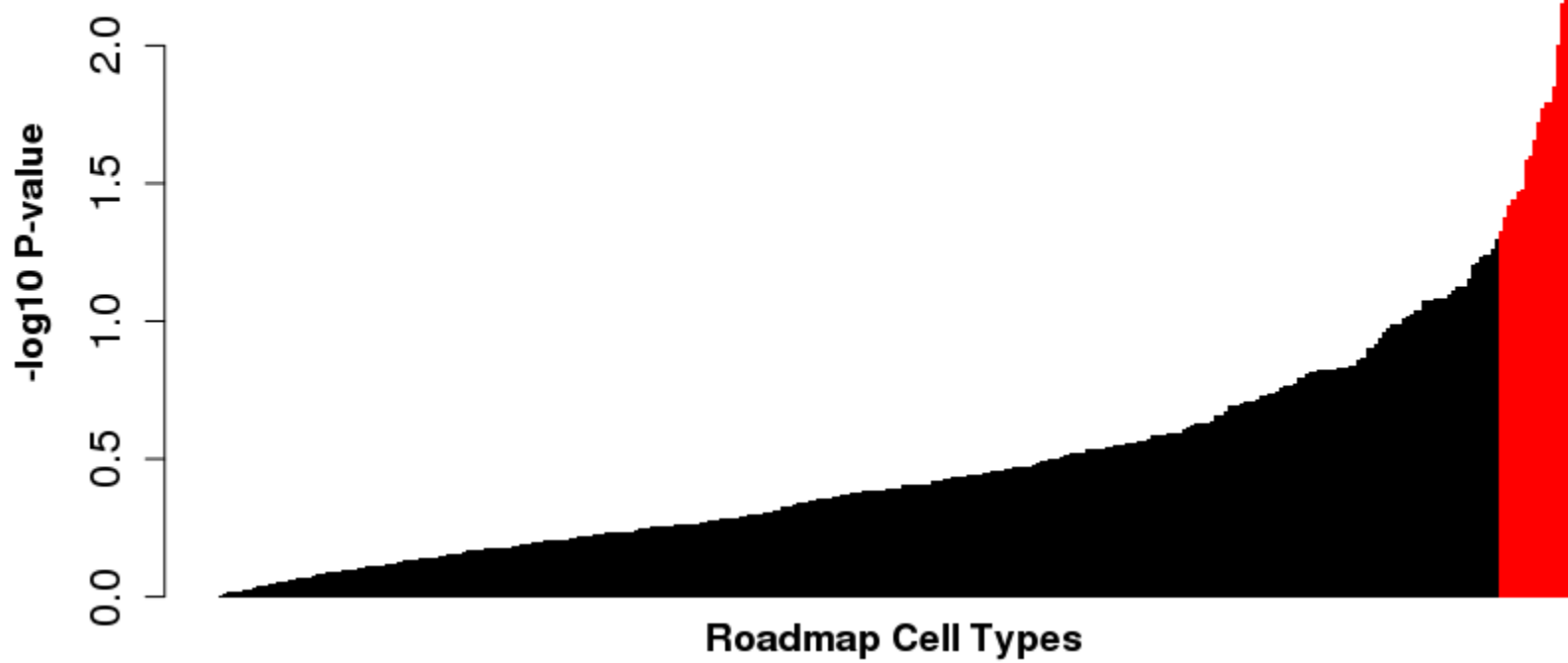
Test significance of overlap:

- For each annotation, assess significance of overlap via local permutations using GoShifter.
    o 17 cell types had significant enrichment of GWAS loci in at least one annotation *(cell-type enrichment P-value < 0.05)*

Prioritizing informative loci:

- For each annotation, overlaps in the lowest 20% of locus score distributions and with a cell type enrichment P-value < 0.05 were considered for further analysis.
    o 5 loci overlapped with DHS peaks in 9 cell types
    o 8 loci overlapped with DHS hotspots in 3 cell types
    o 15 loci overlapped with enhancer marks in 3 cell types
    o 4 loci overlapped with DNaseI footprints in 3 cell types
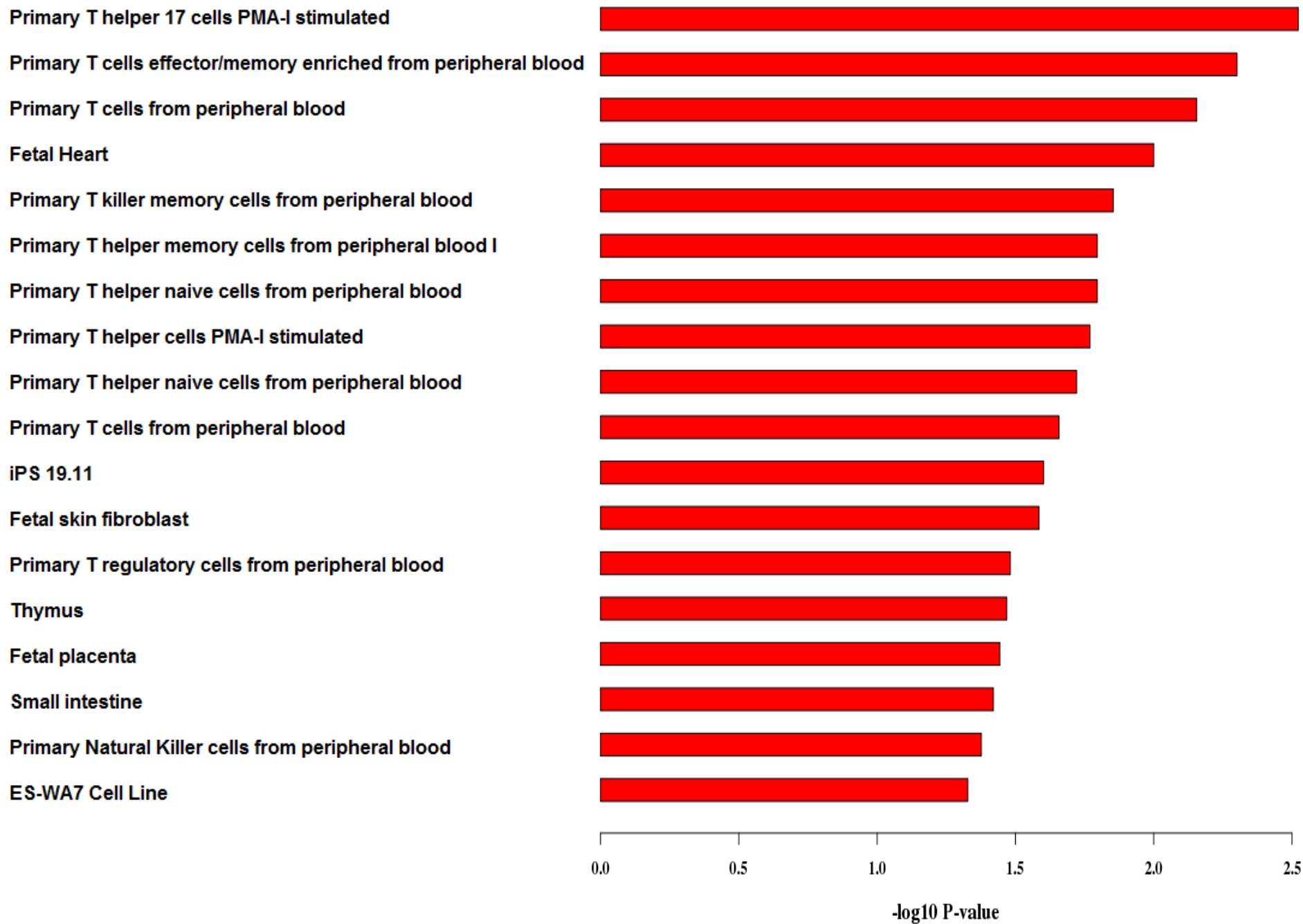   o 21 loci overlapped with at least one annotation in 17 cell types

Figure 4

4A

-log10 P-value

Roadmap Cell Types

4B

Figure 5

ACVR1B SNPs luciferase activity - 16HBE