

PRIMARY RESEARCH

Open Access



Ensemble genomic analysis in human lung tissue identifies novel genes for chronic obstructive pulmonary disease

Jarrett D. Morrow^{1*}, Michael H. Cho^{1,2}, John Platig³, Xiaobo Zhou¹, Dawn L. DeMeo^{1,2}, Weiliang Qiu¹, Bartholome Celli², Nathaniel Marchetti⁴, Gerard J. Criner⁴, Raphael Bueno⁵, George R. Washko², Kimberly Glass¹, John Quackenbush³, Edwin K. Silverman^{1,2} and Craig P. Hersh^{1,2}

Abstract

Background: Genome-wide association studies (GWAS) have identified single nucleotide polymorphisms (SNPs) significantly associated with chronic obstructive pulmonary disease (COPD). However, many genetic variants show suggestive evidence for association but do not meet the strict threshold for genome-wide significance. Integrative analysis of multiple omics datasets has the potential to identify novel genes involved in disease pathogenesis by leveraging these variants in a functional, regulatory context.

Results: We performed expression quantitative trait locus (eQTL) analysis using genome-wide SNP genotyping and gene expression profiling of lung tissue samples from 86 COPD cases and 31 controls, testing for SNPs associated with gene expression levels. These results were integrated with a prior COPD GWAS using an ensemble statistical and network methods approach to identify relevant genes and observe them in the context of overall genetic control of gene expression to highlight co-regulated genes and disease pathways. We identified 250,312 unique SNPs and 4997 genes in the cis(local)-eQTL analysis (5% false discovery rate). The top gene from the integrative analysis was *MAPT*, a gene recently identified in an independent GWAS of lung function. The genes *HNRNPAB* and *PCBP2* with RNA binding activity and the gene *ACVR1B* were identified in network communities with validated disease relevance.

Conclusions: The integration of lung tissue gene expression with genome-wide SNP genotyping and subsequent intersection with prior GWAS and omics studies highlighted candidate genes within COPD loci and in communities harboring known COPD genes. This integration also identified novel disease genes in sub-threshold regions that would otherwise have been missed through GWAS.

Keywords: eQTL, Expression QTL, Integrative genomics, Network medicine, Ensemble methods, Bayesian methods

Background

Chronic obstructive pulmonary disease (COPD) is characterized by progressive airflow obstruction accompanied by chronic inflammation. It is a major cause of morbidity and mortality worldwide [1]. Although environmental exposures such as cigarette smoking are risk factors, a genetic component to susceptibility has been observed [2–5]. Multiple genome-wide association studies (GWAS) have

identified loci associated with COPD susceptibility across various populations [6–9]. However, most of these associations have small effect sizes, so there are likely additional COPD genes to be discovered. Understanding the gene regulatory implications of the significant and sub-genome-wide significant (sub-threshold) GWAS variants in lung tissue may identify genes and loci relevant to COPD for future validation experiments.

Prioritization of previously identified genomic loci enhances the molecular understanding of complex disease [10, 11]. Additionally, sub-threshold genetic loci may play a role in complex diseases [12] such as COPD, as

* Correspondence: jarrett.morrow@channing.harvard.edu

¹Channing Division of Network Medicine, Brigham and Women's Hospital, 181 Longwood Avenue, Boston, MA 02115, USA

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



they likely carry a significant biological signal and may reach significance in later higher powered studies. Increasing the power to identify additional associations often requires a much larger sample size [13], which greatly increases study expense. Integration with omics data can provide insight into the regulatory effects of these variants [12, 14, 15], without increasing sample size. Expression quantitative trait locus (eQTL) analysis tests the association between genetic variants and gene expression and can point to relevant single nucleotide polymorphisms (SNPs) and genes within GWAS loci [15–17] using the observation that trait-associated SNPs are likely to be eQTLs/eSNPs [17] and/or have gene regulatory implications [18].

In this study of genetic control of gene expression, we performed eQTL analysis in lung tissue samples from severe COPD cases and ex-smoker controls and integrated the findings with results from a prior GWAS [8]. We used the Bayesian method Sherlock [19] to identify genes having collective associations within the significant and sub-threshold GWAS SNPs. To observe these genes in the overall context of genetic control of gene expression, we constructed a bipartite network and identified communities [20] harboring the Sherlock-derived genes. We observed that some of these communities contained differentially expressed genes and genes with CpG sites differentially methylated by COPD status. This integration of previous omics studies hones in on the communities demonstrating greater relevance to COPD.

The central hypothesis of this study is that sub-threshold GWAS SNPs, in addition to genome-wide significant SNPs, both influence gene expression and confer disease susceptibility through effects better observed using network and integrative statistical methods. The foundation of this study is the aggregation of the gene expression signals from SNPs identified in prior GWAS, both significant and sub-threshold, using regulatory evidence via an ensemble Bayesian and network approach. This integrative method extracts the additional genetic and genomic signals contained in the sub-threshold SNPs by combining evidence across genotyping, gene expression and DNA methylation datasets and highlights novel genes and loci within regions that may not have been identified through GWAS. This motivates hypotheses regarding the biological role of these findings in disease and informs selection of targets for further functional investigations.

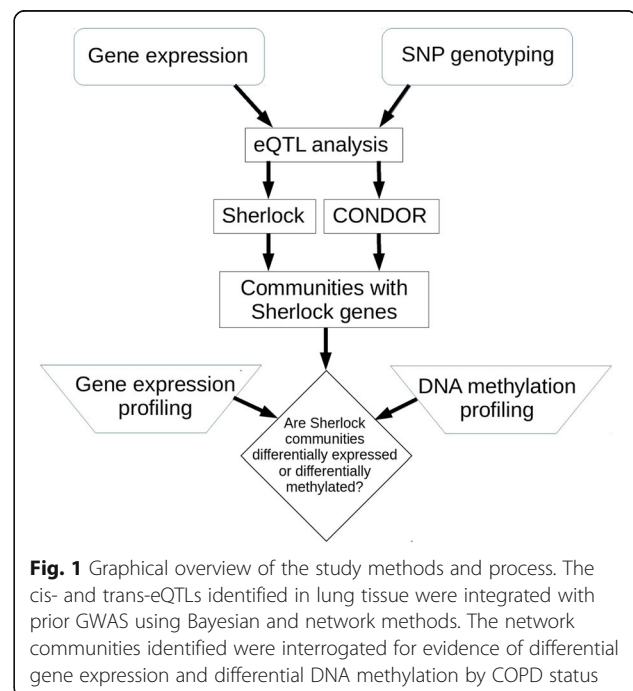
Results

Gene expression data were available for lung tissue samples from 86 severe COPD cases (mean FEV₁ 26.4% predicted) and 31 controls with normal spirometry, all Caucasians (Additional file 1: Table S1). There were no significant differences between cases and controls by sex

or age. The cases had higher lifetime smoking intensity in pack-years and quit smoking on average 8.7 fewer years in the past ($p = 0.0006$). We identified eQTLs using the gene expression and imputed genotyping data and integrated them with prior GWAS and omics studies using an ensemble approach of statistical and network methods (Fig. 1).

Using the lung tissue gene expression profiling and imputed genotyping data from the cases and controls, we performed cis- and trans-eQTL analysis (see the “Methods” section). We identified 347,251 significant cis-eQTL results (FDR < 5%) out of 55,550,191 total tests. Within these results, there were 250,312 unique cis-eQTL SNPs (eSNPs) and 5878 unique eQTL genes (eGenes, 4997 gene symbols) (Additional file 1: Table S2). This represents 4.2% of the SNPs and 24% of the expression probes tested. The trans results contain 8519 significant results (FDR < 5%), out of 146,665,850,054 total tests, with 6930 unique eSNPs and 451 unique eGenes (434 gene symbols) (Additional file 1: Table S3).

We intersected the significant cis-eQTL results with the GWAS at a suggestive level of significance ($p < 10^{-4}$) [8] and observed that 292 of these 1847 significant and sub-threshold GWAS SNPs were eSNPs (4.3 fold enrichment, hypergeometric p value < 0.00001). The top intersection results are shown in (Additional file 1: Table S4). Regional genomic plots of significant cis-eQTLs (FDR < 5%) for 5 of these 13 loci highlight the regulatory information for the top eSNPs and SNPs in linkage disequilibrium (LD) (Additional file 1: Figures S1–S5). Two of the eSNPs from (Additional file 1: Table S4) are located within the



associated eGene (rs1504550-*IREB2* and rs2252518-*ACVR1B*; Additional file 1: Figures S1 and S2). Two others (rs12461383-*C19orf54* and rs11852372-*CHRNA5*; Additional file 1: Figures S3 and S4) are in promoter flanking and transcription factor binding regions within DNase hypersensitivity (DHS) sites. The last eSNP (rs151321-*SULT1A2*; Additional file 1: Figure S5) is in LD (shaded in red) with several SNPs located in regulatory regions. To observe overall genetic control of gene expression in a disease context, we intersected all cis-eQTL results with the nominally significant GWAS SNPs ($p < 0.05$) [8] and plotted the p values from the two sets (Fig. 2). Each point in the plot represents an eQTL result (eSNP-eGene pair); prior COPD gene expression profiling results [21] are overlaid in color. We observed that eQTLs with COPD GWAS associations are generally not enriched for differentially expressed genes; regions with sub-threshold GWAS p values ($p < 10^{-4}$) and significant eQTL p values lack differentially expressed genes (FDR < 5%). Therefore, we used additional statistical and network methods to extract the signal in these results, given this complex relationship between the disease and the genetic control of gene expression.

We integrated the nominally significant cis-eQTLs ($p < 10^{-3}$) and trans-eQTLs ($p < 10^{-6}$) with prior GWAS using the Bayesian method Sherlock [19], seeking genes with collective associations across the significant and sub-threshold GWAS results. The 438,536 SNPs common to the eQTL, GWAS, and GWAS permutation data were the basis for this integrative analysis. A total of 50 Sherlock results had p values < 10^{-3} (Table 1, Additional file 2: Table S5). This p value threshold corresponds to a LBF

(logarithm of Bayes factor) sum of 1.94. Of the 50 genes identified, 13 were previously found in the intersection between cis-eQTLs and GWAS ($p < 10^{-4}$) results. Several genes have been identified in previous COPD GWAS studies. We repeated the Sherlock analysis using the eQTL results from GTEx V7 (using the same p value thresholds) and observed the results for these top 50 genes (Table 1). We further sought to place our 50 Sherlock-derived genes in the context of overall genetic control of gene expression using network methods, since co-regulated genes may have shared function. This process has the potential to reveal additional COPD genes of interest.

We constructed a bipartite network using the cis- and trans-eQTLs with p value thresholds identical to those for Sherlock (cis: $p < 10^{-3}$ and trans: $p < 10^{-6}$). After all filtering steps (see the “Methods” section), 171,490 eSNPs and 11,348 eGenes were used in the construction of the network. The power-law nature of the degree distribution for this network is heavy-tailed (Additional file 1: Figure. S6) and similar to that seen in other bipartite eQTL networks [20], suggesting a scale-free structure characterized by the presence of hubs. We identified 250 communities within this network and focused on the 14 that contain Sherlock-derived genes (Table 2, Additional file 1: Table S6). We also examined two communities that contained putative interactors (*HMGB1* and *CD79A*) of genes near GWAS loci from our previous study [21]. These differentially expressed interactors were identified using gene expression profiling in lung tissue and in vitro, in vivo, and in silico datasets that identified genes with evidence of

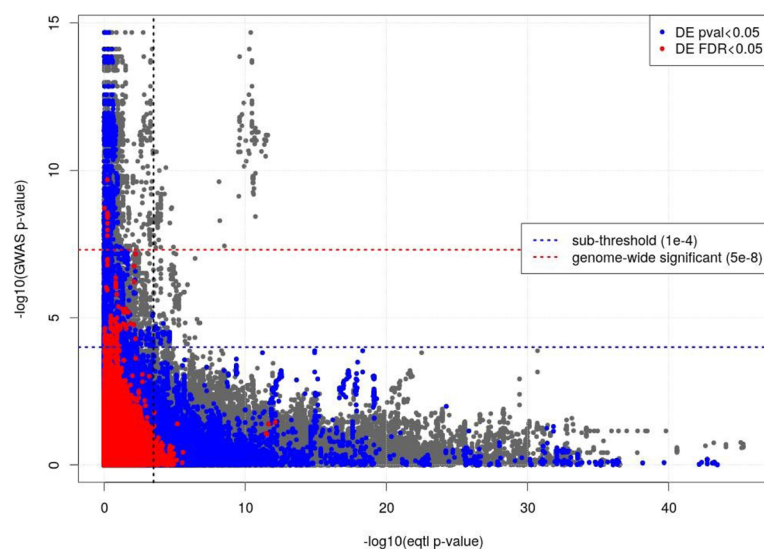


Fig. 2 Plot of COPD GWAS p values vs. the cis-eQTL p values. Each point in the plot represents a cis-eQTL result with an rsID found in the prior GWAS. GWAS p values (y axis) are plotted against the expression QTL p values (x axis). A vertical dotted line indicates the threshold of significance (FDR < 5%) for the eQTL. Horizontal lines delineate genome-wide significant (red) and sub-threshold (blue) GWAS p values. The significant (red; FDR < 5%) and nominally significant (blue; $p < 0.05$) eGenes from gene expression profiling in COPD lung tissue are highlighted

Table 1 COPD genes identified in the Sherlock analysis

Gene symbol	Total LBF score	Sherlock <i>p</i> value	Differentially expressed probe (<i>p</i> < 0.05)	Differentially methylated site (<i>p</i> < 0.05, effect > 5%)	GTEX V7 (cis-only, LBF score)	GTEX V7 (cis-only, <i>p</i> value)
<i>MAPT</i>	7.65	6.91E-07	Yes	No	7.31	6.41E-07
<i>LRRC37A4</i>	7.46	6.91E-07	No	No	2.13	1.8E-3
<i>C17orf69</i>	7.38	6.91E-07	Yes	No	2.39	1.3E-3
<i>IREB2</i> *	6.40	6.91E-07	No	No	4.37	5.89E-05
<i>C19orf54</i> *	5.45	5.53E-06	No	No	5.01	1.79E-05
<i>ACVR1B</i> *	5.40	5.53E-06	No	Yes	4.77	2.82E-05
<i>EIF3CL</i> *	4.45	1.94E-05	No	No	N/A	N/A
<i>TUFM</i> *	4.29	2.49E-05	Yes	No	5.31	8.97E-06
<i>FAM13A</i>	4.09	3.60E-05	No	No	5.69	3.84E-06
<i>PCBP2</i>	3.97	4.43E-05	No	No	N/A	N/A
<i>CYP2B7</i> *	3.87	5.67E-05	No	No	N/A	N/A
<i>SULT1A1</i> *	3.81	6.08E-05	Yes	No	5.38	8.97E-06
<i>SULT1A2</i> *	3.80	6.08E-05	Yes	Yes	5.01	1.79E-05
<i>TIGD2</i>	3.58	7.88E-05	No	No	2.48	1.10E-03
<i>CHRNA5</i> *	3.37	1.05E-04	No	No	5.27	1.15E-05
<i>BZRAP1</i>	3.27	1.20E-04	No	Yes	-0.20	7.28E-01
<i>GPX8</i> *	3.25	1.20E-04	No	No	N/A	N/A
<i>TEKT3</i>	3.19	1.31E-04	No	No	0.00	1.59E-01
<i>SNRPB</i>	3.06	1.63E-04	No	No	0.00	1.61E-01
<i>ZNF652</i>	3.03	1.69E-04	No	No	-0.02	2.69E-01
<i>AHSA2</i> *	2.86	2.14E-04	No	No	2.22	1.63E-03
<i>CDH23</i>	2.82	2.21E-04	No	Yes	0.01	1.38E-01
<i>NOP2</i>	2.69	2.99E-04	No	No	N/A	N/A
<i>AASDH</i>	2.68	3.06E-04	No	No	2.45	1.14E-03
<i>DAGLA</i>	2.68	3.06E-04	No	No	N/A	N/A
<i>IFI27L2</i> *	2.65	3.24E-04	No	No	2.90	5.98E-04
<i>APIP</i>	2.60	3.48E-04	No	No	2.52	1.03E-03
<i>AXIN2</i>	2.59	3.55E-04	No	No	-0.04	4.10E-01
<i>WDR47</i>	2.49	4.20E-04	Yes	No	N/A	N/A
<i>C4orf33</i>	2.41	4.74E-04	No	No	1.68	3.71E-03
<i>HNRNPAB</i>	2.34	5.23E-04	No	No	N/A	N/A
<i>GFPT1</i>	2.33	5.24E-04	Yes	No	-0.02	2.04E-01
<i>LOC644172</i>	2.32	5.30E-04	No	No	N/A	N/A
<i>SNORD25</i>	2.25	5.86E-04	No	No	N/A	N/A
<i>PPAT</i>	2.23	6.06E-04	Yes	No	-0.03	3.22E-01
<i>FBRSL1</i>	2.23	6.08E-04	No	No	-0.09	5.90E-01
<i>FSTL5</i>	2.22	6.14E-04	No	No	N/A	N/A
<i>SMG6</i>	2.18	6.44E-04	No	Yes	-0.01	1.85E-01
<i>CHIAP2</i>	2.09	7.58E-04	No	No	N/A	N/A
<i>RPL23A</i>	2.07	7.77E-04	Yes	No	N/A	N/A
<i>C2orf74</i> *	2.06	7.85E-04	No	No	N/A	N/A
<i>CTSH</i>	2.04	8.17E-04	No	No	1.40	5.59E-03
<i>UBE2J1</i>	2.03	8.27E-04	Yes	No	N/A	N/A

Table 1 COPD genes identified in the Sherlock analysis (*Continued*)

Gene symbol	Total LBF score	Sherlock <i>p</i> value	Differentially expressed probe (<i>p</i> < 0.05)	Differentially methylated site (<i>p</i> < 0.05, effect > 5%)	GTEx V7 (cis-only, LBF score)	GTEx V7 (cis-only, <i>p</i> value)
<i>AEN</i>	2.01	8.50E-04	No	No	0.36	3.61E-02
<i>CUL1</i>	2.00	8.88E-04	No	No	-0.02	3.02E-01
<i>DSP</i>	2.00	8.91E-04	No	No	1.51	4.77E-03
<i>MYCN</i>	1.97	9.32E-04	No	No	-0.04	4.13E-01
<i>TRIM4</i>	1.96	9.43E-04	Yes	No	1.57	4.37E-03
<i>ZNF57</i>	1.94	9.64E-04	No	No	-0.02	2.16E-01
<i>NARS2</i>	1.94	9.74E-04	No	No	2.89	6.10E-04

LBF logarithm of Bayes factor

*Gene identified in the cis-eQTL-GWAS intersection in (Additional file 1: Table S4)

interaction with one of the three genes (*HHIP*, *FAM13A*, and *IREB2*) implicated by in-depth functional studies at COPD GWAS loci.

To validate the disease relevance of the communities, we calculated the differential expression and differential DNA methylation meta-analysis *p* values (see the “Methods” section) for these 16 communities. Seven communities were validated based on nominally significant (meta-*p* < 0.05) differential expression and differential methylation results (Table 2). These communities contain the Sherlock-derived genes *CDH23*, *CHRNA5*, *HNRNPAB*, *IREB2*, *PCBP2*, *ZNF652*, *ACVR1B*, and *RPL23A* (Figs. 3, 4, and 5 and Additional file 1: Figures S7–S8) or the interactors *HMGB1* and *CD79A*

(Additional file 1: Figures S9–S10). There was significant pathway enrichment (FDR *q* value < 0.05) using ConsensusPathDB [22] for two validated communities (ID = 222:*ACVR1B* and ID = 135:*CD79A*) in Table 2 (Additional file 1: Table S7), highlighting cGMP-PKG signaling, focal adhesion, and actin and immune system-related pathways. Six of the nine remaining communities, which were lacking joint evidence, had either nominally significant differential expression or differential methylation.

Discussion

Although many genome-wide significant loci from COPD GWAS were not eSNPs in lung tissue, we found

Table 2 CONDOR communities that contain Sherlock-derived genes or putative COPD GWAS gene interactors

Community ID	Sherlock or interactor gene(s)	Total SNPs	Sub-threshold SNPs	Total genes	Number of differentially expressed genes	Number of differentially methylated genes	Expression meta- <i>p</i> value	Methylation meta- <i>p</i> value
98**	<i>HMGB1</i>	143	0	4	3	1	2.26E-19	0.0028
113**	<i>CDH23</i>	489	0	12	2	1	0.0184	4.95E-05
135**	<i>CD79A</i>	1959	0	162	29	13	9.27E-11	4.27E-33
202**	<i>CHRNA5</i> , <i>HNRNPAB</i> , <i>IREB2</i> , <i>PCBP2</i>	293	57	17	4	1	0.0032	0.0026
218**	<i>ZNF652</i>	410	0	47	8	3	0.0017	4.85E-06
222**	<i>ACVR1B</i>	790	0	67	12	9	0.0003	8.23E-16
223**	<i>RPL23A</i>	509	4	32	10	6	3.26E-06	8.82E-18
20*	<i>WDR47</i>	476	0	6	3	0	0.0021	–
78*	<i>CHIAP2</i>	631	0	18	3	1	0.0907	0.0019
131	<i>AHSA2 C2orf74</i>	599	8	4	1	0	0.0634	–
161*	<i>SMG6</i>	633	0	18	1	2	0.3723	3.38E-09
179*	<i>DSP</i>	68	0	7	3	0	0.0060	–
181*	<i>FSTL5</i>	475	0	23	4	1	0.0503	0.0069
187*	<i>SNRPB</i>	178	0	14	1	2	0.5669	0.0001
210	<i>CTSH</i>	439	5	12	0	0	0.5741	–
249	<i>TRIM4</i>	555	0	11	3	0	0.0957	–

*Communities with either significant differential expression or differential methylation (*p* < 0.05)**Communities with both significant differential expression and differential methylation (*p* < 0.05)

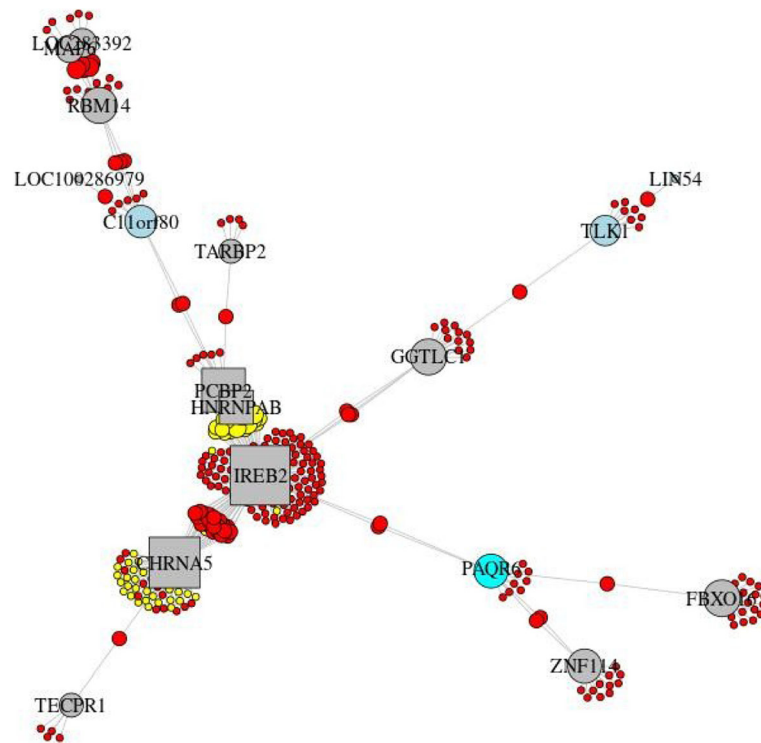


Fig. 3 Community 202 from CONDOR analysis that contains the Sherlock-derived genes *CHRNA5*, *HNRNPAB*, *IREB2*, and *PCBP2*. Community genes are listed in (Additional file 1: Table S6). (Red = SNP, yellow = SNP with GWAS $p < 10^{-4}$, square = Sherlock gene, gray = gene, green = gene with differentially methylated site ($p < 0.05$ and effect $> 5\%$), light blue = gene with differentially expressed probe ($p < 0.05$), and cyan = gene with differentially methylated site and differentially expressed probe)

that the sub-threshold GWAS findings are enriched in eSNPs. We also observed that eQTLs with GWAS associations did not have eGenes significantly differentially expressed in severe COPD cases vs controls, demonstrating the complex nature of genetic control of gene expression. We employed an ensemble approach involving Bayesian and network methods to investigate these eQTL results, which yielded 16 relevant bipartite communities. Based on the differential gene expression and/or differential DNA methylation of all of the genes or CpG sites within each community, we validated the disease relevance for 13 of these communities, highlighting potential COPD genes within the significant and sub-threshold GWAS results.

One of the seven communities (community 202) which was validated by both differential expression and DNA methylation contains two previously identified COPD GWAS genes located in a genome-wide significant region: *IREB2* (iron responsive element binding protein 2) and *CHRNA5* (cholinergic receptor nicotinic alpha 5 subunit) [23–25]. The product of *IREB2* is known to interact with mRNA to influence translation or degradation. Two other Sherlock-derived genes in community 202 also have putative RNA binding activity, *PCBP2* (poly(rC) binding protein 2) and *HNRNPAB* (heterogeneous nuclear

ribonucleoprotein A/B). *PCBP2* plays a role in mRNA stability, and it has been suggested that deregulation of this stability may contribute to COPD pathogenesis [26]. A recent study of breast cancer highlighted the regulatory role of RNA binding by *PCBP3* (paralog of *PCBP1* along with *PCBP2*) on mRNA stability and induction of epithelial-mesenchymal transition (EMT) [27]. Additionally, *HNRNPAB* has been shown to induce EMT [28], a potential contributor to airway disease [29, 30]. Together, this suggests a role for this community in COPD pathogenesis. Community 222 contains the Sherlock-derived gene *ACVR1B* (activin A receptor type 1B), a gene identified in a previous eQTL study in blood and sputum in COPD [31]. *ACVR1B* was a sub-threshold finding in a GWAS of lung function in COPD [32] and was identified in our intersection of eQTLs with the sub-threshold GWAS of case-control status. The genes in community 222 were enriched for cGMP-PKG signaling, bacterial invasion of epithelial cells, and focal adhesion pathways [33], with possible relevance to COPD pathogenesis and exacerbations. Community 113 includes the Sherlock-derived gene *CDH23* (cadherin-related 23), involved in cell-cell adhesion and perhaps EMT as a calcium-dependent cell adhesion molecule [34]. This gene was contained within sub-threshold loci in GWAS of lung

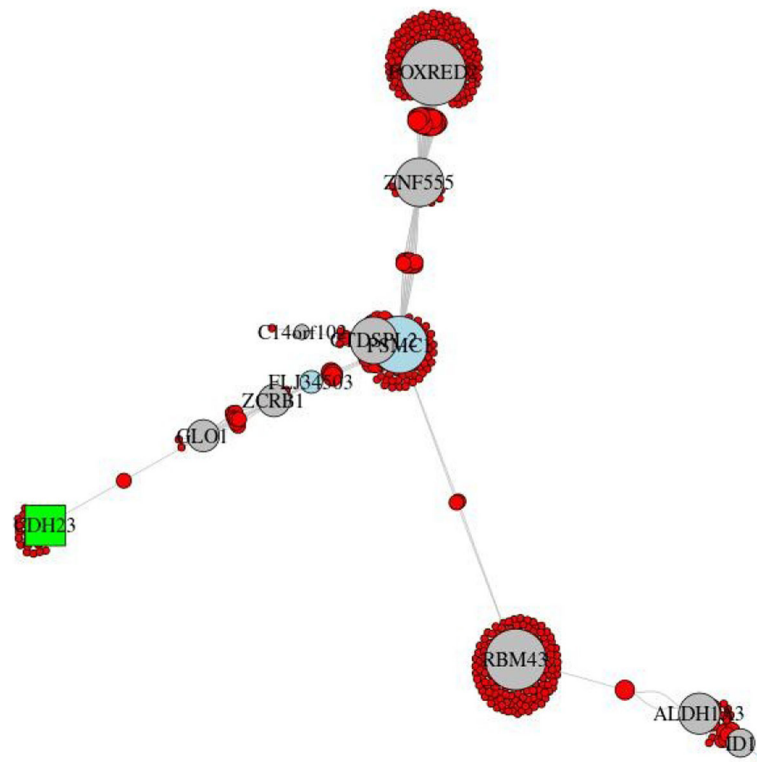


Fig. 5 Community 113 from CONDOR analysis that contains the Sherlock-derived gene *CDH23*. The central genes *PSMC1* and *CTDSPL2* partially overlap and are obstructed in the figure. Community genes are listed in (Additional file 1: Table S6). (Red = SNP, yellow = SNP with GWAS $p < 10^{-4}$, square = Sherlock gene, gray = gene, green = gene with differentially methylated site ($p < 0.05$ and effect $> 5\%$), light blue = gene with differentially expressed probe ($p < 0.05$), cyan = gene with differentially methylated site and differentially expressed probe)

[7]. Further efforts will be required to determine which of these genes is relevant for COPD pathogenesis.

In a previous gene expression profiling study [21], we identified several putative interactors of three known COPD GWAS genes (*HHIP*, *FAM13A*, and *IREB2*). Communities harboring two of these interactors were identified in the current study. Both community 98 with *HMGB1* (high-mobility group box 1) and community 135 with *CD79A* (*CD79a* molecule) had evidence of differential expression and differential methylation. Additionally, there may also be a role for *HMGB1* in the development of EMT in airway epithelial cells [43].

Our study has several limitations. The omics datasets in this study were generated using homogenized lung tissue, so we could not determine the cellular specificity of the eQTLs, differential expression, and differential methylation. Studies in single lung cell types will address this cellular heterogeneity and provide validation of the findings. Our study focused on severe COPD and was enriched for subjects with emphysema and therefore may miss genes relevant for milder disease or other COPD phenotypes such as airway disease. Lastly, future integrative studies using these datasets will explore in more detail the gene regulatory impact of DNA methylation in lung tissue.

This study of the genetic control of gene expression in human lung has revealed potential genes of interest co-regulated with known COPD genes. The ensemble approach using statistical and network methods also pointed to specific genes in complex genomic regions found through prior GWAS, and genes within loci that would not meet strict thresholds for genome-wide significance, thereby extracting additional information from these results and supporting our hypothesis regarding the relevance of sub-threshold SNPs. We integrated three omics datasets, providing regulatory characterization of significant and sub-threshold GWAS variants, and highlighted genes for further functional investigation that may be involved in COPD pathogenesis. These genes would otherwise not have been identified through GWAS and could potentially meet the strict threshold for statistical significance in larger GWAS in COPD.

Methods

Study subjects

We collected lung tissue samples from former smokers undergoing thoracic surgery for lung transplantation, lung volume reduction surgery, or lung nodule resection

at three medical centers; all subjects quit smoking at least 1 month prior to surgery [21, 42]. Distant normal tissue was sourced from lung nodule resection samples. The COPD subjects had severe airflow obstruction, with GOLD grade 3–4 spirometry (FEV1% predicted < 50% and FEV1/FVC < 0.7) and the controls had normal spirometry (FEV1% predicted \geq 80% and FEV1/FVC \geq 0.7). IRB approval was obtained at the three centers (Brigham and Women's Hospital, Boston, MA; St. Elizabeth's Hospital; Boston, MA; and Temple University Hospital, Philadelphia, PA), and subjects provided written informed consent.

eQTL analysis

Microarray expression profiling was available for 111 cases and 40 controls [21] (GEO Series GSE76925). Of the 32,831 expression probes, 24,495 had genomic location information and were retained for integration with genotyping data. Genome-wide SNP genotyping data was obtained from lung tissue DNA using the HumanOmni2.5Exome-8 V1.0 BeadChip (Illumina, Inc., San Diego, CA) as previously described [21]. After quality control, genotypes were phased using SHAPEIT2 [44] and imputed using IMPUTE2 [45, 46] with the 1000 Genomes Phase3 V5 reference. The analyses were performed using only data from the Caucasian subjects. Data for markers with an imputation info metric > 0.5 and minor allele frequency > 5% were retained for the 117 Caucasian subjects that had both high-quality genotyping and gene expression data (86 cases and 31 controls; Additional file 1: Table S8). To account for population stratification, two principal components (PC) based on the Tracy-Widom statistic for the Caucasian population were retained [47]. Both cis- and trans-eQTL analyses were performed using the R/Bioconductor package Matrix eQTL (version 2.1.1) [48]. A total window size of 1 million bases was used for the cis analysis (500 kb upstream and downstream from the gene); trans analysis was performed genome-wide. This analysis identifies associations between genotype dosage and gene expression levels, adjusting for age, sex, pack-years of smoking, and the two ancestry PCs. An iterative method was used to determine the number of PCs for the matrix of expression values to add as covariates to mitigate batch effects [21]; 13 PCs were included in the eQTL analyses. An eQTL association result consists of an eGene (microarray expression probe) and eSNP pair.

Integration using Sherlock

The Sherlock method performs genetic signature matching using a Bayesian statistical framework [19]. The hypothesis is that SNPs associated with expression of disease-relevant genes are also likely to influence disease risk and be identified through GWAS. Using Sherlock, we integrated the cis- and trans-eQTLs with all results from a published COPD GWAS [8]. Sherlock provides a

total score for each gene, along with the score for each of the individual eQTL contributions. This total score is the sum of the LBFs (logarithm of Bayes factor) for each of these contributions. For interpretation of individual results, a value of 4.0 is typically required for significance. To output a p value, we created permuted GWAS results with similar linkage disequilibrium structure to the GWAS using the set of 379 EUR genotypes available in 1000 Genomes Phase1 V3 [49]. Specifically, we randomly permuted the case-control phenotypes 50 times as recommended in the Sherlock method (190 cases and 189 controls) and applied Plink2 [50] to calculate association p values for each iteration and used these results as inputs for Sherlock. Only overlapping SNPs (loci with rsIDs) present across the eQTL, GWAS, and permutation results were included in the analysis; minor allele frequencies for these markers were obtained from 1000 Genomes data. In the ensemble analysis, we applied a p value threshold of 10^{-3} to select a more significant set of Sherlock-derived genes for downstream analysis. We performed a replication of the Sherlock analysis using the GTEx V7 lung tissue eQTL results [51]. Only the GTEx markers found across the COPD GWAS [8] and our permutation results were included in the analysis. The GTEx project produced cis-eQTL results using a window of 1 million bases upstream and downstream. To align this Sherlock input with our study, we labeled eSNPs located 500 kb to 1 million bases from the gene transcription start site as trans-eQTLs.

Network construction

A bipartite network was constructed using the cis- and trans-eQTLs. Network nodes are eGenes represented by their gene symbol annotation and eSNPs represented by their rsIDs. Edges only connect eSNPs to eGenes; no edges are present between pairs of eSNPs or pairs of eGenes. Only eSNPs represented in the GWAS were included in the network. Cis- or trans-eQTLs with only a single edge between an eSNP and eGene were excluded, since they did not create additional connections in the network. We identified communities within this bipartite network using the R package CONDOR [20] and visualized them using the R package igraph [52], with the Fruchterman-Reingold algorithm. A differential expression meta-analysis p value was computed for each community of interest. Specifically, the differential expression p values from prior expression profiling [21] for each expression probe annotated to genes in the community were combined using Fisher's method via the R package metap. For differential DNA methylation, we used a similar approach based on prior methylation profiling results [42] for CpG sites annotated to genes in the community. In order to focus

on CpG sites more likely to be biologically relevant, we required that the mean difference in methylation between cases and controls be greater than 5%.

Regulatory annotation

The R package Sushi [53] was used with gene annotation and regulatory information from Ensemble BioMart [54] (CTCF Binding Site, TF binding site, Open chromatin, Promoter and Enhancer information produced from ENCODE, Roadmap Epigenomics, and Blueprint projects [55] for GRCh37) and DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3) from the UCSC database [56] (GRCh37). Linkage disequilibrium information in these regional plots was produced using correlation r^2 values for SNP pairs from PLINK, using genotyping data from 1000 Genomes Phase3 V5.

Additional files

Additional file 1: Supplemental Data. Supplemental supporting figures (Figures S1–S10) and tables (Tables S1–S8). (PDF 4193 kb)

Additional file 2: Supplemental Table S5. Table containing Sherlock results. (PDF 68 kb)

Acknowledgements

Funding: NIH grants K25 HL136846, P01 HL105339, R01 HL111759, R01 HL125583, R01 HL130512, P01 HL132825, R01 HL089856, R01 HL089897, and K25 HL133599 funded this study.

We thank Drs. Amund Gulsvik, Per Bakke, Augusto Litonjua, Pantel Vokonas, Ruth Tal-Singer, and the GenKOLS, NETT/NAS, ECLIPSE, and COPDGene studies for use of the GWAS meta-analysis data.

R01 HL089856 and R01 HL089897 are from the COPDGene study. NCT00608764 was funded by the National Institutes of Health is also supported by the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer Ingelheim, Novartis, Pfizer, Siemens, GSK, and Sunovion. The National Emphysema Treatment Trial was supported by the NHLBI N01HR76101, N01HR76102, N01HR76103, N01HR76104, N01HR76105, N01HR76106, N01HR76107, N01HR76108, N01HR76109, N01HR76110, N01HR76111, N01HR76112, N01HR76113, N01HR76114, N01HR76115, N01HR76116, N01HR76118, and N01HR76119; the Centers for Medicare and Medicaid Services; and the Agency for Healthcare Research and Quality. The Normative Aging Study is supported by the Cooperative Studies Program/ERIC of the US Department of Veterans Affairs and is a component of the Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC). The Norway GenKOLS study (Genetics of Chronic Obstructive Lung Disease, GSK code RES11080), the ECLIPSE study (NCT00292552; GSK code SCO104960), and the ICGN study were funded by GlaxoSmithKline.

Availability of data and materials

The dataset supporting the conclusions of this article is available in the GEO repository (GSE76925): <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE76925>

Authors' contributions

JM was responsible for the concept and design, analysis and interpretation of the data, manuscript preparation, and approval of final manuscript. MC contributed to the acquisition of the data, analysis and interpretation of the data, manuscript preparation, and approval of the final manuscript. JP contributed to the analysis and interpretation of data, statistical support, and approval of the final manuscript. XZ took part in the analysis and interpretation of the data, manuscript preparation, and approval of final manuscript. DD contributed to the acquisition of data, manuscript preparation, and approval of the final manuscript. WQ contributed to the analysis and interpretation of data,

statistical support, and approval of the final manuscript. BC, NM, GC, and RB contributed to the acquisition of data, approval of final manuscript. GW took part in the acquisition of the data, analysis and interpretation of the data, and approval of the final manuscript. KG contributed to the analysis and interpretation of the data, statistical support, manuscript preparation, and approval of the final manuscript. JQ provided statistical support and approval of the final manuscript. ES contributed to the concept and design, acquisition of the data, manuscript preparation, and approval of the final manuscript. CH was responsible for the concept and design, acquisition of the data, analysis and interpretation of data, manuscript preparation, and approval of the final manuscript.

Ethics approval and consent to participate

Subjects provided written informed consent for the use of excess lung tissue for research. IRB approval was obtained at Partners Healthcare (parent company of Brigham and Women's Hospital), Temple University, and St. Elizabeth's Hospital. The methods for lung tissue research were carried out in accordance with the relevant guidelines.

Competing interests

Drs. Morrow, Platig, Zhou, Qiu, Marchetti, Criner, Celli, Glass, Quackenbush report no competing interests related to this manuscript.

Dr. Cho has received compensation from GSK.

Dr. DeMeo has received compensation from Novartis.

Dr. Bueno has received compensation from Myriad Genetics, Inc., Siemens, Verastem, Inc., Genentech, Inc., Gritstone Oncology, Inc., HTG Molecular Diagnostics, Inc., Neil Leifer, Esq., Morrison Mahoney, David Weiss, LLC, Balick & Balick, LLC, Hartford Hospital, Cleveland Clinic/Conference, Aspen Lung Conference, Case Western Reserve, North Shore University, Castle Bioscience, Novartis Institutes for Biomedical Research, AstraZeneca, imCORE/Roche, Arthur Tuverson, LLC, Ferraro Law Firm, Rice Dolan & Kershaw, Satterly & Kelly, LLC, Exosome, Inc.

Dr. Washko has received compensation from Boehringer Ingelheim, GlaxoSmithKline, Janssen Pharmaceuticals, BTG Interventional Medicine, Regeneron, Quantitative Imaging Solutions and ModoSpira, and his Spouse works for Biogen.

Dr. Silverman has received compensation from COPD Foundation, GlaxoSmithKline, Merck and Novartis,

Dr. Hersh has received consulting fees from AstraZeneca, Concert Pharmaceuticals, Mylan, and grant support from Boehringer Ingelheim.

Publisher's Note

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

Author details

¹Channing Division of Network Medicine, Brigham and Women's Hospital, 181 Longwood Avenue, Boston, MA 02115, USA. ²Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA. ³Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA 02115, USA. ⁴Division of Pulmonary and Critical Care Medicine, Temple University, Philadelphia, PA 19140, USA. ⁵Division of Thoracic Surgery, Brigham and Women's Hospital, Boston, MA 02115, USA.

Received: 20 September 2017 Accepted: 2 January 2018

Published online: 15 January 2018

References

- Vestbo J, Hurd SS, Agustí AG, Jones PW, Vogelmeier C, Anzueto A, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *Am J Respir Crit Care Med.* 2013;187:347–65.
- Hersh CP, Hokanson JE, Lynch DA, Washko GR, Make BJ, Crapo JD, et al. Family history is a risk factor for copd. *Chest.* 2011;140:343–50.
- McCloskey S, Patel B, Hinchliffe S, Reid E, Wareham N, Lomas D. Siblings of patients with severe chronic obstructive pulmonary disease have a significant risk of airflow obstruction. *Am J Respir Crit Care Med.* 2001;164:1419–24.
- Silverman EK. Genetics of chronic obstructive pulmonary disease. In: Organizer DC, Goode JA, editors. *Chronic Obstr. Pulm. Dis. Pathog. Treat.* John Wiley & Sons, Ltd; 2000. p. 45–64.

5. Silverman EK, Chapman HA, Drazen JM, Weiss ST, Rosner B, Campbell EJ, et al. Genetic epidemiology of severe, early-onset chronic obstructive pulmonary disease. Risk to relatives for airflow obstruction and chronic bronchitis. *Am J Respir Crit Care Med*. 1998;157:1770–8.
6. Cho MH, Boutaoui N, Klanderman BJ, Sylvia JS, Ziniti JP, Hersh CP, et al. Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nat Genet*. 2010;42:200–2.
7. Cho MH, Castaldi PJ, Wan ES, Siedlinski M, Hersh CP, Demeo DL, et al. A genome-wide association study of COPD identifies a susceptibility locus on chromosome 19q13. *Hum Mol Genet*. 2012;21:947–57.
8. Cho MH, McDonald M-LN, Zhou X, Mattheisen M, Castaldi PJ, Hersh CP, et al. Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. *Lancet Respir Med*. 2014;2:214–25.
9. Hobbs BD, de Jong K, Lamontagne M, Bossé Y, Shrine N, Artigas MS, et al. Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. *Nat Genet*. 2017;49:426–32.
10. Björnsson H. An integrated epigenetic and genetic approach to common human disease. *Trends Genet*. 2004;20:350–8.
11. Feinberg AP. Epigenomics reveals a functional genome anatomy and a new approach to common disease. *Nat Biotechnol*. 2010;28:1049–52.
12. Wang X, Tucker NR, Rizki G, Mills R, Krijger PH, de Wit E, et al. Discovery and validation of sub-threshold genome-wide association study loci using epigenomic signatures. *elife*. 2016;5:e10557.
13. Bush WS, Moore JH. Chapter 11: genome-wide association studies. Lewitter F, Kann M, editors. *PLoS Comput Biol*. 2012;8:e1002822.
14. Civelek M, Lusk AJ. Systems genetics approaches to understand complex traits. *Nat Rev Genet*. 2013;15:34–48.
15. Li L, Kabesch M, Bouzigon E, Demenais F, Farrall M, Moffatt MF, et al. Using eQTL weights to improve power for genome-wide association studies: a genetic study of childhood asthma. *Front Genet*. 2013;4:103. doi:10.3389/fgenet.2013.00103.
16. Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, Barroso I, et al. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. Gibson G, editor. *PLoS Genet*. 2010;6:e1000895.
17. 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:e1000888.
18. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 2012;337:1190–5.
19. He X, Fuller CK, Song Y, Meng Q, Zhang B, Yang X, et al. Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. *Am J Hum Genet*. 2013;92:667–80.
20. Platig J, Castaldi PJ, DeMeo D, Quackenbush J. Bipartite community structure of eQTLs. Markowitz F, editor. *PLoS Comput Biol*. 2016;12:e1005033.
21. Morrow JD, Zhou X, Lao T, Jiang Z, DeMeo DL, Cho MH, et al. Functional interactors of three genome-wide association study genes are differentially expressed in severe chronic obstructive pulmonary disease lung tissue. *Sci Rep*. 2017;7:44232.
22. Kamburov A, Stelzl U, Lehrach H, Herwig R. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res*. 2013;41:D793–800.
23. DeMeo DL, Mariani T, Bhattacharya S, Srisuma S, Lange C, Litonjua A, et al. Integration of genomic and genetic approaches implicates IREB2 as a COPD susceptibility gene. *Am J Hum Genet*. 2009;85:493–502.
24. Hardin M, Zielinski J, Wan ES, Hersh CP, Castaldi PJ, Schwinder E, et al. CHRNA3/5, IREB2, and ADCY2 are associated with severe chronic obstructive pulmonary disease in Poland. *Am J Respir Cell Mol Biol*. 2012;47:203–8.
25. Pillai SG, Ge D, Zhu G, Kong X, Shianna KV, Need AC, et al. A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet*. 2009;5:e1000421.
26. Navratilova Z, Krsjakova T, Novosadova E, Zatloukal J, Kolek V, Petrek M. Genes affecting mRNA stability and metalloproteinase inhibitor RECK are down-regulated in patients with COPD. *Eur Respir J*. 2013; 42(Suppl 57):3512.
27. Hou P, Li L, Chen F, Chen Y, Liu H, Li J, et al. PTBP3-mediated regulation of ZEB1 mRNA stability promotes epithelial-mesenchymal transition in breast cancer. *Cancer Res*. 2017;canres.0883.2017.
28. Zhou Z-J, Dai Z, Zhou S-L, Hu Z-Q, Chen Q, Zhao Y-M, et al. HNRNPAB induces epithelial-mesenchymal transition and promotes metastasis of hepatocellular carcinoma by transcriptionally activating SNAIL. *Cancer Res*. 2014;74:2750–62.
29. Nowrin K, Sohal SS, Peterson G, Patel R, Walters EH. Epithelial-mesenchymal transition as a fundamental underlying pathogenic process in COPD airways: fibrosis, remodeling and cancer. *Expert Rev Respir Med*. 2014;8:547–59.
30. Sohal SS, Mahmood MQ, Walters EH. Clinical significance of epithelial mesenchymal transition (EMT) in chronic obstructive pulmonary disease (COPD): potential target for prevention of airway fibrosis and lung cancer. *Clin Transl Med*. 2014;3:33.
31. Castaldi PJ, Cho MH, Zhou X, Qiu W, McGeachie M, Celli B, et al. Genetic control of gene expression at novel and established chronic obstructive pulmonary disease loci. *Hum Mol Genet*. 2015;24:1200–10.
32. Lutz SM, Cho MH, Young K, Hersh CP, Castaldi PJ, McDonald M-L, et al. A genome-wide association study identifies risk loci for spirometric measures among smokers of European and African ancestry. *BMC Genet*. 2015;16(1):138.
33. Ezzie ME, Crawford M, Cho J-H, Orellana R, Zhang S, Gelinas R, et al. Gene expression networks in COPD: microRNA and mRNA regulation. *Thorax*. 2012;67:122–31.
34. Gheldof A, Bex G. Cadherins and epithelial-to-mesenchymal transition. *Prog Mol Biol Transl Sci*. 2013;116:317–36.
35. Imboden M, Bouzigon E, Curjurić I, Ramasamy A, Kumar A, Hancock DB, et al. Genome-wide association study of lung function decline in adults with and without asthma. *J Allergy Clin Immunol*. 2012;129:1218–28.
36. Yucesoy B, Kaufman KM, Lummus ZL, Weirauch MT, Zhang G, Cartier A, et al. Genome-wide association study identifies novel loci associated with diisocyanate-induced occupational asthma. *Toxicol Sci*. 2015;146:192–201.
37. Siedlinski M, Cho MH, Bakke P, Gulsvik A, Lomas DA, Anderson W, et al. Genome-wide association study of smoking behaviours in patients with COPD. *Thorax*. 2011;66:894–902.
38. Fingerlin TE, Murphy E, Zhang W, Peljto AL, Brown KK, Steele MP, et al. Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. *Nat Genet*. 2013;45:613–20.
39. Wain LV, Shrine N, Miller S, Jackson VE, Ntalla I, Artigas MS, et al. Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet Respir Med*. 2015;3:769–81.
40. Noth I, Zhang Y, Ma S-F, Flores C, Barber M, Huang Y, et al. Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: a genome-wide association study. *Lancet Respir Med*. 2013;1:309–17.
41. Hobbs BD, Parker MM, Chen H, Lao T, Hardin M, Qiao D, et al. Exome array analysis identifies a common variant in *IL27* associated with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*. 2016;194:48–57.
42. Morrow JD, Cho MH, Hersh CP, Pinto-Plata V, Celli B, Marchetti N, et al. DNA methylation profiling in human lung tissue identifies genes associated with COPD. *Epigenetics*. 2016;11:730–9.
43. Chen Y-C, Statt S, Wu R, Chang H-T, Liao J-W, Wang C-N, et al. High mobility group box 1-induced epithelial mesenchymal transition in human airway epithelial cells. *Sci Rep*. 2016;6:18815. doi:10.1038/srep18815.
44. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. *Nat Methods*. 2011;9:179–81.
45. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. Chakravarti a, editor. *G3amp58 GenesGenomesGenetics*. 2011;1: 457–470.
46. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. Schork NJ, editor. *PLoS Genet*. 2009;5:e1000529.
47. 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;38:904–9.
48. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*. 2012;28:1353–8.
49. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. *Nature*. 2015;526:68–74.
50. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*. 2015;4:7.
51. Aguet F, Ardlie KG, Cummings BB, Gelfand ET, Getz G, Hadley K, et al. Genetic effects on gene expression across human tissues. *Nature*. 2017; 550:204–13.
52. Csardi G, Nepusz T. The igraph software package for complex network research. *InterJournal 2006;Complex Systems*:1695.

53. Phanstiel DH, Boyle AP, Araya CL, Snyder MP. Sushi.R: flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. *Bioinformatics*. 2014;30:2808–10.
54. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*. 2005;21:3439–40.
55. Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. The Ensembl Regulatory Build. *Genome Biol*. 2015;16:56.
56. Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, et al. ENCODE Data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res*. 2013;41:D56–63.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

