Assaying lung-specific accessible chromatin to predict the causal variants in COPD

Christopher J. Benway, Jiangyuan Liu, Fei Du, Feng Guo, Michael H. Cho, Edwin K. Silverman, Xiaobo Zhou
Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA

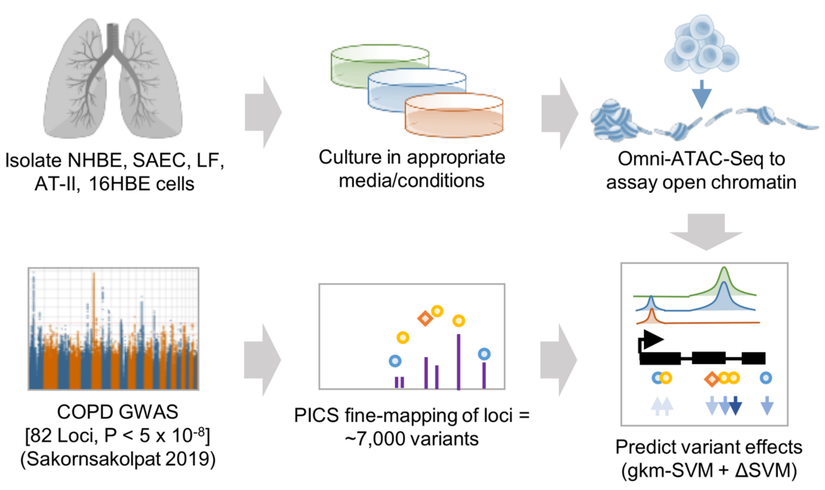
BRIGHAM HEALTH
BRIGHAM AND WOMEN'S HOSPITAL
HARVARD MEDICAL SCHOOL

ABSTRACT

Chronic obstructive pulmonary disease (COPD) is a complex and heterogeneous disorder characterized by irreversible obstruction of airflow in the lungs. COPD susceptibility is determined by environmental factors, mainly cigarette smoke exposure, and genetic factors, as evidenced by family and population studies. The largest genome-wide association study (GWAS) for COPD to date, examining a multi-cohort study population of over 250,000 individuals, identified 82 genetic loci significantly associated with disease susceptibility[1]. However, identifying the causal variants and their functional role in the appropriate cell type remains a major challenge due to the limitations of statistical fine-mapping and insufficient regulatory profiling in human lung cells.

We sought to assess chromatin accessibility by Omni-ATAC-seq[2] in primary human lung cell-types implicated in COPD pathology (large and small airway epithelium, alveolar type II pneumocytes, and lung fibroblasts) to (1) prioritize the putative causal variants at associated loci, (2) determine cell-type enrichment of genetic association, and (3) predict the molecular effects of risk variants in disease.

METHODS



1. We generated ATAC-seq data for four primary lung cell types implicated in disease pathology: alveolar type II pneumocytes (AT-II), bronchial epithelial (NHBE), small airway epithelial (SAEC), and lung fibroblast (LF) cells, as well as the 16HBE human bronchial epithelial cell line.
2. We used these lung cell chromatin accessibility profiles (as well as data from 222 DnaseI HS ENCODE samples) to generate cell type-specific regulatory sequence vocabularies using the machine learning algorithm 'gkm-SVM[4].'
3. Using summary statistics from the largest and most recent GWAS for COPD disease susceptibility, we performed statistical fine-mapping using the 'PICS' algorithm[3] to calculate the most likely causal SNPs given the observed association signal at each locus and create a "credible set" of SNPs for COPD risk.
4. We predicted the effect of 6,087 COPD risk variants in the credible set (corresponding to the 82 COPD loci) on chromatin accessibility in the five lung cell types by determining the induced change in gkm-SVM score, 'deltaSVM.'

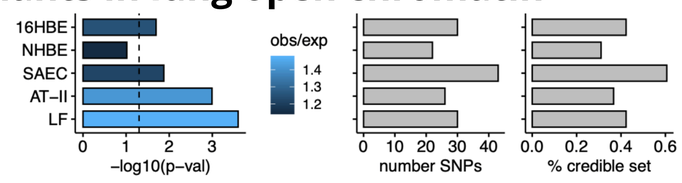
REFERENCES

1. P. Sakornsakolpat et al., *Nature Genetics*, in press, doi:10.1038/s41588-018-0342-2.
2. M. Corces et al., *Nature Methods*, in press, doi:10.1038/nmeth.4396.
3. K. Farh et al., *Nature*, in press, doi:10.1038/nature13835.
4. D. Lee et al., *Nature Genetics*, in press, doi:10.1038/ng.3331.

RESULTS

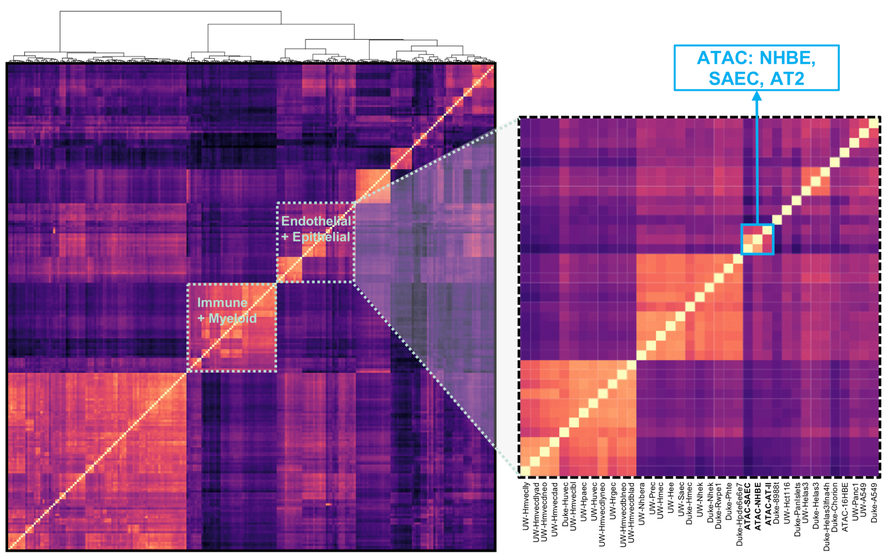
Overlap and enrichment of COPD GWAS variants in lung open chromatin

- PICS fine mapping of the largest GWAS for COPD identified ~6,000 putative causal variants.
- Direct overlap of COPD risk variants was significant in 4 of the 5 cell types (Fisher's exact test).
- 7 of the 82 (8.54%) index SNPs directly overlapped an open chromatin region in any lung cell type.
- Using stratified LD-score regression controlling for 53 functional genomic annotations, we found that COPD heritability was significantly enriched in accessible chromatin in AT-II cells (15.12-fold, p = 3.55×10^{-5}) but not the other lung cell types.



Index SNP	Nearest Gene	P-value (GWAS)	16HBE	NHBE	SAEC	AT-II	LF	P(PICS)
rs7068966	*CDC123*	6.2×10^{-23}			■			26.9%
rs4888379	*CFDP1*	5.9×10^{-21}			■			7.9%
rs2806356	*ARMC2*	2.9×10^{-15}					■	44.1%
rs2579762	*LRMDA*	2.6×10^{-10}					■	11.4%
rs798565	*AMZ1*	3.9×10^{-9}				■		34.1%
rs979453	*CCDC69*	1.4×10^{-8}			■			27.6%
rs34651	*TNPO1*	3.0×10^{-8}	■	■	■	■	■	93.5%

Machine learning algorithm predicts lung cell-specific functional variants in COPD



- Lung deltaSVM scores, in which we predicted the impact of COPD risk variants on chromatin accessibility, demonstrated cell type specific effects when compared to all available ENCODE cell types.
- Our ATAC-derived lung predictions for the COPD credible set correlated with epithelial and endothelial cells, including ENCODE samples from primary lung bronchial and small airway epithelium and lung-derived cell lines.

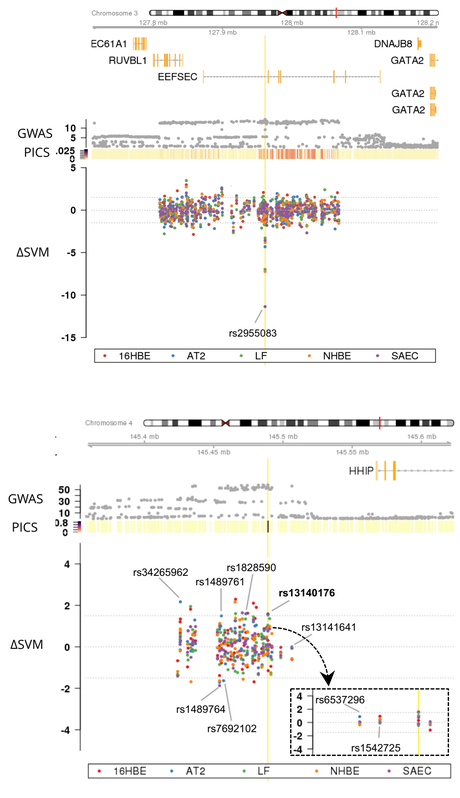
- DeltaSVM predictions identified novel putative functional variants in the COPD credible set.
- Variants with very large predicted effect sizes were generally not cell-type specific (compared to both lung and ENCODE models).
- Most variants with large absolute deltaSVM scores were considered low probability causal SNPs based on PICS fine mapping (P_{PICS} < 1%).
- Notably, the lead risk variant at the *EEFSEC* locus, rs2955083 (P_{PICS} = 2.85 %) produced large deltaSVM scores in all primary lung tissues tested. The COPD risk allele A is predicted to dramatically reduce open chromatin at the locus.

Top 10 COPD risk SNPs with positive deltaSVM scores for chromatin accessibility

	16HBE	AT2	LF	NHBE	SAEC	mean	locus	nearest gene
rs9435746	0.12	2.75	7.76	8.84	14.39	8.44	1p36.13	*SDHB*
rs76817161	-2.58	1.64	7.37	8.15	12.98	7.53	4q24	*GSTCD*
rs2227315	10.12	6.87	7.95	6.82	5.74	6.85	17q21.1	*CSF3*
rs6912639	0.09	3.37	6.73	7.27	9.41	6.70	6q24.2	*ADGRG6*
rs11049484	-0.74	3.45	5.60	6.95	10.76	6.69	12p11.22	*CCDC91*
rs12362635	-2.42	3.82	5.86	6.84	10.01	6.63	11q14.2	*ME3-DT*
rs17365084	-1.43	2.86	5.50	6.77	8.35	5.87	7p21.1	*ITGB8*
rs2579754	-2.33	2.92	5.81	5.68	8.48	5.72	10q22.3	*LRMDA*
rs35957220	6.17	4.69	6.84	5.45	4.79	5.44	7p22.3	*GNA12*
rs11049388	-1.22	1.05	5.90	5.67	7.28	4.97	12p11.22	*CCDC91*

Top 10 COPD risk SNPs with negative deltaSVM scores for chromatin accessibility

	16HBE	AT2	LF	NHBE	SAEC	mean	locus	nearest gene
rs61920227	0.76	-5.10	-7.49	-9.12	-11.37	-8.27	12p11.22	*CCDC91*
rs78729125	0.51	-2.74	-8.11	-8.42	-11.37	-7.66	17q21.31	*MAPT*
rs2955083	0.94	-4.31	-6.98	-7.23	-11.36	-7.47	3q21.3	*EEFSEC*
rs7897626	2.25	-3.34	-7.72	-8.41	-10.30	-7.44	10p14	*CDC123*
rs34134267	-0.67	-2.55	-6.83	-7.54	-10.82	-6.93	7p21.1	*ITGB8*
rs2285224	-8.04	-7.58	-7.45	-5.87	-4.81	-6.43	16q23.1	*CFDP1*
rs10516529	-8.75	-4.82	-7.20	-6.18	-5.34	-5.89	4q24	*NPNT*
rs7895605	0.08	-2.12	-5.73	-6.54	-7.80	-5.55	10q22.3	*LRMDA*
rs9909762	0.75	-1.41	-5.47	-5.46	-9.44	-5.44	17q24.3	*CASC17*
rs2904259	-4.94	-5.25	-5.44	-5.47	-4.31	-5.12	4q22.1	*FAM13A*



- Due to local LD structure at the *EEFSEC* locus, PICS fine mapping cannot resolve the likely causal variant for COPD risk.
- DeltaSVM scores from lung open chromatin identified a subset of SNPs with putative functional consequences, including the index variant, rs2955083, which was predicted to result in the largest effect on open chromatin in small airway cells.
- At one of the most significant and well-studied COPD risk loci, 4q31 (5-10 kb upstream *HHIP*) we profiled the predicted effect of variants on cell-specific chromatin accessibility.
- We found that the lead SNP, rs13140176 (P_{PICS} = >90%), had large deltaSVM scores in AT-II and 16HBE cells.
- The scores predict that the risk allele G leads to decreased chromatin accessibility in these cell types, providing support of a possible mechanism to explain that this risk allele is also associated with lower HHIP expression.
- Further experimental validation will be required to evaluate these candidate causal variants *in vitro* and *in vivo*.