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

Christopher J. Benway, Fei Du, Jiangyuan Liu, Michael H. Cho, Edwin K. Silverman, Xiaobo Zhou

Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA

INTRODUCTION

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 (ATII), 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.'

RESULTS

Omni-ATAC-seq identifies open chromatin in primary lung cell subtypes

- We observed high correlation between technical replicates and observed that the samples clustered according to cell type with peaks in the bronchial epithelium and small airway epithelium sharing the highest degree of similarity (mean r = 0.78, Fig. 1B).
- Analysis of the genomic distribution of the OCRs indicates enrichment in promoters, introns, and distal intergenic regions, as has been previously reported (Fig. 1C) .
- Peaks were enriched at transcription start sites (TSS) and enhancers, genome-wide (Figs. 1D,1E).
- Open chromatin regions (OCRs) corresponded to expression of known lung genes (Fig. 1F).

Overlap and enrichment of COPD GWAS variants in lung open chromatin

- Direct overlap of COPD risk variants was significant in 4 of the 5 cell types (Fig. 2A) and only 7/82 index SNPs directly overlapped an OCR in any lung cell type (Fig. 2B).
- COPD heritability was enriched in accessible chromatin in AT-II cells (15.12x, p = 3.55x10^{-5}, Fig. 2C)

DeltaSVM predicts lung cell-specific functional variants in COPD

- Lung deltaSVM scores, in which we predicted the impact of COPD risk variants on chromatin accessibility, were normally distributed for the credible set (Fig. 3A) and demonstrated cell type specific effects when compared to all ENCODE cells types (Fig. 3B).
- Our ATAC-derived lung predictions for the COPD credible set correlated with epithelial and endothelial cells, including ENCODE lung samples (Fig. 3C).
- DeltaSVM predictions identified novel SNPs within the COPD credible set with putative cell type specific functions, though the SNPs with the largest scores were often predicted to exhibit effects in many cell types (Figs 3D and 4A-F)..

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.

FIGURES

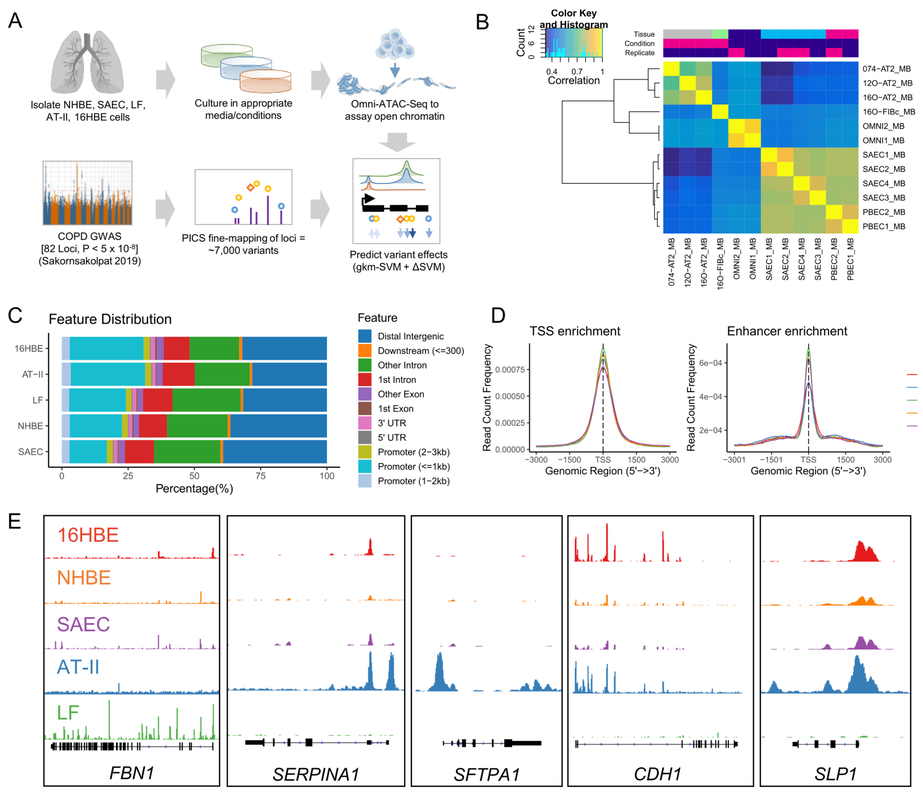

Figure 1. (A) Schematic outline of study design. Briefly, lung cell subtypes were obtained or isolated from donor tissue and cultured in appropriate conditions prior to lysis and transposition according to the Omni-ATAC-Seq protocol. Open chromatin regions (OCRs) were then used to evaluate the enrichment of COPD risk variants in cell-specific regulatory regions and construct predictive models of functional variant effects. (B) Correlation of peak sets from individual Omni-ATAC-seq libraries by Pearson correlation (C) Genomic distribution of OCRs for each cell type specific peak set after combining replicates (D) Profile of OCRs enriched at transcription start sites (TSSs) genome-wide in all five cell types. (E) Profile of OCRs enriched at FANTOM5 enhancers genome-wide in all five cell types. (F) Cell-type-specific signal of chromatin accessibility at genes with known differential expression in lung cells.

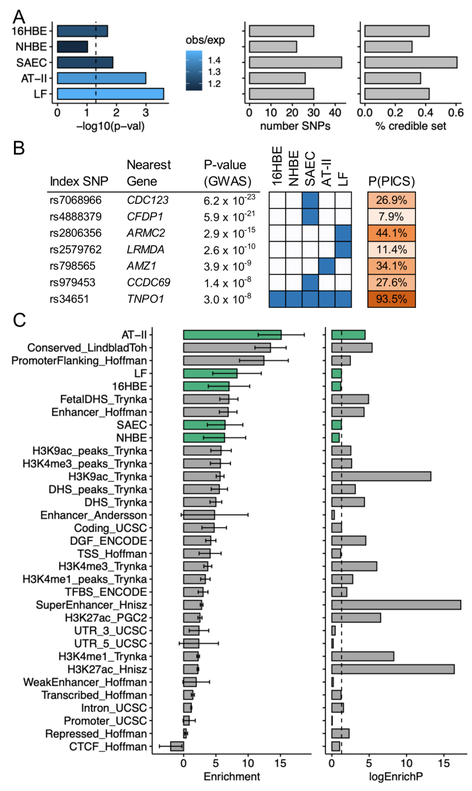

Figure 2. (A) Results of Fisher's exact test for overlap between COPD credible sets and lung cell OCRs. Significance of overlap (-log 10(P)), number of SNPs overlapping OCRs, and proportion of total credible set overlapping OCRs is indicated. Color of significance bars indicate the ratio of observed/expected overlaps. (B) Matrix of index SNPs from the credible sets which directly overlap OCRs in lung cells. The posterior probability (PICS) of the index SNP being the causal SNP is indicated. (C) COPD heritability enrichment (standard error) and significance level (–log 10(P)) of different functional genomic annotations estimated using partitioned LD score regression. Enrichment in lung cell OCRs are highlighted in green.

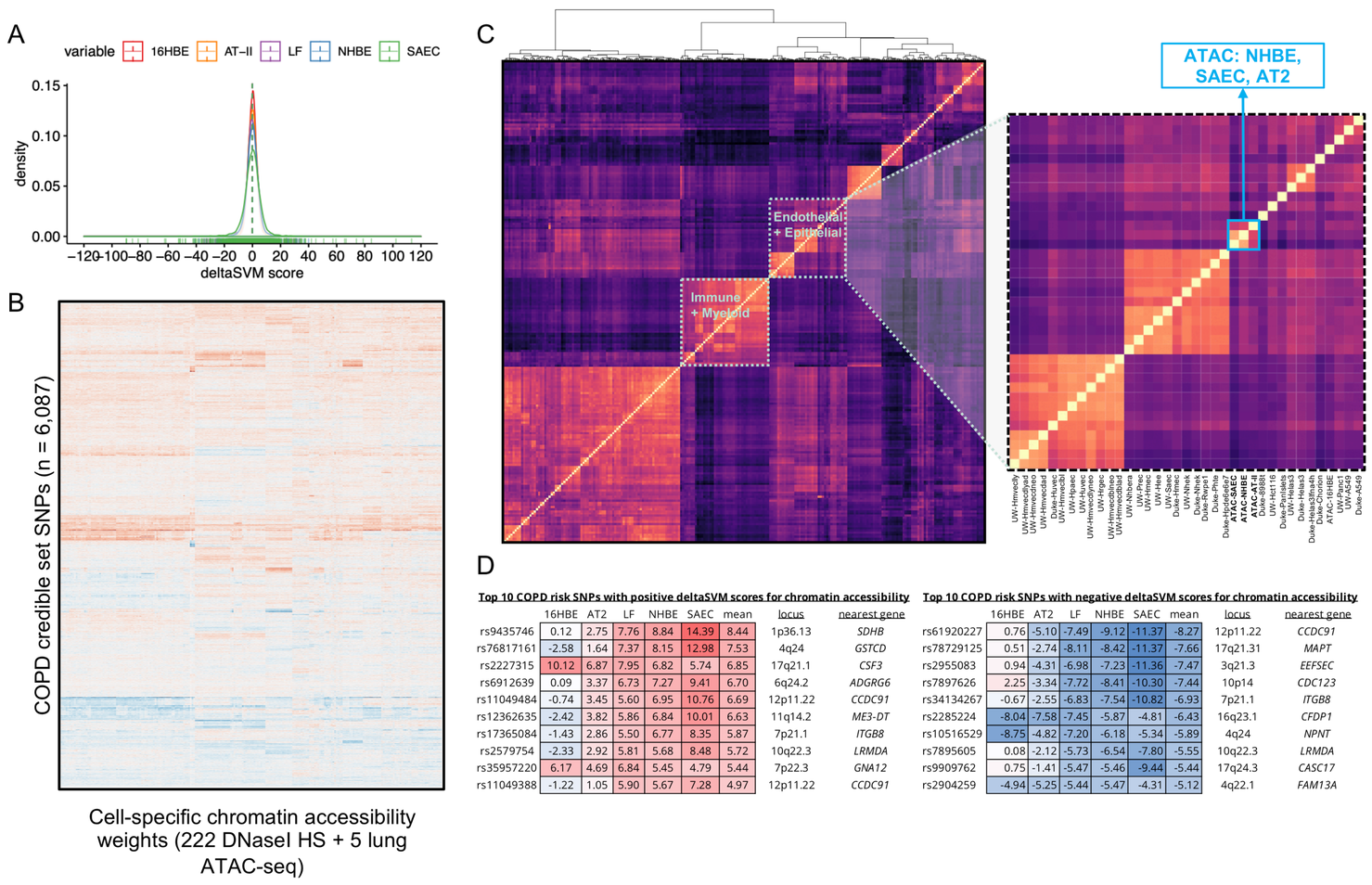

Figure 3. (A) Distribution of raw deltaSVM scores for the five cell types. (B) Cell type-specific deltaSVM scores for chromatin accessibility within the COPD credible set of SNPs across all ENCODE cell types and the lung specific models. (C) Correlation matrix of the deltaSVM scores for each cell type. (D) List of SNPs with the top scaled deltaSVM scores (positive and negative) across the five lung cell types.

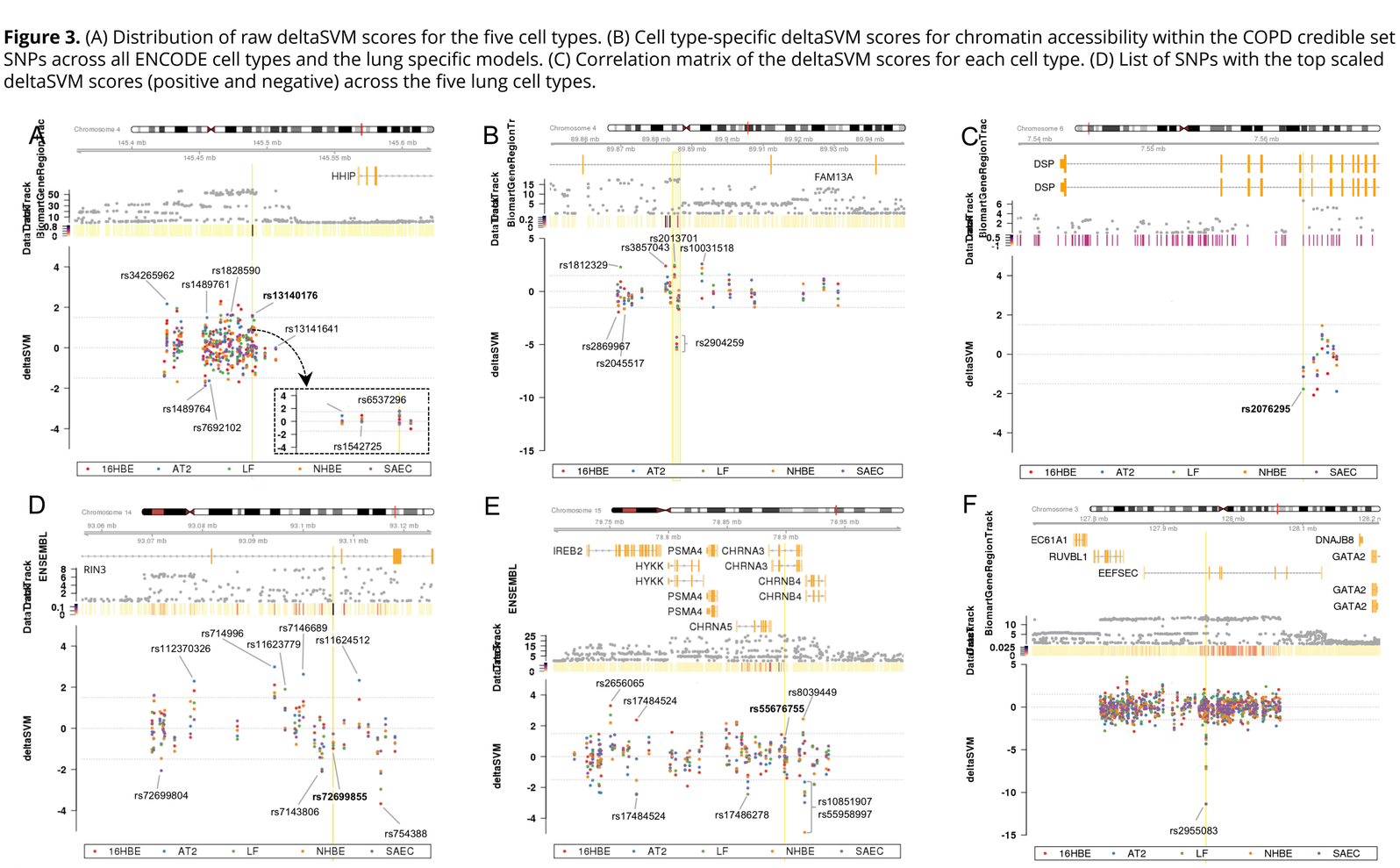

Figure 4. (A-F) Plots of deltaSVM scores for lung-specific chromatin accessibility at 6 COPD loci (nearest genes *HHIP, FAM13A, DSP, RIN3, CHRNA5*, and *EEFSEC*. In (A), the lead GWAS SNP for the *HHIP* locus, rs13140176, had large deltaSVM scores in AT-II cells and 16HBE cells. Our analysis predicted other putative functional SNPs at the locus, albeit with low causal probabilities. In (B), at the *FAM13A* locus, an experimentally validated functional SNP, rs2013701, scored consistent deltaSVM scores in all five cell types. (C) At the *DSP* locus, the lead SNP (and highly probable causal SNP), rs2076295, is predicted to impact chromatin accessibility in all cell types, especially lung fibroblasts. (D and E). At the *RIN3* and *CHRNA3/5* loci, numerous putative candidate SNPs were prioritized for future analysls. (F) At the *EEFSEC* locus, although PICS fine-mapping could not deconstruct the effect of LD, a very large deltaSVM score in lung cells suggests a functional role for the lead SNP, rs2955083.