Analysis of genetic variants in chronic obstructive pulmonary disease using a hierarchical model of the cell

Simon J. Larsen[1,*], Daniel E. Carlin[2], Trey Ideker[2], and Jan Baumbach[1,3]

[1]Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark
[2]Department of Medicine, University of California, San Diego, USA
[3]Chair of Experimental Bioinformatics, Wissenschaftszentrum Weihenstephan, Technical University of Munich, Munich, Germany
[*]Corresponding author, e-mail: sjlarsen@imada.sdu.dk

Abstract

We developed HiGAna, a cell hierarchy-based method for genetic association studies. Our method first constructs a gene hierarchy based on the Gene Ontology and computes a low-dimensional representation of the variants in each term. All terms in the hierarchy are then tested for association with a phenotype using a generalized linear model framework. We also introduce two conditional tests for filtering terms that are redundant or enriched mainly due to a single gene. We apply HiGAna to find cellular mechanisms associated with COPD-status in smokers and discover two novel mechanisms not implicated in the standard GWAS: semaphorin receptor binding and vesicle transport. We also compare our results to other self-contained enrichment tests and find that HiGAna provides comparatively good power and speed, but the results are largely complementary between methods. HiGAna is available as a package for the R programming language at `https://higana.compbio.sdu.dk`.

Introduction

Chronic obstructive pulmonary disease (COPD) is a type of progressive lung disease characterized by long-term breathing problems and airflow obstruction. There were an estimated 384 million COPD cases worldwide in 2010 [1], and an estimated three million deaths annually caused by COPD [21], making it one of the leading causes of death. Tobacco smoking is the most significant risk factor for COPD, but other factors, such as childhood asthma, air pollution, and occupational exposure to dusts and chemical fumes, are also associated with chronic airflow obstruction and COPD [7]. Although the main cause of COPD is smoking, it is believed to arise from an interaction between genetic and environmental factors [7]. A twin study in Danish and Swedish twins suggested as much as 60% of COPD susceptibility is explained by additive genetic factors [15]. Currently, the most well-established inherited COPD risk factor is severe α_1-antitrypsin (AAT) deficiency arising from single-nucleotide variants in the serpin family A member 1 gene (*SERPINA1*) [10].

In the past, genome-wide association studies (GWAS) have been instrumental to the discovery of thousands of single-nucleotide polymorphisms (SNPs) associated with human

disease or phenotypic traits. A recent large-scale GWAS identified 82 genome-wide significant variants in COPD, but these variants still only explain up to 7% of the phenotypic variance [29]. A major challenge with GWAS is the multiple testing problem arising from testing thousands to millions of SNPs for significance. This problem is most commonly overcome by controlling the family-wise error rate using Bonferroni correction, with $p \leq 5 \times 10^{-8}$ being the current *de facto* standard significance cutoff. This strict significance cutoff, in turn, means that strong effect sizes or large sample sizes are necessary in order for causal variants to meet the significance criteria.

Previous work has attempted to tackle the multiple testing problem by using gene-based and gene set-based tests instead, thus vastly reducing the number of tested hypotheses. Gene and gene set-based methods also have the potential to help discover groups of variants with weaker effects that would not be detected when testing individual SNPs. Furthermore, evaluating genetic variants at the gene- or gene set-level also helps with the functional interpretation of the results. PLINK [25, 5] provides a self-contained method for testing groups of SNPs or genes for association. A standard single-variant association analysis is first performed to obtain a test statistic for each SNP. For each gene set, a set of most significant SNPs is then constructed greedily among all nominally significant SNPs while ensuring no two SNPs exhibit strong LD (e.g. $r^2 > 0.5$). Each set is then assigned a score equal to the mean test statistic among the top SNPs and this score is tested for significance using a permutation test. MAGMA [19] introduced a gene set test using a linear regression framework. Each gene is tested for enrichment using principal component regression (PCR) on nearby SNPs. Gene sets are then tested for enrichment by evaluating the distribution of z-values of genes in each set to a null distribution. MAGMA provides both a self-contained and a competitive gene set test. Several other notable competitive test include ALIGATOR [14], MAGENTA [30] and INRICH [18].

Ontologies are commonly used in biological and biomedical sciences for modeling knowledge as a hierarchy where concepts are recursively described by increasingly more specific concepts. For more than two decades, extensive knowledge about the cellular structure, elemental functions and higher-level processes of the cell has been collected in the Gene Ontology (GO) knowledge base [34]. This ontology is complemented by a genome annotation database, describing gene products in terms of GO terms thus providing a structured, functional description of each gene. Recently, researchers have employed ontologies instead of flat networks or separate gene sets for analyzing molecular profiling data. A gene ontology was previously used to propagate genetic perturbations through an ontology in order to model their effect on cell growth in yeast [38]. In a subsequent study, a deep neural network was constructed from the hierarchical organization of an ontology, in order to provide an interpretable model of cell viability while retraining high predictive performance [20].

Here, we introduce HiGAna (Hierarchical Genetic Association Analysis), a new method for cell hierarchy-based association studies in genotype data. We first construct a functional/structural gene hierarchy from the Gene Ontology. Then, we apply our method to compute a low-dimensional representation of the SNPs associated with each term and identify terms associated with COPD outcome using a self-contained enrichment test based on generalized linear models (Figure 1c). A common issue with Gene Ontology-based enrichment or overrepresentation analysis is the prevalence of redundant terms reported as significant. A cellular mechanism significantly associated with outcome will often result in not just one significant term but a several highly related terms located in the same subhierarchy, of which

most are more general descriptions of the relevant mechanism. A single gene containing strongly associated variants may also drive a gene set to be enriched despite there being no significant term-wide effect. To overcome this, we further implemented two conditional tests for identifying redundant terms whose signal is mainly driven by a single gene or child term.

We evaluate our method by analyzing a COPD genotype data set. We identify 27 enriched terms, including three biological processes not previously implicated in COPD GWAS. In this paper, we focus our analysis on gene sets of at least three genes, but constructed hierarchies can theoretically include single genes or even single-nucleotides as well (Figure 1b). We also validate our results in UK Biobank summary statistics and find that a large fraction of the terms replicate (Figure 1a). Finally, we compare our results to those obtained with other self-contained gene set tests implemented in MAGMA and PLINK, as well as a standard gene set overrepresentation test and find that our method has good statistical power and computation time.

Results

Ontology construction

We constructed a gene hierarchy based on the Gene Ontology [34] knowledge base. We built a hierarchy from "is a" and "part of" relations and removed redundant terms (see Methods). The constructed ontology contained 16,451 terms. In this study, we restricted the analysis to terms with at least three and at most 75 genes. We considered terms with fewer than three genes to be too small to be considered a function, process or cellular component, and we considered terms with more than 75 genes too general to meaningfully interpret. The reduced ontology contained 9312 terms (see Figure S3 for term size distribution).

Cell hierarchy-based association analysis of COPD data

The association analysis was performed on the COPDGene genotype data with disease status as outcome (2812 cases vs. 2534 controls). Among the 9312 tested terms, we identified 27 terms that were significantly associated with COPD (FDR < 0.05, Benjamini-Yekulieli). The results are summarized in Table 1 (see Figure S2 for a hierarchical overview, Table S1 for the full table of results). Among the significant terms, five were from the molecular function domain of GO, one was a cellular component, and the remaining 21 were biological processes. Many of the significant terms contained genes near a strong peak region in 15q25.1 (Figure S1), which suggests their signal was driven mainly by this peak, and would thus also be implicated in a standard GWAS.

We identified three terms that did not contain any genome-wide significant SNPs ($p < 5 \times 10^{-8}$) nor any SNPs in LD with such a SNP ($r^2 > 0.2$, within 1 kb): "semaphorin receptor binding" (GO:0030215), "organelle transport along microtubule" (GO:0072384) and "vesicle transport along microtubule" (GO:0047496). We considered these three terms to be novel since they would not be implicated in a standard SNP-wise analysis, and, for all three terms, the term, when considered as a whole, was more significant than any of the constituting SNPs (Figure 2a-b). The "semaphorin receptor binding" term (GO:0030215) consists mainly of genes coding for proteins in the semaphorin protein family. The strongest effects in the term PCR was from SNPs in or near genes *SEMA5A* and *SEMA6A* in chromosome 5, *SEMA5B*

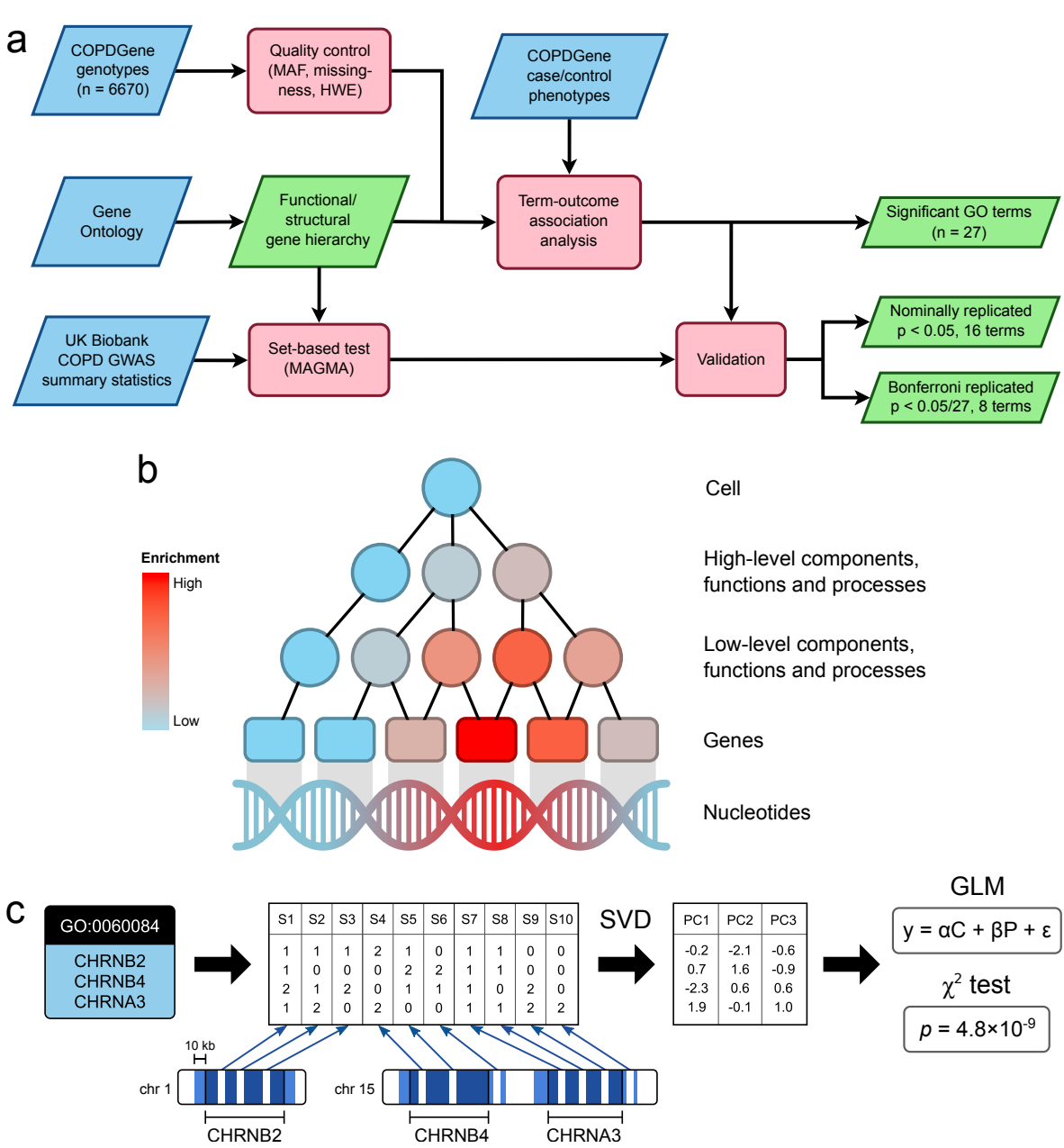

Figure 1: Study design. (a) Study analysis workflow. Blue rhomboids indicate data inputs and green rhomboids indicate analysis outputs. (b) The functional/structural gene hierarchy models the hierarchical organization of the cell. Terms increase in specificity when traversing towards the bottom. The enrichment analysis tests terms for association to identify enriched subhierarchies. (c) Dimensionality reduction and association test workflow. For each term, a SNP matrix is constructed from the variants near its genes. A low-dimensionality representation is computed using singular value decomposition, and the term is tested for enrichment by regressing the outcome on the PCs, P, and other covariates, C.

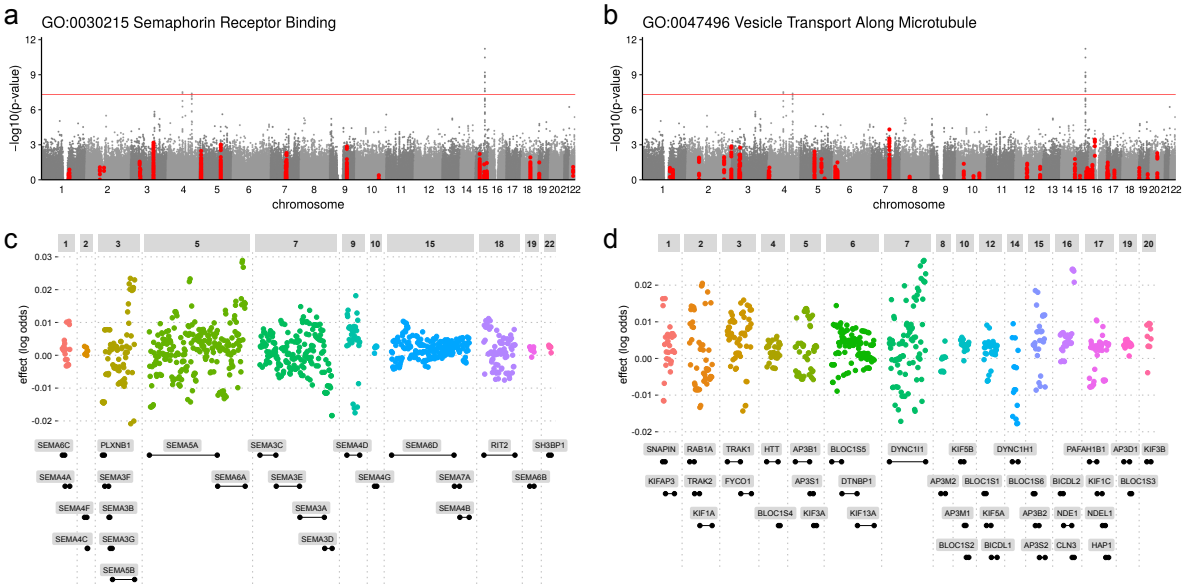

Figure 2: Associated SNPs and contributed effects for two novel, significant GO terms, GO:0030215 and GO:0047496. (a-b) Manhattan plot of standard GWAS analysis with SNPs associated with GO:0030215 (a) and GO:0047496 (b) highlighted in red. The red horizontal line marks $p = 5 \times 10^{-8}$. (c-d) Contributed effects of SNPs in the principal component regression models for GO:0030215 (c) and GO:0047496 (d). SNPs are sorted by genomic position. Horizontal distance between SNPs does not correspond to genomic distance. SNP effect is the change in log odds ratio per minor allele copy contributed to the principal component regression for the entire term.

in chromosome 3, *SEMA3D* in chromosome 7 and *SEMA4D* in chromosome 9 (Figure 2c). The term "vesicle transport along microtubule" (GO:0047496) is a child (is-a relationship) of "organelle transport along microtubule" (GO:0072384), with the more specific term exhibiting the strongest enrichment. The strongest effects came from SNPs in or near genes *DYNC1I1* in chromosome 7, *CLN3* in chromosome 16 and *KIF1A* in chromosome 2 (Figure 2d).

Conditional association tests

We implemented two conditional association tests conditioning each term on its most significant child term or gene, respectively, in order to identify redundant terms (see Methods). The results of the conditional association tests on the 27 significantly associated terms are summarized in Figure 3a).

If one wants to reject terms based on this test, we recommend a p-value cutoff of $0.05/n$ where n is the number of terms tested. In this case, using a cutoff of 0.05/27 results in 13 terms passing the "top child" test and 5 terms passing the "top gene" test. Interestingly, the three terms least affected by removing the most enriched gene were the three novel terms (GO:0047496, GO:0072384, GO:0030215), demonstrating that this test effectively filters terms that are enriched mainly due to a single gene or genetic locus.

To illustrate the utility of the "top child" conditional test, we highlight the subhierarchy near "vesicle transport along microtubule" (GO:0047496). The term and its parent term

term	asp.	genes	p-value	FDR	FWER	description
GO:0060084	BP	3	1.12×10^{-10}	0.00001	0.00000	synaptic transmission involved in micturition
GO:0035095	BP	5	2.55×10^{-9}	0.00012	0.00002	behavioral response to nicotine
GO:0014891	BP	5	2.05×10^{-8}	0.00062	0.00019	striated muscle atrophy
GO:0022848	MF	10	3.38×10^{-8}	0.00076	0.00031	acetylcholine-gated cation-selective channel activity
GO:0014889	BP	6	4.87×10^{-8}	0.00081	0.00045	muscle atrophy
GO:0043501	BP	6	5.40×10^{-8}	0.00081	0.00050	skeletal muscle adaptation
GO:0005892	CC	16	1.09×10^{-7}	0.00141	0.00102	acetylcholine-gated channel complex
GO:0014732	BP	3	2.08×10^{-7}	0.00235	0.00193	skeletal muscle atrophy
GO:0010043	BP	29	3.26×10^{-7}	0.00327	0.00303	response to zinc ion
GO:0060073	BP	4	3.96×10^{-7}	0.00331	0.00368	micturition
GO:0007271	BP	21	4.03×10^{-7}	0.00331	0.00375	synaptic transmission, cholinergic
GO:0014888	BP	12	6.87×10^{-7}	0.00518	0.00639	striated muscle adaptation
GO:0051536	MF	26	7.60×10^{-7}	0.00529	0.00708	iron-sulfur cluster binding
GO:0035094	BP	20	1.48×10^{-6}	0.00958	0.01380	response to nicotine
GO:0003994	MF	3	1.84×10^{-6}	0.01108	0.01710	aconitate hydratase activity
GO:0043500	BP	17	2.03×10^{-6}	0.01146	0.01888	muscle adaptation
GO:0047496	BP	37	2.37×10^{-6}	0.01263	0.02209	vesicle transport along microtubule
GO:0046501	BP	3	2.63×10^{-6}	0.01324	0.02452	protoporphyrinogen IX metabolic process
GO:0010040	BP	3	3.03×10^{-6}	0.01442	0.02820	response to iron(II) ion
GO:0010039	BP	14	3.22×10^{-6}	0.01456	0.02997	response to iron ion
GO:0071364	BP	27	3.43×10^{-6}	0.01476	0.03190	cellular response to epidermal growth factor stimulus
GO:0060079	BP	40	4.55×10^{-6}	0.01870	0.04235	excitatory postsynaptic potential
GO:0071281	BP	7	4.96×10^{-6}	0.01952	0.04620	cellular response to iron ion
GO:0007588	BP	32	9.12×10^{-6}	0.03439	0.08495	excretion
GO:0072384	BP	58	1.02×10^{-5}	0.03705	0.09533	organelle transport along microtubule
GO:0042166	MF	15	1.37×10^{-5}	0.04785	0.12803	acetylcholine binding
GO:0030215	MF	23	1.45×10^{-5}	0.04850	0.13477	semaphorin receptor binding

Table 1: GO terms significantly associated with COPD status. FDR and FWER correction was performed with the Benjamini-Yekutieli and Bonferroni procedures, respectively. Second column refers to the Gene Ontology aspect each term is associated with: biological process (BP), molecular function (MF), cellular component (CC).

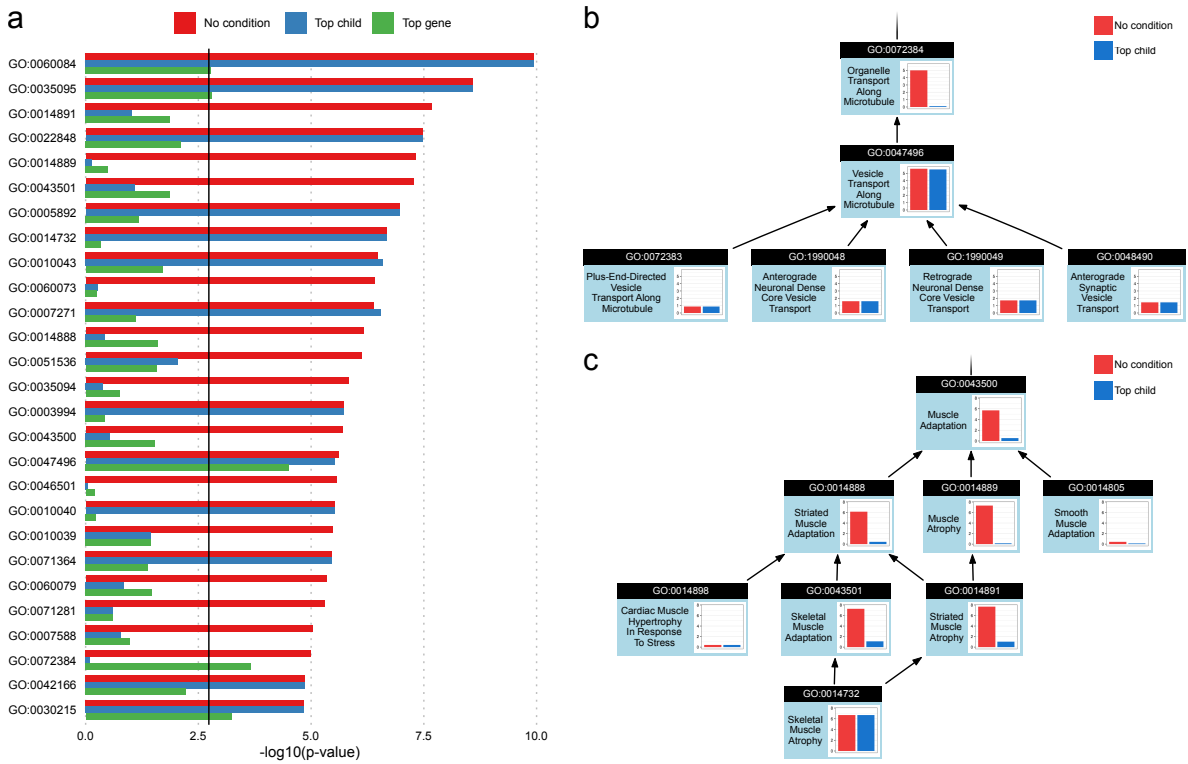

Figure 3: Conditional association test results. (a) Comparison of enrichment results for normal analysis and conditional tests. P-values are unadjusted. Vertical line corresponds to FWER cutoff $p = 0.05/27$. (b,c) Conditional results for the vesicle transport along microtubule and muscle adaptation subhierarchies, respectively. Bar heights indicate -log10(p-value).

(GO:0072384) were both highly significant in the standard test. However, when conditioned on its top child, the parent term lost almost all signal, while the child term was mostly unaffected (Figure 3b). All four children of GO:0047496 were leaf terms and exhibited only a moderate degree of enrichment individually. These results indicate that "vesicle transport along microtubule" is the appropriate level of specificity when considering this biological process in relation to COPD. In other cases, the conditional test makes it clear that one highly specific term is enriched, while ancestor terms appear to be enriched only due to being supersets of the relevant term. For instance, in a subhierarchy related to muscle adaptation, we observed that "skeletal muscle atrophy" (GO:0014732) was enriched while none of its ancestors retained a significant amount of signal when conditioned on the most enriched child (Figure 3c).

Comparison with other methods

We evaluated the power of HiGAna against the self-contained gene set enrichment tests implemented in MAGMA (v1.07b) [19] and PLINK (v1.90b3b3.31) [5] as well as a gene set enrichment (GSE) test using Fisher's exact test (Supplementary text S1.1) on the set of genes within 10 kb of a genome-wide significant SNP ($p < 5 \times 10^{-8}$). These genes were *CHRNA5*, *CHRNA3*, *CHRNB4*, *FAM13A*, *IREB2*, *HYKK* and *PSMA4*. PLINK overall reported the greatest number of significant terms, followed by HiGAna. PLINK reported 62 significant

method	$p < 0.05$	$p < 0.01$	$p < 0.001$	FDR	FWER	Computation time
HiGAna	831	244	67	27	23	2 hours 32 minutes
PLINK	911	305	112	62	39	5 days 15 hours
MAGMA	710	166	21	3	3	2 minutes 29 seconds
GSE	86	59	28	19	19	6 seconds

Table 2: Number of significant terms reported by each method for different significance criteria. FDR and FWER are the number of terms with $p < 0.05$ after Benjamini-Yekutieli and Bonferroni correction, respectively. GSE denotes gene set enrichment using a Fisher's exact test.

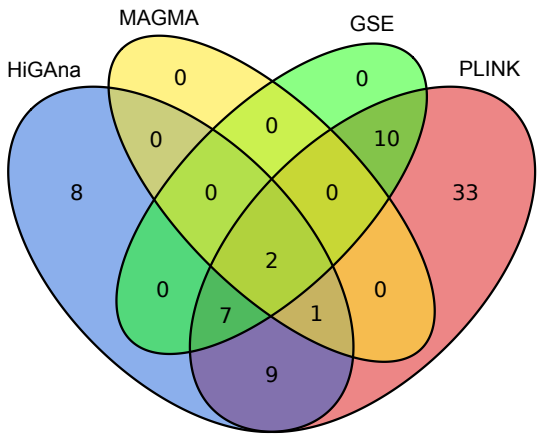

Figure 4: Venn diagram of significant terms (FDR < 0.05) discovered by HiGAna, MAGMA, PLINK and Fisher's exact test gene set enrichment (GSE).

terms (FDR, $p < 0.05$), while HiGAna reported 27 terms, and GSE and MAGMA reported 19 and 3 terms, respectively (Table 2). All terms reported by MAGMA and nine of 19 terms reported by GSE were also found with HiGAna (Figure 4a). All terms reported by MAGMA and GSE were also found by PLINK. Nine terms were shared between HiGAna and PLINK, while 8 were unique to HiGAna and 33 were unique to PLINK.

The three terms we identified as novel were all among the eight terms that were unique to HiGAna: "Semaphorin receptor binding" (GO:0030215), "Organelle transport along microtubule" (GO:0072384) and "Vesicle transport along microtubule" (GO:0047496). Similarly, we observed that three of the terms unique to PLINK fit the same novelty criteria, namely "Positive regulation of cellular extravasation" (GO:0002693), "Acylglycerol O-acyltransferase activity" (GO:0016411) and "Glucose mediated signaling pathway" (GO:0010255). Thus, both PLINK and HiGAna were able to discover novel enriched terms, but their findings were complementary.

We measured the computation time needed by each method on a Linux server with 48 cores at 2.50 GHz and 512 GB RAM. PLINK took 5 days 15 hours 23 minutes, HiGAna took 2 hours 32 minutes, MAGMA took 2 minutes 29 seconds and the gene set enrichment took just 6 seconds. The long computation time of PLINK is due to the use of permutation tests and the large number of permutations needed to get accurate estimates of significance. The

majority of the computation time needed by HiGAna is spent on singular value decomposition of the term SNP matrices. This step can be sped up significantly by using randomized matrix decomposition instead [12] at the cost of a small decrease in accuracy. We have included this as an option in the provided R package.

Type 1 error rate

We estimated the type 1 error rate of HiGAna empirically using permutation tests. We repeated the association test 10,000 times while randomly permuting the outcome variable in order to generate a null distribution with no disease-associated SNPs. The estimated type 1 error rate was the mean fraction of terms with $p < 0.05$ over all permutations. The estimated type 1 error was 0.0514 (0.0077 SD) for HiGAna, suggesting the type 1 error rate was well-controlled.

Validation in UK Biobank summary statistics

We performed external validation of our results using summary-level statistics from the UK Biobank (`http://nealelab.is/uk-biobank`, August 1, 2018) with ICD-10 code J44 as phenotype. Since we did not have access to individual-level genotypes, we could not use HiGAna nor PLINK for validation. Instead, we applied the SNP summary statistics analysis in MAGMA using the 1000 Genomes Project phase 3 European population data as reference population [33]. Of the 27 terms reported by HiGAna, 16 terms validated nominally ($p < 0.05$) and 8 replicated after Bonferroni correction ($p < 0.05/27$) (Table 3; full table in Table S3). None of the novel terms replicated with this method, although one term ("organelle transport along microtubule", GO:0072384) was near nominally significant ($p = 0.0547$).

Discussion

We developed a cell hierarchy-based method for finding associations between genetic variants affecting cellular mechanisms or subcomponents and phenotypic traits. We applied our methodology to a COPD case-control study and found 27 terms significantly associated with COPD outcome. We identified three terms that we considered novel with respect to single-SNP GWAS due to their high term-wide significance despite having no near-genome-wide significant SNPs. We further introduced two conditional tests conditioning terms on their most significant gene or child term and demonstrated their ability to identify redundant terms. Most notably, the "top gene" test confirmed that the three novel terms were not significantly driven by a single gene unlike most other reported terms, and the "top child" test identified that organelle transport along microtubule (GO:0072384) was mainly associated with COPD outcome due to the more specific child process, vesicle transport along microtubule (GO:0047496).

We observed a strong association with several terms related to nicotine response, including behavioral response to nicotine (GO:0035095; adj. $p = 0.00002$) and acetylcholine-gated cation-selective channel activity (GO:0022848; adj. $p = 0.00076$). These terms comprise of nicotinic acetylcholine receptor units (nAChRs), receptor proteins that respond to the acetylcholine neurotransmitter as well as nicotine [2]. Several SNPs in the CHRNA3/CHRNAB4/CHRNA5 gene cluster have been linked to nicotine dependence in genetic association studies [35, 32, 27, 28]. This locus has also previously been associated with COPD [24, 36], but the link between

Table 3: Validation of terms reported by HiGAna in UK Biobank summary statistics. Significance values below 0.05 are highlighted with bold text. Second column refers to the Gene Ontology aspect each term is associated with: biological process (BP), molecular function (MF), cellular component (CC).

term	asp.	genes	p-value	FWER	description
GO:0046501	BP	3	$\mathbf{5.94 \times 10^{-10}}$	$\mathbf{1.60 \times 10^{-8}}$	protoporphyrinogen IX metabolic process
GO:0060084	BP	3	$\mathbf{7.35 \times 10^{-6}}$	$\mathbf{1.99 \times 10^{-4}}$	synaptic transmission involved in micturition
GO:0060073	BP	4	$\mathbf{2.20 \times 10^{-5}}$	$\mathbf{5.95 \times 10^{-4}}$	micturition
GO:0022848	MF	10	$\mathbf{8.34 \times 10^{-5}}$	$\mathbf{2.25 \times 10^{-3}}$	acetylcholine-gated cation-selective channel activity
GO:0035095	BP	5	$\mathbf{1.45 \times 10^{-4}}$	$\mathbf{3.90 \times 10^{-3}}$	behavioral response to nicotine
GO:0014732	BP	3	$\mathbf{4.35 \times 10^{-4}}$	$\mathbf{1.18 \times 10^{-2}}$	skeletal muscle atrophy
GO:0042166	MF	15	$\mathbf{6.94 \times 10^{-4}}$	$\mathbf{1.87 \times 10^{-2}}$	acetylcholine binding
GO:0060079	BP	40	$\mathbf{7.15 \times 10^{-4}}$	$\mathbf{1.93 \times 10^{-2}}$	excitatory postsynaptic potential
GO:0005892	CC	16	$\mathbf{2.98 \times 10^{-3}}$	8.03×10^{-2}	acetylcholine-gated channel complex
GO:0003994	MF	3	$\mathbf{4.86 \times 10^{-3}}$	1.31×10^{-1}	aconitate hydratase activity
GO:0007271	BP	21	$\mathbf{7.74 \times 10^{-3}}$	2.09×10^{-1}	synaptic transmission, cholinergic
GO:0035094	BP	20	$\mathbf{1.42 \times 10^{-2}}$	3.82×10^{-1}	response to nicotine
GO:0010043	BP	29	$\mathbf{1.68 \times 10^{-2}}$	4.55×10^{-1}	response to zinc ion
GO:0010040	BP	3	$\mathbf{1.78 \times 10^{-2}}$	4.80×10^{-1}	response to iron(II) ion
GO:0014889	BP	6	$\mathbf{3.68 \times 10^{-2}}$	9.92×10^{-1}	muscle atrophy
GO:0071364	BP	27	$\mathbf{3.71 \times 10^{-2}}$	1.00	cellular response to epidermal growth factor stimulus
GO:0072384	BP	58	5.47×10^{-2}	1.00	organelle transport along microtubule
GO:0043501	BP	6	6.00×10^{-2}	1.00	skeletal muscle adaptation
GO:0014891	BP	5	8.64×10^{-2}	1.00	striated muscle atrophy
GO:0007588	BP	32	1.11×10^{-1}	1.00	excretion
GO:0071281	BP	7	1.36×10^{-1}	1.00	cellular response to iron ion
GO:0047496	BP	37	1.54×10^{-1}	1.00	vesicle transport along microtubule
GO:0030215	MF	23	2.21×10^{-1}	1.00	semaphorin receptor binding
GO:0051536	MF	26	2.67×10^{-1}	1.00	iron-sulfur cluster binding
GO:0014888	BP	12	6.36×10^{-1}	1.00	striated muscle adaptation
GO:0043500	BP	17	6.92×10^{-1}	1.00	muscle adaptation
GO:0010039	BP	14	7.08×10^{-1}	1.00	response to iron ion

nAChRs and COPD is likely a result of variants affecting nicotine response which in turn produces differences in smoking behavior that affect COPD outcome.

We observed significant enrichment in a subhierarchy related to muscle adaptation, more specifically skeletal muscle atrophy (GO:0014732; adj. $p = 0.00235$), as well as a subhierarchy related to iron ion response (GO:0010039; adj. $p = 0.0146$). These terms were enriched mainly due to variants in the iron-responsive element-binding protein 2 (*IREB2*), as confirmed by the "top gene" test. The protein encoded by *IREB2* is an RNA-binding protein (*IRP2*) that is involved with maintaining iron homeostasis, and has previously been associated with COPD and lung function in genotyping and mRNA expression [24, 8, 4]. While the link between IREB2 and pulmonary disease is still unclear, cigarette smoking has been associated with higher levels of iron in the lungs which may, in turn, contribute to oxidative injury and pulmonary obstruction. [22, 11]. Skeletal muscle dysfunction is also a common comorbidity of COPD [17]. Finally, the association between *IREB2* and COPD may also be explained, at least in part, by linkage disequilibrium with SNPs near *CHRNA3/CHRNA5* [13].

Semaphorin receptor binding (GO:0030215) was found to be significantly enriched (adj. $p = 0.048$) despite containing no individually significant SNPs. Semaphorins are a class of secreted and membrane proteins that were initially characterized for their role in axonal growth and guidance but are now known to play an essential in the development and maintenance of many tissues [37, 3]. Recent studies have found an association between forced vital capacity (FVC) and variants near *SEMA3D* (rs12707691) [31] and *SEMA4D* (rs7036107) [16], and a targeted GWAS of 389 SNPs selected from differential mRNA expression analysis identified several SNPs in SEMA6D to be associated with COPD status [8].

We also found an association with vesicular transport along microtubules (GO:0047496; adj. $p = 0.0126$). While a wide range of genes were associated with this process in the Gene Ontology, the strongest effects were observed in the genes *DYNC1I1* and *DYNC1H1*, coding for proteins that are part of the cytoplasmic dynein 1 complex that moves various cargoes along microtubules, and the gene *CLN3* encoding Battenin, a lysosomal/endosomal transmembrane protein with mostly unknown function [23]. An intron variant in *DYNC1I1* (rs6961619) was previously associated with FVC in the UK Biobank.

Although all subjects in the COPD cohort were long-time smokers, the amount of exposure to tobacco smoking still varies between subjects. Thus, we cannot fully rule out that any of the reported terms are associated with COPD status because they affect smoking behavior. Follow-up work should incorporate smoking-related covariates in the model (e.g. pack-years) to avoid this issue.

When comparing HiGAna to PLINK and MAGMA, we found that HiGAna and PLINK had greater power than MAGMA. This contradicts the evaluation in the original MAGMA manuscript [19] where MAGMA was found to be more powerful than PLINK in a Crohn's disease cohort. This discrepancy may be due to a difference in sample sizes, effect sizes or the sizes and number of tested gene sets. PLINK overall reported more significant terms than HiGAna, but both methods reported three unique terms without any genome-wide significant SNPs, demonstrating that both methods were able to boost the signal of terms with SNPs that would otherwise be missed with a standard GWAS, but that their results were complementary. We also observed an overall greater agreement between HiGAna and PLINK ($\rho = 0.569$) than between HiGAna and MAGMA ($\rho = 0.452$) or between MAGMA and PLINK ($\rho = 0.403$; Figure S5).

While MAGMA was very fast, it provided no other benefit over PLINK and HiGAna

for this data set. The computation time also differed greatly between HiGAna and PLINK. The cohort used in this evaluation was relatively small (5314 subjects), but modern genetic association studies may comprise of tens to hundreds of thousands of subjects. We hypothesize that in significantly larger cohorts, the difference in computation time will increase even further, which may render the PLINK analysis infeasible.

In this study, we limited the analysis to terms with three to 75 genes in order to reduce the multiple testing burden and focus on terms that communicate a specific mechanism or cellular subcomponent. For large cohorts with sufficient power, we suggest including most or all of the hierarchy to provide a more complete picture. While the Gene Ontology overall provides a comprehensive model of the cell, we may still be missing important signal due to poor or missing annotations in the ontology. We examined how many genes were excluded for different term size cutoffs and found that 20% of genes were not annotated to a term with 100 or fewer genes (Figure S4). This is evidenced by the lack of significant terms annotated with the FAM13A gene in chromosome 4, despite there being two genome-wide significant SNPs within 10 kb of FAM13A. Increasing the maximum gene set size would include more genes but at the expense of greater multiple testing burden and may be worthwhile in larger studies.

Computing the low-dimensional representation of the SNP matrices in an unsupervised manner using PCA (i.e. not including any clinical variables) makes it simple to adjust for confounding variables and avoid the use of permutation tests for estimating significance. It also allows testing for association with different clinical variables without recomputing the principal components. One major disadvantage is, however, that we are unlikely to discover signal in large gene sets. The use of principal component analysis may also favor SNPs with greater minor allele frequency due to contributing more variance to the SNP matrix.

Of the 27 terms reported by HiGAna, 16 terms validated nominally ($p < 0.05$) and 8 replicated after Bonferroni correction ($p < 0.05/27$). A major limitation of the validation analysis was that we did not have a validation cohort with individual-level genotype data and thus had to validate using the UK Biobank summary statistics with MAGMA. This may in part explain why many terms did not replicate considering that MAGMA reported relatively few significant terms in the discovery cohort as well. Replication in an individual-level data set will be necessary to meaningfully validate our results.

Methods

Genotype data preparation

We obtained genotype data from 6670 non-Hispanic white subjects from the COPDGene Study. The study enrolls smokers (minimum 10 pack-years) with and without COPD. We refer to [26] for a detailed description of the study design. We removed all variants missing in more than 5% of samples and all samples missing genotypes for more than 5% of variants. Furthermore, we removed variants with a Hardy-Weinberg equilibrium exact test $p < 10^{-6}$ or a minor allele frequency below 0.01. The resulting data set contained 630,369 variants and all 6670 subjects. Case individuals had GOLD grade 2-4 (n=2812) and control individuals had GOLD grade 0 (n=2534), as defined by the Global Initiative for Chronic Obstructive Lung Disease (GOLD) [6]. Individuals with GOLD grade 1 were excluded from the case-control analysis (n=1324).

Preparation of ontology

We built an annotated ontology from the Gene Ontology (GO) knowledge base [34]. We obtained the set of terms and relations from GO (release 2019-01-19) and built a hierarchy of terms using all "is a" and "part of" relationships reported in GO. The three aspects of GO (molecular function, cellular component and biological process) were combined under a common root to construct a single, unified hierarchy encompassing all of GO.

Sets of genes annotated to each term were also obtained from GO (release 2019-01-01). All evidence codes except for "inferred from electronic annotation" (IEA) were used. Annotations were propagated upwards through the hierarchy by adding all annotations of each term to its parents starting from the bottom. After this operation, the set of genes assigned to each term is the union of the genes annotated to that term in GO and all genes annotated to any of its descendants. From this it follows that the root term is assigned all genes with at least one annotation in GO.

After propagation, all terms without annotations were removed. Furthermore, we removed terms considered to be highly redundant with respect to their children using the following procedure: we defined the degree of redundancy of a parent-children relationship as $R(P, C) = |A_C| \,/\, |A_P|$, where A_P and A_C are the sets of genes annotated to the parent and child terms, respectively. All terms with redundancy ≥ 0.9 with respect to one or more children were collapsed by removing the term and connecting all its children to all its parents.

Computing term principal components

For each term, we built a SNP matrix G of all SNPs that were within 10 kb of any gene annotated to that term. Each SNP matrix was projected into its principal component (PC) decomposition $T = GW$ using singular value decomposition. Only the first k PCs were kept, where k is the number of PCs needed to explain 95% of the variance, limited to at most 50 PCs. The dimensionality reduction serves two purposes: by discarding the last 5% of variance, we account for LD between SNPs and avoid collinearity between predictors in the regression model. Secondly, it reduces loss of power when testing large SNP matrices due to having too many degrees of freedom.

In order to reduce the computation time for large data sets we also include the option to use randomized singular value decomposition [12] implemented in the *rSVD* R package [9]. This will drastically reduce the principal component computation at the cost of reduced accuracy.

Association analysis

The association analysis evaluates each term in the ontology for association with the outcome using a generalized linear model framework. For each term, we trained a model $f_1 : Y = \alpha T + \beta C + \epsilon$, where T is the principal component matrix, C is a set of optional covariates, α is the genetic effect, β is the effect of the covariates and ϵ is the vector of residuals. This model was compared to the null model $f_0 : Y = \beta C + \epsilon$ by evaluating the reduction in deviance from including the genetic term using an appropriate test statistic. Since the outcome variable in this study was binary, we compared logistic regression models using a χ^2-test. The covariate matrix included relevant covariates such as age and sex, as well as the first four principal componentscomputed from all variants in order to correct for population structure.

Conditional association tests

We implemented two conditional association tests for identifying terms that were significant due to a single gene or child term. For each term, we identified a set of SNPs to be excluded and then repeated the association analysis with those SNPs removed from the genotype matrix. For the "top child" test we identified the most significant child term for each term under consideration and excluded SNPs near genes annotated to the top child. For the "top gene" test we computed an association score for each gene by repeating the association where each gene set is a single gene (Table S2). We then identified the most significant gene for each term and excluded all SNPs near that gene. In cases where a SNP is near both a gene identified for removal and a gene that should be kept, we chose to be strict and always exclude the SNP regardless.

Computing SNP effect sizes

Given the SNP matrix G for a term, its PC decomposition $T = GW$, and a corresponding regression model $Y = \alpha T_k + \beta C + \epsilon$, one can compute a SNP effect vector $E = W_k\alpha$, where W is a matrix whose columns are the first k eigenvectors of $T^\intercal T$. The vector E captures the total effect contributed by each SNP while accounting for LD.

Genome-wide association analysis

The standard genome-wide association analysis was carried out using the PLINK (v1.90b3.31) logistic regression method (`--logistic`). Genotypes were included as an allelic dosage additive effect. Sex, age and the first four principal components were included as additional covariates.

References

[1] Davies Adeloye et al. "Global and regional estimates of COPD prevalence: Systematic review and meta-analysis". In: *Journal of Global Health* 5.2 (2015). DOI: 10.7189/jogh.05.020415.

[2] Edson X. Albuquerque et al. "Mammalian Nicotinic Acetylcholine Receptors: From Structure to Function". In: *Physiological Reviews* 89.1 (Jan. 2009), pp. 73–120. DOI: 10.1152/physrev.00015.2008.

[3] Laura Taylor Alto and Jonathan R. Terman. "Semaphorins and their Signaling Mechanisms". In: *Methods in Molecular Biology*. Springer New York, Oct. 2016, pp. 1–25. DOI: 10.1007/978-1-4939-6448-2_1.

[4] Soumyaroop Bhattacharya et al. "Molecular Biomarkers for Quantitative and Discrete COPD Phenotypes". In: *American Journal of Respiratory Cell and Molecular Biology* 40.3 (Mar. 2009), pp. 359–367. DOI: 10.1165/rcmb.2008-0114oc.

[5] Christopher C. Chang et al. "Second-generation PLINK: rising to the challenge of larger and richer datasets". In: *GigaScience* 4.1 (2015). DOI: 10.1186/s13742-015-0047-8.

[6] Global Initiative for Chronic Obstructive Lung Disease. *2019 Global Strategy for Prevention, Diagnosis and Management of COPD*. 2019.

[7] Marc Decramer, Wim Janssens, and Marc Miravitlles. “Chronic obstructive pulmonary disease”. In: *The Lancet* 379.9823 (2012), pp. 1341–1351. DOI: 10.1016/s0140-6736(11)60968-9.

[8] Dawn L. DeMeo et al. “Integration of Genomic and Genetic Approaches Implicates IREB2 as a COPD Susceptibility Gene”. In: *The American Journal of Human Genetics* 85.4 (Oct. 2009), pp. 493–502. DOI: 10.1016/j.ajhg.2009.09.004.

[9] N. Benjamin Erichson et al. “Randomized matrix decompositions using R”. In: *arXiv preprint* (2016). DOI: 10.18637/jss.v089.i11.

[10] Marilyn G. Foreman, Michael Campos, and Juan C. Celedón. “Genes and Chronic Obstructive Pulmonary Disease”. In: *Medical Clinics of North America* 96.4 (2012), pp. 699–711. DOI: 10.1016/j.mcna.2012.02.006.

[11] Andrew J. Ghio et al. “Particulate Matter in Cigarette Smoke Alters Iron Homeostasis to Produce a Biological Effect”. In: *American Journal of Respiratory and Critical Care Medicine* 178.11 (Dec. 2008), pp. 1130–1138. DOI: 10.1164/rccm.200802-334oc.

[12] N. Halko, P. G. Martinsson, and J. A. Tropp. “Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions”. In: *SIAM Review* 53.2 (2011), pp. 217–288. DOI: 10.1137/090771806.

[13] Megan Hardin et al. “CHRNA3/5,IREB2, and ADCY2 Are Associated with Severe Chronic Obstructive Pulmonary Disease in Poland”. In: *American Journal of Respiratory Cell and Molecular Biology* 47.2 (Aug. 2012), pp. 203–208. DOI: 10.1165/rcmb.2012-0011oc.

[14] Peter Holmans et al. “Gene Ontology Analysis of GWA Study Data Sets Provides Insights into the Biology of Bipolar Disorder”. In: *The American Journal of Human Genetics* 85.1 (2009), pp. 13–24. DOI: 10.1016/j.ajhg.2009.05.011.

[15] Truls Ingebrigtsen et al. “Genetic influences on chronic obstructive pulmonary disease – A twin study”. In: *Respiratory Medicine* 104.12 (2010), pp. 1890–1895. DOI: 10.1016/j.rmed.2010.05.004.

[16] Gleb Kichaev et al. “Leveraging Polygenic Functional Enrichment to Improve GWAS Power”. In: *The American Journal of Human Genetics* 104.1 (Jan. 2019), pp. 65–75. DOI: 10.1016/j.ajhg.2018.11.008.

[17] Ho Cheol Kim, Mahroo Mofarrahi, and Sabah N. A. Hussain. “Skeletal muscle dysfunction in patients with chronic obstructive pulmonary disease.” In: *International journal of chronic obstructive pulmonary disease* 3 (4 2008), pp. 637–658. ISSN: 1176-9106.

[18] Phil H. Lee et al. “INRICH: interval-based enrichment analysis for genome-wide association studies”. In: *Bioinformatics* 28.13 (2012), pp. 1797–1799. DOI: 10.1093/bioinformatics/bts191.

[19] Christiaan A. de Leeuw et al. “MAGMA: Generalized Gene-Set Analysis of GWAS Data”. In: *PLOS Computational Biology* 11.4 (2015). Ed. by Hua Tang, e1004219. DOI: 10.1371/journal.pcbi.1004219.

[20] Jianzhu Ma et al. “Using deep learning to model the hierarchical structure and function of a cell”. In: *Nature Methods* 15.4 (2018), pp. 290–298. DOI: 10.1038/nmeth.4627.

[21] GBD 2013 Mortality and Causes of Death Collaborators. "Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013". In: *The Lancet* 385.9963 (2015), pp. 117–171. DOI: 10.1016/s0140-6736(14)61682-2.

[22] Michael E. Nelson, Amy R. O'Brien-Ladner, and Lewis J. Wesselius. "Regional variation in iron and iron-binding proteins within the lungs of smokers." In: *American Journal of Respiratory and Critical Care Medicine* 153.4 (Apr. 1996), pp. 1353–1358. DOI: 10.1164/ajrccm.153.4.8616566.

[23] Emelie Perland and Robert Fredriksson. "Classification Systems of Secondary Active Transporters". In: *Trends in Pharmacological Sciences* 38.3 (Mar. 2017), pp. 305–315. DOI: 10.1016/j.tips.2016.11.008.

[24] Sreekumar G. Pillai et al. "A Genome-Wide Association Study in Chronic Obstructive Pulmonary Disease (COPD): Identification of Two Major Susceptibility Loci". In: *PLoS Genetics* 5.3 (Mar. 2009). Ed. by Mark I. McCarthy, e1000421. DOI: 10.1371/journal.pgen.1000421.

[25] Shaun Purcell et al. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses". In: *The American Journal of Human Genetics* 81.3 (2007), pp. 559–575. DOI: 10.1086/519795.

[26] Elizabeth A. Regan et al. "Genetic Epidemiology of COPD (COPDGene) Study Design". In: *COPD: Journal of Chronic Obstructive Pulmonary Disease* 7.1 (Mar. 2010), pp. 32–43. DOI: 10.3109/15412550903499522.

[27] Nancy L. Saccone et al. "Multiple distinct risk loci for nicotine dependence identified by dense coverage of the complete family of nicotinic receptor subunit (CHRN) genes". In: *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 150B.4 (June 2009), pp. 453–466. DOI: 10.1002/ajmg.b.30828.

[28] Nancy L. Saccone et al. "Multiple Independent Loci at Chromosome 15q25.1 Affect Smoking Quantity: a Meta-Analysis and Comparison with Lung Cancer and COPD". In: *PLoS Genetics* 6.8 (Aug. 2010). Ed. by Greg Gibson, e1001053. DOI: 10.1371/journal.pgen.1001053.

[29] Phuwanat Sakornsakolpat et al. "Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations". In: *Nature Genetics* 51.3 (2019), pp. 494–505. DOI: 10.1038/s41588-018-0342-2.

[30] Ayellet V. Segré et al. "Common Inherited Variation in Mitochondrial Genes Is Not Enriched for Associations with Type 2 Diabetes or Related Glycemic Traits". In: *PLoS Genetics* 6.8 (2010). Ed. by Peter M. Visscher, e1001058. DOI: 10.1371/journal.pgen.1001058.

[31] Nick Shrine et al. "New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries". In: *Nature Genetics* 51.3 (2019), pp. 481–493. DOI: 10.1038/s41588-018-0321-7.

[32] V. L. Stevens et al. "Nicotinic Receptor Gene Variants Influence Susceptibility to Heavy Smoking". In: *Cancer Epidemiology Biomarkers & Prevention* 17.12 (Dec. 2008), pp. 3517–3525. DOI: 10.1158/1055-9965.epi-08-0585.

[33] Peter H. Sudmant et al. “A global reference for human genetic variation”. In: *Nature* 526.7571 (2015), pp. 68–74. DOI: 10.1038/nature15393.

[34] The Gene Ontology Consortium. “The Gene Ontology Resource: 20 years and still GOing strong”. In: *Nucleic Acids Research* 47.D1 (2018), pp. D330–D338. DOI: 10.1093/nar/gky1055.

[35] Thorgeir E. Thorgeirsson et al. “A variant associated with nicotine dependence, lung cancer and peripheral arterial disease”. In: *Nature* 452.7187 (Apr. 2008), pp. 638–642. DOI: 10.1038/nature06846.

[36] Jemma B. Wilk et al. “Genome-Wide Association Studies Identify CHRNA5/3 and HTR4 in the Development of Airflow Obstruction”. In: *American Journal of Respiratory and Critical Care Medicine* 186.7 (Oct. 2012), pp. 622–632. DOI: 10.1164/rccm.201202-0366oc.

[37] Umar Yazdani and Jonathan R. Terman. “The semaphorins”. In: *Genome Biology* 7.3 (2006), p. 211. DOI: 10.1186/gb-2006-7-3-211.

[38] Michael Ku Yu et al. “Translation of Genotype to Phenotype by a Hierarchy of Cell Subsystems”. In: *Cell Systems* 2.2 (2016), pp. 77–88. DOI: 10.1016/j.cels.2016.02.003.