Supplementary materials for 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

S1 Supplementary text

S1.1 Gene set enrichment using Fisher's exact test

Let S be the set of selected candidate genes, T be the set of genes annotated to some term (gene set) we want to investigate and U be the set of all genes. From these three sets, one can create a contingency table (Table S1). Given a 2×2 contingency table. The *marginal values* are the totals in the rightmost column and bottommost row. Fisher's exact test computes probability of obtaining such a table provided there is no association between the two sets S and T,

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}.$$

To compute the significance of overrepresentation, we compute the sum of probabilities for all contingency tables where the degree of overlap is greater than or equal to $|S \cap T|$ contingency table as well as every other contingency table that has a greater overlap between S and T while preserving the marginal values.

Table S1: Contingency table for a candidate gene set S and a term T from universe U.

	$\in T$	$\notin T$	Total
$\in S$	$a = \lvert S \cap T \rvert$	$b = \lvert S \setminus T \rvert$	$a + b$
$\notin S$	$c = \lvert T \setminus S \rvert$	$d = \lvert U \setminus (S \cup T) \rvert$	$c + d$
Total	$a + c$	$b + d$	$n = a + b + c + d$

S2 Supplementary figures

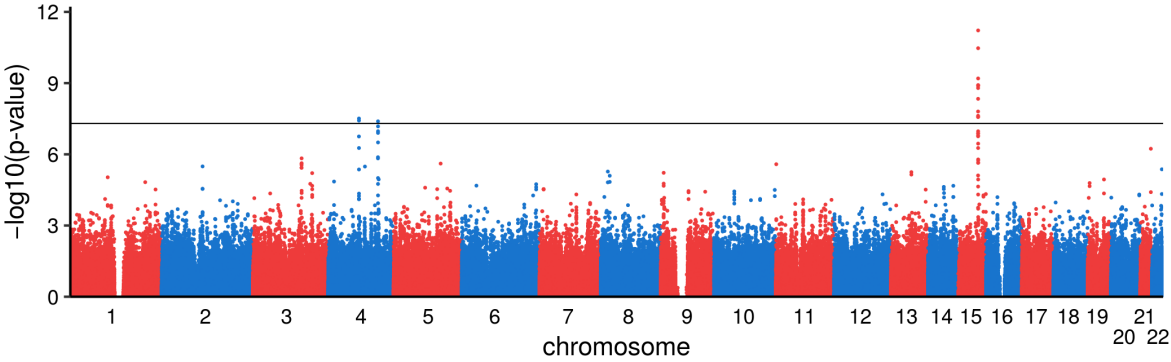


Figure S1: Manhattan plot of case-control association analysis of COPDGene genotype data. Horizontal line denotes $p = 5 \times 10^{-8}$.

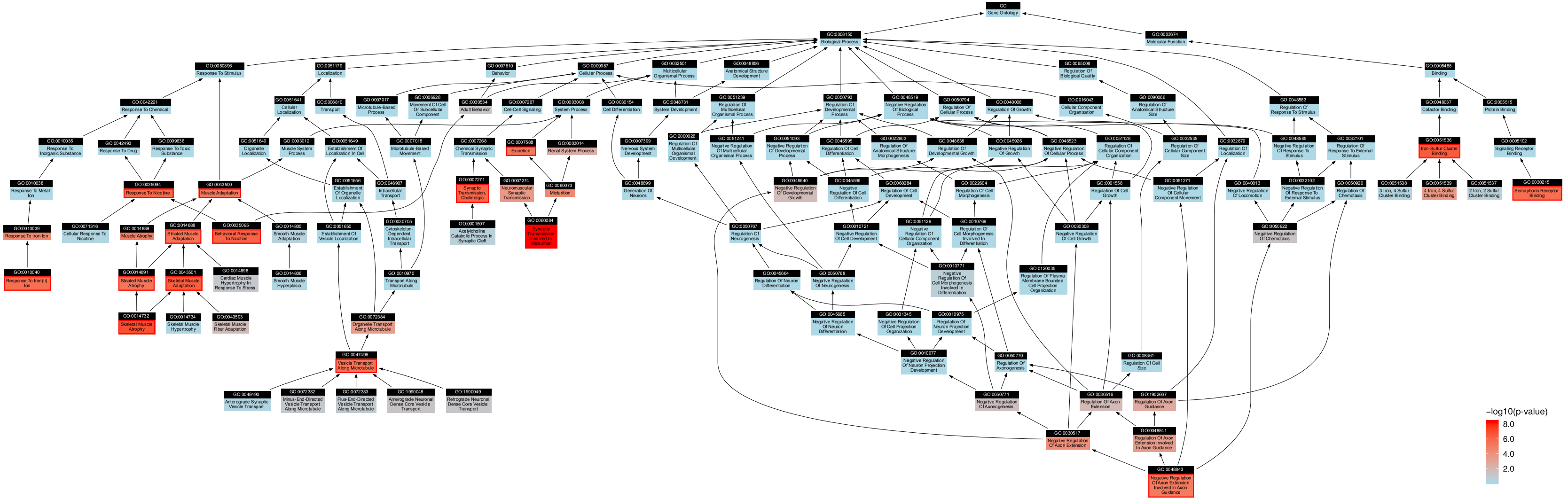
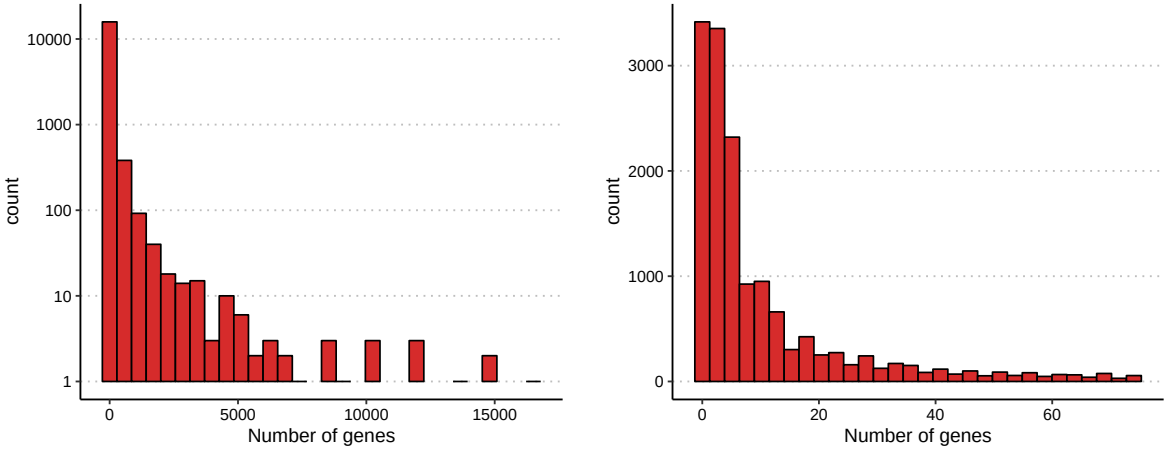


Figure S2: All significantly enriched terms (FDR < 0.05), all their ancestors and their immediate children. Node color denotes unadjusted p-value.



(a) (b)

Figure S3: Distribution of term size of terms in the hierarchy constructed from Gene Ontology. (a) All terms. Horizontal axis is log-scaled. (b) Restricted to terms of up to 75 genes.

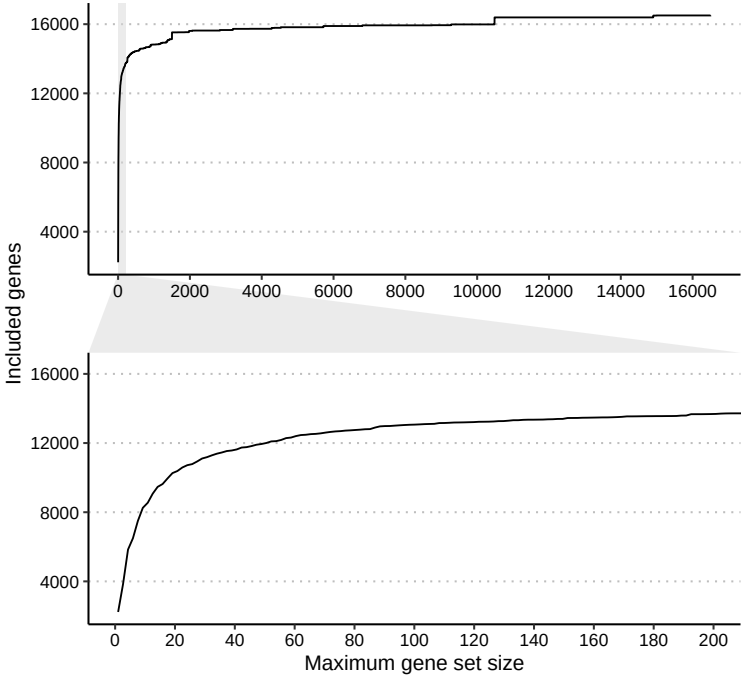
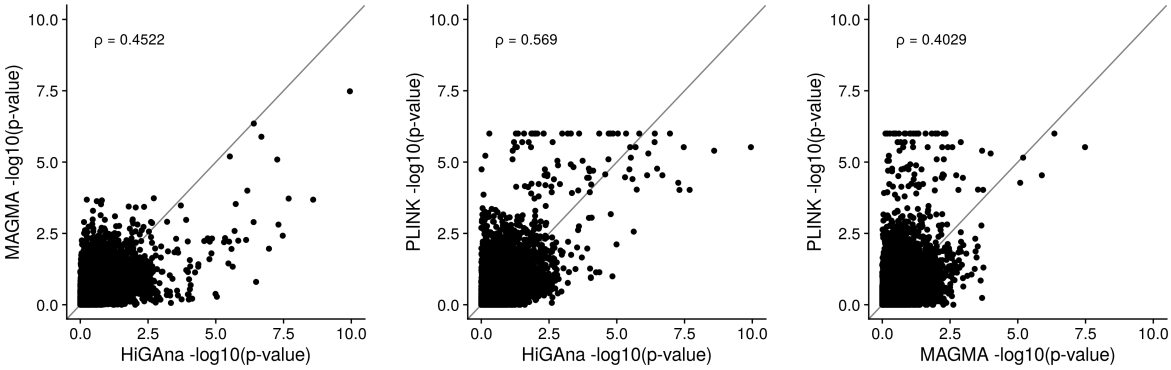


Figure S4: Number of unique genes included in the Gene Ontology-based gene hierarchy when only including terms up to certain size (in terms for genes).



(a) (b) (c)

Figure S5: Comparison of -log10 p-values from enrichment analysis of 9312 terms using HiGAna, MAGMA and PLINK. (a) Comparison between HiGAna and MAGMA. (b) Comparison between HiGAna and PLINK. (c) Comparison between MAGMA and PLINK. ρ values are the Pearson correlation coefficients between the -log10 p-values.