

Characterization of a COPD-Associated *NPNT* Functional Splicing Genetic Variant in Human Lung Tissue via Long-Read Sequencing

Aabida Saferali¹, Zhonghui Xu¹, Gloria M. Sheynkman^{2,3,4}, Craig P. Hersh^{1,5}, Michael H. Cho^{1,5}, Edwin K. Silverman^{1,5}, Alain Laederach⁶, Christopher Vollmers⁷, Peter J. Castaldi^{1,8}

¹Channing Division of Network Medicine, Brigham and Women's Hospital

²Department of Molecular Physiology and Biological Physics, University of Virginia

³Center for Public Health Genomics, University of Virginia

⁴UVA Cancer Center, University of Virginia

⁵Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital

⁶Department of Biology, University of North Carolina at Chapel Hill

⁷Department of Biomolecular Engineering, Cellular, Cellular, and Developmental Biology, University of California Santa Cruz

⁸Division of General Medicine and Primary Care, Brigham and Women's Hospital

Abstract:

Chronic obstructive pulmonary disease (COPD) is a leading cause of death worldwide. Genome-wide association studies (GWAS) have identified over 80 loci that are associated with COPD and emphysema, however for most of these loci the causal variant and gene are unknown. Here, we utilize lung splice quantitative trait loci (sQTL) data from the Genotype-Tissue Expression project (GTEx) and short read sequencing data from the Lung Tissue Research Consortium (LTRC) to characterize a locus in nephronectin (*NPNT*) associated with COPD case-control status and lung function. We found that the rs34712979 variant is associated with alternative splice junction use in *NPNT*, specifically for the junction connecting the 2nd and 4th exons (chr4:105898001-105927336) ($p=4.02 \times 10^{-38}$). This association colocalized with GWAS data for COPD and lung spirometry measures with a posterior probability of 94%, indicating that the same causal genetic variants in *NPNT* underlie the associations with COPD risk, spirometric measures of lung function, and splicing. Investigation of *NPNT* short read sequencing revealed that rs34712979 creates a cryptic splice acceptor site which results in the inclusion of a 3 nucleotide exon extension, coding for a serine residue near the N-terminus of the protein. Using Oxford Nanopore Technologies (ONT) long read sequencing we identified 13 *NPNT* isoforms, 6 of which are predicted to be protein coding. Two of these are full length isoforms which differ only in the 3 nucleotide exon extension whose occurrence differs by genotype. Overall, our data indicate that rs34712979 modulates COPD risk and lung function by creating a novel splice

acceptor which results in the inclusion of a 3 nucleotide sequence coding for a serine in the nephronectin protein sequence. Our findings implicate *NPNT* splicing in contributing to COPD risk, and identify a novel serine insertion in the nephronectin protein that warrants further study.

1 **Introduction**

2 The development of chronic obstructive pulmonary disease (COPD) is influenced by
3 genetic susceptibility factors in addition to environmental exposures. Recent genome-wide
4 association studies (GWAS) have identified over 80 distinct genetic loci that influence
5 susceptibility to COPD and emphysema¹, but for most of these loci the causal genetic variants
6 and effector genes are unknown. By identifying the functional genetic variants in these GWAS
7 loci and elucidating the biological mechanisms through which they influence disease
8 susceptibility, causal mechanisms of COPD may be discovered.

9 The large majority of causal GWAS variants reside in the non-coding genome and disrupt
10 gene regulatory elements. As a result, expression quantitative trait locus (eQTL) studies that
11 associate genetic variants to gene expression values have been used to identify functional gene
12 targets of GWAS-identified loci. In COPD, this approach has identified putative causal variants
13 affecting the expression of *HHIP*², *FAM13A*³, *TGFB2*⁴, and *ACVR1B*⁵. However, eQTL studies
14 do not capture all of the potentially relevant functional mechanisms through which causal
15 variants may alter gene expression. In particular, the alteration of gene splicing and isoform
16 ratios is an important disease-causing gene regulatory mechanism that is not well captured by
17 gene-level eQTL analyses^{6,7}.

18 A genome-wide significant and replicated genetic association signal for respiratory
19 phenotypes near *NPNT* may harbor a causal splicing variant, because the pattern of association at
20 this locus is characterized by a single, clear lead SNP association that is located 5 nucleotides
21 upstream from the 5' splice site of the second exon in *NPNT*. We hypothesized that this region
22 contains a functional variant that alters *NPNT* splicing. We utilized short and long-read RNA
23 sequencing from human lung tissues to identify a causal splicing variant, rs34712979; to catalog
24 the isoform variability of *NPNT* in the human lung; and to characterize the effects of rs34712979
25 on isoform usage rates.

26 **Methods**

27 *Genetic association analysis, splicing QTL, and colocalization results*

28 Summary genome-wide association study (GWAS) statistics were used from our previous
29 study of COPD¹ and two lung function phenotypes, FEV₁ and FEV₁/FVC
30 (<http://ldsc.broadinstitute.org/ldhub/>)⁸. Leafcutter splicing QTL (sQTL) significant results⁹ were

32 obtained for all tissues from the GTEx Portal (<https://www.gtexportal.org/home/datasets>), and
33 complete lung sQTL results were obtained from the Anvil GTEx Terra workspace. Multiple
34 colocalization for GWAS and sQTL results was performed using the moloc R package
35 (<https://github.com/clagiamba/moloc>). Moloc analyses were performed using default parameters
36 for prior variance of the approximate Bayes factor (ABF, $\text{prior_var} = c(0.01, 0.1, 0.5)$) and
37 default parameters for the prior likelihood that a given SNP is causal for one trait, pairs of traits,
38 or all traits ($\text{priors} = c(1e-04, 1e-06, 1e-07)$). Linkage disequilibrium (LD), evolutionary
39 conservation, and overlap with regulatory elements were identified for individual single
40 nucleotide polymorphisms (SNPs) with Haploreg
41 (<https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php>). Posterior causal probabilities
42 based on strength of genetic association and local LD patterns were determined using the PICS
43 algorithm (<https://pubs.broadinstitute.org/pubs/finemapping/pics.php>).¹⁰

45 *Splicing analysis in short read RNA-seq data from GTEx lung tissue samples*

46 With dbGaP approval, RNA-seq BAM files and whole genome sequencing VCF files for
47 lung tissue samples in GTEx V8 release were accessed on Google Cloud via the AnVIL GTEx
48 Terra workspace (https://app.terra.bio/#workspaces/anvil-datastorage/AnVIL_GTEx_V8_hg38).
49 To obtain splicing junctional counts for each genotype and visualize with Sashimi plots, a
50 Docker image (<https://hub.docker.com/repository/docker/pacifly/splice-plot-app>) was built on
51 the Python package SplicePlot¹¹ (<https://github.com/wueric/SplicePlot>) and its dependencies, and
52 a WDL workflow
53 (https://portal.firecloud.org/?return=terra#methods/zxu_spliceplot/spliceplot/18) was created and
54 executed via Cromwell engine to run SplicePlot functionalities on Google Cloud. The splicing
55 junctional count method in SplicePlot was adapted to accommodate novel junctions missed from
56 the alignment.

58 *Lung Tissue Research Consortium Samples, short-read RNA sequencing, and whole genome 59 sequencing*

60 The Lung Tissue Research Consortium (LTRC) is an NHLBI-sponsored collection of
61 lung and blood tissues collected from patients undergoing thoracic surgery who completed a
62 standard questionnaire, pulmonary function testing, and chest computed tomography (CT)

63 imaging. Through the NHLBI Trans-Omics and Precision Medicine (TOPMed) program, LTRC
64 whole-genome sequencing (WGS) data were generated at Broad Genomics and lung RNA-seq
65 were generated at the University of Washington, and 1,335 LTRC samples with RNA-seq and
66 WGS passing quality control filters were analyzed in this study. Briefly, RNA-seq data were
67 generated using paired-end Illumina sequencing from poly-A selected libraries to an average
68 depth of 67 million mapped reads. Methods for TOPMed WGS are available at
69 <https://www.nhlbiwgs.org/topmed-whole-genome-sequencing-methods-freeze-8>.

70

71 *Long read RNA-seq analysis in human lung samples from the LTRC*

72 We conducted targeted Oxford Nanopore Technologies (ONT) long read sequencing on
73 RNA from 10 human lung samples from the LTRC which were selected to include five samples
74 from each homozygous class (i.e., GG genotype, AA genotype) of rs34712979. For each of these
75 ten samples, 100-200ng of total RNA was used to generate full-length cDNA using a modified
76 Smart-seq2 protocol¹². The enrichment and library generation procedures are described in detail
77 in the Supplemental Methods.

78 Enriched re-amplified cDNA from two independent enrichment reactions was sequenced
79 on a MinION 9.4.1 flow cell or a MinION 10.3 flow cell using the R2C2 method¹³⁻¹⁶. For each
80 run, 1ug of DNA was prepared using the LSK-109 kit according to the manufacturer's
81 instructions with only minor modifications. End-repair and A-tailing steps were both extended
82 from 5 minutes to 30 minutes. The final ligation step was also extended to 30 minutes. Each run
83 took 48 hours and the resulting data in Fast5 format was basecalled using the high accuracy
84 model of the gpu accelerated Guppy algorithm (9.4.1 flow cell: version 3.4.5+fb1fbfb with
85 config file dna_r9.4.1_450bps_hac.cfg config file; 10.3 flow cell: version 3.6.1+249406c with
86 config file dna_r10_450bps_hac.cfg). To generate R2C2 consensus reads for each sample, we
87 processed and demultiplexed the resulting raw reads using our C3POa pipeline
88 (<https://github.com/rvolden/C3POa>). R2C2 reads were analyzed to identify and quantify
89 isoforms using version 3.5 of Mandalorion (-O 0,40,0,40 -r 0.01 -i 1 -w 1 -n 2 -R 5)
90 (<https://github.com/rvolden/Mandalorion-Episode-III>). Isoforms were categorized using the
91 sqanti_qc.py script of the SQANTI¹⁷ program with slight modifications to make it compatible
92 with Python3.

93 For *NPNT* isoform quantification (i.e., usage analysis), the proportion of isoform usage
94 for each sample was calculated by dividing the number of reads for each isoform by the total
95 number of isoform reads aligning to the *NPNT* locus. Differences in isoform usage between
96 genotype classes were identified using the Mann-Whitney test. For certain analyses, isoform
97 reads were collapsed based on whether they contained a 3 nucleotide exon extension. To
98 quantify the number of reads containing this TAG sequence at the 5' end of the second exon, we
99 extracted the following 30-mers from the fasta files for each sample using the 'grep' command:

100

101 AGTTCGACGGGAGTAGGTGGCCCAGGCAA

102 CGAGTTCGACGGGAGGTGGCCCAGGCAAAT

103

104 *NPNT* protein sequence analysis

105 Protein sequence analysis was performed using Uniprot¹⁸ to identify *NPNT* protein domains. The
106 Chou and Fasman Secondary Structure Prediction server¹⁹ was used to characterize the impact of
107 sequence changes to *NPNT* structure.

108

109 **Results**

110 *Genetic association signals for respiratory phenotypes in NPNT*

111 Genome-wide significant association signals near *NPNT* have been identified for COPD
112 and various measures of lung function, and previous colocalization and fine mapping analyses
113 have implicated rs34712979 as the most likely causal variant for the COPD association near
114 *NPNT*¹. We obtained the summary GWAS statistics for the most recent GWAS meta-analyses of
115 COPD, FEV₁, and FEV₁/FVC and compared the genetic association signal for each of these
116 phenotypes. In each case, the variant with minimal p-value was rs34712979, a common variant
117 with a minor allele frequency of 23% in the 1,000 Genomes European (EUR) population. Using
118 the PICS algorithm, we determined that the estimated posterior probability that rs34712979 was
119 the causal variant for each of these associations was between 94-100%. This variant is in strong
120 linkage disequilibrium with only one other common variant, rs6828309 ($r^2 = 0.81$ in EUR).
121 Supplemental Table 1 shows that rs34712979 is highly conserved and overlaps promoter and
122 enhancer elements in multiple cell types, which is not the case for rs6828309 (Supplemental

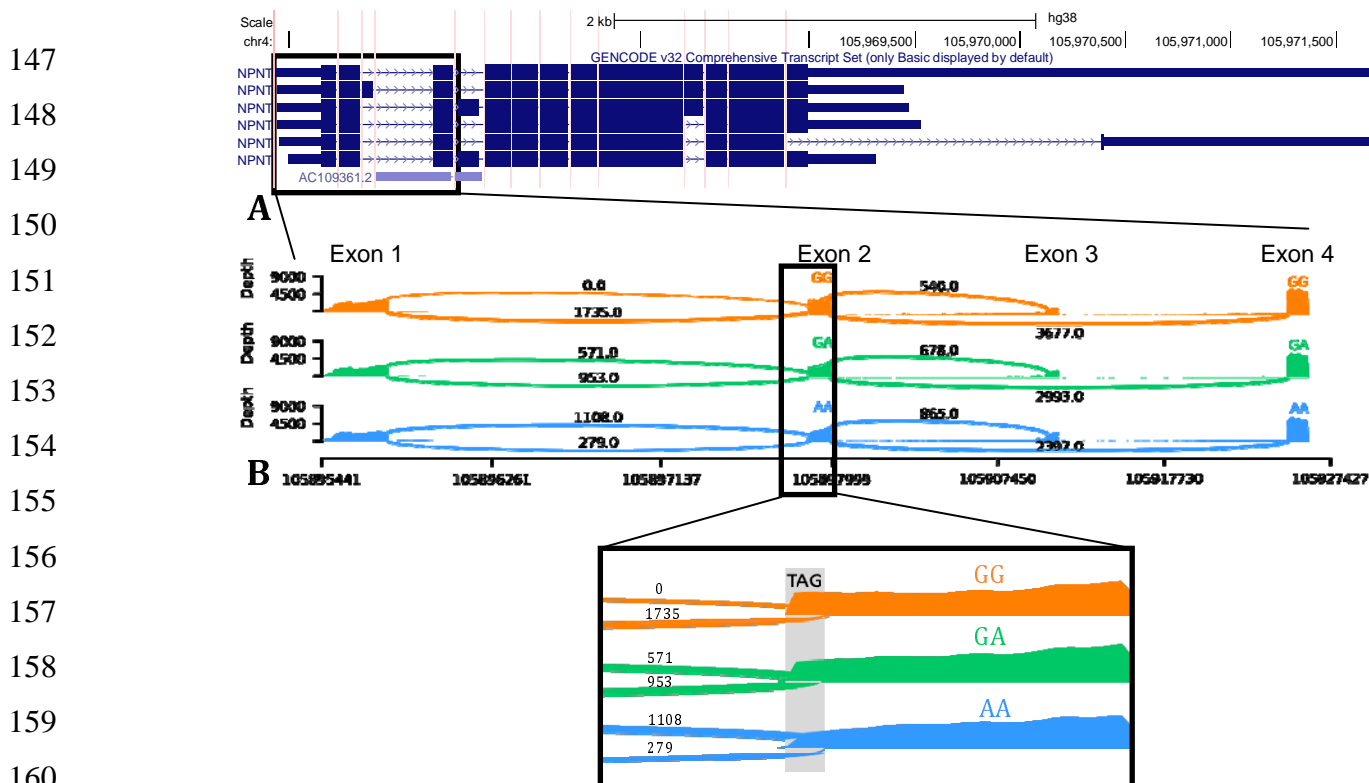
123 Table 1). The minor A allele of rs34712979 is associated with lower measures of lung function
124 and increased risk of COPD (odds ratio=1.18).

125

126 *An NPNT splicing association signal in human lung tissue colocalizes with GWAS associations*

127 To determine whether the genetic associations near *NPNT* to pulmonary phenotypes may
128 be explained by genetic effects on splicing, we examined leafcutter quantitative trait locus (QTL)
129 results from 49 tissues obtained from a total of 838 subjects in GTEx version 8. We observed
130 that rs34712979 was associated with splicing ratios for three exon-exon junctions in a total of 18
131 tissues that connect the first and second, second and fourth, and third and fourth exons of *NPNT*
132 (GENCODE v32) (Supplemental Table 2). Focusing on lung tissue, the most relevant tissue for
133 COPD, we observed that rs34712979 was associated with multiple splicing ratios at a nominal p-
134 value threshold of $p < 0.001$ (Supplemental Table 3), with the association for the junction
135 connecting the 2nd and 4th exons (chr4:105898001-105927336) exceeding genome-wide
136 significance. To better understand the effect of rs34712979 on splicing in human lung, we
137 identified 27 subjects in GTEx homozygous for the A allele of rs34712979 and compared the
138 distribution of junctional reads from lung RNA in these subjects against 27 randomly selected
139 samples from the other two genotype classes (Figure 1), and found that the A allele results in
140 higher rates of inclusion of the 3rd exon. Upon further investigation of junctional reads flanking
141 exon 2 we discovered the presence of a novel splice site acceptor at the 5' end of exon 2
142 associated with the A allele. This novel cryptic splice site was not detected in the sQTL analysis
143 as 99.6% of reads using the cryptic splice site were soft-clipped (ie. portions of the read were
144 masked due to a mismatch with the reference genome) and therefore were not included in the
145 Leafcutter clustering step.

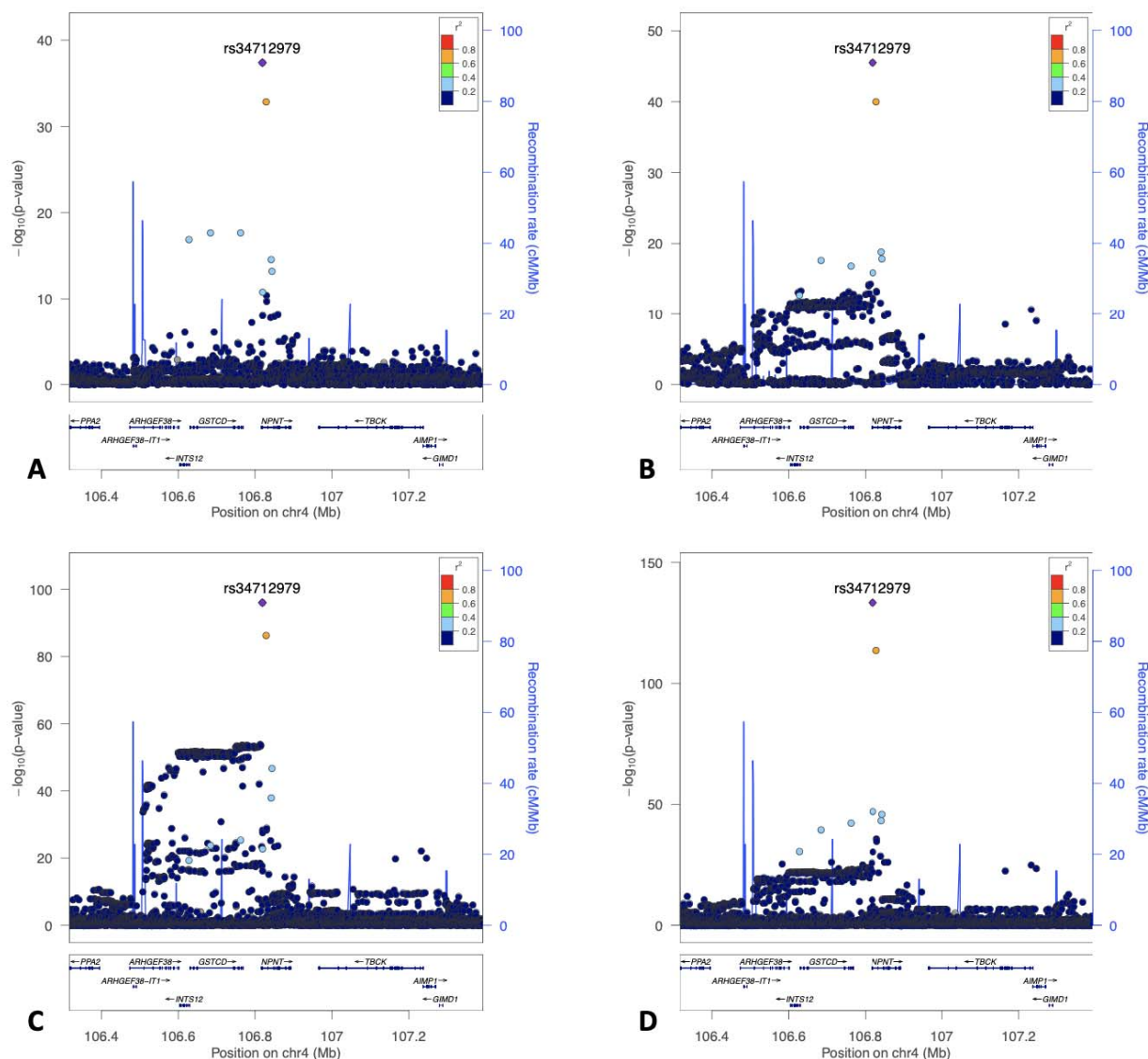
146



161 **Figure 1.** GTEX Leafcutter lung sQTL results for effect of rs34712979 on splice junctions
 162 involving the 2nd-4th exons of *NPNT* in leafcutter splicing analysis from short-read RNAseq in
 163 515 GTEX samples (A) and in junctional reads from 27 samples from each genotype class of
 164 rs34712979 (B). The 3 nucleotide alternatively spliced exonic extension sequence is shown in
 165 the inset window of panel B.

166
 167 To confirm that the lung sQTL signals in this region overlap with the GWAS association
 168 signals, we performed multiple colocalization using the moloc method which resulted in an
 169 estimated 94% probability of a shared causal variant for the three genetic association signals and
 170 the lung splicing signal for the 2nd and 4th exons of *NPNT* (chr4:105898001-105927336).
 171 rs34712979 had the best individual SNP posterior probability across all evaluated scenarios, with
 172 a posterior probability of 94% for the scenario of shared colocalization across all four datasets
 173 (Supplemental Table 4). The local association plots for each of the association signals near
 174 *NPNT* is shown in Figure 2.

175



176

177 **Figure 2.** Local association plots for genetic association near *NPNT* to COPD (A,

178 Sakornsakolpat 2019), FEV1 (B), FEV1/FVC (C, Shrine 2019), and GTEx Leafcutter lung sQTL

179 analyses (D).

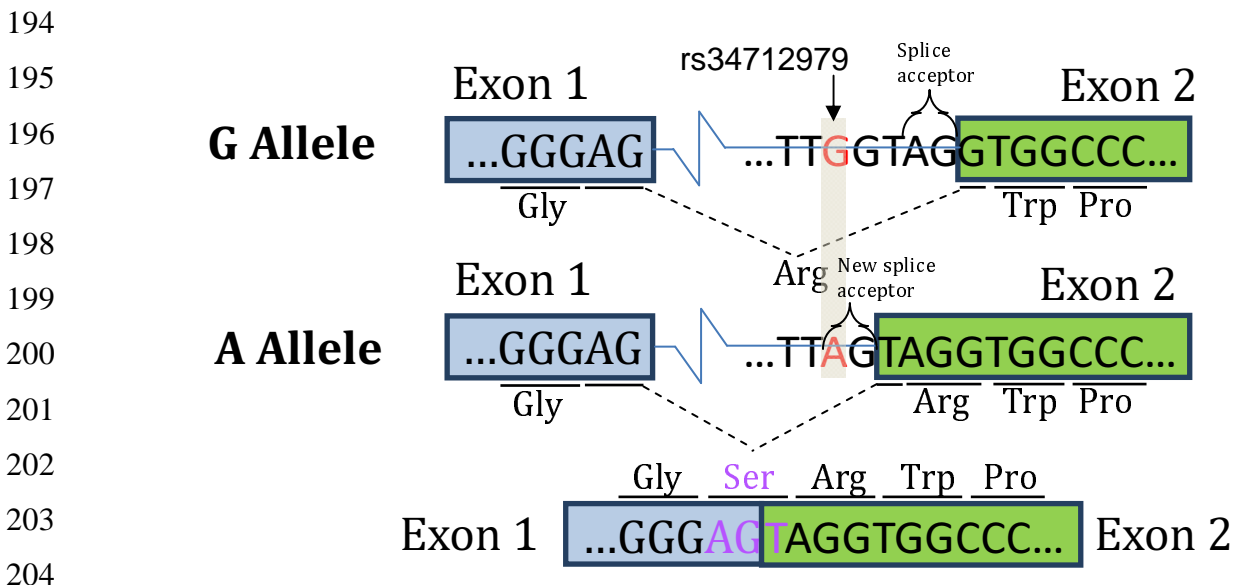
180

181 *rs34712979* creates an alternative splice acceptor site

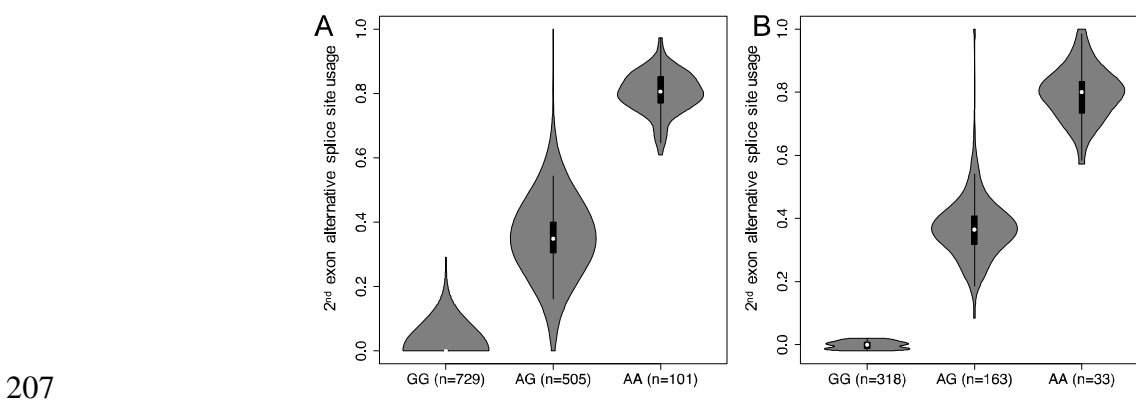
182 *rs34712979* is located 5bp upstream of the second exon of *NPNT*, so we examined the
183 sequence content in this region for motifs that may alter splicing, and we observed that the minor
184 A allele of *rs34712979* creates a cryptic AG splice acceptor that would result in a 3 nucleotide 5'
185 exon extension in exon 2. In other words, the A allele creates a NAGNAG splice site, which
186 contains adjacent AG acceptor sites that can be variably used^{20,21}. Open reading frame analysis

187 indicates that this results in an additional in-frame AGT codon, coding for serine, spanning the
 188 boundaries of the first and second exon (Figure 3).

189 To confirm this effect in human lung samples, we queried short-read RNA-seq data from
 190 human lung samples in GTEx and confirmed that samples with the A allele demonstrate
 191 preferential use of the cryptic versus the annotated splice acceptor site (Figure 4). We further
 192 confirmed that this phenomenon occurs in a larger set of 1,335 lung samples from the LTRC
 193 (Figure 4).



205 **Figure 3.** The A allele of rs34712979 creates an alternative splice acceptor site resulting in
 206 inclusion of a serine at the 5' splice site of exon 2.



208 **Figure 4.** Violin plot of exon 2 alternative splice site usage in short-read RNA-seq from A)
 209 LTRC and B) GTEx Lung Tissue Samples.

210
 211 *Characterization of rs34712979 Allelic Effects on NPNT Isoform Usage*

212 To determine the effect of rs34712979 on full-length NPNT isoforms, we performed
213 targeted enrichment for *NPNT* transcripts followed by ONT long-read sequencing in lung tissue
214 samples from the LTRC selected to include 5 subjects from each homozygote class of
215 rs34712979. The experiment yielded 24,747 reads mapping to *NPNT*. Thirteen high confidence
216 isoforms were detected, 12 of which were novel (Supplemental Table 1). Six of these isoforms,
217 including one antisense isoform, have open reading frames indicating that they have protein-
218 coding potential. Two protein coding isoforms containing exon 2 utilize the cryptic splice
219 acceptor, while two isoforms utilize the annotated splice acceptor (Figure 5, panel A).

220 Isoform usage differed between rs34712979 genotype classes. On a per-isoform basis, 9
221 of the 13 isoforms show differential usage between the two homozygous genotype classes
222 (Mann-Whitney $p < 0.05$). Focusing on the putative protein-coding isoforms, we found that the
223 full length annotated isoform ‘coding_5’ (corresponding to ENST00000379987.6) is more highly
224 expressed in the GG genotype class, while isoform ‘coding_1’, also full length, is highly
225 expressed in 4 out of 5 subjects with the AA genotype (Figure 6). The three short isoforms are
226 also differentially expressed by genotype, with isoforms ‘coding 3’ and ‘coding 4’ more highly
227 expressed in AA. Looking specifically at usage of the alternative or canonical splice acceptor
228 site at the 5’ end of exon 2, we observed 10 isoforms that include the second exon, with five
229 isoforms each using the alternative or canonical splice acceptor. Collapsing these isoforms by
230 splice site usage demonstrates markedly increased usage of the alternative splice site in the AA
231 genotype, with essentially no usage of the alternative site in 4 of the 5 GG genotype class
232 samples (Figure 6, panel E). We confirmed the presence of the 3-nucleotide 5’ exon extension
233 sequence in full length reads by counting the reads containing a 30-bp sequence centered on the
234 TAG triplet located after the novel splice junction and compared this to the number of reads
235 containing a corresponding 30-bp sequence without the TAG. Table 1 shows the clear
236 association between the AA genotype and the 3-nucleotide exon extension event.

237
238
239
240
241
242

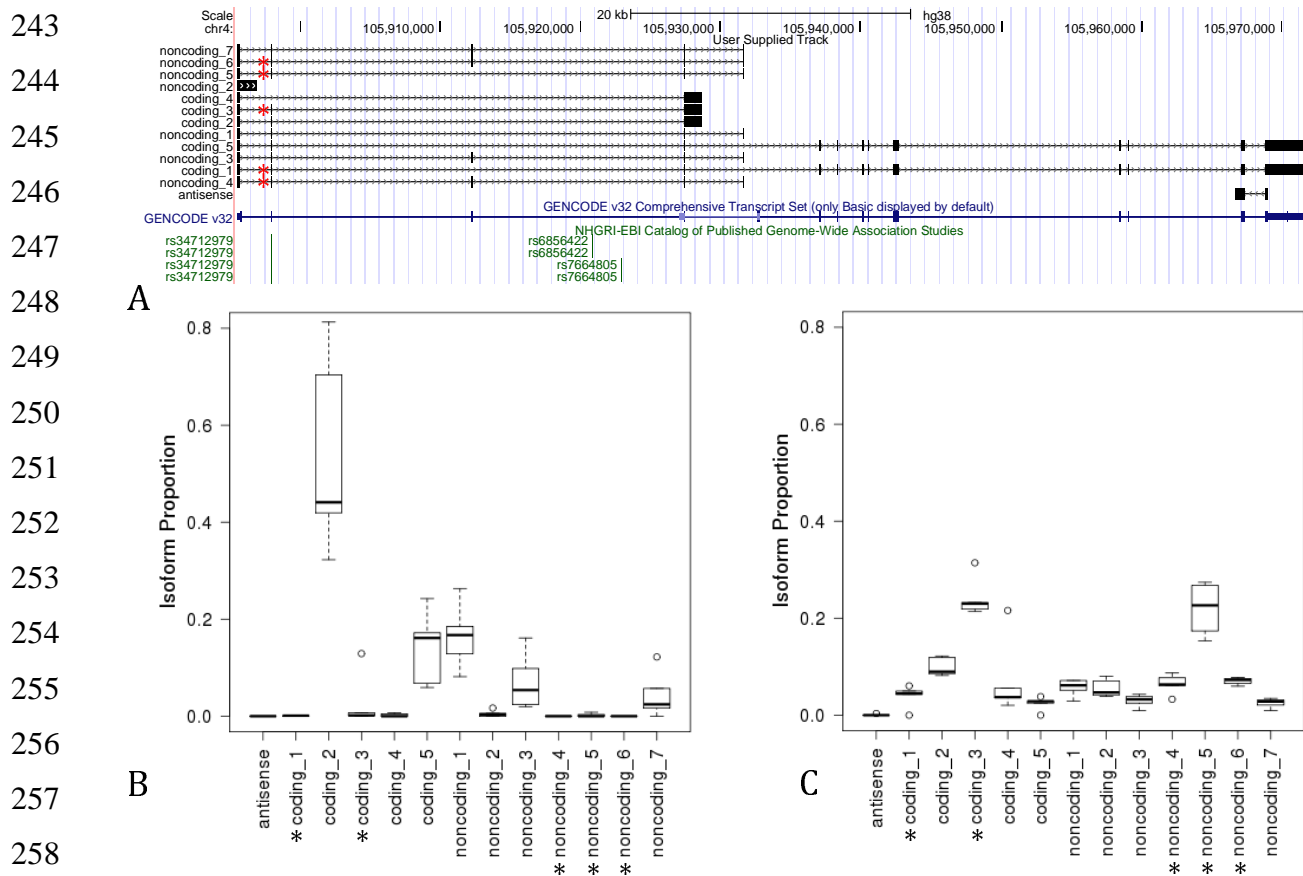
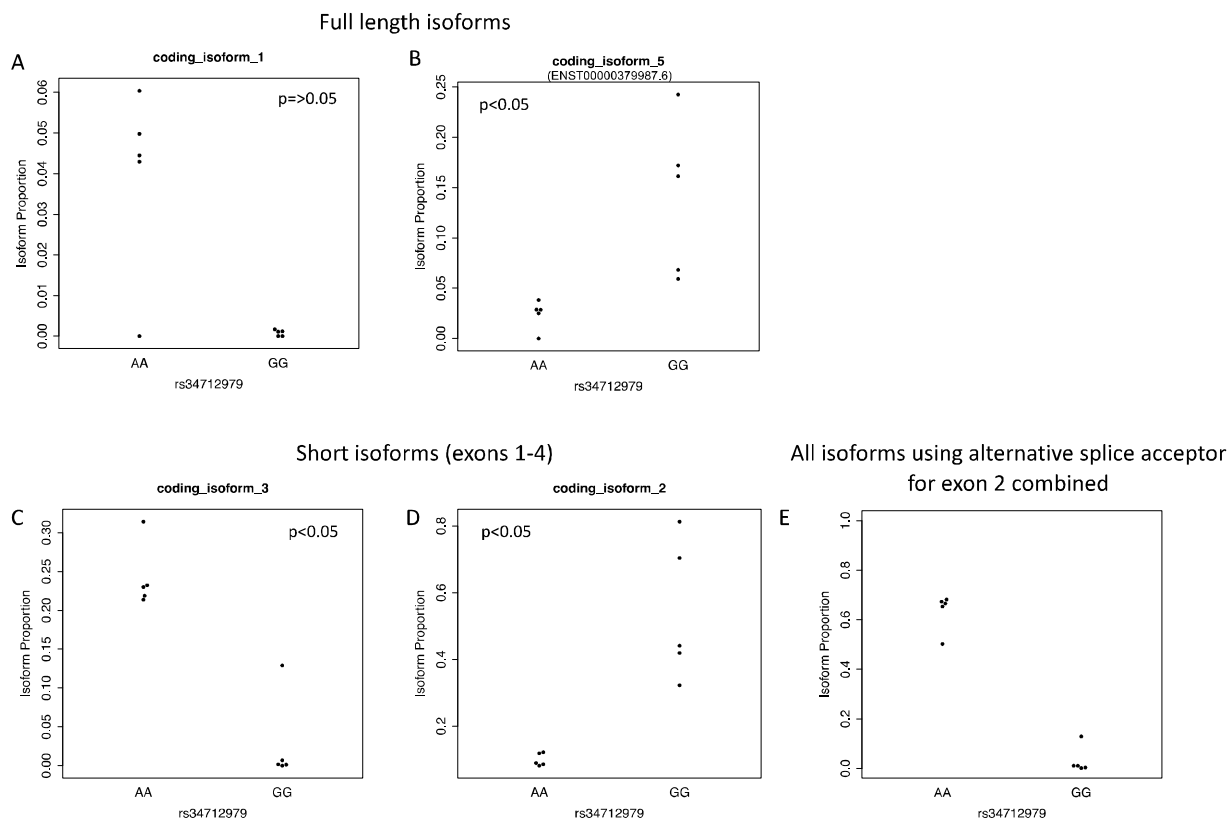


Figure 5. Isoforms detected by long read sequencing in 10 human lung tissues predicted to be coding or noncoding (A) * Indicates isoforms utilizing the novel splice acceptor site resulting in a three nucleotide 5' extension of exon 2. The isoform usage profile differs markedly between rs34712979 GG (B) and AA homozygotes (C).



274
275 **Figure 6.** Long read sequencing confirms the presence of two full length isoforms, one
276 incorporating a 3 bp intronic sequence (A) and the other which is fully annotated (B). Out of the
277 two short isoforms (coding isoforms 2 and 3) which contain exon 2, one includes the upstream
278 cryptic splice site and is more highly expressed in samples homozygous for the rs34712979 AA
279 genotype (C), while the other is predominantly expressed in the GG genotype (D). All isoforms
280 using the cryptic splice site combined have increased expression in the AA genotype with
281 minimal expression in GG.

282

283 *Protein sequence analysis*

284 The serine residue 5' exon extension is inserted 24 residues from the N-terminus of NPNT
285 protein, and four amino-acids after the end of the predicted signal peptide. Protein secondary
286 structure analysis of NPNT isoform sequences with and without the 3 nucleotide exon extension
287 revealed that the serine residue results in the perturbation of an alpha-helical segment with a turn
288 motif (Supplemental Figure 1).

289

290

291 Discussion

292 Our results provide strong support for the hypothesis that rs34712979 is a functional
293 GWAS variant that acts by altering isoform usage of *NPNT*. Our analyses led to three main
294 findings: 1) the genetic association patterns of association to COPD, FEV1 and FEV1/FVC near
295 *NPNT* are highly likely to share rs34712979 as a causal genetic variant, 2) the A allele of
296 rs34712979 creates an alternative splice acceptor site that results in a 3-nucleotide exon
297 extension at the 5' end of exon 2 that is predicted to result in the addition of a serine to the
298 *NPNT* protein, and 3) the A allele alters *NPNT* isoform usage. The genetic association patterns
299 reflect the cumulative data of hundreds of thousands of subjects, and the splicing effects were
300 demonstrated in lung RNA from over 1,000 subjects in two different human cohorts.

301 *NPNT*, or nephronectin, is an extracellular protein involved in tissue development,
302 remodeling, and repair. It was first identified in the context of the search for a novel ligand for
303 integrin $\alpha 8 \beta 1$ ^{22,23} and was shown to be necessary for normal kidney development in murine
304 models²⁴. *NPNT* has also been linked to osteoblast differentiation and bone remodeling²⁵,
305 invasiveness of breast cancer²⁶, and pulmonary silicosis²⁷. Full-length *NPNT* protein includes
306 five epidermal growth factor (EGF)-like functional domains followed by a meprin, A-5 protein,
307 receptor protein-tyrosine phosphatase mu (MAM), and an Arg-Gly-Asp (RGD) integrin-binding
308 domain. The most well-described function of *NPNT* is as an extra-cellular matrix protein that
309 binds integrin $\alpha 8 \beta 1$ through its RGD domain. Recently, *NPNT* has been identified in
310 extracellular vesicles secreted by a murine breast cancer cell line, and full-length *NPNT* has been
311 reported to undergo post-translational modifications including cleavage to a shorter 20-kd
312 isoform²⁸. Full length *NPNT* has also been shown to be highly expressed in human pneumocytes
313 in the Human Protein Atlas project²⁹. Previous studies have identified *NPNT* as a potential
314 effector gene via association between lung function and mRNA levels and *NPNT* staining in
315 pulmonary endothelial and alveolar epithelial cells³⁰. In our study, we identify two full length
316 *NPNT* isoforms, one of which is fully annotated while the other utilizes a cryptic splice acceptor
317 in exon 2, in addition to three shortened isoforms that are likely protein coding. Within the pair
318 of long isoforms and within the pair of short isoforms, they differ only in the inclusion of a three
319 nucleotide 5' exon extension coding for a serine residue near the N-terminus of the protein.
320 Additional studies are needed to confirm the extent to which these differences at the RNA level
321 translate to differences in pulmonary protein isoform content, structure and function.

322 The *NPNT* genetic association lies in a region first described in association with lung
323 function^{31,32}. Subsequent studies confirmed an association with COPD and also identified two
324 independent signals at this locus³³. Several genes in this region have previously been
325 hypothesized to be the effector genes using eQTL studies, including the nearby genes *GSTCD*
326 and *INTS12*. These two genes harbor eQTL in blood and lung, and prior studies have found
327 correlations between *GSTCD* and *INTS12* mRNA with lung function and variable expression
328 during lung development³⁴. Additional studies identified *NPNT* as another potential effector gene
329 at this locus^{1,30}, indicating that the genetic association at this locus likely involves several
330 different genes and mechanisms.

331 From a genetic standpoint, the most clear and compelling genetic associations of *NPNT*
332 to common disease colocalize to an area including the second exon with clear evidence that
333 rs34712979 is a causal variant contributing to the association signals to lung function^{8,35} and
334 COPD¹. Our RNA-seq analyses provide strong evidence that the A allele of rs34712979 (which
335 is associated with decreased pulmonary function and increased COPD risk) creates an alternative
336 splice acceptor site at the 5' end of the second exon, and this site is preferentially used relative to
337 the annotated splice site on haplotypes containing the A allele. Usage of this alternative splice
338 site results in the inclusion of a serine residue near the N-terminus of the protein that is predicted
339 to perturb an alpha-helical segment with a turn motif. In addition to this alternative acceptor site
340 usage, the A allele substantially alters *NPNT* isoform usage patterns. While some of the isoform
341 changes can be directly explained by the second exon alternative splice site usage event, our
342 analysis also identified allelic effects on downstream splicing events as well. The mechanisms
343 that may link usage of alternative splice site to other splicing events remain to be elucidated.

344 As is characteristic of human GWAS associations, the penetrance of the rs34712979 A
345 allele is low, indicating that either the overall effect of this allele on biological function is subtle
346 or that the effect is large but occurs only in restricted sub-populations. Additional research on
347 *NPNT* isoform expression at the RNA and protein levels in large, well-phenotyped human
348 cohorts may identify specific sub-populations most impacted by the functional consequences of
349 this allele.

350 The strengths of this study are that the GWAS and RNA-seq findings are based on the
351 analysis of a very large number of human samples, and the integrated analysis of short and long-
352 read RNA-seq data provides a high level of resolution to observe changes at the isoform level.

353 To our knowledge this is one of the first applications of long-read sequencing to human tissue
354 samples to demonstrate splicing-related effects of a GWAS-identified genetic variant. Important
355 limitations to consider are that long-read sequencing technologies can be affected by biases
356 related to transcript length, therefore direct comparisons of abundance between the short and
357 long isoforms of *NPNT* should be interpreted with caution. The most prominent finding in our
358 study, namely allele-specific alternative splice site usage, is clearly identified in isoforms of
359 equal length suggesting that these allele-specific observations are unlikely to be affected by this
360 bias. While there is strong statistical support for our findings, these are nonetheless correlative
361 findings from large human cohorts and future work is required to define the underlying
362 molecular mechanisms and provide further experimental evidence to demonstrate how these
363 mechanisms may alter COPD risk.

364 In summary, these analyses demonstrate that the pulmonary disease GWAS association
365 near *NPNT* is very likely to be mediated by a common genetic variant that alters splicing,
366 resulting in the insertion of a novel serine residue in the *NPNT* protein immediately downstream
367 of the signal peptide domain. Given the known function of *NPNT* in extracellular matrix biology
368 and the high expression of *NPNT* in lung tissue and pneumocytes³⁰, further investigation of the
369 functional consequences of this splicing variant, including its effects on *NPNT* protein structure
370 and function, are likely to elucidate causal mechanisms of COPD pathogenesis.

371

372 **Table 1.** Number of ONT reads containing 30bp sequences including and excluding the TAG
373 sequence at the 5' end of the 2nd exon of NPNT.

rs34712979	...GATTAGGTG...	...GATGTG...
AA	825	206
AA	735	163
AA	291	92
AA	1638	403
AA	2041	500
GG	10	2216
GG	2	1519
GG	8	51
GG	6	502
GG	1	1553

374

375

376 **Funding/Acknowledgements:** This work was funded by R01 HL124233, R01 HL147326, R01
377 HL111527, U01 HL089897, U01 HL089856, R01HL125583, R01HL130512, T32HL007427.
378 Research reported in this publication was supported by the NHLBI and FDA Center for Tobacco
379 Products (CTP). The content is solely the responsibility of the authors and does not necessarily
380 represent the official views of the NIH or the Food and Drug Administration.
381 The Genotype-Tissue Expression (GTEx) Project was supported by the [Common Fund](#) of the
382 Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA,
383 NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained
384 from: the GTEx Portal between January and August of 2020 and via the GTEx Terra Workspace
385 during the same time interval.

386
387 Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by
388 the National Heart, Lung and Blood Institute (NHLBI). Whole Genome Sequencing and
389 RNASeq for "NHLBI TOPMed: The Lung Tissue Research Consortium (phs001662)" was
390 performed at Northwest Genome Center (NWGC, HHSN268201600032I, RNASeq) and Broad
391 Genomics (HHSN268201600034I, WGS) Core support including centralized genomic read
392 mapping and genotype calling, along with variant quality metrics and filtering were provided by
393 the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract
394 HHSN268201800002I). Core support including phenotype harmonization, data management,
395 sample-identity QC, and general program coordination were provided by the TOPMed Data
396 Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We
397 gratefully acknowledge the studies and participants who provided biological samples and data
398 for TOPMed.

399

400 **COPDGene[®] Investigators – Core Units:**

401 *Administrative Center:* James D. Crapo, MD (PI); Edwin K. Silverman, MD, PhD (PI); Barry J.
402 Make, MD; Elizabeth A. Regan, MD, PhD

403

404 *Genetic Analysis Center:* Terri Beaty, PhD; Ferdouse Begum, PhD; Peter J. Castaldi, MD, MSc;
405 Michael Cho, MD; Dawn L. DeMeo, MD, MPH; Adel R. Boueiz, MD; Marilyn G. Foreman,
406 MD, MS; Eitan Halper-Stromberg; Lystra P. Hayden, MD, MMSc; Craig P. Hersh, MD, MPH;
407 Jacqueline Hetmanski, MS, MPH; Brian D. Hobbs, MD; John E. Hokanson, MPH, PhD; Nan
408 Laird, PhD; Christoph Lange, PhD; Sharon M. Lutz, PhD; Merry-Lynn McDonald, PhD;

409 Margaret M. Parker, PhD; Dmitry Prokopenko, Ph.D; Dandi Qiao, PhD; Elizabeth A. Regan,
410 MD, PhD; Phuwanat Sakornsakolpat, MD; Edwin K. Silverman, MD, PhD; Emily S. Wan, MD;
411 Sungho Won, PhD

412
413 *Imaging Center:* Juan Pablo Centeno; Jean-Paul Charbonnier, PhD; Harvey O. Coxson, PhD;
414 Craig J. Galban, PhD; MeiLan K. Han, MD, MS; Eric A. Hoffman, Stephen Humphries, PhD;
415 Francine L. Jacobson, MD, MPH; Philip F. Judy, PhD; Ella A. Kazerooni, MD; Alex Kluiber;
416 David A. Lynch, MB; Pietro Nardelli, PhD; John D. Newell, Jr., MD; Aleena Notary; Andrea
417 Oh, MD; Elizabeth A. Regan, MD, PhD; James C. Ross, PhD; Raul San Jose Estepar, PhD;
418 Joyce Schroeder, MD; Jered Sieren; Berend C. Stoel, PhD; Juerg Tschirren, PhD; Edwin Van
419 Beek, MD, PhD; Bram van Ginneken, PhD; Eva van Rikxoort, PhD; Gonzalo Vegas Sanchez-
420 Ferrero, PhD; Lucas Veitel; George R. Washko, MD; Carla G. Wilson, MS;

421
422 *PFT QA Center, Salt Lake City, UT:* Robert Jensen, PhD

423
424 *Data Coordinating Center and Biostatistics, National Jewish Health, Denver, CO:* Douglas
425 Everett, PhD; Jim Crooks, PhD; Katherine Pratte, PhD; Matt Strand, PhD; Carla G. Wilson, MS

426
427 *Epidemiology Core, University of Colorado Anschutz Medical Campus, Aurora, CO:* John E.
428 Hokanson, MPH, PhD; Gregory Kinney, MPH, PhD; Sharon M. Lutz, PhD; Kendra A. Young,
429 PhD

430
431 *Mortality Adjudication Core:* Surya P. Bhatt, MD; Jessica Bon, MD; Alejandro A. Diaz, MD,
432 MPH; MeiLan K. Han, MD, MS; Barry Make, MD; Susan Murray, ScD; Elizabeth Regan, MD;
433 Xavier Soler, MD; Carla G. Wilson, MS

434
435 *Biomarker Core:* Russell P. Bowler, MD, PhD; Katerina Kechris, PhD; Farnoush Banaei-
436 Kashani, Ph.D **COPDGene® Investigators – Clinical Centers**
437 *Ann Arbor VA:* Jeffrey L. Curtis, MD; Perry G. Pernicano, MD

438
439 *Baylor College of Medicine, Houston, TX:* Nicola Hanania, MD, MS; Mustafa Atik, MD; Aladin
440 Boriek, PhD; Kalpatha Guntupalli, MD; Elizabeth Guy, MD; Amit Parulekar, MD;

441
442 *Brigham and Women’s Hospital, Boston, MA:* Dawn L. DeMeo, MD, MPH; Alejandro A. Diaz,
443 MD, MPH; Lystra P. Hayden, MD; Brian D. Hobbs, MD; Craig Hersh, MD, MPH; Francine L.
444 Jacobson, MD, MPH; George Washko, MD

445
446 *Columbia University, New York, NY:* R. Graham Barr, MD, DrPH; John Austin, MD; Belinda
447 D’Souza, MD; Byron Thomashow, MD

448
449 *Duke University Medical Center, Durham, NC:* Neil MacIntyre, Jr., MD; H. Page McAdams,
450 MD; Lacey Washington, MD

451
452 *Grady Memorial Hospital, Atlanta, GA:* Eric Flenaugh, MD; Silanth Terpenning, MD

453

454 *HealthPartners Research Institute, Minneapolis, MN*: Charlene McEvoy, MD, MPH; Joseph
455 Tashjian, MD
456
457 *Johns Hopkins University, Baltimore, MD*: Robert Wise, MD; Robert Brown, MD; Nadia N.
458 Hansel, MD, MPH; Karen Horton, MD; Allison Lambert, MD, MHS; Nirupama Putcha, MD,
459 MHS
460
461 *Lundquist Institute for Biomedical Innovation at Harbor UCLA Medical Center, Torrance, CA*:
462 Richard Casaburi, PhD, MD; Alessandra Adami, PhD; Matthew Budoff, MD; Hans Fischer,
463 MD; Janos Porszasz, MD, PhD; Harry Rossiter, PhD; William Stringer, MD
464
465 *Michael E. DeBakey VAMC, Houston, TX*: Amir Sharafkhaneh, MD, PhD; Charlie Lan, DO
466
467 *Minneapolis VA*: Christine Wendt, MD; Brian Bell, MD; Ken M. Kunisaki, MD, MS
468
469 *National Jewish Health, Denver, CO*: Russell Bowler, MD, PhD; David A. Lynch, MB
470
471 *Reliant Medical Group, Worcester, MA*: Richard Rosiello, MD; David Pace, MD
472
473 *Temple University, Philadelphia, PA*: Gerard Criner, MD; David Ciccolella, MD; Francis
474 Cordova, MD; Chandra Dass, MD; Gilbert D'Alonzo, DO; Parag Desai, MD; Michael Jacobs,
475 PharmD; Steven Kelsen, MD, PhD; Victor Kim, MD; A. James Mamary, MD; Nathaniel
476 Marchetti, DO; Aditi Satti, MD; Kartik Shenoy, MD; Robert M. Steiner, MD; Alex Swift, MD;
477 Irene Swift, MD; Maria Elena Vega-Sanchez, MD
478
479 *University of Alabama, Birmingham, AL*: Mark Dransfield, MD; William Bailey, MD; Surya P.
480 Bhatt, MD; Anand Iyer, MD; Hrudaya Nath, MD; J. Michael Wells, MD
481
482 *University of California, San Diego, CA*: Douglas Conrad, MD; Xavier Soler, MD, PhD; Andrew
483 Yen, MD
484
485 *University of Iowa, Iowa City, IA*: Alejandro P. Comellas, MD; Karin F. Hoth, PhD; John
486 Newell, Jr., MD; Brad Thompson, MD
487
488 *University of Michigan, Ann Arbor, MI*: MeiLan K. Han, MD MS; Ella Kazerooni, MD MS;
489 Wassim Labaki, MD MS; Craig Galban, PhD; Dharshan Vummidi, MD
490
491 *University of Minnesota, Minneapolis, MN*: Joanne Billings, MD; Abbie Begnaud, MD; Tadashi
492 Allen, MD
493
494 *University of Pittsburgh, Pittsburgh, PA*: Frank Scieurba, MD; Jessica Bon, MD; Divay Chandra,
495 MD, MSc; Carl Fuhrman, MD; Joel Weissfeld, MD, MPH
496
497 *University of Texas Health, San Antonio, San Antonio, TX*: Antonio Anzueto, MD; Sandra
498 Adams, MD; Diego Maselli-Caceres, MD; Mario E. Ruiz, MD; Harjinder Singh
499

500 **Conflict of Interest Statement:** P. Castaldi has received personal fees and grant support from
501 GlaxoSmithKline and Novartis. C.Hersh has received grants from NHLBI, Bayer, Boehringer-
502 Ingelheim, Novartis and Vertex. M. Cho has received grant support from GSK and Bayer, and
503 speaking or consulting fees from AstraZeneca and Illumina. E. Silverman has received grant
504 support from GSK and Bayer. A. Laederach reports grants from the NIH NHLBI and consultant
505 fees from Ribometrix. C. Vollmers has filed patent applications on aspects of the R2C2 method.
506

507 **References**

508

- 509 1. Sakornsakolpat, P., Prokopenko, D., Lamontagne, M., Reeve, N.F., Guyatt, A.L., Jackson,
510 V.E., Shrine, N., Qiao, D., Bartz, T.M., Kim, D.K., et al. (2019). Genetic landscape of chronic
511 obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations.
512 *Nature Genetics* 51, 494–505.
- 513 2. Zhou, X., Baron, R.M., Hardin, M., Cho, M., Zielinski, J., Hawrylkiewicz, I., Sliwinski, P.,
514 Hersh, C.P., Mancini, J.D., Lu, K., et al. (2012). Identification of a chronic obstructive
515 pulmonary disease genetic determinant that regulates HHIP. *Hum. Mol. Genet.* 21, 1325–1335.
- 516 3. Castaldi, P.J., Guo, F., Qiao, D., Du, F., Naing, Z.Z.C., Li, Y., Pham, B., Mikkelsen, T.S.,
517 Cho, M., Silverman, E., et al. (2018). Identification of Functional Variants in the FAM13A
518 COPD GWAS Locus by Massively Parallel Reporter Assays. *American Journal of Respiratory
519 and Critical Care Medicine* 199, 52–61.
- 520 4. Parker, M.M., Hao, Y., Guo, F., Pham, B., Chase, R., Platig, J., Cho, M., Hersh, C.P.,
521 Thannickal, V.J., Crapo, J.D., et al. (2019). Identification of an emphysema-associated genetic
522 variant near TGFB2 with regulatory effects in lung fibroblasts. *Elife* 8.
- 523 5. Boueiz, A., Pham, B., Chase, R., Lamb, A., Lee, S., Naing, Z.Z.C., Cho, M., Parker, M.M.,
524 Sakornsakolpat, P., Hersh, C.P., et al. (2019). Integrative Genomics Analysis Identifies
525 ACVR1B as a Candidate Causal Gene of Emphysema Distribution. *Am J Respir Cell Mol Biol*
526 60, 388–398.
- 527 6. Li, Y.I., van de Geijn, B., Raj, A., Knowles, D.A., Petti, A.A., Golan, D., Gilad, Y., and
528 Pritchard, J.K. (2016). RNA splicing is a primary link between genetic variation and disease.
529 *Science* 352, 600–604.
- 530 7. Saferali, A., Yun, J.H., Parker, M.M., Sakornsakolpat, P., Chase, R.P., Lamb, A., Hobbs, B.D.,
531 Boezen, M.H., Dai, X., de Jong, K., et al. (2019). Analysis of genetically driven alternative
532 splicing identifies FBXO38 as a novel COPD susceptibility gene. *PLoS Genet* 15, e1008229.
- 533 8. Shrine, N., Guyatt, A.L., Erzurumluoglu, A.M., Jackson, V.E., Hobbs, B.D., Melbourne, C.A.,
534 Batini, C., Fawcett, K.A., Song, K., Sakornsakolpat, P., et al. (2019). New genetic signals for
535 lung function highlight pathways and chronic obstructive pulmonary disease associations across
536 multiple ancestries. *Nature Genetics* 51, 481–493.
- 537 9. The GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects
538 across human tissues. *Science* 369, 1318–1330.
- 539 10. Farh, K.K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shores, N.,
540 Whitton, H., Ryan, R.J.H., Shishkin, A.A., et al. (2015). Genetic and epigenetic fine mapping of
541 causal autoimmune disease variants. *Nature* 518, 337–343.
- 542 11. Wu, E., Nance, T., and Montgomery, S.B. (2014). SplicePlot: a utility for visualizing splicing
543 quantitative trait loci. *Bioinformatics* 30, 1025–1026.

- 544 12. Picelli, S., Faridani, O.R., Björklund, Å.K., Winberg, G., Sagasser, S., and Sandberg, R.
545 (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* 9, 171–181.
- 546 13. Byrne, A., Supple, M.A., Volden, R., Laidre, K.L., Shapiro, B., and Vollmers, C. (2019).
547 Depletion of Hemoglobin Transcripts and Long-Read Sequencing Improves the Transcriptome
548 Annotation of the Polar Bear (*Ursus maritimus*). *Front Genet* 10, 403.
- 549 14. Volden, R., Palmer, T., Byrne, A., Cole, C., Schmitz, R.J., Green, R.E., and Vollmers, C.
550 (2018). Improving nanopore read accuracy with the R2C2 method enables the sequencing of
551 highly multiplexed full-length single-cell cDNA. *Proceedings of the National Academy of*
552 *Sciences* 115, 9726–9731.
- 553 15. Cole, C., Byrne, A., Adams, M., Volden, R., and Vollmers, C. (2020). Complete
554 characterization of the human immune cell transcriptome using accurate full-length cDNA
555 sequencing. *Genome Res.* 30, 589–601.
- 556 16. Volden, R., and Vollmers, C. (2020). Highly Multiplexed Single-Cell Full-Length cDNA
557 Sequencing of human immune cells with 10X Genomics and R2C2. *bioRxiv* 20,
558 2020.01.10.902361.
- 559 17. Tardaguila, M., la Fuente, de, L., Marti, C., Pereira, C., Pardo-Palacios, F.J., Del Risco, H.,
560 Ferrell, M., Mellado, M., Macchietto, M., Verheggen, K., et al. (2018). SQANTI: extensive
561 characterization of long-read transcript sequences for quality control in full-length transcriptome
562 identification and quantification. *Genome Res.* 28, 396–411.
- 563 18. UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids*
564 *Research* 47, D506–D515.
- 565 18. Kumar, T., (2013). CFSSP: Chou and Fasman secondary structure prediction server. *Wide*
566 *Spectrum* 1(9), 15–19.
- 567 20. Hiller, M., Huse, K., Szafranski, K., Jahn, N., Hampe, J., Schreiber, S., Backofen, R., and
568 Platzer, M. (2004). Widespread occurrence of alternative splicing at NAGNAG acceptors
569 contributes to proteome plasticity. *Nature Genetics* 36, 1255–1257.
- 570 21. Bradley, R.K., Merkin, J., Lambert, N.J., and Burge, C.B. (2012). Alternative splicing of
571 RNA triplets is often regulated and accelerates proteome evolution. *PLoS Biol.* 10, e1001229.
- 572 22. Brandenberger, R., Schmidt, A., Linton, J., Wang, D., Backus, C., Denda, S., Müller, U., and
573 Reichardt, L.F. (2001). Identification and characterization of a novel extracellular matrix protein
574 nephronectin that is associated with integrin $\alpha 8\beta 1$ in the embryonic kidney. *J. Cell Biol.*
575 154, 447–458.
- 576 23. Morimura, N., Tezuka, Y., Watanabe, N., Yasuda, M., Miyatani, S., Hozumi, N., and
577 Tezuka, K.-I. (2001). Molecular Cloning of POEM: A Novel Adhesion Molecule That Interacts
578 With $\alpha 8\beta 1$ Integrin. *Journal of Biological Chemistry* 276, 42172–42181.

- 579 24. Linton, J.M., Martin, G.R., and Reichardt, L.F. (2007). The ECM protein nephronectin
580 promotes kidney development via integrin alpha8beta1-mediated stimulation of Gdnf expression.
581 *Development* 134, 2501–2509.
- 582 25. Kahai, S., Lee, S.-C., Seth, A., and Yang, B.B. (2010). Nephronectin promotes osteoblast
583 differentiation via the epidermal growth factor-like repeats. *FEBS Lett.* 584, 233–238.
- 584 26. Steigedal, T.S., Toraskar, J., Redvers, R.P., Valla, M., Magnussen, S.N., Bofin, A.M.,
585 Opdahl, S., Lundgren, S., Eckhardt, B.L., Lamar, J.M., et al. (2018). Nephronectin is Correlated
586 with Poor Prognosis in Breast Cancer and Promotes Metastasis via its Integrin-Binding Motifs.
587 *Neoplasia* 20, 387–400.
- 588 27. Lee, S., Honda, M., Yamamoto, S., Kumagai-Takei, N., Yoshitome, K., Nishimura, Y., Sada,
589 N., Kon, S., and Otsuki, T. (2019). Role of Nephronectin in Pathophysiology of Silicosis. *Int J*
590 *Mol Sci* 20, 2581.
- 591 28. Toraskar, J., Magnussen, S.N., Hagen, L., Sharma, A., Hoang, L., Bjørkøy, G., Svineng, G.,
592 and Steigedal, T.S. (2019). A Novel Truncated Form of Nephronectin Is Present in Small
593 Extracellular Vesicles Isolated from 66cl4 Cells. *J. Proteome Res.* 18, 1237–1247.
- 594 29. Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A.,
595 Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). Proteomics. Tissue-based map
596 of the human proteome. *Science* 347, 1260419–1260419.
- 597 30. Obeidat, M., Hao, K., Bossé, Y., Nickle, D.C., Nie, Y., Postma, D.S., Laviolette, M.,
598 Sandford, A.J., Daley, D.D., Hogg, J.C., et al. (2015). Molecular mechanisms underlying
599 variations in lung function: a systems genetics analysis. *Lancet Respir Med* 3, 782–795.
- 600 31. Hancock, D., Eijgelsheim, M., Wilk, J.B., Gharib, S., Loehr, L., Marciante, K., Franceschini,
601 N., Durme, Y., Chen, T.-H., Barr, R.G., et al. (2010). Meta-analyses of genome-wide association
602 studies identify multiple loci associated with pulmonary function. *Nature Genetics* 42, 45–52.
- 603 32. Parker, M.M., Foreman, M.G., Abel, H.J., Mathias, R.A., Hetmanski, J.B., Crapo, J.D.,
604 Silverman, E., Beaty, T.H., COPDGene Investigators (2014). Admixture mapping identifies a
605 quantitative trait locus associated with FEV1/FVC in the COPDGene Study. *Genet. Epidemiol.*
606 38, 652–659.
- 607 33. Soler Artigas, M., Wain, L.V., Miller, S., Kheirallah, A.K., Huffman, J.E., Ntalla, I., Shrine,
608 N., Obeidat, M., Trochet, H., McArdle, W.L., et al. (2015). Sixteen new lung function signals
609 identified through 1000 Genomes Project reference panel imputation. *Nat Commun* 6, 8658.
- 610 34. Obeidat, M., Miller, S., Probert, K., Billington, C.K., Henry, A.P., Hodge, E., Nelson, C.P.,
611 Stewart, C.E., Swan, C., Wain, L.V., et al. (2013). GSTCD and INTS12 regulation and
612 expression in the human lung. *PLoS ONE* 8, e74630.
- 613 35. Wain, L.V., Shrine, N., Artigas, M.S., Erzurumluoglu, A.M., Noyvert, B., Bossini-Castillo,
614 L., Obeidat, M., Henry, A.P., Portelli, M.A., Hall, R.J., et al. (2017). Genome-wide association

615 analyses for lung function and chronic obstructive pulmonary disease identify new loci and
616 potential druggable targets. *Nature Genetics* 49, 416–425.

617