Supplementary Information

Novel variation at chr11p13 associated with cystic fibrosis lung disease severity

Hong Dang[1], Paul J. Gallins[2], Rhonda G. Pace[1], Xue-liang Guo[1], Jaclyn R. Stonebraker[1], Harriet Corvol[3,4], Garry R. Cutting[5,6], Mitchell L. Drumm[7], Lisa J. Strug[8,9], Michael R. Knowles[1], and Wanda K. O'Neal[1]

[1] Marsico Lung Institute, University of North Carolina at Chapel Hill School of Medicine CF/Pulmonary Research and Treatment Center, Chapel Hill, NC 27599, USA
[2] Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27607, USA
[3] Assistance Publique-Hôpitaux de Paris (AP-HP), Hôpital Trousseau, Pediatric Pulmonary Department; Institut National de la Santé et la Recherche Médicale (INSERM) U938, Paris 75012, France
[4] Sorbonne Universités, Université Pierre et Marie Curie (UPMC) Paris 06, Paris 75005, France
[5] McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21287, USA
[6] Department of Pediatrics, Johns Hopkins University School of Medicine, Baltimore, Maryland 21287, USA
[7] Department of Pediatrics, School of Medicine, Case Western Reserve University, Cleveland, Ohio 44106, USA
[8] Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Ontario, Canada M5G 0A4
[9] Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada M5T 3M7

Correspondence:
Hong Dang (dangh@email.unc.edu)
919-966-2580
7209 Marsico Hall
CB#7248
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599

Supplementary Table 1. Demographics of patients selected for resequencing

	Study Group	Subjects *n*	Enrollment Age (Yrs) Mean (±SD)	Range	Males (%)	European (%)
ReSeqChr11	Mild	191	29.1 (10.2)	11.7 - 57.6	93 (48.7)	191 (100)
	Severe	186	15.5 (4.6)	8.0 - 34.9	96 (51.6)	186 (100)
	Total	**377**	**22.4 (10.5)**	**8.0 - 57.6**	**189 (50.1)**	**377 (100)**
GWAS1+2[a] All	**Total**	**6,365**	**19.5 (9.4)**	**6.0 - 62.2**	**3,368 (52.9)**	**6,079 (95.5)**

[a]Corvol, et al., *Nat Commun*, 2015.

Supplementary Table 2. Summary of rare variant set tests for association with CF lung disease severity by SKAT test and Burden test[a]

Set ID	SNPs (*n*) All	Markers per Test (*n*)	SKAT *P* value	Burden *P* value
APIP exon	19	19	0.49	0.67
APIP gene	221	219	0.36	0.03
EHF exon	10	8	0.57	0.05
EHF gene	180	170	0.58	0.45
MIR1343 miRNA	1	1	0.26	0.26
PDHX exon	14	14	0.20	0.32
PDHX gene	309	308	0.90	0.86
RP11-350D17.3 lincRNA	4	3	0.60	0.63

[a]Wu, et al., *AJHG*, 2011.

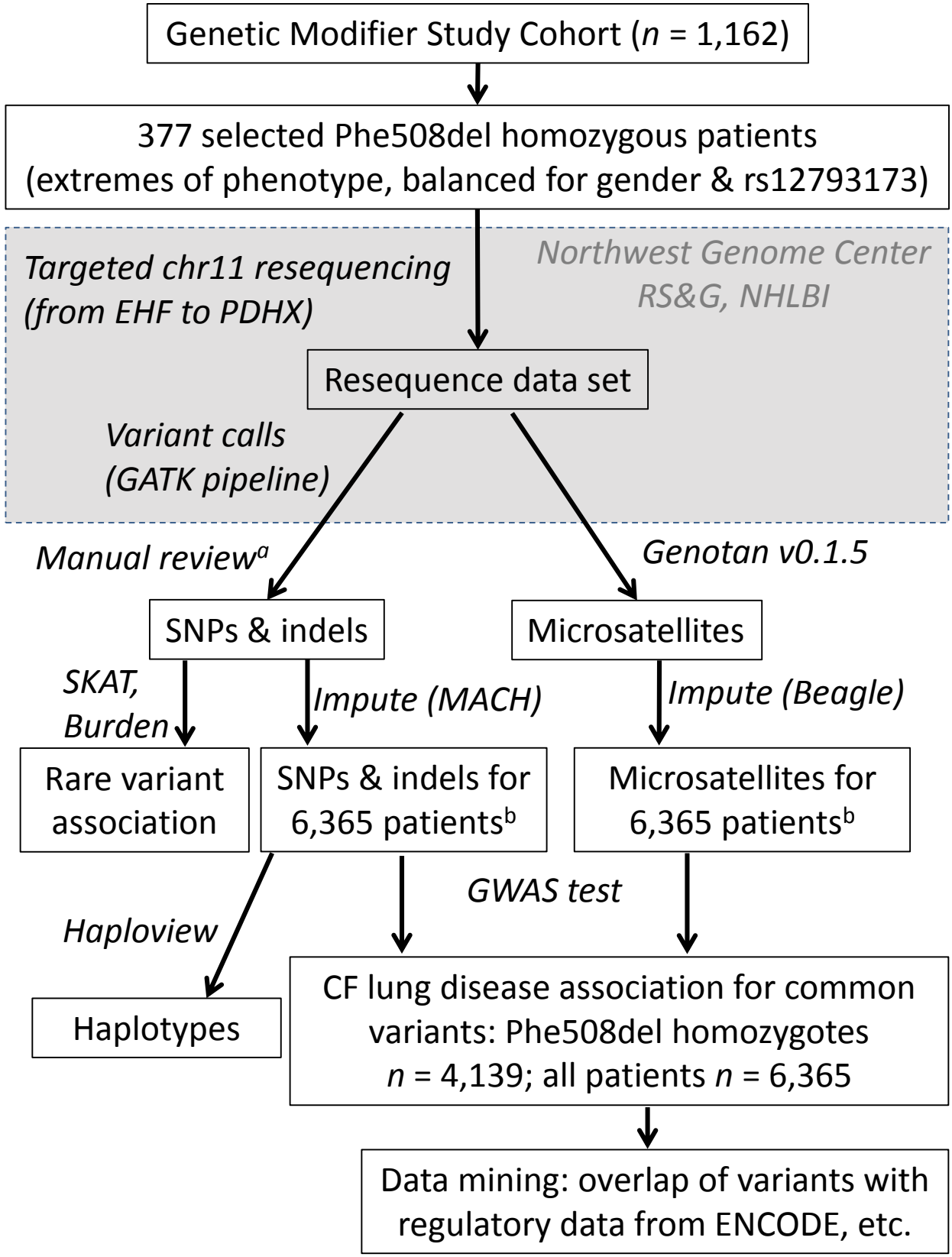
Supplementary Table 3 can be found on the Human Genome Variation website (http://www.nature.com/hgv).

Supplementary Table 4. Transcription Factor ChIP-seq factors from ENCODE for sequenced region from Figure 1L.

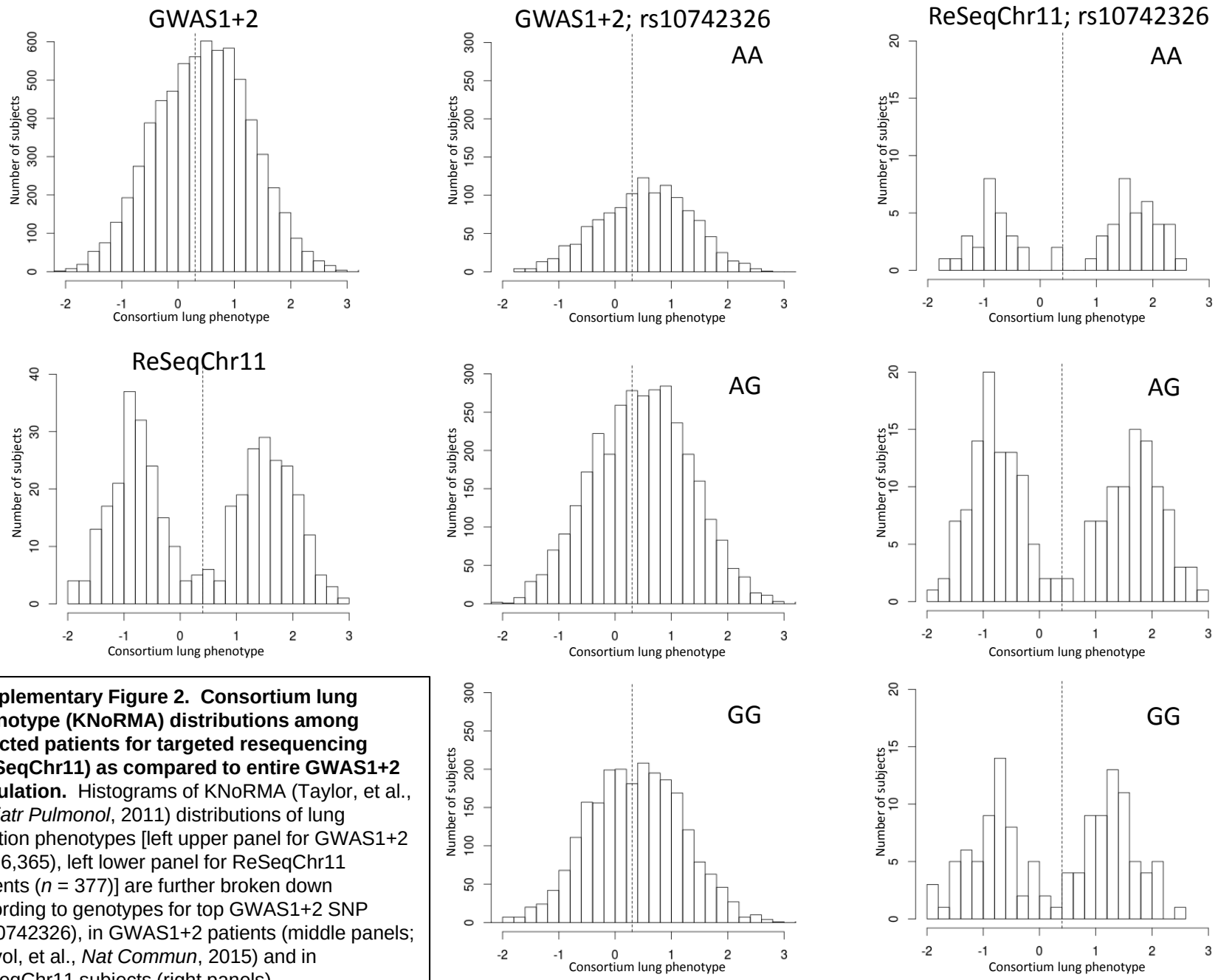
ENCODE derived transcription factor confidence scores are defined as described in the methods of http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=491641535_kyVVua7AaN7Pq4JuatQWgg283ubY&c=chr11&g=wgEncodeRegTfbsClusteredV3

chr	chromosome start	chromosome end	name	score
chr11	34778471	34778898	MYC	170
chr11	34778568	34779095	STAT3	1000
chr11	34778571	34778827	CEBPB	316
chr11	34778585	34778885	NR3C1	162
chr11	34778619	34779159	EP300	364
chr11	34778621	34778889	FOS	405
chr11	34778687	34779097	FOSL2	302
chr11	34778758	34779054	GATA3	260
chr11	34778773	34779043	FOXA1	186
chr11	34778865	34779375	POLR2A	180
chr11	34779737	34780493	GATA2	429
chr11	34779777	34780233	CCNT2	148
chr11	34779786	34780242	JUND	370
chr11	34779801	34780277	STAT1	233
chr11	34779812	34780252	MYC	258
chr11	34779823	34780193	MAZ	163
chr11	34779834	34780198	TEAD4	872
chr11	34779842	34780138	RCOR1	296
chr11	34779850	34780200	STAT5A	245
chr11	34779856	34780172	EP300	501
chr11	34779858	34780062	SPI1	201
chr11	34779875	34780271	MAX	247
chr11	34779891	34780167	TAL1	502
chr11	34779948	34780378	E2F6	144
chr11	34792834	34793124	MAFF	350
chr11	34792839	34793121	MAFK	628
chr11	34794704	34795300	TCF7L2	225
chr11	34794829	34795159	EP300	337
chr11	34794892	34795132	FOXA1	252
chr11	34794900	34795065	GATA3	1000
chr11	34795375	34795705	EP300	146
chr11	34795436	34795732	GATA3	265
chr11	34796919	34797259	RFX5	300
chr11	34797031	34797255	CEBPB	403
chr11	34804857	34805137	FOXA1	214
chr11	34805377	34805927	GATA2	214

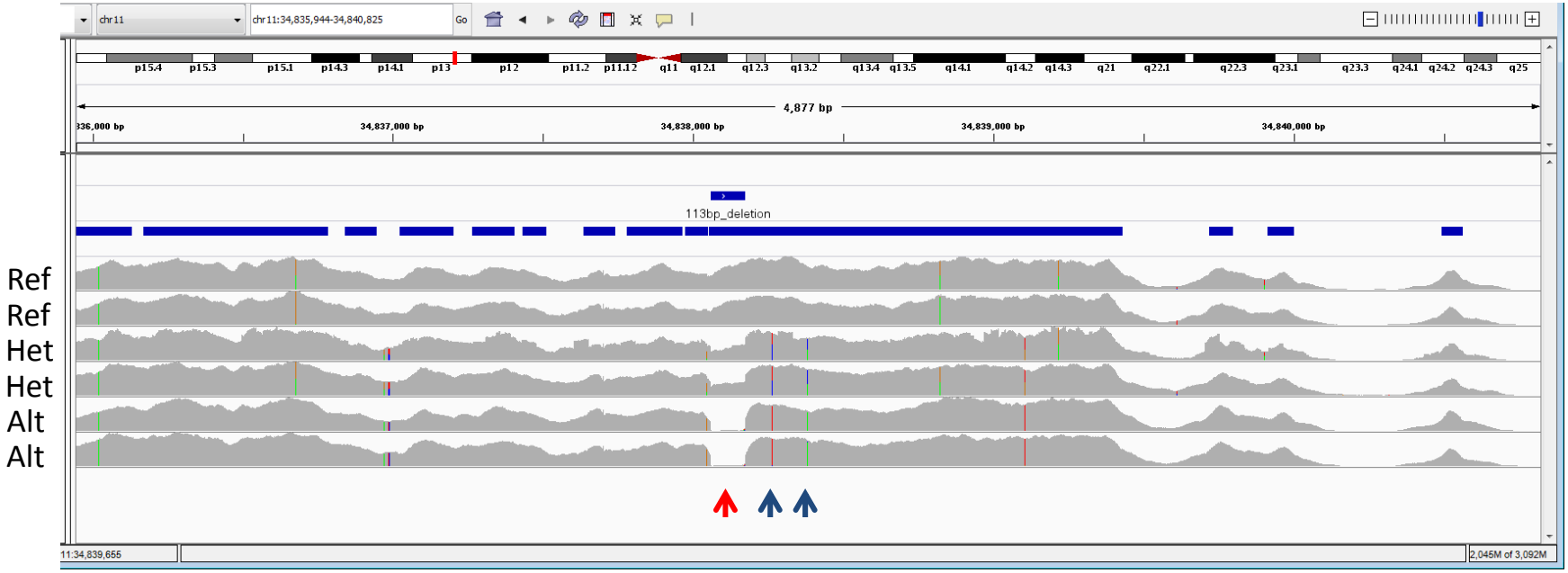
chr	chromosome start	chromosome end	name	score
chr11	34805387	34805827	TEAD4	221
chr11	34805440	34805670	CEBPB	167
chr11	34805469	34805851	EP300	623
chr11	34805477	34806303	MYBL2	328
chr11	34805529	34805865	SP1	384
chr11	34805565	34805861	HNF4A	183
chr11	34805569	34806113	NFIC	305
chr11	34805569	34805825	JUND	204
chr11	34805578	34805778	USF1	209
chr11	34805593	34805809	MEF2A	173
chr11	34805596	34805886	JUN	139
chr11	34805598	34805867	RXRA	1000
chr11	34805679	34805887	FOXA2	572
chr11	34805687	34805878	FOXA1	713
chr11	34806121	34806419	EP300	1000
chr11	34806126	34806426	RXRA	206
chr11	34806128	34806418	HNF4G	301
chr11	34806133	34806355	NFIC	379
chr11	34806160	34806370	FOXA1	629
chr11	34806162	34806386	RAD21	151
chr11	34806166	34806374	FOXA2	582
chr11	34806180	34806435	CEBPB	394
chr11	34806194	34806343	SP1	383
chr11	34806195	34806362	HNF4A	552
chr11	34810366	34810606	FOXA1	207
chr11	34812065	34812565	CTCF	691
chr11	34812166	34812410	RAD21	562
chr11	34812175	34812435	SMC3	219
chr11	34813340	34813576	BATF	186
chr11	34813646	34814089	CTCF	256
chr11	34813731	34813971	RAD21	166
chr11	34814311	34814607	GATA3	200
chr11	34814318	34814642	EP300	222
chr11	34814323	34814613	FOXA2	232
chr11	34816268	34816608	BACH1	413
chr11	34816292	34816612	MAFK	1000
chr11	34816315	34816595	MAFF	1000
chr11	34816668	34816964	GATA3	236
chr11	34819068	34819508	CBX3	327



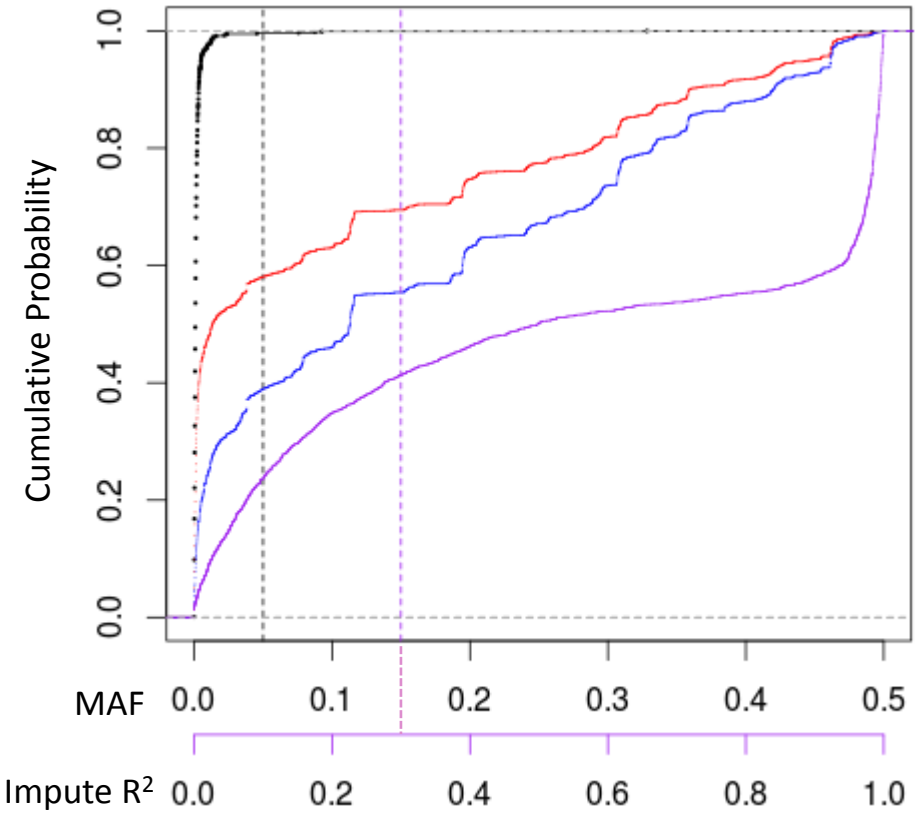
Supplementary Figure 1. Flowchart of Experimental Design and Quality Control.
[a]Especially for insertions/deletions (indels). [b]From the International CF Gene Modifier Consortium.



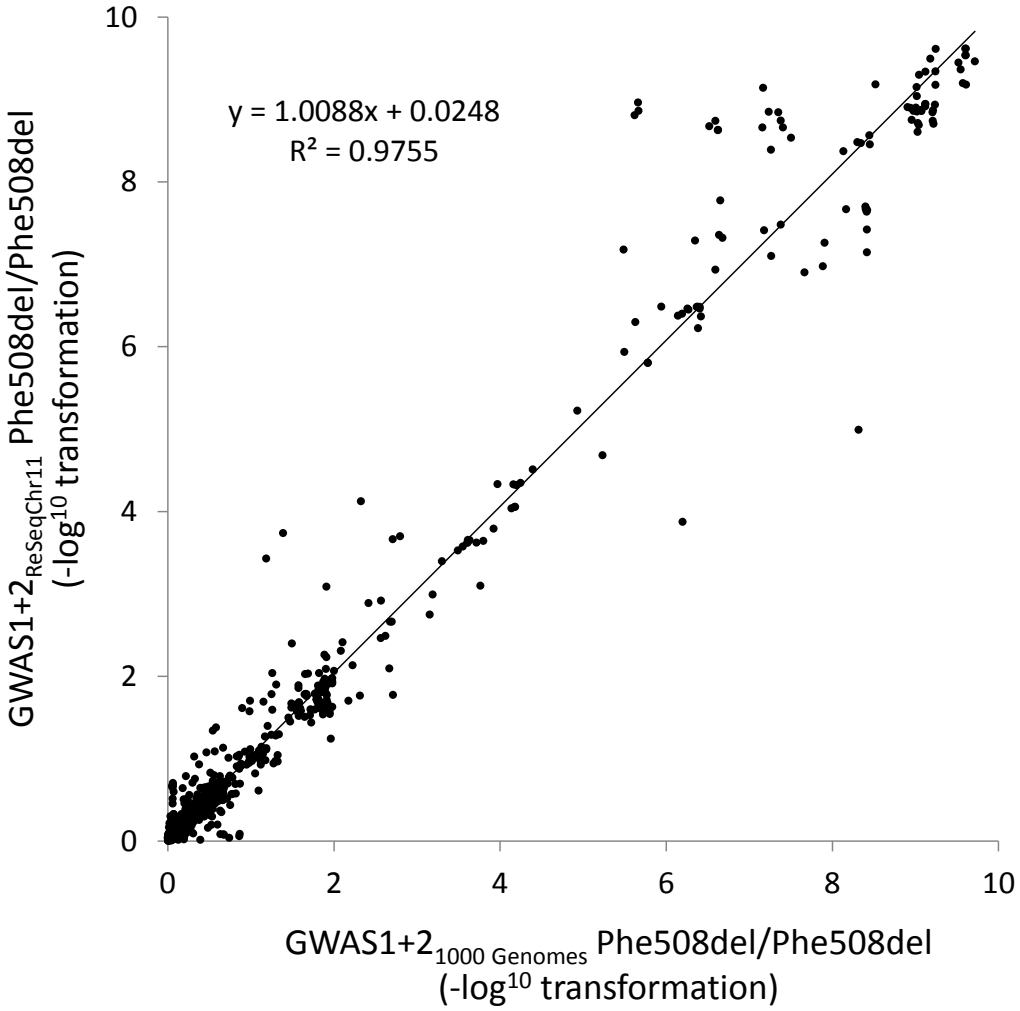
Supplementary Figure 2. Consortium lung phenotype (KNoRMA) distributions among selected patients for targeted resequencing (ReSeqChr11) as compared to entire GWAS1+2 population. Histograms of KNoRMA (Taylor, et al., *Pediatr Pulmonol*, 2011) distributions of lung function phenotypes [left upper panel for GWAS1+2 (*n* = 6,365), left lower panel for ReSeqChr11 patients (*n* = 377)] are further broken down according to genotypes for top GWAS1+2 SNP (rs10742326), in GWAS1+2 patients (middle panels; Corvol, et al., *Nat Commun*, 2015) and in ReSeqChr11 subjects (right panels).



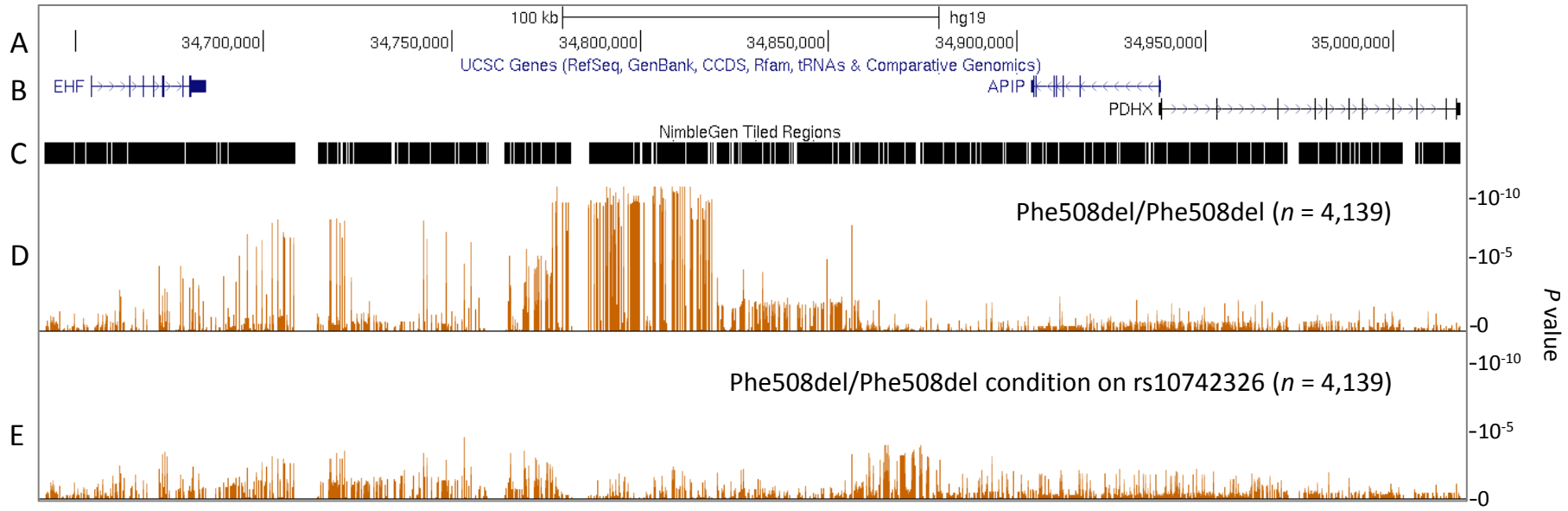
Supplementary Figure 3. Example of resequencing coverage maps around 113 bp deletion. Coverage tracks from six example sequence alignment and mapping results in BAM format were visualized in Integrative Genomics Viewer (IGV; Robinson, et al., *Nat Biotechnol*, 2011) genome browser. Grey mountain ranges show coverage from aligned sequence reads. Two samples each of homozygous reference (Ref), heterozygous (Het), and homozygous deletion (Alt) alleles of a 113 bp LINE element (red arrow), and two SNPs in LD (blue arrows) are shown.



Supplementary Figure 4. Empirical Cumulative Distribution Function (eCDF) plots of MAF and imputation quality among all SNPs/indels from the ReSeqChr11 project. The minor allele frequencies (MAF; ranging from 0 – 0.5; X axis in black) of all variants (red dots) were plotted along the cumulative distribution (Y axis), while novel variants are marked by black dots. Known variants as defined by SNPs/indels with known dbSNP ID (rs number) are shown in blue dots. Vertical grey dashed line represents MAF = 0.05. Purple dots represent the cumulative distribution of variance imputation quality R^2 values (ranging from 0 – 1; X axis in purple). Vertical purple dashed line indicates R^2 value of 0.3 (typically accepted as quality threshold, above which the imputed variants are deemed of sufficient quality).



Supplementary Figure 5. Comparison of CF lung disease association *P* values between GWAS1+2 (Corvol et al, 2015) and ReSeqChr11 imputed data. CF lung disease severity association *P* values plotted compared to the earlier GWAS1+2 analysis (utilizing 1000 Genomes data for imputation; GWAS1+$2_{1000\ Genomes}$; Corvol et al. 2015) and newly imputed *P* values for those same patients, based on the results of the ReSeqChr11 resequencing data reported here (GWAS1+$2_{ReSeqChr11}$) using 377 ReSeqChr11 subjects for imputation, after $-\log_{10}$ transformation. Comparisons are for Phe508del homozygous (*n* = 4,139) subjects.



Supplementary Figure 6. Summary of conditional analysis of CF lung disease association with rs10742326 genotypes as covariate. Remaining CF lung disease association signals, after conditioning on rs10742326 genotype, were compared to signals from Phe508del homozygous patients using the UCSC genome browser. The tracks are: (A) Scale bar and genome coordinates on chr11 of UCSC hg19 reference genome; (B) UCSC genes annotation showing *EHF*, *APIP* and *PDHX* genomic structure; (C) Tiled region of NimbleGen probes used to enrich the local genomic DNA to be resequenced; (D) CF lung disease severity association *P* values for imputed SNPs/indels among Phe508del homozygous patients; (E) CF lung disease severity association *P* values after conditioning on rs10742326 genotype dosage; none of the SNPs (*n* = 3,496) achieved regional significance after Bonferroni correction.