The American Journal of Human Genetics, Volume *99*

Supplemental Data

Increasing Generality and Power of Rare-Variant Tests by Utilizing Extended Pedigrees

Jae Hoon Sul, Brian E. Cade, Michael H. Cho, Dandi Qiao, Edwin K. Silverman, Susan Redline, and Shamil Sunyaev

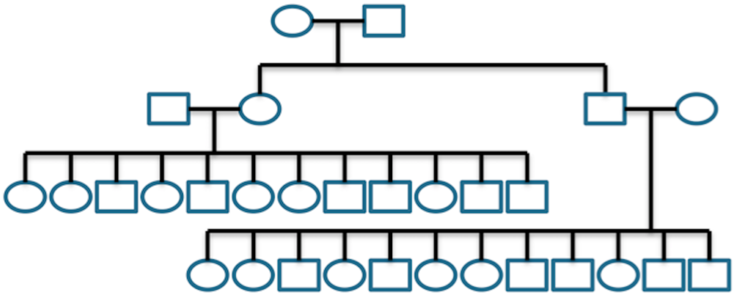
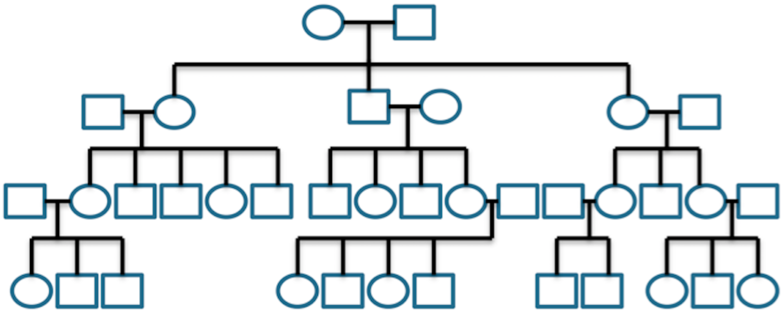
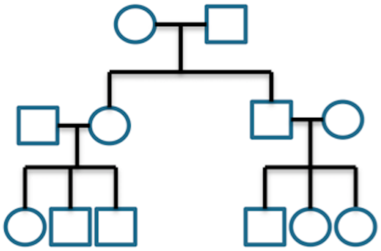
Pedigree type	Pedigree structure
"Wide"	
"Deep"	
"Small"	

Figure S1. Three different pedigree structures used in the false positive rate and power simulations. The first pedigree type is “wide” family that has 30 individuals in three generations. The second pedigree type is “deep” family that has 36 individuals in four generations. The third type is “small” family that has 12 individuals in three generations.

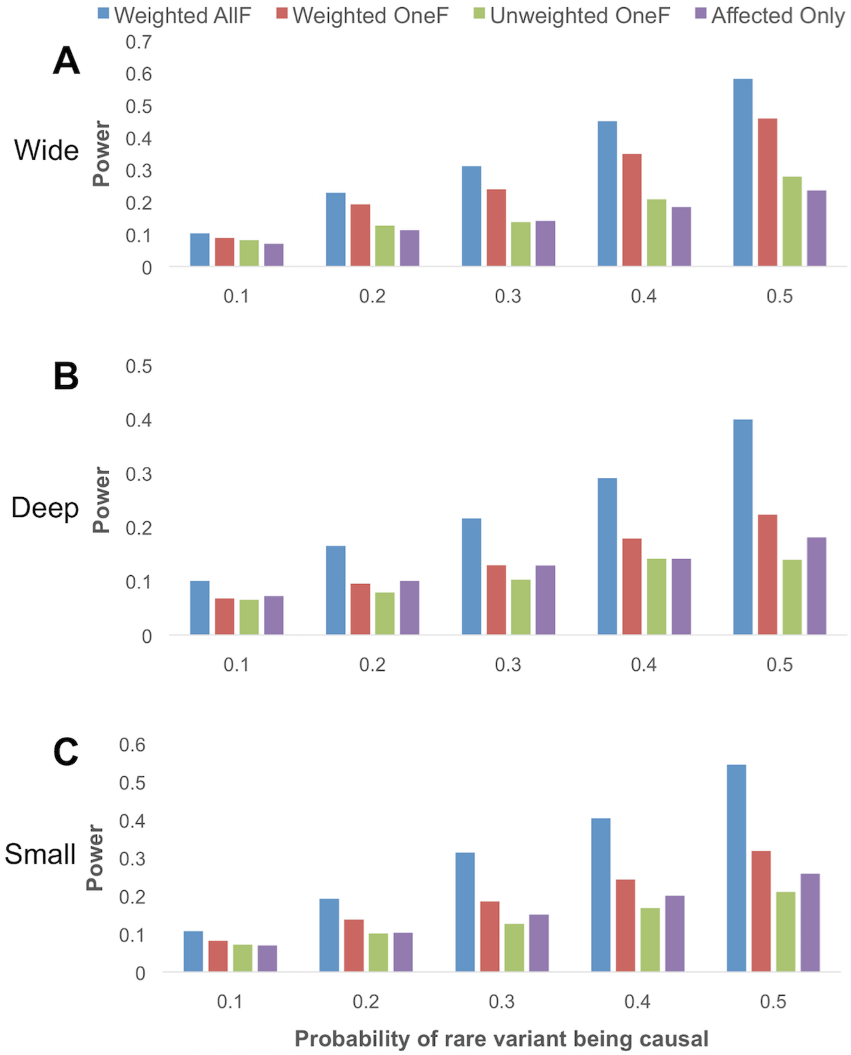
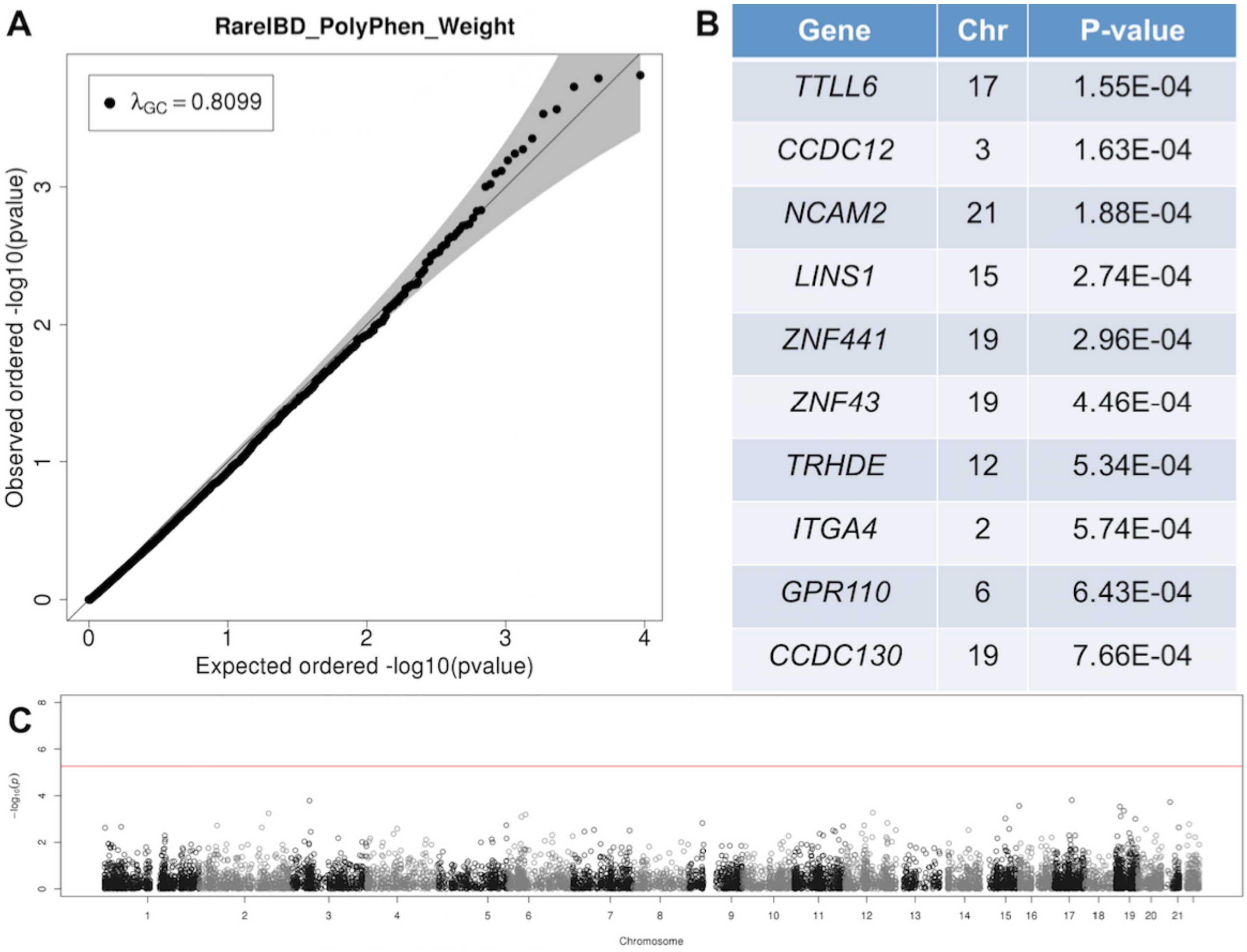
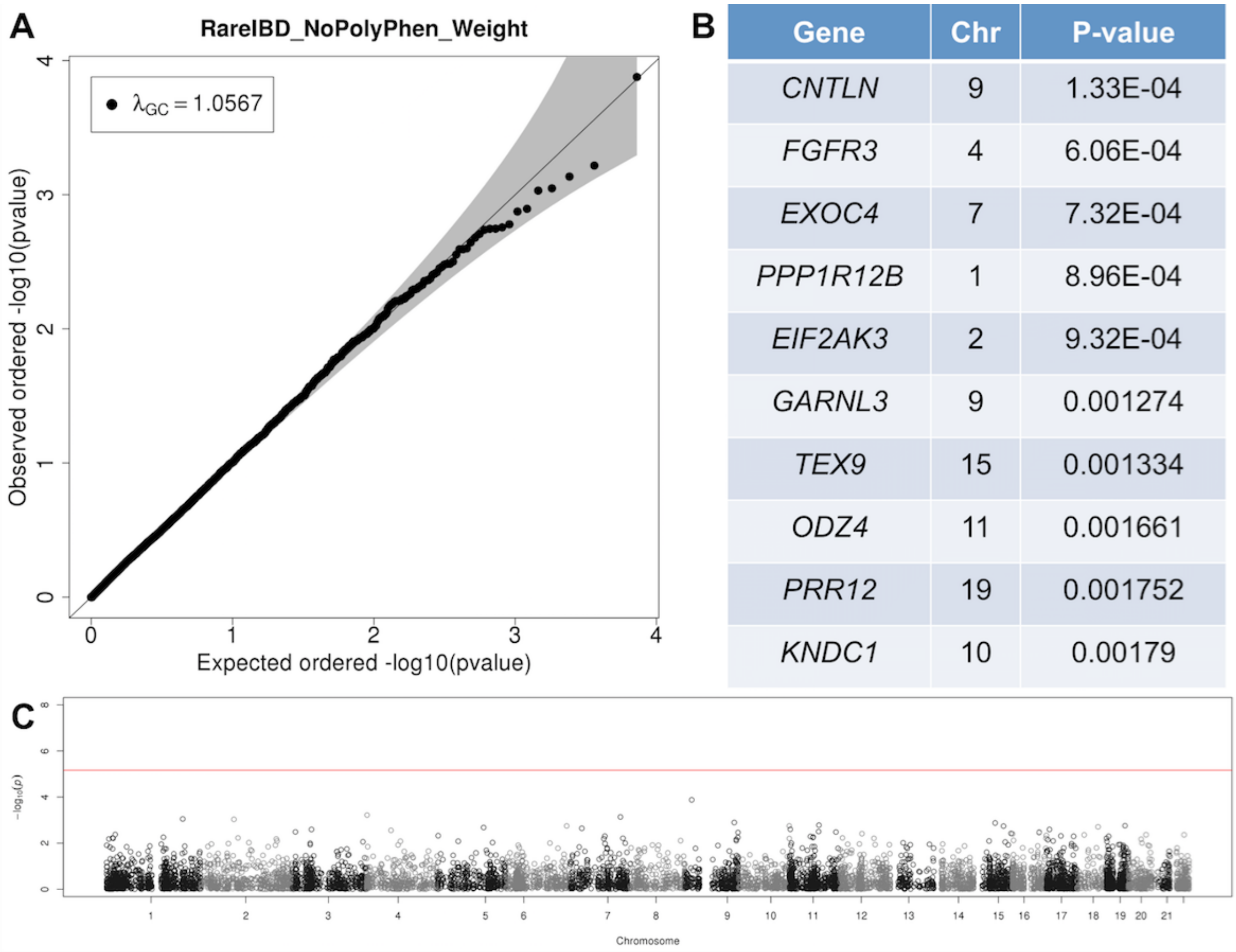


Figure S2. Power comparison of RareIBD with different settings using three different pedigree structures (Figure S1): wide families (A), deep families (B), and small families (C). In this simulation, all individuals are genotyped. We consider 4 different versions of RareIBD. 1) Weighted AllF is RareIBD that computes its statistic using mean and standard deviation (SD) of all founders ("AllF") with frequency-based and effect size-based weights. 2) Weighted OneF is RareIBD that computes its statistic using mean and SD of one founder who carries a mutation ("OneF") with the weights. 3) Unweighted OneF is RareIBD with OneF, but does not include frequency-based and effect size-based weights. 4) Affected Only is RareIBD with weighted AllF, but uses only affected individuals when computing its statistic. Power is measured at $\alpha = 0.05$ from 2,000 replications of simulations.



Gene	Chr	P-value
TTLL6	17	1.55E-04
CCDC12	3	1.63E-04
NCAM2	21	1.88E-04
LINS1	15	2.74E-04
ZNF441	19	2.96E-04
ZNF43	19	4.46E-04
TRHDE	12	5.34E-04
ITGA4	2	5.74E-04
GPR110	6	6.43E-04
CCDC130	19	7.66E-04

Figure S3. Results of applying RareIBD with PolyPhen-2 weighting to whole-exome sequencing data of extended families with EOCOPD. There are 347 individuals in this dataset. (A) is the QQ-plot showing the distribution of p-values of 12,092 genes that contain at least 3 rare variants, and it also indicates λ_{GC} values. (B) shows the top 10 genes with most significant p-values, and (C) is the Manhattan plot of p-values along the chromosomes.



Gene	Chr	P-value
CNTLN	9	1.33E-04
FGFR3	4	6.06E-04
EXOC4	7	7.32E-04
PPP1R12B	1	8.96E-04
EIF2AK3	2	9.32E-04
GARNL3	9	0.001274
TEX9	15	0.001334
ODZ4	11	0.001661
PRR12	19	0.001752
KNDC1	10	0.00179

Figure S4. Results of applying RareIBD without PolyPhen-2 weighting to microarray and exome-chip data of CFS African Americans (AA). There are 632 individuals in this dataset. (A) is the QQ-plot showing the distribution of p-values of 7,267 genes that contain at least 3 rare variants, and it also indicates λ_{GC} values. (B) shows the top 10 genes with most significant p-values, and (C) is the Manhattan plot of p-values along the chromosomes.

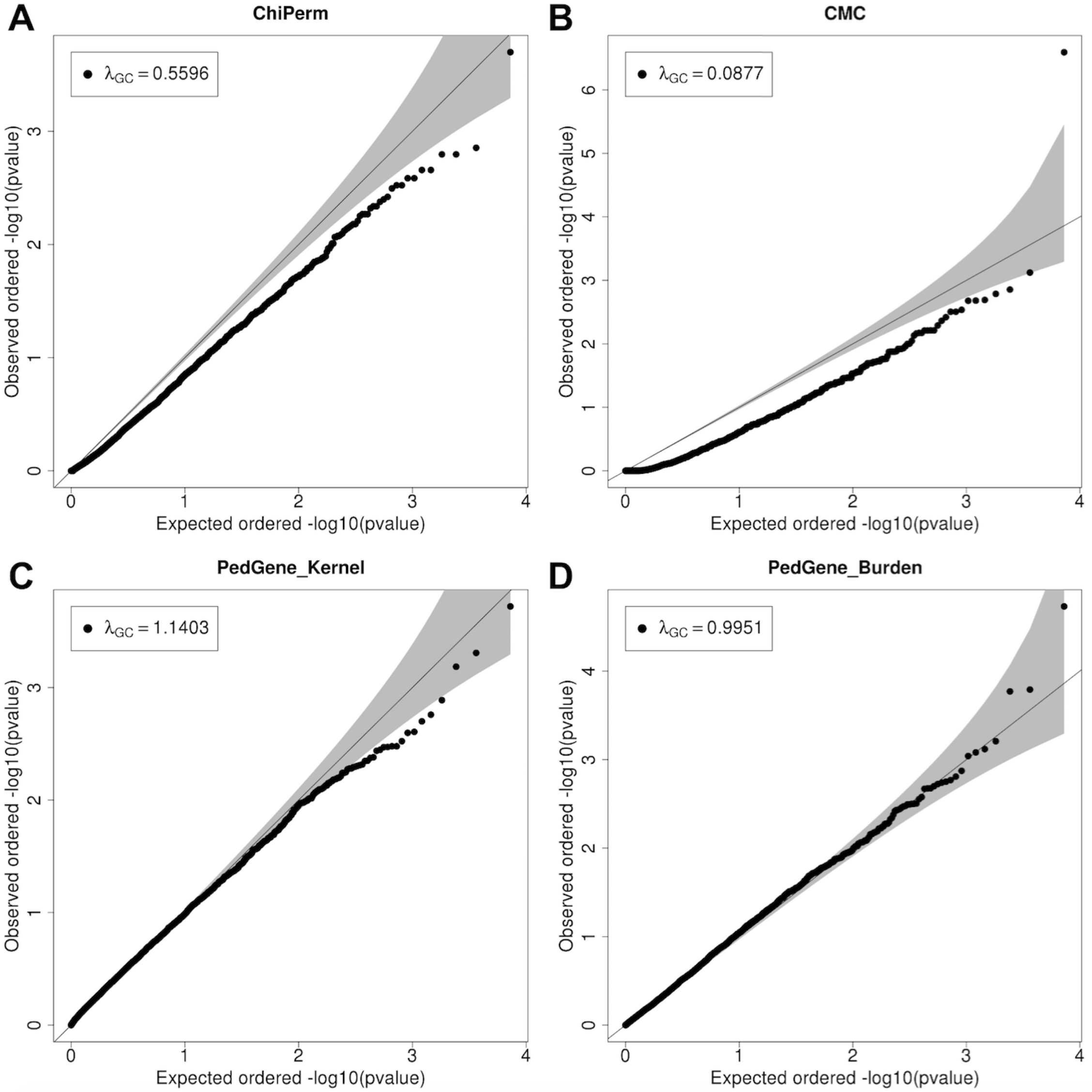
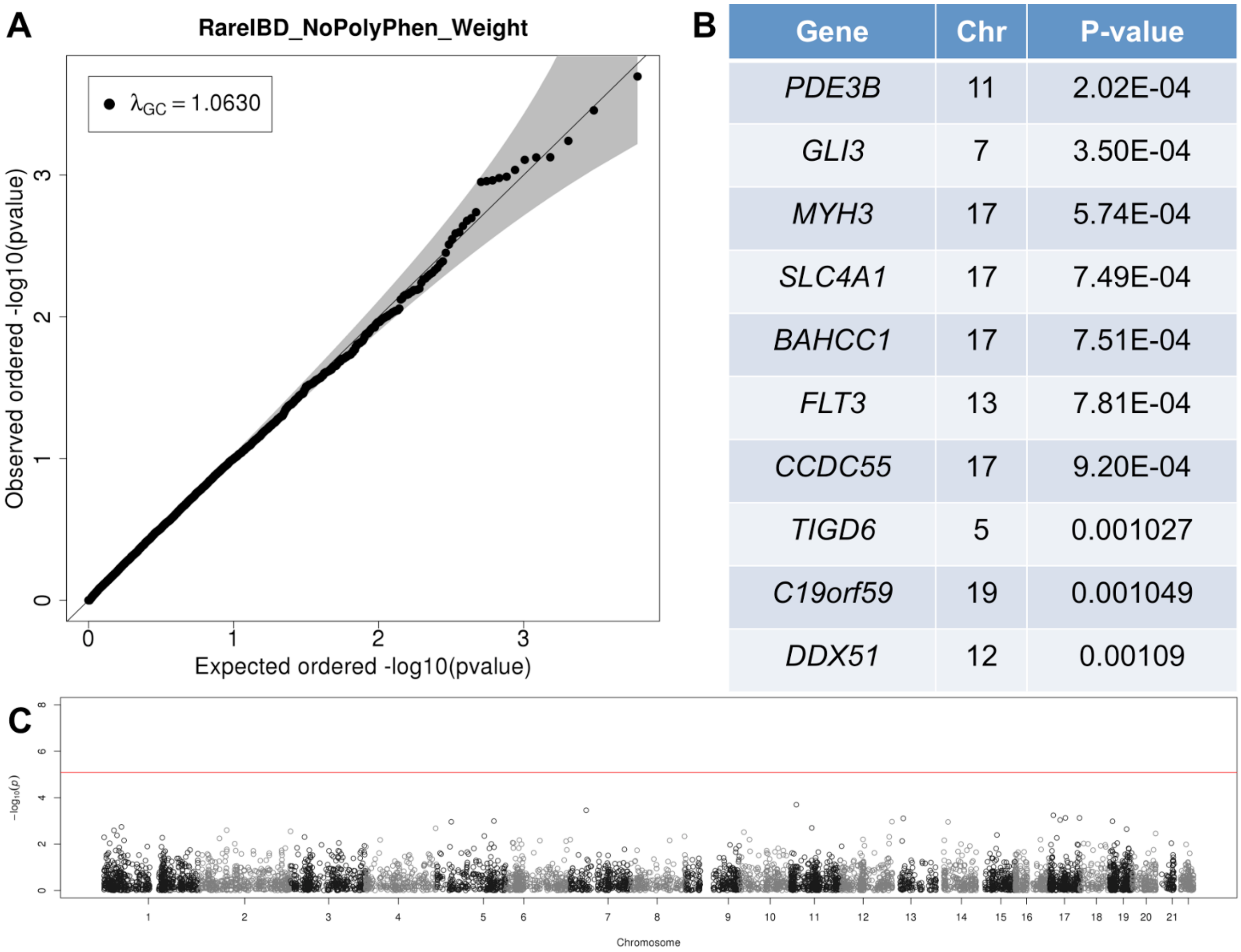


Figure S5. Results of applying FPCA and Pedgene software to microarray and exome-chip data of CFS African Americans (AA). These are QQ-plots from ChiPerm of FPCA (A), CMC of FPCA (B), kernel approach of Pedgene (C), and burden approach of Pedgene (D). All QQ-plots include λ_{GC} values.



Gene	Chr	P-value
PDE3B	11	2.02E-04
GLI3	7	3.50E-04
MYH3	17	5.74E-04
SLC4A1	17	7.49E-04
BAHCC1	17	7.51E-04
FLT3	13	7.81E-04
CCDC55	17	9.20E-04
TIGD6	5	0.001027
C19orf59	19	0.001049
DDX51	12	0.00109

Figure S6. Results of applying RareIBD without PolyPhen-2 weighting to microarray and exome-chip data of CFS Europeans (EA). There are 710 individuals in this dataset. (A) is the QQ-plot showing the distribution of p-values of 6,110 genes that contain at least 3 rare variants, and it also indicates λ_{GC} values. (B) shows the top 10 genes with most significant p-values, and (C) is the Manhattan plot of p-values along the chromosomes.

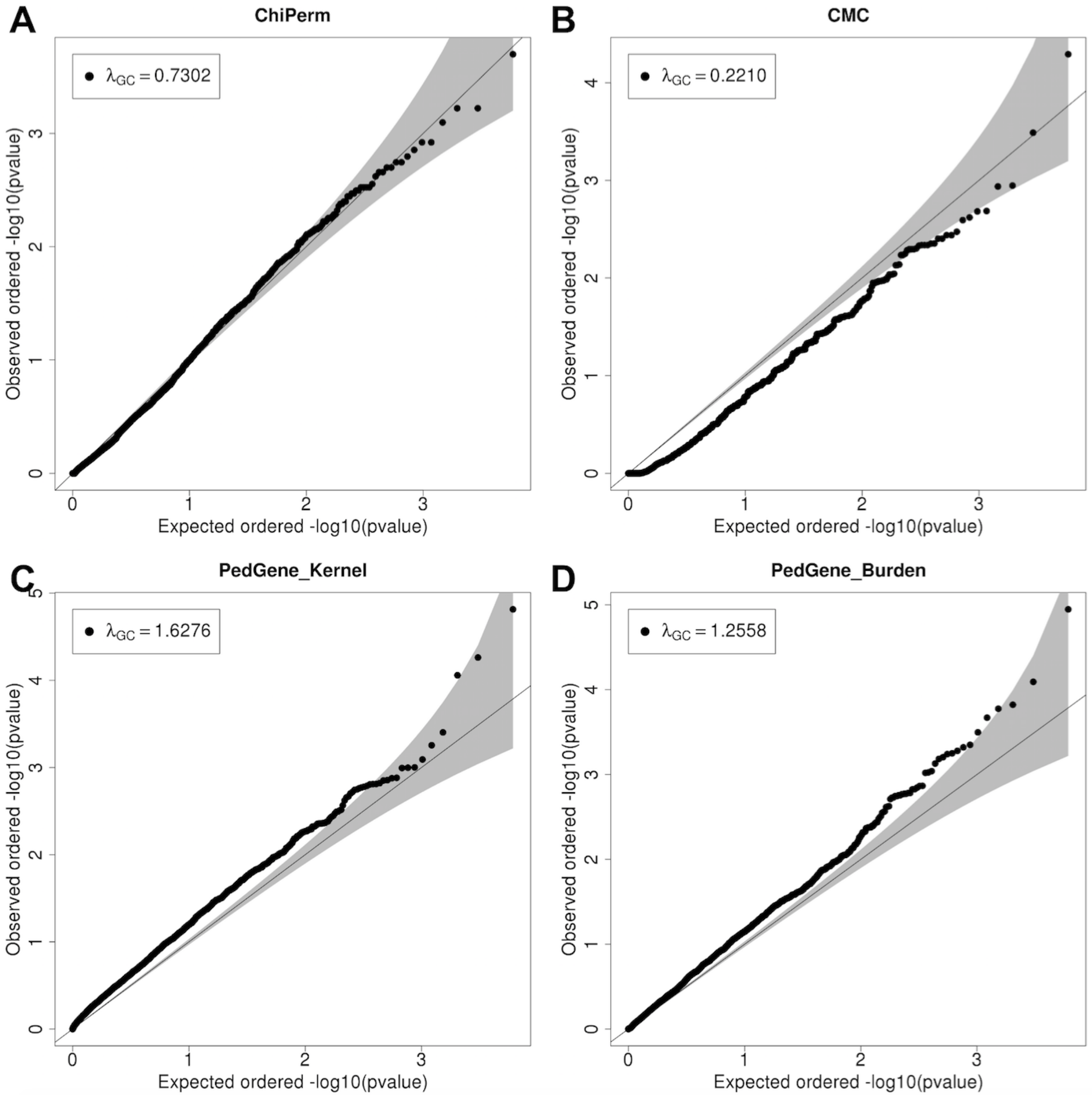


Figure S7. Results of applying FPCA and Pedgene software to microarray and exome-chip data of CFS Europeans (EU). These are QQ-plots from ChiPerm of FPCA (A), CMC of FPCA (B), kernel approach of Pedgene (C), and burden approach of Pedgene (D). All QQ-plots include λ_{GC} values.

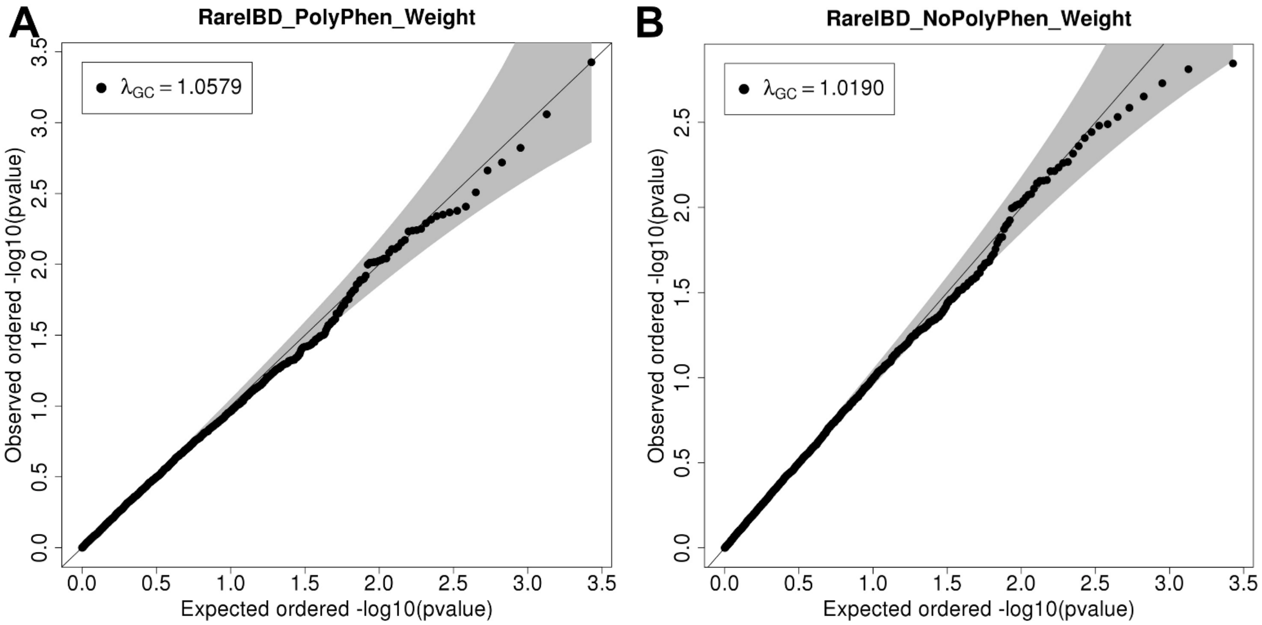


Figure S8. Results of applying RareIBD to the merged dataset of CFS-AA and CFS-EU with PolyPhen-2 weighting (A) and without PolyPhen-2 weighting (B). Because the two datasets were genotyped in different microarray platforms, we merged them by using only SNPs present in both datasets. We removed two families in which both microarray platforms were used to genotype different individuals in those families to remove batch effect within a family. The number of individuals is 1,216 and the number of SNPs is 226,489 after merging the two datasets. We estimated MAF of each variant separately for EU and AA, and used the MAF of population to which a family belongs in determining whether each variant is rare or not for the family (MAF $<1\%$). Only genes with at least 3 rare variants are included in the analysis, and there are 2,680 such genes.

Method	Wide	Deep	Small
Weighted AllF	0.0475	0.0466	0.0517
Weighted OneF	0.0503	0.0462	0.0514
Unweighted OneF	0.0487	0.0504	0.0484
Affected Only	0.0491	0.0471	0.0541

Table S1. Comparison of false positive rate of RareIBD with different improvements discussed in Materials and Method using three different pedigree structures (Figure S1): wide, deep, and small families. In this simulation, all individuals are genotyped. We consider 4 different versions of RareIBD. 1) Weighted AllF is RareIBD that computes its statistic using mean and standard deviation (SD) of all founders ("AllF") with frequency-based and effect size-based weights. 2) Weighted OneF is RareIBD that computes its statistic using mean and SD of one founder who carries a mutation ("OneF") with the weights. 3) Unweighted OneF is RareIBD with OneF, but does not include frequency-based and effect size-based weights. 4) Affected Only is RareIBD with weighted AllF, but uses only affected individuals when computing its statistic. False positive rate is measured at $\alpha = 0.05$ from 10,000 replications of simulations.

Software	Method	Wide	Deep	Small
RareIBD	RareIBD	0.0528	0.0585	0.0631
FPCA	FPCA	4.00E-04	1.70E-03	1.00E-04
	ChiPerm	0.0473	0.0495	0.0455
	ChiMin	0.5432	0.3136	0.2447
	T2	0.0867	0.062	0.2043
	CMC	0.0609	0.0565	0.0534
Pedgene	Kernel	0.0339	0.0394	0.0218
	Burden	0.0666	0.0626	0.100

Table S2. Comparison of false positive rate (FPR) of RareIBD with those of other approaches when two rare variants are present in a family. We measure FPR using three different pedigree structures (Figure S1): wide, deep, and small families. Each family has 30% probability that two founders carry the same rare variant. We assume that top two generations are missing in this simulation. False positive rate is measured at $\alpha = 0.05$ from 10,000 replications of simulations.

Statistic	Summary	EOCOPD	CFS-AA	CFS-EU
family size	minimum	6	3	4
	maximum	23	56	28
	mean	12.6	11.4	11.7
	median	12	10	9.5
percentage of genotyped individuals	all individuals	56.0%	41.0%	50.0%
	founders	6.2%	18.2%	31.4%
	nonfounders	87.4%	59.2%	60.7%
family depth	minimum	2	2	2
	maximum	5	5	5
	mean	3.4	3.0	3.2
	median	3	3	3
relationship among affected pairs	minimum	0.125	0.0625	0.0625
	maximum	0.5	0.5	0.5
	mean	0.42	0.39	0.41
	median	0.5	0.5	0.5

Table S3. Detailed information on family structure of EOCOPD, CFS-AA, and CFS-EU datasets. The "family size" is the number of individuals in a family including individuals who were not genotyped. The "percentage of genotyped individuals" is calculated for all individuals, only founders, and only nonfounders in a family. The "family depth" of 2 is parent-offspring relationship. The "relationship among affected pairs" is the coefficients of relationship of affected pairs who are related in a family.