<div align="center">

Online Supplementary Materials for
FamLBL: Detecting Rare Haplotype Disease Association Based on Common SNPs Using Case-Parent Triads

Meng Wang, Shili Lin

</div>

A.1 Rare Haplotypes Tagging Rare Causal SNVs

We carried out a small simulation to compare famLBL with LBL and several popular "collapsing" methods: CMC (Li and Leal, 2008), fbat-v0 (De et al., 2013) (first-order methods) and SKAT (Wu et al., 2011) (second-order method). The disease model presented in Tables S1 and S2 portraits a cis-acting mechanism in which a rare haplotype tags three underlying rare causal single nucleotide variants (SNVs). The result in Figure S1 shows that famLBL (red curves) can be more powerful than "collapsing" methods. Although it is less powerful than LBL as expected since there is population homogeneity, the drop in power is not substantial.

A.2 Analysis of Framingham Heart Study Data

We consider 9 SNPs (Table S6) that were identified to be associated with hypertension in Han et al. (2013). Note that although there were 10 SNPs listed in the top segment of Table 4 in Han et al. (2013), SNP No. 9 has the same rs number as SNP No. 3, and as such we do not consider SNP No. 9 as we cannot independently verify that they are two different SNPs. These 9 SNPs are referred to as target SNPs in the following discussion. For each target SNV in Table S6, haplotypes in 7-SNP regions are considered, with the target SNP positioned in the 1st, 2nd, or the last spot. All haplotypes that are inferred to be significantly associated with hypertension by famLBL are given in Table S7. As we can see from the table, there are both risk haplotypes (OR > 1) and protective haplotypes (OR < 1), but all the results for each target SNP are completely consistent in that if haplotypes involving the "**0**" allele in the target SNP are risk ones (e.g. haplotype 0100**0**01 involving rs2229188), then those involving the "**1**" allele are protective (e.g. haplotype 0100**0**11 involving rs2229188), and vice versa. It is interesting to see that there are many haplotypes involving rs2229188 that are inferred to be strongly associated (with very large BF values) with hypertension, as this SNP has been implicated to be associated with cardiovascular disease in the literature (Wang et al. 2008). Further, for all risk haplotypes in which rs2229188 is not positioned at the first or last spot, each haplotype involves a 0**01** sub-haplotype, suggesting that this 3-SNP haplotype may be tagging rare causal SNVs interacting in a cis-fashion.

To explore the sensitivity of the results on the selection of the window size, we also consider haplotypes spanning 5-SNP regions, again with the target SNP positioned in the first, second, or the last spot. The results (Table S8) are largely consistent with those in Table S7 in that most of the significant 5-SNP haplotypes are sub-haplotypes of corresponding significant 7-SNP haplotypes, although there are a couple exceptions. This application illustrates the usage of famLBL as a follow-up tool to further investigate regions previously implicated by other methods to better understand the underlying mechanisms.

<div align="center">

1

</div>

Table S1. Distribution and Odds Ratios (ORs) of haplotypes. There is only one causal haplotype (OR > 1) and it is a rare one (frequency < 0.05).

hap	freq	OR
01001	**0.005**	3
11010	**0.025**	1
10100	0.42	1
00011	**0.01**	1
10110	0.3	1
00100	0.24	1

Table S2. SNV frequencies and odds ratios (ORs) corresponding to the model given in Table S1. All three causal SNVs are rare.

SNV	MAF	OR
1	0.255	1
2	**0.03**	1.3
3	**0.04**	-1.2
4	0.335	1
5	**0.015**	1.6

Table S3. Population prevalence corresponding to phenocopy rate of 0.05 (top segment) and 0.1 (bottom segment) under the three haplotype settings (HS1, HS2, HS3) and the three disease models (RR, RC, C) given in Table 1 of main text.

	RR	RC	C
	Phenocopy Rate $= 0.05$		
HS1	0.052	0.061	0.060
HS2	0.052	0.058	0.057
HS3	0.052	0.058	0.057
	Phenocopy Rate $= 0.1$		
HS1	0.103	0.120	0.119
HS2	0.103	0.115	0.113
HS3	0.103	0.115	0.113

Table S4. Comparison of LBL and famLBL type I error rates with the threshold BF $= 2$ intended to control the type I error at the 5% level. Each number (%), if present, is the type I error rate for LBL (before "/") or the type I error rate for famLBL (after "/") for an unassociated haplotype under the particular disease model (with phenocopy rate $= 0.1$). As can be seen from the table, the type I error rates for LBL and famLBL are comparable and they are all around or below 5%.

Setting	Haplotype	Frequency	LBL/famLBL		
			RR	RC	C
HS1	01100	0.3	0.8/0.0	0.8/1.0	1.0/0.2
	10100	0.005	–	–	1.4/1.0
	11011	0.01	–	0.4/1.0	4.8/4.4
	11100	0.155	0.6/0.4	5.4/5.6	2.4/3.4
	11111	0.11	1.6/1.4	–	–
HS2	01010	0.06	2.2/1.6	2.2/1.8	0.8/2.2
	01100	0.25	0.4/0.2	0.8/0.4	1.0/0.8
	10000	0.08	1.0/2.2	–	–
	10100	0.005	–	–	2.8/3.0
	11011	0.01	–	3.4/2.4	3.0/3.6
	11100	0.09	2.4/0.8	1.6/1.6	1.2/1.0
	11101	0.085	1.4/2.6	1.2/1.4	1.4/2.4
	11111	0.1	1.4/1.0	1.2/1.6	1.4/1.0
HS3	00111	0.07	0.6/1.0	1.0/2.0	1.8/1.8
	01000	0.02	3.0/2.2	3.0/3.2	2.4/2.4
	01011	0.05	2.8/2.8	2.4/1.0	1.4/1.2
	01101	0.06	2.2/1.0	2.4/2.2	2.4/1.0
	01110	0.14	1.8/1.0	0.6/0.8	1.6/1.0
	10010	0.08	1.4/1.8	–	–
	10100	0.005	–	–	2.6/4.2
	11011	0.01	–	3.8/1.8	3.6/4.4
	11101	0.09	0.4/1.0	2.4/1.0	0.8/1.6
	11110	0.13	0.6/1.0	1.2/0.8	1.2/1.4
	11111	0.1	1.4/1.4	1.0/1.0	1.4/1.0

Table S5. Comparison of LBL and famLBL type I error rates with the threshold BF = 6 intended to control the type I error at the 1% level. Each number (%), if present, is the type I error rate for LBL (before "/") or the type I error rate for famLBL (after "/") for an unassociated haplotype under the particular disease model (with phenocopy rate = 0.1). As can be seen from the table, the type I error rates for LBL and famLBL are comparable and they are all around or below 1%.

Setting	Haplotype	Frequency	LBL/famLBL RR	RC	C
HS1	01100	0.3	0.0/0.0	0.0/0.0	0.4/0.0
	10100	0.005	–	–	0.0/0.0
	11011	0.01	–	0.4/0.6	0.2/0.4
	11100	0.155	0.2/0.0	0.6/1.2	0.8/1.4
	11111	0.11	0.2/0.2	–	–
HS2	01010	0.06	0.4/0.6	0.8/0.4	0.4/0.6
	01100	0.25	0.0/0.0	0.0/0.0	0.0/0.2
	10000	0.08	0.2/1.0	–	–
	10100	0.005	–	–	1.4/0.2
	11011	0.01	–	1.0/0.2	0.4/0.6
	11100	0.09	0.6/0.2	0.2/0.4	0.4/0.2
	11101	0.085	0.2/0.6	0.2/0.2	1.0/1.0
	11111	0.1	0.2/0.0	0.4/0.2	0.2/0.2
HS3	00111	0.07	0.6/0.2	0.2/0.8	0.4/0.2
	01000	0.02	0.2/0.6	0.2/0.2	0.4/0.6
	01011	0.05	0.2/0.8	0.2/0.8	0.0/0.4
	01101	0.06	0.2/0.2	0.4/0.4	0.6/0.0
	01110	0.14	0.0/0.0	0.6/0.2	0.0/0.4
	10010	0.08	0.4/0.2	–	–
	10100	0.005	–	–	0.4/0.6
	11011	0.01	–	0.6/0.2	0.0/1.0
	11101	0.09	0.2/0.2	0.2/0.2	0.4/0.2
	11110	0.13	0.6/0.4	0.4/0.0	0.0/0.6
	11111	0.1	0.4/0.0	0.2/0.2	0.0/0.2

Table S6. Nine SNPs with most significant associations reported in Han et al. (2013).

SNP	Chrom.	Gene
rs684596	12	TCTN1
rs2229188	7	CYP51A1
rs1112438	3	TTC21A
rs7559838	2	HPCAL1
rs2736483	4	
rs16881524	5	NDUFS4
rs7657817	4	FAM13A
rs11149562	16	CDH13
rs17717907	7	ASB4

Table S7. Hypertension associated haplotypes panning 7-SNP regions, with LB and UB denote the lower and upper bounds of the Odds Ratio (OR). The bold type within each haplotype indicates the position and the allele of the target SNP.

SNP	Haplotype	Frequency	OR	LB	UB	BF
rs684596	0 1 0 1 1 **1** 0	0.006	0.082	0.003	0.806	10.481
rs2229188	0 1 1 0 0 0 **0**	0.023	18.065	2.702	212.512	> 100
	0 1 0 0 0 **0** 1	0.026	12.416	2.030	112.056	43.564
	0 1 0 0 0 **1** 1	0.222	0.334	0.102	0.948	5.363
	1 1 0 0 0 **0** 1	0.022	14.282	2.889	117.331	> 100
	0 1 1 0 **0** 1 1	0.058	3.963	1.157	13.410	8.213
	1 0 0 **0** 0 1 0	0.088	7.360	2.980	19.221	> 100
	0 0 0 **0** 1 0 0	0.022	26.950	5.534	243.471	> 100
	1 1 0 **0** 1 1 1	0.023	9.207	1.322	96.931	11.497
	0 0 **0** 1 0 0 0	0.022	27.058	5.496	311.064	> 100
	1 0 **0** 1 1 1 0	0.021	9.718	1.680	71.308	26.768
	0 **0** 1 0 0 0 0	0.023	31.785	7.092	185.304	> 100
	0 **0** 1 1 1 0 0	0.020	13.943	2.617	96.448	149.616
	0 1 0 0 0 0 1	0.020	37.151	7.029	256.381	> 100
	0 1 1 1 0 0 1	0.023	13.818	2.319	151.866	82.635
rs7559838	0 0 0 0 0 0 **0**	0.079	2.184	1.066	4.923	3.513
	0 **1** 1 0 0 0 0	0.064	0.231	0.040	0.863	6.945
rs2736483	0 **1** 1 1 0 1 0	0.165	0.418	0.169	0.917	4.595
rs16881524	0 **0** 1 0 1 0 0	0.110	0.388	0.145	0.940	4.232
	1 1 1 1 0 0 1	0.028	2.942	1.276	6.228	12.573
rs7657817	1 **0** 1 1 0 0 0	0.024	3.397	1.070	12.305	4.703

Table S8. Hypertension associated haplotypes spanning 5-SNP regions, with LB and UB denote the lower and upper bounds of the Odds Ratio (OR). The bold type within each haplotype indicates the position and the allele of the target SNP.

| SNV	haplotype	frequency	OR	LB	UB	BF
rs684596	1 0 1 0 **0**	0.008	3.089	1.006	2.516	3.916
rs2229188	0 1 1 0 **0**	0.0246	10.340	1.551	124.21	20.717
	1 0 0 0 **0**	0.0289	26.682	5.540	197.552	> 100
	0 0 0 **0** 1	0.026	44.656	9.526	348.975	> 100
	1 1 0 **0** 1	0.025	10.423	1.640	88.588	24.652
	0 0 **0** 1 0	0.027	31.500	6.424	304.600	> 100
	1 0 **0** 1 1	0.023	11.381	1.674	152.475	36.035
	0 **0** 1 0 0	0.022	43.860	8.191	474.376	> 100
	0 **0** 1 1 1	0.020	20.491	3.525	200.537	> 100
	0 1 0 0 0	0.023	38.398	7.486	369.814	> 100
	0 1 1 1 0	0.018	31.000	4.509	456.688	> 100
rs7559838	0 0 0 0 **0**	0.117	2.170	1.091	4.500	4.432
rs16881524	1 1 1 **1** 0	0.024	3.632	1.557	7.760	34.446
rs7657817	**0** 1 1 0 0	0.083	2.645	1.085	6.430	4.725
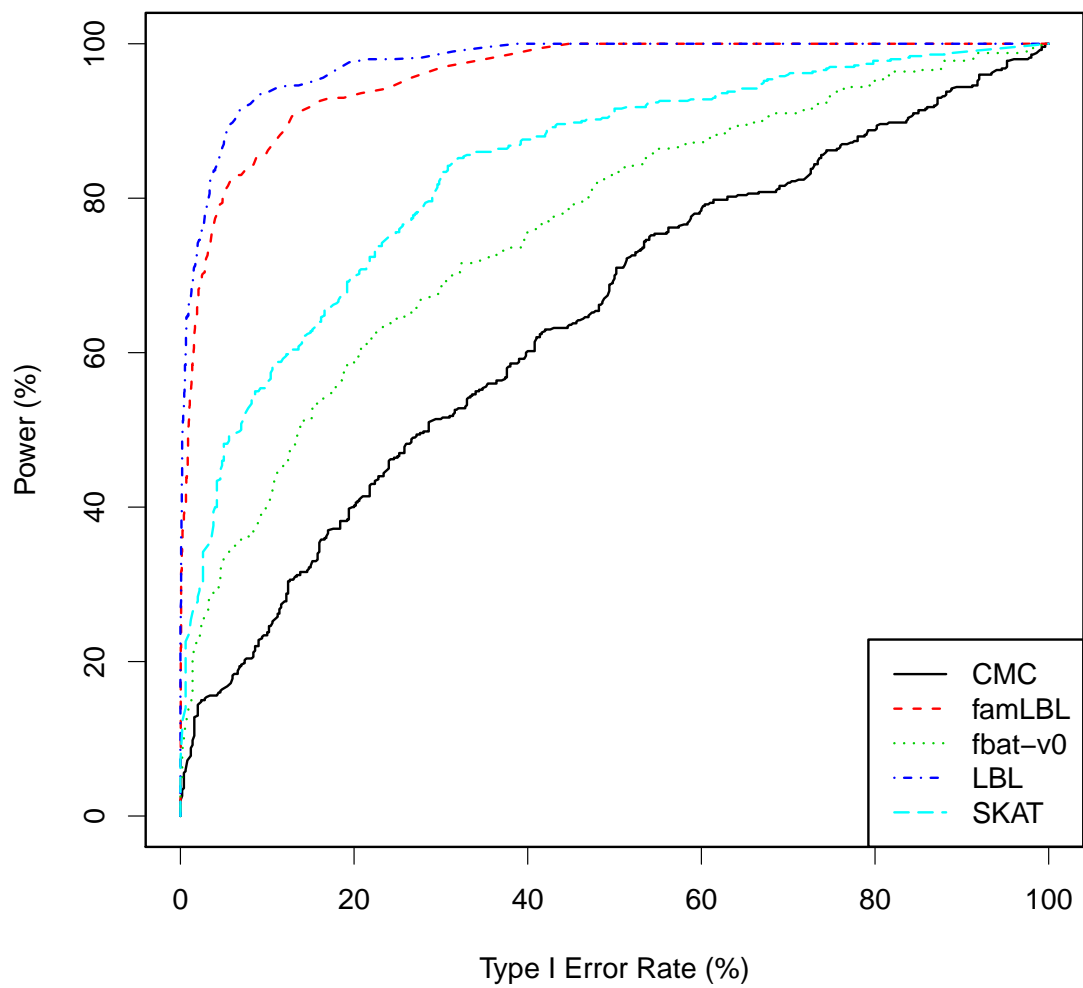
Figure S1. Receiver Operating Characteristic (ROC) curves comparing the performance of haplotype-based methods (famLBL, LBL) versus "collapsing" methods (CMC, fbat-v0, SKAT) under population homogeneity.
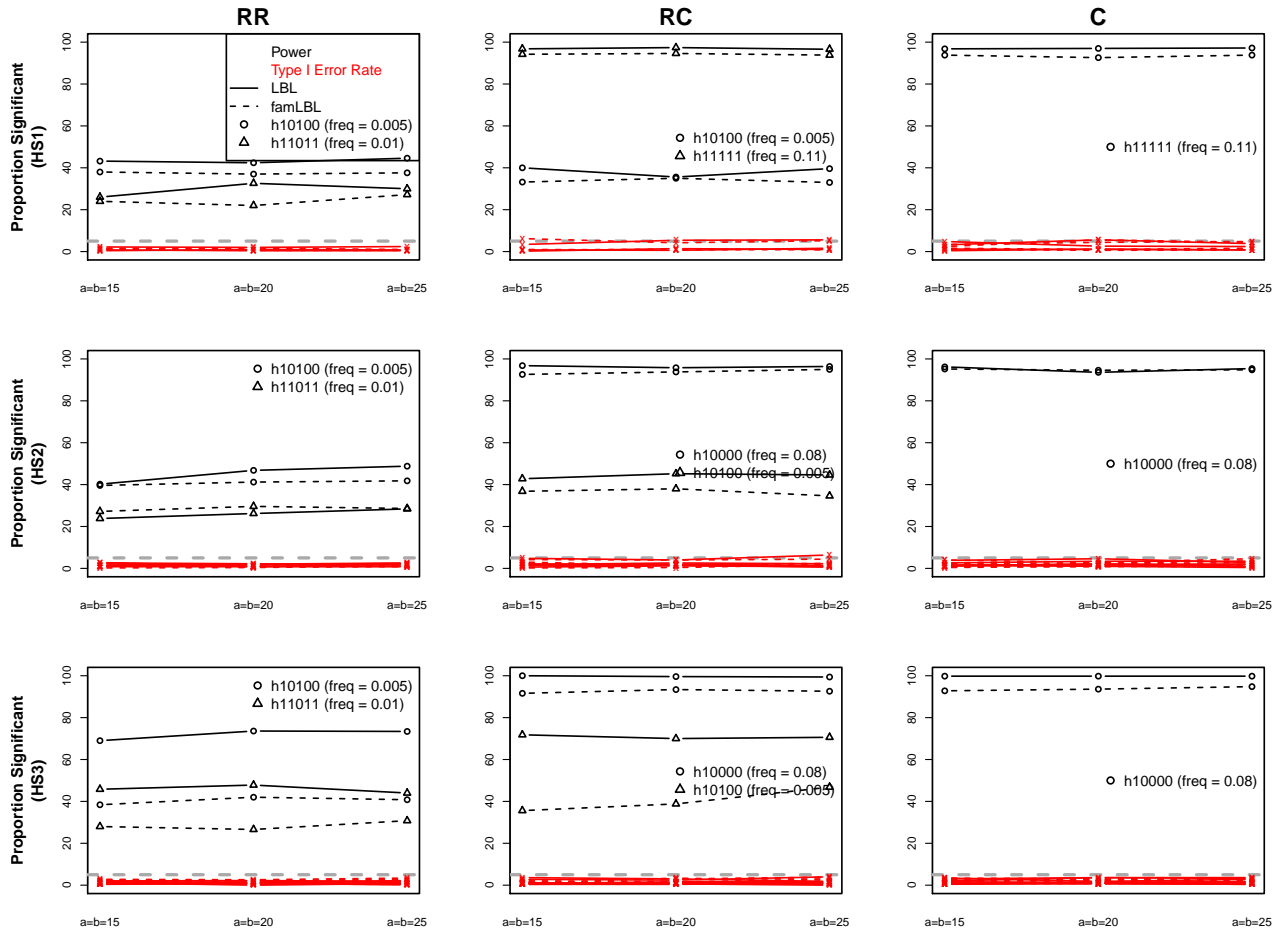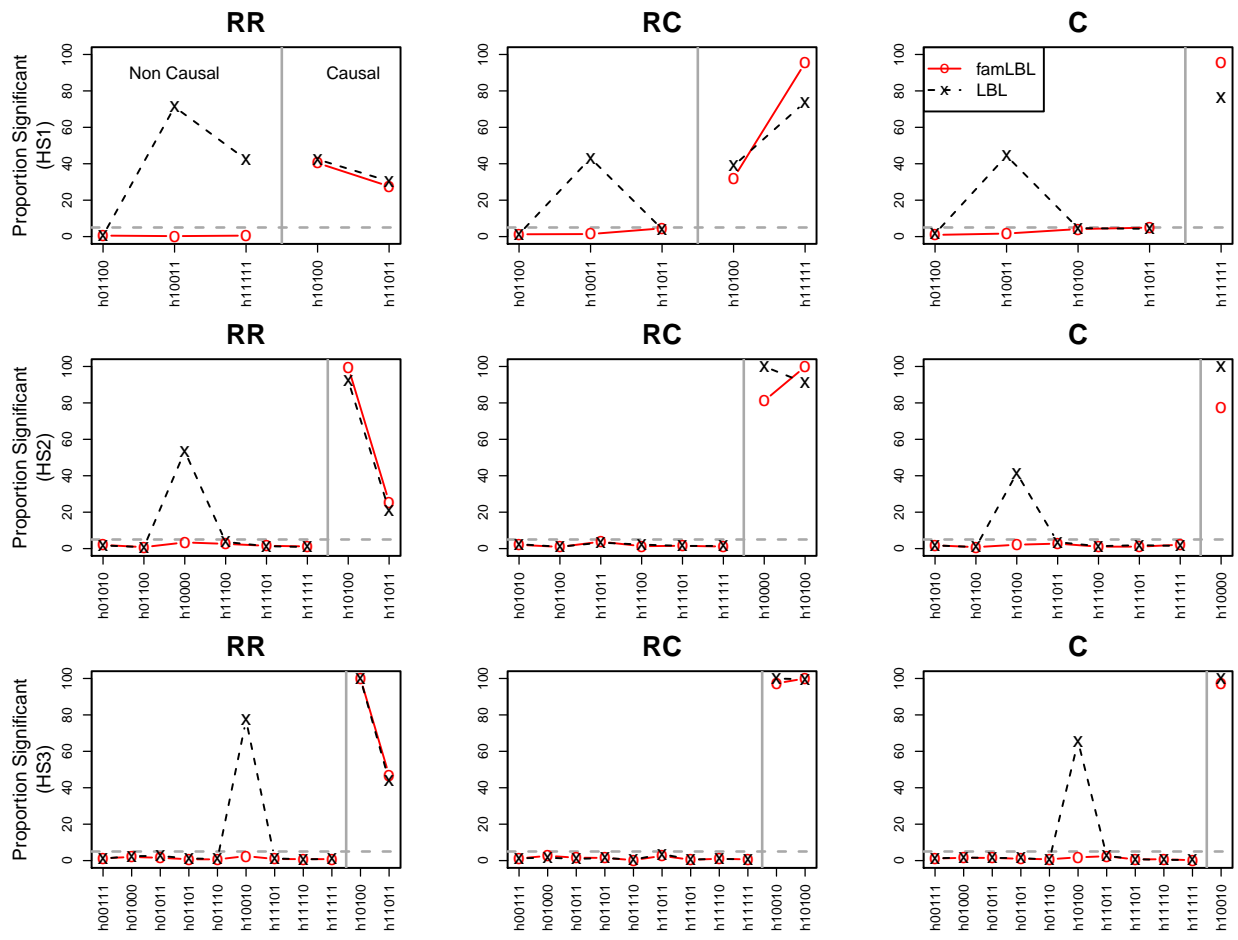
Figure S2. Comparison of power and type I error rate between famLBL (using trio data) and LBL (using case-control data) under population homogeneity. The black lines represent power for detecting specific haplotypes whereas the red lines portrait type I error; solid: LBL, dashed: famLBL. Phenocopy rate for all models is 0.05.

Figure S3. Comparison of type I error rate and power between famLBL (using trio data) and LBL (using case-control data) in the presence of population stratification. Type I error for non-causal haplotypes are plotted on the left and power for causal haplotypes are plotted on the right; black "x": LBL, red "o": famLBL. Phenocopy rate for all models is 0.05.
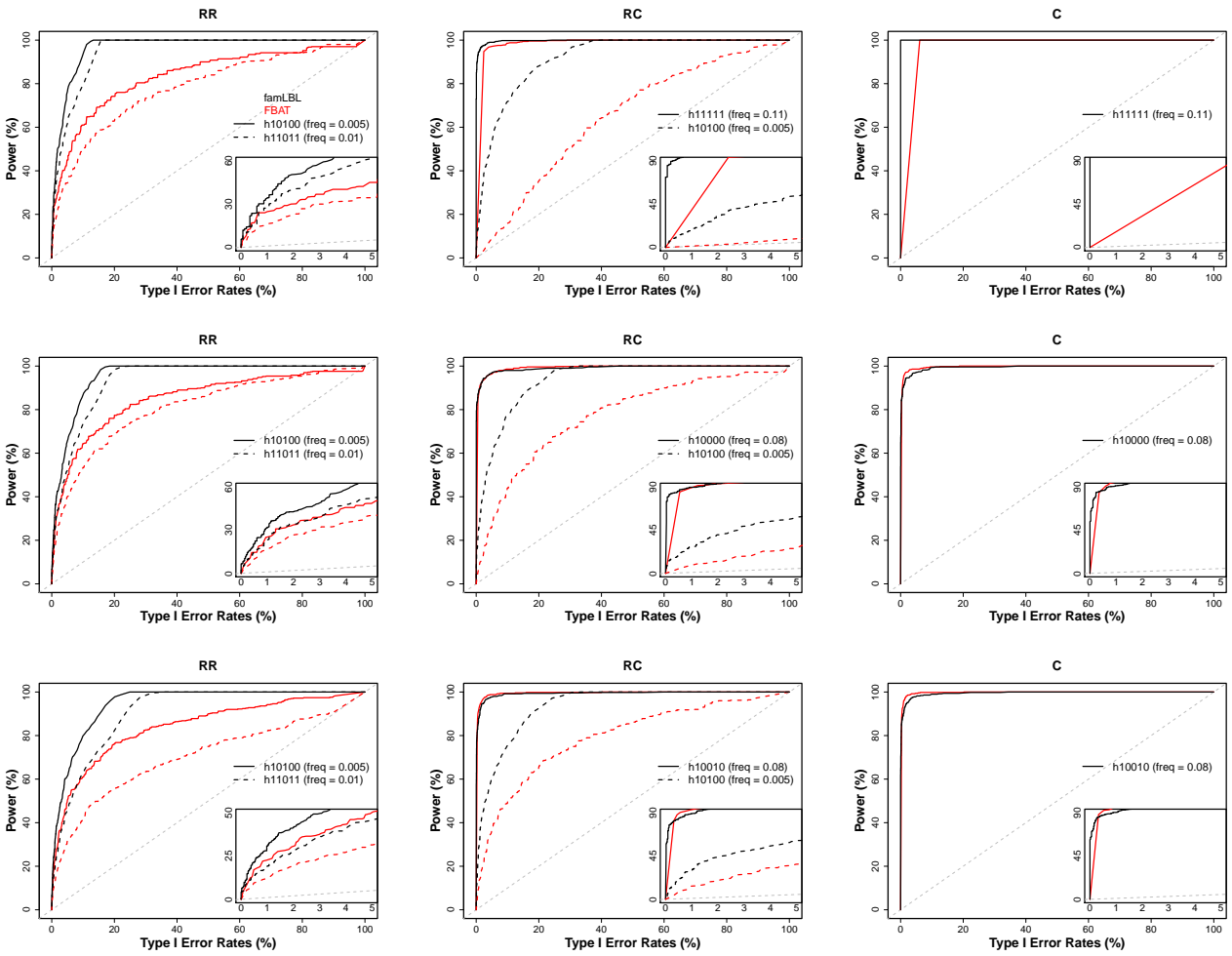
Figure S4. Receiver Operating Characteristic (ROC) curves comparing the performance of famLBL (black curves) with FBAT (red curves). Inset in each plot zooms in on the portion where type I error is at most 5%. A diagonal line is also added for reference. Phenocopy rate for all models is 0.05.

References

1. De, G., Yip, W.-K., Ionita-Laza, I., and Laird, N. (2013) Rare Variant Analysis for Family-Based Design. PLOS ONE, 8(1).

2. Han, M., Hu, Y.-Q., and Lin, S. (2013) Joint detection of association, imprinting and maternal effects using all children and their parents. EUROPEAN JOURNAL OF HUMAN GENETICS, 21, 14491456.

3. Li, B. and Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. The American Journal of Human Genetics, 83, 311 321.

4. Wang, Y., Rogers, P., Su, C., Varga, G., Stayrook, K.,Burris, T. (2008) Regulation of cholesterologenesis by the oxysterol receptor, LXRα. Journal of biological chemistry, 283, 26332-26339.

5. Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. The American Journal of Human Genetics, 89(1), 82 93.