Supplementary Materials for the manuscript "AGGrEGATOr: A Gene-based GEne-Gene interActTiOn test for case-control association studies."

Mathieu Emily*,[1]

1. Agrocampus Ouest - IRMAR UMR CNRS 6625, 65, rue de Saint Brieuc, 35042 Rennes Cedex, France

* To whom correspondence should be addressed: mathieu.emily@agrocampus-ouest.fr

S1 Accuracy of the estimation of the covariance matrix Σ

To perform the minP test, our procedure AGGrEGATOr requires the computation of the covariance matrix Σ of the Wald statistics W_{jk}. We propose the following formulation:

$$Cov(W_{jk}, W_{j',k'}) \approx r_{j,j'} r_{k,k'},$$

where $r_{j,j'}$ is the correlation measure of LD between SNP j and SNP j'. To assess the accuracy of our estimate, we performed simulations as follows. Let j, j', k and k' be four SNPs such as the Minor Allele Frequencies for j and j' are given by π_j and π_k for k and k'. We further assumed that the correlation LDs are $r_{j,j'}$ between j and j' and $r_{k,k'}$ between k and k'. For a given set of values π_j, π_k, $r_{j,j'}$ and $r_{k,k'}$, we estimated the empirical covariance between W_{jk} and $W_{j'k'}$ by simulating 10,000 phenotypes. Figure S1 displays the comparison of the empirical estimate with our estimate for $\pi_j = 0.45$, $\pi_k = 0.45$, $r_{j,j'} \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ and $r_{k,k'} \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. One can remark that our estimate is unbiased for all combinaison of $r_{j,j'}$ and $r_{k,k'}$. Similar accuracies are obtained for other Minor Allelic Frequencies $\pi_j \in \{0.05, 0.25, 0.45\}$ and $\pi_k \in \{0.05, 0.25, 0.45\}$ as shown in Figure S2.

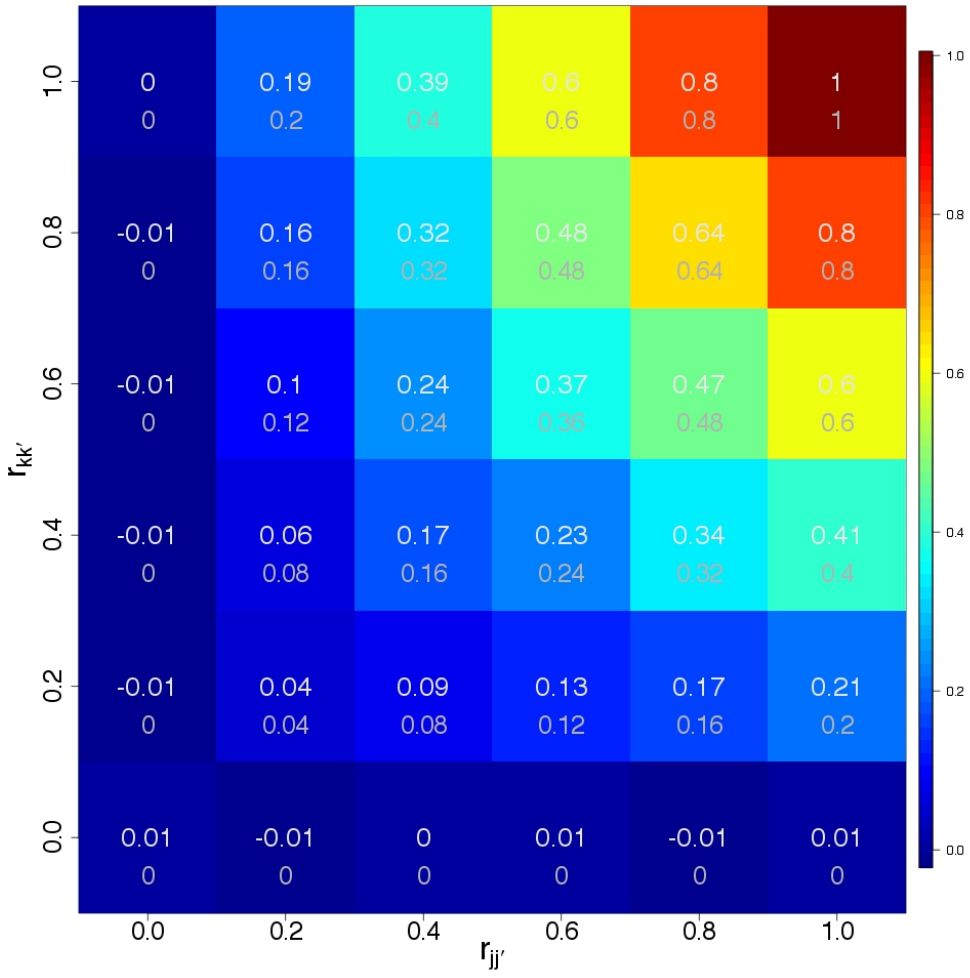

Figure S1: Emipirical estimation of the covariance between W_{jk} and $W_{j'k'}$ for Minor Allelic Frequencies given by $\pi_j = \pi_k = 0.45$. Each cell corresponds to a pair $(r_{jj'}, r_{kk'})$ where $r_{jj'}$ is the correlation LD between j and j'. The two values in each cell correspond to the empirical estimation in white and to our estimate in grey.

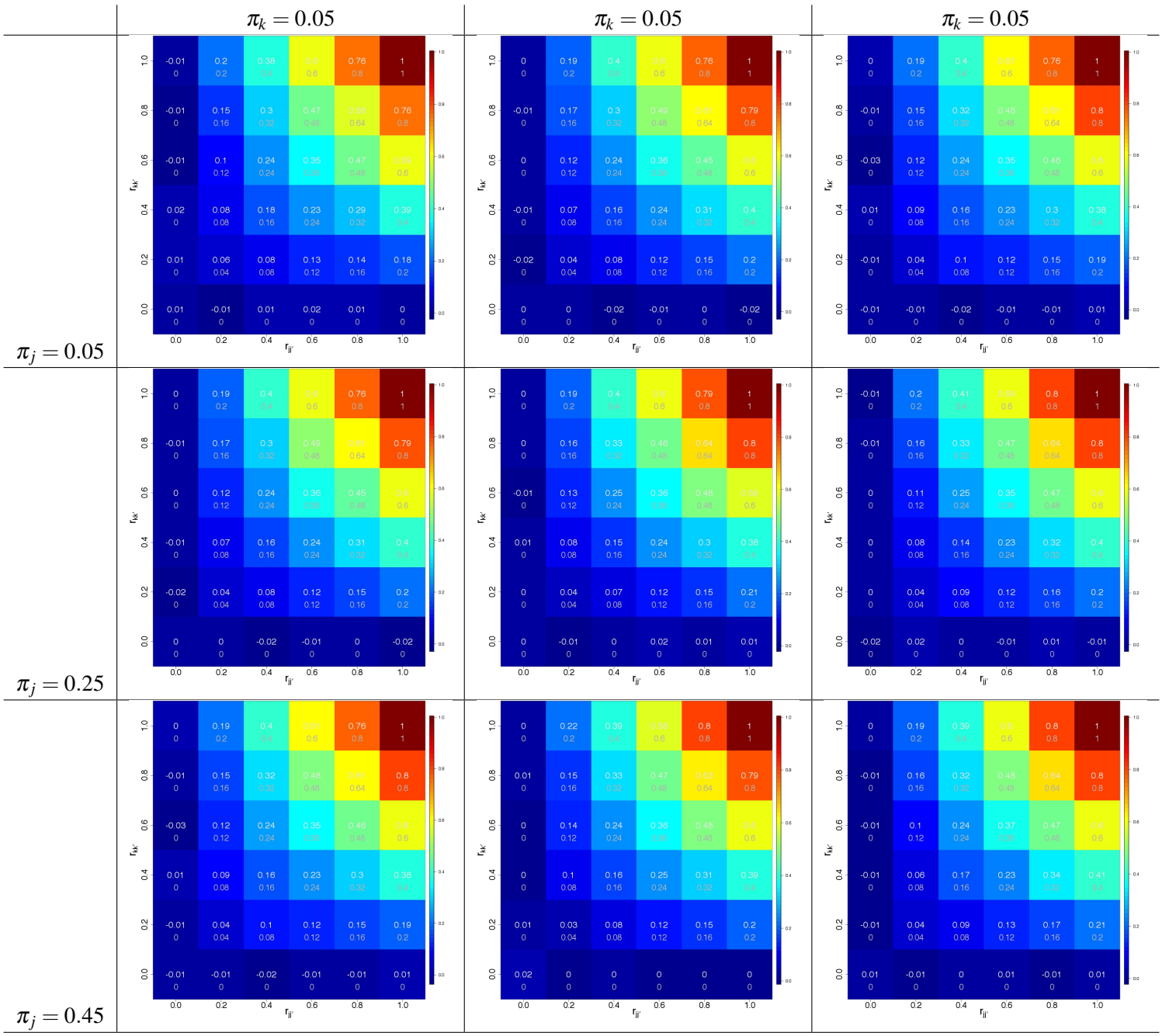

Figure S2: Emipirical estimation of the covariance between W_{jk} and $W_{j'k'}$ for Minor Allelic Frequencies given by $\pi_j \in \{0.05, 0.25, 0.45\}$ and $\pi_k \in \{0.05, 0.25, 0.45\}$. Each cell corresponds to a pair $(r_{jj'}, r_{kk'})$ where $r_{jj'}$ is the correlation LD between j and j'. The two values in each cell correspond to the empirical estimation in white and to our estimate in grey.

S2 Disease genetic models with 1 causal pair

For disease model with 1 causal pair, we used 3×3 table of odds as suggested by [4]. In such table, each cell characterizes the odds of the disease conditional to the corresponding genotype. Each model had two parameters: γ characterizes the baseline odds, *i.e.* the odds conditional to genotype pair *AABB*, and θ quantifies the strength of the model. Table S1 represents the 3 disease model without interaction and Table S2 displays the 8 disease models with epistatic effects.

Table S1: Table of odds for the three models (No effect, one marginal recessive and mulitplicative-multiplicative) without interaction between a pair of SNP.

	No effect			One marginal recessive			Multiplicative-multiplicative		
	BB	Bb	bb	BB	Bb	bb	BB	Bb	bb
AA	γ	γ	γ	γ	γ	γ	γ	$\gamma(1+\theta)$	$\gamma(1+\theta)^2$
Aa	γ	γ	γ	γ	γ	γ	$\gamma(1+\theta)$	$\gamma(1+\theta)^2$	$\gamma(1+\theta)^3$
aa	γ	γ	γ	$\gamma(1+\theta)$	$\gamma(1+\theta)$	$\gamma(1+\theta)$	$\gamma(1+\theta)^2$	$\gamma(1+\theta)^3$	$\gamma(1+\theta)^4$

Table S2: Table of odds for the 8 models with interaction between a pair of SNP.

	Recessive-recessive			Interaction-multiplicative			XOR		
	BB	Bb	bb	BB	Bb	bb	BB	Bb	bb
AA	γ	γ	γ	γ	γ	γ	γ	γ	$\gamma(1+\theta)$
Aa	γ	γ	γ	γ	$\gamma(1+\theta)$	$\gamma(1+\theta)^2$	γ	γ	$\gamma(1+\theta)$
aa	γ	γ	$\gamma(1+\theta)$	γ	$\gamma(1+\theta)^2$	$\gamma(1+\theta)^4$	$\gamma(1+\theta)$	$\gamma(1+\theta)$	γ

	Dominant-dominant			Threshold			Additive-additive		
	BB	Bb	bb	BB	Bb	bb	BB	Bb	bb
AA	γ	γ	γ	γ	γ	γ	γ	$\gamma(1+\theta)$	$2\gamma(1+\theta)$
Aa	γ	$\gamma(1+\theta)$	$\gamma(1+\theta)$	γ	γ	$\gamma(1+\theta)$	$\gamma(1+\theta)$	$2\gamma(1+\theta)$	$3\gamma(1+\theta)$
aa	γ	$\gamma(1+\theta)$	$\gamma(1+\theta)$	γ	$\gamma(1+\theta)$	$\gamma(1+\theta)$	$2\gamma(1+\theta)$	$3\gamma(1+\theta)$	$4\gamma(1+\theta)$

	Recessive-dominant			Special Interaction		
	BB	Bb	bb	BB	Bb	bb
AA	γ	γ	γ	γ	γ	$\gamma(1+\theta)$
Aa	γ	γ	$\gamma(1+\theta)$	γ	γ	γ
aa	γ	γ	$\gamma(1+\theta)$	$\gamma(1+\theta)$	γ	γ

S3 Details of pairs of loci used in the simulation studies

Table S3: Details of the SNPs used in the three pairs of loci (Locus1, Locus2), (PARP1, KRAS) and (GNPDA2, FAIM2).

Ma *et al.* [3]	
Locus 1	Locus 2
rs11589332	rs12712643
rs512854	rs11680220
rs539426	rs1558854
rs593911	rs759853
rs536662	rs10779925
rs668156	rs11891871
rs17186233	rs17467001
rs1441010	rs134202425
rs1441009	rs7585512
rs12563433	rs17532603
rs3828089	
rs1441008	
rs753425	
rs10737757	

Li *et al.* [2]	
PARP1	KRAS
rs7537552	rs3924649
rs7537636	rs12307733
rs10495278	rs7980769
rs9287011	rs11836162
rs12090413	rs11047882
rs12092786	
rs12093044	

Yuan *et al.* [6]	
GNPDA2	FAIM2
rs16857402	rs17201502
rs2709	rs905619
rs10020551	rs637871
rs4484337	rs1027711
rs12643262	rs956864
rs7670601	rs640081
	rs706795

S4 Simulation study based on gene pair PARP1 and KRAS

S4.1 LD pattern for gene pair PARP1 and KRAS

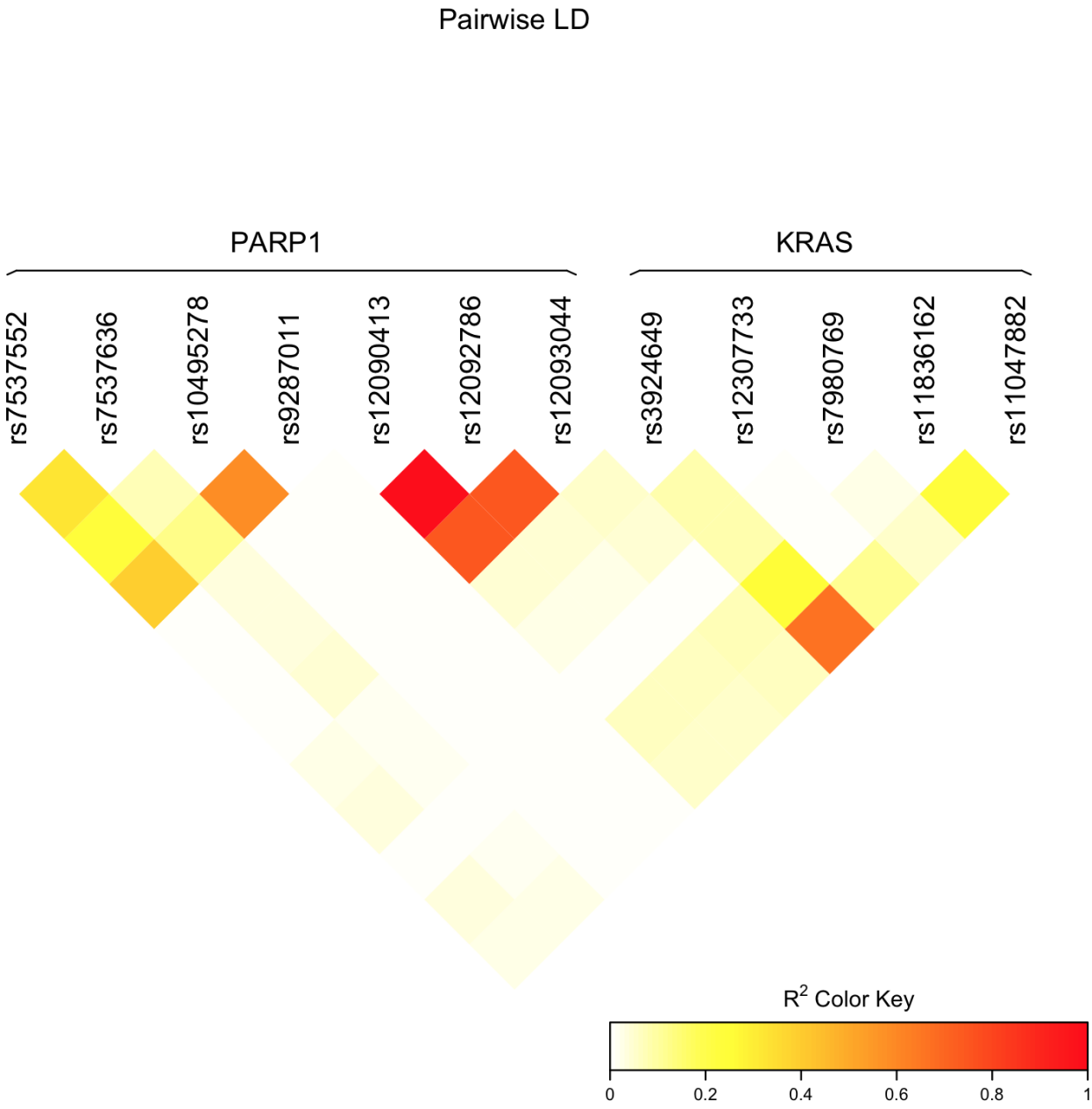

Figure S3: Pairwise r^2 within and between the 7 SNPs from PARP1 and the 5 SNPs from KRAS.

S4.2 Evaluation of the Type-I error rate based on gene pair PARP1 and KRAS

Table S4: Estimation of the false positive rate in several scenarios involving three disease models based on the pair PARP1-KRAS: a model with no effect, a model with one marginal (recessive) effect and a model with two (multiplicative) marginal effects. α is the expected predefined type-I error rate and θ is the parameter of the disease. Results with an $*$ indicate a significant deviation from the expected false positive rate.

Models	α	θ	AGGrEGATOr	CCA	KCCA	CLD	PCA	PSLPM	GBIGM
No effect	0.05		0.060	0.052	0.052	0.056	0.053	0.053	0.046
	0.01		0.010	0.008	0.009	0.009	*0.018**	0.010	0.007
One marginal effect	0.05	1	0.061	0.047	0.055	0.054	0.053	0.053	*0.034**
	0.01	1	0.017	0.013	0.012	0.006	0.010	0.009	0.007
	0.05	4	0.047	0.052	0.061	*0.078**	0.048	0.053	*0.017**
	0.01	4	0.010	0.006	0.014	*0.021**	0.008	0.014	*0.002**
Multiplicative marginal effects	0.05	1	0.044	0.043	0.057	*0.085**	0.039	*0.029**	*0.584**
	0.01	1	0.005	0.007	0.008	*0.021**	0.011	0.007	*0.322**
	0.05	4	0.051	*0.146**	*0.111**	*0.687**	*0.422**	*0.073**	*1.00**
	0.01	4	0.009	*0.037**	*0.02**	*0.373**	*0.164**	*0.059**	*1.00**

S4.3 Power study based on gene pair PARP1 and KRAS

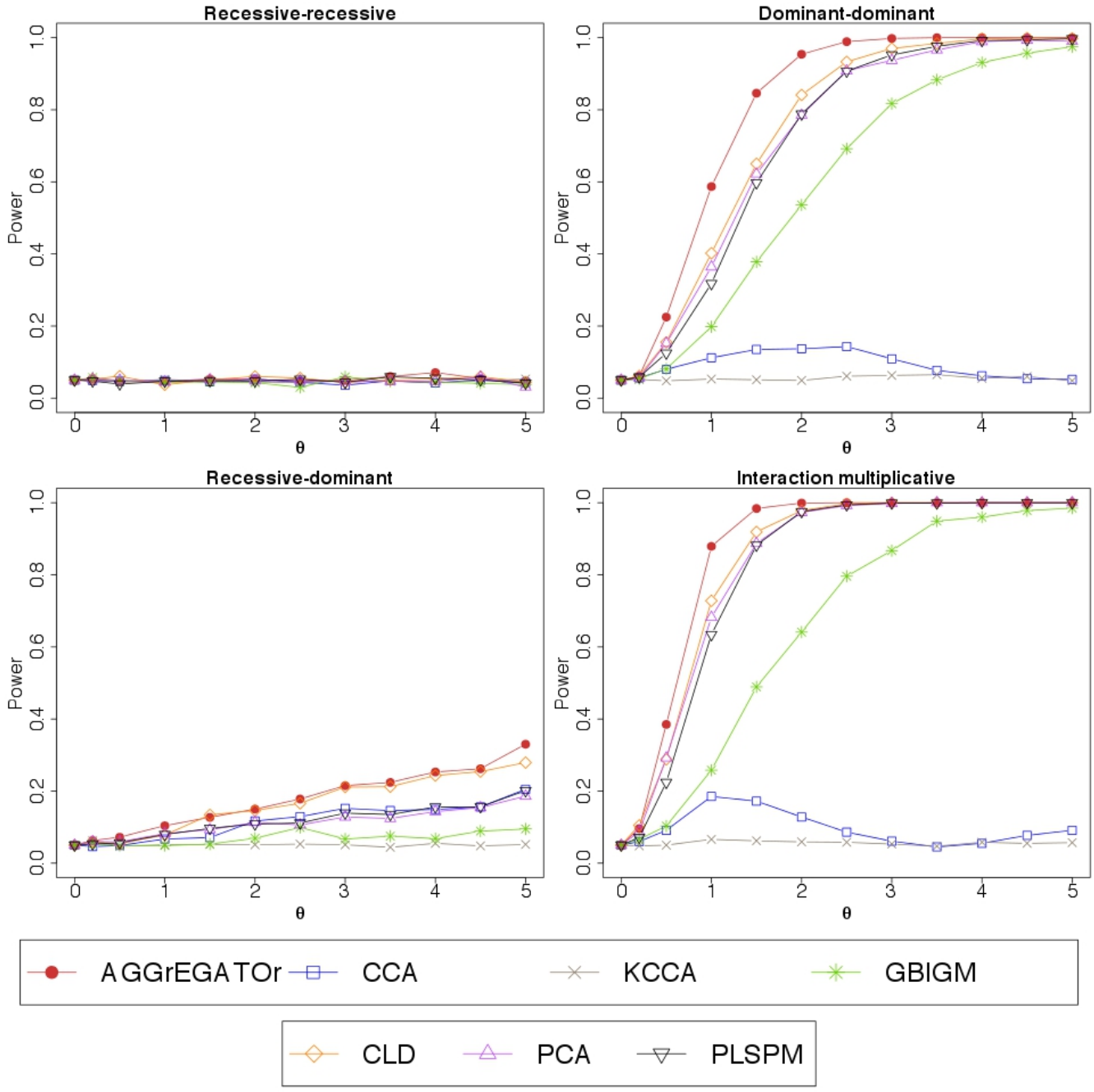

Figure S4: Power comparison between AGGrEGATOr and the six competitive methods (CCA, KKCA, GBIGM, CLD, PCA and PLSPM) under the four disease models (Recessive-recessive, Dominant-dominant, Recessive-dominant and Interaction multiplicative). The horizontal axis corresponds to the value of θ and quantifies the amount of interaction between the phenotype and the causal pair from genes PARP1 and KRAS.

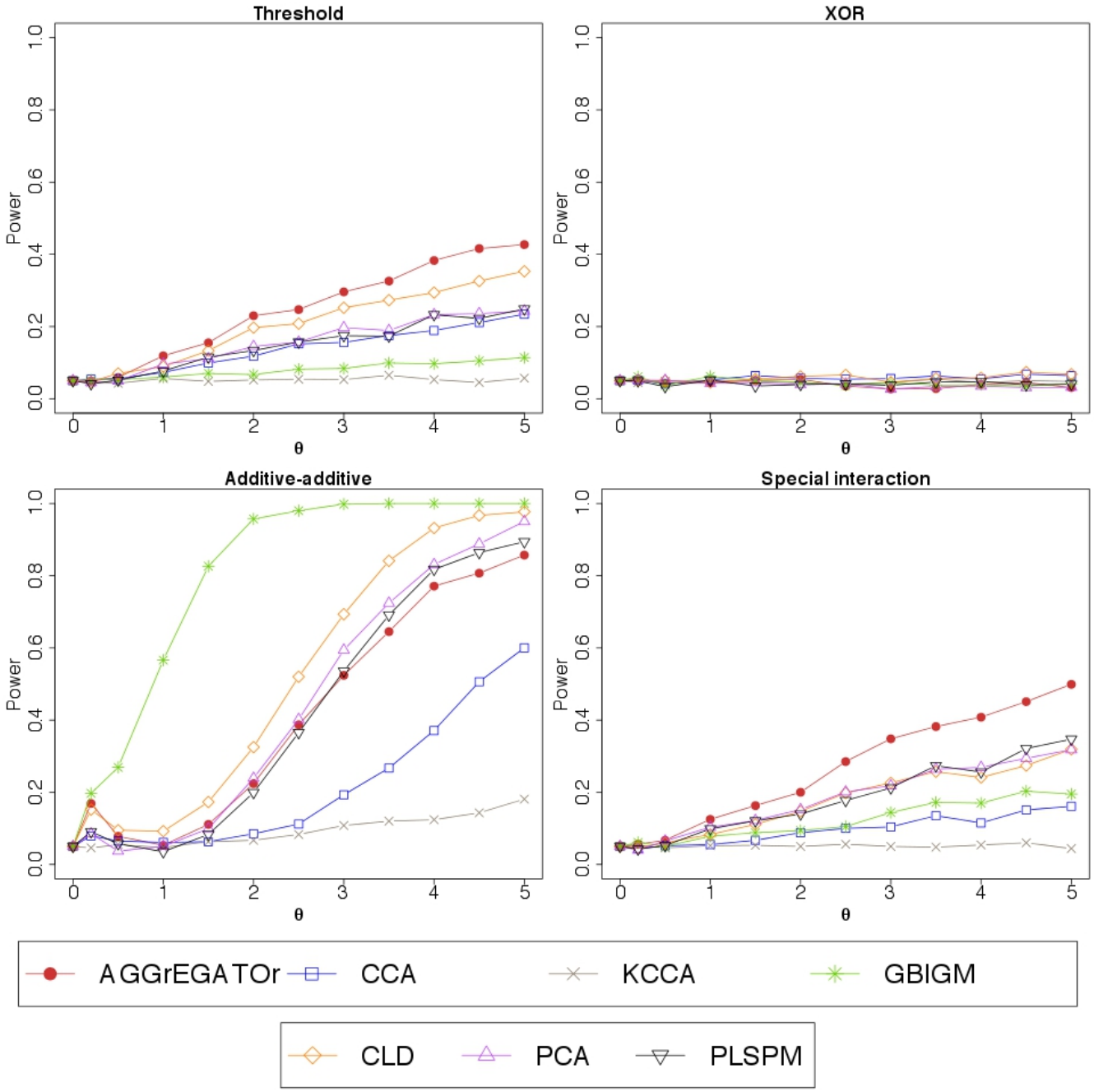

Figure S5: Power comparison between AGGrEGATOr and the six competitive methods (CCA, KKCA, GBIGM, CLD, PCA and PLSPM) under the four disease models (Threshold, XOR, Additive-additive and Special interaction). The horizontal axis corresponds to the value of θ and quantifies the amount of interaction between the phenotype and the causal pair from genes PARP1 and KRAS.

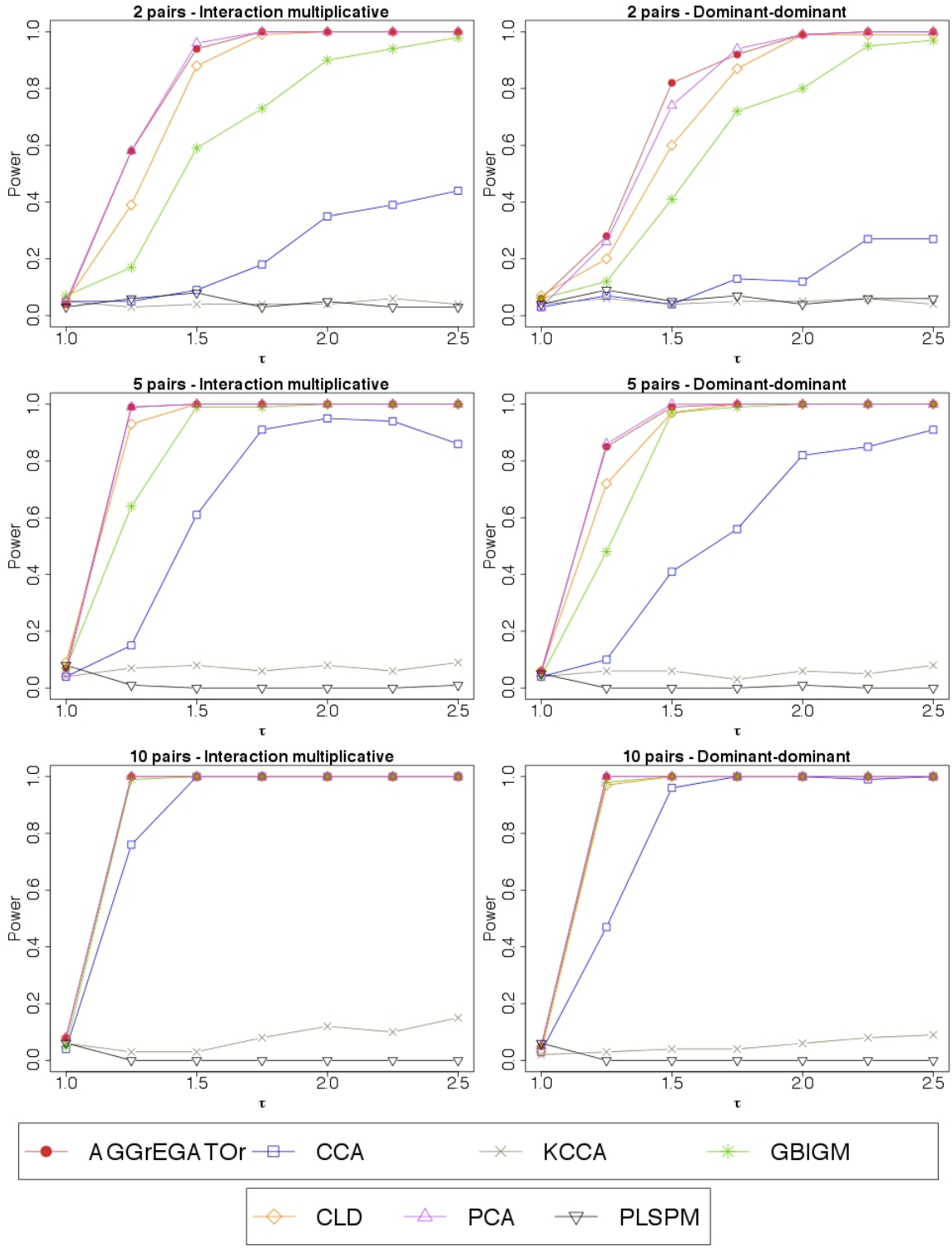

Figure S6: Power comparison between AGGrEGATOr and the six competitive methods (CCA, KKCA, GBIGM, CLD, PCA and PLSPM) under six disease models with multiple causal pairs (from 2 to 10). The horizontal axis corresponds to the value of τ and quantifies the amount of interaction between the phenotype and each causal pair from genes PARP1 and KRAS.

S5 Simulation study based on gene pair GNPDA2 and FAIM2

S5.1 LD pattern for gene pair GNPDA2 and FAIM2

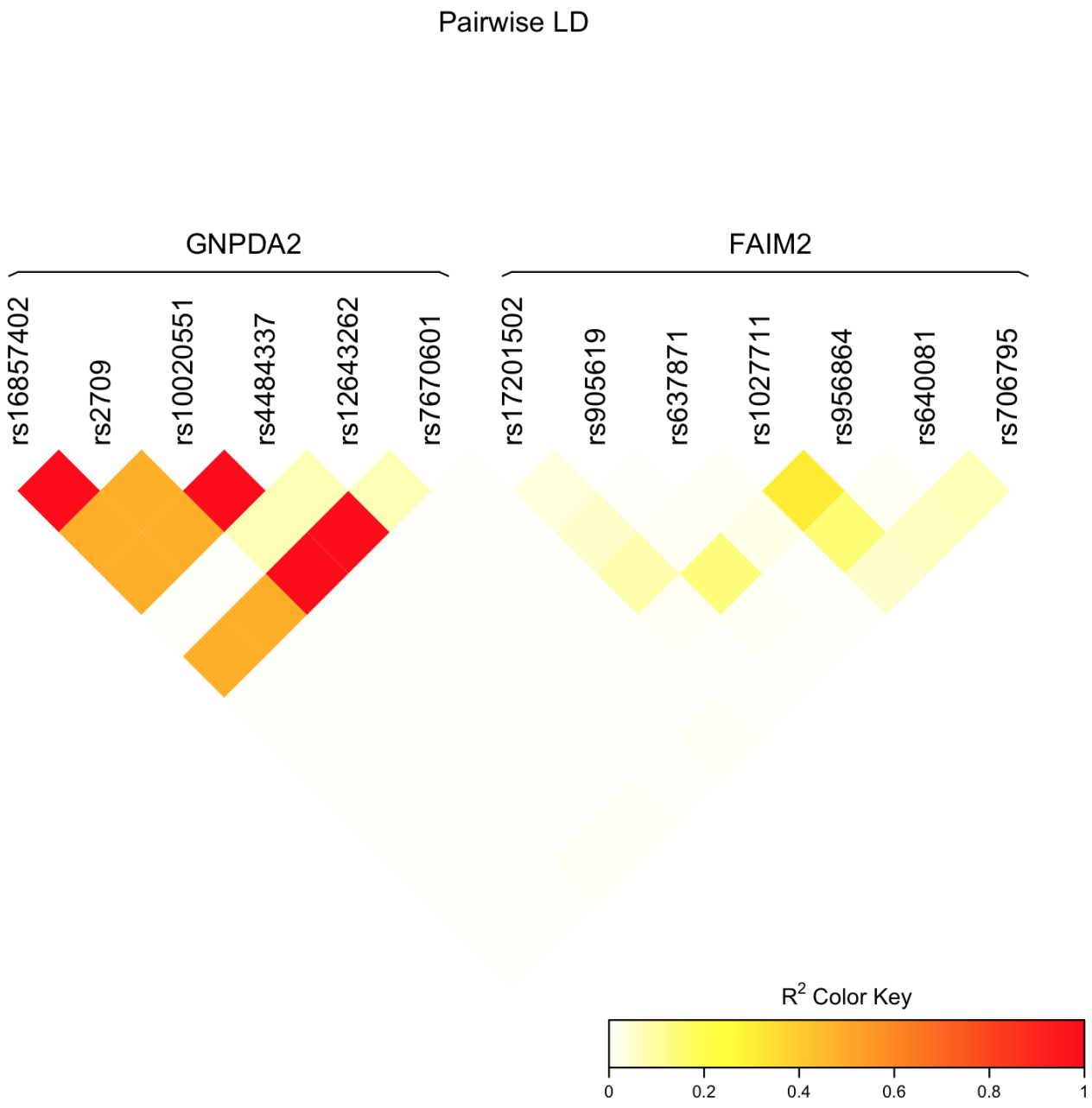

Figure S7: Pairwise r^2 within and between the 6 SNPs from GNPDA2 and the 7 SNPs from FAIM2.

S5.2 Evaluation of the type-I error rate based on gene pair GNPDA2 and FAIM2

Table S5: Estimation of the false positive rate in several scenarios involving three disease models based on the pair GNPDA2-FAIM2: a model with no effect, a model with one marginal (recessive) effect and a model with two (multiplicative) marginal effects. α is the expected predefined type-I error rate and θ is the parameter of the disease. Results with an $*$ indicate a significant deviation from the expected false positive rate.

Models	α	θ	minP	CCA	KCCA	CLD	PCA	PSLPM	GBIGM
No effect	0.05		0.055	0.057	0.042	0.057	0.59	0.051	0.049
	0.01		0.011	0.010	0.007	0.011	0.017	0.013	0.012
One marginal effect	0.05	1	*0.066*	0.046	0.054	0.058	0.060	0.047	*0.112*
	0.01	1	0.015	0.005	0.011	0.009	0.010	0.011	*0.029**
	0.05	4	0.046	0.041	0.048	0.064	0.053	0.059	*0.627*
	0.01	4	0.010	0.006	0.011	*0.018**	0.009	0.010	*0.344**
Multiplicative marginal effects	0.05	1	0.050	*0.118*	*0.087*	0.062	0.037	*0.524*	*0.394*
	0.01	1	0.013	*0.040**	*0.029**	0.010	0.004	*0.291**	*0.184**
	0.05	4	*0.107*	*0.330*	*0.372*	*0.216*	0.042	*0.591*	*1.00*
	0.01	4	*0.024**	*0.174**	*0.197**	*0.076**	0.006	*0.520**	*0.995**

S5.3 Power study based on gene pair GNPDA2 and FAIM2

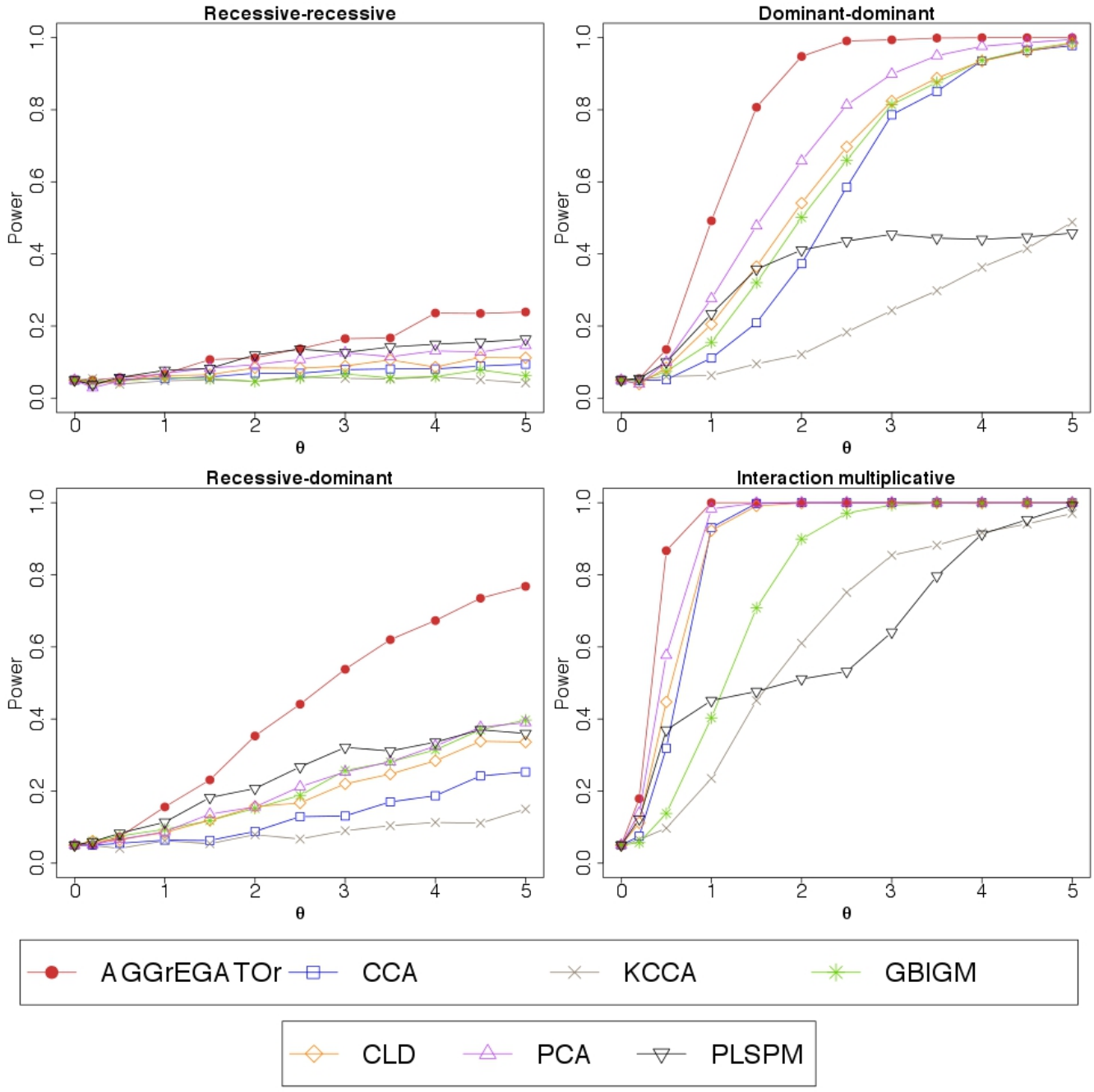

Figure S8: Power comparison between AGGrEGATOr and the six competitive methods (CCA, KKCA, GBIGM, CLD, PCA and PLSPM) under the four disease models (Recessive-recessive, Dominant-dominant, Recessive-dominant and Interaction multiplicative). The horizontal axis corresponds to the value of θ and quantifies the amount of interaction between the phenotype and the causal pair from genes GNPDA2 and FAIM2.

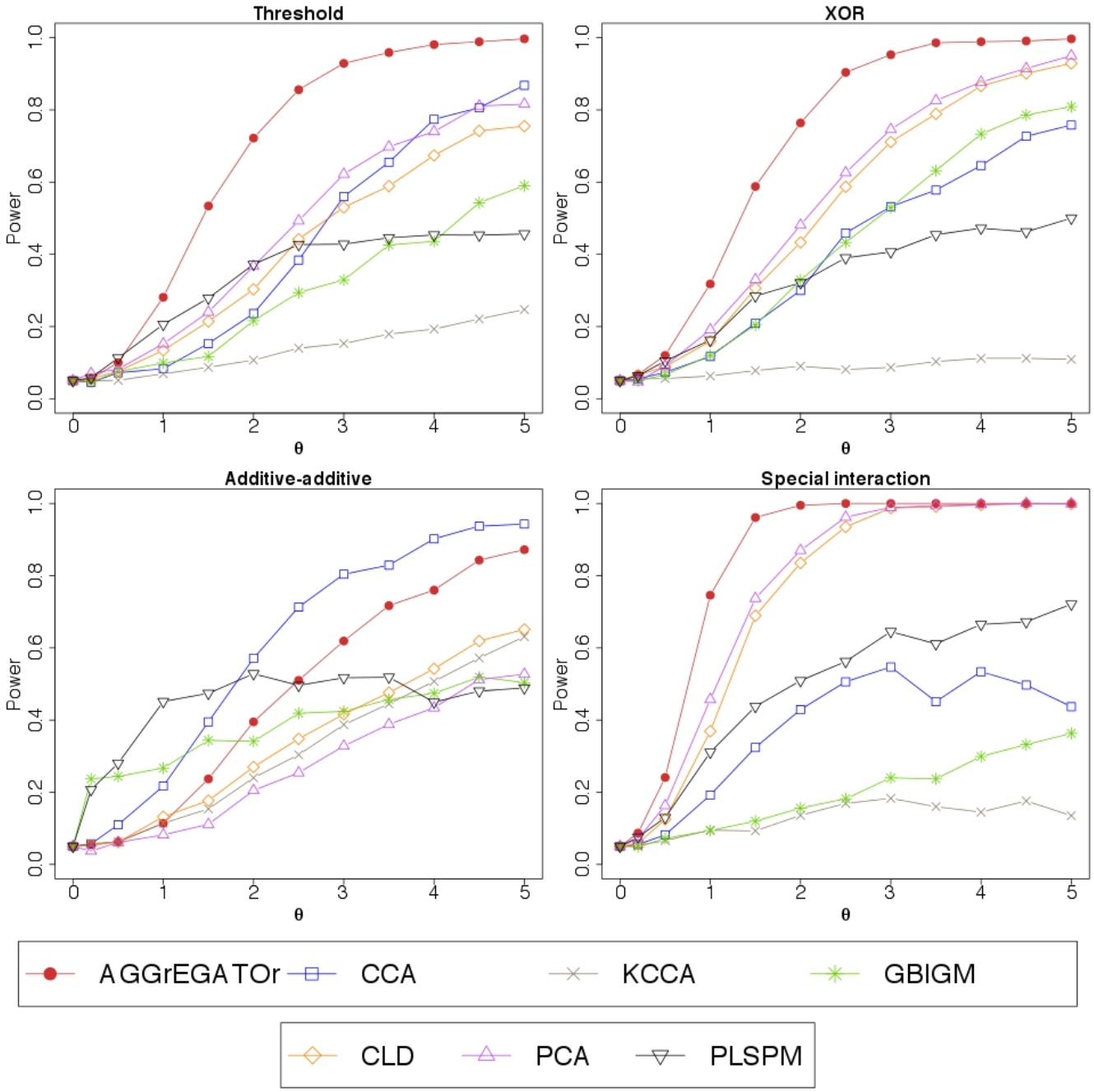

Figure S9: Power comparison between AGGrEGATOr and the six competitive methods (CCA, KKCA, GBIGM, CLD, PCA and PLSPM) under the four disease models (Threshold, XOR, Additive-additive and Special interaction). The horizontal axis corresponds to the value of θ and quantifies the amount of interaction between the phenotype and the causal pair from genes GNPDA2 and FAIM2.

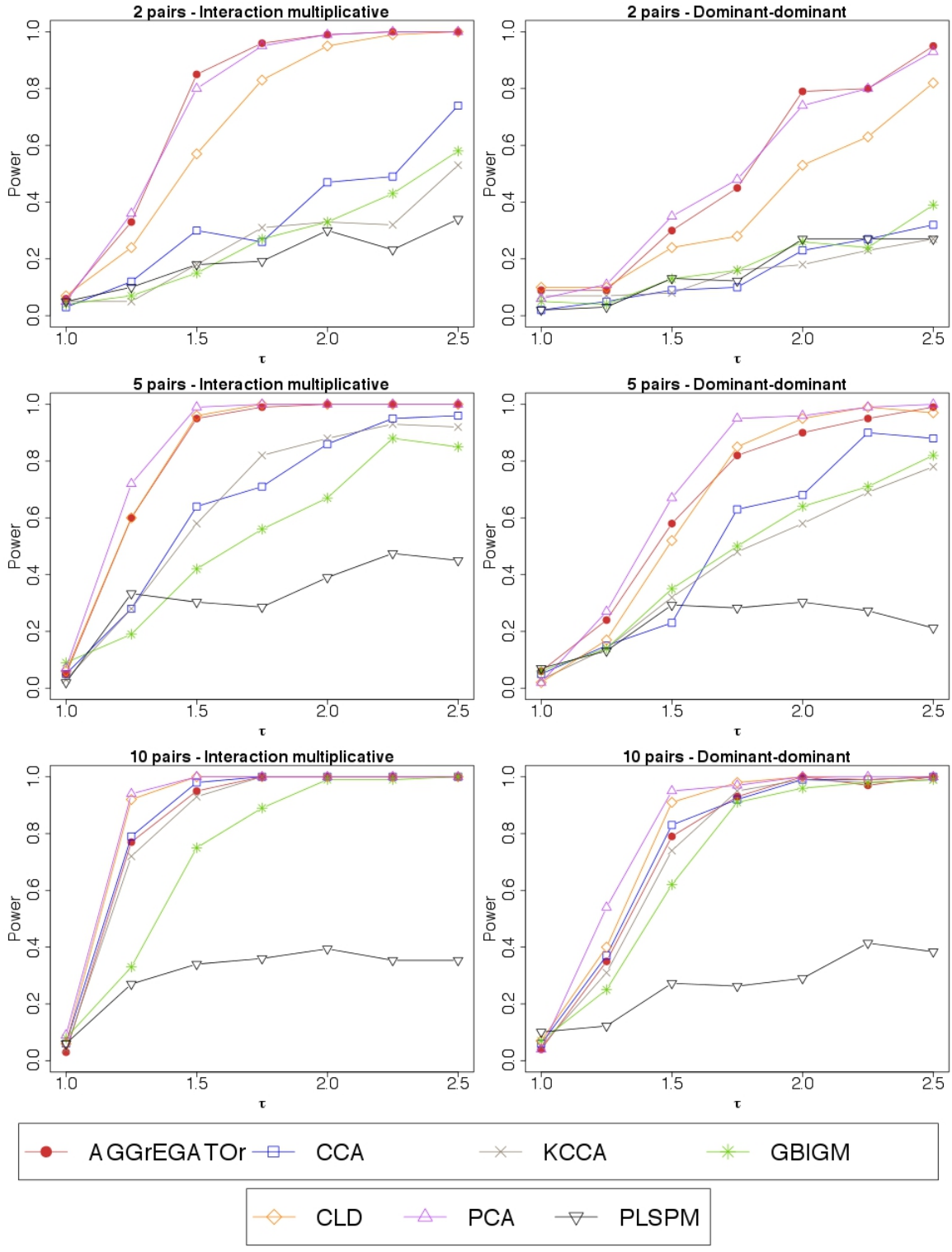

Figure S10: Power comparison between AGGrEGATOr and the six competitive methods (CCA, KKCA, GBIGM, CLD, PCA and PLSPM) under six disease models with multiple causal pairs (from 2 to 10). The horizontal axis corresponds to the value of τ and quantifies the amount of interaction between the phenotype and each causal pair from genes GNPDA2 and FAIM2.

S6 Rheumatoid Arthritis

Table S6 displays the list of the 17 genes used in the analysis of the GSE39428 [1] and WTCCC data sets [5].

Table S6: Description of genes in the Rheumatoid Arthritis data.

Gene	Chromosome
PADI1	1
PADI2	1
PADI4	1
PADI6	1
PRKD3	2
GC	5
GLRX	6
CDSN	6
PSORS1C1	6
TXNDC5	6
CA1	8
BUB3	10
SORBS1	10
VDR	12
SERPINA1	14
PCSK6	15
DNAH9	17

Table S7: Evaluation of the LD between gene pairs. For each gene pair, the maximum of correlation between one SNP from the first gene and one SNP from the other gene is reported. Values with * are higher than 0.1 and the corresponding gene pair has been excluded from the discovery cohort.

	BUB3	CA1	DNAH9	GLRX	PADI1	PADI2	PADI4	PADI6	PRKD3
BUB3	1.00	0.01	0.02	0.00	0.01	0.01	0.01	0.01	0.01
CA1	0.01	1.00	0.02	0.01	0.01	0.01	0.01	0.01	0.02
DNAH9	0.02	0.02	1.00	0.01	0.01	0.02	0.02	0.03	0.02
GLRX	0.00	0.01	0.01	1.00	0.01	0.00	0.01	0.01	0.04
PADI1	0.01	0.01	0.01	0.01	1.00	0.02	**0.12***	0.01	0.04
PADI2	0.01	0.01	0.02	0.00	0.02	1.00	0.01	0.01	0.01
PADI4	0.01	0.01	0.02	0.01	**0.12***	0.01	1.00	**0.79***	0.00
PADI6	0.01	0.01	0.03	0.01	0.01	0.01	**0.79***	1.00	0.01
PRKD3	0.01	0.02	0.02	0.04	0.04	0.01	0.00	0.01	1.00
PSORS1C1	0.02	0.04	0.05	0.01	0.04	0.01	0.01	0.02	0.01
SERPINA1	0.01	0.01	0.02	0.01	0.01	0.00	0.01	0.01	0.05
SORBS1	0.01	0.07	0.03	0.01	0.02	0.01	0.01	0.01	0.04
TXNDC5	0.02	**0.22***	0.03	0.03	0.05	0.03	0.02	0.04	0.03
VDR	0.02	0.02	0.03	0.01	0.01	0.01	0.02	0.01	0.02
GC	0.01	0.02	0.03	0.01	0.01	0.03	0.01	0.02	0.04
CDSN	0.02	0.02	0.02	0.01	0.01	0.01	0.00	0.01	0.02
PCSK6	0.02	0.05	0.04	0.04	0.08	0.03	0.03	0.04	0.03

	PSORS1C1	SERPINA1	SORBS1	TXNDC5	VDR	GC	CDSN	PCSK6
BUB3	0.02	0.01	0.01	0.02	0.02	0.01	0.02	0.02
CA1	0.04	0.01	0.07	**0.22***	0.02	0.02	0.02	0.05
DNAH9	0.05	0.02	0.03	0.03	0.03	0.03	0.02	0.04
GLRX	0.01	0.01	0.01	0.03	0.01	0.01	0.01	0.04
PADI1	0.04	0.01	0.02	0.05	0.01	0.01	0.01	0.08
PADI2	0.01	0.00	0.01	0.03	0.01	0.03	0.01	0.03
PADI4	0.01	0.01	0.01	0.02	0.02	0.01	0.00	0.03
PADI6	0.02	0.01	0.01	0.04	0.01	0.02	0.01	0.04
PRKD3	0.01	0.05	0.04	0.03	0.02	0.04	0.02	0.03
PSORS1C1	1.00	0.02	0.02	0.05	0.03	0.01	**0.38***	0.05
SERPINA1	0.02	1.00	0.02	0.02	0.01	0.02	0.02	0.02
SORBS1	0.02	0.02	1.00	0.07	0.05	0.02	0.02	0.03
TXNDC5	0.05	0.02	0.07	1.00	0.04	0.04	0.02	**0.20***
VDR	0.03	0.01	0.05	0.04	1.00	0.01	0.01	0.03
GC	0.01	0.02	0.02	0.04	0.01	1.00	0.01	0.03
CDSN	**0.38***	0.02	0.02	0.02	0.01	0.01	1.00	0.04
PCSK6	0.05	0.02	0.03	**0.20***	0.03	0.03	0.04	1.00

References

[1] Chang, X., B. Xu, L. Wang, Y. Wang, Y. Wang, and S. Yan (2013): "Investigating a pathogenic role for txndc5 in tumors," *International Journal of Oncology*, 43, 1871–1884.

[2] Li, J., D. Huang, M. Guo, X. Liu, C. Wang, Z. Teng, R. Zhang, Y. Jiang, H. Lv, and L. Wang (2015): "A gene-based information gain method for detecting gene-gene interactions in case-control studies," *European Journal of Human Genetics*, 13, 1566–1572.

[3] Ma, L., A. G. Clark, and A. Keinan (2013): "Gene-based testing of interactions in association studies of quantitative traits," *PLoS Genet*, 9, e1003321.

[4] Marchini, J., P. Donnelly, and L. R. Cardon (2005): "Genome-wide strategies for detecting multiple loci that influence complex diseases," *Nature Genetics*, 37, 413–417.

[5] WTCCC (2007): "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, 447, 661–678.

[6] Yuan, Z., Q. Gao, Y. He, X. Zhang, F. Li, J. Zhao, and F. Xue (2012): "Detection for gene-gene co-association via kernel canonical correlation analysis," *BMC Genetics*, 13, 83.