Supplementary Material for "Ranking causal SNPs and associated regions in genome-wide association studies using the support vector machine and random forest

Usman Roshan[*†] Satish Chikkagoudar [†] Zhi Wei [†] Kai Wang [‡]

Hakon Hakonarson [§]

1 Background

1.1 Composite odds ratio for estimating disease risk

A standard assumption in disease models is that the the probability of disease given i copies of a risk allele is given by $P(D|g_i) = \frac{1}{1+e^{-\alpha+i\beta}}$ for $i = 0, 1, 2$ [1, 2]. This follows naturally by assuming that the log likelihood ratio is linear and is also known as the logistic regression model [3]. Under this assumption we can estimate α and β by maximum likelihood using a simple gradient descent procedure [3]. This usually converges within a few iterations.

After estimating α and β we use e^{β} as the odds ratio for the given SNP. This follows naturally by noting that $P(D|g_i) = \frac{1}{1+e^{-\alpha+i\beta}}$ can be rewritten as $\ln(\frac{P(D|g_i)}{1-P(D|g_i)}) = \alpha + i\beta$. The odds ratio $\frac{\frac{Pr(D|g_1)}{1-Pr(D|g_1)}}{\frac{Pr(D|g_0)}{1-Pr(D|g_0)}}$ is then given by $\lambda = e^{\beta}$. For two copies of the risk allele we obtain $e^{2\beta} = (e^{\beta})^2 = \lambda^2$. The odds ratio calculated in this manner (under the logistic regression model) does not suffer from bias and stratification problems under simpler models [1].

We assume that each SNP is acting independently. Then the composite odds ratio for several SNPs is defined as $\Pi_i \lambda_i$ where λ_i is the odds ratio of SNP i.

[*]To whom correspondence should be addressed; usman@cs.njit.edu, Ph: 973-596-2872, Fax: 973-596-5777

[†]Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102

[‡]Psychiatry & the Behavioral Sciences, Zilkha Neurogenetic Institute, University of Southern California

[§]Center for Applied Genomics, The Childrens Hospital of Philadelphia, Philadelphia, PA 19104

2 Comparison of 1-df chi-square to MAX trend test

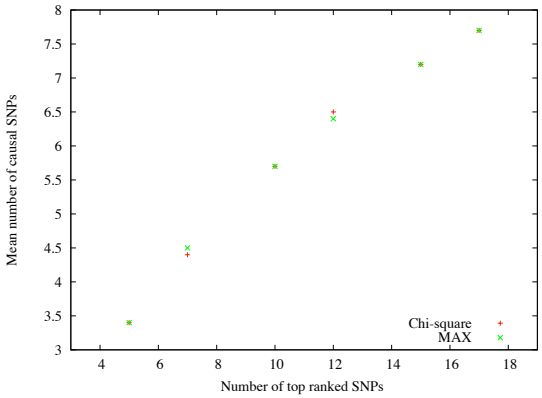
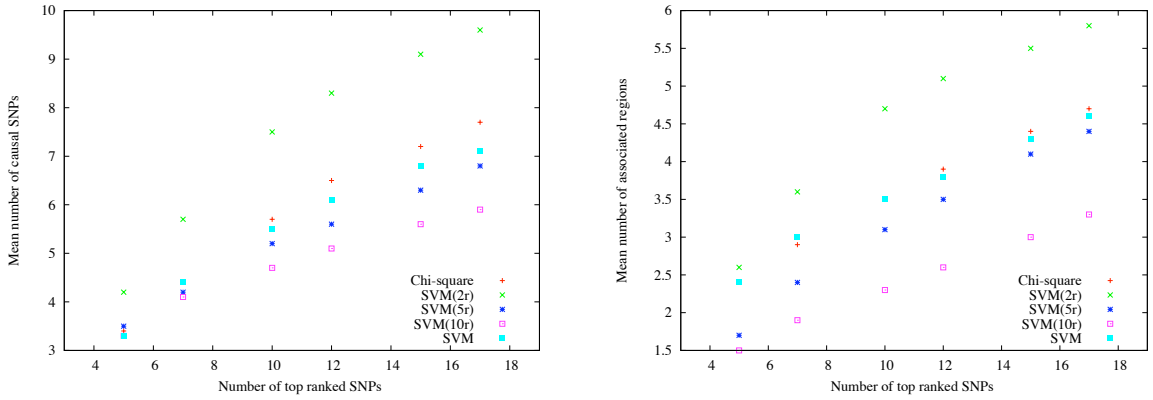


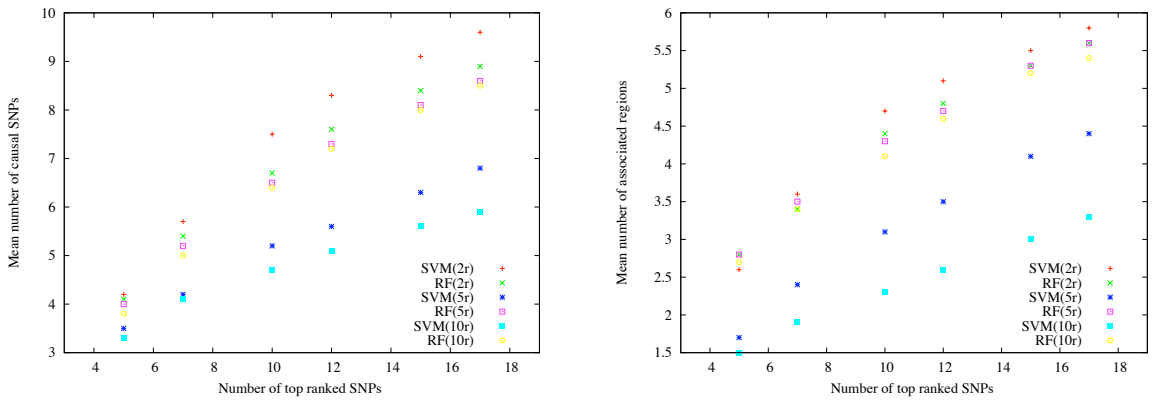
Figure 1: Number of causal SNPs in top k chi-square and MAX trend ranked SNPs on simulated data of relative risk 1.5.

3 Comparison of support vector machine, random forest, and 1-df chi-square on simulated data



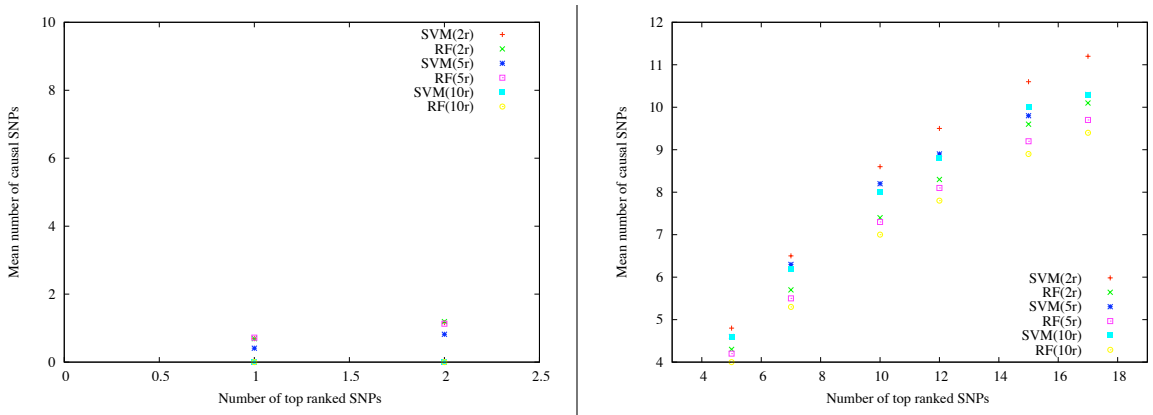
(a) Causal variants (b) Associated regions

Figure 2: Number of causal variants and associated regions in simulated data of relative risk 1.5 identified by the support vector machine applied to the top $2r$, $5r$, and $10r$ chi-square ranked SNPs as input, where r is the number of SNPs with p-values within Bonferroni threshold, and the entire GWAS denoted by just SVM.



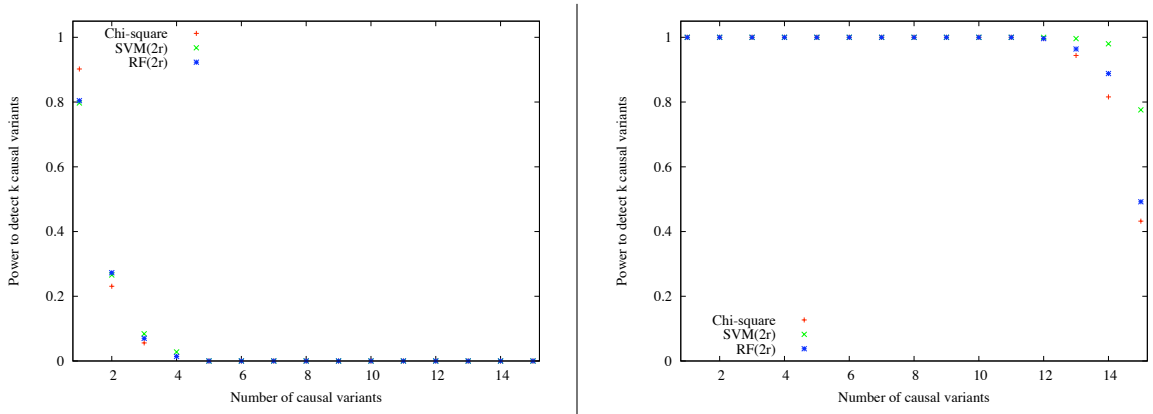
(a) Causal variants (b) Associated regions

Figure 3: Number of causal variants and associated regions in simulated data of relative risk 1.5 identified by SVM and RF at the three different thresholds.



(a) Relative risk 1.25 (b) Relative risk 2

Figure 4: Number of causal variants in simulated data of relative risk 1.25 and 2.



(a) Relative risk 1.25 (b) Relative risk 2

Figure 5: Empirical power of chi-square, SVM($2r$), and RF($2r$) to detect k causal variants in simulated data within the top r ranked SNPs where r is the number of SNPs with p-values within Bonferroni threshold.

4 Running time

Table 1: Running time in seconds on one simulated dataset with causal variants

RR	Sample size	r	χ^2	SVM($2r$)	RF($2r$)	SVM($10r$)	RF($10r$)
1.25	2000	1	11.4	13.7	55.4	11.9	202.2
1.5	2000	33	11.4	13.4	824.1	15.8	1264.2
2	2000	84	11.4	15.3	1203. 7	31.5	2211.7
1.25	4000	13	22.6	24.7	1103.2	25.0	2479.0
1.25	8000	44	64.9	77.1	3781.9	77.4	9236.1
1.25 ($f < 5\%$)	4000	1	23.9	25.8	70.3	25.8	433.4
1.25 ($f < 5\%$)	8000	8	67.5	77.2	1057.7	78.8	3595.2
1.5 ($f < 5\%$)	4000	15	24.1	27.4	708	32.8	2772.5
1.5 ($f < 5\%$)	8000	40	66.5	81.8	3787.6	113	9616.2

Table 2: Running time of SVM and RF applied to all SNPs in seconds on one simulated dataset with causal variants

RR	Sample size	χ^2	SVM	RF
1.25	2000	11.4	784.3	9735.7
1.5	2000	11.4	793.1	9732.3
2	2000	11.4	800.5	9473.7

Table 3: Running time in seconds on WTCCC studies

	χ^2	SVM($2r$)	RF($2r$)	SVM($10r$)	RF($10r$)
T1 diabetes	700	744	8056	920	14891
Arthritis	569	589	5449	657	11018
Crohn's dis.	586	593	3427	617	7144
T2 diabetes	550	554	2068	558	3611

5 Stability at relative risk 1.25 and 2

Table 4: Correlation coefficient between original SNP ranks and mean SNP ranks across 100 jacknifed datasets of relative risk 1.25

	r	$2r$	$5r$	$10r$
χ^2	1	0.95	0.93	0.95
SVM	1	0.66	0.77	0.96
RF(100)	1	0.94	0.94	0.49
RF(10000)	1	0.9	0.97	0.78

Table 5: Correlation coefficient between original SNP ranks and mean SNP ranks across 100 jacknifed datasets of relative risk 2

	r	$2r$	$5r$	$10r$
χ^2	0.99	0.99	0.94	0.92
SVM	0.99	0.97	0.95	0.92
RF(100)	0.83	0.76	0.56	0.38
RF(10000)	0.99	0.98	0.87	0.74

6 Ranking SNPs by SVM first and then followed by chi-square

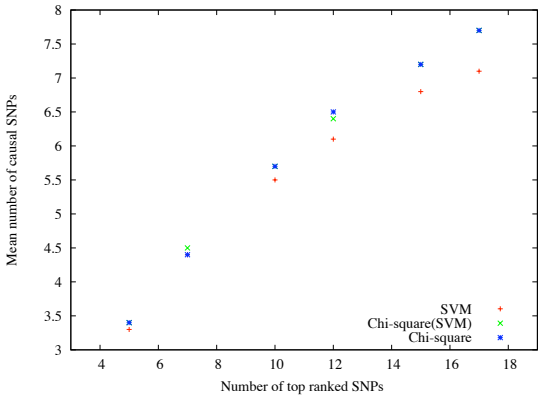


Figure 6: SVM and chi-square denote the rankings given by SVM and chi-square applied to the entire GWAS. Chi-square(SVM) denotes the ranking computed by applying chi-square to the top 100 ranked SNPs given by the SVM applied to the entire GWAS. This is all on simulated data with relative risk 1.5.

7 Real data

We removed all SNPs and samples from cases and controls that are specified as problematic by the WTCCC. This left 1480 individuals from the 1958 British Birth Cohort, 1458 from the UK Blood Service Control Group, 1963 cases for type 1 diabetes, 1860 for arthritis, 1748 for Crohn's disease, and 1924 for type 2 diabetes. We then combined the two control sets with each case set and removed all SNPs with greater than 1% missing entries. Using the plink software package [4, 5] we also removed all SNPs that deviate from the Hardy-Weinberg equilibrium with p-values below 5×10^{-7}. This left us with a total of 422,006 SNPs for type 1 diabetes, 403301 for arthritis, 405306 for Crohn's disease, and 402532 for type 2 diabetes. We confirmed that the same significant SNPs with approximately the same p-values were reported by our programs as published in the original WTCCC study (Table 3 of [6]).

The Genetics of Kidneys in Diabetes (GoKinD) type 1 diabetes dataset is composed of case subjects from New England [7, 8]. The Center for Applied Genomics at CHOP cleaned this dataset first which resulted in 1529 case subjects. Since there are no control subjects we used the WTCCC Coronary Artery Disease case subjects as controls. We computed the common SNPs between the two datasets and removed those with greater than 1% missing data and those that deviated from Hardy Weinberg equilibrium with p-values below 5×10^{-7}. This left us with 303,139 SNPs in total.

We summarize all five studies in Table 6 below.

Table 6: Description of real genome-wide association studies

Data	Cases	Controls	Number of SNPs
WTCCC T1 diabetes	1963	2938	422006
WTCCC arthritis	1860	2938	403301
WTCCC Crohn's disease	1748	2938	405306
WTCCC T2 diabetes	1924	2938	402532
GoKinD T1 diabetes	1529	1924	309139

7.1 Effect of p-value threshold on the support vector machine on real data

Table 7: Ranking of previously replicated SNPs in the WTCCC type 1 diabetes, arthritis, Crohn's disease, and type 2 diabetes studies as given by the support vector machine. SVM($2r$) denotes the support vector machine ranking computed on the top $2r$ chi-square ranked SNPs, SVM($5r$) and SVM($10r$) denote the same with top $5r$ and $10r$, and r denotes the number of SNPs with p-values within Bonferroni threshold.

Type 1 diabetes						Rheumatoid arthritis					
SNP	χ^2 p-val	χ^2	SVM($2r$)	SVM($5r$)	SVM($10r$)	SNP	χ^2 p-val	χ^2	SVM($2r$)	SVM($5r$)	SVM($10r$)
rs9272346	5e-139	0	2	2	2	rs6457617	1e-77	1	43	48	13
rs6679677	3e-26	95	10	13	25	rs6920220	4e-5	242	57	169	521
rs17696736	3e-14	179	75	63	108	rs3890745	5e-5	250	39	112	153
rs705702	4e-6	517	172	305	150	rs1678542	2e-4	293	54	198	49
rs12708716	6e-7	437	210	472	411						
rs17388568	2e-6	487	39	102	104						
rs2542151	1e-5	561	38	137	158						
rs12251307	3e-4	748	74	267	476						
rs3087243	2e-4	721	271	611	1999						
Crohn's disease											
SNP	χ^2 p-val	χ^2	SVM($2r$)	SVM($5r$)	SVM($10r$)	SNP	χ^2 p-val	χ^2	SVM($2r$)	SVM($5r$)	SVM($10r$)
rs11209026	4e-15	0	1	4	44	rs4132670	2e-11	2	11	48	104
rs3828309	5e-12	7	51	129	141	rs8050136	2e-7	10	2	26	69
rs17234657	1e-12	3	0	36	15	rs796158	3e-5	31	NA	25	31
rs9292777	9e-12	9	31	71	134						
rs17221417	3e-11	15	20	75	125						
rs9858542	3e-6	76	26	205	332						
rs2542151	2e-7	46	12	30	8						
rs11747270	2e-7	44	46	151	335						
rs6596075	3e-6	74	66	163	300						
rs6908425	4e-5	111	8	5	31						
rs12035082	2e-5	100	45	125	237						
rs2836754	5e-5	121	7	72	215						

7.2 Comparison of the support vector machine and random forest on real data

Table 8: Ranking of previously replicated SNPs in the WTCCC type 1 diabetes, arthritis, Crohn's disease, and type 2 diabetes as given by chi-square, support vector machine, and random forest. See caption of Table 7 for more details.

Type 1 diabetes				
SNP	χ^2 p-val	χ^2	SVM($2r$)	RF($2r$)
rs9272346	5e-139	0	2	2
rs6679677	3e-26	95	10	121
rs17696736	3e-14	179	75	215
rs705702	4e-6	517	172	238
rs12708716	6e-7	437	210	316
rs17388568	2e-6	487	39	428
rs2542151	1e-5	561	38	390
rs12251307	3e-4	748	74	688
rs3087243	2e-4	721	271	385
Rheumatoid arthritis				
SNP	χ^2 p-val	χ^2	SVM($2r$)	RF($2r$)
rs6457617	1e-77	1	43	1
rs6920220	4e-5	242	57	154
rs3890745	5e-5	250	39	286
rs1678542	2e-4	293	54	167
Crohn's disease				
SNP	χ^2 p-val	χ^2	SVM($2r$)	RF($2r$)
rs11209026	4e-15	0	1	15
rs3828309	5e-12	7	51	14
rs17234657	1e-12	3	0	1
rs9292777	9e-12	9	31	11
rs17221417	3e-11	15	20	29
rs9858542	3e-6	76	26	30
rs2542151	2e-7	46	12	31
rs11747270	2e-7	44	46	71
rs6596075	3e-6	74	66	106
rs6908425	4e-5	111	8	105
rs12035082	2e-5	100	45	65
rs2836754	5e-5	121	7	38
Type 2 diabetes				
SNP	χ^2 p-val	χ^2	SVM($2r$)	RF($2r$)
rs4132670	2e-11	2	11	0
rs8050136	2e-7	10	2	1

8 Type 1 diabetes associated regions

Below we list the replicated SNPs from [9] and the set of SNPs with squared correlation coefficient at least 0.05 with each one. We also list the regions given by the Type 1 Diabetes Consortium [10, 11].

8.1 Replicated SNPs [9] and SNPs in their associated region

- rs3087243: rs3087243 rs231726 rs1427676 rs11571300 rs6748358 rs859839 rs10179639 rs6752680 rs7596727 rs4335928 rs4675374 rs10932036 rs10932038 rs4675386 rs16840488 rs6754914 rs4675397 rs13383369 rs4675399 rs11884257 rs16840552 rs16824662 rs6759546 rs16840628 rs1559933
- rs2165738: rs2165738 rs13035764 rs11125629 rs1545255 rs10432678 rs17046399 rs17791703 rs17046446
- rs13023380: rs13023380 rs12479125 rs2033300 rs3788964 rs7608315 rs17716942 rs16846743
- rs17388568: rs17388568 rs2086346 rs17005728 rs1479923 rs2069774 rs11724582 rs6848139 rs7673567 rs4833830 rs4572894 rs17005850 rs4833836 rs7682481 rs1512971
- rs11755527: rs11755527 rs1847472 rs604912 rs12203124 rs12200345 rs12204886 rs285644 rs285642 rs6940559 rs16882661 rs927297 rs6454806 rs1321858 rs10755490 rs285612 rs16882779 rs2295712 rs790595 rs12530177 rs790604 rs790605 rs790606 rs790607 rs9353726 rs12661853 rs1551115 rs998095 rs9451420 rs12208773
- rs9272346: rs9272346 rs9275134 rs2856688 rs7775228 rs6457617 rs6457620 rs2647015 rs2647046 rs2858308 rs9275418 rs9275523 rs9275572 rs6936863 rs3916765 rs9275765 rs9275772 rs9461799 rs9275793 rs2859090 rs2227127 rs9276429 rs9276431 rs9276432 rs4398729 rs9276435 rs9276440 rs9276448 rs9276490 rs5014418 rs7768538 rs7453920 rs6902723 rs6903130 rs6919798 rs9296044 rs2857212 rs2857210 rs2621416 rs17429127 rs9276712 rs2621384 rs2857161 rs2621383 rs2621382 rs2857154 rs2857129 rs2621330 rs2070121 rs16870880 rs17501267 rs2071474 rs1894407 rs10484565 rs241432 rs241429 rs241428
- rs947474: rs947474 rs744254 rs744253 rs4750316 rs1570527 rs11258455 rs12257237 rs11258500 rs12572911 rs11592152 rs10906529 rs806399 rs648731 rs582052 rs4750439 rs2453 rs2236380 rs681071 rs11596750 rs1889001 rs11258747 rs10796145 rs961203 rs593608 rs630196 rs10508307 rs586457 rs4750491 rs636819 rs7918923 rs4750495 rs10508308 rs12146312 rs591441 rs656715 rs7083579 rs11259097 rs3793730 rs688879 rs501878 rs661891 rs501930 rs6602745 rs4748099 rs11259211 rs650652 rs510745 rs596866 rs500766 rs11259403 rs12249263 rs11259425 rs10752351
- rs12251307: rs12251307 rs7899110 rs9988772 rs2274359 rs1073646 rs10905876 rs4747888 rs11598494 rs3750671 rs11257010 rs652318 rs11257057 rs642104 rs613960 rs3814195 rs669534 rs658386 rs616246 rs678395 rs11257102 rs11257103 rs7911111 rs11257349 rs11257374 rs7916509 rs1341926 rs4750190 rs4750200 rs7894025 rs7910400 rs11257712 rs4747976 rs11257726 rs17152914 rs1544198 rs11257741 rs17152987 rs7896455 rs7896594 rs705702: rs705702 rs772923 rs11171739 rs2292239 rs1689512 rs773643
- rs6603781: rs6679677 rs1217396 rs6537798 rs1217414 rs1235005 rs3811018 rs3006998 rs11810241 rs12402202 SNP_A-1962749 rs1553452 rs2938327
- rs17696736: rs17696736 rs1980364 rs7299227 rs10850003 rs3519 rs1005902 rs11066205
- rs3825932: rs3825932 rs12148343 rs12101606 rs3743200 rs3743201 rs7168964 rs4778773 rs11631671 rs8025124 rs1822471 rs7166598 rs12911414 rs12440502 rs6495365 rs1879648 rs7166639 rs2102999 rs12593555 rs939653 rs939655 rs1402759 rs7165489
- rs416603: rs416603 rs431918 rs393161 rs1559394 rs4451969 rs181694 rs243315 rs1874021 rs2867936 rs12927773 rs12445900 rs2867945 rs9935384 rs12922090 rs7187741 rs7199847 rs7193871 rs7205925 rs7203055 rs4325584 rs8058507 rs11864785 rs10163438 rs4781079 rs8052665 rs4129934 rs9928810 rs11074989 rs12051011 rs8044309 rs4500747 rs6498196 rs8059633

- rs12708716: rs12708716 rs12924729 rs3893660 rs9941107 rs7198004 rs7203150 rs9746695 rs11647011 rs2867879 rs7184083 rs6498169 rs28087 rs767019 rs27908 rs42369 rs166054 rs248836 rs171591 rs248848 rs40448 rs168716 rs2113261 rs1646066 rs1646067 rs11074956 rs151772 rs149310 rs149311 rs3760114 rs243327 rs243325 rs416603 rs431918 rs393161 rs1559394
- rs2542151: rs2542151 rs2542152 rs2542153 rs2542160 rs2847297 rs2542170 rs8085163 rs17597893 rs2847289 rs16939895 rs8087237 rs669822 rs7234029 rs2222138 rs908579 rs908578 rs6505771 rs8088313
- rs9976767: rs9976767 rs2839509 rs11910025 rs17114898 rs17114906 rs883391 rs883390 rs2839519 rs2839530 rs2839531 rs12482483 rs8134499 rs228049 rs228050 rs17115062 rs228077 rs7276861 rs1399922 rs17115129 rs17178310 rs435725 rs17767630 rs17178366 rs408465 rs11701162
- rs229541: rs229541 rs69264 rs4821602 rs13053175 rs5750402 rs5756581 rs4820279 rs4821623 rs2267363 rs727047 rs5756601 rs1981476 rs17824310 rs5995416 rs4821628 rs6000704

8.2 Associated regions given by T1DBase

Each region below contains the start and stop SNP numbers in our WTCCC case control study, the chromosome number and start and stop positions on the chromosome given by genome build NCBI36, the SNP rs ID, and the region name. This data is also available online at the T1DBase website at `http://t1dbase.org/page/RegionsDownload/display/format/table/build/NCBI36`.

- 8866 8889 # Chr1:63868682-63940620 rs2269241 1p31.3
- 16822 16908 # Chr1:113620000-114460000 rs2476601 1p13.2
- 25079 25094 # Chr1:190728079-190816535 rs2816316 1q31.2
- 27174 27212 # Chr1:204869063-205116454 rs3024505 1q32.1
- 36895 36904 # Chr2:12528551-12596884 rs1534422 2p25.1
- 49964 50064 # Chr2:102221730-102575042 rs917997 2q12.1
- 58554 58613 # Chr2:162669119-163101007 rs1990760 2q24.2
- 64282 64314 # Chr2:204381054-204528303 rs3087243 2q33.2
- 78525 78656 # Chr3:45955677-46629136 rs333,rs11711054 3p21.31
- 104201 104227 # Chr4:25637903-25745871 rs10517086 4p15.2
- 117697 117747 # Chr4:123128865-123833732 rs2069763,rs4505848 4q27
- 133662 133726 # Chr5:35835032-36072476 rs6897932 5p13.2
- 160158 161868 # Chr6:24700000-34000000 rs9268645 6p21.32
- 169866 169896 # Chr6:90863556-91103018 rs11755527 6q15
- 174985 175076 # Chr6:126479722-127461527 rs9388489 6q22.32
- 176754 176829 # Chr6:137915383-138379949 rs6920220,rs10499194 6q23.3
- 180130 180155 # Chr6:159237500-159446677 rs1738074 6q25.3
- 186873 186944 # Chr7:26624487-27171807 rs7804356 7p15.2
- 191238 191345 # Chr7:50866662-51640000 rs4948088 7p12.1

- 229625 229659 # Chr9:4218550-4311558 rs7020673 9p24.2
- 249494 249537 # Chr10:6069732-6237542 rs12722495,rs11594656,rs12251307 10p15.1
- 249600 249651 # Chr10:6475380-6585110 rs947474,rs11258747 10p15.1
- 264058 264095 # Chr10:89998027-90268360 rs10509540 10q23.31
- 272638 272657 # Chr11:2025000-2264880 rs689,rs7111341 11p15
- 296656 296713 # Chr12:9512800-9867423 rs4763879 12p13.31
- 303876 303915 # Chr12:54637613-55091576 rs2292239 12q13.2
- 303941 304104 # Chr12:55268557-56819824 rs1678542 12q13.3
- 312682 312867 # Chr12:109772109-111723111 rs3184504 12q24.12
- 340575 340604 # Chr14:68237491-68387815 rs1465788 14q24.1
- 345527 345559 # Chr14:97427667-97601359 rs4900384 14q32.2
- 354985 355009 # Chr15:76773860-77050416 rs3825932 15q25.1
- 361275 361410 # Chr16:10923546-11560000 rs12708716 16p13.13
- 362616 362635 # Chr16:20174576-20276065 rs12444268 16p12.3
- 363783 363801 # Chr16:28191236-28944416 rs4788084 16p11.2
- 368229 368262 # Chr16:73760231-74086012 rs7202877 16q23.1
- 373037 373045 # Chr17:7560000-7660000 rs16956936 17p13.1
- 376306 376362 # Chr17:34634168-35508018 rs2290400,rs12150079 17q12
- 376386 376408 # Chr17:35990900-36132000 rs7221109 17q21.2
- 384100 384126 # Chr18:12726556-12916278 rs45450798,rs478582,rs1893217 18p11.21
- 392665 392680 # Chr18:65630495-65722590 rs763361 18q22.2
- 398736 398748 # Chr19:51843218-52015224 rs425105 19q13.32
- 400294 400356 # Chr20:1444473-1707590 rs2281808 20p13
- 416179 416196 # Chr21:42681878-42761422 rs3788013,rs11203203 21q22.3
- 418667 418771 # Chr22:28137855-28999883 rs5753037 22q12.2
- 420055 420071 # Chr22:35898616-35996732 rs229541 22q13.1

9 Prediction of type 1 diabetes risk on independent studies

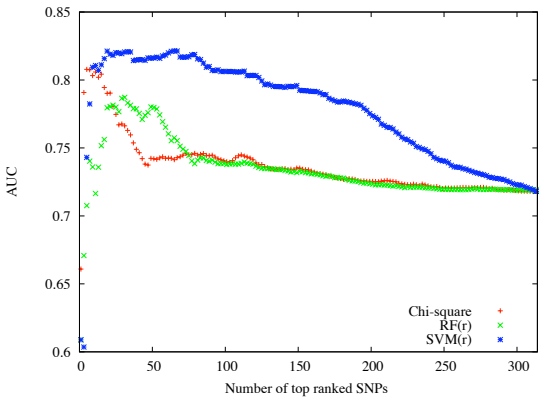
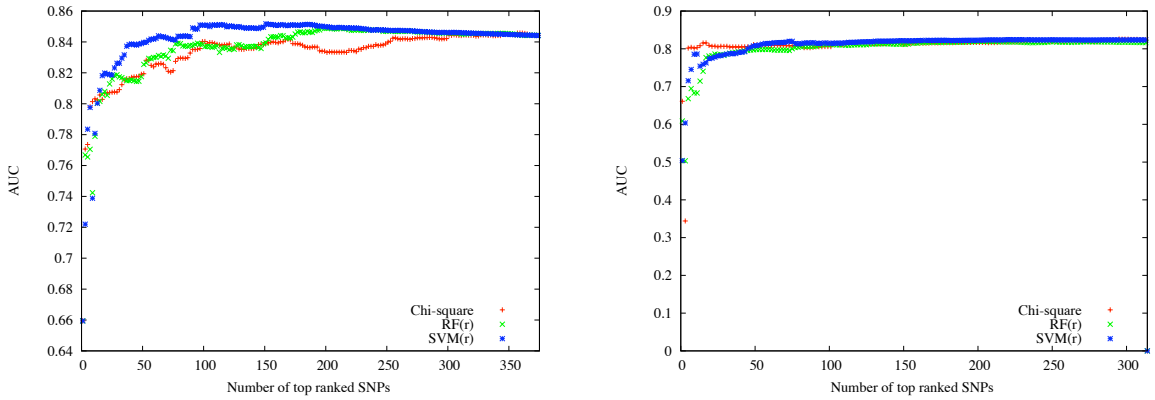


Figure 7: ROC area under curve of the composite odds ratio score on the WTCCC type 1 diabetes study as a function of top ranked SNPs obtained from the GoKinD study by the three different methods



(a) Prediction on GoKinD and learning from WTCCC

(b) Prediction on WTCCC and learning from GoKinD

Figure 8: ROC area under curve of the support vector machine score on the GoKinD type 1 diabetes study as a function of top ranked SNPs obtained from the WTCCC study and vice-versa

10 Prediction of arthritis risk by cross-validation

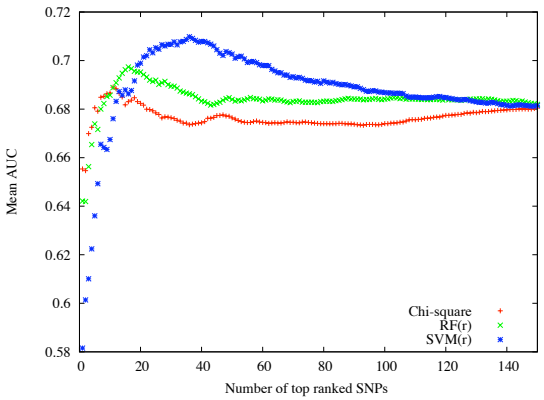
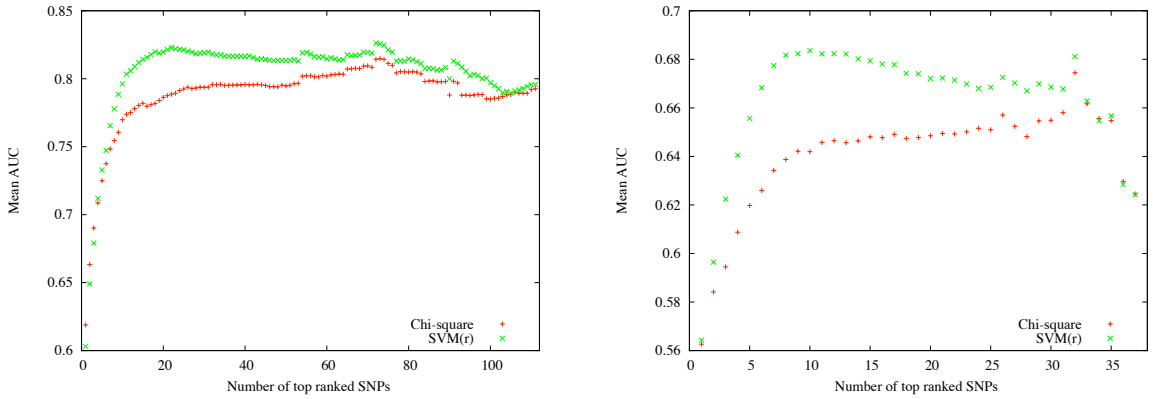


Figure 9: ROC area under curve of the composite odds ratio score as a function of top ranked SNPs given by the three different methods on the WTCCC arthritis study

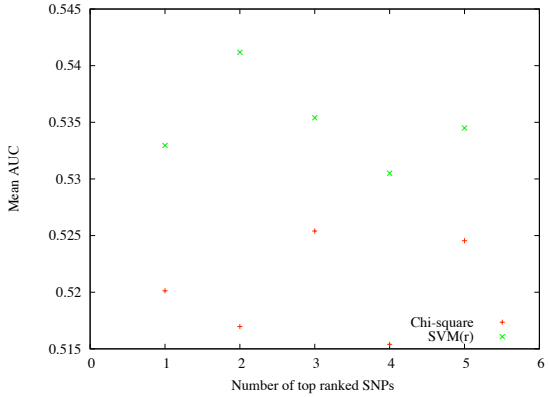
11 Prediction of disease risk on simulated data



(a) Relative risk 2

(b) Relative risk 1.5

Figure 10: ROC area under curve of the composite odds ratio score on simulated data. We computed SNP rankings from the general performance simulated data and predicted on the risk prediction test data which has same settings as training except that there are only a 100 case and 100 control subjects.



(a) Relative risk 1.25

Figure 11: ROC area under curve of the composite odds ratio score on simulated data. See above caption for more.

References

N. P. Jewell. *Statistics for Epidemiology*. Chapman & Hall, 2003.

Naomi R. Wray, Michael E. Goddard, and Peter M. Visscher. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Research*, 17:1520–1528, 2007.

Ethem Alpaydin. *Machine Learning*. MIT Press, 2004.

Shaun Purcell. Plink v1.05 available at http://pngu.mgh.harvard.edu/purcell/plink/, 2009.

S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham. Plink: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, 81, 2007.

Welcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, 2007.

Patricia W. Mueller, John J. Rogus, Patricia A. Cleary, Yuan Zhao, Adam M. Smiles, Michael W. Steffes, Jean Bucksa, Therese B. Gibson, Suzanne K. Cordovado, Andrzej S. Krolewski, Concepcion R. Nierras, and James H. Warram. Genetics of Kidneys in Diabetes (GoKinD) Study: A Genetics Collection Available for Identifying Genetic Susceptibility Factors for Diabetic Nephropathy in Type 1 Diabetes. *J Am Soc Nephrol*, 17(7):1782–1790, 2006.

The Gain Collaborative Research Group. New models of collaboration in genome-wide association studies: the genetic association information network. *Nature*, 39:1045–1051, 2007.

D. M. Evans, P. M. Visscher, and N. R. Wray. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Human Molecular Genetics*, 18(18):3525–3531, 2009.

Erin M. Hulbert, Luc J. Smink, Ellen C. Adlem, James E. Allen, David B. Burdick, Oliver S. Burren, Christopher C. Cavnor, Geoffrey E. Dolman, Daisy Flamez, Karen F. Friery, Barry C. Healy, Sarah A. Killcoyne, Burak Kutlu, Helen Schuilenburg, Neil M. Walker, Josyf Mychaleckyj, Decio L. Eizirik, Linda S. Wicker, John A. Todd, and Nathan Goodman. T1DBase: integration and presentation of complex data for type 1 diabetes research. *Nucl. Acids Res.*, 35(1):D742–746, 2007.

Type 1 Diabetes Consortium. Known type 1 diabetes associated regions, 2009.