OPEN ACCESS Freely available online

PLoS GENETICS

Genome-Wide Interaction-Based Association Analysis Identified Multiple New Susceptibility Loci for Common Diseases

Yang Liu[1], Haiming Xu[2], Suchao Chen[3], Xianfeng Chen[1], Zhenguo Zhang[1], Zhihong Zhu[2], Xueying Qin[3], Landian Hu[1], Jun Zhu[2], Guo-Ping Zhao[4], Xiangyin Kong[1]*

1 The Key Laboratory of Stem Cell Biology, Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences and Ruijin Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, People's Republic of China, **2** Institute of Bioinformatics, Zhejiang University, Hangzhou, People's Republic of China, **3** State Key Lab of CAD&CG, Zhejiang University, Hangzhou, People's Republic of China, **4** National Human Genome Center, Shanghai, People's Republic of China

Abstract

Genome-wide interaction-based association (GWIBA) analysis has the potential to identify novel susceptibility loci. These interaction effects could be missed with the prevailing approaches in genome-wide association studies (GWAS). However, no convincing loci have been discovered exclusively from GWIBA methods, and the intensive computation involved is a major barrier for application. Here, we developed a fast, multi-thread/parallel program named "pair-wise interaction-based association mapping" (PIAM) for exhaustive two-locus searches. With this program, we performed a complete GWIBA analysis on seven diseases with stringent control for false positives, and we validated the results for three of these diseases. We identified one pair-wise interaction between a previously identified locus, *C1orf106*, and one new locus, *TEC*, that was specific for Crohn's disease, with a Bonferroni corrected $P<0.05$ ($P=0.039$). This interaction was replicated with a pair of proxy linked loci ($P=0.013$) on an independent dataset. Five other interactions had corrected $P<0.5$. We identified the allelic effect of a locus close to *SLC7A13* for coronary artery disease. This was replicated with a linked locus on an independent dataset ($P=1.09\times10^{-7}$). Through a local validation analysis that evaluated association signals, rather than locus-based associations, we found that several other regions showed association/interaction signals with nominal $P<0.05$. In conclusion, this study demonstrated that the GWIBA approach was successful for identifying novel loci, and the results provide new insights into the genetic architecture of common diseases. In addition, our PIAM program was capable of handling very large GWAS datasets that are likely to be produced in the future.

Citation: Liu Y, Xu H, Chen S, Chen X, Zhang Z, et al. (2011) Genome-Wide Interaction-Based Association Analysis Identified Multiple New Susceptibility Loci for Common Diseases. PLoS Genet 7(3): e1001338. doi:10.1371/journal.pgen.1001338

Editor: David B. Allison, University of Alabama at Birmingham, United States of America

Received August 22, 2010; **Accepted** February 15, 2011; **Published** March 17, 2011

Funding: This work is supported by the National High Technology Research and Development Program of China (2006AA02Z330, 2006AA02A301), the National Basic Research Program of China (No. 2007CB512202, 2011CBA00400, 2011CB510100), the National Natural Science Foundation of China (No. 30530450, 30871356), and the Knowledge Innovation Program of the Chinese Academy of Sciences (Grant No. KSCX1-YW-R-74). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: xykong@sibs.ac.cn

Introduction

Recent genome-wide association studies (GWAS) have identified many common genetic variants associated with common diseases. This has rapidly expanded our knowledge of the genetic architecture of these diseases. For example, the Wellcome Trust Case Control Consortium (WTCCC) study [1] and other large-scale GWASs (including meta-analyses) have discovered many susceptibility loci for common diseases, including coronary artery disease (CAD) [2], Crohn's disease (CD) [3,4], type 1 diabetes (T1D) [5], and type 2 diabetes (T2D) [6]. However, compared with the successes of single-locus approaches, the achievements of interaction-based approaches, which seek susceptibilities that derive from gene-gene interactions, have lagged behind [7,8]. Thus, gene-gene interactions that are largely undetected may explain some of the heritability of common diseases [9]. Most reported interactions are currently found through candidate approaches, which incorporate prior biological knowledge. Moreover, very few interactions have been confirmed in an independent population.

Genome-wide interaction-based association (GWIBA) analysis uses markers to conduct genome-wide screens without prior candidate selection. In addition, GWIBA incorporates interaction effects among genetic variants. Many interaction-based methods for GWIBA are currently available, including a logistic regression-based method [10]; in addition, several methods have been recently developed [11–15]. However, no studies on real data have successfully identified novel disease-associated loci. Two studies reported non-significant results on small datasets [11,14]; several studies with the WTCCC dataset reported problematic interactions [12,13,15], and those were found to be probable false positives in this study. Thus, the GWIBA methods have identified very few new loci convincingly, and none of the detected interactions have been replicated to date. In addition, the computational time was a major barrier for GWIBA analyses on large-scale GWAS datasets. Most previous studies resorted to stochastic searches, or partial search strategies based on biological knowledge [12–18]. Until recently, genome-wide association studies have followed the traditional single-locus approach and

Author Summary

Recent studies on the genetic basis of common diseases have identified many loci that confer disease susceptibility. However, much of the heritability of these diseases remains unexplained. Loci involved in gene–gene interactions are considered cryptic, because they confer susceptibility, but may not generate a detectable signal on their own. These interactions may account for the "missing heritability" of common diseases. Theoretically, these interactions can be identified with the genome-wide interaction-based association analysis. But, in reality, very few gene–gene interactions have been identified with that method, and most were based on prior biological knowledge. Here, we applied a parallel computing technique that facilitated the identification of multiple new cryptic susceptibility loci involved in common diseases. We applied stringent control for false positives, and we validated our findings with independent datasets. This study demonstrated that interactions between gene loci could be successfully identified with the genome-wide interaction-based approach. With this approach, we also identified cryptic loci with moderate single-locus effects. The identified loci and interactions merit further investigations for fine mapping and functional analyses. Our results extend the current knowledge of common diseases for future studies in genetic mapping. This approach is applicable to current and future genome-wide association datasets.

have investigated gene-gene interactions only through candidate approaches.

In this study, our main aim was to discover novel susceptibility loci by identifying interaction effects in a GWIBA analysis with the large-scale WTCCC dataset [1]. We also aimed to confirm these novel loci in independent datasets. To that end, we identified several novel susceptibility loci with replication/validation evidence, and the results provide new insights into the genetic architecture of common diseases.

Results

Identification of Gene Interactions

We performed a complete GWIBA analysis with validation analyses. We started with the WTCCC dataset [1], which contained ~2,000 cases for seven diseases and ~3,000 shared controls (Materials and Methods). The quality-controlled WTCCC data were used as input for the "pair-wise interaction-based association mapping" (PIAM) program, and we performed an exhaustive two-locus search for each disease (Materials and Methods). We used the single-locus likelihood ratio test (LRT) *p*-value (5×10^{-7}) as a cutoff value for incorporating the single-locus effects in the PIAM searches. The cutoff value was based on the significance threshold set by WTCCC for single-locus analyses. This prevented the marginal effects of a few loci from dominating the interactions. The computation was performed with the PIAM program running in parallel on computer clusters.

In the initial search, we used the cases and the shared controls of the WTCCC data to screen single-nucleotide polymorphism (SNP) pairs that passed a *p*-value threshold of $P<50/L$, where L was the total number of two-locus combinations for each disease. The threshold allowed SNP pairs with *p*-values that were 1,000 times larger than the significance level of $0.05/L$. During the calculation, the distributions of two-locus statistics were evaluated with the approximate statistical distribution method in PIAM (Materials and Methods), which generated genome-wide two-locus quantile-quantile plots (Figure S1). We obtained 2,570 SNP pairs for seven diseases at these screening thresholds (Table S1A), after excluding 20,968 SNP pairs within the major histocompatability complex (MHC) region for rheumatoid arthritis (RA) and T1D (Table S1B).

Although many SNP pairs had rather significant *p*-values, there were an overwhelming number of false positive results observed. We found that the initial SNP quality control performed by the WTCCC was not sufficiently stringent for the interaction searches, due to sparse data and poor genotyping quality. The sparseness of the data was due to the constraint that we used two-locus genotype interaction analyses, instead of the single-locus analysis applied by the WTCCC; this relative sparseness of data conferred a higher sensitivity to genotyping errors. Therefore, a stringent additional SNP quality control was applied (Materials and Methods). A total of 1,392 SNP pairs passed this additional quality control (Table S1C).

After the initial search, these 1,392 SNP pairs were tested with the expanded controls to gain greater statistical power (Materials and Methods). We retained 634 SNP pairs that gave Bonferroni corrected $P<0.5$ (Table S1D), according to the numbers of available two-locus tests (Table S2).

Among the results from the 634 SNP pairs, we observed two major types of problematic results, irrespective of the SNP quality control. The first problem was that we found many "interactions" between known susceptibility loci with large marginal effects. These "interactions" might have resulted from marginal effects, according to the two-locus LRT tests that incorporated both marginal and pure interaction effects. To control for this problem, we used a strategy similar to BEAM [19], where we compared the two-locus *p*-values with the single-locus *p*-values. The second problem was that we found 88 SNP pairs with linked SNPs (Table S1E); most of these gave quite significant *p*-values, but were identified as artificial associations that could be separated into two types, one was a batch effect and the other was a genotype clustering problem. These artificial associations were due to sparse data and genotyping artifacts. Later, we found that some previously reported interactions were probably these kinds of artificial associations [12,13,15] (details in Discussion). Therefore, a stringent result filter was applied to filter out these false positive interactions (Materials and Methods). Thus, we removed 536 SNP pairs with excessive marginal effects, and 85 SNP pairs with the two kinds of artificial associations. Within the 88 SNPs pairs with linked SNPs, 3 pairs were not affected by artificial associations; therefore, these interactions were considered true haplotypic associations. These 3 SNP pairs were located in regions known to be associated with CD, thus, we did not present these results in detail here, except in the corresponding regional signal plots (Figure S2) and odds ratio (OR) tables (Table S3). Finally, 10 SNP pairs with unlinked SNPs remained qualified (Table S1F).

After the result filtering, the simultaneous searches identified an interaction between rs7522462 (on *C1orf106*) and rs11945978 (on *TEC*) for CD with a Bonferroni corrected $P<0.05$, and another five pairs of regions associated with CAD, CD, T1D, and T2D with Bonferroni corrected $P<0.5$ (Table 1; Figure 1). Among the above six pairs of regions, the interaction between rs7522462 and rs11945978 for CD, and the allelic effect of rs6470733 (close to *SLC7A13*) for CAD were replicated by proxy linked SNPs. In addition, we validated one pair of interacting regions around rs153423 (near *SPRY4*) and rs748855 (on *NOD2*) for CD, one single region around rs1501540, and one pair of interacting regions around rs11731175 and rs11236365 (on *SLCO2B1*) for T2D, all with nominal $P<0.05$, through local validation analyses (Materials and Methods; Table 1; Figure 2). We then performed

Table 1. Identified gene interactions.

Disease	SNP	Chr	Nearest Gene	Trend *p*-value	Validation of allelic effect	Pure interaction *p*-value	Validation of interaction	Two-locus *p*-value (expanded)	Test numbers	Corrected *p*-value (expanded)
CAD	rs9397512	6q25.2	*SYNE1*	5.68×10^{-3}	Unavailable	1.54×10^{-8}	Unavailable	8.82×10^{-12}	4.31×10^{10}	0.380
	rs6470733	8q21.3	*SLC7A13*	9.18×10^{-4}	$P_{PR}=1.09\times10^{-7}$					
CD	rs7522462	1q32.1	*C1orf106*	2.36×10^{-5}	Meta-analysis	5.03×10^{-6}	$P_{PR}=0.013$	8.90×10^{-13}	4.33×10^{10}	0.039*
	rs11945978	4p12	*TEC*	0.016	$P_{PR}=0.047$					
CD	rs153423	5q31.3	*SPRY4*	3.21×10^{-3}	Not significant	4.28×10^{-5}	$P_{LV}=0.034$	3.36×10^{-12}	4.33×10^{10}	0.146
	rs748855	16q12.1	*NOD2*	2.63×10^{-7}	Known region					
T1D	rs7310460	12p13.31	*CLEC2D*	9.43×10^{-4}	Meta-analysis	1.09×10^{-8}	Unavailable	5.87×10^{-12}	4.29×10^{10}	0.252
	rs2302270	12q24.32	-	0.673	-					
T2D	rs1501540	1p34.3	-	2.24×10^{-4}	$P_{LV}=0.022$	2.13×10^{-7}	Not significant	1.92×10^{-12}	4.26×10^{10}	0.082
	rs7359782	18p11.21	*C18orf58*	4.51×10^{-3}	Not significant					
T2D	rs11731175	4q35.2	-	0.115	-	1.68×10^{-11}	$P_{LV}=0.029$	7.00×10^{-12}	4.26×10^{10}	0.298
	rs11236365	11q13.4	*SLCO2B1*	0.715	-					

Six pairs of SNPs that represented interacting loci. Chr: chromosome and cytoband information. Nearest Gene: When the nearest annotated gene was >500 kb away from the SNPs, it was not listed. The trend *p*-values were obtained with the shared controls. The validation of allelic effect for loci with original trend *p*-values>0.05 is not presented. Validation status: Unavailable, data unavailable for validation; P_{PR}, *p*-value of the proxy replication; Meta-analysis, susceptibility locus found by meta-analysis studies after the WTCCC study; Not significant, validation was not significant ($P>0.05$); Known region, a susceptibility region that was known at the time of the WTCCC study; P_{LV}, *p*-value of the local validation. The last five columns contain the interaction results. Pure interaction *p*-values were obtained with the shared controls. Two-locus *p*-value (expanded): two-locus LRT *p*-value according to the search situation, and the two-locus *p*-values and the corresponding corrected *p*-values for the final significance were obtained in the expanded control analysis.

*Bonferroni corrected $P<0.05$. Because all of these loci were obtained in the simultaneous searches, in which the two-locus tests took account of all effects of the two loci, therefore the main effects and the interaction effects were examined in the validation analyses, and the following criterion was used to determine the validation status of each locus: (1) if the pure interaction effect was validated (i.e. $P<0.05$), both of the loci and their interaction effect were validated, irrespective of the validations of the main effects; (2) if the pure interaction effect was not validated, then the validation status of a locus was determined by the validation of its main effect.

doi:10.1371/journal.pgen.1001338.t001

the three-locus conditional searches based on the six pairs of SNPs listed in Table 1; this did not produce any significant results.

We did not identify any interactions for bipolar disorder (BD), hypertension (HT), or RA, according to the significance thresholds and result filtering applied (except the interactions within the MHC region for RA). In fact, a single-locus analysis did not identify significant results for HT, and only one significant locus was associated with BD, but this has not been replicated to date [1,20]. This may indicate that the quality control and result filtering we performed was effective for removing random false positives and artificial associations.

Within each SNP pair in Table 1, the SNPs were independent in the controls and dependent in the cases (Table 2). The SNPs in Table 1 showed good genotype clustering (Figure S3), and did not present any significant deviations from Hardy-Weinberg equilibrium (HWE, $P>0.05$). Note that the corrected two-locus *p*-values in Table 1 were only corrected within each disease.

CAD

Only one pair of interacting loci was associated with CAD. The SNPs were rs9397512 and rs6470733, located at the intron of *SYNE1* and 7 kb downstream from *SLC7A13*, respectively. Note that this interaction only gave a moderate corrected *p*-value of 0.380. However, this pair of regions generated wide, block-like interaction signals (Figure 1), with strong linkage disequilibrium (LD) (Figure S4, plotted with Haploview [21]). For this interaction, when the rs6470733 genotypes were paired with the TT genotype stratum of rs9397512, the effects were in the opposite direction compared to those observed when the rs6470733 genotypes were paired with the CC and CT strata (Table 3). The highest OR relative to the most common homozygote combination (2.95) was higher than the OR under the assumption of an additive effect of the two loci (1.95). The two loci also showed moderate single-locus allelic effects, especially rs6470733.

In order to validate the association of the two loci identified for CAD, we used the online results of the German MI Family Study (GerMIFS) [2], which included 875 cases and 1644 controls (Materials and Methods). Surprisingly, we found that rs13262822, which was 1.5 kb downstream from rs6470733 and had an $r^2=0.90$ (based on the WTCCC shared controls), showed a rather significant allelic effect with a trend test $P=1.09\times10^{-7}$ (Table 4). In addition, rs13262822 showed an allelic effect in the WTCCC data with a trend test $P=1.81\times10^{-3}$ and an association in the same direction. Therefore, the allelic effect of the original SNP, rs6470733, was replicated by its proxy SNP rs13262822, in strong LD. Interestingly, in the GerMIFS data, the minor allele frequency (MAF) of rs13262822 was a bit larger than that in the WTCCC data, and the OR of 1.39 was much higher compared to the OR of 1.16 reported in the WTCCC. The previous paper did not identify the rs13262822 locus because the trend *p*-value of rs13262822 in the WTCCC data marginally failed the 0.001 threshold before the combined analysis [2]. At this time, the online result from the GerMIFS data is not sufficient to confirm the marginal effect of rs9397512 or the interaction effect. *SYNE1* was previously suggested as a potential mediator of cardiomyopathy, because it showed muscle-specific inner nuclear envelope expression and a physical interaction with lamin A/C [22]. Furthermore, a recent study suggested that *SYNE1* was involved in the pathogenesis of Emery Dreifuss muscular dystrophy through skeletal muscle cell destruction [23], which emphasized the functional role of *SYNE1* in muscles. *SLC7A13* is a cationic amino acid transporter, and two early studies showed that cationic amino

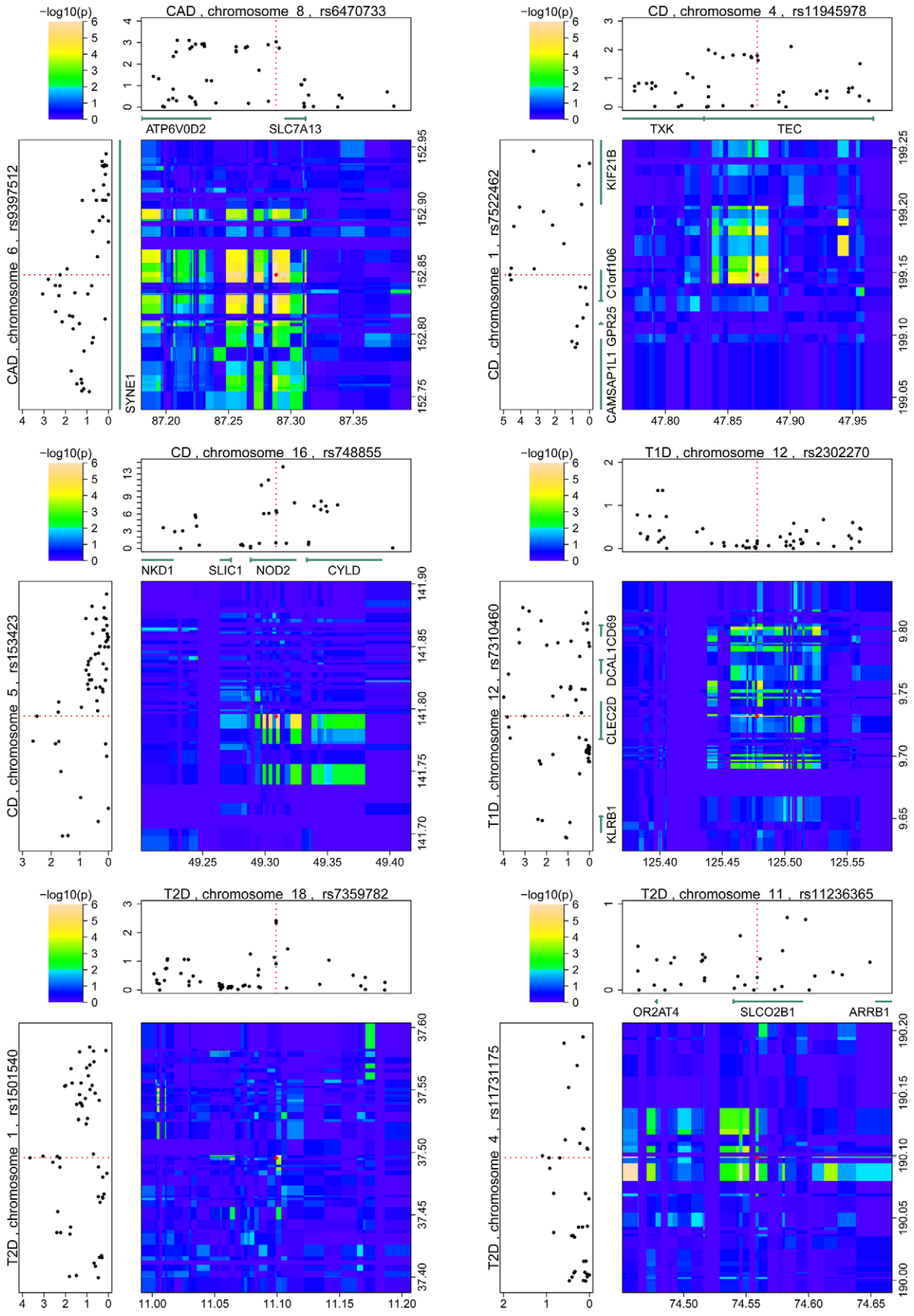
−log10(p)
CAD , chromosome 8 , rs6470733
ATP6V0D2
SLC7A13
CAD , chromosome 6 , rs9397512
SYNE1
CD , chromosome 4 , rs11945978
TXK
TEC
CD , chromosome 1 , rs7522462
KIF21B
C1orf106
GPR25
CAMSAP1L1
CD , chromosome 16 , rs748855
NKD1
SLIC1
NOD2
CYLD
CD , chromosome 5 , rs153423
T1D , chromosome 12 , rs2302270
T1D , chromosome 12 , rs7310460
DCAL1
CD69
CLEC2D
KLRB1
T2D , chromosome 18 , rs7359782
T2D , chromosome 1 , rs1501540
T2D , chromosome 11 , rs11236365
OR2AT4
SLCO2B1
ARRB1
T2D , chromosome 4 , rs11731175

Figure 1. Regional signal plots of the interactions in Table 1 with WTCCC data. SNPs within 100 kb of the most significant SNP pairs are shown. For each panel, the upper left plot is the color key for the interaction signal plot at the lower right; the lower left and upper right plots are single-locus signal plots with gene annotations. These plots are aligned by chromosome positions in Mb, which are based on NCBI build 36. The red dotted lines and the red dot in the middle indicate the position of the most significant pair of SNPs in the corresponding regions. The solid black dots in the single-locus signal plots denote the trend test *p*-values (values indicated by the numbered axes) and the colored images in the interaction signal plots denote the pure interaction *p*-values (values indicated by the color key in the upper left box); these were obtained with the shared controls and transformed by a negative logarithm.
doi:10.1371/journal.pgen.1001338.g001

acid transporters may be related to atherosclerotic lesion formation by regulating L-ornithine transport and polyamine synthesis in vascular smooth muscle [24,25]. Thus, these two genes may be involved in different, but related aspects of CAD pathogenesis. This could explain the statistical interaction between the two regions.

CD

Two pairs of interacting loci were associated with CD. The first interaction was between rs7522462, which is in the region of *C1orf106* gene, and rs11945978, which is in a newly identified region of the *TEC* gene. The *C1orf106* region was previously identified in a meta-analysis after the WTCCC study, which

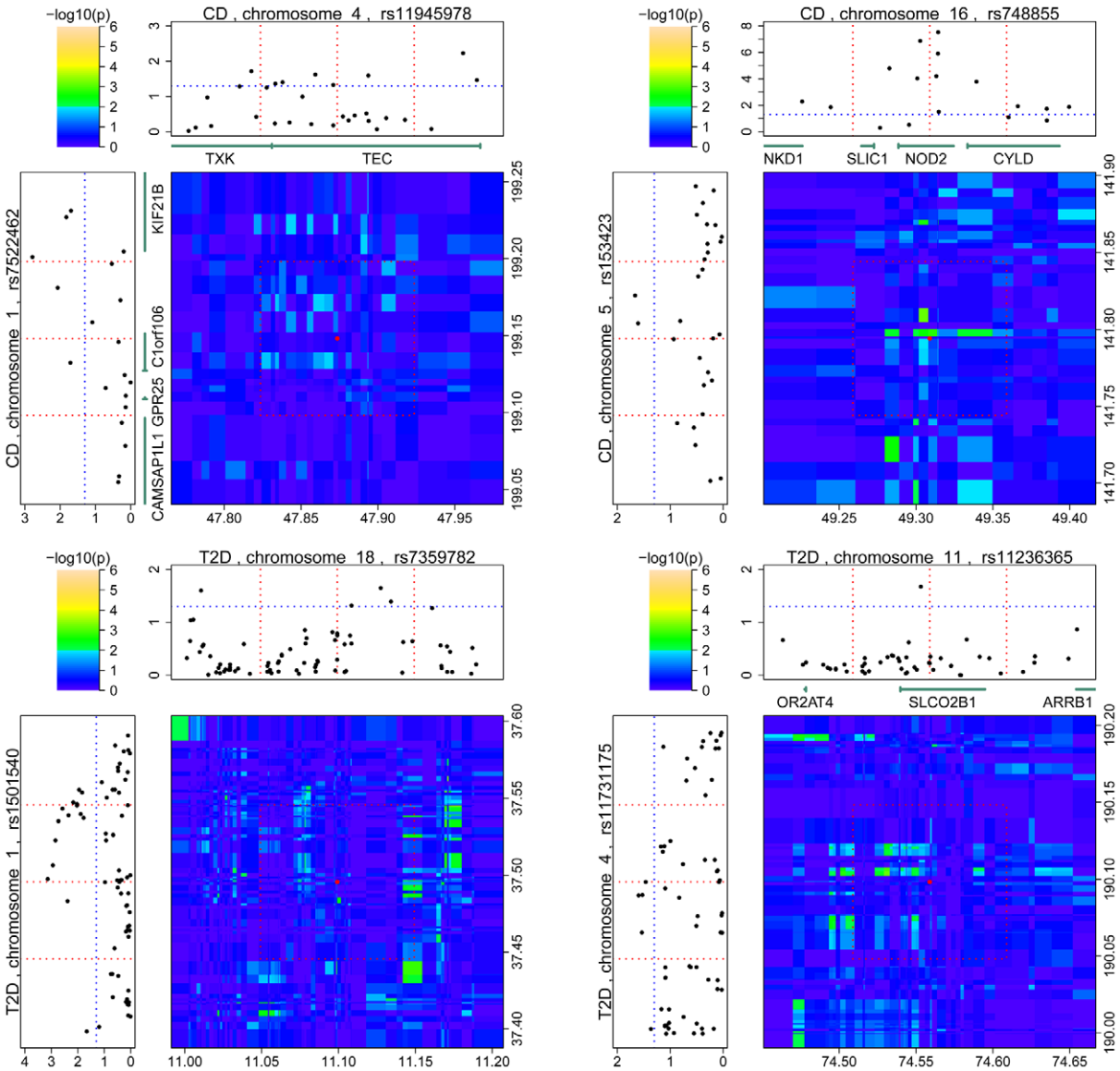

Figure 2. Regional signal plots of the interactions observed in validation datasets. The signal plots of three interactions for CD and two interactions for T2D are shown. The interactions were observed in the IBDGC non-Jewish population data and the GENEVA Diabetes Study data, respectively. The format of this figure is similar to that described in Figure 1; the red dotted lines and the red dot in the middle indicate the positions of the originally identified SNP pair. Two other red dotted lines for each single-locus signal plot and the red box in the interaction signal plot were added to indicate the regions used for the local validation tests; the blue dotted lines indicate the *p*-value thresholds of 0.05. (The interaction between rs7522462 and rs11945978 was replicated by proxy SNPs; therefore, its local validation was disregarded).
doi:10.1371/journal.pgen.1001338.g002

Table 2. Tests for SNP independence in the interactions observed in the case and control groups.

		Simulation 1 *P*		Simulation 2 *P*		Simulation 3 *P*	
Disease	**Interaction**	**Case**	**Control**	**Case**	**Control**	**Case**	**Control**
CAD	rs9397512, rs6470733	0.0015	0.3224	0.0015	0.3288	0.0015	0.3314
CD	rs7522462, rs11945978	0.0027	0.0577	0.0022	0.0588	0.0022	0.0568
CD	rs153423, rs748855	0.0273	0.2240	0.0283	0.2300	0.0267	0.2271
T1D	rs7310460, rs2302270	0.0020	0.1675	0.0033	0.1633	0.0021	0.1613
T2D	rs1501540, rs7359782	0.0021	0.5032	0.0025	0.4957	0.0022	0.5002
T2D	rs11731175, rs11236365	0.0020	0.1453	0.0023	0.1480	0.0020	0.1492

Each row represents one gene-gene interaction listed in Table 1. Chi-square tests of 3 by 3 contingency tables were used to determine whether two SNPs were dependent in the case group or in the control group. For each test, 10,000 Monte Carlo simulations were used to obtain the *p*-value; each test was repeated 3 times for both the case group and the control group. The tests were performed with the R statistical software (http://www.r-project.org/).

doi:10.1371/journal.pgen.1001338.t002

included data from both the WTCCC and from the national institute of diabetes, digestive, and kidney diseases (NIDDK) inflammatory bowel disease genetics consortium (IBDGC) [4]. This interaction gave a Bonferroni corrected $P = 0.039$. The interaction signal showed a clear block that extended over several tens of kb (Figure 1). In the IBDGC data, a weak interaction signal also appeared at the corresponding regions (Figure 2). For this interaction, the single-locus effect of rs7522462 varied significantly among the genotype strata of rs11945978, which indicated the interaction (Table 3), and the effect of rs7522462 was strongest in the rs11945978 CC stratum.

Validation analysis with the IBDGC data supported the interaction between rs7522462 and rs11945978 (Materials and Methods). We selected proxy SNPs in the IBDGC data instead of the original SNPs, according to HapMap [26] CEU r^2 values. For rs7522462, two proxy SNPs were in moderate LD: rs296533, which is 16 kb upstream, with an $r^2 = 0.44$, and rs296547, which is 10 kb downstream, with an $r^2 = 0.79$. For rs11945978, the proxy SNP rs2089509 showed perfect linkage disequilibrium (LD) in the HapMap CEU population. The allelic effects, interaction effect, and combined effect of the proxy SNP combination of rs296533 and rs2089509 were replicated in the IBDGC non-Jewish population data (rs296533 trend $P = 0.020$, rs2089509 trend $P = 0.047$, pure interaction $P = 0.013$, two-locus $P = 0.001$). The ORs showed trends similar to those in the WTCCC data, particularly in the CC genotype stratum of rs11945978 (corresponding to the GG genotype stratum of rs2089509) (Table 5). The trend *P* of rs7522462 stratified by rs11945978 CC, and the trend *P* of rs296533 stratified by rs2089509 GG were 2.05×10^{-8} and 1.35×10^{-3}, respectively. The risk alleles of rs7522462 and its proxy, rs296533, and the risk alleles of rs11945978 and its proxy, rs2089509, comprised the major haplotypes according to the HapMap data. This indicated the same association direction in the WTCCC data and the IBDGC non-Jewish population data. Although the interaction between rs296533 and rs2089509 was not significant in the IBDGC Jewish population data (with quite a small sample size), the interaction showed a similar pattern (Table S4). Nevertheless, the downstream proxy SNP, rs296547, had a larger r^2 value of 0.79 and the interaction was not significant in the IBDGC data. This may be explained by the small sample size and the LD difference between the HapMap data and the IBDGC data for the marker loci and the causing loci. For SNPs that were either ungenotyped in the WTCCC or in the IBDGC non-Jewish population data (rs7522462, rs11945978, rs296533, rs2089509), the corresponding genotypes were imputed (Materials and Methods). We found a consistent interaction between rs296533 and rs2089509, which was significant in both the IBDGC non-Jewish population data ($P = 0.013$) and the imputed WTCCC data ($P = 0.015$), and they showed a similar interaction pattern (Table S4). A previous study found that the expression of *TEC* was up-regulated upon T-cell activation, and *Tec* overexpression in lymphocyte cell lines was sufficient to induce phosphorylation of phospholipase C gamma and activation of nuclear factor of activated T cells [27]; moreover, over-activation of T cells is a typical feature of CD.

The second interaction for CD was between rs153423 and rs748855, which gave a corrected *P* of 0.146. The latter SNP lies in the early identified *NOD2* gene [1]; the former SNP is located about 100 kb upstream from the *SPRY4* gene, and the association signal extended fairly close to the gene (Figure 1). The two-locus pattern showed that rs153423 was epistatic to rs748855, because the most common rs153423 genotype (AA) masked a considerable single-locus effect of rs748855 (Table 3). Locus-based replication for this interaction failed, and local validation of the interaction with the IBDGC non-Jewish population data indicated a nominally significant interaction ($P = 0.034$; Figure 2). A previous study showed that *SPRY4* suppressed vascular epithelial growth factor-induced, *Ras*-independent activation of *Raf1* [28]; moreover, another study suggested that vascular epithelial growth factor-A signaling was related to CD through angiogenesis [29].

T1D

Only one pair of interacting loci was associated with T1D. The SNPs, rs7310460 and rs2302270, interacted with a moderate corrected *P* of 0.252. The 12p13.31 region around rs7310460 was previously found in a meta-analysis study conducted after the WTCCC study [5]. This region harbors many immunoregulatory genes, including *CLEC2D*. In contrast, rs2302270 is mapped to an intergenic region. The association signal for the interaction effect extended about 100 kb for both regions with clear borders, and it included the previously suggested *CD69* gene [5] (Figure 1). The association pattern showed that rs2302270 was epistatic to rs7310460 (Table 3). We currently have no available data to validate the association of the rs2302270 region or the interaction.

T2D

Two pairs of interacting loci were associated with T2D. The first interaction was between rs1501540 and rs7359782, which gave a corrected *P* of 0.082. The rs1501540 SNP is mapped to a region with no annotated genes, and rs7359782 is located 238 kb upstream of *C18orf58*. The interaction signal was very narrow; however it was not restricted to a single SNP (Figure 1). The

Table 3. OR tables for the SNP pairs shown in Table 1.

CAD		rs6470733			CD		rs11945978		
	rs9397512	AA	AG	GG		rs7522462	CC	CT	TT
OR	CC	1	1.52 (1.24,1.87)	1.52 (1.05,2.20)	OR	GG	1	0.64 (0.54,0.76)	0.97 (0.75,1.27)
	CT	1.33 (1.10,1.60)	1.21 (0.99,1.47)	2.95 (2.16,4.02)		GA	0.77 (0.64,0.92)	0.64 (0.53,0.77)	0.64 (0.47,0.86)
	TT	1.81 (1.43,2.28)	1.43 (1.10,1.86)	0.87 (0.49,1.54)		AA	0.32 (0.22,0.48)	0.78 (0.56,1.08)	0.44 (0.23,0.81)
OR1	CC	1	1.52 (1.24,1.87)	1.52 (1.05,2.20)	OR1	GG	1	0.64 (0.54,0.76)	0.97 (0.75,1.27)
	CT	1	0.91 (0.76,1.09)	2.22 (1.65,3.00)		GA	1	0.83 (0.68,1.02)	0.83 (0.61,1.13)
	TT	1	0.79 (0.59,1.06)	0.48 (0.27,0.86)		AA	1	2.43 (1.49,3.96)	1.36 (0.66,2.79)
OR2	CC	1	1	1	OR2	GG	1	1	1
	CT	1.33 (1.10,1.60)	0.80 (0.65,0.97)	1.94 (1.25,3.01)		GA	0.77 (0.64,0.92)	1.00 (0.82,1.22)	0.65 (0.45,0.94)
	TT	1.81 (1.43,2.28)	0.94 (0.72,1.23)	0.57 (0.30,1.10)		AA	0.32 (0.22,0.48)	1.22 (0.88,1.70)	0.45 (0.23,0.86)
CD		**rs748855**			**T1D**		**rs2302270**		
	rs153423	**AA**	**AG**	**GG**		**rs7310460**	**GG**	**GA**	**AA**
OR	AA	1	0.97 (0.83,1.14)	0.71 (0.57,0.90)	OR	TT	1	0.62 (0.48,0.79)	0.62 (0.30,1.31)
	AG	1.76 (1.43,2.16)	1.08 (0.88,1.31)	0.70 (0.51,0.97)		TA	1.12 (0.96,1.32)	1.27 (1.04,1.53)	0.82 (0.48,1.41)
	GG	1.14 (0.69,1.90)	0.29 (0.12,0.70)	1.39 (0.66,2.92)		AA	0.89 (0.72,1.10)	1.59 (1.22,2.09)	3.79 (1.56,9.21)
OR1	AA	1	0.97 (0.83,1.14)	0.71 (0.57,0.90)	OR1	TT	1	0.62 (0.48,0.79)	0.62 (0.30,1.31)
	AG	1	0.61 (0.48,0.78)	0.40 (0.28,0.56)		TA	1	1.13 (0.94,1.34)	0.73 (0.43,1.25)
	GG	1	0.26 (0.10,0.70)	1.22 (0.50,2.95)		AA	1	1.79 (1.33,2.39)	4.24 (1.73,10.4)
OR2	AA	1	1	1	OR2	TT	1	1	1
	AG	1.76 (1.43,2.16)	1.11 (0.91,1.34)	0.98 (0.68,1.41)		TA	1.12 (0.96,1.32)	2.05 (1.58,2.65)	1.31 (0.53,3.24)
	GG	1.14 (0.69,1.90)	0.30 (0.13,0.72)	1.95 (0.91,4.17)		AA	0.89 (0.72,1.10)	2.58 (1.87,3.55)	6.07 (1.93,19.1)
T2D		**rs7359782**			**T2D**		**rs11236365**		
	rs1501540	**CC**	**CT**	**TT**		**rs11731175**	**TT**	**TC**	**CC**
OR	GG	1	0.92 (0.75,1.13)	1.27 (0.90,1.79)	OR	GG	1	1.23 (1.02,1.46)	1.06 (0.69,1.63)
	GA	0.97 (0.81,1.16)	0.82 (0.68,1.00)	0.59 (0.42,0.84)		GT	1.14 (0.98,1.32)	1.28 (1.06,1.55)	0.89 (0.54,1.48)
	AA	0.79 (0.62,1.00)	0.73 (0.56,0.96)	0.03 (0.00,0.21)		TT	1.83 (1.43,2.35)	0.33 (0.20,0.56)	0.19 (0.02,1.51)
OR1	GG	1	0.92 (0.75,1.13)	1.27 (0.90,1.79)	OR1	GG	1	1.23 (1.02,1.46)	1.06 (0.69,1.63)
	GA	1	0.85 (0.71,1.01)	0.61 (0.43,0.86)		GT	1	1.13 (0.93,1.37)	0.79 (0.47,1.30)
	AA	1	0.93 (0.69,1.26)	0.04 (0.01,0.27)		TT	1	0.18 (0.10,0.32)	0.10 (0.01,0.83)
OR2	GG	1	1	1	OR2	GG	1	1	1
	GA	0.97 (0.81,1.16)	0.90 (0.73,1.10)	0.46 (0.29,0.73)		GT	1.14 (0.98,1.32)	1.05 (0.85,1.30)	0.84 (0.44,1.60)
	AA	0.79 (0.62,1.00)	0.80 (0.60,1.06)	0.02 (0.00,0.17)		TT	1.83 (1.43,2.35)	0.27 (0.16,0.46)	0.18 (0.02,1.47)

This table facilitates the interpretation of the statistical interactions. The statistics were obtained with the shared controls. For each SNP pair, there are three odds ratio tables: the OR, OR1, and OR2. Each OR table has 9 odds ratio values for 9 genotype combinations. 95% confidence intervals are shown in parentheses. OR: odds ratios of the two-locus genotype combinations, relative to the most common homozygote combination. OR1 and OR2: odds ratios of one of the two SNPs, where the samples were stratified by the genotypes of the other SNP; the interaction is indicated by different odds ratio values of one SNP between different genotype strata of the other SNP.

doi:10.1371/journal.pgen.1001338.t003

Table 4. Comparison of rs13262822 associations in the WTCCC CAD data and the GerMIFS data.

Study	Minor/risk allele	CC/CG/GG counts in case	CC/CG/GG counts in control	Case/control frequency of minor allele	Trend *P*	OR (95% CI)
WTCCC	C/C	183/752/986	195/1148/1593	0.362/0.326	1.81×10^{-3}	1.16 (1.06,1.27)
GerMIFS	C/C	58/448/229	144/683/745	0.552/0.395	1.09×10^{-7}	1.39 (1.22,1.59)

OR (95%CI): Odds ratios with 95% confidence intervals.

doi:10.1371/journal.pgen.1001338.t004

Table 5. Comparison of OR tables between two datasets for one CD interaction.

CD (WTCCC)		rs11945978			CD (IBDGC)		rs2089509		
	rs7522462	CC	CT	TT		rs296533	GG	GA	AA
OR	GG	1	0.64 (0.54,0.76)	0.97 (0.75,1.27)	OR	GG	1	0.69 (0.47,1.02)	0.60 (0.35,1.05)
	GA	0.77 (0.64,0.92)	0.64 (0.53,0.77)	0.64 (0.47,0.86)		GT	0.94 (0.63,1.39)	0.69 (0.47,1.03)	0.55 (0.30,1.01)
	AA	0.32 (0.22,0.48)	0.78 (0.56,1.08)	0.44 (0.24,0.81)		TT	0.19 (0.08,0.43)	0.57 (0.32,1.02)	0.87 (0.28,2.68)
OR1	GG	1	0.64 (0.54,0.76)	0.97 (0.75,1.27)	OR1	GG	1	0.69 (0.47,1.02)	0.60 (0.35,1.05)
	GA	1	0.83 (0.68,1.02)	0.83 (0.61,1.13)		GT	1	0.74 (0.49,1.11)	0.59 (0.32,1.09)
	AA	1	2.43 (1.49,3.96)	1.36 (0.66,2.79)		TT	1	3.03 (1.19,7.70)	4.67 (1.23,17.8)
OR2	GG	1	1	1	OR2	GG	1	1	1
	GA	0.77 (0.64,0.92)	1.00 (0.82,1.22)	0.65 (0.45,0.94)		GT	0.94 (0.63,1.39)	1.01 (0.67,1.51)	0.91 (0.44,1.90)
	AA	0.32 (0.22,0.48)	1.22 (0.88,1.70)	0.45 (0.23,0.86)		TT	0.19 (0.08,0.43)	0.82 (0.46,1.49)	1.45 (0.44,4.79)

Comparison of OR tables between the interaction of rs7522462 and rs11945978 in the WTCCC data with the shared controls (left) and the interaction of the proxy SNPs, rs296533 and rs2089509 in the IBDGC data (right). The legend to this table is the same as that of Table 3.

doi:10.1371/journal.pgen.1001338.t005

interaction and the region around rs7359782 failed in the validation analysis. However, the region around rs1501540 showed large allelic effects and was validated in the GENEVA Diabetes Study data with the local validation strategy ($P = 0.022$; Figure 2). In contrast, we did not detect any SNPs that were both significant and in the same direction in the two populations. Interestingly, we found that the significant SNPs in each dataset showed different frequencies between the two populations (MAF = 0.27 in the WTCCC data, MAF = 0.14 in the GENEVA data, with respect to the most significant SNPs in the associated region of each dataset, rs1501540 and rs302001); but, within each population, the significant SNPs showed similar frequencies.

The second interaction for T2D was between rs11731175 and rs11236365, which gave a corrected two-locus P of 0.298. Neither of the SNPs showed obvious marginal effect (trend P and genotypic LRT test $P>0.05$). The rs11731175 SNP lies within a region where the nearest annotated gene is more than 500 kb away, and rs11236365 is mapped to the *SLCO2B1* gene (Figure 1). The association pattern clearly showed that rs11731175 was epistatic to rs11236365. The GG and GT genotypes of rs11731175 masked the effect of rs11236365 (Table 3). However, when the genotype of rs11731175 was TT, rs11236365 showed a very strong effect. Moreover, the ORs of the TT and CT genotypes of rs11236365 relative to the most common homozygote combination were 1.83 and 0.33, respectively. It appeared that the C allele of rs11236365 provided a strong protective effect against T2D. The exact replication of this interaction with the GENEVA Diabetes Study data was not significant. Local validation of the interaction was nominally significant ($P = 0.029$), and the interaction signal was very close to the original signal (Figure 2). *SLCO2B1* is an organic anion transporting polypeptide, and one of its substrates, dehydroepiandrosterone-sulfate (DHEA-S, a direct metabolite of DHEA) [30], was found in several early studies to increase insulin sensitivity in a T2D mouse model [31,32], in rats [33–35], and in humans [36].

Discussion

Recent studies on the genetics of common diseases have revealed a lot of susceptibility loci and produced many tools for data analyses. However, the GWIBA approaches, which are prospective methods for discovering novel interacting loci, had not succeeded in identifying convincing interactions. In the present study, we developed an effective GWIBA approach that facilitated the discovery of novel loci. First, we used the parallel search program, PIAM, and implemented a simple statistical method and an optimized algorithm for detecting interactions. This could complete two-locus exhaustive searches on large-scale GWAS data in a short time. Second, in addition to the initial search, we used expanded controls with large sample sizes to gain statistical power for detecting interactions. Third, the results were carefully examined, and we found the artificial associations as well as the "interactions" with excessive single-locus effects. Finally, we employed independent datasets to validate the detected interactions; moreover, we introduced the "local validation" method for the validation of interactions between populations, where confounding factors may affect the consistency of the observed interactions.

Implications for the Genetic Architecture of Common Diseases

In Table 1, two regions were previously identified through meta-analysis studies. One region associated with CD, the *C1orf106*, which did not achieve significance in the WTCCC study, was subsequently identified in a meta-analysis study that included the WTCCC data and the IBDGC data [4]. We identified this region by including only the WTCCC data that showed corrected p-values<0.05. Also, the region on 12p13.31 that was associated with T1D was previously identified in a meta-analysis study of T1D [5]. These results demonstrated that this GWIBA approach enhanced the power of detecting loci with moderate single-locus effects; it also implied that some known susceptibility loci with moderate single-locus effects might be interacting with other loci. Moreover, we reasoned that interaction effects could increase the overall effects of loci that only showed marginal effects, and there were very few examples of large-effect common variants for common diseases [9]; therefore, we speculated that interactions of common variants may prefer to reside on loci with moderate to small single-locus effects. This hypothesis could explain a common phenomenon that there was seldom any significant interaction detected by the means of investigating interactions among loci with certain marginal effects after the single-locus analyses [1,4,5,8].

Our exhaustive searches revealed several two-locus associations, where both the individual loci exhibited relatively small single-locus effects. The most extreme case was the interaction between

rs11731175 and rs11236365 for T2D. Neither of the SNPs showed obvious marginal effect, but they exhibited an excessive interaction effect. This suggested that some interactions might be missed by using methods based on single-locus analysis or interaction-based approaches with non-exhaustive search strategies. On the other hand, we searched for genetic interactions associated with seven diseases and observed only one pair of loci that fell within that extreme situation. Theoretically, because the allele frequency of genetic loci often varies among different populations, it is relatively unlikely that marginal effects of the interactions will be obscured in all populations.

We have noticed that although some loci (or their good proxies) could not be replicated, the corresponding regions showed apparent association/interaction signals in the validation data. The signals were unlikely to be observed by chance given the local validation *p*-values. The replication failure for these loci was unlikely to simply result from insufficient statistical power; because unlike the replicated interaction for CD, we did not identify any consistency of the OR values between the datasets for these loci (data not shown), while the signals were observed. In addition, the signals were unlikely to be affected by genotyping artifacts, because multiple loci were considered and the data were initially quality-controlled. To our knowledge, tens of loci have been identified for some common diseases, but no interaction between exact loci has been confirmed in independent populations to date, despite of the fact that many of the loci are in the same pathway. Based on these observations, one plausible hypothesis is that the genetic heterogeneity may affect the consistency of the interactions. We speculated that many disease-causing interacting loci for common diseases might reside among rare variants that have large effects, and these rare variants could vary in frequency between populations, or they could be on adjacent, but distinct loci between populations. This could appropriately explain the lack of consistently observed interactions for common diseases in current GWASs that used common-variant markers.

The Need of a False Positive Control for Interaction Analyses

In this study, we found an overwhelming number of false positives, including artificial associations, in the raw results. The problem of false positives was more severe in our two-locus analyses than in the single-locus analyses, because our two-locus genotype combinations had insufficient sample sizes, which made them very sensitive to the artificial genotyping errors that were widely present in GWAS data. In addition, sparse data caused inaccuracy on asymptotic tests. Therefore, the results of two-locus analyses require careful examination, and particular attention must be paid to incredibly small *p*-values.

In the raw results with linked SNPs, we identified two kinds of artificial interactions; the batch effect (Figure 3) and the genotype clustering problem (Figure 4). Note that, although these kinds of observations were exaggerated by LD, and therefore, were previously considered as LD effects [8] (rs2532292), they were, in fact, caused by genotyping artifacts (Figure S5). Thus, interactions with unlinked SNPs, particularly SNPs with low MAFs, also require careful examination. In some previous studies, we found probable false positive results of the same kinds. For example, in two previous works [12,13], we conducted experimental searches on the WTCCC RA data without any quality control procedures; all the interactions that were outside the MHC region contained unqualified SNPs, according to the WTCCC study. Careful examination showed that many of these results were SNP pairs with linked SNPs, which were probably artificial associations of the two kinds mentioned in this study. Only one result was not affected by unqualified SNPs; but, when this was tested with quality-controlled samples, we observed a sharp decrease in significance. Moreover, a recent study [15] tested a new program on part of the WTCCC data and reported many interactions; however, almost all those results were interactions with linked SNPs that showed extremely significant *p*-values. We observed a large overlap between those reported SNPs and the SNPs that were filtered out in this study. In particular, two of the SNPs that were reported in that study (rs1065705 and rs1420247) were confirmed in this study to be affected by artificial genotyping errors (Table S1E). We also found that three regions, *PLXNA2*, *PTPRT*, and *PPM1A* that were reported to be "associated" with multiple diseases in the WTCCC data were extremely unlikely to be true interactions; in particular, we found that the *PPM1A* region, with the most significant *p*-value, was "associated" with all diseases except BD, and the association was probably a false positive. Therefore, we suggest that careful false positive control procedures should be adopted in future GWIBA studies to avoid misleading results and unnecessary endeavors in subsequent replication analyses.

Limitations

There are a few limitations of this study. First, although GWIBA permits agnostic searching without the need for prior biological knowledge, it loses substantial power due to the penalty introduced by multiple testing corrections for the huge number of potential pair-wise interactions. Therefore, candidate-gene methods should not be discarded, because they offer promising, well-powered detection of interactions based on biological knowledge. For example, a previous study performed a partial search on genes within certain biological networks and obtained some significant interactions [18]. Second, the contingency tables used for fast computing could not incorporate continuous covariates. However, these might be very important in some genetic analyses. This problem might be partially addressed by incorporating the covariates after an initial screen for interactions with a loose threshold. Third, we had to compromise for the huge computational issue by using general tests that assumed no specific genetic models; this resulted in decreased power compared to a test that conforms to a certain specific model. Furthermore, detection of high-order interactions was restricted to the conditional search, in order to conserve the computational time. Fourth, two-locus associations should be interpreted with caution when the single-locus effect of one SNP is very large; validation analyses should be performed to further confirm pure interaction effects. Fifth, the non-pseudoautosomal region of the X chromosome was not included in this analysis due to the imbalanced proportions of males and females between the case and control groups; however, many susceptibility loci of common diseases may reside on the X chromosome. This problem might be addressed by stratifying the contingency tables with a sex covariate, and then removing the corresponding female individuals with heterozygote genotypes for the tested SNPs on the X chromosome. Finally, this method provided inflated test statistics to detect SNPs with low MAFs, which were removed from the analysis. The removal may have caused us to miss low-frequency variants with relatively large effects, and these loci may be more valuable than common variants with smaller effects [37]. These issues require further studies to be fully addressed. Thus, we do not unreservedly recommend the approach used in this analysis for detecting genetic interactions. Rather, we recommend further improvements to this method, and the use of other methods when appropriate. Nevertheless, we would like to emphasize that the procedures described here are important for ensuring the reliability of interactions.

Computational Efficiency of PIAM for Future Large-Scale GWAS Datasets

We implemented PIAM with a multi-thread/parallel program, rapid tests for two-locus interactions, optional two-stage strategies for interaction searching, fast algorithms for collecting contingency

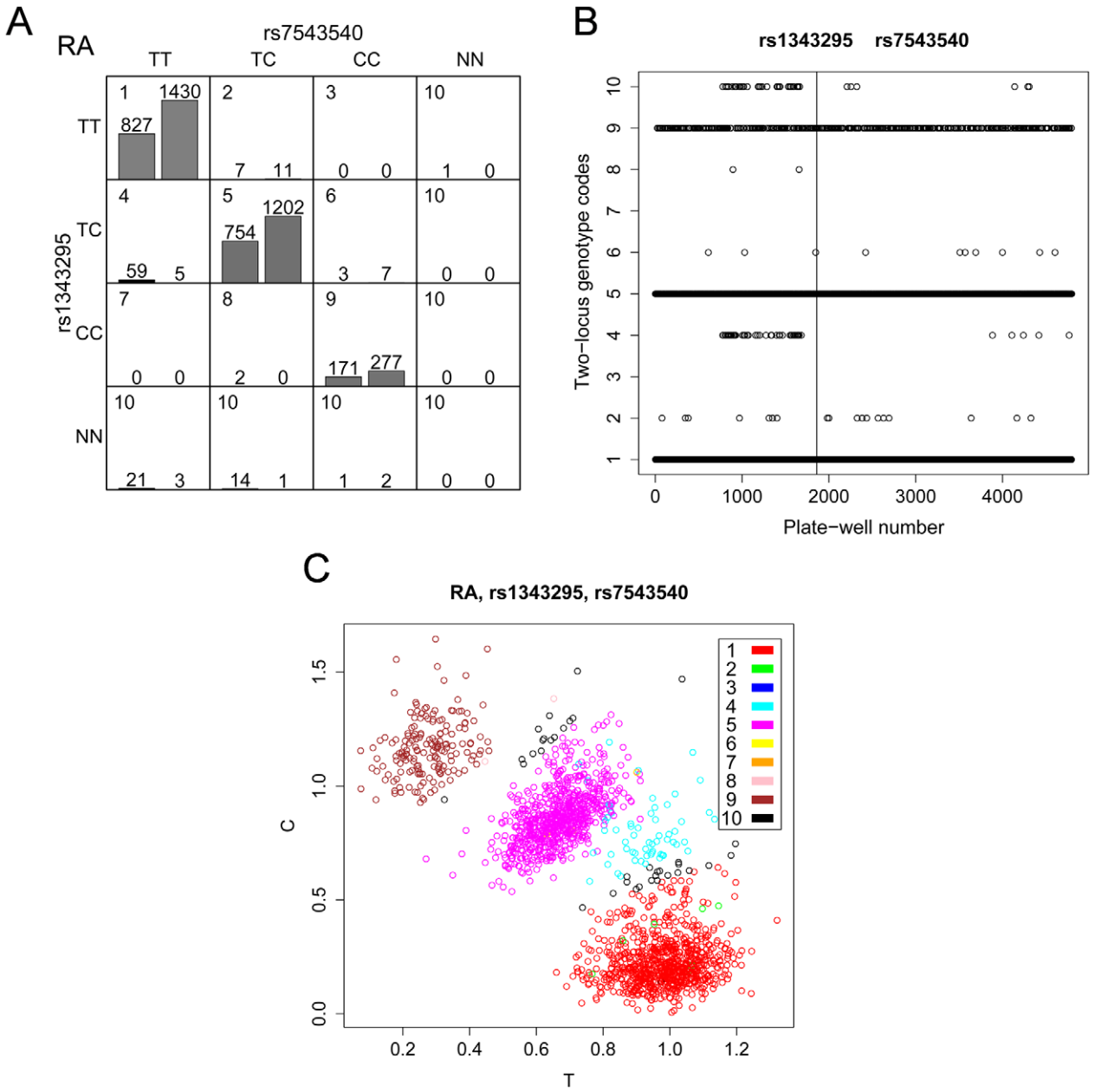

Figure 3. Batch effect observed for SNPs rs1343295 and rs7543540. (A) For each two-locus genotype combination, a genotype code is shown in the upper left corner of each cell. NN denotes missing genotypes. The distribution of RA cases (left bar) and controls (right bar) in each genotype combination is shown with the number of observations indicated above the bars. The samples are mainly distributed on the diagonal of the genotype combinations, where two SNPs are in LD. Note that many genotype combinations are sparse. An excessive number of cases relative to controls was observed for the genotype combination TC for rs1343295 and TT for rs7543540 (code 4), which primarily caused the association. (B) Genotype combination codes (1–10) of samples were plotted against the plate and well numbers of samples in 96-well plates. Codes 1–9 denote the nine non-missing genotypes shown in (A). Samples with missing genotypes were grouped in code 10. The vertical line separates cases (left) and controls (right). The 59 cases of one particular genotype combination (code 4) were not evenly distributed among the wells, but severely aggregated. (C) Cluster plot for RA cases. The coordinates denote the allele intensities of the first SNP in the title (rs1343295) and the 10 colors denote the 10 genotype combinations of the two SNPs. The genotype clustering of 59 cases (plotted in cyan circles) are ambiguous between heterozygotes and homozygotes for rs1343295, and genotypes were considered heterozygotes. In fact, the genotypes of these 59 cases should probably be considered homozygotes, and then no association would exist; however, the batch effect produced this artificial error due to the low-quality genotyping and subsequent artificial clustering.
doi:10.1371/journal.pgen.1001338.g003

tables with a binary genotype coding method [38], and an intrinsic CPU instruction for new types of CPUs. These components made PIAM capable of handling very large GWAS datasets that are anticipated to be commonly available in the future. For example, the WTCCC2 study will include much larger numbers of SNP markers and sample sizes for the identification of susceptibility loci with moderate single-locus effects and interactions. We estimated that, for a dataset with up to 1,000,000 SNPs and 10,000 samples, PIAM could complete an exhaustive, two-locus search within 6 days with one computer equipped with a modern quad-core, 3.0 GHz, desktop CPU and 4 G of memory; this speed could be multiplied with parallel computing on multiple computers.

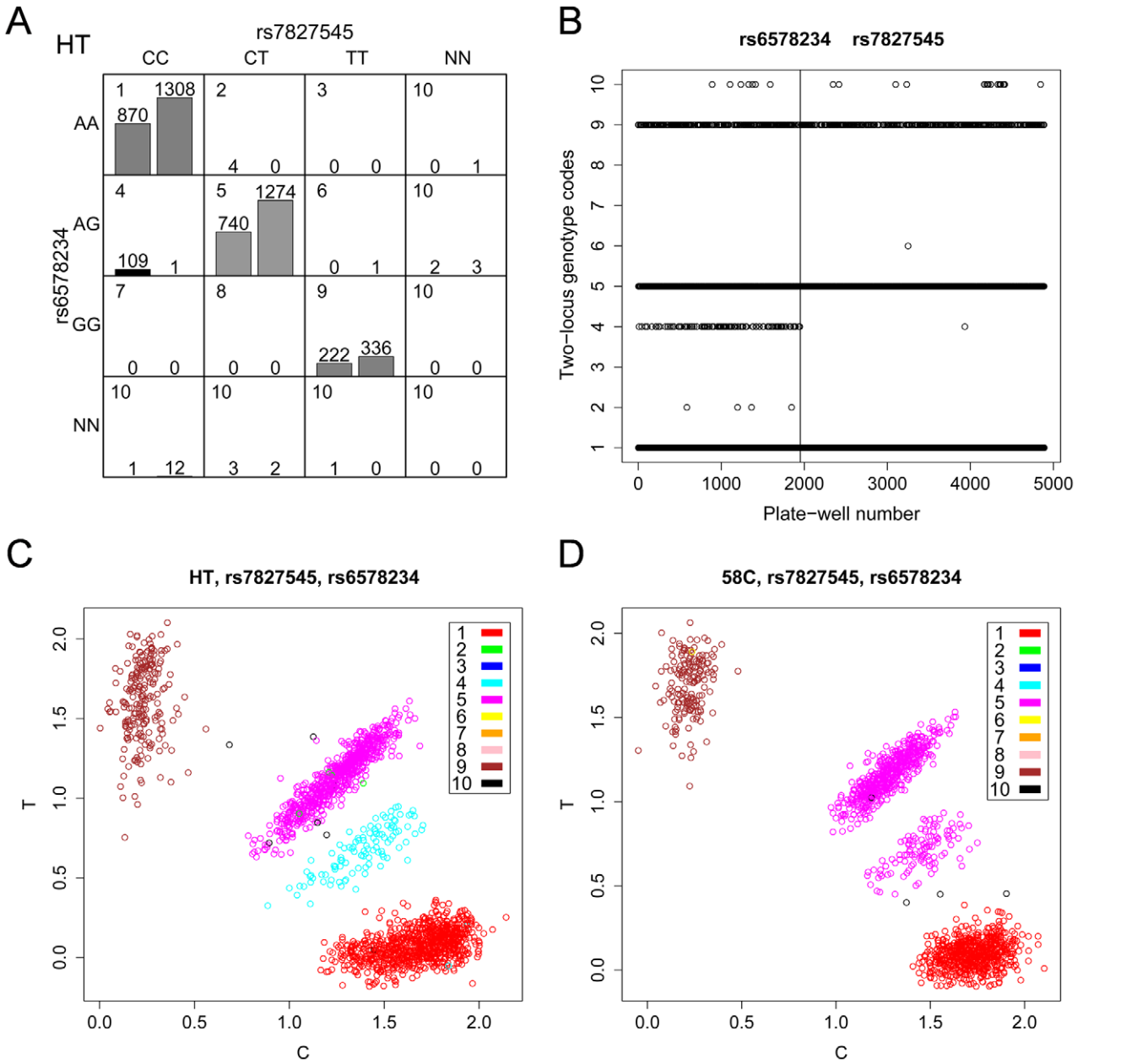

Figure 4. Genotype clustering problem observed for rs6578234 and rs7827545. The legend to this figure is the same as that for Figure 3, except that the associated disease is HT, and the population was divided into two cohorts (C and D). The description of (D) is the same as that for (C), except that the 1958 birth cohort control samples were plotted in (D). (A) The situation here is similar to that shown in Figure 3A, with an excessive number of cases compared to control; in the genotype combination of AG for rs6578234 and CC for rs7827545 (code 4), 109 cases and only 1 control were observed. However, the batch effect did not occur, because all the plates showed association signals for this combination (B, genotype combination 4). Instead, a fourth cluster is observed in both case and control groups (C and D); the pattern in the two groups is the same, but the genotypes were assigned different codes for cases and controls by the genotype calling algorithm CHIAMO. The genotypes of rs7827545 of the 109 HT cases were called homozygotes (C, indicated in cyan, code 4), and the genotypes in the controls were called heterozygotes (D, indicated in pink, code 5). In fact, the genotypes of rs7827545 of the 109 cases should probably be considered heterozygotes, as in the controls, and no association should exist. We do not know the cause of the fourth genotype cluster, because the clustering pattern could not be explained by multiple clusters of copy number variation.
doi:10.1371/journal.pgen.1001338.g004

Prospective

This GWIBA approach can be used routinely, in addition to single-locus analyses in future genome-wide association studies. It is a promising approach for the discovery of novel loci with interaction effects, which may provide important insights into common diseases. By combining various approaches, we could greatly accelerate the discovery of the genetic architecture of common diseases.

Materials and Methods

The WTCCC Data

The initial data were obtained from the WTCCC (http://www.wtccc.org.uk/). This dataset comprised ~2,000 samples each for seven diseases (BD, CAD, CD, HT, RA, T1D, and T2D). For half of these samples, there were ~3,000 shared control samples from

the 1958 birth cohort (58C); for the other half, there were control samples from the National Blood Service (NBS). Genotyping was performed with the Affymetrix GeneChip Mapping 500K Array Set. Genotypes were called with the CHIAMO algorithm with the parameter "posterior probability less than 0.9" set to "missing". The non-pseudoautosomal SNPs on the X chromosome were not included, because the general genotypic test was used in PIAM, and the male/female proportion was imbalanced between the cases and controls. Quality control for the samples and the SNPs was performed as described in the WTCCC. In addition, we excluded SNPs that were significant in the single-locus association analysis, but showed poor clustering according to the WTCCC. After trimming, 459,075 SNPs remained for each disease, and the corresponding data were used as input for the two-locus exhaustive searches.

Pair-Wise Interaction-Based Association Mapping (PIAM)

We developed a fast, multi-thread/parallel program named "pair-wise interaction-based association mapping" (PIAM, available at http://www.ihs.ac.cn/xykong/PIAM.zip) to search for susceptibility SNPs with interaction effects in a set of genome-wide SNPs. PIAM is based on a two-locus logistic regression model and the likelihood ratio test (LRT). For the logistic regression model, the additive effect of a SNP was represented with a variable that was coded 0, 1, and 2 for homozygote, heterozygote, and the other homozygote (e.g., AA, AB, BB), respectively. We added another variable for the heterozygote effect that was coded 1 for heterozygote and 0 otherwise. Therefore, two variables were used for the general effects of one SNP. The interaction was modeled by the multiplication of variables between SNPs; thus, four terms were used for each pair-wise interaction. The interactions can be interpreted by their deviation from the restricted model without the interaction terms. The restricted model only considers the additive effect between the two loci on the log odds ratios, that is, the multiplicative effect between the two loci on the odds ratios. The full two-locus logistic model considers all possible effects of the two loci. Accordingly, the deviation between the two indicates the significance (or relevance) of the interaction.

A previous study proposed the use of a full, two-locus, logistic regression model and evaluated its statistical power [10]. However, when all the SNP pairs were tested with the LRT, and the null model (with only one intercept term) was compared to the full model of two SNPs (with an intercept term and eight terms for all the effects of two SNPs and their interactions), as previously proposed, there were excessive results associated with the single-locus effect of a single SNP. Therefore, for practical use, we modified the previous approach with the following strategy. First, a single-locus LRT for the general effect of each SNP was performed. Then, a family-wise-significant, single-locus, p-value threshold was used to divide the whole set of SNPs into two subgroups. One subgroup was small and significant (subset A with n SNPs) and the other subgroup comprised the remainder of SNPs (subset B with m SNPs). Then, we performed three types of searches:

(1) The epistatic search. For each of the $n(n-1)/2$ combinations of SNP pairs within subset A, we used the 4 *d.f.* LRT of the logistic regression model comparison (a restricted model without four interaction terms compared to the full model of two loci) to test for a pure interaction effect.

(2) The conditional search. For each of the $n \times m$ SNP pairs between subsets A and B, we used a 6 *d.f.* LRT to test for marginal and interaction effects of the SNP in subset B that were conditional on the presence of the SNP in subset A. For this, a submodel was compared to the full model (the submodel contained two terms for the SNP in subset A and one intercept term).

(3) The simultaneous search. For each of the $m(m-1)/2$ SNP pairs within subset B, we used an 8 *d.f.* LRT to test for all the effects of the combination of two SNPs. For this, we compared the null model to the full model.

For the conditional and simultaneous searches, the LRT statistics were calculated by the G^2 test with contingency tables for fast computing. This method was equivalent to the LRT for the logistic regression models, but it did not estimate the parameters. In addition, the genotypes were transformed to a set of binary values to accelerate the collection of contingency tables, as proposed by a recent study [38]. For the simultaneous search, we also implemented in PIAM the previously proposed two-stage strategy [10,39]. The Bonferroni correction was used for N multiple tests, where N was the total number of tests for all search situations. Missing genotypes were addressed by removing the corresponding individual. After the exhaustive two-locus search, the conditional search was extended to high-order interactions. For example, conditional on an existing two-locus interaction, the full two-locus model was compared to a full three-locus model by adding another locus; this resulted in an 18 *d.f.* LRT test.

Approximate Statistical Distribution

The huge number of statistics (up to 1×10^{11}) generated in this study would be extremely computationally demanding to handle directly. Therefore, to check the overall distributions of the observed two-locus test statistics, we implemented the approximate statistical distribution method in PIAM. First, very small, continuous intervals (e.g., 0.001 in length), were predefined for the LRT statistics; a maximum value of the statistic was set to be that with a corresponding p-value equal to $0.01/L$, where L was the total number of comparisons; thus, the last interval was the maximum value to infinity. During the computation, PIAM recorded the number of statistics within each small interval, rather than the exact value of the statistic. When applying the statistics to the quantile-quantile plot, the statistics were treated as equal to the lower bound of the corresponding interval; therefore, the error of the statistic was controlled below 0.001. This method can also be used to handle p-values transformed by a negative logarithm. The approximate statistic distributions can further be used in various multiple test correction approaches. This method can, and should, be adopted in other GWIBA studies in the future to obtain the overall statistical distributions, similar to the single-locus analyses.

Additional SNP Quality Control

We found that the initial SNP quality control performed by the WTCCC was not sufficiently stringent for our interaction analysis. That control yielded an abundance of extremely significant interaction results, but these were subsequently identified as false positives, due to sparse data and/or poor genotyping quality. The sparse data were introduced by comparing the two-locus genotypes for the interaction analysis to the single-locus analysis. These relatively sparse data were more sensitive to genotyping errors. To avoid that problem, an additional, more stringent SNP quality control was applied. We removed any SNPs with a missing data rate that was $>2\%$ of the cases or controls, with a MAF<0.1 in the controls, or with a HWE p-value<0.001 in the controls. The removed SNPs with high missing data rates typically showed poor genotype clustering; the low frequency SNPs often yielded an inflated two-locus LRT statistic, due to sparse data (at least one expected cell count <5 in the two-locus contingency table for

current sample size, assuming HWE for both SNPs with either or both MAFs$<$0.1); and a deviation from HWE in the control population was probably due to genotyping errors. After applying this additional SNP quality control, the test numbers changed for the Bonferroni correction, due to the removal of SNPs.

Expanded Control Analysis

To improve the detection of true interactions that may not initially achieve significance within the cases and shared controls, we applied the expanded control analysis, as performed in the WTCCC study. The enlarged, "expanded control" was used for each disease to test pairs of loci with interactions that passed moderate screening p-value thresholds ($50/L$ in this study, where L was the total number of two-locus combinations) in the initial analysis with the cases and shared controls, but did not necessarily achieve significance (i.e., a Bonferroni corrected $P<0.05$). In the expanded control analysis, the final statistical significance was evaluated. The expanded control for a certain disease was the combination of the shared control and some other disease cohorts. For BD, the expanded control group included the shared control plus the CAD, CD, HT, T1D, and T2D groups. For the three autoimmune diseases (CD, RA, and T1D), the expanded control included the shared control plus all other disease cohorts, except the autoimmune disease cohorts. The same was true for the three metabolism-related diseases CAD, HT, and T2D. These expanded controls were the same as those used in the WTCCC study. Note that, associations caused by diseases other than the disease of interest could be avoided in this expanded control analysis, because the first stage screening with the shared control required a low p-value.

The expanded control analysis was not an independent replication of the initial analysis; therefore, a genome-wide multiple test correction should also be used when testing the interactions retained in the initial analysis. For convenience, we used the same test numbers for correction in the expanded control analysis as those used in the initial analysis. That is, the two SNP subsets (subsets A and B) were the same as those in the initial analysis; therefore, the subset division was not based on the single-locus p-value of the expanded control. This is similar to the "joint analysis" strategy [40] of single-locus analyses in GWAS. However, an additional problem we encountered was the possibility that some SNPs in subset B might not pass the 5×10^{-7} p-value threshold in the initial analysis, but could pass it in the expanded control analysis. These SNPs had to be removed to avoid associations that were caused by a single-locus effect only. An alternative strategy could be to determine different subsets A and B for the expanded control analysis. These would be chosen according to the single-locus p-value threshold of the expanded control. Then, the corresponding search situation and the appropriate numbers of multiple tests would be used in the expanded control analysis.

Result Filtering

First, some interactions detected by tests that incorporate marginal effects may result from marginal effects alone, without any pure interaction effects, and we used a strategy similar to BEAM (the hierarchical significance declaration procedure) [19] to address this problem. We compared two-locus p-values with single-locus p-values, as follows. For SNP pairs obtained in the simultaneous search, we compared the corrected two-locus p-value to the more significant corrected single-locus p-value of the two SNPs; for SNP pairs obtained the conditional search, we compared the corrected two-locus association p-value to the corrected single-locus p-value of the SNP in subset B; we removed SNP pairs that had two-locus p-values that were less significant than the single-locus p-values. The p-values in the expanded control analysis were used for these comparisons. In addition, we also removed SNP pairs with any SNPs that did not pass the 5×10^{-7} p-value threshold in the initial analysis, but passed the threshold in the expanded control analysis.

Second, we examined all SNP pairs that were located within 1 Mb of each other. Two kinds of artificial associations were found; one was a batch effect and the other was a genotype clustering problem. The batch effect was severe aggregation of samples of some individuals with particularly high risk, two-locus genotypes, in the 96-well plates. The genotype clustering problem was observed on genotype clustering plots; this manifested as an ambiguous extra cluster (beyond the normal three clusters) that the genotype calling algorithm classified differently between the case and control groups. SNP pairs with either of these problems were removed from the analyses.

Third, we further checked the regional interaction signals to avoid artificial associations due to errors in genotyping a given SNP. Only results with consecutive interaction signals were retained; i.e., an elevated interaction signal could be observed on at least two nearby SNPs from both regions. No results were excluded based on this check in this study.

Validation Analysis of CAD

We used the online analysis results of the German MI Family Study [2] (http://www.cardiogenics.imbs-luebeck.de/) to test for allelic effects in order to validate the pair of regions that we had associated with CAD. We did not have access to the individual-level genotype data from that study to validate the interaction. The genotyping platform was the Affymetrix GeneChip Mapping 500K Array Set. The SNPs were quality-controlled; only SNPs with a trend test $P<0.001$ were shown on the website. We searched the website for any SNPs that showed significant single-locus effects within 50 kb of the loci. Because only SNPs with trend p-values<0.001 were shown, we could not check the regional signals or validate the interaction, due to the lack of individual-level genotypes.

Validation Analysis of CD

The NIDDK IBDGC data (phs000130.v1.p1) [3] was accessed from the National Center for Biotechnology Information (NCBI) database of genotypes and phenotypes [41] (dbGaP, http://www.ncbi.nlm.nih.gov/dbgap/) to validate the interactions for CD. The dataset was stratified into two populations, the non-Jewish population stratum, which comprised 513 cases and 515 controls, and the Jewish population stratum, which comprised 300 cases and 432 controls. The genotyping platform was the Illumina HumanHap300 Genotyping BeadChip. The SNPs in the association result file (pha002847.1.IBD.analysis.tar.gz) were selected by removing SNPs with call rates <0.9 in cases or controls and SNPs with HWE p-values<0.001 in the controls. Thus, a total of 305,345 SNPs was used as the validation SNP set.

Two subsequent strategies were used for this validation analysis: the proxy replication strategy and the local validation strategy. First, proxy replication was implemented; because a different genotyping platform was used for the IBDGC data compared to the WTCCC data. SNPs in LD with the original SNPs were selected for proxy replication. The measurement of LD was based on the r^2 values from the CEU population data (Phase III release #2) of the International HapMap Project [26] (HapMap, http://www.hapmap.org/). The MaCH imputation method [42] was then used to impute ungenotyped SNPs for validations between the WTCCC and the IBDGC datasets. Interactions were

considered valid when they could be replicated by proxy SNPs. The reference haplotypes for MaCH were obtained from the HapMap CEU population (Phase III release #2).

Second, "local validation" was used when the locus-based replication (e.g., proxy replication) failed. In this local validation method, all SNPs in the validation dataset within 50 kb of the original loci were tested for allelic effects (by the trend test) or pure interaction effects (by the 4 *d.f.* LRT test). This method was based on the notion that confounding factors might affect the consistency of the interactions between the original data and the validation data. The significance was evaluated under the null hypothesis that none of the SNPs (interactions) in the 100 kb region (pair of 100 kb regions) was associated with the disease. In addition, the *p*-values of the SNPs (interactions) should form a uniform distribution for SNPs that were independent; moreover, the number of *p*-values lower than a certain threshold (we used 0.05) from the total number of *p*-values should form a binomial distribution. Therefore, a single-tailed binomial test could be performed to determine the significance in the numbers of SNPs with *p*-values lower than the threshold. However, SNPs were not independent, due to the LD. Thus, to obtain the empirical significance for the tests, we randomly sampled 1,000 pairs of 100 kb regions from this validation dataset and calculated the empirical distribution of the *p*-values for correction (Table S5). A significant local validation was interpreted to reject the null hypothesis, that none of the SNPs or interactions in a certain region was associated with the disease. However, strictly speaking, this cannot be interpreted as a successful replication of the original association or interaction; thus, we used the term "local validation" instead of "local replication" to avoid confusion.

Validation Analysis of T2D

The GENEVA Diabetes Study data (phs000091.v1.p1) was accessed from the dbGaP to validate the T2D association results. The participants in that study were all female. The genotyping platform was the Affymetrix Genome-Wide Human SNP Array 6.0. Caucasian individuals without missing data on the disease status were included, SNPs were quality-controlled, and 496,606 genotypes were set to "missing", as initially recommended. After selection, a total of 1,543 cases and 1,770 controls, with 707,301 SNPs were analyzed. According to the genotyping platforms, the SNPs in the GENEVA Diabetes Study dataset contained most of the SNPs in the WTCCC data. Therefore, we could select the exact SNP combinations in the validation dataset for exact replication, without the need for the proxy replication described above. Upon failure of the exact replication, the local validation method was used with this dataset.

Supporting Information

Figure S1 The distributions of two-locus statistics represented in quantile-quantile plots. Quantile-quantile plots were generated for all two-locus LRT statistics in the simultaneous search (A), in the conditional search (B), and in the epistatic search (C). The LRT statistics (*y* coordinates) were sorted and plotted in black circles against the expected based on the null hypothesis (*x* coordinates); the shaded regions show the 95% concentration bands, and the dashed lines indicate the expected distributions. Statistics that resulted in *p*-values$<0.01/L$ are shown in triangles, (*L* is the total number of comparisons for the corresponding search situation). There were no available conditional statistic plots for HT or epistatic statistic plots for BD and HT. For the simultaneous and conditional searches (A and B, respectively), the two-locus statistical distributions for BD, CAD, HT, and T2D fit the expected quite well; the distributions for the three autoimmune diseases CD, RA, and T1D showed moderate overdispersion; the statistical deviations for CD, RA, and T1D started suddenly from the middle of the dotted lines, and therefore, they did not reflect general overdispersion. This was due to the many single-locus associated SNPs for these three diseases, including SNPs in the MHC region for RA and T1D and multiple associated regions for CD. Therefore, we applied a strategy similar to BEAM to control for the excessive single-locus effects in the two-locus associations (described in Materials and Methods). The statistics for the epistatic search did not present overdispersion, except for the RA and T1D data; this was also due to the many significant SNP pairs within the MHC region. Note that the statistics of artificial associations identified in this study were not removed from these plots, which resulted in the extreme deviations in the tails, particularly in (A).
Found at: doi:10.1371/journal.pgen.1001338.s001 (0.62 MB PNG)

Figure S2 Regional signal plots of the interaction between linked SNPs. The format of this figure is the same as that described in Figure 1.
Found at: doi:10.1371/journal.pgen.1001338.s002 (0.30 MB PNG)

Figure S3 Cluster plots of the SNPs in Table 1 and one other pair of linked SNPs. The three genotypes are indicated in red, green, and blue circles; the black "+" denotes missing genotypes.
Found at: doi:10.1371/journal.pgen.1001338.s003 (1.66 MB JPG)

Figure S4 Linkage disequilibrium plots for the regions in Table 1. One row represents one pair of regions.
Found at: doi:10.1371/journal.pgen.1001338.s004 (0.97 MB JPG)

Figure S5 Genotype clustering problem of rs2532292. The legend to this figure is the same as that of Figure 3. (A) The interaction was yielded by the genotype combination coded as "4", with only a modest effect size; this interaction was detected because BEAM was sensitive to low frequency variants. (B) The batch effect did not exist. (C) The cluster plot of rs2532292 in the cases showed that the four cases with the genotype combination "4" (in cyan) were distributed on the lower edge of the heterozygote cluster, rather than sporadically distributed. Therefore, the rs2532292 genotypes for these four cases should be probably the common homozygotes, and it was the same for the four controls with the genotype combination "4" (data not shown).
Found at: doi:10.1371/journal.pgen.1001338.s005 (0.74 MB PNG)

Table S1 Original results. The table file is called "Table S1.xls", and it is compressed in the zip file. (A) Raw results from PIAM. This table shows the results generated by PIAM, with the corresponding disease and the additional SNP quality control codes in the first and last columns, respectively. The quality control code "1" denotes an unqualified SNP pair, which includes at least one SNP that failed the additional quality control; the quality control code "0" denotes all others. (B) Results excluded from RA and T1D searches. These results were excluded because both SNPs were within the MHC region. The format of this table is the same as that described in (A). (C) Results that passed the additional SNP quality control in (A) and were tested with the expanded controls. The format of this table is similar to that described in (A), with additional columns for information from the Affymetrix annotations (columns D-O), test numbers (column AR), interaction LRT statistics and *p*-values (AW and AX), and expanded control analysis results (AY-BK). (D) Results that passed the *p*-value threshold of Bonferroni corrected $P<0.5$ in the

expanded control analysis (column BI). Filter codes (column BL): 1, a corrected, two-locus p-value that was less significant than the corrected single-locus p-value of either SNP (for the simultaneous search) or the SNP in subset B (for the conditional search); 2, single-locus p-value that exceeded the 5×10^{-7} threshold in the expanded control analysis; 3, SNP pairs that were located within 1 Mb of each other; 0, results that passed filters 1–3. (E) Results that failed filter 3. This format of this table is the same as that described in (D). The results masked in dark grey were false positives due to the batch effect or genotype clustering problem; others were not affected. Results highlighted in yellow were the nearest SNP pairs selected in each associated region. (F) Results that passed filters 1–3. The format of this table is the same as that described in (D). Results highlighted in yellow were the SNP pairs that gave the most significant two-locus p-value within each pair of associated regions in the initial search.
Found at: doi:10.1371/journal.pgen.1001338.s006 (4.73 MB ZIP)

Table S2 Numbers of multiple tests. The total numbers of SNPs and the multiple tests used for the Bonferroni correction. Additional QC: the additional SNP quality control.
Found at: doi:10.1371/journal.pgen.1001338.s007 (0.03 MB DOC)

Table S3 OR tables for the interaction between linked SNPs. The legend to this table is the same as that of Table 3.
Found at: doi:10.1371/journal.pgen.1001338.s008 (0.03 MB DOC)

Table S4 OR tables for the interaction between rs296533 and rs2089509 with the IBDGC non-Jewish population, Jewish population, and the imputed WTCCC data. The legend to this table is the same as that of Table 3.
Found at: doi:10.1371/journal.pgen.1001338.s009 (0.06 MB DOC)

Table S5 Local validation tests. (A) Detailed results of the local validation tests. (B) Sampling p-values with the IBDGC non-Jewish population data. (C) Sampling p-values with the GENEVA T2D data.
Found at: doi:10.1371/journal.pgen.1001338.s010 (0.21 MB XLS)

Acknowledgments

We are grateful to the anonymous reviewers whose comments and suggestions contributed to the significant improvement of this paper. We acknowledge the contributing investigators of the WTCCC study, the German MI Family Study, the NIDDK IBDGC study, and the GENEVA Diabetes Study for generating and providing the data for us. The data of the NIDDK IBDGC study (phs000130.v1.p1) and the GENEVA Diabetes Study (phs000091.v1.p1) were obtained from dbGaP. We thank Bo Liu from the Shanghai Supercomputer Center for facilitating the use of the Dawning 4000 A computer clusters. We also thank Liangliang Zhang and Zhongping Xu for making this manuscript easier to understand for common biological researchers.

Author Contributions

Conceived and designed the experiments: Y Liu, X Kong, G-P Zhao. Analyzed the data: Y Liu. Contributed reagents/materials/analysis tools: Y Liu, H Xu, S Chen, X Chen, Z Zhang, Z Zhu, X Qin, L Hu, J Zhu. Wrote the paper: Y Liu. Interpreted the results: Y Liu, X Kong, L Hu.

References

1. The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661–678.
2. Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, et al. (2007) Genomewide association analysis of coronary artery disease. N Engl J Med 357: 443–453.
3. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, et al. (2006) A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science 314: 1461–1463.
4. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, et al. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 40: 955–962.
5. Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, et al. (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet 41: 703–707.
6. Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, et al. (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 40: 638–645.
7. Williams SM, Canter JA, Crawford DC, Moore JH, Ritchie MD, et al. (2007) Problems with genome-wide association studies. Science 316: 1840–1842.
8. Cordell HJ (2009) Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 10: 392–404.
9. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747–753.
10. Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 37: 413–417.
11. Gayan J, Gonzalez-Perez A, Bermudo F, Saez ME, Royo JL, et al. (2008) A method for detecting epistasis in genome-wide studies using case-control multi-locus association analysis. BMC Genomics 9: 360.
12. Wan X, Yang C, Yang Q, Xue H, Tang NL, et al. (2009) MegaSNPHunter: a learning approach to detect disease predisposition SNPs and high level interactions in genome wide association study. BMC Bioinformatics 10: 13.
13. Yang C, He Z, Wan X, Yang Q, Xue H, et al. (2009) SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. Bioinformatics 25: 504–511.
14. Tang W, Wu X, Jiang R, Li Y (2009) Epistatic module detection for case-control studies: a Bayesian model with a Gibbs sampling strategy. PLoS Genet 5: e1000464. doi:10.1371/journal.pgen.1000464.
15. Wan X, Yang C, Yang Q, Xue H, Tang NL, et al. (2010) Predictive rule inference for epistatic interaction detection in genome-wide association studies. Bioinformatics 26: 30–37.
16. Yang C, Wan X, Yang Q, Xue H, Yu W (2010) Identifying main effects and epistatic interactions from large-scale SNP data via adaptive group Lasso. BMC Bioinformatics 11 Suppl 1: S18.
17. Wongseree W, Assawamakin A, Piroonratana T, Sinsomros S, Limwongse C, et al. (2009) Detecting purely epistatic multi-locus interactions by an omnibus permutation test on ensembles of two-locus analyses. BMC Bioinformatics 10: 294.
18. Emily M, Mailund T, Hein J, Schauser L, Schierup MH (2009) Using biological networks to search for interacting loci in genome-wide association studies. Eur J Hum Genet 17: 1231–1240.
19. Zhang Y, Liu JS (2007) Bayesian inference of epistatic interactions in case-control studies. Nat Genet 39: 1167–1173.
20. Sklar P, Smoller JW, Fan J, Ferreira MA, Perlis RH, et al. (2008) Whole-genome association study of bipolar disorder. Mol Psychiatry 13: 558–569.
21. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263–265.
22. Mislow JM, Kim MS, Davis DB, McNally EM (2002) Myne-1, a spectrin repeat transmembrane protein of the myocyte inner nuclear membrane, interacts with lamin A/C. J Cell Sci 115: 61–70.
23. Zhang Q, Bethmann C, Worth NF, Davies JD, Wasner C, et al. (2007) Nesprin-1 and -2 are involved in the pathogenesis of Emery Dreifuss muscular dystrophy and are critical for nuclear envelope integrity. Hum Mol Genet 16: 2816–2833.
24. Durante W, Liao L, Peyton KJ, Schafer AI (1998) Thrombin stimulates vascular smooth muscle cell polyamine synthesis by inducing cationic amino acid transporter and ornithine decarboxylase gene expression. Circ Res 83: 217–223.
25. Durante W, Liao L, Peyton KJ, Schafer AI (1997) Lysophosphatidylcholine regulates cationic amino acid transport and metabolism in vascular smooth muscle cells. Role in polyamine biosynthesis. J Biol Chem 272: 30154–30159.
26. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437: 1299–1320.
27. Tomlinson MG, Kane LP, Su J, Kadlecek TA, Mollenauer MN, et al. (2004) Expression and function of Tec, Itk, and Btk in lymphocytes: evidence for a unique role for Tec. Mol Cell Biol 24: 2455–2466.
28. Sasaki A, Taketomi T, Kato R, Saeki K, Nonami A, et al. (2003) Mammalian Sprouty4 suppresses Ras-independent ERK activation by binding to Raf1. Nat Cell Biol 5: 427–432.
29. Scaldaferri F, Vetrano S, Sans M, Arena V, Straface G, et al. (2009) VEGF-A links angiogenesis and inflammation in inflammatory bowel disease pathogenesis. Gastroenterology 136: 585–595.
30. Grube M, Kock K, Karner S, Reuther S, Ritter CA, et al. (2006) Modification of OATP2B1-mediated transport by steroid hormones. Mol Pharmacol 70: 1735–1741.

31. Coleman DL, Leiter EH, Schwizer RW (1982) Therapeutic effects of dehydroepiandrosterone (DHEA) in diabetic mice. Diabetes 31: 830–833.
32. Coleman DL, Schwizer RW, Leiter EH (1984) Effect of genetic background on the therapeutic effects of dehydroepiandrosterone (DHEA) in diabetes-obesity mutants and in aged normal mice. Diabetes 33: 26–32.
33. Ladriere L, Laghmich A, Malaisse-Lagae F, Malaisse WJ (1997) Effect of dehydroepiandrosterone in hereditarily diabetic rats. Cell Biochem Funct 15: 287–292.
34. Kimura M, Tanaka S, Yamada Y, Kiuchi Y, Yamakawa T, et al. (1998) Dehydroepiandrosterone decreases serum tumor necrosis factor-alpha and restores insulin sensitivity: independent effect from secondary weight reduction in genetically obese Zucker fatty rats. Endocrinology 139: 3249–3253.
35. Mukasa K, Kanesiro M, Aoki K, Okamura J, Saito T, et al. (1998) Dehydroepiandrosterone (DHEA) ameliorates the insulin sensitivity in older rats. J Steroid Biochem Mol Biol 67: 355–358.
36. Lasco A, Frisina N, Morabito N, Gaudio A, Morini E, et al. (2001) Metabolic effects of dehydroepiandrosterone replacement therapy in postmenopausal women. Eur J Endocrinol 145: 457–461.
37. Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. Science 322: 881–888.
38. Wan X, Yang C, Yang Q, Xue H, Fan X, et al. (2010) BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. Am J Hum Genet 87: 325–340.
39. Evans DM, Marchini J, Morris AP, Cardon LR (2006) Two-stage two-locus models in genome-wide association. PLoS Genet 2: e157. doi:10.1371/journal.pgen.0020157.
40. Skol AD, Scott LJ, Abecasis GR, Boehnke M (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38: 209–213.
41. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, et al. (2007) The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 39: 1181–1186.
42. Li Y, Willer C, Sanna S, Abecasis G (2009) Genotype imputation. Annu Rev Genomics Hum Genet 10: 387–406.