Published in final edited form as:
Hum Genet. 2010 September ; 128(3): 269–280. doi:10.1007/s00439-010-0849-9.

Genome-wide searching of rare genetic variants in WTCCC data

Tao Feng and **Xiaofeng Zhu**
Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio 44106

Abstract

Although they have demonstrated success in searching for common variants for complex diseases, Genome-Wide Association (GWA) studies are less successful in detecting rare genetic variants because of the poor statistical power of most of current methods. We developed a two-stage method that can apply to GWA studies for detecting rare variants. Here we report the results of applying this two-stage method to the Wellcome Trust Case Control Consortium (WTCCC) dataset that include 7 complex diseases: Bipolar disorder, Cardiovascular disease, Hypertension, Rheumatoid Arthritis, Crohn's disease, Type 1 Diabetes and Type 2 Diabetes. We identified 24 genes or regions that reach genome wide significance. 8 of them are novel and were not reported in the WTCCC study. The cumulative risk (or protective) haplotype frequency for each of the 8 genes or regions is small, being at most 11%. For each of the novel genes, the risk (or protective) haplotype set cannot be tagged by the common SNPs available in chips (r^2<0.32). The gene identified in hypertension was further replicated in the Framingham Heart Study (FHS), and is also significantly associated with Type 2 Diabetes. Our analysis suggests that searching for rare genetic variants is feasible in current genome-wide association studies and candidate gene studies, and the results can severe as guides to future resequencing studies to identify the underlying rare functional variants.

Introduction

Despite the success of GWAS in searching for the common variants contributing to complex diseases in recent years, the identified common variants are responsible for only a small fraction of the phenotypic variation (Levy et al. 2009; Newton-Cheh et al. 2009; Visscher 2008). It has been suggested that it is time to shift from searching for common variants of modest effect to rarer variants of large effect by effectively searching the full genome(Goldstein 2009). Rare variants may hold the promise for the prediction of individual risk and personalized medicine because of their large effect, although it has been argued that common variants illuminate the biologic pathways of underlying diseases(Hirschhorn 2009). Large sample based on resequencing studies with carefully selected designs are usually necessary to detect the rare variants(Cohen et al. 2004; Ji et al. 2008). Such studies are greatly welcomed but are still tremendously expensive when searching is on the full genome scale. Several statistical methods have been developed and these methods mainly focus on when resequencing data are available (Cohen et al. 2004; Li and Leal 2008; Madsen and Browning 2009). Our simulation study suggests that searching for rare variants is possible and efficient using current GWA study designs(Zhu et al. 2010)

Corresponding author: Xiaofeng Zhu, PH.D, Department of Epidemiology and Biostatistics, Case Western Reserve University, Wolstein Research Building, 2103 Cornell Road, Cleveland, Ohio 44106, Tel: (216) 368 0201, Fax: (216) 368 4880, xzhu1@darwin.case.edu.

COMPETING INTERESTS STATEMENT
The authors declare that they have no competing financial interests.

by clustering the haplotypes in each gene according to disease risk. Since many GWA studies have been conducted or are ongoing, the results based on haplotype analysis to detect rare variants can be a future guide for in-depth resequencing studies.

The WTCCC study was the first successful large comprehensive GWA study which includes 7 complex diseases: Bipolar disorder, Cardiovascular disease, Hypertension, Rheumatoid Arthritis, Crohn's disease, Type 1 Diabetes and Type 2 Diabetes, with 2,000 cases for each of the diseases and 3,000 shared common controls(2007). There were 24 independent association signals identified and many of them have been replicated in independent replication studies. Here we describe the experience of our searching for rare variants by haplotype analysis across the genome in the WTCCC data.

Materials and methods

A detailed description of study samples can be found in the original WTCCC GWA study paper(2007). In brief the WTCCC dataset includes seven major complex diseases: bipolar disease (BD), coronary artery disease (CAD), Crohn's disease (CD), rheumatoid arthritis (RA), type 1 diabetes (T1D), type 2 diabetes (T2D); each has ~2,000 individuals, and a shared ~3,000 controls. The majority of subjects were of European ancestry. All the individuals were genotyped using Affymetrix GeneChip 500K arrays. We downloaded the genotype data called by the algorithm CHIAMO for all the seven disease cases and the shared controls (which consist of the 1958 Birth Cohort (58C) and UK Blood Service sample (NBS)) from the WTCCC website.

Framingham Heart Study. A detailed description of study samples can be found at Levy et al.(Levy et al. 2009). Our goal is to extract as many as unrelated cases and controls from the available family data. We defined hypertensive case as the systolic blood pressure >140 or diastolic blood pressure >90 or on medication treatments at any one of the four visits, and normtensive controls as the systolic blood pressure <140 and diastolic blood pressure <90 and no medication treatment at any one of the four visits. We then examined each family and chose the youngest case when there are multiple cases in a family, and the oldest control if there are multiple controls in a family. This process results 549 cases and 547 controls in our final analysis.

Quality controls

The individuals dropped in the WTCCC study because of evidence of non-European ancestry or call rate were excluded in the current analysis. We applied the following criteria to call SNPs: 1) CHIAMO probability greater than 0.95; 2) HWE exact test p-value $<5.7\times10^{-7}$ in controls; 3) allele frequency difference test based on 1df Trend Test p-value $<5.7\times10^{-7}$ or genotype frequency difference based on 2df General Test $<5.7\times10^{-7}$ between 58C and NBS. We further excluded the SNPs with missing genotype proportion >1% or minor allele frequencies<1%. We further dropped the SNPs with bad genotype calling, as suggested in the original WTCCC analysis(2007). Supplementary Table 1 shows the numbers of individuals as well as the numbers of SNPs that were analyzed for all the seven diseases and shared controls.

For FHS data, we further performed Mendelian inheritance consistence check. We set a genotype as missing if Mendelian inheritance error was identified.

Inferring haplotypes

In each gene or block, haplotypes were then inferred using the software BEAGLE 3.0 (Browning and Browning 2007) which is based on the localized haplotype-cluster model. To account for the uncertainty of haplotype inference, we sampled an individual's haplotype

conditional on the individual's genotype and the estimated haplotype frequency. We repeated the analysis based on the sampled haplotypes given the estimated haplotype frequencies and individual genotypes.

Two-stage analysis

We hypothesized that a complex disease can be attributed to both common and multiple rare variants. Further, we hypothesized that multiple rare variants can be captured by many haplotypes(Zhu et al. 2010). We applied two-stage analysis to the WTCCC data using the method in Zhu et al(Zhu et al. 2010) to the i^{th} gene or block (G_i), where i=1 to N, and N is the total number of genes and blocks. In stage 1, co-classification of risk haplotype, we randomly selected 400 cases and 1,000 controls. For each disease, we examined whether a haplotype is more frequent in cases than in controls by performing a one-sided Fisher exact test. We defined the risk haplotype set (SR_i) as the set of haplotypes that have one-sided Fisher exact test p-value <0.05 for the i^{th} gene or block. Similarly, we defined the protective haplotype set (SP_i) as the set that included haplotypes more frequent in controls than in cases for the i^{th} gene or block. In the stage 2 association test, we compared the frequency of risk haplotype set SR_i and protective haplotype set SP_i, identified in stage 1, between the remaining cases and controls. Because there was no overlap of samples between stages 1 and 2, the p-values calculated in stage 2 are valid. The Q-Q plot of the −log10(p-value) was used to examine whether there is any effect of population stratification or cryptic relatedness in the analysis. Because the power of two-stage analysis test is dependent on how well the co-classification performed at stage 1 and the sample size at stage 2, the actual power of two stage test will not change much when different sample size at the stage1 is used. We then randomly selected another 400 cases and 1000 controls at stage 1 and kept the rest of the samples for the stage 2 analysis. This process was performed 100 times and the smallest p-value for testing the risk haplotype set SR_i was recorded for the ith gene or block as p_i. Similarly, we recorded the smallest p-value of testing the protective haplotype set SP_i as q_i. We next ranked the p_i and q_i separately. We selected the top-ranked genes or blocks. When increasing the number of times to 1000, the top genes or blocks did not vary much. Thus, we only reported the results based on 100 times.

Evaluating the significance of the selected genes or blocks

Since p_i and q_i are not the true p-values for a gene or block due to the selection of the smallest p-value among 100 resamplings, we use a permutation procedure to evaluate the true p-values for the selected genes or blocks. We were concerned that using only a portion of samples in the stage 1 co-classification might reduce the efficiency of the method. Thus, for the top genes or blocks we selected, we reanalyzed data using the entire sample for the co-classification stage. We then tested the association of each risk haplotype set or protective haplotypes using the entire sample, again by Fisher's exact test. We recorded the p-values for each gene as the observed p-value. For each gene or block, we randomly shuffled the disease status for 1,000,000 replications and the p-values were calculated in the same way; then these p-values were tallied to calculate the empirical p-value for each selected gene. Because our permutation procedure is only for the top genes or blocks selected, this method is computationally efficient.

Examining whether a risk haplotype set or a protective haplotype set can be tagged by a common SNP

We created a pseudo-SNP genotype for an individual according to the number of risk haplotypes carried in the risk (protective) haplotype set. That is, an individual will have genotype 2/2 if he/she carried both haplotypes from a haplotype set, 1/2 if he/she only carried one haplotype from the haplotype set, and 0/0 if he/she carried no haplotypes from the risk haplotype set. We then evaluated the linkage disequilibrium between the pseudo-

SNP and genotyped SNPs in the analyzed samples. A strong LD suggested that the risk haplotype set can be well tagged by a single real SNP. Similar analysis was also performed for the protective haplotype sets.

Results

We used the SNP map annotations provided by Affymetrix 6.0 GeneChip as a reference (https://www.affymetrix.com/support/technical/annotationfilesmain.affx). We mapped a SNP to a particular gene if the SNP is mapped to a gene based on the Affymetrix annotations. If SNPs are located between two neighboring genes, we mapped them in their own block. However, if a gene or block includes a large number of SNPs, we further divided it to small blocks, with each block having less than 100 SNPs. The total numbers of genes and blocks for the seven diseases ranges from 19613-19678 (Supplemental table 1). In each gene or block, we inferred the most likely haplotypes for all the individuals in the WTCCC data using the software BEAGLE 3.0(Browning and Browning 2007) which is based on the localized haplotype-cluster model. We hypothesized that a complex disease can be attributed to both common and multiple rare variants. Further, we hypothesized that multiple rare variants can be captured by many haplotypes(Zhu et al. 2010). We applied the computational efficient two-stage analysis method(Zhu et al. 2010) in the WTCCC data to each gene or block (See Methods). Figure 1 and 2 presents the QQ plots of $-\log_{10}$(p-value) for testing association at stage 2 between the 7 disease cases and the common controls against the uniform distribution, which is the expected distribution under the null hypothesis, and the genome-wide $-\log_{10}$(P value) according to the chromosomal positions of genes in association tests. Overall we did not observe any substantial deviation from the null as suggested by the inflated factor λ (Figure 1 and 2). However, we did observed heavy tails for T1D and RA, which is mainly driven by many known genes in MHC and HLA regions. When we excluded the SNPs in the MHC and HLA region and redraw the QQ plots of RA and T1D (supplemental figure 1 and figure 2), the heavy tails were essentially disappeared. Our analysis results suggest that neither population stratification nor cryptic relatedness play a significant role in the data analysis, which is consistent with the original WTCCC report. Since the power of the two-stage method to detect genes is dependent on the samples selected for stage 1 and 2 analysis, we then repeated the same analysis 100 times, each time taking a new random sample to obtain the stage 1 individuals. We recorded the smallest p-value for each gene or block among the 100 resamplings. We then ranked the p-values for 7 diseases separately. To save computing time, we selected the top 50 risk and protective genes and blocks for each of the diseases except RA and T1D, for which we selected 100 and 150, respectively. To further reduce the type I error because of genotyping quality, we dropped SNPs with CHIAMO probability less than 0.99, as suggested by Browning and Browning(Browning and Browning 2008). We then redid the two-stage analysis with both stages using the entire sample. We performed 1,000,000 permutations to evaluate the p-values for all selected genes and blocks in order to account for the dependence of the two stages. Table 1 summarizes the genes or blocks that reached a genome-wide significance level (nominal $p<2.5\times10^{-6}$) when using the entire sample in the co-classification stage for risk and protective haplotype sets, respectively. For RA and T1D in HLA regions, we only list the most significant genes - the full set of genes is listed in Supplemental Table 2. We used a p-value 2.5×10^{-6} to declare the genome-wide significance because for each disease there is a total of <20,000 independent tests. This significant level corresponds to P-value 0.05 after the Bonferroni correction of 20,000 independent tests. The genes or regions showing moderate evidence of association ($p\leq10^{-4}$) are summarized in Table 2. Although we aimed to detect rare variants, among the 23 strongest association regions reported in the WTCCC study, 13 also reached genome-wide significance in this analysis. Of the remaining 10 regions, 2 also showed moderate association evidence (Table 2). When examining the maximum LD between the SNPs and the pseudo-SNPs clustered by

risk or protective haplotypes in these 15 genes, only half of them have r^2 >0.8 (Table 1). However, we also identified 11 additional genes and blocks which were either not reported or showed only moderate evidence of association in the original WTCCC report (table 1). Among these 11 genes, 3 have a cumulative frequency of risk (or protective) haplotypes <5%. No SNPs in genes or blocks can well tag the rare risk haplotypes (maximum r^2 <0.31 Table 1), indicating the association evidence for these rare haplotypes cannot be driven by any individual SNPs. Further, it is reasonable to believe that rare variants are unlikely to be well tagged by the common SNPs available in chips. The average number of risk (or protective) haplotypes in the significant genes in table 1 is 4.2, suggesting multiple variants may independently contribute to the diseases.

Bipolar disease (BD)

BD is a psychiatric disorder and is still poorly understood genetically. We did not observe any genes reaching genome-wide significance. Of the 7 genes showing moderate association evidence (Table 2, empirical p-value <10E-4), RNPEPL1 (arginyl aminopeptidase-like 1) and TDRD9 (tudor domain containing 9) showed moderate association evidence in the WTCCC report(2007). Among the rest of the genes, POFUT2 is located on 21q22.3, which is an active region for searching genes affecting BD where both linkage and association evidence have been reported(Kato 2007;Straub et al. 1994). The region where ZDHHC13 (zine finger, DHHC domain containing 12 isoform) (McInnis et al. 2003) located has also been reported of linkage evidence.

Coronary artery disease (CAD)

We detected 3 genes and one region showing genome wide significant association evidence to CAD. Among the three genes, CDKN2B is the only gene reaching genome-wide significance in the original WTCCC study and is also identified by our method (Empirical p-value=1.0E-6). The other two novel genes are hemochromatosis type 2 (HFE2, Empirical p-value<1.0E-6) and eukaryotic translation initiation factor 4H (EIF4H, Empirical p-value<1.0E-6). HFE2 was also detected by Browning and Browning (Browning and Browning 2008). The region of HFE2 has shown linkage to juvenile hemochromatosis which is a feature of heart failure(Rivard et al. 2003). We also detected a region located between 87.9-88 Mb (Empirical p-value=1.0E-6), where the nearest genes are gap junction protein and beta 7 (GJB7). The risk haplotype frequencies are rare and they cannot be tagged by common SNPs (r^2<0.32). Among the genes with empirical p-value<1.0E-4, it is interesting that the variants in PSRC1 (empirical p-value=1.6E-5) have been detected to be associated with CAD in a large GWAS analysis(Samani et al. 2007). This gene was not reported in the original WTCCC report although the risk haplotype can be well tagged by a SNP in the gene. The results are in general consistent with the our previous results using different haplotype inference method(Zhu et al. 2010), except that gene ZBTB43 could not be detected by this method.

Crohn's disease (CD)

We observed 6 genes and one block significantly associated with CD (Table 1). These seven regions are either strongly or moderately associated with CD in the WTCCC report. Although the frequency of the cumulative risk haplotypes in each region is not rare (>7%), the maximum r^2 values between SNPs and risk haplotype set are relatively small except for gene NOD2 (nucleotide-binding oligomerization domain) and the block ranged 131.83-131.84Mb on chromosome 5 (r^2 >0.95). Gene PTGER4 (prostaglandin E receptor 4) has also been reported to be associated with CD with possibly multiple variants contributing to disease susceptibility(Libioulle et al. 2007). Among the 6 moderate association evidence regions, two regions were also reported in the WTCCC (Table 2). The association evidence

of BSN (bassoon protein) has been replicated in the Spanish's population(Marquez et al. 2009).

Hypertension (HT)

There was no SNP reaching genome-wide significance for HT in the original WTCCC report. We identified a novel gene ZFAT1 (zinc finger protein 406 isoform, empirical p-value<1.0E-6), which is significantly associated with HT. This result is consistent with that found when we used a slightly different analysis approach(Zhu et al. 2010). In the linkage analysis of large pedigree data from South Italy, genome-wide significant linkage evidence to essential hypertension was reported on chromosome 8q22-23 (Ciullo et al. 2006), where the ZFAT1 gene is located. There are 7 risk haplotypes with total frequency 4.5% in cases and 1.1% in controls. These risk haplotypes form a set that cannot be tagged by common SNPs (r^2=0.107).

We also identified two genes: ABLIM1 (actin binding LIM protein 1, empirical p=1.2E-5) on chromosome 10 and NR2F2 (nuclear receptor subfamily 2, group F, member 2, empirical p=5.5E-5) on chromosome 15, moderately associated with HT and were not reported in Zhu et al. 2010. Interestingly, a study of transcriptional profiling with a blood pressure QTL interval-specific oligonucleotide array using the Dahl salt-sensitive rat has suggested that the homologous gene NR2F2 is associated with blood pressure in the rat(Joe et al. 2005). This gene was also identified by multilocus association testing method in the WTCCC data(Browning and Browning 2008), although that study did not focus on searching for rare variants.

Replication of HT in FHS data

We performed analysis of these three genes: ZFAT1, ABLIMI and NR2F2 in FHS data. We used all the sample in both stage 1 and 2 analysis and evaluated the p-value using 1,000,000 permutations. Since ZFAT1 is a large gene with 278 SNPs genotyped, we partitioned ZFAT1 into 3 blocks with size 100, 100 and 78 SNPs, respectively, and tested each block accordingly. We could not replicate the association evidence when the haplotypes of using the same set of SNPs identified in WTCCC was tested. However, the association evidence was observed when test was performed on the neighbor block. We identified 3 risk haplotypes are moderately associated with HT (empirical p-value=8.2E-5, Table 3). When we combined the SNPs identified from both WTCCC and FHS, the significance of association is reduced (Empirical p-value =0.059). We also identified 4 protective haplotypes in ABLIM1 significantly associated with HT (empirical p-value=0.01, table 3). However, we failed to replicate the association evidence in NR2F2.

Rheumatoid arthritis (RA)

The two significant regions identified in the WTCCC study were also identified in this analysis. The association between RA and the MHC region has been well established inWTCCC study. We also identified many genes in this region associated with RA (Supplemental Table 2). The strongest association evidence is on the block from 32.5-32.9 Mb, where the most significant SNP in WTCCC report, rs6457617, is located. The significance level in this analysis is much higher than that in the WTCCC report. This block includes 307 haplotypes and 10 of them are identified as risk haplotypes with total frequency 49.5% in cases and 25.0% in controls. Other than rs6457617, which was reported as the most significant one in single locus analysis in WTCCC, SNP rs9275418 has the maximum r^2 value with the risk haplotype set (r^2 =0.31), suggesting additional variants, beside rs6457617, independently associated with RA. The other known gene is PTPN22 (protein tyrosine phosphatase, non-receptor type 22). There is only one risk haplotype

among 37 haplotypes identified in this gene and is in strong LD with SNP rs6679677 (r^2 =0.99).

We identified 3 additional genes or regions that reached genome-wide significance. Among them, OLIG3 (oligodendrocyte transcription factor 3) has been reported to be associated with RA(Plenge et al. 2007). We identified 3 risk haplotypes among 46 haplotypes, with the largest r^2 =0.52 between SNPs and risk haplotypes. The remaining 2 genes or blocks are novel, including NDST3 (N-deacetylase/N-sulfotransferase 3, and a block between 16.8-17 Mb on chromosome 17. The maximum r^2 between the risk haplotype set and SNPs is all less than 0.09. NDST3 is located in 4q27, where association evidence has been identified with autoimmune diseases, including RA(Liu et al. 2008; Zhernakova et al. 2007). Among the 2 genes with p-values <1.0E-4, the region including the OS9 (osteosarcoma amplified 9, endoplasmic reticulum lectin) gene has shown replication evidence to RA(Barton et al. 2008).

Type I diabetes (T1D)

We observed 6 regions that reach genome-wide significance for T1D. The strongest region associated with T1D is the major histocompatibility complex (MHC), where there are many genes that have shown association evidence (Supplemental table 2). The strongest genes and block include NOTCH4 (notch4 preproprotein), C6orf10 (chromosome 6 open reading frame 10), BTNL2 (butyrophilin-like 2) and block 32.52-32.89 Mb on chromosome 6. The minimum number of risk haplotypes in these genes and block is 11 and the largest r^2 value is 0.45. The MHC region is well established for association with T1D, but how many independent variants in the MHC region contribute to T1D is still unknown. The other genes, including PHTF1 (putative homeodomain transcription factor 1), RAB5B (member RAS oncogene family), and SH2B3 (lymphocyte adaptor protein), identified in the WTCCC report are also observed in this analysis.

We identified two novel regions, including ADAD1 (adenosine deaminase domain containing 1) and a block in chromosome 16 (0.99-1.03Mb) significantly associated with T1D. ADAD1 is a region showing moderate association evidence in the WTCCC. We did not observe any SNP that can well tag the 2 risk haplotypes identified in the block.

Type 2 diabetes (T2D)

We only identified one gene, ZFAT1, that was not reported in the WTCCC study to be significantly associated with T2D. Among 241 haplotypes in ZFAT1, we identified 4 are risk haplotypes. No single SNP can well tag the haplotype risk set. Interestingly, this gene is also shown to be significantly associated with HT. Among the risk haplotypes detected, 4 are shared by both HT and T2D (Table 4), suggested these rare haplotypes may contribute to both HT and T2D. We failed to identify both the TCF7L2 and FTO genes whose association to T2D has been established(2007; Grant et al. 2006). Further examining these two genes, we observed there are 4841 and 2220 haplotypes in TCF7L2 and FTO, respectively. Simulation studies suggested the current method will have limited power when the number of haplotypes increases and the haplotype frequencies are too rare. Among the genes or regions reaching a p-value <10E-4, linkage evidence has been reported in these genes: PLXNA2 (plexin-A2), TRIP13 (thyroid hormone receptor interactor 13), block (42.75-42.76Mb) on chromosome 15, and block (18.259-18.259Mb) on chromosome 20(Lillioja and Wilton 2009).

Discussion

We have conducted a genome-wide search for rare genetic variants using the GWAS design by reanalyzing the WTCCC data. Although only the common SNPs were tagged in the Affymetrix 500K chip, our findings still detect rare variants by examining haplotypes in each gene and provide further understanding of the genetic patterns underlying complex diseases.

Our first experience is that we identified 8 novel genes or regions independently associated with the diseases. We should caution that these findings are tentative and further independent replication studies are necessary. However, we replicated the association evidence between ZFAT1 and HT in FHS data, although the evidence is not from the same block. We believe the replication for rare variants could be much challenged and it is less likely to have the same variants showing association evidence in two independent studies. The 7 risk haplotypes together occur in 4.5% of hypertensive cases and 1.1% of controls. No common SNPs are in strong LD with the 7 identified rare risk haplotypes, suggesting multiple rare variants in this gene contribute to HT. Interestingly, ZFAT1 is also identified to be associated with T2D, with 4 risk haplotypes shared by both HT and T2D cases. It has been known that HT is extremely common in patients with type 2 diabetes, affecting up to 60%(Varughese and Lip 2005). This analysis suggests ZFAT1 may contribute both HT and T2D. In this study, all the novel genes and blocks identified have small cumulative risk (or protective) haplotype frequencies. These rare risk (protective) haplotypes cannot be well tagged by common SNPs. Multiple rare haplotypes were also observed in 5 of the 8 novel genes or blocks, further suggesting multiple rare variants likely contribute to the variation of the diseases.

For the genes reported in the WTCCC study, we also replicated 8 of their 24 genes. When a common variant is solely the cause, our approach is expected to be less powerful than a single SNP approach. Interestingly, two blocks capture our attention. The block on chromosome 6 ranged from 32.5-32.9Mb, which is in the HLA region, is highly significantly associated with RA. The most significantly associated SNP associated with RA in single SNP analysis is rs6457617 in the WTCCC report, which is located in this block. However, the significance level in the WTCCC report is far less than that in this study, even on comparing with the less efficient two-stage method with independent samples in stage 1 and 2 (single SNP p-value in the WTCCC 3.44×10^{-76} vs 2.94×10^{-94}, after adjusting for 100 multiple comparisons). In addition, SNP rs6457617 is not the SNP having the maximum LD with the risk haplotype set we detected, suggested that rs6457617 may not be a causative variant. The most significant region associated with T1D is the same block as for RA, where the most significant SNP rs9272346 in the WTCCC report is located. Similarly, the significance level of single SNP analysis is far less than the less efficient two-stage method (p-value 2.42×10^{-134} vs 1.12×10^{-279}). SNP rs9272346 is also not the SNP having the maximum LD with the risk haplotype set we detected for T1D, suggesting additional independent variants exist in this region contributing to T1D. This can also be further confirmed in that many genes in the MHC region are strongly associated with RA and T1D (supplemental table 2).

Our analysis was based on the most likely haplotypes inferred from the statistical software BEGEAL(Browning and Browning 2007). To overcome any concern about haplotype uncertainty, we reanalyzed the significant genes we identified by sampling an individual haplotypes conditional on the individual's genotypes and the haplotype frequency. The results are consistent in general, indicating that significant evidence identified in this study is unlikely due to incorrect haplotype inference (Supplementary table 3). However, we did observe a block for T1D, is strongly affected by the uncertainly of haplotype inference.

These blocks may reflect the false positive due to the haplotype uncertainty. We were also concerned about the possibility of genotype errors, which could lead to biased results. We therefore applied stricter QC procedures by dropping SNPs with missing rate>0.01. We compared the missing rates between cases and controls and did not observe any systemic difference (Supplementary table 4). Further, we did not observe any single SNP in strong LD with the risk (or protective) haplotype sets in the novel genes or blocks we identified, suggesting the findings are unlikely driven by SNP genotyping error. We also did not observe any strong effect of population structure in the rare variant analysis, consistent with the original WTCCC study.

The identification of the novel genes and regions by searching for rare risk (protective) haplotypes demonstrates that it is an efficient alternative way, beside single SNP analysis, for common variants in GWA studies. The identified risk (protective) haplotypes can serve as guidance for future resequencing analysis in order to identify the underlying functional variants. Especially, resequencing a region or a gene can make it possible to determine the rare causal variants falling on the risk (protective) haplotypes detected in GWAS.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dr. RC Elston for his constructed reading and comments of the manuscript. The work was supported by the National Institutes of Health, grant numbers HL074166, HL086718 from National Heart, Lung, Blood Institute, HG003054 from the National Human Genome Research Institute, RR03655 from the National Center for Research Resources. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk/info/participants.shtml. Funding for that project was provided by the Wellcome Trust under award 076113. The Framingham Heart Study research was supported by NHLBI Contract: 2 N01-HC-25195-06 and its contract with Affymetrix, Inc for genotyping services (Contract No. N02-HL-6-4278). The authors are grateful to the many investigators within the Framingham Heart Study who have collected and managed the data and especially to the participants for their invaluable time, patience, and dedication to the Study.

References

Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661–78. [PubMed: 17554300]

Barton A, Thomson W, Ke X, Eyre S, Hinks A, Bowes J, Plant D, Gibbons LJ, Wilson AG, Bax DE, Morgan AW, Emery P, Steer S, Hocking L, Reid DM, Wordsworth P, Harrison P, Worthington J. Rheumatoid arthritis susceptibility loci at chromosomes 10p15, 12q13 and 22q13. Nat Genet 2008;40:1156–9. [PubMed: 18794857]

Browning BL, Browning SR. Haplotypic analysis of Wellcome Trust Case Control Consortium data. Hum Genet 2008;123:273–80. [PubMed: 18224336]

Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 2007;81:1084–97. [PubMed: 17924348]

Ciullo M, Bellenguez C, Colonna V, Nutile T, Calabria A, Pacente R, Iovino G, Trimarco B, Bourgain C, Persico MG. New susceptibility locus for hypertension on chromosome 8q by efficient pedigree-breaking in an Italian isolate. Hum Mol Genet 2006;15:1735–43. [PubMed: 16611673]

Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 2004;305:869–72. [PubMed: 15297675]

Goldstein DB. Common genetic variation and human traits. N Engl J Med 2009;360:1696–8. [PubMed: 19369660]

Grant SF, Thorleifsson G, Reynisdottir I, Benediktsson R, Manolescu A, Sainz J, Helgason A, Stefansson H, Emilsson V, Helgadottir A, Styrkarsdottir U, Magnusson KP, Walters GB, Palsdottir E, Jonsdottir T, Gudmundsdottir T, Gylfason A, Saemundsdottir J, Wilensky RL, Reilly MP, Rader DJ, Bagger Y, Christiansen C, Gudnason V, Sigurdsson G, Thorsteinsdottir U, Gulcher JR, Kong A, Stefansson K. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat Genet 2006;38:320–3. [PubMed: 16415884]

Hirschhorn JN. Genomewide association studies--illuminating biologic pathways. N Engl J Med 2009;360:1699–701. [PubMed: 19369661]

Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, Simon DB, Newton-Cheh C, State MW, Levy D, Lifton RP. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat Genet 2008;40:592–9. [PubMed: 18391953]

Joe B, Letwin NE, Garrett MR, Dhindaw S, Frank B, Sultana R, Verratti K, Rapp JP, Lee NH. Transcriptional profiling with a blood pressure QTL interval-specific oligonucleotide array. Physiol Genomics 2005;23:318–26. [PubMed: 16204469]

Kato T. Molecular genetics of bipolar disorder and depression. Psychiatry Clin Neurosci 2007;61:3–19. [PubMed: 17239033]

Levy D, Ehret GB, Rice K, Verwoert GC, Launer LJ, Dehghan A, Glazer NL, Morrison AC, Johnson AD, Aspelund T, Aulchenko Y, Lumley T, Kottgen A, Vasan RS, Rivadeneira F, Eiriksdottir G, Guo X, Arking DE, Mitchell GF, Mattace-Raso FU, Smith AV, Taylor K, Scharpf RB, Hwang SJ, Sijbrands EJ, Bis J, Harris TB, Ganesh SK, O'Donnell CJ, Hofman A, Rotter JI, Coresh J, Benjamin EJ, Uitterlinden AG, Heiss G, Fox CS, Witteman JC, Boerwinkle E, Wang TJ, Gudnason V, Larson MG, Chakravarti A, Psaty BM, van Duijn CM. Genome-wide association study of blood pressure and hypertension. Nat Genet. 2009

Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 2008;83:311–21. [PubMed: 18691683]

Libioulle C, Louis E, Hansoul S, Sandor C, Farnir F, Franchimont D, Vermeire S, Dewit O, de Vos M, Dixon A, Demarche B, Gut I, Heath S, Foglio M, Liang L, Laukens D, Mni M, Zelenika D, Van Gossum A, Rutgeerts P, Belaiche J, Lathrop M, Georges M. Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. PLoS Genet 2007;3:e58. [PubMed: 17447842]

Lillioja S, Wilton A. Agreement among type 2 diabetes linkage studies but a poor correlation with results from genome-wide association studies. Diabetologia 2009;52:1061–74. [PubMed: 19296077]

Liu Y, Helms C, Liao W, Zaba LC, Duan S, Gardner J, Wise C, Miner A, Malloy MJ, Pullinger CR, Kane JP, Saccone S, Worthington J, Bruce I, Kwok PY, Menter A, Krueger J, Barton A, Saccone NL, Bowcock AM. A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci. PLoS Genet 2008;4:e1000041. [PubMed: 18369459]

Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 2009;5:e1000384. [PubMed: 19214210]

Marquez A, Cenit MC, Nunez C, Mendoza JL, Taxonera C, Diaz-Rubio M, Bartolome M, Arroyo R, Fernandez-Arquero M, de la Concha EG, Urcelay E. Effect of BSN-MST1 locus on inflammatory bowel disease and multiple sclerosis susceptibility. Genes Immun. 2009

McInnis MG, Dick DM, Willour VL, Avramopoulos D, MacKinnon DF, Simpson SG, Potash JB, Edenberg HJ, Bowman ES, McMahon FJ, Smiley C, Chellis JL, Huo Y, Diggs T, Meyer ET, Miller M, Matteini AT, Rau NL, DePaulo JR, Gershon ES, Badner JA, Rice JP, Goate AM, Detera-Wadleigh SD, Nurnberger JI, Reich T, Zandi PP, Foroud TM. Genome-wide scan and conditional analysis in bipolar disorder: evidence for genomic interaction in the National Institute of Mental Health genetics initiative bipolar pedigrees. Biol Psychiatry 2003;54:1265–73. [PubMed: 14643094]

Newton-Cheh C, Johnson T, Gateva V, Tobin MD, Bochud M, Coin L, Najjar SS, Zhao JH, Heath SC, Eyheramendy S, Papadakis K, Voight BF, Scott LJ, Zhang F, Farrall M, Tanaka T, Wallace C, Chambers JC, Khaw KT, Nilsson P, van der Harst P, Polidoro S, Grobbee DE, Onland-Moret NC, Bots ML, Wain LV, Elliott KS, Teumer A, Luan J, Lucas G, Kuusisto J, Burton PR, Hadley D, McArdle WL, Brown M, Dominiczak A, Newhouse SJ, Samani NJ, Webster J, Zeggini E, Beckmann JS, Bergmann S, Lim N, Song K, Vollenweider P, Waeber G, Waterworth DM, Yuan

X, Groop L, Orho-Melander M, Allione A, Di Gregorio A, Guarrera S, Panico S, Ricceri F, Romanazzi V, Sacerdote C, Vineis P, Barroso I, Sandhu MS, Luben RN, Crawford GJ, Jousilahti P, Perola M, Boehnke M, Bonnycastle LL, Collins FS, Jackson AU, Mohlke KL, Stringham HM, Valle TT, Willer CJ, Bergman RN, Morken MA, Doring A, Gieger C, Illig T, Meitinger T, Org E, Pfeufer A, Wichmann HE, Kathiresan S, Marrugat J, O'Donnell CJ, Schwartz SM, Siscovick DS, Subirana I, Freimer NB, Hartikainen AL, McCarthy MI, O'Reilly PF, Peltonen L, Pouta A, de Jong PE, Snieder H, van Gilst WH, Clarke R, Goel A, Hamsten A, Peden JF, et al. Genome-wide association study identifies eight loci associated with blood pressure. Nat Genet. 2009

Plenge RM, Cotsapas C, Davies L, Price AL, de Bakker PI, Maller J, Pe'er I, Burtt NP, Blumenstiel B, DeFelice M, Parkin M, Barry R, Winslow W, Healy C, Graham RR, Neale BM, Izmailova E, Roubenoff R, Parker AN, Glass R, Karlson EW, Maher N, Hafler DA, Lee DM, Seldin MF, Remmers EF, Lee AT, Padyukov L, Alfredsson L, Coblyn J, Weinblatt ME, Gabriel SB, Purcell S, Klareskog L, Gregersen PK, Shadick NA, Daly MJ, Altshuler D. Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. Nat Genet 2007;39:1477–82. [PubMed: 17982456]

Rivard SR, Lanzara C, Grimard D, Carella M, Simard H, Ficarella R, Simard R, D'Adamo AP, Ferec C, Camaschella C, Mura C, Roetto A, De Braekeleer M, Bechner L, Gasparini P. Juvenile hemochromatosis locus maps to chromosome 1q in a French Canadian population. Eur J Hum Genet 2003;11:585–9. [PubMed: 12891378]

Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, Dixon RJ, Meitinger T, Braund P, Wichmann HE, Barrett JH, Konig IR, Stevens SE, Szymczak S, Tregouet DA, Iles MM, Pahlke F, Pollard H, Lieb W, Cambien F, Fischer M, Ouwehand W, Blankenberg S, Balmforth AJ, Baessler A, Ball SG, Strom TM, Braenne I, Gieger C, Deloukas P, Tobin MD, Ziegler A, Thompson JR, Schunkert H. Genomewide association analysis of coronary artery disease. N Engl J Med 2007;357:443–53. [PubMed: 17634449]

Straub RE, Lehner T, Luo Y, Loth JE, Shao W, Sharpe L, Alexander JR, Das K, Simon R, Fieve RR, et al. A possible vulnerability locus for bipolar affective disorder on chromosome 21q22.3. Nat Genet 1994;8:291–6. [PubMed: 7874172]

Varughese GI, Lip GY. Antihypertensive therapy in diabetes mellitus: insights from ALLHAT and the Blood Pressure-Lowering Treatment Trialists' Collaboration meta-analysis. J Hum Hypertens 2005;19:851–3. [PubMed: 16079882]

Visscher PM. Sizing up human height variation. Nat Genet 2008;40:489–90. [PubMed: 18443579]

Zhernakova A, Alizadeh BZ, Bevova M, van Leeuwen MA, Coenen MJ, Franke B, Franke L, Posthumus MD, van Heel DA, van der Steege G, Radstake TR, Barrera P, Roep BO, Koeleman BP, Wijmenga C. Novel association in chromosome 4q27 region with rheumatoid arthritis and confirmation of type 1 diabetes point to a general risk locus for autoimmune diseases. Am J Hum Genet 2007;81:1284–8. [PubMed: 17999365]

Zhu X, Feng T, Li Y, Lu Q, Elston RC. Detecting Rare Variants for Complex Traits Using Family and Unrelated Data. Genet Epidemiol 2010;32:171–187. [PubMed: 19847924]

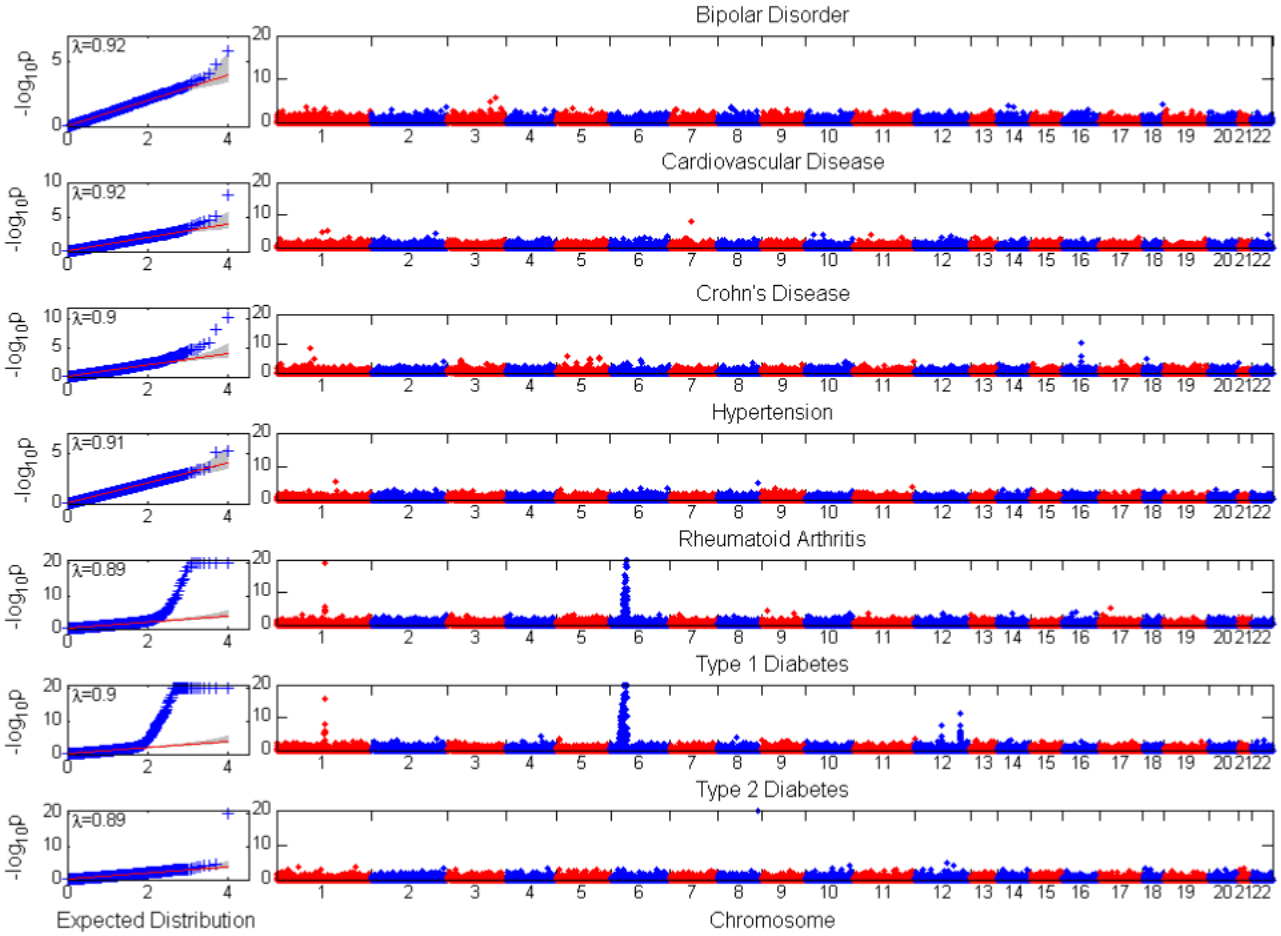

Figure 1.

QQ plots of $-\log_{10}$(p-value) for testing association of risk haplotype set at stage 2 between the 7 disease cases and the common controls against the uniform distribution (left panel), and the Manhattan plot of the genome-wide $-\log_{10}$(P value) according to the chromosomal positions of genes in association tests (right panel).

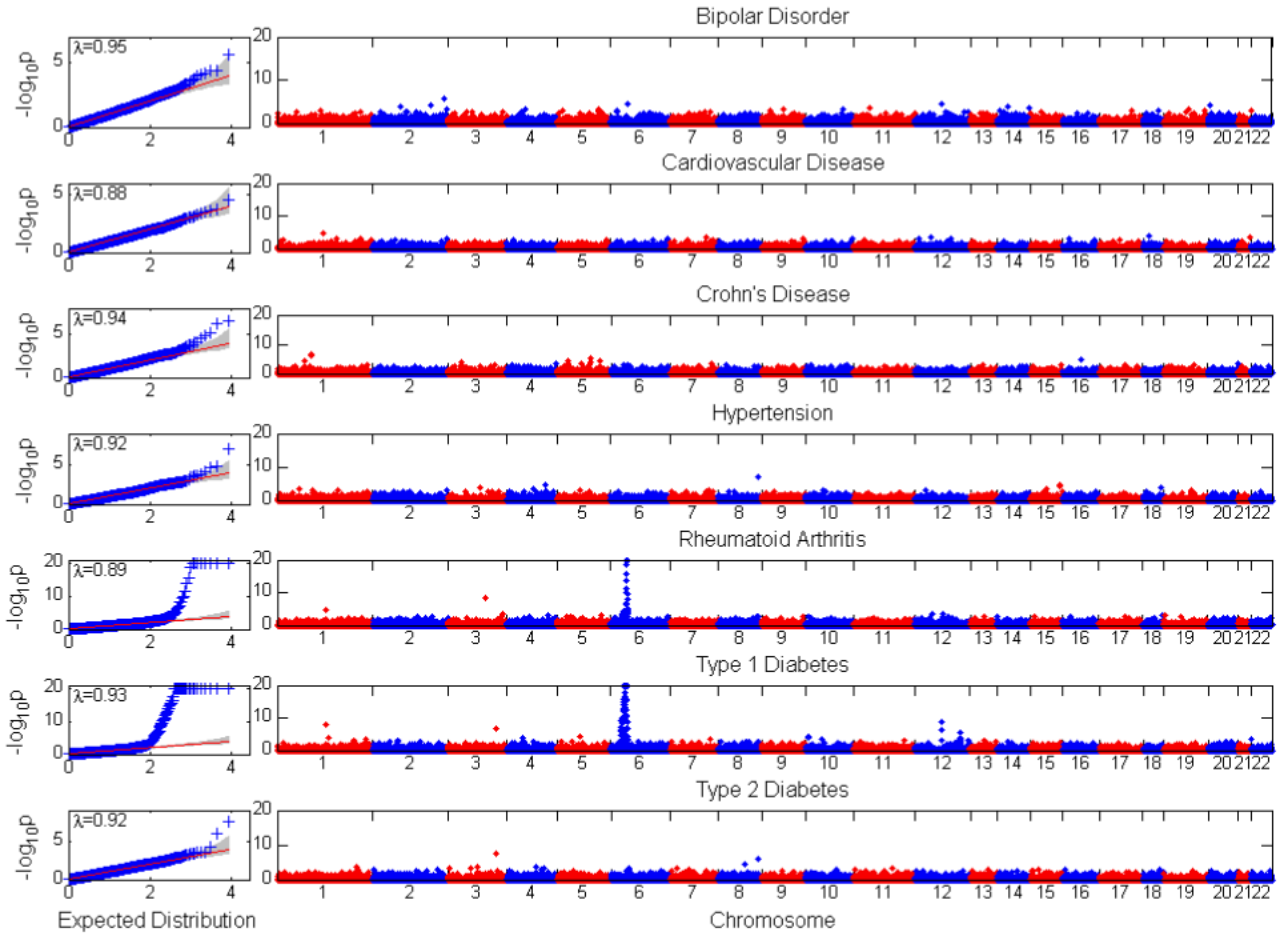

Figure 2.

QQ plots of $-\log_{10}$(p-value) for testing association of protective haplotype set at stage 2 between the 7 disease cases and the common controls against the uniform distribution (left panel), and the Manhattan plot of the genome-wide $-\log_{10}$(P value) according to the chromosomal positions of genes in association tests (right panel).

Table 1

regions of the genes or the genome showing strong association evidence

Disease	chromosome	Gene	Range (MB)	Freq of case haplotype set	Freq of control haplotype set	Smallest P value[a]	observed P value[b]	# of rare haplotypes/ total # of haplotypes	r^{2c}	SNP ID[d]	Empirical P value[e]	Association evidence in WTCCC study[f]
CAD	1	HFE2	144.11-144.15	0.0062	0.0003	1.67E-07	2.65E-08	1/3	0.0645	rs12091564	<1.0E-6	
CAD	6	Block	87.93-88.02	0.0049	0	2.32E-05	2.21E-08	2/21	0.0478	rs9362399	1.0E-06	
CAD	7	EIF4H	73.20-73.25	0.0119	0.0005	5.23E-14	1.13E-15	1/5	0.3191	rs150880	<1.0E-6	
CAD	9	CDKN2B	21.99-22.12	0.3367	0.2692	5.94E-08	7.10E-13	7/232	0.4509	rs1333049	1.00E-06	++
CD	1	IL23R	67.41-67.49	0.2477	0.3407	1.59E-16	9.67E-22	7/99	0.6263	rs11465760	<1.0E-6	++
CD	5	PTGER4	40.31-40.62	0.1144	0. 1749	6.73E-16	2.30E-15	13/446	0.1193	rs16869864	<1.0E-6	++
CD	5	PTGER4	40.31-40.62	0.3838	0.2983	1.06E-06	1.38E-17	9/446	0.2662	rs9292777	2.00E-06	++
CD	5	Block	131.83-131.84	0.1888	0.2306	1.37E-06	9.24E-07	1/4	0.9828	rs4705861	1.0E-06	+
CD	5	ZNF300	150.26-150.32	0.107	0.0752	3.18E-07	1.11E-07	1/7	0.1120	rs1478388	<1.0E-6	++
CD	6	C6orf10	32.34-32.45	0.1384	0.0951	1.64E-05	1.34E-10	6/81	0.3590	rs9391858	<1.0E-6	+
CD	16	NOD2	49.28-49.31	0.3472	0.2775	6.20E-14	9.56E-13	2/37	0.9591	rs17221417	<1.0E-6	++
HT	8	ZFAT1	135.57-135.67	0.0448	0.0112	3.95E-29	2.20E-25	7/247	0.1068	rs7816909	<1.0E-6	
RA	1	PTPN22	114.11-114.22	0.1704	0.0963	1.55E-24	3.17E-26	1/37	0.9923	rs6679677	<1.0E-6	++
RA	4	NDST3	119.26-119.42	0.007	0	3.47E-09	1.89E-11	1/50	0.0251	rs12650031	<1.0E-6	
RA	6	Block*	32.52-32.84	0.4956	0.250	2.94E-96	7.65E-133	10/307	0.3097	rs9275418	<1.0E-6	++
RA	6	OLIG3	138.05-138.14	0.3242	0.2672	2.60E-06	1.53E-09	3/46	0.5230	rs6920220	5.0E-06	+
RA	17	Block	16.89-17.01	0.0062	0	1.32E-06	3.28E-10	2/24	0.0891	SNP_A-1954587	<1.0E-6	
T1D	1	PTPN22*	114.11-114.22	0.1694	0.0965	3.69E-24	3.22E-26	1/32	0.9936	rs6679677	<1.0E-6	++
T1D	4	ADAD1	123.55-123.59	0.3064	0.2602	2.02E-07	3.02E-07	1/8	0.9965	rs17388568	1.00E-06	+
T1D	6	Block	32.52-32.83	0.6749	0.2421	1.12E-281	0	16/294	0.8305	rs9273363	<1.0E-6	++
T1D	12	ERBB3*	54.73-54.77	0.4017	0.3382	6.64E-11	1.55E-10	1/4	0.9943	rs2292239	<1.0E-6	++
T1D	12	C12orf30	110.95-111.00	0.5050	0.4242	5.90E-16	2.10E-15	1/7	0.9983	rs17696736	<1.0E-6	++
T1D	16	Block	0.99-1.03	0.081	0.1125	3.32E-06	1.56E-07	2/7	0.1458	rs535255	2.0E-06	
T2D	8	ZFAT1	135.57-135.67	0.0314	0.0003	1.41E-34	1.72E-46	4/241	0.1741	rs6421008	<1.0E-6	

[a] Smallest P-value was calculated as the smallest p-value of Fisher's exact tests among 100 resamplings when stage 1 co-classification used 400 cases and 1000 controls.

[b] Observed P-value was calculated using Fisher's exact test when stage 1 co-classification used the entire sample. This p-value should be considered as a gene-specific test statistic.

[c] r^2 is the maximum correlation between the risk (protective) haplotype set and the SNPs consisting haplotypes.

[d] SNP ID is the SNP having the maximum correlation with the risk (protective) haplotype set.

[e] Empirical P value was obtained based on 1,000,000 permutations when stage 1 co-classification used the entire sample. This p-value refers the reported p values in the text.

[f] ++ indicates strong association evidence was observed in WTCCC study; + indicates moderate association was observed in WTCCC study

* indicated there are many genes identified and we reported the most significant one.

Table 2

regions of the genes or the genome showing moderate association evidence

Disease	chromosome	Gene	Range (MB)	Freq of case haplotype set	Freq of control haplotype set	Smallest P value[a]	observed P value[b]	# of rare haplotypes in haplotype set	r^{2c}	Rs Name[d]	Empirical P value[e]	Association evidence in WTCCC study[f]
BD	2	RNPEPL1	241.16-241.16	0.1882	0.2253	7.21E-05	6.84E-06	2/4	1	rs6730107	2.60E-05	+
BD	5	FBXO38	147.72-147.81	0.621	0.6647	1.84E-05	5.22E-06	1/18	0.7522	rs6861078	7.10E-05	
BD	7	ZNF680	63.45-63.71	0.0342	0.0582	3.52E-05	3.85E-08	3/153	0.3223	rs633702	0.000131	
BD	11	ZDHHC13	19.07-19.15	0.0008	0.0092	2.48E-05	4.97E-09	3/97	0.0056	rs11025015	4.20E-05	
BD	14	KLHDC1	49.23-49.29	0.6673	0.7054	1.44E-04	4.61E-05	1/4	0.9937	rs12717402	0.000148	
BD	14	TDRD9	103.50-103.61	0.2537	0.2987	2.29E-06	1.12E-06	1/20	0.9896	rs11622475	2.70E-05	+
BD	21	POFUT2	45.51-45.54	0.0096	0.0031	0.000138	3.15E-05	1/4	0.0133	rs2838855	1.03E-05	
CAD	1	PSRC1	109.62-109.62	0.8092	0.7721	9.44E-06	5.61E-06	1/2	1	rs599839	1.40E-05	
CAD	1	Block	220.86-220.89	0.1952	0.2354	3.54E-04	1.46E-06	3/15	0.7341	rs3008613	2.30E-05	
CAD	17	Block	71.44-71.45	0.2461	0.2827	6.14E-05	4.04E-05	1/2	1	rs2608881	6.30E-05	
CD	2	C2orf65	74.66-74.70	0.8790	0.9036	5.09E-05	0.00010	1/2	1	rs363691	0.000219	
CD	3	BSN	49.58-49.68	0.3295	0.2821	3.17E-06	7.93E-07	1/15	0.9955	rs9858542	1.4E-05	++
CD	6	Block	107.50-107.55	0.2414	0.2860	1.23E-06	1.59E-06	2/16	0.5170	rs1665901	8.30E-05	
CD	8	LOC441376	117.99-118.03	0.1553	0.1183	1.02E-05	2.22E-07	2/44	0.1393	rs3020176	2.80E-05	
CD	10	Block	101.25-101.28	0.5066	0.4498	4.26E-06	5.69E-08	2/38	0.9503	rs6584283	6.00E-06	++
CD	20	ZGPAT	61.82-61.84	0.7849	0.7466	2.18E-05	1.34E-05	1/2	1	rs2738758	1.80E-05	
HT	10	ABLIM1	116.18-116.32	0.0236	0.0562	8.83E-06	5.43E-16	11/642	0.0354	rs6585278	1.20E-05	
HT	15	NR2F2	94.64-94.70	0.1176	0.1511	2.26E-05	1.17E-06	3/15	0.4636	rs11073474	5.50E-05	+
RA	12	CLEC1B	10.03-10.04	0.4261	0.4753	4.23E-05	1.29E-06	2/8	0.7233	rs770738	1.80E-05	
RA	12	OS9	56.34-56.39	0.3062	0.3468	3.25E-05	1.97E-05	1/2	1	rs10876991	4.00E-05	
T1D	1	HMGCS2	120.09-120.12	0.1110	0.1429	1.53E-05	2.07E-06	2/12	0.1444	rs3790692	3.90E-05	
T1D	2	Block	74.60-74.62	0.1617	0.1316	2.71E-05	1.84E-05	1/4	0.9992	rs6546909	2.90E-05	
T1D	4	Block	123.24-123.46	0.3054	0.2631	5.10E-06	2.88E-06	1/8	0.718	rs10015924	3.90E-05	+
T1D	6	FAM135A	71.13-71.33	0.2687	0.3139	2.70E-04	7.76E-07	3/26	0.4010	rs10498873	1.90E-05	
T1D	17	Block	36.02-36.03	0.3117	0.3534	3.53E-05	9.84E-06	1/2	1	rs7221109	1.80E-05	

Disease	chromosome	Gene	Range (MB)	Freq of case haplotype set	Freq of control haplotype set	Smallest P value[a]	observed P value[b]	# of rare haplotypes in haplotype set	r^{2c}	Rs Name[d]	Empirical P value[e]	Association evidence in WTCCC study[f]
T2D	15	ZFAND6	78.14-78.22	0.2419	0.283	1.15E-04	3.89E-06	2/8	0. 994375	rs2903265	4.90E-05	+

[a] Smallest P-value was calculated as the smallest p-value of Fisher's exact tests among 100 resamplings when stage 1 co-classification used 400 cases and 1000 controls.

[b] Observed P-value was calculated using Fisher's exact test when stage 1 co-classification used the entire sample. This p-value should be considered as a gene-specific test statistic.

[c] r^2 is the maximum correlation between the risk (protective) haplotype set and the SNPs consisting haplotypes.

[d] SNP ID is the SNP having the maximum correlation with the risk (protective) haplotype set.

[e] Empirical P value was obtained based on 1,000,000 permutations when stage 1 co-classification used the entire sample. This p-value refers the reported p values in the text.

[f] ++ indicates strong association evidence was observed in WTCCC study; + indicates moderate association was observed in WTCCC study

Table 3

Replication analysis of FHS data for the genes identified in WTCCC for HT

Gene	Range (MB)	Freq of case haplotype set	Freq of control haplotype set	observed P value[a]	# of rare haplotypes/total # of haplotypes	r^{2b}	SNP ID[c]	Empirical P value[d]
ZFAT1(1)	135.82-135.86	0.120219	0.0648995	4.88E-06	3	0.142317	rs11776156	0.00009
ZFAT1(2)	135.57-135.86	0.0355191	0.00914077	1.67E-05	3	0.0136243	rs6988000	0.059
ABLIM1	116.35-116.43	0.0346084	0.077697	7.23E-06	4	0.0967802	rs12570718	0.00987

ZFAT1(1) : haplotypes were constructed based on the SNPs in FHS data

ZFAT1(2) : haplotypes were constructed by combining the SNPs identified in WTCCC and FHS together.

[a] Observed P-value was calculated using Fisher's exact test when stage 1 co-classification used the entire sample. This p-value should be considered as a gene-specific test statistic.

[b] r^2 is the maximum correlation between the risk (protective) haplotype set and the SNPs consisting haplotypes.

[c] SNP ID is the SNP having the maximum correlation with the risk (protective) haplotype set.

[d] Empirical P value was obtained based on 1,000,000 permutations when stage 1 co-classification used the entire sample. This p-value refers the reported p values in the text.

Table 4

Risk haplotypes and the corresponding frequencies detected in T2D and HT

Risk haplotype	HT(%)	T2D(%)	Controls(%)
CCGCTAGCGATCTCACGTCGCGTGTGTCTC	0.10	0.10	0.00
CCGGTAGCGATCTCACGTCGCGTGTGTCTC	1.23	1.51	0.00
GTGCTAGCGATCTCACGTCGCGTGTGTCTC	1.20	1.38	0.00
GCAGGAACCGCCTTACGTTGTGTGTACGCC	0.10	0.00	0.00
GCAGGAACCGCCTTACTTCACCCGTACGCC	0.38	0.00	0.15
GCGGGAGCCGTATTACGTCACCCGTATGCC	1.33	0.00	0.95
CCGCGAGCGGTCTCAGGTTGTGCGTACGCC	0.13	0.00	0.02
CCGGGAGCCGTATTACGTCACCCGTATGCC	0.00	0.16	0.03