Presence of Multiple Independent Effects in Risk Loci of Common Complex Human Diseases

Xiayi Ke[1,*]

Many genetic loci and SNPs associated with many common complex human diseases and traits are now identified. The total genetic variance explained by these loci for a trait or disease, however, has often been very small. Much of the "missing heritability" has been revealed to be hidden in the genome among the large number of variants with small effects. Several recent studies have reported the presence of multiple independent SNPs and genetic heterogeneity in trait-associated loci. It is therefore reasonable to speculate that such a phenomenon could be common among loci known to be associated with a complex trait or disease. For testing this hypothesis, a total of 117 loci known to be associated with rheumatoid arthritis (RA), Crohn disease (CD), type 1 diabetes (T1D), or type 2 diabetes (T2D) were selected. The presence of multiple independent effects was assessed in the case-control samples genotyped by the Wellcome Trust Case Control Consortium study and imputed with SNP genotype information from the HapMap Project and the 1000 Genomes Project. Eleven loci with evidence of multiple independent effects were identified in the study, and the number was expected to increase at larger sample sizes and improved statistical power. The variance explained by the multiple effects in a locus was much higher than the variance explained by the single reported SNP effect. The results thus significantly improve our understanding of the allelic structure of these individual disease-associated loci, as well as our knowledge of the general genetic mechanisms of common complex traits and diseases.

Over the past few years, genome-wide association studies (GWASs) have been used for identifying a large number of common genetic loci for many common complex traits and diseases. These loci, however, contribute only a small proportion of the disease variability, leaving a large amount of disease heritability unexplained.[1] This so-called "missing heritability" issue has been partly demystified through methods that take into account genetic information of common variants accumulated across biological pathways[2] or across the entire genome among the large numbers of variants of small effects,[3,4] rather than just the individual confirmed disease-susceptibility loci. For example, with human height, a complex trait with an estimated heritability of 80%, it was shown that genome-wide information of common variants could explain 45% of heritability, whereas only 5% could be explained by the 50 confirmed associated loci at the time[3] and about 10% could be explained by the hundreds of variants clustered in genomic loci and biological pathways affecting the trait.[5]

This improved knowledge of "missing heritability," however, is far from sufficient in our understanding of the underlying biological mechanisms of a trait or disease. Knowledge of local allelic structure of individual associated loci is very important but is still lacking at the moment. It has been noticed that for the majority of individual associated loci, there is usually only a single common SNP or allele to be identified and reported. It is highly possible, however, that multiple independent effects could be present in a gene or locus that is associated with a trait. The major histocompatibility complex (MHC) region is well known to harbor multiple independent effects for autoimmune diseases, such as type 1 diabetes (T1D [MIM 222100])[6] and rheumatoid arthritis (RA [MIM 180300]).[7] Similar observations have also been reported for non-MHC regions, such as the *OLIG3* (MIM 609323)-*TNFAIP3* (MIM 191163) region, where multiple independent alleles were found to be associated with RA.[8,9] More recently, multiple independent effects were observed in 19 loci associated with human height[5] and in six loci associated with Crohn disease (CD [MIM 266600]).[10] Similar observation was also reported for expression quantitative trait loci (eQTLs).[11] It is highly possible that such phenomena are more prevalent among disease-associated genes and genetic loci than what has been reported in the literature. The lack of reports might be due to several limitations. First, the priority of past GWASs was often to identify the most significantly associated SNP for an individual locus at the GWAS stage and to confirm the association through replication. Second, the study samples might also have insufficient statistical power for confidently detecting independent associations of smaller effects. Third, the SNP coverage for many individual loci in early GWAS SNP arrays was very limited, thus reducing the chance of detecting independent effects.

The last limitation can be substantially improved through imputation with SNP genotype information provided by the HapMap Project[12] and the 1000 Genomes Project.[13] The second limitation can also be partially alleviated if a locus under investigation is already confirmed to be associated with a disease or trait. In this study, the presence of multiple independent effects was investigated in more than 100 known disease-associated loci with the use of the Wellcome Trust Case Control Consortium

[1]Medical Research Council Centre of Epidemiology for Child Health and Centre for Paediatric Epidemiology and Biostatistics, Institute of Child Health, University College London, 30 Guilford Street, London WC1N 1EH, UK
*Correspondence: x.ke@ucl.ac.uk
http://dx.doi.org/10.1016/j.ajhg.2012.05.020.

(WTCCC) samples for RA, T1D, T2D (type 2 diabetes [MIM 125853]), and CD.[14]

Approximately 2,000 RA, CD, T1D, and T2D disease samples (genotyped with Affymetrix Genechip 500 SNP chip in the phase I WTCCC study) and 6,000 control samples (genotyped by the WTCCC in its phase II study with Affymetrix v.6.0 chip) from the 1958 Birth Cohort study and British National Blood Service were downloaded from the European Genome-phenotype Archive. Samples and SNPs with low genotyping quality were removed as described by the WTCCC.[14] The final sample sets contained 1,860 RA patients, 1,748 CD patients, 1,963 T1D patients, and 1,924 T2D patients, as well as a total of 5,380 controls.

A list of approximately 30 disease-associated loci was obtained from the literature (mainly from published genome-wide meta-analyses) for RA,[15–17] CD,[18,19] T1D,[20–22] and T2D[23–26] (Table S1, available online). For CD, 23 loci were selected from the Barrett et al. meta-analysis,[18] in which specific genes of interest were assigned. A further six loci (*SCAMP3* [MIM 606913]-*MUC1* [MIM 158340], *THADA* [MIM 611800], *PRDM1* [MIM 603423], *ZFP36L1* [MIM 601264], *GALC* [MIM 606890]-*GPR65* [MIM 604620], and *TYK2* [MIM 176941]-*ICMA1* [MIM 147840]-*ICAM3* [MIM 146631]) with evidence of multiple independent associations from the more recent meta-analysis report by Franke et al.[10] were also selected. This resulted in a total of 117 loci. Next, the genotype data for the individual loci were extracted from the genome-wide data of the above samples on the basis of their National Center for Biotechnology Information build 36 coordinates as downloaded from Ensembl. If a reported disease-associated SNP was located outside the boundary of a locus (e.g., intergenic or downstream or upstream of a gene) as defined by Ensembl, the boundary of the locus was extended to include that SNP for the current study.

For each locus, imputation was carried out with software program IMPUTE2.[27] Reference data of samples of European descent were downloaded from the website of the authors of the software.[27] The reference data used contained 120 phased CEU (Utah residents with ancestry from northern and western Europe from the CEPH collection) haplotypes from the pilot 1000 Genomes Project and 1,910 phased haplotypes from the HapMap 3 CEU data. An effective size of 11,418 was used for imputation as recommended for the population of European descent. So that high quality could be ensured, all imputed genotypes with an information score below 0.90 and/or a minor allele frequency (MAF) $< 5\%$ (except for SNPs whose association was confirmed in the literature) were excluded. After quality control (QC), there were a total of 22,173 SNPs from the 117 loci associated with RA, CD, T1D, and T2D.

For each locus or gene, a frequentist test of additive effect for each SNP was carried out (after being adjusted for sex) with software program SNPTEST on the imputed dosage data. There were a total of 621 SNPs found to have an additive p value $\leq 5 \times 10^{-8}$, resulting in a subset of SNPs with unequivocal evidence of association with the corresponding diseases (Table S1). SNPs whose associations were confirmed in the literature were also added to this subset, resulting in a total of 721 SNPs unevenly distributed across the 117 loci (Table S1). Twenty-three of these loci (three RA-associated loci, seven CD-associated loci, seven T1D-associated loci, and six T2D-associated loci) were found to contain more than one SNP from the subset. After correction for multiple testing of all the 22,173 SNPs after QC, a p value threshold of 2.3×10^{-6} was used for filtering the rest of the SNPs. A further 19 loci also had more than one SNP that either was confirmed in the literature or met this threshold. This included seven CD-associated loci, four RA-associated loci, five T1D-associated loci, and three T2D-associated loci (Table S1). These 42 loci were examined for the presence of multiple effects.

SNPs in a locus were first screened under a penalized logistic regression model with the R package "penalized."[28] The following types of shrinkage were implemented in the package: an L1 absolute value ("lasso" [least absolute shrinkage and selection operator]) penalty,[29] an L2 quadratic ("ridge") penalty,[30] or a combination of the two (the "naive elastic net" of Zou and Hastie[31]). Analysis in this study was conducted with the L1 option because the lasso tends to produce fewer independent modes of the data and thus provides a more conserved estimate of the number of independent signals for a locus through shrinkage. The number of independent modes depends on the value of parameter λ. For loci with only a small number of SNPs, a λ value of 5 was chosen for the penalized regression model, whereas for loci with a large number of SNPs, a λ value of 10 or higher was used so that the number of candidate independent SNPs could be reduced. SNPs with non-zero coefficients in the penalized regression model were then passed onto a conditional logistic regression model for further examination. Under this model, SNPs were conditioned on each other iteratively so that a minimum set of independent SNPs could be obtained. Within this set, each SNP was required to have a minimum p value of 0.05 after being conditioned on other SNPs individually as well as collectively (i.e., the combined effects of all other SNPs if there were more than one). Conditional analysis was also carried out with SNPTEST, and conditional p values are listed in Tables 1 and 2.

First, the penalized and conditional regression-analysis procedures were applied to several disease-associated loci known to have multiple independent associations. This included the RA-associated locus *OLIG3-TNFAIP3*.[8,9] The independent effects of rs6920220 and rs10499194 were first reported by Plenge et al.[8] and later confirmed by Orozco et al.[9] in a different population sample of 3,962 RA patients and 3,531 healthy controls (in this latter study, rs13207033 was used as a surrogate for rs10499194[9]). In the present study, the independence between rs6920220 and rs10499194/rs13207033 was not significant under the regression analysis, indicating that the current sample set was not sufficiently powerful for detecting such

Table 1. Evidence of Multiple Independent Effects in Disease-Associated Loci under Penalized and Conditional Regression Analyses

Chr	Locus	Reported Associations	Independent SNPs (and p Values Conditioned on All Other SNPs)	Genomic Context of Independent SNPs
CD				
1	*IL23R* (MIM 607562)	rs11465804 (intron 7)	rs7517847 (3.7×10^{-15}), rs11465804 (3.7×10^{-15})	rs7517847 (intron 5), rs11465804 (intron 7)
5	*PTGER4* (MIM 601586)	rs4613763 (intergenic)	rs4613763 (8.7×10^{-5}), rs6888952 (1.1×10^{-9}), rs9283753 (0.0016)	rs4613763 (intergenic), rs6888952 (intergenic), rs9283753 (intergenic)
16	*NOD2* (MIM 605956)	rs2076756 (intron 8)	rs2076756 (3.1×10^{-13}), rs8056611 (0.00013)	rs2076756 (intron 8), rs8056611 (downstream)
T1D				
10	*IL2RA* (MIM 147730)	rs12251307 (intergenic)	rs12722495 (5.2×10^{-5}), rs7096384 (9.1×10^{-8})	rs12722495 (intergenic), rs7096384 (upstream)
16	*CLEC16A* (MIM 611303)	rs12708716 (intron 20)	rs7205474 (1.1×10^{-5}), rs2867880 (3.6×10^{-7})	rs7205474 (intron 1), rs2867880 (intron 23)
T2D				
12	*TSPAN8* (MIM 600769)-*LGR5* (MIM 6066670	rs4760790 (intergenic)	rs1705232 (1.8×10^{-11}), rs6581998 (0.0051)	rs1705232 (intergenic), rs6581998 (intergenic)

Only common SNPs with $p \leq 5 \times 10^{-8}$ in the current WTCCC data and those SNPs whose associations were confirmed in the literature were analyzed. The following abbreviations are used: chr, chromosome; CD, Crohn disease; T1D, type 1 diabetes; and T2D, type 2 diabetes.

a relationship. This suggests that similar independent effects in other disease-associated regions could also be missed as a result of insufficient power in the present samples, although it is also possible that this difference was due to allelic heterogeneity between study samples. In the present study, however, other independent SNP alleles were identified. At $p \leq 2.3 \times 10^{-6}$, the effect of rs5029926 ($p = 7.9 \times 10^{-7}$), which had a much stronger association than rs10499194 ($p = 0.0032$), was found to be significantly independent from that of rs6920220. In the Orozco et al. study,[9] an uncommon SNP (rs5029937; MAF < 5%) was reported to have an independent effect in the region. The effect of this SNP (MAF = 3.6% in the control samples) was indeed found to be independent from that of rs6920220 and rs5029926.

Five (*PRDM1*, *SCAMP3-MUC1*, *THADA*, *ZFP36L1*, and *TYK2-ICMA1-ICAM3*) of the six CD-associated loci with evidence of multiple independent effects[10] were initially excluded from the regression analyses as a result of the lack of SNPs meeting the p value thresholds (Table S1). The reported SNP rs8005161 in the *GALC-GPR65* locus had a p value of 2.18×10^{-5} for its association with the disease in the present samples. There were multiple other SNPs strongly associated ($p < 5 \times 10^{-8}$) with the disease. These SNPs were in linkage disequilibrium (LD) with each other, but their effects (as shown by that of SNP 14-87522091) were independent of the effect of rs8005161. Multiple effects were also observed in the *PRDM1* locus at a reduced p value threshold of 1×10^{-5}. The reported SNP rs6568421 was associated with the disease at $p = 7.09 \times 10^{-5}$, whereas SNP rs7746082 in the same locus was associated with the disease at $p = 8.14 \times 10^{-6}$. The effects of these two SNPs were found to be independent of each other, thus confirming the presence of multiple effects in the locus. For the other four loci, there was still a lack of SNPs in each locus even when the p value threshold was reduced to 1×10^{-3}. Therefore, no further tests were carried out.

There were two SNPs reported to be associated with T2D in the *CDKN2A* (MIM 600160)-*CDKN2B* (MIM 600431) locus; these were rs10965252 and rs7020996.[26] In the present study, associations of both SNPs were confirmed with p values at 6.78×10^{-8} (rs10965252) and 0.00029 (rs7020996). The independence of their effects was also confirmed in the regression analysis. These results from the present study were thus largely consistent with and render strong support to what had been reported in the literature.

Our primary concern was to examine whether additional loci with multiple independent effects could be identified. For the three RA-associated loci (*MMEL1-TNFRSF14* [MIM 602746], *PTPN22* [MIM 600716], and *KIF5A* [MIM 602821]) whose SNP associations met the p value threshold of 5×10^{-8} (Table S1), no independent effects were observed. Several CD-associated loci, however, were shown to have evidence of multiple effects. These included *IL23R* (MIM 607562), *PTGER4* (MIM 601586), and *NOD2* (MIM 605956) (Table 1). The reported associated SNPs for these loci were usually among the top associated SNPs (e.g., rs11465804 of the *IL23R* locus with $p = 5.94 \times 10^{-22}$ and rs4613763 of the *PTGER4* locus with $p = 1.08 \times 10^{-15}$) in the present study samples, and they were often found to represent one of the independent effects at the corresponding locus.

For T1D-associated loci, multiple independent effects were observed in the *IL2RA* and *CLEC16A* loci (Table 1).

Table 2. Evidence of Multiple Independent Effects in Disease-Associated Loci under Penalized and Conditional Regression Analyses

Chr	Locus	Reported Associations (p Value in Current Samples)	Independent SNPs (and p Values Conditioned on All Other SNPs)	Genomic Context of Independent SNPs
CD				
6	*CDKAL1*(MIM 611259)	rs6908425 (intron 3)	rs6908425 (2.9×10^{-5}), rs898165 (4.6×10^{-7})	rs6908425 (intron 3), rs898165 (intron 13)
12	*LRRK2* (MIM 609007)–*MUC19* (MIM 612170)	rs11175593 (intergenic)	rs7962370 (9.2×10^{-7}), rs11175593 (0.0025)	rs7962370 (intergenic), rs11175593 (intergenic)
T1D				
2	*AFF3* (MIM 601464)	rs9653442 (intergenic)	rs11685258 (3.0×10^{-5}), rs2309837 (2.1×10^{-5})	rs11685258 (intron 1), rs2309837 (intergenic)
4	*IL2* (MIM 147680)	rs4505848 (intron 18 of *KIAA1109* [MIM 611565])	rs4505848 (0.0059), rs13152362 (0.00029)	rs4505848, rs13152362 (intron 59 of *KIAA1109*)
T2D				
3	*PPARG* (MIM 601487)-*SYN2* (MIM 600755)	rs13081389[a] (intergenic) and rs17036101 (intergenic)	rs6775191 (5.3×10^{-6}), rs17036101 (0.017)	rs6775191 (intergenic), rs17036101 (intergenic)

Only common SNPs with additive $p \leq 2.3 \times 10^{-6}$ in the WTCCC data and those SNPs whose associations were confirmed in the literature were analyzed. The following abbreviations are used: chr, chromosome; CD, Crohn disease; T1D, type 1 diabetes; and T2D, type 2 diabetes.
[a]rs13081389 is in perfect LD ($r^2 = 1$) with rs17036101.

For both loci, however, the original reported SNPs were not represented as independent effects themselves but were replaced with other SNPs with stronger effects in the same directions. For example, rs12722495 in *IL2RA* was in a moderate level of LD ($r^2 = 0.65$) with the reported associated SNP rs12241307 but had a stronger effect in the present samples (odds ratios [OR] = 0.66 for rs12722495 versus 0.75 for rs12241307). A separate effect was identified with rs7096384, which was in low LD ($r^2 < 0.1$) with both rs12241307 and rs12722495 and whose minor allele conferred risk to the disease (OR = 1.23). For T2D-associated loci that met the 5×10^{-8} p value threshold, multiple independent effects were observed only in *TSPAN8-LGR5* (Table 1).

At the p value threshold of 2.3×10^{-6}, several more genetic loci with the presence of multiple effects were identified in the WTCCC samples (Table 2). These included *CDKAL1* and *LRRK2-MUC19*, both associated with CD; *AFF3* and *IL2-IL21*, both associated with T1D; and the T2D-associated locus *PPARG-SYN2*. Out of the 42 loci analyzed under the penalized and conditional regression models (33%) and the 117 loci surveyed in the study (12%), a total of 14 loci were observed to have evidence of multiple effects; these include the three loci (*OLIG3-TNFAIP3*, *GALC-GPR65*, and *CDKN2A-CDKN2B*) that have known multiple effects and that also met the p value thresholds.

Although the present study did not set out to identify the locations of the causal alleles, information about the physical locations and functional relevance of the multiple independent effects can help improve our understanding of the genetic mechanisms of the individual loci in relation to a disease. For example, there were two independent effects, represented by rs2076756 and rs8056611, identified in the *NOD2* locus (Table 1). The former was previously reported[10] and is located in intron 8, whereas the latter is located downstream of the gene. Although this does not mean that there were two separate causal variants located in exactly these two areas, the diversity of locations could indicate a multifaceted mechanism. Interestingly, rs8056611 was found to be in moderate LD (r2 ~ 0.6) with two eQTL SNPs in the eQTL browser,[32] i.e., rs3135499, a putative *cis*-eQTL SNP for *CARD15*, and rs10521209, an exon-QTL SNP for *CYLD* (cylindromatosis [turban tumor syndrome]), which locates immediately downstream of *NOD2*. Similarly, the reported T1D-associated SNP rs4505848 was near *IL2* but located in intron 18 of *KIAA1109* (MIM 611565). A separate effect (represented by rs13152362) for T1D was located in intron 59 of *KIAA1109* in the present study (Table 2), adding more evidence of the importance and complexity of the *KIAA1109-TENR-IL2-IL21* region to autoimmune diseases. For the T1D-associated *AFF3* locus, rs2309837 was located in its intergenic region, whereas rs11685258 was located in intron 1 and had a high conservation score of 766 as annotated in the Openbioinformatics's ANNOVAR Most Conserved Elements database. A high conservation score of 448 was also observed for rs9283753, one of the three independent SNPs in the CD-associated *PTGER4* locus; all three of these SNPs were located in its intergenic region.

One limitation of the present study is the modest sample size (approximately 2,000 cases versus 6,000 controls). This was exemplified by five of the six CD-associated loci (*PRDM1*, *SCAMP3-MUC1*, *THADA*, *ZFP36L1*, and *TYK2-ICMA1-ICAM3*), which had known evidence of multiple effects but failed the assessment in the present study simply as a result of the lack of associated SNPs meeting the minimum p value threshold. Undoubtedly, loci with multiple smaller effects could be identified at a reduced p value threshold at the cost of increased type I error. Despite

Table 3. Variance in Liability Explained by the Independent Effects as Identified in the Conditional Logistic Regression Analysis

Chr	Locus	Independent SNPs and Disease Variance Explained (in Liability Scale)[1]		Reported SNPs and Disease Variance Explained
CD (Prevalence = 0.1%)				
1	*IL23R* (MIM 607562)	rs11465804 (0.87%) rs7517847 (0.47%)	total: 1.34%	rs11465804 (1.12%)
5	*PTGER4* (MIM 601586)	rs4613763 (0.10%), rs6888952 (0.37%), rs9283753 (0.09%)	total: 0.56%	rs4613763 (0.35%)
6	*CDKAL1* (MIM 611259)	rs6908425 (0.13%), rs898165 (0.16%)	total: 0.28%	rs6908425 (0.14%)
12	*LRRK2* (MIM 609007)-*MUC19*)MIM 612170)	rs11175593 (0.04%), rs7962370 (0.14%)	total: 0.18%	rs11175593 (0.06%)
16	*NOD2* (MIM 605956)	rs2076756 (0.33%), rs8056611 (0.11%)	total: 0.44%	rs2067085 (0.25%)
		total: 2.66%		total: 1.92%
T1D (Prevalence = 0.5%)				
2	*AFF3* (MIM 601464)	rs11685258 (0.19%), rs2309837 (0.15%)	total: 0.34%	rs9653442 (0.15%)
4	*IL2* (MIM 147680)	rs4505848 (0.07%) rs13152362 (0.14%)	total: 0.21%	rs4505848 (0.16%)
10	*IL2RA* (MIM 147730)	rs12722495 (0.32%), rs7096384 (0.14%)	total: 0.46%	rs12251307 (0.20%)
16	*CLEC16A* (MIM 611303)	rs7205474 (0.25%), rs2867880 (0.26%)	total: 0.51%	rs12708716 (0.28%)
		total: 1.48%		total: 0.77%
T2D (Prevalence = 5.0%)				
3	*PPARG* (MIM 601487)-*SYN2* (MIM 600755)	rs17036101 (0.12%), rs6775191 (0.36%)	total: 0.48%	rs17036101 (0.20%)
12	*TSPAN8* (MIM 600769)-*LGR5* (MIM 606667)	rs1705232 (0.98%), rs6581998 (0.14%)	total: 1.12%	rs4760790 (0.47%)
		total: 1.54%		total: 0.67%

The following abbreviations are used: chr, chromosome; CD, Crohn disease; T1D, type 1 diabetes; and T2D, type 2 diabetes.
[1]Adjusted ORs were obtained with multiple logistic regressions for individual independent SNPs either within or across loci. Genotype relative risks were estimated according to a multiplicative model, and explained variance was estimated with the R software reported by So et al.[34] Total explained variance was the sum of such estimates from each of the individual SNPs either within or across loci.

such limitations, it was observed that a number of disease-associated loci surveyed in this study were found to have evidence of the presence of multiple independent effects. Although independent replications are needed for individual loci, the results demonstrate that the presence of multiple effects could be common among genetic loci associated with common complex diseases and traits.

The present study was focused on common SNPs (MAF $\geq$ 5%). Independent effects from uncommon or rare SNP alleles (MAF < 5%), e.g., rs5029937 in the *OLIG3-TNFAIP3* locus (associated with RA)[9] and rs35667974, rs35337543, rs35732034, and rs35744605 in the *IFIH1* locus (associated with T1D),[33] were already established and reported in the literature. rs5029937 (MAF = 3.6%) in the *OLIG3-TNFAIP3* locus was strongly associated ($p = 7.48 \times 10^{-8}$) with RA in the present study, and its effect was found to be independent of other independent common SNPs in the locus, as described above. For the *IFIH1* locus, only a single independent association with T1D, as represented by SNP rs1990760 ($p = 1.16 \times 10^{-5}$), was identified among the common SNPs, consistent with the previous report.[33] Two reported uncommon SNPs, rs35337543 (MAF = 2.4%) and rs35732034 (MAF = 1.3%), were strongly associated with the disease in the study samples ($p = 4.50 \times 10^{-8}$ for rs35337543 and $p = 5.73 \times 10^{-5}$ for rs35732034), and their effects were indeed found to be independent of that of rs1990760 (the p values after being conditional on rs1990760 were 2.89×10^{-6} for rs35337543 and 0.00025 for rs35732034), as well as independent of each other's effects. One other such example is rs11175593 (MAF = 1.2% in the control samples of the present study), which was found in the *LRRK2-MUC19* locus associated with CD.[17] The strength of this association was moderate in the current study samples ($p = 0.00046$). Common SNP rs7962370 (MAF = 12% in the control samples) was associated with the disease at $p = 2.84 \times 10^{-7}$, and effects of the two SNPs were found to be independent from each other (Table 2). It is expected that more such findings are likely to be made as sequencing data become available.

With the presence of multiple independent effects in disease-associated loci, the proportion of heritability explained by the known disease-associated loci would probably increase. For the assessment of such increases, all loci from Tables 1 and 2 were selected and heritability explained by individual effects was estimated with the algorithm developed by So et al.[34] under a disease-liability model. As shown in Table 3, in many of the loci, the heritability explained by the multiple independent SNPs was

substantially higher than that explained by the SNPs confirmed in the literature. As a result, there was also a substantial increase in the combined variance explained by all the loci with multiple effects for each disease (Table 3).

Although the total variance explained by GWAS-identified SNPs associated with a trait or disease is usually very small, it is now known that much of this so called "missing heritability" is actually hidden in the genome.[3,35] For traits like human height, it has been found that the additive genetic variance is spread across the genome and the variance explained by each chromosome is proportional to its length.[4] It has also been found that SNPs in or near genes explained more variation than did SNPs between genes, indicating a model of uniform distribution of trait variance on the physical scale, but not on the biological scale. The clustering of large numbers of variants in genetic loci and relevant biological pathways and the detection of multiple independent effects in 19 of these loci in association with height (as reported in a recent meta-analysis[5]) provide further evidence that genetic trait variance is clustered rather than uniformly distributed on the biological scale.

For other traits, distribution of the genetic variance might be different even on the physical scale. For example, for T1D, the variance explained by the genome-wide SNP genotype information in the WTCCC case-control samples was about 30%.[35] However, it was observed that for chromosome 6 alone, where the MHC locus locates, the variance explained was about 19%, whereas the rest of the genome together only explained about 13%.[35] Results from the present study further suggest that much of this 13% might be clustered around known associated loci in the form of multiple independent effects.

The presence of multiple independent associated alleles in a locus significantly increases its total contribution to genetic variance, as well as the total variance explained by all the known associated loci. It also highlights the importance and complexity of the issue of genetic heterogeneity. On the one hand, an independent effect observed in one study might fail to replicate in a different study as a result of genetic heterogeneity between the two study samples (e.g., as a result of a different composition of disease subtypes). On the other hand, the presence of multiple effects could be observed purely because of the presence of different associated alleles between subsets of disease samples in a study or between samples from different studies. With the rapid growth of next-generation-sequencing data, the allelic structure of disease-associated loci will be more refined. Such knowledge will undoubtedly significantly improve our understanding of the genetic and molecular mechanisms of common complex traits.

Supplemental Data

Supplemental Data include one table and can be found with this article online at http://www.cell.com/AJHG.

Acknowledgments

The UK Medical Research Council (MRC) provides funding to the MRC Centre of Epidemiology for Child Health. The author also wishes to thank the Wellcome Trust Case Control Consortium for providing the genotyping data in the study.

Received: January 9, 2012
Revised: April 13, 2012
Accepted: May 23, 2012
Published online: July 5, 2012

Web Resources

The URLs for data presented herein are as follows:

1000 Genomes Project, http://www.1000genomes.org/
ANNOVAR, http://www.openbioinformatics.org/annovar/
Ensembl, http://www.ensembl.org/
eQTL browser, http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/
European Genome-phenotype Archive, http://www.ebi.ac.uk/ega/
Genome Variation Server, http://gvs.gs.washington.edu/GVS131/index.jsp
Variance (or heritability) explained by genetic variants, http://sites.google.com/site/honcheongso/software/varexp
International HapMap Project, http://hapmap.ncbi.nlm.nih.gov/
IMPUTE2, http://mathgen.stats.ox.ac.uk/impute/impute_v2.html
Online Mendelian Inheritance in Man (OMIM), http://www.omim.org
SNPTEST, https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html
The R Project for Statistical Computing, http://www.r-project.org/
UCSC Genome Browser, http://genome.ucsc.edu
Wellcome Trust Case Control Consortium, http://www.wtccc.org.uk/

References

1. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.
2. Wang, K., Li, M., and Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. Am. J. Hum. Genet. *81*, 1278–1283.
3. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.
4. Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., Cunningham, J.M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M.G., et al. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. Nat. Genet. *43*, 519–525.
5. Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832–838.
6. Nejentsev, S., Howson, J.M., Walker, N.M., Szeszko, J., Field, S.F., Stevens, H.E., Reynolds, P., Hardy, M., King, E.,

Masters, J., et al.; Wellcome Trust Case Control Consortium. (2007). Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A. Nature *450*, 887–892.
7. Plenge, R.M., Cotsapas, C., Davies, L., Price, A.L., de Bakker, P.I., Maller, J., Pe'er, I., Burtt, N.P., Blumenstiel, B., DeFelice, M., et al. (2007). Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. Nat. Genet. *39*, 1477–1482.
8. Orozco, G., Hinks, A., Eyre, S., Ke, X., Gibbons, L.J., Bowes, J., Flynn, E., Martin, P., Wilson, A.G., Bax, D.E., et al.; Wellcome Trust Case Control Consortium; YEAR consortium. (2009). Combined effects of three independent SNPs greatly increase the risk estimate for RA at 6q23. Hum. Mol. Genet. *18*, 2693–2699.
9. Ding, B., Padyukov, L., Lundström, E., Seielstad, M., Plenge, R.M., Oksenberg, J.R., Gregersen, P.K., Alfredsson, L., and Klareskog, L. (2009). Different patterns of associations with anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis in the extended major histocompatibility complex region. Arthritis Rheum. *60*, 30–38.
10. Franke, A., McGovern, D.P., Barrett, J.C., Wang, K., Radford-Smith, G.L., Ahmad, T., Lees, C.W., Balschun, T., Lee, J., Roberts, R., et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat. Genet. *42*, 1118–1125.
11. Wood, A.R., Hernandez, D.G., Nalls, M.A., Yaghootkar, H., Gibbs, J.R., Harries, L.W., Chong, S., Moore, M., Weedon, M.N., Guralnik, J.M., et al. (2011). Allelic heterogeneity and more detailed analyses of known loci explain additional phenotypic variation and reveal complex patterns of association. Hum. Mol. Genet. *20*, 4082–4092.
12. Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al.; International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.
13. 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.
14. 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.
15. Raychaudhuri, S., Thomson, B.P., Remmers, E.F., Eyre, S., Hinks, A., Guiducci, C., Catanese, J.J., Xie, G., Stahl, E.A., Chen, R., et al.; BIRAC Consortium; YEAR Consortium. (2009). Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. Nat. Genet. *41*, 1313–1318.
16. Gregersen, P.K., Amos, C.I., Lee, A.T., Lu, Y., Remmers, E.F., Kastner, D.L., Seldin, M.F., Criswell, L.A., Plenge, R.M., Holers, V.M., et al. (2009). REL, encoding a member of the NF-kappaB family of transcription factors, is a newly defined risk locus for rheumatoid arthritis. Nat. Genet. *41*, 820–823.
17. Stahl, E.A., Raychaudhuri, S., Remmers, E.F., Xie, G., Eyre, S., Thomson, B.P., Li, Y., Kurreeman, F.A., Zhernakova, A., Hinks, A., et al.; BIRAC Consortium; YEAR Consortium. (2010). Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat. Genet. *42*, 508–514.
18. Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., Brant, S.R., Silverberg, M.S., Taylor, K.D., Barmada, M.M., et al.; NIDDK IBD Genetics Consortium; Belgian-French IBD Consortium; Wellcome Trust Case Control Consortium. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat. Genet. *40*, 955–962.
19. McGovern, D.P., Jones, M.R., Taylor, K.D., Marciante, K., Yan, X., Dubinsky, M., Ippoliti, A., Vasiliauskas, E., Berel, D., Derkowski, C., et al.; International IBD Genetics Consortium. (2010). Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. Hum. Mol. Genet. *19*, 3468–3476.
20. Todd, J.A., Walker, N.M., Cooper, J.D., Smyth, D.J., Downes, K., Plagnol, V., Bailey, R., Nejentsev, S., Field, S.F., Payne, F., et al.; Genetics of Type 1 Diabetes in Finland; Wellcome Trust Case Control Consortium. (2007). Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat. Genet. *39*, 857–864.
21. Cooper, J.D., Smyth, D.J., Smiles, A.M., Plagnol, V., Walker, N.M., Allen, J.E., Downes, K., Barrett, J.C., Healy, B.C., Mychaleckyj, J.C., et al. (2008). Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. Nat. Genet. *40*, 1399–1401.
22. Barrett, J.C., Clayton, D.G., Concannon, P., Akolkar, B., Cooper, J.D., Erlich, H.A., Julier, C., Morahan, G., Nerup, J., Nierras, C., et al.; Type 1 Diabetes Genetics Consortium. (2009). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat. Genet. *41*, 703–707.
23. Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., Willer, C.J., Li, Y., Duren, W.L., Erdos, M.R., Stringham, H.M., Chines, P.S., Jackson, A.U., et al. (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science *316*, 1341–1345.
24. Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., Marchini, J.L., Hu, T., de Bakker, P.I., Abecasis, G.R., Almgren, P., Andersen, G., et al.; Wellcome Trust Case Control Consortium. (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.
25. Qi, L., Cornelis, M.C., Kraft, P., Stanya, K.J., Linda Kao, W.H., Pankow, J.S., Dupuis, J., Florez, J.C., Fox, C.S., Paré, G., et al.; Meta-Analysis of Glucose and Insulin-related traits Consortium (MAGIC); Diabetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium. (2010). Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. Hum. Mol. Genet. *19*, 2706–2715.
26. Voight, B.F., Scott, L.J., Steinthorsdottir, V., Morris, A.P., Dina, C., Welch, R.P., Zeggini, E., Huth, C., Aulchenko, Y.S., Thorleifsson, G., et al.; MAGIC investigators; GIANT Consortium. (2010). Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat. Genet. *42*, 579–589.
27. Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. *5*, e1000529.
28. Goeman, J.J. (2010). L1 penalized estimation in the Cox proportional hazards model. Biom. J. *52*, 70–84.
29. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J Royal Stat Soc Series B Stat Methodol. *58*, 267–288.
30. Hoerl, A.E., and Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics *12*, 55–67.

31. Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. J. R. Stat. Soc. Series B Stat. Methodol. *67*, 301–320.
32. Veyrieras, J.B., Kudaravalli, S., Kim, S.Y., Dermitzakis, E.T., Gilad, Y., Stephens, M., and Pritchard, J.K. (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet. *4*, e1000214.
33. Nejentsev, S., Walker, N., Riches, D., Egholm, M., and Todd, J.A. (2009). Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science *324*, 387–389.
34. So, H.C., Gui, A.H.S., Cherny, S.S., and Sham, P.C. (2011). Evaluating the heritability explained by known susceptibility variants: A survey of ten complex diseases. Genet. Epidemiol. *35*, 310–317.
35. Lee, S.H., Wray, N.R., Goddard, M.E., and Visscher, P.M. (2011). Estimating missing heritability for disease from genome-wide association studies. Am. J. Hum. Genet. *88*, 294–305.