Published in final edited form as:

J Mol Biol. 2015 July 3; 427(13): 2271–2289. doi:10.1016/j.jmb.2015.04.014.

Genetic basis of common human disease: Insight into the role of Missense SNPs from Genome Wide Association Studies

Lipika R. Pal¹ and John Moul^{1,2,*}

¹Institute for Bioscience and Biotechnology Research, University of Maryland, 9600 Gudelsky Drive, Rockville, MD 20850

²Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20742

Abstract

Recent genome wide association studies have led to the reliable identification of SNPs at a number of loci associated with increased risk of specific common human diseases. Each such locus implicates multiple possible candidate SNPs for involvement in disease mechanism. A variety of mechanisms may link the presence of a SNP to altered *in vivo* gene product function and hence contribute to disease risk. Here we report an analysis of the role of one of these mechanisms, missense SNPs (msSNPs) in proteins in seven complex trait diseases. Linkage disequilibrium information was used to identify possible candidate msSNPs associated with increased disease risk at each of 356 loci for the seven diseases. Two computational methods were used to estimate which of these SNPs has a significant impact on *in vivo* protein function. 69% of the loci have at least one candidate msSNP and 33% have at least one predicted high impact msSNP. In some cases, these SNPs are in well-established disease related proteins, such as MST1 (macrophage stimulating 1) for Crohn's disease. In others, they are in proteins identified by GWAS as likely candidates for disease relevance, but previously without known mechanism, such as ADAMTS13 (ADAM metalloproteinase with thrombospondin type 1 motif, 13) for Coronary Artery Disease. In still other cases, the missense SNPs are in proteins not previously suggested as disease candidates, such as TUBB1 (tubulin, beta 1, class VI) for Hypertension. Together, these data support a substantial role for this class of SNPs in susceptibility to common human disease.

Graphical abstract

*Corresponding author: ; Email: jmoul@umd.edu, Phone: (240) 314-6241, FAX: (240) 314-6255

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

GWAS; SNPs; complex trait disease; missense; protein structure

INTRODUCTION

Microarray based Genome Wide¹ Association Studies (GWAS) have transformed our knowledge of the genetic basis of common human diseases. For many common diseases, dozens of genetic loci where the presence of one or more marker SNPs (Single Nucleotide Polymorphisms) is associated with disease risk have been identified. For some diseases, combination of data from multiple studies (meta-analyses) has resulted in over a hundred risk loci being identified, for example, 163 loci for Inflammatory Bowel Disease [1]. Each such locus must harbor at least one underlying mechanism – a genetic variant that in some way alters the *in vivo* function of a molecular level process so as to affect disease risk. Individual effects on disease risk are small, and it is the combination of variants in many loci together with environmental factors that result in a disease phenotype. Identification of these mechanism variants and the affected processes is essential for exploiting GWAS results to improve understanding of overall disease mechanism, and for deriving new therapeutic strategies.

While GWA studies identify the approximate location of disease related variants, they rarely pinpoint exactly which variant is involved in mechanism nor do they provide direct information on those mechanisms. A current typical microarray in GWAS contains complementary oligonucleotides for only 500,000 to a million common SNPs, while the number of known SNPs down to a frequency of about 1% is more than 40 million [2]. GWAS sparse sampling of the total set of variants is effective because of correlation between the presence of a particular SNP and other variants up to about 200 Kilobases (Kb) away, as a consequence of incomplete recombination across genomes within the human population. The Hapmap [3] and 1000 genomes [4] projects have provided extensive data on this linkage disequilibrium (LD), so that it is now possible to consider the potential mechanistic role of all SNPs in a disease-associated locus, extrapolating from the GWAS observations.

¹Abbreviations used: Genome wide association study: GWAS; Linkage disequilibrium: LD; Single nucleotide variant: SNV; Single nucleotide polymorphism: SNP; Human Gene Mutation Database: HGMD; Support vector machine: SVM; minor allele frequency: maf; Caucasian: CEU

SNPs found to be associated with disease risk in a GWA study (marker SNPs) and SNPs in strong LD with these (those with a high correlation between the presence of a candidate and the presence of the marker) are clear candidates for involvement in disease mechanism. SNPs in weaker LD with markers may also be candidates, and there are reasons to expect many mechanism SNPs will be among these. First, weak LD is often a consequence of relatively low frequency of the candidate SNP, and low frequency SNPs are most likely to have a fitness impact, and so to be involved in disease mechanism [5,6]. Second, as shown later, there are many candidates in relatively weak LD with markers, so that on simple statistical grounds, mechanism SNPs are more likely to be among this set.

The range of LD varies widely across the genome, but a typical single locus discovered in GWAS may contain thousands of candidate mechanism SNPs. As noted earlier, it is likely that just one of those is directly implicated in a disease related mechanism. It is also possible that none of the candidate SNPs is in fact the mechanism variant – the causal variant may be a rare single base variant [7], an INDEL [8], or a copy number variant [9]. Nevertheless, as illustrated in this study, there is evidence that common SNPs are involved in a significant number of underlying mechanisms.

Given the full set of genotypes for all participants in a study, imputation methods [10–13] allow the evaluation of risk status for each candidate SNP, using the data for those experimentally observed (the tag SNPs). When many GWA studies are included in an analysis, as in this work, obtaining genotype data for all studies is impractical, because of the hurdles in obtaining access to the data, necessitated by privacy considerations. We employ an alternative procedure for identifying candidates for involvement in disease mechanism that does not require full access to the genotype data, allowing much more facile use of the extensive information available through the GWAS catalog [14]. We compare the method with imputation results for one study, the WTCCC1 seven disease study [15], and then use it to identify candidate SNPs in all studies of those diseases in the GWAS catalog [14].

About 5% of GWAS loci included in this study are in so-called gene deserts, but the large majority encompass known protein encoding genes, and so it is likely that in most cases disease mechanism involves change of *in vivo* activity of a gene product. A SNP may perturb the function of a gene product through a range of mechanisms, including transcription factor binding; miRNA interactions; messenger RNA splicing, structure and half-life; translation efficiency; and non-synonymous substitution effects. In this paper we explore the role of SNPs that result in an amino acid substitution – missense SNPs (msSNPs). These missense variants may alter *in vivo* activity of the relevant protein by affecting folding, ligand binding, catalysis, allosteric regulation, localization, post-translational modification, or aggregation and half-life.

A number of computational methods have been developed to estimate the *in vivo* effects of amino acid substitutions resulting from single base changes [16–22]. Most methods, for example, SIFT [22], Evolutionary Trace (ET) [16] and SNPs3D profile [23], use sequence profile and phylogenetic information to identify substitutions that affect fitness, and are in this sense deleterious. Methods are trained on a variety of disease related variants and

controls, particularly variants causing monogenic disease. For this class of disease there is evidence that causative mutations have a high impact on *in vivo* protein function, typically in excess of five fold [24,25]. The sequence based methods have the advantage of detecting all effects on function, providing there is an adequate sequence profile available for the protein domain, and the change in function affects fitness. They have the disadvantage of not providing any direct insight into the specific molecular mechanism involved. A few methods make use of protein structure to directly identify high impact variants that affect protein stability, for example SNPs3D structure [26]. Use of this method to study monogenic disease [26] and cancer [27] found the majority of mutations to result in a change in protein thermodynamic stability, implying a role in folding or half-life. One of the most popular methods, Polyphen2 [18] makes use of a combination of profile and structure information. Although structure based methods provide information on a specific molecular mechanism, they are restricted to cases where three-dimensional structure is available.

Here, we have used two methods incorporated in the SNPs3D resource [28] to identify high impact missense SNPs. As noted above, one method detects effects on protein function based on the level of amino acid conservation at a substitution position [23], and the other uses the three-dimensional structural context of a substitution to estimate effects on thermodynamic stability, related to changes in folding and half-life [26]. 86% of missense variants in the dbSNP database, dbSNP137 [2], could be analyzed with the profile method and 25% by structure method, either using direct experimental structural coverage or a sufficiently accurate comparative model (one based on 40% sequence identity or better [26]). To address the issue of false positives, we also compared the SNPs3D high impact assignments to those from two other methods, SIFT [22] and Polyphen2 [18].

Missense single base changes play a major role in monogenic disease [29], and usually have a high impact on protein function [25]. Missense mutations, usually somatic, also play a major role in cancer [30], and application of the computational methods provides evidence that these also often have a high impact on protein function [27,31]. The role of high impact missense variants in common, complex trait, disease is less well established. Unlike monogenic disease and cancer, each disease locus makes only a small contribution to the disease phenotype (a small increase in disease risk, for example), so that subtle effects at the protein level may be expected to play a role. Indeed, there is now strong evidence that the largest role is played by variants that in some way affect protein expression [32-34], and these usually are not more than two fold in size of impact on protein function. Subtle effects at the phenotype level do not necessarily imply subtle effects at the protein level. The phenotype effect of a variant depends on two factors: the impact on protein function, and how tightly that function is coupled to the phenotype. Mouse knockout data for proteins involved in complex trait disease provide evidence that in many cases, complete loss of protein function results in small changes of phenotype. For example, for blood pressure, knockout of even the genes most centrally involved (ACE, REN, and others) produces a change of not more than 30 mm of Hg [35]. For most blood pressure related genes, the knockout effect is much smaller. In such cases, a variant must have a high impact on protein function to have a detectable impact on the phenotype. Although knockout data provide many examples of weak coupling to a complex trait phenotype, the information is very

incomplete, and the extent to which high impact variants play a role in common disease is not yet known.

Application of computational methods to common human missense SNPs predicts a approximately a fifth to a quarter to be deleterious, implying high impact on protein function (for the 16,182 missense SNPs with a frequency > 5% in the 1000 genomes data [4] in the Caucasian population, 22% are predicted high impact by SNPs3D [23,36] and 24% by Polyphen2 [18]). In most cases, these have no known relationship to disease. The large fraction of high impact msSNPs found in the human population suggests the possibility of a major role in complex disease. The advent of GWAS data now makes it possible to assess the role of these SNPs. We focus on predicted high impact missense SNPs, since the computational methods are able to identify these. Since smaller impact at the protein level can also contribute to complex trait disease, lower impact missense must also contribute. We estimate an upper limit for their role, although many will be effectively neutral and not affect disease risk.

The analysis presented here provides the first estimates of the extent of involvement of missense SNPs in general and high impact ones in particular. For each locus, we determined which missense SNPs are candidates for involvement in disease mechanism and evaluated which of these are expected to have a high impact on *in vivo* protein function. Analysis of the results reveals a wide range of implicated protein functions. Missense SNPs in a small number of proteins have already been suggested as involved in disease mechanism. A number of the proteins have previously identified as candidates for involvement in disease, usually through GWAS, but a missense mechanism has not been suggested. An additional set of proteins with predicted high impact missense SNPs have not previously been suggested as disease candidates. Altogether, putative causal missense SNPs are proposed for 118 of the 356 loci considered. Generally, follow-up experimental studies are required to confirm these mechanisms. Overall, the results show a significant role for missense mechanisms in complex trait disease.

RESULTS

Loci associated with altered risk for the WTCCC1 seven diseases

The analysis was performed on risk loci for seven complex trait diseases, using data from the first large scale genome wide association study by the Wellcome Trust Case Control Consortium [15] and follow-up studies. The seven complex diseases are Bipolar Disorder (BD), Coronary Artery Disease (CAD), Crohn's Disease (CD), Hypertension (HT), Rheumatoid Arthritis (RA), Type 1 Diabetes (T1D) and Type 2 Diabetes (T2D).

The original WTCCC1 seven disease GWA study identified 21 loci where there is a strong association ($P\text{-value} < 5.0 \times 10^{-7}$) between the presence of at least one, usually more SNPs and altered disease risk for one or more of these seven diseases. Subsequent studies and meta-analyses collected in the GWAS catalog [37] have identified a further 335 loci with $P\text{-values} < 1.0 \times 10^{-5}$. A list of included studies from the GWAS catalog is provided in Supplementary Table 1. The number of loci for each disease ranges from 17 for

Hypertension to 90 for Crohn's disease (Table 1). A complete list of all included loci is given in Supplementary Table 2.

A representative marker SNP was selected in each of the 356 loci. For loci with several markers, reported by different studies, the one with the highest sample size was selected. Increased disease risk may be associated with either the minor or the major allele of a marker SNP. Overall the risk alleles of these markers are approximately equally divided between major and minor (48% major and 52% minor). Supplementary Table 3 shows the risk allele type distribution for the selected markers for each disease.

Selection of candidate SNPs for involvement in disease mechanism

We assume that each disease-associated locus contains some variant or variants that directly influence the *in vivo* function of a gene in that region, and identify the set of SNPs that could be causative in this sense, based on LD relationships to the relevant marker. As noted earlier, because of sparse sampling, these causal variants are rarely represented directly on the genotyping microarray used in a study. All SNPs within 200Kb of a marker and with linkage disequilibrium $r^2 > 0$ were assessed as potential candidates for involvement in disease mechanism, a total of 588,751 SNPs across 356 loci. (Some SNPs are candidates in more than one disease, and the total number of unique SNPs involved is 498,536).

LD relationships vary across populations. A high fraction of the included GWA studies have been conducted on Caucasian populations, and so we made use of the CEU 1000genomes data (87 individuals) [4] and the CEU Hapmap (120 individuals) [3] LD data. The small number of individuals on which the LD data is based results in unreliable LD values for SNPs with less than 5% maf (minor allele frequency). 41% of potential candidates cannot be analyzed for this reason. Additionally, for weaker LD relationships, even at higher frequencies, it is not always possible to accurately calculate the implied case risk allele frequency (see Methods). A further 13% of potential candidates fall into this category.

Thus, altogether, 46% of the potential candidates can be evaluated with current LD data. Each of the accepted candidates was examined to ascertain whether or not the LD relationship to the associated marker is sufficiently strong to generate the observed frequency difference between case and control for the marker. A total of 235,253 fulfill this condition and so are potentially causative. Supplementary Table 2 shows the number of candidate SNPs for all diseases.

Risk alleles for the accepted candidate SNPs show a stronger bias towards minor than found for the marker SNPs (42% major and 58% minor for candidates versus 48% major and 52% minor for markers). Supplementary Table 4 shows the values for each disease.

Candidate missense SNPs and high impact missense SNPs

Among the 235,253 qualifying candidate SNPs, there are 1259 unique missense SNPs, occurring in 540 unique proteins, and 69% (244) of the 356 loci contain at least one candidate missense SNP. Each of these candidates provides a possible mechanism underlying the corresponding locus. The missense SNPs may be essentially neutral, or have a mild effect on the *in vivo* function of the corresponding protein, or have a high impact on

function. The two methods in the SNPs3D analysis package [26,23] were used to identify putative high impact SNPs. As noted earlier, one method specifically looks only for effects on stability of the protein three-dimensional structure [26]. The other method uses the profile of amino acids found at the substitution position within the protein family as a basis for assigning impact, and in principle will detect any SNP that has a high impact on protein function [23]. Of the 1259 unique missense candidate SNPs, 1065 are amenable to analysis, using either the profile (84% of candidates) or stability (27%) methods or both. Profile coverage is less than 100% because of limited availability of sufficiently deep sequence alignments. Low coverage for stability analysis reflects current experimental structural data for human proteins, together with models based on 40% or higher sequence identity [26]. A total of 323 unique missense SNPs, in 175 different proteins are assigned as high impact, corresponding to 30% of those tested, higher than the rate for all missense SNPs noted earlier (22%). Some SNPs occur in loci common to two or more of the seven diseases considered, so that these 323 predicted high impact missense SNPs provide a total of 432 putative mechanisms, in a total of 118 of the 356 loci.

Figure 1 shows the fraction of loci with missense SNPs for the seven diseases, and Supplementary Table 5 gives the full data. The fraction of loci with at least one candidate missense SNP varies across diseases from 58 – 80%. Strikingly, the fraction of loci with a high impact candidate mechanism SNP varies considerably across the diseases, from 46% for the 90 loci of Crohn's disease down to 19% for the 43 loci in Type 2 diabetes. Supplementary Figure 4 shows the distribution of the number of proteins containing missense and high impact missense SNPs in each locus. Most loci have only a single protein containing a missense or high impact missense SNP.

Six out of these seven diseases have marker SNPs in the MHC region of chromosome 6, and that locus has a number of unusual properties. Very extensive LD results in a single marker producing a large number of candidate SNPs (375 of the total 1259) in 48 proteins, so that it is not possible to identify mechanism. On the other hand, because there are many well studied proteins in this locus, particularly MHC proteins. As a result, structural coverage of these SNPs is quite high (57%). However, there are a number of instances of MHC isoforms with multiple co-dependent (epistatic within the same protein) variants. For example, in two isoforms (NP_002114, NP_001230890) of HLA-DQB1, substantial steric clashes are created by a missense SNP substitution L58F. But in another isoform (NP_001230891), the presence of two other substitutions, T60S and E106A, allow a PHE at position 58 to be accommodated without strain. It may be that in the general population, this SNP occurs mostly in only the latter isoform, and so is neutral with respect to stability. For these reasons, we exclude MHC region SNPs from some of the analysis. (We have found that epistatic effects between missense variants within the same protein are rare in other classes of proteins (Shi and Moulton, unpublished)).

We compared the SNPs3D high impact assignments with those of SIFT and Polyphen2. In 59% of the instances where SNPs3D assigns high impact to a missense SNP, one or both of the other methods also does so. Agreement between the methods increases the higher the SNPs3D high SVM score (corresponding to more robust assignments), and is 74% for the 34% of scores greater than 1.0. Supplementary Table 6 shows the agreement between the

methods and all predictions for all high impact missense SNPs are given in Supplementary Table 7.

Allele properties of high impact candidate missense SNPs

As noted earlier, over all candidate SNPs, we found that risk alleles are more commonly minor (42% major and 58% minor) compared to markers. Over the subset of all missense SNPs this trend increases slightly, to 39% major, 61% minor and is most pronounced for those classified as high impact: 34% major and 66% minor.

Figure 2 shows the distribution of LD parameters (D' and r^2) of the high impact SNPs. While almost half of these SNPs (48%) have high D' values (> 0.8), the majority have low r^2 (75% ≤ 0.2), and would not be included as candidates using a high r^2 criterion. Nevertheless, these are robust candidates by the criteria used in this analysis, and should not be ignored.

Figure 3 shows the distribution of implied disease associated odds ratios for these high impact candidates. 51% (220) odds ratios are less than 2, and only 19% (82) are greater than 5 (most of these are in the MHC region, and only 16 are from the other loci). These values are larger than the odds ratios for the markers, where 351 out of 356 are less than 2, as expected when the LD between the putative causative SNP and the marker is less than complete. Odds ratio for the full set of candidate SNPs are on average higher than for the set of high impact ones (Supplementary Figure 1).

Role of destabilization of protein structure

As noted earlier, the SNPs3D stability method putatively identifies those missense SNPs that appear to destabilize the three dimensional structure of the affected protein. Stability calculations were possible for 334 missense SNPs, and of these 115 are predicted high impact. Detailed analysis for the 35 of these in non-MHC region proteins shows a variety of destabilization mechanisms. Most common is loss of salt bridges, affecting 10 cases. These often involve other electrostatic effects as well, such as loss of hydrogen bonds. There are seven cases of over-packing, five cases with substantial reduction in burial of hydrophobic area, and four instances of introduction of backbone strain, either through change of a GLY to some other residue or introduction of a PRO.

The subset of variants for which structure analysis was possible occur in proteins with significantly higher *in vivo* maximum expression compared to the larger set with a profile analysis (mean $\log_2(\text{maximum expression})$ of 6.3 versus 8.1, paired T test $P = 0.0098$), probably reflecting the fact that the higher the expression level of a protein, the more likely it is to be successfully expressed/have its structure determined. There is also a marginally significant higher mean number of protein interaction partners for proteins with structure (mean $\log_{10}(\text{partners})$ of 1.00 versus 0.87, paired T test $P = 0.046$), likely reflecting the fact that proteins with known partners are of more general interest and so more studied. We did not find any significant difference in the profile SVM scores for the two sets, suggesting no overall difference in the severity of impact in protein function; or in the minor allele frequency, or in whether the proteins concerned had already been implicated in the relevant

disease. There is a difference in the fraction of cases for which the minor allele is the risk variant (67% of those with structure versus 58% for the profile set, $P=0.0007$).

Where a missense SNP is assigned high impact by the SNPs3D profile method, but assigned low impact by the stability method, some mechanism other than structural destabilization is implied. Such SNPs may affect a variety of possible functions including binding to another protein or small molecule ligand, catalysis (in the case of enzymes), allosteric mechanisms, post-translational modification and transport. Out of 76 high impact profile assignments that also have structure assignments, 48 have low impact stability assignments. In some of these cases, it is possible to suggest a specific mechanism based on structural data and consulting the literature, and examples are discussed below.

Function of proteins containing high impact potential mechanism msSNPs and relevance to disease mechanism

The presence of a high impact candidate SNP in a locus does not necessarily mean it is actually involved in complex disease mechanism – the SNP may incidentally affect the function of a protein not relevant to the disease. Thus each identified candidate represents evidence for a possible mechanism, to be considered together with other data, such as the function of the protein concerned and suggestions from other GWA studies. On this basis, we divide the 227 disease specific proteins containing high impact SNPs (175 different proteins, some involved in more than one disease) into three primary categories: (A) Four proteins where a missense mechanism has already been recognized for the corresponding diseases; (B) 104 proteins already proposed as candidate genes for involvement in disease mechanism through GWA and other studies, but for which the molecular mechanism has yet to be established; and (C) 77 proteins without a proposed relationship with any of these diseases. The 42 MHC region proteins are categorized separately (D) since, as noted earlier, the locus has a well-established connection with the immune component of the diseases concerned, but the very high LD makes assignment of specific mechanism difficult in many cases. Below, we describe some examples for the three primary categories, and Figure 4 shows the structural context for some of these. With the exception of PTPN22 (protein tyrosine phosphatase non-receptor type 22), at least one other method – SIFT and/or Polyphen2 – also predicts the missense SNP is high impact. For PTPN22, there is experimental evidence of impact. All candidate SNPs have minor risk alleles except where noted. The full list of protein categories is given in Supplementary Table 8.

(A) Proteins containing high impact missense SNPs with already established data on a missense mechanism—In Crohn's disease, MST1 (macrophage stimulating 1) on chr3: 49.50–49.90 Mb contains a missense SNP, approximate 20Kb from the marker SNP and in high LD with it. The SNP introduces a R703C substitution in the corresponding protein with a SNPs3D profile assignment of high impact and a low impact stability score, indicating a functional role other than destabilization. We had previously identified this candidate SNP and our subsequent experimental studies have shown that this variant results in reduced affinity between MST1 and the cell surface receptor RON [38]. Figure 4a shows the structure of the complex, with the R703C substitution lying close to the protein-protein interface. Consistent with this, SPR binding studies show an approximately five-fold

reduction in protein-protein affinity in the presence of the SNP. Thus, the experimental data strongly suggest that lower strength of interaction between MST1 and RON is the underlying molecular mechanism by which disease risk is increased via this locus. RON is involved in a macrophage stimulation pathway, and some Crohn's patients exhibit low macrophage activity [39], consistent with this contribution to disease risk. Interestingly, a recent experimental study suggests the missense SNP also affects expression [40].

A second Category A protein that is also implicated in Crohn's disease is NOD2 (nucleotide-binding oligomerization domain containing 2) on chr16: 50.59–50.92 Mb. Here the candidate is about 6Kb from the marker SNP, with $D' = 1.0$ and $r^2 = 0.14$ (the low r^2 reflects the substantial difference in maf of the marker (0.37) and the candidate (0.07)). The SNPs3D profile score for the resulting R702W is borderline deleterious (−0.04) but both SIFT and Polyphen2 assign high impact (damaging and probably damaging respectively). NOD2 is an innate immune recognition factor that interacts with muramyl dipeptide fragments of bacterial cell walls, so playing a role in the innate immune response [41]. This variant and two other coding variants (G908R and a frame-shift insertion, L1007fsinsC) have previously been noted as directly associated with Crohn's disease risk [42]. The G908 substitution is rare (maf 0.03) and so was not included in our analysis. These authors suggested that the variants in some way disrupt structure, and so interfere with interaction with the muramyl dipeptide, inhibiting recognition of bacteria/pathogens and contribute towards disease risk. There is currently no adequate structural model to assess that, but the profile methods do support a high impact role of some kind for this variant.

Another Category A protein involved in Type 1 diabetes, as well as other autoimmune diseases including Rheumatoid arthritis, is PTPN22 (protein tyrosine phosphatase non-receptor type 22), chr1: 114.1–114.5 Mb. A missense SNP, R620W, in high LD with the marker and 73Kb away, has a high impact assignment by the profile method. There is no structure assignment available. PTPN22 is a negative regulator of T-cell activation and the risk variant tryptophan is associated with lower β -cell function for Type 1 diabetes [43]. This variant has been shown to disrupt the interaction of the P1 proline rich motif with the SH3 binding domain of Csk (c-src tyrosine protein kinase) [44] (Figure 4b), reducing the affinity by approximately 2.6–2.9 fold, thereby reducing the ability of PTPN22 to down-regulate T cell activation [44, 45]. Here, then, the missense SNP appears to act by weakening a protein-protein association.

(B) Previously proposed disease related proteins, now found to contain a high impact missense SNP—The 104 proteins in this category have previously been suggested as candidates for involvement in disease mechanism from GWAS or other studies. These candidate proteins were selected by study authors using a variety of criteria such as functional relevance, some times using tools such as GRAIL [46], or a known eQTL relationship, or some times simply that the protein reading frame is nearest to the marker SNP, or in some cases using 1000 genomes data [4] to identify genes with non-synonymous variants that are in strong LD with the disease markers [47]. However, specific missense SNP mechanisms have not previously been proposed. Some examples are:

For Coronary Artery Disease, on the chr9: 135.95–136.35 Mb locus, a msSNP predicted as high impact by the profile method, and low impact by the structure method, produces a P618A substitution in ADAMTS13 (ADAM metalloproteinase with thrombospondin type 1 motif, 13). This residue is in the $\beta 6$ – $\beta 7$ loop of the spacer domain (Figure 4c). The loop interacts with the C_A domain, and mutations at other positions in this loop have been shown to dramatically reduce protein secretion, suggesting direct interaction between the domains is necessary for complete folding [48]. ADAMTS13 specifically cleaves VWF and thereby controls VWF-mediated platelet thrombus formation [49], of relevance to Coronary Artery Disease [48]. In this case, then, we identify a putative missense mechanism that likely impairs maturation of the folded structure by changing a domain-domain interaction, and so reducing the level of VWF cleavage.

There is a report [50] that the major (risk) allele of the marker SNP in a Type 1 Diabetes locus (chr 5: 35.67–36.07 Mb) is associated with a higher level of the soluble isoform of the IL7 receptor. This soluble isoform is an antagonist to IL7 signaling and thus the larger amounts associated with the risk allele decrease immune signaling [50]. It has been shown that the presence of this variant affects splicing, resulting in increased skipping of exon 6 [51], and so the amount of soluble receptor produced, providing a plausible molecular mechanism basis to explain the association results. A number of auto-immune disease related effects have been found in the IL7 pathway, including increased levels of IL7 [52], consistent with the naïve expectation of increased immune signaling, particularly T cell proliferation, being associated with increased disease risk, and contrary to the signal from this SNP. We find a missense SNP, H165Q, assigned high impact by both the profile and stability methods, in LD with this marker, with the minor candidate allele associated with increased disease risk. Histidine at 165 is highly conserved, and interacts with a surface loop that is involved in interactions with the gamma chain, necessary for cell signaling (Figure 4d). Thus the risk allele likely reduces IL7 signaling, reinforcing the effect of altered splicing, but with higher penetrance.

For Type 2 diabetes on chr11:17.2–17.6 Mb, in ABCC8 (ATP-binding cassette, sub family C, member 8), the H562Q substitution in the transmembrane domain of this component of an ABC potassium transporter is assigned high impact by the SNPs3D profile model. There is currently no adequate structural template with which to assess stability impact. Inactivating mutations in ABCC8 are responsible for reduced efflux of potassium, leading to uncontrolled insulin secretion, and resulting in neonatal hyperinsulinism and hypoglycaemia. Truncating mutations are common in these cases. Interestingly, a synonymous variant in the same codon (H562H) has been observed in patients with hyperinsulinism, but its relevance to the disease is not established [53,54]. Activating mutations lead to lower than normal secretion of insulin, as in monogenic diabetes [55]. On this basis, we expect H562Q to have a moderately activating effect on channel activity, so contributing in a low penetrance manner to Type 2 Diabetes risk. Structural data is needed to support this suggestion.

For Crohn's disease, in PLCL1 (Phospholipase C-like 1) on chr2: 198.7–199.1 Mb, the V667I substitution was originally suggested as the putative mechanism variation on the basis of strong LD to the marker [47], and a relationship of PLCL1 to other Crohn's related genes

determined by GRAIL [46]. The substitution is predicted high impact by the SNP3D profile method, but found not to be structurally destabilizing with the SNPs3D structure based method. This residue is highly conserved and lies one level below the substrate binding site (Figure 4e). Likely, the bulkier isoleucine side chain distorts the binding site, so affecting a catalytic efficiency.

An E881G missense substitution in MIA3 (Melanoma inhibitory activity protein, member 3) for Coronary Artery Disease on chr1: 222.62–223.02 Mb, with a major risk allele, is predicted high impact by the profile method. A follow-up validation GWA study has shown [56] a direct association between this SNP and Coronary Artery Disease risk with the major allele as the risk allele. These authors suggest that since MIA3 is a collagen VII exporter from the endoplasmic reticulum and may participate in the migration of monocytic cells through fibrinogen or human microvascular endothelial cells, so that increased activity may increase the risk for plaque formation [56] consistent with risk associated with the major allele.

(C) Proteins containing high impact missense SNPs without a previously suggested association to disease—The 77 proteins in this category have not previously been suggested as disease relevant but contain high impact candidate missense SNPs. In some cases there is circumstantial evidence to support involvement of these proteins in disease mechanism.

For Type 1 Diabetes, at the locus on chr20: 1.41–1.81 Mb, no candidate genes have previously been suggested. There are three genes here that contain predicted high impact missense SNPs. One is in NSFL1C, a protein involved in golgi membrane fusion, and not obviously relevant to Type 1 Diabetes. The other two high impact SNPs are in members of the SIRP family. Members of this family are involved in immune regulation [57], consistent with a possible role in Type 1 Diabetes, and bind to a variety of other immune related proteins. However, the prediction of high impact in one of these, SIRPD, is marginal, and both SIFT and Polyphen2 report low impact. The other protein, SIRPB1 (signal regulatory protein beta 1), has a missense SNP, I229M (Figure 4f), predicted high impact both by profile and structure based methods. Although the exact role of this member of the family is not clear, the presence of the high impact SNP together with the immune regulation function makes it a likely candidate for involvement in Type 1 Diabetes disease mechanism.

For the hypertension associated locus on chr20: 57.55–57.95 Mb, GNAS and EDN3 are reported as the candidate genes [35]. Authors in this study suggested EDN3 as a strong candidate gene for regulation of blood pressure and GNAS as related with heart rate and smooth muscle tone. EDN3 is also linked to a QTL for blood pressure and heart weight in a rat model [58]. Although these genes, particularly EDN3, are relevant from a biology standpoint, variant related mechanisms have yet been identified. In this locus there is a predicted high impact (both by profile and structure methods) missense SNP Q43P in another gene, TUBB1, a member of the tubulin family and related to acute coronary syndrome, premature myocardial infarction and hemorrhagic stroke [59,60]. No direct role in hypertension is so far established, but the presence of a high impact candidate SNP and the connection to coronary syndromes suggest that experimental follow-up is merited.

For Rheumatoid Arthritis, at locus chr12: 57.77 – 58.17 Mb, the marker lies in KIF5A [61] and authors of the GWA study suggested this gene and another nearby, PIP4K2C, as candidates. KIF5A encodes a member of kinesin family of microtubule motors and mutations in this gene cause spastic paraplegia [62], but there is no apparent connection with Rheumatoid Arthritis risk. PIP4K2C (phosphatidylinositol-5-phosphate 4-kinase, type II, gamma) is expressed in B-cells and implicated in signaling through the B-cell antigen receptor [63]. While the latter provides a plausible candidate gene, no variant related mechanism has been identified. We find two other proteins in this locus that do contain high impact missense SNPs, GLI1 and INHBC. GLI1 encodes a member of the Kruppel family of zinc finger proteins and is expressed at a higher level in Rheumatoid Arthritis synovial tissue together with activation of sonic hedgehog signaling pathway [64]. A missense SNP, E192D is predicted high impact, and lies in an interdomain linker region, predicted to be disordered. The other high impact missense SNP, R322Q, is in INHBC and results in loss of a salt bridge so causing destabilization (figure 4h). INHBC encodes a member of the TGF beta superfamily, the beta C chain of inhibin. INHBC has been found to be related with serum uric acid level in a GWAS [65], without any direct evidence of increasing Rheumatoid Arthritis disease risk. However, INHBC is involved in the TGF-beta signaling pathway (KEGG) [66] which is upregulated in rheumatoid synovium [67,68]. Presence of high impact missense SNPs in these proteins together with the relationship to rheumatoid conditions makes these likely candidates for involvement in Rheumatoid Arthritis.

Role of low impact missense SNPs

Predicted high impact SNPs are expected to have major impact on the level of *in vivo* protein function, with a five fold or more reduction in activity [25,24]. It may be that in some cases, lower levels of function change are sufficient to contribute to disease risk of particular proteins. In support of that, there is extensive evidence of association of disease risk with eQTL relationships (SNPs associated with a change in expression), where the loss of function is usually less than two fold [69]. In this study, we find a further 124 loci that have a candidate SNP with predicted low impact. Many of these are likely effectively neutral, and so not disease relevant, but some will be relevant. Two examples of possible mechanisms are as follows:

For Bipolar disorder, we find six low impact missense SNPs in SYNE1 (nesprin1) on chr6: 152.59 – 152.99 Mb. SYNE1 is a member of the nesprin family and is expressed in central nervous system neurons. Mutations in this protein are also associated with a number of other mental diseases including autism, cerebellar ataxia, Emery-Dreifuss muscular dystrophy [70]. A GWA study has shown an association between other variants in this gene and risk of bipolar disorder [71]. Thus, one or more of these low impact SNPs may be involved in disease mechanism.

For Type 2 diabetes, we found a total of five candidate low impact missense substitutions at three positions in THADA (thyroid adenoma associated) on chr2: 43.53 – 43.93 Mb, including the marker (T1187A) itself. THADA is expressed in the pancreas. It has been shown that the major (risk) allele of the marker SNP [72] is associated with lowering of

pancreatic β -cell function, consistent with a role in Type 2 Diabetes risk, and one or more of these five low impact substitutions may be causative.

In some loci, the evidence provided by high impact missense SNPs is less convincing than that for low impact missense ones. For example, in a Crohn's locus on chr1: 200.68–201.08 Mb, Franke et al [47] suggested two candidate genes, C1orf106 and KIF21B (kinisin family member 21B). KIF21B has been suggested as associated with multiple sclerosis [73, 74] and inflammatory bowel disease [1], but no functional relevance to Crohn's has been identified. We found two rare ($\text{maf} < 5\%$) high impact missense SNPs in this gene, but because of uncertainty in the LD data we cannot ascertain whether these could generate the observed marker signal. We also found a high impact SNP in another gene in the locus, CACNA1S. However, CACNA1S is a calcium channel expressed in skeletal muscle cells, so does not look promising for involvement in Crohn's. We found two low impact missense SNPs in the other proposed candidate gene, C1orf106.

Role of rarer missense SNPs

As discussed earlier, over half of all potential candidate SNPs are excluded on the grounds of insufficient LD accuracy, most with frequencies less than 5%. The rare variants may in fact be involved in disease mechanisms. The fraction of high impact missense SNPs in these (42%) is higher than found for the ones included in the analysis (30%), as expected – the lower the frequency of a variant, the more likely it is to be deleterious [5, 6, 16]. We can obtain an upper limit for their potential contribution by assuming all are in appropriate LD with markers, and could generate the observed marker case/control frequency difference. 40% of the common SNPs satisfied these conditions, but this fraction is likely lower for rare variants. Out of the 356 loci considered, there are 218 (61%), which contain at least one rare high impact missense variant. In all, 648 unique rare missense variants in 436 different proteins are assigned as high impact, corresponding to a potential 784 mechanisms across the seven diseases (some variants are candidates for more than one disease). Some of these occur in loci also containing a common high impact candidate SNP. There are 110 loci that do not contain a common high impact missense SNP, but do contain at least one rare high impact missense SNP. (Of these, in 41 loci the rare high impact SNP falls in a previously suggested candidate protein). If we assume that the same fraction of rare missense variants qualify as candidates as found for common SNPs (40%), that would imply a further 44 loci with a rare missense mechanism, increasing total loci with a putative missense mechanism to 64%. However, this should be regarded as an upper bound, since a higher proportion of rare variants will not meet the candidate SNP criteria.

Discussion and conclusions

Although GWA studies have now discovered over 2000 loci where one or more common SNPs is associated with risk of a common disease, in only a small number of those loci is there any detailed understanding of the underlying mechanism. It is not possible to reap of the full benefits of GWAS, particularly with regard to new therapies, without such mechanistic understanding. The aim of this study was to investigate the extent to which missense SNPs play a role in these complex trait human disease loci, and where that is the

case, to examine the underlying molecular mechanisms. Taking into consideration all relevant GWA studies, we find that approximately 33% of 356 loci associated with increased risk in one or more of seven diverse complex trait diseases have a candidate missense SNP with a predicted high impact on the function of a protein (usually a decrease in *in vivo* function similar to that typically found for monogenic disease mutations [26]), indicating a significant role for this mechanism in complex trait disease.

Each of the candidate high impact missense SNPs provides a hypothesis for involvement in disease mechanism, and a basis for further experimental verification. Candidates have LD relationships to marker SNPs such that they may generate the observed frequency difference between case and control, but since the actual candidate SNP frequency in the case population is not usually known, this evidence is not definitive. Agreement with imputation methods for a subset of the data confirms the likely relevance of most of the candidates (Supplementary Figure 3). There may also be false positive assignments of high impact from the SNPs3D methods, and here agreement with assignments from SIFT or Polyphen2 provide support, and allow an estimate of likely reliability for each candidate. Further supporting evidence for role of high impact missense SNPs is the observation that the fraction of minor risk alleles increases from roughly equal for marker SNPs (48% major and 52% minor) to strongly biased to minor for predicted high impact candidates (34% major and 66% minor). Additionally, although only 30% of the proteins in the loci have previously been proposed as disease relevant, 58% of the candidate high impact SNPs fall in these, a very significant enrichment (Fisher's exact test P-value < 0.0001). There is also a higher fraction of high impact missense SNPs in GWAS loci than in all missense SNPs (30% versus 22%).

Missense molecular mechanisms

Missense SNPs may affect *in vivo* protein function in a number of ways, including effects on catalysis, ligand binding, allosteric regulation, post-translational modification, and protein folding and half-life. The SNPs3D structural analysis method identifies those missense SNPs that destabilize tertiary structure, thus likely to affect folding and/or half life. For the 50% (163) of high impact missense SNPs where adequate three dimensional structure information is available to apply this method, 71% (115) are predicted destabilizing. This large role for destabilization is consistent with that found for high impact mutations in monogenic disease (~70%) [26]. A limited analysis of cancer driver mutations also suggests a higher fraction there (64%) [27]. For the 48 cases where three-dimensional structure is available and stability is not predicted to be involved, a variety of molecular mechanisms are found. Examples include likely distortion of an enzyme active site (PLCL1), disruption of protein-protein interactions (MST1, PTPN22), and indirect effects on protein folding and maturation (ADAMTS13). The low coverage of candidate SNPs by structure partly reflects the state of experimental work for human proteins, but 38% of candidate missense SNPs are predicted to lie in intrinsically disordered regions [75] of the proteins concerned. The fraction predicted disorder for candidate high impact SNPs is lower, at 23%.

Enrichment tests

As noted above, we find that high impact missense SNPs are enriched with minor allele risk variants, proteins in which they occur are enriched for those suggested as disease relevant, and there is an enrichment of high impact SNPs. We also can estimate whether there is a significantly higher occurrence of missense mechanisms in disease loci than would be expected by chance. There is a protein carrying a candidate high impact missense SNP in about 1/3 of risk loci. An estimate of the number of loci expected to have such high impact missense SNP carrying proteins by chance is as follows: As noted earlier, for 16,182 missense SNPs with a frequency > 5% in the 1000 genomes data [4] in the Caucasian population, 22% are predicted high impact by SNPs3D. Thus, on average, one in 10 proteins is expected to carry a high impact missense SNP by chance. The risk loci contain an average of about five proteins, thus approximately one in two loci is expected to include a high impact msSNP bearing protein by chance. We observe one in three, slightly less than chance, no enrichment. This is a very approximate calculation, but a more precise one is in fact quite tricky to make, requiring careful weighting of the probabilities of finding a hitting particular genes (a function of LD and length) in each locus. In fact, we do not expect to see enrichment of this type, given that the potential role of missense mechanisms is so very common. The activity of rather few genes is sufficiently tightly coupled to a specific disease phenotype for contemporary GWAS to detect (up to about 100). Thus, to a first approximation, what determines whether a gene is implicated in a complex trait disease is not whether there is a mechanism SNP available (there often will be) but whether altering the activity of the affected protein significantly impacts the disease phenotype.

Potential mechanisms involving SNPs that change gene expression are also common, with eQTL studies suggesting an average of up to one SNP affecting expression per gene [69]. Thus for the same reasons as with missense, we would not expect a significantly higher occurrence of eQTL mechanisms in disease loci than chance. There is one report of this type of enrichment [32]. In that case, the random expectation was based on the number of SNPs on the microarray chip use in the GWA study. Most eQTL relationships involve SNPs very close (~85% within 200Kb) to genes [69], whereas the set on the microarray are much more broadly distributed. On that basis, the apparent eQTL enrichment may be an artifact of the choice of control set.

We also checked to see if the genes containing predicted high impact SNPs are enriched for those under unusually high selection pressure, using the mutability results of Samocha et al. [76]. For the 162 proteins containing high impact msSNPs that have mutability data, only seven (4.3%) are categorized as ‘constrained’ by these authors (containing significantly fewer common missense SNPs than expected), similar to the rate for all genes (~5%). This is the expected result, since most genes influencing complex traits are probably under relatively mild selection pressure, as each contributes so little to the disease phenotype.

Role of variants affecting expression

Other types of enrichment related tests have focused on the relative roles of expression and missense by comparing the contributions to complex traits from different classes of genome regions – DNAase-I hypersensitive (DHS, correlated with involvement in expression

regulation), coding, introns, UTRs, and intergenic – with that expected by chance. There is very significant enrichment in coding regions, for example 14 fold in Gusev et al. [33], supporting a role for missense variants. There is also enrichment in DHS regions, though to a lesser extent (5 fold in [33]). However, because of the much larger number of bases classified as DHS compared to those involved in coding, DHS usually emerges as the source of the largest fraction of total GWAS signal. Overall, these methods have found 10 to 20% of the total GWAS signal is associated with coding regions, while up to 79% are associated with DHS regions [33,34,77]. These numbers cannot be directly compared with the 30% missense value obtained here. The most robust of these other studies integrate over all SNPs included in a GWAS experiment, and have been found to account for a much larger fraction of heritability than when considering just GWAS loci. For example, Torres et al. [77], found that over 50% of Type 2 Diabetes heritability is explained when including all SNPs compared with only 10% from SNPs in GWAS identified loci. These results imply that there are a large number of very weak contributions to complex trait phenotypes (too weak to detect individually by GWAS), and many of those are associated with expression changes. The relative roles of missense and expression for the strongest contributions that are identified by GWAS is still unclear, although most studies still find a higher fraction for expression.

Role of rare and low impact missense SNPs

The high impact missense SNPs discussed above are those most likely to be involved in disease mechanism, but there are two sorts of additional missense that are also relevant. As discussed earlier, a further 31% of loci have predicted high impact missense SNPs that may be involved in disease mechanism, but available linkage disequilibrium information makes this difficult to determine. Most of these are of lower frequency (less than 5%). Two analyses [5,16] using different methodologies, have shown that the lower the frequency of a missense variant in the human population, the more likely it is to have significant impact on fitness, and in one of the studies, the impact on fitness has been shown to correlate with impact at the protein level [16]. It has also been suggested that combinations of rare variants may be sufficient for many complex disease effects [5,6], although some more recent data suggest otherwise [78]. Clearly, the role of these lower frequency missense SNPs should not be ignored.

In addition to the 33% of loci that contain high impact missense SNPs, a further 35% have candidate missense SNPs that are not predicted to have a major impact on protein function. The impact of these on function will range from none (effectively neutral) up to close to high impact, and the distribution over that range is not yet fully understood. One analysis [5] shows over 50% of very low frequency missense variants ($\text{maf} < 1\%$) have a mildly deleterious effect on fitness, but the fraction at higher frequencies is not clear. Another analysis [16], using a continuous fitness impact scale, shows low levels of missense variants with intermediate fitness impact in the range of interest. As noted earlier, there is considerable data demonstrating a role for SNPs affecting expression in complex trait disease (for instance [79,80]), and these generally have a low (less than 2 fold) effect on expression, suggesting that lower impact missense SNPs will also play a role in these diseases.

METHODS

As shown in Figure 5, a five step workflow is used to determine which missense SNPs are candidates for involvement in disease mechanism.

Markers from Genome Wide association data

Two sets of loci were compiled for the seven diseases included in the WTCCC1 GWA study [15]: Bipolar Disorder, Coronary Artery Disease, Crohn's Disease, Hypertension, Rheumatoid Arthritis, Type 1 Diabetes, and Type 2 Diabetes. Set 1 consists of 21 loci reported to contain SNPs with significant disease association in the WTCCC1 study [15]. Set 2 consists of a further 335 loci for these diseases containing significant marker SNPs (P -value $< 1E-05$) as reported in the GWAS catalog [37], derived from later GWA studies and meta-analyses for these seven diseases. For Set 1, representative marker SNPs for each locus are those in table 3 of the WTCCC1 paper [15]. For Set 2, markers are taken from the GWAS catalog entries for the relevant diseases. Where a locus has multiple markers resulting from different studies, the one based on the highest sample size was selected.

Linkage disequilibrium (LD) data

Linkage disequilibrium data for the selected markers were downloaded from the International Hapmap Project (hapmap release#27 – merged I+II: rel #24 and III: release #2, NCBI build 36, February 2009). 1000 genomes data were obtained from the interim Phase 1 release (Nov, 2011) [81]. Frequency and genotype data for the Caucasian (CEU) population were used. Conversion of 1000 genomes hg19 coordinates to hg18 coordinates was done with the UCSC liftOver tool [82]. 1000genomes linkage disequilibrium data for the selected markers were generated using the PLINK pairwise LD measures module in the LD calculations toolset [83]. If a potential candidate SNP has different LD relationships with a marker in Hapmap and in the 1000genomes CEU population, the one with higher D' was selected for further consideration.

Pre-phasing, Imputation and association testing of linked SNPs

Impute2 [84] was used to impute the genotypes of SNPs in linkage disequilibrium with each selected marker. The reference dataset for imputation is Hapmap 3+1000genomes Pilot haplotypes, NCBI build 36 (hg18). Imputation was based on WTCCC1 case control genotype data. Pre-phasing of the data was performed using ShapeIT [85] and disease association P values of imputed SNPs were calculated using SNPTEST v2 [86].

Calculation of marker risk allele probabilities in Case individuals

For studies where full genotype data were not available, $P(r_m|D)$, the probability of the marker risk allele ' r_m ' (the allele with higher frequency in case ('D') versus control ($\bar{D}$) individuals, whether major or minor), is derived from the marker odds ratio ' OR_m ' (obtained from the GWAS catalog [37]) as follows:

$$OR_m = O_r / O_p$$

Where O_r , the odds for the risk allele of the marker SNP is:

$$O_r = P(r_m|D)/P(r_m|\bar{D})$$

$P(r_m|\bar{D})$, the probability of the risk allele occurring in Control individuals, is assumed to be equal to ' f_m ', the risk allele frequency of the marker in the Hapmap or the 1000genomes CEU populations. Similarly O_p , the odds for a protective allele 'p' of a marker SNP is:

$$O_p = P(p_m|D)/P(p_m|\bar{D})$$

Then the desired quantity, $P(r_m|D)$, is obtained from:

$$P(r_m|D) = 1/(1+A)$$

where,

$$A = (1 - f_m)/f_m \cdot OR_m$$

Relationship between mechanism and marker risk alleles

As noted earlier, we assume the frequency difference between Case and Control populations for the marker risk allele, r_m , is a consequence of its LD relationship with a variant that directly influences disease risk. If that mechanism variant is one of the candidate SNPs, r_c , then, for the Case population, the relationship between the presence of the candidate risk allele and the probability of also observing the marker risk allele is:

$$P(r_m|D) = P(r_m|r_c) \cdot P(r_c|D) + P(r_m|p_c) \cdot [1 - P(r_c|D)] \quad (1)$$

where $P(r_m|r_c)$ is the conditional probability of the marker risk allele being present in a Case individual, given the presence of the candidate risk allele, and $P(r_m|p_c)$ is the conditional probability of the marker risk allele being present given the presence of the protective candidate allele ' p_c '. These probabilities are calculated using the haplotype frequencies of the marker and candidate alleles as derived from the CEU population LD and SNP frequency information. Thus, $P(r_m|r_c) = f_{mc}/f_c$, where f_{mc} is the haplotype frequency of the marker and candidate risk alleles, and $f_{mc} = f_m \cdot f_c + D$, where D is the linkage disequilibrium between the marker and candidate alleles, derived from the LD data, and f_m and f_c are the control risk allele frequencies of the marker and candidate SNPs respectively.

Similarly, in control individuals, the probability of the risk marker allele is:

$$P(r_m|\bar{D}) = P(r_m|r_c) \cdot P(r_c|\bar{D}) + P(r_m|p_c) \cdot [1 - P(r_c|\bar{D})] \quad (2)$$

Subtracting equation (2) from equation (1):

$$[P(r_m|D) - P(r_m|\overline{D})] = [P(r_m|r_c) - P(r_m|p_c)] \cdot [P(r_c|D) - P(r_c|\overline{D})]$$

that is

$$\delta P(r_m) = [P(r_m|r_c) - P(r_m|p_c)] \cdot \delta P(r_c) \quad (3)$$

where $\delta P(r_m)$ and $\delta P(r_c)$ are the differences in probability of occurrence of the risk allele in Case and Control for the marker and candidate SNPs respectively.

Eqn (3) provides a convenient expression for obtaining the difference in probability of occurrence of the risk allele between Case and Control for any candidate SNP, using the LD data and the frequency properties of the marker SNP:

$$\delta P(r_c) = K \cdot \delta P(r_m)$$

where,

$$K = 1/[P(r_m|r_c) - P(r_m|p_c)] \quad (4)$$

K is a proportionality constant dependent on the LD relationship between marker and risk alleles. For complete LD between the candidate and marker SNPs, $P(r_m|r_c) = 1$ and $P(r_m|p_c) = 0$, so $K = 1$. For incomplete LD $[P(r_m|r_c) - P(r_m|p_c)] < 1$, thus

$$|K| \geq 1.0, \text{ and } |\delta P(r_c)| \geq |\delta P(r_m)|$$

That is, the difference in probability of occurrence of the marker SNP risk allele between Case and Control is no larger than that for the candidate SNP and for incomplete LD this difference is smaller.

Selection of candidate mechanism SNPs

For each locus, candidate mechanism SNPs were selected from the set of all Hapmap and 1000genomes SNPs that are in non-zero LD with the marker SNP and within 200Kb of the marker. The marker SNP itself is also included as a candidate. Two classes of criteria are used for selection:

1. Whether the LD information is adequate for accurate calculation of the properties of the candidate SNP. Two criteria are used:
 - a. Any candidate SNP with a minor allele frequency (maf) of less than 5% is discarded – current Hapmap and 1000genomes CEU populations are not large enough to reliably determine LD at these lower frequencies. About 41% of the total data are removed using this criterion.

- b. Any candidate for which the implied case and control frequency difference cannot be reliably calculated because of numerical instability in the value of K , eqn (4) above, is discarded. As LD weakens, the magnitude of the denominator in eqn (4) decreases, and K becomes poorly determined. All candidates for which $K > 10.0$ are rejected. Supplementary Figure 2 gives the distribution of K values for all candidate SNPs. About 12% of the total data are removed using this criterion.
2. Whether the properties of the candidate SNP are such that if it is the SNP involved in mechanism in that locus, it would generate the observed frequency difference between case and control for the marker SNP. A candidate SNP is considered potentially causative if its LD relationship to the marker is such that the observed frequency difference between case and control for the marker could be a consequence of a difference in frequency of the candidate between case and control. For this purpose the implied case frequency of the candidate risk is obtained from

$$P(r_c|D) = f_c + \delta P(r_c)$$

where $\delta P(r_c)$ is the difference in frequency of the candidate between case and control that is required to produce the observed frequency difference for the marker (calculated as described above), and f_c is the frequency of the candidate in the control population, taken to be the Caucasian frequency in dbSNP.

The estimated $P(r_c|D)$ must lie in the range $0.0 \rightarrow 1.0$. SNPs that do not meet this condition are not in sufficiently strong LD with the marker to produce the GWAS observations. After the extensive filtering above, a further 7% of SNPs do not qualify by this criterion.

Functional class and gene assignment

The functional class of each SNP (missense or otherwise) was taken from dbSNP137 [2]. Association of genes to a SNP is assigned according to dbSNP 137. The chromosomal position of each SNP is according to NCBI build 37, hg19.

Locus boundaries

A locus is defined as the region spanning 200Kb on either side of the marker (LD calculation range), 0.4Mb in total for each marker.

Identification of high impact missense SNPs

Two computational methods for analyzing the impact of coding region missense SNPs on *in vivo* protein function (SNPs3D) [26,23] were used for initial analysis. These methods use support vector machine (SVM) models, trained on monogenic disease mutations and a control set, to classify each single base variant as deleterious or non-deleterious to *in vivo* protein function. One method [26] uses structural information to analyze the effect of an amino acid substitution on protein stability. The second method uses sequence conservation within the protein family and the probability of the particular amino acid substitution

introduced by the base variant [23]. For both models, missense SNPs with SVM scores less than zero are considered to represent a high impact on the *in vivo* function of the corresponding protein, and those with zero or higher, a low impact. This threshold corresponds to that used in classifying monogenic disease mutations in training, and is estimated to correspond to an approximately five fold impact on activity [25,24]. The high impact SNPs were considered as candidates for increasing disease risk.

Two other methods were also used to further assess SNPs predicted to be high impact by one or both of the SNPs3D methods: SIFT [22] and PolyPhen2 [18]. For SIFT predictions, we considered any variant predicted as ‘damaging’ to be high impact. For Polyphen2, predictions of ‘probably damaging’ or ‘possibly damaging’ were considered to be high impact.

Human mRNA expression data

Human mRNA expression data for 79 human tissues [87] was downloaded from <http://biogps.org/downloads/>. For each gene, the maximum expression level across all tissues was used to compare proteins with structure with others.

Human Protein Functional Interaction Network

The Human protein functional interaction network [88] was downloaded from <http://genomebiology.com/content/supplementary/gb-2010-11-5-r53-s3.zip> on April 2012. The functional interaction set, FI, was used to compare proteins with structure with others.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Yizhou Yin for assistance with the protein stability and disorder calculations and for useful discussions. This work was supported in part by NIH grants R01GM102810 and R01GM104436.

References

1. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*. 2012; 491:119–24. DOI: 10.1038/nature11582 [PubMed: 23128233]
2. dbSNP n.d. <http://www.ncbi.nlm.nih.gov/projects/SNP/>.
3. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, et al. Integrating common and rare genetic variation in diverse human populations. *Nature*. 2010; 467:52–8. DOI: 10.1038/nature09298 [PubMed: 20811451]
4. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, et al. A map of human genome variation from population-scale sequencing. *Nature*. 2010; 467:1061–73. DOI: 10.1038/nature09534 [PubMed: 20981092]
5. Kryukov GV, Pennacchio LA, Sunyaev SR. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet*. 2007; 80:727–39. DOI: 10.1086/513473 [PubMed: 17357078]

6. Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei L-J, et al. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet.* 2010; 86:832–8. DOI: 10.1016/j.ajhg.2010.04.005 [PubMed: 20471002]
7. Weedon MN, Frayling TM. Insights on pathogenesis of type 2 diabetes from MODY genetics. *Curr Diab Rep.* 2007; 7:131–8. [PubMed: 17425917]
8. Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, et al. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res.* 2013; 23:749–61. DOI: 10.1101/gr.148718.112 [PubMed: 23478400]
9. Cook EH, Scherer SW. Copy-number variations associated with neuropsychiatric conditions. *Nature.* 2008; 455:919–23. DOI: 10.1038/nature07458 [PubMed: 18923514]
10. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. *G3 (Bethesda).* 2011; 1:457–70. DOI: 10.1534/g3.111.001198 [PubMed: 22384356]
11. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet.* 2007; 39:906–13. DOI: 10.1038/ng2088 [PubMed: 17572673]
12. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet.* 2012; 44:955–9. DOI: 10.1038/ng.2354 [PubMed: 22820512]
13. 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. DOI: 10.1086/521987 [PubMed: 17924348]
14. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 2014; 42:D1001–6. DOI: 10.1093/nar/gkt1229 [PubMed: 24316577]
15. 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. DOI: 10.1038/nature05911 [PubMed: 17554300]
16. Katsonis P, Lichtarge O. A formal perturbation equation between genotype and phenotype determines the evolutionary action of protein coding variations on fitness. *Genome Res.* 2014; doi: 10.1101/gr.176214.114
17. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet.* 2006; 7:61–80. DOI: 10.1146/annurev.genom.7.080505.115630 [PubMed: 16824020]
18. Adzhubei, I.; Schmidt, S.; Peshkin, L.; Ramensky, V.E.; Gerasimova, A.; Bork, P., et al. A method and server for predicting damaging missense mutations. *Nat Methods.* 2010; 7:248–9. DOI: 10.1038/nmeth0410-248 [PubMed: 20354512]
19. Bromberg Y, Yachdav G, Rost B. SNAP predicts effect of mutations on protein function. *Bioinformatics.* 2008; 24:2397–8. DOI: 10.1093/bioinformatics/btn435 [PubMed: 18757876]
20. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics.* 2009; 25:2744–50. DOI: 10.1093/bioinformatics/btp528 [PubMed: 19734154]
21. Reva B, Antipin Y, Sander C. Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol.* 2007; 8:R232.doi: 10.1186/gb-2007-8-11-r232 [PubMed: 17976239]
22. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc.* 2009; 4:1073–81. DOI: 10.1038/nprot.2009.86 [PubMed: 19561590]
23. Yue P, Moulton J. Identification and analysis of deleterious human SNPs. *J Mol Biol.* 2006; 356:1263–74. DOI: 10.1016/j.jmb.2005.12.025 [PubMed: 16412461]
24. Yampolsky LY, Stoltzfus A. The exchangeability of amino acids in proteins. *Genetics.* 2005; 170:1459–72. DOI: 10.1534/genetics.104.039107 [PubMed: 15944362]
25. Shi Z, Sellers J, Moulton J. Protein stability and in vivo concentration of missense mutations in phenylalanine hydroxylase. *Proteins.* 2012; 80:61–70. DOI: 10.1002/prot.23159 [PubMed: 21953985]

26. Yue P, Li Z, Moulton J. Loss of protein structure stability as a major causative factor in monogenic disease. *J Mol Biol.* 2005; 353:459–73. DOI: 10.1016/j.jmb.2005.08.020 [PubMed: 16169011]
27. Shi Z, Moulton J. Structural and functional impact of cancer-related missense somatic mutations. *J Mol Biol.* 2011; 413:495–512. DOI: 10.1016/j.jmb.2011.06.046 [PubMed: 21763698]
28. SNPs3D <http://www.snps3d.org/>.
29. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet.* 2014; 133:1–9. DOI: 10.1007/s00439-013-1358-4 [PubMed: 24077912]
30. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. *Nat Genet.* 2013; 45:1127–33. DOI: 10.1038/ng.2762 [PubMed: 24071851]
31. Stehr H, Jang S-HJ, Duarte JM, Wierling C, Lehrach H, Lappe M, et al. The structural impact of cancer-associated missense mutations in oncogenes and tumor suppressors. *Mol Cancer.* 2011; 10:54.doi: 10.1186/1476-4598-10-54 [PubMed: 21575214]
32. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* 2010; 6:e1000888.doi: 10.1371/journal.pgen.1000888 [PubMed: 20369019]
33. Gusev A, Lee SH, Trynka G, Finucane H, Vilhjálmsson BJ, Xu H, et al. Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. *Am J Hum Genet.* 2014; 95:535–52. DOI: 10.1016/j.ajhg.2014.10.004 [PubMed: 25439723]
34. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science (80-).* 2012; 337:1190–6.
35. Ehret GB, Munroe PB, Rice KM, Bochud M, Johnson AD, Chasman DI, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature.* 2011; 478:103–9. DOI: 10.1038/nature10405 [PubMed: 21909115]
36. Allali-Hassani A, Wasney GA, Chau I, Hong BS, Senisterra G, Loppnau P, et al. A survey of proteins encoded by non-synonymous single nucleotide polymorphisms reveals a significant fraction with altered stability and activity. *Biochem J.* 2009; 424:15–26. DOI: 10.1042/BJ20090723 [PubMed: 19702579]
37. GWAS Catalog n.d. <http://www.genome.gov/gwastudies> (accessed September, 2013).
38. Gorlatova N, Chao K, Pal LR, Araj RH, Galkin A, Turko I, et al. Protein characterization of a candidate mechanism SNP for Crohn's disease: the macrophage stimulating protein R689C substitution. *PLoS One.* 2011; 6:e27269.doi: 10.1371/journal.pone.0027269 [PubMed: 22087277]
39. Korzenik JR. Is Crohn's disease due to defective immunity? *Gut.* 2007; 56:2–5. DOI: 10.1136/gut.2006.095588 [PubMed: 17172582]
40. Goyette P, Lefebvre C, Ng A, Brant SR, Cho JH, Duerr RH, et al. Gene-centric association mapping of chromosome 3p implicates MST1 in IBD pathogenesis. *Mucosal Immunol.* 2008; 1:131–8. DOI: 10.1038/mi.2007.15 [PubMed: 19079170]
41. Strober W, Asano N, Fuss I, Kitani A, Watanabe T. Cellular and molecular mechanisms underlying NOD2 risk-associated polymorphisms in Crohn's disease. *Immunol Rev.* 2014; 260:249–60. DOI: 10.1111/imr.12193 [PubMed: 24942694]
42. Inohara N, Ogura Y, Fontalba A, Gutierrez O, Pons F, Crespo J, et al. Host recognition of bacterial muramyl dipeptide mediated through NOD2. Implications for Crohn's disease. *J Biol Chem.* 2003; 278:5509–12. DOI: 10.1074/jbc.C200673200 [PubMed: 12514169]
43. Petrone A, Spoletini M, Zampetti S, Capizzi M, Zavarella S, Osborn J, et al. The PTPN22 1858T gene variant in type 1 diabetes is associated with reduced residual beta-cell function and worse metabolic control. *Diabetes Care.* 2008; 31:1214–8. DOI: 10.2337/dc07-1158 [PubMed: 18252906]
44. Bottini N, Musumeci L, Alonso A, Rahmouni S, Nika K, Rostamkhani M, et al. A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. *Nat Genet.* 2004; 36:337–8. DOI: 10.1038/ng1323 [PubMed: 15004560]

45. Hinks A, Barton A, John S, Bruce I, Hawkins C, Griffiths CEM, et al. Association between the PTPN22 gene and rheumatoid arthritis and juvenile idiopathic arthritis in a UK population: further support that PTPN22 is an autoimmunity gene. *Arthritis Rheum.* 2005; 52:1694–9. DOI: 10.1002/art.21049 [PubMed: 15934099]
46. Raychaudhuri S, Plenge RM, Rossin EJ, Ng ACY, Purcell SM, Sklar P, et al. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* 2009; 5:e1000534.doi: 10.1371/journal.pgen.1000534 [PubMed: 19557189]
47. Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet.* 2010; 42:1118–25. DOI: 10.1038/ng.717 [PubMed: 21102463]
48. Bongers TN, de Bruijne ELE, Dippel DWJ, de Jong AJ, Deckers JW, Poldermans D, et al. Lower levels of ADAMTS13 are associated with cardiovascular disease in young patients. *Atherosclerosis.* 2009; 207:250–4. DOI: 10.1016/j.atherosclerosis.2009.04.013 [PubMed: 19439298]
49. Akiyama M, Takeda S, Kokame K, Takagi J, Miyata T. Crystal structures of the noncatalytic domains of ADAMTS13 reveal multiple discontinuous exosites for von Willebrand factor. *Proc Natl Acad Sci U S A.* 2009; 106:19274–9. DOI: 10.1073/pnas.0909755106 [PubMed: 19880749]
50. Monti P, Brigatti C, Krasmann M, Ziegler AG, Bonifacio E. Concentration and activity of the soluble form of the interleukin-7 receptor α in type 1 diabetes identifies an interplay between hyperglycemia and immune function. *Diabetes.* 2013; 62:2500–8. DOI: 10.2337/db12-1726 [PubMed: 23454692]
51. Gregory SG, Schmidt S, Seth P, Oksenberg JR, Hart J, Prokop A, et al. Interleukin 7 receptor alpha chain (IL7R) shows allelic and functional association with multiple sclerosis. *Nat Genet.* 2007; 39:1083–91. DOI: 10.1038/ng2103 [PubMed: 17660817]
52. McKay FC, Swain LI, Schibeci SD, Rubio JP, Kilpatrick TJ, Heard RN, et al. Haplotypes of the interleukin 7 receptor alpha gene are correlated with altered expression in whole blood cells in multiple sclerosis. *Genes Immun.* 2008; 9:1–6. DOI: 10.1038/sj.gene.6364436 [PubMed: 17928869]
53. Flanagan SE, Clauin S, Bellanné-Chantelot C, de Lonlay P, Harries LW, Gloyn AL, et al. Update of mutations in the genes encoding the pancreatic beta-cell K(ATP) channel subunits Kir6.2 (KCNJ11) and sulfonylurea receptor 1 (ABCC8) in diabetes mellitus and hyperinsulinism. *Hum Mutat.* 2009; 30:170–80. DOI: 10.1002/humu.20838 [PubMed: 18767144]
54. Fernández-Marmiesse A, Salas A, Vega A, Fernández-Lorenzo JR, Barreiro J, Carracedo A. Mutation spectra of ABCC8 gene in Spanish patients with Hyperinsulinism of Infancy (HI). *Hum Mutat.* 2006; 27:214.doi: 10.1002/humu.9401 [PubMed: 16429405]
55. Tarasov AI, Nicolson TJ, Riveline J-P, Taneja TK, Baldwin SA, Baldwin JM, et al. A rare mutation in ABCC8/SUR1 leading to altered ATP-sensitive K⁺ channel activity and beta-cell glucose sensing is associated with type 2 diabetes in adults. *Diabetes.* 2008; 57:1595–604. DOI: 10.2337/db07-1547 [PubMed: 18346985]
56. Wang AZ, Li L, Zhang B, Shen G-Q, Wang QK. Association of SNP rs17465637 on chromosome 1q41 and rs599839 on 1p13.3 with myocardial infarction in an American caucasian population. *Ann Hum Genet.* 2011; 75:475–82. DOI: 10.1111/j.1469-1809.2011.00646.x [PubMed: 21463265]
57. Van Beek EM, Cochrane F, Barclay AN, van den Berg TK. Signal regulatory proteins in the immune system. *J Immunol.* 2005; 175:7781–7. [PubMed: 16339510]
58. Cicila GT, Choi C, Dene H, Lee SJ, Rapp JP. Two blood pressure/cardiac mass quantitative trait loci on chromosome 3 in Dahl rats. *Mamm Genome.* 1999; 10:112–6. [PubMed: 9922389]
59. Navarro-Núñez L, Roldán V, Lozano ML, Rivera J, Marin F, Vicente V, et al. TUBB1 Q43P polymorphism does not protect against acute coronary syndrome and premature myocardial infarction. *Thromb Haemost.* 2008; 100:1211–3. [PubMed: 19132255]
60. Navarro-Núñez L, Lozano ML, Rivera J, Corral J, Roldán V, González-Conejero R, et al. The association of the beta1-tubulin Q43P polymorphism with intracerebral hemorrhage in men. *Haematologica.* 2007; 92:513–8. [PubMed: 17488662]

61. Raychaudhuri S, Remmers EF, Lee AT, Hackett R, Guiducci C, Burtt NP, et al. Common variants at CD40 and other loci confer risk of rheumatoid arthritis. *Nat Genet.* 2008; 40:1216–23. DOI: 10.1038/ng.233 [PubMed: 18794853]
62. Fichera M, Lo Giudice M, Falco M, Sturnio M, Amata S, Calabrese O, et al. Evidence of kinesin heavy chain (KIF5A) involvement in pure hereditary spastic paraplegia. *Neurology.* 2004; 63:1108–10. [PubMed: 15452312]
63. Carpenter C. Btk-dependent regulation of phosphoinositide synthesis. *Biochem Soc Trans.* 2004; 32:326–9. [PubMed: 15046600]
64. Wang M, Zhu S, Peng W, Li Q, Li Z, Luo M, et al. Sonic hedgehog signaling drives proliferation of synoviocytes in rheumatoid arthritis: a possible novel therapeutic target. *J Immunol Res.* 2014; 2014:401903.doi: 10.1155/2014/401903 [PubMed: 24741597]
65. Yang Q, Köttgen A, Dehghan A, Smith AV, Glazer NL, Chen M-H, et al. Multiple genetic loci influence serum urate levels and their relationship with gout and cardiovascular disease risk factors. *Circ Cardiovasc Genet.* 2010; 3:523–30. DOI: 10.1161/CIRCGENETICS.109.934455 [PubMed: 20884846]
66. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000; 28:27–30. [PubMed: 10592173]
67. Gonzalo-Gil E, Criado G, Santiago B, Dotor J, Pablos JL, Galindo M. Transforming growth factor (TGF)- β signalling is increased in rheumatoid synovium but TGF- β blockade does not modify experimental arthritis. *Clin Exp Immunol.* 2013; 174:245–55. DOI: 10.1111/cei.12179 [PubMed: 23869798]
68. Pohlert D, Beyer A, Koczan D, Wilhelm T, Thiesen H-J, Kinne RW. Constitutive upregulation of the transforming growth factor-beta pathway in rheumatoid arthritis synovial fibroblasts. *Arthritis Res Ther.* 2007; 9:R59.doi: 10.1186/ar2217 [PubMed: 17594488]
69. Yu C-H, Pal LR, M J. Analysis of consensus genome-wide expression-QTLs and their relationship to Human complex trait disease. Submitted 2015.
70. Tulgren ED, Turgeon SM, Opperman KJ, Grill B. The Nesprin family member ANC-1 regulates synapse formation and axon termination by functioning in a pathway with RPM-1 and β -Catenin. *PLoS Genet.* 2014; 10:e1004481.doi: 10.1371/journal.pgen.1004481 [PubMed: 25010424]
71. Green EK, Grozeva D, Forty L, Gordon-Smith K, Russell E, Farmer A, et al. Association at SYNE1 in both bipolar disorder and recurrent major depression. *Mol Psychiatry.* 2013; 18:614–7. DOI: 10.1038/mp.2012.48 [PubMed: 22565781]
72. Simonis-Bik AM, Nijpels G, van Haeften TW, Houwing-Duistermaat JJ, Boomsma DI, Reiling E, et al. Gene variants in the novel type 2 diabetes loci CDC123/CAMK1D, THADA, ADAMTS9, BCL11A, and MTNR1B affect different aspects of pancreatic beta-cell function. *Diabetes.* 2010; 59:293–301. DOI: 10.2337/db09-1048 [PubMed: 19833888]
73. McCauley JL, Zuvich RL, Beecham AL, De Jager PL, Konidari I, Whitehead PL, Aubin C, Ban M, Pobywajlo S, Briskin R, Romano S, Aggarwal NT, Piccio L, McArdle WL, Strachan DP, Evans D, Cross AH, Cree B, Rioux JD, Barcellos LF, Iverson AJ, Compston A, Hafler HJ. Comprehensive follow-up of the first genome-wide association study of multiple sclerosis identifies KIF21B and TMEM39A as susceptibility loci. *Hum Mol Genet.* 2010; 19:953–62. DOI: 10.1093/hmg/ddp542 [PubMed: 20007504]
74. Goris A, Boonen S, D'hooghe M-B, Dubois B. Replication of KIF21B as a susceptibility locus for multiple sclerosis. *J Med Genet.* 2010; 47:775–6. DOI: 10.1136/jmg.2009.075911 [PubMed: 20587413]
75. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol.* 2004; 337:635–45. DOI: 10.1016/j.jmb.2004.02.002 [PubMed: 15019783]
76. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, et al. A framework for the interpretation of de novo mutation in human disease. *Nat Genet.* 2014; 46:944–50. DOI: 10.1038/ng.3050 [PubMed: 25086666]
77. Torres JM, Gamazon ER, Parra EJ, Below JE, Valladares-Salgado A, Wacher N, et al. Cross-Tissue and Tissue-Specific eQTLs: Partitioning the Heritability of a Complex Trait. *Am J Hum Genet.* 2014; 95:521–34. DOI: 10.1016/j.ajhg.2014.10.001 [PubMed: 25439722]

78. Hunt KA, Mistry V, Bockett NA, Ahmad T, Ban M, Barker JN, et al. Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. *Nature*. 2013; 498:232–5. DOI: 10.1038/nature12170 [PubMed: 23698362]
79. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. *Nat Rev Genet*. 2009; 10:184–94. DOI: 10.1038/nrg2537 [PubMed: 19223927]
80. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KCC, et al. A genome-wide association study of global gene expression. *Nat Genet*. 2007; 39:1202–7. DOI: 10.1038/ng2109 [PubMed: 17873877]
81. 1000 genome database n.d. <http://www.1000genomes.org/data>.
82. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, et al. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res*. 2006; 34:D590–8. DOI: 10.1093/nar/gkj144 [PubMed: 16381938]
83. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007; 81:559–75. DOI: 10.1086/519795 [PubMed: 17701901]
84. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*. 2009; 5:e1000529.doi: 10.1371/journal.pgen.1000529 [PubMed: 19543373]
85. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. *Nat Methods*. 2012; 9:179–81. DOI: 10.1038/nmeth.1785 [PubMed: 22138821]
86. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet*. 2010; 11:499–511. DOI: 10.1038/nrg2796 [PubMed: 20517342]
87. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*. 2004; 101:6062–7. DOI: 10.1073/pnas.0400782101 [PubMed: 15075390]
88. Wu G, Feng X, Stein L. A human functional protein interaction network and its application to cancer data analysis. *Genome Biol*. 2010; 11:R53.doi: 10.1186/gb-2010-11-5-r53 [PubMed: 20482850]

Highlights

- We evaluate the importance of missense SNPs (those changing an amino acid) in complex trait disease.
- The analysis covers 356 GWAS loci associated with risk of seven diseases.
- 33% of loci have at least one predicted high impact msSNP, providing a putative molecular mechanism.
- These data support a substantial role for this class of SNPs in susceptibility to common human disease.

Figure 1. Fraction of loci containing putative high impact missense (red bars) and any missense SNPs (blue bars) for each of the seven diseases. The fraction of high impact SNPs varies widely.

Figure 2. Distribution of D' and r^2 for the high impact candidate missense SNPs. Most high impact SNPs have high D' values, but small r^2 .

Figure 3. Distribution of odds ratio of risk alleles of high impact missense SNPs. Most odds ratios are reasonably small, with 51% less than 2.

Figure 4.

Examples of proposed high impact missense SNPs involved in complex trait disease mechanism.

(a) Crohn's Disease candidate SNP in MST1 (Macrophage stimulating protein) (green chain), resulting in reduced binding affinity to the RON receptor (cyan) and so reduced macrophage signaling. The ARG 703 (yellow) substitution by CYS (purple) lies close to the protein-protein interface (modeled from the hetero complex, PDB ID: 4QT8).

(b) Type 1 Diabetes and Rheumatoid Arthritis candidate SNP in PTPN22 (protein tyrosine phosphatase non-receptor type 22), weakening a protein-protein association, affecting T-cell activation. The ARG 620 (yellow) to TRP substitution in PTPN22 (cyan chain) introduces a steric clash (red disks) with W47 of the SH3 of Csk (c-src tyrosine protein kinase) (green chain) (from NMR model, PDB ID: 1JEG).

(c) Coronary Artery Disease candidate SNP in ADAMTS13 (ADAM metalloproteinase with thrombospondin type 1 motif, 13), reducing the level of mature protein, and hence reducing cleavage of VWF (Von Willebrand factor) and downstream processes related to plaque formation. The PRO 618 (yellow) to ALA (purple) substitution lies in the $\beta 6$ – $\beta 7$ loop of the spacer domain (green) and interacts with the C_A domain (cyan) (modeled from PDB ID: 3GHM).

(d) Type 1 Diabetes candidate SNP in IL7R (interleukin 7 receptor), putatively impairing interaction with IL7, and so affecting immune signaling. The HIS 165 (yellow) to GLU (purple) substitution affects a surface loop involved in interaction with the gamma chain of the signaling complex. (modeled from PDB ID: 3UP1).

(e) Crohn's Disease candidate SNP in PLCL1 (Phospholipase C-like 1). The VAL 667 (yellow) substitution to the bulkier ILE (purple) likely distorts the substrate binding site, affecting the catalytic efficiency (modeled from PDB ID: 1DJX).

(f) Type 1 Diabetes candidate SNP in SIRPB1 (signal regulatory protein beta 1) involved in immune regulation. The ILE 229 (yellow) substitution by bulkier MET (purple) results in destabilization of the protein structure (PDB ID: 2WNG). The exact role of this protein immune regulation is not yet clear.

(g) Hypertension candidate SNP in TUBB1 (tubulin beta 1, class VI) related to Coronary Syndrome and Myocardial Infarction. The GLU 43 (yellow) to PRO (purple) substitution results in predicted destabilization of the protein structure through loss of polar-polar residue interactions, expected to result in lower *in vivo* concentration (modeled from PDB ID: 3RYC).

(h) Rheumatoid Arthritis candidate SNP in INHBC (inhibin, beta C), related to the serum uric acid level. Substitution of ARG 322 (yellow) by GLN (purple) in INHBC (inhibin, beta C) is predicted to destabilize the protein structure through loss of salt bridges (red broken lines) with a spatially close conserved aspartic acid, expected to result in lower *in vivo* concentration (modeled from PDB ID: 2ARV).

Figure 5. Workflow for the identification of candidate high impact msSNPs potentially involved in disease mechanism.

Table 1

Number of risk loci for each of the seven diseases.

Disease	Locus source type		Total loci
	Set 1: WTCCC1 loci showing the strongest association signals	Set 2: Loci from the GWAS catalog for other follow-up studies and meta-analyses	
Bipolar Disorder	1	68	69
Coronary Artery Disease	1	45	46
Crohn's Disease	9	81	90
Hypertension	0	17	17
Rheumatoid Arthritis	2	35	37
Type 1 Diabetes	5	49	54
Type 2 Diabetes	3	40	43