Published online 8 June 2016

Nucleic Acids Research, 2016, Vol. 44, No. 18 ***8641–8654***
doi: 10.1093/nar/gkw519

Explaining the disease phenotype of intergenic SNP through predicted long range regulation

Jingqi Chen[1,2] and Weidong Tian[1,2,*]

[1]State Key Laboratory of Genetic Engineering and Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University, Shanghai 200436, P.R. China and [2]Department of Biostatistics and Computational Biology, School of Life Sciences, Fudan University, Shanghai 200436, P.R. China

Received January 08, 2016; Revised May 05, 2016; Accepted May 29, 2016

ABSTRACT

Thousands of disease-associated SNPs (daSNPs) are located in intergenic regions (IGR), making it difficult to understand their association with disease phenotypes. Recent analysis found that non-coding daSNPs were frequently located in or approximate to regulatory elements, inspiring us to try to explain the disease phenotypes of IGR daSNPs through nearby regulatory sequences. Hence, after locating the nearest distal regulatory element (DRE) to a given IGR daSNP, we applied a computational method named INTREPID to predict the target genes regulated by the DRE, and then investigated their functional relevance to the IGR daSNP's disease phenotypes. 36.8% of all IGR daSNP-disease phenotype associations investigated were possibly explainable through the predicted target genes, which were enriched with, were functionally relevant to, or consisted of the corresponding disease genes. This proportion could be further increased to 60.5% if the LD SNPs of daSNPs were also considered. Furthermore, the predicted SNP-target gene pairs were enriched with known eQTL/mQTL SNP-gene relationships. Overall, it's likely that IGR daSNPs may contribute to disease phenotypes by interfering with the regulatory function of their nearby DREs and causing abnormal expression of disease genes.

INTRODUCTION

Genome-wide Association Studies (GWAS) are designed to simultaneously examine the association of millions of genetic variants with target traits. To date, more than 2000 GWAS studies have contributed to the discovery of thousands of disease-associated variants, including inserts, deletions, copy number variations (CNV) and single nucleotide polymorphisms (SNP) (1). Elucidation of the phenotypic-association mechanisms for these variants is of great importance for understanding the molecular details of disease onset and progression and developing novel therapeutic approaches (1,2). So far, most work has been focused on disease-associated SNPs (daSNPs) located in coding regions, especially the non-synonymous SNPs which may alter the biochemical function of coded proteins. Nevertheless, the majority of daSNPs determined so far are located in non-coding regions (93% as reported by Maurano *et al*. (3)), including introns, Long Terminal Repeats (LTRs) and intergenic regions, presenting major challenges to the community on interpreting their involvement in diseases.

Increasing evidence has shown that non-coding daSNPs are often located in or closely linked to regulatory regions (3), suggesting that they may interfere with the normal functioning of their host regulatory elements (HREs). Recent methods for prioritizing daSNPs have incorporated the association of non-coding daSNPs with regulatory sequences. For example, Trynka *et al*. (4) have shown that H3KEme3 could help prioritize disease-causal SNPs. Claussnitzer *et al*. (5) and Khurana *et al*. (6) both incorporated transcription factor (TF) binding information to screen for disease-driving non-coding SNPs. Huang *et al*. (7) used the ChIP-Seq intensity variation of TFs and histone markers to identify functional non-coding SNPs that could disrupt enhancer activities. The mechanisms for daSNPs to affect the function of their HREs could be through the interruption of TF binding to regulatory sequences. For instance, Tokuhiro *et al*. (8) found that a daSNP for rheumatic arthritis is in the intron of *SLC22A4*, and may affect the binding of RUNX1. Pomerantz *et al.* found that a daSNP associated with prostate cancer and located in an enhancer of 8q24 affected the binding of Tcf-4 to this enhancer, which was needed for the regulation of its target gene- -*MYC* (9,10). Regulatory sequences may also be transcribed. Accumulating lines of evidence have suggested that enhancers could encode RNA transcripts with regulatory roles (11). It is therefore likely that daSNPs located in intergenic regulatory sequences such as enhancers might result in the transcription of abnormal RNA transcripts, which could then influence disease-related pathways and trigger disease phenotypes. Indeed, recent studies have found that the intergenic

*To whom correspondence should be addressed. Tel: +86 21 51630723; Fax: +86 21 51630723; Email: weidong.tian@fudan.edu.cn

trans-regulatory RNA transcripts that harbor a daSNP and target the cell cycle progression and differentiation pathways could result in multiple common diseases (12,13). This thus highlighted the functional significance of non-coding daSNPs. Intergenic regulatory sequences may also be involved in inter-chromosome interactions that may have important roles in maintaining chromosome conformations, which could be disrupted by daSNPs, causing disease phenotypes. Given that the regulatory sequences mainly assist to regulate their target genes' expression through TF binding, the encoded RNA transcripts or other mechanisms, in order to interpret the direct phenotypic association of non-coding daSNPs, we have to identify the target genes that are regulated by the host regulatory sequences potentially affected by the daSNPs.

Regulatory elements within coding gene regions typically regulate the host coding genes, though exceptions exist. However, for those regulatory elements outside coding gene regions, i.e. distal regulatory elements (DREs), determining the target genes they regulate is a challenging task. This is because DREs can regulate target genes from quite a long distance (14) and their relationships with the target genes may be beyond a one-to-one pattern (15). Given that DREs may regulate target genes through a process called DNA looping (16), Chromosome Conformation Capture (3C) (17) and the further developed Hi-C (18) and ChIA-PET (19) techniques aimed at capturing long-range chromatin interactions are potentially useful for identifying DREs-target gene relationships. However, DREs may regulate target genes in a dynamic and tissue-specific way, while current Hi-C or ChIA-PET data were applied to only a few cell lines and may capture a subset of DRE-target gene relationships. Computational predictions can help to fill in the missing subsets. For example, based on the hypothesis that there may be a consistency of DRE-target gene relationships in multiple species given their essentiality, Lu *et al.* (20) combined the phylogenetic profile correlation with Hi-C/ChIA-PET data to predict target genes for DREs. Corradin *et al.* (21) also developed a method based on the association of epigenetic markers on enhancers and the cell-line specific expression of genes, while He *et al.* (22) employed a probabilistic approach named IM-PET based on a number of genomics features. Ernst *et al.* (23) and Thurman *et al.* (24) also provided predictions of enhancer-target gene pairs through multi-cell activity profile correlation and DNaseI signal correlation, respectively. Each of the above-mentioned methods has their own merit, and a combination of them may provide more coverage for the DREs and their target genes.

In this study, we focused on a challenging category of non-coding daSNPs—the daSNPs located in intergenic regions (IGR), by using the predicted target genes of their HREs to explain their disease-associated mechanisms. Our rationale was as followed: target genes regulated by the HREs of IGR daSNPs can be predicted using computational methods; if the predicted genes are strongly functionally associated with the corresponding diseases, it is likely that the IGR daSNPs may contribute to disease phenotypes by causing abnormal expression of the predicted target genes through the disruption of the HREs. In fact, this scenario has been reported in a study on aniridia (25). Here, we predicted the target genes regulated by the HRE of IGR daSNPs using different source data including phylogenetic profile correlation and Hi-C/ChIA-PET data, and also combined other published predictions (21–24). Our results revealed that disease phenotypes of a high percentage of IGR daSNPs could be interpreted through the above rationale.

MATERIALS AND METHODS

Data downloading and processing

NHGRI catalog of GWAS SNPs (1), GWASdb (26), HuGE Navigator (27) and Johnson and O'Donnell's collection (28) were downloaded and combined. Non-coding LD SNPs ($r^2 > 0.8$) to the daSNPs in IGR were identified using the CEPH population-based HapMap data (ftp://ftp.ncbi.nlm.nih.gov/hapmap/, release 27). Disease genes were retrieved from the GAD (29) database (http://geneticassociationdb.nih.gov), and were required to have the evidence of 'experiment'. ICD-10 (http://www.who.int/classifications/icd/en/) annotations were used to make disease names from different sources comparable. Dnase I hypersensitivity site (DHS) annotations were downloaded from the UCSC genome browser (30). Protein-coding gene annotations were obtained from Lu *et al.* (20). The human eQTL/mQTL data sets used in this study were downloaded from the GEO database (http://www.ncbi.nlm.nih.gov/geo); these data sets included all that became publicly available by the end of 2014 (the list of these data is shown in Supplementary Table S1). The eQTL/mQTL data consisted of SNP IDs and corresponding gene IDs or gene regions. The gene regions in the eQTL/mQTL data were converted to gene IDs used in this analysis if they overlapped with the regions of the gene IDs for at least 1 bp. In March 2016, we searched in the ENCODE (15) website (http://www.encodeproject.org) and found 72 samples of human Small RNA-seq data (20–200 nt, ribominus) in bam format, corresponding to a total of 17 human tissues and 15 human cell lines. We downloaded and processed these bam data. Read counts for HREs were calculated by featureCounts (32). Only uniquely mapped reads were used for the analysis. Super-enhancer and typical enhancer data for 36 tissues and 50 cell lines were downloaded from the Supplementary Data of the paper by Hnisz *et al.* (33).

The component methods in INTREPID

INtegrated TaRget gEne PredItion (INTREPID) combined the predictions made by five component methods: HIC, PPC, IM-PET, PreSTIGE and ENCODE. Below, we briefly described the methodology of each of these methods.

The HIC method. HIC utilized Hi-C and ChIA-PET data to predict the target genes for DREs. Hi-C data and ChIA-PET data of human cell lines generated from 8 studies (Supplementary Table S2) were downloaded from the GEO database (http://www.ncbi.nlm.nih.gov/gds). For each Hi-C/ChIA-PET data, we first identified the peak regions that were enriched with Hi-C/ChIA-PET reads. To do it, we recorded all chromosome positions mapped by Hi-C reads and determined the read count for each of these positions.

The read counts were converted into z-scores, and a peak region was defined as a region of continuous positions with z-scores ≥ 1.645. Nearby peak regions between which the distance was smaller than 1 kb were further merged together. Next, we identified the peak regions that were interacting with each other. To do it, we first recorded all pairs of peak regions in the same chromosome that were connected by one or more Hi-C/ChIA-PET reads. Then, for every candidate pair we computed an Odds Ratio (OR) value = observed number of reads (normalized)/expected number of reads. The observed number of reads was computed as the number of observed Hi-C/ ChIA-PET reads connecting the two peak regions divided by the geometric average of the length of the two peak regions. The expected number of reads was defined as the average of the observed number of reads for all candidate pairs of peak regions in which the distance between the two peaks was greater than or equal to that between the two tested peaks. A candidate pair of peak regions was considered positive if its OR was larger than 2. Finally, a positive pair of peak region was predicted as a DRE-target gene pair if one peak region overlapped with a DRE, and the other peak region overlapped with a DHS within a protein-coding gene region (−1 kb - Transcription End Site). The target genes predicted from each of the 8 Hi-C/ChIA-PET data sets were then combined to form the final predictions.

The PPC method. PPC predicted the target genes regulated by a DRE based on the phylogenetic profile correlation between a DRE and a protein-coding gene. The phylogenetic profiles for DREs and protein-coding genes were constructed following the procedures described in Lu *et al.* (20). For protein-coding genes, the phylogenetic profile was constructed using the promoter sequence (−1 kb to Transcription Start Site (TSS)). The Pearson correlation coefficient (PCC) between the phylogenetic profiles of a DRE and a protein-coding gene located in the same chromosome was computed. Because the distance between a DRE and a protein-coding gene would affect the calculated PCC, for each chromosome we divided all pairs of DREs-protein-coding genes into three groups: within 50 kb, 50–500 kb, or above 500 kb and then calculated the average and standard deviation of the PCCs for each group. Then, for each DRE all protein-coding genes within the same chromosome were ranked by the z-score of their PCCs with the DRE in each distance group, and the top five ranked genes (z-score ≥ 2) were predicted as the target gene of the DRE.

PreSTIGE, IM-PET and ENCODE. PreSTIGE (21), IM-PET (22) and ENCODE (23,24) all used epigenetic data to predict the target genes for DREs. We collected the predicted DRE-target gene pairs from the respective publications of these three methods. Below, we briefly described their methodologies. PreSTIGE (21) predicted the target genes for enhancers by pairing cell-type specific H3K4me1 marker in enhancers and the cell-type specific expressed genes in multiple cell types, and incorporating CTCF binding sites as boundaries. IM-PET (22) integrated a number of genomic features including promoter-enhancer activity correlation and transcription factor-promoter correlation to build a probabilistic model for predicting regulatory enhancer-promoter pairs. For ENCODE, we combined the predictions made by two ENCODE publications. One ENCODE paper utilized the correlation between enhancer mark intensities and gene expression profiles to train logistic regression classifier for identifying potential enhancer-target gene relationships (23). The other ENCODE paper was based on the correlation between cross cell-type DHS signals of enhancers and nearby promoters (24).

The combination of the predicted target genes by the five methods. Given a DRE, we applied HIC and PPC separately to predict the target genes. For each of PreSTIGE, IM-PET and ENCODE, based on the predicted regulatory relationships downloaded from their respective publications we identified the regulatory sequence that overlaps with the DRE (at least 1 bp overlap), and then selected the corresponding target genes as the predictions for the DRE by that method. The gene IDs used by that method were converted to the IDs we used in this analysis based on the annotations of RefSeq (http://www.ncbi.nlm.nih.gov/refseq). In case a method predicted enhancer-promoter interactions instead of regulatory sequence-target gene relationships, we converted the promoters into gene IDs by requiring the predicted promoter to overlap with the annotated promoter of a gene ID (at least 1 bp overlap). Finally, we combined the predicted target genes made by each of the five methods as the final prediction for that DRE, i.e. the union of the predicted target genes, and named this approach as INTREPID.

Evaluation of the functional relevance between the predicted target genes of an IGR daSNP and the corresponding disease genes

Given an IGR daSNP-disease associations, we collected two groups of genes: the target genes predicted by INTREPID, and the disease genes annotated by the GAD (29) database. Then, we applied three methods to evaluate the functional relevance between the two groups of genes, which are the occurrence analysis, the over-representation analysis (ORA) and the relevance analysis. The occurrence analysis simply investigated whether there were any overlaps between the two groups. The ORA analysis was done using the fisher.test function in R, and the background genes were defined as the genes located in the same chromosome as that of IGR daSNP. The significance level was set at 0.05. When multiple tests were performed, the function p.adjust (method = 'fdr') in R was used to correct the *P*-values and the significance level was set at 0.05.

The relevance analysis was based on a functional association network—the STRING network (34). The functional relevance between a pair of genes was defined as their shortest path length in the STRING network using the Dijkstra algorithm (35). For each predicted target gene, we computed the minimum shortest path length between it and all disease genes (termed $SP_{\text{gene-disease}}$), and then transformed it into a relevance score between 0 and 1: Relevance $\text{Score}_{\text{gene-disease}} = 1/(1+SP_{\text{gene-disease}})$. The mean relevance score of all predicted target genes was then computed to indicate the functional relevance between the predicted target genes and the disease genes. To assess its significance,

we randomly selected the same number of genes as that of the predicted target genes, and computed a mean relevance score following the above-described procedure. This process was repeated 1000 times, and the *P*-value of the observed mean relevance score was derived thereafter. The significance level was set at 0.05. When multiple tests were performed, the function p.adjust (method = 'fdr') in R was used for correcting the *P*-values, and the significance level was set at 0.05.

Disease-disease similarity

The similarity between two diseases was defined as the Jaccard similarity calculated as the ratio of the number of overlapped disease genes to the number of the union of disease genes.

RESULTS

A combined approach to predict the target genes for the HREs of intergenic daSNPs

A total number of 5639 daSNPs were collected from GWAS catalog (1), GWASdb (26), HuGE Navigator (27) and Johnson and O'Donnell's collection (28). Among these daSNPs, 1834 were located in IGR (Figure 1A), and were associated with 128 diseases (Supplementary Table S3). A total of 32% of the IGR daSNPs were associated with two or more diseases. In total, there were 2774 IGR daSNP-disease associations. We also collected all linkage dis-equilibrium (LD) SNPs ($r^2 > 0.8$ in the CEU population according to HapMap (36)) to IGR daSNPs, and found that 1472 (80.3%) IGR daSNPs had LD SNPs in the intergenic regions, with a median of 10.5 LD SNPs each. To determine the relationships between IGR daSNPs and regulatory sequences, we downloaded annotation of Dnase I hypersensitivity sites (DHSs) from UCSC Genome Browser, and defined the IGR DHSs as putative DREs. A total of 85% (1562) of IGR daSNPs were within ±1 kb distance to the boundaries of one or more DREs, and this proportion increased to 95.6% if the LD SNPs of IGR daSNPs were also considered (Figure 1B), i.e. if any one of the LD SNPs were within ±1 kb distance to the boundaries of one or more DREs. Given such a strong association, it was reasonable to hypothesize that IGR daSNPs may interfere with the normal function of their nearby regulatory sequences, a consequence of which may cause abnormal expression of potential target genes, leading to disease phenotypes. For convenience, we defined the nearest DRE (within ±1 kb) to an IGR daSNP as its host regulatory element (HRE). Hnisz *et al.* (33) defined super-enhancer and typical enhancer regions in 86 types of human tissues and cell lines. This can be used as additional information to see whether the HREs of IGR daSNPs could be functional. We downloaded these defined enhancer regions. Among the 1562 IGR daSNPs that had HREs (not counting in LD information), we found that the HREs of 1011 of them were overlapping with the enhancer regions defined by Hnisz *et al.* (33). The high level of overlaps indicated that the IGR daSNPs used in our later analysis were functional with a high possibility, and that the majority of their HREs might be enhancers. The above-proposed disease-association mechanism of IGR daSNP would be strongly supported if the target genes regulated by the HRE were known disease genes or were functionally associated with known disease genes. Thus, a key step toward understanding the disease-association mechanism of an IGR daSNP was then to determine the target genes regulated by its HRE.

Lu *et al.* previously developed a method that combined phylogenetic profile correlation and Hi-C data to predict the target genes of DREs (20). Recently, there were two studies that used other genome features for similar purposes (21,22). In addition, the ENCODE project has also provided lists of DRE-target gene relationships (23,24). Thus, we adopted an integrated approach named INtegrated TaRget gEne PredItion (INTREPID) for DREs that combined the predictions made from Hi-C/ChIA-PET data (named HIC), Phylogenetic Profile Correlation (named PPC), IM-PET (22), PreSTIGE (21) and ENCODE (23,24) (see Materials and Methods). Note that the predictions were made only for genes located in the same chromosome with the DREs. Here, for 1562 IGR daSNPs whose HREs could be identified, we applied INTREPID to predict the target genes for their corresponding HRE. The predicted target genes were considered as the target genes of the IGR daSNP, i.e. they were likely to be affected by the IGR daSNP. A median number of nine target genes were predicted for each IGR daSNP.

Recently, a number of studies have been conducted to identify expression quantitative trait loci (eQTL) or methylation quantitative trait loci (mQTL) at genome-scale (37). We downloaded eight sets of eQTL/mQTL data from the GEO database (http://www.ncbi.nlm.nih.gov/geo), which included the eQTL/mQTL data from blood, liver, intestine and brain tissues of Caucasian or European populations (Supplementary Table S1). In addition, we collected the eQTL data from GTEx (38) produced by the BROAD Institute. From these data, we identified a total number of 950 SNP-gene pairs where the SNP was an IGR daSNP serving as either an eQTL or an mQTL to the paired gene. These SNP-gene pairs corresponded to 404 IGR daSNPs and 166 genes. Among these pairs, 349 were between an IGR daSNP and its predicted target gene. In contrast, an average of 9 would be obtained if we randomly selected the same number of genes as the predicted target genes for each IGR daSNP (*P*-value < 0.001, experiments were repeated 1000 times). This result strongly supported that the expression of the predicted target genes were likely affected by IGR daSNPs, partly validating INTREPID's predictions.

The predicted target genes of IGR daSNPs can help to explain their disease phenotypes

We obtained disease gene annotations from GAD (29) (http://geneticassociationdb.nih.gov). For each IGR daSNP-disease association, we prepared two groups of genes: the predicted target genes, and the disease genes annotated with the corresponding disease phenotype. Then, we investigated whether the predicted target genes (i) consisted of one or more disease genes (the occurrence analysis), (ii) were over-represented with disease genes (the ORA analysis), or (iii) were functionally relevant to disease genes inferred using a network-based approach (the

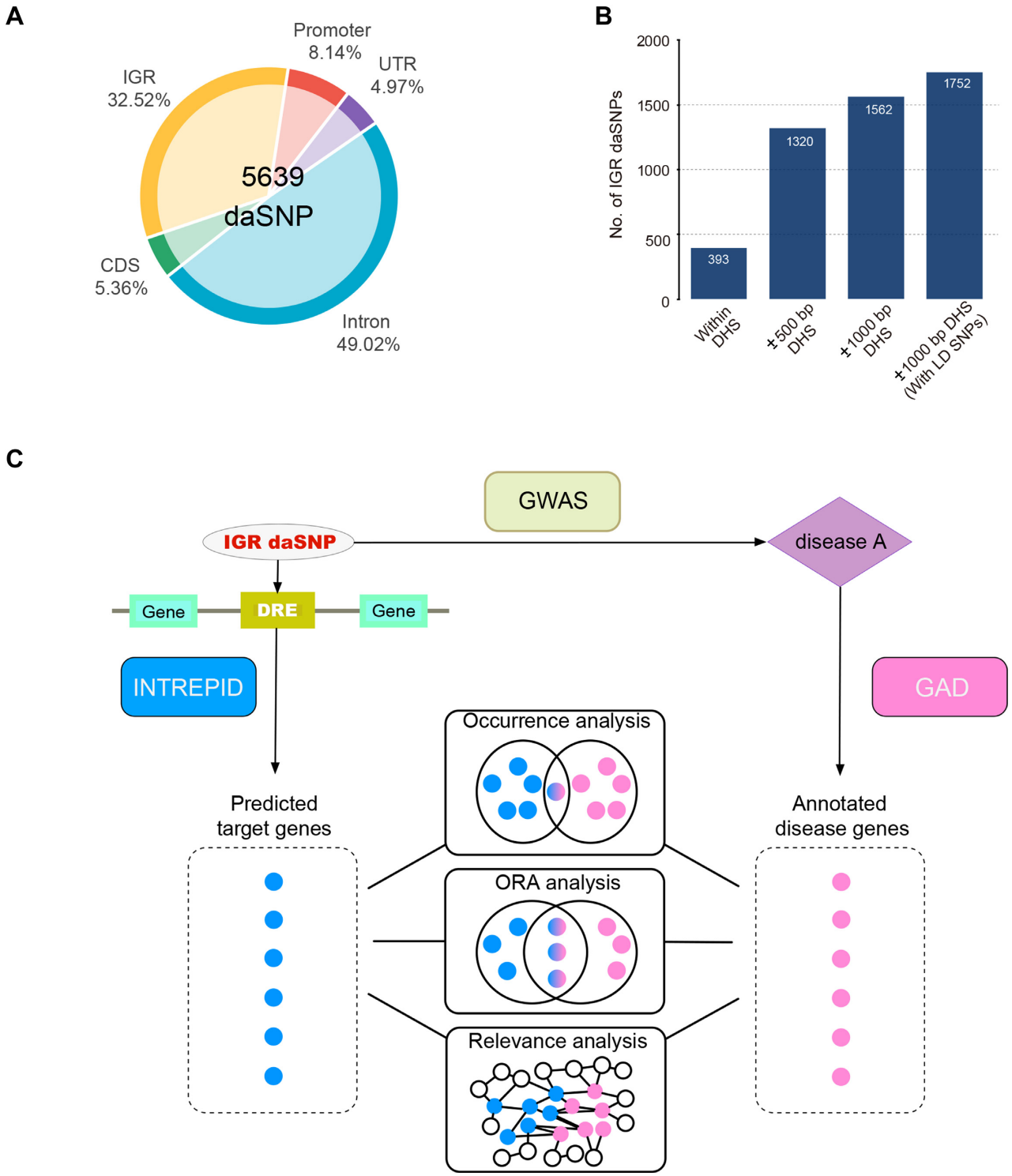

Figure 1. (**A**) The proportions of daSNPs with respect to their relative locations to protein-coding genes in the genome. IGR refers to intergenic region. (**B**) The numbers of IGR daSNPs located at different distance cutoff to their nearest DHSs. (**C**) The workflow for explaining the disease phenotype of IGR daSNPs.

relevance analysis) (see Materials and Methods and Figure 1C).

Out of 2774 IGR daSNP-disease associations being investigated, the predicted target genes produced positive results for 753 (27.1%), 524 (18.9%), 823 (29.7%) by the occurrence, the ORA (P-value ≤ 0.05, Fisher's exact test) and the relevance analysis (P-value ≤ 0.05 by permutation), respectively (Figure 2A, Supplementary Table S3). Nearly all positive results found by the ORA analysis were also found by the relevance analysis. In contrast, with randomly selected genes as the target genes we observed an average of 352 and 24 IGR daSNP-disease associations using the occurrence and the ORA analysis, respectively, both significantly smaller than the observed numbers (P-value < 0.001, experiments repeated for 1000 times, Figure 2A). The relevance analysis already integrated the comparison with randomly selected genes (see Materials and Methods). The corresponding numbers of IGR daSNPs with positive results were 672 (36.6%), 342 (17.7%) and 549 (29.9%), respectively (Figure 2B). Since some daSNPs may be tag SNPs whose closely linked SNPs might be the true daSNPs (21), we also investigated the predicted target genes of the LD SNPs, and considered an IGR daSNP-disease association positive if the predicted target genes of either the IGR daSNP or any of its LD SNPs generated a positive result. The numbers of IGR daSNP-disease associations with positive results increased to 1168 (42.1%), 1004 (36.2%) and 1504 (54.2%) for the occurrence, the ORA, and the relevance analysis, respectively (Figure 2A, Supplementary Table S4), and the corresponding numbers of IGR daSNPs with positive results increased to 995 (54.3%), 657 (35.8%) and 997 (54.4%), respectively (Figure 2B).

Positive results obtained using different analyses represented different confidence on explaining the disease-association mechanism for IGR daSNPs. Accordingly, we classified an IGR daSNP-disease association as 'highly likely', 'mechanistically likely' and 'potentially likely' explainable by the predicted target genes if the ORA analysis, the relevance but not the ORA analysis, or only the occurrence analysis produced a positive result. The numbers of IGR daSNP-disease associations falling into these three categories are 524, 302 and 194, respectively (Figure 2C), altogether covering 36.8% of all IGR daSNP-disease associations. When LD SNPs included, the above three numbers became 1004, 502 and 173, respectively (Figure 2D), together covering 60.5% of all IGR daSNP-disease associations. However, because LD SNPs provided indirect evidence, those IGR daSNP-disease associations supported only by LD SNPs were considered with less confidence.

To conclude, the above results strongly supported our proposed disease-association mechanism for IGR daSNPs: a daSNP affected the function of its HRE, causing abnormal expression of target genes that were relevant to the disease, and consequently leading to the disease phenotype.

Case reports for the 'explainable' IGR daSNP-disease associations

Below, we presented a couple of examples for each of the 'highly likely', 'mechanistically likely' and 'potentially likely' explainable categories of IGR daSNP-disease associations with and without the inclusion of LD SNPs.

The association of SNP rs2857161 (Chromosome 6) with multiple sclerosis (39) was considered 'highly likely' explainable by the nine predicted target genes, among which five (*TAP2*, *HLA-DOB*, *MICA*, *MICB* and *TAP1*) were known associated to this disease (29) (ORA enrichment P-value = 9.3e-4). Moreover, this SNP was an eQTL of both *TAP2* and *HLA-DOB* (38) (Figure 3A). In another example, rs4779584 (chromosome 15) was associated with colorectal cancer based on a GWAS study (40). This association was 'highly likely' explainable by the eight predicted target genes, in which *CYP19A1* and *GREM1* were known disease genes (ORA enrichment P-value = 0.026). In addition, rs4779584 was an eQTL of the gene *GREM1* (41) (Figure 3B). A further example was rs10757278 in Chromosome 9. Its associations with myocardial infarction and ischaemic heart disease (29) were 'highly likely' explainable by its 10 predicted target genes, among which *MTAP*, *CDKN2A* and *CDKN2B* were all known to associate with both diseases (Figure 3C). The ORA enrichment P-values were 3.61e-5 and 6.92e-5 for myocardial infarction and ischaemic heart disease, respectively. Experimental evidence has shown that the enhancers where rs10757278 is located interacted with the *MTAP* gene and the *CDKN2A/B* locus (42), supporting our prediction for the target genes affected by rs10757278.

SNP rs16940202 (Chromosome 16) was associated with colitis disease (43). None of the six predicted target genes was a known colitis gene. But they were significantly functionally associated with known colitis genes in the STRING network (P-value = 0.05), making this SNP-disease association 'mechanistically likely' explainable. In addition, this SNP was an eQTL of one of the predicted target genes—*IRF8* (38) (Figure 3D). Asthma associated SNP rs12950743 (44) (Chromosome 17) could also be 'mechanistically likely' explainable by the predicted target genes. This SNP was predicted to affect seven genes among which only *PRKCA* was a known gene for asthma (29), which failed to yield a positive result by the ORA analysis. Yet, the relevance analysis produced a positive result (P-value = 0.03) due to the functional association of the other predicted target genes with known asthma genes (Figure 3E).

An example for 'potentially likely' explainable IGR daSNP was rs6983267 (Chromosome 8). This SNP was associated with nasopharyngeal cancer, UADT cancer, colorectal cancer and prostate cancer (26). Of the 10 predicted target genes for this SNP, *NSMCE2* was associated with prostate cancer (29) and *MYC* was associated with both colorectal cancer and prostate cancer (29), making this SNP 'potentially likely' explainable for the two phenotypes. This SNP was located in an enhancer that was shown by *in vitro* experiment to physically interact with the promoter region of *MYC* (9,10), supporting our prediction that *MYC* was a target gene of rs6983267. Moreover, experiments with transgenic mice showed that the activity of the enhancer and the expression of *MYC* were both significantly increased with the presence of the risk allele of rs6983267 (45), suggesting that this SNP was functional. Therefore, it is likely that this SNP affects the function of its host enhancer, causing enhanced expression of *MYC* in prostate and then increasing the chance of having prostate cancer. Another exam-

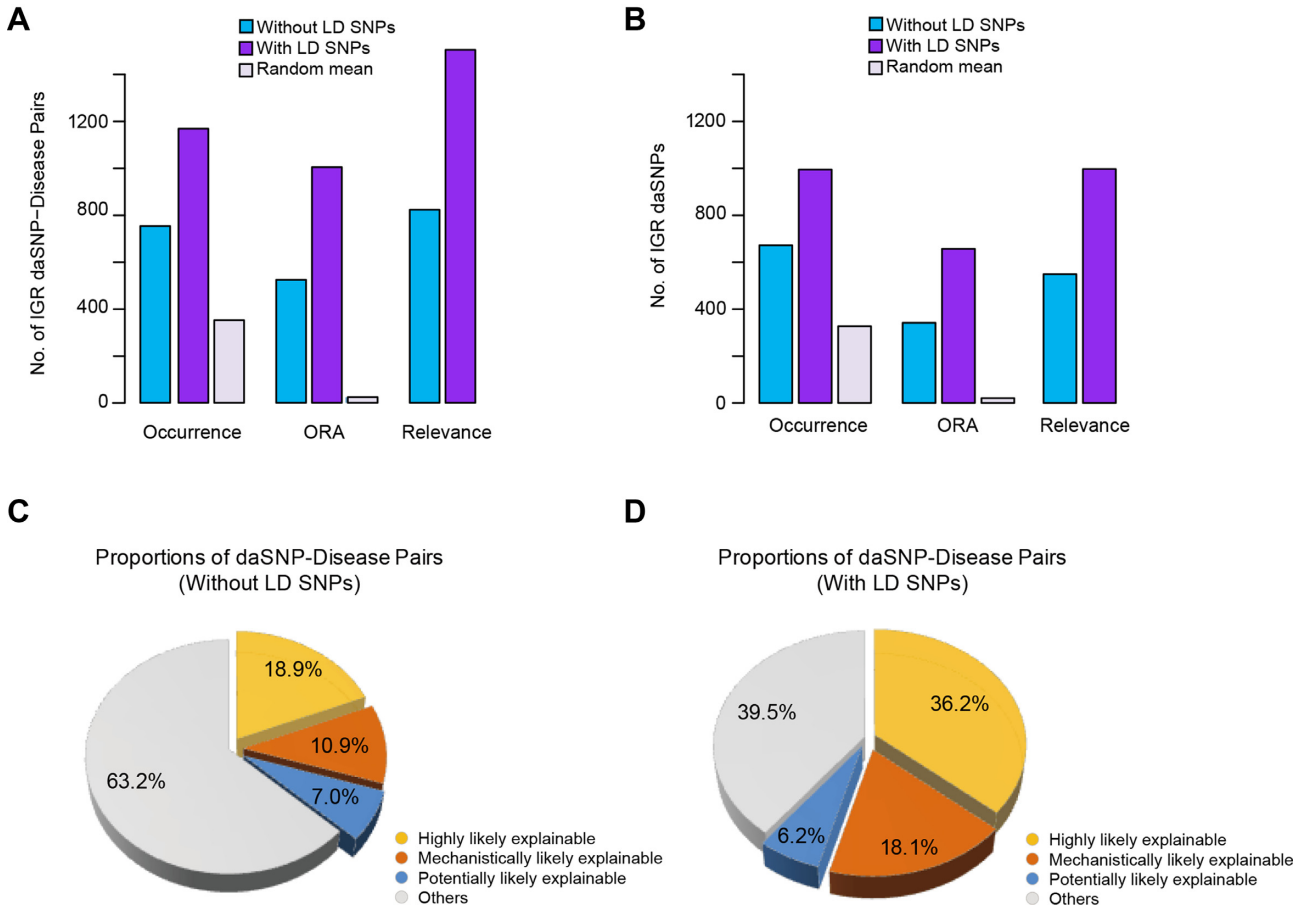

Figure 2. (**A**) The numbers of IGR daSNP-disease associations found positive by different methods (the occurrence, the ORA and the relevance analysis) without or with considering LD SNPs. Random refers to the results using the same number of randomly selected genes as the predicted target genes (LD SNPs were not considered for Random). (**B**) Similar to (A) except that the numbers of positive IGR daSNPs were reported. (**C** and **D**) show the proportions of different categories of explainable IGR daSNP-disease associations without or with considering LD SNPs, respectively.

ple for 'potentially likely' explainable SNPs was the SNP rs11257655 in Chromosome 10. This SNP was associated with type II diabetes (46), and two of its 17 predicted target genes, *CDC123* and *CAMK1D*, were associtaed with type II diabetes (29). In addition, rs11257655 was an eQTL site of the gene *CAMK1D* (38).

The association of SNP rs2292627 (Chromosome 10) with macular degeneration (MD) (26) could be explained by the predicted target genes of one of its LD SNPs—rs11200583, among the seven predicted target genes of which *PLEKHA1* and *HTRA1* were known MD genes (*P*-value = 4.3e-3). This LD SNP was also an eQTL of *PLEKHA1* (38) (Figure 3F). SNP rs6537296 (Chromosome 4) was associated with 'pulmonary heart disease and diseases of pulmonary circulation' (47). The ORA analysis found the predicted target genes of two LD SNPs (rs7697189 and rs1489762) of this daSNP were enriched with the corresponding disease genes, while the relevance analysis identified another four LD SNPs (rs6842889, rs11100860, rs995758 and rs1489759) with positive results. Thus, the disease phenotype of rs6537296 could be indirectly explained by the predicted target genes of its LD SNPs.

Prediction of candidate disease-contributing IGR daSNPs

In the above analysis, we have predicted a number of 'explainable' IGR daSNPs whose predicted target genes are associated with the corresponding disease phenotypes. However, these predictions only indicated associations. We hope to identify among these IGR daSNPs the disease-contributing ones. However, this is not an easy task. First of all, we need to know which tissues or cell types develop the disease. Secondly, we need to determine whether the HRE of the IGR daSNP is functional in the tissues or cell types. Thirdly, we need to verify that the predicted target genes are regulated by the HRE in that tissue or cell type, and demonstrate that the expression of one or more of them, particularly those related to the disease, could be altered by the daSNP. Finally, we need to prove that the altered expression of the target genes could indeed contribute to the disease. Unfortunately, there are no experimental data readily available for us to perform these analyses. Therefore, here we employed a simple strategy to approximate the process of finding the disease-contributing IGR daSNPs.

Recent studies have shown that intergenic regulatory sequences could encode small RNA transcripts of regulatory roles (14). Therefore, a regulatory sequence with transcribed

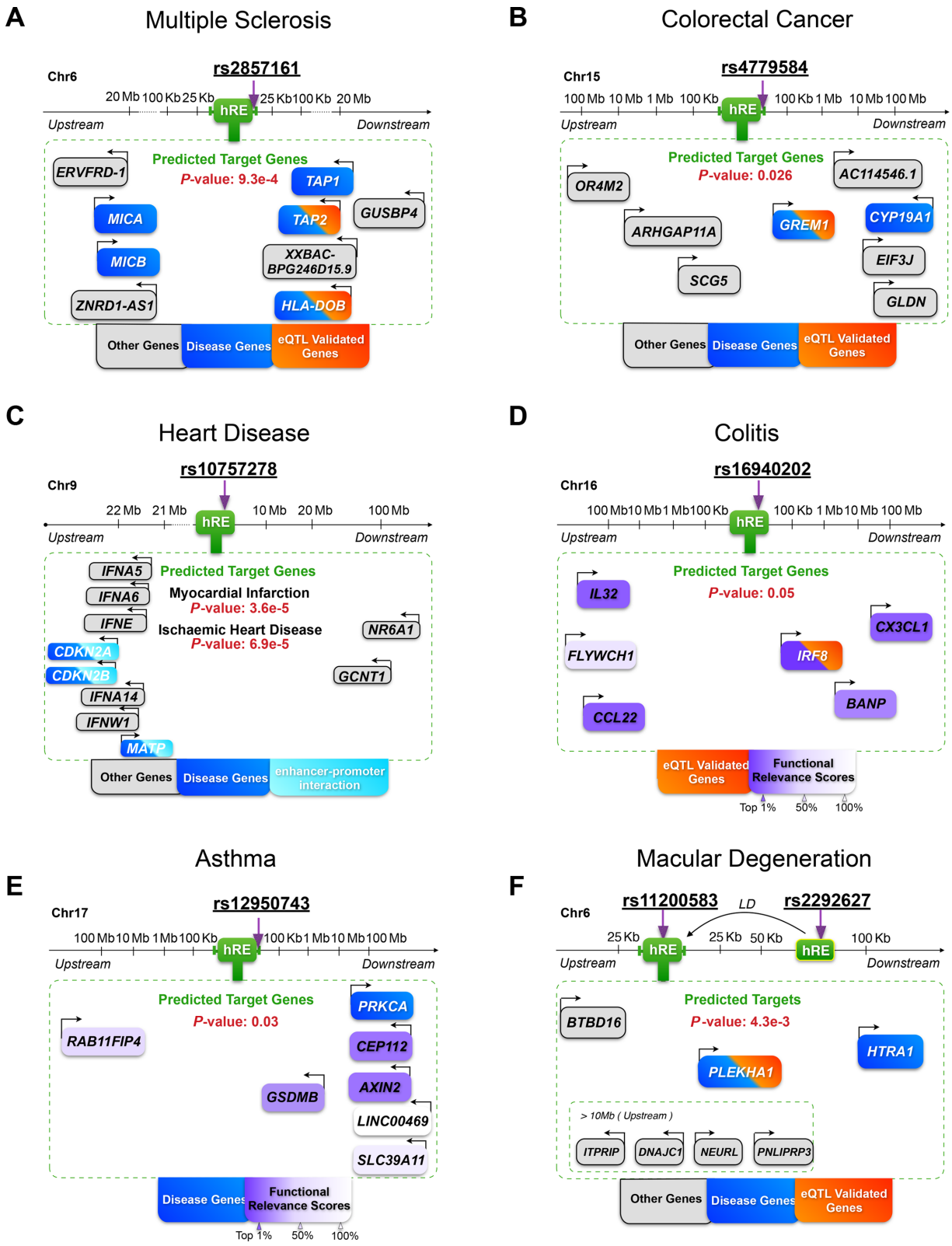

Figure 3. Examples of explainable IGR daSNP-disease pairs. (**A**–**C**) were three 'highly likely' explainable IGR daSNP-disease associations. (**D** and **E**) were two 'mechanistically likely' explainable IGR daSNP-disease associations. (**F**) was an IGR daSNP-disease association explained through the LD SNPs. In (**A**–**F**), predicted target genes were presented schematically according to their relative distances to the corresponding daSNP. Blue color referred to known disease genes, and orange color referred to eQTL/mQTL-validated genes. Grey color represented other target genes. The arrows upon a rectangle illustrated the transcription directions of the gene. In (**D** and **E**), rectangles in different concentration of purple represented genes with different ranks of functional relevance scores with the corresponding disease.

small RNA in a cell line or tissue could be considered functional in that cell line or tissue. Here, we downloaded the small RNA-seq data from the ENCODE project, which were generated from a total number of 17 human tissues and 15 cell lines (14). By mapping the small RNA-seq reads to the HREs of IGR daSNPs, we then identified the functional HREs in specific tissues or cell lines (at least two reads were mapped to the HRE in at least one sample of the tissue or cell line). Besides the small RNA information, we also used the previously defined enhancer regions in multiple cell lines and tissues (13) to determine the functional HREs in specific tissues or cell lines by checking whether they overlapped with the defined enhancer regions in that tissue or cell line. If the HRE of an IGR daSNP were functional in a tissue or a cell line where the disease develops, then the IGR daSNP would more likely be disease contributing. However, there were only a limited number of tissues or cell lines with small RNA or enhancer information. Therefore, we selected only three types of diseases—immunological diseases, cancers and neurological diseases whose corresponding tissues or cell lines had available small RNA-seq data or enhancer annotations (refer to Supplementary Table S5 for our definition of the tissues or cell lines corresponding to these three types of diseases). Since no large scale experimental data were available to verify the predicted target genes for the HREs of IGR daSNPs, here we assumed our predictions were reliable and also assumed the eQTL/mQTL data could support the regulatory relationship between an IGR daSNP and its target genes in the disease tissues or cell lines. Finally, we predicted an 'explainable' IGR daSNP as a candidate disease-contributing daSNP if it satisfied the following criteria: (i) its association with the disease was either highly likely explainable or mechanistically likely explainable; (ii) its HRE was functional in one or more disease tissues or cell lines; (iii) its regulatory relationship with at least one target gene was supported by eQTL/mQTL data. Based on these criteria, we identified 64, 5 and 2 candidate disease-contributing IGR daSNPs for immunological disease, cancers and neurological diseases, respectively (for details, refer to Supplementary Table S5). Since there were not much eQTL/mQTL data available, many of the 'explainable' IGR daSNPs were excluded from the candidate lists. Without requiring the support of eQTL/mQTL data, we obtained 139, 31 and 8 potentially disease-contributing IGR daSNPs for immunological disease, cancers and neurological diseases, respectively (Supplementary Table S5).

Below, we give an example of candidate disease-contributing IGR daSNP for each of the three types of diseases, respectively. SNP rs2284178 in Chromosome 6 was associated with three autoimmune-related diseases (diabetes, Bechet's disease and type I diabetes, respectively) (6) and all of them could be explained by the ORA analysis. The HRE of this SNP could transcribe small RNAs in five cell lines/tissues (15), in which two were lymphocytes–GM12878 (normal) and Karpas-422 (tumor). The HRE of this SNP also overlapped with the enhancer regions that were only active in immune cells (13). rs2284178 was also an eQTL to two predicted target genes—*HLA-B* and *HLA-C* (16). Therefore, it is very likely that rs2284178 may be contributing to immunological diseases by interrupting the corresponding enhancer function and resulting in the abnormal expression of *HLA-B* and *HLA-C*. rs4779584 in Chromosome 15 was associated with colorectal cancer (6), and was predicted to be disease-contributing. Its HRE overlapped with the defined enhancer regions in three cancer cell lines (VACO_400 (a colorectal cancer cell line), VACO_9M (a colon cancer cell line) and u87 (a primary glioblastoma cell line)). In addition, this SNP was an eQTL of one of its predicted target genes–*GREM1* (16). As a result, rs4779584 might contribute to cancers by interrupting the expression of its target genes. rs3101942 in Chromosome 6 was associated with narcolepsy (6), and was considered a disease-contributing SNP for neurological diseases. Its HRE could transcribe small RNAs in three types of nervous system related tissues/cell lines—bipolar spindle neuron, frontal cortex and SK-N-DZ. This SNP was an eQTL of the target gene *PSMB9* (38). Thus, it is likely that this SNP may be contributing to neurological diseases.

As we mentioned earlier, the identified candidate disease-contributing variants were obtained in a simplified and approximate way, and may not represent the real cases. Nevertheless, the variants and the information obtained through the process, particularly the predicted target genes, made it possible for experimentalists to focus on only a small number of genes to start with, and would be worthy of further exploitation.

The 'one-SNP-multiple-diseases' phenomena

When attempting to explain the disease phenotype of IGR daSNPs earlier, we analyzed the functional relevance between the predicted target genes of an IGR daSNP and the genes known to be associated with the corresponding disease phenotype. As the methods for functional relevance analyses were general, here we also applied both the ORA and the relevance analyses to inspect the functional relevance between the predicted target genes of an IGR daSNP and the disease genes from each of the 200 diseases in the GAD database. The ORA and the relevance analyses identified 270 and 548 IGR daSNPs whose predicted target genes were functionally relevant to one or more diseases (multiple test correction, FDR ≤ 0.05, Supplementary Table S6 and S7), respectively. In comparison, with randomly selected genes as the target genes the ORA analysis only found an average of 3.6 IGR daSNPs with one or more relevance diseases, which were significantly lower than the number obtained by using the predicted target genes (experiments were repeated 1000 times, P-value < 0.001). Note that the relevance analysis already incorporated the comparison with randomly selected genes. Among the IGR daSNPs with relevant diseases found by the ORA or the relevance analysis, 158 and 295 were relevant to their annotated diseases, respectively. Both numbers were less than what we obtained in the previous section because of multiple test correction.

Most IGR daSNPs with positive results were found relevant to multiple diseases, with a median number of 13 and 24.5 diseases each by the ORA and the relevance analyses, respectively (Figure 4A). A major reason for the 'one-SNP-multiple-diseases' phenomena was the frequent occurrence of 'one-gene-multiple-diseases' association. For example, in GAD more than 50% of disease genes were annotated to more than one disease, with 18% annotated to

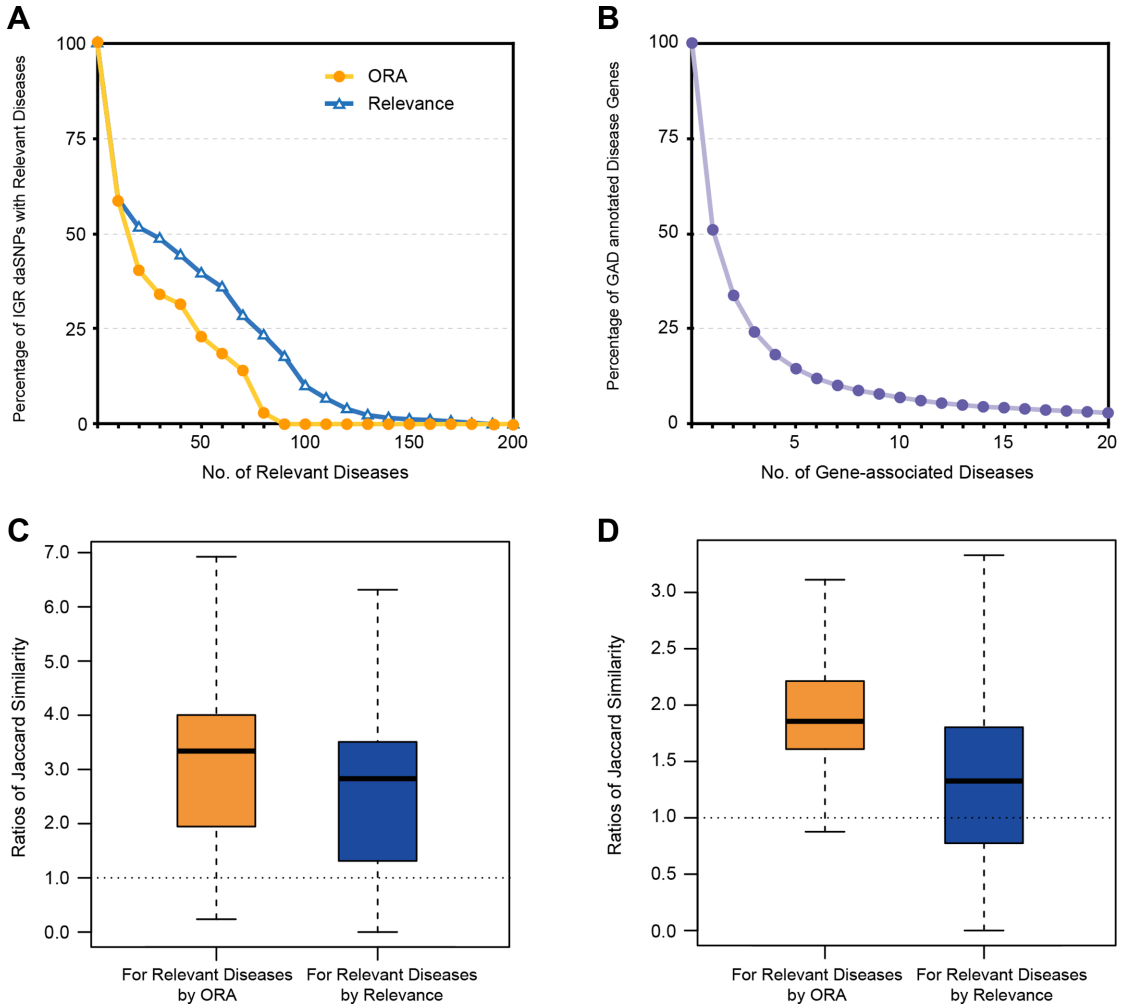

Figure 4. (**A**) The distributions of the numbers of IGR daSNPs in terms of the number of relevant diseases (e.g. > 10 relevant diseases) determined by the ORA or the relevance approaches. (**B**) The distribution of the numbers of disease genes in terms of the number of associated diseases. (**C**) Boxplots for the relative similarity among the relevant diseases found by the ORA or the relevance analysis for each IGR daSNP. The relative similarity was defined as the average Jaccard similarity among the relevant diseases divided by the average Jaccard similarity between all pairs of diseases. (**D**) Boxplots for the relative similarity between the relevant disease and the annotated diseases of an IGR daSNPs. Here, only those IGR daSNPs whose relevant diseases did not include the annotated diseases were considered, and the relative similarity was defined as the average Jaccard similarity between the relevant diseases and the annotated disease divided by the average Jaccard similarity between the annotated disease and all other diseases.

five or more diseases (Figure 4B). For this reason, the diseases found relevant to a given IGR daSNP by the ORA or the relevance analysis were highly related with each other: the average Jaccard similarity (computed using annotated disease genes) between the relevant diseases found by either the ORA or the relevance analysis were significantly higher than background Jaccard similarity (see Materials and Methods) (Figure 4C). There were 112 and 253 IGR daSNPs whose relevant diseases did not include the annotated diseases based on the ORA and the relevance analysis, respectively. For these IGR daSNPs, we found that for both ORA and the relevance analysis the average Jaccard similarity between the relevant diseases and the annotated diseases were significantly higher than background similarity (Figure 4D). The above evidence showed that the multiple relevant diseases to IGR daSNPs found by using the predicted target genes were not random, but were related with each other and with the annotated diseases. On the other hand, based on current GWAS annotations about 32% of IGR daSNPs already had more than one disease phenotype. Thus, the 'one-SNP-multiple-diseases' phenomena may be more common than expected.

While investigating the 'one-SNP-multiple-diseases' phenomena, we discovered a special 4 Mb region (29–33 Mb) in Chromosome 6. About 24% (439) of all IGR daSNPs were located in this region, and they tended to have significantly more numbers of disease phenotypes than the rest of the daSNPs did (T test, P-value = 7.6e-10). For example, 12 out of the 16 IGR daSNPs with more than five annotated disease phenotypes were located in this region; 19 and 8 out of the top 20 IGR daSNPs ranked by the numbers of relevant diseases according to the ORA and the relevance analysis, respectively, were located in this region. The diseases annotated or found to be relevant to IGR daSNPs in this region were often autoimmune diseases, such as rheumatoid arthritis, multiple sclerosis, lu-

pus erythematosus, etc. The predicted target genes for IGR daSNPs in this region also tended to be located in this region, and be enriched in immune-related functions. For example, 10 out of the top 20 most frequently predicted target genes for IGR daSNPs in this region were located in this region, including *HLA-B*, *HLA-DQA1* and *HLA-DRB1*. Interestingly, these *HLA* genes were annotated with more than 50 disease phenotypes. Further inspection of this region revealed that it was in fact a super hot spot of regulatory sequences and immune-related genes in the genome: it consisted of 5622 DHSs and 186 genes that were functionally enriched in immune-related processes. Thus, genetic variations in this region were very likely to contribute to multiple immune-related diseases. Here we presented one example—SNP rs9268832. This SNP was annotated to be associated with both lupus erythematosus and rheumatoid arthritis (26,48). The relevance analysis associated it with 109 diseases, including the two annotated diseases. This was because two of the predicted target genes—*HLA-DRB1* and *HLA-DQB1* were annotated not only to the two target diseases (49,50), but also to very high numbers of other diseases (90 and 83 for *HLA-DRB1* and *HLA-DQB1*, respectively). Note that rs9268832 was an eQTL to both genes (38,51–52).

The 'one-SNP-multiple-disease' phenomena also made it possible for us to predict the potential phenotypes of IGR daSNPs, which might be of use for prioritizing SNPs identified by GWAS studies. For instance, SNP rs9469220 was not considered to be associated with Crohn's disease because its *P*-value found by GWAS was 2e-6 (53), which was below the widely-used significance cutoff of 5e-8. The relevance analysis now provided evidence that it might be associated with the Crohn's disease. As such, the prediction of novel disease phenotypes for IGR daSNPs was of value for helping retrieve weak disease associations that would otherwise be discarded due to low statistical significance.

The contribution of different target gene prediction methods to explain the disease phenotypes of IGR daSNPs and the necessity of combining five methods

In this study, we used INTREPID to predict the target genes for IGR daSNPs. INTREPID combined the predictions from five component methods: HIC, PPC, IM-PET, PreSTIGE and ENCODE, by simply taking the union of their predictions. The reasons why we combined the predictions made by these five methods instead of using one best method were listed as follows. Firstly, the predictions made by each component method were useful for explaining the disease phenotypes of IGR daSNPs. For all five methods, the proportions of daSNP-predicted target gene pairs among the collected daSNP-gene eQTL/mQTL pairs were all significantly higher than random (all *P*-values $<$ 0.001, experiments were done similar to that on INTREPID predictions), partly validating the quality of their predictions. Secondly, although different component methods predicted different numbers of predictions: HIC predicted the largest number of target genes for IGR daSNPs, followed by PPC, ENCODE, PreSTIGE and IM-PET (Figure 5A), each made unique predictions (Figure 5B). Most of the predictions made by HIC and PPC were unique. This is especially true for PPC, because it could predict the target genes located far from the daSNPs (the median distance was 19 Mb) while the other methods usually predicted the target genes within relatively shorter distances to the daSNPs (the median distances were generally within 400 kb) (Supplementary Figure S1). Although about 40–70% of the predictions made by ENCODE, PreSTIGE or IM-PET could be confirmed by at least one other method, these three methods each also made a significant number of unique predictions. Thirdly, the combination of the predictions resulted in a larger number of explainable daSNP-disease associations than any individual method did alone. Here, we started from HIC's predicted target genes and then added the target genes predicted by PPC, ENCODE, PreSTIGE and IM-PET in succession. We observed that each addition increased the numbers of explainable IGR daSNP-disease associations based on either the occurrence, the ORA or the relevance analysis (Figure 5C). The same trends were also observed for the numbers of highly likely, mechanistically likely and potentially likely explainable SNP-disease associations (Figure 5D). For instance, the association of rs9273363 with narcolepsy (54) was not 'highly likely' explainable until the predicted target genes of ENCODE were added. In conclusion, each of the five methods has its own merit, and by combining their predictions we could better explain the daSNP-disease asscoiations. This also suggested that the addition of more predicted target genes in high-quality might help explain more IGR daSNP-disease associations.

DISCUSSION

In this study, we presented a novel approach to explain the disease phenotypes of IGR daSNPs. Given an IGR daSNP, we first identified its HRE whose function might be affected by the daSNP. Then, we applied a combined method named INTREPID to predict the target genes regulated by the HRE. Evidence from eQTL/mQTL data showed that the expression of the predicted target genes was likely affected by the corresponding IGR daSNPs. Next, we applied three levels of functional analysis to investigate whether the predicted target genes were functionally related to known disease genes. Results showed that about 36.8% of IGR daSNP-disease associations were 'highly-likely', 'mechanistically likely' or 'potentially likely' explainable by the predicted target genes. When LD SNPs were considered, this proportion increased to about 60.5%. These lines of evidence thus strongly supported the following scenario for the disease-association mechanism of IGR daSNPs: an IGR daSNP may disrupt the normal regulatory role of its HRE, causing abnormal expression changes to disease genes or genes functionally relevant to disease genes, and consequently leading to disease phenotypes.

For those IGR daSNPs that could not be explained using the predicted target genes, there were several reasons. Firstly, the target genes predicted for these IGR daSNPs by INTREPID might be incomplete. For example, distal regulation may be dynamic and cell-type specific, while the Hi-C and ChIA-PET data used by HIC and the epigenetic data used by ENCODE (23,24), PreSTIGE (21) and IM-PET (22) were generated from a limited number of cell lines.

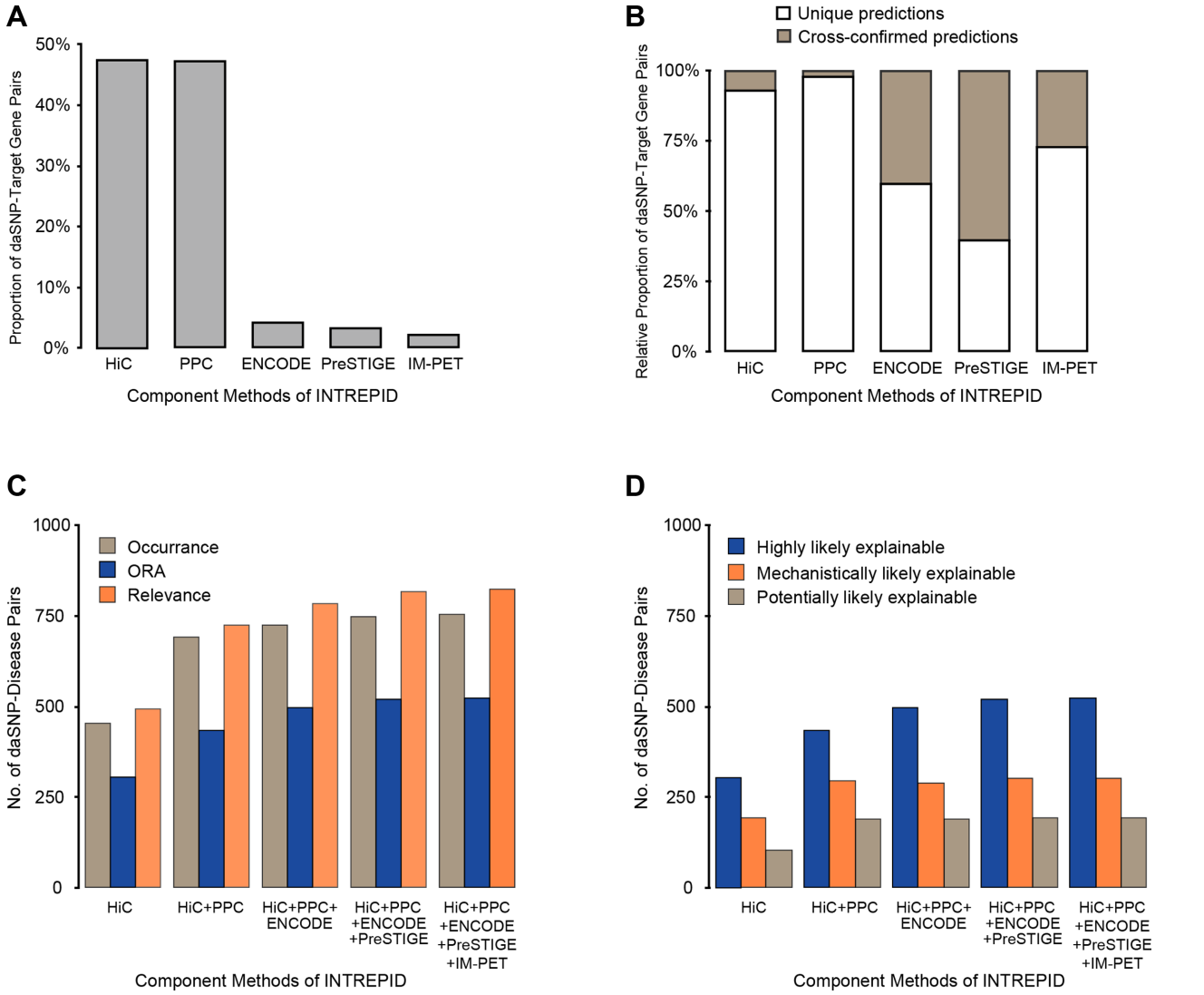

Figure 5. (**A**) The proportions of the predicted IGR daSNP-target gene pairs by each of the five component methods among the predicted pairs by INTREPID. (**B**) The relative proportions of IGR daSNP-target gene pairs that were predicted by only the investigated method (unique predictions) or by the investigated method and also the other methods (cross-confirmed predictions) among the predicted pairs by the investigated method. (**C**) The numbers of IGR daSNP-disease associations identified by the occurrence, the ORA, or the relevance analysis using the predicted target genes from only HIC, or by successively adding the predicted target genes from each of PPC, ENCODE, PreSTIGE and IM-PET. (**D**) Similar to (C) except that the numbers of different categories of 'explainable' IGR daSNP-disease associations were reported.

On the other hand, in this study the predicted target genes were limited to genes located in the same chromosome as the daSNPs, while inter-chromosome distal regulation may be more common than expected. Secondly, we defined the nearest DRE to an IGR daSNP as its HRE, and assumed that the DRE may be functionally affected by the daSNP, which may not always be a real case. Thirdly, the annotation of disease genes was not complete. Finally, there may be disease-association mechanisms for IGR daSNPs other than the one proposed in this study.

When using the predicted target genes to explain the disease phenotypes of IGR daSNP, we implicitly suggested that there might be a causative chain of relationships: the IGR daSNP affected the normal function of its HRE, which then altered the expression of the target genes regulated by the HRE, and eventually contributed to the disease phenotype. However, in reality we did not have experimental data to support the above causal relationships, and the so-called 'explainable' should be interpreted as the proof of association between an IGR daSNP and its corresponding disease phenotype. A step toward inferring whether an IGR daSNP might be disease-contributing was to put the above relationships into cell line or tissue-specific context. In this study, we used small RNA data and annotated enhancer information to first determine whether the HRE of an IGR daSNP was functional in the corresponding disease cell lines or tissues, and then used eQTL/mQTL data to verify the daSNP-target gene relationships. Based on these lines of evidence, we then determined a number of candidate disease-contributing IGR daSNPs for three types of complex diseases—immunological disease, cancer and neurological diseases. Though this is still far from demonstrating

that an IGR daSNP is disease-contributing, the list of candidate disease-contributing IGR daSNPs provided in this study was of value for designing specific experiments to investigate the disease-contributing mechanism of these daSNPs.

Besides explaining the disease phenotypes of IGR daSNPs, we also used the predicted target genes to investigate the association of IGR daSNPs with other diseases, and discovered the 'one-SNP-multiple-diseases' phenomena. The phenomena were mainly caused by the frequent occurrence of 'one-gene-multiple-disease' associations and the high relatedness between complex diseases. However, the 'one-SNP-multiple-diseases' phenomena only represented a possibility, and did not necessarily indicate that the multiple diseases would occur at the same time for one genotype of daSNP. In reality, an IGR daSNP may only affect one or a few of the predicted target genes under a specific biological circumstance, during which the daSNP would only be associated with one or a subset of diseases found by using all predicted target genes. To reduce the number of associated diseases for an IGR daSNP and to identify circumstance-specific target genes affected by an IGR daSNP, more information will be needed. Nevertheless, the predicted target genes by INTREPID have provided a useful pool of candidate target genes and disease phenotypes for future exploration.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

The authors thank Wei Ning for his help with plotting the figures.

FUNDING

National Natural Science Foundation of China [31471245, 91231116, 31071113, 30971643]; National Basic Research Program of China [2012CB316505]; Specialized Research Fund for the Doctoral Program of Higher Education of China [20120071110018]; Innovation Program of Shanghai Municipal Education Commission [13ZZ006]; Shanghai Shuguang Program [13SG05]. Funding for open access charge: National Natural Science Foundation of China [31471245, 91231116, 31071113, 30971643]; National Basic Research Program of China [2012CB316505]; Specialized Research Fund for the Doctoral Program of Higher Education of China [20120071110018]; Innovation Program of Shanghai Municipal Education Commission [13ZZ006]; Shanghai Shuguang Program [13SG05].
Conflict of interest statement. None declared.

REFERENCES

1. Welter,D., MacArthur,J., Morales,J., Burdett,T., Hall,P., Junkins,H., Klemm,A., Flicek,P., Manolio,T. and Hindorff,L. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001–D1006.
2. Stranger,B.E., Stahl,E.A. and Raj,T. (2011) Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, **187**, 367–383.
3. Maurano,M.T., Humbert,R., Rynes,E., Thurman,R.E., Haugen,E., Wang,H., Reynolds,A.P., Sandstrom,R., Qu,H., Brody,J. *et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science*, **337**, 1190–1195.
4. Trynka,G., Sandor,C., Han,B., Xu,H., Stranger,B.E., Liu,X.S. and Raychaudhuri,S. (2013) Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.*, **45**, 124–130.
5. Claussnitzer,M., Dankel,S.N., Klocke,B., Grallert,H., Glunk,V., Berulava,T., Lee,H., Oskolkov,N., Fadista,J., Ehlers,K. *et al.* (2014) Leveraging cross-species transcription factor binding site patterns: from diabetes risk Loci to disease mechanisms. *Cell*, **156**, 343–358.
6. Khurana,E., Fu,Y., Colonna,V., Mu,X.J., Kang,H.M., Lappalainen,T., Sboner,A., Lochovsky,L., Chen,J., Harmanci,A. *et al.* (2013) Integrative annotation of variants from 1092 humans: Application to cancer genomics. *Science*, **342**, 1235587.
7. Huang,D. and Ovcharenko,I. (2015) Identifying causal regulatory SNPs in ChIP-seq enhancers. *Nucleic Acids Res.*, **43**, 225–236.
8. Tokuhiro,S., Yamada,R., Chang,X., Suzuki,A., Kochi,Y., Sawada,T., Suzuki,M., Nagasaki,M., Ohtsuki,M., Ono,M. *et al.* (2003) An intronic SNP in a RUNX1 binding site of SLC22A4, encoding an organic cation transporter, is associated with rheumatoid arthritis. *Nat. Genet.*, **35**, 341–348.
9. Pomerantz,M.M., Ahmadiyeh,N., Jia,L., Herman,P., Verzi,M.P., Doddapaneni,H., Beckwith,C.A., Chan,J.A., Hills,A., Davis,M. *et al.* (2009) The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat. Genet.*, **41**, 882–884.
10. Sotelo,J., Esposito,D., Duhagon,M.A., Banfield,K., Mehalko,J., Liao,H., Stephens,R.M., Harris,T.J., Munroe,D.J. and Wu,X. (2010) Long-range enhancers on 8q24 regulate c-Myc. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 3001–3005.
11. Li,W., Notani,D. and Rosenfeld,M.G. (2016) Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nat. Rev. Genet.*, **17**, 207–223.
12. Glinskii,A.B., Ma,J., Ma,S., Grant,D., Lim,C.U., Sell,S. and Glinsky,G.V. (2009) Identification of intergenic trans-regulatory RNAs containing a disease-linked SNP sequence and targeting cell cycle progression/differentiation pathways in multiple common human disorders. *Cell Cycle*, **8**, 3925–3942.
13. Glinskii,A.B., Ma,S., Ma,J., Grant,D., Lim,C.U., Guest,I., Sell,S., Buttyan,R. and Glinsky,G.V. (2011) Networks of intergenic long-range enhancers and snpRNAs drive castration-resistant phenotype of prostate cancer and contribute to pathogenesis of multiple common human disorders. *Cell Cycle*, **10**, 3571–3597.
14. Vokes,S.A., Ji,H., Wong,W.H. and McMahon,A.P. (2008) A genome-scale analysis of the cis-regulatory circuitry underlying sonic hedgehog-mediated patterning of the mammalian limb. *Genes Dev.*, **22**, 2651–2663.
15. Ferretti,E., Cambronero,F., Tumpel,S., Longobardi,E., Wiedemann,L.M., Blasi,F. and Krumlauf,R. (2005) Hoxb1 enhancer and control of rhombomere 4 expression: complex interplay between PREP1-PBX1-HOXB1 binding sites. *Mol. Cell. Biol.*, **25**, 8541–8552.
16. Mossing,M.C. and Record,M.T. Jr (1986) Upstream operators enhance repression of the lac promoter. *Science*, **233**, 889–892.
17. Dekker,J., Rippe,K., Dekker,M. and Kleckner,N. (2002) Capturing chromosome conformation. *Science*, **295**, 1306–1311.
18. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
19. Fullwood,M.J., Liu,M.H., Pan,Y.F., Liu,J., Xu,H., Mohamed,Y.B., Orlov,Y.L., Velkov,S., Ho,A., Mei,P.H. *et al.* (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, **462**, 58–64.
20. Lu,Y., Zhou,Y. and Tian,W. (2013) Combining Hi-C data with phylogenetic correlation to predict the target genes of distal regulatory elements in human genome. *Nucleic Acids Res.*, **41**, 10391–10402.
21. Corradin,O., Saiakhova,A., Akhtar-Zaidi,B., Myeroff,L., Willis,J., Cowper-Sal,R., Lupien,M., Markowitz,S. and Scacheri,P.C. (2014) Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res.*, **24**, 1–13.

22. He,B., Chen,C., Teng,L. and Tan,K. (2014) Global view of enhancer-promoter interactome in human cells. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, E2191–E2199.
23. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.
24. Thurman,R.E., Rynes,E., Humbert,R., Vierstra,J., Maurano,M.T., Haugen,E., Sheffield,N.C., Stergachis,A.B., Wang,H., Vernot,B. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.
25. Bhatia,S., Bengani,H., Fish,M., Brown,A., Divizia,M.T., de Marco,R., Damante,G., Grainger,R., van Heyningen,V. and Kleinjan,D.A. (2013) Disruption of autoregulatory feedback by a mutation in a remote, ultraconserved PAX6 enhancer causes aniridia. *Am. J. Hum. Genet.*, **93**, 1126–1134.
26. Li,M.J., Wang,P., Liu,X., Lim,E.L., Wang,Z., Yeager,M., Wong,M.P., Sham,P.C., Chanock,S.J. and Wang,J. (2011) GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.*, **40**, D1047-D1054.
27. Yu,W., Gwinn,M., Clyne,M., Yesupriya,A. and Khoury,M.J. (2008) A navigator for human genome epidemiology. *Nat. Genet.*, **40**, 124–125.
28. Johnson,A.D. and O'Donnell,C.J. (2009) An open access database of genome-wide association results. *BMC Med. Genet.*, **10**, 1–17.
29. Becker,K.G., Barnes,K.C., Bright,T.J. and Wang,S.A. (2004) The genetic association database. *Nat. Genet.*, **36**, 431–432.
30. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
31. Consortium,E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
32. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
33. Hnisz,D., Abraham,B.J., Lee,T.I., Lau,A., Saint-Andre,V., Sigova,A.A., Hoke,H.A. and Young,R.A. (2013) Super-enhancers in the control of cell identity and disease. *Cell*, **155**, 934–947.
34. Franceschini,A., Szklarczyk,D., Frankild,S., Kuhn,M., Simonovic,M., Roth,A., Lin,J., Minguez,P., Bork,P., von Mering,C. *et al.* (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**, D808–D815.
35. Dijkstra,E.W. (1959) A note on two problems in connexion with graphs. *Numer. Math.*, **1**, 269–271.
36. Gibbs,R.A., Belmont,J.W., Hardenbol,P., Willis,T.D., Yu,F., Yang,H., Ch'ang,L.-Y., Huang,W., Liu,B. and Shen,Y. (2003) The international HapMap project. *Nature*, **426**, 789–796.
37. Rockman,M.V. and Kruglyak,L. (2006) Genetics of global gene expression. *Nat. Rev. Genet.*, **7**, 862–872.
38. Consortium,G.T. (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.
39. International Multiple Sclerosis Genetics, C., Hafler,D.A., Compston,A., Sawcer,S., Lander,E.S., Daly,M.J., De Jager,P.L., de Bakker,P.I., Gabriel,S.B., Mirel,D.B. *et al.* (2007) Risk alleles for multiple sclerosis identified by a genomewide study. *N. Engl. J. Med.*, **357**, 851–862.
40. Peters,U., Hutter,C.M., Hsu,L., Schumacher,F.R., Conti,D.V., Carlson,C.S., Edlund,C.K., Haile,R.W., Gallinger,S., Zanke,B.W. *et al.* (2012) Meta-analysis of new genome-wide association studies of colorectal cancer risk. *Hum. Genet.*, **131**, 217–234.
41. Schadt,E.E., Molony,C., Chudin,E., Hao,K., Yang,X., Lum,P.Y., Kasarskis,A., Zhang,B., Wang,S., Suver,C. *et al.* (2008) Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.*, **6**, e107.
42. Harismendy,O., Notani,D., Song,X., Rahim,N.G., Tanasa,B., Heintzman,N., Ren,B., Fu,X.D., Topol,E.J., Rosenfeld,M.G. *et al.* (2011) 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. *Nature*, **470**, 264–268.
43. Anderson,C.A., Boucher,G., Lees,C.W., Franke,A., D'Amato,M., Taylor,K.D., Lee,J.C., Goyette,P., Imielinski,M., Latiano,A. *et al.* (2011) Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.*, **43**, 246–252.
44. Torgerson,D.G., Ampleford,E.J., Chiu,G.Y., Gauderman,W.J., Gignoux,C.R., Graves,P.E., Himes,B.E., Levin,A.M., Mathias,R.A., Hancock,D.B. *et al.* (2011) Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. *Nat. Genet.*, **43**, 887–892.
45. Wasserman,N.F., Aneas,I. and Nobrega,M.A. (2010) An 8q24 gene desert variant associated with prostate cancer risk confers differential in vivo activity to a MYC enhancer. *Genome Res.*, **20**, 1191–1197.
46. Replication, D.I.G., Meta-analysis, C., Asian Genetic Epidemiology Network Type 2 Diabetes, C., South Asian Type 2 Diabetes, C., Mexican American Type 2 Diabetes, C., Type 2 Diabetes Genetic Exploration by Nex-generation sequencing in muylti-Ethnic Samples, C., Mahajan,A., Go,M.J., Zhang,W., Below,J.E. *et al.* (2014) Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.*, **46**, 234–244.
47. Hancock,D.B., Eijgelsheim,M., Wilk,J.B., Gharib,S.A., Loehr,L.R., Marciante,K.D., Franceschini,N., van Durme,Y.M., Chen,T.H., Barr,R.G. *et al.* (2010) Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nat. Genet.*, **42**, 45–52.
48. Plenge,R.M., Seielstad,M., Padyukov,L., Lee,A.T., Remmers,E.F., Ding,B., Liew,A., Khalili,H., Chandrasekaran,A., Davies,L.R. *et al.* (2007) TRAF1-C5 as a risk locus for rheumatoid arthritis–a genomewide study. *N. Engl. J. Med.*, **357**, 1199–1209.
49. Brennan,P., Hajeer,A., Ong,K.R., Worthington,J., John,S., Thomson,W., Silman,A. and Ollier,B. (1997) Allelic markers close to prolactin are associated with HLA-DRB1 susceptibility alleles among women with rheumatoid arthritis and systemic lupus erythematosus. *Arthritis Rheum*, **40**, 1383–1386.
50. Hrycek,A., Siekiera,U., Cieślik,P. and Szkróbka,W. (2005) HLA-DRB1 and-DQB1 alleles and gene polymorphisms of selected cytokines in systemic lupus erythematosus. *Rheumatol. Int.*, **26**, 1–6.
51. Sasayama,D., Hori,H., Nakamura,S., Miyata,R., Teraishi,T., Hattori,K., Ota,M., Yamamoto,N., Higuchi,T. and Amano,N. (2013) Identification of single nucleotide polymorphisms regulating peripheral blood mRNA expression with genome-wide significance: An eQTL study in the Japanese population. *PloS One*, **8**, e54967.
52. Kabakchiev,B. and Silverberg,M.S. (2013) Expression quantitative trait loci analysis identifies associations between genotype and gene expression in human intestine. *Gastroenterology*, **144**, U1488–U1284.
53. Wellcome Trust Case Control, C. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.
54. Koike,A., Nishida,N., Inoue,I., Tsuji,S. and Tokunaga,K. (2009) Genome-wide association database developed in the Japanese Integrated Database Project. *J. Hum. Genet.*, **54**, 543–546.