Human Molecular Genetics, 2010, Vol. 19, No. 22 **4345–4352**
doi:10.1093/hmg/ddq356
Advance Access published on August 24, 2010

Identification of epistatic effects using a protein–protein interaction database

Yan V. Sun* **and Sharon L.R. Kardia**

Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, USA

Received June 2, 2010; Revised August 4, 2010; Accepted August 17, 2010

Epistasis (i.e. gene–gene interaction) has long been recognized as an important mechanism underlying the complexity of the genetic architecture of human traits. Definitions of epistasis range from the purely molecular to the traditional statistical measures of interaction. The statistical detection of epistasis usually does not map onto or easily relate to the biological interactions between genetic variations through their combined influence on gene expression or through their interactions at the gene product (i.e. protein) or DNA level. Recently, greater high-dimensional data on protein–protein interaction (PPI) and gene expression profiles have been collected that enumerates sets of biological interactions. To better align statistical and molecular models of epistasis, we present an example of how to incorporate the PPI information into the statistical analysis of interactions between copy number variations (CNVs). Among the 23 640 pairs of known human PPIs and the 1141 common CNVs detected among HapMap samples, we identified 37 pairs of CNVs overlapping with both genes of a PPI pair. Two CNV pairs provided sufficient genotype variation to search for epistatic effects on gene expression. Using 47 294 probe-specific gene expression levels as the outcomes, five epistatic effects were identified with *P*-value less than 10^{-6}. We found a CNV–CNV interaction significantly associated with gene expression of *TP53TG3* (*P*-value of 2×10^{-20}). The proteins associated with the CNV pair also bind TP53 which regulates the transcription of *TP53TG3*. This study demonstrates that using PPI data can assist in targeting statistical hypothesis testing to biological plausible epistatic interaction that reflects molecular mechanisms.

INTRODUCTION

The term 'epistasis' was first used by Bateson (1) to describe the blocking of the phenotypic effect of one genetic mutation by another genetic locus. Fisher created the term 'epistacy' to define the statistical deviation of the linear combination of multiple genetic effects for a trait (2). Epistasis has long been used as a general term to describe the complex interactions among genetic loci (3). For example, 'functional epistasis' (the molecular interactions among proteins and other functional elements) and 'statistical epistasis' (the deviation from the additive contributions of multiple genetic loci within a population) have distinct meanings but have been referred to by the same term of 'epistasis' (3). Although we would expect a phenotypic effect to arise from the functional interruption of the relationship between two genes, the translation of the statistical interactions to the functional interactions is not straightforward and has been extensively debated (4). The common genetic variants identified by the genome-wide association studies (GWAS) using additive genetic models without considering dominance or potential epistatic associations only explain a small fraction of the heritability (5). Although it has been suggested that the epistatic associations could be responsible for a portion of the unexplained heritability of in GWAS (5,6), the challenge remains to identify how to conduct genome-wide epistatic studies with prior biological knowledge (7) to connect the statistical evidence to a testable molecular model. Developing databases that curate and annotate information about all the molecular interactions within a cell (e.g. DNA–DNA, DNA–RNA, DNA–protein, RNA–RNA, RNA–protein and protein–protein) would greatly assist in providing an *a priori* biological hypothesis space for testing for epistasis (8,9). It would also greatly facilitate faster validation of the putative molecular epistatic mechanisms. Given the current state of this type of molecular interaction knowledge, we have chosen to focus on

*To whom correspondence should be addressed at: Department of Epidemiology, School of Public Health, University of Michigan, 1415 Washington Heights No. 4605, Ann Arbor, MI 48109, USA. Tel: +1 7346156279; Fax: +1 7347641357; Email: yansun@umich.edu

demonstrating the use of a protein–protein interaction (PPI) database on the identification of epistasis.

The physical interactions between proteins are important for a wide range of biological functions—e.g. biochemical reactions, cytoskeletal structures, cellular transport systems, transcriptional activation and regulation. With the advancement of molecular technologies, there are now several biochemical and biophysical high-throughput methods (e.g. two-hybrid system and affinity purification-mass spectrometry) that are being used to identify the physical interactions among proteins. Combining the high-throughput affinity purification, mass spectrometry and computational algorithms (10,11), or using protein-fragment complementation arrays (12), the global maps of PPI have been constructed in the budding yeast (*Saccharomyces cerevisiae*). The genetic interaction network of yeast has also been constructed by a synthetic genetic array method which screens the synthetic lethal genetic interactions through high-density arrays of double mutants (13,14). The overlap between the genetic interactions and the PPI in yeast ranges 10–20%, which is significantly higher than the random expectation (14). The yeast two-hybrid system was used to screen over 4000 human proteins and identified over 3000 mostly novel PPI interactions (15). Other high-throughput studies of human interactome also identified a large number of PPI pairs (16,17). PPI databases, such as BioGRID (18) and MIPS (19), have been developed to document all known PPI information from multiple organisms including human. As a result, we are able to query thousands of identified interactions among human proteins using these databases, to define the hypothesis testing space for a new type of evaluation of gene–gene interaction that integrate aspects of statistical and functional epistasis.

While the set of measured genetic variations (and thus gene–gene interaction) continues to expand with advances in sequencing, in this paper we focus on copy number variations (CNVs) which are a less common type of genetic variant in human compared with single nucleotide variations. A CNV is a segment of DNA that is present at a variable number of copies compared with a reference genome sequence. CNVs have been demonstrated to be associated with quantitative gene expression levels (20,21) that in some cases are likely to have causative, functional effects (21). CNVs have also been reported to be associated with human disorders such as Parkinson's disease (22), schizophrenia (23–25), autism (26,27) and Crohn's disease (28). When a CNV overlaps with a gene coding region, it may influence the function of the gene product by shifting the amino acid structure of the encoded protein.

Integrating the rich information available from PPI databases and the high-throughput measurements of genetic variants, we have a unique opportunity to relate the functional interaction between gene products (i.e. proteins), and the statistical interactions (epistasis) associated with human traits. In this study, we established an analytical framework to demonstrate that the functional espistasis and statistical epistasis can be integrated to identify interacting genetic loci. We combined the human PPI and CNV data into a statistical model and identified epistatic associations with gene expression levels (the intermediate traits influencing protein production) in transformed lymphocytes of human.

RESULTS

Following our integrative approach to identifying epistatic associations described in Figure 1, we first filtered and merged the PPI and the CNV databases. Thirty-seven pairs of CNVs were identified as representing potential functional gene–gene interaction that could be manifested at the protein levels. Of these 37 pairs, 20 represented homodimeric pairs (i.e. a protein interacting with itself and functions as a homodimer) that were excluded from this study because they represent potential intragenic epistasis rather than intergenic epistasis that would represent the impact of PPI on gene expression. The annotations of the remaining 17 pairs of CNVs that overlap with genes that form heterodimeric protein pairs were listed in Table 1. Limited by HapMap sample size of 210 and the low minor allele frequency (MAF) of some CNVs, we finally selected two CNV pairs (CNP109–CNP10282 and CNP865–CNP10140) that have a non-zero count in at least four out of nine genotype combinations to conduct the following epistatic association analysis.

We tested the association of the epistatic interaction between the two CNV pairs and 47 293 gene expression levels using the model described in the Materials and Methods section. The epistatic associations with the *P*-value lower than 10^{-4} were listed in Table 2. Using Bonferroni corrected threshold of *P*-value ($0.05/47\,293 = 1.06 \times 10^{-6}$), we identified five statistically significant epistatic associations after correcting for multiple testing (Table 2). The most significant epistatic association (*P*-value of 1.98×10^{-20}) is between CNV pair, CNP865 (*PCDHA4*)–CNP10140 (*SETDB1*), and the gene expression level of *TP53TG3* (probe GI_7706742-A). This epistatic association was illustrated in Figure 2. The other four epistatic associations that passed the genome-wide significance threshold are between CNP865–CNP10140 and the gene expression level of *MGC20553* (probe GI_34222248-S), CNP865–CNP10140 and the gene expression level of *Hs.522383* (probe Hs.522383-S), CNP865–CNP10140 and the gene expression level of *PPIA* (probe GI_45439310-I) and CNP109 (*ARNT*)-CNP10282 (*EPAS1*) and the gene expression level of *hmm31138* (probe hmm31138-S).

In order to better understand the effects of the most significant epistatic association (*P*-value of 1.98×10^{-20}) between CNV pair, CNP865 (*PCDHA4*)–CNP10140 (*SETDB1*), and the gene expression level of *TP53TG3* (probe GI_7706742-A), we also examined the main effect of the CNVs on their gene's expression level. Specifically, we examined the associations between CNP865 and the gene expression level of its overlapping gene *PCDHA4*, as well as the association between CNP10140 and the gene expression level of its overlapping gene *SETDB1*. Furthermore, we evaluated whether the two CNV had any statistical evidence of a main effect, knowing completely that it could be a single-dimensional reflection of the gene–gene epistatic effect on *TP53TG3* levels. The statistics of these association tests were summarized in Table 3. After adjusting for the population structure by the first 10 principal components (PCs), we identified that CNP10140 (*SETDB1*) was significantly associated with *TP53TG3* gene expression with a *P*-value of 0.0065. The other three associations were not statistically significant at an alpha level of 0.05.

Table 1. Seventeen pairs of CNVs overlapping with genes involved in interacting protein pairs

CNV1 CNV ID	Chromosome	CNV start (bp)	CNV end (bp)	Gene	CNV2 CNV ID	Chromosome	CNV start (bp)	CNV end (bp)	Gene
CNP10045	chr1	22 193 098	22 209 830	*HSPG2*	CNP10479	chr3	13 682 416	13 684 365	*FBLN2*
CNP109	chr1	150 822 330	150 853 218	*ARNT*	CNP10282	chr2	46 549 602	46 551 188	*EPAS1*
CNP11463	chr8	54 122 300	54 154 739	*OPRK1*	CNP12537	chr17	72 728 774	72 743 351	*SLC9A3R1*
CNP11164	chr6	162 658 558	162 660 430	*PARK2*	CNP12039	chr12	125 388 402	125 395 161	*UBC*
CNP11164	chr6	162 658 558	162 660 430	*PARK2*	CNP10878	chr5	886 628	928 018	*TRIP13*
CNP10375	chr2	128 538 342	128 543 406	*WDR33*	CNP12349	chr15	84 142 284	84 151 045	*SH3GL3*
CNP865	chr5	140 204 020	140 223 940	*PCDHA4*	CNP10140	chr1	150 915 736	150 923 166	*SETDB1*
CNP10145	chr1	151 940 298	151 959 872	*S100A10*	CNP10140	chr1	150 915 736	150 923 166	*SETDB1*
CNP11948	chr12	6 407 756	6 453 667	*SCNN1A*	CNP12589	chr18	55 803 727	55 815 999	*NEDD4L*
CNP11538	chr9	2 138 837	2 140 663	*SMARCA2*	CNP2563	chr22	23 993 985	24 248 712	*SMARCB1*
CNP12014	chr12	79 349 306	79 352 242	*SYT1*	CNP10981	chr5	141 998 202	142 000 197	*FGF1*
CNP12303	chr15	29 798 781	30 027 160	*TJP1*	CNP104	chr1	147 303 148	147 526 040	*GJA8*
CNP12303	chr15	29 798 781	30 027 160	*TJP1*	CNP11167	chr6	168 078 125	168 338 764	*MLLT4*
CNP11931	chr11	123 592 514	123 601 125	*ZNF202*	CNP12670	chr19	50 542 545	50 583 764	*ZNF473*
CNP12010	chr12	69 785 308	69 796 518	*YEATS4*	CNP2563	chr22	23 993 985	24 248 712	*SMARCB1*
CNP10878	chr5	886 628	928 018	*TRIP13*	CNP11670	chr10	5 705 468	5 757 133	*ASB13*
CNP11633	chr9	80 628 360	80 658 853	*GNAQ*	CNP12537	chr17	72 728 774	72 743 351	*SLC9A3R1*

DISCUSSION

The most statistically significant epistatic interaction between CNV pairs is associated with the TP53 target 3 (*TP53TG3*) gene expression level. *TP53TG3* is located on the short arm of chromosome 16. The gene expresses several transcripts by alternative splicing. The two major transcripts, *TP53TG3A* and *TP53TG3B*, encode 124- and 132-amino acid peptides that are expressed predominantly in testis. The *TP53TG3* gene is induced in a TP53-dependent manner (29). CNP865 and CNP10140 overlap with *PCDH4* and *SETDB1* correspondingly. The protein product of *PCDH4* not only binds SETDB1 protein but also binds TP53 protein (15). In this example, we started with the evidence of physical interaction from a PPI database (SETDB1–PCHDA4). Then we identified the statistically significant epistatic association of CNP10140 (*SETDB1*) and CNP865 (*PCDHA4*) with *TP53TG3* gene expression. Combining additional experimental data, we proposed a plausible molecular mechanism underlying the epistatic association illustrated in Figure 3. The pair of proteins, SETDB1–PCDHA4, affects *TP53TG3* gene expression through PCDHA4 binding with TP53, which is known to induce the gene expression of *TP53TG3*. This molecular model may be experimentally validated in human cell lines by mutating *SETDB1* and *PCDHA4* to mimic the CNV.

Limited by the sample size, we could only test the epistatic associations of two CNV pairs. Fifteen additional CNV pairs (Table 1) showed evidence of physical interaction between the protein products of their overlapping genes. Several of the genes containing these CNV pairs are associated with the pathogenesis of human diseases. For instance, the CNV pair CNP11948–CNP12589 (overlaps with *SCNN1A* and *NEDD4L* correspondingly) is a candidate pair for testing the epistatic association with hypertension and diseases associated with electrolyte transport. *SCNN1A* encodes the alpha subunit of an epithelial sodium channel (ENaC) which controls fluid and electrolyte transport across epithelia in many organs. Mutations in this gene have been associated with the autosomal recessive form of pseudohypoaldosteronism type 1, a salt wasting disease resulting from target organ unresponsiveness to mineralocorticoids. Genetic variants of *SCNN1A* had been reported as a genetic risk factor of hypertension (30). NEDD4L had strong affinity binding to SCNN1A and potentially regulate the activity of ENaC (31). Therefore, the genetic effect of *SCNN1A* may be modified by *NEDD4L* to affect human diseases such as hypertension or kidney disease. With larger samples from epidemiological studies, testing the epistatic associations of all these CNV pairs with disease traits may help us identify additional genetic mechanisms underlying human diseases.

Although we focused on the application of combining PPI data and CNV data in this study, our approach to identifying epistatic associations can be expanded to other types of molecular interaction [e.g. between transcriptional factors (TFs) and *cis*-regulatory elements (CREs)] as well as other types of genetic variants (e.g. SNPs, especially non-synonymous variations). Both experimental and computational approaches have been used to identify the interactions between the TFs and CREs of genes in human (32,33). The database derived from these studies will provide high-quality cell-type-specific interaction information between TFs and CREs to facilitate the similar approach to incorporating TF–CRE interactions in the epistatic analysis in human. Potential epistatic associations may also exist between CNVs and SNPs. In this study, we tested the epistatic associations with gene expression levels in transformed lymphocytes. Because different gene expression patterns exist in different cell types (34–36), the profile of epistatic associations is expected to be different in various cell lines/tissues even though the primary sequence of DNA remains the same across human cells. Therefore, selecting the appropriate tissue, organ or model system is critical for studying the epistatic associations of intermediate phenotypes such as gene expression levels. It should also be noted that many different kinds of phenotypes (e.g. proteomic or metabolomic profiles) could easily be examined using this methodology.

Animal and plant studies have shown more gene–gene (epistatic) interactions than previously expected (37). The

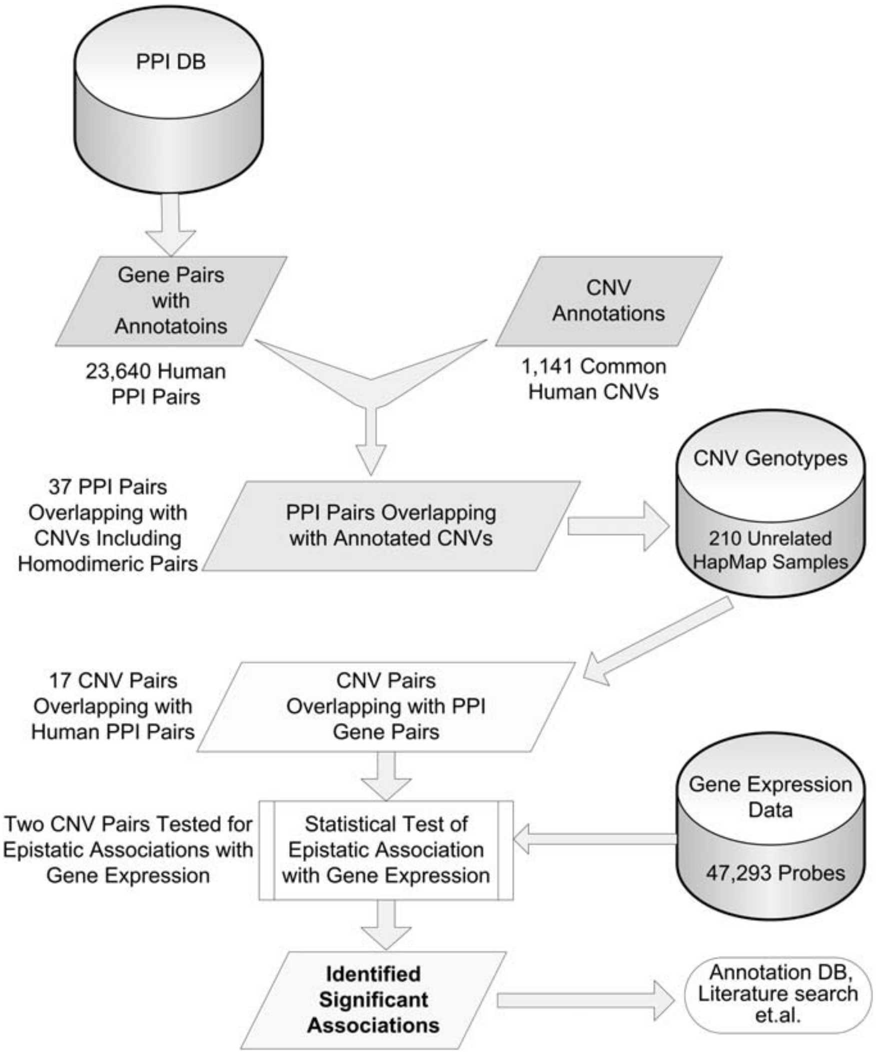
PPI DB
Gene Pairs with Annotatoins
23,640 Human PPI Pairs
CNV Annotations
1,141 Common Human CNVs
37 PPI Pairs Overlapping with CNVs Including Homodimeric Pairs
PPI Pairs Overlapping with Annotated CNVs
CNV Genotypes
210 Unrelated HapMap Samples
17 CNV Pairs Overlapping with Human PPI Pairs
CNV Pairs Overlapping with PPI Gene Pairs
Two CNV Pairs Tested for Epistatic Associations with Gene Expression
Statistical Test of Epistatic Association with Gene Expression
Gene Expression Data
47,293 Probes
Identified Significant Associations
Annotation DB, Literature search et.al.

Figure 1. The workflow of the integrated analysis of epistasis between CNV pairs with evidence of physical interaction.

evidence of epistatic effects has been documented in yeast (38), *A. thaliana* (39), fruit fly (40) and mice (41). Given the biological complexity of most disease phenotypes, it is not surprising that epistatic interactions play a major role and explain a larger proportion of variation in the phenotype than main genetic effects alone (42). Although computationally efficient methods of screening statistical epistatic interactions are available (43–45), conducting a complete genome-wide survey of epistatic interactions (2×10^{12} tests for 2 million genetic variants) requires a huge amount of resources and suffers from a serious multiple testing problem. Instead of targeting the genome-wide discovery of epistatic interactions, our approach focused on a limited set of gene pairs with strong evidence of physical interaction. The combined evidence of physical interactions and statistical interactions will greatly strengthen the confidence of the findings and leads to practical targets for validating biological functions.

In this study, we did not observe significant associations between the CNVs and the gene expression levels of their overlapping gene, namely between CNP865 and Illumina probe GI_14165412-A (*PCDHA4*), and between CNP10140 and Illumina probe GI_41281392-S (*SETDB1*). The 50 bp long GI_14165412-A specifically targets mRNA NM_031500 which is the shorter isoform of *PCDHA4*. NM_031500 spans from 140 186 672 to 140 189 167 bp on chromosome 5. CNP865 spans a 20 kb genomic region (140 204 020–140 223 940 bp) on chromosome 5. It is located about 15 kb downstream of NM_031500 and overlaps with the longer isoform of *PCDHA4* mRNA, NM_018907. It is likely that the CNV of CNP865 affects the function of PCDHA4 protein through NM_018907 expression rather than NM_031500. Therefore, no significant association is observed between CNP865 and NM_031500 expression. Illumina probe GI_41281392-S targets the last exon of gene *SETDB1* (NM_012432: 150 898 815–150 937 220 bp on chromosome 1), and is located over 13 kb downstream of CNP10140 (150 915 736–150 923 166 bp). The association between CNP10140 and *SETDB1* gene expression is marginal with a *P*-value of 0.06. It is unclear whether this CNP affects the expression level.

Table 2. The epistatic effects of CNV pairs with *P*-value lower than 10^{-4}

Outcome	Gene name	CNV1	CNV2	$\text{Beta}_{\text{CNP1*CNP2}}$	$\text{SE}_{\text{CNP1*CNP2}}$	*P* (CNP1*CNP2)
GI_7706742-A	***TP53TG3***	**CNP865 (*PCDHA4*)**	**CNP10140 (*PCTDB1*)**	**−1.994**	**0.192**	**1.98×10^{-20}**
GI_34222248-S	***MGC20553***	**CNP865 (*PCDHA4*)**	**CNP10140 (*PCTDB1*)**	**−0.997**	**0.177**	**5.99×10^{-08}**
Hs.522383-S	***Hs.522383***	**CNP865 (*PCDHA4*)**	**CNP10140 (*PCTDB1*)**	**−0.297**	**0.058**	**8.96×10^{-07}**
GI_45439310-I	***PPIA***	**CNP865 (*PCDHA4*)**	**CNP10140 (*PCTDB1*)**	**1.236**	**0.244**	**9.41×10^{-07}**
GI_6005877-S	*SLC22A1LS*	CNP865 (*PCDHA4*)	CNP10140 (*PCTDB1*)	−0.952	0.200	3.89×10^{-06}
GI_15011899-A	*SYTL2*	CNP865 (*PCDHA4*)	CNP10140 (*PCTDB1*)	−1.759	0.391	1.19×10^{-05}
GI_37675279-A	*JUB*	CNP865 (*PCDHA4*)	CNP10140 (*PCTDB1*)	−0.914	0.216	3.54×10^{-05}
GI_34147589-S	*LOC56901*	CNP865 (*PCDHA4*)	CNP10140 (*PCTDB1*)	−1.102	0.261	3.70×10^{-05}
Hs_462524-S	*Hs.462524*	CNP865 (*PCDHA4*)	CNP10140 (*PCTDB1*)	−0.206	0.050	5.27×10^{-05}
Hmm10297-S	*hmm10297*	CNP865 (*PCDHA4*)	CNP10140 (*PCTDB1*)	−0.290	0.073	9.36×10^{-05}
hmm31138-S	***hmm31138***	**CNP109 (*ARNT*)**	**CNP10282 (*EPAS1*)**	**−0.097**	**0.014**	**2.06×10^{-10}**
Hs.143656-S	*Hs.143656*	CNP109 (*ARNT*)	CNP10282 (*EPAS1*)	−0.054	0.013	3.44×10^{-05}
Hs.459819-S	*Hs.459819*	CNP109 (*ARNT*)	CNP10282 (*EPAS1*)	−0.057	0.014	4.15×10^{-05}
GI_42657279-S	*LOC401208*	CNP109 (*ARNT*)	CNP10282 (*EPAS1*)	−0.108	0.027	9.10×10^{-05}

Epistatic interactions with significant *P*-value after adjusted for multiple testing are highlighted in bold.

Table 3. CNV main effects of gene expression levels of *TP53TG3, PCDHA4* and *SETDB1*

Outcome		Predictor		
GE Probe	Gene	CNV	Gene	*P*-value
GI_7706742-A	*TP53TG3*	CNP865	*PCDHA4*	0.351
GI_7706742-A	*TP53TG3*	CNP10140	*SETDB1*	**0.0065**
GI_14165412-A	*PCDHA4*	CNP865	*PCDHA4*	0.302
GI_41281392-S	*SETDB1*	CNP10140	*SETDB1*	0.060

The first 10 PCs were used as covariates to adjust for the population structure of the HapMap population. Bold value indicates $P < 0.05$.

Although technologies have been greatly improved in measuring PPIs and CNVs in human, the current databases of human PPIs and CNVs are still relatively new and still being developed. Several databases host PPI data from human (e.g. BioGRID, MIPS etc.). Each database chooses its own standard to include PPI data from published results and has its own style of annotating and presenting the data. Therefore, each PPI database is likely to be limited both in size and through the experimental methods used to obtain the information about the protein interactions. Both of these are likely to improve with each new release. Comparing all PPI databases is a very ambitious task and is out of the scope of this study. We selected one PPI database that we are familiar with to demonstrate the utility of combining the physical interaction data with statistical epistasis test to identify novel molecular interactions. Admittedly, the bias and incompleteness of the current PPI database restricts us from knowing or testing the entire set of possible epistatic effects associated with PPIs. We expect to identify more testable molecular mechanisms using our approach as the identification and availability of PPI evolves. A recent study of a large population identified 11 700 CNVs and revealed a more complete but complex picture of the human CNV map (46). Being aware of the limitations of the imperfect measurements and incomplete database, we tried to demonstrate that we could identify novel epistatic effects by combining the high-throughput experimental measurements and statistical/bioinformatic methods, rather than claiming an approach to screening genome-wide epistatic associations.

Testing epistatic associations usually requires a larger sample size; however, human CNVs overlapping with genic regions tend to be less common (46). Because of the limited size of the HapMap sample and the low allele frequency of the measured CNVs, we were only able to test the epistatic associations of two pairs of CNVs with putative interaction effect. For the same reason, we were not able to test the race-specific epistatic associations within each racial group. Despite these limitations, our approach illustrates the use of empirical data about the physical interaction among human proteins to formulate a statistical model of testing epistatic associations of complex traits such as gene expression levels. This approach to investigating potential epistasis creates a foundation for a whole range of future experimental and epidemiological studies focused on understanding both the molecular mechanisms and its impact of human disease risk.

MATERIALS AND METHODS

Sample and data

Two hundred and seventy subjects (210 unrelated) in three racial groups were available from the HapMap project for the current study. There were 60 unrelated subjects from Utah; these individuals represent the United States Caucasian population with Northern and Western European ancestry (parents in 30 trios). There were 60 unrelated subjects collected from the Yoruba people in Nigeria (parents in the 30 trios). Forty-five unrelated Han Chinese in Beijing, China and 45 unrelated Japanese in Tokyo, Japan were collected for the Asian group. All subjects gave specific consent for their inclusion in the HapMap project (47).

The mRNA gene expression data for the HapMap samples were obtained from Wellcome Trust Sanger Institute. The Illumina's Sentrix Human-6 Expression BeadChip (Illumina, San Diego, CA, USA) was used to measure 47 294 human transcripts. The quantification and normalization of the gene expression data were described in a previous report (20).

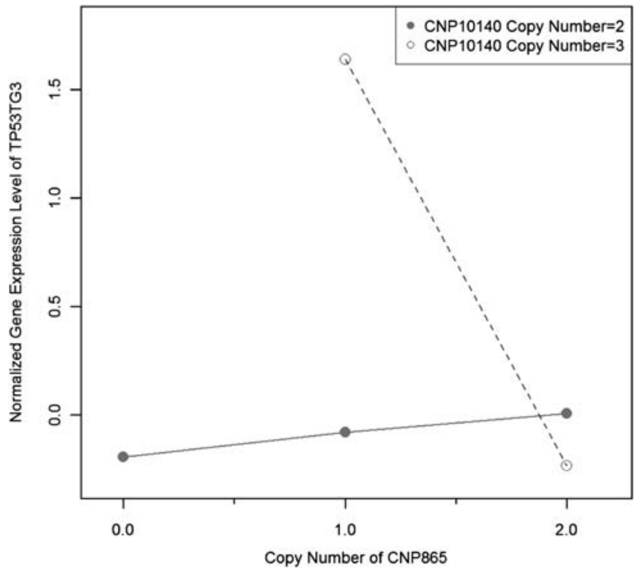

Figure 2. Genotype-specific means of the gene expression level of TP53TG3.

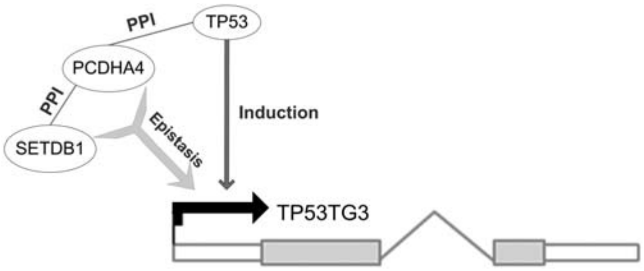

Figure 3. A plausible molecular mechanism of the epistatic effect of SETDB1–PCDHA4 associated with gene expression of TP53TG3.

The CNV genotype data for the 210 unrelated HapMap subjects were merged with their gene expression data using the anonymous, unique identifiers.

Copy number variation analysis

The genotyping data for the 270 HapMap subjects were produced and distributed by Affymetrix® using the Genome-Wide Human SNP Array 6.0 platform. Prior to CNV analysis, the Contract QC (CQC) value was calculated for each chip (Affymetrix 6.0) for each of the HapMap subjects using Affymetrix Genotyping Console™. All the studied chips passed the CQC threshold of 0.4, which is recommended by Affymetrix for controlling genotyping quality.

We used the Affymetrix® Genotyping Console (GTC) 3.0.1 to generate a reference genome for comparing copy numbers generated using all 270 raw intensity files (CEL files) from the human SNP 6.0 array (Affymetrix® Genotyping Console 3.0.1 User Manual). Using the common reference genome for comparisons, the intensity ratio of each probe (both SNP probe and copy number probe) on each array was calculated. The boundaries of the common CNV segments were determined using the predefined CNV regions (48). A hidden Markov model (HMM) was utilized to call the copy number state (i.e. the number of DNA copies, two for a diploid genome) for each identified CNV. Genotype calls for the common CNVs (observed in multiple unrelated subjects) were determined using the Canary algorithms (48). The chromosomal boundaries as well as the copy number states were exported and utilized in the association analysis of gene expression levels.

Identification of potential interacting CNV pairs

We retrieved 23 640 pairs of human PPI from BioGRID database version 2.0.51 (http://www.thebiogrid.org/). The BioGRID database was developed to house and distribute collections of protein and genetic interactions from major model organism species including human (18). Using 1141 CNV regions identified in 210 unrelated HapMap samples, we identified 37 CNV pairs where both CNVs overlapped with the genic region represented in a PPI pair. In this study, we define the 'overlap' as more than 1 bp in common between the CNV region and the genomic region (including the 5 kb flanking regions) of a given gene represented in the PPI pair. The annotations (e.g. gene symbol and chromosomal locations) of human genes were obtained from NCBI Built 36.1.

Statistical analysis

Population genetic parameters for CNVs (copy number states) were calculated, including MAFs, genotype frequencies and a chi-square test for departures from expectations under Hardy–Weinberg equilibrium for the 210 unrelated subjects in three racial groups. Summary statistics for CNVs and least-squares linear regression models were estimated using the statistical software R. Single CNV associations and CNV–CNV epistatic associations with gene expression levels were evaluated in the pooled HapMap samples using the linear regression model. A single variable was used to represent the additive effect of a CNV genotype (i.e. the copy number states of 0, 1 or 2 exported from Affymetrix® GTC 3.0.1 using Canary).

In this study, we pooled the three racial groups to increase the power of the analysis and to adjust for population stratification we used PC representations of the genetic variation in this pooled sample in the linear regression models (49). Briefly, 906 602 SNPs were genotyped in the HapMap samples using the Affymetrix® Genome-Wide Human SNP Array 6.0 platform. SNPs were excluded if they had an unknown chromosomal location, a call rate less than 95% or a MAF less than 0.05. These quality control filters resulted in 752 286 autosomal SNPs available for analysis in 210 independent HapMap subjects. The top 10 PCs of the 752 286 autosomal SNPs (MAF $>$ 0.05 and call rate $>$95%) were calculated and used as covariates in the multivariable linear regression model of testing epistatic associations of CNV pairs.

$$\begin{aligned}\mathrm{eQTL} = {} & \beta_1\mathrm{PC1} + \cdots + \beta_{10}\mathrm{PC10} + \beta_{11}\mathrm{CNV1} + \beta_{12}\mathrm{CNV2} \\ & + \beta_{13}\mathrm{CNV1} * \mathrm{CNV2} + \epsilon\end{aligned}$$

β_{13} was tested to determine the statistical significance of the epistatic associations of the CNV pairs.

Conflict of Interest statement. None declared.

FUNDING

This study was supported by National Institute of Health grant HL100245 and NS041558.

REFERENCES

1. Bateson, W. (1909) *Mendel's Principles of Heredity*. Cambridge University.
2. Fisher, R.A. (1918) The correlation between relatives on the supposition of Mendelian inheritance. *Trans. Roy. Soc. Edinb.*, **52**, 399–433.
3. Phillips, P.C. (2008) Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.*, **9**, 855–867.
4. Cordell, H.J. (2009) Detecting gene–gene interactions that underlie human diseases. *Nat. Rev. Genet.*, **10**, 392–404.
5. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
6. Moore, J.H. and Williams, S.M. (2009) Epistasis and its implications for personal genetics. *Am. J. Hum. Genet.*, **85**, 309–320.
7. Herold, C., Steffens, M., Brockschmidt, F.F., Baur, M.P. and Becker, T. (2009) INTERSNP: genome-wide interaction analysis guided by a priori information. *Bioinformatics*, **25**, 3275–3281.
8. Pattin, K.A. and Moore, J.H. (2008) Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases. *Hum. Genet.*, **124**, 19–29.
9. Pattin, K.A. and Moore, J.H. (2009) Role for protein–protein interaction databases in human genetics. *Expert Rev. Proteomics*, **6**, 647–659.
10. Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P. *et al.* (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, **440**, 637–643.
11. Valente, A.X., Roberts, S.B., Buck, G.A. and Gao, Y. (2009) Functional organization of the yeast proteome by a yeast interactome map. *Proc. Natl Acad. Sci. USA*, **106**, 1490–1495.
12. Tarassov, K., Messier, V., Landry, C.R., Radinovic, S., Serna Molina, M.M., Shames, I., Malitskaya, Y., Vogel, J., Bussey, H. and Michnick, S.W. (2008) An in vivo map of the yeast protein interactome. *Science*, **320**, 1465–1470.
13. Tong, A.H., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., Chang, M. *et al.* (2004) Global mapping of the yeast genetic interaction network. *Science*, **303**, 808–813.
14. Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L., Toufighi, K., Mostafavi, S. *et al.* (2010) The genetic landscape of a cell. *Science*, **327**, 425–431.
15. Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S. *et al.* (2005) A human protein–protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–968.
16. Sowa, M.E., Bennett, E.J., Gygi, S.P. and Harper, J.W. (2009) Defining the human deubiquitinating enzyme interaction landscape. *Cell*, **138**, 389–403.
17. Gandhi, T.K., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K.N., Mohan, S.S., Sharma, S., Pinkert, S., Nagaraju, S., Periaswamy, B. *et al.* (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat. Genet.*, **38**, 285–293.
18. Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A. and Tyers, M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.
19. Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stumpflen, V., Mewes, H.W. *et al.* (2005) The MIPS mammalian protein–protein interaction database. *Bioinformatics*, **21**, 832–834.
20. Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., de Grassi, A., Lee, C. *et al.* (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, **315**, 848–853.
21. Sun, Y.V., Peyser, P.A. and Kardia, S.L. (2009) A common copy number variation on chromosome 6 association with the gene expression level of endothelin 1 in transformed B lymphocytes from three racial groups. *Circ. Cardiovasc. Genet.*, **2**, 483–488.
22. Singleton, A.B., Farrer, M., Johnson, J., Singleton, A., Hague, S., Kachergus, J., Hulihan, M., Peuralinna, T., Dutra, A., Nussbaum, R. *et al.* (2003) Alpha-synuclein locus triplication causes Parkinson's disease. *Science*, **302**, 841.
23. The International Schizophrenia Consortium Manuscript preparation, Stone, J.L., O'Donovan, M.C., Gurling, H., Kirov, G.K., Blackwood, D.H., Corvin, A., Craddock, N.J., Gill, M. *et al.* (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*, **455**, 237–241.
24. Walsh, T., McClellan, J.M., McCarthy, S.E., Addington, A.M., Pierce, S.B., Cooper, G.M., Nord, A.S., Kusenda, M., Malhotra, D., Bhandari, A. *et al.* (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science*, **320**, 539–543.
25. Xu, B., Roos, J.L., Levy, S., van Rensburg, E.J., Gogos, J.A. and Karayiorgou, M. (2008) Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat. Genet.*, **40**, 880–885.
26. Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J. *et al.* (2007) Strong association of de novo copy number mutations with autism. *Science*, **316**, 445–449.
27. Weiss, L.A., Shen, Y., Korn, J.M., Arking, D.E., Miller, D.T., Fossdal, R., Saemundsen, E., Stefansson, H., Ferreira, M.A., Green, T. *et al.* (2008) Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.*, **358**, 667–675.
28. McCarroll, S.A., Huett, A., Kuballa, P., Chilewski, S.D., Landry, A., Goyette, P., Zody, M.C., Hall, J.L., Brant, S.R., Cho, J.H. *et al.* (2008) Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat. Genet.*, **40**, 1107–1112.
29. Ng, C.C., Koyama, K., Okamura, S., Kondoh, H., Takei, Y. and Nakamura, Y. (1999) Isolation and characterization of a novel TP53-inducible gene, TP53TG3. *Genes Chromosomes Cancer*, **26**, 329–335.
30. Iwai, N., Baba, S., Mannami, T., Ogihara, T. and Ogata, J. (2002) Association of a sodium channel alpha subunit promoter variant with blood pressure. *J. Am. Soc. Nephrol.*, **13**, 80–85.
31. Harvey, K.F., Dinudom, A., Cook, D.I. and Kumar, S. (2001) The Nedd4-like protein KIAA0439 is a potential regulator of the epithelial sodium channel. *J. Biol. Chem.*, **276**, 8597–8601.
32. Ravasi, T., Suzuki, H., Cannistraci, C.V., Katayama, S., Bajic, V.B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N. *et al.* (2010) An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, **140**, 744–752.
33. Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S.M., Habegger, L., Rozowsky, J., Shi, M., Urban, A.E. *et al.* (2010) Variation in transcription factor binding among humans. *Science*, **328**, 232–235.
34. Mehdy, M.C., Ratner, D. and Firtel, R.A. (1983) Induction and modulation of cell-type-specific gene expression in dictyostelium. *Cell*, **32**, 763–771.
35. Whitfield, M.L., Finlay, D.R., Murray, J.I., Troyanskaya, O.G., Chi, J.T., Pergamenschikov, A., McCalmont, T.H., Brown, P.O., Botstein, D. and Connolly, M.K. (2003) Systemic and cell type-specific gene expression patterns in scleroderma skin. *Proc. Natl Acad. Sci. USA*, **100**, 12319.
36. Palmer, C., Diehn, M., Alizadeh, A.A. and Brown, P.O. (2006) Cell-type specific gene expression profiles of leukocytes in human peripheral blood. *BMC Genomics*, **7**, 115.
37. Cheverud, J.M. (2000) Detecting epistasis among quantitative trait loci. *Epistasis and the Evolutionary Process.* Oxford University Press, New York, pp. 58–81.
38. Steinmetz, L.M., Sinha, H., Richards, D.R., Spiegelman, J.I., Oefner, P.J., McCusker, J.H. and Davis, R.W. (2002) Dissecting the architecture of a quantitative trait locus in yeast. *Nature*, **416**, 326–330.
39. Kroymann, J. and Mitchell-Olds, T. (2005) Epistasis and balanced polymorphism influencing complex trait variation. *Nature*, **435**, 95–98.
40. Mackay, T.F. and Anholt, R.R. (2006) Of flies and man: drosophila as a model for human complex traits. *Annu. Rev. Genomics Hum. Genet.*, **7**, 339–367.
41. Flint, J. and Mackay, T.F. (2009) Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Res.*, **19**, 723–733.

42. Smith, J.A., Turner, S.T., Sun, Y.V., Fornage, M., Kelly, R.J., Mosley, T.H., Jack, C.R., Kullo, I.J. and Kardia, S.L. (2009) Complexity in the genetic architecture of leukoaraiosis in hypertensive sibships from the GENOA study. *BMC Med. Genomics*, **2**, 16.
43. Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F. and Moore, J.H. (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.*, **69**, 138–147.
44. Zhang, Y. and Liu, J.S. (2007) Bayesian inference of epistatic interactions in case-control studies. *Nat. Genet.*, **39**, 1167–1173.
45. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
46. Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P. *et al.* (2010) Origins and functional impact of copy number variation in the human genome. *Nature*, **7289**, 704–712.
47. The International HapMap Consortium. (2003) The international HapMap project. *Nature*, **426**, 789–796.
48. Korn, J.M., Kuruvilla, F.G., McCarroll, S.A., Wysoker, A., Nemesh, J., Cawley, S., Hubbell, E., Veitch, J., Collins, P.J., Darvishi, K. *et al.* (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.*, **40**, 1253–1260.
49. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.