中央研究院
資訊科學研究所
Institute of Information Science, Academia Sinica • Taipei, Taiwan, ROC

TR-IIS-12-001

Construction of Gene Clusters Resembling Genetic Causal Mechanisms for Common Complex Disease with an Application to Young-Onset Hypertension

Ke-Shiuan Lynn, Chen-Hua Lu, Han-Ying Yang,
Wen-Lian Hsu and Wen-Harn Pan

Mar. 20, 2012 || Technical Report No. TR-IIS-12-001

http://www.iis.sinica.edu.tw/page/library/TechReport/tr2012/tr12.html

Construction of Gene Clusters Resembling Genetic Causal Mechanisms for Common Complex Disease with an Application to Young-Onset Hypertension

Ke-Shiuan Lynn[1], Chen-Hua Lu[1], Han-Ying Yang[1], Wen-Lian Hsu[1,*] and Wen-Harn Pan[2,*]

[1]Institute of Information Science, Academia Sinica, Taipei, Taiwan

[2] Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan

ABSTRACT

Motivation: Lack of power and reproducibility are caveats of genetic association studies of common complex diseases. Indeed, the heterogeneity of disease etiology demands that causal models consider the simultaneous involvement of multiple genes. Rothman's sufficient-cause model, which is well known in epidemiology, provides a framework for such a concept. In the present work, we developed a three-stage algorithm to construct gene clusters resembling Rothman's causal model for a complex disease, starting from finding influential gene pairs followed by grouping homogeneous pairs.

Result: The algorithm was trained and tested on 2,772 hypertensives and 6,515 normotensives extracted from four large Caucasian and Taiwanese databases. The constructed clusters, each featured by a major gene interacting with many other genes and identified a distinct group of patients, reproduced in both ethnic populations and across three genotyping platforms. We present the 14 largest gene clusters which were capable of identifying 19.3% of hypertensives in all the datasets and 41.8% if one dataset was excluded for lack of phenotype information. Although a few normotensives were also identified by the gene clusters, they usually carried less risky combinatory genotypes (insufficient causes) than the hypertensive counterparts. After establishing a cut-off percentage for risky combinatory genotypes in each gene cluster, the 14 gene clusters achieved a classification accuracy of 82.8% for all datasets and 98.9% if the information-short dataset was excluded. Furthermore, not only 9 of the 14 major genes but also many other contributing genes in the clusters are associated with hypertension-related functions. Our results provide insights into polygenic aspect of hypertension etiology.

Availability: Supplementary Data Files and MATLAB files that generate Figs. 3-5 are available at http://ms.iis.sinica.edu.tw/genetic_causal_pies /index.htm.

Contact: pan@ibms.sinica.edu.tw or hsu@iis.sinica.edu.tw

Keywords: genetic causal pie, sufficient cause, data-mining, young-onset hypertension, complex disease

1 INTRODUCTION

Effective mapping of complex disease genes is one of the major goals of genomic research. With advancements in genomic technology, the genome-wide association study (GWAS) approach has been adopted to identify novel genes for common complex diseases owing to its ability to simultaneously examine a large number of polymorphism-phenotype associations (Altshuler *et al*., 2008; McCarthy *et al*., 2008; Frazer *et al*., 2009; Hardy and Singleton, 2009). Although GWAS have indeed identified certain susceptibility genes for many diseases, the genes thus far discovered mostly have been associated with small to modest effects (Altshuler *et al*., 2008; McCarthy *et al*., 2008; Frazer *et al*., 2009; Hardy and Singleton, 2009; Moore *et al*., 2010). For very complex diseases, such as hypertension, GWAS have revealed very few genes despite a large number of patients that have been studied. It is generally accepted that common complex disease etiologies are heterogeneous in nature (Pan *et al*., 2006; Kohara *et al*., 2008; Lynn *et al*., 2009; Moore *et al*., 2010). In "state of art" GWAS approach, however, inheritance models involving gene-gene interactions and gene-environment interactions (Zerba *et al*., 1996; Sing *et al*., 2003; Musani *et al*., 2007; Cordell, 2009; Moore and Williams, 2009) have not been taken into consideration.

Rothman's concept of sufficient causes (Rothman, 1976, 2005) describes scenarios in which multiple causal mechanisms can all lead to the development of a disease. Each mechanism is depicted as a causal pie, composed of several component causes, and the number of causes varies in these mechanisms. These component causes—genetic or environmental—can be shared or completely different across causal mechanisms. Thus, the probability of ascertaining a disease is increased as a person carries more and more component causes. Under such a conceptual framework, if a gene is only involved in few of the causal pies, its effect toward a disease could be insignificant when all patients are considered. This model provides an explanation for low reproducibility across studies. Although Rothman's causal pie model is well known in epidemiology, few attempts have been made to construct such pies, not to mention its recognition and application in the genetic field.

In the present study, we focused on constructing gene clusters resembling genetic causal pies using genome-wide single-nucleotide polymorphism (SNP) data for young-onset hypertension (YOH), which has a stronger genetic attribute than its late-onset counterpart (Mongeau, 1987; Pan *et al*., 2000). We made use of two large Caucasian databases, the Framingham Heart Study (FHS, http://www.framinghamheartstudy.org, Dawber *et al*., 1951) and Wellcome Trust Case

Control Consortium (WTCCC, http://www.wtccc.org.uk, Wellcome Trust Case Control Consortium, 2007), and two large Taiwanese databases, the Taiwan Young-Onset Hypertension Study (Taiwan YOH, Pan *et al*., 2000) and Taiwan Han Chinese Cell and Genome Bank (THCCG, http://ncc.sinica.edu.tw/han-chinese_genomebank, Pan *et al*., 2006). We aimed to find either single SNPs or multiple SNP sets each of which resembles a genetic causal pie and could distinguish a certain proportion of hypertensives (HTs) from normotensive controls (NCs). Owing to limited databases and many gene clusters found in the databases, we intended to demonstrate the existence of such causal pie-like gene clusters rather than to construct all the genetic causal pies. We thus developed an algorithm to construct influential (as many as patients being identified) and effective (cluster components identifying the same group of patients) gene clusters. We first searched for pair-wise gene-gene interactions primarily observable in FHS and Taiwan YOH patients via an exhaustive search. Gene (SNP) pairs that identified similar patients were further merged into clusters following the logic of the multiple genetic causal pies framework. The resulting gene clusters were then tested for reproducibility on various platforms (including gene expression data) and examined for robustness in varied algorithm parameters. Crucial gene pairs that represented minimum and sufficient component causes in each of the genetic causal pies were searched, and their effects to hypertension onset were discussed. Moreover, influential functions, process and pathways of these genes were collated to shed light on hypertension etiology.

2 METHODS

This study was approved by the Internal Review Board of Academia Sinica. All four databases used in this paper were approved by local institutional review boards or equivalent committees and all participants in the databases signed a written informed consent at all institutions/hospitals where they were recruited and human experimentation was conducted.

2.1 Characteristics of the four employed databases

The FHS database contains 7,126 subjects (Framingham, Massachusetts, U.S.A., predominantly Caucasian) among whom 6,748 were assayed by the Affymetrix500k platform with detailed information on blood pressure measurements and medications. The WTCCC database currently consists of datasets from three studies (WTCCC1~WTCCC3). However, only the dataset from WTCCC1 was available at the time our experiment was conducted. The dataset includes 2,001 hypertensive cases and 3,004 NCs (1504 from the 1958 British Birth Cohort and 1500 from the UK Blood

Service Control Group), all from the British population and assayed by the Affymetrix500k platform. The Taiwan YOH database contains 1,023 well-characterized YOH subjects, among which 175 were assayed by the Affymetrix100k platform, 200 were assayed by the Affymetrix500k platform, and 400 were assayed by the Illumina550k platform. The THCCG database involved 3,435 sampled residents with detailed clinical information. Among them, 175 were assayed by the Affymetrix100k platform, 468 were assayed by the Affymetrix500k platform, and 1,000 were assayed by the Illumina550k platform.

2.2 Training and test datasets

We extracted suitable samples from the four databases to construct our training and test datasets. To prevent ambiguous data from disrupting our data mining–based approach, NC subjects in FHS and THCCG subjects with multiple high blood pressure readings (≥120/80 mmHg) were removed from the datasets. To ensure a strong genetic effect on the onset of hypertension, late-onset (onset >50 years) and secondary HT patients were also excluded. Detailed inclusion criteria for HT patients and for NC subjects are listed in Supplementary Method 1. In addition, we adopted the "SNP Finder" in SNPper (http://snpper.chip.org/bio/snpper-enter, Riva and Kohane, 2002) to search for intragenic SNPs and their corresponding gene symbols in each genotyping platform. The resultant training and test datasets are summarized below and detailed subject IDs are provided in Supplementary Data File 1.

Training datasets:

(1) Caucasian subset (FHS_Affy500k): Affymetrix500k genotype data extracted from FHS, including 214,383 intragenic SNPs for 3,186 Framingham residents, among whom 305 developed hypertension and 2881 remained normotensive during follow-up

(2) Taiwanese subset (Taiwan_Affy500k): Affymetrix500k genotype data, including 213,353 intragenic SNPs for 200 HT cases from the Taiwan YOH study and 184 NC subjects from THCCG

Test datasets:

(1) Caucasian subset (WTCCC_Affy500k): Affymetrix500k genotype data extracted from WTCCC, including 214,383 intragenic SNPs for 2,001 HT cases and 3,004 NC subjects

(2) Taiwanese subsets:

(a) Taiwan_Illu550k: Illumina550k genotype data, including 221,828 intragenic SNPs for 200 HT cases from the Taiwan YOH study and 400 NC subjects from THCCG

(b) Taiwan_Affy100k: Affymetrix100k genotype data, including 47,038 intragenic SNPs for 129 HT cases from the Taiwan YOH study and 129 NC subjects from THCCG

Some of the subjects overlapped in the Taiwan_Illu550k and Taiwan_Affy100k, leaving a total of 266 unique HT cases and 446 unique NC subjects in the Taiwan test dataset (see Supplementary Fig. 1 for detailed calculations). To demonstrate reproducibility among genotyping platforms, these overlapped subjects were not removed from the two test datasets because the adopted SNPs differed between the two platforms, and some patients may have been identified by one of them. However, the overlapped subjects were counted only once for the evaluation of classification performance.

More importantly, because we do not have phenotype information for WTCCC, late-onset (WTCCC recruited HT patients < 60 yr of age but we required ≤ 50 yr of age) may have been included in the HT subset, whereas high body mass index, high blood sugar, or borderline blood pressure (120/80~140/90 mmHg) subjects may have been included in the NC subset. For comparison, in FHS_Affy500k, only 305 of 557 (54.8%) HT patients and 2,881 of the remaining 6191 (46.5%) subjects who had genotype data and satisfied our inclusion criteria were selected from the FHS database. Therefore, although the WTCCC_Affy500k was used as one of the test datasets, focus should be placed on the reproducibility of the constructed gene sets in its HT population instead of on its classification accuracy.

2.3 Detection of gene-gene interaction

Several definitions of gene-gene interaction (or epistasis) have been proposed in the literature (Musani *et al*., 2007; Cordell, 2009; Moore and Williams, 2009; Neuman and Rice, 1992). Based on these definitions, many methods have also been developed to detect gene-gene interactions. These methods can be roughly categorized into three classes: exhaustive search, regression-based approach, and data-mining approach. Exhaustive search, which performs a certain test for all possible pairs in the dataset, is the simplest way to detect interactions (Marchini *et al*., 2005). However, such a method is not suitable for higher-order interactions since the number of tests grows exponentially and soon becomes computationally infeasible. Popular in statistical analysis packages, regression-based approaches attempt to fit a regression model (linear, logistic, or logic) between subjects' multilocus genotypes and their outcomes and to test whether the effect

of multiplicative terms is negligible (Fisher, 1918; Armitage *et al*., 2002; Kooperberg *et al*., 2005). In contrast to the previous two approaches, data-mining approaches are preferred for detecting high dimensional interactions. They focus on selecting a minimal subset of loci so that, in the subspace spanned by the loci, a hyperplane or a hypersurface can be constructed to distinguish different outcome groups. Examples of this category are multifactor-dimensionality reduction (MDR) (Ritchie *et al*., 2001), combinatorial partitioning method (CPM) (Nelson *et al*., 2001), genetic programming (Nunkesser *et al*., 2007), neural networks (Motsinger-Reif *et al*., 2008) and support vector machines (Chen *et al*., 2008). Other methods, including Bayesian model-based approach (Zhang *et al*., 2007) and entropy-based approach (Kang *et al*., 2008), have also been developed.

In our preliminary studies, we observed that many interacting genes have shared genes. Also, gene pairs with a shared gene often identified a similar group of individuals and thus can be organized together to form a gene cluster anchored by a major gene. To detect all such clusters and their component genes in a genome-wide data, a method that can quickly detect all the possible interacting gene pairs is needed. To this end, we adopted an exhaustive search with simple testing criteria to detect single genes and interacting gene pairs that are associated with increased risk. We first define the following terms that were used in our detection method:

Risky genotype set: certain genotypes (as illustrated in Fig. 1, each as a risky genotype) that are observable in at least C_{HT}% ($C_{HT} > 0$) of a diseased population and at most C_{NC}% ($C_{HT} > C_{NC} \geq 0$) of a non-diseased population

Single disease gene: a single gene that exhibits a risky genotype set

Risky combinatory genotype set: certain combinatory genotypes (as illustrated in Fig. 2, each as a risky combinatory genotype) that are observable in at least C_{HT}% ($C_{HT} > 0$) of a diseased population and at most C_{NC}% ($C_{HT} > C_{NC} \geq 0$) of a non-diseased population

Gene-gene interaction: a pair of genes that exhibit a risky combinatory genotype set without either of them being a disease gene

Disease gene pair: a pair of genes that exhibit a gene-gene interaction

To identify disease genes and disease gene pairs, we exhaustively searched for all the risky genotype sets for all SNPs and then search for all the risky combinatory genotype sets for all SNP pairs. We noted that, the value of C_{NC} was set to a small value instead of zero in real applications to tolerate possible genotyping and sampling errors in the dataset. In addition, we adopted the ceiling function, $\lceil x \rceil = \min\{m \in Z | m \geq x\}$, in our algorithm to deal with the fraction resulting from the product of the criterion and sample size. Such a design allowed more qualified gene and gene pairs for datasets with a small NC

population where genotyping and sampling qualities usually exhibit large variations. Also, although we used SNP data to construct genetic clusters, we will merge them by the associated genes for the subsequent cross-platform comparisons and functional analysis.

	SNP: dominant				SNP: recessive		
genotype	AA	AG	GG	genotype	AA	AG	GG
HT/NC	HT_{AA}/NC_{AA}	HT_{AG}/NC_{AG}	HT_{GG}/NC_{GG}	HT/NC	HT_{AA}/NC_{AA}	HT_{AG}/NC_{AG}	HT_{GG}/NC_{GG}
Risky genotype set = {AG, GG}, if $(HT_{AG}+HT_{GG}) \geq C_{HT}$% in HT cases and $(NC_{AG}+NC_{GG}) \leq C_{NC}$% in NC controls				Risky genotype set = {GG}, if $HT_{GG} \geq C_{HT}$% in HT cases and $NC_{GG} \leq C_{NC}$% in NC controls			

Fig. 1. Criteria for a risky genotype set of a single disease gen (SNP). In the above example, the SNP is AG polymorphism with disease allele G.

HT/NC		SNP1: dominant			HT/NC		SNP1: dominant		
		AA	AG	GG			AA	AG	GG
SNP2: dominant	CC	HT_{AACC}/NC_{AACC}	HT_{AGCC}/NC_{AGCC}	HT_{GGCC}/NC_{GGCC}	SNP2: recessive	CC	HT_{AACC}/NC_{AACC}	HT_{AGCC}/NC_{AGCC}	HT_{GGCC}/NC_{GGCC}
	CT	HT_{AACT}/NC_{AACT}	HT_{AGCT}/NC_{AGCT}	HT_{GGCT}/NC_{GGCT}		CT	HT_{AACT}/NC_{AACT}	HT_{AGCT}/NC_{AGCT}	HT_{GGCT}/NC_{GGCT}
	TT	HT_{AATT}/NC_{AATT}	HT_{AGTT}/NC_{AGTT}	HT_{GGTT}/NC_{GGTT}		TT	HT_{AATT}/NC_{AATT}	HT_{AGTT}/NC_{AGTT}	HT_{GGTT}/NC_{GGTT}
Risky combinatory genotype set = {AGCT, GGCT, AGTT, GGTT}, if $(HT_{AGCT}+HT_{GGCT}+HT_{AGTT}+HT_{GGTT}) \geq C_{HT}$% in HT cases and $(NC_{AGCT}+NC_{GGCT}+NC_{AGTT}+NC_{GGTT}) \leq C_{NC}$% in NC controls					Risky combinatory genotype set = {AGTT, GGTT}, if $(HT_{AGTT}+HT_{GGTT}) \geq C_{HT}$% in HT cases and $(NC_{AGTT}+NC_{GGTT}) \leq C_{NC}$% in NC controls				
HT/NC		SNP1: recessive			HT/NC		SNP1: recessive		
		AA	AG	GG			AA	AG	GG
SNP2: dominant	CC	HT_{AACC}/NC_{AACC}	HT_{AGCC}/NC_{AGCC}	HT_{GGCC}/NC_{GGCC}	SNP2: recessive	CC	HT_{AACC}/NC_{AACC}	HT_{AGCC}/NC_{AGCC}	HT_{GGCC}/NC_{GGCC}
	CT	HT_{AACT}/NC_{AACT}	HT_{AGCT}/NC_{AGCT}	HT_{GGCT}/NC_{GGCT}		CT	HT_{AACT}/NC_{AACT}	HT_{AGCT}/NC_{AGCT}	HT_{GGCT}/NC_{GGCT}
	TT	HT_{AATT}/NC_{AATT}	HT_{AGTT}/NC_{AGTT}	HT_{GGTT}/NC_{GGTT}		TT	HT_{AATT}/NC_{AATT}	HT_{AGTT}/NC_{AGTT}	HT_{GGTT}/NC_{GGTT}
Risky combinatory genotype set = {GGCT, GGTT}, if $(HT_{GGCT}+HT_{GGTT}) \geq C_{HT}$% in HT cases and $(NC_{GGCT}+NC_{GGTT}) \leq C_{NC}$% in NC controls					Risky combinatory genotype set = {GGTT}, if $HT_{GGTT} \geq C_{HT}$% in HT cases and $NC_{GGTT} \leq C_{NC}$% in NC controls				

Fig. 2. Criteria for a risky combinatory genotype set of a disease gene (SNP) pair. In the above example, SNP1 is the AG polymorphism with disease allele G whereas SNP2 is the CT polymorphism with disease allele T.

2.4 The gene cluster construction algorithm

Two problems were encountered as we attempted to organize the detected gene pairs into gene clusters: (i) value assignment for C_{HT} and C_{NC}, and (ii) removal of false positive gene pairs. For stringent detection criteria, i.e. a very large C_{HT} with a very small C_{NC}, the

detected disease genes and gene pairs can be too conservative to provide clear information about the underlying disease mechanism. However, as the criteria were relaxed, the detected disease genes and gene pairs increased quickly and soon became unmanageable. To solve this dilemma, we proposed first using stringent criteria to generate a manageable amount of candidates, and then relaxing the criteria to search for additional gene pairs for each gene cluster. On the other hand, false positive gene pairs in a gene cluster degenerate its classification performance and provide false information to the underlying disease mechanism. Although two gene pairs with a shared gene may not identified identical individuals, those identified by a gene cluster usually carry more risky combinatory genotypes in the gene cluster than the others (see Supplementary Fig. 2 for a demonstration). Therefore we proposed accumulating multiple gene pairs that have a shared gene so as to locate the frequently identified subjects (FIS) and then to remove false positive gene pairs that identified subjects other than these FIS. The gene clusters formed by our algorithm is rather intrinsic in the datasets and may resemble Rothman's genetic causal pies. Furthermore, we have proven in Supplementary Method 2 that the probability of a false positive gene cluster containing k non-LD SNP pairs and identifying m subjects in a population of n subjects is bounded above by $(m/n)^k$.

Our algorithm consists of three stages: cluster selection, component growth and component pruning. During the first stage, we set up a set of stringent criteria to identify influential disease gene pairs and grouped them with shared genes. Then at the second stage, we iteratively relaxed the criteria to encourage effective gene clusters to include additional gene pairs until new gene pairs started to identify different groups of HT patients. Finally, at the third stage, all disease gene pairs that identified different groups of patients were removed from the cluster. In this work, we used $C_{HT} = 2.0$ and $C_{NC} = 0.1$ to produce manageable cluster size in the first stage. Let the sample size of the HT and NC populations be S_{HT} and S_{NC}, respectively. The proposed gene cluster construction algorithm comprises the following steps.

Step 1 Cluster selection. Select a conservative set of gene clusters using stringent criteria:

- Step 1.1 Set $t_{HT} = t_{HT0} = \lceil C_{HT} \times S_{HT} \rceil$ and $t_{NC} = t_{NC0} = \lceil C_{NC} \times S_{NC} \rceil$ and use them to replace C_{HT} and C_{NC} in Figs. 1 and 2.
- Step 1.2 Search for all single genes with risky genotype sets (as illustrated in Fig.1) from the training datasets.
- Step 1.3 Search for all gene pairs with risky combinatory genotype sets (as illustrated in Fig.2) from the training datasets.
- Step 1.4 Find shared genes among the qualified gene pairs and use them to group the gene pairs.

Step 2 Component growth: For each constructed gene cluster, repeat the following steps until the HT patients identified by the new disease gene pairs differ from those by the existing ones:

Step 2.1 Set $t_{HT} = t_{HT} - 1$ and $t_{NC} = t_{NC} + 1$.

Step 2.2 Search for additional gene pairs with risky combinatory genotype sets from the training datasets.

Step 2.3 For each constructed gene cluster, record subject IDs that are frequently identified by its gene pair components:

Step 2.3.1 Locate the most frequently identified subjects (MFIS).

Step 2.3.2 Select subjects that were identified at least half the time compared with the MFIS, and categorize them as frequently identified subjects (FIS).

Step 2.3.3 Select the t_{HT} most frequently identified subjects if the number of FIS is less than t_{HT}.

Step 3 Component pruning:

Step 3.1 Select gene clusters with sufficient number ($\geq t_{HT0}$) of FIS.

Step 3.2 For each gene cluster, remove the gene pairs which identify subjects not in FIS.

2.5 Identification of influential genes using gene expression data

The gene clusters constructed from the SNP data usually consist of many genes which is disadvantageous for etiology analysis. We attempted to identify the influential genes using expression data. We selected 253 (135 in Taiwan_Affy500k and 118 in Taiwan_Illu550k) HT patients and 232 (36 in Taiwan_Affy500k and 196 in Taiwan_Illu550k) NC subjects who had gene expression data for the demonstration. For each subject, three replicates of genome-wide expression data were generated by the following three steps: (i) lymphocytes were isolated from the fasting blood immediately after it was drawn; (ii) the lymphoblastoid cell line was established via Epstein-Barr virus transformation; (iii) total RNA was extracted and hybridized onto three Phalanx Human OneArrays (HOA v5.1, Phalanx Biotech Group, Taiwan), each of which contains 39,200 polynucleotide probes with 25,215 of them mapped to the latest draft of the human genome.

Before merging the three replicates for each subject, we checked the consistency among them. We first computed the Pearson correlation coefficient for every two replicates and removed those with at least one correlation less than 0.9. We then checked the consistency for each gene if more than one of the replicates were available. The values of a gene were set to 0 (missing) if its minimum was less than 60% of its maximum. After

such an adjustment, replicates were merged using median values. A base-2 logarithm and Z-score global normalization were applied to the merged data. In the resultant data, we further set those values higher than 6 to 0 (missing) since they were outliers or represented false signals.

We developed an algorithm to identify influential genes in each gene cluster using the above gene expression data. Starting with the shared gene in a gene cluster, the algorithm iteratively added a gene in the gene cluster such that the HT patients carrying risky combinatory genotypes can be maximally discriminated (in terms of adjusted *p* values) from HT patients without carrying risky combinatory genotype, from NC subjects carrying risky combinatory genotype and from NC subjects without carrying risky combinatory genotype. This process is stopped if no better discrimination can be achieved by adding any other gene in the gene clusters. Pseudo code of this algorithm is provided in Supplementary Method 3. Although such a sequential search may not obtain the best discrimination among the above subject groups, it was adopted for its capability of selecting a small set of influential genes so that the underlying disease mechanisms in a gene cluster can be easily revealed.

3 RESULTS

3.1 The constructed gene clusters

No single gene was found to fulfill the stringent criteria, i.e., carrying risky genotype sets in at least 2.0% of HT patients and at most 0.1% of NC subjects in the training datasets. However, allele "CC" of *rs16854417*, an intronic SNP in *SLC9A9*, identified 3/305 (0.98%) and 4/200 (2%) of HT patients in FHS_Affy500k and in Taiwan_Affy550k, respectively (see Supplementary Table 1 for its allele frequencies in various datasets). In contrast, no NC subject in FHS and only one NC subject in Taiwan_Affy550k carried the CC allele for this SNP. Although this subject was a 53-year-old female with three normal blood pressure readings (120/78, 118/76 and 118/78 mmHg), she had a family history of hypertension.

In search for disease gene pairs, at the cluster-selection stage, we applied the stringent criteria and obtained 264 gene pairs in 360 genes, of which 24 were shared by multiple gene pairs. The 264 gene pairs were then grouped into 103 gene clusters, of which 24 consisted of multiple gene pairs and the remaining 79 contained only one gene pair. At the component growth stage, the criteria were relaxed accordingly for each of the 103 gene clusters to search for additional gene pairs. For the 79 two-gene clusters, the expansion was carried out twice, each assuming that one of the two genes was a shared

gene. As a result, the original 103 gene clusters were expanded to 196 gene clusters. Because small gene clusters were more likely to be false positives (see Supplementary Method 2 for the proof), we selected the 14 largest gene clusters which contained 17,170 gene pairs in 8,559 genes for the subsequent analysis. The 14 gene clusters were finally reduced to 17,115 gene pairs in 8,524 genes at the component pruning stage. We listed in Supplementary Table 2 the numbers of overlapping genes between gene clusters as a distance measure. The average percentage of overlapping genes in the 14 gene clusters is 4.9%. Such a low overlapping ratio is expected because the patients identified by the gene clusters exhibited few overlaps. We also provided in Supplementary Data Files 2 and 3 the gene symbols and SNP rs numbers in the gene clusters obtained at the cluster select stage and at the component pruning stage, respectively.

In Figs. 3 and 4, we demonstrate how the constructed gene clusters (denoted by their shared genes) identified different groups of subjects in training datasets and in test datasets, respectively, using a cluster visualization procedure (Supplementary Method 4). In both figures, the horizontal axis denotes the number of subjects, and the vertical axis denotes the number of gene pairs. A black pixel in the figures represents a subject who carried a risky combinatory genotype in the corresponding gene pair. Moreover, the gray areas in the figures indicate the portion of subjects carrying risky combinatory genotypes in the 14 gene clusters, whereas the light-blue horizontal lines denote that no corresponding gene pairs could be found in the dataset (due to differences among platforms).

Summarizing from Figs. 3 and 4, the percentages of the HT population carrying risky combinatory genotypes in the 14 gene clusters were 36.4% (184/505 of which 87/305 in Caucasian and 97/200 in Taiwanese) in training datasets and 15.5% (352/2,267 of which 214/2,001 in Caucasian and 138/266 in Taiwanese) in test datasets. The lower percentage in the test Caucasian may be due to the inclusion of late-onset patients in the WTCCC_Affy500k dataset. On the other hand, the percentages of NC population carrying the risky combinatory genotypes were 10.1% (309/3,065 of which 239/2,881 in Caucasian and 70/184 in Taiwanese) in the training datasets and 12.3% (425/3,450 of which 336/3,004 in Caucasian and 89/446 in Taiwanese) in test datasets. Detailed percentages for each of the 14 gene clusters in the five datasets are presented in Supplementary Table 3.

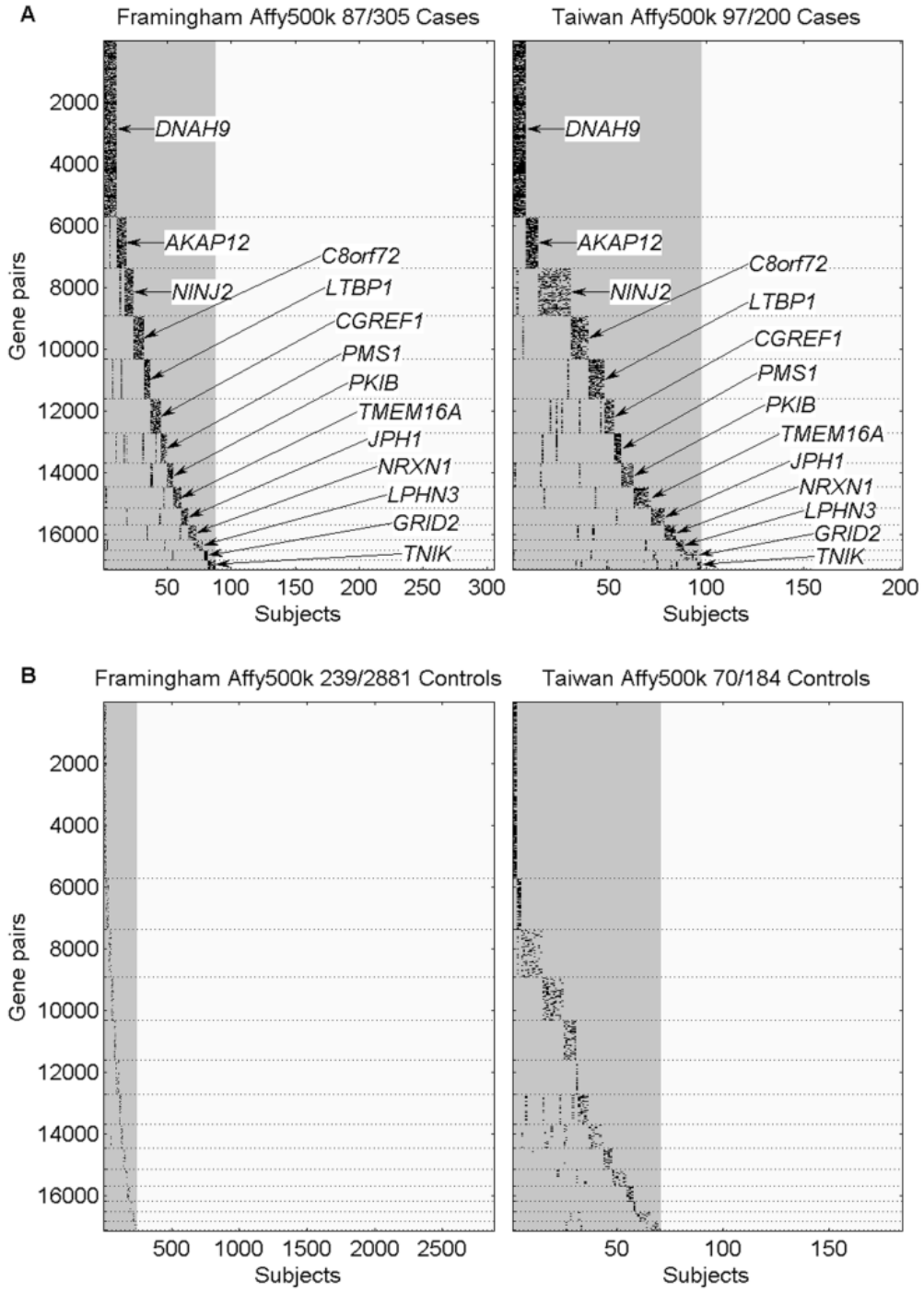


Fig. 3. The 14 gene-subject clusters (denoted by their shared genes) for (A) HT patients and for (B) NC subjects in the two training datasets, FHS_Affy500k (left) and Taiwan_Affy500k (right). The numerator in the title indicates the number of subjects identified by all gene clusters, whereas the denominator denotes the total number of subjects in the dataset. The gray areas indicate the total portion of identified subjects.

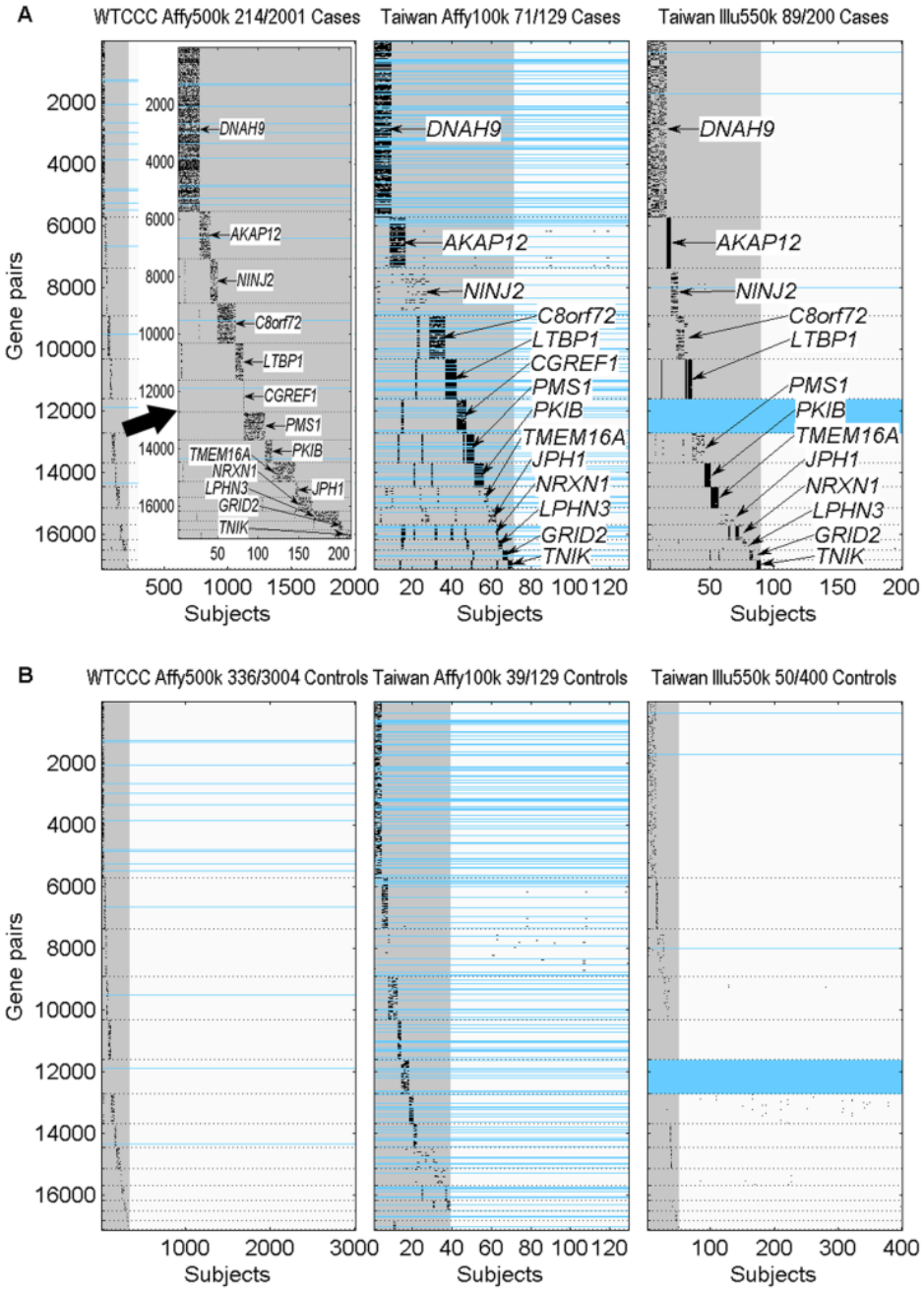


Fig. 4. The 14 gene-subject clusters (denoted by their shared genes) for (A) HT patients and for (B) NC subjects in the three test datasets, WTCCC_Affy500k (left), Taiwan_Affy100k (middle) and Taiwan_Illu550k (right). The numerator in the title indicates the number of subjects identified by all gene clusters, whereas the denominator denotes the total number of subjects in the dataset. The gray areas indicate the total portion of identified subjects, whereas the light-blue horizontal lines denote that no corresponding gene pairs could be found in the dataset.

3.2 Gene clusters resembling genetic causal pies

We have shown in Figs. 3 and 4 that the 14 constructed gene clusters were capable of identifying higher percentage of HT population than that of NC population and each gene cluster seemed to identify a distinct group of subjects. Further computing the number of risky combinatory genotypes carried in each subject, we found that HT patients usually carried more risky combinatory genotypes than NC subjects. This can be seen in Figs. 3 and 4 that the gene-subject clusters for HT patients (part (A)) usually exhibit darker blocks than those for NC subjects (part (B)). In Fig. 5, we used box plots to show the distributions of carried risky combinatory genotypes for HT patients (red) and for NC subjects (blue) that are identified by the same gene cluster in a dataset. Due to gene diversity among platforms, we used percentage (with respect to the size of the corresponding gene cluster), rather than number, of carried risky combinatory genotypes to demonstrate the difference between HT patients and HC subjects in the figure. Moreover, for each gene cluster in a dataset, we selected a percentage from the HT patients which resulted in minimum classification error as a threshold for disease onset. In Fig. 5, the threshold is represented by a dashed line between HT patients and NC subjects in a box plot, whereas the classification error is denoted by ER.

In addition to platform differences, our datasets also exhibited ethnic differences, i.e. Caucasian and Taiwanese. Although FHS_Affy500k, Taiwan_Affy500k, and WTCCC_Affy500k were all assayed on the Affymetrix500k platform, they represented different ethnic groups and therefore may have different thresholds for disease onset. We adopted two scenarios, S1 and S2, to compute the thresholds in the three datasets: the former assuming different thresholds for different ethnic groups, whereas the latter anticipating one threshold for a platform. With S1, we separated HT patients from NC subjects with 82.8% classification accuracy (sensitivity = 0.68, specificity = 0.93) or with 98.9% accuracy (Sen. = 0.98, Spec. = 1.0) if WTCCC_Affy500k was excluded. In accuracy calculations, a true positive was a HT patient with sufficient risky combinatory genotypes in at least one gene cluster, whereas a true negative indicated a NC subject with insufficient risky combinatory genotypes in all gene clusters. Detailed classification results are provided in Table 1. From the above results, we found that the sufficient risky combinatory genotype was similar to the sufficient cause in Rothman's causal pie model in that a subject must carry sufficient risky combinatory genotypes (component causes) in a gene cluster (a causal pie) for the onset of HT. In addition, most of the gene clusters were not only consistently observed in all the datasets but also clustered HT patients into distinct groups, suggesting multiple causal mechanisms for HT.

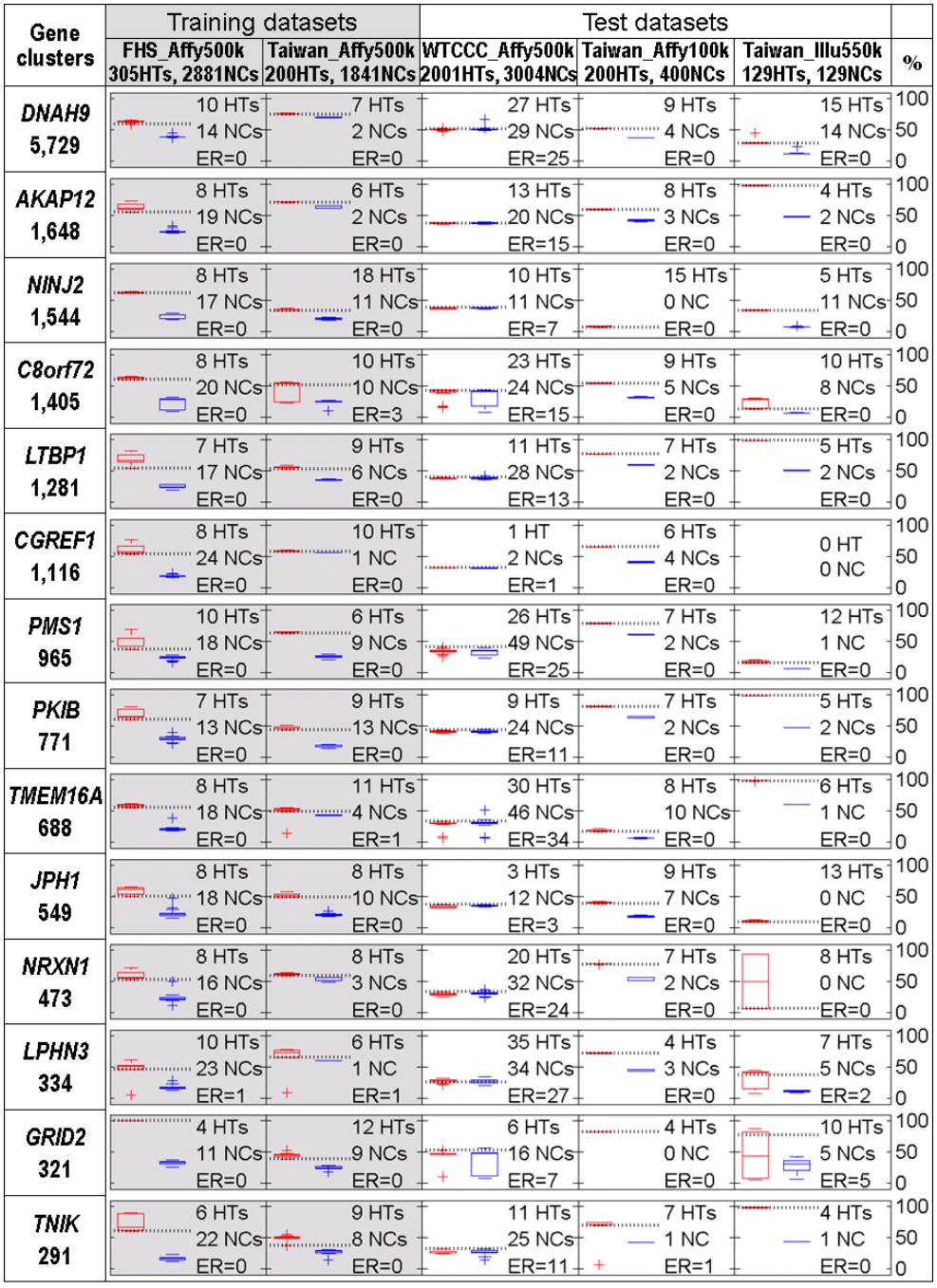


Fig. 5. Box plots of percentage of risky combinatory genotypes carried in the subjects who were identified by one of the fourteen gene clusters in the five datasets. Each box in the left-hand side shows the shared gene and the number of involved genes in a gene clusters. Red box plots represent HT patients, whereas blue box plots are for NC subjects. The horizontal dashed line represents the cut-off percentage of risky combinatory genotypes for defining HT computed for each data set. ER denotes classification error.

Table 1. Classification accuracy after establishing a cut-off percentage of risky combinatory genotypes in each gene cluster

Datasets	Population	Classification accuracy S1(S2)	Sensitivity S1(S2)	Specificity S1(S2)	Subjects with risky genotypes	
					HT	NC
FHS_Affy500k (training)	Caucasian	99.7%$^{+}$ (99.1%)	0.99 (0.98)	1.0 (0.996)	87	239
Taiwan_Affy500k (training)	Taiwanese	98.2%$^{+}$ (93.4%)	0.97 (0.97)	1.0 (0.89)	97	70
WTCCC_Affy500k (test)	Caucasian	61.6% (62.9%)	0.25 (0.26)	0.85(0.87)	214	336
Taiwan_Affy100k & Taiwan_Illu550k (test)	Taiwanese	98.2% (98.2%)	0.97 (0.94)	1.0 (1.0)	138*	89*
Overall	**Both**	**82.8% (82.5%)**	**0.68 (0.69)**	**0.93 (0.93)**	**536**	**734**
Overall except WTCCC_Affy500k	**Both**	**98.9% (97.5%)**	**0.98 (0.97)**	**1.0 (0.98)**	**322**	**398**

S1: threshold is computed for each dataset; S2: threshold is computed for each platform; $^{+}$Classification accuracy of the dataset evaluated using a five-fold validation procedure exhibit similar result and is presented in Supplementary Table 4; *see Supplementary Fig. 1 for detailed calculations.

3.3 Advantages in comparison with existed gene-gene interaction algorithms

Most of the algorithms for detecting gene-gene interaction dealt with complex model fitting and therefore their application to GWAS data can be very time consuming or simply infeasible. Also, methods that detect gene-gene interactions among multiple loci cannot guarantee global optimal solutions since only limited combinations are explored. In comparison with conventional methods, our algorithm has the following advantages:

Our detection algorithm is fast. We used simple testing criteria to detect gene-gene interactions which allowed us to quickly perform exhaustive search for all gene pairs. Using 19779 SNPs in 505 cases and 3065 controls (cluster1 in our training dataset) as an example, our algorithm took 5 hours and 51 minutes to finish all pairwise tests whereas the PLINK (Purcell et al., 2007) with the "fast-epistasis function" spent 8 hours and 43 minutes (more than 10 days for the "epistasis" function)

Our gene clustering algorithm is robust. We tested the robustness of our algorithm to criteria and to sample size changes, respectively. Detailed testing procedures were

described in Supplementary Method 5. We showed in Supplementary Fig. 3 that our algorithm was capable of detecting the same gene clusters either the criterion C_{HT} was increased from 2.0 to 2.5 (25% increased) or was decreased from 2.0 to 1.8 (10% decreased) or even to 1.5 (25% decreased). However when the criterion was increased to 2.5, it became too stringent and many gene clusters were no longer recoverable from the component growth stage. Thus, our gene cluster construction algorithm was very robust to criteria changes. We also showed in Supplementary Fig. 4 that the top 15 gene clusters remained >95% unchanged if 90% of the sample size was used (i.e., 10% of the data was randomly removed). The similarity decreased to around 55% if 70% of the samples were used, but the similarity remains >50% when 50% of the samples were used. These results showed that our gene cluster construction algorithm was robust to small (<10%) changes in sample size in comparison with many single-SNP analyses (Neale and Sham, 2004).

Our detection algorithm can deal with risky factor or protective factor or both. Unlike regression models that detect interactions that are associated with both risk factor and protective factor, our method can be assigned to detect interactions that are associated with either risky factor (as demonstrated in this manuscript) or protective factor (i.e., by setting $C_{NC} > C_{HT} \geq 0$).

Our gene clustering algorithm constructs reproducible clusters. Our test results also showed that the constructed gene clusters were reproduced in both ethnic populations and across three genotyping platforms. We have proven that the probability of a false positive gene cluster detected by our algorithm is very low. For a gene cluster containing k non-LD SNP pairs and identifying m subjects in a population of n subjects, such a probability is bounded above by $(m/n)^k$.

Our gene clustering algorithm can detect patient subgroups. Our observations showed that many interacting pairs identified similar group of patients. Accumulation of these gene pairs allowed us to identify patient subgroups whereas the identified patients help us to eliminate false positive gene pairs. Most of all, we found the gene clusters resemble different genetic causal pies in that subjects carrying sufficient number of risky combinatory genotype sets in the pie have very high possibility of disease onset.

3.4 Minimum and sufficient component causes

From Fig. 5, the number of risky combinatory genotypes involved in a genetic causal pie which ranged from hundreds to thousands in the 14 gene clusters is rather high. By analyzing the functions of genes involved in a gene cluster, we found that many of them perform similar functions, and thus some of the genes may be redundant to the causal pie. An intuitive guess for such a redundancy is LD among SNPs (in different genes). We

thus checked the LD between all SNPs in each of the 14 clusters using PLINK and retained only one of the gene pairs if their associated SNPs were found to have LD (D' ≥ 0.9). After the LD reduction, the sizes of the 14 gene clusters were reduced in an average percentage of 96.82% without changing their classification accuracies.

To further remove the redundancy, we computed the minimum number of genes in each of the 14 clusters (via the genetic algorithm) without disrupting the classification accuracy of the LD-reduced gene set. We found substantial reduction in number of genes (23–42%, as detailed in Supplementary Table 5) for the 14 clusters, leaving the resultant number of genes in each cluster ranged from a few dozen to a few thousand. Furthermore, after the reduction, the cut-off percentage remained similar to that of the original gene set. However, some genes that seemed irrelevant or redundant in one dataset may have been crucial for HT identification in the other datasets. Therefore, whether such reductions sustain in larger datasets warrants further investigation.

We also used the gene expression data to compute the minimum genes in each cluster via the algorithm proposed in section 2.5 and Supplementary Method 3. We found the number of genes in each gene cluster can be tremendously reduced to around a couple of dozen while HT patients carrying risky combinatory genotypes can still be significantly discriminated (adjusted $p < 10^{-5}$) from HT patients without carrying risky combinatory genotype, from NC subjects carrying risky combinatory genotypes and from NC subjects without carrying risky combinatory genotype (Supplementary Fig. 5). Using the selected subsets of genes to repeat the previous genetic classification, we found the accuracies only decrease slightly (82.8%→78.9% for all datasets and 98.9→93.0% for all datasets but WTCCC_Affy500k, refer to Table 1), which implies that these gene may actually be important in each gene cluster. The selected gene symbols in the 14 gene clusters are provided in Supplementary Data File 4.

3.5 Functional analysis of the gene clusters

Identifying key functions in the 14 gene clusters can help biologists to better understand the etiology of hypertension. The most intuitive approach is to look for the gene ontology (GO, Barrell *et al*., 2009) of the shared (major) genes (Supplementary Table 6). Among the 14 major genes, *LTBP1*, *CGREF1*, *TMEM16A* and *JPH1* are associated with calcium ion binding/transport, *AKAP12* and *LPHN3* are involved in G-protein-coupled receptor signaling pathways, *NINJ2* is related to nervous system development and is associated with an increased risk of stroke (Ikram *et al*., 2009), *TNIK* is involved in Wnt receptor signaling pathway, nervous system development and response to stress, and *GRID2* is involved in glutamate signaling pathway, neuroactive ligand-receptor interaction and

long-term depression. The remaining five major genes seem less hypertension-related. *DNAH9* is responsible for ATP and nucleotide bindings and microtubule motor activity. Ectopic expression of *C8orf72* (also known as *FAM110B*) proteins impaired cell cycle progression in G1 phase (Hauge *et al*., 2007). *PMS1* is responsible for ATP and DNA bindings and is involved in repair of DNA mismatches. *PKIB* encodes a protein which is a member of the cAMP-dependent protein kinase inhibitor family. *NRXN1* functions in the vertebrate nervous system as cell adhesion molecules and receptors.

We also analyzed the functions, processes, and pathways of the other genes involved in the 14 gene clusters. We compared the GO information of the 14 gene clusters with that of 14 randomly generated, equal-sized, ones. Supplementary Table 7 lists the mechanisms in the 14 clusters that were significantly more abundant ($p < 0.05$) than those in the random sets. The majority of the listed mechanisms are known to highly related to hypertension, such as magnesium ion binding, calcium ion transport, central nervous system development, metabolic process and sodium ion transport. In addition, we utilized the genes selected from the gene expression data to find influential mechanisms in each individual gene cluster. We list in Supplementary Table 8 the influential pathways in which multiple genes in a cluster were involved whereas in Supplementary Table 9 we present the abundant functions, processes, and pathways in the individual gene clusters. From these tables, we found that gene clusters anchored by the major genes *CGREF1*, *PMS1* and *TNIK* all had multiple genes in several cardiomyopathy-related pathways. Among the involved genes, a hypertension-candidate gene, *CACNA1C*, interacted with all three major genes suggesting its important role in cardiomyopathy-related hypertension. Moreover, multiple genes in gene clusters anchored by *C8orf72*, *PMS1* and *NRXN1* were found to involve in the metabolic pathways implying its influential role in these clusters. In addition, gene clusters anchored by *TMEM16A*, *LPHN3* and *GRID2* involved multiple neurotransmitter receptor genes contributing to the neuroactive ligand-receptor interaction pathway suggesting its possible link with hypertension. Finally, pathways such as axon guidance and Alzheimer's disease also frequently involved in several gene clusters and their relationships with hypertension warrant further investigations.

Another attempt to identify influential disease mechanisms was to compare biomedical profiles among patient groups identified by the 14 gene clusters in Taiwan_Afft500k. We found that patients identified by the *PKIB* gene cluster had lower blood sodium levels ($P = 0.038$) than other patients, which coincided with situations in which the cluster consisted of more sodium channel activity genes than the others. In addition, the patients identified by the *JPH1* gene cluster had lower blood potassium levels ($P = 0.026$) than

other patients in the dataset, and this cluster happened to include more outward rectifier potassium channel activity genes than the others. Moreover, the *LPHN3* gene cluster that identified six patients whose urine potassium levels were higher than other patients in the dataset ($P = 0.026$) contained abundant genes with clustering of voltage-gated potassium channels, potassium channel regulator activity, and calcium-activated potassium channel activity.

3.6 Existence of alternative component causes in a genetic causal pie

We demonstrated that the 14 gene clusters resulting from the training datasets were reproducible in test datasets of different platforms through gene match (see Figs. 3 and 4). Although the different sets of 14 clusters obtained from different platforms involved the same genes and mostly formed distinct gene-subject clusters, there was no direct proof that these gene clusters were the same across platforms. For example, in the *DNAH9* gene cluster, we did not know whether the SNP pairs obtained from different platforms could identify the same group of HT patients.

To address this issue, we tested the 14 gene clusters on 46 HT patients for whom there existed both Affymetrix500k and Affymetrix100k data and on 200 HT patients who had both Affymetrix500k and Illumina550k data (not used in the test datasets). As shown in the upper panel of Supplementary Figure 4, the HT patients identified in the Affymetrix500k data differed from those identified in the Affymetrix100k data. Similarly, in the lower panel, the HT patients identified in the Affymetrix500k data differed from those identified in the Illumina550k data. Both results indicated that, even with the same gene pairs, a gene cluster that consisted of different SNP pairs (used by different platforms) identified different groups of patients. The above results seemed to suggest that a genetic causal pie that involves multiple genes can involve different genetic variants (i.e., SNPs). If all the influential SNPs were genotyped, however, then the percentage of HT patients identified by the gene clusters could be increased.

4 DISCUSSION

We have developed a gene cluster construction algorithm for complex diseases, starting from finding influential gene pairs followed by grouping them into gene clusters. Most of the gene clusters consisted of multiple gene pairs that identified a similar group of patients and thus were highly susceptible to link with certain disease mechanisms. On an application to young-onset hypertension, our algorithm successfully constructed multiple reproducible gene clusters; each identified a distinct group of subjects. Furthermore, the

algorithm exhibited robustness (> 90% of the top 15 gene clusters remained unchanged) to criterion change and to small (≤ 10%) sample size change.

The constructed gene clusters resemble Rothman's causal pie model in that each gene cluster can be regarded as a causal pie with each risky combinatory genotype set of a gene pair in the cluster representing a component cause in the pie (a slice of the pie). And for each subject, the probability of ascertaining a disease increase dramatically as sufficient number of risky combinatory genotypes is carried. Multiple gene clusters, each of which identified a distinct group of subjects, imply multiple causal pies (disease mechanisms) for young-onset hypertension and may help to identify disease subtypes. Such a multi-causal pie-multi-component model provides an explanation of why conventional GWAS approaches in which all hypertensives were considered as a single group in comparison with the normotensives usually resulted in few significant genetic markers with poor reproducibility.

In this work, we presented the 14 large gene clusters constructed by our algorithm. These gene clusters were reproducible not only in Taiwanese and Caucasian populations but also across multiple genotyping platforms. In addition, they identified 19.3% of HT patients in all the datasets and 41.8% if the WTCCC_Affy500k was excluded for lack of biomedical profiles. Although 11.3% (with or without WTCCC_Affy500k) of NC subjects also carried risky combinatory genotypes in the gene clusters, they carried less risky combinatory genotypes than HT patients. After applying a suitable threshold to the number of risky combinatory genotypes in each gene cluster, we can further discriminate the HT patients from the NC subjects with an accuracy of 82.8% (sensitivity = 0.68 and specificity = 0.93) for all datasets and with an accuracy of 98.9% (sensitivity = 0.98 and specificity = 1.0) if the WTCCC_Affy500k was excluded.

The number of genes involved in the 14 gene clusters ranged from a few hundred to a few thousand. The meaning of such large number of genes is not clear. However, since multiple genes with similar functions are often involved in a given cluster and the influence of SNP variation is usually small, it is likely that it takes accumulative effects of multiple genes of the same functions and those of multiple pathways to lead to development of hypertension. Canalization (Waddington 1959), which measures the ability of a population to produce the same phenotype regardless of variability in its environment or genotype, may provide an explanation for the large number of genetic causal pies as well as the large number of component causes in a pie. Indeed, canalization values are high in most biological systems, implying that evolutionary forces select for traits that promote canalization which would ensure a normal blood pressure. Therefore, minor/moderate genetic or environment perturbations may not substantively impact

biological systems. They need to be accumulated in some particular fashion and amount so as to cause malfunction in a biological system.

We have also listed in Supplementary Table 6-9 some important functions, processes and pathways that are related to the 14 gene clusters. According to our gene-gene interaction model, the mechanisms that are related to a shared gene (Supplementary Table 6) should have strong associations with those that abundantly appeared in the gene cluster (Supplementary Table 8). To name a few, in the *LTBP1* gene cluster, calcium ion binding/transport regulated by *LTBP1* may be related to metabolism of lipids and lipoproteins that is attributed by two other genes in the cluster; in the *TMEM16A* gene cluster, calcium and chloride ion binding regulated by *TMEM16A* may be associated with purine metabolism, neuroactive ligand-receptor interaction and signaling by GPCR; in the *LPHN3* gene cluster, G-protein-coupled receptor activity regulated by *LPHN3* may interact with axon guidance, diabetes pathways and neuroactive ligand-receptor interaction; in the *TNIK* gene cluster, the stress response regulated by *TNIK* may be linked to arrhythmogenic right ventricular cardiomyopathy, dilated cardiomyopathy, hypertrophic cardiomyopathy and vascular smooth muscle contraction. However, further gene mapping endeavors are needed to depict detailed mechanisms in these gene clusters.

Owing to the difficulty of incorporating environment-gene with gene-gene interactions, in the present study, we only focused on constructing genetic causal pies for young-onset hypertension. With the genetic causal pies identified, we can then combine environmental factors, for example using the algorithm proposed by Hoffmann (Hoffmann *et al*., 2006) or by Liao (Liao and Lee 2010), to further explore the interactions between genetic and environmental factors and thus to better depict the hypertension etiology.

ACKNOWLEDGEMENTS

We would like to thank three anonymous reviewers for thoughtful critiques and comments that have resulted in a much improvement of this manuscript. In addition, the authors thank Kuang-Mao Chiang for data organization and Jhih-Siang Lai for technical support. The Framingham Heart Study project is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (N01 HC25195). Our present study also made use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113.

Funding: This work was supported by the National Science Council (NSC 99-3112-B-001-005, NSC 98-3112-B-001-009 and NSC 97-3112-B-001-011) and by the Academia Sinica Postdoctoral Training Grant of Taiwan.

Conflict of Interest: none declared.

REFERENCES

Altshuler,D. et al. (2008) Genetic mapping in human disease. *Science*, 322, 881–888.

Armitage, P. et al. (2002) *Statistical Methods in Medical Research 4th edn*, Blackwell Science, Chichester.

Barrell, D. et al. (2009) The GOA database in 2009--an integrated Gene Ontology Annotation resource. *Nucleic Acids Res.*, 37, D396–403.

Chen, S.H. et al. (2008) A support vector machine approach for detecting gene-gene interaction. *Genet. Epidemiol*. 32, 152–167.

Cook, N.R. et al. (2004) Tree and spline based association analysis of genexgene interaction models for ischemic stroke. *Stat. Med*., 23, 1439–1453.

Cordell, H.J. (2009) Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.*, 10, 392–404.

Dawber, T.R. et al. (1951) Epidemiological approaches to heart disease: the Framingham Study. *Am J Public Health Nations Health*, 41, 279–281.

Fisher, R. (1918) The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edin*., 52, 399–433.

Frazer,K.A. et al. (2009) Human genetic variation and its contribution to complex traits. *Nature Rev. Genet.*, 10, 241–251.

Hardy,J. and Singleton,A. (2009) Genomewide association studies and human disease. *N. Engl. J. Med.*, 360, 1759–1768.

Hauge, H. et al. (2007) Characterization of the FAM110 gene family. *Genomics*, 90, 14–27.

Hoffmann, K. et al. (2006) Estimating the proportion of disease due to classes of sufficient causes. *Am. J. Epidemiol.*, 163, 76–83.

Ikram, M.A. et al. (2009) Genomewide association studies of stroke. *N. Engl. J. Med.*, 360, 1718–1728.

Kang, G. et al. (2008) An entropy-based approach for testing genetic epistasis underlying complex diseases. *J. Theor. Biol.*, 250, 362–374.

Kooperberg, C. and Ruczinski, I. (2005) Identifying interacting SNPs using Monte Carlo logic regression. *Genet. Epidemiol.*, 28, 157–170.

Kohara, K. et al. (2008) Identification of hypertension-susceptibility genes and pathways by a systemic multiple candidate gene approach: the millennium genome project for hypertension. *Hypertens. Res.*, 31, 203–212.

Liao, S.F. and Lee, W.C. (2010) Weighing the causal pies in case-control studies. *Ann. Epidemiol.*, 20,568–573.

Lynn, K.S. et al. (2009) A neural network model for constructing endophenotypes of common complex diseases: an application to male young-onset hypertension microarray data. *Bioinformatics*, 25, 981–988.

Marchini, J. et al. (2005)Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genet.*, 37, 413–417.

McCarthy,M.I. et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Rev. Genet.*, 9, 356–369.

Mongeau, J.G. (1987) Heredity and blood pressure in humans: an overview. *Pediatr. Nephrol.*, 1, 69-75

Moore, J.H. and Williams, S.M. (2009) Epistasis and its implications for personal genetics. *Am. J. Hum. Genet.*, 85, 309–320.

Moore, J.H. et al. (2010) Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, 26, 445–455.

Motsinger-Reif, A. A. et al. (2008) Comparison of approaches for machine-learning optimization of neural networks for detecting gene–gene interactions in genetic epidemiology. *Genet. Epidemiol*., 32, 325–340.

Musani, S.K. et al. (2007) Detection of gene x gene interactions in genome-wide association studies of human population data. *Hum. Hered.*, 63, 67–84.

Neale, B.M. and Sham, P.C. (2004) The future of association studies: gene-based analysis and replication. *Am. J. Hum. Genet*., 75, 353–362.

Nelson, M.R. et al. (2001) A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res.*, 11, 458–470.

Neuman, R.J. and Rice, J.P. (1992) Two-locus models of disease. *Genet. Epidemiol.* 9: 347–365.

Nunkesser, R. et al. (2007) Detecting high-order interactions of single nucleotide polymorphisms using genetic programming. *Bioinformatics*, 23, 3280–3288.

Pan, W.H. et al. (2000) Linkage analysis with candidate genes: the Taiwan young-onset hypertension genetic study. *Hum. Genet.* 107, 210–215.

Pan, W.H. et al. (2006) Han Chinese cell and genome bank in Taiwan: purpose, design and ethical considerations. *Hum. Hered.* 61, 27–30.

Pan, W.H. et al. (2006) Using endophenotypes for pathway clusters to map complex disease genes. *Genet. Epidemiol.*, 30, 143–154.

Purcell, S. et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet*., 81, 559–575.

Ritchie M.D. et al. (2001) Multifactordimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.*, 69, 138–147.

Riva, A. and Kohane, I.S. (2002) SNPper: retrieval and analysis of human SNPs. *Bioinformatics*, 18, 1681-1685.

Rothman, K.J. (1976) Causes. *Am. J. Epidemiol*., 104, 587–592.

Rothman, K.J. (2005) Causation and causal inference in epidemiology. *Am. J. Public Health*, 95, Suppl 1:S144–150.

Sing, C.F. et al. (2003) Genes, environment, and cardiovascular disease. *Arterioscler. Thromb. Vasc. Biol.*, 23, 1190–1196.

Waddington, C.H. (1959) Canalization of development and genetic assimilation of acquired characters. *Nature*, 183, 1654–1655.

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447, 661–678.

Zerba, K.E. et al. (1996) Genotype-environment interaction: apolipoprotein E (Apo E) gene effects and age as an index of time and spatial context in the human. *Genetics*, 143, 463–478.

Zhang, Y. and Liu, J. S. (2007) Bayesian inference of epistatic interactions in case–control studies. *Nature Genet.*, 39, 1167–1173.

Supplementary Materials for:

Construction of Gene Clusters Resembling Genetic Causal Mechanisms for Common Complex Disease with an Application to Young-Onset Hypertension

Ke-Shiuan Lynn[1], Chen-Hua Lu[1], Han-Yin Yang[1], Wen-Lian Hsu[1] and Wen-Harn Pan[2,*]

[1]Institute of Information Science, Academia Sinica, Taipei, Taiwan
[2] Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan

Overview of the Supplementary Materials:

Supplementary Methods

1. Inclusion criteria HT patients and for NC subjects
2. Validation algorithm of the 14 gene clusters using gene expression data
3. Cluster visualization
4. Robustness evaluation of the gene cluster construction algorithm
5. Probability of a false positive gene cluster

Supplementary Figures

1. Detailed aspects of subjects identified by the 14 gene clusters in the Taiwanese test datasets
2. Demonstration of a gene cluster construction process
3. Percentage of overlaps in the top-*n* gene clusters (*n* = 5, 10, 15, 20, 30, 50, 100) with respect to changes in the proportion of the case population used in the analysis
4. Percentage of overlaps in the top-*n* gene clusters (*n* = 5, 10, 15, 20, 30, 50, 100) with respect to changes in sample size in FHS_Affy500k
5. Box plots of the combined gene expression values for four subject groups in the fourteen gene clusters.

Supplementary Tables

1. Allele frequencies of the SNP *rs16854417* (*SLC9A9*) in different datasets
2. Numbers of overlapping genes between the 14 gene clusters
3. Percentage of HT patients who carried risky combinatory genotypes in each gene cluster among all patients in each dataset and percentage of NC subjects who carried risky combinatory genotypes in each gene cluster among all NC subjects in each dataset
4. Number of classification errors in the two training datasets evaluated at the validation sets of a five-fold validation procedure
5. Number of component causes in each gene cluster after LD reduction and after redundancy removal
6. Selected gene ontology of the 14 major genes
7. Mechanisms that were observed more/less frequently ($P < 0.05$) in the 14 gene clusters than in the 14 randomly generated, equal-sized, gene sets
8. Influential pathways in the individual gene cluster
9. Abundant functions, processes, and pathways in the individual gene clusters

Supplementary Methods

1. Inclusion criteria HT patients and for NC subjects

- HT patients: Subjects who satisfied all the listed criteria were included.
  - FHS
    1. Hypertension diagnosis: (i) Subject systolic blood pressure (SBP) ≥ 140 mmHg or diastolic blood pressure (DBP) ≥ 90 mmHg in at least two of four measurements (no subject was selected in the gen3 cohort under this criterion because only one BP measurement was available). (ii) Subject diagnosed as an HT patient at any examination.
    2. Body mass index (BMI) ≤ 35, blood sugar < 126 (not checked for gen3 cohort), no hard congenital heart disease (hard CHD), and not a diabetes patient.
    3. Age at onset, 20–50 years.
  - Taiwan YOH study
    1. Hypertension diagnosis: (i) Subject SBP ≥ 140 mmHg or DBP ≥ 90 mmHg in at least two consecutive visits in 2 months. (ii) Subject taking at least one anti-hypertensive medication. (iii) Not a secondary HT patient.
    2. BMI ≤ 35, blood sugar < 126, and hemoglobin A1c (HbAIC) < 7 (not an obese or diabetes patient).
    3. Age 20–50 years.
- NC subjects: Subjects who satisfied all the listed criteria were included.
  - FHS
    1. BP: Subject with a normal mean BP and has no more than one measure of SBP/DBP exceeding 120/80 mmHg.
    2. BMI ≤ 35.
    3. Blood sugar < 126 (not checked for gen3 cohort in FHS), and no hard CHD.
  - THCCG
    1. BP: Same criteria as used for FHS.
    2. BMI ≤ 35.
    3. HbAIC < 7 (not a diabetes patient).

2. Probability of a false positive gene cluster

We attempt to calculate the probability of a false positive gene cluster that contains k non-LD SNP pairs and identifies m subjects in a population of n subjects. Assumining that the k non-LD SNP pairs according identify m_1, m_2, m_3, …, m_k subjects with all m_i, i = 1, 2, 3, …, k subjects being subsets of the m subjects,

For a SNP pair that identifies m_1 subjects in a population of n, there are $C(n, m_1)$ possible combinations where $C(n, m_1) = n!/(m_1!(n-m_1)!)$ and therefore the probability for the first SNP pair in the gene cluster to identify the m_1 subjects by chance is $1/C(n, m_1) \le m_1/n \le m/n$. For the gene cluster of k non-LD SNP pairs to be formed by chance, the probability is $1/(C(n, m_1)*C(n, m_2)* C(n, m_3)*…* C(n, m_k)) \le (m_1* m_2*m_3* …*m_k)/n^k \le (m/n)^k$. In conclusion, the probability of a false positive gene cluster that contains k non-LD SNP pairs and identifies m subjects in a population of n subjects bounded above by $(m/n)^k$.

3. Validation algorithm of the 14 gene clusters using gene expression data

Let *hyp_mtx* be the hypertensive part and *nor_mtx* be the normotensive part of the data in which rows represent different genes and columns represent different subjects.

For each gene cluster

Let PID_{hyp_risky} be the indices of HT patients in *hyp_mtx* who carried risky combinatory genotypes and $PID_{hyp\_norisky}$ be those of HT patients who did NOT carry risky combinatory genotypes.

Let PID_{nor_risky} be the indices of NC subjects in *nor_mtx* who carried risky combinatory genotypes and $PID_{nor\_norisky}$ be those of NC subjects who did NOT carry risky combinatory genotypes.

Let GID_{shared} be the index of shared gene.

Set *hyp_vec* = *hyp_mtx*(GID_{shared},:).

Set *nor_vec* = *nor_mtx*(GID_{shared},:).

p1 = t-test(*hyp_vec*(PID_{hyp_risky}), *hyp_vec*($PID_{hyp\_norisky}$));

p2 = t-test(*hyp_vec*(PID_{hyp_risky}), *nor_vec*(PID_{nor_risky}));

p3 = t-test(*hyp_vec*(PID_{hyp_risky}), *nor_vec*($PID_{nor\_norisky}$));

$p_t = p1 + p2 + p3$;

Set *minp* =*tmpp*= p_t;

While $minp \geqq tmpp$

For each gene $i \neq GID_{shared}$ in the gene list

hyp_tmp = *hyp_vec*+*hyp_mtx*(*i*,:);

nor_tmp = *nor_vec*+*nor_mtx*(*i*,:);

$pp1_i$ = t-test(*hyp_tmp*(PID_{hyp_risky}), *hyp_tmp*($PID_{hyp\_norisky}$));

$pp2_i$ = t-test(*hyp_tmp*(PID_{hyp_risky}), *nor_tmp*(PID_{nor_risky}));

$pp3_i$ = t-test(*hyp_tmp*(PID_{hyp_risky}), *nor_tmp*($PID_{nor\_norisky}$));

hyp_tmp = *hyp_vec*-*hyp_mtx*(*i*,:);

nor_tmp = *nor_vec*-*nor_mtx*(*i*,:);

$np1_i$ = t-test(*hyp_tmp*(PID_{hyp_risky}), *hyp_tmp*($PID_{hyp\_norisky}$));

$np2_i$ = t-test(*hyp_tmp*(PID_{hyp_risky}), *nor_tmp*(PID_{nor_risky}));

$np3_i$ = t-test(*hyp_tmp*(PID_{hyp_risky}), *nor_tmp*($PID_{nor\_norisky}$));

p_i = min($pp1_i + pp2_i + pp3_i$, $np1_i + np2_i + np3_i$);

If $p_i < tmpp$

$tmpp = p_i$;

best_id=i;

End If

End For

If *tmpp* < *minp*

Minp = tmpp;

hyp_vec = hyp_vec+hyp_mtx(*best_id*,:);

nor_vec = nor_vec+nor_mtx(*best_id*,:);

End If

End While

End For

4. Cluster visualization

We developed the following steps to generate gene-subject cluster plots for the demonstration of Rothman's genetic causal pies:

Step 1 Construct a binary matrix for each dataset in which each row in the matrix represents a SNP pair and a non-zero element indicates a subject carrying a risky combinatory genotype associated with the SNP pair.

Step 2 Reorder rows in the matrix such that SNP pairs with a shared gene are grouped together.

Step 3 Sort the resulting gene clusters by their size in descending order.

Step 4 Merge rows (SNP pairs) of the same gene pairs in a gene cluster into a single row using the "OR" operator if a similar group of subjects is identified.

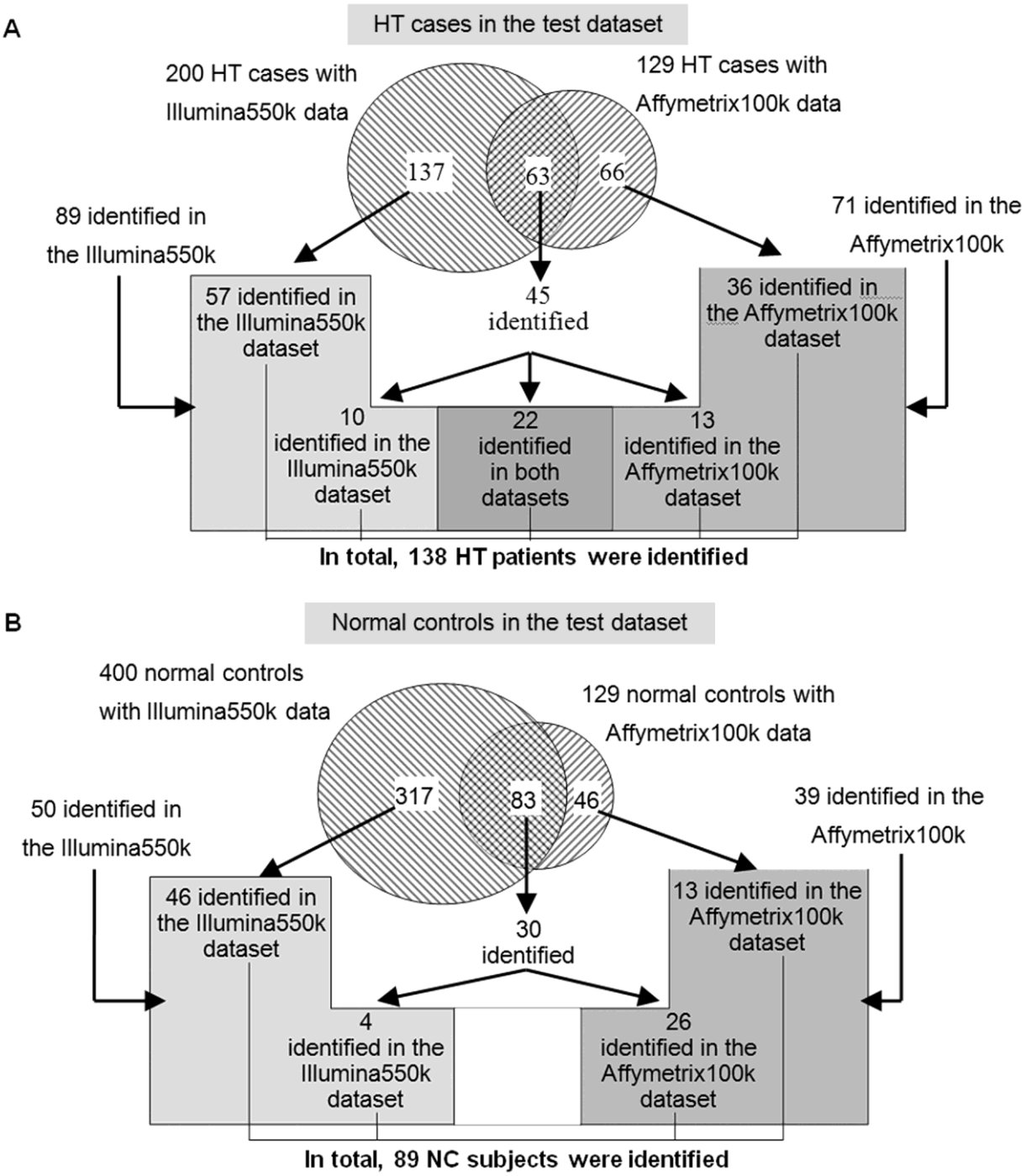
Step 5 Starting from the largest gene clusters, group columns that represent subjects carrying risky combinatory genotypes in the gene cluster.

5. Robustness evaluation of the gene cluster construction algorithm

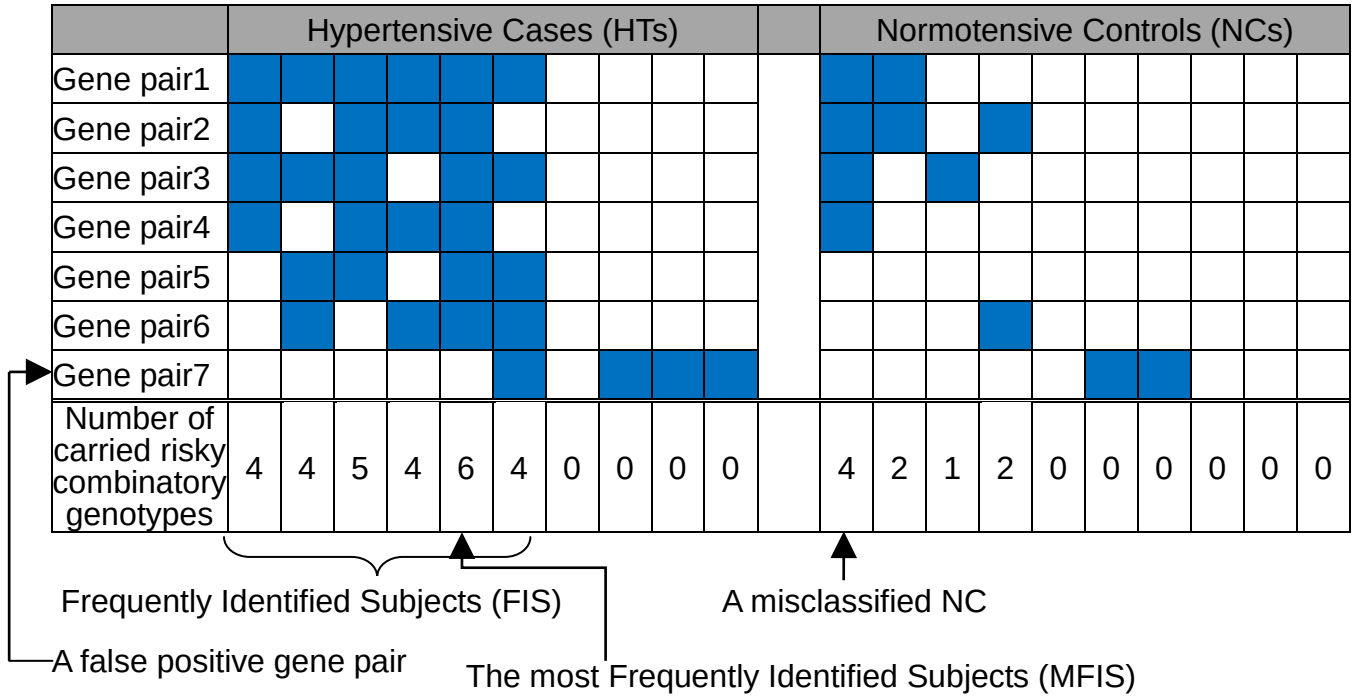
We tested the robustness of our gene cluster construction algorithm to small changes in criteria of the risky combinatory genotype. Because the clusters were selected at the first stage, we changed the criteria used at this stage. Apart from the original setting of 2.0% of the case population, we first tested the algorithm after increasing the setting to 2.5% and then decreasing it to 1.8% and 1.5%. These proportions were selected so as to change the numbers of cases in the two training datasets. That is, compared with the original 2% of the case population (7 and 4 cases in FHS_Affy500k and Taiwan_Affy500k, respectively), the 2.5%, 1.8% and 1.5% of case population corresponded to (8 and 5), (6 and 4) and (5 and 3) cases in the two datasets. To compute the similarity of the two lists of gene clusters, we first ranked the gene clusters in descending order based on the cluster size and then compared the shared genes corresponding to the top n gene clusters (n = 5, 10, 15, 20, 30, 50, 100). The top n gene clusters of the two lists were said to have 100% overlap if their corresponding shared genes were the same (regardless of the ranking order).

We also tested the robustness of the developed gene cluster construction algorithm to changes in sample size. Because the Taiwan_Affy500k was already small in terms of NC sample size (184 NC subjects), and further reducing it could introduce a considerable amount of false-positive SNP pairs, we only reduced the sample size of FHS_Affy500k to 90%, 70%, and 50% of its original size in our experiments. For each size, three sub-datasets were constructed, each of which was randomly drawn from the original dataset. The similarity between the gene clusters computed from the reduced sub-dataset and those computed from the original dataset was evaluated via the same procedure as that used to evaluate the effect of criterion changes.

Supplementary Figures

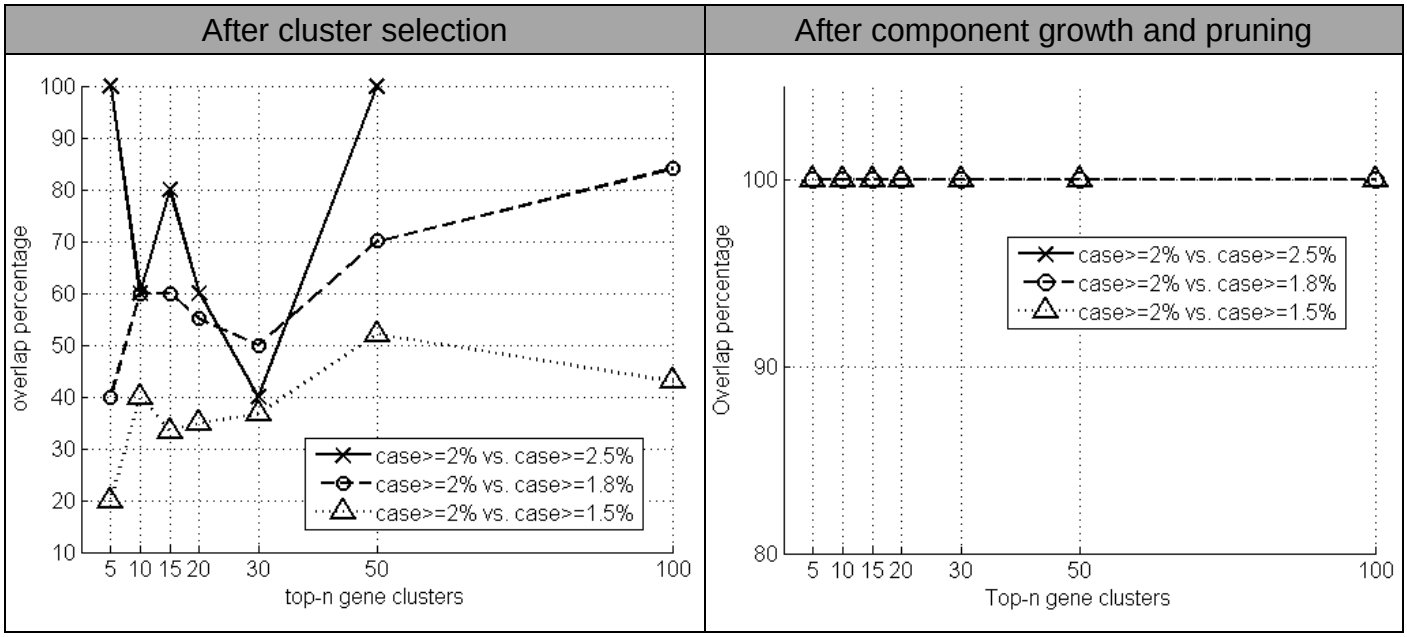


Supplementary Fig. 1. Detailed aspects of subjects identified by the 14 gene clusters in the Taiwanese test datasets; The Taiwanese test datasets include Taiwan_Affy100k and Taiwan_Illu550k. The upper panel (A) is for HT cases and the lower panel (B) is for NC subjects.



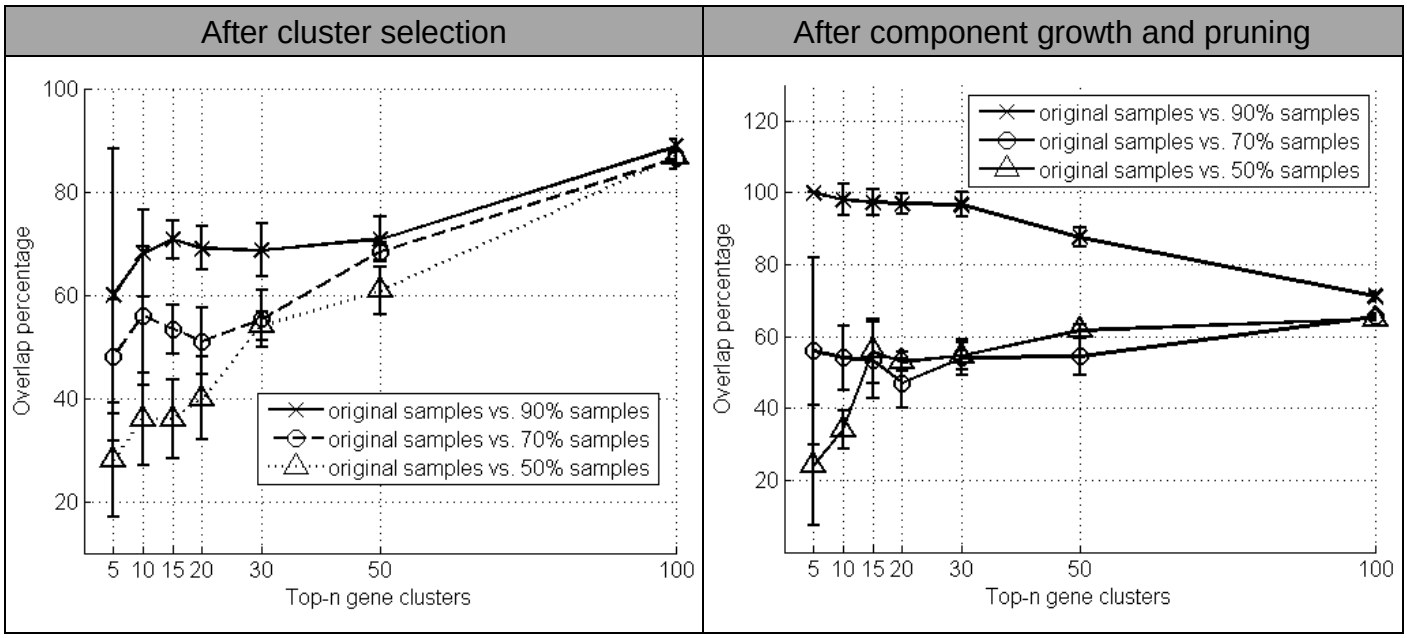
Supplementary Fig. 2. Demonstration of a gene cluster construction process

The top-6 hypertensive patients are frequently identified subjects (FISs) whereas the fifth one is the most frequently identified subjects (MFIS). On the other hand, there are 7 gene pairs in the above gene cluster. Of the 7 gene pairs, the gene pair 1, which identifies 6 patients, is the most effective, whereas the gene pair 7 identifies a different group of patients and thus will be removed from the gene cluster. According to the number of risky combinatory genotypes carried by the FISs, the sufficient number of component causes (risky combinatory genotypes) is 4. Under such a threshold, the first normotensive control is misclassified.



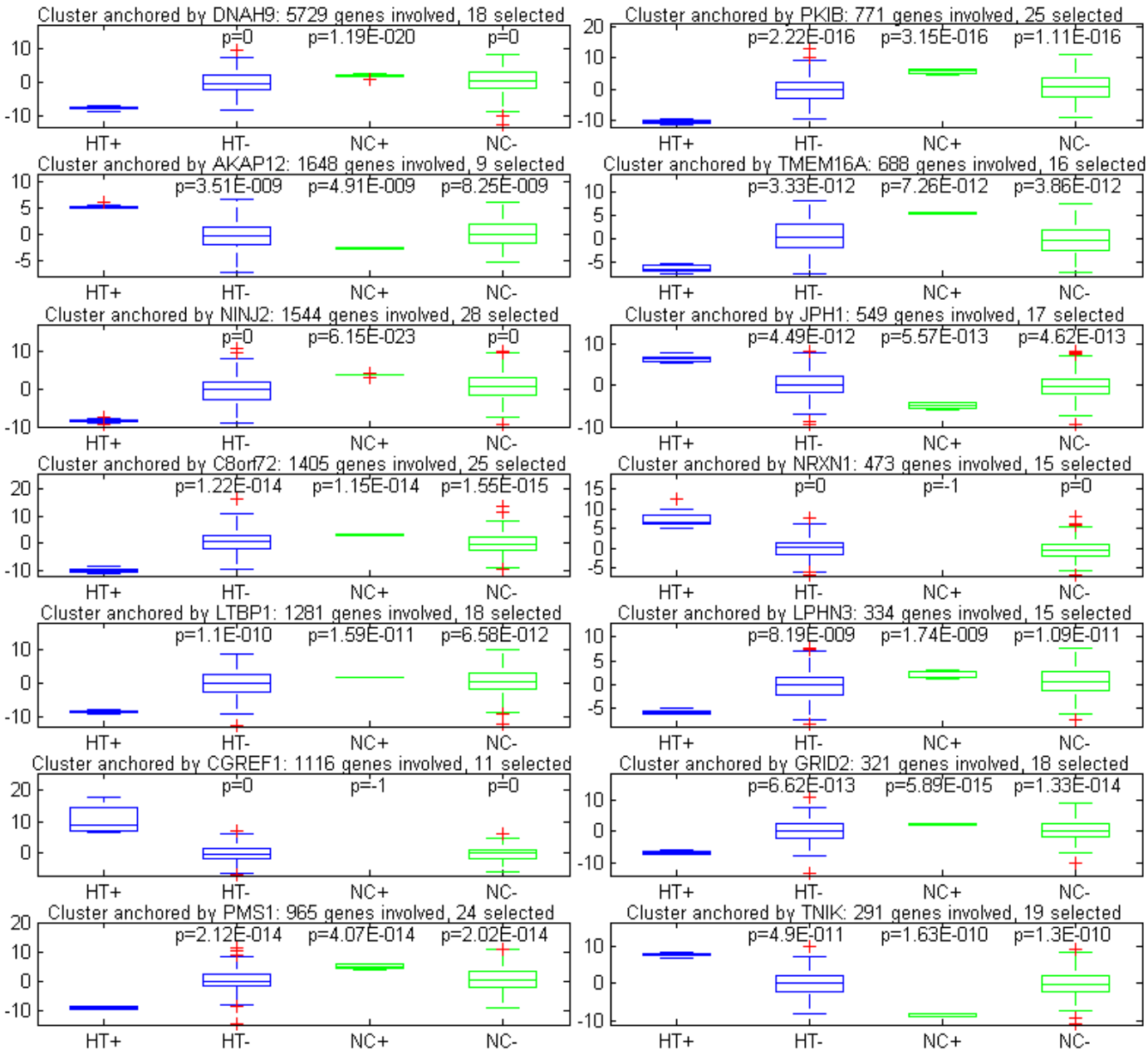
Supplementary Fig. 3. Percentage of overlaps in the top-*n* gene clusters (*n* = 5, 10, 15, 20, 30, 50, 100) with respect to changes in the proportion of the case population used in the analysis

The left panel shows results after the cluster selection stage of our cluster construction algorithm whereas the right panel shows results after component growth and pruning stages. It should be noted that the solid data line was cut off at *n* = 50 because the criterion became too stringent and only 51 gene clusters were qualified.

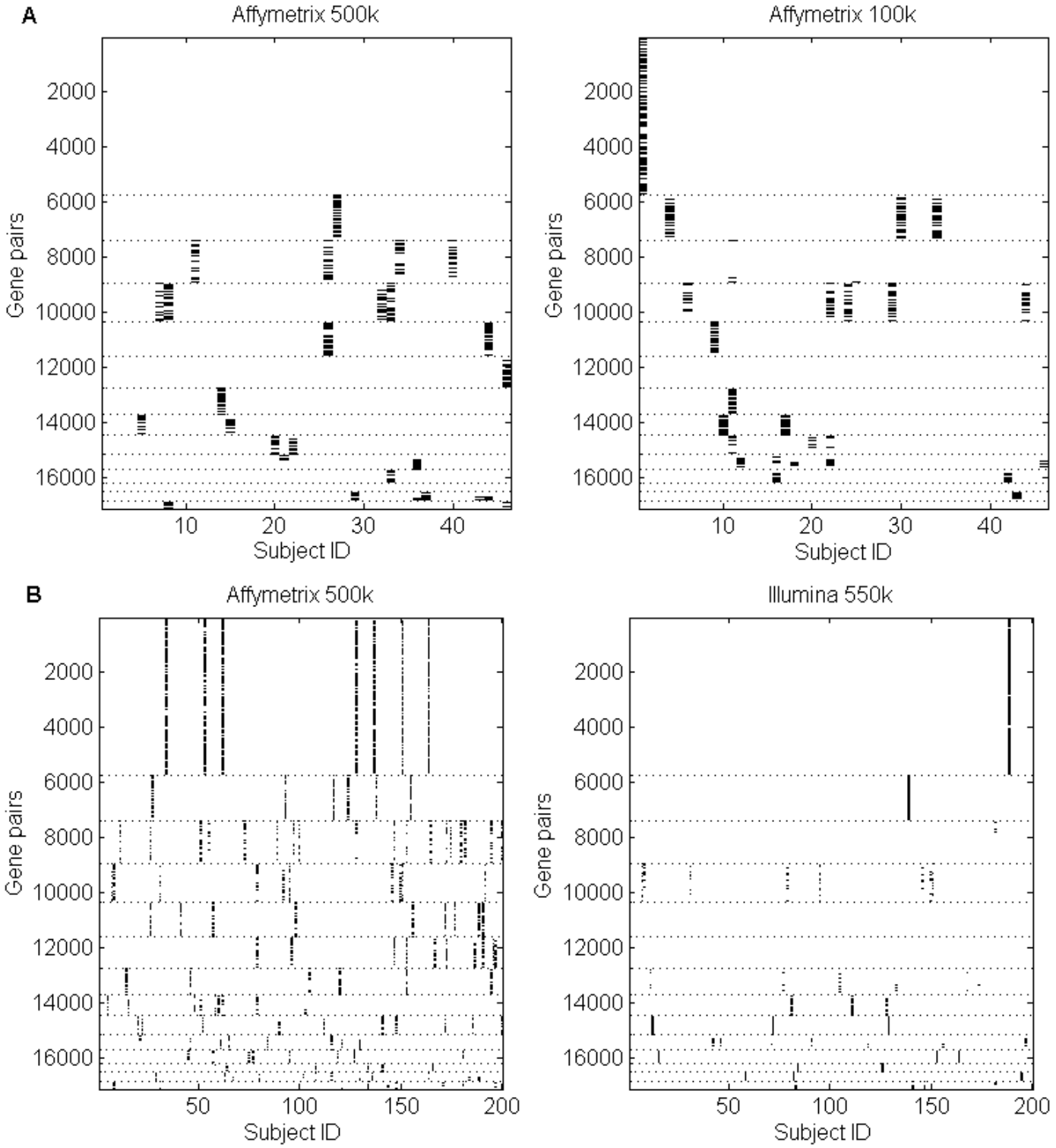


Supplementary Fig. 4. Percentage of overlaps in the top-*n* gene clusters (*n* = 5, 10, 15, 20, 30, 50, 100) with respect to changes in sample size in FHS_Affy500k

The left panel shows results after the cluster selection stage of our cluster construction algorithm whereas the right panel shows results after component growth and pruning stages.



Supplementary Fig. 5. Box plots of the combined gene expression values for four subject groups in the fourteen gene clusters. Of the four subject groups, HT+ represents hypertensives carrying risky combinatory genotypes, HT- represents hypertensives without carrying risky combinatory genotypes, NC+ denotes normotensives carrying risky combinatory genotypes and NC- denotes normotensives without carrying risky combinatory genotypes. The *p* value indicates the t-test result of the corresponding subject group with respective to the HT+ group (*p* = -1 indicates no subject in the corresponding group).



Supplementary Fig 6 Gene clusters consisting of the same gene pairs but different SNP pairs may identify different groups of patients

Overlapped patients in different genotyping platforms were used to test whether the same gene clusters detected from different platforms identify the same group of patients (i.e., whether the different SNP pairs selected from different platforms have LD): (A) 46 HT overlapped patients in both Affymetrix 500k and Affymetrix 100k data were tested; (B) 200 HT overlapped patients in both Affymetrix 500k and Illumina550k data were tested.

Supplementary Tables

Supplementary Table 1. Allele frequencies of the SNP *rs16854417* (*SLC9A9*) in different datasets

Ethnicity		Caucasian			Asian	
Datasets		HapMap-CEU	FHS_Affy500k	WTCCC_Affy500k	HapMap-HCB	Taiwan_Affy500k
Sample size		120	Case:305, Control:2881	Case:2001, Control:3004	90	Case:200, Control:184
Allele frequency	**CC**	**0%**	**total: 0.09% (case:0.98%, control:0%)**	**total: 0.04% (case:0.05%, control:0.03%)**	**0%**	**total: 1.3% (case:2.0%, control:0.54%)**
	CG	3.3%	total: 3.7% (case:3.28%, control:3.75%)	total: 2.29% (case:1.95%, control:2.73%)	22.2%	total: 11.7% (case:11.5%, control:11.96%)
	GG	96.7%	total: 94.4% (case:93.1%, control:94.1%)	total: 97.20% (case:97.35%, control:97.07%)	77.8%	total: 86.2% (case:85.0%, control:87.5%)

CEU: U.S. Utah residents with ancestry from northern and western Europe, HCB: Han Chinese in Beijing, China

Supplementary Table 2. Numbers of overlapping genes between the 14 gene clusters

	DNAH9	*AKAP12*	*NINJ2*	*C8orf72*	*LTBP1*	*CGREF1*	*PMS1*	*PKIB*	*TMEM16A*	*JPH1*	*NRXN1*	*LPHN3*	*GRID2*	*TNIK*
DNAH9	5729	403	222	258	260	184	183	128	125	84	88	72	64	44
AKAP12	403	1648	60	47	69	44	41	17	28	10	14	16	14	10
NINJ2	222	60	1544	63	58	24	48	37	27	36	10	6	15	8
C8orf72	258	47	63	1405	58	38	45	29	24	16	11	4	21	10
LTBP1	260	69	58	58	1281	31	28	26	17	16	15	8	8	7
CGREF1	184	44	24	38	31	1116	31	30	17	8	9	9	7	8
PMS1	183	41	48	45	28	31	965	15	23	16	2	4	6	4
PKIB	128	17	37	29	26	30	15	771	7	9	8	2	5	3
TMEM16A	125	28	27	24	17	17	23	7	688	11	8	3	4	10
JPH1	84	10	36	16	16	8	16	9	11	549	1	3	1	3
NRXN1	88	14	10	11	15	9	2	8	8	1	473	12	3	0
LPHN3	72	16	6	4	8	9	4	2	3	3	12	334	0	0
GRID2	64	14	15	21	8	7	6	5	4	1	3	0	321	2
TNIK	44	10	8	10	7	8	4	3	10	3	0	0	2	291

Supplementary Table 3. Percentage of HT patients who carried risky genotypes in each gene cluster among all patients in each dataset and percentage of NC subjects who carried risky genotypes in each gene cluster among all NC subjects in each dataset

No. of genes	Shared gene	HT carrying risky genotypes (%)					NC carrying risky genotypes (%)				
		Training		Test		Total	Training		Test		Total
		Cau	TW	Cau	TW	% (No.)	Cau	TW	Cau	TW	% (No.)
1	*SLC9A9*	0.98	2.00	0.05	N/A	0.32 (8)	0	0.54	0.03	N/A	0.03 (2)
2–55	Omitted due to insufficient gene pairs for evaluation of frequently identified subjects										
291	*TNIK*	1.97	4.50	0.55	4.14	1.33 (37)	0.76	4.35	0.83	0.45	0.87 (57)
321	*GRID2*	1.31	6.00	0.30	5.26	1.30 (36)	0.38	4.89	0.53	1.12	0.63 (41)
334	*LPHN3*	3.28	3.00	1.75	3.76	2.20 (61)	0.80	0.54	1.13	1.79	1.01 (66)
473	*NRXN1*	2.62	4.00	1.00	5.64	1.84 (51)	0.56	1.63	1.07	0.45	0.81 (53)
549	*JPH1*	2.62	4.00	0.15	7.52	1.41 (39)	0.62	5.43	0.40	1.79	0.74 (48)
688	*TMEM16A*	2.62	5.50	1.50	5.26	2.27 (63)	0.62	2.17	1.53	3.14	1.26 (82)
771	*PKIB*	2.30	4.50	0.45	4.51	1.33 (37)	0.45	7.07	0.80	0.90	0.83 (54)
965	*PMS1*	3.28	3.00	1.30	6.02	2.09 (58)	0.62	4.89	1.63	0.90	1.23 (80)
1,116	*CGREF1*	2.62	5.00	0.05	2.26	0.90 (25)	0.83	0.54	0.07	0.90	0.48 (31)
1,281	*LTBP1*	2.30	4.50	0.55	3.76	1.33 (37)	0.59	3.26	0.93	1.12	0.86 (56)
1,405	*C8orf72*	2.62	5.00	1.15	6.02	2.06 (57)	0.69	5.43	0.80	2.47	1.00 (65)
1,544	*NINJ2*	2.62	9.00	0.50	7.52	2.02 (56)	0.59	5.98	0.37	2.47	0.77 (50)
1,648	*AKAP12*	2.62	3.00	0.65	4.51	1.41 (39)	0.66	1.09	0.67	1.35	0.72 (47)
5,729	*DNAH9*	3.28	3.50	1.35	9.02	2.45 (68)	0.49	1.09	0.97	3.81	0.95 (62)

Cau: Caucasian, TW: Taiwanese

Supplementary Table 4 Number of classification errors in the two training datasets evaluated at the validation sets of a five-fold validation procedure

	FHS_Affy500k					Taiwan_Affy500k				
	Validation set1	Validation set2	Validation set3	Validation set4	Validation set5	Validation set1	Validation set2	Validation set3	Validation set4	Validation set5
DNAH9	0	0	0	0	1	0	0	0	0	1
AKAP12	1	0	0	0	0	0	1	0	0	0
NINJ2	0	0	0	0	1	0	0	0	0	1
C8orf72	0	1	0	0	0	0	2	1	0	2
LTBP1	0	0	0	0	1	0	0	1	0	0
CGREF1	1	0	0	0	0	1	0	0	0	0
PMS1	0	0	0	1	0	0	0	0	1	0
PKIB	0	0	1	0	0	1	0	0	0	0
TMEM16A	0	1	0	0	0	0	1	0	1	0
JPH1	0	0	1	0	0	0	1	0	0	0
NRXN1	0	0	0	1	0	0	0	0	0	1
LPHN3	0	0	0	1	1	0	0	0	2	0
GRID2	0	0	0	0	0	0	1	0	0	0
TNIK	0	1	0	0	0	1	0	0	0	0
Overall Classification accuracy	99.91%					98.96%				

Supplementary Table 5 Number of component causes in each gene cluster after LD reduction and after redundancy removal

	Number of component causes		
	After cluster construction	After LD reduction	After redundancy removal
DNAH9	5729	5521 (96.21%)	2427 (42.36%)
AKAP12	1648	1585 (96.18%)	624 (37.86%)
NINJ2	1544	1497 (96.96%)	578 (37.44%)
C8orf72	1405	1371 (97.58%)	524 (37.30%)
LTBP1	1281	1236 (96.49%)	453 (35.36%)
CGREF1	1116	1080 (96.77%)	410 (36.74%)
PMS1	965	943 (97.72%)	347 (35.96%)
PKIB	771	754 (97.80%)	265 (34.37%)
TMEM16A	688	666 (96.80%)	207 (30.09%)
JPH1	549	530 (96.54%)	170 (30.97%)
NRXN1	473	458 (96.83%)	114 (24.10%)
LPHN3	334	316 (94.61%)	82 (24.55%)
GRID2	321	310 (96.57%)	82 (25.55%)
TNIK	291	286 (98.28%)	68 (23.37%)

Supplementary Table 6. Selected gene ontology of the 14 major genes

Shared gene	Location	No. of genes	Selected gene ontology
DNAH9	17p12	5729	**Function**: ATP binding, microtubule motor activity
AKAP12	6q24-q25	1648	**Process**: G-protein-coupled receptor protein signaling pathway
NINJ2	12p13	1544	**Process**: nervous system development, tissue regeneration **Phenotype**: genome-wide association studies of stroke
C8orf72	8q12.1	1405	
LTBP1	2p22-p21	1281	**Function**: calcium ion binding, growth factor binding **Pathway**: TGF-beta signaling pathway
CGREF1	2p23.3	1116	**Function**: calcium ion binding **Process**: response to stress
PMS1	2q31-q33; 2q31.1	965	**Function**: ATP binding **Process**: DNA mismatch repair
PKIB	6q22.31	771	**Function**: cAMP-dependent protein kinase inhibitor activity **Process**: negative regulation of protein kinase activity
TMEM16A	Chromosome 5	688	**Function**: calcium ion binding, chloride ion binding
JPH1	8q21	549	**Function**: structural constituent of muscle **Process**: calcium ion transport into cytosol, regulation of ryanodine-sensitive calcium-release channel activity
NRXN1	2p16.3	473	**Function**: metal ion binding **Process**: axon guidance **Pathway**: cell adhesion molecules (CAMs) **Phenotype**: susceptibility to autism
LPHN3	4q13.1	334	**Function**: G-protein-coupled receptor activity, sugar binding **Process**: G-protein-coupled receptor protein signaling pathway
GRID2	4q22	321	**Function**: extracellular-glutamate-gated ion channel activity **Process**: glutamate signaling pathway **Pathway**: neuroactive ligand-receptor interaction, long-term depression
TNIK	3q26.2-q26.31	291	**Function**: protein serine/threonine kinase activity, small GTPase regulator activity **Process**: Wnt receptor signaling pathway, activation of JNKK activity, nervous system development, response to stress

Supplementary Table 7. Mechanisms that were observed more/less frequently ($p < 0.05$) in the 14 gene clusters than in the 14 randomly generated, equal-sized, gene sets

	Functions, processes and pathways	Gene ratios in the 14 gene clusters	Gene ratios in the 14 random sets	*p*-value
Functions	Acyltransferase activity	0.0049±0.0019	0.0074±0.0017	1.05×10^{-3}
	Calmodulin binding	**0.0136±0.0037**	**0.0055±0.0049**	**3.40×10^{-5}**
	Kinase activity	0.0039±0.0041	0.0072±0.0026	1.77×10^{-2}
	Magnesium ion binding	0.0061±0.0059	0.0099±0.0032	4.13×10^{-2}
	Motor activity	0.0052±0.0016	0.0020±0.0030	1.31×10^{-3}
	Olfactory receptor activity	0.0056±0.0082	0.0152±0.0120	2.04×10^{-2}
	Receptor binding	0.0127±0.0020	0.0073±0.0083	2.68×10^{-2}
Processes	**Activation of protein kinase C activity by G-protein-coupled receptor protein signaling pathway**	**0.0040±0.0020**	**0.0001±0.0020**	**1.90×10^{-5}**
	Axonogenesis	0.0062±0.0026	0.0016±0.0030	1.92×10^{-4}
	Calcium ion transport	0.0130±0.0034	0.0056±0.0057	3.05×10^{-4}
	Central nervous system development	**0.0129±0.0033**	**0.0037±0.0050**	**5.00×10^{-6}**
	Ion transport	**0.0477±0.0080**	**0.0281±0.0066**	**1.67×10^{-7}**
	Metabolic process	**0.0217±0.0048**	**0.0071±0.0072**	**1.00×10^{-6}**
	Mitosis	0.0028±0.0050	0.0105±0.0045	2.20×10^{-4}
	Protein homo-oligomerization	0.0049±0.0012	0.0022±0.0036	1.31×10^{-2}
	Response to drug	0.0080±0.0097	0.0149±0.0042	2.25×10^{-2}
	Sensory perception of smell	0.0081±0.0101	0.0196±0.0037	4.55×10^{-4}
	Sensory perception of sound	0.0085±0.0072	0.0014±0.0045	4.16×10^{-3}
	Sodium ion transport	0.0088±0.0041	0.0052±0.0049	4.24×10^{-2}
Pathways	Adherens junction	0.0070±0.0030	0.0020±0.0059	8.67×10^{-3}
	Cell adhesion molecules (CAMs)	0.0131±0.0047	0.0057±0.0059	1.10×10^{-3}
	Dilated cardiomyopathy	0.0083±0.0076	0.0027±0.0041	2.22×10^{-2}
	Huntington's disease	0.0046±0.0059	0.0088±0.0022	2.05×10^{-2}
	Hypertrophic cardiomyopathy (HCM)	0.0080±0.0076	0.0029±0.0040	3.64×10^{-2}
	Natural killer cell mediated cytotoxicity	0.0027±0.0064	0.0069±0.0031	3.72×10^{-2}
	Phosphatidylinositol signaling system	0.0066±0.0056	0.0023±0.0051	4.24×10^{-2}
	Signaling in Immune system	0.0099±0.0034	0.0132±0.0038	2.14×10^{-2}
	Signaling by NGF	0.0144±0.0038	0.0093±0.0051	6.00×10^{-3}
	Tight junction	0.0113±0.0053	0.0057±0.0068	2.32×10^{-2}

Note: In the above list, we only present mechanisms that are sufficient abundant (at least 6 of the 14 gene clusters had to contain ≥ 0.5% of all genes associated with a particular mechanism) in the 14 gene clusters.

Supplementary Table 8. Influential pathways in the individual gene cluster

Major gene	No. of Genes	Identified Patients	Influential Pathways
DNAH9	18	15	Axon guidance(2), Hemostasis(2)
AKAP12	9	8	
NINJ2	28	16	Axon guidance(2), ErbB signaling pathway(2), Focal adhesion(2), Regulation of actin cytoskeleton(3)
C8orf72	25	10	Alzheimer's disease(2), Calcium signaling pathway(2), Metabolic pathways(2)
LTBP1	18	9	Metabolism of lipids and lipoproteins(2)
CGREF1	11	4	Arrhythmogenic right ventricular cardiomyopathy(2), Axon guidance(2), Calcium signaling pathway(2), Cardiac muscle contraction(2), Dilated cardiomyopathy(2), Glutamatergic synapse(2), Hemostasis(2), Hypertrophic cardiomyopathy(2)
PMS1	24	11	Arrhythmogenic right ventricular cardiomyopathy(2), Axon guidance(2), Dilated cardiomyopathy(2), Hypertrophic cardiomyopathy(2), Metabolic pathways(3)
PKIB	25	11	Axon guidance(4)
TMEM16A	16	13	Neuroactive ligand-receptor interaction(2), Purine metabolism(3), Signaling by GPCR(2)
JPH1	17	13	
NRXN1	15	10	Axon guidance(2), Metabolic pathways(2)
LPHN3	15	9	Axon guidance(3), Diabetes pathways(2), Neuroactive ligand-receptor interaction(3)
GRID2	18	14	Axon guidance(2), Neuroactive ligand-receptor interaction(2), Synaptic Transmission(2)
TNIK	19	8	Alzheimer's disease(2), Arrhythmogenic right ventricular cardiomyopathy(2), Axon guidance(3), Dilated cardiomyopathy(2), Hypertrophic cardiomyopathy(2), Vascular smooth muscle contraction(2)

Note: Numbers in the parenthesis indicate the number of genes involved in the corresponding pathway.

Supplementary Table 9 Abundant functions, processes, and pathways in the individual gene clusters

Shared gene	No. of genes	Abundant mechanisms
DNAH9	5729	**Process**: positive regulation of insulin secretion++, positive regulation of stress-activated MAPK cascade++ **Pathway**: ubiquinone and other terpenoid-quinone biosynthesis++
AKAP12	1648	**Function**: glucuronosyltransferase activity++, copper ion binding++, sodium ion binding++ **Process**: positive regulation of cholesterol storage++

		Pathway: starch and sucrose metabolism+
NINJ2	1544	**Function**: G-protein-coupled receptor activity*, long-chain fatty acid-CoA ligase activity++, folic acid binding++, nucleoside: sodium symporter activity++, bile acid:sodium symporter activity++, lipid transporter activity++, very long-chain fatty acid-CoA ligase activity++ **Process**: homocysteine metabolic process++, energy reserve metabolic process++, folic acid and derivative metabolic process++, G-protein signaling, coupled to cGMP nucleotide second messenger++ **Pathway**: PPAR signaling pathway++, glyoxylate and dicarboxylate metabolism++, muscle contraction++
C8orf72	1405	**Function**: G-protein-coupled photoreceptor activity++ **Process**: T cell receptor signaling pathway++, positive regulation of T cell differentiation++, negative regulation of G-protein-coupled receptor protein signaling pathway++, regulation of T cell receptor signaling pathway++ **Pathway**: metabolism of lipids and lipoproteins*
LTBP1	1281	**Function**: SH3 domain binding** **Process**: G-protein-coupled receptor protein signaling pathway*, positive regulation of inflammatory response++, glycogen metabolic process++, gluconeogenesis++, positive regulation of systemic arterial blood pressure++, folic acid and derivative biosynthetic process++, G-protein signaling, coupled to cGMP nucleotide second messenger++, glycerol transport++, glycerol metabolic process, response to fatty acid++ **Pathway**: signaling by VEGF++
CGREF1	1116	**Function**: magnesium ion binding*, voltage-gated sodium channel activity++ **Process**: muscle contraction*, Wnt receptor signaling pathway++, response to glucocorticoid stimulus++, positive regulation of fatty acid oxidation++, positive regulation of potassium ion transport++ **Pathway**: galactose metabolism++
PMS1	965	**Function**: insulin binding++, G-protein-coupled receptor binding++, sodium:dicarboxylate symporter activity++, insulin-like growth factor I binding++, insulin-like growth factor receptor binding++ **Process**: regulation of G-protein-coupled receptor protein signaling pathway++, insulin receptor signaling pathway++, negative regulation of insulin receptor signaling pathway++, blood vessel maturation++, negative regulation of glucose import++, insulin-like growth factor receptor signaling pathway++ **Pathway**: diabetes pathways**, metabolic pathways*, signaling by Wnt++
PKIB	891	**Function**: sugar binding*, calcium channel regulator activity++, triglyceride lipase activity++ **Process**: regulation of fatty acid oxidation++, detection of calcium ion++, negative regulation of inflammatory response++, JAK-STAT cascade++, low-density lipoprotein particle remodeling++, positive regulation of lipid storage++ **Pathway**: fructose and mannose metabolism++
TMEM16A	771	**Process**: regulation of muscle contraction++, cholesterol esterification++, negative regulation of smooth muscle cell proliferation++, negative regulation of blood coagulation++, lipopolysaccharide biosynthetic process++ **Pathway**: signaling by insulin receptor++, lipid digestion, mobilization, and transport++

JPH1	688	**Function**: calcium ion binding*, extracellular-glycine-gated chloride channel activity++, calcium: sodium antiporter activity++ **Process**: inflammatory response**, release of sequestered calcium ion into cytosol++, response to glucose stimulus++, patterning of blood vessels++, regulation of calcium ion transport++ **Pathway**: calcium signaling pathway*, glycerophospholipid metabolism++
NRXN1	549	**Function**: voltage-gated calcium channel activity**, sodium:phosphate symporter activity++, calcium-release channel activity++ **Process**: ventricular cardiac muscle cell differentiation++, regulation of cardiac muscle contraction by regulation of the release of sequestered calcium ion++ **Pathway**: arrhythmogenic right ventricular cardiomyopathy (ARVC)*, cardiac muscle contraction*, signaling by Rho GTPases*
LPHN3	473	**Function**: chloride channel activity**, voltage-gated chloride channel activity++, voltage-gated chloride channel activity++ **Process**: heart development**, carbohydrate metabolic process**, lipid metabolic process*, glucose metabolic process++, circadian rhythm++, response to calcium ion++, blood circulation++ **Pathway**: Wnt signaling pathway**, fatty acid metabolism++
GRID2	334	**Function**: phospholipid binding**, lipid binding* **Process**: lipid catabolic process**, nervous system development*, potassium ion transport*, negative regulation of calcium ion transport via voltage-gated calcium channel activity++, cholesterol homeostasis++ **Pathway**: vascular smooth muscle contraction**, T cell receptor signaling pathway**, Jak-STAT signaling pathway**
TNIK	291	**Function**: voltage-gated calcium channel activity**, lipid binding**, voltage-gated ion channel activity++, cholesterol monooxygenase (side-chain-cleaving) activity++, inward rectifier potassium channel activity++, large conductance calcium-activated potassium channel activity++ **Process**: negative regulation of lipid catabolic process++, negative regulation of Wnt receptor signaling pathway++ **Pathway**: dilated cardiomyopathy**, hypertrophic cardiomyopathy (HCM)**, arrhythmogenic right ventricular cardiomyopathy (ARVC)*, cardiac muscle contraction*, T cell receptor signaling pathway*, MAPK signaling pathway++

Note: Three symbols indicate different levels of influence: "**" indicates that the ratio of associated genes is *at least 0.01* in the cluster and is *much higher (≥ mean + 2 SD)* than in the other gene clusters; "*" is similar to "**" except that the ratio is only *slightly higher (≥ mean + 1.5 SD)* than in the other gene clusters; "++" denotes that the ratio of associated genes is *between 0.002 and 0.01* in the cluster and is *much higher (≥ mean + 2 SD)* than in the other gene clusters.