Quantitative mapping of genetic similarity in human heritable diseases by shared mutations

Huiying Zhao[1], Yuedong Yang[2*], Matthew Mort[3], David N. Cooper[3], Yaoqi Zhou[2*]

[1.] Institute of Health and Biomedical Innovation, Queensland University of Technology, Queensland 4222, Australia; [2.] Institute for Glycomics and School of Information and Communication Technology, Griffith University, Gold Coast, QLD 4222, Australia; [3.] Institute of Medical Genetics, Cardiff University, Heath Park, Cardiff CF14 4XN, UK.

Corresponding author: yuedong.yang@griffith.edu.au; yaoqi.zhou@griffith.edu.au

Abstract

Many genetic diseases exhibit considerable epidemiological comorbidity and common symptoms, which provokes debate about the extent of their etiological overlap. The relationships between inherited diseases have often been studied in relation to the common genes or pathways involved. However, the sharing of disease-associated genes is not automatically indicative of similar molecular mechanisms because different mutations may affect different functional aspects of genes/proteins leading to unrelated pathological conditions. The rapid growth in the number of known disease-causing mutations in the Human Gene Mutation Database (HGMD) has allowed us to characterise disease genetic similarity (or disease-disease relationships) by ascertaining the extent to which identical genetic mutations are shared between diseases. Using this approach, we show that 41.6% of disease pairs in all possible pairs (42,083) exhibit a significantly sharing of mutations (P-value<0.05). These mutation-related disease pairs are in agreement with heritability-based disease-disease relations in 48 neurological and psychiatric disease pairs (Spearman's correlation coefficient=0.50; P-value=3.4×10^{-5}), and share over-expressed genes significantly more often

than unrelated disease pairs (1.5 to 1.8-fold higher; P-value $\leq 1.6\times10^{-4}$). We further demonstrate the utility of mutation-related disease pairs in predicting novel mutations and improving the identification of individuals susceptible to Crohn's disease. Moreover, the mutation-based disease network concurs closely with that based on phenotypes. In all these applications, disease-disease relations inferred from shared mutations perform significantly better than those derived from shared genes. This work highlights the potential for inferring disease-disease relationships at the mutational level.

Introduction

The co-occurrence of apparently distinct diseases is frequently observed as clinical phenotypic overlaps between different genetic diseases. For example, an increased incidence of Attention Deficit Hyperactivity Disorder (ADHD) has been found in families with bipolar disorder (Faraone et al. 2012). Patients with Friedreich ataxia appear to be at increased risk of developing type 2 diabetes mellitus (Ristow 2004). Type I diabetes and celiac disease often co-exist in populations because of seven common associated [HUIYING: what do you mean by 'associated'? GWAS?] loci (Smyth et al. 2008). Inflammatory bowel disease is known to be associated with an increased incidence of psychological disorders such as depression [HUIYING: major depressive disorder?] (Szigethy et al. 2015). Such co-occurrences are being increasingly discovered between disease pairs through the statistical analysis of the growing number of large electronic clinical records (Lee et al. 2008; Park et al. 2009; Blair et al. 2013).

The genetic mechanisms underlying disease co-occurrences are being uncovered by means of increasingly powerful sequencing technologies. In particular, Genome-wide association studies (GWAS) are capable of revealing common risk variants that transcend traditional diagnostic boundaries. Examples include high correlations in genetics between migraine and stroke (Debette et al. 2015), and across psychiatric diseases such as schizophrenia,

major depressive disorder and bipolar disorder (Cross-Disorder Group of the Psychiatric Genomics Consortium et al. 2013). Causative mutations associated with multiple diseases have also been discovered through sequencing (Carvill et al. 2013). For example, *SCN2A* gene mutations were found to be associated not only with self-limited autosomal dominant syndrome of benign familial neonatal-infantile seizures but also epileptic encephalopathies (Liao et al. 2010; Carvill et al. 2013). In similar vein, *CACNA1A* gene mutations have been implicated in a range of neurological conditions (Jouvenceau et al. 2001; Damaj et al. 2015; Subramanian et al. 2015).

Obviously, shared genetic architecture across different diseases is suggestive of possible common genetic aetiologies. Disease interactions inferred from shared genes have often improved our understanding of the molecular mechanisms underlying inherited disease (Davis and Chawla 2011; Lechner et al. 2012). Moreover, shared genes have led directly to the identification of pan-cancer driver genes (Melamed et al. 2015) and the discovery of common genetic mechanisms of four neuropsychiatric disorders: autism spectrum disorder (ASD), epileptic encephalopathy, intellectual disability and schizophrenia (Li et al. 2016).

However, sharing disease-causing genes does not automatically indicate the identity of the underlying molecular mechanisms because genes (and their protein products) have multiple functions and different mutations may impact different functions and cause different diseases. For example, many mutations in the *PRKAR1A* gene cause Carney complex due to decay of mRNA [HUIYING: do you mean nonsense-mediated mRNA decay due to frameshift and truncating variants?] and the absence of protein translation (Kirschner et al. 2000) whereas another mutation in exon 11 (c.1101C→T) was found to cause acrodysostosis by truncating the last 14 amino acids in the coded cAMP-binding protein domain B protein due to the introduction of a stop codon (Linglart et al. 2011).

Meanwhile, the same genetic mutation can cause different heritable diseases. For example, the c.1932delC mutation in the *HPS1* gene has been implicated in both Gitelman syndrome (Maki et al. 2004) and Hermansky–Pudlak syndrome (Wei et al. 2009). Similarly, the c.727-1G>T mutation in the *KAL1* gene was found to be relevant to both diastrophic dysplasia (Barbosa et al. 2011) and renal agenesis (Tsai and Gill 2006), whilst another *de novo* mutation (c.1267C>T) in the same gene has been reported to be associated with both epilepsy and ASD (Jiang et al. 2013). When identical mutations are shared by different diseases, it suggests that these diseases share molecular mechanisms at least in part, for example the activation or deactivation of a specific function associated with the gene involved. Thus, shared mutations, rather than shared disease-causing genes, might be a more specific and hence more accurate indicator of genetic similarity between diseases.

In this study, we identified disease relationships between inherited human diseases by ascertaining the number of shared genetic mutations. The disease-causing mutations were taken from the Human Gene Mutation Database (HGMD) (Stenson et al. 2017). The mutations relevant to the genetic diseases were added to the mutations associated with their 'sub-diseases' as defined by the Human Phenotype Ontology project (HPO) (Kohler et al. 2014). The resulting relationships identified between brain and psychiatric diseases are consistent with heritability-based studies, and related diseases identified by this mutation-based approach share a significantly greater number of over-expressed genes than unrelated ones. The mutation-based approach was shown to be useful for the discovery of novel disease-causing mutations and for the identification of patients with Crohn's disease. Moreover, the mutation-based disease network was shown to be consistent with phenotypic clusters. These results highlight the potential importance of using mutations to detect disease-disease relationships.

Results

Pairwise genetic similarity between human diseases

We measured pairwise genetic similarity between two diseases according to the statistical significance of the sharing of disease-causing mutations as compared to random chance. This method was termed DRAM (Disease RelAtionship by shared Mutations). For comparison, we also mapped the mutations to genes, and assessed the genetic similarity according to the number of overlapping genes (namely DRAG, Disease RelAtionship by shared Genes).

We assessed genetic disease-disease relations for a total of 42,083 disease pairs in 2,572 diseases successfully mapped to Human Phenotype Ontology (HPO, see Methods). By using a binomial test P-value cutoff of 0.05, DRAM detected 17,514 related disease pairs (41.6%) involving 2,100 diseases (81.6%). This is somewhat fewer than the 25,692 related pairs yielded by the gene-based approach (DRAG). The smaller number of pairs of related diseases identified by the DRAM was as expected because sharing exactly the same mutation is a much stricter requirement than simply sharing the same gene. However, the disease-disease relationship ascertained from DRAM only moderately correlates with that from DRAG in terms of its significance (Spearman's correlation coefficient (SPC) = 0.47, P-value is $< 2.6 \times 10^{-6}$). The correlation is however slightly better (SPC=0.52) for related disease pairs from both DRAG and DRAM. If the most significant 17,514 pairs are selected [HUIYING: Why the most significant? Isn't this the number of related disease pairs detected by DRAM?], the proportion of the pairs shared by both methods is only 60.7% (10,628/17,514), highlighting a significant difference between them in identifying related disease pairs.

Genetic correlation between brain disorders according to shared heritability and shared mutations

To assess our methods, we compared disease relationships based upon shared mutations (DRAM) with those based upon shared genes (DRAG) as well as those based on shared

heritability (Anttila et al. 2016) in common brain diseases. As shown in Supplementary Figure S2, the correlation between heritability-based and mutation-based disease-disease relationships (from DRAM) is higher (SPC=0.505, P-value=2.5×10^{-4}) than that from DRAG (0.343, P-value=7.7×10^{-3}). When an adjusted P-value cutoff of 1.0×10^{-3} (=0.05/48) was used, 22 and 12 out of 48 pairs were detected as related by DRAM and the heritability-based estimation, respectively. Of the 12 most significantly related pairs by DRAM, 8 pairs overlapped with those obtained by the heritability-based approach (P-value < 1.0×10^{-3}), which was a significant enrichment (P-value = 2.8×10^{-3}) over pairs not showing relationships according to DRAM [HUIYING: Is this what you wanted to say?]. Meanwhile, DRAM discovered 13 related pairs that were not revealed by the heritability-based approach as listed in Table S1.

Figure 1 depicts disease-disease relationships detected by both heritability and DRAM (in blue), by heritability only (in grey), and by DRAM only (in red) for 5 psychiatric diseases (in yellow) and 8 neurological diseases (in green). Six pairs of psychiatric diseases were connected based on either sharing of heritability (P-value threshold 0.05/48) or mutations (DRAM, P-value threshold 0.05/48). For psychiatric diseases, DRAM only missed one connection between schizophrenia (SCZ) and ASD from heritability-based approach while indicating a mutation-based connection between ADHD and ASD. For neurological diseases, all three pairs connected by sharing of heritability were also connected by mutations. DRAM, however, discovered four connections missed by the heritability-based approach. Those connections were Alzheimer disease and ischemic stroke, ischemic stroke and migraine, migraine with aura and ischemic stroke, and Parkinson disease and Alzheimer disease. More importantly, DRAM detected six connections between neurological and psychiatric diseases (P-value < 0.05/48) whereas the heritability-based approach only connected migraine with Major Depressive Disorder (MDD) and SCZ. As we shall see below in the Discussion, these disease relationships,

uncovered by DRAM but missed by the heritability-based approach, often share similar clinical phenotypes.

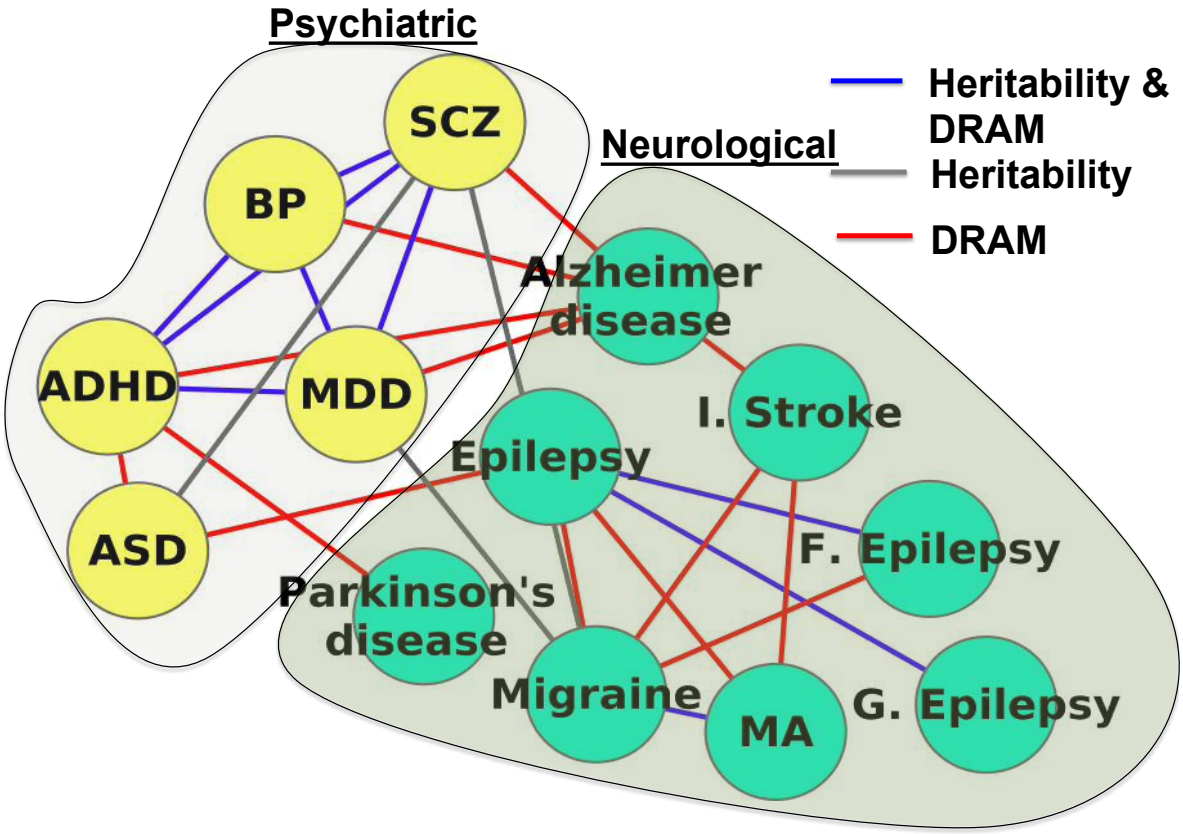

Figure 1. **Mutation-similarity network reveals relationships in brain diseases.** Disease-disease relationship networks in eight neurological diseases (green nodes) and five psychiatric diseases (yellow nodes), assessed through sharing mutations by DRAM and sharing heritability in a previous study. The red edges connect disease pairs showing significant genetic similarity (P-value < 0.05/48, 1.0×10^{-3}) by the DRAM method but not by heritability. The blue edges connect disease pairs showing genetic similarity by both methods (P-value < 0.05/48, 1.0×10^{-3}). The grey edges connect disease pairs showing genetic similarity by heritability (P-value < 0.05/48, 1.0×10^{-3}) but not by DRAM. F. Seizures stands for Focal seizures; G. Seizures - Generalized seizures; I. Stroke - Ischemic stroke; SCZ - Schizophrenia, BPD - Bipolar disorder, MDD - Major depressive disorder; ADHD - Attention deficit hyperactivity disorder; ASD - Autism spectrum disorder.

Related diseases identified by DRAM showing significant sharing of over-expressed genes

If two diseases are related, it is reasonable to suppose that these diseases might share over-expressed genes. We defined over-expressed genes using three different thresholds (5, 7 or 10-fold higher than the average levels in all tissues as described in the Methods section). As shown in Table 1, 82%, 89% and 80% pairs in 45 related disease pairs detected by DRAM were found to share significantly more over-expressed genes than under random expectation under three thresholds, respectively (all P-values<1.6×10^{-4} by binomial test). They were also significantly higher than 49%, 50% and 53% in 1363 unrelated disease pairs as defined by DRAM (P-values<2×10^{-4}).

Table 1. Disease pairs with significant sharing of highly expressed genes (P-value ≤ 0.05) are enriched with pairs significantly sharing mutations. The enrichment is significantly higher than for the disease pairs sharing of genes.

	Mutations			Genes	
Threshold	Significant [a]	Not Significant[b]	P-value[c]	Significant[d]	P-value[e]
5-fold	0.89 (40/45)	0.49 (673/1363)	1.9×10^{-8}	0.52 (30/58)	1.3×10^{-7}
7-fold	0.82 (37/45)	0.50 (677/1363)	7.9×10^{-6}	0.58 (33/58)	3.0×10^{-4}
10-fold	0.80 (36/45)	0.53 (723/1363)	1.6×10^{-4}	0.54 (31/58)	2.0×10^{-4}

[a] Number of disease pairs with significant sharing of highly expressed genes and sharing of mutations divided by the total number of disease pairs with significant sharing of mutations (45 pairs) only;
[b] Number of disease pairs with significant sharing of highly expressed genes but not significant sharing of mutations divided by total number of disease pairs without significant sharing of mutations (1363 pairs);
[c] Binominal test results on difference between the number of disease pairs showing significant sharing of highly expressed genes among disease pairs significantly sharing mutations (DRAM) and the number of disease pairs showing significant sharing of highly expressed genes among disease pairs not significantly sharing mutations;
[d] Number of disease pairs with significant sharing of highly expressed genes and sharing of genes in the total number of disease pairs with significant sharing of genes only (DRAG) (58 pairs);
[e] Binomial test results on difference between the number of disease pairs showing significant sharing of highly expressed genes among disease pairs significantly sharing mutations (DRAM)

and the number of disease pairs showing significant sharing of highly expressed genes among disease pairs significantly sharing of genes (DRAG).

Figure 2 shows disease-disease connections based on significant sharing of over-expressed genes and significant sharing of mutations. There are 62 disease pairs having significant sharing of over-expressed genes for any one of three thresholds, among which 45 pairs exhibit significant sharing of mutations. These diseases were color-coded according to the affected cell types (epithelial cells, nerve cells, Leydig cells, blood cells, embryonal cells, and muscle cells). As the Figure shows, disease-disease connections are not biased toward the same cell type. For example, the diseases affecting epithelial cells are shown sharing mutations with 31 diseases, 19 of which are from the diseases not affecting epithelial cells. Eleven diseases affecting Leydig cells do not share mutations with any other disease affecting Leydig cells. In the Figure, the node size of a disease is proportional to the number of diseases that share mutations significantly with the node. Prostate cancer has the most (7) connections in sharing significant mutations with other diseases, and it also has 24 connections in terms of sharing a significant number of over-expressed genes. Thyroid carcinoma, squamous cell carcinoma, neuroectodermal neoplasm, muscular dystrophy, and leukemia, share significantly mutations with six other diseases.

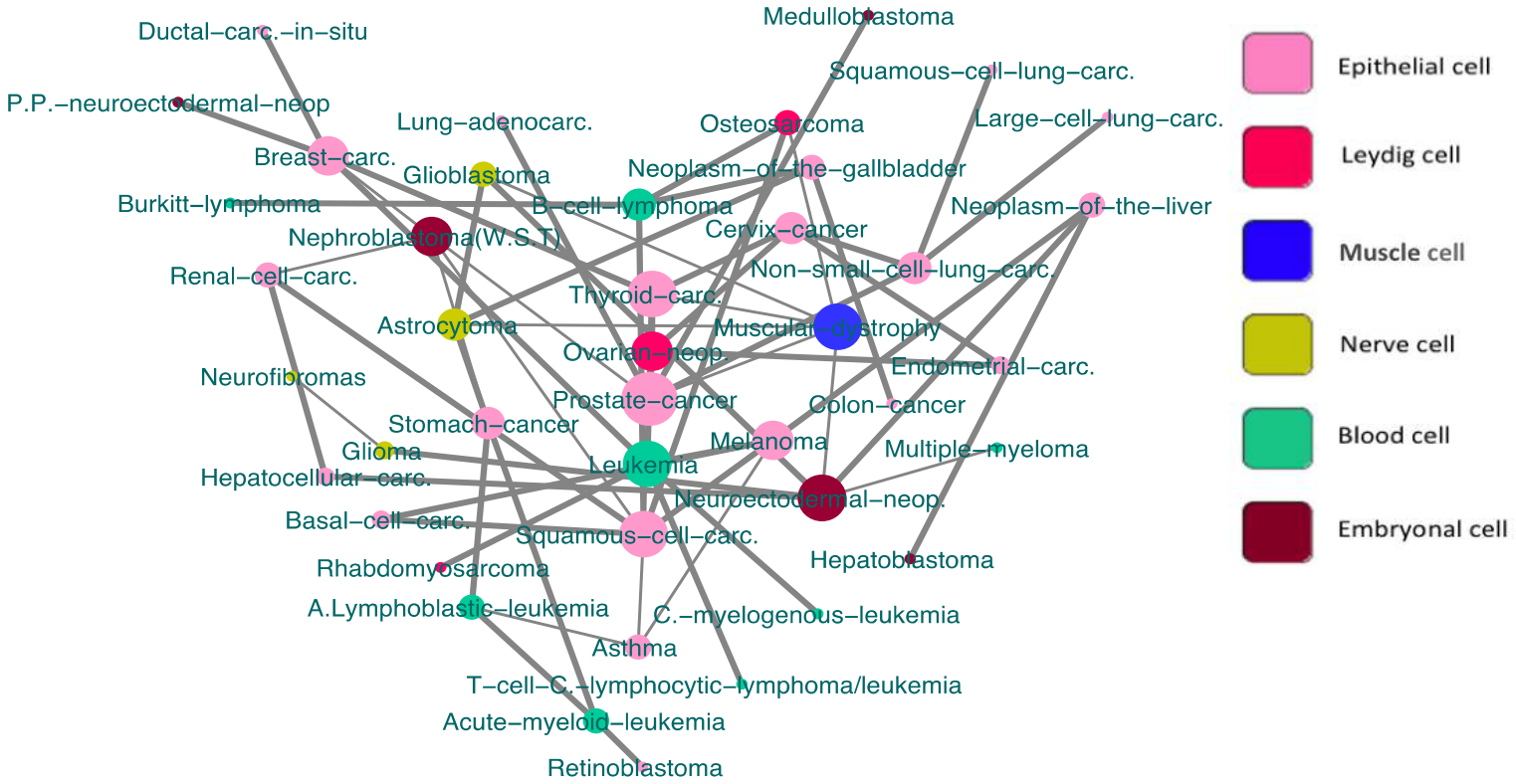

Figure 2. **Disease pairs with significant sharing of over-expressed genes are enriched with disease pairs with significant sharing of mutations**. Each node is a disease colour-coded in terms of its affected cell type. 62 disease pairs with significant sharing of over-expressed genes by 5-, 7-, or 10-fold are compared to the pairs with significant sharing of mutations. Larger nodes indicate a disease sharing mutations with more diseases. The thin edges represent disease pairs (only 17 pairs) not detected to be related (P-value > 0.05) by DRAM whereas the thick edges (45 pairs) represent disease pairs sharing over-expressed genes and mutations significantly at the same time. Different cell types affected by diseases were shown in different colours as labelled.

As a comparison, DRAG identified 58 disease pairs as being related based on the same threshold ($P \leq 0.05$) for defining disease-disease relationships. However, only 52%, 58% and 54% of disease pairs were found to significantly share over-expressed genes according to the three thresholds for defining over-expression. These percentages are significantly lower than those yielded by DRAM (Table 1), but not significantly (with all P-values >0.1) different from the percentages in unrelated pairs as ascertained by DRAG (Table S2).

Predicting novel mutations from mutations associated with relevant diseases

The ability to predict novel mutations by DRAM was compared to DRAG methods based either on the HGMD database or based on the DisGeneT database. As shown in Figure 3a, DRAM achieved the highest average MCC values for all diseases in ten-fold cross-validation. DRAG based on HGMD was found to exhibit higher performance than DRAG based on the DisGeneT database. This could be due to the fact that DisGeneT acquired gene-disease associations from multiple databases, which may be inconsistent with each other. Individually, DRAM achieved higher MCC values than DRAG (HGMD) in 248 out of 445 diseases (56%). It also achieved higher MCC values than DRAG (DisGeneT) in 313 (313/445, 70%) diseases. Application of the pairwise t-test indicated that these differences were significant (P-value=0.002 and 5.1×10^{-6}, respectively). This trend was confirmed by enrichment factors in top ranked predicted mutations (0.1%, 0.5% and 1% predictions) in Figure 3b. For example, DRAM represents an improvement over DRAG (HGMD) by 35% for enrichment factors for the top-ranked 0.1% predictions (Table 2). On average, DRAM has a sensitivity of 45.8% with a precision of 13.4% for all predicted mutations, as compared to 42.6% sensitivity and 12.6% precision by DRAG (HGMD), and 40.3% sensitivity and 13.1% precision by DRAG (DisGeneT). DRAM represents an improvement over DRAG (HGMD and DisGeneT) in terms of both sensitivity and precision. As a control, we examined the prediction by random sampling. Briefly, for each disease we randomly sampled the same number of predicted mutations by DRAM from HGMD,

and counted the number of mutations consistent with the known mutations relevant to the disease. As a result, the precision by random sampling was 0.483%, much lower than that by DRAM (13.4%).

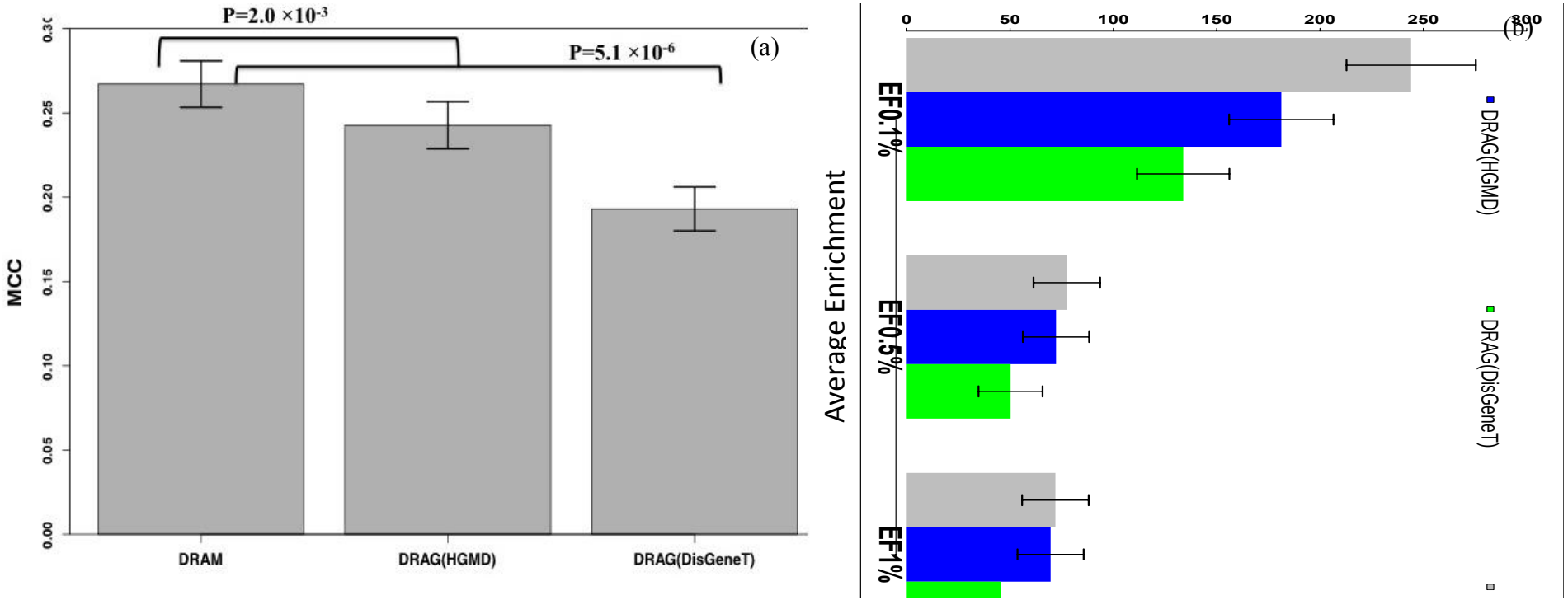

Figure 3. **Predicting disease-causing mutations by mutation (DRAM) or gene sharing (DRAG)**. Comparison of DRAM and DRAG based on two datasets (HGMD and DisGeneT) by (a) the average Matthews correlation coefficient (MCC) and (b) average enrichment factors in top-ranked 0.1%, 0.5%, and 1% predicted mutations (across 445 diseases). The error bars are based on the standard error of the [HUIYING: text missing?]mean

Table 2. Predicting disease-causing mutations by applying genetic similarities across diseases according to sharing of mutations (DRAM) or genes (DRAMG) onto HGMD and DisGeneT datasets. The result is based on 10-fold cross-validation.

	DRAM (HGMD)[a]	DRAG (DisGeneT)[b]	DRAG(HGMD)[c]
Sensitivity[d]	45.8%	40.3%	42.6%
Precision[e]	13.4%	13.1%	12.6%
0.1% EF[f]	244.1	133.8	181.3
0.5% EF[g]	77.5	50.2	72.2
1% EF[h]	71.9	45.7	69.6

[a] Genetic similarity defined by sharing mutations between diseases;
[b] Genetic similarity defined by sharing genes collected in DisGeneT database;
[c] Genetic similarity defined by sharing genes assigned by mutations;
[d] The average number of correctly predicted mutations divided by the number of mutations in the test set;
[e] The average number of correctly predicted mutations divided by the average number of predicted mutations;
[f] Enrichment factor in top 0.1%;
[g] Enrichment factor in top 0.5%;
[h] Enrichment factor in top 1%.

As an example, the Peripheral Arterial Disease (PAD) was associated with 1307 mutations according to HGMD. We randomly selected 1177 mutations (90%), by which 36 diseases were found to exhibit significant sharing of mutations with PAD. These 36 related diseases contained 1367 unique mutations that were scored according to genetic similarity between these diseases with PAD (described in Method). At the threshold with the maximum MCC (0.78), 114 mutations overlapped with the 130 remaining mutations, and 205 new mutations were predicted [HUIYING: what do you mean? Unclear!]. These new mutations were associated with 14 diseases as shown in Table S3. By comparison, DRAG (HGMD) and DRAG (DisGeneT) achieved MCC values of 0.63 and 0.22, respectively.

Identifying individuals susceptible to Crohn's disease

Although the overall precision in predicting novel mutations was only moderate, these predicted mutations are nevertheless useful for assisting disease diagnostics. To demonstrate this, we employed DRAM to identify individuals susceptible to Crohn's disease by utilizing both the number of known Crohn's disease-causing mutations listed in HGMD and the relation scores of mutations predicted from diseases relevant to Crohn's disease[HUIYING: Unclear!]

(see Methods). As shown in Figure 4, the gene-based approach, DRAG (HGMD) yielded essentially a random prediction with AUC=0.53, only slightly higher than the random selection (AUC=0.5). Relying on mutations from HGMD yields an AUC of 0.59. A recently reported

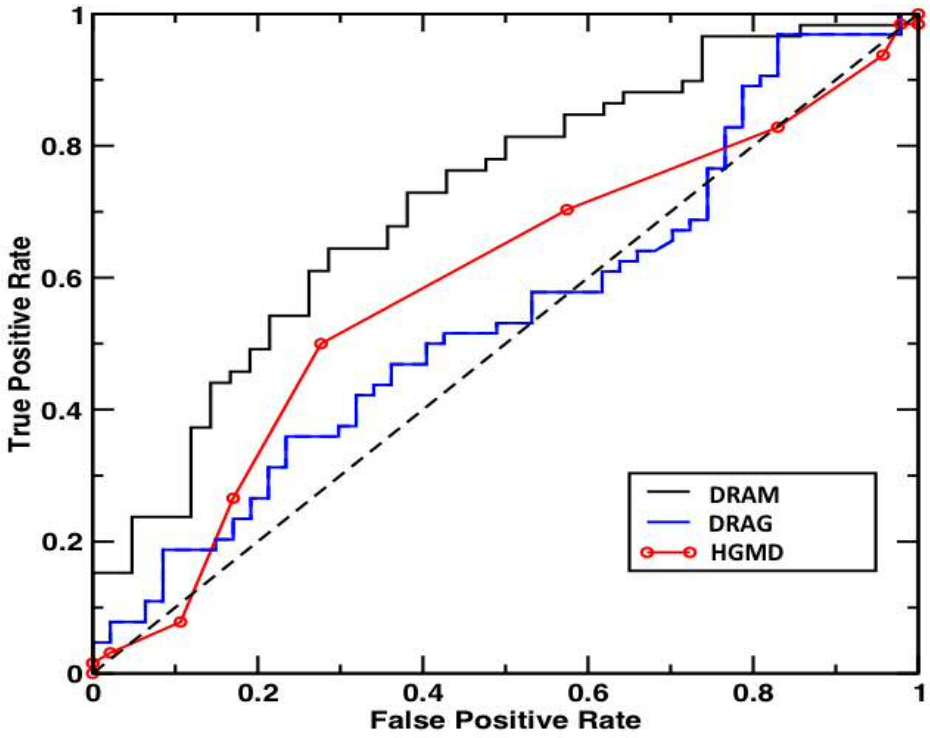

Figure 4. Receiver operating characteristic curves (ROC) for discriminating Crohn's disease affecting from healthy individuals [HUIYING: what do you mean? Unclear!]by counting the number of known Crohn's disease-related mutations in the HGMD directly, by adding predicted mutations from DRAG, and from DRAM, as labelled. Dashed line indicates random prediction.

"Ensemble" method (Giollo et al. 2017), a meta-analytical method, achieved AUC of 0.66 for the same datasets. All of these AUC values are lower than 0.69 achieved by our method DRAM. When measured by the highest MCC values, DRAG (HGMD), HGMD and DRAM have MCC values of 0.17, 0.22, and 0.35, respectively. Meanwhile, 64 individuals susceptible to Crohn's disease were predicted correctly (100%) by DRAM, as compared to 59 and 32 susceptible individuals predicted correctly by DRAG (92.1%) and HGMD (50%), respectively.

Disease network according to DRAM

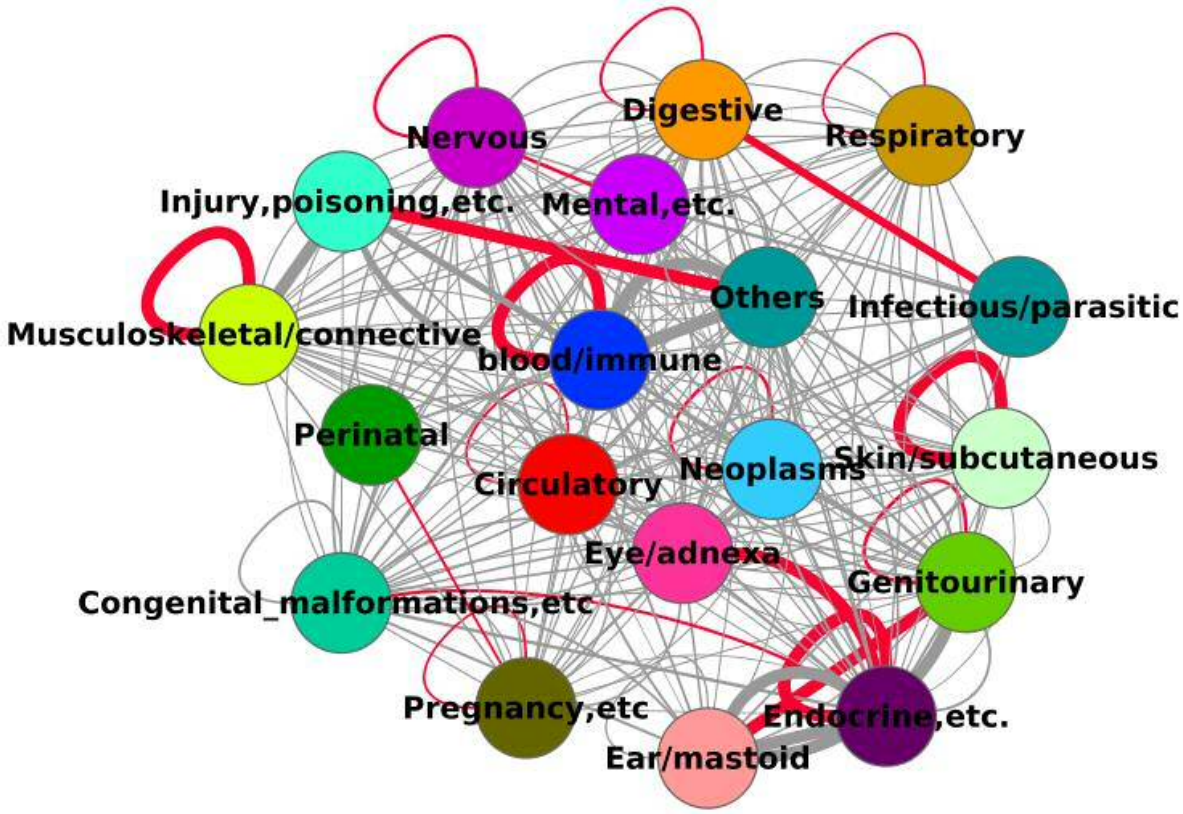

Figure 5. **The consistency between phenotypic and mutation-based clustering**. There are 296 mutation-sharing disease connections (in grey) between 19 phenotypic chapters classified by ICD-10. Each node represents one disease phenotypic chapter detailed in Supplementary Table S5. The width of the edge represents the median –log(P) [P: genetic similarity] between diseases within two chapters. The most significant association within a single chapter (in a loop) or between the chapters (in a line) is shown in red. 11 loops in red indicate the strongest association within the chapter itself.

To further examine the consistency between phenotype-based and mutation-based disease-disease relations, we mapped diseases from human phenotype ontology (HPO ID) to the International Classification of Diseases (ICD-10 ID). Totally, 1611 out of 2572 HPO IDs were mapped to 19 “Chapter” IDs (clusters) of ICD-10. The mutation-based similarities of one ICD-10 cluster with another cluster or itself can be measured by calculating median –log

(P-values) of all pairs of disease from two clusters. We found that most ICD-10 clusters (11 out of 19 clusters) are likely to be similar to themselves in terms of mutation-based similarity (Detailed in Table S5 and Figure 5). In other words, in the majority of the cases, phenotypic classification was consistent with mutation-based classification. However, eight clusters are more similar to at least one other cluster than themselves (more details given in the Discussion).

Discussion

In this study, we systematically quantified genetic similarity across 2,572 human inherited diseases according to the extent of sharing of disease-causing mutations between diseases. This study was made possible by the availability of a comprehensive collection of disease-causing mutations represented by the HGMD and the hierarchical definitions of diseases by the HPO. We showed that disease-disease relations are more accurately described by shared specific mutations than by shared disease-causing genes based on multiple independent tests. This includes consistency with genetic correlation in brain diseases according to shared heritability, the sharing of over-expressed genes in different diseases, an improvement in predicting novel disease-causing mutations by ten-fold cross-validation, and more accurate discrimination of Crohn's disease-susceptible individuals from healthy individuals.

Shared genes are commonly employed for building disease-disease relationships. Inferring disease-disease relationships by shared mutations is a much stricter criterion than by shared genes. This is because shared mutations automatically imply shared genes, but not *vice versa*. As a result, shared genes will identify a larger number of related disease pairs. However, using a statistical test as a ranking tool allows us to identify significantly different related disease pairs. In fact, from the same number of top ranked pairs (17,514), only 60.7% are shared between DRAM and DRAG. This highlights the importance of using mutation-based inference, rather than gene-based inference, for revealing disease-disease associations.

We have demonstrated that DRAM is useful for identifying disease-disease relationships. When DRAM was applied to 13 brain diseases, a significant number of overlapping mutations was found between ASD and ADHD. This result is consistent with the observation that ADHD and ASD exhibit behavioural overlaps accompanied by abnormalities in similar brain systems (Lim et al. 2014; Polderman et al. 2014; Reilly et al. 2014). The similarities between diseases

as revealed by DRAM are also reflected in the overlap of their clinical features. For example, the relationship of ischemic strokes (I. Stroke) and migraine with aura identified by DRAM but missed by heritability, is consistent with the fact that ischemic strokes can be miscategorised as cerebral infarction during the course of a typical migraine with aura attack (Lee et al. 2016). The relationship between migraine and focal epilepsy concurs with the fact that both diseases can be associated with Cortical Spreading Depression (CSD), a wave of profound cellular depolarization (Rogawski 2012). Moreover, genetic similarity between Alzheimer disease and psychiatric diseases may reflect a shared cognitive effect spectrum. For example, major depressive disorder may be observed in the early stage of Alzheimer disease (Niida et al. 2013), and Alzheimer disease shares some of the same symptoms with late onset bipolar disorder (Lebert et al. 2008; Besga et al. 2015). Future studies that aim to connect both overlapping mutations and overlapping clinical features across genetically similar diseases could reveal additional insights into underlying molecular mechanisms.

Another useful application of DRAM lies in its ability to disclose potential disease-causing mutations from the comparison of genetically related diseases. This is evident from the ten-fold cross-validation using annotated disease-causing mutations in HGMD as well as the improved identification of Crohn's disease-susceptible individuals through consideration of predicted disease-causing mutations. The overall low accuracy is likely because only a relatively small number of disease-causing mutations have so far been identified in the majority of heritable diseases. As more and more disease-causing mutations are discovered, we expect that DRAM will become even more powerful in uncovering both novel disease-disease relationships and disease-causing mutations, and therefore has the potential to aid disease diagnosis.

We further tested the disease relationship by clustering diseases according to ICD-10 IDs, and built an association network of 19 disease chapters. The disease chapters of ICD-10 were

based on the phenotypic characteristics of diseases (2010). We found that most disease clusters showed the closest relatedness with themselves based on shared mutations, which reflected consistency between the genetic and phenotypic similarity between diseases. On the other hand, the finding of eight phenotype clusters showing genetic similarity with other phenotypes suggests a shared genetic susceptibility, and hence could provide new information regarding the underlying biological mechanisms of these phenotypes by leveraging what is known about genetically similar phenotypes. We also examined disease-disease relationships at the level of individual diseases. If selected diseases containing at least nine associated diseases with a P-value$<1\times10^{-6}$ (~0.05/40,000 [total number of disease pairs in the test]), 767 diseases (3,568 pairs) from 14 disease clusters were involved. Figure S4 demonstrates how various diseases from different ICD-10 chapters connect with each other within and between chapters. The disease category with the highest number of connections is the Morphological abnormalities of the central nervous system" belonging to the "Congenital malformations, deformations and chromosomal abnormalities" cluster (shortened to "Congenital malformations, etc" in Figure S4). "Morphological abnormality of the central nervous system" has 62 connections within the chapter but 46 connections to diseases in other chapters.

Methods

HGMD dataset: We obtained 157,581 mutations [HUIYING: Causing human inherited disease? Associated with human inherited disease? In other words, did you use just the DMs or DM?s and polymorphic variants as well?] (including single base-pair substitutions and micro-insertions/deletions) from the Human Gene Mutation Database (HGMD Professional v.2015.4) (Stenson et al. 2017). In order to obtain a hierarchical definition of diseases, we mapped the "Cui" ID (defined by Unified Medical Language System, UMLS) of the diseases annotated in the HGMD into phenotypes as defined in the Human Phenotype Ontology (HPO) (Kohler et al. 2014). The HPO was downloaded from http://human-phenotype-

ontology.github.io/downloads.html, including 11,785 HPO phenotypes. A total of 2,572 diseases with 124,452 annotated mutations were mapped to HPO. The number of mutations associated with a disease ranges from 1 to as high as 20,685 for abnormalities of nervous system morphology, with an average of 16 mutations per disease and a median of 2. After including mutations assigned to subclasses of diseases, the average number increased to 48 and the median to 18.

Brain disorders dataset: The genetic relationships of five psychiatric diseases and eight neurological diseases were obtained from a recent study carried out by the Brain Storm consortium (Anttila et al. 2016). These 13 diseases have 48 pairs sharing mutations in the HGMD database.

DisGeNET dataset: In order to compare with a gene-based method, we downloaded the DisGeNET (Pinero et al. 2015), which collected associations between 15,093 phenotypes and 17,381 genes by integrating several public data sources and literature, and clinical or abnormal human phenotypes. Of these, 11,103 disease phenotypes could be mapped to the HPO, and 927 diseases were found to overlap with the diseases in the HGMD dataset. Because this dataset was to be used to assess our ability to predict associated mutations, we retained those diseases associated with >20 mutations; finally, 445 diseases remained. These diseases were associated with 14,085 genes from the DisGeNET, and 5102 genes harbouring 122,191 genetic mutations from the HGMD.

Evaluating genetic similarity between diseases

We evaluated genetic similarity between diseases by the number of shared mutations through the exact binominal test. Mutations were assigned to a disease if they associated either with the disease itself or its descendant phenotypes according to the hierarchical relationships of diseases. For two diseases that are associated with m mutations and n mutations, respectively,

the expected number of mutations shared by random chance is m*n/N, where N is the total number of mutations in the HGMD. We excluded pairs of diseases where one disease was a descendant of the other.

Clustering diseases according to ICD-10 ID

Diseases that shared mutations were mapped onto the disease clusters defined by ICD-10. ICD-10 is the 10th version of the International Statistical Classification of Diseases and Related Health Problems (ICD), a medical classification list drawn up by the World Health Organization (WHO). ICD-10 contains 22 disease "Chapters" that are divided into 263 disease "blocks" and 12,131 disease "codes". In this study, the HPO IDs were mapped to disease "Chapters".

Evaluating disease relations by cell-specific expression profiles

The FANTOM5 (Lizio et al. 2015) human gene expression data cover a diverse range of 889 cellular states and assess the promoter activity of every known gene in each cell or tissue type. We downloaded the human CAGE peak data qualified by transcripts per million (TMP) (http://fantom.gsc.riken.jp/5/datafiles/latest/extra/CAGE_peaks/hg19.cage_peak_tpm_ann.osc.txt.gz). In this file, the terms describing cell types were queried in "Medical dictionary for Regular activity (MedDRA®)" to obtain "Cui" IDs that were subsequently mapped to HPO IDs, which led to 64 unique HPO IDs. The expression peaks assigned to each gene were averaged over multiple experiments for the same cell type, which yielded an estimated aggregate expression value for each gene in each cell type (disease). A gene was considered to be enriched in a single cell type if the gene expression was at least *n* (5, 7, or 10)-fold higher in that cell type than the average over all other cell types in the dataset (Yu et al. 2015). We defined genes enriched in a cell type as being over-expressed in that specific cell type (disease).

We measured disease relations according to the number of overlapping over-expressed genes in cells. For each pair of diseases, the expected number of overlapping over-expressed genes by chance would be m*n/N, where m and n are the numbers of over-expressed genes in two cell types, respectively; N is total number of genes (24,383, RefSeq gene information [hg19]) in the human genome (Zhao et al. 2016). The overexpression similarity was evaluated by the exact binomial test.

Predicting novel mutations according to genetic similarity across diseases

A cross-validation test was performed to assess how well disease-causing mutations from a particular disease can be inferred from known mutations associated with genetically similar diseases. For each query disease, we randomly divided all of its reported mutations into 10 folds. Each time, nine folds of mutations were employed as training sets to calculate the genetic similarity with all other diseases. The score of a mutation in other diseases was calculated by using , where is a converted Chi-square with one degree of freedom from the P-value of a relevant disease *i*, and is 1.0 only if mutation *j* is known to be associated with disease *i,* and is 0, otherwise. Mutations existing in the nine folds were excluded for testing. All mutations above a given threshold were predicted to be associated with the disease, whilst the others were not. The predicted disease-causing mutations were true positives (TP) if they appeared in the remaining fold and false positives (FP) if not; predicted non-disease-causing mutations were false negatives (FN) if they appeared in the test fold and true negatives (TN) if not. The schematic diagram for ten-fold cross-validation is shown in Figure S1. Performance was measured by sensitivity [SN=TP/(TP+FN)], precision [PR=TP/(TP+FP)], and the Matthews correlation coefficient [

The threshold was individually optimized to maximise MCC for each fold of mutations per disease.

Additionally, we also calculated the enrichment factor (EF) in the top of all mutations:

———————— , where is the number of correctly predicted mutations in the top predicted mutations in the test fold, is the number of predicted mutations in the top scored mutations, is the number of mutations in the test fold relevant to a disease; and is the total number of mutations in the HGMD dataset.

Discriminating Crohn's disease-susceptible individuals from healthy persons

We further utilized the mutation-based genetic similarity approach to identify Crohn's disease-susceptible individuals. The whole exome sequencing data in VCF format of 111 persons was downloaded from Critical Assessment of Genome Interpretation (CAGI). These samples were derived from 47 healthy individuals and 64 individuals who were affected by Crohn's disease. For each individual, we counted the number (N) of harboured mutations considered to be relevant to Crohn's disease according to the HGMD. We also employed mutations predicted as being relevant to Crohn's disease by DRAM as described in the section above. The final disease susceptibility score by *D-score*= , where S_i represents the predicted Chi-square statistics for novel Crohn's disease-related mutations, and w is a normalisation factor determined by trial and error (w=30). All putatively neutral mutations were removed by excluding those mutations with a minor allele frequency >0.01 in the Exome Sequencing Project (ESP) (http://evs.gs.washington.edu/EVS/). The performance of each method was assessed by the Area Under the receiver operating characteristic Curve (AUC).

Authors' Contributions HZ and YY designed the study; HZ, YY implemented the methods and produced results; HZ, YY, MM, DC and YZ wrote the manuscript; all authors reviewed the manuscript.

Acknowledgments

We would like to thank National Health and Medical Research Council Early Career Fellowship of Australia (1091816) to HZ and National Health and Medical Research Council (1059775 and 1083450) of Australia to YZ. We also gratefully acknowledge the use of the High Performance Computing Cluster "Gowonda" to complete this research. This research/project has also been undertaken with the aid of the research cloud resources provided by the Queensland Cyber Infrastructure Foundation (QCIF). MM and DNC acknowledge the financial support of Qiagen Inc through a license agreement with Cardiff University.

Reference

2010. The International Classification of Diseases, 10th Edition (ICD-10) changeover is coming. *Optometry* **81**: 551-553.

Anttila V, Bulik-Sullivan B, Finucane HK, Bras J, Duncan L, Escott-Price V, Falcone G, Gormley P, Malik R, Patsopoulos N. 2016. Analysis of shared heritability in common disorders of the brain. *bioRxiv*: 048991.

Barbosa M, Sousa AB, Medeira A, Lourenco T, Saraiva J, Pinto-Basto J, Soares G, Fortuna AM, Superti-Furga A, Mittaz L et al. 2011. Clinical and molecular characterization of diastrophic dysplasia in the Portuguese population. *Clin Genet* **80**: 550-557.

Besga A, Gonzalez I, Echeburua E, Savio A, Ayerdi B, Chyzhyk D, Madrigal JL, Leza JC, Grana M, Gonzalez-Pinto AM. 2015. Discrimination between Alzheimer's disease and late onset bipolar disorder using multivariate analysis. *Front Aging Neurosci* **7**: 231.

Blair DR, Lyttle CS, Mortensen JM, Bearden CF, Jensen AB, Khiabanian H, Melamed R, Rabadan R, Bernstam EV, Brunak S et al. 2013. A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk. *Cell* **155**: 70-80.

Carvill GL, Heavin SB, Yendle SC, McMahon JM, O'Roak BJ, Cook J, Khan A, Dorschner MO, Weaver M, Calvert S et al. 2013. Targeted resequencing in epileptic encephalopathies identifies *de novo* mutations in *CHD2* and *SYNGAP1*. *Nat Genet* **45**: 825-830.

Cross-Disorder Group of the Psychiatric Genomics C Lee SH Ripke S Neale BM Faraone SV Purcell SM Perlis RH Mowry BJ Thapar A Goddard ME et al. 2013. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet* **45**: 984-994.

Damaj L, Lupien-Meilleur A, Lortie A, Riou E, Ospina LH, Gagnon L, Vanasse C, Rossignol E. 2015. *CACNA1A* haploinsufficiency causes cognitive impairment, autism and epileptic encephalopathy with mild cerebellar symptoms. *Eur J Hum Genet* **23**: 1505-1512.

Davis DA, Chawla NV. 2011. Exploring and exploiting disease interactions from multi-relational gene and phenotype networks. *PLoS One* **6**: e22670.

Debette S, Kamatani Y, Metso TM, Kloss M, Chauhan G, Engelter ST, Pezzini A, Thijs V, Markus HS, Dichgans M et al. 2015. Common variation in *PHACTR1* is associated with susceptibility to cervical artery dissection. *Nat Genet* **47**: 78-83.

Faraone SV, Biederman J, Wozniak J. 2012. Examining the comorbidity between attention deficit hyperactivity disorder and bipolar I disorder: a meta-analysis of family genetic studies. *The Am J Psych* **169**: 1256-1266.

Giollo M, Jones DT, Carraro M, Leonardi E, Ferrari C, Tosatto SC. 2017. Crohn disease risk prediction-Best practices and pitfalls with exome data. *Hum Mutat* doi:10.1002/humu.23177.

Jiang YH, Yuen RK, Jin X, Wang M, Chen N, Wu X, Ju J, Mei J, Shi Y, He M et al. 2013. Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *Am J Hum Genet* **93**: 249-263.

Jouvenceau A, Eunson LH, Spauschus A, Ramesh V, Zuberi SM, Kullmann DM, Hanna MG. 2001. Human epilepsy associated with dysfunction of the brain P/Q-type calcium channel. *Lancet* **358**: 801-807.

Kirschner LS, Sandrini F, Monbo J, Lin JP, Carney JA, Stratakis CA. 2000. Genetic heterogeneity and spectrum of mutations of the *PRKAR1A* gene in patients with the carney complex. *Hum Mol Genet* **9**: 3037-3046.

Kohler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, Black GC, Brown DL, Brudno M, Campbell J et al. 2014. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res* **42**: D966-974.

Lebert F, Lys H, Haem E, Pasquier F. 2008. [Dementia following bipolar disorder]. *Encephale* **34**: 606-610.

Lechner M, Hohn V, Brauner B, Dunger I, Fobo G, Frishman G, Montrone C, Kastenmuller G, Waegele B, Ruepp A. 2012. CIDeR: multifactorial interaction networks in human diseases. *Genome Biol* **13**: R62.

Lee DS, Park J, Kay KA, Christakis NA, Oltvai ZN, Barabasi AL. 2008. The implications of human metabolic network topology for disease comorbidity. *Proc Natl Acad Sci U S A* **105**: 9880-9885.

Lee MJ, Lee C, Chung CS. 2016. The migraine-stroke connection. *J Stroke* **18**: 146-156.

Li J, Cai T, Jiang Y, Chen H, He X, Chen C, Li X, Shao Q, Ran X, Li Z et al. 2016. Genes with de novo mutations are shared by four neuropsychiatric disorders discovered from NPdenovo database. *Molecular Psychiatry* **21**: 290-297.

Liao Y, Anttonen AK, Liukkonen E, Gaily E, Maljevic S, Schubert S, Bellan-Koch A, Petrou S, Ahonen VE, Lerche H et al. 2010. *SCN2A* mutation associated with neonatal epilepsy, late-onset episodic ataxia, myoclonus, and pain. *Neurology* **75**: 1454-1458.

Lim L, Chantiluke K, Cubillo AI, Smith AB, Simmons A, Mehta MA, Rubia K. 2014. Disorder-specific grey matter deficits in attention deficit hyperactivity disorder relative to autism spectrum disorder. *Psychological Medicine* doi:10.1017/S0033291714001974: 1-12. [HUIYING: Please update reference]

Linglart A, Menguy C, Couvineau A, Auzan C, Gunes Y, Cancel M, Motte E, Pinto G, Chanson P, Bougneres P et al. 2011. Recurrent *PRKAR1A* mutation in acrodysostosis with hormone resistance. *N Engl J Med* **364**: 2218-2226.

Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, Abugessaisa I, Fukuda S, Hori F, Ishikawa-Kato S et al. 2015. Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol* **16**: 22.

Maki N, Komatsuda A, Wakui H, Ohtani H, Kigawa A, Aiba N, Hamai K, Motegi M, Yamaguchi A, Imai H et al. 2004. Four novel mutations in the thiazide-sensitive Na-Cl co-transporter gene in Japanese patients with Gitelman's syndrome. *Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association* **19**: 1761-1766.

Melamed RD, Emmett KJ, Madubata C, Rzhetsky A, Rabadan R. 2015. Genetic similarity between cancers and comorbid Mendelian diseases identifies candidate driver genes. *Nat Commun* **6**: 7033.

Niida A, Niida R, Kuniyoshi K, Motomura M, Uechi A. 2013. Usefulness of visual evaluation of the anterior thalamic radiation by diffusion tensor tractography for differentiating between Alzheimer's disease and elderly major depressive disorder patients. *Int J Gen Med* **6**: 189-200.

Park J, Lee DS, Christakis NA, Barabasi AL. 2009. The impact of cellular networks on disease comorbidity. *Molecular Systems Biology* **5**: 262.

Pinero J, Queralt-Rosinach N, Bravo A, Deu-Pons J, Bauer-Mehren A, Baron M, Sanz F, Furlong LI. 2015. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database (Oxford)* **2015**: bav028.

Polderman TJ, Hoekstra RA, Posthuma D, Larsson H. 2014. The co-occurrence of autistic and ADHD dimensions in adults: an etiological study in 17,770 twins. *Translational Psychiatry* **4**: e435.

Reilly C, Senior J, Murtalgh L. 2014. ASD, ADHD, mental health conditions and psychopharmacology in neurogenetic syndromes: parent survey. *J Intellect Disabil Res* [HUIYING: Please update reference!] doi:10.1111/jir.12147.

Ristow M. 2004. Neurodegenerative disorders associated with diabetes mellitus. *J Mol Med* **82**: 510-529.

Rogawski MA. 2012. Migraine and epilepsy-shared mechanisms within the family of episodic disorders. In *Jasper's Basic Mechanisms of the Epilepsies*, (ed. JL Noebels, et al.), Bethesda (MD).

Smyth DJ, Plagnol V, Walker NM, Cooper JD, Downes K, Yang JH, Howson JM, Stevens H, McManus R, Wijmenga C et al. 2008. Shared and distinct genetic variants in type 1 diabetes and celiac disease. *New Engl J Med* **359**: 2767-2777.

Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, Hussain M, Phillips AD, Cooper DN. 2017. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet*. Mar 27. doi: 10.1007/s00439-017-1779-6. [Epub ahead of print]

Subramanian M, Senthil N, Sujatha S. 2015. Idiopathic generalized epilepsy and hypokalemic periodic paralysis in a family of South Indian descent. *Case Rep Neurol Med* **2015**: 906049.

Szigethy E, Youk AO, Gonzalez-Heydrich J, Bujoreanu SI, Weisz J, Fairclough D, Ducharme P, Jones N, Lotrich F, Keljo D et al. 2015. Effect of 2 psychotherapies on depression and disease activity in pediatric Crohn's disease. *Inflamm Bowel Dis* **21**: 1321-1328.

Tsai PS, Gill JC. 2006. Mechanisms of disease: Insights into X-linked and autosomal-dominant Kallmann syndrome. *Nature Clinical Practice Endocrinology & Metabolism* **2**: 160-171.

Wei A, Lian S, Wang L, Li W. 2009. The first case report of a Chinese Hermansky-Pudlak syndrome patient with a novel mutation on *HPS1* gene. *J Dermatol Sci* **56**: 130-132.

Yu NY, Hallstrom BM, Fagerberg L, Ponten F, Kawaji H, Carninci P, Forrest AR, Fantom C, Hayashizaki Y, Uhlen M et al. 2015. Complementing tissue characterization by integrating transcriptome profiling from the Human Protein Atlas and from the FANTOM5 consortium. *Nucleic Acids Res* **43**: 6787-6798.

Zhao H, Eising E, de Vries B, Vijfhuizen LS, International Headache Genetics C, Anttila V, Winsvold BS, Kurth T, Stefansson H, Kallela M et al. 2016. Gene-based pleiotropy across migraine with aura and migraine without aura patient groups. *Cephalalgia* **36**: 648-657.