Published in final edited form as:
Cell. 2013 September 26; 155(1): . doi:10.1016/j.cell.2013.08.030.

A Non-Degenerate Code of Deleterious Variants in Mendelian Loci Contributes to Complex Disease Risk

David R. Blair[1], **Christopher S. Lyttle**[2], **Jonathan M. Mortensen**[3], **Charles F. Bearden**[4], **Anders Boeck Jensen**[5], **Hossein Khiabanian**[6], **Rachel Melamed**[6], **Raul Rabadan**[6], **Elmer V. Bernstam**[4], **Søren Brunak**[5], **Lars Juhl Jensen**[5], **Dan Nicolae**[8,9], **Nigam H. Shah**[3], **Robert L. Grossman**[9,10], **Nancy J. Cox**[9], **Kevin P. White**[9,10,*], and **Andrey Rzhetsky**[9,10,*]

[1]Committee on Genetics, Genomics, and Systems Biology, the University of Chicago, Chicago, IL 60637, USA

[2]The Center for Health and the Social Sciences, the University of Chicago, Chicago, IL 60637, USA

[3]Stanford Center for Biomedical Informatics Research, Stanford, CA 94305, USA

[4]School of Biomedical Informatics, Department of Internal Medicine, the University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[5]Center for Biological Sequence Analysis, Technical University of Denmark, Copenhagen, Denmark DK-2800

[6]Department of Biomedical Informatics, Center for Computational Biology and Bioinformatics, Columbia University, New York, NY, 10032, USA

[7]Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, 2200, Denmark

[8]Department of Statistics, the University of Chicago, Chicago, IL 60637, USA.

[9]Departments of Medicine and Human Genetics, the University of Chicago, Chicago, IL 60637, USA.

[10]Computation Institute, Institute for Genomics and Systems Biology, the University of Chicago, Chicago, IL 60637, USA.

Summary

Whereas countless highly penetrant variants have been associated with Mendelian disorders, the genetic etiologies underlying complex diseases remain largely unresolved. Here, we examine the extent to which Mendelian variation contributes to complex disease risk by mining the medical records of over 110 million patients. We detect thousands of associations between Mendelian and complex diseases, revealing a non-degenerate, phenotypic code that links each complex disorder to a unique collection of Mendelian loci. Using genome-wide association results, we demonstrate that common variants associated with complex diseases are enriched in the genes indicated by this "Mendelian code." Finally, we detect hundreds of comorbidity associations among Mendelian disorders, and we use probabilistic genetic modeling to demonstrate that Mendelian variants likely

* Correspondence: kpwhite@uchicago.edu (K.P.W.), arzhetsky@uchicago.edu (A.R.)..

contribute non-additively to the risk for a subset of complex diseases. Overall, this study illustrates a complementary approach for mapping complex disease loci and provides unique predictions concerning the etiologies of specific diseases.

Introduction

Clinicians and geneticists have previously observed that rare, Mendelian disorders such as thalassemia and cystic fibrosis, certain chromosomal abnormalities (such as Down and Kleinfelter syndromes), and severely deleterious copy number variants (CNV) often predispose patients to more common, apparently non-Mendelian diseases. For example, patients with beta-thalassemia, Huntingtons disease and Friederichs ataxia often develop type 2 diabetes mellitus (De Sanctis et al., 1988; Podolsky et al., 1972; Ristow, 2004), and carriers of the genetic variants associated with Lujan-Fryns and DiGeorge (velo-cardio-facial) syndromes display an increased risk for schizophrenia (De Hert et al., 1996; Sinibaldi et al., 2004). Additionally, bearers of the 16p11.2 microdeletions and microduplications often develop autism (Kumar et al., 2008; Tabet et al., 2012). In such cases, it has been long suspected that the simple and complex diseases share genetic architecture; whether there is a broader pattern of such associations, however, remains unclear.

A large and growing number of Mendelian and chromosomal diseases have been precisely assigned to particular causal genetic events. Although Mendelian disorders often manifest many of the same complexities that are associated with multi-genic diseases, such as incomplete penetrance and genetic modification (Badano et al., 2006), they remain the best understood in terms of their underlying genetic etiologies. This is because the variants underlying Mendelian diseases are generally highly penetrant and nearly unaffected by the environment. Furthermore, their physiologic effects are often severe, allowing for very early diagnosis, sometimes even prenatally. Therefore, in contrast to more complex human disorders, the clinical diagnosis of a Mendelian disease reveals unique insight into the genotype of the affected patient. Consequently, we hypothesize that statistically significant comorbidities between complex and Mendelian illnesses represent a type of genetic association, in which a non-Mendelian phenotype is mapped to the genetic loci that cause the Mendelian disease.

By analyzing millions of electronic clinical records obtained from distinct regions of the United States and Denmark, we demonstrate that such “transitive” genetic associations are consistent and ubiquitous, yielding novel insight into the etiology of complex diseases. Furthermore, we observe that each complex disease possesses a unique Mendelian disease allelic architecture, creating a non-degenerate code that identifies each illness by its associated Mendelian loci. In support of our transitive association hypothesis, we demonstrate that complex disease genome-wide association signals are specifically enriched within the genetic loci indicated by this code. Finally, we use mathematical modeling to demonstrate that the variants underlying Mendelian disorders likely interact with one another to contribute to complex disease risk, highlighting the potential of clinical data for uncovering complicated genetic architectures.

Results

Clinical Record Analysis

We mined the administrative data associated with millions of clinical records for evidence of comorbidity among Mendelian and complex diseases. As a rule, such records are maintained in order to facilitate patient billing rather than academic research, and therefore, they may be incomplete and variably biased (van Walraven and Austin, 2012). However, this does not

diminish their overall utility for making accurate inferences about clinical phenotypes in large populations. The key to such analyses is to carefully consider how missing data and biases may affect the conclusions of the intended research and, if required, introduce appropriate corrections. Because we conditioned our inferences on the observed disease incidence counts, our comorbidity estimates did not depend on the accurate estimation of marginal disease prevalence. Therefore, we assumed a "missing at random" model for undocumented records, which is common practice for epidemiological studies with uninformatively missing data (Lyles and Allen, 2002). Finally, we took great care to focus our data analysis on clearly identifiable phenotypes (see Experimental Procedures), and we detected disease comorbidity using a sophisticated statistical pipeline that accounted for a large set of potentially confounding demographic, socioeconomic, and environmental factors (see Extended Experimental Procedures and Figure S1 for details).

We judged the quality of our statistical inferences by comparing the results generated from multiple, distinct clinical datasets. In the present study, we examined eight datasets, with the smallest and largest describing approximately 150 thousand and 100 million unique patients respectively (see Table 1 and Figure 1A). We found that our estimates of the comorbidity odds ratios for the complex-Mendelian disease pairs were remarkably consistent (see Figure 1F and 1G, all correlation p-values $< 5 \times 10^{-8}$), which is reassuring considering that the datasets represent populations in different geographic regions with variable ethnic structure and disease prevalence (Figure 1B and 1C). While it is possible that the *USA* dataset partially overlaps with the smaller, North American ones (*CU*, *NYPH*, *SU*, *TX*, and *UC*), the smaller datasets should be nearly completely disjoint from one another and from *DK*, indicating that duplicate records do not drive this result (see Extended Experimental Procedures for a more detailed treatment of potentially confounding factors). Although other groups have mined clinical record datasets for disease comorbidities in the past (Hidalgo et al., 2009; Lee et al., 2008), the vast majority of the relationships detected in this study are likely to be novel, as associations among complex and Mendelian diseases have never been analyzed at this scale (over 100 million unique patients) (see Figure 1D and 1E for a comparison to previously published results).

A Non-Degenerate Mendelian Phenotypic Code for Complex Diseases

Figure 2 summarizes all of the significant comorbidities that were detected among the complex and Mendelian disorders within our compendium of clinical records (see Table S4 for detailed results). Each colored cell in the matrix indicates the logarithm of the relative risk associated with a significant clinical signal, and the complex diseases are grouped according to our current understanding of their pathophysiology. It is reassuring to observe that many of the known comorbidities are replicated within our dataset. For example, we detected significant comorbidity between lipoprotein deficiencies and myocardial infarction (Strong and Rader, 2012) and ataxia telangiectasia and breast cancer (Sellers, 1997). However, the majority of the 2,909 associations shown in Figure 2 have not been previously reported. For example, our analysis uncovered significant clinical comorbidities between Marfan syndrome and several neuropsychiatric diseases (autism, bipolar disorder, and depression), and it determined that fragile X is significantly associated with asthma, psoriasis, and viral infection, highlighting a potential immune system dysfunction in these patients (Ashwood et al., 2010).

In Figure 3A, the rows and columns of the comorbidity matrix have been rearranged such that disorders with similar comorbidity structure are placed adjacent to one another. Importantly, this rearrangement demonstrates that each complex disease was comorbid with a diverse and unique combination of Mendelian phenotypes. Despite extensive variation within this "Mendelian code," it nonetheless recapitulates much of our current understanding of the pathophysiology of complex diseases (see Figure S2). To illustrate, we

computed the Euclidean distance between every pair of shared risk profiles and produced the Neighbor-Joining tree (Saitou and Nei, 1987) that best approximates this set of statistics (Figure 3B). Not surprisingly, the resulting tree contained many groupings that are highly consistent with our current knowledge of disease etiology. For example, autism, intellectual disability, and epilepsy form a tight cluster in the tree (replicated in 96% of bootstrap pseudosamples), consistent with previous genetic studies that have uncovered variants underlying the risk for all three neuropsychiatric traits (Shinawi et al., 2010).

Complex disease GWA signals are enriched within the genetic loci implicated by the Mendelian Code

We conjectured that the significant complex-Mendelian comorbidities displayed in Figure 2 indicate that the genes and pathways perturbed in the Mendelian disorders also play a role in the etiology of the corresponding complex diseases. Thus, we hypothesized that the "Mendelian code" could used to pinpoint loci harboring complex-disease predisposing genetic variants. To test this prediction, we probed legacy genome-wide association (GWA) results (NIH, 2012) and asked whether common variants associated with the complex diseases were enriched within the loci implicated by the Mendelian comorbidities. Overall, we observed that complex disease GWA signals were globally enriched in Mendelian loci (106 observed, 55.3 expected, 1.92-fold enrichment, $p = 4.0 \times 10^{-10}$), an observation that has been previously highlighted by others (Lupski et al., 2011). Furthermore, when we restricted our analysis to unique signals only (i.e. removed duplicate signals that were replicated in subsequent studies), the enrichment fell to 1.6-fold but remained highly significant (63 observed, 40.4 expected, $p = 4.6 \times 10^{-5}$). Importantly, complex disease-specific GWA signals were specifically enriched in the precise loci indicated by the Mendelian phenotypic code (1.97-fold enrichment, 40 observed, 20.1 expected, $p = 5.7 \times 10^{-5}$, see Table S1 for detailed results), suggesting that the comorbidities highlighted in Figure 2 reflect a shared complex-Mendelian genetic architecture. Moreover, the GWA signals enriched in comorbid Mendelian loci were more likely to be detected in multiple studies than other protein-coding SNPs, including those that lie within non-comorbid Mendelian loci (replication rates: 0.8 vs. 0.36, $p = 0.026$, Mann-Whitney-U Test). Overall, these results suggest that the loci implicated by the Mendelian code are likely to contain a spectrum of complex disease predisposing variants, providing testable hypotheses for future gene re-sequencing and exome analyses (see Discussion for details).

Mendelian Disorders Share Significant Clinical Comorbidity

Our analysis generated a surprisingly large number of statistically significant clinical associations between pairs of Mendelian disorders (462 after conservative statistical filtering, see Extended Experimental Procedures, Figure 4, Figures S4 and S5, and Table S5). We propose that these associations represent interactions among genetic variants in distinct Mendelian loci, and we found that it was possible to map individual interactions to specific biological hypotheses. As an example, we observed significant shared risk between fragile X and glycogenosis (odds ratio = 859.09), and this effect remained highly significant after controlling for a wide variety of potentially confounding factors, including disease similarity, age, gender, ethnicity, and environment (see Extended Experimental Procedures). A link between fragile X and glycogenosis has been previously proposed in the molecular genetics literature (De Boulle et al., 1993; Zang et al., 2009), and glycogen metabolism has been suggested to play an important role in fragile X pathophysiology and treatment (Min et al., 2009). A few anecdotal cases aside, however, most of the relationships in Figure 4 represent totally undocumented interactions among rare and highly deleterious genetic variants.

We do acknowledge that some of the apparently significant comorbidities could be due to confounding factors. First, miscoding errors during medical billing could create false signals of comorbidity. This could happen, for example, if two distinct physicians examined the same patient, but erroneously entered different billing codes due to the clinical ambiguity of the Mendelian disease. Second, the co-occurrence of Mendelian phenotypes could be an artifact of a cryptic population structure. As a result of assortative mating, some sub-populations could be enriched with multiple Mendelian diseases, increasing the apparent rate of rare disease co-occurrence. Although these biases seem plausible, we do not believe that they contribute significantly to the comorbidities depicted in Figure 4 for the following reasons. First, while medical billing errors were likely present in the datasets, we went through great lengths to estimate and remove their effects (see Extended Experimental Procedures). Second, our statistical analysis procedure included a variety of demographic and environmental covariates, and we found that these potential confounders contributed only marginally to the shared risk among Mendelian disorders, casting doubt on the cryptic population structure hypothesis.

Perhaps more importantly, there are additional, orthogonal pieces of evidence that indicate that the previous two confounders are unlikely to contribute pervasively to Mendelian-Mendelian comorbidity. For example, we found that comorbid Mendelian disorders, even after removing all clinically similar disease pairs, tended to map to genetic loci that are significantly more functionally alike than expected by chance, as measured by their distances within a large human gene network (Lee et al., 2011), see Extended Experimental Procedures, p-value < 0.00001. This result fits naturally with the theory of widespread epistasis among Mendelian variants, but it cannot be easily explained using either of the other two hypotheses. Second, cryptic population structure, billing code errors, and genetic interactions make very different predictions with respect to complex disease risk in patients diagnosed with multiple, comorbid Mendelian disorders (see Experimental Procedures). In the next section, we use probabilistic modeling to provide direct statistical evidence that the risk for several complex diseases is highly consistent with the genetic modifier hypothesis described above.

Mendelian Loci Contribute to Complex Disease Risk in a Non-Additive Manner

Examining the complex disease risk in patients with compound Mendelian phenotypes offered an additional avenue for assessing the likelihood of the three mechanisms proposed in the previous section. As a simple example, assume that the relationships in Figure 4 were dominated by miscoding errors. If this were true, then an individual diagnosed with one comorbid Mendelian disorder should have the same average risk for the complex disease as an individual diagnosed with two. Instead, we observed that individuals diagnosed with two comorbid, Mendelian phenotypes had a higher average risk for the complex disease in 62 out of the 65 of the illnesses considered in this study (p-value = 6.2×10^{-12}, Wilcoxon signed-rank test). Such analyses provide only indirect evidence for the genetic modifier hypothesis. To provide direct statistical evidence, we formulated two probabilistic genetic models for complex disease risk in patients diagnosed with compound Mendelian phenotypes. The first, termed the additive model (Risch, 1990), is consistent with cryptic population structure and assumes that the Mendelian variants contribute independently to complex disease risk. The second, called the combinatorial model, invokes a simple mechanism for genetic epistasis among the Mendelian variants. By fitting each model to the clinical datasets, we formally tested whether the genetic modifier hypothesis was supported by the observed risk profiles of the complex diseases.

The two genetic models that we considered shared several assumptions in common. First, both assumed that each complex disease is associated with a set of genetic loci, some of which are linked to Mendelian phenotypes as well. This assumption ensured that each model

was capable of accounting for the comorbidity structure that was observed within the clinical data. Second, the models assumed that the genetic loci under consideration possessed only dominant, recessive, or X-linked (haploid) variants, although the frequency and penetrance of such variants could vary freely. Third, they assumed that the penetrance values for the complex diseases, at both Mendelian and other loci, were sampled from some population-level distribution. Similarly, both models assumed that the frequencies of the deleterious genotypes were sampled from a population-level distribution as well. Finally, the models assumed that the total number of loci associated with any complex disease was finite and fixed.

The two models differed in one important assumption only: the additive genetic model assumed that the effects of the deleterious genotypes contributed independently (additively) to complex disease risk (Risch, 1990), while our non- additive model broke this assumption by introducing “communities” of loci. Essentially, such communities represented loci that normally function in a coordinated manner, and our non-additive model assumed that at least one adverse genetic event must be present within multiple communities in order to generate significant complex disease risk. Thus, this community-based genetic model required *combinations* of particular deleterious genotypes, so we refer to it as the combinatorial model to differentiate it from other non-additive genetic mechanisms. In the present study, the combinatorial model was constructed to be as simple as possible and included only two communities of loci.

Although the assumptions outlined above are simple, they generated two models that make distinctly different predictions in terms of the average complex disease risk in patients with multiple comorbid Mendelian phenotypes (see Extended Experimental Procedures for details). Specifically, the additive model predicted that the average complex disease risk should increase linearly as function of the number of comorbid Mendelian phenotypes, while the combinatorial model predicted a *super-linear* (polynomial) increase. Furthermore, if billing record miscoding errors were included into the additive model, the increase in complex disease risk would become *sub-linear*. All three signatures were visually apparent in the risk profiles for the complex diseases (see Figure S3), although sub-linear increases were rare (approximately 5 out of 65 illnesses). To formally quantify the evidence in favor of each model, we took a Bayesian approach and computed their posterior probabilities conditioned on the clinical data (see Extended Experimental Procedures).

Due to the computational burden associated with fitting genetic models to over one hundred million patients, we selected a representative sample of 20 complex diseases for analysis. In practice, the population-level mean of the genotype frequencies and the total number of complex disease predisposing loci were not jointly identifiable, so we repeated the model selection procedure for a range of potential loci numbers (see Experimental Procedures). Each model was clearly favored for a subset of diseases, but the combinatorial model had stronger overall support across the entire set (see Figure 3C). For diseases that displayed a sub-linear increase in risk (consistent with possible miscoding errors), the additive model was supported over the combinatorial by a wide margin (see diabetes mellitus type II in Figure S3). Overall, this result provides additional and orthogonal support for the hypothesis that Mendelian-Mendelian comorbidities were driven by genetic interactions. It also suggests that certain complex diseases (such as Addisons disease, acute glomerulonephritis, and malignant brain neoplasms, but not the two forms of diabetes or bipolar disorder) have a non-additive (epistatic) genetic architecture with respect to Mendelian disease variants.

Discussion

Highly penetrant mutations have not been found for most common, complex diseases, despite intensive search. While rare single nucleotide and copy number variants have been implicated in some complex disorders, including Intellectual Disability (Vissers et al., 2010), Schizophrenia (Bassett et al., 2008) and Autism (Iossifov et al., 2012), these results appear to be the exception rather than the norm. The fact that we observed widespread comorbidity among Mendelian and complex diseases suggests that rare, highly penetrant variants do in fact play a significant role in complex disease risk, but their deleterious effects do not result in single, isolated diseases. Instead, highly deleterious genetic variants likely induce a variety of pathological consequences, consistent with the Mendelian code displayed in Figure 2 and Figure 3A. Such analysis resonates with the results of recent genetic dissections of oligogenic traits, such as Bardet-Biedl syndrome, which appears to harbor a diverse genetic architecture that produces a variety of clinical phenotypes (Katsanis et al., 2001; Zaghloul et al., 2010).

In addition to these direct associations, we also observed that common risk variants associated with complex diseases were specifically enriched in comorbid Mendelian loci. The most obvious explanation for this is that some of the patients included in GWA studies carried genetic variation that predisposed them to both the Mendelian and complex diseases. However, there are several reasons to be skeptical of this hypothesis. First, subjects with Mendelian disorders are typically, by design, excluded from GWAS (Zhao et al., 2010). Second, Mendelian diseases are rare and have overt clinical presentations, so it is highly improbable that such carriers were included in the studies unintentionally. Finally, even if the rate of accidental sampling of Mendelian phenotypes were aberrantly high, we do not believe that “synthetic” genome-wide associations, in which the detected common variants are in linkage disequilibrium with Mendelian disease alleles, drive our results (Dickson et al., 2010). As discussed at length by others (Visscher et al., 2012), numerous empirical and theoretical analyses are simply not consistent with this interpretation.

As an alternative explanation, we and others (Lupski et al., 2011) propose that Mendelian genes carry both rare and common deleterious variants, such that alleles from both ends of the frequency spectrum contribute to disease risk. Rare, highly penetrant variants cause Mendelian disorders, while common variants with milder effects contribute to the complex phenotypes. By design, GWAS detect only the latter end of the frequency spectrum, and the former is typically uncovered through linkage analysis and sequencing. When the Mendelian and complex phenotypes are similar, we can think of the two disorders as different ends of the same genetic and phenotypic spectrum, known as the allelic series hypothesis. In fact, there are several well-documented examples of this phenomenon, such as the familial and common forms of Parkinsonism and blood lipid disorders (Manolio et al., 2009).

However, a few special cases, this straightforward definition of allelic series is not very helpful when explaining Mendelian and complex phenotypes that are comorbid and share genetic loci but are biologically dissimilar. For example, asthma and systemic primary carnitine deficiency share clinical risk and are both associated with variants in the *SLC22A5* locus, but there is no obvious relationship between the biology underlying these two diseases. Instead, we suggest a modification to the allelic series hypothesis that considers the multifactorial nature of gene function. On one end of the spectrum, we hypothesize that very rare, Mendelian disease variants completely or nearly completely abolish all of a gene's physiological functions. Therefore, their effects are highly penetrant and pleiotropic, resulting in overt pathologies (like Mendelian disease) while at the same time increasing a carrier's risk for a variety of other disorders. On the other end, less deleterious mutations may perturb the same genes, but their effects are more limited, perhaps modifying only a

subset of a gene's functions. In such instances, the resulting deleterious effects may be quite subtle, allowing the variants to reach relatively high population frequencies. Moreover, their ultimate pathological manifestations may be very different than those that are observed in patients harboring Mendelian variants, reflecting the different subsets of physiological functions perturbed by each mutation type.

With this in mind, we hypothesize that the loci underlying comorbid Mendelian disorders represent strong candidates for harboring complex disease predisposing genetic variants with moderate and weak effects, as the Mendelian associations have already suggested that the underlying gene is involved in the pathophysiology of the complex disorder. This theory is supported by our GWAS enrichment results, but we believe that it extends to rare variants with larger effects as well. Because they have already been shown to contain a variety of complex disease predisposing variants, we propose that the best candidates for testing this hypothesis are perhaps those loci that were found to contain both common risk and Mendelian disease causing variants (see Table S1). Consistent with this hypothesis, we note that 4 out of the 7 neoplasms for which GWAS results were available were found to associate with both common and rare Mendelian genetic variants in the *TERT* locus, which encodes the human telomerase reverse transcriptase. Mendelian variants within this locus completely abolish reverse transcriptase enzymatic activity, resulting in several overt, pathological symptoms (combined into a syndrome called dyskeratosis congenita) (Kirwan and Dokal, 2009). Recently, a rare, germline variant in the promoter region of *TERT* was linked to a familial form of melanoma, although carriers of the allele may have increased risk for other neoplasms as well (Horn et al., 2013). In support, somatic variants within the promoter region of *TERT* were also found in a variety of human cancer cell lines (Huang et al., 2013) and solid tumors (Killela et al., 2013). Such results raise the intriguing possibility that a spectrum of *TERT*-associated variants, both rare and common, somatic and germline, increase one's risk for neoplastic disease.

Furthermore, our complex-Mendelian comorbidity analysis predicted that schizophrenia, bipolar disorder, autism and depression are all associated with the following four Mendelian loci: *SYNE1*, *PRPF3*, *CACNA1C*, and *PPP2R2B*. Consistent with their hypothesized shared genetic architecture (Cross-Disorder Group of the Psychiatric Genomics Consortium et al., 2013), these four loci were also found to harbor common genetic variants influencing risk for this same set of diseases. Interestingly, exome sequencing in autism patients has uncovered both *de novo* and inherited potentially deleterious variants in *SYNE1* (O'Roak et al., 2011; Yu et al., 2013). We find this result particularly interesting, as it suggests that these four genes may also harbor rare variants that predispose carriers to multiple neuropsychiatric disorders. If this is correct, then pooling strategies that combine sequence data from patients with these different but related complex phenotypes could offer a simple approach for increasing the power to identify rare variants with modest effects.

In the second part of our study, we discovered approximately 450 comorbidity associations among pairs of Mendelian disorders, suggesting that genetic interactions among Mendelian variants are quite common. Consistent with this hypothesis, we used genetic modeling to demonstrate that epistatic effects could be detected in the complex disease risk profiles of patients diagnosed with multiple, comorbid Mendelian disorders. At the very least, our results suggest that strongly deleterious variants have a high propensity for modifying the effects of other deleterious alleles in functionally similar genes. However, the existence of non-additive effects among rare genetic variants could have practical consequences as well. For example, undocumented epistasis among rare variants in distinct loci could negatively impact the power of targeted re-sequencing studies.

Although our inference of widespread, non-additive genetic effects is novel, it is well known that highly penetrant genetic variants are subject to modification by other alleles that exist *in trans*. For example, at first glance, the Mendelian disorder retinitis pigmentosa appears to follow the "independent effects" assumption of genetic additivity quite well (Parmeggiani, 2011), as several, highly penetrant mutations in distinct genes have been associated with the phenotype. However, this disease was also one of the first Mendelian phenotypes with clearly demonstrated di-genic inheritance (Kajiwara et al., 1994), and epistatic interactions among multiple loci have been reported for other Mendelian phenotypes as well, such as Bardet-Biedl syndrome (Badano et al., 2006). There are also known examples in which *trans* genetic variants modify the specific symptoms of Mendelian disorders. More specifically, several suspected genetic modifiers have been previously identified for cystic fibrosis (CF) (Cutting, 2010), a recessive disease caused by mutations in the *CFTR* gene. CF patients display a variety of symptoms, including mucus congestion in the lungs, intestinal obstruction, diabetes, abnormal gut microflora, and liver disease, and nearly a dozen loci have been identified that appear to modulate the strength of these clinical symptoms (Cutting, 2010). For example, variation in *EDNRA* appears to affect the pulmonary function of CF patients, while *MSRA* alleles modulateintestinal obstruction.

In summary, we detected thousands of instances of comorbidity between complex-Mendelian and Mendelian-Mendelian disease pairs. The existence of such associations was not unexpected; however, their widespread nature was surprising. Furthermore, although there is a growing body of evidence that genetic interactions are common across both Mendelian and complex traits, such as Alzheimer's disease (Badano and Katsanis, 2002), facioscapulohumeral dystrophy type 2 (Lemmers et al., 2012), and Hirschsprungs disease (Wallace and Anderson, 2011), we believe that this is the first instance in which such relationships have been uncovered systematically across multiple complex diseases. Ultimately, we demonstrate that digital phenotypic data can be utilized to infer genetic and genomic architectures, potentially allowing for extensive, novel analyses in the field of human disease genetics. Moreover, this work highlights the importance of documenting a wider spectrum of Mendelian and other disease traits in a very large population of humans, perhaps the entire United States or even multiple countries, in order to uncover the pathophysiology associated with very rare genetic events.

Experimental Procedures

Phenotype Curation and Billing Code Assignments

To identify the clinical phenotypes of interest, we used the disease codes provided by the International Disease Classification (ICD) system (WHO, 2010), see Table 1. The mappings between billing codes (both ICD9 and ICD10) and diseases were obtained from (Rzhetsky et al., 2007) and by manual curation, first by a Ph.D.-level contractor trained in a biomedical field followed by two of the authors, iteratively. All billing code mappings for the complex and Mendelian diseases are provided in Table S2 and Table S3 respectively. The billing codes enabled the identification of 65 specific complex disorders and 95 Mendelian disease groups (representing 213 disorders) (see Tables S2 and S3 respectively). Note, this reduction of 213-to-95 was not a choice of experimental design but necessitated by the ICD9 code taxonomy. See Extended Experimental Procedures for additional details.

Clinical Record Analysis

Each clinical record database was first parsed (see Table 1), removing duplicate records and identifying patients that harbored the diseases of interest. In theory, a small fraction of these records could be shared between *USA* and the other, smaller US datasets (*CU*, *NYPH*, *SU*, *TX*, *UC*) since some patients could have been documented in multiple databases. Because

duplicate records would strongly bias the results for rare diseases, we decided against simply combining the information from different datasets into a single meta-analysis. Instead, we performed an independent statistical analysis for each dataset and then combined the results according to a conservative procedure (see Extended Experimental Procedures for details). For the complex-Mendelian comorbidity analysis, any disease pair containing a complex or Mendelian disease that was specific to males or females (indicated by ♂ and ♀ respectively in Figure 2) was analyzed after conditioning on the appropriate gender; gender-specific diseases were not included in the Mendelian-Mendelian analysis. The *MED* dataset (Hidalgo et al., 2009; Lee et al., 2008) was excluded from the meta-analysis, as we were unable to consistently identify our phenotypes of interest. Specifically, the *MED* dataset provides individual ICD9 code counts only, but many of the disorders used in our analysis map to multiple such codes. Additional details concerning our statistical procedures for the analysis of complex-Mendelian and Mendelian-Mendelian disease pairs are provided in the Extended Experimental Procedures.

Neighbor-Joining Tree Inference

The complex disease tree was constructed from the Mendelian comorbidity relationships using the Neighbor-Joining method (Saitou and Nei, 1987). See Extended Experimental procedures for additional details.

GWAS Enrichment Analysis

To test for an enrichment of common, complex disease risk variants in Mendelian loci, we aligned legacy genome-wide association results (NIH, 2012) with the SNP-to-gene annotations provided by SCAN (Gamazon et al., 2010). Binomial tests that specifically controlled for gene length and SNP annotation biases were used to assess enrichment (see Extended Experimental Procedures for details).

The Additive and non-Additive Genetic Models for Complex Disease Risk

In the main text, we briefly described two competing genetic models that specify distinct mechanisms for how multiple Mendelian disease variants combine to affect complex disease risk. Ultimately, the additive and combinatorial models make very different predictions with respect to the increase in complex disease risk as a function of the number of comorbid Mendelian phenotypes, allowing them to be differentiated within our massive clinical datasets. The mathematical details concerning this prediction are somewhat involved, and the interested reader should consult the Extended Experimental Procedures. In the following section, we simply introduce our competing genetic models using standard notation.

Consistent with common practice (Risch, 1990), each of our genetic models treats the genotype (g) and phenotype (φ) of an individual as random variables. Their joint probability is equivalent to the expected population frequency of individuals that possess both a particular genotype (G) and disease of interest (D). It is computed by taking the product of the genotype frequency and its corresponding penetrance:

$$P\left(\phi{=}D, g{=}G|\Theta\right) = P\left(g{=}G|\Theta\right) P\left(\phi{=}D|g{=}G, \Theta\right) = F\left(G\right) \times W_{D}\left(G\right),$$

where $F(G)$ is the probability of observing genotype G and $W_D(G)$ is the genetic penetrance of G with respect to phenotype D (i.e. the probability of D given G) (Risch, 1990). The overall expected prevalence of the disease within the population is computed by summing the previous probability over all possible genotypes:

$$P(\phi=D|\Theta)=\sum_{G}F(G)\times W_{D}(G).$$

Although not included for the sake of simplicity, environmental factors can be easily incorporated into this framework through the inclusion of additional random variables.

Our additive genetic model is specified within the previous framework by defining the following simple penetrance function (Risch, 1990):

$$W_{D}(G)=1-\prod_{i=1}^{n}[1-W_{D}(G_{i})],$$

where n is the number of independent loci affecting phenotype D, and $W_D(G_i)$ is the marginal penetrance function of the genotype at the i^{th} locus (Risch, 1990), which may take a variety of forms (dominant, recessive, additive, etc). Technically, the model assumes that each locus contributes independently to complex disease risk, and this is generally the assumption that underlies most "additive" models in human genetics. That said, it also approximates a stricter definition of "additivity" in which the probability of the complex disease is simply the linear combination of the penetrance probabilities of the individual loci (Risch, 1990).

Our non-additive genetic model assumes that the deleterious genotypes belong to a different "communities" of loci that act coordinately, and at least one adverse genetic event must be present within multiple communities in order to generate significant complex disease risk. Because this model requires combinations of deleterious alleles, we call it the "combinatorial" model. To illustrate, imagine two disjoint groups of loci, or "communities," each harboring a set of genotypes that pre-dispose an individual to the disease of interest. We denote the two communities using circle and square subscripts, such that $\{g_{\bigcirc,1}, g_{\bigcirc,2}, ..., g_{\bigcirc,n_{\bigcirc}}\}$ and $\{g_{\square,1}, g_{\square,2}, ..., g_{\square,n_{\square}}\}$ denote the genetic loci that belong to each community and $n_{\bigcirc}$ and $n_{\square}$ denote community sizes. To simplify notation, we will indicate either the square or the circle community, depending on context, using the $\mathscr{C}$ symbol $(\mathscr{C}=\{\circ,\square\})$. Assuming an additive model both within and across communities, the penetrance function for the two-community combinatorial model is:

$$W_{D}(G)=\prod_{\mathscr{C}\in\{\circ,\square\}}\left(1-\prod_{i=1}^{n_{\mathscr{C}}}[1-W_{D}(G_{i})]\right),$$

Note, more general formulations of the model could allow for more than two communities and a variety of different community- and loci-specific penetrance functions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to Steven Bagley, Richard R. Hudson, Ivan Iossifov, Ravinesh Kumar, Simon Lovestone, Fabiola Rivas, Gregory Gibson, Jason Pitt, Michael Wigler, and anonymous reviewers for helpful comments on earlier versions of the manuscript, and to GeneXplain, GmbH, for help with annotation of Mendelian disorders. This work

was supported by NIH grants 1P50MH094267, NHLBI MAPGen U01HL108634-01, P50GM081892-01A1, and GM007281.

References

Ashwood P, Nguyen DV, Hessl D, Hagerman RJ, Tassone F. Plasma cytokine profiles in Fragile X subjects: is there a role for cytokines in the pathogenesis? Brain Behav Immun. 2010; 24:898–902. [PubMed: 20102735]

Badano JL, Katsanis N. Beyond Mendel: an evolving view of human genetic disease transmission. Nature reviews Genetics. 2002; 3:779–789.

Badano JL, Leitch CC, Ansley SJ, May-Simera H, Lawson S, Lewis RA, Beales PL, Dietz HC, Fisher S, Katsanis N. Dissection of epistasis in oligogenic Bardet-Biedl syndrome. Nature. 2006; 439:326–330. [PubMed: 16327777]

Bassett AS, Marshall CR, Lionel AC, Chow EW, Scherer SW. Copy number variations and risk for schizophrenia in 22q11.2 deletion syndrome. Human molecular genetics. 2008; 17:4045–4053. [PubMed: 18806272]

Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech. 2008; 10:10008–10020.

Cross-Disorder Group of the Psychiatric Genomics Consortium. Smoller JW, Craddock N, Kendler K, Lee PH, Neale BM, Nurnberger JI, Ripke S, Santangelo S, Sullivan PF. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. Lancet. 2013; 381:1371–1379. [PubMed: 23453885]

Cutting GR. Modifier genes in Mendelian disorders: the example of cystic fibrosis. Annals of the New York Academy of Sciences. 2010; 1214:57–69. [PubMed: 21175684]

De Boulle K, Verkerk AJMH, Reyniers E, Vits L, Hendrickx J, Van Roy B, Van Den Bos F, de Graaff E, Oostra BA, Willems PJ. A point mutation in the FMR-1 gene associated with fragile X mental retardation. Nature genetics. 1993; 3:31–35. [PubMed: 8490650]

De Hert M, Steemans D, Theys P, Fryns JP, Peuskens J. Lujan-Fryns syndrome in the differential diagnosis of schizophrenia. American journal of medical genetics. 1996; 67:212–214. [PubMed: 8723050]

De Sanctis V, Zurlo MG, Senesi E, Boffa C, Cavallo L, Di Gregorio F. Insulin dependent diabetes in thalassaemia. Archives of disease in childhood. 1988; 63:58–62. [PubMed: 3348650]

Gamazon ER, Zhang W, Konkashbaev A, Duan S, Kistner EO, Nicolae DL, Dolan ME, Cox NJ. SCAN: SNP and copy number annotation. Bioinformatics. 2010; 26:259–262. [PubMed: 19933162]

Hidalgo CA, Blumm N, Barabasi AL, Christakis NA. A dynamic network approach for the study of human phenotypes. PLoS Comput Biol. 2009; 5:e1000353. [PubMed: 19360091]

Horn S, Figl A, Rachakonda PS, Fischer C, Sucker A, Gast A, Kadel S, Moll I, Nagore E, Hemminki K, et al. TERT Promoter Mutations in Familial and Sporadic Melanoma. Science. 2013; 339:959–961. [PubMed: 23348503]

Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, Garraway LA. Highly Recurrent TERT Promoter Mutations in Human Melanoma. Science. 2013; 339:957–959. [PubMed: 23348506]

Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J, Yamrom B, Lee YH, Narzisi G, Leotta A, et al. De novo gene disruptions in children on the autistic spectrum. Neuron. 2012; 74:285–299. [PubMed: 22542183]

Kajiwara K, Berson EL, Dryja TP. Digenic retinitis pigmentosa due to mutations at the unlinked peripherin/RDS and ROM1 loci. Science. 1994; 264:1604–1608. [PubMed: 8202715]

Katsanis N, Ansley SJ, Badano JL, Eichers ER, Lewis RA, Hoskins BE, Scambler PJ, Davidson WS, Beales PL, Lupski JR. Triallelic inheritance in Bardet-Biedl syndrome, a Mendelian recessive disorder. Science. 2001; 293:2256–2259. [PubMed: 11567139]

Killela PJ, Reitman ZJ, Jiao Y, Bettegowda C, Agrawal N, Diaz LA, Friedman AH, Friedman H, Gallia GL, Giovanella BC, et al. TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. PNAS. 2013; 110:6021–6026. [PubMed: 23530248]

Kirwan M, Dokal I. Dyskeratosis congenita, stem cells and telomeres. Biochimica et biophysica acta. 2009; 1792:371–379. [PubMed: 19419704]

Kumar RA, KaraMohamed S, Sudi J, Conrad DF, Brune C, Badner JA, Gilliam TC, Nowak NJ, Cook EH Jr. Dobyns WB, et al. Recurrent 16p11.2 microdeletions in autism. Human molecular genetics. 2008; 17:628–638. [PubMed: 18156158]

Lee DS, Park J, Kay KA, Christakis NA, Oltvai ZN, Barabasi AL. The implications of human metabolic network topology for disease comorbidity. Proc Natl Acad Sci U S A. 2008; 105:9880–9885. [PubMed: 18599447]

Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome research. 2011; 21:1109–1121. [PubMed: 21536720]

Lemmers RJ, Tawil R, Petek LM, Balog J, Block GJ, Santen GW, Amell AM, van der Vliet PJ, Almomani R, Straasheijm KR, et al. Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. Nature genetics. 2012; 44:1370–1374. [PubMed: 23143600]

Lupski JR, Belmont JW, Boerwinkle E, Gibbs RA. Clan genomics and the complex architecture of human disease. Cell. 2011; 147:32–43. [PubMed: 21962505]

Lyles RH, Allen AS. Estimating crude or common odds ratios in case-control studies with informatively missing exposure data. American journal of epidemiology. 2002; 155:274–281. [PubMed: 11821253]

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461:747–753. [PubMed: 19812666]

Min WW, Yuskaitis CJ, Yan Q, Sikorski C, Chen S, Jope RS, Bauchwitz RP. Elevated glycogen synthase kinase-3 activity in Fragile X mice: key metabolic regulator with evidence for treatment potential. Neuropharmacology. 2009; 56:463–472. [PubMed: 18952114]

NIH. 2012. http://www.genome.gov/admin/gwascatalog.txt.

O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S, Karakoc E, Mackenzie AP, Ng SB, Baker C, et al. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nature genetics. 2011; 43:585–589. [PubMed: 21572417]

Parmeggiani F. Clinics, epidemiology and genetics of retinitis pigmentosa. Current genomics. 2011; 12:236–237. [PubMed: 22131868]

Podolsky S, Leopold N, Sax D. Increased frequency of diabetes mellitus in patients with huntington's chorea. The Lancet. 1972; 299:1356–1359.

Risch N. Linkage strategies for genetically complex traits. I. Multilocus models. American journal of human genetics. 1990; 46:222–228. [PubMed: 2301392]

Ristow M. Neurodegenerative disorders associated with diabetes mellitus. J Mol Med. 2004; 82:510–529. [PubMed: 15175861]

Rzhetsky A, Wajngurt D, Park N, Zheng T. Probing genetic overlap among complex human phenotypes. Proc Natl Acad Sci U S A. 2007; 104:11694–11699. [PubMed: 17609372]

Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987; 4:406–425. [PubMed: 3447015]

Sellers TA. Genetic factors in the pathogenesis of breast cancer: their role and relative importance. The Journal of nutrition. 1997; 127:929S–932S. [PubMed: 9164266]

Shinawi M, Liu P, Kang SH, Shen J, Belmont JW, Scott DA, Probst FJ, Craigen WJ, Graham BH, Pursley A, et al. Recurrent reciprocal 16p11.2 rearrangements associated with global developmental delay, behavioural problems, dysmorphism, epilepsy, and abnormal head size. Journal of medical genetics. 2010; 47:332–341. [PubMed: 19914906]

Sinibaldi L, De Luca A, Bellacchio E, Conti E, Pasini A, Paloscia C, Spalletta G, Caltagirone C, Pizzuti A, Dallapiccola B. Mutations of the Nogo-66 receptor (RTN4R) gene in schizophrenia. Human mutation. 2004; 24:534–535. [PubMed: 15532024]

Strong A, Rader DJ. Sortilin as a regulator of lipoprotein metabolism. Current atherosclerosis reports. 2012; 14:211–218. [PubMed: 22538429]

Tabet AC, Pilorge M, Delorme R, Amsellem F, Pinard JM, Leboyer M, Verloes A, Benzacken B, Betancur C. Autism multiplex family with 16p11.2p12.2 microduplication syndrome in monozygotic twins and distal 16p11.2 deletion in their brother. European journal of human genetics : EJHG. 2012; 20:540–546. [PubMed: 22234155]

van Walraven C, Austin P. Administrative database research has unique characteristics that can risk biased results. Journal of clinical epidemiology. 2012; 65:126–131. [PubMed: 22075111]

Visscher, Peter M.; Brown, Matthew A.; McCarthy, Mark I.; Yang, J. Five Years of GWAS Discovery. The American Journal of Human Genetics. 2012; 90:7–24.

Vissers LE, de Ligt J, Gilissen C, Janssen I, Steehouwer M, de Vries P, van Lier B, Arts P, Wieskamp N, del Rosario M, et al. A de novo paradigm for mental retardation. Nature genetics. 2010; 42:1109–1112. [PubMed: 21076407]

Wallace AS, Anderson RB. Genetic interactions and modifier genes in Hirschsprung's disease. World journal of gastroenterology : WJG. 2011; 17:4937–4944. [PubMed: 22174542]

WHO. 2010. http://www.who.int/classifications/icd/en/

Yu, Timothy W.; Chahrour, Maria H.; Coulter, Michael E.; Jiralerspong, S.; Okamura-Ikeda, K.; Ataman, B.; Schmitz-Abe, K.; Harmin, David A.; Adli, M.; Malik, Athar N., et al. Using Whole-Exome Sequencing to Identify Inherited Causes of Autism. Neuron. 2013; 77:259–273. [PubMed: 23352163]

Zaghloul NA, Liu Y, Gerdes JM, Gascue C, Oh EC, Leitch CC, Bromberg Y, Binkley J, Leibel RL, Sidow A, et al. Functional analyses of variants reveal a significant role for dominant negative and common alleles in oligogenic Bardet-Biedl syndrome. Proc Natl Acad Sci U S A. 2010; 107:10602–10607. [PubMed: 20498079]

Zang JB, Nosyreva ED, Spencer CM, Volk LJ, Musunuru K, Zhong R, Stone EF, Yuva-Paylor LA, Huber KM, Paylor R, et al. A Mouse Model of the Human Fragile X Syndrome I304N Mutation. PLoS genetics. 2009; 5

Zhao J, Bradfield JP, Zhang H, Annaiah K, Wang K, Kim CE, Glessner JT, Frackelton EC, Otieno FG, Doran J, et al. Examination of all type 2 diabetes GWAS loci reveals HHEX-IDE as a locus influencing pediatric BMI. Diabetes. 2010; 59:751–755. [PubMed: 19933996]

Highlights

- Analyzed over 100 million unique patient records from the USA and Denmark.
- Discovered a non-degenerate, Mendelian comorbidity code for complex diseases.
- Results predict genetic loci enriched with a spectrum of complex disease variants.
- Inferred widespread epistasis among loci harboring deleterious Mendelian variants.

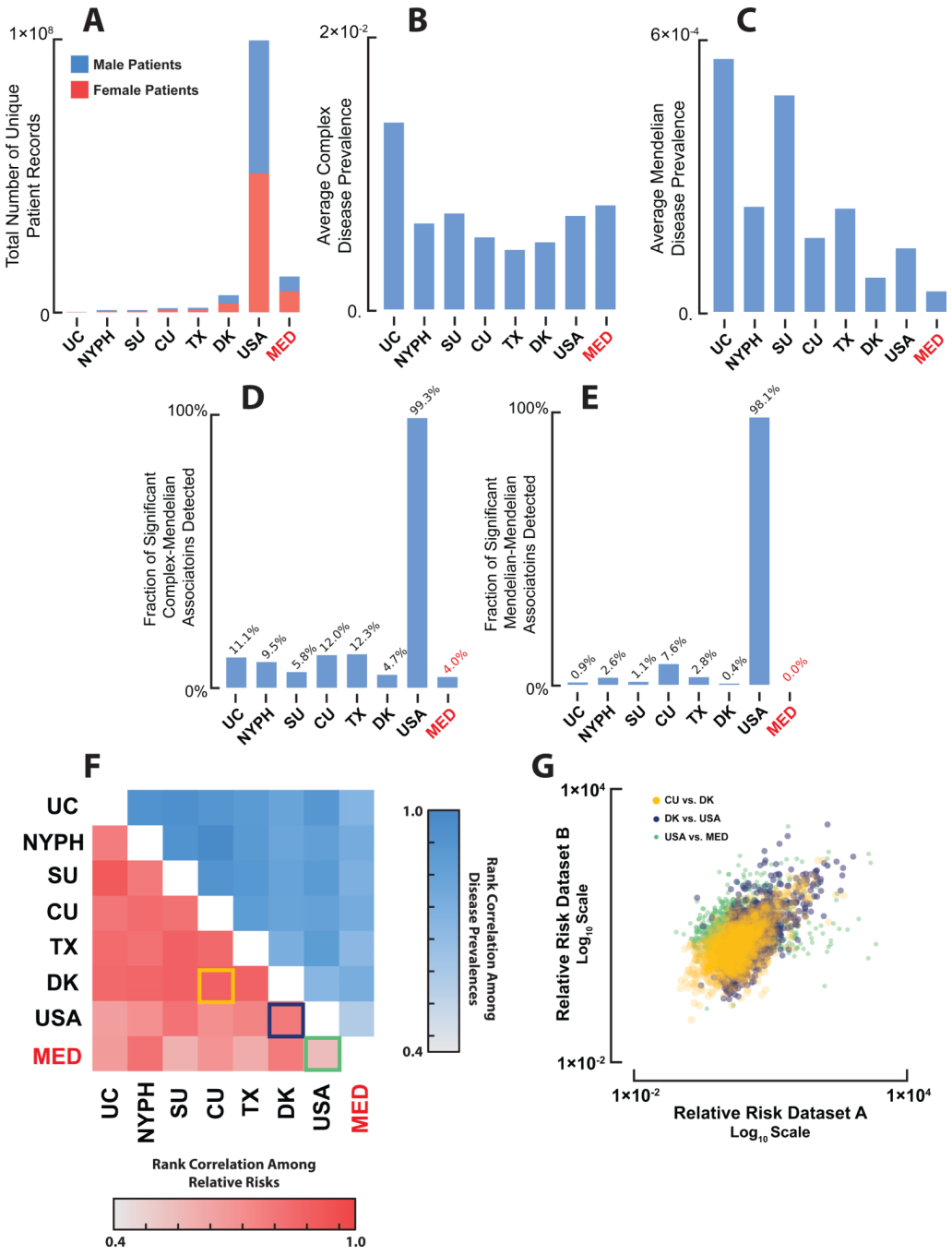

Figure 1. A systematic comparison of the eight clinical record datasets analyzed in this study

(A) The total number of records in each dataset, broken down by gender. Panels (B) and (C) display the average prevalence for the complex and Mendelian diseases across the 8 datasets. Using the superset of the discovered associations (based on the original 7 datasets, see Extended Experimental Procedures for details), we compared the number of association signals that were detected in each dataset independently, depicted as the percentage of all associations discovered in the union of the 7 datasets (excluding *MED*): (D) Mendelian-complex and (E) Mendelian-Mendelian associations. (F) The rank correlation among relative risk estimates (lower diagonal) and disease prevalence (upper diagonal) for each significantly comorbid complex-Mendelian disease pair across the eight distinct datasets.

(G) Scatter plots depicting the relative risk correlations for three pairs of datasets, indicated using the colored boxes in panel F. See also Tables S2 and S3.

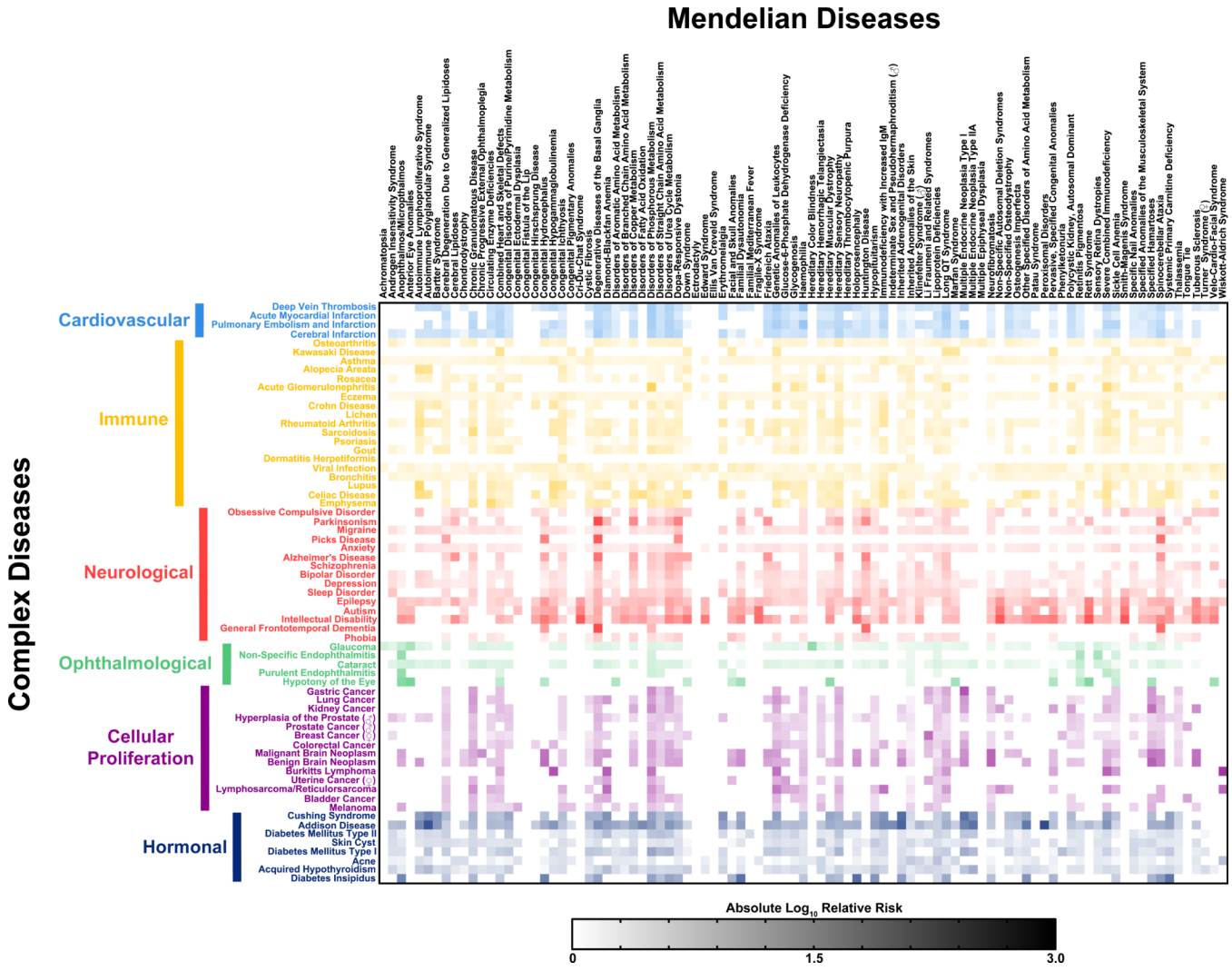

Figure 2. The significant comorbidity relationships among the complex and Mendelian disease pairs

Entries in the matrix indicate the $\log_{10}$-transformed relative risk associated with each significantly comorbid complex-Mendelian disease pair. The complex phenotypes are grouped by our current understanding of their pathophysiology. The symbols ♂ and ♀ indicate male and female-specific diseases, respectively. The numerical values underlying each association are provided Table in S4. The statistical procedure for generating these values is outlined in Figure S1. See also Tables S1-S3.

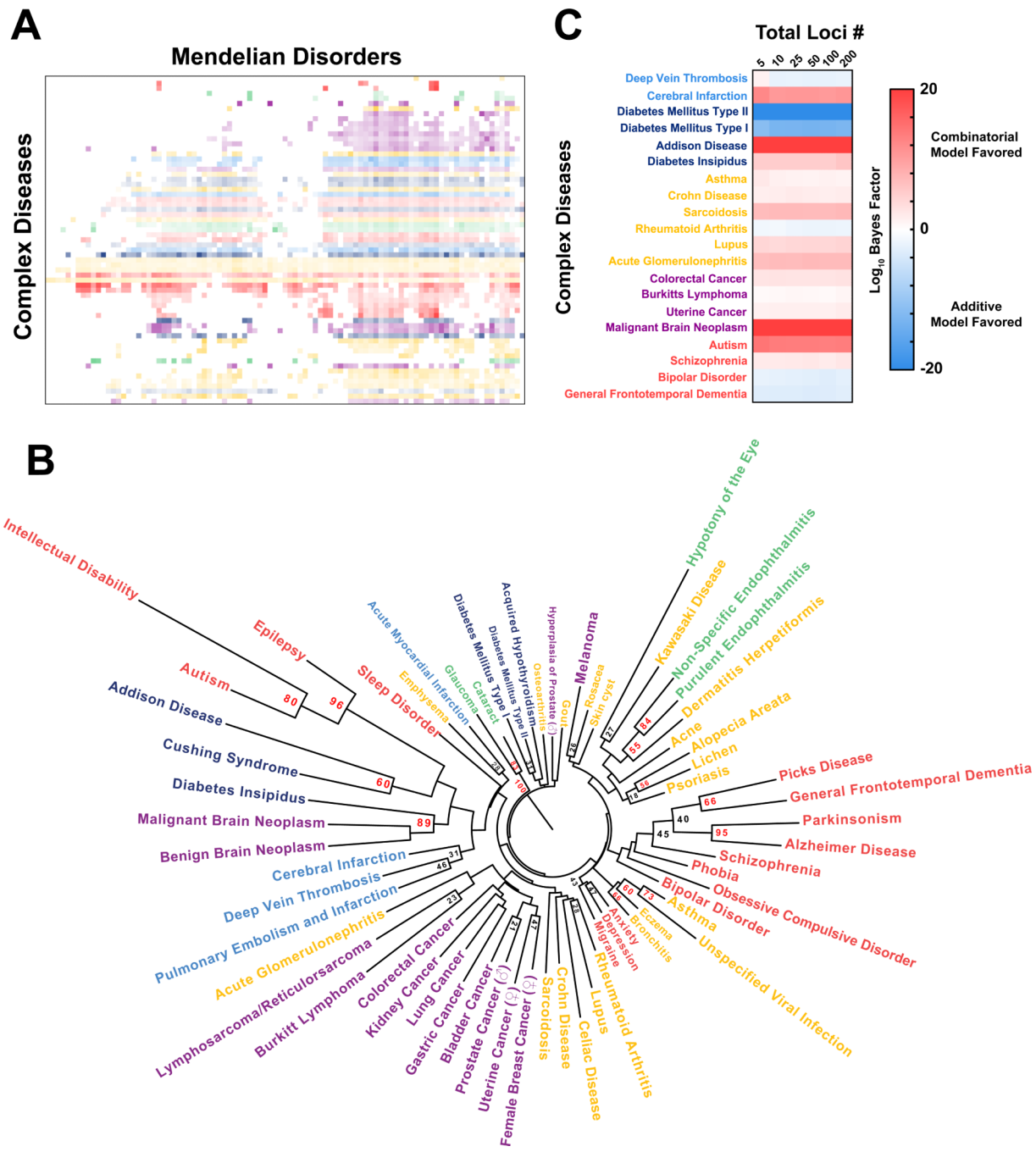

Figure 3. Complex-Mendelian comorbidities provide unique insight into the etiology of complex diseases

(A) The data matrix from Figure 2 is re-ordered such that similar rows and columns are adjacent to one another (accomplished using greedy clustering). (B) The neighbor-joining tree for the complex phenotypes was constructed from the Euclidean distances among the relative risks displayed in Figure 2 and Panel A. The bootstrap numbers (10,000 replicates) over tree arcs indicate the reliability of the corresponding partitions, with 100 being the most reliable and 0 the least. The color of the tree labels is preserved with regard to the groupings of the phenotypes depicted in Figure 2. (C) Heat map comparing the qualities of fit for the two multi-locus genetic models discussed in the main text over a range of loci numbers. The

value of the $\log_{10}$-Bayes Factor indicates the support for the combinatorial model in comparison to the additive model. A $\log_{10}$-Bayes Factor of 1 indicates that the combinatorial model is 10 times more likely than the heterogeneity model given the data. See Figure S3 for a graphical comparison of the model fits to the complex disease risk data. See also Table S2.

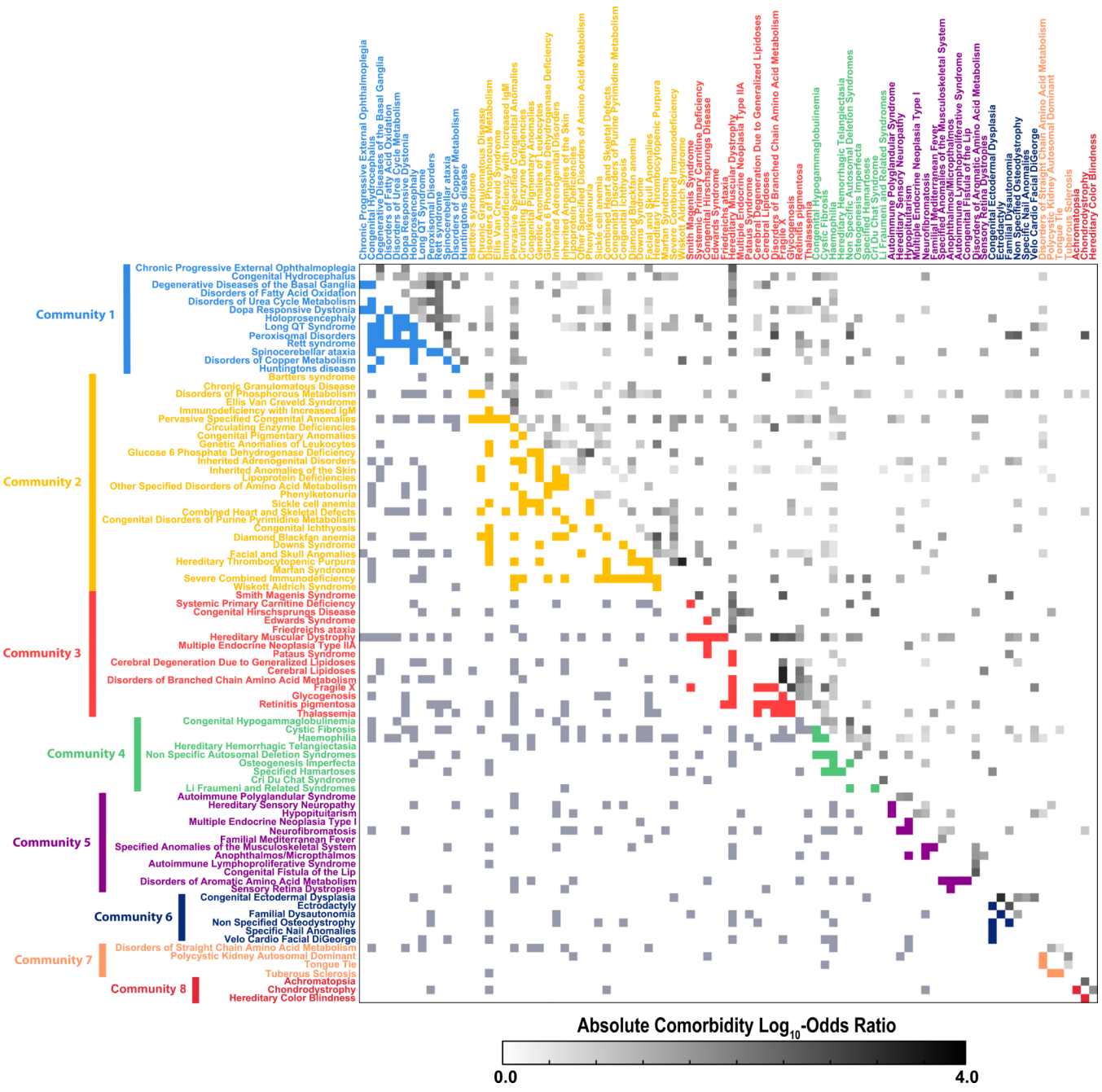

Figure 4. The significant comorbidity relationships detected among all pairs of Mendelian diseases

The upper diagonal of the matrix displays the $\log_{10}$-transformed odds ratios for the significant associations, with gray-scale intensity indicating the effect size of the association. The lower diagonal displays the community structure determined using a network-clustering algorithm (Blondel et al., 2008), with each community corresponding to a unique color and associations between diseases within the same community colored accordingly. The numerical values underlying each association are provided in Table S5. The statistical procedure for generating these values is depicted in Figure S4. An unfiltered version of the matrix is displayed in Figure S5.

Table 1
The clinical record datasets utilized in this study

This table provides a brief description, the ICD encoding type, and the size of each dataset. The MED dataset, highlighted in red, was used for comparison and was not included in the full meta-analysis.

Dataset	Description	Encoding Type	Number of unique patients
CU	Columbia University, 1985-2003, New York, NY	ICD9	1,505,822
DK	Denmark; database covering most of the country's population	ICD10	6,214,312
NYPH	New York Presbyterian Hospital and Columbia University; 2004-present, New York, NY	ICD9	767,978
SU	Stanford University, San Francisco, CA	ICD9	806,369
TX	University of Texas at Houston, Houston, TX	ICD9	1,599,528
UC	University of Chicago, Chicago, IL	ICD9	146,989
USA	*MarketScan* insurance claims dataset	ICD9	99,143,849
MED	*Medicare* database	ICD9	13,039,018
	Total:		123,223,865