Supplemental Figure Legends: Figure legends for five supplemental figures, each of which has direct correspondence to the results displayed the main figures.

Figure S1. *Flow chart for the statistical procedures, see main Figure 2 and Table S4 for results*.
Figure S2. *The Mendelian phenotypic code replicates are current understanding of complex disease pathophysiology, see main Figure 3A and 3B.*
Figure S3. *Complex disease risk profiles in patients harboring multiple Mendelian phenotypes, see main Figure 3C for comparison of model fits.*
Figure S4. *Flow chart for the statistical procedures implemented during the analysis of the Mendelian-Mendelian disease pairs, see main Figure 4 and Table S5 for results.*
Figure S5. *The matrix of comorbidity log-odds for the significant Mendelian-Mendelian comorbidities, see main Figure 4 for filtered version.*

Supplemental Tables: Tables containing pertinent information referenced by both the Experimental Procedures and main Figures/Results. Tables S2-S5 are included as excel files due to their size.

Table S1. *Common risk variants enriched within the loci indicated by the Mendelian code, see main Figure 2 and 3A.*
Table S2. *ICD9 and ICD10 codes used to identify the complex diseases, see main Figures 1, 2, 3, and 4 for results.*
Table S3. *Curated data pertaining to the Mendelian diseases used in this study, see main Figures 1, 2, 3, and 4 for results.*
Table S4. *The summary statistics for the complex-Mendelian comorbidity analysis, see Figure 2 for results matrix.*
Table S5. *The summary statistics for the Mendelian-Mendelian comorbidity analysis, see Figure 4 and Figure S5 for results.*

Extended Experimental Procedures: Details concerning the statistical analysis procedures and genetic modeling that described briefly in the Experimental Procedures and Results.

Supplemental References: References that are restricted to the Extended Experimental Procedures.

Supplemental Data

Supplemental Figures

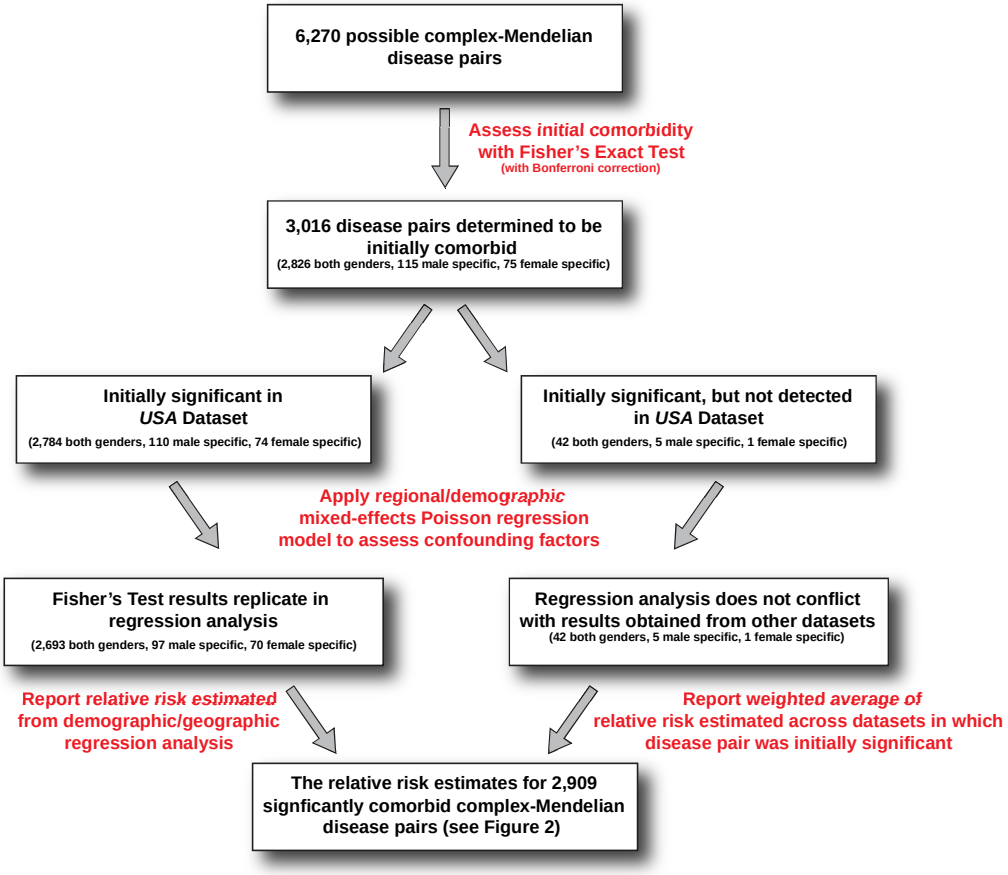

Figure S1. *Flow chart for the statistical procedures implemented during the analysis of the complex-Mendelian disease pairs, see main Figure 2 and Table S4 for results*.

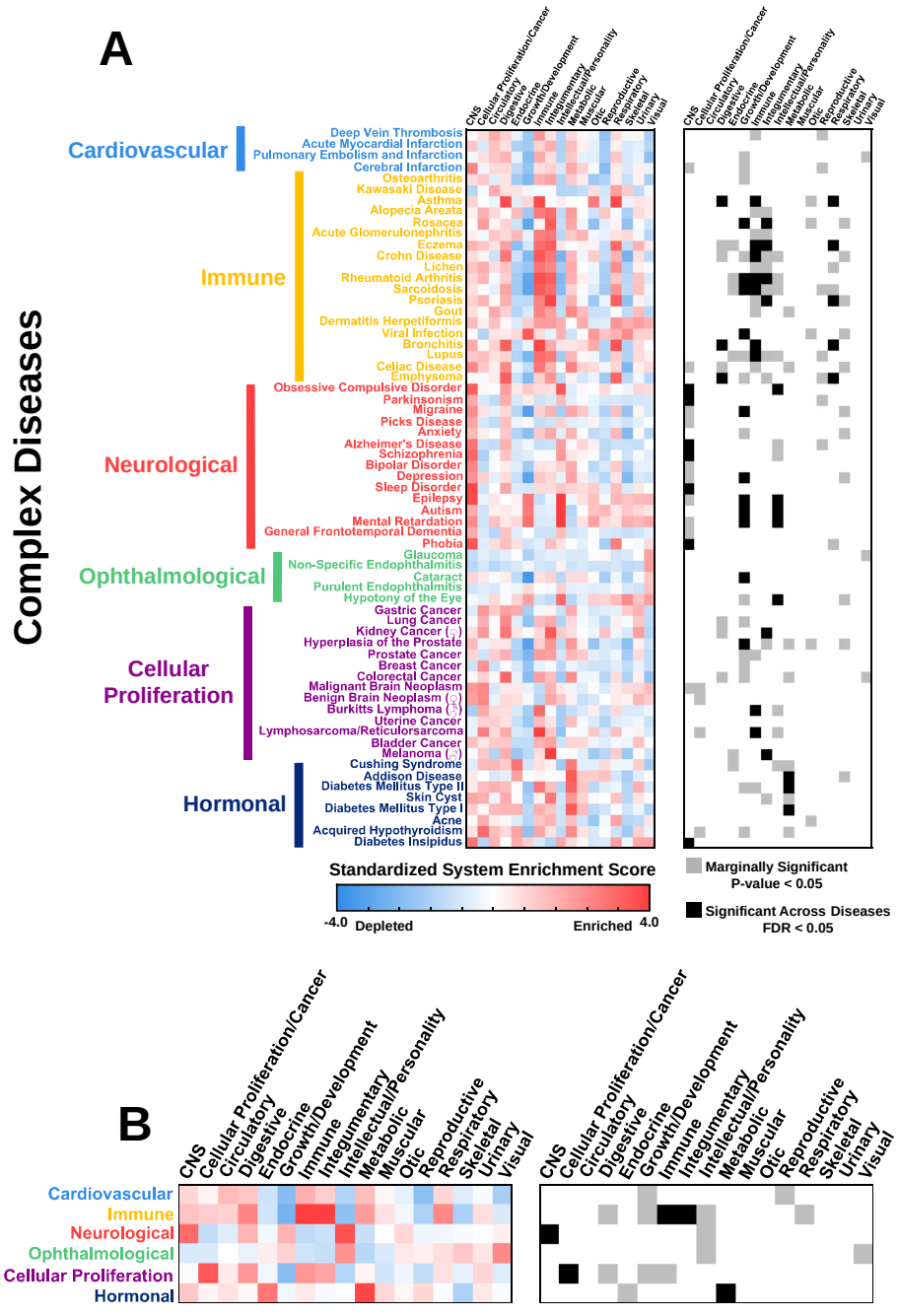

Figure S2. *The Mendelian phenotypic code replicates are current understanding of complex disease pathophysiology, see main Figure 3A and 3B.* (A), Left panel: the physiological systems enriched within the comorbidities for each complex phenotype. System enrichments were computed by constructing a null distribution of enrichment scores obtained by randomly shuffling the systems annotated to each Mendelian disorder. (A), Right Panel: The system enrichments that are marginally (*p<0.05*) and globally (*FDR<0.05*) significant. (B) The systems enriched in each group of complex diseases (left) and those that reached marginal and global significance (right).

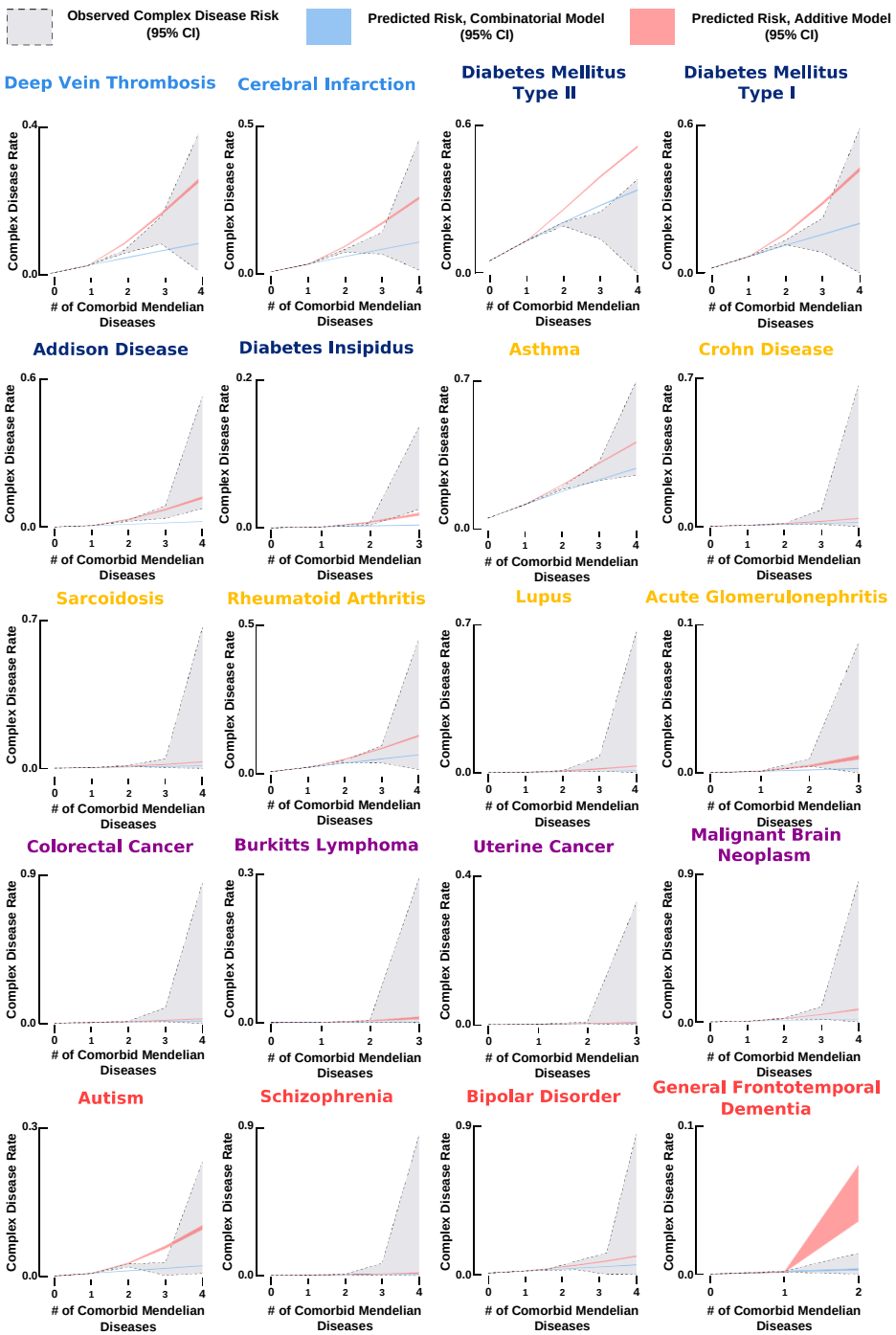

Figure S3. *Complex disease risk profiles in patients harboring multiple Mendelian phenotypes, see main Figure 3C for comparison of model fits.* Each panel depicts the observed incidence rate (95% confidence intervals, gray) for one of the twenty diseases listed in Figure 3C, plotted against the number of comorbid Mendelian diseases diagnosed per patient. The predicted risk curves for the combinatorial and additive genetic models are shown in red and blue respectively (95% credible intervals, generated through Markov Chain Monte Carlo sampling, see Extended Experimental Procedures for details).

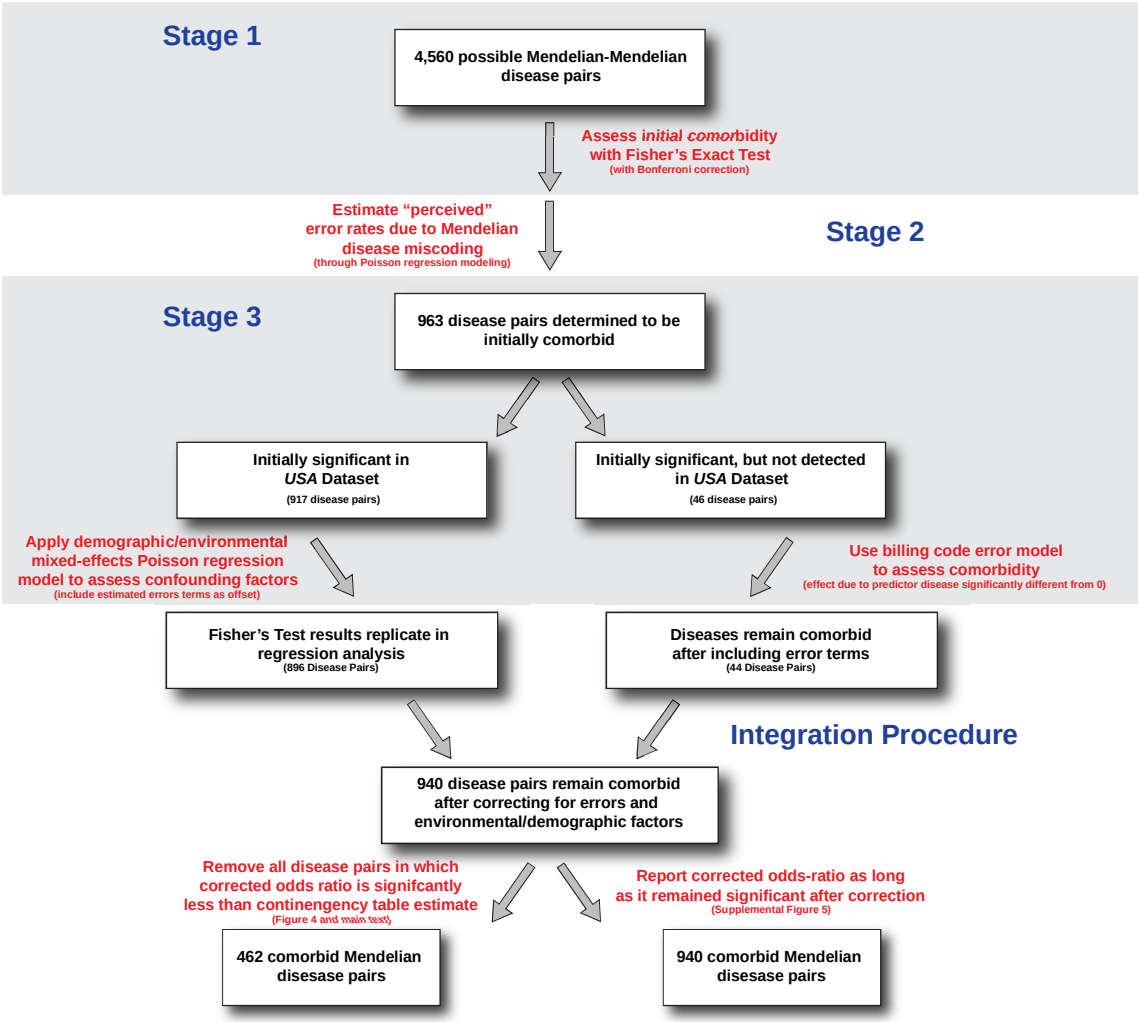

Figure S4. *Flow chart for the statistical procedures implemented during the analysis of the Mendelian-Mendelian disease pairs, see main Figure 4 and Table S5 for results.*

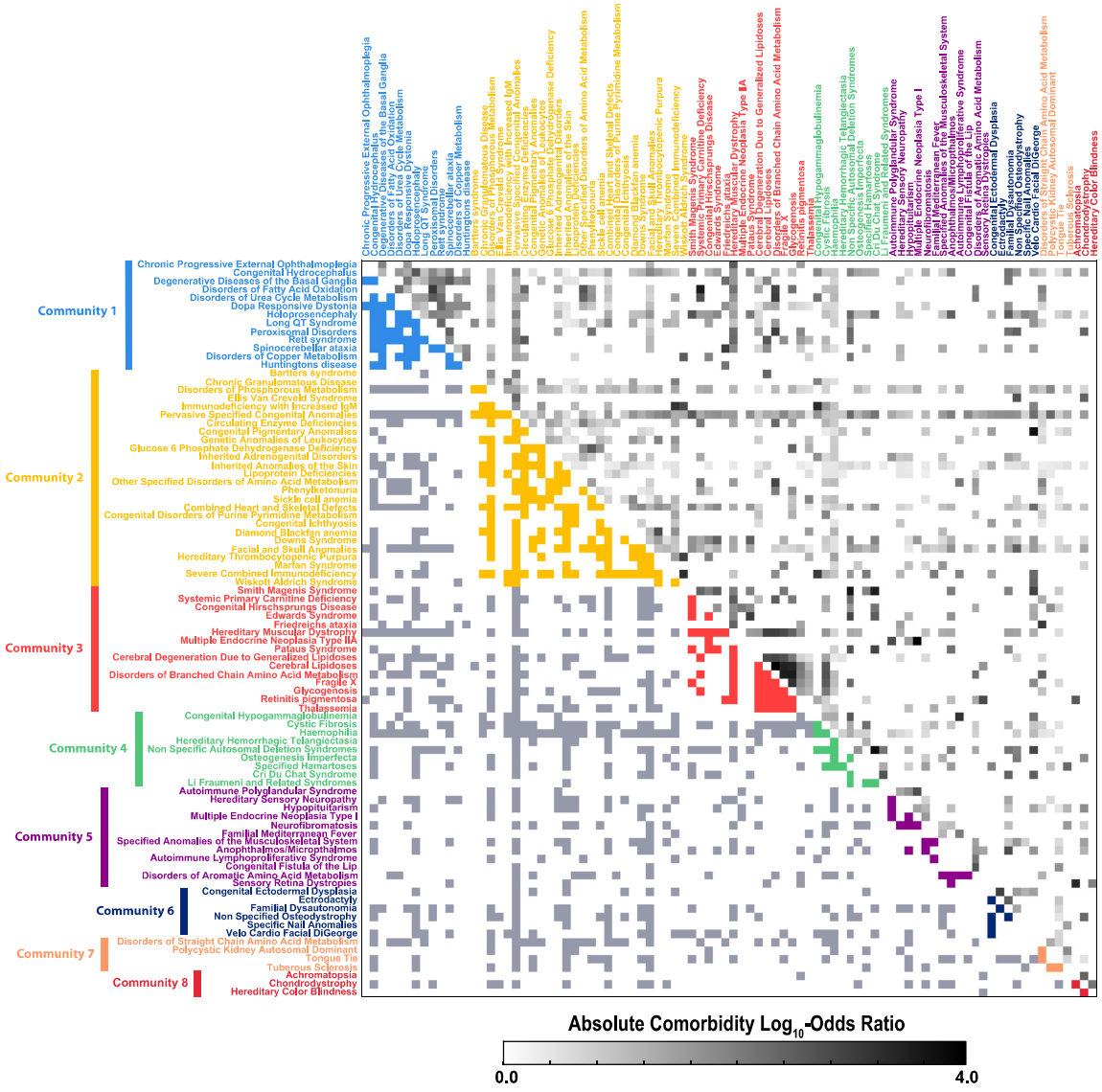

Figure S5. *The matrix of comorbidity log-odds for the significant Mendelian-Mendelian comorbidities, including the effects of shared pathology and ICD9 coding distance, see main Figure 4 for filtered version.*

Table S2. *ICD9 and ICD10 codes used to identify the complex diseases, see main Figures 1, 2, 3, and 4 for results.* This table is provided as an excel file.

Table S3. *Curated data pertaining to the Mendelian diseases used in this study, see main Figures 1, 2, 3, and 4 for results.* This includes ICD9 and ICD10 billing codes, the specific Mendelian disorders indicated by such billing codes (may be multiple), the genes associated with said disorders, and the primary and secondary biological systems affected in diseased patients. This table is provided as an excel file.

Table S4. *The summary statistics for the complex-Mendelian comorbidity analysis, see Figure 2 for results matrix.* This table contains the following dataset-specific (order of the entries provided in first line of the table) and global statistics for each initially significant complex-Mendelian disease pair: dataset-specific relative risk estimates, dataset-specific conditional odds ratio, 95% confidence intervals for said odds ratios, Bonferroni-corrected *p*-values for the association, global weighted average of the relative risk estimates, global weighted average of the odds ratio estimates, relative risk predicted by the linear model (*USA* only), and 95% confidence interval for the previous estimate. This table is provided as an excel file.

Table S5. *The summary statistics for the Mendelian-Mendelian comorbidity analysis, see Figure 4 and Figure S5 for results.* This table contains the following dataset-specific (order of the entries provided in first line of the table) and global statistics for each initially significant Mendelian-Mendelian disease pair: dataset-specific standardized risk ratio estimates, dataset-specific conditional odds ratio estimates, 95% confidence intervals for said odds ratios, Bonferroni-corrected *p*-values for the association, global weighted average of the risk ratio estimates, global weighted average of the odds ratio estimates, miscoding error corrected odds ratios, miscoding error corrected 95% confidence intervals, odds ratio predicted by the linear model (*USA* only), 95% confidence interval for the previous estimate, and the error corrected versions of the previous two statistics. This table is provided as an excel file.

Extended Experimental Procedures

Mendelian Disease Billing Code Curation

To construct the set of simple genetic disorders used in the analysis, we began with a large list of human diseases whose etiology is thought to be almost entirely due to genetic variation. This list was generated using the Online Mendelian Inheritance in Man (OMIM) (Antonarakis and McKusick, 2000; Hamosh et al., 2005) knowledge base, Orphanet (Liem, 2008; Weinreich et al., 2008) and other resources (NIH). It included both Mendelian diseases, defined as those that have been mapped to a specific set of genes/loci, and "chromosomal" disorders (monosomy, trisomy, large gene duplications and deletions), which provide a much less specific signal of genetic association. Most of the diseases in the original list could not be accurately mapped to the ICD taxonomies because they were either: 1) not covered by the taxonomies or 2) were grouped with complex diseases. Of the initial list (containing well over 5000 diseases), we were able to reliably map (once again, by iterative, manual curation) 213 specific disorders to 95 billing codes (see Table S3). Ultimately, 90 of these billing codes represent diseases linked to specific loci (Mendelian disorders) and 5 are associated with chromosomal disorders. We also annotated each Mendelian disease with the genes assigned to its respective loci by consulting a variety of resources within the National Library of Medicine (NIH) and the primary literature. The genes associated with each simple genetic phenotype are listed in Table S3.

Additional Details Concerning Disease Tree Inference

As described in the main text, the complex disease tree was inferred using the Neighbor-Joining method (Saitou and Nei, 1987). The disease-disease similarity matrix was computed using the pair-wise Euclidean distances among the complex-Mendelian disease relative risks. To assess tree reliability, we

performed bootstrapping by resampling the Mendelian disorders, with replacement, 10,000 times and estimating a separate tree for each of the samples (Efron, 1982; Felsenstein, 1985, 1993). The bootstrap numbers reported in the main text were computed as the percent of bootstrapped replicates that contain a given tree partition. The above analysis was performed using the DendroPy software package (Sukumaran and Holder, 2010) and rapidNJ (Simonsen et al., 2008). Tree visualization was performed using FigTree (2013).

Additional Details Concerning GWAS Enrichment Analysis

To test for Mendelian enrichment in genome-wide association (GWA) results, we first obtained a list of all protein-coding genes that harbor single nucleotide polymorphisms (SNPs). This list was generated by cross-referencing the human gene set maintained by ENSEMBL (Flicek et al., 2011) with the SNP-annotated genes in SCAN (Gamazon et al., 2010). This resulted in a set of 17,341 genes that harbor SNPs within their chromosomal boundaries, 575 of which were included in our list of Mendelian loci (see Table S3). On average, the Mendelian genes were longer and harbored more SNPs than their non-Mendelian counterparts (Mann-Whitney U Test, $p=4.4\times 10^{-6}$ and $p=8.8\times 10^{-8}$ respectively); however, our statistical analysis procedure explicitly accounted for this bias. Specifically, we tested whether Mendelian loci were significantly enriched in GWA signals using a Binomial Test. In our version of the test, a significant signal within a Mendelian gene represented a "success," and we tested whether such successes occurred at a greater rate than would be expected under a null model in which all SNPs were equally likely to be signals. In other words, the probability of a success under the null hypothesis was computed by dividing the number of SNPs in Mendelian genes (177,735) by the total number of SNPs in all protein coding genes (5,008,300). The number of trials for the test was equivalent to the total number of significant signals observed within protein-coding genes. To test for enrichment in the loci specifically indicated by the "Mendelian code," we first tested whether the GWA signals provided for each complex disease were enriched in the precise genes that were linked to their comorbid Mendelian disorders. This was performed using the same Binomial Test outlined above, except that a success was defined as a significant signal in a comorbid gene, the number of trials was given by the number of complex disease-specific protein-coding signals, and the success rate under the null was computed by dividing the number of SNPs contained within the comorbid genes by the total number of SNPs in all protein-coding genes. Unfortunately, these individual tests lacked sufficient power, so we computed the *global* enrichment of GWA signals in comorbid loci by summing the complex disease-specific "successes." The probability mass function for the null model of this statistic (and *p-value*) was computed exactly by taking the convolution of the individual null models.

Computing the Functional Gene Network Similarity Among Mendelian Phenotypes

To evaluate the Mendelian-Mendelian associations within the context of a large, molecular-genetic network, we used HumanNet, an undirected network in which the nodes represent genes and the edges denote functional relationships, each of which is weighted by the log-likelihood ratio of the evidence in favor of a true relationship (Lee et al., 2011). To compute the functional "similarity" between any two genes, we first converted the likelihood-ratios (LogLR) into probabilities according to:

$$P(\text{Gene}_1 \leftrightarrow \text{Gene}_2) = \frac{\exp[\text{LogLR}]}{1+\exp[\text{LogLR}]},$$

where $\leftrightarrow$ indicates a direct functional relationship between two genes. Next, we used Dijkstra's shortest path algorithm (Dijkstra, 1959) to approximate the probability of an *indirect* functional relationship between any two genes:

$$P(\text{Gene}_1 \leftsquigarrow \text{Gene}_2) = \max_{\text{All Paths}} \prod_{\forall \text{Gene}_i \leftrightarrow \text{Gene}_j \in \text{Path}} P(\text{Gene}_i \leftrightarrow \text{Gene}_j),$$

where $\leftsquigarrow$ denotes an indirect relationship. To compute the genetic network functional similarity between any two Mendelian phenotypes, these approximate, indirect association probabilities were combined by computing the probability of *at least one* indirect, functional relationship among the genes underlying each disorder:

$$P(\text{Disease}_A \leftsquigarrow \text{Disease}_B) = 1 - \prod_{\text{Gene}_i \in \mathbb{A},\, \text{Gene}_j \in \mathbb{B}} 1 - P(\text{Gene}_i \leftsquigarrow \text{Gene}_j),$$

where $\mathbb{A}$ and $\mathbb{B}$ denote the sets of genes linked to Disease_A and Disease_B respectively. Finally, the overall functional similarity between two phenotypes was computed using the log-odds that at least one indirect, functional relationship exists between at least one of pair of their associated genes:

$$\text{Log-Odds}\left[\text{Disease}_A \leftsquigarrow \text{Disease}_B\right] = \log \frac{P(\text{Disease}_A \leftsquigarrow \text{Disease}_B)}{1 - P(\text{Disease}_A \leftsquigarrow \text{Disease}_B)}.$$

To compute whether the genes underlying the comorbid Mendelian disorders were more functionally similar than expected, we first summed together the gene network similarities for all of the comorbid Mendelian disease pairs. Then, we randomly shuffled the significantly comorbid pairs 100,000 times and re-computed their overall functional similarity. The *p*-value for the statistic was generated by counting the total number of shuffled datasets that had a functional similarity at least as high as that of the observed.

Mendelian Disease Biological Systems Annotation and Enrichment

We annotated each Mendelian disorder with the set of biological systems affected in patients harboring the disease. To keep the subsequent analysis as

simple as possible, we considered the following set of 17 systems: *Circulatory*, *Otic*, *Respiratory*, *CNS*, *Reproductive*, *Intellectual/Personality*, *Skeletal*, *Immune*, *Integumentary*, *Urinary*, *Metabolic*, *Endocrine*, *Muscular*, *Visual*, *Digestive*, *Growth/Development*, and *Cellular Proliferation/Cancer*. Each disorder was annotated by consulting the National Library of Medicine's *Genetics Home Reference* (2012), Up-to-Date , and the primary literature. The systems annotated to each Mendelian disease are provided in Table S3.

To determine whether the comorbid Mendelian disorders associated with each complex disease were enriched/depleted for particular systems, we first computed enrichment/depletion scores for each complex disorder by summing the number of comorbid Mendelian phenotypes that were annotated with each system, weighted by the relative risks of the associations. To evaluate the significance of the scores, we randomly shuffled the systems annotated to each Mendelian disorder 10,000 times and re-computed the enrichment/depletion scores. A *p*-value for each complex disease was generated by determining the fraction of randomized annotations that resulted in an enrichment/depletion score at least as extreme as the observed value. Multiple testing was controlled using the False Discovery Rate (Benjamini and Hochberg, 1995). The results of this analysis, depicted in Figure S2, are highly consistent with the known pathophysiology of the complex diseases examined in this work.

Statistical Analysis Procedures for Complex-Mendelian Disease Pairs

Contingency Table Analysis

After parsing the clinical records, we first constructed 2×2 contingency tables for all possible complex-by-Mendelian disease pairs. Using these contingency tables, we then computed the following statistics for each pair: the relative risk for the complex disease associated with the Mendelian disorder, the conditional maximum likelihood estimate of the disease comorbidity odds ratio (with 95% confidence interval), and the *p*-value for a null model in which the two diseases occur independently of one another (Fisher's Exact Test). The analyses were performed using custom scripts written in Python and R. We considered a comorbidity relationship to be *initially* significant given that: 1) it passed the 0.05-significance level in *at least one dataset* after a strict Bonferroni correction (i.e. *p*-value × number of tests × number of datasets) and 2) none of the datasets predicted discordant effects (<0.5% of the disease pairs). To account for potentially confounding factors in our analysis, we subjected this set of initially significant complex-Mendelian disease pairs to another round statistical modeling, described in detail in below.

Accounting for the Potentially Confounding Effects of Demographic and Environmental Covariates

In addition to age, gender, and insurance billing codes, the *USA* dataset also provided the county of origin for each patient. We took advantage of this fact and aligned the county-annotated clinical records with information gathered by the United States census (U.S.A. Health Resources and Services Administration, 2013). This allowed us to combine our patient-specific data (age, gender, and phenotype) with a variety of county-level confounding factors, including average per capita income, percent ethnicity (separately for American Indians, Asians, White Hispanics, White non-Hispanics, Black Hispanics, Black non-Hispanics, and Pacific Islanders), and the proportion of various socioeconomic groups (poor, urban, and insurance status). We then incorporated both the patient-specific and county-level data into a single Poisson regression analysis for each initially significant complex-Mendelian disease pair, which allowed us to estimate the relative risks associated with the Mendelian disorders after accounting for a variety of demographic and environmental (i.e. county-level) factors.

Specifically, we modeled the incidence counts for each complex disease within every county of the United States after conditioning on the presence/absence of its comorbid Mendelian partners. Let $y_{i,j,k,l}$ denote the total number patients diagnosed with the complex disease of interest who were treated in county i contained within state j and are of age k, where age is measured in decades and subsumes one of 11 possible values (0-10). The index l indicates whether the comorbid Mendelian disorder is absent (0) or present (1) within this particular population. Finally, let $N_{i,j,k,l}$ denote the total number of patients with these same attributes, regardless of their complex disease status. We modeled the incidence counts $y_{i,j,k,l}$ using the following Poisson mixed-effects regression model:

$$P(y_{i,j,k,l} \mid \lambda_{i,j,k,l}) = \frac{\lambda_{i,j,k,l}^{\,y_{i,j,k,l}} \exp[-\lambda_{i,j,k,l}]}{y_{i,j,k,l}!},$$

$$\lambda_{i,j,k,l} = N_{i,j,k,l} \times \exp\left[\alpha + \mathbf{bX}_{i,j} + \mathbf{mZ}_{i,j}\right]$$

where α denotes the baseline rate for the complex disease, $\mathbf{X}_{i,j}$ and **b** are vectors of fixed effects and their coefficients respectively, and $\mathbf{Z}_{i,j}$ and **m** are random effects and their coefficients respectively.

The specific fixed effects ($\mathbf{X}_{i,j}$) included into the model were: *Mendelian Disease Status* (binary), *Gender* (binary), *Average Per Capita Income* ($11,362-$89,471), *Percent Ethnicity* (separately for American Indian, Asian, White Hispanic, White Non-Hispanic, Black Hispanic, Black Non-Hispanic, and Pacific Islander), *Percent Insured*, *Percent Poor*, and *Percent Urban.* These factors were chosen from a larger superset using standard model selection procedures. The coefficient $\mathbf{b}_0$ corresponds to the fixed effect of the comorbid Mendelian disorder, and therefore, the complex-Mendelian pair was determined to be comorbid if $\mathbf{b}_0$ was found to be significantly greater than 0. The random effects included into the model ($\mathbf{Z}_{i,j}$) were: 1) a single random intercept for every county within each

state and 2) an additional random intercept for every age group (0-11) within each state. An additional state-level random intercept term was initially included as well but was found to have vanishing effects after incorporating the other two variables. The model parameters (α, **b**, and **m**) were fit to the data for each initially significant complex-Mendelian disease pair independently using an approximate maximum likelihood method, see *lme4* R package for details (Bates et al., 2013; R Core Team, 2013).

Integrating the Contingency Table Analysis with the Regression Modeling Output

Given that the confounding factors described above do not significantly contribute to the observed comorbidities, we hypothesized that the Poisson regression modeling results should replicate the simple contingency table analyses. We tested this using the *USA* dataset by comparing the relative risks estimated from the Poisson regression model with those obtained from the contingency tables. We assumed that a comorbidity relationship replicated in the Poisson model given that: 1) the 95% confidence interval for the relative risk estimate did not overlap 0, and 2) the direction of the effect was concordant with the simpler analysis. Overall, 96.7% of the comorbidity relationships that were not specific to a particular gender were replicated by the linear modeling, with 2.2% failing to remain significant and the remaining 1.1% removed due to discordancy. The replication rates for both the male- and female-specific disorders were 88% and 95% respectively. All relationships that failed to replicate in the linear modeling were excluded from downstream analyses. Finally, those initially significant comorbidity relationships that were not detected in USA (<2% of total) were maintained as long as the linear modeling did not predict a significant, discordant effect in the *USA* dataset.

We found that the covariates that we included into the Poisson regression analysis did in fact have statistically significant effects on the complex-Mendelian disease comorbidity risks, but overall, these effects were marginal. The average decrease in relative risk between the mixed effects model and the contingency table analysis was approximately 8%. Because of this slight but significant bias, the relative risks estimated from the Poisson regression models were used in subsequent analyses provided that the association was detected within *USA*; otherwise, the global weighted average of the relative risk estimate was used instead. Table S4 contains a summary of the statistics described above for each complex-Mendelian disease pair that reached initial significance. Figure S1 displays a flow diagram for the statistical procedures implemented during the analysis of the complex-Mendelian disease pairs.

Statistical Analysis Procedures for Mendelian-Mendelian Disease Pairs

Contingency Table Analysis

The first stage of the statistical analysis procedure for the Mendelian-Mendelian disease pairs was conducted by constructing 2×2 contingency tables for all possible pairs. To detect *initially* significant comorbid Mendelian disorders, we computed the following summary statistics for each disease pair: 1) the conditional maximum likelihood estimate of the disease-disease comorbidity odds ratio, 2) the *p*-value for the association according to Fisher's Exact Test, and in place of relative risk, 3) we computed the symmetric standardized shared risk ratio, defined as:

$$SRR = \frac{\text{Observed } \#(M_1,\ M_2)}{\text{Expected } \#(M_1,\ M_2)},$$

where $\#(M_1, M_2)$ denotes the number of patients diagnosed with Mendelian diseases M_1 and M_2 simultaneously. The expected number of patients with both diseases was computed by assuming disease independence and multiplying the product of their marginal incidence rates by the total number of patients in the dataset:

$$\text{Expected } \#(M_1,\ M_2) = f_1 \times f_2 \times T,$$

where T denotes the total number of patients and f_1 (or f_2) denotes the marginal incidence rate for M_1 (or M_2). Consistent with the complex-Mendelian disease pair analysis, we considered a comorbidity relationship to be initially significant given that its Bonferroni-corrected Fisher's Exact Test *p*-value was less than 0.05 and none of the datasets produced discordant results.

Assessing and Accounting for the Effects of Medical Billing Errors

As discussed in the main text, insurance billing errors could potentially create false signals of Mendelian-Mendelian disease comorbidity. Mendelian disorders are rare and many of them share similar symptoms. Thus, medically similar conditions have the potential to be miscoded within clinical records, possibly creating false signals of disease comorbidity. This can happen, for example, when an erroneous code is corrected by a later diagnosis. In the end, it will appear as if the patient was diagnosed with multiple Mendelian diseases when in fact they harbor a single illness. In practice, we found that it was difficult to differentiate diagnostic errors from true Mendelian disease co-occurrences, as there was no direct way to distinguish between the two scenarios. Therefore, we attempted to remove the effects of billing code errors from our datasets by conservatively filtering false-positives using a statistical modeling approach.

According to the diagnostic error hypothesis outlined above, billing codes corresponding to clinically related Mendelian diseases should be miscoded more often than those corresponding to more distinct phenotypes. Therefore, the comorbidity false positive rate should be higher for Mendelian diseases with similar pathophysiology. We measured the similarity between Mendelian diseases using two metrics. First, we computed the distance between disease codes within the ICD9 taxonomy. This taxonomy is organized hierarchically according to pathology, so the distance between two codes in the ICD9 tree

directly reflects the clinical similarity of their corresponding phenotypes. Second, we annotated each Mendelian disorder with the biological systems it affects (see Additional Analysis Procedures for details). We provided two types of annotations for each illness: a primary affected system and a list of secondary affected systems (see Table S3). Most Mendelian diseases are severely debilitating and highly pleiotropic. In other words, they often affect a wide variety of systems, but similar Mendelian diseases tend to share the same primary system. Therefore, we used the shared primary biological system as a predictor of false positives in the Mendelian diagnostic error analysis.

We incorporated the previous two disease similarity metrics into the following simple logistic regression model, which measured their effect on comorbidity detection in the clinical datasets. Let $\Delta_{i,j}$ denote the taxonomical distance between two Mendelian disorders M_i and M_j ($\Delta_{i,j} \in \{1,\ldots,5\}$), where each distance value was treated as a unique factor rather than an integer. Furthermore, let $\Omega_{i,j}$ denote another factor variable that indicated whether the two diseases shared the same primary biological system. $\Omega_{i,j}$ instantiated one of the 17 annotations (biological systems) listed in Table S3 given that two diseases shared a primary system; otherwise, $\Omega_{i,j}$ instantiated the value *Null.* For each disease pair, we constructed a 21-dimensional, binary design vector (17 primary systems plus 4 taxonomical distances), denoted $\mathbf{D}_{i,j}$, such that each element in the vector indicated whether one of the previously defined similarity factors was present. We set $\Delta_{i,j} = 5$ and $\Omega_{i,j} = Null$ to be the baseline values for this model, and thus, $\mathbf{D}_{i,j} = \mathbf{0}$ if a disease pair was completely dissimilar. Finally, let $\mathcal{S}_{i,j} = 1$ indicate that diseases M_i and M_j were detected as comorbid in the initial contingency table analysis. We modeled the probability of this event according to:

$$P(\mathcal{S}_{i,j} = 1 \mid \Delta_{i,j}, \Omega_{i,j}) = \frac{\exp\left[\mathbf{e} \times \mathbf{D}_{i,j}\right]}{1 + \exp\left[\mathbf{e} \times \mathbf{D}_{i,j}\right]},$$

where $\mathbf{e}$ is a vector of parameters indicating the effects associated with the factors in $\mathbf{D}_{i,j}$. As expected, the two measures of Mendelian disease similarity were strongly predictive of comorbidity (Likelihood Ratio Tests, *p*-values < 2.2×10^{-16}). Furthermore, the estimated effects associated with these predictors were rather large. For example, Mendelian disease pairs with the lowest and highest taxonomical distances had a 17-fold difference in the probability of comorbidity detection.

Thus far, we have only considered the most conservative interpretation of this analysis, which is that similar Mendelian diseases were more likely to be miscoded. An alternative to the “miscoding bias” interpretation is that Mendelian phenotypes with similar biology are more likely to be truly comorbid. In fact, we would expect the same result under the genetic modifier hypothesis discussed in

the main text, as Mendelian disease variants with similar biological effects should have a higher probability of apparent genetic interaction. However, because we were unable to disentangle these competing hypotheses, we chose to favor the conservative interpretation and assumed that disease similarity only resulted in miscoding bias. In the remainder of this section, we describe our approach for estimating the rates of Mendelian-Mendelian co-occurrence due to shared biology (and thus miscoding). These "error" rates were then used to conservatively filter perceived biases from the Mendelian-Mendelian comorbidity results.

Ideally, we would have estimated and removed the effects of diagnostic errors on Mendelian disease co-occurrence rates while simultaneously including other potentially confounding covariates, such as ethnicity, socio-economic status, and environment. For the complex-Mendelian disease pairs, the effects associated with the latter set of covariates were estimated using mixed-effects Poisson regression modeling. Unfortunately, adding the shared pathology terms to the Poisson regression model proved intractable, as this necessitated the joint inference of all Mendelian-Mendelian regression models simultaneously. Instead, we estimated the effects of ICD9 taxonomical distance and the shared primary biological systems through an independent Poisson regression procedure. The resulting parameter estimates were then subsequently included as offsets into another round of Poisson regression analysis, which accounted for the same demographic/environmental covariates that were modeled during the complex-Mendelian analysis (see below for more details). We note that this independent modeling approach should be valid as long as the effects of shared pathology and the demographic/environmental covariates are uncorrelated.

We inferred the effects of shared pathology on Mendelian disease co-occurrence rates independently for each of the eight clinical datasets provided in Table 1. The model applied to each dataset was as follows. Let $\mathcal{T}_{i,j}$ denote the 2×2 contingency table for the disease pair consisting of M_i and M_j. For the sake of modeling convenience, we randomly assigned one disorder to be the *response disease* (in this case M_i) and the other disorder to be the *predictor* (M_j), noting that this decision ultimately had no impact on the statistical inference results. Let $y_{i,j}$ denote the number patients with Mendelian disease M_i, and let j index the predictor disease status, which varies over $\{0,1\}$ (indicating the presence/absence of M_j respectively). Finally, let N_j denote the total number of patients in the dataset with the predictor disease status j, regardless of whether they were also diagnosed with M_i. We modeled the conditional counts of the response of disease ($y_{i,j}$) using the following Poisson distribution:

$$P(y_{i,j} \mid \lambda_{i,j}, N_j) = \frac{\left(\lambda_{i,j}\right)^{y_{i,j}} \exp\left[\lambda_{i,j}\right]}{y_{i,j}!}$$

$$\lambda_{i,j} = N_j \times \exp\left[\alpha_i + X_{i,j} \times (\theta_{i,j} + \mathbf{e} \times \mathbf{D}_{i,j})\right],$$

where $\alpha_{i,j}$ denotes the baseline incidence rate for disease M_i, X_j indicates the presence/absence of the predictor disease, $\theta_{i,j}$ indicates its effect on M_i, and $\mathbf{e}$ is a vector of perceived "error" effects due to disease similarity. Thus, the full, conditional probability model for the contingency table $\mathcal{T}_{i,j}$ is:

$$P\left(\mathcal{T}_{i,j} \mid \mathbf{D}_{i,j}, \alpha_{i,j}, \theta_{i,j}, \mathbf{e}\right) \equiv \prod_{j=\{0,1\}} P(y_{i,j} \mid \lambda_{i,j}, N_j).$$

According to this formalism, M_i and M_j are comorbid if $\theta_{i,j}$ is significantly greater than zero. For reference, the previous statistical model can be viewed as a Poisson approximation to the product binomial model used to analyze 2-way contingency tables.

To infer the model from the data, we assumed that each contingency table in the dataset was conditionally independent of the others, resulting in the following likelihood:

$$P(\mathbf{y} \mid \mathbf{D}, \vec{\alpha}, \vec{\theta}, \mathbf{e}) = \prod_{i=1}^{T} \prod_{j=1}^{2} P(y_{i,j} \mid \lambda_{i,j}, N_j),$$

where T denotes the total number of Mendelian-Mendelian disease pairs (4,560 in our study). Generally, inference for Poisson regression models proceeds by obtaining the parameters that maximize the previous function. However, we found that the present model was under-constrained, so we took a Bayesian approach and added another hierarchy. Specifically, we assumed that $\vec{\alpha}$, $\vec{\theta}$, and $\mathbf{e}$ were sampled from the following prior distributions:

$$\alpha_{(M_i,M_j)} \sim \mathcal{N}\left(\mu_\alpha, {\sigma_\alpha}^2\right)$$

$$\theta_{(M_i,M_j)} \sim \mathcal{N}\left(\mu_\theta, {\sigma_\theta}^2\right)$$

$$\mathbf{e} \sim \mathcal{N}\left(\mu_\mathbf{e}, {\sigma_\mathbf{e}}^2\right),$$

where $\mathcal{N}\left(\mu, \sigma^2\right)$ denotes a Gaussian distribution with mean μ and variance σ^2. Furthermore, we assumed that the prior parameters specified above were in turn sampled from the following fixed hyper-priors:

$$\mu_\alpha \sim \mathcal{N}(-8.0, 1.0),$$
$$\mu_\theta \sim \mathcal{N}(0.0, 0.5),$$
$$\mu_\mathbf{e} \sim \mathcal{N}(3.0, 2.0),$$
$$\sigma_\alpha{}^2, \sigma_\theta{}^2, \sigma_\mathbf{e}{}^2 \sim \Gamma^{-1}(5.0, 5.0),$$

where Γ^{-1} denotes the inverse gamma distribution. These hyper-priors were chosen by examining the disease incidence rates in each dataset coupled with

the results of the previously described contingency table and logistic regression analyses. Posterior distributions for the model and prior parameters were inferred through Markov Chain Monte Carlo (3-independent chains, each with 100,000 burn-in iterations, followed by 10,000 samples, thinned by 10 iterations) using the Gibbs sampling procedure outlined in (Doss and Narasimhan, 1994). We checked algorithmic convergence by comparing the inference results obtained from the three independently initialized Markov chains. On simulated data, the choice of hyper-priors had relatively little impact on inference, as long as they were specified within reason (i.e. the hyper-priors assigned non-negligible probability mass to the correct order of magnitude). However, the large size of the clinical record databases precluded a detailed analysis of their effects on the actual datasets.

Whether driven by miscoding errors or genetic interactions, we found that our estimates of the disease similarity effects on the co-incidence rates for Mendelian disorders were remarkably consistent across datasets (average multiplicative effect approximately equal to 4.3). Taking the conservative interpretation, we removed the disease similarity effects from our estimates of the comorbidity odds ratios as follows. For the associations detected within *USA* (approximately 95% of them), we simply included the estimated error effect terms as offsets in another Poisson mixed effects model, which accounted for demographic and environmental covariates (see below for details). For associations that were detected in datasets other than *USA*, we estimated the disease-disease comorbidity odds ratio using the following equation:

$$\text{Odds Ratio}_{i,j} = \frac{\exp\left[\bar{\alpha}_{i,j} + \bar{\theta}_{i,j}\right] / \left(1 - \exp\left[\bar{\alpha}_{i,j} + \bar{\theta}_{i,j}\right]\right)}{\exp\left[\bar{\alpha}_{i,j}\right] / \left(1 - \exp\left[\bar{\alpha}_{i,j}\right]\right)},$$

where $\bar{\alpha}_{i,j}$ and $\bar{\theta}_{i,j}$ denote the Monte Carlo approximations to these parameters.

Accounting for the Potentially Confounding Effects of Demographic and Environmental Covariates

To assess the effects of environmental and demographic factors on Mendelian-Mendelian disease comorbidity, we once again applied a mixed effects Poisson regression model to the disease incidence counts contained within *USA*, as demographic/environmental information could be obtained for this dataset (see *Statistical Analysis Procedures for Complex-Mendelian Disease Pairs* for details). Briefly, for each initially significant Mendelian-Mendelian disease pair, we arbitrarily assigned one disorder to be the response disease and the other to be the predictor, consistent with the procedure described in the previous section. Next, we applied the same mixed effects Poisson regression model that was used to analyze the complex-Mendelian associations to each Mendelian pair. To account for any “perceived” billing code errors, we also included any non-zero effects due to disease similarity (see previous section) as offset terms. Finally, after fitting the models using an approximate maximum likelihood procedure, we

converted the asymmetric relative risk estimates into symmetric odds ratios using the following equation:

$$\text{Odds Ratio}_{i,j} = \frac{\exp\left[\bar{\alpha} + \bar{\mathbf{b}}_0\right] / \left(1 - \exp\left[\bar{\alpha}_{i,j} + \bar{\mathbf{b}}_0\right]\right)}{\exp\left[\bar{\alpha}\right] / \left(1 - \exp\left[\bar{\alpha}\right]\right)},$$

where $\bar{\alpha}$ and $\bar{\mathbf{b}}_0$ denote the expected values for the baseline and predictor disease effect rates respectively (see *Statistical Analysis Procedures for Complex-Mendelian Disease Pairs* for more details). We note that nearly 98% of the initially comorbid Mendelian-Mendelian relationships that were detected in *USA* remained significant after accounting for demographic and environmental covariates.

Integrating the Various Statistical Analysis Procedures

To summarize, our statistical approach for detecting significant Mendelian-Mendelian comorbidity relationships consisted of three stages: 1) an initial analysis that relied on simple contingency tables, 2) a regression analysis which attempted to estimate and remove the effects of Mendelian billing code errors, and 3) an additional regression analysis which included potentially confounding socioeconomic, demographic, and environmental covariates. The latter could only be performed for those relationships detected in *USA* (approximately 95% of the initially comorbid pairings). Ultimately, our integration procedure for the previous three statistical analyses came in two flavors. Our most conservative analysis, which is presented in the main text and in Figure 4, removed all initially significant Mendelian-Mendelian associations in which the odds ratio, after correcting for demographic, geographic, and billing code errors, was significantly lower than the estimate obtained using the contingency table analysis (as determined from the 95% confidence intervals for the corrected estimates). However, we also performed all subsequent analyses (minus the genetic modeling) using a less conservative filtering procedure, which maintained all comorbidity relationships that replicated in the final stage of the analysis, regardless of odd-ratio differences (see Figure S5 and Table S5 for comorbidity results). All subsequent analyses produced qualitatively identical results. Figure S4 displays a flow diagram for the statistical procedures implemented during the analysis of the Mendelian-Mendelian disease pairs.

Genetic Modeling Procedures

Multi-locus Genetic Models for Complex-Mendelian Disease Comorbidity

As a simple illustration, consider a Mendelian phenotype, *M*, and a complex disease, *D*, that are apparently comorbid and share one genetic locus in common, denoted μ. Assume that phenotype *D* follows the additive model and is linked to deleterious variation harbored by *n* loci: [g_1, …, g_{n-1}, μ]. To keep this illustration as simple as possible, we limit the model to include only three types of genotypes: *wild type* (indicated with superscript *w*), *mildly deleterious* (*m*), and

severely deleterious (*s*). The wild-type variation represents the functional norm for each genetic locus, and both severe and mild genotypes harbored by the loci in [g_1, …, g_{n-1}] predispose the bearer to the complex disease *D*. Severe variation in locus μ predisposes carriers to both phenotypes *M* and *D*, but mild variation in μ predisposes the bearer only to the complex phenotype. For mathematical simplicity, we also assume that the population frequencies for these three genotype classes are the same across every gene in the set [g_1, …, g_{n-1}, μ], and we impose the following relationships among the genotype frequencies:

$$0 \le p^s \le p^m \ll p^w \le 1,$$

$$p^s + p^m + p^w = 1.$$

Of course, more general formulations of this model are possible, but they are associated with less transparent equations (see below for details).

Now, assuming complete penetrance of the severely deleterious variants, the apparent prevalence of the Mendelian disease, given either a dominant or recessive inheritance pattern, is:

$$P(\phi = M \mid \Theta) = p_S = \begin{cases} 1 - \left[1 - f_S\right]^2, & M \text{ is dominant,} \\ \left[f_S\right]^2, & M \text{ is recessive,} \end{cases}$$

where f_S is the population frequency of severely deleterious alleles. If we also invoke complete penetrance for the genotypes linked to the complex disease, its prevalence is given by:

$$P(\phi = D \mid \text{additive}, \Theta) = 1 - \left[1 - p^s - p^m\right]^{2n}.$$

Consequently, the expected frequency of individuals in the population that are affected by both diseases is simply:

$$P(\phi = D \wedge M \mid \text{additive}, \Theta) = P(\phi = M \mid \Theta).$$

The standardized shared risk ratio for *D* and *M*, defined as their observed joint prevalence divided by their expected joint prevalence given independence, is:

$$SRR_{additive} = \frac{P(\phi = D \wedge M \mid \Theta)}{P(\phi = D \mid \Theta) P(\phi = M \mid \Theta)} = \frac{1}{P(\phi = D \mid \Theta)}.$$

Therefore, under this simple model, the shared risk ratios between complex and Mendelian disorders should follow the inverse frequency of the complex disease. Of course, this result is not consistent with the observed data, which could be due to a variety of reasons, including the oversimplified assumptions upon which it relies.

Under the two-community model, a complex disease *D* transpires if and only if both gene communities harbor *penetrant,* deleterious variation. Furthermore, the community-specific penetrance functions can be chosen from the full range of

multi-locus genetic models: additive, multiplicative, threshold, and many others (Risch, 1990). For the purpose of our illustration, the simplest such choice is a per-community additive model, which was defined in the Experimental Procedures. Given this assumption, the disease prevalence equation for the combinatorial model is very similar to that specified for the additive model, with the exception of an additional term that enforces the requirement that two communities of loci must be affected by deleterious variation simultaneously in order for the complex disease to occur. Let $\kappa_{\mathcal{C}}$ denote the total number of mildly or severely deleterious locus-specific genotypes harbored by the $\mathcal{C}$ community. According to the assumptions made in the previous paragraph,

$$P(\kappa_{\mathcal{C}} \geq 1) = 1 - \left[1 - p_{\mathcal{C}}^{s} - p_{\mathcal{C}}^{m}\right]^{2n_{\mathcal{C}}}.$$

Assuming complete penetrance, the probability that both communities are affected by deleterious variants is:

$$P(\varphi = D \mid \text{combinatorial}, \Theta) = \prod_{\mathcal{C}=\{\square,\bigcirc\}} P(\kappa_{\mathcal{C}} \geq 1 \mid \Theta)$$

$$= \prod_{\mathcal{C}=\{\square,\bigcirc\}} \left(1 - \left[1 - p_{\mathcal{C}}^{s} - p_{\mathcal{C}}^{m}\right]^{2n_{\mathcal{C}}}\right),$$

which is equivalent to the expected prevalence of the complex disease in the population.

As with the additive model, we account for the co-occurrence of Mendelian (*M*) and complex (*D*) diseases by assuming that *M* and *D* share one locus in common, μ, and that this locus belongs to one of the two communities (say □). According to the combinatorial model, the joint prevalence of the two diseases and their standardized shared risk ratio are given by the following two equations:

$$P(\varphi = D \wedge M \mid \text{combinatorial}, \Theta) = P(\varphi = M \mid \Theta) \times \left(1 - \left[1 - p_{\bigcirc}^{s} - p_{\bigcirc}^{m}\right]^{2n_{\bigcirc}}\right).$$

$$SRR_{combinatorial} = \frac{P(\phi = D \wedge M \mid \Theta)}{P(\phi = D \mid \Theta) P(\phi = M \mid \Theta)} = \frac{1 - \left[1 - p_{\bigcirc}^{s} - p_{\bigcirc}^{m}\right]^{2n_{\bigcirc}}}{P(\phi = D \mid \Theta)}.$$

Importantly, the risk ratio under the combinatorial model is always less than or equal to the inverse of the complex disease prevalence, in contrast to the heterogeneity model. Therefore, in principle, these two models could be distinguished by examining shared risk in our actual dataset, although a more rigorous analysis requires the introduction of additional modeling complexities.

The Additive and Combinatorial Models Under More General Assumptions

We can formulate more general versions of the two models described in the main text and in the previous section, at the expense of more complicated equations. However, the more general assumptions allow us to make more realistic inferences from the clinical datasets. Below, we extend the previous two models to include general genotype frequencies and penetrance parameters. In subsequent sections, we demonstrate how these more general assumptions allow us to make novel and interesting inferences from the clinical record datasets.

To begin, we assume that the genotype at the i^{th} *locus* in both models is either harmful ($I_i = 1$) or normal ($I_i = 0$). Therefore, the genetic model at locus *i* can be any one of the following: dominant ($I_i = 1$if $G_i = \{Aa, aA, aa\}$), recessive ($I_i = 1$ if $G_i = \{aa\}$), or haploid ($I_i = 1$ if $G_i = \{a\}$). Importantly, the penetrance of the harmful genotypes is allowed to be incomplete. Consistent with previous notation, we use $F(G)$ to indicate the frequency of the genotype in the population and $W_D(G)$ (or $W(G)$ when the disease of interest in unambiguous) to indicate the penetrance of genotype *G* with respect to disease *D*. Under the additive model (the n loci contribute independently to disease risk), the joint probability of the complex disease D and its underlying genotype is:

$$\begin{aligned} P(\varphi = D, g = G \mid \text{Additive}, \Theta) &= F(G) \times W(G) \\ &= \prod_{G_i \in G} F(G_i) \times \left[1 - \prod_{G_j \in G} [1 - W(G_j)] \right] \\ &= \prod_{I_i \in \mathbf{I}} p_i^{I_i} (1 - p_i)^{1 - I_i} \times \left[1 - \prod_{I_j \in \mathbf{I}} [1 - x_j \delta(I_j = 1)] \right], \end{aligned}$$

where $F(G_i) = p_i < 1$ is the frequency of the harmful genotype(s), $W(G_i) = x_i > 0$ if $I_i = 1$ and $W(G_i) = 0$ otherwise, and the vector Θ contains all parameters defined in the model (i.e. $\Theta = \{p_i, x_i\} \forall I_i \in \mathbf{I}$). Similarly, the joint phenotype-genotype probability under the two-community combinatorial model is:

$$\begin{aligned} P(\varphi = D, g = G \mid \text{Combinatorial}, \Theta) &= \prod_{\mathcal{C} = \{\square, \bigcirc\}} F(G_{\mathcal{C}}) \times W(G_{\mathcal{C}}) \\ &= \prod_{\mathcal{C} = \{\square, \bigcirc\}} \left(\prod_{G_i \in G_{\mathcal{C}}} F(G_i) \times \left[1 - \prod_{G_j \in G_{\mathcal{C}}} [1 - W(G_j)] \right] \right) \\ &= \prod_{\mathcal{C} = \{\square, \bigcirc\}} \left(\prod_{I_i \in \mathbf{I}_{\mathcal{C}}} p_i^{I_i} (1 - p_i)^{1 - I_i} \times \left[1 - \prod_{I_j \in \mathbf{I}_{\mathcal{C}}} [1 - x_j \delta(I_j = 1)] \right] \right). \end{aligned}$$

Consistent with previous notation, one community in the model is indicated using a square, and the other is indicated with a circle. In the equation defined above, we have assumed that functional perturbation of both communities deterministically causes the complex disease. We could relax this assumption by including an overall two-community penetrance parameter, but this assumption

would simply linearly rescale the per-loci parameters, negating its utility. Furthermore, although we do not explicitly include environmental factors in the previous two models, they could be introduced as additional "loci" that represent environmental insults.

The previous joint phenotype-genotype probability equations can be further simplified by expressing them in terms of the underlying distributions over the penetrance and genotype frequency parameters. First, assume that these parameters are each drawn identically and independently from two distinct Beta densities, with each p_i following a beta-distribution with parameters a and b, and each x_j following a beta-distribution with parameters α and β:

$$P(p_i \mid a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p_i^{a-1}(1-p_i)^{b-1}$$
$$P(x_j \mid \alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x_j^{\alpha-1}(1-x_j)^{\beta-1}.$$

By multiplying the joint phenotype-genotype prevalence equations by the beta densities and rearranging terms, we obtain the following probability density over genotype, phenotype, and model parameters:

$$P(\varphi = D, g = G, \vec{x}, \vec{p} \mid \text{Additive}, \Theta) =$$
$$\prod_{I_i \in \mathbf{I}} p_i^{I_i}(1-p_i)^{1-I_i} P(p_i \mid a,b) \times \left[1 - \prod_{I_j \in \mathbf{I}} [1 - x_j \delta(I_j = 1)] P(x_j \mid \alpha,\beta) \right].$$

To express the previous equation in terms of the hyper-parameters a, b, α and β only, we simply integrate the previous equation over the penetrance and genotype frequency parameters as follows:

$$P(\varphi = D, g = G \mid \text{Additive}, \Theta) =$$
$$\prod_{I_i \in \mathbf{I}} \int_0^1 p_i^{I_i}(1-p_i)^{1-I_i} P(p_i \mid a,b)\, dp_i \times \left[1 - \prod_{I_j \in \mathbf{I}} \int_0^1 [1 - x_j \delta(I_j = 1)] P(x_j \mid \alpha,\beta)\, dx_j \right].$$

This equation is further simplified by noting each integral corresponds to the expectation of a Beta density, resulting in:

$$P(\varphi = D, g = G \mid \text{Additive}, \Theta)$$
$$= \prod_{I_i \in \mathbf{I}} \left[\frac{a}{(a+b)} \right]^{I_i} \left[1 - \frac{a}{(a+b)} \right]^{1-I_i} \times \left[1 - \prod_{I_j \in \mathbf{I}} \left(1 - \left(\frac{\beta}{\alpha+\beta} \right) \delta(I_j = 1) \right) \right]$$
$$= \prod_{I_i \in \mathbf{I}} \langle p \rangle^{I_i} \left[1 - \langle p \rangle \right]^{1-I_i} \times \left[1 - \prod_{I_j \in \mathbf{I}} \left(1 - \langle x \rangle \delta(I_j = 1) \right) \right],$$

where $\langle p \rangle = \frac{a}{a+b}$ and $\langle x \rangle = \frac{\alpha}{\alpha+\beta}$ denote the expected values for the genotype frequency and penetrance distributions respectively.

At this point, each genomic locus in the additive model is indistinguishable from the others (as each locus-specific term in the previous equation is identical), and therefore, we rewrite the joint genotype-phenotype frequency in terms of the number of deleterious variants present, denoted k, such that:

$$P(\varphi = D, g = G \mid \text{Additive}, \Theta) =$$
$$P(\varphi = D, g' = k \mid \text{Additive}, \Theta) =$$
$$\binom{n}{k} \langle p \rangle^k \left[1 - \langle p \rangle\right]^{n-k} \left[1 - \left(1 - \langle x \rangle\right)^k\right],$$

where the binomial term accounts for the number of ways to select k loci from the total pool of n possible. Finally, the frequency of the complex disease, independent of genotype, can be derived by marginalizing previous equation over all possible genotypes capable of producing disease, resulting in:

$$P(\varphi = D \mid \text{Additive}, \Theta) = \sum_{k=1}^{n} \binom{n}{k} \langle p \rangle^k \left[1 - \langle p \rangle\right]^{n-k} \left[1 - (1 - \langle x \rangle)^k\right].$$

The same logic outlined above can be applied to the combinatorial model, as an independent additive model governs whether each community is affected by deleterious genetic variation. Thus, the overall prevalence of the complex disease given the assumptions underlying the combinatorial model is:

$$P(\varphi = D \mid \text{Combinatorial}, \Theta) = \prod_{\mathcal{C}=\{\square,\bigcirc\}} \left(\sum_{k=1}^{n_{\mathcal{C}}} \binom{n_{\mathcal{C}}}{k} \langle p \rangle^k \left[1 - \langle p \rangle\right]^{n_{\mathcal{C}}-k} \left[1 - (1 - \langle x \rangle)^k\right] \right),$$

where $n_{\mathcal{C}}$ denotes the number of loci in community $\mathcal{C} \in \{\square,\bigcirc\}$.

Complex Disease Risk in Patients with Multiple Mendelian Disorders

Thus far, we have demonstrated how the two genetic models outlined in the main text can be used to account for complex-Mendelian disease comorbidity, and we specified their predictions for complex disease risk under more general assumptions. Below, we combine the previous two results and specify the models' predicted marginal risks for complex disease conditional on the co-occurrence of comorbid Mendelian disorders. This derivation allows us to assess the likelihood of each model by applying them to a rather novel (although not unheard of) phenomenon: the appearance of multiple comorbid Mendelian diseases within individual patients.

Let $\mathcal{M}_K = \{M_1, M_2 \ldots M_{K-1}, M_K\}$ denote the set of K Mendelian disorders harbored by some patient, and assume each of them can predispose this individual to some complex disease D. For the sake of simplicity, assume that each Mendelian disorder maps to a single locus and that none of them share the same locus. Therefore, if a patient harbors the Mendelian disease set $\mathcal{M}_K$, then we assume that this individual also harbors at least K deleterious genetic variants (one for each of the K observed diseases plus any additional unobserved variants). Consistent with the notation used in the previous section,

our goal is to specify the following two probabilities: $P(\varphi = D \mid \mathcal{M}_K, \text{Additive}, \Theta)$ and $P(\varphi = D \mid \mathcal{M}_K, \text{Combinatorial}, \Theta)$. For the additive genetic model, this derivation is fairly straightforward.

Given an observed set of K Mendelian disorders and the additive genetic modeling assumptions outlined above, there are only three ways that an individual can acquire the complex disease:

1) The patient can acquire the disease as a result of Mendelian disorder variation in his/her genome only (red)
2) The patient can acquire the disease due to variation in both Mendelian and non-Mendelian loci (green)
3) The patient can acquire the disease due to variation in non-Mendelian loci only (blue)

In mathematical terms, the previous three conditions are expressed as follows:

$$
\begin{aligned}
&P(\varphi = D \mid \mathcal{M}_K, \text{Additive}, \Theta) = \\
&\left[1-\left(1-\langle x\rangle\right)^K\right] \times \left[1-\sum_{k=1}^{n'} \langle p\rangle^k \left(1-\langle p\rangle\right)^{n'-k} \left[1-\left(1-\langle x\rangle\right)^k\right]\right] \\
&\ldots + \left[1-\left(1-\langle x\rangle\right)^K\right] \times \left[\sum_{k=1}^{n'} \langle p\rangle^k \left(1-\langle p\rangle\right)^{n'-k} \left[1-\left(1-\langle x\rangle\right)^k\right]\right] \\
&\ldots + \left(1-\langle x\rangle\right)^K \times \left[\sum_{k=1}^{n'} \langle p\rangle^k \left(1-\langle p\rangle\right)^{n'-k} \left[1-\left(1-\langle x\rangle\right)^k\right]\right] \\
&= \left[1-\left(1-\langle x\rangle\right)^K\right] + \left(1-\langle x\rangle\right)^K \times \left[\sum_{k=1}^{n'} \langle p\rangle^k \left(1-\langle p\rangle\right)^{n'-k} \left[1-\left(1-\langle x\rangle\right)^k\right]\right],
\end{aligned}
$$

where n' denotes the additional number of loci (excluding the Mendelian loci under consideration) associated with the complex disease ($n' = n - K$).

For the two-community combinatorial model, the derivation of the probability $P(\varphi = D \mid \mathcal{M}_K, \text{Combinatorial}, \Theta)$ proceeds similarly, although with an additional complication. As before, we assume that a patient with the disease set $\mathcal{M}_K$ harbors K deleterious variants in Mendelian loci; however, we do not know the community assignments for these variants. Therefore, we introduce an additional set of latent variables (one for each of the K "observed" Mendelian variants), denoted $\mathbf{Z}$, where each $Z_j \in \mathbf{Z}$ indicates whether a variant belongs to the first ($Z_j = \square$) or second ($Z_j = \bigcirc$) community. Finally, for the sake of simplicity, we assume that the remaining n' deleterious variants are evenly divided between the two communities (although this assumption could be relaxed at the expense of additional computation):

$$n'_{\square} = \left\lfloor \frac{n'}{2} \right\rfloor$$
$$n'_{\bigcirc} = \left\lceil \frac{n'}{2} \right\rceil.$$

As noted in the previous section, the combinatorial model is simply the product of two additive models, each of which governs the affected status of a different community. With this in mind, let $\nu_{\mathcal{C}} = 1$ denote that community $\mathcal{C}$ is affected by deleterious genetic variation. The probability of the complex disease under the combinatorial model, conditional on the Mendelian disease set $\mathcal{M}_K$ and the community assignment indicators $\mathbf{Z}$, is:

$$P(\varphi = D \mid \mathbf{Z}, \mathcal{M}_K, \text{Combinatorial}, \Theta) = \prod_{\mathcal{C}=\{\square,\bigcirc\}} P\left(\nu_{\mathcal{C}} = 1 \mid \mathbf{Z}, \mathcal{M}_K, \Theta\right)$$

$$P\left(\nu_{\mathcal{C}} = 1 \mid \mathbf{Z}, \mathcal{M}_K, \Theta\right) =$$

$$\left[1 - \prod_{j=1}^{K}\left(1 - \delta\left(Z_j = \mathcal{C}\right)\langle x\rangle\right)\right] + \prod_{j=1}^{K}\left(1 - \delta\left(Z_j = \mathcal{C}\right)\langle x\rangle\right) \times \left[\sum_{k=1}^{n'_{\mathcal{C}}} \langle p\rangle^k \left(1 - \langle p\rangle\right)^{n'_{\mathcal{C}}-k}\left[1 - \left(1 - \langle x\rangle\right)^k\right]\right],$$

where $\delta\left(Z_j = \mathcal{C}\right)$ is an indicator function which returns 1 given that the community assignment of the jth locus is equivalent to $\mathcal{C}$.

Ideally, we would express the previous probability without conditioning on the community assignments of the Mendelian loci, as these are not known *a priori*. In the case of a single patient, this turns out to be relatively simple. First, after integrating the locus-specific genotype frequency and penetrance parameters out of the model, the K Mendelian disease loci are statistically indistinguishable from one another. Therefore, we introduce two new variables, denoted $K_{\square}$ and $K_{\bigcirc}$, which indicate the number of "observed" Mendelian loci assigned to each community:

$$K_{\square} = \sum_{j=1}^{K} \delta\left(Z_j = \square\right)$$
$$K_{\bigcirc} = \sum_{j=1}^{K} \delta\left(Z_j = \bigcirc\right).$$

This allows us to rewrite the probability for the complex disease as:

$$P(\varphi = D \mid \mathcal{M}_K, \text{Combinatorial}, \Theta) =$$

$$P(\varphi = D \mid K_{\square}, K_{\bigcirc}, \text{Combinatorial}, \Theta) = \prod_{\mathcal{C}=\{\square,\bigcirc\}} P\left(\nu_{\mathcal{C}} = 1 \mid K_{\mathcal{C}}, \Theta\right),$$

$$P\left(\nu_{c} = 1 \mid K_{\mathcal{C}}, \Theta\right) = \left[1 - \left(1 - \langle x\rangle\right)^{K_{\mathcal{C}}}\right] + \left(1 - \langle x\rangle\right)^{K_{\mathcal{C}}} \times \left[\sum_{k=1}^{n'_{\mathcal{C}}} \langle p\rangle^k \left(1 - \langle p\rangle\right)^{n'_{\mathcal{C}}-k}\left[1 - \left(1 - \langle x\rangle\right)^k\right]\right].$$

We can remove the dependence on $K_{\square}$ and $K_{\bigcirc}$ by multiplying the previous equation by a prior distribution for the community assignments, denoted $P(K_{\square},K_{\bigcirc}\,|\,\mathcal{M}_K)$, and summing over all possible assignments $\mathcal{K}$:

$$P(\varphi = D\,|\,\mathcal{M}_K,\text{Combinatorial},\Theta) = \sum_{\{K_{\square},K_{\bigcirc}\}\in\mathcal{K}} P(K_{\square},K_{\bigcirc}\,|\,\mathcal{M}_K)\prod_{\mathcal{C}=\{\square,\bigcirc\}} P\left(\nu_{\mathcal{C}}=1\,|\,K_{\mathcal{C}},\Theta\right).$$

For simplicity, we assume a uniform prior:

$$P(K_{\square},K_{\bigcirc}\,|\,\mathcal{M}_K) = 2^{-K}\ ,$$

and therefore,

$$P(\varphi = D\,|\,\mathcal{M}_K,\text{Combinatorial},\Theta) = \sum_{\{K_{\square},K_{\bigcirc}\}\in\mathcal{K}} 2^{-K}\prod_{\mathcal{C}=\{\square,\bigcirc\}} P\left(\nu_{\mathcal{C}}=1\,|\,K_{\mathcal{C}},\Theta\right).$$

Because each of the K variants are statistically indistinguishable and $K_{\bigcirc} = K - K_{\square}$, the previous summation can be rewritten as:

$$P(\varphi = D\,|\,\mathcal{M}_K,\text{Combinatorial},\Theta) = 2^{-K}\times\sum_{K_{\square}=0}^{K}\begin{pmatrix} K \\ K_{\square}\end{pmatrix} P\left(\nu_{\mathcal{C}}=1\,|\,K_{\square},\Theta\right)P\left(\nu_{\mathcal{C}}=1\,|\,K-K_{\square},\Theta\right),$$

where the binomial coefficient accounts for the number elements in $\mathcal{K}$ that are consistent with $K = K_{\square} + K_{\bigcirc}$.

To summarize, the following two equations provide the probability for the complex disease, conditional on its co-occurrence with K Mendelian disorders, under the additive and combinatorial genetic models:

$$P(\varphi = D\,|\,\mathcal{M}_K,\text{Additive},\Theta) =$$

$$\left[1-\left(1-\langle x\rangle\right)^K\right]+\left(1-\langle x\rangle\right)^K\times\left[\sum_{k=1}^{n'}\langle p\rangle^k\left(1-\langle p\rangle\right)^{n'-k}\left[1-\left(1-\langle x\rangle\right)^k\right]\right]$$

$$P(\varphi = D\,|\,\mathcal{M}_K,\text{Combinatorial},\Theta) = 2^{-K}\times\sum_{K_{\square}=0}^{K}\begin{pmatrix} K \\ K_{\square}\end{pmatrix} P\left(\nu_{\mathcal{C}}=1\,|\,K_{\square},\Theta\right)P\left(\nu_{\mathcal{C}}=1\,|\,K-K_{\square},\Theta\right),$$

$$P\left(\nu_{c}=1\,|\,K_{\mathcal{C}},\Theta\right) = \left[1-\left(1-\langle x\rangle\right)^{K_{\mathcal{C}}}\right]+\left(1-\langle x\rangle\right)^{K_{\mathcal{C}}}\times\left[\sum_{k=1}^{n'_{\mathcal{C}}}\langle p\rangle^k\left(1-\langle p\rangle\right)^{n'_{\mathcal{C}}-k}\left[1-\left(1-\langle x\rangle\right)^k\right]\right].$$

Importantly, these two functions make strikingly different predictions with respect to the probability of D as a function of the number comorbid Mendelian disorders. To see this, assume that the average penetrance for each variant is relatively small ($\langle x\rangle << 1$). Given this assumption, it is easy to demonstrate (Risch, 1990) that the probability for the complex disease under the additive model can be approximated as:

$$P(\varphi = D \mid \mathcal{M}_K, \text{Additive}, \Theta) =$$

$$\left[1-\left(1-\langle x\rangle\right)^K\right]+\left(1-\langle x\rangle\right)^K \times \left[\sum_{k=1}^{n'}\langle p\rangle^k\left(1-\langle p\rangle\right)^{n'-k}\left[1-\left(1-\langle x\rangle\right)^k\right]\right]$$

$$\approx K \times \langle x\rangle + R.$$

In natural language, the previous equation states that the average risk for the disease should approximately increase according to a linear function of the number of comorbid Mendelian phenotypes. Conversely, if we invoke this same approximation for the combinatorial model, we obtain:

$$P(\varphi = D \mid \mathcal{M}_K, \text{Combinatorial}, \Theta) = 2^{-K} \times \sum_{K_\square=0}^{K}\begin{pmatrix} K \\ K_\square \end{pmatrix} P\left(v_C = 1 \mid K_\square, \Theta\right) P\left(v_C = 1 \mid K - K_\square, \Theta\right)$$

$$\approx 2^{-K} \times \sum_{K_\square=0}^{K}\begin{pmatrix} K \\ K_\square \end{pmatrix}\left[K_\square \times \langle x\rangle + R\right] \times \left[\left(K - K_\square\right) \times \langle x\rangle + R\right].$$

This indicates that the combinatorial model predicts a polynomial increase in disease risk as a function of K. In the main text, we illustrate that most of the complex diseases that we examined (barring, for example, Type I and Type II Diabetes) appear to display a super-linear increase in average disease risk as a function of the number of comorbid Mendelian disorders. This result suggests that the combinatorial model more realistically captures disease risk in patients with compound Mendelian variants. To formally test this hypothesis, we performed Bayesian model selection to determine which genetic model best accounted for the variation in complex disease risk observed across patients with differing numbers of comorbid Mendelian disorders. This approach is described in greater detail below.

Applying the Genetic Models to Clinical Data: A Bayesian Approach

As discussed above (and in the main text), we observed that the combinatorial and additive genetic models make distinct predictions with respect to the effects of compound deleterious Mendelian variants on complex disease risk. To formally test which model best accounts for the disease risks observed within the clinical datasets, we took a Bayesian approach and computed the posterior probability for each model conditional on the complex and Mendelian disease incidence counts harvested from the patient records. This procedure involved three steps: 1) specifying the likelihoods for the patient data under each model, 2) integrating these likelihoods over the unknown model parameters, and 3) using the resulting marginal likelihoods to compute the desired posterior probabilities.

Let $\vec{\varphi}$ denote the complex disease status for all T patients contained within the clinical record dataset, where $\varphi_i = D$ ($\varphi_i \neq D$) indicates that the ith patient is (is not) diagnosed with complex disease D. Let $\vec{K}$ denote a vector containing the number of comorbid Mendelian disorders assigned to each patient. Our goal is

specify the likelihood of $\vec{\varphi}$ conditioned on the number of comorbid Mendelian disorders harbored by each patient, denoted $P(\vec{\varphi} \mid \vec{K}, \text{Additive}, \Theta)$ and $P(\vec{\varphi} \mid \vec{K}, \text{Combintorial}, \Theta)$ for the additive and combinatorial models respectively. In the present analysis, $\vec{\varphi}$ included disease incidence counts for all patients contained within the datasets *USA* and *DK*, as these were very unlikely to contain overlapping information.

With respect to the additive model, the likelihood is fairly straightforward. Let $\delta(\varphi_i = D)$ denote an indicator function which returns 1 if patient i has the complex disease and 0 otherwise. With this notation in place, the likelihood for the patient data is simply:

$$P(\vec{\varphi} \mid \vec{K}, \text{Additive}, \Theta) =$$

$$\prod_{i=1}^{T} \left[P(\varphi_i = D \mid K_i, \text{Additive}, \Theta) \right]^{\delta(\varphi_i = D)} \times \left[1 - P(\varphi_i = D \mid K_i, \text{Additive}, \Theta) \right]^{1-\delta(\varphi_i = D)},$$

where the $P(\varphi_i = D \mid K_i, \text{Additive}, \Theta)$ is equivalent to $P(\varphi = D \mid \mathcal{M}_{K_i}, \text{Additive}, \Theta)$, as defined in the previous section. The above likelihood can be further simplified by noting that the per-patient likelihood terms are identical among individuals with the same number of comorbid Mendelian diseases. Let $T(K)$ denote the total number of patients with K comorbid Mendelian phenotypes, and let $D(K)$ denote the number of these patients who also suffer from the complex disease D. The previous likelihood can be concisely rewritten as:

$$P(\vec{\varphi} \mid \vec{K}, \text{Additive}, \Theta) =$$

$$\prod_{k=1}^{K_{max}} \left[P(\varphi = D \mid k, \text{Additive}, \Theta) \right]^{D(K)} \times \left[1 - P(\varphi = D \mid k, \text{Additive}, \Theta) \right]^{T(K) - D(K)},$$

where K_{max} is largest number of comorbid Mendelian disorders harbored by a single patient within the dataset.

The corresponding likelihood equation for the combinatorial genetic model is a bit more complicated. Because the community assignments of the comorbid Mendelian disorders should be the same for all of the patients, this likelihood must first be expressed conditional on the community assignments of the Mendelian disease loci. Reusing notation from the previous section, let $\mathbf{Z}$ denote the community assignments for *all* Mendelian loci that are comorbid with complex disease D. The data likelihood, conditional on these community assignments, is:

$$P(\vec{\varphi} \mid \mathbf{Z}, \vec{K}, \text{Combintorial}, \Theta) =$$

$$\prod_{i=1}^{T} \left[\prod_{\mathcal{C}=\{\square,\bigcirc\}} P\left(\nu_{\mathcal{C}} = 1 \mid \mathbf{Z}, K_i, \Theta\right) \right]^{\delta(\varphi_i = D)} \times \left[1 - \prod_{\mathcal{C}=\{\square,\bigcirc\}} P\left(\nu_{\mathcal{C}} = 1 \mid \mathbf{Z}, K_i, \Theta\right) \right]^{1-\delta(\varphi_i = D)},$$

where $P\left(\nu_{\mathcal{C}} = 1 \mid \mathbf{Z}, K_i, \Theta\right)$ is equivalent to $P\left(\nu_{\mathcal{C}} = 1 \mid \mathbf{Z}, \mathcal{M}_{K_i}, \Theta\right)$. Previously, we removed the dependence of the combinatorial model on the community

assignments $\mathbf{Z}$ by marginalizing them out of the probability distribution. This can be performed once again by applying the same uniform prior. Let L denote the number of Mendelian disease loci associated with the disorder D, and let $P(\mathbf{Z} \mid L)$ represent our prior distribution for the community assignments of these loci, where:

$$P(\mathbf{Z} \mid L) = 2^{-L} \ .$$

Plugging this prior into the previously defined likelihood and marginalizing over all possible community assignments (denoted $\mathcal{Z}$) yields:

$$P(\vec{\varphi} \mid \vec{K}, \text{Combintorial}, \Theta) =$$

$$\sum_{\mathbf{Z} \in \mathcal{Z}} P(\mathbf{Z} \mid L) \times \prod_{i=1}^{T} \left[\prod_{C=\{\square,\bigcirc\}} P\left(v_C = 1 \mid \mathbf{Z}, K_i, \Theta\right) \right]^{\delta(\varphi_i = D)} \times \left[1 - \prod_{C=\{\square,\bigcirc\}} P\left(v_C = 1 \mid \mathbf{Z}, K_i, \Theta\right) \right]^{1-\delta(\varphi_i = D)}$$

$$= \sum_{\mathbf{Z} \in \mathcal{Z}} 2^{-L} \times \prod_{i=1}^{T} \left[\prod_{C=\{\square,\bigcirc\}} P\left(v_C = 1 \mid \mathbf{Z}, K_i, \Theta\right) \right]^{\delta(\varphi_i = D)} \times \left[1 - \prod_{C=\{\square,\bigcirc\}} P\left(v_C = 1 \mid \mathbf{Z}, K_i, \Theta\right) \right]^{1-\delta(\varphi_i = D)} .$$

Unlike in the previous section, this marginalization requires the summation over 2^L different assignments, rendering it computationally intractable. As an approximation, we instead performed the marginalization independently for each patient, which results in following, approximate likelihood:

$$P(\vec{\varphi} \mid \vec{K}, \text{Combintorial}, \Theta) =$$

$$\sum_{\mathbf{Z} \in \mathcal{Z}} P(\mathbf{Z} \mid L) \times \prod_{i=1}^{T} \left[\prod_{C=\{\square,\bigcirc\}} P\left(v_C = 1 \mid \mathbf{Z}, K_i, \Theta\right) \right]^{\delta(\varphi_i = D)} \times \left[1 - \prod_{C=\{\square,\bigcirc\}} P\left(v_C = 1 \mid \mathbf{Z}, K_i, \Theta\right) \right]^{1-\delta(\varphi_i = D)}$$

$$\approx \prod_{i=1}^{T} \left[\sum_{\mathbf{Z} \in \mathcal{Z}} P(\mathbf{Z} \mid L) \prod_{C=\{\square,\bigcirc\}} P\left(v_C = 1 \mid \mathbf{Z}, K_i, \Theta\right) \right]^{\delta(\varphi_i = D)} \times \ldots$$

$$\ldots \left[1 - \sum_{\mathbf{Z} \in \mathcal{Z}} P(\mathbf{Z} \mid L) \prod_{C=\{\square,\bigcirc\}} P\left(v_C = 1 \mid \mathbf{Z}, K_i, \Theta\right) \right]^{1-\delta(\varphi_i = D)} .$$

The marginal probabilities contained within each bracket are equivalent to the per-patient complex disease probabilities derived in the previous section, and thus, the approximate likelihood for the data can be concisely written as:

$$P(\vec{\varphi} \mid \vec{K}, \text{Combinatorial}, \Theta) \approx$$

$$\prod_{i=1}^{T} \left[P(\varphi_i = D \mid K_i, \text{Combinatorial}, \Theta) \right]^{\delta(\varphi_i = D)} \times \left[1 - P(\varphi_i = D \mid K_i, \text{Combinatorial}, \Theta) \right]^{1-\delta(\varphi_i = D)}$$

$$= \prod_{k=1}^{K_{max}} \left[P(\varphi = D \mid k, \text{Combinatorial}, \Theta) \right]^{D(K)} \times \left[1 - P(\varphi = D \mid k, \text{Combinatorial}, \Theta) \right]^{T(K) - D(K)},$$

where $P(\varphi_i = D \mid K_i, \text{Combinatorial}, \Theta)$ is equivalent to $P(\varphi = D \mid \mathcal{M}_{K_i}, \text{Combinatorial}, \Theta)$. Although imperfect, we found that the previous approximation worked very well on simulated data, introducing an unbiased error that was always less than 1% of the total likelihood while requiring several orders of magnitude less computational time.

With the data likelihood for each model fully defined, it is straightforward to specify the desired marginal likelihoods. Let $P(\Theta)$ denote a prior distribution over the parameter set Θ, where $\Theta = \{\langle p \rangle, \langle x \rangle\}$ with respect to both the combinatorial and additive genetic models. In practice, we used independent, non-informative beta densities as the prior distributions for both $\langle p \rangle$ and $\langle x \rangle$. The marginal likelihoods for the models were computed by multiplying the data likelihoods defined above by the prior distributions followed by integrating the resulting product over all possible values:

$$P(\vec{\varphi} \mid \vec{K}, \text{Combinatorial}) = \int P(\vec{\varphi} \mid \vec{K}, \text{Combinatorial}, \Theta) P(\Theta) d\Theta$$

$$P(\vec{\varphi} \mid \vec{K}, \text{Additive}) = \int P(\vec{\varphi} \mid \vec{K}, \text{Additive}, \Theta) P(\Theta) d\Theta.$$

Although straightforward in principle, the above integrals are analytically intractable, so we employed the following thermodynamic approximation, described in detail in (Friel and Pettitt, 2008).

Briefly, let $\ln P(\vec{\varphi} \mid \vec{K}, \text{Combinatorial})$ denote the logarithm of the marginal likelihood under the combinatorial model (the procedure for the additive model is equivalent), and let $P_t(\Theta \mid \vec{\varphi}, \vec{K}, \text{Combinatorial})$ denote the posterior density over the model parameters evaluated at some temperature t:

$$P_t(\Theta \mid \vec{\varphi}, \vec{K}, \text{Combinatorial}) = \frac{\left[P(\vec{\varphi} \mid \vec{K}, \text{Combinatorial}, \Theta) \right]^t P(\Theta)}{\int \left[P(\vec{\varphi} \mid \vec{K}, \text{Combinatorial}, \Theta) \right]^t P(\Theta) d\Theta}.$$

The previous equation, termed the *power posterior*, is equivalent to the standard posterior distribution over model parameters when the temperature t is set equal to 1. Importantly, the desired model marginal likelihood can be defined in terms of the power posterior as follows:

$$\ln P(\vec{\varphi} \mid \vec{K}, \text{Combinatorial}) =$$

$$\int_0^1 \left[\int_\Theta \ln P(\vec{\varphi} \mid \vec{K}, \text{Combintorial}, \Theta) \times P_t(\Theta \mid \vec{\varphi}, \vec{K}, \text{Combinatorial}) d\Theta \right] dt$$

$$= \int_0^1 \mathbf{E}_{\Theta \mid \vec{\varphi}, \vec{K}, t} \left[\ln P(\vec{\varphi} \mid \vec{K}, \text{Combintorial}, \Theta) \right] dt.$$

At first glance, the previous equation may not seem that useful, as it requires integrating over the power posterior expectation of the data likelihood, which in fact requires computing the very integral that we were originally trying to solve (the model marginal likelihood). However, the previous equation establishes as an accurate (although computationally intensive) approach to approximating the desired marginal likelihood.

First, the one-dimensional integral over the power posterior expectation can be approximated using a discrete set of V temperature points (denoted $\vec{t} = \{t_1,\ldots,t_V\}$) and the trapezoid rule, resulting in:

$$\ln P(\vec{\varphi}\,|\,\vec{K},\text{Combinatorial}) = \int_0^1 \mathbf{E}_{\Theta|\vec{\varphi},\vec{K},t}\left[\ln P(\vec{\varphi}\,|\,\vec{K},\text{Combintorial},\Theta)\right] dt$$

$$\approx \sum_{i=1}^{V-1}(t_{i+1}-t_i)\times\ldots$$

$$\ldots\frac{\mathbf{E}_{\Theta|\vec{\varphi},\vec{K},t_{i+1}}\left[\ln P(\vec{\varphi}\,|\,\vec{K},\text{Combintorial},\Theta)\right]+\mathbf{E}_{\Theta|\vec{\varphi},\vec{K},t_i}\left[\ln P(\vec{\varphi}\,|\,\vec{K},\text{Combintorial},\Theta)\right]}{2}.$$

Importantly, the posterior expectation at each temperature can be approximated as well. In practice, we did so by drawing samples from the power posteriors using the Hamiltonian Markov Chain Monte Carlo algorithm implemented in the STAN programming environment (Stan Development Team, 2013). These samples were then used to obtain Monte Carlo estimates for the desired expectations of the data likelihoods, which were then plugged into the previous equation. The convergence and performance of the MCMC sampling procedure was assessed at each temperature by comparing two independent Markov chains and assessing their combined effective sample size using the methods and procedures outlined in the STAN reference manual (Stan Development Team, 2013).

As noted by others (Friel and Pettitt, 2008, Calderhead and Girolami 2009), we found that the accuracy of the marginal likelihood approximatation was strongly dependent on the temperature points $\vec{t}$. More specifically, we found that our estimates required accurate estimation of the one-dimensional integral at very low temperatures. This was due to the fact that our prior distributions were very diffuse and the corresponding posteriors were sharply peaked. Therefore, to adequately cover both low and high temperature ranges, we used the following piecewise function to define $\vec{t}$:

$$\vec{t}^{\,1} = \left(\frac{i}{V_1}\right)^c \;\forall i = 1\ldots V_1$$

$$\vec{t}^{\,2} = \exp\left[i\times\frac{t_0^1-\log(t_{min})}{V_2}\right]\forall i = 1\ldots V_2-1$$

$$\vec{t} = \left\{0,\vec{t}^{\,2},\vec{t}^{\,1}\right\}$$

In practice, we set c=$V_1 = 39$, $c = 3.0$, $V_2 = 9$, and $t_{min} = 1\times10^{-10}$, for a total of 49 different temperatures.

After approximating the model marginal likelihoods, the posterior distribution over the competing genetic models was given by:

$$P(\Phi|\vec{\varphi},\vec{K}) = \frac{P(\vec{\varphi}\,|\,\vec{K},\Phi)P(\Phi)}{\sum_{\Phi\in\{\text{Add, Comb}\}} P(\vec{\varphi}\,|\,\vec{K},\Phi)P(\Phi)},$$

where Φ denotes one of the two genetic models considered in the present study. In practice, we assumed a uniform prior over models, and thus,

$$P(\Phi|\vec{\varphi},\vec{K}) = \frac{P(\vec{\varphi}\,|\,\vec{K},\Phi)}{\sum_{\Phi\in\{\text{Add, Comb}\}} P(\vec{\varphi}\,|\,\vec{K},\Phi)}.$$

In the main text, we reported model posterior probability in terms of the $\log_{10}$-Bayes Factor favoring the combinatorial genetic model, defined as:

$$\log_{10}\text{-Bayes Factor} = \log_{10}\frac{P(\text{Combinatorial}|\vec{\varphi},\vec{K})}{P(\text{Additive}|\vec{\varphi},\vec{K})}.$$

Additional Details Concerning Model Inference

We fit the additive and combinatorial models to the complex-Mendelian comorbidity data for 20 distinct diseases (see main Figure 4C and Figure S3). For each of the two models, there are three unknown parameters: the population-level mean frequency of the deleterious genotypes ($\langle p\rangle$), the population-level mean of the penetrance parameters ($\langle x\rangle$), and the total number of loci associated with the disease (n). In practice, without additional sources of data, it is impossible to jointly infer both the population-level mean genotype frequency parameter and the total number of loci. Therefore, we repeated estimation of parameters for each model over a range of potential loci numbers. This was accomplished by computing the posterior densities over the mean genotype penetrance and frequency parameters, conditional on the significantly comorbid complex-Mendelian disease co-occurrence counts (see above for details). Disease co-occurrences were obtained by combining the two largest, non-overlapping clinical datasets (*USA* and *DK*), and to prevent confounding factors from biasing our results, only Mendelian diseases whose comorbidity odds ratios were accurately estimated by the marginal disease counts were included into the analysis. We used non-informative, Beta distributions as priors for the unknown model parameters.

Supplemental References

Antonarakis, S.E., and McKusick, V.A. (2000). OMIM passes the 1,000-disease-gene mark. Nature genetics *25*, 11.

Bates, D., Maechler, M., and Bolke, B. (2013). lme4: Linear mixed-effects models using S4 classes. In Secondary lme4: Linear mixed-effects models using S4 classes, Secondary Author, ed.^eds.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological *57*, 289-300.

Dijkstra, E.W. (1959). A note on two problems in connection with graphs. Numerische mathematik *1*, 269-271.

Doss, H.J., and Narasimhan, B. (1994). Bayesian Poisson regression: Sensitivity analysis through dynamic graphics (Penn State Erie: The Behrend College).

Efron, B. (1982). The Jackknife, the Bootstrap and Other Resampling Plans (Philadelphia, PA: Society for Industrial and Applied Mathematics).

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. Evolution *39*, 783-791.

Felsenstein, J., Kishino H. (1993). Is there something wrong with the bootstrap on phylogenies? a reply to Hillis and Bull. Systematic biology, 193-200.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S.*, et al.* (2011). Ensembl 2012. Nucleic Acids Research *40*, D84-D90.

Friel, N., and Pettitt, A.N. (2008). Marginal likelihood estimation via power posteriors. Journal of the Royal Statistical Society, Series B *70*, 589-607.

Gamazon, E.R., Zhang, W., Konkashbaev, A., Duan, S., Kistner, E.O., Nicolae, D.L., Dolan, M.E., and Cox, N.J. (2010). SCAN: SNP and copy number annotation. Bioinformatics *26*, 259-262.

Genetics Home Reference, N. (2012). http://ghr.nlm.nih.gov/.

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., and McKusick, V.A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res *33*, D514-517.

Lee, I., Blom, U.M., Wang, P.I., Shim, J.E., and Marcotte, E.M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome research *21*, 1109-1121.

Liem, S.L. (2008). [Orphanet and the Dutch Steering Committee Orphan Drugs. A European and Dutch databank of information on rare diseases]. Ned Tijdschr Tandheelkd *115*, 621-623.

NIH, N. http://www.nlm.nih.gov/.

R Core Team (2013). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).

Rambaut, A. (2013). http://tree.bio.ed.ac.uk/software/figtree/%5D.

Risch, N. (1990). Linkage strategies for genetically complex traits. I. Multilocus models. American journal of human genetics *46*, 222-228.

Rzhetsky, A., Seringhaus, M., and Gerstein, M. (2008). Seeking a new biology through text mining. Cell *134*, 9-13.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol *4*, 406-425.

Simonsen, M., Mailund, T., and Pedersen, C.N.S. (2008). Rapid Neighbour Joining. Paper presented at: Proceedings of the 8th Workshop in Algorithms in Bioinformatics (WABI) (Springer Verlag).

Stan Development Team (2013). Stan: A C++ Library for Probability and Sampling, Version 1.1. In Secondary Stan: A C++ Library for Probability and Sampling, Version 1.1., Secondary Author, ed.^eds.

Sukumaran, J., and Holder, M.T. (2010). DendroPy: a Python library for phylogenetic computing. Bioinformatics *26*, 1569-1571.

U.S.A. Health Resources and Services Administration (2013). http://www.arf.hrsa.gov/.

Weinreich, S.S., Mangon, R., Sikkens, J.J., Teeuw, M.E., and Cornel, M.C. (2008). [Orphanet: a European database for rare diseases]. Ned Tijdschr Geneeskd *152*, 518-519.