Finding genes in Mendelian disorders using sequence data: methods and applications

Iuliana Ionita-Laza[1], Vlad Makarov[2], Seungtai Yoon[2], Benjamin Raby[3], Joseph Buxbaum[2], Dan L. Nicolae[4], Xihong Lin[5]

[1] Department of Biostatistics, Columbia University, New York, NY 10032

[2] Department of Psychiatry, Mount Sinai School of Medicine, New York, NY 10029

[3] Channing Laboratory, Brigham and Women's Hospital, Harvard Medical School, Boston MA 02115

[4] Departments of Medicine and Statistics, University of Chicago, Chicago

[5] Department of Biostatistics, Harvard University, Boston, MA 02115

Corresponding author:

Iuliana Ionita-Laza

722 W 168th St 6th Floor

New York, NY, 10025

E-mail: ii2135@columbia.edu

Phone: 212-304-5551

Abstract Many sequencing studies are now underway to identify the genetic causes for both Mendelian and complex traits. Using exome-sequencing, genes for several Mendelian traits including Miller Syndrome, Freeman-Sheldon Syndrome and Kabuki Syndrome have already been identified. The underlying methodology in these studies is a multi-step algorithm based on filtering variants identified in a small number of affected individuals depending on whether they are novel (not yet seen in public resources such as dbSNP, and 1000 Genomes Project), shared among affected (possibly related) individuals, and other external functional information available on the variants.

While intuitive, these filter-based methods are non-optimal and do not provide any measure of statistical uncertainty. We describe here a formal statistical approach that has several distinct advantages: (1) provides fast computation of approximate P-values for individual genes, (2) adjusts for the background variation in each gene so that large genes do not rise to the top based on their sheer size alone, (3) allows for natural incorporation of functional or linkage-based information, and (4) accommodates designs based on both affected relative pairs and unrelated affected individuals. We show via simulations that the proposed approach can be used in conjunction with the existing filter-based methods to achieve a substantially better ranking of a disease gene when compared with currently used filter-based approaches, especially so in the presence of disease locus heterogeneity.

We revisit recent studies on three Mendelian diseases that used filter-based approaches to identify the corresponding disease genes and show how the proposed approach can be applied in such cases, resulting in the disease gene being ranked first in all three studies, and

approximate P-values of 10^{-6} for the gene for the Miller Syndrome, $1.0 \cdot 10^{-4}$ for the gene for the Freeman-Sheldon Syndrome, and $3.5 \cdot 10^{-5}$ for the gene for the Kabuki Syndrome.

1 Introduction

Spurred by recent advances in high-throughput sequencing technologies, sequencing studies for varied Mendelian and complex traits are currently underway. Such studies will provide an unprecedented view of the genetic variation, rare and common, that influences risk to these diseases. Genes for several Mendelian diseases have already been identified[1, 2, 3], using exome-sequencing of a small number of affected individuals and additional information from public resources such as dbSNP and the 1000 Genomes Project.

The large number of genetic variants in the human genome, and the low population frequency of the majority of these variants create challenges for the computational and statistical analysis of these data. In particular, traditional testing strategies based on individual variant testing can have low power, and new statistical methods that aggregate information across multiple variants in a genetic region have been proposed[4, 5, 6, 7, 8, 9, 10, 11, 12, 13].

For Mendelian diseases, traditional methods for gene identification range from candidate gene studies (where candidates were selected based, for example, on functional similarity to already established genes, and in many situations their exons were sequenced in a small number of subjects) to positional cloning strategies (where small regions discovered using linkage analysis were followed-up with denser genotyping that led to identification of haplotypes thought to harbor causal mutations). Recently, several studies have been published on using

whole-exome sequencing data on a small number of (mostly unrelated) affected individuals to identify the genes for several Mendelian traits[1, 2, 3]. Unlike traditional linkage methods, the underlying gene could be identified directly, and by using unrelated subjects. More precisely, in each case the causal gene was identified using a filter-based methodology, where variants identified in cases were checked for novelty (not identified before), functionality (e.g. non-synonymous variants), and sharing among affected (possibly related) individuals. Such an approach is intuitive and reasonable; however, from an inferential perspective it has several disadvantages including: (1) it does not produce any measure of statistical uncertainty (e.g. gene level P-values), making it unfeasible to assess consistency with the null hypothesis (2) it does not adjust for background variation in each gene, therefore allowing large genes to rank high based on their size alone, and (3) it does not properly account for the different levels of variant sharing expected among relatives of different types, which can affect the rank of the disease gene. Although the filter-based approach can take into account external information such as functional predictions or linkage scores, such information needs to be provided in a dichotomized fashion (e.g. linkage or no linkage) rather than original scores (or transformations thereof).

In what follows, we discuss a formal statistical framework that aims to address the aforementioned limitations of the filter-based approach, and show applications to simulated data and recent studies for three Mendelian traits. For these previously published Mendelian studies, we show that the proposed approach ranks the true disease gene first in all three studies, and assigns significant P-values to the respective disease genes.

2 Methods

We start by reviewing the filter-based approach that is currently being used to identify Mendelian genes from sequence data. Then we propose a weighted-sum statistic and an analytical approximation of the P-value for a gene. We then discuss an omnibus method that combines this weighted-sum approach with the currently-used filter-based method to achieve a more sensible gene ranking procedure.

2.1 Filter-based approach

The filter-based approach is based on computing for each gene a statistic equal to the number of affected individuals that are carriers of at least one non-synonymous variant that is novel, that is, not seen in controls. For unrelated affected individuals computing this statistic is straightforward. Let G be a gene of interest, and M_U be the number of novel variant positions observed in a set of A affected individuals sequenced at gene G. Let X_{ij} be the coded genotype (i.e. number of the minor allele) for affected individual $i \leq A$ at novel variant position $j \leq M_U$. Then for each affected individual i we calculate the load (or burden) of novel non-synonymous variants as:

$$L_i = \sum_{j=1}^{M_U} w_j X_{ij},$$

where w_j is 1 for non-synonymous variant, and 0 otherwise. Then the filter-based method is based on the following statistic:

$$S_{\text{filter}} = \sum_{i=1}^{A} I_{\{L_i>0\}}, \tag{1}$$

where $I(\cdot)$ is an indicator function.

For affected relative pairs and Mendelian diseases, it is reasonable to assume that both affected individuals in a pair share the disease variant. If each pair of affected relatives is treated as a unit, the score for each unit (i.e. the equivalent of $I_{\{L_i>0\}}$ above) is taken to be 1 if there is at least one novel, non-synonymous variant in gene G shared between both relatives, and 0 otherwise. However, this definition fails to account for the different levels of expected sharing among relatives of different types. Ideally, one would like to assign a higher score if two cousins share a variant vs. two siblings. Later on we discuss such an alternative scoring scheme.

As the number of sequenced controls increases, restricting attention to only the novel variants runs the risk of disregarding rare disease mutations that are in fact present in control individuals as well (possibly due to reduced penetrance and/or recessive mode of inheritance). A simple extension of the filter-based approach is to also consider variants that have a frequency in controls less than some threshold, say 0.01, rather than only the novel ones. We refer to this approach Filter-R (all rare variants are included), and the existing filter-based approach based on novel variants only is referred to as Filter-N.

2.2 Weighted-sum statistic for Mendelian traits

We describe here a weighted-sum statistic that resembles statistics that have been proposed before for case-control designs[6]. However, unlike existing weighted-sum statistics, for the proposed statistic (1) an approximate analytical P-value can be calculated for each gene, and (2) both affected relative pairs and unrelated affected individuals can be accommodated.

Let G be a gene of interest, and M be the number of rare variant positions observed in a set of individuals (both affected and unaffected) sequenced at gene G. We assume for now that all individuals are unrelated. A rare variant is defined as a variant with population frequency less than some pre-specified threshold, e.g. 0.01. The optimal threshold is not known and necessarily depends on the underlying frequency spectrum for disease mutations in Mendelian diseases. However, extensive data available on the frequency spectrum for Mendelian mutations suggests that the total mutation frequency is $<<$ 1% for most Mendelian diseases[14]. For each rare variant position j, with $j \leq M$, let $T(j)$ be the total number of variants in affected individuals (note that this corresponds to an additive model). One simple statistic we can define is:

$$S = \sum_{j=1}^{M} T(j).$$

Moreover, incorporation of external weights such as those from Polyphen[15] or SIFT[16] can be done easily. For example,

$$S_w = \sum_{j=1}^{M} w_j T(j),$$

where w_j is the weight for variant j which can be any real positive number (derived independently of the data). For example, if only non-synonymous variants are to be included then $w_j = 1$ for such variants and 0 otherwise. A similar weighting scheme works if only variants that are not in dbSNP are to be considered.

Let N_a be the total number of chromosomes in affected individuals, and N_u be the corresponding number for controls. For variant j let $\widehat{f_j}$ be the estimated frequency based on controls. If we assume that the underlying frequency distribution of the variants in a region can be approximated by Beta(α, β), then we estimate f_j by:

$$\widehat{f_j} = \frac{x_j + \alpha}{N_u + \alpha + \beta},$$

where x_j is the observed number of occurrences of the minor allele in controls at variant position j (The parameters α and β can be estimated from data available on controls using standard maximum likelihood estimation [19]. We also note that results are robust to the choice of α and β especially as N_u becomes large.). If we assume for now that the rare variants under consideration are in linkage equilibrium then we show in the Supplemental Materials S1.1 and S1.2 that:

$$\begin{aligned} \widehat{\mathrm{E}}(S_w) &= \sum_{j=1}^{M} w_j N_a \widehat{f_j} \text{ and} \\ \widehat{\mathrm{Var}}(S_w) &= \sum_{j=1}^{M} w_j^2 N_a \left[\frac{N_a - 1}{N_u} + 1 \right] \widehat{f_j}(1 - \widehat{f_j}). \end{aligned}$$

In the general case when variants are allowed to be correlated, a suitable variance estimator

has also been derived (Supplemental Material S1.2).

We use the following gamma-based approximation for the probability density function of the weighted-sum statistic of Poisson-like random variables (Supplemental Table S5; see also Fay and Feuer[17]):

$$P_{\text{null}}(a) \quad = \quad P(S_w \geq a) = 1 - Q\left(\frac{a}{\widehat{w}_{\text{equiv}}}, \frac{\widehat{\text{E}}(S_w)}{\widehat{w}_{\text{equiv}}}\right), \tag{2}$$

where $\widehat{w}_{\text{equiv}} = \frac{\widehat{\text{Var}}(S_w)}{\widehat{\text{E}}(S_w)}$ and Q is the incomplete gamma function: $Q(a,x) = \frac{1}{\Gamma(a)} \int_x^\infty e^{-t} t^{a-1} dt$.

This approximation becomes very accurate as the observed number of variants M in a region increases. It can however be conservative when M is small (Supplemental Table S5).

2.2.1 Only novel variants in cases are considered.

Previous studies on several Mendelian traits[1, 2, 3] have used public resources such as dbSNP and 1000 Genomes Project data, as well as sequence data on a small number of controls to filter out variants that are common and only keep those that are novel (do not appear in these existing databases). This is indeed a reasonable approach if disease mutations are assumed to be very rare and highly penetrant. We can modify our weighted-sum statistic above as follows:

$$S_w^{\text{novel}} = \sum_{j=1}^{M_U} w_j T(j),$$

where M_U is the number of *novel* variants in affected individuals. Note that M_U is a subset of M, and that $\text{E}(S_w^{\text{novel}}) \leq \text{E}(S_w)$ and $\text{Var}(S_w^{\text{novel}}) \leq \text{Var}(S_w)$. In order to calculate $\text{E}(S_w^{\text{novel}})$

and $\mathrm{Var}(S_w^{\mathrm{novel}})$ one would need to estimate the number of novel variants in cases based on the observed variants in controls, and both parametric and non-parametric methods can be applied to obtain such estimates[18, 19]. However, it can be difficult to obtain accurate estimates on the number of novel variants in a gene if only a small number of variants is observed in controls, as would be the case for many genes of small to moderate length. Therefore we use the same gamma-based approximation as in eq. (2) to obtain an upper bound on the P-value for this scenario.

In what follows we refer to the weighted-sum approach with all rare variants as WS-R, and to the above approach with only the novel variants as WS-N.

2.2.2 Affected-relative pairs

For Mendelian diseases data on affected relatives, e.g. affected siblings or affected cousins, may be available. It would be desirable to extend both the filter-based approach and the weighted-sum approach discussed above to be able to handle relative pairs. A simple solution adopted in the current filter-based approach is to score each pair of affected relatives as 1 if they share at least one novel and non-synonymous variant, and 0 otherwise. A potential weakness of such a scoring scheme is that it fails to account for the different levels of expected sharing among relatives of different types. In particular, we would like to assign a higher score when such sharing happens between more distant relatives, e.g. cousins, compared with siblings.

In Ionita-Laza and Ottman[20] we have developed such a scoring scheme. Namely, for a

pair of relatives we derive an *effective* number of variants in the pair, that is, the number of variants at a fixed segregating or variant position adjusted for the familial correlation. We have denoted this number by k_{eff} and showed there that for a pair of relatives k_{eff} can be calculated as follows:

$$k_{\text{eff}} = \begin{cases} \log_{2f}\left[4f\varphi + 4f^2(1 - 4\varphi + 4\delta\varphi^2)\right], & \text{if both relatives carry a rare variant} \\ 1, & \text{if only one of the two relatives carries a rare variant} \\ 0, & \text{if neither of the two relatives carries a rare variant} \end{cases}$$

where f is the frequency of the variant at the given position, φ is the kinship coefficient; δ is 0 if the two relatives can share a maximum of one allele IBD (e.g. first cousins) and 1 if they can share 2 alleles IBD (e.g. siblings).

When two heterozygous individuals are unrelated , $\varphi = 0$ and we obtain the expected result that $k_{\text{eff}} = 2$. For identical twins $\varphi = 0.5$, $\delta = 1$ and $k_{\text{eff}} = 1$. For two sibs, when $f = 0.01$ we obtain $k_{\text{eff}} = 1.17$. Similarly for two second cousins, $k_{\text{eff}} = 1.76$. These and other examples are summarized in Supplemental Table S1. With this scoring scheme, the filter-based approach can be modified to assign higher scores to sharing among cousins compared with siblings.

It is also possible to extend the weighted-sum approach to take into account data on affected relatives in addition to unrelated affected individuals. For a variant position and a pair of relatives, instead of the observed number of variants we use the *effective* number k_{eff} defined above. Then for variant position j we replace $T(j)$, the total number of vari-

ants at position j in the affected individuals, with $T_{\text{eff}}(j)$ and the weighted-sum statistic is correspondingly defined as:

$$S_w = \sum_{j=1}^{M} w_j T_{\text{eff}}(j).$$

As for the scenarios with only unrelated individuals, we derive a gamma-based approximation for the distribution of S_w (Supplemental Material S2).

For Mendelian diseases, it is reasonable to assume that affected relatives within the same family are likely to share the disease mutation. The approach discussed above can be modified easily to reflect this assumption, by setting k_{eff} to be zero unless both relatives share a variant (that can be non-synonymous and novel). This is the default setting in our handling of affected relatives, and the one illustrated in the examples that follow.

2.3 Joint-Rank approach

We describe here how the weighted-sum approach above can be combined with the currently-used filter-based method to produce an overall better ranking for the true disease gene(s) in a study. Both approaches discussed in the previous sections attempt to quantify the increase in rare variant burden in affected individuals, although in slightly different ways. The weighted-sum approach aggregates information across all affected individuals and adjusts for the underlying variation in controls, but does not always distinguish whether the variants that enter the calculation of S_w occur in many or just a few of the individuals. On the contrary, the existing filter-based approach essentially exploits the information on the *number* of affected individuals that carry at least one novel variant, but fails to distinguish whether variants

occur recurrently at the same position, or different positions, and does not take into account the number of novel variants an individual carries, unlike the weighted-sum approach.

For the purpose of ranking genes, we propose to combine the two approaches to calculate for each gene a combined rank, henceforth called the Joint-Rank, that represents the average of the ranks from the weighted-sum and filter-based approaches. For a true disease gene both ranks should be high, and the Joint-Rank approach may lead to an overall better ranking of the true disease gene. The filter-based rank is not adjusted for the background variation, and hence the Joint-Rank can be viewed as adjusting the filter-based rank for the length of the gene, and the background variation in each gene.

The various approaches discussed in this section are summarized in Table 1.

3 Applications

Next, we investigate via simulations the properties of the proposed approaches. We also use real high-coverage sequence data on 310 control individuals randomly selected from the large collection of unaffected individuals that have been sequenced as part of the ARRA Autism Project (see Supplemental Material S4 for more details on these data) to illustrate applications to three Mendelian disease examples recently reported in the literature: Miller Syndrome[2], Freeman-Sheldon Syndrome[1], and Kabuki Syndrome[3].

3.1 Simulated data

We first use simulations to investigate the underlying properties of the proposed approaches. We simulated 10 independent genomic regions each of length 1 Mb under a coalescent model using the software package COSI[21]. The model used in the simulation was the calibrated model for the European population, and was an option available in the COSI package. A total of $10,000$ haplotypes were generated for each region. We then randomly sample small subregions of the size of individual genes. The size of each gene is sampled from the length distribution of real exonic regions (as available from the refGene table).

Type-1 Error

We evaluated the Type-1 error of the proposed approaches for several different scenarios, including two different designs: (1) case-control, and (2) affected sib-pairs and unrelated controls. The results for the case-control design are shown in Table 2. We show there that the proposed gamma-based approximation is valid and leads to a good control of the Type-1 error.

When only novel variants (i.e. not seen in a set of independent controls) are considered, the approximation can be very conservative. Despite this conservativeness, since the magnitude of the effect at the disease gene is expected to be large for Mendelian diseases, the approximation is expected to be powerful for such effects.

Similar results hold for datasets containing affected relative pairs (Supplemental Table S2).

Disease Gene Ranking

We investigate here the performance of the various approaches as measured by the overall ranking of the disease gene in a genome-scan with $20,000$ genes. A genome-scan is simulated by sampling $20,000$ regions with region length selected from the gene (exonic) length distribution in refGene table. The genes are sampled independently from the ten 1 Mb regions we have simulated. We assume 10 affected individuals and a number of controls between 100 and 500. One gene at random is selected to be the disease gene and a small number of affected individuals (between 2 and 6) are assumed to each carry a different novel disease mutation. We simulate 1000 such genome-scans, and calculate the median rank for the disease gene across the 1000 simulations.

We show in Figure 1 that the Joint-Rank-N approach outperforms both the WS-N and the Filter-N methods in terms of the rank assigned to the disease gene. The performance of the filter-based approach decreases as the genetic heterogeneity at the disease gene increases, and it is in these situations that a formal approach such as the weighted-sum method discussed in this paper becomes particularly necessary. The extension of the filter-based approach to include rare variants rather than only novel variants (i.e. Filter-R) does not perform very well, especially as the number of affected individuals that carry a disease mutation at a disease locus decreases (Supplemental Figure S1). In such situations the proposed weighted-sum approach alone is expected to perform better.

Results for affected sib-pairs are shown in Supplemental Figure S2, and are similar to those for the case-control design.

3.2 Applications to three Mendelian diseases

For these applications, we use high-coverage sequence data with spiked-in mutations to resemble the original disease studies as closely as possible. In particular, we assume that the same number of affected individuals as in the original studies are carriers of novel non-synonymous disease mutations, and these mutations are artificially added to the corresponding disease gene, above and beyond the existing variation in our real data. We also disregard variants with a known *rs* number by simply setting their weights to 0. The next set of results are based on these spike-in datasets.

Miller Syndrome

In Ng et al.[2] the authors performed exome-sequencing of four affected individuals with Miller Syndrome, two siblings and two unrelated affected individuals. All four affected individuals were compound heterozygotes for novel and non-synonymous mutations in one gene, *DHODH*, and the two siblings shared the disease mutations. Since the sequence data available to us contains only unrelated individuals, we emulated the original study by using data on only 3 unrelated individuals as cases, and 300 unrelated individuals as controls; all individuals are part of the same exome-sequencing study (Supplemental Material S4). For the disease gene *DHODH* we make the additional assumption that each of the three affected individuals is compound heterozygote for unique mutations in this gene.

In Figure 2 we plot the P-values (WS-N) for all genes, as well as the value of the filter-based statistic (i.e. the number of affected individuals carriers of novel non-synonymous

variants). With only three affected individuals, we identify gene *DHODH* as the leading gene, with an approximate P-value of 10^{-6} (analytic gamma-approximation).

Freeman-Sheldon Syndrome

For the Freeman-Sheldon syndrome example, Ng et al.[1] performed exome-sequencing of four unrelated affected individuals. Two different novel and non-synonymous variant positions in the same gene, *MYH3*, were detected in all four individuals. Three individuals had a mutation at the first variant position, while the fourth individual had a mutation at the second variant position. Based on our spike-in dataset, the resulting approximate P-value (WS-N) in this case is $1.0 \cdot 10^{-4}$. This was the highest ranked gene in the study (Figure 2).

Kabuki Syndrome

For the Kabuki Syndrome example, exome-sequencing was performed in ten unrelated affected individuals (Ng et al.[3]). Nine different novel and non-synonymous mutations in gene *MLL2* were identified in the 10 affected individuals. Based on our spike-in dataset, the resulting approximate P-value is $3.5 \cdot 10^{-5}$, and again this is the highest ranked gene (Figure 2).

Results for these three Mendelian diseases are summarized in Table 3.

4 Discussion

Recent studies have shown how genes for Mendelian diseases can be identified using whole-exome sequence data for a small number of affected individuals. The underlying approach is based on filtering variants based on novelty, functionality, and sharing among multiple affected individuals. Such filter-based approaches are intuitive and powerful for Mendelian diseases, but suffer from several shortcomings. Notable among them are: (1) the lack of statistical uncertainty assessment (e.g. in the form of P-values), (2) the lack of adjustment for the background variation in each gene, so that large genes can rank high based on their size alone, and (3) genetic distance between biological relatives is not modeled properly. We have shown here that such a filter-based approach can be complemented by a formal statistical procedure that has several distinct advantages: (1) evaluates statistical significance by calculating approximate P-values, (2) can handle both related and unrelated affected individuals, (3) can incorporate external weights about the functionality of variants or linkage-based scores, and (4) importantly, adjusts for background variation so that more variable regions do not rise to the top based on noise alone. The resulting procedure leads to an overall better ranking of the disease gene, and allows for untieing genes that otherwise have the same number of affected individuals that carry a novel mutation in the gene.

We have investigated two distinct scenarios: one that considers all rare variants in the population, regardless of whether they have been seen before or not (WS-R); and a second scenario where only *novel* variants in cases are included (WS-N). We have derived a gamma-based approximation for the null distribution of the weighted-sum statistic WS-R and have

shown that this approximation is good. Also, we have shown that the same approximation can be used for WS-N to derive an upper bound on the P-value. Via applications to both simulated and real data, we have shown that a combination of the weighted-sum approach and the filter-based approach, a procedure we call Joint-Rank, provides a more robust way to rank genes in Mendelian diseases compared with filter-based approaches alone. In particular, the Joint-Rank approach adjusts for the background variation in each gene (as does the weighted-sum approach) and at the same time favors genes with larger number of affected individuals that are carriers of novel variants (as does the filter-based approach). Throughout our examples we have assumed that causal variants are novel and hence not present in unaffected individuals. Under such a scenario, the optimal approach is indeed to only consider novel variants. However, if causal variants could be present in unaffected individuals (for example, for a recessive mode of inheritance, or reduced penetrance scenarios), the weighted-sum approach WS-R should also be considered.

We revisited recent exome-sequencing studies on several Mendelian diseases and showed how the approach works concretely in these examples. The proposed approach produced significant P-values for each of the disease genes in the three Mendelian traits, while properly adjusting for the background variation in each gene, as estimated from exome-sequencing data available to us for 300 controls. Due to the lack of even modest-sized sequence datasets in the past, the filter-based approach used a variety of variant databases to filter out already discovered variants, including dbSNP and 1000 Genomes Project data. With the proposed approach, it is still possible to use these databases to filter out variants by simply setting the

weights for variants in the databases to 0, and this is especially useful when the number of controls available is rather small. For our own examples, we have presented results based on a relatively small number of controls (i.e. 300); however, increasing the number of controls will naturally lead to smaller P-values and improved overall ranking for the true disease gene.

As with any association study, good experimental design is essential. The validity of the P-values obtained from the weighted-sum approach, and of the Joint-Rank procedure overall is contingent on having a control dataset that is well matched to the affected individuals, for both ethnic background, as well as sensitivity and specificity for variant detection. Other potential issues such as hidden relatedness among individuals can lead to inflated Type-1 error. Principal component analysis or mixed-model methods can be used to adjust for relatedness of subjects by extending the current method to a regression-framework.

One strength of the proposed weighted-sum approach is that the P-values can be obtained in an analytical fashion. This fact makes the proposed approach to be computationally very fast compared to a permutation-based procedure, and also allows inclusion of affected relative pairs, situations where resampling-based procedures are non-trivial.

The proposed methods implicitly assume an additive model for the effect of mutations at a position. This model is optimal for additive, dominant, and compound heterozygous modes of inheritance (as true for the three Mendelian disease examples we considered). Even for a recessive model with two mutations required at the same position, the proposed approach is expected to be very powerful.

For Mendelian diseases, results from previous linkage-based scans may be available. In

that case, Roeder et al.[24] proposed an exponential weighting scheme, whereby linkage scores are translated into weights that can be used to weight the gene level P-values calculated with the proposed approach, as in a weighted-hypothesis testing procedure[25].

In summary, we have discussed an analytic framework to identify disease genes for Mendelian diseases, and have shown that it performs well in simulations and applications to previous exome-sequencing studies for three Mendelian traits.

Software Software implementing the proposed approaches is available freely on our website at: http://www.columbia.edu/∼ii2135/.

Acknowledgments The research was partially supported by NSF grant DMS-1100279 and NIH grant 1R03HG005908 (to II-L), and R37-CA076404 and P01-CA134294 from the National Cancer Institute (to XL).

Control exomes were sequenced as part of an ongoing ARRA-funded autism exome-sequencing grant (MH089025, Mark Daly, communicating PI, Joseph Buxbaum, Bernie Devlin, Richard Gibbs, Gerard Schellenberg, James Sutcliffe, collaborating PI's).

Web Resources

PolyPhen http://genetics.bwh.harvard.edu/pph2/

SIFT http://sift.jcvi.org/

dbSNP http://www.ncbi.nlm.nih.gov/projects/SNP/

1000 Genomes Project http://www.1000genomes.org/

refGene http://genome.ucsc.edu/cgi-bin/hgTables/

Supplemental Material

S1 Expectation and Variance of S_w for unrelated cases

S1.1 Expectation and Variance of $T(j)$

We assume we have sequenced $N_a/2$ affected individuals, and $N_u/2$ unaffected individuals. For an observed variant position j, let $\widehat{f_j}$ be the estimated frequency of f_j based on N_u chromosomes. Then we use the following to estimate the expected value of $T(j)$.

$$\widehat{\mathrm{E}}(T(j)) = N_a \widehat{f_j}.$$

For the variance, we have:

$$\begin{aligned}
\widehat{\mathrm{Var}}(T(j)) &= \mathrm{Var}(\mathrm{E}(T(j)|\widehat{f_j})) + \mathrm{E}(\mathrm{Var}(T(j)|\widehat{f_j})) = \mathrm{Var}(N_a \widehat{f_j}) + \mathrm{E}(N_a \widehat{f_j}(1-\widehat{f_j})) \\
&= N_a^2 \frac{\widehat{f_j}(1-\widehat{f_j})}{N_u} + N_a \widehat{f_j} - N_a \mathrm{E}(\widehat{f_j}^2) = \\
&= N_a \left[\frac{N_a - 1}{N_u} + 1 \right] \widehat{f_j}(1-\widehat{f_j}).
\end{aligned}$$

S1.2 Expectation and Variance of S_w

We recall here that for each gene we calculate the following weighted-sum statistic:

$$S_w = \sum_{j=1}^{M} w_j T(j).$$

Then $\widehat{\mathrm{E}}(S_w) = \sum_{j=1}^{M} w_j \widehat{\mathrm{E}}(T(j))$. For the variance of S_w we have:

$$\widehat{\mathrm{Var}}(S_w) = \sum_{j=1}^{M} w_j^2 \widehat{\mathrm{Var}}(T(j)) + \sum_{1 \le j \ne j' \le M} w_j w_{j'} \widehat{\mathrm{Cov}}(T(j), T(j')).$$

The covariance can be estimated as follows[27]. Let V_e be the $M \times M$ empirical variance estimator with $v_{jj'} = \frac{A}{N} \sum_{i=1}^{N} (X_{ij} - E(X_{ij}))(X_{ij'} - E(X_{ij'}))$, where $N = A + U$ is the total number of individuals (affected and unaffected). Let D be the $M \times M$ diagonal matrix with $d_{jj} = \widehat{\mathrm{Var}}(T(j))$. Also we define an adjusted variance matrix: $V_A = D^{1/2}[\mathrm{Diag}(V_e)^{-1/2} V_e \mathrm{Diag}(V_e)^{-1/2}] D^{1/2}$. Then an estimate for $\mathrm{Var}(S_w)$ is $\sum_{j,j'} V_A[j, j']$.

S2 Expectation and Variance of S_w when affected individuals are related

S2.1 Expectation and Variance for $T(j)$

We show here how to derive the expected value and variance of T_{eff} at a variant position when affected relatives are considered. Let A be the total number of affected relative pairs (of same type). If f is estimated based on N_u chromosomes, then we can get for

$$\widehat{\mathrm{E}}[T_{\text{eff}}] = A \left[k_{\text{eff}|2} 4\widehat{f}\varphi + 4\widehat{f}(1 - 2\varphi) \right].$$
$$\widehat{\mathrm{Var}}[T_{\text{eff}}] = A^2 \left(k_{\text{eff}|2} 4\varphi + 4(1 - 2\varphi) \right)^2 \frac{\widehat{f}(1 - \widehat{f})}{N_u} + A \cdot \left(k_{\text{eff}|2} 4\widehat{f}\varphi + 4\widehat{f}(1 - 2\varphi) \right).$$

To assess the covariance between T_{eff} at two different positions, we need to know the

joint distribution of genotypes at two positions in two relatives. Lange[28] has derived the relative-to-relative transition probabilities for two linked genes, and we make use of these transition probabilities and the observed genotype distribution at two positions in unrelated controls to derive the joint distribution in relatives that we need. We then use a gamma-based approximation for the weighted-sum of Poisson random variables.

We claim here that the distribution of T_{eff} under the null hypothesis of no association with disease can be approximated by an overdispersed Poisson distribution with mean $\sum_{i=1}^{A} E[k_{\text{eff}}(i)]$, and an index of dispersion very close to 1. It is easy to verify this claim by simple simulation experiments. We have simulated datasets of affected sib-pairs and controls at one single variant position of frequency $0.001 \leq f \leq 0.01$. For each dataset we calculate T_{eff} assuming (1) the true value of f, and (2) the estimated value of f from controls. We report the mean and variance for $T_{\text{eff}}(f)$ and $T_{\text{eff}}(\widehat{f})$ based on 10000 random simulations, as well as the correlation between $T_{\text{eff}}(f)$ and $T_{\text{eff}}(\widehat{f})$. Results are shown in Supplemental Table S3. For more distant relatives, such as first and second cousins, we only report the theoretical mean and variance for $T_{\text{eff}}(f)$ (Supplemental Table S4). As shown, the theoretical and empirical results match very well. There is a slight inflation in the variance over the mean for sib-pairs and when $f = 0.01$ (dispersion index < 1.06), although this inflation disappears for more distant relatives. In Figure S3 we also show the distribution of T_{eff} against a Poisson with the same mean for a scenario with 100 affected sib-pairs and 500 controls and $f = 0.005$.

S3 Gamma-based approximation for a sum of weighted-Poisson random variables

We have done some simple calculations in R to assess the accuracy of the gamma-based approximation for the weighted-sum of Poisson random variables. We assume M Poisson random variables are included, and for each a weight w_i is chosen from $U(0,1)$. The results for different values for M are shown in Table S5.

S4 Sequence Data

To illustrate applications to real sequence data, we used exome-level data on 310 control individuals randomly selected from the large collection of unaffected individuals that have been sequenced as part of the ARRA Autism Project (AAP). The AAP involves whole-exome sequencing of 1000 autism cases and 1000 controls, and several hundred trios. Whole-exome sequencing of controls was carried out at the Broad Institute and at Baylor College of Medicine using standard approaches. Following QC, variants were called using several approaches (including the Genome Analysis Toolkit[26]), and variant call files with all variants and relevant QC metrics were made available to us. For our applications we considered data on 310 randomly chosen control individuals.

References

[1] Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461: 272–276.

[2] Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA et al. (2010a) Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* 42: 30–35.

[3] Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC et al. (2010b) Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* 42: 790–793.

[4] Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 83: 311–321.

[5] Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 5: e1000384.

[6] Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR (2010) Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 86: 832–838.

[7] Liu DJ, Leal SM (2010) A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet* 6: e1001156.

[8] King CR, Rathouz PJ, Nicolae DL (2010) An evolutionary framework for association testing in resequencing studies. *PLoS Genet* 6: e1001202.

[9] Bhatia G, Bansal V, Harismendy O, Schork NJ, Topol EJ, Frazer K, Bafna V (2010) A covering method for detecting genetic associations between rare variants and common phenotypes. *PLoS Comput Biol* 6: e1000954.

[10] Han F, Pan W (2010) A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered* 70: 42–54.

[11] Ionita-Laza I, Buxbaum J, Laird NM, Lange C (2011) A new testing strategy to identify rare variants with either risk or protective effect on disease. *Plos Genet*, 7: e1001289.

[12] Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ (2011) Testing for an unusual distribution of rare variants. *PLoS Genet* 7: e1001322.

[13] Wu M, Lee S, Cai T, Li Y, Boehnke M, Lin X (2011) Rare Variant Association Testing for Sequencing Data Using the Sequence Kernel Association Test (SKAT) *Am J Hum Genet*, in press.

[14] Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common disease-common variant...or not? *Hum Mol Genet* 11: 2417–2423.

[15] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. *Nat Methods* 7: 248–249.

[16] Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols* 4: 1073–1081.

[17] Fay MP, Feuer EJ (1997) Confidence intervals for directly standardized rates: a method based on the gamma distribution, *Stat Med*, 16: 791–801.

[18] Efron B, Thisted R (1976) Estimating the number of unknown species: How many words did Shakespeare know? *Biometrika* 63: 435–437

[19] Ionita-Laza I, Lange C, Laird NM (2009) Estimating the number of unseen variants in the human genome. *Proc Natl Acad Sci USA* 106: 5008–5013.

[20] Ionita-Laza I, Ottman R (2011) On study designs for identification of rare disease variants in complex diseases. *Genetics. In press.*

[21] Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* 15: 1576–1583.

[22] Lemire M (2011) Defining rare variants by their frequencies in controls may increase type I error. *Nat Genet* 43: 391–392.

[23] Pearson RD (2011) Bias due to selection of rare variants using frequency in controls. *Nat Genet* 43: 392–393.

[24] Roeder K, Bacanu SA, Wasserman L, Devlin B (2006) Using linkage genome scans to improve power of association in genome scans. *Am J Hum Genet* 78: 243–252.

[25] Ionita-Laza I, McQueen MB, Laird NM, Lange C (2007) Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100K scan. *Am J Hum Genet* 81: 607–614.

[26] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20: 1297–1303.

[27] Rakovski CS, Xu X, Lazarus R, Blacker D, Laird NM (2007) A new multimarker test for family-based association studies. *Genet Epidemiol* 31: 9–17.

[28] Lange K (1974) Relative-to-relative transition probabilities for two linked genes. *Theoretical Population Biology* 6: 92–107.

Figure Legends

Figure 1: The median rank (novel-variants only) of a disease gene in genome scans with $20,000$ genes, with gene length sampled from the real gene length distribution. 1000 such genome-scans are simulated. $2-6$ of 10 affected individuals are assumed to carry a novel disease mutation in the disease gene (with fewer mutations for larger number of controls). The following methods are compared: WS-N, Filter-N, and Joint-Rank-N.

Figure 2: Applications to three Mendelian diseases: Miller Syndrome, Freeman-Sheldon Syndrome and Kabuki Syndrome. On the left we show the P-values (WS-N) for $19,811$ genes surveyed (manhattan plot). On the right we show for each gene the number of affected individuals that are carriers of novel disease variants, and the gene P-value.

Table 1: Summary of methods discussed in text.

Approach	Description
WS-R	Weighted-sum with all rare variants (e.g. MAF$\leq$ 0.01)
WS-N	Weighted-sum with only novel variants (not seen before)
Filter-R	Filter-based approach with all rare variants (e.g. MAF$\leq$ 0.01)
Filter-N	Filter-based approach with only novel variants (not seen before)
Joint-Rank-R	For each gene: the average of the ranks from approach WS-R and Filter-R
Joint-Rank-N	For each gene: the average of the ranks from approach WS-N and Filter-N

Table 2: Type-1 error for the Case-Control Design.

Approach	A[a]	U[b]	α			
			10^{-4}	10^{-3}	10^{-2}	$5 \cdot 10^{-2}$
WS-R	5	100	$1.5 \cdot 10^{-4}$	$6.0 \cdot 10^{-4}$	$4.0 \cdot 10^{-3}$	$1.7 \cdot 10^{-2}$
		500	$1.3 \cdot 10^{-4}$	$7.0 \cdot 10^{-4}$	$5.0 \cdot 10^{-3}$	$2.1 \cdot 10^{-2}$
		1000	$1.1 \cdot 10^{-4}$	$5.7 \cdot 10^{-4}$	$5.0 \cdot 10^{-3}$	$2.1 \cdot 10^{-2}$
	10	100	$1.0 \cdot 10^{-4}$	$4.0 \cdot 10^{-4}$	$3.0 \cdot 10^{-3}$	$1.6 \cdot 10^{-2}$
		500	$1.2 \cdot 10^{-4}$	$7.1 \cdot 10^{-4}$	$4.8 \cdot 10^{-3}$	$2.3 \cdot 10^{-2}$
		1000	$1.1 \cdot 10^{-4}$	$8.0 \cdot 10^{-4}$	$5.0 \cdot 10^{-3}$	$2.3 \cdot 10^{-2}$
	20	100	$1.7 \cdot 10^{-5}$	$1.4 \cdot 10^{-4}$	$1.4 \cdot 10^{-3}$	$1.0 \cdot 10^{-2}$
		500	$1.1 \cdot 10^{-4}$	$6.4 \cdot 10^{-4}$	$5.0 \cdot 10^{-3}$	$2.5 \cdot 10^{-2}$
		1000	$1.1 \cdot 10^{-4}$	$6.5 \cdot 10^{-4}$	$5.0 \cdot 10^{-3}$	$2.6 \cdot 10^{-2}$
WS-N	5	100	$7.8 \cdot 10^{-5}$	$3.0 \cdot 10^{-4}$	$1.5 \cdot 10^{-3}$	$6.7 \cdot 10^{-3}$
		500	$2.6 \cdot 10^{-5}$	$7.4 \cdot 10^{-5}$	$4.3 \cdot 10^{-4}$	$3.0 \cdot 10^{-3}$
		1000	$2.1 \cdot 10^{-5}$	$1.2 \cdot 10^{-4}$	$2.9 \cdot 10^{-4}$	$1.1 \cdot 10^{-3}$
	10	100	$3.3 \cdot 10^{-5}$	$1.4 \cdot 10^{-4}$	$1.1 \cdot 10^{-3}$	$6.1 \cdot 10^{-2}$
		500	$7.0 \cdot 10^{-6}$	$5.2 \cdot 10^{-5}$	$2.5 \cdot 10^{-4}$	$2.0 \cdot 10^{-3}$
		1000	$1.3 \cdot 10^{-5}$	$3.0 \cdot 10^{-5}$	$1.1 \cdot 10^{-4}$	$8.6 \cdot 10^{-4}$
	20	100	$8.0 \cdot 10^{-6}$	$4.2 \cdot 10^{-5}$	$3.0 \cdot 10^{-4}$	$2.7 \cdot 10^{-3}$
		500	$3.0 \cdot 10^{-6}$	$2.1 \cdot 10^{-5}$	$1.7 \cdot 10^{-4}$	$1.7 \cdot 10^{-3}$
		1000	$2.3 \cdot 10^{-6}$	$2.3 \cdot 10^{-5}$	$8.7 \cdot 10^{-5}$	$7.8 \cdot 10^{-4}$

[a] #unrelated affected individuals
[b] #unrelated unaffected individuals

Table 3: Summary results for the applications to three Mendelian traits.

Syndrome	Gene Length (kb)	Dataset		MOI[a]	P-value[b]
		A[c]	U[d]		(WS-N)
Miller	16.0	3	300	CH	1.0E-06
Freeman-Sheldon	28.7	4	300	D	1.0E-04
Kabuki	36.3	10	300	D	3.5E-05

[a]Mode of Inheritance: compound heterozygote (CH) or dominant (D)
[b]Analytical P-value
[c]#unrelated affected individuals
[d]# unaffected individuals

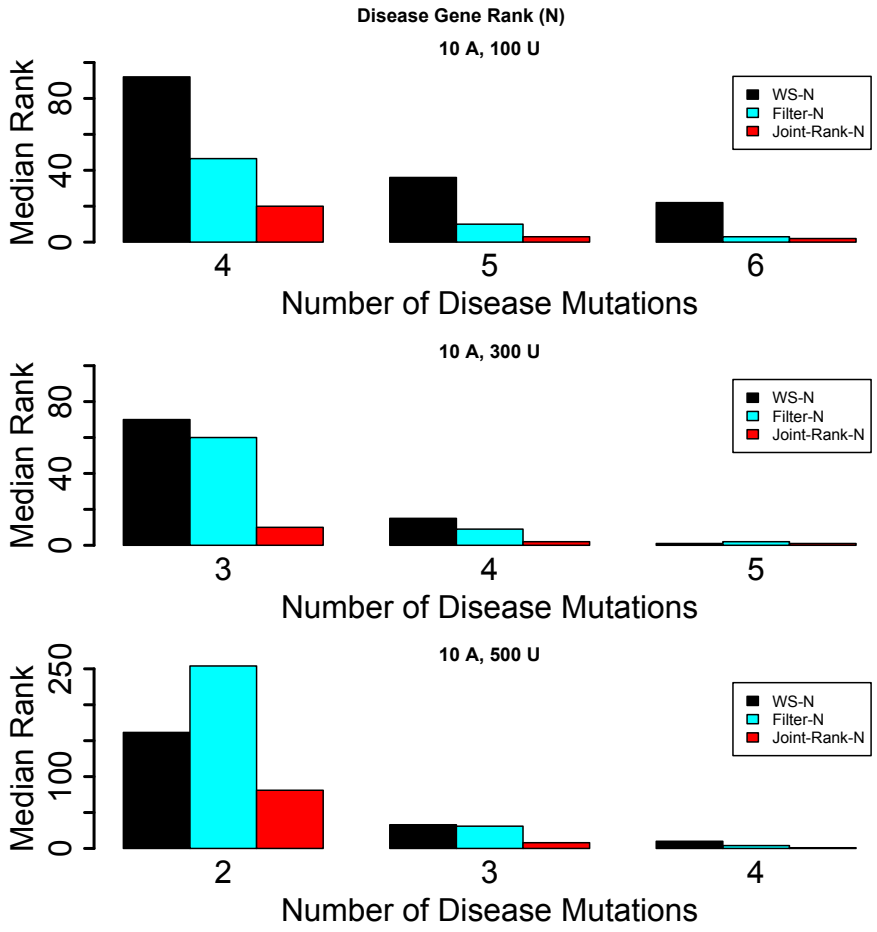
Disease Gene Rank (N)
10 A, 100 U
Median Rank
0
40
80
WS-N
Filter-N
Joint-Rank-N
4
5
6
Number of Disease Mutations
10 A, 300 U
Median Rank
0
40
80
WS-N
Filter-N
Joint-Rank-N
3
4
5
Number of Disease Mutations
10 A, 500 U
Median Rank
0
100
250
WS-N
Filter-N
Joint-Rank-N
2
3
4
Number of Disease Mutations

Figure 1:

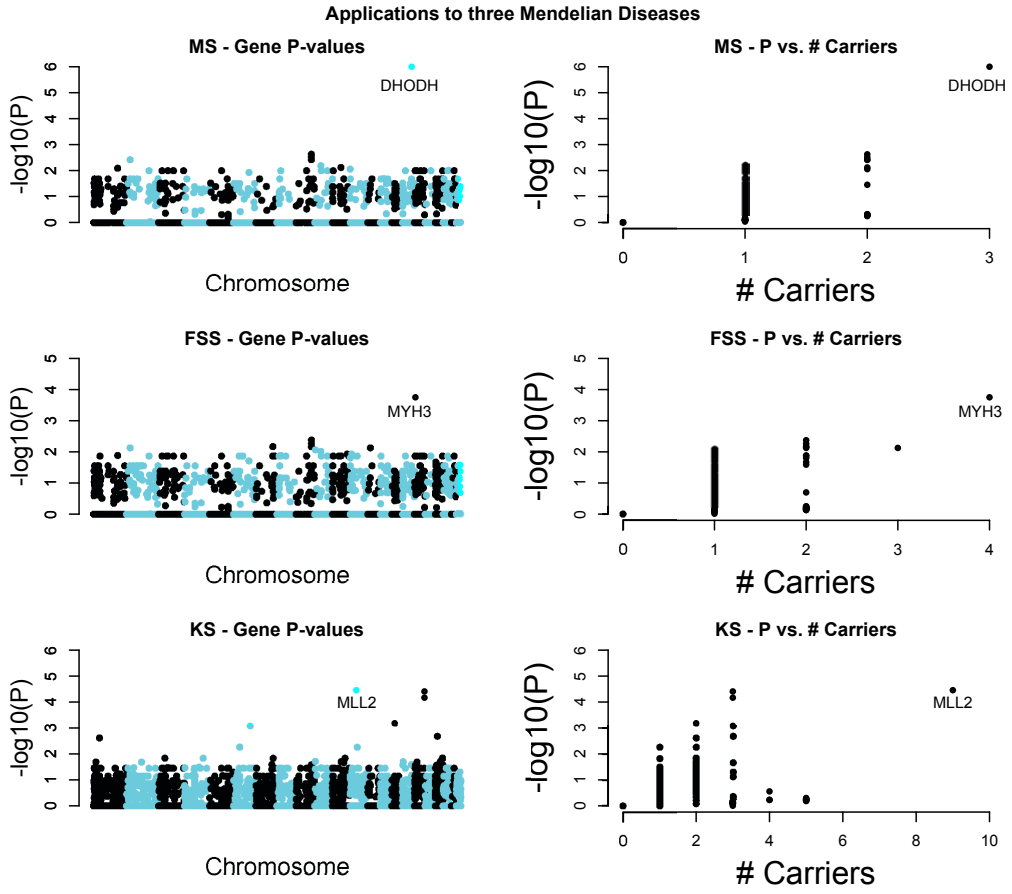
Applications to three Mendelian Diseases
MS - Gene P-values
DHODH
-log10(P)
Chromosome
MS - P vs. # Carriers
DHODH
-log10(P)
Carriers
FSS - Gene P-values
MYH3
-log10(P)
Chromosome
FSS - P vs. # Carriers
MYH3
-log10(P)
Carriers
KS - Gene P-values
MLL2
-log10(P)
Chromosome
KS - P vs. # Carriers
MLL2
-log10(P)
Carriers

Figure 2:

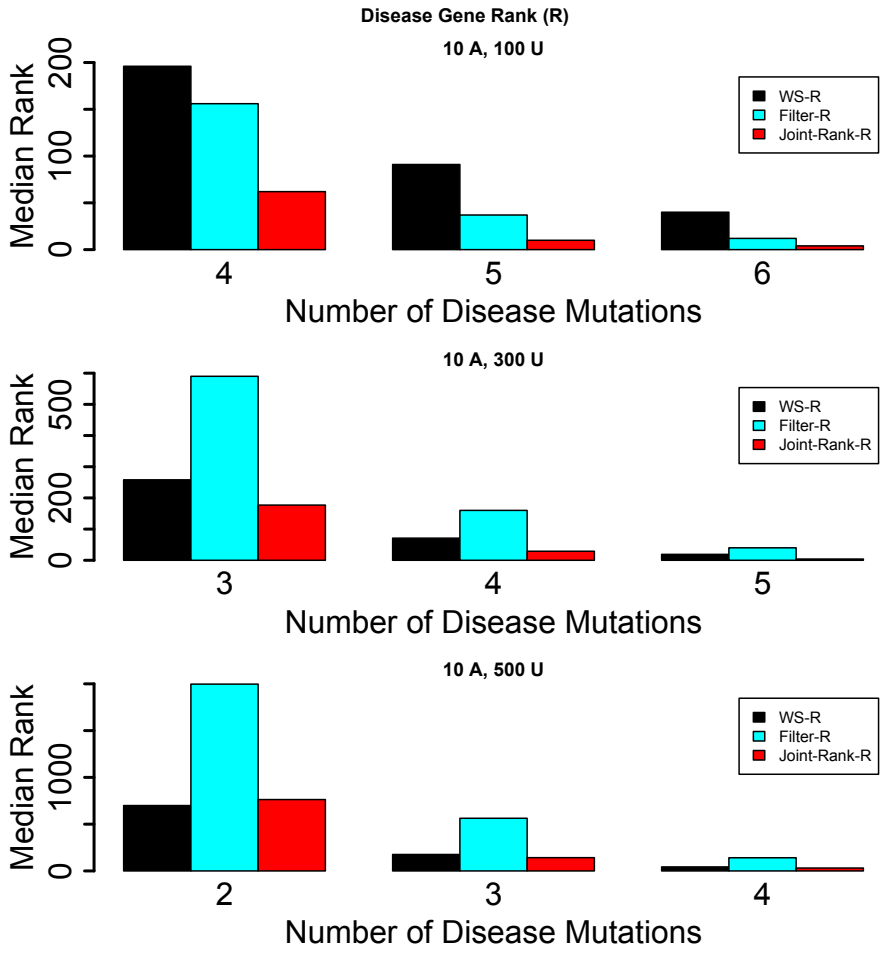

Figure S1: The median rank (with all rare variants considered) of a disease gene in genome scans with $20,000$ genes, with gene length sampled from the real gene length distribution. 1000 such genome-scans are simulated. $2-6$ of 10 affected individuals are assumed to carry a novel disease mutation in the disease gene. The following methods are compared: WS-R, Filter-R, and Joint-Rank-R.

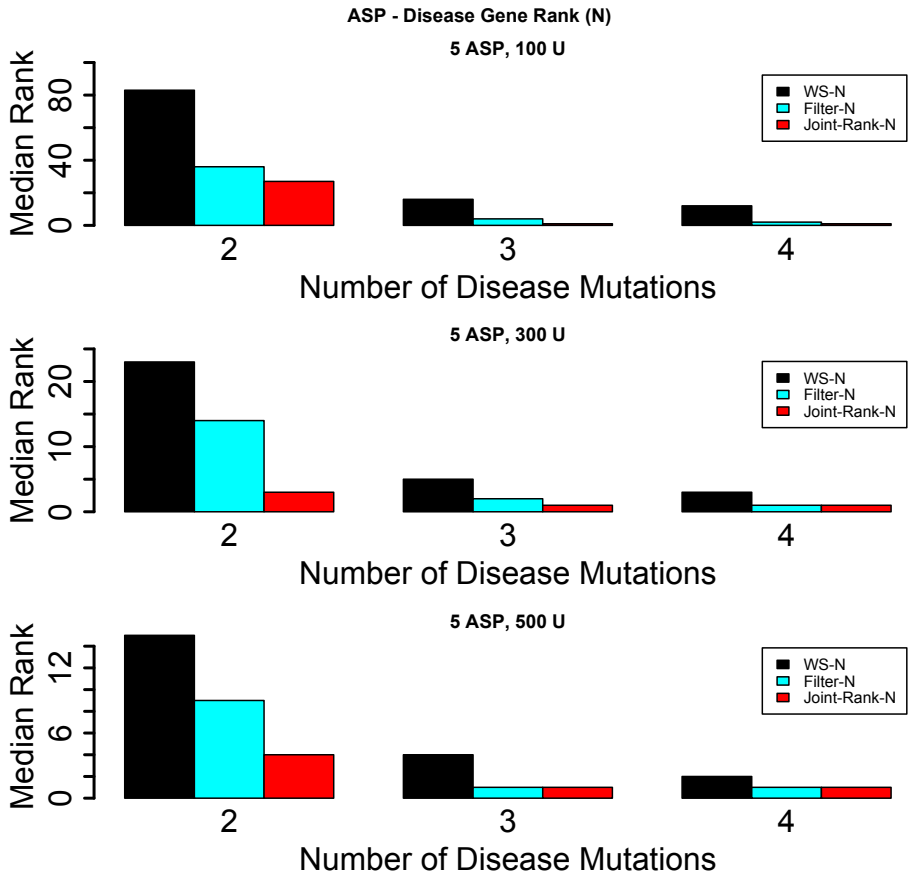

Figure S2: The median rank of a disease gene in genome scans with $20,000$ genes, with gene length sampled from the real gene length distribution. 1000 such genome-scans are simulated. $2-4$ of 5 affected sib-pairs (ASP) are assumed to share a novel disease mutation in the disease gene. The following methods are compared: WS-N, Filter-N, and Joint-Rank-N.

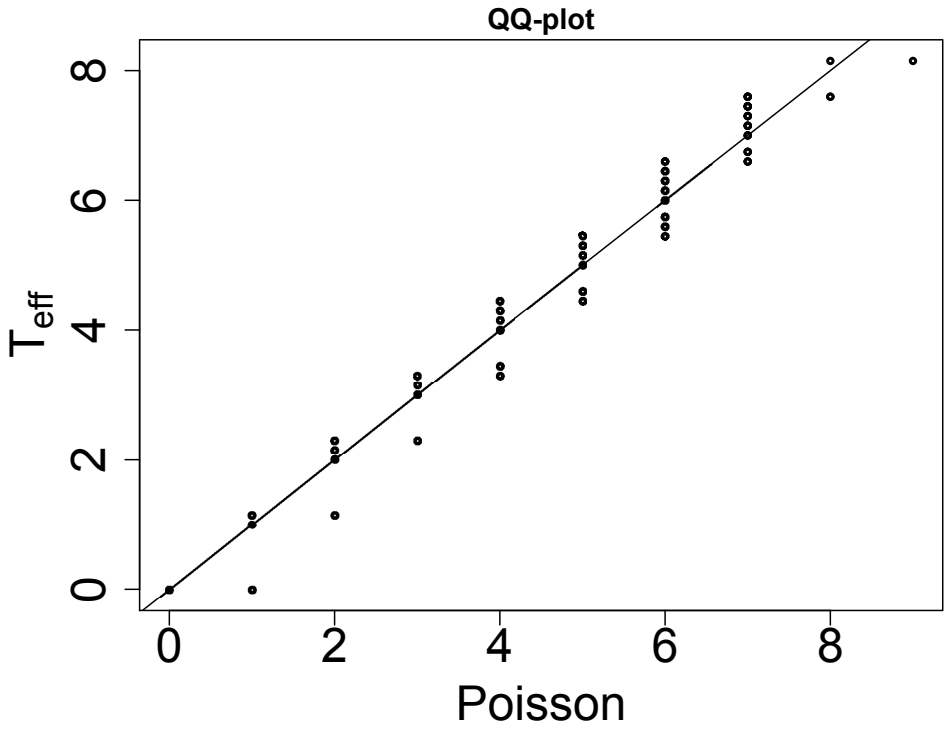

Figure S3: QQ-plot showing distribution of T_{eff} vs. Poisson(E(T_{eff})). 100 ASPs and 500 controls are simulated for a total of 30000 simulations.

Table S1: The *effective* number of variants at a rare variant position in two related heterozygous individuals, as defined in text; φ is the kinship coefficient. Results for $f = 0.01$ are shown.

Relationship	φ	k_{eff}
Identical twins	1/2	1.00
Parent-child	1/4	1.17
Sibs	1/4	1.17
Half-sibs	1/8	1.34
Uncle-nephew	1/8	1.34
First Cousins	1/16	1.50
First Cousins-1 (once removed)	1/32	1.64
Second Cousins	1/64	1.76
Unrelateds	0	2.00

Table S2: Type-1 error for the Sib-Pair Design.

Approach	A[a]	U[b]	α			
			10^{-4}	10^{-3}	10^{-2}	$5 \cdot 10^{-2}$
WS-R	5	100	$1.7 \cdot 10^{-4}$	$8.0 \cdot 10^{-4}$	$4.7 \cdot 10^{-3}$	$2.0 \cdot 10^{-2}$
		500	$1.0 \cdot 10^{-4}$	$7.4 \cdot 10^{-4}$	$5.5 \cdot 10^{-3}$	$2.6 \cdot 10^{-2}$
		1000	$1.4 \cdot 10^{-4}$	$7.0 \cdot 10^{-4}$	$4.9 \cdot 10^{-3}$	$2.5 \cdot 10^{-2}$
	10	100	$1.0 \cdot 10^{-4}$	$5.0 \cdot 10^{-4}$	$3.8 \cdot 10^{-3}$	$1.8 \cdot 10^{-2}$
		500	$1.1 \cdot 10^{-4}$	$9.8 \cdot 10^{-4}$	$6.0 \cdot 10^{-3}$	$2.7 \cdot 10^{-2}$
		1000	$1.5 \cdot 10^{-4}$	$9.9 \cdot 10^{-4}$	$5.9 \cdot 10^{-3}$	$2.7 \cdot 10^{-2}$
WS-N	5	100	$1.0 \cdot 10^{-4}$	$4.5 \cdot 10^{-4}$	$2.2 \cdot 10^{-3}$	$8.0 \cdot 10^{-3}$
		500	$2.7 \cdot 10^{-5}$	$2.7 \cdot 10^{-4}$	$4.9 \cdot 10^{-4}$	$2.4 \cdot 10^{-3}$
		1000	$2.4 \cdot 10^{-5}$	$5 \cdot 10^{-5}$	$3.0 \cdot 10^{-4}$	$1.5 \cdot 10^{-3}$
	10	100	$4.9 \cdot 10^{-5}$	$2.5 \cdot 10^{-4}$	$1.4 \cdot 10^{-3}$	$6.7 \cdot 10^{-3}$
		500	$2.0 \cdot 10^{-5}$	$1.0 \cdot 10^{-4}$	$3.8 \cdot 10^{-4}$	$1.7 \cdot 10^{-3}$
		1000	$4.9 \cdot 10^{-5}$	$1.0 \cdot 10^{-4}$	$2.9 \cdot 10^{-4}$	$1.4 \cdot 10^{-3}$

[a] #affected sib-pairs
[b] #unrelated unaffected individuals

Table S3: Simulation results for T_{eff}.

f	N_{sibs}	N_{controls}	f		$\widehat{f}$		Cor[a]	Theoretical	
			$\widehat{\mu}$	$\widehat{\text{var}}$	$\widehat{\mu}$	$\widehat{\text{var}}$		μ	var
0.01	5	100	0.152	0.163	0.152	0.163	0.999915	0.156	0.161
		500	0.153	0.151	0.153	0.151	0.999968	0.156	0.161
		1000	0.162	0.168	0.162	0.168	0.999986	0.156	0.161
0.001	5	100	0.017	0.018	0.017	0.018	0.999864	0.016	0.016
		500	0.014	0.015	0.014	0.015	0.999984	0.016	0.016
		1000	0.016	0.017	0.016	0.017	0.999985	0.016	0.016

[a]Correlation between $T_{\text{eff}}(f)$ and $T_{\text{eff}}(\widehat{f})$

Table S4: Theoretical results for T_{eff}.

Relationship	f	N	Theoretical	
			μ	var
Siblings	0.01	5	0.156	0.161
	0.001		0.016	0.016
First Cousins	0.01	5	0.191	0.194
	0.001		0.019	0.019
Second Cousins	0.01	5	0.197	0.196
	0.001		0.019	0.020

Table S5: Gamma-based approximation of weighted-sum of M Poisson RVs.

M	α				
	10^{-5}	10^{-4}	10^{-3}	10^{-2}	0.05
3	$4.7 \cdot 10^{-6}$	$5.3 \cdot 10^{-5}$	$4.3 \cdot 10^{-4}$	$6.3 \cdot 10^{-3}$	$2.8 \cdot 10^{-2}$
5	$1.0 \cdot 10^{-5}$	$9.3 \cdot 10^{-5}$	$8.3 \cdot 10^{-4}$	$7.0 \cdot 10^{-3}$	$3.3 \cdot 10^{-2}$
20	$1.0 \cdot 10^{-5}$	$9.3 \cdot 10^{-5}$	$8.4 \cdot 10^{-4}$	$8.0 \cdot 10^{-3}$	$3.9 \cdot 10^{-2}$
40	$1.0 \cdot 10^{-5}$	$9.7 \cdot 10^{-5}$	$8.8 \cdot 10^{-4}$	$8.4 \cdot 10^{-3}$	$4.1 \cdot 10^{-2}$