

Inferring Disease Risk Genes from Sequencing Data in Multiplex Pedigrees Through Sharing of Rare Variants

Running title: Sharing of Rare Variants to Infer Disease Genes

Alexandre Bureau^{1,2*}, Ferdouse Begum³, Margaret A. Taub⁴,
Jacqueline Hetmanski³, Margaret M. Parker⁵, Hasan Albacha-Hejazi⁶,
Alan F. Scott⁷, Jeffrey C. Murray⁸, Mary L. Marazita⁹,
Joan E. Bailey-Wilson¹⁰, Terri H. Beaty³, Ingo Ruczinski⁴

¹ Département de Médecine Sociale et Préventive, Université Laval, Québec, Canada.

² Centre de recherche CERVO, Québec, Canada.

³ Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA.

⁴ Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA.

⁵ Channing Division of Network Medicine, Harvard Medical School, Boston, MA, USA.

⁶ Prime Health Clinic Jeddah, Riyadh, Saudi Arabia.

⁷ Institute of Genetic Medicine, Johns Hopkins Medical Institutions, Baltimore, MD, USA.

⁸ Department of Pediatrics, School of Medicine, University of Iowa, Iowa City, IA, USA.

⁹ Department of Oral Biology, School of Dental Medicine, University of Pittsburgh, Pittsburgh, PA, USA.

¹⁰ Inherited Disease Research Branch, National Human Genome Research Institute, Baltimore, MD, USA.

*To whom correspondence should be addressed: Département de Médecine Sociale et Préventive, Université Laval, Québec,

1 Abstract

2 We previously demonstrated how sharing of rare variants (RVs) in distant affected relatives can be used to
3 identify variants causing a complex and heterogeneous disease. This approach tested whether single RVs
4 were shared by all sequenced affected family members. However, as with other study designs, joint analysis
5 of several RVs (e.g. within genes) is sometimes required to obtain sufficient statistical power. Further,
6 phenocopies can lead to false negatives for some causal RVs if complete sharing among affecteds is required.
7 Here we extend our methodology (Rare Variant Sharing, RVS) to address these issues. Specifically, we
8 introduce gene-based analyses, a partial sharing test based on RV sharing probabilities for subsets of affected
9 relatives and an haplotype-based RV definition. RVS also has the desirable features of not requiring external
10 estimates of variant frequency or control samples, provides functionality to assess and address violations of
11 key assumptions, and is available as open source software for genome-wide analysis. Simulations including
12 phenocopies, based on the families of an oral cleft study, revealed the partial and complete sharing versions
13 of RVS achieved similar statistical power compared to alternative methods (RareIBD and the Gene-Based
14 Segregation Test), and had superior power compared to the pedigree Variant Annotation, Analysis and
15 Search Tool (pVAAST) linkage statistic. In studies of multiplex cleft families, analysis of rare single
16 nucleotide variants in the exome of 151 affected relatives from 54 families revealed no significant excess
17 sharing in any one gene, but highlighted different patterns of sharing revealed by the complete and partial
18 sharing tests.

19 **KEYWORDS:** Family studies; identity by descent; oral clefts; variant sharing.

1 Sequencing distant relatives is an established approach to identify causal variants for Mendelian disorders
2 [e.g., Ng et al. 2010, Bamshad et al. 2011, Ionita-Laza et al. 2011]. Typically external databases are
3 combined with variant filtering strategies to identify causal variants under the assumption of complete
4 penetrance and the absence of phenocopies. Sequencing related individuals has also become an option to
5 identify causal variants in non-Mendelian complex disorders [Aida et al. 1998, Dahl et al. 2001, Daoud et al.
6 2009], although the rationale and strategies employed are more complicated. When familial phenotype
7 aggregation is observed at a rate much higher than the prevalence in the general population, possible
8 explanations include a shared familial environment or a relatively large familial "gene burden" [Isobe et al.
9 2013, Mescheriakova et al. 2016]. Another explanation could be the presence of a rare but highly penetrant
10 (Mendelian or near-Mendelian) causal variant. This phenomenon has been observed in many common
11 complex diseases and disorders such as chronic obstructive pulmonary disease, breast and ovarian cancer,
12 birth defects, Alzheimer's disease, and amyotrophic lateral sclerosis [Miki et al. 1994, Szabo and King 1995,
13 Stratton 1996, Papassotiropoulos et al. 2006, Johnston et al. 2012, Bureau et al. 2014a].

14 We recently devised a statistical framework for such a setting, based on the notion that sequencing DNA in
15 extended multiplex families can help to identify high penetrance causal variants too rare in the population to
16 be detected through tests of association in population based studies, but co-segregating with disease within
17 families [Bureau et al. 2014b]. Specifically, when only a few affected subjects per family are sequenced,
18 evidence any one rare variant (RV) may be causal can be quantified from the probability of sharing
19 alleles by all affected relatives, given it was seen in any one family member under the null hypothesis of
20 complete absence of linkage and association. We presented a general framework for calculating such sharing
21 probabilities when two or more affected subjects per family are sequenced, and showed how information
22 from multiple families can be combined by calculating a p-value as the sum of the probabilities of sharing
23 events at least as extreme [Bureau et al. 2014b]. We refer to this approach as RVS for Rare Variant Sharing.
24 By sequencing three affected second cousins from a multiplex oral cleft family, we successfully employed
25 this approach to identify a causal nonsense mutation in the gene *CDH1* [Bureau et al. 2014a].

1 Alternative approaches have also been proposed for the setting where the causal variant is rare, in the
2 sense that when it is seen identical by state (IBS) in multiple affected relatives it has to be identical
3 by descent (IBD). Instead of using exact sharing probabilities, Sul et al. [2016] proposed to use as test
4 statistic the sum of the number of affected subjects sharing a RV and the number of unaffected subjects
5 without the RV (either of which can be zero), standardized by subtracting its expectation and dividing
6 by its standard deviation within each family under the null hypothesis and the assumption that only one
7 founder in the family introduced a causal variant. Further, in both approaches inference is conditional on
8 the introduction of the RV by a single founder, and as a consequence we do not need to know or estimate
9 its population allele frequency. This is a great benefit, as the allele frequency is commonly unknown,
10 especially when sequenced families come from a genetic background not well represented in the standard
11 reference databases such as the Exome Sequencing Project [Fu et al. 2013] or the 1000 Genomes Projects
12 [1000 Genomes Project Consortium et al. 2015] , or the sample is a mix of families from different genetic
13 backgrounds. Qiao et al. [2017] proposed GESE, a gene-based segregation test requiring an estimate of
14 variant frequencies (as compared to calculating the probability of sharing conditional on the variant being
15 observed), but otherwise relying on very similar assumptions as our previously introduced RVS approach
16 [Bureau et al. 2014a]. Specifically, GESE (like RVS and RareIBD) assumes only one founder in the family
17 introduced a causal variant in a gene, and Qiao et al. [2017] recommend limiting the tests to variants with
18 high functional impact. Further, GESE also calculates the p-value as the sum of the probabilities of all
19 events as or less likely as the observed event. However, in addition to absence of phenocopies (i.e., all
20 affected subjects in a family are carriers), GESE assumes complete penetrance (i.e., all unaffected subjects
21 do not carry the putative causal variant), while our original variant sharing approach was based only
22 on sharing among affected subjects, and did not make any assumptions about unaffected subjects. The
23 latter feature can be critically important, as was the case in detection of a nonsense mutation in *CDH1*
24 shared among three affected second cousins, but also present in the unaffected parents who transmitted
25 this variant [Bureau et al. 2014a, Figure 1]. We note GESE and RareIBD can be applied in "affected-only"

1 mode by setting the phenotype of unaffected subjects to unknown, even though not intended as such.

2 When reliable information about allele frequencies is available, some authors have argued for combining
3 a linkage signal with the association signal derived using known allele frequencies to increase statistical
4 power. Hu et al. [2014] proposed an extension of their previously developed Variant Annotation, Analysis
5 and Search Tool (VAAST), a likelihood-ratio-based RV association test that incorporates case/control
6 allele frequency differences and functional annotation into the likelihood. The aptly named pedigree-
7 VAAST (pVAAST) employs a sequence-based model, where variants are tested for being directly causal
8 instead of merely linked to some unobserved disease variant (as in classical linkage analysis) in variant and
9 gene-based linkage analysis, and offers the option to combine this linkage information with the VAAST
10 statistic. GESE calculates a probability of segregation combining association and linkage signals requiring
11 knowledge (or an accurate estimate) of the variant frequency in the population. This feature of GESE
12 has the potential to increase power, but misspecification of variant frequencies, which is likely in family
13 samples with distinctly different or mixed genetic backgrounds, may yield spurious signals. Methods
14 relying on filtering of variants based on frequencies in external databases such as our RVS approach and
15 RareIBD are also susceptible to false positive signals from variants common in a study population but not
16 well represented in the utilized reference databases. However, the haplotype structure around a variant
17 contains information on the population frequency of variant that can be exploited to filter out common
18 variants without producing a frequency estimate as required by pVAAST and GESE.

19 GESE, pVAAST and RareIBD are designed for analysis of all RVs in a gene or genomic region, as is standard
20 when analyzing RVs to increase the proportion of subjects or families with at least one RV [Li and Leal 2008,
21 Lee et al. 2014]. Here we present an extension of our previously published single-variant RVS method for a
22 gene-based approach. Acknowledging phenocopies, diagnosis error and intra-familial genetic heterogeneity
23 exist in complex disorders, we further extend our method by relaxing the previous assumption that all
24 sequenced affected subjects must be carriers of the same causal variant, and by introducing an approach
25 based on haplotypes of known variants to detect which variants are not actually rare in the study population

1 despite being rare or completely absent in reference databases. Comparing our gene-based RV sharing
2 approach to alternative methods, we show in a simulation study that knowledge and use of allele frequencies
3 of rare variants in approaches such as GESE and the pVAAST linkage statistic does not lead to power
4 gains over methods such as RVS and RareIBD, which do not require knowledge of such frequencies and are
5 therefore more universally applicable. An implementation of our method RVS is available as open source
6 software from the Bioconductor project at bioconductor.org/packages/release/bioc/html/RVS.html.

7 **Methods**

8 **Gene-based analysis:** We initially presented the RVS approach for single variants [Bureau et al. 2014b].
9 As long as there is a single RV within a gene in the same family, the RVs are independent since the families
10 are unrelated, and the information can simply be pooled together and analyzed jointly. The abundance of
11 RVs in the human genome, however, implies that multiple RVs are likely to occur on the same haplotype
12 over a region such as a gene. Such RVs have identical sharing patterns, and are indistinguishable in
13 genetic analysis. Therefore, we redefine the units of analysis as the haplotypes of RVs over each genomic
14 region instead of individual RVs themselves. Simply taking the minimal RV sharing probability among
15 all RVs in the same gene in a family has the effect of merging RVs on the haplotype with the lowest
16 sharing probability. We detail in the section "Recoding rare variants haplotypes for rare variant sharing
17 computations" on page 7 a systematic approach to recode RVs into haplotypes based on genomic sequence.
18 After this recoding, when two or more RVs (or haplotypes of several RVs) remain in the same gene in a
19 family, we retain one RV per family to compute the RV sharing probability. We propose using those RVs
20 with the sharing pattern yielding the lowest probability among all RVs present in the same gene. When we
21 do this, the test is no longer exact; the resulting p-value becomes an approximation of the exact p-value.
22 We examine the impact of this practice on Type I error in a simulation study, and in sequencing data from
23 individuals drawn from multiplex cleft families.

1 **Partial sharing:** We define the following random variables:

2 C_i represents the number of copies of the RV observed in the sequence of subject i ,

3 F_j is the indicator variable that founder j introduced one copy of the RV into the pedigree, and

4 D_{ij} is the number of generations (meioses) between subject i and his or her ancestor j .

5 For a set of n sequenced subjects for which the pedigree structure limits to one the number of copies of
6 the RV that they can share, we compute the probability that any subset of size $k \leq n$ shares a given RV.

7 Without loss of generality, we assume the n subjects are ordered such that subjects $1, \dots, k \leq n$ share the
8 RV, and thus:

$$\begin{aligned}
 & P(\text{ the subset shares the RV }) \\
 = & P(C_1 = \dots = C_k = 1, C_{k+1} = \dots = C_n = 0 \mid C_1 + \dots + C_n \geq 1) \\
 = & \frac{P(C_1 = \dots = C_k = 1, C_{k+1} = \dots = C_n = 0)}{P(C_1 + \dots + C_n \geq 1)}. \tag{1}
 \end{aligned}$$

9 To simplify the notation, we define $K = C_1 + \dots + C_n$. For a single family, we define a RV sharing
10 configuration where k subjects share a RV as $G_k = (C_1, \dots, C_n) \mid K = k$. The p-value of this configuration
11 G_k is the sum of probabilities of all sharing configurations g with probability $P_g = P(g \mid K \geq 1)$ lower or
12 equal to the probability of the observed configuration $P_{G_k} = P(G_k \mid K \geq 1)$, and with size $k^{(g)} \geq k$, i.e.:

$$p = \sum_{g \mid \{ P_g \leq P_{G_k} \text{ and } k^{(g)} \geq k \}} P_g \tag{2}$$

13 We note it is computationally advantageous to identify classes of all equiprobable configurations, to compute
14 their probability only once for a member of each class, and to multiply the probability by the number of
15 equiprobable configurations. Classes of equiprobable configurations are defined by exchangeable relatives
16 (e.g. siblings), and subsets of exchangeable relatives (e.g. sibships who represent sets of first cousins).

17 With M families where the same RV is observed, or where M distinct RVs are observed (a different one
18 in each family) in the same region (such as a gene), the configuration $G_k = (G_{k_1}, \dots, G_{k_M})$ is a vector

1 of family-specific configurations containing the k_1, \dots, k_M subjects sharing a RV in the M families, with
2 $k = \sum_{m=1}^M k_m$. The p-value is then computed by applying the same criteria used for a single family, with
3 the only modification to the probability of a global configuration $g = (g_1, \dots, g_M)$ with $k^{(g)} = \sum_{m=1}^M k_m^{(g)}$
4 affected subjects being the product over the probabilities of all family-specific configurations g_1, \dots, g_M
5 under the independence assumption:

$$p = \sum_{g \mid \{ P_g \leq P_{G_k} \text{ and } k^{(g)} \geq k \}} \prod_{m=1}^M P_{g_m} \quad (3)$$

6 In the current implementation, M is limited to 10 as the summation grows exponentially with M . This
7 M is the number of families with a RV, not the maximum size of the family sample, as only a fraction of
8 families harbor RVs in any given gene.

9 **Refining the definition of RVs:** The assumption that all copies of a RV seen in related members of a
10 pedigree are IBD is crucial to the validity of the RVS tests. Instead of relying solely on filtering of common
11 variants based on their frequencies in external reference databases, genome-wide genotype data enables
12 the identification of variants actually introduced multiple times to the family from unrelated founders and
13 most likely not rare in the population from which the family is drawn. Our approach is based on haplotypes
14 of known variants (common and rare). RVs seen on two or more haplotypes are discarded as introduced
15 multiple times into the family, e.g. being IBS without being IBD. We infer haplotypes by phasing the
16 sequence data in families prior to analysis. Given the need to include a sufficient number of common variants
17 within a genomic interval spanning a gene, and to ensure the quality of phasing, this approach is advisable
18 only with genomic sequence and not with capture-based exome sequencing data alone. Our implementation
19 uses Shapeit2 available at mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html, with
20 the duoHMM option to improve phasing where parent-offspring pairs are sequenced [O'Connell et al. 2014].
21 Haplotypes for the purpose of determining IBD status are comprised of variants previously observed in
22 genomic sequence databases (such as the 1000 Genomes project), as these variants are less likely to be
23 sequencing errors. They may include RVs included in the analysis, depending on the filter used to define

1 a RV. We have opted to include all variants within the RefSeq gene boundaries in the inferred haplotypes.

2 **Recoding rare variant haplotypes for rare variant sharing computations:** Once RVs have been
3 assigned to haplotypes, those haplotypes containing at least one RV meeting the filtering criteria are recoded
4 as a new synthetic RV representing all the RVs on that haplotype. Hence, there are as many recoded RVs in
5 the analysis region as there are different haplotypes bearing at least one RV. The RV sharing probabilities
6 are then computed for each recoded RV. In the same spirit, Sul et al. [2016] considered variants with the
7 same genotype in all family members as duplicates and used one of those variants in the computation of
8 the RareIBD statistic.

9 **Implementation:** Our RVS approach is freely available through the RVS Bioconductor package [Sherman
10 et al. 2018]. Briefly, the main function `RVgene` takes as input a data frame with pedigree and genotype
11 data in standard formats (two alleles of a variant on two consecutive columns or minor allele count on
12 one column) In addition, lists of RV sharing probabilities (pre-computed by the `RVsharing` function)
13 and numbers of affected subjects for each possible sharing configuration of the sequenced affected members
14 from every family must be provided to the `RVgene` function to perform the test allowing for partial sharing.
15 Due to the computational demand of the convolution of the RV sharing event distribution for all families
16 involved, the number of families with RVs in the same gene is currently limited to 10, or fewer depending
17 on available RAM (the complete sharing test does not have this limitation). Another new feature of
18 the RVS package is its correction of sharing probabilities to account for cryptic relatedness using the
19 analytical approximation described in Bureau et al. [2014b], based on a reimplement of the RV
20 sharing probability computation using the `gRain` package for general computations on Bayesian networks.
21 The previous Monte Carlo approximation for cryptic relatedness correction has also been reimplemented.

22 **Simulation study:** We used 594 phased sequences from the CEU, TSI and GBR samples of the 1000
23 Genomes Project to define a population of haplotypes for each gene in the genome. Single nucleotide
24 variants (SNVs) seen more than once in this sample were used to define recoding haplotypes. RVs were

1 defined as SNVs with frequency $< 1\%$ among the 594 sequences (up to 5 copies). To evaluate haplotype
2 recoding in the context of a sequencing study with a family sample, we assumed a genetic origin to the
3 disease in all families and used the 47 simple pedigree structures (i.e., not allowing inbreeding or marriage
4 loops, removing 7 inbred pedigrees to limit computational complexity) from the second set of multiplex
5 cleft families (see section "Multiplex oral cleft families" on page 10). To evaluate statistical properties of
6 this test, focusing on distant affected relatives for which the RVS tests are designed, we further removed
7 the 14 pedigrees with affected 1st degree relatives, leaving 33 simple pedigrees including 93 affected 2nd –
8 9th degree relatives. We assumed the disease had a genetic origin in 16 families (randomly sampled at each
9 replicate) or about 50% of all available families. For each RefSeq gene, each family where the disease had a
10 genetic origin was assigned a RV as potential disease causing variant (a distinct RV for each family if there
11 were enough RVs in the gene, otherwise RVs were reused across multiple families). The genotype at the
12 causal RV site was generated conditional on disease status under the Risch 2-locus heterogeneity model
13 of disease [Risch 1990, and Table 1], which means the potential causal RV was not necessarily present
14 in the family (the disease could be caused by the other unlinked gene). The disease prevalence implied
15 by this model was 2.5%. Founder haplotypes were then sampled from the population of all haplotypes,
16 and transmission to descendants was simulated without recombination. Both haplotype sampling and
17 transmission were conditional on genotypes at the causal RV site. In families where the disease did not
18 have a genetic origin (a subset of families for power evaluations, all families for assessment of Type I error
19 under the null hypothesis), founder haplotype sampling was performed unconditionally, and transmission
20 was simulated under Mendelian segregation instead.

21 We compared the statistical power of gene-based tests including the proposed RVS test allowing for partial
22 sharing, the original RVS test (complete sharing), RareIBD [Sul et al. 2016], as well as two competing tests
23 requiring external variant frequency estimates, namely GESE [Qiao et al. 2017] and the pVAASST LOD
24 score [Hu et al. 2014]. We define RVs as variants with a minor allele frequency (MAF) less than 1%, and
25 evaluated power for a gene with few coding RVs (*PEAR1*, 7 coding RVs) and a gene with many coding RVs

Locus	MAF	Penetrance
Tested	10^{-4}	$x \in \{0.25, 0.50, 0.75\}$
Other	0.1	0.04
None (phenocopy)	NA	0.02

Table 1: Model parameters used in the simulations (MAF: minor allele frequency).

1 (*CDH13*, 20 coding RVs). Separate simulation using RV penetrances of 0.25, 0.5 and 0.75 were conducted,
2 corresponding to relative risks of about 10, 20 or 30 under the respective genetic models (Table 1). We
3 also assessed Type I error of the gene-based tests in the context of a gene with many rare coding RVs
4 (*CDH13*), where there were usually multiple RV-carrying haplotypes in the same gene in the same family.
5 Both GESE (in the case where no external estimates are available, for example from public databases) and
6 pVAAST require control samples, so we created samples of 5,000 control subjects by randomly sampling
7 two haplotypes for each subject from the 594 phased 1000 Genome sequences. This is the minimal size
8 recommended by Qiao et al. [2017] to ensure a correct Type I error for GESE. Variants with MAF less
9 than 1% in the control data were selected. For pVAAST, we specified a genetic model where the tested RV
10 was assumed to be dominant with a penetrance in the range 0.5 – 1.0 (the interval over which pVAAST
11 maximizes the LOD score). Since pVAAST requires specifying the amino acid substitution at the tested RV,
12 we assumed a substitution with a mild effect (alanine to valine). The severity of the mutation is actually
13 irrelevant for the linkage LOD score statistic, but is taken into account in the case-control composite
14 likelihood ratio test (CLRT) which is added to the linkage LOD score to obtain the CLRT for pedigrees
15 (CLRTp). Due to high computational demands of pVAAST, the number of Monte Carlo simulations to
16 estimate the p-value was set to 10,000. For RareIBD and GESE, the number of resampling simulations
17 was adaptively selected, increasing up to 10^7 for very small p-values. For GESE, we evaluated the Type
18 I error with control samples drawn from a population with the correct MAF and a population where the
19 true MAF of the RVs was 10 times lower to assess the impact of MAF misspecification. Again, due to

1 excessive computational burden, assessment of the Type I error for pVAAST was limited to 200 replicates
2 and was performed only under the latter control sample definition.

3 **Multiplex oral cleft families:** The first set of 55 multiplex cleft families was described in detail in
4 Bureau et al. [2014a]. In brief, the families were mostly comprised of pairs of distant affected relatives
5 with a few triples of distant affected relatives, on which whole exome sequencing (WES) was performed.
6 We refer to this sample as the "WES sample". Two of the multiplex families in the WES sample were
7 expanded to additional affected relatives, and included in the second set of families described below. For
8 that reason, these two families were removed from the WES sample in the current analysis, leaving 53
9 multiplex families. The second set included 54 multiplex cleft families from the Philippines, the United
10 States (European ancestry), Guatemala and the Syrian Arab Republic. Whole genome sequencing (WGS)
11 data were generated for 153 affected relatives and 7 unaffected relatives (all unaffected individuals were from
12 Filipino families). We thus call this set the "WGS sample". In one Syrian pedigree, the 8 affected relatives
13 did not have any known common ancestor, so a sub-pedigree including 6 affected relatives descending from
14 a common couple of ancestors was used in the analysis, reducing the total number of affected subjects to
15 151 (Table 2). The sequencing, alignment, and variant calling process was described in Holzinger et al.
16 [2017], who also reported on RVs observed in the WGS data from the Filipino and Syrian families.

17 In addition to a large proportion of families with four or more affected relatives, the WGS sample differs
18 from the WES sample by the presence of first degree relatives (Supplementary Table 1). Thus, the distri-
19 bution of the $-\log_{10}$ probabilities of sharing a RV by all affected relatives within the respective families
20 (which is the potential $-\log_{10}$ p-values when a RV is present in a single family) was more dispersed in the
21 WGS than the WES sample (Supplementary Figure 4).

22 SNVs from both WES and WGS data were annotated using wAnnovar (wannovar.wglab.org) in Novem-
23 ber 2016. Exonic and splice site SNVs were extracted and filtered using the same criteria as described in
24 Bureau et al. [2014b], with the additional requirement of a maximum frequency of 1% in all population

	WES		WGS	
	S	F	S	F
Syrian	16	8	35	14
Filipino	22	11	76	18
Indian	26	12	12	6
German	38	19	0	0
Chinese/Asian	4	2	31	16
European Ancestry	2	1	12	5
Guatemalan	0	0	4	2
Total	108	53	151	54

Table 2: Number of affected individuals and families with nonsyndromic oral clefts, by DNA sequence approach (S: subjects; F: families).

1 samples from the gnomAD exome database (gnomad.broadinstitute.org), but dropping the previous
2 step of filtering against the internal database of the Center for Inherited Disease Research. For the WGS
3 data the above described haplotype-based approach was applied to ensure rare SNVs were introduced
4 only once in each family. The duoHMM algorithm of Shapeit2 [O’Connell et al. 2014] made use of 37
5 parent-child duos of sequenced subjects to improve phasing, while the remaining 112 sequenced subjects
6 were treated as unrelated (more distant sequenced relatives are not considered by the duoHMM algo-
7 rithm). In the WES data, haplotyping of SNVs on a SNP array (Illumina Omni Express) with Merlin
8 was used to infer IBD sharing from common variants, but did not use rare SNVs. Gene-level p-values
9 were computed for both partial and complete sharing tests, separately for the WES and WGS samples,
10 using the above described approach. Genome-wide association study results in the top genes detected by
11 the RVS analysis were extracted from the Facebase Human Genomics Analysis Interface, Pittsburgh, PA
12 (<http://facebase.sdmgenetics.pitt.edu>) [accessed June 20, 2018].

1 Results

2 **Recoding of rare variant haplotypes:** It is possible for two or more RVs on distinct haplotypes to be on
3 the same inferred haplotype (due to failure of the inferred haplotypes to distinguish all actual haplotypes),
4 thus being incorrectly recoded together as one RV. To assess the frequency of such an event, we applied
5 our RV haplotype recoding procedure to intervals defined from 5 kb upstream to the transcription end of
6 each RefSeq gene (hg19 assembly) in a dataset simulated as described in the "Simulation study" section
7 above. There were 18,272 genes with at least one coding RV in the CEU, TSI and GBR samples. Taking
8 the original estimated haplotypes as the truth and processing each family separately, we generated a total
9 of 450,114 family-specific haplotypes with at least one coding RV (an average of 24.6 haplotypes per gene,
10 or 0.54 per gene per family). For 99.7% of these RV-bearing haplotypes, all coding RVs recoded to that
11 haplotype were on the same actual haplotype.

12 **Type I error, power, and scalability:** We computed the potential p-values as defined by Bureau et al.
13 [2014b] for every replicate at every penetrance level. These potential p-values were generally less than
14 10^{-5} , and thus the simulated datasets were sufficiently informative to potentially reject the null hypothesis
15 at the adopted significance levels. In the simulation under the null hypothesis, computation of the partial
16 sharing gene-based test succeeded in 97% of 1000 replicates (those where RVs were seen in 10 or fewer
17 families), and computation of the complete sharing test always succeeded in this simulation. Type I error
18 was generally well controlled for RareIBD, the two versions of the RVS approach, and GESE when the
19 MAF was correctly estimated. A slight inflation at the most extreme p-values was observed however,
20 particularly for GESE (Figure 1). Type I error inflation was severe for the pVAAST CLRTp test, and
21 somewhat substantial for GESE and pVAAST LOD when the MAF was underestimated by a factor of 10.
22 The performance of the gene-based tests at detecting causal RVs are compared for *PEAR1* (Figure 2, left)
23 and *CDH13* (Figure 2, right). Computation of the subset sharing gene-based test succeeded in 96% and
24 70% of replicates (where RVs were seen in 10 or fewer families) for *PEAR1* and *CDH13*, respectively. Since

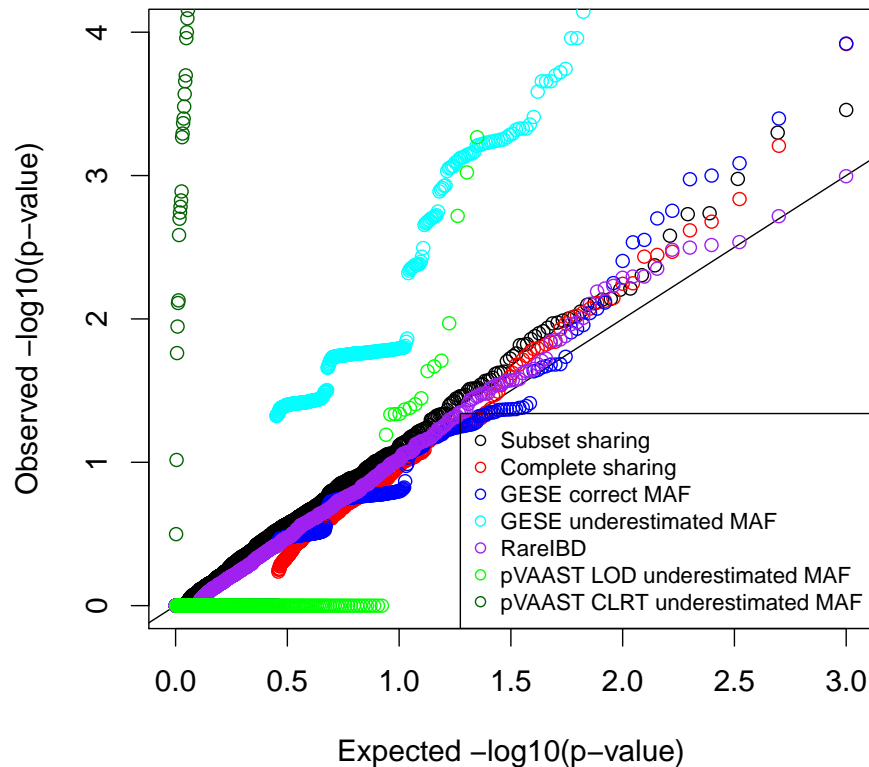


Figure 1: Expected and observed p-values under the null hypothesis in 1000 replicates (200 for pVAAST).

1 for pVAAST 10,000 simulations under the null did not allow us to estimate p-values below $\alpha = 1 \times 10^{-5}$,
2 we compared power of all methods at significance level $\alpha = 1 \times 10^{-4}$. In addition, Supplementary Figures 1
3 and 2 show the power at significance level $\alpha = 1 \times 10^{-5}$ for methods conditioning on the presence of at least
4 one variant in a gene, and for GESE the more stringent $\alpha = 2.5 \times 10^{-6}$ level, as suggested by Qiao et al.
5 [2017] to correct for the total number of genes in the genome. Similarly, we divided the significance level by
6 4 for the stringent GESE (Figure 2 with $\alpha = 2.5 \times 10^{-5}$). Generally, all tests performed similarly, except the
7 LOD test of pVAAST. RareIBD had in most instances slightly higher power than its nearest competitors.
8 The test allowing for sharing by a subset of subjects had slightly greater power than the complete sharing
9 test for the small *PEAR1* gene and slightly lower power than the complete sharing test for the larger
10 *CDH13* gene. GESE had lower power than RareIBD and the RVS tests when applying the stringent

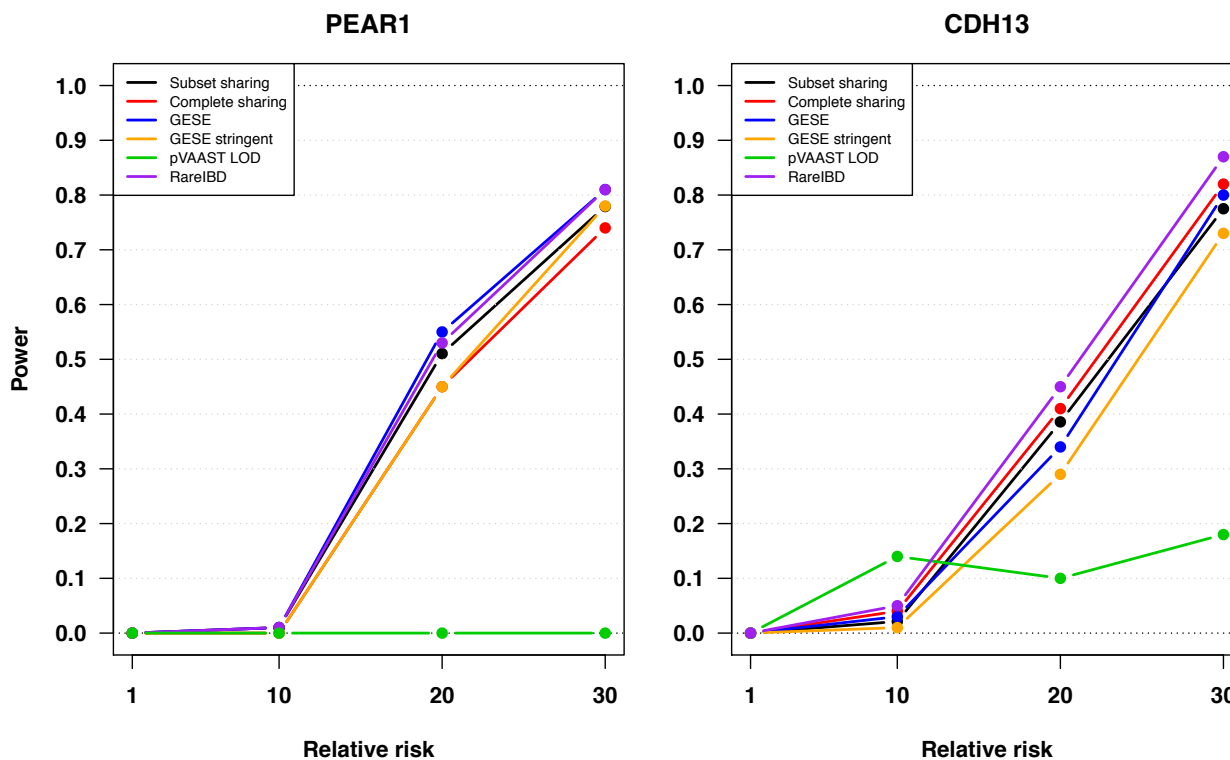


Figure 2: Power for genes *PEAR1* (left) and *CDH13* (right) under a genetic heterogeneity model with significance level $\alpha = 10^{-4}$.

1 significance level suggested for GESE by Qiao et al. [2017], for a fair comparison with tests conditioning
 2 on the presence of at least one variant. Only when using the same $\alpha = 1 \times 10^{-5}$ significance level as the
 3 other tests did GESE have slightly higher power in the small *PEAR1* gene (Supplementary Figure 1). The
 4 LOD test of pVAAST had low power, with the LOD score maximizing at 0 for all replicates in the small
 5 *PEAR1* gene and a majority of replicates in the larger *CDH13* gene. When the relative risk of causal RVs
 6 was equal to 10 and thus many unaffected carriers are expected, the linkage LOD test of pVAAST had
 7 higher power than the other methods in the large *CDH13* gene, but this power did not increase further at
 8 higher relative risks. The CLRTp combining linkage and case-control association signals gave the highest
 9 power under all alternative models for both genes, which is due to additional information beyond variant
 10 segregation within families (examined here). Computing times for testing coding RVs in the gene *CDH13*
 11 in one replicate of the simulated dataset exhibited dramatic differences in scalability (Table 3), with the

1 RVS complete sharing test taking the least computing time (even including the prior haplotype recoding)
2 followed by GESE and the RVS partial sharing test. We also examined the correlation of the $-\log_{10}$ p-
3 values of the five tests and found the partial and complete sharing tests, GESE and RareIBD to be highly
4 correlated, while these four tests are only weakly correlated to the pVAAST LOD test (Supplementary
5 Figure 3, illustrated using gene *CDH13* with relative risk 20).

Test	Time [sec]
RVS: haplotype recoding	2
RVS: complete sharing	1
RVS: partial sharing	37
GESE	18
RareIBD	2,392
pVAAST	20,617

Table 3: Running times (in seconds) for analyzing rare coding variants in the gene *CDH13* in one simulated dataset, using a single 2.8 GHz Intel E7-4870 processor of a HP DL580R07 computer. For RareIBD and GESE, the number of resampling simulations was 10^7 . For pVAAST, the number of Monte Carlo simulations to estimate the p-value was 10,000.

6 **Analysis of the oral cleft exome sequence data in the WGS sample:** There were 69,719 rare exonic
7 and splice site SNVs passing filtering criteria. The number of genes with rare SNVs for which the RVS
8 p-value computation succeeded in the WGS dataset was 12,706 for the partial, and 14,050 for the complete
9 sharing test. We calculated Bonferroni-corrected significance thresholds using the number of genes with
10 RVs for which the potential p-value as defined by Bureau et al. [2014b] would remain below 0.05 after
11 applying the correction, and obtained 6,372 genes with potential p-values below the threshold $0.05/6,372$
12 $= 7.8 \times 10^{-6}$ for a family-wise Type I error rate of 0.05 under the partial sharing test, and 7,647 values
13 below the threshold $0.05/7,647 = 6.5 \times 10^{-6}$ for a family-wise Type I error rate of 0.05 under the complete
14 sharing test.

Gene	Based on pedigree structure		Adjusted for unknown relationships	
	Partial sharing	Complete sharing	Partial sharing	Complete sharing
EP400NL	4.8×10^{-6}	1.8×10^{-6}	0.0028	0.0038
SPATA21	2.3×10^{-5}	3.6×10^{-6}	3.7×10^{-5}	8.0×10^{-6}
IL17RE	7.6×10^{-5}	0.037	0.0015	0.0540

Table 4: The three most significant genes based on the partial sharing test applied to exonic SNVs with frequency $\leq 1\%$ found in the whole genome sequence data of 151 relatives from a sample of 54 multiplex oral cleft families.

1 Among the top three genes based on the partial sharing test (Table 4), E1A binding protein p400 pseudo-
2 gene 1 (*EP400NL*) and spermatogenesis associated 21 (*SPATA21*) were also the top hits with the complete
3 sharing test, which returned smaller p-values than the partial sharing test. In both genes, shared variants
4 appear in families of Syrian origin, where unknown relationships are likely. We adjusted the RV sharing
5 probabilities based on a mean kinship of 0.013 among founders, estimated from the Syrian families of the
6 WES sample as we did previously [Bureau et al. 2014b]. This increased the sharing probabilities above
7 the Bonferroni corrected significance threshold. The partial sharing test also detected the gene interleukin
8 17 receptor E (*IL17RE*), with eight families carrying a RV (Supplementary Figure 5). Four of these eight
9 families were Syrian, and the adjustment for unknown relationships performed as before increased the p-
10 value to 0.0015. Sharing by two or more subjects was observed in five families, but sharing by all affected
11 relatives occurred in only two families, so the complete sharing p-value was much higher. *IL17RE* (OMIM:
12 614995) appears to play a role in host mucosal defense against infection. Some evidence of association
13 between common SNVs in *SPATA21* and cleft lips with or without cleft palate ($p = 2.9 \times 10^{-4}$) was
14 reported in a multiethnic genome-wide association study [Leslie et al. 2016]. No evidence of association
15 was reported for *EP400P1* and *IL17RE* in any study included in the Facebase database.

16 **False signals due to sharing IBS without IBD in the WES sample:** In the WES sample, we did
17 not phase rare SNVs with common variants since we only had the exome sequence, and relied solely on
18 filtering by variant frequency in the 1000 Genomes, Exome Sequencing Project and gnomAD (exome data)

1 databases. Where there were multiple RVs in the same gene in a family, we retained the minimal RV
2 sharing probability, effectively merging RVs with the lowest sharing probability. Excess RV sharing was
3 detected in the gene *ADAMTS9* using the complete sharing test. A supposedly rare SNV was present in
4 at least one affected individual from 26 families: the variant was shared by the two affected subjects in
5 seven families while the variant was seen in only one out of two affected subjects in 18 families and in only
6 one out of three affected subjects in one family ($p = 4.0 \times 10^{-6}$). Upon inspection of all SNVs from the
7 genotyping array and Merlin-inferred haplotypes in *ADAMTS9*, there was evidence against SNVs being
8 IBD between the affected relatives in five of the seven families where the variant was shared IBS (due
9 to homozygous genotypes for opposite alleles in two affected relatives), including the three families where
10 Bureau et al. [2014b] reported the G allele at rs149253049 was shared by two affected relatives.

11 **Rare SNVs in WES sample within genes with a signal in the WGS sample** We examined rare
12 SNVs found in the WES sample within genes reported in Table 4. There were two rare SNVs in *EP400NL*,
13 each seen in a single individual. There were two SNVs in *SPATA21* shared by two affected relatives (in
14 two separate families), one of these two inferred IBD and the other not. Finally, a SNV in *IL17RE* shared
15 IBS by two out of three relatives in one family could be inferred not to be IBD.

16 Discussion

17 This article introduced a number of improvements to the RVS approach for linking RVs to disease intro-
18 duced by Bureau et al. [2014b]. The approach can now be applied to a gene-based analysis, the commonly
19 used strategy to jointly test all RVs in a gene and increase the number of variant carriers [Li and Leal
20 2008, Lee et al. 2014]. This requires a strategy to work with multiple RVs occurring in the same family.
21 When genomic sequence is available, inferring haplotypes of common and previously known RVs within
22 any one gene enables identification and merging of all RVs on the same haplotype prior to analysis, and
23 avoids recomputation of the same RV sharing probability for each of these RVs. There may still remain

1 RVs on different haplotypes within the same family. This is a rare occurrence even in genes with multiple
2 RVs such as *CDH13* used in our simulation, and taking the minimal sharing probability among them did
3 not lead to inflated Type I errors in our simulation study.

4 Phenocopies, diagnosis errors and intra-familial genetic heterogeneity clearly exist, so causal RVs may
5 not be shared by all affected relatives in any one pedigree. One way to detect excess sharing among
6 sequenced family members is then to examine partial sharing, i.e. sharing by a subset of affected subjects
7 in a family, and summing the probability of sharing patterns as or more extreme as the observed one within
8 and between families when computing the p-value. The test allowing partial sharing had a slight power
9 advantage over the complete sharing test for a gene with few RVs in our simulation based on the oral cleft
10 families, but a slight power disadvantage for a gene with many RVs, despite simulating data under genetic
11 models with phenocopies. This difference can be explained by an increased chance of partial sharing under
12 the null hypothesis when there are many RVs, while sharing by all affected subjects remains rare. The
13 exponential growth in computational complexity of the partial sharing test p-value calculation with the
14 number of affected subjects in families with a RV restricts the size and number of families included in the
15 analysis. Nonetheless, the partial sharing test may detect excess RV sharing patterns not detected by the
16 complete sharing test, as exemplified by the sharing pattern observed in the *IL17RE* gene in our WGS
17 sample (although cryptic relatedness in the population increased the p-value to a non significant level, so
18 the observed sharing patterns may well result from RVs not involved in susceptibility to oral cleft).

19 Another way to detect excess sharing without requiring all affected subjects to share a RV is the RareIBD
20 approach based on the number of affected subjects carrying a RV (and unaffected subjects not carrying the
21 RV when such subjects are available). The RareIBD and RVS tests coincide in the special case where all
22 pedigrees have only one affected relative pair of the same type (e.g. an affected first cousin pair), founders
23 are not sequenced, and there is at most one RV per family in a given gene. The RareIBD score can then
24 only take two values (2 or 1) in all families where a RV is seen, the same standardization of that score is
25 applied in all these families [Equation 4 of Sul et al. 2016], and all families have the same weight in the

1 weighted summation over families [Equation 5 of Sul et al. 2016]. The RareIBD Z-score is thus a linear
2 function of the number of families where the score is 2, which under the null hypothesis follows a binomial
3 distribution with success probability being the probability of sharing a RV by both affected relatives given
4 at least one is a carrier (e.g. 1/15 for affected first cousin pairs). In this setting both RVS tests simplify to
5 the exact tail probability of the same null binomial distribution (and while our method provides a closed-
6 form solution, the gene-dropping simulation used in RareIBD is sampling from the binomial distribution to
7 compute the p-value). When all pedigrees have the same configuration of three or more affected subjects,
8 there will be slight differences between RareIBD and each of the RVS tests due to differences in the
9 ordering of the combinations of complete and partial sharing events across families, and hence of their null
10 distributions. Nonetheless, the most extreme event where all affected subjects in all families share a RV has
11 the same p-value, because the p-value is then again the probability of that most extreme event, whether it
12 is computed exactly or by gene dropping. When different configurations of affected subjects are found in
13 different pedigrees, as in our simulated pedigree sample, standardization and weighting of the pedigrees in
14 Rare IBD introduces further differences with the RVS tests. The correlation between the $-\log_{10}p$ of the
15 RareIBD and each RVS test remains high, as evidenced in Supplementary Figure 3, but this figure also
16 reveals datasets simulated with causal RVs where the RV sharing tests rejected the null hypothesis at a
17 given significance threshold while RareIBD failed to reject it, and vice-versa. Computational complexity
18 of evaluating the RareIBD statistic is linear in the number of subjects once the expectation and standard
19 deviation of the number of subjects sharing a RV have been precomputed, but must be repeated on a large
20 number of replicates of gene-dropping under the null hypothesis to obtain a p-value. As a result, for a
21 dataset of typical size like our simulated dataset of 33 extended families, the time to compute the exact
22 p-value of the partial sharing test can be substantially shorter than the time to compute the RareIBD
23 p-value by gene-dropping simulations (Table 3).

24 The pVAAST approach combines the case-control and pedigree-based information in a single statistic
25 (CLRTp), which achieved greater power than all other methods assessed (although the pVAAST linkage

1 LOD score alone had low power). However, the CLRT_p statistic is highly sensitive to inflation due to
2 variant frequency underestimation, as it creates a frequency difference between the cases and controls
3 captured by the CLRT part of this statistic. The LOD score and, in a different way, the GESE statistic
4 also involve the variant frequency and are inflated to a lesser extent by its underestimation. When families
5 in a sample come from various populations, as was the case with the oral cleft families, methods that do
6 not need variant frequencies can be validly applied to the whole sample, provided the analyzed variants
7 are rare or absent in all populations of origin of the families, while methods requiring variant frequency
8 estimates need to be applied separately in each population or use variant frequencies that are incorrect
9 for most families if applied to all families together. Methods requiring variant frequency estimates also
10 have the disadvantage that all genes with RVs seen in the control sample or reference database must be
11 considered in the multiple testing correction, instead of only those genes with RVs seen in affected subjects.
12 This reduces power, as exemplified in Figure 2 when a significance level 4 times lower was used following
13 the suggestion of Qiao et al. [2017].

14 All variant sharing methods we evaluated show good power to detect genes containing causal RVs with
15 high relative risks (20 or higher) but much lower power when the relative risk is smaller at the sample
16 size used in our simulation study. It is plausible that very rare variants with high relative risks remain to
17 be discovered for complex traits because family-based and case-control association studies cannot detect
18 variants carried by a very small number of affected subjects. Even if most undiscovered variants have
19 relative risks at which power of variant sharing methods is fairly low, if there are many such variants the
20 power to detect at least one of them can be good.

21 The second purpose of inferring haplotypes in the vicinity of a putative rare variant is to confirm the IBD
22 status of the variants, and hence their rarity in the population. While no estimate of variant frequency
23 is needed to compute RV sharing probabilities, the inference is nonetheless sensitive to the actual allele
24 frequency in the population [Bureau et al. 2014b]. Variants rare in public databases of diverse human
25 populations (such as the 1000 Genomes or gnomAD) may be not so rare in specific populations. For

1 instance, observing the G allele at rs149253049 on six distinct haplotypes in as many affected subjects
2 from three multiplex families suggested this allele is fairly common in the Bengali population from which
3 these families were sampled. Placing RVs on haplotypes and keeping in the analysis only those found on
4 a single haplotype within a pedigree, as we proposed here and applied to the WGS sample, increases our
5 confidence that these RVs are indeed shared IBD. This is most feasible with whole genome sequence, where
6 all variants are detected with the same technology. Phasing RVs from exome sequence data with SNVs
7 from genotyping arrays would pose a challenge, and we did not implement such a procedure.

8 Haplotypes of known variants do not perfectly distinguish all haplotypes, and may lead to merging RVs
9 actually on distinct haplotypes. Our assessment by simulation revealed a small 0.3% probability of such
10 events occurring. Also, there may be recombination events occurring within a gene in recent generations
11 resulting in IBD copies of a RV on distinct haplotypes spanning a single gene. Future development will
12 include an improvement of our approach to identify recent recombination events, and define haplotypes
13 over the intervals between recombination events instead of only relying on gene boundaries.

14 Software user-friendliness is an important factor driving the adoption of methods in statistical genetics. Our
15 RV sharing tests and GESE are available as R packages, and interface with data classes for pedigrees and
16 genotype data defined in the R statistical environment (cran.r-project.org). Users familiar with this
17 environment already know how to use functions from other R packages to read data in the most common
18 format such as variant call format (VCF) and ped, and convert them to these data classes. RareIBD
19 and pVAAST are available as stand-alone software packages, and require data input files in their own
20 format. External scripting or calls to provided formatting scripts are needed to reformat data files, which
21 may represent important hurdles for some users. The handling of missing genotypes also distinguishes
22 the evaluated RV analysis packages: our methods and GESE focus on RVs seen in the sequence with
23 missing genotypes treated as the reference allele homozygous genotype, while pVAAST sums over all
24 possible genotypes and RareIBD requires complete genotype data, so missing genotypes must be imputed
25 externally prior to calling the program.

1 The example of application of RVS to a familial study of oral cleft was limited to the exome. RVS can also
2 be applied to whole genome sequence when it is available, as in one of the two familial samples presented
3 here, to test rare variants in the non-coding part of the genome, where a large proportion of GWAS signals
4 are located. Whole genome analysis presents challenges such as defining meaningful analysis windows
5 outside of genes and selecting variants more likely to be functional based on biological annotations which
6 will be addressed in future work.

7 In conclusion, the additional features of a gene-based analysis of RV sharing with confirmation of IBD
8 status of variants by inferring haplotypes from genomic sequence data does enhance the ability of our
9 RV sharing approach to detect causal RVs in studies of extended multiplex families, while maintaining
10 its original advantage of not requiring variant frequency estimates from a population sample. Testing
11 for partial sharing patterns offers an additional option to detect RVs involved in complex diseases where
12 phenocopy, mis-diagnosis or intra-familial genetic heterogeneity exist, but this approach did not always
13 provide a power gain over computationally cheaper alternatives.

14 **Acknowledgements**

15 We thank the members of the families who participated in the oral cleft sequencing studies, and the
16 field and laboratory staff who made this analysis possible. We thank Thomas Sherman (Johns Hopkins
17 Bloomberg School of Public Health) for programming the RVS package and Saeed Sabbah (Université Laval)
18 for performing the haplotyping of SNVs in the WES sample. The statistical analysis work was partly funded
19 by the Canadian Statistical Sciences Institute through a Collaborative Research Team grant of which AB
20 is the leader. National Institute of Dental and Craniofacial Research grant R03-DE-02579 to IR. The oral
21 cleft sequencing studies were supported by the National Institutes of Health (NIH) (R01-DE-014581, U01-
22 DE-018993, and U01 DE020073 to THB.; R01-DE-016148 and R01-DE-009886 to MLM.; P50-DE-016215
23 and R37-DE-08559 to JCM). Recruitment of Syrian families was supported by the Intramural Research

1 Program of the National Human Genome Research Institute, National Institutes of Health, USA, and the
2 Ibn Al-Nafees Hospital, Syrian Arab Republic. Additional support from X01HG006177 to THB, MLM, and
3 JCM for whole exome sequencing at the Center for Inherited Disease Research, which is funded through
4 a federal contract from the NIH to Johns Hopkins University (contract no. HHSN268200782096C). The
5 FaceBase database was funded by NIH grant U01-DE024425 to MLM.

1 **References**

- 2 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel
3 JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. 2015. A global reference for human genetic
4 variation. *Nature* 526:68–74.
- 5 Aida H, Takakuwa K, Nagata H, Tsuneki I, Takano M, Tsuji S, Takahashi T, Sonoda T, Hatae M, Takahashi
6 K, Hasegawa K, Mizunuma H, Toyoda N, Kamata H, Torii Y, Saito N, Tanaka K, Yakushiji M, Araki
7 T, Tanaka K. 1998. Clinical features of ovarian cancer in japanese women with germ-line mutations of
8 *brca1*. *Clinical cancer research : an official journal of the American Association for Cancer Research*
9 4:235–240.
- 10 Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. 2011. Exome
11 sequencing as a tool for mendelian disease gene discovery. *Nature reviews Genetics* 12:745–755.
- 12 Bureau A, Parker MM, Ruczinski I, Taub MA, Marazita ML, Murray JC, Mangold E, Noethen MM, Ludwig
13 KU, Hetmanski JB, Bailey-Wilson JE, Cropp CD, Li Q, Szymczak S, Albacha-Hejazi H, Alqosayer K,
14 Field LL, Wu-Chou YH, Doheny KF, Ling H, Scott AF, Beaty TH. 2014a. Whole exome sequencing of
15 distant relatives in multiplex families implicates rare variants in candidate genes for oral clefts. *Genetics*
16 197:1039–1044.
- 17 Bureau A, Younkin SG, Parker MM, Bailey-Wilson JE, Marazita ML, Murray JC, Mangold E, Albacha-
18 Hejazi H, Beaty TH, Ruczinski I. 2014b. Inferring rare disease risk variants based on exact probabilities
19 of sharing by multiple affected relatives. *Bioinformatics (Oxford, England)* 30:2189–2196.
- 20 Dahl M, Nordestgaard BG, Lange P, Vestbo J, Tybjaerg-Hansen A. 2001. Molecular diagnosis of inter-
21 mediate and severe alpha(1)-antitrypsin deficiency: Mz individuals with chronic obstructive pulmonary
22 disease may have lower lung function than mm individuals. *Clinical chemistry* 47:56–62.
- 23 Daoud H, Valdmanis PN, Kabashi E, Dion P, Dupr N, Camu W, Meininger V, Rouleau GA. 2009. Con-

- 1 tribution of tardbp mutations to sporadic amyotrophic lateral sclerosis. *Journal of medical genetics*
2 46:112–114.
- 3 Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure
4 J, Nickerson DA, Bamshad MJ, Project NES, Akey JM. 2013. Analysis of 6,515 exomes reveals the recent
5 origin of most human protein-coding variants. *Nature* 493:216–220.
- 6 Holzinger ER, Li Q, Parker MM, Hetmanski JB, Marazita ML, Mangold E, Ludwig KU, Taub MA, Begum
7 F, Murray JC, Albacha-Hejazi H, Alqosayer K, Al-Souki G, Albasha Hejazi A, Scott AF, Beaty TH,
8 Bailey-Wilson JE. 2017. Analysis of sequence data to identify potential risk variants for oral clefts in
9 multiplex families. *Molecular Genetics and Genomic Medicine* 5:570–579.
- 10 Hu H, Roach JC, Coon H, Guthery SL, Voelkerding KV, Margraf RL, Durtschi JD, Tavtigian SV,
11 Shankaracharya, Wu W, Scheet P, Wang S, Xing J, Glusman G, Hubley R, Li H, Garg V, Moore
12 B, Hood L, Galas DJ, Srivastava D, Reese MG, Jorde LB, Yandell M, Huff CD. 2014. A unified test of
13 linkage analysis and rare-variant association for analysis of pedigree sequence data. *Nature biotechnology*
14 32:663–669.
- 15 Ionita-Laza I, Makarov V, Yoon S, Raby B, Buxbaum J, Nicolae DL, Lin X. 2011. Finding disease variants
16 in mendelian disorders by using sequence data: methods and applications. *American journal of human*
17 *genetics* 89:701–712.
- 18 Isobe N, Damotte V, Re VL, Ban M, Pappas D, Guillot-Noel L, Rebeix I, Compston A, Mack T, Cozen
19 W, Fontaine B, Hauser SL, Oksenberg JR, Sawcer S, Gourraud PA. 2013. Genetic burden in multiple
20 sclerosis families. *Genes and immunity* 14:434–440.
- 21 Johnston JJ, Rubinstein WS, Facio FM, Ng D, Singh LN, Teer JK, Mullikin JC, Biesecker LG. 2012.
22 Secondary variants in individuals undergoing exome sequencing: screening of 572 individuals identifies
23 high-penetrance mutations in cancer-susceptibility genes. *American journal of human genetics* 91:97–108.

- 1 Lee S, Abecasis G, Boehnke M, Lin X. 2014. Rare-variant association analysis: Study designs and statistical
2 tests. *The American Journal of Human Genetics* 95:5 – 23.
- 3 Leslie EJ, Carlson JC, Shaffer JR, Feingold E, Wehby G, Laurie CA, Jain D, Laurie CC, Doheny KF,
4 McHenry T, Resick J, Sanchez C, Jacobs J, Emanuele B, Vieira AR, Neiswanger K, Lidral AC, Valencia-
5 Ramirez LC, Lopez-Palacio AM, Valencia DR, Arcos-Burgos M, Czeizel AE, Field LL, Padilla CD,
6 Cutiongco-de la Paz EM, Deleyiannis F, Christensen K, Munger RG, Lie RT, Wilcox A, Romitti PA,
7 Castilla EE, Mereb JC, Poletta FA, Orioli IM, Carvalho FM, Hecht JT, Blanton SH, Buxo CJ, Butali A,
8 Mossey PA, Adeyemo WL, James O, Braimah RO, Aregbesola BS, Eshete MA, Abate F, Koruyucu M,
9 Seymen F, Ma L, de Salamanca JE, Weinberg SM, Moreno L, Murray JC, Marazita ML. 2016. A multi-
10 ethnic genome-wide association study identifies novel loci for non-syndromic cleft lip with or without
11 cleft palate on 2p24.2, 17q23 and 19q13. *Hum Mol Genet* 25:2862–2872.
- 12 Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application
13 to analysis of sequence data. *Am J Hum Genet* 83:311–21.
- 14 Mescheriakova JY, Broer L, Wahedi S, Uitterlinden AG, van Duijn CM, Hintzen RQ. 2016. Burden
15 of genetic risk variants in multiple sclerosis families in the netherlands. *Multiple sclerosis journal -*
16 *experimental, translational and clinical* 2:2055217316648721.
- 17 Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett
18 LM, Ding W. 1994. A strong candidate for the breast and ovarian cancer susceptibility gene *brca1*.
19 *Science (New York, NY)* 266:66–71.
- 20 Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW,
21 Nickerson DA, Shendure J, Bamshad MJ. 2010. Exome sequencing identifies the cause of a mendelian
22 disorder. *Nature genetics* 42:30–35.
- 23 O’Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, Traglia M, Huang J, Huffman JE,
24 Rudan I, McQuillan R, Fraser RM, Campbell H, Polasek O, Asiki G, Ekoru K, Hayward C, Wright AF,

- 1 Vitart V, Navarro P, Zagury JF, Wilson JF, Toniolo D, Gasparini P, Soranzo N, Sandhu MS, Marchini
2 J. 2014. A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS genetics*
3 10:e1004234.
- 4 Papassotiropoulos A, Fountoulakis M, Dunckley T, Stephan DA, Reiman EM. 2006. Genetics, transcrip-
5 tomics, and proteomics of alzheimer's disease. *The Journal of clinical psychiatry* 67:652–670.
- 6 Qiao D, Lange C, Laird NM, Won S, Hersh CP, Morrow J, Hobbs BD, Lutz SM, Ruczinski I, Beaty
7 TH, Silverman EK, Cho MH. 2017. Gene-based segregation method for identifying rare variants in
8 family-based sequencing studies. *Genetic epidemiology* 41:309–319.
- 9 Risch N. 1990. Linkage strategies for genetically complex traits. i. multilocus models. *American journal of*
10 *human genetics* 46:222–228.
- 11 Sherman T, Fu J, Scharpf RB, Bureau A, Ruczinski I. 2018. Detection of rare disease variants in extended
12 pedigrees using RVS. (in revision) .
- 13 Stratton MR. 1996. Recent advances in understanding of genetic susceptibility to breast cancer. *Human*
14 *molecular genetics* 5 Spec No:1515–1519.
- 15 Sul JH, Cade BE, Cho MH, Qiao D, Silverman EK, Redline S, Sunyaev S. 2016. Increasing generality and
16 power of rare-variant tests by utilizing extended pedigrees. *The American Journal of Human Genetics*
17 99:846–859.
- 18 Szabo CI, King MC. 1995. Inherited breast and ovarian cancer. *Human molecular genetics* 4 Spec No:1811–
19 1817.