Sharing of rare variants by affected relatives: building evidence for causal variants based on exact sharing probabilities

Alexandre Bureau[1,2], Margaret M. Parker[3], Samuel Youngkin[3], Terri H. Beaty[3], Ingo Ruczinski[3]

Received _____; accepted _____

[1]Centre de recherche de l'Institute universitaire en santé mentale de Québec

[2]Département de médecine sociale et préventive, Université Laval

[3]Johns Hopkins Bloomberg School of Public Health

ABSTRACT

Family based study designs are regaining popularity because large scale sequencing can help to interrogate the relationship between disease and variants too rare in the population to be detected through tests of association in a conventional case-control study, but may nonetheless co-segregate with disease within families. Where only a few affected subjects per family are sequenced, evidence that a rare variant may be causal can be quantified from the probability any variant would be shared by all affected relatives given it was seen in any one family member under the null hypothesis of complete absence of linkage and association. For variants seen in M families and shared by affected relatives in m of them, a p-value can be obtained as the sum of the probabilities of sharing events as (or more) extreme. We generalized the expression for the sharing probability to more than two subjects per family. We also examined the impact of unknown relationships and proposed approximation of sharing probability based on empirical estimates of kinship between family members obtained from genome-wide marker data. A simulation study demonstrate the accuracy of the approximation for low levels of kinship between founders. We applied this method to a study of 55 multiplex families with apparent non-syndromic forms of oral clefts from four distinct populations. Whole exome sequencing was performed by the Center for Inherited Disease Research (CIDR) on two or three affected members per family. The rare single nucleotide variant (SNV) rs149253049 in the gene ADAMTS9 was shared by affected relatives in three Indian families ($p = 2 \times 10^{-6}$), illustrating the power of this sharing approach. Another SNV was shared in three out of four families, among which two families from the Syrian sample where excess sharing was detected. In that case the evidence against the null hypothesis was reduced after applying a correction for unknown relationships.

1. Introduction

The advent of high-throughput sequencing of whole exomes and even whole genomes opens the possibility of detecting rare variants (RVs, from unique to a family to one percent frequency in a population) impacting human health. The first successful applications of exome sequencing have been with rare Mendelian traits (Gilissen et al. 2012). A widespread study design to discover such rare highly penetrant variants in families where previous genotyping has not been performed is to sequence the exome (or increasingly the whole genome) of two or three affected subjects, what Gilissen et al. call the "linkage" approach (Gilissen et al. 2012). The identification of the likely causal variant is in fact conducted by focusing on novel variants predicted to be functional that are shared by all sequenced family members.

RVs may also explain a part of the so-called "missing heritability" of complex diseases and even be responsible for association signals detected with common variants (Cirulli and Goldstein 2010). In a family with a high concentration of disease cases, there is a high probability that multiple affected members carry the same rare disease predisposing variant if such a variant exists and its penetrance is high (Cirulli and Goldstein 2010; Wijsman 2012). This confers an advantage to family samples over samples of unrelated individuals where disease causing RVs may be so seen only once or twice in tens of thousands of subjects.

Contrary to monogenic Mendelian traits, considerable genetic heterogeneity is expected with complex diseases. Oral clefts are common craniofacial malformations representing a good example of a genetically heterogeneous disorder with at least a dozen different genes identified as genetic risk factors in genome-wide association studies (GWAS), a few of which may be directly causal (Beaty et al. 2013; Ludwig et al. 2012). Sequencing studies need to include larger numbers of families to provide enough power to identify RVs.

Results reported here are from a whole exome study of 55 multiplex families with apparent non-syndromic forms of oral clefts.

As with Mendelian disorders, it has initially been proposed to use the RV sharing information to filter out RVs that are not shared in at least one family. (Feng et al. 2011). For variants sufficiently rare that the copies in the sequenced relatives are almost certainly identical by descent (IBD), the probability that a RV independent of the disease and detected in at least one sequenced subject would not be shared by other sequenced affected relatives was computed by Feng et al. (Feng et al. 2011) to quantify the effectiveness of what they call the concordance filter to discard irrelevant RVs. We adopt the view that the probability a rare variant would be shared by all affected relatives in a family given it was seen in any one of them, computed under the null hypothesis of absence of linkage and association to the disease, can be used to quantify the evidence against the null hypothesis and therefore establish that a RV may be predisposing to the disease. It is important to stress that more evidence is extracted from each family than only testing for linkage: the null probability of sharing a RV by two first cousins is $\frac{1}{15}$ while their null IBD sharing probability is $\frac{1}{8}$. The evidence can be combined across all the families where the RV is seen, if there are more than one.

RV sharing probabilities between relatives rest on the assumptions that the variant is rare enough to have been introduced only once in the family and that the known family structure is correct, in particular that family founders are unrelated. Cryptic relatedness can often be detected from dense genotype data. When founders of a pedigree are related, a RV may be introduced more than once in a family, leading to greater actual sharing probabilities between relatives than the value computed based on the known pedigree

structure, and an overstatement of the evidence against the null.

In this article, we generalize to more than two subjects per family the expression for the probability that a variant is shared by all relatives sequenced in a family given that it was seen in any of them. We also examine the impact of unknown relationships and propose two methods to approximate the sharing probability using kinship coefficients between founders, either based on genealogical knowledge or empirical estimates obtained from genome-wide marker data on family members. The validity of the approximations is evaluated in a simulation of small populations. We applied the sharing probability computation to a study of multiplex families with apparent non-syndromic forms of oral clefts from four distinct populations. Whole exome sequencing was performed on two or three affected members per family. The rare single nucleotide variant rs149253049 in gene ADAMTS9 shared in three families from India provided substantial evidence against the null. We also illustrate with another RV shared in three out of four families where it was found that cryptic relatedness may importantly lower the evidence.

2. Material and Methods

Our goal is to compute the probability that a set of related subjects whose DNA sequence is observed through sequencing or other means (sequenced subjects) share a rare variant (RV) identical-by-descent given that a RV has been observed at a site in the sequence, under the null hypothesis of no linkage and no association to the disease of the sequenced subjects. We assume that the variant for which we compute a sharing probability is rare enough that there exists a single copy of that variant among the alleles present in the n_f founders of the pedigree relating the subject for which we want to compute a sharing probability. In the basic setting, all founders are unrelated and a

single copy of the variant is present among the founders. In a generalization, we allow founders to be related, and two copies of the allele to be introduced in the pedigree by a pair of related founders. We finally demonstrate how RV sharing probabilities computed in a single family can be combined across multiple families where the same variant is detected.

2.1. Computation assuming all founders are unrelated

We define the following random variables and constants:

C_i Number of copies of the RV received by sequenced subject i

F_j Indicator variable that founder j introduces one copy of the RV in the pedigree

B_k Number of copies of the RV in subject k where a line of descent from a founder branches into two separate lines of descents to a subset of sequenced subjects

D_{ij} Number of generations (meioses) between subject i and his ancestor j

For a set of n sequenced subjects for which the pedigree structure limits to one the number of copies of the rare variant that they can share, we want to compute the probability

$$\begin{aligned} P[\text{RV shared}] &= P[C_1 = \ldots = C_n = 1 | C_1 + \ldots + C_n \geq 1] = \frac{P[C_1 = \ldots = C_n = 1]}{P[C_1 + \ldots + C_n \geq 1]} \quad (1) \\ &= \frac{\sum_{j=1}^{n_f} P[C_1 = \ldots = C_n = 1 | F_j] P[F_j]}{\sum_{j=1}^{n_f} P[C_1 + \ldots + C_n \geq 1 | F_j] P[F_j]} \end{aligned}$$

where the expression on the second line results from our assumption that there exists a single copy of that variant among the alleles present in the n_f founders. The probabilities

$P[F_j] = \frac{1}{n_f}$ cancel from the numerator and denominator. For the other terms, we first derive expressions for the special case where all the sequenced subjects descend from every founder among their ancestors through independent lines of descent. In that case,

$$P[C_1 = \ldots = C_n = 1|F_j] = \begin{cases} \prod_i \left(\frac{1}{2}\right)^{D_{ij}} = \left(\frac{1}{2}\right)^{D_j} & \text{if } F_j \text{ is a common ancestor to the } n \text{ sequenced subjects} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

and

$$P[C_1 + \ldots + C_n \geq 1|F_j] = 1 - P[C_1 = \ldots = C_n = 0|F_j] = 1 - \prod_{i \in d(j)} \left(1 - \left(\frac{1}{2}\right)^{D_{ij}}\right) \tag{3}$$

where $D_j = \sum_i D_{ij}$ and $d(j)$ is the subset of sequenced individuals who descend from founder j.

The global expression is then

$$P[\text{RV shared}] = \frac{\sum_{j=1}^{n_f} \left(\frac{1}{2}\right)^{D_j} I(F_j \text{ is a common ancestor to the } n \text{ sequenced subjects})}{\sum_{j=1}^{n_f} \left[1 - \prod_{i \in d(j)} \left(1 - \left(\frac{1}{2}\right)^{D_{ij}}\right)\right]} \tag{4}$$

We note here that equation 4 covers the general case of pedigrees without inbreeding, including individuals marrying multiple times and marriage loops as in the family depicted in Figure 3B for instance. It is a generalization of the sharing probability for two subjects, $P[\text{RV shared}] = \frac{1}{2^{(D+1)1}}$ where D is the degree of relationship between the two subjects, given by Feng et al. (Feng et al. 2011) For the common special case of a pedigree with a founder couple ancestral to all descendants in the pedigree, the numerator simplifies and we obtain the following expression:

$$P[\text{RV shared}] = \frac{\left(\frac{1}{2}\right)^{D_f - 1}}{\sum_{j=1}^{n_f} \left[1 - \prod_{i \in d(j)} \left(1 - \left(\frac{1}{2}\right)^{D_{ij}}\right)\right]} \tag{5}$$

where f is any of the two founders forming the ancestral couple.

When the lineages of sequenced individuals "coalesce" at a branching individual k who descends from founders of the pedigree, we can no longer write a general expression like 4, and recursive computations are required. Without loss of generality, let k be the branching individual who has sequenced subjects $1, \ldots, i_k$ as descendants through independent lines of descent. We have

$$P[C_1 = \ldots = C_n = 1] = P[C_1 = \ldots = C_{i_k} = 1 | B_k = 1] P[B_k = C_{i_k+1} = \ldots = C_n = 1] \quad (6)$$

because $P[C_1 = \ldots = C_{i_k} = 1 | B_k = 0] = 0$. The term $P[C_1 = \ldots = C_{i_k} = 1 | B_k = 1]$ is computed from equation 2 replacing F_j by B_k. The term $P[B_k = C_{i_k+1} = \ldots = C_n = 1]$ is computed by reapplying equation 6 recursively with every branching individual.

Also, for a founder above a branching individual in the pedigree, we have

$$\begin{aligned} P[C_1 = \ldots = C_n = 0 | F_j] &= P[C_1 = \ldots = C_{i_k} = 0 | B_k = 1, F_j] P[B_k = 1, C_{i_k+1} = \ldots = C_n = 0 | F_j] \\ &\quad + P[C_1 = \ldots = C_{i_k} = 0 | B_k = 0, F_j] P[B_k = C_{i_k+1} = \ldots = C_n = 0 | F_j] \\ &= P[C_1 = \ldots = C_{i_k} = 0 | B_k = 1] P[B_k = 1, C_{i_k+1} = \ldots = C_n = 0 | F_j] \quad (7) \\ &\quad + P[B_k = C_{i_k+1} = \ldots = C_n = 0 | F_j] \end{aligned}$$

The term $P[C_1 = \ldots = C_{i_k} = 0 | B_k = 1]$ is computed from the right-hand side of equation 3 replacing F_j by B_k. The two terms $P[B_k = a, C_{i_k+1} = \ldots = C_n = 0 | F_j], a = 0, 1$ require recursive computations. If h is a branching individual who is an ancestor of k and a descendant of founder j, then

$$\begin{aligned} & P[B_k = 1, C_{i_k+1} = \ldots = C_n = 0 | F_j] \\ = \; & P[B_k = 1, C_{i_k+1} = \ldots = C_{i_h} = 0 | B_h = 1] P[B_h = 1, C_{i_h+1} = \ldots = C_n = 0 | F_j] \end{aligned}$$

$$+P[B_k = 1, C_{i_k+1} = \ldots = C_{i_h} = 0|B_h = 0]P[B_h = C_{i_h+1} = \ldots = C_n = 0|F_j]$$
$$= \left(\tfrac{1}{2}\right)^{D_{kh}} P[C_{i_k+1} = \ldots = C_{i_h} = 0|B_h = 1]P[B_h = 1, C_{i_h+1} = \ldots = C_n = 0|F_j] \quad (8)$$

and similarly to 8

$$P[B_k = C_{i_k+1} = \ldots = C_n = 0|F_j]$$
$$= \left(1 - \left(\tfrac{1}{2}\right)^{D_{kh}}\right) P[C_{i_k+1} = \ldots = C_{i_h} = 0|B_h = 1]P[B_h = 1, C_{i_h+1} = \ldots = C_n = 0|F_j]$$
$$+P[B_h = C_{i_h+1} = \ldots = C_n = 0|F_j] \quad (9)$$

where the computation of the term $P[C_{i_k+1} = \ldots = C_{i_h} = 0|B_h = 1]$ can itself involve other branching individuals who are descendants of h.

2.2. Computation allowing for relatedness between founders or inbreeding loops within a pedigree

We generalize our computation to the setting where founders are related, while still excluding that the founders are themselves inbred (only their children will be). This includes the setting where inbreeding loops are known and are included in the pedigree structure. One can then define a noninbred subpedigree by removing some familial links. The relatedness between the "founders" of that subpedigree can be captured by their kinship coefficient based on the removed links, and the first approximation described below can then be applied. When familial links between founders are unknown, they sometime can be estimated from genotype data on these founders. Other times, genotype data is only available on the sequenced subjects.

We propose two methods to approximate sharing probabilities between sequenced subjects in presence of IBD sharing in excess of what is expected based on the pedigree

structure. With the first method, only one founder allele (not necessarily the RV considered in the computation) can be shared by only one pair of founders. This method gives an exact sharing probability when only two founders are related, and a good approximation when a few founders are related. Known founder pair-specific kinship coefficients can be used. With the second method, up to T alleles can be shared by two pairs of founders, with T set to 2 or 3 depending on the level of sharing between founders. It requires to assume that all founders are related to the same extent, i.e. all pairs of founders have the same kinship coefficient calculated to explain the excess sharing between sequenced subjects. The method gives a good approximation for more extensive hidden relatedness than the first method. Note that in this second approximation, we still assume that only two founders introduce a copy of the RV considered in the computation.

The elements that we need to implement either approach are:

1. The probability that a pair of related founders introduce the RV in the pedigree.
2. The sharing probabilities conditional on the introduction of the RV by two of the founders.

The two methods to approximate the probability that a pair of related founders introduce the RV in the pedigree are described below. The formulas for the sharing probabilities conditional on the introduction of the RV by two of the founders are given in Appendix A. As an alternative to computing these conditional sharing probabilities, they can be estimated using Monte Carlo simulation of the transmission of the RV down the pedigree from the founder introducing it.

Once the required elements have been computed, we get an adjusted estimate of sharing probability with the following formula:

$$P[\text{RV shared}] = \frac{\sum_{j=1}^{n_f} P[C_1 = \ldots = C_n = 1|F_j^U]P[F_j^U] + \sum_j \sum_{k>j} P[C_1 = \ldots = C_n = 1|F_j, F_k]P[F_j, F_k]}{\sum_{j=1}^{n_f} P[C_1 + \ldots + C_n \geq 1|F_j^U]P[F_j^U] + \sum_j \sum_{k>j} P[C_1 + \ldots + C_n \geq 1|F_j, F_k]P[F_j, F_k]} \quad (10)$$

where F_j^U is the event that founder j is the only one to introduce the RV in the family.

2.2.1. Method 1

The probability that two related founders, say j and k, introduce the RV in the pedigree is expressed as follows:

$$\begin{aligned} P[F_j, F_k] &= P[\text{Allele shared is RV}|j\&k \text{ share allele IBD}]P[j\&k \text{ share allele IBD}] \quad (11) \\ &= \frac{1}{2n_f - 1} 2\phi_{jk} = \frac{2\phi_{jk}}{2n_f - 1} \end{aligned}$$

where ϕ_{jk} is the kinship coefficient between founders j and k. The first term represents the probability that the RV is the allele IBD between the two founders among the $2n_f - 1$ distinct alleles in all founders. The marginal probability that any founder h introduces the RV needs to be adjusted compared to the unrelated case. In that computation, we make the simplifying assumption that the probability that 3 or more founders share an allele IBD is 0 so that the event "i and j share an allele IBD" means that they are the only ones to do so. This assumption is true only when a single pair of founders are related. While the formula allows all pairs of founders to be related, we recommend using this approximation when only a few of the ϕ_{jk} are non-zero.

$$\begin{aligned} P[F_h] &= \sum_j \sum_{k>j} P[F_h|j\&k \text{ share allele IBD}]P[j\&k \text{ share allele IBD}] \quad (12) \\ &+P[F_h|\text{no founder pair shares allele IBD}]P[\text{no founder pair shares allele IBD}] \end{aligned}$$

$$= \frac{2}{2n_f - 1} \sum_j \sum_{k>j} P[j\&k \text{ share allele IBD}] + \frac{1}{n_f} \left(1 - \sum_j \sum_{k>j} P[j\&k \text{ share allele IBD}] \right)$$
$$= \frac{4 \sum_j \sum_{k>j} \phi_{jk}}{2n_f - 1} + \frac{1}{n_f} \left(1 - \sum_j \sum_{k>j} 2\phi_{jk} \right)$$

We obtain the probability of F_j^U, the event that founder j is the only one to introduce the RV in the family, as

$$P[F_j^U] = P[F_j] - \sum_{k \neq j} P[F_j, F_k] \tag{13}$$

If we know which founders j and k are related, then their degree of relatedness is usually also known, and specifies their kinship coefficient ϕ_{jk}. If it is possible to identify a subset of founders that are suspected to be related, with the other founders unrelated to that subset and between themselves, then this method can still be applied, with the kinship coefficient between the subset of founders suspected to be related estimated as described in section 2.2.2. If instead familial links between founders are completely unknown, we generally recommend to apply the second method.

2.2.2. Method 2

For the second method, we assume $\phi_{jk} = \phi^f \forall j, k$. This is an assumption that we prefer to make when the relatedness between specific pairs of founders is unknown and we need to rely on genotye data to estimate it. Even with perfect information on IBD sharing between subjects, there is considerable variation in the kinship coefficient based on IBD sharing estimated for pairs of subjects with the same degree of relatedness due to variation in the length of genome shared from pair to pair (Manichaikul et al. 2010), and reliable inference can only be obtained for the mean or other central tendency parameter.

Two situations can occur with respect to the genotype data available to estimate kinship between founders:

1. Polymorphic markers have been genotyped on the pedigree founders, typically a genomewide SNP array. Then ϕ_{jk} can be estimated for each founder pair j and k, and a global estimate $\hat{\phi}^f$ obtained by averaging the $\hat{\phi}_{jk}$ over all founder pairs from the same population.

2. Genotype data is only available on the sequenced subjects (either from the sequencing data itself of from other genotyping). The common ϕ^f is estimated based on the estimated kinship coefficients between sequenced subjects and the relationship between the sequenced subjects and all founders.

$$\begin{aligned}\phi_{i_1 i_2} &= \phi^f \sum_j \sum_{k>j} \left[\left(\frac{1}{2}\right)^{D_{i_1 j}+D_{i_2 k}} I(j\&k \text{ not mating}) + \left(\frac{1}{2}\right)^{D_{i_1 j}+D_{i_2 k}-1} I(j\&k \text{ mating}) \right] + \phi^p_{i_1 i_2} \\ &= \phi^f \kappa_{i_1 i_2} + \phi^p_{i_1 i_2} \end{aligned} \tag{14}$$

An estimate of ϕ^f is then obtained for every pair i_1, i_2 as

$$\hat{\phi}^f_{i_1,i_2} = \frac{(\hat{\phi}_{i_1 i_2} - \phi^p_{i_1 i_2})}{\kappa_{i_1 i_2}} \tag{15}$$

These pair-specific estimates can then be averaged over all pairs of sequenced subjects from the same population to obtain a global $\hat{\phi}^f$.

This second method of approximation relates the estimated mean kinship $\hat{\phi}^f$ to the distribution of the number of alleles distinct by descent in the founders. Then, $P[F_j, F_k]$ and $P[F_j]$ are derived from that distribution. The rest of this sub-subsection explains in

detail how to compute the approximate values of these quantities.

The number of alleles A distinct by descent in the founders can take values $1, \ldots, 2n_f$. We will assume only the values $2n_f - d, \ldots, 2n_f$ have nonzero probability, and that among the a distinct alleles present $2n_f - a$ of them are present twice and the remaining $2(a - n_f)$ are present once.

We parameterize the probabilities $P[A]$ to be proportional to

$$\begin{array}{cccc} 2n_f - d & \ldots & 2n_f - 1 & 2n_f \\ \frac{1}{d!}\theta^d & \ldots & \theta & 1 \end{array} \tag{16}$$

inspired from a truncated Poisson distribution. The expected kinship coefficient among the n_f founders is then

$$E[\Phi] = \frac{\sum_{a=2n_f-d}^{2n_f-1} \frac{1}{(2n_f-a)!}\theta^{(2n_f-a)}\bar{\phi}_a}{\sum_{a=2n_f-d}^{2n_f} \frac{1}{(2n_f-a)!}\theta^{(2n_f-a)}} \tag{17}$$

where $\bar{\phi}_a$ is the mean kinship coefficient among the n_f founders when there are a alleles distinct by descent. Assuming no inbreeding among the founders, we show in Appendix B that:

$$\begin{aligned} \bar{\phi}_a \;=\; & P[\text{Alleles from two founders are IBD} \big| \text{One of the founders shares an allele IBD with 2 other founders}] \\ & P[\text{One of the founders shares an allele IBD with 2 other founders}] \\ & + P[\text{Alleles from two founders are IBD} \big| \text{One of the founders shares an allele IBD with 1 other founders}] \\ & P[\text{One of the founders shares an allele IBD with 1 other founder}] \end{aligned}$$

$$= \frac{1}{2(n_f-1)}\frac{2n_f-a}{n_f}\frac{2n_f-a-1}{n_f-1}+\frac{1}{4(n_f-1)}\left[\frac{(2n_f-a)(a-n_f)}{n_f(n_f-1)}+\frac{2(2n_f-a)(a-n_f)}{n_f(2n_f-1)}\right] \quad (18)$$

Equating $E[\Phi]=\hat{\phi}^f$, we solve the polynome for θ. The value of d required to obtain a good approximation depends on the value of $\hat{\phi}^f$. When less than $2n_f-5$ distinct alleles must be allowed to obtain a real positive root of the polynome in θ, we have found in practice that we obtain a poor approximation, since the probability that any of the distinct alleles (including the RV of interest) is present more than twice become non-negligible. This is why we propose setting $d=5$. When $\hat{\phi}^f$ is small, values the approximation is almost identical with values of $d=4$, 3 or even 2 than with $d=5$. When $d=2$, we have the explicit solution:

$$\hat{\theta}=\frac{-(\hat{\phi}^f-\bar{\phi}_{2n_f-1})-\sqrt{(\hat{\phi}^f-\bar{\phi}_{2n_f-1})^2-2(\hat{\phi}^f-\bar{\phi}_{2n_f-2})\hat{\phi}^f}}{\hat{\phi}^f-\bar{\phi}_{2n_f-2}} \quad (19)$$

What we need finally is the probability P_2 every founder pair introduces the RV and the probability P_U every founder is alone to introduce the RV. Assuming only one or two founders introduce the RV, $n_fP_U+\frac{1}{2}n_f(n_f-1)P_2=1$ and we only need to obtain P_U. We can obtain P_U by developing equation 13 into:

$$P_U=\sum_{a=2n_f-d}^{2n_f}P[A=a]\left(P[F_j|A=a]-\sum_{k\neq j}P[F_j,F_k|A=a]\right)$$

The probability that any founder j introduces the RV under our model assuming his genotype is composed of two distinct alleles drawn from among the a distinct alleles of the founders is $P[F_j|A=a]=P_a=\frac{2}{a}$. We also note that $P[F_j,F_k|A=a]$ depends only on a and we note it R_a.

$$P_U = \sum_{a=2n_f-d}^{2n_f}P[A=a]\left(P_a-\sum_{k\neq j}R_a\right)$$

$$= \sum_{a=2n_f-d}^{2n_f} P[A=a]\left(\frac{2}{a} - (n_f-1)R_a\right)$$

To solve for R_a, we use the result from probability theory that

$$\begin{aligned} 1 &= P[F_1 \cup \ldots \cup F_{n_f}] \\ &= \sum_j^{n_f} P[F_j] - \sum_j^{n_f}\sum_{k\neq j} P[F_j, F_k] \\ &= \sum_a P[A=a]\left(\sum_j^{n_f} P[F_j|A=a] - \sum_j^{n_f}\sum_{k>j} P[F_j, F_k|A=a]\right) \\ &= \sum_a P[A=a]\left(n_f P_a - \frac{1}{2}n_f(n_f-1)R_a\right) \end{aligned} \quad (20)$$

assuming at most two founders can introduce the RV. To find a solution for R_a, we assume that $n_f P_a - \frac{1}{2}n_f(n_f-1)R_a = 1$, which obviously satisfies 20. We obtain

$$R_a = \frac{2(\frac{2n_f}{a} - 1)}{n_f(n_f-1)}$$

and

$$P_U = \sum_a P[A=a]\left(\frac{2}{n_f} - \frac{2}{a}\right) \quad (21)$$

Given the constant $P[F_j]$ implied by method 2, we can simplify equation 10 to:

$$P[\text{RV shared}] = \frac{w\sum_{j=1}^{n_f} P[C_1 = \ldots = C_n = 1|F_j^U] + (1-w)\sum_j\sum_{k>j} P[C_1 = \ldots = C_n = 1|F_j, F_k]}{w\sum_{j=1}^{n_f} P[C_1 + \ldots + C_n \geq 1|F_j^U] + (1-w)\sum_j\sum_{k>j} P[C_1 + \ldots + C_n \geq 1|F_j, F_k]} \quad (22)$$

where $w = n_f P_U$.

When a Monte Carlo simulation is used to compute the sharing probabilities conditional on the introduction of the RV by one or two of the founders, the last two steps are included in the simulation: The number of distinct alleles a is sampled from distribution 16, then the RV is sampled among the a alleles. The RV is introduced twice if it is one of the first $2n_f - a$ alleles and introduced once otherwise. If it is introduced twice, the pair of founders introducing it is sampled with equal probability for all pairs. If it is introduced once, the founder introducing it is sampled instead.

2.3. Combining RV sharing probabilities across multiple families

For variants seen in only one family, the RV sharing probability can be interpreted directly as a p-value from a Bernoulli trial. For variants seen in M families and shared by affected relatives in a subset S_o of them, the p-value is obtained as the sum of the probability of events as or more extreme as the observed sharing between m out of M families. If we note p_m the sharing probability between the subjects in family m, then the p-value is obtained as:

$$p = \sum_{u \in U} \prod_{m=1}^{M} p_m^{I(m \in S_u)} (1 - p_m)^{I(m \in S_u^c)} \tag{23}$$

where U is the subset of family sets S_u such that

$$\prod_{m=1}^{M} p_m^{I(m \in S_u)} (1 - p_m)^{I(m \in S_u^c)} \leq \prod_{m=1}^{M} p_m^{I(m \in S_o)} (1 - p_m)^{I(m \in S_o^c)}$$

.

2.4. Defining the set of rare variants tested

The lowest possible p-value for a RV seen in only one or very few families depends on the family structures. The sharing probabilities between sequenced subjects in small or densely inbred families are high, and so is the potential p-value of a variant seen only in one such family (for instance, it is 1/7 for a grand-parent - grand-child pair). We propose to test the null hypothesis of absence of linkage and association only among the variants achieving a sufficiently low p-value if shared by all affected subjects in the family (or families) in which they are seen. We define the p-value threshold as the familywise Type I error level α divided by the number of variants included and solve this p-value threshold using the RV sharing probabilities based on the reported pedigree structure.

3. Results

3.1. Validation of the approximation of the sharing probabilities

We simulated small populations from which we sampled founders of a pedigree to validate the quality of the approximation of the sharing probabilities in presence of relatedness between the founders.

3.1.1. Simulated populations

The entire pedigree of small populations was simulated over 6 generations using the computer package Spip (Anderson 2005). The initial size of the population was set to 100, 200 or 400, with equal number of males and females. Population size increased at an average rate of 10 percent per generation. Although Spip allows the simulation of age-structured populations, we simulated non overlapping generations by specifying a single reproduction

time. Each subject had an 80 percent probability of reproducing and each reproducing female mated with only one male selected randomly from the same generation. The number of offspring per female followed a Poisson distribution. At the end of each simulation, we sampled 8 subjects from the sixth generation to be the founders of the pedigree for three second cousins shown on Figure 1. We chose that family structure which we encountered in our sample **?** to have three sequenced subjects with symmetric relationships. The simulation was repeated 100 times for each population size.

Kinship coefficients were estimated using the R package kinship2 (www.r-project.org). The distribution of kinship coefficients between subjects from the same generation had stabilized around the fifth generation (not shown). Table 1 shows the mean and standard deviation (SD) of the mean kinship coefficient between pairs of subjects from the population at the 6^{th} generation and the number of copies of the RV in the 8 subjects sampled to be the founders of the pedigree. With a population of 100 founders, the probability that the RV is introduced by more than two founders (given that it was seen in at least one founder) is too high (0.10) to obtain a good approximation of the RV sharing probability when assuming that the RV can only be introduced once or twice. The approximation was therefore only computed with 200 and 400 founders.

3.1.2. Approximation of sharing probabilities

The first step in applying approximation method 2 was to estimate the parameter θ of the distribution of the number of distinct alleles in the founders. We used two different values for $\hat{\phi}^f$: the mean kinship coefficient between the 8 sampled subjects and the mean kinship coefficient in the population. The former is a best case scenario where ϕ^f is estimated without error which is not possible in practice, while the latter can be approached with a

sufficiently larger sample from the population. We checked the quality of the approximations with the simulated data from one replicate (Figure 2). All approximations are reasonably good although the events of observing only 8 to 10 distinct alleles among the pedigree founders from the 200 founder population are not captured by the approximate distributions.

We approximated the RV sharing probability using the analytical formulas 21, 22 and those from Appendix A, and by Monte Carlo sampling 100,000 realizations of RV transmission in each replicate. The probability $P[F_j, F_k]$ that two founders introduce the variant derived from the approximate distribution overestimates on average the value in the simulated populations, particularly when the number of founders is 200 (Table 1). This overestimation compensates to some extent for ruling out the events three or more founders introduce the variant, and the events one or more founders introduce two copies of the variant. To evaluate the quality of the RV sharing approximation, we estimated the root mean squared error (RMSE) and bias over the simulation replicates. If we denote the RV sharing probability by β and its approximation by $\hat{\beta}$, then these quantities are defined in absolute terms as:

$$\begin{aligned} \text{RMSE} &= \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\beta}_r - \beta_r)^2} \\ \text{bias} &= \frac{1}{R}\sum_{r=1}^{R}(\hat{\beta}_r - \beta_r) \end{aligned}$$

where in our case $R = 100$. We also computed these quantities relative to the true value in each replicate:

$$\text{RMSE*} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\frac{\hat{\beta}_r - \beta_r}{\beta_r}\right)^2}$$

$$\text{bias}^* \quad = \quad \frac{1}{R}\sum_{r=1}^{R}\left(\frac{\hat{\beta}_r - \beta_r}{\beta_r}\right)$$

The approximation RMSE improves from 200 to 400 founders in absolute terms and to a lesser extent in relative terms (Table 2). The RV sharing probability approximation is accurate for the population with 400 founders with a negligible bias, where the mean kinship coefficient is approximately equal to second cousins once removed ($\frac{1}{128}$). The RV sharing probability is underestimated (negative bias) for the population with 200 founders where the mean kinship is close to first cousins once removed ($\frac{1}{32}$), indicating the limits of an approximation restricted to two founders introducing the rare variant. The loss of precision and accuracy from using the population average ϕ^f instead of the average of the sampled subjects is smaller in the population with 400 founders than in the population with 200 founders, both in absolute and relative terms. Sampling 100,000 realizations of RV transmission in the Monte Carlo simulation was sufficient to achieve the same level of error as the analytical approximation with the relatively larger RV sharing probability in populations of 200 founders, but the Monte Carlo error remained slightly larger to approximate the smaller RV sharing probability in populations of 400 founders.

3.2. Whole exome sequencing study of nonsyndromic oral cleft

We computed the sharing probability for all rare single nucleotide variants (SNVs) detected in exons and splice junctions in a whole exome sequencing study of affected relative pairs and trios drawn from 55 multiplex nonsyndromic oral cleft families from diverse sites [Germany, Philippines, India, Syria, China (Taiwan and Shanghai) and one European American family]. The study sample and sequencing methodology has been described elsewhere (Bureau et al., submitted). Briefly, 51 families provided 2 affected subjects and 4 families provided 3 affected subjects for a total of 114 sequenced subjects. Exon capture

was performed using the Agilent SureSelect Human All Exon Target Enrichment system on libraries of 150 to 200bp fragments prepared by fragmentation of 200ng of genomic DNA per subject. DNA sequencing was performed on an Illumina HiSeq 2500 instrument using standard protocols for a 100 bp paired-end run. Base calls and quality scores were obtained from Illuminas Real-Time Analysis (RTA) software, and reads were aligned to a reference genome with the Burrows-Wheeler Alignment (BWA) tool. Post-processing of the alignment, multi-sample variant calling and variant quality score recalibration were performed using the Genome Analysis Tool Kit (GATK) and only variants passing this step were included in analyses.

We defined a rare SNV as a SNV with a minor allele frequency (MAF) ¡0.01 based on the Exome Sequencing Project (ESP) database (esp.gs.washington.edu/drupal/), and a MAF ¡0.01 in the April 2012 release of the 1000 Genomes data (www.1000genomes.org). Variants not seen in 1000 Genomes data were retained if their MAF was ¡0.1 in an internal database of all exomes previously sequenced at CIDR, to distinguish variant calls resulting from technical artifacts. Variants seen in more than 20 percent of the families were excluded. After applying the above criteria 60,997 rare exonic and splice site SNVs were detected in the autosomal genome.

The p-value threshold to achieve a familywise Type I error rate of 0.05 was 2.2×10^{-5}, and 2,292 SNVs had the potential to yield a p-value below that threshold when sharing probabilities were computed based on the known pedigree structure. The SNV rs149253049 in gene ADAMTS9 had a p=2.0×10^{-6}. The G allele was shared by affected relatives in three families from India (Table 3) and was not seen in any other family. The G nucleotide is the rarest of the three alleles of rs149253049. It was not found in the Exome Sequencing

Project (ESP) database nor the 1000 Genomes project data (April 2012 release). It was seen once in the ClinSeq project for a frequency of 0.0001.

In our Indian sample, kinship estimates between affected subjects from genomewide SNP genotypes were based on the estimator of Manichaikul Manichaikul et al. (2010) robust to population stratification. There was no evidence of IBD sharing in excess of the known degree of relatedness, nor of relatedness between subjects from distinct Indian families, as we reported elsewhere (Bureau et al., submitted), and using equation 15 we obtain an estimated mean kinship of the founders $\hat{\phi}^f = 0.0$. All this suggests that sharing probabilities for rs149253049 computed based on known pedigree structures and indepedence between families are accurate.

Another SNV had a p-value below the above threshold: rs117883393 in the gene OR2A2 ($p = 6.1 \times 10^{-6}$). That SNV was shared in heterozygous state by all sequenced subjects in three families and present in heterozygous state in one of the two sequenced subject of a fourth family. Its frequency in the ESP database is 0.0063 for the whole sample and 0.0081 for the European American subsample. We have reasons to suspect the sharing probabilities may be underestimated in two of the families where the variant is shared because these families are from the Syrian sample, where cultural and demographic factors make relationships between pedigree founders more likely. In our Syrian sample we used the moment estimator of Manichaikul et al. (2010) based on population allele frequencies estimated in that sample instead of the robust estimator because the latter tended to give negative estimates when the level of estimated inbreeding differed substantially between the two subjects (results not shown). We then inferred ϕ^f using equation 15 and obtained $\hat{\phi}^f = 0.013$, close to the kinship coefficient of second cousins ($\frac{1}{64}$).

For the family shown on Figure 3A, the RV sharing probability obtained from equation 4 was 0.0030. The probability that the rare allele of SNV rs117883393 was introduced by two founders was equal to 0.092 using approximation method 2 with $\hat{\phi}^f = 0.013$, leading to an adjusted RV sharing probability of 0.0047. For the family shown on Figure 3B, the RV sharing probability obtained from equation 4 was 0.011. The probability that the rare allele was introduced by two founders was also equal to 0.092, leading to an adjusted RV sharing probability of 0.018. With these adjusted RV sharing probabilities for the two Syrian families, and assuming no unknown relationships in the two German families, the p-value for the four families increased to 1.4×10^{-5}.

4. Discussion

In this paper we propose using the probability of sharing of a RV by affected subjects under the null hypothesis of absence of linkage and association of the RV with a disease to build the evidence against that null hypothesis in the context of exome sequencing studies of complex diseases in family samples. We have presented formulas to compute exactly the probability of sharing of a RV by any number of affected subjects in arbitrary non-inbred pedigrees under the assumption that the variant is sufficiently rare to be introduced only once in the pedigree, generalizing a previous formula applicable to two affected subjects. These formulas are implemented in the `RVsharing` R package.

We re-emphasize that RV sharing probabilities are not the same as IBD sharing probabilities, which only captures linkage information without taking into account that the IBD sharing involves a RV. This is most easily illustrated with two unilinealy related subjects, for which the probability of IBD sharing is $\frac{1}{2^D}$ while the RV sharing probability is $\frac{1}{2^{(D+1)}-1}$. The ratio P[RV shared]/P[IBD] tends to $\frac{1}{2}$ as D tends to infinity.

In our application of the proposed approach to an exome sequencing study of oral cleft in 55 multiplex families, we analyzed all exonic and splicing variants with a MAF < 1 percent, with the only additional restriction that the families where the variant was detected needed to be sufficiently informative to produce a p-value that would remain significant after Bonferroni correction. This allowed us to reject the null hypothesis of absence of linkage and association to oral cleft for the G allele of SNV rs149253049 in gene ADAMTS9 shared by two distantly related oral cleft cases in three families from India. Given that rs149253049 is a synonymous nucleotide change, it would be discarded under filtering strategies keeping only nonsynonymous or truncating variants. Interestingly, the G allele shared in the three families was the rarest of the three nucleotides A,C and G of that SNV, and has not been reported in the ESP and the 1000 Genomes. We provide statistical evidence warranting further investigation of the role this variant and the gene ADAMTS9 may have in causing oral cleft.

A potential pitfall with RV sharing probabilities based on a known pedigree structure is the possibility of cryptic relatedness between founders of the pedigree that would make the actual null sharing probability greater than the one computed. We have developed two methods to adjust the RV sharing probability based on estimates of the kinship coefficients between founders of the known pedigree. The first approximation method can be applied to obtain RV sharing probabilities in inbred pedigrees, by removing familial links to obtain a non-inbred pedigree and using the known kinship coefficient between the subjects whose familial links were removed. It can also be applied when relatedness is suspected between a subset of founders only. The second method allows more extensive relatedness between all founders but assumes equal kinship coefficients between all pairs of founders. Our

simulation study on a pedigree where the founders were drawn from the 6^{th} generation of larger pedigrees representing small populations showed that the second method gives accurate approximations when the mean kinship coefficient between founder of the known pedigree is of the order of second cousins once removed, but underestimates the RV sharing probability with closer relationships.

An important aspect of these approximation methods is that they are based solely on estimated kinship between founders, and do not require an estimate of the RV frequency in the population from which the pedigree founders come from. We have proposed a formula to estimate mean kinship among founders based on the kinship estimates between the sequenced subjects. A number of methods have been implemented to estimate kinship coefficients from genomewide genotype data (Manichaikul et al. 2010; Yang et al. 2011; Speed et al. 2012; Thornton et al. 2012) and an appraisal of these methods is beyond the scope of this paper. Since our approximation method requires only a mean kinship coefficient between founders, variation in the length of genome shared by pairs of subjects is smoothed by the averaging. Using a population average instead of an average over the founder pairs of the pedigree had a moderate impact on the error of the approximation in our simulation study. In our Syrian sample where we suspected relationships between founders due to cultural practices, the application of the second approximation method to the two families whose sequenced affected members share the rare allele of SNV rs117883393 reduced the evidence against the null for that SNV.

For this work, we have implemented the formulas to compute the approximation of the RV sharing probability based on our method 2 for the family structures for which we needed to obtain such approximation, shown on Figures 1 and 3. Developing a general

implementation of the formulas of Appendix A applicable to general pedigree structures is however challenging. We have shown in our simulation study that an accurate Monte Carlo approximation can be achieved with a reasonnable number of draws for sharing probabilities of the order of 10^{-3}. The Monte Carlo approximation is implemented for arbitrary pedigree structures in our R package `RVsharing`.

In our extension of the RV sharing probability to more than two subjects, we have considered only the probability that all affected sequenced subjects share a RV. This is appropriate for three affected subjects sequenced in a pedigree as in the oral cleft study, where causal RV not shared by all sequenced subjects are indistinguishable from benign RVs. As sequencing of more affected subjects from large multiplex families becomes more common with decreasing sequencing costs, the requirement that all affected sequenced subjects share a RV becomes a too stringent requirement, given the intra-familial heterogeneity in disease causation that characterizes complex traits (Feng et al. 2011). At the same time, with $n > 3$ sequenced subjects in a family, the event that n-1 or n-2 affected subjects out of n share a RV is evidence of the variant involvement in the disease. The computation of the probability of such events in pedigrees of arbitrary shape will require new developments.

Non-affected family members may also be included in future sequencing studies. While sequencing non-affected family members has been used to exclude private benign variation in studies of Mendelian traits (Gilissen et al. 2012), this would risk excluding causal variants with incomplete penetrance in studies of complex traits. An affected only analysis of RV sharing protects against unaffected carriers reducing evidence for a variant in the same way as it does in linkage analysis (McPeek 1999). Sequence data on non-affected family members, in particular subjects marrying in the pedigree, will be useful to narrow down the

number of founders that could have introduced a given RV in the pedigree and refine the RV sharing probabilities.

The methods and analyses presented are limited to a single RV at a time. Our results illustrate that with few families it is possible to obtain substantial evidence of co-segregation of a rare variant with disease. Yet very rare causal variants found in a single family were not considered in our analysis of the oral cleft family sample because of the insufficient informativeness of individual families. A combined analysis of multiple RVs from the same functional unit, typically the same gene, will be needed to detect significant RV sharing at the level of that functional unit. Various issues will need to be resolved to implement such analysis, in particular dealing with multiple RVs in the same family. While an exact p-value can be computed for the same variant seen in multiple families assuming independence between families, obtaining p-values from the asymptotic distribution of a RV sharing summary statistic seems a more promising approach for large numbers of RVs in a gene. This will be the object of future work.

5. Web resources

The URLs for data presented herein are as follows:

`RVsharing` R package:

Table 1. Founder relatedness and distribution of number of copies of a rare variant for three second cousins in small populations

N founders	100	200		400	
mean (SD) of mean ϕ^f	0.043 (0.004)	0.0216 (0.0015)		0.0108 (0.0006)	
mean (SD) of P[RV sharing]	0.0073 (0.0026)	0.0039 (0.0023)		0.0022 (0.0007)	
N founders with RV	Simulated	Simulated	Approx	Simulated	Approx
1	0.62	0.83	0.74	0.95	0.93
2	0.28	0.15	0.26	0.05	0.07
3+	0.10	0.02	0	0.002	0

Table 2. Approximation of rare variant sharing probabilities for three second cousins in small populations

	N founders	200		400	
Parameter	type	sample ϕ^f	pop. ϕ^f	sample ϕ^f	pop. ϕ^f
Analytical approximation					
RMSE	absolute	0.0015	0.0026	0.00056	0.00068
	relative	0.27	0.34	0.24	0.28
Bias	absolute	-0.0009	-0.0012	-0.00017	-0.00015
	relative	-0.18	-0.18	-0.022	0.010
Monte Carlo approximation					
RMSE	absolute	0.0015	0.0026	0.00061	0.00075
	relative	0.27	0.35	0.26	0.32
Bias	absolute	-0.0009	-0.0012	-0.00015	-0.00014
	relative	-0.17	-0.17	-0.011	0.019

Table 3. Sharing probabilities for rs149253049

Relationship between affecteds	degree	sharing probability
first cousins	3	0.067
third cousins	7	0.0039
second cousins once removed	6	0.0079
Product		2.0×10^{-6}

REFERENCES

Anderson, E. C. (2005). An efficient monte carlo method for estimating ne from temporally spaced samples using a coalescent-based likelihood. *Genetics 170*(2), 955–67.

Beaty, T. H., M. A. Taub, A. F. Scott, J. C. Murray, M. L. Marazita, H. Schwender, M. M. Parker, J. B. Hetmanski, P. Balakrishnan, M. A. Mansilla, E. Mangold, K. U. Ludwig, M. M. Noethen, M. Rubini, N. Elcioglu, and I. Ruczinski (2013). Confirming genes influencing risk to cleft lip with/without cleft palate in a case-parent trio study. *Hum Genet*.

Cirulli, E. T. and D. B. Goldstein (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet 11*(6), 415–25.

Feng, B. J., S. V. Tavtigian, M. C. Southey, and D. E. Goldgar (2011). Design considerations for massively parallel sequencing studies of complex human disease. *PLoS One 6*(8), e23221.

Gilissen, C., A. Hoischen, H. G. Brunner, and J. A. Veltman (2012). Disease gene identification strategies for exome sequencing. *Eur J Hum Genet 20*(5), 490–7.

Ludwig, K. U., E. Mangold, S. Herms, S. Nowak, H. Reutter, A. Paul, J. Becker, R. Herberz, T. AlChawa, E. Nasser, A. C. Bohmer, M. Mattheisen, M. A. Alblas, S. Barth, N. Kluck, C. Lauster, B. Braumann, R. H. Reich, A. Hemprich, S. Potzsch, B. Blaumeiser, N. Daratsianos, T. Kreusch, J. C. Murray, M. L. Marazita, I. Ruczinski, A. F. Scott, T. H. Beaty, F. J. Kramer, T. F. Wienker, R. P. Steegers-Theunissen, M. Rubini, P. A. Mossey, P. Hoffmann, C. Lange, S. Cichon, P. Propping, M. Knapp, and M. M. Nothen (2012). Genome-wide meta-analyses of nonsyndromic cleft lip with or without cleft palate identify six new risk loci. *Nat Genet 44*(9), 968–71.

Manichaikul, A., J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, and W. M. Chen (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics 26*(22), 2867–73.

McPeek, M. S. (1999). Optimal allele-sharing statistics for genetic mapping using affected relatives. *Genet Epidemiol 16*(3), 225–49.

Speed, D., G. Hemani, M. R. Johnson, and D. J. Balding (2012). Improved heritability estimation from genome-wide snps. *Am J Hum Genet 91*(6), 1011–21.

Thompson, E. A. (1986). *Pedigree Analysis in Human Genetics*. Baltimore: Johns Hopkins University Press.

Thornton, T., H. Tang, T. J. Hoffmann, H. M. Ochs-Balcom, B. J. Caan, and N. Risch (2012). Estimating kinship in admixed populations. *Am J Hum Genet 91*(1), 122–38.

Wijsman, E. M. (2012). The role of large pedigrees in an era of high-throughput sequencing. *Hum Genet 131*(10), 1555–63.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher (2011). Gcta: a tool for genome-wide complex trait analysis. *Am J Hum Genet 88*(1), 76–82.

This manuscript was prepared with the AAS LaTeX macros v5.2.

A. Sharing probabilities conditional on the introduction of the RV by two of the founders

We need to introduce an additional type of subjects, the descendants that are common to the two founders introducing the RV, and who can therefore receive two copies of the variant. We note the number of copies of the RV in such a subject h by T_h.

As before, we begin by the expressions for the special case where all the sequenced subjects descend from every founder among their ancestors through independent lines of descent. With two founders introducing the RV, we further need to distinguish four events.

A.1. The lines of descent to every sequenced subject are common to the two founders introducing the variant

This implies that the two founders introducing the RV are mates and their descendants in common are their children. With the assumption of independent lines of descent, the n sequenced individuals descend from n children of the founders and

$$\begin{aligned}
P[C_1 = \ldots = C_n = 1|F_j, F_k] &= \sum_{x=0}^{n} P[C_1 = \ldots = C_n = 1|\sharp\{i : T_i = 2\} = x, \sharp\{i : T_i = 1\} = n - x, F_j, F_k] \\
&\quad P[\sharp\{i : T_i = 2\} = x, \sharp\{i : T_i = 1\} = n - x|F_j, F_k] \qquad \text{(A1)} \\
&= \sum_{x=0}^{n} \left(\frac{1}{2}\right)^{\sum_{\{i:T_i=2\}} D_{ij}-2} \left(\frac{1}{2}\right)^{\sum_{\{i:T_i=1\}} D_{ij}-1} \binom{n}{x} \left(\frac{1}{4}\right)^{x} \left(\frac{1}{2}\right)^{n-x} \\
&= \sum_{x=0}^{n} \left(\frac{1}{2}\right)^{D^s-n-x} \binom{n}{x} \left(\frac{1}{2}\right)^{2x} \left(\frac{1}{2}\right)^{n-x} = \left(\frac{1}{2}\right)^{D^s} \sum_{x=0}^{n} \binom{n}{x} \\
&= \left(\frac{1}{2}\right)^{D^s-n}
\end{aligned}$$

where $D^s = \sum_i D_{ij}$ and $D_{ij} = D_{ik} \forall i$. This expression applies if all $D_{ij} \geq 2$, i.e. the sequenced subjects are grand-children or more distant descendants of the founders. When

a sequenced subject is a children of the founders, then $C_i = T_i$. We adapt the formula to distinguish the n_c sequenced subjects who are children of the founders from the others.

$$\begin{aligned} P[C_1 \geq 1, \ldots, C_{n_c} \geq 1, C_{n_c+1} = \ldots = C_n = 1 | F_j, F_k] &= P[C_1 \geq 1, \ldots, C_{n_c} \geq 1 | F_j, F_k] \\ & \quad P[C_{n_c+1} = \ldots = C_n = 1 | F_j, F_k] \\ &= \left(\frac{3}{4}\right)^{n_c} \left(\frac{1}{2}\right)^{(D^s - n_c) - (n - n_c)} \\ &= \left(\frac{3}{4}\right)^{n_c} \left(\frac{1}{2}\right)^{D^s - n} \end{aligned} \tag{A2}$$

The expression for the probability of not seeing the variant in any sequenced individual when all $D_{ij} \geq 2$ is:

$$\begin{aligned} P[C_1 = \ldots = C_n = 0 | F_j, F_k] &= \sum_{x=0}^{n} \sum_{y=0}^{n-x} P[C_1 = \ldots = C_n = 0 | \sharp\{i : T_i = 2\} = x, \sharp\{i : T_i = 1\} = y, F_j, F_k] \\ & \quad P[\sharp\{i : T_i = 2\} = x, \sharp\{i : T_i = 1\} = y | F_j, F_k] \\ &= \sum_{x=0}^{n} \sum_{y=0}^{n-x} \prod_{\{i:T_i=2\}} \left(1 - \left(\frac{1}{2}\right)^{D_{ij}-2}\right) \prod_{\{i:T_i=1\}} \left(1 - \left(\frac{1}{2}\right)^{D_{ij}-1}\right) \\ & \quad \binom{n}{x, y, n-x-y} \left(\frac{1}{4}\right)^{x} \left(\frac{1}{2}\right)^{y} \left(\frac{1}{4}\right)^{n-x-y} \end{aligned} \tag{A3}$$

without obvious simplification. The modification for sequenced subjects who are children of the founders is similar to that for the joint sharing probability, with probability equal to $\frac{1}{4}$ of not receiving the variant instead of $\frac{3}{4}$ of receiving it.

A.2. One founder is ancestor of all sequenced subjects and the other is ancestor of only one subject

We note j the founder who is ancestor of all sequenced subjects and 1 the sequenced subject descendant of the two founders j and k. There is only one child of founder k who can receive two copies of the variant (possibly subject 1 himself) and we note that child h. The number of copies he received is noted T.

$$\begin{aligned} P[C_1 = \ldots = C_n = 1|F_j, F_k] &= P[C_1 = \ldots = C_n = 1|T = 2, F_j, F_k]P[T = 2|F_j, F_k] \quad \text{(A4)} \\ &\quad + P[C_1 = \ldots = C_n = 1|T = 1, F_j, F_k]P[T = 1|F_j, F_k] \\ &= \left(\frac{1}{2}\right)^{D_{1h}-1+\sum_{i=2}^{n} D_{ij}} \left(\frac{1}{2}\right)^{D_{hj}} \frac{1}{2} \\ &\quad + \left(\frac{1}{2}\right)^{D_{1h}+\sum_{i=2}^{n} D_{ij}} \left[\left(\frac{1}{2}\right)^{D_{hj}} \frac{1}{2} + \left(1 - \left(\frac{1}{2}\right)^{D_{hj}}\right) \frac{1}{2}\right] \\ &= \left(\frac{1}{2}\right)^{D_{1h}+\sum_{i=2}^{n} D_{ij}} \left[\left(\frac{1}{2}\right)^{D_{hj}} + \frac{1}{2}\right] \end{aligned}$$

This expression applies if $D_{1h} \geq 1$, i.e. subject 1 is not h himself, he or she is a grand-child or more distant descendant of the founder k. When subject 1 is a child of founder k, the expression becomes:

$$\begin{aligned} P[C_1 \geq 1, C_2 = \ldots = C_n = 1|F_j, F_k] &= P[C_1 = 2, C_2 = \ldots = C_n = 1|F_j, F_k] \quad \text{(A5)} \\ &\quad + P[C_1 = \ldots = C_n = 1|F_j, F_k] \\ &= \left(\frac{1}{2}\right)^{D^s} \frac{1}{2} + \left(\frac{1}{2}\right)^{\sum_{i=2}^{n} D_{ij}} \frac{1}{2} \\ &= \left(\frac{1}{2}\right)^{D^s+1} \left[1 + 2^{D_{1j}}\right] \end{aligned}$$

The expression for the probability of not seeing the variant in any sequenced subject

when $D_{ih} \geq 1$ is:

$$\begin{aligned} P[C_1 = \ldots = C_n = 0|F_j, F_k] &= \prod_{i=1}^{n} P[C_i = 0|F_j, F_k] \qquad \text{(A6)} \\ &= \left[\begin{array}{c} P[C_1 = 0|T = 2, F_j, F_k]P[T = 2|F_j, F_k] \\ +P[C_1 = 0|T = 1, F_j, F_k]P[T = 1|F_j, F_k] \\ +P[C_1 = 0|T = 0, F_j, F_k]P[T = 0|F_j, F_k] \end{array} \right] \prod_{i=2}^{n} P[C_i = 0|F_j] \\ &= \left[\begin{array}{c} \left(1 - \left(\frac{1}{2}\right)^{D_{1h}-1}\right)\left(\frac{1}{2}\right)^{D_{hj}}\frac{1}{2} \\ + \left(1 - \left(\frac{1}{2}\right)^{D_{1h}}\right)\frac{1}{2} \\ + \left(1 - \left(\frac{1}{2}\right)^{D_{hj}}\right)\frac{1}{2} \end{array} \right] \prod_{i=2}^{n} \left(1 - \left(\frac{1}{2}\right)^{D_{ij}}\right) \end{aligned}$$

The same probability when subject 1 is a child of founder k is

$$P[C_1 = \ldots = C_n = 0|F_j, F_k] = \prod_{i=1}^{n} P[C_i = 0|F_j, F_k] = \frac{1}{2}\prod_{i=1}^{n} \left(1 - \left(\frac{1}{2}\right)^{D_{ij}}\right) \qquad \text{(A7)}$$

A.3. Each founder is ancestor of one sequenced subject

We assume that founder j is ancestor of subject 1 and founder k is ancestor of subject 2. The formula applies equally to the case where both j and k are ancestor of the same subject, as long as one is ancestor of his mother and the other ancestor of his father (or j or k are themselves either father or mother of the subject). If there are $n > 2$ sequenced subjects, then $P[C_1 = \ldots = C_n = 1|F_j, F_k] = 0$. If $n = 2$, then

$$P[C_1 = C_2 = 1|F_j, F_k] = P[C_1 = 1|F_j]P[C_2 = 1|F_k] = \left(\frac{1}{2}\right)^{D_{1j}+D_{2k}} \qquad \text{(A8)}$$

The expression for the probability of not seeing the variant in any sequenced subject is

$$\begin{aligned} P[C_1 = \ldots = C_n = 0|F_j, F_k] &= P[C_1 = 0|F_j]P[C_2 = 0|F_k] \\ &= \left(1 - \left(\frac{1}{2}\right)^{D_{1j}}\right)\left(1 - \left(\frac{1}{2}\right)^{D_{2k}}\right) \end{aligned} \quad \text{(A9)}$$

A.4. Extension to branching in the pedigree

Having two founders introducing the variant require to adapt the formulas of section 2.1. Equation 6 becomes

$$\begin{aligned} P[C_1 = \ldots = C_n = 1] &= P[C_1 = \ldots = C_{i_k} = 1|B_k = 1]P[B_k = C_{i_k+1} = \ldots = C_n = 1] \\ &\quad + P[C_1 = \ldots = C_{i_k} = 1|B_k = 0]P[B_k = 0, C_{i_k+1} = \ldots = C_n = 1] \end{aligned} \quad \text{(A10)}$$

The term $P[C_1 = \ldots = C_{i_k} = 1|B_k = 1]$ is not directly computable, and we instead compute terms $P[C_1 = \ldots = C_{i_k} = 1|F_j, B_k = 1]$ for every founder j below the branching subject k in the pedigree (in the sense defined in chapter 4 of Thompson (Thompson 1986)), which can be done using equations A1, A2, A4 or A5, depending on the relationship between j and k. These terms can then be summed over all founders j below k, with equal weight under approximation method 2, or weigthed by $P[F_j|B_k = 1]$ computed from the kinship matrix between founders under approximation method 1. The term $P[C_1 = \ldots = C_{i_k} = 1|B_k = 0]$ is computed by applying equation 2 with every founders j below k. The other terms are computed by reapplying equation A10 recursively with the other branching individuals, with slight modification for the terms where $B_k = 0$ instead of 1.

In equation 8, the term $P[C_1 = \ldots = C_{i_k} = 0|B_k = 1, F_j]$ does not simplify anymore when j is a founder below k, but can be computed using equations A3, A6 or A7.

Similarly, the term $P[C_1 = \ldots = C_{i_k} = 0|B_k = 0, F_j]$ no longer equals 1 but equals $P[C_1 = \ldots = C_{i_k} = 0|F_j]$ which can be computed using equation 3. If instead founders h and j introducing the variant are both ancestors of branching individual k (*e*.g. his parents), then one must consider the event $B_k = 2$. Additional terms are then computed as follows:

$$P[C_1 = \ldots = C_{i_k} = 0|B_k = 2] = \prod_{i=1}^{i_k} \left(1 - \left(\frac{1}{2}\right)^{D_{ik}-1}\right) \tag{A11}$$

If there is no other branching individual between either founder h or j and branching individual k, then

$$P[B_k = 2, C_{i_k+1} = \ldots = C_n = 0|F_h, F_j] = \left(\frac{1}{2}\right)^{D_{kh}+D_{kj}} P[C_{i_k+1} = \ldots = C_n = 0|F_h, F_j] \tag{A12}$$

$$P[B_k = 1, C_{i_k+1} = \ldots = C_n = 0|F_h, F_j] = \left[\left(\frac{1}{2}\right)^{D_{kh}} \left(1 - \left(\frac{1}{2}\right)^{D_{kj}}\right) + \left(\frac{1}{2}\right)^{D_{kj}} \left(1 - \left(\frac{1}{2}\right)^{D_{kh}}\right)\right] P[C_{i_k+1} = \ldots = C_n = 0|F_h, F_j] \tag{A13}$$

$$P[B_k = 0, C_{i_k+1} = \ldots = C_n = 0|F_h, F_j] = \left(1 - \left(\frac{1}{2}\right)^{D_{kh}+D_{kj}}\right) P[C_{i_k+1} = \ldots = C_n = 0|F_h, F_j] \tag{A14}$$

With other intervening branching individuals a recursion similar to equation 8 would be needed.

B. Computation of $\bar{\phi}_a$

We explain here the expressions for the terms of equation 18. The probability that alleles from two founders are IBD given one of the founders shares an allele IBD with m other founders where $m = 1$ or 2 is simply the probability of randomly sampling one of these m founders times the probability of sampling the allele shared IBD in the two

founders, that is

$$P[\text{Alleles from two founders are IBD}|\text{One of the founders shares an allele IBD with } m \text{ other founders}] = \frac{m}{4(n_f - 1)}$$

The probability that one of the founders shares an allele IBD with 2 other founders is the probability for that founder to have received as his first allele one of the two copies of the $2n_f - a$ alleles present in two copies among the $2n_f$ founder alleles, and as his second allele one of the two copies of the $2n_f - a - 1$ remaining alleles with two copies among the $2n_f - 2$ remaining eligible alleles (excluding sampling the second copy of the same allele, because we assume founders are not inbred). Each of these alleles is present in two copies, so

$$\begin{aligned} P[\text{One of the founders shares an allele IBD with 2 other founders}] &= \frac{2(2n_f - a)}{2n_f}\frac{2(2n_f - a - 1)}{2n_f - 2} \\ &= \frac{2n_f - a}{n_f}\frac{2n_f - a - 1}{n_f - 1} \end{aligned}$$

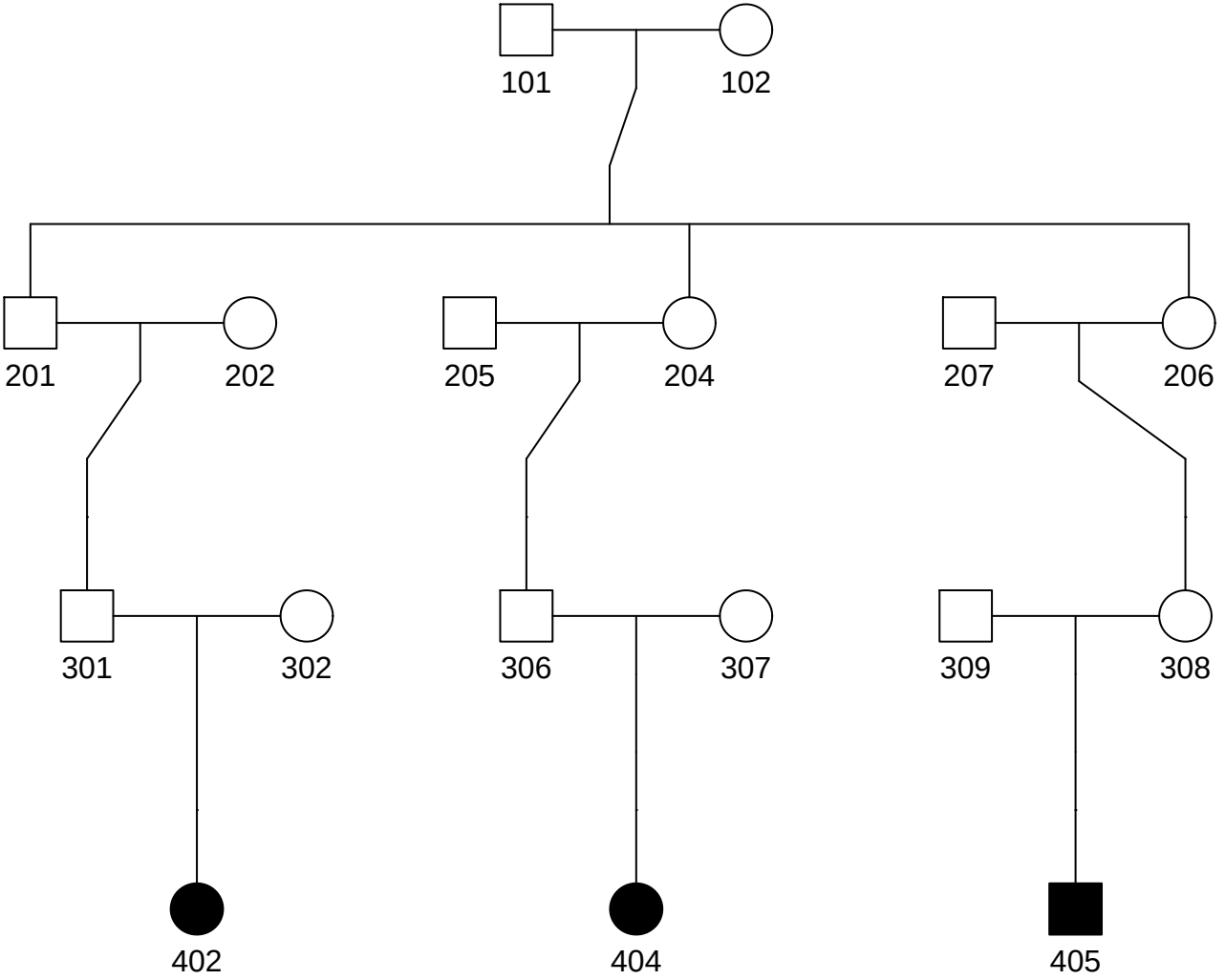
The probability that one of the founders shares an allele IBD with 1 other founder is the probability for that founder to have received as his first allele one of the two copies of the $2n_f - a$ alleles present in two copies among the $2n_f$ founder alleles, and as his second allele one of the $2a - 2n_f$ alleles present in a single copy among the $2n_f - 2$ remaining eligible alleles, or the reverse, that is to have received as his first allele one of the $2a - 2n_f$ alleles present in a single copy among the $2n_f$ founder alleles, and as his second allele one of the two copies of the $2n_f - a$ alleles present in two copies among the $2n_f - 1$ remaining alleles. The probability of the event of interest is then:

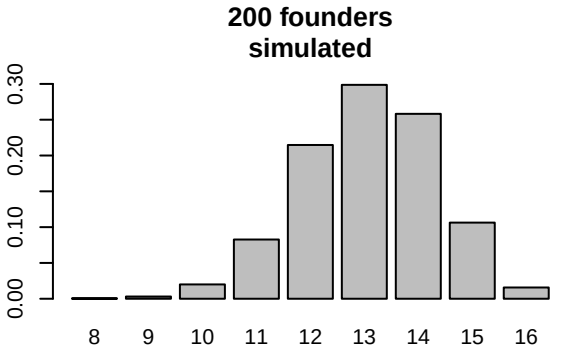
$$
\begin{aligned}
&P[\text{One of the founders shares an allele IBD with 1 other founder}] \\
&= \frac{2(2n_f - a)}{2n_f} \frac{2a - 2n_f}{2n_f - 2} + \frac{2a - 2n_f}{2n_f} \frac{2(2n_f - a)}{2n_f - 1} \\
&= \frac{(2n_f - a)(a - n_f)}{n_f(n_f - 1)} + \frac{2(2n_f - a)(a - n_f)}{n_f(2n_f - 1)}
\end{aligned}
$$

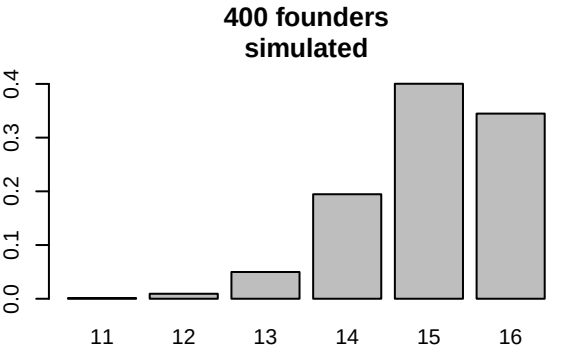
Fig. 1.— Pedigree of three second cousins used in simulation study

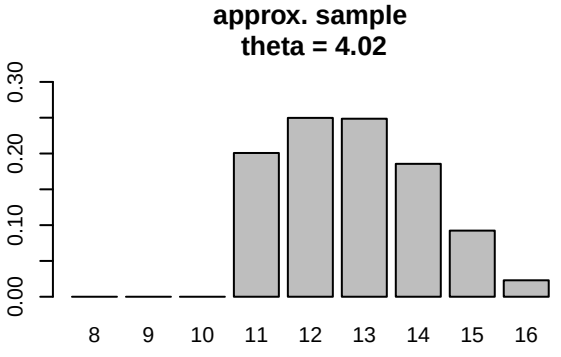
Fig. 2.— Number of distinct alleles in a sample of eight subjects from small populations

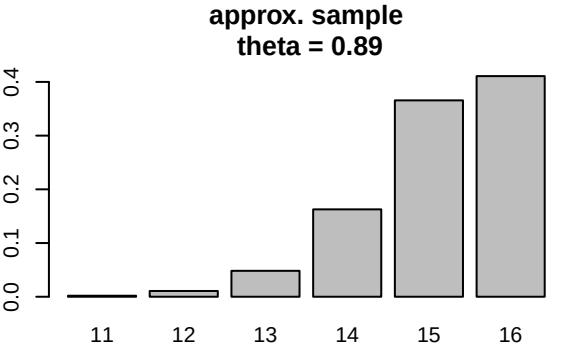
Fig. 3.— Syrian families sharing rare allele of rs117883393

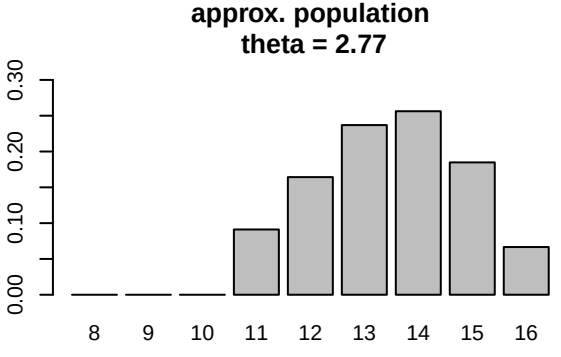

101
102
201
202
205
204
207
206
301
302
306
307
309
308
402
404
405

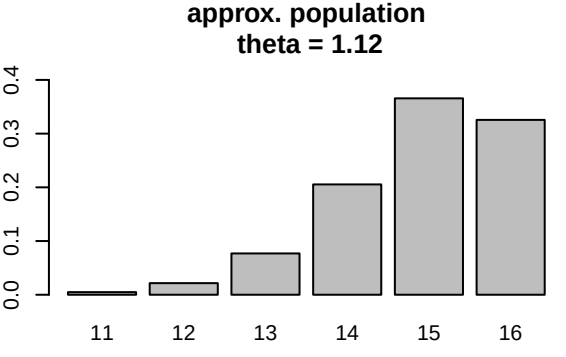

200 founders
simulated
0.00
0.10
0.20
0.30
8
9
10
11
12
13
14
15
16

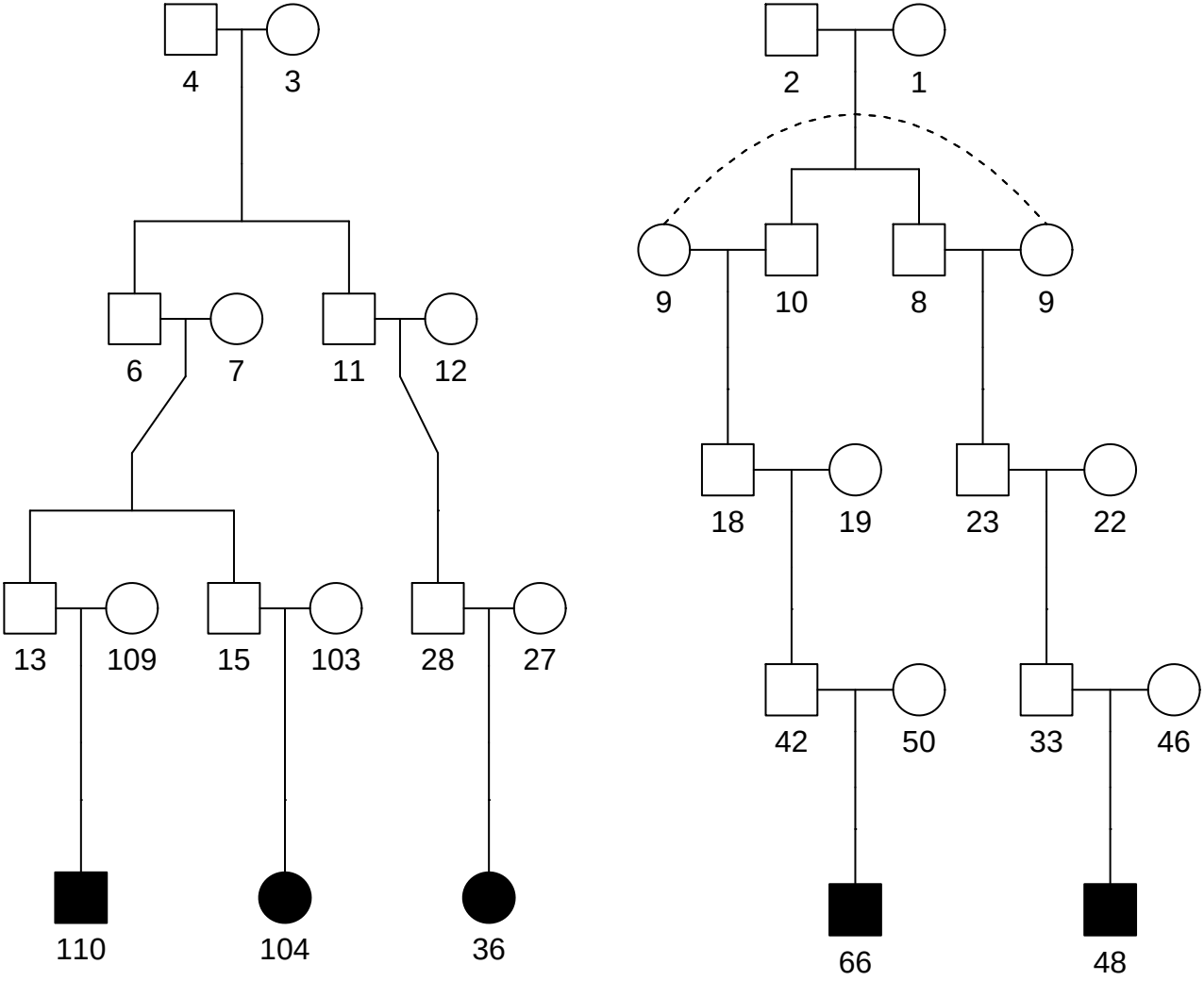

400 founders
simulated
0.0
0.1
0.2
0.3
0.4
11
12
13
14
15
16


approx. sample
theta = 4.02
0.00
0.10
0.20
0.30
8
9
10
11
12
13
14
15
16


approx. sample
theta = 0.89
0.0
0.1
0.2
0.3
0.4
11
12
13
14
15
16


approx. population
theta = 2.77
0.00
0.10
0.20
0.30
8
9
10
11
12
13
14
15
16


approx. population
theta = 1.12
0.0
0.1
0.2
0.3
0.4
11
12
13
14
15
16


4
3
6
7
11
12
13
109
15
103
28
27
110
104
36
2
1
9
10
8
9
18
19
23
22
42
50
33
46
66
48