Inferring rare disease risk variants based on exact probabilities of sharing by multiple affected relatives: Supplementary Data

Alexandre Bureau[1,2], Samuel G. Younkin[3], Margaret M Parker[4], Joan E Bailey-Wilson[5], Mary L Marazita[6], Jeffrey C Murray[7], Elisabeth Mangold[8], Hasan Albacha-Hejazi[9], Terri H Beaty[4], Ingo Ruczinski[3].

[1] Centre de Recherche de l'Institut Universitaire en Santé Mentale de Québec, Québec, G1J 2G3, Canada.

[2] Département de Médecine Sociale et Préventive, Université Laval, Québec, G1V 0A6, Canada.

[3] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, 21205, USA.

[4] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, 21205, USA.

[5] Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, MD, 21224, USA.

[6] Department of Oral Biology, Center for Craniofacial and Dental Genetics, School of Dental Medicine, University of Pittsburgh, PA, 15219, USA.

[7] Department of Pediatrics, School of Medicine, University of Iowa, IA, 52242, USA.

[8] Institute of Human Genetics, University of Bonn, Bonn, D-53127, Germany.

[9] Dr. Hejazi Clinic, P.O. Box 2519, Riyadh 11461, Saudi Arabia.

A. Recursive computation of rare variant sharing probabilities in pedigrees with lineages coalescing below the founders

Let B_k denote the number of copies of the RV in subject k where a line of descent from a founder branches into two separate lines of descent to a subset of sequenced subjects. 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 for the numerator of equation (1)

$$P[C_1 = \cdots = C_n = 1] = P[C_1 = \cdots = C_{i_k} = 1 | B_k = 1] P[B_k = C_{i_k+1} = \cdots = C_n = 1] \quad \text{(A1)}$$

because $P[C_1 = \cdots = C_{i_k} = 1 | B_k = 0] = 0$. The term $P[C_1 = \cdots = C_{i_k} = 1 | B_k = 1]$ is computed from equation (2.1) replacing F_j by $B_k = 1$. The term $P[B_k = C_{i_k+1} = \cdots = C_n = 1]$ is then computed by reapplying equation (A1) recursively for every branching individual.

Also, for a founder above a branching individual in the pedigree, we have in equation (1)

$$\begin{aligned} & P[C_1 = \cdots = C_n = 0 | F_j] \\ = \; & P[C_1 = \cdots = C_{i_k} = 0 | B_k = 1, F_j] \times P[B_k = 1, C_{i_k+1} = \cdots = C_n = 0 | F_j] \\ & + P[C_1 = \cdots = C_{i_k} = 0 | B_k = 0, F_j] \times P[B_k = C_{i_k+1} = \cdots = C_n = 0 | F_j] \\ = \; & P[C_1 = \cdots = C_{i_k} = 0 | B_k = 1] \times P[B_k = 1, C_{i_k+1} = \cdots = C_n = 0 | F_j] \\ & + P[B_k = C_{i_k+1} = \cdots = C_n = 0 | F_j] \qquad \text{(A2)} \end{aligned}$$

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

$$\begin{aligned}
& P[B_k = 1, C_{i_k+1} = \cdots = C_n = 0|F_j] \\
&= P[B_k = 1, C_{i_k+1} = \cdots = C_{i_h} = 0|B_h = 1] \times P[B_h = 1, C_{i_h+1} = \cdots = C_n = 0|F_j] \\
&= \left(\frac{1}{2}\right)^{D_{kh}} P[C_{i_k+1} = \cdots = C_{i_h} = 0|B_h = 1] \times P[B_h = 1, C_{i_h+1} = \cdots = C_n = 0|F_j] \quad \text{(A3)}
\end{aligned}$$

and similarly to equation (A2),

$$\begin{aligned}
& P[B_k = C_{i_k+1} = \cdots = C_n = 0|F_j] \\
&= \left(1 - \left(\frac{1}{2}\right)^{D_{kh}}\right) \times P[C_{i_k+1} = \cdots = C_{i_h} = 0|B_h = 1] \times P[B_h = 1, C_{i_h+1} = \cdots = C_n = 0|F_j] \\
&\quad + P[B_h = C_{i_h+1} = \cdots = C_n = 0|F_j] \qquad \text{(A4)}
\end{aligned}$$

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

B. Estimating the common kinship coefficient among founders based on estimated kinship coefficients between sequenced subjects

We express the kinship coefficient between subjects i_1 and i_2 as a function of the kinship coefficients between founder pairs as follows:

$$\begin{aligned}
\phi_{i_1 i_2} &= \phi^f \sum_j \sum_{k>j} \left[\left(\frac{1}{2}\right)^{D_{i_1 j} + D_{i_2 k}} I(j\&k \text{ not mating}) + \left(\frac{1}{2}\right)^{D_{i_1 j} + D_{i_2 k} - 1} I(j\&k \text{ mating}) \right] + \phi^p_{i_1 i_2} \\
&= \phi^f \kappa_{i_1 i_2} + \phi^p_{i_1 i_2} \qquad \text{(B1)}
\end{aligned}$$

where $I(C)$ is the indicator function taking value 1 if the condition C is true and 0 otherwise, and $\phi^p_{i_1 i_2}$ is the expected kinship for the pair i_1, i_2 based on the known pedigree alone. An estimate of ϕ^f is then obtained for every pair i_1, i_2 as

$$\hat{\phi}^f_{i_1,i_2} = \frac{(\hat{\phi}_{i_1 i_2} - \phi^p_{i_1 i_2})}{\kappa_{i_1 i_2}} \tag{B2}$$

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

C. Computation of P_U

In general the probability of F_j^U, the event that founder j is the only one to introduce the RV into the family, is

$$P[F_j^U] = P[F_j] - \sum_{k \neq j} P[F_j, F_k] \tag{C1}$$

We can obtain P_U by developing equation (C1) 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.

$$\begin{aligned} P_U &= \sum_{a=2n_f-d}^{2n_f} P[A=a] \left(P_a - \sum_{k \neq j} R_a \right) \\ &= \sum_{a=2n_f-d}^{2n_f} P[A=a] \left(\frac{2}{a} - (n_f - 1) R_a \right) \end{aligned}$$

To solve for R_a, we use a basic result from probability theory:

$$\begin{aligned}
1 &= P[F_1 \cup \cdots \cup F_{n_f}] \\
&= \sum_{j=1}^{n_f} P[F_j] - \sum_{j=1}^{n_f} \sum_{k=j+1}^{n_f} P[F_j, F_k] \\
&= \sum_a P[A=a] \left(\sum_{j=1}^{n_f} P[F_j|A=a] - \sum_{j=1}^{n_f} \sum_{k=j+1}^{n_f} P[F_j, F_k|A=a] \right) \\
&= \sum_a P[A=a] \left(n_f P_a - \frac{1}{2} n_f (n_f - 1) R_a \right) \qquad \text{(C2)}
\end{aligned}$$

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 (C2). We obtain

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

and

$$P_U = \sum_a P[A=a] \left(\frac{2}{n_f} - \frac{2}{a} \right). \qquad \text{(C3)}$$

The expected kinship coefficient among the n_f founders with respect to the distribution (2) of A is

$$E[\Phi] = \frac{\sum_{a=2n_f-d}^{2n_f-1} \frac{1}{(2n_f-a)!} \theta^{(2n_f-a)} \bar{\phi}_a}{\sum_{a=2n_f-d}^{2n_f} \frac{1}{(2n_f-a)!} \theta^{(2n_f-a)}} \qquad \text{(C4)}$$

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 have that:

$$\begin{aligned}
\bar{\phi}_a &= P[\text{Alleles from two founders are IBD}|\text{One of the two shares allele IBD with 2 other founders}] \\
&\quad \times P[\text{One of the two founders shares allele IBD with 2 other founders}] \\
&\quad + P[\text{Alleles from two founders are IBD}|\text{One of the two shares allele IBD with 1 other founder}] \\
&\quad \times P[\text{One of the two founders shares allele IBD with 1 other founder}] \\
&= \frac{1}{2(n_f-1)}\frac{2n_f-a}{n_f}\frac{2n_f-a-1}{n_f-1} + \\
&\quad \frac{1}{4(n_f-1)}\left[\frac{(2n_f-a)(a-n_f)}{n_f(n_f-1)} + \frac{2(2n_f-a)(a-n_f)}{n_f(2n_f-1)}\right] \qquad \text{(C5)}
\end{aligned}$$

Proof 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 by the two founders, that is

$$\begin{aligned}
&P[\text{alleles from two founders are IBD}|\text{one of the founders shares an allele IBD with } m \text{ other founders}] \\
&= \frac{m}{4(n_f-1)}
\end{aligned}$$

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:

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

∎

Equating $E[\Phi] = \hat{\phi}^f$, we solve the polynomial equation (C4) for θ. The value of d required to obtain a good approximation depends on the value of $\hat{\phi}^f$. When we need to allow fewer than $2n_f - 5$ distinct alleles to obtain a real positive root of the polynome in θ, we obtain a poor approximation, since the probability that any one of the distinct alleles (including the RV of interest) is present more than twice becomes non-negligible (see Results section). This is why we propose setting $d = 5$. When $\hat{\phi}^f$ is small, the approximation is almost identical with values of d from 2 or 3 to 5. When $d = 2$, we have the explicit solution:

$$\hat{\theta} = \frac{-(\hat{\phi}^f - \bar{\phi}_{2n_f-1}) - \sqrt{(\hat{\phi}^f - \bar{\phi}_{2n_f-1})^2 - 2(\hat{\phi}^f - \bar{\phi}_{2n_f-2})\hat{\phi}^f}}{\hat{\phi}^f - \bar{\phi}_{2n_f-2}} \tag{C6}$$

D. 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 three events when no marriage loops are present.

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

This implies that the two founders i and j 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

$$
\begin{aligned}
&P[C_1 = \cdots = C_n = 1 | F_j, F_k] \\
&= \sum_{x=0}^{n} P[C_1 = \cdots = C_n = 1 | \sharp\{i : T_i = 2\} = x, \sharp\{i : T_i = 1\} = n - x, F_j, F_k] \\
&\quad \times P[\sharp\{i : T_i = 2\} = x, \sharp\{i : T_i = 1\} = n - x | F_j, F_k] \\
&= \sum_{x=0}^{n} \left(\frac{1}{2}\right)^{\sum_{\{i:T_i=2\}} D_{ij} - 2} \left(\frac{1}{2}\right)^{\sum_{\{i:T_i=1\}} D_{ij} - 1} \binom{n}{x} \left(\frac{1}{4}\right)^{x} \left(\frac{1}{2}\right)^{n-x} \\
&= \sum_{x=0}^{n} \left(\frac{1}{2}\right)^{D^s - n - x} \binom{n}{x} \left(\frac{1}{2}\right)^{2x} \left(\frac{1}{2}\right)^{n-x} = \left(\frac{1}{2}\right)^{D^s} \sum_{x=0}^{n} \binom{n}{x} \\
&= \left(\frac{1}{2}\right)^{D^s - n} \qquad \text{(D1)}
\end{aligned}
$$

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 grandchildren or more distant descendants of the founders. When a sequenced subject is a child 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.

$$
\begin{aligned}
&P[C_1 \geq 1, \ldots, C_{n_c} \geq 1, C_{n_c+1} = \cdots = C_n = 1 | F_j, F_k] \\
&= P[C_1 \geq 1, \ldots, C_{n_c} \geq 1 | F_j, F_k] \times P[C_{n_c+1} = \cdots = C_n = 1 | F_j, F_k] \\
&= \left(\frac{3}{4}\right)^{n_c} \left(\frac{1}{2}\right)^{(D^s - n_c) - (n - n_c)} \\
&= \left(\frac{3}{4}\right)^{n_c} \left(\frac{1}{2}\right)^{D^s - n} \qquad \text{(D2)}
\end{aligned}
$$

The expression for the probability of not seeing the variant in any sequenced individual when all

$D_{ij} \geq 2$ is:

$$\begin{aligned}
&P[C_1 = \cdots = C_n = 0 | F_j, F_k] \\
&\quad= \sum_{x=0}^{n} \sum_{y=0}^{n-x} P[C_1 = \cdots = C_n = 0 | \sharp\{i : T_i = 2\} = x, \sharp\{i : T_i = 1\} = y, F_j, F_k] \\
&\qquad \times P[\sharp\{i : T_i = 2\} = x, \sharp\{i : T_i = 1\} = y | F_j, F_k] \\
&\quad= \sum_{x=0}^{n} \sum_{y=0}^{n-x} \prod_{\{i:T_i=2\}} \left(1 - \left(\frac{1}{2}\right)^{D_{ij}-2}\right) \prod_{\{i:T_i=1\}} \left(1 - \left(\frac{1}{2}\right)^{D_{ij}-1}\right) \\
&\qquad \times \binom{n}{x, y, n-x-y} \left(\frac{1}{4}\right)^{x} \left(\frac{1}{2}\right)^{y} \left(\frac{1}{4}\right)^{n-x-y} \qquad \text{(D3)}
\end{aligned}$$

without obvious simplification. The modification for sequenced subjects who are children of the founders is similar to that for the joint sharing probability D2, with probability equal to $\frac{1}{4}$ of not receiving the variant instead of $\frac{3}{4}$ of receiving it.

D.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.

$$\begin{aligned}
&P[C_1 = \cdots = C_n = 1 | F_j, F_k] \\
&= P[C_1 = \cdots = C_n = 1 | T = 2, F_j, F_k] P[T = 2 | F_j, F_k] \\
&\quad + P[C_1 = \cdots = C_n = 1 | T = 1, F_j, F_k] P[T = 1 | F_j, F_k] \\
&= \left(\frac{1}{2}\right)^{D_{1h} - 1 + \sum_{i=2}^{n} D_{ij}} \left(\frac{1}{2}\right)^{D_{hj}} \frac{1}{2} \\
&\quad + \left(\frac{1}{2}\right)^{D_{1h} + \sum_{i=2}^{n} D_{ij}} \left[\left(\frac{1}{2}\right)^{D_{hj}} \frac{1}{2} + \left(1 - \left(\frac{1}{2}\right)^{D_{hj}}\right) \frac{1}{2}\right] \\
&= \left(\frac{1}{2}\right)^{D_{1h} + \sum_{i=2}^{n} D_{ij}} \left[\left(\frac{1}{2}\right)^{D_{hj}} + \frac{1}{2}\right] \qquad \text{(D4)}
\end{aligned}$$

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

$$\begin{aligned}
&P[C_1 \geq 1, C_2 = \cdots = C_n = 1 | F_j, F_k] \\
&= P[C_1 = 2, C_2 = \cdots = C_n = 1 | F_j, F_k] + P[C_1 = \cdots = C_n = 1 | F_j, F_k] \\
&= \left(\frac{1}{2}\right)^{D^s} \frac{1}{2} + \left(\frac{1}{2}\right)^{\sum_{i=2}^{n} D_{ij}} \frac{1}{2} \\
&= \left(\frac{1}{2}\right)^{D^s + 1} \left[1 + 2^{D_{1j}}\right] \qquad \text{(D5)}
\end{aligned}$$

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

$D_{ih} \geq 1$ is:

$$\begin{aligned}
P[C_1 = \cdots = C_n = 0|F_j, F_k] &= \prod_{i=1}^{n} P[C_i = 0|F_j, F_k] \\
&= \{P[C_1 = 0|T = 2, F_j, F_k] \times P[T = 2|F_j, F_k] \\
&\quad + P[C_1 = 0|T = 1, F_j, F_k] \times P[T = 1|F_j, F_k] \\
&\quad + P[C_1 = 0|T = 0, F_j, F_k] \times P[T = 0|F_j, F_k]\} \\
&\quad \times \prod_{i=2}^{n} P[C_i = 0|F_j] \\
&= \left\{\left(1 - \left(\frac{1}{2}\right)^{D_{1h}-1}\right)\left(\frac{1}{2}\right)^{D_{hj}} \frac{1}{2}\right. \\
&\quad \left. + \left(1 - \left(\frac{1}{2}\right)^{D_{1h}}\right)\frac{1}{2} + \left(1 - \left(\frac{1}{2}\right)^{D_{hj}}\right)\frac{1}{2}\right\} \\
&\quad \times \prod_{i=2}^{n}\left(1 - \left(\frac{1}{2}\right)^{D_{ij}}\right) \qquad \text{(D6)}
\end{aligned}$$

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

$$P[C_1 = \cdots = C_n = 0|F_j, F_k] = \prod_{i=1}^{n} P[C_i = 0|F_j, F_k] = \frac{1}{2}\prod_{i=1}^{n}\left(1 - \left(\frac{1}{2}\right)^{D_{ij}}\right) \qquad \text{(D7)}$$

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

We assume that founder j is an ancestor of subject 1 and founder k is an ancestor of subject 2. The formula applies equally to the case where both j and k are ancestors of the same subject, as long as one is ancestor of that subject's 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 $P[C_1 = \cdots = C_n = 1|F_j, F_k] = 0$. If $n = 2$, then

$$P[C_1 = C_2 = 1|F_j, F_k] = P[C_1 = 1|F_j]P[C_2 = 1|F_k] = \left(\frac{1}{2}\right)^{D_{1j}+D_{2k}} \qquad \text{(D8)}$$

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

$$\begin{aligned} P[C_1 = \cdots = C_n = 0|F_j, F_k] &= P[C_1 = 0|F_j]P[C_2 = 0|F_k] \\ &= \left(1 - \left(\frac{1}{2}\right)^{D_{1j}}\right)\left(1 - \left(\frac{1}{2}\right)^{D_{2k}}\right) \qquad \text{(D9)} \end{aligned}$$

D.4. Extension to branching in the pedigree

Having two founders introducing the variant requires adaptation of the formulas of the Methods section. Equation (A1) becomes

$$\begin{aligned} & P[C_1 = \cdots = C_n = 1] \\ = \quad & P[C_1 = \cdots = C_{i_k} = 1|B_k = 1] \times P[B_k = C_{i_k+1} = \cdots = C_n = 1] \\ & +P[C_1 = \cdots = C_{i_k} = 1|B_k = 0] \times P[B_k = 0, C_{i_k+1} = \cdots = C_n = 1] \qquad \text{(D10)} \end{aligned}$$

The term $P[C_1 = \cdots = C_{i_k} = 1|B_k = 1]$ is not directly computable, and we instead compute terms $P[C_1 = \cdots = C_{i_k} = 1|F_j, B_k = 1]$ for every founder j below the branching subject k in the pedigree (in the sense defined in chapter 4 of Thompson (1986)), which can be done using equations (D1), (D2), (D4) or (D5), depending on the relationship between j and k. These terms can then be summed over all founders j below k, with equal weight when a common kinship coefficient is estimated for all founders, or weighted by $P[F_j|B_k = 1]$ computed from the pair-specific kinship coefficient between founders. The term $P[C_1 = \cdots = C_{i_k} = 1|B_k = 0]$ is computed by applying equation (2.1) to every founders j below k. The other terms are computed by reapplying equation (D10) recursively with the other branching individuals, with slight modification for the terms where $B_k = 0$ instead of 1.

In equation (A2), the term $P[C_1 = \cdots = C_{i_k} = 0|B_k = 1, F_j]$ can no longer be computed from the right-hand side of equation (1) when j is a founder below k, but can be computed using equations (D3), (D6) or (D7). Similarly, the term $P[C_1 = \cdots = C_{i_k} = 0|B_k = 0, F_j]$ no longer equals 1 but equals $P[C_1 = \cdots = C_{i_k} = 0|F_j]$ which can be computed using equation (1). 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:

$$P[C_1 = \cdots = C_{i_k} = 0|B_k = 2] = \prod_{i=1}^{i_k} \left(1 - \left(\frac{1}{2}\right)^{D_{ik}-1}\right) \tag{D11}$$

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

$$\begin{aligned} &P[B_k = 2, C_{i_k+1} = \cdots = C_n = 0|F_h, F_j] \\ = \ &\left(\frac{1}{2}\right)^{D_{kh}+D_{kj}} P[C_{i_k+1} = \cdots = C_n = 0|F_h, F_j] \end{aligned} \tag{D12}$$

$$\begin{aligned} &P[B_k = 1, C_{i_k+1} = \cdots = C_n = 0|F_h, F_j] \\ = \ &\left[\left(\frac{1}{2}\right)^{D_{kh}} \left(1 - \left(\frac{1}{2}\right)^{D_{kj}}\right) + \left(\frac{1}{2}\right)^{D_{kj}} \left(1 - \left(\frac{1}{2}\right)^{D_{kh}}\right)\right] \\ &\times P[C_{i_k+1} = \cdots = C_n = 0|F_h, F_j] \end{aligned} \tag{D13}$$

$$\begin{aligned} &P[B_k = 0, C_{i_k+1} = \cdots = C_n = 0|F_h, F_j] \\ = \ &\left(1 - \left(\frac{1}{2}\right)^{D_{kh}+D_{kj}}\right) P[C_{i_k+1} = \cdots = C_n = 0|F_h, F_j] \end{aligned} \tag{D14}$$

With other intervening branching individuals a recursion similar to equation (A3) would be needed.

E. Approximating sharing probabilities in presence of inbreeding

The inbred pedigree is trimmed to obtain a reduced pedigree without inbreeding loops, and the term "founder" refers from now on to the founders of this reduced pedigree. We assume that only one founder allele (not necessarily the RV considered in the computation) can be shared by one pair of founders and the event "j and k share an allele IBD" means they are the only ones to do so. This assumption is always satisfied when only two founders are related and then this method gives an exact sharing probability. The method however allows all pairs of founders to be related, and can still give a good approximation when the kinship coefficient between a few of the founders are non-zero.

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

$$\begin{aligned} P[F_j, F_k] &= P[\text{Allele shared is RV}|j\&k \text{ share allele IBD}]P[j\&k \text{ share allele IBD}] \\ &= \frac{1}{2n_f - 1} 2\phi_{jk} = \frac{2\phi_{jk}}{2n_f - 1} \end{aligned} \tag{E1}$$

where ϕ_{jk} is the kinship coefficient between founders j and k. The first term represents the probability the RV is the allele IBD between the two founders among $2n_f - 1$ distinct alleles in all founders. If we know which founders j and k are related, then their degree of relatedness is usually also known, and this specifies their kinship coefficient ϕ_{jk}. If a subset of the founders are suspected to be related (with other founders unrelated to that subset and among themselves) then this method can still be applied, with the kinship coefficient among the subset of founders suspected to be related estimated as described in the Methods section.

The marginal probability that any founder h introduces the RV needs to be adjusted compared to the unrelated case.

$$\begin{aligned}
P[F_h] &= \sum_j \sum_{k>j} P[F_h | j\&k \text{ share allele IBD}] P[j\&k \text{ share allele IBD}] \\
&\quad + P[F_h|\text{no founder pair shares allele IBD}] P[\text{no founder pair shares allele IBD}] \\
&= \frac{2}{2n_f - 1} \sum_j \sum_{k>j} P[j\&k \text{ share allele IBD}] + \frac{1}{n_f} \left(1 - \sum_j \sum_{k>j} P[j\&k \text{ share allele IBD}] \right) \\
&= \frac{4 \sum_j \sum_{k>j} \phi_{jk}}{2n_f - 1} + \frac{1}{n_f} \left(1 - 2 \sum_j \sum_{k>j} \phi_{jk} \right) \qquad \text{(E2)}
\end{aligned}$$

We obtain the probability of F_j^U, the event that founder j is the only one to introduce the RV into the family, from equation (C1). Once the required elements have been computed, we get an adjusted estimate of sharing probability from the following formula:

$$P[\text{RV shared}] = \qquad \text{(E3)}$$

$$\frac{\sum_j P[C_1 = \cdots = C_n = 1 | F_j^U] P[F_j^U] + \sum_j \sum_{k>j} P[C_1 = \cdots = C_n = 1 | F_j, F_k] P[F_j, F_k]}{\sum_j P[C_1 + \cdots + C_n \geq 1 | F_j^U] P[F_j^U] + \sum_j \sum_{k>j} P[C_1 + \cdots + C_n \geq 1 | F_j, F_k] P[F_j, F_k]}$$

F. Simulation of small populations

The entire pedigree of small populations was simulated over 6 generations using the computer package Spip (Anderson and Dunham 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 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. Kinship coefficients were computed using the R package `kinship2`. The distribution of kinship coefficients between subjects from the same generation had converged to its equilibrium distribution around the fifth generation (data not shown).

In each simulated population, $2n_p$ distinct alleles were assigned to the n_p population founders ancestral to any of the eight sampled pedigree founders and the transmission of these RVs to the pedigree founders was simulated 2000 times under Mendel's laws using the software package Simulate (Terwilliger *et al.* 1993).

G. Whole exome sequencing study of nonsyndromic oral clefts

G.1. Genotyping and DNA sequencing

Whole exome sequencing and genotyping was done at the Center for Inherited Disease Reseach (CIDR). Genomic DNA was isolated by the original research team, and aliquots of DNA were sent to the CIDR for sequencing. All affected subjects included in the sequencing study were genotyped using the Human OmniExpress SNP array from Illumina as a quality control step. Genotypes were called using Illumina's software package GenomeStudio (v. 2010.2, Genotyping Module version 1.7.4, GenTrain version 1.0). Six subjects were genotyped in duplicate (4 family members and 2 HapMap controls). Single nucleotide polymorphic (SNP) markers with call rate $< 98\%$, with cluster separation value < 0.2 or with discrepant genotypes in more than one duplicate pair were removed from subsequent analyses.

DNA fragmentation was performed on 200ng of genomic DNA using a Covaris E210 system, which shears DNA into fragments 150 to 200 bp in length with 3' or 5' overhangs. End repair was performed where 3' to 5' exonuclease activity of enzymes removes 3' overhangs, and the polymerase activity fills in the 5' overhangs. An A base is then added to the 3' end of the blunt phosphorylated DNA fragments to prepare DNA fragments for ligation to the sequencing adapters, which have a single T base overhang at their 3' end. Ligated fragments are subsequently size selected through purification using SPRI beads and undergo PCR amplification techniques to prepare the libraries. The Caliper LabChip GX was used for quality control of libraries to ensure adequate concentration and appropriate fragment size.

Exon capture was done using the Agilent SureSelect Human All Exon Target Enrichment system (Kit S0297201), which results in 51Mb of targeted sequence capture per sample. Under standard procedures, biotinylated RNA oligonucleotides were hybridized with 500ng of the library. Magnetic bead selection was used to capture the resulting RNA-DNA hybrids. RNA is digested and remaining DNA capture PCR-amplified. Sample indexing was introduced at this step. The Agilent Bioanalyzer (HiSensitivity) was used for quality control of adequate fragment sizing and quantity of DNA capture.

DNA sequencing was performed on an Illumina HiSeq 2500 instrument using standard protocols for a 100 bp paired-end run. Six samples were run in per flowcell, guaranteeing $> 90 - 95\%$ completeness at a minimum of 20X coverage. Variant Calling: Illumina HiSeq reads were processed through Illumina's Real-Time Analysis (RTA) software generating base calls and corresponding quality scores. Resulting data were aligned to a reference genome with the Burrows-Wheeler Alignment (BWA) tool creating a SAM/BAM file. Post processing of the aligned data includes local realignment around indels, base call quality score recalibration performed by the Genome Analysis Tool Kit (GATK) and flagging of molecular/optical duplicates using software from the Picard program suite. Multi-sample variant calling was performed using GATK2.0's Unified Genotyper. Variant Quality Score Recalibration (VQSR) was done in GATK2.0 and only variants passing this step were included. CIDR required a minimum mean of 8x coverage before calling any SNV, but the overall coverage averaged 84X over all exons.

G.2. Rare single nucleotide variant analysis

For the application of our proposed approach, we defined a rare SNV as an allele with frequency < 0.01 based on the sequence of 5,379 subjects in the Exome Sequencing Project (ESP, esp.gs.washington.edu/drupal/) database at the time of analysis, and a frequency

< 0.01 in the 1,092 subjects from the April 2012 release of the 1000 Genomes data (www.1000genomes.org). SNVs not seen in either of the above databases were retained if they were either absent or their frequency was ≤ 0.1 in an internal database of all exomes previously sequenced at the Center for Inherited Disease Research (Baltimore, MD), to increase confidence variant calls did not result from technical artifacts. Among SNVs passing the above filtering criteria, SNVs seen in more than 20 percent of the families were excluded.

The G allele of rs149253049 was called unambiguously in our study: among the reads covering that position in the sequenced subjects from these three families, the number with the G allele ranged from 16 out of 33 to 34 out of 69. The G nucleotide is the rarest of the three alleles of rs149253049. It was not found in the ESP database nor the 1000 Genomes project data. It was seen once in the 662 participants of European descent from the ClinSeq project (0.001 frequency reported in www.ncbi.nlm.nih.gov/snp/). The subjects in the Indian oral cleft families are from the Bengali population which shares partially its ancestry with the Gujarati Indian from Houston, Texas (GIH) and the Indian Telugu from the UK (ITU) included in the 1000 Genome project. Nonetheless, the G allele could be present in the Indian Bengali population at a small but appreciable frequency, which could violate the underlying assumption of IBS without IBD being negligible, and thus, render the reported sharing probability too optimistic. We carried out a sensitivity analysis by calculating the true sharing probabilities as a function of allele frequency, and found that as long as the true allele frequency in the Indian population is below 1.1 percent our finding retains statistical significance after multiple comparison correction (Supplementary Figure S4). In our Indian sample, kinship estimates between affected subjects from genome-wide SNP genotypes were based on the estimator of Manichaikul *et al.* (2010) robust to population stratification. There was no evidence of excess IBD sharing given the known degree of relatedness, nor of relatedness between subjects from distinct Indian families, and using equation (B2) on all 18 pairs of sequenced subjects we obtained an estimated mean kinship of the founders $\hat{\phi}^f = -0.006$. All this suggested sharing probabilities for rs149253049 computed based on known pedigree structures are accurate and these families are unrelated.

By contrast, suspected relationships among founders and a higher variant frequency cast doubts on the significance of the sharing observed at rs117883393. The T allele 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 sharing probabilities may be underestimated in two of the families where this SNV is shared because these families are from the Syrian sample, where cultural and demographic factors make relationships between 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 relatives (results not shown). We then inferred ϕ^f using equation (B2) on all 13 pairs of sequenced subjects and obtained $\hat{\phi}^f = 0.013$, close to the kinship coefficient of second cousins ($\frac{1}{64}$). The estimates of kinship between subjects from distinct Syrian and German families were close to 0 or slightly negative, indicating no evidence of relatedness across families (not shown).

For the family shown on Figure 1A, the RV sharing probability obtained using a recursive computation of the terms of equation (1) was 0.0028. The probability $1 - w$ that the rare T allele at SNV rs117883393 was introduced by two founders equaled 0.092 using $\hat{\phi}^f = 0.013$, leading to an adjusted RV sharing probability of 0.0044. For the family shown on Figure 1B, the RV sharing probability obtained from equation (2.1) was 0.011. The probability that this 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 these two Syrian families, and assuming no unknown relationships in the two other families of German origin, the p-value for all four families increased to 1.3×10^{-5}. An additional analysis of sensitivity to the assumption of no IBS without IBD yielded p-values above the multiple comparison corrected significance threshold of 2.1×10^{-5} even for population allele frequencies much below 1 percent in the Syrians (see Supplementary Figure S5), rendering the significance of rs117883393 in *OR2A2* somewhat doubtful.

We compared these results to those of a standard variant filtering strategy. We retained nonsynonymous or truncating SNVs not found in build 137 of dbSNP, and predicted to be damaging based on a SIFT score < 0.05. We found 10,589 novel SNVs predicted to be damaging in the entire exome. Of that number, 656 were shared by all sequenced relatives in at least one family (only 7 were shared in two families, all others were shared in only one family). Using rare variant sharing as a filter as proposed by Feng *et al.* (2011) therefore yields a large number of variants to follow up.

REFERENCES

Anderson, E. C. and Dunham, K. K. (2005). Spip 1.0: a program for simulating pedigrees and genetic data in age-structured populations. *Molecular Ecology Notes*, **5**, 459–61.

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

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

Terwilliger, J., Speer, M., and Ott, J. (1993). Chromosome-based method for rapid computer simulation in human genetic linkage analysis. *Genet Epidemiol*, **10**(4), 217–24.

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

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

Table S1: 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[a]) of mean ϕ^{f}[b]	0.043 (0.004)	0.0216 (0.0015)		0.0108 (0.0006)	
mean (SD) of P[RV sharing][c]	0.007 (0.003)	0.0039 (0.0023)		0.0022 (0.0007)	
N founders with RV	Simulated mean (SD)	Simulated mean (SD)	Approx mean (SD)	Simulated mean (SD)	Approx mean (SD)
1	0.56 (0.07)	0.72 (0.07)	0.80 (0.02)	0.86 (0.05)	0.91 (0.01)
2	0.29 (0.03)	0.22 (0.04)	0.20 (0.02)	0.13 (0.04)	0.09 (0.01)
3+	0.15 (0.05)	0.06 (0.04)	0	0.01 (0.01)	0

[a] SD: standard deviation

[b] ϕ^{f}: kinship coefficient among founders

[c] P[RV sharing]: probability of sharing of a rare variant by the three second cousins

Table S2: Sharing probabilities for rs149253049

Relationship between affecteds	Degree	Sharing probability
first cousins	3	0.0667
third cousins	7	0.0039
second cousins once removed	6	0.0079
Product		2.0×10^{-6}

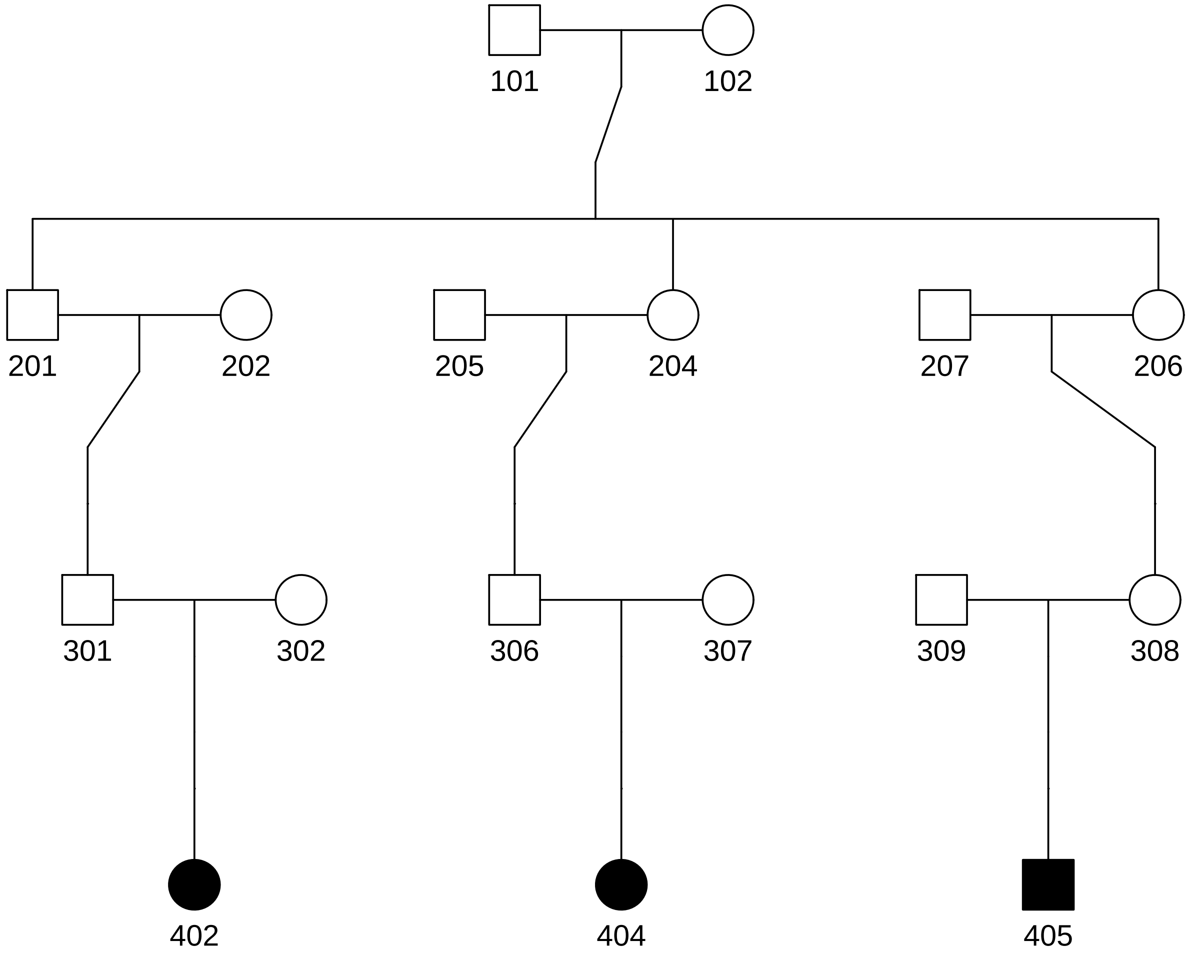
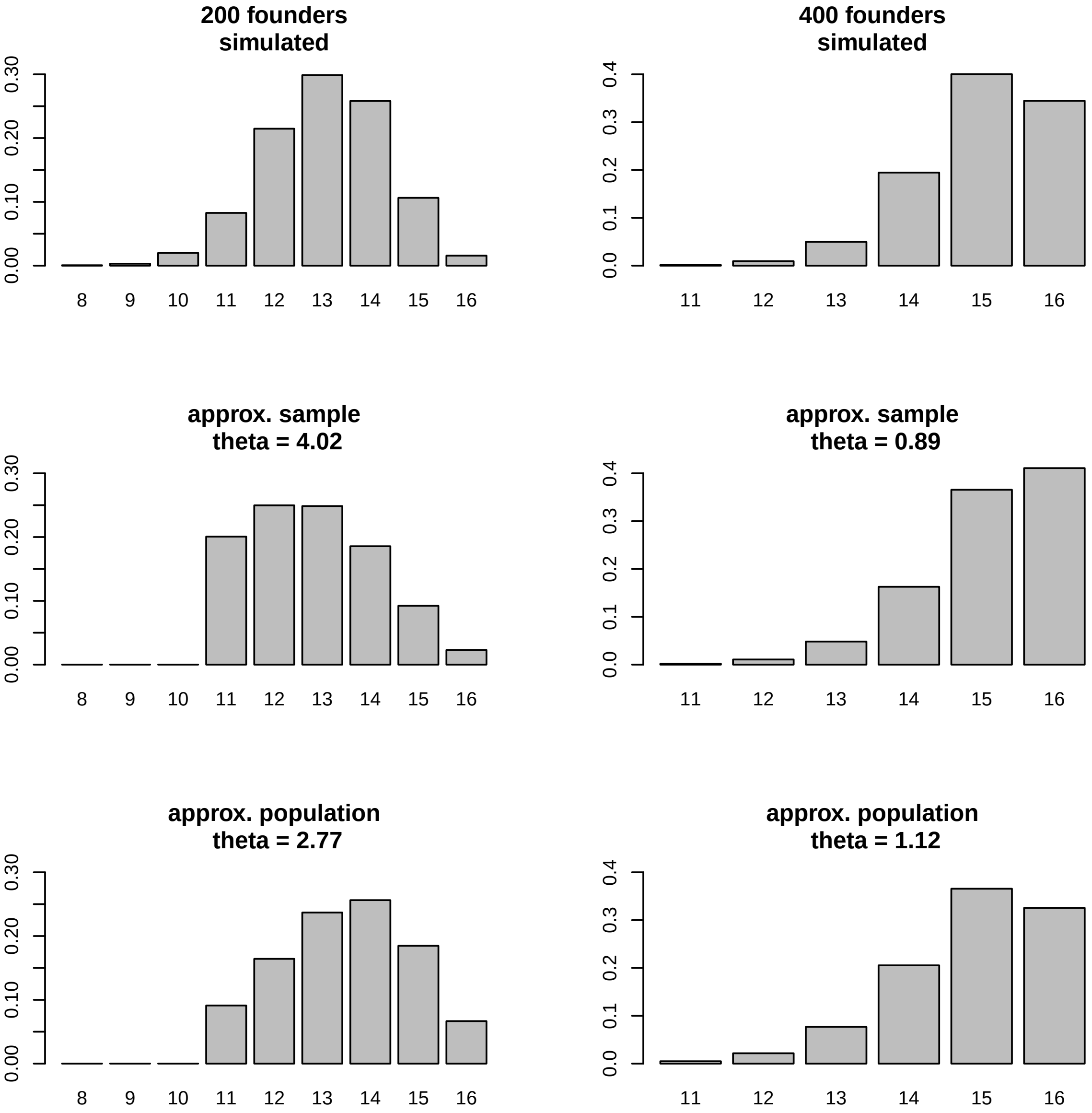


Fig. S1.— Pedigree of three second cousins used in simulation study. Filled symbols represent affected members who have been sequenced.



200 founders simulated

400 founders simulated

approx. sample theta = 4.02

approx. sample theta = 0.89

approx. population theta = 2.77

approx. population theta = 1.12

Fig. S2.— Number of distinct alleles in a sample of eight subjects from small populations. The approximation of the distribution of the number of distinct alleles was obtained using either the mean kinship coefficient among the eight sampled founders (approx. sample) or the mean kinship coefficient in the population (approx. population)

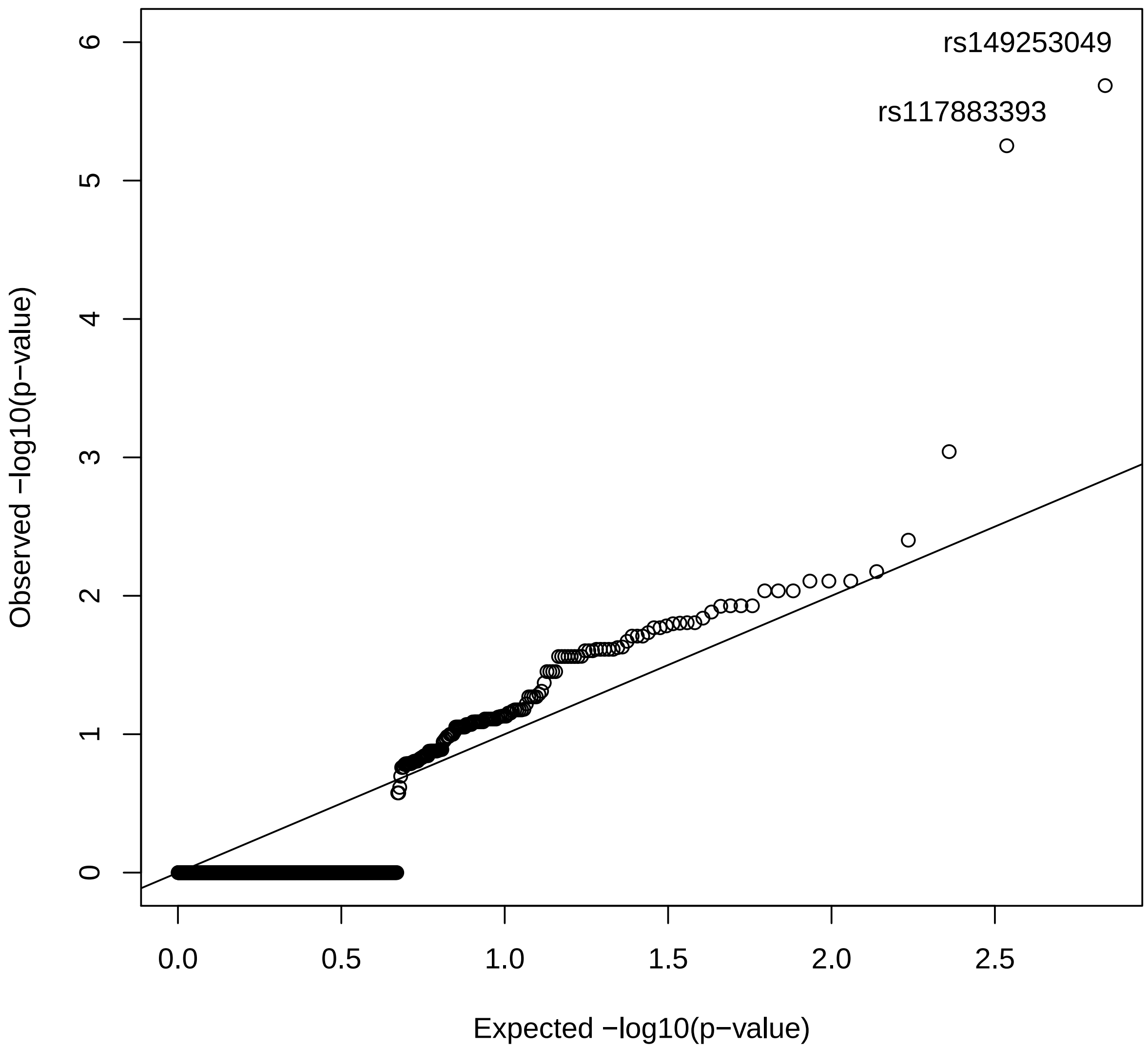


Fig. S3.— Observed versus expected distribution of $-\log_{10}$ p-values of the SNVs from the oral clefts exome sequencing study with a potential p-value $< 2.1 \times 10^{-5}$.

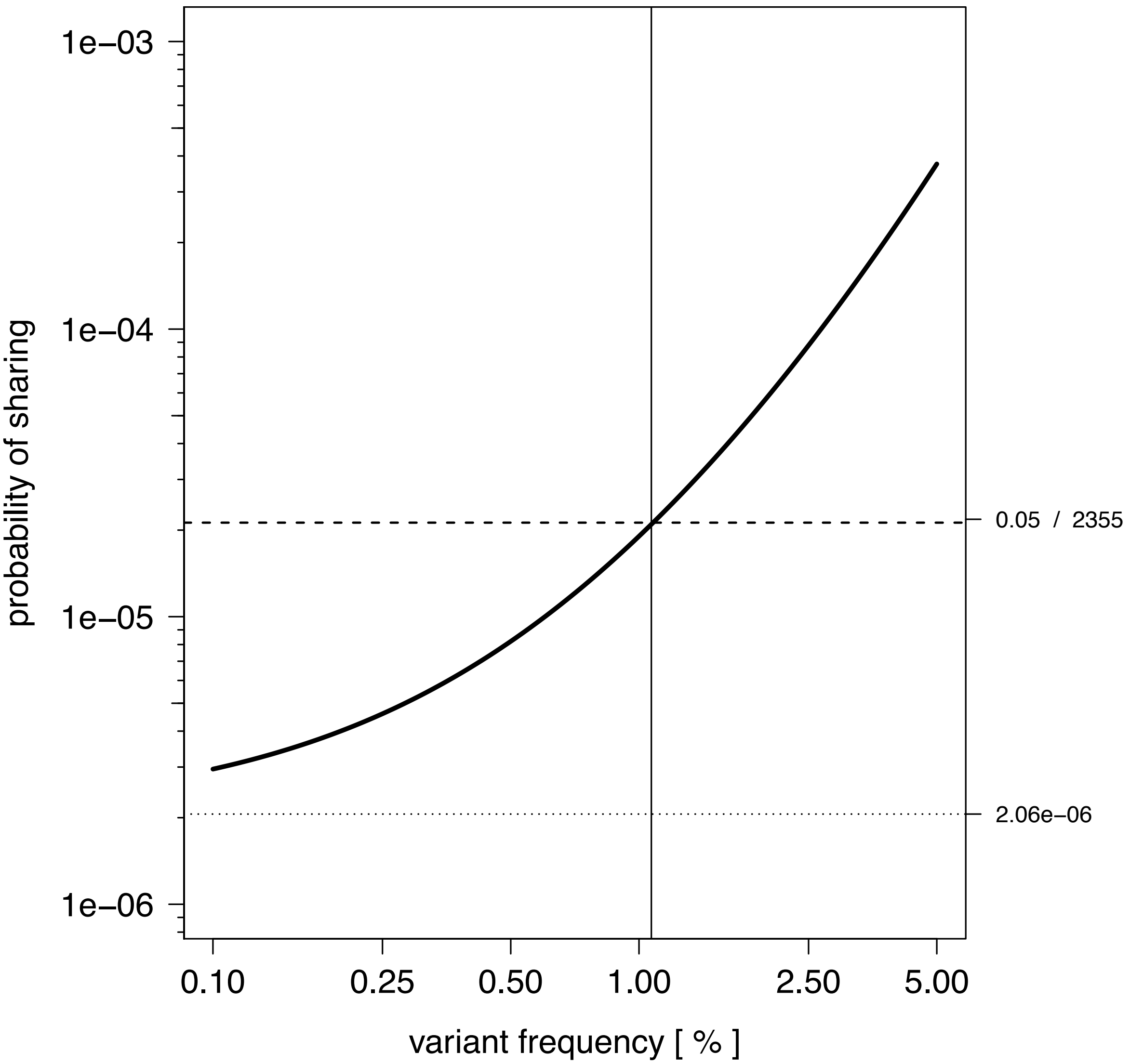


Fig. S4.— P-values based on IBS sharing probabilities for rs149253049 in *ADAMTS9*. Derived p-values (y-axis, log-scale) for the G allele of rs149253049 in *ADAMTS9* calculated using conditional probabilities and Mendel's laws, as a function of variant allele frequency (x-axis), based on the three Indian familiy pedigrees (Table 3). The sharing probabilities calculated under the assumption of no IBS without IBD is $\frac{1}{15\times127\times255} = 2.06 \times 10^{-6}$, indicated by the dotted horizontal line. The multiple comparisons corrected significance threshold is $\frac{0.05}{2{,}355} = 2.1\times10^{-5}$ indicated by the dashed horizontal lines.

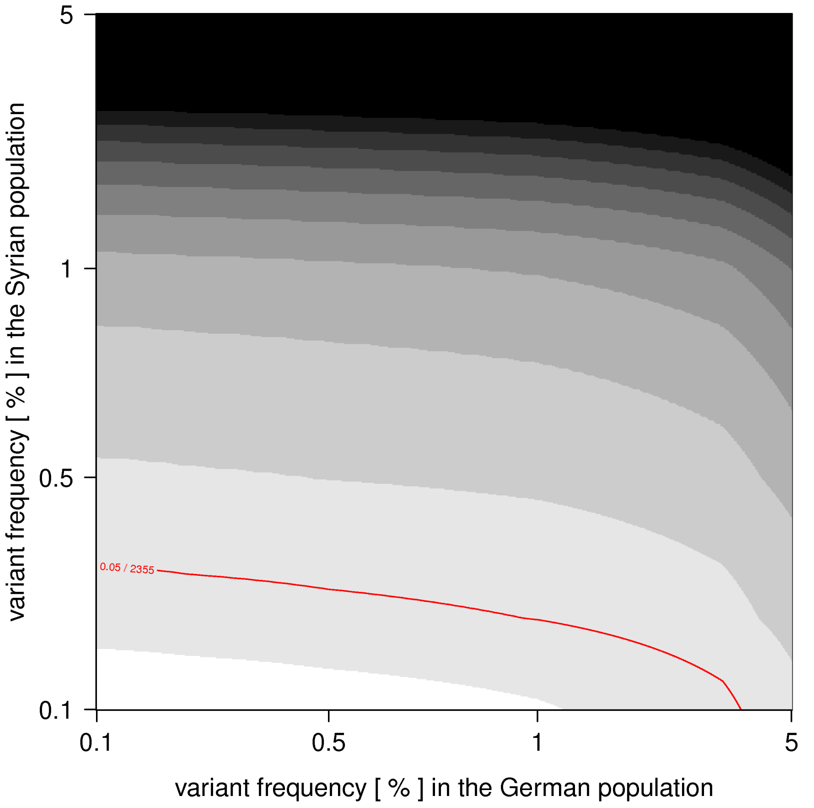


Fig. S5.— P-values based on IBS sharing probabilities for rs117883393 in *OR2A2*. Derived p-values (z-axis, darker shades indicating higher p-values) for the T allele of rs117883393 in *OR2A2* as a function of assumed true variant allele frequencies in the German (x-axis) and Syrian (y-axis) populations. The p-values are based on the calculated sharing probabilities of the two Syrian families and the two German families that contained this variant allele. The cryptic founder relatedness observed in the Syrian population was taken into account by approximating the sharing probabilities using the same inflation factor derived in the analysis of the Syrian families under the assumption of no IBS without IBD. A Monte Carlo simulation combining an allele frequency of 0.01 and an adjustment for unknown relationships as decribed in the Methods indicated this approximation was good at that allele frequency (results not shown). The multiple comparison corrected significance threshold ($\frac{0.05}{2{,}355} = 2.1 \times 10^{-5}$) is shown as a red line. Due to pedigree structures, sharing probabilities (and thus, p-values) are much more sensitive to population allele frequencies in the Syrians. Even allele frequencies as low as 0.5% in the Syrians would render this finding non-significant.