Sharing of Rare Nucleotide and Copy Number Variants in Extended Multiplex Families

Ingo Ruczinski

Johns Hopkins Bloomberg School of Public Health

Joint work with Alexandre Bureau, Jack Fu, Rob Scharpf, and Tom Sherman

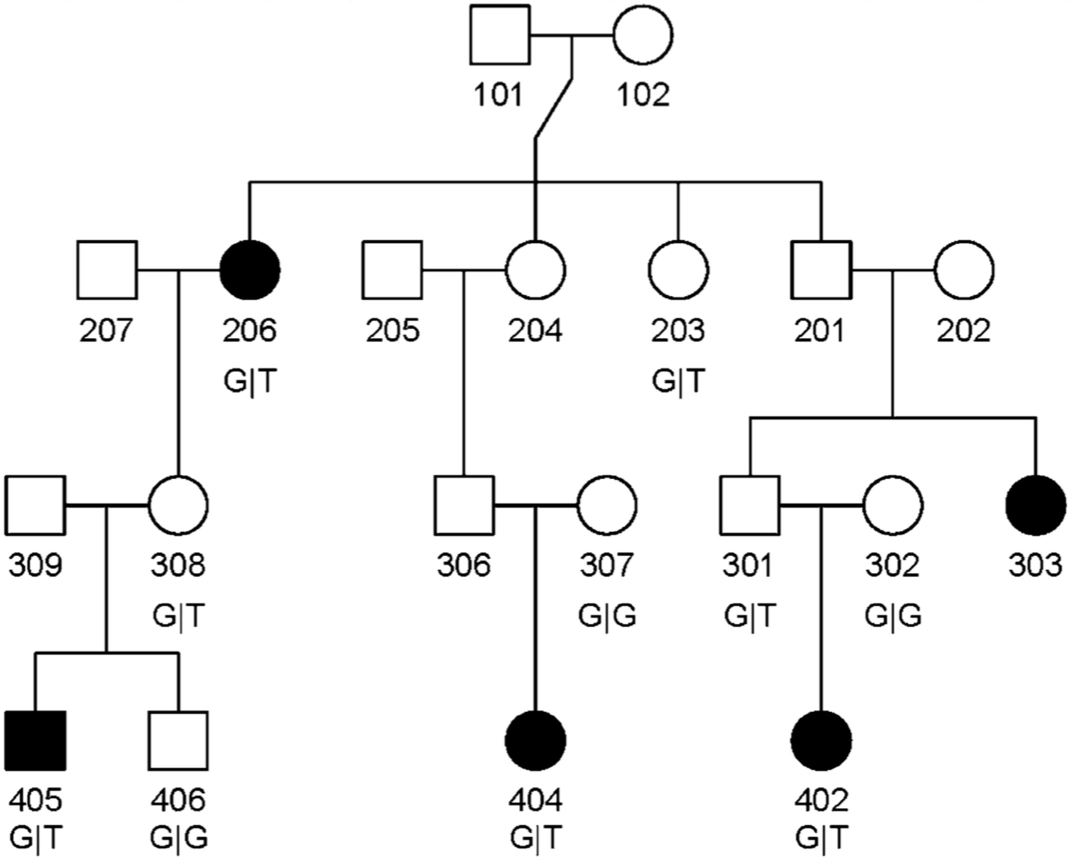
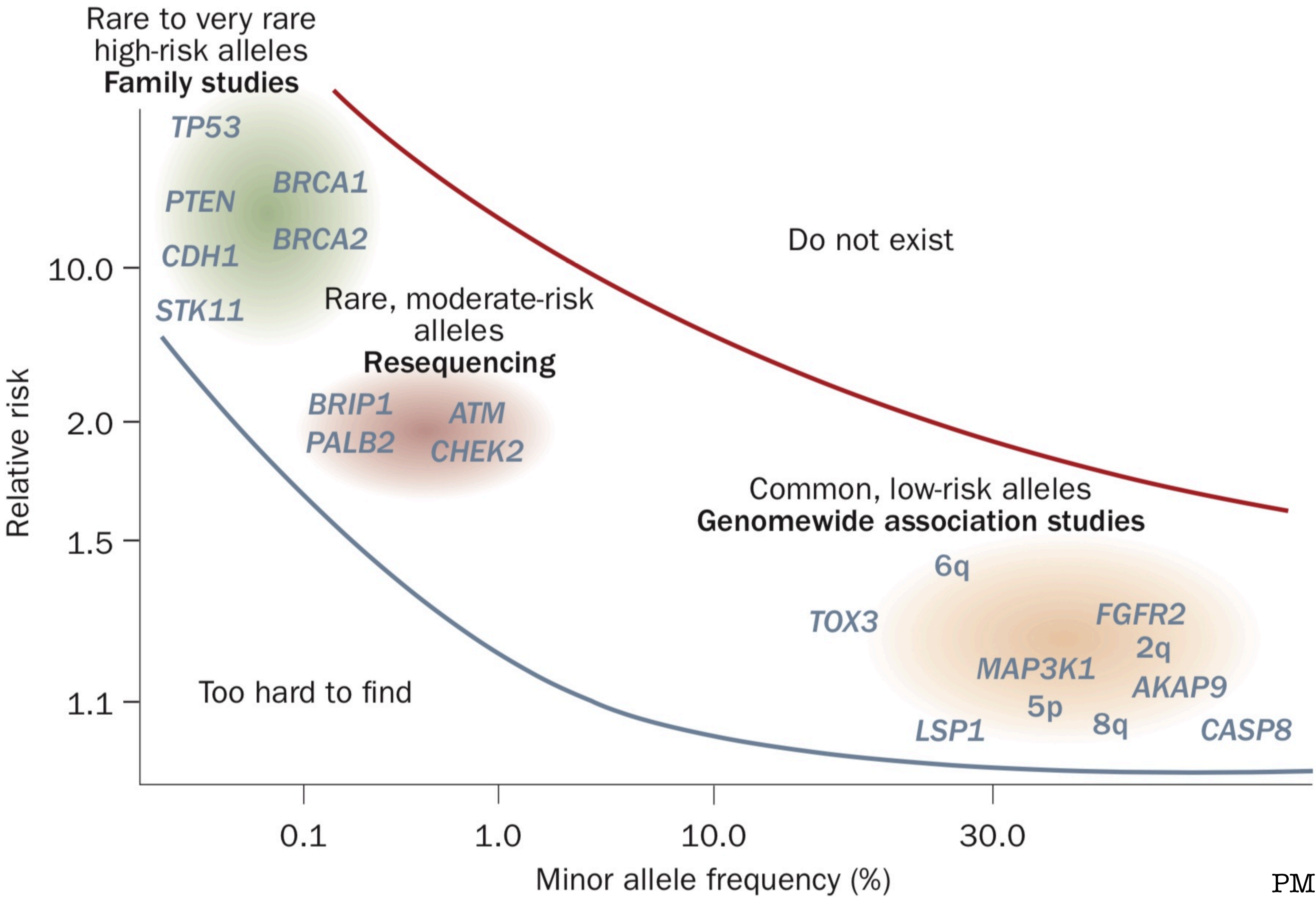
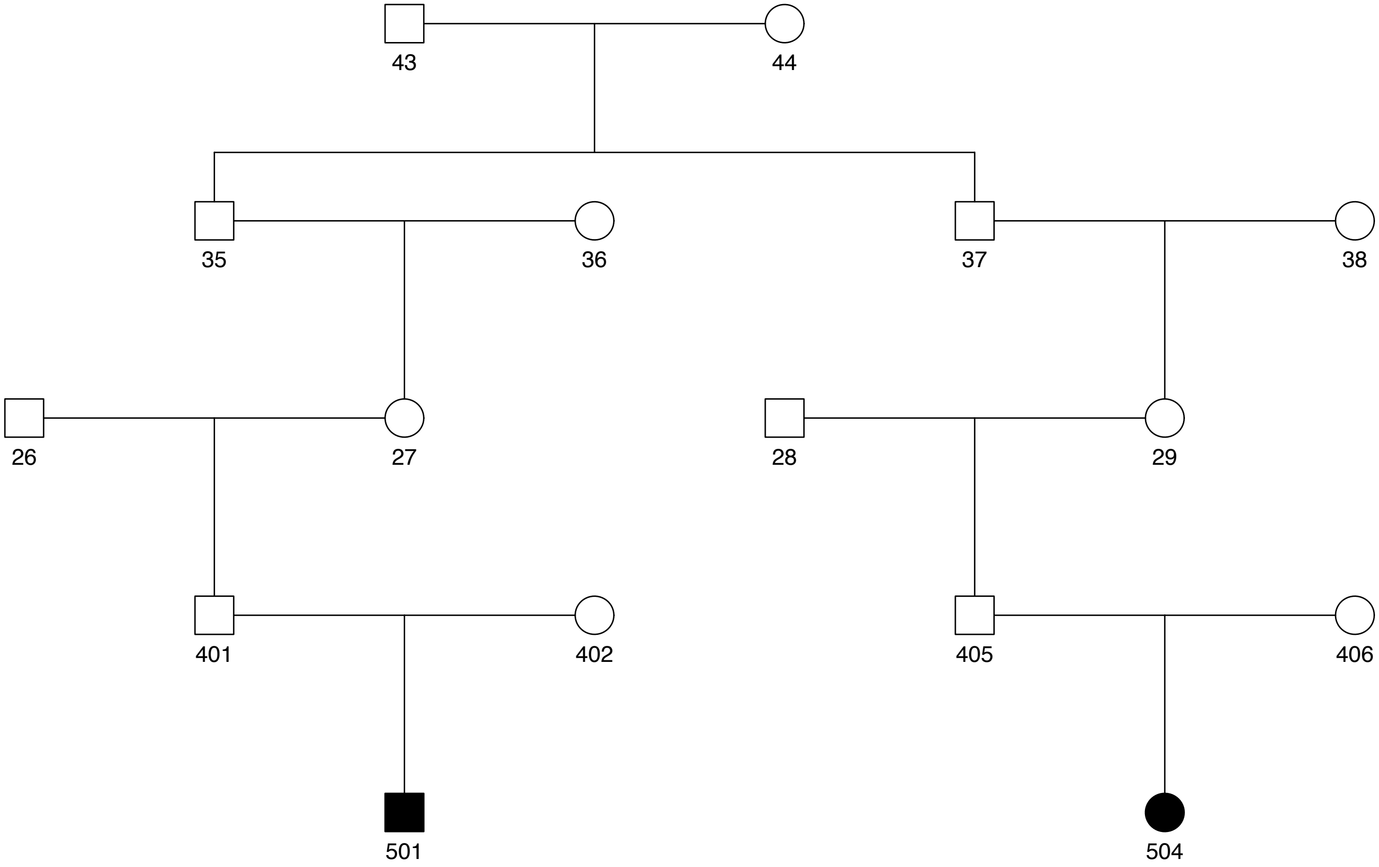


Figure 1 Structure of pedigree where three affected second cousins shared a rare variant in *CDH1*. Affected subjects are represented by filled symbols. Individuals 402, 404, and 405 were sequenced.

PMID 24793288



PMID 20351699


43
44
35
36
37
38
26
27
28
29
401
402
405
406
501
504

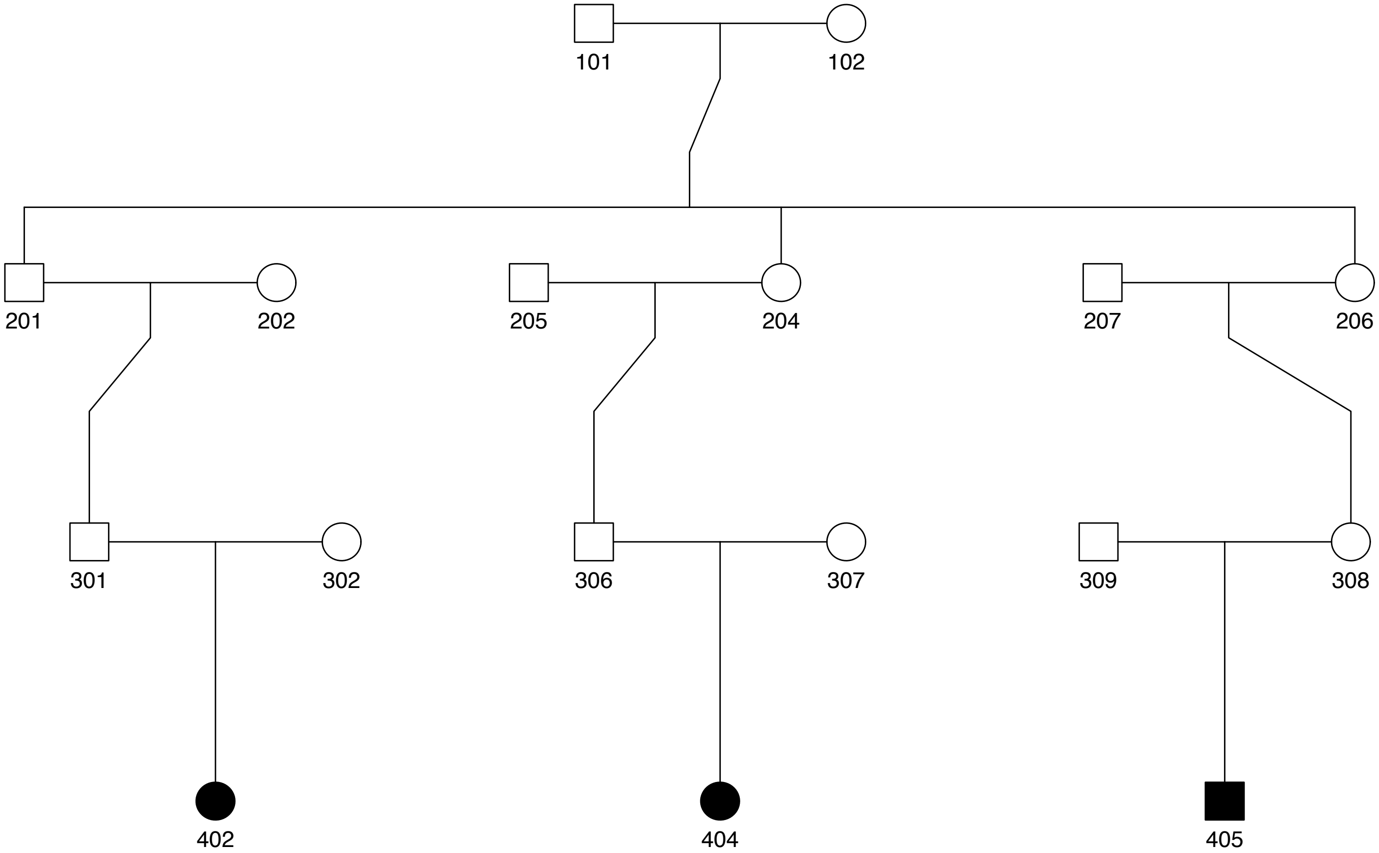
For two sequenced relatives we have

$$P(\text{ rare variant is shared }) = \frac{1}{2^D - 1} = \frac{\theta}{1 - \theta}.$$

D: degree of relationship between the two subjects.

θ: kinship coefficient.

PMID 21850262


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

For a set of n sequenced subjects we want to compute

$$
\begin{aligned}
& P\,(\text{ rare variant is shared })\\
= \quad & P\left(C_1 = \cdots = C_n = 1 | C_1 + \cdots + C_n \geq 1\right)\\
= \quad & \frac{P\left(C_1 = \cdots = C_n = 1\right)}{P\left(C_1 + \cdots + C_n \geq 1\right)}\\
= \quad & \frac{\sum_{j=1}^{n_f} P\left(C_1 = \cdots = C_n = 1 | F_j\right) \times P\left(F_j\right)}{\sum_{j=1}^{n_f} P\left(C_1 + \cdots + C_n \geq 1 | F_j\right) \times P\left(F_j\right)}
\end{aligned}
$$

assuming no IBS without IBD.

PMID 24740360

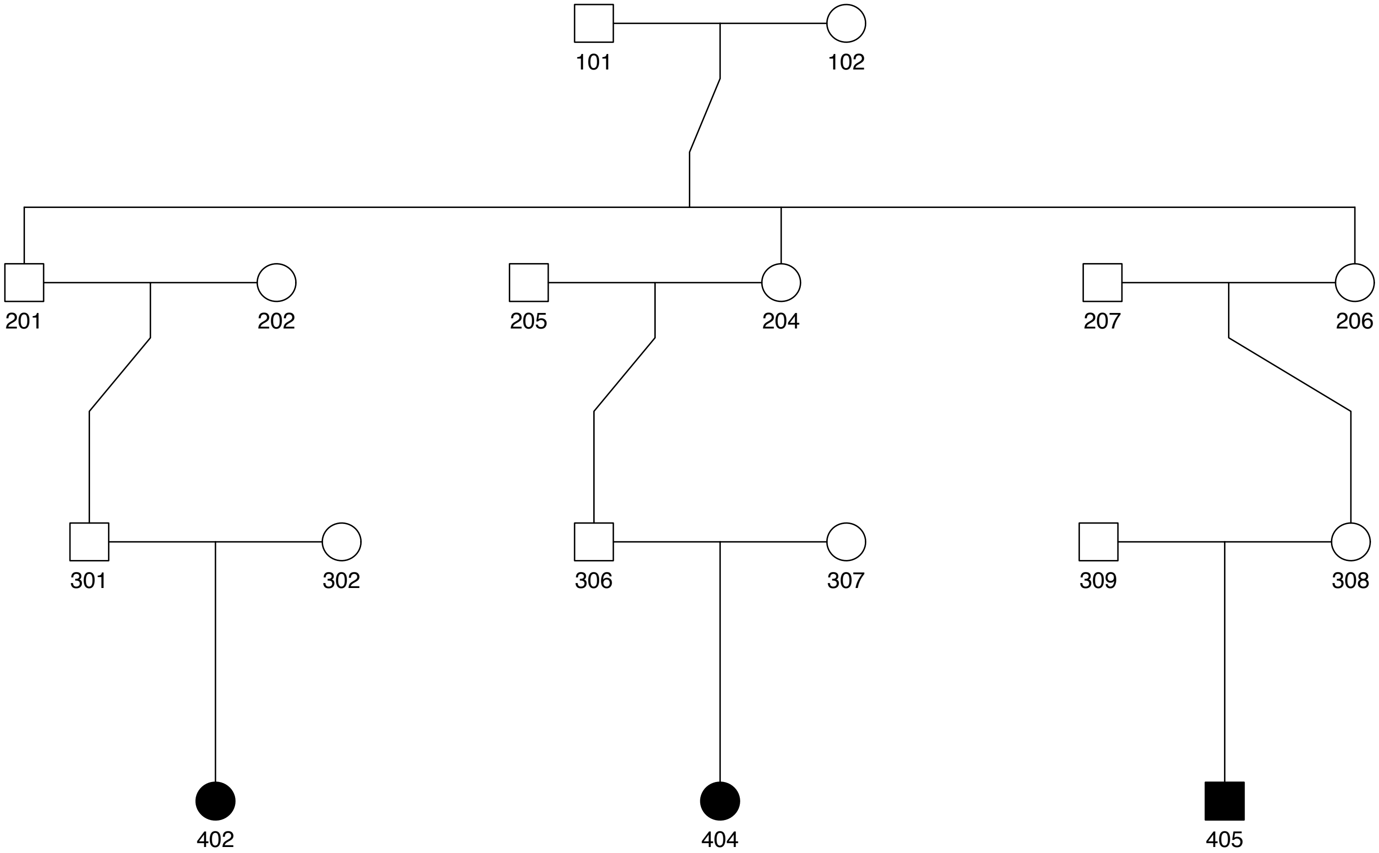
For the common special case of a pedigree with a founder couple ancestral to all sequenced subjects in the pedigree

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

where

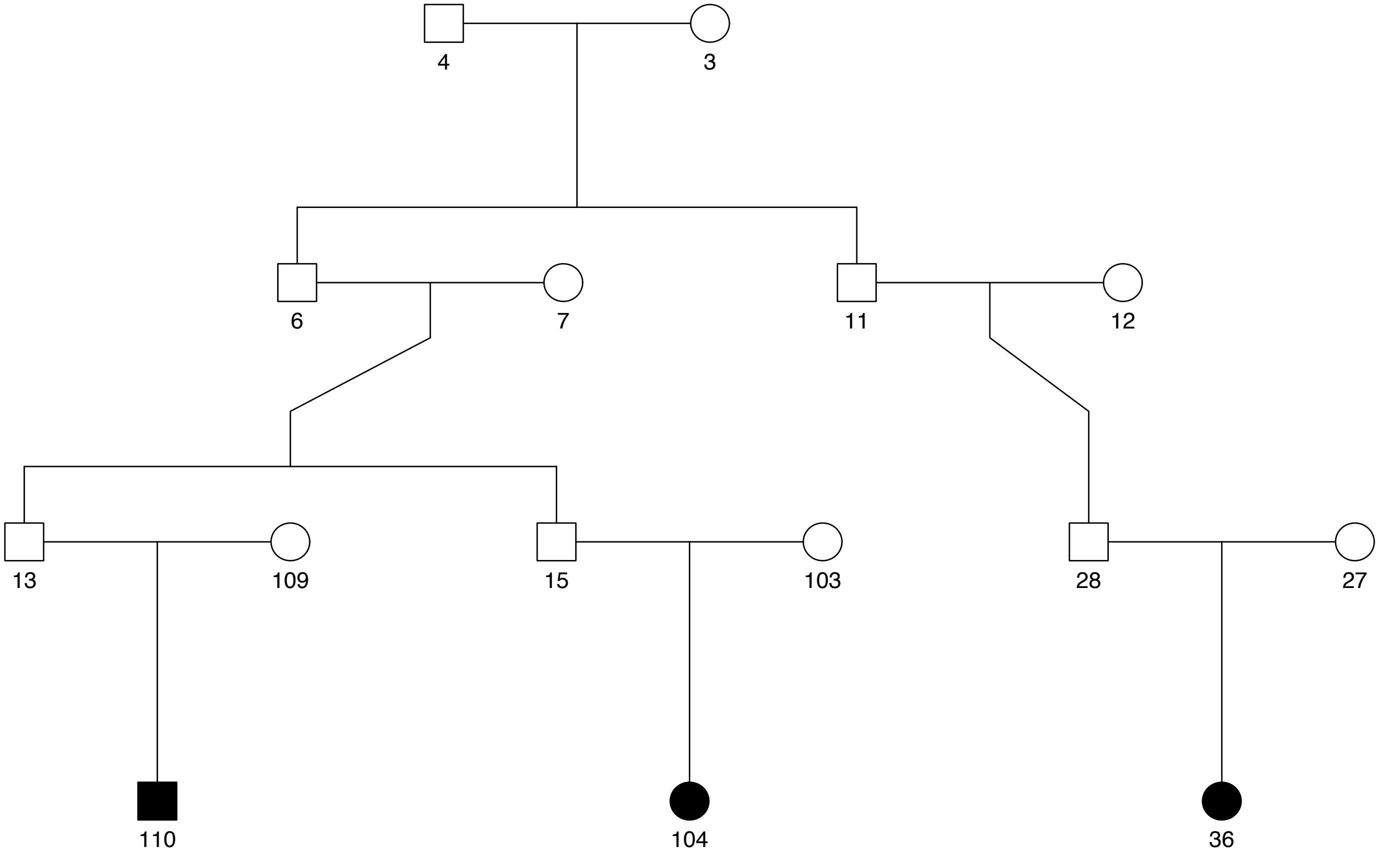
- D_{ij} is the number of generations (meioses) between subject i and his or her ancestor j,
- $D_j = \sum_i D_{ij}$,
- $d(j)$ is the subset of sequenced subjects who descend from founder j,
- f is any of the two founders forming the ancestral couple.

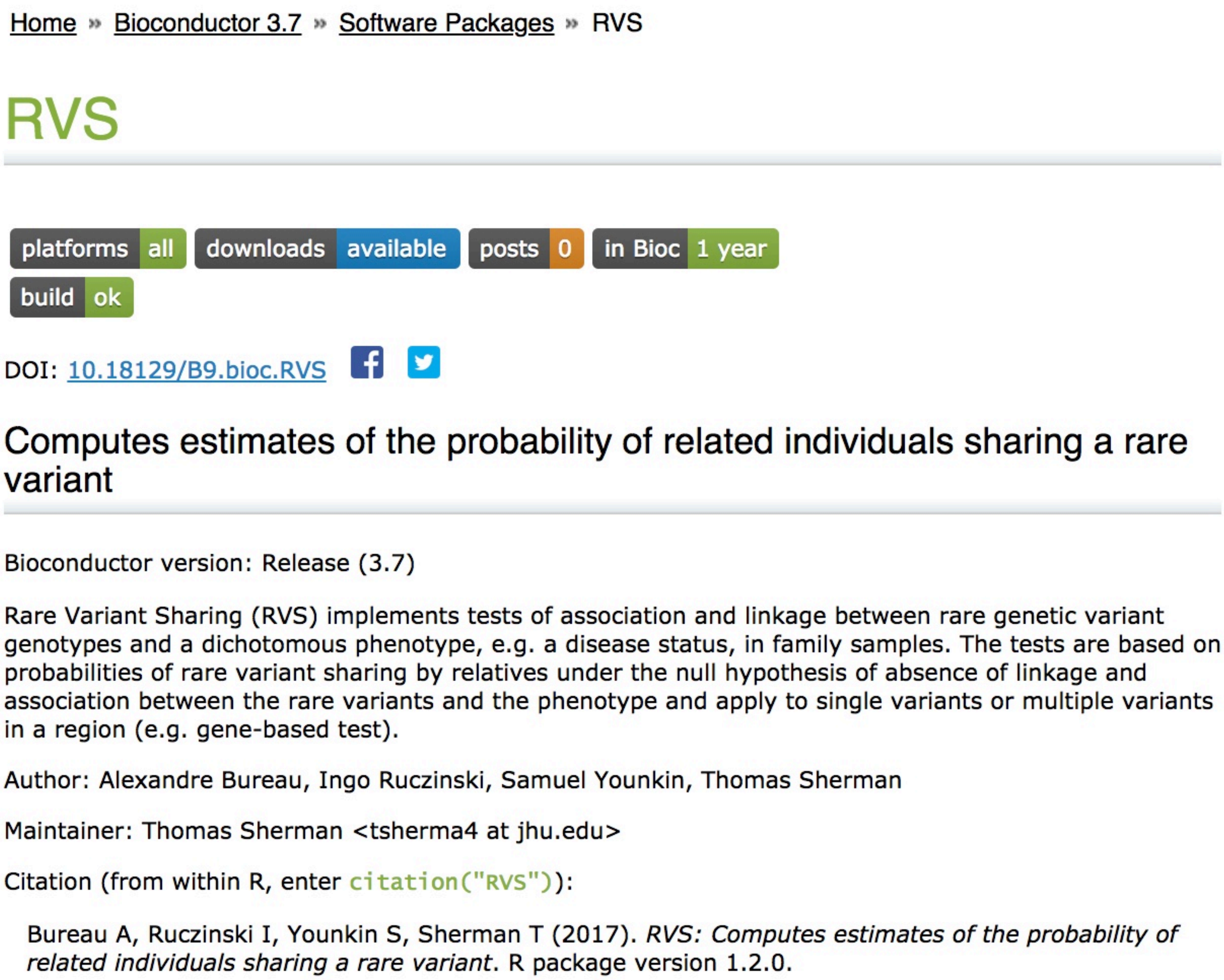
PMID 24740360


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

$$\frac{2\Bigg(\bigg(\dfrac{1}{2}\bigg)^{3}\Bigg)^{3}}{\Bigg(2\Bigg(1-\Bigg(1-\bigg(\dfrac{1}{2}\bigg)^{3}\Bigg)^{3}\Bigg)+3\Bigg(1-\Bigg(1-\bigg(\dfrac{1}{2}\bigg)^{2}\Bigg)\Bigg)+3\,\dfrac{1}{2}\Bigg)\Bigg)}=\frac{1}{745}.$$

PMID 24793288


4
3
6
7
11
12
13
109
15
103
28
27
110
104
36



Home » Bioconductor 3.7 » Software Packages » RVS

RVS

DOI: 10.18129/B9.bioc.RVS

Computes estimates of the probability of related individuals sharing a rare variant

Bioconductor version: Release (3.7)

Rare Variant Sharing (RVS) implements tests of association and linkage between rare genetic variant genotypes and a dichotomous phenotype, e.g. a disease status, in family samples. The tests are based on probabilities of rare variant sharing by relatives under the null hypothesis of absence of linkage and association between the rare variants and the phenotype and apply to single variants or multiple variants in a region (e.g. gene-based test).

Author: Alexandre Bureau, Ingo Ruczinski, Samuel Younkin, Thomas Sherman

Maintainer: Thomas Sherman <tsherma4 at jhu.edu>

Citation (from within R, enter `citation("RVS")`):

> Bureau A, Ruczinski I, Younkin S, Sherman T (2017). *RVS: Computes estimates of the probability of related individuals sharing a rare variant*. R package version 1.2.0.

GESE

Qiao et al 2017. *Gene-based segregation method for identifying rare variants in family-based sequencing studies*. Genetic Epidemiology 41: 309-319.

PBAT

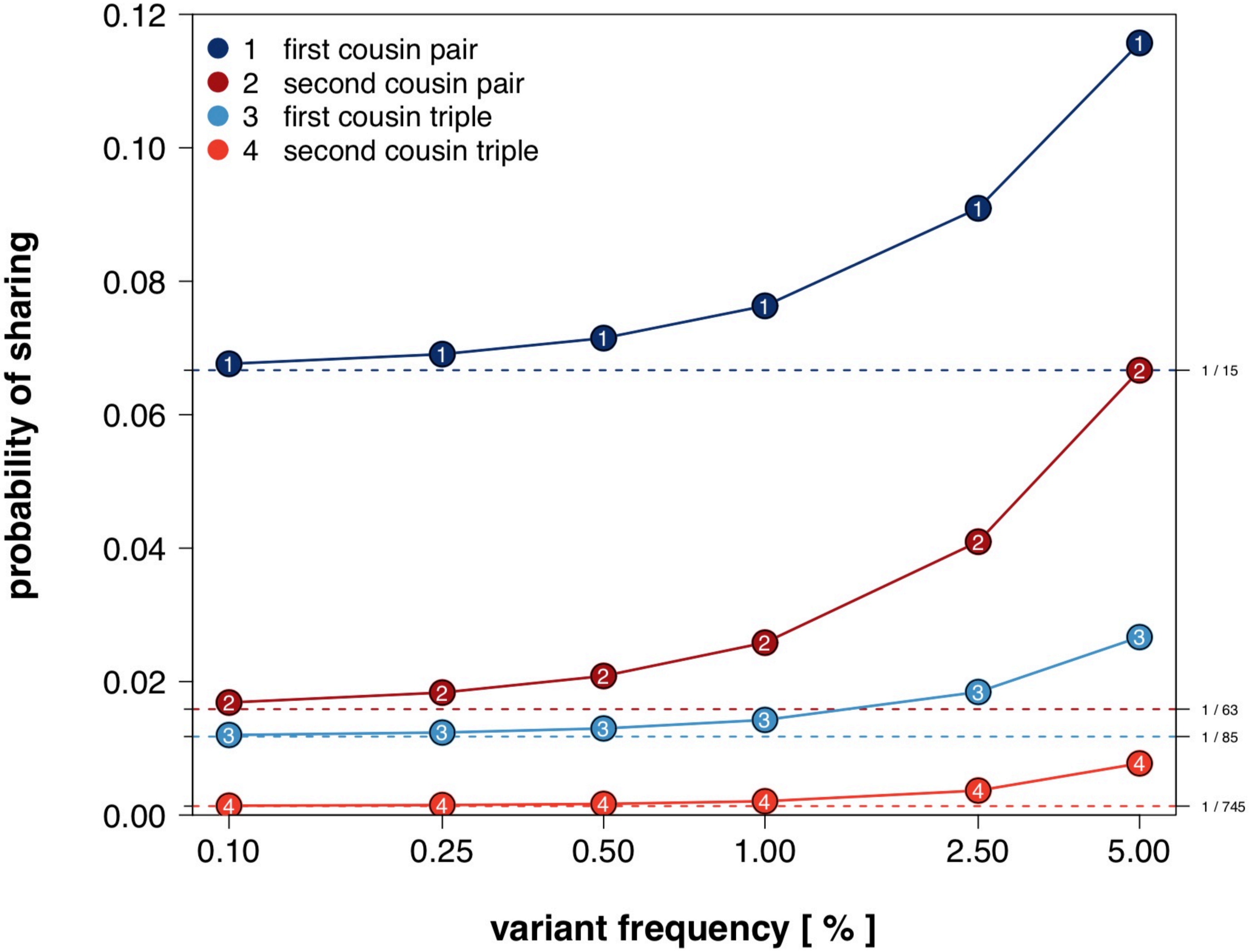
Lange et al 2004. *PBAT: Tools for family-based association studies.* The American Journal of Human Genetics 74: 367-369.

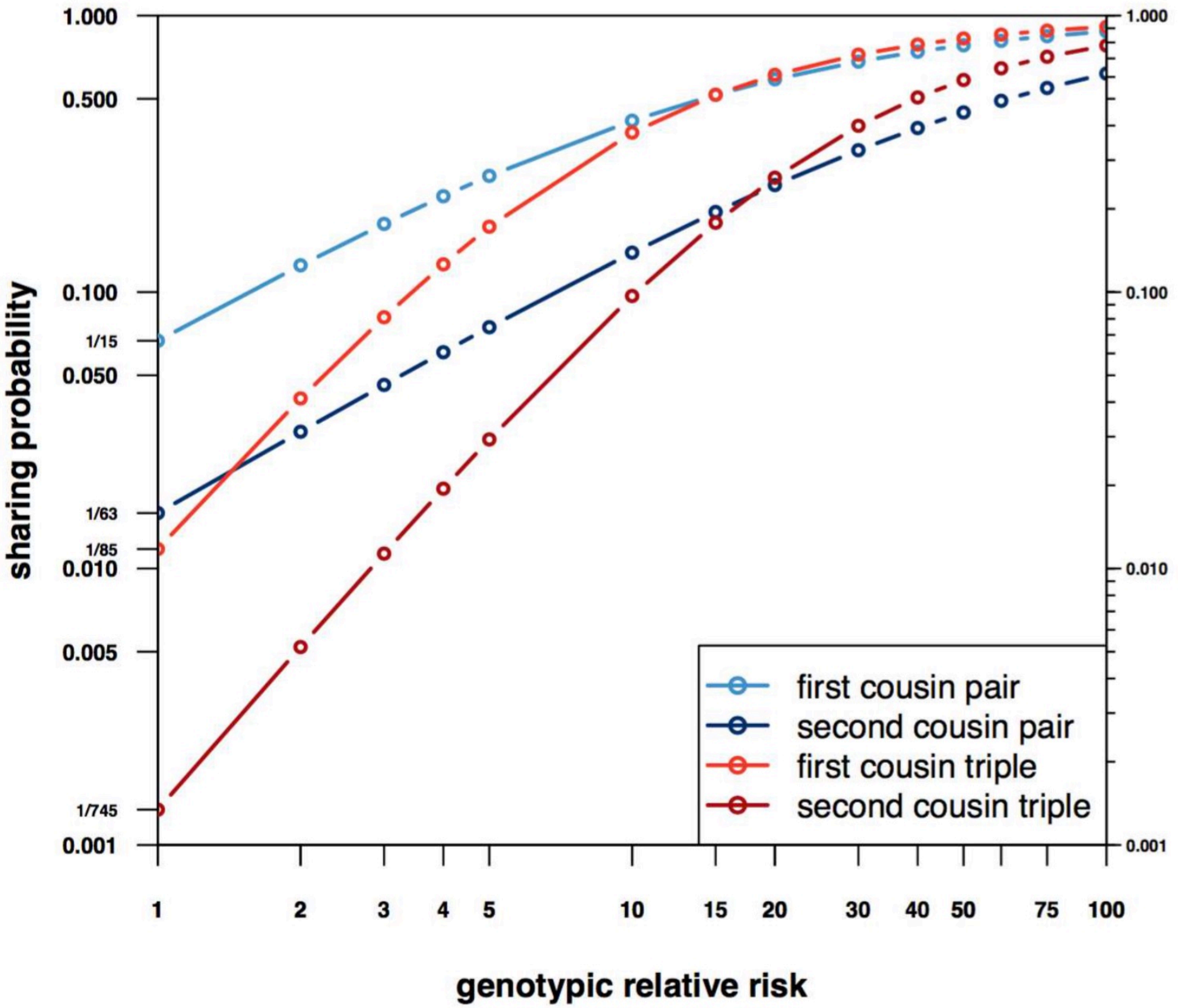
pVAAST

Hu et al. 2014. *A unified test of linkage analysis and rare-variant association for analysis of pedigree sequence data*. Nature Biotechnology 32: 663-669.

RareIBD

Sul et al 2016. *Increasing generality and power of rare-variant tests by utilizing extended pedigrees*. The American Journal of Human Genetics 99: 846-859.


0.12
0.10
0.08
0.06
0.04
0.02
0.00
probability of sharing
● 1 first cousin pair
● 2 second cousin pair
● 3 first cousin triple
● 4 second cousin triple
1 / 15
1 / 63
1 / 85
1 / 745
0.10
0.25
0.50
1.00
2.50
5.00
variant frequency [%]



sharing probability
1.000
0.500
0.100
1/15
0.050
1/63
1/85
0.010
0.005
1/745
0.001
1.000
0.100
0.010
0.001
1
2
3
4
5
10
15
20
30
40
50
75
100
genotypic relative risk
first cousin pair
second cousin pair
first cousin triple
second cousin triple

Sharing probabilities for rs149253049.

Relationship	D	Pr(sharing)
1st cousins	3	0.0667
3rd cousins	7	0.0039
2nd cousins o/rem	6	0.0079
Product		2.0×10^{-6}

PMID 24740360

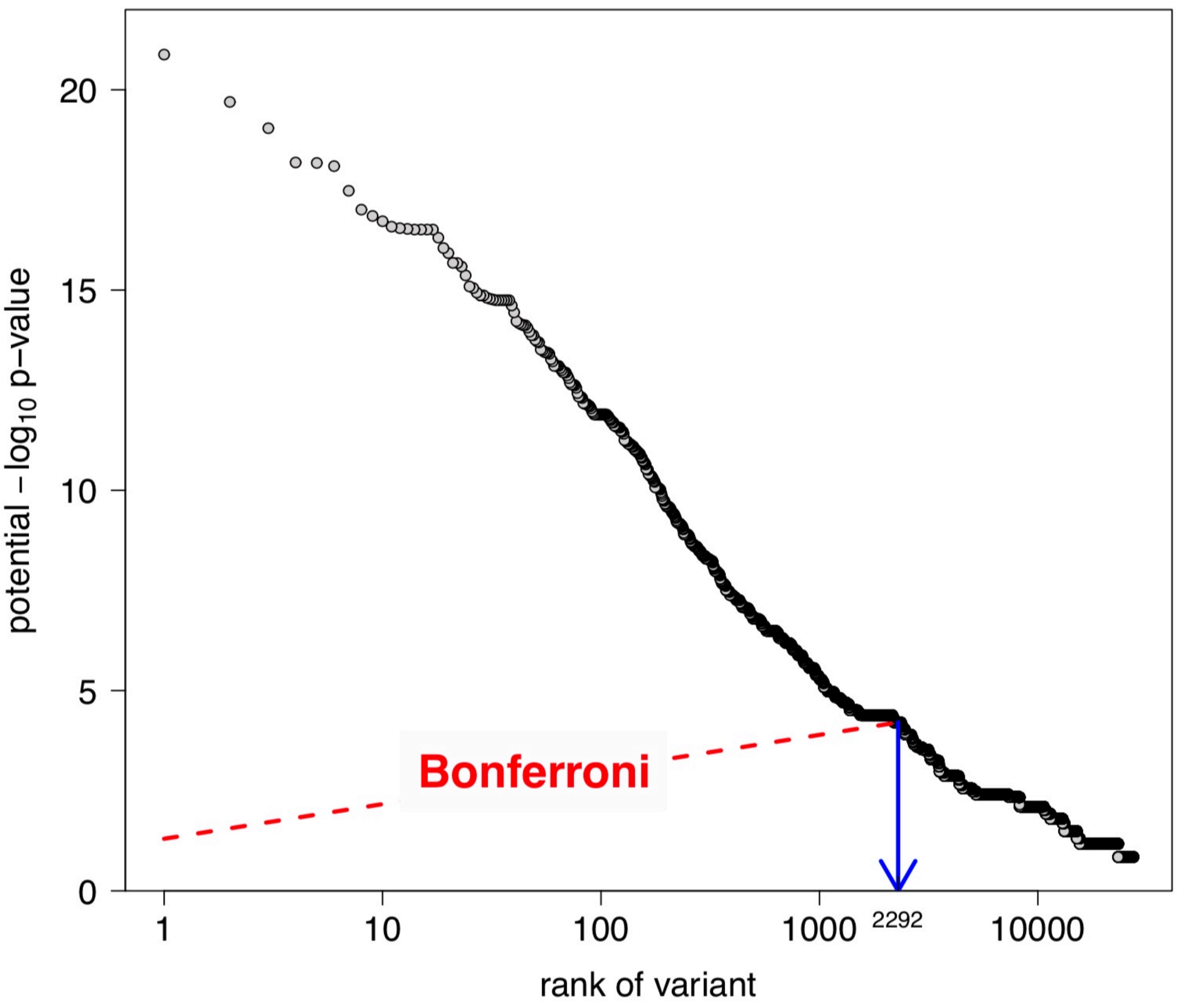
For variants seen in only one family, the RV sharing probability can be interpreted directly as a *P*-value from a Bernoulli trial. For variants seen in M families and shared by affected relatives in a subset S_o of them, the *P*-value can be obtained as the sum of the probability of events as (or more)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

$$p = \sum_{v \in V} \prod_{m=1}^{M} p_m^{I(m \in S_v)} (1 - p_m)^{I(m \notin S_v)}$$

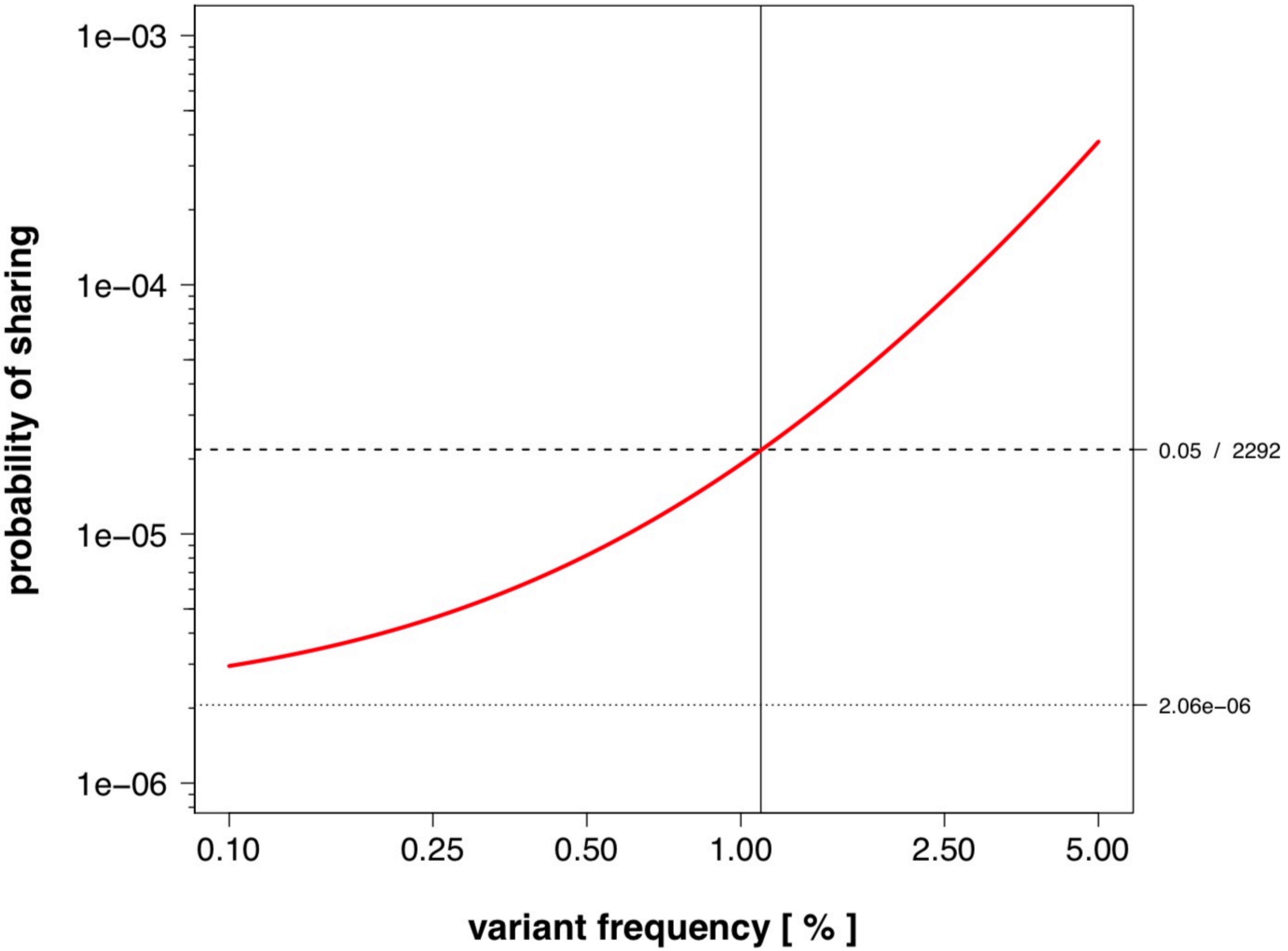
where V is the subset of family sets S_v such that

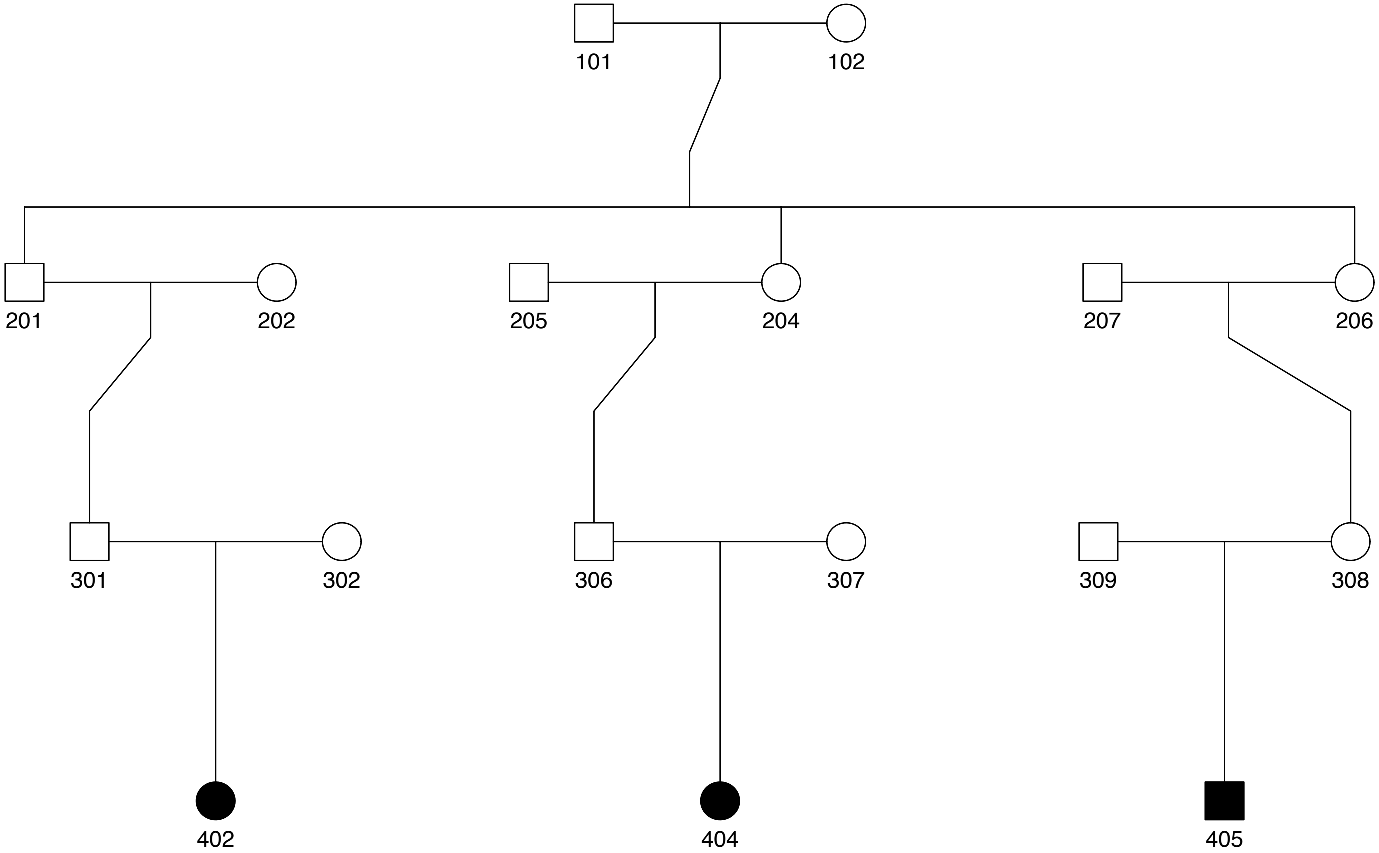
$$\prod_{m=1}^{M} p_m^{I(m \in S_v)} (1 - p_m)^{I(m \notin S_v)} \leq \prod_{m=1}^{M} p_m^{I(m \in S_o)} (1 - p_m)^{I(m \notin S_o)}$$

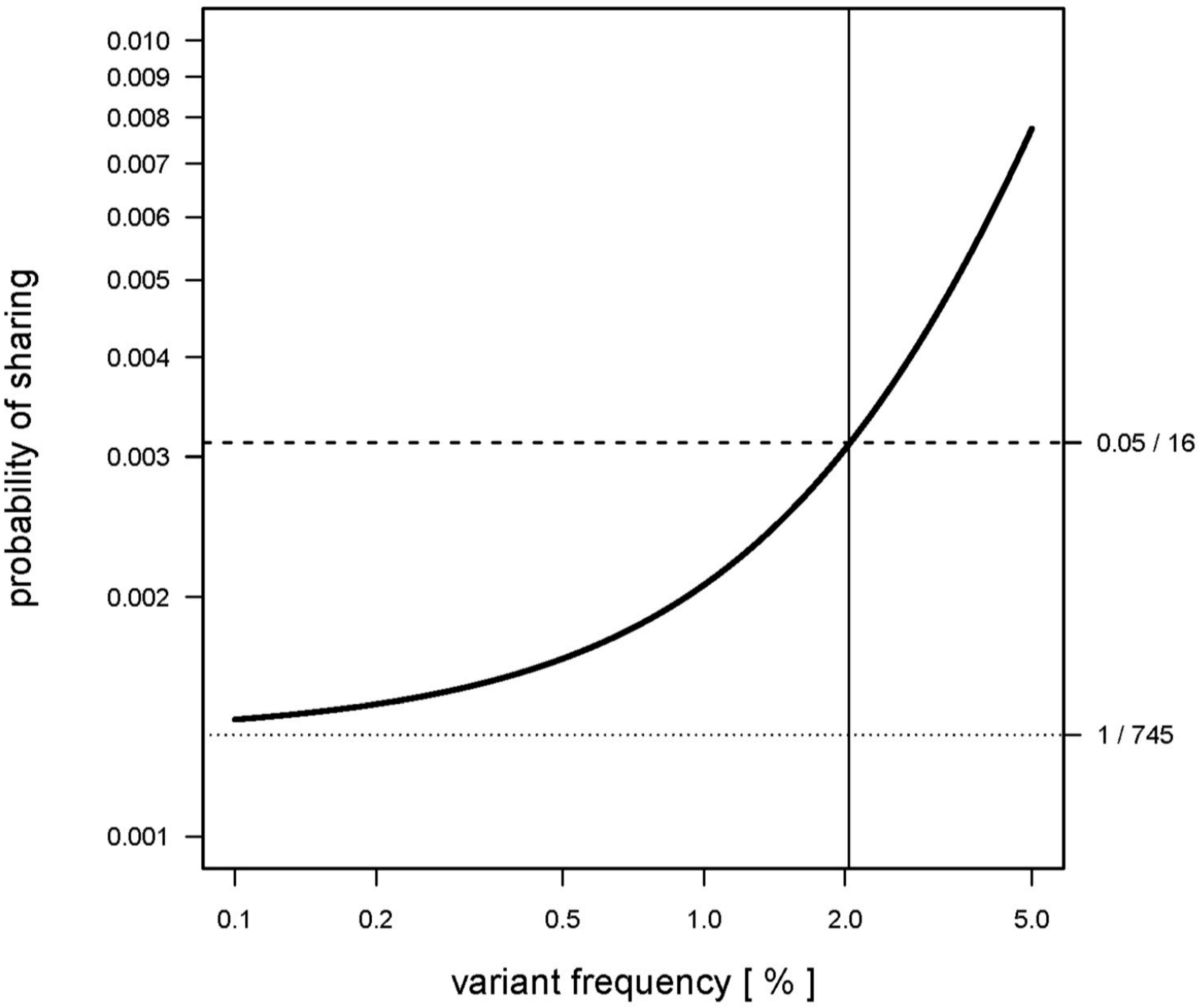
PMID 24740360



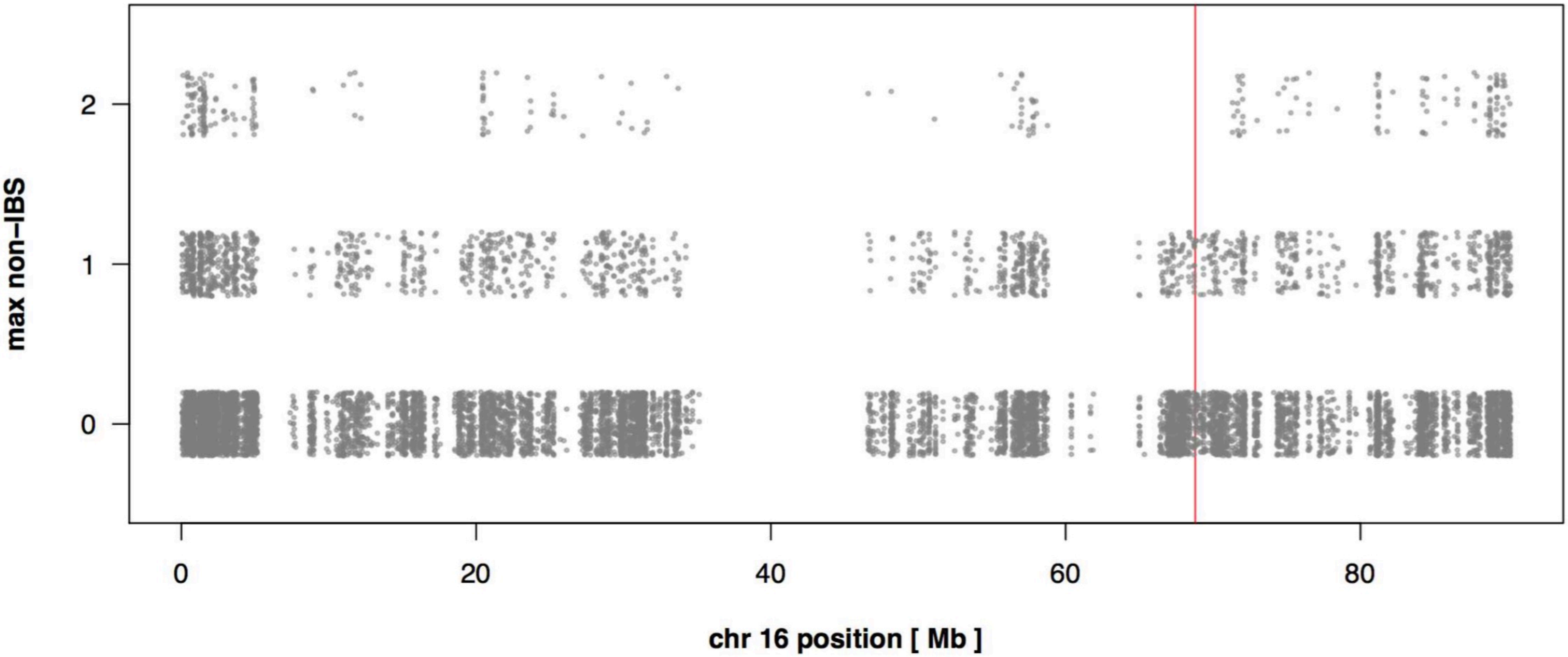
potential $-\log_{10}$ p−value
0
5
10
15
20
1
10
100
1000
2292
10000
rank of variant
Bonferroni


1e−03
1e−04
1e−05
1e−06
probability of sharing
0.05 / 2292
2.06e−06
0.10
0.25
0.50
1.00
2.50
5.00
variant frequency [%]


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



PMID 24793288


max non–IBS
2
1
0
0
20
40
60
80
chr 16 position [Mb]

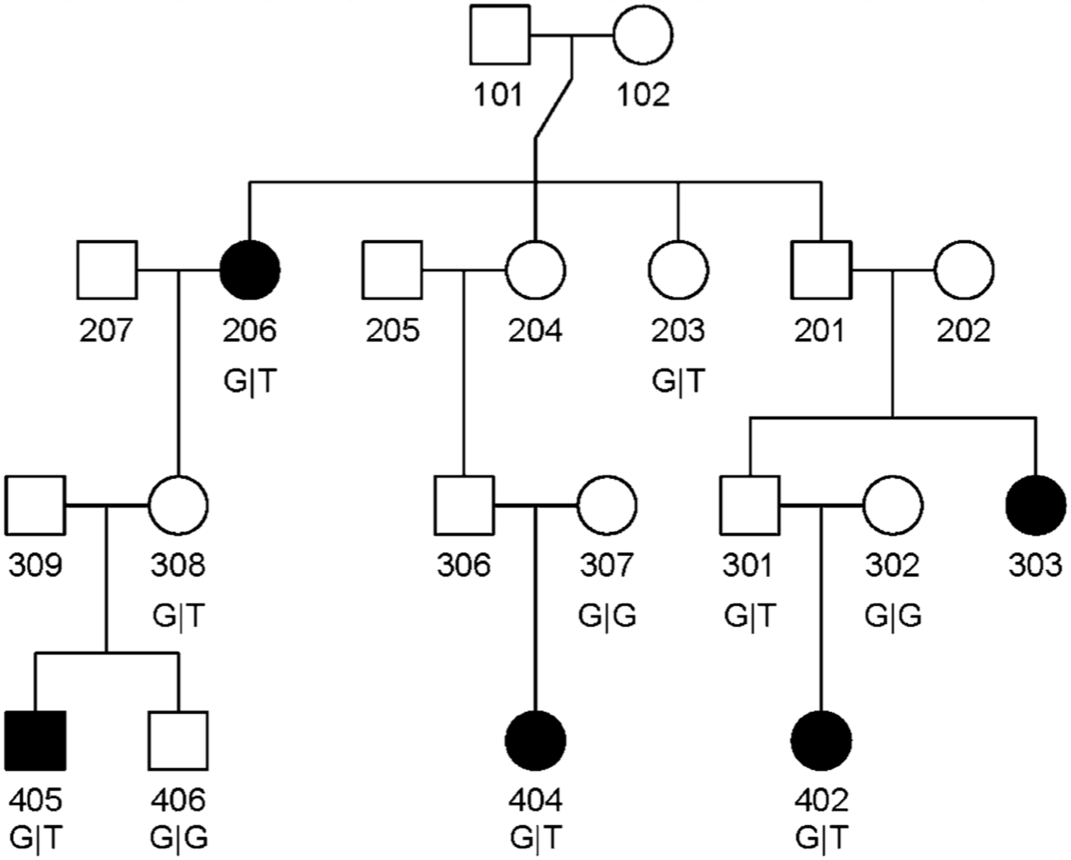
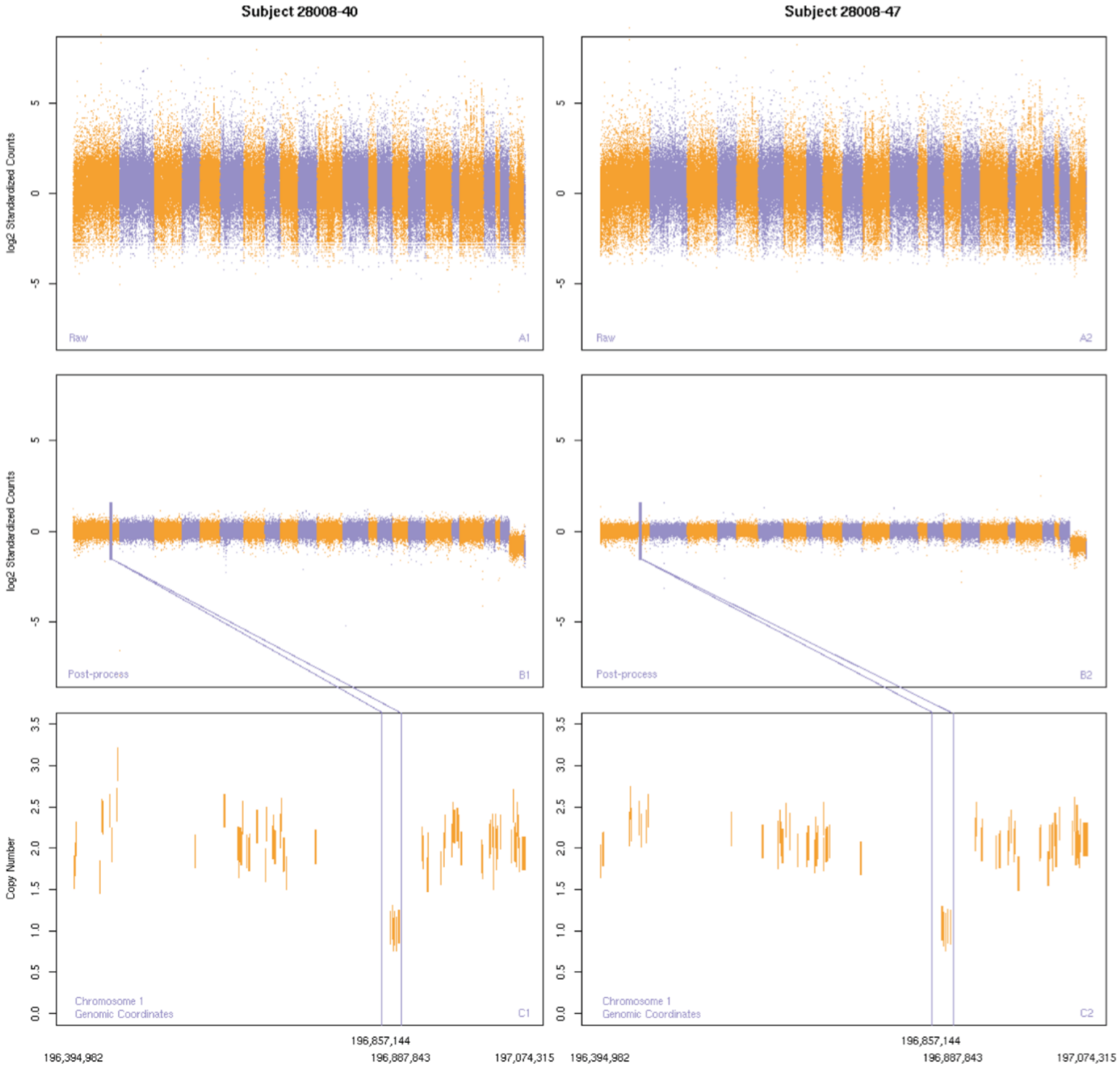


Figure 1 Structure of pedigree where three affected second cousins shared a rare variant in *CDH1*. Affected subjects are represented by filled symbols. Individuals 402, 404, and 405 were sequenced.

PMID 24793288


Subject 28008-40
Subject 28008-47
log2 Standardized Counts
5
0
-5
Raw
A1
A2
Post-process
B1
B2
Copy Number
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Chromosome 1
Genomic Coordinates
C1
C2
196,394,982
196,857,144
196,887,843
197,074,315

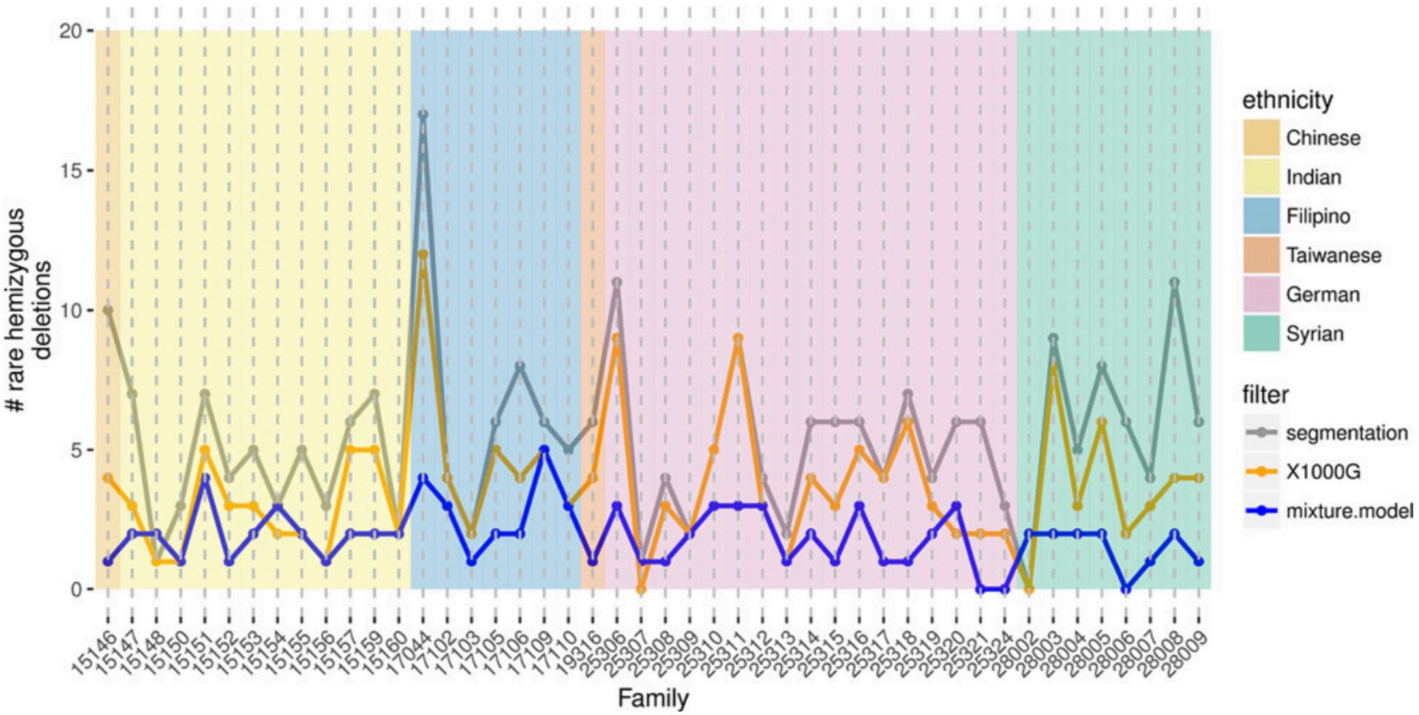


FIGURE 1 The number of autosomal hemizygous deletions (*y*-axis) identified among 95 participants across 46 mulitiplex families (*x*-axis). Candidate deletions were first identified by segmentation of *M* values (gray). Excluding deletions overlapping with homozygous deletions and copy number polymorphisms in the 1000G project, we obtained an initial estimate of the frequency of rare, autosomal hemizygous deletions per family (orange). At each region with a potentially rare deletion, we fit Bayesian mixture models with and without a mixture component for the hemizygous copy number state to the average *M* values. For regions where the log Bayes factor comparing the model with deletion to the model without deletion was at least 2, a sample was considered hemizygous if the posterior probability for the hemizygous component was at least 0.9. Excluding regions with more than five families identified as hemizygous under this mixture model, a total of 88 rare deletions were identified in the 46 families with a median frequency per family of 2 (blue)

PMID 27910131

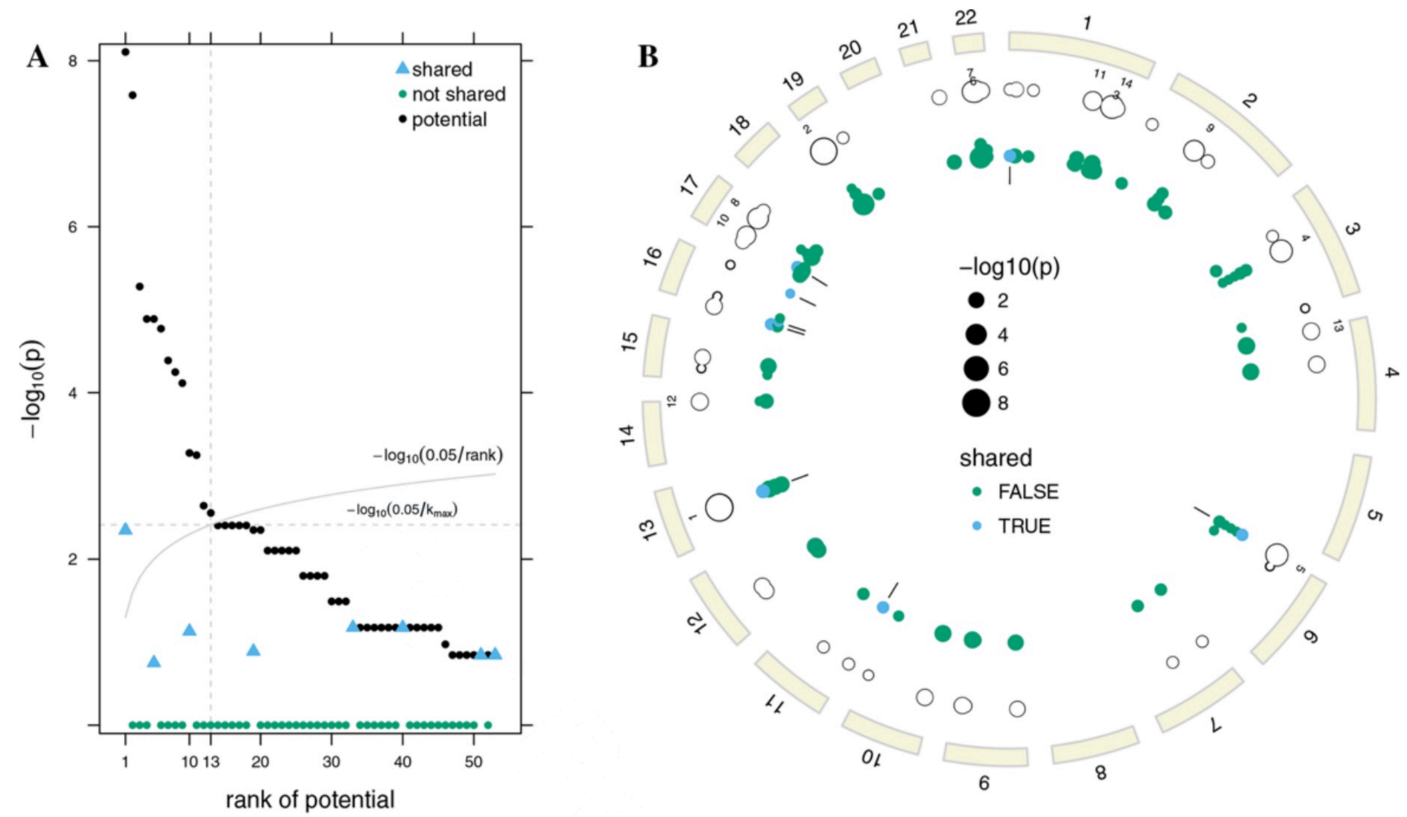


FIGURE 2 Ranks of the potential *P*-values are plotted against the -log10 potential *P*-value (A). Of the 53 regions with one or more rare deletion alleles, the first 13 ranked regions have *potential* for a statistically significant association with oral cleft. Observed sharing probabilities for the first 13 regions were less than their potential *P*-values and are not statistically significant. A circos plot displays these data for each deleted region by genomic position (B). The tracks starting from the outermost ring are the ideograms (beige), the top 13 ranks of the potential sharing probabilities, the potential sharing probabilities (unfilled circles), and the contribution of each family to the potential sharing probabilities (solid circles). Families with a shared deletion are indicated in blue with tick marks on the innermost track highlighting the eight regions with shared deletions

PMID 27910131

Enumeration	Locus 1 Family A	Locus 1 Family B	Locus 2 Family A	Locus 2 Family C	Locus 3 Family C	P shared
1	🟦	🟦	🟦	🟦	🟦	P1
2	🟦	🟦	🟦	🟦	⬜	P2
3	🟦	🟦	🟦	⬜	🟦	P3
4	🟦	🟦	⬜	🟦	🟦	P4
5	🟦	⬜	🟦	🟦	🟦	P5
6	⬜	🟦	🟦	🟦	🟦	P6
7	🟦	🟦	🟦	⬜	⬜	P7
			⋮			
27	⬜	⬜	⬜	⬜	🟦	P27
28	⬜	⬜	⬜	🟦	⬜	P28
29	⬜	⬜	🟦	⬜	⬜	P29
30	⬜	🟦	⬜	⬜	⬜	P obs
31	🟦	⬜	⬜	⬜	⬜	P31
32	⬜	⬜	⬜	⬜	⬜	P32

🟦 Deletion shared by family at location

⬜ Not shared

🟧 Observed sharing. Shared only by Locus 1 in Family B

PMID 27910131

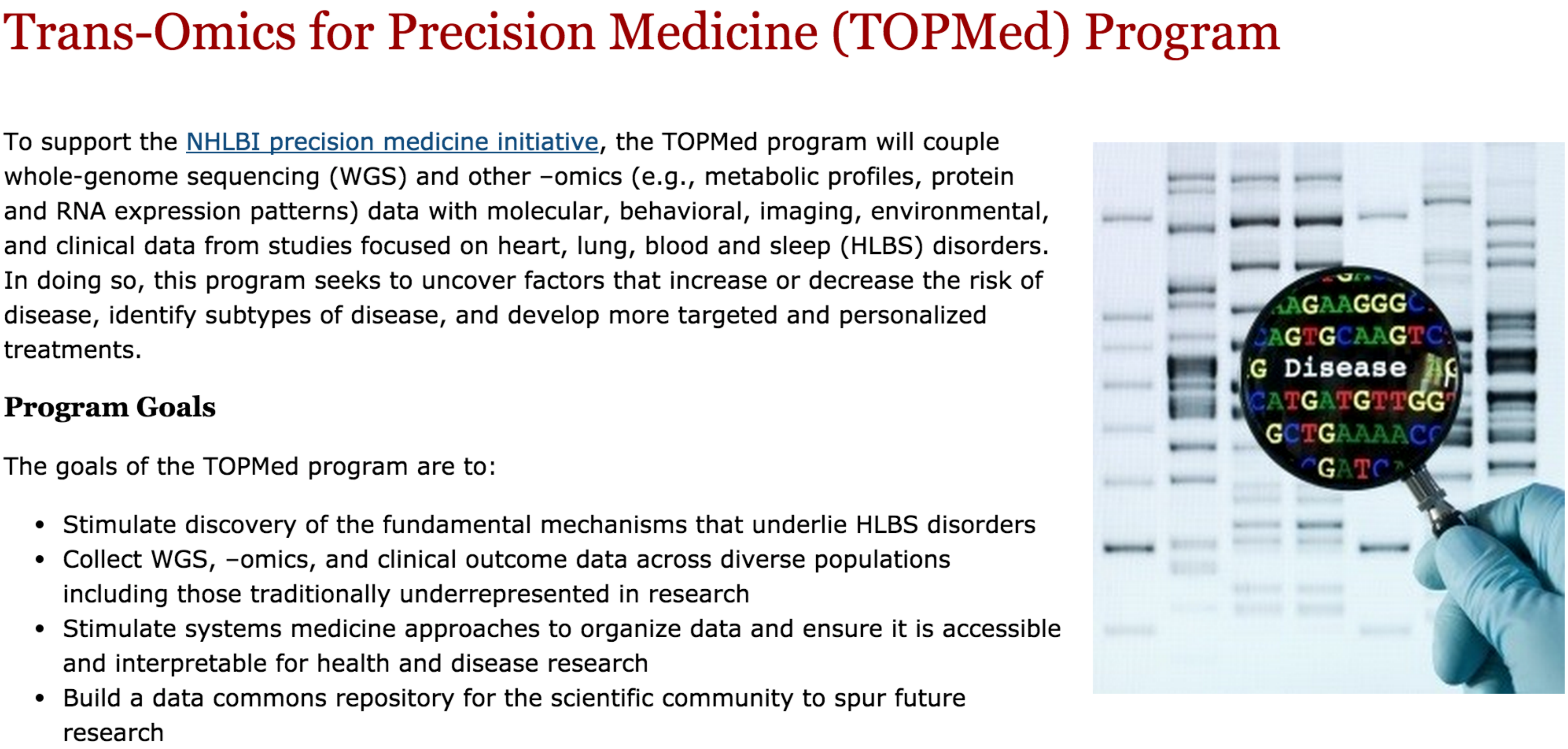
New Results

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

Alexandre Bureau, Ferdouse Begum, Margaret A. Taub, Jacqueline Hetmanski, Margaret M. Parker, Hasan Albacha-Hejazi, Alan F. Scott, Jeffrey C. Murray, Mary L. Marazita, Joan E. Bailey-Wilson, Terri H. Beaty, Ingo Ruczinski

doi: https://doi.org/10.1101/285874

www.biorxiv.org/content/early/2018/07/11/285874



Trans-Omics for Precision Medicine (TOPMed) Program

To support the NHLBI precision medicine initiative, the TOPMed program will couple whole-genome sequencing (WGS) and other –omics (e.g., metabolic profiles, protein and RNA expression patterns) data with molecular, behavioral, imaging, environmental, and clinical data from studies focused on heart, lung, blood and sleep (HLBS) disorders. In doing so, this program seeks to uncover factors that increase or decrease the risk of disease, identify subtypes of disease, and develop more targeted and personalized treatments.

Program Goals

The goals of the TOPMed program are to:

- Stimulate discovery of the fundamental mechanisms that underlie HLBS disorders
- Collect WGS, –omics, and clinical outcome data across diverse populations including those traditionally underrepresented in research
- Stimulate systems medicine approaches to organize data and ensure it is accessible and interpretable for health and disease research
- Build a data commons repository for the scientific community to spur future research

www.nhlbi.nih.gov/research/resources/nhlbi-precision-medicine-initiative/topmed

METHODS

Bureau A, Younkin SG, Parker MM, Bailey-Wilson JE, Marazita ML, Murray JC, …, Ruczinski I (2014).
Inferring rare disease risk variants based on exact probabilities of sharing by multiple affected relatives.
Bioinformatics (Oxford, England) 30: 2189-2196.

Bureau A, Parker MM, Ruczinski I, Taub MA, Marazita ML, Murray JC, Mangold E, …, Beaty TH (2014).
Whole exome sequencing of distant relatives in multiplex families implicates rare variants in candidate genes for oral clefts.
Genetics 197: 1039-1044.

Fu J, Beaty TH, Scott AF, Hetmanski J, Parker MM, Wilson JE, Marazita ML, …, Scharpf RB (2017).
Whole exome association of rare deletions in multiplex oral cleft families.
Genetic Epidemiology 41(1): 61-69.

Bureau A, Begum F, Taub MA, Hetmanski J, Parker MM, Albacha-Hejazi H, Scott AF, …, Ruczinski I (2018).
Inferring disease risk genes from sequencing data in multiplex pedigrees through sharing of rare variants.
Genetic Epidemiology (to appear).

SOFTWARE

Sherman T, Fu J, Scharpf RB, Bureau A, Ruczinski I (2018).
Detection of rare disease variants in extended pedigrees using RVS.
Bioinformatics (in revision).

bioconductor.org/packages/release/bioc/html/RVS.html