Rare variant sharing methods with genealogies

Alexandre Bureau

Joint work with Ingo Ruczinski

and Marie-Hélène Roy-Gagnon

BIRS, August 6th, 2018

Outline

- Rare variant sharing probability reminder
- Cryptic relatedness problem and genotype-based solution
- Exploiting extensive genealogical databases to improve power and deal with cryptic relatedness
- Computational considerations

Rare variant sharing by *n* sequenced relatives

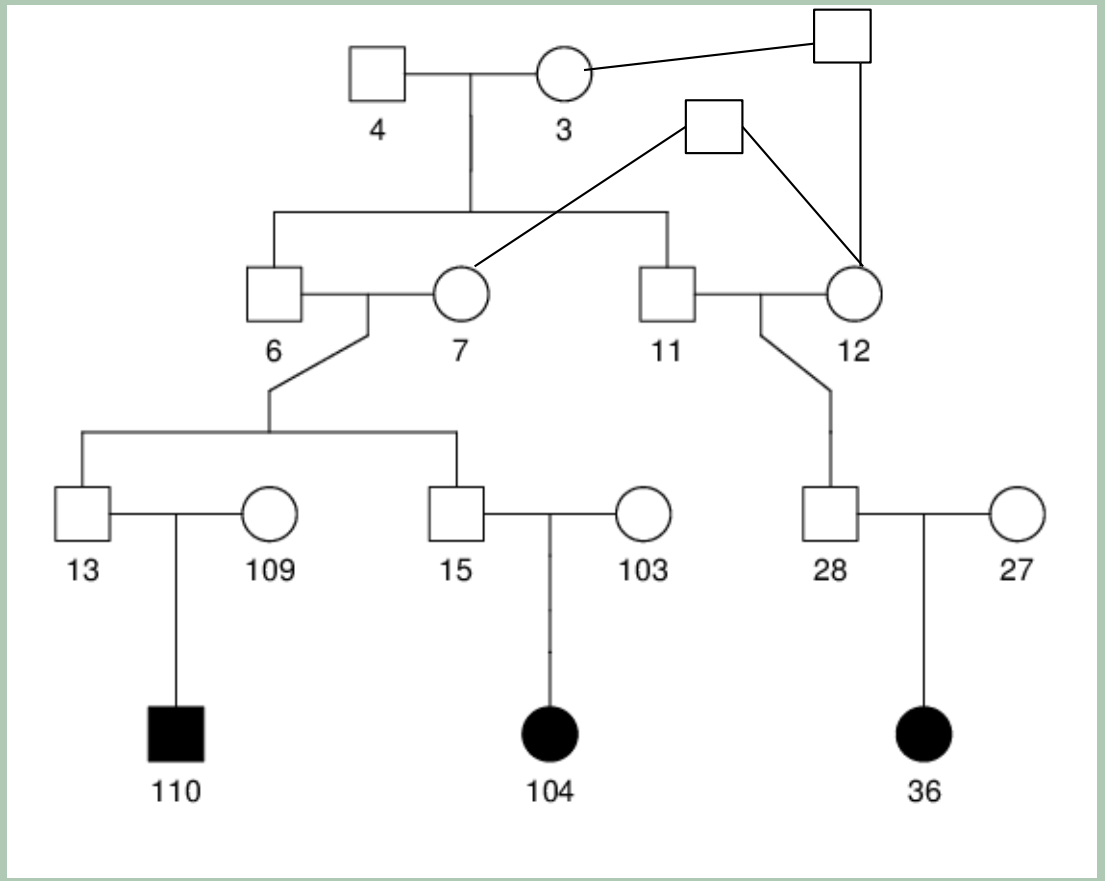
C_i : Number of copies of the rare variant in subject i

F_j : Indicator variable that founder j introduced one copy of the rare variant (RV) into the pedigree (among n_f)

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

Key point: it is a joint probability among affected subjects

Cryptic relatedness increases sharing probabilities



Genotype-based solutions

- Cryptic relatedness is often accounted for by replacing pedigree-based measures of relatedness by genome-wide genotype-based estimates
  - Typically done for pairwise relationships (kinship coefficients)
- With the RV sharing approach, we have proposed such approach (implmented in the RVS package). It requires converting kinship estimates among founders in distribution of number of distinct alleles among founders (Bureau et al. 2014).

Generalization to RV introduced by 2 founders

$$P[\text{RV shared}] =$$

$$\frac{w\frac{1}{n_f}\sum_{j=1}^{n_f} P[C_1 = \cdots = C_n = 1|F_j^U] + (1-w)\frac{2}{n_f(n_f-1)}\sum_j\sum_{k>j} P[C_1 = \cdots = C_n = 1|F_j, F_k]}{w\frac{1}{n_f}\sum_{j=1}^{n_f} P[C_1 + \cdots + C_n \geq 1|F_j^U] + (1-w)\frac{2}{n_f(n_f-1)}\sum_j\sum_{k>j} P[C_1 + \cdots + C_n \geq 1|F_j, F_k]}$$

Where $w = n_f P_U$, P_U is the probability each single founder introduces the RV.

Computing P_U

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

Where *A* is the number of alleles distinct by descent among the founders. We parameterize P[*A*] to be proportional to

$2n_f - d$	$\ldots$	$2n_f - 1$	$2n_f$
$\frac{1}{d!}\theta^d$	$\ldots$	θ	1

Poisson distribution truncated at d, the maximum number of alleles present twice among the founders.

Estimating θ

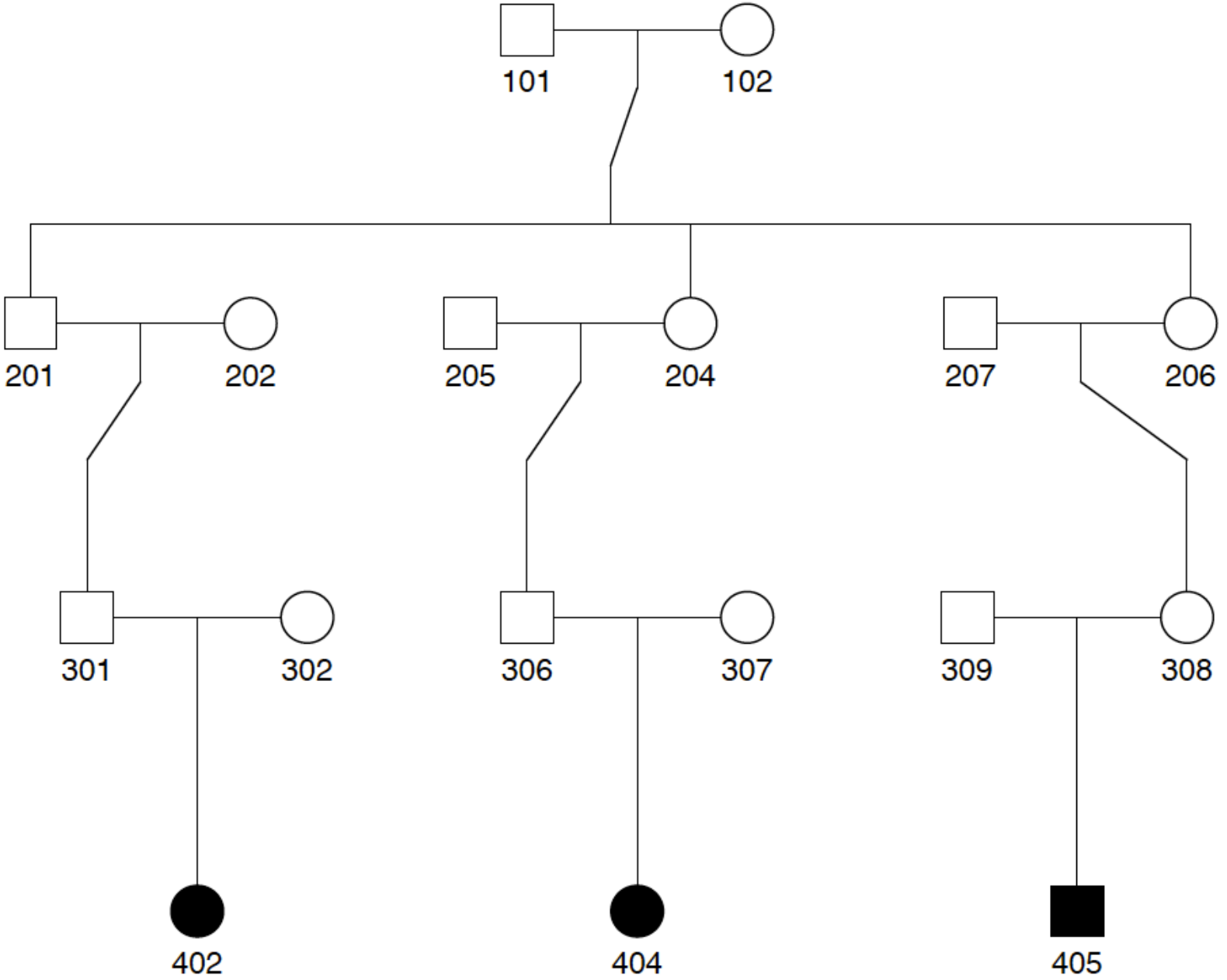
The expected kinship coefficient among the n_f founders with respect to the previous distribution is

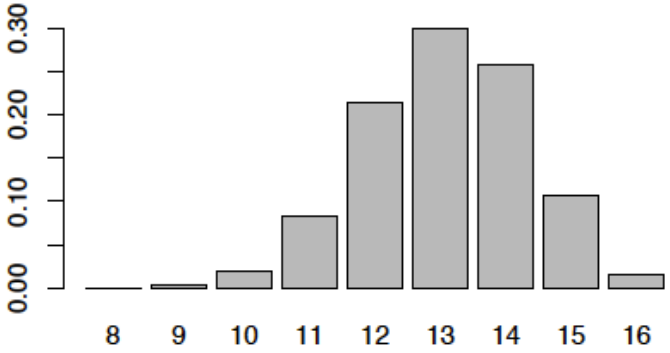
$$E[\Phi] = \frac{\sum_{a=2n_f-d}^{2n_f-1} \frac{1}{(2n_f-a)!}\theta^{(2n_f-a)}\bar{\phi}_a}{\sum_{a=2n_f-d}^{2n_f} \frac{1}{(2n_f-a)!}\theta^{(2n_f-a)}}$$

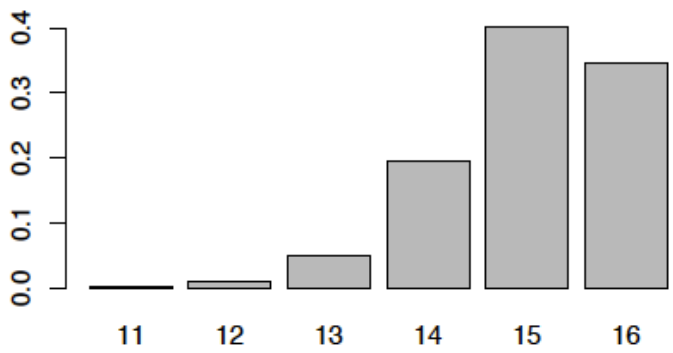
$$\bar{\phi}_a = \frac{1}{2(n_f-1)}\frac{2n_f-a}{n_f}\frac{2n_f-a-1}{n_f-1} + \frac{1}{4(n_f-1)}\left[\frac{(2n_f-a)(a-n_f)}{n_f(n_f-1)} + \frac{2(2n_f-a)(a-n_f)}{n_f(2n_f-1)}\right]$$

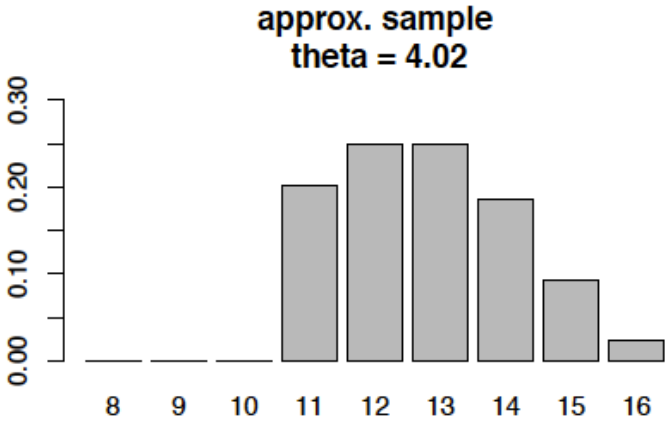
Set $E[\Phi] = \hat{\phi}^f$ the estimated mean kinship among founders and solve for θ.

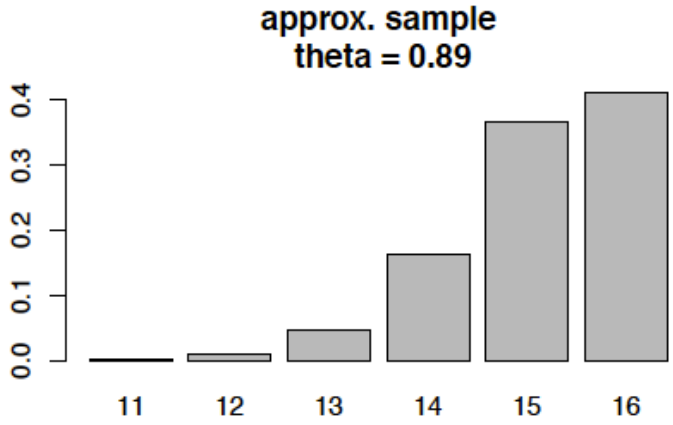
Pedigree with 8 founders used in simulation study

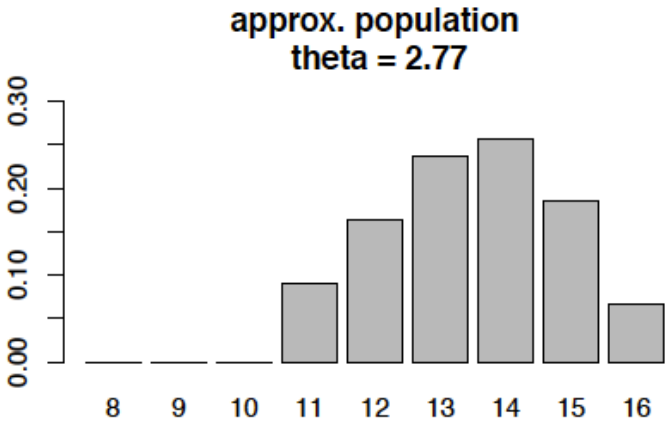


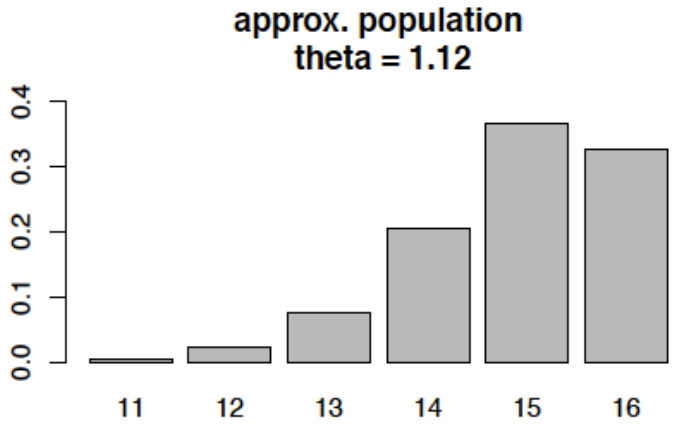






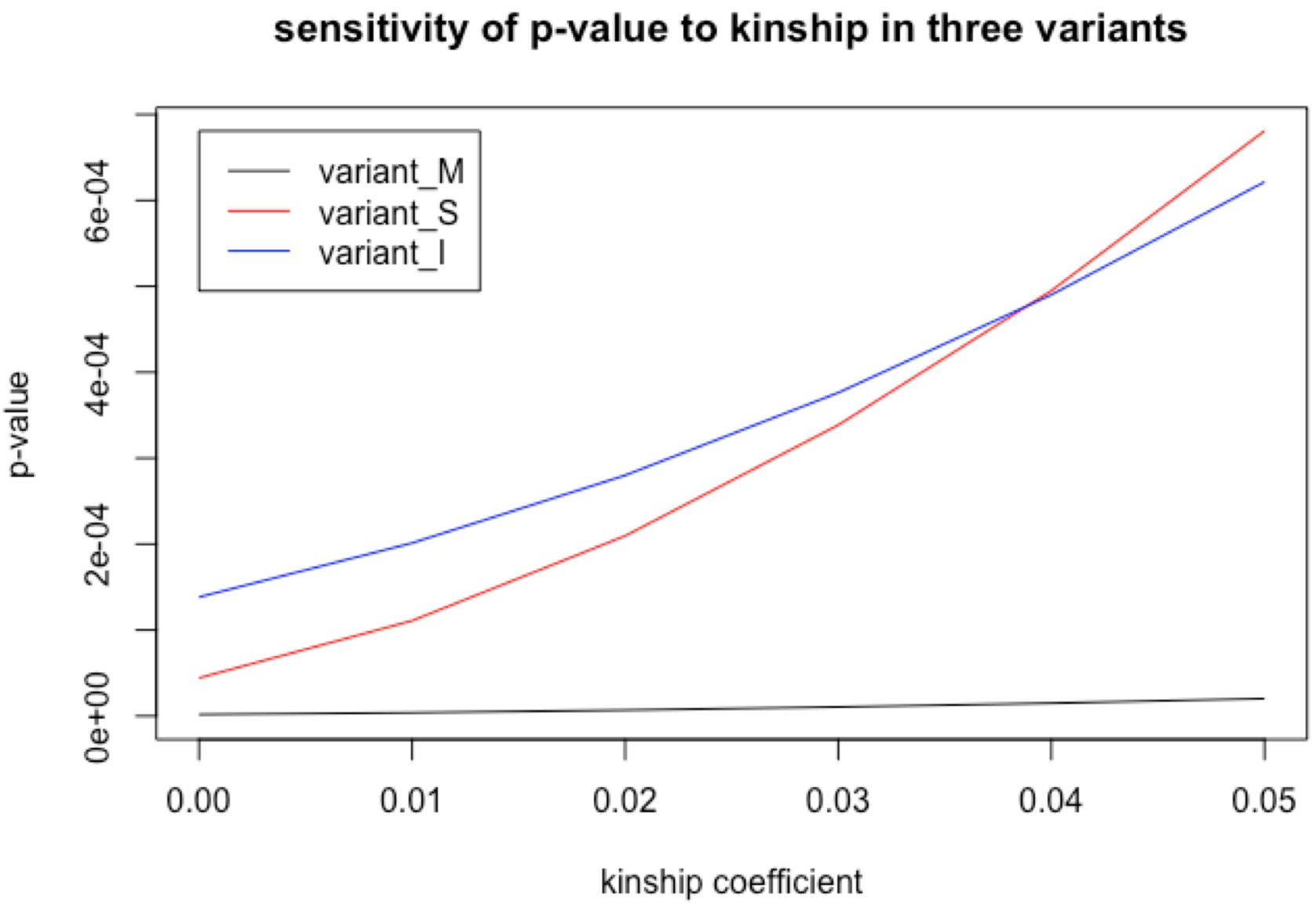






Distribution of number of distinct alleles among samples of 8 subjects from small population and approximation with d = 5.

Sensitivity to kinship

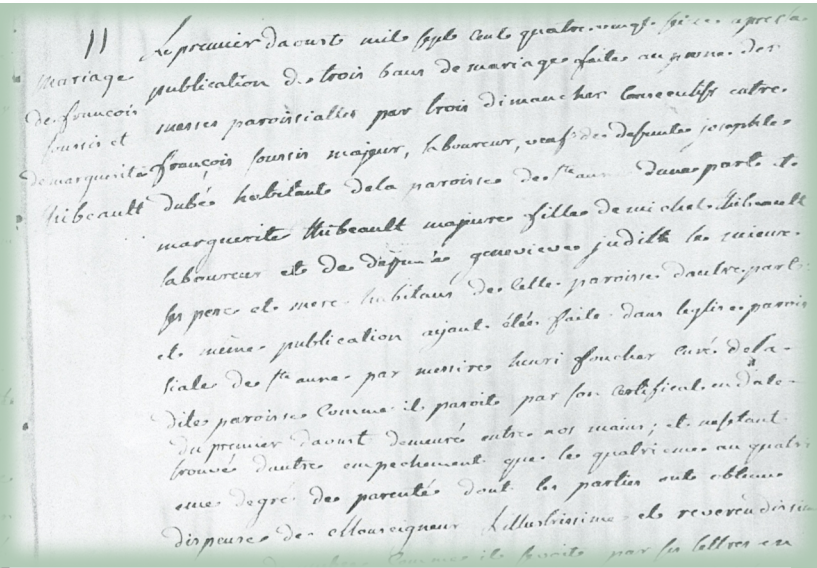


Improvement with extensive genealogical database

- If all relationships contributing substantially to sharing probabilities are contained in the database, then use the database as a huge pedigree.
- Merge all families originally considered unrelated but where in fact all affected relatives have one or more ancestor in common. This is potentially more powerful than treating the families as independent.
- Cryptic relatedness is imbedded in the genealogy.

BALSAC genealogy project

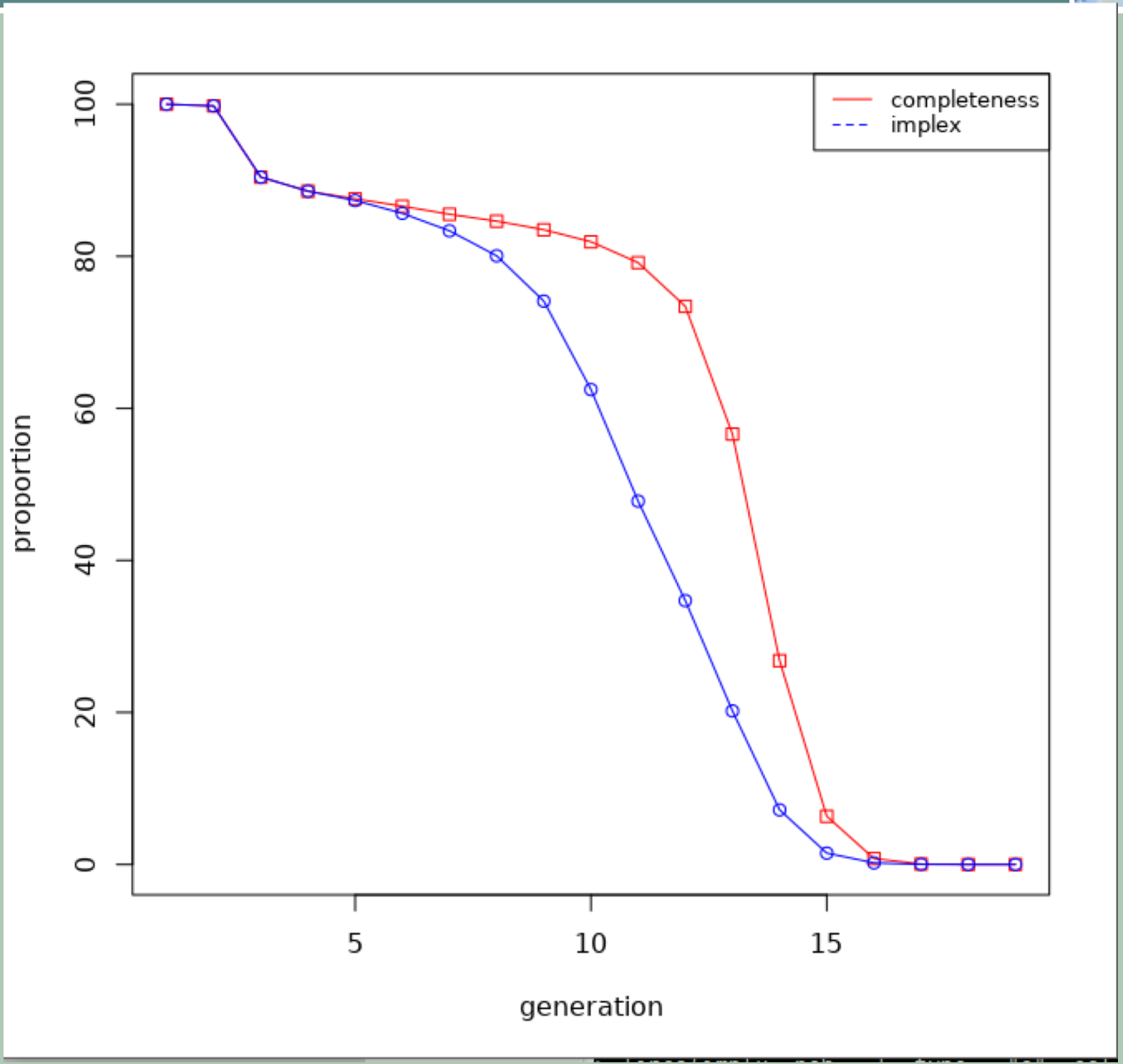
- Vital data records of Quebec
- Automatic construction of ascending or descending genealogies, family histories and individual life courses
- Currently covers >3 million computerized and linked records (mostly catholic marriages) over >3 centuries (1620-1965)
- Aims to cover the entire population of Quebec from the onset of settlement to recent years



Example of asthma family study in Saguenay-Lac-St-Jean

- Family recruitment : at least one asthmatic proband with at least one unaffected parent (PI Catherine Laprise, UQAC)
- 217 families from the Saguenay-Lac-St-Jean (SLSJ) region of Quebec comprising 1018 individuals (430 family founders)
- Connected in a single genealogy comprising 56,815 individuals (7,709 population founders) using the BALSAC database.
- Highly complete over 12 generations, extending maximally to 19 generations.

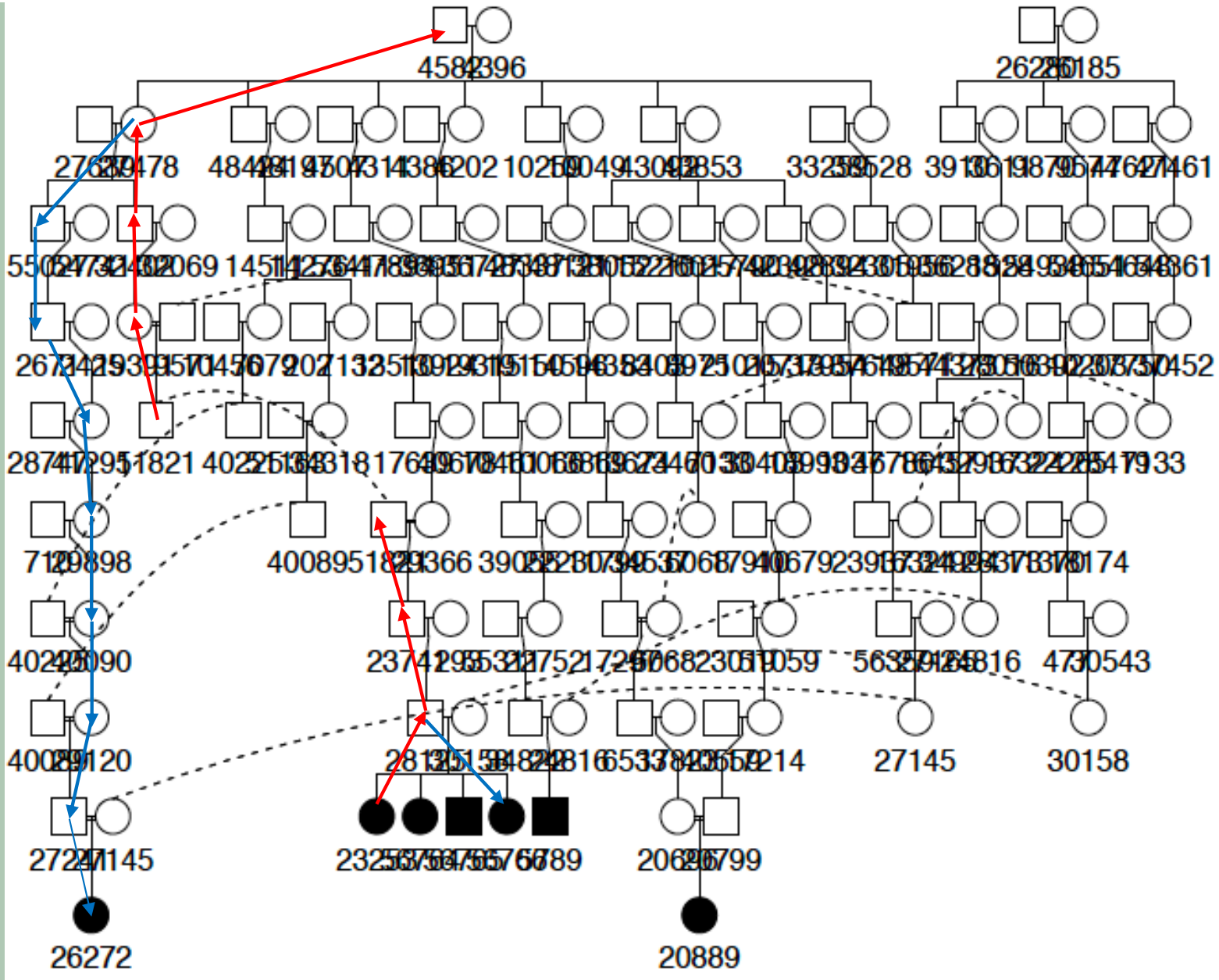
SLSJ ashtma study genealogy completeness



Monte Carlo simulation conditional on carrier

- Genealogies such as the SLSJ asthma study are too complex for exact computations using the RVS package.
- Forward simulation of variants introduced by a single population founder (gene dropping) rarely results in the variant appearing in current affected subjects.
- Instead, perform simulation conditionnal on one affected subject carrying the variant, simulating the transmission path back to a founder, then performing gene dropping from the subjects in that path.

Example backward and forward simulation



Recovering the distribution of RV sharing events

Distribution we want

$$P[C_1,\ldots,C_n|\sum_{i=1}^{n} C_i \geq 1] = \frac{P[C_1,\ldots,C_n]}{P[\sum_{i=1}^{n} C_i \geq 1]} = \frac{P[C_1,\ldots,C_n]}{P[C_i=1]} \times \frac{P[C_i=1]}{P[\sum_{i=1}^{n} C_i \geq 1]}$$

$$= P[C_1,\ldots,C_n|C_i=1] \times \frac{P[C_i=1]}{P[\sum_{i=1}^{n} C_i \geq 1]} \text{if } C_i = 1$$

Distribution we sample from

The target distribution is proportional to the distribution we sample from. Assuming $P[C_i = 1]$ are equal for all subjects i = 1…*n*, we estimate it by averaging the simulations over the conditionning subjects.

Example of averaging over conditioning subject

Consider configuration S: $C_1 = C_2 = C_3 = 1, C_4 = 0$

Results of simulations conditioning on each subject carrying the variant:

$C_1 = 1$	$C_2 = 1$	$C_3 = 1$	$C_4 = 1$	T_S
1 0 0 0	1 1 0 0	0 0 1 0	0 0 0 1	0
1 1 1 0	0 1 0 1	0 0 1 0	0 0 0 1	1
1 0 0 0	0 1 0 0	0 1 1 0	1 1 1 1	1
1 1 0 0	0 1 1 0	0 0 1 0	0 0 0 1	0
1 0 1 0	0 1 0 0	1 1 1 0	0 0 1 1	1
$N_{S1} = 1$	$N_{S2} = 0$	$N_{S3} = 1$	$N_{S4} = 0$	

Probability estimate: $(N_{S1} + N_{S2} + N_{S3})/(4x5) = 0.1$

Estimating sharing configuration distribution

For each configuration S, estimate of probability is given by

$$\hat{P}_S = \frac{1}{n \times n_r} \sum_{i \in S} N_{Si}$$

With Monte Carlo standard error estimated by

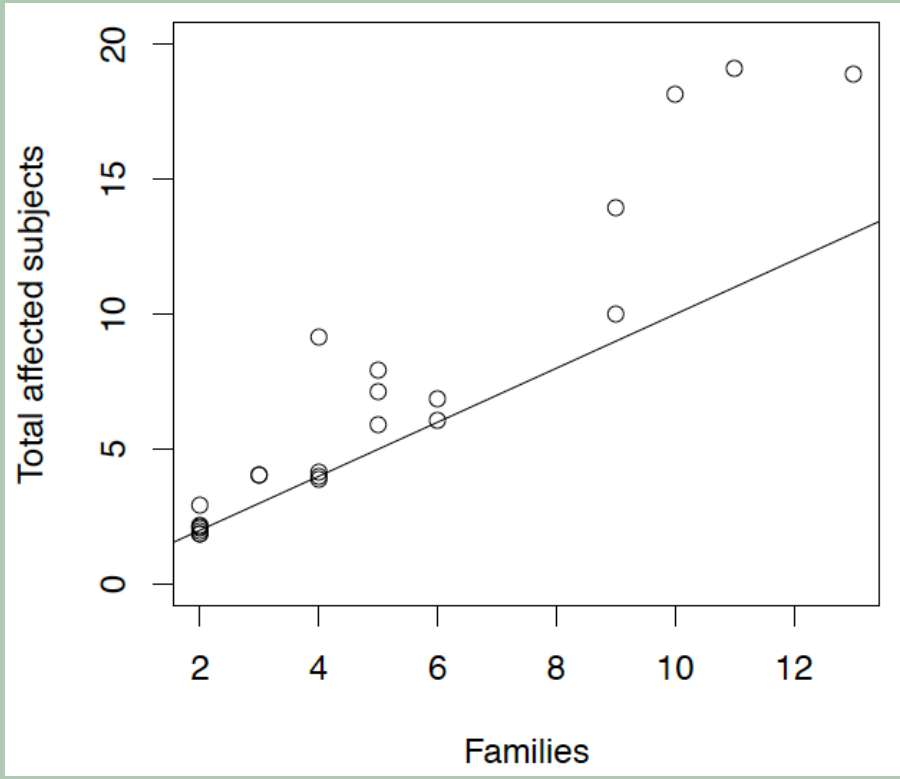
$$SE(\hat{P}_S) = \frac{1}{\sqrt{n \times n_r}} SD(T_S)$$

Setting number of replicates

- Simple example with 5 affected subjects from 2 families from the SLSJ asthma study.
- Estimated probability a particular configuration S of 3 affected sujects from the 2 families share a RV: 0.0019
- $SD(T_S) = 0.099$
- With $n_r = 10^5$, get SE = 0.00013 (7% of estimate value)

Computing time issue

- Simulations conditional on each subject ran in parallel
- Genealogy size grows rapidly with number of affected families, so does gene dropping time
- Trimming genealogy according to ancestor carrying variant speeds up gene dropping in GENLIB package (Gauvin et al. 2015), but trimming costs time.



Sample size in replicates of disease simulation in SLSJ asthma family sample

Conclusion

- Genealogies highly informative for inferring IBD sharing of rare variants…
- But computing cost remains a serious issue…

Acknowledgements

- Catherine Laprise (UQAC) and the participants of the SLSJ asthma family study for phenotype, pedigree structures and genealogical data.
- Jordie Croteau, Thomas Sherman and Saeed Sabbah for their programming work
- The BALSAC project for access to genealogical data
- Funding
  - Fonds de recherche du Québec, Santé
  - NIH R03-DE-02579
  - CANSSI Collaborative Research Team 8

References

- Bureau, A. et al. (2014) Inferring rare disease risk variants based on exact probabilities of sharing by multiple affected relatives. *Bioinformatics* **30**(15): 2189-2196.
- Gauvin, H. et al. (2015) GENLIB: an R package for the analysis of genealogical data. BMC Bioinformatics 16:160.
- Sherman T, Fu J, Scharpf RB, Bureau A, Ruczinski I (2018). Detection of rare disease variants in extended pedigrees using RVS. *Bioinformatics* (in revision).

Exome sequencing in the Eastern Quebec SZ and BD kindred study

- 24 sequenced subjects (7 with schizophrenia, 17 with bipolar disorder).
- 11 families (9 with 2 subjects, 2 with 3 subjects)
- Selection informed by previous linkage studies (Maziade et al. 2005, Mol Psychiatry. **10**: 486-99).
- Exome capture and sequencing using Illumina Hi-Seq at the Genome Quebec McGill Innovation Centre.
- Data processing using the Genome Quebec dnaseq pipeline, following Broad Institute best practice guidelines.

Monte Carlo alternative to numerical approximation

1. Sample A from P[A].
2. Sample which of the $A = a$ alleles is the rare variant.
3. If rare variant is among first $2n_f - a$ alleles then it is introduced twice
   - sample pair of founders introducing it with uniform probabilities,

   otherwise
   - sample the sole founder introducing it with uniform probabilities.
4. Perform gene dropping simulation down the pedigree.

Estimating the mean kinship among founders

- If founders genotypes are measured, use them to estimate kinship for each pair of founders and take the mean.
- If founders are not genotyped, we express the kinship coefficient between genotyped subjects i_1 and i_2 as

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

Then reverse:

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

and take the mean over all pairs of genotyped subjects.

1st sequencing study of multiplex oral cleft families

- 54 multiplex cleft families ascertained through non-syndromic oral clefts in distant relatives
  - Sequenced 2 affected subjects in 50 families, 3 in 4 families
- Families recruited from Germany, Philippines, India, Syria, Taiwan, China, USA
- Exon capture using Agilent SureSelect
- Sequencing of 100 bp paired-end reads on Illumina Hi-Seq
- Multi-sample variant calling using GATK
- Defined rare SNVs as < 1% frequency in Exome sequencing project (ESP) and 1000 Genomes, and seen in < 20% of families (60,038 exonic and splice site SNVs).

rs117883393 in *ORA2*

- T allele shared in 3 families out of 4 where it occurred
- Population frequency 0.8% in European Americans (ESP)
- 2 Syrian families shared T allele where cryptic relatedness among founders was suspected (estimated mean kinship = 0.013)
- P-value increased from 6.1×10^{-6} to 1.4×10^{-5} after correction for cryptic relatedness (not taking allele frequency into account)

2e étude de séquençage de familles avec fentes labio-palatines

- 54 familles avec cas multiples de fentes labio-palatines
  - De 2 à 6 sujets séquencés par famille, total 155
- Familles recrutées aux Philippines, États-Unis, Guatemala et Syrie
- Séquençage du génome entier par Illumina, mais analyse initiale des SNVs exoniques ou d'épissage
- Rareté définie par fréquence < 1% dans le Exome sequencing project (ESP), Exome Aggregation Consortium (ExAC) et 1000 Génomes (73 000 SNVs)

Test allowing for sharing of a RV by subset of affected relatives

- Single family m:
  - Number of subjects sharing RV:
  - RV sharing configuration:

$$K = C_1 + \cdots + C_n$$

$$G_{k_m} = (C_1, \ldots C_{n_m})_m \mid K_m = k_m$$

With probability

$$P_{G_{k_m}} = P(G_{k_m} \mid K_m \geq 1)$$

- M families:

$$G_k = (G_{k_1}, \ldots G_{k_M}) \quad g = (g_1, \cdots g_M)$$

$$k = \sum_{m=1}^{M} k_m$$

- P-value

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