Institut universitaire
en santé mentale
de Québec

Analyse des études génétiques familiales : établir le lien variants rares - maladie et intégrer l'expression des gènes

Alexandre Bureau

Plan de la présentation

- ✓ **Variation génétique et séquençage à haut débit**
- Analyse de variants génétiques rares dans les familles basée sur les probabilités de partage de variants
- Calcul des probabilités de partage
- Étude de puissance de l'approche de partage de variants rares
- Intégrer l'expression des gènes aux analyses génétiques dans les familles

Variation dans la séquence d'ADN

		Génotypes :
Individu 1	CTCCGAATACGAATGGCCGT CTCCGAATAGGATTGGCCGT	C/G, A/T
Individu 2	CTCCGAATACGAATGGCCGT CTCCGAATAGGAATGGCCGT	C/G, A/A
Individu 3	CTCCGAATACGATTGGCCGT CTCCGAATACGATTGGCCGT	C/C, T/T

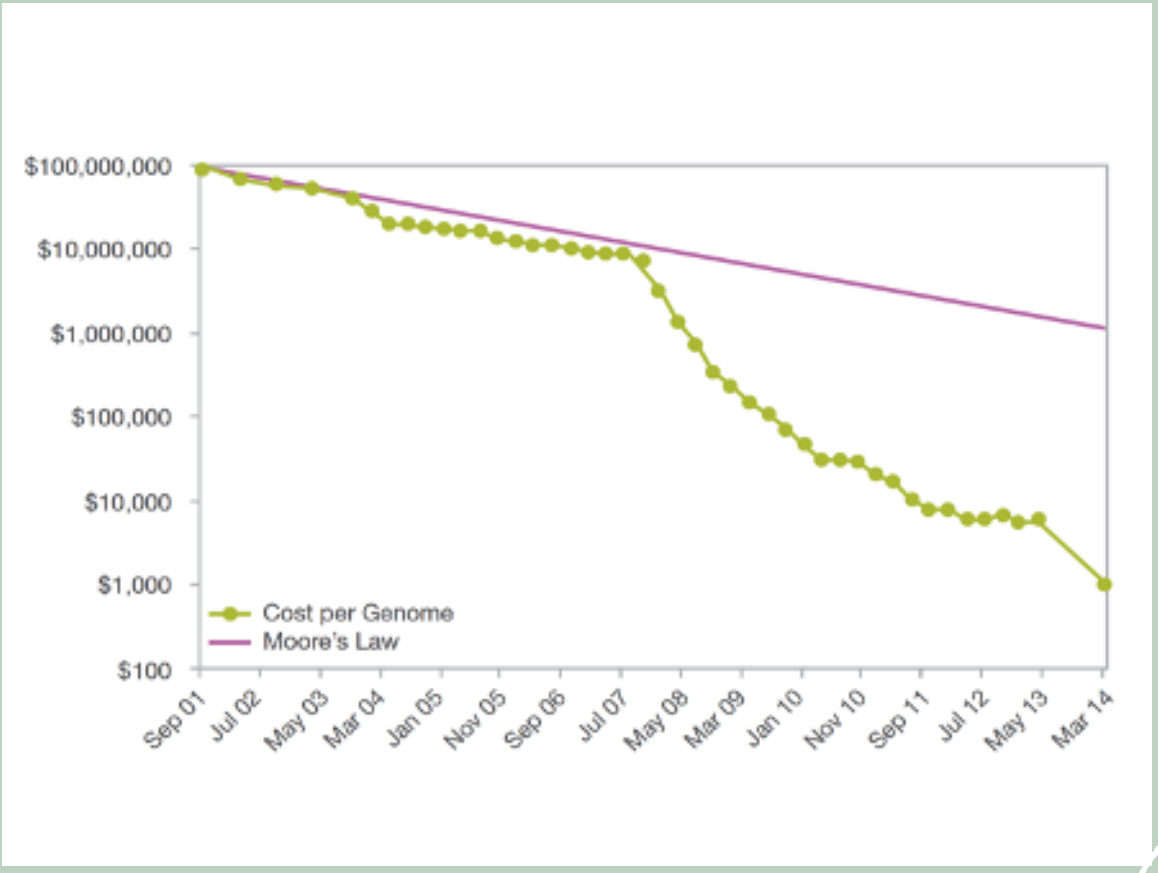
SNV = Single Nucleotide Variant
(Variant de nucléotide), identifiés par no. rsXXX

L'évolution du séquençage

- Déterminer la séquence d'ADN par la méthode Sanger était un processus coûteux et fastidieux, réservé aux grands projets de génomique ou à des régions ciblées dans les études de maladies.
- À partir de 2007, diverses compagnies de génomique ont introduit des méthodes de séquençage de nouvelle génération atteignant des débits beaucoup plus élevés que la méthode Sanger.
- Cela a permis l'avènement d'études de séquençage à grande échelle visant la détection des variants rares impliqués dans des maladies.

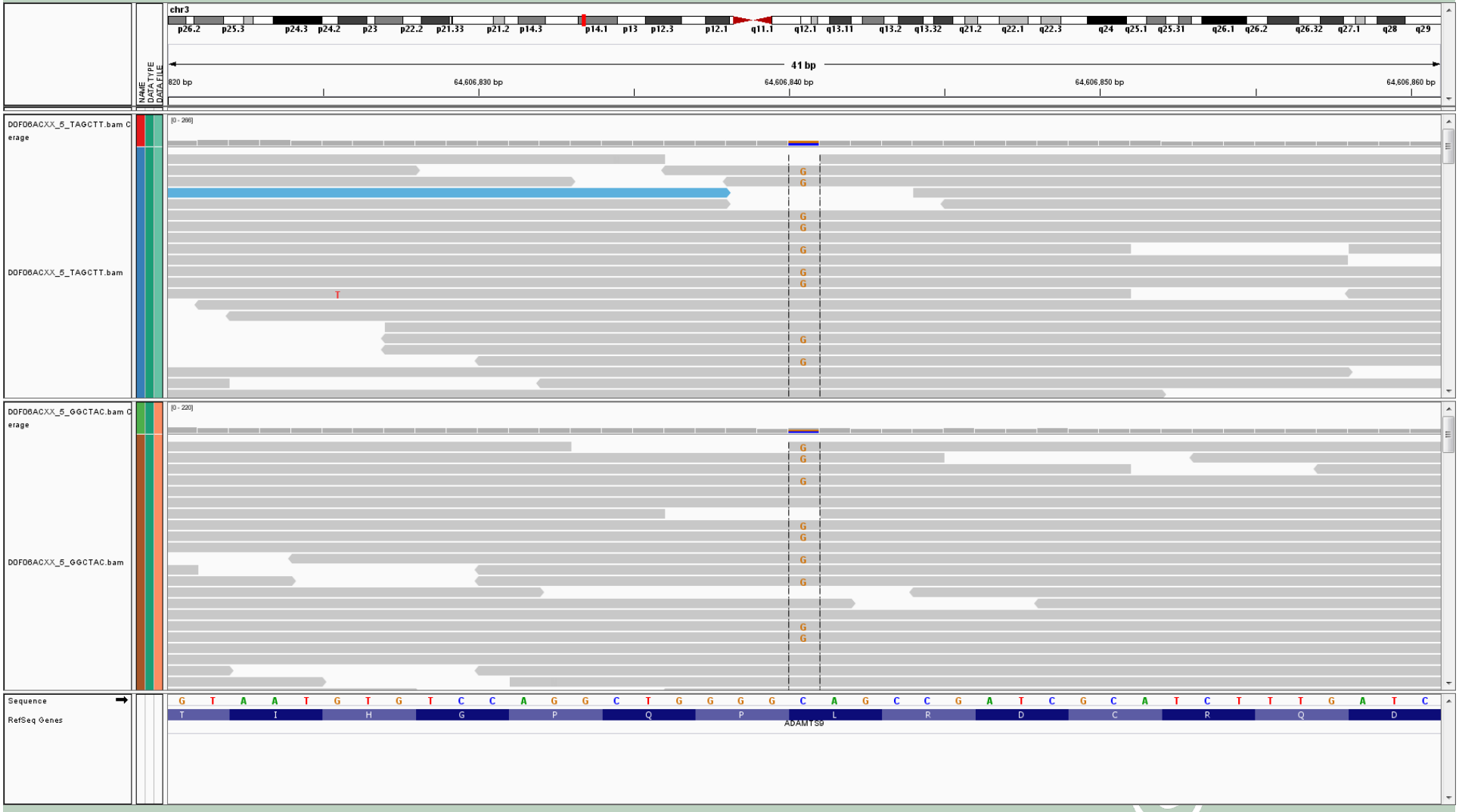
La séquence du génome humain pour 1000$!

Évolution du coût de séquençage d'un génome humain

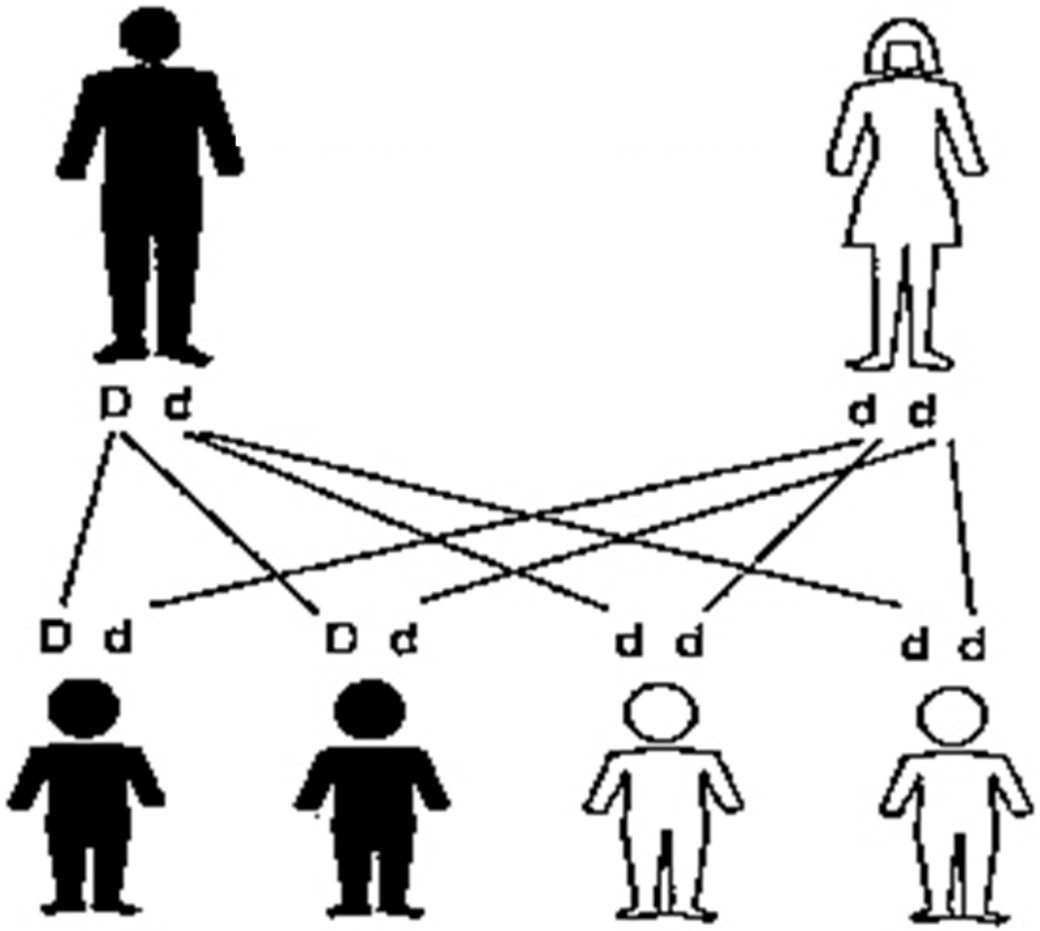


Source: Illumina

Résultat de l'alignement ds Integrated Genome Viewer



Transmission génétique

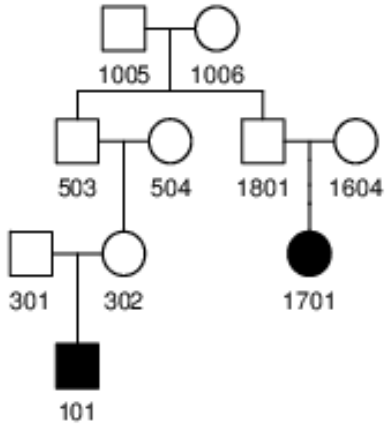


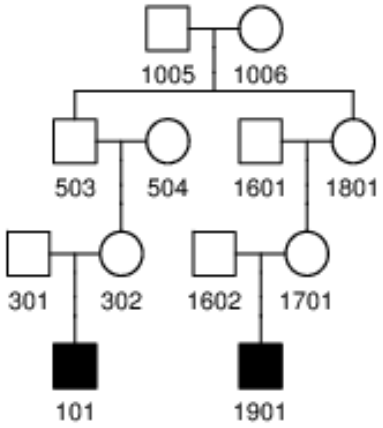
Source: A. Labbe

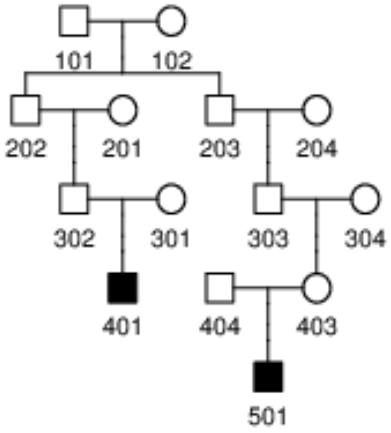
Plan de la présentation

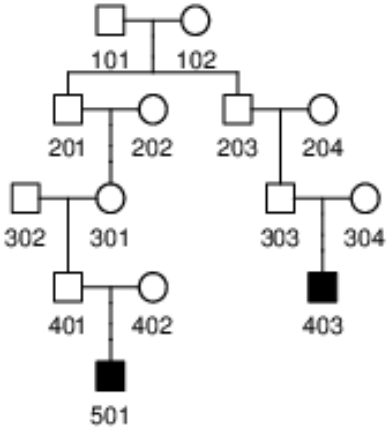
- Variation génétique et séquençage à haut débit
- ✓ **Analyse de variants génétiques rares dans les familles basée sur les probabilités de partage de variants**
- Calcul des probabilités de partage
- Étude de puissance de l'approche de partage de variants rares
- Intégrer l'expression des gènes aux analyses génétiques dans les familles

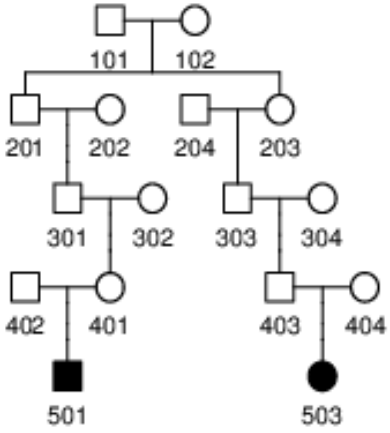
Séquençage d'apparentés distants

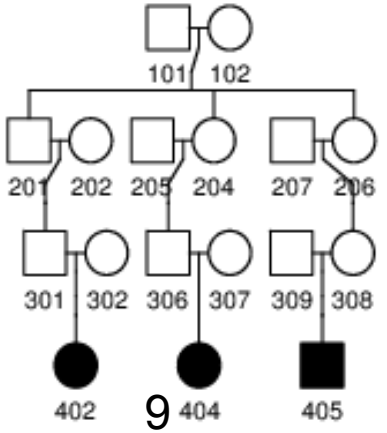












Whole exome sequencing study of multiplex cleft families

- 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).

BIOINFORMATICS ORIGINAL PAPER

2014, pages 1–8
doi:10.1093/bioinformatics/btu198

Genetics and population analysis

Advance Access publication April 16, 2014

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

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] and Ingo Ruczinski[3,*]

[1]Centre de Recherche de l'Institut Universitaire en Santé Mentale de Québec, G1J 2G3, [2]Département de Médecine Sociale et Préventive, Université Laval, Québec, G1V 0A6 Canada, [3]Department of Biostatistics, [4]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, [5]Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, MD 21224, [6]Department of Oral Biology, Center for Craniofacial and Dental Genetics, School of Dental Medicine, University of Pittsburgh, PA 15219, [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 and [9]Dr. Hejazi Clinic, P.O. Box 2519, Riyadh 11461, Saudi Arabia

Associate Editor: Jeffrey Barrett

HIGHLIGHTED ARTICLE
INVESTIGATION

Whole Exome Sequencing of Distant Relatives in Multiplex Families Implicates Rare Variants in Candidate Genes for Oral Clefts

Alexandre Bureau,[*] Margaret M. Parker,[†] Ingo Ruczinski,[‡] Margaret A. Taub,[‡] Mary L. Marazita,[§] Jeffrey C. Murray,[**] Elisabeth Mangold,[††] Markus M. Noethen,[††] Kirsten U. Ludwig,[††] Jacqueline B. Hetmanski,[†] Joan E. Bailey-Wilson,[‡‡] Cheryl D. Cropp,[‡‡] Qing Li,[‡‡] Silke Szymczak,[‡‡] Hasan Albacha-Hejazi,[§§] Khalid Alqosayer,[***] L. Leigh Field,[†††] Yah-Huei Wu-Chou,[‡‡‡] Kimberly F. Doheny,[§§§] Hua Ling,[§§§] Alan F. Scott,[****] and Terri H. Beaty[†,1]

*Centre de Recherche de l'Institut Universitaire en Santé Mentale de Québec and Département de Médecine Sociale et Préventive, Université Laval, Québec, QC G1V 0A6, Canada, [†]Department of Epidemiology and [‡]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland 21205, [§]Department of Oral Biology, School of Dental Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania 15219, **Department of Pediatrics, School of Medicine, University of Iowa, Iowa City, Iowa 52242, [††]Institute of Human Genetics, University of Bonn, Bonn, Germany D-53111, [‡‡]Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore Maryland 21121, [§§]Hejazi Clinic, Riyadh, Saudia Arabia 11461, ***Prime Health Clinic Jeddah, Riyadh, Saudi Arabia 21511, [†††]Department of Medical Genetics, University of British Columbia, Vancouver, Canada V6T1Z3, [‡‡‡]Laboratory of Human Molecular Genetics, Chang Gung Memorial Hospital, Taoyuan, Taiwan 333, [§§§]Center for Inherited Disease Research and ****Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland 21224

Méthode basée sur le partage de variants rares

- Sachant qu'on observe un variant rare chez un sujet séquencé, il y a une probabilité qu'on l'observe aussi chez son ou ses apparentés séquencés
  - Cette probabilité dépend seulement de leur degré d'apparentement.
  - Elle est calculées sous l'hypothèse que le variant n'est ni lié ni associé à la maladie (hypothèse nulle).
- La probabilité que tous les sujets atteints partagent un variant dans la ou les familles où il est détecté est une valeur p pour cette hypothèse nulle.
- Même principe que l'analyse de liaison basée sur le partage d'allèles identiques (identity-by-descent ou IBD)

Probabilités de partage sous l'hypothèse nulle

	Probabilité de partage	
	IBD	Variant rare
Formule:	$1/2^{D-1}$	$1/(2^{D+1} - 1)^{*}$
Relation		
Parent-enfant (D=1)	1	1/3
Oncle-neveu (D=2)	1/2	1/7
Cousins germains (D=3)	1/4	1/15
Petits cousins (D=5)	1/16	1/63

D: degré de parenté

*Feng et al., 2011. PLoS One 6, e23221

Plan de la présentation

- Variation génétique et séquençage à haut débit
- Analyse de variants génétiques rares dans les familles basée sur les probabilités de partage de variants
- ✓ **Calcul des probabilités de partage**
- Étude de puissance de l'approche de partage de variants rares
- Intégrer l'expression des gènes aux analyses génétiques dans les familles

Problèmes à résoudre

1. Calcul de la probabilité de partage d'un variant rare par n=3+ apparentés et 2+ apparentés de familles complexes.
2. Calcul de la probabilité de partage d'un variant rare par k parmi n>k apparentés.

1. 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 RV into the pedigree

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

Expression in a special case

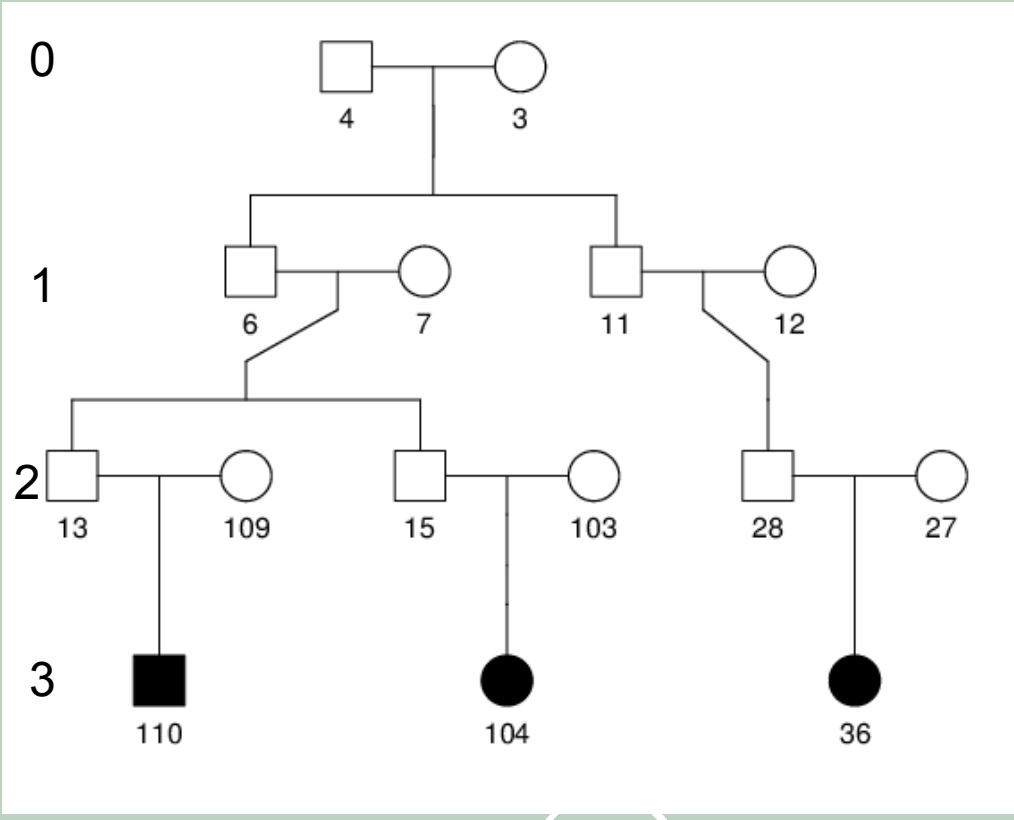
- Sequenced subjects descend from common founder couple through independent lines of descent

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

Where D_{ij} is degree of relatedness between sequenced subject i and founder j, $d(j)$ is subset of sequenced subjects descending from founder j and D_f is the sum of D_{ij} for a common founder

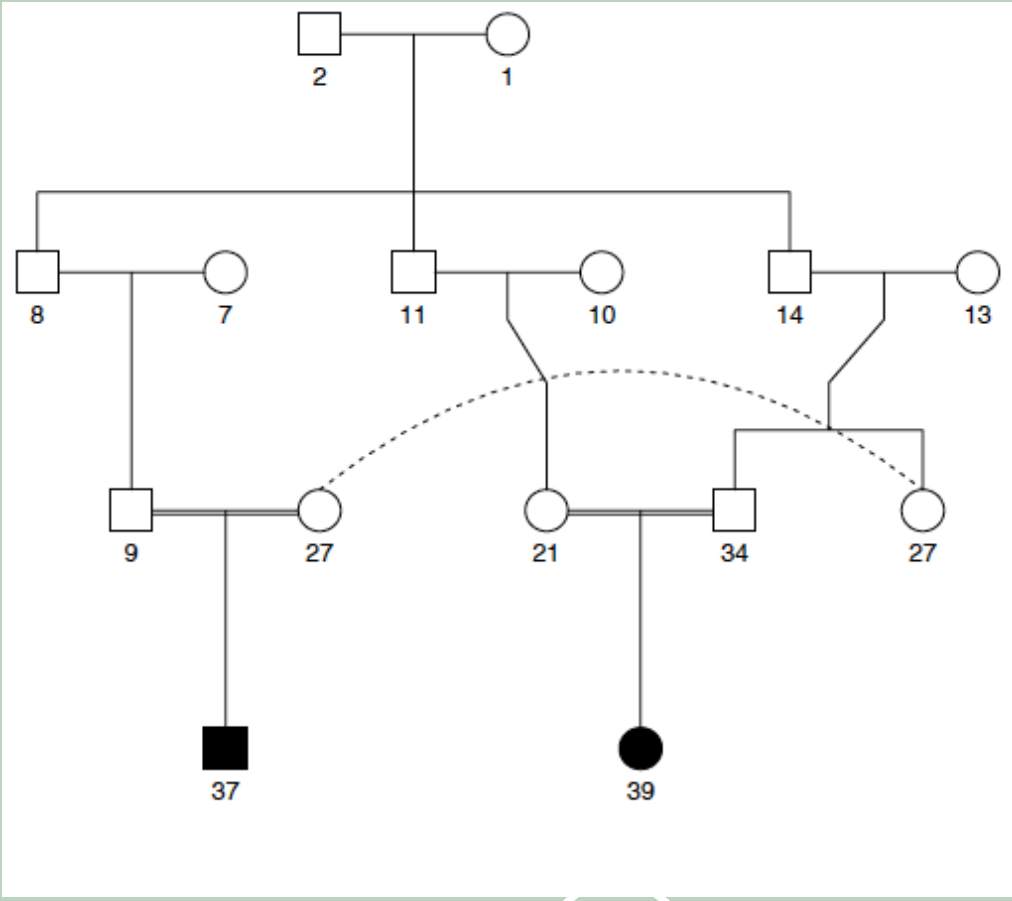
Calcul exact sur arbres généalogiques simples

- Entrées: objet de classe S3 "pedigree" (liste de vecteurs)
- Attribuer profondeur de chaque sujet.
- De profondeur R à 0, incrémenter nombre de générations des branches de l'arbre, combinant les branches qui se rejoignent.
- Sortie: classe S4 "RVsharingProb ».



Approximation Monte Carlo pour arbres complexes

- Définition d'une classe S4 "trio" (parents et enfant)
- Codage de l'arbre en liste récursive de trios
- Introduit un variant chez un fondateur
- Simule transmission avec prob. ½ de parent à enfant



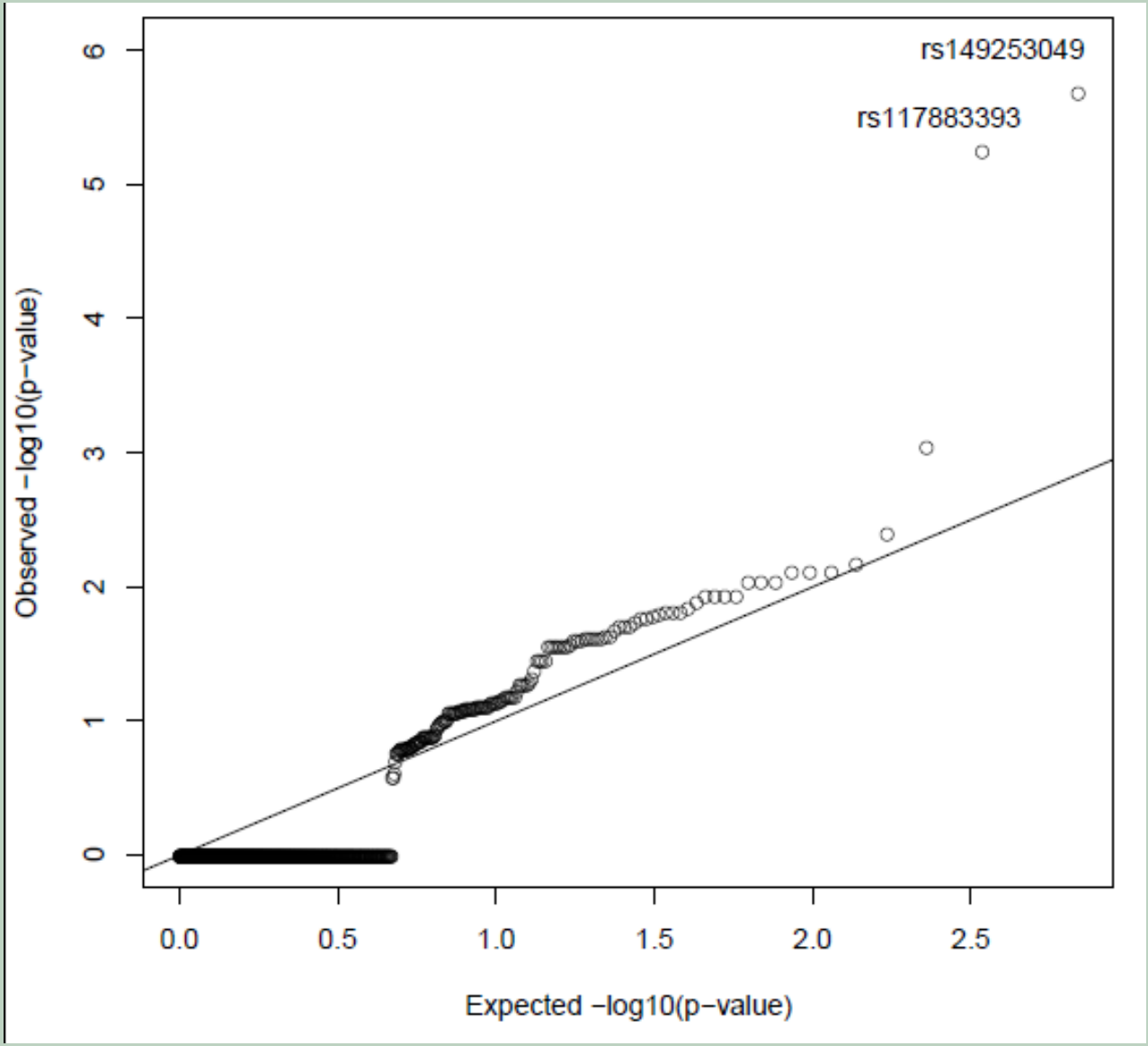
Implantation

- Module R RVsharing
  - Calcul exact sur arbres généalogiques simples
  - Approximation Monte Carlo pour arbres généalogiques complexes

Probabilités de partage observées vs. attendues

Étude des fentes labio-palatines

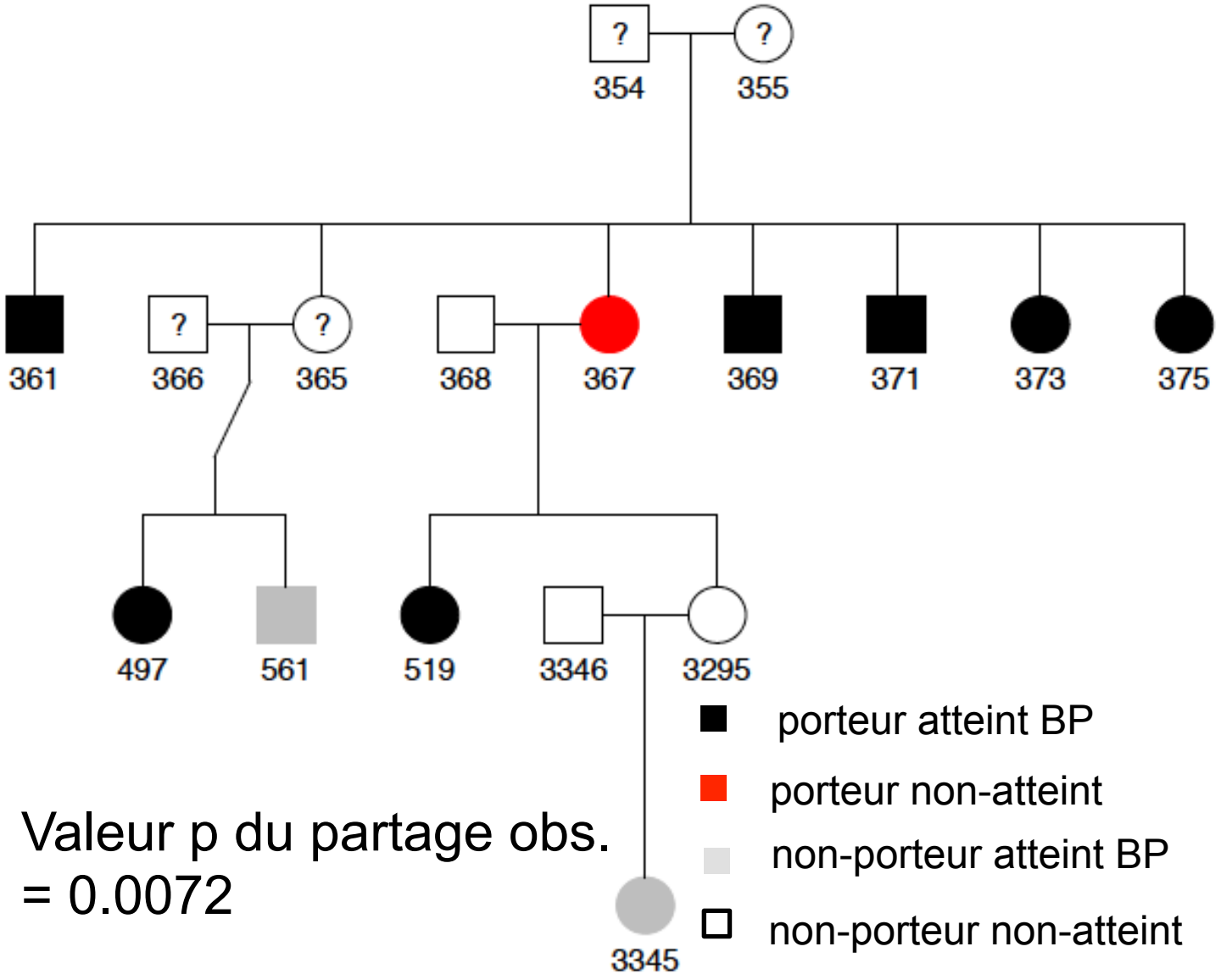
Sous-ensemble de 2 355 variants avec puissance suffisante.



2. Partage par sous-ensemble des atteints

- Hétérogénéité intra-familiale, erreur de diagnostic
  - Variant causal absent de certains atteints
- Avec grand nombre d'atteints dans une famille, partage par un sous-ensemble d'atteint est informatif

Ex: famille québécoise avec trouble bipolaire



Valeur p du partage obs. = 0.0072

Partage d'un variant rare par k parmi n>k sujets

Pour le numérateur $P[C_1=\ldots=C_k=1,C_{k+1}=\ldots=C_n=0]$, on se sert des relations:

$P[A,B,C,D^c] = P[A,B,C] - P[A,B,C,D]$

$P[A,B,C^c,D^c] = P[A,B] - P[A,B,C] - P[A,B,D] + P[A,B,C,D]$

…

Chacun des termes calculés avec algorithme pour partage par tous les sujets appliqué à k, …, n sujets.

Dénominateur $P[C_1+\ldots+C_n>=1]$ inchangé.

Disponible dans prochaine version du module RVsharing.

Plan de la présentation

- Variation génétique et séquençage à haut débit
- Analyse de variants génétiques rares dans les familles basée sur les probabilités de partage de variants
- Calcul des probabilités de partage
- ✓ **Étude de puissance de l'approche de partage de variants rares**
- Intégrer l'expression des gènes aux analyses génétiques dans les familles

Power assessment

Simulation setup

Variant frequency = 0.0001

Dominant effect

Disease prevalence = 1%

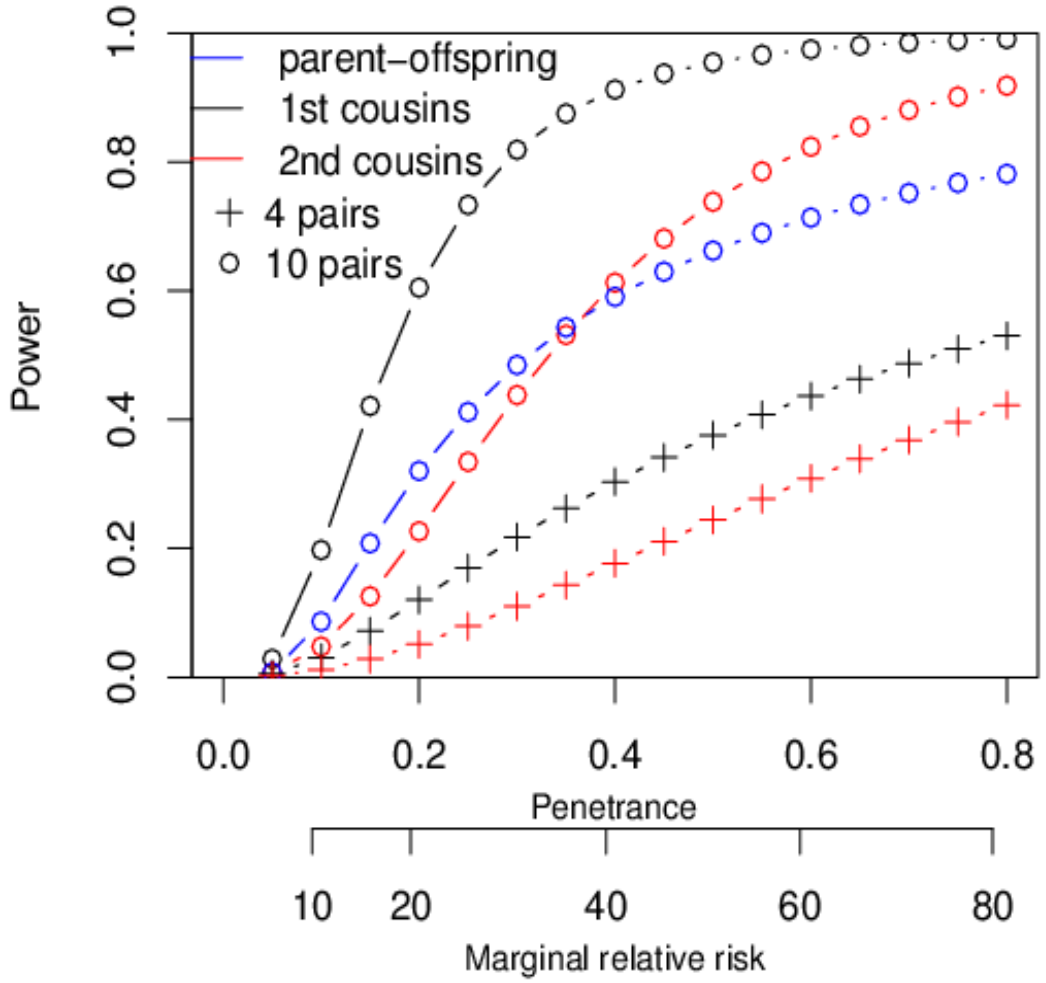
Heterogeneity model

Recurrence risk

1st cousins: 2

2nd cousins: 1.25

$\alpha = 2.2 \times 10^{-5}$



Alternative: comparer avec un groupe témoin en plus

- En plus du partage par apparentés atteints, compare fréquence du variant chez les atteints p_1 à fréquence dans un groupe témoin p_0.
- Avantage:
  - utilise information de tous les atteints porteurs (incluant atteints non-apparentés)
- Désavantages:
  - Groupe témoin doit être de même origine ethnique que le groupe des cas, car les fréquences de variants fluctuent d'une population à l'autre. Échantillons de référence (1000 génomes, Exome sequencing project) souvent inappropriés.
  - Le nombre de variants considéré inclut tous les variants observés, incluant ceux seulement chez les témoins (pénalité plus grande pour la mutiplicité des tests)

Simulation

- 100 familles avec paire de petits cousins atteints + 1000 témoins non-atteints non reliés
- Variant avec risque relatif = 50, de fréquence = 0,001
- Prévalence de la maladie dans la population = 1%
- Compare
  - test de probabilité de partage
  - test association + liaison et test de liaison seule (modèle dominant) implanté dans pVAAST (Hu et al. Nat Biotechnol, 2014. **32**(7): p. 663-9.)
- 100 réplicats

Comparaison de puissance

Test statistique	Puissance*
Test exact de partage de variant rare	54%
Test cas-témoin + liaison (pVAAST CLRT)	100%
Test liaison seulement (pVAAST LOD)	19%

* Seuil $\alpha = 1 \times 10^{-5}$

- Pour CRLT, puissance reste 100% jusqu'à $\alpha = 10^{-15}$
- Score LOD calculé par pVAAST capture liaison seulement, pas le partage d'un allèle spécifique.

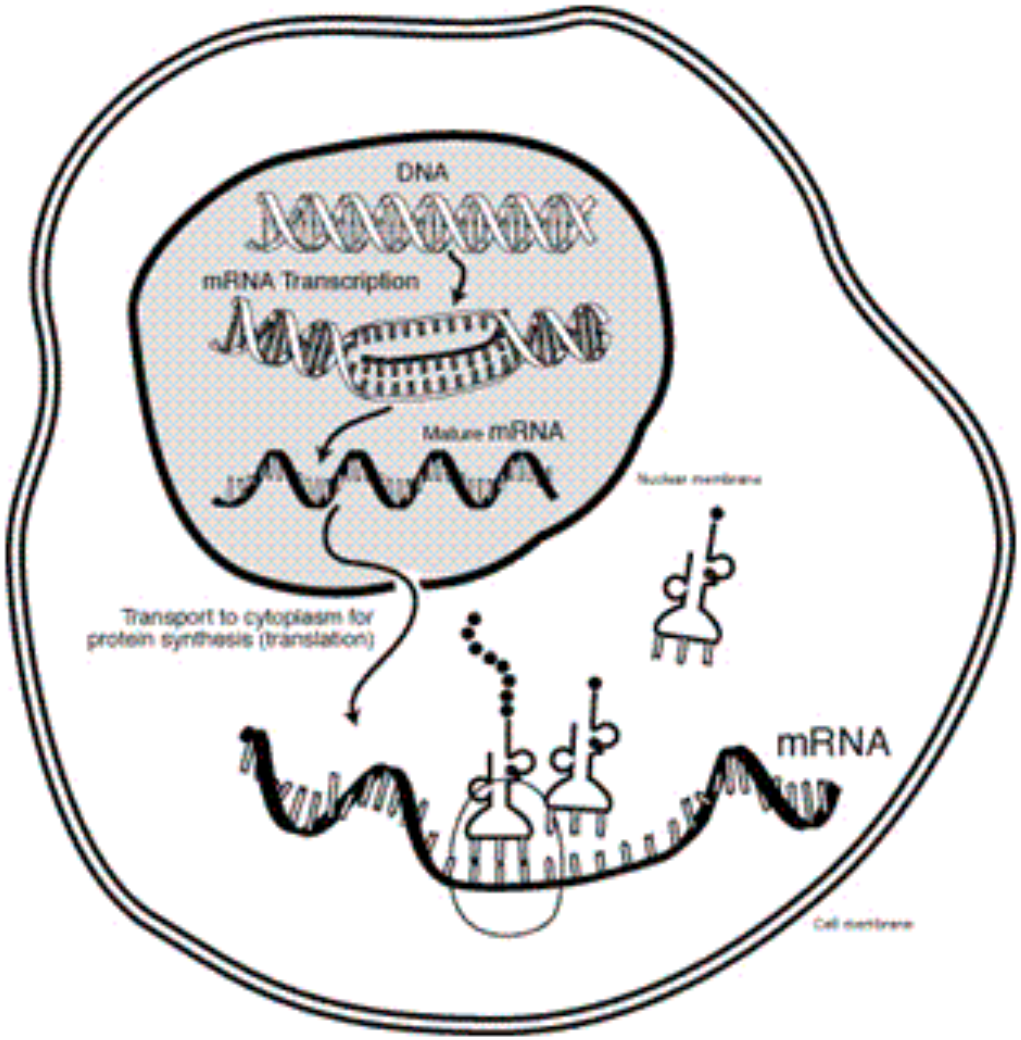
Extension: analyse au niveau du gène

- Analyse d'un variant à la fois est limitée aux variants retrouvés dans plus d'une famille (ou dans une très grande famille).
  - Dans l'étude des fentes labiales, 57 583 des 60 038 variants rares n'ont pas pu être analysés faute de puissance.
- Analyse simultanée de tous les variants d'un même gène permet d'inclure variants vus dans une seule famille.
- Problèmes à surmonter:
  - Variants de risque noyés parmi variants bénins
  - Traitement de plusieurs variants dans une même famille

Plan de la présentation

- Variation génétique et séquençage à haut débit
- Analyse de variants génétiques rares dans les familles basée sur les probabilités de partage de variants
- Calcul des probabilités de partage
- Étude de puissance de l'approche de partage de variants rares
- ✓ **Intégrer l'expression des gènes aux analyses génétiques dans les familles**

L'expression des gènes



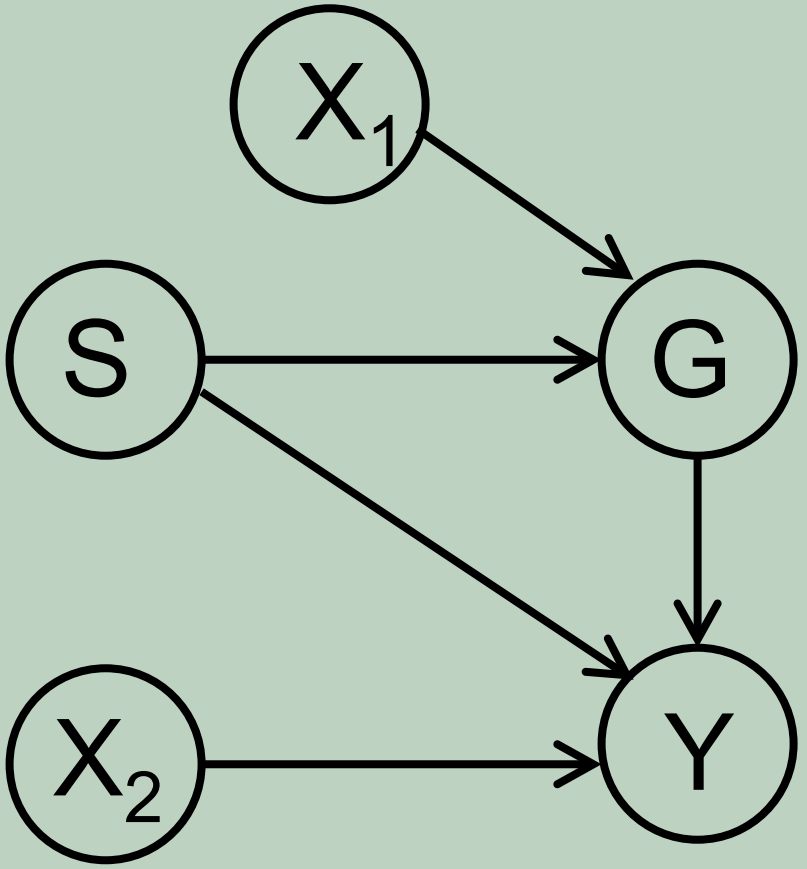
Source: A. Labbe

Eastern Quebec major psychosis kindreds

- 48 multigenerational kindreds densely affected by schizophrenia (SZ) and bipolar disorder (BP).
- Ascertainment criterion: at least 4 relatives with SZ or BP.
- Over 1000 subjects with diagnosis, gene expression and genotype.
- Gene expression measured on immortalized lymphocytes using Illumina chip (19 000 expressed transcripts).
- 720 000 SNVs genotyped using Illumina Omni Express chip.

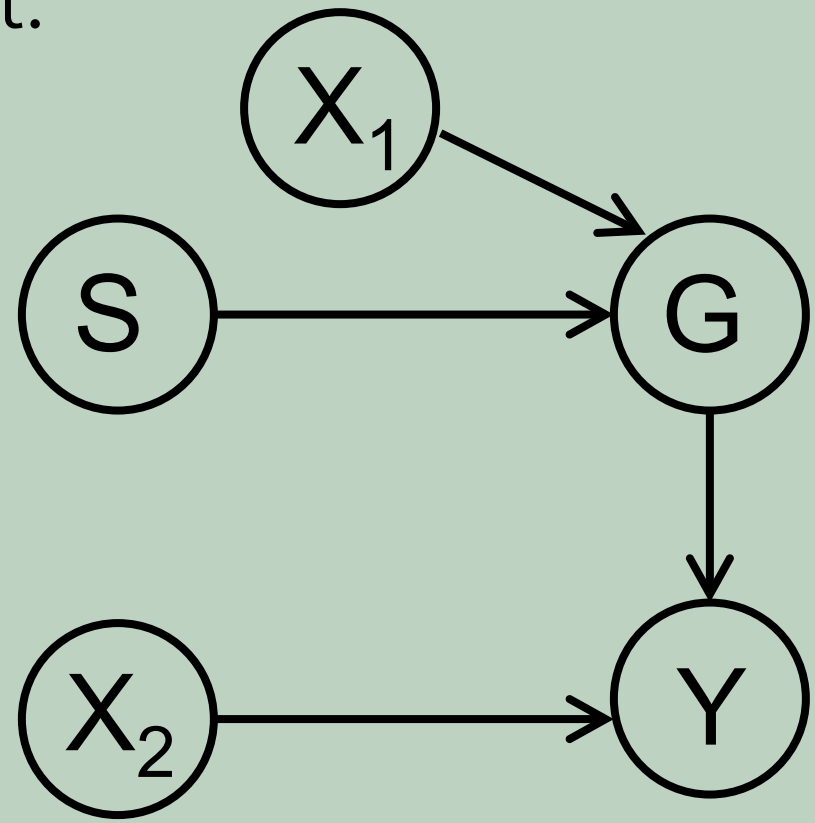
Représentation graphique

- S: Génotype (SNVs)
- G: Expression de gène
- Y: Maladie (oui/non)
- X: Covariables



Objectif

- Détecter SNVs impliqués indirectement dans la maladie via l'expression de gènes sans que l'association avec la maladie soit détectable directement.



Approche de Zhao et al.

- Applicable à des sujets non apparentés
- Modèle de la maladie

$$g(E[Y_i|G_i,X_{1i}]) = \alpha_0 + G_i^T\alpha_G + X_{1i}^T\alpha_X$$

Peut inclure expression de plusieurs gènes

- Modèle de l'expression de gènes

$$G_i^T\alpha_G = \beta_0 + S_i^T\beta_S + X_{2i}^T\beta_X + \varepsilon_i$$

- Estime α et β en résolvant équations d'estimation
- Estimation empirique de la variance

Zhao et al. Biometrics, 2014. **70**(4):881-90

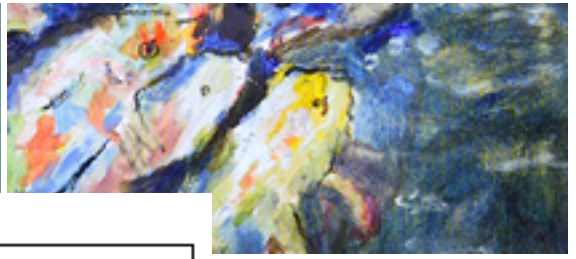
Analyses familiales

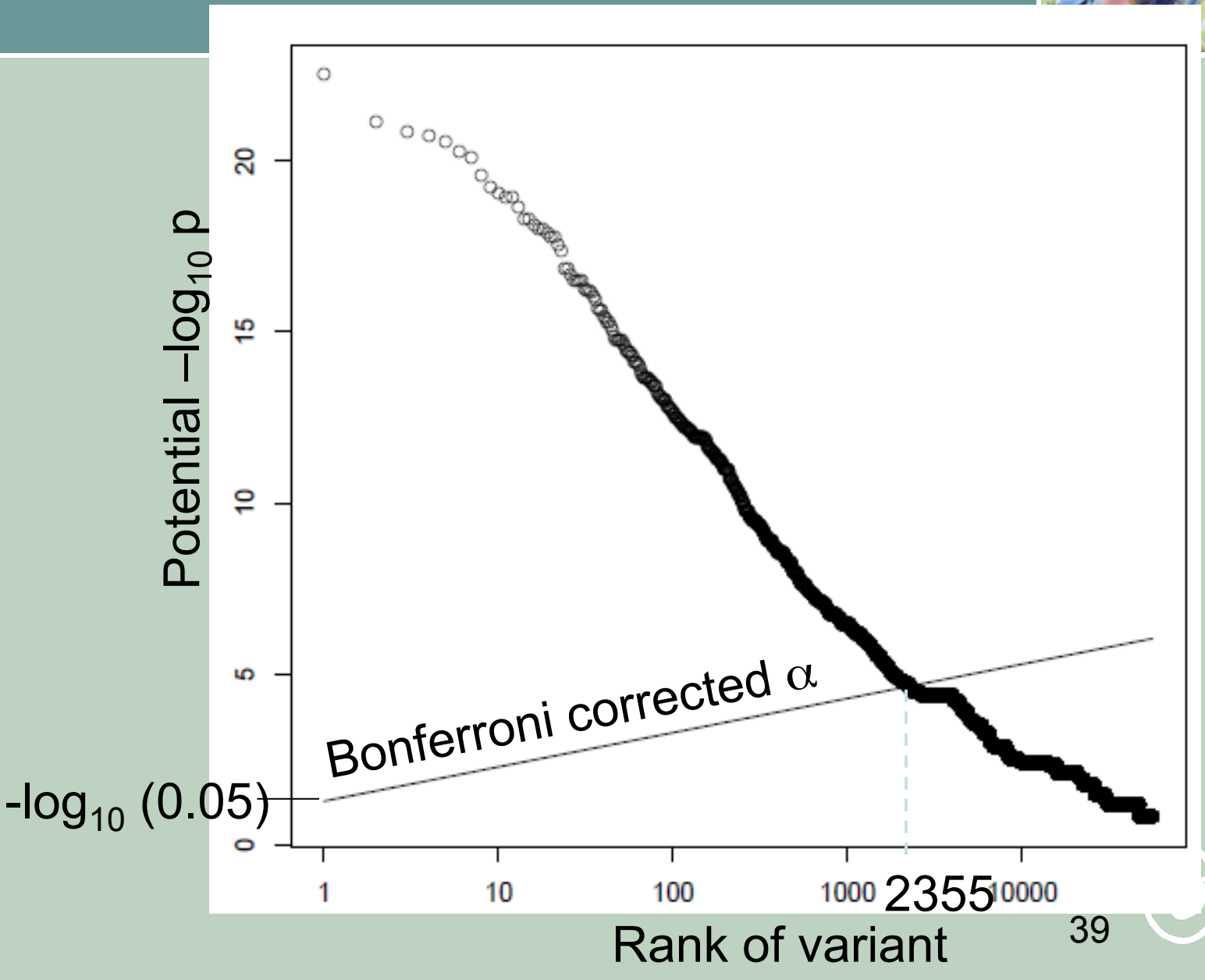
- Appliquer le même modèle que Zhao et al.
  - Tenir compte de l'apparentement dans l'estimation de la variance (estimateur robuste à la dépendance familiale)
- Tenir compte de l'apparentement dans le modèle
  - Modèle linéaire mixte généralisé?
  - Difficile de conditionner sur la condition de recrutement des familles (ex. au moins 4 sujets atteints)
- À la recherche d'étudiants et collaborateurs!

Remerciements

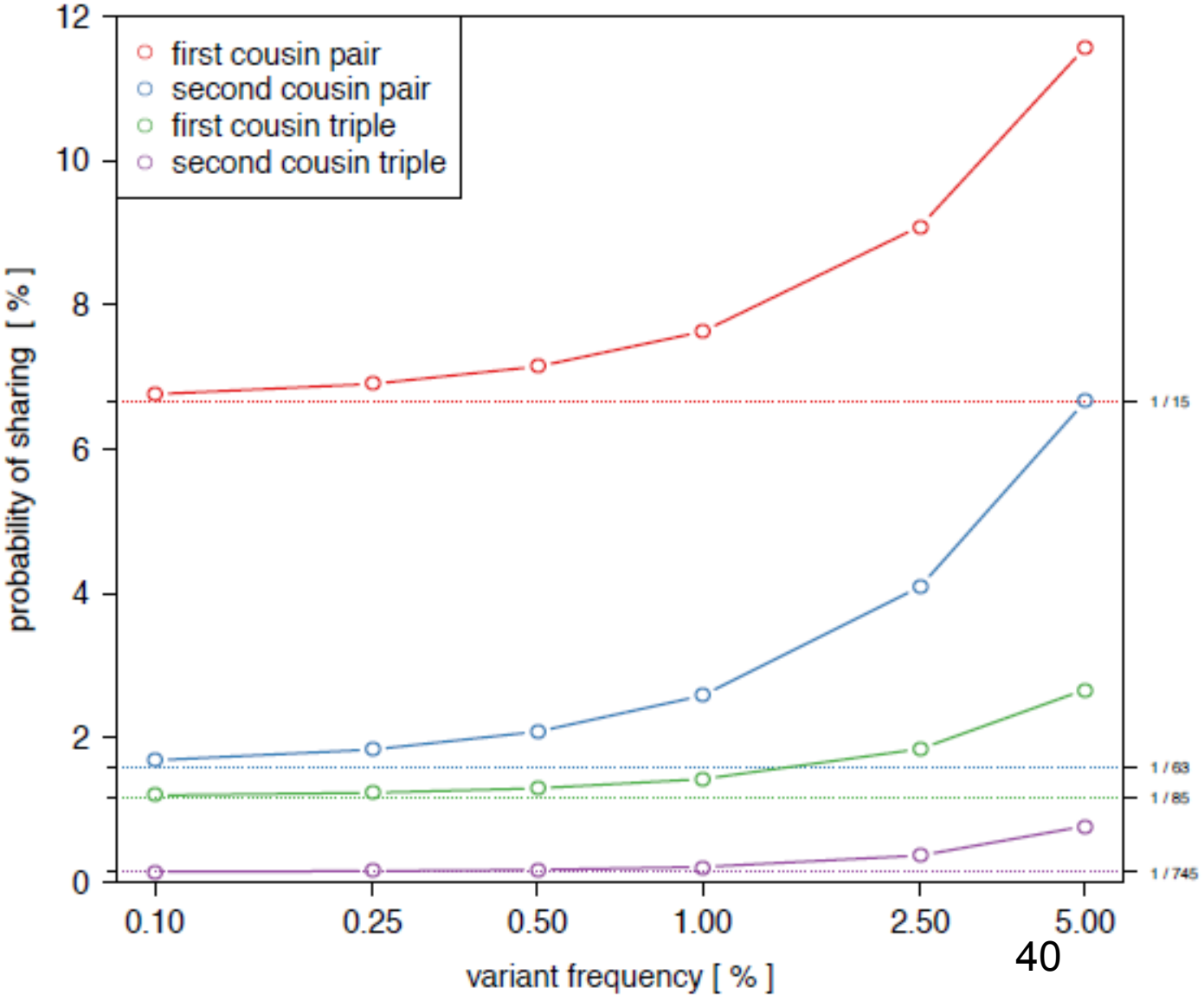
- Financement :
  - Fonds de recherche du Québec, Santé
  - NIH X01-HG-006177, R01-DE-014581, U01-DE-020073
  - Intramural Research Program of NHGRI
  - Deutsche Forschungsgemeinschaft
  - Center for Inherited Disease Research
  - The Eastern Quebec Kindred Study is funded by CIHR (grants MT-12854, MOP-74430 and MOP-1194089) and by a Canada Research Chair (# 950-200810) in the genetics of neuropsychiatric disorders of which Michel Maziade is the Chair.
- Recrutement:
  - Deutsche Selbsthilfevereinigung für Lippen-Gaumen-Fehlbildungen
- Programmation: Jordie Croteau (CRIUSMQ)

Selection of rare SNVs to test





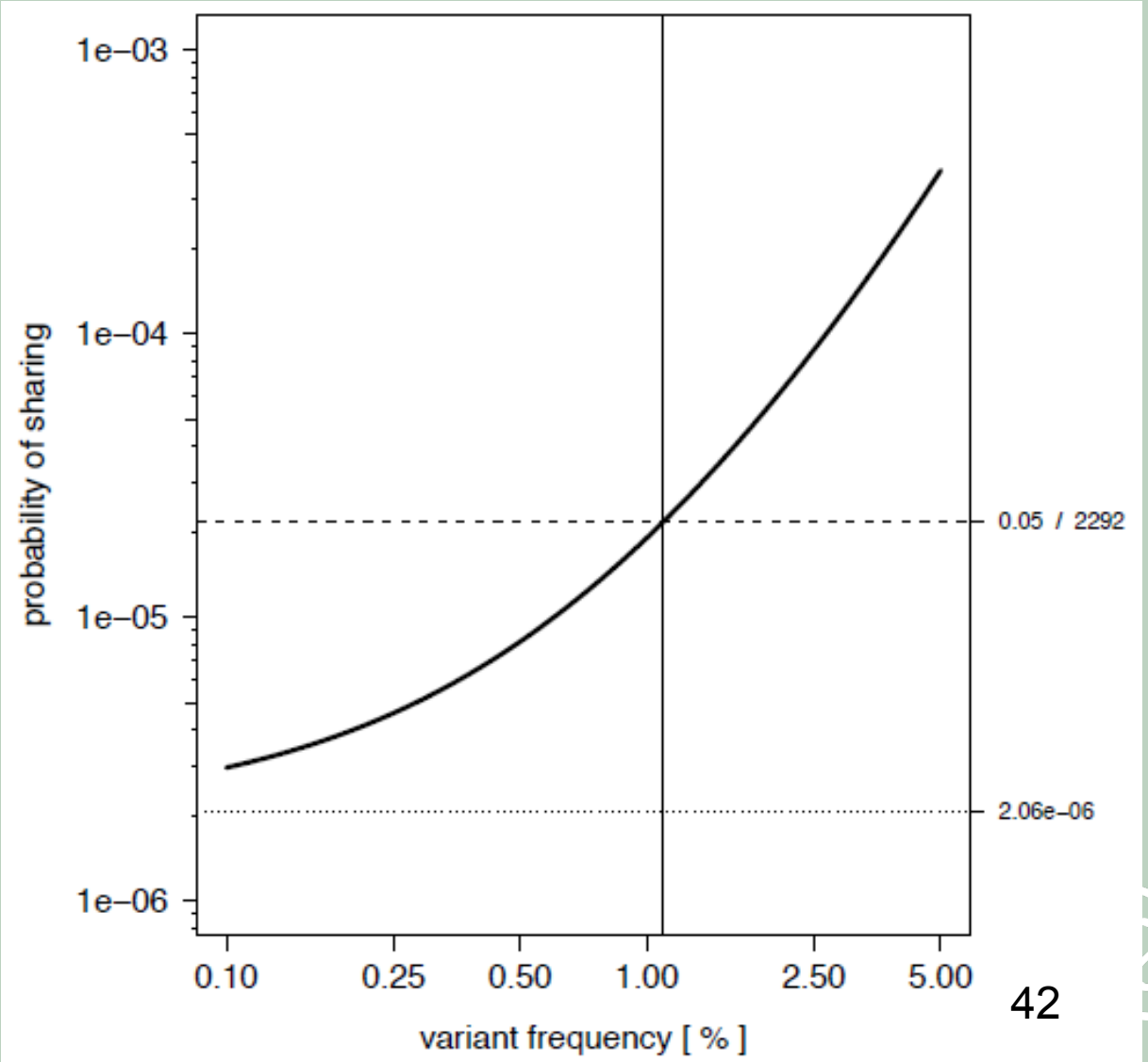
Impact of variant frequency on actual identical-by-state sharing probability



rs149253049 in *ADAMTS9*

- Synonymous triallelic SNV
- G allele shared by 3 relative pairs from India is rarest (not seen in ESP and 1000 Genomes)
- No evidence of relatedness among founders of these 3 families
- Performed assessment of sensitivity to population frequency in an Indian population

Sharing remained significant up to allele frequency of 1.1% for G allele



rs117883393 in *ORA2*

- T allele shared in 3 families out of 4 where it occurred
- Population frequency 0.8% in Caucasians ESP
- 2 Syrian families shared T allele where cryptic relatedness among founders was suspected (estimated mean kinship = 0.0013)
- P-value increased from 6.1×10^{-6} to 1.4×10^{-5} after correction for cryptic relatedness (not taking allele frequency into account)

Poissons solubles
Acrylique sur toile
Jean-Claude Bélanger

Programme d'accompagnement
artistique *Vincent et moi*

Institut universitaire
en santé mentale
de Québec