Comparaison de tests d'association pour des variants génétiques rares

Unité Inserm UMR 1087-CNRS UMR 6291

Elodie PERSYN

Encadrants :
Lise BELLANGER[1]
Christian DINA[2]

[1] Laboratoire de Mathématiques Jean Leray UMR CNRS 6629 , Université de Nantes, France
[2] L'institut du thorax, Inserm UMR 1087 / CNRS UMR 6291, CHU de Nantes, France

PLAN

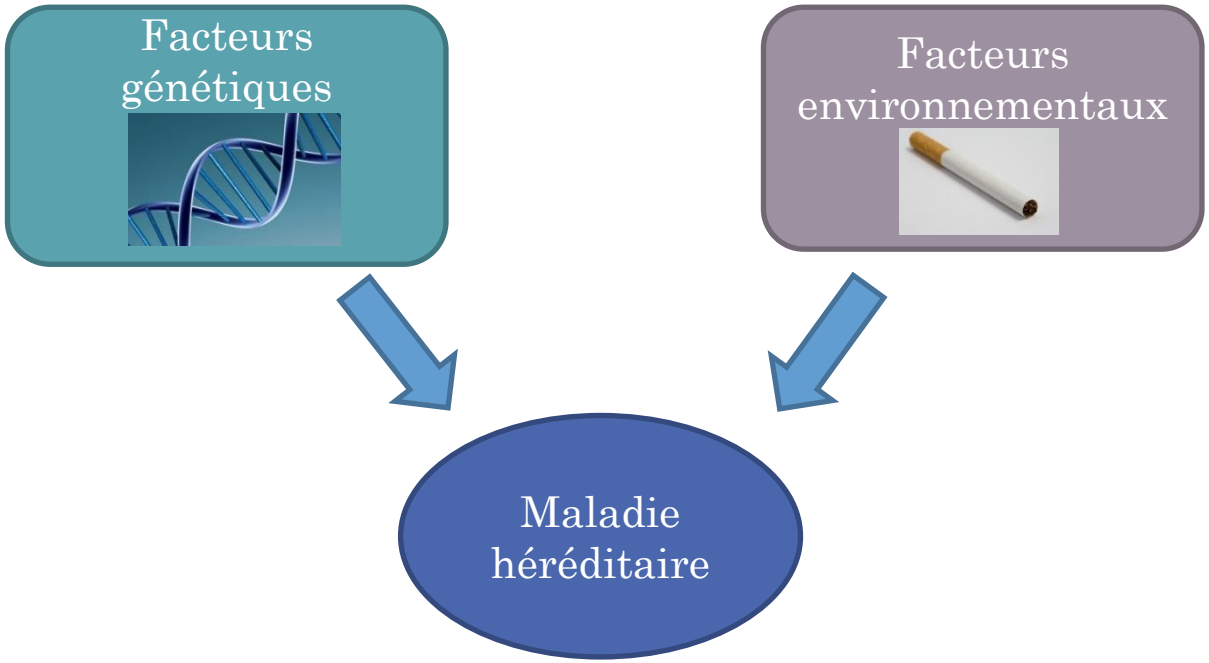
- Contexte
  - Epidémiologie génétique et projet VaCaRMe
  - Les études d'association génétique pour des maladies
  - De l'étude des variants fréquents à l'étude des variants rares
- Tests d'association génétique dans le cadre des variants rares
  - Structure des données
  - Tests « poolés »
  - Propriétés des tests d'association étudiés
- Simulations de données
  - Scénarios génétiques
  - Premières observations
  - Analyses des profils de puissance des méthodes
- Etude de données réelles :le syndrome de Brugada
  - Données
  - Analyse des profils de significativité des gènes
- Discussion

PLAN

- Contexte
  - Epidémiologie génétique et projet VaCaRMe
  - Les études d'association génétique pour des maladies
  - De l'étude des variants fréquents à l'étude des variants rares
- Tests d'association génétique dans le cadre des variants rares
  - Structure des données
  - Tests « poolés »
  - Propriétés des tests d'association étudiés
- Simulations de données
  - Scénarios génétiques
  - Premières observations
  - Analyses des profils de puissance des méthodes
- Etude de données réelles :le syndrome de Brugada
  - Données
  - Analyse des profils de significativité des gènes
- Discussion

Epidémiologie génétique

- Discipline dont l'un des objectifs est de mettre en évidence les facteurs génétiques impliqués dans le trait étudié.



- Projet de recherche VaCaRMe

VACARME
VAincre les maladies CArdiovasculaires, Respiratoires et MEtaboliques

→ Applications : diagnostic précoce et médecine personnalisée

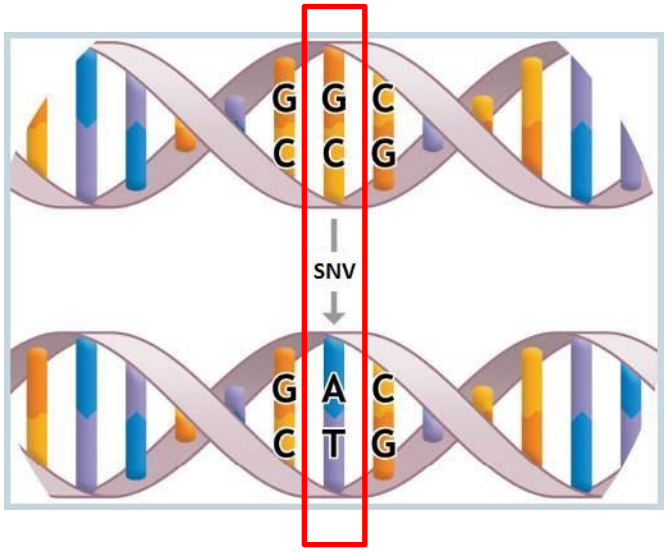
→ Financé par la région Pays de la Loire

VARIATIONS GÉNÉTIQUES ÉTUDIÉES

- Cartographie génétique des gènes → marqueurs génétiques

- Single Nucleotide Variant (**SNV**)

→ Modification d'un nucléotide dans la séquence d'ADN



2 allèles : G et A

G : allèle **majeur** (le plus fréquent)
A : allèle **mineur** (le moins fréquent)

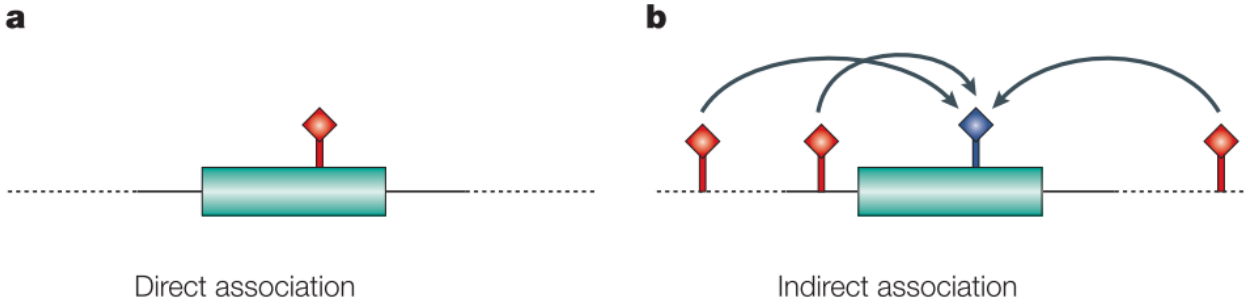
Fréquence de l'allèle mineur : MAF

SNV fréquent :
MAF>1%

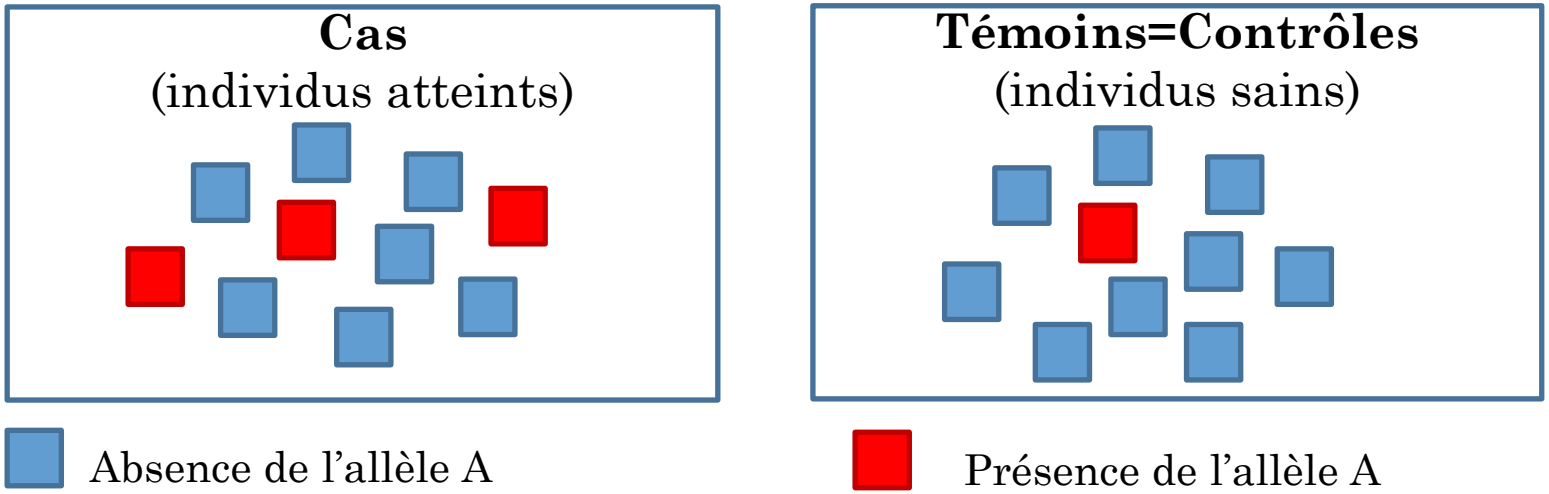
SNV rare :
MAF≤1%

Les études d'association génétique

- **Objectif :** Tester l'association entre des variants génétiques et une maladie donnée. → trouver des gènes



- Etudes cas-témoins



- Test d'association pour un variant :
- → Comparaison des fréquences alléliques chez les cas et les témoins

DES VARIANTS GÉNÉTIQUES FRÉQUENTS AUX VARIANTS GÉNÉTIQUES RARES

- De nombreuses études d'association génome-entier (GWASs) menées pour des variants fréquents

  - Une faible part de la composante génétique attendue est expliquée par ces variants fréquents

- Les variants rares : rôle important

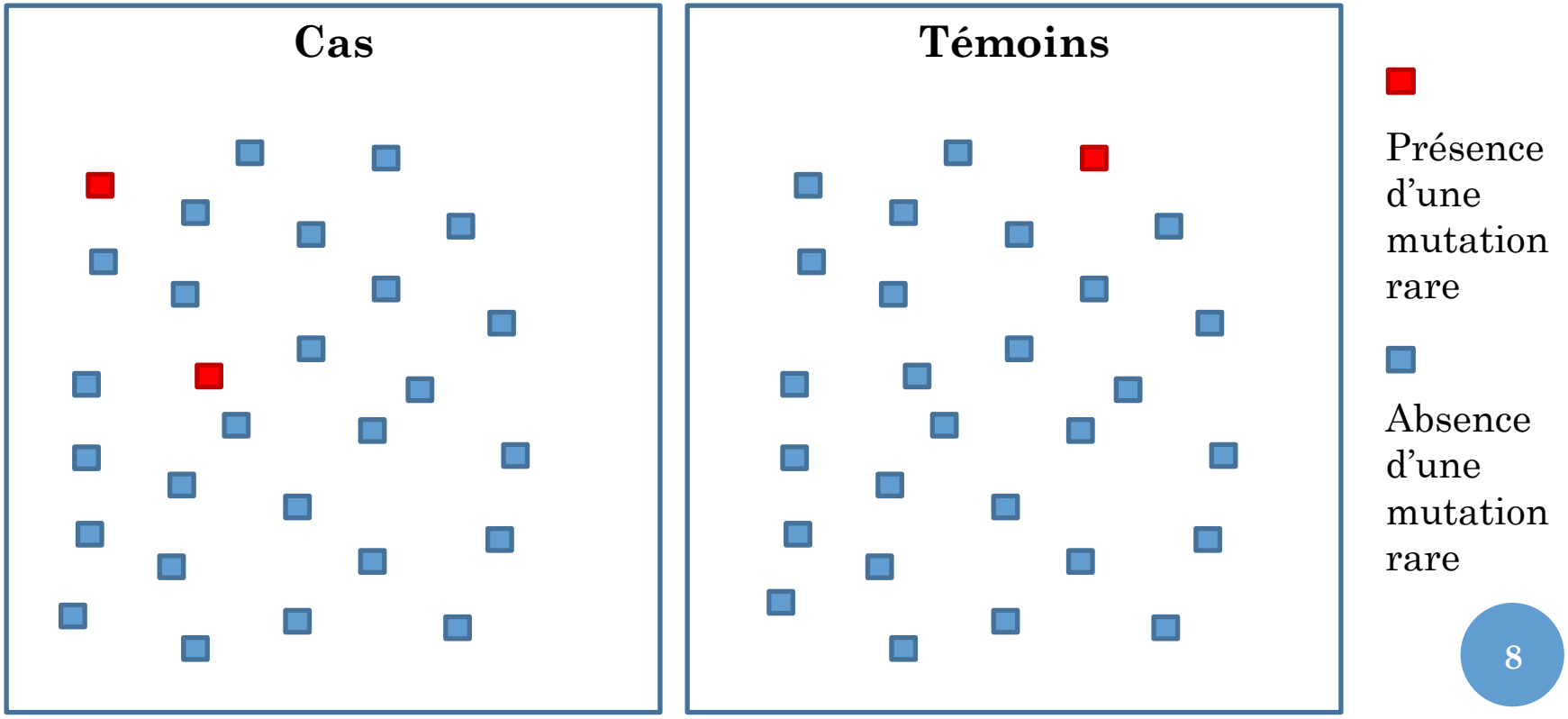
CDCV
Common Disease – Common Variant

→

CDRV
Common Disease – Rare Variant

De nouveaux tests statistiques pour les variants génétiques rares

- Tester individuellement les variants
- → Pas adapté pour des variants très peu fréquents
- Nouvelle stratégie : utiliser l'information génétique de plusieurs variants rares. (situés sur le même gène)

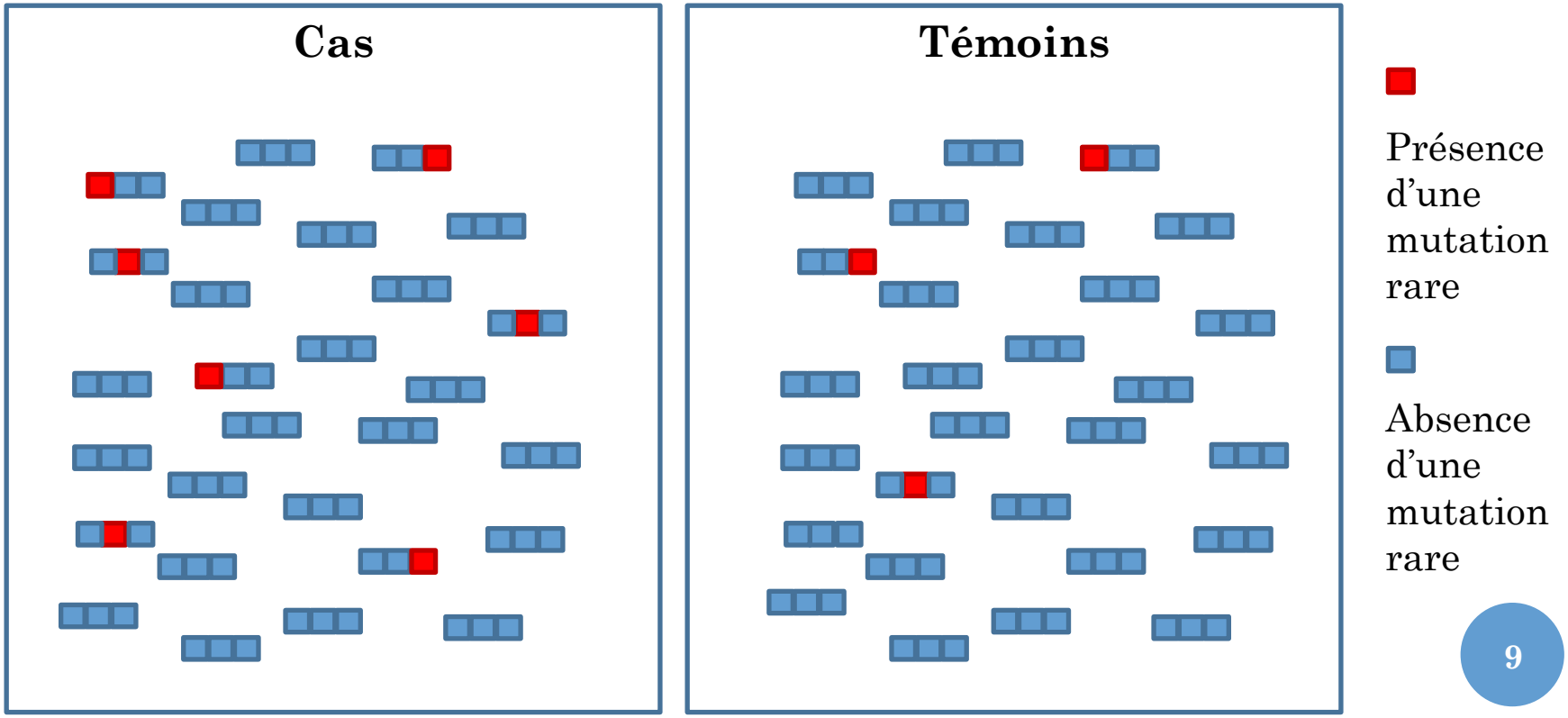


2 sur 28
ont une mutation rare

1 sur 28
ont une mutation rare

Exemple : CAST

- CAST= « cohort allellic sum test » (Morgenthaler and Thilly, 2007)
- Comparaison des proportions de personnes porteuses d'au moins une mutation rare pour un gène donné entre cas et témoins



7 sur 28
ont au moins une mutation rare

3 sur 28
ont au moins une mutation rare

De nombreux tests d'association pour les variants génétiques rares…

2007	•CAST (Morgenthaler and Thilly)
2008	•CMC (Li and Leal)
2009	•WS (Madsen and Brown)
2010	•aSum (Han and Pan) VT (Price et al.) CMAT (Zawistowski et al.) SCORE (Hoffmann et al.) KBAC (Liu and Leal)
2011	•RWAS (Sul et al.) PWST (Zhang et al.) RBS (Ionita-Laza et al.) EREC (Lin and Tang) C-alpha (Neale et al.) SKAT (Wu et al.)
2012	•SKAT-O (Lee et al.) sigma P (Cheung et al.)
2013	•BOMP (Chen et al.) Methode de Fisher (Derkach et al.)
2014	•ADA (Lin et al.) CLUSTER (Lin)

Objectifs du stage

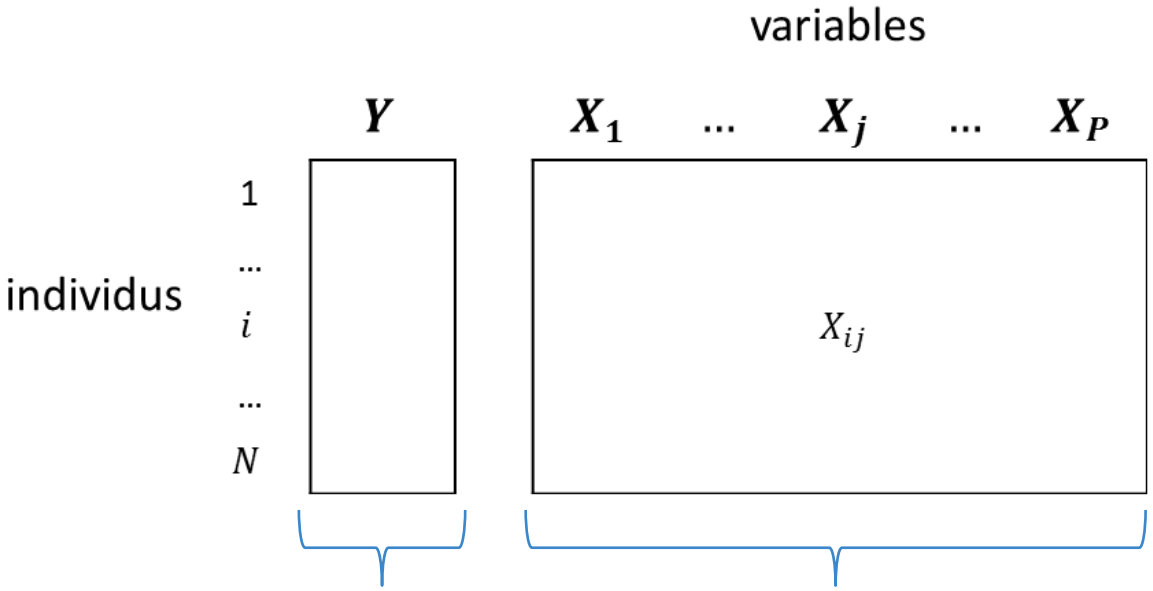
- Identifier les principales stratégies pour tester l'association entre variants génétiques rares et maladie

- Comparer un ensemble représentatif des tests statistiques au moyen de :
  - Données simulées selon différents scénarios génétiques
  - Données réelles issues de l'étude du syndrome de Brugada

PLAN

- Contexte
  - Epidémiologie génétique et projet VaCaRMe
  - Les études d'association génétique pour des maladies
  - De l'étude des variants fréquents à l'étude des variants rares
- Tests d'association génétique dans le cadre des variants rares
  - Structure des données
  - Tests « poolés »
  - Propriétés des tests d'association étudiés
- Simulations de données
  - Scénarios génétiques
  - Premières observations
  - Analyses des profils de puissance des méthodes
- Etude de données réelles :le syndrome de Brugada
  - Données
  - Analyse des profils de significativité des gènes
- Discussion

Structure des données et notations



Phénotype

Y_i=1 si individu i atteint

Y_i=0 sinon

Génotypes

X_{ij} = 0 allèles mineurs

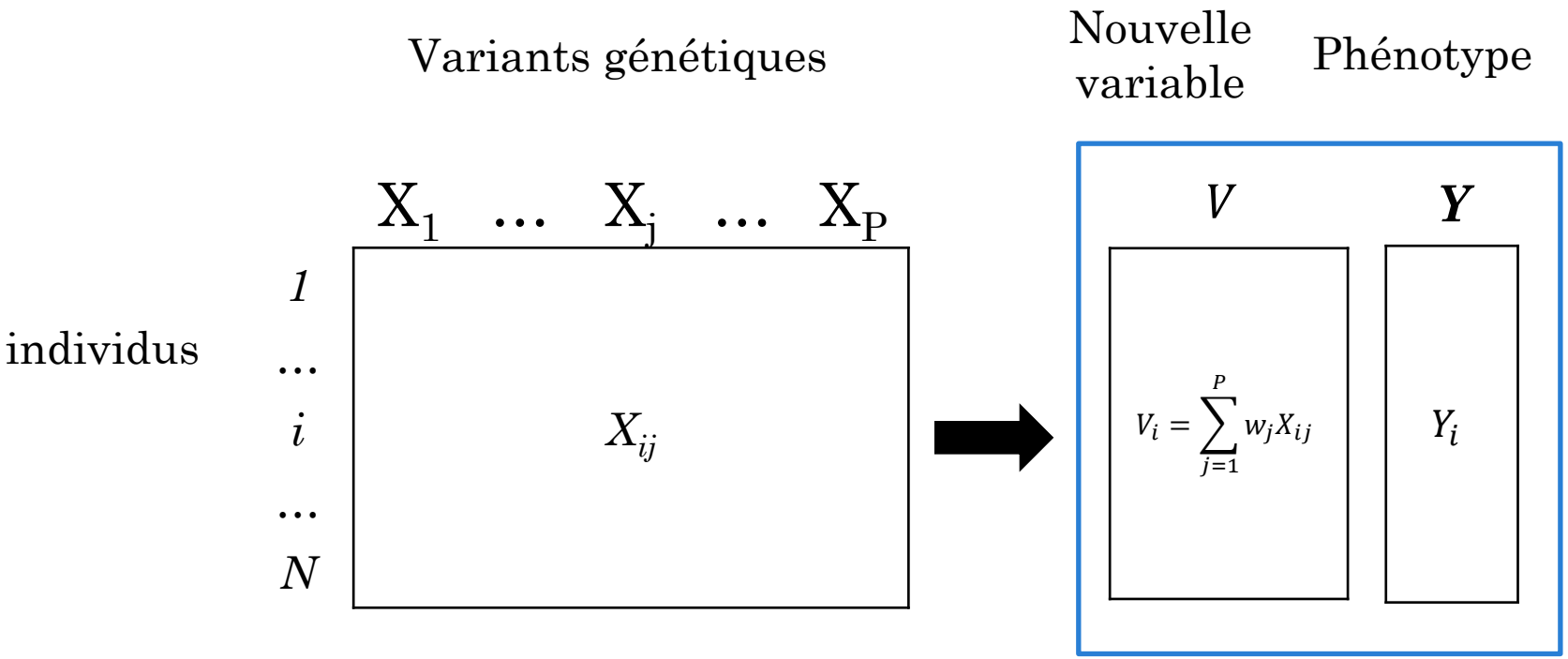
1

2

génotype de l'individu i pour le SNV j

TESTS « POOLÉS »

- Résumer l'information génétique d'un individu par un score génétique et tester l'association entre ce score et la maladie



Test de l'association

TEST D'ASSOCIATION ENTRE LE SCORE GÉNÉTIQUE ET LA MALADIE

- Modèle de régression logistique univarié

$$logit\big(P(Y_i = 1)\big) = \beta_0 + \boldsymbol{\beta V_i}$$

- Distribution de la statistique de test sous H_0
  - Loi connue
  - Test complexe → procédure de permutations des labels des individus

LES TESTS COMPARÉS

Catégorie	Description de la stratégie utilisée	Méthodes
Tests « poolés »	Calcul d'un score génétique.	**WS, aSum, VT,** RWAS, PWST, RBS, EREC, CMAT, SCORE
	Calcul d'un score génétique et création d'une variable binaire indiquant si le score est supérieur à un seuil.	**CAST,** CMC, BOMP, RARECOVER
Tests « non-poolés »	Tests de décomposition de la variance	**C-alpha, SKAT, SKAT-O**
	Combinaison de p-values de tests simple-marqueur.	Méthode de Fisher, Sigma-P, ADA, CLUSTER
	Considération d'un génotype multi-locus	**KBAC**

PLAN

- Contexte
  - Epidémiologie génétique et projet VaCaRMe
  - Les études d'association génétique pour des maladies
  - De l'étude des variants fréquents à l'étude des variants rares
- Tests d'association génétique dans le cadre des variants rares
  - Structure des données
  - Tests « poolés »
  - Propriétés des tests d'association étudiés
- Simulations de données
  - Scénarios génétiques
  - Premières observations
  - Analyses des profils de puissance des méthodes
- Etude de données réelles :le syndrome de Brugada
  - Données
  - Analyse des profils de significativité des gènes
- Discussion

Scénarios génétiques

- Les simulations permettent de calculer les puissances et les erreurs de type I des méthodes.

Simulation d'un groupe de variants

- Composition et structure
  - Seulement des variants rares ou avec aussi des variants plus fréquents
  - Des variants indépendants ou corrélés

- Modèle génétique
  - Des proportions différentes de variants causaux
  - Variants causaux :
    - présence de variants à risque
    - et mélange de variants à risque et protecteurs

Scénarios génétiques

VR : variant rare
VPF : variant peu fréquent
VF : variant fréquent

- Composition et structure

Scénario	Nombre de VRs	Nombre de VPFs	Nombre de VFs	Corrélation entre les variants
1	8 causaux + $P_{VR_{nc}}$ non-causaux	0	0	Indépendants
2				Corrélés
3				VRs causaux corrélés, VRs non causaux corrélés, VRs causaux et les VRs non causaux non corrélés
4		4 non-causaux	0	Indépendants
5		4 non-causaux	2 non-causaux	
6		3 non-causaux et 1 causal	0	

$P_{VR_{nc}}$ prenant les valeurs 0,4,8,16 ou 32.

30 scénarios

Scénarios génétiques

- Modèle de risque

$$logit\big(P(Y_i = 1)\big) = \beta_0 + \boldsymbol{X_i\beta}$$

  - $\beta_0 = \log\left(\frac{0.05}{1-0.05}\right)$ pour une prévalence environ égale à 5%
  - $\boldsymbol{\beta}$ **:** coefficients de régression des variants génétiques

- Gène non associé à la maladie

$$OR_j = 1 \text{ pour tout variant } j \ (\beta_j = 0)$$

Erreur de type I

- Gène associé à la maladie
  - **8** variants rares causaux à risque

30 scénarios

$$OR_{VR_c} = 2\mathbf{1_8}$$

  - **5** variants rares causaux à risque et **3** protecteurs

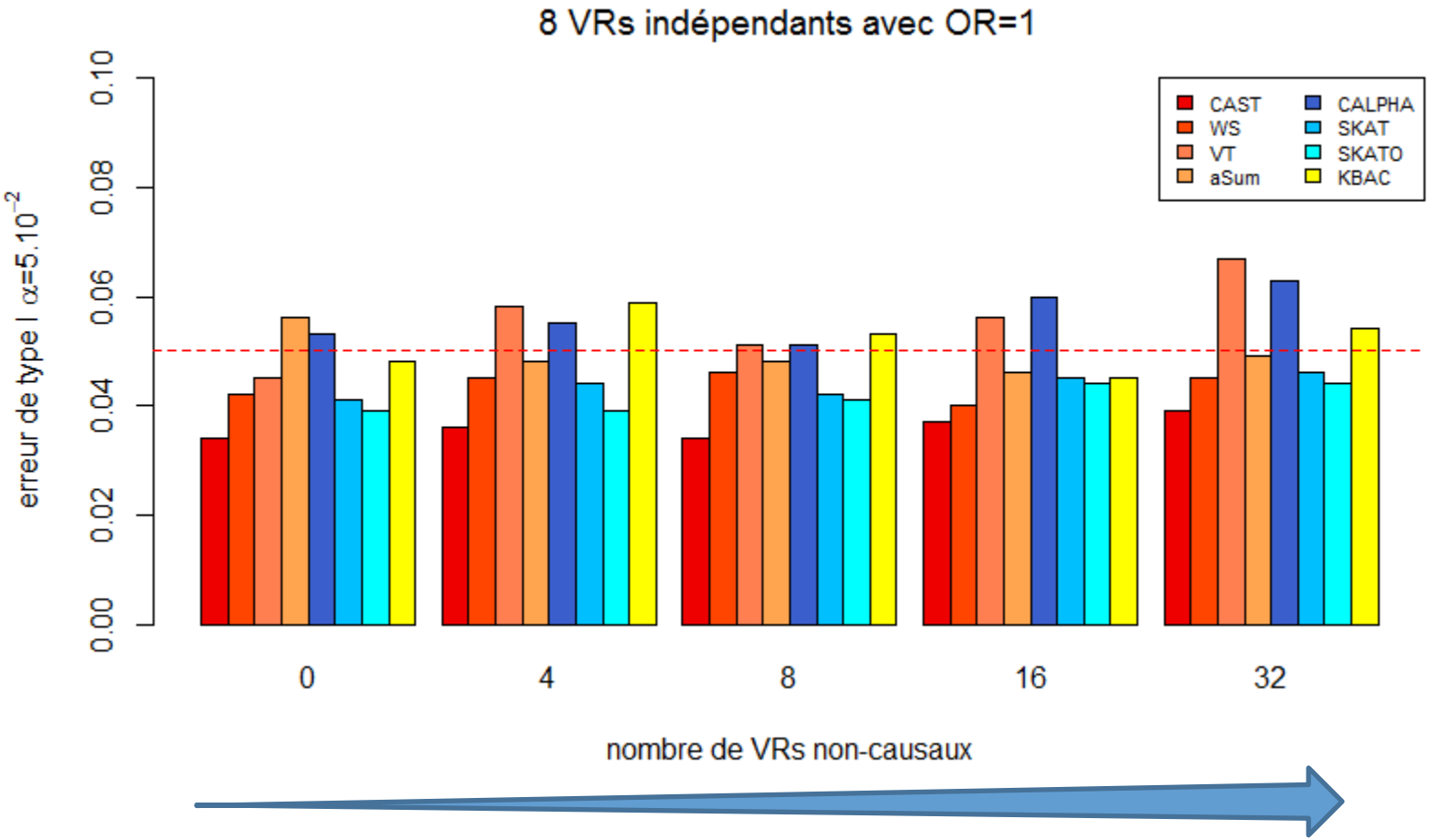
30 scénarios

$$OR_{VR_c} = \left(3,3,2,2,2,\frac{1}{2},\frac{1}{2},\frac{1}{2}\right)$$

Puissance

Contrôle des erreurs de type I

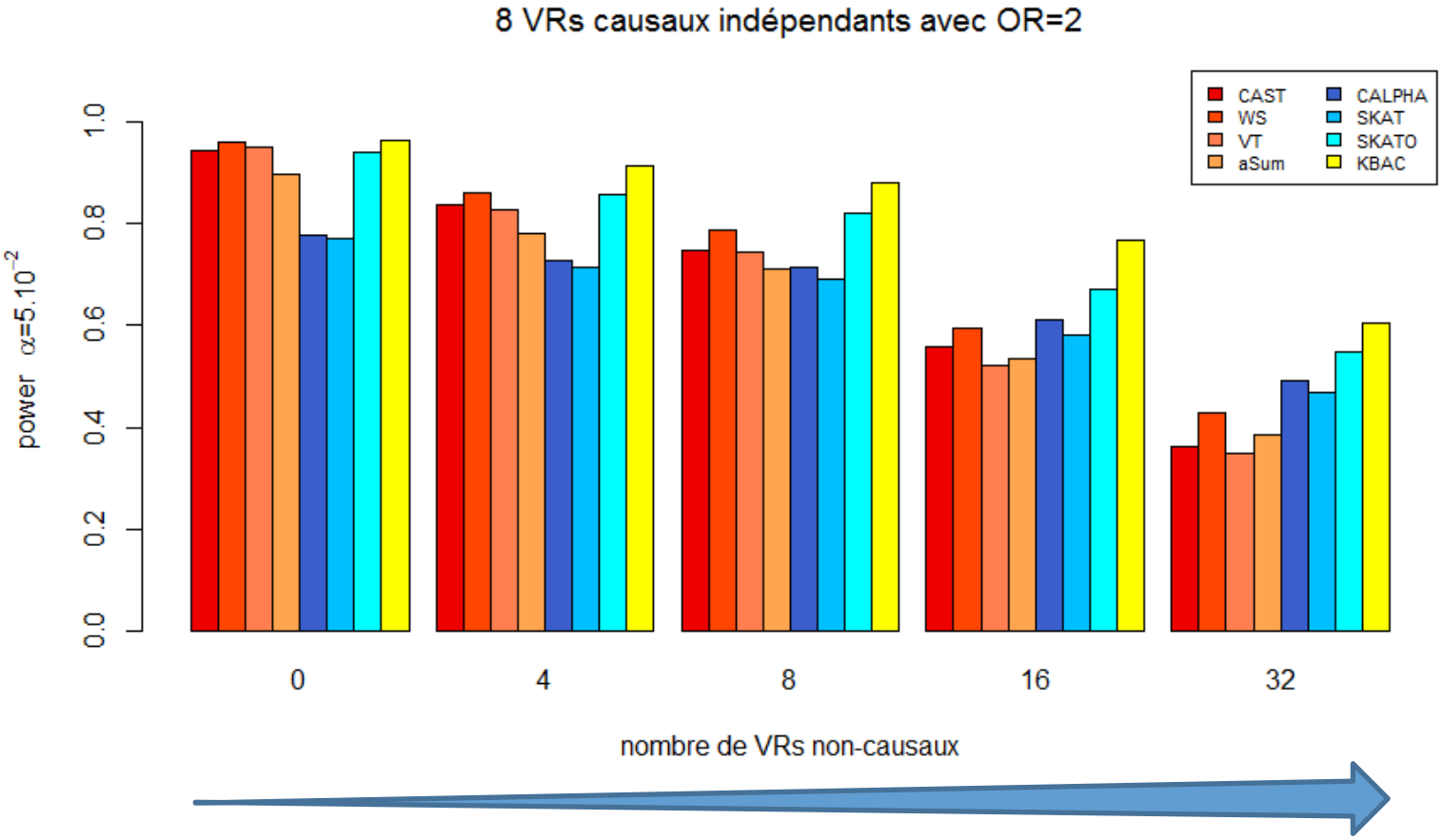
- Les erreurs de type I sont correctes pour l'ensemble des sous-scénarios.



Nombre de simulations : 1000

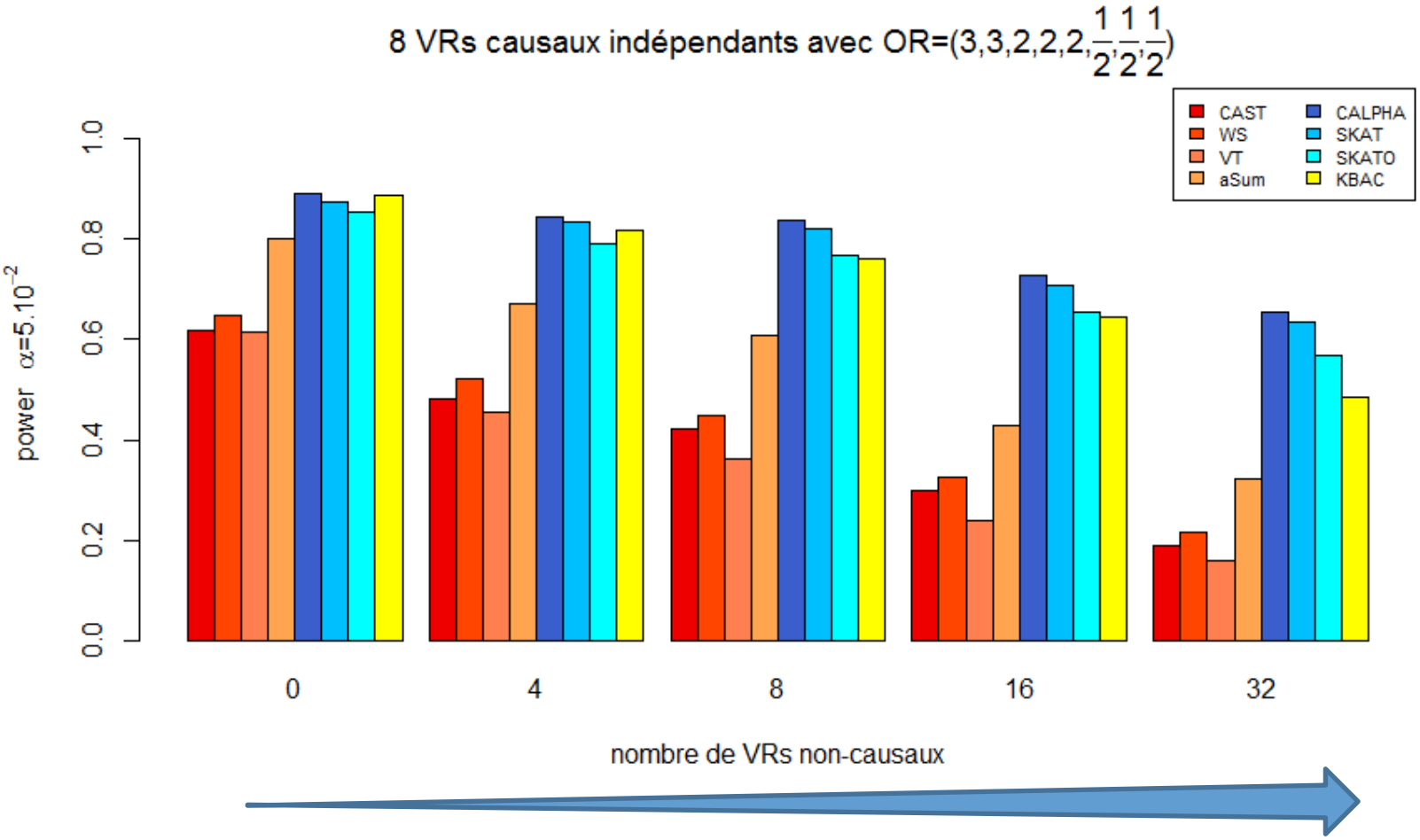
Premières observations pour la puissance

- Les puissances diminuent fortement avec l'introduction de variants rares non-causaux.



En présence de variants rares protecteurs

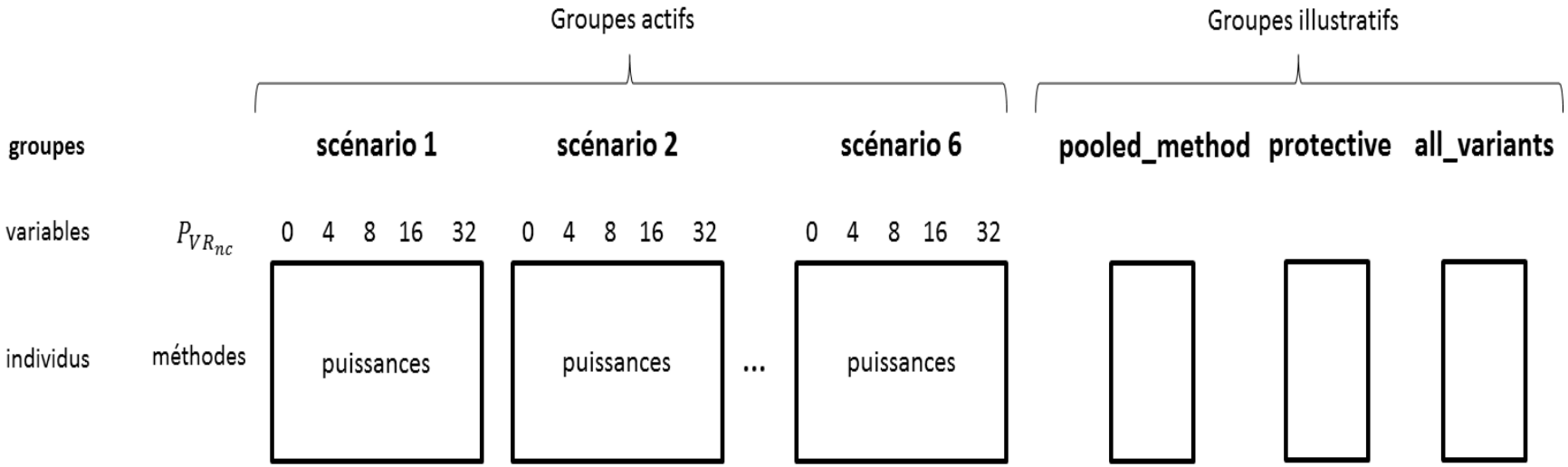
- Comportements des tests complètement différents avec la présence de VRs protecteurs.



Analyse des profils de puissance des méthodes

- Au total 60 sous-scénarios
- → Utilisation de méthodes d'analyse multidimensionnelle pour une vision synthétique

- Choix d'effectuer une analyse factorielle multiple (AFM)



→ Quelles méthodes ont des profils de puissance similaires?

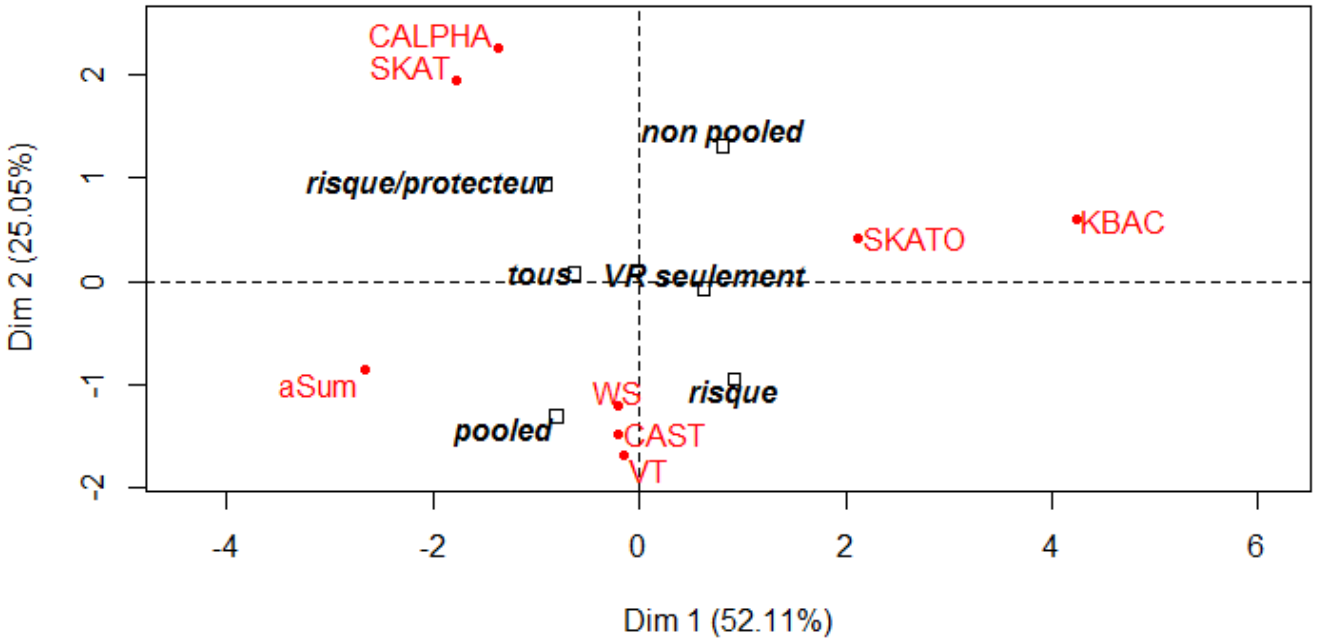
- Analyses séparées pour :
  - Variants rares tous à risque
  - Variants rares à risque et protecteurs

RÉSULTATS : VARIANTS À RISQUE

Peu sensibles au bruit

Bruit=ajout de variants rares non causaux

Sensibles au bruit



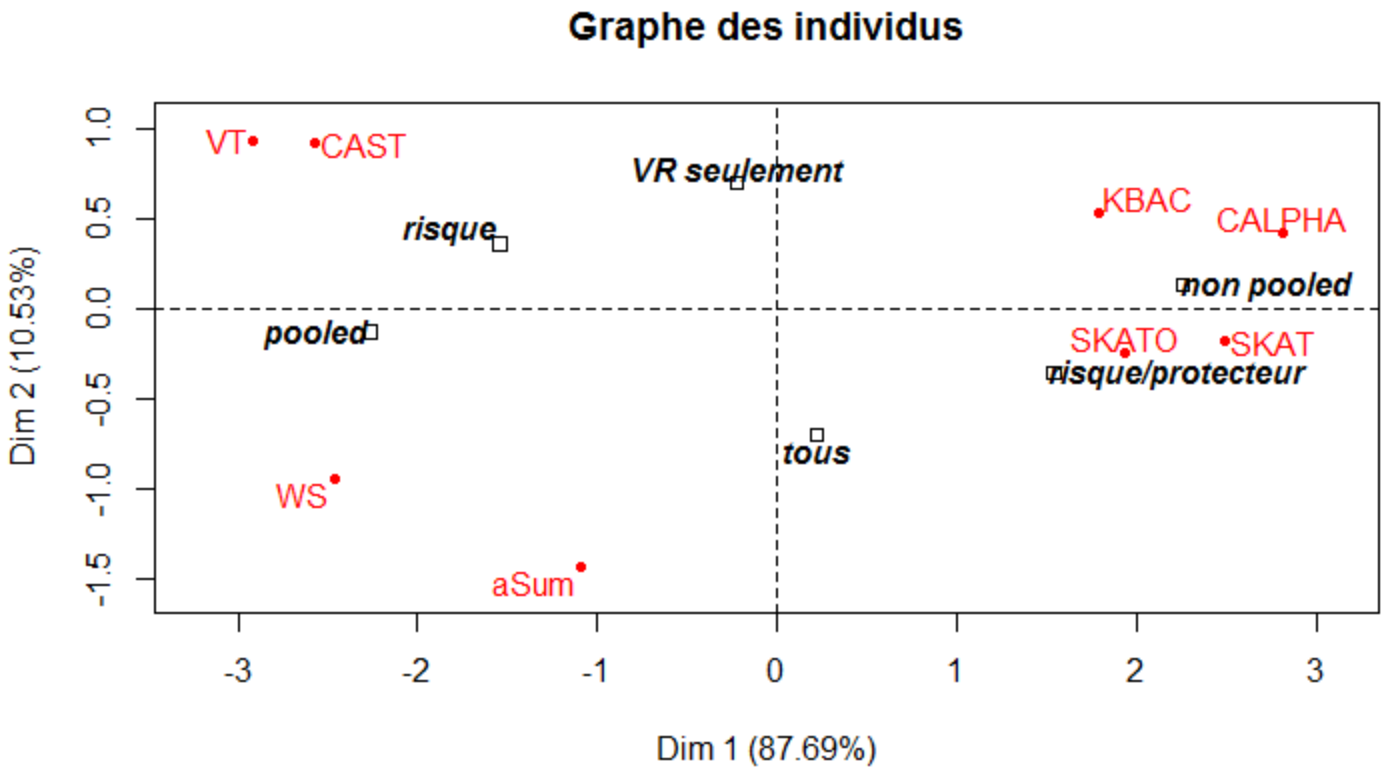
Peu puissantes → Puissantes

Pour une majorité de sous-scénarios

Résultats : variants à risque et variants protecteurs

Puissantes pour les scénarios 4 et 5 (VPFs non-causaux)

Puissantes pour le scénario 6 (VPFs non-causaux et causaux)



Peu puissantes → Puissantes

Pour la totalité des sous-scénarios

Conclusions

De nombreux tests statistiques existent

- Il n'existe pas de test nettement plus puissant que les autres pour l'ensemble des scénarios considérés.

- Cependant 2 tests se démarquent : SKAT-O et KBAC

PLAN

- Contexte
  - Epidémiologie génétique et projet VaCaRMe
  - Les études d'association génétique pour des maladies
  - De l'étude des variants fréquents à l'étude des variants rares
- Tests d'association génétique dans le cadre des variants rares
  - Structure des données
  - Tests « poolés »
  - Propriétés des tests d'association étudiés
- Simulations de données
  - Scénarios génétiques
  - Premières observations
  - Analyses des profils de puissance des méthodes
- Etude de données réelles :le syndrome de Brugada
  - Données
  - Analyse des profils de significativité des gènes
- Discussion

Données pour le syndrome de Brugada

Populations et régions génétiques d'intérêt

Recrutement de cas	Recrutement de témoins	Choix des régions génétiques d'intérêt
167 patients atteints du syndrome de Brugada	167 patients atteints de rétrécissement aortique calcifié	Exons de 163 gènes candidats pour diverses maladies cardiaques Dont 21 spécifiques au syndrome de Brugada

Séquençage ciblé et détection des variants génétiques

Filtres

Variants rares	Variant ayant un potentiel impact fonctionnel	Vérifiés afin d'éviter des faux-positifs

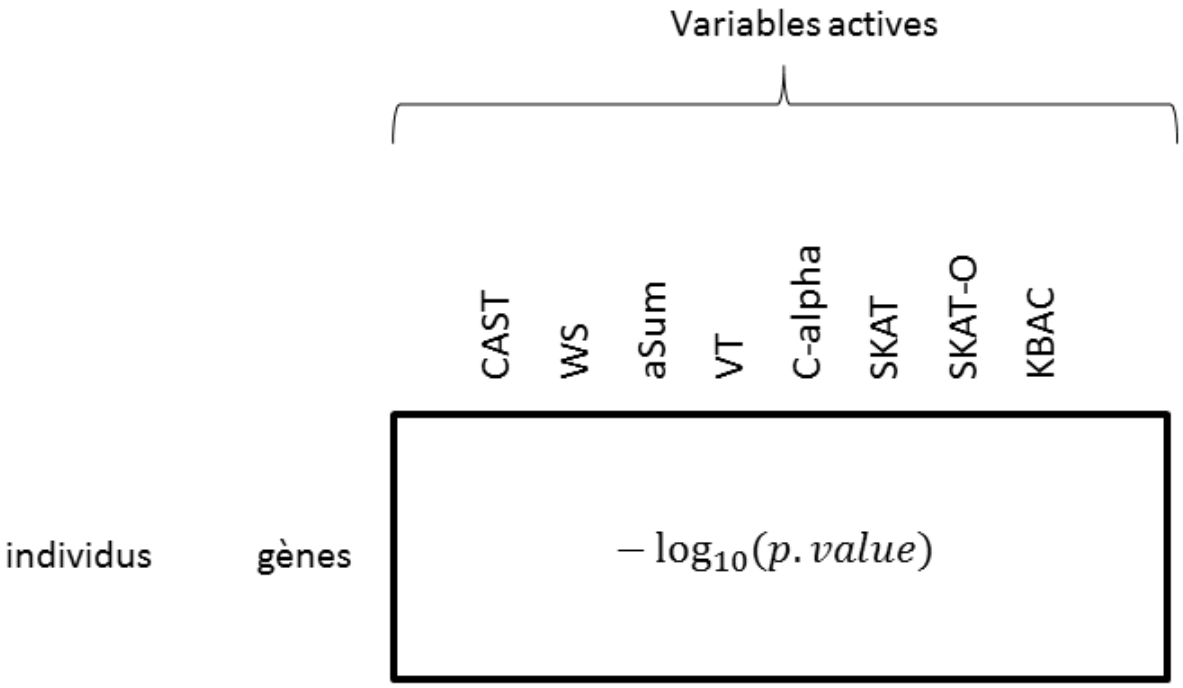
RÉSULTATS

Le gène *SCN5A* connu depuis 1998 pour être le gène majeur impliqué dans le syndrome de Brugada

	Gène candidat le plus significatif		Nombre de gènes avec une p-value inférieure à		
	nom	p-value	0.01%	1%	5%
CAST	*SCN5A*	3,38E-05	1	1	4
WS	*SCN5A*	6,46E-05	1	2	14
aSum	*SCN5A*	2,18E-05	1	3	16
VT	*SCN5A*	0,000999001	1	2	9
C-alpha	*TRDN*	0,007152939	0	1	4
SKAT	*TRDN*	0,07416087	0	0	0
SKAT-O	*SCN5A*	5,72E-05	1	1	8
KBAC	*SCN5A*	0,000999001	1	2	4

- L'ensemble des tests (sauf SKAT et C-alpha) ont permis de confirmer l'implication de variants génétiques rares présents sur le gène *SCN5A* dans le syndrome de Brugada.

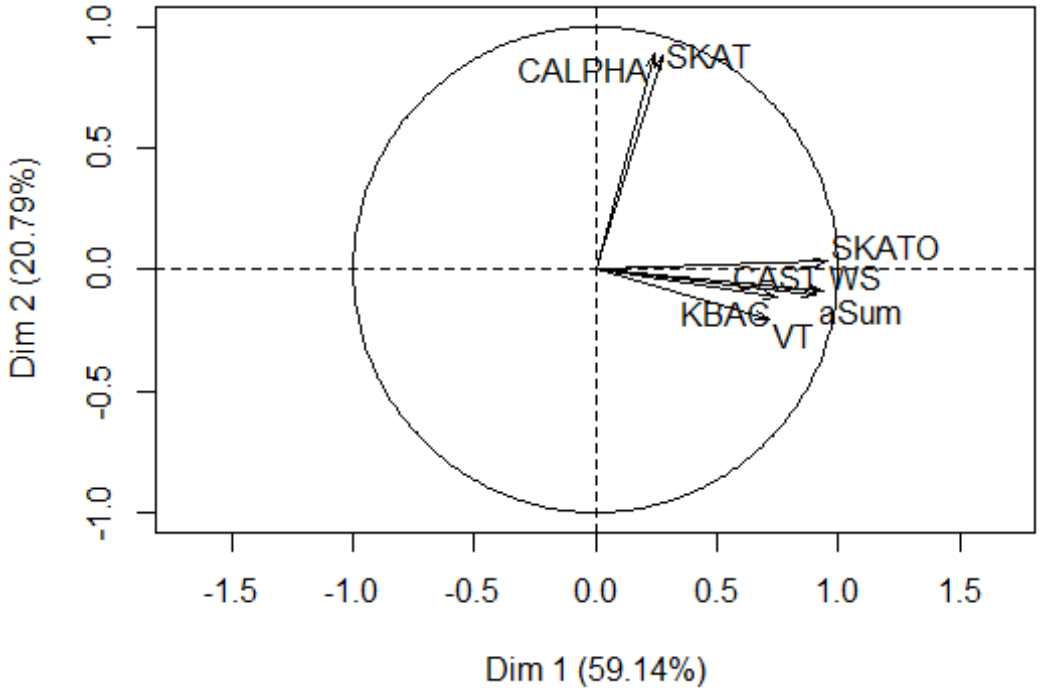
ACP normée sur les profils de significativité des gènes



Représentation des profils de significativité des gènes pour l'ensemble des tests étudiés.

→Peut-on identifier des groupes de tests fournissant des résultats très corrélés?

Résultats pour les données réelles



- Des groupes de méthodes fournissant des résultats similaires.
  - SKAT et C-alpha
  - CAST, WS, aSum, VT, KBAC et SKAT-O

PLAN

- Contexte
  - Epidémiologie génétique et projet VaCaRMe
  - Les études d'association génétique pour des maladies
  - De l'étude des variants fréquents à l'étude des variants rares
- Tests d'association génétique dans le cadre des variants rares
  - Structure des données
  - Tests « poolés »
  - Propriétés des tests d'association étudiés
- Simulations de données
  - Scénarios génétiques
  - Premières observations
  - Analyses des profils de puissance des méthodes
- Etude de données réelles :le syndrome de Brugada
  - Données
  - Analyse des profils de significativité des gènes
- Discussion

Conclusions

- De nombreux tests d'association pour les variants génétiques rares, lesquels utiliser?

SKAT-O et KBAC

- Simulations : Puissants pour la quasi-totalité des scénarios envisagés
- Données réelles : Fournissent des résultats globalement très corrélés.

Perspectives

Continuité en thèse

Analyse d'association de variants génétiques rares dans une population démographiquement stable

- Tests d'association pour les variants génétiques rares.
  →Test en développement pour détecter des regroupements de variants à risque dans de petites régions génomiques.

- Utilisation de la notion d' **identité par descendance** (IBD) pour détecter des associations de variants rares.

- Modélisation de la **structure génétique d'une population** à l'échelle régionale à l'aide d'analyses multivariées adaptées.

Merci pour votre attention!

l'institut du thorax

Equipe Génétique des maladies héréditaires
Jean-Jacques SCHOTT
Solena LE SCOUARNEC
Thème Statistique Génétique
Christian DINA
Matilde KARAKACHOFF
Floriane SIMONET

Equipe Variabilité génétique et mort subite
Richard REDON

Laboratoire de Mathémariques Jean Leray
UMR 6629 - Nantes

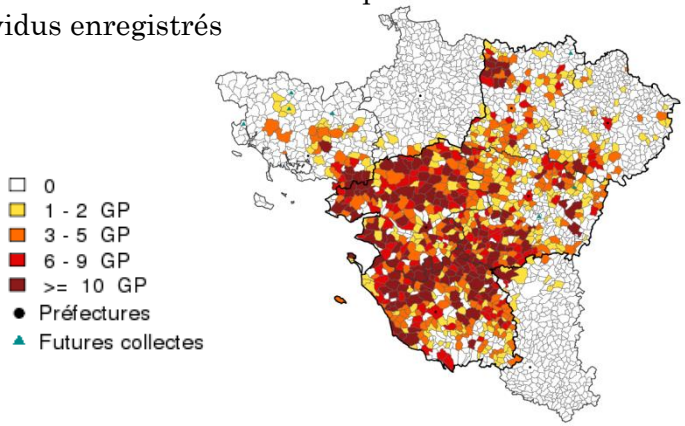
Lise BELLANGER

CCIPL : Centre de **C**alcul **I**ntensif des **P**ays de la **L**oire

Guy MOEBS

Vaincre les maladies Cardiovasculaires, Respiratoires et Métaboliques
VACARME

Lieux d'origine des grands-parents de 1235 individus nés à moins de 15km parmi les 1938 individus enregistrés



Direction scientifique : Richard REDON
Chef de projet : Stéphanie CHATEL

Région PAYS DE LA LOIRE

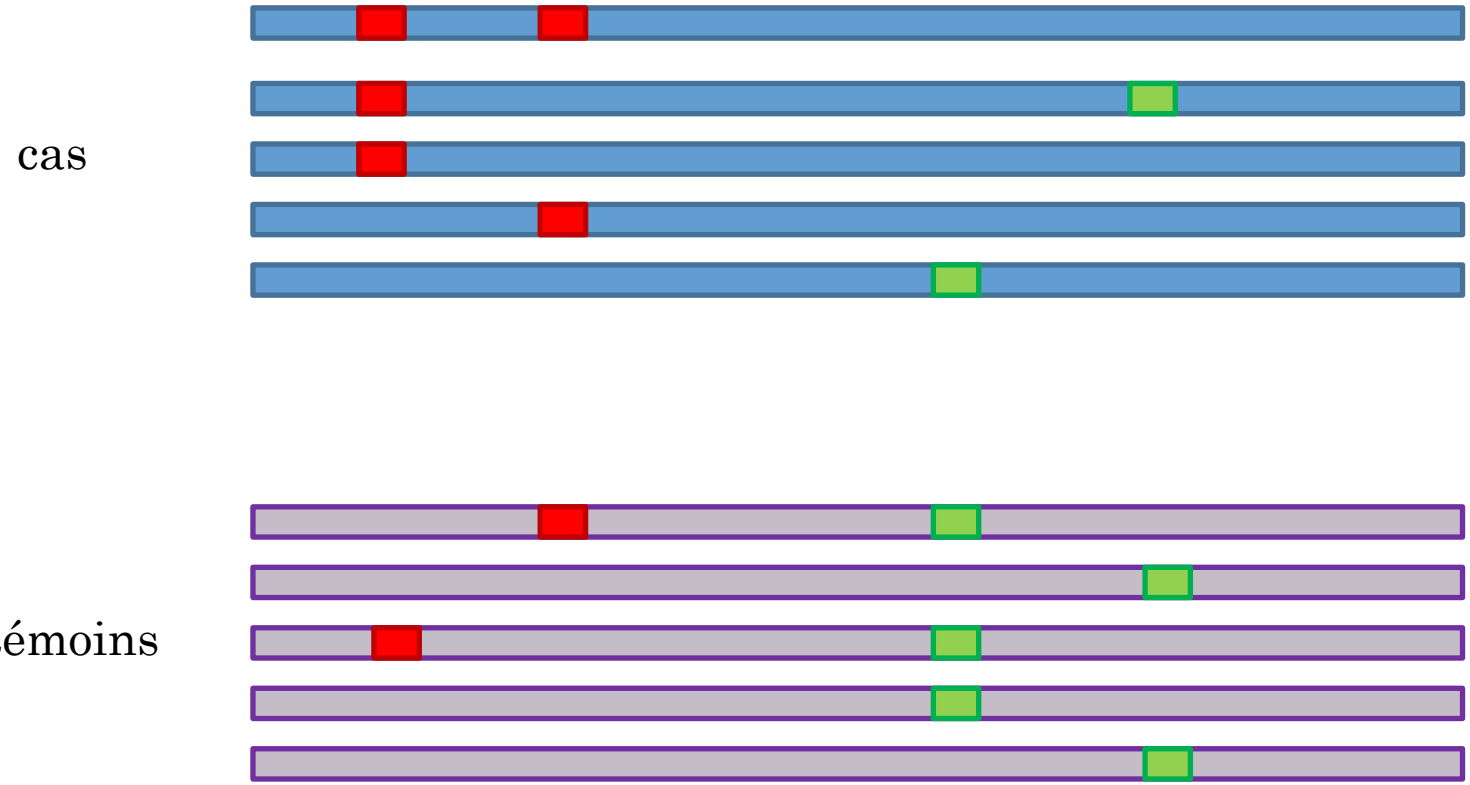
Références

- Chen, Y.-C., Carter, H., Parla, J., Kramer, M., Goes, F.S., Pirooznia, M., Zandi, P.P., McCombie, W.R., Potash, J.B., Karchin, R., 2013. A hybrid likelihood model for sequence-based disease association studies. PLoS Genet. 9, e1003224. doi:10.1371/journal.pgen.1003224
- Cheung, Y.H., Wang, G., Leal, S.M., Wang, S., 2012. A fast and noise-resilient approach to detect rare-variant associations with deep sequencing data for complex disorders. Genet. Epidemiol. 36, 675–685. doi:10.1002/gepi.21662
- Derkach, A., Lawless, J.F., Sun, L., 2013. Robust and Powerful Tests for Rare Variants Using Fisher's Method to Combine Evidence of Association From Two or More Complementary Tests. Genet. Epidemiol. 37, 110–121. doi:10.1002/gepi.21689
- Han, F., Pan, W., 2010. A data-adaptive sum test for disease association with multiple common or rare variants. Hum. Hered. 70, 42–54. doi:10.1159/000288704
- Hoffmann, T.J., Marini, N.J., Witte, J.S., 2010. Comprehensive approach to analyzing rare genetic variants. PloS One 5, e13584. doi:10.1371/journal.pone.0013584
- Ionita-Laza, I., Buxbaum, J.D., Laird, N.M., Lange, C., 2011. A new testing strategy to identify rare variants with either risk or protective effect on disease. PLoS Genet. 7, e1001289. doi:10.1371/journal.pgen.1001289
- Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., Nickerson, D.A., Christiani, D.C., Wurfel, M.M., Lin, X., 2012. Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies. Am. J. Hum. Genet. 91, 224–237. doi:10.1016/j.ajhg.2012.06.007
- Li, B., Leal, S.M., 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. 83, 311–321. doi:10.1016/j.ajhg.2008.06.024
- Lin, D.-Y., Tang, Z.-Z., 2011. A general framework for detecting disease associations with rare variants in sequencing studies. Am. J. Hum. Genet. 89, 354–367. doi:10.1016/j.ajhg.2011.07.015
- Lin, W.-Y., 2014. Association testing of clustered rare causal variants in case-control studies. PloS One 9, e94337. doi:10.1371/journal.pone.0094337

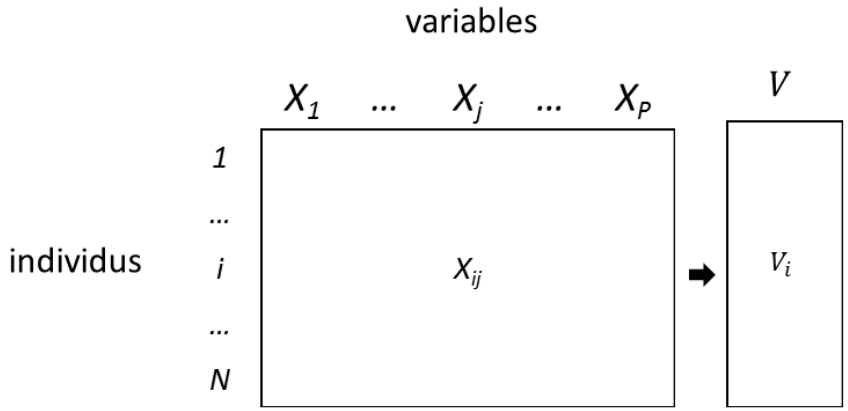
Références

- Liu, D.J., Leal, S.M., 2010. A Novel Adaptive Method for the Analysis of Next-Generation Sequencing Data to Detect Complex Trait Associations with Rare Variants Due to Gene Main Effects and Interactions. PLoS Genet 6, e1001156. doi:10.1371/journal.pgen.1001156
- Madsen, B.E., Browning, S.R., 2009. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 5, e1000384. doi:10.1371/journal.pgen.1000384
- Morgenthaler, S., Thilly, W.G., 2007. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat. Res. 615, 28–56. doi:10.1016/j.mrfmmm.2006.09.003
- Neale, B.M., Rivas, M.A., Voight, B.F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S.M., Roeder, K., Daly, M.J., 2011. Testing for an Unusual Distribution of Rare Variants. PLoS Genet 7, e1001322. doi:10.1371/journal.pgen.1001322
- Price, A.L., Kryukov, G.V., de Bakker, P.I.W., Purcell, S.M., Staples, J., Wei, L.-J., Sunyaev, S.R., 2010. Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet. 86, 832–838. doi:10.1016/j.ajhg.2010.04.005
- Sul, J.H., Han, B., He, D., Eskin, E., 2011. An Optimal Weighted Aggregated Association Test for Identification of Rare Variants Involved in Common Diseases. Genetics 188, 181–188. doi:10.1534/genetics.110.125070
- Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., Lin, X., 2011. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. Am. J. Hum. Genet. 89, 82–93. doi:10.1016/j.ajhg.2011.05.029
- Zawistowski, M., Gopalakrishnan, S., Ding, J., Li, Y., Grimm, S., Zöllner, S., 2010. Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. Am. J. Hum. Genet. 87, 604–617. doi:10.1016/j.ajhg.2010.10.012
- Zhang, Q., Irvin, M.R., Arnett, D.K., Province, M.A., Borecki, I., 2011. A data-driven method for identifying rare variants with heterogeneous trait effects. Genet. Epidemiol. 35, 679–685. doi:10.1002/gepi.20618

Variants à risque et protecteurs



TESTS POOLES



Nouvelle variable V	Valeur V_i	Poids du variant j
S	$\sum_{j=1}^{P} w_j X_{ij}$	w_j
S_{ws}	$\sum_{j=1}^{P} w_{WS_j} X_{ij}$	$w_{WS_j} = \frac{1}{\sqrt{N.\widehat{MAF}_j(1-\widehat{MAF}_j)}}$
S_{aSum}	$\sum_{j=1}^{P} w_{aSum_j} X_{ij}$	$w_{aSum_j} = \begin{cases} -1 & \text{si le variant } j \text{ est protecteur} \\ 1 & \text{sinon} \end{cases}$
$S_{VT}(t)$	$\sum_{j=1}^{P} I(\widehat{MAF}_j \leq t) X_{ij}$	$I(\widehat{MAF}_j \leq t)$, avec t seuil variable
C_{CAST}	$I(S_i \geq 1)$	$w_j = I(\widehat{MAF}_j \leq t)$, avec t seuil fixe
$C_{BOMP}(T)$	$I(S_i \geq T)$	$w_j = 1$

C-ALPHA

C-ALPHA

Hypothèse effectuée

Pour un SNP j

$$m_j^A \sim \text{Binom}(m_j, p_j)$$

Où

- m_j^A: nombre de mutations rares chez les cas pour le SNV *j*
- p_j: probabilité d'avoir un allèle mutant du SNP j chez un cas

Sous H_0: $m_j^A \sim \text{Binom}(m_j, p_0)$

$p_0 = \bar{Y}$ est la proportion de cas

H_0: $\forall$ j	$p_j = p_0$
H_1: $\exists$ j	$p_j \neq p_0$

Calcul de la statistique T

$$T = \sum_{j=1}^{p} \left[\left(m_j^A - m_j p_0 \right)^2 - m_j p_0 (1 - p_0) \right]$$

Compréhension

$$T = \sum_{j=1}^{P} V_{obs}\left(m_j^A\right) - V_{H_0}\left(m_j^A\right)$$

Calcul de la p-value avec procédure de permutations des phénotypes.

SKAT

Modèle linéaire généralisé à effets aléatoires

$$\text{logit}\left(P(Y_i = 1 | X_i = x_i)\right) = \alpha_0 + \boldsymbol{\beta x_i}$$

$\boldsymbol{\beta} = (\beta_1, \dots, \beta_j, \dots, \beta_P)'$ **effets génétiques aléatoires**

Hypothèses sur le modèle:

- β_j est une variable **aléatoire** suivant une distribution de moyenne 0 et de variance $w_j \tau$.
- $\forall (j,k) \in [\![1, P]\!]^2 \ \ \text{corr}(\beta_j, \beta_k) = 0$

w_j est un poids donné au variant j.

fonction de facteurs comme la MAF ou un score de fonctionnalité

$$w_j = \frac{1}{MAF(1-MAF)}$$

τ est une composante de la variance

Hypothèse

$$H_0: \beta = 0$$

$$H_0: \beta_1 = \cdots = \beta_j = \cdots = \beta_P = 0$$

Autre écriture : $H_0: \tau = 0$

Statistique de test

$$\mathbf{Q = (Y - \hat{\mu})' K (Y - \hat{\mu})}$$

Avec

- $\hat{\mu} = \widehat{\alpha_0} + \hat{\alpha} Z_m$
- K
  - $\mathbf{K = XWWX'}$
  - Est de dimension (N x N)
  - L'élément (i, i') de la matrice est $K(X_i, X_{i'})$
  - K(.,.) est appelée fonction noyau et mesure la similarité génétique entre 2 individus i et i'

Evaluation de la p-value

- Sous H0

Q suit un mélange de lois de chi2

Estimation de la distribution sous H0 à l'aide de

- La méthode Davies : estimation des paramètres de la loi
- Ou par méthode de permutations

SKAT-O

- Jusqu'à présent hypothèse:

$$\forall\,(j,k) \in [\![1,P]\!]^2 \;\; corr(\beta_j, \beta_k) = 0$$

Weighted linear kernel

$$\mathbf{K = XWWX'}$$

- Incorporation d'une structure de corrélation entre les effets génétiques :

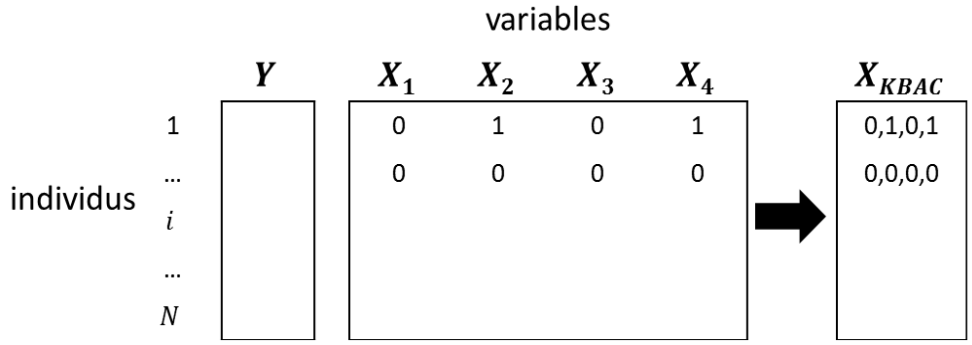
β suit une distribution multivariée avec comme structure de corrélation:

$$\mathbf{R_\rho = (1-\rho)I + \rho 11'}$$

$$\mathbf{K_\rho = XWR_\rho WX'}$$

- SKAT-O: recherche de ρ maximisant la puissance

KBAC



H_0 Les fréquences des génotypes multilocus sont identiques chez et chez les témoins.

La statistique dans le cadre du test bilatéral est :

$$KBAC = \left(\sum_{l=1}^{L} w_{KBAC\,l} \left(\frac{N_l^A}{N^A} - \frac{N_l^U}{N^U} \right) \right)^2$$

Le risque de tirer un génotype l chez les cas

$$\widehat{R_l} = \frac{N_l^A}{N_l}.$$

Le nombre de cas présentant le génotype l, N_l^A,sous H_0 suit la loi hypergéométrique $\mathcal{H}\left(N^A, \frac{N_l}{N}, N\right)$ et la fonction de densité de la variable aléatoire R_l est :

$$k_l^0(r_l) = P(R_l = r_l) = \frac{\binom{N_l}{N_l r_l}\binom{N - N_l}{N^A - N_l r_l}}{\binom{N}{N^A}}$$

Le poids w_{KBAC_l} est :

$$w_{KBAC_l} = \sum_{r_l \in \left\{\frac{0}{N_l}, \ldots, \widehat{R_l}\right\}} k_l^0(r_l) dr_l$$

RÉSULTATS : VARIANTS À RISQUE

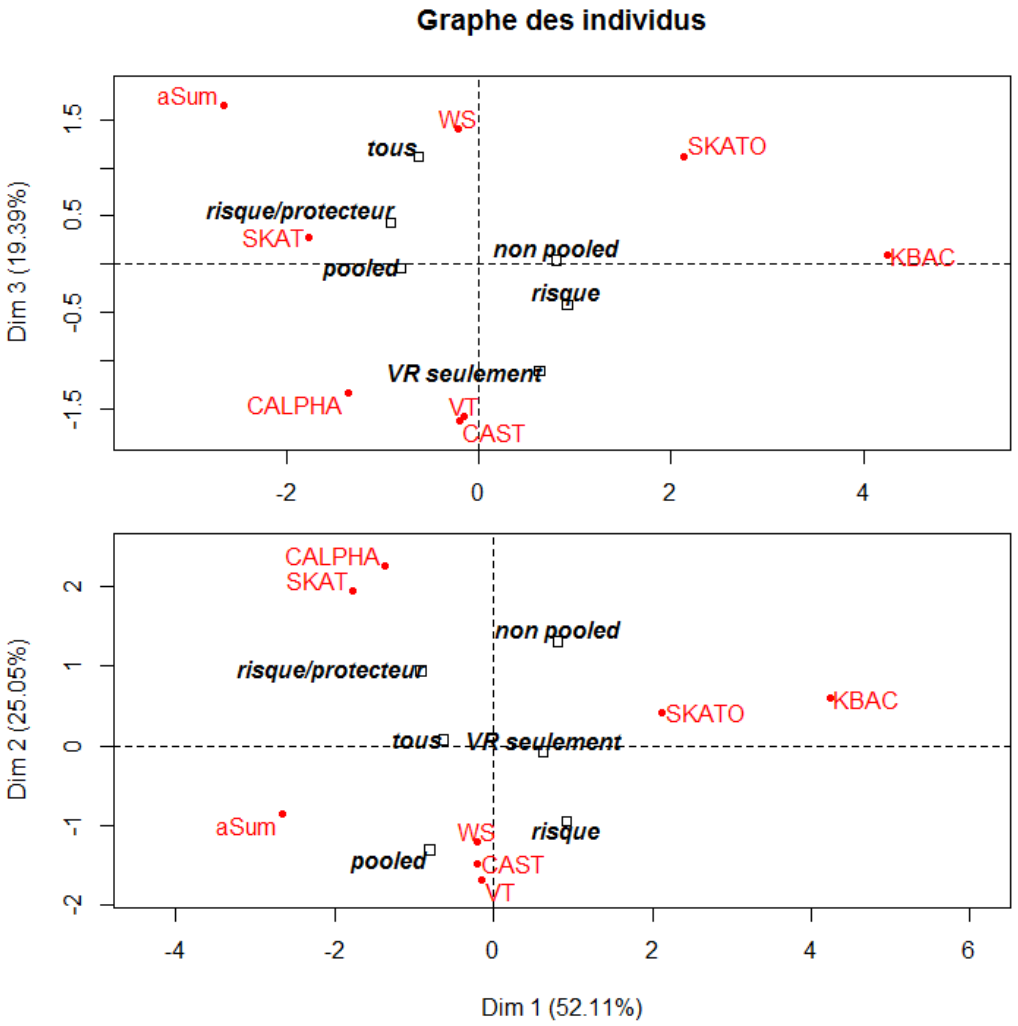
Puissantes pour le scénario 6 (VPFs non-causaux et causaux)

Puissantes pour les scénarios 4 et 5 (VPFs non-causaux)

Peu sensibles au bruit

Bruit=ajout de variants rares non causaux

Sensibles au bruit



Peu puissantes → Puissantes

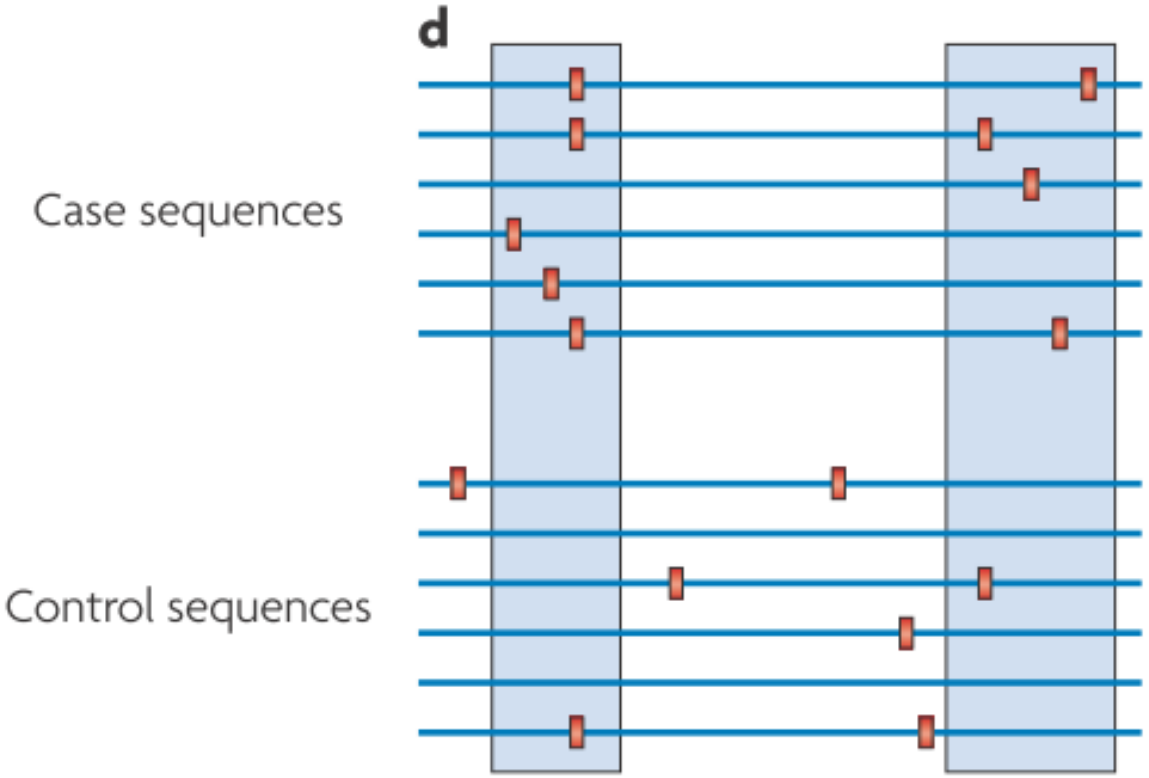
Pour une majorité de sous-scénarios

Test DoEstRer

Regroupements de variants rares à risque

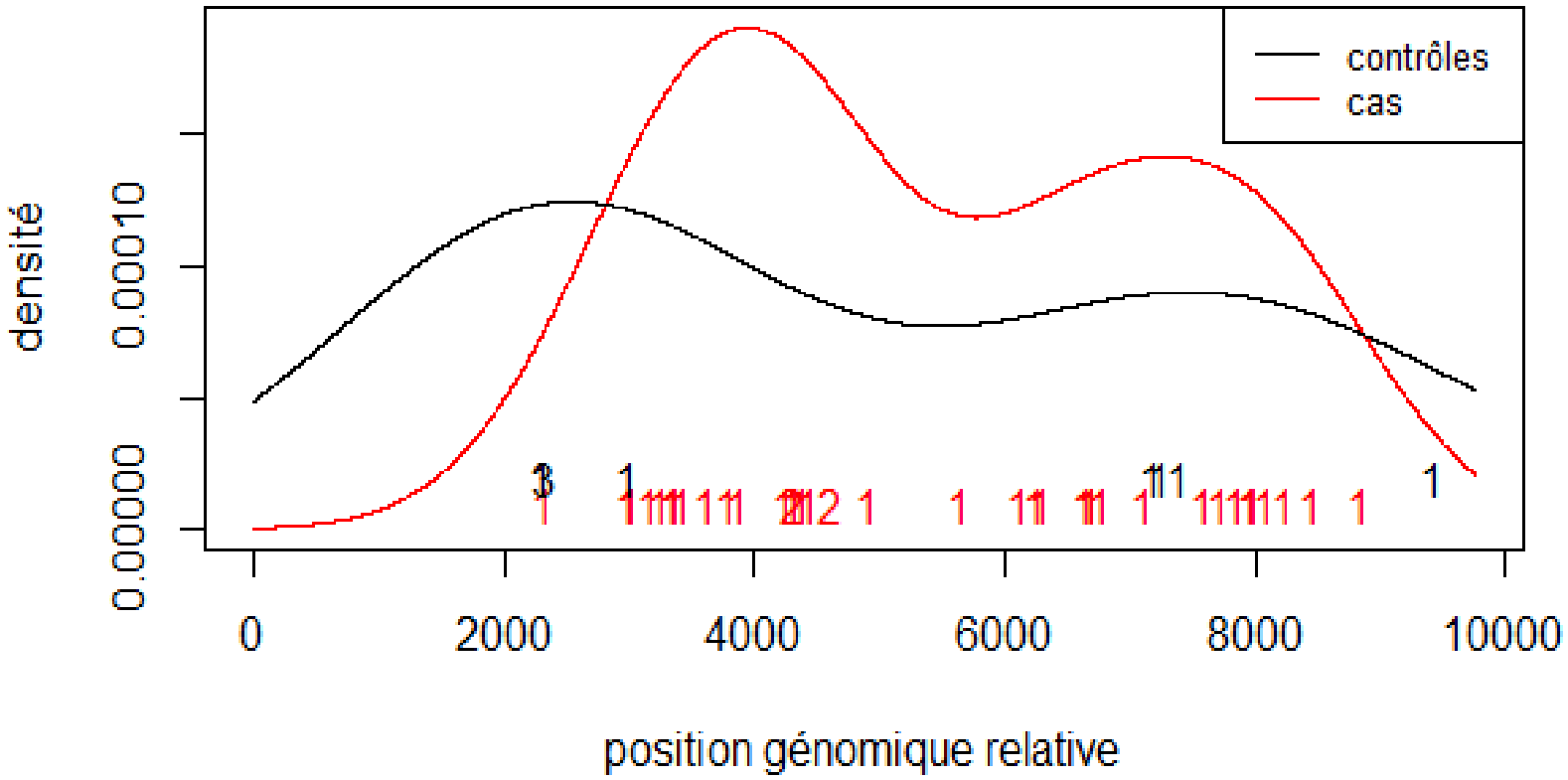
Image tirée de :

Bansal, V. et al., 2010. Statistical analysis strategies for association studies involving rare variants. Nat. Rev. Genet.



d | De multiples variations rares contribuent au phénotype et ces variations se situent dans des régions génomiques spécifiques.

EXEMPLE *SCN5A*



Construction du test

- **Hypothèses de test**

H_0 : Les **fonctions de densité des positions des mutations** sont identiques chez les cas (**A**ffected) et les témoins (**U**naffected).

ET

Les **fréquences moyennes de mutations** rares sont égales chez les cas et les témoins .

$$H_0 : f^A = f^U = f \;\; et \; p^A = p^U = p$$

$$H_1 : f^A \neq f^U \; ou \; p^A \neq p^U$$

Construction du test

- **Statistique de test**

$$STAT = \int_{1}^{\max d} |\widehat{p^A} \times \hat{f}_{bw}^{A}(pos) - \widehat{p^U}\,\hat{f}_{bw}^{U}(pos)|\, dpos$$

Avec

- $maxd$: longueur du gène d'intérêt
- $\hat{f}^A$ et $\hat{f}^U$: les fonctions de densités estimées chez les cas et les témoins avec l'estimation par noyau.
- $\widehat{p^A}$ et $\widehat{p^U}$: estimateurs des fréquences moyennes de mutations chez les cas et les témoins.