

Analyse de variants rares issus de données de séquençage à haut débit

Réseau Interdisciplinaire autour de la Statistique (RIS) : Groupe Statistiques et Génomique

Mathilde Boissel

UMR8199 CNRS/EGID



Valentin Harter

Centre de Traitement des Données du Cancéropôle Nord-Ouest

Centre François Baclesse, Caen



Plan

Introduction

- les séquences génomiques
- les variants
- origine des variants
- conséquences des variants
- sources de données
- données publiques
- les études portant sur les variants
- problématique des variants rares

QC pour « Quality controls »

- QC des individus
- QC des variants

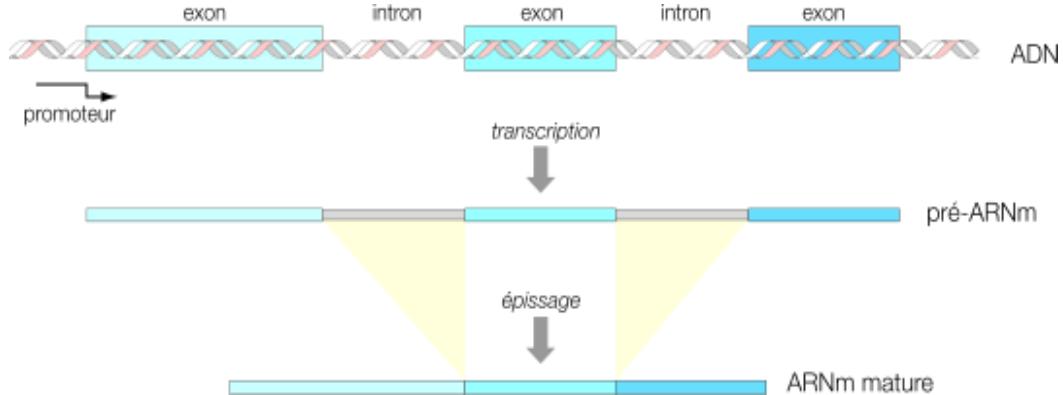
Analyses de données

- Variants fréquents
- Variants rares : Tests d'agrégation de variants
- Avantages & inconvénients

Retours d'expérience

Introduction, les séquences génomiques

Une **séquence génomique** est une représentation de l'enchaînement des désoxyribonucléotides le long d'un brin d'une macromolécule d'**ADN**. Elle est représentée par une chaîne de caractères utilisant l'alphabet A,C,G et T, initiales des bases azotées : Adénine, Cytosine, Guanine et Thymine.



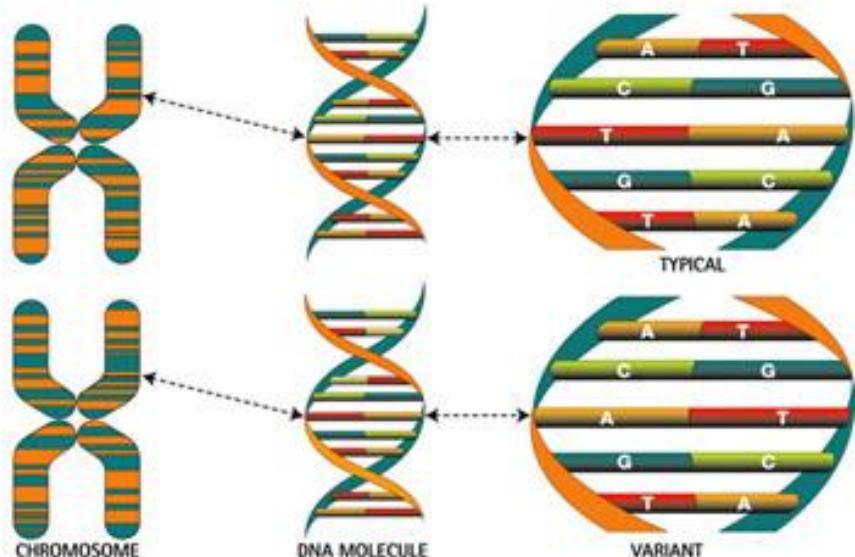
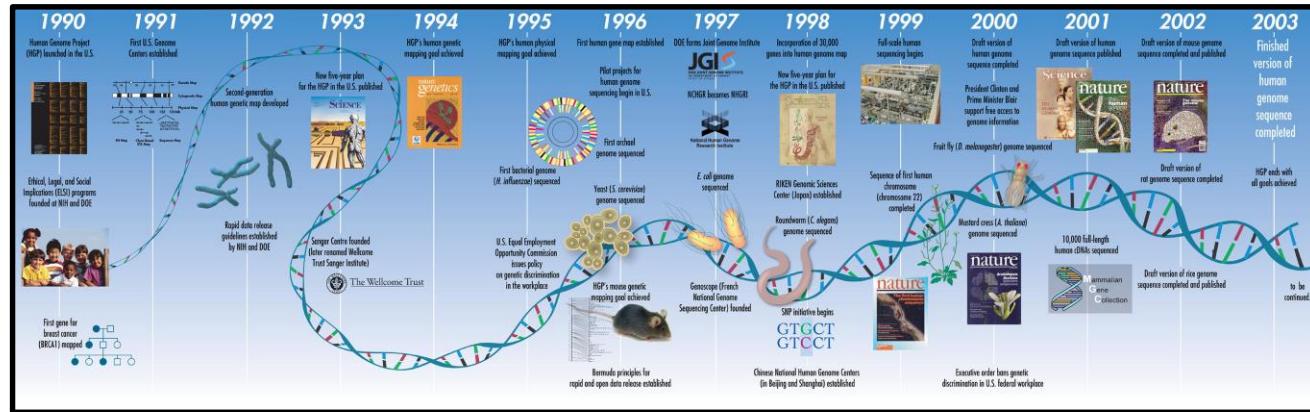
Un **gène** est une séquence génomique susceptible d'être transcrète en ARN, voire ensuite traduite en protéine selon le **code génétique** et une combinaison de ses **exons** (un transcript). Les **introns** sont des régions dites non-codantes.

La **séquençage de nouvelle génération (NGS)** ou **séquençage à haut débit** permet de lire des séquences génomiques rapidement et de façon ciblée : lecture d'un ensemble de gènes, lecture de l'exome complet (WES) ou du génome entier (WGS).



Introduction, les variants

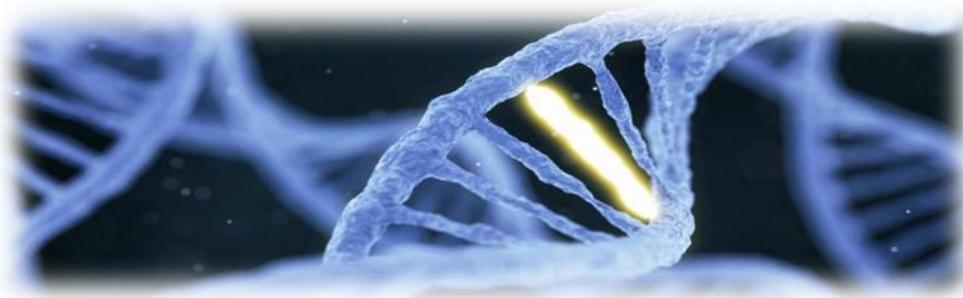
Le séquençage complet du génome humain a été réalisé en 2003 par le consortium international «**Human Genome Project**». Depuis son achèvement, nous disposons d'un **génome de référence** établi sur la base du séquençage de quelques volontaires Nord-américains.



On parlera de **variant génétique**, ou variant, pour définir une variation de l'ADN par rapport à cette référence.

Un variant dont la fréquence dans une population (MAF) est supérieure ou égale à 1% est dit **polymorphisme**.

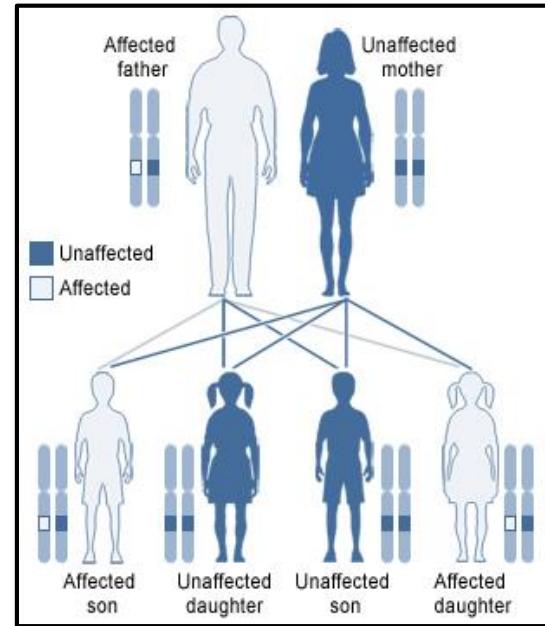
Introduction, origine des variants



Les **mutations génétiques**, soit des modifications rares, accidentnelles ou provoquées, de l'information génétique dans le génome. Elles jouent un rôle dans les processus biologiques normaux et anormaux, y compris: l'évolution, le cancer et le développement du système immunitaire. On distinguera :

- Les **mutations germinales ou « de novo »**, portant sur l'ADN des cellules souches d'un gamète.
- Les **mutations somatiques ou acquises**, portant au contraire sur l'ADN de cellules somatiques.
 - Elles peuvent intervenir à tout moment dans la vie de l'organisme.
 - Une mutation somatique sur une cellule de l'œuf après fécondation peut être à l'origine d'une **mosaïque**.
 - La prolifération par mitose de cellules portant des mutations somatiques particulières peut former un cancer.

Héritage : les variants sont hérités d'un chromosome de chaque parent lors de la fécondation.

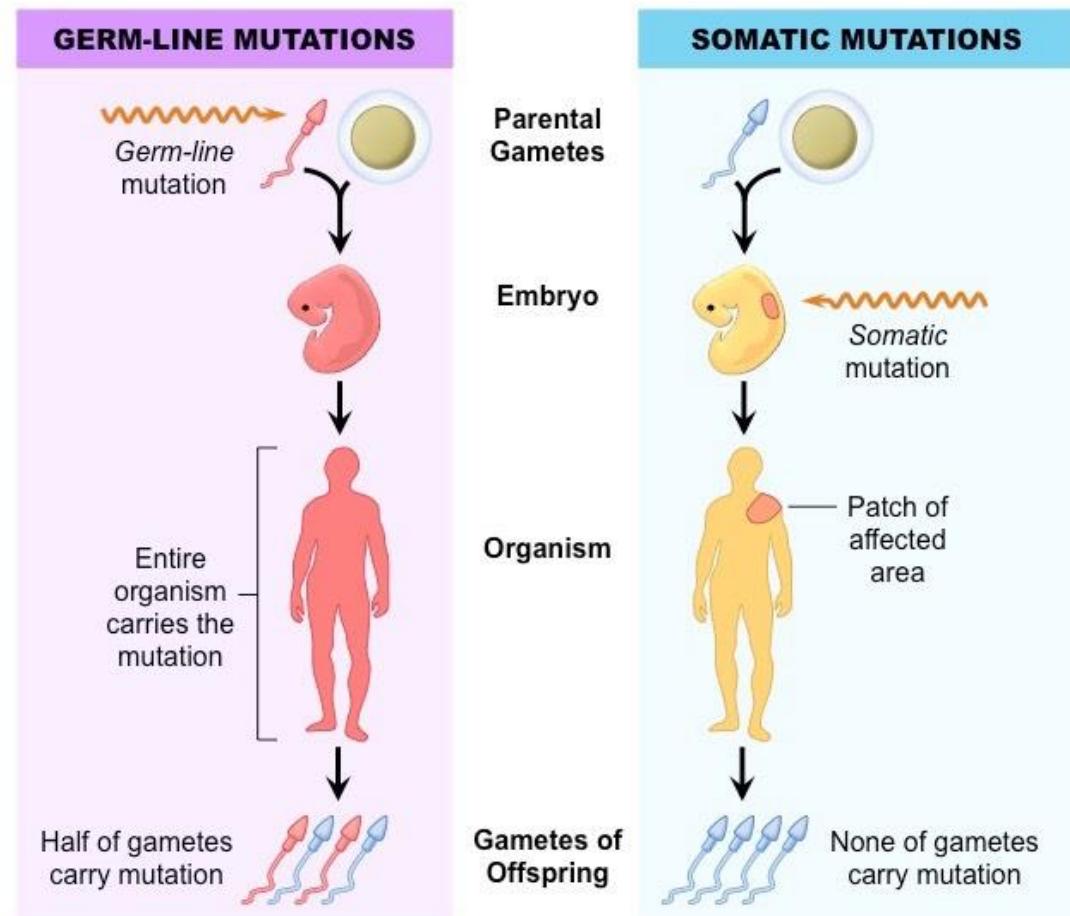


Introduction, origine des variants

Les études sur **ADN constitutionnel** permettent d'étudier en particulier les causes d'une pathologie et son caractère héréditaire (variants hérités, mutations *de novo*).

Le séquençage est généralement réalisé à partir d'un prélèvement sanguin sur tube EDTA.

Les études sur **ADN somatique**, portant sur des échantillons de tissu somatique (ex. tissu tumoral), permettent en particulier d'étudier les caractéristiques d'une pathologie.



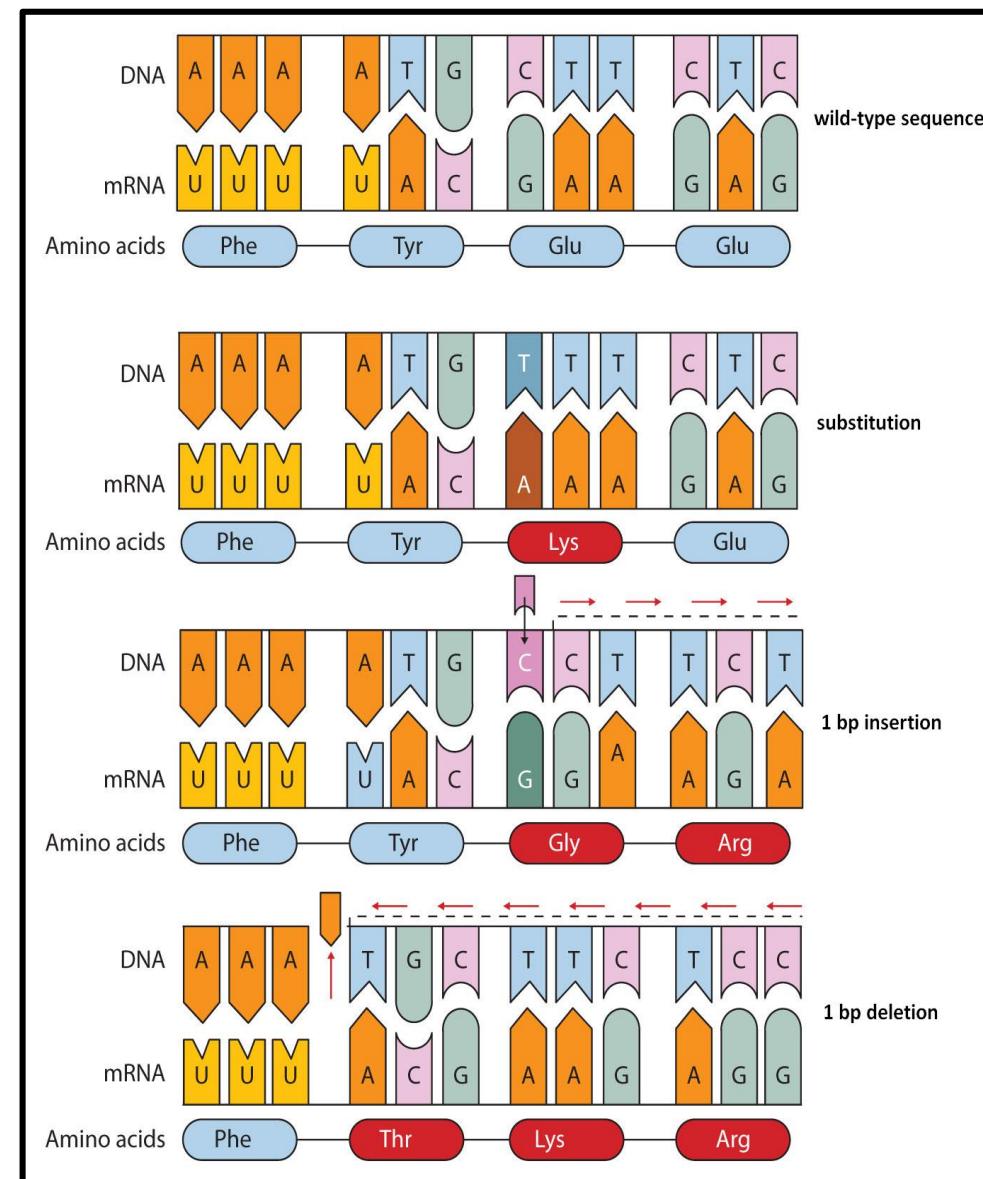
Introduction, conséquences des variants

Les « **SNV** » (Single Nucleotide Variant), variation d'un seul nucléotide, pour lesquels on distinguera entre autres :

- **Synonyme** : mutation silencieuse sur un exon, sans changer la séquence de la protéine ;
- **Faux sens** (missense) : mutation sur un exon, induisant le changement de l'acide aminé associé → protéine modifiée ;
- **Non sens** : mutation sur un exon, induisant le changement du codon associé par un codon stop → protéine tronquée ;
- **Transversion** : remplacement d'une base purine (A ou G) par une base pyrimidine (C ou T), ou inversement ;
- **Transition** : inversion entre deux bases purines, ou deux bases pyrimidines.

Les « **InDels** », insertion ou délétion de nucléotides dans la séquence, pour lesquels on distinguera :

- **In-frame** : insertion ou délétion d'un multiple de 3 bases, n'induisant pas de décalage du cadre de lecture ;
- **Frameshift** : insertion ou délétion induisant un décalage du cadre de lecture.

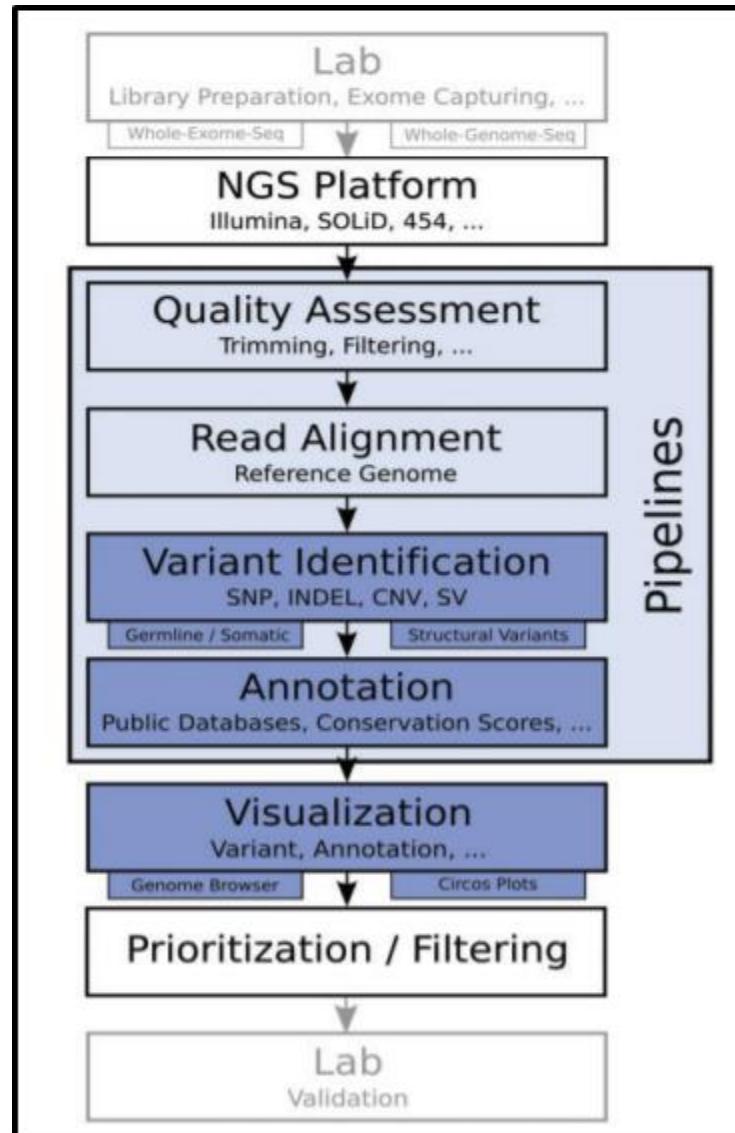


Introduction, sources de données

Pipeline Bio-Informatique

Des données de sortie du NGS à la production de données exploitables annotées:

- filtres de données
- identification des séquences
- identification des variants
- annotations de gène, AA change, conséquences, scores de qualité, scores de « dangerosité » , ...



fichiers .fastq

fichiers .bam / fichier de couverture

fichiers .vcf / fichiers .gvcf

fichiers d'annotation
par variant ou par variant*transcrit

En sortie, les biostatisticiens pourront se baser sur les fichiers de couverture, les .vcf ou .gvcf et les fichiers d'annotation

Introduction, sources de données

Les **fichiers .fastq** : servent aux QC, pour enlever les bases de mauvaises qualités.

fastQ

```
@HWI-ST1136:117:HS055:3:1101:1134:2244 1:N:0:GCCAAT  
GCCCGCCGAGCCGGGCCGTGGCCCGCCGGTCCCCGTCCCAGGGTTGG  
+  
@CCFFFDFHHHHJJIGIJJJGGICHEBB<@67=BBB2<@DD6@BB5<@D
```

Identifiant
Séquence
Qualité

Les **fichiers .vcf** (Variant Calling File) : listent les variations détectées par rapport aux génomes de référence.

Clé : position génomique

Les **fichiers .gvcf** (genomic VCF) sont construits sur le même standard que les .vcf mais donnent des informations sur toutes les positions, même les non variantes

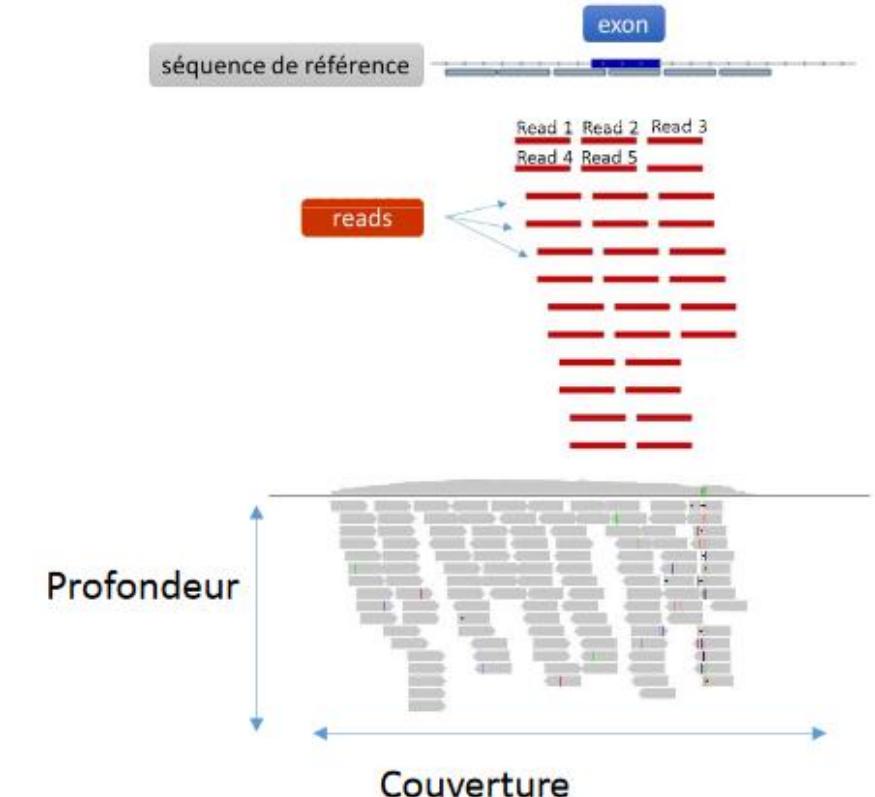
Clé : position génomique

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
1	45792777	.	A	T	374.78	PASS	AC=1;AC_AFR=0;AC_AMR=1;AC_Adj=1;AC
1	45792807	.	C	T	1527.07	AC_Adj_0_Filter	AC=1;AC_AFR=0;AC_AMR=0;AC_A
1	45792815	rs72890563	A	G	129554.81	PASS	AC=152;AC_AFR=138;AC_AM
1	45793074	.	TG	T	604.96	PASS	AC=1;AC_AFR=0;AC_AMR=0;AC_Adj=1;AC
1	45793078	.	GGACGCCGGCGCTGCAACCCGGGAGCT	G	92.01	VQSRTTrancheINDEL99.	
1	45793079	.	G	A	57.98	AC_Adj_0_Filter	AC=1;AC_AFR=0;AC_AMR=0;AC_A
1	45793262	.	C	T	1608.10	PASS	AC=2;AC_AFR=0;AC_AMR=1;AC_Adj=2;AC
1	45793264	rs148226888	C	T,G	2671.88	PASS	AC=1,4;AC_AFR=0,2;AC_AMR=0,
1	45793286	.	G	T	1229.30	PASS	AC=2;AC_AFR=1;AC_AMR=0;AC_Adj=2;AC
1	45793289	.	T	C	1768.19	PASS	AC=2;AC_AFR=0;AC_AMR=0;AC_Adj=2;AC

Les **fichiers de couverture / profondeur** :

La couverture est la zone couverte par au moins une lecture, exprimée en % et la profondeur est le nombre de lectures de chaque base, exprimée en X.

Clé : individu*(position génomique)



Introduction, données publiques

1000 genomes project, phase 3

<http://www.internationalgenome.org/>

84,4 millions de variants



2 504 individus dont 669 européens

Données individuelles téléchargeables, génome complet

The Haplotype Reference Consortium

<http://www.haplotype-reference-consortium.org/home>

64 976 haplotypes

Serveur d'imputation <https://imputationserver.sph.umich.edu/>

Ensembl

http://grch37.ensembl.org/Homo_sapiens/Info/Index



Base d'annotation avec choix de l'espèce, de l'assemblage (attention au choix de la version)

Nom des gènes et alias connus

Liste des transcrits RefSeq ou transcrits « hypothétiques » (la colonne Flags indique le niveau)

Exome Aggregation Consortium

<http://exac.broadinstitute.org/>

60 706 individus

Séquençage d'exons (seulement)

Annotation, MAF, prédiction d'effet et conséquence,...



Données agrégées téléchargeables (MAF par population)

Genome Aggregation Database (gnomAD)

<http://gnomad.broadinstitute.org/>

Développement du projet ExAC

123 136 exomes et 15 496 genomes



Données agrégées téléchargeables (MAF par population)

GWAS Catalog

<https://www.ebi.ac.uk/gwas/>

Base de SNPs identifiés par GWAS

Liste par étude, maladie, gène, publication...



Utile pour confronter ses résultats

Introduction, les études portant sur les variants

Les **études de liaison** localisent les régions contenant les gènes responsables du trait ou de la maladie sur le génome au moyen d'observations sur des individus **apparentés** (échantillons de familles). Cette stratégie permet de mesurer la co-ségrégation due à l'existence de liaisons génétiques entre des loci.

Les **études d'association** identifient des allèles de variants associés à un phénotype (risque accru de développer une maladie en présence de ce variant, avec une pénétrance pouvant être faible).

Le lien de cause à effet entre cet allèle et le risque accru n'est pas connu à l'issue de ces études.

Les études Cas-Témoin (ou Cas-Contrôle)

« *Les études cas-témoin sont utilisées pour mettre en évidence des facteurs qui peuvent contribuer à l'apparition d'une maladie en comparant des sujets qui ont cette maladie (les cas) avec des sujets qui n'ont pas la maladie mais qui sont similaires par ailleurs (les contrôles)* »

Mann (2003) Emerg Med J;20:54–60

Biais d'information

Les cas sont bien tous des cas, les témoins tous des témoins ? Vérifier le mode de saisie et la cohérence des valeurs.

Biais de sélection / Biais de confusion

Les 2 groupes doivent être comparables (vis-à-vis des variables confondantes selon le sujet d'étude).

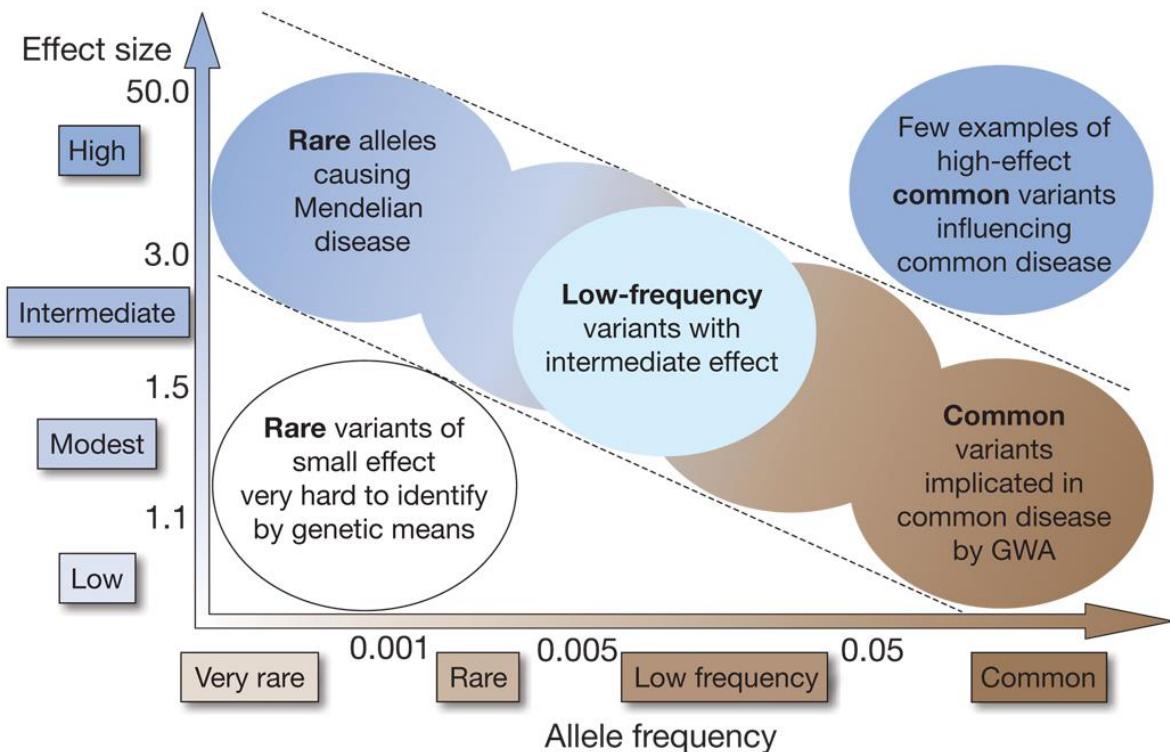
Exemple 1 : Le risque héréditaire de cancer du sein et de l'ovaire ne dépend ni de l'âge, ni du sexe. Il peut par contre dépendre de l'origine ethnique/géographique.

Exemple 2 : le diabète de type II est tout à fait lié à l'âge il sera donc important d'équilibrer les 2 groupes cas et contrôle.

→ d'où la nécessité de contrôler les facteurs confondants disponibles dans les jeux de données.

Introduction, problématique des variants rares

Identification de variants rares associés



Manolio et al. Finding the missing heritability of complex diseases
Nature. 2009 Oct 8; 461(7265): 747–753

Dans une analyse **variant par variant** :

- Avec un variant d'une fréquence allélique de 0,1% en population générale
- Au degré de significativité $5 \cdot 10^{-6}$ (correspondant à 100.000 tests indépendants)
- Pour une puissance de 80%

Odds-ratio Cas/Témoin	Nb de sujets nécessaires par groupe
2	83 000
3	27 000
5	10 000
10	4 000

Par une méthode d'**agrégation de variants** lorsque MAF $\leq 1\%$

- Le principe est de scorer la présence ou l'absence de variants rares par individus.
- On recherche ensuite des associations entre ces scores et les traits phénotypiques.

Introduction, problématique des variants rares

Limites de l'étude des variants rares

Même par une méthode d'agrégation, l'étude des variants rares nécessite de nombreuses données individuelles

- pour disposer de suffisamment d'informations sur les variants
- pour pouvoir quantifier leur association avec un risque

Il n'existe pas de consensus pour quantifier le caractère délétère d'un groupe de variants.

Les grandes cohortes de données individuelles en population générale mises à disposition par des consortia ne sont pas forcément comparables aux cas

- méthode de séquençage
- filtres qualité
- origines ethniques diverses
- ...

La réalisation des contrôles qualité (QC) des données est une étape primordiale.

Le moindre « bruit de fond » peut entraîner une conclusion erronée.

Quality controls, QC des individus

QC réalisés sur les données filtrées (certains variants exclus préalablement) par des critères qualité du pipeline

Ex. GATK VQSLOD=PASS ou GQ > 20

Pourcentage de génotypes manquants par individu

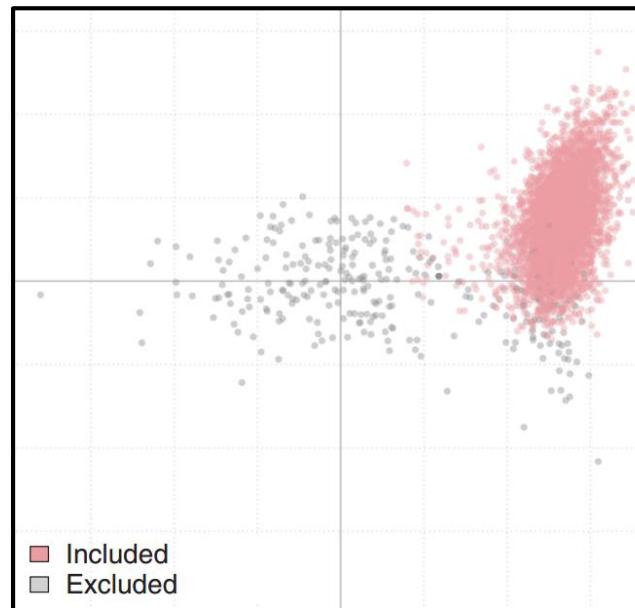
Ex. > 5% → exclusion de l'individu

DP (depth) minimum sur la zone de capture

Ex. 8X pour le SNP ou 20X pour InDels

Homogénéité de la population

- Analyse des distances entre individus
Ex. ACP sur les variants fréquents (MAF > 5%)
- Adéquation avec une population cible :
Ex. ACP + LDA sur les variants fréquents (MAF > 5%)
- Ratio nombre de variants hétérozygotes/homozygotes
- Ratio transition/transversion
- Nombre de singletons (variants uniques dans la population)



Il s'agit de contrôles graphiques, des outliers pourraient soit ne pas appartenir à la population d'étude, soit indiquer un problème lors du séquençage ou sur le pipeline.

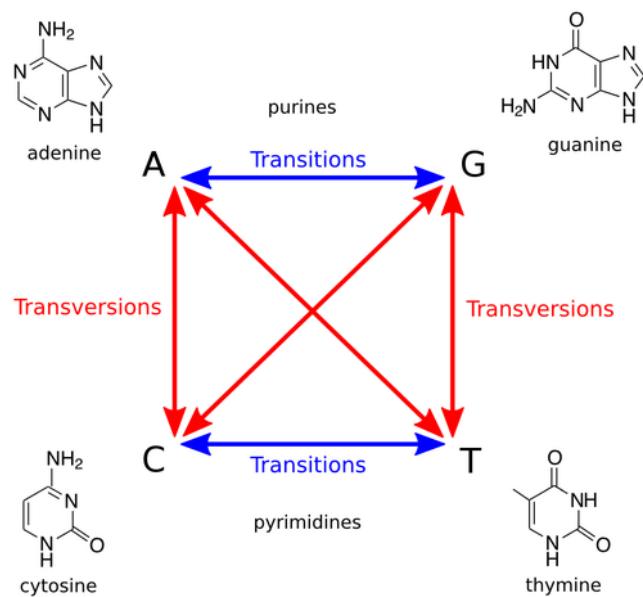
Quality controls, QC des individus

Ratio transition/transversion

2.8 sur le genome

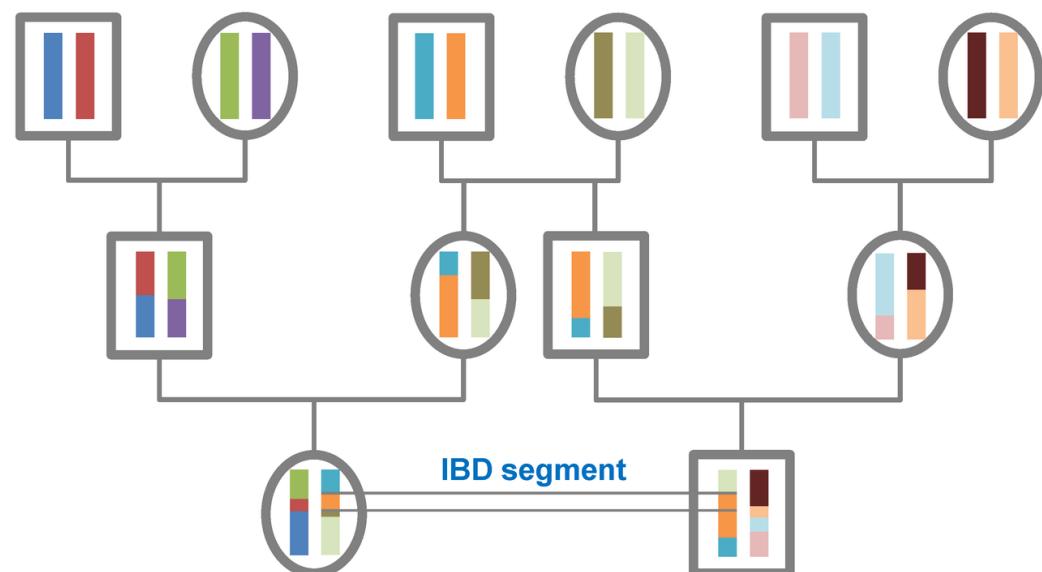
2.1 sur l'exome

Des valeurs inférieures pourraient indiquer des faux positifs.



Calcul des distances IBS (Identical By State) entre les individus

- Le nombre de segments identiques permet d'estimer le degré d'apparenté après multiples recombinaisons (générations)
- Deux segments IBS hérités du même ancêtre sont dit IBD (Identical By Descent)
- PLINK method for pairwise IBD segment detection*



Quality controls, QC des variants

Score Q ou PHRED : probabilité d'erreur d'identification d'une base

permet un contrôle qualité « technique » à chaque position génomique

Calculé à partir de plusieurs paramètres relatifs à la forme et la résolution du pic d'électrophorèse de chaque base.

Ex : un score de 10 \Leftrightarrow Erreur = $10^{-10}/10 = 10\%$; un score de 20 \Leftrightarrow Erreur = $10^{-20}/10 = 1\% \dots$

QUAL et DP du fichier .vcf

la qualité (a Phred-scaled probability that a REF/ALT polymorphism exists at this site given sequencing data.) et la profondeur (DP the filtered depth, at the sample level) des reads pour un site donné

Pourcentage de génotypes manquants

Ex : > 5% => exclusion du variant car trop mal couvert

Test de Fisher sur le nombre de génotypes manquants cas/contrôles

Contrôles manuels des variants

- classés délétères en dehors des zones codantes (CDS)
- avec une MAF incohérente avec une population de référence

VQSLOD>=99.5% pour les SNV, VQSLOD>99% pour les Indels

Seuils conseillés par GATK

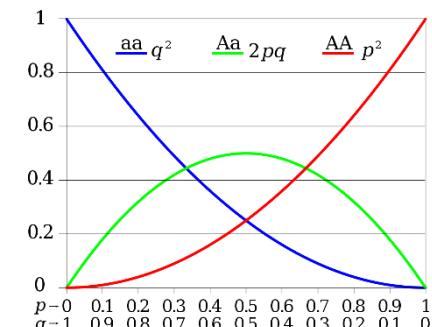
But : Filtrer les variants d'une façon qui permet d'équilibrer la sensibilité (essayant de découvrir les vrais variants) et la spécificité (essayant de limiter les faux positifs qui s'insinuent quand les filtres sont tolérants).

Balance allélique

Ex : ABHom >90% ou 30%<ABHet <70%

Idéalement, quand le genotype est heterozygote la valeur AB devrait être proche de 0.5 : car la moitié des reads auraient l'allèle REF et l'autre moitié auraient l'allele ALT.

Test d'équilibre de Hardy-Weinberg



Analyses de données, variants fréquents

SNP fréquent (MAF > 1%) : Étude locus par locus

Modèle linéaire généralisé:
$$Y = f(\alpha X + \beta SNP) + \varepsilon$$

où

Y est un trait phénotypique,

X est une matrice de covariables (Ex. Age, Sex, BMI...),

SNP est le variant (ex codé 0,1,2 pour un modèle additif, pour AA, Aa et aa avec A allele de référence),

f est la fonction de lien du modèle (ex linéaire ou logistique),

α et β sont les paramètres à estimer.

Le test du paramètre β : $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$

permet de conclure sur une relation entre le SNP et Y .

Analyses de données, variants rares

Pourquoi ne pas faire pareil avec les variants rares?

“ While single variant tests are typically conducted to investigate associations of common variants and phenotypes, the same approach has little power for testing for rare variant effects due to their low frequencies and large numbers.

Instead, the statistical development of rare variants analysis has been focused on testing cumulative effects of rare variants in genetic regions or SNP sets, such as genes. ”

[Optimal unified approach for rare variant association testing with application to small sample case-control whole-exome sequencing studies. Lee et al.]

=> Tests d'agrégation de variants

Adaptative burden test

L'effet cumulatif des variants rares est toujours testé mais de façon moins restrictive (non linéaire, $w_j < 0$, ...) et donc plus adapté à certains cas de figure

Exemple : KBAC

Les motifs de génotypes identifiés au sein de la région génomique sont pondérés par un noyau de densité hypergéométrique "hypergeometric density kernel function", puis comparés pour identifier des différences entre cas et témoins.

Les génotypes sont recodés en m motifs P_m

$$f(G_i) = \gamma K_m$$

K_m : noyau défini pour le pattern P_m .
Ex. pour un phénotype dichotomique, un noyau hypergéométrique est donc approprié

Exemple : aSum

Le test de somme adaptative reprend la structure linéaire de WST mais s'affranchit de la contrainte des $w_j \geq 0$

Han and Pan's weighting scheme :

$$w_j = 1 \text{ si } |\gamma_j| \leq 0, \text{ et } w_j = -1 \text{ sinon,}$$

Où γ_j est une estimation de l'effet de l'allèle mineur pour le $j^{\text{ème}}$ variant sur le phénotype

Burden test

Collapse rare variants in a genetic region into a single burden variable

- Test l'effet cumulatif des variants rares regroupés par région génétique

- WST : $f(\mathbf{G}_i) = \beta \sum_j w_j G_{ij}$
- CAST : $f(\mathbf{G}_i) = I[\sum_j w_j G_{ij}]$

β : l'effet, sur le trait, de la somme des variants rares, chacun pouvant être pondéré par un poids

$w_j \in [0,1]$: le poids donné au $j^{\text{ème}}$ variant rare

Exemple : Le poids w_j peut donner plus ou moins d'importance selon la fréquence allélique.

Madsen and Browning's weighting scheme :

$$w_j = \frac{1}{\sqrt{q_j(1-q_j)}} \text{ avec } q_j \text{ la MAF du } j^{\text{ème}} \text{ variant}$$

Exemple :
CAST, CMC, WST

Dans tous les cas, on aura le modèle suivant :

$$Y_i = \mathbf{a} + f(\mathbf{G}_i)$$

f est la fonction appliquée au génotype G_i du $i^{\text{ème}}$ individu, typiquement codé $G_{ij} = \{0, 1, \text{ or } 2\}$ pour le $j^{\text{ème}}$ variant rare

Instead of aggregating variants, aggregates individual variant score test statistics with weights

- Test la variance au sein de la région génétique comprenant les variants rares

$$f(\mathbf{G}_i) = \sum_j \beta_j G_{ij}$$

β_j : Effet de chaque copie de l'allèle mineur du $j^{\text{ème}}$ variant rare et $\beta_j \sim N(\mathbf{0}, \tau w_j^2)$,

$w_j \in [0,1]$: le poids donné au $j^{\text{ème}}$ variant rare

τ : la variance inconnue de la région génétique

Combined test



Sous l'hypothèse nulle de non association entre le trait et les variants rares de la région génétique, $\beta_j = 0$ pour tout j , est donc équivalent à $\tau = 0$

Exemple :
SKAT, C-alpha

Combiner les deux types de tests :
Tester l'effet agrégé des variants et leur variance au sein de la région génétique

Exemple :

MiST (Mixed effects Score Test), SKAT-O, C-alpha generalised.

MiST propose : π score statistique de l'effet moyen, τ score statistique de l'hétérogénéité de l'effet et une combinaison des scores (p-valeurs) avec la procédure de Fisher.

Variance-component test

SKAT, C-alpha

- + Permet l'interaction SNV-SNV
- + Spécifiquement puissant en présence à la fois de variants protecteurs et délétères ou de beaucoup de variants non-causaux

- Peut être moins puissant que les burden tests si une large proportion des variants rares de la région sont causaux et influencent le phénotype dans la même direction.

Combined test

MiST, SKAT-O, C-alpha generalised

- + Applicable à des traits binaires ou continues
- + SKAT-O et MiST ont été présentés comme performants pour différents types de phénotypes, avec peu de variants causaux et pouvant aller dans des directions différentes
- Suppose que la classification des variants est bien connue (matrice Z)
- Exemple : définition des clusters, au sein de la région génétique, basée sur des annotations de conséquences de variants ou sur des scores de prédition d'un effet délétère

Burden test

CAST, CMC, WST

- + les Burden Tests supposent implicitement que tous les variants rares de la région génétique sont causaux et affectent le phénotype dans la même direction
- Souffre d'une perte de puissance substantielle quand ces hypothèses ne sont pas respectées ou qu'une petite partie seulement des variants sélectionnés sont causaux

Adaptative burden test

KBAC, aSum

- + Permet les interactions entre variants et gènes
- + Développé pour surmonter les problèmes de détection d'association de variants rares en présence d'une mauvaise classification de ces derniers
- + Permet aux variants rares d'une région génétique d'avoir des effets de différentes directions et/ou magnitudes sur un trait binaire
- Adapté à quelques cas de figure seulement (cadre bien défini à vérifier avant application)

Avantages & inconvénients

Retours d'expérience 1

Méthode

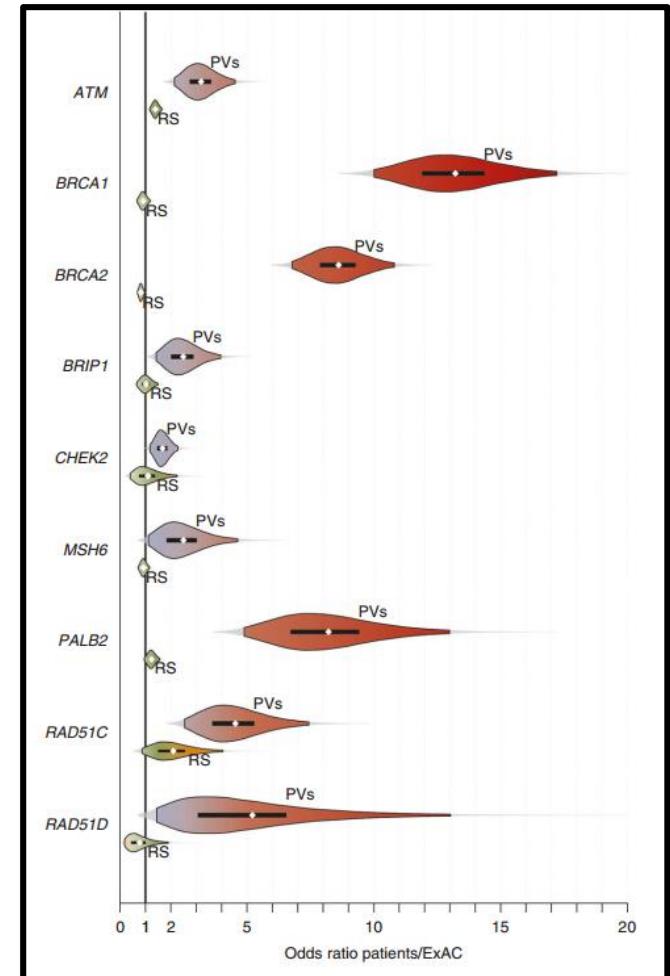
- Étude d'un cas/contrôle sur le risque héréditaire de cancer du sein et de l'ovaire
- Stratification des populations selon une analyse factorielle des données 1000 genomes project
- Sélection des variants classés fortement délétères
 - frameshift
 - non sense
 - base de connaissances
 - algorithmes SIFT, polyphen-2, mutationTaster
 - abolition des sites d'épissage (MaxEntScan, SplitSiteFinder)
- Calcul d'OR entre Cas/Témoins
- Estimations d'OR entre Cas et population Européenne ExAC non TCGA
- Application des tests d'agrégation CAST et WST à partir du package R 'AssotesteR'

Publication :

Landscape of pathogenic variations in a panel of 34 genes and cancer risk estimation from 5131 HBOC families.

Genet Med. 2018 Jul 10

Castéra L, Harter V, Muller E, Krieger S, Goardon N, Ricou A, Rousselin A, Paimparay G, Legros A, Bruet O, Quesnelle C, Domin F, San C, Brault B, Fouillet R, Abadie C, Béra O, Berthet P; French Exome Project Consortium, Frébourg T, Vaur D



Retours d'expérience 2

Méthode :

Étude d'un cas/contrôle sur le diabète

Analyse des variants rares avec la méthode KBAC couplée à une régression logistique pour tester l'association des patterns avec le phénotype binaire « diabète »

Publication :

Type 2 diabetes-associated variants of the MT₂ melatonin receptor affect distinct modes of signaling.

Sci Signal. 2018 Aug 28;11(545).

Karamitri A, Plouffe B, Bonnefond A, Chen M, Gallion J, Guillaume JL, Hegron A, Boissel M, Canouil M, Langenberg C, Wareham NJ, Le Gouill C, Lukasheva V, Lichtarge O, Froguel P, Bouvier M, Jockers R.

Remarques :

Élargissement de l'analyse impossible à d'autres traits quantitatifs avec cette méthode

Retours d'expérience 3

Méthode :

Analyse des variants rares avec la méthode MiST :

- Étude d'association avec différents traits (binaires : diabète et obésité, quantitatifs : FG et BMI)
- Le cluster (région génomique) est à l'échelle d'un gène = un transcrit
(l'étude par transcrit est plus adapté si on pense à l'exploitation fonctionnelle qui suivra l'analyse)
- Utilisation de « subcluster » possible pour regrouper des gènes ou des variants rares selon la prédition de leur conséquence ou selon un niveau de pathogénicité (recherche en cours sur les scores *in silico* et usage de la base dbNSFP - annotation database about deleteriousness of SNVs)
- Stratification ethnique des données autorisée mais corrigée : ACP réalisée sur les données d'études conjointement à des données de référence (1000 genomes project) et ajout des PC1 et PC2 dans l'analyse

Publication : En cours.

Quelques sources

- Broadway et al. Kernel Approach for Modeling Interaction Effects in Genetic Association Studies of Complex Quantitative Traits. *Genet Epidemiol.* 2015 Jul;39(5):366-75
- Lee et al. Optimal unified approach for rare variant association testing with application to small sample case-control whole-exome sequencing studies. *Am J Hum Genet.* 2012 Aug 10;91(2):224-37
- Lee et al. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *Am J Hum Genet.* 2014 Jul 3; 95(1): 5–23
- Manolio et al. Finding the missing heritability of complex diseases. *Nature.* 2009 Oct 8; 461(7265): 747–753
- Moutsianas and Morris. Methodology for the analysis of rare genetic variation in genome-wide association and re-sequencing studies of complex human traits. *Brief Funct Genomics.* 2014 Sep; 13(5): 362–370
- Moutsianas et al. The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLoS Genet.* 2015 Apr 23;11(4)
- Sun J et al. A unified mixed-effects model for rare-variant association in sequencing studies. *Genet Epidemiol.* 2013 May;37(4):334-44
- Wang et al. Genome measures used for quality control are dependant on gene function and ancestry. *Bioinformatics.* 2015 Feb 1; 31(3): 318–323

Références des illustrations

Les séquences génomiques

- <http://blog.illumina.com/blog/illumina/2014/04/30/next-generation-sequencing-challenges-and-clinical-translation>
- <http://fr.academic.ru/dic.nsf/frwiki/49111>
- Mardis ER. Next-generation sequencing platforms. *Annu Rev Anal Chem*(Palo Alto Calif). 2013;6:287-303

Origine de variants

- <http://www.microbiologybook.org/French-immuno/immchapter19.htm>
- <http://www.cell.com/pb-assets/marketing/sliders/19192%20CPN%20CRISPR%20gene%20editing%20592x200.jpg>
- <http://ib.bioninja.com.au/standard-level/topic-3-genetics/33-meiosis/somatic-vs-germline-mutatio.html>

Sources de données

- Pabinger S et al. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform.* 2014 Mar;15(2):256-78
- Images «fastQ» et «couverture/profondeur» :
http://www.canceropole-gso.org/download/fichiers/3606/NGS+Canceropole+GSO+mai+2016+AGros_DIFF.pdf

Problématique des variants rares

- Manolio et al. Finding the missing heritability of complex diseases *Nature*. 2009 Oct 8; 461(7265): 747–753

Les variants

- https://www.mun.ca/biology/scarr/Human_Genome_Project_timeline.html
- <https://cdn1.sph.harvard.edu/wp-content/uploads/sites/21/2012/11/dnavarant2.jpg>

Conséquences des variants

- https://catalog.flatworldknowledge.com/bookhub/reader/2547?e=gob-ch19_s05

QC des individus

- Modèle transition/transversion: http://drdk.me/transition_transversion.html
- https://en.wikipedia.org/wiki/Identity_by_descent

QC des variants

- <https://ipfs.io/ipns/tr.wikipedia-on-ipfs.org/l/m/Hardy-Weinberg.svg.png>

MERCI de votre attention !



Merci également à

Amélie Bonnefond

Mickaël Canouil

Iandry Rabearivelo

Franck De Graeve

Olivier Sand

Et l'ensemble de l'équipe NGS

Laurent Castéra

Etienne Muller

Et l'ensemble de l'équipe du LBGC du centre François Baclesse