

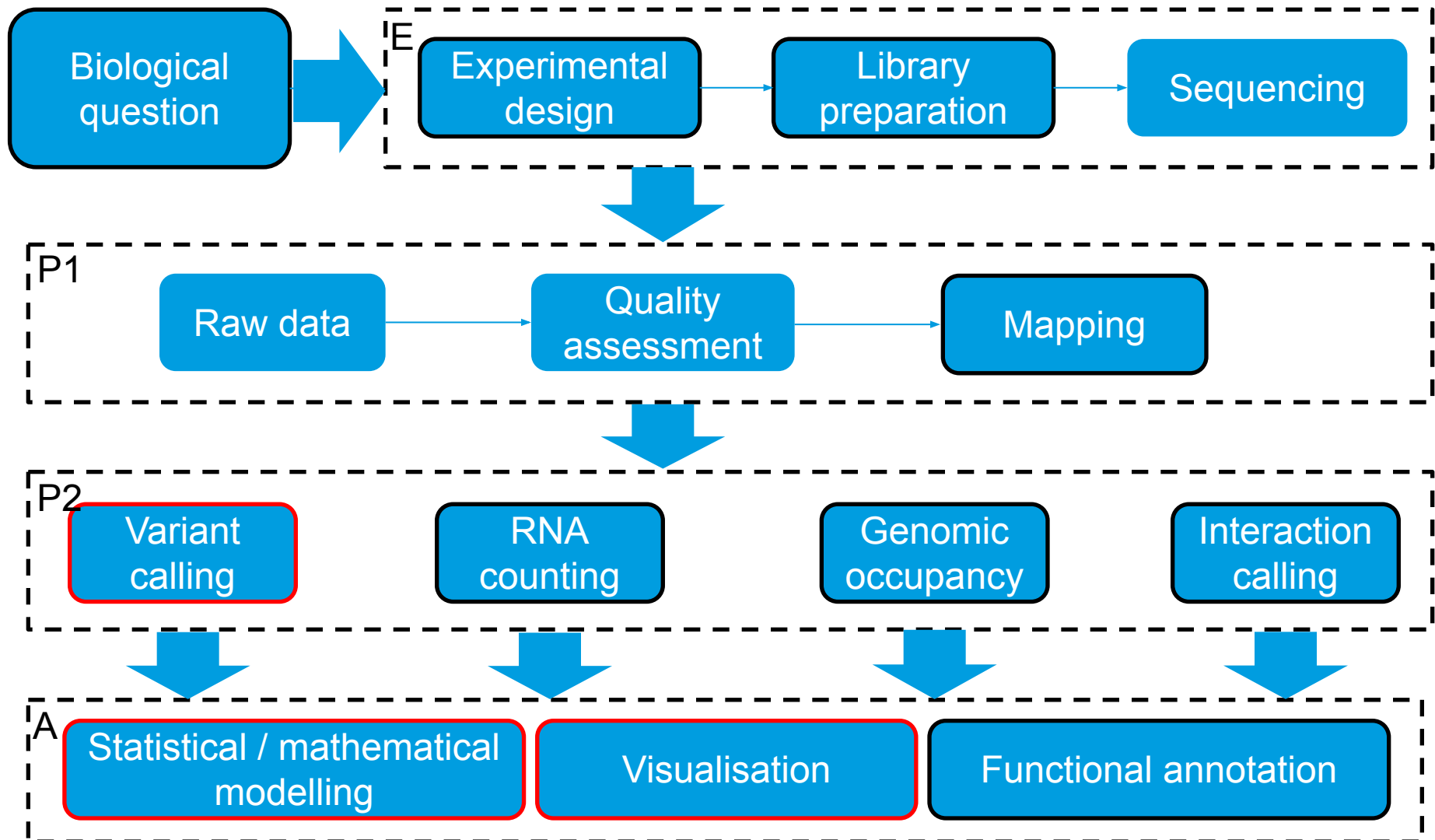
The logo of the University of Bordeaux is displayed against a background with a blue diagonal stripe in the top left and a dark grey diagonal stripe in the bottom right. The text 'université' is in a dark brown, lowercase, sans-serif font, with a blue square above the 'i' and a blue 'e'. Below it, the word 'de' is in a smaller, dark brown, lowercase, sans-serif font. To the right of 'de', the word 'BORDEAUX' is in a bold, dark brown, uppercase, sans-serif font.

université
de **BORDEAUX**

Introduction: Analyse de variants génomiques et de données RNA-seq

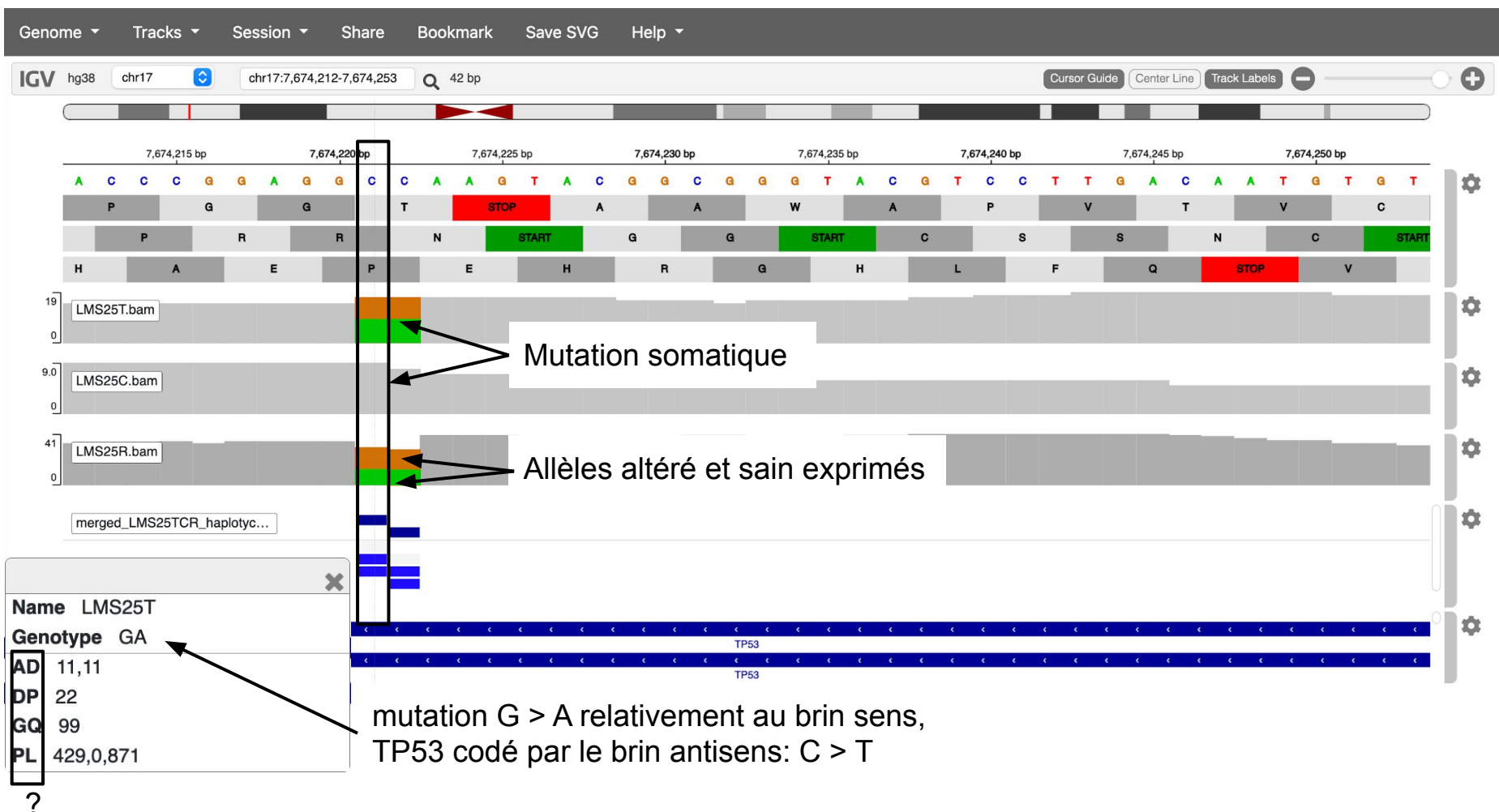


Pipeline bioinformatique



Integrative Genome Viewer

- Visualisation des reads et détection du variant à la position 7674221 sur le chromosome 17



Recherche d'informations

ce que je ne
connaissais pas



DP GQ AD PL mutation GATK

Taper les bons mots clés



contexte



Environ 4 900 résultats (0,35 secondes)

Conseil : [Recherchez des résultats uniquement en français](#). Vous pouvez indiquer votre langue de recherche sur la page [Préférences](#).

gatk.broadinstitute.org › articles › 3... ▼ [Traduire cette page](#)

VCF - Variant Call Format – GATK

4 sept. 2020 — AC=2;AF=1.00;AN=2;DP=30;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=29.49;SOR=1.765 GT:AD:DP:GQ:PL 1/1:0 ...

gatk.broadinstitute.org › articles › 3... ▼ [Traduire cette page](#)

Genotype Refinement workflow for germline short ... - GATK

15 sept. 2020 — In this sense it serves as an optional extension of the **variant** calling workflow ... GT:AD:DP:GQ:PL 0/0:11,0:11:0:0,0,249 0/0:10,0:10:24:0,24,360 ...

ressources.france-bioinformatique.fr › ... ▼ [PDF](#) [Traduire cette page](#)

Variant Filtering

##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for ...

##reference=file:///tmp/13905.1.galaxy.q/tmp-gatk-MPGS7G/gatk_input.fasta ... LowQual AC=1;

AF=0.500;AN=2;DB:DP=46;FS=0.000;MQ=27.49 GT:AD:DP:GQ:PL ...

Elodie Darbo / rNGSclass part 3

Recherche d'informations

PL

"Normalized" [Phred-scaled](#) likelihoods of the possible genotypes. For the typical case of a monomorphic site (where there is only one ALT allele) in a diploid organism, the PL field will contain three numbers, corresponding to the three possible genotypes (0/0, 0/1, and 1/1). The PL values are "normalized" so that the PL of the most likely genotype (assigned in the GT field) is 0 in the Phred scale. We use "normalized" in quotes because these are not probabilities. We set the most likely genotype PL to 0 for easy reading purpose. The other values are scaled relative to this most likely genotype.

Keep in mind, if you are not familiar with the statistical lingo, that when we say PL is the "Phred-scaled likelihood of the genotype", we mean it is "How much less likely that genotype is compared to the best one". Have a look at [this article](#) for an example of how PL is calculated.

GQ

The Genotype Quality represents the [Phred-scaled](#) confidence that the genotype assignment (GT) is correct, derived from the genotype PLs. Specifically, the GQ is the difference between the PL of the second most likely genotype, and the PL of the most likely genotype. As noted above, the values of the PLs are normalized so that the most likely PL is always 0, so the GQ ends up being equal to the second smallest PL, unless that PL is greater than 99. In GATK, the value of GQ is capped at 99 because larger values are not more informative, but they take more space in the file. So if the second most likely PL is greater than 99, we still assign a GQ of 99.

Basically the GQ gives you the difference between the likelihoods of the two most likely genotypes. If it is low, you can tell there is not much confidence in the genotype, i.e. there was not enough evidence to confidently choose one genotype over another. See the [FAQ article on the Phred scale](#) to get a sense of what would be considered low.

Not to be confused with the site-level annotation QUAL; see [this FAQ article](#) for an explanation of the differences in what they mean and how they should be used.

A few examples

recherche
dans la
page

Recherche d'informations

AD and DP

Allele depth (AD) and depth of coverage (DP). These are complementary fields that represent two important ways of thinking about the depth of the data for this sample at this site.

AD is the unfiltered allele depth, *i.e.* the number of reads that support each of the reported alleles. All reads at the position (including reads that did not pass the variant caller's filters) are included in this number, except reads that were considered uninformative. Reads are considered uninformative when they do not provide enough statistical evidence to support one allele over another.

DP is the filtered depth, at the sample level. This gives you the number of filtered reads that support each of the reported alleles. You can check the variant caller's documentation to see which filters are applied by default. Only reads that passed the variant caller's filters are included in this number. However, unlike the AD calculation, uninformative reads are included in DP.

See the Tool Documentation for more details on [AD \(DepthPerAlleleBySample\)](#) and [DP \(Coverage\)](#) for more details.

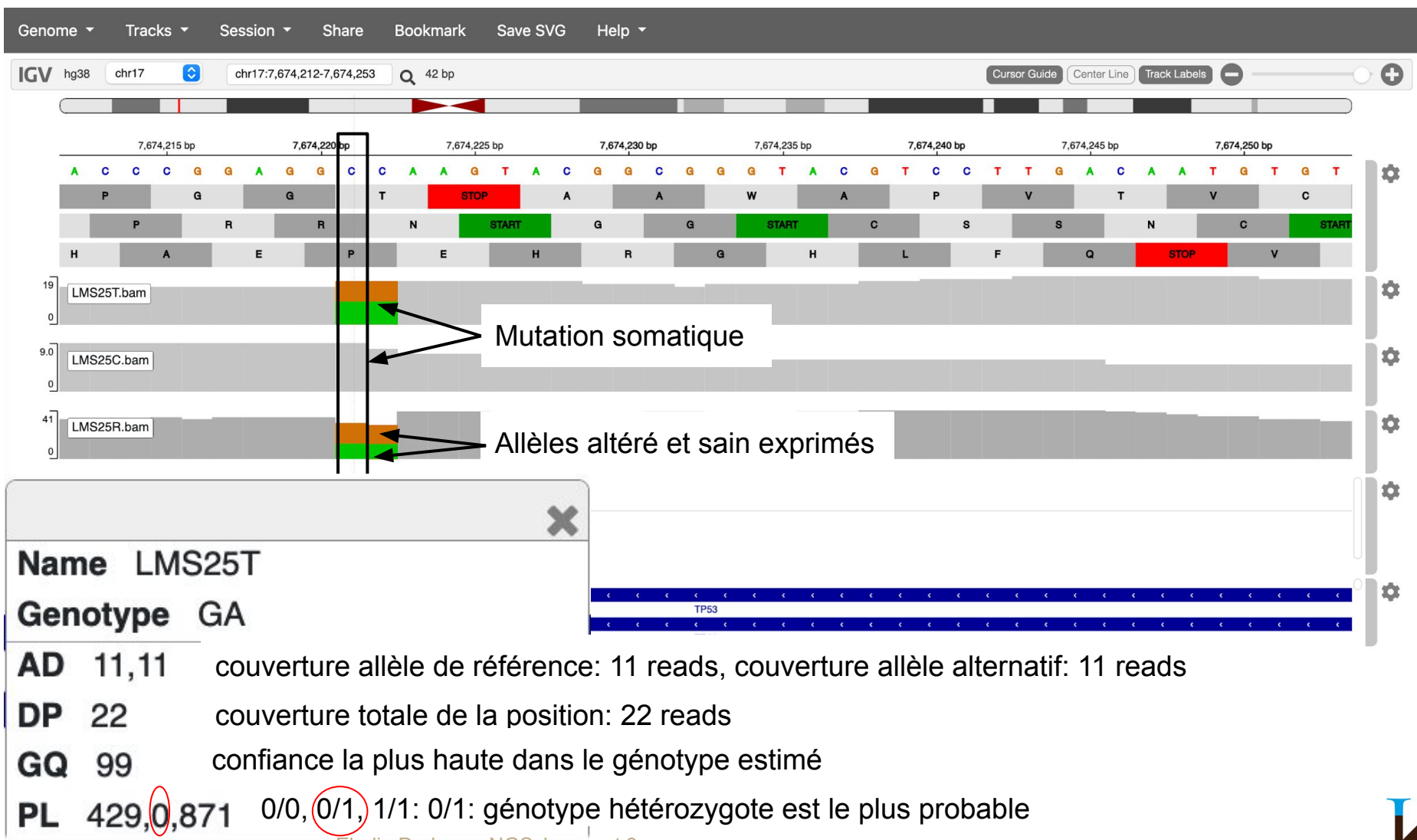
PL

"Normalized" [Phred-scaled](#) likelihoods of the possible genotypes. For the typical case of a monomorphic site (where there is only one ALT allele) in a diploid organism, the PL field will contain three numbers, corresponding to the three possible genotypes (0/0, 0/1, and 1/1). The PL values are "normalized" so that the PL of the most likely genotype (assigned in the GT field) is 0 in the Phred scale. We use "normalized" in quotes because these are not probabilities. We set the most likely genotype PL to 0 for easy reading purpose. The other values are scaled relative to this most likely genotype.

Keep in mind, if you are not familiar with the statistical lingo, that when we say PL is the "Phred-scaled likelihood of the genotype", we mean it is "How much less likely that genotype is compared to the best one". Have a look at [this article](#) for an example of how PL is calculated.

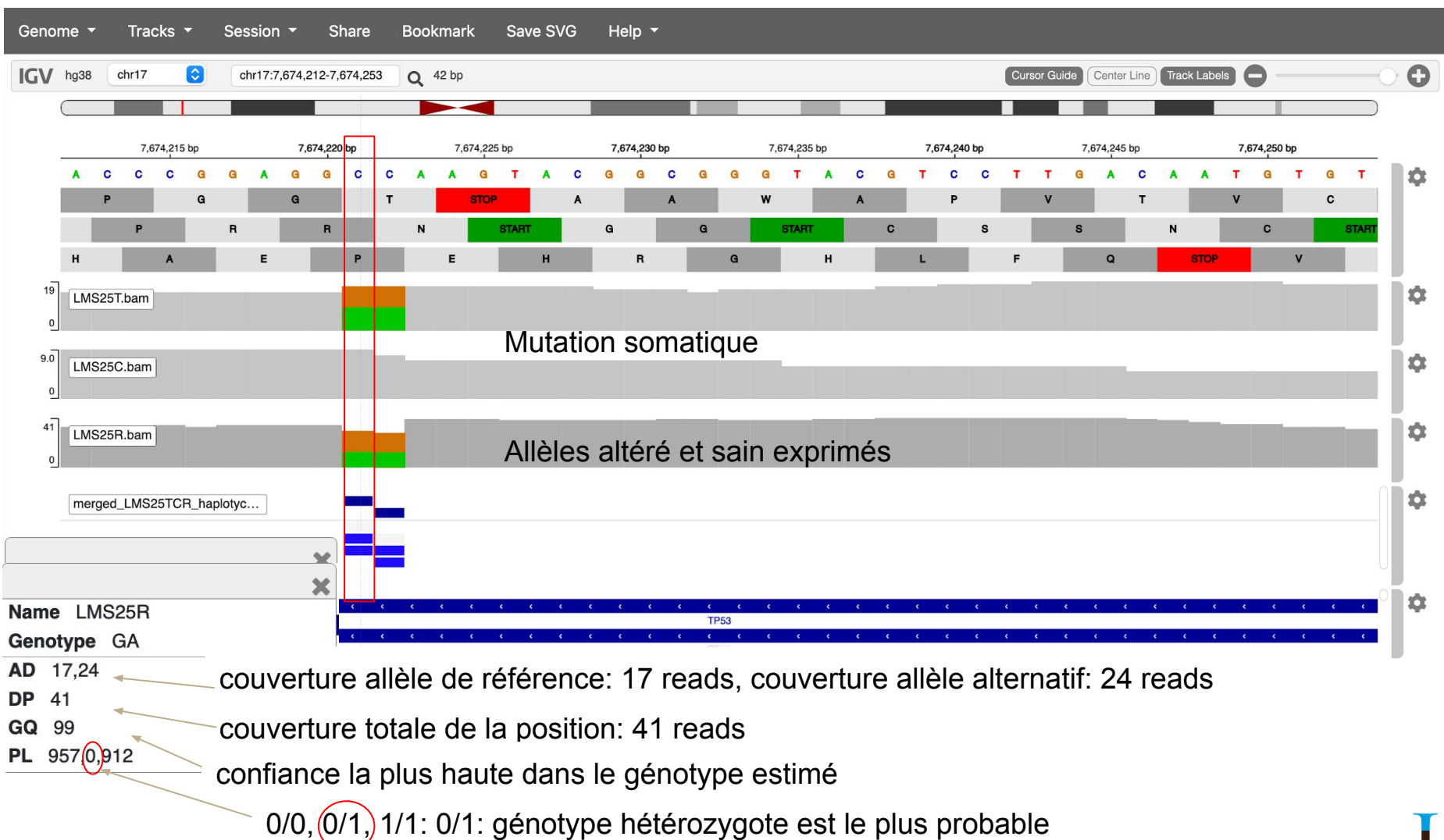
Integrative Genome Viewer

- Visualisation des reads et détection du variant à la position 7674221 sur le chromosome 17



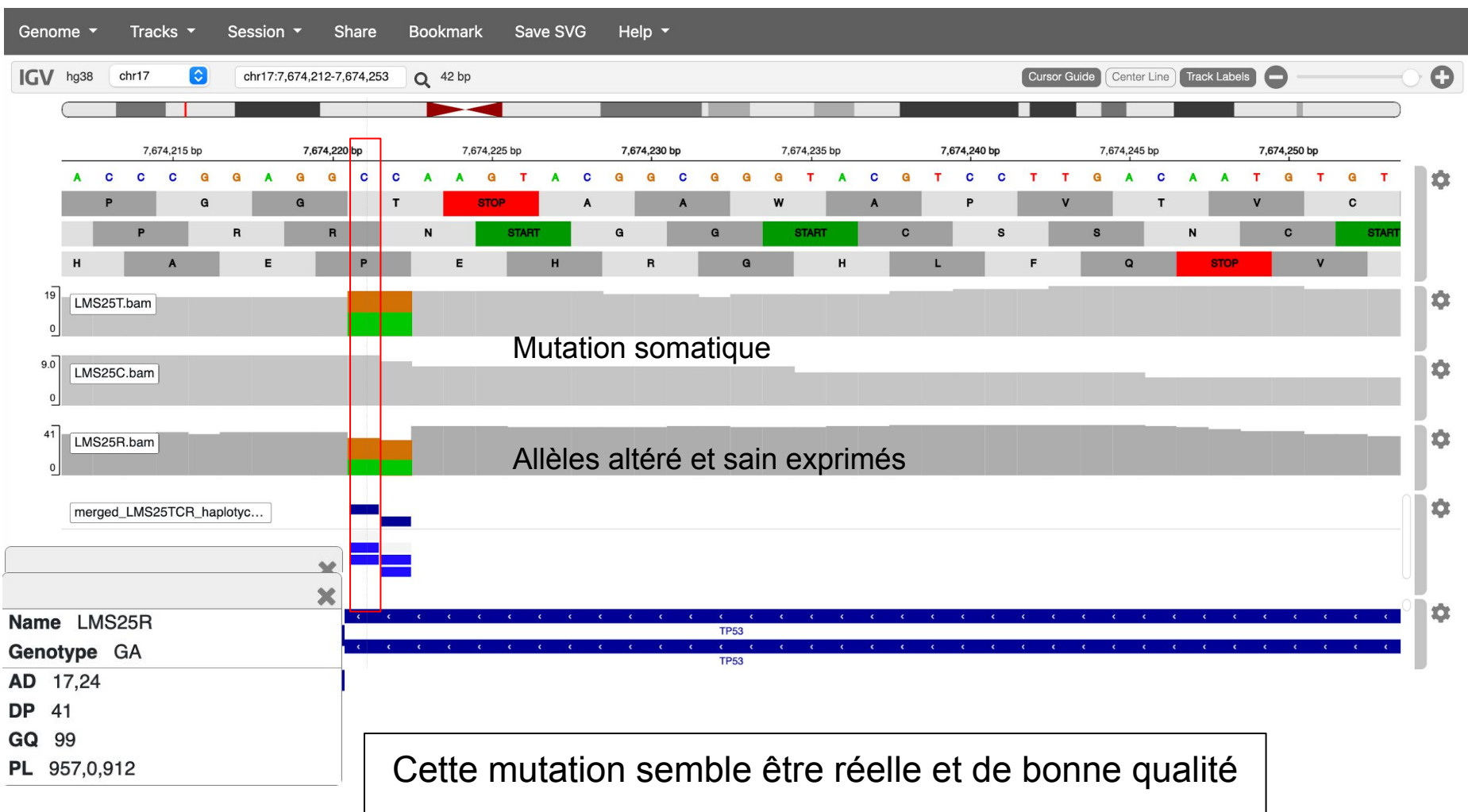
Integrative Genome Viewer

- Visualisation des reads et détection du variant à la position 7674221 sur le chromosome 17



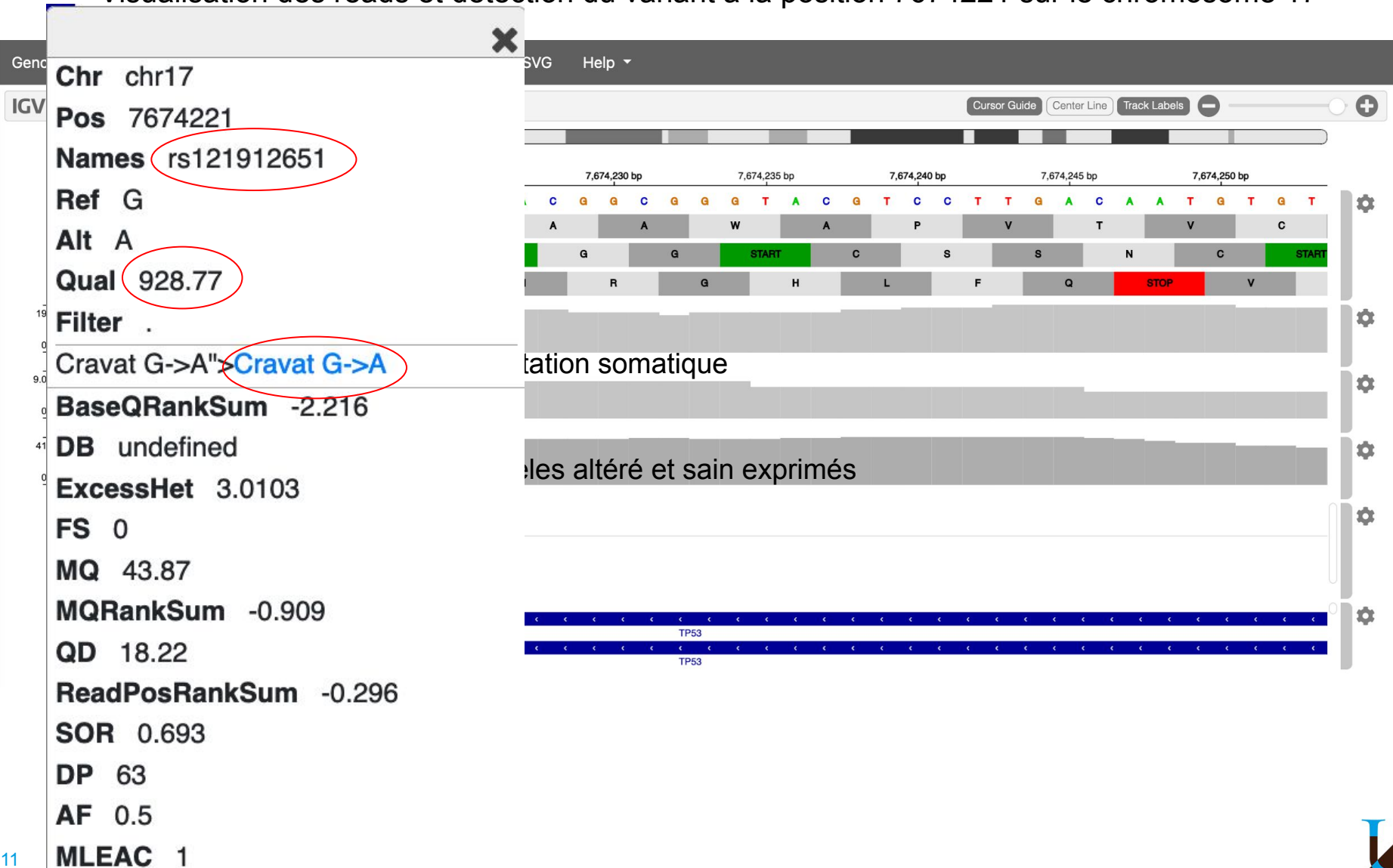
Integrative Genome Viewer

- Visualisation des reads et détection du variant à la position 7674221 sur le chromosome 17



Integrative Genome Viewer

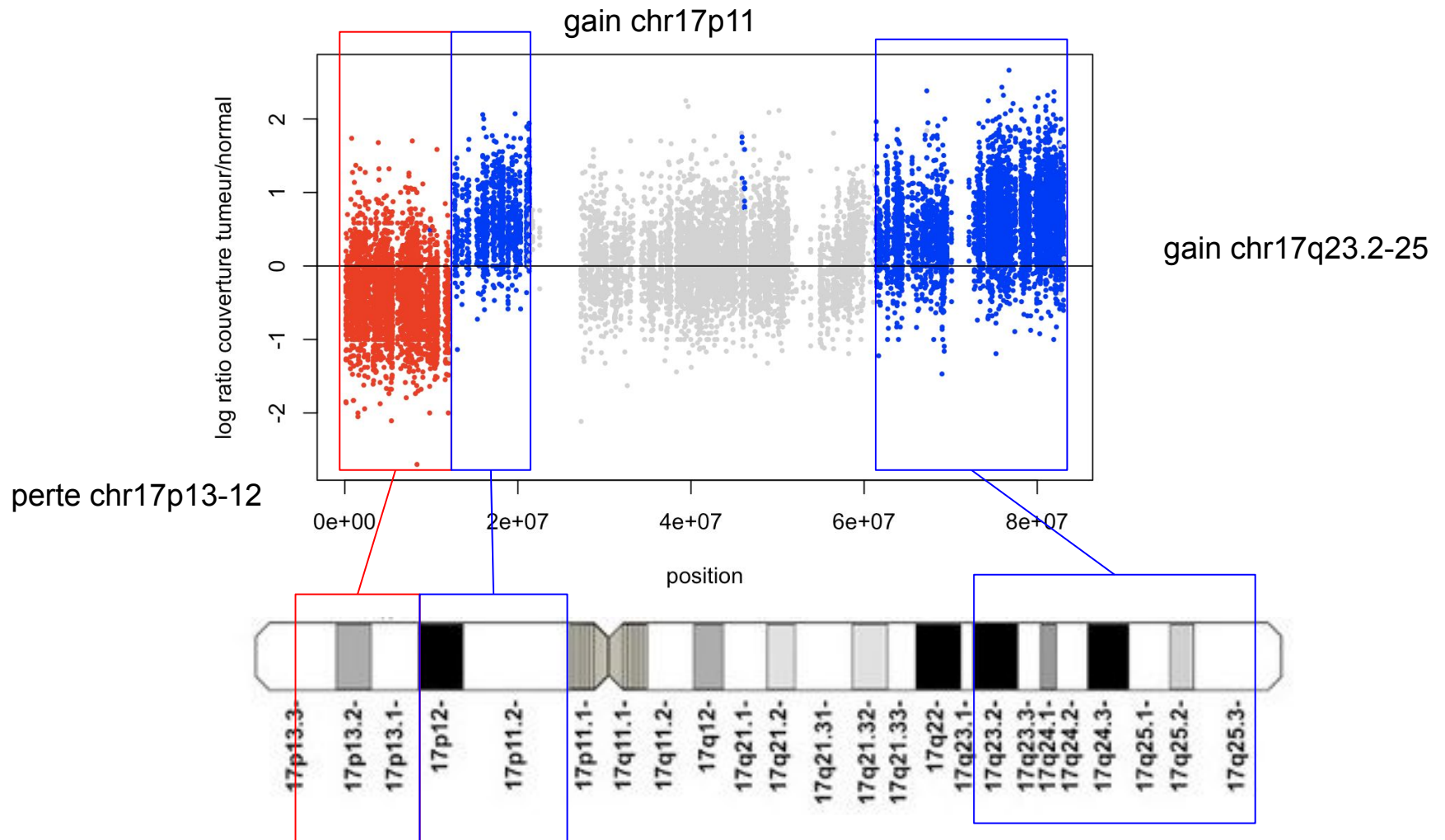
— Visualisation des reads et détection du variant à la position 7674221 sur le chromosome 17



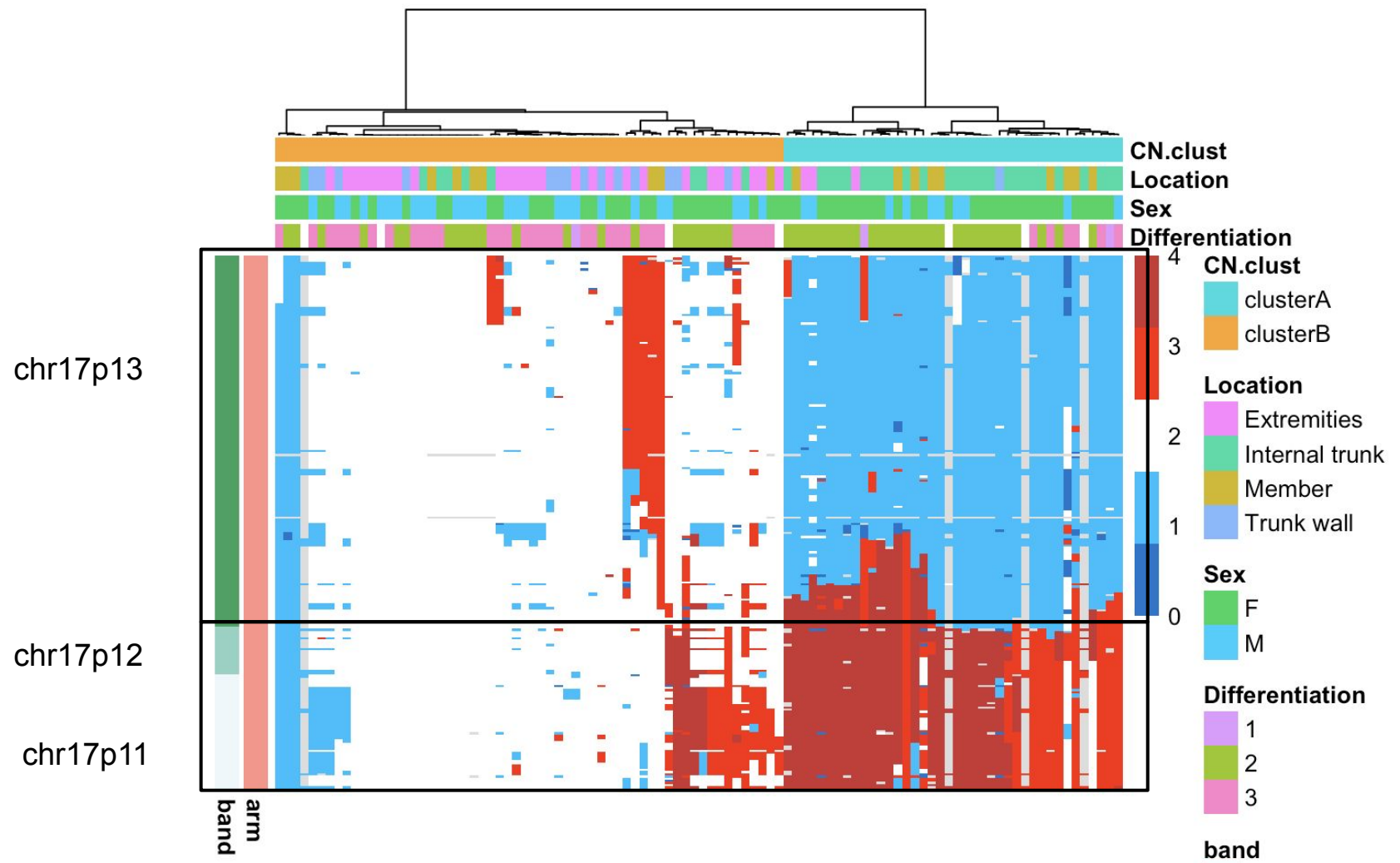
- ID: rs121912651, COSV52662035
- Changement protéique: p.Arg237Trp
- Effet de la mutation: perte du rôle supresseur de tumeur et gain de fonction qui peut promouvoir la tumorigénèse.
- Fréquence dans la population faible
- Zhang, Y., Coillie, S., Fang, JY. et al. Gain of function of mutant p53: R282W on the peak?. Oncogenesis 5, e196 (2016):

p53 est un facteur de transcription, notamment impliqué dans les mécanismes de mitose et de mort programmée. La mutation de son gène codant en position 7 674 221pb est **dominante autosomique**. Cette mutation non synonymes présente dans de nombreux cancers est ainsi qualifiée d'Hot spot. Elle amène un GOF (Gain of function) à la protéine p53 : cette mutation affecte la structure de l'hélice H2 de la protéine, résultant à une perte de liaison H et à la dissociation du motif boucle-feuillet-hélice. Cela amène à une **déstabilisation** thermodynamique de la protéine en plus de toucher le repliement de la protéine. p53 est ainsi plus dénaturée à 37°C, la température corporelle. Enfin la mutation touche la partie de **liaison de p53 avec les protéines Bcl**. Cette liaison est primordiale pour l'exercice de leur fonction puisqu'elle résulte à la perméabilisation de la membrane externe des mitochondries puis à l'apoptose. (4)

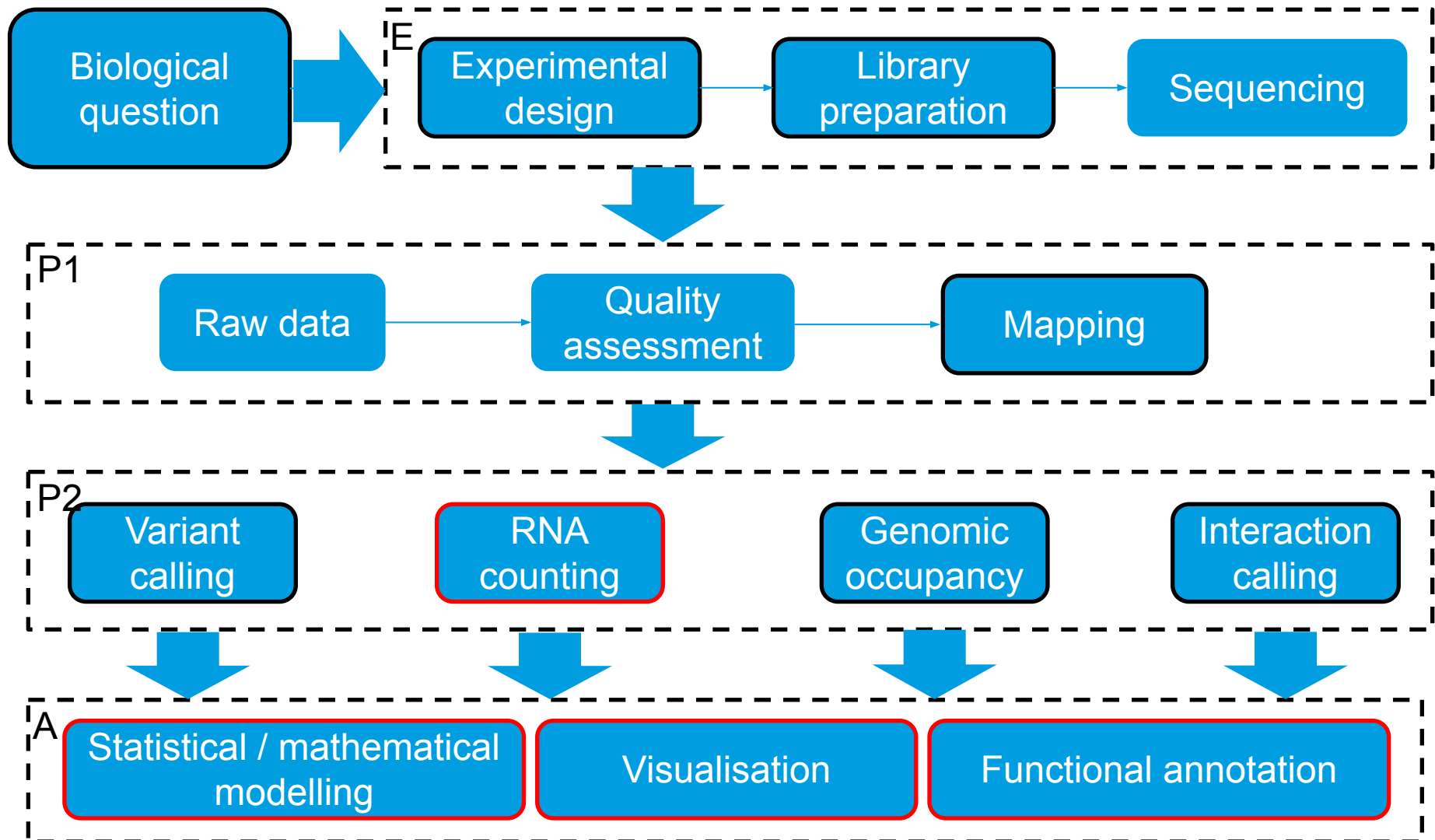
Copy Number Variation



Copy Number Variation

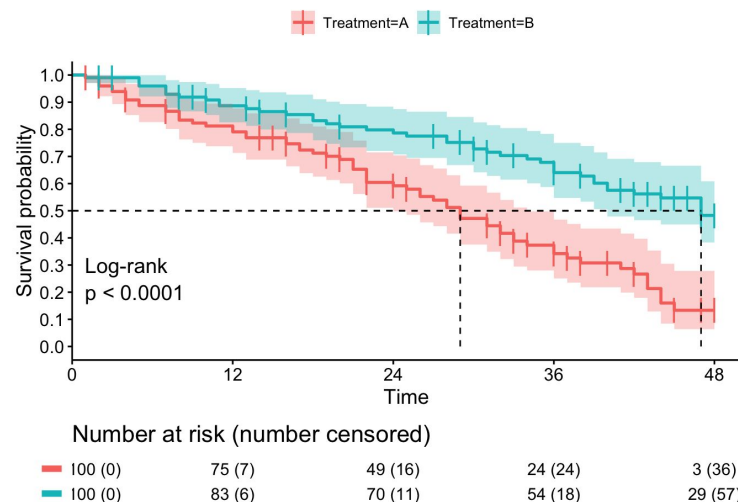


Pipeline bioinformatique



Analyse de survie: courbe de Kaplan Meier

L'estimateur de Kaplan-Meier est utilisé pour estimer la fonction de survie. La représentation visuelle de cette fonction est généralement appelée la courbe de Kaplan-Meier, et elle montre la probabilité d'un événement (par exemple, la survie) à un certain intervalle de temps. Si la taille de l'échantillon est suffisamment importante, la courbe doit se rapprocher de la véritable fonction de survie pour la population étudiée. Elle compare généralement deux groupes dans une étude (par exemple, un groupe qui a reçu le traitement A et un groupe qui a reçu le traitement B).



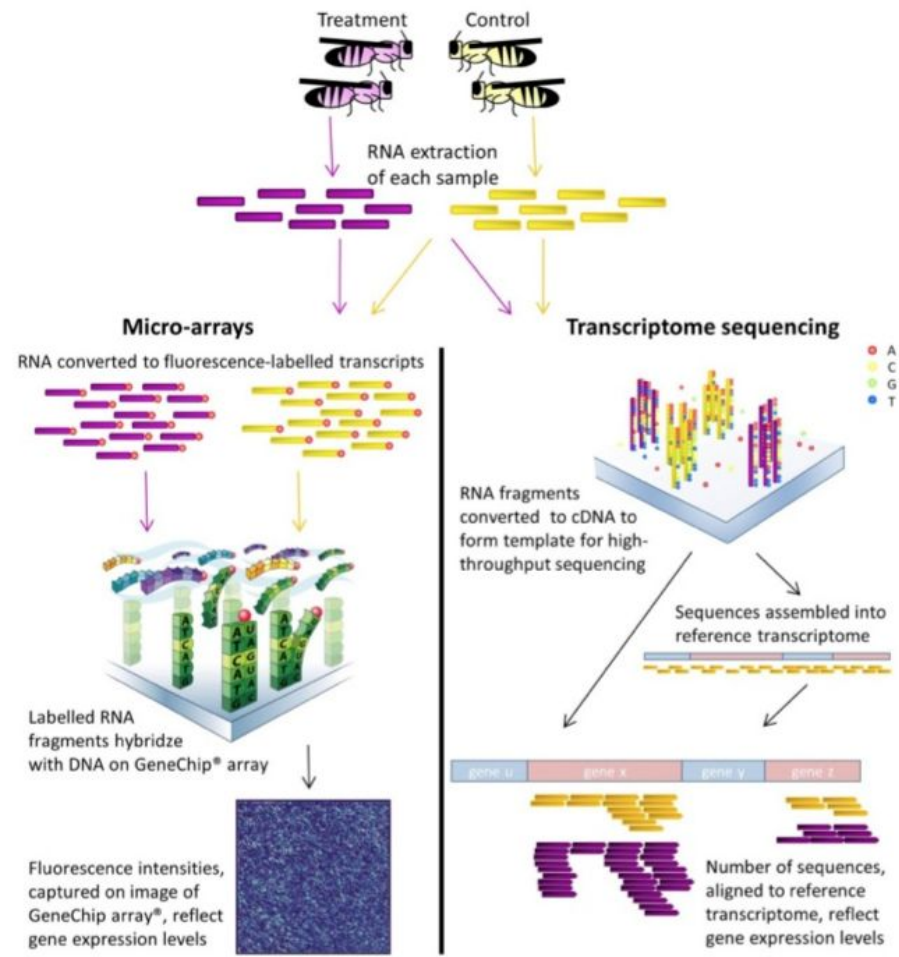
Le traitement B semble donner de meilleurs résultats que le traitement A (durée médiane de survie de +/- 47 mois contre 30 mois avec une valeur p significative).

<https://towardsdatascience.com/kaplan-meier-curves-c5768e349479>

A propos du RNA-seq

- Pourquoi utilise-t'on du RNA-seq?
- Normalisation

cDNA micro-arrays vs RNA-seq



From: Wertheim B. *in* Functional Genomics (Eds. G. Meroni and F. Petrera), 2012

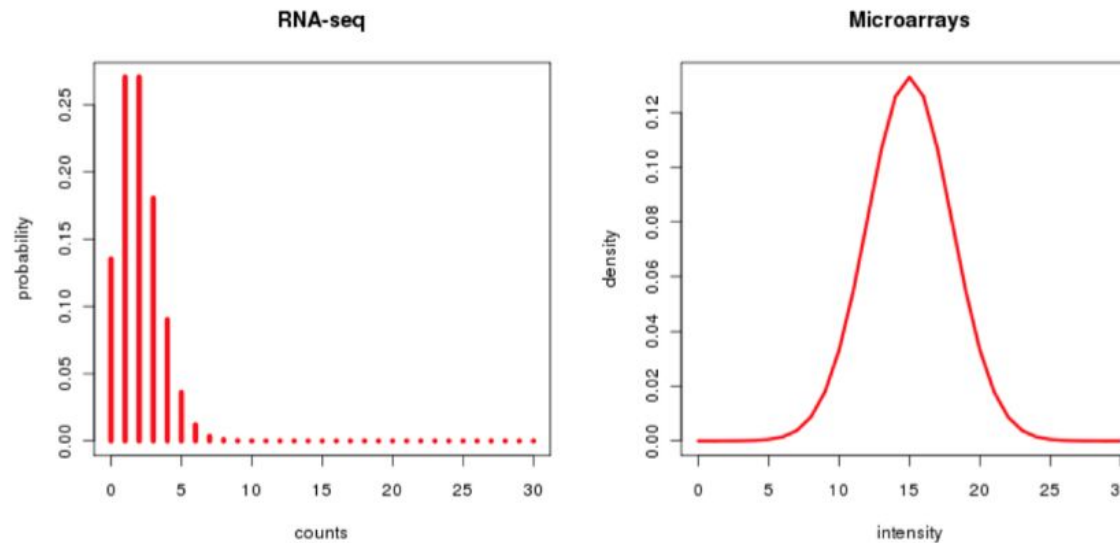
cDNA micro-arrays vs RNA-seq

- ❑ avantages du RNA-seq
 - ❑ Transcriptome de référence non nécessaire
 - ❑ Identification de nouveaux transcrits
 - ❑ Intervalles de mesure plus important (plus précis & plus sensible)
 - ❑ Reproductible
 - ❑ Peut-être adapté à de nouvelles questions

- ❑ avantage des micro-arrays
 - ❑ Prix
 - ❑ Plus petite quantité d'ARN
 - ❑ Plus de recul

cDNA micro-arrays vs mesures de RNA-seq

Ce sont des comptes discrets mappés sur des régions d'intérêt



Les méthodes développées pour les micro-arrays ne sont pas applicables au RNA-seq

Pourquoi utilise-t'on le RNA-seq?

- ❑ 2 objectifs principaux:

- ❑ Quantitatif:

- ❑ Expression différentielle

- ❑ Epissage alternatif

- ❑ TSS / polyA alternatif

- ❑ Qualitatif :

- ❑ structure des mRNAs : jonction exon/intron, identification des sites TSS et polyA, transcrits de fusion)

- ❑ Assemblage du transcriptome

- ❑ Analyse des SNPs

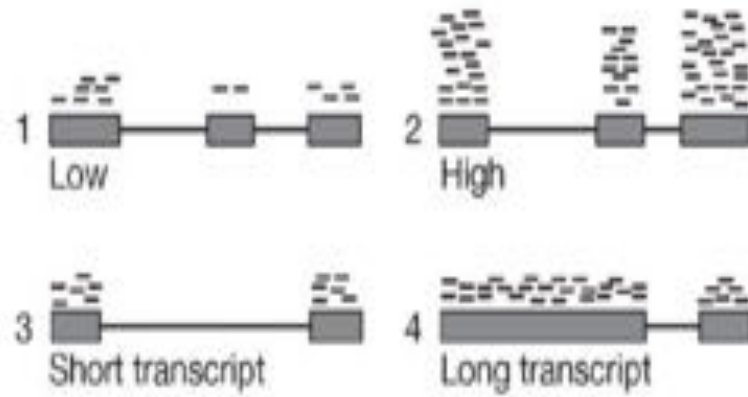
Pourquoi doit-on normaliser?

- ❑ Il y a des biais systématiques (techniques et biologiques) qui impactent les résultats
- ❑ Vrai pour toutes les analyses de données NGS mais certains biais sont spécifiques aux technologies.
- ❑ Essentiel pour les analyses quantitatives

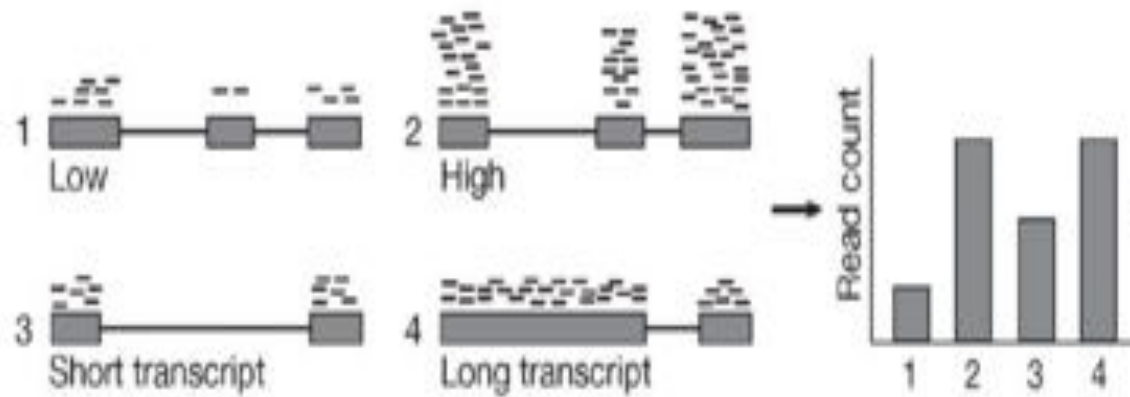
Quels biais?

- ❑ Composition des séquences (GC content)
- ❑ PCR amplification
- ❑ Effet de batch
- ❑ Profondeur (nombre total de reads séquencés et mappés)
- ❑ Longueur des gènes *
- ❑ La taille de la librairie

Longueur des gènes



Longueur des gènes



Le nombre attendu de reads dans une librairie est proportionnel au nombre total de transcrits fois la longueur des gènes.

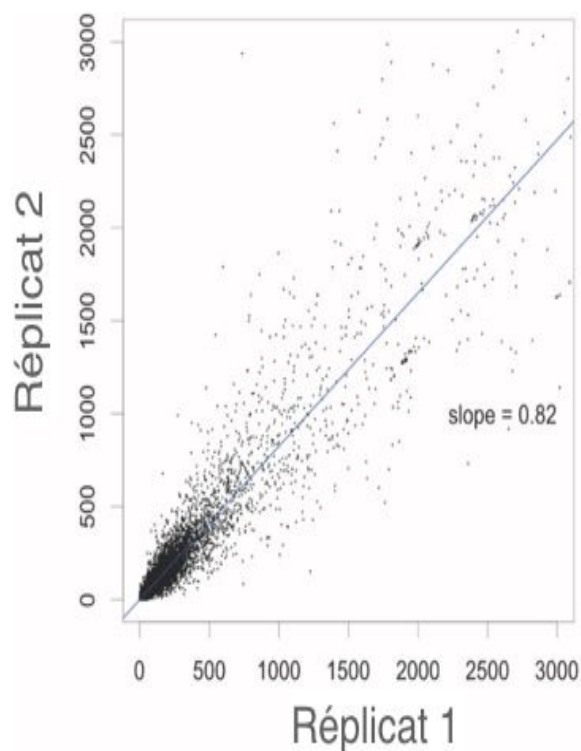
Taille de la librairie

	Sample 1	Sample 2
Gene 1	15	30
Gene 2	24	48
.	.	.
Gene 20,000	345	690

Taille de la librairie

	Sample 1	Sample 2
Gene 1	15	30
Gene 2	24	48
.	.	.
Gene 20,000	345	690
Total number of reads	10,000,000	20,000,000

Taille de la librairie (entre réplicats)

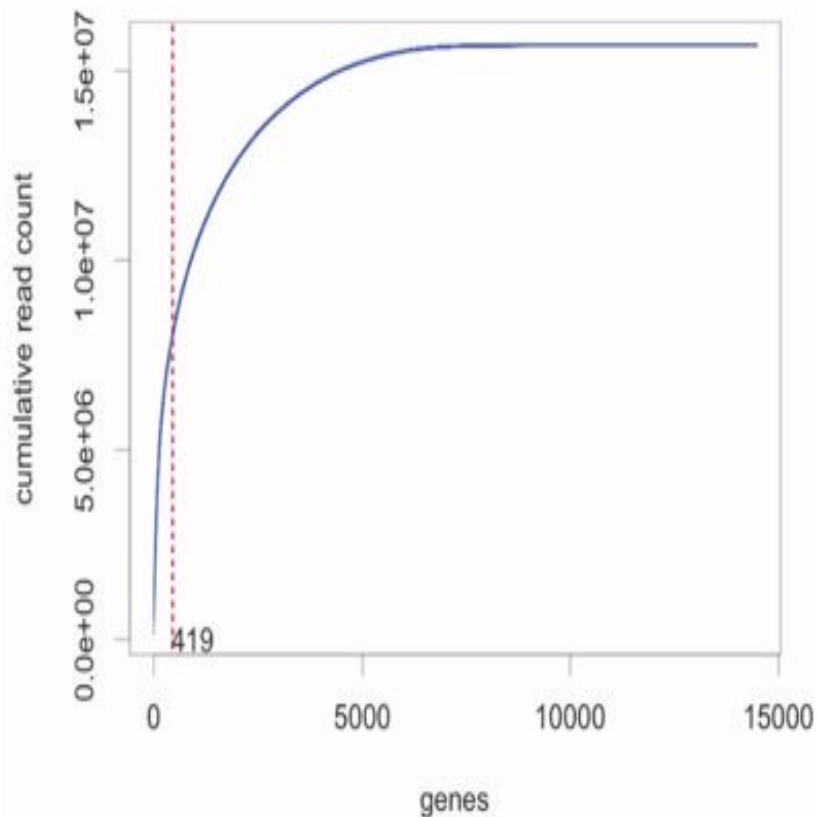


Idéalement tous les points devraient se trouver sur la ligne de pente 1

Le ratio entre les 2 réplicats devrait être 1
----> En fait c'est 0.82, nous devons normaliser par ce nombre.

Utiliser la pente réelle comme *facteur de normalisation ou size factor*

Taille de la librairie (entre conditions)



Dans différentes conditions biologiques, différents ARNs sont exprimés, ce qui mène à une quantité totale d'ARNs différente

La majorité des gènes n'est pas exprimée

Le but de la normalisation est de minimiser l'effet des séquences majoritaires

Concept de taille de librairie *effective*: edgeR

Trimmed Mean of M-values (Robinson et al. 2010) (edgeR)

Meilleure hypothèse: l'output entre échantillon est similaire pour un set de gènes G

Objectif: identifier un sous-jeu de gènes non exceptionnel

Etape 1: Calcule le log ratio et compte total entre deux échantillons

Etape 2: Exclut les gènes avec des log ratio extrêmes (i.e. changements massifs entre deux échantillons) ou avec des comptes extrêmes (ceux qui tombent dans le x% quantile pour une de ces mesures)

Etape 3: Calcule un facteur de normalisation basé sur les gènes restants

Enrichissement fonctionnel à partir de gènes différentiellement exprimés

A partir des données d'expression des gènes, nous pouvons détecter les gènes dont l'expression diffère entre 2 groupes.

Est-ce que ces gènes sont impliqués dans des voies biologiques particulières ?

2 solutions

- A partir de la liste des gènes ordonnée par la différence d'expression
- A partir de listes de gènes significativement différentiellement exprimés

A Broad institute tool (<http://software.broadinstitute.org/gsea/index.jsp>)



Gene Set Enrichment Analysis

[GSEA Home](#) [Downloads](#) [Molecular Signatures Database](#) [Documentation](#) [Contact](#)

Overview

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).

From this web site, you can:

- Download the GSEA software and additional resources to analyze, annotate and interpret enrichment results.
- Explore the **Molecular Signatures Database (MSigDB)**, a collection of annotated gene sets for use with GSEA software.
- View **documentation** describing GSEA and MSigDB.

What's New

19-Oct-2017: MSigDB 6.1 released. See [release notes](#) for more information, including important corrections to gene sets in the C3 collection.

11-Aug-2017: Four new CHIP files are now available for use with data specified with Ensembl IDs, which are commonly used for gene expression derived from RNA-Seq data. More details are [here](#).

01-Jul-2017: The production version of GSEA Desktop v3.0 is now available! It's open-source on [GitHub](#), features SVG plots, Cytoscape 3.3+ support for Enrichment Maps, heatmap dataset export, and more.

06-Apr-2017: Version 6.0 of the Molecular Signatures Database (MSigDB) is now available under a Creative Commons license, with additional terms for some sub-collections of gene sets. The release also includes updates to the C3 motif gene sets, and some other minor additions and corrections. See the [Release Notes](#) for details.

06-Oct-2016: Version 5.2 of the Molecular Signatures Database (MSigDB) is now available. It contains the overhauled C5 collection of 6,166 sets of recent gene ontology annotations, as well as a number of additions, updates and corrections. See the [Release Notes](#) for details.



The diagram illustrates the GSEA workflow. On the left, 'Molecular Profile Data' (represented by a heatmap) and 'Gene Set Database' (represented by a database icon) are inputs. These feed into a central box labeled 'Run GSEA'. The output of this process is 'Enriched Sets', shown as a GSEA plot with a green line and a red bar chart below it.

License Terms

GSEA and MSigDB are available for use under these [license terms](#).

Please [register](#) to download the GSEA software, access our web tools, and view the MSigDB gene sets. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.

Contributors

GSEA and MSigDB are maintained by the GSEA team. Our thanks to our many contributors. Funded by: National Cancer Institute, National Institutes of Health, National Institute of General Medical Sciences.

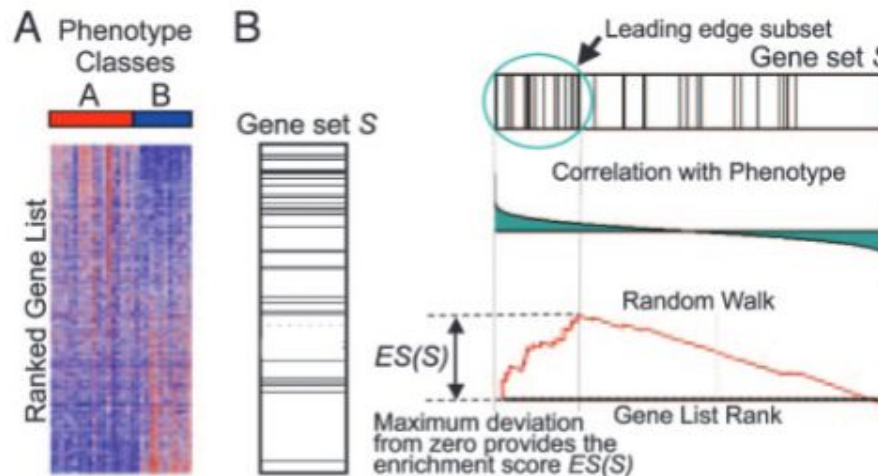


Citing GSEA

To cite your use of the GSEA software, please reference Subramanian, et al. [Bioinformatics](#) 2005;21:971-973.

How does work GSEA?

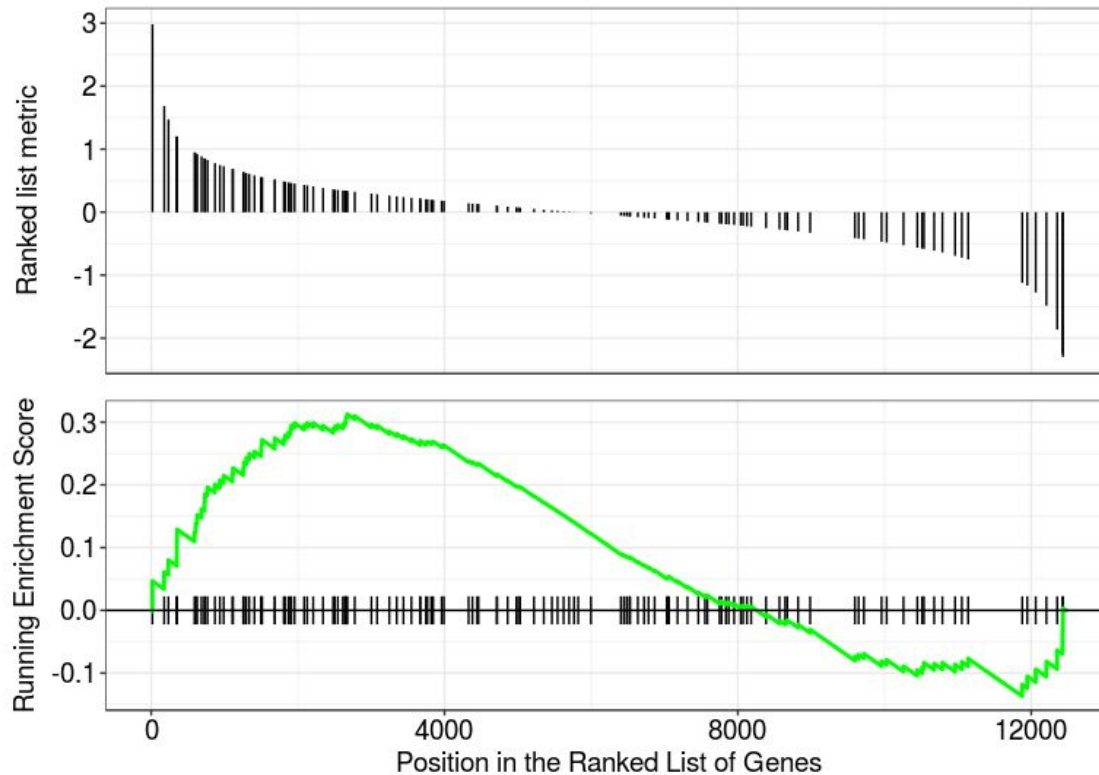
(http://software.broadinstitute.org/gsea/doc/subramanian_tamayo_gsea_pnas.pdf)



- ❑ Est ce que la distribution du groupe de gènes suis une distribution continue de probabilités?
 - pondérée (taille du groupe de gènes) statistique Kolmogorov-Smirnov-like
- ❑ Est ce significatif?
 - permutation des labels et p-valeur nominale + correction multi-testing

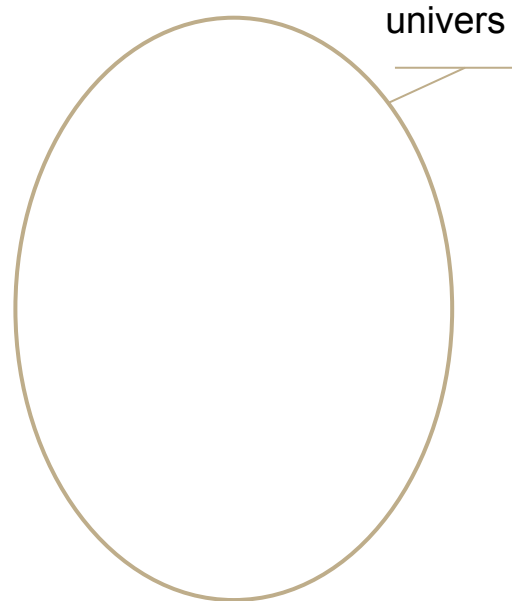
How does work GSEA?

(http://software.broadinstitute.org/gsea/doc/subramanian_tamayo_gsea_pnas.pdf)



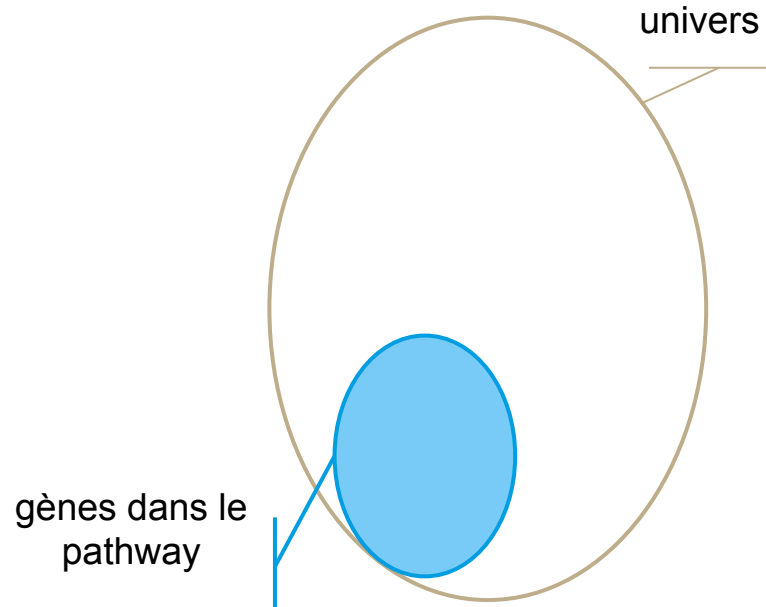
Enrichissement fonctionnel dans une liste de gènes

- ❑ Est ce que la proportion de gènes observée d'un pathway est due au hasard?
test hypergéométrique



Enrichissement fonctionnel dans une liste de gènes

- ❑ Est ce que la proportion de gènes observée d'un pathway est due au hasard?
test hypergéométrique



Enrichissement fonctionnel dans une liste de gènes

- ❑ Est ce que la proportion de gènes observée d'un pathway est due au hasard?
test hypergéométrique

