Tuto MARIO GALAXY

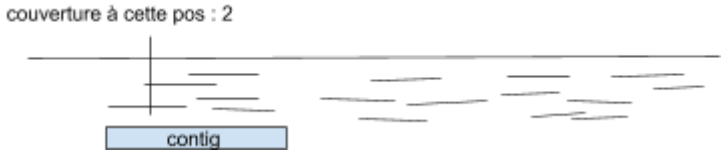
Pourquoi GALAXY ?
https://usegalaxy.org/

user-friendly,

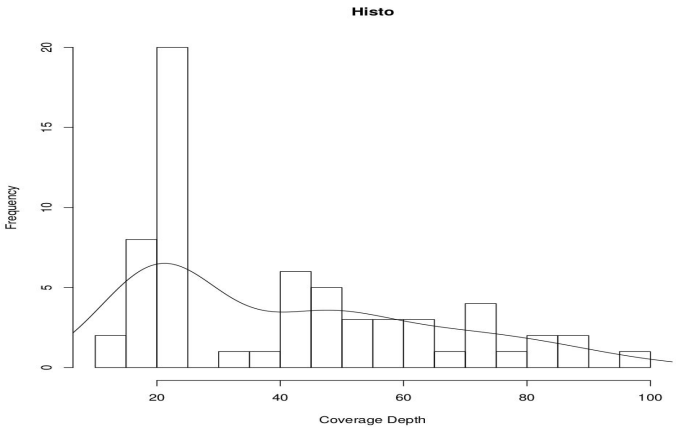
différentes instances galaxy, de base, celle par défaut des USA.
Données peuvent être donnés à différents groupes de scientifique tenant des instances galaxy (inra, cnrs, …), chaque instance fournissent des outils différents.
Avoir une trace des outils utilisés et de leur version => outils dynamiques changeant souvent.
Outils sur colonne de gauche

Télécharger des données : Get Data -> Upload File
REnommer -> icône stylo sur les données

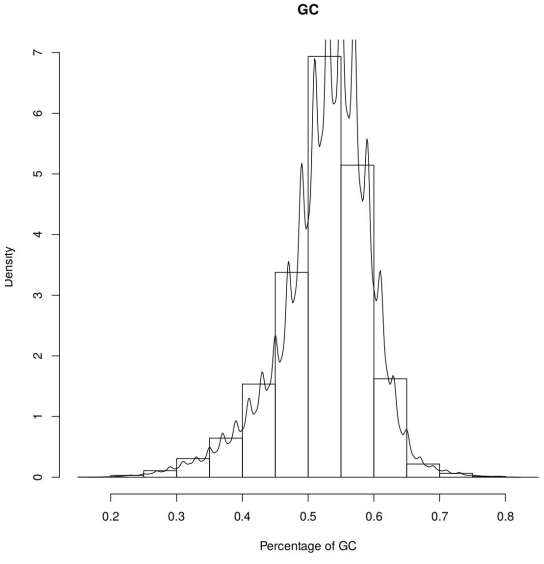
Histogramme
Afficher données -> oeil (affiche une partie des fichiers seulement)
Fichier Contigs-stats => 63 contigs, définis par leur longueur et couverture
Couverture =



couverture/profondeur de contig = 12X veut dire que ces positions en moyenne ont été lues 12 fois
Graph/Display -> Histogram



On remarque que de nombreux contigs ont une couverture de 20%.
Résultat de séquençage + assemblage.
Pourquoi grande disparité ? On devrait s'attendre à ce que coverage soit à peu près égale partout. Donc variation d'efficacité.
Alors que si ech = métagénomique, variation peut être due aux différentes espèces présentes et leur quantité dans l'échantillon.



Fichier fastq contient différents reads, avec identifiant, séquence, qualité.

Calcul GC

EMBOSS -> geecee
Donne fichier avec 1 seule colonne => cut
Histogram

Beaucoup de reads ont une quantité de GC d'environ 50%

Suit une loi normale => donc échantillon homogène => donc échantillon plus ou moins pur (car qu'un seul pic)
Rq: plus on a un GC élevé plus on résiste aux fortes températures

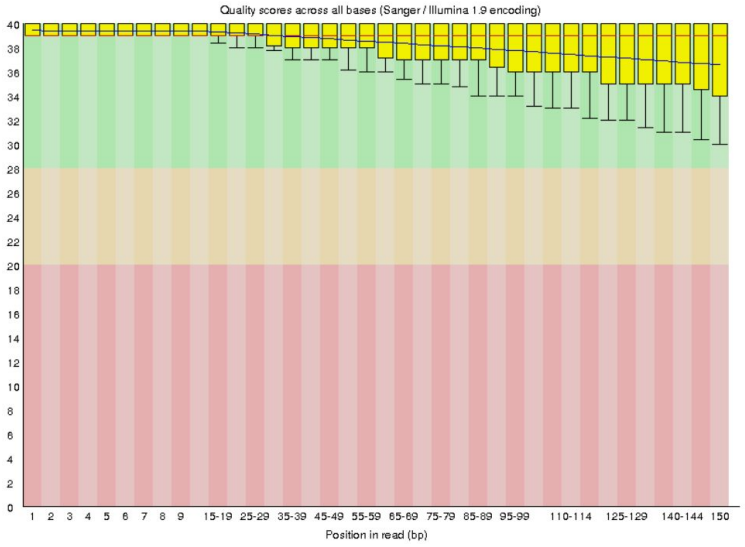
Qualité de Séquençage

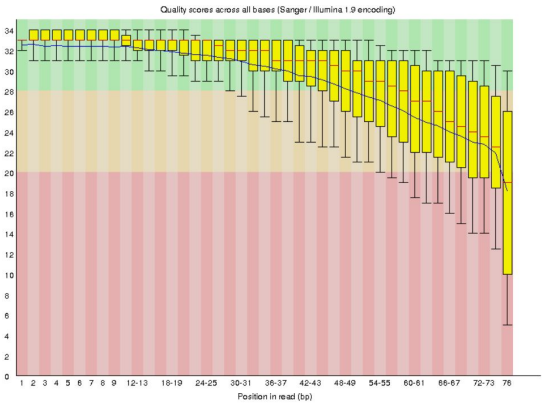
Qualité de séquence = proba que les nucléotides sont bien vérifiés.
Cette probabilité est séquenceur-dépendant. (capacité du séquenceur à donner et lire la bonne couleur (étape de base-calling).

NGS -> Fastq groomer : permet de vérifier utilisabilité du fichier
NGS -> Fastqc

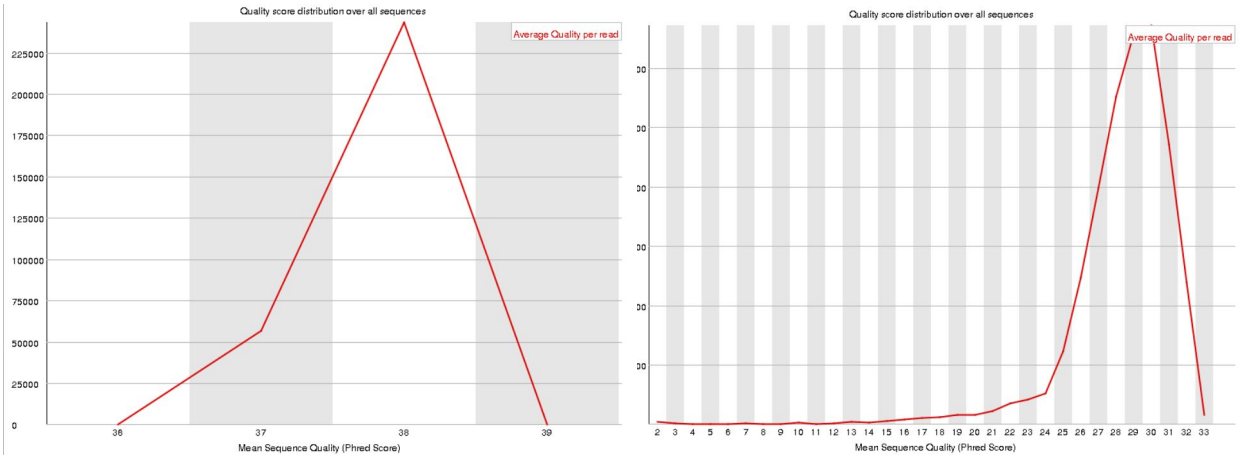
Filename	FASTQ_Groomer_on_data_2
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	300000
Sequences flagged as poor quality	0
Sequence length	150
%GC	52

Per base sequence quality

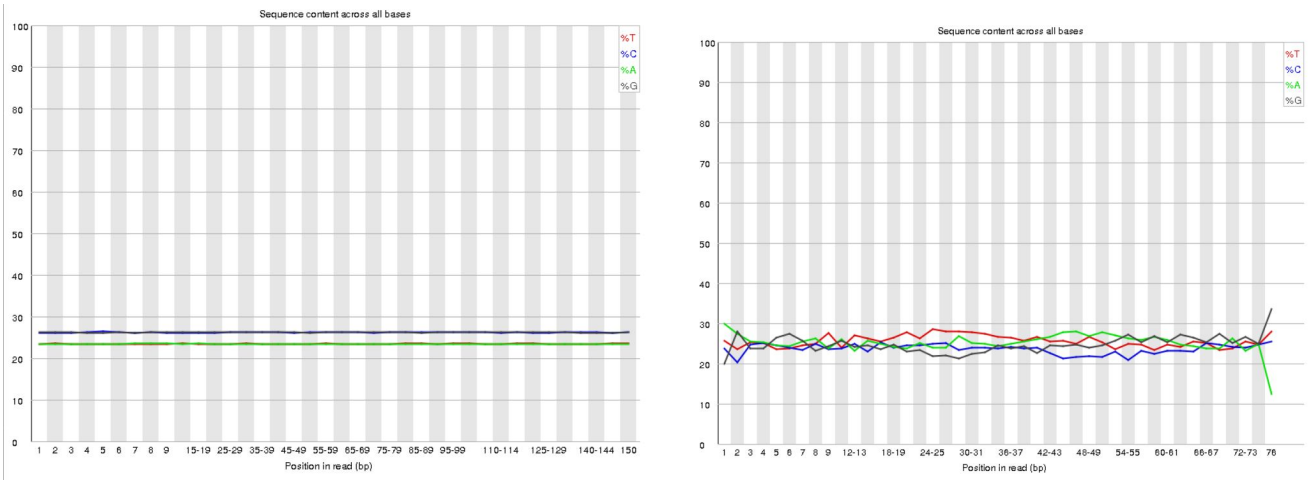




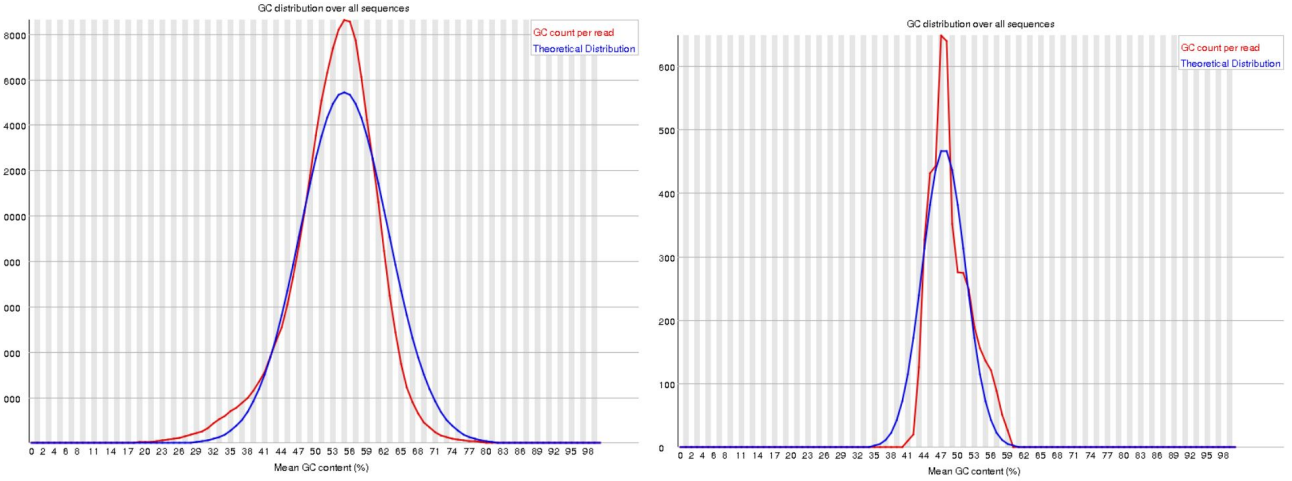
Sur le per-base sequence quality, on regarde sur la position1 jusqu'à la 150eme (longueur 150) la qualité des reads pour chaque position.
Moins bonne qualité vers la fin de séquence (plus d'erreurs possibles) car avec le temps, enzymes peuvent incorporer des mauvais nucléotides.
Donc , sur la fin de la séquence, le cluster (pcr) va présenter de nombreuses différences de couleur. Donc le signal sera de moins en moins pur.
Ici, très bonne qualité d'environ Q36 en moyenne vers la fin de la séquence.
2e graphe montre l'inverse.(dégringole vite, en dessous de q20, écart type grands)



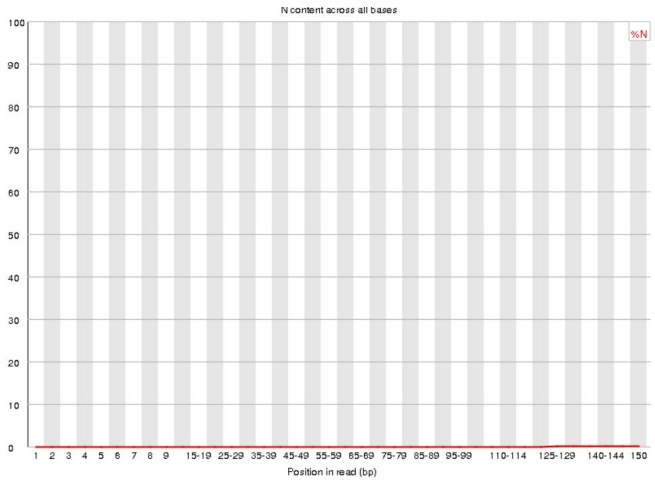
Sur ce graphe, on regarde la distribution de score sur la qualité moyenne de chaque read (indépendamment de la position)
On obtient alors une courbe de faible largeur (36-39), l'idéal serait d'obtenir un pic.
Le petit écart entre 36 et 37 est dû à la machine/ à l'échantillon (préparation du flowcell (oligo sur puits) par exemple).
2e graphe: large distribution, mais pic à 30 donc ok.



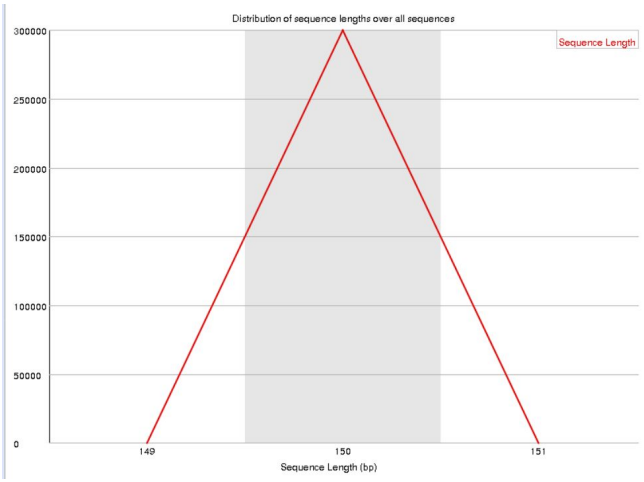
Ici, même pourcentage pour GC que AT. Ici, le graphe semble donc bon. (graphe 2 dégueu, car 1000 fois moins de données donc moins écrasé, on voit plus la fluctuation mais sinon plutôt ok, malgré grand biais sur la fin de seq 76).



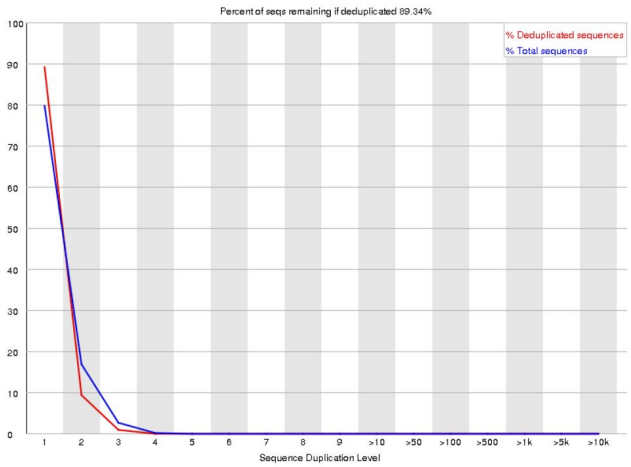
Ici, pourcentage en GC ne suit pas parfaitement courbe théorique. Graphe 2: comme échantillon moins grand, courbe moins lisse, plus haute, mais sinon moyenne reste la même que théorique. petit plateau sur la droite: dû à une contamination ?)

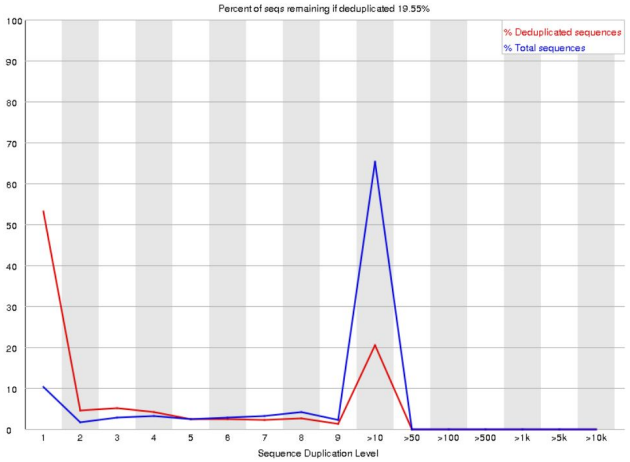


Quantité de N (non connu) au lieu de ATCG dans les reads. Ici 0 donc très bien.

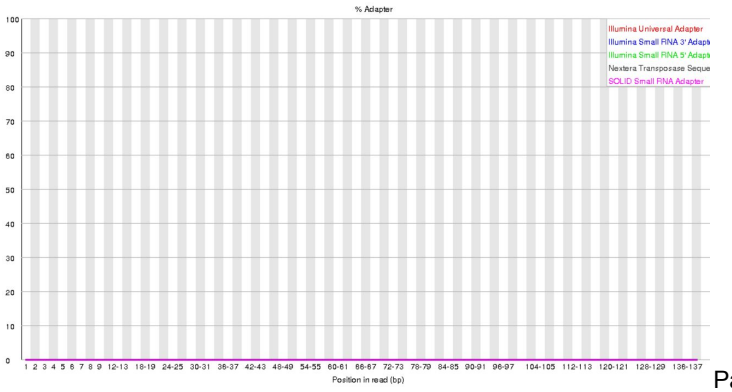


Ok, juste une longueur.

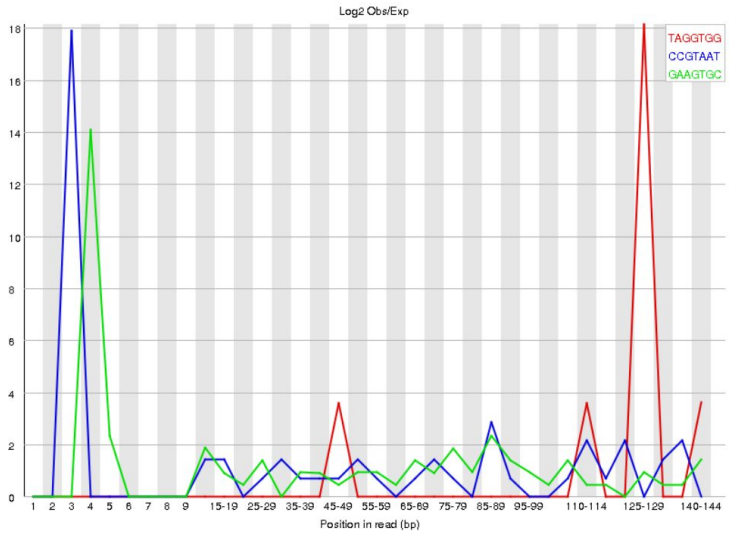




Nombre de lectures exactement identiques. Surreprésentation de certaines séquences si PCR surexprime/copie trop une même séquence.
Si biais important, PCR ne s'est pas bien passée (surreprésentation, non aléatoire, donc non représentatif). C'est le cas sur le graphe 2 mais biais augmenté par le fait de la taille de l'échantillon (le peu de surrpz sur le peu de nombre total fait vite grimper les courbes).



Pas d'adaptateur séquencé, OK.



KMER content

kmer = sequence de longueur k
Ici certaines séquences ont été surreprésentées.

Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position	
TAGGTGG	40	0.007897543	18.133062	125-129	
CCGTAAT	200	0.0084517	17.915308	3	***
GAAGTGC	305	0.005655977	14.097292	4	***

Amélioration du dataset foiré
D'après l'analyse du 2e jeu de données, la fin des reads biaise les analyses, on va donc filtrer les reads selon leur qualité.
2 moyens :

NGS -> Filter by quality
Enlève tous les reads dont la qualité est inf à 20. Ne change pas grand chose et réduit fortement le nombre de données (1000 seq en moins)

NGS -> FastQC trimmer by sliding window
pour chaque read, applique un masque qui enlève les parties des reads dont la qualité est inférieure à 20.
Change les longueurs des séquences, mais meilleurs résultats et moins de perte de données.

Traitement secondaire des données:
Mapping reads on a reference genome

Récupération d'un dataset, évaluation qualité et transformation si nécessaire puis mapping des reads avec logiciel BWA.

Dataset = exome du chromosome 22 d'un humain
=> donc seq ciblé, donc seq Amplicon

Qualité très bonne pour chaque graphe, malgré de petites digressions. (gros écart-types vers pos 66-76 sur per seq) (biais aux premières pos sur per base sequence content : primers mal enlevés ?)

NGS -> Map with BWA for illumina
SAM file:
@SQ = commentaires,
beaucoup de commentaires en début de fichier

SAM = lisible 

BAM = binaire
BAI = index 

@SQ	SN:chrY	LN:59373566						
@PG	ID:bwa	PN:bwa	VN:0.5.9-r16					
61CC3AAXX100125:7:118:2538:5577	16	chr22	16050954	23	76M	*	0	0	GGGGA
61CC3AAXX100125:7:1:17320:13701	16	chr14	19790660	0	76M	*	0	0	CTCATA
61CC3AAXX100125:7:93:5100:14497	16	chr22	16052858	0	76M	*	0	0	TCTTCT
61CC3AAXX100125:6:92:7549:15004	16	chr22	16052936	0	76M	*	0	0	CATTCA
61CC3AAXX100125:5:7:1488:7780	16	chr22	16053177	23	76M	*	0	0	AGTGAT
61CC3AAXX100125:7:72:14903:20386	16	chr22	16053702	23	76M	*	0	0	ATGAA(
61CC3AAXX100125:7:88:9942:19183	0	chr14	19788896	0	76M	*	0	0	CCTGT(
61CC3AAXX100125:7:76:1585:2024	4	*	0	0	*	*	0	0	AAATTA
61CC3AAXX100125:6:26:17654:5573	0	chr22	16053945	23	76M	*	0	0	AGGGC
61CC3AAXX100125:7:117:7805:10957	16	chr22	16054452	0	76M	*	0	0	GCCGG
61CC3AAXX100125:7:36:11248:16392	4	*	0	0	*	*	0	0	GGTCT(
61CC3AAXX100125:6:80:10088:8830	16	chr22	16054924	0	76M	*	0	0	CCCTG(
61CC3AAXX100125:6:115:5701:20053	0	chr22	16055354	23	76M	*	0	0	CTCTAT

Analyse du premier read:
-> 16 : read reverse strand ,
chr22, position 16050954,
MAPQ (Phred quality) = 23,
CIGAR : 76M (76 matchs = le read a été aligné sur 76 positions continues (pas forcément les 76 identique à la ref))
TAGS: XT:A:U NM:i:0 X0:i:1 X1:i:1 XM:i:0 XO:i:0 XG:i:0 MD:Z:76 XA:Z:chr14,+19791972,76M,1;

NM:I donne le nombre de changements nécessaires pour transformer le read en la seq de ref:
ici NM:I:0 => 0 modifs donc pas de mismatchs.

Analyse de l'alignement:
NGS -> SAMtools -> Flagstat

% de reads ayant passés les filtres check de qualité : 100%
% de reads ayant été mappés: 97.45% (+ pour ceux qui ont trim les seq ayant un pb sur 3')

NGS -> SAMtools -> IdxStats

Column Description
------ ----------
1 Reference sequence identifier
2 Reference sequence length
3 Number of mapped reads
4 Number of placed but unmapped reads
(typically unmapped partners of mapped reads)

Example output from a *de novo* assembly:
contig_1 170035 98397 0
contig_2 403835 199564 0
contig_3 553102 288189 0

La plupart des reads sont associés au chromosome 22: number of mapped reads => 1074953 pour le chr22

Grep pour récupération des données:
Join, Substract and Group -> Group (right setting parameters)

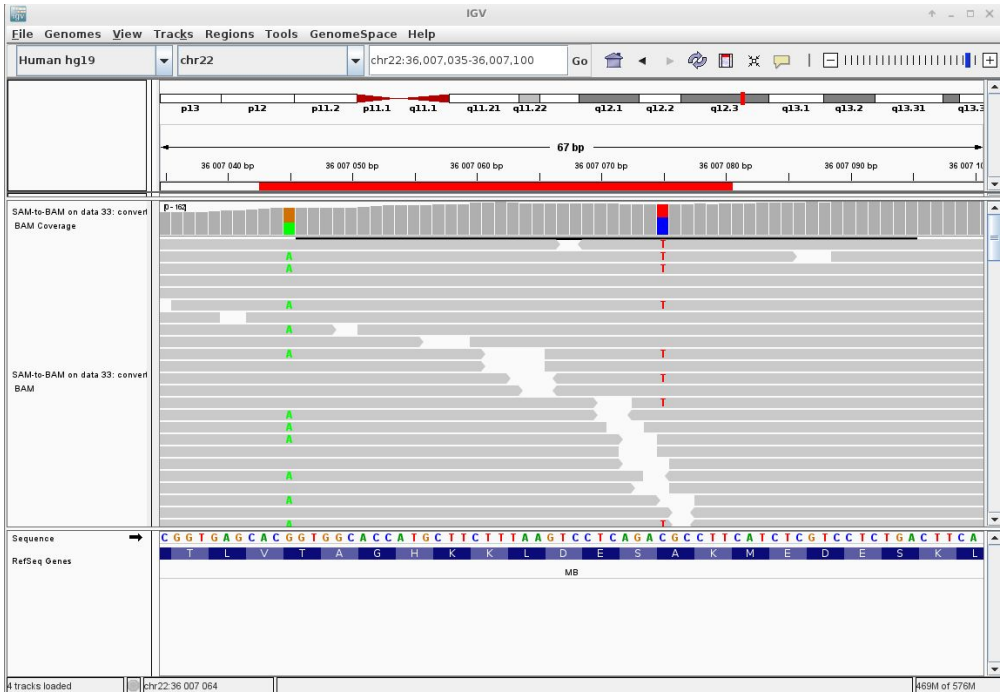
Unmapped reads : 29334 (* a la fin de IdxStat),
Mapped on forward strand :542628
Mapped on reverse strand: 579740

Number of reads having all positions mapped: 1118960 (Filter + Group)

Number of reads having the same seq than ref: 953581 (Filter-> Select)

Detecting mutations

IGV : see reads alignment on the referenced genome



nt colorés = différence entre nt de la ref et nt de l'ensemble des reads

NGS: SAM Tools > Sort BAM dataset
Construction d'un consensus à partir d'un dataset avec Mpileup
Puis utilisation de VARSCAN comparant consensus avec genome de référence.
Fichier VCF
Nombre de SNP : 2507
Nombre de Test échoués :0

HET : hétérozygote, HOM = homozygote (seuil de HOM si on veut travailler sur mutations dominante ou)
HET => 2 copies des chromosomes.

TD Mycobacterium tuberculosis

Récupération des données de séquençage Pair-ended
Groomer sur les fastq

FastQC sur les groomer

R1paired => per base seq qual = très bon (>24); qual score = 36 en moy; per base seq content = très bon mis à part toutes premières et dernières séquences; per seq gc content = très très bon, suit quasiment courbe théorique; per base N content = 0; sequence length dist = très bonne, un warning car les seq varient entre 50 et 100 ce qui est normal (attendu); seq duplication très bon (chute vite); pas de surreprésentation, d'adaptateur ou de kmers.

R2paired =>per base seq qual = très bon(>24); qual score = 34 en moy; per base seq content = très bon idem que R1; per seq content gc = très bon suit quasi courbe théorique; pas de N, length distrib entre 50 et 100 mais normal; seq duplication lvl = très bon (chute vite); pas de surreprésentation, pas d'adaptateurs, quelques kmers.

BWA sur les 2 jeux de séquences

Analyses stats:
Flagstat -> 0 QC failed reads, 99.08% mapped, 43.48% properly paired (paired mais pas dans le bon sens forcément)
Cela semble très bon au niveau des pourcentages.

IdxStats -> 3182+ 1270 unmapped reads, donc tous ne sont pas mappés, mais en revanche, une très grande partie est mappée (4 millions).

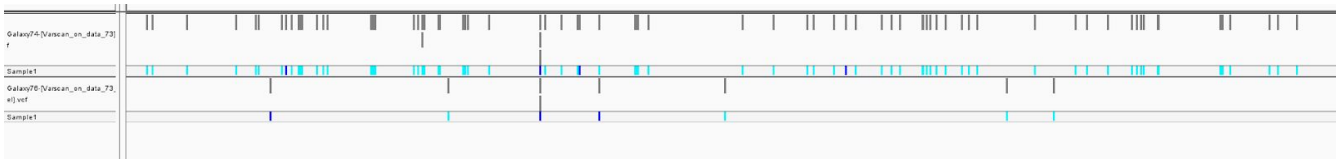
Stats -> reads mapped:481570, reads mapped and paired:478388

Conversion en BAM : SamTools -> SAM-to-BAM
Sort: Samtools -> Sort, chromosomal coordinates
-> MPileup pour construire séquence consensus :

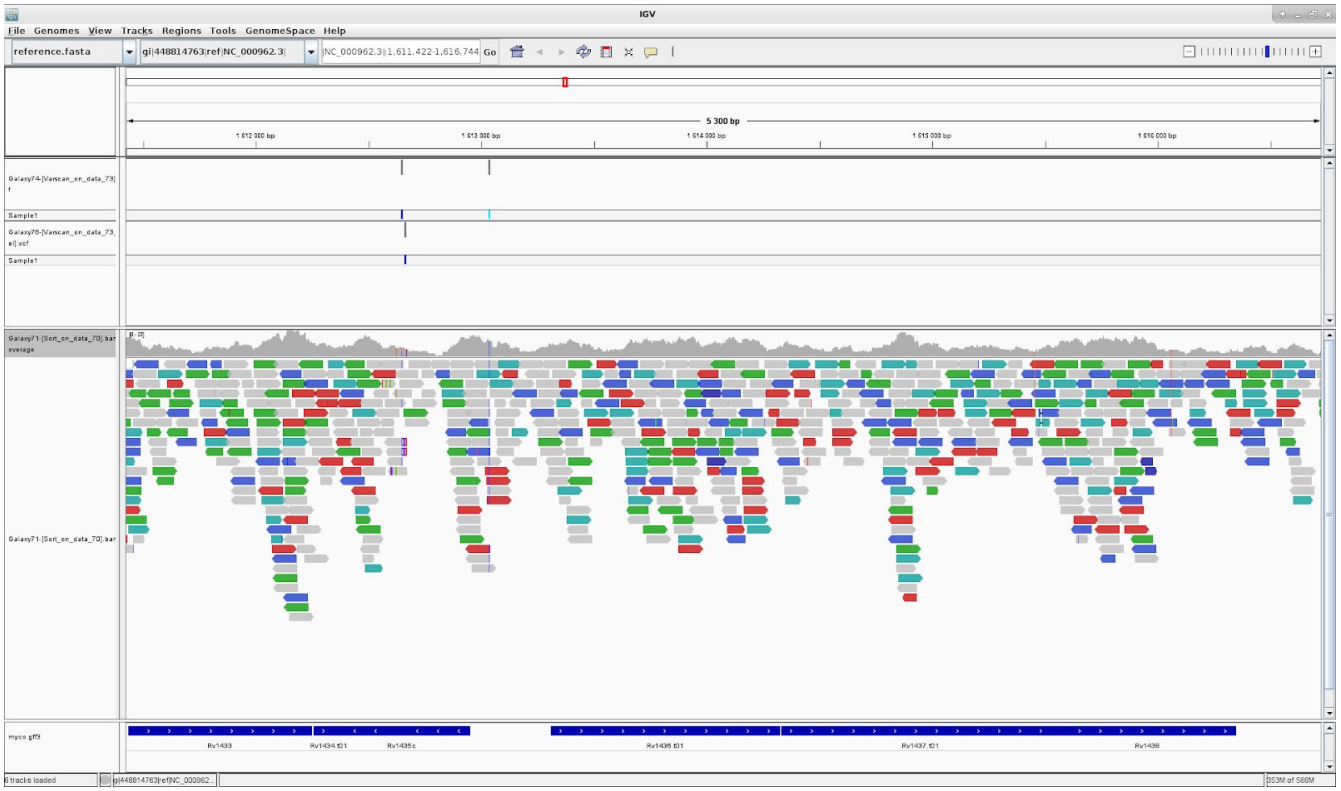
Analyse Tertiaire

-> Varscan

Visualisation dans IGV (avec génome de référence):
IGV: go to http://software.broadinstitute.org/software/igv/download
Then, click on one of the two following button: launch with 750 MB
Once the applet is launched, you can add the result files given by Galaxy (to save time, they
are already available in Transfert/SHDA/Myco/Myco_for_igv) using the menu.
1- Genomes menu/ Load Genome from file : reference.fasta
2- Files menu/ Load from file : file.bam, file.vcf and file.gff.



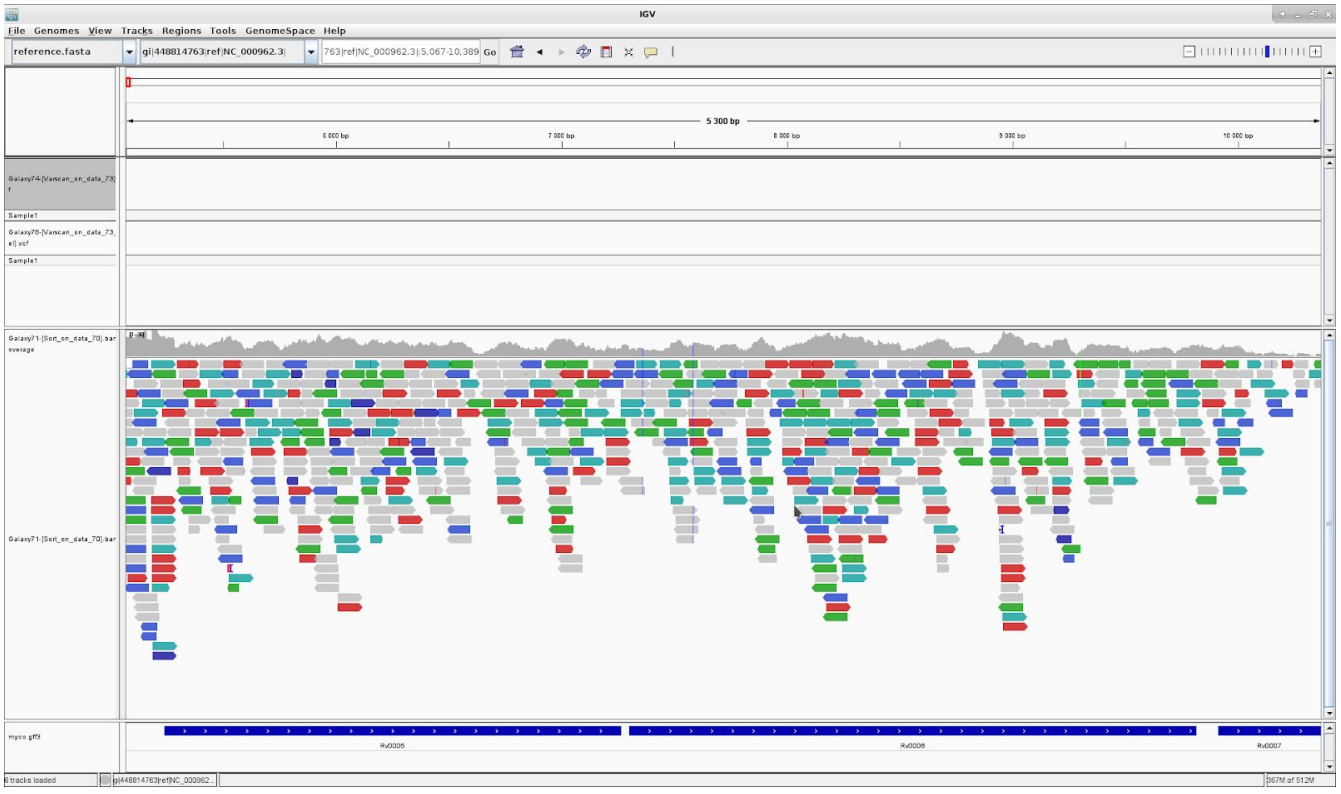
VARSCAN SNP (bien dézoomer) & Insert/Deletion, paramètres par défaut.
149 SNP, 14 Indel



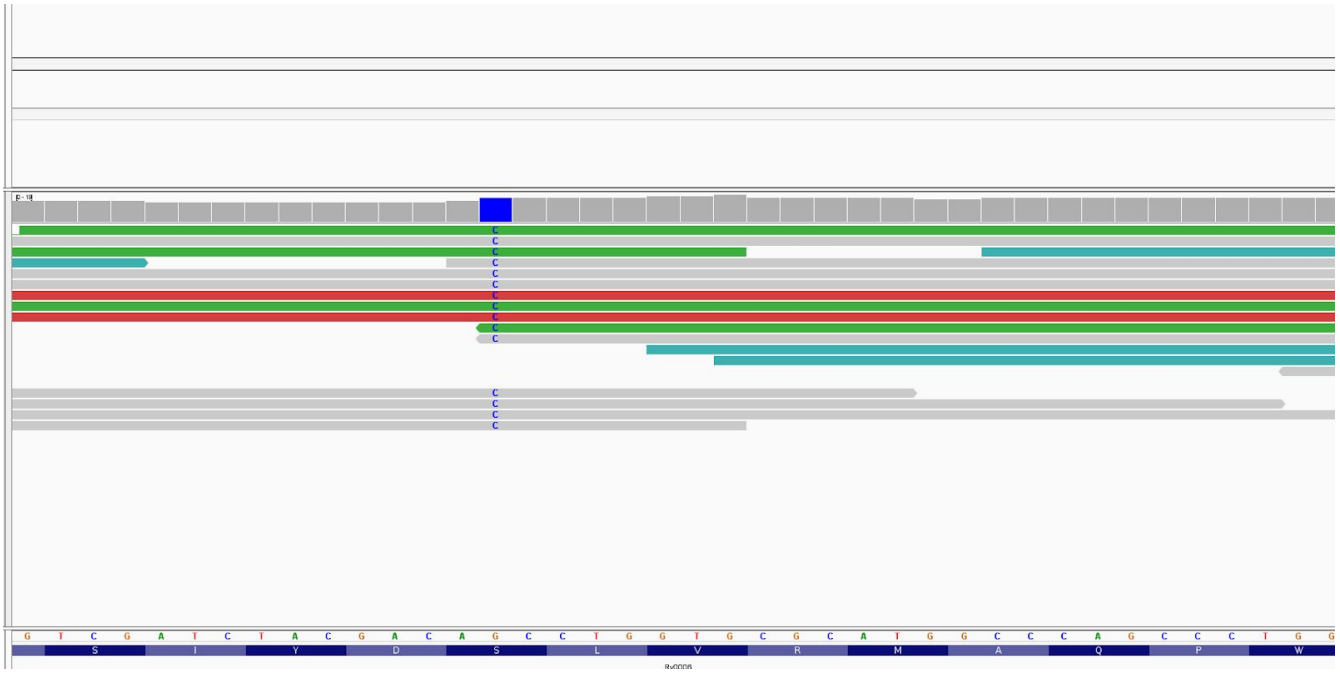
Comment caractériser cette souche multi drogues- résistante?
Connaît-on des gènes impliqués dans la résistance à certains antibiotiques ?
(diminution synthèse paroi, diminue fixation des bact, destruction petite ss unité ribosome)

Mécanismes défense:
ajouts de gènes par ex : des ponts de flux, transporteur dans memb faisant sortir l'antibio

Vérifier les variances au niveau de ces gènes.
ex : gène gyra gyrb (dans .gff -> Rv005, Rv006)



En effet, on peut observer ici quelques SNP et insertions sur quelques reads
Se fier à la pile, pas à un seul read.



Par exemple ici, mutation sur tous les reads d'un C à la place d'un G, modifiant peut-être acide aminé S -> devenu ACC (Thréonine) => substitution mais pas grave S + ou - ⇔ T
2 autres mutations du même type.
Peu de risque d'insertion car sinon, Bactérie ne serait pas viable ?

2ème TD (step6) : Les glandes medulo-surrénales du cerveau => Transcriptomique (RNAseq)
Comparaison niveau expression des gènes entre le cerveau et le réseau medulo-surrénal
Etapes:

Biopsie, récolte d'une coupe de cerveau

Extraction ARN (plus compliqué qu'ADN car - stable et plus fragile) => Aller vite

Plonger ARN dans azote liquide, ou dégradation lysine, blocage enzymes. => Pour gagner du temps et éviter les dégradations.

Pour vérifier état des ARNs => faire plusieurs fois la même expérience (réplicats biologiques et techniques)
Biologique => 3 échantillons au départ, même processus sur les 3
Techniques => même échantillon de départ, puis 3 banques ou 1 banque et séquençage 3 fois.
OU/ET
dopper ech avec des arn connus qui n'ont rien à voir (on sait ce qu'on doit retrouver à la fin) = "spikes"
N'importe quelle pollution ou problème atteindra aussi ces arns connus

Extraction ARN : mRNA, rRNA, tRNA, ncRNA, smRNA (noyau, non codants, +prot=spliceosome(=qui réalise l'épissage))
Ici, plus logique de se concentrer sur les rRNA (ayant une queue polyAAAA, on prends un adaptateur = TTTT et on effectue une Reverse Transcriptase)
Séquençage Paired-end
ADNc obtenu, fragmentation, synthèse, construction librairie.

2 fichiers FastQ pour chaque tissu (2 pour cerveau, 2 pour glande surrénale, Forward-Reverse)
Plus réplicas!

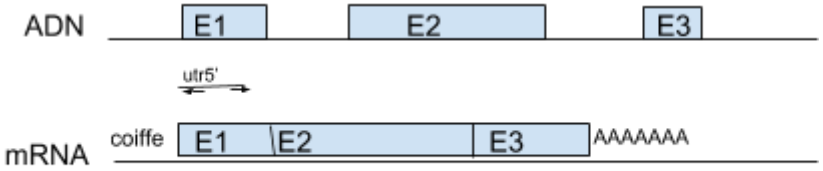
Analyse qualité (FastQC)

Adrenal 1 : per base seq q -> Très bien, rien en dessous de Q20,
per seq Q score -> Bien , autour de Q38
per base seq content -> Mauvais au début des reads (3'?)

Globalement : beaucoup de surreprésentation, problème aux extrémités, biais sur les GC, beaucoup de kmers
Dataset pourri, pas beaucoup de reads, qualité avec grands écart-types
Cause : ARNr restés ? (mauvaise technique avec polyA ?)
Contamination?

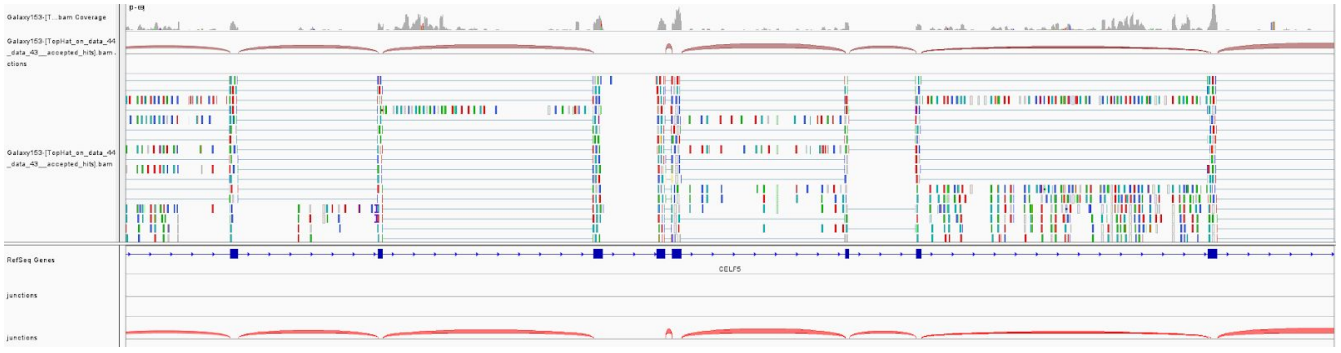
Donc Amélioration: Trimming (en 5'et 3')
Pas de perte de seq mais seul 1er graphe bon.
-> Meilleur au niveau de la qualité

Mapping: ici, pas BWA car non adapté pour les eucaryotes
5' utr = non traduite
ARNm commence au niveau de l'exon 1 dans la partie utr non traduite

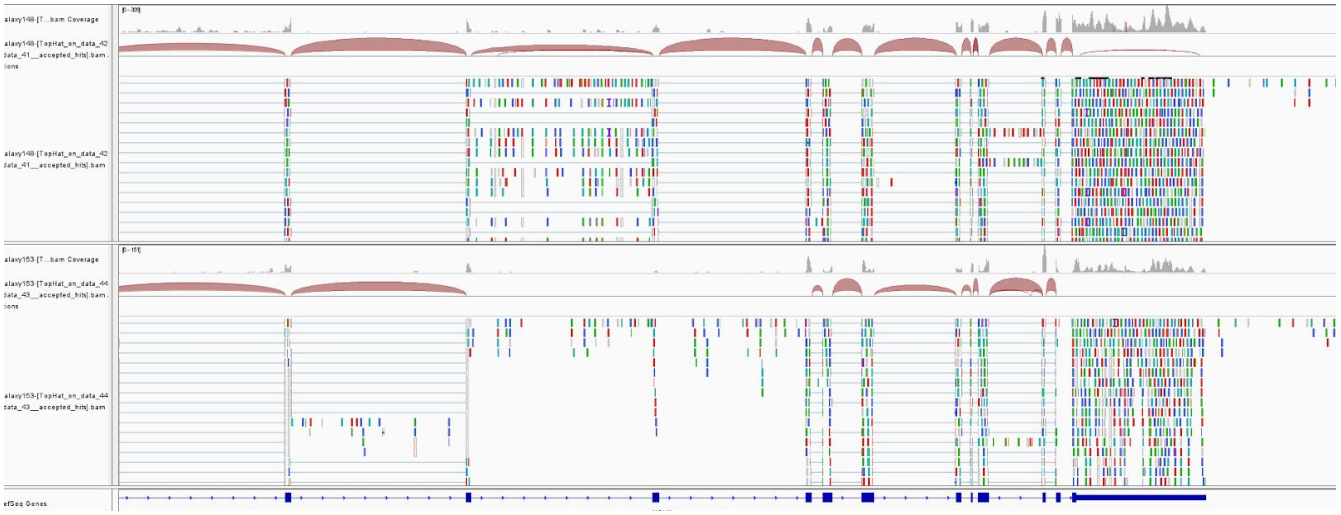


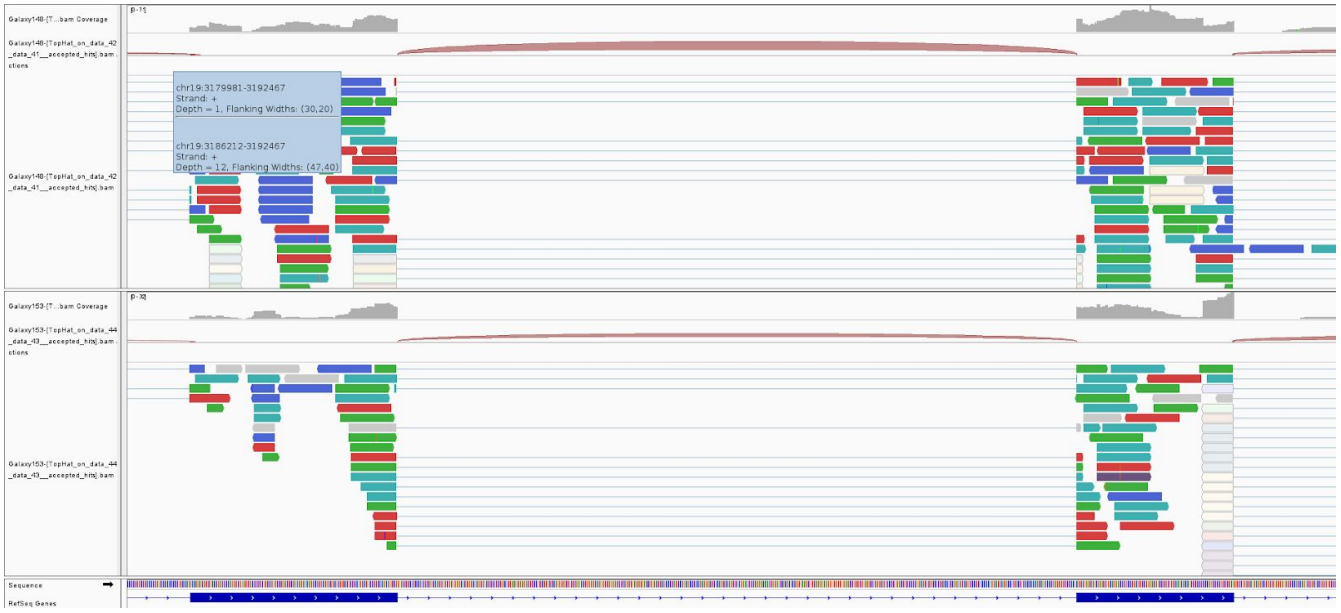
On utilise Top Hat au lieu de BWA pour savoir où les introns ont été coupés, les jonctions intron-exon, pour en tenir compte lors du mapping des reads..
TopHat 1 fois sur Brain, 1 fois sur Surrénal
Plusieurs fichiers générés par TopHat:
align summary (stats)
insertions
deletions
splicejunctions (jonctions intron-exon)
accepted hits ⇔ SAM

Visualisation accepted hits du surrénal dans IGV (dl accepetd hits + junction splices et les charger)



Les régions exoniques de ref sont représentées par des carrés bleus.
Les traces rouges représente les jonction intron-exon.
(y a quelque chose qui concorde, carrés bleus et début/fin des traces rouges représentant les jonctions)





De temps en temps des reads map dans des régins d'introns. Soit ARN non maturé encore, soit problème d'épissage.
Epissage différent entre cerveau et surrénal ?

Utilisation de logiciel qui à partir du mapping vont essayer d'interpréter ce mapping pour reconstruire un mapping testant différentes combinaisons d'épissages alternatifs (tous les transcripts possibles pour chaque gène).
On regarde sur quels exons se chevauchent les reads. Certains vont chevaucher 2 exons qui ne sont pas côtes à côtes donc il s'agira d'un autre épissage.
Quantification des différentes formes des ARNm (stats et proba sur volume, longueur, nombre et distribution des reads).
-> Cufflinks

Comparaison entre les 2, fusion des données pour cerveau et surrénal. et s'en sert de référence en re-superposant des reads ? Comparatif entre les deux. (stats)
-> Cuffmerge