ARCHITECTURE BBRIC

BBRIC est un réseau de bioinformaticiens principalement du département SPE. Ce réseau a mis en place une architecture vous permettant de vous accompagner de la donnée brute à son analyse.

Toutes ces ressources sont accessibles via le portail BBRIC:

https://bbric.toulouse.inra.fr

COMPOSANTS:

ARCHIVE

Gérer sur le long terme les **données brutes** et les métadonnées associées nécessaire à leur analyse et leur soumission

REFERENCE

Gérer sur le long terme les **données de référence** et d'intérêt (**assemblage**, **annotations**) pour la communauté

PROTOCOLS

Référencer les **outils** permettant d'analyser vos données et dont nous assurons la maintenance

WORKSPACE

Gérer **vos résultats d'analyses** réalisées par vos bioinformaticiens

Bioinformatique Biodiversité Représentation & Intégration des Connaissances

Données brutes de séquence

Le format des données brutes de séquence Illumina est FASTQ compressé.

Le format des données brutes de séquence PacBio est HD5.

Ces données brutes sont le point de départ de nombreuses analyses bioinformatiques et en ce sens il est très important de connaître le type de librairie afin de les analyser correctement et produire le meilleur résultat possible.

TYPES:

1) **Single-End** : seule une extrémité d'un fragment d'ADN/ARN est séquencé
Limitations: certaines lectures issues de régions répétées sont alignées non spécifiquement avec les mêmes scores.
Recommandation: Dans le cas de manip de **RNA-Seq sur** des **bactéries**, la librairie doit être **ORIENTEE** à cause des gènes chevauchants.

2) **Paired-End**: les deux extrémités d'un fragment d'ADN/ARN de taille connue et limitée (~400nt) sont séquencées. Les deux séquences produites peuvent être chevauchantes.
Avantage: pour le mapping, l'information de distance permet de lever l'ambiguité sur une des deux reads quand cette dernière s'aligne à plusieurs endroits du génome.
Limitation: les éléments répétés ne peuvent être assemblés avec les algorithmes actuels

3) **Mate-Pair**: les deux extrémités d'un grand fragment (plusieurs kb) d'ADN sont séquencées
Avantage: l'information de longue distance permet d'ordonner et d'orienter des contigs séparés le plus souvent par des éléments répétés
Limitations: il faut produire différentes tailles de banques pour obtenir un résultat satisfaisant; qui plus est ces banques sont polluées par des séquences Paired-End

4) **Single-Molecule (PacBio, MinIon)**: de grands fragments d'ADN/ARN sont séquencés
Avantage: les régions répétées de "*taille standard*" sont directement séquencées. Idem pour les Isoformes qui sont directement séquencés (Iso-Seq).

Type de librairie	Assemblage génome simple	Assemblage génome complexe	Mesure d'expression	Construction Transcriptome de référence	Detection de polymorphisme
Single-End	+/-	--	*	-	+
Paired-End	+	+/-	*	+	+
Mate-Pair	++	+	--	--	--
Single-Molecule	+++	+++	-	+++	En test

--	pas adapté
-	insuffisant
+/-	résultat dépendant fortement de la complexité du génome (présence d'IS, TE)
+	résultat suffisant (ex: couverture du genespace)
+++	haute qualité espérée

Bioinformatique Biodiversité Représentation & Intégration des Connaissances

GFF3

http://www.sequenceontology.org/gff3.shtml

Le format de fichier GFF3 (Generic Feature Format) permet de représenter des features (gene, exon, intron, ncRNA, repeat_region, EST_match, protein_match, …) sur une séquence génomique.

Lisible avec Excel ou dans un genome browser (Jbrowse, IGV, …)

FORMAT :

Fichier texte tabulé, composé obligatoirement de 9 colonnes :

1) SeqId Identifiant de la séquence génomique.
2) Source Origine. Exemple : nom d'un logiciel ou d'une base de données
3) Type Type de la feature. *Doit être un terme SOFA*
http://www.sequenceontology.org/resources/intro.html
4) Start Position de début de la feature.
5) End Position de fin de la feature.
6) Score Exemple : la e-value pour des similarités de séquences
7) Strand [+ ou -] Brin de la feature
8) Phase [0, 1 ou 2] Phase de la feature
9) Attributes Liste d'attributs de la forme clé=valeur

EXEMPLE :

```
##gff-version 3
##sequence-region chr5 1 100000
Chr5 GBK gene 1000 9000 . + . ID=gene01;Name=EDEN
Chr5 GBK mRNA 1000 9000 . + . ID=mRNA01;Parent=gene01;Name=EDEN.1
Chr5 GBK CDS  1201 1500 . + 0 ID=cds01;Parent=mRNA01;Name=EDENPROT.1
Chr5 GBK CDS  1600 8500 . + 0 ID=cds02;Parent=mRNA01;Name=EDENPROT.1
```

VCF
https://vcftools.github.io/specs.html

Le format VCF (Variant Call Format) est un fichier tabulé utilisé pour lister les variations de séquence (SNPs et InDels). Il peut être lu par Excel ou R par exemple mais aussi par des outils de visualisation comme Tablet, IGV ou JBROWSE.

FORMAT: Fichier texte composé de 2 parties : un en-tête et un corps

En-tête : métadonnées: Lignes commençant par ##.
Description des données et des colonnes utilisées dans le corps.

Corps : données Format tabulé avec les colonnes suivantes :

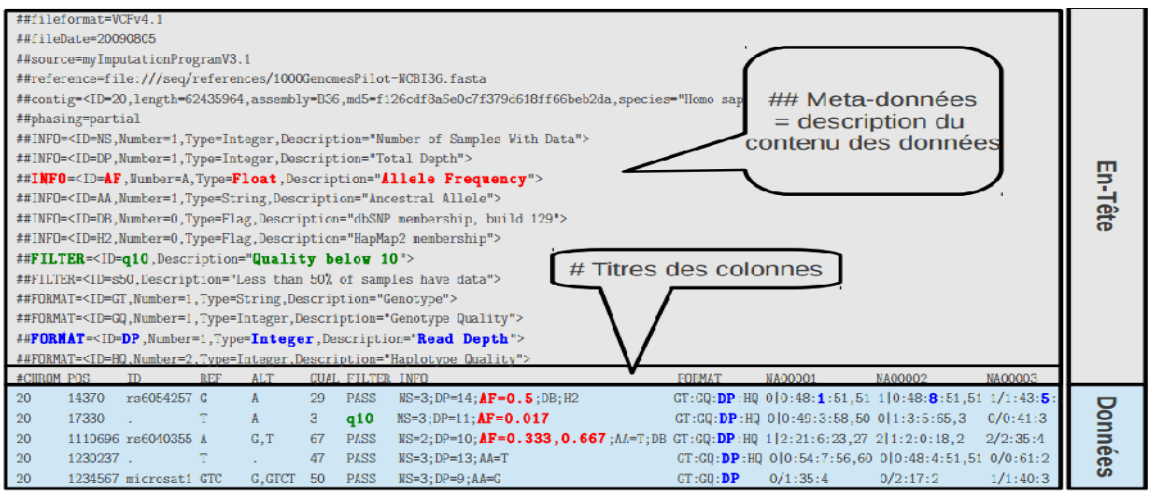
Informations générales sur chaque site :

1. Identifiant du **scaffold** ou du chromosome
2. **Position** du SNP ou de l'INDEL
3. Identifiant du SNP s'il est connu
4. Allèle sur la **référence** (ex: A)
5. Allèle(s) **alternatifs** (ex : G,T)
6. **Score** de qualité de l'allèle (Phred Score : https://fr.wikipedia.org/wiki/Score_de_qualit%C3%A9_phred)
7. Colonne **FILTER** qui indique quels sont les filtres (décrits dans l'en-tête) que l'allèle n'a pas passé
8. Colonne **INFO** dont les champs sont décrits dans l'en-tête (ex:NS=3;DP=14;AF=0.5)
9. Colonne **FORMAT** qui va servir à décoder les colonnes suivantes et dont les champs sont décrits dans l'en-tête (ex:GT:GQ:DP:HQ)

Description des génotypes de chaque échantillon

10 .. Nombre d' échantillons : une colonne par échantillon contenant les **valeurs** correspondantes aux champs indiqués dans la colonne format (ex : 0/1:21:6:23,27).

Ex : ici, la colonne GT renseigne sur le génotype de l'échantillon et la valeur 0/1 indique que l'échantillon a l'allèle référence (0) et l'allèle indiquée comme première alternative (1)




```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sap
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF ALT    QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G   A      29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:
20     17330   .         T   A      3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A   G,T    67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T   .      47   PASS   NS=3;DP=13;AA=T                    GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC G,GTCT 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

InterProScan

https://code.google.com/p/interproscan/wiki/InterProScan5OutputFormats

Le format brut (RAW/TSV) de fichier InterProScan permet de représenter des signatures (DOMAINES, SITES, REGIONS) et leurs annotations fonctionnelles sur une séquence protéique.

Ces données sont générées par le programme InterProScan qui scanne un ensemble de bases de données (PFAM, ProDom, HAMAP) fédérées au sein d'InterPro (http://www.ebi.ac.uk/interpro/about.html).

FORMAT:

Fichier texte tabulé, composé de 11 à 15 colonnes :

1) Identifiant protéine
2) Code unique qui identifie la séquence protéique
3) Longueur de la séquence protéique
4) Banque de signatures cible (ex. Pfam / PRINTS / Gene3D)
5) Numéro d'accession de la signature (ex: PF09103 / G3DSA:2.40.50.140)
6) Description de la signature (e.g. BRCA2 repeat profile)
7) Début de la signature sur la séquence de la protéine
8) Fin de la signature sur la séquence de la protéine
9) Score (e-value) du match retourné par la méthode utilisée pour la banque cible
10) Status : non pertinent en mode automatique
11) Date de l'analyse
12) Numéro d'accession InterPro (ex: IPR002093)
13) Description InterPro (ex: BRCA2 repeat)
14) Annotations GO (Gene Ontology) (ex. GO:0005515)
15) Annotations voies métaboliques (Reactome/KEGG/UniPathWay) (ex: REACT_71)

EXEMPLE:

KL638872	62daab23d12 71f12095d2da 269f76b2e	332	Hamap	MF_00059	DNA-directed RNA polymerase subunit alpha [rpoA].	5	316	34.74	T	06/10/2015	IPR011773	DNA-directed RNA polymerase, alpha subunit	GO:0003677 GO:0003899 GO:0006351	KEGG: 00230+2.7.7.6 KEGG: 00240+2.7.7.6
KL638872	62daab23d12 71f12095d2da 269f76b2e	332	Pfam	PF01000	RNA polymerase Rpb3/RpoA insert domain	60	154	1.7E-24	T	06/10/2015	IPR011262	DNA-directed RNA polymerase, insert domain	GO:0003899 GO:0006351 GO:0046983	KEGG: 00230+2.7.7.6 KEGG: 00240+2.7.7.6 Reactome: REACT_1788