Bioinformatique
Biodiversité
Représentation
& Intégration
des Connaissances

Session de Formation
« annotation fonctionnelle & analyse du polymorphisme »
Vendredi 16 Octobre 2015

LIPM

Programme

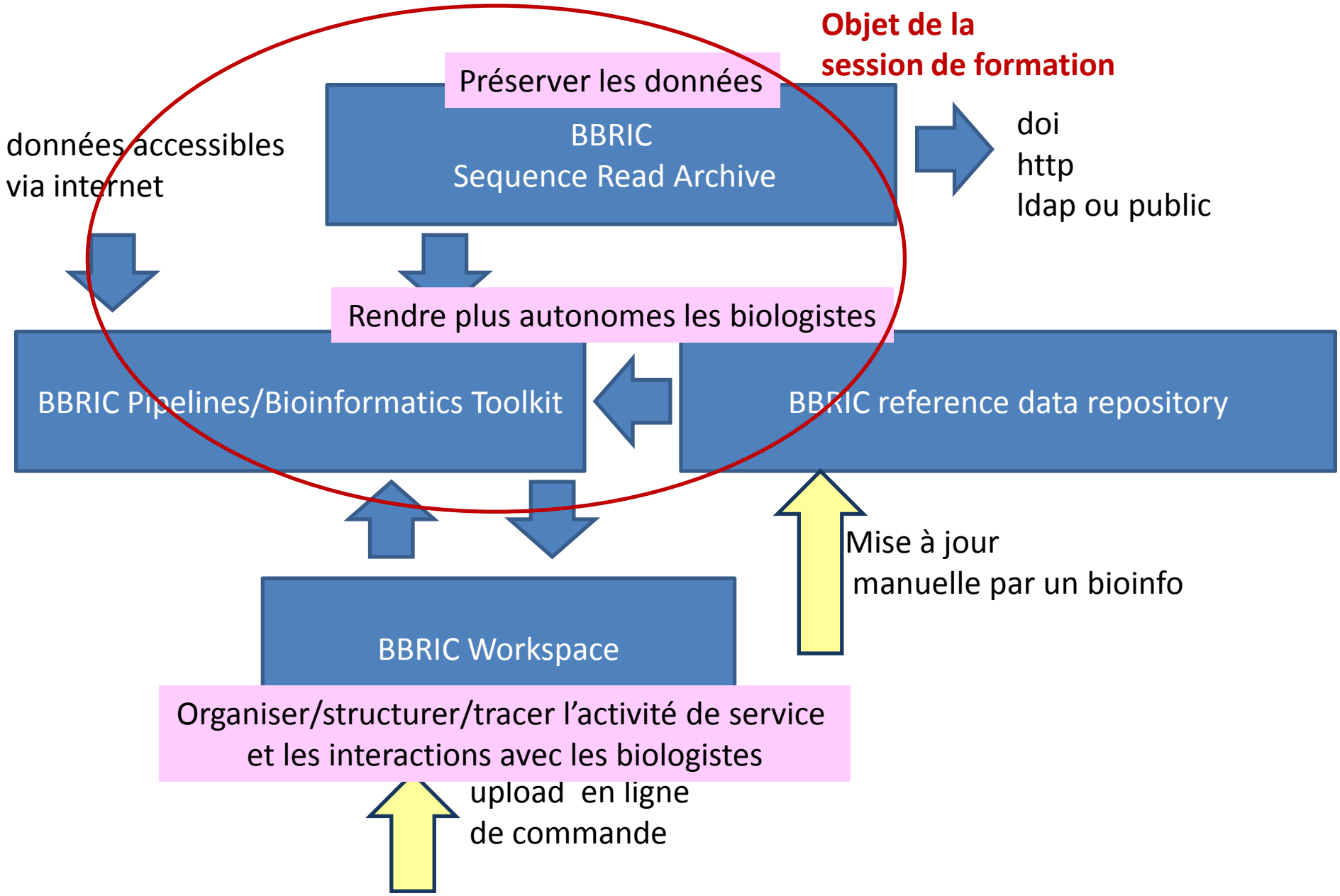
MODULE	RESPONSABLE	PLAGE
Introduction Galaxy - Archive - Reference	S. Carrere	9h-10h
Analyse du polymorphisme - Détection de SNPs à partir d'un génome de référence	L. Cottret	10h-11h30
PAUSE	10h45-11h	
Analyse du polymorphisme - Prédiction de l'effet	S. Carrere	11h30-12h
Analyse du polymorphisme - Construction d'une matrice de fréquences alléliques		12h-12h30
PAUSE	12h30-13h30	
RNASeq - Mesure de l'expression & Contrôle Qualité	E. Sallet	13h30-15h15
PAUSE	15h15-15h30	
Génomique comparative - Analyse de familles de protéines orthologues	L. Cottret	15h30-16h30
Métabolisme - Identification de voies métaboliques	L. Cottret	16h30-17h30

Modules	Expert
Détection des SNPs à partir de données de re-séquençage (*GATK*)	Fabrice Legeai (IGEPP)
Détection des SNPs (*samtools/VarScan*) pour analyse de leurs effets	Ludovic Legrand (LIPM)
Détection des SNPs (*samtools/VarScan*) pour calcul des fréquences alléliques	Ludovic Legrand (LIPM)
Mesure de l'expression a partir de données RNAseq	Ludovic Legrand (LIPM)
Indicateurs qualité d'une expérience de RNAseq	Joseph Tran (IJPB) Adeline Simon (BIOGER)
Analyse comparative de protéomes d'espèces apparentées (*orthoMCL*)	Martial Briand (IRHS) Corinne Rancurel (ISA) Sébastien Carrere & Ludovic Cottret (LIPM)

Présentation de l’architecture bioinformatique et du portail BBRIC

Sébastien Carrere

Architecture bioinformatique cible



https://bbric.toulouse.inra.fr

https://bbric.toulouse.inra.fr

En termes d'analyse de données

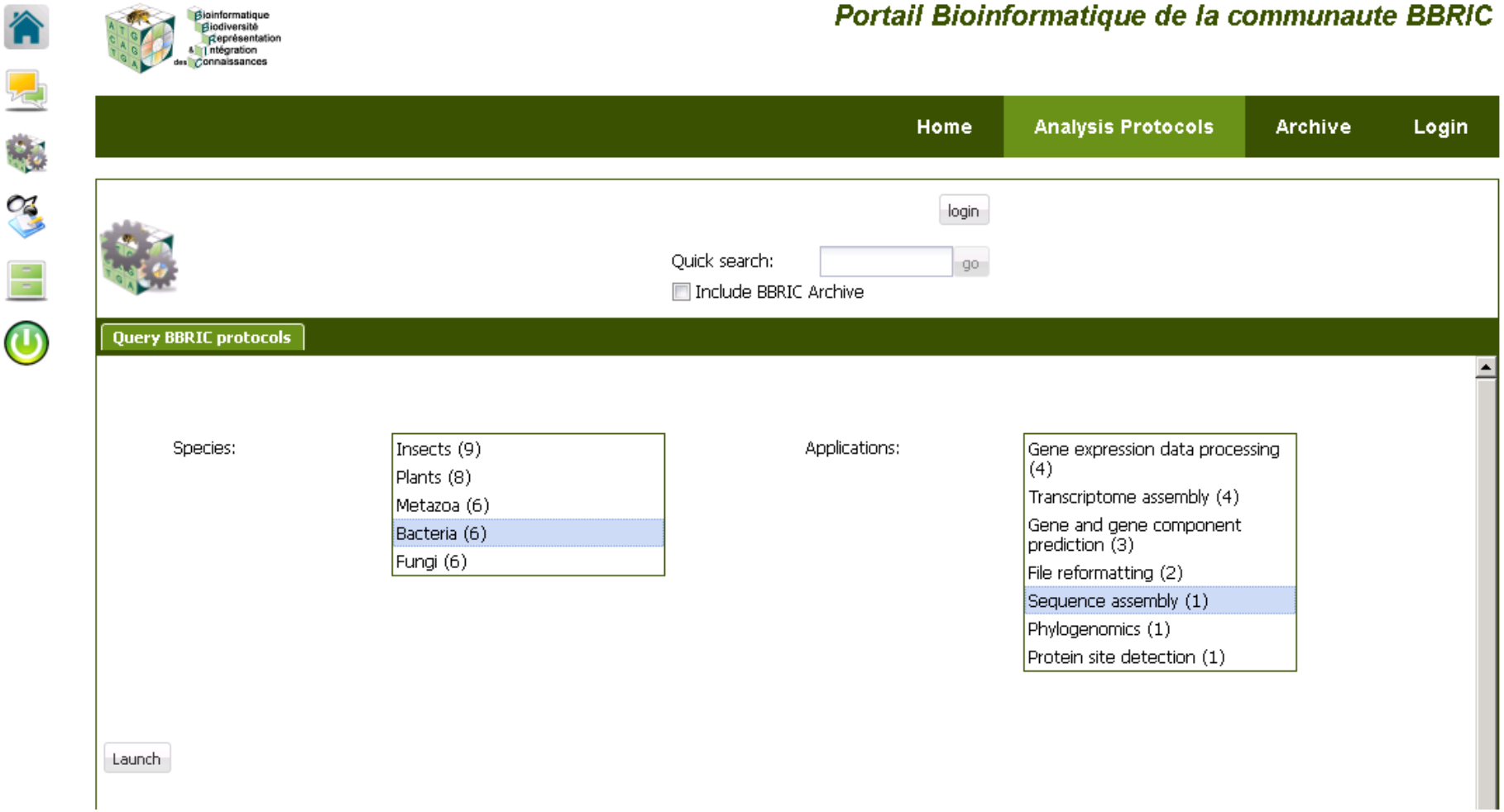
➔Il n'y a pas UNE façon de faire les choses

➔Pour répondre à une question, les données et les programmes changent au fil du temps

➔On n'a pas forcément *a priori* connaissance de là où se trouve l'outil qui va permettre de répondre à une question

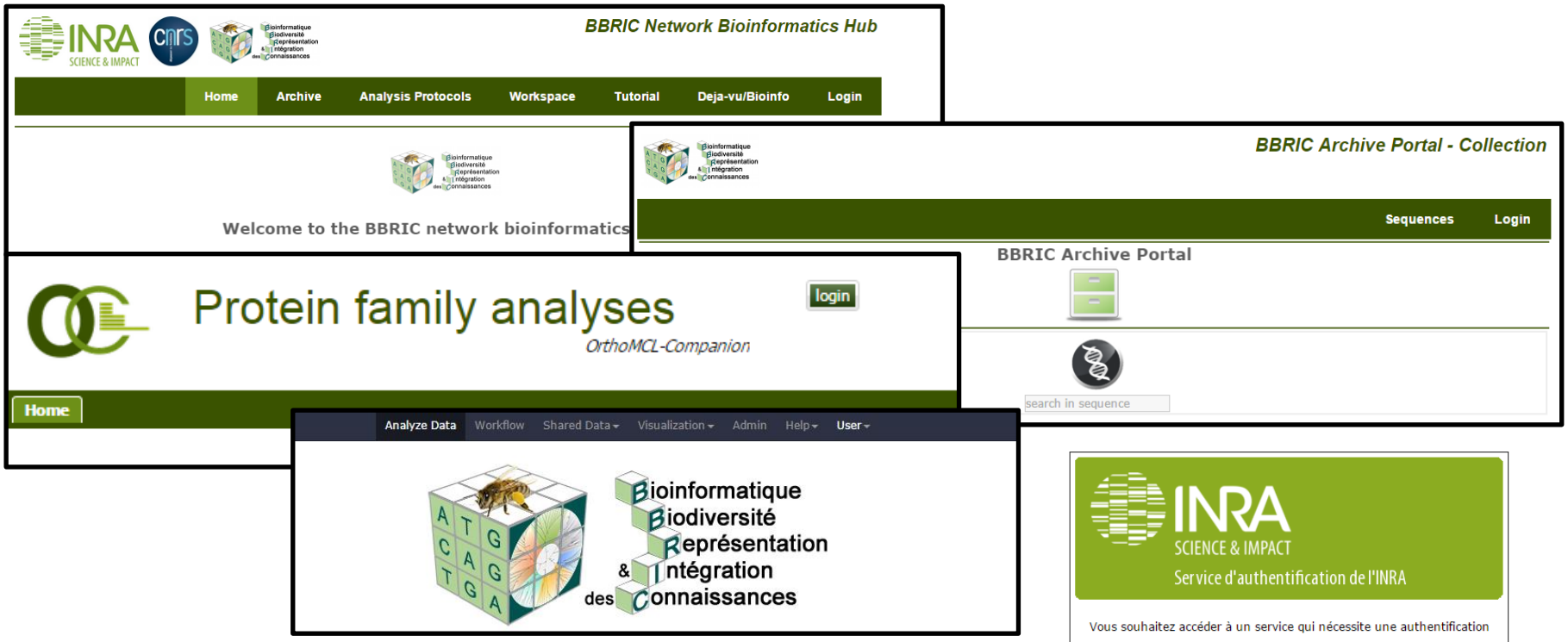
➔ On interroge le système BBRIC a partir d'un problème (ex: Bactérie/Assemblage)

Protocoles d'analyses BBRIC



Le système propose la liste des programmes potentiellement pertinents avec un lien hypertexte qui renvoie vers le formulaire permettant de faire l'analyse

Authentification simplifiée



Présentation de l’Archive Séquences BBRIC

Sébastien Carrere

Archive BBRIC

Objectifs

- Conserver sur le **long terme** les données « brutes » de séquence et les informations associées à la génération des données (métadonnées)
- Faciliter la **soumission** des séquences aux **banques publiques** lors des publications
- Fournir un **minimum d'information** permettant l'analyse (automatique) des données

BBRIC Archive : 182 espèces ; 2518 échantillons (une centaine de publiés!) ; 69 projets ; 7.3 Tb compressées (oct 2015)

jerome.gouzy@toulouse.inra.fr
jgouzy@inra.fr

Bioinformatique Biodiversité Représentation & Intégration des Connaissances

BBRIC Archive Portal - Collection

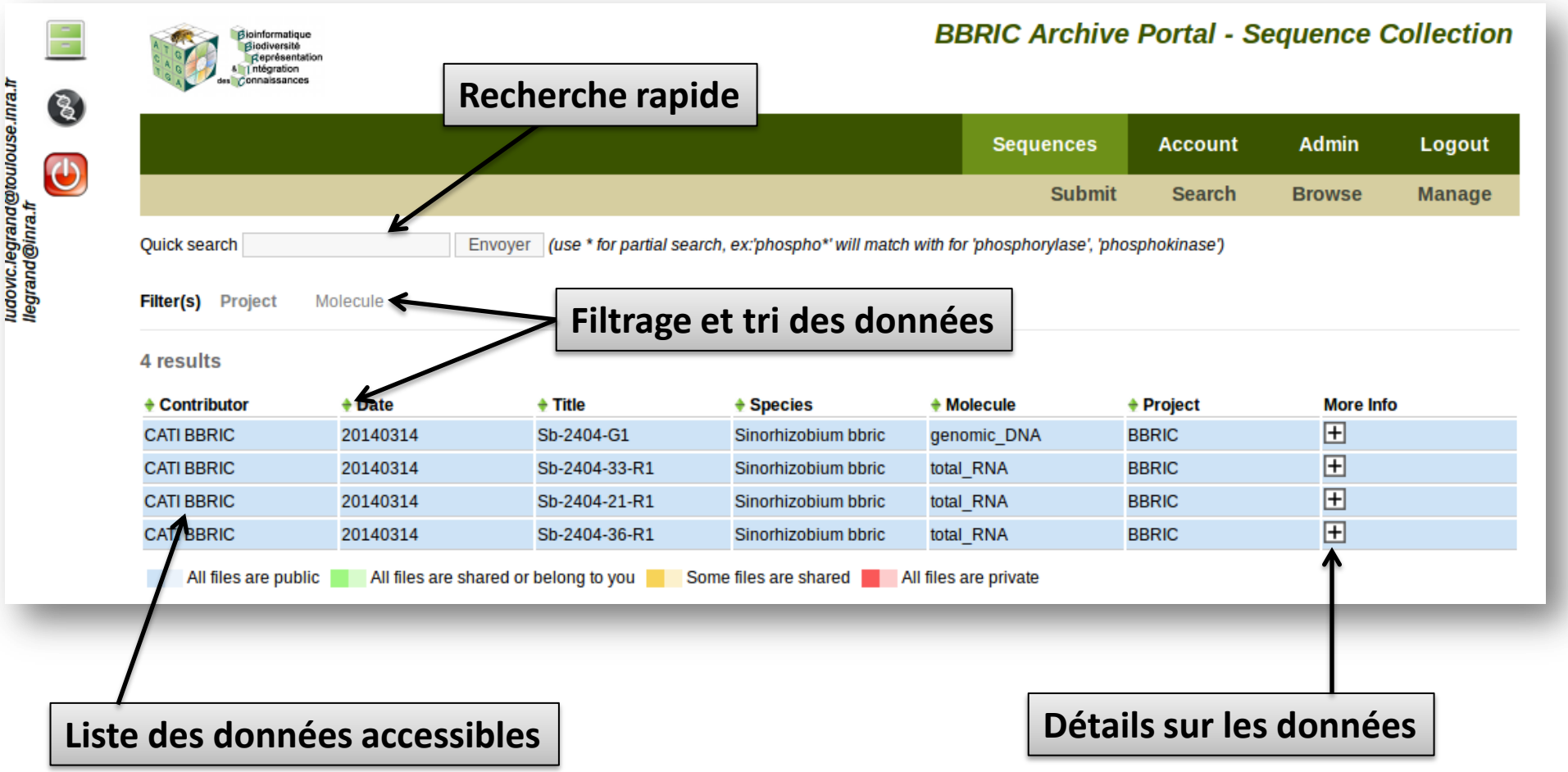
Sequences | Account | Admin | Logout

Statistics | Users

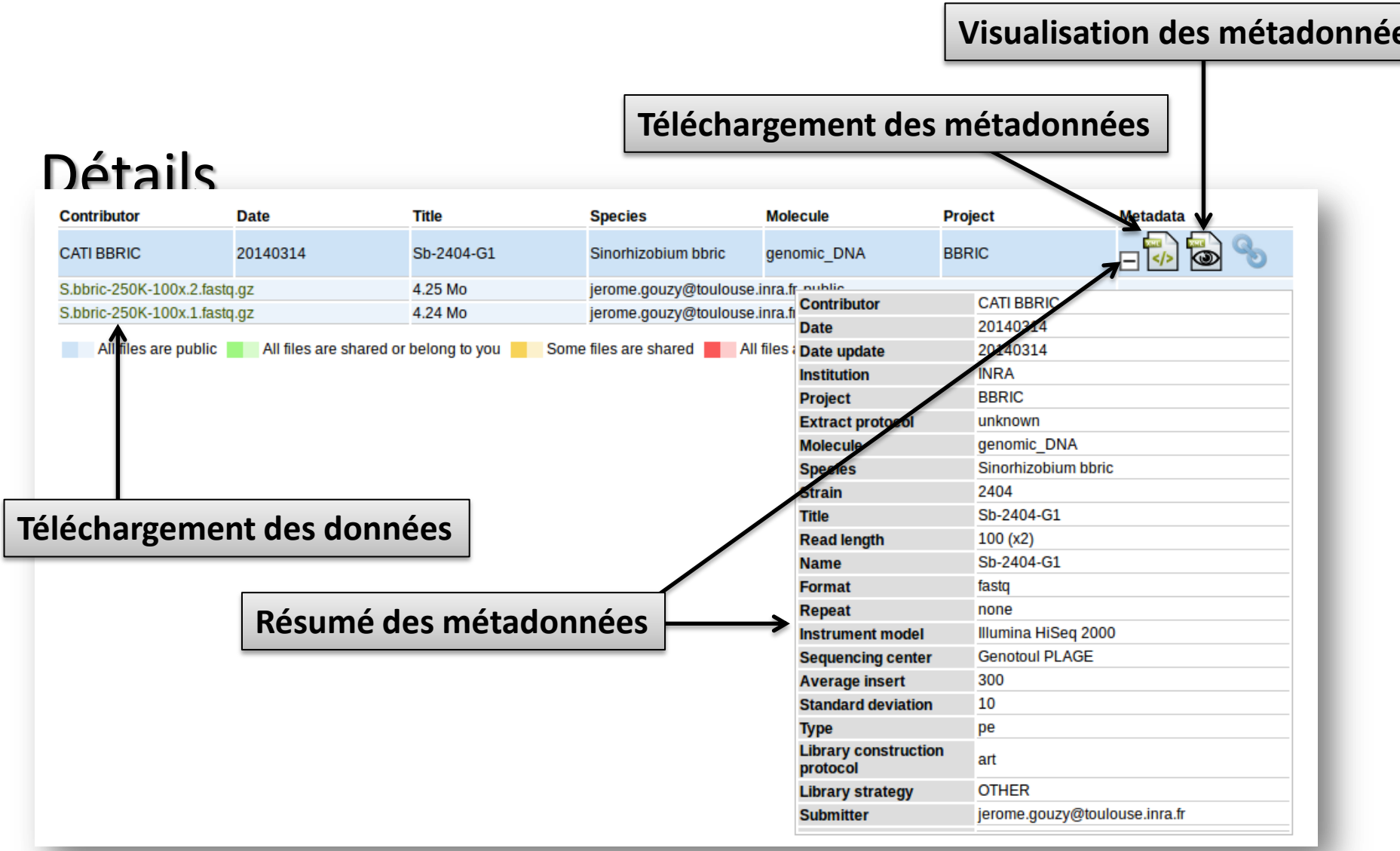
Statistics by species

Species	Data size (Go)	Entry count	Molecule	Projects
Acidovorax citrulli	0.71	1	genomic_DNA	AIP Bioressources
Aeschynomene evenia	21.83	2	polyA_RNA	SESAM2
Alopecurus myosuroides	92.43	26	total_RNA	ALOMY
Arabidopsis thaliana	1235.26	216	genomic_DNA,polyA_RNA	PoolSeq48,QUANTIREX,Xanthomix,MYBLYON,ARATH-DamSeq-RRS1,RESURRECTION,CBMK,MYBSeq
Bacillus sp.	4.97	7	genomic_DNA	RAINS-Project
Brassica oleracea	214.20	22	other,total_RNA	metaSEED,Brassichax
Burkholderia caryophylli	0.78	1	genomic_DNA	AIP Bioressources
Clavibacter michiganensis sbsp insidiosus	0.52	1	genomic_DNA	AIP Bioressources
Clavibacter michiganensis sbsp michiganensis	0.39	1	genomic_DNA	AIP Bioressources
Cryphonectria parasitica	16.16	18	genomic_DNA	Cryphomic
Curtobacterium flaccumfaciens pv falccumfaciens	0.34	1	genomic_DNA	AIP Bioressources
Dickeya dianthicola	0.52	1	genomic_DNA	AIP Bioressources
Diplocarpon rosae	5.79	1	genomic_DNA	ROADMOVIE
Ditylenchus dipsaci	7.10	1	total_RNA	Nematargets
Erwinia amylovora	0.56	1	genomic_DNA	AIP Bioressources
Erwinia sp.	1.86	2	genomic_DNA	RAINS-Project
Lolium sp.	197.82	30	total_RNA	LOLSS
Medicago sativa	386.64	4	genomic_DNA	SEQLuz
Medicago truncatula	519.18	126	other,total_RNA,polyA_RNA	DIAGNOGENE,MIRMED,Symbimics,NFYAKIN,ANR-09-BLAN-0241 MycSignalling
Meloidogyne arenaria	6.25	1	total_RNA	Nematargets
Meloidogyne incognita	22.08	9	total_RNA	Nematargets
Meloidogyne javanica	6.11	1	total_RNA	Nematargets
Microbacteriacae sp.	0.65	1	genomic_DNA	RAINS-Project

Interface Archive (1/2)

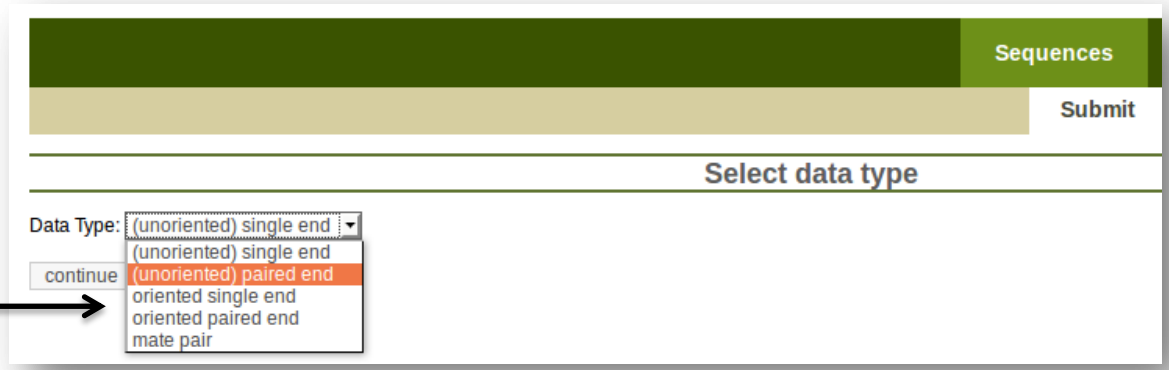


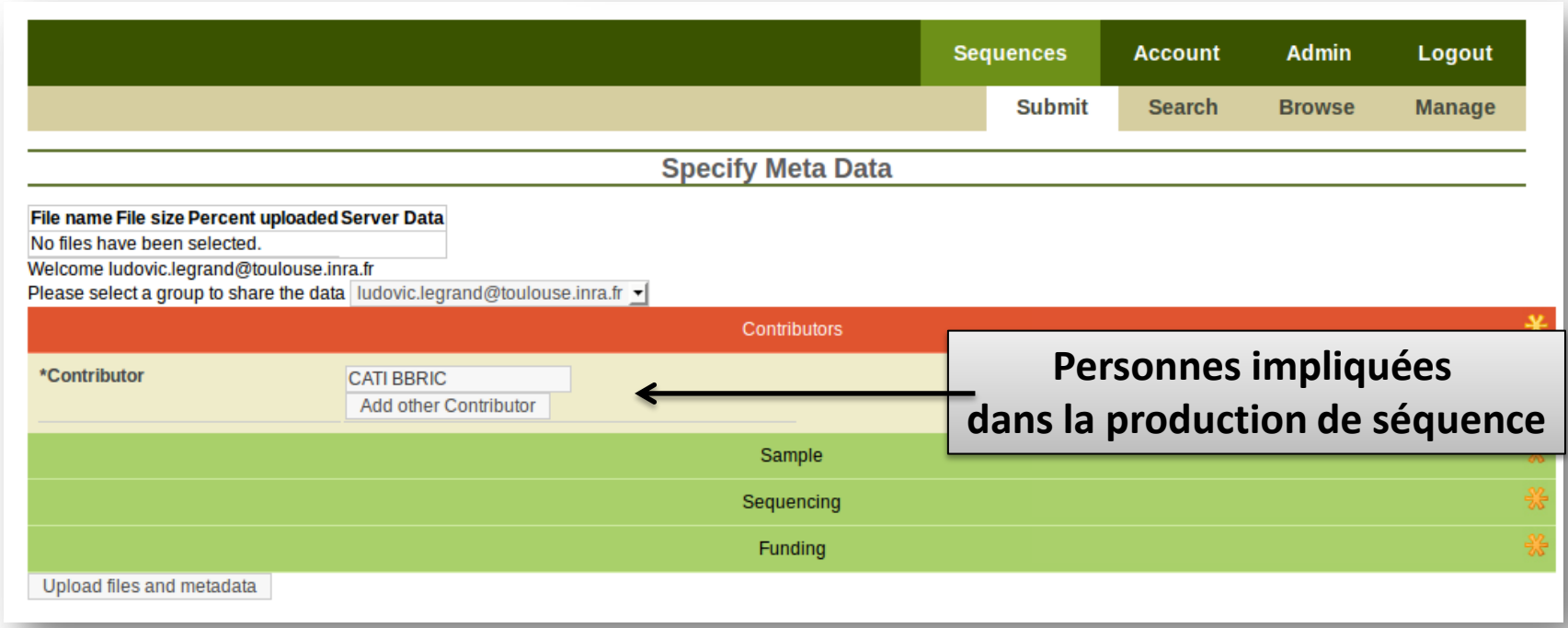
Interface Archive (2/2)



Soumission de séquences à l'Archive (1/4)

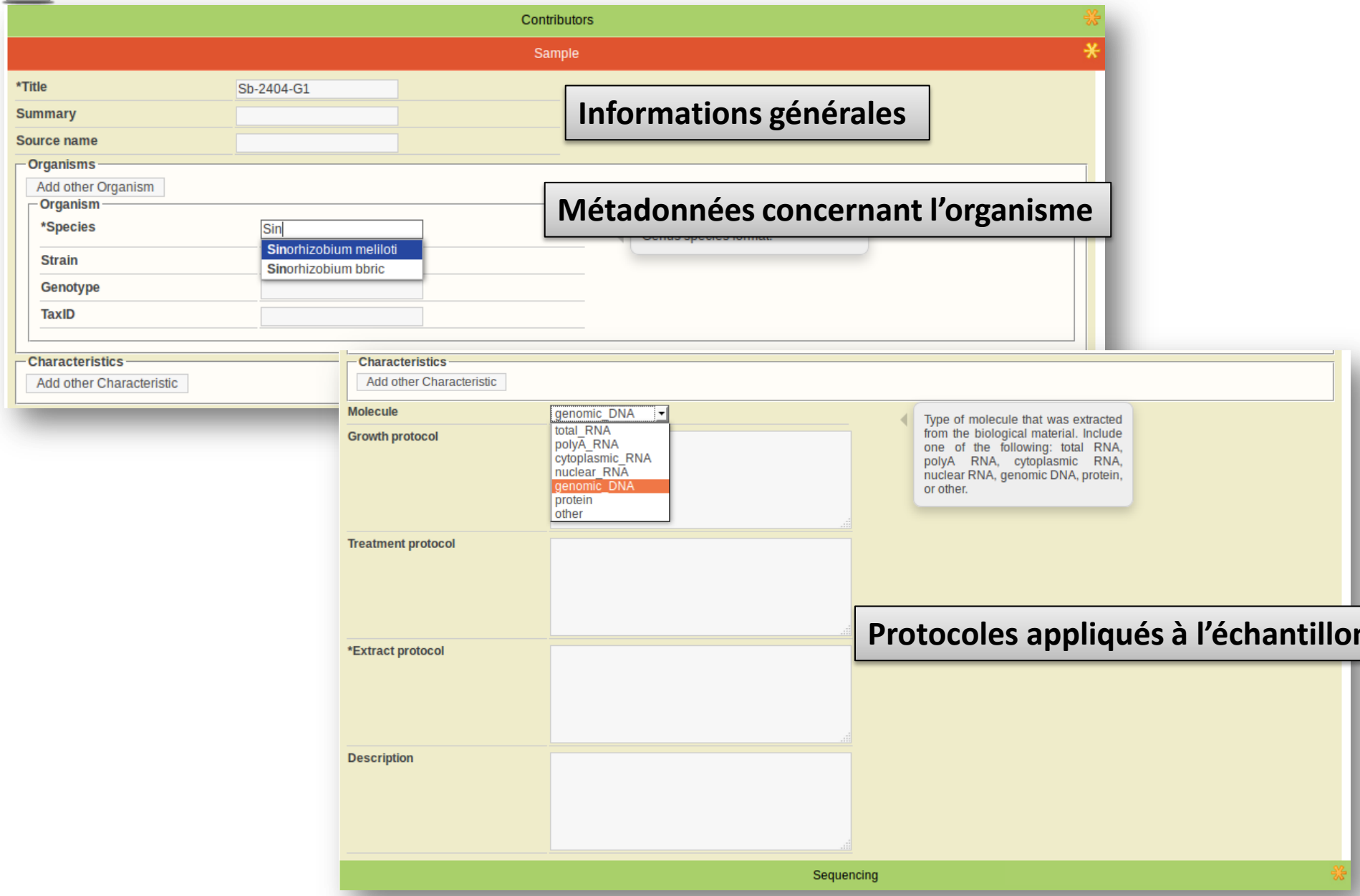
Choix du type de librairie



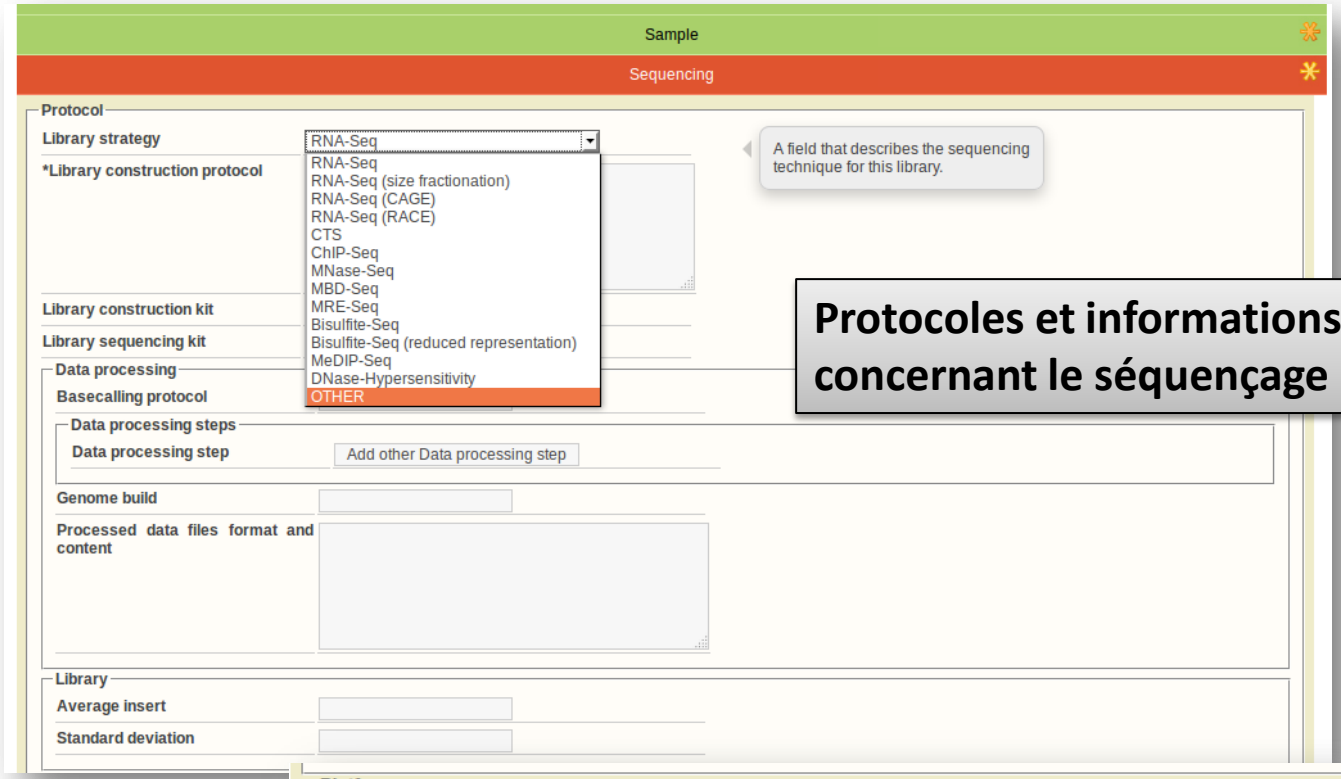


Personnes impliquées dans la production de séquence

Soumission (2/4): Description de l'échantillon biologique

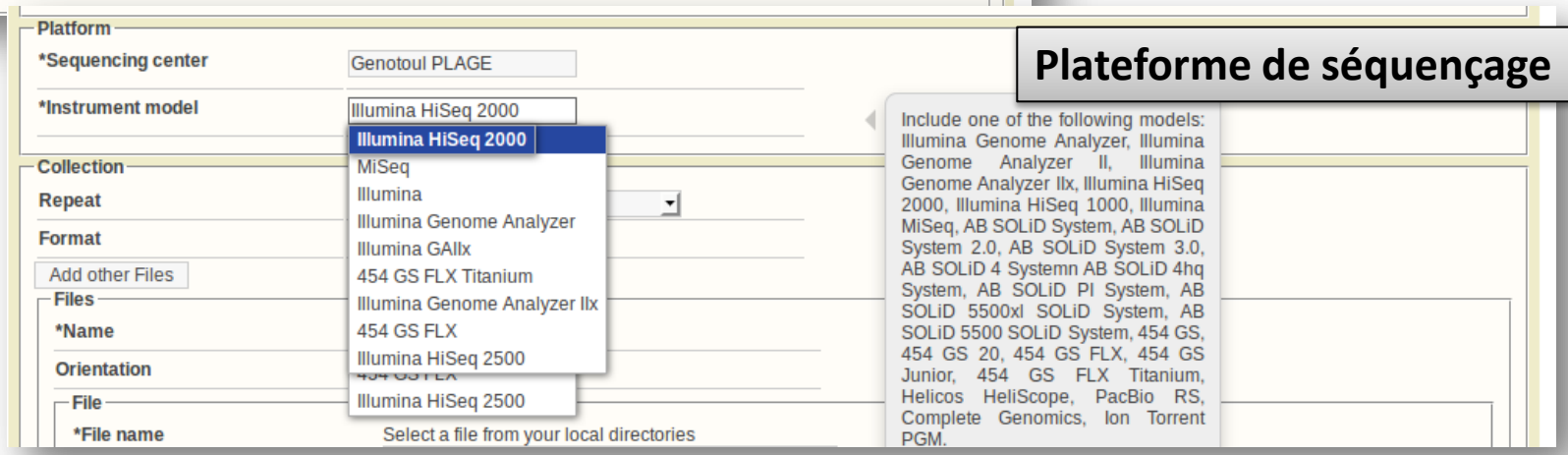


Soumission (3a/4): Description du protocole de séquençage



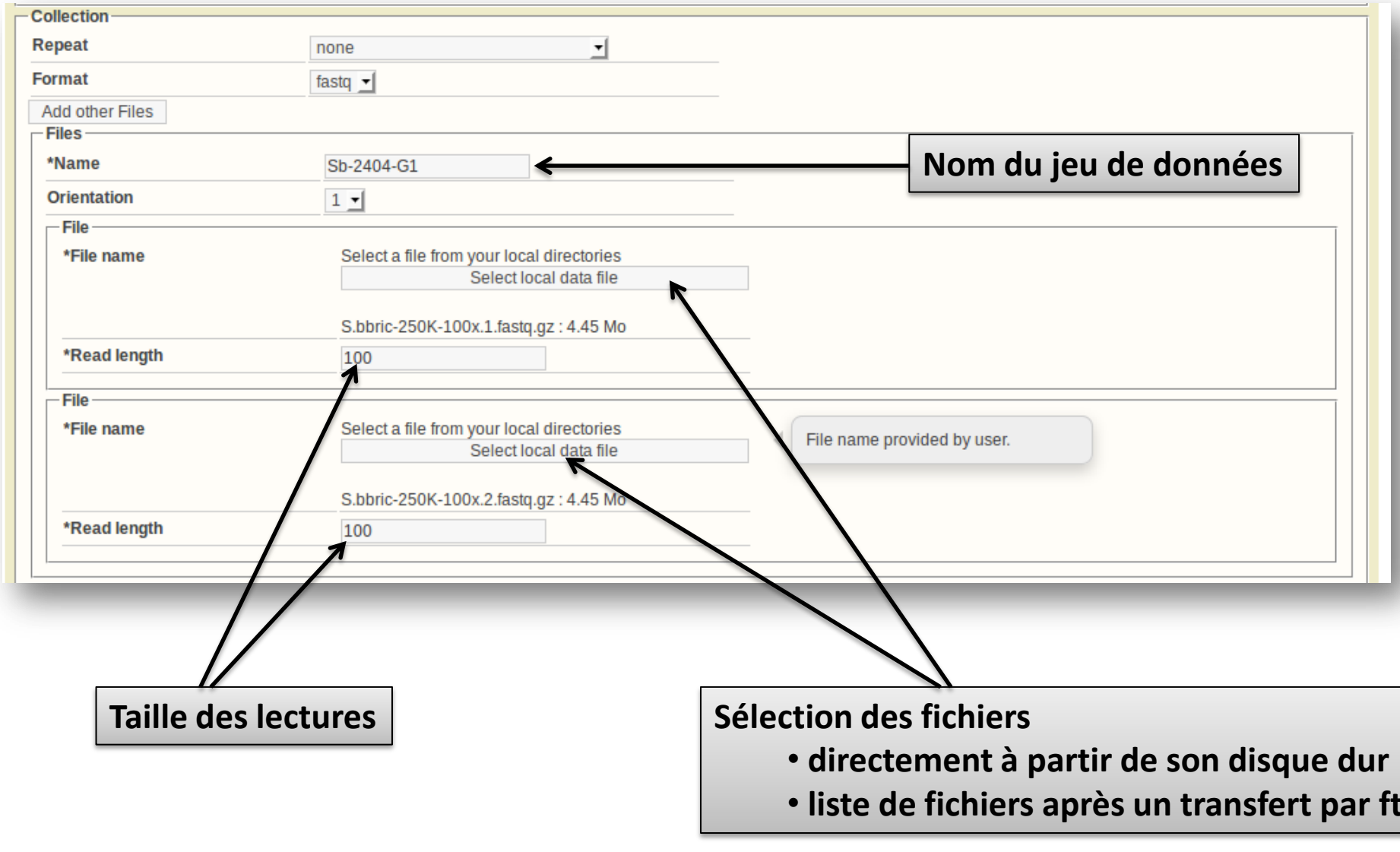
Protocoles et informations concernant le séquençage

Informations provenant du centre de séquençage

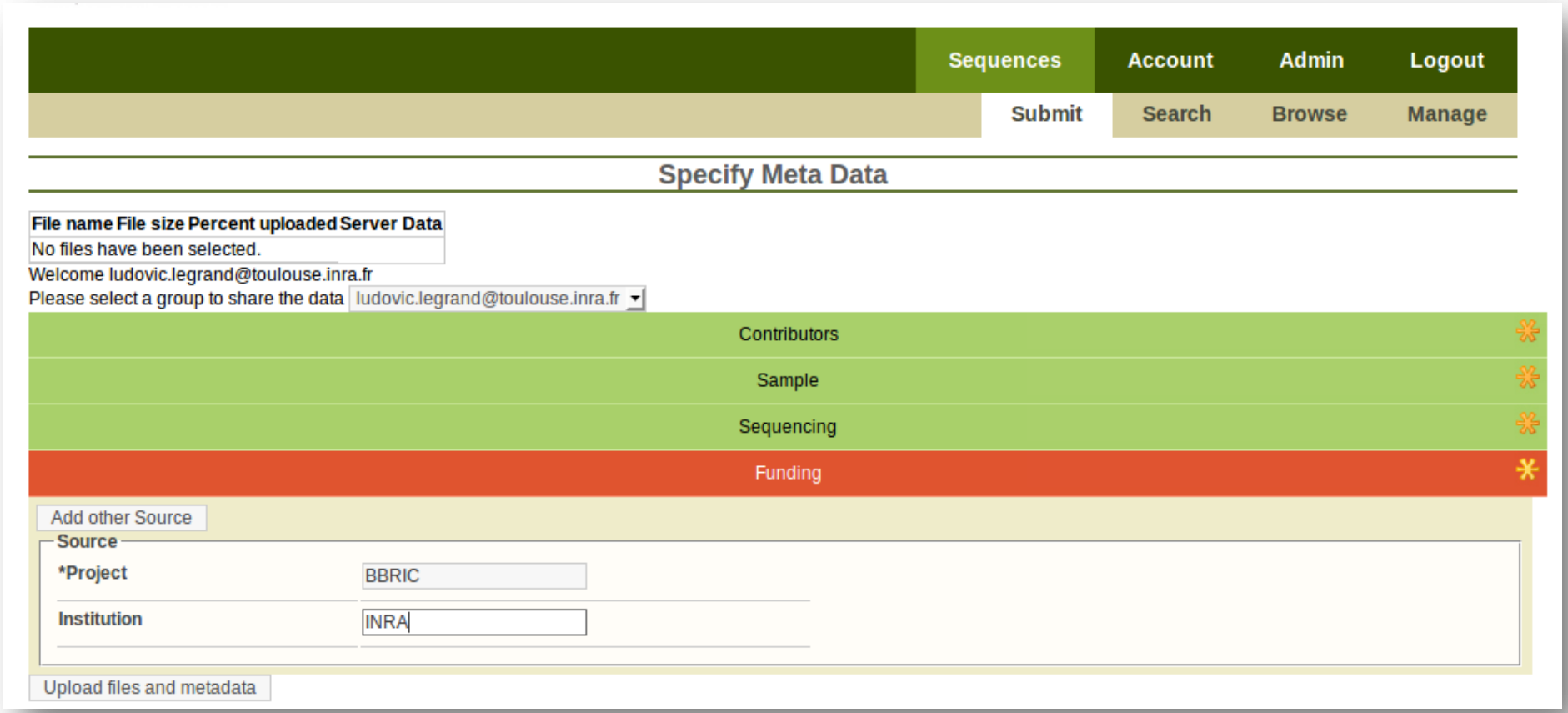


Plateforme de séquençage

Soumission (3b/4): Upload des données



Soumission (4/4) Financement et soumission



Informations sur le financement

Contrôles d'intégrité et droits d'accès

- Validation des métadonnées et des données
  - contrôle de la taille des lectures
  - contrôle d'intégrité sur les paires pour les données paired-end et mate-pairs

  ➔ Envoi d'un email (échec ou succès)

- Possibilité d'associer les données à des groupes d'utilisateurs

Best-practice @ LIPM

- → *Mail fournisseur: données disponibles*
  1. Vous transférez à votre bioinformaticien
  2. Votre bioinformaticien charge les fichiers dans votre espace
  3. Vous saisissez 1 première fiche
  4. Le bioinformaticien valide avec vous
  5. Vous utilisez cette fiche comme *template* pour vos 300 échantillons suivants

Présentation des Références BBRIC

Sébastien Carrere

Reference BBRIC

Objectifs

- **Centraliser** les données de **référence** qui sont **utiles** pour vos analyses
  - Assemblages (génomes, transcriptomes)
  - Annotations structurales
  - Annotations fonctionnelles
  - Extractions de séquences (genes/mRNA/CDS/prot/ncRNA)
- Fournir un **minimum d'information** permettant de tracer leur origine

References Login

Quick search search in reference Valider *(use * for partial search, ex:'Xantho*' will match all data produced on Xanthomonas)*

Filter(s) Type Species

8 results

Species	Title	Type	Date	Accession	XRef	More Info
Helianthus annuus	HaT13l	Annotation Sequence:Proteins Sequence:mRNA	20150806		www.heliagene.org/HaT13l	⊞
Medicago truncatula	Mt20120830-Symbimics	Annotation Sequence:CDS Sequence:Genes Sequence:Genome Sequence:Proteins Sequence:mRNA Sequence:ncRNA	20150810		doi:10.1111/tpj.12442 www.ncbi.nlm.nih.gov/sra/SRP028599	⊞
Rosa chinensis	ROSCH-Transcriptome	Annotation Sequence:Proteins Sequence:mRNA	20150810		doi:10.1186/1471-2164-13-638 bbric-archive.toulouse.inra.fr:sequence	⊞
Lolium sp.	LOLSS-Transcriptome	Annotation Sequence:Proteins Sequence:mRNA	20150806		doi:10.1007/s11103-015-0292-3 bbric-archive.toulouse.inra.fr:sequence	⊞
Helianthus annuus	Ha412v1r1-bronze	Annotation Sequence:CDS Sequence:Genes Sequence:Genome Sequence:Proteins Sequence:mRNA Sequence:ncRNA	20150806		sunflowergenome.org:HA412.v1.1.bronze.20141015.fasta.gz www.heliagene.org:about_annotation.html	⊞
Medicago truncatula	JCVI-Mt4.0	Annotation Sequence:CDS Sequence:Genes Sequence:Genome Sequence:Proteins Sequence:mRNA Sequence:ncRNA	20150806		ftp.jcvi.org:Mt4.0v1_genes_20130731_1800.gff3 ftp.jcvi.org:JCVI.Medtr.v4.20130313.fasta	⊞
Alopecurus myosuroides	ALOMY-Transcriptome	Annotation Sequence:Proteins Sequence:mRNA	20150806		doi:10.1186/s12864-015-1804-x	⊞
Arabidopsis thaliana	TAIR10	Annotation Sequence:CDS Sequence:Genes Sequence:Genome Sequence:Proteins Sequence:mRNA Sequence:ncRNA	20150812		www.arabidopsis.org:README_TAIR10.txt	⊞

All files are public | All files are shared or belong to you | Some files are shared | All files are private

Visualisation des métadonnée

Téléchargement des métadonnées

Species	Title	Type	Date	Accession	XRef	Metadata
Medicago truncatula	JCVI-Mt4.0	Annotation Sequence:CDS Sequence:Genes Sequence:Genome Sequence:Proteins Sequence:mRNA Sequence:ncRNA	20150806		ftp.jcvi.org:Mt4.0v1_genes_20130731_1800.gff3 ftp.jcvi.org:JCVI.Medtr.v4.20130313.fasta	
JCVI-Mt4.0.gff3		116.95 Mo	lipm_bioinfo, sebastien.carrere@toulouse.inra.fr, public			
JCVI-Mt4.0_cds.fasta		96.52 Mo	lipm_bioinfo, sebastien.carrere@toulouse.inra.fr, public			
JCVI-Mt4.0_ncrna.fasta		6.03 Ko	lipm_bioinfo, sebastien.carrere@toulouse.inra.fr, public			
JCVI-Mt4.0_genome.fasta		399.33 Mo	lipm_bioinfo, sebastien.carrere@toulouse.inra.fr, public			
JCVI-Mt4.0_mrna.fasta		119.32 Mo	lipm_bioinfo, sebastien.carrere@toulouse.inra.fr, public			
JCVI-Mt4.0.gtbl		639.40 Mo	lipm_bioinfo, sebastien.carrere@toulouse.inra.fr, public			
JCVI-Mt4.0_gene.fasta		169.22 Mo	lipm_bioinfo, sebastien.carrere@toulouse.inra.fr, public			
JCVI-Mt4.0_prot.fasta		40.09 Mo	lipm_bioinfo, sebastien.carrere@toulouse.inra.fr, public			

Téléchargement direct des données

Liens vers source des données:
- Données brutes ayant été utilisées (Archive)
- Publication
- Origine des données (TAIR/NCBI)

Présentation de l'environnement web « Galaxy »

Sébastien Carrere

- Portail web
- Accès simplifié à de nombreux outils bioinformatiques
- Moins puissant que la ligne de commande, mais suffisant pour beaucoup d'analyses
- Projet international très actif, grande communauté
- Aide : tutoriaux, vidéos d'exemples
  - http://wiki.galaxyproject.org/Learn

- Beaucoup de fonctionnalités
  - Nombreuses formations dédiées à Galaxy
- Aujourd'hui
  - Utilisation de workflows conçus par BBRIC
  - Lien avec l'architecture BBRIC
  - Traitement de données issues de l'Archive

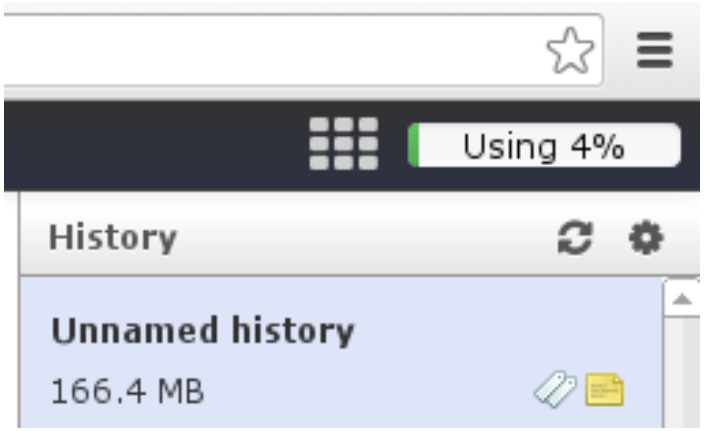
Liste d'outils

Historique qui contient des Datasets

Lancement des outils
Visualisation des résultats
...

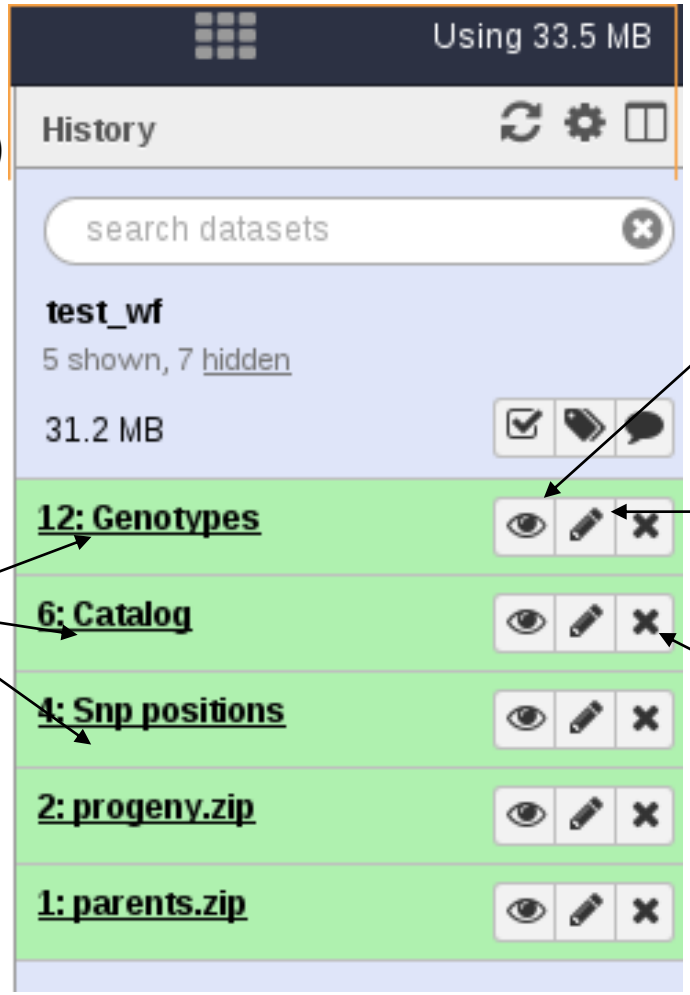
- Analyze Data
  - Lancement d'outil, visualisation de résultats
- Workflow
  - Conception et **utilisation** de workflows
- Shared Data
  - Accès à des données mises à disposition par les administrateurs

- Sources de données multiples
  - Upload depuis le portail web
  - **Shared Data**
  - **Archive BBRIC**
  - FTP
- Volume total limité par utilisateur
  - Quota défini par l'administrateur



History

Chaque dataset :
- Résultat d'un outil
- Donnée ajoutée (upload, archive)



Aperçu du contenu

Modification
-nom
-type

Suppression

Les datasets s'empilent
au fur et à mesure

Gris = outil pas encore lancé
Jaune = l'outil s'exécute
Vert = terminé avec succès
Rouge = terminé avec une erreur

History

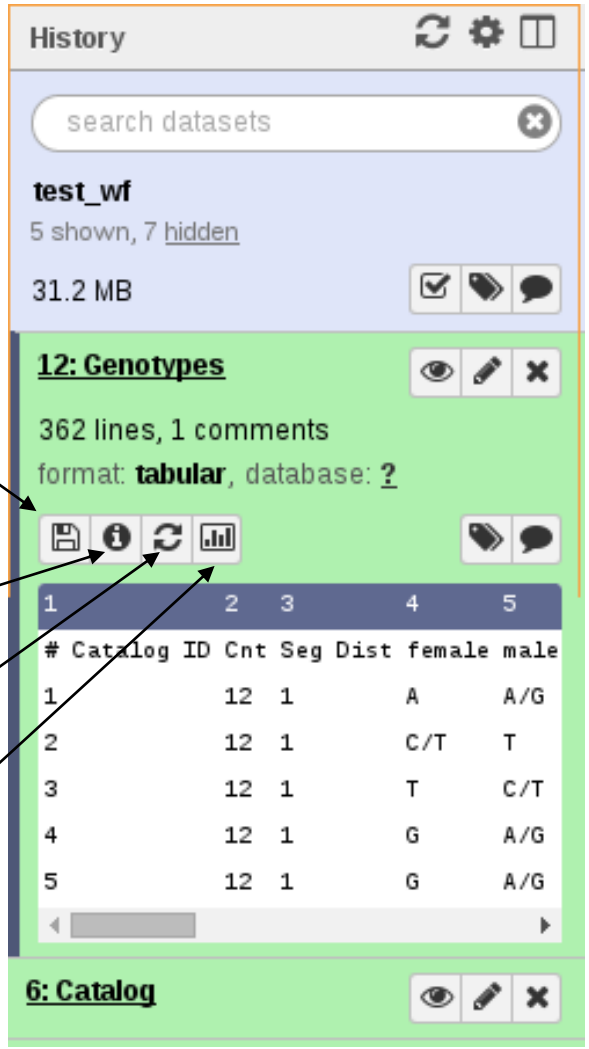
Détails d'un dataset : clic sur le titre

Télécharger le fichier

Infos sur l'origine

Relancer l'outil
avec les mêmes paramètres

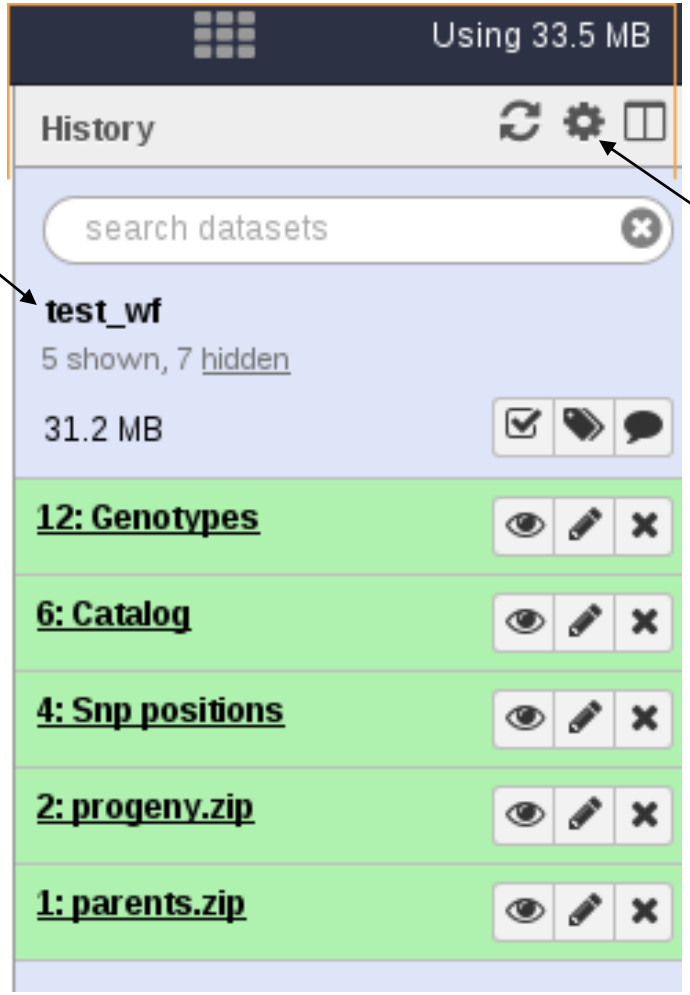
Visualisation



Statistiques/aperçu

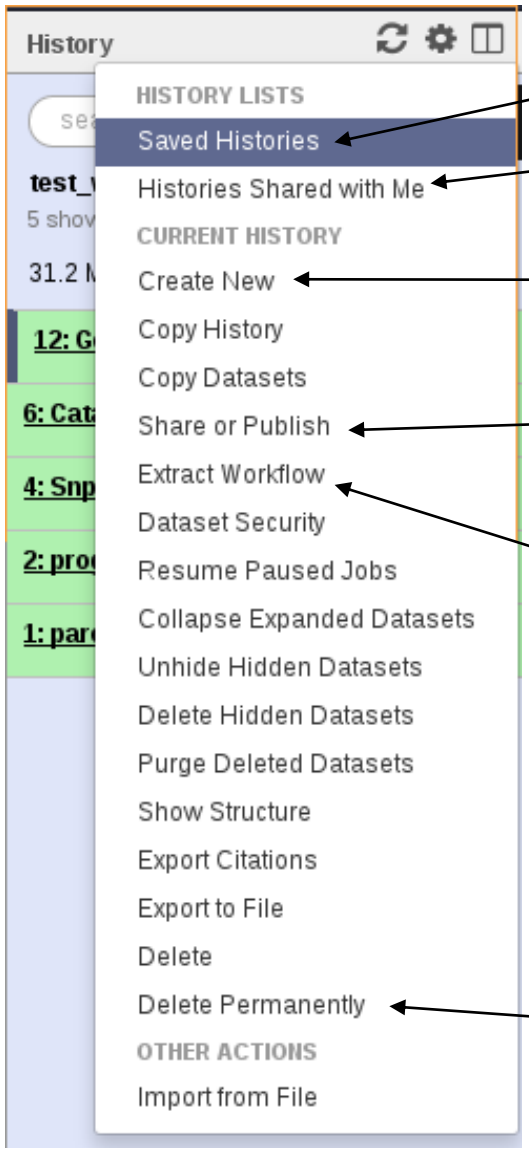
History

Nom de l'historique



Menu de l'historique

History



Liste de nos historiques

Historiques partagés par d'autres

Nouvel historique

Partage de l'historique actuel

Création de workflow

Suppression de l'historique actuel

History

Saved Histories

search history names and tags

Advanced Search

	Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated ↑	Status
☐	test_wf	5	0 Tags		31.2 MB	Mar 27, 2015	~4 days ago	**current history**
☐	Unnamed history	2	0 Tags		4.2 MB	Apr 17, 2014	Apr 17, 2014	
☐	Unn		0 Tags		0 bytes	Mar 06, 2014	Mar 06, 2014	

Switch
View
Share or Publish
Copy
Rename
Delete
Delete Permanently

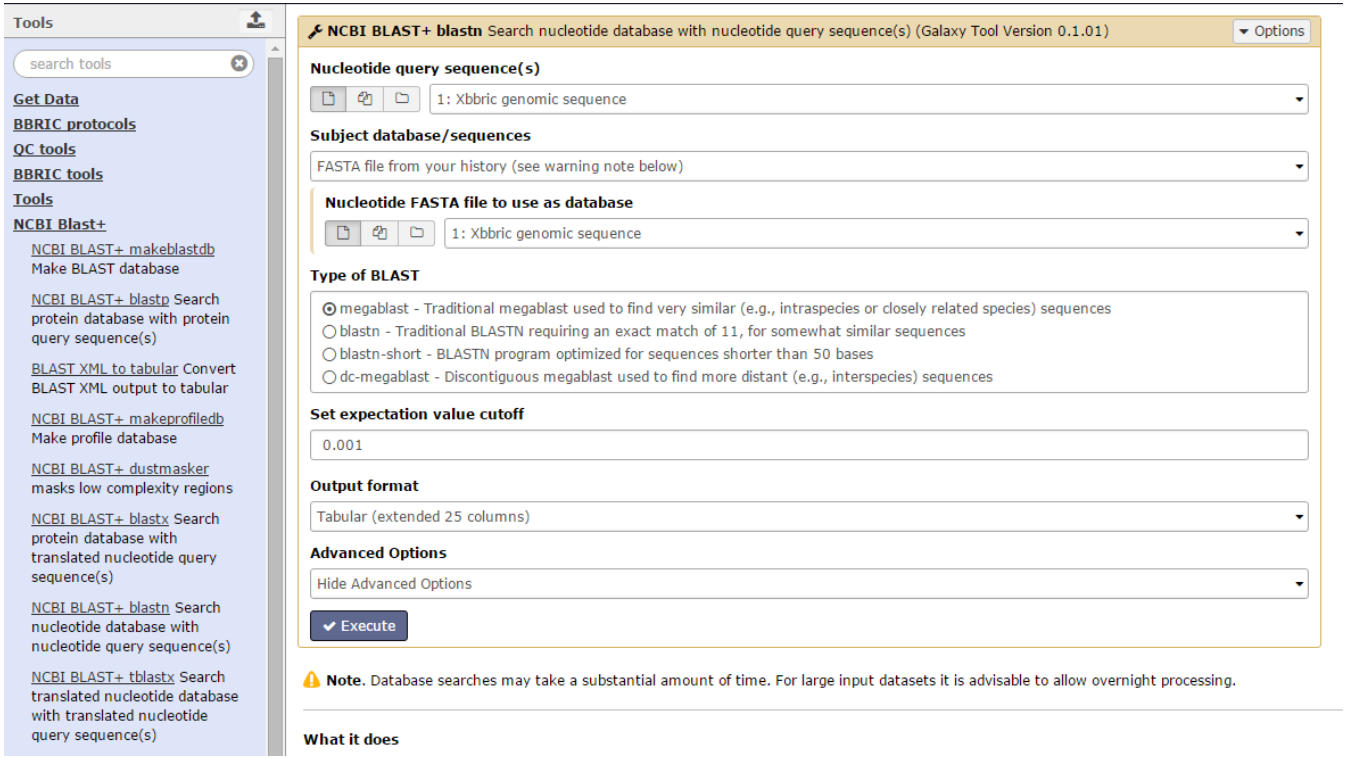
For 0 s — Delete | Delete Permanently | Undelete

Histories tha … than a time period specified by the Galaxy administrator(s) may be permanently deleted.

Tools

Lancement d'un outil

1. Choix d'un outil



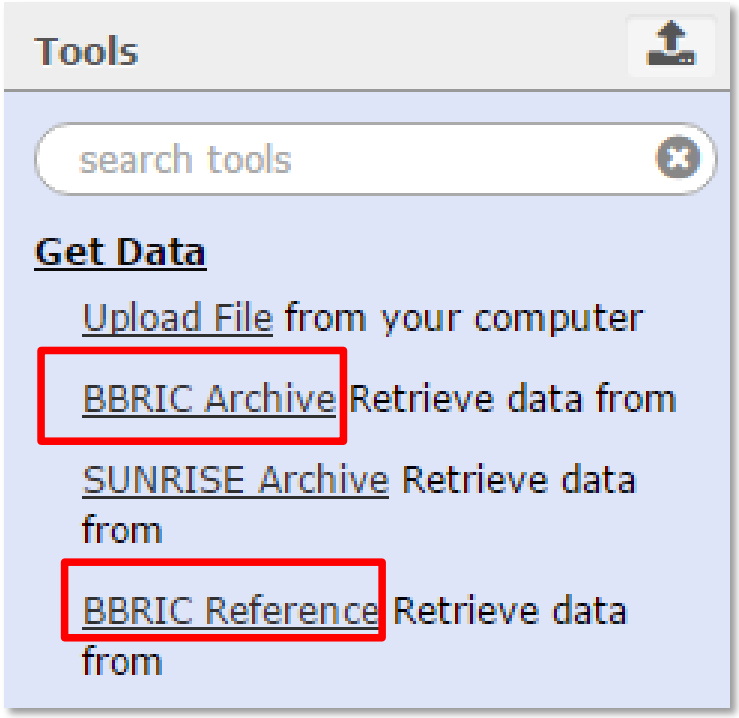
2. Réglages des options

3. Lancement du calcul (crée un nouveau dataset)

- Comment créer les premiers datasets ?
  - **Depuis l'archive BBRIC**
  - **Depuis les Shared data**
  - Upload d'un fichier
  - ...

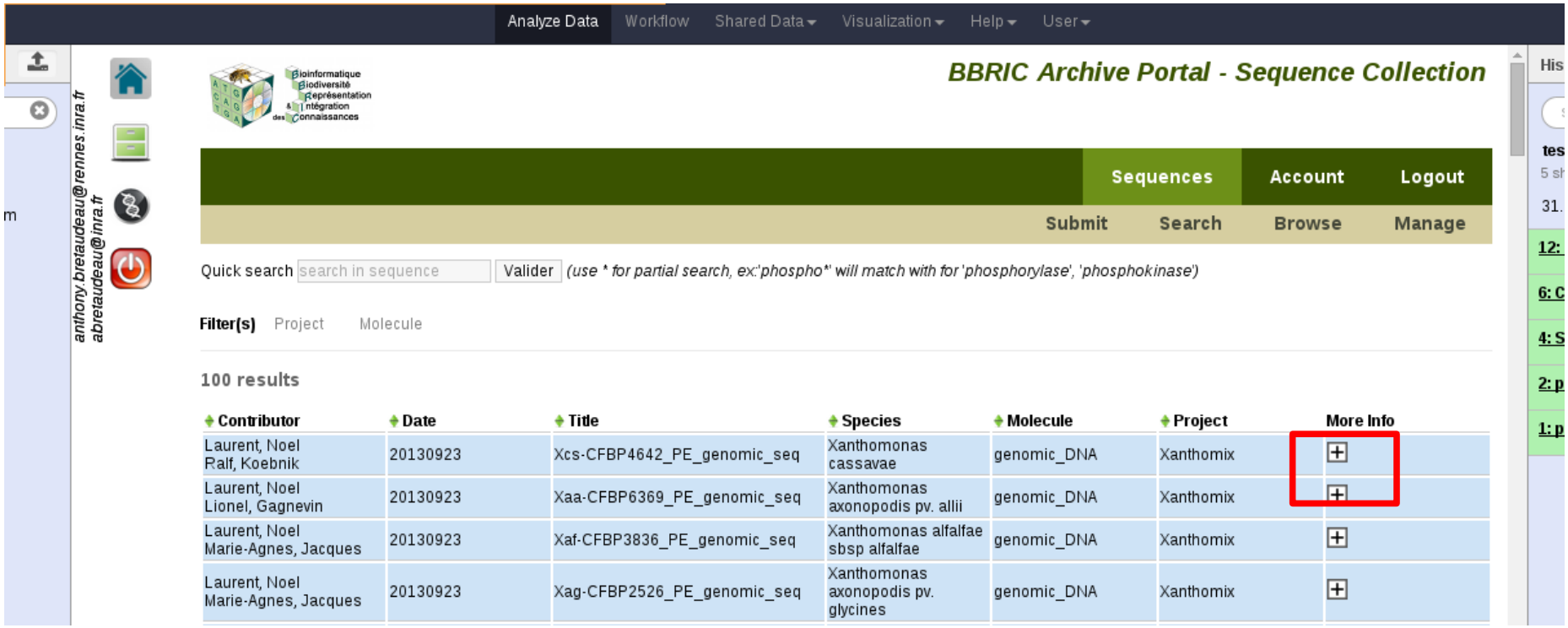
Data

Depuis l'archive : outil « BBRIC Archive »
Depuis les références: outil « BBRIC Reference »



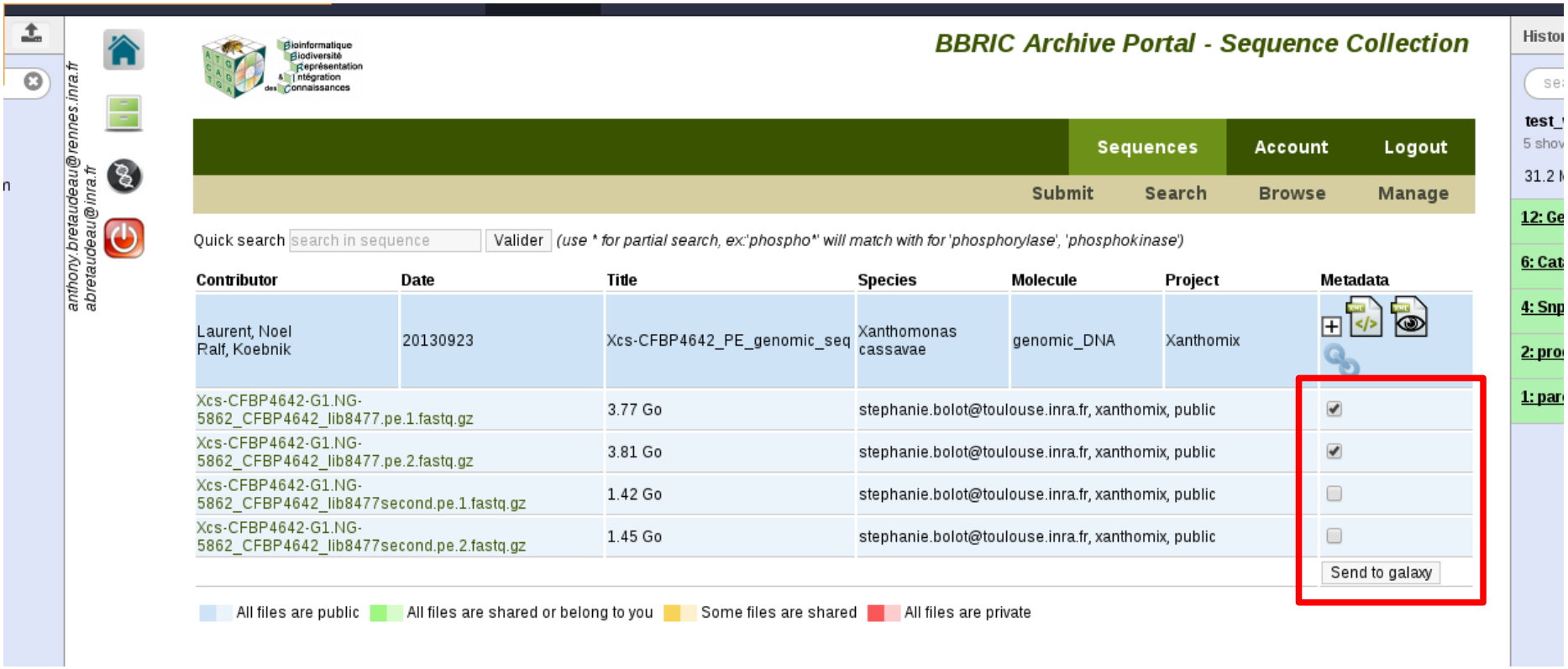
Data

Depuis l'archive : choix d'un jeu de données



Data

Depuis l'archive : import dans galaxy

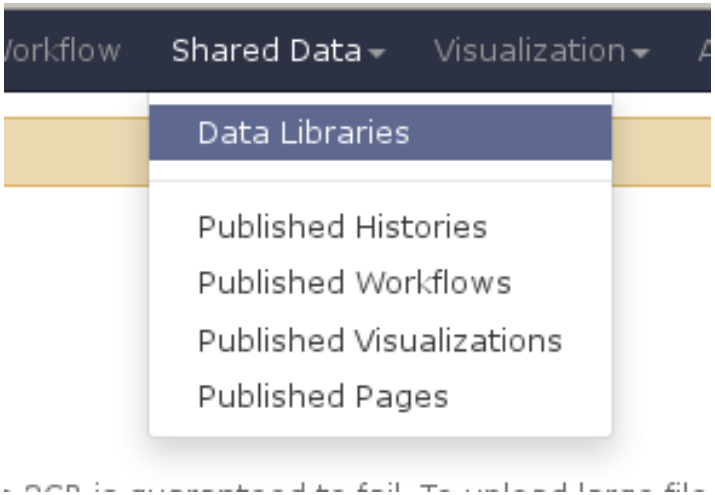


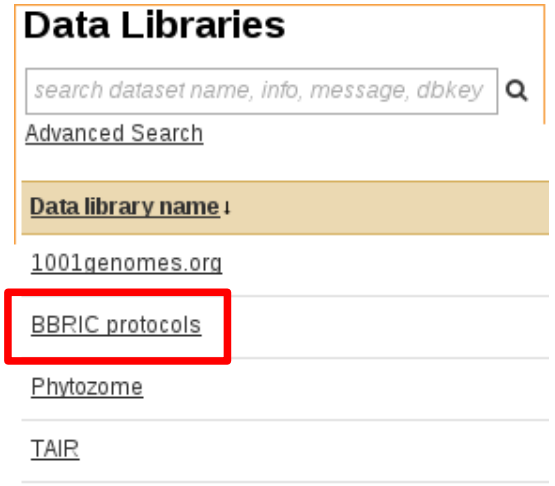
Data

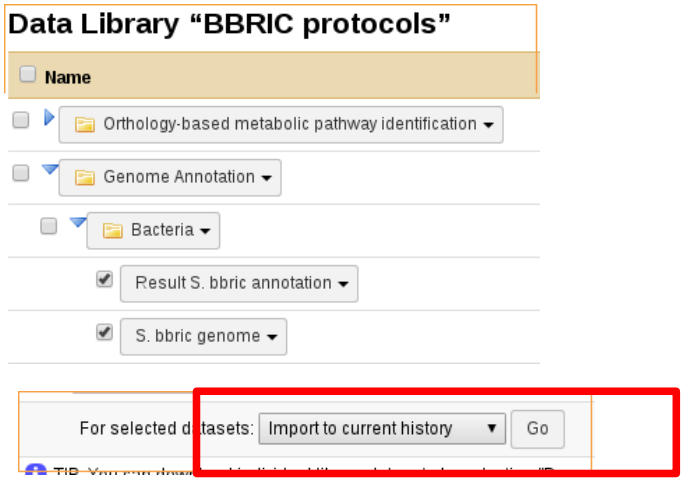
Depuis l'archive : nouveau dataset

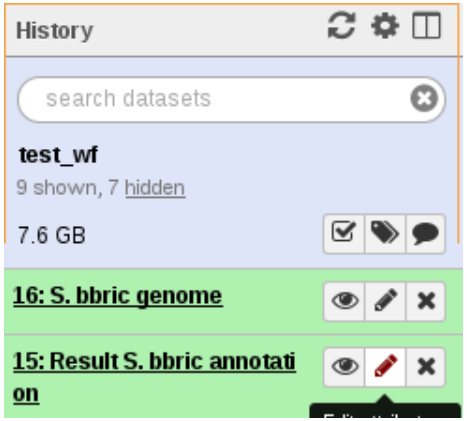
Data

Depuis les data libraries





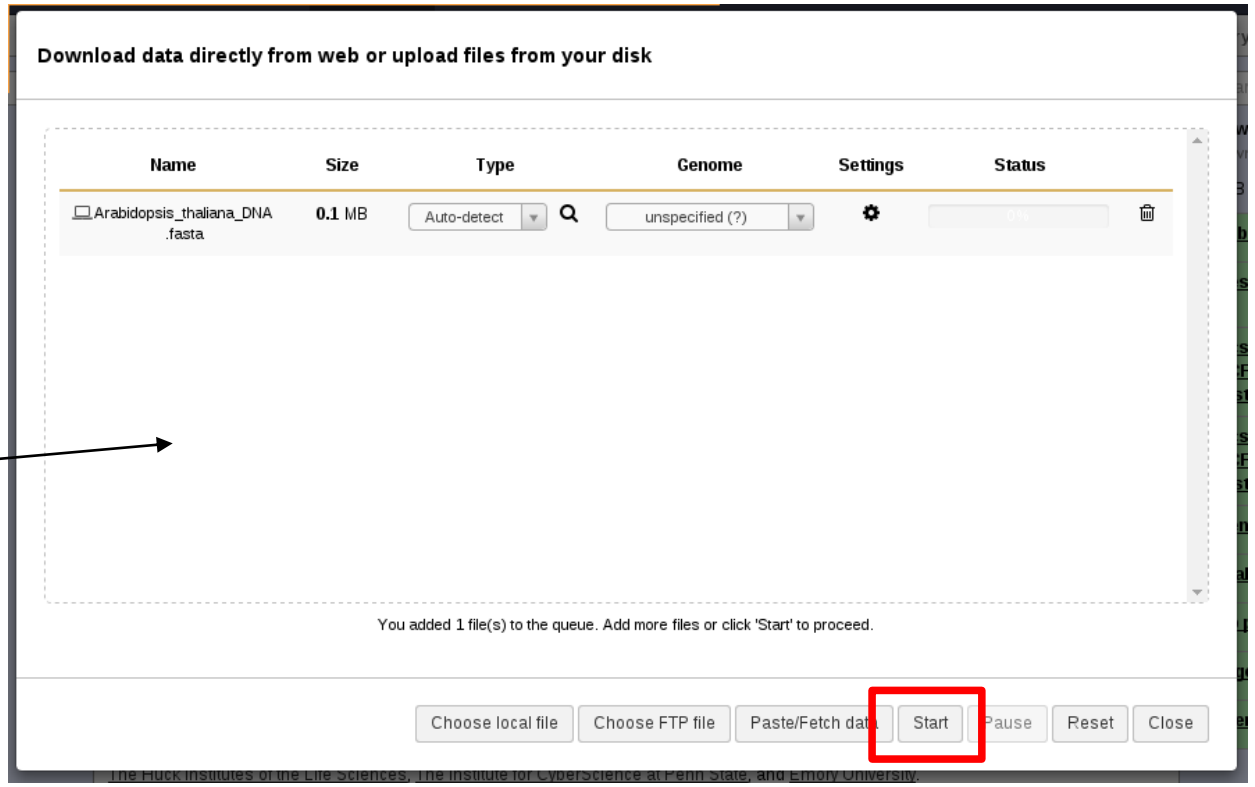


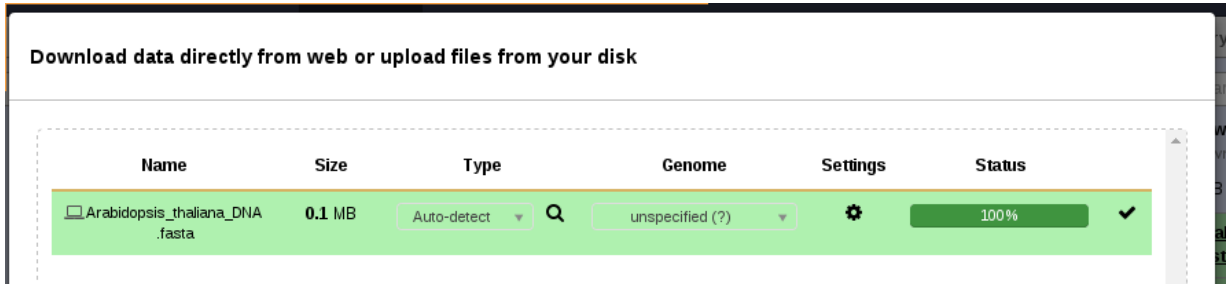


Data

Upload de fichier

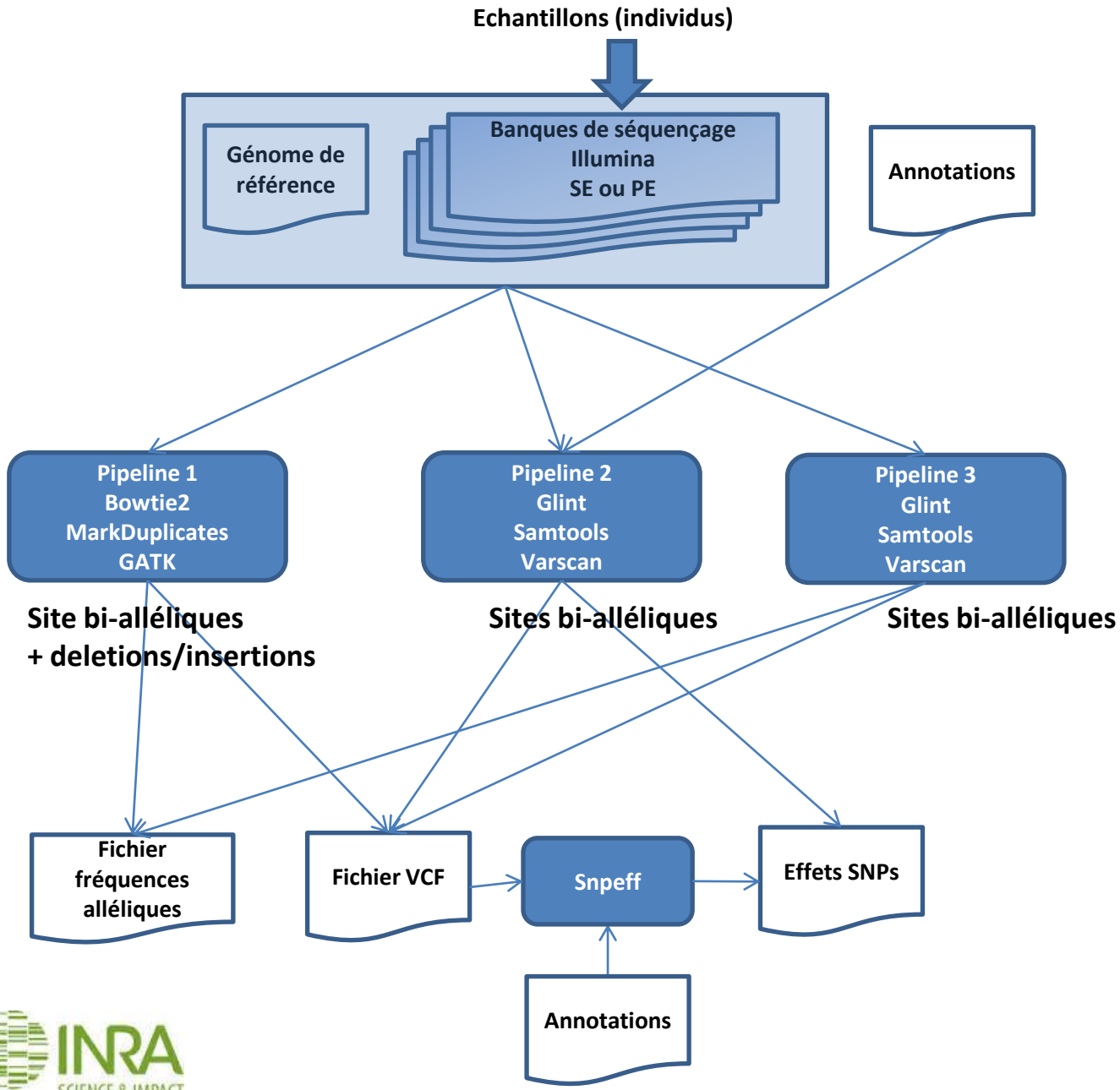
Déposer les fichiers





Session Polymorphisme

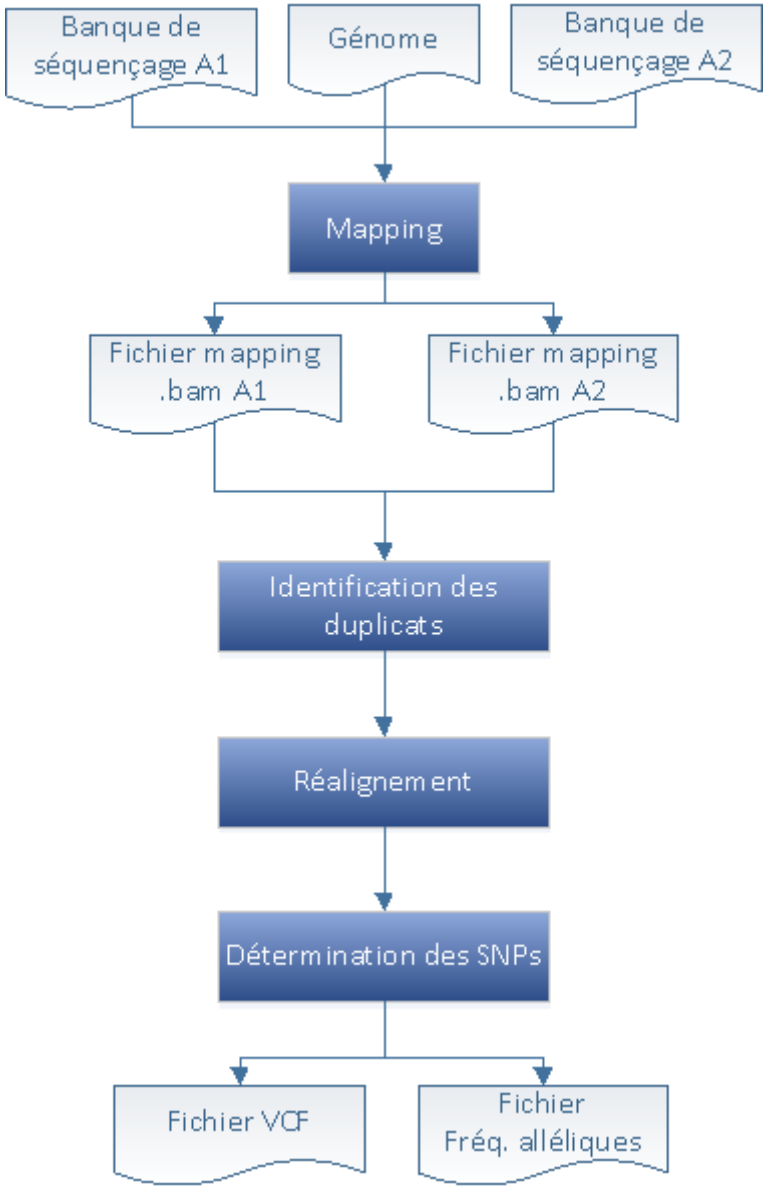
Détection des SNPs : 3 pipelines



Détection des SNPs à partir de données de re-séquencage (GATK)

Expert : Fabrice Legeai (INRA Rennes)

Objectif du pipeline / données d'entrées



Données entrées :

- Banques de séquençage de type Illumina PE ou SE
- Génome

Objectifs du pipeline :

Identifier les SNP sur un ou plusieurs individus par rapport à un génome de référence.

Limitation du pipeline :

Assez long
Assez peu documenté (voir doc GATK)
Ne fonctionne que sur des données Illumina, Single-end ou Pair-ends

Définitions :

Un **SNP** (Single Nucleotide Polymorphism) est un type de polymorphisme de l'ADN constitué par la variation d'une seule paire de base du génome entre individus d'une même espèce.

Individus 1 : …..CATATGCGCATA…..
Individus 2 : …..CATATACGCATA…..

Un **allèle** représente l'une des formes que peut prendre le polymorphisme (G ou A dans le cas précédent).

Un polymorphisme est dit **hétérozygote** chez un individus ou dans une population lorsque plusieurs allèles y sont présents.
Un polymorphisme est dit **homozygote** chez un individus ou dans une population lorsque un unique allèle y est présent.

Un **génotype** est un ensemble des allèles identifiés chez un individu

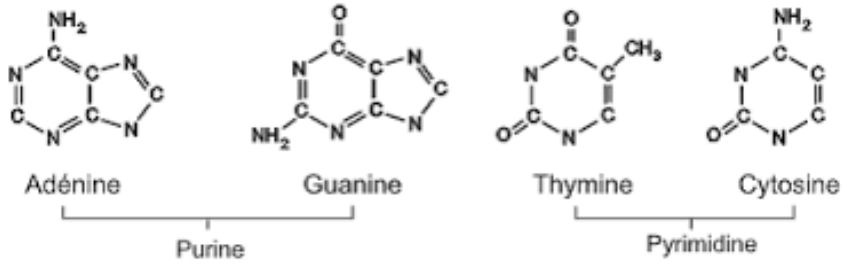
Un **haplotype** est un ensemble d'allèles situés sur un même chromosome. Un individu diploïde comporte 2 haplotypes.

Fréquence :

- Un SNP tous les 100 à 1000 nucléotides. (Génome humain de l'ordre 10^7).
- Distribution au hasard
- Moyenne 10 SNP par gènes mais certain peuvent en avoir 0.

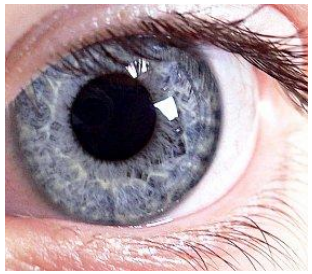
Les différents types de SNP :

- Les **Transitions** : changement d'un purine par une purine (A ↔ G) ou d'une pyrimidine par une pyrimidine (C ↔ T).
- Les **Transversions** : changement d'un purine par une pyrimidine et vice-versa (A ou G ↔ C ou T).



- Les **Insertions** : ajout d'un nucléotide dans la séquence.
- Les **Délétions** : suppression d'un nucléotide dans la séquence.

Lien avec d'un phénotype



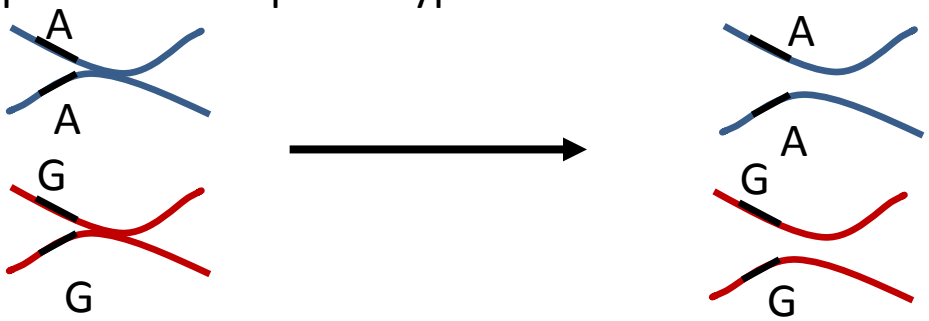
Origine des yeux bleus ?

Analyse des SNPs sur plus de 3000 individus -> le polymorphisme de type SNP sur le gène HERC2 qui régule le gène OCA2 est responsable de la pigmentation des yeux.

A global view of the OCA2-HERC2 region and pigmentation. Donnelly MP & al . Hum Genet. 2012 May;131(5):683-96. doi: 10.1007/s00439-011-1110-x. Epub 2011 Nov 8

Des **méthodes statistiques** dites d'**association** ou de **scan génomique**, permettent d'associer des régions chromosomiques à des phénotypes.

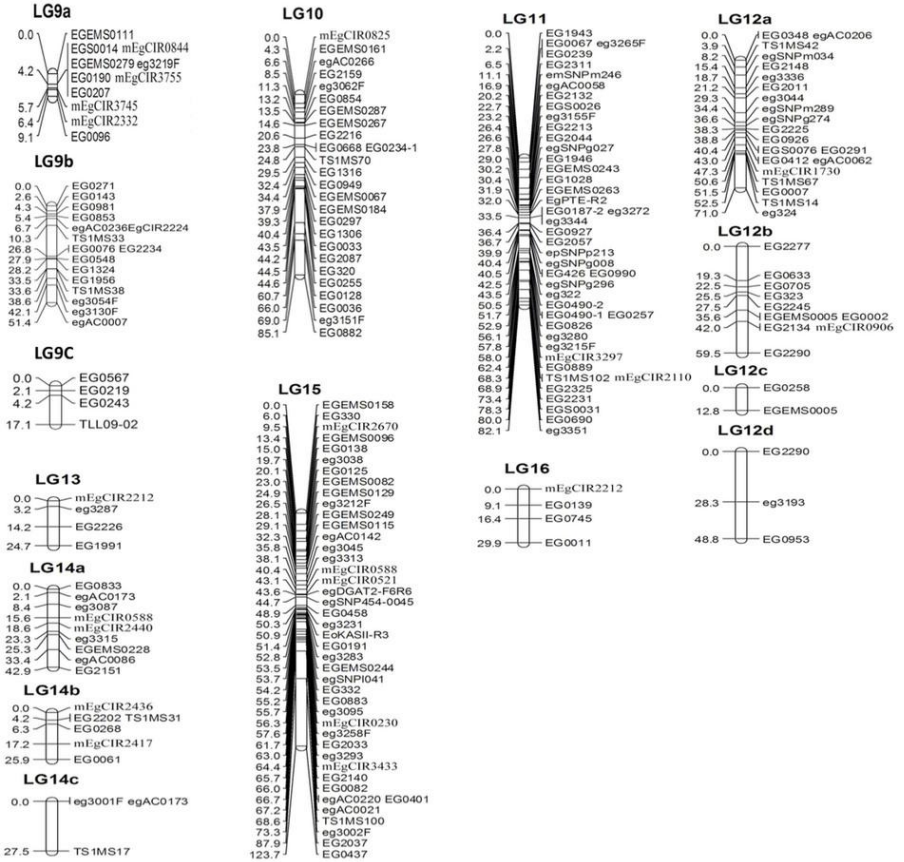
En effet, les allèles localisés à proximité de gènes responsables de phénotypes ségrégent avec les allèles responsables du phénotype.



On cherche alors à identifier des haplotypes caractéristiques dans des populations portant un caractère phénotypique particulier.

Cartes génétiques

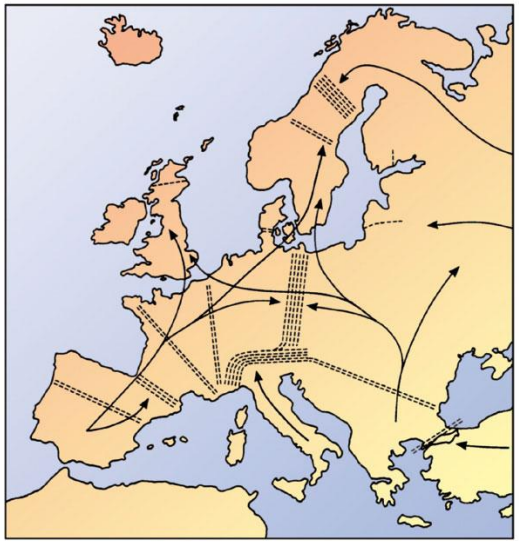
En étudiant la ségrégation des allèles (génotype) des SNPs chez des individus apparentés, des méthodes statistiques permettent de connaître la distance entre ces marqueurs et ainsi d'élaborer des cartes génétiques



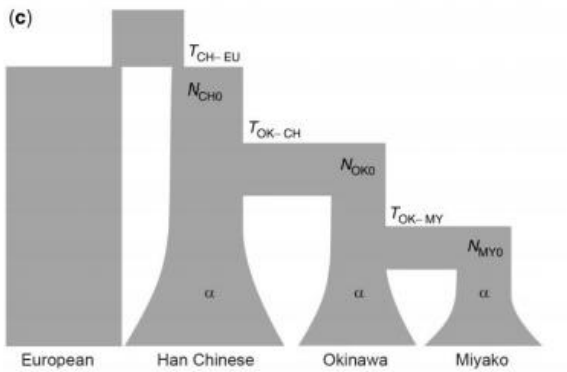
Lee et al.. A consensus linkage map of oil palm and a major QTL for stem height. Sci Rep.

En génomique des populations : Analyse des SNPs pour comprendre le déplacement et l'évolution des populations.

Le polymorphisme ou plutôt **les fréquences des polymorphismes dans les populations** sont des **marqueurs des croisements** (transferts de gènes, des échanges entre les populations), et peuvent être interprétés pour illustrer des traits d'histoire de vie ou des épisodes démographiques.



The genetic legacy of the Quaternary ice ages
Godfrey Hewitt
Nature 405, 907-913(22 June 2000)



Genome-wide SNP analysis reveals population structure and demographic history of the ryukyu islanders in the southern part of the Japanese archipelago. Sato &al. Mol Biol Evol. 2014

Mapping :

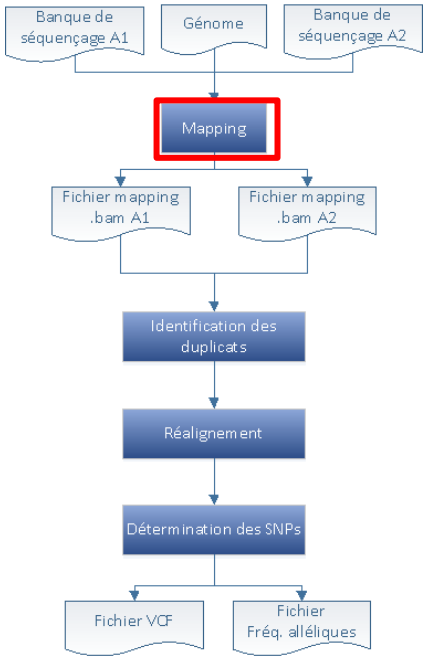
- Positionner les lectures de séquençage sur le génome de référence.

Paramétrage :

- Alignement global End to End

Ref: ---ACTGTCGACTGCGATCTCGACATCGGTGC---
Read: GACTGGGCGATCTCGACTTCG

Ref: ---ACTGTCGACTG - - CGATCTCGACATCGGTGC---
Read: GACTGGGCGATCTCGACTTCG



- Conserve le meilleur alignement

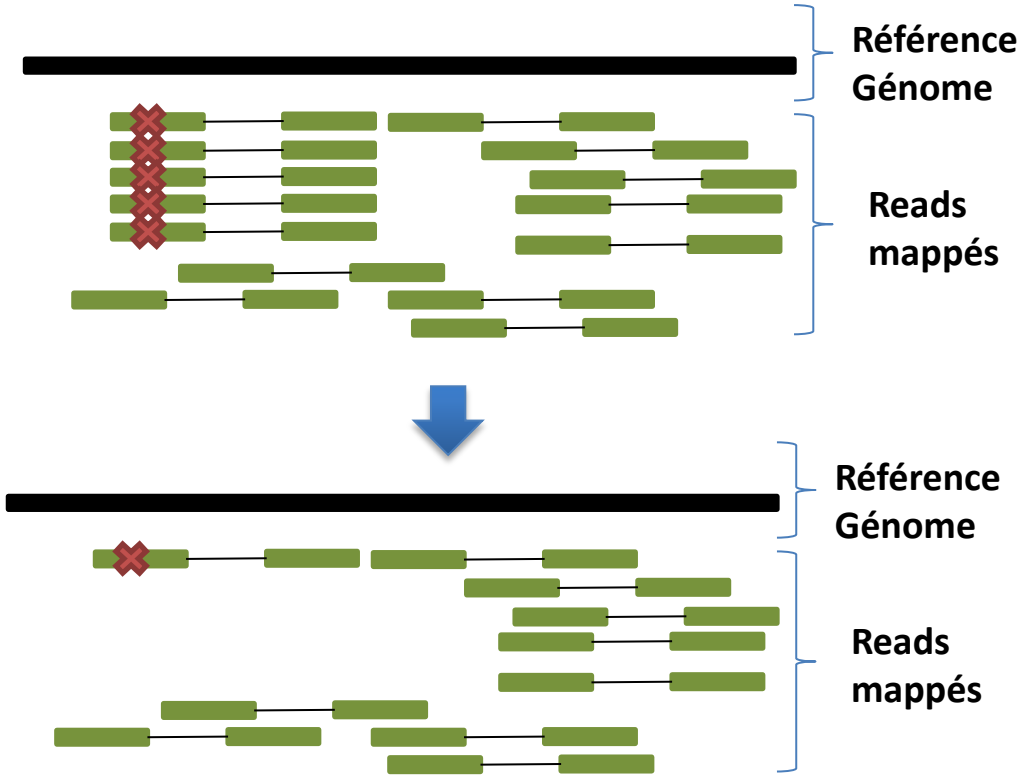
- Si les fragments d'ADN ont été séquencés par paires (Paired-Ends), et que les alignements des 2 lectures sont discordants, c'est à dire que la distance entre les séquences de la paire ou les orientations sont incohérentes, on élimine les alignements

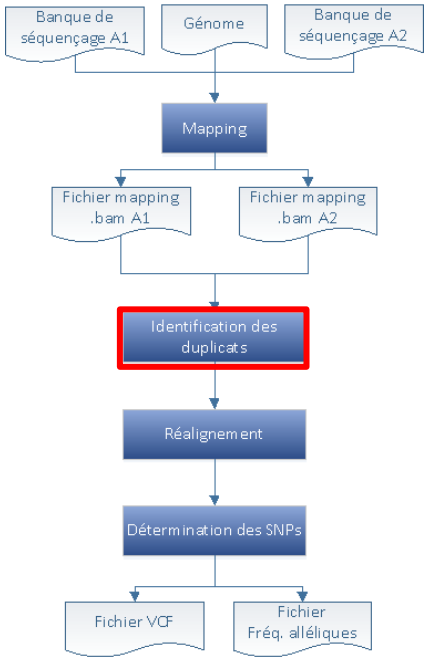
Tool : Bowtie2

Identification des duplicats

-> Réduire le biais lié à la PCR d'amplification réalisée pendant la construction de la banque.

Enlever les duplicats techniques (même couple de séquences lu deux fois) pour éviter la surestimation des profondeurs



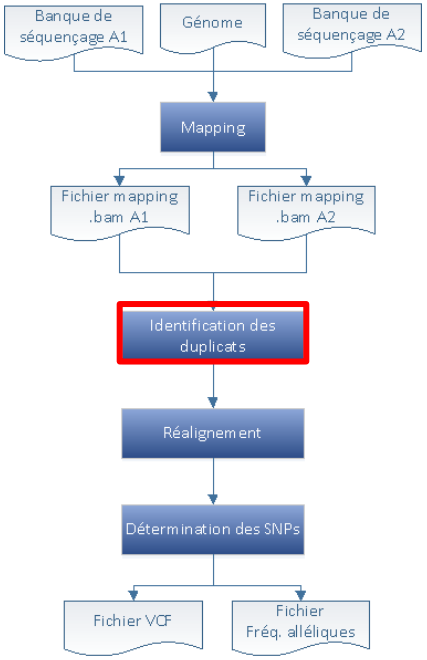


Tool : Mark duplicates (picards Tools)

Réalignement pour l'identification des insertions/délétions

```
Réf:  TAGGCCGCCCGCGCTATGCCGTGCA
R1:         CGC*CGCGCTATGCCGTGC
R2:           CCGCGCTATGCCGTGC
R3:         GCC*GCGCTATGCCGTGC
R4        GCC****GCGCTATGCCGTGC
```

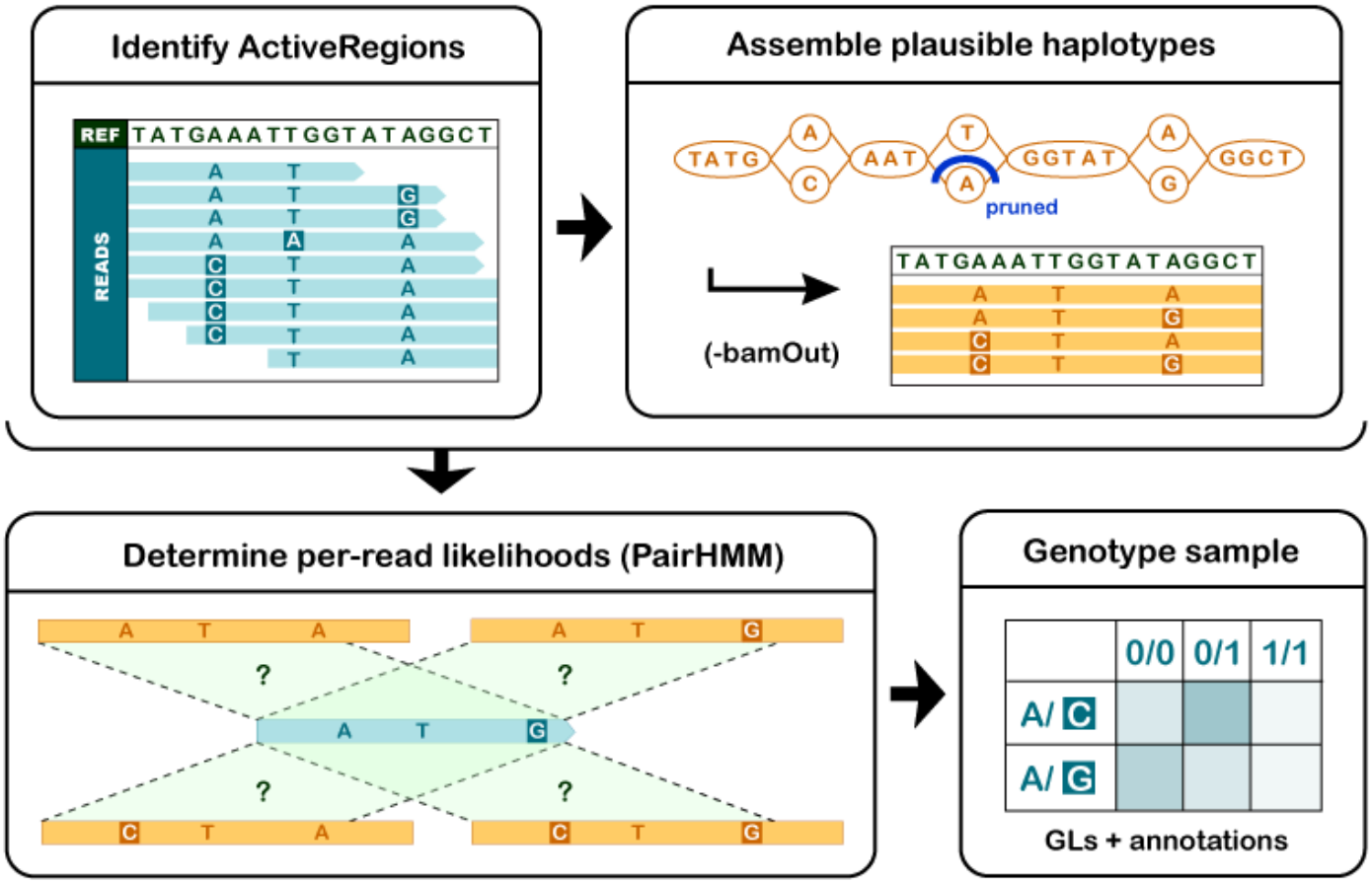
```
Réf:  TAGGCCGCCCGCGCTATGCCGTGCA
R1:         CGCC*GCGCTATGCCGTGC
R2:           CC*GCGCTATGCCGTGC
R3:         GCC*GCGCTATGCCGTGC
R4        GCC*GCGCTATGCCGTGC
```

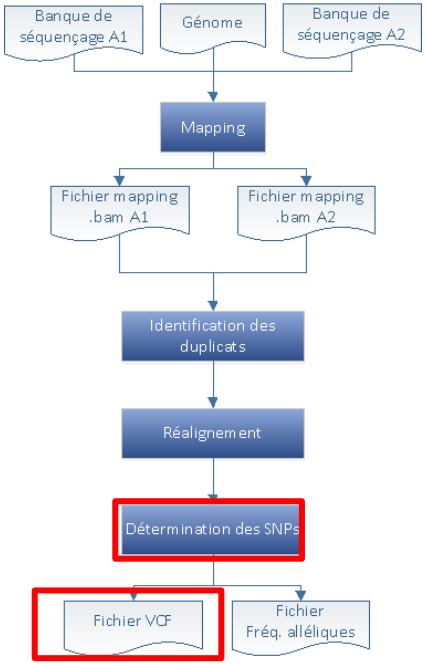


Réalignement pour l'identification des insertions /délétions sur les zones de faible complexité.

Tool : RealignerTargetCreator /IndelRealigner (GATK)

Identification des SNPs : HaplotypeCaller



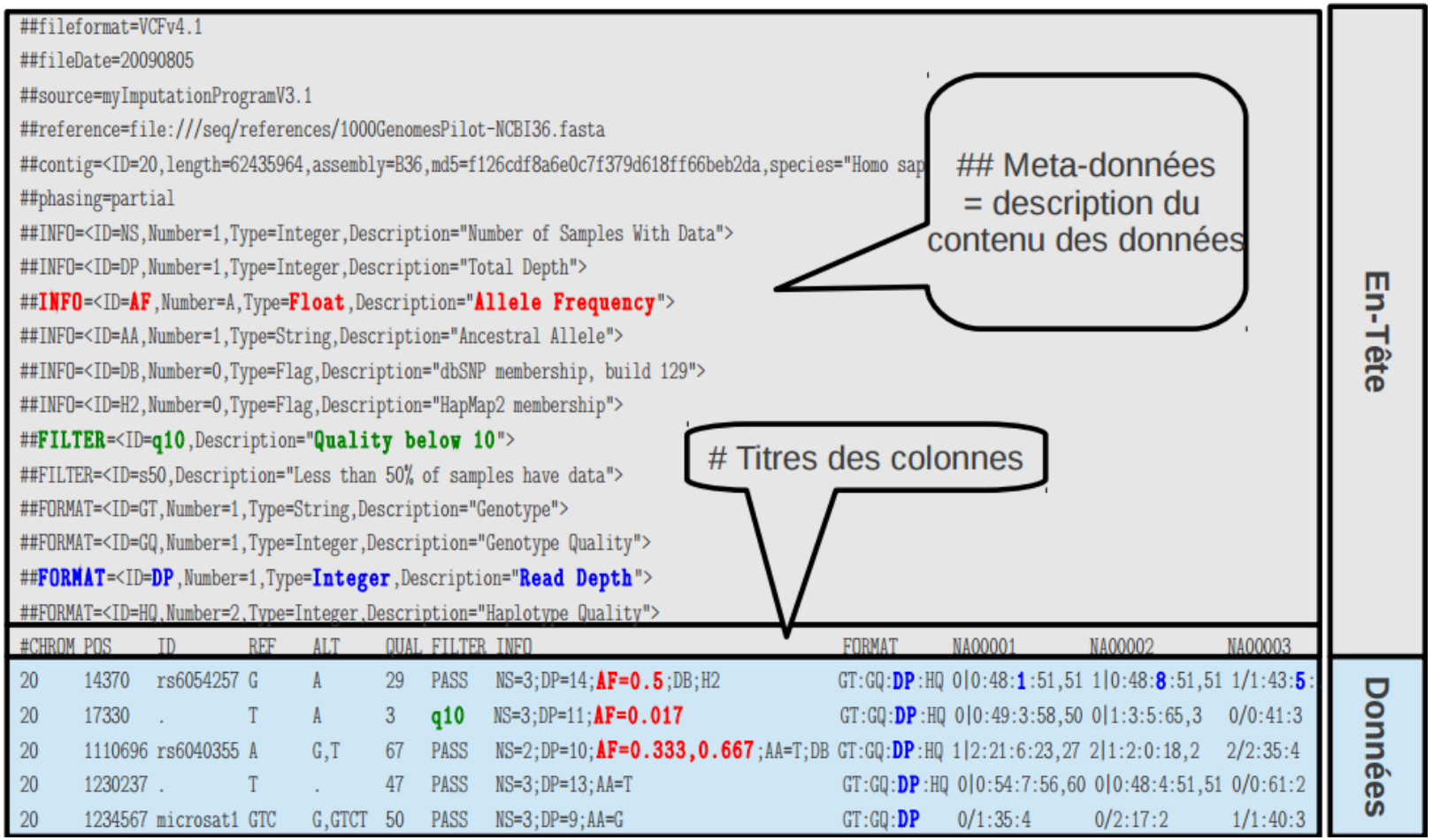


https://www.broadinstitute.org/gatk/guide/article?id=4148

Résultats : Fichier VCF (Variant Call Format)

Fichier Format VCF :

Format de fichier permettant de stocker la liste des variants



Descriptif complet du format : http://samtools.github.io/hts-specs/VCFv4.2.pdf

FICHIER DE SORTIE : FORMAT VCF

Informations sur chaque site								Génotypes: format + n échantillons			
#chr	pos	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00004
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	0/01:41:3
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

- Chromosome
- Position
- Nom
- Allèle sur la référence
- Allèle(s) alternatifs
- **QUAL** Qualité
- **FILTER**
- **INFO**

Informations sur le génotype (site x échantillon)

Critères de sélection ou de filtrage :

- Les informations générales du site (lignes)
- Les échantillons (colonnes)
- Les génotypes (sites x échantillons)

Exemple	**REF**	**ALT**
Substitution	**G**	**A**
Insertion	**C**	**CT**
DéletionCC	**C**	
2 Génotypes	**T**	**A,G**

Score Phred

Score de qualité phred	Probabilité d'une identification incorrecte	Précision de l'identification d'un base
10	1 pour 10	90 %
20	1 pour 100	99 %
30	1 pour 1000	99.9 %
40	1 pour 10000	99.99 %
50	1 pour 100000	99.999 %

FICHIER DE SORTIE : FORMAT VCF

Informations sur chaque site								Génotypes: format + n échantillons			
#chr	pos	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00004
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	0/01:41:3
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

INFO : **Liste de champs + valeurs**
- **champ1=xx:champ2=yy:champ3=zz**
- Exemple ici : NS, DP, AF, AA, DB, H2
  - Signification : cf méta données

```
##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">
```

Les champs ont des formes différentes
- DP (Total depth) contient 1 entier (number=1, type=Integer)
- AF (Allele Frequency) contient une liste de valeurs décimales car il peut y avoir plus d'1 allèle alternatif (number=Array, type=float)
- H2 (Hapmap2 membership) est présent ou absent (number=0, type=flag)

Les sites n'ont pas tous les mêmes champs

FICHIER DE SORTIE : FORMAT VCF

Informations sur chaque site								Génotypes: format + n échantillons			
#chr	pos	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00004
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	0/01:41:3
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

Format des informations des colonnes génotype de tous les échantillons
- **champ1:champ2:champ3**
- Exemple ici : GT, GQ, DP, HQ
  - Signification : cf méta données

Valeurs des génotypes ; site x échantillon

```
##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">
```

Les informations ont des formes différentes

- DP (Read deepth) contient 1 entier (number=1, type=Integer)
- GT (Génotype) contient un 1 mot (number=1, type=string)
- HQ (Haplotype Quality) contient 2 entiers (number=2, type=Integer)

Les sites n'ont pas tous les mêmes champs

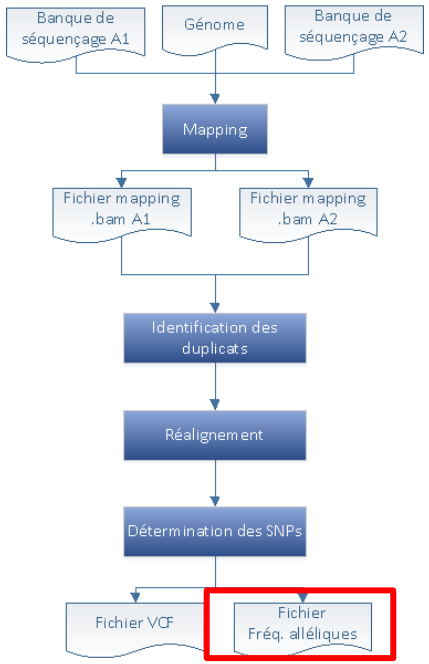
- Comment traiter les données manquantes lors du filtrage ? => option des outils

GT :
Unphased : allèle1 / allèle 2
Phased : allèle1 | allèle 2

Allèle :
0=REF, 1=ALT1, 2=ALT2 ...

FICHIER DES FREQUENCES ALLELIQUES

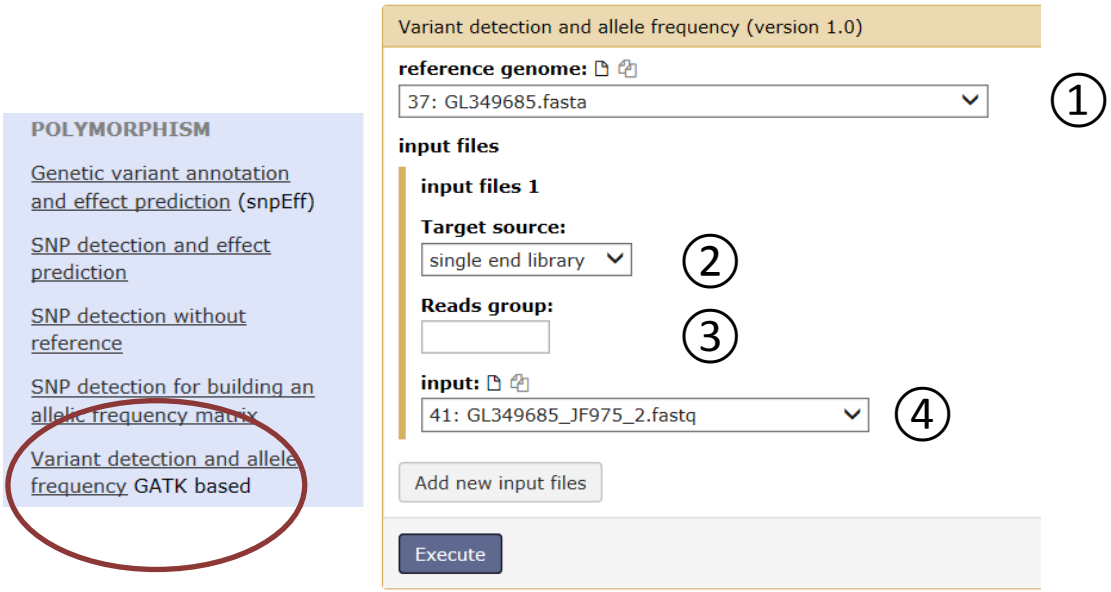
CHROM	POS	REF	ALT	#ALLELES	ind1 #REF	ind1 #TOT	ind2 #REF	ind2 #TOT
GL349685	5347	G	A	1	0	2	0	0
GL349685	5353	G	T	1	0	2	1	1
GL349685	5395	G	T	1	0	2	3	3
GL349685	5397	T	C	1	0	2	3	3
GL349685	5458	T	G	1	0	2	0	8
GL349685	5513	T	TAATCGCC	1	0	2	9	9
GL349685	5522	T	A	1	0	2	9	9
GL349685	5644	GAC	G	1	0	0	0	0
GL349685	6425	T	C	1	0	2	4	4
GL349685	6431	G	GGT	1	0	2	4	4
GL349685	6661	C	T	1	0	10	7	7
GL349685	6672	C	T	1	2	11	7	7



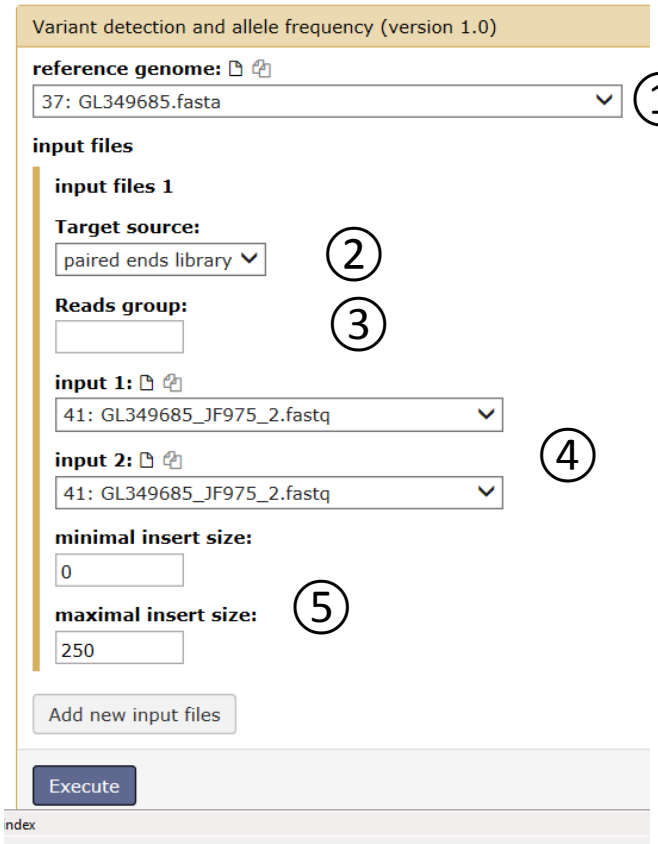
CHROM : scaffold ou chromosome
POS : position sur la référence
REF : la séquence sur la référence
ALT : la séquence alternative
#ALLELES : Nombre d'allèles différents de la référence
ind1 #REF : Nombre de reads correspondant à la référence pour échantillon 1
ind1 #TOT : Nombre de reads total sur ce site pour échantillon 1
ind2 #REF : Nombre de reads correspondant à la référence pour échantillon 2
ind2 #TOT : Nombre de reads total sur ce site pour échantillon 2

Interface dans Galaxy

Banques Single-End



Banques Paire-End



① Génome de référence
② Choix du type de librairie
③ Nom de l'échantillon
④ Choix de la librairie
⑤ taille du fragment séquencé (en cas de séquençage pair-end)

Je veux identifier des SNPs et des indels à partir de données de séquençage et d'un génome de référence

Je vais sur	L'outil permet	L'outil ne permet pas

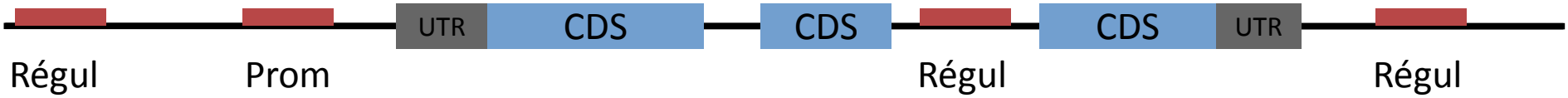
Paramètres clés	Pièges	Formats et fichiers

Détection des SNPs (samtools/VarScan) pour analyse de leurs effets

Responsable et Intervenant principal : Sébastien Carrere
Expert : Ludovic Legrand

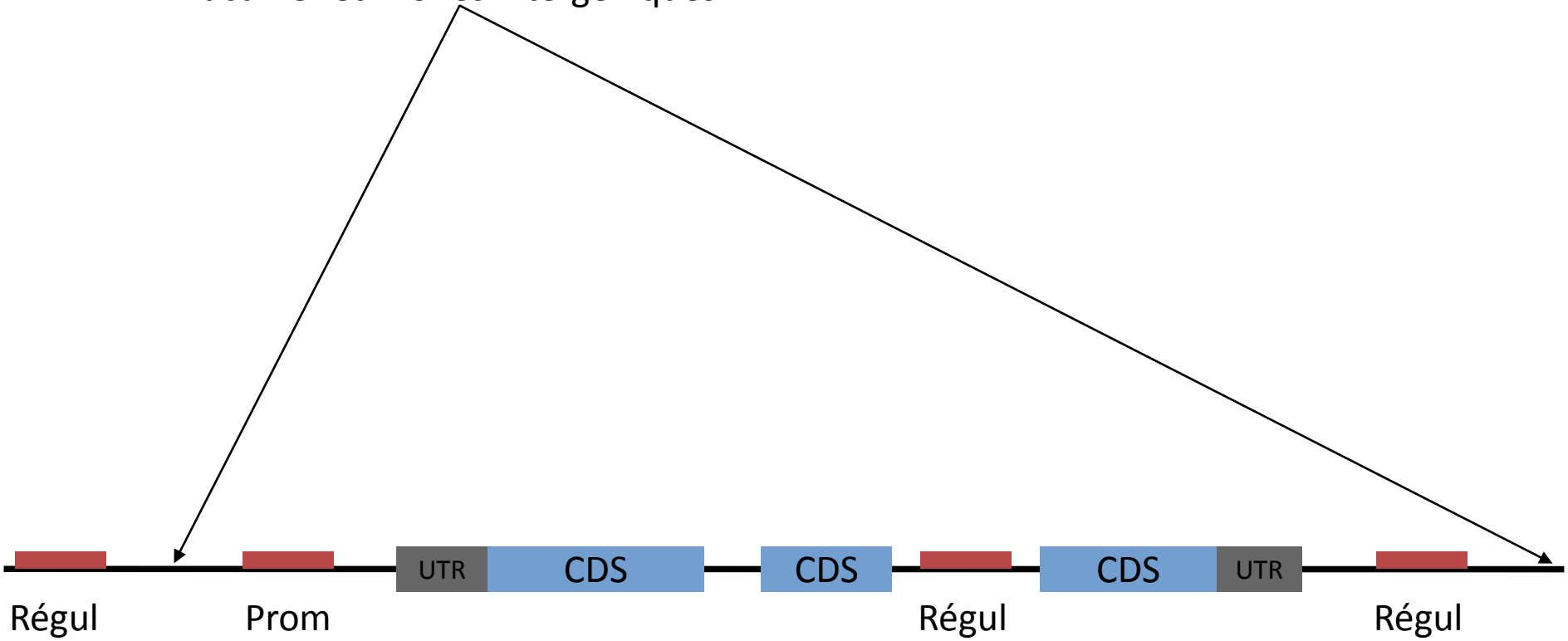
Impact des SNPs

- Impact de chaque SNP variable selon sa position et son environnement :
  - Aucun effet
  - Modification de la protéine produite
  - Effet sur la régulation de l'expression (« eSNP » = expression SNP)
    - •Sites de fixation de facteurs de transcription
    - •Sites d'épissage
    - •Dégradation des mRNA



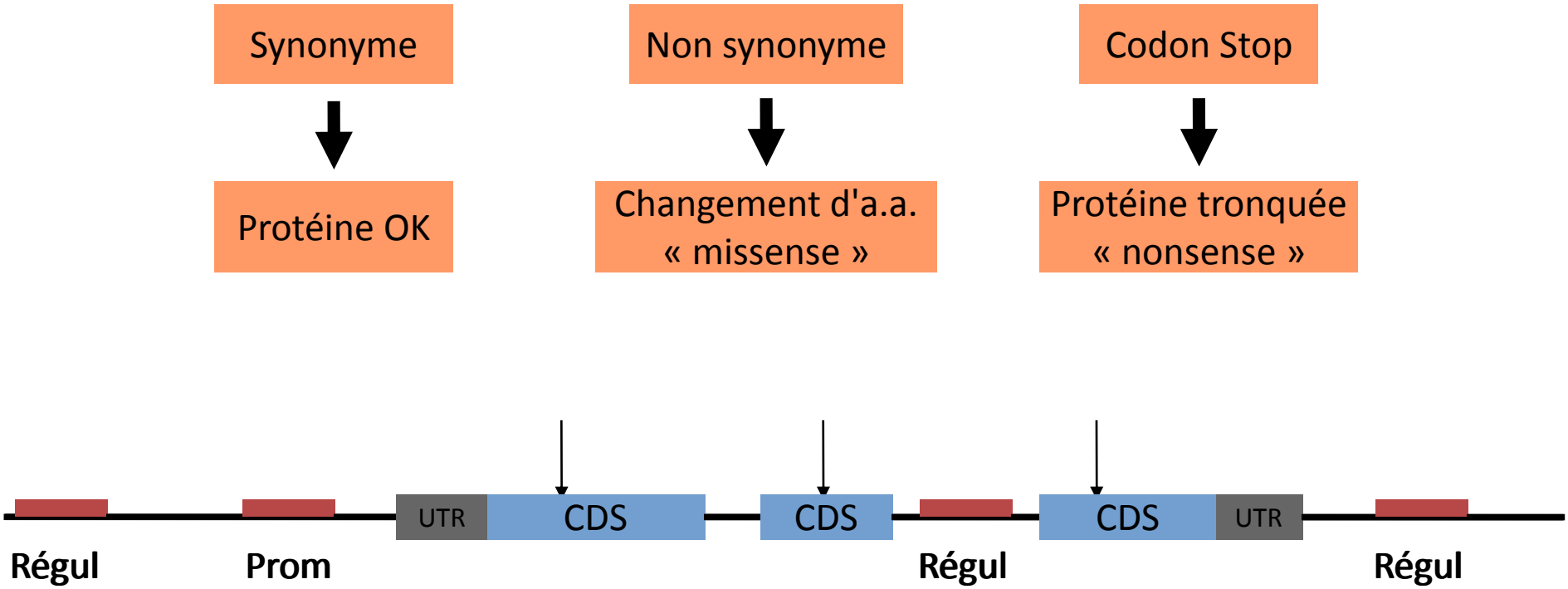
Impact des SNPs

- Impact de chaque SNP variable selon sa position et son environnement :
  - Aucun effet : zones intergéniques



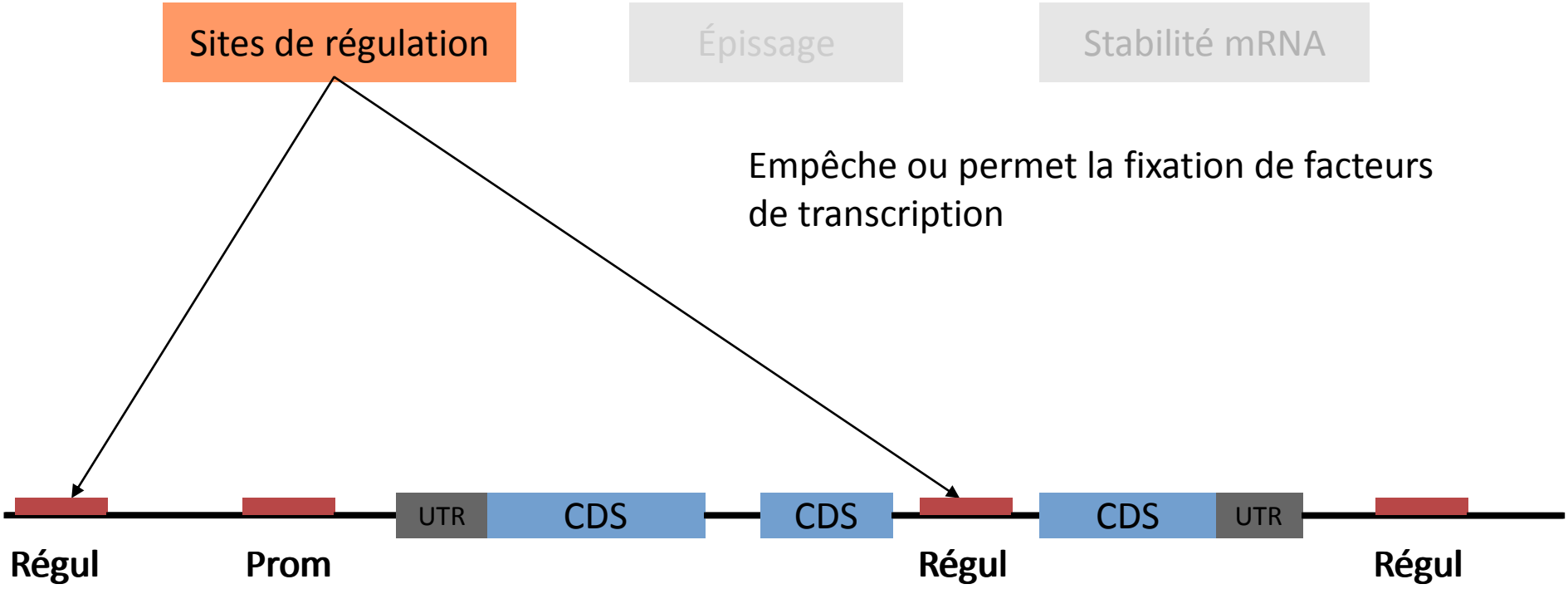
Impact des SNPs

- Impact de chaque SNP variable selon sa position et son environnement :
  - Modification de la protéine produite (nonsense, missense)



Impact des SNPs

- Impact de chaque SNP variable selon sa position et son environnement :
  - Effet sur la régulation de l'expression (« eSNP » = expression SNP)

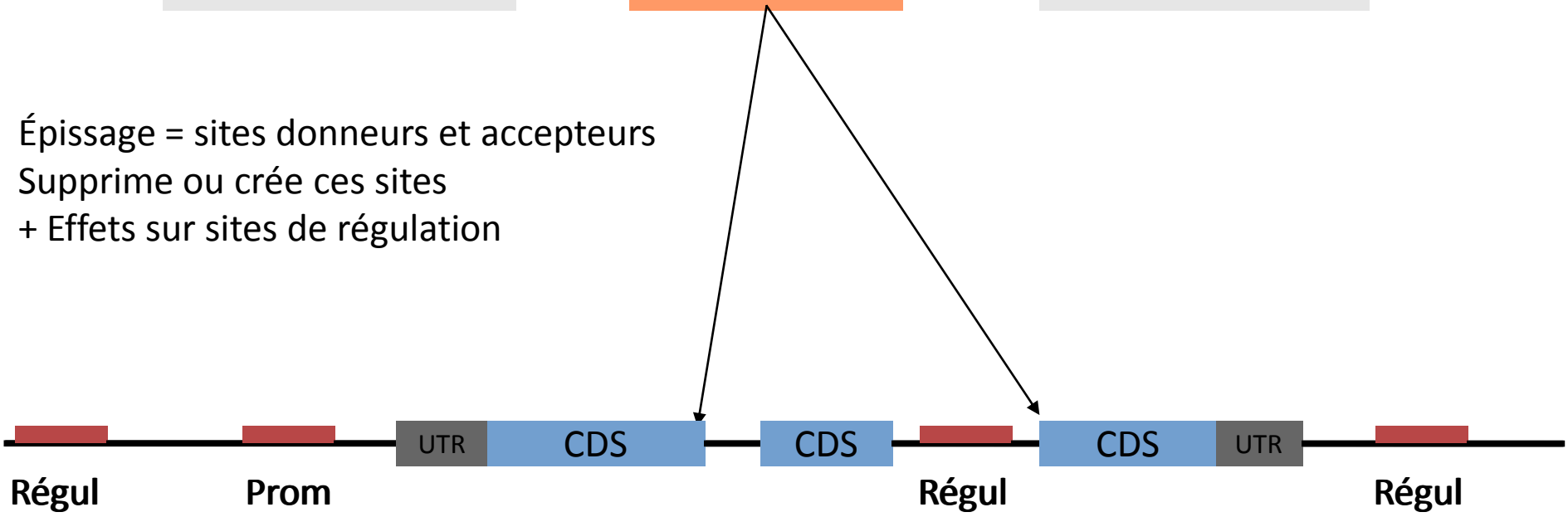


Impact des SNPs

- Impact de chaque SNP variable selon sa position et son environnement :
  - Effet sur la régulation de l'expression (« eSNP » = expression SNP)

Sites de régulation | Épissage | Stabilité mRNA

Épissage = sites donneurs et accepteurs
Supprime ou crée ces sites
+ Effets sur sites de régulation



=> Impact sur les isoformes exprimées

Impact des SNPs

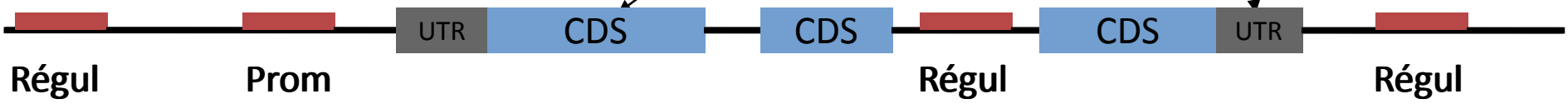
- Impact de chaque SNP variable selon sa position et son environnement :
  - Effet sur la régulation de l'expression (« eSNP » = expression SNP)

Sites de régulation

Épissage

Stabilité mRNA

Durée de vie des mRNA
Liaison à des RNA-binding proteins

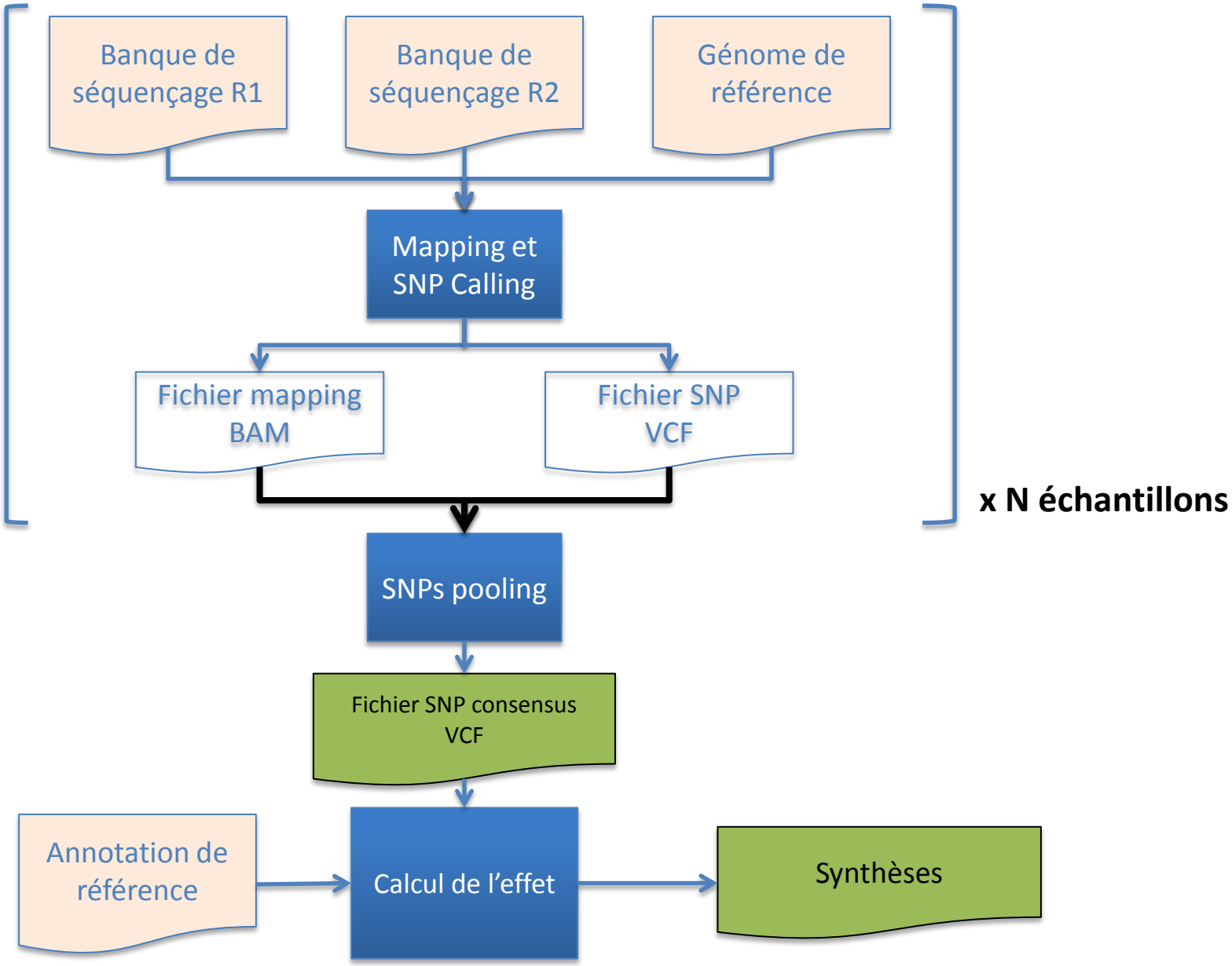


=> Impact sur la durée de vie des mRNA => expression

Le workflow

- Entrées:
  - Génome de référence connu et annotation structurale disponible
  - Re-séquençage du génome de plusieurs individus : Illumina pairé (ou single)
- Etapes:
  - Mapping
  - Calling de SNPs pour chaque échantillon
  - Génotypage de tous les individus pour tous les SNPs
  - Prédiction des effets induits par ces SNPs
- But :
  - Trouver des SNPs pouvant expliquer les différences de phénotypes observées
- Limites :
  - Importance de l'annotation de référence
  - Effets cachés des SNPs : SNP intergénique ou synonyme peut avoir un fort impact

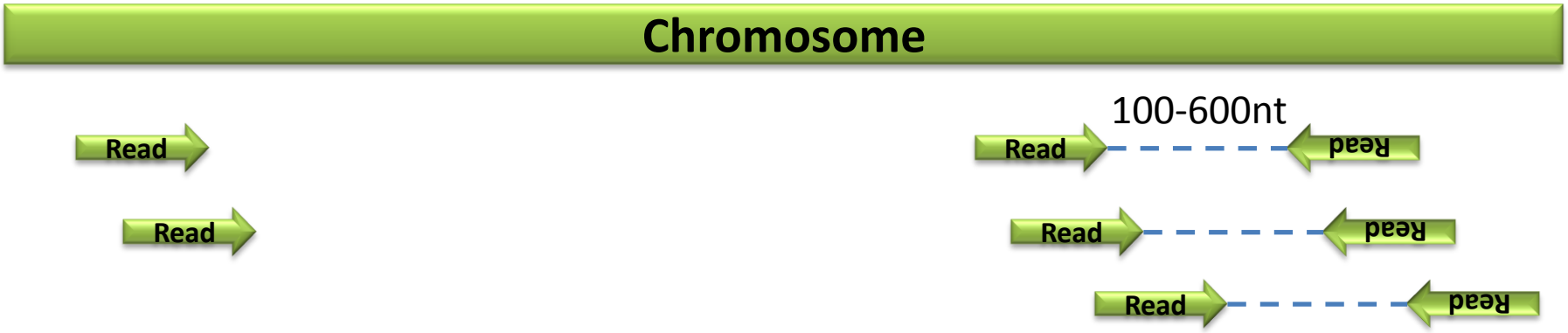
Fonctionnement du pipeline



Fonctionnement du pipeline: mapping et SNP calling

Mapping

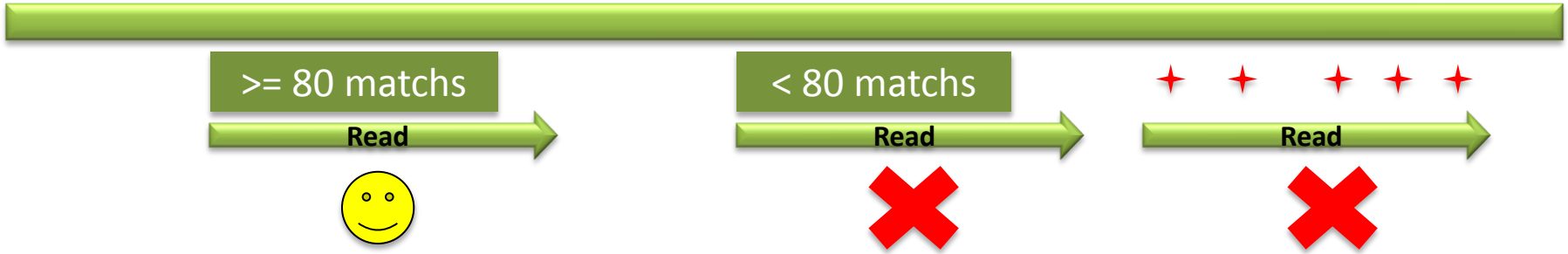
Par échantillon



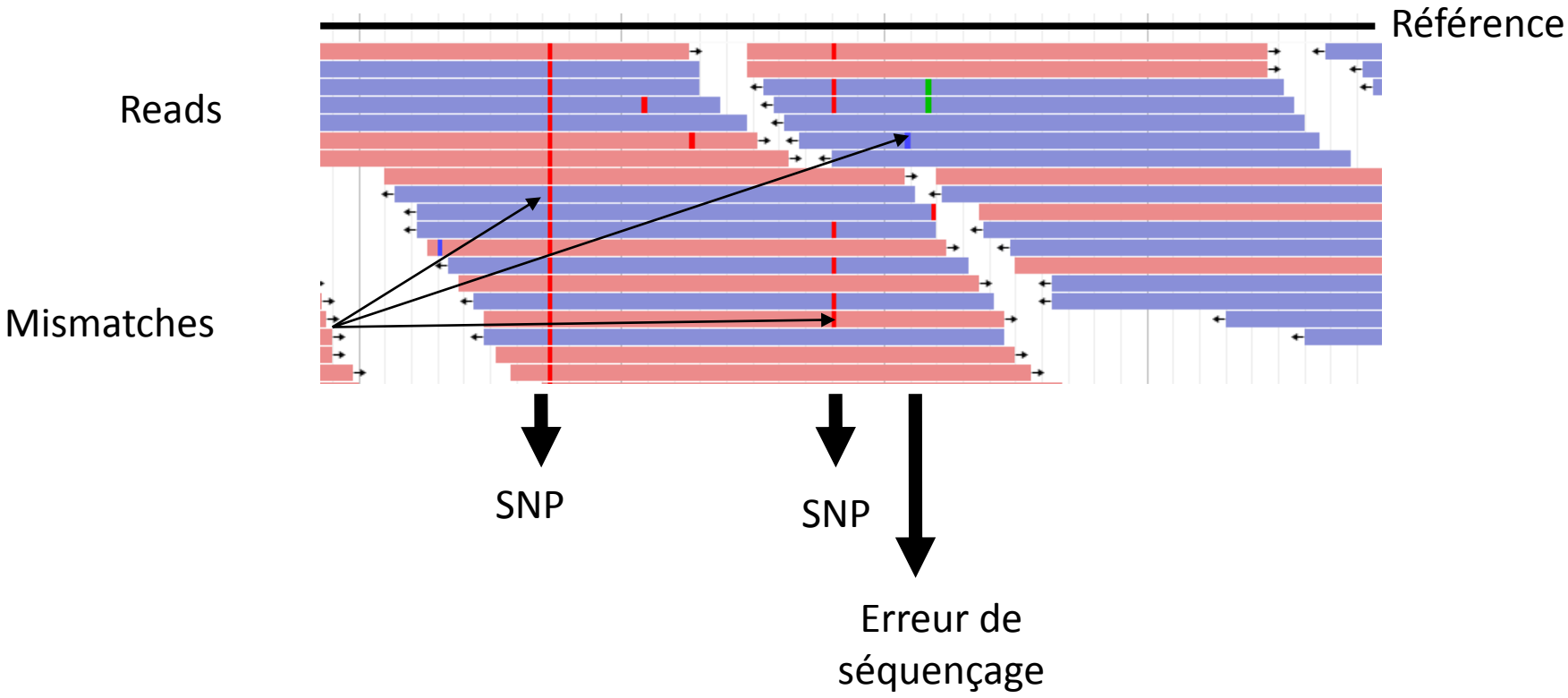
Librairie Single-end

Librairie Paired-end

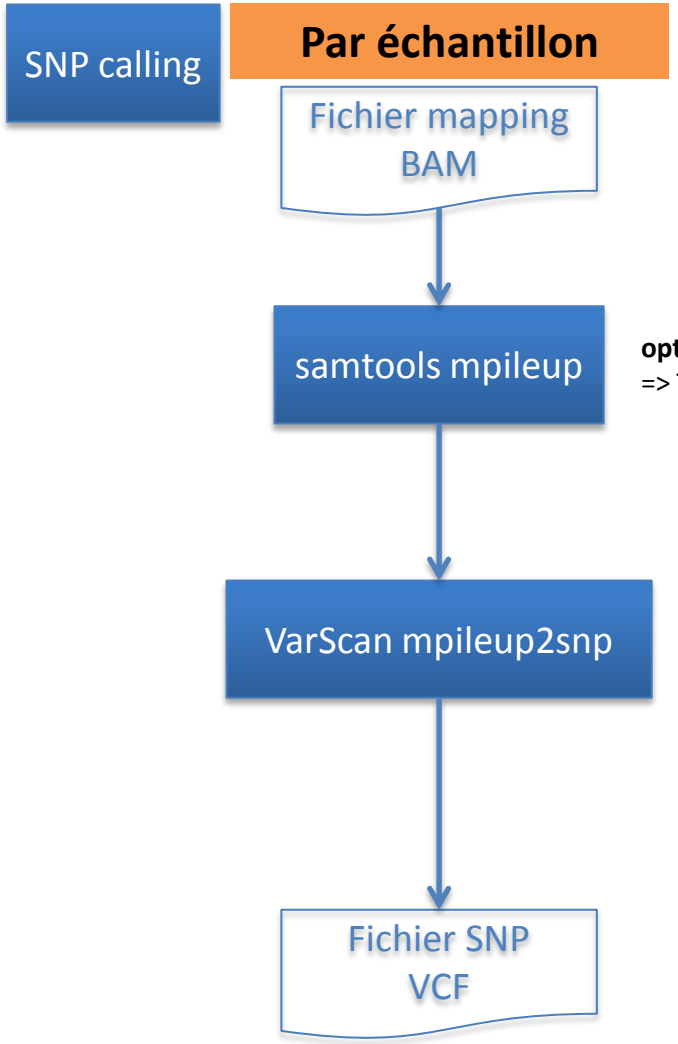
Paramètres par défaut de glint: -lmin=80nt –mmis=4



SNP calling



Fonctionnement du pipeline: mapping et SNP calling



option -B: désactive le calcul du BAQ (Base Alignment Quality)
=> Trop strict, perte de SNPs vrais positifs

options:

Minimum position coverage
Profondeur minimum en reads à une position donnée pour faire le calling
valeur par défaut du pipeline: 20

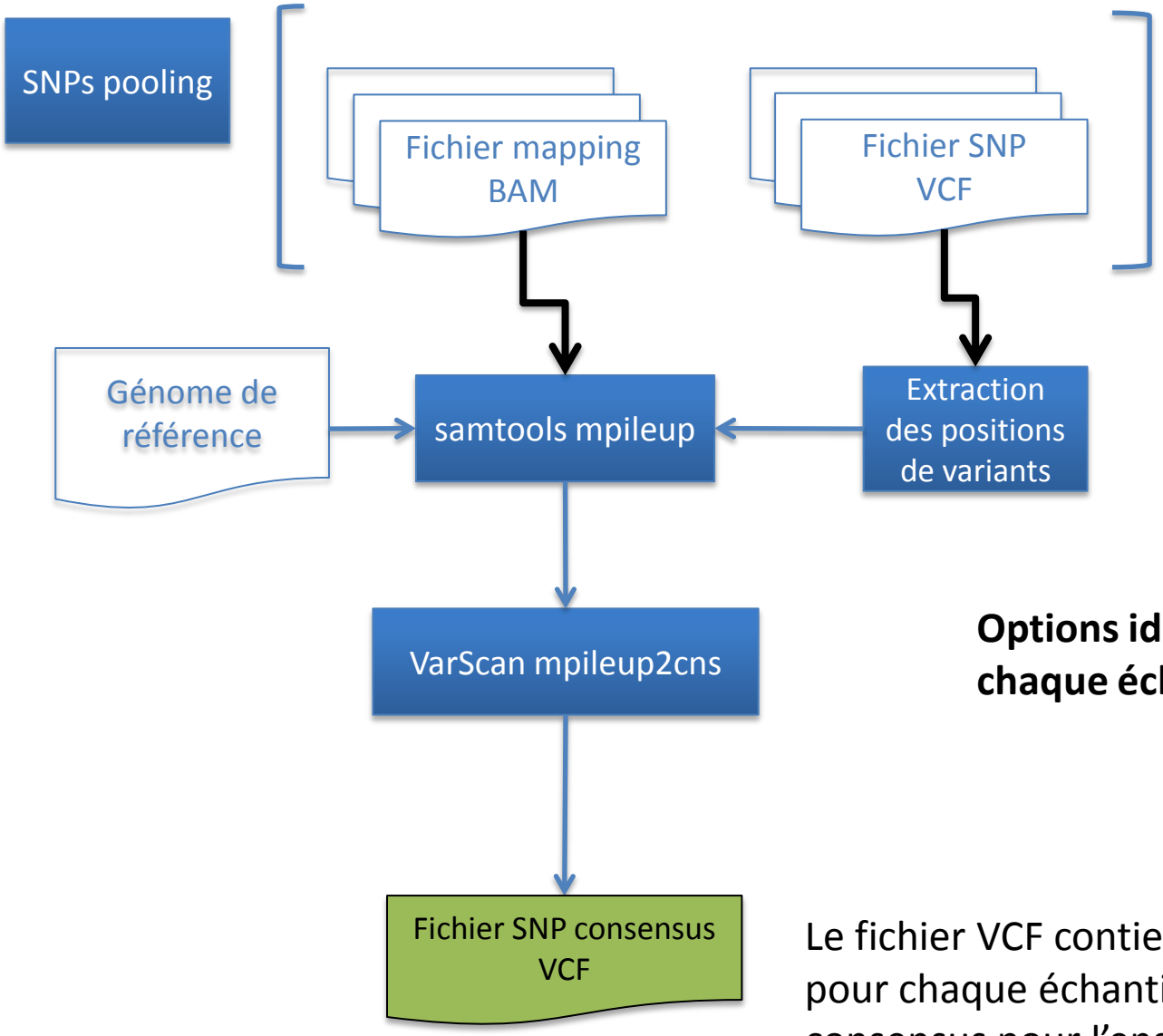
Minimum variant coverage
Nombre minimum de reads supportant un variant pour faire le calling
valeur par défaut du pipeline: 10

Minimum variant frequency
Fréquence allélique minimum du variant pour le conserver
valeur par défaut du pipeline: 0,2

Minimum homozygous frequency
Fréquence allélique minimum pour attribuer un génotype homozygote
valeur par défaut du pipeline: 0,75

http://samtools.sourceforge.net/mpileup.shtml
http://varscan.sourceforge.net/using-varscan.html

Fonctionnement du pipeline: SNP pooling



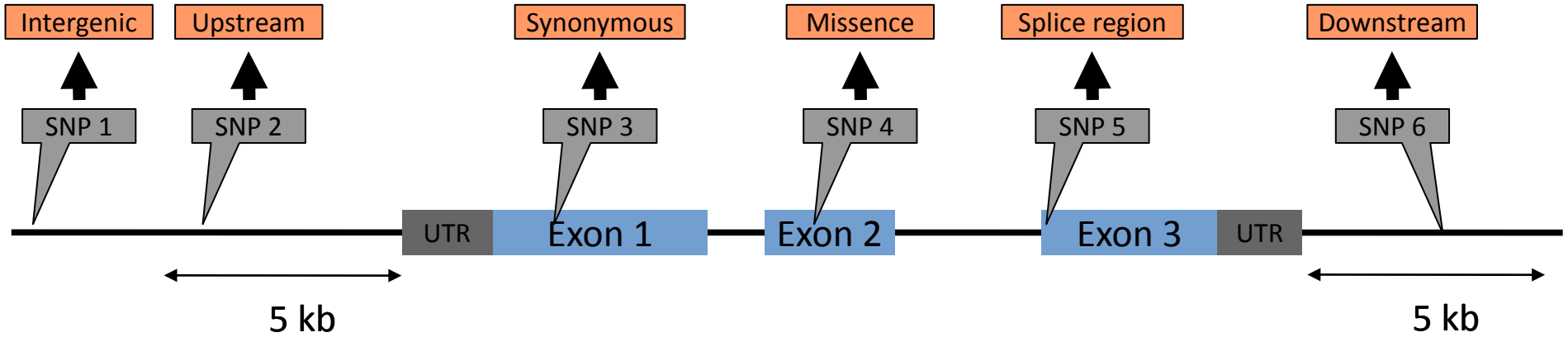
Options identiques au call sur chaque échantillon

Le fichier VCF contient les variants et génotypes pour chaque échantillon ainsi que le génotype consensus pour l'ensemble des échantillons.

SNPeff

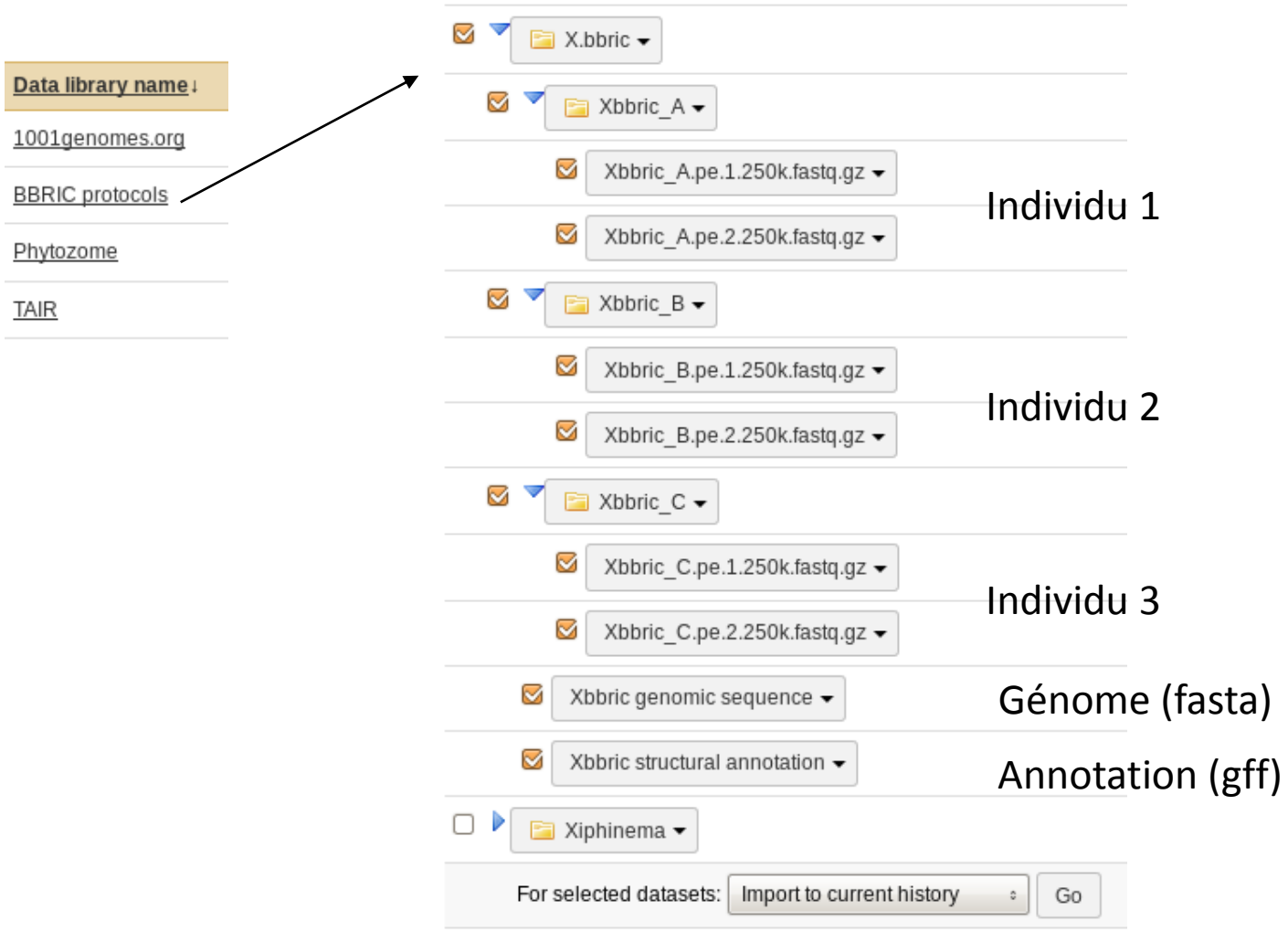
- Prédiction des effets des SNPs avec SNPEff
  - Se base sur une annotation de référence
  - Résultat
    - •VCF annoté
    - •Rapport HTML
    - •Tableau des effets par gène

SNPEff déduit un effet probable en fonction de la position du SNP par rapport à l'annotation



Galaxy : import des données

TP



Galaxy : Formulaire

POLYMORPHISM

Genetic variant annotation and effect prediction (snpEff)

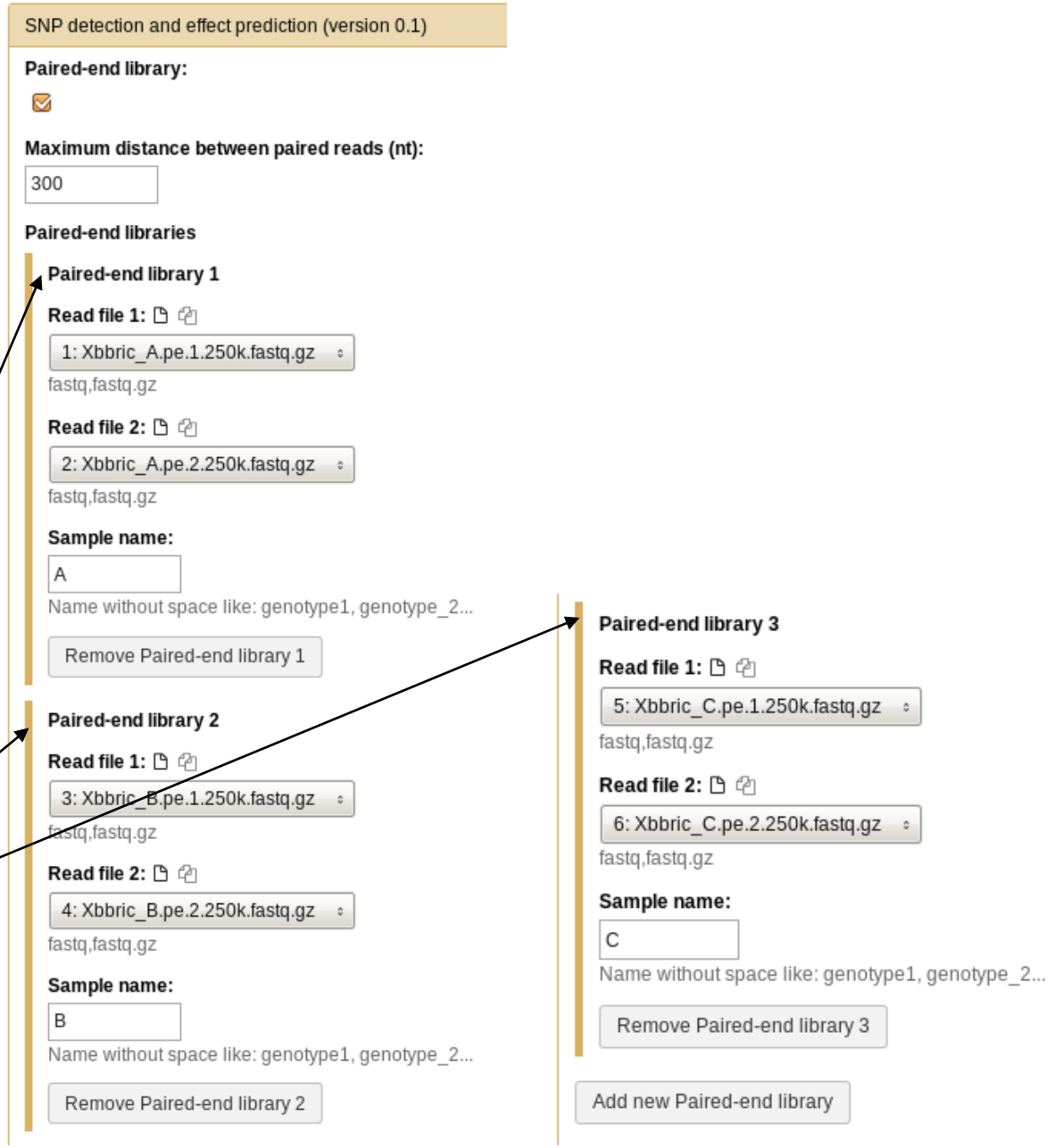
SNP detection and effect prediction

SNP detection without reference

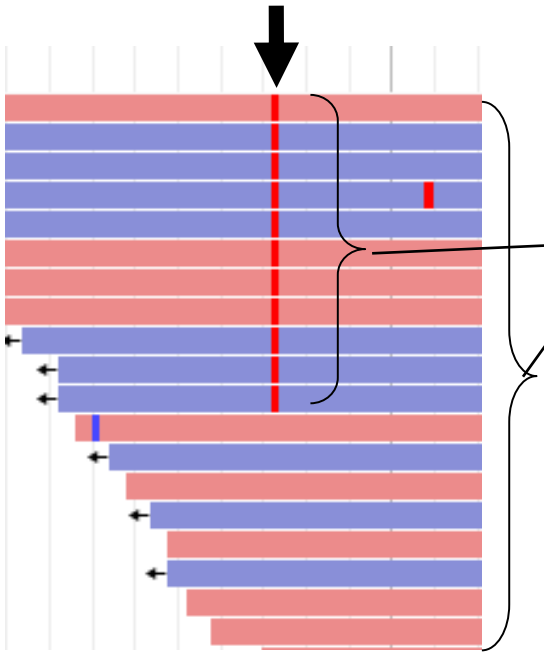
SNP detection for building an allelic frequency matrix

Variant detection and allele frequency GATK based

3 échantillons



Galaxy : Formulaire



Reference genome file (fasta):

7: Xbbric genomic sequence

Genome annotation file (GFF3):

8: Xbbric structural annotation

Minimum hit length:

80

Maximum number of mismatches:

4

Mapping (échantillons = 76~100bp)

Minimum position coverage:

20

Minimum variant coverage:

10

Minimum variant frequency:

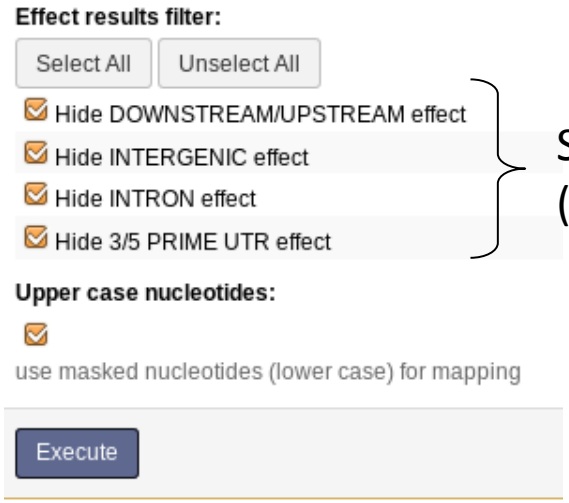
0.2

Minimum homozygous frequency:

0.75

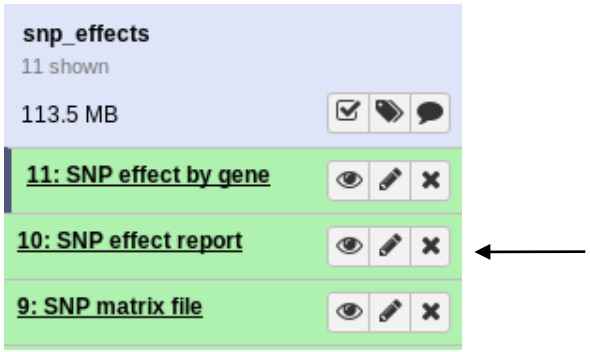
SNP calling

Galaxy : Formulaire



SNP non annotés dans le VCF
(a priori les moins intéressants)

Galaxy : Résultats



Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	5	0.012%
LOW	2,818	6.755%
MODERATE	747	1.791%
MODIFIER	38,148	91.443%

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	747	20.924%
NONSENSE	5	0.14%
SILENT	2,818	78.936%

Missense / Silent ratio: 0.2651

Number of effects by type and region

Génome compact

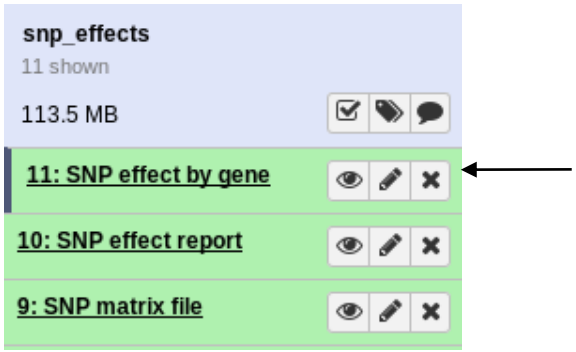
Type

Type (alphabetical order)	Count	Percent
downstream_gene_variant	18,220	43.674%
intergenic_region	479	1.148%
missense_variant	747	1.791%
splice_region_variant+stop_retained_variant	3	0.007%
stop_gained	5	0.012%
synonymous_variant	2,815	6.748%
transcript	8	0.019%
upstream_gene_variant	19,441	46.601%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	18,220	43.674%
EXON	3,567	8.55%
INTERGENIC	479	1.148%
NONE	8	0.019%
SPLICE_SITE_REGION	3	0.007%
UPSTREAM	19,441	46.601%

Galaxy : Résultats



The following table is formatted as tab separated values.

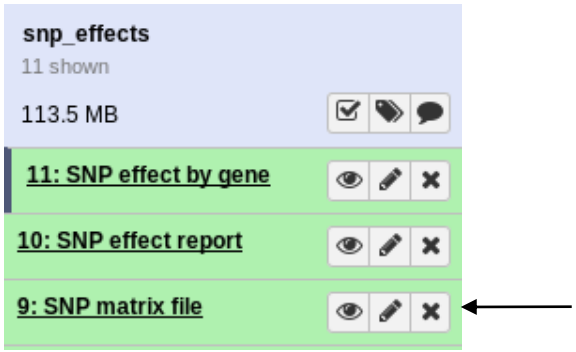
#GeneName	GeneId	variants_effect_downstream_gene_variant	variants_effect_missense_variant
GNL	gene:Xbbric.3991	6	1
V	gene:Xbbric.3969	24	0
Xbbric-rRNA-16s_rRNA-4561300-4562834	gene:Xbbric-rRNA-16s_rRNA-4561300-4562834	66	0
Xbbric-rRNA-23s_rRNA-4557919-4560799	gene:Xbbric-rRNA-23s_rRNA-4557919-4560799	38	0
Xbbric-rRNA-5s_rRNA-4557675-4557789	gene:Xbbric-rRNA-5s_rRNA-4557675-4557789	4	0
Xbbric.3908	gene:Xbbric.3908	88	3
Xbbric.3909	gene:Xbbric.3909	90	8
Xbbric.3911	gene:Xbbric.3911	93	1
Xbbric.3912	gene:Xbbric.3912	80	9
Xbbric.3913	gene:Xbbric.3913	50	6
Xbbric.3915	gene:Xbbric.3915	35	1
Xbbric.3916	gene:Xbbric.3916	35	0
Xbbric.3919	gene:Xbbric.3919	1	1
Xbbric.3921	gene:Xbbric.3921	16	1
Xbbric.3922	gene:Xbbric.3922	26	0
Xbbric.3923	gene:Xbbric.3923	3	0
Xbbric.3925	gene:Xbbric.3925	11	0
Xbbric.3926	gene:Xbbric.3926	66	1

(...)

Nom/id des gènes

Nombres de SNP trouvés dans chaque catégorie

Galaxy : Résultats

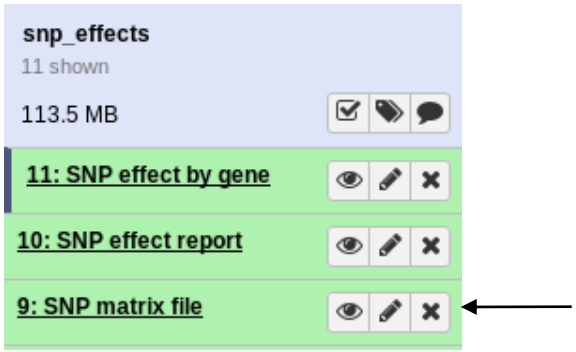


Fichier VCF : standard, VCFTools

#CHROM	POS	ID	REF	ALT	QUAL	FILTER
Xbbric	29309	.	G	A	.	PASS

Chromosome
Position
SNP ID (. = indéfini)
Allèle de référence
Allèle alternatif
Qualité (. = indéfinie)
Qualité suffisante ou non (indéfini ici)

Galaxy : Résultats



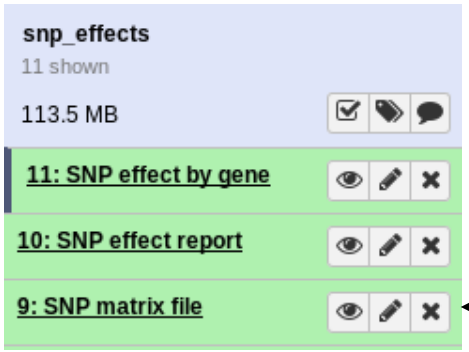
Fichier VCF : standard, VCFTools

INFO
ADP=49;WT=1;HET=0;HOM=2;NC=0;ANN=A|stop_gained|HIGH|Xbbric.3932|gene:Xbbric.3932|transcript|mRNA:Xbbric.3932|

|Coding|1/1|c.23G>A|p.Trp8*|23/339|23/339|8/112||WARNING_TRANSCRIPT_NO_START_CODON

Liste de paires de « clé=valeur » séparées par des « ; »
ADP : Profondeur moyenne par échantillon pour les bases ayant un score Phred = 20
WT : nombre d'échantillons wild-type
HET : nombre d'échantillons hétérozygotes
HOM : nombre d'échantillons homozygotes (=ont le snp x 2)
NC : échantillons « not called »
ANN : annotation ajoutée par SNPEff (ici apparition d'un codon stop)

Galaxy : Résultats



Fichier VCF : standard, VCFTools

FORMAT
GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR

A	B	C
1/1:255:49:49:0:49:100%:3.925E-29:0:38:0:0:31:18	1/1:255:47:47:0:47:100%:6.1512E-28:0:36:0:0:33:14	0/0:100:53:53:53:0:0%:1E0:69:0:23:30:0:0

Infos par échantillons selon codage défini dans la colonne FORMAT :

GT : genotype (1 = allele alt)

GQ : genotype quality

SDP : Raw Read Depth

DP : Quality Read Depth of bases with Phred score >= 20

RD : Depth of reference-supporting bases

AD : Depth of variant-supporting bases

FREQ : Variant allele frequency

PVAL : P-value from Fisher's Exact Test (not computed here : default value)

RBQ : Average quality of reference-supporting bases

ABQ : Average quality of variant-supporting bases

RDF/RDR : Depth of reference-supporting bases on forward/reverse strand

ADF/ADR : Depth of variant-supporting bases on forward/reverse strand

Je veux identifier des SNPs et évaluer leurs effets sur l'annotation d'un génome

Je vais sur	L'outil permet	L'outil ne permet pas

Paramètres clés	Pièges	Formats et fichiers

Détection des SNPs et calcul des fréquences alléliques pour les positions bi-alléliques

Responsable et intervenant principal: Sébastien Carrere
Expert: Ludovic Legrand

Plan

- Une matrice de comptage/fréquences alléliques pour quoi faire?
  - Contexte et objet d'étude nécessitant ce type de matrice
- Le pipeline pour la détection des SNPs et le calcul des fréquences alléliques (positions bi-alléliques)
  - Objectif et périmètre du pipeline
  - Fonctionnement général du pipeline
    - Mapping et calling de variants (SNPs)
    - Génération de la matrice de SNPs poolés
    - Filtre des sites bi-alléliques et génération de la matrice de comptages/fréquences alléliques
  - Fichiers de sortie:
    - VCF
    - Fichier de comptage allélique
  - Interface galaxy du pipeline
- Perspectives: exploitation de la matrice de fréquences alléliques

Qu'est-ce qu'une fréquence allélique?

- Un **allèle** est un variant, une des versions d'un même gène/locus (R ou r).
- Pour un organisme diploïde, une paire d'allèles, hérités de chacun des parents, représente le **génotype** d'un gène spécifique (RR, Rr ou rr).

Estimation des fréquences alléliques dans le polymorphisme floral chez les Gueules-de-loup (*Antirrhinum majus*)

Echantillon de **400** plantes d'une population **diploïde**:
165 Rouges; 190 Roses; 45 Blanches

P = fréquence de l'allèle R dans l'échantillon
= (2x165+190)/(400*2) = 0.65

Q = fréquence de l'allèle r dans l'échantillon
= (190+2x45)/(400*2) = 0.35

Vérification: p+q = 0.65 + 0.35 = 1.00

Qu'est-ce qu'une fréquence allélique?

Fréquence Allélique

C'est une mesure de la variabilité génétique au sein d'une population. Elle quantifie à un locus donné la probabilité d'observer un allèle (variant) donné par rapport à l'ensemble des allèles connus à ce locus dans une population.

Si on considère un gène A possédant de multiples allèles A1, …, An en un locus d'une population diploïde,

$$p(i) = (2n(AiAi) + sum(i \neq j, nAiAj)) / 2N$$

Où p(i) représente la fréquence de l'allèle Ai en ce locus.

Fréquence Génotypique

Une mesure de la diversité génétique au sein d'une population. Elle quantifie à un locus donné la probabilité d'observer un génotype spécifique par rapport à l'ensemble des génotypes recensés à ce même locus dans la population.

$$F(AA) = n(AA)/N$$

Où F(AA) représente la fréquence des individus ayant le génotype AA, n(AA) est le nombre d'individus ayant ce génotype, et N le nombre total d'individus.

Les fréquences alléliques pour quoi faire?

- **Génétique/Génétique des populations**

  Traite de l'impact des différentes forces évolutives (sélection, mutations, migration) sur la répartition de la diversité génétique entre les populations et dans les populations.

- **Les fréquences alléliques permettent de pouvoir:**
  - Quantifier et typer des **variations génétiques** présentes dans les populations (MAF, Minor Allele Frequency): genique ⇔ genotype
  - Chercher des **associations** entre variants et des traits phénotypiques (GWAS): genotype ⇔ phenotype

- Les **changements dans la fréquence des allèles** dans le temps peuvent indiquer la présence d'une **dérive génétique** ou bien que de **nouvelles mutations** ont été introduites dans la population.

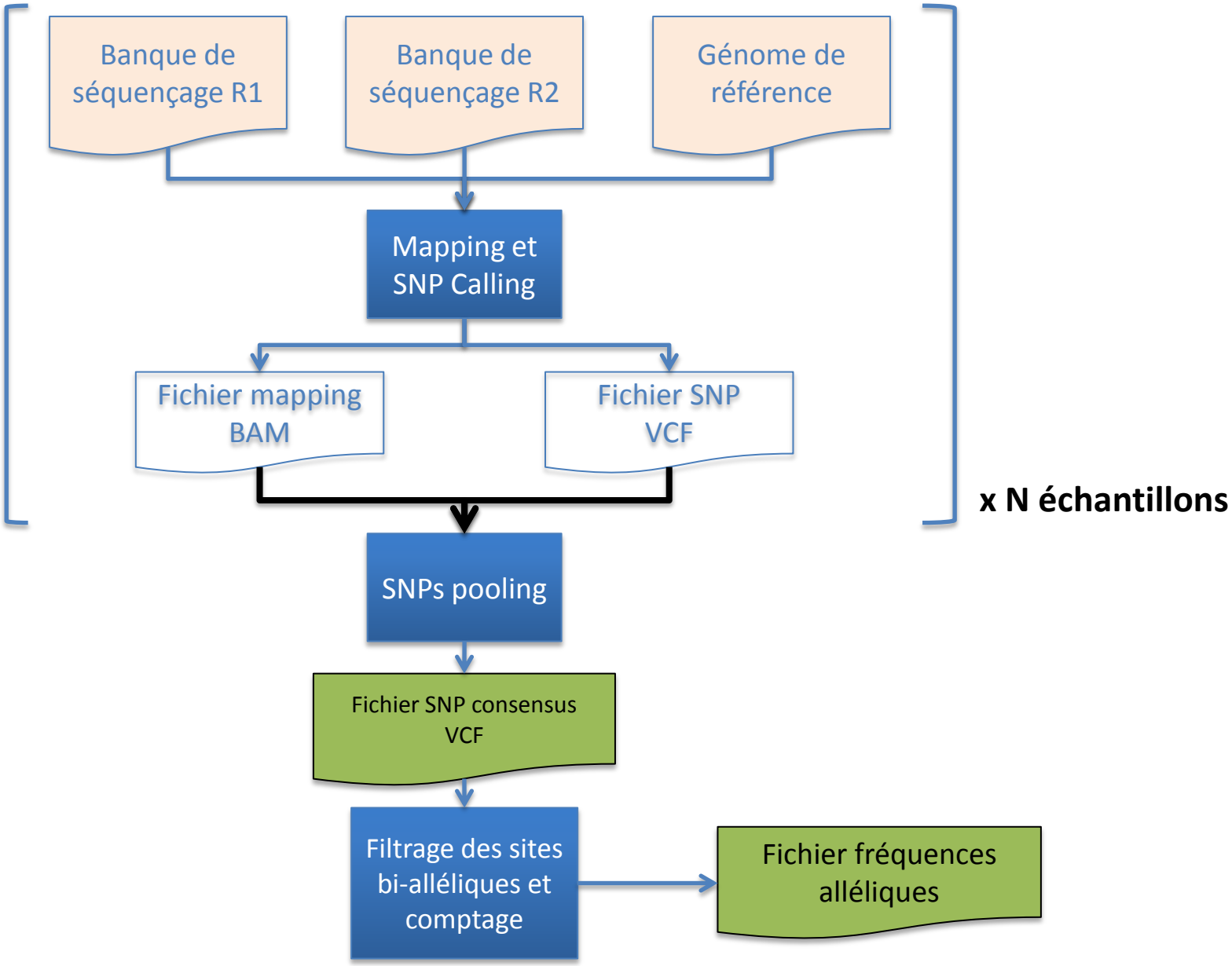
Objectif et périmètre du pipeline

- **Objectif**: Détecter les variants de type SNP afin de calculer sur un ensemble d'échantillons les fréquences alléliques associées aux SNPs uniquement aux positions bi-alléliques.

- **Périmètre**: principalement pour réaliser des études d'association (GWAS) en génétique/génétique des populations et/ou caractériser des allèles peu fréquents dans une population (Pool-seq).

- **Limitations**: SNPs bi-alléliques uniquement, exclusion des sites multi-alléliques et des InDels

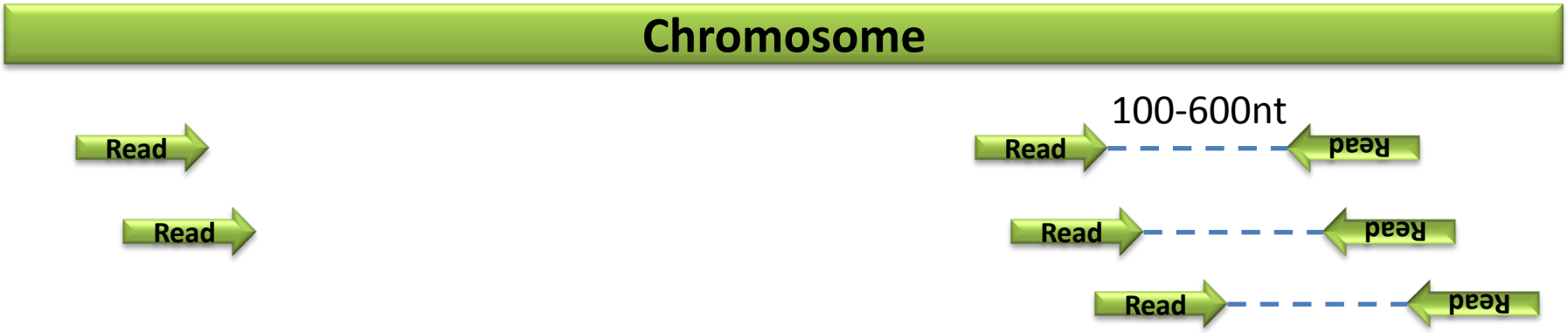
Fonctionnement du pipeline



Fonctionnement du pipeline: mapping et SNP calling

Mapping

Par échantillon

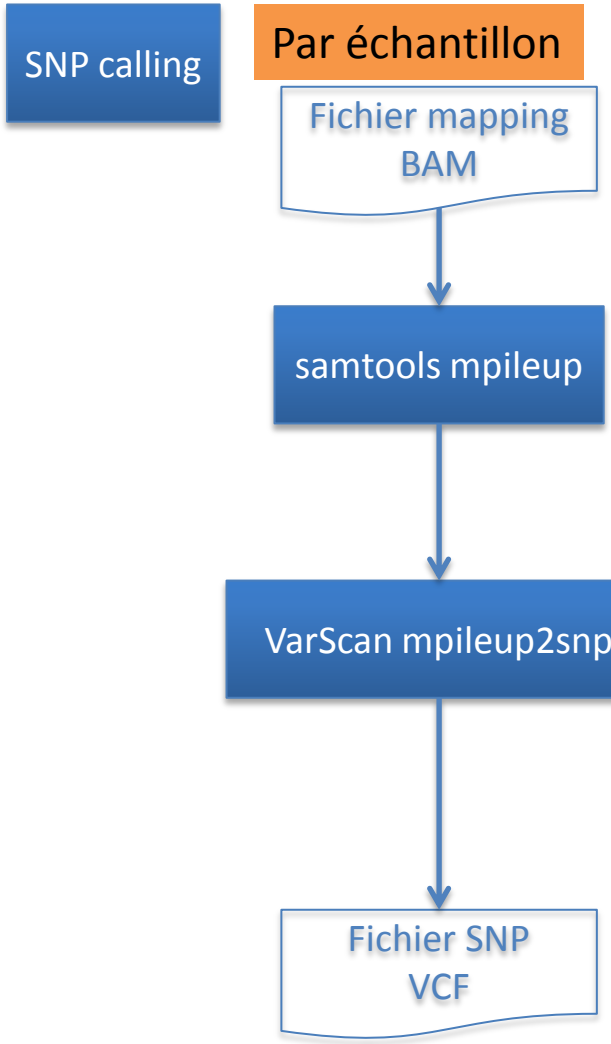


Librairie Single-end

Librairie Paired-end

Paramètres par défaut de glint: -lmin=50nt –mmis=4

Fonctionnement du pipeline: mapping et SNP calling



option -B: désactive le calcul du BAQ (Base Alignment Quality)
=> Trop strict, perte de SNPs vrais positifs

options:

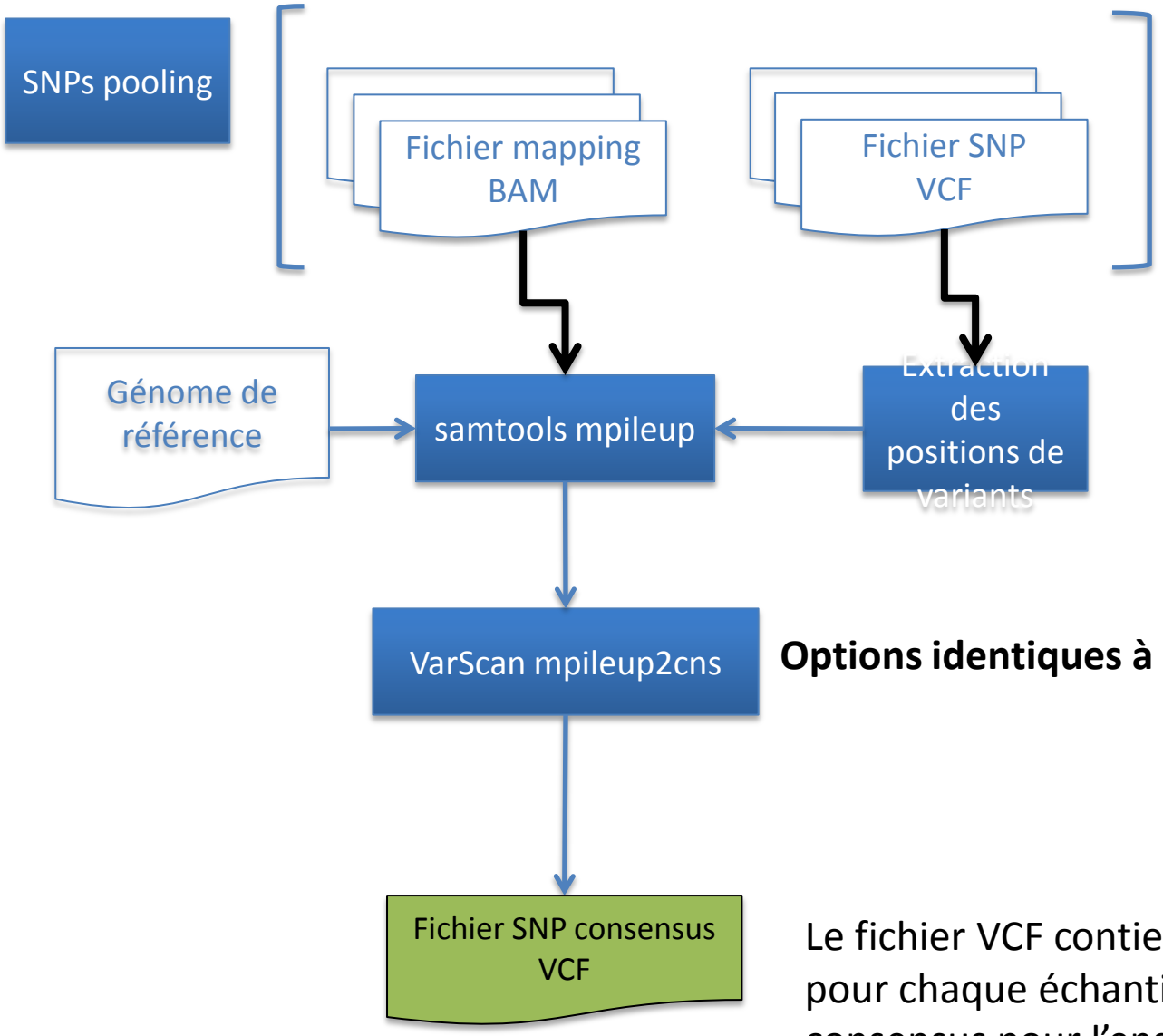
Minimum position coverage
Profondeur minimum en reads à une position donnée pour faire le calling
valeur par défaut du pipeline: 10

Minimum variant coverage
Nombre minimum de reads supportant un variant pour faire le calling
valeur par défaut du pipeline: 2

Minimum variant frequency
Fréquence allélique minimum du variant pour le conserver
valeur par défaut du pipeline: 0,01

http://samtools.sourceforge.net/mpileup.shtml
http://varscan.sourceforge.net/using-varscan.html

Fonctionnement du pipeline: SNP pooling

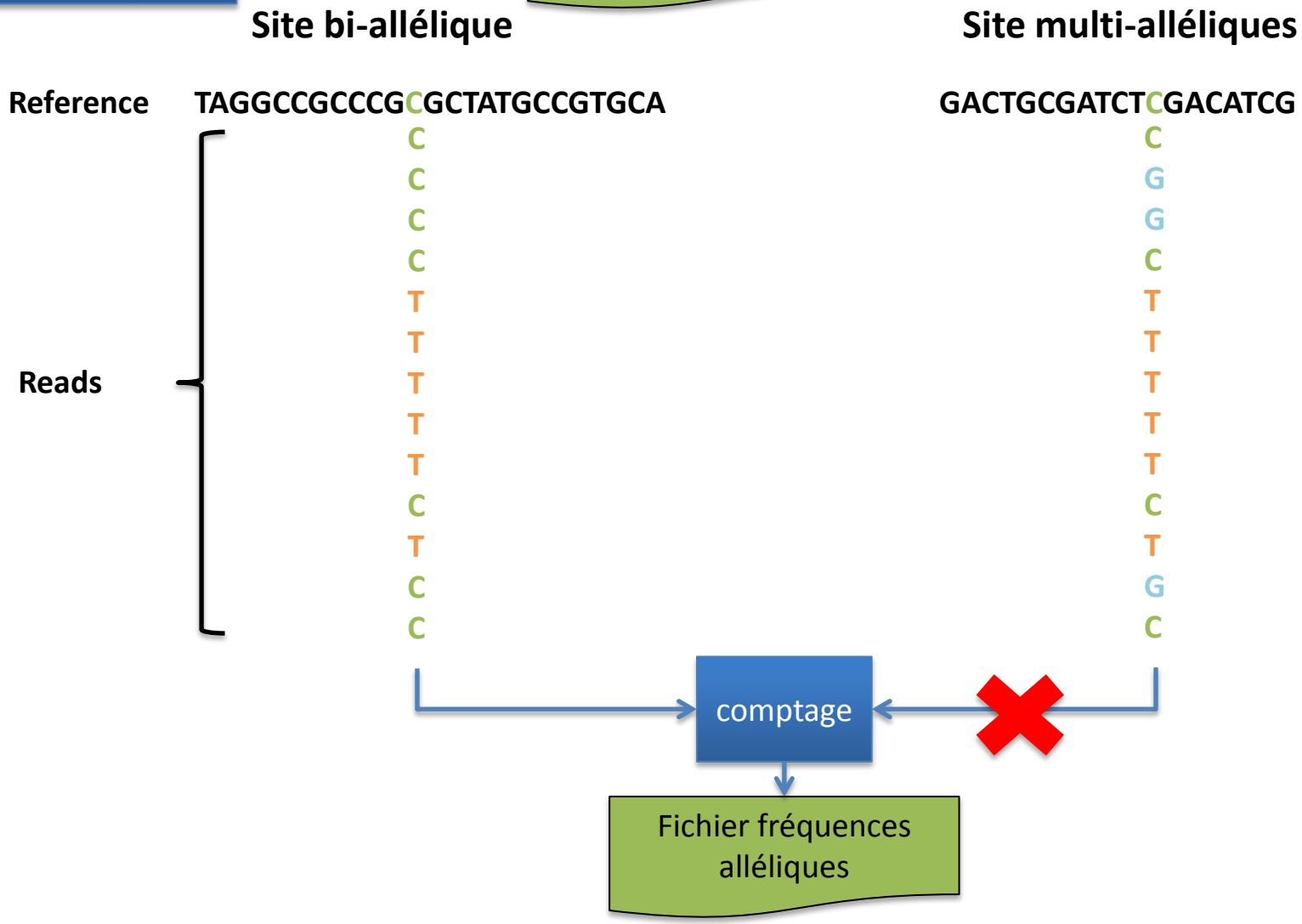


Le fichier VCF contient les variants et génotypes pour chaque échantillon ainsi que le génotype consensus pour l'ensemble des échantillons.

Fonctionnement du pipeline: Filtrage et comptage

Filtrage des sites bi-alléliques et comptage

Fichier SNP consensus VCF



Les sorties: VCF et fichier de fréquences alléliques

TP

Fichier fréquences alléliques

Allelic count

Tabular file with Allelic count by sample.

CHROM	POS	REF	ALT	#ALLELES	Spl1 #REF	Spl1 #TOT	Spl2 #REF	Spl2 #TOT
ID1	475	G	A	1	NA	NA	0	13
ID1	781	A	G	1	0	10	15	0

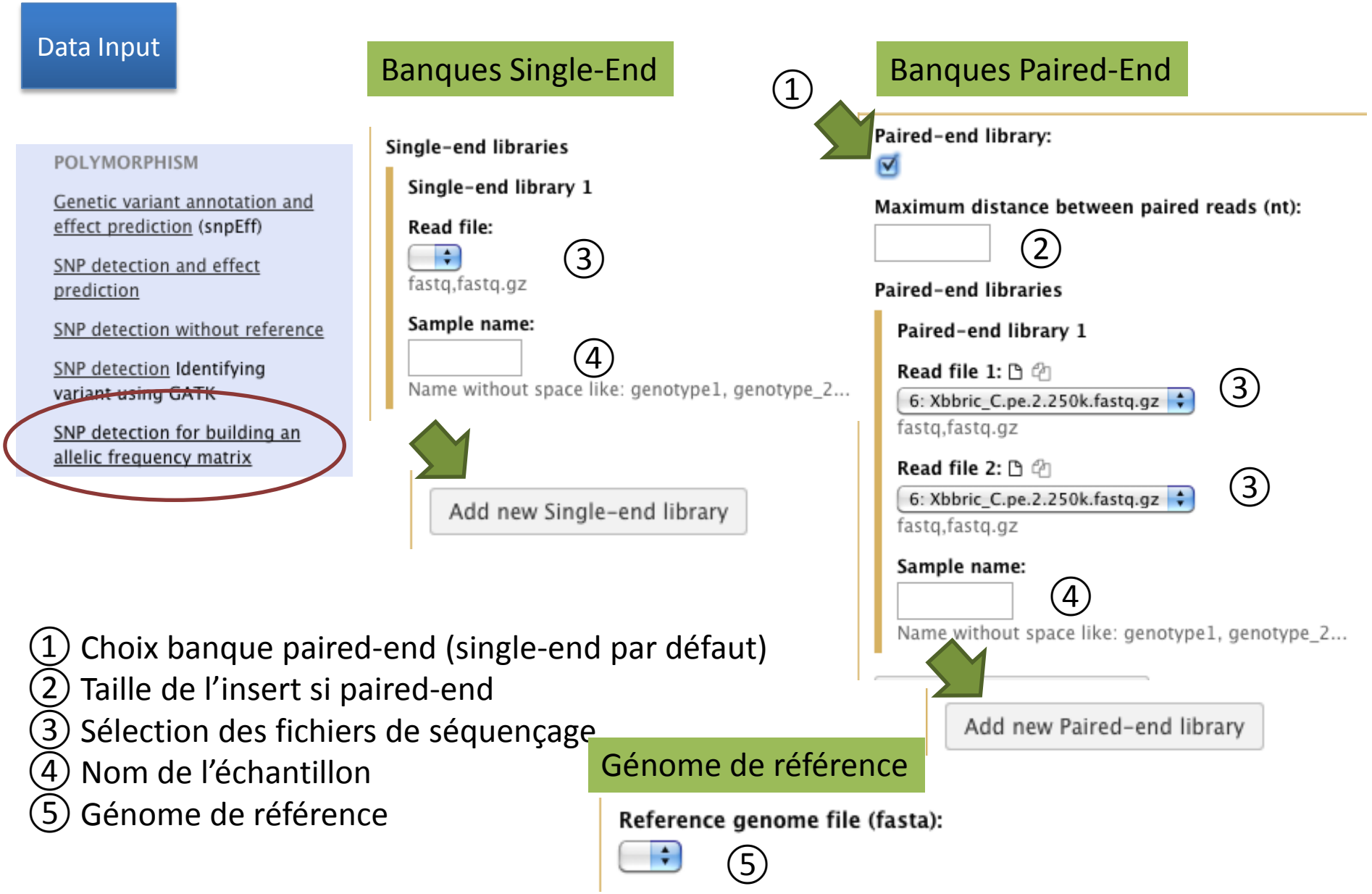
NA: position not in sample (missing data)

CHROM: chromosome or scaffold ID
POS: variant position
REF: base on reference sequence
ALT: alternative base
#ALLELE: number of allele
Spl1 #REF: number of reads with same base than reference for sample Spl1
Spl1 #TOT: number of reads for sample Spl1
Spl2 #REF: number of reads with same base than reference for sample Spl2
Spl2 #TOT: number of reads for sample Spl2

Fichier SNP consensus VCF

Cf. la fiche de description du format VCF.

Interface Galaxy du pipeline



① Choix banque paired-end (single-end par défaut)
② Taille de l'insert si paired-end
③ Sélection des fichiers de séquençage
④ Nom de l'échantillon
⑤ Génome de référence

Interface Galaxy du pipeline

Mapping and SNP calling/SNP pooling

Glint

Minimum hit length:

50 Couverture minimum le long de la read

Maximum number of mismatches:

4 Nombre maximum de substitutions autorisées

VarScan mpileup2snp/mpileup2cns

Minimum position coverage:

10 Profondeur minimum en reads à une position donnée pour faire le calling

Minimum variant coverage:

2 Nombre minimum de reads supportant un variant pour faire le calling

Minimum variant frequency:

0.01 Fréquence allélique minimum du variant

Execute

Perspectives: post-traitements

- GWAS avec PLINK/gPLINK/Haploview
  - PLINK: toolset pour l'analyse de données genotype/phenotype
    - http://pngu.mgh.harvard.edu/~purcell/plink/
  - gPLINK: interface graphique pour la gestion de commandes PLINK
    - http://pngu.mgh.harvard.edu/~purcell/plink/gplink.shtml
  - Haploview: analyse d'haplotypes
    - http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haploview
- GWAS avec Efficient Mixed-Model Association eXpedited (EMMAX)
  - Transformation/conversions de format de génotypes
    - http://pngu.mgh.harvard.edu/~purcell/plink/
  - GWAS
    - http://genetics.cs.ucla.edu/emmax/

Je veux identifier des SNPs et calculer des fréquences alléliques

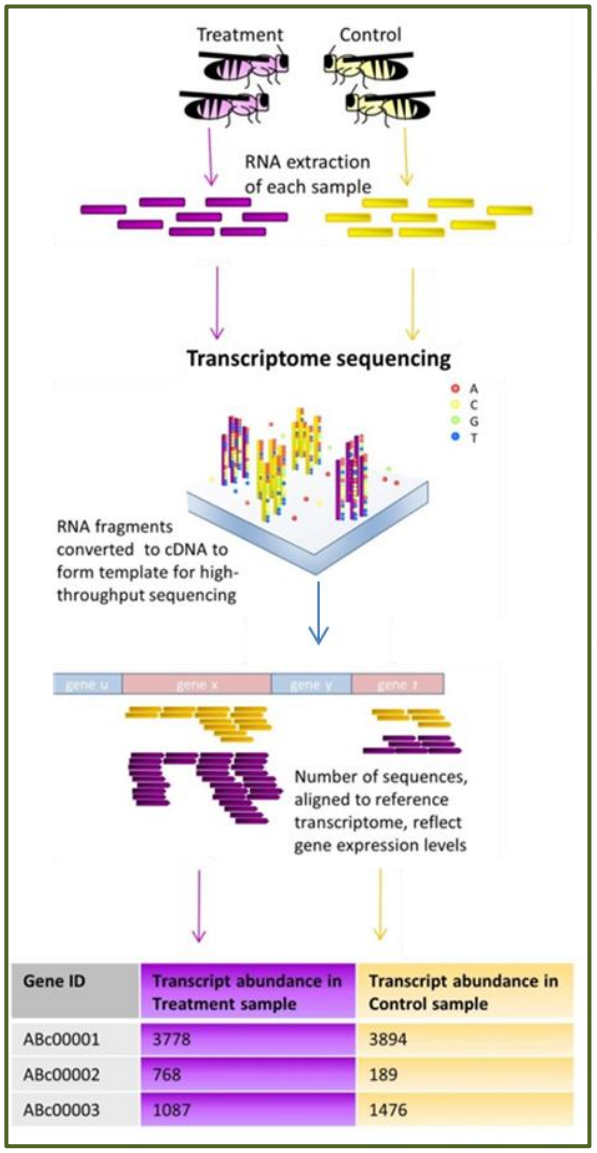
Je vais sur	L'outil permet	L'outil ne permet pas

Paramètres clés	Pièges	Formats et fichiers

MESURE DE L'EXPRESSION A PARTIR DE DONNÉES RNASEQ

Responsable et intervenant principal: Erika Sallet
Expert: Ludovic Legrand

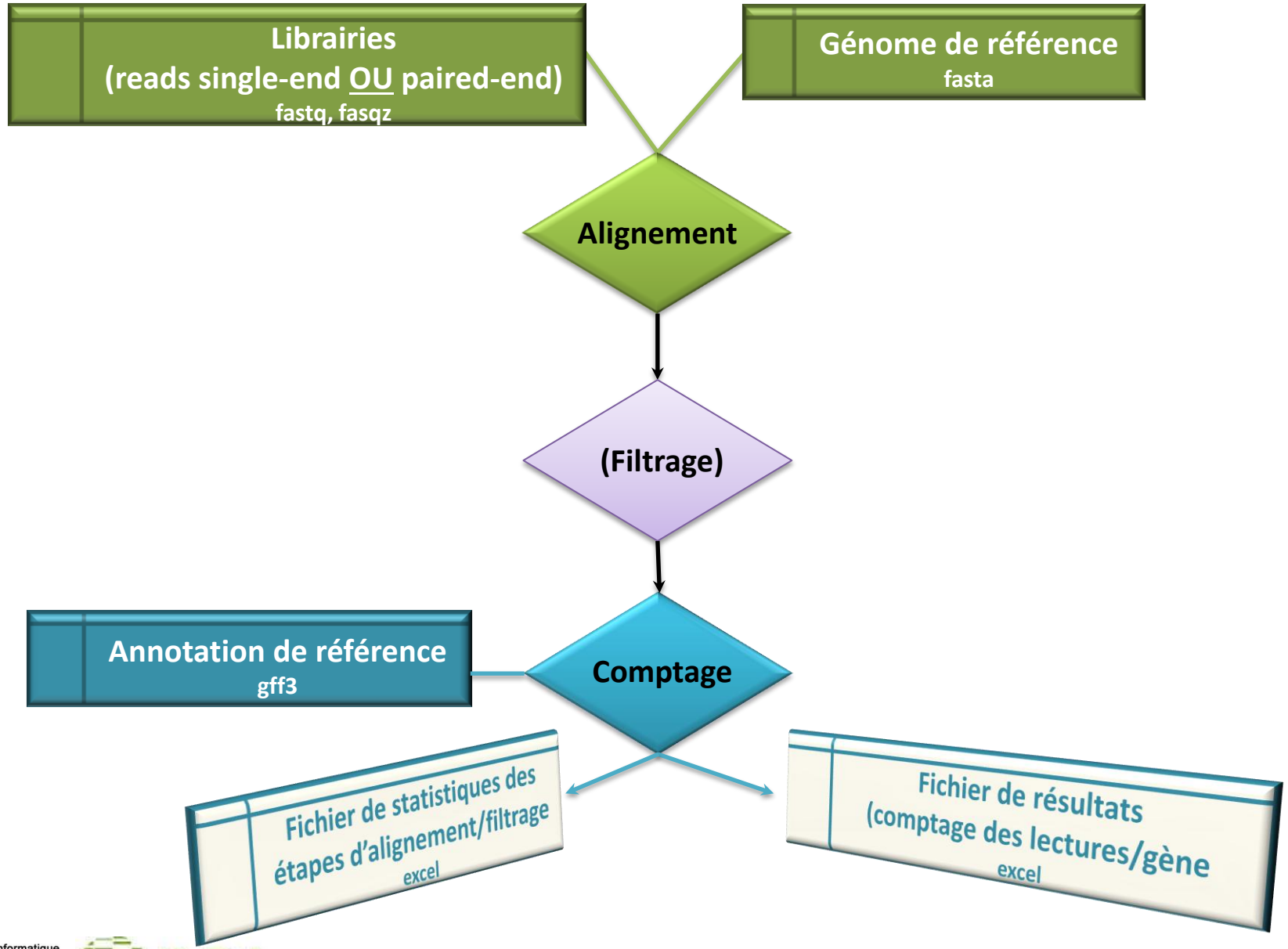
Objectif et périmètre



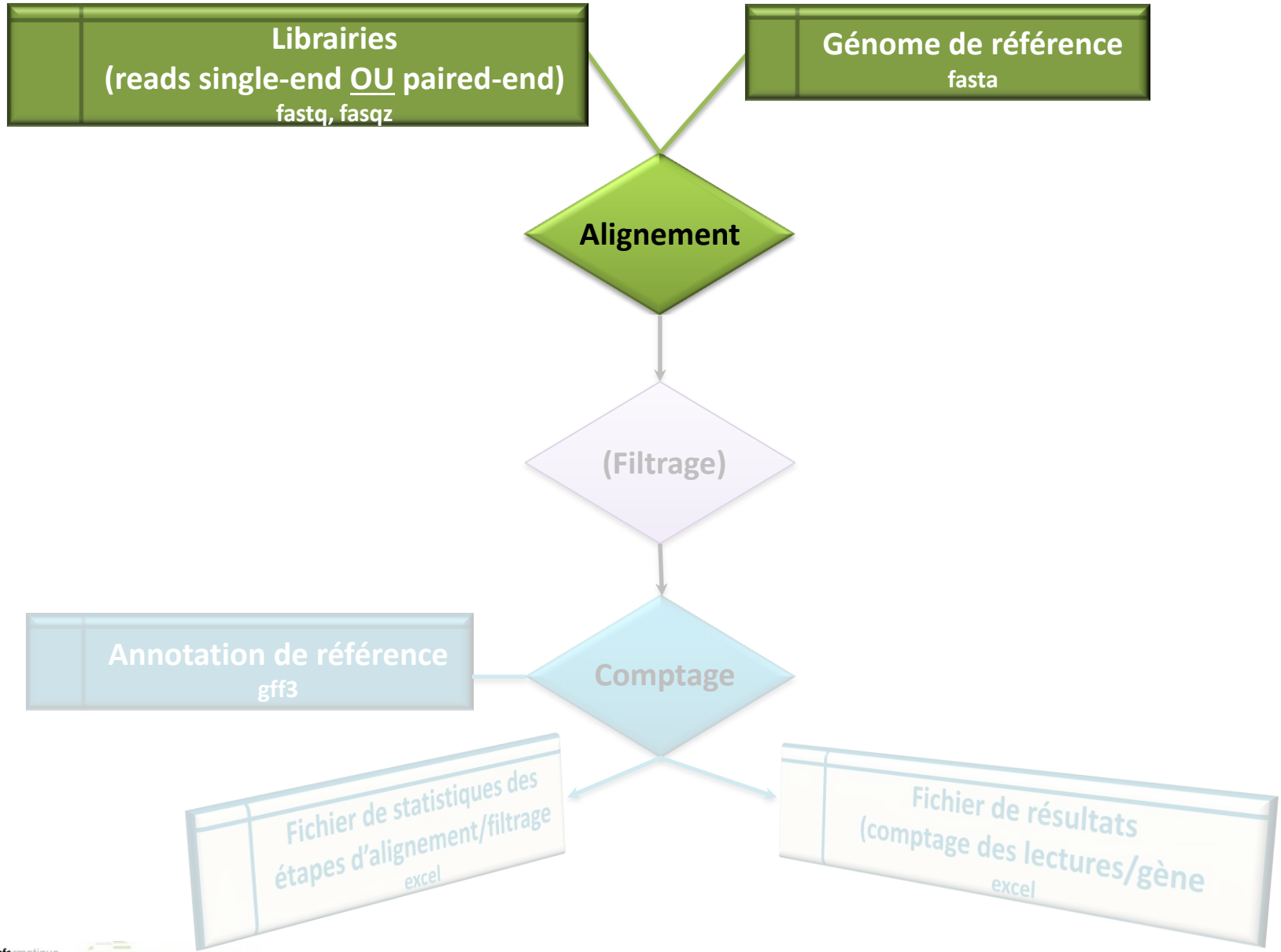
- Objectif : Mesure de l'expression
- Périmètre du pipeline :
  - Comptage des lectures sur des gènes définis dans le fichier d'annotation (pas de découverte de nouveaux transcrits)
  - Pro /eucaryotes
- Fichiers requis :
  - Librairies RNASeq paired-end ou single-end (fastq ou fastqz)
  - Génome ou transcriptome (fasta)
  - Annotation (GFF3)

Adapté de : Beyond the Gene List: Exploring Transcriptomics Data in Search for Gene Function, Trait Mechanisms and Genetic Architecture - By Bregje Wertheim

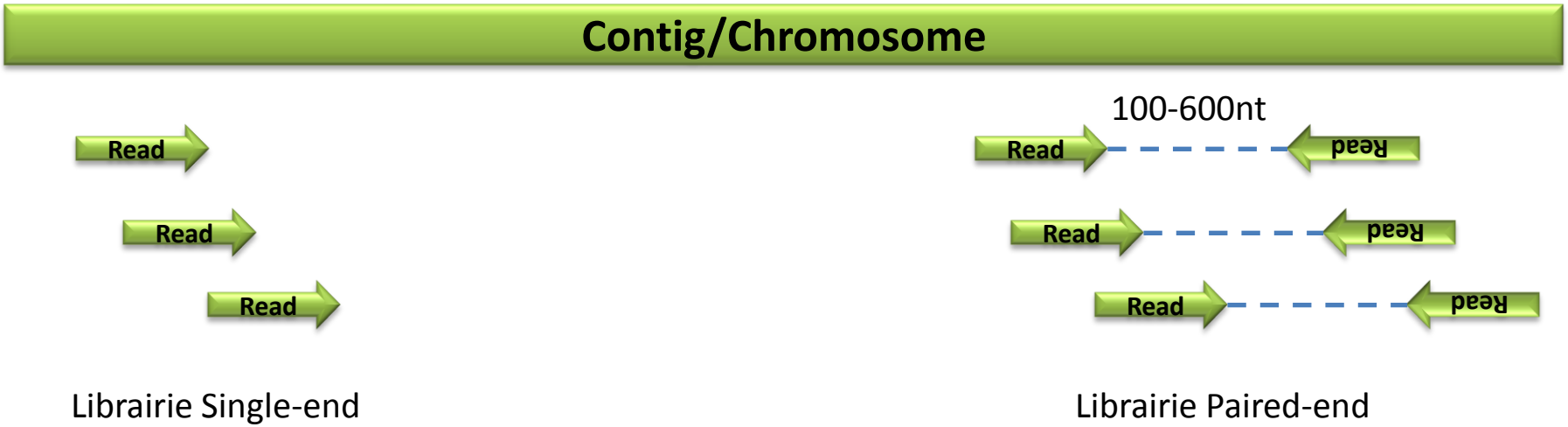
Fonctionnement



Alignement


Librairies
(reads single-end OU paired-end)
fastq, fasqz
Génome de référence
fasta
Alignement
(Filtrage)
Annotation de référence
gff3
Comptage
Fichier de statistiques des étapes d'alignement/filtrage
excel
Fichier de résultats (comptage des lectures/gène
excel

Alignement



Librairie Single-end

Librairie Paired-end

Paramètres influant sur l'alignement :

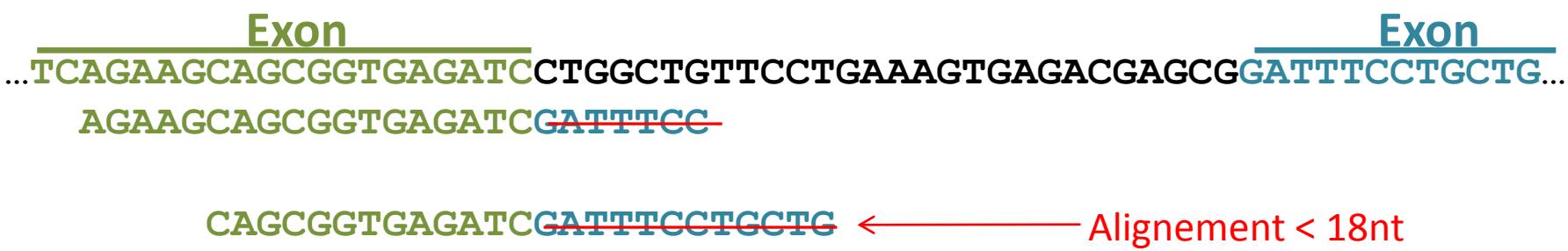
- longueur minimale du hit
- nombre maximal de mismatches
- distance maximale entre les paires
- petit ARN non codant (small RNA)

Alignement : influence des paramètres

Longueur minimale du hit

- compromis entre bruit et information
- perte des exons plus petits que ce paramètre
- valeur par défaut : 18 nt

Exemple read: 25nt lmin: 18nt



Alignement : influence des paramètres

Nombre maximal de mismatches

- erreurs de séquençage (~1%)
- variabilité avec la référence
- hétérozygotie
- valeur par défaut : 0

Exemple: lecture de 100nt paired-end

- 5 mismatches autorisés
  - prise en compte des variations alléliques et des erreurs
  - perte de spécificité d'alignement sur les lectures compensée par l'alignement de la paire

Alignement : influence des paramètres

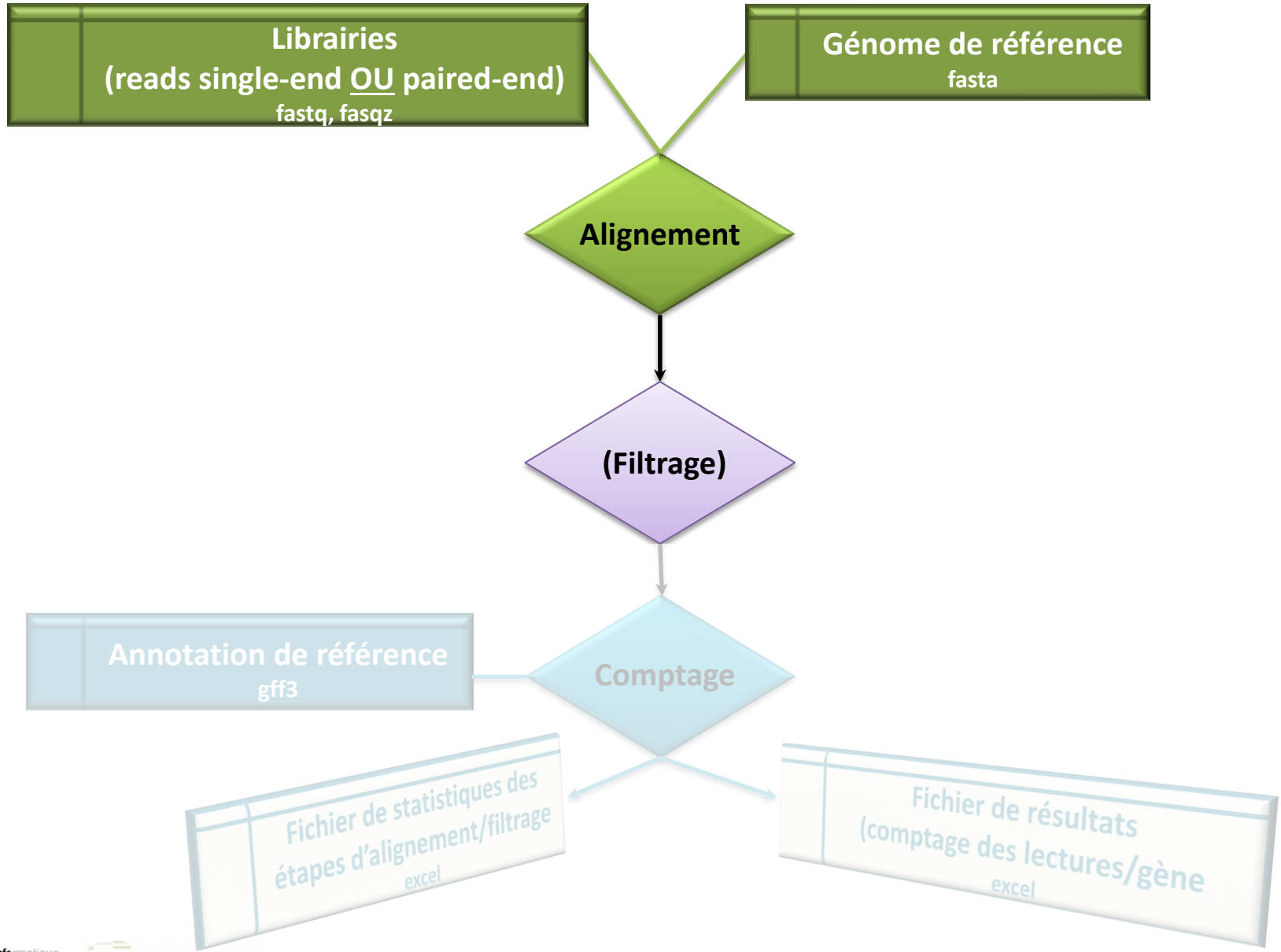
Distance maximale entre les paires

- cohérence de l'alignement de 2 reads d'une paire
- généralement entre 100 et 600 nt

Petit ARN non codant (small RNA)

- lecture entre 20 et 50 nt mais hits entre 18 et 30 nt
- 0 mismatch autorisé
  - pas d'informations complémentaires avec la paire
  - limite les alignements dus au hasard
- adaptation des seuils pour les petits alignements

Filtrage


Librairies
(reads single-end OU paired-end)
fastq, fasqz
Génome de référence
fasta
Alignement
(Filtrage)
Annotation de référence
gff3
Comptage
Fichier de statistiques des étapes d'alignement/filtrage
excel
Fichier de résultats (comptage des lectures/gène
excel

Filtrage

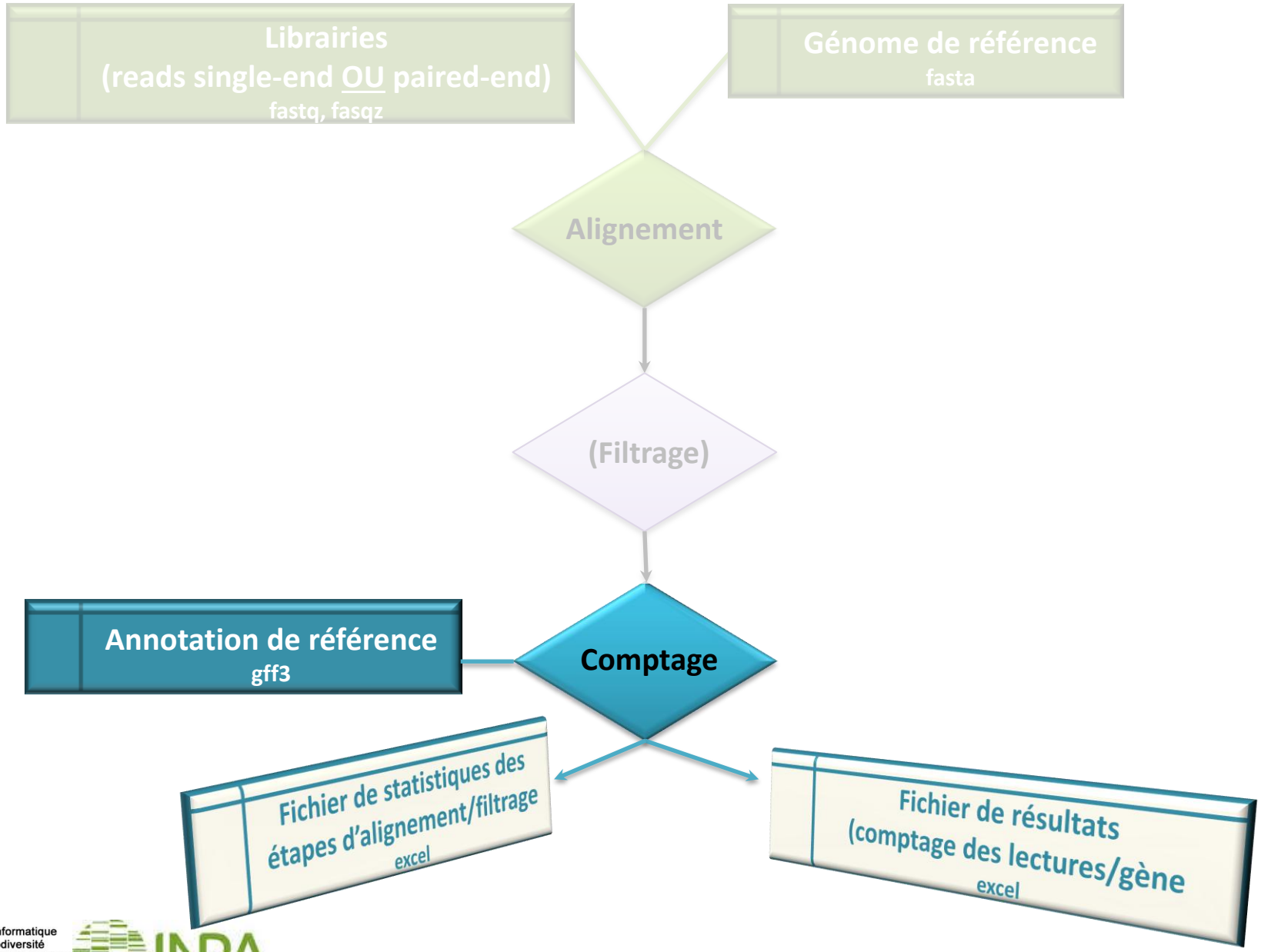
Filtrage des lectures non spécifiques (ambiguës)

- alignement identique de score maximum sur plusieurs positions
- duplication de gènes, transposons, gènes conservés

$\Rightarrow$on préfère perdre de l'information en écartant les hits ambigus

$\Rightarrow$garder les hits ambigus, c'est additionner le niveau d'expression de N objets biologiques très probablement **régulés différemment**

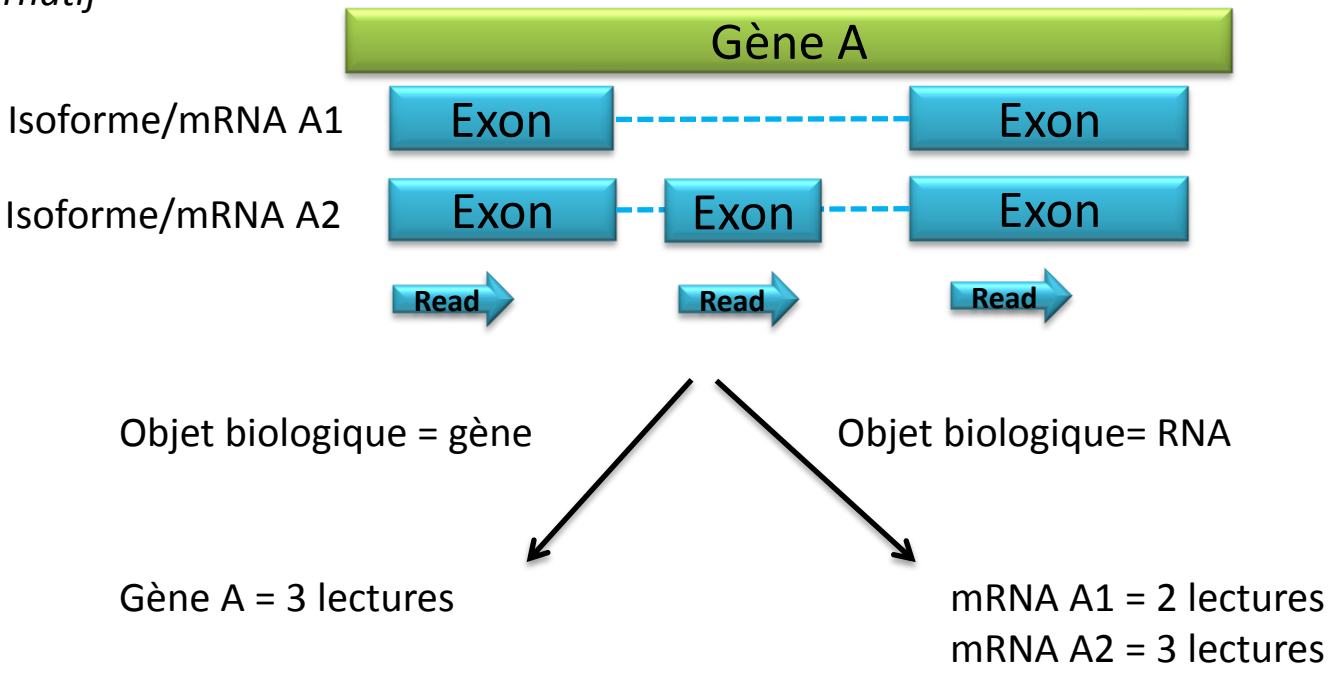
Comptage


Librairies
(reads single-end OU paired-end)
fastq, fasqz
Génome de référence
fasta
Alignement
(Filtrage)
Annotation de référence
gff3
Comptage
Fichier de statistiques des étapes d'alignement/filtrage
excel
Fichier de résultats
(comptage des lectures/gène
excel

Comptage : choix de l'objet biologique

Comptage du nombre de lectures par objet biologique

Exemple d'un gène A
avec épissage alternatif



Comptage : choix de l'objet biologique

Comptage du nombre de lectures par gène

- moins sensible à la qualité de l'annotation
- ne tient pas compte de l'épissage alternatif

Comptage du nombre de lectures par RNA

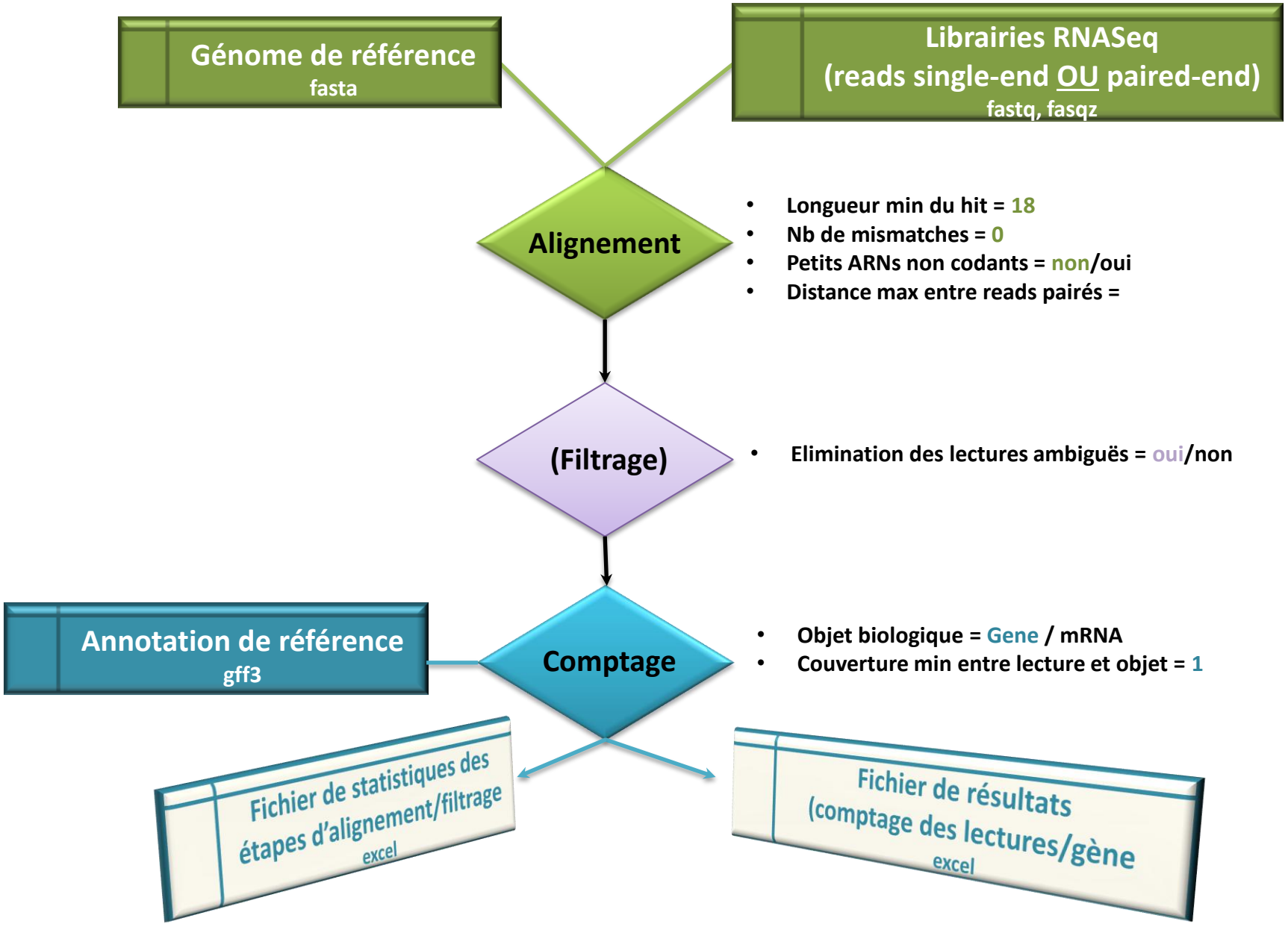
- sensible à l'annotation des exons
- tient compte de l'épissage alternatif

Comptage : influence des paramètres

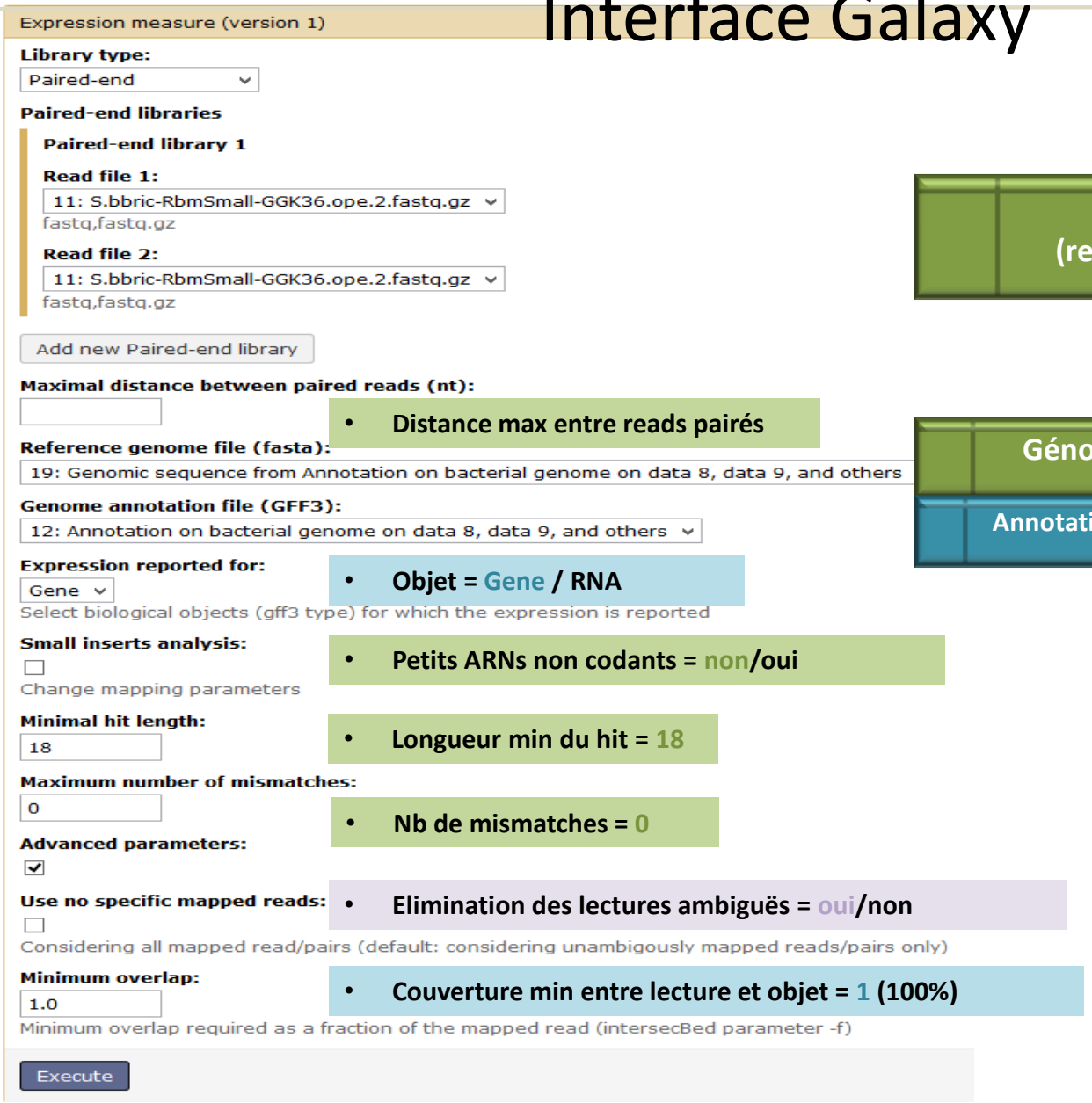
Couverture min de la lecture sur l'objet biologique

- Valeur par défaut = 1 (100%)
  - suppose une annotation de bonne qualité

- diminuer la couverture
  - autoriser les hits partiels sur les objets
  - annotation de faible ou moyenne qualité


Génome de référence
fasta
Librairies RNASeq
(reads single-end OU paired-end)
fastq, fasqz
Alignement
Longueur min du hit = 18
Nb de mismatches = 0
Petits ARNs non codants = non/oui
Distance max entre reads pairés =
(Filtrage)
Elimination des lectures ambiguës = oui/non
Annotation de référence
gff3
Comptage
Objet biologique = Gene / mRNA
Couverture min entre lecture et objet = 1
Fichier de statistiques des étapes d'alignement/filtrage
excel
Fichier de résultats (comptage des lectures/gène
excel

Interface Galaxy



Librairies
(reads single-end OU paired-end)
fastq, fasqz

- **Distance max entre reads pairés**

Génome de référence
fasta

Annotation de référence
gff3

- **Objet = Gene / RNA**
- **Petits ARNs non codants = non/oui**
- **Longueur min du hit = 18**
- **Nb de mismatches = 0**
- **Elimination des lectures ambiguës = oui/non**
- **Couverture min entre lecture et objet = 1 (100%)**

Fichier de statistiques

Nombre d'alignements spécifiques sur génome

Nombre d'alignements sur génome

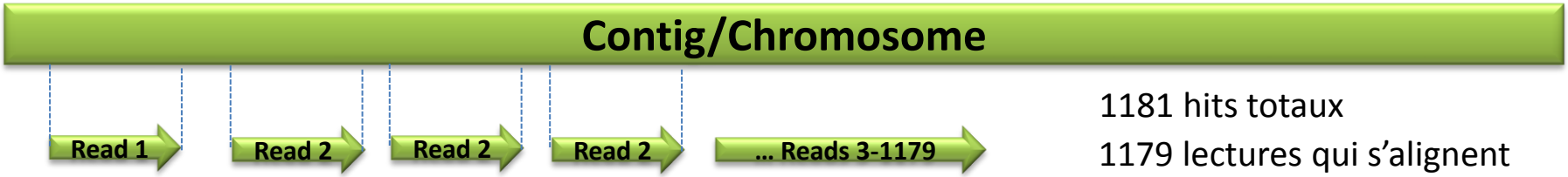
Nombre de lectures brutes

Nombre de lectures alignées sur le génome

Nombre d'alignements sur objet biologique

alignements sur objet biologique / alignements spécifiques sur génome

lib	specific_hits	mapping_hits	raw_reads/pairs_count	mapped_reads/pairs_count	feature_overlapping_hits	feature_overlapping_hits/specific_hits_percent
S.bbric_RbmLong_GGK21.ope	1178	1181	9697	1179	1153	97.88
S.bbric_RbmSmall_GGK36.ope	24405	24411	64272	24407	20674	84.71



1181 hits totaux

1179 lectures qui s'alignent

Fichier de statistiques

Nombre d'alignements spécifiques sur génome

Nombre d'alignements sur génome

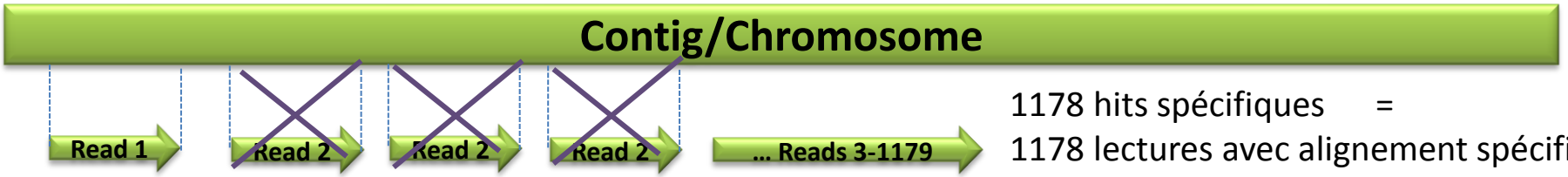
Nombre de lectures brutes

Nombre de lectures alignées sur le génome

Nombre d'alignements sur objet biologique

alignements sur objet biologique / alignements spécifiques sur génome

lib	specific_hits	mapping_hits	raw_reads/pairs_count	mapped_reads/pairs_count	feature_overlapping_hits	feature_overlapping_hits/specific_hits_percent
S.bbric_RbmLong_GGK21.ope	1178	1181	9697	1179	1153	97.88
S.bbric_RbmSmall_GGK36.ope	24405	24411	64272	24407	20674	84.71



1178 hits spécifiques =
1178 lectures avec alignement spécifique

Fichier de résultats

Gene/RNA ID | Contig Id | Positionnement du gène/RNA sur le contig: Début, Fin, Brin | Taille du Gène/RNA | Type (Gène/RNA) | Comptage brut (Lectures/Objet) | Comptage normalisé (RPKM)

id	0_seqid	1_start	2_end	3_strand	4_length	5_type	6_Note	S.bbric_RbmLong_GGK21.ope-count	S.bbric_RbmLong_GGK21.ope-rpkm	S.bbric_RbmSmall_GGK36.ope-count	S.bbric_RbmSmall_GGK36.ope-rpkm
SBBRIC1.1	SBBRIC1	384	685	-	302	gene		5	14054.58	2	260.41
SBBRIC1.10	SBBRIC1	9162	11163	+	2002	gene		14	5936.34	107	2101.63
SBBRIC1.100	SBBRIC1	90298	90594	+	297	gene		0	0.00	9	1191.58

Fichier de résultats

Comptage

- nombre de lectures alignées par objet
- utilisé pour les analyses d'expression différentielle après normalisation

Nat Methods. 2008 Jul;5(7):621-8. doi: 10.1038/nmeth.1226. Epub 2008 May 30.

Mapping and quantifying mammalian transcriptomes by RNA-Seq.

Mortazavi A[1], Williams BA, McCue K, Schaeffer L, Wold B.

Normalisation RPKM

- **R**ead **P**er **K**ilobase per **M**illion mapped reads

$$\text{RPKM (X)} = \frac{\text{Nb de lectures/gène (comptage brut)}}{\text{million lectures alignées x taille objet biologique(kb)}}$$

Normalisation entre librairies

Normalisation entre objets

Limites : normalisation RPKM

BRIEFINGS IN BIOINFORMATICS. VOL 14. NO 6. 671–683
Advance Access published on 17 September 2012
doi:10.1093/bib/bbs046

A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis

Marie-Agnès Dillies, Andrea Rau*, Julie Aubert*, Christelle Hennequet-Antier*, Marine Jeanmougin*, Nicolas Servant*, Céline Keime*, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloë, Caroline Le Gall, Brigitte Schaëffer, Stéphane Le Crom*, Mickaël Guedj*, Florence Jaffrézic* and on behalf of The French StatOmique Consortium*

Key points

- Normalization of RNA-seq data in the context of differential analysis is essential in order to account for the presence of systematic variation between samples as well as differences in library composition.
- The Total Count and RPKM normalization methods, both of which are still widely in use, are ineffective and should be definitively abandoned in the context of differential analysis.
- Only the DESeq and TMM normalization methods are robust to the presence of different library sizes and widely different library compositions, both of which are typical of real RNA-seq data.

Table 3: Summary of comparison results for the seven normalization methods under consideration

Method	Distribution	Intra-Variance	Housekeeping	Clustering	False-positive rate
TC	–	+	+	–	–
UQ	++	++	+	++	–
Med	++	++	–	++	–
DESeq	++	++	++	++	++
TMM	++	++	++	++	++
Q	++	–	+	++	–
RPKM	–	+	+	–	–

A '–' indicates that the method provided unsatisfactory results for the given criterion, while a '+' and '++' indicate satisfactory and very satisfactory results for the given criterion.

Aller plus loin…

- Normalisation et analyses statistiques pour identifier les gènes différentiellement exprimés avec R
  - DESeq ou DESeq2 (bioconductor)
  - TMM (edgeR)

Je veux mesurer l'expression de gènes connus à partir de données de séquençage

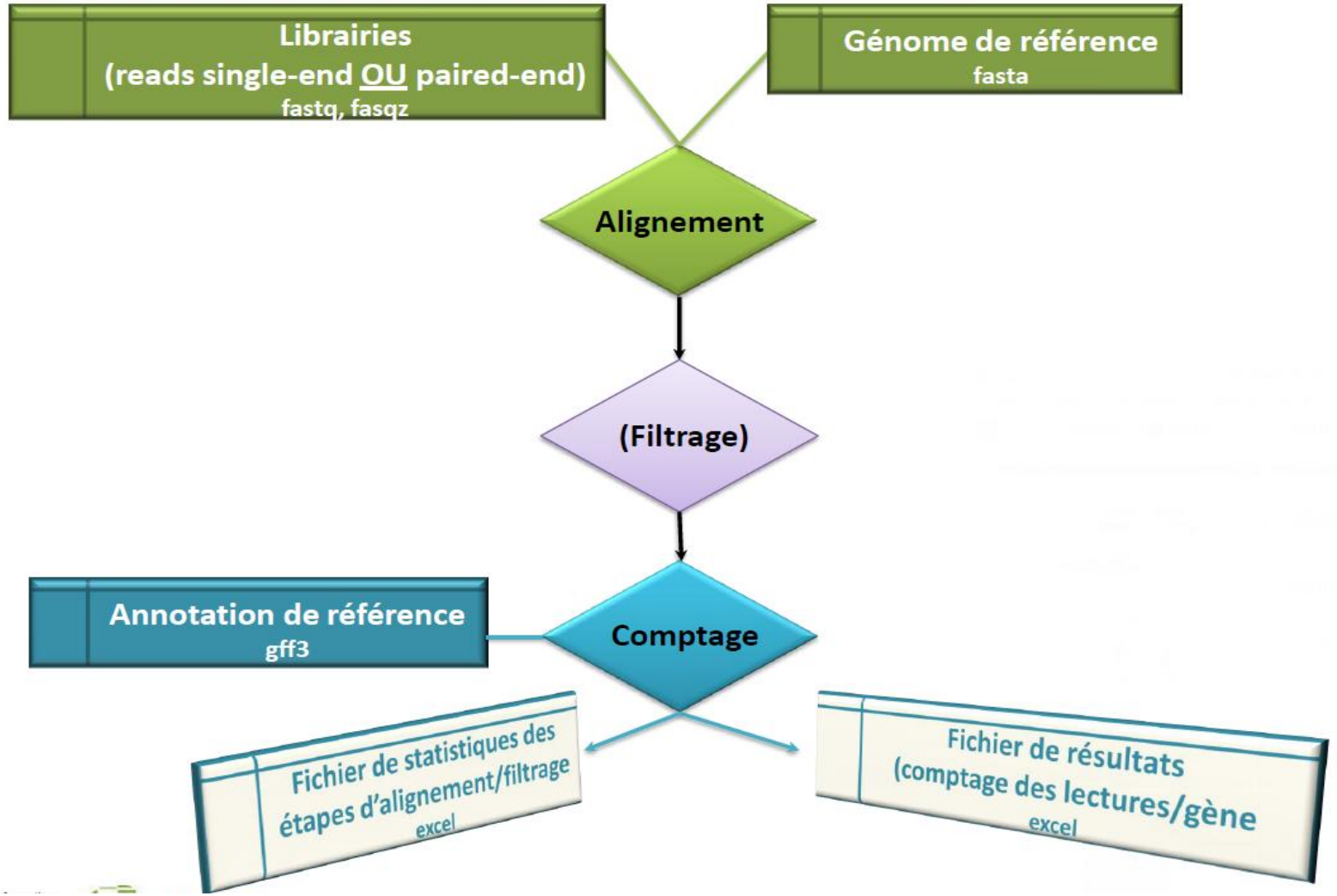
Je vais sur	L'outil permet	L'outil ne permet pas

Paramètres clés	Pièges	Formats et fichiers

CONTRÔLE QUALITÉ DE MESURES D'EXPRESSION

Responsable et intervenant principal: Erika Sallet
Expert: Adeline Simon et Joseph Tran

Rappel : pipeline Mesures de l'expression

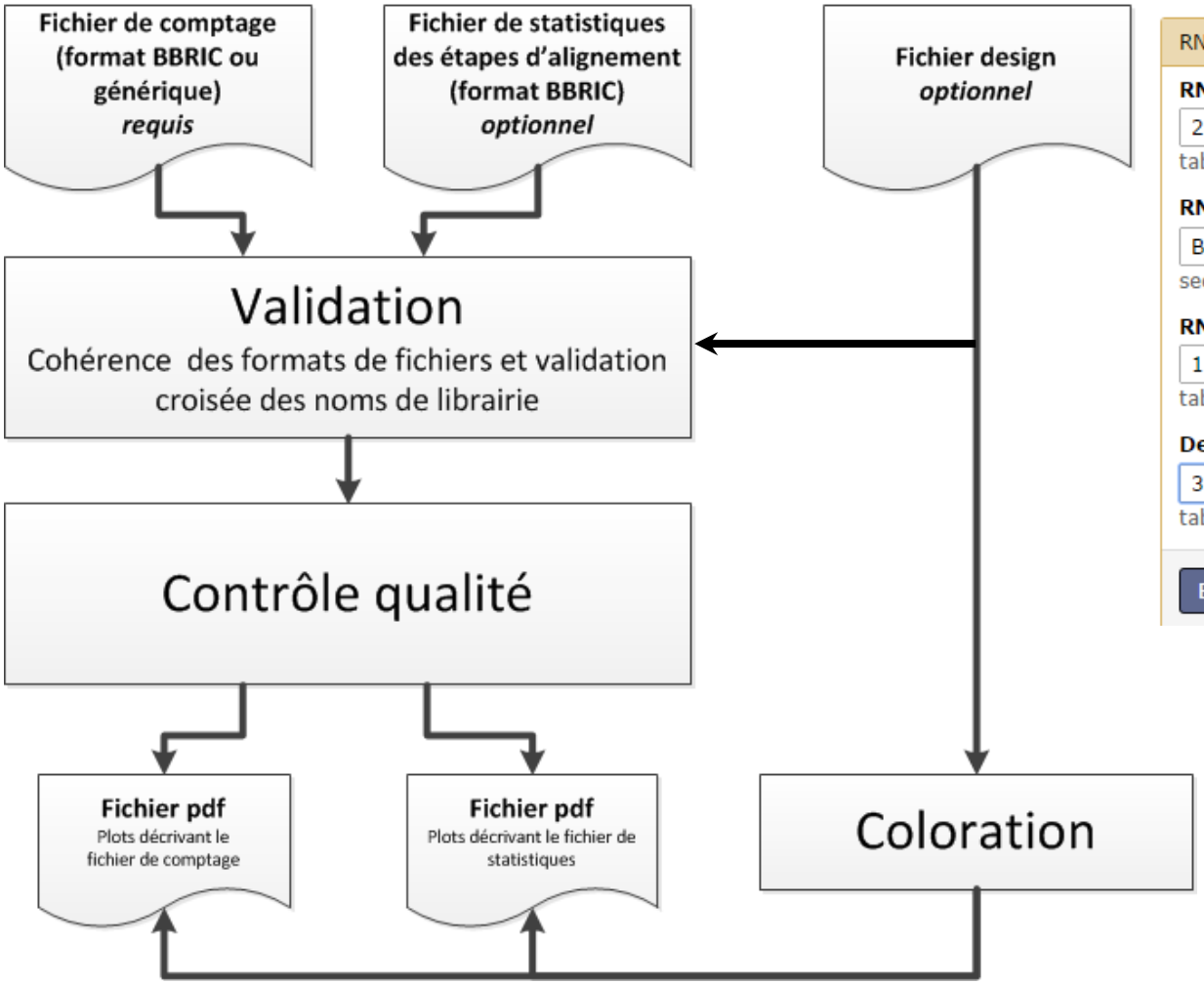


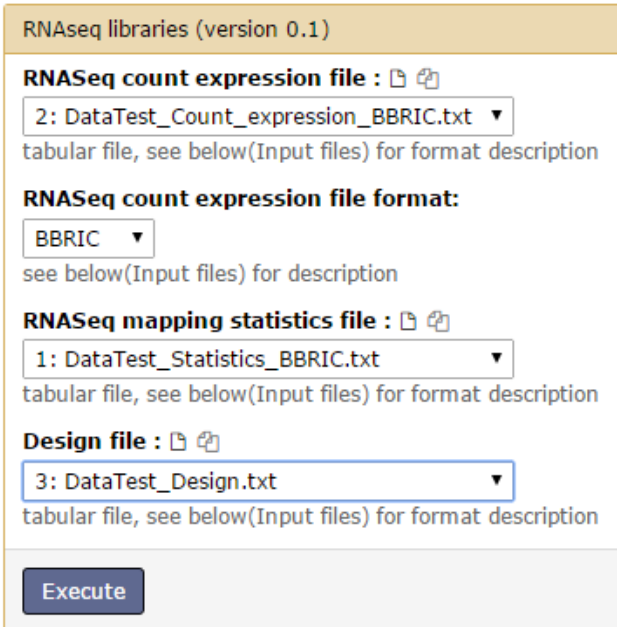
Objectif et périmètre du pipeline

- Objectif : identifier facilement des problèmes de qualité dans les résultats issus d'une mesure d'expression

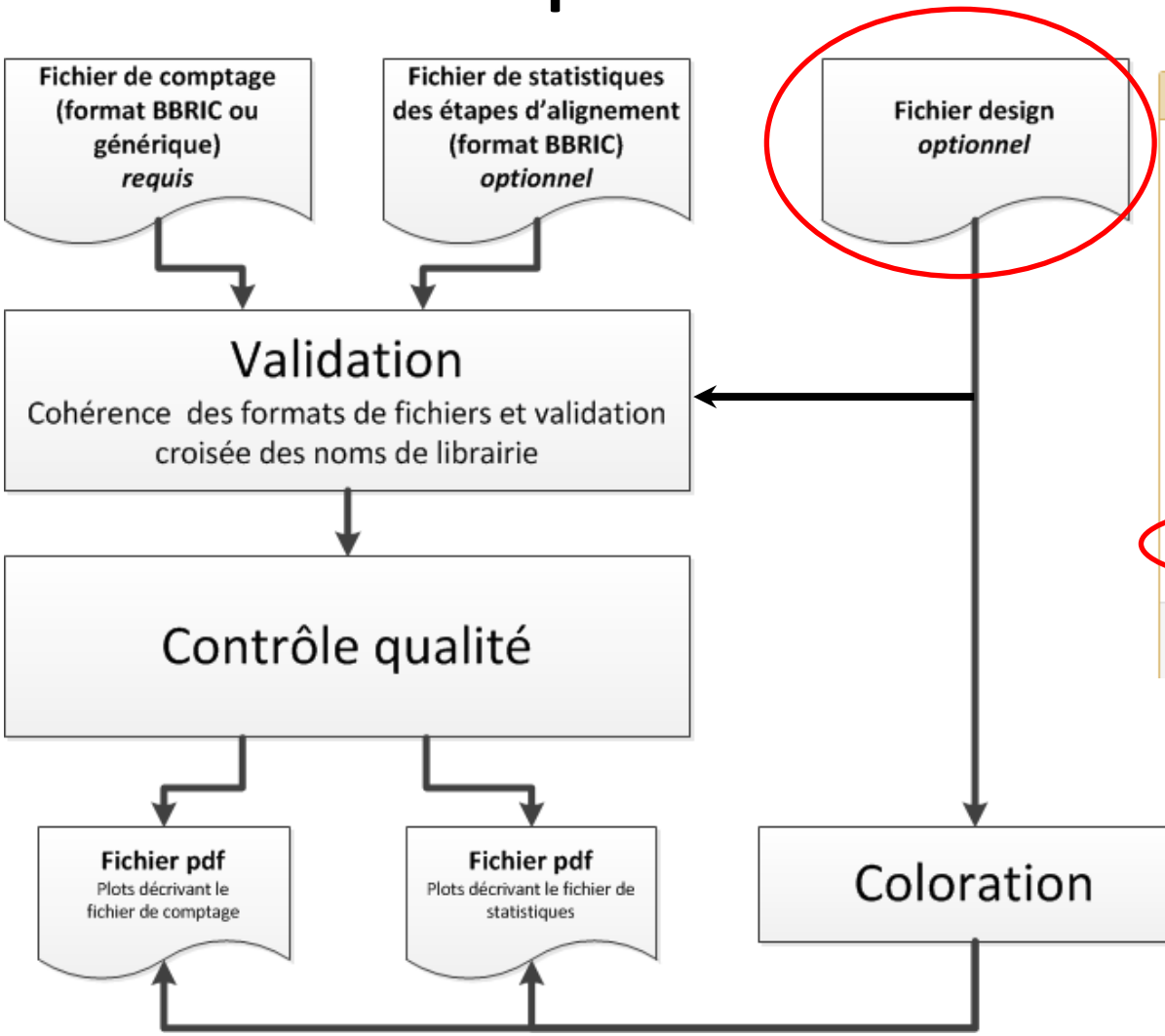
- Périmètre : sortie du pipeline BBRIC ou pipelines externes équivalents

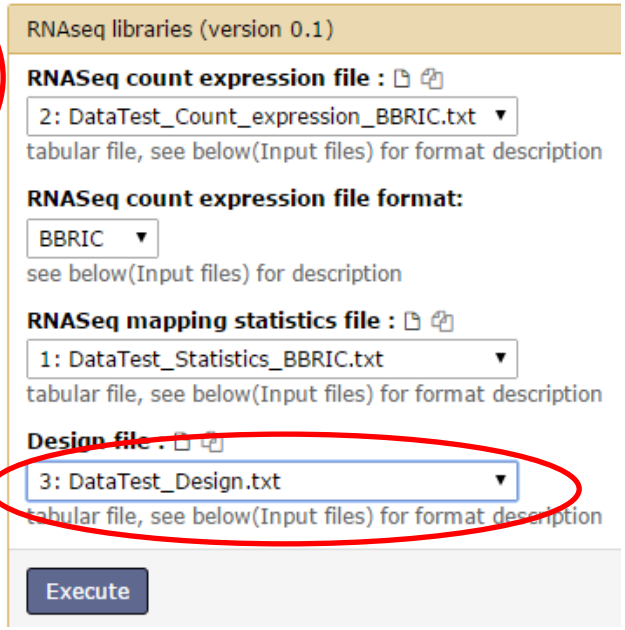
Description du Workflow





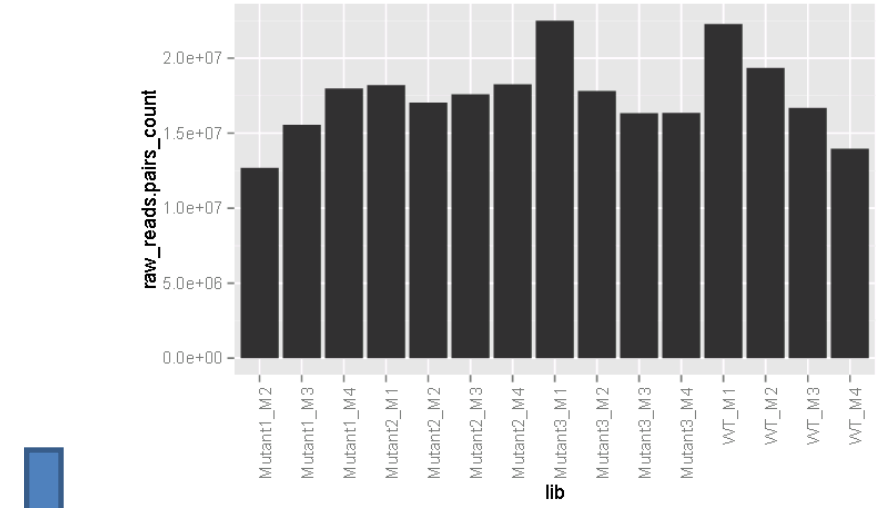
Description du Workflow



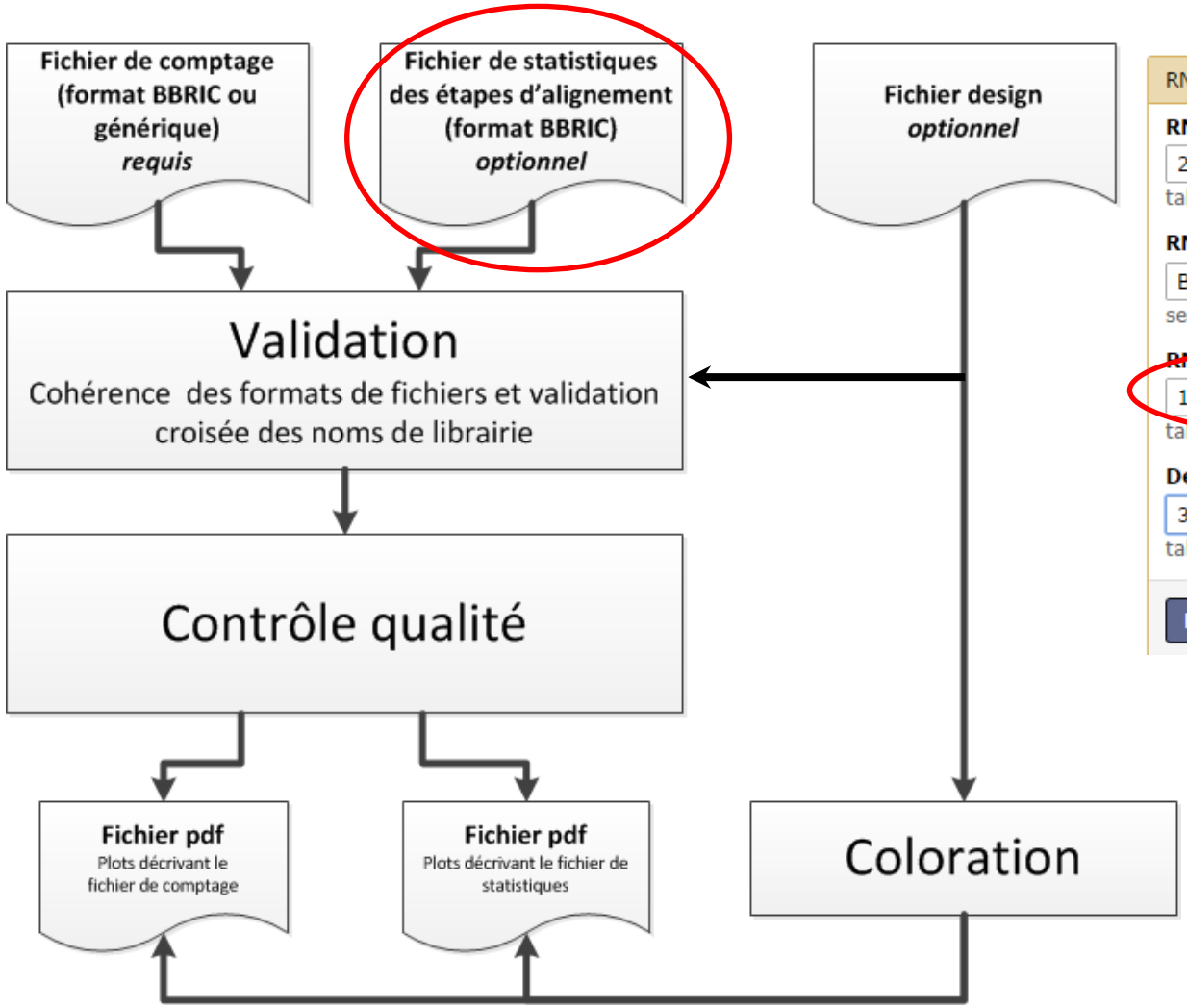


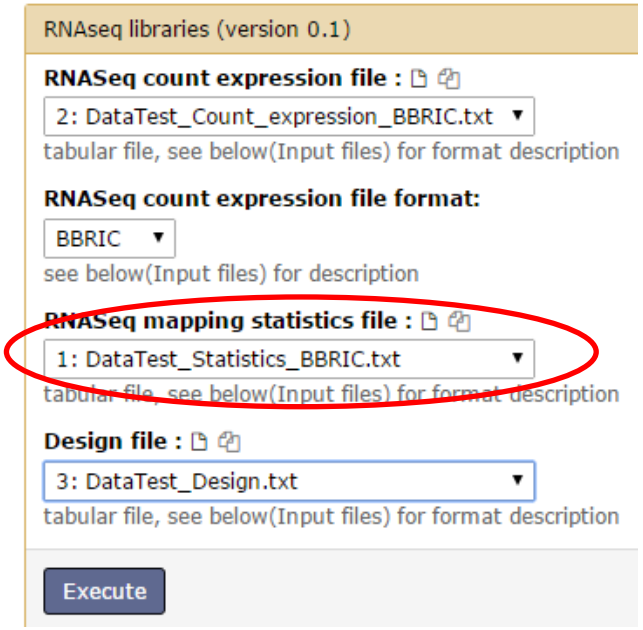
Format du fichier design

lib	Strain
Mutant1_M2	Mutant1
Mutant1_M3	Mutant1
Mutant1_M4	Mutant1
Mutant2_M1	Mutant2
Mutant2_M2	Mutant2
Mutant2_M3	Mutant2
Mutant2_M4	Mutant2
Mutant3_M3	Mutant3
Mutant3_M4	Mutant3
Mutant3_M1	Mutant3
Mutant3_M2	Mutant3
WT_M1	WT
WT_M2	WT
WT_M3	WT
WT_M4	WT



Description du Workflow





Fichier de comptage

lib	specific_ hits	mapping_hits	raw_reads/ pairs_count	mapped_reads/ pairs_count	feature_overlapp ing_hits	feature_overlapping_hits /specific_hits_percent
S.bbric_Rb mLong_GGK 21.ope	1178	1181	9697	1179	1153	97.88
S.bbric_Rb mSmall_GG K36.ope	24405	24411	64272	24407	20674	84.71

Nombre d'alignements spécifiques sur génome

Nombre d'alignements sur génome

Nombre de lectures brutes

Nombre de lectures alignées sur le génome

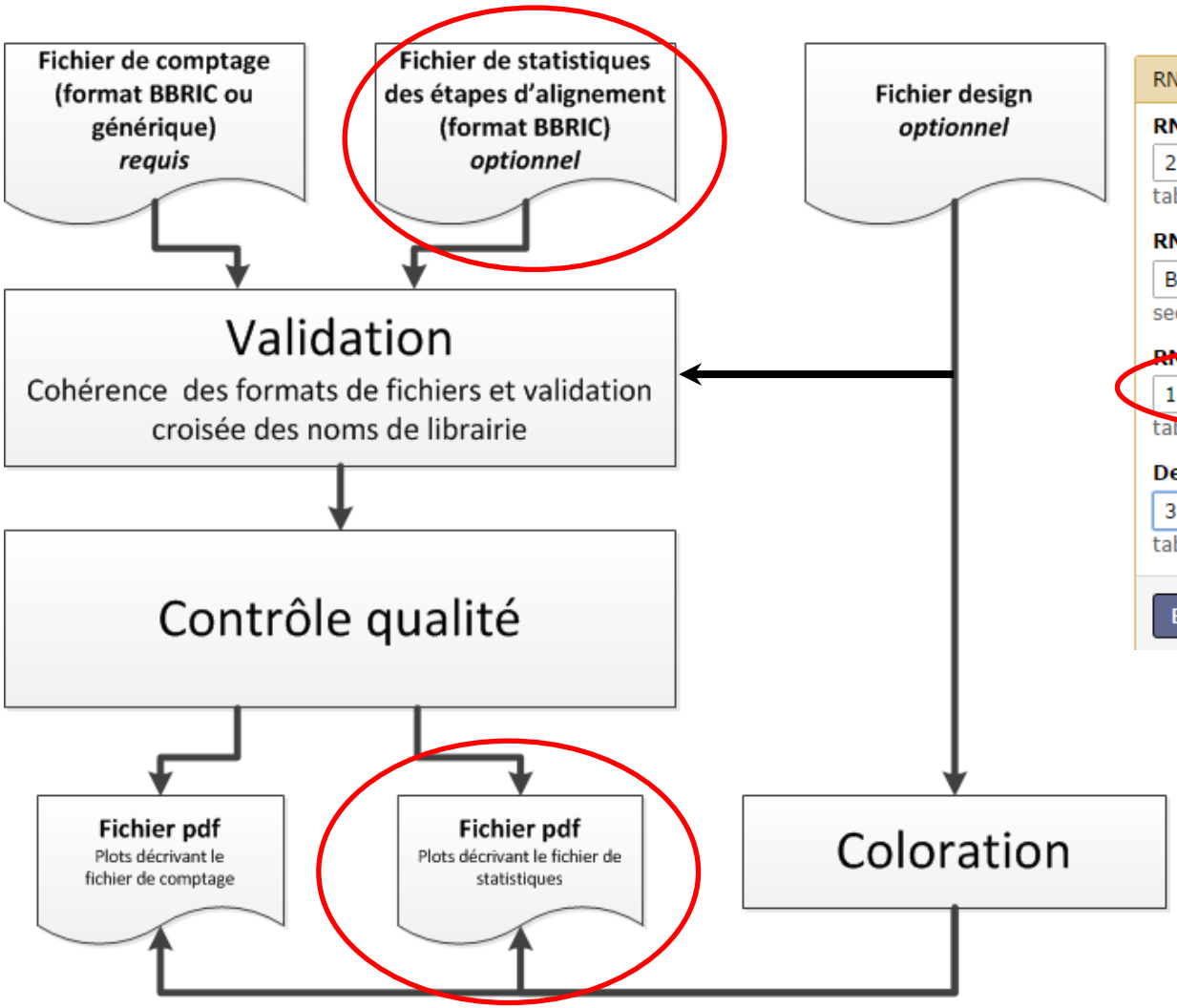
Nombre d'alignements sur objet biologique

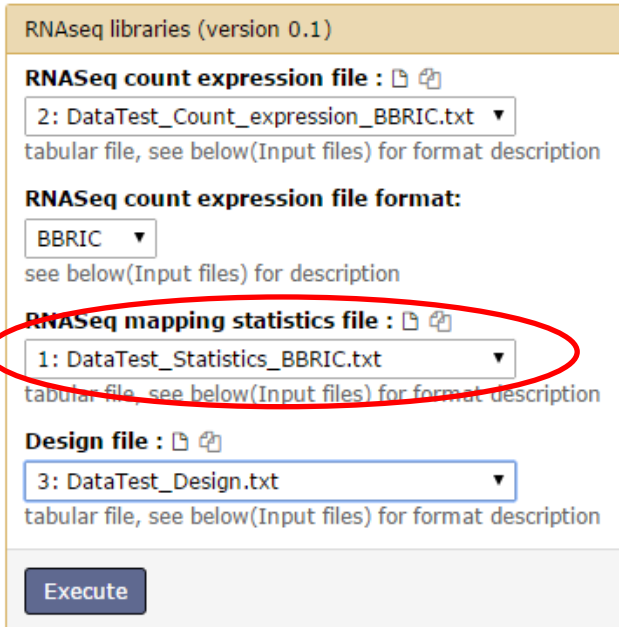
alignements sur objet biologique / alignements spécifiques sur génome

1181 hits totaux

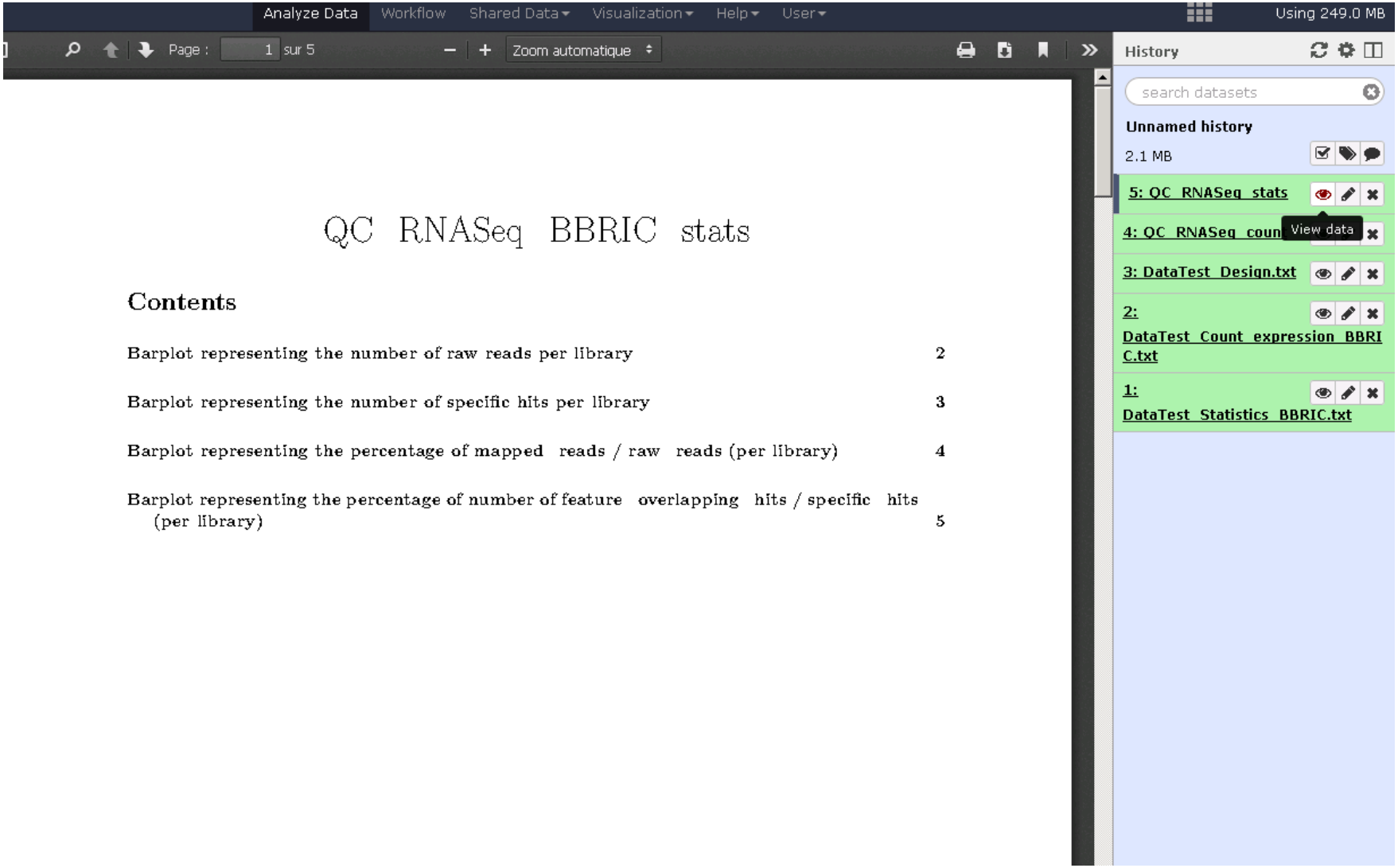
1179 lectures qui s'alignent

Description du Workflow

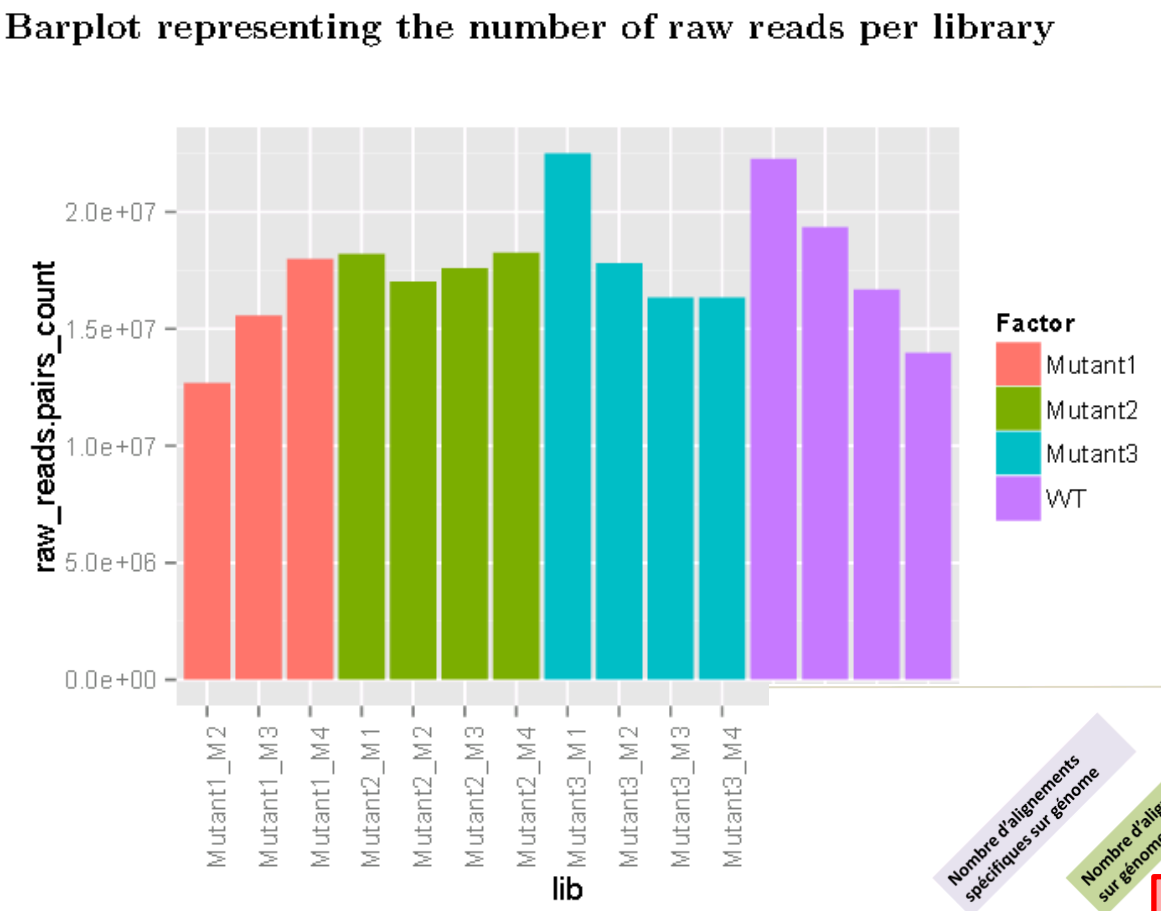




Plots à partir du fichier de comptage



Nombre de lectures brutes par librairie



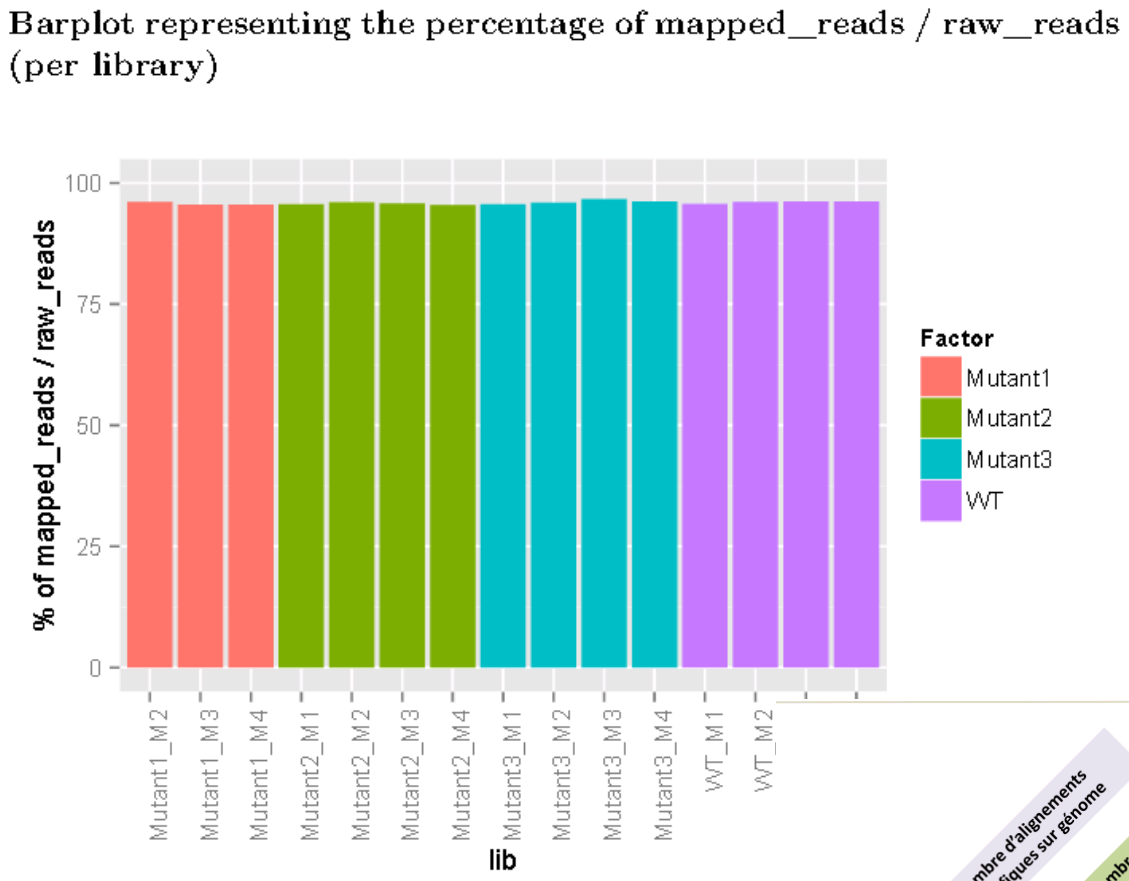
Objectif :
Vérifie que le nombre de reads est équivalent entre les échantillons pour repérer les problèmes expérimentaux

Question :
Le prestataire a-t-il bien fourni le nombre de reads « garanti » pour tous les échantillons ?

Nombre d'alignements spécifiques sur génome
Nombre d'alignements sur génome
Nombre de lectures brutes
Nombre de lectures alignées sur le génome
Nombre d'alignements sur objet biologique
alignements sur objet biologique / alignements spécifiques sur génome

lib	specific_ hits	mapping_hits	raw_reads/ pairs_count	mapped_reads/ pairs_count	feature_overlapp ing_hits	feature_overlapping_hits /specific_hits_percent
S.bbric_Rb mLong_GGK 21.ope	1178	1181	9697	1179	1153	97.88
S.bbric_Rb mSmall_GG K36.ope	24405	24411	64272	24407	20674	84.71

Proportion de lectures mappées



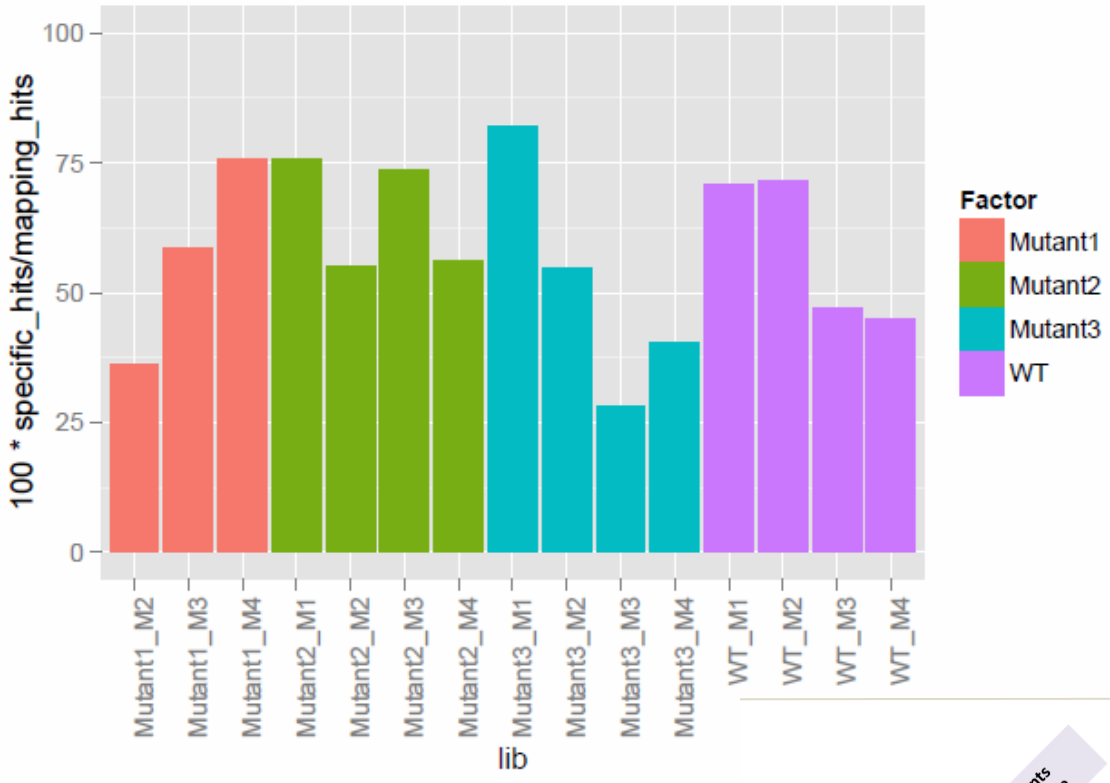
Objectif :
Vérifie que la qualité du mapping est la même pour toutes les librairies.
Si le pourcentage est trop bas, peut être la preuve d'une contamination

Questions :
• Un échantillon est-il contaminé par un autre organisme ?
• Les paramètres de mapping sont-ils corrects pour tous les échantillons ?

lib	specific_ hits	mapping_hits	raw_reads/ pairs_count	mapped_reads/ pairs_count	feature_overlapp ing_hits	feature_overlapping_hits /specific_hits_percent
S.bbric_Rb mLong_GGK 21.ope	1178	1181	9697	1179	1153	97.88
S.bbric_Rb mSmall_GG K36.ope	24405	24411	64272	24407	20674	84.71

Nombre d'alignements spécifiques sur génome
Nombre d'alignements sur génome
Nombre de lectures brutes
Nombre de lectures alignées sur le génome
Nombre d'alignements sur objet biologique
alignements sur objet biologique / alignements spécifiques sur génome

Proportion de hits avec alignement spécifique



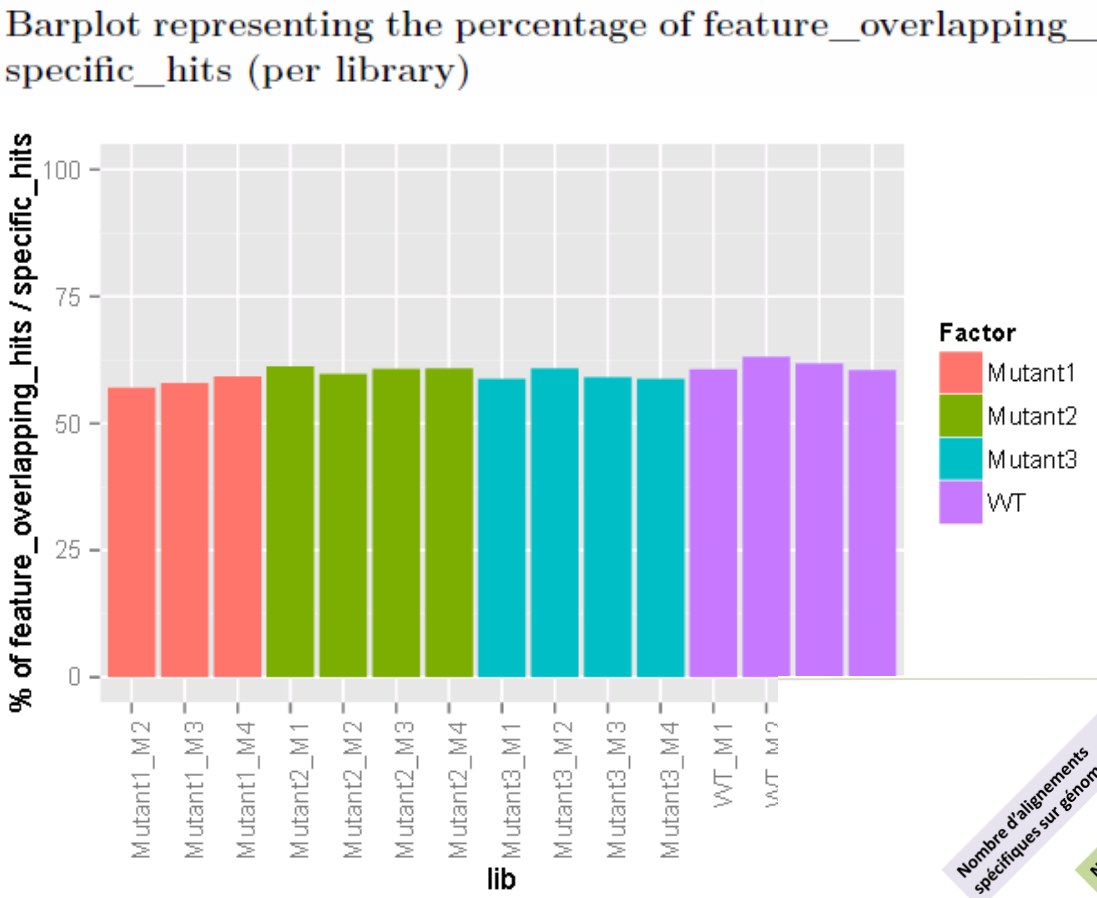
Objectif :
Déterminer la proportion de hits spécifiques (qui ne s'alignent qu'à un seul endroit dans le génome) par rapport à l'ensemble des hits mappant sur le génome

Questions :
- Les paramètres de mapping permettent-ils d'assurer des hits spécifiques ?
- Un échantillon est-il contaminé (ARN ribosomique par exemple) ?

Nombre d'alignements spécifiques sur génome
Nombre d'alignements sur génome
Nombre de lectures brutes
Nombre de lectures alignées sur le génome
Nombre d'alignements sur objet biologique
alignements sur objet biologique / alignements spécifiques sur génome

lib	specific_hits	mapping_hits	raw_reads/pairs_count	mapped_reads/pairs_count	feature_overlapping_hits	feature_overlapping_hits/specific_hits_percent
S.bbric_RbmLong_GGK21.ope	1178	1181	9697	1179	1153	97.88
S.bbric_RbmSmall_GGK36.ope	24405	24411	64272	24407	20674	84.71

Proportion de hits recouvrant une feature



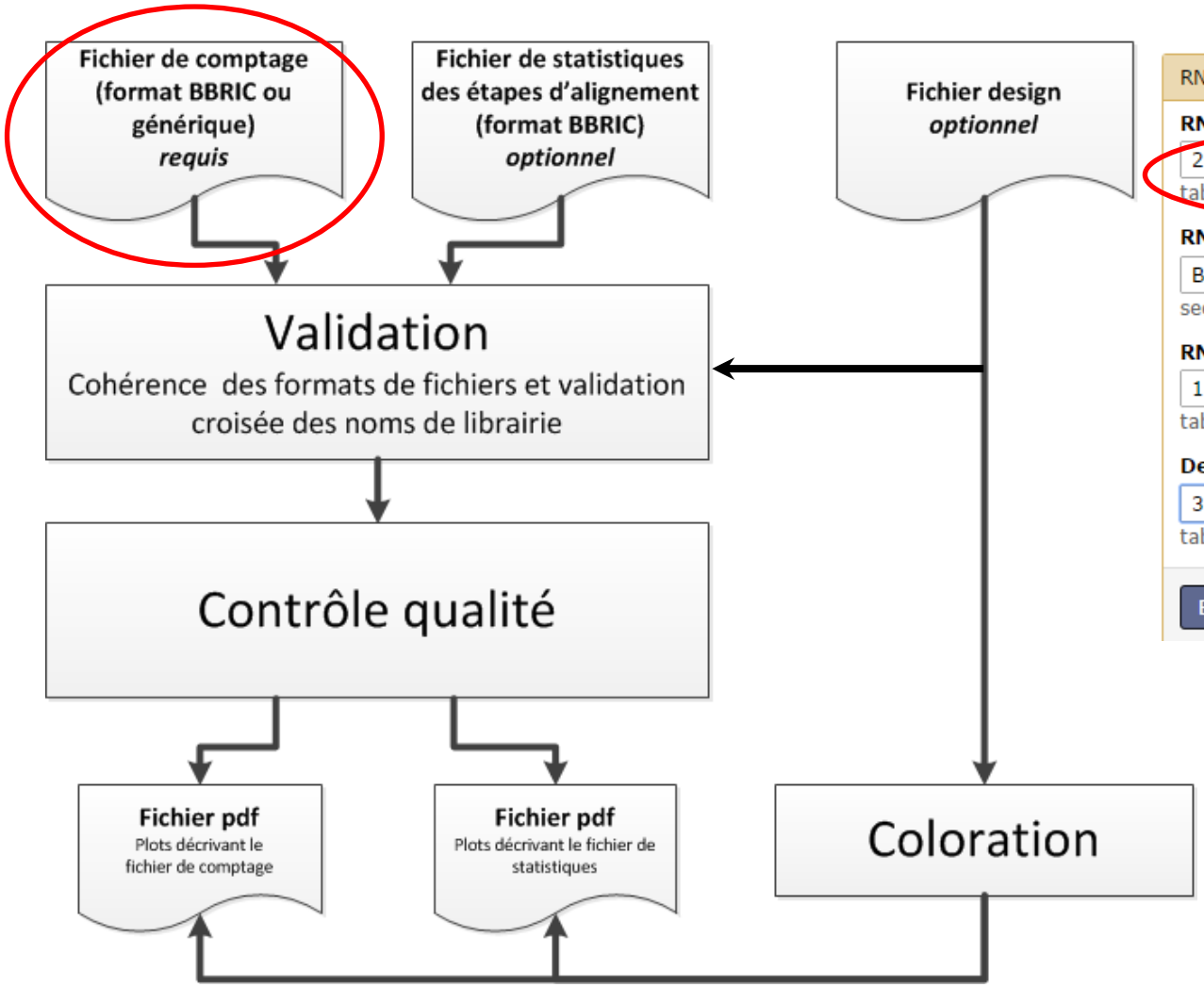
Objectif :
Vérifie que la qualité du mapping est la même pour toutes les librairies

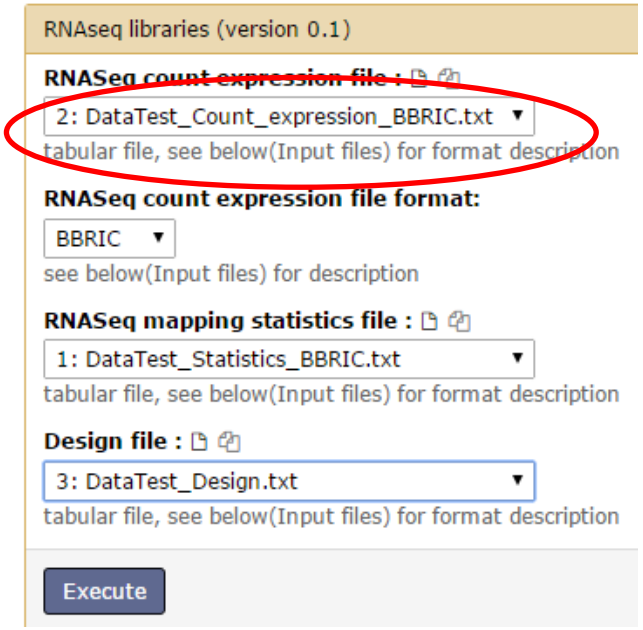
Questions :
- L'annotation structurale est-elle valide ?
- Les fichiers pair-end sont-ils utilisés dans la bonne orientation ?

lib	specific_ hits	mapping_hits	raw_reads/ pairs_count	mapped_reads/ pairs_count	feature_overlapp ing_hits	feature_overlapping_hits /specific_hits_percent
S.bbric_Rb mLong_GGK 21.ope	1178	1181	9697	1179	1153	97.88
S.bbric_Rb mSmall_GG K36.ope	24405	24411	64272	24407	20674	84.71

Nombre d'alignements spécifiques sur génome
Nombre d'alignements sur génome
Nombre de lectures brutes
Nombre de lectures alignées sur le génome
Nombre d'alignements sur objet biologique
alignements sur objet biologique / alignements spécifiques sur génome

Description du Workflow





Fichier de comptage (format BBRIC)

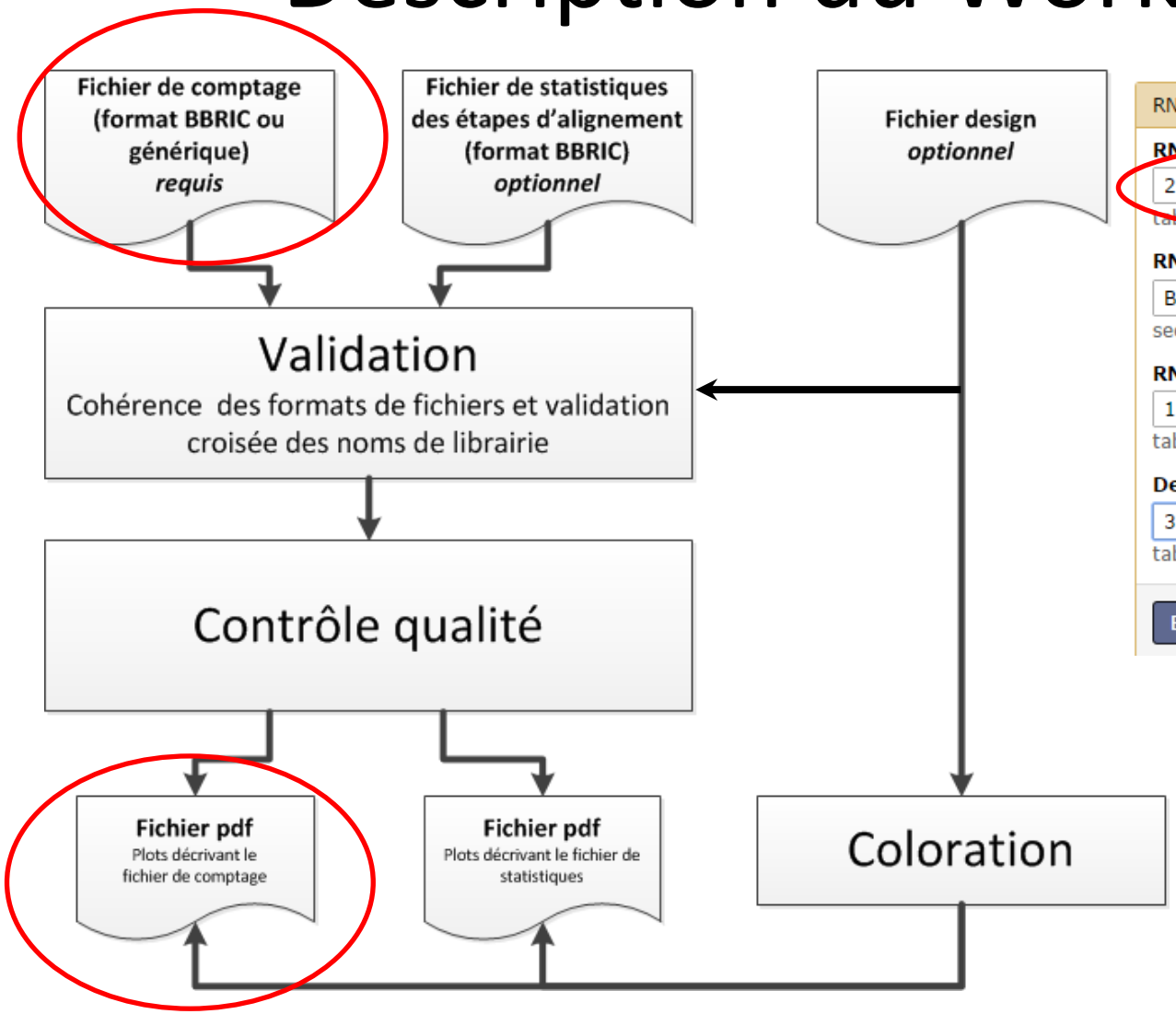
Gene/RNA ID
Contig Id
Positionnement du gène/RNA sur le contig
Début
Fin
Brin
Taille du Gène/RNA
Type (Gène/RNA)
Comptage brut (Lectures/Objet)
Comptage normalisé (RPKM)

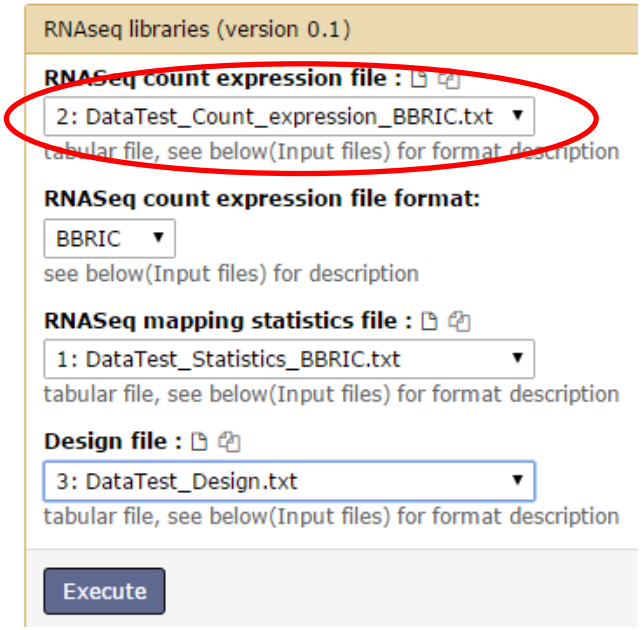
id	0_seqid	1_start	2_end	3_strand	4_length	5_type	6_Note	S.bbric_RbmLong_GGK21.ope-count	S.bbric_RbmLong_GGK21.ope-rpkm	S.bbric_RbmSmall_GGK36.ope-count	S.bbric_RbmSmall_GGK36.ope-rpkm
SBBRIC1.1	SBBRIC1	384	685	-	302	gene		5	14054.58	2	260.41
SBBRIC1.10	SBBRIC1	9162	11163	+	2002	gene		14	5936.34	107	2101.63
SBBRIC1.100	SBBRIC1	90298	90594	+	297	gene		0	0.00	9	1191.58

Fichier de comptage (format générique)

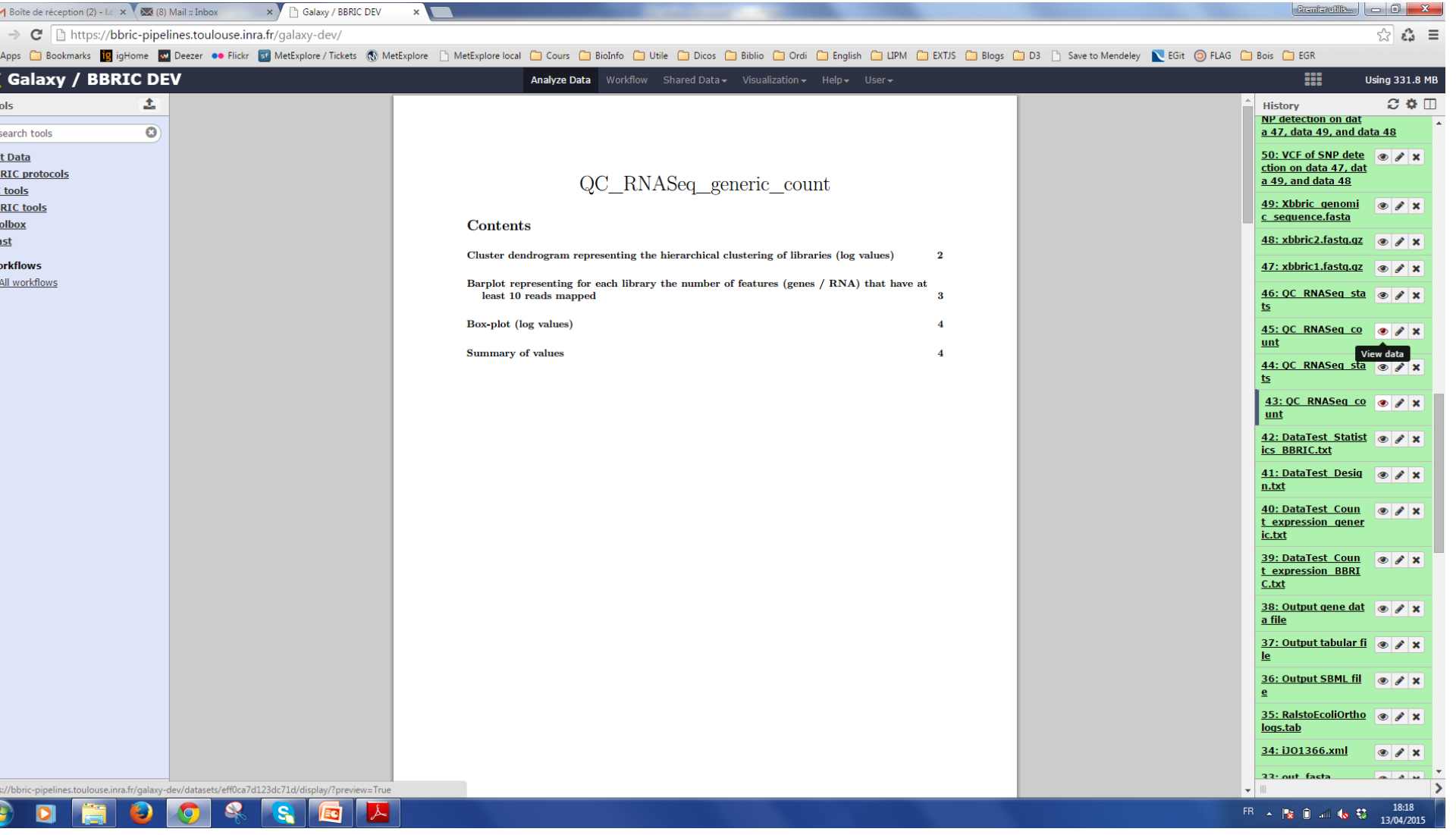
id	Lib1	Lib2
gene1	0	0
gene2	314	497
gene3	10991	11553

Description du Workflow



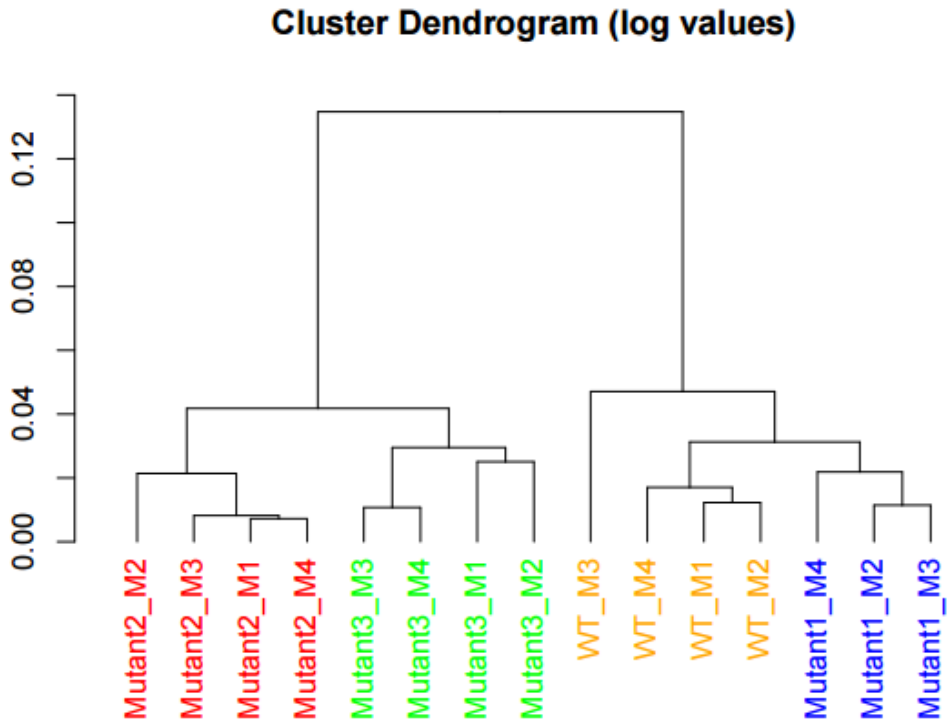


Résultats de comptage



Clustering hiérarchique des librairies

Cluster dendrogram representing the hierarchical clustering of libraries (log values)



Objectif:

• Vérifier que les duplicats sont bien dans le même groupe

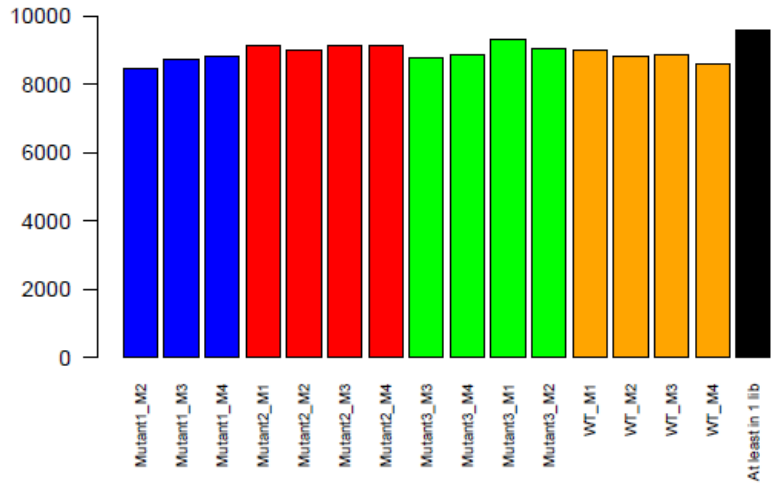
• Classification des conditions entre elles

Question:

Y-a-t-il une inversion des échantillons ?

Barplot representing the number of features mapped

Barplot representing for each library the number of features (genes / RNA) that have at least 10 reads mapped



	lib	nb features with at least 10 reads	nb features total
1	Mutant1_M2	8471	10345
2	Mutant1_M3	8730	10345
3	Mutant1_M4	8803	10345
4	Mutant2_M1	9112	10345
5	Mutant2_M2	9001	10345
6	Mutant2_M3	9128	10345
7	Mutant2_M4	9121	10345
8	Mutant3_M3	8796	10345
9	Mutant3_M4	8872	10345
10	Mutant3_M1	9300	10345
11	Mutant3_M2	9024	10345
12	WT_M1	9006	10345
13	WT_M2	8815	10345
14	WT_M3	8857	10345
15	WT_M4	8605	10345
16	At least in 1 lib	9563	10345

Objectifs :

• Permet de connaître le nombre de transcrits exploitables par librairie

• Permet de valider la prédiction des transcrits

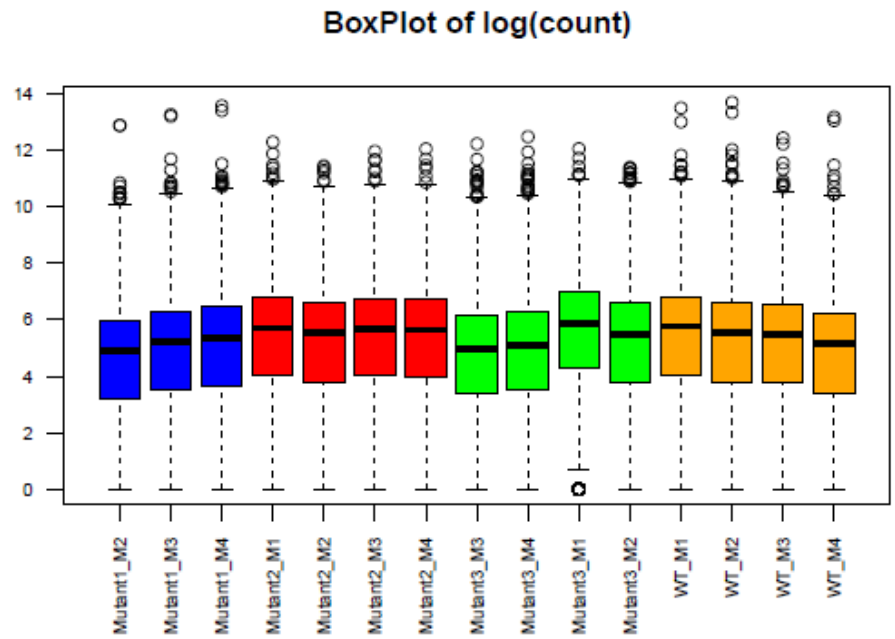
• Permet de repérer un problème expérimental si l'histogramme est déséquilibré

Question :

Y-a-t-il des soucis avec l'annotation structurale?

Trancrits prédits sans reads : 10345 – 9563 = 782

Box plot du nombre de reads par feature



Summary of values

```
##    Mutant1_M2        Mutant1_M3        Mutant1_M4
##  Min.   :     0.0   Min.   :     0.0   Min.   :     0.0
##  1st Qu.:    23.0   1st Qu.:    32.0   1st Qu.:    37.0
##  Median :   134.0   Median :   189.0   Median :   211.0
##  Mean   :   529.2   Mean   :   751.1   Mean   :   936.7
##  3rd Qu.:   386.0   3rd Qu.:   548.0   3rd Qu.:   640.0
##  Max.   :390252.0   Max.   :573790.0   Max.   :780274.0
##    Mutant2_M1        Mutant2_M2        Mutant2_M3
##  Min.   :     0.0   Min.   :     0.0   Min.   :     0.0
##  1st Qu.:    54.0   1st Qu.:    43.0   1st Qu.:    54.0
##  Median :   297.0   Median :   254.0   Median :   289.0
##  Mean   :   982.1   Mean   :   838.7   Mean   :   936.1
##  3rd Qu.:   884.0   3rd Qu.:   744.0   3rd Qu.:   843.0
##  Max.   :216767.0   Max.   : 93211.0   Max.   :156509.0
```

Objectifs :

• Vérifier que la distribution des valeurs de comptage est homogène à travers les librairies

• Permet de vérifier que le regroupement par la classification hiérarchique n'est pas dû à une ressemblance de la distribution des valeurs

Question :

• La distribution des valeurs de comptage nous permet-elle de lancer une analyse statistique efficace ?

Je veux contrôler la qualité de mes banques RNAseq et de mes mesures d'expression

Je vais sur	L'outil permet	L'outil ne permet pas

Paramètres clés	Pièges	Formats et fichiers

ANALYSE COMPARATIVE DE PROTEOMES D'ESPECES APPARENTEES (ORTHOMCL)

Experts : Martial Briand, Sébastien Carrère, Ludovic Cottret, Corinne Rancurel

La génomique comparative

C'est déterminer ce qui est **spécifique** à chaque espèce et ce qui est **commun** à chaque espèce

La comparaison génomique sous-entend que les génomes à comparer ont un ancêtre commun :
C'est la recherche de parenté entre les gènes de différents génomes

Un génome n'est pas statique :
il évolue au cours du temps avec des additions et des pertes de gènes

⇒ Il y a une évolution des gènes ancestraux avec des différences qualitatives (changement de la nature des gènes) et des différences quantitatives (variations du nombre de gènes)

C'est comparer les gènes (protéines) homologues entre eux au niveau de leur séquence

→ Identifier les gènes orthologues et paralogues

la motivation première

C'est inférer des connaissances sur une séquence
à partir des connaissances attachées à une autre

La génomique comparative

Quelques applications :

- constituer des familles de gènes
- identifier les gènes uniques de chaque génome
- aider à l'annotation fonctionnelle
- étudier l'expansion/ réduction de familles de gènes au cours du temps
- identifier les orthologues de gènes décrits chez les organismes modèles (ex. Arabidsopsis, *C. elegans*, Drosophile) dans nos organismes d'intérêt (ex. Légumineuses, Meloidogyne, insectes ravageurs)
- ...

Qu'est-ce qu'une relation d'orthologie

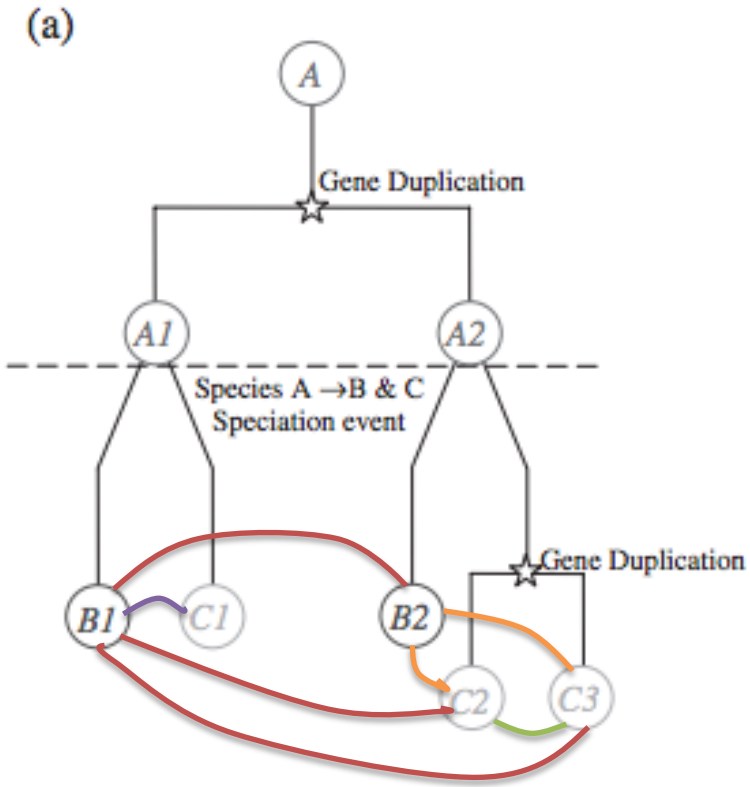
Définitions :

Paralogues : gènes qui ont co-évolué suite à un événement de duplication dans un génome

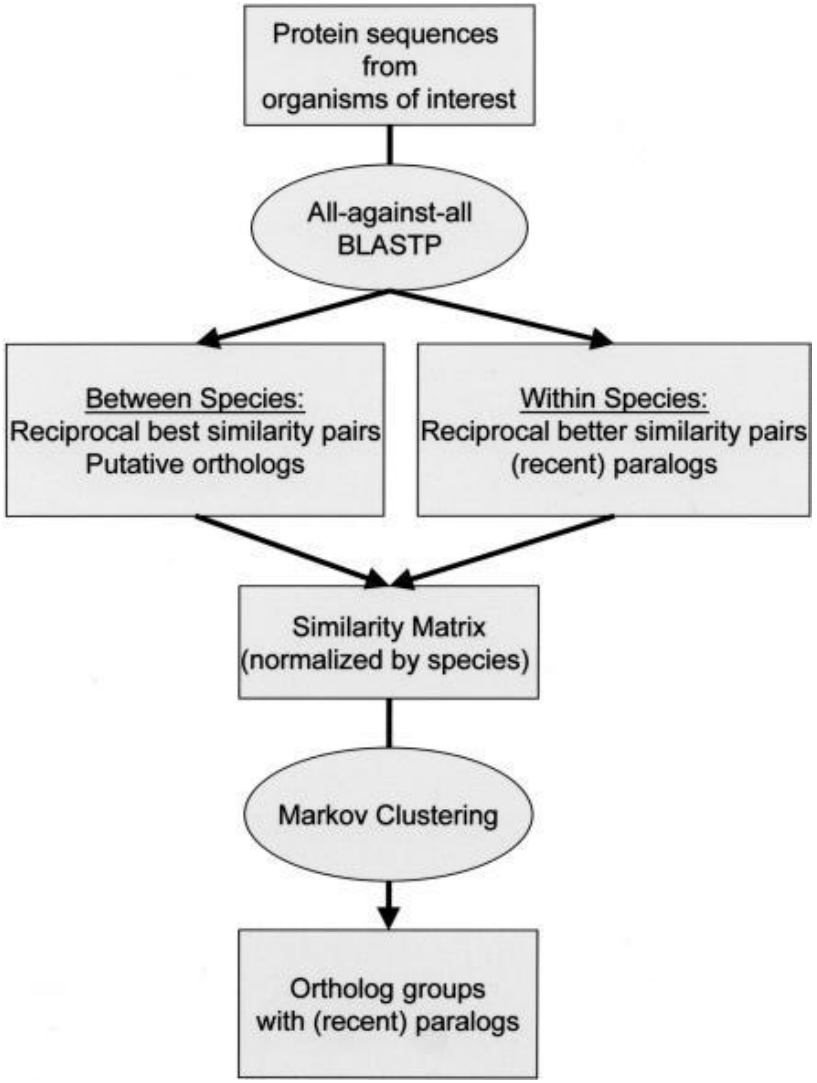
Orthologues : gènes qui ont un même ancêtre commun et qui ont co-évolué suite à un événement de spéciation

In-paralogues : paralogues qui ont été dupliqués suite à un événement de spéciation, ils sont alors co-orthologues aux gènes de l'autre espèce

Out-paralogues : paralogues qui ont été dupliqués avant un événement de spéciation (ils n'appartiennent pas nécessairement à la même espèce)



L'outil OrthoMCL : comment ça marche

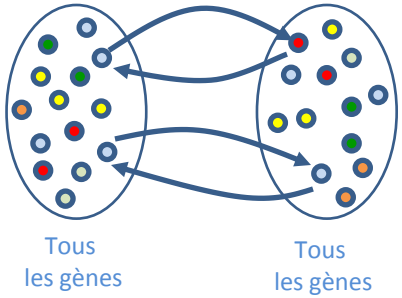


Li et al., 2003

OrthoMCl est un logiciel qui construit des clusters d'orthologues (incluant les in-paralogues récents) à partir de fichiers multifasta contenant des CDS.

La méthode utilisée est basée sur :

la méthode du « blast hit réciproque » (BHR)
recherche des meilleures similitudes réciproques

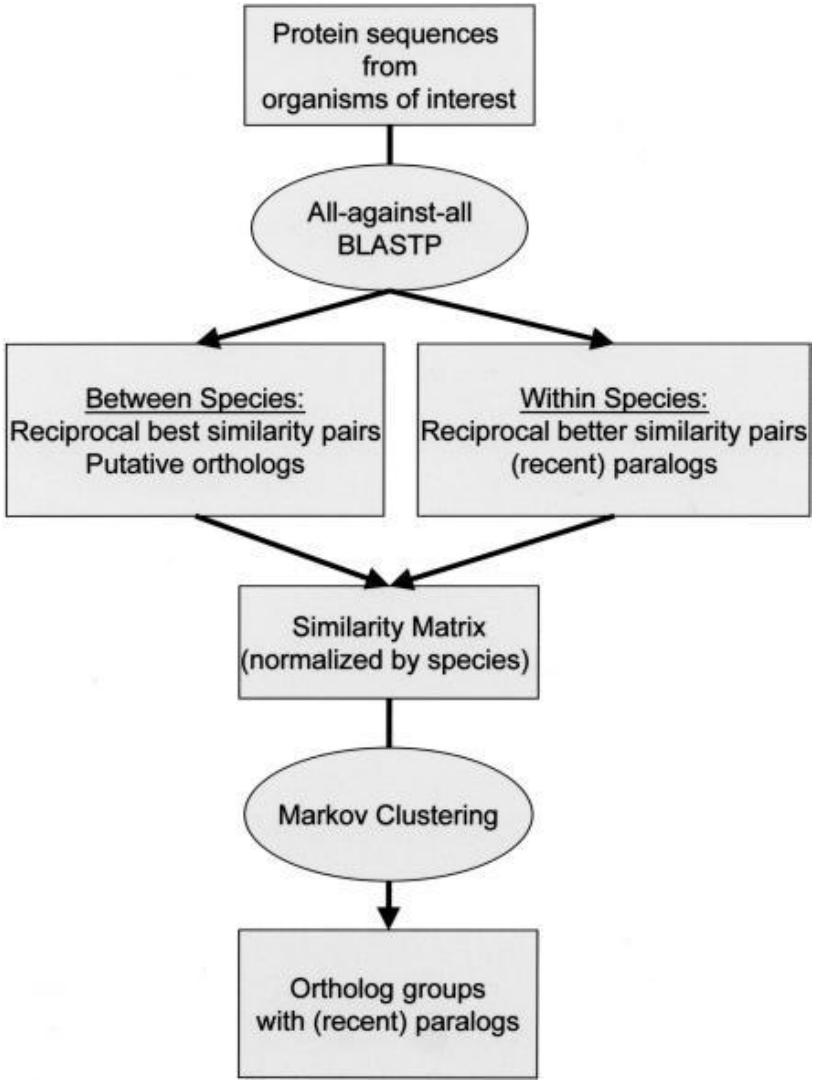


+

l'algorithme MCL (Markov Cluster Algorithm) pour trouver des regroupements d'éléments similaires

Approche = création du graphe de similarité puis partitionnement (clustering)

L'outil OrthoMCL : paramétrage



Li et al., 2003

pv_cutoff=<Float> **1e-5 (DEFAULT)**

P-Value or E-Value Cutoff in BLAST search and/or ortholog clustering

=> Plus la e-value est faible, plus l'alignement est significatif. On fixe un seuil ("*cutoff*") au dessus duquel on ne veut pas conserver les résultats.

pi_cutoff=<Int> **0 (DEFAULT)**

Percent Identity Cutoff <0-100> in ortholog clustering

=> Utiliser pour éliminer les hits avec de petites valeurs de pourcentage d'identité (pi). Le pi est la moyenne des pi de tous les hsps pour cette paire (query-subject).

pmatch_cutoff=<Int> **0 (DEFAULT)**

Percent Match Cutoff <0-100> in ortholog clustering

=> Pour former une paire, le nombre d'acide aminés de la protéine la plus courte doit être inclus d'au moins **pmatch_cutoff %** de la longueur de l'autre séquence.

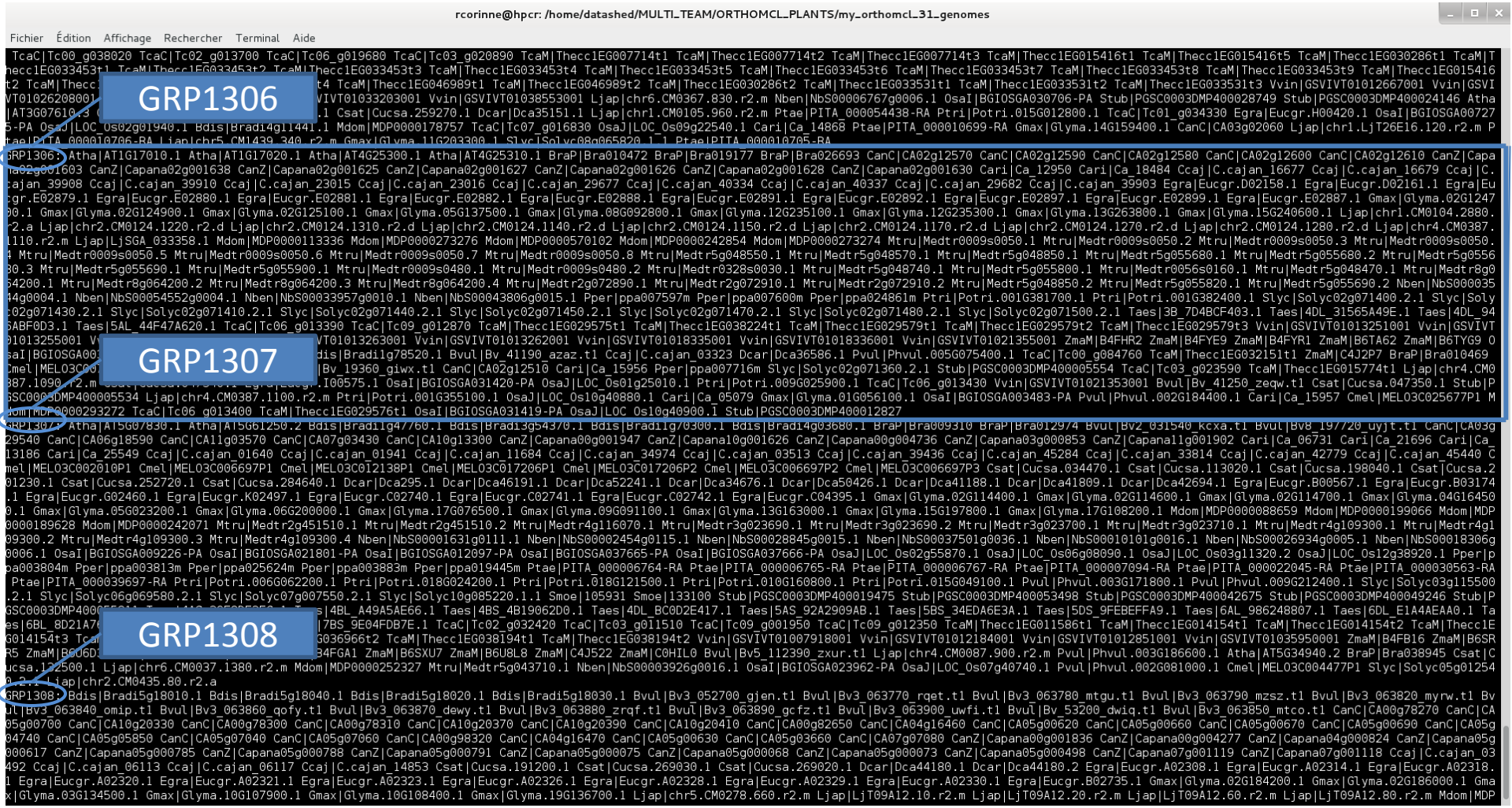
inflation=<Float> **1.5 (DEFAULT)**

Markov Inflation Index, used in MCL algorithm,
Increasing this index increases cluster tightness, and the number of clusters

=> Une valeur de 1,5 donne un équilibre entre sensibilité et sélectivité.

Le résultat brut d'OrthoMCL « orthomcl.out »

orthomcl-companion



Pas très digeste …

Focus sur le fichier « orthomcl.out »

Extrait de quelques groupes d'un fichier orthomcl.out

GRP22264: Mdom|MDP0000120086 Mdom|MDP0000255483 Mdom|MDP0000287109 Mdom|MDP0000388642 Mdom|MDP0000554052 Mdom|MDP0000298540 Mdom|MDP0000570301

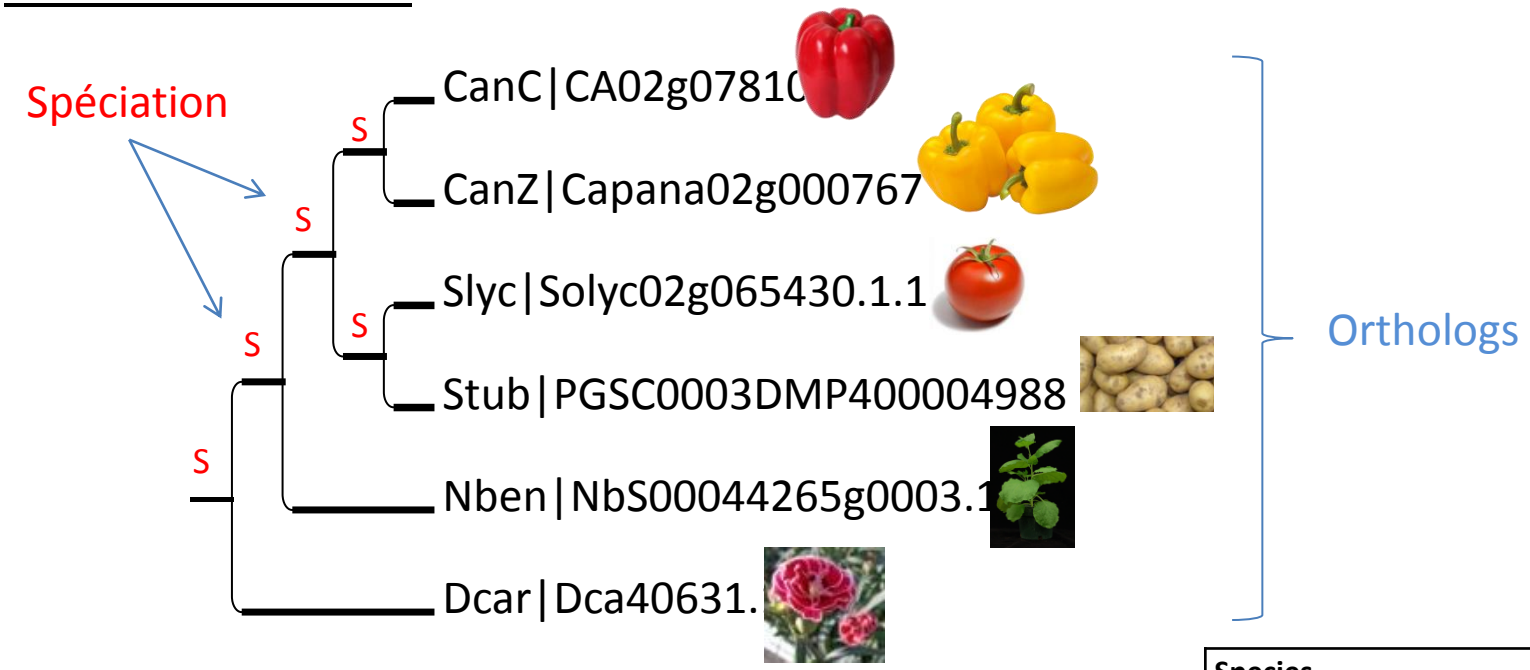
GRP26560: Nben|NbS00044265g0003.1 Dcar|Dca40631.1 CanC|CA02g07810 CanZ|Capana02g000767 Slyc|Solyc02g065430.1.1 Stub|PGSC0003DMP400004988

GRP27024: Bdis|Bradi4g23995.1 Bdis|Bradi4g24001.1 OsaI|BGIOSGA001239-PA Ljap|chr3.CM0786.230.r2.d OsaJ|LOC_Os11g15620.1

Exemple de groupes OrthoMCL obtenus - Traduction

GRP26560: Nben|NbS00044265g0003.1 Dcar|Dca40631.1 CanC|CA02g07810 CanZ|Capana02g000767 Slyc|Solyc02g065430.1.1 Stub|PGSC0003DMP400004988

Traduction en arbre:

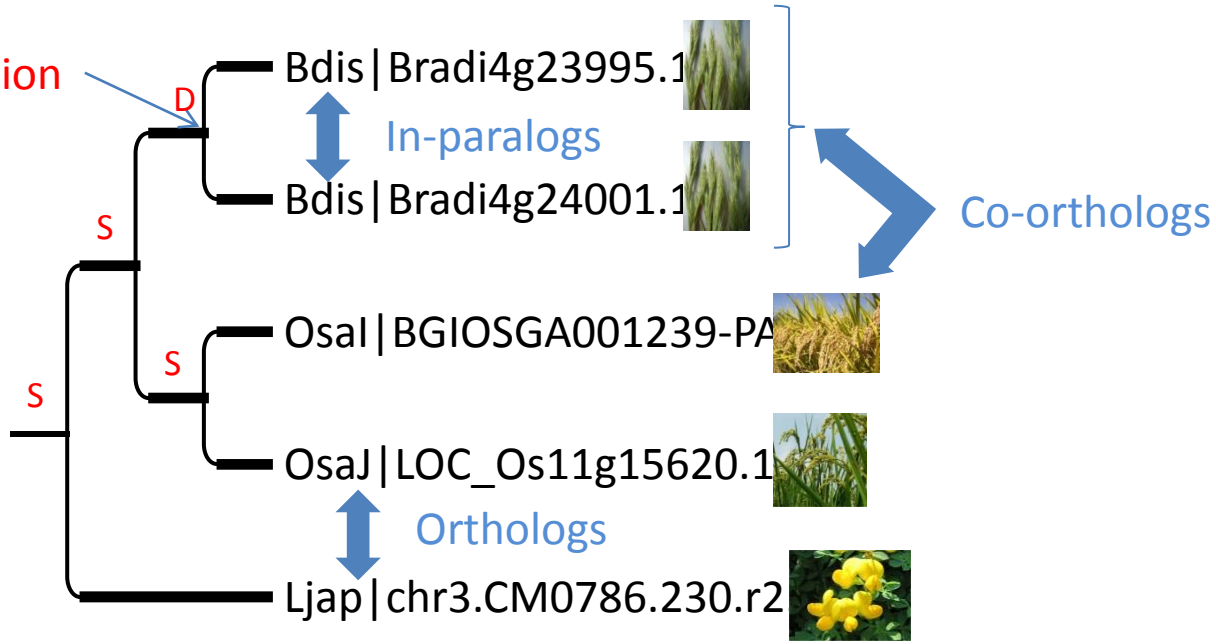


Species	Abbrev
Capsicum annuum C	CanC
Capsicum annuum Z	CanZ
Dianthus caryophyllus	Dcar
Nicotiana benthamiana	Nben
Solanum lycopersicum	Slyc
Solanum tuberosum	Stub

Exemple de groupes OrthoMCL obtenus - Traduction

GRP27024: Bdis|Bradi4g23995.1 Bdis|Bradi4g24001.1 OsaI|BGIOSGA001239-PA Ljap|chr3.CM0786.230.r2.d OsaJ|LOC_Os11g15620.1

Traduction en arbre:

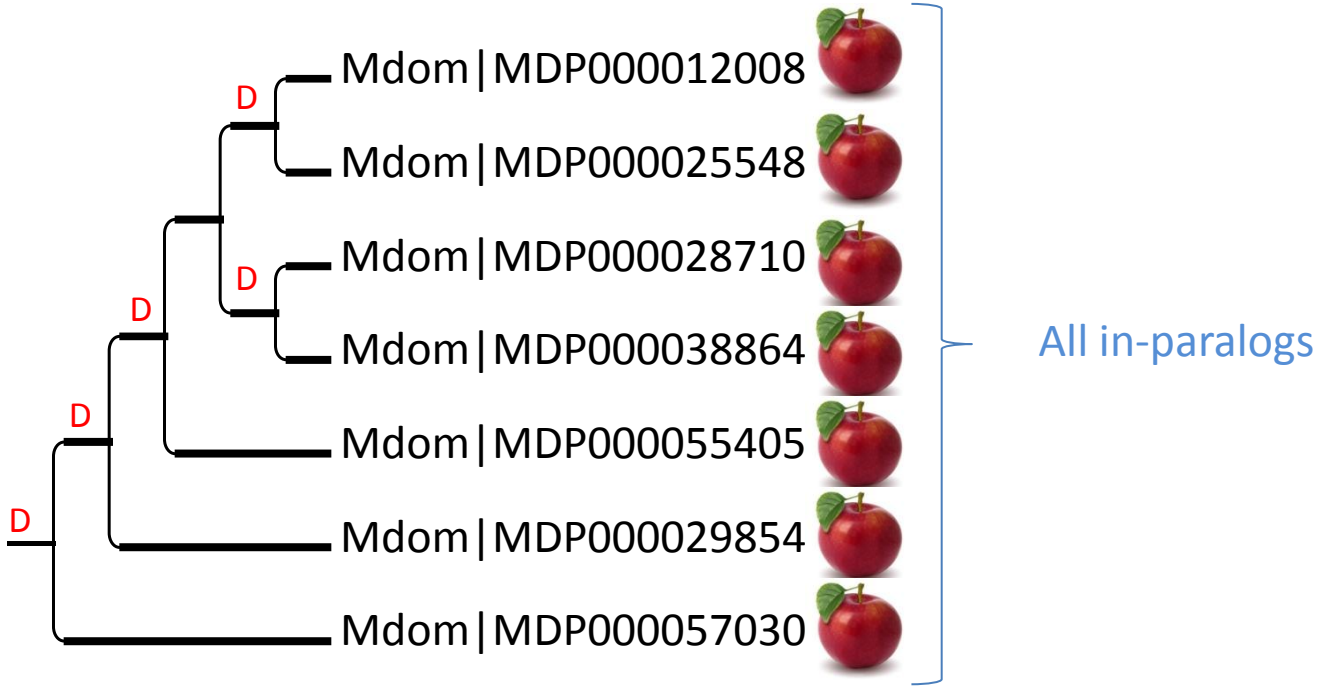


Species	Abbrev
Brachypodium distachyon	Bdis
Lotus japonicus	Ljap
Oryza sativa I	OsaI
Oryza sativa J	OsaJ

Exemple de groupes OrthoMCL obtenus - Traduction

GRP22264: Mdom|MDP0000120086 Mdom|MDP0000255483 Mdom|MDP0000287109 Mdom|MDP0000388642 Mdom|MDP0000554052 Mdom|MDP0000298540 Mdom|MDP0000570301

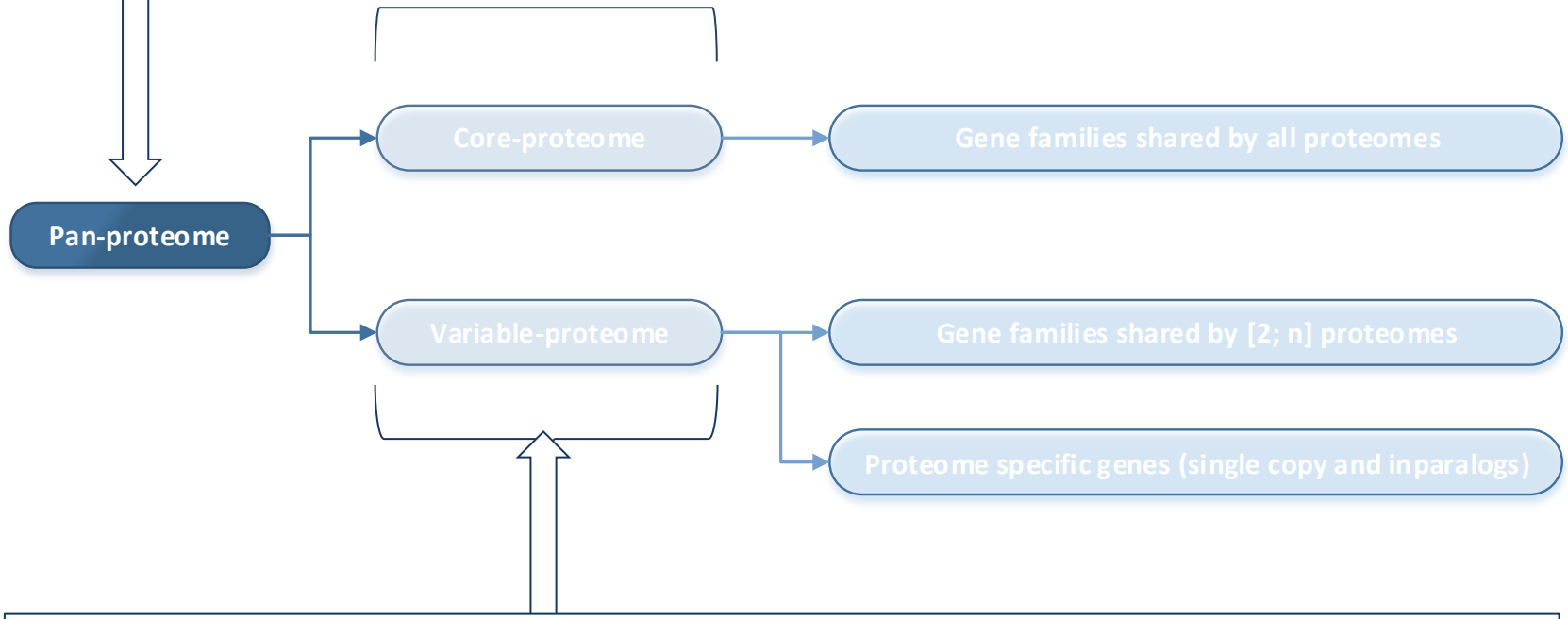
Traduction en arbre:



Species	Abbrev
Malus domestica	Mdom

Possibilité d'accéder au pan/core-genome

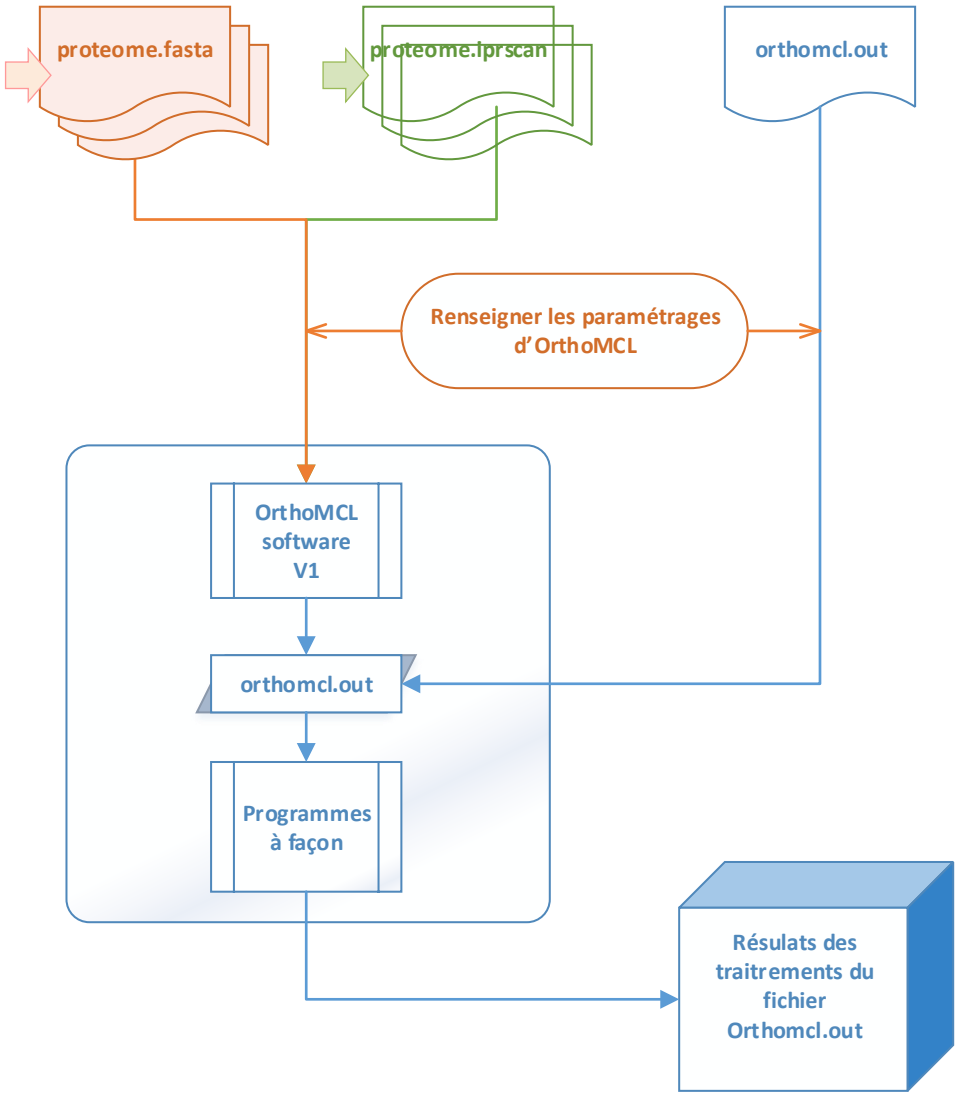
Pan = 1 représentant de chaque groupe orthomcl
+ toutes les protéines spécifiques en copie unique de toutes les espèces



Core = protéines <u>en copie unique</u> retrouvées chez tous les organismes

Variable = protéines partagées par au moins 2 organismes
+ les gènes spécifiques à chaque organisme

Présentation générale de l'outil



Données minimales :
les fichiers de séquences des protéomes

Valeur ajoutée :
Possibilité de fournir les fichiers d'annotation au format InterProScan

L'outil accepte en entrée un fichier orthomcl.out version : 1 ou 2

Nomenclature obligatoire :

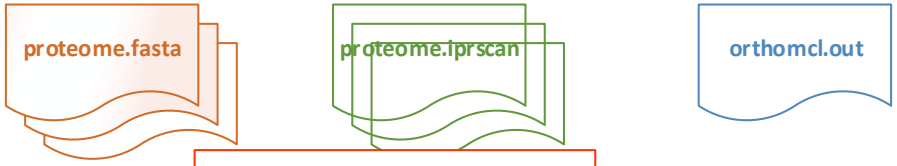
filename . extension

extension

.fa .faa .fas .fasta
.iprscan .ipr

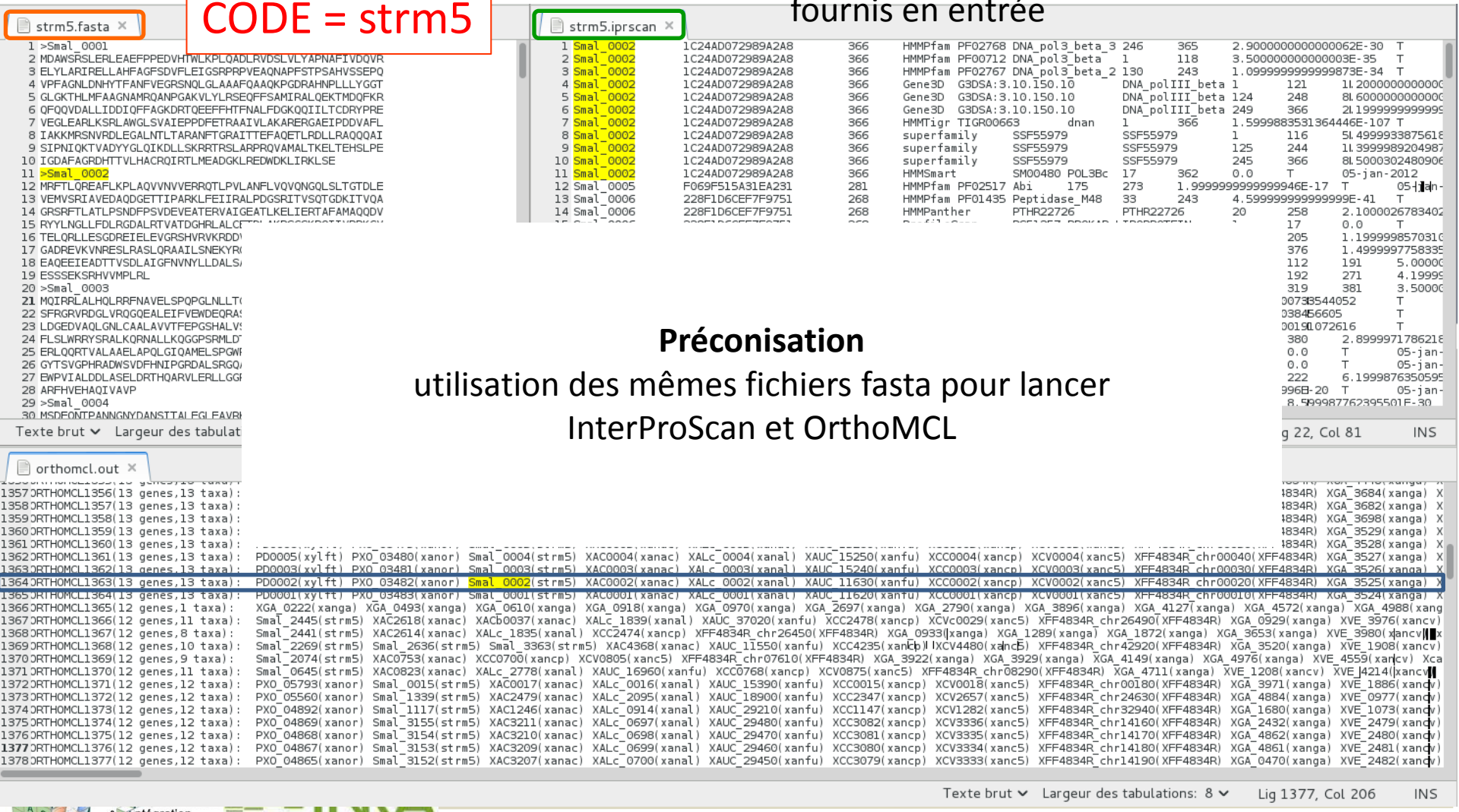
filename = **code court et concis**
Il sert d'identifiant de votre protéome

Présentation générale de l'outil



CODE = strm5

Obligatoirement: les mêmes identifiants (IDs) des séquences doivent être utilisés dans les 3 types de fichier pouvant être fournis en entrée



Préconisation

utilisation des mêmes fichiers fasta pour lancer InterProScan et OrthoMCL

L'interface

https://bbric-pipelines.toulouse.inra.fr/orthomcl-companion/web/index.html

Protein family analyses

OrthoMCL-Companion

login

Home

Perform OrthoMCL analyses, browse, visualise and share the results!

Nowadays, comparative genomics is a classical tool for biologists to address various scientific questions such as pathogeny determinism, bacteria host specificity, duplication and gene family expansion in a species.
The large number of available finished or draft genome sequences associated with the full automatisation of structural annotation pipelines leads to an era of high throughput comparative genomics.
Tools like **OrthoMCL** provide an easy way to compare tens of species at a time and build homologous gene families.
However, output results, classically a matrix with groups as lines and genes as columns, remains too raw to be interpreted by an end user without programming skills.
In order to help users in the post process of these data, we developed a web user interface that automatically extract frequently asked datasets such as core and pan proteomes, specific proteins, abundance and presence/absence matrix (aka. phylogenetic profiles).
Providing optional InterPro annotation will also lead to build tables and charts representing terms (eg : PFAM, GO) count in each previously described datasets.
Our tool uses standard formats as input and outputs to ease post-process analyses.

Test Dataset

3 proteomes of about 400 proteins each annotated with InterPro 51.0

Results

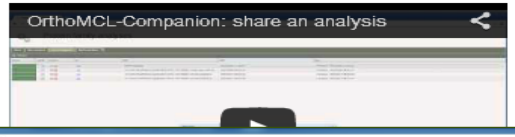
Test dataset analyses results (OrthoMCL with hight coverage parameter - pvmatch cut off = 80)

Analyse raw proteomes

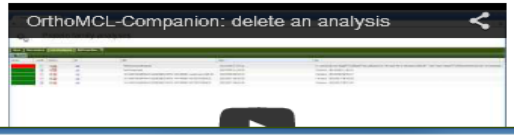
Analyse annotated proteomes

Browse

Share

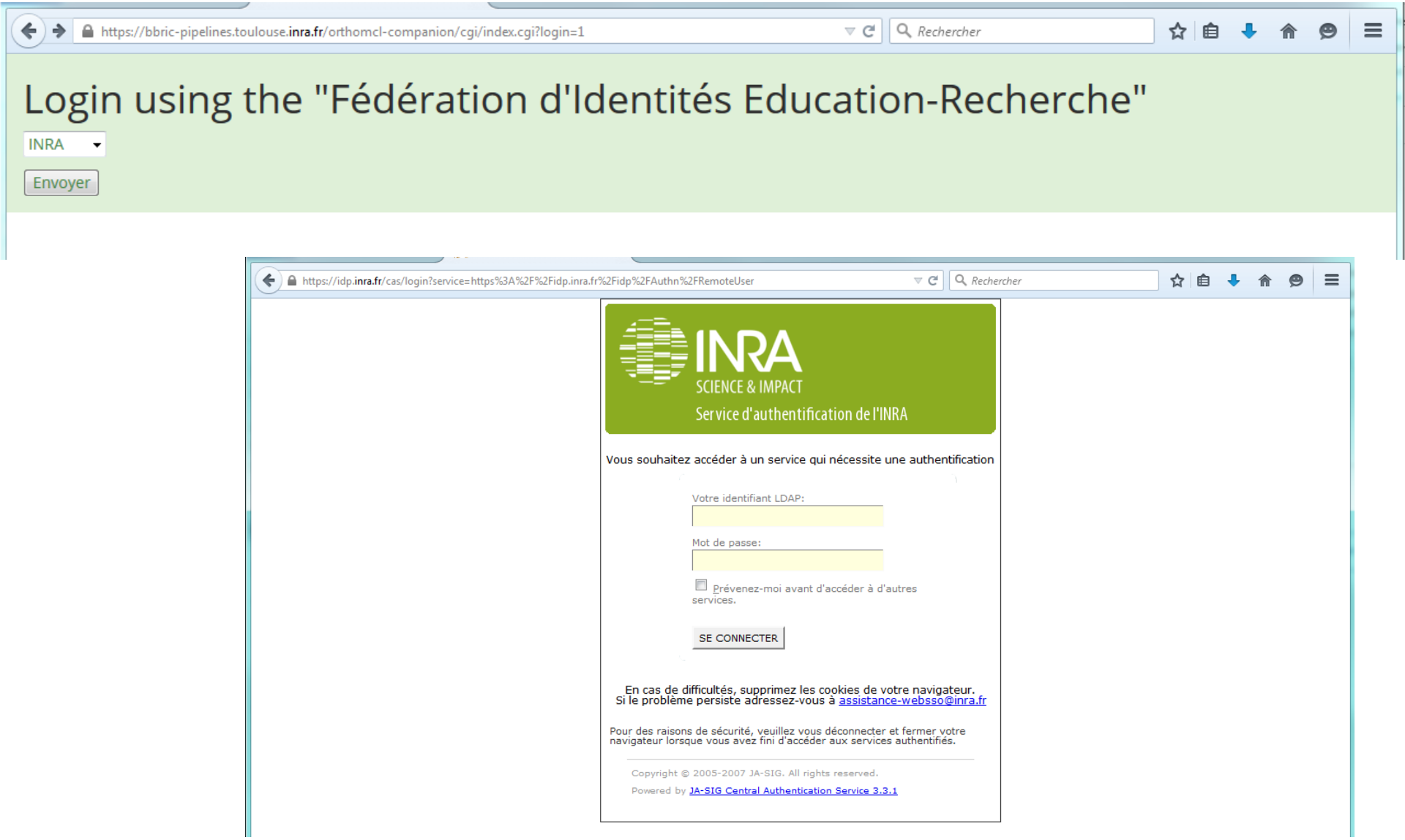


Delete



Fast launch

L'interface : se connecter



L'interface : se connecter

https://bbric-pipelines.toulouse.inra.fr/orthomcl-companion/web/index.html

Protein family analyses

OrthoMCL-Companion

Home | New analysis | ...es

Perform OrthoMCL analyses, browse, visualise and share the results!

Nowadays, comparative genomics is a classical tool for biologists to address various scientific questions such as pathogeny determinism, bacteria host specificity, duplication and gene family expansion in a species.
The large number of available finished or draft genome sequences associated with the full automatisation of structural annotation pipelines leads to an era of high throughput comparative genomics.
Tools like **OrthoMCL** provide an easy way to compare tens of species at a time and build homologous gene families.
However, output results, classically a matrix with groups as lines and genes as columns, remains too raw to be interpreted by an end user without programming skills.
In order to help users in the post process of these data, we developed a web user interface that automatically extract frequently asked datasets such as core and pan proteomes, specific proteins, abundance and presence/absence matrix (aka. phylogenetic profiles).
Providing optional InterPro annotation will also lead to build tables and charts representing terms (eg : PFAM, GO) count in each previously described datasets.
Our tool uses standard formats as input and outputs to ease post-process analyses.

Test Dataset

3 proteomes of about 400 proteins each annotated with InterPro 51.0

Results

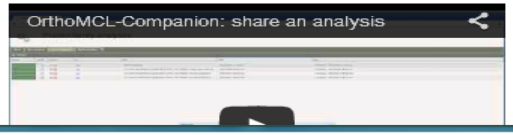
Test dataset analyses results (OrthoMCL with hight coverage parameter - pvmatch cut off = 80)

Analyse raw proteomes

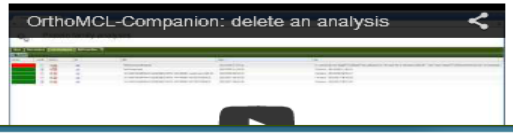
Analyse annotated proteomes

Browse

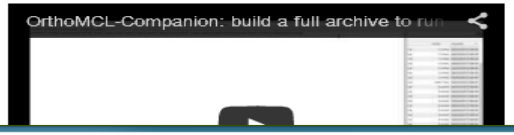
Share



Delete

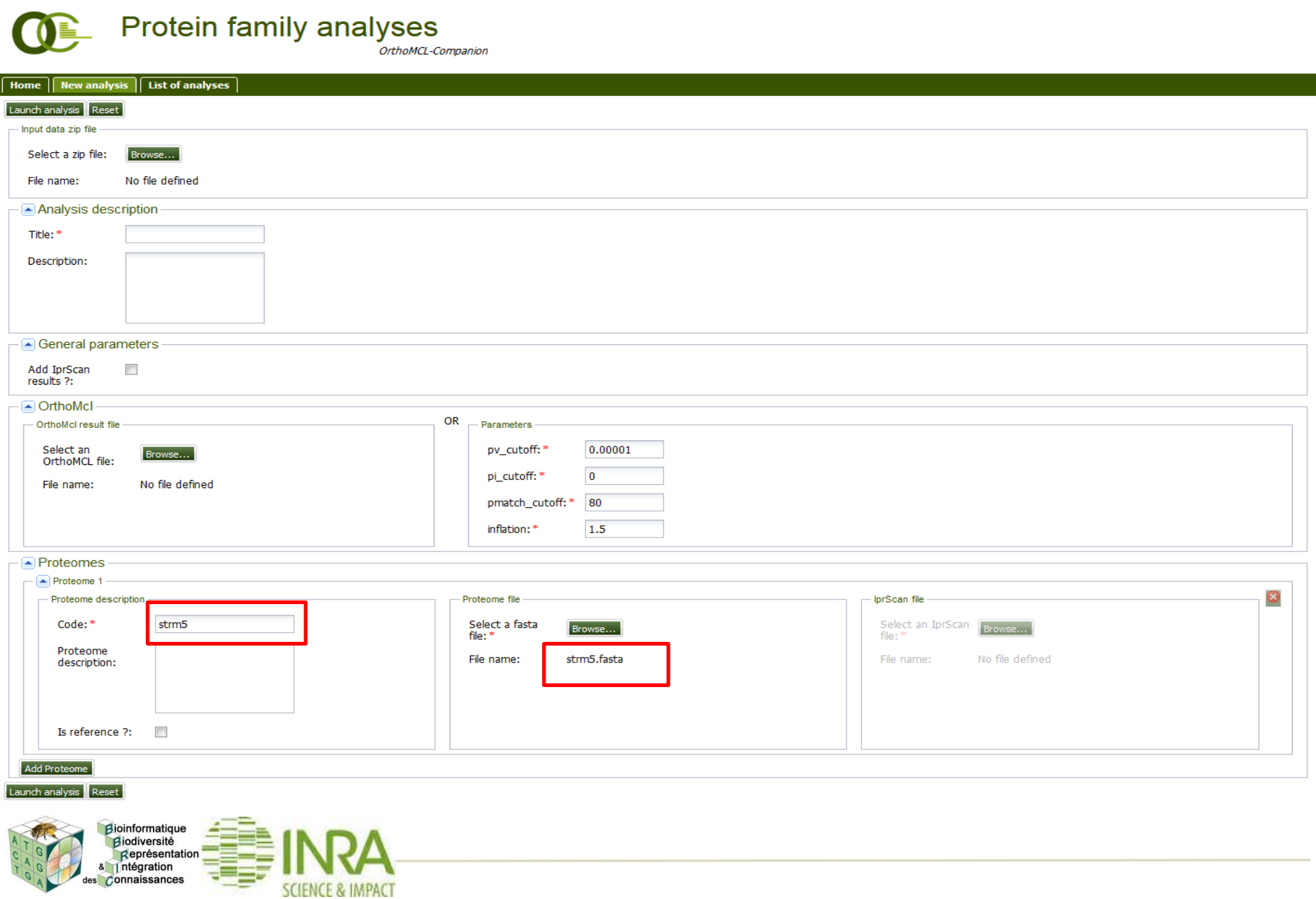


Fast launch



L'interface : lancer une analyse

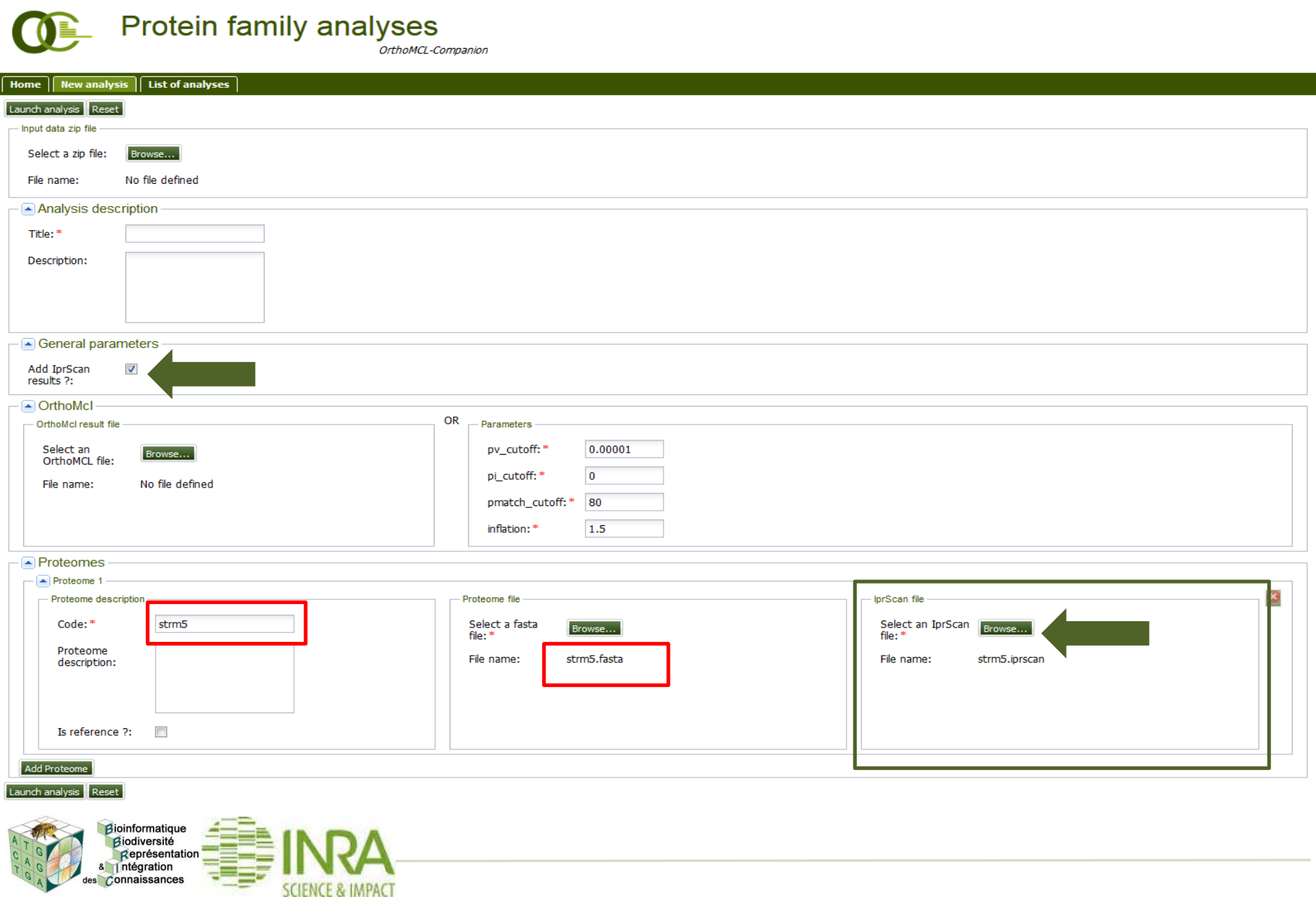
L'interface : lancer une analyse



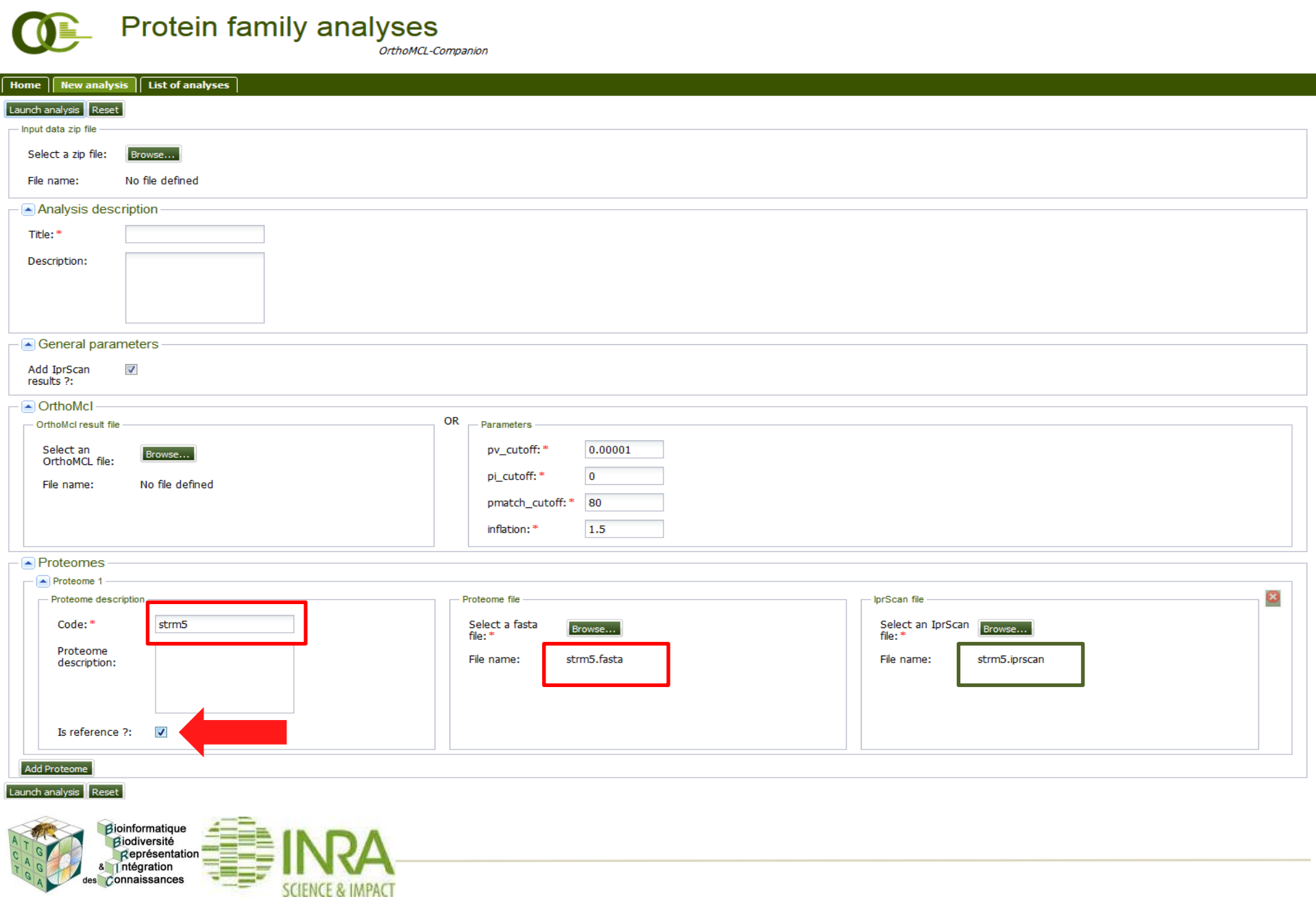
Bioinformatique Biodiversité Représentation & Intégration des Connaissances

INRA SCIENCE & IMPACT

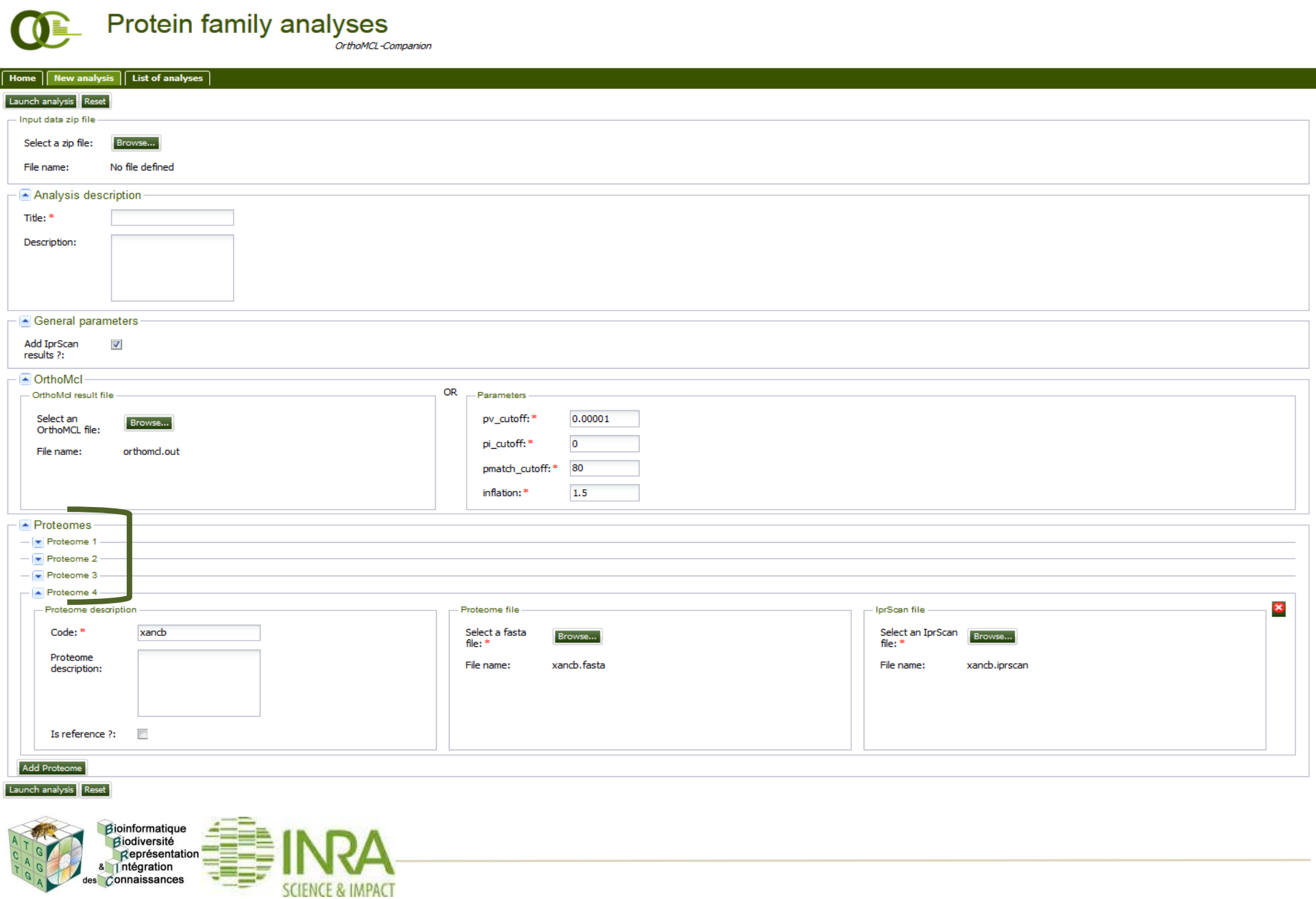
L'interface : lancer une analyse



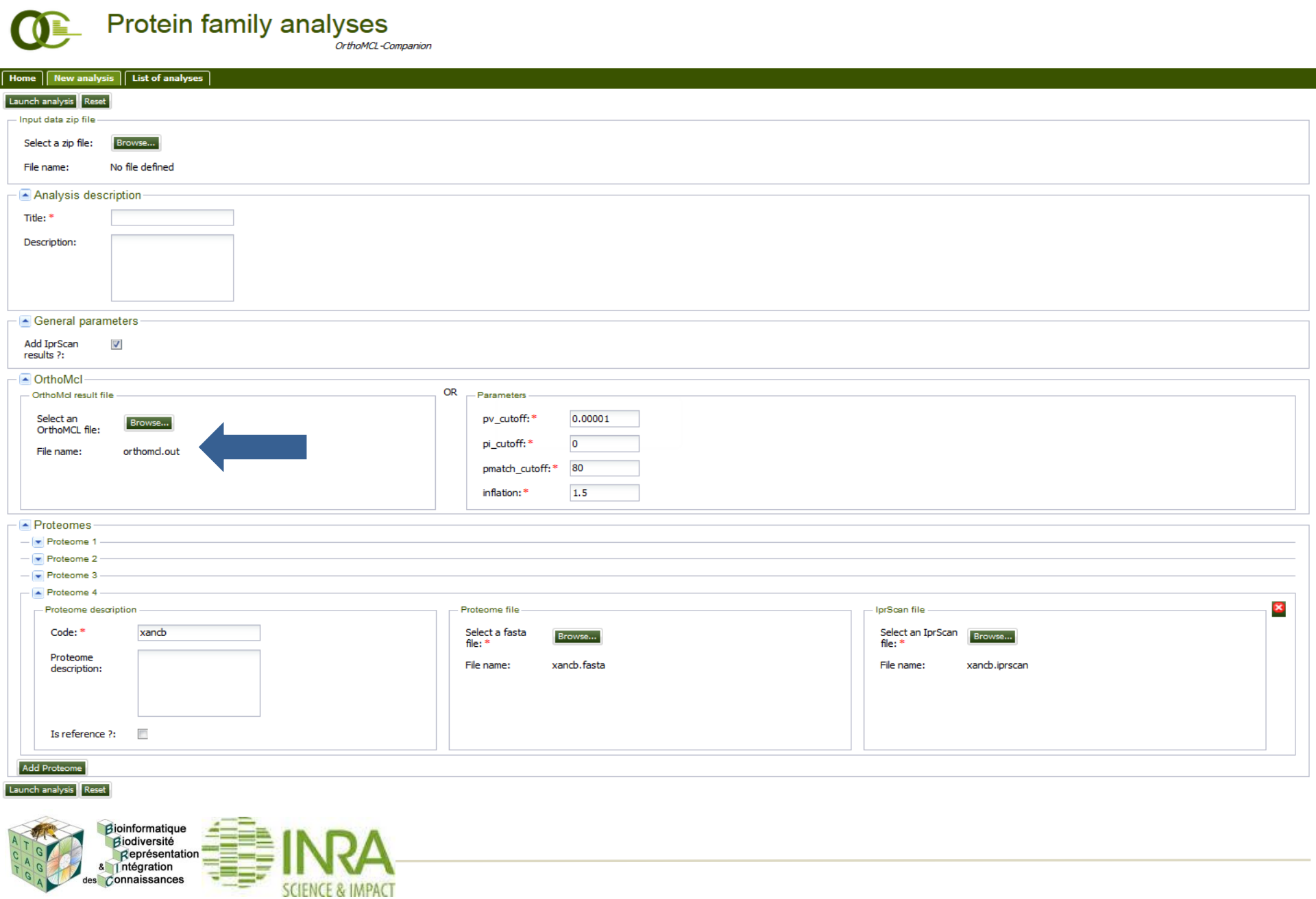
L'interface : lancer une analyse



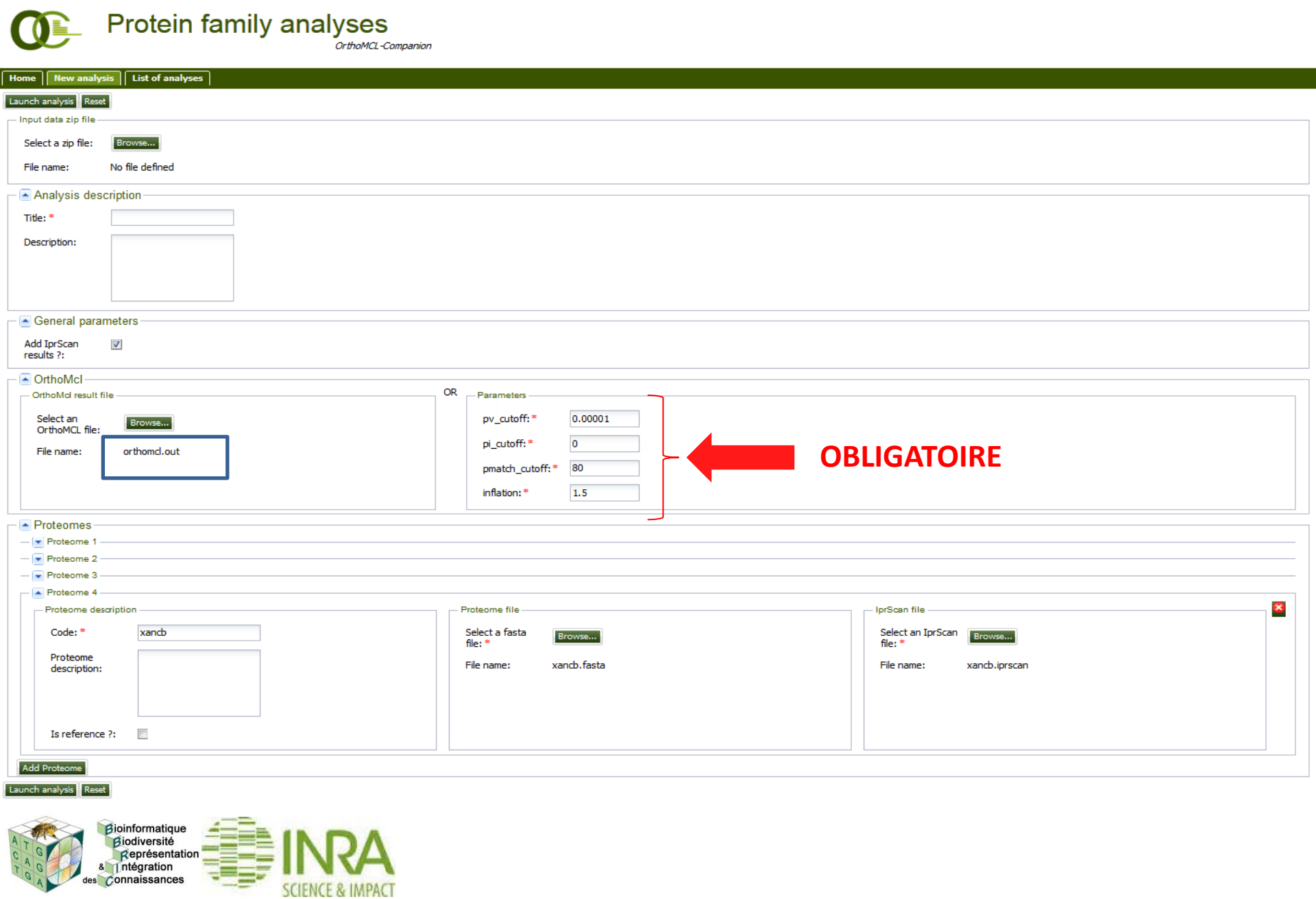
L'interface : lancer une analyse



L'interface : lancer une analyse



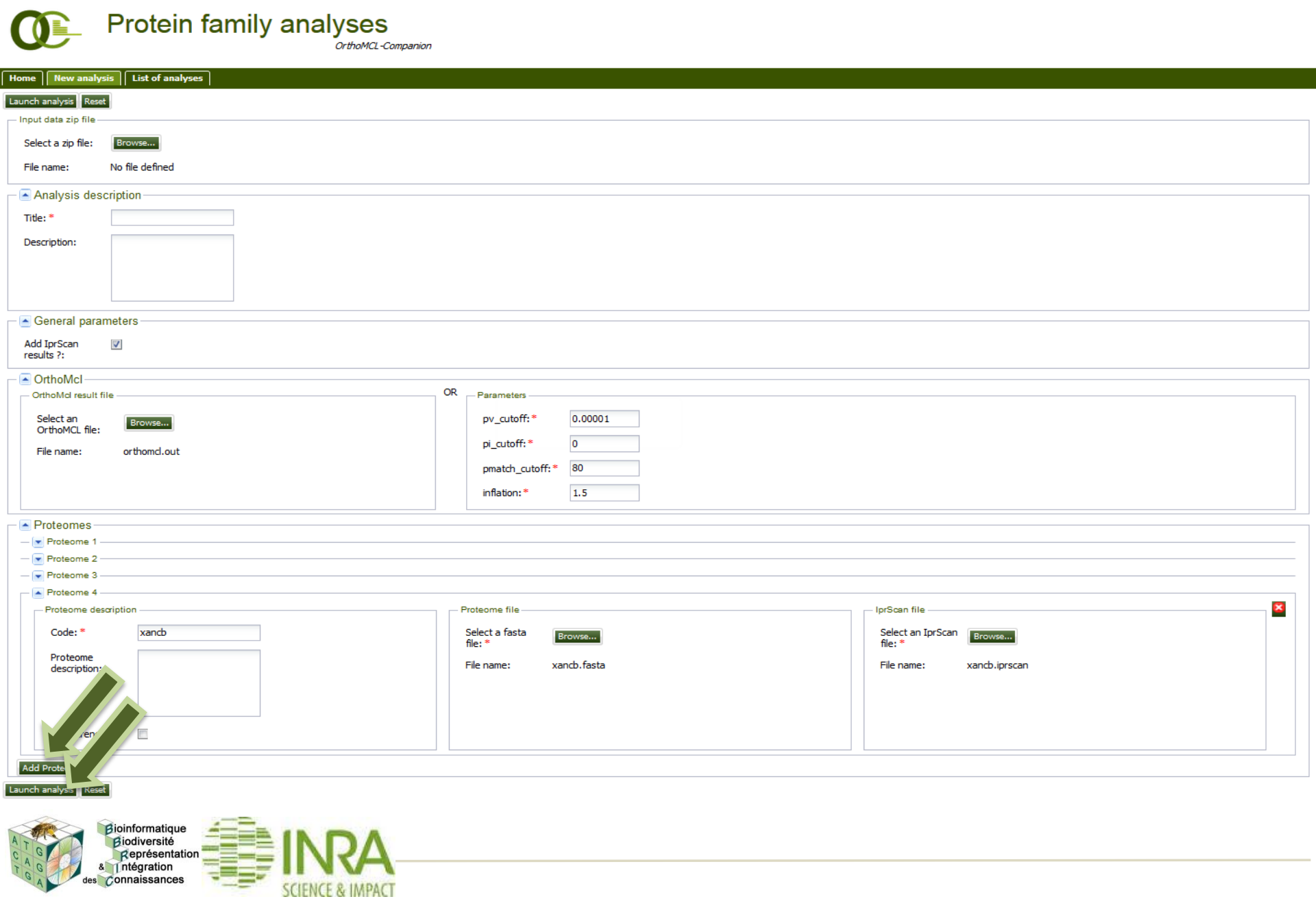
L'interface : lancer une analyse



Bioinformatique Biodiversité Représentation & Intégration des Connaissances

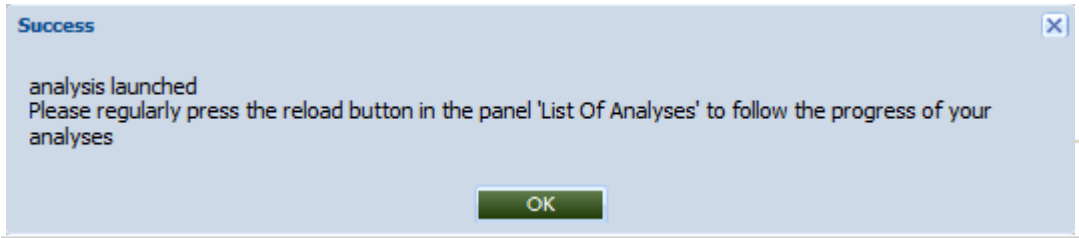
INRA SCIENCE & IMPACT

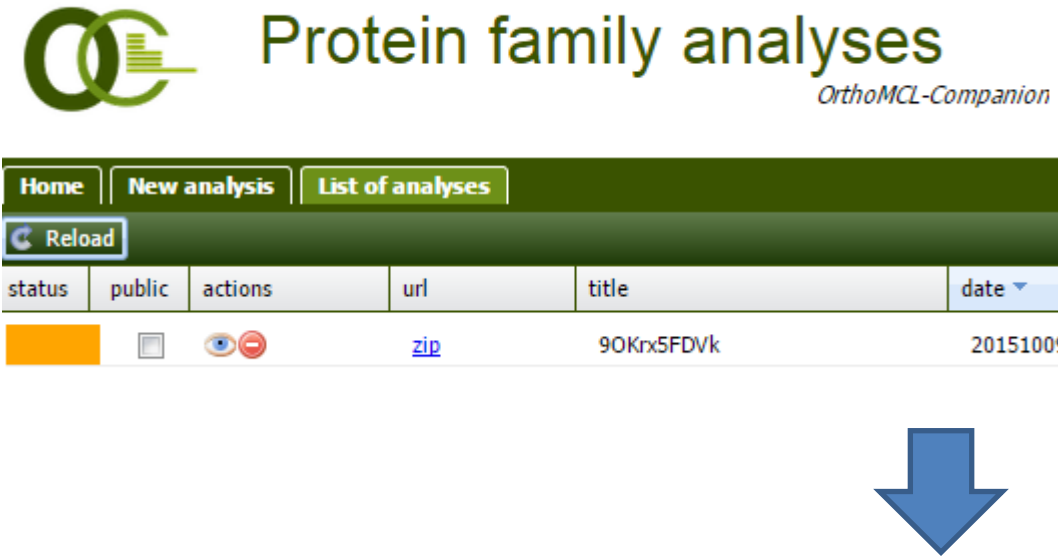
L'interface : lancer une analyse



Les résultats obtenus

orthomcl-companion





Bioinformatique Biodiversité Représentation & Intégration des Connaissances

INRA SCIENCE & IMPACT

Les résultats obtenus

Protein family analyses

OrthoMCL-Companion

Home | New analysis | List of analyses | 9OKrx5FDVk

Global parameters

Contact	Ludovic.Cottret@toulouse.inra.fr
Date	20151009 15:40:44

OrthoMcl parameters

Outfile	orthomcl_output/orthomcl.out
Version	1.4 (blast+)
Parameters	--pi_cutoff=0 --pv_cutoff=0.00001 --pmatch_cutoff=80 --inflation=1.5

Proteomes

code	reference	title
Xbbric	false	X. bbric str. mini
Ecoli	false	E. coli str. mini
Bacsu	false	B. subtilis str. mini

Results

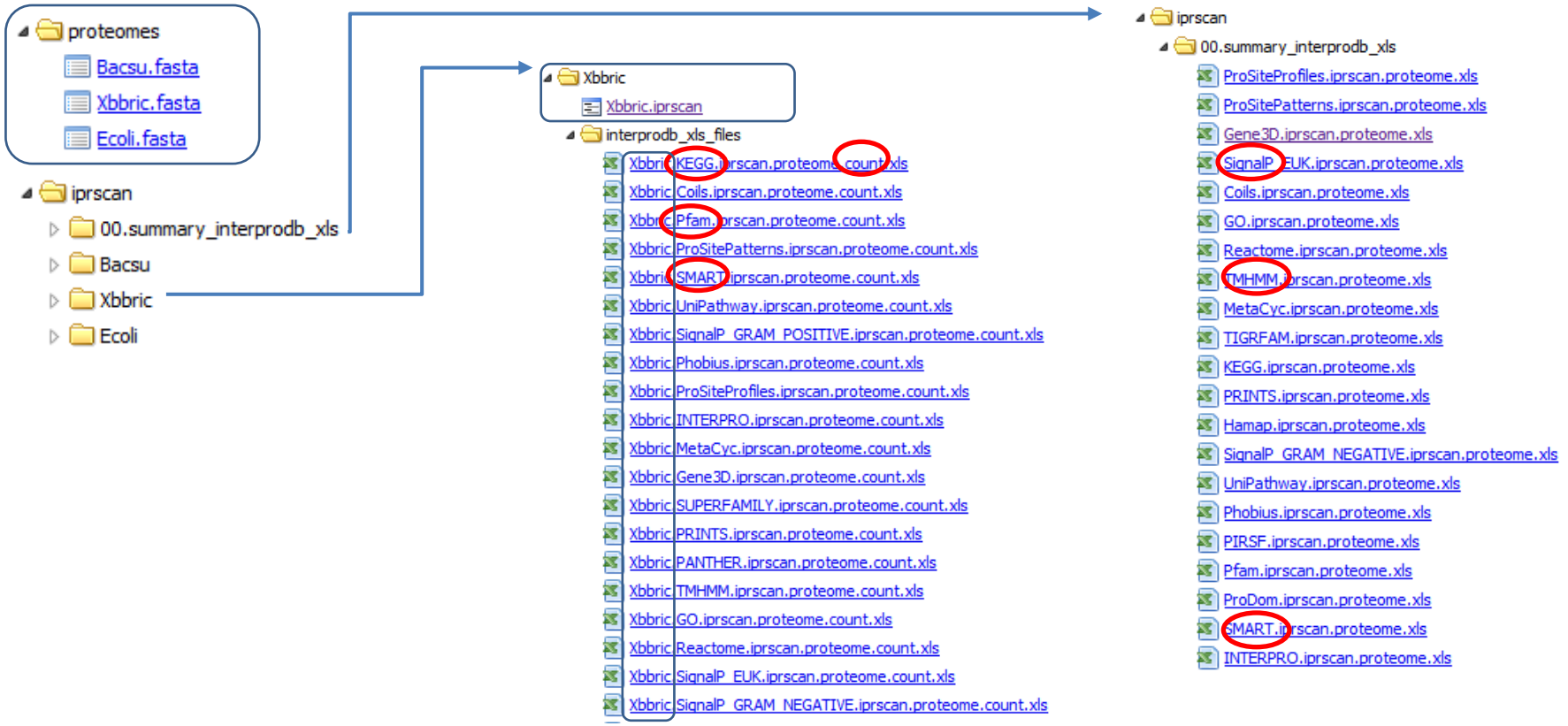
Status	Name	Description
	input_data	Input data
	orthomcl_output	OrthoMCL Group description and annotation
	summary	Analysis summary - Global counts
	specific_proteins	Extraction of Specific protein sequences and annotation
	orthologous_proteins	Extraction of "orthologous" protein sequences and annotation
	core_proteome	Core proteome pseudomolecule
	venn_diagrams	Lists of all groups with a protein of this proteome + unique protein by proteome
	pan_proteome	Directory with fasta multi-sequence of pan-proteome
	statistics_matrices	Descriptive Statistics: number of OrthOMCL group - Number of proteins by OrthOMCL group - Number of taxa by OrthOMCL gro...

Les résultats obtenus

Results

Status	Name	Description
	input_data	Input data
	summary	Analysis summary - Global counts
	specific_proteins	Extraction of Specific protein sequences and annotation
	orthologous_proteins	Extraction of "orthologous" protein sequences and annotation
	core_proteome	Core proteome pseudomolecule
	venn_diagrams	Lists of all groups with a protein of this proteome + unique protein by proteome
	pan_proteome	Directory with fasta multi-sequence of pan-proteome
	statistics_matrices	Descriptive Statistics: number of OrthOMCL group - Number of proteins by OrthOMCL group - Number of taxa by OrthOMCL group; Phylogenetics matrices : Presence/Absence martix - ...

input_data



Les résultats obtenus

Results

Status	Name	Description
	input_data	Input data
	summary	Analysis summary - Global counts
	specific_proteins	Extraction of Specific protein sequences and annotation
	orthologous_proteins	Extraction of "orthologous" protein sequences and annotation
	core_proteome	Core proteome pseudomolecule
	venn_diagrams	Lists of all groups with a protein of this proteome + unique protein by proteome
	pan_proteome	Directory with fasta multi-sequence of pan-proteome
	statistics_matrices	Descriptive Statistics: number of OrthOMCL group - Number of proteins by OrthOMCL group - Number of taxa by OrthOMCL group; Phylogenetics matrices : Presence/Absence martix - ...

input_data

- proteomes
  - Bacsu.fasta
  - Xbbric.fasta
  - Ecoli.fasta
- iprscan
  - 00.summary_interprodb_xls
  - Bacsu
  - Xbbric
  - Ecoli

BG13696	Bacsu	PF07733	Bacterial DNA polymerase III alpha subunit
BG13696	Bacsu	PF02811	PHP domain
BG13696	Bacsu	PF14579	Helix-hairpin-helix motif
b0532	Ecoli	PF13953	PapC C-terminal domain
b4210	Ecoli	PF00892	EamA-like transporter family
b4293	Ecoli	PF04542	Sigma-70 region 2
b4293	Ecoli	PF08281	Sigma-70, region 4
b0604	Ecoli	PF13098	Thioredoxin-like domain
Xbbric.3982	Xbbric	PF05598	Transposase domain (DUF772)
Xbbric.4079	Xbbric	PF00535	Glycosyl transferase family 2
Xbbric.3926	Xbbric	PF13659	Methyltransferase domain
Xbbric.4259	Xbbric	PF01551	Peptidase family M23

#TERM	DESCRIPTION	HITS	RATIO
PF00106	short chain dehydrogenase	6	0.0159574468085106
PF01609	Transposase DDE domain	5	0.0132978723404255
PF00005	ABC transporter	5	0.0132978723404255
PF00126	Bacterial regulatory helix-turn-helix protein, lysR family	4	0.0106382978723404
PF00072	Response regulator receiver domain	4	0.0106382978723404
PF04542	Sigma-70 region 2	4	0.0106382978723404
PF13276	HTH-like domain	4	0.0106382978723404
PF00535	Glycosyl transferase family 2	4	0.0106382978723404

Les résultats obtenus

Results

Status	Name	Description
	input_data	Input data
	summary	Analysis summary - Global counts
	specific_proteins	Extraction of Specific protein sequences and annotation
	orthologous_proteins	Extraction of "orthologous" protein sequences and annotation
	core_proteome	Core proteome pseudomolecule
	venn_diagrams	Lists of all groups with a protein of this proteome + unique protein by proteome
	pan_proteome	Directory with fasta multi-sequence of pan-proteome
	statistics_matrices	Descriptive Statistics: number of OrthOMCL group - Number of proteins by OrthOMCL group - Number of taxa by OrthOMCL group; Phylogenetics matrices : Presence/Absence martix - ...

summary

summary
- summary.xls
- summary.count.html
- summary.percentage.html

#CODE	TITLE	NB OF PROTEINS	PROTEINS IN GROUP	INPARALOGS	INPARALOGS SPECIFIC PROTEINS	SINGLE COPY SPECIFIC PROTEINS
Bacsu	B. subtilis str. mini	338	175	104	40	163
Ecoli	E. coli str. mini	406	210	103	15	196
Xbbric	X. bbric str. mini	364	159	21	11	205

summary

Percentage of every kind of datasets for all species

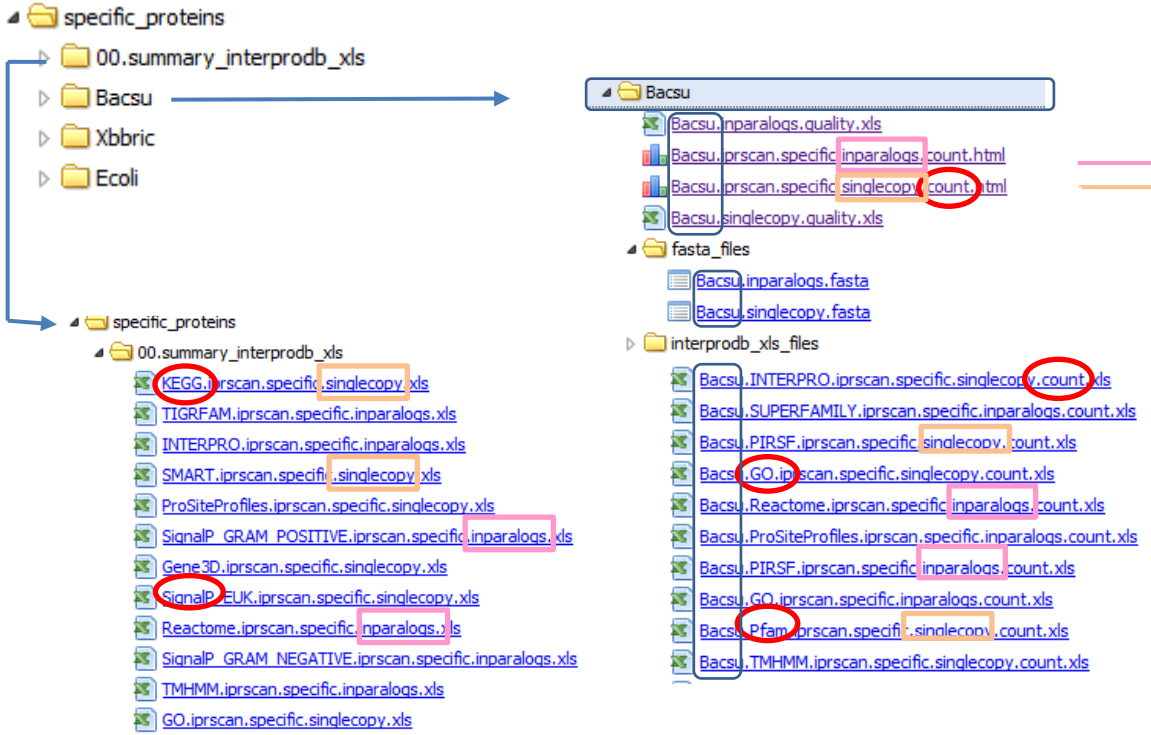
Count of every kind of datasets for all species

Les résultats obtenus

Results

Status	Name	Description
	input_data	Input data
	summary	Analysis summary - Global counts
	specific_proteins	Extraction of Specific protein sequences and annotation
	orthologous_proteins	Extraction of "orthologous" protein sequences and annotation
	core_proteome	Core proteome pseudomolecule
	venn_diagrams	Lists of all groups with a protein of this proteome + unique protein by proteome
	pan_proteome	Directory with fasta multi-sequence of pan-proteome
	statistics_matrices	Descriptive Statistics: number of OrthOMCL group - Number of proteins by OrthOMCL group - Number of taxa by OrthOMCL group; Phylogenetics matrices : Presence/Absence martix - ...

specific-proteins



#seqid	groupid	length	%of undetermined AA (X or *)
BG12077	ORTHOMCL157	308	0
BG11300	ORTHOMCL159	254	0
BG12932	ORTHOMCL70	513	0
BG12005	ORTHOMCL159	253	0
BG11346	ORTHOMCL168	641	0
BG11867	ORTHOMCL74	358	0
BG13442	ORTHOMCL158	363	0
BG11392	ORTHOMCL70	544	0

Bacsu.inparalogs.quality

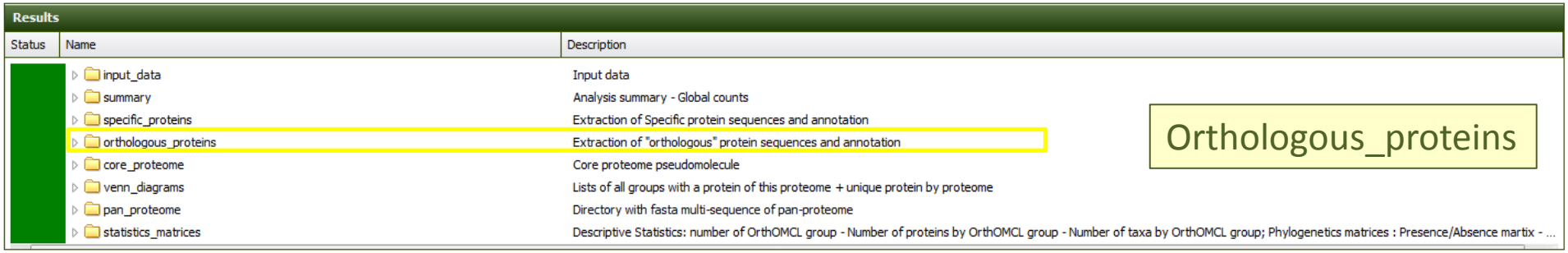
Protéines d'un groupe orthomcl mono-espèce

#seqid	groupid	length	%of undetermined AA (X or *)
BG12780	null	73	0
BG11314	null	184	0
BG10712	null	181	0
BG13114	null	154	0
BG13882	null	80	0
BG10239	null	106	0
BG12369	null	438	0
BG12620	null	157	0

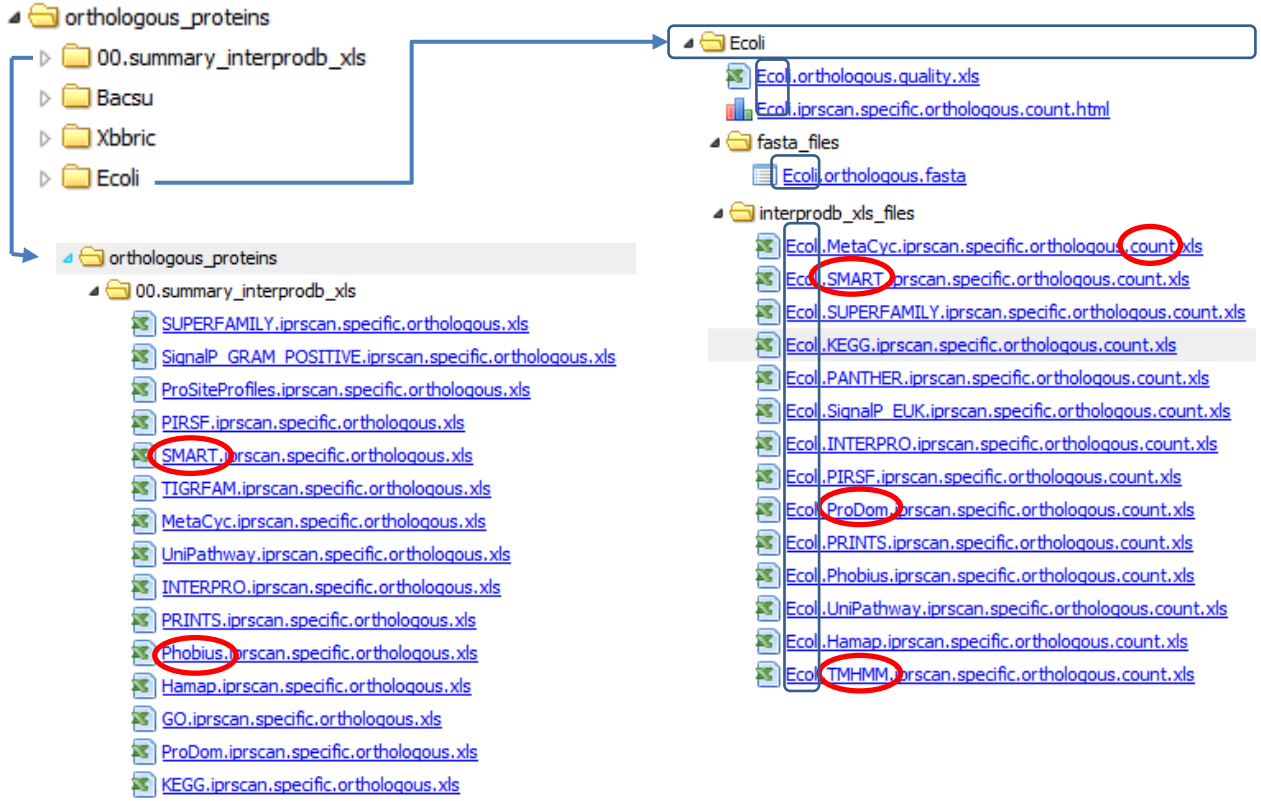
Bacsu.singlecopy.quality

Protéines spécifiques à une espèce et en une seule copie

Les résultats obtenus



Orthologous_proteins



Protéines provenant de groupes orthomcl possédant une et une seule séquence par espèce et où toutes les espèces sont présentes.

Les résultats obtenus

Results

Status	Name	Description
	input_data	Input data
	summary	Analysis summary - Global counts
	specific_proteins	Extraction of Specific protein sequences and annotation
	orthologous_proteins	Extraction of "orthologous" protein sequences and annotation
	core_proteome	Core proteome pseudomolecule
	venn_diagrams	Lists of all groups with a protein of this proteome + unique protein by proteome
	pan_proteome	Directory with fasta multi-sequence of pan-proteome
	statistics_matrices	Descriptive Statistics: number of OrthOMCL group - Number of proteins by OrthOMCL group - Number of taxa by OrthOMCL group; Phylogenetics matrices : Presence/Absence martix - ...

Core-genome

core_proteome	Core proteome pseudomolecule
aligned.core.proteome.fasta	Core proteome pseudomolecule

Protéines orthologues présentes en une seule copie dans toutes les espèces

↓

Alignements

↓

Concaténation des alignements

Les résultats obtenus

Results

Status	Name	Description
	input_data	Input data
	summary	Analysis summary - Global counts
	specific_proteins	Extraction of Specific protein sequences and annotation
	orthologous_proteins	Extraction of "orthologous" protein sequences and annotation
	core_proteome	Core proteome pseudomolecule
	venn_diagrams	Lists of all groups with a protein of this proteome + unique protein by proteome
	pan_proteome	Directory with fasta multi-sequence of pan-proteome
	statistics_matrices	Descriptive Statistics: number of OrthOMCL group - Number of proteins by OrthOMCL group - Number of taxa by OrthOMCL group; Phylogenetics matrices : Presence/Absence martix - ...

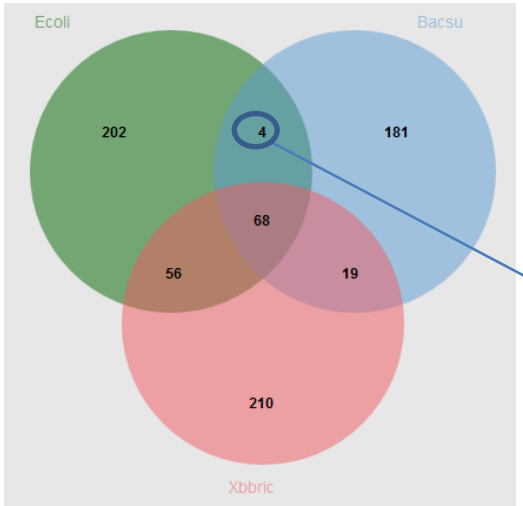
venn_diagrams

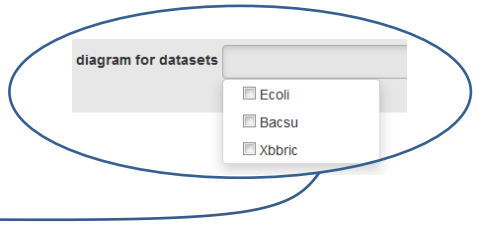
- venn_diagrams
  - venn.html
  - lists
    - Xbbric.list.txt
    - Ecoli.list.txt
    - Bacsu.list.txt

Orthomcl Groups content analysis

Generate Venn diagrams that show shared and specific families proteins of selected proteomes.
On these Venn diagrams :
- intersection between proteomes contain number of orthomcl groups shared by these proteomes.
- number of specific elements of a proteome is the sum of :
  * specific inparalogs groups of this proteome
  * specific singlecopy protein of this proteome.

Sub directory lists contain, for each proteome, the list of each group that contain a protein of this proteome and specific singlecopy protein of this proteome.





Click on a venn diagram figure to display the linked elements:

Common elements in Ecoli Bacsu :
ORTHOMCL161
ORTHOMCL153
ORTHOMCL148
ORTHOMCL160

Les résultats obtenus

Results

Status	Name	Description
	input_data	Input data
	summary	Analysis summary - Global counts
	specific_proteins	Extraction of Specific protein sequences and annotation
	orthologous_proteins	Extraction of "orthologous" protein sequences and annotation
	core_proteome	Core proteome pseudomolecule
	venn_diagrams	Lists of all groups with a protein of this proteome + unique protein by proteome
	pan_proteome	Directory with fasta multi-sequence of pan-proteome
	statistics_matrices	Descriptive Statistics: number of OrthOMCL group - Number of proteins by OrthOMCL group - Number of taxa by OrthOMCL group; Phylogenetics matrices : Presence/Absence martix - ...

Pan-genome

pan_proteome

PanProteomeDescription.xls

PanProteome.fasta

#Species	Nb seq in panproteome	Nb seq rep group in panproteome	Nb seq uniq in panproteome
Xbbric	286	81	205
Ecoli	249	53	196
Bacsu	205	42	163

PanProteomeDescription

```
>G0-10541|Ecoli|ORTHOMCL0
VIVLILVFRLVIGEQMIDVLGPEKRRRRTTQEKIAIVQQSFEPGMTVSLVARQHGVAASQLFLWRKQYQEGSLTAVAAGEQVVPASELAAAMKQIKELQRLLGKKTMENELLKEAVEYGRAKKVDSARALIARGWGVSLVSRCLRVSRAQLHVILRRTDDWMDGRRSRHTDDTDVLLRIHHVIGELPTYGYRRVWALLRR
QAELDGMPAINAKRVYRIMRQNALLLERKPAVPPSKRAHTGRVAVKESNQRWCSDGFEFCCDNGERLRVTFALDCCDREALHWAVTIGGFNSETVQDVMLGAVERRFGNDLPSSPVEWLTDNGSCYRANETRQFARMLGLEPKNTAVRSPESNGIAESFVKTIKRDYISIMPKPDGLTAAKNLAEAFEHYNEWHPHSALG
YRSPREYLRQRACNGLSDNRCLEI
>BG12652|Bacsu|ORTHOMCL1
MKRQLKSPLSEGQKGLWWLQKMSPGMSAYNIPLCFRFSKPIHAETFKKALLFVQRQYPVLASVIQEENGIPFQSVQLSKDLYFVEEDISAMKSADIMPFLKEKAKEPFQLEAGPLWRTHLFHRLEECIVLITIHHIIFDGVSMLTLISALFEAYQQLLNGIEPLQQPSTADYYDFVDWENRMLTGREGEEHLAYWKEQLS
GSLPVLDLPADRPRSSARKFKGQAYKSLLPHHLRNQIKSFARTNHVNESVVFLSIYKVLLHHYTKQKDIIVGVPTMGRQEDRFETLIGYFINMMAVRSKNIGSQPLTAFIRELQLTVAVGLDHAAFPFPALVRELNVDRSAADSPVFQTAFLYQNFFQATGLQKVLEPYQTLIGIEYIEDIRQEGEFELALEIYEQENETV
LHLLYNPDLYELSSIESMMENYMKLAQHMMEDPSLPLEAYSLQLNQEQTSLLEQWNATGTNIANDKCIHEVFEEEKAKQTPDAVAVMFEDRSLTYKEVDEKSTSVAVYLQHQGVRPEQPVGICAERSFDMIIGILGILKAGGAYVPLDPSFPQERLKYMLKDSQASIVLTQPNVHDRISGLTGSHVKAINIELACRNGYTD
QQSSGLKREVKPEHLAYIIYTSGSTGEPKGVMVEHRSIMNTLNFLESHYPVTAEDAYLLKTNYVFDVSISELFGWFIGDGRLVILPPNGEKSPQLCMDYIETYKVTHINFVPAMLHVFLEMAKDNKRFTEDGPLKYMMVAGEAFPKVLVKKAVSLFTNCRVENIYGPTEASIYAAYFGCGKGDIASHHTPIGKPVSNTKI
YIVDQHLKPVPIGKPGELCIAGAGLARGYFKKPGLTAEKFIDNPFESGTKLYKSGDSARWLPDGNIEYLGRIDSQVKIRGFRVELGAIETKLGEFPGILDQAVVVKQLEGHQQLAAYYTEESGHASANPKDLRLHLKSSLPEYMIPSHFIRLDELPLSPSGKVNRKELEKREIVFNRRKPNHLQLTEIEDQVLRIWEETL
KVSGFGPEDGFFDAGGDSLLAVAVAERIKKEFDCEFHVTELFEYSTIRAISEYILEMKNSDLAGTQNEDDHDDKKDGKYPKQKIPPYFDDSVAIVGISCQFPGAKNHDFWNNIKEGKESIRFFSEEDVRANGVPEELIQHPDIVFVQSVIEGKDLFDPGFFQISPKDAEYMDPQLRLLLLHSWKAIEDAGYVAKEIPAT
SVYMSASSNSYRTLLPKETTEGHESRDDGVYSWVLAQSSTIFTMISHKLGLKGFSYFVHSNCSSSLVGLVQAYKSLTSGESQYALVSGATLHAQSAIGVFRQNGLNFSSDGHTVKATDASADSMAGSEGVAVLLLKKAVDAVKDGDHIYAIMGSIGINNDGAEKAGFYAPSVKSQTEVIQHVLDTTKIHPETVSYIEAHGTG
TKLGDPIEMSALNKVYKQVTDKTQFCGIGSVKTNIGHLDTAAGLAGCIKVAMSLYHNELAPTINCTEPNPDIKFESSPFVVPRERKSLEKHAGVHRAALSSFGLGGTNAHAIFEQVENISDAGAENEGNQPVIIPISAKNSERLQVYAKEMLSYISQDEQRHFSLRDIAYTFQVGREAMDNRIVFIVNDLEEWKHQLEAF
VTGKPLAEGCIQGEKTRMTSAEQLLGNAEADDMASSRISKEELRKLAEMWANGFHVEWRRLYPNIKPRRISLPTYPFAEERYWPESSTGAITTIEPSRLHPLEHHNTSVLSEQRFSSIFTGQEYFIAEHIIKGMAILPAAVTLEMARAAIEQGIGGLEDHETGIRLKWVVWVRPVVAGSEPVQVNIGLYDEDGGHIAYRM
YGDPESADAEPVVYNQGKAELIQLKREKALDLSKIKKQCDQSKMDAASFYEGMIGADYGPGYKSVEAVYKGDGQLLAKLSLPESVAHTLGDYVLHPSVMDGALQAAEYLQNVVRAELSDTEDFKAALPFALEELEVFRQCVSDMWVYVQFNSKNKPGDLIQKVDIHLCDEHGMICVRLKGFSTRVMEADIQTEPSKINAE
TLLLQFVWQEQKAANSLAAKKYAEHLVFLCEYDHETRKQIEAAIEDVHVYSLEARPSSVDGRFHSYTEQVFKKVQEIIRTKPKDGILVQIVTSAEGEQQLFSGLTGLLKTACQENAKLTGQMIEVSSEESGESIAGKLLENQMSSDSYVKYQNGTRYIADWREIKQAKGDGSKPWKDNGVYLISGGAGGLGHIFAKEIAE
QTKNATVILAGRSPLSESKSKKLKELHSKGADITYRQTDVTNKIEVYQLIDDIQKRYGRLNGILHSAGIIKDSYLVNKQAKDLHDVLAPKVKGLVYLDEASKDLPLDFFILFSSLSGSLGSIGQSDYAAANVFMDMYAGYRNRLADLSQRHGQTLSVNWPLWRDGGMQVDQETEKRLVQLAGIVPMRAEKGIQALVQALH
SEANQVMVIEGDVQKIKQNMLAKNASAPMEKKEAEHMTEQINSIDADSLLDKVKAMLKREIAKLLKVKLETIDDHAEMTVYGFDSISMTEFTNHINRAYQLELTPTVFFDHPTIHAFGKHLSEEYQSVFAKTFAVRAVSAQLQPAAKQEQAVRAKAKRRRKQQVMLPNAIQSDAGPEPIAIVGISGIFPMAKDVEAYWNI
LKEGKDCMTEIPKDRWDWREYEGDPAKEVNKTNVKWGGFIDGIADFDPLFFGISPREAEQMEPQQRLLLTYAWKAIEDAGYSAKRLSGTKTGVFIGTGNTGYSSLLSKANSAIEGSAAANTSPSVGPNRVSYFLNLHGPSEPVDTACSSSLVAIHHAISSIEEGTCDMALAGGVNTIILPEVYISFDKAGALSKEGKCKT
FSNQADGFAHGEGAGILFLKKLKAAEEAGDHIYGVIKGSAINHGGRAASLTTPNPKAQADVIQSAYQKAGIDPKTVTYIEAHGTGTELGDPVEINGLKSAFKALGVNEGDTSANPYCGLGSVKTNIGHLSLAAGAAGVIKILLQLKHKTLVKSLHCENVNPYIQLKNSPFYIVRETEEWKALKNEQGEELPRRAGVSSFG
IGGVNAHVIIEEYIPEASDENIPSIAPEHPGIFVLSAKNEARLKEHAQQLADALDKQTYSDVNLARIAYTLQAGRDAMEERLGIISGSIEDLQKKLKDFAAEKSGVEDVFKGRIDKGTLQMLTEDEEIQEAVEKWMERGKYAKLLELWVKGLDVDWTKLYGENLPKRISLPTYPFAKDRYWISDHIEKSGSIDANQAASR
LGGAVLHPLMHQNTSNLSEQRFSSIYTGEEFFLADHVVKGQRILPGVAHLELARAAVEQAAEVQGVPRIMKLKNAVWVRPIVVEDQPQQVHIRLLPGENGEISYEIYGHSDVTGEQSIVYSQGSAVLNPAENLPAVDLQSLREQCQESHFSVNEVYDTYRMIGFEYGPAYRGVKKIYTAEQFVLAKLSLHPSAADTLSQY
KMHPGLMDSALQASSILTGAGDNQLTLPFAVQELEVFGACSSEMWVYARYSQGSKATDKVQKRDMDILDESGNVCVRMKGLSFRAAEGGSGSAESDQTLATLMFEEKWVPKDFKKESPEPHYERHIVMLCDMNGLSKDRIESRMIGAECIVLESFREGLAERFQDYAEQALETVQGLLKSRPQGNVLIQLLTSAQRKQYS
FSGLSALLKTAGLENKKLIGQTIEIDSHENVESVIEKLKENKRHTEDQHIKYEKGKRYINDLDEMQIDDREISMPWRDKGVYLITGGAGGLGFIFAKEIARQAEQPVLILTGRSALNADQQAELNELQQLGARAEYRQVDVTQTEAASELITSITSDYEDLNGVIHSAGLIKDNYLMSKTNEELTQVLAPKVKGLVNVDE
ATEHLALDFFILFSSISSVAGSAGQADYAMANAFMDSYAAYRNALVTAMYRHGQTLSINWPLWKEGGMRANKEIENMTLKNTGVTPMRTETGIQALYKGLAFGKDQVIVMEGFKDMMREKLTQKPSSDDVPMKTVQVRVTSEARMDQGNMFDHIQEVLKQTISQLLKIKPEEIDPDMEFNQYGFDSITILTEFANTLNEKC
KLDLTPTVFFEHATVYAFAGYLSEEYPNAFTAQTPAKAEVLMQFVEQNIKNMTFSTENRFVKPSVTPMQKEADHKPEPIAIVGMSGVFPKAKDVEEYWKNLSSGADCITEVPKDRWDWQEYYGDPLKEANKTNVKWGGFIDEVADFDPLFFGISPLEAEQMEPQQRLLMTYAWKAVEEAGHSARSLAGTKTGIFIGTGNT
GYSSLLSNVDIEGSAAANMSPSAGPNRVSYFLNIHGPSEPIDTACSSSLVAIHHAVCAIENGNCEMAIAGGVNTVVTPQGHIAYDKAGALSKEGRCKTFSDKADGFAVSEGAGILFLKKLTAAERDGDHIYGVIKGSAVNHGGRANSLITPNFKAQADVVKTAYEKAGIDPRTVTYIEAHGTGTELGDPVEINGLKAAFK
ELYEKTGDPAVHGSHCGLGSAKTNIGHLSLAAGVAGVIKVLLQLKHKTLVKSLYSETVNEYIRLDDSPFYIVQESREWQALRDEAGREPRRAGISSFGIGGVNAHVVIEEYIPKETTHPATAPAVTIAQHPGIFILSAKDEDRLKDQARQLADFISKRSITARDLTDIAYTLQEGRDAMEERLGIIAVSTGDLLEKLNLF
IEGGTNAKYMYRGRAEKGIAQTLRSDDEVQKTLNNSWEPHIYERLLDLWVKGMEIGWSKLYGDKQPKRISLPTYPFAKERVWITDTKEEAAAHQTALKTVESAALHPLIHVNTSDLSEQRFSSAFTGAEFFFADHKVKGKPVMPGVAYLEMVHAAVTRAVRRTEDQGSVHIKNVVWVQPIVADGQPVQVDISLNPQQDG
EIAFNVVTTEAAHNDRKIHCQGSASIRGAGDIPVQDISALQDQCSLSTLSHDQCYELFKAIGIDYGPGFQGIDRLYIGRNQALAELSLPAGVTHTLNEFVLHPSMADSALQASIGLKLNSGDEQLSLPFALQELEIFSPCTNKMWVSVTSRPNEDKTQRLDIDLCDEQGRVCVRIKGITSRLLEEGIQPFDGPTSLGNSKA
FIAGALMAPIWNRVQLEKRSISPADEVVILGGDDNSRKAVQREFFFAKELYIEFNASIHRITGQLEALGSFDHIVWNSPSRVTECEVSDEMIEAQDQQVIQMYKLIKGMLSLGYSQKELSWSITVTNYQVVDQHDIVDPFDAGVMGLIGSMSKEYPNWQTKLIDVKKYEDLPLSQLLSLPADQEGNTWAYRNKIWHKL
RLIPVHNNQPVHTKYKHGGVYVVIGGAGGIGEAWSEYMIRTVQAQIVWIGRRKKDAAIQSKLDRFARLGRAPYYIQADAANREELERAYETMKQTHREINGIIHSAIVLQDRSLMNMSEECFRNVLAAKVDVSVRMAQVFRHEPLDFVLFSSSVQSFARASGQSNYAAGCSFKDAFAQRLSQVWPCTVAVMNWSYWGSIG
VVSSSPDYQRRMAQAGIGSIEAPEAMEALELLLGGPLKQLVMMKMANETNDEAEQTEETIEVYPETHGSAIQKLRSYHPGDNTKIQQLL
>acrD|Xbbric|ORTHOMCL10
MGFSTIFIRRPIATSLLMAGILLLGILGYRQLPVSALPEIDAPSLVVSTQYPGANATTMASLVTIPLERQLGQISGLQMMTSDSSAGLSTIILQFSMDRDIDIAAQDVQAAIRQATLPSSLPYQPVYNRVNPADAAILTLKLTSDTLPLREVNRYADAILAQRLSQVPGVGLVSIAGNVRPAVRIQVNPAQLSNMGLTME
SLRSALTQTINVSAPKGSLNGKTQSYSIGTNDQLTDAAQYRETIISYNNGAPVRLADVAKVVDGVENDQLAAWADGKPAVLLEIRRQPGANIVQTVEQIRSILPQLQGVLPADVHLDVFSDRTETIRASVHEVKFTLVLTIFLVVAVIFVFLRRLWATIIPSVAVPLSLAGTFGVMAFAGMSLDNLSLMALVVATGFVVDD
AIVMIENIVRYIEQGKSGPEAAEIGARQIGFTVLSLTVSLVAVFLPLLLMPGVIGRLFHEFAWVLSIAVVISMLVSLTLTPMMCAYLLKPDALPEGEDAHERAAAEGKTNLWTRTVGVYERSLDWVLDHQPLTLAVAIGAVALTVVLYVAIPKGLLPEQDTGLITGVVQVDQNVAFPQMEQRIQAVAAALRKDPAVTGVA
AFIGAGSMNPTLNQGQLSIVLKTRGDRDDLETVVARLQKAVSGIPGVALFLKPVQDVTLDTRVAATEYQYSMADVDSTELASWATRMTEAMRKLPELADVDNNLANQGRALELSIDRDKASMLGVPMQTIDDTLYDAFGQRQISTIFTELNQYRVVLEVAPEFRSSTALMNQLAVASNGSGALTGTNATSFGQVTSSNSS
```

Pan = 1 représentant de chaque groupe orthomcl

\+ toutes les protéines spécifiques en copie unique de toutes les espèces

✓ **Son intérêt est de produire une banque blast de référence**

Les résultats obtenus

Results

Status	Name	Description
	input_data	Input data
	summary	Analysis summary - Global counts
	specific_proteins	Extraction of Specific protein sequences and annotation
	orthologous_proteins	Extraction of "orthologous" protein sequences and annotation
	core_proteome	Core proteome pseudomolecule
	venn_diagrams	Lists of all groups with a protein of this proteome + unique protein by proteome
	pan_proteome	Directory with fasta multi-sequence of pan-proteome
	statistics_matrices	Descriptive Statistics: number of OrthOMCL group - Number of proteins by OrthOMCL group - Number of taxa by OrthOMCL group; Phylogenetics matrices : Presence/Absence martix - ...

statistics-matrices

- statistics_matrices
  - Number_Of_Proteins_by_OrthoMCL_group.classes.html
  - Number_Of_Taxa_by_OrthoMCL_group.classes.html
  - MatrixAbundance.txt
  - MatrixPresenceAbscence.txt
  - xls_files
    - Number_Of_Taxa_by_OrthoMCL_group.classes.count.xls
    - Number_Of_Proteins_by_OrthoMCL_group.classes.count.xls
    - Number_Of_Taxa_by_OrthoMCL_group.xls
    - Number_Of_Proteins_by_OrthoMCL_group.xls

GROUP	Bacsu	Ecoli	Xbbric
ORTHOMCL0	0	1	1
ORTHOMCL1	1	1	1
ORTHOMCL10	1	1	1
ORTHOMCL100	0	1	1
ORTHOMCL101	0	1	0
ORTHOMCL102	0	1	1
ORTHOMCL103	0	1	0
ORTHOMCL104	0	1	1
ORTHOMCL105	0	1	1
ORTHOMCL106	0	1	1
ORTHOMCL107	0	1	1
ORTHOMCL108	0	1	0
ORTHOMCL109	0	1	1
ORTHOMCL11	1	1	1
ORTHOMCL110	0	1	1
ORTHOMCL111	0	1	0
ORTHOMCL112	0	1	1
ORTHOMCL113	0	1	1
ORTHOMCL114	0	1	1
ORTHOMCL115	0	1	1
ORTHOMCL116	0	1	1
ORTHOMCL117	0	1	1
ORTHOMCL118	0	1	1
ORTHOMCL119	0	1	1
ORTHOMCL12	1	1	1

GROUP	Bacsu	Ecoli	Xbbric
ORTHOMCL0	0	19	2
ORTHOMCL1	11	2	1
ORTHOMCL10	1	2	2
ORTHOMCL100	0	1	1
ORTHOMCL101	0	2	0
ORTHOMCL102	0	1	1
ORTHOMCL103	0	2	0
ORTHOMCL104	0	1	1
ORTHOMCL105	0	1	1
ORTHOMCL106	0	1	1
ORTHOMCL107	0	1	1
ORTHOMCL108	0	2	0
ORTHOMCL109	0	1	1
ORTHOMCL11	3	1	1
ORTHOMCL110	0	1	1
ORTHOMCL111	0	2	0
ORTHOMCL112	0	1	1
ORTHOMCL113	0	1	1
ORTHOMCL114	0	1	1
ORTHOMCL115	0	1	1
ORTHOMCL116	0	1	1
ORTHOMCL117	0	1	1
ORTHOMCL118	0	1	1
ORTHOMCL119	0	1	1
ORTHOMCL12	3	1	1

Je veux comparer des protéomes complets

Je vais sur	L'outil permet	L'outil ne permet pas

Paramètres clés	Pièges	Formats et fichiers