Alignement de séquences, manipulation, contrôle-qualité et analyse de fichiers SAM/BAM

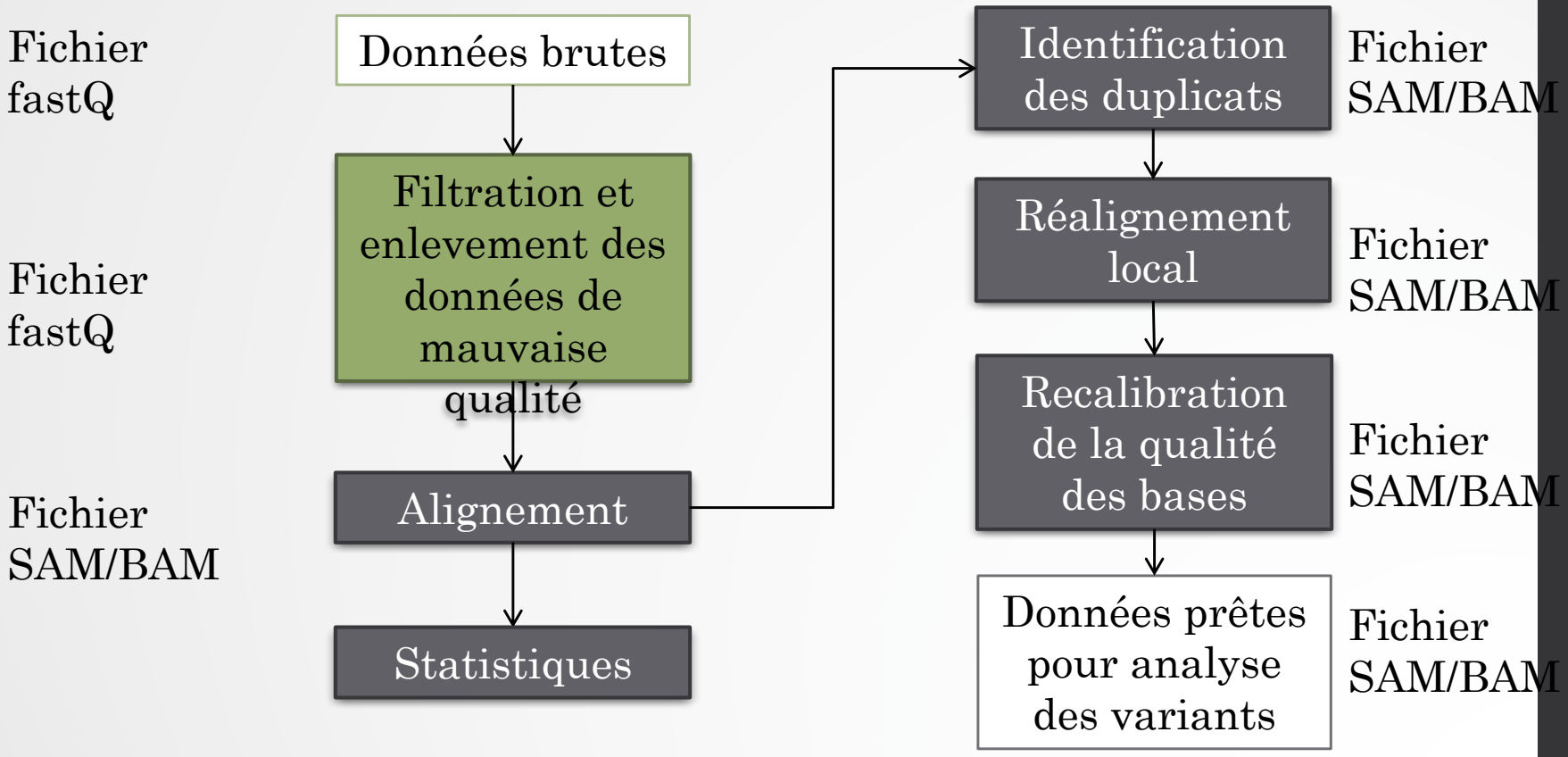
Stéphanie Le Gras

DU Dijon

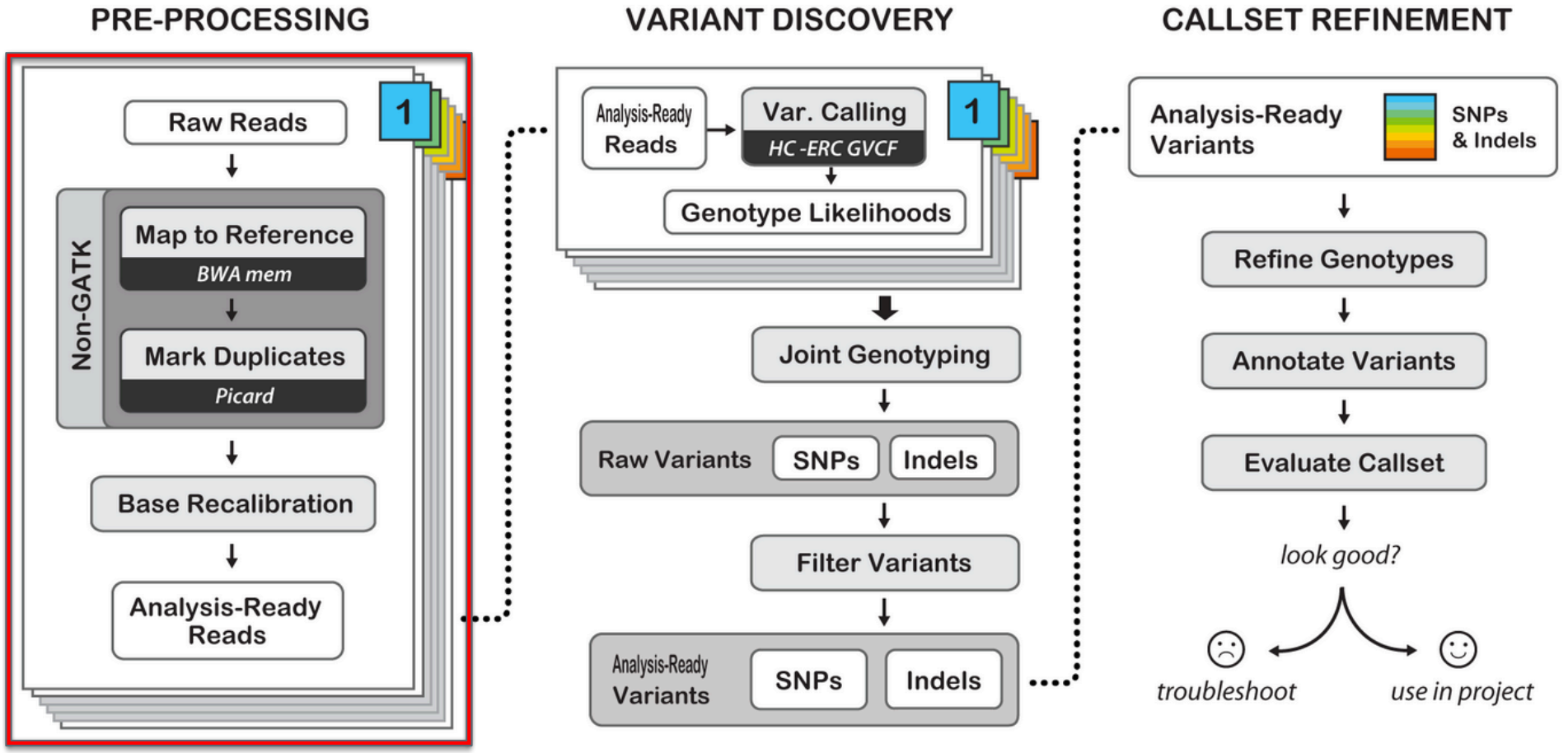
Objectifs

- Préparer les données avant de faire l'analyse de variants
  - Comprendre à quoi sert un alignement
  - Réaliser un alignement
  - Comprendre les biais qu'il peut y avoir dans un alignement de lectures
  - Corriger les biais
  - Connaitre le format SAM/BAM
  - Estimer l'efficacité de capture
  - Calcule la couverture nucléotidique

Analyse pré-détection des variants



Analyse de données pour l'application de re-séquençage



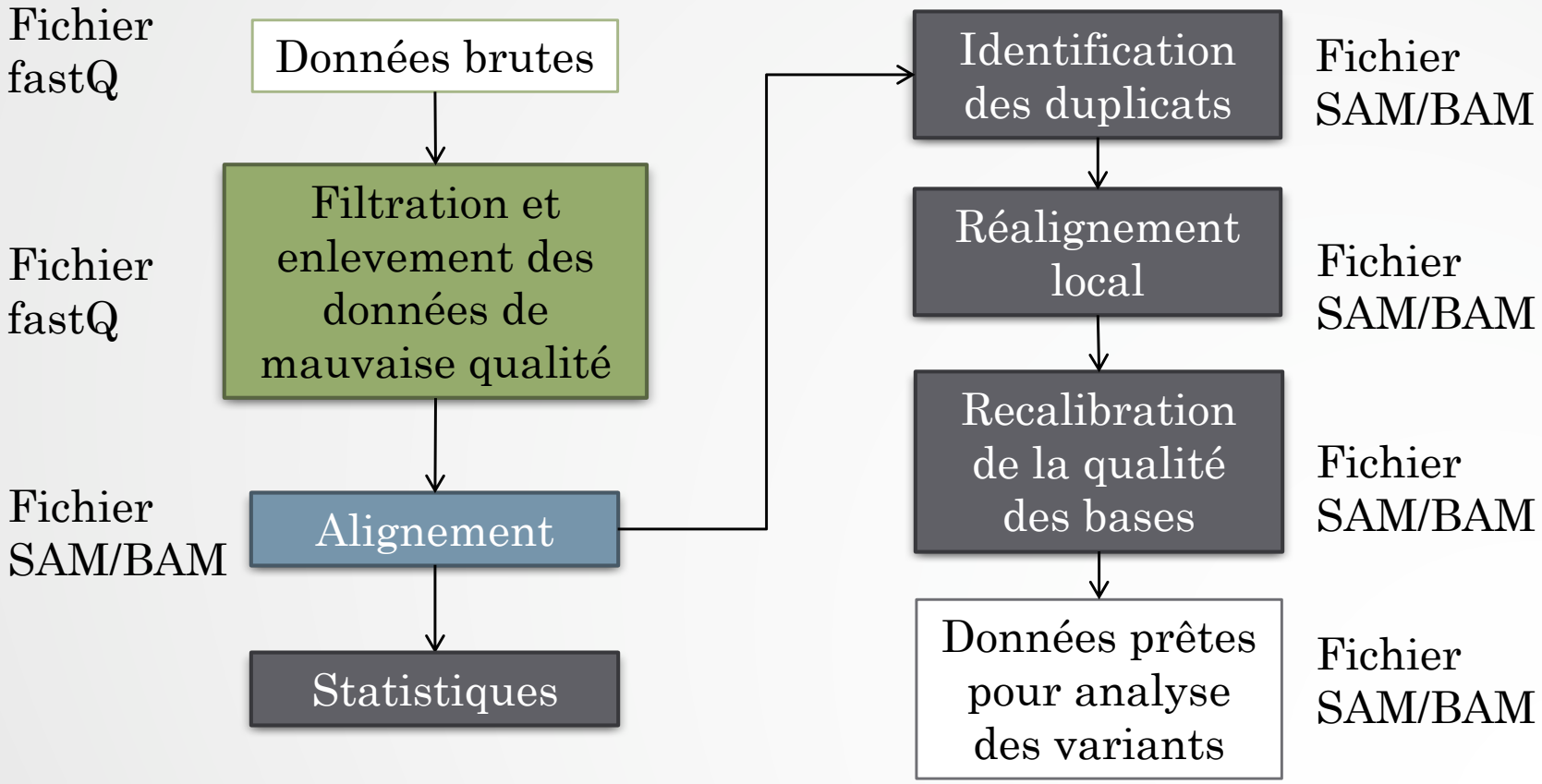
Best Practices for Germline SNPs and Indels in Whole Genomes and Exomes - June 2016

Plan

- Alignement
  - Le format SAM/BAM
  - Les régions posant problèmes
  - BWA
- La couverture nucléotidique
- Estimation de l'efficacité de capture
- Raffinement des alignements
- Recalibration des bases

ALIGNEMENT

Processus

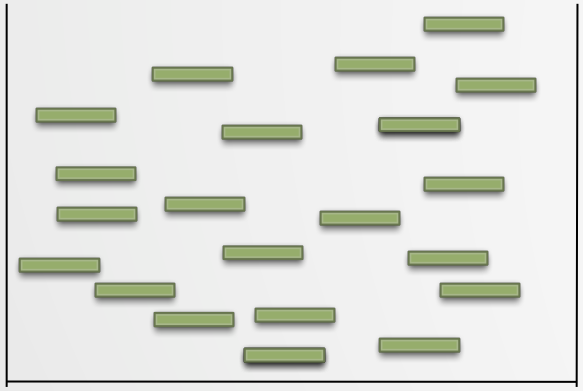

Fichier fastQ
Données brutes
Filtration et enlevement des données de mauvaise qualité
Fichier fastQ
Fichier SAM/BAM
Alignement
Statistiques
Identification des duplicats
Fichier SAM/BAM
Réalignement local
Fichier SAM/BAM
Recalibration de la qualité des bases
Fichier SAM/BAM
Données prêtes pour analyse des variants
Fichier SAM/BAM

Alignement

- Trouver la position des lectures dans le génome de référence

Génome de référence

Lecture



- Une seule position dans le génome de référence
- Plusieurs positions possibles (Répétition, régions dupliqués, pseudogènes…)

Alignement de données NGS

- Défi NGS :
  - Aligner rapidement des millions de lectures courtes en utilisant le minimum de ressources informatiques
  - Gestion des données pairées
  - ~~BLAST~~, ~~Blat~~
- Outils NGS
  - BWA (Li et al, 2009)
  - Bowtie
  - SOAP
  - … (Rufallo et al, Bioinformatics, 2011)
- Format SAM/BAM

SAM/BAM format

- SAM : Sequence Alignment/Map
- Format d'alignement générique
  - Avant SAM/BAM : 1 format de fichier par aligneur!
- Convient aux reads courts et longs (Illumina, AB/Solid et Roche/454)
- Utilisé comme fichiers de sortie par le projet 1000 génomes
- Fichier texte tabulé (SAM)
- Contient deux sections:
  - Entête (optionnel)
  - Alignement

Entête

- Entête commence par @
- Se trouve au début du fichier
- Tag
  - @HG : (version du format, …)
  - @SQ : Liste des séquences de référence (une ligne par séquence de référence utilisée)
  - @RG : group de lecture
  - @PG : nom du programme

Alignement

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

Alignement FLAG

Bit	Description
0x1	template having multiple segments in sequencing
0x2	each segment properly aligned according to the aligner
0x4	segment unmapped
0x8	next segment in the template unmapped
0x10	SEQ being reverse complemented
0x20	SEQ of the next segment in the template being reversed
0x40	the first segment in the template
0x80	the last segment in the template
0x100	secondary alignment
0x200	not passing quality controls
0x400	PCR or optical duplicate
0x800	supplementary alignment

- Utilisé pour filtrer un fichier SAM/BAM
- http://picard.sourceforge.net/explain-flags.html

Alignement

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

Alignement : CIGAR

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

- Comprendre l'alignement

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001  163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001   83 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Alignement :Tags additionnels

Tag	Meaning
NM	Edit distance
MD	Mismatching positions/bases
AS	Alignment score
BC	Barcode sequence
X0	Number of best hits
X1	Number of suboptimal hits found by BWA
XN	Number of ambiguous bases in the referenece
XM	Number of mismatches in the alignment
XO	Number of gap opens
XG	Number of gap extentions
XT	Type: Unique/Repeat/N/Mate-sw
XA	Alternative hits; format: (chr,pos,CIGAR,NM;)*
XS	Suboptimal alignment score
XF	Support from forward/reverse alignment
XE	Number of supporting seeds

Exemple

Entête

Alignement

```
@HD VN:1.0  SO:coordinate
@SQ SN:chr18    LN:78077248
@RG ID:1    PL:illumina PU:TGACCA   LB:R52  SM:CRN-100_4
@PG ID:bwa  PN:bwa  VN:0.6.1-r112-master
HWI-ST1136:79:HS026:1:1101:1345:2086    163 chr18   12145691    60
23M =   12145945    355 AGAGGGAGAGGCGGGCCATCTTT CCCFFFFFHHHHHJH0
@;778=D XT:A:U  NM:i:0  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0
MD:Z:23
```

<QNAME> <FLAG> <RNAME> <POS> <MAPQ> <CIGAR>
<MRNM> <MPOS> <ISIZE> <SEQ>
<QUAL>[<TAG>:<VTYPE>:<VALUE> [...]]

Le fichier BAM

- Le format de fichier BAM est la version compressée du fichier SAM. (Format compatible avec GZIP)
- Indexer les fichiers BAM (*.bam.bai) :
  - accélérer la recherche des alignements à une position donnée
- Ordonné par coordonnées chromosomiques

Manipulation de fichiers SAM/BAM

- Manipulation de fichiers SAM/BAM avec les API (Application Programming Interface)
  - Samtools (en C)
  - Picard (en Java)
  - Pysam (en python)
  - …
- Attention: les différentes API ne proposent pas toutes les mêmes fonctionnalités

Samtools

- Permet de créer et d'indexer des fichiers BAM à partir de fichier SAM
- Calculer des statistiques d'alignement
- Enlever les duplicats de PCR
- Fusionner des fichiers SAM/BAM
- Visualiser des alignement à partir des fichiers BAM
- Détecter des SNP
- Détecter des petits indels

Picard

- Modules complémentaires à Samtools
  - Identifications de duplicats
  - Ordonner des fichiers BAM
  - Ajouter des informations de groupe de lecture
- Contient plus d'outils de conversion de format
- Pas de visualisation d'alignements possible
- Pas de détection de variants

Manipulation des fichiers SAM

Programs processing SAM/BAM

- BAMTools, C++ APIs (not based on C APIs) for processing BAM files.
- BamView, BAM alignment viewer. It can be integrated to Artemis.
- BEDTools, a software package for manipulating BED files, with some utilities working with BAM. Built upon BAMTools.
- BreakDancer, structural variation caller for paired-end data.
- DNAA, DNA Analysis package including various post-alignment processing.
- Gambit, graphical BAM alignment viewer.
- GAP5, sequence assembly viewer, editor and analyzer. Capable of importing BAM files and outputing SAM.
- GATK, the Genome Analysis Toolkit. Rich funtionality including an accurate SNP caller. Built upon Picard.
- GBrowse, generic genome browser. Experimental SAM/BAM alignment viewing. Built upon Perl APIs.
- GenomeView, a Java based genome browser.
- IGB, the Integrated Genome Browser for various data formats.
- IGV, the Integrative Genomics Viewer, supporting multiple tracks and genome annotations. Built upon Picard.
- LookSeq, web-based alignment/annotation viewer.
- MagicViewer, graphical BAM alignment viewer.
- samToBed by Aaron Quinlan. Converting alignments in the SAM format to the BED format.
- Savant, a Java based genome browser.
- Tablet, alignment viewer. It also supports tons of other alignment/assembly formats.
- Vancouver Short Read Analysis Package (in particular FindPeaks), post alignment processing of new sequencing data.

SAM: un format universel

Aligners natively generating SAM

- BFAST, `Blat-like Fast Accurate Search Tool' for Illumina and SOLiD reads.
- Bowtie. Highly efficient short read aligner. Natively support SAM output in recent version. A convertor is also available in samtools-C.
- BWA, Burrows-Wheeler Aligner for short and long reads.
- GEM library. Short read aligner. Convertor provided by the developers.
- Karma, the K-tuple Alignment with Rapid Matching Algorithm.
- LASTZ, aligner for both short and long reads.
- Mosaik. The latest version support SAM output.
- Novoalign. An accurate aligner capable of gapped alignment for Illumina short reads. Academic free binary. Convertor is also available in samtools.
- SNP-o-matic, short read aligner and SNP caller.
- SOLiD BaseQV Tool. Developed by Applied Biosystems for converting SOLiD output files.
- SSAHA2 (since v2.4). Classical aligner for both short and long reads.
- Stampy, by Gerton Lunter. An accurate read aligner capable of gapped alignment for Illumina short reads. Used for indel discovery on the 1000 genomes data.
- TopHat for mapping short RNA-seq reads bridging exon junctions.

BWA

- Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60. [PMID: 19451168]
- Rapide et peu gourmand en ressources
- Supporte l'alignement de lectures avec des insertions/délétions (indels)
- Supporte les séquençages simples (single end) et pairés (paired end)
- Nécessite des données de bonne qualité
- Fonctionne avec un nombre limité d'erreurs (2 pour 32bp, 4 pour 100 bp, ...)
- Nécessite d'indexer les séquences de référence (accélérer la recherche)

Partie pratique n°1

Partie pratique n°2

Partie pratique n°3

Partie pratique n°4

QC : Alignement

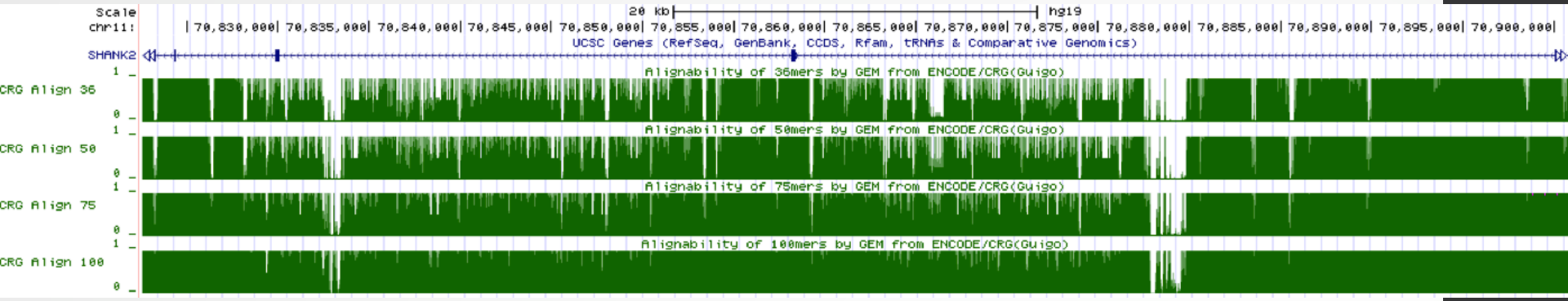
- Pourcentage de lectures alignées sur le génome de référence
  - Si trop faible: Contamination? Mauvais génome utilisé? Mauvaise qualité des lectures?

Partie pratique n°5

Régions pouvant poser problème

- Régions de faibles compléxités (homopolymères)
- Régions dont les séquences sont représentées plus d'une fois dans le génome (répétitions…)
  - Alignabilité
    - dépend de la longueur des lectures
    - meilleure si données pairées

Alignabilité

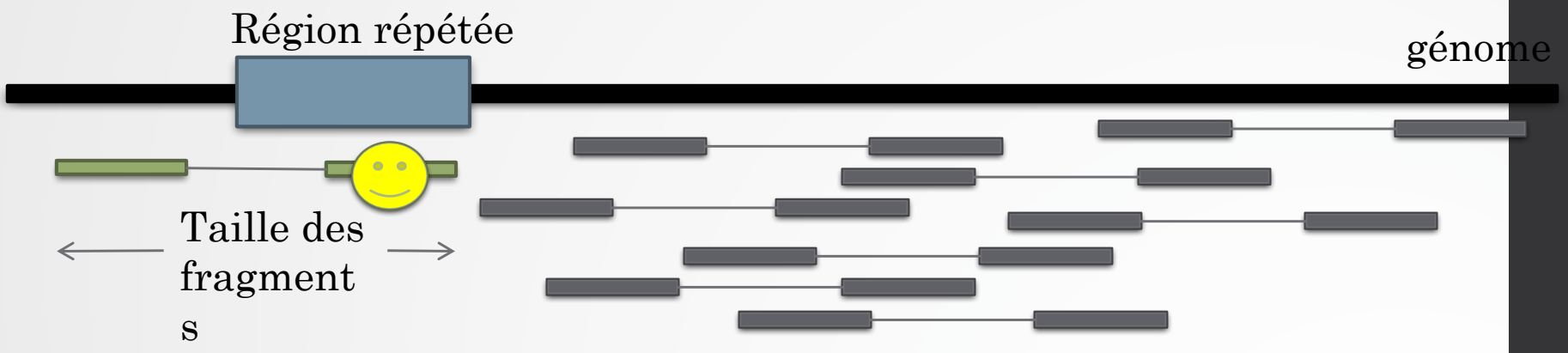


Les régions répétées

- Les lectures s'alignant dans les régions répétées ne peuvent pas être gardées pour l'analyse (introduction de biais)

Les régions répétées

- Les lectures s'alignant dans les régions répétées ne peuvent pas être gardées pour l'analyse (introduction de biais)



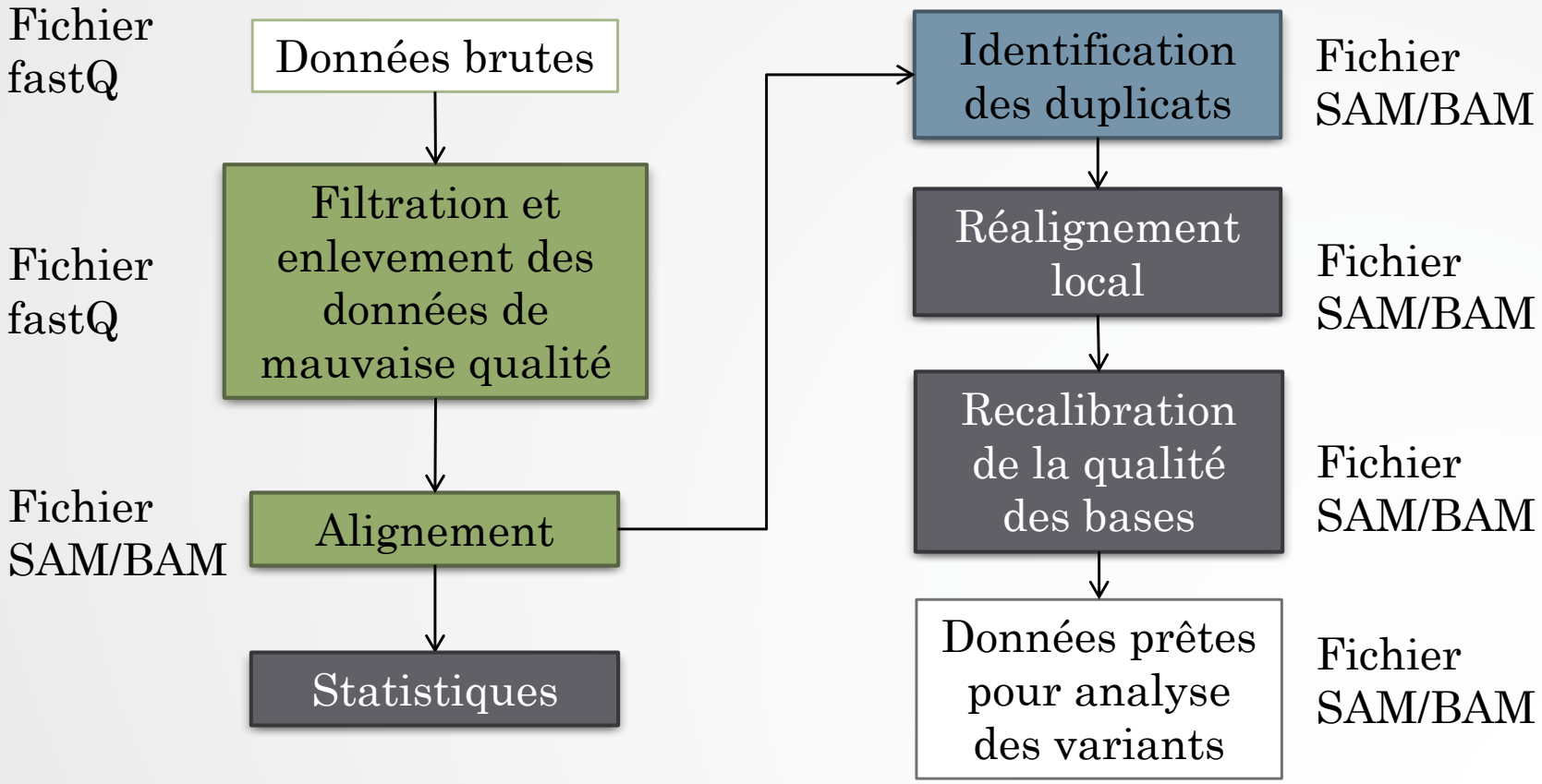
- Avantage des lectures pairées!

Les lectures issues de régions répétées

- Comment les détecter?
  - BWA donne une qualité d'alignement de 0 à des lectures s'alignant à plus d'une position
  - En utilisant les flags des fichiers SAM

Partie pratique n°6

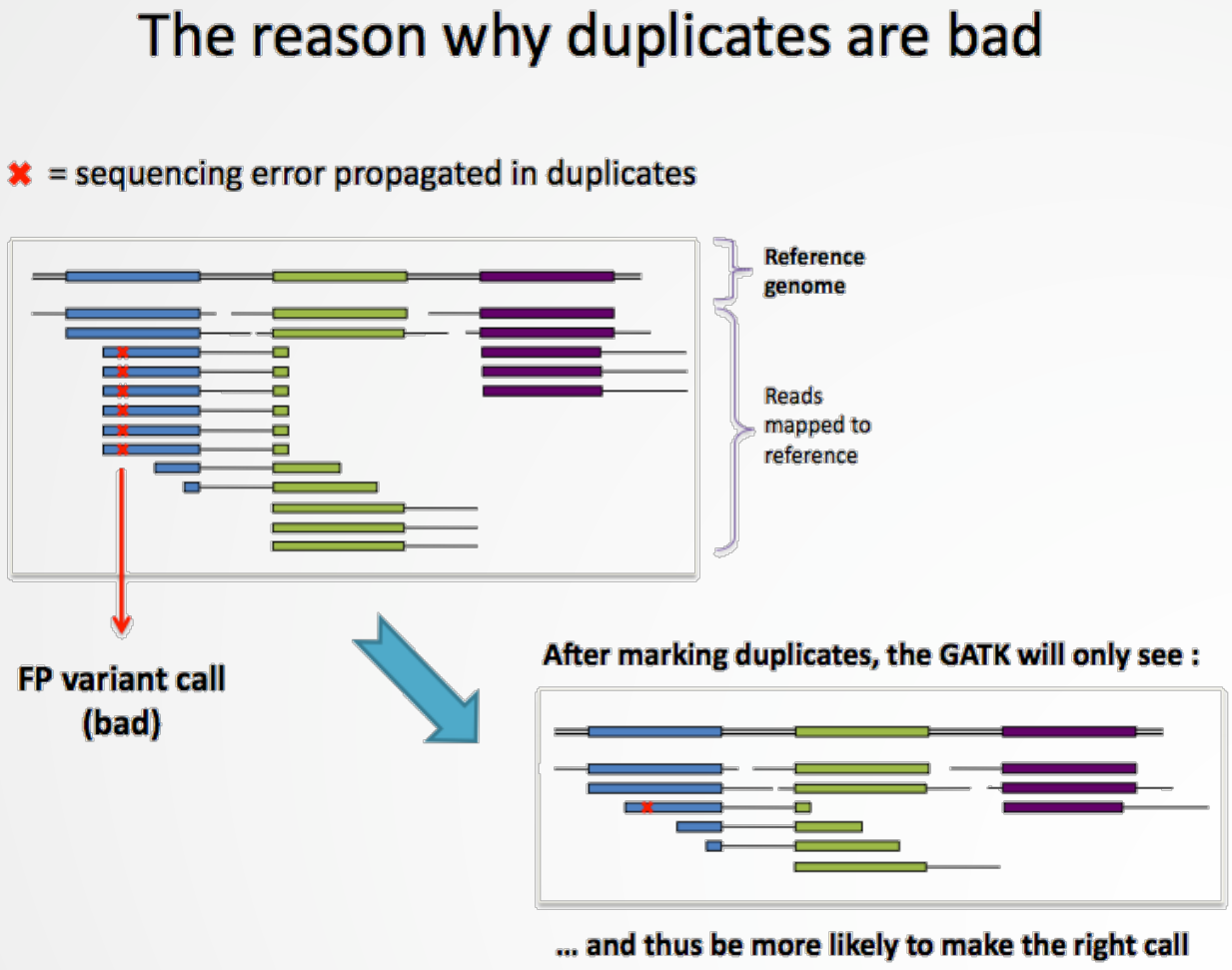
Processus


Fichier fastQ
Données brutes
Filtration et enlevement des données de mauvaise qualité
Fichier fastQ
Fichier SAM/BAM
Alignement
Statistiques
Identification des duplicats
Fichier SAM/BAM
Réalignement local
Fichier SAM/BAM
Recalibration de la qualité des bases
Fichier SAM/BAM
Données prêtes pour analyse des variants
Fichier SAM/BAM

Les lectures dupliquées

- Lectures dupliquées :
  - Séquences ayant la même séquence nucléotidique
  - Alignées sur le même chromosome avec la même position de début et de fin d'alignement et dans le même sens de lecture
  - Ont le même CIGAR
  - Cause :
    - PCR pendant la préparation de la librairie (duplicats moléculaire)
    - Même cluster lu deux fois (duplicats optiques)

Les lectures dupliquées



Source: GATK

Les lectures dupliquées

- Lectures dupliquées :
  - Avantage des lectures pairées

Les lectures dupliquées

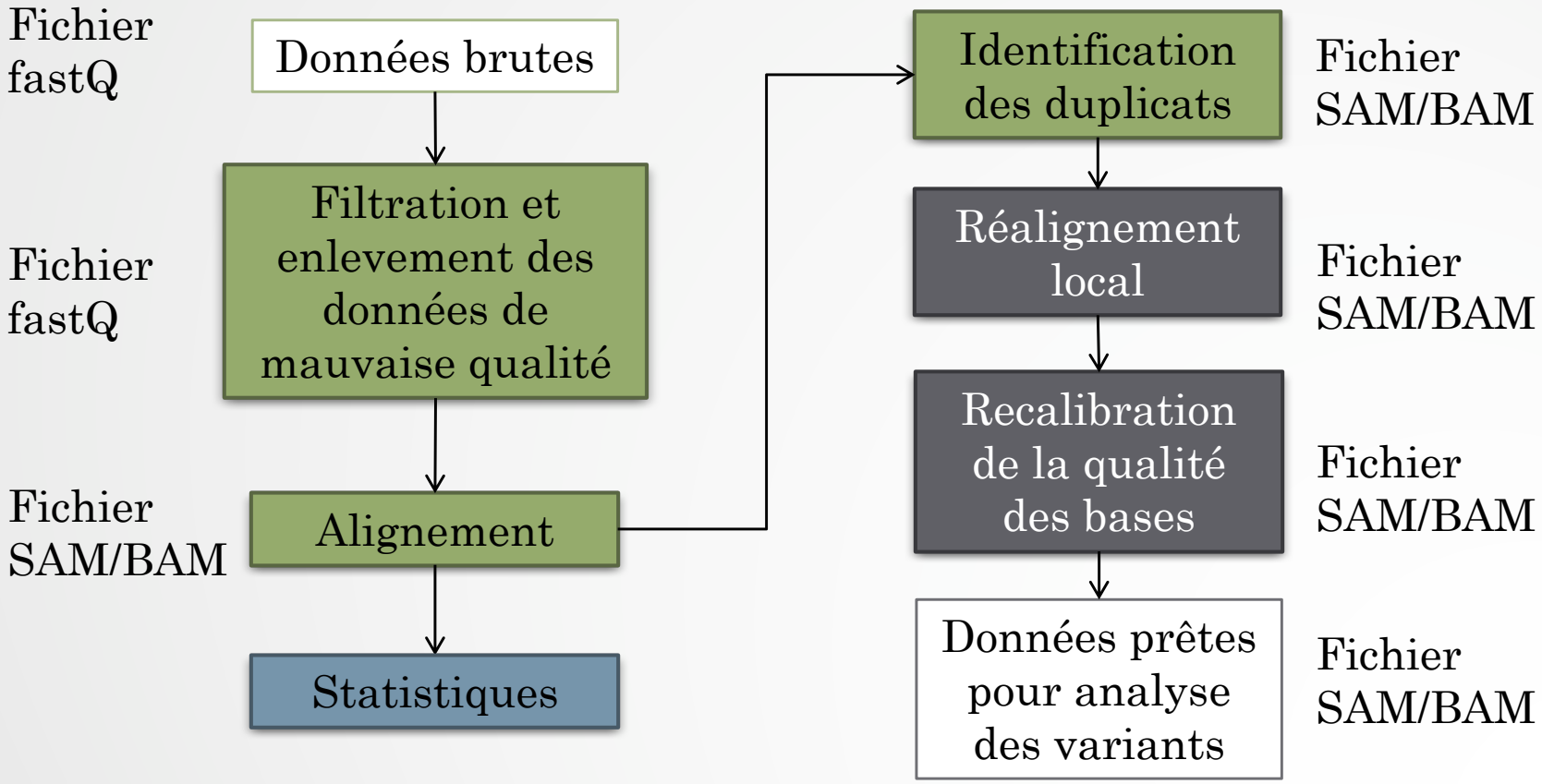
- Lectures dupliquées :
  - Avantage des lectures pairées

Partie pratique n°7

Partie pratique n°8

Estimation de la couverture

Processus


Fichier fastQ
Données brutes
Filtration et enlevement des données de mauvaise qualité
Fichier fastQ
Fichier SAM/BAM
Alignement
Statistiques
Identification des duplicats
Fichier SAM/BAM
Réalignement local
Fichier SAM/BAM
Recalibration de la qualité des bases
Fichier SAM/BAM
Données prêtes pour analyse des variants
Fichier SAM/BAM

Couverture nucléotidique

- Nombre de lectures présentes à une position donnée
- Meilleure sera la couverture, meilleure sera la détection des variants (Encore plus vrai pour les variants hétérozygotes)

Couverture minimale pour la détection des hétérozygotes

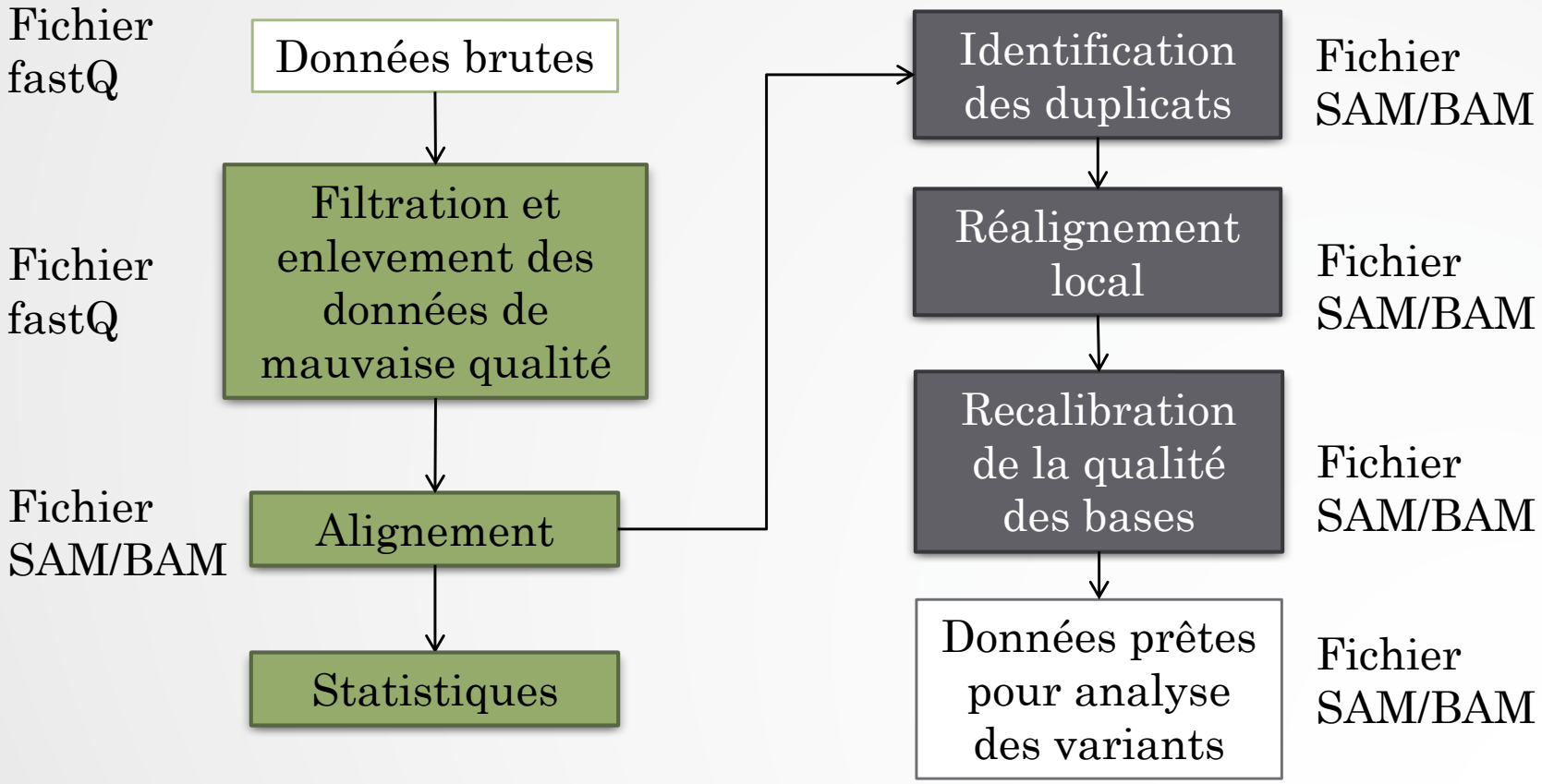
		sequencing error filter level						
Power (Q)		5%	10%	15%	20%	25%	30%	35%
90.00%	(10)	4	4	4	7	7	12	24
95.00%	(13)	5	5	8	8	11	18	30
99.00%	(20)	7	7	11	17	19	35	54
99.50%	(23)	8	12	12	18	26	42	71
99.90%	(30)	10	14	18	27	38	61	110
99.95%	(33)	11	15	19	28	42	68	117
99.99%	(40)	14	18	28	34	54	83	148
99.995%	(43)	15	19	30	42	61	92	165
100.00%	(50)	17	25	36	51	70	109	194

doi:10.1371/journal.pone.0025531.t001

Partie pratique n°9

Estimation de l'efficacité de capture

Processus


Fichier fastQ
Données brutes
Filtration et enlevement des données de mauvaise qualité
Fichier fastQ
Fichier SAM/BAM
Alignement
Statistiques
Identification des duplicats
Fichier SAM/BAM
Réalignement local
Fichier SAM/BAM
Recalibration de la qualité des bases
Fichier SAM/BAM
Données prêtes pour analyse des variants
Fichier SAM/BAM

Efficacité de capture

- Comment évaluer l'efficacité de capture?
  - Nombre de lectures tombant dans les régions que l'on a cherché à capturer
- Besoin d'un fichier avec les coordonnées chromosomiques des régions que l'on a cherché à capturer (fichier au format BED)

Le format BED

- Fichier texte tabulé
- Minimum de 3 colonnes
  - Col 1 : Chromosome
  - Col 2 : Position de début de la région
  - Col 3 : position de fin de la région
  - Col 4 : Nom de la région (optionnel)
  - Col 5 : Score (optionnel)
  - Col 6 : Orientation (optionnel)
  - … jusqu'à 12 colonnes

Et comme pour SAM il y a SAMtools…

- Pour les fichiers BED il y a… BEDtools
  - Calculer l'intersection entre deux fichiers BED
  - Calculer le nombre de lecture par annotation génomique (exon, intron, …)
  - …
- (BEDtools peut également gérer les fichiers BAM)

Partie pratique n°10

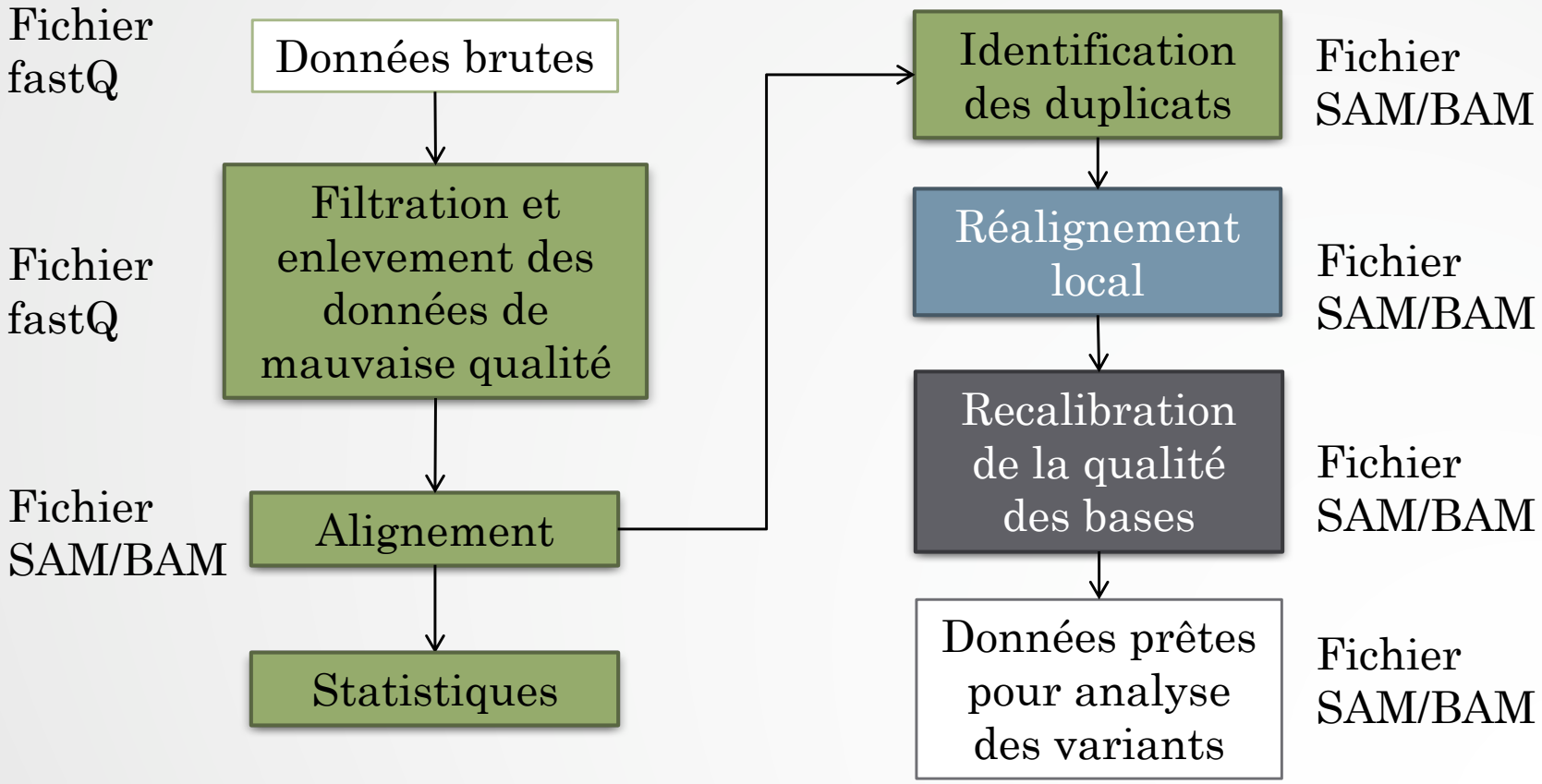
- Le fichier de sortie est de la forme :
  - Col 1 : chromosome
  - Col 2 : début de la région
  - Col 3 : fin de la région
  - Col 4 : Nombre de lecture dans la région
  - Col 5 : Taille de la région
  - Col 6 : Nombre de nucléotide couvert dans la région
  - Col 7 : Pourcentage de bases couvertes
- Question :
  - Est ce que toutes les régions sont couvertes?
  - Combien de lectures sont capturées dans la région?
  - Quel est le taux d'efficacité de la capture?

Format de fichier

- Il existe un grand nombre de format de fichiers pour stocker des données génomiques
- http://genome.ucsc.edu/FAQ/FAQformat.html
- Conversion d'un fichier avec un grand nombre d'informations vers un fichier contenant la quantité d'information nécessaire
  - Gain de place
- Pour convertir d'un format vers un autre, il faut que le fichier d'origine contienne les informations nécessaires
  - Ex : Bam -> BED (Perte d'information)
  - ~~Ex : BED -> BAM~~ (Information manquante)
  - ~~Ex : Fastq -> BED~~ (Il faut aligner les séquences!)

Raffinement des alignements

Processus


Fichier fastQ
Données brutes
Filtration et enlevement des données de mauvaise qualité
Fichier fastQ
Fichier SAM/BAM
Alignement
Statistiques
Identification des duplicats
Fichier SAM/BAM
Réalignement local
Fichier SAM/BAM
Recalibration de la qualité des bases
Fichier SAM/BAM
Données prêtes pour analyse des variants
Fichier SAM/BAM

Réalignement autour des indels

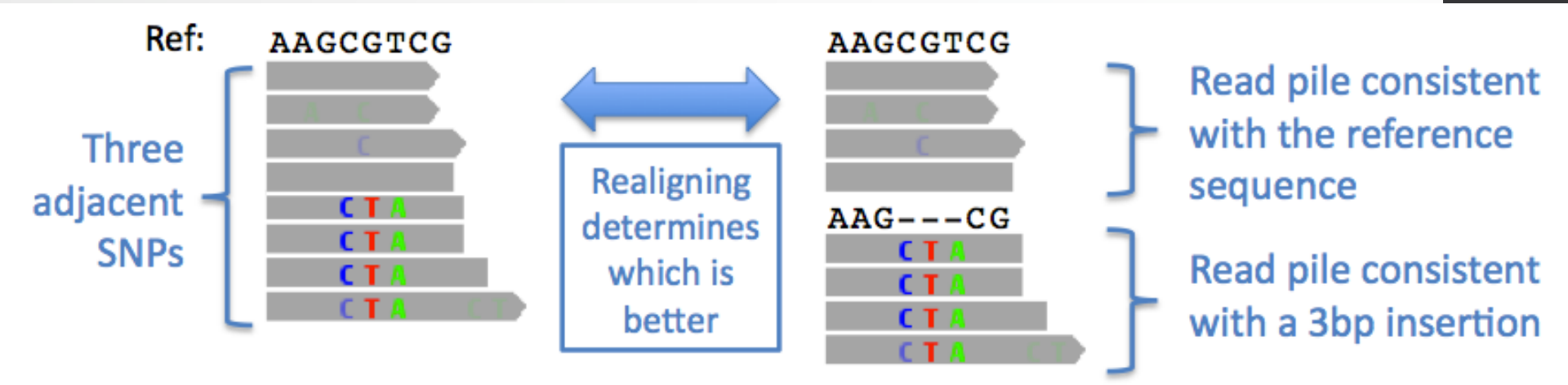
- Problème d'alignement :
  - Dans les régions de faibles complexités
  - Autour des indels
- Ces mesappariements détectés comme des variants peuvent biaiser les modèles statistiques utilisées lors de la détection des variants

Réalignement autour des indels

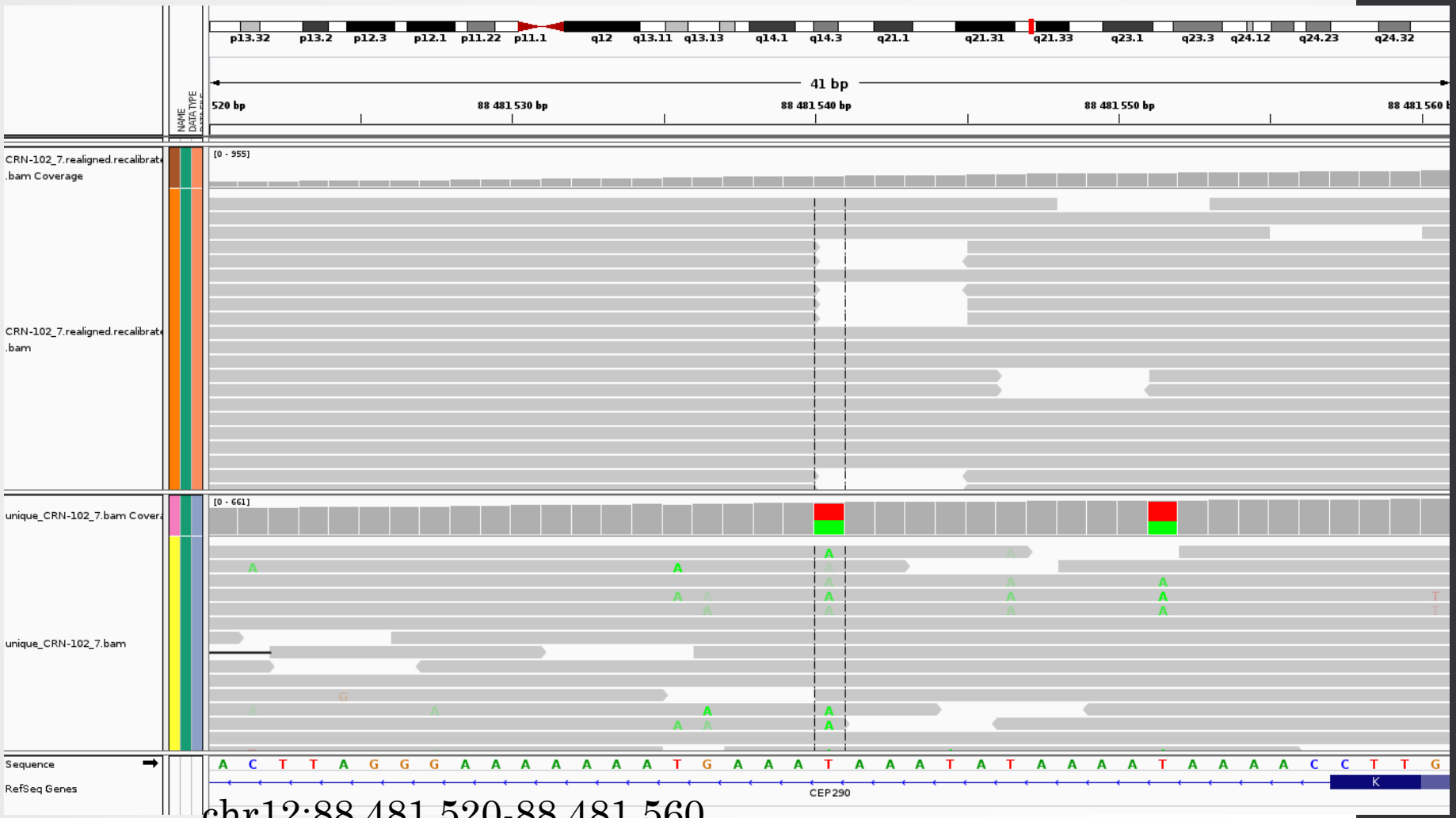
- 3 différents type de régions sont ciblées par le réalignement des indels
  - Les indels connus (1000 génomes, dbSNP…)
  - Les indels détectés dans les échantillons analysés (en utilisant le CIGAR)
  - Sites avec des indels supposés

Comment ça marche

- 1. Trouver la meilleur séquence consensus dans une région donnée (la région inclue un maximum d'un indel)
- 2. Le score de la séquence consensus est égal à la somme des qualités des bases ayant un mesappariement
- 3. Si le score du consensus est meilleur que celui de l'alignement originel alors le nouvel alignement est conservé.



Exemple sur un autre échantillon



chr12:88,481,520-88,481,560

IGV

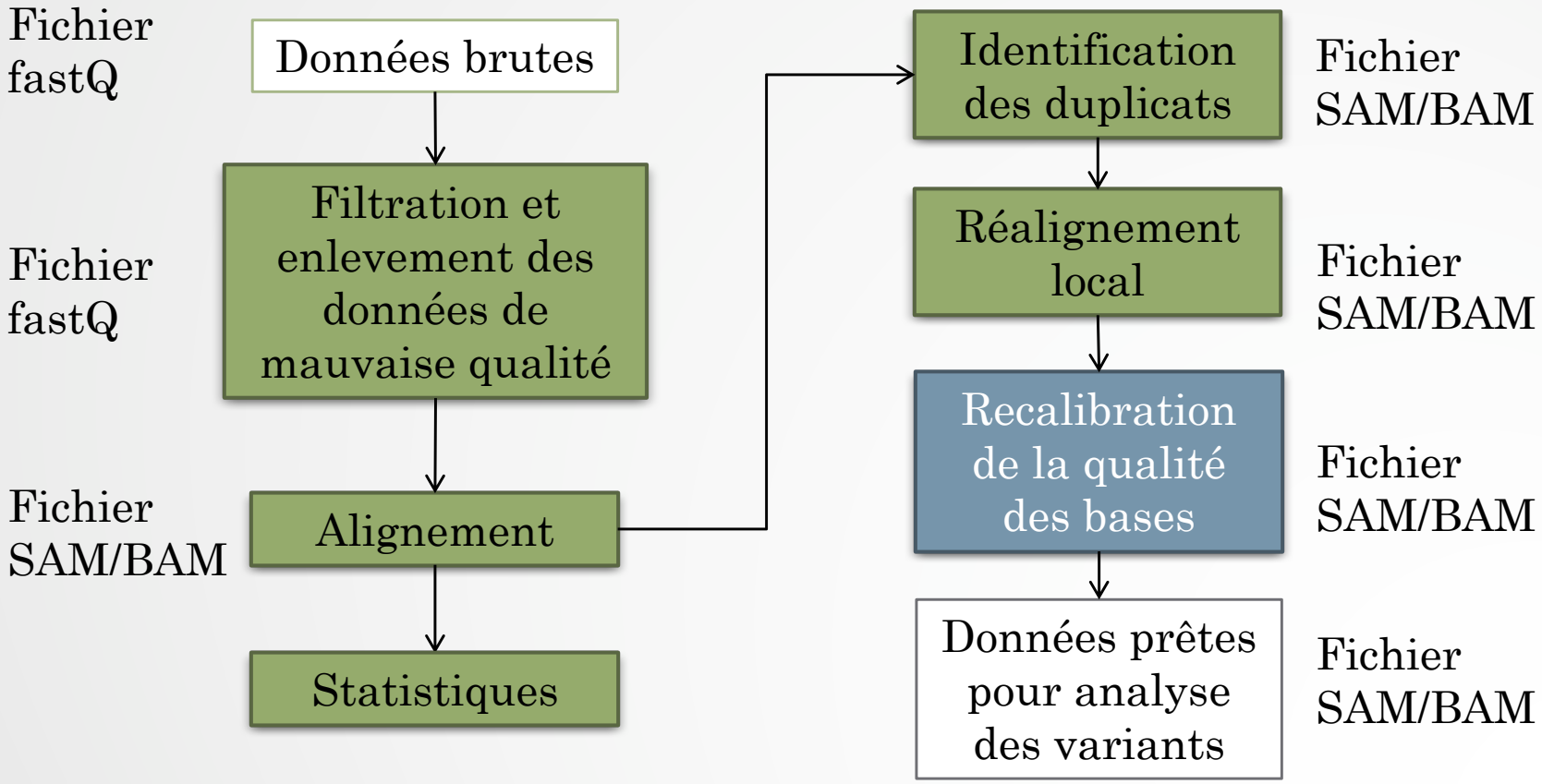
- Navigateur de Génomes
- Développé par le Broad Institute
- Le meilleur pour la visualisation de données de reséquençage

IGV

- Génome à charger : hg19 (Genomes/Load from server)
- Fichier à charger (File/Load from File):
  - CRN-107_RG.bam (avant réalignement local)
  - CRN-107.realigned.bam (après réalignement local)
- Pour visualiser un fichier Bam il faut avoir son fichier bai (index) correspondant
- chr4:122,766,732-122,766,861

Recalibration de la qualité des bases

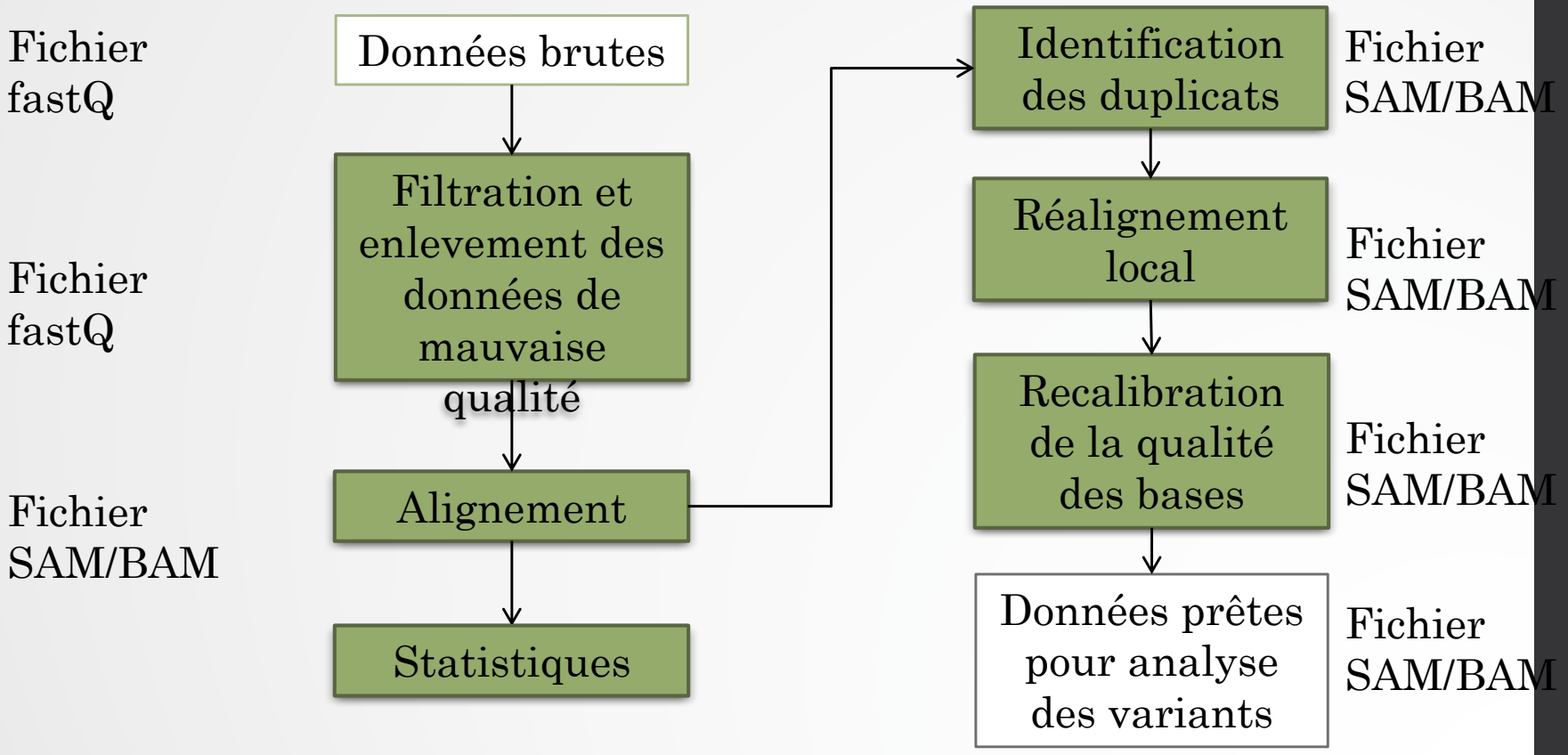
Processus


Fichier fastQ
Données brutes
Filtration et enlevement des données de mauvaise qualité
Fichier fastQ
Fichier SAM/BAM
Alignement
Statistiques
Identification des duplicats
Fichier SAM/BAM
Réalignement local
Fichier SAM/BAM
Recalibration de la qualité des bases
Fichier SAM/BAM
Données prêtes pour analyse des variants
Fichier SAM/BAM

Recalibration des qualités des bases

- A quoi ça sert?
  - Corriger les biais d'assignement des scores de qualités des séquenceurs
- Comment ça marche?
  - Les outils scannent les lectures alignées et cherchent les positions qui sont des mesappariements (nucléotides différents du génome)
  - Si les mesappariements ne sont pas des variants connus alors l'outil considère que c'est une erreur de séquençage
  - Calcul de statistiques sur tous les mesappariements détectés et sur leur contexte dans la lecture (dinucléotide, position dans la lecture…)
  - Correction de la valeur de qualité pour qu'elle corresponde à la qualité observée

Processus


Fichier fastQ
Données brutes
Filtration et enlevement des données de mauvaise qualité
Fichier fastQ
Alignement
Fichier SAM/BAM
Statistiques
Identification des duplicats
Fichier SAM/BAM
Réalignement local
Fichier SAM/BAM
Recalibration de la qualité des bases
Fichier SAM/BAM
Données prêtes pour analyse des variants
Fichier SAM/BAM

Références

- Mark A. DePristo, Eric Banks, Ryan Poplin, Kiran V. Garimella, Jared R. Maguire, Christopher Hartl, Anthony A. Philippakis, Guillermo del Angel, Manuel A. Rivas, Matt Hanna, Aaron McKenna, Tim J. Fennell, Andrew M. Kernytsky, Andrey Y. Sivachenko, Kristian Cibulskis, Stacey B. Gabriel, David Altshuler, and Mark J. Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics , 43(5):491{498, May 2011.
- Heng Li and Richard Durbin. Fast and accurate short read alignment with BurrowsWheeler transform. Bioinformatics , 25(14):1754{1760, July 2009.
- Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence Alignment/Map format and SAMtools. Bioinformatics , 25(16):2078{2079, August 2009.