Big data pour les données épigénétiques : illustration par des données pan-génomiques de méthylation de l'ADN

Adebiotech / EPIGEN 2018

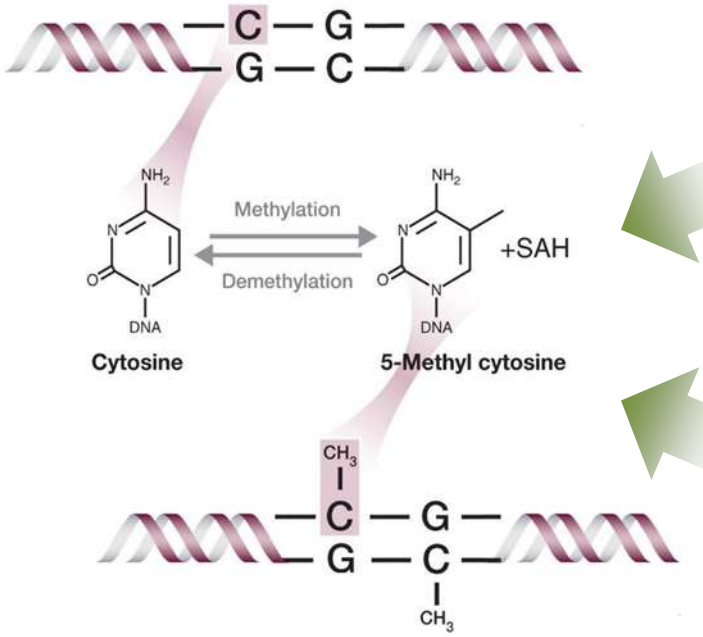
Hélène KIEFER BDR UMR 1198 Jouy-en-Josas

14 / 03 / 2018

Les facteurs déterminant la méthylation de l'ADN

1 **Séquence (Animaux : CpG prépondérant)**

2 **Etat de la cellule (tissu, environnement, physiologie…)**

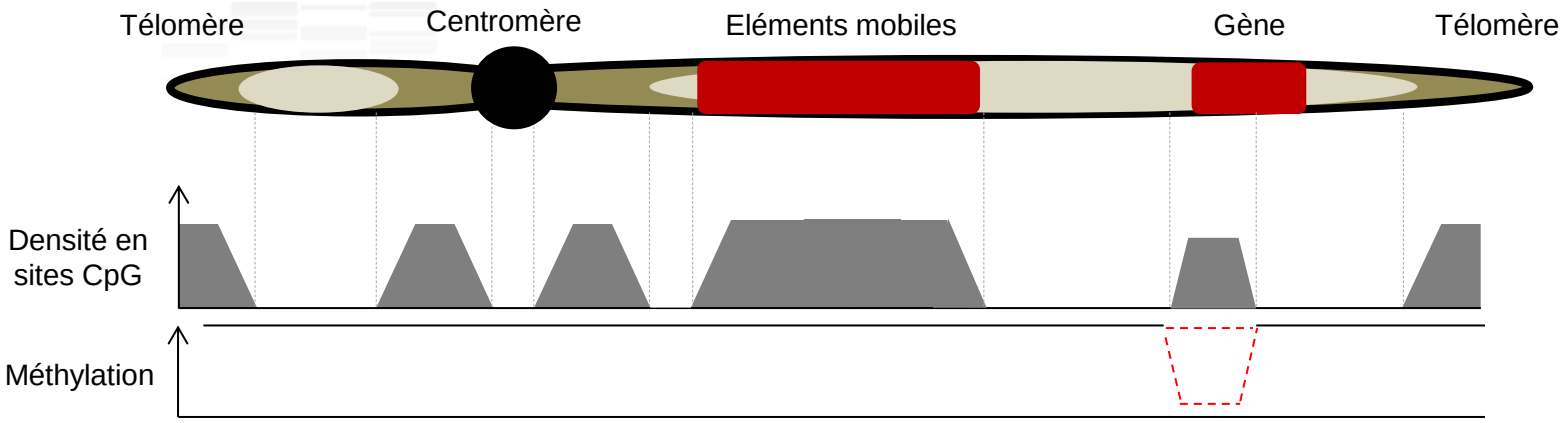


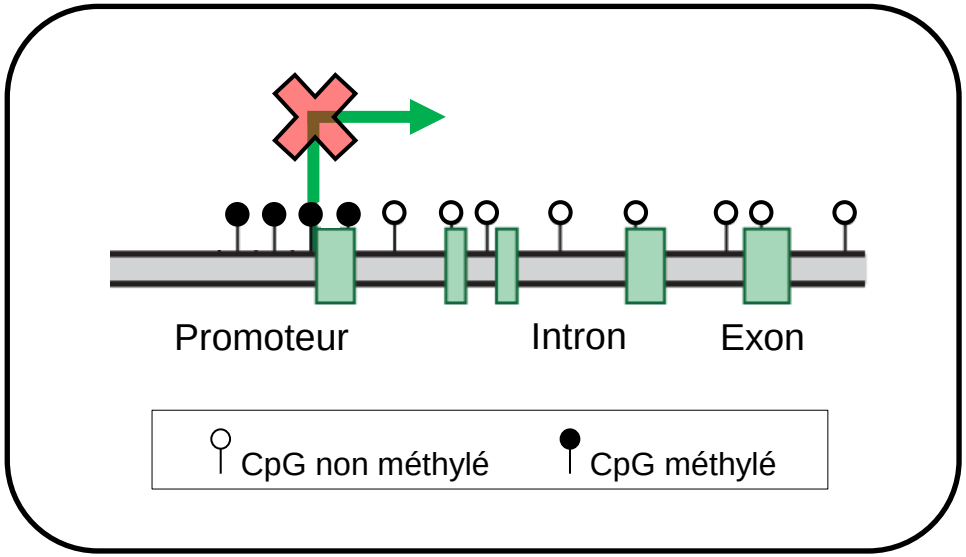
3 **Déméthylation active et passive**

4 **Facteurs stochastiques**

Adapté de: https://www.caymanchem.com/images/articles/screen/2153-1.jpg

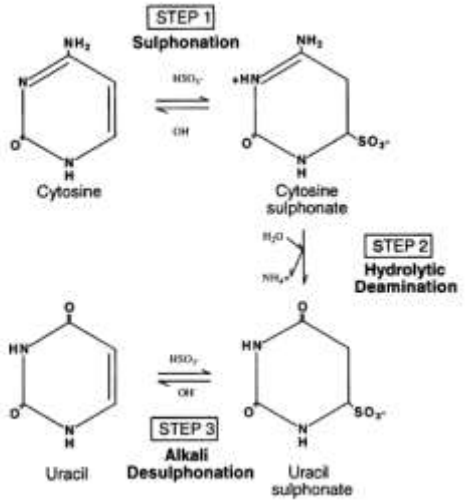
Méthylation de l'ADN et structure/fonction du génome





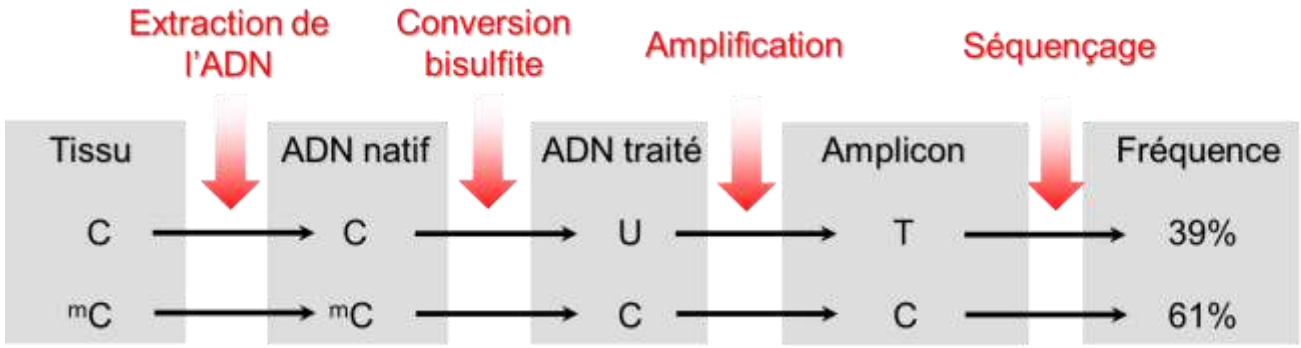
Propriétés de biomarqueur de la méthylation de l'ADN

- Intervient dans de nombreuses fonctions biologiques
- Equilibre entre stabilité et plasticité → « archivage » à long terme d'évènements affectant le génome
- Modification covalente de la molécule d'ADN génomique → information relativement facile d'accès par rapport à d'autres marques épigénétiques (ex. modifications d'histones)
- Information précise « à la base près » comme pour le génotype
- Technologie PacBio de séquençage haut-débit → information directe sur l'état de méthylation
- Technologie Illumina de séquençage haut-débit → traitement chimique de l'ADN génomique au bifulfite de sodium pour distinguer les C méthylées des C non méthylées

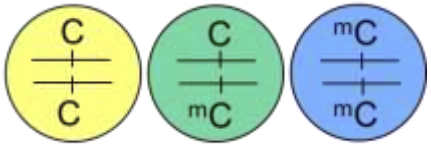


Clark et al, Nucleic Acid Research 1994

Challenge 1 : la méthylation de l'ADN est une variable continue

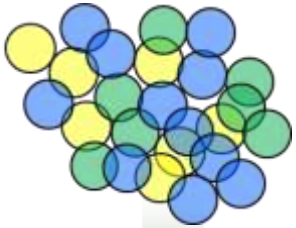


Cellules

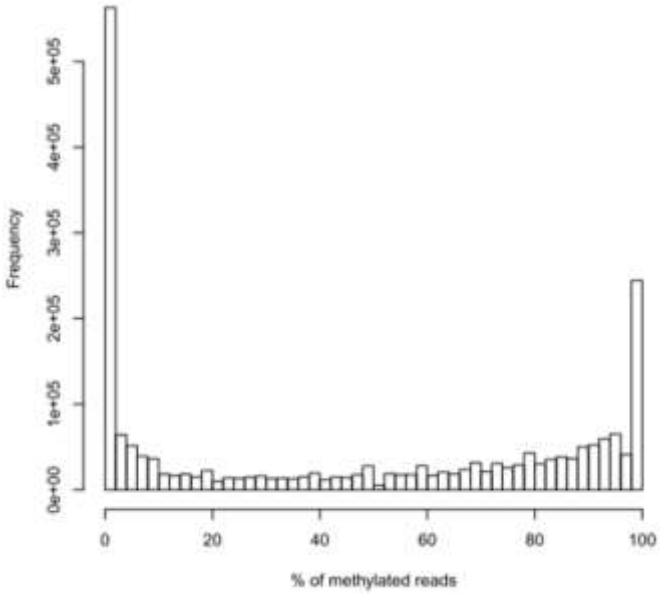


A l'échelle d'une cellule, peu de molécules séquencées suffisent

Tissu



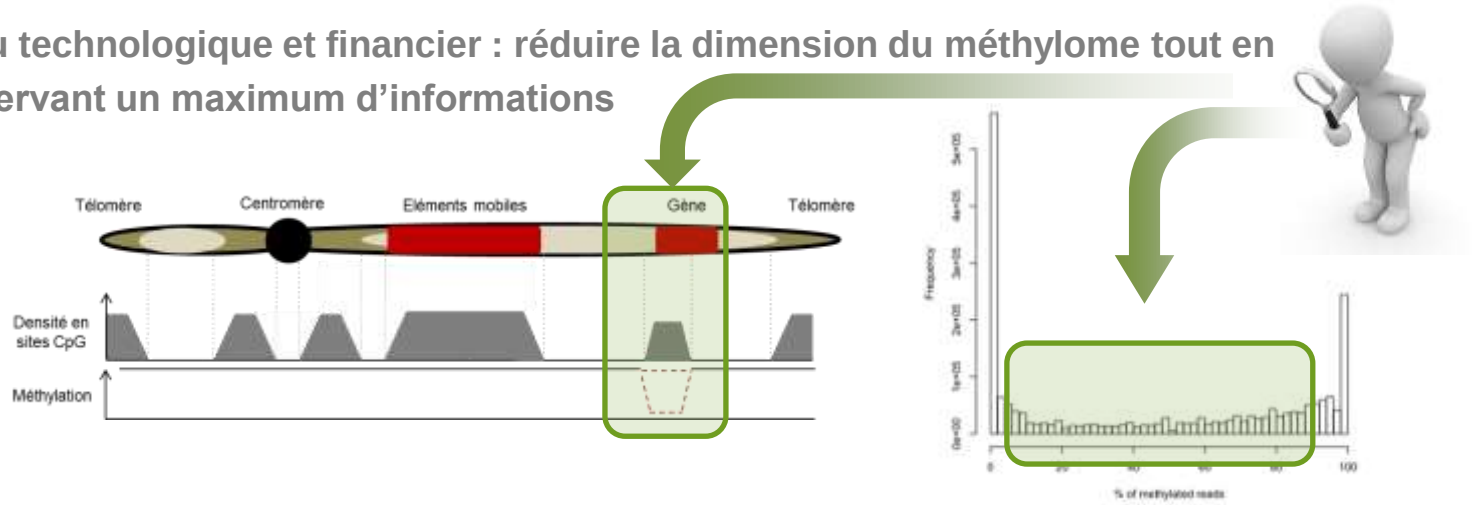
≥10 molécules séquencées sont nécessaires



Challenge 2 : « Big Data » !

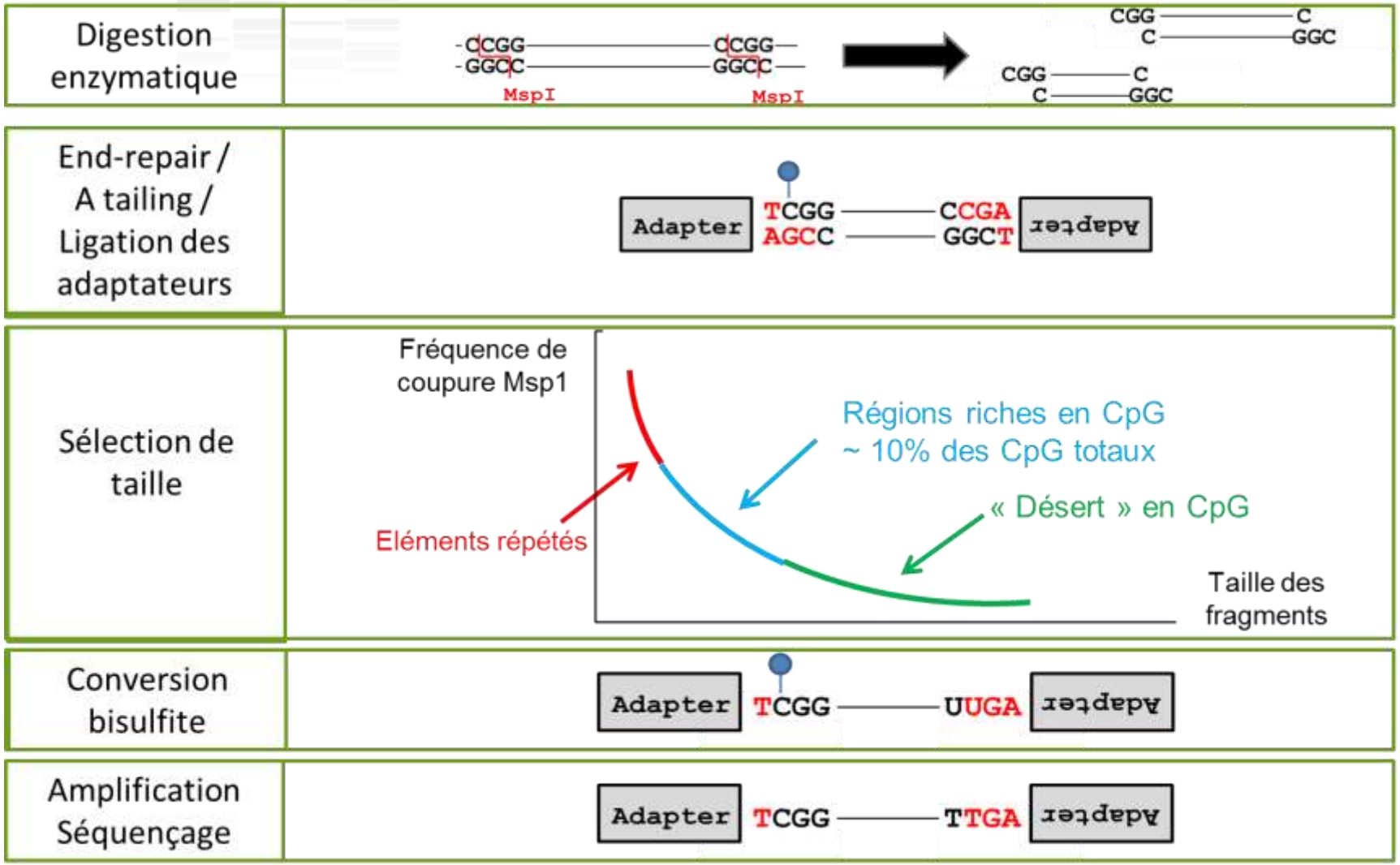
Species	Human	Mouse	Pig	Sheep	Horse	Cattle
Whole genome size (Gb)	3,1	2,7	2,8	2,6	2,5	2,7
Total number of CpG sites	29 345 332	21 867 837	30 460 432	26 376 870	29 873 125	27 203 575
CpG sites per Mb whole genome	9 466	8 099	10 840	10 068	12 094	10 075

- Méthylome : patron de méthylation de l'ensemble / d'une partie des CpG du génome
- Enjeu technologique et financier : réduire la dimension du méthylome tout en conservant un maximum d'informations



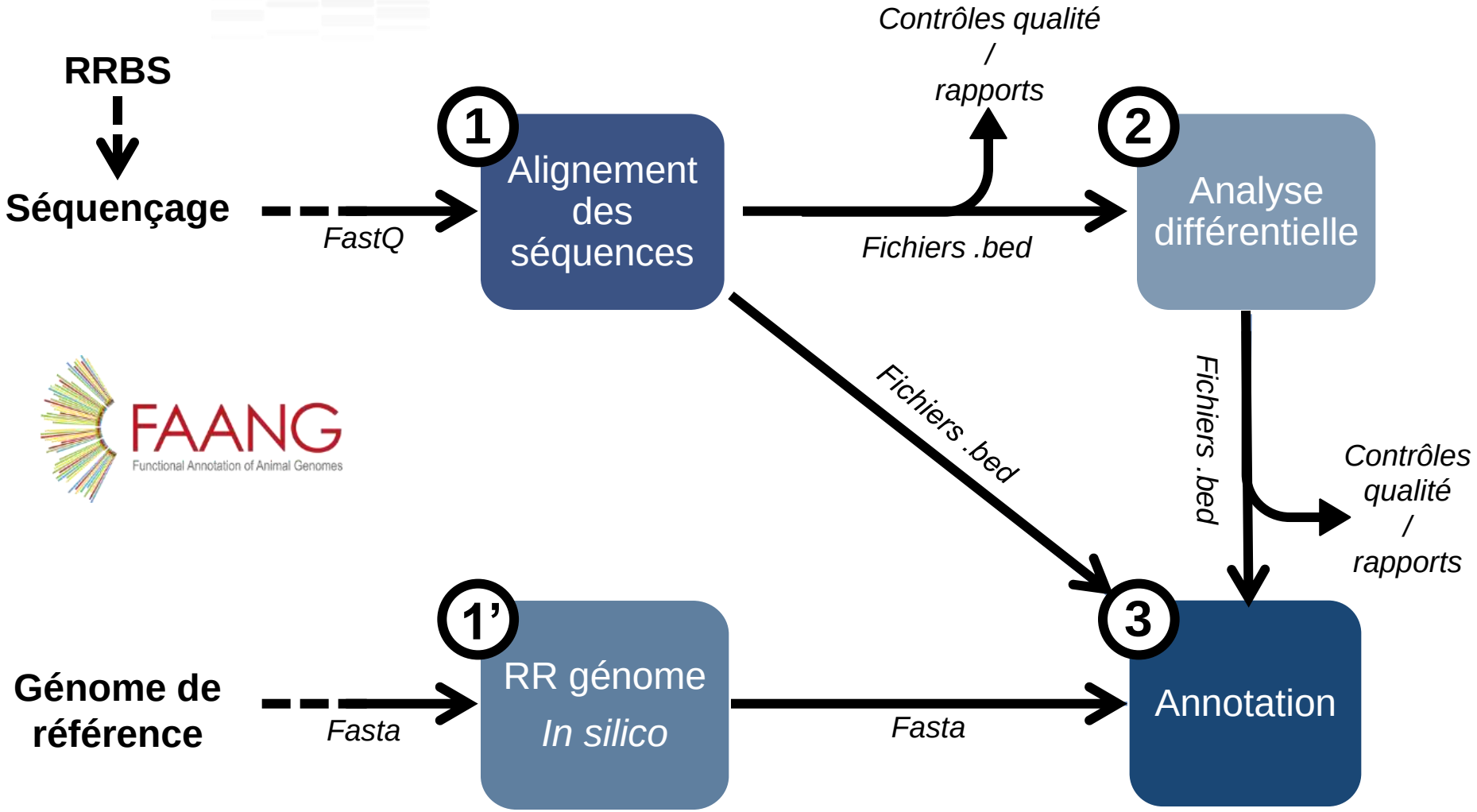
- Ex : RRBS (Reduced Representation Bisulfite Sequencing, Gu *et al*. 2011)

Vue d'ensemble de la construction de librairies RRBS



Vue d'ensemble du pipeline d'analyses bioinformatiques / biostatistiques


L. Jouneau et F. Piumi



Infrastructure informatique

GENOTOUL BIOINFORMATICS HOME

The GenoToul bioinformatics facility is part of the Genotoul GIS. It has been set up in 2000. Since 2009, it is one of the 13 IBISA bioinformatics platforms. Since 2008, the plateform collaborates with the local genomic platform and processes huge volumes of data produced by second and third generation of sequencers and makes them available to biologists (ng6).

EQUIPMENT

- A computer farm : about 5000 cores (INTEL-2014, AMD-2012), 34 Tera Byte memory (3TB on a SMP machine), Infiniband interconnection (QDR), parallel file system (GPFS)
- Web servers and virtual machines hosting infrastructure
- More than 1Peta Byte disk space

SERVICES

- Access to public biological banks
- Access to generic and specific bioinformatics software pieces
- Access to web resources
- Projects (Web/VM) hosting (ask for a project hosting)
- Training

Use this link to create your user account. All questions about technical issues can be sent using one of the Ask for forms. The platform staff can help you to process your data or to develop specific databases or software packages. For any specific request please send a mail to anim.bioinfo(at)toulouse.inra.fr.

NEWS

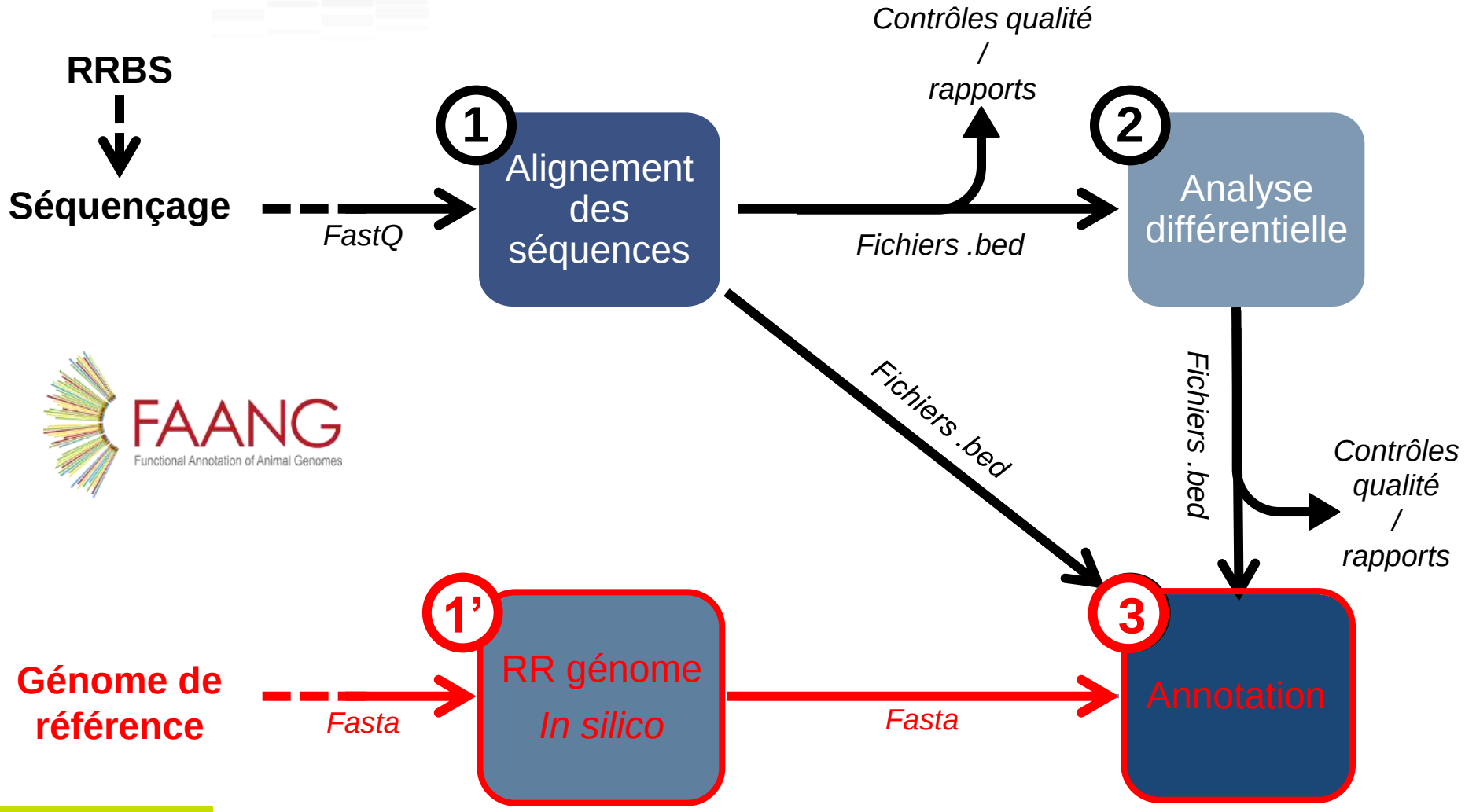
Newsletter #28
15 November 2017

Finally, a twitter account !
15 September 2017

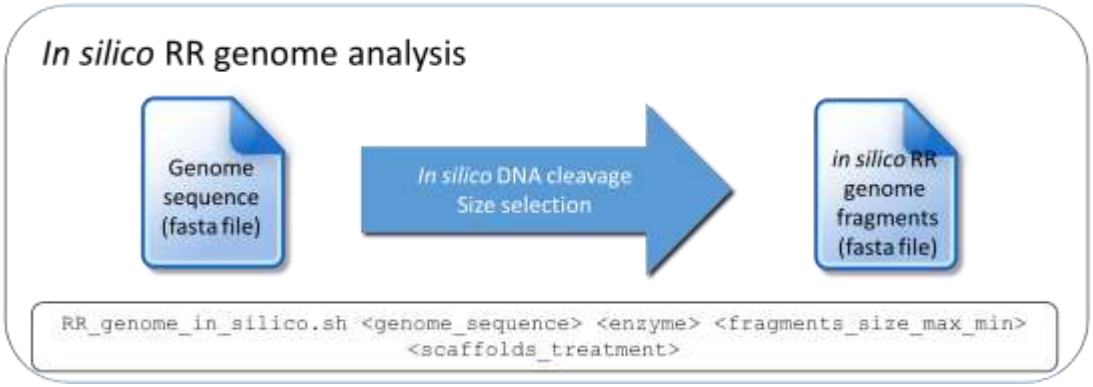
Newsletter #27
Special new cluster
29 June 2017

Vue d'ensemble du pipeline d'analyses bioinformatiques / biostatistiques


L. Jouneau et F. Piumi

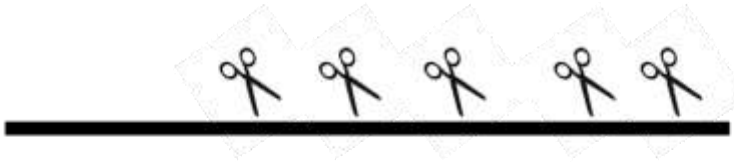


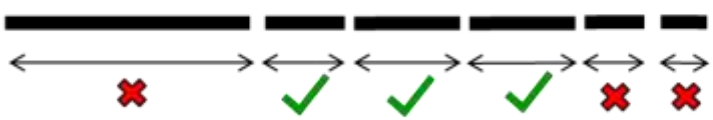
Production d'un « Reduced Representation » (RR) génome *in silico*



Fichiers d'entrée et de sortie

Commande

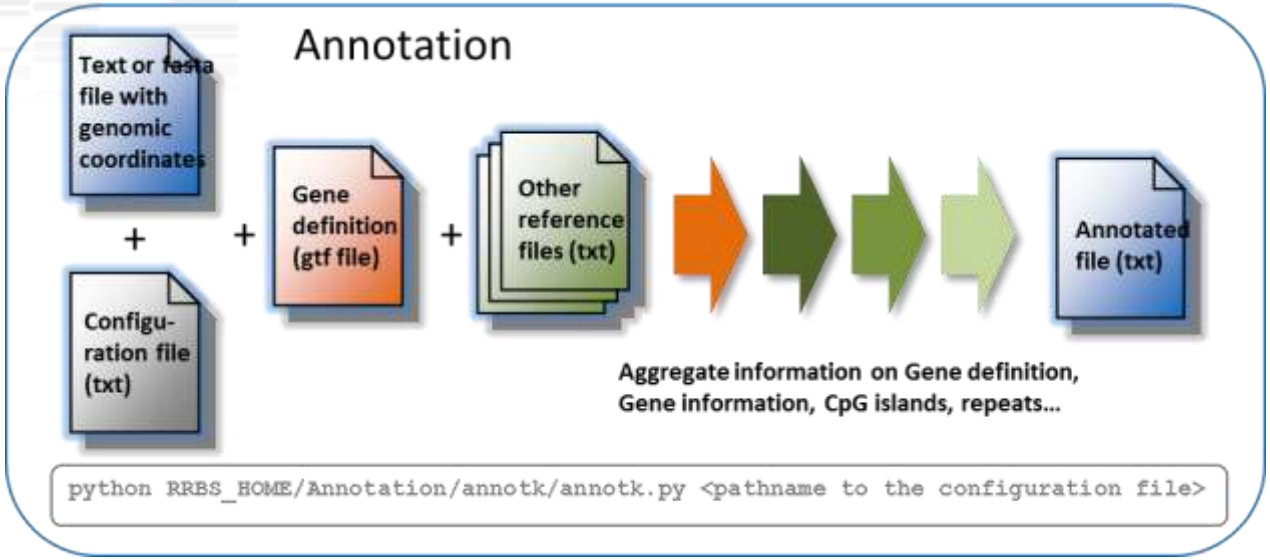


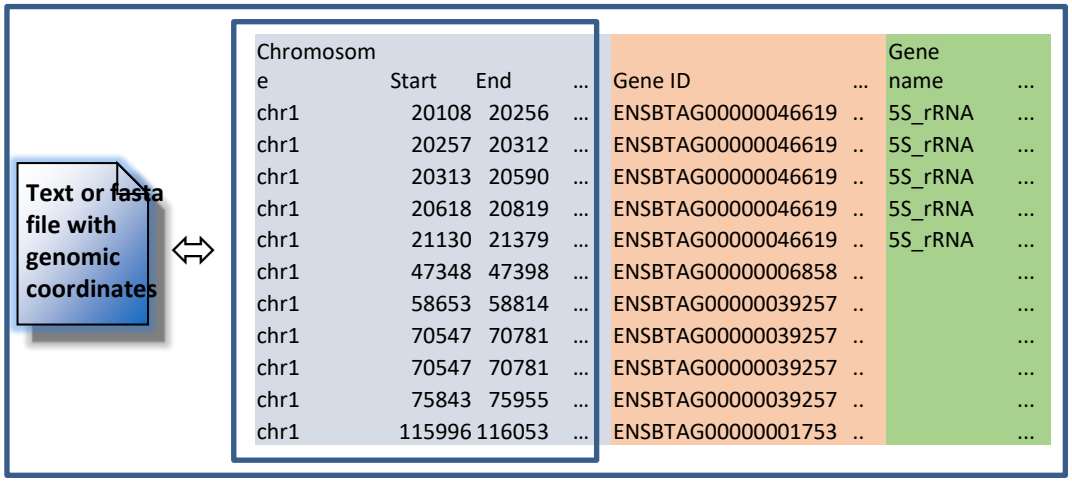


Processus

Pipeline d'annotation

Très rapide (de quelques minutes à quelques heures)





Chromosome	Start	End	...	Gene ID	...	Gene name	...
chr1	20108	20256	...	ENSBTAG00000046619	..	5S_rRNA	...
chr1	20257	20312	...	ENSBTAG00000046619	..	5S_rRNA	...
chr1	20313	20590	...	ENSBTAG00000046619	..	5S_rRNA	...
chr1	20618	20819	...	ENSBTAG00000046619	..	5S_rRNA	...
chr1	21130	21379	...	ENSBTAG00000046619	..	5S_rRNA	...
chr1	47348	47398	...	ENSBTAG00000006858	..		...
chr1	58653	58814	...	ENSBTAG00000039257	..		...
chr1	70547	70781	...	ENSBTAG00000039257	..		...
chr1	70547	70781	...	ENSBTAG00000039257	..		...
chr1	75843	75955	...	ENSBTAG00000039257	..		...
chr1	115996	116053	...	ENSBTAG00000001753	..		...

Pipeline d'annotation

Configu-ration file (txt)

```
file_to_annotate        in/DMRs_between_C1_C2.txt
file_format             tab
output_file             out/DMRs_between_C1_C2_annotated.txt
keep_scaffolds          No

theme                           Gene features
        join_type               gtf
        target_keys             Chromosome,Position
        reference_file          reference/Bos_taurus.UMD3.1.81.gtf
        nb_max_results          3
        max_dist_nearest_gene   10kb

theme                   Gene
        join_type       value
        target_keys     Gene ID
        reference_file  reference/bovine_biomart.txt
        reference_keys  Ensembl Gene ID

theme                   Repeats
        join_type       location
        target_keys     Chromosome,Position
        reference_file  reference/bovine_repeats.txt
        reference_keys  1,2,3
        min_overlap     75%


theme                   CpG islands
        join_type       location
        target_keys     Chromosome,Position
        reference_file  reference/bovine_CGI.txt
        reference_keys  1,2,3
        min_overlap     0%
        nb_max_results  all
        interval_shift  0       island
        interval_shift  2000    shore
        interval_shift  4000    shelves
```

Global parameters

GTF annotation

Gene information annotation

Annotation of overlaps with repeat regions

Annotation of overlaps with CpG islands

Pratique !!! Archivage + manipulation par des non bioinformaticiens

Analyse *in silico* appliquée au génome bovin

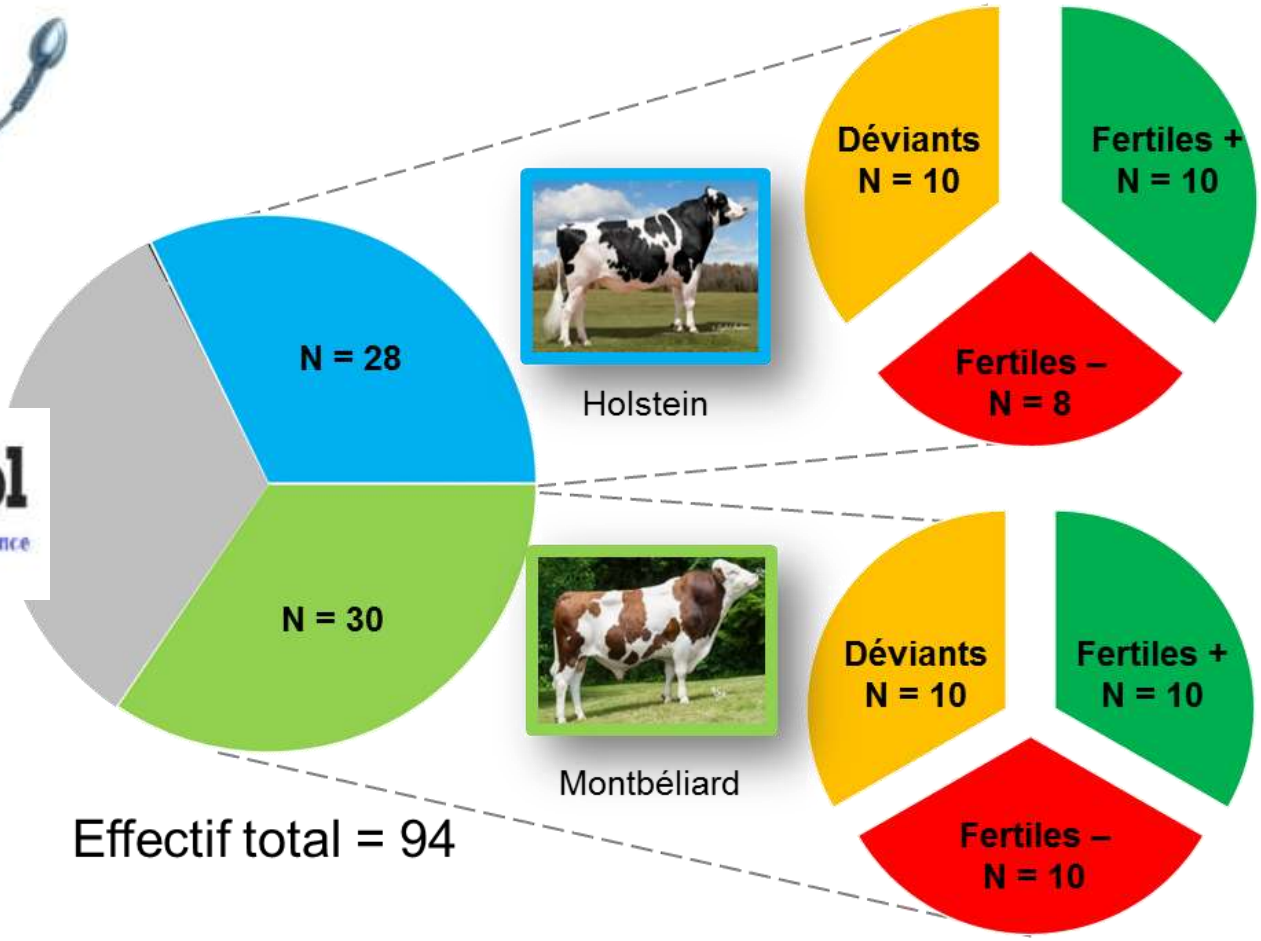
Size selection (bp)	No	0-250	40-290	80-330	120-370	160-410
RR genome size (Mb)	2 670	71	79	83	85	86
Per cent of whole genome	100.0	2.7	3.0	3.1	3.2	3.2
Number of MspI fragments	1 990 837	810 994	585 584	456 504	376 199	317 974
Number of CpG sites	27 540 276	3 588 657	3 454 028	3 127 005	2 854 907	2 666 727
Percent of total genomic CpG sites	100.0	13.0	12.5	11.4	10.4	9.7
Percent in 3'UTR	0.4	0.6	0.6	0.6	0.6	0.6
Percent in 5'UTR	0.2	0.5	0.4	0.3	0.2	0.2
Percent in exon	4.6	9.2	8.7	7.6	6.4	5.4
Percent in intron	31.1	30.7	32.0	33.0	33.8	34.5
Percent in intergenic	57.0	46.9	48.3	50.1	51.7	52.7
Percent in promoter-TSS	4.8	9.7	7.7	6.1	5.0	4.3
Percent in TTS	1.9	2.4	2.4	2.4	2.3	2.3
Percent in CpG islands	13.4	31.4	22.9	16.1	11.5	8.3
Percent in overlapping repeats	61.9	26.2	31.2	37.8	43.4	47.9

Fenêtre choisie pour la production de librairies : 40-290 pb
(0-250 : fragments trop petits)

Application : Recherche de biomarqueurs de la fertilité dans le méthylome de spermatozoïdes de taureaux


J-P. Perrier

SeQuaMol
Qualité Moléculaire de la semence

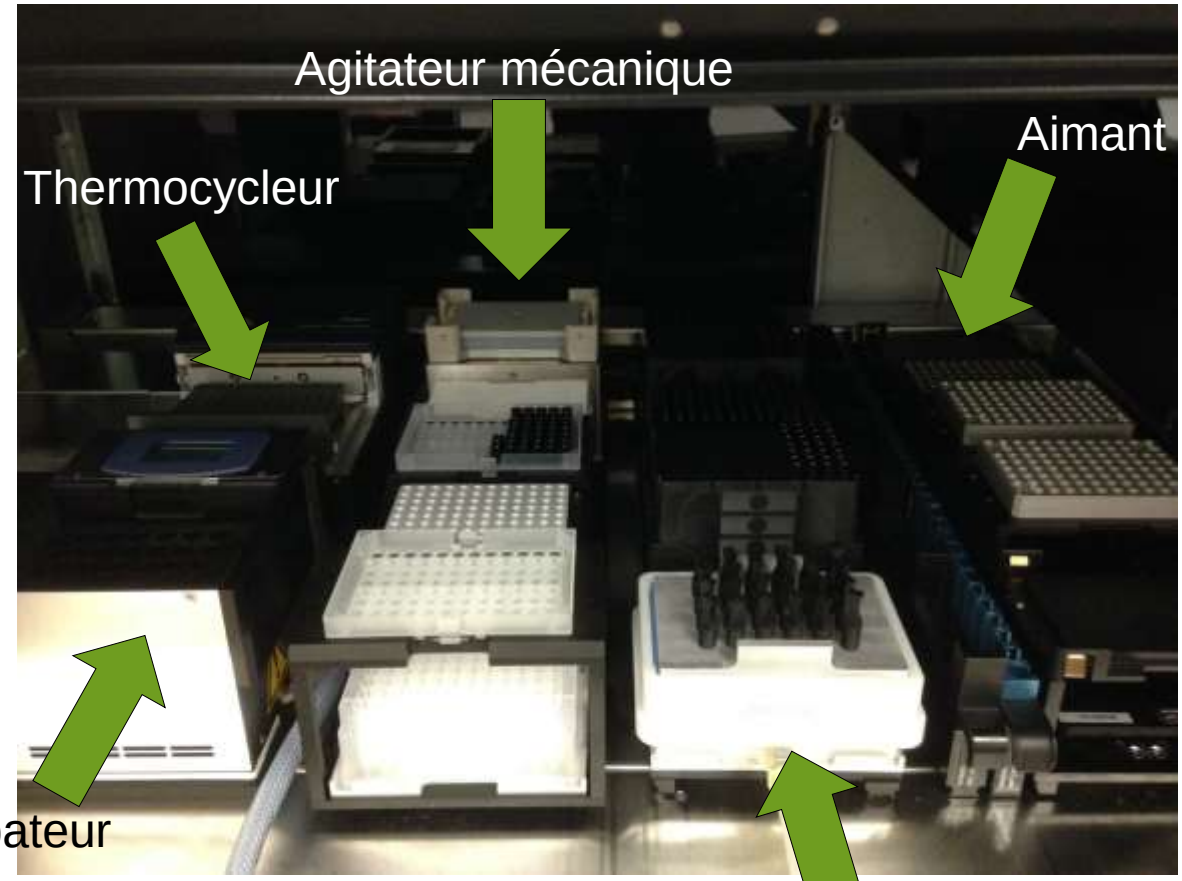


Automatisation de la construction de librairies RRBS


J-P. Perrier et A. Chaulot-Talmon

- Programme de >3000 lignes de commandes
- Production de 12 librairies simultanément
- Indépendance vis-à-vis des kits
- Contrôles qualité à chaque étape

From: http://www.hamiltoncompany.com/



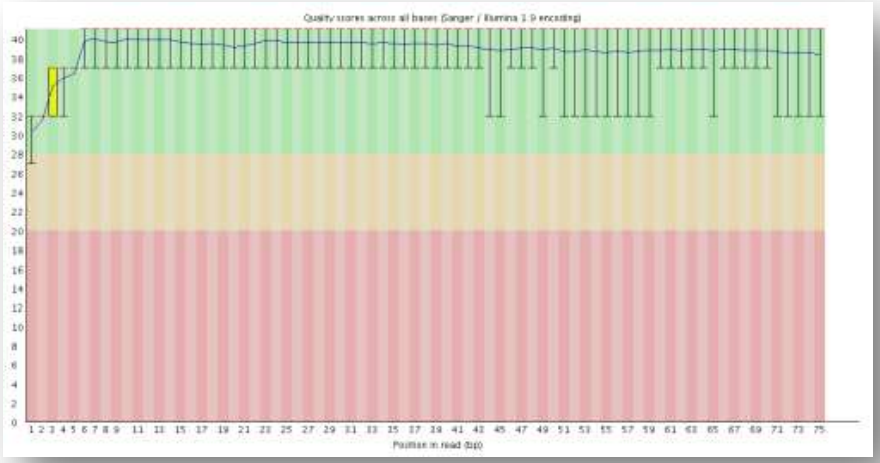
Résultat du séquençage

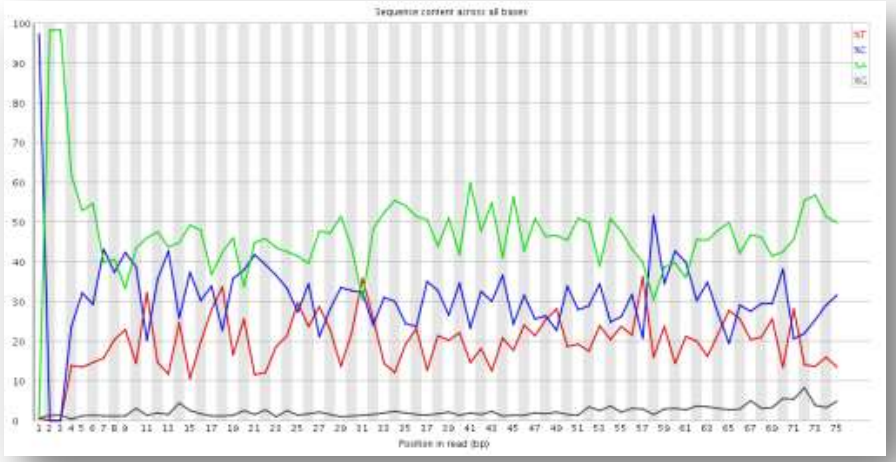
INTEGRAGEN GENOMICS


J-P. Perrier

- Paired-end 2x75 pb
- 35 millions de paires de séquences obtenus en moyenne par échantillon → 4 milliards de séquences totales (**~990 Go**)
- Contrôles qualité conformes à l'attendu :

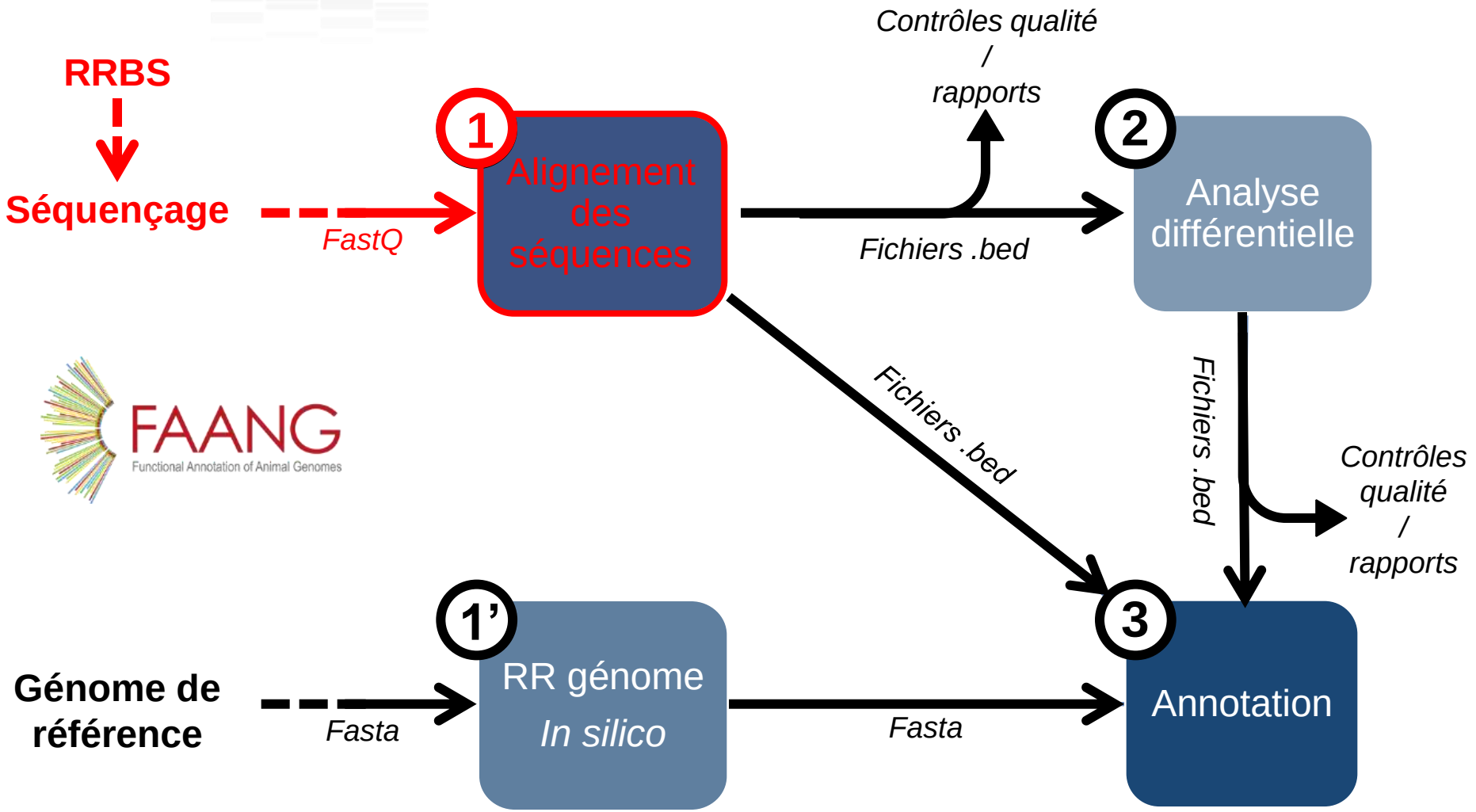
	Nombre de paires de séquences analysées	Taux de conversion bisulfite (%)
Moyenne	35 223 758	99,3
Minimum	26 370 072	98,0
Maximum	45 869 353	99,9



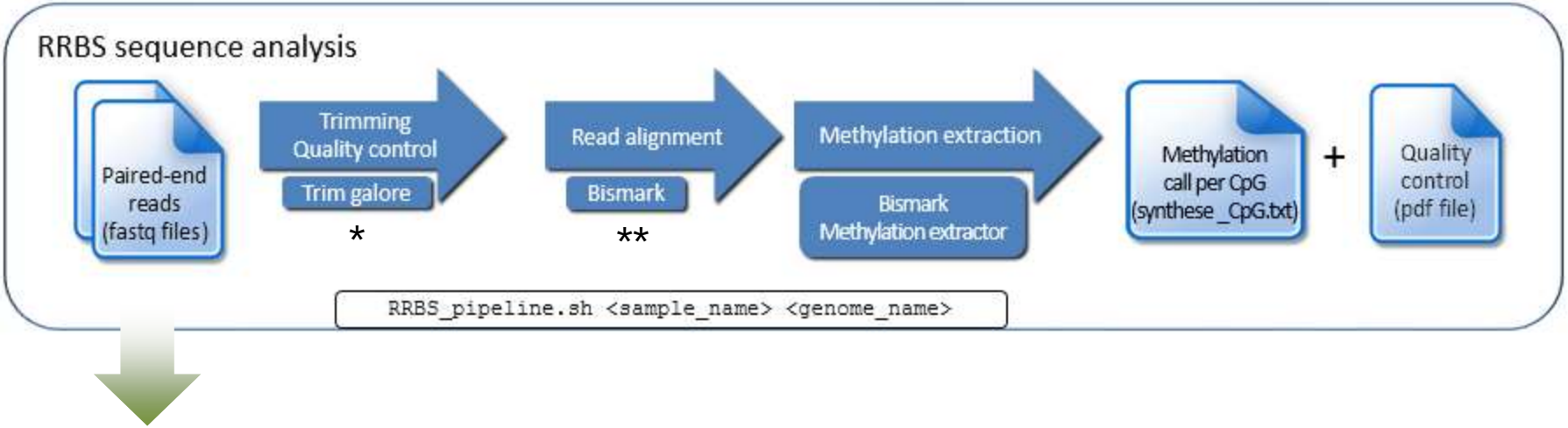


Vue d'ensemble du pipeline d'analyses bioinformatiques / biostatistiques


L. Jouneau et F. Piumi



Des séquences brutes à l'information de méthylation



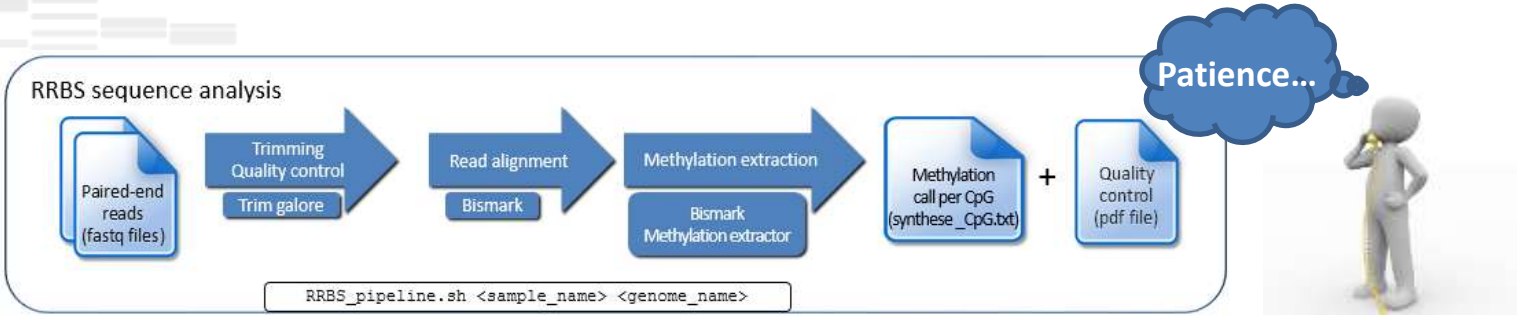
2 fichiers contenant ~35 millions de séquences (2 x ~8,5 Go) :

```
@K00103:24:H2MG5BBXX:2:1101:3862:1088 2:N:0:GATCGC
CAACTCAACTCTAAATTACCTTCCTCTATAACTTAACTCATAAAATCCCTACAAAAAACATAAAAACACACCAAG
+
<<,,A7<FKKKKKKFKKK<FFAKKKKKAFAFFKKFKKKKKKAFKF7KFKKKFFK,AFKKKKKFKKKKKK7FF<FK
```

* http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

** Krueger F. and Andrews S., 2011

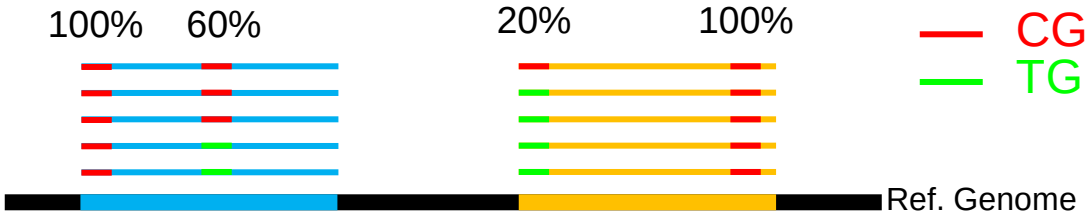
Des séquences brutes à l'information de méthylation



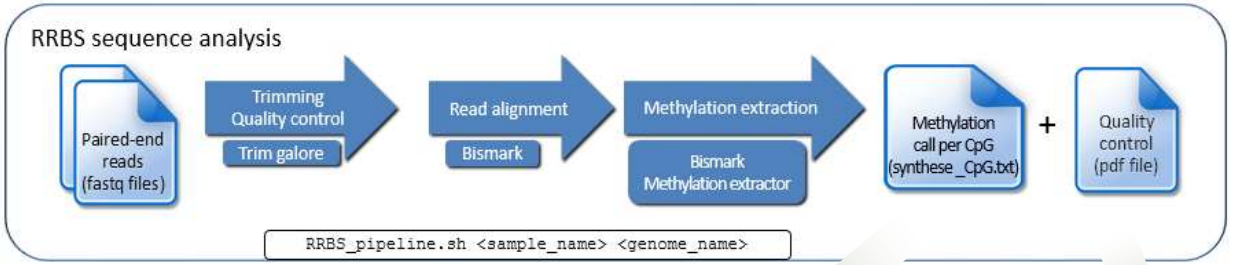
Trimming

Alignement on reference genome

Methylation call

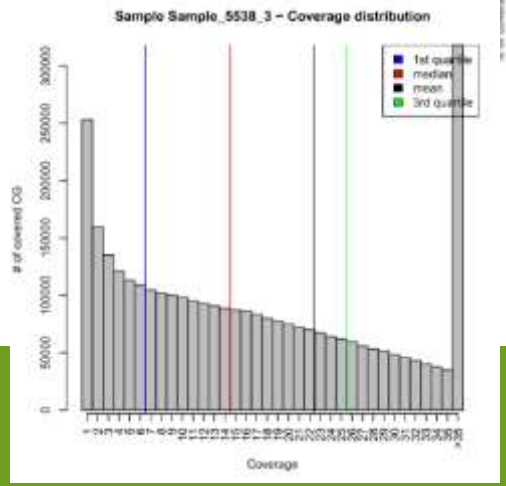


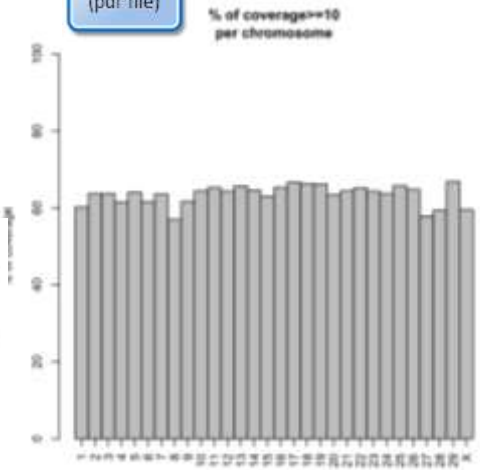
Des séquences brutes à l'information de méthylation



1 fichier .txt (~63 Mo)

Chromosome	Position	Coverage	# methylated	% methylated
1	20178	4	2	50
1	20192	4	2	50
1	20235	4	0	0
1	20250	4	2	50
1	20258	4	2	50
1	20353	10	2	20
1	20357	10	2	20
1	20377	10	2	20
1	20388	10	2	20

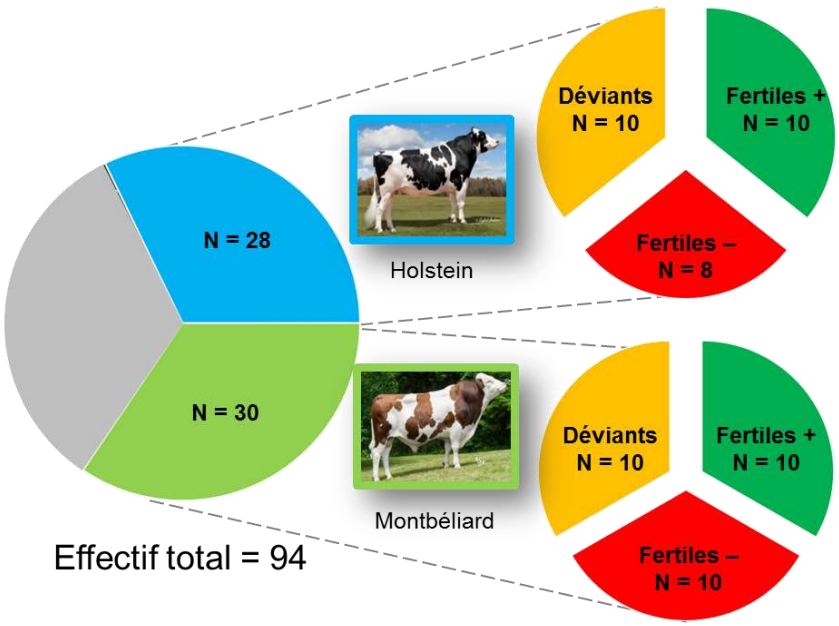




Statistiques après alignement


J-P. Perrier

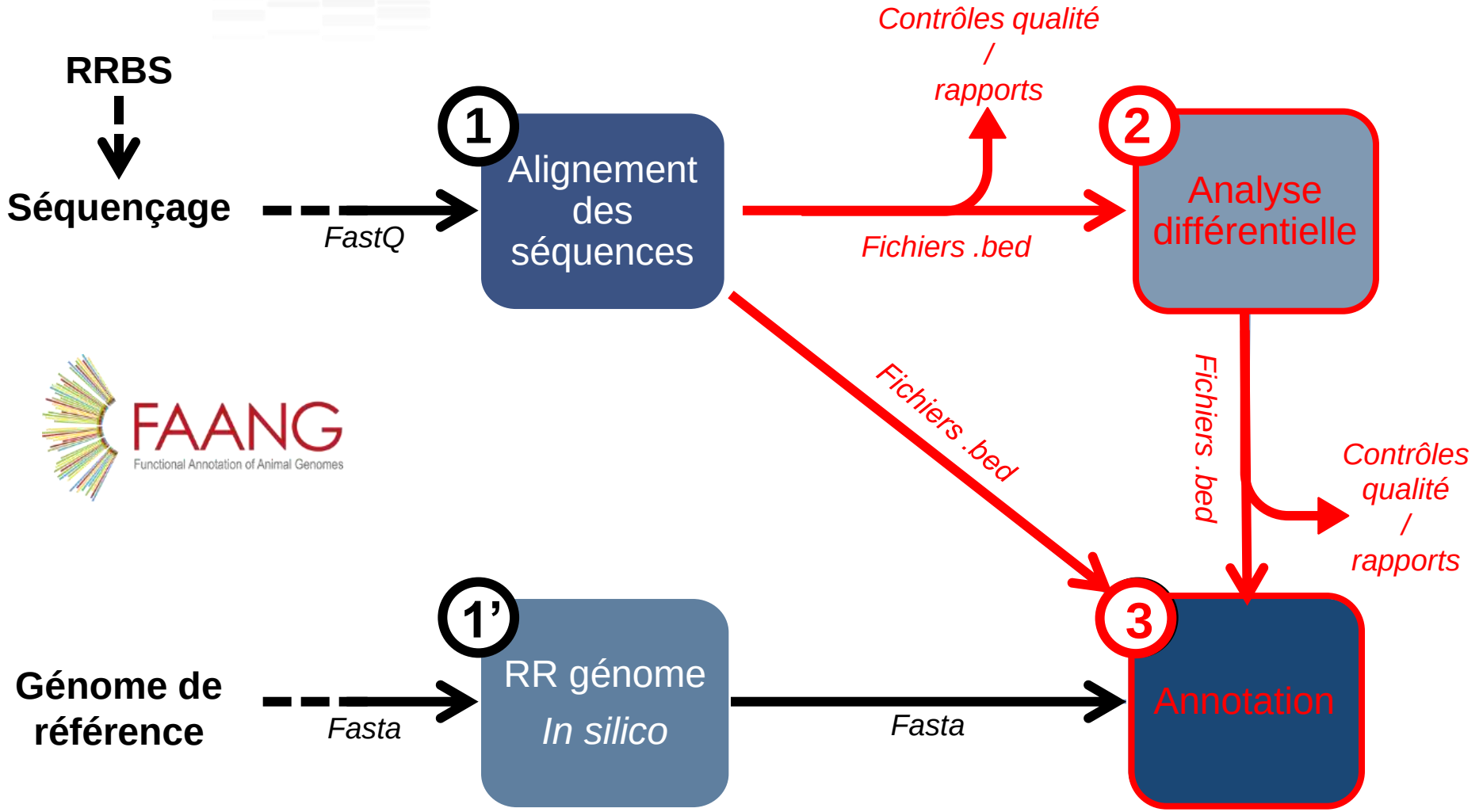
SeQuaMol
Qualité Moléculaire de la semence



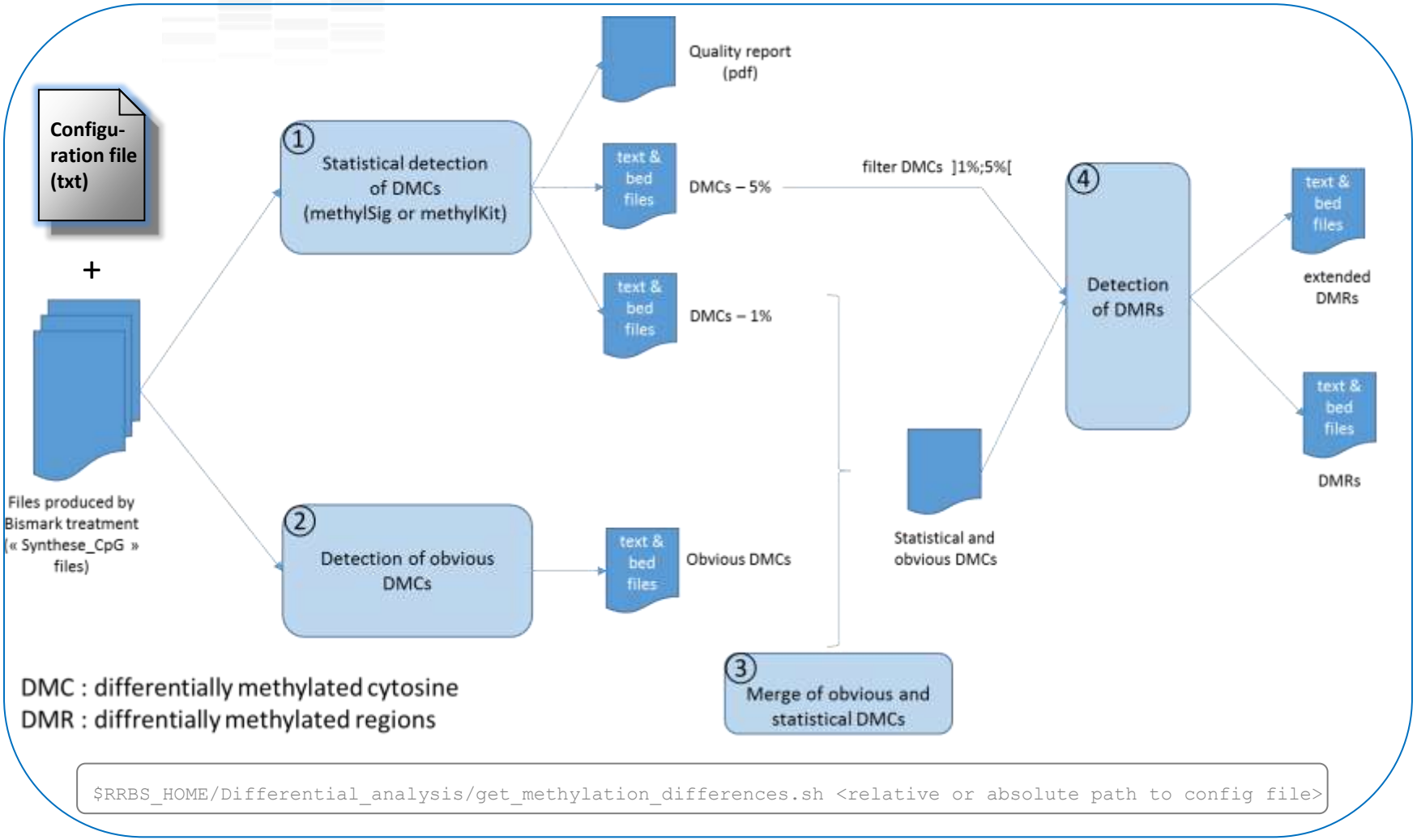
	Nombre de paires de séquences analysées	Taux de conversion bisulfite (%)	Alignements uniques (%)	Alignements multiples (%)	Pas d'alignement (%)	Nombre de CpG couverts	Nombre de CpG sur le chromosome Y	Nombre de CpG avec couverture ≥ 10 (%)
Moyenne	35 223 758	99,3	39,3	51,1	9,6	3 216 778	3896	61,0
Minimum	26 370 072	98,0	33,5	44,8	8,9	2 956 237	3010	54,8
Maximum	45 869 353	99,9	45,5	56,7	10,5	3 458 347	4738	66,4

Vue d'ensemble du pipeline d'analyses bioinformatiques / biostatistiques


L. Jouneau et F. Piumi



Identification de cytosines/régions différenciellement méthylées entre 2 conditions




```
$RRBS_HOME/Differential_analysis/get_methylation_differences.sh <relative or absolute path to config file>
```

Identification de cytosines/régions différenciellement méthylées entre 2 conditions

Configu-ration file (txt)

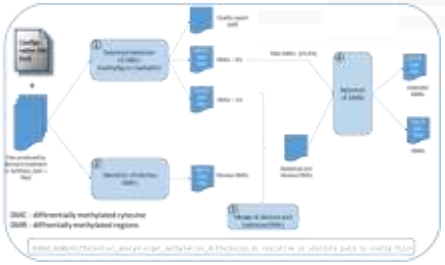
```
#Global parameters:
#----------------------
#title            HO_CMCP_MethylKit_qvalue001
#output_dir       /work/jperrier/ARTICLE_FERTILITE/3-Differential_analysis/MethylKit/HO_CMCP_qvalue001
#
#MethylKit parameters:
#------------------------- *
#stat_method      methylKit
#min_coverage1    10
#min_per_group    4
#stat_value       qvalue
#stat_threshold1 0.01
#methdiff_threshold1              0.25
#
#Obvious DMCs parameters:
#------------------------
#min_coverage2    10
#max_coverage2    1000000000
#methdiff_threshold2              0.8
#
#DMCs -> DMRs parameters:
#------------------------
#nb_min_DMCs_in_DMRs              3
#max_distance_between_DMCs        100
#stat_threshold2 0.05
#
Sample            File             Condition
17                /work/jperrier/ARTICLE_FERTILITE/synthese_CpG_sqmY/synthese_CpG_sqmY_17.txt          F_Neg
18                /work/jperrier/ARTICLE_FERTILITE/synthese_CpG_sqmY/synthese_CpG_sqmY_18.txt          F_Neg
20                /work/jperrier/ARTICLE_FERTILITE/synthese_CpG_sqmY/synthese_CpG_sqmY_20.txt          F_Neg
22                /work/jperrier/ARTICLE_FERTILITE/synthese_CpG_sqmY/synthese_CpG_sqmY_22.txt          F_Neg
28                /work/jperrier/ARTICLE_FERTILITE/synthese_CpG_sqmY/synthese_CpG_sqmY_28.txt          F_Neg
36                /work/jperrier/ARTICLE_FERTILITE/synthese_CpG_sqmY/synthese_CpG_sqmY_36.txt          F_Neg
66                /work/jperrier/ARTICLE_FERTILITE/synthese_CpG_sqmY/synthese_CpG_sqmY_66.txt          F_Neg
77                /work/jperrier/ARTICLE_FERTILITE/synthese_CpG_sqmY/synthese_CpG_sqmY_77.txt          F_Neg
12                /work/jperrier/ARTICLE_FERTILITE/synthese_CpG_sqmY/synthese_CpG_sqmY_12.txt          F_Pos
14                /work/jperrier/ARTICLE_FERTILITE/synthese_CpG_sqmY/synthese_CpG_sqmY_14.txt          F_Pos
16                /work/jperrier/ARTICLE_FERTILITE/synthese_CpG_sqmY/synthese_CpG_sqmY_16.txt          F_Pos
21                /work/jperrier/ARTICLE_FERTILITE/synthese_CpG_sqmY/synthese_CpG_sqmY_21.txt          F_Pos
54                /work/jperrier/ARTICLE_FERTILITE/synthese_CpG_sqmY/synthese_CpG_sqmY_54.txt          F_Pos
56                /work/jperrier/ARTICLE_FERTILITE/synthese_CpG_sqmY/synthese_CpG_sqmY_56.txt          F_Pos
62                /work/jperrier/ARTICLE_FERTILITE/synthese_CpG_sqmY/synthese_CpG_sqmY_62.txt          F_Pos
63                /work/jperrier/ARTICLE_FERTILITE/synthese_CpG_sqmY/synthese_CpG_sqmY_63.txt          F_Pos
76                /work/jperrier/ARTICLE_FERTILITE/synthese_CpG_sqmY/synthese_CpG_sqmY_76.txt          F_Pos
86                /work/jperrier/ARTICLE_FERTILITE/synthese_CpG_sqmY/synthese_CpG_sqmY_86.txt          F_Pos
```

Part dedicated to analysis parametrization

Definition of the two conditions and localization of analysis input files

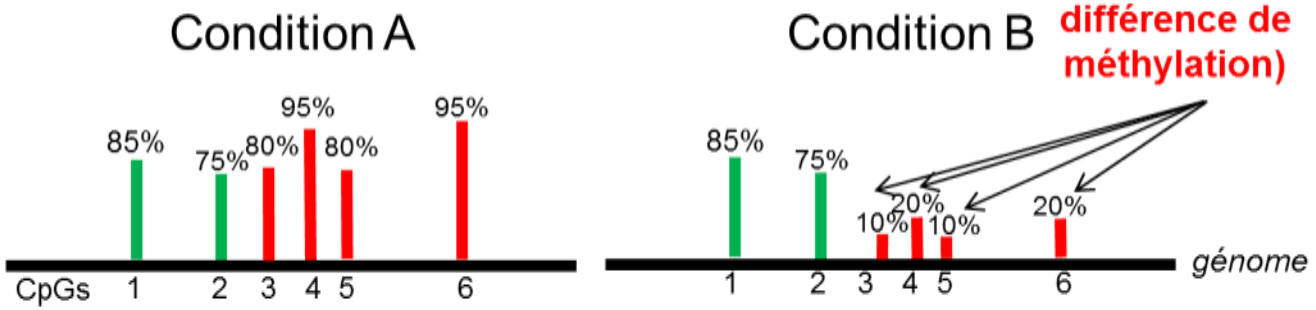
* Akalin *et al.*, 2012

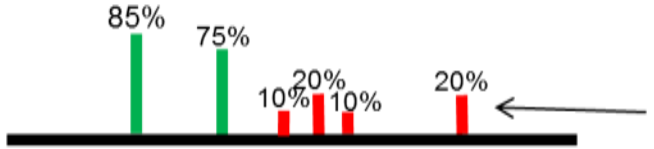
Identification de cytosines/régions différenciellement méthylées entre 2 conditions



Sélection des CpGs avec profondeur de séquençage appropriée (≥10 x)

DMCs
(minimum de 25% de différence de méthylation)





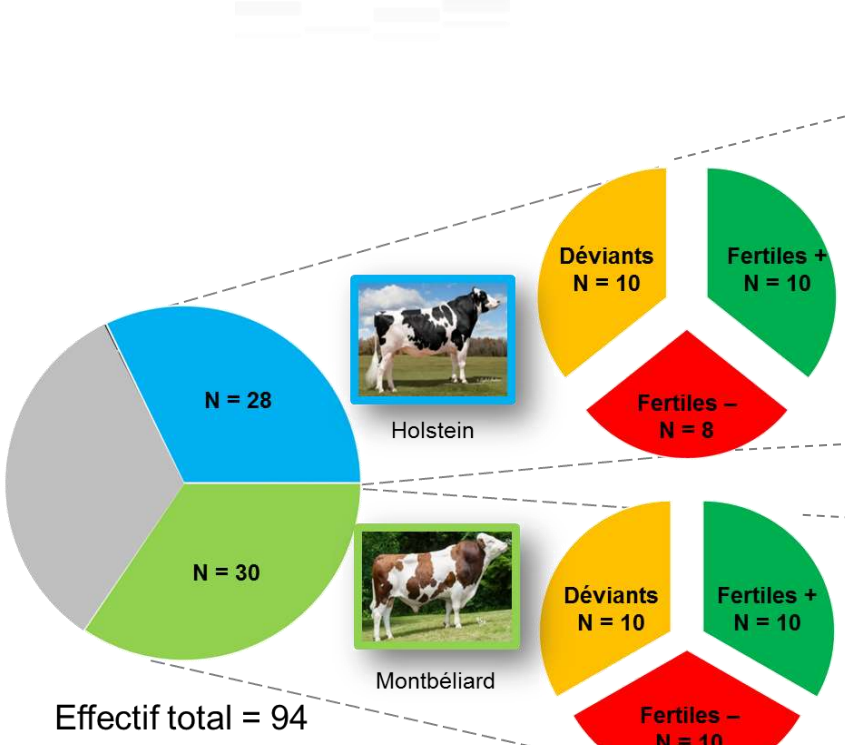
CpG trop loin de la DMR

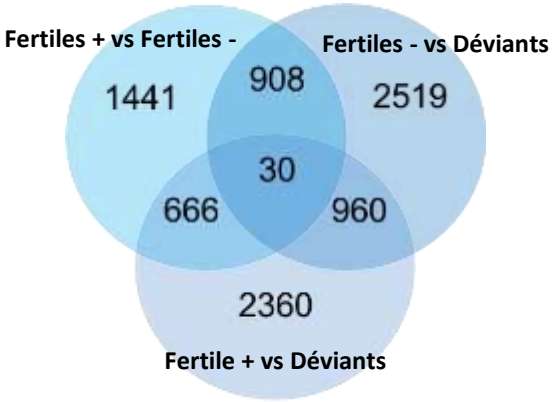
DMR

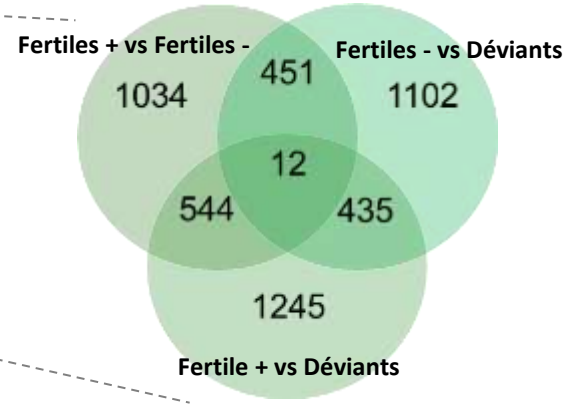
(un minimum de 3 DMCs avec moins de 100pb entre chaque)

Identification de cytosines différenciellement méthylées (DMC) liées à la fertilité


J-P. Perrier





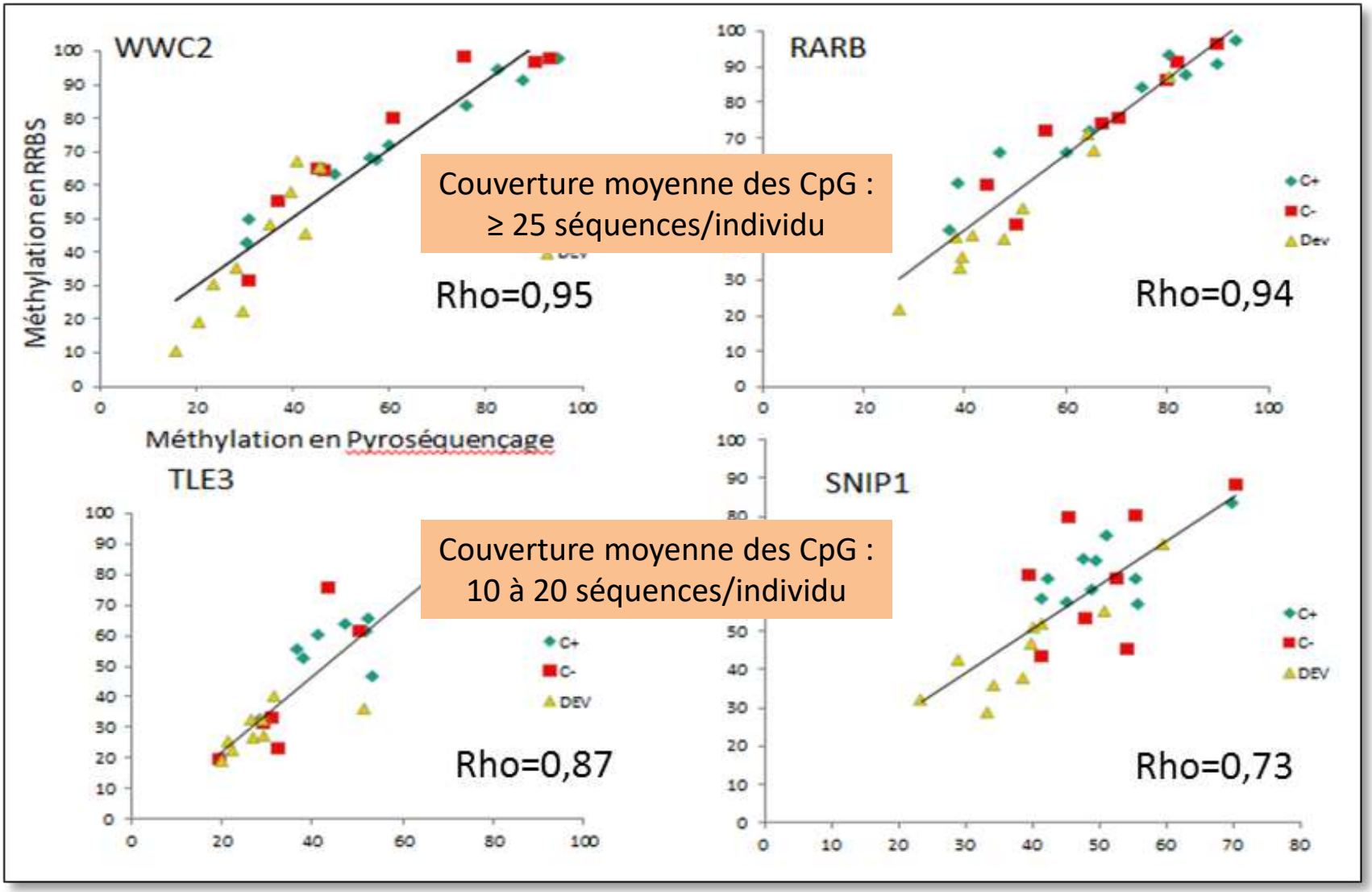


- methylKit
- qvalue 0,01
- Couverture minimale : 10 séquences/CpG pour ≥ 4 échantillons par catégorie
- Différentiel de méthylation >25%

Perrier et al., en préparation

Validation expérimentale des données de RRBS

C. Mehalla, M2



En résumé et en conclusion

- Peut s'analyser comme du polymorphisme de séquence, mais pas avec les mêmes outils bioinformatiques/biostatistiques (variable continue entre 0% et 100%)
- >20 millions de sites CpG dans les génomes de mammifères, dont beaucoup sont invariants en termes de méthylation (cellules somatiques)
- Il est possible de restreindre l'analyse à quelques millions de CpG stratégiques (technologie RRBS) ou à <1 million chez l'Homme (puce Illumina 450K), néanmoins le coût est élevé
- Une **plateforme bioinformatique** et un espace de stockage de l'ordre du To est nécessaire pour analyser plusieurs conditions et réplicats biologiques
- Entre 2 conditions proches (même type cellulaire), seuls quelques milliers de CpG sont différenciellement méthylés (entre types cellulaires : plusieurs centaines de milliers) → **importance de la construction d'un outil ciblé à mettre à disposition des filières**
- Les outils d'analyse et les paramètres choisis peuvent conditionner les résultats finaux → **importance de la prise en main des outils d'analyse par les biologistes et de la validation expérimentale !**

- **Jean-Philippe Perrier → en recherche d'emploi !**
- **Luc Jouneau**
- **François Piumi**
- **Maxime Gasselin**
- **Hala Al Adhami**
- **Aurélie Chaulot-Talmon**
- **Hélène Jammes**

- **Eli Sellem**
- **Chrystelle Le Danvic**
- **Sébastien Fritz**
- **Laurent Schibler**

Comparaison d'aligneurs pour l'analyse du RRBS

Séq. de réf. **C C G A T G A T G T C G C T G A C G C A A C G A**

Niveau de méthylation: 100% 50% 50% 0%

Fragmentation ADN
Conversion des C non méthylés en T
Séquençage

Reads bisulfités

T C G A
T C G A

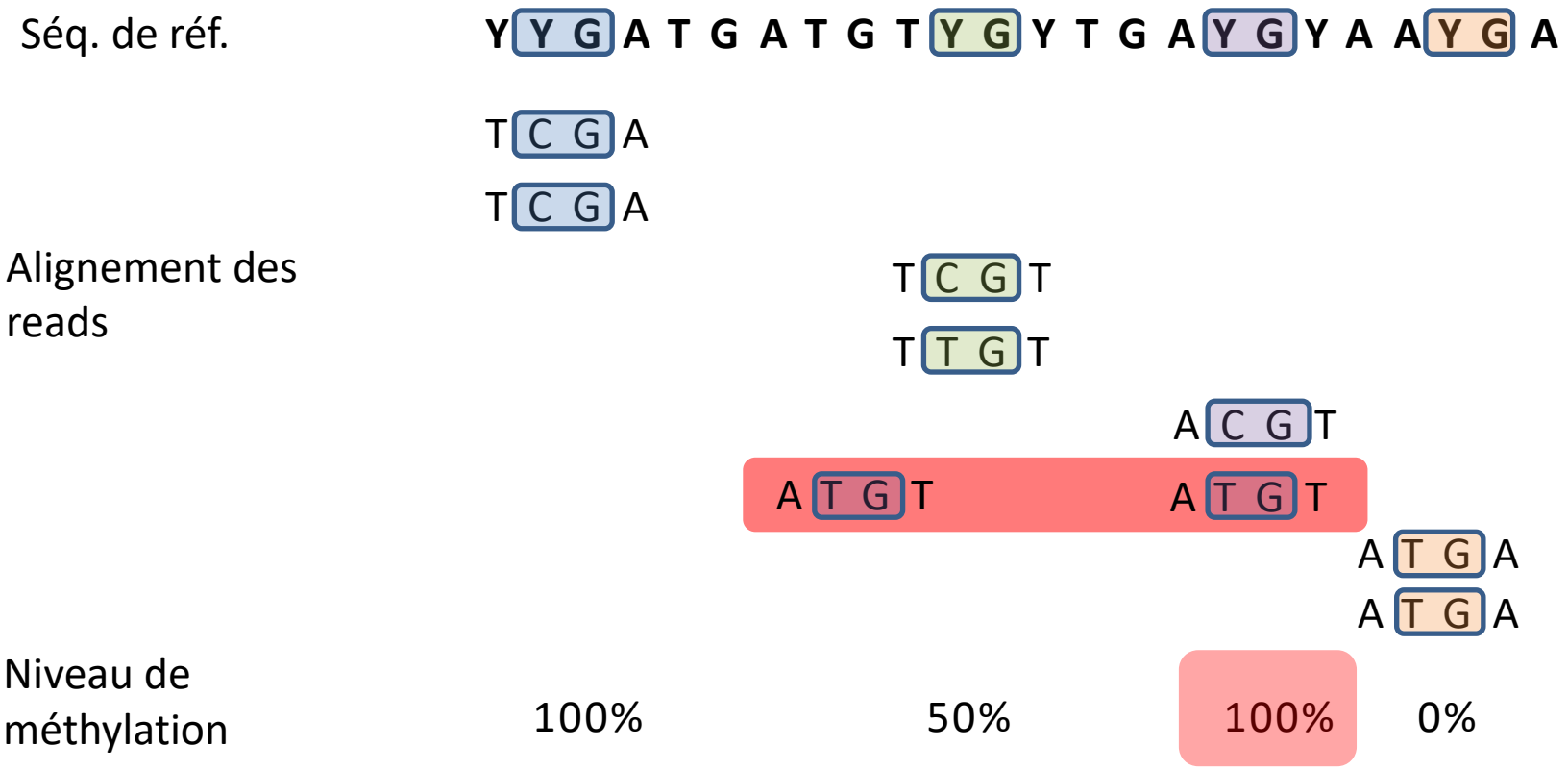
T C G T
T T G T

A C G T
A T G T

A T G A
A T G A

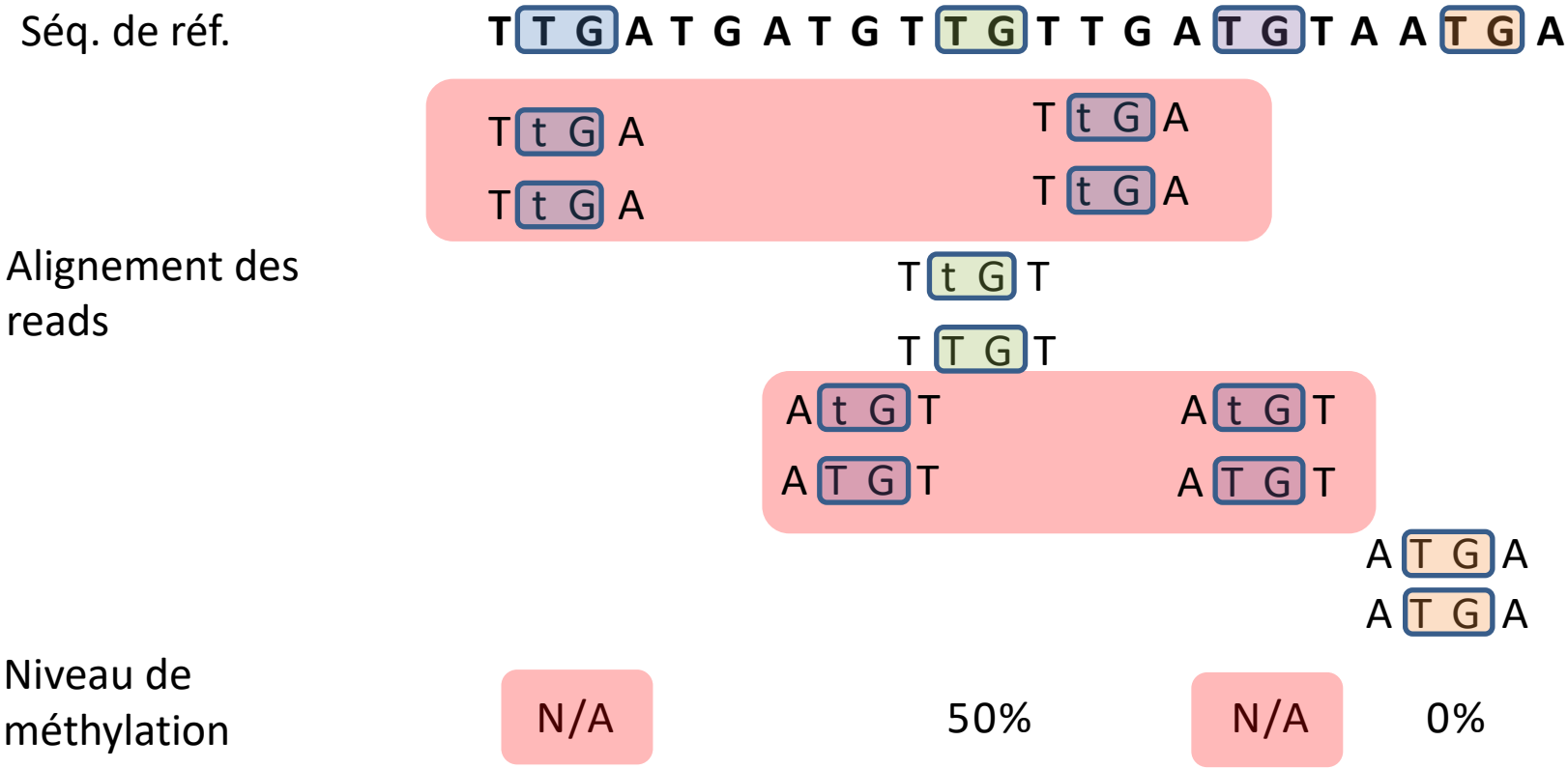
Bock C., Nature Reviews, oct 2012

Aligneur « wild-card » (exemple BSMAP)



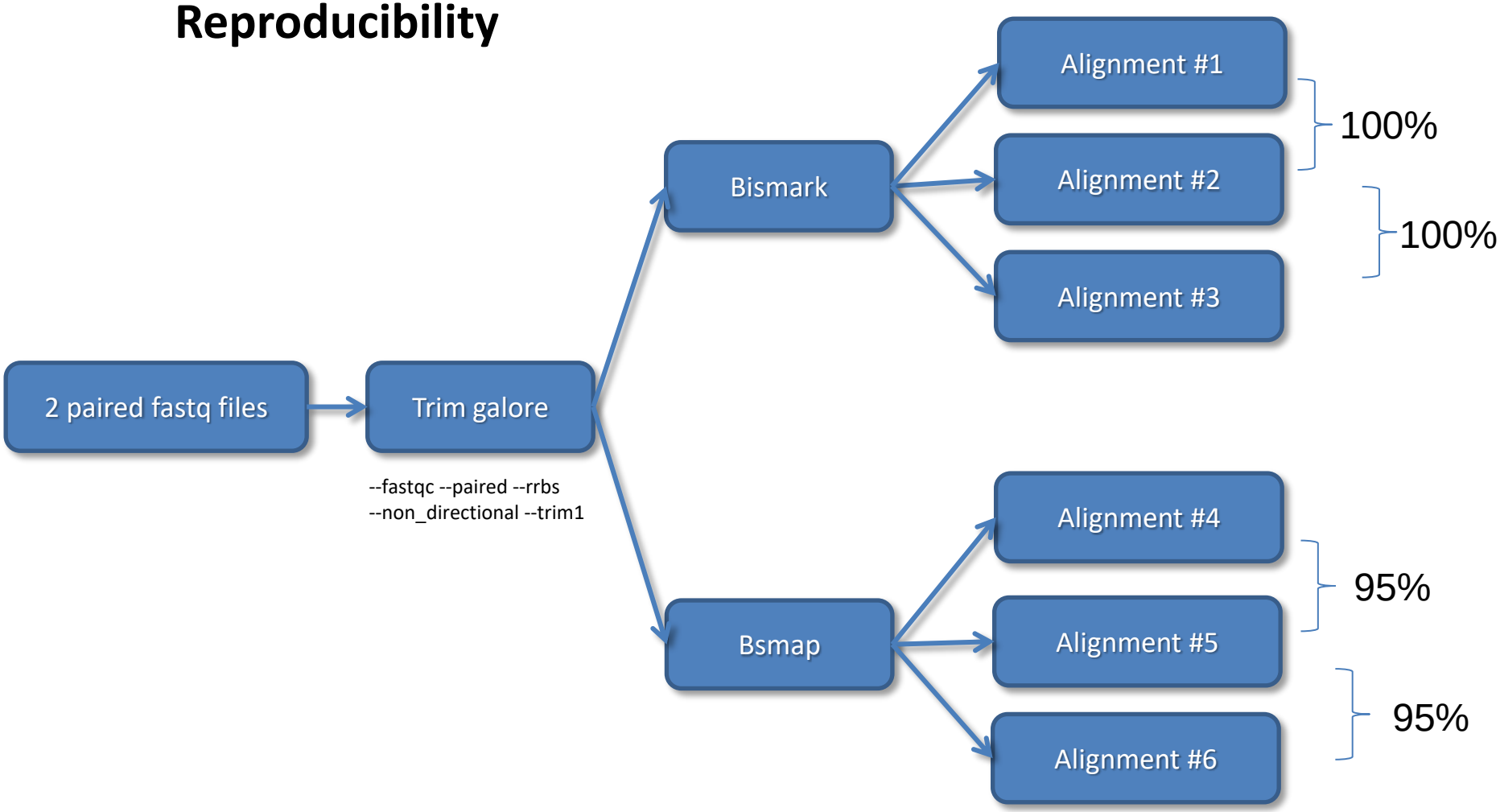
Bock C., Nature Reviews, oct 2012

Aligneur « 3 lettres » (exemple Bismark)



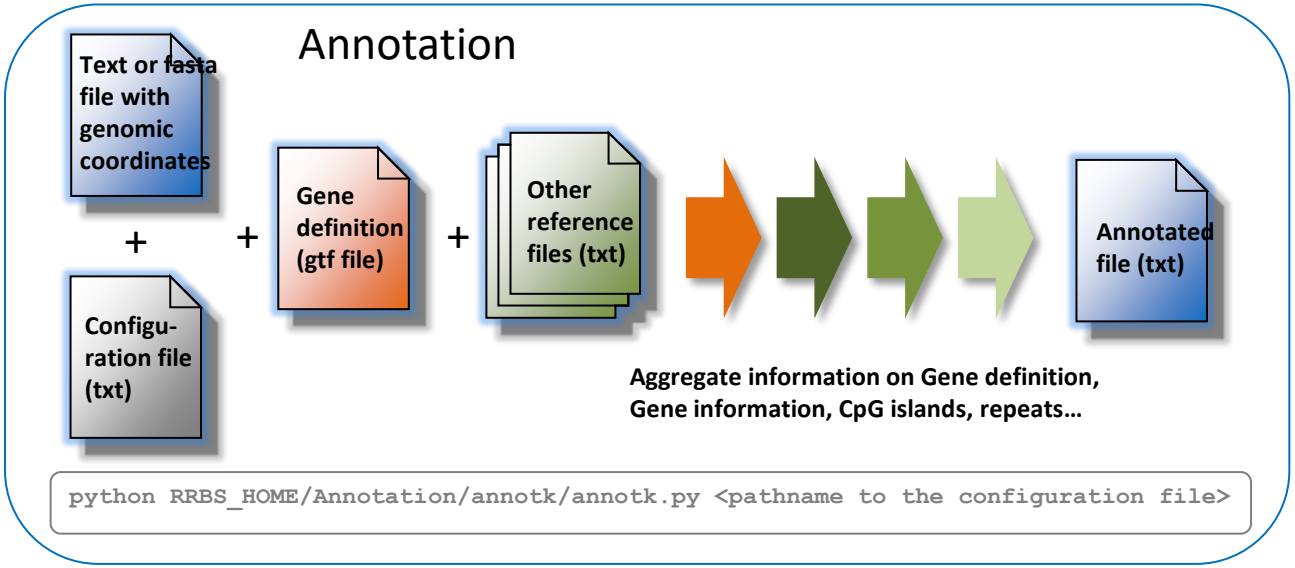
Bock C., Nature Reviews, oct 2012

Bismark - Bsmap comparison : Reproducibility



Bsmap doesn't find the same mapped reads between two identical runs

=> Bismark is reproducible, Bsmap is not.


Annotation
Text or fasta file with genomic coordinates
+
Configu-ration file (txt)
+
Gene definition (gtf file)
+
Other reference files (txt)
Annotated file (txt)
Aggregate information on Gene definition,
Gene information, CpG islands, repeats…
python RRBS_HOME/Annotation/annotk/annotk.py <pathname to the configuration file>