Initiation aux NGS

Galaxy

Christian Siatka

Directeur Général de L'Ecole de l'ADN

Professeur - Université de Nîmes

siatka@ecole-adn.fr

Informations fondamentales !

- This tool extracts reads in BAM format from the SRA at the NCBI. It is based on the sam-dump utility of the SRA Toolkit.

Sémantique ou jargon courant en Bioinformatique

Next Generation Sequencing - technology

École de l'ADN Mines

Abbreviations

Les génomes de référence sont généralement désignés par leurs abréviations, telles que:

hg19 = human genome, version 19
mm9 = Mus musculus genome, version 9
dm3 = Drosophila melanogaster, version 3
ce10 = Caenorhabditis elegans, version 10

Acronym	full phrase	Synonyms/Explanation
<ANYTHING>-seq	-sequencing	indicates that an experiment was completed by DNA sequencing using NGS
ChIP-seq	chromatin immunoprecipitation sequencing	NGS technique for detecting transcription factor binding sites and histone modifications (see entry *Input* for more information)
DNase	deoxyribonuclease I	DNase I digestion is used to determine active ("open") chromatin regions
HTS	high-throughput sequencing	next-generation sequencing, massive parallel short read sequencing, deep sequencing
MNase	micrococcal nuclease	MNase digestion is used to determine sites with nucleosomes
NGS	next-generation sequencing	high-throughput (DNA) sequencing, massive parallel short read sequencing, deep sequencing
RPGC	reads per genomic content	normalize reads to 1x sequencing depth, sequencing depth is defined as: (mapped reads x fragment length) / effective genome size
RPKM	reads per kilobase per million reads	normalize read numbers: RPKM (per bin) = reads per bin / (mapped reads (in millions) x bin length (kb))

Abbreviations

NOTE :base de données de séquences de référence (RefSeq) - Reference Sequence

- caccès libre, annotée et organisée de séquences nucléotidiques accessibles au public (ADN, ARN) et de leurs produits protéiques.
- construite par le Centre National d'Information en biotechnologie (NCBI)
- contrairement à GenBank, n'enregistre qu'un seul enregistrement pour chaque molécule biologique (c'est-à-dire l'ADN, l'ARN ou les protéines) des principaux organismes, allant des virus aux bactéries en passant par les eucaryotes.

Category	Description
NC	Complete genomic molecules
NG	Incomplete genomic region
NM	mRNA
NR	ncRNA
NP	Protein
XM	predicted mRNA model
XR	predicted ncRNA model
XP	predicted Protein model (eukaryotic sequences)
WP	predicted Protein model (prokaryotic sequences)

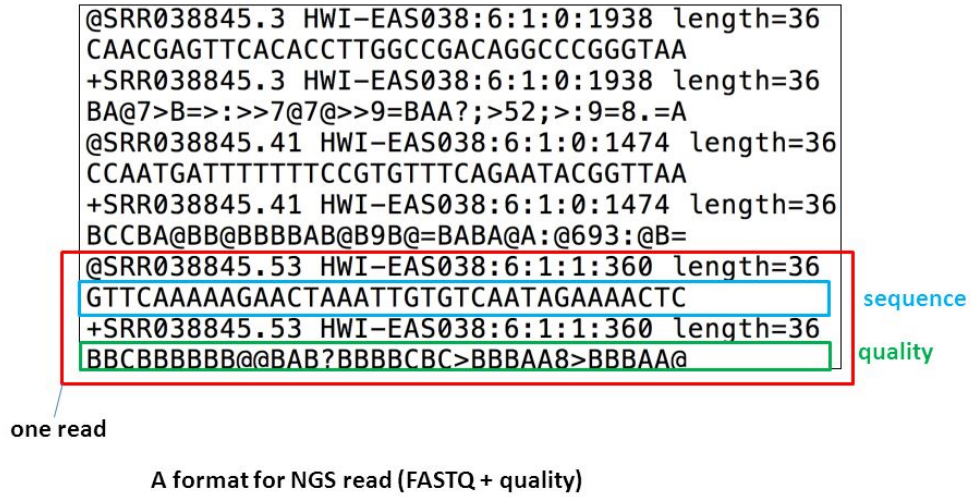
Un « read » (tag)

Une lecture, une séquence de données avec, les infos associées

Single read (SR)

le séquençage de la séquence d'ADN d'une seule extrémité de chaque fragment d'ADN.

FASTQ format (NGS raw data)

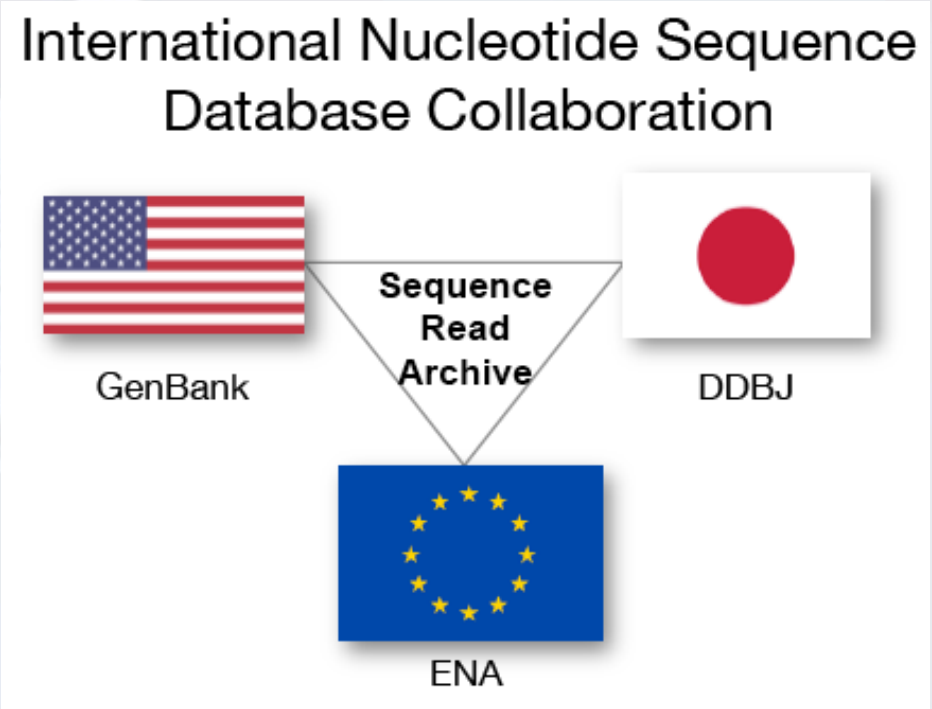


A format for NGS read (FASTQ + quality)

Paired-end read (PE)

le séquençage donne les deux extrémités de chaque fragment d'ADN. Les exécutions de PE sont plus chères (vous produisez deux fois plus de lectures d'ADN qu'avec SR), mais elles augmentent la cartographie pour les régions répétitives - et permettent une identification plus facile des variations structurelles et des indels.

SRA



The Sequence Read Archive (SRA) is the main repository for nucleic acid sequences and it has been growing tremendously in the past years.

There are three copies of the SRA which are maintained by the NCBI, the European Bioinformatics Institute, and the DNA Databank of Japan

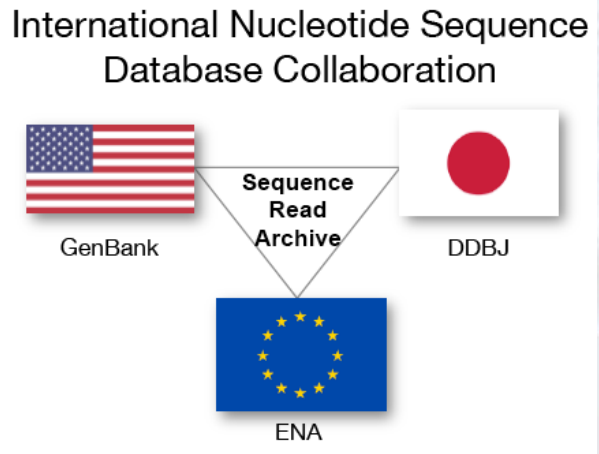
Trace Archive (TRACE) (SRA)

http://www.ncbi.nlm.nih.gov/Trace

Stocke toutes les séquences brutes Sanger (~300-800bp)

..

Taille (2017) : ~ 3,3 milliards de séquences (stable de puis 2010



FASTA

- Le format FASTA (ou format Pearson) est un format de fichier texte utilisé pour stocker des séquences biologiques de nature nucléique ou protéique.
- Ces séquences sont représentées par une suite de lettres codant pour des acides nucléiques ou des acides aminés selon la nomenclature IUPAC. Chaque séquence peut être précédée par un nom et des commentaires.
- Ce format est originellement issu de la suite de programmes FASTA mais, de par son utilisation très répandue, est devenu un standard de facto en bioinformatique
- .La simplicité du format FASTA rend la manipulation et la lecture (ou analyse syntaxique) des séquences aisée par l'utilisation d'outils de traitement de texte et de langages de script tels que Python, R, Ruby ou Perl.

FASTA

Voici un exemple de séquence nucléique :

```
>gi|373251181|ref|NG_001742.2| Mus musculus olfactory receptor GA_x5J8B7W2GLP-600-794
chromosome 2
AGCCTGCCAAGCAAACTTCACTGGAGTGTGCGTAGCATGCTAGTAACTGCATCTGAATCTTTCAGCTGCT
TGTTGGGCCTCTCACAAGGCAGAGTGTCTTCATGGGACTTTGATATTTATTTTTGTACAACCTAAGAGGA
ACAAATCCTTTGACACTGACAAATTGGCTTCCATATTTTATACCTTAATCATCTCCATGTTGAATTCATT
GATCAACAGTTTAAGAAAAAAAGATGTAAAAATGCTTTTAGAAAGAGAGGCAAAGTTATGCACAATAACT
TCTCATGAAGTCACAGTTTGTTAAAAGTTGCCTTAGTTCACAATAAATAATTATGTATGCTCTATAATTT
CAGTGA
```

Voici un exemple de séquence protéique :

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

FASTQ

- The FASTQ file format was derived from the simple text format for nucleic acid or protein sequences, FASTA.
- FASTQ bundles the sequence of every single read produced during a sequencing run together with the quality scores
- FASTQ files are uncompressed and quite large because they contain the following information for every single sequencing read:
- 1. @ followed by the read ID and possibly information about the sequencing run
- 2. sequenced bases
- 3. + (perhaps followed by the read ID again, or some other description)
- 4. quality scores for each base of the sequence (ASCII-encoded, see below)

```
$ zcat ERR459145.fastq.gz | head
@ERR459145.1 DHKW5DQ1:219:D0PT7ACXX:2:1101:1590:2149/1
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC
+
@7<DBADDDBH?DHHI@DH>HHHEGHIIIGGIFFGIBFAAGAFHA'5?B@D
@ERR459145.2 DHKW5DQ1:219:D0PT7ACXX:2:1101:2652:2237/1
GCAGCATCGGCCTTTTGCTTCTCTTTGAAGGCAATGTCTTCAGGATCTAAG
+
@@;BDDEFGHHHHIIIGBHHEHCCHGCGIGGHIGHGIGIIGHIIAHIIIGI
@ERR459145.3 DHKW5DQ1:219:D0PT7ACXX:2:1101:3245:2163/1
```

STAR

- (Dobin et al., 2013) Numerous alignment programs have been published in the past (and will be published in the future), and depending on your specific project, some aligners may be preferred over others. For example, detection of structural variants and fusion transcripts will require very specific settings or a dedicated alignment tool for that particular task.

- STAR (Spliced Transcripts Alignment to a Reference) is a fast NGS read aligner for RNA-seq data.
- STAR has more than 100 parameters
- handling of multi-mapped reads (e.g., how the best alignment score is assigned and the number and
- order in which secondary alignments are reported);
- • optimization for very small genomes;
- • defining the minimum and maximum intron sizes that are allowed (the default setting for the maximum
- intron size is 1,000,000 bp);
- • handling of genomes with more than 5,000 scaffolds (usually reference genomes in a draft stage);
- • using STAR for the detection of chimeric (fusion) and circular transcripts.

SAM/BAM format .sam ou .bam

- The output option of STAR already indicates that the results of the alignment will be stored in a SAM or BAM file.

- **The Sequence Alignment/Map (SAM)** format is, in fact, a generic nucleotide alignment format that describes the alignment of sequencing reads (or query sequences) to a reference.

- The human readable, TABdelimited SAM files can be compressed into the **Binary Alignment/Map** format.
- These BAM files are bigger than simply gzipped SAM files, because they have been optimized for fast random access rather than size reduction. Position-sorted BAM files can be indexed so that all reads aligning to a locus can be efficiently retrieved without loading the entire file into memory.

bedGraph .bed

- text file
- used for genomic intervals, e.g. genes, peak regions etc.
- the format can be found at UCSC
- for deepTools, the first 3 columns are important: chromosome, start position of the region, end position of the genome
- do not confuse it with the bedGraph format (although they are related)
- example lines from a BED file of mouse genes (note that the start position is 0-based, the end-position 1-based, following UCSC conventions for BED files):

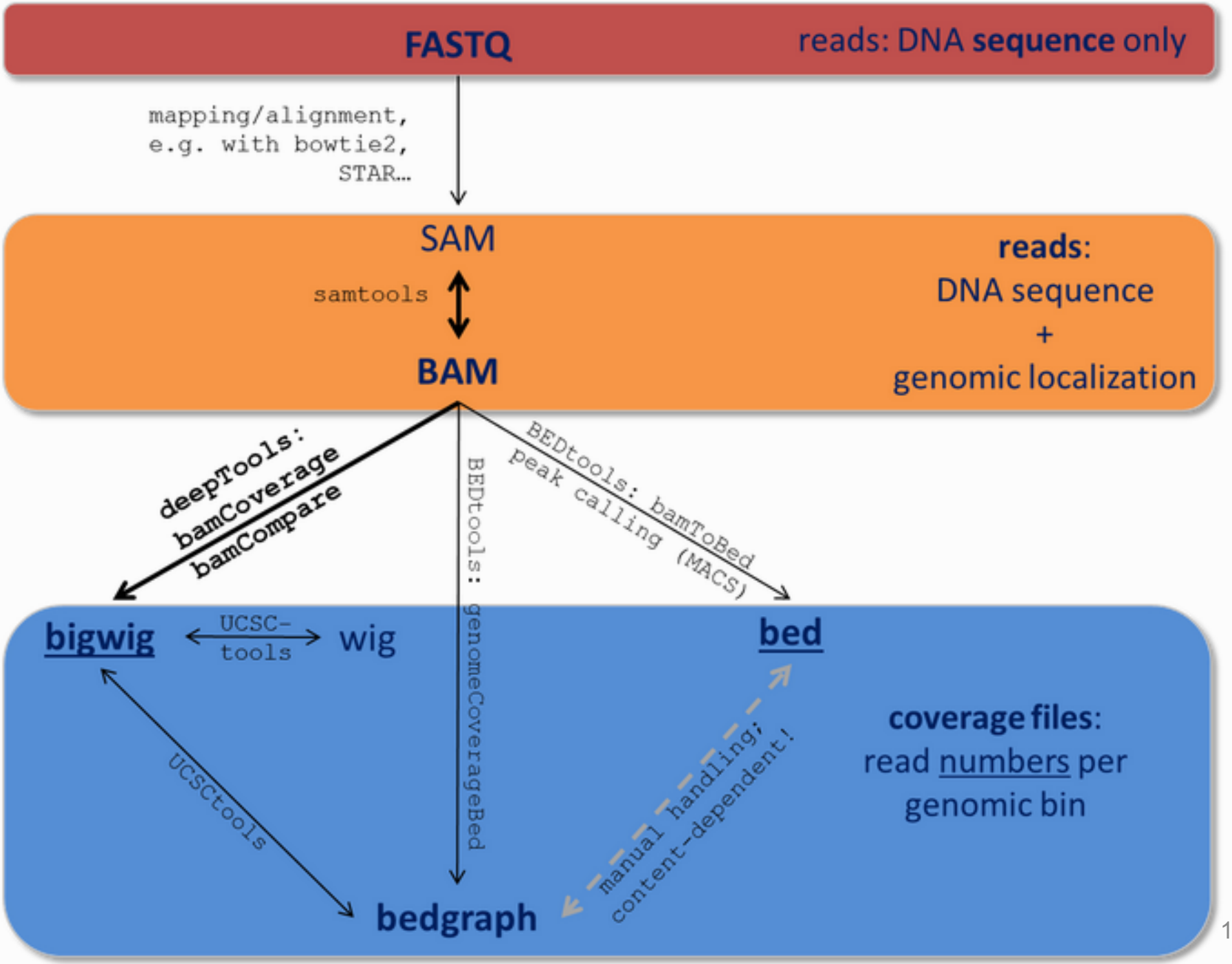
```
chr1    3204562 3661579 NM_001011874 Xkr4    -
chr1    4481008 4486494 NM_011441    Sox17   -
chr1    4763278 4775807 NM_001177658 Mrpl15  -
chr1    4797973 4836816 NM_008866    Lypla1  +
```

BED format .bg ou .bedgraph

- text file
- similar to BED file (not the same!), it can only contain 4 columns and the 4th column must be a score
- again, read the UCSC description for more details

bigWig .bw, .bigwig

- binary version of a bedGraph or wig file
- contains coordinates for an interval and an associated score
- the score can be anything, e.g. an average read coverage
- UCSC description for more details


FASTQ
reads: DNA sequence only
mapping/alignment,
e.g. with bowtie2,
STAR…
SAM
samtools
BAM
reads:
DNA sequence
+
genomic localization
deepTools:
bamCoverage
bamCompare
BEDtools: genomeCoverageBed
BEDtools: bamToBed
peak calling (MACS)
bigwig
UCSC-
tools
wig
bed
UCSCtools
manual handling;
content-dependent!
bedgraph
coverage files:
read numbers per
genomic bin

Phred Quality Score

The score is called Phred score, Q

$$q = -10 \log_{10}(p)$$

- p=error probability for the base
- if p=0.01 (1% chance of error), then q=20
- p = 0.00001, (99.999% accuracy), q = 50
- Phred quality values are rounded to the nearest integer

- proportional to the probability p that a base call is incorrect,
- Examples:
- Phred score of 10 corresponds to one error in every ten base calls (Q = -10 log10(0.1)), or 90% accuracy;
- Phred score of 20 corresponds to one error in every 100 base calls, or 99% accuracy.
- A higher Phred score thus reflects higher confidence in the reported base.

Phred Quality Score

The score is called Phred score, Q

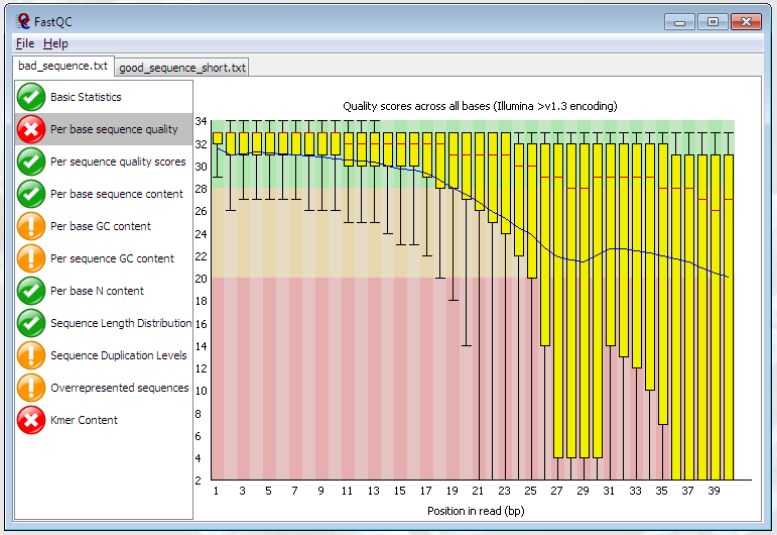
$$q = -10 \log_{10}(p)$$

Different Illumina (formerly Solexa) versions used different scores and ASCII offsets. Starting with Illumina format 1.8, the score now represents the standard Sanger/Phred format that is also used by other sequencing platforms and the sequencing archives

Description	**ASCII characters**		**Quality score**	
	Range	Offset	Type	Range
Solexa/early Illumina (1.0)	59 to 126 (; to ~)	64	Solexa	-5 to 62
Illumina 1.3+	64 to 126 (@ to ~)	64	Phred	0 to 62
Sanger standard/Illumina 1.8+	33 to 126 (! to ~)	33	Phred	0 to 93

Multi QC / fast QC

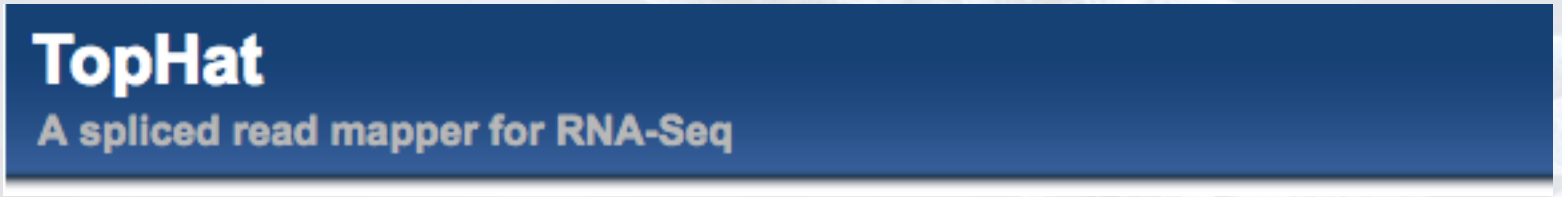
- A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.
- FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.



TopHat

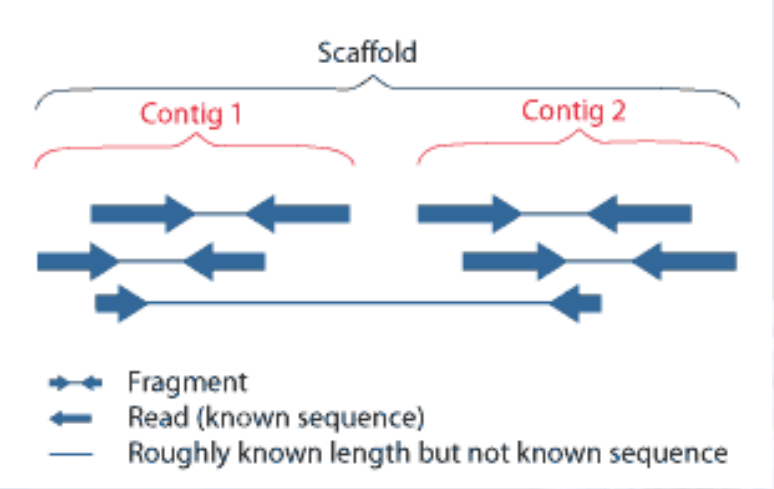
- Another popular aligner is TopHat, which is basically a sophisticated wrapper around the genomic aligner Bowtie (Kim et al., 2013).

- TopHat is a program that aligns RNA-Seq reads to a genome in order to identify exon-exon splice junctions. It is built on the ultrafast short read mapping program Bowtie. TopHat runs on Linux and OS X.



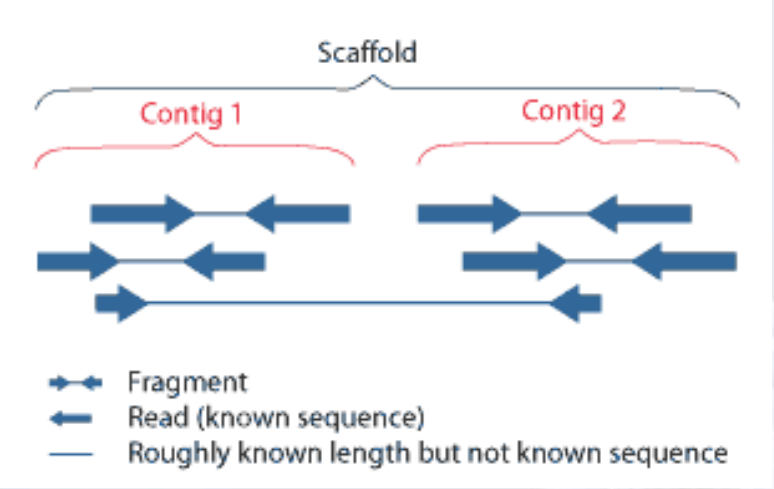
Contig

- Un contig est une séquence génomique continue et ordonnée générée par l'assemblage des clones d'une bibliothèque génomique (sous forme de plasmides, cosmides, BAC ou YAC) qui se chevauchent.

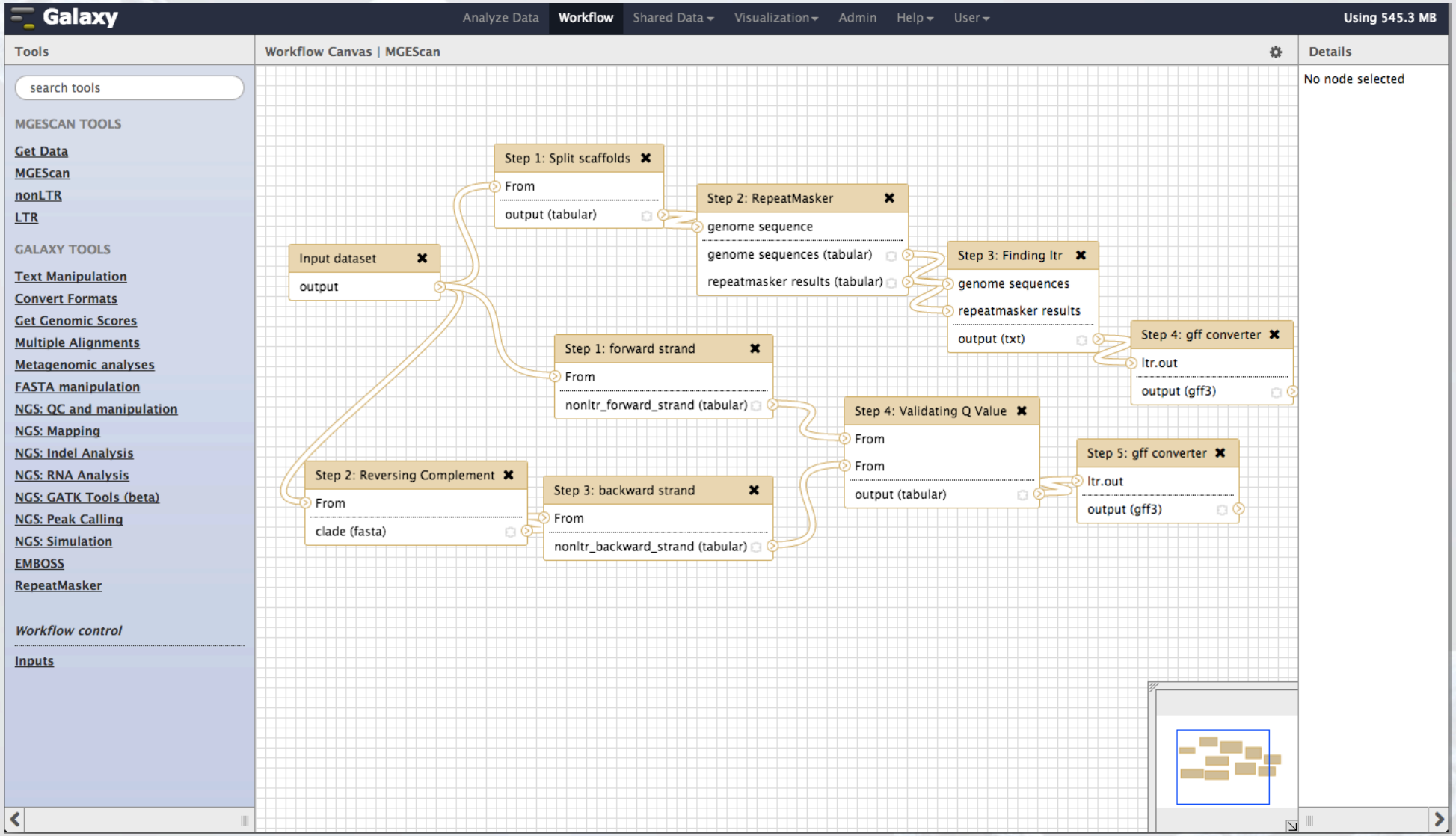


Scafold

- Relier ensemble une série non-contigüe de séquences génomiques dans un ensemble, constitué de séquences séparées par des intervalles de longueur connue.
- Les séquences qui sont liées sont typiquement des séquences contiguës correspondant à des chevauchements de lecture.



Pipeline


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 545.3 MB
Tools
search tools
MGESCAN TOOLS
Get Data
MGEScan
nonLTR
LTR
GALAXY TOOLS
Text Manipulation
Convert Formats
Get Genomic Scores
Multiple Alignments
Metagenomic analyses
FASTA manipulation
NGS: QC and manipulation
NGS: Mapping
NGS: Indel Analysis
NGS: RNA Analysis
NGS: GATK Tools (beta)
NGS: Peak Calling
NGS: Simulation
EMBOSS
RepeatMasker
Workflow control
Inputs
Workflow Canvas | MGEScan
Details
No node selected
Input dataset
output
Step 1: Split scaffolds
From
output (tabular)
Step 2: RepeatMasker
genome sequence
genome sequences (tabular)
repeatmasker results (tabular)
Step 3: Finding ltr
genome sequences
repeatmasker results
output (txt)
Step 4: gff converter
ltr.out
output (gff3)
Step 1: forward strand
From
nonltr_forward_strand (tabular)
Step 4: Validating Q Value
From
From
output (tabular)
Step 5: gff converter
ltr.out
output (gff3)
Step 2: Reversing Complement
From
clade (fasta)
Step 3: backward strand
From
nonltr_backward_strand (tabular)

Autre termes plus évident

- Base calling
- Mapping to a reference genome
- *De novo* or assisted genome assembly
- Single-molecule sequencing
- Pairing
- …