Université de Lille

File formats for NGS data

sylvain.legrand@univ-lille.fr

Plan du cours

1 – Sequence formats
2 – Data quality
3 – Alignment formats
4 – Annotations formats
5 – Graphical data visualization

1 – Sequence formats

Fastq format

- 4 lines per sequence

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

- Line 1 : ID
- Line 2 : sequence
- Line 3 : +, sometimes followed by the repetition of the sequence identifier
- Line 4 : quality
- Format used for reads

Fastq fastq, ID line

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*
Y	Y if the read is filtered (did not pass), N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

Source Wikipedia

Fastq format, quality

- Quality score= Phred Score
  - For each base, measure the probability that the assigned base is false
  - The score of a given base (Q) is given by the following equation: :

$$Q = -10log10(p)$$

p: Estimated probability that the given base is wrong
So a high score indicates a lower probability of error.

Q10: 1 error per 10 bases → 90% accuracy
Q20: 1 error per 100 bases → 99% accuracy
Q30: 1 error per 1000 bases → 99.9% accuracy

- The score is encoded in ASCII to lighten the file: Illimina = Phred+33

```
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                          |    |        |                              |                     |
33                          59   64       73                             104                   126
 0........................26...31.......40
```

Fasta format

- An ID line starting with « > » followed by the sequence with or without line breaks (example: every 80 bp)

```
>949344 pacid=16033748 polypeptide=917831 locus=949344 ID=949344.v1.107 annot-version=v1.0
ATGGCGACAAATAAATTTTCATCTTGTTTCTTATTTTCGTTAATGGTGTTTTTTCTCATCCTTTTGCCACTGATTTCAGG
GCAAATGATACCATGTCTACTGGGGAAATGTAAAAACACAAGGACATGCAATGCATCTTGCAAATCTAGAGGATACAAAG
GAGGGGCTTGTATAAGCATGGACGTTCGCTCAAAAACCGGTGCTTATTGCTGCAAAGTTAGATTTGAATAA
>945418 pacid=16033749 polypeptide=913905 locus=945418 ID=945418.v1.107 annot-version=v1.0
ATGCCACCAAATATCTACAGACTCTCCTTCTTCTTATCCCTACTCTGTTTCTTCTTCATTCCCTGTTTCTTCTCTCTCGA
CGAACAAGGTCAAGCTCTTTTGGCATGGAAGTCTCAACTGAACATCTCCGGCGACGCTTTTTCCTCCTGGCACGTCGCCG
ACACATCTCCCTGCAACTGGGTCGGCGTAAAATGTAACCGTAGAGGTGAAGTTTCGGAGATACAACTCAAAGGCATGGAC
TTGCAAGGTTCTCTGCCGGTGACTAGTCTCCGGAGCCTCAAGTCTCTTACTTCCCTCACTTTATCTTCACTCAATCTTAC
CGGAGTAATCCCCAAGGAGATAGGAGACTTTATTGAGCTTGAATTACTCGATTTATCGGATAATTCTCTCTCAGGCGATA
TCCCTGTGGAAATCTTCAGGCTCAAGAAACTCAAGACTCTGTCTTTGAACACTAACAATCTCGAAGGTCGGATTCCGATG
GAGATTGGGAATCTTTCGGGTCTCCTAGAGCTTATGCTTTTCGATAACAAGCTATCCGGAGAGATCCCGAGGAGTATCGG
```

- Used in particular for assembled sequences

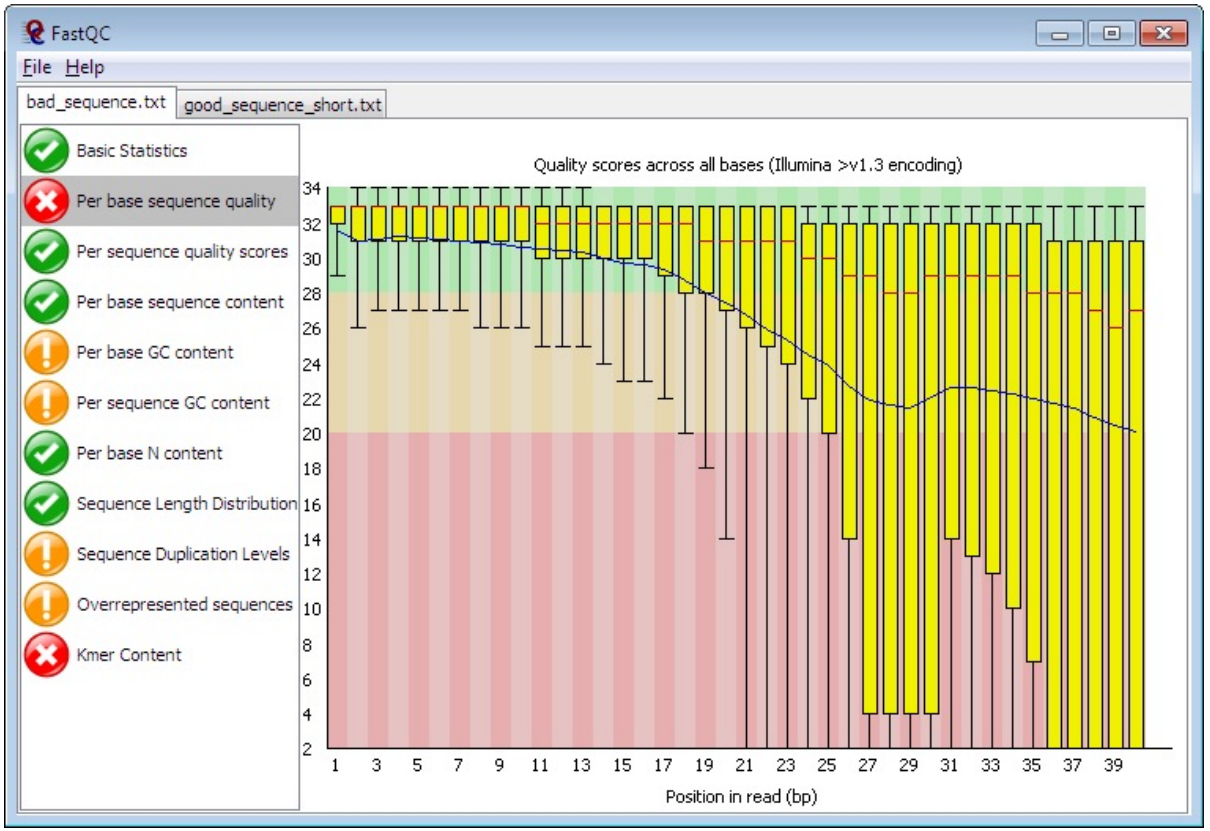
2 – Data quality

QC software example : fastqc

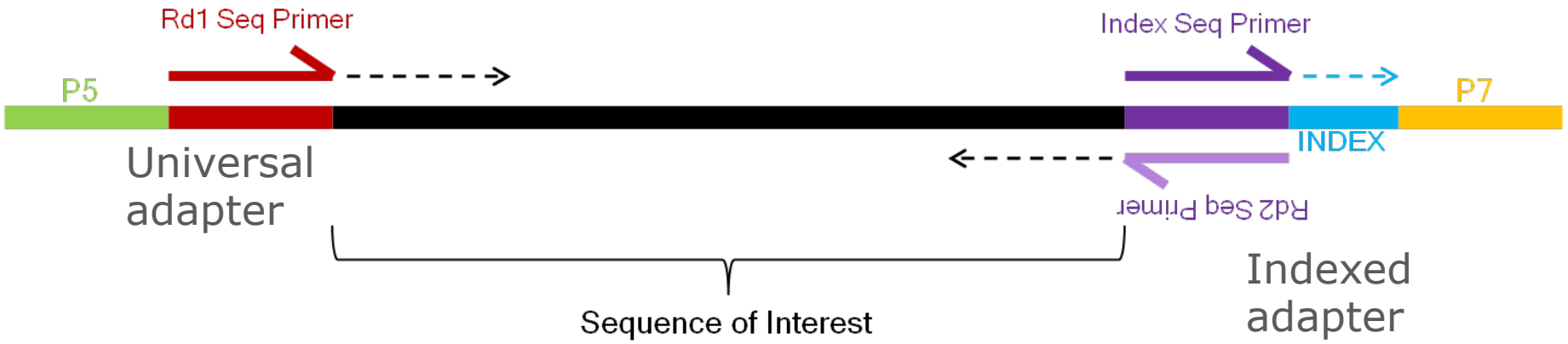
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Input format : fastq; output format: html report

Command line : `fastqc file_in.fastq`



Adapters trimming



TruSeq® v1/v2/LT Sample Prep Kits [2,5]

TruSeq® DNA (v1/v2/LT), TruSeq® DNA PCR-Free, TruSeq® Nano DNA, TruSeq® RNA (v1/v2/LT), TruSeq® Stranded RNA LT, TruSeq® RNA Access, and TruSeq® ChIP

TruSeq Universal Adapter
5′ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

TruSeq Adapter, Index 1 [5]
5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 2
5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 3
5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACTTAGGCATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 4
5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 5
5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 6
5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCCAATATCTCGTATGCCGTCTTCTGCTTG

http://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/experiment-design/illumina-customer-sequence-letter.pdf

Adapters trimming, software example : cutadapt

https://cutadapt.readthedocs.org/en/stable/

Input format: fastq, Output format: fastq

Commande line example: `cutadapt -b adaptateur1 -b adaptateur2 -O 10 file_in.fastq >file_out.fastq`

```
CP99_inf_1.fastq en cours...
cutadapt version 1.0
Command line parameters: -e 0.1 -O 10 -a TGGAATTCTCGGGTGCCAAGG CP99_inf_1.fastq
Maximum error rate: 10.00%
   Processed reads: 12201246
     Trimmed reads: 564676 (  4.6%)
   Too short reads: 0 (  0.0% of processed reads)
    Too long reads: 0 (  0.0% of processed reads)
        Total time:    489.68 s
     Time per read:      0.04 ms

=== Adapter 1 ===

Adapter 'TGGAATTCTCGGGTGCCAAGG', length 21, was trimmed 564676 times.

Histogram of adapter lengths
length   count
10       47388
11       39176
12       30270
13       22786
14       16764
15       11535
16       7717
17       5232
18       3710
19       2830
20       2352
21       374916
```

Cleaning of the sequences according to their quality, software example: prinseq

http://prinseq.sourceforge.net/

Input file: fastq; output file: 2 fastq files, one with the poor quality sequences and another with the cleaned good quality sequences

Commande line : `perl prinseq-lite.pl -fastq file_in.fastq -min_len 50 -min_qual_mean 25 -trim_qual_right 20 -ns_max_n 0 -noniupac`

```
Cutadapt_CP99_miR_1.fastq  en cours...
Input and filter stats:
        Input sequences: 7,781,005
        Input bases: 193,463,851
        Input mean length: 24.86
        Good sequences: 2,820,137 (36.24%)
        Good bases: 55,524,434
        Good mean length: 19.69
        Bad sequences: 4,960,868 (63.76%)
        Bad bases: 137,904,528
        Bad mean length: 27.80
        Sequences filtered by specified parameters:
        trim_qual_right: 1746
        min_len: 1026091
        max_len: 3921798
        min_qual_mean: 2526
        ns_max_n: 8707
```

3 – Alignment formats

SAM/BAM formats

- https://samtools.github.io/hts-specs/SAMv1.pdf
- SAM: Sequence Alignment/Map format
- BAM: compressed version of the file
- TAB-delimited text format consisting of a header section, which is optional, and an alignment section
- Header lines start with '@', while alignment lines do not

```
@HD     VN:1.0    SO:unsorted
@SQ     SN:Chr1   LN:30427671
@SQ     SN:Chr2   LN:19698289
@SQ     SN:Chr3   LN:23459830
@SQ     SN:Chr4   LN:18585056
@SQ     SN:Chr5   LN:26975502
@SQ     SN:chloroplast      LN:154478
@SQ     SN:mitochondria     LN:366924
@PG     ID:BowtieVN:0.12.8CL:"bowtie -v 0 --all --best --strata -S Athaliana_TAIR10.fa ./SRR9602/Cutadapt_SRR9602_prinseq_good_e9wD.fastq"
```

Header

```
SRR960237.9     16    Chr3          12098843  255  21M  *  0  0  AGGCCTCTACGAATTCATGAT  JIIJJJJHFHDHHFFFFFC@C  XA:i:0  MD:Z:21  NM:i:0
SRR960237.20    0     Chr2          8658085   255  23M  *  0  0  ACGGAATAATGTAAAACTGTACA  CCCFFFFFHHHHHJJJJJJIJJJ  XA:i:0  MD:Z:23  NM:i:0
SRR960237.28    16    chloroplast   136311    255  20M  *  0  0  GAATTCACCGCCGTATGGCT  JJJJJIGHHHHHFFFFFCCC  XA:i:0  MD:Z:20  NM:i:0
SRR960237.28    0     chloroplast   102319    255  20M  *  0  0  AGCCATACGGCGGTGAATTC  CCCFFFFFHHHHHGIJJJJJ  XA:i:0  MD:Z:20  NM:i:0
```

...

Alignment

Header in SAM/BAM

- Each header line begins with the character '@' followed by one of the two-letter header record type codes

- Examples:
  - @HD: VN: version of the file; SO: sorting order of alignments
  - @SQ: Reference sequence dictionary. The order of @SQ lines defines the alignment sorting order
  - @PG (Program): ID; VN: version; CL: command line
  - Others fields possible…

Alignment in SAM/BAM

- Each alignment line has 11 mandatory fields

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0, 2^{16} - 1]$	bitwise FLAG
3	RNAME	String	`\*|[:rname:^*=][:rname:]*`	Reference sequence NAME[11]
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	`\*|=|[:rname:^*=][:rname:]*`	Reference 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

SAM/BAM handling

- samtools – Utilities for the Sequence Alignment/Map (SAM) format
- http://www.htslib.org/doc/samtools.html
- Samtools is a set of utilities that manipulate alignments in the SAM and BAM formats.
- It converts between the formats, does sorting, merging and indexing, and can retrieve reads in any regions swiftly.

4 – Annotations format

BED format

- Browser Extensible Data

- Utilisé pour stocker des régions génomiques sous forme de coordonnées ainsi que les annotations associées

- The first three required BED fields are:
  - chrom: name of the chromosome (e.g. chr3, chrY, chr2_random) or scaffold (e.g. scaffold10671).
  - chromStart: starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered 0.
  - chromEnd: ending position of the feature in the chromosome or scaffold (exclusive)

- Note: The chromEnd base is not included in the display of the feature, however, the number in position format will be represented. For example, the first 100 bases of chromosome 1 are defined as chrom=1, chromStart=0, chromEnd=100, and span the bases numbered 0-99 in our software (not 0-100), but will represent the position notation chr1:1-100

BED format

- 9 additional optional BED fields can be used (name, score, strand…)

```
chr7    127471196  127472363  Pos1  0  +
chr7    127472363  127473530  Pos2  0  +
chr7    127473530  127474697  Pos3  0  +
chr7    127474697  127475864  Pos4  0  +
chr7    127475864  127477031  Neg1  0  -
chr7    127477031  127478198  Neg2  0  -
chr7    127478198  127479365  Neg3  0  -
chr7    127479365  127480532  Pos5  0  +
chr7    127480532  127481699  Neg4  0  -
```

GFF format

- GFF (General Feature Format)

- GFF lines have nine required fields that must be tab-separated
  - seqname: The name of the sequence (chromosome or scaffold...)
  - source: The program that generated this feature.
  - feature: The name of this type of feature (ex: "CDS", "start_codon", …)
  - start: The starting position of the feature in the sequence. The first base is numbered 1
  - end: The ending position of the feature (inclusive).
  - score: A score between 0 and 1000
  - strand: Valid entries include "+", "-", or "."
  - frame: If the feature is a coding exon, frame should be a number between 0-2 that represents the reading frame of the first base.
  - group: All lines with the same group are linked together into a single item.

- If a field is empty, enter "."

GFF format

```
##gff-version 3
##annot-version v1.0
##species Arabidopsis lyrata
scaffold_1	phytozomev11	gene	29681	31614	.	+	.	ID=311229.v1;Name=311229
scaffold_1	phytozomev11	mRNA	29681	31614	.	+	.	ID=311229.v1.107;Name=311229;pacid=16057706;longest=1;Parent=311229.v1
scaffold_1	phytozomev11	exon	29681	29746	.	+	.	ID=311229.v1.107.exon.1;Parent=311229.v1.107;pacid=16057706
scaffold_1	phytozomev11	CDS	29681	29746	.	+	0	ID=311229.v1.107.CDS.1;Parent=311229.v1.107;pacid=16057706
scaffold_1	phytozomev11	exon	29864	29938	.	+	.	ID=311229.v1.107.exon.2;Parent=311229.v1.107;pacid=16057706
scaffold_1	phytozomev11	CDS	29864	29938	.	+	0	ID=311229.v1.107.CDS.2;Parent=311229.v1.107;pacid=16057706
scaffold_1	phytozomev11	exon	30016	30117	.	+	.	ID=311229.v1.107.exon.3;Parent=311229.v1.107;pacid=16057706
scaffold_1	phytozomev11	CDS	30016	30117	.	+	0	ID=311229.v1.107.CDS.3;Parent=311229.v1.107;pacid=16057706
scaffold_1	phytozomev11	exon	30232	30274	.	+	.	ID=311229.v1.107.exon.4;Parent=311229.v1.107;pacid=16057706
scaffold_1	phytozomev11	CDS	30232	30274	.	+	0	ID=311229.v1.107.CDS.4;Parent=311229.v1.107;pacid=16057706
scaffold_1	phytozomev11	exon	30444	30448	.	+	.	ID=311229.v1.107.exon.5;Parent=311229.v1.107;pacid=16057706
scaffold_1	phytozomev11	CDS	30444	30448	.	+	0	ID=311229.v1.107.CDS.5;Parent=311229.v1.107;pacid=16057706
scaffold_1	phytozomev11	exon	30605	30685	.	+	.	ID=311229.v1.107.exon.6;Parent=311229.v1.107;pacid=16057706
scaffold_1	phytozomev11	CDS	30605	30685	.	+	0	ID=311229.v1.107.CDS.6;Parent=311229.v1.107;pacid=16057706
```

Handling of BED and GFF files

- Bedtools: A swiss-army knife of tools that allow a fast and flexible way of comparing large datasets of genomic features

- https://bedtools.readthedocs.io/en/latest/index.html

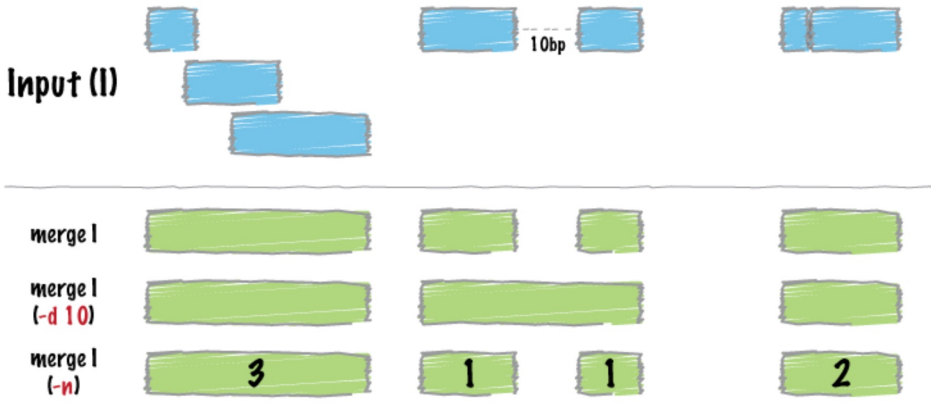
- Bedtools allows one to intersect, merge, count, complement, and shuffle genomic intervals from multiple files in widely-used genomic file formats such as BAM, BED, GFF/GTF, VCF

- The full list of bedtools sub-commands: annotate, bamtobed, bamtofastq, bed12tobed6, bedpetobam, bedtobam, closest, cluster, complement, coverage, expand, flank, fisher, genomecov, getfasta, groupby, igv, intersect, jaccard, links, makewindows, map, maskfasta, merge, multicov, multiinter, nuc, overlap, pairtobed, pairtopair, random, reldist, shift, shuffle, slop, sort, subtract, tag, unionbedg, window

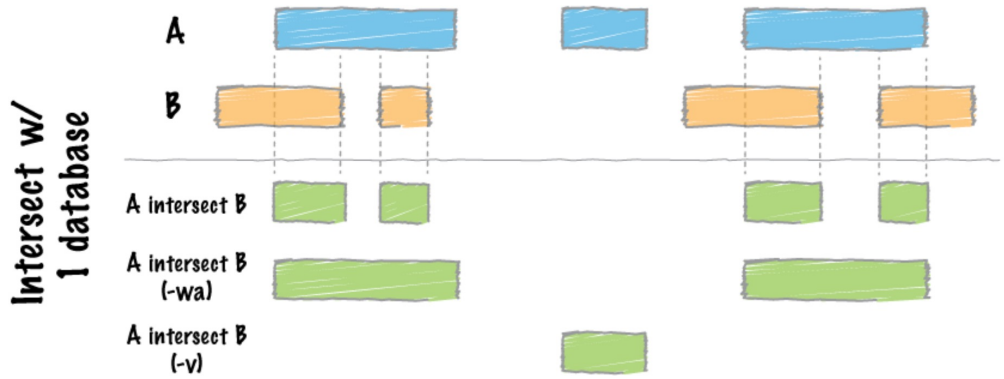
Handling of BED and GFF files

- Examples:

merge

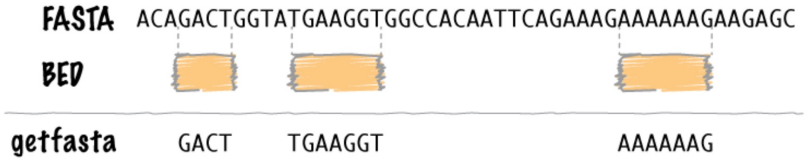


intersect



shuffle

getfasta



5 – Graphical data visualization – example: IGV

IGV

- The Integrative Genomics Viewer (IGV) is a high-performance, easy-to-use, interactive tool for the visual exploration of genomic data
- https://software.broadinstitute.org/software/igv/
- Formats

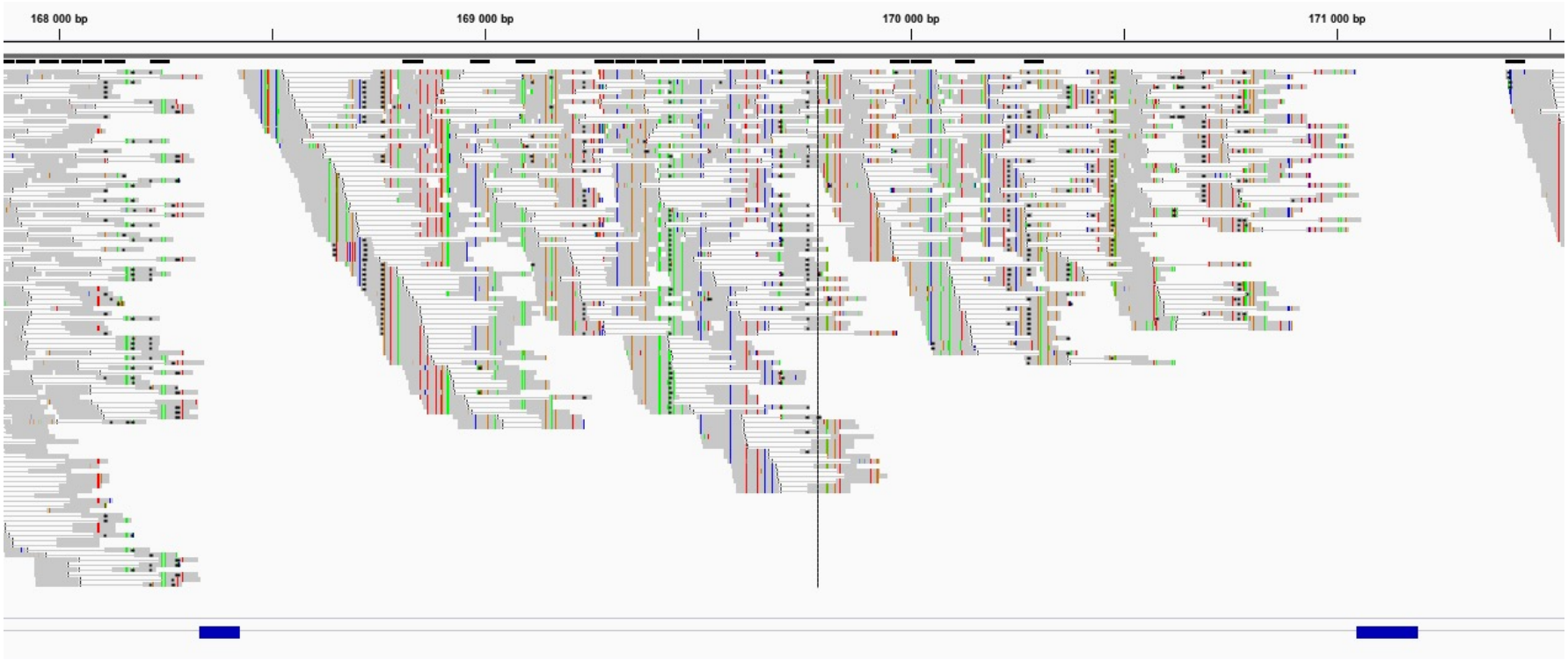
- BAM
- BED
- BEDPE
- BedGraph
- bigBed
- bigWig
- Birdsuite Files
- broadPeak
- CBS
- Chemical Reactivity Probing Profiles
- chrom.sizes
- CN
- Custom File Formats
- Cytoband
- FASTA
- GCT
- CRAM
- genePred
- GFF/GTF
- GISTIC
- Goby
- GWAS
- IGV
- LOH
- MAF (Multiple Alignment Format)
- MAF (Mutation Annotation Format)
- Merged BAM File
- MUT
- narrowPeak
- PSL
- RES
- RNA Secondary Structure Formats
- SAM
- Sample Info (Attributes) file
- SEG
- TDF
- Track Line
- Type Line
- VCF
- WIG

IGV

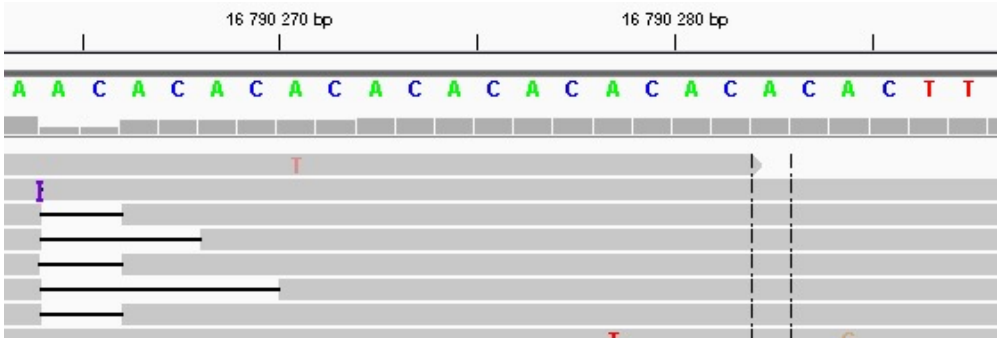
Fasta

BAM

GFF



Zoom →



Sylvain Legrand
Maître de Conférences
UMR CNRS 8198 EVO-ECO-PALEO
Evolution, Ecologie et Paléontologie
Université de Lille - Faculté des Sciences et Technologies
Bât SN2, bureau 208 - 59655 Villeneuve d'Ascq

sylvain.legrand@univ-lille.fr | http://eep.univ-lille.fr/
Tél. +33 (0)3 20 43 40 16