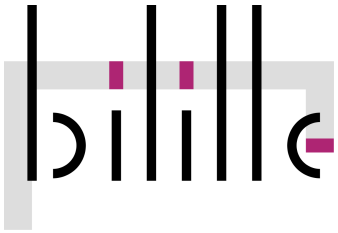


Short read mapping

3rd Oct. 2023

Pierre Péricard

pierre.pericard@univ-lille.fr

Bilille Platform - PLBS

Hélène Touzet

CNRS, Bonsai, CRIStAL

DNA-seq

raw reads

trimmed and cleaned reads

- reference genome → **mapping**
  find location of the reads and align them with respect to the reference
- no reference genome → **assembly**
  reconstruct the initial sequence

Short read mapping

```
ACAACTGTCTGCTTCAGGAGTTAAATCTTACA-GGATGA     reference

ACAACTGTCTGCTTCAGGAGT                       read1  +
 CAACTGTCTG-TTCAGGAGTT                      read2  +
 CAACTGTCTGCTTCAGGAGTT                      read3  -
    TGTCTGCTTCGGGAGTTAAATCTT                read4  +
                GAGTTAAATCTTACAGGGATGA      read5  -
```

Multiple applications

- detection of genomic variation (SNPs) → Variants
- detection of peaks : small RNA-seq, ChIP-seq → ChIP-seq
- metagenomics analysis → Metagenomics
- ...

RNA-seq : alternative approaches → RNA-seq

```
Read:      GACTGGGCGATCTCGACTTCG
           |||||  |||||||||| |||
Reference: GACTG--CGATCTCGACATCG
```

Matches/mismatches/insertions/deletions


merida


FUJI

How to align sequences, as many tools as applications

- BLAST
  - database search for homology detection
  - fast, accurate up to 85% identity
- short read mapping : Bowtie2, BWA
  - comparison of billions of short sequences against a genome
  - very fast, accurate up to 95% identity
- long read mapping : Minimap2
  - comparison of long reads against a genome
  - takes advantage of the length of reads (sampling)
  - very fast, accurate up to 85% identity
- tradeoff for speed versus sensitivity

Why is it difficult ?

- volume of the data (reads and reference genome)
- existence of sequencing errors in the reads
- orientation of read relative to reference genome not known
- existence of repetitive elements in the reference sequence
- divergence between the sequenced genome and the reference genome

How to choose a read mapper

- input data : read length, error profile, paired end
- hardware requirements : RAM, multithreading, ...
- ease of installation and use : configurability, options, ...
- quality of results : speed, sensitivity, multiple matches, paired-end matches
- documentation, user community, maintenance

Bowtie2

- optimized for Illumina reads
- large user-community
- well-documented and actively maintained
- suitable for all kinds of genomes

Nat Methods. 2012 Mar 4;9(4):357-9. doi: 10.1038/nmeth.1923.

Fast gapped-read alignment with Bowtie 2.

Langmead B[1], Salzberg SL.

+ 30 000 citations (Google Scholar)

`https://doi.org/10.1038/nmeth.1923`

Bowtie2

- Input : FASTA file or FASTQ file
- Output

```
20000 reads; of these:
  20000 (100.00%) were unpaired; of these:
    1247 (6.24%) aligned 0 times
    18739 (93.69%) aligned exactly 1 time
    14 (0.07%) aligned >1 times
93.77% overall alignment rate
```

SAM/BAM file containing all alignments found and their scores

SAM format

Sequence Alignment/Map format

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001   99 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  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

- text-based format
- header section (@)
- alignment section : one line per alignment (11 mandatory columns + optional fields)

Alignment section

```
r001   99 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;
```

1	QNAME	Query template NAME (read ID)
2	FLAG	information about the read (see next slide)
3	RNAME	References sequence NAME (chr, transcript,...)
4	POS	1-based leftmost mapping POSition
5	MAPQ	MAPping Quality (see later)
6	CIGAR	summary of alignment (see later)
7	RNEXT	Ref. name of the mate/NEXT read
8	PNEXT	Position of the mate/NEXT read
9	TLEN	observed Template LENgth
10	SEQ	read SEQuence
11	QUAL	read QUALity (Phred-score)

+ optional fields

SAM FLAG

Combination (sum) of properties of the read and its alignment

- template having multiple segments in sequencing → 1
- each segment properly aligned according to the aligner → 2
- segment unmapped → 4
- next segment in the template unmapped → 8
- SEQ being reverse complemented → 16
- SEQ of the next segment in the template being reversed → 32
- the first segment in the template → 64
- the last segment in the template → 128
- secondary alignment → 256
- not passing quality controls → 512
- PCR or optical duplicate → 1024
- supplementary alignment → 2048

`https://broadinstitute.github.io/picard/explain-flags.html`

SAM FLAG

Examples

template having multiple segments in sequencing + the first segment in the template + next segment in the template unmapped $\rightarrow$ flag ?

99

64 (first in pair) + 32 (mate reverse strand) + 2 (read mapped in proper pair) + 1 (read paired)

2064

2048 (supplementary alignment) + 16 (read reverse strand)

147

CIGAR string

Compact Idiosyncratic Gapped Alignment Report

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
                ||||||||  ||||  |||
+r001/1         TTAGATAAAGGATA*CTG
+r002           aaaAGATAA*GGATA
+r003        gcctaAGCTAA
+r004                          ATAGCT..............TCAGC
-r003                                ttagctTAGGC
-r001/2                                          CAGCGGCAT
```

```
@SQ SN:ref LN:45
r001   99 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  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

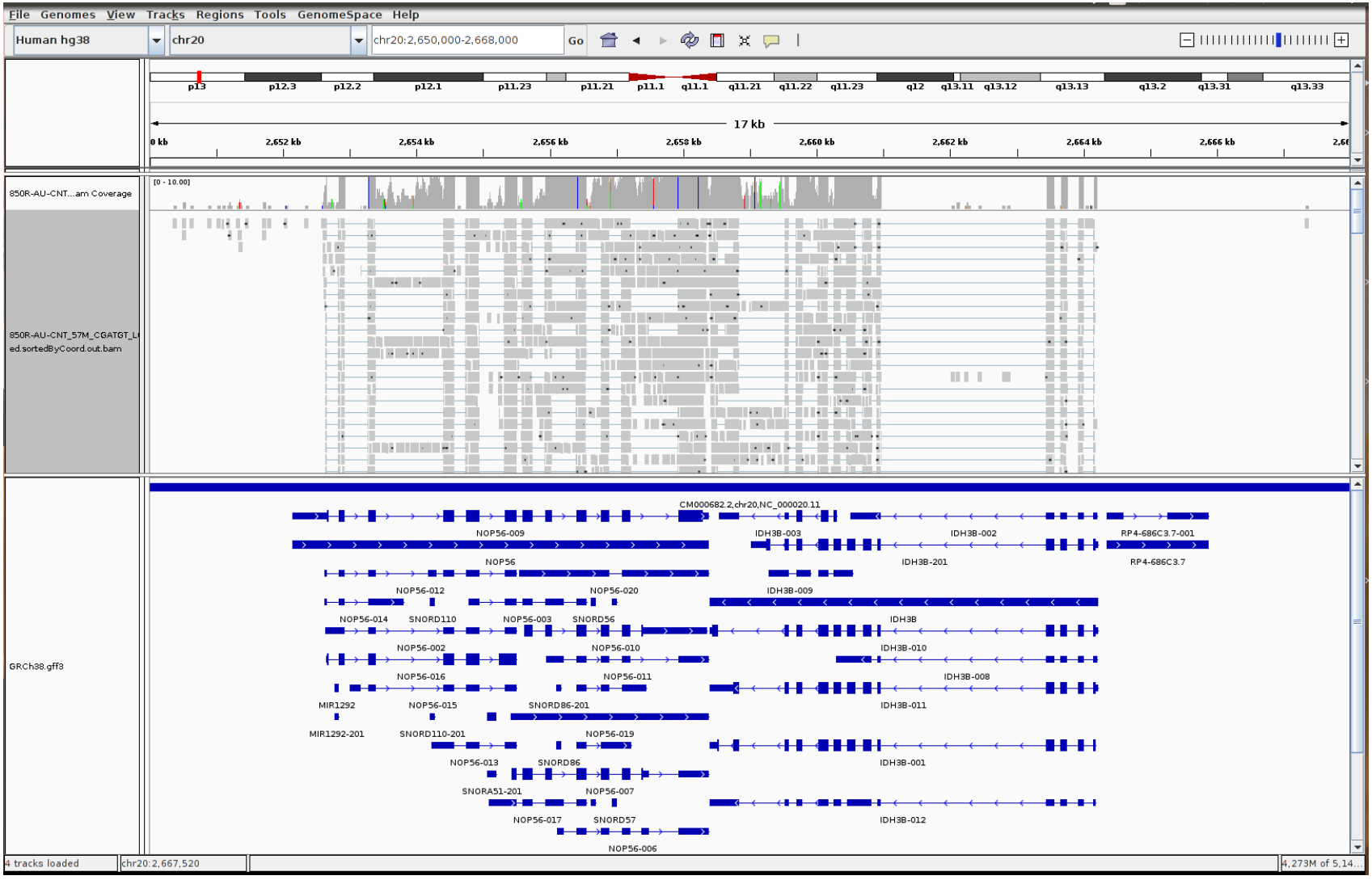
Op	BAM	Description	Consumes query	Consumes reference
M	0	alignment match (can be a sequence match or mismatch)	yes	yes
I	1	insertion to the reference	yes	no
D	2	deletion from the reference	no	yes
N	3	skipped region from the reference	no	yes
S	4	soft clipping (clipped sequences present in SEQ)	yes	no
H	5	hard clipping (clipped sequences NOT present in SEQ)	no	no
P	6	padding (silent deletion from padded reference)	no	no
=	7	sequence match	yes	yes
X	8	sequence mismatch	yes	yes

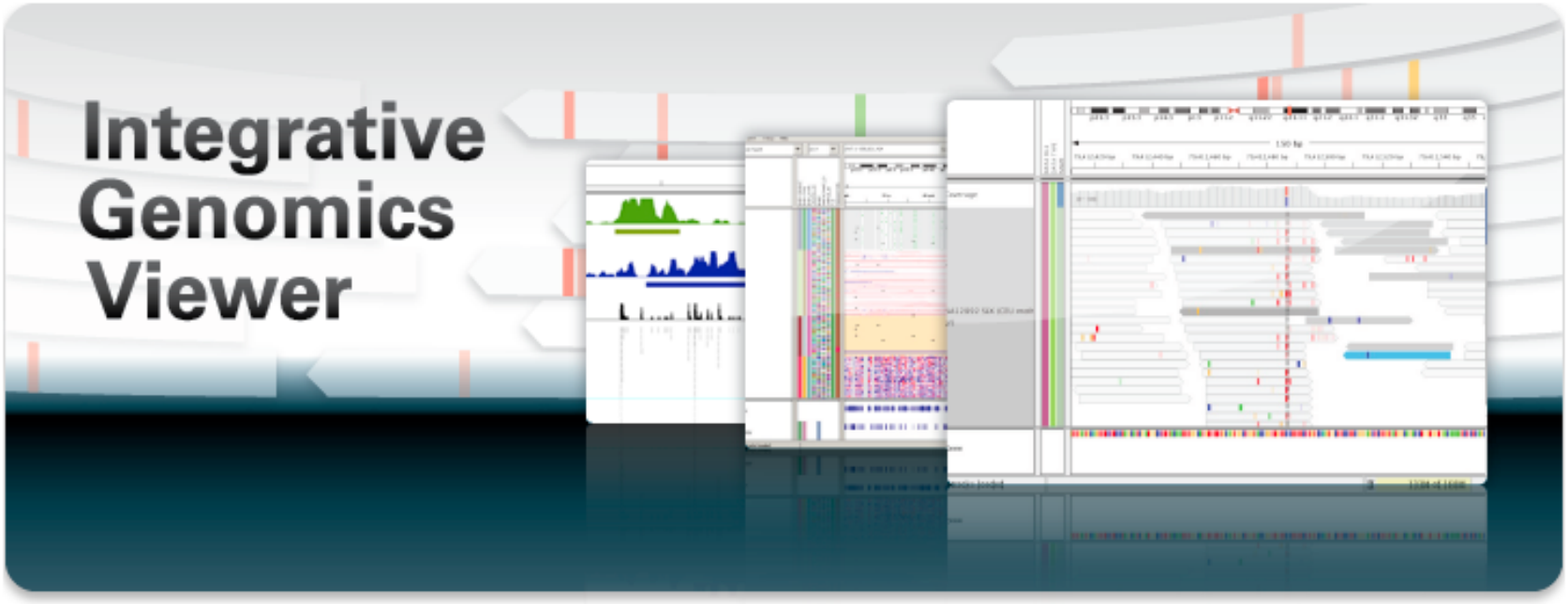
BAM format

- BAM : Binary sAM
- same data as in SAM
- *binary* format, more compact
  - smaller files
  - faster treatment for computers
  - easier transfer
- BAI : Index for BAM files
  - speed up data search and retrieve in a BAM file

Visualization of BAM files : genome browsers

- interactive visualization and exploration of genomes
- tracks : BAM files (alignments), annotation (GFF), variants (VCF), ...
- interconnection with external resources


File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg38
chr20
chr20:2,650,000-2,668,000
Go
p13 p12.3 p12.2 p12.1 p11.23 p11.21 p11.1 q11.1 q11.21 q11.22 q11.23 q12 q13.11 q13.12 q13.13 q13.2 q13.31 q13.33
17 kb
0 kb 2,652 kb 2,654 kb 2,656 kb 2,658 kb 2,660 kb 2,662 kb 2,664 kb 2,666 kb
850R-AU-CNT...am Coverage
[0 - 10.00]
850R-AU-CNT_57M_CGATGT_L
ed.sortedByCoord.out.bam
GRCh38.gff3
CM000682.2,chr20,NC_000020.11
NOP56-009
IDH3B-003
IDH3B-002
RP4-686C3.7-001
NOP56
IDH3B-201
RP4-686C3.7
NOP56-012
NOP56-020
IDH3B-009
NOP56-014
SNORD110
NOP56-003
SNORD56
IDH3B
NOP56-002
NOP56-010
IDH3B-010
NOP56-016
NOP56-011
IDH3B-008
MIR1292
NOP56-015
SNORD86-201
IDH3B-011
MIR1292-201
SNORD110-201
NOP56-019
NOP56-013
SNORD86
IDH3B-001
SNORA51-201
NOP56-007
NOP56-017
SNORD57
IDH3B-012
NOP56-006
4 tracks loaded
chr20:2,667,520
4,273M of 5,14...


Integrative
Genomics
Viewer

IGV – Integrative Genomics Viewer

- developed by the Broad Institute
- popular and versatile
- standalone application (2008)
- web application (2018)
- Galaxy

Paired inputs

Pair of files

```
@1/1
AGGGATGTGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTA
+
EGGEGGGDFGEEEAEECGDEGGFEEGEFGBEEDDECFEFDD@CDD<ED
@2/1
AGGGATGTGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTA
+
HHHHHHEGFHEEFEEHEEHHGGEGGGGEFGFGGGGHHHHFBEEEEEFG
```

```
@1/2
CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC
+
GHHHDFDFGFGEGFBGEGGEGEGGGHGFGHFHFHHHHHHHEF?EFEFF
@2/2
CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC
+
HHHHHHHHHHHHHGHHHHHHGHHHHHHHHHHHHFHHHFHHHHHHHHHH
```

Arguments : -1 -2

One single interleaved dataset

```
@1/1
AGGGATGTGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTA
+
EGGEGGGDFGEEEAEECGDEGGFEEGEFGBEEDDECFEFDD@CDD<ED
@1/2
CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC
+
GHHHDFDFGFGEGFBGEGGEGEGGGHGFGHFHFHHHHHHHEF?EFEFF
@2/1
AGGGATGTGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTA
+
HHHHHHEGFHEEFEEHEEHHGGEGGGGEFGFGGGGHHHHFBEEEEEFG
@2/2
CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC
+
HHHHHHHHHHHHHGHHHHHHGHHHHHHHHHHHHFHHHFHHHHHHHHHH
```

Argument : `--interleaved`

Description of the pairs

- relative orientation of the mates

  `--ff` forward forward
  `--fr` foward reverse
  `--rf` reverve forward

- fragment length (mate 1 + inner distance + mate 2)

  `-l` minimum length (default 0)
  `-X` maximum length (default 500)

Bowtie2 output - paired reads

```
10000 reads; of these:
  10000 (100.00%) were paired; of these:
    650 (6.50%) aligned concordantly 0 times
    8823 (88.23%) aligned concordantly exactly 1 time
    527 (5.27%) aligned concordantly >1 times
    ----
    650 pairs aligned concordantly 0 times; of these:
      34 (5.23%) aligned discordantly 1 time
    ----
    616 pairs aligned 0 times concordantly or discordantly
      1232 mates make up the pairs; of these:
        660 (53.57%) aligned 0 times
        571 (46.35%) aligned exactly 1 time
        1 (0.08%) aligned >1 times
96.70% overall alignment rate
```

- concordant

  the pair aligns with the expected relative mate orientation and with the expected range of distances between mates

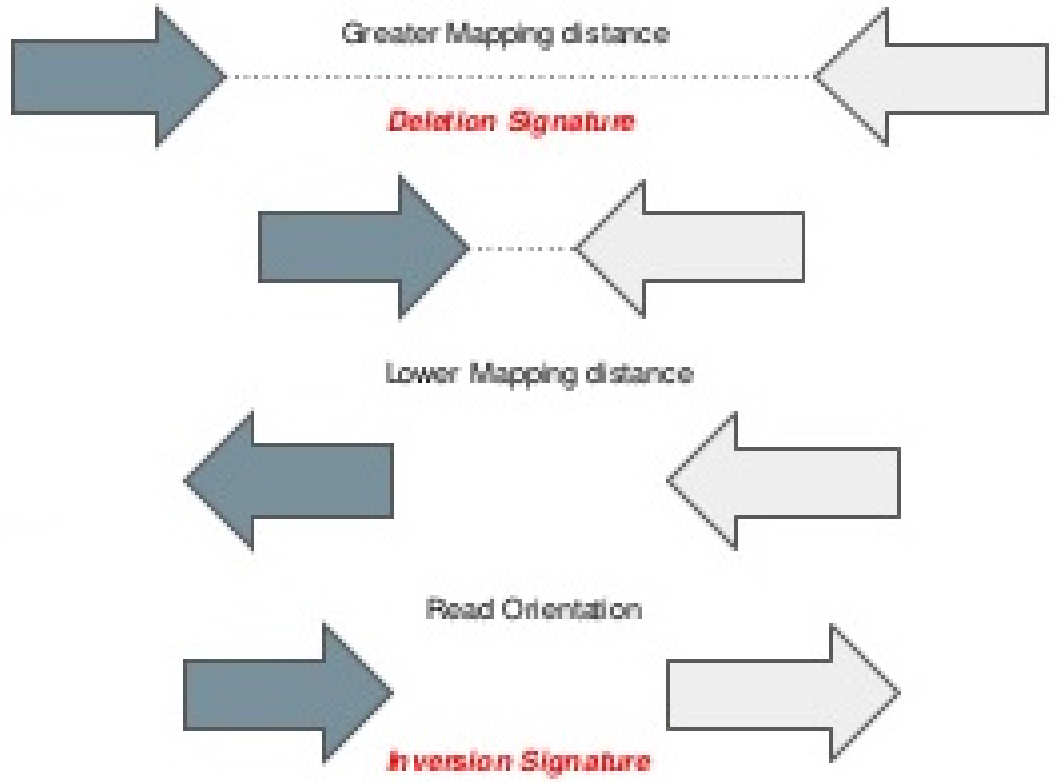
- discordant

  both mates have unique alignments, but the alignments do not match paired-end expectations

- the alignment score for a paired-end alignment equals the sum of the alignment scores of the individual mates.

template

discordant

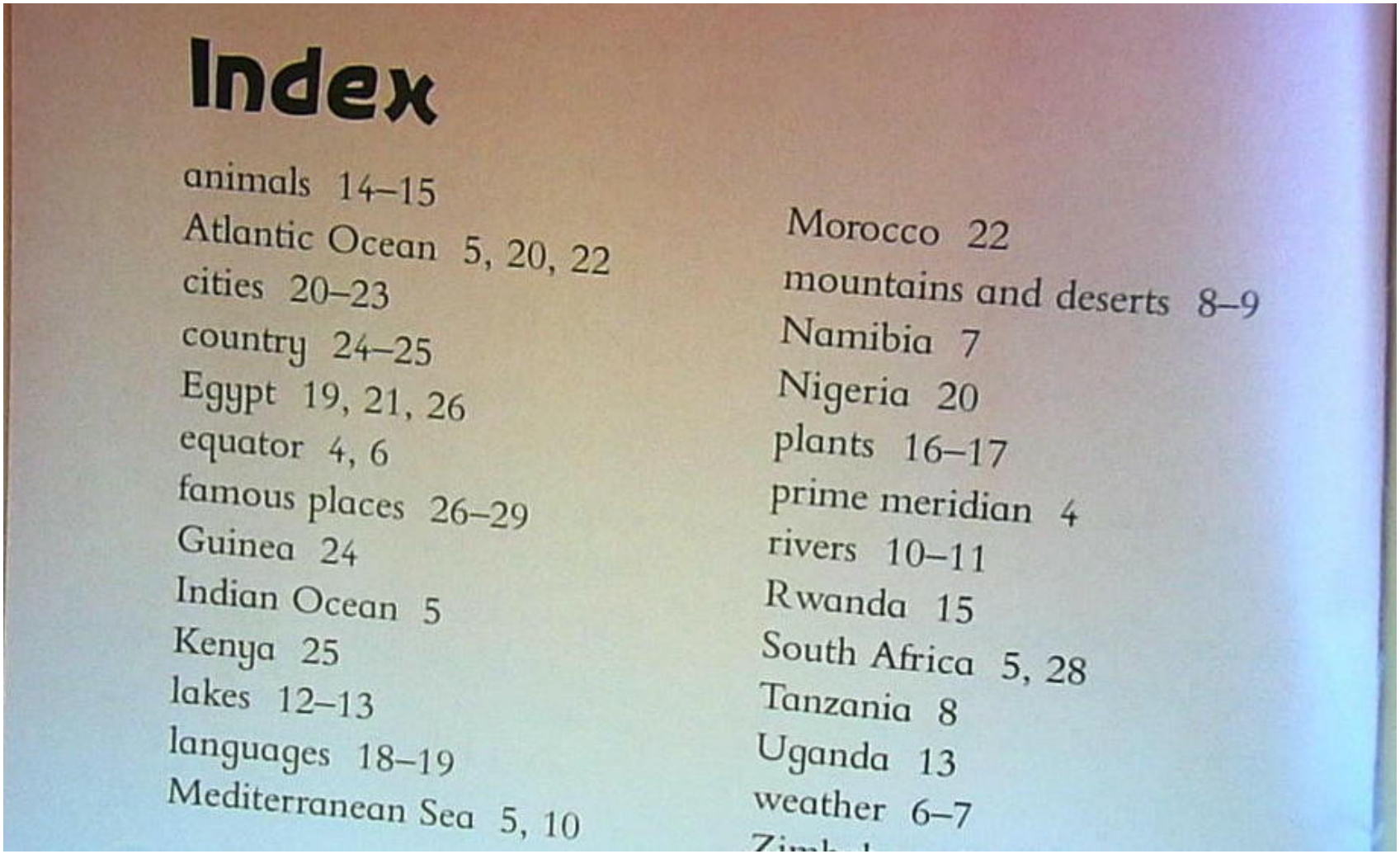


Bowtie2

Two main ingredients

- index for the reference sequence
- seed-and-extend strategy

Step 0 : build an index for the reference sequence



Index

Step 0 : build an index for the reference sequence

- compressed representation of the sequence, that is kept in memory

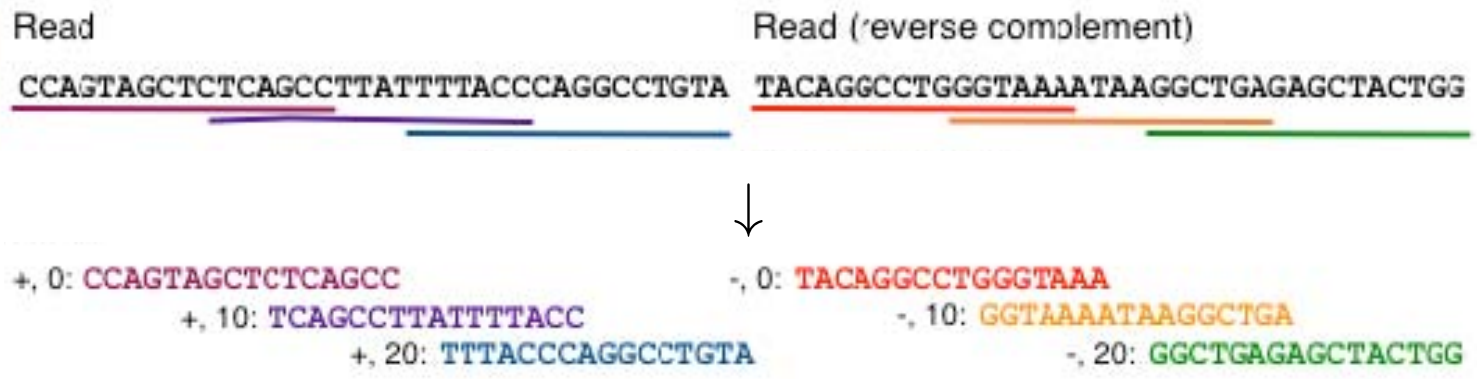
  "Where is this k-mer present in the sequence ?"

- Bowtie2 index : Burrows-Wheeler Transform and FM-index

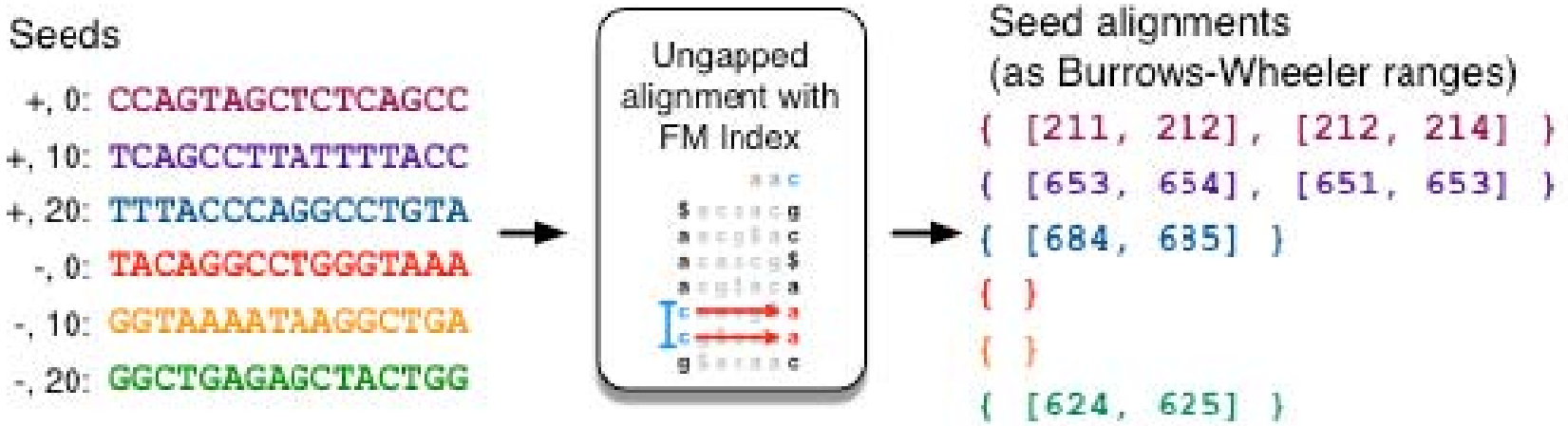
  Size of the index for the human genome : 3.2 Gb

- Other indexes : hashtable, suffix array, ...

Step 1 : extraction of *seed* substrings from the read and its reverse complement



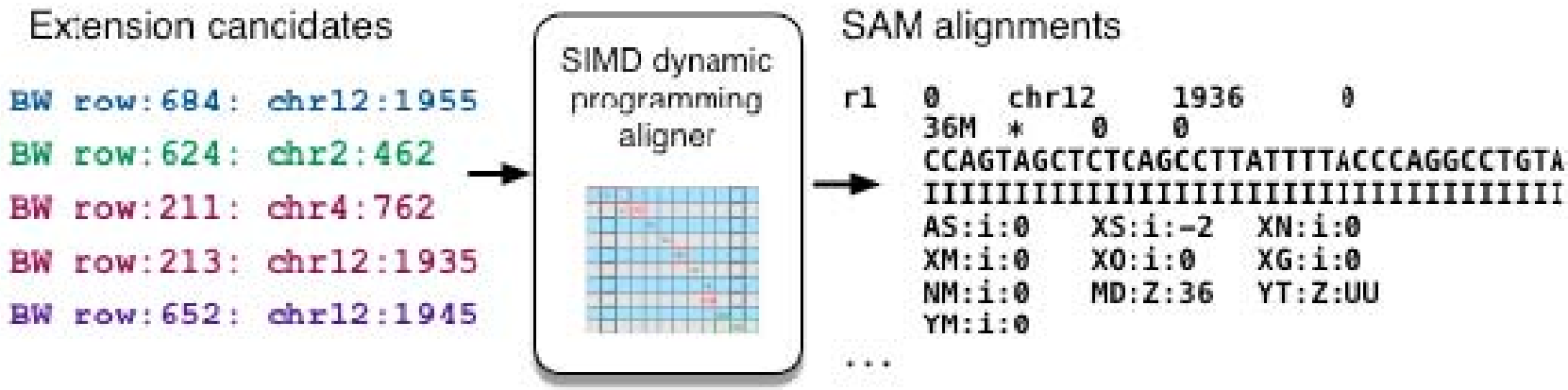
Step 2 : seed substrings are aligned to the indexed genome



no gaps, no ambiguous character in the reference

Step 3 : extension of seeds

```
... CGTCGTG CACTGCACG CATGGA ...
          |||||||||
... TCCACGT CACTGCACG CTGGAC ...
     <-----   seed    ----->
```



heuristics choice of seeds (randomized)

Alignment modes

`--end-to-end` (default)

align the entire read from one end to the other

`--local`

some characters may be trimmed ("soft clipped") from the ends in order to achieve the greatest possible alignment score

Seed options

Step 1 :

`-L <int>` length of the seed

`-i <func>` interval between extracted seeds

Step 2 :

`-N 0 or 1` number of mismatches permitted per seed.

Default 0

Seed options

Step 3 :

`-D <int>`

number of times Bowtie2 will try to extend the seed in order to find a new best matching location

`-R <int>`

number of times Bowtie2 will start a substring with a different offset from that at the beginning before reporting the best hit.

Alignment score

higher = more similar

```
Read:      GACTGGGCGATCTCGACTTCG
           |||||  |||||||||| |||
Reference: GACTG--CGATCTCGACATCG
```

- base mismatch penalty
- gap open penalty
- gap extension penalty
- match reward

Alignment score

higher = more similar

```
Read:      GACTGGGCGATCTCGACTTCG
           |||||  |||||||||| |||
Reference: GACTG--CGATCTCGACATCG
```

- base mismatch penalty -1
- gap open penalty -5
- gap extension penalty -0.5
- match reward 1

Score ?

Alignment score

higher = more similar

```
Read:      GACTGGGCGATCTCGACTTCG
           |||||  |||||||||| |||
Reference: GACTG--CGATCTCGACATCG
```

- base mismatch penalty -1
- gap open penalty -5
- gap extension penalty -0.5
- match reward 1

Score : 18 matches + 1 mismach + 1 gap of length 2
18 -1 - 5 -0.5
11.5

Alignment score

higher = more similar

```
Read:      GACTGGGCGATCTCGACTTCG
           |||||  |||||||||| |||
Reference: GACTG--CGATCTCGACATCG
```

- base mismatch penalty → depends on the quality value
- gap open penalty
- gap extension penalty
- match reward in local mode only

MAPQ, mapping quality score

- related to "uniqueness" of the alignment

  The greater the gap between the best alignment's score and the second-best alignment's score, the higher its mapping quality should be.

- ranges between 0 and 42

- poorly documented

+ many more options

Do you want to tweak input options?

No

See "Input Options" section of Help below for information

Do you want to tweak alignment options?

No

See "Alignment Options" section of Help below for information

Do you want to tweak scoring options?

No

See "Scoring Options" section of Help below for information

Do you want to use -a or -k options

No, do not set

Make sure you understand implications of setting -k and -a. See "Reporting Options" section of Help below for information on -k and -a options

Do you want to tweak effort options?

No

See "Effort Options" section of Help below for information

Do you want to tweak SAM/BAM Options?

No

See "Output Options" section of Help below for information

Do you want to tweak Other Options?

No

See "Other Options" section of Help below for information

Would you like the output to be a SAM file

Yes | No

By default, the output from this Bowtie2 wrapper is a sorted BAM file.

+ more in command-line mode

Preset options

In end-to-end mode :

`--very-fast` -D 5 -R 1 -N 0 -L 22 -i S,0,2.50

`--fast` -D 10 -R 2 -N 0 -L 22 -i S,0,2.50

`--sensitive` -D 15 -R 2 -N 0 -L 22 -i S,1,1.15 (default)

`--very-sensitive` -D 20 -R 3 -N 0 -L 20 -i S,1,0.50

In local mode :

`--very-fast-local` -D 5 -R 1 -N 0 -L 25 -i S,1,2.00

`--fast-local` -D 10 -R 2 -N 0 -L 22 -i S,1,1.75

`--sensitive-local` -D 15 -R 2 -N 0 -L 20 -i S,1,0.75 (default)

`--very-sensitive-local` -D 20 -R 3 -N 0 -L 20 -i S,1,0.50

+ default values for all other parameters

Do you want to use presets?

- (•) No, just use defaults
- ( ) Very fast end-to-end (--very-fast)
- ( ) Fast end-to-end (--fast)
- ( ) Sensitive end-to-end (--sensitive)
- ( ) Very sensitive end-to-end (--very-sensitive)
- ( ) Very fast local (--very-fast-local)
- ( ) Fast local (--fast-local)
- ( ) Sensitive local (--sensitive-local)
- ( ) Very sensitive local (--very-sensitive-local)

Reporting options

Default

Bowtie2 returns one good alignment

No guarantee that this alignment is the best possible in terms of alignment score.

`-k <int>`

specify how many alignments to return

`-a`

return all of the found alignments
very slow

Can I trust the results of Bowtie2 or BWA ?

- Generally, yes

- Be careful with
  - very short reads (<50 nt) with high sequencing error rate
  - low-complexity reads
  - genomes with repeats
  - genomes with high GC-content bias