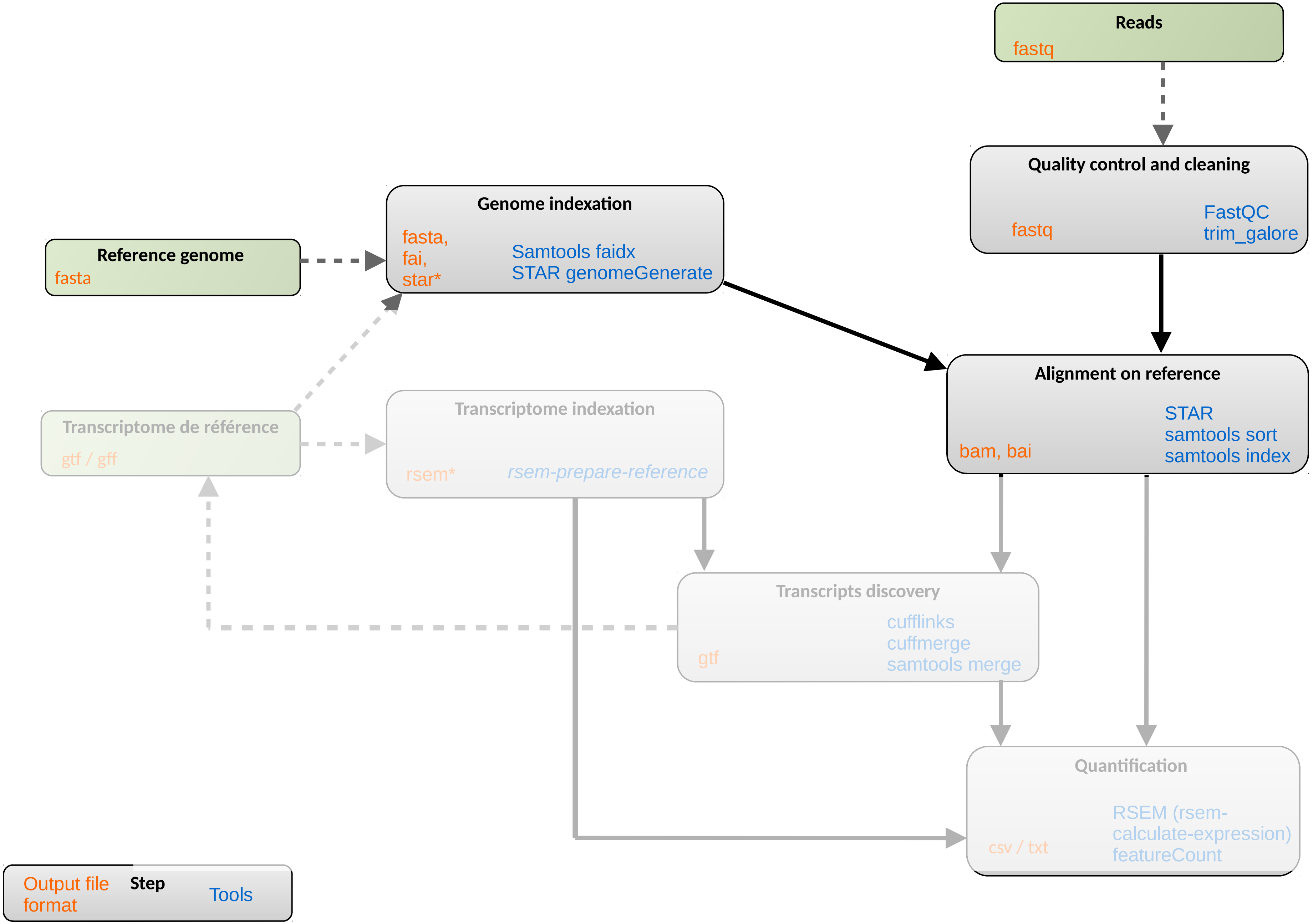
Analysis workflow



Reads
fastq

Quality control and cleaning
fastq
FastQC
trim_galore

Reference genome
fasta

Genome indexation
fasta,
fai,
star*
Samtools faidx
STAR genomeGenerate

Alignment on reference
bam, bai
STAR
samtools sort
samtools index

Transcriptome de référence
gtf / gff

Transcriptome indexation
rsem*
rsem-prepare-reference

Transcripts discovery
gtf
cufflinks
cuffmerge
samtools merge

Quantification
csv / txt
RSEM (rsem-calculate-expression)
featureCount

Output file format
Step
Tools

Summary -

Spliced read mapping & Visualisation

1. What is a spliced aligner?

2. Reference genome & transcriptome files formats

3. STAR principle and usage

4. BAM & Bed files formats

5. Visualisation with IGV

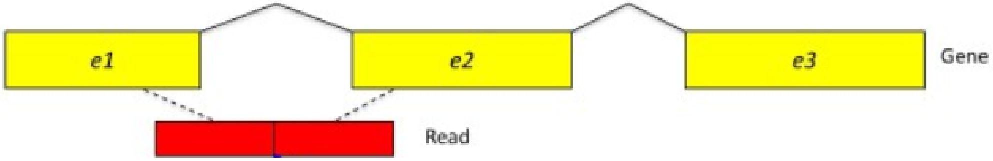
Aim -

Spliced read mapping & Visualisation

Aim: Discover the true location (origin) of each read on the reference.

Problems:

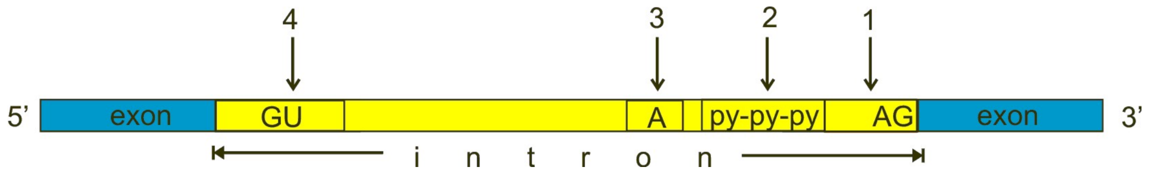
- Some features (repetitive regions, assembly errors, missing information) make it impossible for some reads.
- Reads may be split by potentially thousands of bases of intronic sequence.



And:
Do it in/with reasonable time/resources.

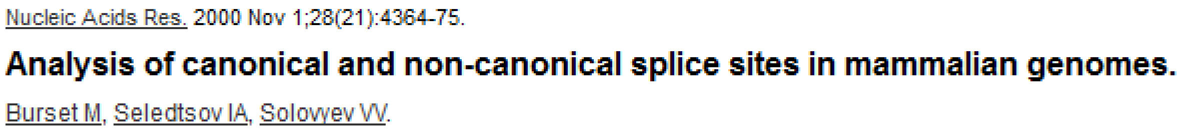
Splice sites

- Canonical splice site:
- which accounts for more than 99% of splicing
- GT and AG for donor and acceptor sites



http://en.wikipedia.org/wiki/RNA_splicing

- Non-canonical site:
- GC-AG splice site pairs, AT-AC pairs



Nucleic Acids Res. 2000 Nov 1;28(21):4364-75.

Analysis of canonical and non-canonical splice sites in mammalian genomes.

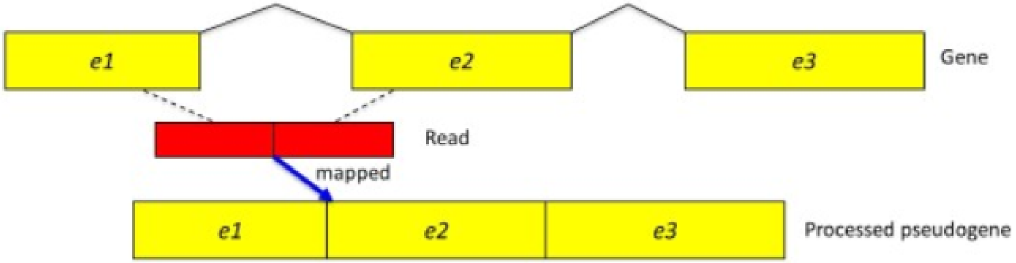
Burset M, Seledtsov IA, Solovyev VV.

- Trans-splicing:

  splicing that joins two exons that are not within the same RNA transcript

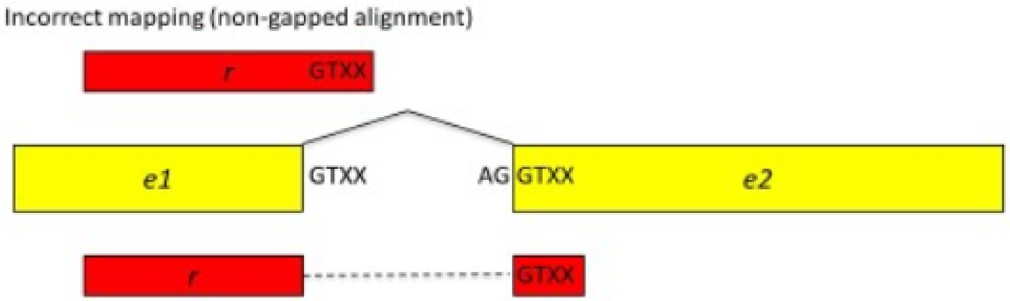
Hard case

- Lot of variations (sequencing errors, mutations)
- Repeats
- Reads spanning 3+ exons
- Gene or pseudogene



Kim et al, Genome Biology, 2013

- Small end “anchor”

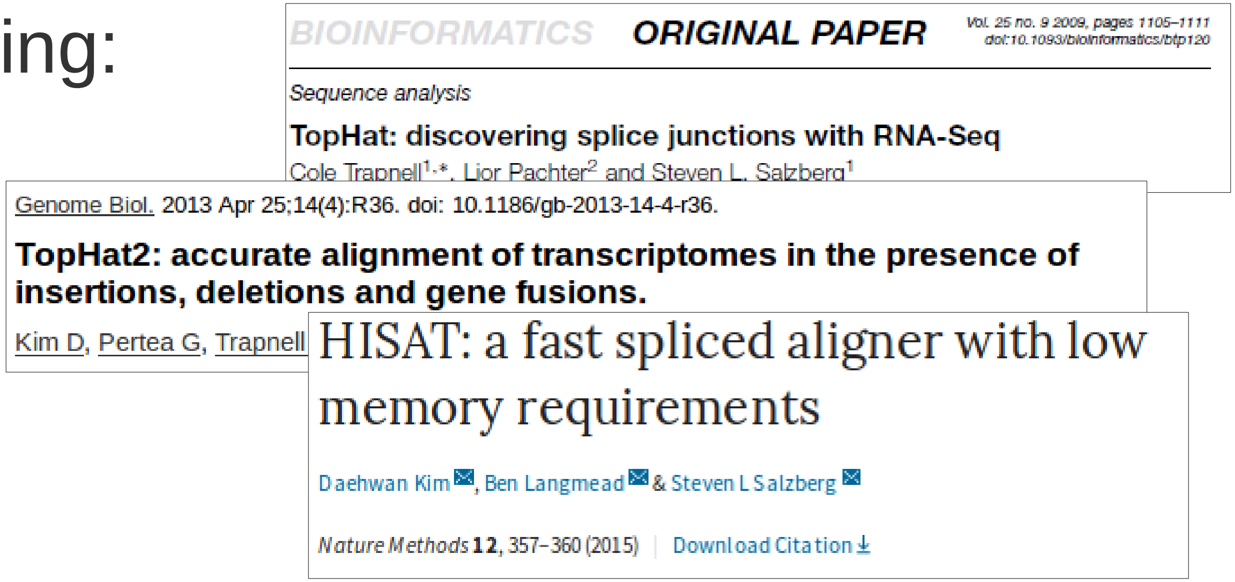


- Unknown junction inside poorly rarely expressed gene

Most used tools

Tools for splice-mapping:

- ~~Tophat:~~
- HISAT



BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 9 2009, pages 1105–1111 doi:10.1093/bioinformatics/btp120

Sequence analysis

TopHat: discovering splice junctions with RNA-Seq

Cole Trapnell[1,*], Lior Pachter[2] and Steven L. Salzberg[1]

Genome Biol. 2013 Apr 25;14(4):R36. doi: 10.1186/gb-2013-14-4-r36.

TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.

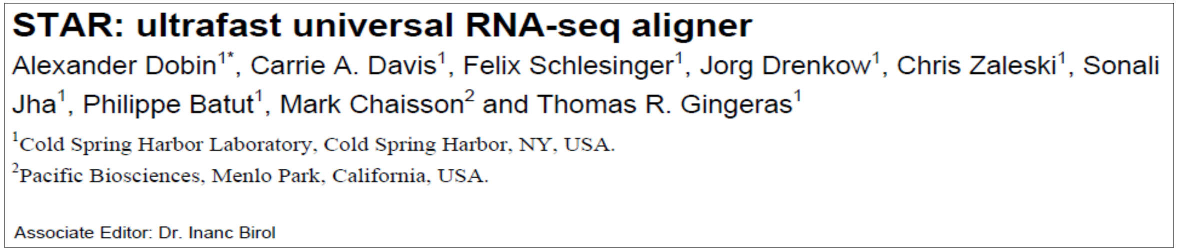
Kim D, Pertea G, Trapnell

HISAT: a fast spliced aligner with low memory requirements

Daehwan Kim, Ben Langmead & Steven L Salzberg

Nature Methods **12**, 357–360 (2015) | Download Citation

- STAR:



STAR: ultrafast universal RNA-seq aligner

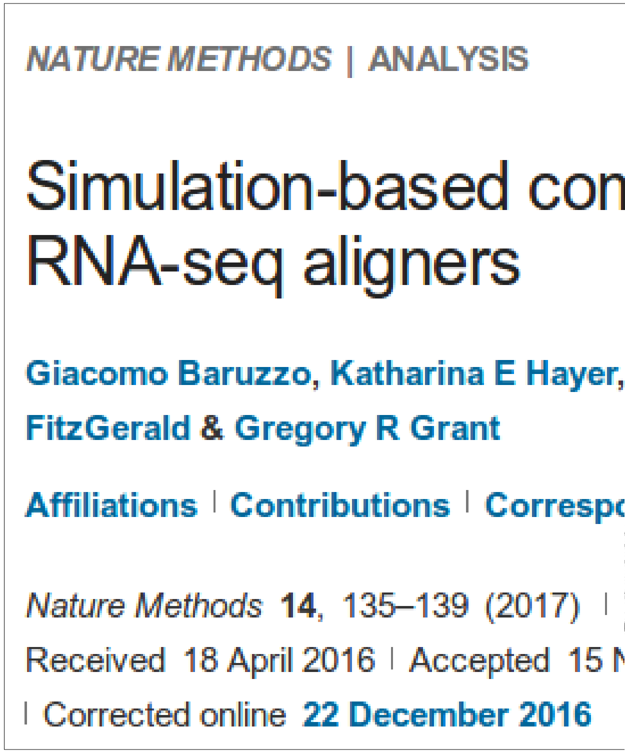
Alexander Dobin[1*], Carrie A. Davis[1], Felix Schlesinger[1], Jorg Drenkow[1], Chris Zaleski[1], Sonali Jha[1], Philippe Batut[1], Mark Chaisson[2] and Thomas R. Gingeras[1]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA.

[2]Pacific Biosciences, Menlo Park, California, USA.

Associate Editor: Dr. Inanc Birol

Benchmarks



NATURE METHODS | ANALYSIS

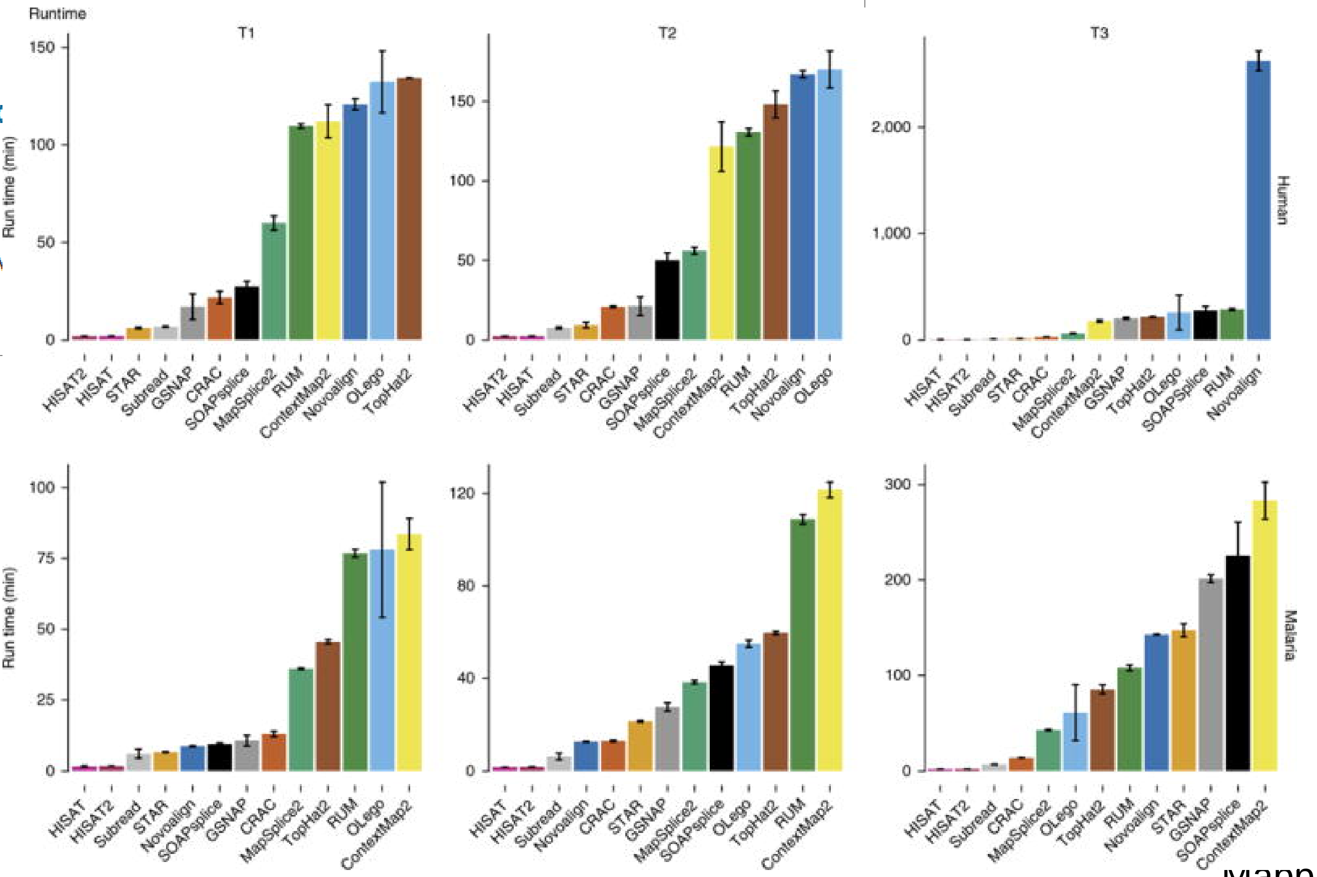
Simulation-based comprehensive benchmarking of RNA-seq aligners

Giacomo Baruzzo, Katharina E Hayer, Eun Ji Kim, Barbara Di Camillo, Garret A FitzGerald & Gregory R Grant

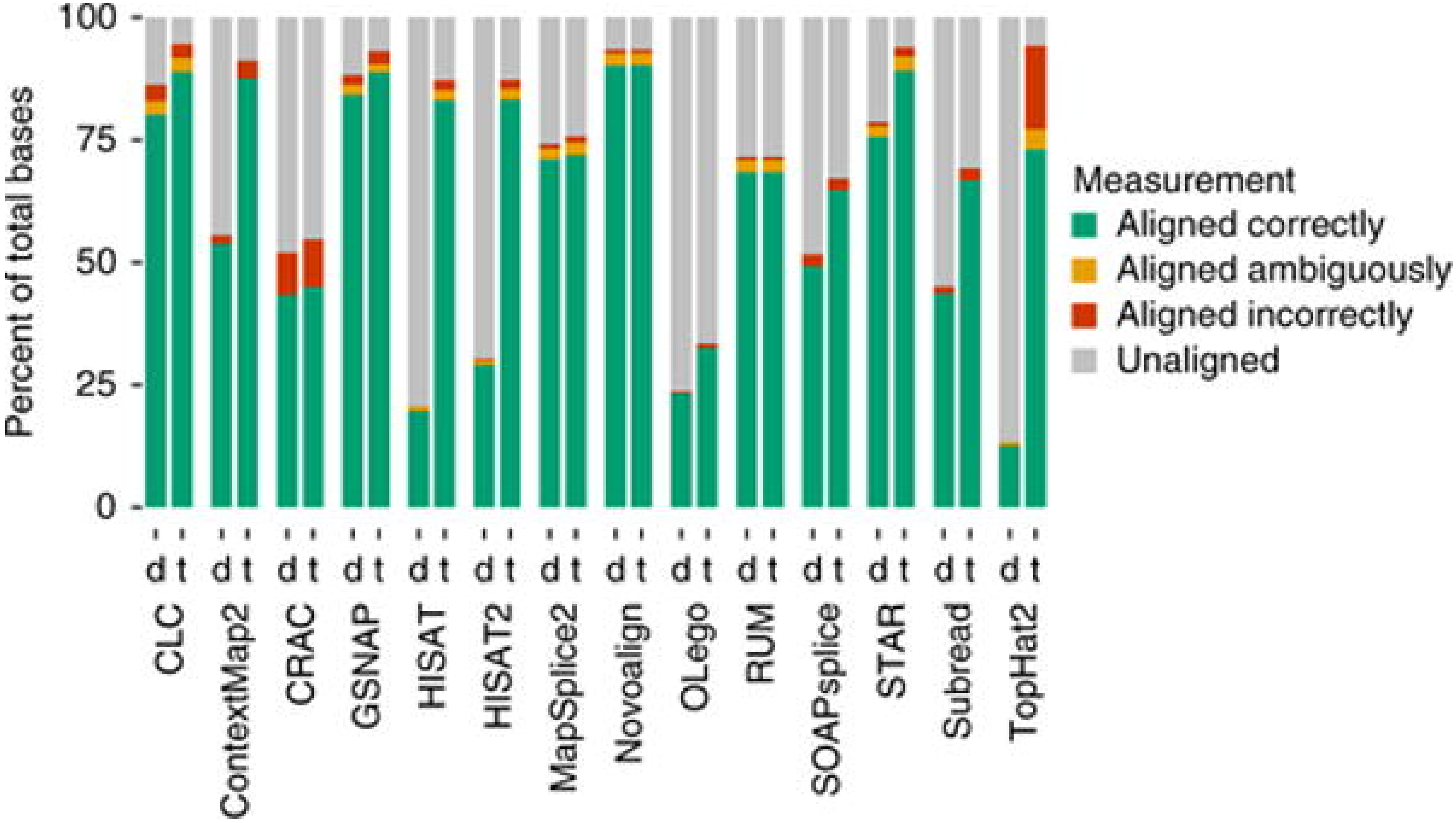
Affiliations | Contributions | Correspo

Nature Methods **14**, 135–139 (2017) |
Received 18 April 2016 | Accepted 15 M
| Corrected online 22 December 2016

Run time:



Tuning parameters

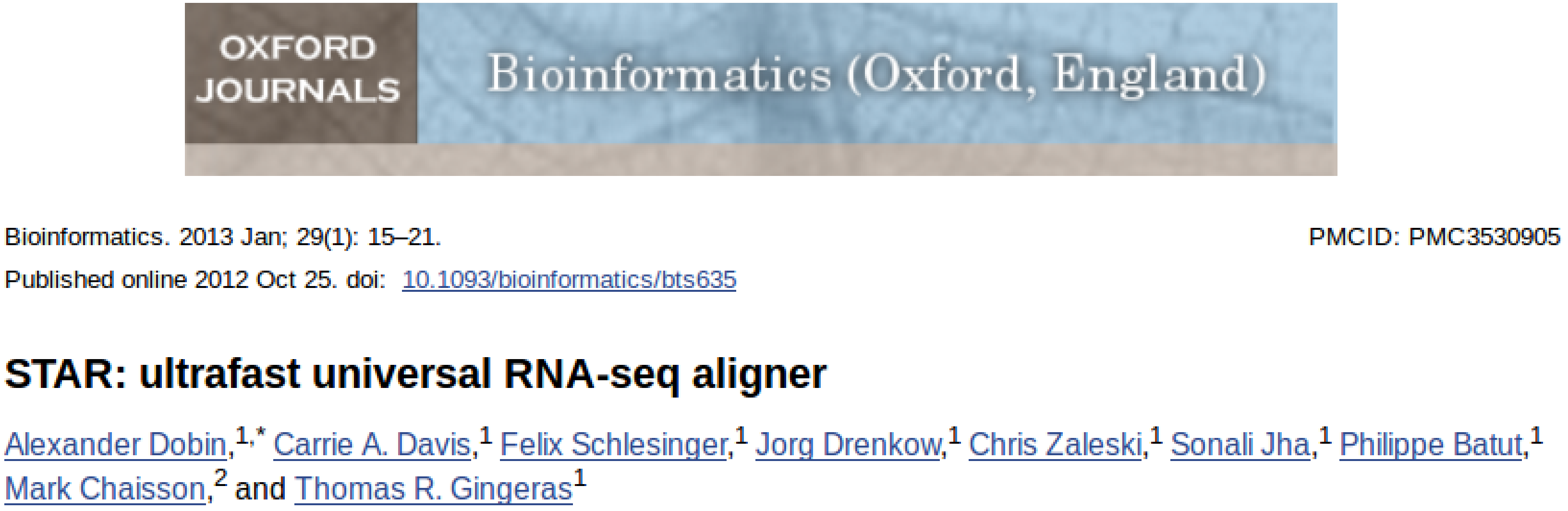


on the human-T3-data base-level statistics.

« Therefore, an algorithm that is robust to parameter settings and exhibits good performance using defaults is desirable »

« most reliable general-purpose aligners appear to be CLC, Novoalign, GSNAP, and STAR. »

rnaSTAR



OXFORD JOURNALS

Bioinformatics (Oxford, England)

Bioinformatics. 2013 Jan; 29(1): 15–21.
Published online 2012 Oct 25. doi: 10.1093/bioinformatics/bts635

PMCID: PMC3530905

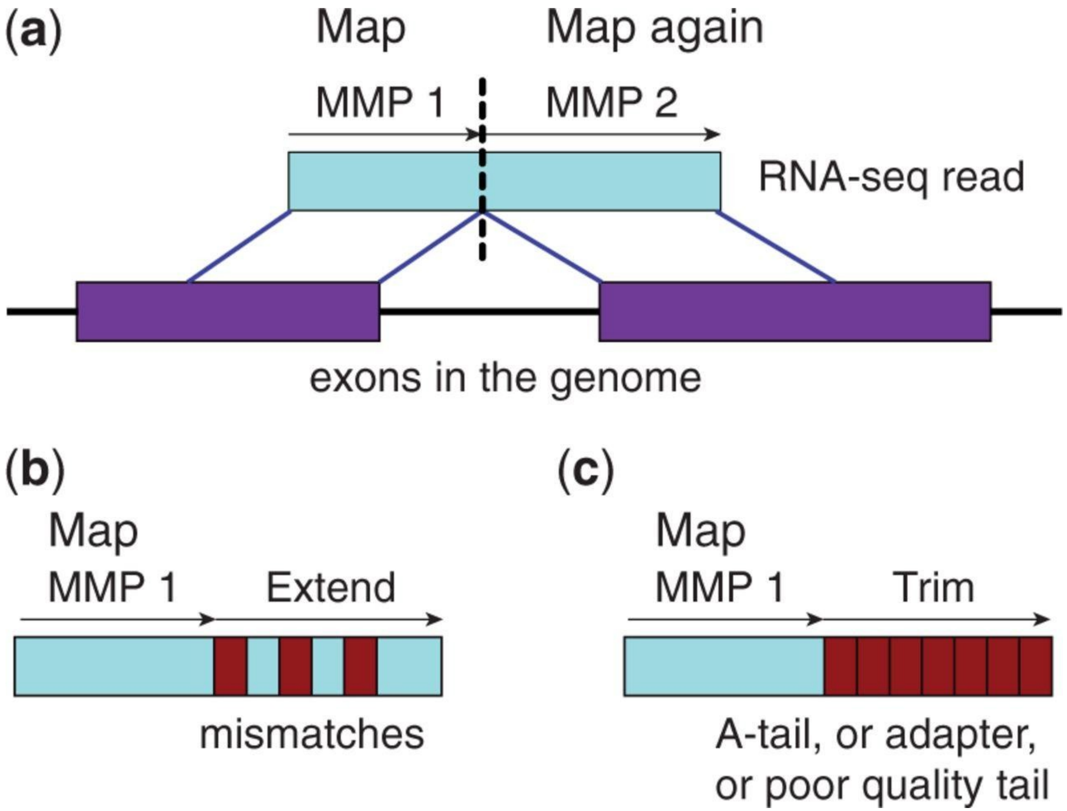
STAR: ultrafast universal RNA-seq aligner

Alexander Dobin,[1,*] Carrie A. Davis,[1] Felix Schlesinger,[1] Jorg Drenkow,[1] Chris Zaleski,[1] Sonali Jha,[1] Philippe Batut,[1] Mark Chaisson,[2] and Thomas R. Gingeras[1]

- Spliced Transcripts Alignment to a Reference
- Outperforms other aligners by more than a factor of 50 in mapping speed

rnaSTAR strategy

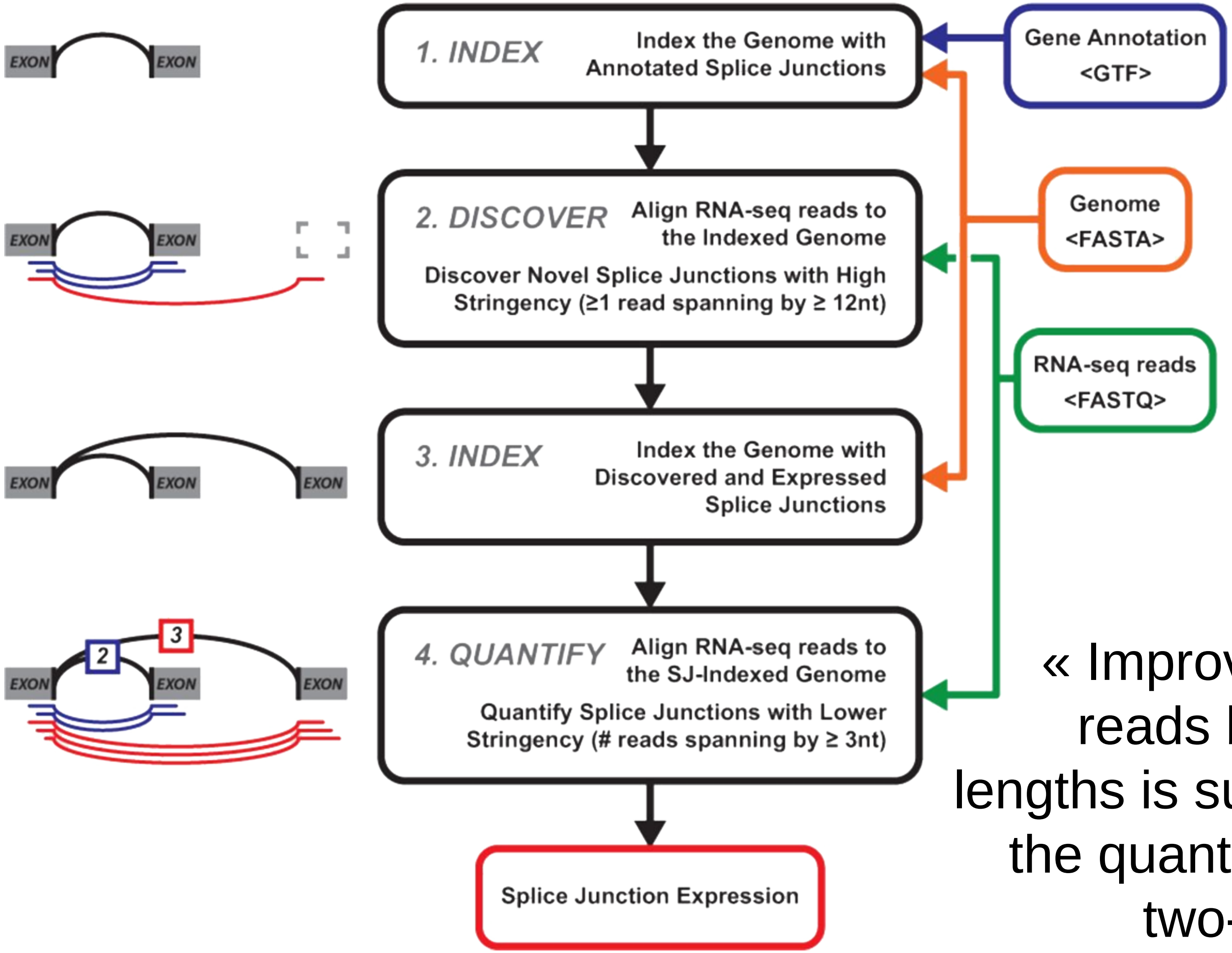
- search for a MMP from the 1st base
- MMP search repeated for the unmapped portion next to the junction
- do it in both fwd and rev directions
- cluster seeds from the mates of paired-end RNA-seq reads



Dobin *et al*, Bioinformatics, 2011

Soft-clipping is the main difference between Tophat and STAR

STAR : two passes strategy


EXON
EXON
1. INDEX
Index the Genome with Annotated Splice Junctions
Gene Annotation <GTF>
2. DISCOVER
Align RNA-seq reads to the Indexed Genome
Discover Novel Splice Junctions with High Stringency (≥1 read spanning by ≥ 12nt)
Genome <FASTA>
RNA-seq reads <FASTQ>
3. INDEX
Index the Genome with Discovered and Expressed Splice Junctions
2
3
4. QUANTIFY
Align RNA-seq reads to the SJ-Indexed Genome
Quantify Splice Junctions with Lower Stringency (# reads spanning by ≥ 3nt)
Splice Junction Expression

« Improved ability to align reads by short spanning lengths is sufficient to explain the quantification benefit of two-pass alignment »

Veeneman et al, Bioinformatics, 2016

STAR indexing

```
module load bioinfo/starXXX
STAR --runMode genomeGenerate --genomeDir
genome_dir --genomeFastaFiles genome.fasta
```

To use *N* CPUs, add: `--runThreadN N`

With an annotation: `--sjdbGTFfile annot.gtf`

Some precomputed indices are already available:
http://labshare.cshl.edu/shares/gingeraslab/www-data/dobin/STAR/STARgenomes
or on your preferred platform: /bank/STARdb

Where to find a reference genome?

Retrieving the genome file (fasta):

- The Genome Reference Consortium

http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/

- NCBI chromosome naming with « | » not well supported by mapping software

- Prefer EMBL

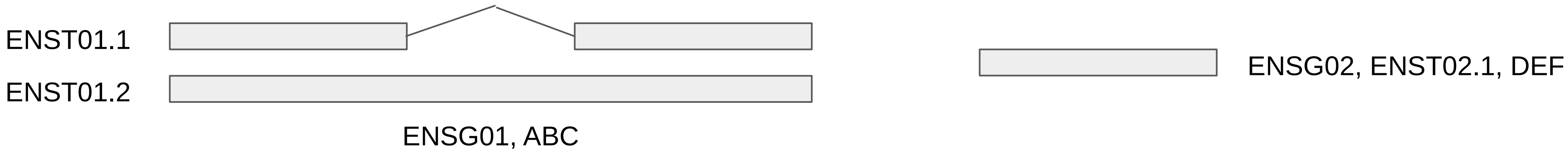
http://www.ensembl.org/info/data/ftp/index.html

Reference transcriptome file

What is a **GTF** file ?

- An annotation file: loci of coding genes (transcripts, CDS, UTRs), non-coding genes, etc.
- Gene Transfer Format (derived from GFF)

```
chr source  feature start end  score strand frame [attributes]
1   ENSEMBL exon    1000  2000 .     +      .     gene_id "ENSG01"; transcript_id “ENST01.1”; gene_name "ABC";
1   ENSEMBL exon    3000  4000 .     +      .     gene_id "ENSG01"; transcript_id “ENST01.1”; gene_name "ABC";
1   ENSEMBL exon    1000  4000 .     +      .     gene_id "ENSG01"; transcript_id “ENST01.2”; gene_name "ABC";
1   ENSEMBL exon    5000  6000 .     +      .     gene_id "ENSG02"; transcript_id “ENST02.1”; gene_name "DEF";
```



- `gene_id` *value* : unique identifier for the gene.
- `transcript_id` *value* : unique identifier for the transcript.

The chromosome names MUST be the same in the gtf file and fasta files (e.g. `chr1` vs `Chr1` vs `1`).

http://genome.ucsc.edu/FAQ/FAQformat.html#format4

Hands-on : STAR

Exercise n°3

Create a directory for the genome and annotation files.

Get the FASTA and GTF files from:
http://genoweb.toulouse.inra.fr/~formation/19_Rnaseq_Cli/data/reference/

Create the STAR index.
Tip: you can allocate *N* CPUs with the `sbatch -c 8`

STAR alignment

```
module load bioinfo/starXXX
STAR --genomeDir genome_dir
--readFilesIn read1.fastq.gz read2.fastq.gz
--readFilesCommand zcat
--sjdbGTFfile transcriptome.gtf
--alignIntronMin 20 --alignIntronMax 500000
--outSAMtype BAM SortedByCoordinate → sort
--outSAMstrandField intronMotif → for cufflinks
--alignSoftClipAtReferenceEnds No → for cufflinks
--outSAMattrIHstart 0 → for cufflinks or StringTie
--outFilterType BySJout → filter by splice site
--outFilterIntronMotifs RemoveNoncanonical → filter
--quantMode TranscriptomeSAM GeneCounts → for RSEM
 --outSAMattributes All → more information
--outFileNamePrefix sampleName
--runThreadN 4
```

STAR options

Intron size

```
 --alignIntronMin  20
 --alignIntronMax 500000
```

Allow soft-clipping past the end of chr (for cufflinks No)

`--alignSoftClipAtReferenceEnds No` [default Yes]

Output format:

```
--outSAMtype BAM SortedByCoordinate [SAM]
```

Output SAM/BAM alignments to transcriptome into a separate file (for RSEM)

```
--quantMode TranscriptomeSAM
```

→ need `:--sjdbGTFfile annot.gtf`

Output read unmapped

```
--outReadsUnmapped Fastx
```

STAR options

Add more tags:

`--outSAMattributes All`

Default output file name: `Aligned.bam` Modify prefix:

`--outFileNamePrefix` *`prefix`*

Infer strand using intron motifs (for Cufflinks)

`--outSAMstrandField intronMotif [None]`

Start IH at `--outSAMattrIHstart 0 [1]` (for Cufflinks)

STAR options

Remove reads that did not pass the junction filter:

```
--outFilterType BySJOut [Normal]
```

Filter out alignments with non-canonical intron motifs

```
--outFilterIntronMotifs RemoveNoncanonical
```

Mismatches :

```
--outFilterMismatchNmax [default: 10]
```

Limit multimap outputed:

```
--outFilterMultimapNmax [Default: 10]
```

> Flag of secondary alignment 0x100

Too short alignemnt

```
--outFilterMatchNminOverLread 0.66
--outFilterScoreMinOverLread 0.66
```

STAR - two passes mode

- First pass: discover new junctions.

- Second pass: run again with knowing the new junctions.
(most useful for poorly annotated genomes.)

```
--twopassMode [None|Basic]
```

Defines the number of reads to be mapped in the 1st pass :

```
--twopass1readsN [-1]
```

STAR Output files

Outputs (w/o specific options except `BAM SortedByCoordinate`):

- `Aligned.sortedByCoord.out.bam`: list of read alignments in SAM format compressed
- `Log.out`: main log file with a lot of detailed information about the run (for troubleshooting)
- `Log.progress.out`: reports job progress statistics
- `Log.final.out`: summary mapping statistics after mapping job is complete, very useful for quality control.
- `SJ.out.tab`: contains high confidence collapsed splice junctions in tab-delimited format

(chr, intron start, end, strand, intron motif, in database, # uniquely mapping reads, # multi, max. overhang)

STAR technical issues

- Temporary disk space:
  - Indexing the mouse genome requires 128GB and 1 hour on 6 slots.
  - Mapping a 16M paired-end reads requires 110GB and 4 mins on 6 slots.

- Available cluster:
  - New : 48 nodes with 32 cores and 256 GB of ram per node
  - Old : 68 nodes with 20 cores and 256 GB of ram per node

Hands-on : STAR

Exercise n°3

Map the 2 FASTQ files.

Do not forget to provide a different output file name for each set.

Index the output BAM files with:

```
samtools index file.bam
```

→ Then BAM format presentation.

SAM / BAM formats

Sequence Alignment/Map format:

- Each line stores an alignment/map

```
Coor    12345678901234  5678901234567890123456789012345
ref     AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1       TTAGATAAAGGATA*CTG
+r002        aaaAGATAA*GGATA
+r003      gcctaAGCTAA
+r004                    ATAGCT..............TCAGC
-r003                           ttagctTAGGC
-r001/2                                       CAGCGGCAT

name flag chr start mapQ  cigar   nNext sNext tlen     seq         qual   tags
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
```

- Header stores genome information

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
```

Fields

```
Coor    12345678901234  5678901234567890123456789012345
ref     AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1       TTAGATAAAGGATA*CTG
+r002        aaaAGATAA*GGATA
+r003      gcctaAGCTAA
+r004                    ATAGCT..............TCAGC
-r003                           ttagctTAGGC
-r001/2                                       CAGCGGCAT

name flag chr start mapQ  cigar   nNext sNext tlen     seq         qual   tags
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
```

- Flags: https://broadinstitute.github.io/picard/explain-flags.html
- MapQ: similar to a phred score
- nNext: = means same chr
- In general, * means NA

CIGAR

```
Coor    12345678901234  5678901234567890123456789012345
ref     AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1       TTAGATAAAGGATA*CTG
+r002        aaaAGATAA*GGATA
+r003      gcctaAGCTAA
+r004                    ATAGCT..............TCAGC
-r003                           ttagctTAGGC
-r001/2                                       CAGCGGCAT

name flag chr start mapQ  cigar   nNext sNext tlen     seq          qual   tags
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
```

- 30M means 30 matches or mismatches
- I and D : insertion/deletion
- S and H : soft/hard clipping

Tags

```
Coor    12345678901234  5678901234567890123456789012345
ref     AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1       TTAGATAAAGGATA*CTG
+r002        aaaAGATAA*GGATA
+r003      gcctaAGCTAA
+r004                    ATAGCT..............TCAGC
-r003                           ttagctTAGGC
-r001/2                                       CAGCGGCAT

name flag chr start mapQ  cigar   nNext sNext tlen     seq         qual   tags
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
```

- Format: *2-letter name*:*format*:*value* (many different)
- NM: # mismatches
- SA: chimeric reads
- NH, HI: # hits for this sequence, hit index
- AS: alignment score
- nM: # mismatches per fragment

SAM / BAM

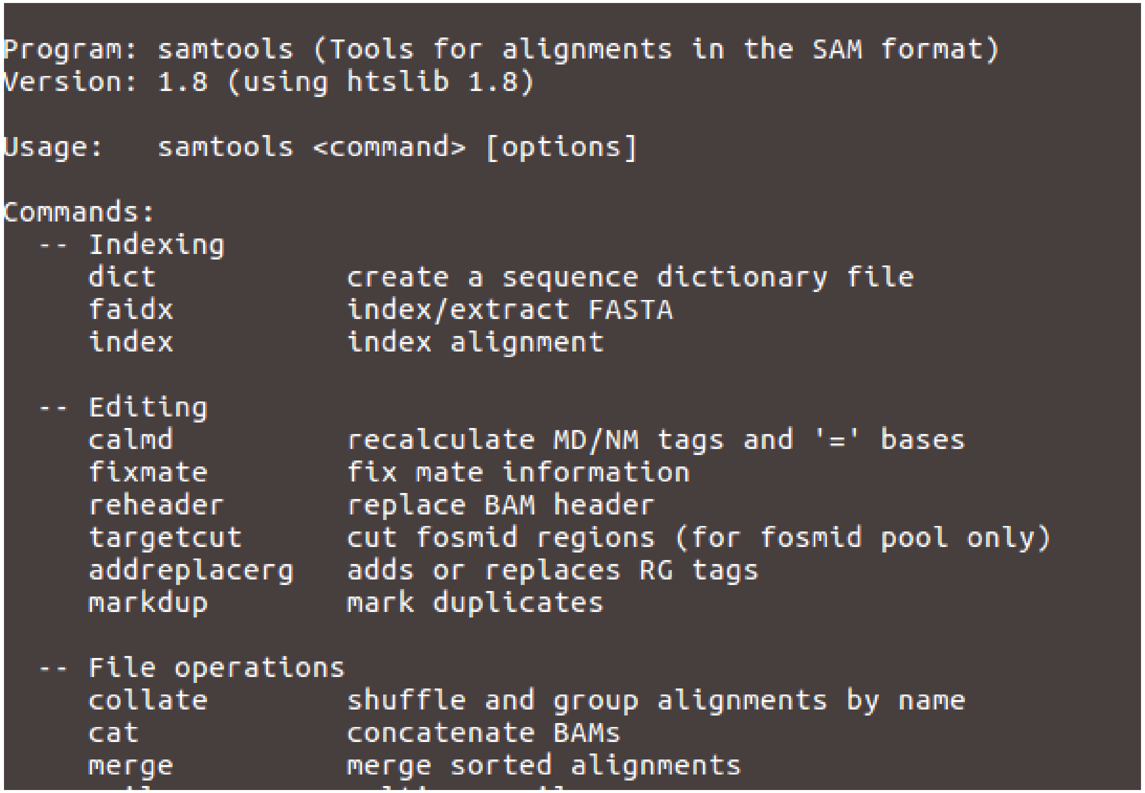
BAM (Binary Alignment/Map) format:

- Compressed binary representation of SAM
- Greatly reduces storage space requirements to about 27% of original SAM
- samtools: reading, writing, and manipulating BAM files
- Most tools require a sorted and indexed BAM file.

- To be viewed a bam file must be indexed :

```
samtools index
```

samtools



```
Program: samtools (Tools for alignments in the SAM format)
Version: 1.8 (using htslib 1.8)

Usage:   samtools <command> [options]

Commands:
  -- Indexing
     dict           create a sequence dictionary file
     faidx          index/extract FASTA
     index          index alignment

  -- Editing
     calmd          recalculate MD/NM tags and '=' bases
     fixmate        fix mate information
     reheader       replace BAM header
     targetcut      cut fosmid regions (for fosmid pool only)
     addreplacerg   adds or replaces RG tags
     markdup        mark duplicates

  -- File operations
     collate        shuffle and group alignments by name
     cat            concatenate BAMs
     merge          merge sorted alignments
```

```
module load bioinfo/samtools-1.8
```

Bam → sam

```
samtools view in.bam
```

Sam → bam

```
samtools view in.sam > out.bam
```

Sort

```
samtools sort -o out.bam in.bam
```

Index

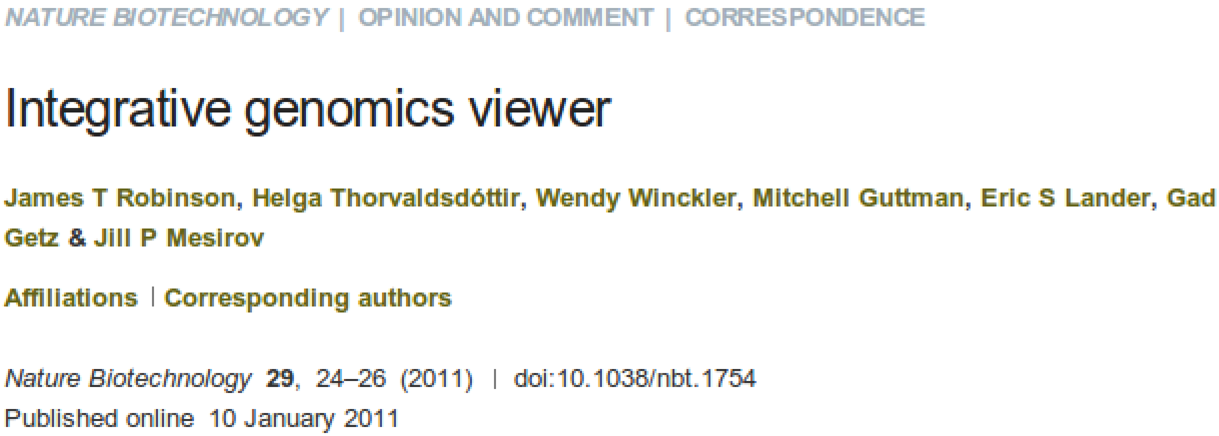
```
samtools sort in.bam
```

Global options nb threads:

```
-@ 4
```

Visualizing alignments on IGV

http://www.broadinstitute.org/igv/home



NATURE BIOTECHNOLOGY | OPINION AND COMMENT | CORRESPONDENCE

Integrative genomics viewer

James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz & Jill P Mesirov

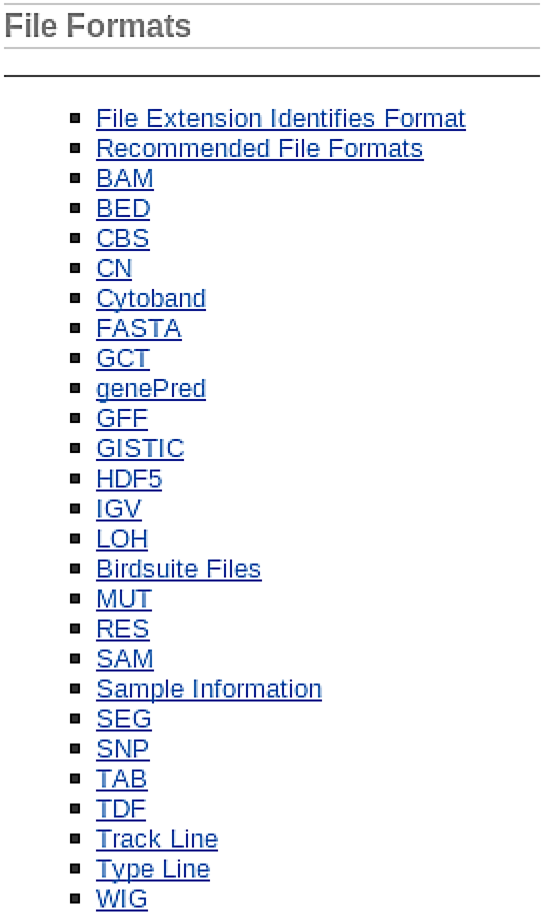
Affiliations | Corresponding authors

Nature Biotechnology **29**, 24–26 (2011) | doi:10.1038/nbt.1754

Published online 10 January 2011

Visualizing alignments on IGV

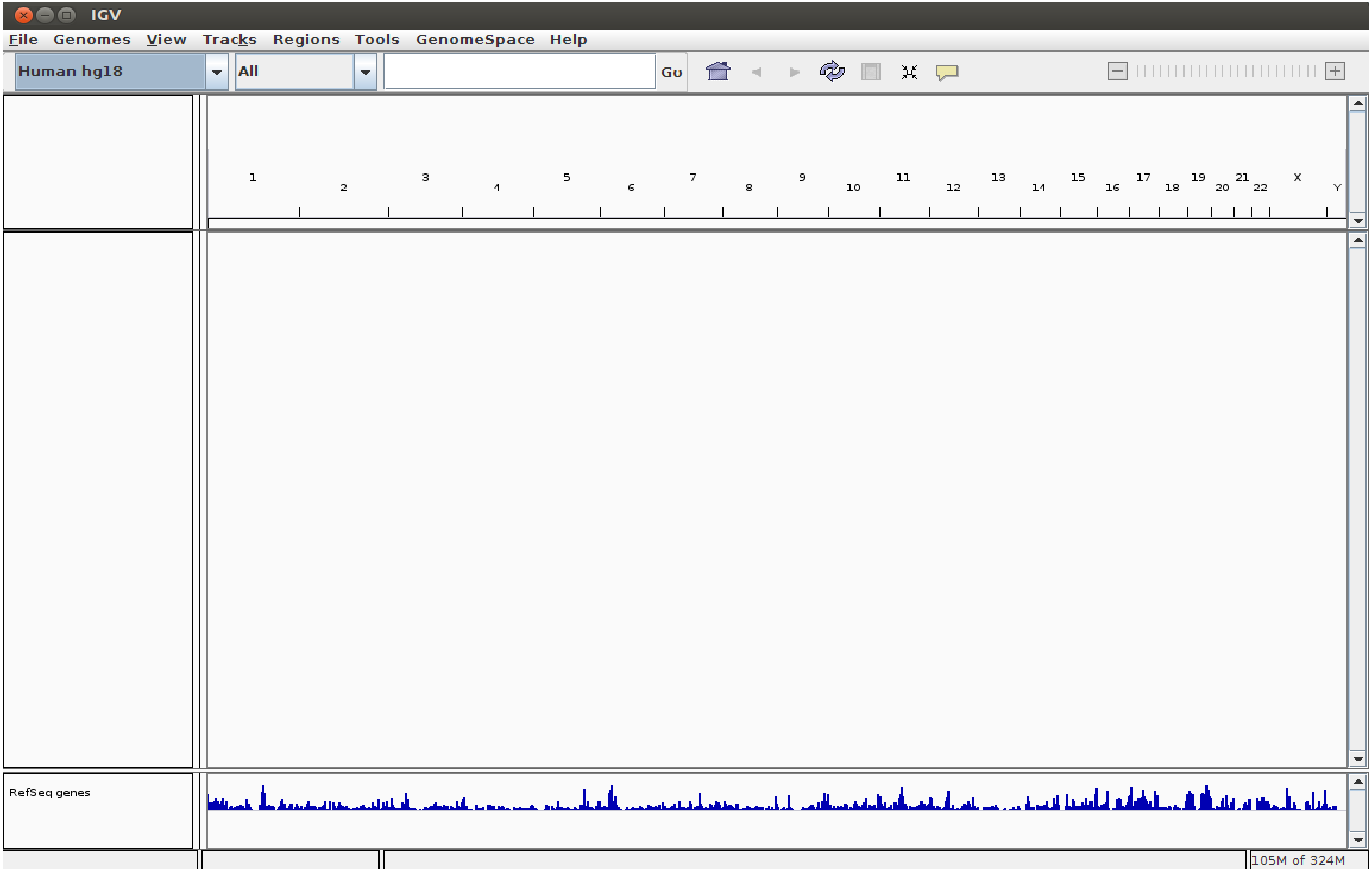
- High-performance visualization tool
- Interactive exploration of large, integrated datasets
- Supports a wide variety of data types
- Documentations
- Developed at the Broad Institute of MIT and Harvard



File Formats

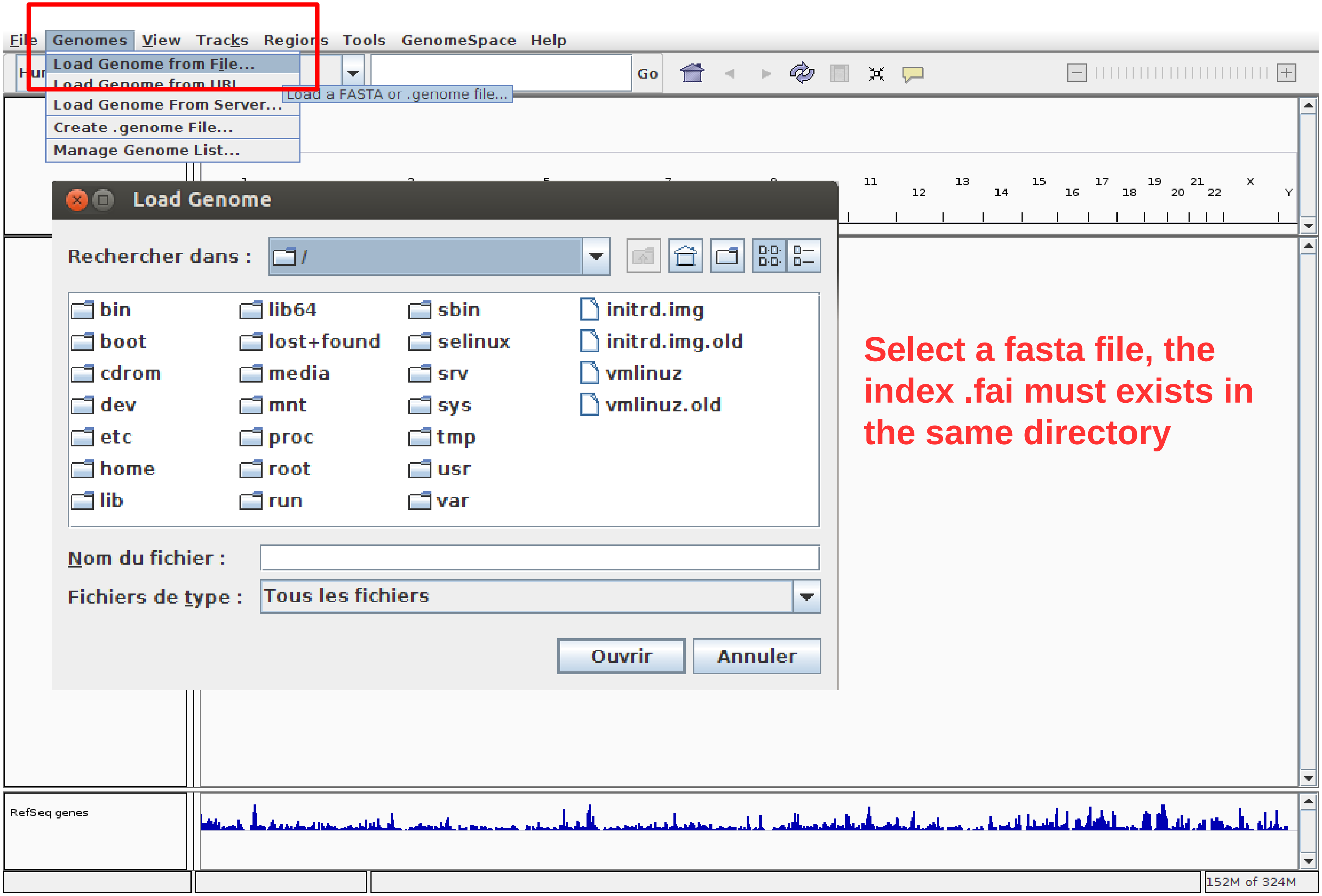
- File Extension Identifies Format
- Recommended File Formats
- BAM
- BED
- CBS
- CN
- Cytoband
- FASTA
- GCT
- genePred
- GFF
- GISTIC
- HDF5
- IGV
- LOH
- Birdsuite Files
- MUT
- RES
- SAM
- Sample Information
- SEG
- SNP
- TAB
- TDF
- Track Line
- Type Line
- WIG

Visualizing alignments on IGV



IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg18
All
Go
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
RefSeq genes
105M of 324M

IGV : Load reference genome

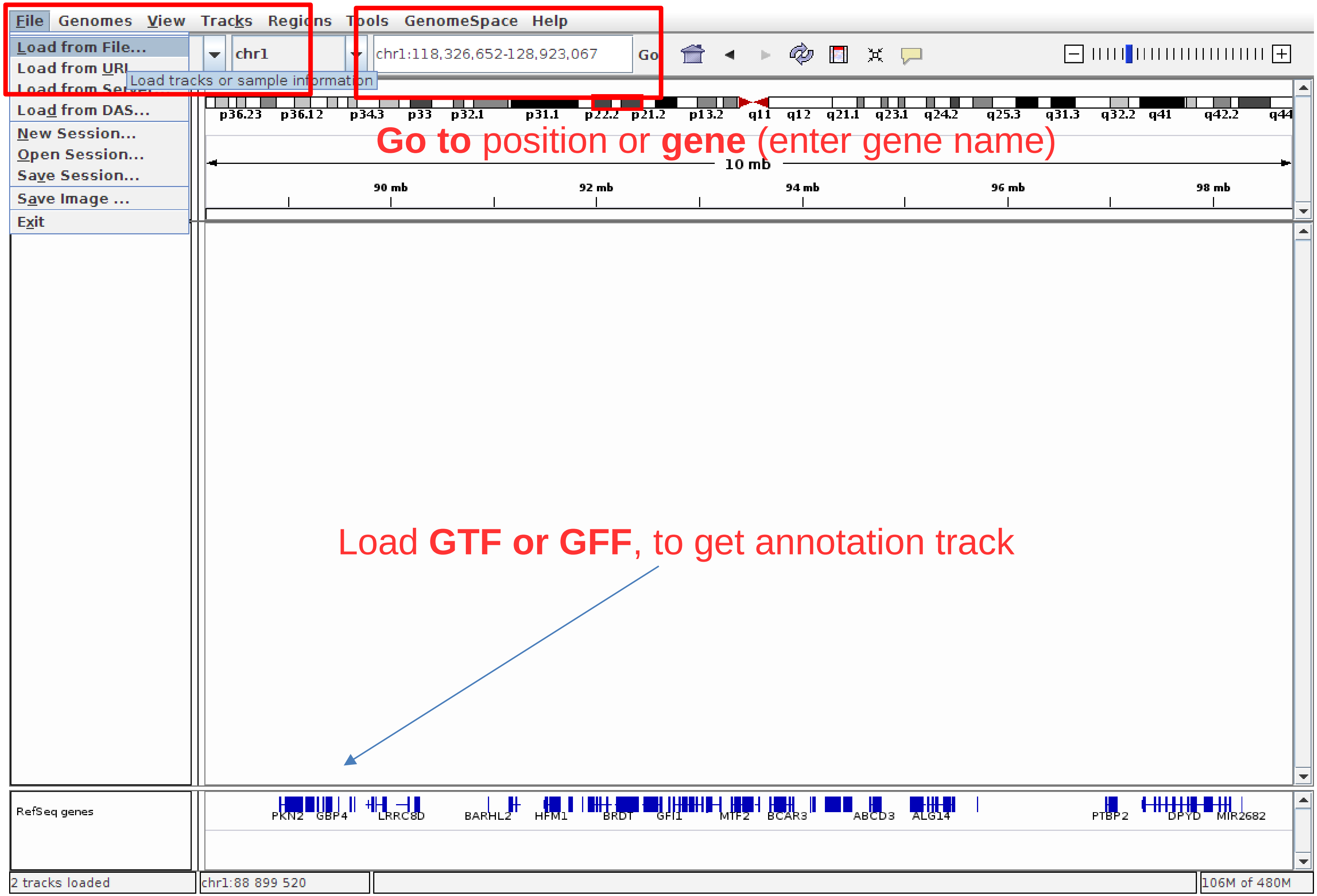


Rechercher dans : /

Fichiers de type : Tous les fichiers

Select a fasta file, the index .fai must exists in the same directory

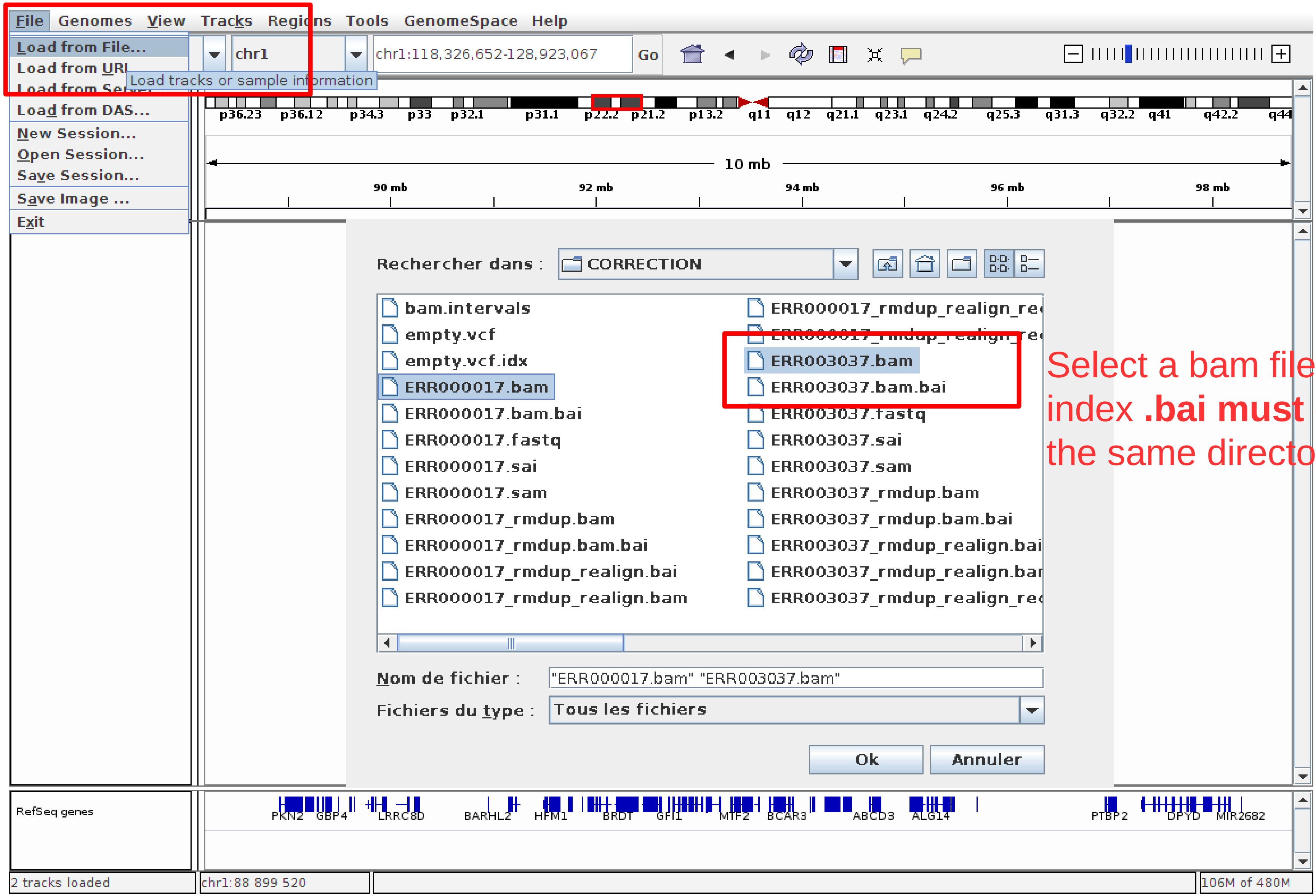
IGV : Load annotation



Go to position or **gene** (enter gene name)

Load **GTF or GFF**, to get annotation track

IGV : Load alignment



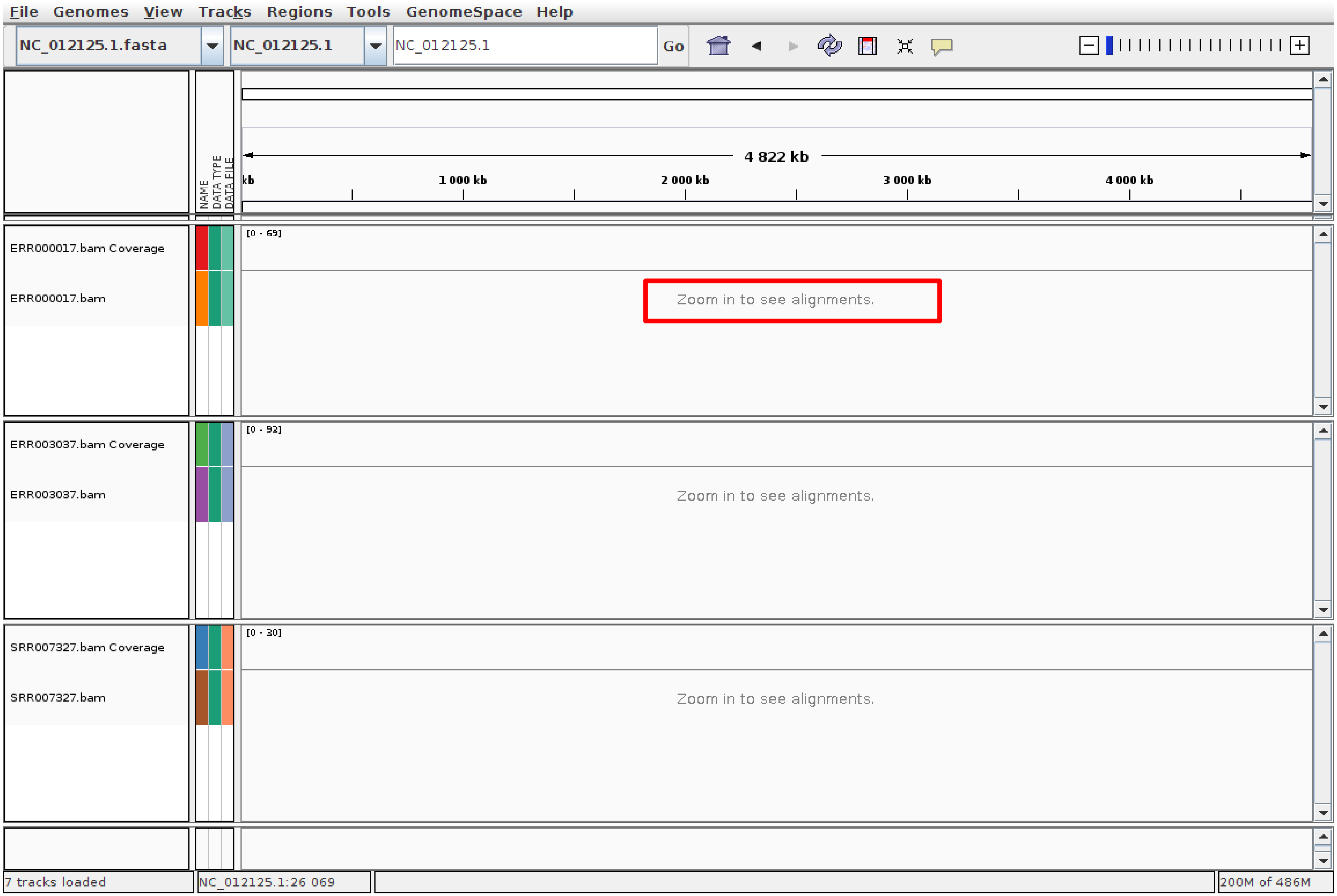
Rechercher dans : CORRECTION

Nom de fichier : "ERR000017.bam" "ERR003037.bam"

Fichiers du type : Tous les fichiers

Select a bam file, the index **.bai must exists** in the same directory

IGV : Load alignment



ERR000017.bam Coverage

ERR000017.bam

Zoom in to see alignments.

ERR003037.bam Coverage

ERR003037.bam

Zoom in to see alignments.

SRR007327.bam Coverage

SRR007327.bam

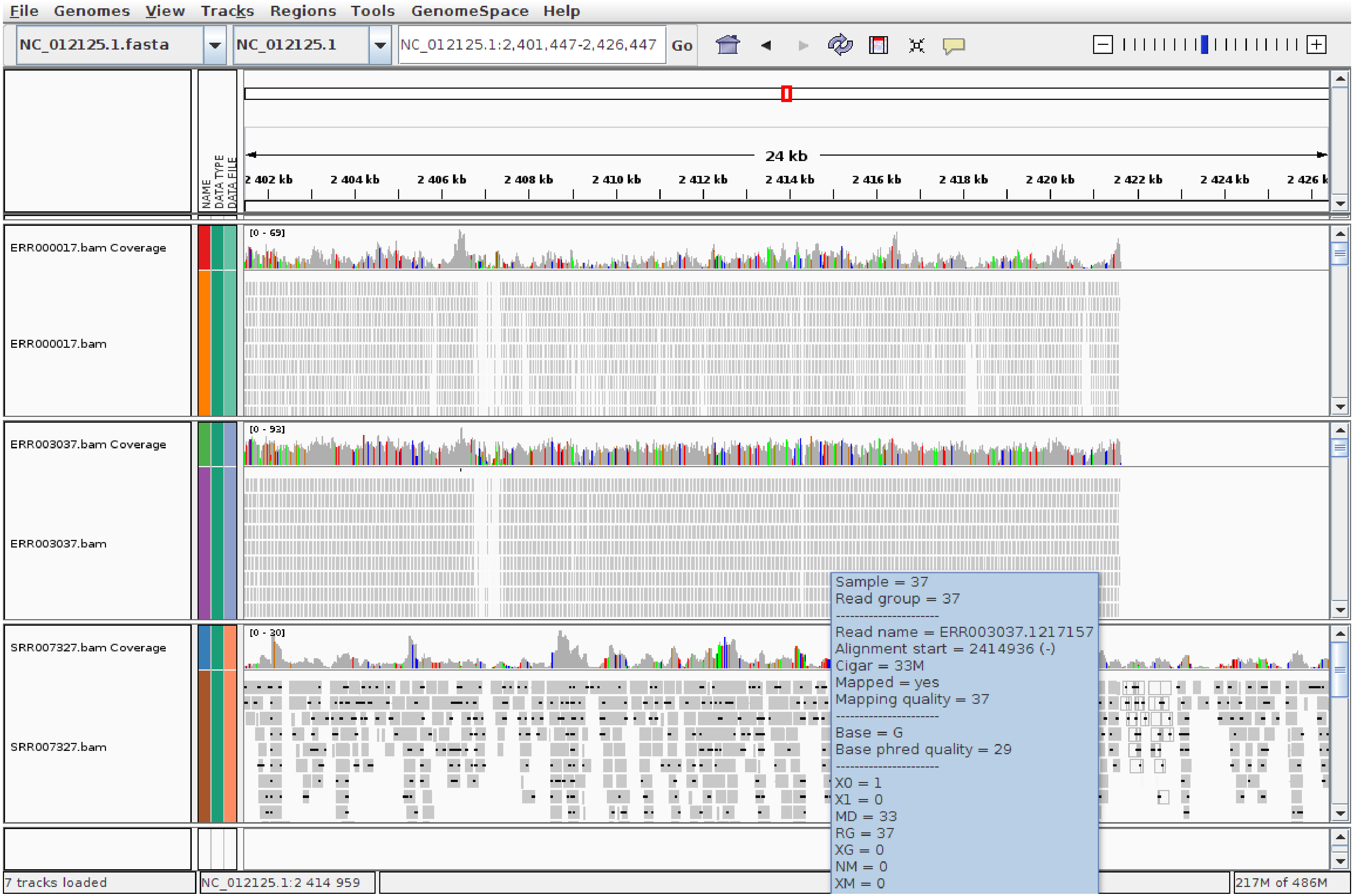
Zoom in to see alignments.

7 tracks loaded

NC_012125.1:26 069

200M of 486M

IGV : Load alignment



Sample = 37
Read group = 37

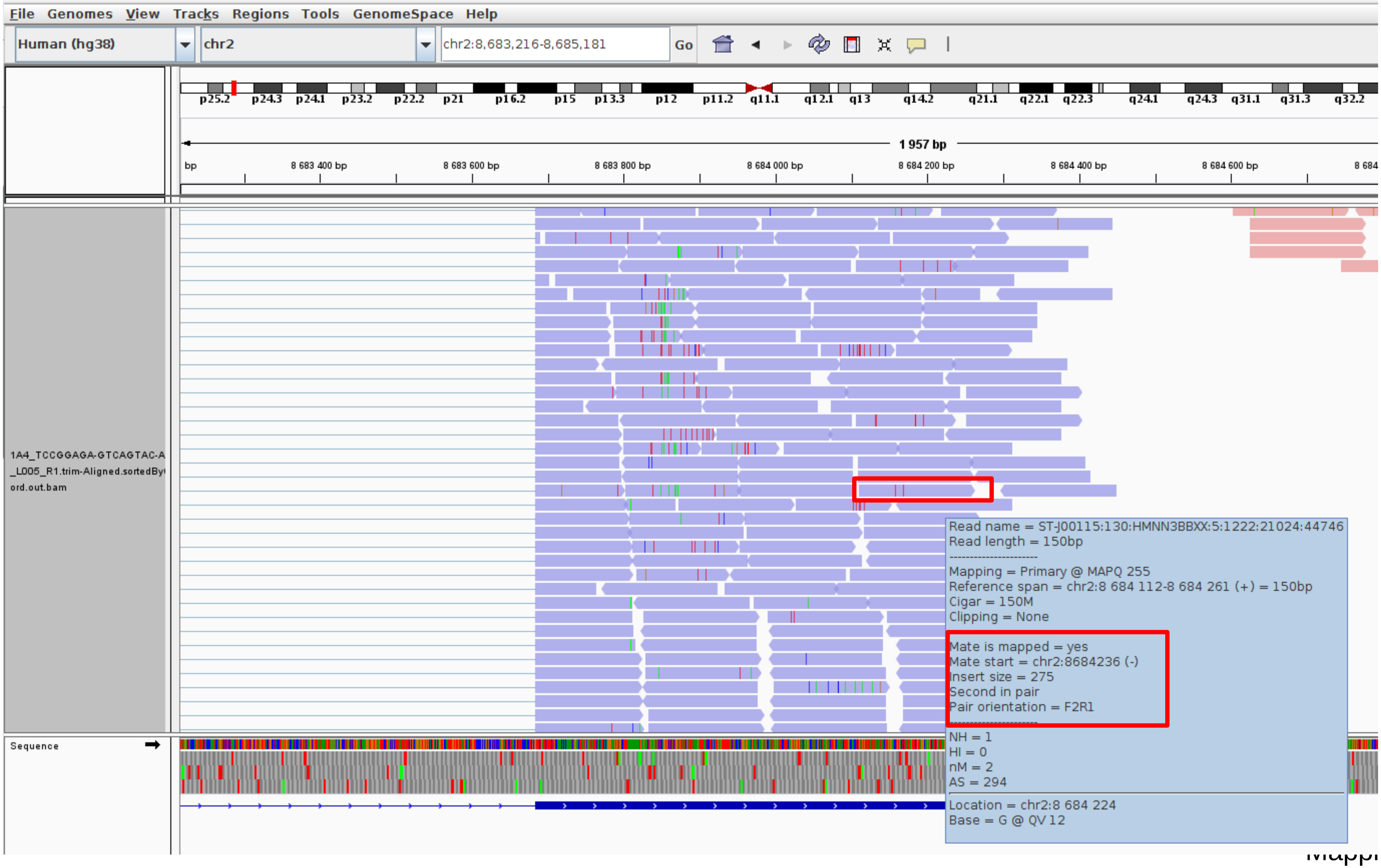
Read name = ERR003037.1217157
Alignment start = 2414936 (-)
Cigar = 33M
Mapped = yes
Mapping quality = 37

Base = G
Base phred quality = 29

X0 = 1
X1 = 0
MD = 33
RG = 37
XG = 0
NM = 0
XM = 0

Find library orientation

Color alignment by > first-of-pair strand


File Genomes View Tracks Regions Tools GenomeSpace Help
Human (hg38)
chr2
chr2:8,683,216-8,685,181
Go
p25.2 p24.3 p24.1 p23.2 p22.2 p21 p16.2 p15 p13.3 p12 p11.2 q11.1 q12.1 q13 q14.2 q21.1 q22.1 q22.3 q24.1 q24.3 q31.1 q31.3 q32.2
1 957 bp
bp 8 683 400 bp 8 683 600 bp 8 683 800 bp 8 684 000 bp 8 684 200 bp 8 684 400 bp 8 684 600 bp 8 684
1A4_TCCGGAGA-GTCAGTAC-A _L005_R1.trim-Aligned.sortedBy ord.out.bam
Sequence

```
Read name = ST-J00115:130:HMNN3BBXX:5:1222:21024:44746
Read length = 150bp
----------------------
Mapping = Primary @ MAPQ 255
Reference span = chr2:8 684 112-8 684 261 (+) = 150bp
Cigar = 150M
Clipping = None

Mate is mapped = yes
Mate start = chr2:8684236 (-)
Insert size = 275
Second in pair
Pair orientation = F2R1
----------------------
NH = 1
HI = 0
nM = 2
AS = 294
Location = chr2:8 684 224
Base = G @ QV 12
```

Exercices 5