

Tuesday – part 2

Mapping RNA-Seq data

Michał Szcześniak, PhD

Faculty of Biology , Adam Mickiewicz University, Poznań
ideas4biology Ltd.

Pipelines

1. FASTQ → QC and filtering → **mapping** → *ab initio* assembly
2. FASTQ → QC and filtering → (**mapping**) → expression estimation → differential expression analysis
3. FASTQ → QC and filtering → *de novo* assembly

Preparing the data

We already have the filtered data

TRIMMED/ERR990413_bbduk2_R1.fastq

TRIMMED/ERR990413_bbduk2_R1.fastq

Removing rRNA-mapping reads

mkdir index

bowtie2-build human_rRNA.fasta index/human_rRNA

bowtie2 -t -p 4 -X 1000 -1 TRIMMED/ERR990413_bbduk2_R1.fastq -2 TRIMMED/ERR990413_bbduk2_R2.fastq -x index/human_rRNA --fast --un-conc ERR990413.fastq > /dev/null

-x: index filename prefix (minus trailing .X.bt2)

--fast: run faster but less precise

--un-conc: write pairs that didn't align concordantly to <path>

-X: maximum fragment length (500)

changing names

mv ERR990413.1.fastq TRIMMED/ERR990413_clean_R1.fastq

mv ERR990413.2.fastq TRIMMED/ERR990413_clean_R2.fastq

Preparing the data

Downloading chr 22

wget

[ftp://ftp.ensembl.org/pub/release-88/fasta/homo_sapiens/dna/
Homo_sapiens.GRCh38.dna.chromosome.22.fa.gz](ftp://ftp.ensembl.org/pub/release-88/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.22.fa.gz)

gunzip Homo_sapiens.GRCh38.dna.chromosome.22.fa.gz

Building an index for chr 22

hisat/hisat-build Homo_sapiens.GRCh38.dna.chromosome.22.fa index/chr22

Preparing a file with known splice sites

wget ftp://ftp.ensembl.org/pub/release-88/gtf/homo_sapiens/
Homo_sapiens.GRCh38.88.gtf.gz

gunzip Homo_sapiens.GRCh38.88.gtf.gz

python hisat/**extract_splice_sites.py** Homo_sapiens.GRCh38.88.gtf >
human_splice_sites.txt

The **extract_splice_sites.py** script is a part of HISAT package.

Mapping

```
hisat/hisat -q -p 4 -X 1000 --time --met-file ERR990413.met_file --phred33  
--rna-strandness RF --known-splicesite-infile human_splice_sites.txt  
--novel-splicesite-outfile novel.splice_sites.txt -x index/chr22  
-1 TRIMMED/ERR990413_clean_R1.fastq -2 TRIMMED/ERR990413_clean_R2.fastq -S  
ERR990413.sam > STATUS/ERR990413_hisat.txt
```

- 1, -2: input data
- S: File for SAM output
- q: query input files are FASTQ .fq/.fastq (default)
- p: number of alignment threads to launch
- X: maximum fragment length (500)
- time: print wall-clock time taken by search phases
- met-file: send metrics to file at <path> (off)
- rna-strandness: Specify strand-specific information (unstranded)
- known-splicesite-infile: provide a list of known splice sites
- novel-splicesite-outfile: report a list of splice sites
- x: Index filename prefix (minus trailing .X.bt2)
- phred33: qualities are Phred+33 (default)

SAM and BAM formats

SAM (Sequence Alignment/Map)

1. Tabular text file
2. Large file sizes
3. Also known as TAM (Text SAM format)

BAM (Binary SAM format)

1. Data compressed with a BGZF algorithm
2. Smaller file sizes
3. Provides fast and easy access to data (e.g. read mappings from region chr1:1,000,000–2,000,000)

Both file types store exactly the same data.

The stored data includes position of read mapping, mapping quality (MAPQ) and status (*concordant / discordant*)

More about SAM:

<https://samtools.github.io/hts-specs/SAMv1.pdf>

Operations on BAM/SAM files

Converting SAM to BAM

```
samtools view -bS ERR990413.sam > ERR990413.bam  
rm ERR990413.sam
```

-bS: to convert SAM to BAM

Correcting the names of paired reads (optional)

```
samtools fixmate -O bam ERR990413.bam ERR990413_fixmate.bam  
rm ERR990413.bam
```

-O: output file name

Sorting BAM files

```
samtools sort ERR990413_fixmate.bam -o ERR990413.sorted.bam  
samtools index ERR990413.sorted.bam
```

Filtering BAM files

All mappings to chr 22

```
samtools view ERR990413.sorted.bam 22
```

All mappings from the given region

```
samtools view ERR990413.sorted.bam 22:16614517-17614517
```

Only display the number of mappings (option -c)

```
samtools view -c ERR990413.sorted.bam 22:16614517-17614517
```

Only return mappings from the provided sets of genomic coordinates in a BED file (option -L FILE)

```
samtools view -c -L RefSeq_chr22.bed ERR990413.sorted.bam 22
```


Visualizing the mappings

```
samtools view -b -h ERR990413.sorted.bam 22:16614517-17614517 >  
ERR990413.fragment.bam  
samtools index ERR990413.fragment.bam # → ERR990413.fragment.bam.bai  
-b    output BAM  
-h    include header in SAM output
```

You can download IGV genome browser from:

http://data.broadinstitute.org/igv/projects/current/igv_mm.jnlp

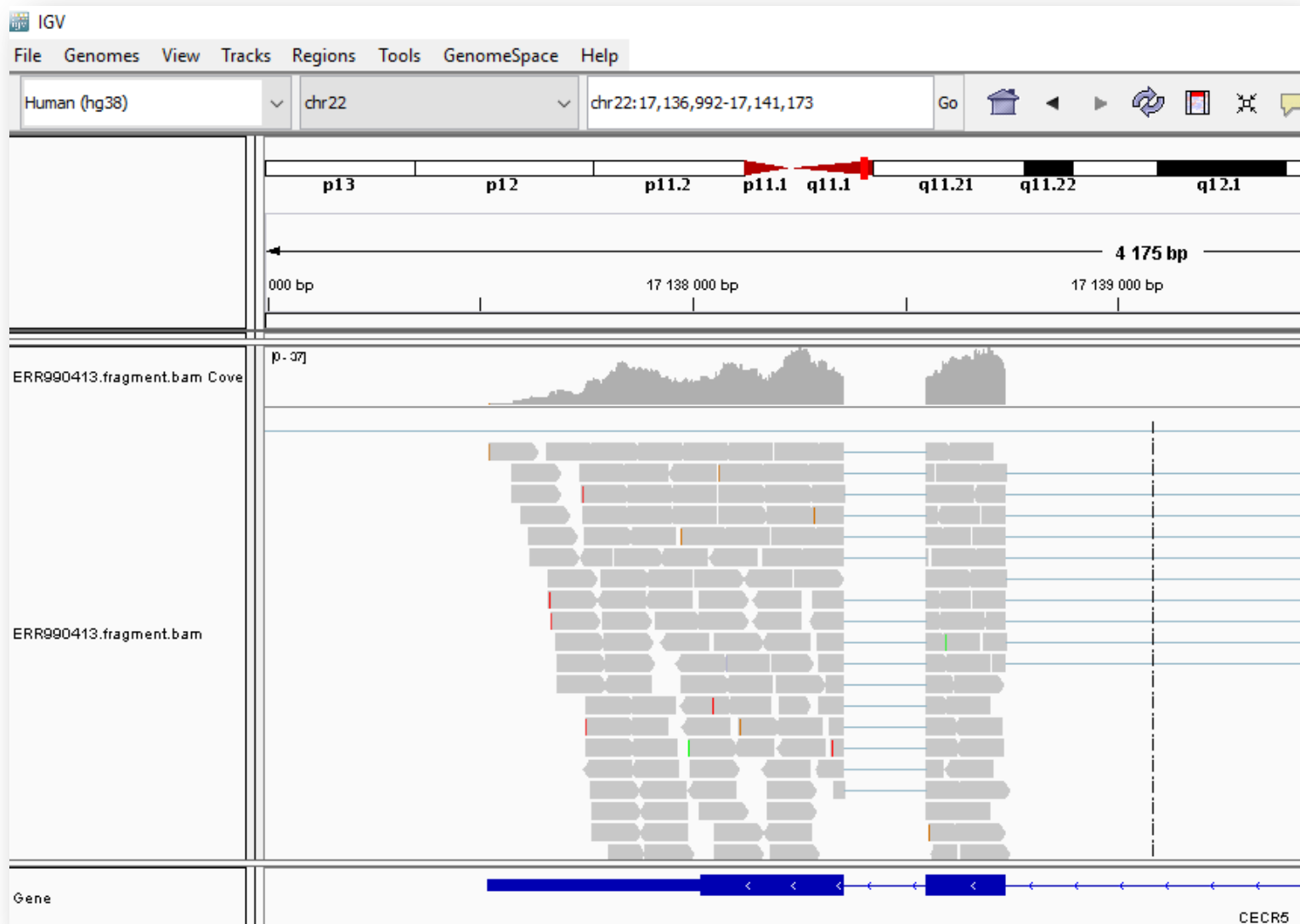
Once IGV is open, choose the genome (hg38) and load a BAM file (the *.bam.bai should be located in the same directory).

Look into the following genomic position: **chr22:17,136,992-17,141,173**

Conventions for data visualization in IGV:

<http://software.broadinstitute.org/software/igv/AlignmentData>

IGV



Reads mapping: quality control

BAM files quality control

(the BAM file should be sorted)

```
mkdir QUALIMAP
```

```
mkdir QUALIMAP/ERR990413/
```

```
qualimap bamqc -bam ERR990413.sorted.bam -nt 2 --java-mem-size=2G  
-outdir QUALIMAP/ERR990413
```

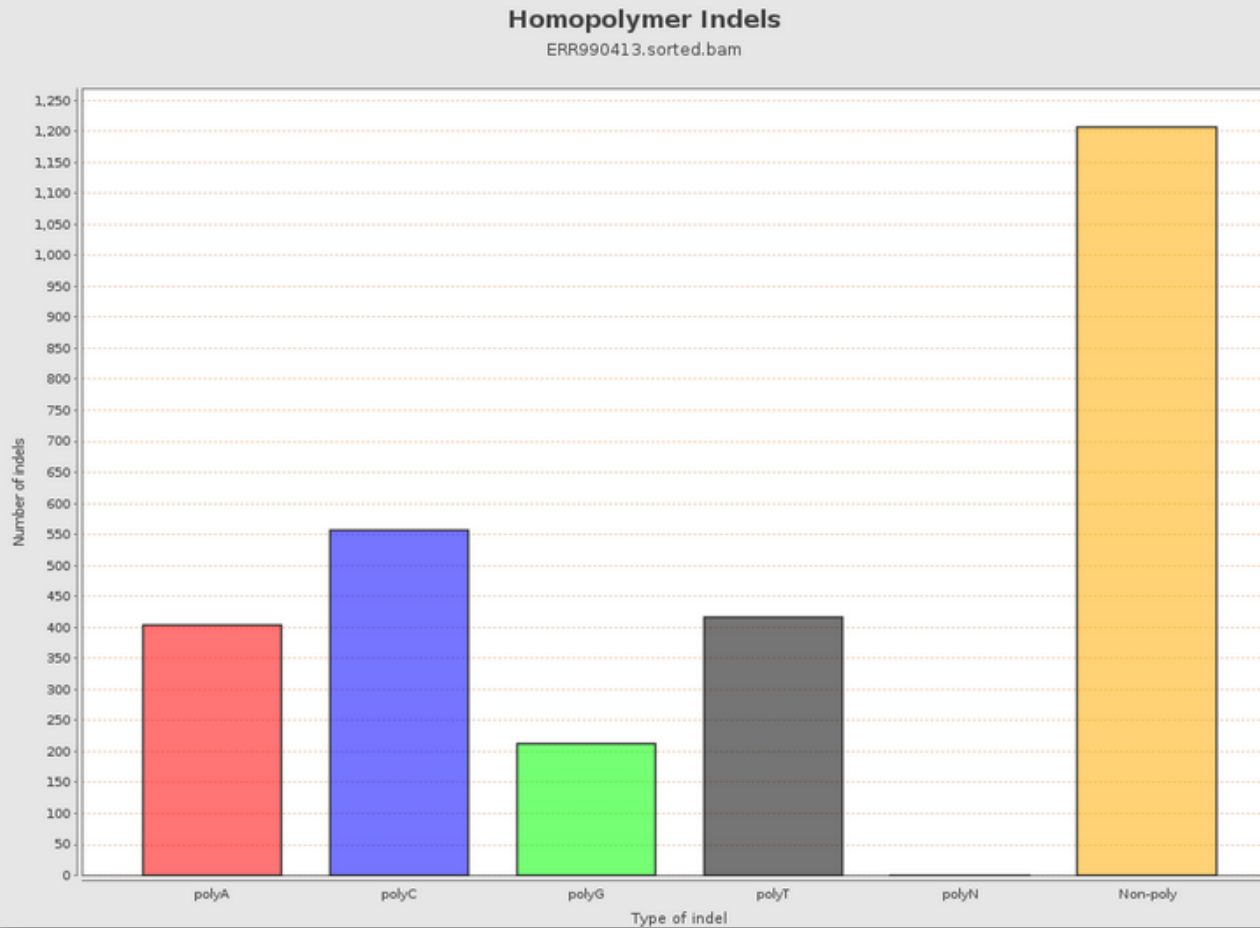
-bam: input BAM file

-nt: numer of threads

--java-mem-size: memory size for Java virtual machine

Qualimap: report

Homopolymer Indels



Mapping smallRNA-Seq data

Mapping

Our clean reads: **SRR1586016_QC.fastq**

building the genome index

bowtie-build Homo_sapiens.GRCh38.dna.chromosome.22.fa index/chr22

mapping

bowtie -t -p 2 -v 2 -a -S -q index/chr22 **SRR1586016_QC.fastq** SRR1586016.sam

-t: write out the time

-p: number of threads

-v 2: two mismatches per read allowed

-a: return all mappings

-S: output is in SAM format

-q: input data is in FASTQ format

Mapping results: postprocessing

converting SAM to BAM

```
samtools view -bS SRR1586016.sam > SRR1586016.bam  
rm SRR1586016.sam
```

sorting the BAM file

```
samtools sort SRR1586016.bam -o SRR1586016.sorted.bam
```

quality control for the BAM file

```
mkdir QUALIMAP/SRR1586016/
```

```
qualimap bamqc -bam SRR1586016.sorted.bam -nt 2 --java-mem-size=2G -outdir  
QUALIMAP/SRR1586016/
```