Hands-on
RNA-seq training session

The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation.
http://www.ebi.ac.uk/ena/

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

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.
http://tophat.cbcb.umd.edu/

Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples. It accepts aligned RNA-Seq reads and assembles the alignments into a parsimonious set of transcripts. Cufflinks then estimates the relative abundances of these transcripts based on how many reads support each one, taking into account biases in library preparation protocols.
http://cufflinks.cbcb.umd.edu/

SAM (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments.
http://samtools.sourceforge.net/

The **Integrative Genomics Viewer (IGV)** is a high-performance visualization tool for interactive exploration of large, integrated datasets. It supports a wide variety of data types including sequence alignments, microarrays, and genomic annotations.
http://www.broadinstitute.org/igv/

Aims:
This training session aims to help you deal with sequences from SGS (Second Generation Sequencing) especially Illumina platforms (GAIIx, HiSeq). You will find the new sequence formats, known biases and will learn to use spliced alignment, genes and transcript discovery and quantification software packages.

Prerequisites : ability to use a Unix environment.

To achieve all of the exercises, log on to your Unix account using "putty" from a MS-Windows PC or ssh from a linux station.

Exercise n°1 : Data Quality

Links:

- EMBL-ENA (European Nucleotide Archive) http://www.ebi.ac.uk/ena/

Study of publicly available data to EMBL ENA:
On this website search for the following entries ERR022486 ERR022488 (in one query if possible).
ERR022486 || ERR022488

• What is the subject of the study, which tissues are being studied?

ERR022486 : RNA from Zebrafish 1day, embryo (31 504 560 reads)
ERR022488 : RNA from Zebrafish 3 days, embryo (24 920 613 reads)

• What type of sequencer has been used?

Illumina Genome Analyzer I I

• What type library sequencing (protocol) has been used?

PAIRED

Abstract: Paired-end sequence data has been generated using polyA selected RNA from a range of zebrafish tissues and developmental stages using the Illumina Genome Analyzer. These data have been used to improve the gene annotation of the zebrafish genome. Study description: Zebrafish total RNA was extracted from embryonic and adult tissue, then polyA selected. After fragmentation and reverse transcription Illumina sequencing libraries were prepared. Paired-end sequence runs were performed with 36, 37, 54 and 76 base reads on the Illumina Genome Analyzer.

• Explore the structure of a study of ENA.

A project is splitied in Study (project information + abstract)→ Sample (the tissu / the studied sample) → Experiment (experimental protocol used, sequencer, libraries preparation...)→ Run (result files)

On you AMI instance, in the training directory create a rnaseq directory and in this one create a two directories : ERR022486 and ERR022488.
mkdir ERR022486 ; mkdir ERR022488

Recover data re-formatted for the study of chromosome 22 from the web page of the course:
http://bioinfo.genotoul.fr/index.php?id=119

You can download the fastq files directly to your account using the "wget"command, remember to place it in the appropriate directory on your AMI:
wget http://snp.toulouse.inra.fr/~contig/F11d/ERR022486_chr22_read1.fastq.gz

Data quality analysis:

- Using FastQC in command line (to avoid overloading the server!)
- Launch FastQC to analyze the quality of the reads without direct visualization, but through the creation of output files:

fastqc -nogroup ERR022486_chr22_read1.fastq&

NB. If you use the command line to run fastqc, you can use the option -nogroup in order not to group de values for base pair positions after 10.

Analyze the results:

If you have launch fastqc from the command line in you AMI you have to download the result files to you PC to visualize them. If you want to run fastqc on your local PC , you must install FastQC on your computer.

- Download and install FastQC:

http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/fastqc_v0.10.1.zip

-instructions available here:

http://www.bioinformatics.bbsrc.ac.uk/projects/download.html#fastqc

- Load fastq (it is possible to upload multiple files at once) You can save a report for each analysis using the corresponding option in the file menu.

What is the length of the reads?
What is the quality of reads?
Look through the results for graphics described during the course
basic statistics
per base sequence quality
per base sequence content
kmer content

Exercise # 2: alignment / visualization

Some links:

- Tophat: http://tophat.cbcb.umd.edu/
- Samtools: http://samtools.sourceforge.net/
- Bowtie: http://bowtie-bio.sourceforge.net/index.shtml
- FTP download from Ensembl: http://www.ensembl.org/info/data/ftp/index.html

Alignement

Search for the Ensembl zebrafish gtf file on FTP site. Download the file and indicate how many lines are in the file.

(complet reference genome ftp://ftp.ensembl.org/pub/release-62/gtf/danio_rerio/Danio_rerio.Zv9.62.gtf.gz)

ftp://ftp.ensembl.org/pub/release-62/fasta/danio_rerio/dna/Danio_rerio.Zv9.62.dna.chromosome.22.fa.gz

For the remaining exercises we are going to work on a reduced files you can get form the training page on the web.

Download the sequence of chromosome 22 associated :

http://snp.toulouse.inra.fr/~contig/F11d/Danio_rerio.Zv9.62.dna.chromosome.22.fa.gz

Download also retrieve the gtf containing only chromosome 22 :

http://snp.toulouse.inra.fr/~contig/F11d/Danio_rerio_chr22.Zv9.62.gtf.gz

Create a directory (mkdir ./bowtie-index) and generate the index with bowtie using the chromosome

mkdir bowtie-index; cd bowtie-index

bowtie-build Danio_rerio.Zv9.62.dna.chromosome.22.fa Danio_rerio.Zv9.62_chr22

If you use the latest version of tophat you can index the file with bowtie2 (bowtie2-index).

some indexes for some organisms are already built and available at: ftp://ftp.cbcb.umd.edu/pub/data/bowtie_indexes/

There are two possible approaches:

- alignment with the basic parameters of tophat (without reference transcriptome)
- alignment with GTF to help tophat find the known junctions.

Today we will focus on alignment without reference transcriptome (GTF file of chromosome 22). In order to verify that tophat is able to find known junctions. The read fastq files can be downloaded from : http://bioinfo.genotoul.fr/index.php?id=119 Raw data files section.

- Create a directorie for each dataset to store tophat analysis results (eg tophat_ERR022486).

- Run tophat on each data sets with an insert size of 200bp and the previously created directory to store the produced files.

```
tophat –output-dir=tophat_ERR022486 -r 200 -G Danio_rerio_chr22.Zv9.62.gtf bowtie-index/Danio_rerio.Zv9.62_chr22 ERR022486_chr22_read1.fastq ERR022486_chr22_read2.fastq
```

to speed up tophat you can decrease the -I parmeter to limit possible intron size.

- Index the resulting bam file with samtools (samtools index) to obtain the corresponding bai file.

```
samtools index tophat_ERR022486 /accepted_hits_ERR022486_chr22.bam
```

- How many reads are uniquely aligned (using the commands cut, sort, uniq)?

```
samtools view accepted_hits_ERR022486_chr22.bam | cut -f 1 | sort | uniq -c
```

- Use "samtools flagstat" to produce the statistics of alignment.

```
samtools flagstat accepted_hits_ERR022486_chr22.bam
```

Visualization of results:

- Use IGV to visualize the resulting alignment.

- Start IGV in your web-browser using the links at the bottom of page : http://bioinfo.genotoul.fr/index.php?id=119

- The zebrafish genome is already integrated into IGV (zebrafish Zv9) but you can also create your own genome using the chromosome 22 fasta file. You can also load the gtf correspondence chromosome 22 as a file.

- Load the bam and bed files one by one to be able to rename files bed and bam output by adding the name of the sample from the IGV Interface (right-click on the name "accepted_hits .....").

- Explore the interface using right-click to expand or colapse features (to view all isoforms, the pairs of associated reads of the same fragment ....)

- Look at the areas shown in the course as well as the following regions:

chr22 :586,901596, 104 is this a new exon?

Chr22 :671,043680, 246 exon boundaries

chr22 :20,729,86920 739.072: UTR

chr22 :24,314,52324 350.044: isoforms

chr22 :25,962,02825 970.244: new transcript?

Be aware that the names of the chromosomes in your fasta genome file could be different from the genome included in IGV and/or your gtf file. In this case, the GTF file elements will not be shown in IGV. The genome and the files opened in IGV have to have the same chromosome names to be shown together.

If it is the case, you can use a mapping file, for example: Zv9_alias.tab. You have to put it in the IGV genome directory of your PC: /home/.... /igv/genomes X:\...\igv\genomes A description on how-to create an alias file can be found at : http://www.broadinstitute.org/software/igv/LoadData/#aliasfile This file must contain the correspondence between the names used by IGV and the names of your GTF:

1 chr1

2 CHR2

3 CHR3

4 CHR4

5 CHR5 ... .

Exercise # 3: assembly, discovery of new elements and differential expression

Handling the GTF file, familiarize yourself with its reference. Using thee reference chromosome 22 Danio_rerio_chr22.Zv9.62.gtf file, find out :

- How many genes are described in the file (column 9 using cut, cut by ";", sort and uniq)?

```
cat Danio_rerio_chr22.Zv9.62.gtf | cut -f9 | cut -d";" -f1 | sort -u | wc -l
```

réponse : 1276

- How many transcripts are described in the file?

```
cat Danio_rerio_chr22.Zv9.62.gtf | cut -f9 | cut -d";" -f2 | sort -u | wc -l
```

réponse : 2133

Transcripts assembly using a GTF reference:

- Create a directorie for each dataset to store cufflinks analysis results (eg cufflinks_ERR022486)

```
mkdir cufflinks_ERR022486; mkdir cufflinks_ERR022488
```

- Run cufflinks (or sigcufflinks) on each data sets using the GTF description file and the previously created directory to store the produced files. Cufflinks input file is the accepted_hits.bam file produced by tophat.

```
cufflinks –output-dir=/work/.../cufflinks_ERR022486 -g Danio_rerio_chr22.Zv9.62.gtf tophat_ERR022486/accepted_hits_ERR022486.bam
sigcufflinks_ERR022486 -g Danio_rerio_chr22.Zv9.62.gtf tophat_ERR022486/accepted_hits_ERR022486.bam
```

-

Analysis of results cufflinks (or sigcufflinks)

Open the output file transcrits.gtf (which describes each transcript assembled) with a text editor or a spreadsheet:

- Can you tell if a transcript has multiple exons? Watch column 9.

gene_id "CUFF.3"; transcript_id "CUFF.3.1"; FPKM "115.8021820938"; frac "0.772636"; conf_lo "83.849676"; conf_hi "147.754688"; cov "10.344032"; full_read_support "yes";
gene_id "CUFF.3"; transcript_id "CUFF.3.1"; exon_number "1"; FPKM "115.8021820938"; frac "0.772636"; conf_lo "83.849676"; conf_hi "147.754688"; cov "10.344032";
gene_id "CUFF.3"; transcript_id "CUFF.3.1"; exon_number "2"; FPKM "115.8021820938"; frac "0.772636"; conf_lo "83.849676"; conf_hi "147.754688"; cov "10.344032";
gene_id "CUFF.3"; transcript_id "CUFF.3.1"; exon_number "3"; FPKM "115.8021820938"; frac "0.772636"; conf_lo "83.849676"; conf_hi "147.754688"; cov "10.344032";
gene_id "CUFF.3"; transcript_id "CUFF.3.1"; exon_number "4"; FPKM "115.8021820938"; frac "0.772636"; conf_lo "83.849676"; conf_hi "147.754688"; cov "10.344032";
_id "CUFF.3"; transcript_id "CUFF.3.1"; exon_number "5"; FPKM "115.8021820938"; frac "0.772636"; conf_lo "83.849676"; conf_hi "147.754688"; cov "10.344032";
gene_id "CUFF.3"; transcript_id "CUFF.3.1"; exon_number "6"; FPKM "115.8021820938"; frac "0.772636"; conf_lo "83.849676"; conf_hi "147.754688"; cov "10.344032";
gene_id "CUFF.4"; transcript_id "CUFF.4.1"; FPKM "99.4095500676"; frac "0.227364"; conf_lo "48.845404"; conf_hi "149.973696"; cov "8.879760"; full_read_support "yes";

- Can we know if there are multiple transcripts for a given locus (several isoforms for a gene given)?

Multiple cuff.x.1 cuff.x.2..... that we can find in isoforms.fpkm.tracking as well

- How many genes a in the GTF (is the figure different for the reference file)?

on genes.fpkm.tracking
cat genes_ERR022486.fpkm_tracking | cut -f1 | sort -u | wc -l
réponse : 1933 pour ERR022486

- How many transcripts a in the GTF (is the figure different for the reference file)?

on isoforms_ERR022486.fpkm_tracking
cat isoforms_ERR022486.fpkm_tracking | cut -f1 | sort -u | wc -l
réponse :3489 pour ERR022486:
ou :
on transcripts_ERR022486.gtf
cat transcripts_ERR022486.gtf | cut -f9 | cut -d";" -f2 | sort -u | wc -l
réponse :3489 pour ERR022486

- Download and load the transcripts.gtf file in IGV to view the examples presented in the alignment section of this hands-on.

- Use cuffcompare to compare the reference gtf and assembly of cufflinks from ERR022486.

```
cuffcompare -o cuffcompare_refVSERR022486 -r Danio_rerio_chr22.Zv9.62.gtf
cufflinks_ERR022486/transcripts_ERR022486.gtf
```

- Look at the connections between genes and those of cufflinks GTF, ditto for isoforms: Watch cuffcompare.tracking files, class code column and cuffcompare.combined.gtf

- Download and view Cuffcompare.combined.gtf in IGV.

- Run the script cuffcompare_quant_summary.pl: Explore files results.

```
cuffcompare_quant_summary.pl -p quant_86_88 -o cuffcompare_quantSummary -r
Danio_rerio_chr22.Zv9.62.gtf --tracking cuffcompare_sigcufflinks/cuffcompare_86_88.tracking
sigcufflinks_ERR022486/ERR022486_transcripts.gtf
sigcufflinks_ERR022488/ERR022488_transcripts.gtf
```