

# Hands-on short reads alignment and SNP discovery



Babraham Bioinformatics

FastQC

“**FastQC** is a quality control tool for high throughput sequence data.”

<http://www.bioinformatics.bbsrc.ac.uk/>

BWA

“**Burrows-Wheeler Aligner (BWA)** is an efficient program that aligns relatively short nucleotide sequences against a long reference sequence such as the human genome.” <http://bio-bwa.sourceforge.net>

SAMtools

“**SAM** (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments.”

<http://samtools.sourceforge.net>



Integrative  
Genomics  
Viewer

“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 aims to help you process sequences produced by NGS sequencers (platforms Illumina Solexa and Roche 454). You will learn how to check sequence quality, align reads versus a reference genome, visualize the corresponding alignment and call SNPs and INDELs (Single Nucleotide Polymorphisms and Insertions/Deletions).

**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.

First enter the training directory and create a new F12c directory:

```
cd training; mkdir F12c; cd F12c
```

## Exercise # 1: Quality analysis

### Some links:

NCBI: <http://www.ncbi.nlm.nih.gov>

EBI EMBL : <http://www.ebi.ac.uk/embl/index.html>

FastQC: <http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc>

### NCBI data retrieval:

On the NCBI site find entries corresponding to the identifiers "SRX002048", "ERR003037" and "ERR000017" (in a single query, if possible).

What type of sequencer has been used to produce each of these three data sets?

Download the three "sra-lite" as follows: "SRR007327", "ERR003037" and "ERR000017"

**NB.** To optimize disk space, it is not possible to directly access the file "fastq" at the NCBI. They have established the "sra" format. It is possible to find the same data sets at the EBI/EMBL in fastq.gz format.

### Handling "sra", then "fastq" using the sra toolkit :

Convert three files "sra" in "fastq"

Conversion into "fastq" a file "sra" is performed with the following command (available on "snp" through the "sra toolkit"):

```
fastq-dump [...]. sra
```

**NB :** if the sra toolkit is not installed you can also download the fastq.gz files from the EBI EMBL website.

- What is the number of reads per file "fastq" ?
- What is the number of reads containing one or more 'N' for the file ERR000017 ?
- What is the number of reads containing zero, one, two, ... N for the file ERR000017 ?

### Statistics with FastQC:

Launch "fastqc"

If you run fastqc on your local PC, open the three "fastq" and navigate the interface.

## Exercise # 2: Sequence alignment

### Some links:

BWA: <http://bio-bwa.sourceforge.net>

BWA man: <http://bio-bwa.sourceforge.net/bwa.shtml>

### Alignment readings with "BWA"

Retrieving the "NC\_012125.1" reference sequence from the NCBI web site.

**NB.** The steps for viewing "text fasta" and able to make a 'Ctrl + A' Ctrl + C ':

- "Display" Genbank
- "Display" FASTA

- "Sent to" Text

### Getting familiar with the bwa software package :

Display help to see the available commands

```
bwa
```

Display help for commands display a

```
bwa index
```

Index the reference sequence (see course slides).

Perform alignment (see course slides).

Check the data type (read length) before launching the alignment :

- bwa aln "for short queries (~ 200bp) with low error rate (<3%)". Which algorithm are you going to use?
- bwa bwsw "for long reads with more errors"

Produce SAM alignments performed with "bwa aln" (see course slides).

## Exercise # 3: Formats, conversions and manipulations

### Some links:

Samtools: <http://samtools.sourceforge.net>

Picard: <http://picard.sourceforge.net>

For each SAM file produced with BWA (Solexa: ERR000017.sam - ERR003037.sam and 454 SRR007327.sam)

- View the first three lines of each SAM in a terminal and identify the various fields
- What are the different "flags" contained in one of the three SAM file?
- What do they mean (<http://picard.sourceforge.net/explain-flags.html>)?
- How many reads are present in the ERR000017 set?
- How many reads of ERR000017 were not aligned by BWA?
- How many reads of ERR000017 have a "36M" cigarline?
- How many reads of ERR000017 are perfectly aligned?
- Why are the answers to the last two questions different (reminder: the reads length of ERR000017 is 36 bp)?
- What is the maximum "mapping quality" in ERR000017 alignment?
- How many reads have this quality value?

### Using samtools:

Getting familiar with the samtools software package :

- Display help to see the available commands  
samtools
- View using one of the commands  
samtools view
- Convert files SAM BAM.

- Show the first lines of the created file BAM.
- Part of the SAM file is missing, why? Modify the above command to display it?
- Sort the BAM file.
- Merge the two BAM files ERR003037 ERR000017
- Index the sorted BAM.

## **Exercise # 4: Views**

### **Some links:**

Samtools tview: <http://samtools.sourceforge.net/tview.shtml>

Interactive Genomics Viewer - IGV: <http://www.broadinstitute.org/igv>

### **Interactive Genomics Viewer - IGV:**

Go to the following url: <http://www.broadinstitute.org/igv/log-in>

### **Two ways to launch IGV:**

1. Download the software package on your PC, uncompress it and execute it by double clicking on the file igv\_win.bat (Note: it is possible to change the memory allocated by editing this file: Xmx1g-example)
2. Launch "webstart"
  - Import the reference genome (at this stage it is necessary to have the file "NC\_012125.1.fasta" on your local PC).
  - Use the "igvtools" menu to produce the coverage file the coverage (this step is optional, but allows you to view coverage when zooming very large)
    - Menu "File" => "Run ... igvtools "
    - Then import the generated tdf file.
  - For the Solexa data sets "merged" and the 454 data set:
    - Load "sorted.bam"
    - Load the corresponding tdf file (coverage file : in blue usually)
  - Explore interface (slide, zoom, label information, right click, ...)
  - load the GFF annotations file.
  - On the NCBI web site search for "NC\_012125.1" .
    - Click on the identifier to display the detailed view. Then click in the column "links" to "RefSeq FTP" and copy link address "NC\_012125.gff" to download the file with "wget".
    - `ftp://ftp.ncbi.nih.gov/genomes/ wget [...] / NC_012125.gff`
  - Load the annotation GFF file.
  - Locate and mark a few regions of interest. Export these regions, erase and reload.
  - Test sessions:
    - Save your session
    - Delete all tracks
    - Reload your saved session using "Open Session".

## Exercise # 5: SNP / indel discovery

### Some links:

Samtools pileup size: <http://samtools.sourceforge.net/mpileup.shtml>

VCF format: <http://vcftools.sourceforge.net/specs.html>

### Samtools mpileup:

Using the "mpileup" command of samtools on the 454 file :

- Produce the "pileup" file .
- How many lines has the generated file? Why, knowing the length of the genome which is 4,833,080 bp?
- Perform SNP calling (produce the vcf file) .

Using the command "vcfutils.pl" filter "vcf" and building "consensus"

- Display help to see the available commands  
`vcfutils.pl`
- Display using the command "varFilter"  
`vcfutils.pl varFilter`

For this same 454 dataset, filter the "vcf" file with the default options .

Build the consensus corresponding to the sequences in the alignment.

### Add the variant track in IGV:

IGV does not recognize the format "pileup" it is necessary to convert the file to GFF . For this use the following line.

```
perl-lane '{next if ($_ =~ /^#/); $cmp++ print "$ F [0]\  
tmpileup\t SNP_$cmp\t $F[1]\t $F[1]\t .\t .\t .\t id = $F[0]"}'  
'raw.f.vcf'> raw.f.gff
```

Load the produced gff file in IGV.

It is possible to move in the reference by "jumping" from "feature" to "feature":

- Click on the track that contains the "features"
- Move with 'Ctrl + F' (below) and 'Ctrl + B' (previous)