Exome-Seq Analysis using GALAXY (95%)

(also answer for the session's exercise)

Course on Methods and tools to analyze Next Generation Sequencing data
(IRB Lleida)

Carles Hernandez-Ferrer and Juan R González
BRGE - Bioinformatics Research Group in Epidemiology
Center for Research in Environmental Epidemiology (CREAL)
`http://www.creal.cat/jrgonzalez/software.htm`

Contents

Introduction

The steps we will follow to extract the annotated variants from the original `.fastq` file are:

1. Quality Control
2. Alignment (BWA)
3. Mark PCR Duplicates
4. Local Realignments Around Indels
5. Quality Recalibration
6. SNP calling
7. Annotation

1 Quality Control

First, we need to upload the `.fastq` file stored in `material/exercise` called `qG13006451.chr22.fastq`.

We do that with the tool `Upload File` in the group `Get Data`. We set the option "`File Format:`" as `fastqillumina` and we us the "`File:`" to select the file we want to upload (`qG13006451.chr22.fastq`).

Having the `.fastq` uploaded into Galaxy we can use the tool `FastQC`, into the `NGS: QC and manipulation` group. But first we need to convert the `.fastq` from *illumina* to *sanger* format.
To do that, we will use the `FASTQ Groomer` from the same groupNGS: QC and manipulation, filling the option "`File to groom:`" with the uploaded `.fastq` file (`qG13006451.chr22.fastq`) and the option "`Input FASTQ quality scores type:`" with "*`Sanger`*".

Having a new `.fastq` file (`qG13006451.chr22.fastq - sanger`) we can run the `FastQC`. This tool generates a `.html` report. For this exercise, we must take care of the section entitled "Per base sequence quality", seen in Figure 1.

The quality of the reads, for this case, is really good so no filtering it is needed. Remember that the per base quality report we obtained for the file `qG13006290` was not as good as for `qG13006451`. And a `Filter by quality` (from the `NGS: QC and manipulation` group) was required. We can see it in Figure 2.

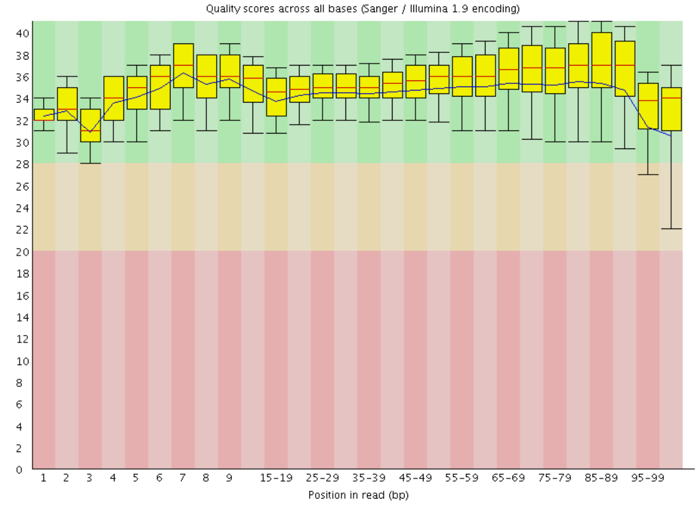

Figure 1: Per base sequence quality report from `FastQC` for `qG13006451`.

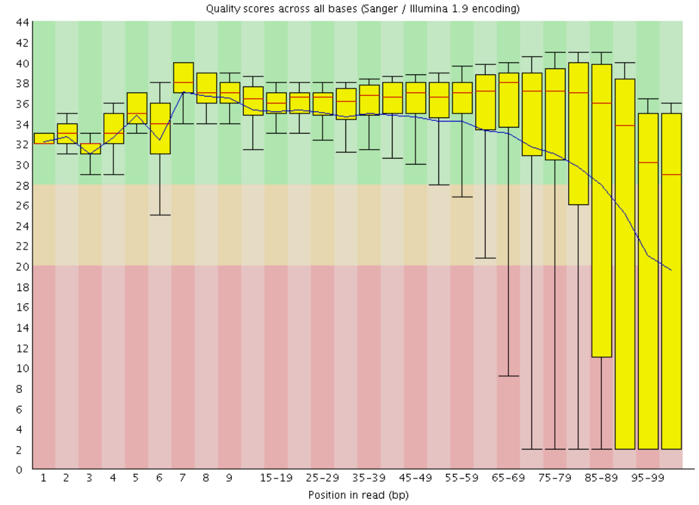

Figure 2: Per base sequence quality report from `FastQC` for `qG13006290`.

2 Alignment (BWA)

For this step, the *Alignment*, we need the genome reference file. So, we upload all the files for `GRCh37.73.dna.22` using the ToolUpload File from `Get Data`:

1. GRCh37.73.dna.22.fa
2. GRCh37.73.dna.22.fa.fai
3. GRCh37.73.dna.22.dict

One the reference is uploaded we can start the alignment using the BWA algorithm. For that fill the tool `Map with BWA for Illumina` from `NGS: Mapping` with:

- "`... reference genome from your history ...?:`" with "*`Use one from the history`*"
- "`Select a reference from history:`" with "*`GRCh37.73.dna.22.fa`*"
- "`Is this library mate-paired?:`" with "*`Single-end`*"
- "`FASTQ file:`" with "*`qG13006451.chr22.fastq - sanger`*"

The tool `Map with BWA for Illumina` has generate a `.bam` file. Now we will add the header of this file using the tool `Add or Replace Groups` from `NGS: Picard (beta)`. In the case you where studding trios, here where you should specify which sample corresponds to each individual.

As this is a simple test just fill the options with:

- "`SAM/BAM dataset to add or ... in:`" with "*`qG13006451.chr22 - align`*"
- "`ID tag`" with "*`IRB-L1`*"
- "`SM tag`" with "*`qG13006451`*"
- "`LB tag`" with "*`hg19`*"
- "`PL tag`" with "*`ILLUMINA`*"
- "`Read group platform unit`" with "*`run barcode`*"

The next step is to sort the reads after being aligned. For that we will use the tool `SortSam` from the Picard Tools, but this tool is not available in the `NGS: Picard (beta)` group from Galaxy. So to get the reads aligned and ordered we will convert the `.bam` file to a `.sam` file using the tool `BAM-to-SAM` from `NGS: SAM Tools` (remember to check the option "`Include header in output`") and download the resulting file (in our case `qG13006451 - Align SAM`).

Having the file (`qG13006451.align.sam`) in the desktop we will use the Picard Tools to sort the reads. At the same time we will convert the `.sam` file to a `.bam` file. This will be done as:

```
java -Xmx4g -jar ~/Software/picard/SortSam.jar \
    I=~/Desktop/exercise/qG13006451.align.sam \
    O=~/Desktop/exercise/qG13006451.align.sort.bam \
    SO=coordinate
```

Finally we upload the generated file (`qG13006451.align.sort.bam`) into Galaxy.

3 Mark PCR Duplicates

This step is really easy since Galaxy has the perfect tool to mark the (reads) PCR duplicates, results of the amplification biases in PCR.
Use the tool `Mark Duplicate reads` from `NGS: Picard (beta)` fillig the options as:

- "`SAM/BAM dataset...`" with "*`qG13006451 { align.sort`*"
- "`The maximum offset ... duplicates`" with "*`10`*"

4 Local Realignments Around Indels

Indels within reads often lead to false positive SNPs at the end of sequence reads. To prevent this artifact, "local realignment around indels" is done using local realignment tools from the Genome Analysis Tool Kit, set as Galaxy tools.

First we need to create a table with the possible indels. This is done with the tool `Realigner Target Creator` in the group `NGS: GATK Tools (beta)`. Fill its options as follows:

- "`Choose the source for the reference list:`" with "*`History`*"
- "`BAM file:`" with "*`qG13006451.align.sort.mark`*"
- "`Using reference file:`" with "*`GRCh37.73.dna.22.fa`*"

Having the table of possible indels we start the realignment using the tool `Indel Realigner` from the same group, filling the options as:

- "`Choose the source for the reference list:`" with "*`History`*"
- "`BAM file:`" with "*`qG13006451.align.sort.mark`*"
- "`Using reference file:`" with "*`GRCh37.73.dna.22.fa`*"
- "`Restrict realignment to provided intervals:`" with "*`possible indels`*"

To fix all this information we will use the tool `FixMateInformation`, from Picard Tools. Download the generated `.bam` file () and call the tool as:

```
java -Xmx4g -jar ~/Software/picard/FixMateInformation.jar \
    I=~/Desktop/exercise/qG13006451.align.sort.mark.real.bam \
    O=~/Desktop/exercise/qG13006451.align.sort.real.fix.bam \
    SO=coordinate \
    VALIDATION_STRINGENCY=LENIENT \
    CREATE_INDEX=true
```

We must upload the new `.bam` file (`qG13006451.align.sort.real.fix.bam`) to Galaxy.

5 Quality Recalibration

Now we will recalibrate the quality data generated from the sequencer. For that we will use the tools from the Genome Analysis Toolkit, inserted into Galaxy.

First we will use the tool `Count Covariates` from the group `NGS: GATK Tools (beta)` to create a table recalibrate the quality data, making reference to the data from dbSnp (that is done by Galaxy).
Put the first options of the tool as:

- "`Choose the source for the reference list:`" with "*History*"
- "`BAM file:`" with "*qG13006451.align.sort.real.fix.bam*"
- "`Using reference file:`" with "*GRCh37.73.dna.22.fa*"

Then, into the option "`Covariates to be used in the recalibration`", check: "*ReadGroupCovariate*", "*QualityScoreCovariate*", "*CycleCovariate*" and "*DinucCovariate*".

Having the `.csv` file from the previous call now we will use the tool from the same group, setting its options as:

- "`Covariates table recalibration file`" with "*Count Covariates - Covariate File*"
- "`Choose the source for the reference list:`" with "*History*"
- "`BAM file:`" with "*qG13006451.align.sort.real.fix.bam*"
- "`Using reference file:`" with "*GRCh37.73.dna.22.fa*"

6 SNP calling

SNP calling is done using some tools from `NGS: GATK Tools (beta)`. The tool `Unified Genotyper` calls SNPs and short indels at the same time and gives a well annotated VCF file as output.
The options of `Unified Genotyper` tool must be filled as follows:

- "`Choose the source for the reference list:`" with "*History*"
- "`BAM files`" → "`BAM file 1`" → "`BAM file:`" with "*qG13006451.align.sort.fix.recal.bam*"
- "`Using reference file`" with "*GRCh37.73.dna.22.fa*"

Having the annotated SNPs it is time to filter the possible wrong SNP calls. The filtering is done using the tool `Variant Filtration` from the same group `NGS: GATK Tools (beta)`:

- "`Choose the source for the reference list:`" with "*History*"
- "`Variant file to annotate:`" with "*snps.raw*"
- "`Using reference file`" with "*GRCh37.73.dna.22.fa*"

7 Annotation

For annotating SPS calls we will use the program annovar. It annotates a lot of different data to the SNPs and is specially suited for exome-level data-set.

The first step consists in downloading the .vcf file we generated from Gaalxy to our desktop.
The second step is to convert the .vcf file to a file compatible with annovar. This is done through the same annovar as follows:

```
˜Software/annovar/convert2annovar.pl -format vcf4
    -includeinfo snps.filtered > snp.annovar
```

Finally, to perform the annotation we will run the following command:

```
summarize_annovar.pl --buildver hg19 \
    --verdbsnp 137 \
    snp.annovar \
    humandb/ \
    -outfile snp
```

Reference to the tools

- Galaxy
  `http://galaxyproject.org`
  `https://usegalaxy.org`
  (Alternative mirror for Exome-Seq analysis: `http://orione.crs4.it/`)

- SAMtools
  `http://samtools.sourceforge.net`

- Picard
  `http://picard.sourceforge.net`

- ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data
  `http://www.openbioinformatics.org/annovar`