Heidi Lischer
University of Zurich

URPP Evolution

23 April 2014

Exercises NGS Tutorial – Part 3

Download the files:

https://www.dropbox.com/s/wk92nvj3jeri35a/human_chr20sub.bam
https://www.dropbox.com/s/12h5m53fyx2orlh/chr20.fa

Work locally on your laptop
The following instructions have been tested on Ubuntu. Mac and Windows users can use Ubuntu in the Virtual Machine (setup distributed by Stefan Wyder)

1. Copy the .bam and .fa file to your computer:
   ```
   wget https://www.dropbox.com/s/wk92nvj3jeri35a/human_chr20sub.bam
   wget https://www.dropbox.com/s/12h5m53fyx2orlh/chr20.fa
   ```
2. Install PicardTools:
   Download it from http://Sourceforge.net/projects/picard
3. Install GATK:
   Download it from the website: http://www.broadinstitute.org/gatk/download (you have to be registered!). Thus I provide the newest version as download:
   ```
   wget https://www.dropbox.com/s/gn89ads8qprio4q/GenomeAnalysisTK-3.1-1.zip
   ```

Exercise 1: Experiment Setup and BAM file

Today we will work on the low coverage data set of the 1000 genomes project. Two samples (HG00096 and HG00978) were sequenced with ILLUMINA and the reads were aligned against the human reference genome hg19.

The provided BAM file is a subset of the original data (to safe disk space and execution time) and covers the chr20 from position 62’000’000-63’000’000.

Use IGV to explore the data set (get IGV by `sudo apt-get install igv`). Load reference sequences (FASTA file) and the alignment file (BAM). IGV requires an index reference and bam file, which can be done using samtools:

```
samtools faidx REF_FILE.fa

samtools index BAM_FILE.bam
```

IGV also allows one to group reads according to read groups (samples): right click on the reads | Group alignments by | read group

Exercise 2: Realignment with GATK (not for 454 data)

Alignments need to be corrected for alignment artifacts introduced by assemblers to be able to call SNPs reliable. This normally includes duplicate removal, indel realignment, base recalibration, etc. We will focus here on the indel realignment with GATK, which is a two-step process. Additionally, GATK requires a dictionary file of your reference sequence, which can be created using PicardTools:

```
java -jar picard-tools/CreateSequenceDictionary.jar R=chr20.fa
O=chr20.dict
```

1. Create target list of intervals to be realigned:
   ```
   java -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -R
   REF_FILE.fa -I BAM_FILE.bam -o target_intervals.list
   ```

2. Perform realignment of the target intervals
   ```
   java -jar GenomeAnalysisTK.jar -T IndelRealigner -R REF_FILE.fa -
   I BAM_FILE.bam -targetIntervals target_intervals.list -o
   BAM_FILE_realigned.bam
   ```

Exercise 3: Variant calling with GATK

Here we aim to identify SNPs and Indels using GATK. The newest version of GATK includes two programs to call variants: UnifiedGenotyper and HaplotypeCaller (see the handouts for a short summery of the differences between the two programs).

1. Run UnifiedGenotyper
   ```
   java -jar GenomeAnalysisTK.jar -T UnifiedGenotyper -R REF_FILE.fa
   -I BAM_FILE_realigned.bam -ploidy 2 -glm BOTH -stand_call_conf 30
   -stand_emit_conf 10 -mbq 10 -o raw_variants_UG.vcf
   ```

2. Run HaplotypeCaller
   ```
   java -jar GenomeAnalysisTK.jar  -T HaplotypeCaller -R REF_FILE.fa
   -I BAM_FILE_realigned.bam -stand_call_conf 30 -stand_emit_conf 10
   -o raw_variants_HC.vcf
   ```

Inspect the vcf files and check some SNPs using IGV. Do you see any differences between SNP callers (e.g.: check the positions 62’067’338 and 62’067’354)?

How would you filter SNPs? Which thresholds would you use (SNP quality, Genotype quality,…

Exercise 4: Variant Effect Prediction

Next we want to identify which of the identified variants are likely to have a phenotypic effect. We use the VEP web interface on http://www.ensembl.org/info/docs/tools/vep/index.html .

1. Upload your vcf file raw_variants_UG.vcf
2. Download the annotated variants as vcf file

Inspect the annotated variants. Which variant is most likely to have an effect on phenotype? Try to understand what the different columns mean (to get a short description move the cursor to the column title).

Annotate the following variants (by pasting the text into the data field). The variant contain some nonsynonymous mutations, do you expect them to be harmful or a benign?

```
1  909238  909238  G/C  +
3  361464  361464  A/-  +
5  121187650  121188519  DUP
```