Variant analysis hands-on tutorial using Chipster

Eija Korpelainen & Maria Lehtivaara, CSC –IT Center for Science, Finland, chipster@csc.fi

1. Start Chipster and open a session

Go to http://chipster.csc.fi/, click "Launch Chipster" and log in.
Select **Open Example session** and choose **course_variant_analysis_NA12878**. Inspect the session description. In this session we have only one sample (two fastq-files).

2. Quality control of raw reads with FastQC and PRINSEQ

Select the file **NA12878_1.fastq.gz** and the tool **Quality control / Read quality with FastQC** and click **Run**. To inspect the results, select the html file and **"Open in external web browser"** as the viewing option in the Visualisation panel. Repeat the same for the file **NA12878_2.fastq.gz**.
-How many reads are there? How long are they? What quality encoding is used?
-How does the base quality look like? Is there a sub-population of particularly bad reads?

Select **both** fastq files (use ctrl / cmd key to select multiple files) and the tool **Read quality statistics with PRINSEQ**. Click **Run for each**. Visualize the resulting html files as before.
-Inspect the sequence duplication level. Are there a lot of reads which have identical sequence? How many times does the most duplicated sequence occur?

3. Trim reads based on quality with PRINSEQ (note: this is not always needed in real life)

Select **both** fastq files and the tool **Preprocessing / Trim reads for several criteria with PRINSEQ.** In the parameter panel, set the parameters **Trim 3-prime end by quality** = **5** and **Minimum length = 50,** and assign the input files 1 and 2 correctly.
-Select the result file **trimlog.txt** and **View text**. How many read pairs are left after trimming? How many reads were filtered out because they were shorter than 50 bp after trimming?

Select the **trimmed fastq files** and run the tool **Read quality with FastQC** again.
-How does the base quality look now?

4. Align reads to reference genome using BWA MEM

Select **both trimmed fastq files** (use ctrl/cmd key) and the tool **Alignment / BWA MEM for single or paired end reads.** In the parameter panel, select the human reference genome **GRCh38** and assign the input files 1 and 2 correctly. Inspect the resulting BAM file.
-Is the file sorted?
-Maximize the visualization panel and look at the first read. To which chromosome does this read map? Does its pair map to the same chromosome? What is the mapping quality?

5. Get information about the alignment results with Samtools

Select the **BAM file** and run the tool **Utilities / Count alignments in BAM**. Repeat the run so that you request mapping quality to be at least 5.
-How many alignments are there and how many have a mapping quality higher than 5?

Select the **BAM file and the index file** (use ctrl/cmd key), and the tool **Count alignments per chromosome in BAM** (check in the parameter panel that the files have been correctly assigned).
-How many alignments are there in chromosome 20?

Select the **BAM file** and run the tool **Count alignment statistics for BAM**.
-How many duplicates are there?
-How many reads have a mate which didn't map, or mapped to a different chromosome?

6. Mark duplicates in BAM with Picard
Select the **BAM file** and run the tool **Mark duplicates in BAM**.
-Open the resulting BAM file. How can you tell that the duplicates have been marked?

Select **NA12878_1_dedup.bam** and run again the tool **Count alignment statistics for BAM.**
-Are any duplicates detected now?

7. Call variants with Samtools and BCFtools (note: in real life you call variants in all samples together)
Select **NA12878_1_dedup.bam** and run the tool **Variants / Call SNPs and short INDELS** using the **human reference genome GRCh38**. Rename the result file to **variants_default.vcf**.
-How many variants do you get?

Repeat the run so that you use the following parameters:
Minimum mapping quality for an alignment to be used =50
Minimum base quality for a base to be considered = 20
Minimum read depth = **10**
Maximum read depth = **300**
Output per sample number of high quality non-reference bases = yes
-How many variants are there now? How many of them are INDELs (sort the INFO column by clicking on the title row)?

8. Filter variants based on quality with VCFtools
Select **variants.vcf** and run the tool **Variants / Filter variants** so that you set **Minimum quality = 20**.
-How many variants are left?

9. Annotate variants with Ensembl Variant Effect Predictor
Select **filtered.vcf** and run the tool **Variants / Ensembl Variant Effect Predictor**. Retrieve variants which have moderate impact: Select **vep.tsv** and the tool **Utilities / Filter table by column term** and set the parameters **Column to filter by = IMPACT** and **Term to match = MODERATE**.
-What kind of consequences do these variants have and how many genes do they affect?

10. Annotate variants with Bioconductor
Select **filtered.vcf** and run the tool **Variants / Annotate variants**.
-Sort the result file coding-variants.tsv by the consequence column and drag the SYMBOL column next to it. Are the genes with non-synonymous mutations the same as the ones found in exercise 9?

11. View reads and variants in genomic context
Open **filtered.vcf** as a **spreadsheet**, click **Detach**, and sort the file by variant quality score (QUAL).

Select the **filtered.vcf** and **NA12878_1.bam.** Choose **Genome browser** from the visualization panel and click **Maximize**. Set **genome** = **hg38 (GRCh38)** and **coverage scale** = **250**, click **Go** and zoom in to nucleotide level. Use the detached VCF file to navigate from variant to variant by clicking on the locations in the **POS** column. Scrutinize the first variant which has the highest quality in more detail:
-Is the variant heterogeneous or homogeneous? Exonic or intronic? What is the read depth (DP)? How many high-quality bases support it in forward/reverse reads (DP4)?

12. Save a workflow and run it
Select **variants.vcf** and **Workflow / Save starting from selected** from the upper panel. Give your workflow a name (but don't change the .bsh ending).
Select **variants_default.vcf** and **Workflow / Run recent** and select your workflow file.
-Does the number of rows in the result file all-variants.tsv differ from the one obtained before?