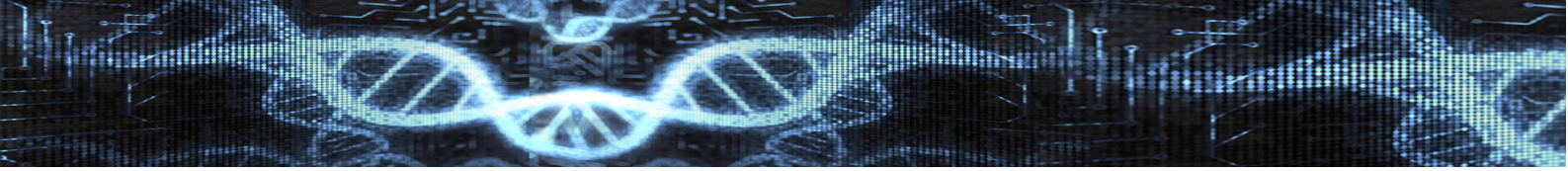

Introduction to Variant Analysis with NGS data

Practical hands-on training compiled by: Date: Lecture series: Study program:	Dr. Christian Rausch 10 November 2015 Tumor Biology and Clinical Behavior VUmc Master of Oncology

Introduction

Galaxy

In this practical session we will use Galaxy, an open, web-based platform for data intensive biomedical research. The website of this open-source project is: www.galaxyproject.org. In Galaxy, many bioinformatics tools are available by default and many more can be installed from public repositories, called tool sheds. It is also possible to integrate your own favorite software programs into Galaxy (provided they have a command line interface).

Input data

We will use a dataset of paired Illumina 151 bp reads obtained through sequencing of a TruSeq Amplicon - Cancer Panel (TSACP) of a colon cancer cell line.
The setup of TSACP has been discussed in the lecture yesterday.

Analysis workflows

After quality control, the read datasets will be mapped to the human reference genome (hg19). The variants will be analyzed and called (determined) using the program Varscan2. Called variants will be further annotated using snpEff.
Visualization of read mapping results and annotated variants will be done using IGV genome browser

Preparations

Log-in PC and startup Galaxy

- To log into the individual PCs use your VU-student account.
- Open Firefox and browse to the website as instructed:
  - Person 1-7: http://145.100.58.57:8080
  - Person 8-14: http://145.100.58.58:8080
  - Person 15-21: http://145.100.58.59:8080
  - Person 22-28: http://145.100.58.61:8080
  - Person 29-35: http://145.100.58.62:8080

- Create a log-in on Galaxy with a valid email-address.
- Log-in.

Download Sequencing Reads and SNP-Database

- Browse to: http://tinyurl.com/p8njlk5 , download and unzip the ZIP file

Import Data Into Galaxy

- On the Galaxy website, upload all 3 files.
- Select filetype fastqsanger for the fastq files, vcf for the vcf file as well as hg19 as the reference genome

Analysis

Click on Analyze data and in the 'search tools' window, search for FastQC.

Run FastQC for both fastq files.

Analyze the FastQC results (click on the eye to view a file ). *Take notes* to explain each of the graphs. If some of the graphs are unclear, take your chance on the web. If that doesn't help ask one of the assistants.

Is the quality of the reads OK? Or need parts of the reads to be trimmed? Are there still adapter sequences found? If all that is fine, please continue with mapping (=aligning) the reads against the reference genome.

BWA read mapping

BWA was discussed in the lecture, it is one of the most popular NGS read mapping programs.

- Carefully select the required input-values: The right read files and hg19 **(select 'hg19 full' if available!)** as reference should be selected.
- When you're done, click 'run': It will take a few minutes to map 200 000 x 150 x 2 reads to the human genome of 3 billion bp.
- While you are waiting: Find more information about cell line CaCo-2. Can you find mutations that have been confirmed to exist in this cell line? What is their effect?
- Rename the final output BAM file to a reasonable meaningful name:
- Click on the pencil symbol of this file: change the name and check that the Database/Build are set to hg19 and Human Feb. 2009 (GRCh37/hg19)…
- Click Save
- Now 'Convert Format' Bam to Bai. Note: Bai is an index file of the BAM that will allow fast access.
- Download the Bam and the Bai file to one new folder on your local computer e.g. on the Desktop. Rename both files to the same name, ending .bam and .bai respectively.

Visualization in IGV Integrative Genome browser

- Download IGV from www.broadinstitute.org/igv (register with name, email, VU Amsterdam, download the binary distribution (zip)).
- Unzip the file on your Desktop (or Download folder) and double click the .bat file.
- Once started, in the upper left Genome window you probably will see Human hg18.
- Click and select 'more' and then select Human hg19, which will take a moment.

- Now load your BAM file (the associated BAI file will be loaded automatically from the same directory, therefore it had to have the same name).
- Check out the following loci which have been found to be altered in a previous study of this cell line:

    - chr17:7578156-7578325
    - chr17:7579324-7579507
    - chr5:112175306-112175475

- Make sure to scroll down to see more or all of the reads.
- Do you see colored reads? What does that mean? Please check the web/manual.
- Are these mutations homozygous or heterozygous?

Variant Calling and Annotation

- Process your BAM file with "flagstat" and look at the mapping statistics
- Run Varscan2 "SNP detection directly from BAM"
- Run SnpSift Filter and apply a filter: (SDP>=30)
    - SDP stands that is the Samtools read depth. Samtools is called automatically by Varscan to calculate various statistics.
- Run SnpSift Annotate: You need to load the SNP-database file 'clinvar' that has been provided as input file. It contains annotations to clinically relevant SNPs.
- Run SnpEff

Take some time to look at the Results

- Analyze the summary html file of snpEff.
- Look at the top part of the resulting VCF file from snpEff.
- Show the resulting VCF to an assistant
- snpEff's output is unfortunately not very human-readable.
- Navigate to the 3 loci given above and this additional one:
- chr18:48591796-48591982
- According to snpEff's output, what can you report over these mutations?
- Download snpEff's output VCF and rename it to the same name as your downloaded BAM and BAI files .vcf and load it in IGV too. Now navigate to 2 of the indicated loci. Play a little bit in IGV to find out how to navigate and discover new mutations.
- Some of the mutations seen in the regions above were not called by Varscan2. What could be the reason why?
- How could you change the parameters so that only mutated positions are reported?
- Optional question: find at least 2 more predictions that can be done with snpEff (which are not default)?
- Optional question: snpEff output as VCF is not very human readable. Can you find another tool that could reformat snpEff's output more user-friendly?