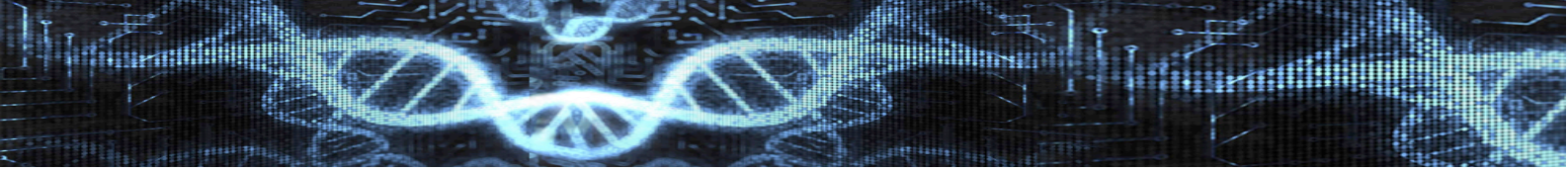

Introduction to Variant Analysis with NGS data

Practical hands-on training compiled by: Date: Lecture series: Study program:	Dr. Christian Rausch 3 November 2014 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 this morning.

Analysis workflows

After quality control, the read datasets will be mapped to the human reference genome (hg19). The variants will be analyzed and called using SAMtools Mpileup and Bcftools. 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

- Log-in and password for the individual PC is provided during the course.
- Open Firefox and browse to the following website:
  - At computers ending with numbers
    - 1 - 5: galaxy.drylab.nl
    - 6 - 10: galaxy-training1.trait-ctmm.cloudlet.sara.nl
    - 11-15: galaxy-training2.trait-ctmm.cloudlet.sara.nl
    - 16-20: galaxy-training3.trait-ctmm.cloudlet.sara.nl
    - 21-25: galaxy-training4.trait-ctmm.cloudlet.sara.nl
    - 26-30: galaxy-training5.trait-ctmm.cloudlet.sara.nl
    - 31-35: galaxy-training6.trait-ctmm.cloudlet.sara.nl (if needed).

- Create a log-in on Galaxy with your valid email-address (top menu, right side under 'user'). You will be automatically logged-in.

- Open a file browser and check if file extensions of know file types are displayed. If this is not the case, change the settings accordingly (for advanced users) or as one of the friendly assistants for help.

Import Data, History and Workflows

- On the Galaxy servers on sara.nl, data and workflows should already be available. Please check under Workflow/Workflows shared with you by others and under Shared Data.
- On galaxy.drylab.nl please import the following history, by opening this link:
    - http://galaxy.drylab.nl/u/galaxyatdrylab/h/practicalcourse followed by 'import'. The history contains the 2 read files together with clinvar-latest.vcf from dbSNP.
    - On Workflow → Upload workflow you can paste in the following links to upload 2 workflows (one at a time):
    - http://galaxy.drylab.nl/u/galaxyatdrylab/w/fastqcx2-bwa-sort-bai
    - http://galaxy.drylab.nl/u/galaxyatdrylab/w/mpileup-bcftoolsview-snpsiftfilterannotatesnpeff-imported-from-uploaded-file

Alternatively, you can also download all data from here: tinyurl.com/ppvy5el

Galaxy has an upload button in the upper left side () to upload the data. Workflows must be imported in Workflows -> Upload or Import Workflow …

Analysis

Click on Analyze data and in the 'search tools' window search for FastQC. There might be 2 versions installed, just take the second one.

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 you chance on the web. If that doesn't help ask one of the assistants.

BWA read mapping

Actually, a complete workflow containing FastQC – BWA – and sorting of the resulting BAM file by coordinates is provided.

- Go to the Workflows menu. Right-mouse-click the first workflow (FastQC-BWA…) and select edit. Here you can see a graphical view of the workflow. Click on the Workflow menu again (confirm 'Leave Page'), right-mouse-click and run the workflow.
- Carefully select the required input-values: The right read files and hg19 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 also the Database/Build where you search for hg19 and select 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, ending .bam and .bai respectively.

Visualization in IGV Integrative Genome browser

- Download IGV from www.broadinstitute.org/igv (-> binary distribution).
- 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 homozygotic or heterocygotic?

Variant Calling and Annotation

- Load the second workflow (MPileup, bcftools, snpSift filter & annotate, snpEff).
- Take a look at this workflow and the run it. Make sure, input files are set correctly.
- Analyze the summary html file of snpEff.
- Look at the top part of the resulting VCF file from snpEff.
- Download this file and import it to Excel, show this result 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.
- In our variant call and annotation pipeline, all positions that were sequenced were reported even these that have no mutations.
- How could you change the pipeline 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?