Bioinformatics variant calling analysis pipeline (version 1.0) guidelines for Next generation sequencing (exome) of human samples using Illumina NextSeq500 platform

Theodosiou Athina, PhD

Bioinformatician
Department of Cytogenetics and Genomics
The Cyprus Institute of Neurology & Genetics
Nicosia, Cyprus

A typical exome sequencing experiment aims for the identification of pathogenic mutations or small copy number variation (CNV) in exonic regions of each patient. Depending on the platform used and the flow-cell it supports, a different number of patients are multiplexed in each run. When Illumina platform is used for sequencing, the samples of a single experimental run can be automatically de-multiplexed in FASTQ format, as long as the cloud BaseSpace option of Illumina is selected for post processing. In the case that data are saved locally, this can be de-multiplexed by a specific algorithm on site. Here we describe the necessary steps of the bioinformatics pipeline needed in order to get as accurate variant calls as possible using as input FASTQ files generated by Illumina NextSeq500 sequencer from a paired-end exome sequencing experiment.

The pipeline is mainly based on the Best Practices provided by the Genome Analysis Toolkit (GATK) (https://software.broadinstitute.org/gatk/)[1] (Figure 1). Once the single nucleotide polymorphisms (SNPs) and small insertions or deletions (indels) are identified, variant effect prediction (VEP) tool[2] is applied and further annotation and interpretation is handled by GEMINI tool[3]. The pipeline described here (Figure 2) is based on the current available resources and tools and runs on Cytera HPC clusters of the Cyprus Institute (https://www.cyi.ac.cy/index.php/cyi/about-us/about-the-cyprus-institute/the-cytera-hpc-facility.html). Cytera is an IPM Hybrid CPU/GPU cluster with 98 twelve core compute nodes and 18 dual-GPU nodes and is provided as part of preparatory access project[†].

[†] "This work was supported by the Cy-Tera Project (ΝΕΑ ΥΠΟΔΟΜΗ/ΣΤΡΑΤΗ/0308/31), which is co-funded by the European Regional Development Fund and the Republic of Cyprus through the Research Promotion Foundation."

The pipeline is wrapped within Perl and bash scripts in order to run on the clusters. The tools and versions used in the analysis are:

- GATK (v3.3)
- BWA (BWA-mem) (0.7.4)
- Picard (1.109)
- Trim_Galore (0.3.7)
- Cutadapt (1.5)
- FastQC (0.10.1)
- Samtools (0.1.18)
- VEP (version 81)
- Gemini (20.0)
- Perl (v5.18.2)

The input files needed for the analysis are:

- 8 Fastq files (8 for each sample)
- Manifest bed format file with targeted human exome region coordinates for hg19
- Indexed Human Reference genome (hg19)
- dbSNP version 138
- Mills gold standard indel reference (hg19.sites)
- 1000G gold standard indel reference (hg19.sites)

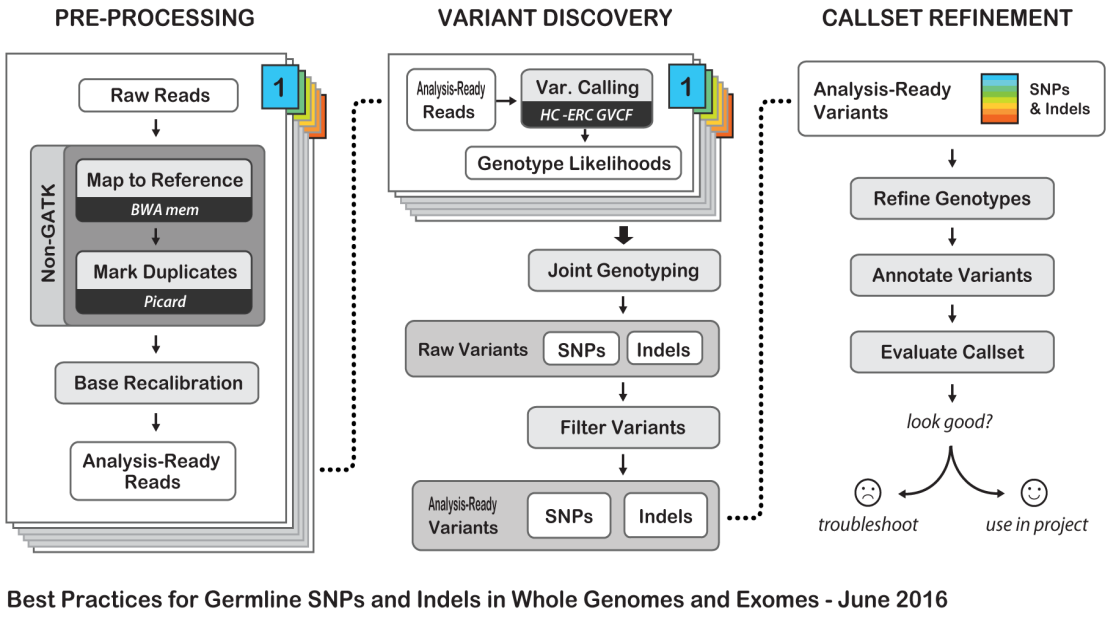

Figure 1. Best Practices guidelines for the detection of germline SNPS and indels in whole genome and exomes (based on current GATK version 3.7 and retrieved on 11-07-17).

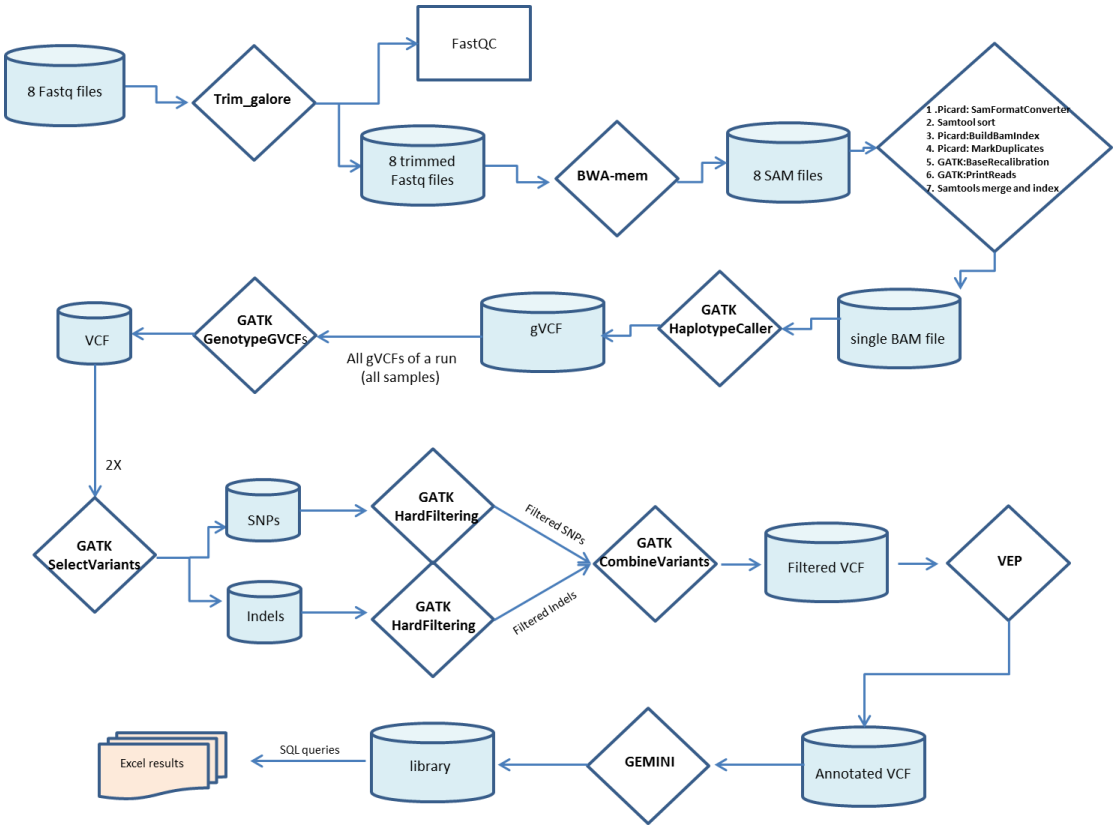

Figure 2. In house bioinformatics pipeline for variant calling of an exome sequencing paired-end experiment using Illumina NextSeq500 platform.

A. Step 1. Preprocessing

1. *Quality control.* Initially, the reads produced by each sequencing lane of the flowcell need to go through a quality control before the next step of the analysis. The FastQC tool (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) is freely available and can provide a simple way to perform some quality control checks on raw sequence data coming from high throughput sequencing. In addition to giving a quick impression of whether the data are problematic or not, the FastQC tool can provide systematic errors produced by platforms or library kits. The tool provides summary graphs and tables to assess the data and exports results as an HTML report.
2. *Trimming of the Fastq files for adapters and quality*. The data after FastQC inspection can be trimmed for adapters and quality with cutadapt tool[4] (in order to improve the next critical step of alignment.

   The above mentioned tools are used within a single wrapper tool the trimgalore tool (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) in a single command.

B. Step 2. Alignment to the genome

In the current pipeline, bwa algorithm (http://bio-bwa.sourceforge.net/) and specifically BWA-mem is used for sequence alignment, in order to map the reads to the reference genome. Currently we use the GRCh37/hg19 version of human reference genome and default parameters with flag parameters –M (this tells it to consider split reads as secondary) and -R (Read group information which is essential for downstream GATK functionality). Moreover –t flag is used when the pipeline is run on the clusters. The output of bwa mem is in SAM format [5].

C. Step 3. Alignment post-processing

1. Picard SamFormatConverter tool is then used to convert SAM format to the binary representation of SAM, the BAM which keeps exactly the same information as SAM. BAM is compressed by the BGZF library.

2. Samtool sort function is used to sort the BAM files.
3. Picard BuildBamIndex is used to index the BAM and create its dictionary (the BAI format).
4. Reads are marked (not removed) for optical and PCR duplicates with Picard MarkDuplicates tool.
5. GATK BaseRecalibration tool is used to perform Base Recalibration using as known sites dbSNP version 138 and gold indel reference the Mills and 1000G gold standard indels for hg19.
6. GATK PrintReads is used to apply the recalibration to the reads.
7. Samtools merge and index in order to get one BAM file for each sample
8. Statistical analysis based on the aligned reads is performed with several Picard tools and report is generated, for example with the mean target coverage, uniformity, percentage of bases achieving 2,10,20,30,50 and 100X coverage and several other measures crucial for interpretation of the sample and the whole experimental run.

D. Step 4. Variant Discovery

1. GATK HaplotypeCaller is used in ERC mode for each sample separately, in order to perform accurate variant calling and call SNPs and small INDELs. The output is in genomic VCF format (gVCF) containing extra information that enhances the variant analysis.
2. All samples of a run are used as input in the GATK GenotypeGVCFs tool that merges gVCF records. This tool performs the multi-sample joint aggregation step and merges the records together in a sophisticated manner in order to produce correct genotype likelihoods, re-genotype the newly merged record, and then re-annotate it. The output is in vcf format.
3. GATK Hardfiltering is then performed separately on SNPs and INDELs (GATK SelectVariants) using the parameters recommended by GATK for exome sequencing analysis:
For SNPs: "QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0."

(QD: Quality By Depth; FS: Fisher strand; MQ: Mapping Quality; MQRankSum: Base Quality Rank Sum Test and ReadPosRankSum:Read position Rank Sum Test)
For Indels: "QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0"
4. Last, GATK CombineVariants is used to combine filtered INDELs and SNPs into a single VCF file.

E. Step 5. Annotation of variants

For the annotation of the variants we use GEMINI[3], a flexible framework for exploring variation. GEMINI currently supports only human genetic variation mapped to build 37 (hg19) of the human genome. Prior loading to GEMINI, the vcf file is annotated using VEP[2]. GEMINI creates a library for single vcf that is given. The library can be then searched using SQL queries and results can be retrieved and visualized in excel format.

F. Step 6. Interpretation of results (semi-automated procedure)

In most of the cases, samples are grouped and analyzed in trios (patient with both parents), therefore a separate analysis per case/family has to be performed. With GEMINI tool our front line walkthrough is:

1. Filtering the variants based on population frequency. We apply a cut-off of minor allele frequency |(MAF) less than 1% using the datasets of 1000 genome, NHLBI GO Exome Sequencing Project (ESP) and Exome Aggregation Consortium (ExAc).
2. Most of our cases are trios, therefore we use an inheritance pattern such as *de novo*, autosomal recessive, X-linked (GEMINI provides the tools for these queries).
3. Depending on the disease phenotype we use *in silico* panels of genes in order to limit our search pool according to the phenotype described by the clinician. For example we created *in silico* panels with intellectual disability related genes, with genes related to microcephaly, and with genes related to cohensinopathies based on current available commercial panels and the literature. These panels can be updated regularly when new genes related to the diseases are described. In addition to our panel design, we ask our referring clinicians to recommend potential genes for more targeted search.

4. Our first line strategy is to limit the search pool on the variants with high impact severity prediction such as frameshifts, exon_deleted, splice_acceptor, splice_donor, start_loss, stop_gain, stop_loss and medium impact severity predictions such as missense, codon_gain, inframe_codon_loss and inframe_codon_change. Synonymous, variants on 5' and 3' UTRs and introns are considered with low impact severity, therefore they are not evaluated further with the exception of synonymous affecting splicing if they are withing genes of interest.

5. Through GEMINI we annotate variants that are found in OMIM (https://www.omim.org/) as well as those that are described in ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), therefore we will not miss a known pathogenic variant.

6. We use a minimum cutoff of at least 15X depth in order to consider a variant potential candidate and we critically inspect variants found in low complexity regions. NGS analysis process produces many false positives within low complexity regions, therefore in most cases these variants are not evaluated further.

7. Potential candidate variants from the vcf are inspected along with the aligned reads within the bam files using the IGV tool (http://software.broadinstitute.org/software/igv/) and variants are compared with the variants found in local population to exclude local polymorphisms.

8. Last, the best candidate variants are confirmed/or rejected with Sanger sequencing.

Concluding, the strategy on the interpretation step of the analysis differs depending on the family, phenotype and inheritance mode we are searching for.

References

1. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 2010;20:1297-303.
2. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. Genome biology 2016;17:122.
3. Paila U, Chapman BA, Kirchner R, Quinlan AR. GEMINI: integrative exploration of genetic variation and genome annotations. PLoS computational biology 2013;9:e1003153.
4. M. M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal 2011;17:10-2.
5. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078-9.