Variants analysis of a family trio 50x WGS dataset

Stéphane Plaisance

November 14th, 2017 - version 1.1

Contents

1 Aim 2
2 Material and Method 2
3 Download Illumina read data from a normal and tumor paired samples 2
4 Map reads to the human reference genome (GRCh38.p10) using BWA mem 5
5 Call variants with Varscan2 trio 6
6 References 8

illustration\textsuperscript{1}

\textsuperscript{1}VIB - Nucleomics Core, nucleomics@vib.be
\textsuperscript{1}https://abm-website-assets.s3.amazonaws.com/biosciencetechnology.com/s3fs-public/shutterstock_125006387.jpg
1 Aim

This report is a partial revisit of a workflow described in the **bcbio nextgen** tutorial page\(^2\) using **Varscan 2.** The same data was analised by a number of groups including **De Pristo e al** (DePristo et al. 2011) who reported benchmarking results on several NGS dataset obtained from these Coriel cell lines on Illumina HiSeq 2000 sequencers. More information about this trio and benchmarking was also posted on the Broad GATK forum\(^3\) that should be read if you consider reproducing and extending the work presented in this report.

The Illumina paired-end reads were obtained from the links provided in the **Bebio Nextgen** page and aligned to the human GRCh38 reference genome using **BWA mem.** After QC on the BAM data, variants were analyzed using the dedicated **varsan2 trio** workflow as detailed in the original paper (Koboldt, Larson, and Wilson 2013). Other callers than Varscan 2 can be used among others the classical **GATK** toolbox\(^4\) but we limited this analysis to Varscan which will be deployed in the VIB BITS Genepattern server.

2 Material and Method

3 Download Illumina read data from a normal and tumor paired samples

This tutorial uses a **WGS** trio (NA12891 [father], NA12892 [mother], NA12878 [child]) part of the CEPH Utah-Pedigree-1463 and sequenced as part of the Illumina platinium genomes.\(^5\)

- NA12878: https://www.ncbi.nlm.nih.gov/sra/ERX168836%5Baccn%5D
- NA12891: https://www.ncbi.nlm.nih.gov/sra/ERX168849%5Baccn%5D
- NA12892: https://www.ncbi.nlm.nih.gov/sra/ERX168850%5Baccn%5D

---

\(^2\)http://bcbio-nextgen.readthedocs.io/en/latest/contents/testing.html

\(^3\)https://gatkforums.broadinstitute.org/gatk/discussion/1292/which-datasets-should-i-use-for-reviewing-or-benchmarking-purposes

\(^4\)https://software.broadinstitute.org/gatk/best-practices/bp_3step.php?case=GermShortWGS

Figure 1: SRA CEPH trio data-sets

(a) NA12878 (child)  
(b) NA12891 (father)  
(c) NA12892 (mother)

Figure 2: CEPH CEU Trio

```
#!/bin/bash
set -eu -o pipefail

BASE=/data/NA12878-trio-eval
```
cd $BASE
reads=reads
mkdir -p ${reads} && cd ${reads}

# Genome data
wget -c -O NA12878_1.fastq.gz ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR194/ERR194147/ERR194147_1.fastq.gz
wget -c -O NA12878_2.fastq.gz ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR194/ERR194147/ERR194147_2.fastq.gz
wget -c -O NA12891_1.fastq.gz ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR194/ERR194160/ERR194160_1.fastq.gz
wget -c -O NA12891_2.fastq.gz ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR194/ERR194160/ERR194160_2.fastq.gz
wget -c -O NA12892_1.fastq.gz ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR194/ERR194161/ERR194161_1.fastq.gz
wget -c -O NA12892_2.fastq.gz ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR194/ERR194161/ERR194161_2.fastq.gz

# QC on all read files
# create folder for QC-results inside the read folder
outfolder= readQC
mkdir -p ${outfolder}

# on each reads file
for fq in NA*.fastq.gz; do
    echo "analysing $(fq)"
    # perform QC test using 1 thread and zip results
    fastqc -t 1 -o ${outfolder} --noextract $(fq)
done

All three read sets (read-1 shown only) were found of good quality as reported by fastqc.
4 Map reads to the human reference genome (GRCh38.p10) using BWA mem

Read mapping was performed as usual with BWA mem (Li and Durbin 2009) default parameters and adding a @RG field to the BAM header with the -R argument, the -M argument was also used although its effect is not known in the context of the current analysis. Both normal and tumor reads were mapped separately to create two SAM files. The SAM data was sorted, converted to BAM, and duplicates were marked using Picard tools.6

BASE=/data/NA12878-trio-eval
cd $BASE

# create folder for mappings
out=bwa_mapping
mkdir -p $out

refidx=/data/biodata/bwa_indexes/GRCh38.p10

for sample in NA12878 NA12891 NA12892; do

6http://broadinstitute.github.io/picard/
```bash
# loop
rgstring="@RG\tID:${sample}\tLB:lib-${sample}\tPU:unkn-0.0\tPL:ILLUMINA\tSM:${sample}\n" \\
$[refidx]\nreads/${sample}_1.fastq.gz \\
reads/${sample}_2.fastq.gz \\
gzip -3 > ${out}/${sample}_GRCh38.p10.sam.gz

echo "# ${cmd}"

eval ${cmd}

[ $?==0 ] || exit 1

# convert to sorted BAM
java -jar $PICARD/picard.jar SortSam \\
I=${out}/${sample}_GRCh38.p10.bam \\
O=${out}/${sample}_GRCh38.p10.bam \\
SO=coordinate \nCREATE_INDEX=TRUE \nTMP_DIR= . \nrm ${out}/${sample}_GRCh38.p10.sam.gz

# QC and plots
mkdir -p ${out}/${sample}_QC
$BIOTOOLS/samtools/bin/samtools stats ${out}/${sample}_GRCh38.p10.bam | \\
> ${out}/${sample}_QC/${sample}_GRCh38.p10.bam_stats.txt

plot-bamstats -p ${out}/${sample}_QC/${sample}_QC \\
${out}/${sample}_QC/${sample}_GRCh38.p10.bam_stats.txt

# MarkDuplicates
# ! original read names were not well formatted
# => Illumina position info was not stored by BWA
# ==> only PCR duplicates will be identified based on mapping coordinates
java -jar $PICARD/picard.jar MarkDuplicates \\
READ_NAME_REGEX=null \nVALIDATION_STRINGENCY=LENIENT \nCREATE_INDEX=true \\
I=$BASE/${out}/${sample}_GRCh38.p10.bam \\
O=$BASE/${out}/${sample}_GRCh38.p10_mrkdup.bam \\
M=$BASE/${out}/${sample}_GRCh38.p10.bam_marked_dup_metrics.txt

# end loop
done

# perform Qualimap BAM-QC on the final BAM data
# prepare sample list
echo -e "NA12878\t$BASE/${out}/NA12878_GRCh38.p10.bam\tCEPH-Trio-1463" > $BASE/${out}/samples.txt

echo -e "NA12891\t$BASE/${out}/NA12891_GRCh38.p10.bam\tCEPH-Trio-1463" >> $BASE/${out}/samples.txt

echo -e "NA12892\t$BASE/${out}/NA12892_GRCh38.p10.bam\tCEPH-Trio-1463" >> $BASE/${out}/samples.txt

# qualimap QC
$QUALIMAP/qualimap multi-bamqc -c -r \\
-d $BASE/${out}/samples.txt \\
-gff /data/biodata/references/GRCh38.p10/Homo_sapiens.GRCh38.p10.gff \\
-outd $BASE/${out} \\
-outfile qualimap_qc.pdf \\
-outformat pdf

5 Call variants with Varscan2 trio

Some of the Varscan 2 (v2.4.3) command used to call trio variants were taken from the github support pages\(^7\) since the inline documentation is not up-to date despite some info on the readme file of the latest build.\(^8\) A more

\(^7\)http://dkoboldt.github.io/varscan/
\(^8\)https://github.com/dkoboldt/varscan/blob/master/VarScan.v2.4.3.description.txt
elaborated tutorial on how to use Varscan 2 and how to filter the raw Varscan results can be found in the 2013 Varscan2 *Curr. Protoc. Bioinformatics* publication (Koboldt, Larson, and Wilson 2013).

**USAGE:** java -jar VarScan.jar trio [mpileup file] [output-basename] OPTIONS

mpileup file - The SAMtools mpileup file for father, mother, child in that order

OPTIONS:

--output-name An output base name for VCF files of results. Required for piped input
--min-coverage Minimum read depth at a position to make a call [20]
--min-reads2 Minimum supporting reads at a position to call variants [2]
--min-avg-qual Minimum base quality at a position to count a read [15]
--min-var-freq Minimum variant allele frequency threshold [0.20]
--min-avg-qual Minimum frequency to call homozygote [0.75]
--p-value Default p-value threshold for calling variants [0.05]
--adj-var-freq Adjusted minimum VAF when recalling at variant site [0.05]
--adj-p-value Adjusted p-value when recalling at variant site [0.10]
--vcf-sample-list For VCF output, a list of sample names in order, one per line
--variants Report only variant (SNP/indel) positions [0]

The Varscan trio calling and post processing steps were applied as follows:

```
BASE=/data/NA12878-trio-eval
cd $BASE
  # create folder for varscan calls
  out=varscan_trio
  mkdir -p $out
  REFFILE=/data/biodata/references/GRCh38.p10/Homo_sapiens.GRCh38.p10.dna.primary_assembly.fa
  FATHER=bwa_mapping/NA12878_GRCh38.p10.bam
  MOTHER=bwa_mapping/NA12891_GRCh38.p10.bam
  CHILD=bwa_mapping/NA12892_GRCh38.p10.bam
  OUTFILE=$out/varscan_mpileup_trio

  # the first ‘-‘ receives the pipe, the next argument $2 becomes the output prefix
  /opt/biotools/bin/samtools mpileup -B -f ${REFFILE} ${FATHER} ${MOTHER} ${CHILD} \
    | java -jar /opt/biotools/varscan/varscan.jar trio - ${OUTFILE} --mpileup 1 \
    --variants
```

The classification of the obtained calls was:

The above command will produce two VCF output files: one for SNPs (trio.mpileup.output.snp.vcf) and one for indels (trio.mpileup.output.indel.vcf). Relevant INFO fields include:

- FILTER - mendelError if MIE, otherwise PASS
- STATUS - 1=untransmitted, 2=transmitted, 3=denovo, 4=MIE
- DENOVO - if present, indicates a high-confidence de novo mutation call

At this point we have a VCF output that need additional filtering as described in the Varscan 2 publication (Koboldt et al. 2012). Genomic annotations should also be added in order to identify potential effective variants in the child (either de-novo or acquired).

Although not all tools presented here were yet ported to the BITS genePattern server, this could be done is sufficient interest is shown by our users (please contact **bits@vib.be**).
6 References


