ADBS Whole Exome Sequencing (WES) Analysis Pipeline

Ravi More[1]

[1]ADBS, National Centre for Biological Sciences (NCBS)

dx.doi.org/10.17504/protocols.io.vrhe536

Ravi More

ABSTRACT

Integrated Whole Exome Sequencing (WES) analysis pipline using various tools and databases, developed as part of the Accelerator program for Discovery in Brain disorders using Stem cells (ADBS) program at National Centre for Biological Sciences (NCBS), Bangalore.

THIS PROTOCOL ACCOMPANIES THE FOLLOWING PUBLICATION

Suhas Ganesh, Husayn Ahmed P, Ravi Kumar Nadella, Ravi Prabhakar More, Manasa Sheshadri, Biju Viswanath, Mahendra Rao, Sanjeev Jain, The ADBS consortium, Odity Mukherjee, 2018. Exome sequencing in families with severe mental illness identifies novel and rare variants in genes implicated in Mendelian neuropsychiatric syndromes. Psychiatry and Clinical Neurosciences. doi:10.1111/pcn.12788

PROTOCOL STATUS

Working

We use this protocol in our group and it is working

SAFETY WARNINGS

Define paths and directories

1

COMMAND

```
SAMPLE_PATH="/path/to/sample"
SAMPLE_NAME="test_sample"
SOFTWARE_PATH="/path/to/software"
DATABASES_PATH="/path/to/databases"
TEMP_DIR="/path/to/temp"
```

LINUX

Unzip the raw reads files from .gz to fastq format

2

COMMAND

```
gunzip $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME*.fq.gz
```

Linux

QC check of R1 and R2 paired-end raw reads using FASTQC, Trimming poor quality reads using Prinseq-lite, and Adapter contimination removal using AfterQC,

3 Software versions used:

FASTQC version 0.10.1
Prinseq-lite version 0.20.4
AfterQC version 0.9.6

COMMAND

```
$SOFTWARE_PATH/FastQC/fastqc $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_R1.fq

$SOFTWARE_PATH/FastQC/fastqc $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_R2.fq

cd $SAMPLE_PATH/$SAMPLE_NAME/

python $SOFTWARE_PATH/AfterQC-master/after.py -f -1 -t -1 -q 20 -1 $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_R1.fq -2 $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_R2.fq

$SOFTWARE_PATH/prinseq-lite-0.20.4/prinseq-lite.pl -fastq $SAMPLE_PATH/$SAMPLE_NAME/good/$SAMPLE_NAME\_R1.good.fq -fastq2 $SAMPLE_PATH/$SAMPLE_NAME/good/$SAMPLE_NAME\_R2.good.fq -out_good $SAMPLE_PA1

mv $SAMPLE_PATH/$SAMPLE_NAME/cleaned_1.fastq $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_cleaned_R1.fastq

mv $SAMPLE_PATH/$SAMPLE_NAME/cleaned_2.fastq $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_cleaned_R2.fastq

$SOFTWARE_PATH/FastQC/fastqc $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_cleaned_R1.fastq

$SOFTWARE_PATH/FastQC/fastqc $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_cleaned_R2.fastq

mkdir -p $SAMPLE_PATH/$SAMPLE_NAME/Report_$SAMPLE_NAME\_4_FASTQC

mv $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_cleaned_R1_fastqc.zip $SAMPLE_PATH/$SAMPLE_NAME/Report_$SAMPLE_NAME\_4_FASTQC/$SAMPLE_NAME\_cleaned_R1_fastqc.zip

mv $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_cleaned_R2_fastqc.zip $SAMPLE_PATH/$SAMPLE_NAME/Report_$SAMPLE_NAME\_4_FASTQC/$SAMPLE_NAME\_cleaned_R2_fastqc.zip
```

Linux

Alignment of clened raw reads against Human Reference Genome hg19 GRCh37.p13 build using BWA and SAMTOOLS.

4 BWA version 0.5.9
Samtools version 1.3

COMMAND

```
$SOFTWARE_PATH/bwa-0.5.9/bwa aln -t 30 $DATABASES_PATH/hg19_fa-chrMlast/hg19_chrM-last.fa $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_cleaned_R1.fastq > $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_R1.s

$SOFTWARE_PATH/bwa-0.5.9/bwa aln -t 30 $DATABASES_PATH/hg19_fa-chrMlast/hg19_chrM-last.fa $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_cleaned_R2.fastq > $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_R2.s

$SOFTWARE_PATH/bwa-0.5.9/bwa sampe $DATABASES_PATH/hg19_fa-chrMlast/hg19_chrM-last.fa $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_R1.sai $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_R2.sai $SAMPLE_

$SOFTWARE_PATH/samtools1.3/bin/samtools view -bS $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME.sam > $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME.bam

$SOFTWARE_PATH/samtools1.3/bin/samtools sort $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME.bam -o $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_sorted.bam

$SOFTWARE_PATH/samtools1.3/bin/samtools flagstat $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_sorted.bam > $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_sorted_flagstat.txt

$SOFTWARE_PATH/samtools1.3/bin/samtools index $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_sorted.bam > $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_sortedbam.bai
```

Linux

Mark PCR duplicates and sorting BAM using PICARD Tools

5 Picard version 2.0.1

COMMAND

```
java -Djava.io.tmpdir=$TEMP_DIR -Xmx50g -jar $SOFTWARE_PATH/picard/build/libs/picard.jar AddOrReplaceReadGroups I="$SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_sorted.bam" O="$SAMPLE_PATH/$SAMPLE_NAME/$SAM

java -Djava.io.tmpdir=$TEMP_DIR -Xmx50g -jar $SOFTWARE_PATH/picard/build/libs/picard.jar MarkDuplicates I="$SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_coordsort.bam" O="$SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_N
```

Linux

Index the coordinate sorted bam file using SAMTOOLS

6 Samtools version 1.3

COMMAND

```
$SOFTWARE_PATH/samtools1.3/bin/samtools index $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_RMDUP.bam
```

Linux

INDEL re-alignment using GATK tools

7 GATK version 3.6

COMMAND

```
java -Xmx8g -jar $SOFTWARE_PATH/GenomeAnalysisTK-3.6/GenomeAnalysisTK.jar -T RealignerTargetCreator -R $DATABASES_PATH/hg19_fa-chrMlast/hg19_chrM-last.fa -I $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_RMDUI

java -Xmx30g -jar $SOFTWARE_PATH/GenomeAnalysisTK-3.6/GenomeAnalysisTK.jar -T IndelRealigner -R $DATABASES_PATH/hg19_fa-chrMlast/hg19_chrM-last.fa -I $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_RMDUP.bam -k
```

Linux

Check the alignment QC of the bam file using Qualimap

8 Qualimap version 2.2.1

COMMAND

```
mkdir -p $SAMPLE_PATH/$SAMPLE_NAME/Report_$SAMPLE\_NAME_5_ALIGNMENT_QC

$SOFTWARE_PATH/qualimap_v2.2.1/qualimap bamqc -bam $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_realignedBam.bam -gff $DATABASES_PATH/TruSeq_exome_targeted_regions.hg19.bed -outdir $SAMPLE_PATH/$SAMPL
```

Linux

SNP and INDEL variant calling using VarScan and SAMTOOLS

9 VarScan version 2.3.9
Samtools version 1.3

COMMAND

```
$SOFTWARE_PATH/samtools1.3/bin/samtools mpileup -f $DATABASES_PATH/hg19_fa-chrMlast/hg19_chrM-last.fa $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_realignedBam.bam > $SAMPLE_PATH/$SAMPLE_NAME/$SAMPL

mkdir -p $SAMPLE_PATH/$SAMPLE_NAME/Report_$SAMPLE_NAME\_13_VARIANT_CALLING

java -jar $SOFTWARE_PATH/VarScan.v2.3.9.jar mpileup2snp $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_raw.mpileup --min-var-freq 0.0025 --p-value 0.001 --min-avg-qual 20 --output-vcf > $SAMPLE_PATH/$SAMPLE_NAME/Rep

java -jar $SOFTWARE_PATH/VarScan.v2.3.9.jar mpileup2indel $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_raw.mpileup --min-var-freq 0.0025 --p-value 0.001 --min-avg-qual 20 --output-vcf > $SAMPLE_PATH/$SAMPLE_NAME/Rep
```

Linux

VCF QC of SNP and INDEL files using rtg-tools

10 rtg-tools version 3.7.1

COMMAND

```
mkdir -p $SAMPLE_PATH/$SAMPLE_NAME/Report_$SAMPLE_NAME\_13_VARIANT_CALLING/VCF_QC

$SOFTWARE_PATH/rtg-tools-3.7.1/rtg vcfstats $SAMPLE_PATH/$SAMPLE_NAME/Report_$SAMPLE_NAME\_13_VARIANT_CALLING/$SAMPLE_NAME\_snp_0.25freq_p0.001_qual_20.vcf > $SAMPLE_PATH/$SAMPLE_NAME/Report_$SA

$SOFTWARE_PATH/rtg-tools-3.7.1/rtg vcfstats $SAMPLE_PATH/$SAMPLE_NAME/Report_$SAMPLE_NAME\_13_VARIANT_CALLING/$SAMPLE_NAME\_indel_0.25freq_p0.001_qual_20.vcf > $SAMPLE_PATH/$SAMPLE_NAME/Report_$SA
```

Linux

SNP AND INDEL variant annotation using ANNOVAR

11 ANNOVAR reference assembly 65) with reference hg19

COMMAND

```
mkdir -p $SAMPLE_PATH/$SAMPLE_NAME/Report_$SAMPLE_NAME\_13_VARIANT_CALLING/annotated_annovar

perl $SOFTWARE_PATH/annovar/convert2annovar.pl -format vcf4 $SAMPLE_PATH/$SAMPLE_NAME/Report_$SAMPLE_NAME\_13_VARIANT_CALLING/$SAMPLE_NAME\_snp_0.25freq_p0.001_qual_20.vcf > $SAMPLE_PATH/$SAMPLE_

perl $SOFTWARE_PATH/annovar/convert2annovar.pl -format vcf4 $SAMPLE_PATH/$SAMPLE_NAME/Report_$SAMPLE_NAME\_13_VARIANT_CALLING/$SAMPLE_NAME\_indel_0.25freq_p0.001_qual_20.vcf > $SAMPLE_PATH/$SAMPLE_

perl $SOFTWARE_PATH/annovar/table_annovar.pl $SAMPLE_PATH/$SAMPLE_NAME/Report_$SAMPLE_NAME\_13_VARIANT_CALLING/$SAMPLE_NAME\_snp_0.25freq_p0.001_qual_20.vcf $SOFTWARE_PATH/annovar/humandb/ -buildv

perl $SOFTWARE_PATH/annovar/table_annovar.pl $SAMPLE_PATH/$SAMPLE_NAME/Report_$SAMPLE_NAME\_13_VARIANT_CALLING/$SAMPLE_NAME\_indel_0.25freq_p0.001_qual_20.vcf $SOFTWARE_PATH/annovar/humandb/ -build
```

Linux

Delete inter-mediate files after varifying the final results files (OPTIONAL STEP AS PER USER)

12

COMMAND

```
rm $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_R1.sai

rm $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_R2.sai

rm $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_RMDUP.bam

rm $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME.sam

rm $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_RMDUP.bam.bai

rm $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_sorted.bam

rm $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_sorted.bam.bai

rm $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_sortedbam.bai

rm $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_IndelRealigner.intervals

rm $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_raw.mpileup

rm $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_coordsort.bam

rm $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_*_R1_0*.fastq

rm $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_*_R2_0*.fastq

rm $SAMPLE_PATH/$SAMPLE_NAME/cleaned_1_singletons.fastq

rm $SAMPLE_PATH/$SAMPLE_NAME/cleaned_2_singletons.fastq

rm $SAMPLE_PATH/$SAMPLE_NAME/$SAMPLE_NAME\_part.sam
```

Linux