**Genomic alterations in Triple Negative Breast Cancer after chemotherapy**

Enas Helmy, Mai Ahmed, Sahar Mostafa, Nefertiti El-Nikhely

**Introduction**

Triple-negative breast cancer (TNBC) is an aggressive subtype that frequently develops resistance to chemotherapy. An unresolved question is whether resistance is caused by the selection of rare pre-existing clones or alternatively through the acquisition of new genomic aberrations (Kim *et al*, Cell 2018). To investigate this question, the research group did, among other approaches, exome sequencing on tumor samples before and after chemotherapy. In addition, they sequenced also matching blood samples, which will allow identifying tissue-specific mutations and exclude germline mutations.

*Reference*

Kim et al, Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single Cell Sequencing, Cell 2018 May 03; 173(4): 879–893.

**Commands**

**Part I**

1. **Sample selection and retrieval**

# To get data

wget <http://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/current/sratoolkit.current-centos_linux64.tar.gz>

fastq-dump SRR5969410

fastq-dump SRR5969460

# Output

(ngs1) ngs@bioinfo\_nu:~/workdir/Project1$ fastq-dump SRR5969460

Read 2734585 spots for SRR5969460

Written 2734585 spots for SRR5969460

(ngs1) ngs@bioinfo\_nu:~/workdir/Project1$ fastq-dump SRR5969410

Read 1892386 spots for SRR5969410

Written 1892386 spots for SRR5969410

#File sizes

|  |  |
| --- | --- |
| File | Size |
| SRR5969410 | 442405582 |
| SRR5969460 | 640289924 |

Files were renamed to show treatment status so they are called:

PreT\_SRR5969410.fastq

PostT\_ SRR5969460.fastq

As ouput is unzipped file, we zipped before alignment using

gzip PreT\_SRR5969410.fastq

Issue 1: Data selection WGS vs WXS & Downloading samples with ftp, trace, wget, and windows download

# Samples selected are from the website <https://www.ncbi.nlm.nih.gov/Traces/study/?page=13&acc=SRP114962&f=assay_type_s%3An%3Awgs%3Ac&o=sample_name_s%3Ad&s=SRR5969409>

SRR5969410 KTN102 pretreatment #800 Sample KTN102\_0\_25 (48 spots)

SRR5969460 KTN102 posttreatment #645 Sample KTN102\_OP\_1 (48spot)

# Download reference genome

From Ensembl database <http://www.ensembl.org/info/data/ftp/index.html>

<ftp://ftp.ensembl.org/pub/release-99/fasta/homo_sapiens/dna/>

wget ftp://ftp.ensembl.org/pub/release-99/fasta/homo\_sapiens/dna/Homo\_sapiens.GRCh38.dna.toplevel.fa.gz

Issue 2: Downloading reference

wget <ftp://ftp.ensembl.org/pub/release-99/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.X.fa.gz>

gunzip Homo\_sapiens.GRCh38.dna.chromosome.X.fa.gz

Rename to GRCh38.dna.chromosome.X.fa

1. **Quality Check**

conda install -c bioconda fastqc

conda install -c bioconda multiqc

for f in ~/workdir/Project1/\*.fastq.gz;do fastqc -t 1 -f fastq -noextract $f;done

mv ../ workdir/Project1/\*html ./

mv ../ workdir/Project1/\*zip ./

multiqc -z -o . .

1. **Alignment**

# Indexing of Chromosome 10 (based on breast cancer in Japanese population)

wget ftp://ftp.ensembl.org/pub/release-99/fasta/homo\_sapiens/dna/Homo\_sapiens.GRCh38.dna.chromosome.10.fa.gz

gunzip Homo\_sapiens.GRCh38.dna.chromosome.10.fa.gz

mkdir -p ~/workdir/Project1/bwa\_align10/bwaIndex10 && cd ~/workdir/Project1/bwa\_align10/bwaIndex10

ln -s ~/workdir/Project1/GRCh38.dna.chromosome.10.fa .

bwa index -a bwtsw GRCh38.dna.chromosome.10.fa

# Sequence Alignment

cd ~/workdir/Project1/bwa\_align10

R1="$HOME/workdir/Project1/PreT\_SRR5969410.fastq.gz"

/usr/bin/time -v bwa mem bwaIndex10/GRCh38.dna.chromosome.10.fa $R1 >PreT\_SRR5969410\_Chr10.sam

1. **Visualization of mapping**

conda install samtools

#Convert sam into bam

samtools view -hbo PreT\_SRR5969410\_Chr10.bam PreT\_SRR5969410\_Chr10.sam

# Sort bam file

samtools sort PreT\_SRR5969410\_Chr10.bam -o PreT\_SRR5969410\_Chr10.sorted.bam

# Index sorted file

samtools index PreT\_SRR5969410\_Chr10.sorted.bam

# Visualize

samtools tview -p chr10:62155107 PreT\_SRR5969410\_Chr10.sorted.bam bwaIndex10/GRCh38.dna.chromosome.10.fa

1. **Variant Calling**

conda install bcftools

bcftools mpileup -Ou -f bwaIndex10/GRCh38.dna.chromosome.10.fa PreT\_SRR5969410\_Chr10.sorted.bam |\

bcftools call -Ov -mv > PreT\_SRR5969410\_Chr10.vcf

1. **VCF statistics**

conda install -c bioconda tabix

bgzip -c PreT\_SRR5969410\_Chr10.vcf > PreT\_SRR5969410\_Chr10.vcf.gz

tabix -p vcf PreT\_SRR5969410\_Chr10.vcf.gz

# For statistics

conda install -c bioconda rtg-tools

rtg vcfstats PreT\_SRR5969410\_Chr10.vcf.gz > stats.txt

1. **Data**

Join variant calling was not a feature of Mutect2. However, Mutect2 v4.1.0.0 onwards enables joint analysis of multiple samples.

**Part II**

# To compare variants in blood and tumour samples in order to identify tissue-specific variants can we use GATK.

picard\_path=$CONDA\_PREFIX/share/picard-\*

java -Xmx2g -jar $picard\_path/picard.jar MergeSamFiles I=PreT\_SRR5969410.sorted.bam I=align.sorted.bam OUTPUT=merged.sorted.bam

**List of useful resources**

<http://www.sixthresearcher.com/list-of-helpful-linux-commands-to-process-fastq-files-from-ngs-experiments/>

<https://bioinformaticsworkbook.org/dataAcquisition/fileTransfer/sra.html>

<https://github.com/ncbi/sra-tools/wiki/02.-Installing-SRA-Toolkit>

<http://www.ensembl.org/info/data/ftp/index.html> (To download reference genome or variantions)

<https://bioinf.shenwei.me/seqkit/faq/>

Low SK et. al,Genome-wide association study of breast cancer in the Japanese population. PlosOne 2013