# Reanalysis the BRCA data

The RAW bam data download from GDC is extracted to subbam for recheck. And the calling process was done through GATK. Ignoring the hard filter provided by the GATK. Now I walk through the whole process for the project.

The outline follows the instruction:  
preprocessing:

1. Download BRCA WXS data from [GDC]('https://gdc-portal.nci.nih.gov/projects/TCGA-BRCA').
2. Dwonload Regulatory Build from Ensembl (release-v86).
3. Call mutation at regulatory regions by GATK (**-L regulatoryRegion.bed**)

Filteration:

1. R script for obtain somatic mutation for every sample.

* 02.getSomatic.R
* manifest statistics for file size and mutation.
* **Normal: MQ > 20, DP >= 5, altDP >= 3**; write to 02.filtersnp
* **Tumor: MQ > 20, DP >= 10, altDP >= 3**; write to 02.filtersnp
* **Somatic:** ; write to 03.somatic/01.annotation

1. Bash script ANNOVAR to **filter dbSNP**

* removedbSNP.sh

1. Annotate with **regulatory region**.

* regulatoryFeture.py

1. Transform file to rds for analysis

* 03.loadAnnoSomatic.R

1. Reduce false positive site

* 04.somaticMutationAnalysis.R
* **normalMutation recurrent >=1**
* draw **01.SomaticMutationStat.png**
* **Filter position which recurrent >= 5**

1. Refined recurrent mutation.

* 37 mutation positions recur >=5 and save them to scan all mpileups.
* extractCoverage.sh extract position from all samples.
* The process is working for the ratio and ensure the position is a real somatic mutation.
* **Mutation point has count >= 10 and alt\_count >=3 in normal**
* **filter out normal false mtuation get depth >= 10 normal sample**
* **02.depth\_for\_tumor\_normal\_barplot.png**
* **03.mutated\_samples\_barplot.png**

1. Use biomaRt to get target genes for mutation

* 05.targetGeneForMutation.R get nearest target genes.
* mark as Oncogene or tumor supressor genes.

1. Target gene expression

* 06.targetGeneExpression.R calculate gene expression.

1. Recheck the mutation point.

* Get mpileup coverage extractCoverage.sh
* Load point mpileup file for recheck, the position cov don't attached with base quality.