Purpose :

- to localize public data – integrate the data from different resources for fast query
- implement existing scripts or develop new scripts for analysis of the data sets

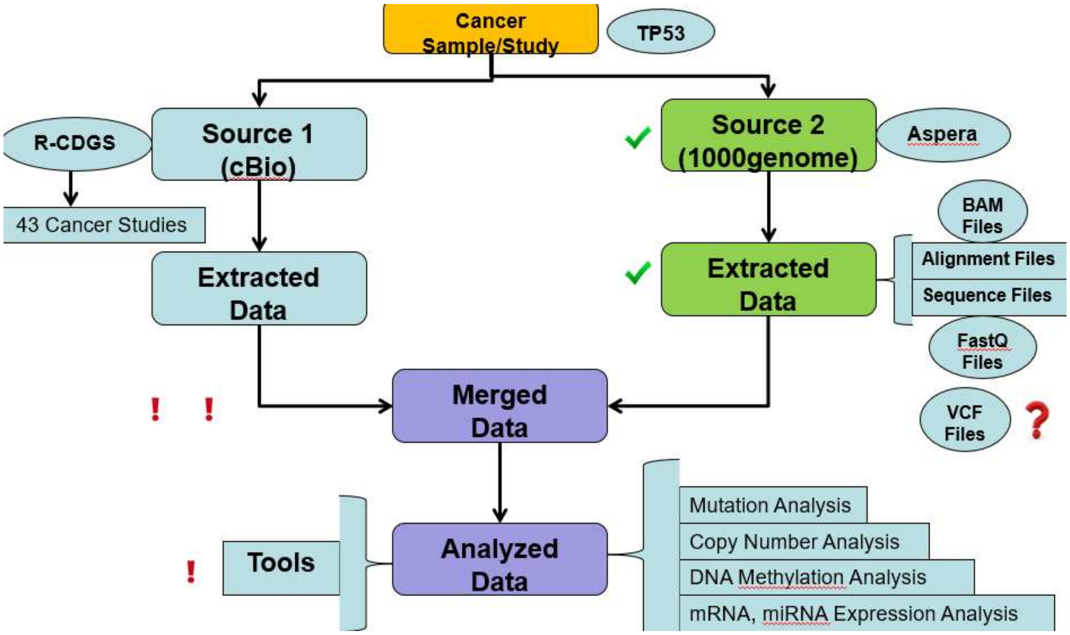
Public Resources :

- cBio@MSKCC
- TCGA Data
- NIH(Cancer)
- COSMIC Sanger database
- The Cancer Atlas
- Gene Expression Omnibus
- 1000 Genome Project

- Harness these datasets
- Add value or resolve specific problems
- Analyzing and deriving some meaningful information

Genes to focus :

- TP53
- CD47



Cancer Sample/Study

TP53

R-CDGS

Source 1 (cBio)

43 Cancer Studies

✓ Source 2 (1000genome)

Aspera

Extracted Data

✓ Extracted Data

BAM Files

Alignment Files

Sequence Files

FastQ Files

VCF Files ?

! ! Merged Data

! Tools

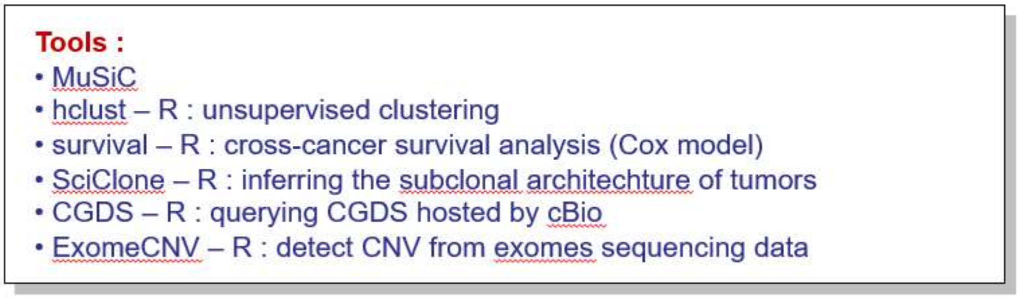
Analyzed Data

Mutation Analysis

Copy Number Analysis

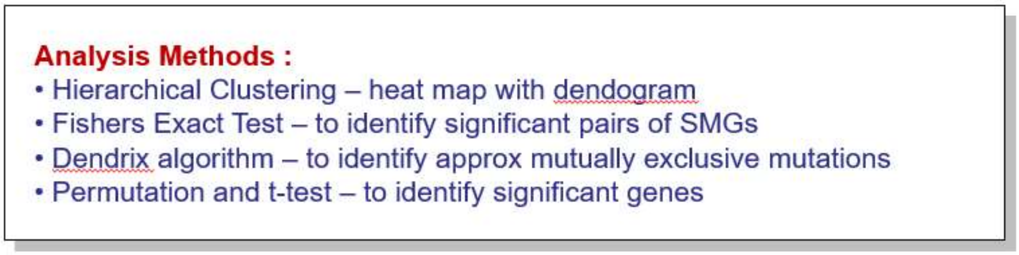
DNA Methylation Analysis

mRNA, miRNA Expression Analysis



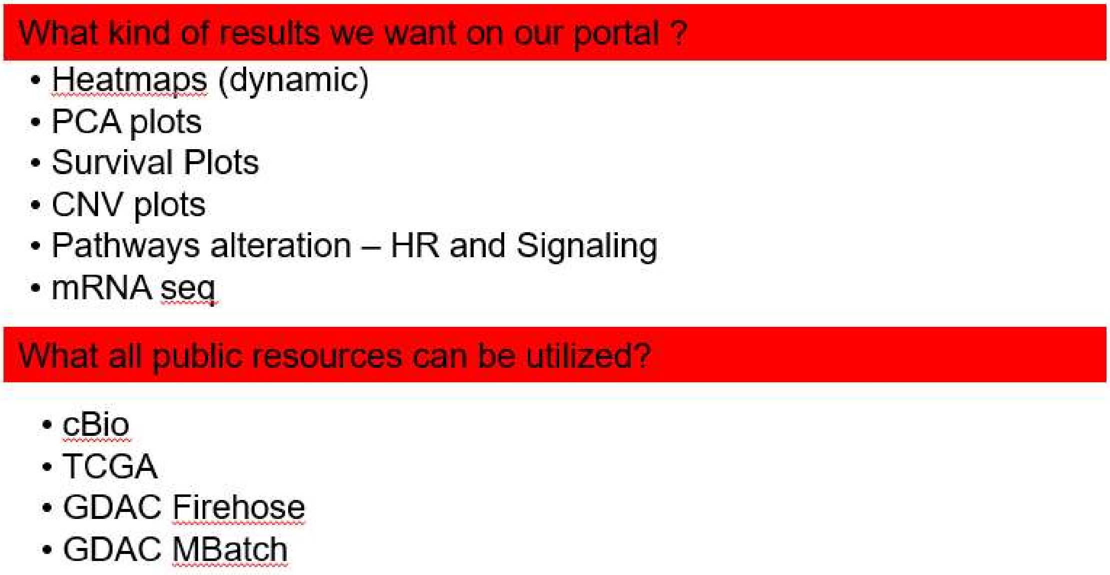
Tools :

- MuSiC
- hclust – R : unsupervised clustering
- survival – R : cross-cancer survival analysis (Cox model)
- SciClone – R : inferring the subclonal architechture of tumors
- CGDS – R : querying CGDS hosted by cBio
- ExomeCNV – R : detect CNV from exomes sequencing data



Analysis Methods :

- Hierarchical Clustering – heat map with dendogram
- Fishers Exact Test – to identify significant pairs of SMGs
- Dendrix algorithm – to identify approx mutually exclusive mutations
- Permutation and t-test – to identify significant genes



What kind of results we want on our portal ?

- Heatmaps (dynamic)
- PCA plots
- Survival Plots
- CNV plots
- Pathways alteration – HR and Signaling
- mRNA seq

What all public resources can be utilized?

- cBio
- TCGA
- GDAC Firehose
- GDAC MBatch