Bridging the gap between data scientist, clinicians and biologists using CRDC and CGC cloud resources

Dr. Daoud Meerzaman
Computational Genomics and Bioinformatics Branch (CGBB)
NCI Center for Biomedical Informatics and IT

April, 28, 2022
Computational Genomics and Bioinformatics Branch (CGBB) members
What does CBIIT-Computational Genomics and Bioinformatics Branch (CGBB) do?

The Computational Genomics and Bioinformatics Branch (CGBB)

- Comprehensive and integrative data analyses NGS and proteogenomic data
- Consulting and Training in NGS Data Analyses
- Develop bioinformatics tools and algorithms for genomics data analysis

Provides state of the art bioinformatics support for NCI Intramural and extramural communities. NCI initiative projects including (APOLLO) and (CIMAC).

https://myncti.cancer.gov/topics/computational-genomics-bioinformatics-branch-cgbb
Genomic Medicine and Bioinformatics Applications

https://www.slideshare.net/JTADrexel/bioinformatics-2512758?next_slideshow=2
https://www.youtube.com/watch?v=GLwCs370IGI
Transfer of information is Unidirectional and linear approach among experts
WGS is more suitable for FFPE samples than WES.

Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing

https://www.nature.com/articles/s41587-021-00994-5.pdf
WGS is more suitable for FFPE samples than WES.

Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing
Multidimensional Transfer of information among experts in genomic medicine pipeline
a network approach

Daoud Meerzaman, and
Barbara K. Dunn
Cancer Res 2019;79:5140-5145
Getting NCI scientists to use the tools and the resources available within the CRDC

The current bioinformatics tools require strong programming skills

Many scientists and clinicians do not have the expertise or the resources

CGC provides state-of-the-art bioinformatic expertise for physicians and biologists

Developing simplified UI-based tools, provide training, collaborations, etc.
Challenges for Data Analysis in Cancer Research

Huge amount of related data stored in different database, format and even in different system at different institutes.

Limited resources facing investigators to access (storage space and internet bandwidth) and analysis these data (lack of software tools and experienced bioinformaticians).

Different strategies during data generation such as genomic, proteomic transcriptomic and epigenomic. With the progress in cancer research, it is being increasingly acknowledged that a research question can not be answered by one form of omic analysis.
Getting NCI scientists to use the tools and the resources available within the CRDC

NCI CRDC, a cloud based data science infrastructure that hosts data from different projects and connect them with analytical tools to allow users to share, integrate, visualize and analyze these data.

Investigators can access up to petabytes of data on the cloud without physically download them to local system.

Investigators can analyze these data with analysis tools provided by the three cloud resources, FireCloud, ISB Cloud and Seven Bridge Cancer Genomic Cloud.
CGBB tested and evaluated all three cloud platform

Seven Bridge CGC (SBG)

- Access to many different data sets no need to download data
- Offers one on one demo and training to use the system.
- More than 600 analytical and bioinformatics tools and workflows readily available
- It requires minimal or no coding experience to use these tools and workflows.
FastQC and DESeq codes for RNA Seq Analysis

```r
# If doesn't need to do alignment, simply analyze the raw counts file
snvmi <- function(cpus, task, time) {
  library(edgeR)
  f <- read.delim("RawCountFile_rsemgenes.txt", header=T, row.names=1)
 grp <- split(colnames(f), ",")
  grps <- do.call(rbind, grp)
  colnames(f) <- paste0(grps[,3], ",", , grp, ,)
  g <- order(colnames(f))
  f <- f[g]
  grps <- do.call(rbind, grp)
  group <- grps[1,]
  cds$"DEGList(counts=f, group=group)
  dim(cds)
  keep <- filterByExpr(cds)
  cds <- cds[keep, , keep.lib.sizes=FALSE]
  dim(cds)
  # check data distribution
  dat <- log10(cdscounts + 1)
  pdf("data_boxplot.pdf", 8,6);
  par(mfrow=c(2,1), col.lab=0,0)
  boxplot(dat, main="log2 transformed counts", las=2)
  #quantile normalization
dat.m <- normalizeBetweenArrays(dat, method="quantile");
  boxplot(dat.m, main="quantile normalized raw counts", las=2)
  dev.off();
  # TMM normalization
cds <- calcNormFactors(cds)
  write.csv(as.matrix(cds), "data_filtered.edger.csv", quote=F, row.names=T);
  tam <- cpm(cds, log=T/2)
  write.csv(as.matrix(tam), "data_filtered_TMM.csv", quote=F, row.names=T);
  y <- cds
  cols <- rainbow(10, alpha=0.8);
  points <- c(55, 50)
  colors <- c(cols[1], cols[2], cols[3])
  pdf("fig1000.pdf")
  ```
CGC RNA Seq Pipelines
DESeq2 workflow

- Input archive file
- SBG Decompressor CWL 1.0
- DESeq2
- Normalized counts
- HTML report
- DESeq2 analysis results
- RData files
- pheno_out

PCA plot

MA plot
Moving Away from Reductionism to Integrative Biology and network Modeling Approaches

http://legomenon.com/russian-matryoshka-nesting-dolls-meaning.html


https://www.genome.jp/kegg-bin/show_pathway?hsa05200
Multi-omics gene-set analysis (MOGSA)

A multivariate single sample gene-set analysis method that integrates multiple experimental and molecular data types measured over the same set of samples.

Perform MOGSA Analysis in 3 Steps

Preparation of Inputs
A matrix of data types
At least of 2 datatypes

Determine the # of PCs based on MFA
To be used in Step3

Perform MOGSA analysis
How is MOGSA being implemented on CGC

App Info | Docker Repository
---|---
cgc-images.sbgenomics.com/data

**BASE COMMAND**

```r
Rscript
-e
'source("mogsa.R", echo=TRUE)'
```

- **Make Docker image**
- **Upload to Docker hub or SBG docker repository**
- **Pull docker image and create own tools**
- **Create workflow**
How is MOGSA working on CGC

Pull data using data browser

- mRNA TCGA
- Protein CPTAC
Intersect function in CGC
MOGSA Workflow Generation on CGC
MOGSA Workflow Generation on CGC
Multi-omics Gene-set Scores Analysis (MOGSA)

subtype-mean gene set scores
All RNA Protein Phosphoprotein

low high
OmicCircos is Next tool to be uploaded on the SBG

We will work cooperatively with support from Seven Bridges to implement more tools which will be useful to end users, such as OmicCircos, a R application package for generating high-quality circular plots for omics data developed by Ying Hu from our group, which has more than 36,700 downloads since it was first published in 2013.
Acknowledgement

• Seven Bridge
  • Dave Roberson
  • Manisha Ray
• CBIIT CGBB team
• Nguyen Trinh
• Xiaopeng Bian
• Qingrong Chen
• Chunhua Yan
• Ying Hu

CGC webinar occurring April 28th at 2pm.