COMPUTATIONAL GENOMICS WITH R
CHECKLIST

---

1. SOFTWARE INFORMATION AND CONVENTIONS

- ☐ Download the R packages needed to reproduce analysis of text using BocManager::install function
- ☐ Install the **compGenomRData** package via devtools::install_github("compgenomr/compGenomRData")
- ☐ When working with compGenomRData package obtain file paths using system.file()
- ☐ End of chapter solutions can be found at htps://gihub.com/compgenomr/excercises

2. HIGH THROUGHPUT EXPERIMENTAL METHODS IN GENOMICS

- ☐ Aim to quantify or locate all or most of the genome tha harbors biological feature of interest
- ☐ Most methods rely on some sort of enrichment of the biological feature
- ☐ RNA-seq (RNA Sequencing) experiments measure protein coding gene expression by extracting mRNA molecules w/special post-transcriptional alterations that protein coding genes acquire
- ☐ ChIP-seq experiments enrich for DNA fragments bound by protein of interest to find transcription factor binding (final product could be DNA or RNA)
- ☐ Microarrays were standard tool for quantification of nucleic acid fragments where complementary bases called "oligos" or "probes" designed to bind to the genetic material. Light signal produced if the genetic material is complementary to oligos w/light intensity proportional to amount of genetic material pairing with that oligo.
- ☐ Microarrays now being replaced with sequencing
- ☐ High Throughput Sequencing, or massively parallel sequencing, sequences thousands/millions of fragments at a time. Requires an enrichment step to enrich for feature we're interested in.
- ☐ Throughput refers to the number of sequenced bases per hour
- ☐ Library is RNA or DNA fragments to be sequenced

3. GENOMIC DATA VISUALIZATION

- ☐ 2 requirements: need a species w/presequenced genome and need annotation on that genome (at the very least know where genes are)

- ☐ Genome browsers are websites or apps that help you visualize genome and all available data associate with it. Allows visualization of proximity of genes to one another, gene structure, and SNPs. Example: http://genome.ucsc.edu, http://ensembl.org, and https://www.broadinstiute.org/igv/
- ☐ Genome sequencing data can be found in public archives like http://www.ncbi.nlm.nih.gov/geo/ and http://www.ebi.ac.uk/ena

Consortium	What is it for?
ENCODE	Transcription factor binding sites, gene expression and epigenomics data for cell lines
Epigenomics Roadmap	Epigenomics data for multiple cell types
The Cancer Genome Atlas (TCGA)	Expression, mutation and epigenomics data for multiple cancer types
1000 genomes project	Human genetic variation data obtained by sequencing 1000s of individuals

4. GENOMICS DATA ANALYSIS STEP: DATA COLLECTION

- ☐ Use publicly available data sets or high throughput experimental data

5. GENOMICS DATA ANALYSIS STEP: DATA QUALITY CHECK/CLEAN UP

- ☐ Process data into format suitable for exploratory analysis and modeling
- ☐ May need to transform data points (log transforming, normalizing etc.) or subset data se w/some arbitrary or pre-defined condition
- ☐ Must align reads to the genome and quantify over genes or regions of interest (count how many reads cover your region of interest)
- ☐ Once count obtained can normalize
- ☐ R/Bioconductor package has tools for sequencing read quality checks and HT-read alignments

6. EXPLORATORY DATA ANALYSIS AND MODELING

- ☐ Take in processed or semi-processed data an apply ML or statistical methods to explore data
- ☐ Want to see relationship between variables measured and a relationship between samples based on measured variables. Are the samples grouped as expected by experimental design? If there are outliers/anomalies may need additional clean-up processing.

- ☐ Predictive Modeling: may try to predict disease status of patients from expression of genes measured in tissue samples using regression based ML
- ☐ Statistical Modeling: using statistical methods like linear regression, conduct hypothesis testing, compare 2 data sets
- ☐ Unsupervised data analysis: clustering(k-means, hierarchical), matrix factorization(PCA, ICA, etc.)
- ☐ Supervised data analysis: generalized linear models, SVM, random forests

7. VISUALIZATION AND REPORTING

- ☐ Create figures tables and text describing outcome of anlaysis
- ☐ Want to see relationship between variables measured and a relationship between samples based on measured variables. Are the samples grouped as expected by experimental design? If there are outliers/anomalies may need additional clean-up processing.
- ☐ Predictive Modeling: may try to predict disease status of patients from expression of genes measured in tissue samples using regression based ML
- ☐ Statistical Modeling: using statistical methods like linear regression, conduct hypothesis testing, compare 2 data sets