Genomics and Genetics Data Skills

The amount of biological data sets, especially the next generation sequencing (NGS) data, has been accumulating astoundingly. As labs get access to big data, often it is in the hands of individual students. It has becoming critical to learn some data skills to manage NGS project, to track the project progress and to make the research more reproducible. However, most of the biological data are not adequately analyzed due to either the lacking of genomics/genetics tools or limited computational skills to analyze such data. The goal of this course is to teach students the biological data skills for quantitative genetic studies. Essentially, the course is composed of three components: 1) basic data science skills, 2) NGS data skills, and 3) quantitative genetic data skills.

At the end of this course, you will be able to:

- Manage NGS projects with full documentation and version control
- Process huge amount of sequencing data, perform quality checking, and summarize basic statistics
- Conduct short read alignment, including genomic sequencing, RNA-seq, and bisulfite sequencing data
- Learn SAM/BAM, VCF/BCF, BED format and get familiar with the tools for genomic variant analysis
- Conduct GWAS and GS experiments, Manhattan plot, population structure control

Schedule:

1. **Unix Shell**

   Basic unix commands, pipe, grep, vi, shell script, working with remote machines, unix data tools.

2. **Introduction to R and RStudio**

   Simple calculations in R, files IO, variable assignment, basic plotting, and using Rmarkdown for documentation.

3. **Git for Geneticist**

   Basic git, collaborating with git, working with branches, overleaf.

4. **Introduction to Python and Jupyter Notebook**

   Strings, lists, sorting, dicts and files. Regular expressions.

5. **Data Visualization (ggplot2)**

   Histogram, dot plot, beanplot, heatmap, correlation plot, **ggplot2**, **data.table**.

6. **Working with Sequence Data**

   Fasta, fastq, base qualities, indexed fasta files.

7. **Working with Alignment Data**

   SAM/BAM, mapping qualities, bwa mapping, samtools, bedtools.

8. **Working with Genetic Variants**

   GATK variant calling pipeline, bcftools, PLINK.

9. **SNP phasing and imputation**

   GBS SNP imputation, fastPhase, FILLIN, LD-based approach, cluster based approach and likelihood based approach.

10. **Quantitative genetics analysis tools**

    LD, IBD, GWAS, introgression, genetic variance partitioning.

11. **Genome Selection Approaches**

    Bayesian-based approaches, i.e. BayesA, BayesB, BayesC, BayesCpai, GBLUP.

Textbooks and resources:

- Buffalo, V., 2015. **Bioinformatics Data Skills: Reproducible and Robust Research with Open Source Tools**. " O'Reilly Media, Inc.".
- Newham, C. and Rosenblatt, B., 2005. **Learning the bash shell: Unix shell programming**. " O'Reilly Media, Inc.".
- Lutz, M., 2013. **Learning python**. " O'Reilly Media, Inc.".
- Teetor, P., 2011. **R cookbook**. " O'Reilly Media, Inc.".
- **Google's Python Class**, https://developers.google.com/edu/python/
- **R Programming on Coursera**, Johns Hopkins University, https://www.coursera.org/learn/r-programming/