

# Statistical Bioinformatics

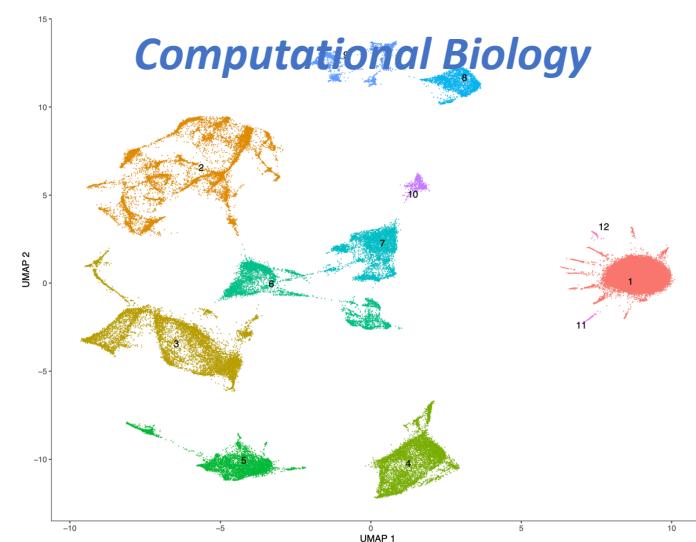
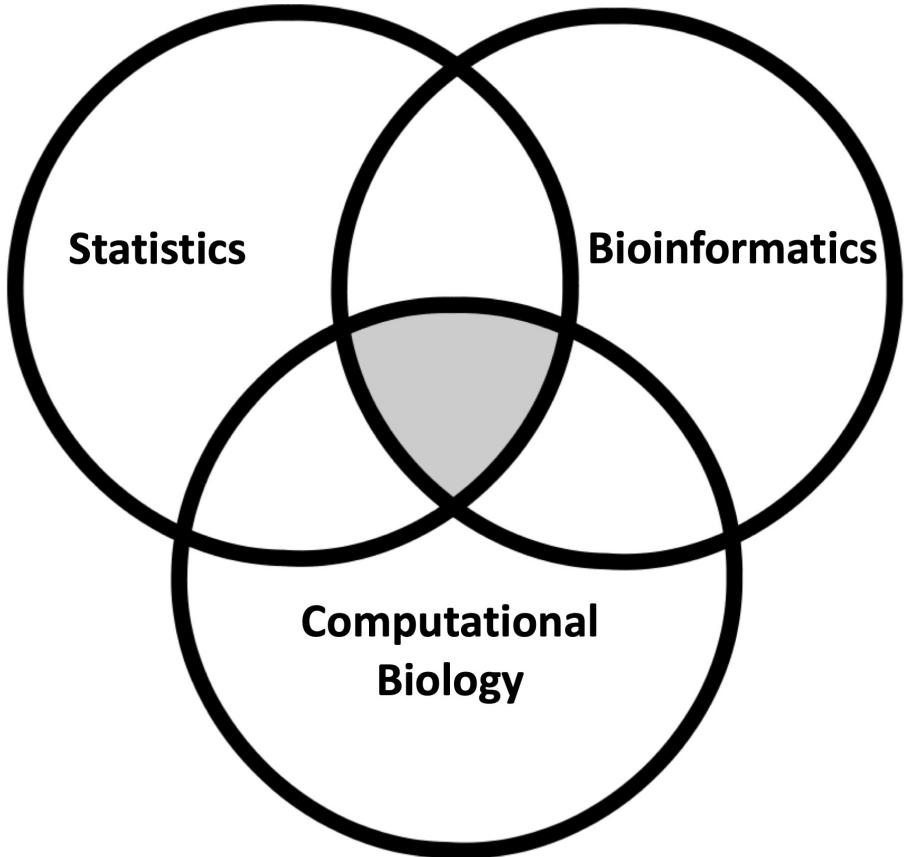
**Harry Feng**

Assistant Professor in Biostatistics

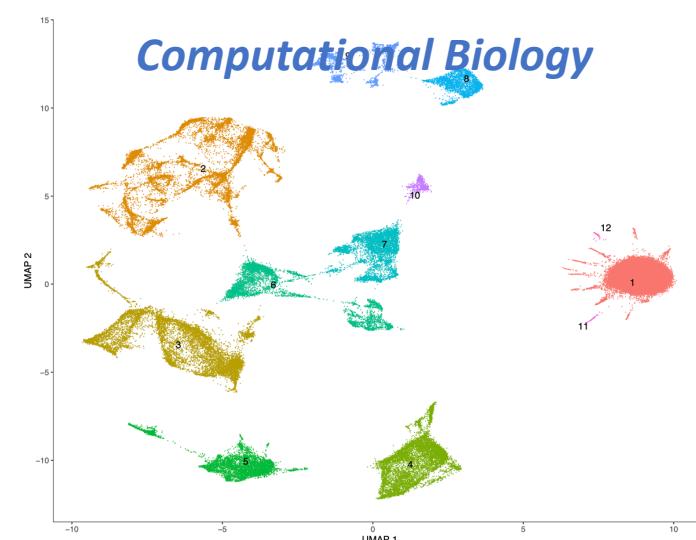
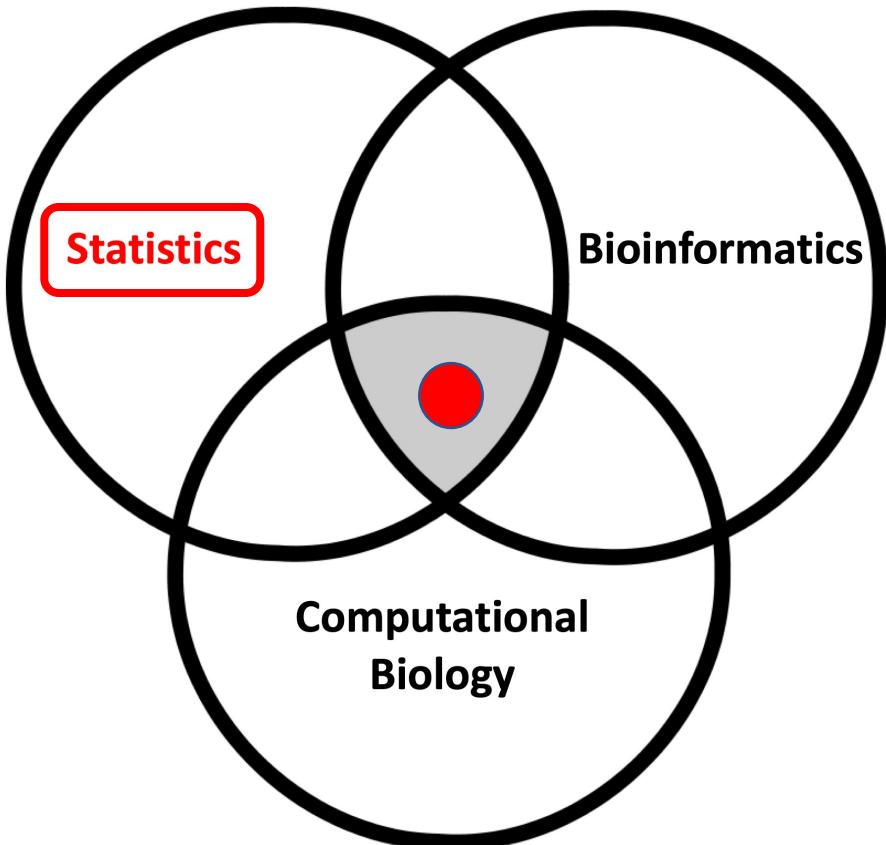
Department of Population and Quantitative Health Sciences (PQHS)

03/01/2024

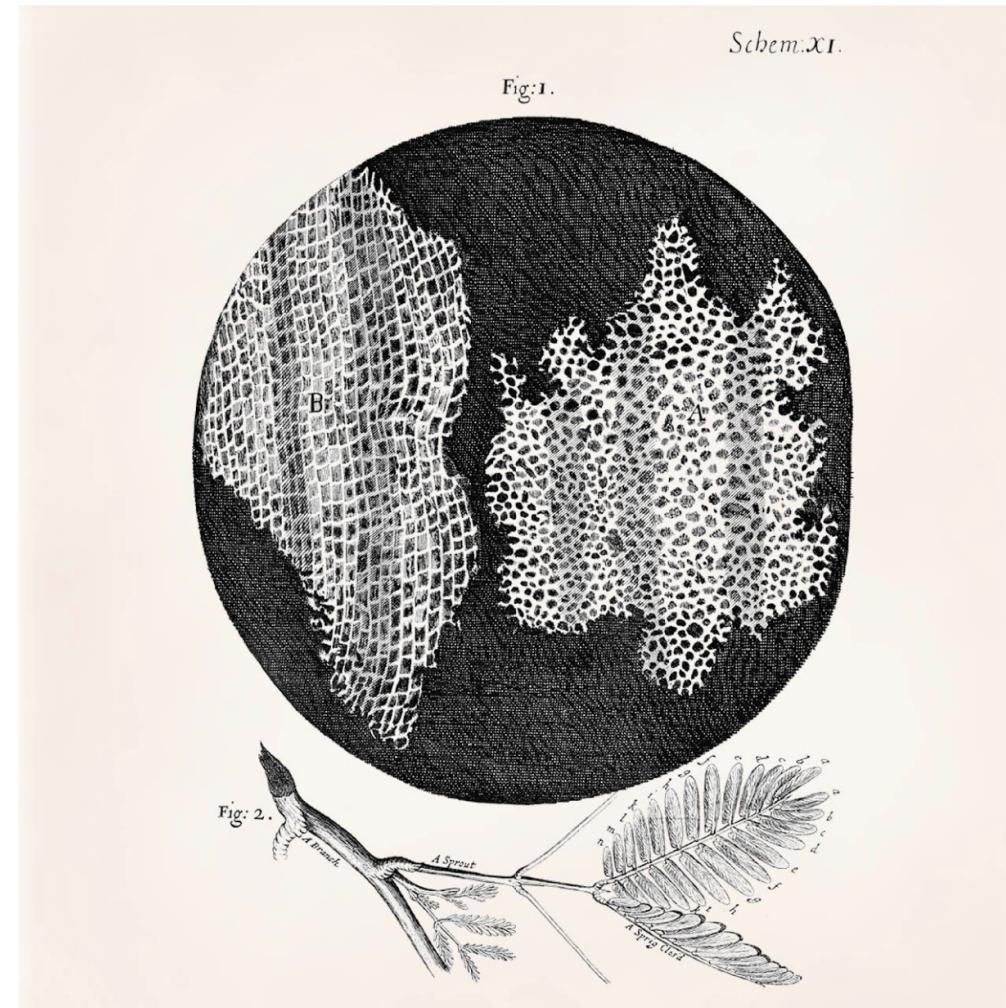
# My research



# My research

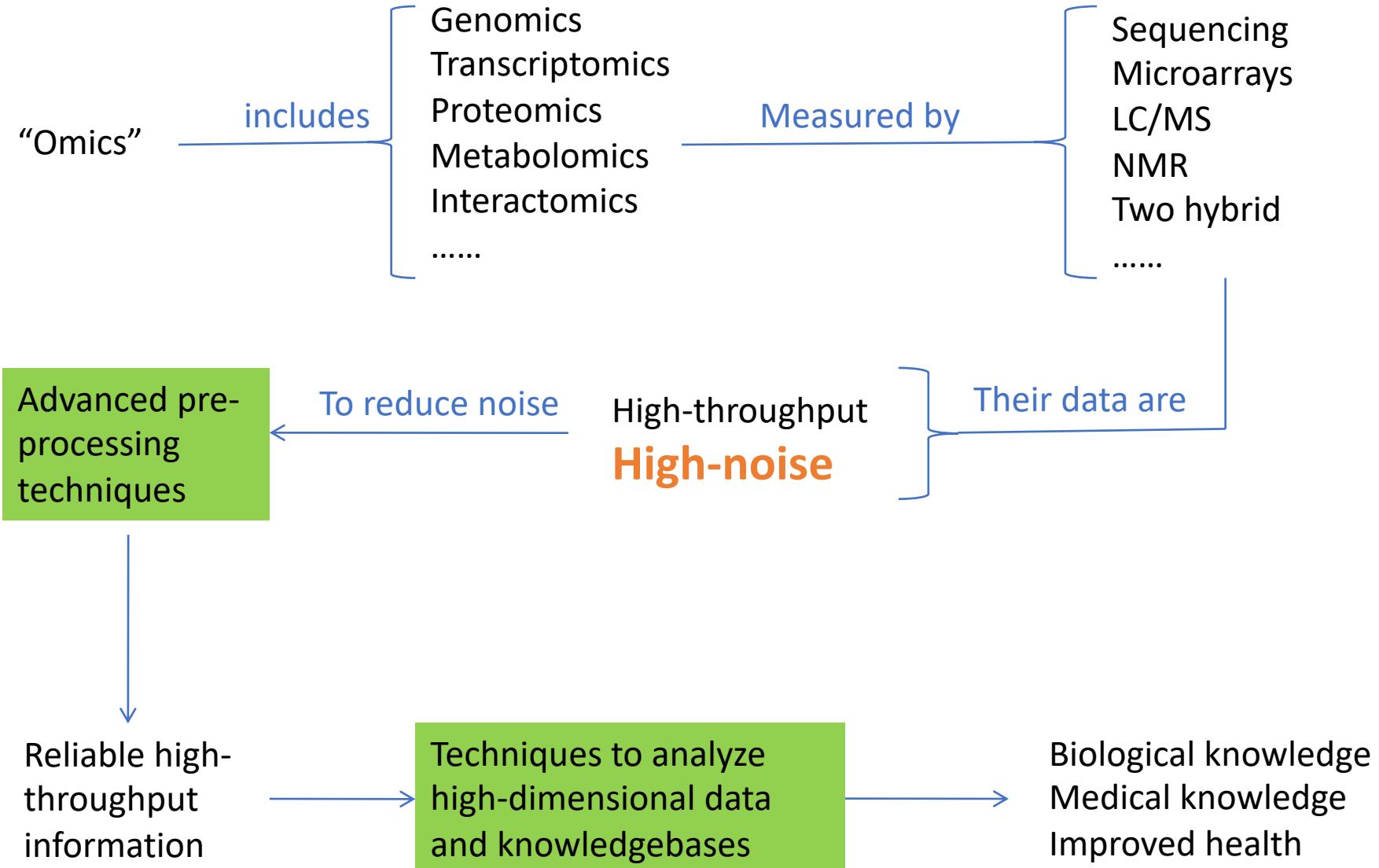


# The building block of life



Robert Hooke's drawing of cork cells. Image obtained from Micrographia.

# Transforming data to knowledge



# Coursework training

In addition to core courses at PQHS:

- ▶ Probability Theory (basic + advanced)
- ▶ Statistical Inference (basic + advanced)
- ▶ Linear Regression (basic + advanced)
- ▶ Generalized Linear Regression
- ▶ Bayesian Statistics
- ▶ Introduction to Bioinformatics
- ▶ Statistical Computing in R
- ▶ Machine Learning & Data Mining
- ▶ Deep Learning
- ▶ ...

Outside courses ARE encouraged!

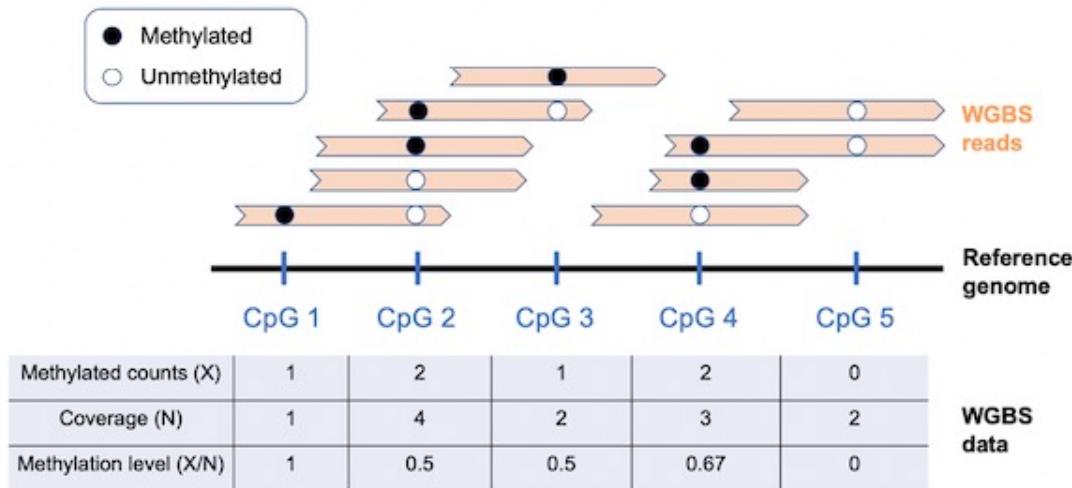
Statistics, Computer Science, Math...

# (Bio)Statistical methodology research

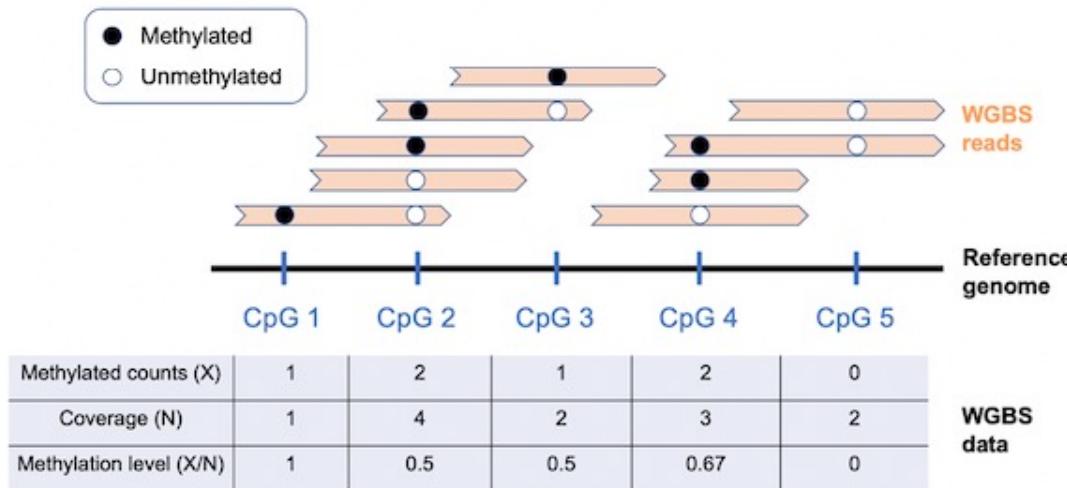
*“Development of mathematical formulas, models, and techniques that are used in statistical analysis of raw research data.”*

- New biotechnology
- Complex study design
- Advancement in data science
- Drawbacks of existing tools

# DNA methylation study



# DNA methylation study



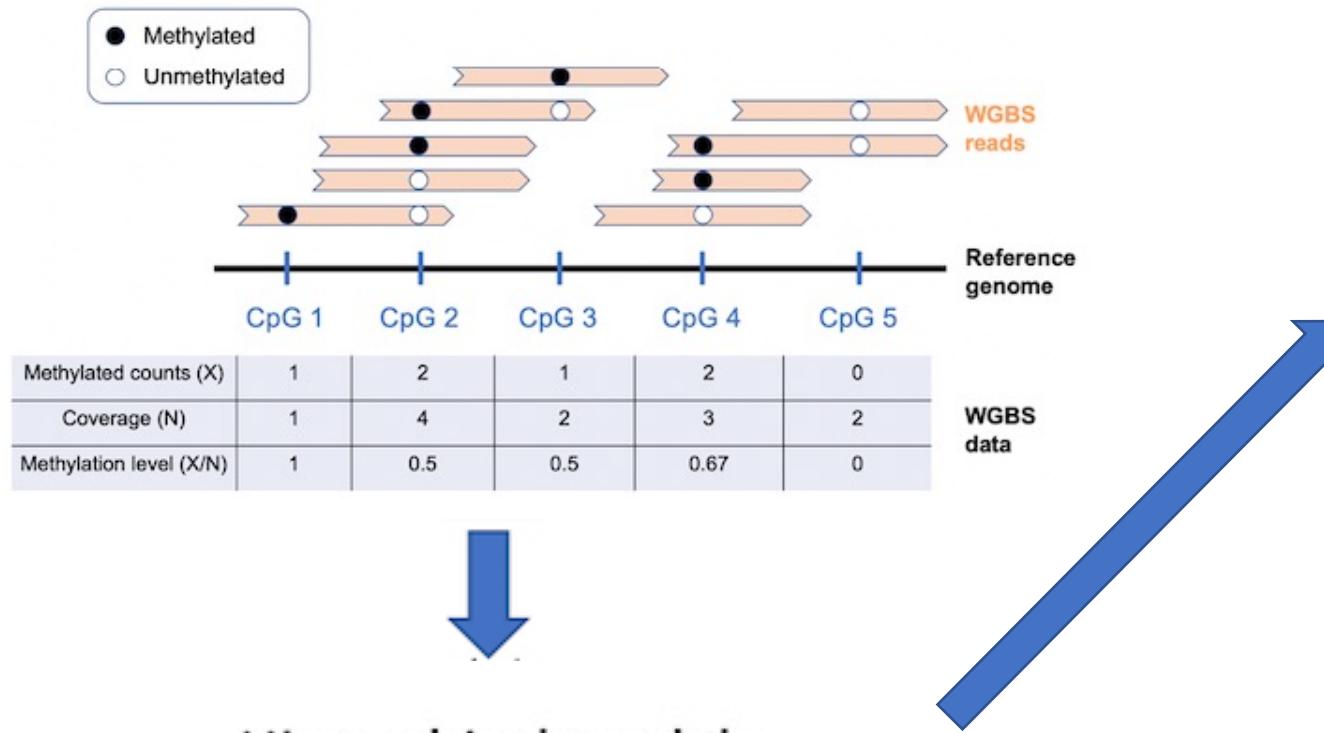
Hierarchical model:

$$X_{ijk} | p_{ijk}, N_{ijk} \sim \text{Binomial}(N_{ijk}, p_{ijk})$$

$$p_{ijk} \sim \text{Beta}(\mu_{ij}, \vartheta_{ij})$$

Prior:  $\vartheta_{ij} \sim \text{log-normal}(m_{0j}, r_{0j}^2)$

# DNA methylation study



Hierarchical model:

$$X_{ijk} | p_{ijk}, N_{ijk} \sim \text{Binomial}(N_{ijk}, p_{ijk})$$

$$p_{ijk} \sim \text{Beta}(\mu_{ij}, \phi_{ij})$$

Prior:  $\phi_{ij} \sim \text{log-normal}(m_{0j}, r_{0j}^2)$

- ① Estimate the methylation level  $\hat{\mu}_{ij}$  by combining all the replicates:  

$$\hat{\mu}_{ij} = \frac{\sum_k X_{ijk}}{\sum_k N_{ijk}}$$
- ② Adopt the method of moment estimator (MME) to obtain the estimation of  $m_{0j}$  and  $r_{0j}^2$  for prior.
- ③ Given the lognormal prior, estimate  $\phi_{ij}$  by Maximize A Posterior (MAP) on the conditional posterior distribution:

$$p(\phi_{ij} | X_{ij}, N_{ij}, \mu_{ij}) \propto f(\phi_{ij}) \prod_k g(X_{ijk} | N_{ijk}, \mu_{ij}, \phi_{ij})$$

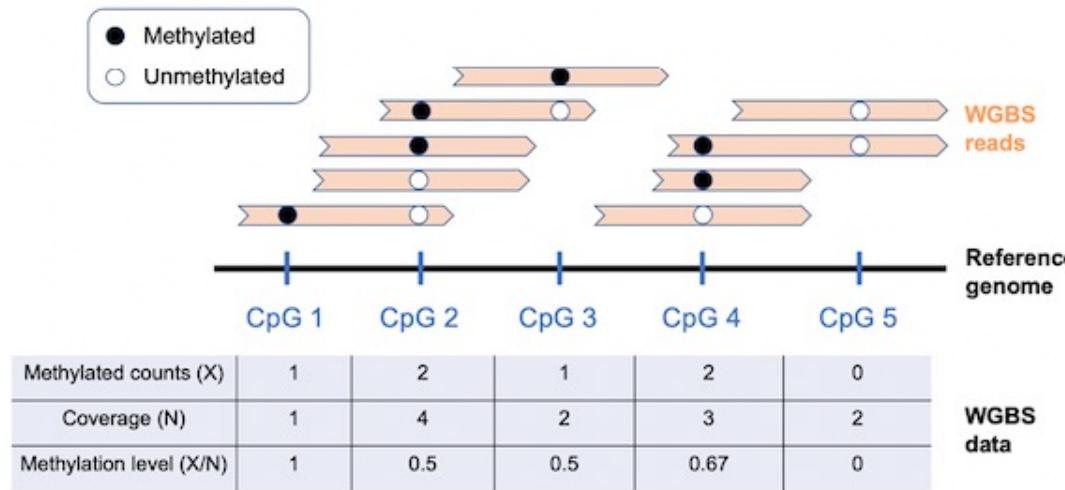
- ④ Calculate the variance of  $\hat{\mu}_{ij}$  by plugging in  $\hat{\mu}_{ij}$  and  $\hat{\phi}_{ij}$ :

$$\text{var}(\hat{\mu}_{ij}) = \left( \frac{1}{\sum_k N_{ijk}} \right)^2 \sum_k \{ N_{ijk} \hat{\mu}_{ij} (1 - \hat{\mu}_{ij}) (1 + (N_{ijk} - 1) \hat{\phi}_{ij}) \}$$

- ⑤ Perform the Wald test at each CpG site as:

$$t_i = \frac{\hat{\mu}_{i1} - \hat{\mu}_{i2}}{\sqrt{\text{var}(\hat{\mu}_{i1}) + \text{var}(\hat{\mu}_{i2})}}$$

# DNA methylation study



Hierarchical model:

$$X_{ijk} | p_{ijk}, N_{ijk} \sim \text{Binomial}(N_{ijk}, p_{ijk})$$

$$p_{ijk} \sim \text{Beta}(\mu_{ij}, \phi_{ij})$$

Prior:  $\phi_{ij} \sim \text{log-normal}(m_{0j}, r_{0j}^2)$

- ① Estimate the methylation level  $\hat{\mu}_{ij}$  by combining all the replicates:  

$$\hat{\mu}_{ij} = \frac{\sum_k X_{ijk}}{\sum_k N_{ijk}}$$
- ② Adopt the method of moment estimator (MME) to obtain the estimation of  $m_{0j}$  and  $r_{0j}^2$  for prior.
- ③ Given the lognormal prior, estimate  $\phi_{ij}$  by Maximize A Posterior (MAP) on the conditional posterior distribution:

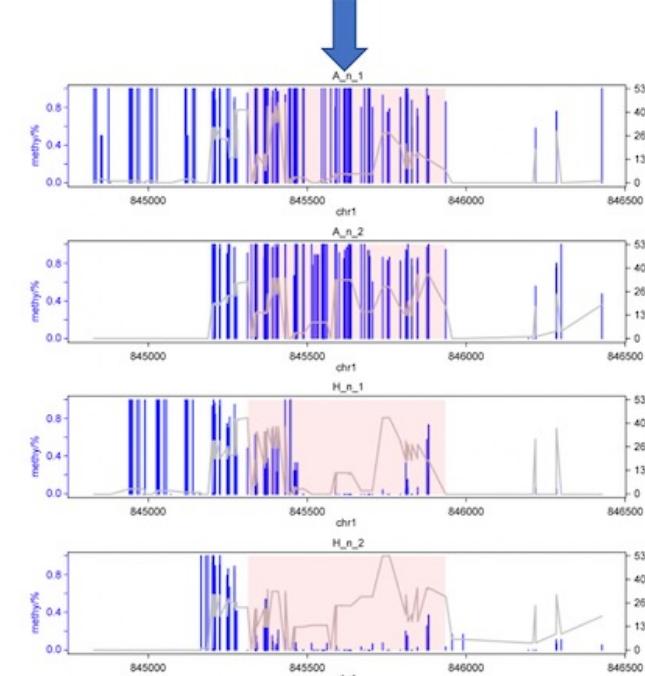
$$p(\phi_{ij} | X_{ij}, N_{ij}, \mu_{ij}) \propto f(\phi_{ij}) \prod_k g(X_{ijk} | N_{ijk}, \mu_{ij}, \phi_{ij})$$

- ④ Calculate the variance of  $\mu_{ij}$  by plugging in  $\hat{\mu}_{ij}$  and  $\hat{\phi}_{ij}$ :

$$\text{var}(\hat{\mu}_{ij}) = \left( \frac{1}{\sum_k N_{ijk}} \right)^2 \sum_k \{ N_{ijk} \hat{\mu}_{ij} (1 - \hat{\mu}_{ij}) (1 + (N_{ijk} - 1) \hat{\phi}_{ij}) \}$$

- ⑤ Perform the Wald test at each CpG site as:

$$t_i = \frac{\hat{\mu}_{i1} - \hat{\mu}_{i2}}{\sqrt{\text{var}(\hat{\mu}_{i1}) + \text{var}(\hat{\mu}_{i2})}}$$



# DNA methylation study

The screenshot shows the RStudio interface with the following components:

- Console Tab:** Displays the R startup message and basic usage instructions.
- Terminal Tab:** Shows the command `> library(DSS)` and its output, indicating the loading of required packages: Biobase, BiocGenerics, and parallel.
- Packages Tab:** Shows the "System Library" with a list of available R packages. The package `base` is checked, while others like abind, AICmodavg, ape, askpass, assertthat, backports, base64enc, batchelor, beachmat, beeswarm, BH, bibtex, Biobase, BiocGenerics, BiocManager, and BiocNeighbors are unselected.
- Viewer Tab:** Not visible in the screenshot.

```
R version 3.6.2 (2019-12-12) -- "Dark and Stormy Night"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> library(DSS)
Loading required package: Biobase
Loading required package: BiocGenerics
Loading required package: parallel
```

# DNA methylation study

R version 3.6.2 (2019-12-12) -- "Dark and Stormy Night"  
Copyright (C) 201  
Platform: x86\_64-  
  
R is free software.  
You are welcome to  
Type 'license()'!  
  
Natural language  
  
R is a collaborat  
Type 'contributor'  
'citation()' on h  
  
Type 'demo()' for  
'help.start()' fo  
Type 'q()' to qui  
  
> library(DSS)  
Loading required  
Loading required  
Loading required

Console Terminal × Jobs ×

Files Plots Packages Help Viewer

Install Update

Name	Description
System Library	

1 Introduction  
2 Using DSS for RNA-seq differential expression analysis  
3 Using DSS for BS-seq differential methylation analysis  
3.1 Overview  
**3.2 Input data preparation**  
3.3 DML/DMR detection from two-group comparison  
3.4 DML/DMR detection from general experimental design  
4 Frequently Asked Questions  
5 Session Info  
References

## 3.2 Input data preparation

DSS requires data from each BS-seq experiment to be summarized into following information for each CG position: chromosome number, genomic coordinate, total number of reads, and number of reads showing methylation. For a sample, this information are saved in a simple text file, with each row representing a CpG site. Below shows an example of a small part of such a file:

chr	pos	N	X
chr18	3014904	26	2
chr18	3031032	33	12
chr18	3031044	33	13
chr18	3031065	48	24

One can follow below steps to obtain such data from raw sequence file (fastq file), using **bismark** (version 0.10.0, commands for newer versions could be different) for BS-seq alignment and count extraction. These steps require installation of **bowtie** or **bowtie2**, **bismark**, and the fasta file for reference genome.

1. Prepare Bisulfite reference genome. This can be done using the **bismark\_genome\_preparation** function (details in bismark manual). Example command is:

```
bismark_genome_preparation --path_to_bowtie /usr/local/bowtie/ \
--verbose /path/to/refgenomes/
```

2. BS-seq alignment. Example command is:

```
bismark -q -n 1 -l 50 --path_to_bowtie \
/path/bowtie/ BS-refGenome reads.fastq
```

# DNA methylation study

```
Console Terminal × Jobs × ~/ ↗
```

R version 3.6.2 (2019-12-12) -- "Dark and Stormy Night"  
Copyright (C) 201  
Platform: x86\_64-  
  
R is free software.  
You are welcome to redistribute it.  
Type 'license()' or  
 Natural language  
  
R is a collaborat  
Type 'contributor  
'citation()' on h  
  
Type 'demo()' for  
'help.start()' fo  
Type 'q()' to qui  
  
> library(DSS)  
Loading required  
Loading required  
Loading required

Files Plots Packages Help Viewer

Install Update

Name	Description
System Library	

3.2 Input data preparation



Bioconductor  
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home Install Help Developers About

Home » Bioconductor 3.10 » Software Packages » DSS

## DSS

platforms all rank 151 / 1823 posts 3 / 0.3 / 0 / 0 in Bioc 7.5 years

build warnings updated before release dependencies 66

DOI: [10.18129/B9.bioc.DSS](https://doi.org/10.18129/B9.bioc.DSS) [f](#) [t](#)

### Dispersion shrinkage for sequencing data

Bioconductor version: Release (3.10)

DSS is an R library performing differential analysis for count-based sequencing data.

Search:

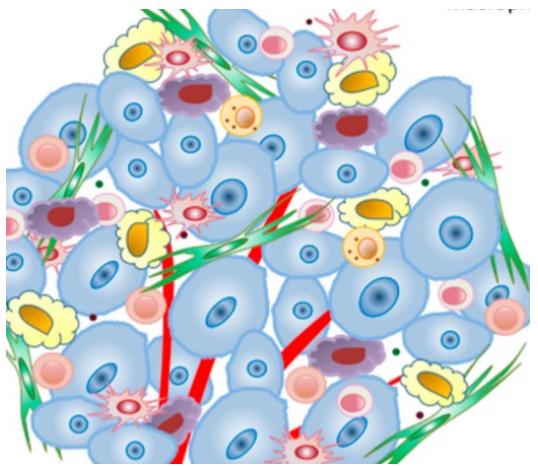
Documentation »

*Bioconductor*

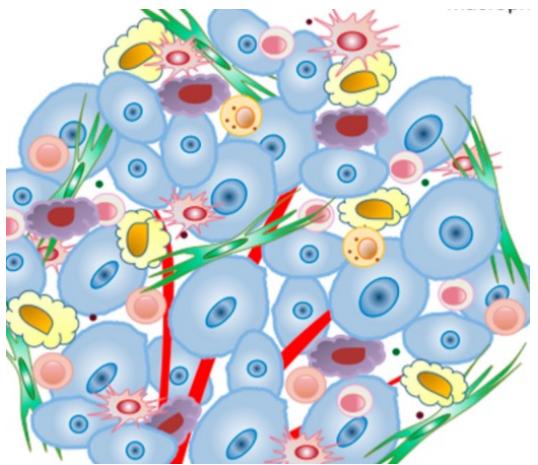
- Package [vignettes](#) and manuals.
- [Workflows](#) for learning and use.
- [Course and conference](#) material.
- [Videos](#).
- Community [resources](#) and [tutorials](#).

*R / CRAN packages and documentation*

# Cell mixture in tumor



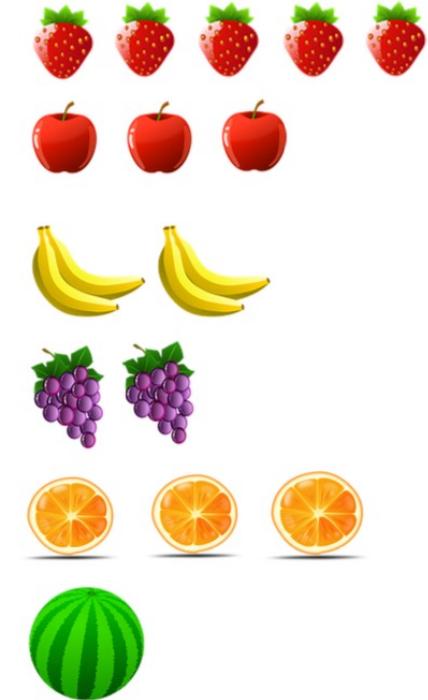
# Cell mixture in tumor



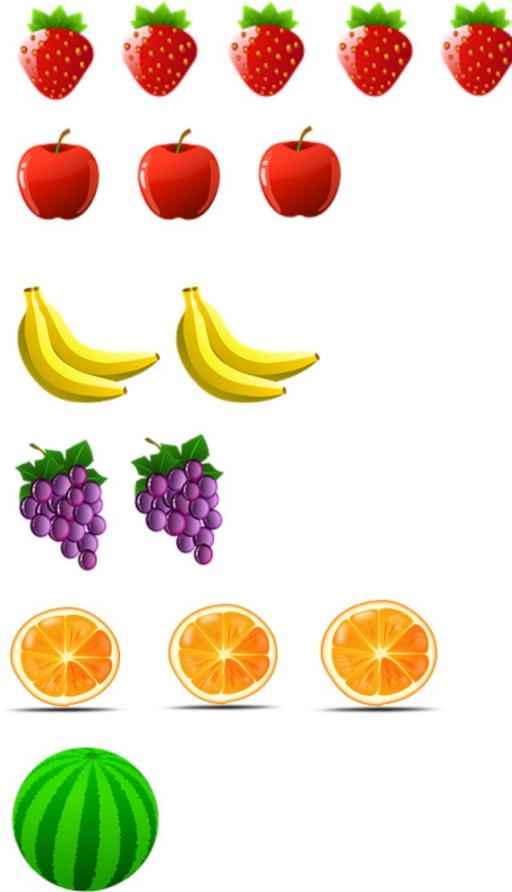
Bulk RNA-seq

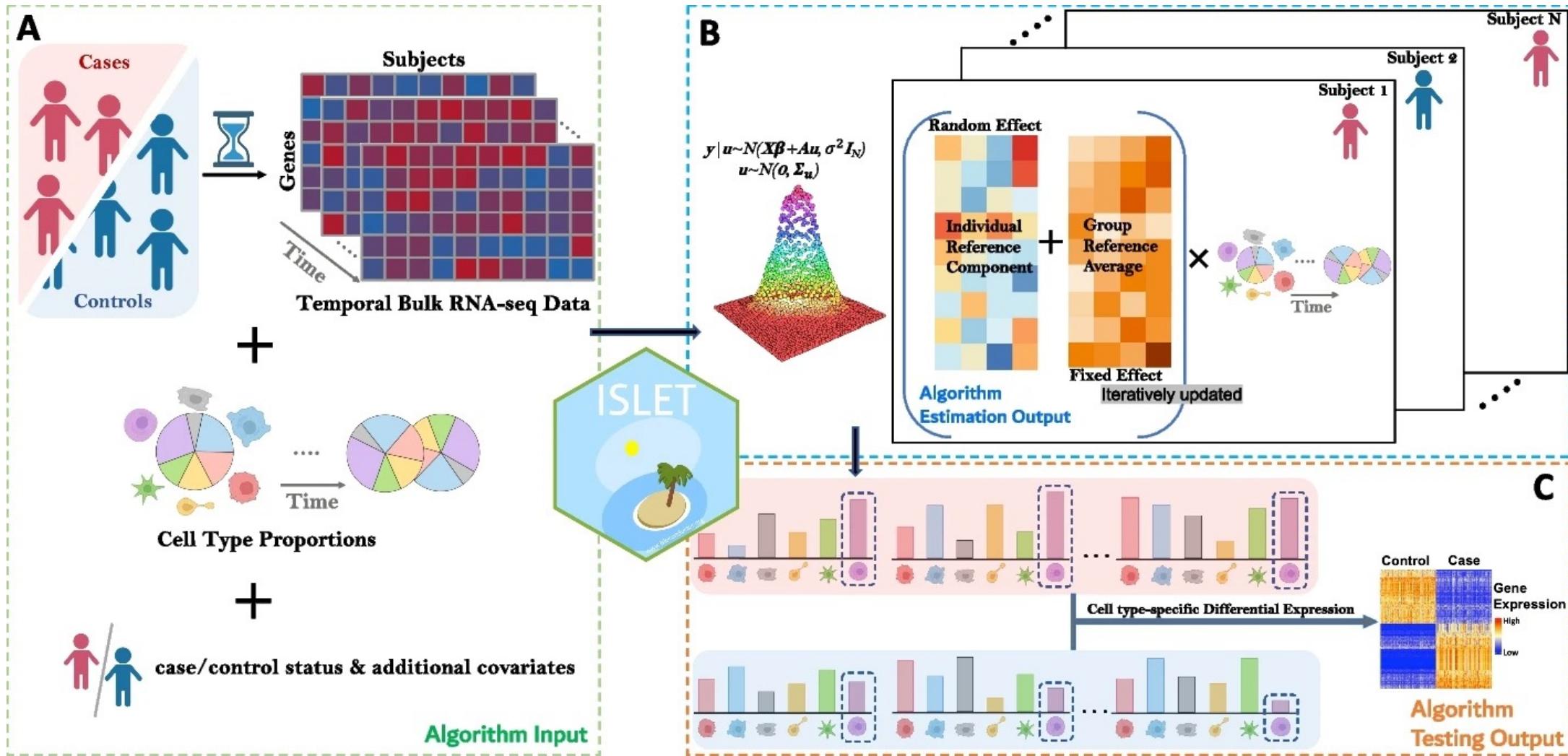


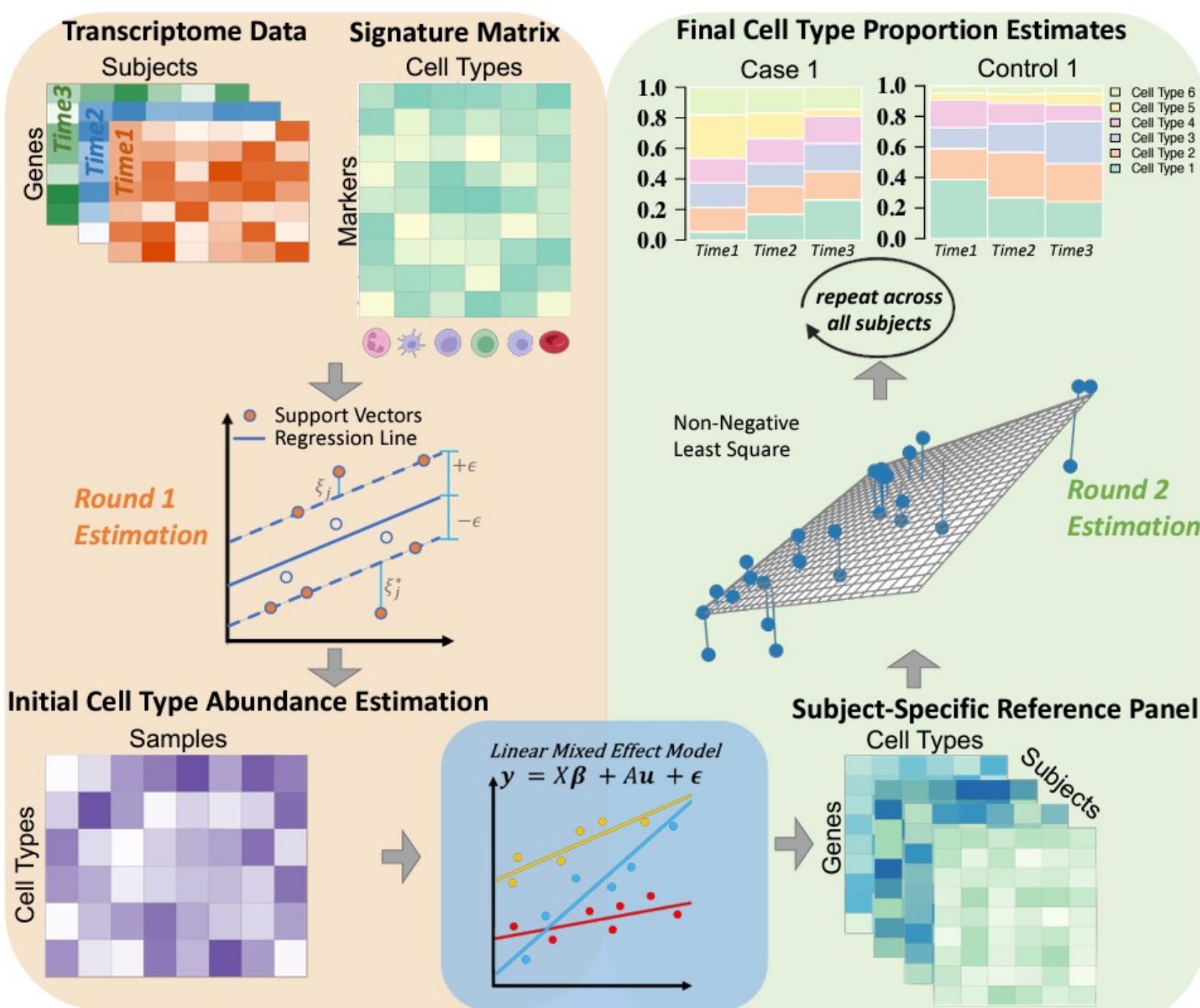
Single-cell RNA-seq

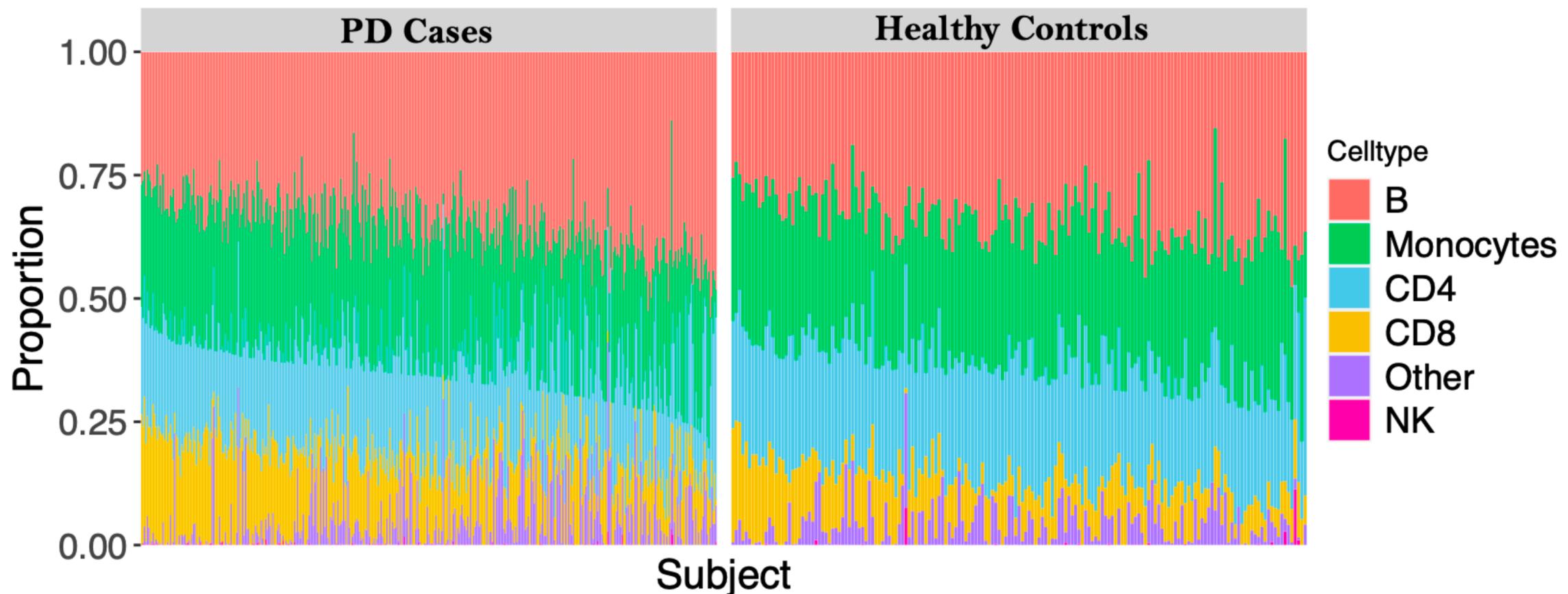


# Deconvolution









# Summary

- Bioinformatics data-driven research
  - Methodology AND Application
- Widely applicable
  - Various cancer/disease types
  - Different study designs



hxf155@case.edu



@HHarryFeng



<https://hfenglab.org/>