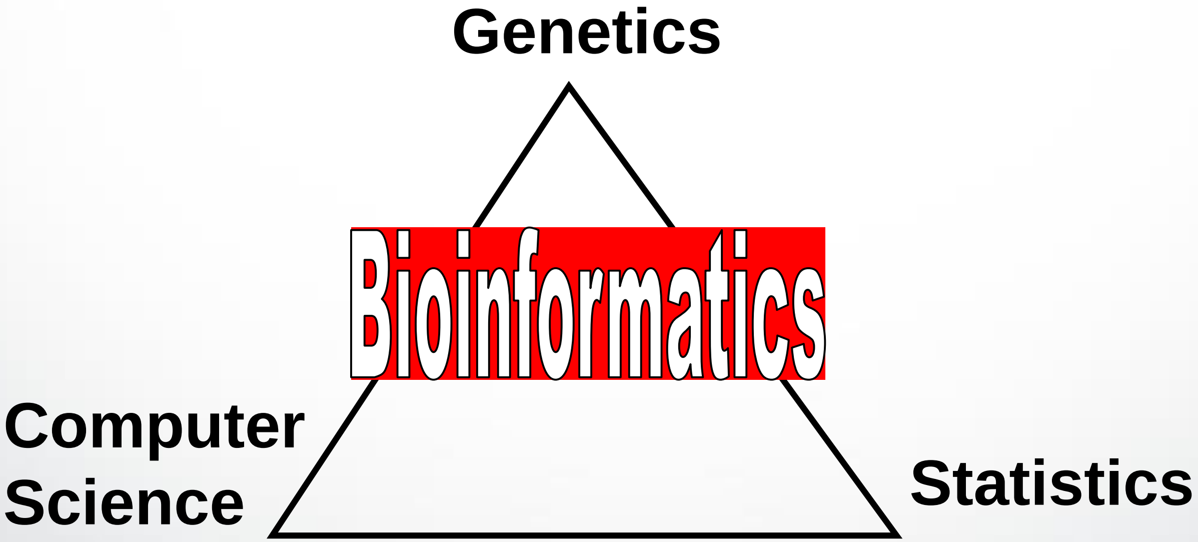
Research Projects

Jung-Ying Tzeng

September 26, 2018

About me

- Professor, Department of Statistics and Bioinformatics Research Center, North Carolina State University, Raleigh NC, USA
- Office: Ricks 305
- Courses: ST511, PP810, ST721 Genetic Data Analysis (2019 Fall)
- Research focus:
  - Statistical genetics and genetic epidemiology



Genetics

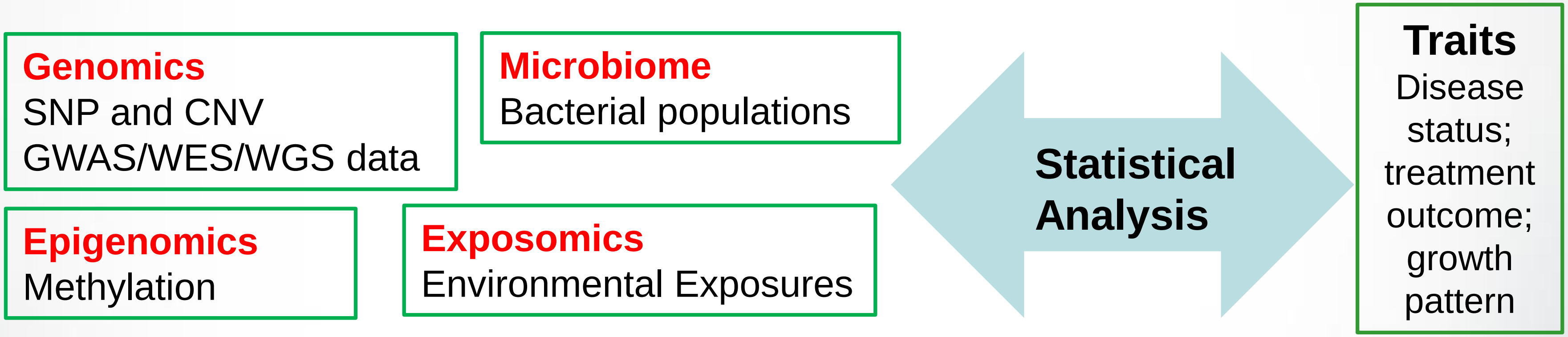
Bioinformatics

Computer Science

Statistics

Research Focus

- Focus heavily on developing and applying statistical methods for **association analysis** to understand the impact of genomic variables on traits



Genomics
SNP and CNV
GWAS/WES/WGS data

Microbiome
Bacterial populations

Epigenomics
Methylation

Exposomics
Environmental Exposures

Statistical Analysis

Traits
Disease status; treatment outcome; growth pattern

- Two general themes
  - To deal with predictors that are correlated, high dimensional, weak signals
    - E.g., Similarity based regression, random effects modeling, kernel machine regression, variance component tests, penalized regression
  - Dimension reduction, latent feature extraction
    - E.g., factor analysis, principal component analysis, penalized least square regression, tensor regression

Rotation projects

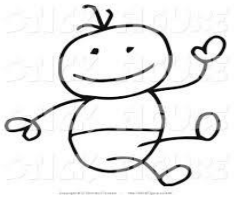
- Need to talk with you individually and assign projects based on your research interests and stat/programming background
- Typically involves
  - Literature review
  - Data analysis
  - Simulation studies
- Willing to learn and devote time
- Even greater if you know
  - Some programming skills, e.g., R
  - Basic statistic analysis approaches, e.g., regressions (and wouldn’t mind to take more advanced statistical courses)

Newborn Epigenetic Study (with PI Dr. Cathrine Hoyo)

Newborn Epigenetic Study (NEST)

- Goal: The impact of maternal exposure during pregnancy on child health

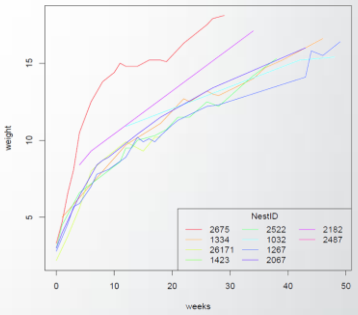
- EWAS/GWAS/mQTL
- GxE / MxE interaction
- Mediation analysis
- Functional data analysis
- Mendelian randomization
- EHR analysis
- PCA on data with excessive 0



GWAS data

Anesthesia exposure during laboring (from EHR)

Epigenetic profile from cord blood



Toxic metal
Trace metal

Child health (growth curve, asthma, obesity)

Microbiome data analysis

Microbiome association analysis

- Motivating study:

  A clinical trial to evaluate if home care can preserve the natural gut microbiota and decrease treatment-related mortality for hematopoietic cell transplantation (HCT)

  - Evaluate the association between longitudinal microbiome profile and treatment outcome
  - Identify specific taxa associated with outcome

- Challenges:

  a) Bacterial taxa abundance are measured in relative compositions (i.e., the abundance of all taxa in a sample sum to 1)

  b) Many taxa with zero counts within a sample

  c) Hundreds or thousands of taxa

  d) Taxa are phylogenetically related

Analysis on multi-platform data

Multi-platform Gene-Set Analysis

- **<u>“Meta”-based approaches</u>** can be applied on studies with different subjects
- **<u>Joint modeling approaches</u>** tend to have higher power and more informative than “meta”-based approaches
- Challenges of joint modeling methods: Moderate sample sizes with large number of variables

X_i for multi-platform data

Gene Platform	1	2	$\cdots$	G
1				
2				
$\vdots$				
P				

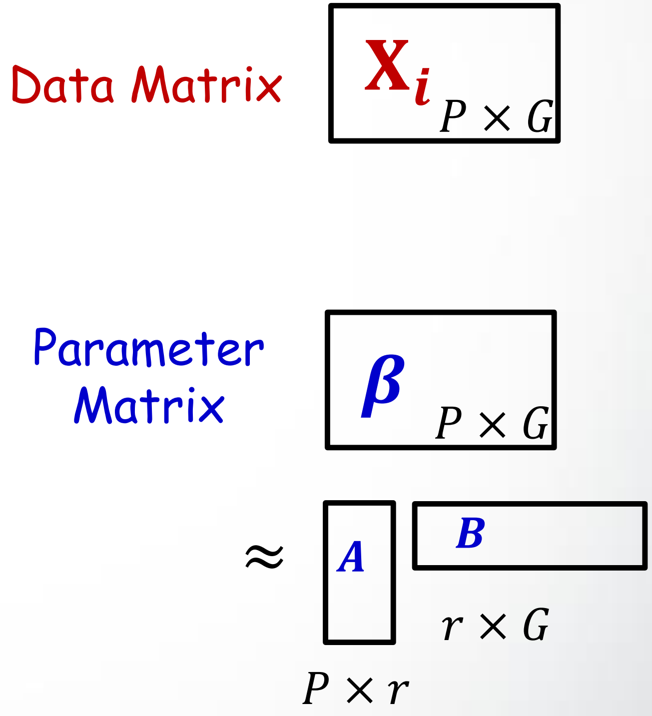
$$g(EY_i) = Z_i\gamma + \sum_{p=1}^{P}\sum_{g=1}^{G} X_{pg,i} \times \beta_{pg}$$

Low-Rank Tensor Model for Multi-platform Gene-Set Analysis

- Low-rank tensor regression (Zhou et al. 2013)
  - Genomic variable X_i of multi-platform data has a matrix structure
  - Parameter β_{pg}'s also form a $P \times G$ matrix, i.e.,

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{11} & \cdots & \beta_{1G} \\ \vdots & \cdots & \vdots \\ \beta_{P1} & \cdots & \beta_{PG} \end{bmatrix}$$

  - Can use low-rank approximation to $\boldsymbol{\beta}$ reduce the # of parameters



Data Matrix $\mathbf{X}_{\boldsymbol{i}}$ $P \times G$

Parameter Matrix $\boldsymbol{\beta}$ $P \times G$

$\approx$ $\boldsymbol{A}$ $P \times r$ $\boldsymbol{B}$ $r \times G$

Other projects

- Association analysis with rare CNVs
- GWAS with summary statistics, e.g., meta analysis, mediation analysis
- Rare variant association analysis incorporating 3D structural information
  - Variants within a gene --- incorporating 3D protein structure
  - WGS data --- incorporating Hi-C information

Thank you