SIMULATION OF CROSS-PHENOTYPIC EFFECTS OF RARE VARIANTS ACROSS TIME

PRATYAYDIPTA RUDRA

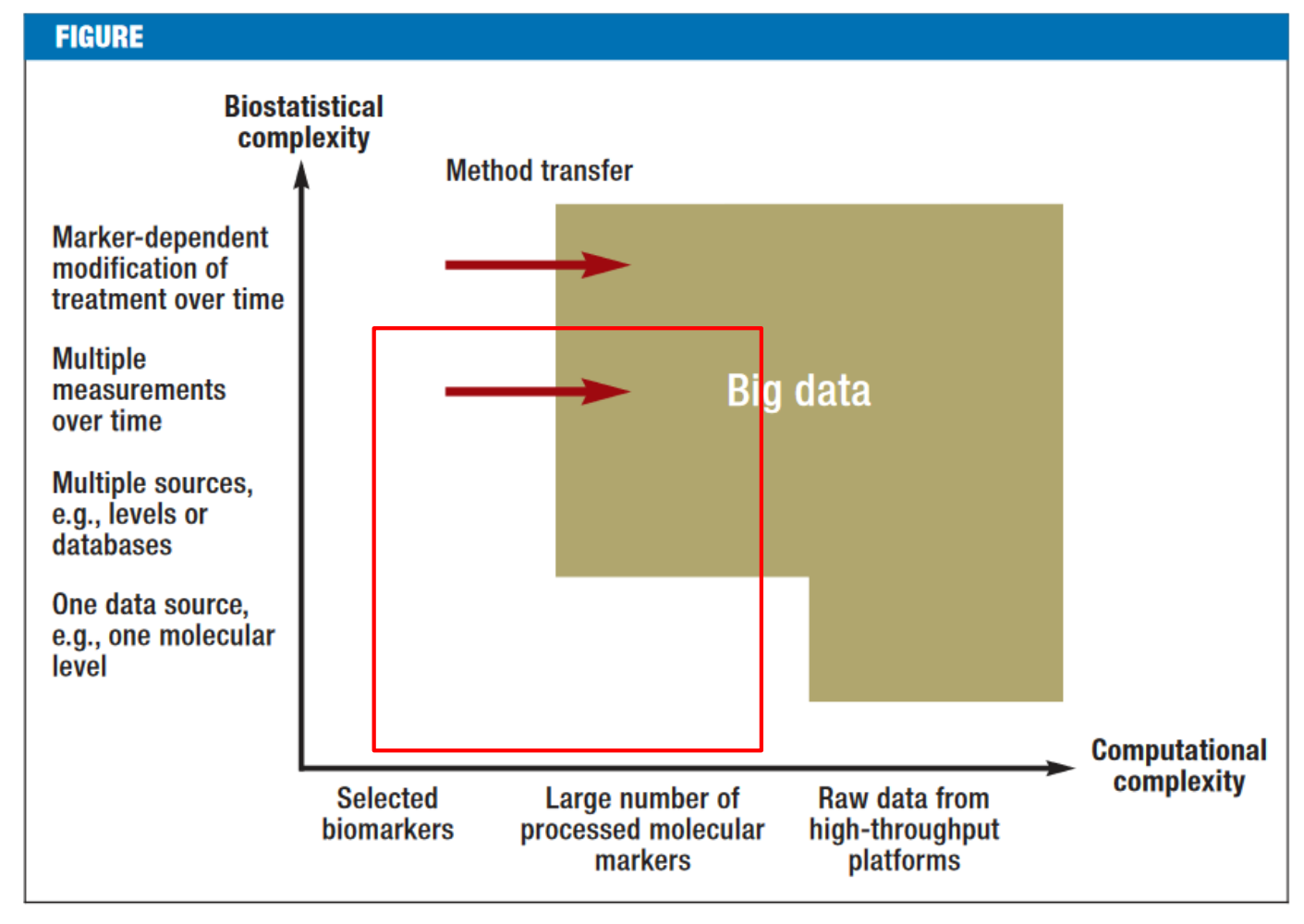
DEPARTMENT OF BIOSTATISTICS AND INFORMATICS

UNIVERSITY OF COLORADO, DENVER

HIGH THROUGHPUT COMPUTING AND BIG GENOMIC DATA

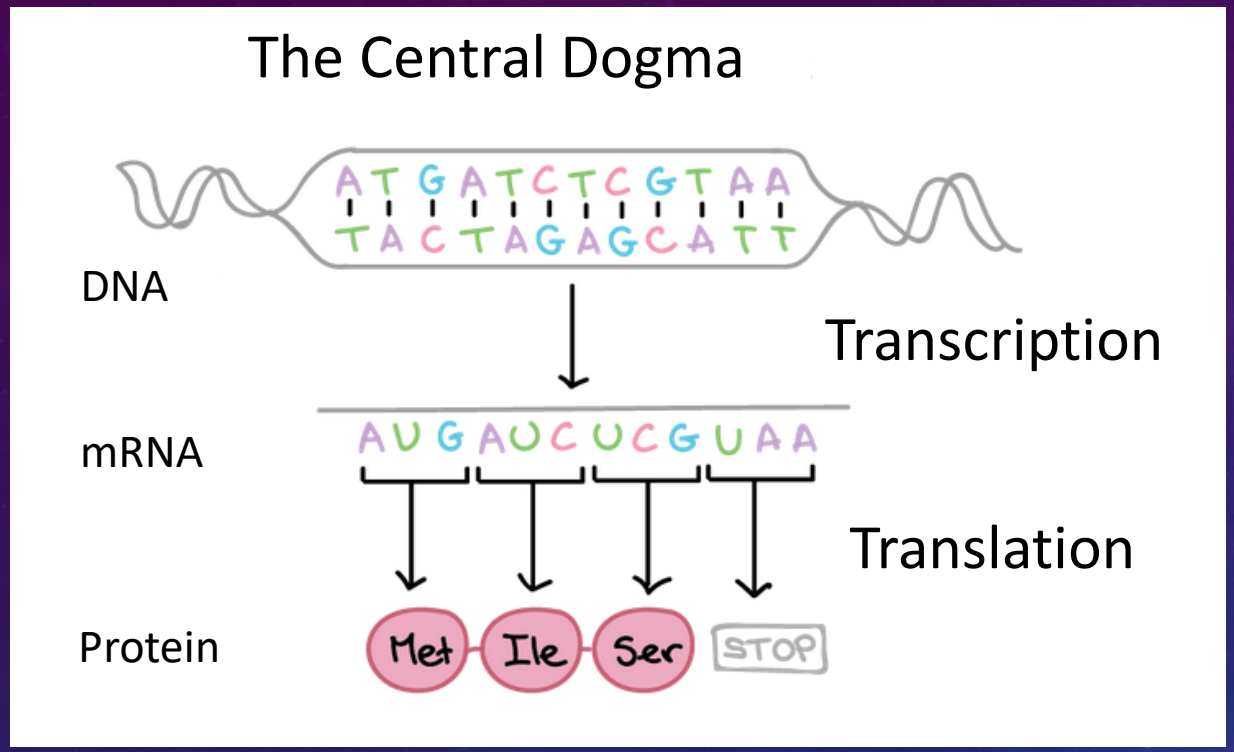
- Statistical Genomics: Potential to generate huge data sets.
  - Storage
  - Analysis
- Often needs a large memory just to read the data prior to analysis.
- Development of statistical methodology – simulation studies.

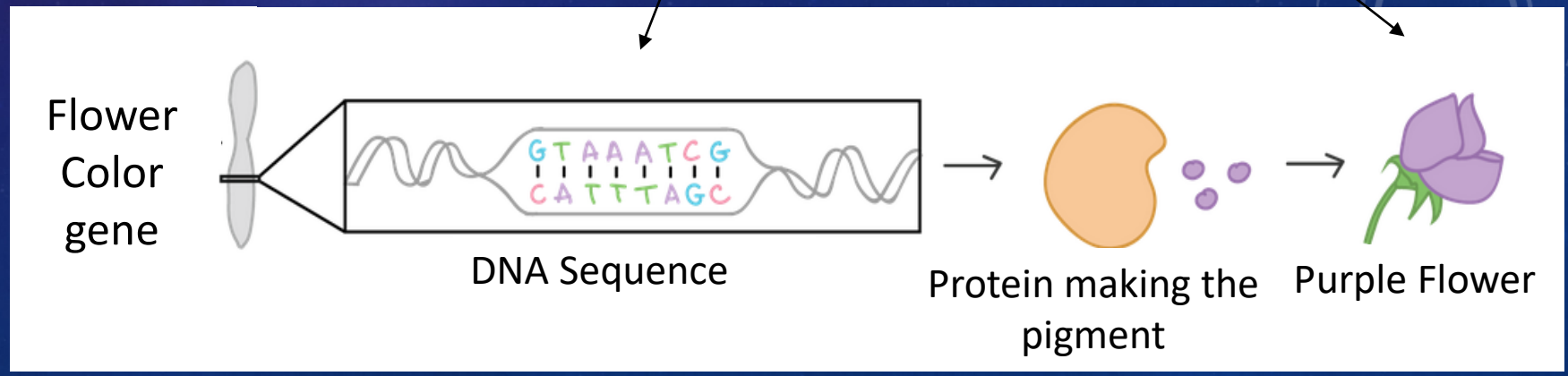
INCREASING SIZE vs INCREASING COMPLEXITY



Binder and Blettner, 2015. *Deutsches Ärzteblatt International*, *112*(9), p.137.

A BRIEF INTRODUCTION TO GENOMIC DATA

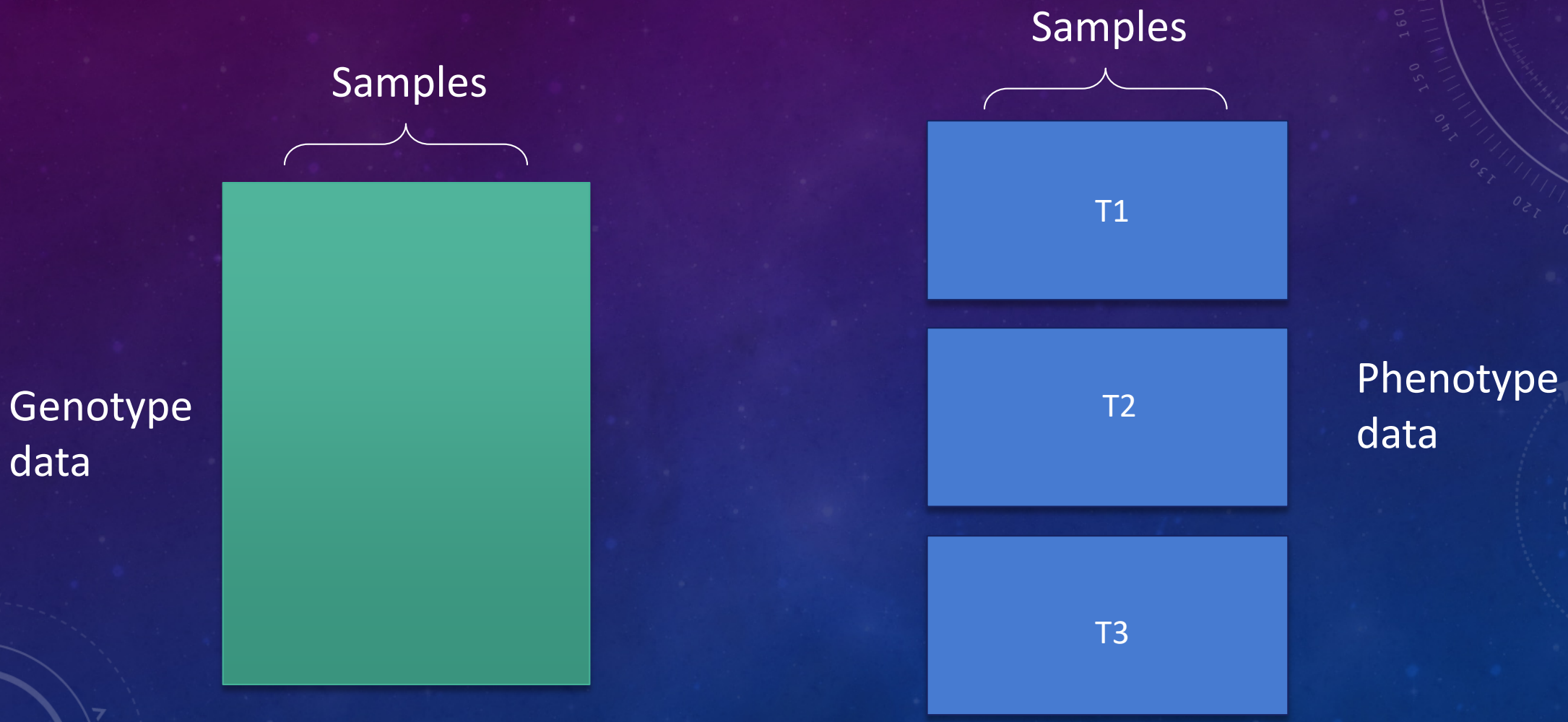




CROSS-PHENOTYPIC GENETIC EFFECTS

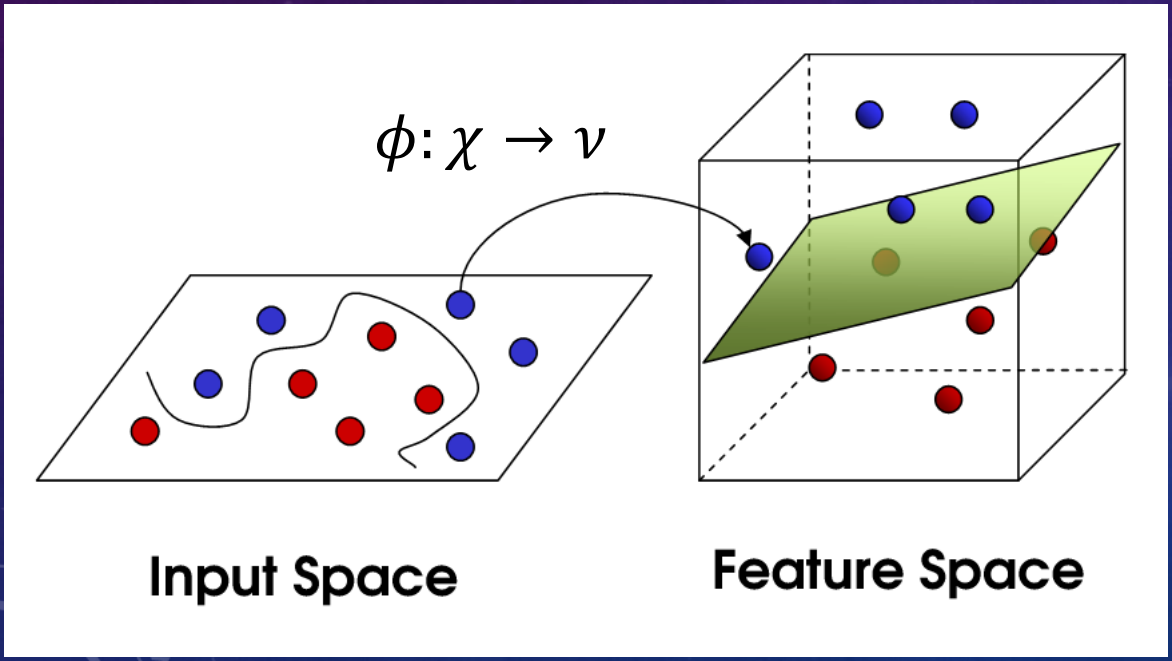
- Crossphenotypic association: When a genetic variant is associated with multiple phenotypes.
- Rare variants: Genetic variants that are present only in a small percentage of people.
- Rare variant testing: Statistical methods designed for common variants are usually less powerful for rare variants.
- Detection of cross-phenotypic effects can be enhanced by utilizing longitudinal phenotype data collected over time.

STRUCTURE OF THE DATA



LITTLE BIT ON THE STATISTICAL METHOD – THE KERNEL TRICK

- Kernel trick avoids the explicit mapping to get a linear boundary for classification problems, or a linear relationship for association problems.



A Kernel $K(x, y) = < \phi(x), \phi(y) >_\nu$ provides the inner product in the feature space.

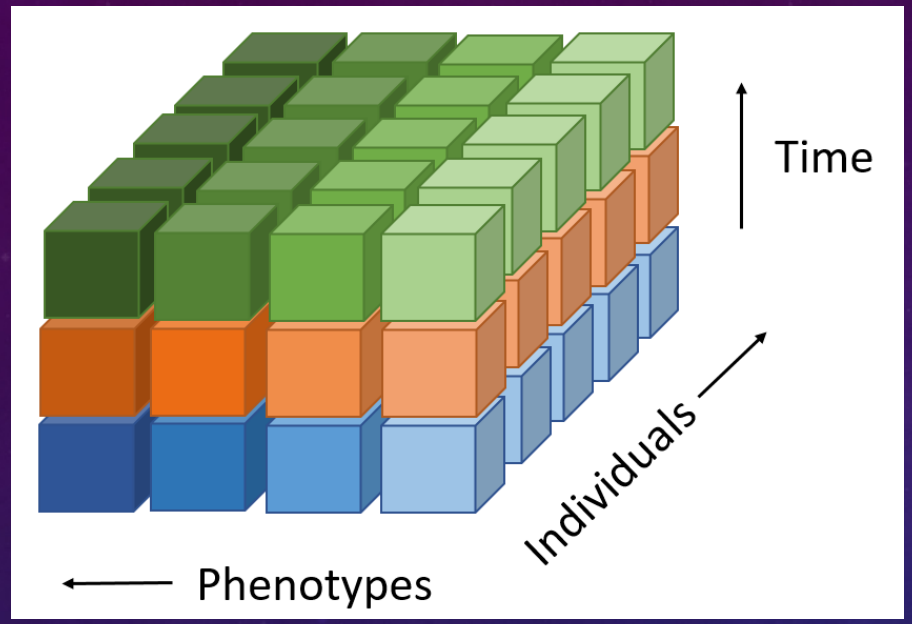
Any method that is based on certain inner product can be solved using this trick.

GENE ASSOCIATION WITH MULTIPLE TRAITS (GAMuT)

- A test based on Kernel Distance Covariance (KDC) framework. (Broadaway et al, 2016)
- It allows for arbitrary number of genotypes and phenotypes, and therefore it is ideal for testing rare variants.
- Obtains exact p-values using Davies' method. (Davies, 1980)
- Can be extended it for longitudinal data by making a simple modification. (Rudra et al, 2018)

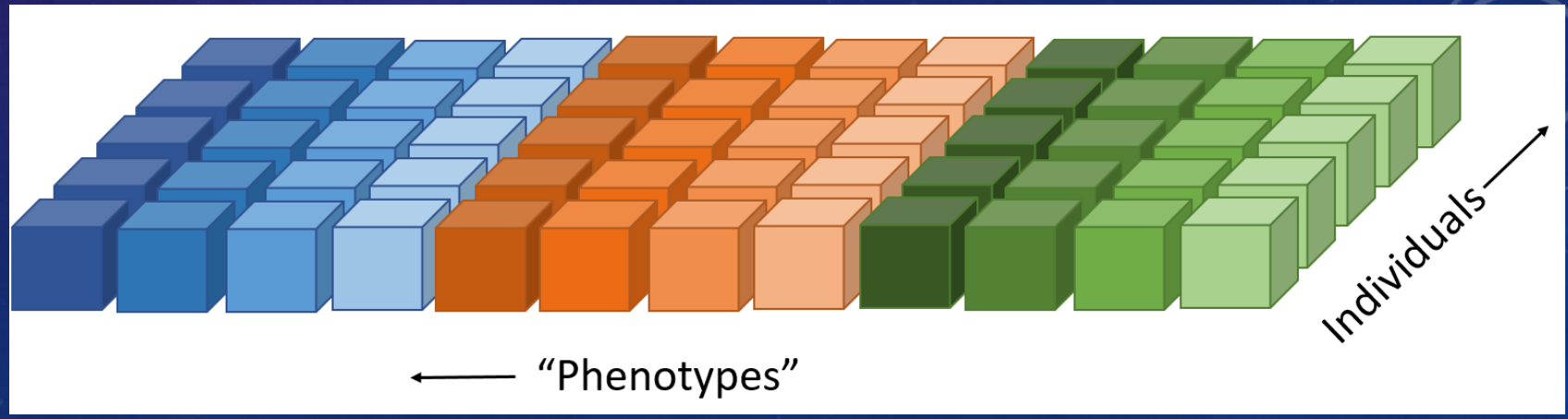
GAMuT FOR LONGITUDINAL DATA

Original data

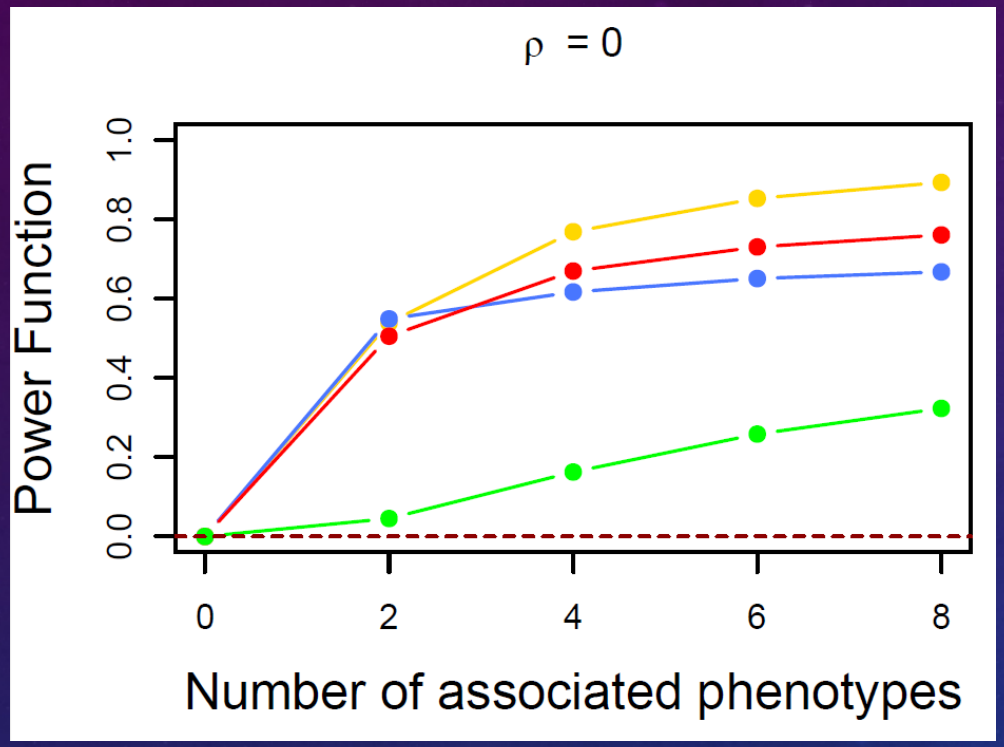


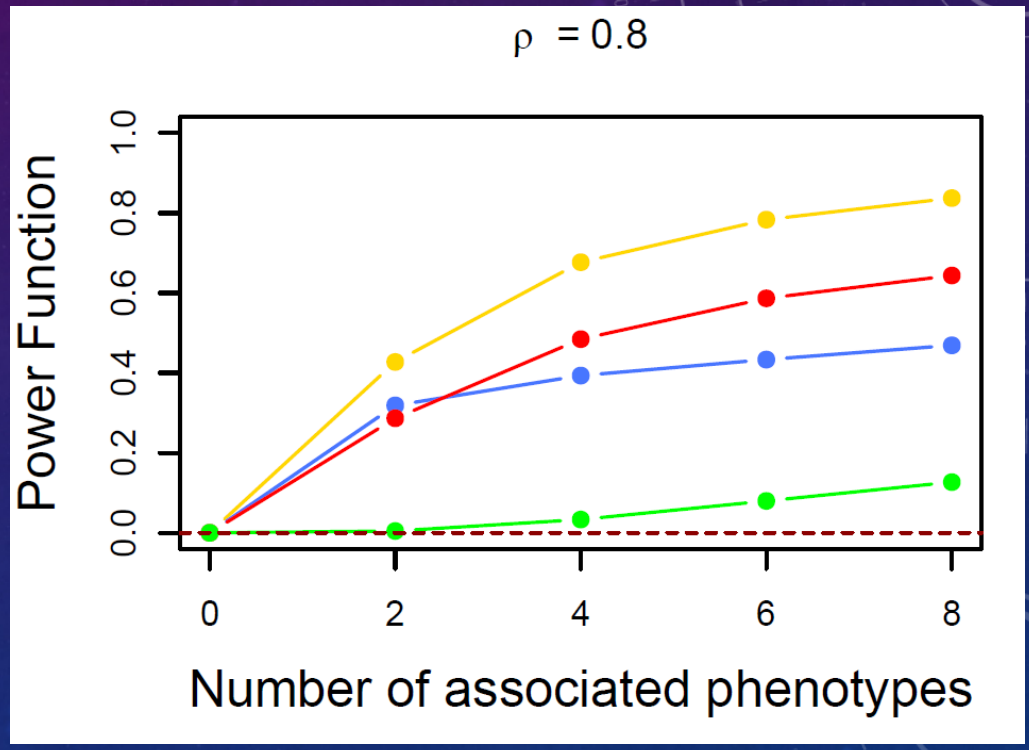
Slice the data at each time point and concatenate the slices treating the observations corresponding to same phenotypes for different time points as different phenotypes

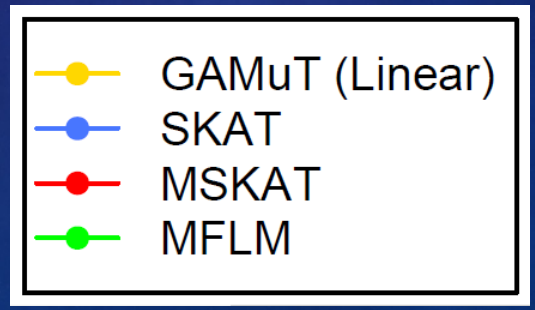
Concatenated data



SIMULATION STUDY TO COMPARE DIFFERENT METHODS







SIMULATING REALISTIC DATA: DIFFERENT PARAMETERS

- What are the correlations between the phenotypes?
- What are the correlations across time for the same phenotype?
- How many phenotypes are associated with a gene?
- How strong is the effect size?
- Is there a trend over time? If yes, what is the nature of it?

WHY HIGH THROUGHPUT COMPUTING?

- Many simulations – time consuming.
- 400 parameter combinations for simulations.
- 1000000 simulations/replications for each combination.
- 6 different statistical models to be fitted for each of the 400 x 1000000 replications.
- On a single machine, takes more than two years.

USING OPEN SCIENCE GRID

- OSG workshop at 2017 RMACC symposium.
- A large number of R-codes needed to be run for the simulations.
- Not much data management was required.
- Paper under revision, very little time.

USING OPEN SCIENCE GRID

- Set up code for each simulation scenario (parameter combination) and (manually) divide it into 100 jobs.
- Codes were more or less parallelizable. Each code is independent of the others, therefore suitable for HTC.
- Simple scripts used for job scheduling.

CHALLENGES

- I am no linux/unix expert.
- Workload management and job scheduling: I have used LSF and SLURM, but not HTCondor.
- Somewhat steep learning curve for using HTCondor in OSG.
- Understanding the error messages can be tough.

OVERALL EXPERIENCE

- Fairly fast and able to handle high memory jobs.
- Works great when each individual job is not too long.
- Awesome support!
- Certain bugs kept occurring and I was not able to resolve them with the help of the support.
- Job management could be tricky with the technique I was using.

FUTURE PLAN

- Use OSG for other research projects in statistical genomics.
- Learn to better resolve the bugs.
- Learn better ways of job scheduling.
- Use checkpointing in a better way.

RESULTS

- Was able to develop a statistical method for testing rare variant cross-phenotypic effects with longitudinal data.
- Used many simulations for different parameter combinations to show the favorable performance of our method compared to other methods.
- Performed some real data analysis using our method along with some resampling techniques.

Thank You!