Quantitative Genomics and Genetics
BTRY 4830/6830; PBSB.5201.01

Jason Mezey
Biological Statistics and Computational Biology (BSCB)
Department of Genetic Medicine
Institute for Computational Biomedicine
jgm45@cornell.edu

Cornell TA: Mahya Mehrmohamadi
mm2489@cornell.edu

WCMC TA: Jin Hyun Ju
jj328@cornell.edu

Spring 2014: Jan. 28 - May 10
T/Th: 8:40-9:55
Quantitative Genomics and Genetics

Professor: Jason Mezey
Biological Statistics and Computational Biology (Cornell)
Department of Genetic Medicine (Weill)

Time: Tues., Thurs. 8:40 am - 9:55 am

Room for Cornell, Ithaca: 224 Weill Hall

Room for WCMC: Belfer 204 B/C or Weill -Greenberg 2nd A/B

COURSE DESCRIPTION: A rigorous treatment of analysis techniques used to understand the genetics of complex phenotypes when using genomic data. This course will cover the fundamentals of statistical methodology with applications to the identification of genetic loci responsible for disease, agriculturally relevant, and evolutionarily important phenotypes. Data focus will be genome-wide data collected for association, inbred, and pedigree experimental designs. Analysis techniques will focus on the central importance of generalized linear models in quantitative genomics with an emphasis on both Frequentist and Bayesian computational approaches to inference.

GRADING: S/U or Letter Grade.

CREDITS: 4 (lecture + computer lab).

SUGGESTED PREREQUISITES: At least one class in Genetics and one class in probability and / or statistics.
Today

• Logistics (time/locations, registering, syllabus, schedule, requirements, computer labs, video-conferencing, etc.)

• Intuitive overview of the goals and the field of quantitative genomics

• The foundational connection between biology and probabilistic modeling

• Begin our introduction to modeling and probability
Times and Locations

- This is a “distance learning” class taught in two locations: Cornell, Ithaca and Weill, NYC
- I will teach all lectures from EITHER Ithaca or NYC (all lectures will be video-conferenced)
- I expect questions from both locations
- Lectures will be recorded:
  - These will be posted along with slides / notes
  - These will also function as backup (if needed)
- I encourage you to come to class...
Times and Locations II

- Lectures are (almost) every Tues. / Thurs. 8:40-9:55AM - see class schedule

- Ithaca lecture will always be 224 Weill Hall

- DEPENDING ON THE DATE, the Weill lecture location will be:
  - Belfer 204B or 204C
  - Weill-Greenberg, 2nd floor A or B
  - A spreadsheet will be made available with these locations (please read it carefully!!)
Times and Locations III

- There is a REQUIRED computer lab for this course (if you take the course for credit)

- Note that the computer lab for both Cornell and WCMC, the lab will meet 5-6PM on Thurs. (!!) - if you have an unavoidable conflict at this time, please send me an email (we will do our best to accommodate but...)

- In Ithaca will be taught by Mahya in room MNLB30A (!!) Mann library

- In NYC will be taught by Jin - same issue as class, depends on the week (see same spreadsheet...)

- Please bring your own laptop the first week (please email me if this is an issue)

- THE FIRST COMPUTER LAB IS NEXT WEEK (!!)
Times and Locations IV

- Office hours:
  - Jason will hold office hours on both campuses by video-conference each Thurs. 3-5PM - locations will be in 101 Biotech in Ithaca and in NYC, the main conference room of the Dept. of Genetic Med., 13th floor, Weill-Greenberg (subject to change!)
  - Mahya will hold office hours for Ithaca students only on Tues. 3-5PM in 101 Biotechnology Building
  - Jin will not have official office hours

- NOTE: unofficial help sessions can be scheduled with Jason, Mahya, or Jin by appointment

- NO office hours this week (!!)
Email list

- There is an official class email list that you must be on (officially registered or not): mezey-groupm-l@cornell.edu
- All information (short notice change in classrooms, homework announcements, etc.) will be distributed using this list (!!) so please make sure you are on it!
- To get on this list (or to be removed)
  - In Ithaca email Mayha: mm2489@cornell.edu
  - In NYC email Jin: jj328@cornell.edu
Website

- The class website will be under the “Classes” link on my site: http://mezeylab.cb.bscb.cornell.edu/
Website resources

• We will post information about the course and a schedule updated during the semester (check back often!!)

• There is no textbook for the class but I will post slides for all lectures

• I will post detailed notes for most lectures - there may be a significant delay for these posts (!!)

• There will also be supplementary readings (and other useful documents) that will be posted

• We will post videos of lectures and lecture slides (1-2 day delay in most cases)

• We will post all homeworks, exams, keys, etc.

• We will post slides for the computer labs and code
Registering for the class I

- You may take this class for a letter grade, S/U, or Audit

- If you can register for this class, please do so (even if you plan to audit!!)

- If you cannot register (you are a student at MSKCC, have a conflict, you are a postdoc, lab tech, etc.) or do not wish to register you are still welcome to sit in the class

- If you audit or do not register officially, I strongly recommend that you do the work for the class, i.e. homework/exams/project/lab (we will grade your work!)

- My observation is that you are likely to be wasting your time if you do not do the work but I leave this up to you...
Registering for the class II

- In Ithaca:
  - You must register for both the lecture (3 credits) and computer lab (1 credit) if you take the course for a letter grade
  - If you are an undergraduate, register for BTRY 4830 (lecture and lab); graduate student, register for BTRY 6830 (same)

- In NYC:
  - Weill: the course (PBSB.5021.01) should be available in the Graduate School drop-down at learn.weill.cornell.edu (2015-2016 Spring, Graduate-Quarter 3-4)
  - Rockefeller: email Kristen Cullen cullenk@mail.rockefeller.edu

- Please contact me if there are any issues with registering (!!!)
Grading

• We will grade undergraduates and graduates separately (!!)

• Grading: problem sets (20%), computer lab attendance (5%), project (25%), mid-term (20%), final (30%)

• A short problem set (almost) every week

• Exams will be take-home (open book)

• A single project (~1 month)
Should I be in this class?

- No probability or statistics: not recommended
- Limited probability or statistics (high school, a long time ago, etc.): if you take the class be ready to work (!!!)
- Prob / stats (e.g. BTRY 4080+4090 or BTRY 6010+6020 in Ithaca, Quantitative understanding in biology at Weill, etc.): you’ll be fine
- No or limited exposure to genetics: you’ll be fine
- No or limited exposure to programming: you’ll be fine (we will teach you “programming” in R from the ground up)
- Strong quantitative background (e.g. stats or CS graduate student): you may find the intuitive discussion of quantitative subjects and the applications interesting
What you will learn in this class I

- A rigorous introduction to basics of probability and statistics that is intuition based (not proof based)

- Foundational concepts of how probability and statistics are at the core of genetics, which are complete enough to build additional / more advance understanding (i.e., enough to “get your hooks into the subject”)

- Exposure to many advanced probability / statistics / genetics / algorithmic concepts that will allow you to build additional understanding beyond this class (as brief as a mention to entire lectures - depending on the subject)

- Clear explanations for convincing yourself that the basics of mathematics and programing are not hard (i.e. anyone can do it if they devote the time)
What you will learn in this class II

- An intuitive and practical understanding of linear models and related concepts that are the foundation of many subjects in statistics, machine learning, and computational biology
- The computational approaches necessary to perform inference with these models (EM, MCMC, etc.)
- The statistical model and frameworks that allow us to identify specific genetic differences responsible for differences in organisms that we can measure
- You will be able to analyze a large data set for this particular problem, e.g. a Genome-Wide Association Study (GWAS)
- You will have a deep understanding of quantitative genomics that from the outside seems diffuse and confusing
Questions about logistics?
Subject overview

- We know that aspects of an organism (measurable attributes and states such as disease) are influenced by the genome (the entire DNA sequence) of an individual.

- This means difference in genomes (genotype) can produce differences in a phenotype:
  
  - Genotype - any quantifiable genomic difference among individuals, e.g. Single Nucleotide Polymorphisms (SNPs). Other examples?

- Phenotype - any measurable aspect of an organisms (that is not the genotype!). Examples?
For any two people, there are millions of differences in their DNA, a subset of which are responsible for producing differences in a given measurable aspect.

Example: People are different...

We know that environment plays a role in these differences ...

...and for many, differences in the genome play a role

An illustration

Physical, metabolism, disease, countable ways.
An illustration continued...

• The problem: for any two people, there can be millions of differences their genomes...

• How do we figure out which differences are involved in producing differences and which ones are not?

• This course is concerned with how we do this.

• Note that the problem (and methodology) applies to any measurable difference, for any type of organism!!
Why do we want to know this?

If you know which genome differences are responsible:

- From a child’s genome we could predict adult features
- We target genomic differences responsible for genetic diseases for gene therapy
- We can manipulate genomes of agricultural crops to be disease resistant strains
- We can explain why a disease has a particular frequency in a population, why we see a particular set of differences
- These differences provide a foundation for understanding how pathways, developmental processes, physiological processes work
- The list goes on...
Quantitative genetics and connection to other disciplines

- **Quantitative genomics** is a field concerned with the *modeling* of the relationship between genomes and phenotypes and using these models to *discover and predict*

- Broad Classification of Fields of Genetics:
  - *Modeling Genetic Fields*: quantitative genetics; system genetics; population genetics; etc.
  - *Mechanism Genetic Fields*: Molecular Genetics; Cellular Genetics; etc.
  - *Model System Genetic Fields*: Human Genetics; Yeast Genetics; etc.
  - *Extension Genetic Fields*: Medical genetics; Developmental Genetics; Evolutionary Genetics; Agricultural Genetics, etc.

Note (!!) Take Prof. Alon Keinan’s Population Genetics Class (!!)
T/Th 10:10-11:25 Comstock B108
http://keinanlab.cb.bscb.cornell.edu/content/btry-6820-4820-2016
History of genetics (relevant to Quantitative Genetics)

- Relevant history:
  - 1900-1980: statistical analysis of the patterns of inheritance (i.e. the resemblance between relatives).
  - 1980-2002: mapping (= identification) of the genetic loci responsible for most Mendelian diseases (e.g. diseases where alleles at a 'single' genetic locus determines disease).
  - 2002-present: 'age of genomics' first convincing mapping of genetic loci for complex traits (i.e. cases where genotype cannot be inferred directly from the phenotype).

In sum: during the last decade, the greater availability of DNA sequence data has completely changed our ability to make connections between genome differences and phenotypes.
Connection of genomics-genetics

• Traditionally, studying the impact / relationship of the genome to phenotypes was the province of fields of “Genetics”

• Given this dependence on genomes, it is no surprise that modern genetic fields now incorporate genomics: the study of an organism’s entire genome (wikipedia definition)

• However, one can study genetics without genomics (i.e. without direct information concerning DNA) and the merging of genetics-genomics is quite recent
Present / future: advances in next-generation sequencing driving the field
Why this is a good time to be learning about this subject

• Mapping (identifying) genotypes (genetic loci) with effects on important phenotypes is fast becoming the major use of genomic data and a major focus of genomics

• However, the data collection, experimental, and statistical analysis techniques for doing this are still being developed

• The current statistical approaches are the focus of this course (i.e., you will have a solid foundation by the end)

• The importance is just now starting to permeate broadly (i.e., we are entering the “internet generation” for genomics and the impact of genomics on biology)
In this class, we will use statistical modeling to say something about biology, specifically the relationships between genotype (DNA) and phenotype.

Let’s start with the biology by asking the following question: why DNA?

The structure of DNA has properties that make it worthwhile to focus on...
It’s the same in all cells

- In multicellular organisms, the structure of the genome is (almost) perfectly copied during the replication of cells.
- The genome is the same in every non-cancer cell of a multicellular organism, with just a few exceptions. So, we may refer to the genome of an individual. In cancers, the genome differs from cell to cell, such that it is more problematic to refer to the genome of a cancer.
- The genome provides instructions for how biological processes proceed (e.g., development, metabolism, environmental response); So, the genome is an important determinant of the measurable characteristics of an organism or cancer.
- In the production of a new organism or offspring, either the entire genome (e.g., bacteria) or a subset of the genome (e.g., half from each parent in humans) is copied almost perfectly from parent to offspring; The copying of genomes from parents to offspring is the primary reason why offspring tend to resemble their parents.

Figure 1: A simplified schematic showing genome organization in human cells. The DNA of a genome is located within the nucleus of a cell. The genome is organized in long strings that are tightly coiled around protein structures to form chromosomes. Each string is a double helix where the building blocks are A-T and G-C nucleotide pairs. with a few exceptions (e.g. cancer, immune system...
It’s passed on to the next generation
It has convenient structure for quantifying differences.
It’s responsible for the construction and maintenance of organisms.

Note: other regions of genomes can impact phenotypes...
Statistics and probability I

• **Quantitative genomics** is a field concerned with the modeling of the relationship between genomes and phenotypes and using these models to discover and predict.

• We will use frameworks from the fields of probability and statistics for this purpose.

• Note that this is not the only useful framework (!!) - and even more generally - mathematical based frameworks are not the only useful (or even necessarily “the best”) frameworks for this purpose.

• So, why use a probability and statistics framework? Let’s start by considering a definition of probability.
Statistics and probability II

- A non-technical definition of probability: a mathematical framework for modeling under uncertainty
- Such a system is particularly useful for modeling systems where we don’t know and/or cannot measure critical information for explaining the patterns we observe
- This is exactly the case we have in quantitative genomes when connecting differences in a genome to differences in phenotypes
We will therefore use a probability framework to model, but we are also interested in using this framework to discover and predict.

More specifically, we are interested in using a probability model to identify relationships between genomes and phenotypes using DNA sequences and phenotype measurements.

For this purpose, we will use the framework of statistics, which we can (non-technically) define as a system for interpreting data for the purposes of prediction and decision making given uncertainty.
Examples of successful applications of the framework

First Generation
Second Generation
Third Generation
Fourth Generation

Female
Male
Person with cystic fibrosis
Linking parents
Linking parents to children

rs1908530 genotype
ERAP2 expression
T/T
T/C
C/C

rs27290 genotype
ERAP2 expression
A/A
A/G
G/G

cis eQTL
No eQTL

NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

Published Genome-Wide Associations through 12/2012

-8

for 17 trait categories
That’s it for today

- Next lecture, we will begin our formal and technical introduction to probability
- We will start by defining the concepts of a “system”, “experiments” and “experimental trials”, and “sample outcomes” and “sample spaces”