Module 1 - Introduction and Overview

(Fundamentals of) Statistical Genetics

Lei Sun

Department of Statistical Sciences, FAS
Division of Biostatistics, DLSPH
University of Toronto

Online Learning Notes I

CHL 5224 (Fall 2020, Tuesdays 10-1pm) will use the online synchronous format. This is to create the best and most equitable learning experience for all students.

Lecture notes will be posted to the Quercus course site one week ahead of each lecture. Students are expected to study the lecture notes prior to each lecture and identify the key content, so that the synchronous teaching can focus on the common and most challenging parts of the lecture. Each 3-hour lecture will be divided into two parts: ≈2-hour for lecturing, during which discussion is encouraged to cultivate engaged learning, and ≈1-hour for additional discussion.

Homework consists of two components: non-graded exercises and graded projects. The non-graded exercises (math stat in the context of statistical genetics) will be assigned after (almost) each lecture, and solution will be provided two weeks later. There will be two projects involving programming and data analysis, and the first one serves as the midterm (40%) and the second one as the final (60%). Each project requires the use of R markdown.

Online Learning Notes II

- Online synchronous via Zoom:

  Meeting ID: 843 1098 8408

  Passcode: 239403
- Study the lecture notes prior to each lecture; in general notes in blue have hyperlinks.
- Cannot emphasize this enough: Join the lectures!
- Have your video on and actively participate the lectures.
- ≈Four 25+5 minutes blocks of teaching+break.
- Do the non-graded take-home assignments, independently; do not wait for the solution sheets.
- There will be a TA.

Outline

- What is Statistical Genetics?
- Big Data and Data Science

- Textbook and Materials
- Teaching Objectives
- Intended Audience
- Course Format and Evaluations

- Resources: books, journals, seminar, STAGE

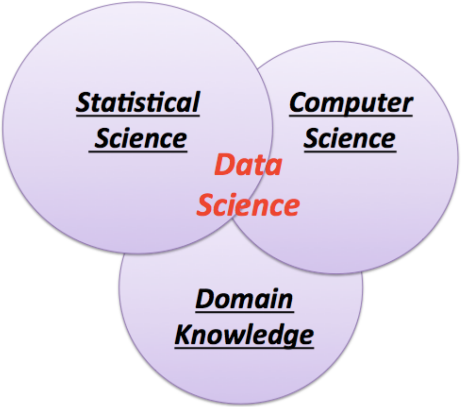
- Overview of Human Genetic Studies
- Designs/Stages of Genetic Studies

- Course Outline (Tentative)

- What would you know at the end of the class?

What is Statistical Genetics and Genomics?

Interdisciplinary and Collaborative



Big Data: $n > 10^3, p > 10^6$ (high-level processed data in GB and 'raw data' in TB) and Complex: (e.g. multiple causal factors, interactions, pathway/network...)

Recent Examples I

- Multiple hypothesis testing
  - Sun et al. (2006). Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. *Genetic Epidemiology* 30:519-530.
  - Sun et al. (2012). Hypothesis-driven GWAS (GWAS-HD): Multiple apical plasma membrane constituents are associated with susceptibility to meconium ileus in individuals with cystic fibrosis. *Nature Genetics* 44(5):562-569.
  - Ting Zhang, Sun (2019). Beyond the traditional simulation design for evaluating type 1 error control: from the 'theoretical' null to 'empirical' null. *Genetic Epidemiology*.
  - Jianhui Gao, Sun (working paper). Data integration methods comparison: meta-analysis, Fisher's method, weighted-p-value approach, and stratified FDR control.

Recent Examples II

- Selection bias and selective inference
  - Sun, Bull (2005). Reduction of selection bias in genome-wide genetic studies by resampling. *Genetic Epidemiology* 28:352-367.
  - Lizhen Xu, Craiu, Sun (2011). Bayesian methods to overcome the winner's curse in genetic studies. *Annals of Applied Statistics* 5(1):201-231.
  - Emery Goossens, Sun (working paper). **BOSS: best ordered subset selection for testing the global null hypothesis.**

Recent Examples III

- Joint analysis of multiple genetic variants

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \sim \begin{bmatrix} X_{1,1} & \cdots & X_{1,J} \\ \vdots & \ddots & \vdots \\ X_{n,1} & \cdots & X_{n,J} \end{bmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_J \end{pmatrix} + \begin{bmatrix} Z_{1,1} & \cdots & Z_{1,K} \\ \vdots & \ddots & \vdots \\ Z_{n,1} & \cdots & Z_{n,K} \end{bmatrix} \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_K \end{pmatrix}$$

MANY approaches: **CAST**(Morgenthaler et al. 2007), **Weighted Sum** (Madsen and Browning 2009), **Variable Threshold** (Price et al. 2010), **KBAC** (Liu and Leal 2010), **EREC** (Lin and Tang 2011, Danyu Ling's group), **C-Alpha** (Neale et al. 2011, Kathryn Roeder's group), **SKAT** (Wu et al. 2011, Xihong Lin's group).

- Andriy Derkach, Lawless, Sun (2014). Pooled association tests for rare genetic variants: a review and some new results. *Statistical Science* 29(2):302-321.
- Andriy Derkach, Lawless, Sun (2013). Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. *Genetic Epidemiology* 37(1):110-121.
- Andriy Derkach, Lawless, Sun (2015). Score tests for association under response-dependent sampling designs for expensive covariates. *Biometrika* 102(4):988-994.

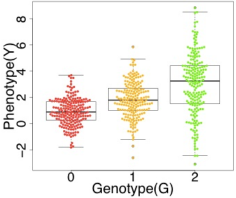
Recent Examples IV

- Joint analysis of multiple outcomes
  - Lizhen Xu, Craiu, Sun, Paterson (2016). Parameter expanded algorithms for Bayesian latent variable modeling of genetic pleiotropy data. *Journal of Computational and Graphical Statistics* 25(2):405-425.

- X-chromosome: model uncertainty, confounding etc.
  - Bo Chen, Craiu, Sun (2019). Bayesian model averaging for the X-chromosome inactivation dilemma in genetic association studies. *Biostatistics.*
  - Bo Chen et al. (working paper). The X Factor: A Robust and Powerful Approach to X-chromosome-Inclusive Whole-genome Association Studies.
  - Wei Deng et al. (2019). Analytical strategies to include the X-chromosome in variance heterogeneity analyses: Evidence for trait-specific polygenic variance structure. *Genetic Epidemiology*.
  - Zhong Wang, Paterson, Sun (working paper). The landscape of X-chromosome analysis and reporting: challenges and some remedies.

Recent Examples V

Indirect modelling of interaction effects



- <u>David Soave</u> et al. (2015). A joint location-scale test improves power to detect associated SNPs, gene-sets and pathways. *The American Journal of Human Genetics* 97(1):125-138.
- <u>David Soave</u>, Sun (2017). A generalized Levene's scale test for variance heterogeneity in the presence of sample correlation and group uncertainty. *Biometrics* 73(3):960-971.
- <u>Ting Zhang</u>, Greenwood, Sun (working paper). Multivariate generalized Levene's test for detecting latent gene-gene or gene-environment interactions.

Recent Examples VI

- 'Simple' regression for complex data
  - Lin Zhang, Sun (working paper). On 'Reverse' Regression for Robust Genetic Association Studies and Allele Frequency Estimation with Related Individuals.
  - Lin Zhang, Sun (working paper). A generalized robust allele-based genetic association test.
  - Yanyan Zhao, Sun (2020). On set-based association tests: insights from a regression using summary statistics. *Canadian Journal of Statistics*.
  - Yanyan Zhao, Sun (working paper). A stable and adaptive polygenic signal detection method based on repeated sampling.

Recent Examples VII

- Method implementation, computing, and software development
  - BR^2: Sun et al. (2012). BR-squared: a practical solution to the winner's curse in genome-wide scans. *Human Genetics*. 129:545-552.
  - TRUFFLE: Apostolos Dimitromanolakis, Paterson, Sun (2019). Fast and accurate shared segment detection and relatedness estimation in un-phased genetic data using TRUFFLE. *American Journal of Human Genetics*105(1):78-88.

Why Still Linkage?

- The types of data can change drastically in short period but the essence of the corresponding statistical methodologies often remains the same.
- Learning linkage, families/pedigrees and other genetic specific concepts gives you an advantage; they require more training in genetics than association.
- State-of-art current research topics will be briefly discussed when appropriate.

- Running a GWAS is much easier than understanding the related statistical concepts!

Big Data - Some Cautionary Words

Big data: are we making a big mistake?

- *While big data promise much to scientists, entrepreneurs and governments, they are doomed to disappoint us if we ignore some very familiar statistical lessons.*
- *There are a lot of small data problems that occur in big data. They don't disappear because you've got lots of the stuff. They get worse.*
- *Data contain systematic biases and it takes careful thought to spot and correct for those biases. Big data sets can seem comprehensive but the "N = All" is often a seductive illusion.*
- *The multiple-comparisons problem. Statistical correlation/patterns≠Causation.*
- *Big data do not solve the problem that has obsessed statisticians and scientists for centuries: the problem of insight, of inferring what is going on.*
- *Proving the value of statistics would also come from interdisciplinary working. Teaming up with computer scientists, astronomers, the bioinformatics people.*
- *"Big data" has arrived, but big insights have not. The challenge now is to solve new problems and gain new answers - without making the same old statistical mistakes on a grander scale than ever.*

Big Data - Examples of Some Specific Issues

- Multiple hypothesis testing (STA4515)
- Accuracy of a particular statistical test: e.g. interested in the extreme tail of a distribution because of multiple hypothesis testing but type 1 errors were often evaluated for 0.05-0.0001 range.
- Statistical evidence measure: e.g. p-value interpretation in the context of large sample or heterogeneous sample sizes across different tests.
- Heterogeneity: e.g. different effect sizes or models across different populations.
- Selective inference: e.g. winner curse, selection bias, 'honest' post-selection statistical significance.

Data Science I

David Donoho's take on Data Science

- *Data Science is statistics*

> *When physicists do mathematics, they don't say they're doing number science. They're doing math. If you're analyzing data, you're doing statistics. You can call it data science or informatics or analytics or whatever, but it's still statistics. ... You may not like what some statisticians do. You may feel they don't share your values. They may embarrass you. But that shouldn't lead us to abandon the term "statistics". (Karl Broman, Univ. Wisconsin)*

- *The activities of Greater Data Science are classified into 6 divisions*

  1. *Data Exploration and Preparation*
  2. *Data Representation and Transformation*
  3. *Computing with Data*
  4. *Data Modeling*
  5. *Data Visualization and Presentation*
  6. *Science about Data Science*

Data Science II

- *In 2065, mathematical derivation and proof will not trump conclusions derived from state-of-the-art empiricism. Instead of deriving optimal procedures under idealized assumptions within mathematical models, we will rigorously measure performance by empirical methods, based on the entire scientific literature or relevant subsets of it.*
- *(BUT,) I am not arguing for a demotion of mathematics. I personally believe that mathematics offers the only way to create true breakthroughs. The empirical method is simply a method to avoid self deception and appeals to glamor.*

The challenges are endless but so are the opportunities!

Data Science Training at UofT

- Interested in Data Science?
- Specialized program in Data Science: MSC in Applied Computing (MSCAC) with concentration in Data Science.
- Courses: e.g. STA414/STA2104 - Statistical Methods for Machine Learning

Teaching Objectives

Statistical genetics is an important data science research area with direct impact on population health, and this course provides an INTRODUCTION to its concepts and fundamentals.

- *Learn about statistical methods for genetic analysis (better analyze genetic data or research in methodology)*.
- Integrative by nature: statistical methods+genetic data for gene discovery (more on interdisciplinary training later).
- Key: basic concepts; illustrative examples and some specific statistical inference techniques.

'Textbook' and Materials

- Loosely based on the book by Laird N and Lange C (2011) The fundamentals of modern statistical genetics.

  (Free .pdf version available via UofT library)

- Mainly lecture notes and some additional reading materials.

- Lecture notes and other materials are posted on the course Quercus.

Intended Audience

- Graduate students from programs in statistics and biostatistics.
- Graduate students from other programs are welcome as long as the background training is sufficient.

Background Needed

- *Assume no formal training in genetics.* Basic concepts in molecular genetics will be reviewed in the class.
- Assume statistical knowledge at the STA303 - Methods of Data Analysis level.
- How quantitative: *familiarity with elementary probability and statistical inference and methods*.
  - Distributions of some basic random variables (binomial, normal etc); conditional, marginal and joint distributions.
  - Likelihood methods, estimation and hypothesis testing.
  - Basic multivariate regressions: linear, logistic etc.
- Rule of thumb: are you comfortable with the Exercises in the textbook. And are you comfortable with this Test Example?

Other Useful Books

- Statistical Genetics
  - Sham P (1998). Statistics in Human Genetics. Arnold, London. (more statistical)
  - Lange K (2002). Mathematical and Statistical Methods for Genetic Analysis. 2nd edition. Springer-Verlag, New York. (more statistical)
  - Zieglier A, Koenig I (2006). A Statistical Approach to Genetic Epidemiology: Concepts and Applications. Wiley-VCH. (more statistical)
  - Thomas DC (2004). Statistical Methods in Genetic Epidemiology. Oxford University Press. (more epidemiological)
  - Ott J (1999). Analysis of Human Genetic Linkage. 3rd edition. Johns Hopkins University Press, Baltimore. (mostly focus on linkage)
- Genetics Background
  - Gonick L, Wheelis M (1991). Cartoon Guide to Genetics. Revised edition. HarperCollins.
  - Virtually any genetics textbook.

Journals to Follow for Statistical Genetics Research

More Genetic

- Nature Genetics
- American Journal of Human Genetics
- PLoS Genetics
- Genetic Epidemiology
- Nature Review Genetics
- Genome Research
- Human Genetics
- European Journal of Human Genetics
- Many more...

More Statistical

- Biostatistics
- Bioinformatics
- Annals of Applied Statistics
- Journal of the American Statistical Association
- Biometrika
- Biometrics
- Statistics in Medicine
- A few more

Also useful: review papers from journals such as Nature Reviews Genetics.

Statistical Methods for Genetics & Genomics (SMG)

- SMG: a journal club and research seminar ongoing for > 20 years.
- co-organized by
  - Dr. Shelley Bull, Senior Scientist of Lunenfeld Research Institute.
  - Dr. Andrew Paterson, Senior Scientist of Sickkids Research Institute.

- Purpose of the seminar:
  - Be exposed to current problems and responding statistical genetic methodologies.
  - A great opportunity for the students to join the community.

- Format of the seminar:
  - Local researchers, post-doc fellows and graduate students in the field participate, and occasionally outside speakers contribute.
  - Discuss interesting journal papers or current research work.

- Time: Fridays 10 - 11am.
- Location: Online via MS Team

Strategic Training for Advanced Genetic Epidemiology

STAGE

- *CIHR STAGE is a formal and comprehensive training program in Genetic Epidemiology and Statistical Genetics, the first of its kind in Canada and one of few in the world.*
- *The program offers new training and career development opportunities designed to cross-train individuals at the interface of genetics and population health sciences in genetic epidemiology and statistical genetics – two disciplines currently facing a massive shortage of qualified individuals in Canada and elsewhere.*
- *The overall goal is to improve prevention and management of complex diseases by increasing capacity in genetic epidemiology and statistical genetics research.*

- Tri-Model:
  - Genetic and Molecular Epidemiology
  - Statistical and Computational Genetics
  - Bio-Medical Genetics

Overview of Human Genetic Studies I

- Identify the genetic factors responsible for human heritable diseases/traits, taking into account environmental factors and interaction between them (GxG and GxE).

- Many branches of genetics:
  - **Statistical genetics** conduct statistical (and often computational) analysis of genetic data.

  - **Genetic epidemiology** *is a branch of epidemiology that deals with both genetic and environmental contributions to disease. Genetic epidemiology uses methods from statistical genetics and epidemiology to understand the interplay between genes, environment and disease. Sometimes data on geographic, spatial, temporal and/or racial, as well as familial, variation in disease rates can provide insight into the genetic nature of disease.*

Overview of Human Genetic Studies II

- **Population genetics** *is concerned with the genetic variation within and between populations, over time and space. This includes modeling variation in genes due to many factors: selection of certain variants due to response to environmental conditions, in- and out-migration, drift occurring in small populations, and mutations, as well as understanding genetic differences in populations. There are some key principles of population genetics, namely Hardy-Weinberg equilibrium, linkage equilibrium and population substructure, which are important in association analysis and will be covered in a short introductory chapter.*

- Others: medical genetics, behavioural genetics, molecular genetics, bioinformatics and computational biology etc.

- Human genetics vs. plant and animal genetics.

Designs/Stages of Genetic Studies I

- Is the trait heritable? Find evidence for the role of genetic factors. Degree of inheritance. **Familial Aggregation**.

- Evidence for mode of inheritance. Simple Mendelian genetic models (one locus dominant or recessive) vs. complex traits. Determine the underlying model that explains the relationship between the phenotype and genetic factors. **Segregation Analysis**.

- Identify new susceptibility gene (two general approaches).

- Approach I: **Candidate Gene Study.**

  Know the basic function of a gene (say related to lung function); want to know if it is important to the trait of interest (say asthma).

Designs/Stages of Genetic Studies II

- Approach II: **Genome-Wide Mapping Study** (focus of this course).
  - Traditionally, a sequential two-staged design:
    - Coarse, genome-wide mapping studies. Which chromosome and which region? **Linkage Analysis using families**.
    - Fine LD mapping. Refine the region and further narrow down the region. **population-based association studies or families-based TDT**, both utilizing the concept of **LD**.
  - More recently, **Genome-Wide Association Studies (GWAS)** and **Next Generation Sequencing (NGS)!**
    - Driven by technology: cost of genotyping reduced more than 100 fold in the past few years!
    - *Over the years, the methodological focus of statistical genetics has changed to keep pace with the different kinds of genetic data that technology has made available. Most recently, new technologies arising from the Human Genome Project and HapMap Project have generated a surge of methodological development to address unsolved problems in human genetics.*
    - Many issues as well as opportunities: computational (storage, manipulation, etc.), statistical (QC, analysis, multiplicity/fishing, power, rare variants, etc.)

Designs/Stages of Genetic Studies III

- **Find the functional/causal genetic variants**.
  - Only biological studies ultimately determine certain genetic variation affecting susceptibility to disease, etiologic variant cannot be established on the basis of statistical analysis.
  - However, statistical analysis can detect variants that merit the next level of biological studies, e.g. Re-ranking sequencing variants in the post-GWAS era.
  - **Data integration** of 'omics' data!

- **Assess important characteristics of known genes**: direct study the function of a gene and its protein product in cell or tissue. May use animal models to mimic the trait of interest.
- Learn how to **predict, prevent, manage and treat the disease.**

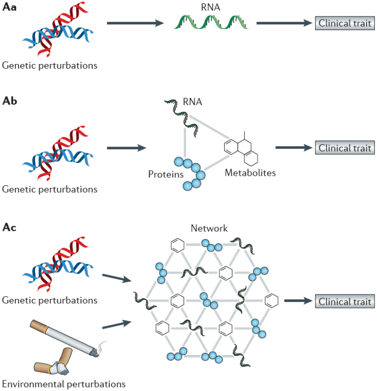
Simple (Mendelian) vs. Complex Diseases/Traits

- Simple: *Mendelian disorders and diseases follow simple Mendelian patterns of inheritance in families and generally do not have any other causes other than the genetic disease variant. Linkage analysis has been very successful in finding genes for Mendelian disorders. Diseases or disorders which are initiated by variants in a single gene are typically rare and severe conditions, e.g., Cystic Fibrosis, Duchenes' Muscular Dystrophy and Sickle Cell Anemia.*

- Complex: *Most common diseases, e.g., asthma, obesity, Alzheimer's disease, bipolar disorder, etc., fall into the category of multi-factorial diseases or complex genetic diseases. Here, disease risk is thought to be influenced by a set of genes and environmental factors which may interact with each other. Linkage analysis has been less successful with finding genes for complex disorders. [However], the basic genetic principles of inheritance of genetic material are the same for both Mendelian and complex diseases.*

- Examples from the textbook: Sickle Cell Anemia vs. Alzheimer's Disease.

Complex Phenotype and Genotype Relationship



- Figure 1 of Civelek and Lusis (2014). Systems genetics approaches to understand complex traits. Nature Review.
- **Systems biology** *is an approach to understand the flow of biological information that underlies complex traits.*
- *It uses a range of experimental and statistical methods to quantitate [how to standardize] and integrate intermediate phenotypes, such as transcript, protein or metabolite levels, in populations that vary for traits of interest.*

Data integration of different kinds of 'omics' data and beyond!
Higher order and multivariate analysis of pathway and network and interaction!

“Unrelated” Population Data vs. Family/Pedigree Data

- Linkage analyses rely on families/pedigrees and need deep understanding of how genetic materials are transmitted from one generation to the next (inheritance).

- Association studies often rely on population data with “unrelated” samples and use regression framework. Many bio/statisticians can do a good job at association studies, BUT, **only those with deep understanding of inheritance tend to develop novel statistical genetics methods**.

- In fact, family-based studies are coming back!

- Thus, this course will spend substantial time on linkage, pedigrees and related concepts!

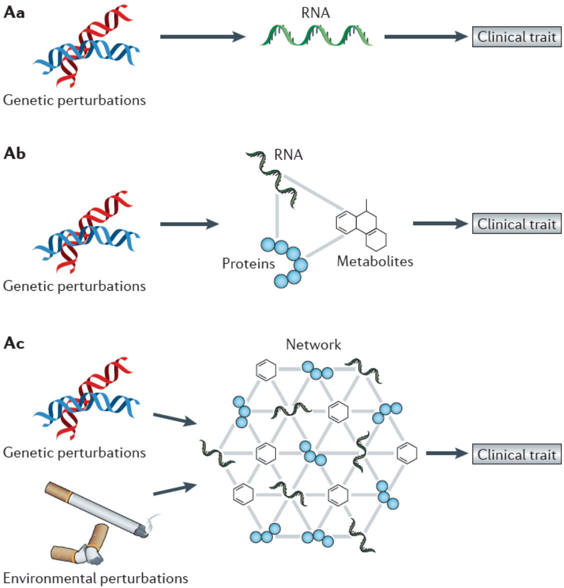
Practical Implications and Ethical Issues

- Research Ethics Boards (REBs)
- Who should be screened for the gene (say breast cancer)?
- What to do next if the disease variant was found (surgery or not)?
- Can a company patent a gene?

- Any taboo or controversial traits, e.g.?
  - Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence.
  - Genome-wide association study of male sexual orientation.

- Many others.

COMPLEX Traits!



Tentative Course Outline (tentative) I

1. Chapter 1: Overview of Genetic Studies and Introduction of Genetic Terminologies.
2. Chapter 3: Some Basic Concepts from Population Genetics

   Review of Some Basic Concepts from Statistical Inference
3. Chapter 2: Principles of Inheritance: Mendels Laws and Genetic Models

   Chapter 4.1: Likelihood for Pedigree Data
4. Chapter 4: Aggregation, Heritability and Segregation Analysis
5. Expanding Chapter 2.3: The Biology Underlying Mendelian Inheritance

   Part of Chapter 5 - Map and Linkage
6. Chapter 6: Linkage Analysis (and GENEHUNTER)

Tentative Course Outline (tentative) II

7. Chapter 7: LD and Association Analysis 1 (regression, logistic regression)
8. Chapter 8: Population Substructure

   Chapter 9: Association Analysis 2 (TDT)
9. Chapter 9: Association Analysis 3 (FBAT) and Haplotype Analysis
10. Chap 11: Genome-Wide Association Studies (GWAS) (and PLINK)
11. ???Chapter 10: Multiple Hypothesis Testing
12. ???Next-Generation Sequencing (NGS)

What would you know at the end of the class? I

- **Some basic genetic terminologies**, e.g. Human genome, Chromosomes (autosomes and sex chromosome), Double helix structure (strand issue), Genetic markers/polymorphisms (SNPs, Microsattelites etc), Alleles, Genotypes (homozygous and heterozygous), Mutations (synonymous and missense), Gene (exons and introns), HWE, LE, LD, Haplotype, Genetic map, Linkage and Association, etc.

- **Population genetics**, e.g. what is the MLE (why MLE?) of the population allele frequency of allele A? How do you perform a LRT (or Score and Wald tests), testing the null hypothesis that the frequency is say 0.75?

Genotypes	AA	Aa/aA	aa	Total
Observed Counts	189	89	9	287

What would you know at the end of the class? II

- **Principle of Inheritance and likelihood over pedigrees**, e.g. how do you calculate the following probability?

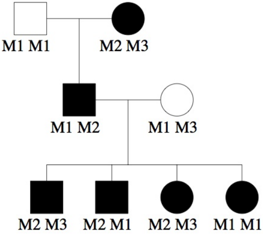
$$P(\text{sib } 1 = Aa, \text{sib } 2 = AA, \text{father} = AA, \text{mother} = Aa)$$

$$P(\text{sib } 1 = Aa, \text{sib } 2 = AA)$$

- **Useful for genetic counselling**, e.g Cystic Fibrosis (CF) is a recessive disease.
  - If both parents are carriers, what's the chance that their first born will have CF?
  - If both parents are carriers, what's the chance that both their kids will have CF?
  - If both parents do not have CF, what's the chance that their first born will have CF?

What would you know at the end of the class? III

- **Linkage analysis**, e.g.
  - Shaded individuals are affected by the disease of interest.
  - The disease is caused by a genetic variant, but we don't know where it is.
  - We do know and collected genotype data for this marker *M*.
  - Is the causal genetic variant 'close' or linked to this marker M?
  - How do we measure the distance and linkage evidence?
  - Do we have sufficient information?



What would you know at the end of the class? IV

- **Association analysis**, e.g.
  - *In a genetic association study for late-onset Alzheimer's Disease in a Japanese population (Takeietal. 2009), a number of SNPs have been genotyped in the APOE region.*
  - Is this SNP/genetic marker (rs394221) associated with Alzheimer's Disease?

	MM	Mm	mm	Total
Cases	149	269	91	509
Controls	153	325	180	658

- **GWAS**.

What would you know at the end of the class? V

How do you solve a problem in (Genomic) Data Science?