HIGH DIMENSIONAL STATISTICS

LECTURE 1: MOTIVATION FOR THE COURSE

Msc programme in Mathematics and Statistics

University of Helsinki

Matti Pirinen

"HIGH DIMENSIONAL", WHAT IS IT AND WHY BOTHER?

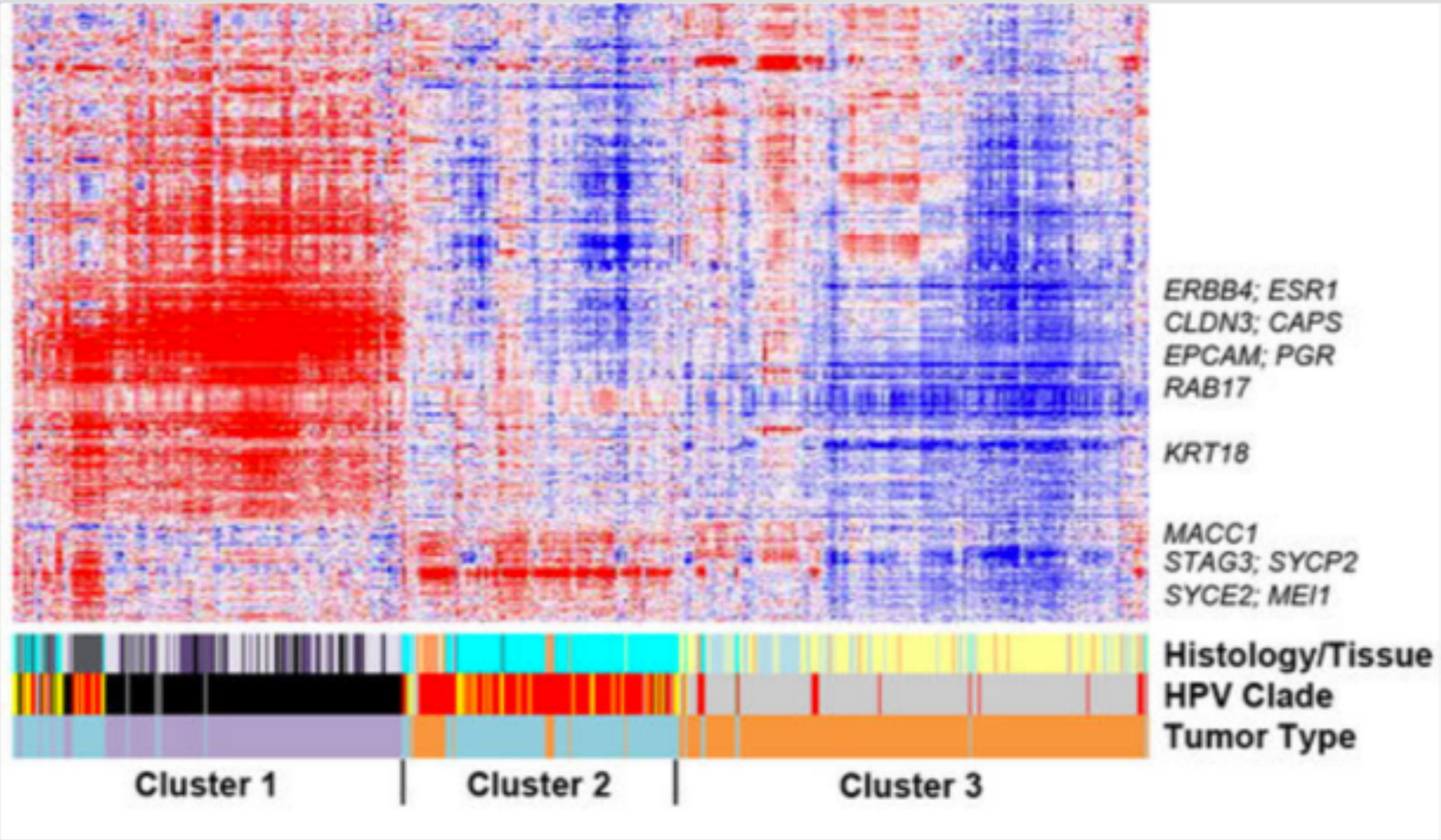
Each row is a gene (~10^4) and each column a cancer patient (~10^3)

Red/blue colors represent gene expression levels

1. Are different subtypes of cancer different in gene expression?

2. Can we tell from gene expression Which subtype patient has?

3. Can we choose treatment based on gene expression levels?



From: "Integrated genomic and molecular characterization of cervical cancer." Nature 543.

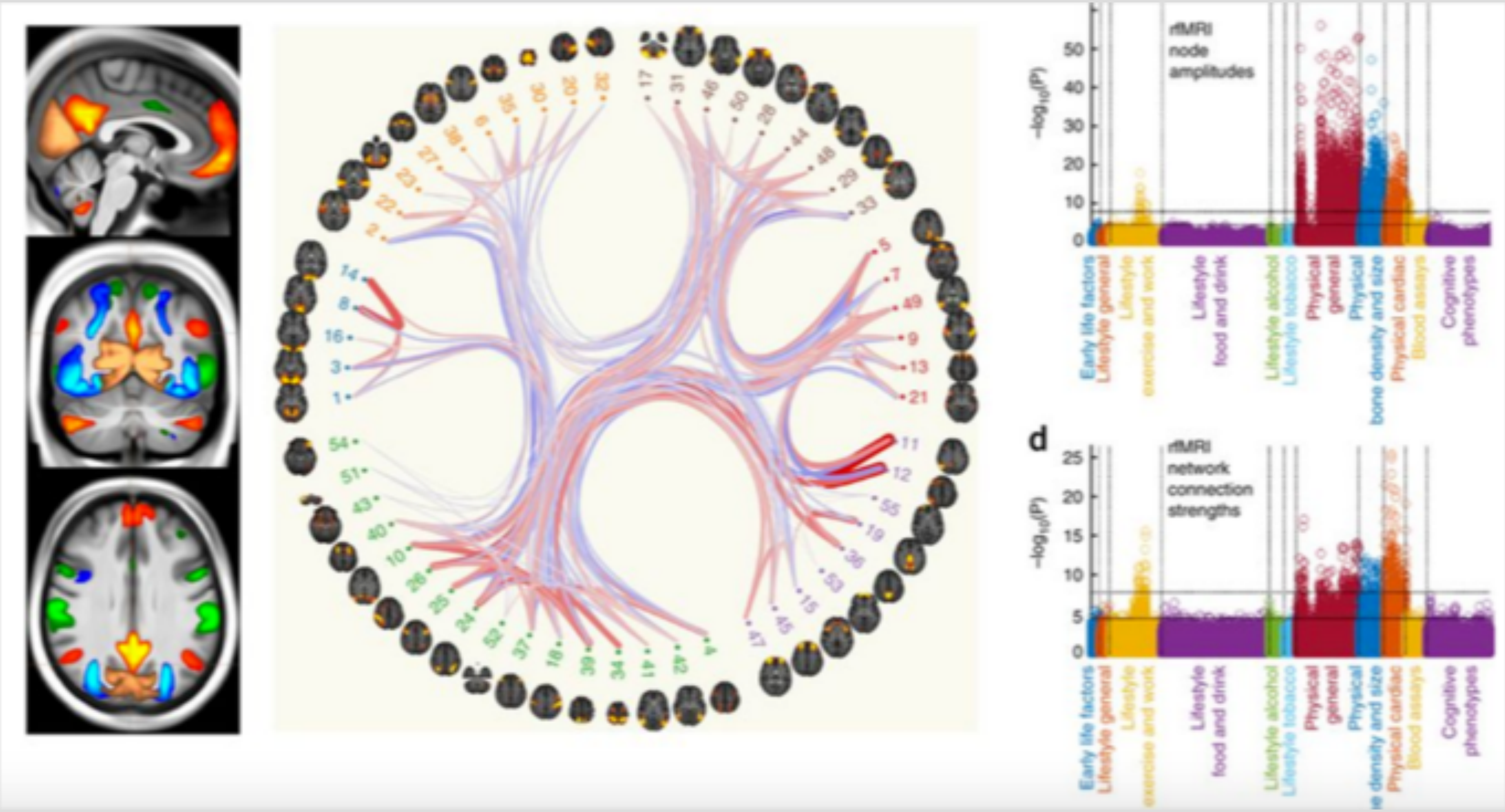
"HIGH DIMENSIONAL", WHAT IS IT AND WHY BOTHER?

UK Biobank releases brain Images of 100,000 participants

1. How brain activity changes between rest and tasks?

2. Can we tell from brain activity what is the context of the individual ?

3. What are the statistical associations between brain activity patterns and 1000s of measured congnitive, behavioral, lifestyle or genetic variables?



Nature Neuroscience 2015 **19**, pages 1523–1536

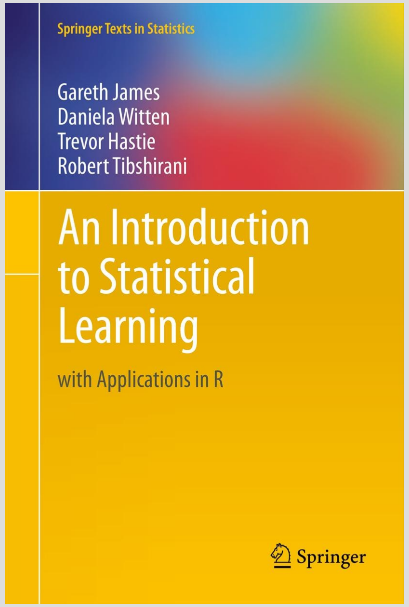
"HIGH DIMENSIONAL", WHAT IS IT AND WHY BOTHER?

- Examples of ongoing data explosion
    - Life sciences with high-throughput technologies e.g. genomics, metabolomics
    - Physics and engineering e.g. CERN, astronomy, robotics, sensors
    - Humanities by digitalization e.g. libraries of historical texts
    - Internet e.g. images, videos, sound, text, social media
- **Working definition:** *High Dimensional* (HD) data set X has a lot of observations (n x p is large)
    - n units (or samples) as rows of X (e.g. individuals n ~10^5)
    - p variables (or features) as columns of X (e.g. genetic variants p~10^6)
    - often HD means that "p >> n is large", but more general definition for us is "n x p is large"
- Unprecedented potential for new knowledge

STATISTICS AND MACHINE LEARNING

- Methods we will consider have (some of) following properties
  - Role of the variables is interpretable in fitted model
  - Simple and complete description in terms of a probability model
  - Conceptually straightforward quantification of uncertainty of parameters and predictions (although not always easy to compute in HDs)
- Our methods are instances of **statistical learning** subset of machine learning (ML) methods
  - ML has also powerful methods for prediction that are more of "black boxes" and are given by algorithms rather than by probability models
    - Deep learning, random forests etc.
- In modern data science we should know all types of learning methods
  - But in this course we will concentrate on statistical learning

EXCELLENT BOOKS AVAILABLE ONLINE



Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

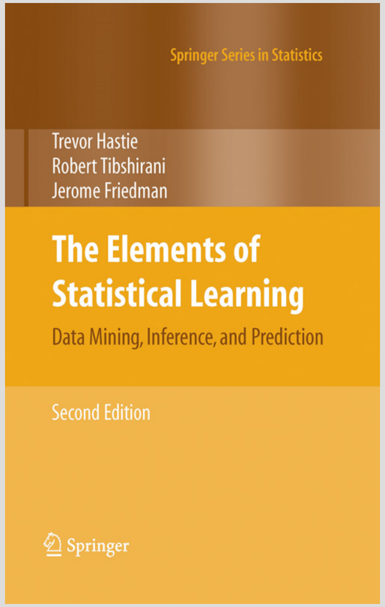
An Introduction to Statistical Learning

with Applications in R

Springer

- Shorthand: ISL
- Very easy to read
- Thorough examples in R
- Little maths
- Excellent book to get intuition behind the concepts and models
- Also video lectures of chapters available

- Shorthand: ESL
- Comprehensive collection of methods
- Mathematical descriptions included



Springer Series in Statistics

Trevor Hastie
Robert Tibshirani
Jerome Friedman

The Elements of Statistical Learning

Data Mining, Inference, and Prediction

Second Edition

Springer

http://www-bcf.usc.edu/~gareth/ISL/

https://web.stanford.edu/~hastie/ElemStatLearn/

CONTENTS

- Weeks 1-2: Large-scale inference, i.e., what are the statistical ideas and measures used when we carry out thousands of tests/comparisons simultaneously
- Weeks 3-5: Regression with a large numbers of predictors, variable selection
- Weeks 6-7: Dimension reduction
- Week 7: Summary
- Exam

PASSING THE COURSE

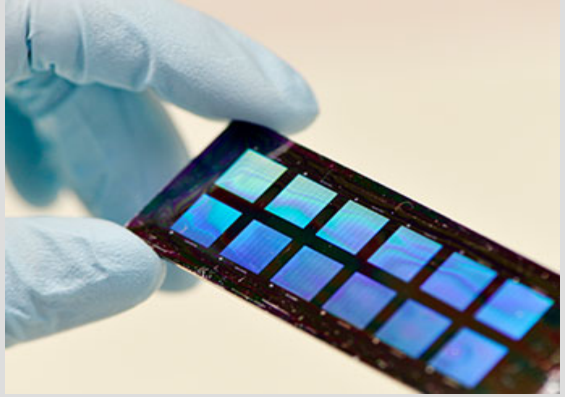
- 6 sets of exercises to be returned through Moodle area "High Dimensional Statistics 2019"
  - Half of the exercise points needed to pass
  - Computer exercises to be done with R (preferably R Markdown)
  - Exercise sessions Tuesdays 10.15-12.00, btw 5.11 … 10.12
- Exam ?.12, 3hrs, with paper and pencil
- Lectures on Tuesdays 12.15 -14.00 and Thursdays 10.15-12.00.
- Course material in Moodle

EXAMPLE FROM GENOMICS

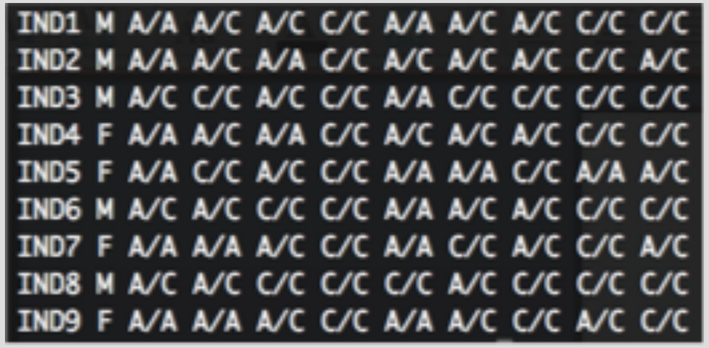
- Which genetic variants are **associated** with cardiovascular disease (CVD), #1 cause of death in western world? (weeks 1-2)
- Which genetic variants are **causal** for CVD? (weeks 3-5)
- How can we best **predict** genetic risk for CVD? (weeks 3-5)
- How can we **visualize** and extract the **main structure** of very high-dimensional genetic data in just a few dimensions? (weeks 6-7)

GENOME-DISEASE ASSOCIATION STUDY

- Collect 10,000s of cases (individuals with the disease) and controls (individuals from the general population who do not have the disease)
- Genotype everyone in 1,000,000s of genomic positions
- Do a statistical test at each position to see whether genotype distributions are different between cases and controls

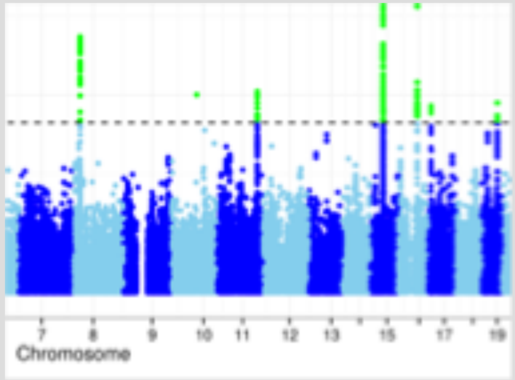


$n = 10^5$

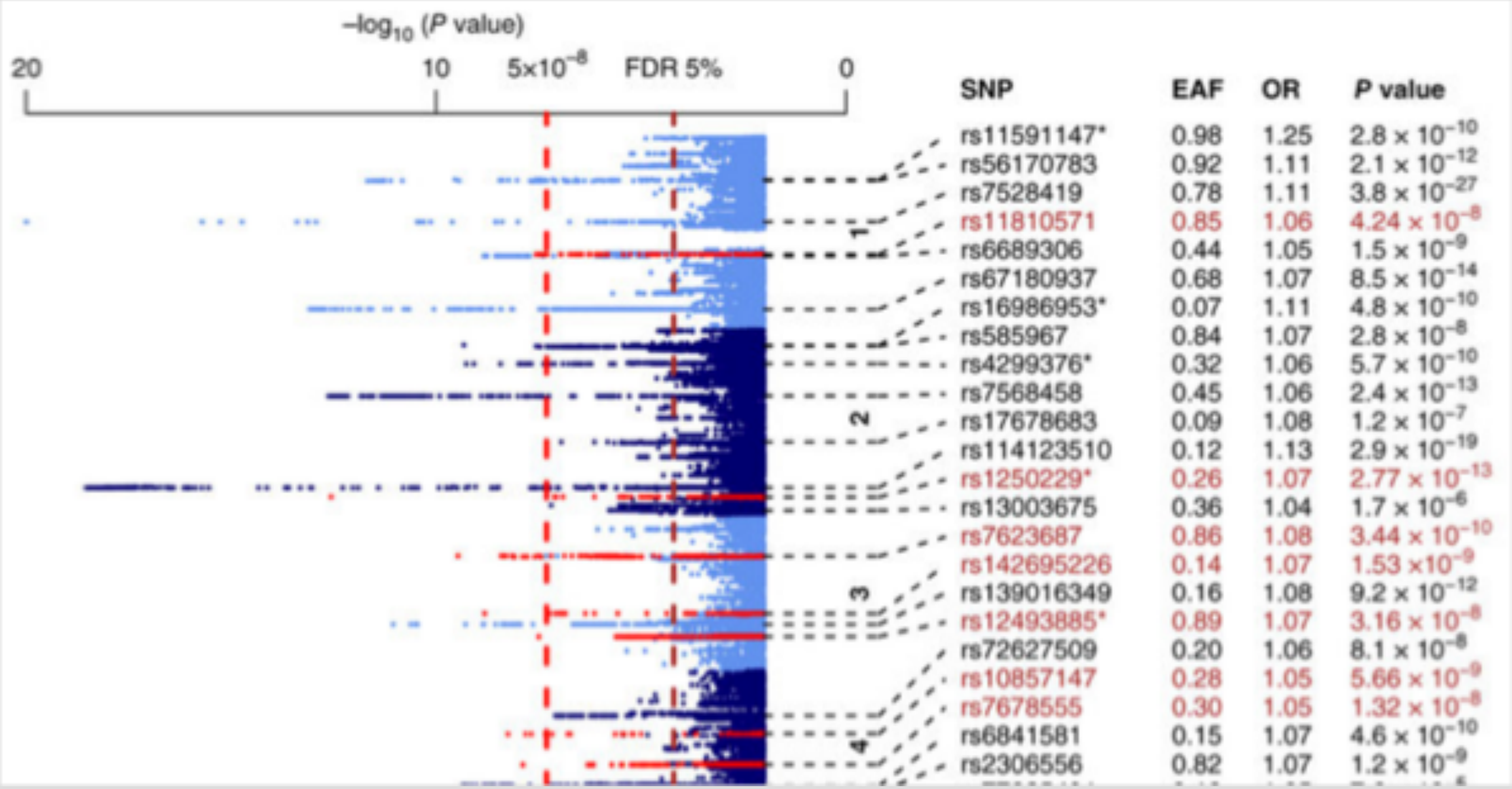


```
IND1 M A/A A/C A/C C/C A/A A/C A/C C/C C/C
IND2 M A/A A/C A/A C/C A/C A/C A/C C/C A/C
IND3 M A/C C/C A/C C/C A/A C/C C/C C/C C/C
IND4 F A/A A/C A/A C/C A/C A/C A/C C/C C/C
IND5 F A/A C/C A/C C/C A/A A/A C/C A/A A/C
IND6 M A/C A/C C/C C/C A/A A/C A/C C/C C/C
IND7 F A/A A/A A/C C/C A/A C/C A/C C/C A/C
IND8 M A/C A/C C/C C/C C/C A/C C/C C/C C/C
IND9 F A/A A/A A/C C/C A/A A/C C/C A/C C/C
```

$p = 10^6$



WHICH VARIANTS ARE INTERESTING?



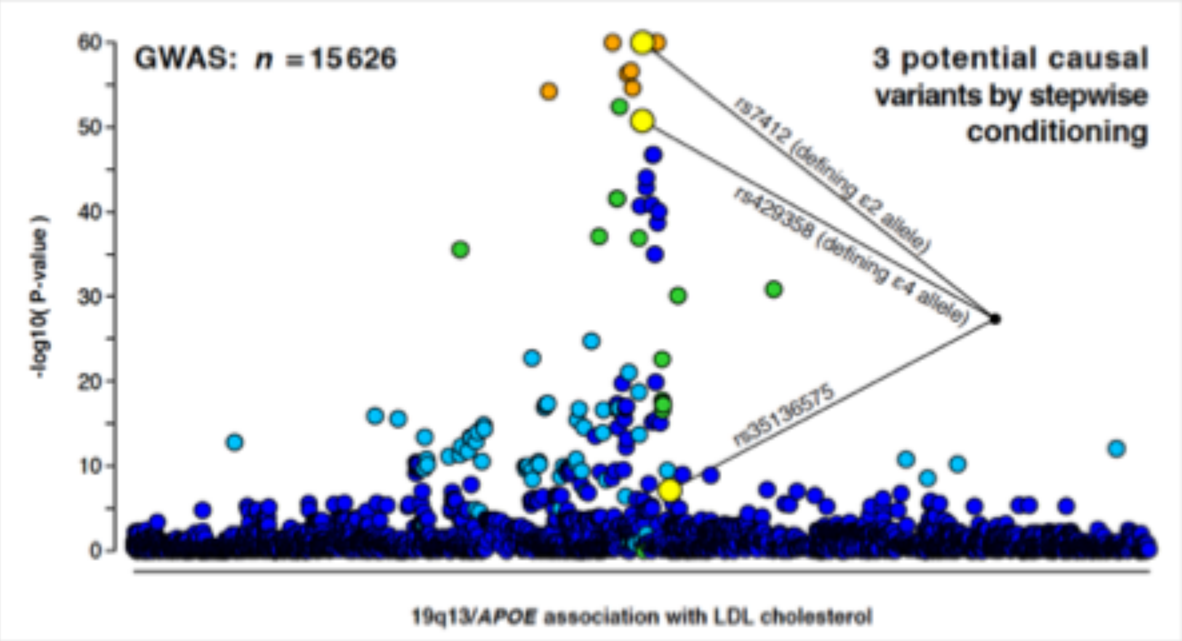
SNP	EAF	OR	P value
rs11591147*	0.98	1.25	2.8×10^{-10}
rs56170783	0.92	1.11	2.1×10^{-12}
rs7528419	0.78	1.11	3.8×10^{-27}
rs11810571	0.85	1.06	4.24×10^{-8}
rs6689306	0.44	1.05	1.5×10^{-9}
rs67180937	0.68	1.07	8.5×10^{-14}
rs16986953*	0.07	1.11	4.8×10^{-10}
rs585967	0.84	1.07	2.8×10^{-8}
rs4299376*	0.32	1.06	5.7×10^{-10}
rs7568458	0.45	1.06	2.4×10^{-13}
rs17678683	0.09	1.08	1.2×10^{-7}
rs114123510	0.12	1.13	2.9×10^{-19}
rs1250229*	0.26	1.07	2.77×10^{-13}
rs13003675	0.36	1.04	1.7×10^{-6}
rs7623687	0.86	1.08	3.44×10^{-10}
rs142695226	0.14	1.07	1.53×10^{-9}
rs139016349	0.16	1.08	9.2×10^{-12}
rs12493885*	0.89	1.07	3.16×10^{-8}
rs72627509	0.20	1.06	8.1×10^{-8}
rs10857147	0.28	1.05	5.66×10^{-9}
rs7678555	0.30	1.05	1.32×10^{-8}
rs6841581	0.15	1.07	4.6×10^{-10}
rs2306556	0.82	1.07	1.2×10^{-9}

Nelson et al. Nat Genetics 2017

SNP Name of the variant
EAF Effect allele frequency
OR Odds ratio

- Each variant is tested for statistical difference between cases and controls
- Millions of tests, how to do inference?
- In Fig. what kind of measures are
  - P-value
  - FDR 5%

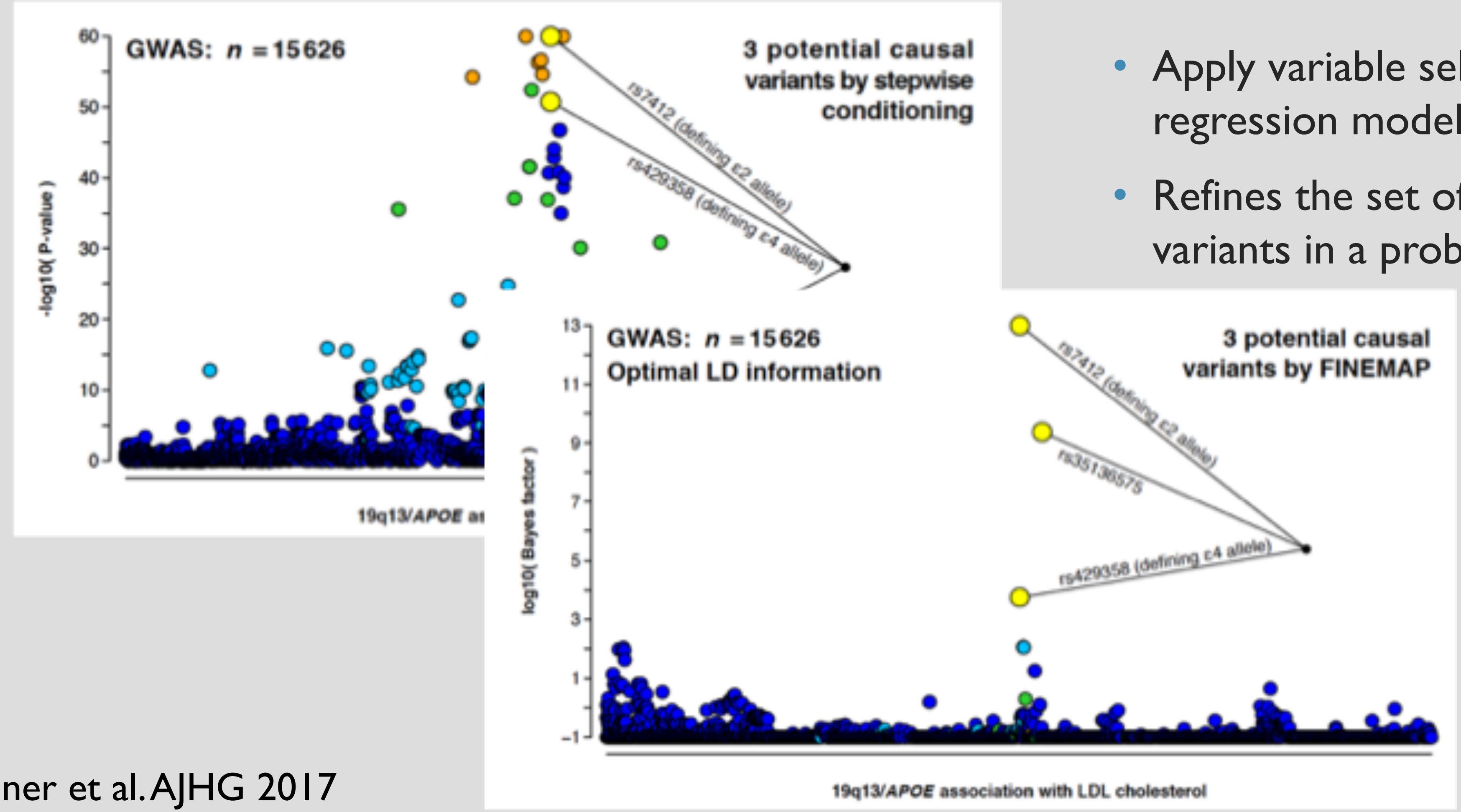
WHICH VARIANTS ARE *CAUSAL* (NOT JUST ASSOCIATED)?



- Variants physically near each other are highly correlated and show similar effect sizes / P-values
- Which one(s) of them is truly driving the signal and which are just passengers?
- We need to analyze them jointly, which becomes a HD regression problem

Benner et al. AJHG 2017

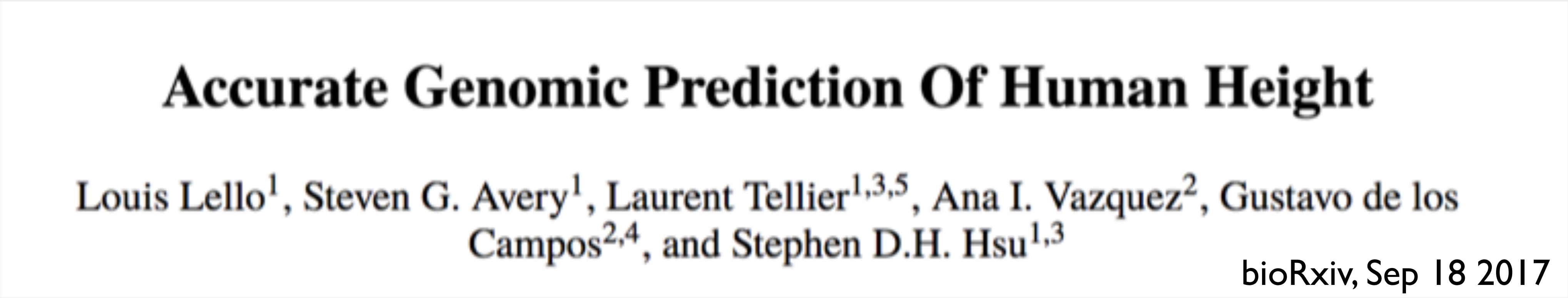
WHICH VARIANTS ARE *CAUSAL* (NOT JUST ASSOCIATED)?



- Apply variable selection in regression model
- Refines the set of potential causal variants in a probabilistic way

Benner et al. AJHG 2017

PREDICTION MODEL



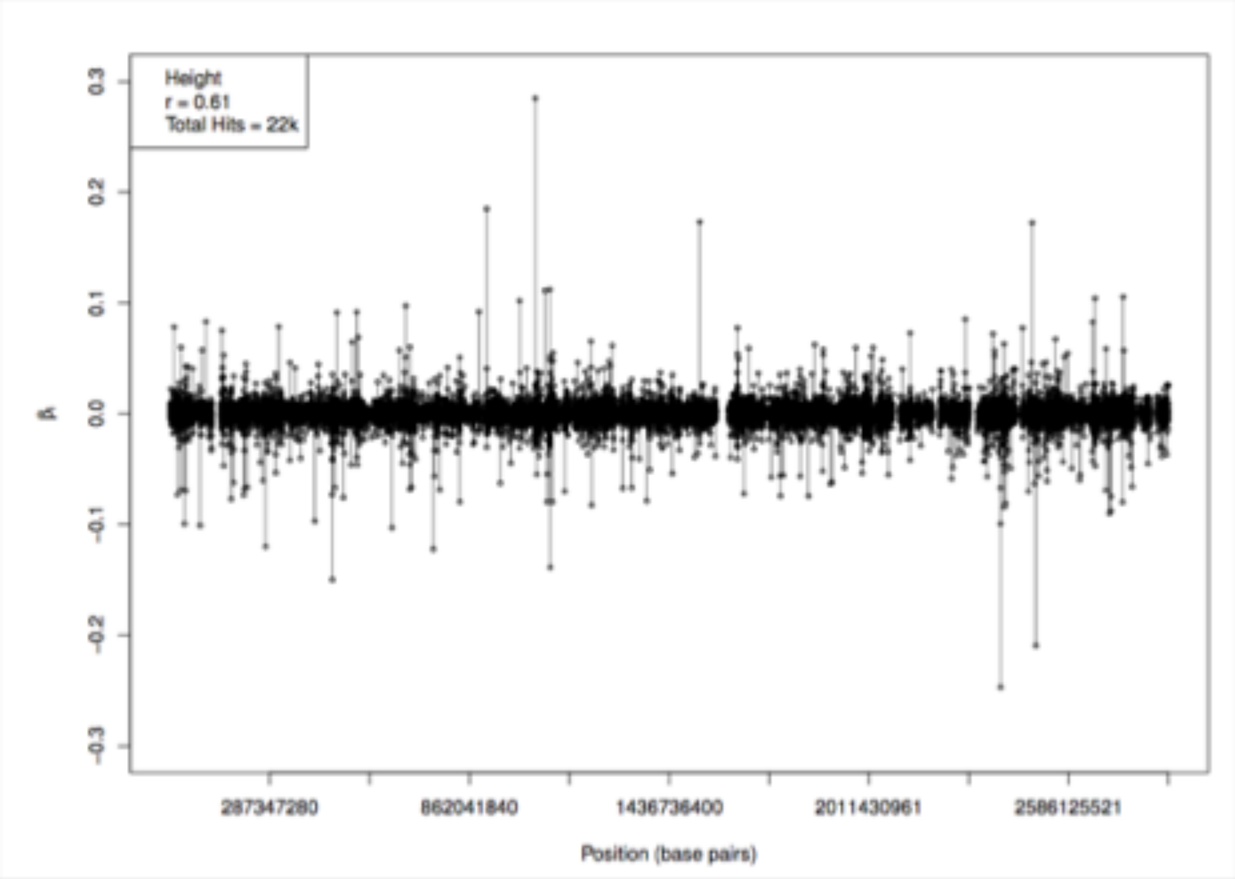
Accurate Genomic Prediction Of Human Height

Louis Lello[1], Steven G. Avery[1], Laurent Tellier[1,3,5], Ana I. Vazquez[2], Gustavo de los Campos[2,4], and Stephen D.H. Hsu[1,3]

bioRxiv, Sep 18 2017

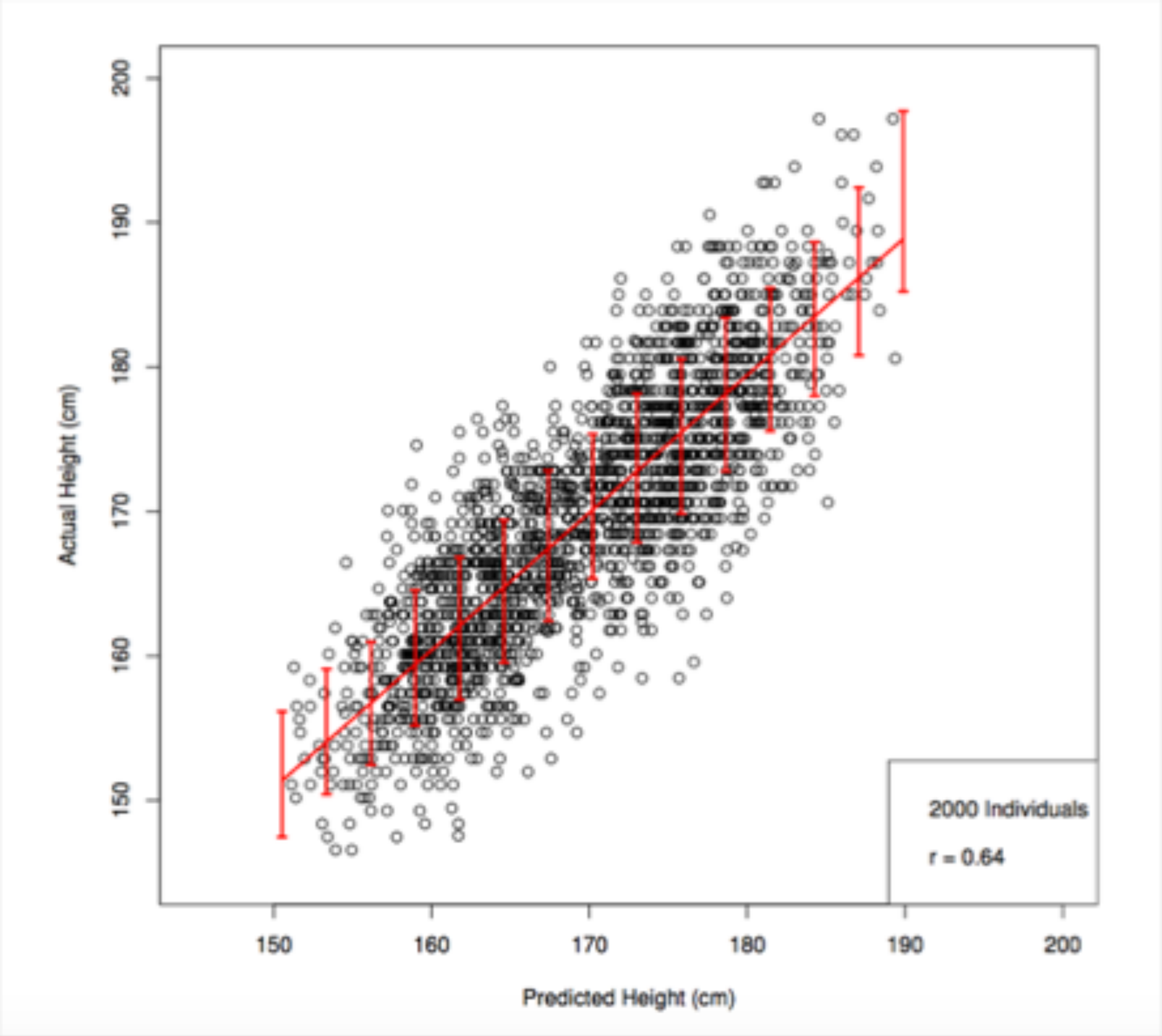
- Start with 650,000 genetic variants and 420,000 individuals with height measurements
- Use LASSO method for building the predictive model (same method we will look at next weeks)

IDENTIFYING RELEVANT VARIANTS



- About 20,000 variants are identified by LASSO and each with its effect size will be used in predicting the height of a new test individual

TESTING THE PREDICTOR



Predictor has correlation of 0.64 with actual height, that is, it explains about 40% (=0.64^2) of variation of height in **the test sample**.

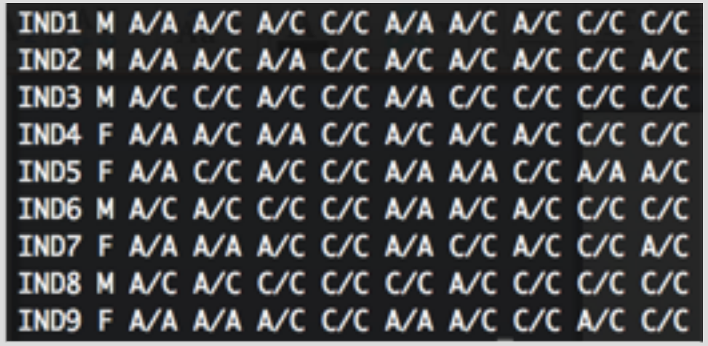
PREDICTION FROM GENOME

- Who is in high risk of getting disease and how can we lower that risk?
- What is the prognosis of the disease and how does that affect treatment?
- Which medication is suitable for a particular individual?

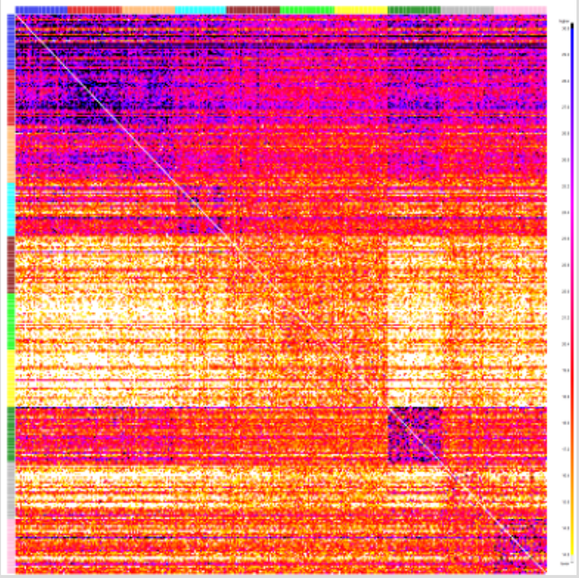
- Predict visible / measurable characteristics from DNA, e.g., for forensic purposes

- Predicting other characteristic of an individual from DNA?
    - Needs also ethical considerations

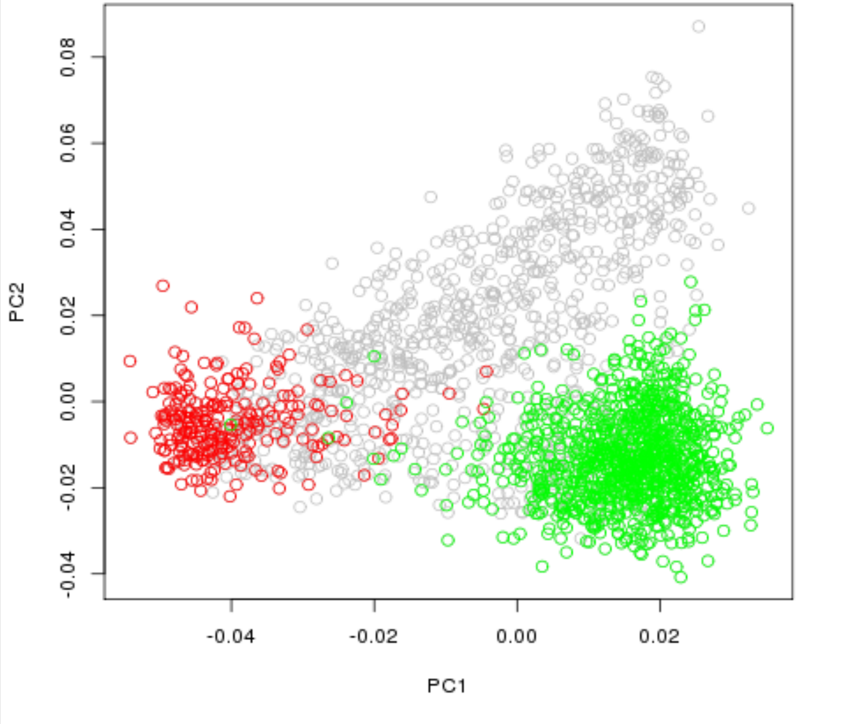
DIMENSION REDUCTION



```
IND1 M A/A A/C A/C C/C A/A A/C A/C C/C C/C
IND2 M A/A A/C A/A C/C A/C A/C A/C C/C A/C
IND3 M A/C C/C A/C C/C A/A C/C C/C C/C C/C
IND4 F A/A A/C A/A C/C A/C A/C A/C C/C C/C
IND5 F A/A C/C A/C C/C A/A A/A C/C A/A A/C
IND6 M A/C A/C C/C C/C A/A A/C A/C C/C C/C
IND7 F A/A A/A A/C C/C A/A C/C A/C C/C A/C
IND8 M A/C A/C C/C C/C C/C A/C C/C C/C C/C
IND9 F A/A A/A A/C C/C A/A A/C C/C A/C C/C
```



Red from Western Finland
Green from Eastern Finland



- From genotypes (10^3 x 10^6) to pairwise covariances (10^3 x 10^3) to first 2 principal components (2 x 10^3)
- Reduction is of order 10^5 and main structure is not only preserved but has also become more easily visible