Introduction to Statistical Genetics

Max Turgeon

STAT 4690–Applied Multivariate Analysis

Overview i

- We will look at three papers that use PCA in slightly different ways:
  1. Price *et al*. "Principal components analysis corrects for stratification in genome-wide association studies." *Nature genetics* (2006).
  2. Leek & Storey. "Capturing heterogeneity in gene expression studies by surrogate variable analysis." *PLoS genetics* (2007).
  3. Gao *et al*. "A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms." *Genetic epidemiology* (2008).

Overview ii

- The main purpose of this lecture is to:
  - Introduce you to important concepts in applied statistics (e.g. confounding and multiple testing).
  - Give you a sense of the versatility of PCA.
  - Give an overview of the interplay between theoretical, methodological and applied research in statistics.
- All three papers can be found on UM Learn (or online).

Introduction to Genetics

DNA

- Long molecule, double-stranded, made of four types of nucleotides:
  - **T**hymine
  - **C**ytosine
  - **G**uanine
  - **A**denine
- Nucleotides are paired:
  - A-T and C-G
- This pairing allows *replication*:
  - DNA molecule opens up
  - From complimentarity, we can reconstruct two molecules.

Central Dogma

- Explains how DNA leads to proteins
- DNA $\Longrightarrow$ RNA $\Longrightarrow$ Protein
  - $(T, C, G, A) \Longrightarrow (U, C, G, A)$
  - Codon (i.e. triple) $\Longrightarrow$ Amino acid
- **Gene**: sequence of nucleotides that encodes a protein
  - Other gene products are possible: microRNA, tRNA, etc.

Genetic variation

- Random mutations
- After fertilization, a zygote has a copy of each chromosome from each parent
  - *Assortment* is random
- Before that, at meiosis, there is *recombination*
- At the population level:
  - Population bottleneck
  - Founder effect
  - Natural selection
- The most studied genetic variation: *Single Nucleotide Polymorphism* (SNP)
  - A location in the genome where in the population we observe at least different nucleotides

Some vocabulary

- **Allele**: Sequence observed at a specific location
  - One basepair for SNP
  - Can be longer
- **Minor/Major Allele**: Least/Most observed allele in a population
- **MAF**: Minor Allele Frequency
  - Frequency at which the minor allele is observed in the population
  - *Population specific*
- **Phenotype**: Observable characteristic or trait

Gene Expression

- All cells have the same DNA, but they produce different proteins.
- Same cell type, under different conditions, can also produce different proteins.
- Different mechanisms:
  - Transcription factors
  - Epigenetics

Population Stratification

High-throughput technologies

- Since the mid-2000s, SNP data is routinely collected at hundreds of thousands, or even millions, of genetic loci.
- There are two basic types of technologies:
  1. *Micro-arrays*: Designed to identify the allele at pre-selected loci
  2. *Next-generation sequencing*: Sequence large portions of DNA.
- The data is similar: high-dimensional data (i.e. more variables than observations).

Genome-Wide Association Studies

- **GWAS**: Every genetic measurement is tested for association with a single (or a few) *phenotype* of interest.
- *Goal*: Find genetic locations with evidence of causal effect on disease of interest
  - Or at least genetic locations that inherited together with causal locus
- Two main challenges:
  - Multiple testing (we'll come back to it)
  - Population stratification (i.e. confounding)

Confounding

- **Confounder**: common cause of both the exposure and outcome of interest
  - E.g. Obesity is a cause of diabetes and cardiovascular diseases.
- Failure to adjust for confounding can lead to spurious correlations
- Three main methods for confounder adjustment:
  - Randomisation
  - Regression model
  - Weighting

Population stratification as a confounder

- Because of migration patterns and natural selection, some alleles are preferentially selected in certain populations
  - E.g. *LCT* gene and lactose intolerance.
- **Population stratification**: "allele frequency differences between cases and controls due to systematic ancestry differences" (Price et al)
- If a given allele and the phenotype of interest are more prevalent in a certain population, this may give rise to spurious correlation.
- **Major problem**: Population stratification is very hard (if not impossible) to measure accurately.
- *Solution*: Estimate it from the collected genetic data.

EIGENSTRAT i

- Price et al. (2006) proposed a method to adjust for population stratification in GWASs.
- Essentially, the population stratification is estimated using the principal components of the genetic data.
- More precisely, let G be the $n \times p$ matrix of genotypes
  - The (i, j)-th entry g_{ij} is the value at the j-th locus for the i-th sample.
  - $g_{ij} \in \{0, 1, 2\}$ counts the number of copies of the minor allele.

EIGENSTRAT ii

- Create matrix X by normalizing G
  - Subtract the mean
  - Divide by binomial standard deviation $\sqrt{p_j(1-p_j)}$.
- Select first k eigenvalues of the covariance matrix of X.
- Adjust for confounding by including the PCs into a regression model.

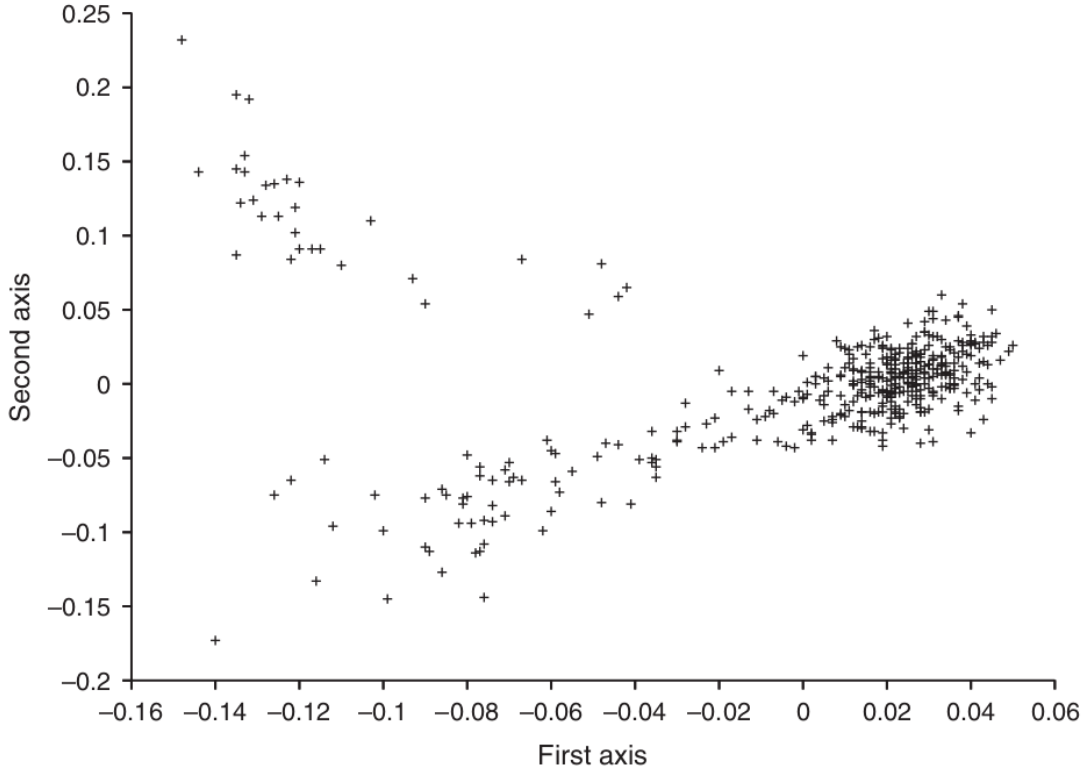

Figure 2 The top two axes of variation of European American samples. We hypothesize that the first axis reflects genetic variation between northwest and southeast Europe, with a fraction of the samples showing southeast European ancestry (first axis < 0; see text). It follows that the second axis separates two southeast European subpopulations.

Figure 1

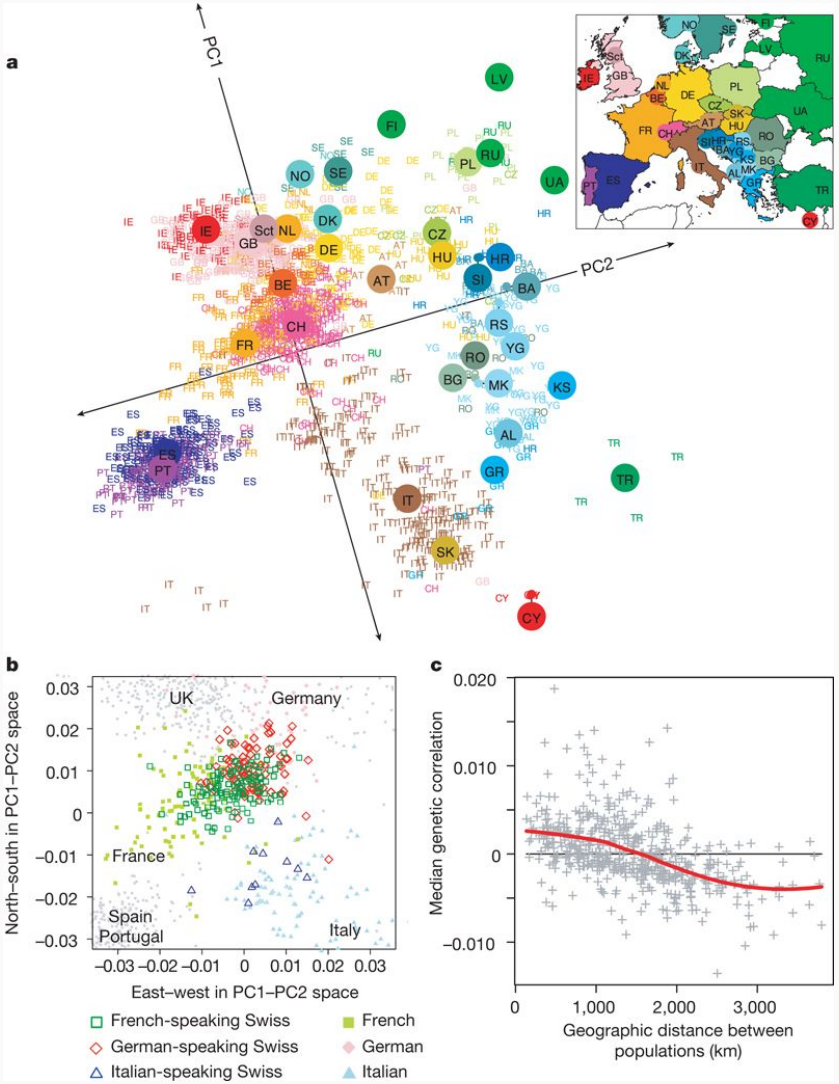

Figure 2

Novembre *et al*. "Genes mirror geography within Europe."

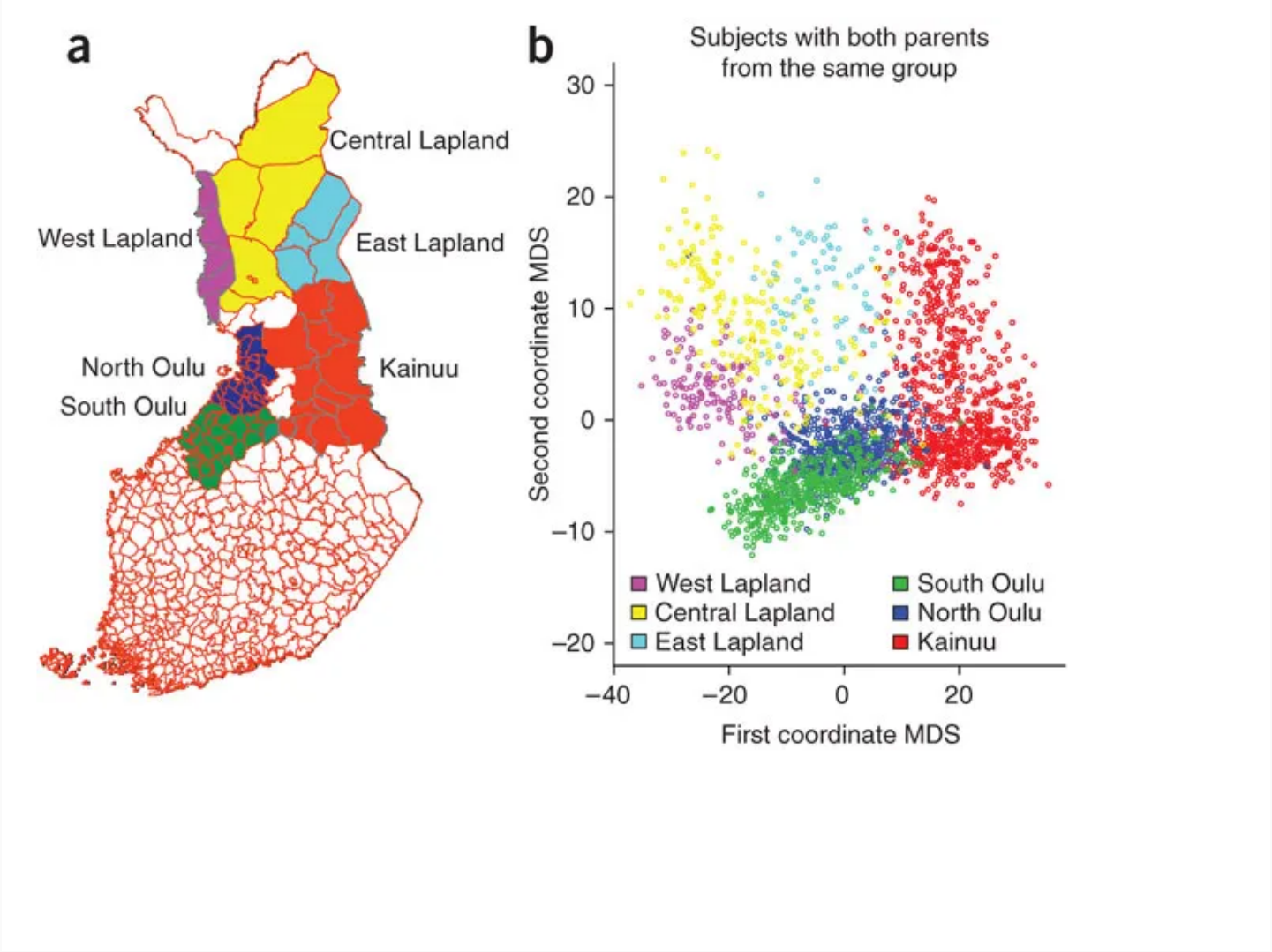

Figure 3

Sabatti *et al*. "Genome-wide association analysis of metabolic traits in a birth cohort from a founder population."

Further comments

- There is a vast literature around how to use PCA to account for population stratification
  - How many PCs to retain.
  - Theoretical justification.
  - Power analysis.
  - How granular can you get.
- *Note*: This is not how 23andMe and AncestryDNA estimate your ethnicity!
- PCA can also be used to estimate under population substructures in your data.
  - E.g. Cryptic relatedness