Faculty of Mathematics and Statistics
University of Ottawa

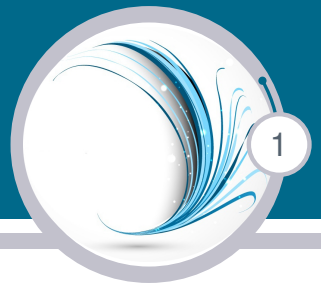
A comparison of kernel-based association statistics, with application to Crohn's disease (soon!)

Honour's Project

Peter Tea
ptea035@uottawa.ca

October 31, 2018

Content



Introduction
Background and Rationale

- The first DNA sequences were obtained in the early 1970s by academic researchers.

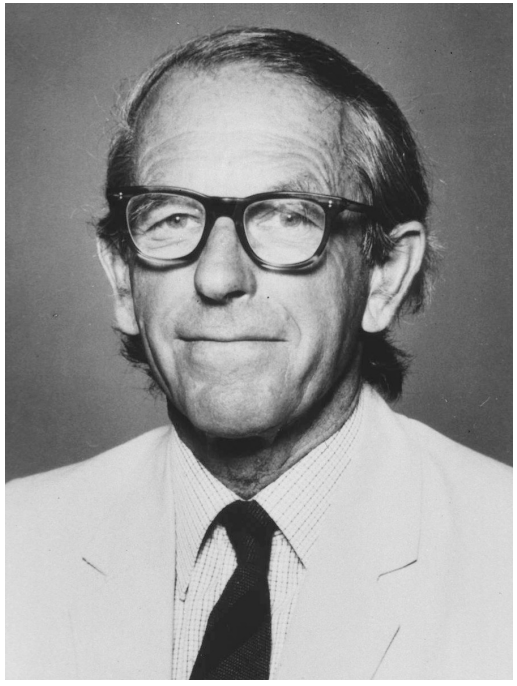
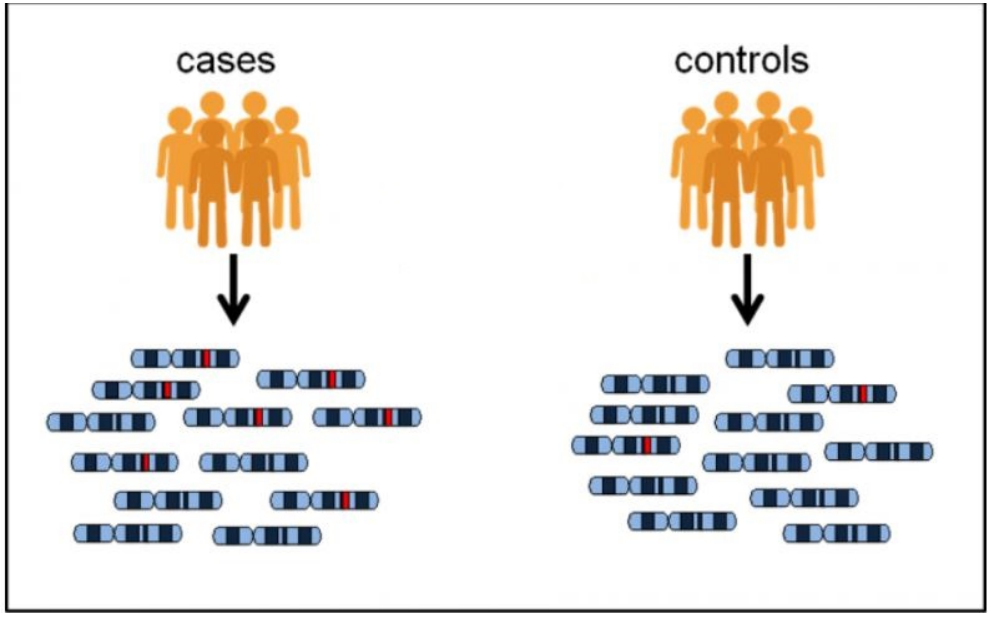


Figure: Frederick Sanger, 1975

- Since then, DNA sequencing has become much easier and faster to execute.
- Today, the human genome can be sequenced in one hour (Illumina).
- Human genome composed of $\sim$ 3 billion base pairs!

Introduction
Genetic Association Studies

- We would like to detect association between genetic variation and a phenotype of interest.

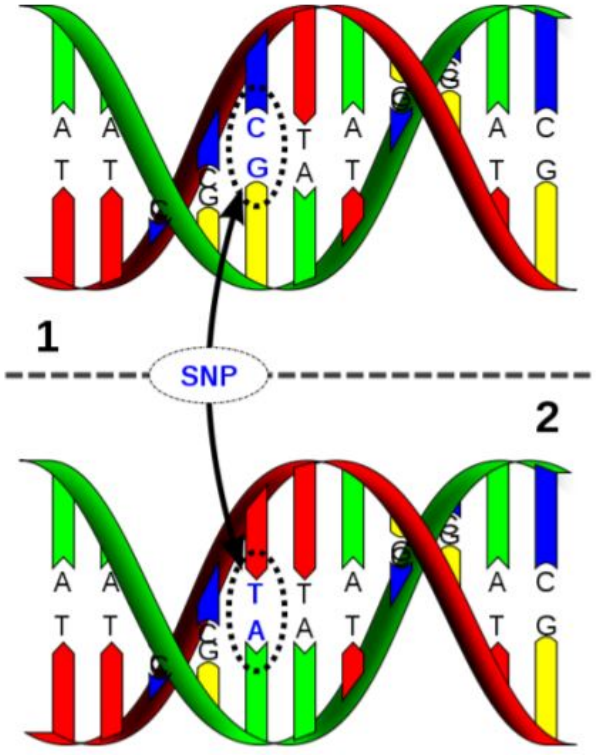


Source: [1]

[1] https://www.ebi.ac.uk/training/online/course/gwas-catalog-exploring-snp-trait-associations/why-do-we-need-gwas-catalog/what-are-genome

How?
Genetic Association Studies

Single **N**ucleotide **P**olymorphism



- Humans only share $\sim$ %99.9 of their genomes.
- Much of human genome variation comes in the form of SNPs. These are variations that involve just one nucleotide.

[2]

[2]Nyholt, Dale. (2017) “Gene-environment interaction in migraine". Queensland University of Technology. Institute of Health and Biomedical Innovation.

Kernel Based Association Statistics

Kernel-based association methodologies

1. Specification of kernel function - (outputs a map that describes the degree of genetic similarity between pairs of individuals).
2. Application of a kernel based association statistic to measure the strength of association between genetic similarity with a trait of interest.

Kernel Based Association Statistics

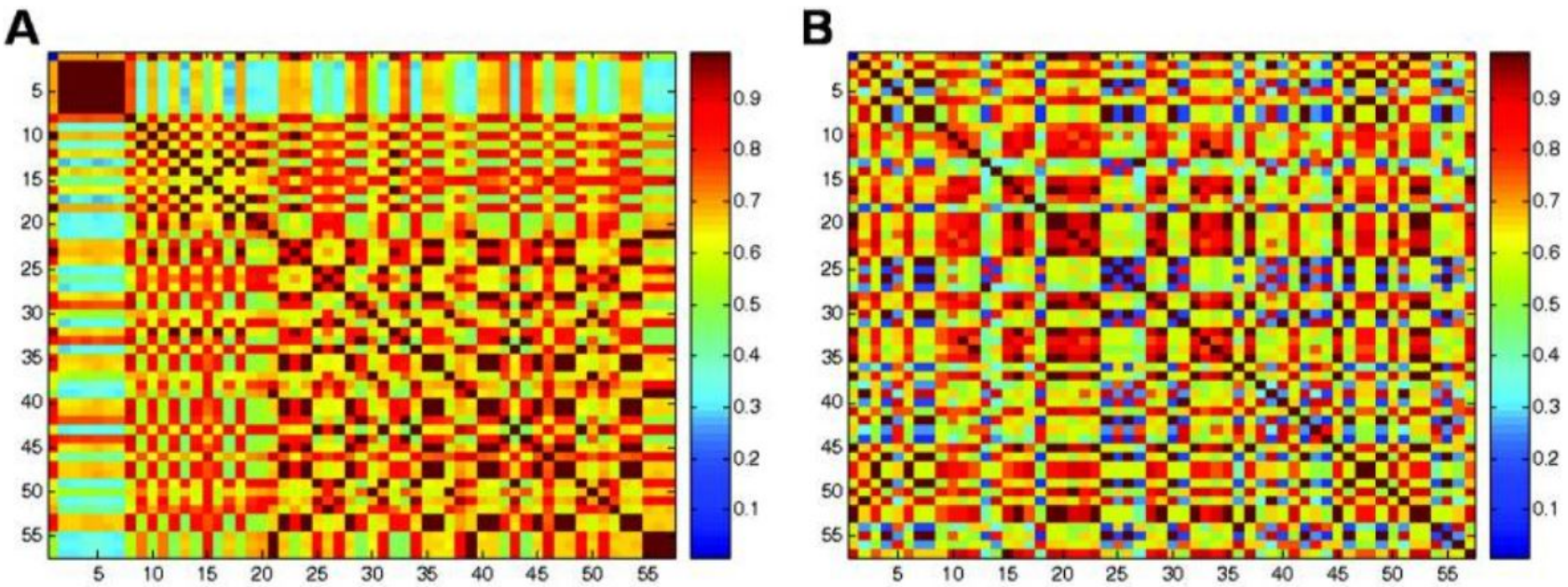
Kernel function

- A kernel function outputs a map that describes the degree of genetic similarity between pairs of individuals.
- Two main strategies: Scoring genotype similarity and tree-based approach.

Kernel Based Association Statistics

Kernel function: Scoring genotype similarity

- Count the number of shared alleles across all SNP sites

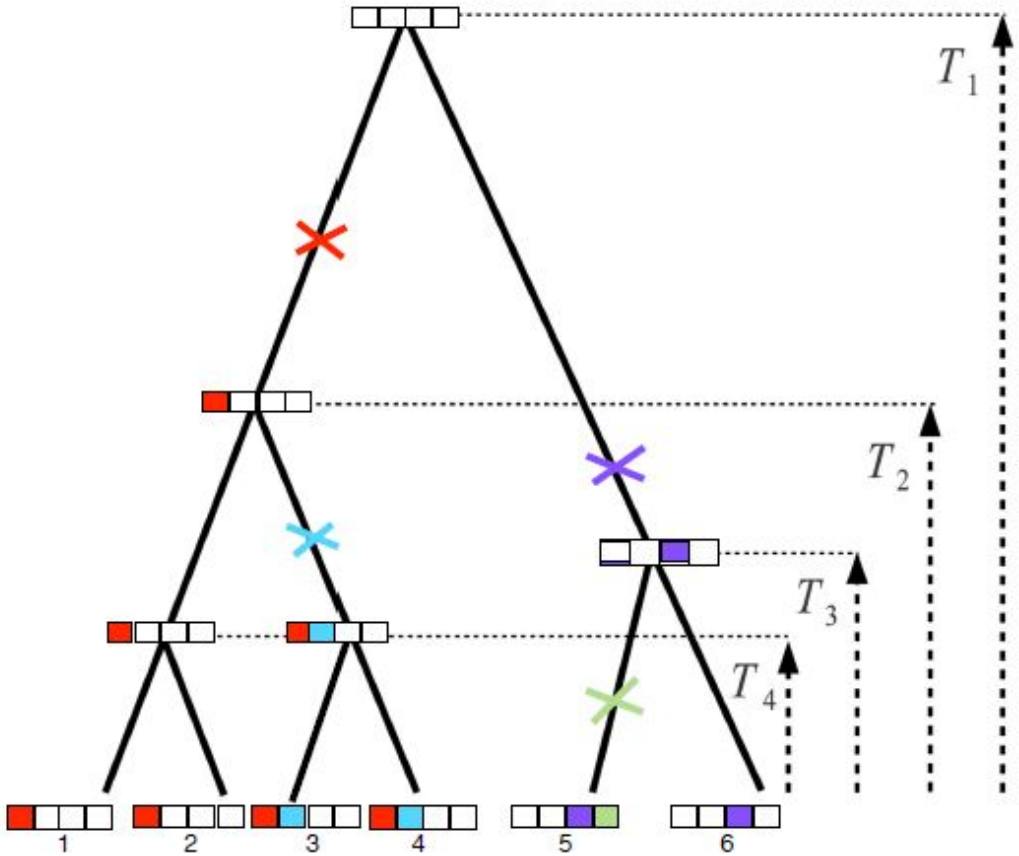


3

[3]Wessel, J.(2006). Generalized Genomic Distance-Based Regression Methodology for Multilocus Association Analysis. American Journal of Human Genetics.

Kernel Based Association Statistics

Kernel function: Tree kernel



4

[4] https://www.ebi.ac.uk/training/online/course/gwas-catalog-exploring-snp-trait-associations/why-do-we-need-gwas-catalog/what-are-genome

Kernel-based association statistics

- Many statistic approaches assume a regression model:

$$Y_i = \beta_0 + \beta_1^T G_{ij} + \epsilon_i$$

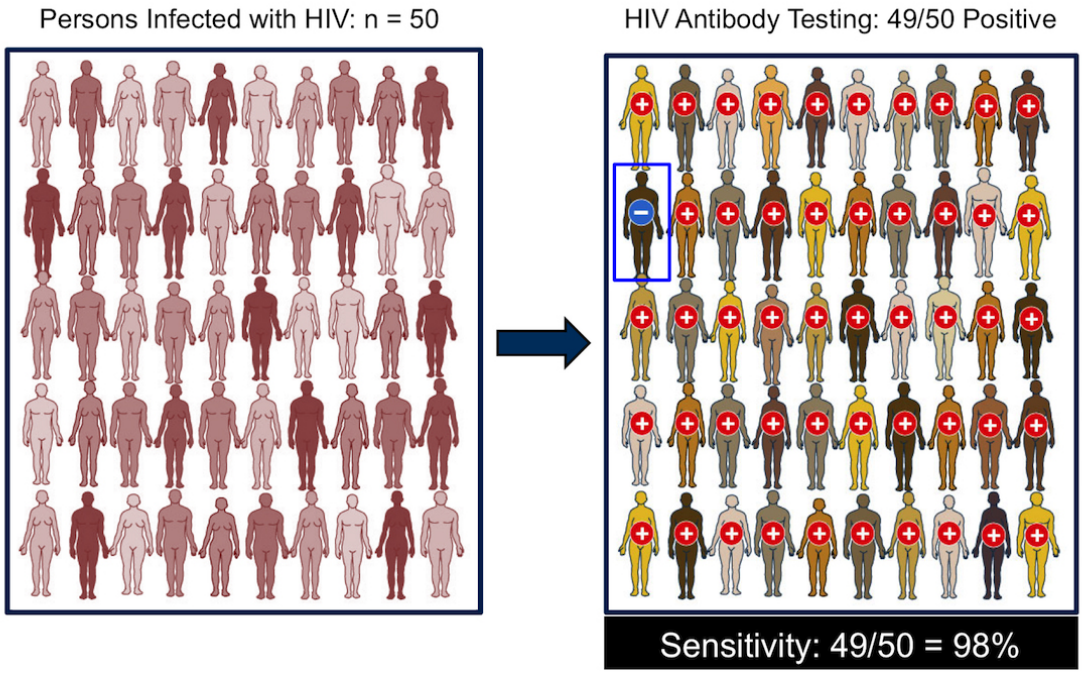
- Hypothesis tests are all the same:

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0$$

  - Gene-Trait Similarity Regression (GTSR) - a generalized linear regression model
  - Multi-locus association analysis (MDMR) - a genomic distance-based regression
  - Sequence kernel association test (SKAT) - a variance-component score statistic

Simulation
Sensitivity

Sensitivity: A diagnostic test's ability to correctly diagnose patients with the disease (i.e. true positive rate).



5

[5]https://www.hiv.uw.edu/go/screening-diagnosis/diagnostic-testing/core-concept/all

Simulation
Data

Phenotypes were simulated from a normal distribution where the mean depends on the number of causal variants.
Phenotype $\sim$ N(μ_*, 1)

$$\mu_0 = 0 \cdot \beta \quad \text{[Homozygous Dominant]}$$
$$\mu_1 = 1 \cdot \beta \quad \text{[Heterozygous]}$$
$$\mu_2 = 2 \cdot \beta \quad \text{[Homozygous Recessive]}$$

Simulation
Data

Phenotype $\sim$ N(μ_*, 1)
Similarly for phenotype 2, the means for each individual was calculated by multiplying the marginal number of rare variants across all causal sites an individual possesses by the parameter β.

To illustrate this calculation, please consider the following:
Assume that the rare causal SNP sites are at these locus: 1,3,5,7, and 9. If an individual is homozygous dominant at sites 1,3, and 5, but is heterozygous at sites 7 and 9 then this individual possesses in total 2 rare causal variants.

Results
Power calculations

Phenotype \ β	0.25	0.3	0.35	0.4	0.45	0.5
P1	0.345	0.465	0.586	0.714	0.808	0.881
P2	0.134	0.175	0.236	0.287	0.352	0.413

Phenotype \ β	0.55	0.6	0.65	0.7	0.75	0.8
P1	0.934	0.964	0.984	0.994	0.998	0.999
P2	0.485	0.564	0.637	0.690	0.751	0.805

Phenotype \ β	0.85	0.9	0.95	1
P1	0.999	0.999	1	1
P2	0.846	0.883	0.914	0.940

Figure: Reported power of single locus tests on detecting genotype - phenotype associations. Parameters that gave an estimate power of 0.8 were chosen for simulation studies to be conducted in the honour's project.

Results: Phenotype 1

True positive rate

Statistic \ Kernel	IBS	AM	AS	LIN	REC	QUAD	H1	Skat
GTSR	0.040	0.040	0.049	0.020	0.040	0.024	0.038	0.033
MDMR	0.492	0.497	0.412	0.348	0.180	0.322	0.519	0.071
SKAT	0.488	0.488	0.488	0.000	0.007	0.004	0.514	0.080

Table 1.2: SENSITIVITY 1/2

Statistic \ Kernel	012	123	124	Tree1	Tree2	Tree3	Tree4	Tree5
GTSR	0.047	0.029	0.040	0.042	0.042	0.047	0.042	0.042
MDMR	0.455	0.490	0.475	0.419	0.426	0.160	0.175	0.169
SKAT	0.514	0.514	0.503	0.448	0.448	0.122	0.124	0.124

Table 1.3: SENSITIVITY 2/2

Results: Phenotype 2

True positive rate

Statistic \ Kernel	IBS	AM	AS	LIN	REC	QUAD	H1	Skat
GTSR	0.038	0.038	0.038	0.031	0.038	0.038	0.035	0.049
MDMR	0.302	0.297	0.310	0.282	0.080	0.233	0.255	0.703
SKAT	0.295	0.295	0.295	0.009	0.002	0.022	0.248	0.789

Table 1.4: SENSITIVITY 1/2

Statistic \ Kernel	012	123	124	Tree1	Tree2	Tree3	Tree4	Tree5
GTSR	0.044	0.040	0.038	0.047	0.047	0.053	0.049	0.049
MDMR	0.233	0.259	0.184	0.213	0.204	0.304	0.310	0.299
SKAT	0.248	0.248	0.197	0.217	0.217	0.149	0.155	0.155

Table 1.5: SENSITIVITY 2/2