

# Privacy with genetic data

Caroline Uhler

Department of Statistics

UC Berkeley

CDI Kickoff Meeting

June 10, 2011

# Main problem

**Aug. 2008 PLOS Genetics paper by Homer et al.:**

Possible to identify individual in a mix of DNA samples (containing only aggregate data)

## **Consequences:**

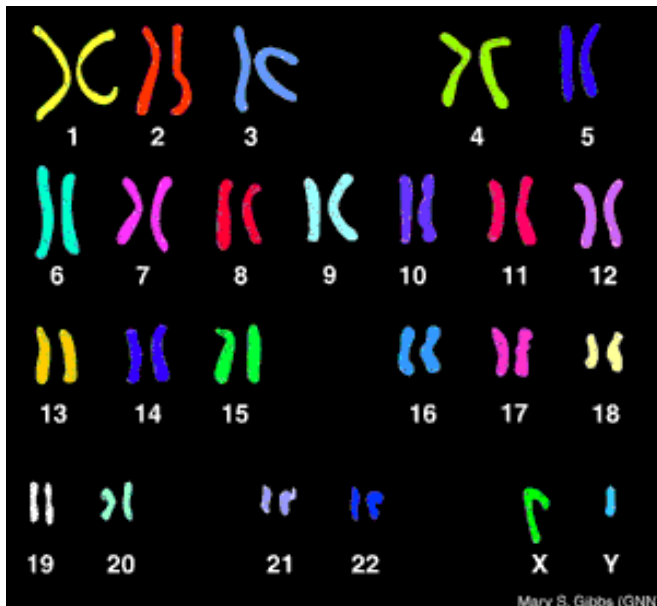
- NIH removed all aggregate genomic data from public websites
- Approval process in order to get access

# Outline

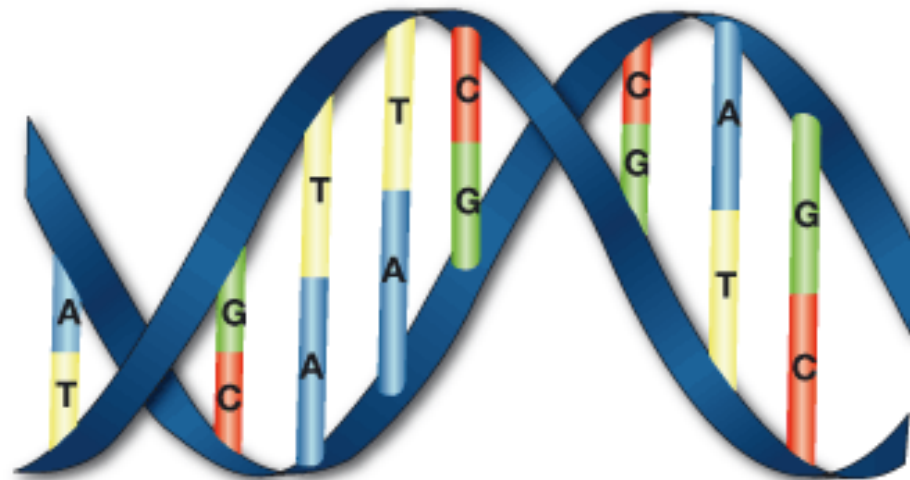
- Genetics background
- Genetic data
- Aug. 2008 PLOS Genetics paper by Homer et al.
- Proposed project: Privacy with genetic data

# Chromosomes / DNA

Chromosomes:



DNA:



- ! Every human has almost identical DNA: 1 difference in 1200 sites (600'000 – 1'000'000 total) ➔ **SNPs**

# SNPs

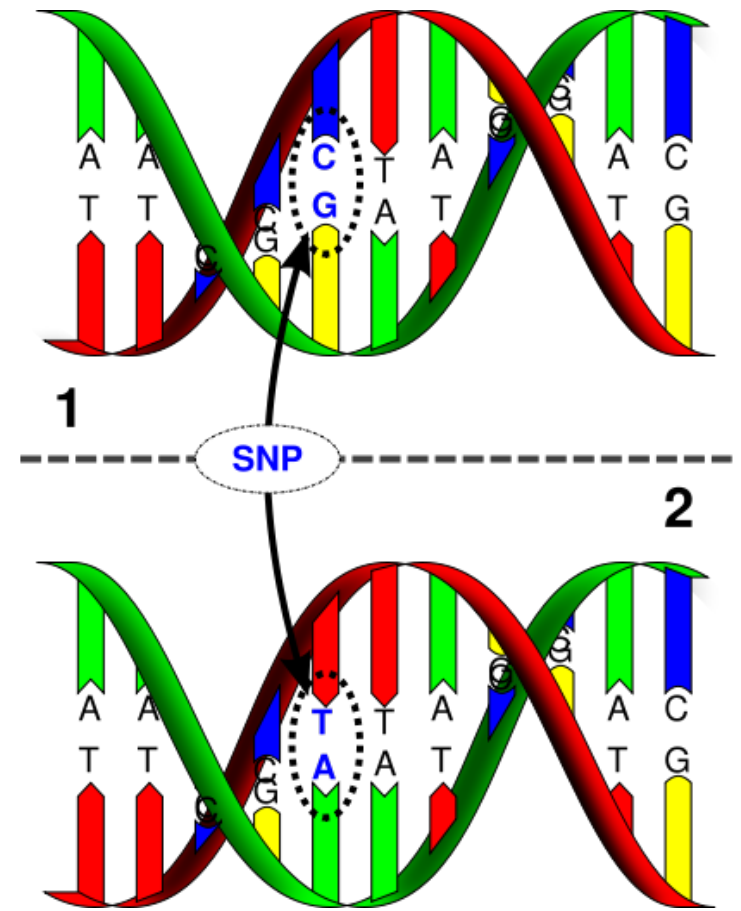
! In SNPs usually only 2 different nucleotides occur ➔ **alleles**

Ex: Common allele: C

Minor allele: T

➔ Possible allele combinations and encoding:

CC	CT	TC	TT
0	1	1	2



# Genetic data

## HapMap:

- SNP data from 4 populations (i.e. Nigeria, CEU, Japan, Han Chinese) of ~60 individuals each
- Not linked to any phenotype data
- Publicly available before and after Aug. 2008

## GWAS (genome-wide association studies):

- SNP data including phenotype
- Sharing only upon request and approval
- ! Before Aug. 2008 minor allele frequencies for cases and controls publicly available (data sharing even encouraged through funding)

# 2008 PLOS Genetics paper (Homer et al.)

## Data:

- Minor allele frequencies of reference population, e.g. HapMap:  $g_i$
- Minor allele frequencies of test sample, e.g. cases in GWAS:  $h_i$
- Minor allele frequencies of specific individual:  $y_i$

## Test:

- For each SNP compute distance  $D_i(Y) = |y_i - g_i| - |y_i - h_i|$
- Test  $H_0 : D(Y) = 0$  versus  $H_1 : D(Y) > 0$   
using one-sample t-test (based on central limit theorem)
- ➡ Can identify individual even if its genome composes only 0.1% of test sample (using 10'000 – 50'000 SNPs)

# Follow-up research

- In practice, distribution of  $D(Y)$  for null samples deviates strongly from standard normal
- High false-positive rate (Braun et al., PLOS Genetics, Oct. 2009)
- Several authors propose more powerful test statistics, e.g. based on likelihood ratio
- Upper bound on power achievable by any method given by likelihood ratio test; safe to expose set of SNPs for which likelihood ratio test does not achieve sufficient power; power scales as:

$$z_\alpha + z_{1-\beta} \approx \sqrt{m/n}, \quad \text{where } m = |\text{SNPs}|, \quad n = |\text{test sample}|$$

(Sankararaman et al., Nature Genetics, Sep. 2009)

# Proposed project

## Data:

Very sparse contingency tables ( $n \sim 50 - 2000$ ) of size  $3^m \times k$ , where  $m \sim 10'000 - 50'000$  and  $k$  is small, usually  $k = 2$ .

## Aim:

Make 'data' available such that one can

- do association studies, in particular find SNPs which interact to cause a disease (i.e. detect higher-order interaction)
- but at the same time guarantee privacy

# References

- ! Homer et al. (2008), Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays, *PLOS Genetics*.
- Braun et al. (2009), Needles in the haystack: identifying individuals present in pooled genomic data, *PLOS Genetics*.
- Jacobs et al. (2009), A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies, *Nature Genetics*.
- Visscher & Hill (2009), The limits of individual identification from sample allele frequencies: theory and statistical analysis, *PLOS Genetics*.
- Sankararaman (2009), Genomic privacy and limits of individual detection in a pool, *Nature Genetics*.