Genome-wide association study

Segun Fatumo PhD

UNIVERSITY OF CAMBRIDGE

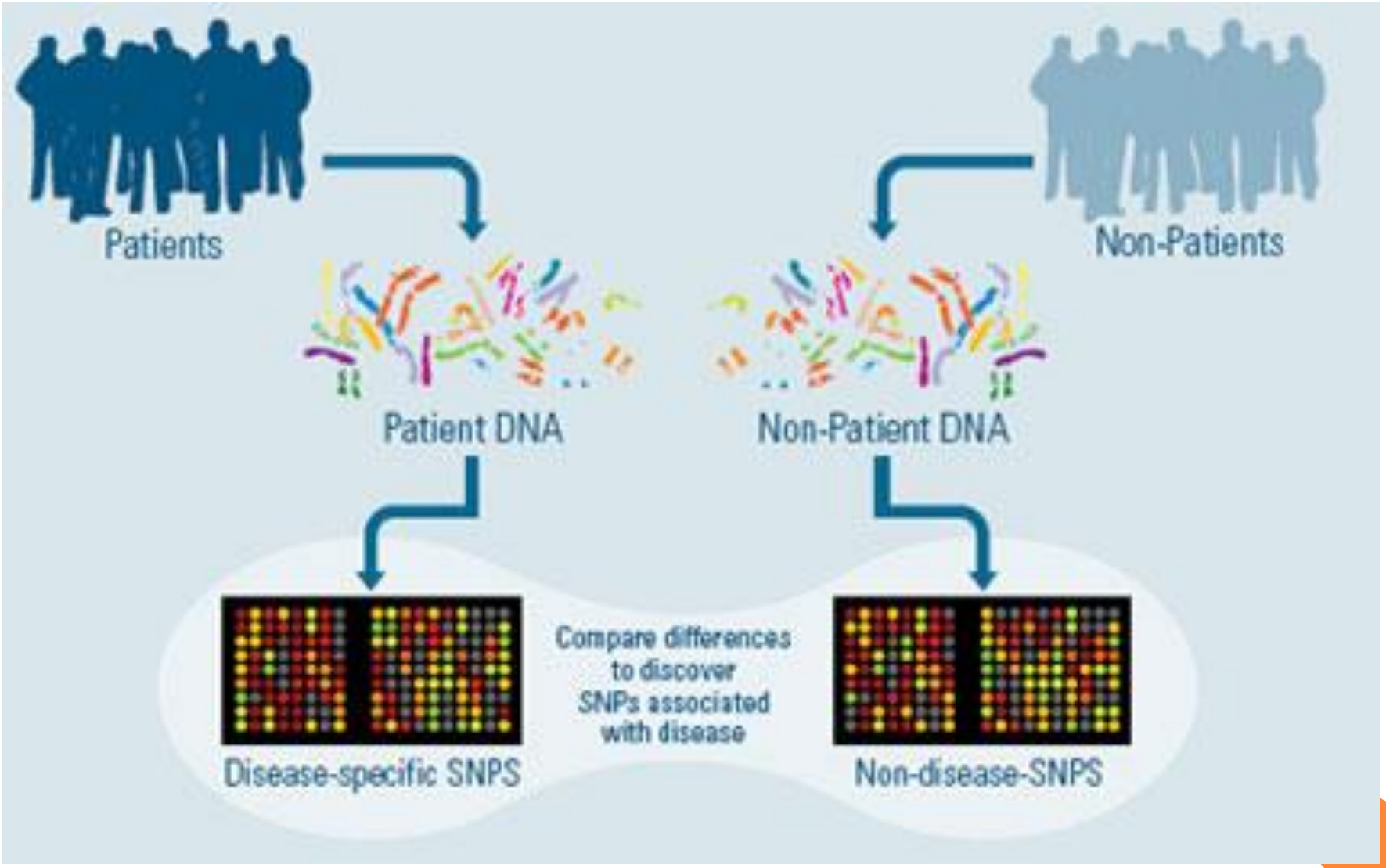
wellcome trust sanger institute

Sessions

- Session I: Overview of GWAS
- Session II: GWAS: Understanding the Data format
- Session III: Quality Control in GWAS
- Session IV: Association Analyses
- Session V: GWAS in Africa
- Session VI: Imputation
- Session VII: Meta-Analysis

Session I: Overview of GWAS

GWAS: THE BIG PICTURE



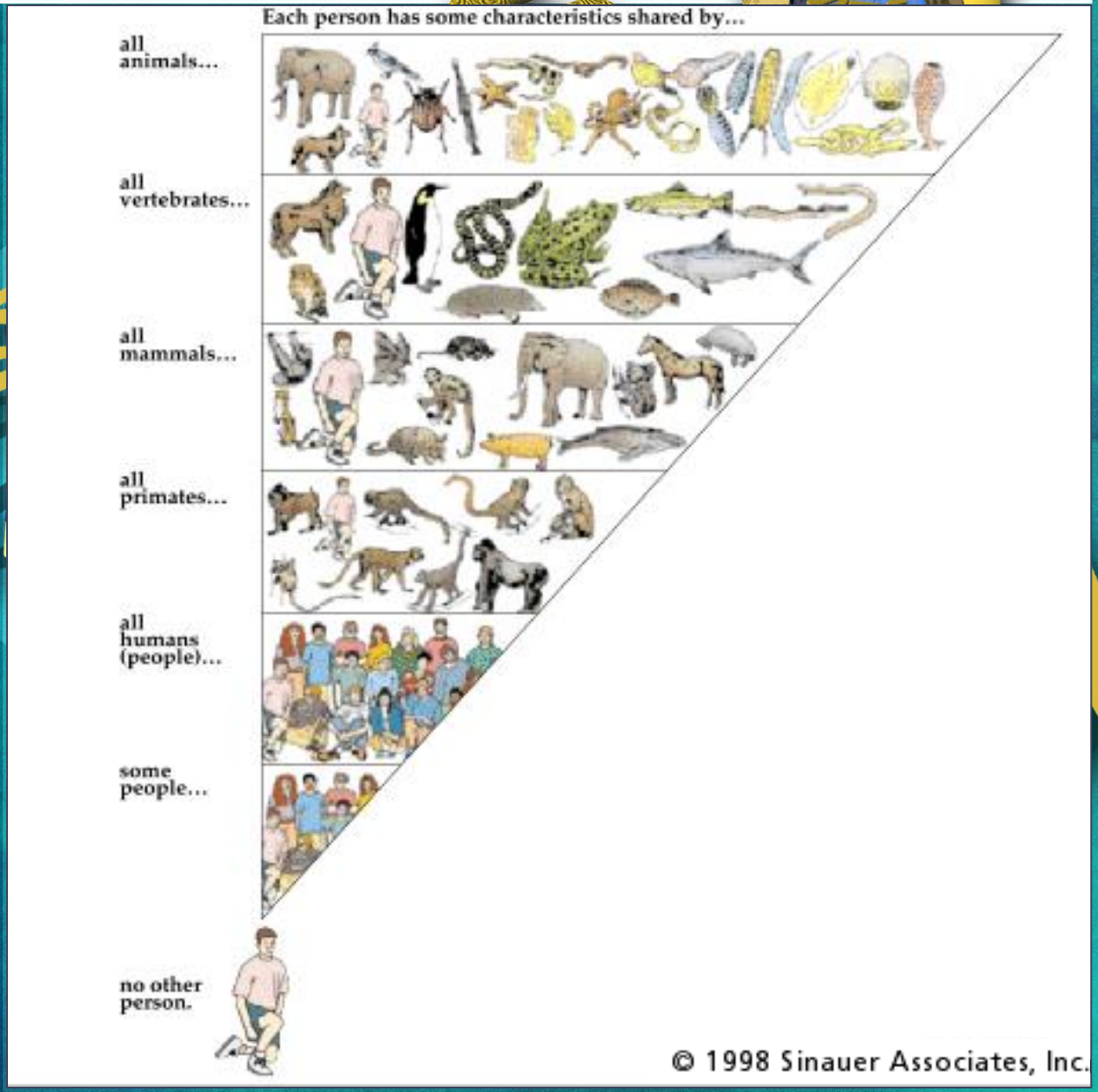
Let's start all over to understand the genesis -Why was Gwas not possible 10 years ago ?

The Human Genome

Trillions of cells

Each cell:

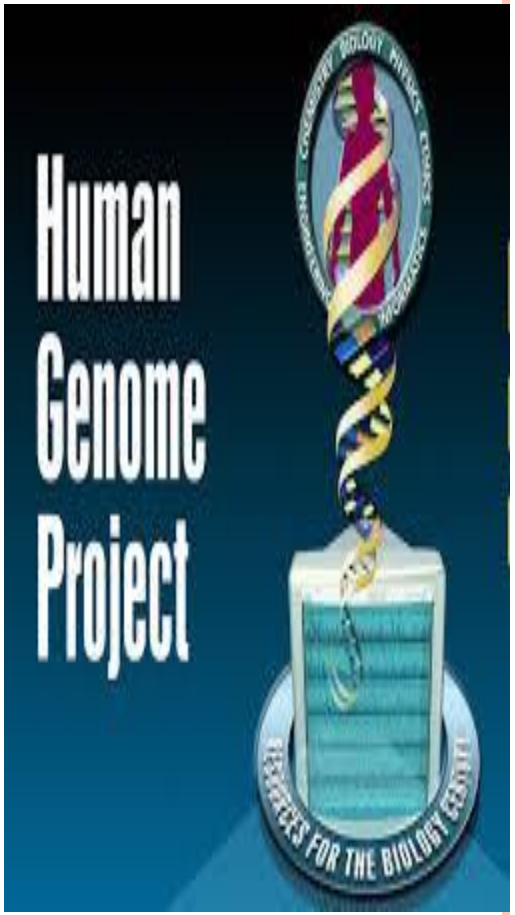
- 46 human chromosomes
- 2 meters of DNA
- 3 billion DNA subunits (the bases: A, T, C, G)
- Approximately 21 000 genes code for proteins that perform most life functions



Human Genome Project

Goals:

- identify all the approximate 30,000 genes in human DNA,
- determine the sequences of the 3 billion chemical base pairs that make up human DNA,
- store this information in databases,
- improve tools for data analysis,
- transfer related technologies to the private sector, and
- address the ethical, legal, and social issues (ELSI) that may arise from the project.

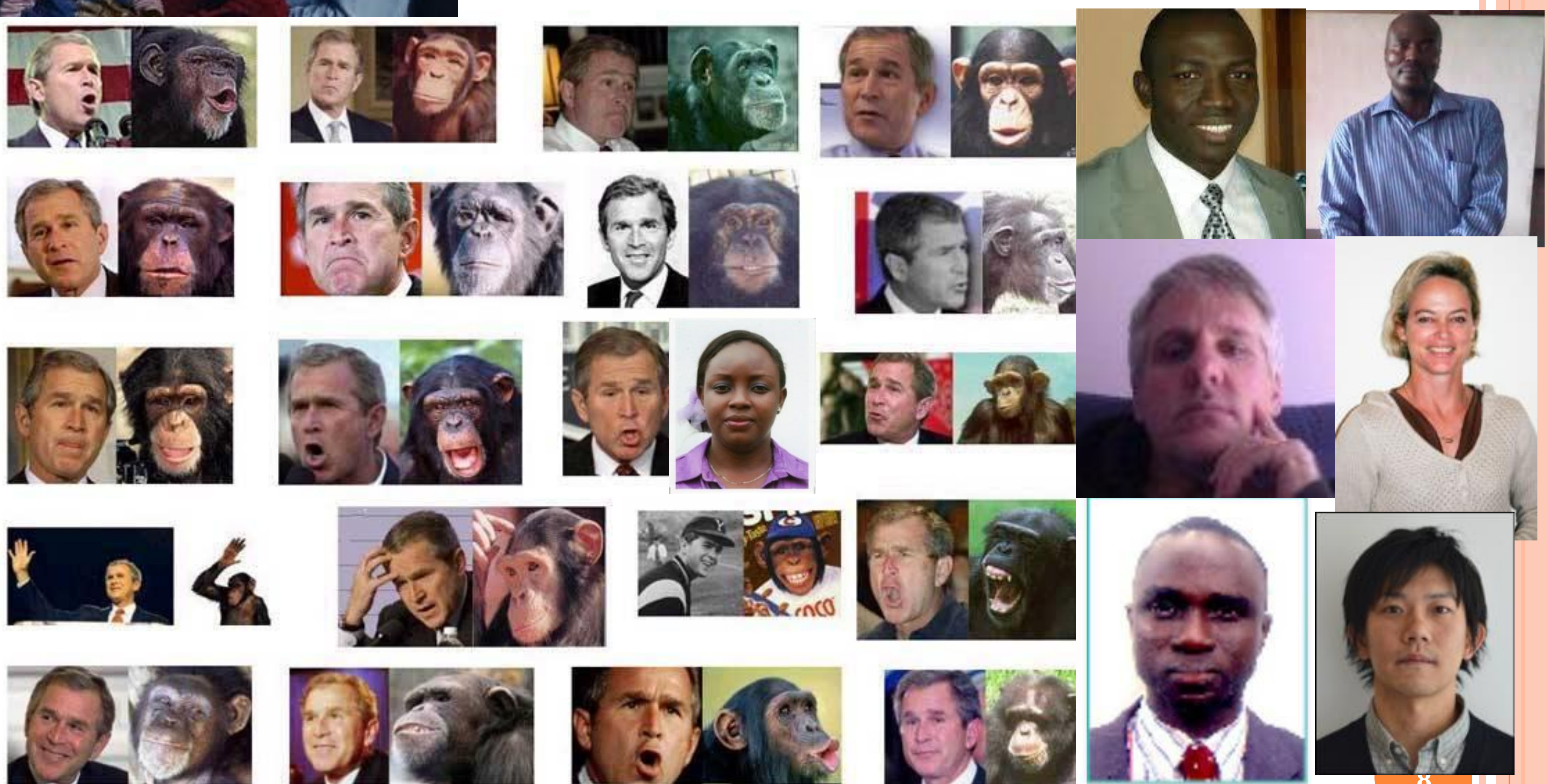


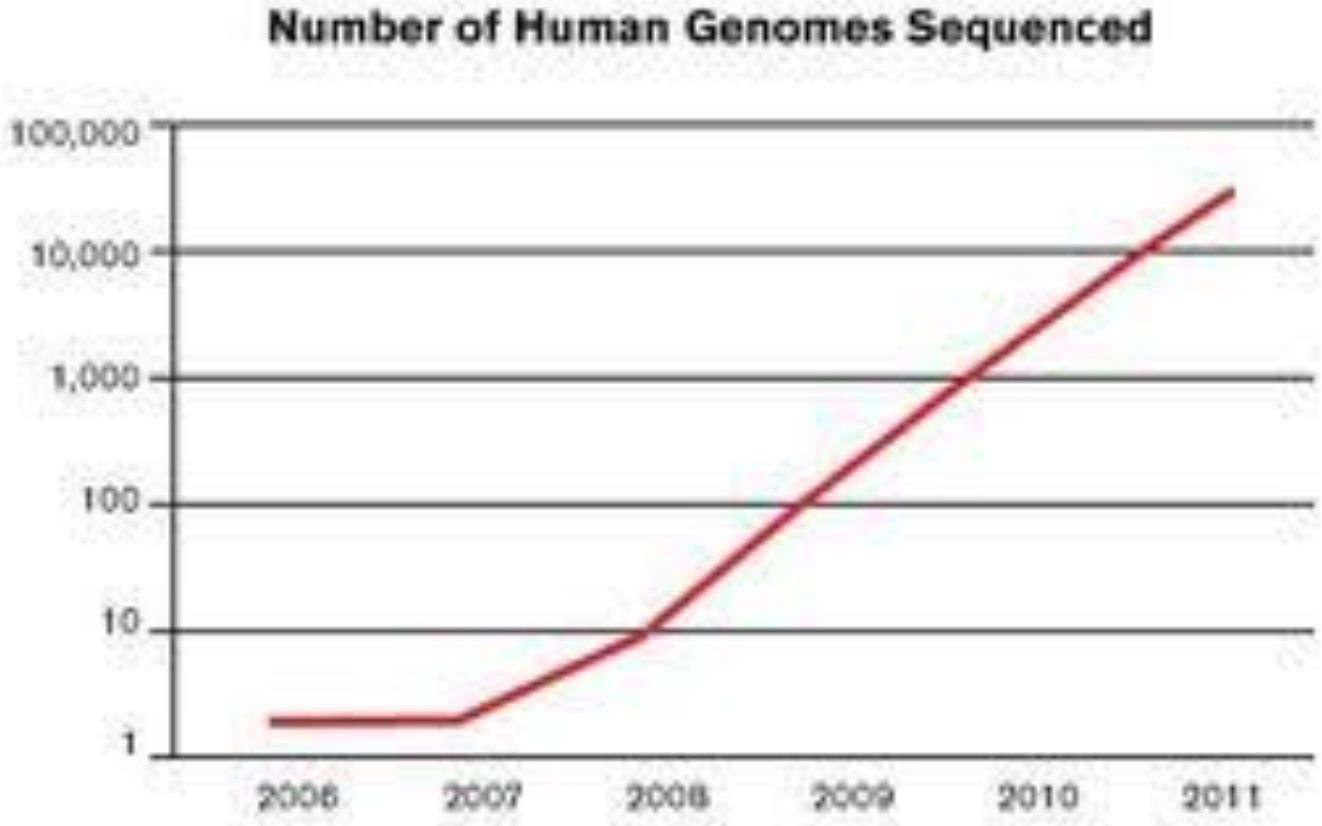
Milestones:

- 1990: Project initiated as joint effort of U.S. Department of Energy and the National Institutes of Health
- June 2000: Completion of a working draft of the entire human genome (covers >90% of the genome to a depth of 3-4x redundant sequence)
- February 2001: Analyses of the working draft are published
- April 2003: HGP sequencing is completed and Project is declared finished two years ahead of schedule

http://doegenomes.org
http://www.sanger.ac.uk/HGP/overview.shtml U.S. Department of Energy Genome Programs, Genomics and Its Impact on Science and Society, 2003

99.9% IDENTITY




Number of Human Genomes Sequenced
100,000
10,000
1,000
100
10
1
2006
2007
2008
2009
2010
2011

SOME BASIC: DNA VARIATION

- >99.9 % of the sequence is identical between any two chromosomes.
  - Compare maternal and paternal chromosome 1 in single person
  - Compare Y chromosomes between two unrelated males

- Even though most of the sequence is identical between two chromosomes, since the genome sequence is so long (~3 billion base pairs), there are still many variations.

- Some DNA variations are responsible for biological changes, others have no known function.

Types of genetic variations

CCTAGTTGACTGATCGCGGGATTCACACACATGG

CCTGGTTGAC..ATCGCGGGATTCACACACACACATGG

InDels
(insertions/deletions)
- two alleles
- > 1,000,000

SSR – short sequence repeats
(VNTR - variable number tandem repeats)
- **many alleles**
- **microsatellites (1-5)**
- **minisatellites (6-100)**
- **…**

> 1,000,000

Single (point) base changes
- two alleles

SNPs

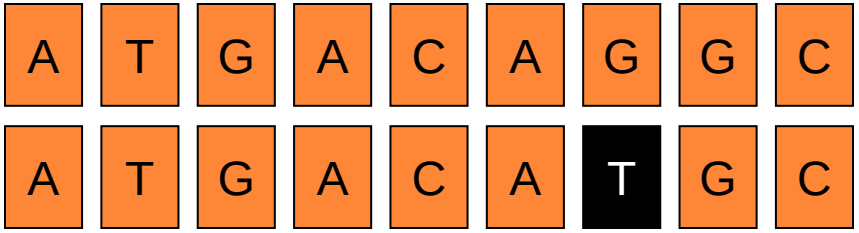
Single Nucleotide Polymorphisms; > 10, 000, 000

- **Inversions**
- **Duplications**
- **Translocations**
- **Transposon insertions**

Variations exceeding 1000bp - STRUCTURAL VARIATIONS
- **less than 3 million bp - submicroscopic; larger– microscopic**
- **InDels and duplications are called CNVs (copy number variations)**

Single Nucleotide Polymorphisms: SNPs

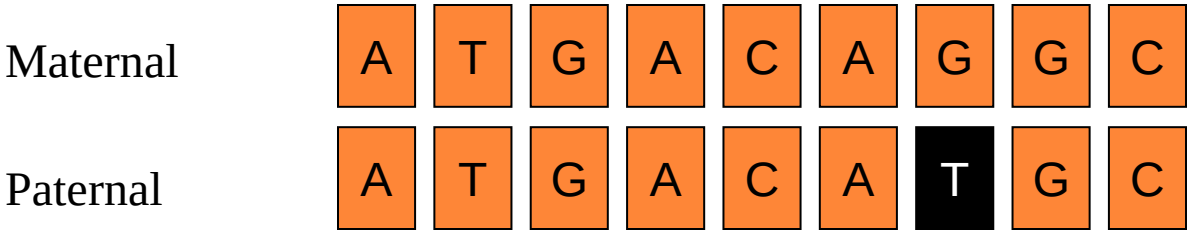
- SNPs – DNA sequence variations that occur when a single nucleotide is altered



- Alleles at this SNP are “G” and “T”
- SNPs are the most common form of variation in the human genome
- SNPs catalogued in several databases

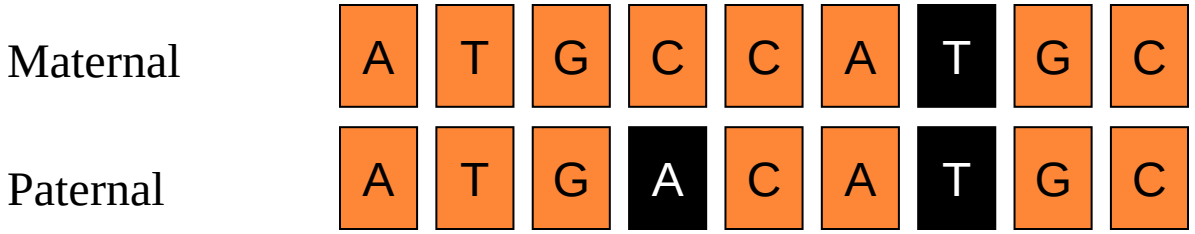
Genotypes and Haplotypes

- **Genotype**: pair of alleles (one paternal, one maternal) at a locus



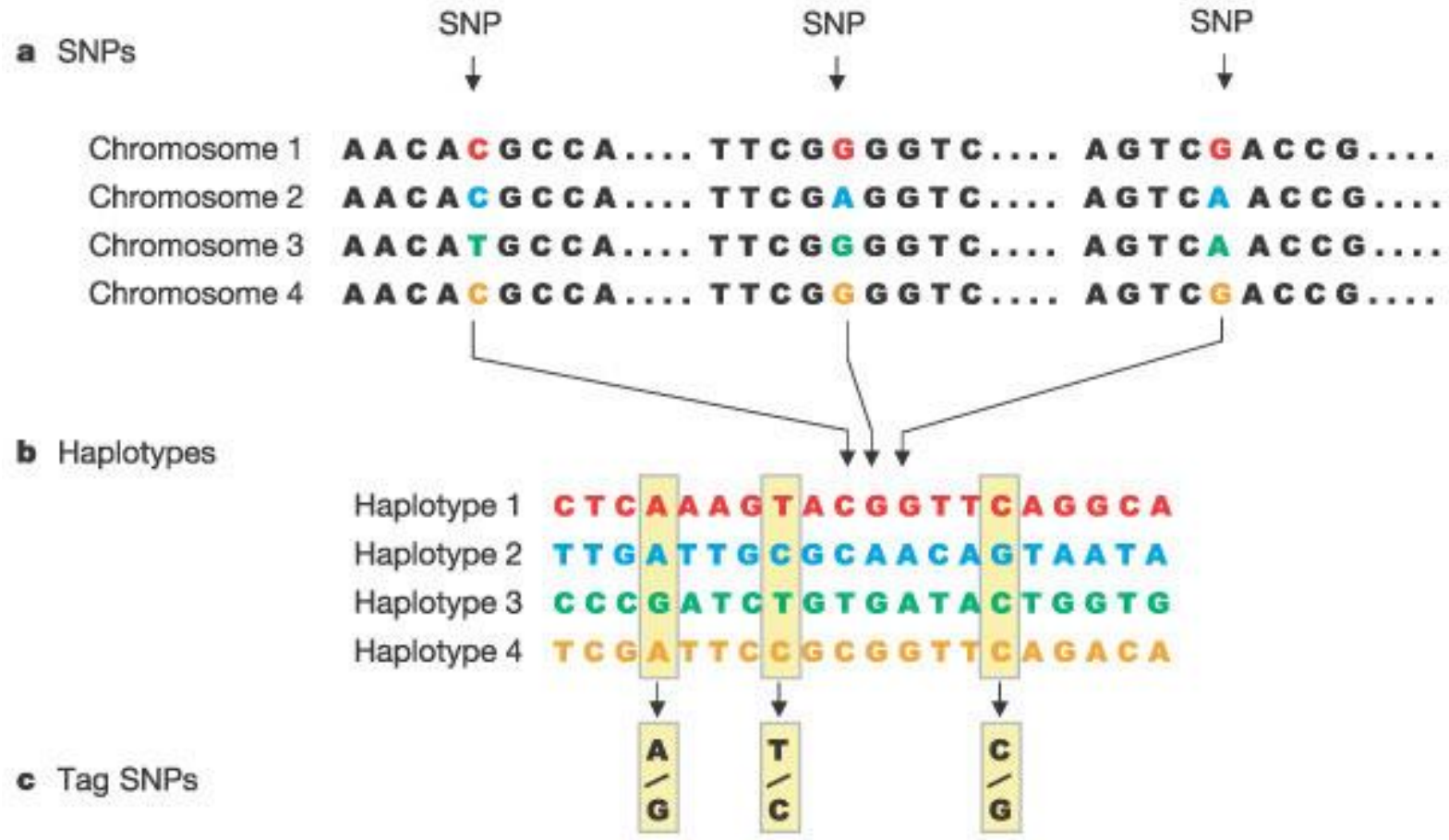
Genotype for this individual is GT

- **Haplotype**: sequence of alleles along a single chromosome



Genotypes for this individual (vertical) : CA and TT

Haplotypes (horizontal): CT and AT



The construction of the HapMap occurs in three steps. (a) Single nucleotide polymorphisms(SNPs) are identified in DNA samples from multiple indivduals. (b)Adjacent SNPs that are inherited together are compiled into "haplotypes." (c)"Tag" SNPs within haplotypes are identified that uniquely identify those haplotypes. By genotyping the three tag SNPs shown in this figure, researchers can identify which of the four haplotypes shown here are present in each individual

GWAS ?

- Genetic association studies test for a correlation between disease status and genetic variation to identify candidate genes or genome regions that contribute to a specific disease or trait.

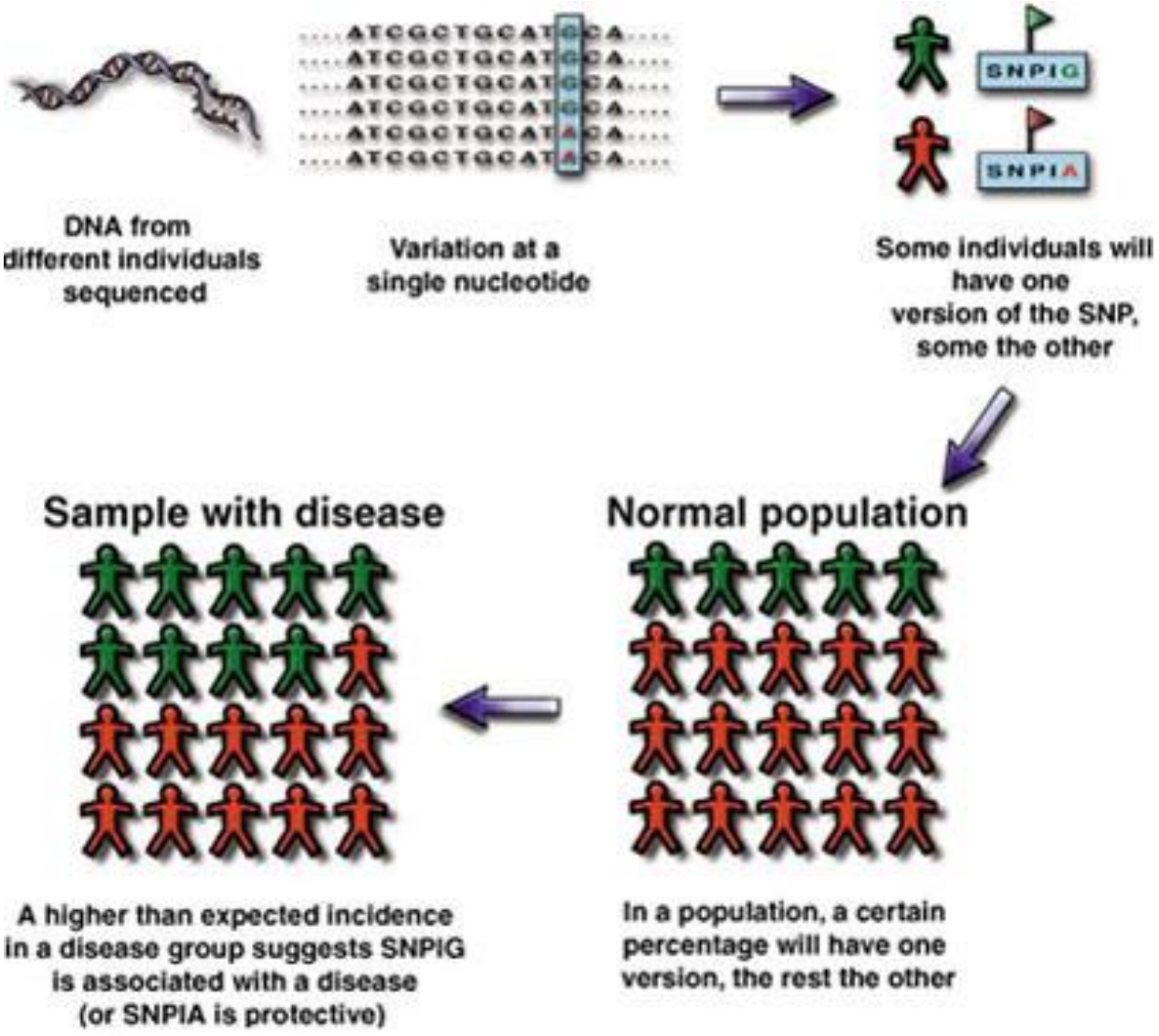
- **Main idea**: look for genetic differences of people with vs. without a disease

- First, need a method able to genotype thousands/millions of polymorphisms at once– microarrays

- Second, need to know where are the common polymorphisms– HapMap project

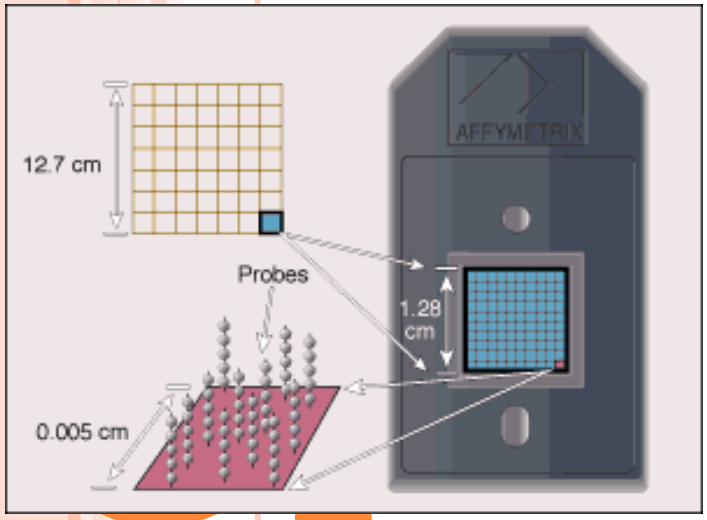
- Third, need HUGE cohorts of people to find subtle allele frequency differences

Overview -Using SNPs to Track Predisposition to Disease and other Genetic Traits

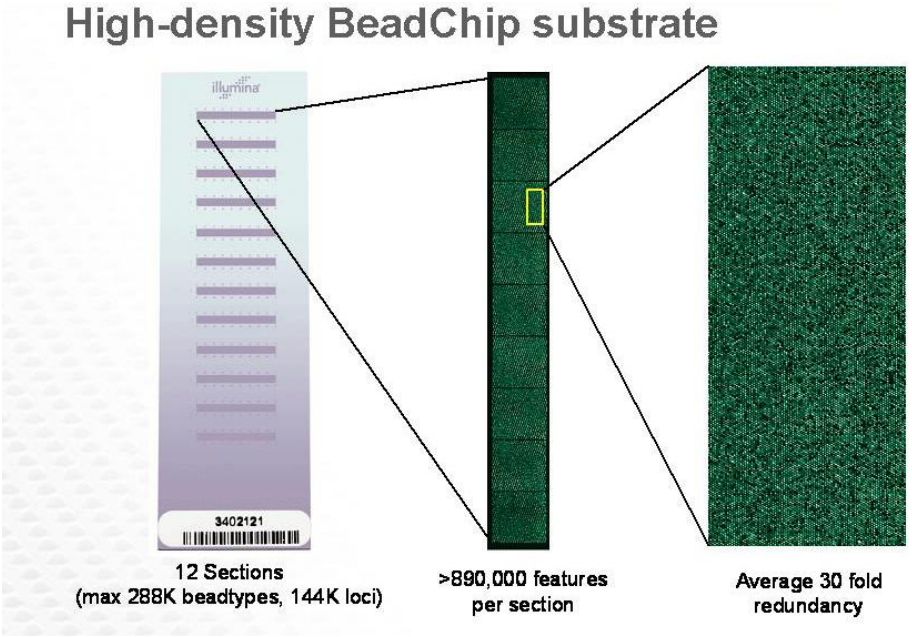


Genotyping Systems

Affymetrix



Illumina



5 Million Products are here and on the way!

A significant proportion of common SNPs can be captured

Genome-wide association studies (GWAS)

- End result of a successful GWAS:

```
rs398925
rs8029466
rs4966230
rs4966231
rs17137796
```

Genome-wide association studies (GWAS)

- End result of a successful GWAS:

```
rs398925
rs8029466
rs4966230
rs4966231
rs17137796
```

- What does this actually tell us?

Genome-wide association studies (GWAS)

- End result of a successful GWAS:

```
rs398925
rs8029466
rs4966230
rs4966231
rs17137796
```

- What does this actually tell us?
  - How to predict disease risk from genotype?
  - What polymorphisms cause disease?
  - What genes are involved?

HAPMAP

- The HapMap is a catalog of common genetic variants that occur in human beings.

- It describes what these variants are, where they occur in our DNA, and how they are distributed among people within populations and among populations in different parts of the world

Haplotype Map of the Human Genome

Goals:

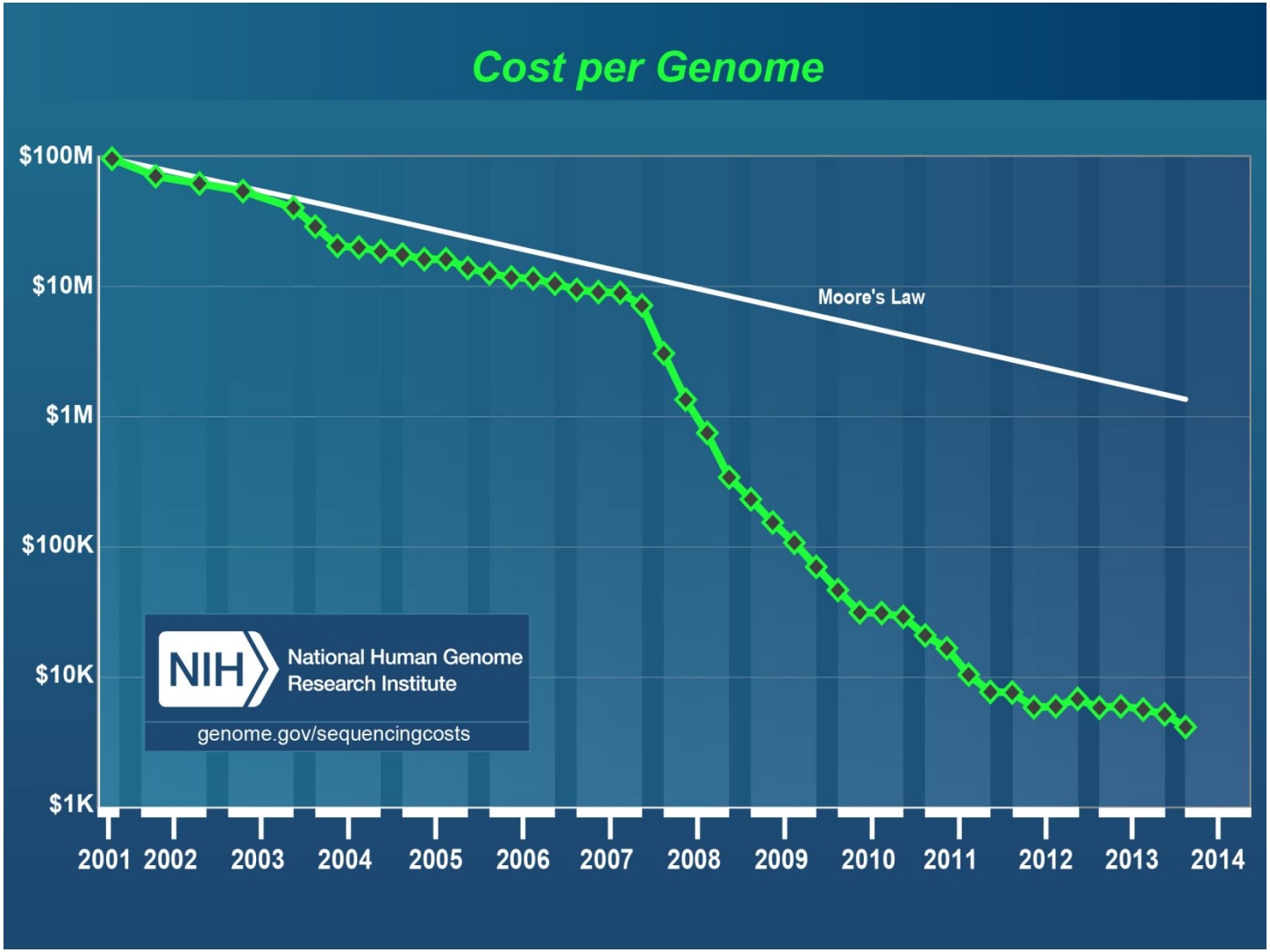
- Define patterns of genetic variation across human genome
- Guide selection of SNPs efficiently to "tag" common variants
- Public release of all data (assays, genotypes)

Phase I: 1.3 M markers in 269 people
Phase II: +2.8 M markers in 270 people

HapMap Samples

- 90 Yoruba individuals (30 parent-parent-offspring trios) from Ibadan, Nigeria (YRI)
- 90 individuals (30 trios) of European descent from Utah (CEU)
- 45 Han Chinese individuals from Beijing (CHB)
- 45 Japanese individuals from Tokyo (JPT)

Cost of Human Genome sequencing



Genome-Wide Association Approach to Common and Complex Diseases

- Identify all 10 million common SNPs
- Collect 1,000 cases and 1,000 controls
- Genotype all DNAs for all SNPs
- That adds up to 20 billion genotypes

- In 2002, this approach cost 50 cents a genotype.
- That's $10 billion for each disease – completely out of the question

Genome-Wide Association Approach to Common and Complex Diseases

- Identify an optimum set of 300,000 tag SNPs
- Collect 1,000 cases and 1,000 controls
- Genotype all DNAs for all SNPs
- That adds up to 600 million genotypes

- In 2008, genotyping dropped to $0.0010, amounting to $600,000 for each disease

- The Human Genome Project cost ~$3 billion
- Illumina now offers a complete genome sequence for $ 5,000
- HiSeq X Ten (by Illuminal), one that can deliver “full coverage human genomes for less than $1,000”.
- Illumina says that each HiSeq X Ten will therefore be capable of sequencing 18,000 human genomes per year.
- Each genome will be sequenced to the gold standard of 30x

Conclusion

- GWASs were made possible by the availability of chip-based microarray technology for assaying one million or more SNP

- HapMap genotype data allowed the examination of *linkage disequilibrium* and thus make GWAS possible