CS2220 Introduction to Computational Biology

Week 7: Single (Simple) Nucleotide Polymorphisms (SNPs)

Dr. Mengling FENG
Institute for Infocomm Research
Massachusetts Institute of Technology
mfeng@mit.edu

Plans for Week 7 and Week 8

- Week 7, 1st Oct 2015
  - 2 hours class: Single (Simple) Nucleotide Polymorphism
  - 1 hour briefing on project and forming of project teams
- Week 8, 7th Oct 2015
  - 2 hours class: Genome-wide Association Study (GWAS)
  - 1 hour Q&A on the lectures and project

Week 7's Learning Objectives

- After the class, students should be able to
  - Define the concept of SNP
  - Elaborate various types of SNPs and their functions
  - Explain the applications of SNPs
  - Know the major initiatives and projects related to SNP
  - Use online resources to find out information about SNPs

Single (simple) Nucleotide Polymorphism The Definition

- SNP is a DNA sequence variation occurring commonly within a population
  - A single nucleotide – A, T, C &G, mutation
  - Must be common
  - Minor Allele Frequency (MAF) >1%

Single (simple) Nucleotide Polymorphism

- ~15 million possible SNP sites in human genome, ~10 million common SNPs (MAF >5%)
- ~12 million SNPs have been identified (dbSNP 2012 release 137)
- Each individual may carry 3~5 million common SNPs (inherited) and ~120 new mutations
- SNPs VS Individual Mutations
  - Natural Selection
  - Founder Effect

SNPs as An Evidence for Nature Selection

- Many Africans carry SNPs around gene G6PD and CD40 ligand, which may lead to resistance to malaria

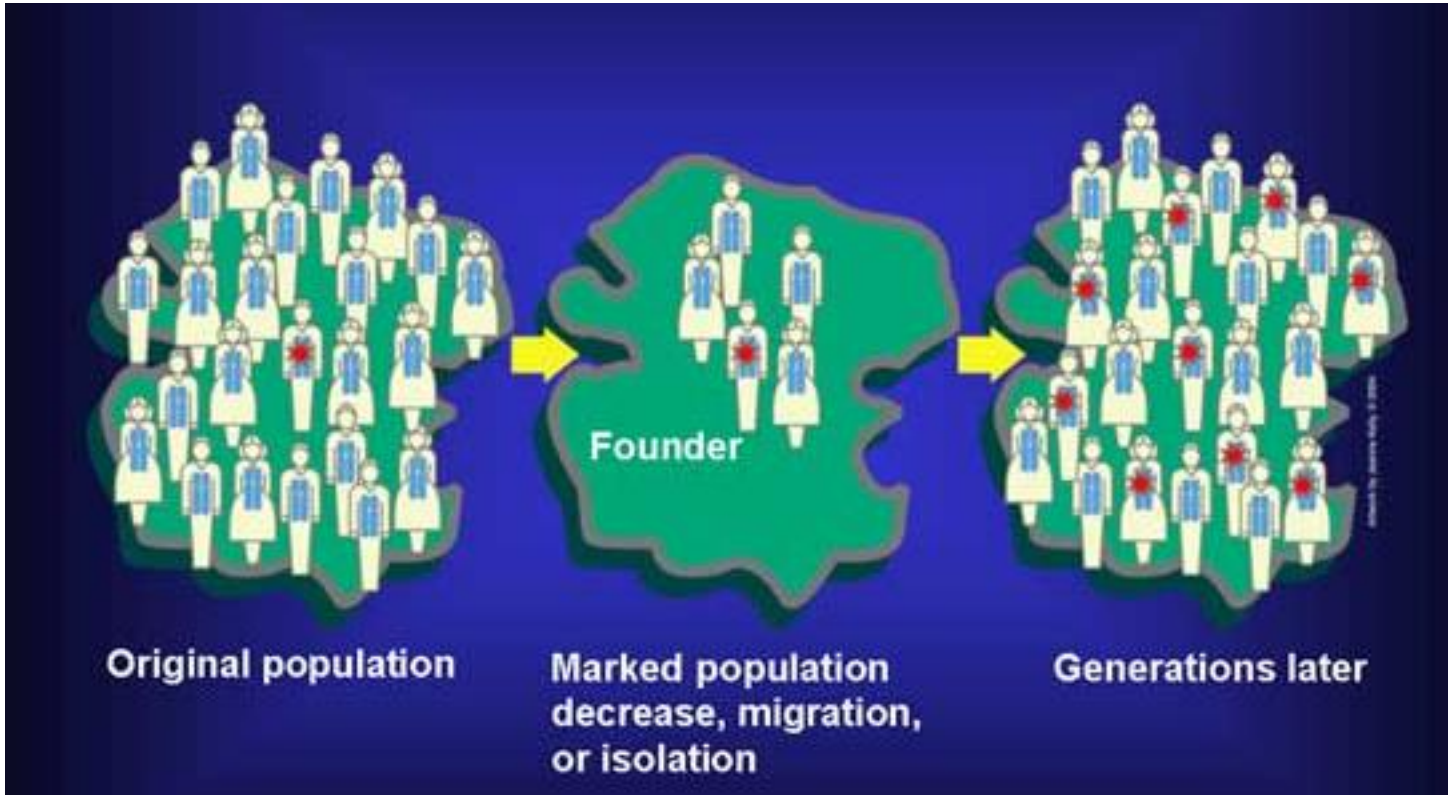
nature International weekly journal of science

Letters to Nature

Nature **419**, 832-837 (24 October 2002) | doi:10.1038/nature01140; Received 7 June 2002; Accepted 19 September 2002; Published online 9 October 2002

Detecting recent positive selection in the human genome from haplotype structure

Founder Effect



- Examples:
  - The Amish group
  - Ashkennazi Jews after the Holocaust

Types of SNPs

- Non-coding SNPs
  - 5’ Un-Translated Regions (UTR)
  - 3’ Un-Translated Regions (UTR)
  - Introns
  - Intergenic Regions (IGR)
  - Pseudogenes
- Coding SNPs
  - Synonymous substitution
  - Non-synonymous substitution
    - Missense
    - Nonsense

Functions of SNPs

- Take home message:
  - We still know very little about them
  - Genome-wide Association and other studies to identify associations and causations
- Majority of SNPs are believed to be silent
- Non-coding SNPs: regulatory functions
  - Splicing
  - Transcriptional regulation (promoter & TF binding sites)
  - Translational regulation (initiation or termination)
  - Regulate mRNA target sites

Functions of SNPs
Synonymous Substitutions

- Do not trigger amino acid change in protein sequence
- Were believed to be “silent” mutations
- Recent studies shown that they can affect
  - Messenger RNA (mRNA) splicing, stability, structure and protein folding => protein functions

	U	C	A	G	
U	UUU → Phe F	UCU → Ser S	UAU → Tyr Y	UGU → Cys C	U
	UUC → Phe F	UCC → Ser S	UAC → Tyr Y	UGC → Cys C	C
	UUA → Leu L	UCA → Ser S	UAA → Stop	UGA → Stop	A
	UUG → Leu L	UCG → Ser S	UAG → Stop	UGG → Trp W	G
C	CUU → Leu L	CCU → Pro P	CAU → His H	CGU → Arg R	U
	CUC → Leu L	CCC → Pro P	CAC → His H	CGC → Arg R	C
	CUA → Leu L	CCA → Pro P	CAA → Gln Q	CGA → Arg R	A
	CUG → Leu L	CCG → Pro P	CAG → Gln Q	CGG → Arg R	G
A	AUU → Ile I	ACU → Thr T	AAU → Asn N	AGU → Ser S	U
	AUC → Ile I	ACC → Thr T	AAC → Asn N	AGC → Ser S	C
	AUA → Ile I	ACA → Thr T	AAA → Lys K	AGA → Arg R	A
	AUG → Met M	ACG → Thr T	AAG → Lys K	AGG → Arg R	G
G	GUU → Val V	GCU → Ala A	GAU → Asp D	GGU → Gly G	U
	GUC → Val V	GCC → Ala A	GAC → Asp D	GGC → Gly G	C
	GUA → Val V	GCA → Ala A	GAA → Glu E	GGA → Gly G	A
	GUG → Val V	GCG → Ala A	GAG → Glu E	GGG → Gly G	G

Functions of SNPs
Non-Synonymous Substitutions

- Missense: change in amino acid of protein sequence

```
    DNA: 5' - AAC AGC CTG CGT ACG GCT CTC - 3'
         3' - TTG TCG GAC GCA TGC CGA GAG - 5'
   mRNA: 5' - AAC AGC CUG CGU GCG ACG CUC - 3'
Protein:      Asn Ser Leu Arg Thr Ala Leu
```

```
    DNA: 5' - AAC AGC CTG CTT ACG GCT CTC - 3'
         3' - TTG TCG GAC GAA TGC CGA GAG - 5'
   mRNA: 5' - AAC AGC CUG CUU GCG ACG CUC - 3'
Protein:      Asn Ser Leu Leu Thr Ala Leu
```

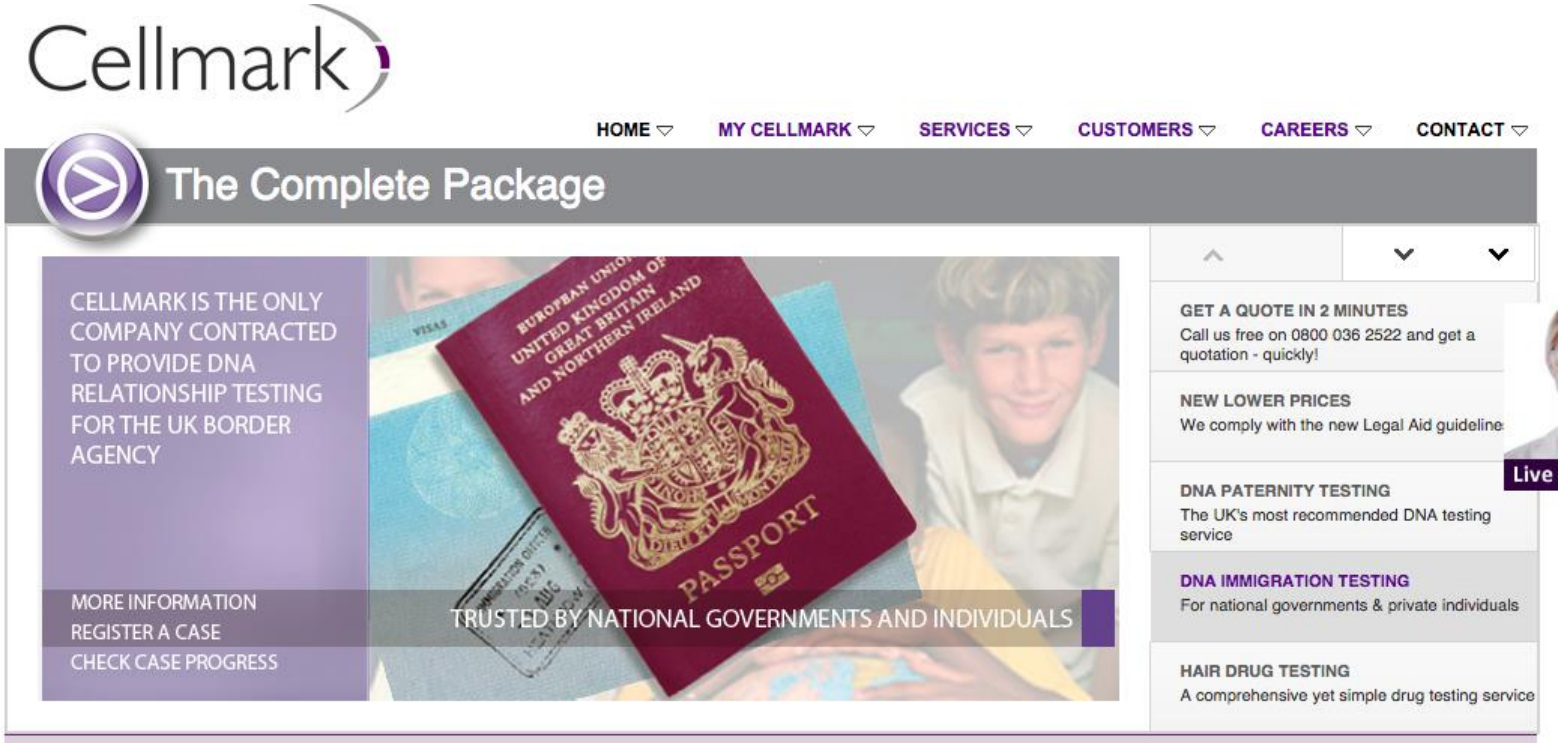
- Nonsense: change in amino acid that lead to premature stop codon

```
    DNA: 5' - ATG ACT CAC CGA GCG CGA AGC TGA - 3'
         3' - TAC TGA GTG GCT CGC GCT TCG ACT - 5'
   mRNA: 5' - AUG ACU CAC CGA GCG CGA AGC UGA - 3'
Protein:      Met Thr His Arg Ala Arg Ser Stop
```

```
    DNA: 5' - ATG ACT CAC TGA GCG CGA AGC TGA - 3'
         3' - TAC TGA GTG ACT CGC GCT TCG ACT - 5'
   mRNA: 5' - AUG ACU CAC UGA GCG CGU AGC UGA - 3'
Protein:      Met Thr His Stop
```

Applications of SNPs

- General Applications
  - Forensics
  - Paternity tests
  - Ancestry trace: immigration to the United Kingdom
  - Follow ethnic migrations



Applications of SNPs

- Genetic marker for distinguishing traits
  - Predisposition for disease

Disease Risks (100)

Elevated Risks	Your Risk	Average Risk
Gallstones new	11.1%	7.0%
Restless Legs Syndrome	2.5%	2.0%

more »

Decreased Risks	Your Risk	Average Risk
Prostate Cancer ♂	12.7%	17.8%
Alzheimer's Disease new	4.9%	7.2%
Colorectal Cancer	4.2%	5.6%

more »

See all 100 risk reports...

Carrier Status (24)

Hemochromatosis	Variant Present
Alpha-1 Antitrypsin Deficiency	Variant Absent
Bloom's Syndrome	Variant Absent
BRCA Cancer Mutations (Selected)	Variant Absent
Canavan Disease	Variant Absent
Cystic Fibrosis	Variant Absent
Familial Dysautonomia	Variant Absent
Factor XI Deficiency	Variant Absent

See all 24 carrier status...

Applications of SNPs

- Genetic marker for distinguishing traits
  - Predisposition for disease
  - Drug efficacy
  - Drug adverse effect

Drug Response (19)

Warfarin (Coumadin®) Sensitivity	Increased
Abacavir Hypersensitivity	Typical
Alcohol Consumption, Smoking and Risk of Esophageal Cancer	Typical
Clopidogrel (Plavix®) Efficacy	Typical
Fluorouracil Toxicity	Typical

See all 19 drug response...

Applications of SNPs

- Genetic marker for distinguishing traits
  - Predisposition for disease
  - Drug efficacy
  - Drug adverse effect
  - Other traits

Applications of SNPs

- Genetic marker for distinguishing traits
  - Predisposition for disease
  - Drug efficacy
  - Drug adverse effect
  - Other traits
- Preventive medicine
- Personalized and targeted medicine
- Profession selection
- etc

Population Genetics of SNPs for Forensic An Individual Identification Panel

NIJ Final Report

September 1, 2007 to February 28, 2011

Population Genetics of SNPs for Forensic Purposes

NIJ Grant# 2007-DN-BX-K197, including supplement

Kenneth K. Kidd (PI), Yale University School of Medicine

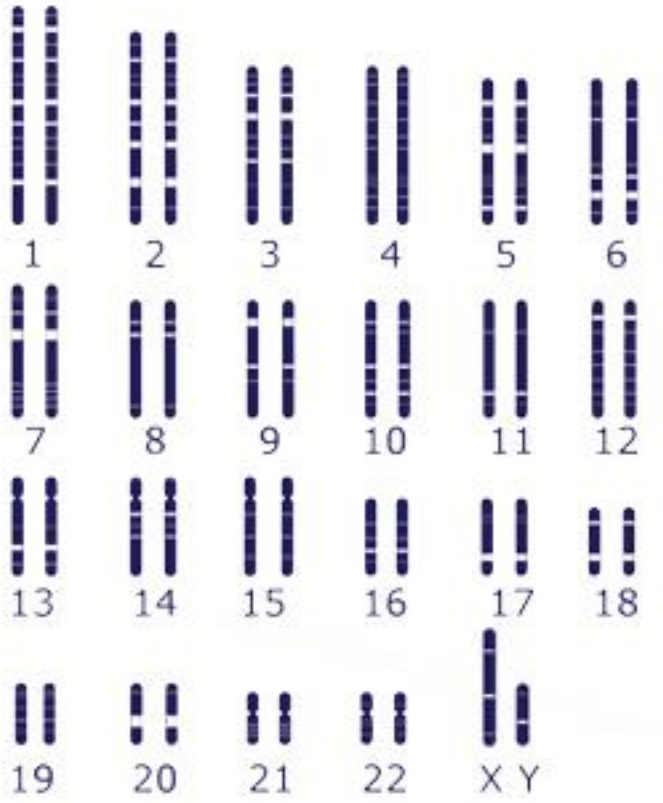
Portions of this report are taken from ten research publications. two submitted manuscripts, and a number of poster presentations--all supported by this grant or the preceding grant (NIJ 2004-DN-BX-K025).

SNP Panel Selection

- SNPs data from 44 populations
- Selection criteria
  - A small panel is preferred
    - Incomplete or damaged DNA samples
    - Reduce cost
  - For individual SNP
    - Average Heterozygosity > 0.4
    - Average Fixation Index Fst < 0.06
    - Linkage Disequilibrium ~ 0.01

Heterozygosity

- Human beings are diploid organism
  - We carry two copies of a gene
  - For a gene having two alleles: A & a
    - Homozygote: AA and aa
    - Heterozygote: Aa
- Heterozygosity
  - Percentage of heterozygot in the population
- SNP selection criterion:
  - Average heterozygosity > 0.4
  - High genetic variations among individuals are preferred



Estimation of Heterozygosity
The Hardy-Weinberg Theorem

frequency of heterozygous genotype

$$p^2 + 2pq + q^2 = 1$$

frequency of homozygous dominant genotype

frequency of homozygous recessive genotype

		Females	
		A (*p*)	**a (*q*)**
Males	**A (*p*)**	AA (p^2)	Aa (pq)
	a (*q*)	Aa (qp)	aa (q^2)

$$H_E = 2pq = 1 - p^2 - q^2 = 1 - \sum_{i=1}^{2} p_i^2$$

$$H_E = 1 - \sum_{i=1}^{k} p_i^2$$

Fixation Index Fst

- A measure of differentiation of subpopulations

$$F_{st} = \frac{\sigma_s^2}{\bar{p}(1-\bar{p})}$$

σ_s^2 is the variance of allele frequencies among different subpopulations

$\bar{p}$ is the average allele frequency across the population

- Selection Criterion:
  - Fst < 0.06
  - Similar genetic profiles among subpopulations are preferred

Linkage Disequilibrium (LD)

- Measures the non-random association of alleles at different loci
- In the study, r2 measure was used
- Selection criterion:
  - LD ~ 0.01
  - Avoid picking up highly linked SNPs
  - Minimize redundancy

An Individual Identification SNP Panel The Results

- Identified two sets of SNPs
- Set I: 45 SNPs
  - Estimated average matching probability $< 10^{-15}$
  - An two random individuals to have the same genotype will be very unlikely
- Set II: 89 SNPs
  - Estimated average matching probability $< 10^{-33}$

SNP As A Disease Bio-maker
Cystic Fibrosis

- A genetic disorder that affects mostly the lungs
- Inherited in an autosomal recessive manner
- Most common among people of Northern European ancestry

Carrier Frequency for Mutant *CFTR* Alleles

Population Group	Approximate Carrier Frequency	Reference
Ashkenazi Jewish	1:29	Kerem et al [1997]
North American of northern European heritage	1:28	Hamosh et al [1998]
African American	1:61	Hamosh et al [1998]

http://www.ncbi.nlm.nih.gov/books/NBK1250/

SNP As A Disease Bio-maker
Gaucher Disease

- A genetic disease in which fatty substances accumulate in cells and certain organs
- Inherited in an autosomal recessive manner

Prevalence

A study from Australia reported a disease frequency of 1:57,000 [Meikle et al 1999]; a similar study from the Netherlands reported 1.16:100,000 [Poorthuis et al 1999].

A founder effect for specific alleles underlies the observed occurrence of GD in specific populations:

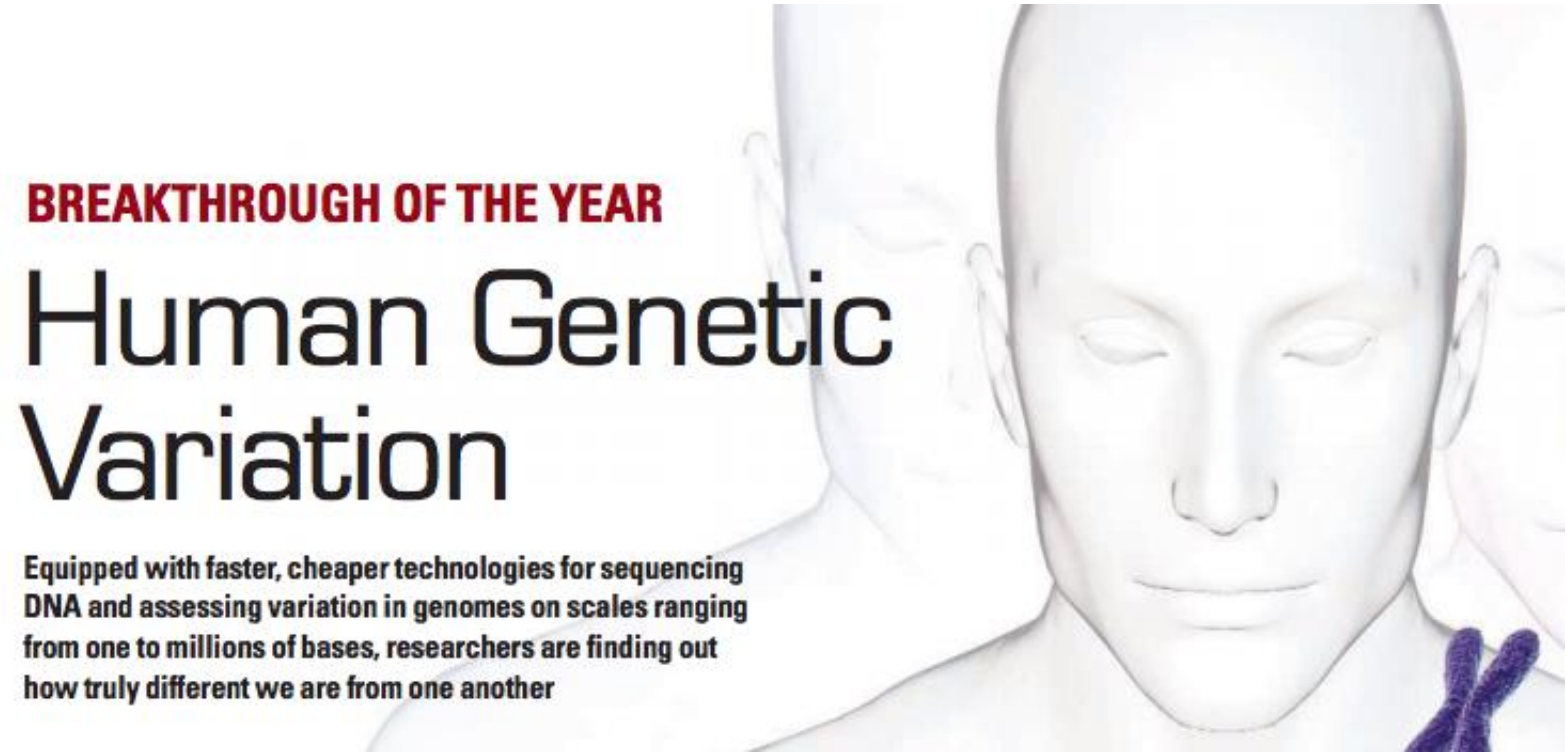
- Ashkenazi Jewish, Spanish, and Portuguese (N370S)
- Swedish (L444P)
- Jenin Arab, Greek, and Albanian (D409H). Among Greeks and Albanians, D409H has been found in *cis* with H255Q.

Non-neuropathic GD (type 1) is prevalent in the Ashkenazi Jewish population, with a disease prevalence of 1:855 and an estimated carrier frequency of 1:18.

The prevalence of neuropathic GD (types 2 and 3) varies across ethnic groups but appears to be higher among those who are not of European origin.

http://www.ncbi.nlm.nih.gov/books/NBK1269/

Breakthrough of The Year 2007
Human Genome Variation

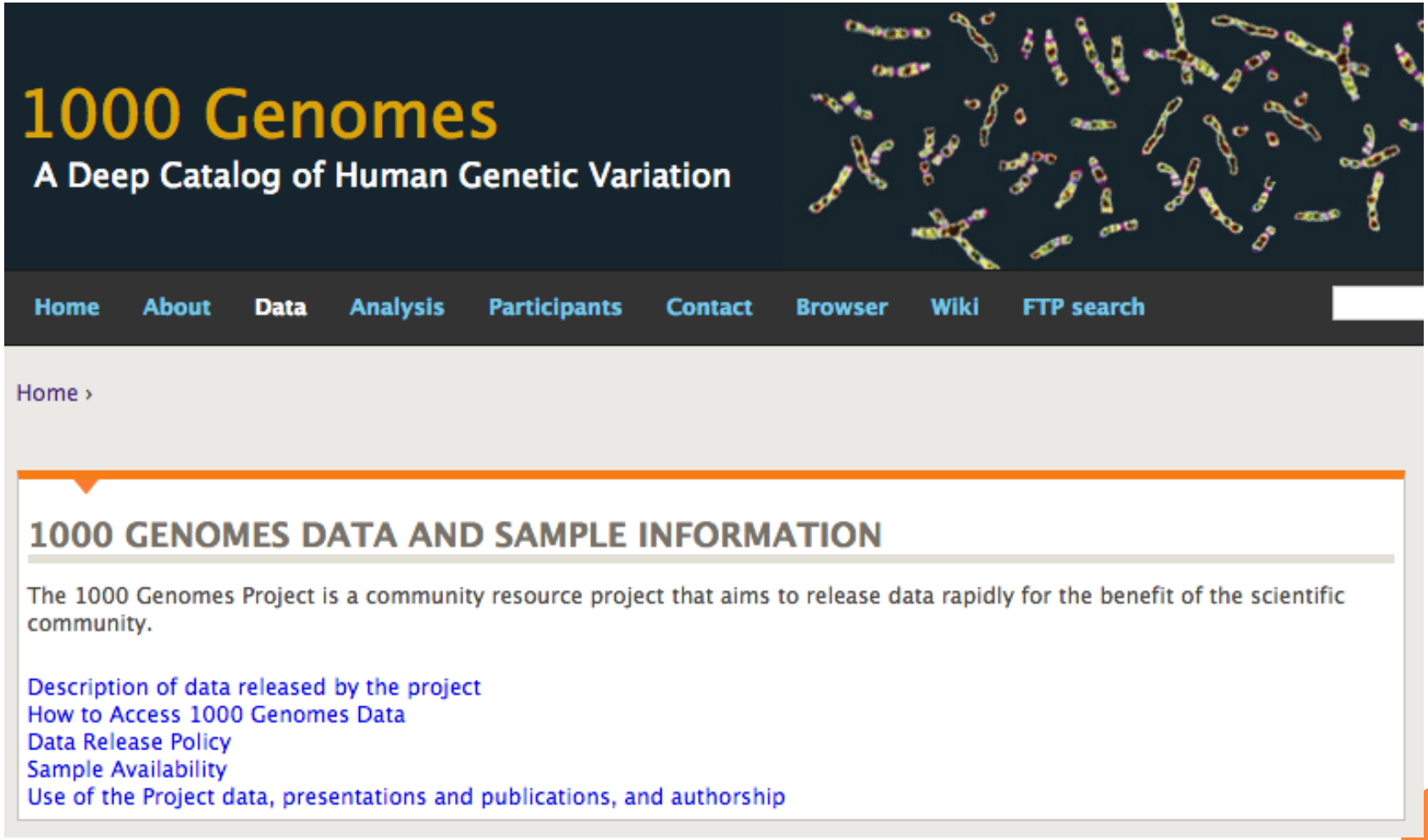


The HapMap Project

http://hapmap.ncbi.nlm.nih.gov/index.html.en

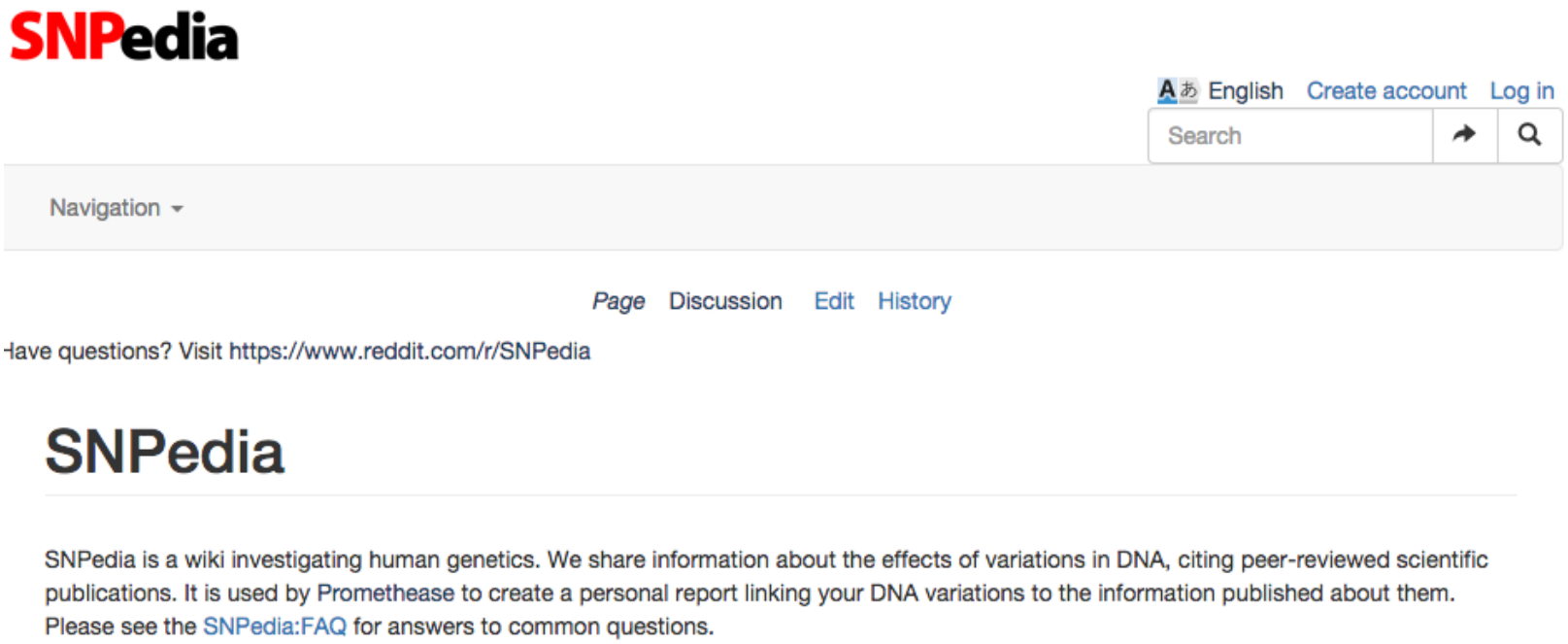
1000 Genome Project

http://www.1000genomes.org/data



Online Resouces: SNPedia

http://snpedia.com/index.php/SNPedia



Online Resouces: dbSNP

http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=487989

Week 7's Learning Obejectives

- After the class, students should be able to
  - Define the concept of SNP
  - Elaborate various types of SNPs and their functions
  - Explain the applications of SNPs
  - Know the major initiatives and projects related to SNP
  - Use online resources to find out information about SNPs
  - Understand the concept of haplotype and linkage disequilibrium