

CS2220 Introduction to Computational Biology

WEEK 7: SINGLE (SIMPLE) NUCLEOTIDE POLYMORPHISMS (SNPs)

1

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 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

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

Access

To read this story in full you will need to login or make a payment (see right).

[nature.com](#) > [Journal home](#) > [Table of Contents](#)

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

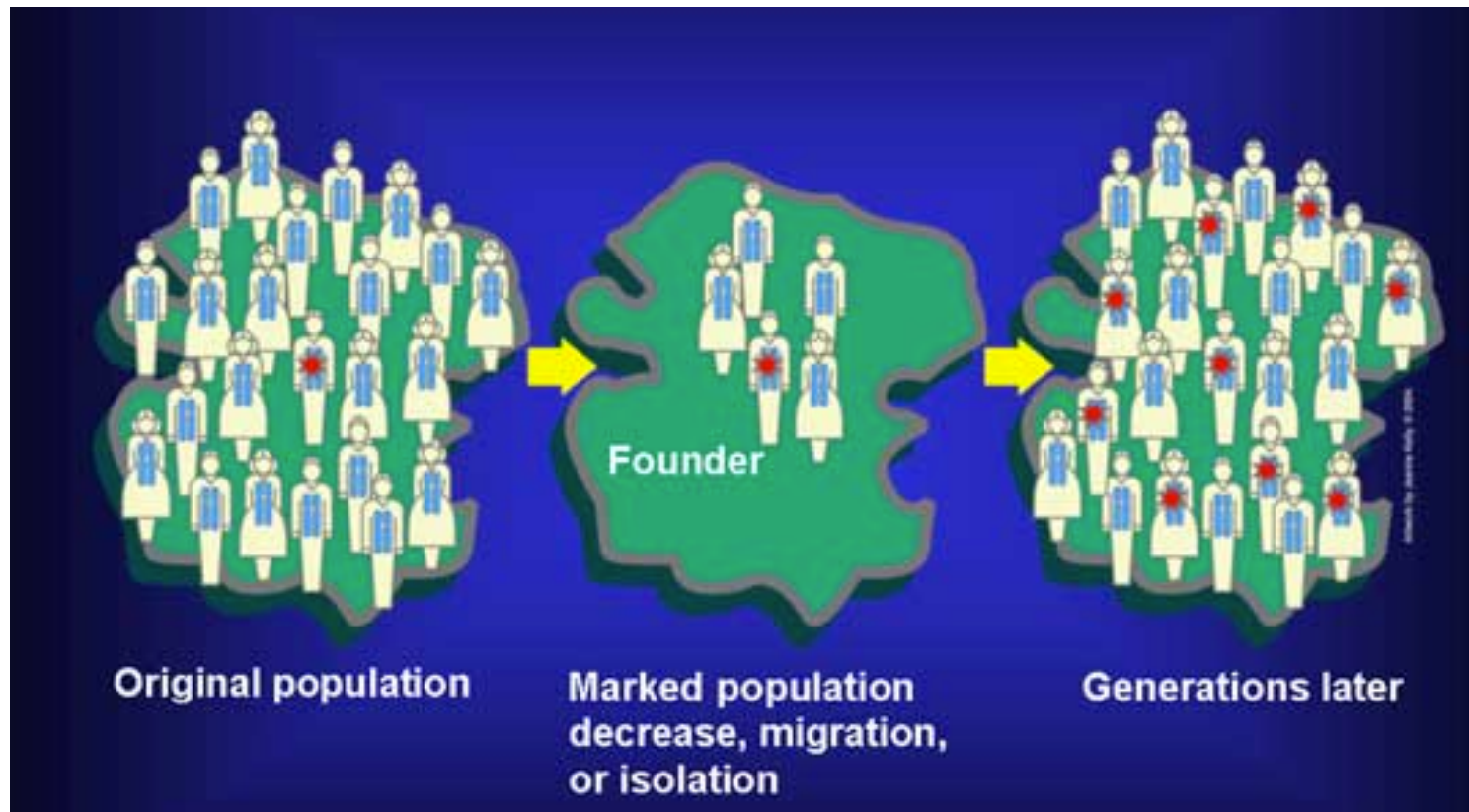
Detecting recent positive selection in the human genome from haplotype structure

Pardis C. Sabeti^{1,2,7}, David E. Reich¹, John M. Higgins¹, Hanin Z. P.

ARTICLE LINKS

- Figures and tables
- Supplementary info

FOUNDER EFFECT



© National Cancer Institute

○ 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
 - Integenic Regions (IGR)
 - Psuedogenes
- Coding SNPs
 - Synonymous substitution
 - Non-synonymous substitution
 - Missense
 - Nonsense

FUNCTIONS OF SNPs

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



A screenshot of the Cellmark website homepage. The header includes a navigation menu with links: HOME, MY CELLMARK, SERVICES, CUSTOMERS, CAREERS, and CONTACT. Below the header is a banner titled "The Complete Package" with a play button icon. The main content area features a large image of a UK passport and a child's face. Text on the left states: "CELLMARK IS THE ONLY COMPANY CONTRACTED TO PROVIDE DNA RELATIONSHIP TESTING FOR THE UK BORDER AGENCY". Below this are links for "MORE INFORMATION", "REGISTER A CASE", and "CHECK CASE PROGRESS". A central text overlay reads "TRUSTED BY NATIONAL GOVERNMENTS AND INDIVIDUALS". On the right, a sidebar lists services: "GET A QUOTE IN 2 MINUTES", "NEW LOWER PRICES", "DNA PATERNITY TESTING", "DNA IMMIGRATION TESTING", and "HAIR DRUG TESTING". A "Live" chat bubble is visible on the right side of the sidebar.

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

Traits (50) ?

Alcohol Flush Reaction

Does Not Flush

Bitter Taste Perception

Can Taste

Earwax Type

Wet

Eye Color

Likely Brown

Hair Curl 

Slightly Curlier Hair on Average

[See all 50 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

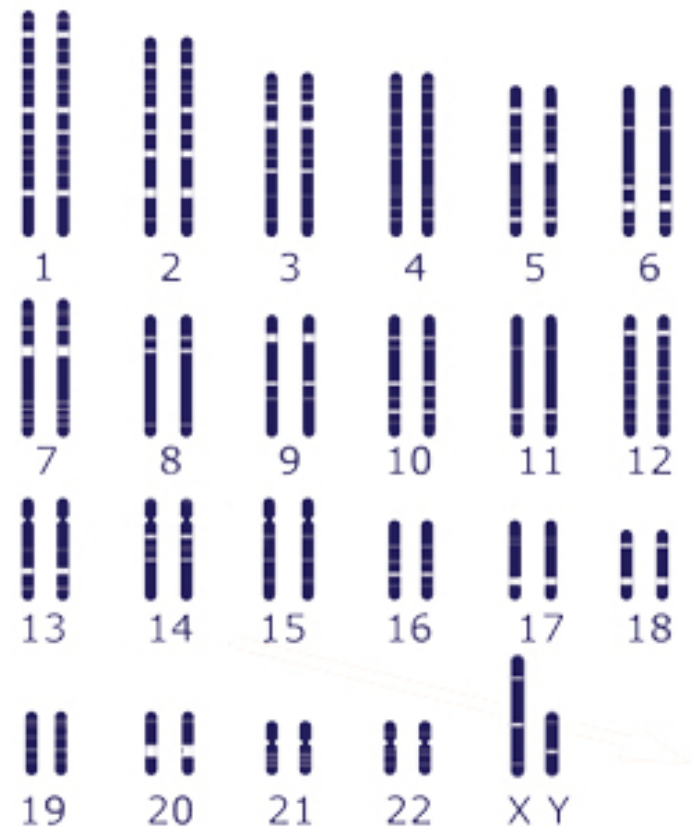
Portions of this report are taken from ten research publications. two submitted manuscripts, and a number of poster presentations--all supported by this grant or the preceding grant (NIJ 2004-DN-BX-K025).

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 $F_{st} < 0.06$
 - 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 heterozygote in the population
- SNP selection criterion:
 - Average heterozygosity > 0.4
 - High genetic variations among individuals are preferred



ESTIMATION OF HETEROZYGOSITY

THE HARDY-WEIBERG THEOREM

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

frequency of heterozygous genotype

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 F_{ST}

- 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:
 - $F_{st} < 0.06$
 - Similar genetic profiles among subpopulations are preferred

LINKAGE DISEQUILIBRIUM (LD)

- Measures the non-random association of alleles at different loci
- In the study, r^2 measure was used
- Selection criterion:
 - $LD \sim 0.01$
 - 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
<u>Ashkenazi Jewish</u>	1:29	<u>Kerem et al [1997]</u>
North American of northern European heritage	1:28	<u>Hamosh et al [1998]</u>
African American	1:61	<u>Hamosh et al [1998]</u>

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.

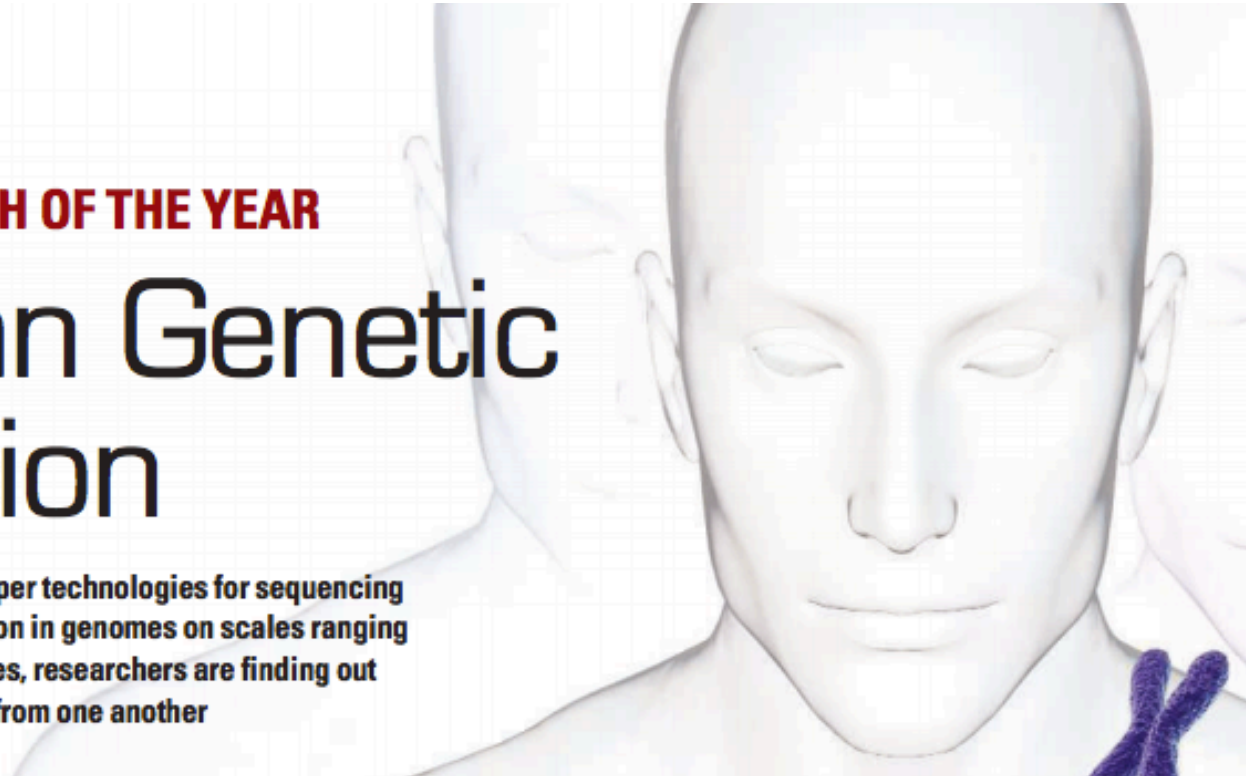
BREAKTHROUGH OF THE YEAR 2007

HUMAN GENOME VARIATION

BREAKTHROUGH OF THE YEAR

Human Genetic Variation

Equipped with faster, cheaper technologies for sequencing DNA and assessing variation in genomes on scales ranging from one to millions of bases, researchers are finding out how truly different we are from one another



THE HAPMAP PROJECT

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



中文 | [English](#) | Français | 日本語 | Yoruba

The International HapMap Project is a partnership of scientists and funding agencies from Canada, China, Japan, Nigeria, the United Kingdom and the United States. The project is a resource that will help researchers find genes associated with human disease and response to pharmaceuticals. See "[About the International HapMap Project](#)"

Project Information

[About the Project](#)
[HapMap Publications](#)
[HapMap Tutorial](#)
[HapMap Mailing List](#)
[HapMap Project Participants](#)

Project Data

[HapMap Genome Browser release #28 \(Phases 1, 2 & 3 - merged genotypes & frequencies\)](#)
[HapMap3 Genome Browser release #3 \(Phase 3 - genotypes & frequencies\)](#)

News

- 2013-06-14: **HapMap data conversion tool**

There are several inquiries for a conversion tool to convert HapMap data into the VCF format. Please see [Analysis Toolkit](#) (by Broad Institute).

- 2012-12-06: **Downtime for hardware maintenance**

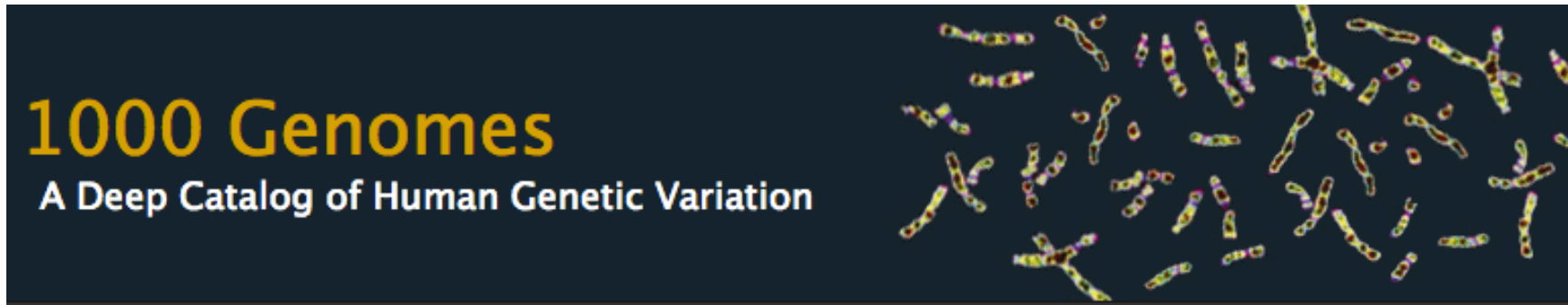
From December 15 - 16, Hapmap site will be taken offline for an internal hardware maintenance. See [HapMap Downtime](#).

- 2011-06-13: **HapMap help desk announcement**

There was a problem with the HapMap help desk system. In the past several weeks, emails sent to the help desk were not reaching the help desk, and thus user requests were not addressed. Please resend your email request to the HapMap help desk in the past several weeks. Sorry for the inconvenience.

1000 GENOME PROJECT

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



1000 Genomes
A Deep Catalog of Human Genetic Variation

[Home](#) [About](#) [Data](#) [Analysis](#) [Participants](#) [Contact](#) [Browser](#) [Wiki](#) [FTP search](#)

[Home](#) >

1000 GENOMES DATA AND SAMPLE INFORMATION

The 1000 Genomes Project is a community resource project that aims to release data rapidly for the benefit of the scientific community.

- [Description of data released by the project](#)
- [How to Access 1000 Genomes Data](#)
- [Data Release Policy](#)
- [Sample Availability](#)
- [Use of the Project data, presentations and publications, and authorship](#)

ONLINE RESOURCES: SNPEDIA

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

SNPedia

 English [Create account](#) [Log in](#)

Search



Navigation ▾

[Page](#) [Discussion](#) [Edit](#) [History](#)

Have questions? Visit <https://www.reddit.com/r/SNPedia>

SNPedia

SNPedia is a wiki investigating human genetics. We share information about the effects of variations in DNA, citing peer-reviewed scientific publications. It is used by [Promethease](#) to create a personal report linking your DNA variations to the information published about them. Please see the [SNPedia:FAQ](#) for answers to common questions.

ONLINE RESOURCES: DBSNP

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

**dbSNP**
Short Genetic Variations

dbVarClinVarGaPPubMedNucleotideProtein

Search small variations in dbSNP or large structural variations in dbVar

Search Entrez dbSNP for Go

Have a question about dbSNP? Try searching the SNP FAQ Archive!
Go

We need your feedback. Please help us evaluate changes and improvements to NCBI variation resources by taking this short two minute survey. Thank you for participating.
Click Here

GENERAL
RSS Feed

ANNOUNCEMENT

Interested in structural variations?
Visit NCBI [dbVar](#)



dbVar
Database of genomic structural variation

05/21/2014 RELEASE NCBI dbSNP Human Build 141

Search by IDs on All Assemblies

Note: rs# and ss# must be prefixed with "rs" or "ss", respectively (i.e. rs25, ss25)

ID: Reference cluster ID(rs#) 

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