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
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%

![SNP Diagram](image)
~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.
Examples:
- The Amish group
- Ashkenazi 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
FUNCTIONS OF SNPs
NON-SYNONYMOUS SUBSTITUTIONS

- Missense: change in amino acid of protein sequence
  
  ![](image1)

- Nonsense: change in amino acid that lead to premature stop codon
  
  ![](image2)
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
APPLICATIONS OF SNPs

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

Drug Response (19)

<table>
<thead>
<tr>
<th>Warfarin (Coumadin®) Sensitivity</th>
<th>Increased</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abacavir Hypersensitivity</td>
<td>Typical</td>
</tr>
<tr>
<td>Alcohol Consumption, Smoking and Risk of Esophageal Cancer</td>
<td>Typical</td>
</tr>
<tr>
<td>Clopidogrel (Plavix®) Efficacy</td>
<td>Typical</td>
</tr>
<tr>
<td>Fluorouracil Toxicity</td>
<td>Typical</td>
</tr>
</tbody>
</table>

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
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 organisms
  - We carry two copies of a gene
  - For a gene having two alleles: A & a
    - Homozygote: AA and aa
    - Heterozygote: Aa

- Heterozygosity
  - Percentage of heterozygotes in the population

- SNP selection criterion:
  - Average heterozygosity > 0.4
  - High genetic variations among individuals are preferred
ESTIMATION OF HETEROZYGOSITY

THE HARDY-WEINBERG THEOREM

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

\[ H_E = 1 - \sum_{i=1}^{k} p_i^2 \]
**Fixation Index Fst**

- A measure of differentiation of subpopulations

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

\( \sigma_s^2 \) 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
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

<table>
<thead>
<tr>
<th>Population Group</th>
<th>Approximate Carrier Frequency</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ashkenazi Jewish</td>
<td>1:29</td>
<td>Kerem et al [1997]</td>
</tr>
<tr>
<td>North American of northern European heritage</td>
<td>1:28</td>
<td>Hamosh et al [1998]</td>
</tr>
<tr>
<td>African American</td>
<td>1:61</td>
<td>Hamosh et al [1998]</td>
</tr>
</tbody>
</table>

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.

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

1000 Genome Project

http://www.1000genomes.org/data
ONLINE RESOURCES: SNPedia

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

**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
- Understand the concept of haplotype and linkage disequilibrium