SNP I

SNP background

張學偉  助理教授
SNP = Single Nucleotide Polymorphism

(read in SNiP)
• **Genetic Polymorphism:**
  A difference in DNA sequence among individuals, groups, or populations.

• **Genetic Mutation:**
  A change in the nucleotide sequence of a DNA molecule. Genetic mutations are a kind of genetic polymorphism.
SNPs are an abundant form of common genome variation, distinguished from rare variations by a requirement for the least abundant allele to have a frequency of 1% or more.
4 states of a SNP are possible (A,T,C,G)
different alleles

But the diallelic type is the most common SNP (e.g. T->C)
SNP discovery

Figure 1

Alternative methods for sequence-based SNP discovery. BAC, bacterial artificial chromosome; RRS, reduced representation sequencing.

Current Opinion in Chemical Biology
Genome Sequencing finds SNPs

- The Human Genome Project involves sequencing DNA cloned from a number of different people.
  
  [The Celera sequence comes from 5 people]

- Even in a library made from one person’s DNA, the homologous chromosomes have SNPs

- This inevitably leads to the discovery of SNPs - any single base sequence difference

- These SNPs can be valuable as the basis for diagnostic tests
If a matching SNP is found, then it can be directly located on the Genome map.
A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms

The International SNP Map Working Group*  
* A full list of authors appears at the end of this paper.

0.3% of genome
SNPs are the most simple form and most common source of genetic polymorphism in the human genome (90% of all human DNA polymorphisms).
SNP markers

• SNPs are very common in the human population.

• SNPs can be found that are linked to any disease alleles.

• These mutations are likely to be neutral - they have no direct effect on phenotype

• Linked SNPs can be used as markers for the disease in diagnostic tests.
Why are SNPs Important?

- Most common form of genetic variation
  - Genetic linkage studies
  - Genome-wide association studies
- Indicate predisposition to
  - Disease predisposition and onset
  - Drug tolerance
  - Drug efficacy
- Genome SNP scans will uncover gene function, and define new drug targets
- SNPs will enable physicians to personalize therapy
Classification of SNP- I

1. Most commonly changes
   Transitions  $\rightarrow$ pu to pu/ py to py
   Transversion $\rightarrow$ py to pu/ pu to py

2. Single-base insertions & deletion (indel)

3. Simple nucleotide polymorphism
   two-nucleotide changes &
   small indels up to a few nucleotides
Nature of genetic code in transitions & transversion

1. transitions are less likely to affect amino acids than are transversion
2. Transitions might have high probability of retention in coding region.
3. The transition/transversion ratio is affected by constraints on sequence evolution.
Classification of SNP-II

1. Non-coding SNPs
   - 3',5'-NTR (non-transcribed region)
   - 3',5'-UTR (untranslated region)
   - Intron
   - Intergenic

2. Coding SNPs
   - Replacemental polymorphism (change a.a.)
   - Synonymous polymorphism (unchange a.a.)
Non-replacemental polymorphism

Include:
synonymous & non-coding SNP

Function:
may affect gene function through
  1. transcriptional regulation
  2. translational regulation
  3. splicing
  4. RNA stability
Apolipoprotein E gene promoter (−219G/T) polymorphism is associated with premature coronary heart disease
A Single Nucleotide Polymorphism in the Matrix Metalloproteinase-1 Promoter Enhances Lung Cancer Susceptibility

Yong Zhu, Margaret R. Spitz, Lei Lei, Gordon B. Mills, and Xifeng Wu

Departments of Epidemiology [Y. Z., M. R. S., L. L., X. W.] and Molecular Therapeutics [G. B. M.]. The University of Texas M. D. Anderson Cancer Center, Houston, Texas 77030
Human promoter SNP that affect gene expression
Distribution of SNP

SNPs will be evenly distributed across the human genome

- ~ 25,000 non-synonymous cSNPs
- ~ 50,000 synonymous cSNPs
- ~ 25,000 regulatory region SNPs
- ~ 50,000 intragenic non-coding SNPs
- ~ 50,000 distributed intergenic SNPs
**Table 1. Publicly available SNP databases on the internet**

<table>
<thead>
<tr>
<th>Public SNP databases</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCBI database of deposited SNPs</td>
<td></td>
</tr>
<tr>
<td>The SNP Consortium&lt;sup&gt;ab&lt;/sup&gt;</td>
<td>snp.cshl.org</td>
</tr>
<tr>
<td>Cancer Genome Anatomy Project: GAf</td>
<td>lpg.nci.nih.gov/GAI</td>
</tr>
<tr>
<td>HGVSbase&lt;sup&gt;ab&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>Database of Japanese Single Nucleotide Polymorphisms&lt;sup&gt;ab&lt;/sup&gt;</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Deposited in dbSNP.
<sup>b</sup>International database.
<sup>c</sup>Search EST sequences for SNPs.

Abbreviations: EST, expressed sequence tag; GAi, Genetic Annotation Initiative; HGV, Human Genome Variation; NCBI, National Center for Biotechnology Information; SNP, single nucleotide polymorphism.
SNP RELATED DATABASES

- Androgen Receptor Mutation Database
- Ataxia-Telangiectasia Mutation Database
- Breast Cancer Mutation Database
- Cystic Fibrosis Mutation Database
- Cytokine Polymorphism Database
- dbSNP
- Factor VIII Database
- Fanconi Anemia Mutation Database
- GM2 Gangliosidoses Database
- Human Ornithine Transcarbamylase Database
- Human Type I and Type III Collagen Mutation Database
- Hypertension Candidate Gene SNP Database
Cancer Genome Anatomy Project

NCI > CGAP > GAI > SNP Maps

Annotations from UniGene 159

What's New

Imagemap Guide

CGAP SNP Index

Cytogenetic Search

RH Map Search

Gene Viewer
| Consensus                  | AGCAAGAATCATAGACAGCTACTACACGAGGTCTAGTTCAGAAATATACGAGGAGGCTACAGAGGATATTAGTTACACGATGATCCTACCTACCCACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTACCTA
The SNP Consortium is an unlikely alliance of pharmaceutical and computer companies managed by Lincoln Stein at Cold Spring Harbor Lab.

“The SNP Consortium Ltd. is a non-profit foundation organized for the purpose of providing public genomic data. Its mission is to develop up to 300,000 SNPs distributed evenly throughout the human genome and to make the information related to these SNPs available to the public without intellectual property restrictions. The project started in April 1999 and is anticipated to continue until the end of 2001.”

The current release (Jan 2001) consists of 856,666 SNPs, all of which have been anchored to the human genome by "in silico" mapping to the genomic working draft (UCSC “Golden Path”).
Search for SNPs in your gene

“an average density ... of one SNP every 1.9 kilobases”

But that does not guarantee a SNP in your favorite gene!
SNP Mapping

• Identify SNP sites along the genome to track disease genes.

• A human SNP map would specify the contributions of individual genes to diseases and other phenotypes.
Digital DNA Signatures – principle

Each SNP locus is queried for the presence or absence of a specific base as shown below. The four possible results are coded in digital form – 10: homozygous for allele1, 11: heterozygous, 01: homozygous for allele2, 00: assay failure – and presented in a defined order. The resulting string of binary digits is termed Digital DNA Signature.

allele1
ACTCTGGTAAATCATGIGTTGCTACTTACTGTGACTCTATATGT
ACTCTGGTAAATCATGIGTTGCTACTTACTGTGACTCTATATGT

allele2

base_1 present?
Yes → 1
No → 0

base_2 present?
Yes → 1
No → 0

heterozygous animal → 11
homozygous base_1 → 10
homozygous base_2 → 01
no information → 00

11 11 10 11 11 00 11 01 01 01 10 11 ...

Locus 2 ...
Locus 1: Sex specific SNP (polymorphism between the homologous genes ZFX and ZFY genotype 11 accords to a male individual)