Studying Genetic Variation I: Computational Techniques

Jim Mullikin, PhD
Genome Technology Branch
NHGRI

Some points from other lectures

- Population Genetics: Practical Applications by Lynn Jorde
  - Described patterns of human genetic variation among and within populations, linkage disequilibrium and HapMap and how all this relates to the search for complex disease genes.

- Identification of Cancer Susceptibility Genes by Elaine Ostrander
  - Genome wide scans to find cancer susceptibility genes and apply haplotype analyses to identify founder haplotypes.

- Genetic Variation II: Laboratory Techniques by Karen Mohlke
  - Focusing primarily on SNP genotyping methods
Overview of Topics

• Genome variation origins
• Types of polymorphisms
• Polymorphism discovery methods
• Access to genetic variation data
• How to find SNPs in a region of interest
• Haplotype Map project
• Extra topics, time permitting
Genome variation origins

• Mutations are fundamentally produced by errors in DNA replication.

• DNA is replicated in the production of the egg and sperm cells.

• Thus, a child does not receive exact copies of information from mother and father.

Types of polymorphisms

• Single Nucleotide Polymorphisms (SNPs) are single base changes and occur at a rate of about 30 - 60 sites per genome per generation.

ACTCCTCT\textcolor{red}{T}ATCCCTGC
ACTCCTCT\textcolor{red}{C}ATCCCTGC
ACTCCTCT\textcolor{red}{[C/T]}ATCCCTGC
Types of polymorphisms

• Short Tandem Repeats (STRs) are specific repeated segments of sequence.

```
GGTTTTTTGCC------TATATATATAAGTAGGA
GGTTTTTTGCC----TATATATATAAGTAGGA
GGTTTTTTGCC--TATATATATATATAAGTAGGA
GGTTTTTTGCC-TATATATATATATATAAGTAGGA
TTGCC[(TA)5/(TA)6/(TA)7/(TA)8]AGT
```

Types of polymorphisms

• Deletion/Insertion Polymorphisms (DIPs) are deletions or insertions of 1 base to as large as a few kilobases.

```
CATAAAAAAAAGAACAAATC
CATAAAAA--AACAAAAATC
CATAAAAAAA [G/-]AACAAAATC
```
Beyond polymorphisms

• When a mutational event is sufficiently large, these events are classified as chromosomal rearrangements.

• There are many examples of these as seen in karyotypes.

• These larger scale rearrangements, duplications or deletions are often associated with various diseases and severe abnormalities.

Overview of Topics

• Genome variation origins
• Types of polymorphisms
• SNP discovery methods
• Access to genetic variation data
• How to find SNPs in a region of interest
• Haplotype Map project


**Discovery methods**

- The primary method for discovering polymorphisms is by sequencing DNA and comparing the sequences.

Trace Archive Search query:
```
trace_name='S213P602260RC9.T0' OR trace_name='50766946'
```
Mining SNPs from sequence

• EST mining
• Clone overlap
• The SNP Consortium (TSC)
• Targeted resequencing
• Haplotype Map Project (HapMap)
• Chip based sequencing arrays

Expressed Sequence Tag Mining

• These sequences are primarily associated with coding regions of genes.

• By clustering these sequences, selected differences are identified as SNPs.

• There are over 100,000 SNPs in dbSNP from a variety of species detected from clustered ESTs.

• The following example is from the CGAP SNP project (see refs).
Clone Overlap

- The human genome was sequenced from BAC clones (containing about 150kb of sequence each).

- These overlapped to various levels, and within the overlap regions, high quality base differences indicated the position and alleles of SNPs.
Clone Overlap

• About 1.3M SNPs in dbSNP come from mining of clone overlaps.

• Special care was required to insure that the overlapping clones came from different haploids. (see references)

• This can be accomplished by looking at the source DNA for the two clones to see that it originated from different individuals, or if from the same individual, that the variation rate within the overlapping regions indicated that the DNA was from different haploids of one individual.

The SNP Consortium

• A two year effort funded by the Wellcome Trust and 11 pharmaceutical and technological companies to discover 300,000 SNPs randomly distributed across the human genome.

• At its initiation in April 1999, the genome was only 10% finished and 20% in draft form.

• The SNPs were developed from a pool of DNA samples obtained from 24 individuals representing several ethnic groups.
The SNP Consortium

- With the rapid increase in genome coverage from the public Human Genome Project, the strategies changed to take full advantage of the draft and finished sequence.

- The initial target of 300,000 SNP was passed quickly, and now the sequence generated from that project contributes over 1.3M SNPs to the public archives.
More SNPs for HapMap Project

• This project required many more SNPs than were available when it started in October 2002, which totaled about 2M.

• Additional random shotgun sequencing has brought this to 8.2M SNPs for the HapMap Project.

• It has been estimated that there are perhaps 10M common SNPs (> 5% MAF), so there are many more SNPs yet to discover.

Targeted Resequencing (Medical Sequencing)

• Any region of the genome can be targeted for resequencing. From the finished sequence, PCR primers can be designed to amplify a target followed by sequencing.

• This method generally works from a 1:1 mixture of an individuals two haploids, so the special case of heterozygous base positions must be properly processed.
Targeted Resequencing

- JSNP database contains 190,562 SNPs detected from resequencing genomic regions containing genes in DNA from 24 Japanese individuals.

- Many groups use this technique for either SNP discovery in their region of interest, or as a way to validate SNPs.

- PolyPhred (see web links) is commonly used for analyzing resequencing traces.
SNP detection by PolyPhred. View of a Censored window with a tag (red=highest ranking SNP tag) marking the consensus position of the SNP in the traces and genotype tags marking each of the samples below (purple=homozygote, pink=heterozygote). On the right trace windows for alternate homozygotes (C/C (top) and G/G (bottom)) and a heterozygote (C/G) middle.

PolyPhred example from their web site.
Perlegen used Affymetrix’s chip design process to place 60M probes on a 5x5” chip. From 20 single haploid chromosome 21 chromosomes, they discovered 36k SNPs.

**Distribution properties**

- **EST mining**
  - Locates SNPs primarily within coding regions.

- **Clone overlap**
  - High density of SNPs within overlap regions, absent elsewhere.

- **The SNP Consortium (TSC)**
  - Randomly distributed across the genome, however, total sequence only covers 50% of the genome.
**Distribution properties**

- **Haplotype Map Project (HapMap)**
  - Random, like TSC, for first phase that reached 2X coverage
  - Chromosome sorted phase increased coverage from 1X-6X

- **Targeted resequencing**
  - Focused discovery that has been applied to 100s of individuals

- **Chip based resequencing**
  - Repetitive elements in the genome are masked

---

**SNPs detected from 48 HapMap individuals gives an estimate dbSNP build 121 completeness**

![SNP distribution chart](chart.png)

- Fraction of all SNPs in dbSNP
- Allele frequency estimated from 210 individuals
Overview of Topics

• Genome variation origins
• Types of polymorphisms
• SNP discovery methods
• Access to genetic variation data
• How to find SNPs in a region of interest
• Haplotype Map project

NCBI dbSNP database of genetic variation

• This is the main repository of publicly available polymorphisms.

• You’ll also find information on allele frequencies, populations, genotypes assays and much more.

• Most groups submit SNPs to dbSNP and only a few maintain web access to their SNPs.
Submitting SNPs to dbSNP

- From their main web page, they have extensive information on how to submit SNPs, genotypes, validation experiments, population frequencies, etc., for any species.

- SNPs that you submit are called Submitter SNPs and get ssIDs.

- If there is a reference sequence available for the species submitted, they will map SNPs to this reference using the flank information you provide.

- SNPs that cluster at the same locus, are merged into Reference SNPs which have unique rsIDs.
### Fasta sequence

```
GAGAGAAGA TGGTCTGG GGGTGTGG CAGCAGCA CTGACGCT T
CTAGCTGG TGGCTGCTG CTGCTGCTG TGGCTGCTG TGGCTGCTG
GAGAGAAGA TGGTCTGG GGGTGTGG CAGCAGCA CTGACGCT T
CTAGCTGG TGGCTGCTG CTGCTGCTG TGGCTGCTG TGGCTGCTG
5
```

### GeneView

**GeneView via analysis of curation annotation:** ADCCIR associates related protein 2D complex, volume 1B, VRU

<table>
<thead>
<tr>
<th>Group Label</th>
<th>Curation mRNA</th>
<th>Gene Model (curation mRNA transcript)</th>
<th>Curation Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cede</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CEA_TCA_GAr76</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Integrated Maps

- **NCBI MapViewer:** eEn451/12 maps match one on NCB1 human chromsome 7

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Curation accession</th>
<th>Curation position</th>
<th>Curation position</th>
<th>Curation orientation</th>
<th>Curation Allele</th>
<th>Assembly Type</th>
<th>Group label</th>
<th>Curation label</th>
<th>Neighbor SNP_haplotype position</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>HPW_253542</td>
<td>22373509 59718053</td>
<td>59718053</td>
<td>forward</td>
<td>250</td>
<td>C1</td>
<td>C1</td>
<td>HPW_253542</td>
<td>59718053</td>
</tr>
<tr>
<td>7</td>
<td>NT_255992</td>
<td>24516105 98222950</td>
<td>98222950</td>
<td>forward</td>
<td>140</td>
<td>C1</td>
<td>C1</td>
<td>NT_255992</td>
<td>98222950</td>
</tr>
</tbody>
</table>

### NCBI Resource Links

- **GeneBank:** RefSeq, dbSNP Blast Analysis, UniGene Cluster ID, 3D structure mapping

### Population Diversity

<table>
<thead>
<tr>
<th>sample</th>
<th>Population</th>
<th>Sample Assortment</th>
<th>Sample Founder (N)</th>
<th>Genotypes</th>
<th>Alleles</th>
</tr>
</thead>
<tbody>
<tr>
<td>e242707M</td>
<td>AFD EUR PANEL</td>
<td>European</td>
<td>48 24</td>
<td>55</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>African American</td>
<td>46</td>
<td>23</td>
<td>77</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td>48</td>
<td>24</td>
<td>51</td>
<td>51</td>
</tr>
<tr>
<td>e2472229</td>
<td>AFD EUR PANEL</td>
<td>African American</td>
<td>46</td>
<td>23</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td>48</td>
<td>24</td>
<td>51</td>
<td>51</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>sample</th>
<th>Population</th>
<th>Sample Assortment</th>
<th>Sample Founder (N)</th>
<th>Genotypes</th>
<th>Alleles</th>
</tr>
</thead>
<tbody>
<tr>
<td>e242707M</td>
<td>AFD EUR PANEL</td>
<td>European</td>
<td>48 24</td>
<td>55</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>African American</td>
<td>46</td>
<td>23</td>
<td>77</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td>48</td>
<td>24</td>
<td>51</td>
<td>51</td>
</tr>
<tr>
<td>e2472229</td>
<td>AFD EUR PANEL</td>
<td>African American</td>
<td>46</td>
<td>23</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td>48</td>
<td>24</td>
<td>51</td>
<td>51</td>
</tr>
</tbody>
</table>
Viewing SNPs in Browsers

NCBI  Ensembl  UCSC

Overview of Topics

- Genome variation origins
- Types of polymorphisms
- SNP discovery methods
- Access to genetic variation data
- How to find SNPs in a region of interest
- Haplotype Map project
How to find SNPs in a region of interest

- Gene based example
- A 2 Mbp region
- From a list of candidate genes


http://innateimmunity.net/IIPGA/PGAs/InnateImmunity/CLCA1
Many submissions, however, possibly all from same source sequences.
How to find SNPs in a region of interest

• Gene based example

• A 2 Mbp region

• From a list of candidate genes

http://genome.ucsc.edu
### Table Browser

Use this program to get the data associated with a track in text format, to calculate intrinsic description of the controls in this form.

**clade:** Vertebrate  
**genome:** Human  
**assembly:** Mar. 2008  
**group:** Variation and Repeats  
**track:** SNPs  
**table:** snp126  
**region:**  
**identifiers (names/accessions):** paste list  
**filter:**  
**intersection:** create  
**correlation:** create  

---

### Filter on Fields from hg18.snp126

- `bm` is ignored  
- `chrom` does match  
- `chromStart` is ignored  
- `chromEnd` is ignored  
- `name` does match  
- `score` is ignored  
- `strand` does match  
- `refStart` does match  
- `refEnd` does match  
- `observed` does match  
- `missType` does match  
- `class` does match  
- `qual` does match  
- `avai` does match  
- `avaiStart` is ignored  
- `avaiEnd` is ignored  
- `gencode` does include  
- `isoType` does match  
- `weight` is ignored  

---

### Table Browser

Use this program to get the data associated with a track in text format, to calculate intrinsic description of the controls in this form.

**clade:** Vertebrate  
**genome:** Human  
**assembly:** Mar. 2008  
**group:** Variation and Repeats  
**track:** SNPs  
**table:** snp126  
**region:**  
**identifiers (names/accessions):** paste list  
**filter:**  
**intersection:** create  
**correlation:** create  
**output format:** hypers to Gnome Browser  
**output file:** (Enter file name to keep output in browser)  
**file type returned:** csv text  

---

**submit**
How to find SNPs in a region of interest

• Gene based example

• A 2 Mbp region

• From a list of candidate genes
Selecting SNPs from a list of candidate genes

- Use the Entrez SNP query:
  coding nonsynonymous[FUNC] AND CLCA*[Gene name] AND human[orgn]

- Download dbSNP database and cross reference with candidate gene list coordinates

Overview of Topics

- Genome variation origins
- Types of polymorphisms
- SNP discovery methods
- Access to genetic variation data
- How to find SNPs in a region of interest
- Haplotype Map project
**Haplotype Map project**

- What is a Haplotype?
- What is Linkage Disequilibrium (LD)?
- What is the Haplotype Map Project?

---

**What is a Haplotype?**

- A set of closely linked genetic markers present on one chromosome which tend to be inherited together (not easily separable by recombination).

- Recombination occurs between homologous chromosomes when cells divide.

- It is believed that recombination is not equally likely across the genome, but that it is punctuated by hot-spots.
What is Linkage Disequilibrium?

- When the observed frequencies of genetic markers in a population does not agree with haplotype frequencies predicted by multiplying together the frequency of individual genetic markers in each haplotype.
The origins of haplotypes in the human genome have been produced by the historical events that have shaped the population.

With the exception of the sex cells, the chromosomes in a chromosome pair are inherited from a person's father and mother. Each chromosome contains many markers that may be shared or not shared with each other. The markers in a chromosome pair come together and exchange pieces. The recombination process occurs in the ovaries or testes of each individual as the DNA is copied from the mother and father.
Identification of Haplotypes Through Genotyping

International HapMap Project

- Goal: to develop a haplotype map covering 80 - 90% of the genome
- The map should be usable in all populations
- Three year project started October 2002 and completed in October 2005 (Phase I)
- International collaboration, involving Canada, China, Nigeria, Japan, the United Kingdom, and the United States
- All data publicly accessible at www.hapmap.org
International HapMap Project: Sample Collection

- Similarity in haplotypes worldwide limits the need to collect samples from many populations
- No clinical information collected, samples anonymous
- Individual consent and extensive community consultation
- 270 samples collected and genotyped
  - Africa (Yoruba in Ibadan, Nigeria)
  - Asia (Japanese in Tokyo, Han Chinese in Beijing)
  - Europe (CEPH family samples, Utah)
- Samples are available as DNA or cell lines from Coriell
- Additional populations being studied in a pilot phase

International HapMap Project: Experimental Strategy

- Participating centers have divided up the genome, according to capacity of each center
- Different centers use different platforms: Illumina, Third Wave, Sequenom, TaqMan, ParAllele
- Data Coordination Center provides lists of SNPs, and receives genotypes
- Phase I HapMap – Obtain genotypes from a working SNP every 5 kb across the genome
- Phase II – Fill in gaps in linkage disequilibrium map: completed by Perlegen
**HapMap Milestones**

- Fall 2004 – Phase I map of 600,000 SNPs in European samples
- Early 2005 – Phase I map in Asian and African samples
- Fall 2005 – Perlegen contributes another 3M SNPs to the map
- Fall 2005 – Final HapMap, including gap filling
- “HapTag” SNPs able to represent 80-90% of common variation with
  - 200,000 SNPs for European or Asian samples
  - 400,000 SNPs for African samples

**HapMap Gbrowse**

http://www.hapmap.org/cgi-perl/gbrowse/hapmap_B35/
http://www.broad.mit.edu/mpg/tagger/
Overview of Topics

- Genome variation origins
- Types of polymorphisms
- SNP discovery methods
- Access to genetic variation data
- How to find SNPs in a region of interest
- Haplotype Map project
- Medical Sequencing
- SNPs for Other Species
- New Sequencing Technologies
A Brief Tour of a Medical Sequencing Pipeline

Choice of Genomic Regions

The regions of interest (ROIs) are typically defined by their biological context (coding, conservation, regulatory function, known variation). When features are in close proximity, the number of amplifiers is automatically reduced, maintaining optimal coverage.
Primer Ordering and Tracking

The design coverage of the ROIS and the status of amplifiers are tracked with the interfaces above. Once the design coverage is considered satisfactory, the primer pairs can be ordered automatically.

Exploring the data

List of projects and progress overview

Q20 per individual

List of subjects
The system keeps track of analysis performed on the data and coverage attained for each ROI. It also allows a user to browse the detected genotypes.
We are developing interfaces that allow exploring the results and identify interesting results as well flag problems.

Three examples of same SNP detected in overlapping amplimers. This information is used to assess accuracy of the detection.

Some of the challenges of variation detection
SNPs for Other Species

• Mouse
  – The reference strain sequenced, C57BL/6J, was inbred for sufficient generations to result in a homozygous genome, however, 15 mouse strains have been sequenced and the variations are available from dbSNP (http://www.nih.gov/news/pr/oct2006/niehs-25.htm)
  – This is a great resource for mouse genetics. For example, crossing two different mouse strains where one mouse has given disease causing mutation.

• Dog
  – The reference dog genome sequence comes from a fairly inbred individual (a boxer named Tasha). This individual is 60% homozygous with the heterozygous regions showing 1 SNP per 900 bases, giving 770k SNPs.
  – Celera sequenced a poodle, Shadow, and comparing this genome to Tasha’s sequence give 1.46M SNPs
  – The public sequencing effort also generated whole genome shotgun sequence from 9 other dog breeds as well as 4 wolves and a coyote

• Chimpanzee
  – The reference sequence in based on Clint along with light WGS of four other West African and three central African chimpanzees giving a total of 1.66M SNPs.
  – Chimpanzee sequence can also be used together with human SNPs to determine the ancestral allele state, as noted in many of the dbSNP records.

• Cat
  – The reference cat sequence, like dog, comes from an inbred individual (an Abyssinian named Cinnamon) which is also about 60% homozygous, with the heterozygous regions showing 1 SNP per 600 bases.
Cat SNP Analysis

- Cinnamon is of the Abyssinian breed, and its genome is diploid
- Thus, when two sequence traces overlap, there is a 50% chance that these two traces came from different chromosomes
- If Cinnamon were an out-bred cat, then traces that arise from different chromosomes should exhibit sequence polymorphisms
- However, due to inbreeding, the locus of these two chromosomes may have been derived from an ancestor’s chromosome only a few generations back, thus exhibiting no polymorphisms
Cinnamon’s Polymorphism Statistics

- 57% of Cinnamon’s autosomes are homozygous
- Within the heterozygous segments of this individual, we discovered over 325,000 SNPs and over 37,000 deletion/insertion polymorphisms
- The heterozygosity level of heterozygous regions is 0.17%, or about 70% higher than human heterozygosity levels
- Comparing Cinnamon to another cat (Gus), a brown classic tabby (RPCI-86), yields a heterozygosity level of about 0.2%, or about twice the level of humans.
Linkage Disequilibrium Across Cat Breeds

- Selected SNPs detected from Cinnamon’s genome within heterozygous regions on 10 different chromosomes.
- 35 SNPs were selected per chromosome, with the first 8 SNPs within a 15kb window and rest selected every approximately every 15kb away from the previous SNP.
- These SNPs were genotyped across 97 cats from 24 breeds, 7 outbred “alley” cats and 12 wild species.
- Linkage disequilibrium (LD) was calculated for those individuals that were homozygous within the first 15kb window, and the length of LD was derived from the extent of the homozygous interval.

![Linkage Disequilibrium Graph](image)
Summary of Cat LD Results

• ~60% of 10 kb regions are homozygous within an individual. This is very similar to dogs.

• Conditional on being homozygous within the 10 kb region, 50% of cases are still homozygous at 150 kb. The extent of linkage disequilibrium is roughly a third that in dogs.

• The number of markers needed for genome-wide association: current estimate about 45k markers.

New Sequencing Technologies

• 454 Life Sciences
  – 100-200 base reads
  – 20-40Mb per run
  – 2 runs per day

• Solexa
  – 25-40 base reads
  – 8*125Mb per run
  – 2 runs per week

• ABI SOLiD
  – Similar to Solexa
  – Run performance like Solexa
SNP Detection with New Sequencing Technologies

- Need to greatly over-sample each base to insure high quality SNP detection, about 30 fold redundancy
- To sequence an entire individual's genome requires 3Gb*30/1Gb/run or about 90 runs on a Solexa machine (45 weeks)
- Targeted sequencing requires additional preparation, e.g. long range (10kb) PCR
  - Introduces variable product amplification levels requiring greater average sequencing redundancy to ensure a minimum redundancy of 30 fold
  - Allelic PCR dropout resulting in missed genetic diversity
  - Approach has been successfully applied to a 140kb genomic interval

Concluding remarks

- Along with the emergence of the human genome, we also have a growing database of variations that are critical to the overall value of the human genome sequence.
- These variations are what make us all (phenotypically) different, and impart different levels of resistance and susceptibility to disease.
- The collection of human sequence variation as well as that for other species will continue to evolve rapidly.
References

EST SNPs

Clone Overlaps/TSC

Targeted Resequencing

Chip based SNP discovery

Haplotype Map Project
snp.cshl.org : The SNP Consortium web pages
http://droog.mbt.washington.edu/PolyPhred.html
http://www.ensembl.org : Ensembl home page
http://www.ucl.ac.uk/~ucbhdjm/courses/b242/2+Gene/2+Gene.html
http://www.hapmap.org/ : Haplotype Map Project home page
http://www.hapmap.org/cgi-perl/gbrowse/gbrowse/hapmap
http://www.broad.mit.edu/personal/jcbarret/haploview/
http://genome.perlegen.com/browser/index.html: Perlegen’s HapMap