SNP Resources: Variation Discovery, HapMap and the EGP

Mark J. Rieder
Department of Genome Sciences
mrieder@u.washington.edu
NIEHS SNPs Workshop
Jan 10-11, 2008

Complex inheritance/disease

Many Other Genes
Variant Gene
Environment
Disease
- Diabetes
- Obesity
- Heart Disease
- Multiple Sclerosis
- Cancer
- Schizophrenia
- Celiac Disease
- Asthma
- Autism

Two hypotheses:
1- common disease/common variant?
2- common disease/many rare variants?

Strategies for Genetic Analysis

Families Linkage Studies
Simple Inheritance
Single Gene
Rare Variants
~600 Short Tandem Repeat Markers

Populations Association Studies
Cases
- Complex Inheritance
- Multiple Genes
- Common Variants
- Polymorphic Markers > 300,000 - 1,000,000
- Single Nucleotide Polymorphisms (SNPs)

Controls
- 40% T, 60% C
- 15% T, 85% C

Strategies for Genetic Analysis (GWAS)

Samples with phenotype data (continuous, case-control)
(n = 100's - 1000's)

Genotype samples with commercial "chips"
- Affymetrix - Random SNP design (v.5, v.6)
- Illumina - preselected tagSNP design (650Y, 1M)

Perform statistical association with each SNP
Calculate p-value for each SNP

Region with plausible statistical association
Development of a genome-wide SNP map: How many SNPs?

Table 1 - Occurrence of SNPs in the human population

<table>
<thead>
<tr>
<th>Minimal allele frequency</th>
<th>Expected SNP number (millions)</th>
<th>Expected SNP frequency (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>11.0</td>
<td>290</td>
</tr>
<tr>
<td>5%</td>
<td>7.1</td>
<td>450</td>
</tr>
<tr>
<td>10%</td>
<td>5.3</td>
<td>600</td>
</tr>
<tr>
<td>20%</td>
<td>3.3</td>
<td>960</td>
</tr>
<tr>
<td>30%</td>
<td>2.0</td>
<td>1,570</td>
</tr>
<tr>
<td>40%</td>
<td>0.97</td>
<td>3,280</td>
</tr>
</tbody>
</table>

Nickerson and Kruglyak, Nature Genetics, 2001

~ 10 million common SNPs (> 1-5% MAF) - 1/300 bp

How has SNP discovery progressed toward this goal?

SNP Resources: SNP discovery and cataloging

1. SNP discovery/genotyping: Genome-wide approaches
   - SNP Consortium
   - HapMap

2. The current state of SNP resources (dbSNP)
3. Comprehensive SNP discovery
   - NIEHS SNPs - Environmental Genome Project

SNP Databases - “How to” Manual for finding SNPs
   - In class - Tutorial

HapMap Project: Create a genome-wide SNP map

Genotype SNPs in four populations:
- CEPH (CEU) (Europe - n = 90, trios)
- Yoruban (YRI) (Africa - n = 90, trios)
- Japanese (JPT) (Asian - n = 45)
- Chinese (HCB) (Asian - n = 45)

To produce a genome-wide map of common variation

Common Variant/Common Disease

Phase I - 1M SNPs
- Density ~ 1 SNP/kb
- 626 genes
- 15 Mbp
- 91,000 SNPs
- Density - 1 SNP/166 bp

Phase II - 4M SNPs
SNP discovery has proceeded in two distinct phases:

1 - SNP Identification/Discovery
   Define the alleles
   Map this to a unique place in the genome

2 - SNP Characterization (HapMap genotyping)
   Determination of the genotype in many individuals
   Population frequency of SNPs
**SNP Discovery: dbSNP database**

- dbSNP - NCBI SNP database

**SNP data submitted to dbSNP: Clustering**

- SNPs submitted by research community (submitted SNPs = rs#)
- Unique mapping to a genome location (reference SNP = rs#)

**Validation status description**

- Validated
- Unvalidated

**dbSNP processing of SNPs**

- Submission information
  - Search by IDs on All Assemblies
  - Submission information
    - By Individual
    - By Study
    - By Project
    - Other (description, release, and ID)
    - Data (based on geographic location)
    - Details
    - Detailed

**HapMap Discovery Increased SNP Density and Validated SNPs**

- 11+ million rs SNPs
- 5.6 million validated rs SNPs

**Finding SNPs: Marker Discovery and Methods**

SNP discovery has proceeded in two distinct phases:

1 - SNP Identification/Discovery
   - Define the alleles
   - Map this to a unique place in the genome

2 - SNP Characterization (HapMap genotyping)
   - Determination of the genotype in many individuals
   - Population frequency of SNPs
HapMap Project: Create a genome-wide SNP map

Genotype SNPs in four populations:
- CEPH (CEU) (Europe - n = 90, trios)
- Yoruban (YRI) (Africa - n = 90, trios)
- Japanese (JPT) (Asian - n = 45)
- Chinese (HCB) (Asian - n = 45)

To produce a genome-wide map of common variation

Finding SNPs: Genotype Data Adds Value to SNPs

- Confirms SNP as “real” and “informative”
- Minor Allele Frequency (MAF) - common or rare
- MAF differs by different population
- Detection of SNP x SNP correlations (Linkage Disequilibrium)
- Determine tagging SNPs (tagSNPs)
- Determine haplotypes

dbSNP: All validated SNPs now have been genotyped!

HapMap release = 4 M validated QC SNPs
Development of a genome-wide SNP map: How many SNPs?

Table 1: Occurrence of SNPs in the human population

<table>
<thead>
<tr>
<th>Minimal allele frequency</th>
<th>Expected SNP number (millions)</th>
<th>Expected SNP frequency (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>11.0</td>
<td>290</td>
</tr>
<tr>
<td>5%</td>
<td>7.1</td>
<td>450</td>
</tr>
<tr>
<td>10%</td>
<td>5.3</td>
<td>600</td>
</tr>
<tr>
<td>20%</td>
<td>3.3</td>
<td>960</td>
</tr>
<tr>
<td>30%</td>
<td>2.0</td>
<td>1,570</td>
</tr>
<tr>
<td>40%</td>
<td>0.97</td>
<td>3,280</td>
</tr>
</tbody>
</table>

Nickerson and Kruglyak, Nature Genetics, 2001

- 10 million common SNPs (> 1-5% MAF) - 1/300 bp
- Feb 2001 - 1.42 million (1/1900 bp)
- Nov 2003 - 2.0 million (1/1500 bp)
- Feb 2004 - 3.3 million (1/900 bp)
- Mar 2007 - 5.6 million (validated - 1/535 bp)

When will we have them all?

Finding SNPs: Sequence-based SNP Mining

Genomic
- BAC Library
- RRS Library
- Random Shotgun
- mRNA Library

DNA SEQUENCING
- BAC Overlap
- Shotgun Overlap
- Align to Reference
- EST Overlap

RANDOM Sequence Overlap - SNP Discovery

GTTACGCCAATACAGATCCAGGAGATTACC
GTTACGCCAATACACATCCAGGAGATTACC

SNP Characterization/Genotyping

Table 1: Occurrence of SNPs in the human population

<table>
<thead>
<tr>
<th>Minimal allele frequency</th>
<th>Expected SNP number (millions)</th>
<th>Expected SNP frequency (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>11.0</td>
<td>290</td>
</tr>
<tr>
<td>5%</td>
<td>7.1</td>
<td>450</td>
</tr>
<tr>
<td>10%</td>
<td>5.3</td>
<td>600</td>
</tr>
<tr>
<td>20%</td>
<td>3.3</td>
<td>960</td>
</tr>
<tr>
<td>30%</td>
<td>2.0</td>
<td>1,570</td>
</tr>
<tr>
<td>40%</td>
<td>0.97</td>
<td>3,280</td>
</tr>
</tbody>
</table>

Nickerson and Kruglyak, Nature Genetics, 2001

- 10 million common SNPs (>1-5% MAF) - 1/300 bp
- Mar 2007 - 5.6 million (validated/mapped - 1/535 bp)

When will we have them all?

HapMap data doesn’t capture low frequency alleles
Used as a proxy for all common variation through linkage disequilibrium
**Targeted SNP Discovery**

**Directed analysis:** cSNPs

5'   Arg-Cys   Val-Val   3'

PCR amplicons

**Complete analysis:** cSNPs, Linkage Disequilibrium and Haplotype Data

5'   Arg-Cys   Val-Val   3'

PCR amplicons

- Generate SNP data from complete genomic resequencing (i.e. 5' regulatory, exon, intron, 3' regulatory sequence)

---

**Increasing Sample Size Improves SNP Discovery**

2 chromosomes

{ GTTACGCAATACAGGATCCAGGAGATTACC
  GTTACGCAATACAGCATCCAGGAGATTACC

Resequencing (n=48)

HapMap Based on ~8 chromosomes

Minor Allele Frequency (MAF)

-0.0  0.0  0.1  0.2  0.3  0.4  0.5

Fraction of SNPs Discovered

---

**Increasing SNP Density: HapMap ENCODE Project**

**ENCOD...**

ENCOD...  ENCyclopedia Of DNA Elements
Catalog all functional elements in 1% of the genome (30 Mb)

10 Regions x 500 kb/region (Pilot Project)
David Altschuler (Broad), Richard Gibbs (Baylor)
16 CEU, 16 YRI, 8 HCB, 8 JPT

Comprehensive PCR based resequencing across these regions

- 15,367 dbSNP
- 16,248 New SNPs
- 50% of SNPs in dbSNP

5 Mb/31,500 SNPs = 1/160 bp

**NIEHS SNPs**

Goal: Comprehensively identify all common sequence variation in candidate genes

Initial biological focus: Candidate environmental response genes involved in DNA repair, cell cycle, apoptosis, metabolism, cell signaling, and oxidative stress.

Approach: Direct resequencing of genes

Samples:

- PDR-90 ethnically diverse individuals representative of U.S. population (397 genes)
- EGP95-95 samples from four ethnic groups (227 genes)
  - (24 HapMap Asians, 22 HapMap Europeans, 12 HapMap Yorubans, 15 African Americans, 22 Hispanic)
Summary of NIEHS SNPs genotypes in dbSNP

<table>
<thead>
<tr>
<th>Data set</th>
<th>Genotypes</th>
<th>SNPs</th>
<th>Populations</th>
<th>Individuals</th>
<th>Genotype Density</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIEHS 2004</td>
<td>110,382,776</td>
<td>1,536,322</td>
<td>3</td>
<td>270</td>
<td>3516</td>
<td>(International HapMap Consortium 2005)</td>
</tr>
<tr>
<td>DBSNP336</td>
<td>81,186,466</td>
<td>525,278</td>
<td>6</td>
<td>176</td>
<td>24,529</td>
<td>(Kazazian et al. 2004)</td>
</tr>
<tr>
<td>HG_1575142</td>
<td>4,232,962</td>
<td>339,484</td>
<td>15</td>
<td>1860</td>
<td>318,740</td>
<td>(International SNP Map Working Group 2001)</td>
</tr>
<tr>
<td>NIEHS 2005</td>
<td>521,776</td>
<td>40,000</td>
<td>8</td>
<td>50</td>
<td>57,500</td>
<td>(International HapMap Consortium 2005)</td>
</tr>
<tr>
<td>HapMap YRI</td>
<td>112,949</td>
<td>19,883</td>
<td>3</td>
<td>448</td>
<td>1,410,123</td>
<td>(International HapMap Consortium 2005)</td>
</tr>
<tr>
<td>HapMap JPT</td>
<td>24,532</td>
<td>520</td>
<td>1</td>
<td>143</td>
<td>5,226,155</td>
<td>(International HapMap Consortium 2005)</td>
</tr>
</tbody>
</table>

*The NIEHS data contains a single mixed population.

626 genes sequenced
15 Mb scanned
> 90,000 genotyped SNPs identified
> 8 million genotypes deposited in dbSNP

Nov 2005 - Zaitlen et al. Genome Research 15:1594-1600

<table>
<thead>
<tr>
<th>Minimal allele frequency</th>
<th>Expected SNP number (millions)</th>
<th>Expected SNP frequency (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>11.0</td>
<td>290</td>
</tr>
<tr>
<td>5%</td>
<td>7.1</td>
<td>450</td>
</tr>
<tr>
<td>10%</td>
<td>3.3</td>
<td>900</td>
</tr>
<tr>
<td>20%</td>
<td>1.9</td>
<td>1,570</td>
</tr>
<tr>
<td>30%</td>
<td>0.87</td>
<td>5,280</td>
</tr>
</tbody>
</table>

Nickerson and Kruglyak, Nature Genetics, 2001

~ 10 million SNPs (>1% MAF-estimated) - 1/300 bp

NIEHS SNPs = 1/166 bp (n = 95, 4 pops)
HapMap ENCODE = 1/160 bp (n = 48, 3 pops)

Comprehensive resequencing can identify the vast majority of SNPs in a region

SNP Discovery: dbSNP database

dbSNP (Perlegen/HapMap)

NIEHS SNPs

SNP Distribution

Minor Allele Freq. (MAF)

Rarer and population specific SNPs are found by resequencing
nsSNPs are found in ~60% of genes

Protein disrupting SNPs in EGP data

Computational inference of nsSNP function

Deleterious nsSNPs are preferentially low frequency
EGP and HapMap Genotyping

- PDR = 90 ethnically diverse individuals representative of U.S. population
- EGP p1 panel
- Select SNPs ~7600 SNPs Illumina
- 397 genes/55,000 SNPs

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>European</td>
<td>CEU, n=60</td>
</tr>
<tr>
<td>African</td>
<td>YRI, n=60</td>
</tr>
<tr>
<td>Asian</td>
<td>HCB, n=45</td>
</tr>
<tr>
<td></td>
<td>JPT, n=45</td>
</tr>
<tr>
<td>African-American</td>
<td>n=62</td>
</tr>
<tr>
<td>Hispanic</td>
<td>n=60</td>
</tr>
</tbody>
</table>

Illumina NIEHS SNPs Genotyping

- Each well samples 1536 SNPs in one individual
- For each HapMap sample 5 x 1536 (7680 genotyped SNPs)
- 3,000,000 genotypes generated (total ~400 samples)

<table>
<thead>
<tr>
<th>Array (1536)</th>
<th>Site Conversion Rate (%)</th>
<th>Average Site Coverage (%)</th>
<th>Concordance (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>85</td>
<td>96.6</td>
<td>99.7</td>
</tr>
<tr>
<td>2</td>
<td>91</td>
<td>97.7</td>
<td>99.5</td>
</tr>
<tr>
<td>3</td>
<td>82</td>
<td>98.5</td>
<td>99.3</td>
</tr>
</tbody>
</table>

NIEHS SNPs Genotype Data

- PDR (397 genes)
- SNPs characterized in six different major populations.

Population Allele Frequency Correlations

Illumina NIEHS SNPs Genotyping
Summary: The Current State of SNP Resources

- Approximately 10 million common SNPs exist in the human genome (1/300 bp).
- Random SNP discovery processes generate many SNPs (HapMap)
- Random approaches to SNP discovery have reached limits of discovery and validation (1/600 bp; 50% SNP validation)
- Most validated SNPs (5+ million) have been genotyped by the HapMap (3 pops)
- Resequencing approaches continue to catalog important variants (rare and common not captured by the HapMap)
- NIEHS SNPs has generated SNP data on >600 candidate genes and 90 K SNPs along with large-scale HapMap+ characterization data for of SNPs generated using the PDR