Modern mapping of monogenic diseases

Colombo, April 2013
Magnus Dehli Vigeland
Overview

- Disease mapping by high throughput sequencing
- Homozygosity mapping
- More general family based sequencing
Traditional mapping of monogenic disorders

1. Collect (large) affected families

2. Parametric linkage analysis

3. Sequence genes in linkage peak \rightarrow identify causal mutation
Modern mapping (since 2009)

1. Find patients
2. Sequence their genomes/exomes → identify causal gene

How does this work?

List all genes where all patients have variants (compared with reference genome)

Filter out known variants (dbSNP, 1000 Genomes, etc)

Hopefully, only one gene remains.

If not, rank according to predicted harmfulness
Modern mapping (since 2009)

1. Find patients
2. Sequence their genomes/exomes $\rightarrow$ identify causal gene

Why does this work now?

Databases of human variation
1000G, dbSNP, ++

High throughput sequencing
The human genome

• Complete genome:
  – ~ 3 000 000 000 base pairs
  – 20 000 - 25 000 genes

• The exome:
  – all exons (protein-coding DNA regions)
  – ~ 1.5% of the genome
  – ~ 180 000 exons
  – Faster and cheaper than whole-genome sequencing
Targeted capture and massively parallel sequencing of 12 human exomes

Sarah B. Ng¹, Emily H. Turner¹, Peggy D. Robertson¹, Steven D. Flygare¹, Abigail W. Bigham², Choli Lee¹, Tristan Shaffer¹, Michelle Wong¹, Arindam Bhattacharjee⁴, Evan E. Eichler¹,³, Michael Bamshad², Deborah A. Nickerson¹ & Jay Shendure¹
Example: Freeman-Sheldon syndrome

- 12 exomes were sequenced
  - 4 patients with FSS (not related)
  - 8 HapMap-individuals, used as controls

- Proof of concept:
  - FSS is dominant, very rare
  - Known gene: \textit{MYH3} (chr. 17)

The big question: \textit{Are 4 individuals enough to identify \textit{MYH3}?}
Example: Freeman-Sheldon syndrome

### Table 2 | Coding variation across 12 human exomes

<table>
<thead>
<tr>
<th>Individual</th>
<th>cSNP calls</th>
<th>Number in dbSNP</th>
<th>Percentage in dbSNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA18507 (YRI)</td>
<td>19,720</td>
<td>17,577</td>
<td>89.1</td>
</tr>
<tr>
<td>NA18517 (YRI)</td>
<td>19,737</td>
<td>17,326</td>
<td>87.8</td>
</tr>
<tr>
<td>NA19129 (YRI)</td>
<td>19,761</td>
<td>17,298</td>
<td>87.5</td>
</tr>
<tr>
<td>NA19240 (YRI)</td>
<td>19,517</td>
<td>17,168</td>
<td>88.0</td>
</tr>
<tr>
<td>NA18555 (CHB)</td>
<td>16,047</td>
<td>14,894</td>
<td>92.8</td>
</tr>
<tr>
<td>NA18956 (JPT)</td>
<td>16,011</td>
<td>14,848</td>
<td>92.7</td>
</tr>
<tr>
<td>NA12156 (CEU)</td>
<td>16,119</td>
<td>15,250</td>
<td>94.6</td>
</tr>
<tr>
<td>NA12878 (CEU)</td>
<td>15,970</td>
<td>15,051</td>
<td>94.2</td>
</tr>
<tr>
<td>FSS10066 (Eur)</td>
<td>16,229</td>
<td>15,144</td>
<td>93.3</td>
</tr>
<tr>
<td>FSS10208 (Eur)</td>
<td>16,073</td>
<td>15,018</td>
<td>93.4</td>
</tr>
<tr>
<td>FSS22194 (Eur)</td>
<td>16,094</td>
<td>15,128</td>
<td>94.0</td>
</tr>
<tr>
<td>FSS24895 (Eur)</td>
<td>15,986</td>
<td>15,027</td>
<td>94.0</td>
</tr>
</tbody>
</table>
Example: Freeman-Sheldon syndrome

![Table and Figure 2]

**Figure 2 | Direct identification of the causal gene for a monogenic disorder**
Overview

• Disease mapping by high throughput sequencing

• Homozygosity mapping

• More general family based sequencing
Terminology: Autozygosity

Heterozygous $\rightarrow$ AB $\rightarrow$ BB $\leftarrow$ Homozygous

Autozygous = homozygous + IBD

IBD = "identical by descent"
IBS = "identical by state"
The inbreeding coefficient $F$

$F = \text{probability of autozygosity at random locus}$

$= \text{expected fraction of genome being autozygous}$

<table>
<thead>
<tr>
<th>Parental relation</th>
<th>$F$ (of offspring)</th>
</tr>
</thead>
<tbody>
<tr>
<td>father/daughter; siblings</td>
<td>1/4</td>
</tr>
<tr>
<td>uncle/niece; half sibs</td>
<td>1/8</td>
</tr>
<tr>
<td>1st cousins</td>
<td>1/16</td>
</tr>
<tr>
<td>2nd cousins</td>
<td>1/64</td>
</tr>
<tr>
<td>3rd cousins</td>
<td>1/256</td>
</tr>
</tbody>
</table>
Homozygosity mapping

• Fundamental assumptions:
  – Recessive disorder
  – Disease alleles are IBD
    (not compound heterozygous)

• Implies:
  – Disease locus lies in autozygous region

Better name: Autozygosity mapping
Homozygosity mapping

Fundamental idea:

Reduce the genomic haystack by restricting the search to homozygous regions

Alternative workflows:

**Exomes only**
- Sequence all patient exomes
- Filter!
- Find shared homozygous variants
- Only those in homozygous regions
- Validate in parents/unaffected sibs

**SNP-array + 1 exome**
- Genotype all patients with dense SNP chip
- Shared homozygous regions (PLINK)
- Sequence 1 patient exome
- Restrict to identified regions + filter!

Linkage analysis!!

Challenge: When is homozygous = autozygous
In healthy, Norwegian genomes: Homozygous regions of 5 Mb are not unusual (+ many smaller!)
Example:
Homozygosity mapping + exome sequencing

- Example from our department
- Family from Afghanistan
- Severe mental retardation
Example:
Homozygosity mapping + exome sequencing

### Strategy:
- Sequence one or more.
- Look for variants that are:
  - homozygous in 3 and 5
  - heterozygous or not present in 1, 2 and 4

<table>
<thead>
<tr>
<th>Filters</th>
<th>3</th>
<th>3, 1, 2, 4</th>
<th>3, 5</th>
<th>3, 5 + one</th>
<th>3, 5, 1, 2, 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>All homozygous</td>
<td>4677</td>
<td>698</td>
<td>3656</td>
<td>*</td>
<td>199</td>
</tr>
<tr>
<td>Nonsyn/ indel/splice</td>
<td>2704</td>
<td>384</td>
<td>2105</td>
<td>*</td>
<td>113</td>
</tr>
<tr>
<td>&lt; 1% in 1000G</td>
<td>92</td>
<td>22</td>
<td>55</td>
<td>*</td>
<td>2</td>
</tr>
<tr>
<td>134 control exomes</td>
<td>29</td>
<td>19</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Guidelines for homozygosity mapping

• Fundamental assumption: Consanguinity
  – distant relationship is good
  – 1st cousins: Ok, but may give long regions

• Chance of success increases if
  – more than 1 affected
  – at least one control in the family

• The validation problem with rare syndromes in small families
  – How to deal with a final gene list of 20 genes? 10 genes? 5 genes?
Overview

• Disease mapping by high throughput sequencing

• Homozygosity mapping

• More general family based sequencing
More general family based sequencing

choose individuals

analyze data
use linkage info

exclude incompatible regions

Find variants in remaining regions
dbSNP + +

Causal variant
Summary

• High throughput sequencing: A new era in disease mapping

• Especially effective with recessive disorders in consanguineous pedigrees (homozygosity mapping)

• Family based sequencing
  – Linkage analysis is crucial (even if LOD < 3)
The end
Example: Genetic diagnosis

Genetic diagnosis by whole exome capture and massively parallel DNA sequencing

Murim Choia, Ute I. Schollb, Weizhen Jia, Tiewen Liua, Irina R. Tikhonovab, Paul Zumbob, Ahmet Nayirc, Ayşin Bakkaloğlüd, Seza Özend, Sami Sanjad, Carol Nelson-Williams, Anita Farhi, Shrikant Mane, and Richard P. Lifton.1

Department of Genetics, Howard Hughes Medical Institute, bKeck Foundation for Biotechnology Resources, Yale University School of Medicine, New Haven, CT 06510; cDepartment of Pediatric Nephrology, Istanbul Medical Faculty, Istanbul 34390, Turkey; dDepartment of Pediatric Nephrology and Rheumatology, Hacettepe University Faculty of Medicine, Ankara 06100, Turkey; and eAmerican University of Beirut, Beirut 11072020, Lebanon

Contributed by Richard P. Lifton, September 17, 2009 (sent for review September 8, 2009)
Example: Genetic diagnosis

- **Homozygous regions:**
  - 26 segments $> 1$ Mb
    - 2-57 Mb
  - Total length = 462 Mb

- **5.3 Mb of exome homozygous**
  - 2495 genes
  - 1493 homozygous cSNPs
  - 29 unknown NS/SS/I
  - 10 in conserved regions
  - 1 "strongly stood out": *SLC26A3*

Theoretical $F = 0.078$

$F = 15\%$ (too high!)
Example: Genetic diagnosis

- *SLC26A3*
- Known gene for **Congenital chloride diarrhea (CCD)**
- CCD consistent with phenotype

Diagnosis = CCD

theoretical F = 0.078