Strategies for finding disease genes.

Aaron Quinlan
EGAG
3-Feb-2016
quinlanlab.org | aaronquinlan@gmail.com
Mendel’s second law
"independent assortment of traits"

We are lucky that Mendel happened to select two traits that were not physically linked!
chromosomal theory of inheritance: that genes are located on chromosomes like beads on a string, and that some genes are linked (meaning they are on the same chromosome and always inherited together)
Sturtevant realized that if the frequency of crossing over was related to distance, one could use this information to map out the genes on a chromosome. After all, the farther apart two genes were on a chromosome, the more likely it was that these genes would separate during recombination.

Therefore, as Sturtevant explained it, the "proportion of crossovers could be used as an index of the distance between any two factors" (Sturtevant, 1913). Collecting a stack of laboratory data, Sturtevant went home and spent most of the night drawing the first chromosomal linkage map for the genes located on the X chromosome of fruit flies (Weiner, 1999).

Sturtevant then worked out the order and the linear distances between these linked genes, thus forming a linkage map. In doing so, he computed the distance in an arbitrary unit he called the "map unit," which represented a recombination frequency of 0.01, or 1%. Later, the map unit was renamed the centimorgan (cM), in honor of Thomas Hunt Morgan.
Studying disease in families

**Autosomal dominant**
- Affected father
- Unaffected mother
- Unaffected daughter
- Affected son

**Autosomal recessive**
- Carrier father
- Carrier mother
- Unaffected son
- Carrier daughter
- Carrier son
- Affected daughter

*U.S. National Library of Medicine*

*Examples:*
- Huntington’s disease
- Cystic fibrosis
Studying disease in families

X-linked recessive, carrier mother

X-linked recessive, affected father

U.S. National Library of Medicine
Linkage mapping: use linkage to track down disease genes in families

Figure 1. Three-generation pedigree segregating an autosomal dominant trait. Alleles at 3 marker loci designated A, B, and C are shown. Squares indicate males; circles, females; open symbols, normal phenotype; and solid symbols, disease phenotype.
Linkage mapping: use linkage to track down disease genes in families

Marker B perfectly co-segregates (linked) with disease

Figure 1. Three-generation pedigree segregating an autosomal dominant trait. Alleles at 3 marker loci designated A, B, and C are shown. Squares indicate males; circles, females; open symbols, normal phenotype; and solid symbols, disease phenotype.
Linkage mapping: use linkage to track down disease genes in families

Marker B perfectly co-segregates (linked) with disease
Marker A partially (1 recomb.) co-segregates with disease

**Figure 1.** Three-generation pedigree segregating an autosomal dominant trait. Alleles at 3 marker loci designated A, B, and C are shown. Squares indicate males; circles, females; open symbols, normal phenotype; and solid symbols, disease phenotype.
Linkage mapping: use linkage to track down disease genes in families

Marker B perfectly co-segregates (linked) with disease
Marker A partially (1 recomb.) co-segregates with disease
Marker C is unlinked (multiple recomb.) with disease

Figure 1. Three-generation pedigree segregating an autosomal dominant trait. Alleles at 3 marker loci designated A, B, and C are shown. Squares indicate males; circles, females; open symbols, normal phenotype; and solid symbols, disease phenotype.
Linkage mapping successes

BRCA1, ~1994
M.C. King, Ray White, others

BRCA2, ~1994


Abstract
A small proportion of breast cancer, in particular those cases arising at a young age, is due to the inheritance of dominant susceptibility genes conferring a high risk of the disease. A genomic linkage search was performed with 15 high-risk breast cancer families that were unlinked to the BRCA1 locus on chromosome 17q21. This analysis localized a second breast cancer susceptibility locus, BRCA2, to a 6-centimorgan interval on chromosome 13q12-13. Preliminary evidence suggests that BRCA2 confers a high risk of breast cancer but, unlike BRCA1, does not confer a substantially elevated risk of ovarian cancer.

Genetic Linkage Map of Six Polymorphic DNA Markers around the Gene for Familial Adenomatous Polyposis on Chromosome 5

APC, ~1990
Ray White, C. C. Bird
# Linkage mapping strengths and weaknesses

<table>
<thead>
<tr>
<th>Property of mapping approach</th>
<th>Linkage analysis</th>
<th>Association analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data type studied</td>
<td>Relatives</td>
<td>Unrelated or related individuals</td>
</tr>
<tr>
<td>Relevant parameter</td>
<td>Recombination fraction</td>
<td>Association statistic</td>
</tr>
<tr>
<td>Range of effect detected</td>
<td>Long (≤5 Mb)</td>
<td>Short (≤100 kb)</td>
</tr>
<tr>
<td>(linkage or association)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of markers required for</td>
<td>Moderate (500–1,000)</td>
<td>Large (&gt;100,000)</td>
</tr>
<tr>
<td>genome-wide coverage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statistics used</td>
<td>Cumbersome (requires tailor-made likelihood methods)</td>
<td>Elegant; can use the range of classical statistical tools</td>
</tr>
<tr>
<td>Dealing with correlated markers</td>
<td>Pose problems in presence of ungenotyped individuals</td>
<td>Can be handled efficiently</td>
</tr>
<tr>
<td>Biological basis of approach</td>
<td>Observe (or infer) recombination in pedigree data</td>
<td>Exploit unobserved recombination events in past generations</td>
</tr>
<tr>
<td>Dealing with allelic heterogeneity</td>
<td>Not a problem</td>
<td>Reduces power</td>
</tr>
<tr>
<td>Detecting genotyping errors</td>
<td>Potentially detected as Mendelian inconsistencies</td>
<td>Potentially detected only in family data, but not in case–control data</td>
</tr>
<tr>
<td>Most suitable application</td>
<td>Rare, dominant traits</td>
<td>Common traits</td>
</tr>
</tbody>
</table>
Association Mapping

Common disease / common variant hypothesis
Genome-wide association studies

Cases (have disease)

Controls (no disease)

The advent of high-throughput DNA sequencing
Exome sequencing: sequence the protein-coding portion (2%) of the genome
Exome sequencing: sequence the protein-coding portion (2%) of the genome.
Exome sequencing to solve Mendelian diseases

Exome sequencing identifies the cause of a mendelian disorder

Sarah B Ng1,9, Kati J Buckingham2,9, Choll Lee1, Abigail W Bigham2, Holly K Tabor2,3, Karin M Dent1, Chad D Huff3, Paul T Shannon4, Ethyllin Wang Jabs3,8, Deborah A Nickerson1, Jay Shendure1 & Michael J Bamshad1,3,9

Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome

Sarah B Ng1,9, Abigail W Bigham2, Kati J Buckingham2, Mark C Hannibal2,3, Margaret J McMillin2, Heidi I Gildersleeve2, Anita E Becc3,3, Holly K Tabor2,3, Gregory M Cooper1, Heather C Mefford2, Choll Lee1, Emily H Turner1, Joshua D Smith1, Mark J Rieder1, Koh-ichiro Yoshiura1, Naomichi Mataumoto5, Tohru Ohta6, Norio Niikawa6, Deborah A Nickerson1, Michael J Bamshad1,3,9 & Jay Shendure1

Now also the Broad Institute (MacArthur)
Identifying rare variation in the human genome and exome

1000 Genomes
A Deep Catalog of Human Genetic Variation

~2500 WGS samples

NHLBI Grand Opportunity Exome Sequencing Project (ESP)

~6500 WES samples

ExAC Browser (Beta) | Exome Aggregation Consortium

~65000 WES samples.
Case study: exome sequencing of a familial disease
Case study: exome sequencing of a familial disease
Case study: exome sequencing of a familial disease

Step 1: annotated functional consequence

**synonymous (silent)**

<table>
<thead>
<tr>
<th>Normal</th>
<th>Mutated</th>
</tr>
</thead>
<tbody>
<tr>
<td>ctg cag act</td>
<td>ctg caa act</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Normal</th>
<th>Mutated</th>
</tr>
</thead>
<tbody>
<tr>
<td>L Q T</td>
<td>L Q T</td>
</tr>
</tbody>
</table>

**non-synonymous (missense)**

<table>
<thead>
<tr>
<th>Normal</th>
<th>Mutated</th>
</tr>
</thead>
<tbody>
<tr>
<td>ctg cag act</td>
<td>ctg cgg act</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Normal</th>
<th>Mutated</th>
</tr>
</thead>
<tbody>
<tr>
<td>L Q T</td>
<td>L R T</td>
</tr>
</tbody>
</table>

**stop-gain (nonsense)**

<table>
<thead>
<tr>
<th>Normal</th>
<th>Mutated</th>
</tr>
</thead>
<tbody>
<tr>
<td>ctg cag act</td>
<td>ctg tag act</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Normal</th>
<th>Mutated</th>
</tr>
</thead>
<tbody>
<tr>
<td>L Q T</td>
<td>L STOP T</td>
</tr>
</tbody>
</table>

**stop-loss**

<table>
<thead>
<tr>
<th>Normal</th>
<th>Mutated</th>
</tr>
</thead>
<tbody>
<tr>
<td>ctg cag act</td>
<td>ctg cag act</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Normal</th>
<th>Mutated</th>
</tr>
</thead>
<tbody>
<tr>
<td>L Q T</td>
<td>L Q T</td>
</tr>
</tbody>
</table>

Impact sometimes hard to predict.
Case study: exome sequencing of a familial disease

Variant Effect Predictor

The VEP determines the effect of your variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions. Simply input the coordinates of your variants and the nucleotide changes to find out the:

- genes and transcripts affected by the variants
- location of the variants (e.g. upstream of a transcript, in coding sequence, in non-coding RNA, in regulatory regions)
- consequence of your variants on the protein sequence (e.g. stop gained, missense, stop lost, frameshift)
- known variants that match yours, and associated minor allele frequencies from the 1000 Genomes Project
- SIFT and PolyPhen scores for changes to protein sequence
- ... And more!

SnpEff

Genetic variant annotation and effect prediction toolbox.

Download Snpeff

Important: This version implements the new VCF annotation standard 'ANN' field.

Latest version 4.2 (2015-12-05)
Requires Java 1.7

ANNOVAR Documentation

ANNOVAR is an efficient software tool to utilize update-to-date information to functionally annotate genetic variants detected from diverse genomes (including human genome hg18, hg19, hg38, as well as mouse, worm, fly, yeast and many others). Given a list of variants with chromosome, start position, end position, reference nucleotide and observed nucleotides, ANNOVAR can perform:
Loss of function mutations

- Functional gene
- Nonsense SNP
- Frame-shift indel
- Splice site SNP
- Exon deletion
- Whole gene deletion

Interesting things are more likely to be wrong b/c they are rare

Many tools + many transcript annotations = many answers
Annotating software matters

### Table 2 Same transcripts, different software: ANNOVAR and VEP annotations for exonic variants

<table>
<thead>
<tr>
<th></th>
<th>ANV+VEP</th>
<th>ANV</th>
<th>VEP</th>
<th>Exact match</th>
<th>Category match</th>
<th>ANV match rate (%)</th>
<th>VEP match rate (%)</th>
<th>Overall category match rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOF total</td>
<td>104,915</td>
<td>77,527</td>
<td>96,761</td>
<td>68,284</td>
<td>69,373</td>
<td>88.08</td>
<td>70.57</td>
<td>66.12</td>
</tr>
<tr>
<td>Frameshift</td>
<td>19,021</td>
<td>15,822</td>
<td>16,685</td>
<td>13,486</td>
<td>-</td>
<td>85.24</td>
<td>80.83</td>
<td>-</td>
</tr>
<tr>
<td>Stop gained</td>
<td>16,758</td>
<td>14,960</td>
<td>16,146</td>
<td>14,348</td>
<td>-</td>
<td>95.91</td>
<td>88.86</td>
<td>-</td>
</tr>
<tr>
<td>Stop lost</td>
<td>1,113</td>
<td>906</td>
<td>1,077</td>
<td>870</td>
<td>-</td>
<td>96.03</td>
<td>80.78</td>
<td>-</td>
</tr>
<tr>
<td>All splicing</td>
<td>69,112</td>
<td>45,839</td>
<td>62,853</td>
<td>39,580</td>
<td>-</td>
<td>86.35</td>
<td>62.97</td>
<td>-</td>
</tr>
<tr>
<td>MISSENSE total</td>
<td>350,806</td>
<td>324,242</td>
<td>347,752</td>
<td>318,056</td>
<td>321,188</td>
<td>98.09</td>
<td>91.46</td>
<td>91.56</td>
</tr>
<tr>
<td>Inframe indel</td>
<td>9,455</td>
<td>8,650</td>
<td>6,600</td>
<td>5,795</td>
<td>-</td>
<td>66.99</td>
<td>87.80</td>
<td>-</td>
</tr>
<tr>
<td>Missense</td>
<td>343,284</td>
<td>315,592</td>
<td>339,953</td>
<td>312,261</td>
<td>-</td>
<td>98.94</td>
<td>91.85</td>
<td>-</td>
</tr>
<tr>
<td>Initiator codon</td>
<td>1,199</td>
<td>0</td>
<td>1,199</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>0.00</td>
<td>-</td>
</tr>
<tr>
<td>SYNONYMOUS and OTHER CODING total</td>
<td>182,120</td>
<td>172,463</td>
<td>175,483</td>
<td>165,643</td>
<td>165,826</td>
<td>96.05</td>
<td>94.39</td>
<td>91.05</td>
</tr>
<tr>
<td>Synonymous</td>
<td>181,873</td>
<td>172,463</td>
<td>175,053</td>
<td>165,643</td>
<td>-</td>
<td>96.05</td>
<td>94.62</td>
<td>-</td>
</tr>
<tr>
<td>Stop retained</td>
<td>203</td>
<td>0</td>
<td>203</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>0.00</td>
<td>-</td>
</tr>
<tr>
<td>Other coding</td>
<td>227</td>
<td>0</td>
<td>227</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>0.00</td>
<td>-</td>
</tr>
<tr>
<td>ALL LOF</td>
<td>104,915</td>
<td>77,527</td>
<td>96,761</td>
<td>68,284</td>
<td>69,373</td>
<td>88.08</td>
<td>70.57</td>
<td>66.12</td>
</tr>
<tr>
<td>ALL LOF and MISSENSE</td>
<td>455,721</td>
<td>401,769</td>
<td>444,513</td>
<td>386,340</td>
<td>390,561</td>
<td>96.16</td>
<td>86.91</td>
<td>85.70</td>
</tr>
<tr>
<td>ALL EXONIC</td>
<td>637,841</td>
<td>574,232</td>
<td>619,996</td>
<td>551,983</td>
<td>556,387</td>
<td>96.13</td>
<td>89.03</td>
<td>87.23</td>
</tr>
</tbody>
</table>

McCarthy et al. Genome Medicine 2014
Example of annotation complexity

Insertion of a single A. What is the impact?

a. frameshift
b. stopgain
c. synonymous
d. yes
An allele underlying a rare disease should be rare!

- Always filter by frequency separately in every available population
  - do NOT filter for frequency in only one population
  - do NOT filter on average worldwide frequency
- If variant causes severe phenotype, should ALWAYS be rare in every population
Rare disease discovery with GEMINI (Genome Mining)

Goal: make rare disease research as simple and reproducible as possible

Code:  github.com/arq5x/gemini
Docs:   gemini.readthedocs.org
GEMINI integrates variants, annotations, relationships and genotypes into a simple database.

Annotation every variant in VCF with information from (subset):
- dbSNP
- UCSC
- OMIM
- KEGG
- Ensembl
- ESP
- 1000G
- CADD
- Polyphen
- SIFT
- ENCODE
- HPRD
- GERP
- FitCons
- VISTA

Variants, annotations, phenotypes & genotypes together in a database

Prioritize genetic variants in various disease contexts based on genome annotations, sample genotypes, and sample relationships.

Powerful ad hoc queries

```
$ gemini query --query
"SELECT chrom, start, end, ref,
alt, gene, impact, aaf
FROM variants
WHERE in_dbsnp = 0
and aaf < 0.01
and cadd_scaled_score >= 30
and is_loss_of_function = 1
and my_disease_regions = 1
and gt_types.mom == HET
and gt_types.dad == HET
and gt_types.child == HOM_ALT"
```

Family studies

Tumor/Normal

Cohort studies
- w/ disease
- w/o disease
 Compound heterozygotes

$ gemini comp_hets
   --columns "chrom, start, end, gene, impact"
   --min-kindreds 2
   --max-priority 1
   --filter "impact_severity != 'LOW' and max_aaf_all < 0.001"
disease.db

<table>
<thead>
<tr>
<th>chr</th>
<th>start</th>
<th>end</th>
<th>gene</th>
<th>isoform</th>
<th>impact</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
<th>T6</th>
<th>T7</th>
<th>T8</th>
<th>T9</th>
<th>T10</th>
<th>T11</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr17</td>
<td>78081692</td>
<td>78081693</td>
<td>GAA</td>
<td>non_syn_coding</td>
<td>4.33</td>
<td>14401 family1 1805;unaffected,1847;unaffected4805;affected T/T,T/C,T</td>
<td>C</td>
<td>4805</td>
<td>1</td>
<td>16_14401_14406</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr17</td>
<td>78084748</td>
<td>78084749</td>
<td>GAA</td>
<td>non_syn_coding</td>
<td>5.51</td>
<td>14406 family1 1805;unaffected,1847;unaffected4805;affected G/A,G/G,A</td>
<td>G</td>
<td>4805</td>
<td>1</td>
<td>16_14401_14406</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr2</td>
<td>129025757</td>
<td>129025758</td>
<td>HS6ST1</td>
<td>non_syn_coding</td>
<td>2.67</td>
<td>3657 family1 1805;unaffected,1847;unaffected4805;affected C/A,C/C,A</td>
<td>C</td>
<td>4805</td>
<td>1</td>
<td>45_3657_3660</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr2</td>
<td>129075743</td>
<td>129075744</td>
<td>HS6ST1</td>
<td>non_syn_coding</td>
<td>3.25</td>
<td>3660 family1 1805;unaffected,1847;unaffected4805;affected G/G,G/C,G</td>
<td>C</td>
<td>4805</td>
<td>1</td>
<td>45_3657_3660</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr22</td>
<td>45255643</td>
<td>45255644</td>
<td>PRR5-ARHGAP8</td>
<td>non_syn_coding</td>
<td>1.88</td>
<td>16838 family1 1805;unaffected,1847;unaffected4805;affected G/G,G/C,G</td>
<td>C</td>
<td>4805</td>
<td>1</td>
<td>169_16838_16839</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr22</td>
<td>45255687</td>
<td>45255688</td>
<td>PRR5-ARHGAP8</td>
<td>non_syn_coding</td>
<td>0.37</td>
<td>16839 family1 1805;unaffected,1847;unaffected4805;affected C/T,C/C,T</td>
<td>T</td>
<td>4805</td>
<td>1</td>
<td>169_16838_16839</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr15</td>
<td>71548994</td>
<td>71548995</td>
<td>THSD4</td>
<td>non_syn_coding</td>
<td>3.68</td>
<td>8777 family1 1805;unaffected,1847;unaffected4805;affected G/A,G/G,A</td>
<td>G</td>
<td>4805</td>
<td>1</td>
<td>181_8777_8780</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr15</td>
<td>71952898</td>
<td>71952899</td>
<td>THSD4</td>
<td>non_syn_coding</td>
<td>4.52</td>
<td>8780 family1 1805;unaffected,1847;unaffected4805;affected A/G,A/G,A</td>
<td>G</td>
<td>4805</td>
<td>1</td>
<td>181_8777_8780</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
GEMINI is popular for rare disease research.

Homozygous mutation of MTPAP causes cellular radiosensitivity and persistent DNA double-strand breaks

Homzygous mutation of MTPAP causes cellular radiosensitivity and persistent DNA double-strand breaks.

Germline Mutations in MAP3K6 Are Associated with Familial Gastric Cancer

Daniel Gaston, Samantha Hansford, Carla Oliveira, Mathew Nightingale, Hugo Pinheiro, Christine Macgillivray, Pardeep Kaurah, Andrea L. Rideout, Patricia Steele, Gabriela Soares, Wee-Yuarn Huang, Scott Whitehouse, Sarah Blowers, Marissa A. LeBlanc, Haiyan Jiang, Wenda Greer, Mark E. Samuels, Andrew Orr, Conrad V. Fernandez, Jacek Majewski, Mark Ludman, Sarah Dyack, Lynette S. Penney, Christopher R. McMaster, David Huntsman, Karen Bedard

New splicing mutation in the choline kinase beta (CHKB) gene causing a muscular dystrophy detected by whole-exome sequencing

Jorge Oliveira, Luís Negrão, Isabel Fineza, Ricardo Talpa, Manuel Melo-Pires, Ana Maria Fortuna, Ana Rita Gonçalves, Hugo Froufe, Conceição Egas, Rosário Santos and Mário Sousa

High-sensitivity sequencing reveals multi-organ somatic mosaicism causing DICER1 syndrome


Mutations in NOTCH1 Cause Adams-Oliver Syndrome

Anna-Barbara Stützlich, Anna Lehman, Dale L. Bodian, Justin Ashworth, Zheyuan Zong, Hong Li, Patricia Lam, Alina Khromykh, Ramaswamy K. Iyer, Joseph G. Vockley, Rajiv Baveja, Emilka Maria Santos Silva, Joanne Dixon, Eby L. Leon, Benjamin D. Solomon, Gustavo Giusman, John E. Niederhuber, Jared C. Roach, and Millan S. Patel

Diagnosis of an imprinted-gene syndrome by a novel bioinformatics analysis of whole-genome sequences from a family trio


1 Inova Translational Medicine Institute, Inova Health System, Falls Church, Virginia
2 Department of Pediatric Endocrinology, Inova Children’s Hospital, Falls Church, Virginia
3 Fairfax Neonatal Associates PC, Inova Children’s Hospital, Falls Church, Virginia
Rare disease genetics in Utah

Family genetics

Clinical collaborators

Utah Genome Project
<table>
<thead>
<tr>
<th>Studies</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>201</td>
<td>Congenital Heart Disease</td>
</tr>
<tr>
<td>96</td>
<td>Suicide</td>
</tr>
<tr>
<td>96</td>
<td>Longevity</td>
</tr>
<tr>
<td>80</td>
<td>Amyotrophic Lateral Sclerosis</td>
</tr>
<tr>
<td>78</td>
<td>Congenital Diaphragmatic Hernia</td>
</tr>
<tr>
<td>72</td>
<td>Autism</td>
</tr>
<tr>
<td>45</td>
<td>Early Infantile Epileptic Encephalopathy</td>
</tr>
<tr>
<td>40</td>
<td>Preterm Birth</td>
</tr>
<tr>
<td>37</td>
<td>Hereditary Hemorrhagic Telangiectasia</td>
</tr>
<tr>
<td>30</td>
<td>Tuberous Sclerosis</td>
</tr>
<tr>
<td>30</td>
<td>Chiari Malformation</td>
</tr>
<tr>
<td>10</td>
<td>Primary Ovarian Insufficiency</td>
</tr>
<tr>
<td>9</td>
<td>Familial Metatarsophalangeal Joint Osteoarthritis</td>
</tr>
<tr>
<td>9</td>
<td>Ataxia</td>
</tr>
<tr>
<td>8</td>
<td>Idiopathic Hypogonadotrophic Hypogonadism</td>
</tr>
<tr>
<td>6</td>
<td>Brown Fat Metabolism</td>
</tr>
<tr>
<td>3</td>
<td>Mitochondrial Pyruvate Insufficiency</td>
</tr>
</tbody>
</table>
Summary

- Two basic and complementary approaches: family based and case-control based.

- Modern DNA sequencing has opened the flood gates for discovery.

- Our molecular and computation tools for disease genetics research have advanced dramatically in recent years. However, much of the low hanging fruit has been picked. Many of the unsolved rare diseases exhibit incomplete penetrance, phenotypic heterogeneity and germline and somatic mosaicism. These will be tough nuts to crack.