Analysis of Human Genome Variation (HGV)

Stefan Gräf <sg550@cam.ac.uk>

Computational Genomic and Medicine

Division of Respiratory Medicine

TrainMalta, 15th / 16th September 2016
Course Tutors / Acknowledgements

Marta Bleda
Computational Biologist
PostDoc

Stefan Gräf
Computational Biologist
Senior Research Associate

Matthias Haimel
Software Engineer
PhD student
Analysis of Human Genome Variation (HGV)

Sequence Alignment

Variant Calling

Variant Annotation and Filtering

Variant Prioritisation

Variant Association

Number of variants

- 3.5M
- 100s
Analysis of DNA Sequence Variation

- Introduction
  - Next-generation sequencing
  - Human genetics
- Identification of genetic variation
- Experimental Design
- Analysis pipeline overview
Next-Generation Sequencing

Definition

Non-Sanger-based high-throughput DNA sequencing technologies. Millions or billions of DNA strands can be sequenced in parallel, yielding substantially more throughput and minimising the need for the fragment-cloning methods that are often used in Sanger sequencing of genomes.
1953

Structure of the DNA
Evolution of Sequencing

DNA Sequencing

First Generation
Single sequence
500-1,000bp

Sanger Sequencing
Sanger Sequencing

PCR in presence of fluorescent, chain-terminating nucleotides

Fragments run through gel electrophoresis

Fluorescent fragments detected by laser and represented on a chromatogram
More than 60 Years of Genome Research

1953: Structure of the DNA

2001: Reference genome

Sanger Sequencing
Evolution of Sequencing

DNA Sequencing

First Generation
Single sequence
500-1,000bp

Next-Generation
Gbp of sequence
25-500bp

Sanger Sequencing

Illumina (Solexa)

Roche 454

ABI SOLiD

Ion Torrent

Clonally amplified
Illumina (Solexa)

Illumina Nextera library preparation, paired-end sequencing and analysis

https://youtu.be/womKfikWlxM
Sanger vs. Next-Generation Sequencing
More than 60 Years of Genome Research

1953: Structure of the DNA
1953: Sanger Sequencing

2001: Reference genome
2001: High-throughput Next-generation Sequencing

2007: Personal genomes
2007: Decoding the Blueprint
A flow cell contains **8 lanes**
- Each lane is subdivided into **100 image tiles**
- Per cycle **4 images** (A, G, T, C) are taken
- 8 lanes x 100 tiles x 4 bases x 50 cyles => **160,000 images**
- An image has a size of 7.3 MB => **1.2 TB per run**
- The imaging takes up most of the time
- **HiSeq X**: dual flow cell, 2x150bp reads, 5.3-6 billion reads pass QC, run takes less than 3 days, >75% of bases are above Q30
- Never write images to disc …
Throughput Growth Over 10 Years

Mardis, Nature (2011)
Cost of Sequencing

Cost per Raw Megabase of DNA Sequence

Moore's Law

NIH National Human Genome Research Institute
genome.gov/sequencingcosts
Cost of Sequencing

Cost per Genome

Moore's Law

NIH National Human Genome Research Institute

genome.gov/sequencingcosts
<table>
<thead>
<tr>
<th>Personal Genome</th>
<th>Platform</th>
<th>Genomic template libraries</th>
<th>No. of reads (millions)</th>
<th>Read length (bases)</th>
<th>Base coverage (fold)</th>
<th>Assembly</th>
<th>Genome coverage (%)</th>
<th>SNVs in millions (alignment tool)</th>
<th>No. of runs</th>
<th>Estimated cost (US$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>J. Craig Venter</td>
<td>Automated Sanger</td>
<td>MP' from BACs, fosmids &amp; plasmids</td>
<td>31.9</td>
<td>800</td>
<td>7.5</td>
<td>De novo</td>
<td>N/A</td>
<td>3.21</td>
<td>&gt;340,000</td>
<td>70,000,000</td>
</tr>
<tr>
<td>James D. Watson</td>
<td>Roche/454</td>
<td>Frag: 500 bp</td>
<td>93.2</td>
<td>250</td>
<td>7.4</td>
<td>Aligned*</td>
<td>95</td>
<td>3.32 (Blat)</td>
<td>734</td>
<td>1,000,000*</td>
</tr>
<tr>
<td>Yoruban male (NA18507)</td>
<td>Illumina/Solexa</td>
<td>93% MP: 200 bp</td>
<td>3,410</td>
<td>35</td>
<td>40.6</td>
<td>Aligned*</td>
<td>99.9</td>
<td>3.83 (MAQ)</td>
<td>40</td>
<td>250,000*</td>
</tr>
<tr>
<td>Han Chinese male</td>
<td>Illumina/Solexa</td>
<td>66% Frag: 150–250 bp</td>
<td>1,921</td>
<td>35</td>
<td>36</td>
<td>Aligned*</td>
<td>99.9</td>
<td>3.07 (SOAP)</td>
<td>35</td>
<td>500,000*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3% MP: 135 bp &amp; 440 bp</td>
<td>1,029</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Korean male (AK1)</td>
<td>Illumina/Solexa</td>
<td>21% Frag: 130 bp &amp; 440 bp</td>
<td>393</td>
<td>36</td>
<td>27.8</td>
<td>Aligned*</td>
<td>98.8</td>
<td>3.45 (GSNAP)</td>
<td>30</td>
<td>200,000*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>79% MP: 130 bp, 390 bp &amp; 2.7 kb</td>
<td>1,156</td>
<td>36, 88,</td>
<td>106</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Korean male (SJ)</td>
<td>Illumina/Solexa</td>
<td>MP: 100 bp, 200 bp &amp; 6300 bp</td>
<td>1,64</td>
<td>35, 74</td>
<td>29.0</td>
<td>Aligned*</td>
<td>99.9</td>
<td>3.44 (MACQ)</td>
<td>15</td>
<td>250,000**</td>
</tr>
<tr>
<td>Yoruban male (NA18507)</td>
<td>Life/AGP</td>
<td>9% Frag: 100–500 bp</td>
<td>211</td>
<td>50</td>
<td>17.9</td>
<td>Aligned*</td>
<td>98.6</td>
<td>3.87 (Corona-lite)</td>
<td>9.5</td>
<td>60,000***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>91% MP: 600–3,500 bp</td>
<td>2,075</td>
<td></td>
<td>25, 50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stephen R. Quake</td>
<td>Helicos Biosciences</td>
<td>Frag: 100–500 bp</td>
<td>2,725</td>
<td>32</td>
<td>28</td>
<td>Aligned*</td>
<td>90</td>
<td>2.81 (Index-OP)</td>
<td>4</td>
<td>48,000*</td>
</tr>
<tr>
<td>AML female</td>
<td></td>
<td>Frag: 150–200 bp</td>
<td>2,730</td>
<td>32</td>
<td>32.7</td>
<td>Aligned*</td>
<td>91</td>
<td>3.81 (MACQ)</td>
<td>98</td>
<td>1,600,000***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Frag: 150–200 bp</td>
<td>1,081</td>
<td>35</td>
<td>13.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AML male</td>
<td></td>
<td>MP: 200–250 bp</td>
<td>1,620</td>
<td>35</td>
<td>23.3</td>
<td>Aligned*</td>
<td>98.5</td>
<td>3.46 (MACQ)</td>
<td>16.5</td>
<td>500,000**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MP: 200–250 bp</td>
<td>1,351</td>
<td>50</td>
<td>21.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>James R. Lupski</td>
<td>Life/AGP</td>
<td>16% Frag: 100–500 bp</td>
<td>238</td>
<td>35</td>
<td>29.6</td>
<td>Aligned*</td>
<td>98.8</td>
<td>3.42 (Corona-lite)</td>
<td>3</td>
<td>75,000**</td>
</tr>
<tr>
<td>CMT male</td>
<td></td>
<td>84% MP: 600–3,500 bp</td>
<td>1,271</td>
<td></td>
<td>25, 50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*A minimum of one read aligning to the National Center for Biotechnology Information build 36 reference genome. **A total of reads from aligned assemblies.

Next-Generation Sequencing Helps Interrogating Many Omic Features of a Cell

The ENCODE Project Consortium (adapted)
<table>
<thead>
<tr>
<th>Method</th>
<th>Sequencing to determine:</th>
<th>Subway route as defined in next figure</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA-Seq</td>
<td>A genome sequence</td>
<td>Comparison, ‘anatomic’ (isolation by anatomic site), flow cytometry, DNA extraction, mechanical shearing, adaptor ligation, PCR and sequencing</td>
</tr>
<tr>
<td>Targeted DNA-Seq</td>
<td>A subset of a genome (for example, an exome)</td>
<td>Comparison, cell culture, DNA extraction, mechanical shearing, adaptor ligation, PCR, hybridization capture, PCR and sequencing</td>
</tr>
<tr>
<td>Methyl-Seq</td>
<td>Sites of DNA methylation, genome-wide</td>
<td>Perturbation, genetic manipulation, cell culture, DNA extraction, mechanical shearing, adaptor ligation, bisulfite conversion, PCR and sequencing</td>
</tr>
<tr>
<td>Targeted methyl-Seq</td>
<td>DNA methylation in a subset of the genome</td>
<td>Comparison, cell culture, DNA extraction, bisulfite conversion, molecular inversion probe capture, circularization, PCR and sequencing</td>
</tr>
<tr>
<td>DNase-Seq, Sono-Seq and FAIRE-Seq</td>
<td>Active regulatory chromatin (that is, nucleosome-depleted)</td>
<td>Perturbation, cell culture, nucleus extraction, DNase I digestion, DNA extraction, adaptor ligation, PCR and sequencing</td>
</tr>
<tr>
<td>MAINE-Seq</td>
<td>Histone-bound DNA (nucleosome)</td>
<td>Comparison, cell culture, MNase I digestion, DNA extraction, adaptor ligation, PCR and sequencing</td>
</tr>
<tr>
<td>ChIP-Seq</td>
<td>Protein-DNA interactions (using chromatin immunoprecipitation)</td>
<td>Comparison, 'anatomic', cell culture, cross-linking, mechanical shearing, immunoprecipitation, DNA extraction, adaptor ligation, PCR and sequencing</td>
</tr>
<tr>
<td>RIP-Seq, CLIP-Seq, HITS-CLIP</td>
<td>Protein-RNA interactions</td>
<td>Variation, cross-linking, 'anatomic', RNase digestion, immunoprecipitation, RNA extraction, adaptor ligation, reverse transcription, PCR and sequencing</td>
</tr>
<tr>
<td>RNA-Seq</td>
<td>RNA (that is, the transcriptome)</td>
<td>Comparison, 'anatomic', RNA extraction, poly(A) selection, chemical fragmentation, reverse transcription, second-strand synthesis, adaptor ligation, PCR and sequencing</td>
</tr>
<tr>
<td>FRT-Seq</td>
<td>Amplification-free, strand-specific transcriptome sequencing</td>
<td>Comparison, 'anatomic', RNA extraction, poly(A) selection, chemical fragmentation, adaptor ligation, reverse transcription and sequencing</td>
</tr>
<tr>
<td>NET-Seq</td>
<td>Nascent transcription</td>
<td>Perturbation, genetic manipulation, cell culture, immunoprecipitation, RNA extraction, adaptor ligation, reverse transcription, circularization, PCR and sequencing</td>
</tr>
<tr>
<td>Hi-C</td>
<td>Three-dimensional genome structure</td>
<td>Comparison, cell culture, cross-linking, proximity ligation, mechanical shearing, affinity purification, adaptor ligation</td>
</tr>
<tr>
<td>Chia-PET</td>
<td>Long-range interactions mediated by a protein</td>
<td>Perturbation, cell culture, cross-linking, mechanical shearing, immunoprecipitation, proximity ligation, affinity purification, adaptor ligation, PCR and sequencing</td>
</tr>
<tr>
<td>Ribo-Seq</td>
<td>Ribosome-protected mRNA fragments (that is, active translation)</td>
<td>Comparison, cell culture, RNase digestion, ribosome purification, RNA extraction, adaptor ligation, reverse transcription, rRNA depletion, circularization, PCR and sequencing</td>
</tr>
<tr>
<td>TRAP</td>
<td>Genetically targeted purification of polysomal mRNAs</td>
<td>Comparison, genetic manipulation, ‘anatomic’, cross-linking, affinity purification, RNA extraction, poly(A) selection, reverse transcription, second-strand synthesis, adaptor ligation, PCR and sequencing</td>
</tr>
<tr>
<td>PARS</td>
<td>Parallel analysis of RNA structure</td>
<td>Comparison, cell culture, RNA extraction, poly(A) selection, RNase digestion, chemical fragmentation, adaptor ligation, reverse transcription, PCR and sequencing</td>
</tr>
<tr>
<td>Synthetic saturation mutagenesis</td>
<td>Functional consequences of genetic variation</td>
<td>Variation, genetic manipulation, barcoding, RNA extraction, reverse transcription, PCR and sequencing</td>
</tr>
<tr>
<td>Immuno-Seq</td>
<td>The B-cell and T-cell repertoires</td>
<td>Perturbation, ‘anatomic’, DNA extraction, PCR and sequencing</td>
</tr>
<tr>
<td>Deep protein mutagenesis</td>
<td>Protein binding activity of synthetic peptide libraries or variants</td>
<td>Variation, genetic manipulation, phage display, in vitro competitive binding, DNA extraction, PCR and sequencing</td>
</tr>
<tr>
<td>PhiT-Seq</td>
<td>Relative fitness of cells containing disruptive insertions in diverse genes</td>
<td>Variation, genetic manipulation, cell culture, competitive growth, linear amplification, adaptor ligation, PCR and sequencing</td>
</tr>
</tbody>
</table>

Shendure & Aiden (2012), adapted
Subway Map of Core Techniques

Shendure & Aiden (2012)
Evolution of Sequencing

DNA Sequencing

First Generation
Single sequence
500-1,000bp

Sanger Sequencing

Next-Generation
Gbp of sequence
25-500bp

Illumina (Solexa)
Roche 454
ABI SOLiD
Ion Torrent

Pacific Biosciences
Helicos HeliScope
Oxford Nanopore
base4

Clonally amplified

Single molecule
Oxford Nanopore Technology

Nanopore DNA sequencing:
https://vimeo.com/127689053?from=outro-embed
Analysis of Human Genome Variation (HGV)

- Sequence Alignment
- Variant Calling
- Variant Annotation and Filtering
- Variant Prioritisation
- Variant Association

Number of variants: 3.5M

Number of variants: 100s
Human Genome Variation
More than 60 Years of Genome Research

- **1953**: Structure of the DNA
- **2001**: Reference genome
- **2007**: Personal genomes
- **2010**: Personal genomes x1000+

**Sanger Sequencing**

**High-throughput Next-generation Sequencing**
Human Genome Variation

1000 Genomes Project (Nature, 1 Nov 2012)
• Aims to understand the genetic contribution to disease
• 1092 individuals from 14 populations
• Low-coverage whole-exome and whole-genome sequencing
• Validated haplotype map of
  • 38 million single nucleotide polymorphisms
  • 1.4 million short insertions and deletions
  • more than 14,000 larger deletions

http://www.1000genomes.org
Rare Genetic Variants in Health and Disease

• Better understand link between low-frequency and rare genetic changes and human disease caused by harmful changes to the proteins the body makes.

• Study the genetic code of 10,000 people in much finer detail than ever before.
  • 4,000 whole genomes of deeply phenotyped cohorts (i.e. TwinsUK and ASLPAC) at 6x depth
  • 6,000 whole exomes of extreme phenotypes of specific conditions
• Provide a sequence variation resource for future studies

http://www.uk10k.org
New initiative will sequence 10,000 whole genomes of people with rare genetic diseases

Project will lay foundation for genomic medicine.

University of Cambridge, Genomics England and Illumina, Inc. today announced the start of a three-year project that will sequence 10,000 whole genomes of children and adults with rare genetic diseases. The project represents a plot for Genomics England, which will provide 2,000 samples, and marks the beginning of the national endeavor to sequence 100,000 genomes in the UK National Health Service (NHS), announced recently by the Prime Minister, David Cameron.

“This project will bring enormous improvements to the care of patients with rare genetic diseases. It will shorten the gap between the first signs of ill-health in a person and providing a conclusive diagnosis by using the power of modern DNA sequencing methods,” said Dr John Bradley,
Genomics England, with the consent of participants and the support of the public, is creating a lasting legacy for patients, the NHS and the UK economy through the sequencing of 100,000 genomes: the 100,000 Genomes Project.

Genomics England was set up by the Department of Health to deliver the 100,000 Genomes Project. Initially the focus will be on rare disease, cancer and infectious disease.

Read more...

http://www.genomicsengland.co.uk/
A Roadmap of Sequencing Science

Shendure & Aiden (2012)
Human Genetics
Human Genetics

Diagram showing the process of fertilization involving the fusion of a sperm and an egg to form a diploid cell with 46 chromosomes.

- Father's sperm contains haploid chromosomes (23 chromosomes).
- Mother's egg contains haploid chromosomes (23 chromosomes).
- Fertilization results in a diploid egg with 46 chromosomes.

http://www.genome.gov/glossary/
**Human Genetics**

![Diagram showing chromosome diagrams with alleles A and a, illustrating homoyzgous AA, heterozygous Aa, and homoyzgous aa states.](http://www.genome.gov/glossary/)
Huntington's Disease

Genotypes
- DD Homozygous
- Dd and dD Heterozygous
- dd Homozygous

Phenotypes
- Affected Dominant
- Unaffected Recessive

http://www.genome.gov/glossary/
Sickle Cell Anemia or Cystic Fibrosis

Alleles
- D
- d

Genotypes
- DD Homozygous
- Dd Heterozygous
- dd Homozygous

Phenotypes
- DD Unaffected
- Dd dD Unaffected
- dd Affected

Recessive

http://www.genome.gov/glossary/
Human Genetics

- Two sets of chromosomes (diploid), one from each parent
- Two alternative copies (alleles) of each gene
- Alleles can be
  - identical (homozygous) or
  - dissimilar (heterozygous)
- Only one allele (dominant), or both alleles (recessive) need to be mutated to be causative
- Genetic configuration (genotype) varies amongst individuals and populations
- Results in a observable trait (phenotype)
Identification of Genetic Variation

GWASs
Linkage and sequencing
Exome
Genome

Cooper et al., 2011
Genetic Variants have Different Functional Consequences

Cooper et al., 2011
Exome-sequencing Interrogates the Protein-coding Portion of the Genome

Bamshad et al., 2011
Whole Exome vs. Whole Genome

Regions covered by WES (64 MB, 2%)

Regions covered by WGS (~3000 MB, 98%)
Consequences

Raised demands for resources
• Storage (talking peta (1015) bytes)
• Computation

Data security
Sample requirements
Teasing out Disease-causing Variants

Long list of candidate variants

Comparative Genomics

Protein Structure / Biochemistry

Experimental Assay

Cooper et al., 2011
## Assessing Deleteriousness

<table>
<thead>
<tr>
<th>Name</th>
<th>Category</th>
<th>Score used for analysis</th>
<th>Deleterious threshold</th>
<th>Information used</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT</td>
<td>Function prediction</td>
<td>1 − Score Score</td>
<td>&gt;0.95</td>
<td>Protein sequence conservation among homologs</td>
</tr>
<tr>
<td>PolyPhen-2</td>
<td>Function prediction</td>
<td>Score</td>
<td>&gt;0.5</td>
<td>Eight protein sequence features, three protein structure features</td>
</tr>
<tr>
<td>LRT</td>
<td>Function prediction</td>
<td>Score * 0.5 (if Omega ≥1) or 1 − Score * 0.5 (if Omega &lt;1)</td>
<td>P</td>
<td>DNA sequence evolutionary model</td>
</tr>
<tr>
<td>MutationTaster</td>
<td>Function prediction</td>
<td>Score (if A or D) or 1 − Score (if N or P)</td>
<td>&gt;0.5</td>
<td>DNA sequence conservation, splice site prediction, mRNA stability prediction and protein feature annotations</td>
</tr>
<tr>
<td>Mutation Assessor</td>
<td>Function prediction</td>
<td>(Score-Min)/(Max − Min)</td>
<td>&gt;0.65</td>
<td>Sequence homology of protein families and sub-families within and between species</td>
</tr>
<tr>
<td>FATHMM</td>
<td>Function prediction</td>
<td>1 − (Score-Min)/(Max − Min)</td>
<td>≥0.45</td>
<td>Sequence homology</td>
</tr>
<tr>
<td>GERP++ RS</td>
<td>Conservation score</td>
<td>Score</td>
<td>&gt;4.4</td>
<td>DNA sequence conservation</td>
</tr>
<tr>
<td>PhyloP</td>
<td>Conservation score</td>
<td>Score</td>
<td>&gt;1.6</td>
<td>DNA sequence conservation</td>
</tr>
<tr>
<td>SiPhy</td>
<td>Conservation score</td>
<td>Score</td>
<td>&gt;12.17</td>
<td>Inferred nucleotide substitution pattern per site</td>
</tr>
<tr>
<td>PON-P</td>
<td>Ensemble score</td>
<td>Score</td>
<td>P</td>
<td>Random forest methodology-based pipeline integrating five predictors</td>
</tr>
<tr>
<td>PANTHER</td>
<td>Function prediction</td>
<td>Score</td>
<td>P</td>
<td>Phylogenetic trees based on protein sequences</td>
</tr>
<tr>
<td>PhD-SNP</td>
<td>Function prediction</td>
<td>Score</td>
<td>P</td>
<td>SVM-based method using protein sequence and profile information</td>
</tr>
<tr>
<td>SNAP</td>
<td>Function prediction</td>
<td>Score</td>
<td>P</td>
<td>Neural network-based method using DNA sequence information as well as functional and structural annotations</td>
</tr>
<tr>
<td>SNPs&amp;GO</td>
<td>Function prediction</td>
<td>Score</td>
<td>P</td>
<td>SVM-based method using information from protein sequence, protein sequence profile and protein function</td>
</tr>
<tr>
<td>MutPred</td>
<td>Function prediction</td>
<td>Score</td>
<td>&gt;0.5</td>
<td>Protein sequence-based model using SIFT and a gain/loss of 14 different structural and functional properties</td>
</tr>
<tr>
<td>KGGSeq</td>
<td>Ensemble score</td>
<td>Score</td>
<td>P</td>
<td>Filtration and prioritization framework using information from three levels: genetic level, variant-gene level and knowledge level</td>
</tr>
<tr>
<td>CONDEL</td>
<td>Ensemble score</td>
<td>Score</td>
<td>&gt;0.49</td>
<td>Weighted average of the normalized scores of five methods</td>
</tr>
<tr>
<td>CADD</td>
<td>Ensemble score</td>
<td>Score</td>
<td>&gt;15</td>
<td>63 distinct variant annotation retrieved from Ensembl Variant Effect Predictor (VEP), data from the ENCODE project and information from UCSC genome browser tracks</td>
</tr>
</tbody>
</table>

*Dong et al., 2015*
Experimental Design
Sir Ronald A. Fisher

“To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.”

(1890 – 1962)

Evolutionary biologist, geneticist and statistician
Andrew Lang

“An unsophisticated forecaster uses statistics as a drunken man uses lamp-posts - for support rather than for illumination.”

(1844 — 1912)

Writer (poet, novelist), literary critic and anthropologist
Variant Discovery Strategies and Sample Selection

- Select study design to achieve adequate statistical power (i.e. trios for de novo mutations, pedigree analysis, cohort of multiple unrelated patients)

- Focus on cases with extreme outcome

- Population stratification important for rare variant detection
Genetic and Phenotypic Heterogeneity Reduces Power

Number of samples to achieve 80% power

<table>
<thead>
<tr>
<th>% Carriers</th>
<th>100</th>
<th>50</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recessive</td>
<td>4</td>
<td>9</td>
<td>170</td>
</tr>
<tr>
<td>Dominant</td>
<td>6</td>
<td>20</td>
<td>1100</td>
</tr>
</tbody>
</table>

http://exomepower.ssg.uab.edu
Phenotype-based Clustering Can Restore Power

(Deep) Phenotyping

Clustering

Electronic health records and other clinical information (i.e. demographics, laboratory tests, human phenotype ontology (HPO), imaging, etc.)
Analysis Pipeline Overview
GATK’s Best Practises

http://www.broadinstitute.org/gatk/
Analysis Pipeline Tasks and Tools

- Read mapping
  - BWA
  - Novoalign
  - ... (more tools)

- Alignment post-processing
  - Picard
  - SAMtools
  - ... (more tools)

- Variant calling
  - GATK
  - VCFtools
  - ... (more tools)

- Variant annotation and filtering
  - SnpEff/Sift
  - Ve!p
  - ... (more tools)
Analysis Pipeline Tasks and Tools

- Read mapping
- Alignment post-processing
- Variant calling
- Variant annotation and filtering

There is NO gold standard!

Tools:
- BWA
- Novoalign
- Picard
- SAMtools
- GATK
- VCFtools
- Ve!p
- SnpEff/Sift
NGS-Course Analysis Pipeline

Alignment of reads (bwa) -> Mark duplicates (Picard) -> Local Indel Realignment (samtools/GATK)

Variant calling (samtools/GATK) -> Base Quality Score Recalibration (samtools/GATK) -> Variant annotation and prioritisation (Ve!p) -> Variant association
Data Formats

File extensions:

- `.fa` — reference sequence (fasta), i.e. GRCh37_chr19.fa
- `.fastq` — raw sequencing reads, i.e. NA12878_1.fq.gz
- `.sam` — aligned sequencing reads, i.e. NA12878.sam
- `.bam` — aligned reads (binary), i.e. NA12891.bam
- `.vcf` — called variants, i.e. trio_mpileup.vcf
- `.tbi` — files indexed with tabix
- `.gz` — compressed files
Offspring trio of central European ancestry

NGS-Course Data

NA12891

NA12878

NA12892
Variant Annotation, Filtering and Prioritisation
Variant Annotation and Effect Prediction

(a) genomic DNA → DNA fragments → sequencer → sequenced fragments

(b) Variant Call Format (VCF) file

- OMIM
- KEGG Pathways
- UniProt
- Ensembl
- Mouse Genome Informatics
- COSMIC
- Human Protein Atlas
- Human Gene Mutation Database
- MitoCarta
- RefGene

Remote data sources → local database → Variant Effect Predictor → rich annotations

Mapping → Alignment → Genotyping → Variant calling

Yourshaw et al., Brief Bioinform (2015)
Exome Aggregation Consortium (ExAC)

- Aggregation of high-quality exome (protein-coding region) sequence data for **60,706 individuals** of diverse ethnicities
- Resolution of **one variant every eight bases** of coding sequence
- Allows calculation of objective metrics of pathogenicity for sequence variants
- Can be used for **efficient filtering** of candidate disease-causing variants

Contributing projects
- 1000 Genomes
- Bulgarian Trios
- Finland-United States Investigation of NIDDM Genetics (FUSION)
- GoT2D
- Inflammatory Bowel Disease
- METabolic Syndrome In Men (METSIM)
- Jackson Heart Study
- Myocardial Infarction Genetics Consortium:
  - Italian Atherosclerosis, Thrombosis, and Vascular Biology Working Group
  - Ottawa Genomics Heart Study
  - Pakistan Risk of Myocardial Infarction Study (PROMIS)
  - Precocious Coronary Artery Disease Study (PROCARDIS)
- Registre Gironi del COR (REGICOR)
- NHLBI-Go Exome Sequencing Project (ESP), incl. 96 PAH cases
- National Institute of Mental Health (NIMH) Controls
- SIGMA-T2D
- Sequencing in Suomi (SISu)
- Swedish Schizophrenia & Bipolar Studies
- T2D-GENES
- Schizophrenia Trios from Taiwan
- The Cancer Genome Atlas (TCGA)
- Tourette Syndrome Association International Consortium for Genomics (TSAICG)

http://biorxiv.org/content/early/2015/10/30/030338
http://exac.broadinstitute.org
Exome Aggregation Consortium (ExAC)
Exome Aggregation Consortium (ExAC)
Variant Annotation and Consequence Prediction

• Deleteriousness scores
  
  • **SIFT**: functional prediction, protein sequence conservation among homologs; score: 1 (tolerated) - 0 (deleterious)
  
  • **PolyPhen**: functional prediction, protein sequence and structure features; score: 0 (benign) - 1 (damaging)
  
  • **CADD**: ensemble score, combines 63 distinct variant annotation features retrieved from Ensembl VEP, Encode, UCSC genome browser; Phred score (i.e. 30 = 99.9% accurate or 1 in 1000 is incorrect)

• DNA sequence conservation scores
  
  • **GERP**: maximum likelihood evolutionary rate estimation, predicts sites under evolutionary constraints
  
  • **PhyloP**: base-wise conservation score derived from Multiz alignment of 100 vertebrate species
  
  • **PhastCons**: evolutionary conserved elements derived from Multiz alignment of 100 vertebrate species (phylogenetic hidden Markov model)

Yourshaw et al., Brief Bioinform (2015), adapted
Variant Annotation and Consequence Prediction

- **Deleteriousness scores**
  - **SIFT:** functional prediction, protein sequence conservation among homologs; score: 1 (tolerated) - 0 (deleterious)
  - **PolyPhen:** functional prediction, protein sequence and structure features; score: 0 (benign) - 1 (damaging)
  - **CADD:** ensemble score, combines 63 distinct variant annotation features retrieved from Ensembl VEP, Encode, UCSC genome browser; Phred score (i.e. 30 = 99.9% accurate or 1 in 1000 is incorrect)

- **DNA sequence conservation scores**
  - **GERP:** maximum likelihood evolutionary rate estimation, predicts sites under evolutionary constraints
  - **PhyloP:** base-wise conservation score derived from Multiz alignment of 100 vertebrate species
  - **PhastCons:** evolutionary conserved elements derived from Multiz alignment of 100 vertebrate species (phylogenetic hidden Markov model)

Yourshaw et al., Brief Bioinform (2015), adapted

proteins sequence and structure based prediction

score based on various informative genome-wide annotations

measures DNA sequence conservation
Variant Annotation Tools

- Ensembl Variant Effect Predictor (VEP)
- SnpEff / SnpSift
- AnnoVar

- Rich annotation of DNA sequencing variants by leveraging the Ensembl Variant Effect Predictor with plugins (Yourshaw et al., 2015)
- Choice of transcripts and software has a large effect on variant annotation (McCarthy et al., 2014)
Ensembl Variant Effect Predictor (VEP)
Ensembl Variant Effect Predictor (VEP)
Including External Resources

- Custom annotation
  — http://www.ensembl.org/info/docs/tools/vep/script/vep_custom.html

- VEP plugins
  — https://github.com/ensembl-variation/VEP_plugins

- Examples
  — http://www.ensembl.org/info/docs/tools/vep/script/vep_example.html
External Resources

- 1000 Genome Project
  — http://www.1000genomes.org/

- Exome Aggregation Consortium (ExAC) Database
  — http://exac.broadinstitute.org/

- dbNSFP
  — https://sites.google.com/site/jpopgen/dbNSFP
## Assessing Deleteriousness

<table>
<thead>
<tr>
<th>Name</th>
<th>Category</th>
<th>Score used for analysis</th>
<th>Deleterious threshold</th>
<th>Information used</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT</td>
<td>Function prediction</td>
<td>1 – Score Score</td>
<td>&gt;0.95</td>
<td>Protein sequence conservation among homologs</td>
</tr>
<tr>
<td>PolyPhen-2</td>
<td>Function prediction</td>
<td>Score</td>
<td>&gt;0.5</td>
<td>Eight protein sequence features, three protein structure features</td>
</tr>
<tr>
<td>LRT</td>
<td>Function prediction</td>
<td>Score * 0.5 (if Omega ≥1) or 1 – Score * 0.5 (if Omega &lt;1)</td>
<td>P</td>
<td>DNA sequence evolutionary model</td>
</tr>
<tr>
<td>MutationTaster</td>
<td>Function prediction</td>
<td>Score (if A or D) or 1 – Score (if N or P)</td>
<td>&gt;0.5</td>
<td>DNA sequence conservation, splice site prediction, mRNA stability prediction and protein feature annotations</td>
</tr>
<tr>
<td>Mutation Assessor</td>
<td>Function prediction</td>
<td>(Score-Min)/(Max – Min)</td>
<td>&gt;0.65</td>
<td>Sequence homology of protein families and sub-families within and between species</td>
</tr>
<tr>
<td>FATHMM</td>
<td>Function prediction</td>
<td>1 – (Score-Min)/(Max – Min)</td>
<td>≥0.45</td>
<td>Sequence homology</td>
</tr>
<tr>
<td>GERP++ RS</td>
<td>Conservation score</td>
<td>Score</td>
<td>&gt;4.4</td>
<td>DNA sequence conservation</td>
</tr>
<tr>
<td>PhyloP</td>
<td>Conservation score</td>
<td>Score</td>
<td>&gt;1.6</td>
<td>DNA sequence conservation</td>
</tr>
<tr>
<td>SiPhy</td>
<td>Conservation score</td>
<td>Score</td>
<td>&gt;12.17</td>
<td>Inferred nucleotide substitution pattern per site</td>
</tr>
<tr>
<td>PON-P</td>
<td>Ensemble score</td>
<td>Score</td>
<td>P</td>
<td>Random forest methodology-based pipeline integrating five predictors</td>
</tr>
<tr>
<td>PANTHER</td>
<td>Function prediction</td>
<td>Score</td>
<td>P</td>
<td>Phylogenetic trees based on protein sequences</td>
</tr>
<tr>
<td>PhD-SNP</td>
<td>Function prediction</td>
<td>Score</td>
<td>P</td>
<td>SVM-based method using protein sequence and profile information</td>
</tr>
<tr>
<td>SNAP</td>
<td>Function prediction</td>
<td>Score</td>
<td>P</td>
<td>Neural network-based method using DNA sequence information as well as functional and structural annotations</td>
</tr>
<tr>
<td>SNPs&amp;GO</td>
<td>Function prediction</td>
<td>Score</td>
<td>P</td>
<td>SVM-based method using information from protein sequence, protein sequence profile and protein function</td>
</tr>
<tr>
<td>MutPred</td>
<td>Function prediction</td>
<td>Score</td>
<td>&gt;0.5</td>
<td>Protein sequence-based model using SIFT and a gain/loss of 14 different structural and functional properties</td>
</tr>
<tr>
<td>KGGSeq</td>
<td>Ensemble score</td>
<td>Score</td>
<td>P</td>
<td>Filtration and prioritization framework using information from three levels: genetic level, variant-gene level and knowledge level</td>
</tr>
<tr>
<td>CONDEL</td>
<td>Ensemble score</td>
<td>Score</td>
<td>&gt;0.49</td>
<td>Weighted average of the normalized scores of five methods</td>
</tr>
<tr>
<td>CADD</td>
<td>Ensemble score</td>
<td>Score</td>
<td>&gt;15</td>
<td>63 distinct variant annotation retrieved from Ensembl Variant Effect Predictor (VEP), data from the ENCODE project and information from UCSC genome browser tracks</td>
</tr>
</tbody>
</table>

Dong et al., 2015
Phred Quality Scores

- Assess/measure accuracy of base calling
- Defined as a property related to the **base calling error probabilities** ($P$):
  \[ Q = -10 \log_{10}(P) \]
- Reaching Q30, virtually all bases in a read are called correctly:

<table>
<thead>
<tr>
<th>Phred Quality Score</th>
<th>Probability of Incorrect Base Call</th>
<th>Base Call Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1 in 10</td>
<td>90%</td>
</tr>
<tr>
<td>20</td>
<td>1 in 100</td>
<td>99%</td>
</tr>
<tr>
<td>30</td>
<td>1 in 1,000</td>
<td>99.9%</td>
</tr>
<tr>
<td>40</td>
<td>1 in 10,000</td>
<td>99.99%</td>
</tr>
<tr>
<td>50</td>
<td>1 in 100,000</td>
<td>99.999%</td>
</tr>
</tbody>
</table>
Exploring the Raw Data (m: Mapping Quality)
Exploring the Raw Data (r: read names)
Exploring the Raw Data (b: Base Quality)
Exploring the Raw Data (m: Mapping Quality)
Exploring the Raw Data (n: Nucleotides Coloured)
Exploring the Raw Data (.: Dot View)
Transition / Transversion

SNV Transition/Transversion (Ts/Tv) Ratio

Plate

Ratio

Count

1e6

0

A C A G A T A C G T A C G

2 3 4 5 6 7

2.075 2.080 2.085 2.090

## Variant Call Format (VCF)

```plaintext
#fileformat=VCFv4.2
#fileDate=20090805
#source=myImputationProgramV3.1
#reference=file:////seq/references/1000GenomesPilot-NCBI36.fasta
#contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxon...>
#phasing=partial
#INFO=<ID:NS,Number=1,Type=Integer,Description="Number of Samples With Data">  
#INFO=<ID:DP,Number=1,Type=Integer,Description="Total Depth">  
#INFO=<ID:AF,Number=A,Type=Float,Description="Allele Frequency">  
#INFO=<ID:AA,Number=1,Type=String,Description="Ancestral Allele">  
#INFO=<ID:DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">  
#INFO=<ID:H2,Number=0,Type=Flag,Description="HapMap2 membership">  
#FILTER=<ID=q10,Description="Quality below 10">  
#FILTER=<ID=s50,Description="Less than 50% of samples have data">  
#FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">  
#FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">  
#FORMAT=<ID:DP,Number=1,Type=Integer,Description="Read Depth">  
#FORMAT=<ID:HQ,Number=2,Type=Integer,Description="Haplotype Quality">  

## CHROM POS ID REF ALT QUAL FILTER INFO

<table>
<thead>
<tr>
<th>CHROM</th>
<th>POS</th>
<th>REF</th>
<th>ALT</th>
<th>QUAL</th>
<th>FILTER</th>
<th>INFO</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>14370</td>
<td>rs6054257</td>
<td>G</td>
<td>A</td>
<td>PASS</td>
<td>NS=3;DP=14;AF=0.5;DB;H2</td>
</tr>
<tr>
<td>20</td>
<td>17330</td>
<td>.</td>
<td>T</td>
<td>3</td>
<td>q10</td>
<td>NS=3;DP=11;AF=0.017</td>
</tr>
<tr>
<td>20</td>
<td>1110696</td>
<td>rs6040355</td>
<td>A</td>
<td>G,T</td>
<td>PASS</td>
<td>NS=2;DP=10;AF=0.333,0.667;AA=T;DB</td>
</tr>
<tr>
<td>20</td>
<td>1230237</td>
<td>.</td>
<td>T</td>
<td>47</td>
<td>PASS</td>
<td>NS=3;DP=13;AA=T</td>
</tr>
<tr>
<td>20</td>
<td>1234567</td>
<td>microsat1</td>
<td>GTC</td>
<td>G,GTCT</td>
<td>PASS</td>
<td>NS=3;DP=9;AA=G</td>
</tr>
</tbody>
</table>

GT:GQ:DP:HQ 0|0:48.1:51.51 1|48.|8:51.51 1|1:43.|5:3:...;DB;H2
GT:GQ:DP:HQ 0|0:49.3:58.50 0|1:35.6:65.3 0|0:41.|3
GT:GQ:DP:HQ 0|0:54.7:56.60 0|0:48:4:51.51 0|0:61.|2
GT:GQ:DP 0|1:35:4 0|0:2:17:2 1|4:0:3

https://samtools.github.io/hts-specs/VCFv4.2.pdf
```