Finding and Calling Genome Variants
Outline

• Genome variants overview
• Mining variants from databases
  ❖ dbSNP
  ❖ HapMap
  ❖ 1000 Genomes
  ❖ Disease/Clinical variants databases
• Calling variants using your own data
  ❖ Samtools (mpileup/bcftools)
Genomic Variation

• Population genetics
  - Measure/explain diversity

• Disease susceptibility
  - GWAS
  - Biomarkers

• Variants may cause a particular trait
  - Regulatory element (e.g. promoter, enhancer, etc.)
  - Protein coding sequence (e.g. silent, missense, or nonsense mutation)

Palstra, RJ. et al (2012)

http://evolution.berkeley.edu/evolibrary/article/mutations_06
Genomic Variation: Sequence vs Structural Variation

### Sequence Variants

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Example (Reference / Alternative)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP</td>
<td>Single Nucleotide Polymorphism</td>
<td>Ref: ...TTGACGTA... Alt: ...TTGGCGTA...</td>
</tr>
<tr>
<td>Insertion</td>
<td>Insertion of one or several nucleotides</td>
<td>Ref: ...TTGACGTA... Alt: ...TTGATCGTA...</td>
</tr>
<tr>
<td>Deletion</td>
<td>Deletion of one or several nucleotides</td>
<td>Ref: ...TTGACGTA... Alt: ...TTGGTA...</td>
</tr>
<tr>
<td>Substitution</td>
<td>A sequence alteration where the length of the change in the variant is the same as that of the reference.</td>
<td>Ref: ...TTGACGTA... Alt: ...TTGTAGTA...</td>
</tr>
</tbody>
</table>

### Structural Variants (>50 bases or more)

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Example (Reference / Alternative)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNV</td>
<td>Copy Number Variation: increases or decreases the copy number of a given region</td>
<td>&quot;Gain&quot; of one copy: &quot;Loss&quot; of one copy:</td>
</tr>
<tr>
<td>Inversion</td>
<td>A continuous nucleotide sequence is inverted in the same position</td>
<td></td>
</tr>
<tr>
<td>Translocation</td>
<td>A region of nucleotide sequence that has translocated to a new position (eg. BCR-ABL fusion gene)</td>
<td></td>
</tr>
</tbody>
</table>
Genome Variation: Individual and Population

• SNP vs SNV (Single Nucleotide Variant)

• SNP: present in more than 1% of the population
  – MAF* > 1% common SNP
  – MAF* < 1% rare SNP
  – Some definitions use 5% as threshold

• On average one variant every 1200 bases (based on HapMap)

*Minor Allele Frequency
## Genome Variation: Reference

<table>
<thead>
<tr>
<th>Organism</th>
<th>Description/Strain</th>
<th>Assembly*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>DNA isolated from WBC of 4 anonymous individuals (2 males and 2 females). However, the majority of the sequence came from one of the male donors</td>
<td>GRCh37/GRCh38</td>
</tr>
<tr>
<td>Mouse</td>
<td>C57BL/6J</td>
<td>GRCm37/GRCm38</td>
</tr>
<tr>
<td><em>C. elegans</em></td>
<td>N2</td>
<td>WormBase v WS220</td>
</tr>
<tr>
<td>Fruit fly</td>
<td>ISO1</td>
<td>BDGP Release 5</td>
</tr>
<tr>
<td>Yeast</td>
<td>S288C</td>
<td>SGD Feb 2011</td>
</tr>
<tr>
<td><em>A. thaliana</em></td>
<td>Col ecotype</td>
<td>TAIR10</td>
</tr>
</tbody>
</table>

*Available in /nfs/genomes
Genome Variation: Distribution of Variants*

*Based on dbSNP 135 [http://massgenomics.org](http://massgenomics.org)

MNP: Multiple Nucleotide Polymorphism
VCF Format

- Variant Call Format (VCF); BCF ➔ binary version of VCF
- Text file format with meta-information and header lines, followed by data lines containing information about a position in the genome.

Example

```plaintext
##fileformat=VCFv4.0
##fileDate=20100607
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">  
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">  
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">  
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">  
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">  
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">  
##ALT=<ID=DEL,Description="Deletion">  
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">  
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

Example data:

```
1 1 1 . ACG . A,AT . PASS .
1 2 rs1 C T,CT . PASS H2;AA=T
1 5 . A . . .
1 100 T <DEL> . . .
```

**Reference alleles (GT=0)**

- `GT:DP` 1/2:13 0/0:29
- `GT:GQ` 0 1:100 2/2:70
- `GT:GQ` 1 0:77 1/1:99

**Alternate alleles (GT>0 is an index to the ALT column)**

- `GT:DP` 1/1:12:3 0/0:20

Phased data (G and C above are on the same chromosome)

- Large SV: Deletion, SNP
- Insertion
- Other event

www.1000genomes.org
Describing/Annotating Variants

• General guidelines*
  ➢ no position 0
  ➢ range indicated by “_” (eg. 586_591)

• DNA
  ➢ g.957A>T (to include chromosome use chr9:g.957A>T)
  ➢ g.413delG
  ➢ g.451_452insT
  ➢ In CDS,
    ➢ c.23G>C
    ➢ +1 is A of ATG (start codon); -1 is the previous/upstream nucleotide
    ➢ “*” is the stop codon (eg. *1 is the first nucleotide of the stop codon)

• RNA
  ➢ r.957a>u

• Protein (three/one letter aa)
  ➢ p.His78Gln

<table>
<thead>
<tr>
<th>Chr</th>
<th>Position</th>
<th>Ref</th>
<th>Alt</th>
<th>Source</th>
<th>g.change : rsID : Depth=AvgSampleReadDepth : FunctionGVS : hgvsProteinVariant</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>89824989</td>
<td>G</td>
<td>T</td>
<td>EVS</td>
<td>g.89824989G&gt;T : rs140823801 : Depth=141 : missense : p.Q993K</td>
</tr>
</tbody>
</table>

*For complete list/guidelines see hgvs.org
Genome Variation Databases: dbSNP

- Repository for SNPs and short sequence variation (<50 bases)
- Current build: dbSNP 143 (Mar 2015)
  - Approx. 88M validated rs#’s for human
    - Mostly germline mutations (smaller subset of somatic)
    - Contains rare variants as well
  - Various organisms
- Each SNP, or record, is identified by an rs# that includes
  - Summary attributes
  - NCBI resources (linked to ClinVar, GenBank, etc.)
  - External resources (linked to OMIM and NHGRI GWAS)
- Submissions are made from public laboratories and private organizations (ss#’s), and identical records are clustered into a single record (rs#’s).
- rs id is same for different assemblies (eg. GRCh37/38), but chromosomal coordinates may differ!
Hands-on: dbSNP

• Finding SNPs for your favorite gene in dbSNP and UCSC Genome Browser
Genome Variation Databases: HapMap Project

- Created to find distribution of SNPs and other genetic variants (GRCh37 coordinates)
- Phase I: 269 samples (2005)
- Phase II: 270 samples (2007)
- Phase III: 1115 samples (2008)
- Samples were from various populations
  - CHB (Han Chinese)
  - MEX (Mexican ancestry in Los Angeles, CA)
  - JPT (Japanese)
  - etc.
Genome Variation Databases:
1000 Genomes Project

• Extension of the HapMap in 2008 to catalogue genetic variation by sequencing at least 1000 participants
• Discover population level human genetic variations
• Initially consisted of whole genome low coverage (4X) and high coverage exome (20X) sequencing
• VCF file was developed, and initially maintained, for the project
## Mining Disease/Clinical Variants

<table>
<thead>
<tr>
<th>Database</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catalog of Published GWAS (NHGRI)</td>
<td>genome.gov/26525384</td>
</tr>
<tr>
<td>GWAS Central</td>
<td>gwascentral.org</td>
</tr>
<tr>
<td>ClinVar (NCBI)</td>
<td>ncbi.nlm.nih.gov/clinvar</td>
</tr>
<tr>
<td>PheGenI (NCBI)</td>
<td>ncbi.nlm.nih.gov/gap/phegeni</td>
</tr>
<tr>
<td>SNPedia</td>
<td>snpedia.com</td>
</tr>
</tbody>
</table>
Mining Disease/Clinical Variants in Cancer: cBioPortal

• Visualization and analysis of cancer genomics data sets

• Currently 89 data sets (including TCGA) consists of various studies

• Tools/Features:
  - OncoPrint: graphical summary of genomic alterations
  - Mutual Exclusivity
  - Correlation Plots
  - MutationMapper (“lollipop plots”)
Hands-on: cBioPortal

- Mining cancer variants in cBioPortal website
Mining Disease/Clinical Variants in Cancer: COSMIC

• Catalog of Somatic Mutations in Cancer (COSMIC) created in 2005
• v70 (Aug 2014) had ~2M coding point mutations
• Datasets are curated from published literature and other databases (eg. TCGA, ICGC)
• Available in both GRCh37/38 coordinates
• Tools/Features
  ➢ Cancer Gene Census (currently 572 genes)
  ➢ Browser: Cancer/Cell Line
  ➢ COSMIC Mart (similar to BioMart)
Calling Variants

• Align reads to reference* and call variants
• Popular tools include Samtools and GATK (from Broad)
• Germline vs Somatic mutations
• Samtools:
  ➢ Samtools’s mpileup (formerly pileup) computes genotype likelihoods supported by the aligned reads (BAM file) and stores in binary call format (BCF) file.
  ➢ Bcftools applies the priors (from above) and calls variants (SNPs and indels). Bcftools can be used to filter VCF files.

*reference-free variant calling software are available (eg. CORTEX)
Calling Variants: Samtools (cont.)

• Removes duplicate reads (e.g. from PCR)
• Both unique and multi-mapped reads are used for calling variants
• Recalibrates quality scores to take into account sequencing errors
Calling Variants: Workflow

1. QC Reads and Align
2. Evaluate Mapping
3. Call Variants (e.g., Samtools' mpileup/bcftools, GATK)
4. Evaluate & Filtering Variants
5. Annotate Variants (e.g., snpEff and VEP)
6. Downstream Analysis

Note: Use a sensitive (gapped) aligner eg. BWA
Calling Variants
Calling Variants:
Questionable Calls
Calling Variants: Evaluating

• Percent, or number of, reads containing variant vs reference
  ❖ View in a browser (eg. IGV)
• Base quality (eg. at least Q30)
• Mapping quality (depends on aligner)
• Coverage across region(s) of interest
  ➢ depth of coverage (eg. at least 5X)
  ➢ 20X considered high-coverage
• Strand bias
Calling Variants: Filtering/Querying

• vcf-annotate (from VCFtools*)
  ♦ vcf-annotate -f + myFile.vcf > myFile_annot.vcf
  ♦ “+” applies several filters with default values, eg.
    ➢ Strand bias: test if variant bases tend to come from one strand (Fisher’s Test)
    ➢ End distance bias: test if variant bases tend to occur at a fixed distance from the end of reads (t-test)

• Remove common SNPs (eg. use dbSNP)
  ❖ bedtools

• bcftools
  ❖ bcftools view -i 'DP>100' myFile.vcf
  ❖ bcftools view -i 'GT[0]=="1/1"' myFile.vcf

*vcftools.sourceforge.net
Hands-on: Samtools’ mpileup/bcftools

• Calling variants from an alignment (bam) file
  - samtools mpileup [options] [file1.bam file2.bam ...]
    samtools mpileup -d100000 -Buf
    /nfs/genomes/human_gp_feb_09_no_random/fasta/chr1.fa -o
    chr1_Subset_1000Genomes_CEU.bcf chr1_Subset_1000Genomes_CEU.bam

  - bcftools call [options] [file | STDIN]
    bcftools call -vmO v -o chr1_Subset_1000Genomes_CEU.vcf
    chr1_Subset_1000Genomes_CEU.bcf

For list of options see: https://htslib.org/doc/samtools-1.1.html and http://samtools.github.io/bcftools/bcftools.html
Calling Variants: Annotation

• Annotate variants with (functional) consequence
eg. chr12:g25232372A>G is a missense variant
• Popular tools include snpEff, and Variant Effect Predictor (VEP) from Ensembl
• Choice of annotation may affect variant annotation
  ➢ RefSeq
  ➢ Ensembl
  ➢ Gencode
Further/Downstream Analysis

Levo, M. and Segal, E. (2014)
BaRC SOP

• Variant calling using Samtools and GATK. Manipulating/interpreting VCF files
http://barcwiki/wiki/SOPs under

Variant calling and analysis
# Resources For Mining Variants

<table>
<thead>
<tr>
<th>Database</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 Genomes</td>
<td><a href="1000genomes.org">1000genomes.org</a></td>
</tr>
<tr>
<td>UK10K</td>
<td><a href="uk10k.org">uk10k.org</a></td>
</tr>
<tr>
<td>Exome Variant Server (EVS)</td>
<td><a href="evs.gs.washington.edu/EVS">evs.gs.washington.edu/EVS</a></td>
</tr>
<tr>
<td>Personal Genome Project (Harvard)</td>
<td><a href="personalgenomes.org">personalgenomes.org</a></td>
</tr>
<tr>
<td>ExAC Browser (Broad)</td>
<td><a href="exac.broadinstitute.org">exac.broadinstitute.org</a></td>
</tr>
</tbody>
</table>
# Resources For Mining Variants: Cancer

<table>
<thead>
<tr>
<th>Database</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>International Cancer Genome Consortium (ICGC)</td>
<td>icgc.org</td>
</tr>
<tr>
<td>Catalogue of Somatic Mutation in Cancer (COSMIC)</td>
<td>cancer.sanger.ac.uk</td>
</tr>
<tr>
<td>cBioPortal for Cancer Genomics</td>
<td>cbiportal.org</td>
</tr>
<tr>
<td>Cancer Cell Line Encyclopedia (CCLE)</td>
<td>broadinstitute.org/ccle</td>
</tr>
</tbody>
</table>
Resources For Mining Variants: Plants

• 1001 Genomes (A. thaliana 1001 strains)
  ➢ 1001genomes.org

• 1000 Genomes (large-scale gene sequencing of at least 1000 plant species)
  ➢ www.onekp.com