CS681: Advanced Topics in Computational Biology

Can Alkan
EA509
calkan@cs.bilkent.edu.tr

http://www.cs.bilkent.edu.tr/~calkan/teaching/cs681/
GENOMIC VARIATION: CHANGES IN DNA SEQUENCE
Human genome variation

- Genomic variation
  - Changes in DNA sequence
- Epigenetic variation
  - Methylation, histone modification, etc.
Human genetic variation

Types of genetic variants

- Single nucleotide changes
- Copy number variants (CNVs)
- Trisomy monosomy

Size of variant

- 1 bp
- 1 kb
- 1 Mb
- 1 chr

Frequency

Throughput

How do we assay them?

- SNP genotyping/Sanger sequencing
- Array-CGH
- Karyotyping
- High throughput sequencing

Size of variant

- 1 bp
- 1 kb
- 1 Mb
- 1 chr
Size range of genetic variation

- Single nucleotide (SNPs)
- Few to ~50bp (small indels, microsatellites)
- >50bp to several megabases (structural variants):
  - Deletions
  - Insertions
    - Novel sequence
    - Mobile elements (Alu, L1, SVA, etc.)
  - Segmental Duplications
    - Duplications of size ≥ 1 kbp and sequence similarity ≥ 90%
  - Inversions
  - Translocations
- Chromosomal changes
**Genetic variation**

If a mutation occurs in a codon:

- Synonymous mutations: Coded amino acid doesn’t change
- Nonsynonymous mutations: Coded amino acid changes

\[
\begin{align*}
\text{GTT} & \quad \rightarrow \quad \text{Valine} \\
\text{GTA} & \quad \rightarrow \quad \text{Valine} \\
\text{SYNONYMOUS} & \\
\text{GT} & \quad \rightarrow \quad \text{GT} \\
\text{G} & \quad \rightarrow \quad \text{G} \\
\text{Valine} & \quad \rightarrow \quad \text{Valine} \\
\text{GTT} & \quad \rightarrow \quad \text{Valine} \\
\text{GCA} & \quad \rightarrow \quad \text{Alanine} \\
\text{NONSYNONYMOUS} & 
\end{align*}
\]
Genetic variation

Where in the genome?

ALLELIC VARIATION

Person 1

Person 2

NONALLELIC (PARALOGOUS) VARIATION

Duplication (duplicons)

Where in the body?

Germ cells or gametes (sperm egg) -> Transmittable -> Germline Variation

Other (somatic cells) -> Not transmittable -> Somatic Variation
SNPs & indels

**SNP**: Single nucleotide polymorphism (substitutions)
**Short indel**: Insertions and deletions of sequence of length 1 to 50 basepairs

**reference**: C A C A G T G C G - T

**sample**: C A C C G T G - G C A T

- **Neutral**: no effect
- **Positive**: increases fitness (resistance to disease)
- **Negative**: causes disease
- **Nonsense mutation**: creates early stop codon
- **Missense mutation**: changes encoded protein
- **Frameshift**: shifts basepairs that changes codon order
Short tandem repeats

**reference:** C A G C A G C A G C A G C A G C A G

**sample:** C A G C A G C A G C A G C A G C A G

- Microsatellites (STR=short tandem repeats) 1-10 bp
  - Used in population genetics, paternity tests and forensics
- Minisatellites (VNTR=variable number of tandem repeats): 10-60 bp
- Other satellites
  - Alpha satellites: centromeric/pericentromeric, 171bp in humans
  - Beta satellites: centromeric (some), 68 bp in humans
  - Satellite I (25-68 bp), II (5bp), III (5 bp)
- Disease relevance:
  - Fragile X Syndrome
  - Huntington’s disease
Structural Variation

DELETION

Autism, mental retardation, Crohn’s

NOVEL SEQUENCE INSERTION

Haemophilia

MOBILE ELEMENT INSERTION

Alu/L1/SVA

TANDEM DUPLICATION

Schizophrenia, psoriasis

INTERSPERSED DUPLICATION

INVERSION

TRANSLOCATION

Chronic myelogenous leukemia
Chromosomal changes

- “Microscope-detectable”
- Disease causing or prevents birth
- Monosomy: 1 copy of a chromosome pair
- Uniparental disomy (UPD): Both copies of a pair comes from the same parent
- Trisomy: Extra copy of a chromosome
  - chr21 trisomy = Down syndrome
## Genetic variation among humans

### Single nucleotide variants in four human genomes

<table>
<thead>
<tr>
<th></th>
<th>(n)</th>
<th>In dbSNP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>J. Craig Venter’s genome</td>
<td>3,213,401</td>
<td>91.0</td>
</tr>
<tr>
<td>James D. Watson’s genome</td>
<td>3,322,093</td>
<td>81.7</td>
</tr>
<tr>
<td>Asian genome</td>
<td>3,074,097</td>
<td>86.4</td>
</tr>
<tr>
<td>Yoruban genome</td>
<td>4,139,196</td>
<td>73.6</td>
</tr>
</tbody>
</table>

### Structural variants in the Venter genome

<table>
<thead>
<tr>
<th></th>
<th>(n)</th>
<th>length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block substitutions</td>
<td>53,823</td>
<td>2–206</td>
</tr>
<tr>
<td>Indels (heterozygous)</td>
<td>851,575</td>
<td>1–82,711</td>
</tr>
<tr>
<td>Inversions</td>
<td>90</td>
<td>7–670,345</td>
</tr>
<tr>
<td>Copy number variants</td>
<td>62</td>
<td>8,855–1,925,949</td>
</tr>
</tbody>
</table>
Genetic variation are “shared”

Haplotype

- “Haploid Genotype”: a combination of alleles at multiple loci that are transmitted together on the same chromosome.
Haplotyping resolution

- Variation discovery methods do not directly tell which copy of a chromosome a variant is located.
- For heterozygous variants, it gets messy: 
  
  ![Diagram showing haplotypes]
  
  Chromosome 1, #1
  Chromosome 1, #2
  Discovered variants in Chromosome 1

Haplotyping resolution or haplotype phasing: finding which groups of variants “go together”
Discovery vs. genotyping

- Discovery: no *a priori* information on the variant
- Genotyping: test whether or not a “suspected” variant occurs
Variation discovery & genotyping

- Targeted methods:
  - SNP:
    - PCR
    - SNP microarray (genotyping)
  - Indel
    - PCR
    - “Indel microarray” (genotyping)
  - Structural variation
    - Quantitative PCR
    - Array Comparative Genomic Hybridization (array CGH)
    - Fluorescent in situ Hybridization (FISH) if variant > 500 kb
  - Chromosomal:
    - Microscope
Variation discovery & genotyping

- Targeted methods are:
  - Cheap(er), but limited:
    - Variants that are not in reference genome cannot be found
    - One experiment yields one type of variant
    - Not always genome-wide

- Alternative:
  - Whole genome resequencing
    - More expensive – getting cheaper
    - (Theoretically) comprehensive
    - Computational challenges
PROJECTS FOR GENOMIC VARIATION DISCOVERY
International HapMap Project

- Determine genotypes & haplotypes of 270 human individuals from 3 diverse populations:
  - Northern Americans (Utah / Mormons)
  - Africans (Yoruba from Nigeria)
  - Asians (Han Chinese and Japanese)
- 90 individuals from each population group, organized into parent-child *trios*.
- Each individual genotyped at ~5 million roughly evenly spaced markers (SNPs and small indels)

[http://www.hapmap.org](http://www.hapmap.org)
HapMap Project

Step 1: SNPs are identified in DNA samples from multiple individuals

Step 2: Adjacent SNPs that are inherited together are compiled into "haplotypes."

Step 3: "Tag" SNPs within haplotypes are identified that uniquely identify those haplotypes

By genotyping just the three tag SNPs shown above, one can identify which of the four haplotypes shown here are present in each individual.
Human Genome Diversity Panel

- More extensive set of genomic variation
- One aim is to build DNA resource libraries for large scale discovery & genotyping projects
- 1,050 human individuals from 52 populations

Initial HapMap and HGDP did not sequence the genomes of any samples.

ARTICLE

The Simons Genome Diversity Project: 300 genomes from 142 diverse populations

Mallick et al., 2016
Why?

- To understand “normal” human genomic variation
- To understand genetic transmission properties
- To understand *de novo* mutations
- To understand population structure, migration patterns
- To understand human disease
  - Find causal variants
  - Diagnose
  - Guide treatment
Human disease

- Rare variant common disease:
  - Most “complex” diseases, including neuropsychiatric diseases

- Common variant common disease
  - More “common”; diseases that follow Mendelian inheritance
    - If a common disease is caused by a recessive mutation, it can be found at high frequency in a population
      - MAF (minor allele frequency) > 5%
Why sequence whole genomes?

- SNP/indel/arrayCGH platforms are mainly designed for individuals of West European descent
- For a disease common in somewhere else, like India:
  - Variants at high frequency in India may not be represented in the available platforms
  - Genome is a big entity; SNP/indel/arrayCGH can not cover the entire genome:
    - Largest has 2.1 million markers (compare to 3 billion)
High Throughput Sequencing

- 2007: “Sanger”-based capillary sequencing; one human genome (WGS): ~ $10 million (Levy et al., 2007)
- 2008: First “next-generation” sequencer 454 Life Sciences; genome of James Watson: ~$2 million (Wheeler et al., 2008)
- 2008: The Illumina platform; genome of an African (Bentley et al, 2008) and an Asian (Wang et al., 2008): ~$200K each
- 2009: The SOLiD platform: ~$200K
- Today with the Illumina platform: ~$1K/ genome
Sequencing-based projects

- The 1000 Genomes Project Consortium (www.1000genomes.org)
  - Large consortium: groups from USA, UK, China, Germany, Canada
  - 2,504 humans from 29 populations

- Independent
  - South African (Schuster et al., 2010), Korean, Japanese, UK (UK100K project), Ireland, Netherlands (GoNL project), France, US All of Us, ...

- Ancient DNA: Neandertal (Green et al., 2010); Denisova (Reich et al., 2010)
DNA sequencing

How we obtain the sequence of nucleotides of a species

ACGTGACTGAGGACCGTG
CGACTGAGACTGACTGGGT
CTAGCTAGACTACGTTTTA
TATATATATACGTCGTCGT
ACTGATGACTAGATTACAG
ACTGATTTTAGATAACCTGAC
TGATTTTTTTTTTTTTTTATT...
DNA Sequencing

GENERAL CONCEPTS AND CAPILLARY (SANGER) SEQUENCING
DNA Sequencing: History

**Sanger method** (1977): labeled ddNTPs terminate DNA copying at random points.


Both methods generate labeled fragments of varying lengths that are further electrophoresed.
1. Start at primer (restriction site)
2. Grow DNA chain
3. Include dideoxynucleotide (modified a, c, g, t)
4. Stops reaction at all possible points
5. Separate products with length, using gel electrophoresis
Capillary (Sanger) sequencing

Capillary sequencing (Sanger):
Can only sequence ~1000 letters at a time
Traditional DNA Sequencing

DNA

Shear

DNA fragments

Vector
Circular genome (bacterium, plasmid)

Known location (restriction site)
Double-barreled / paired-end sequencing

[Diagram of genomic segment being cut many times at random (Shotgun) and getting two reads from each segment (paired-end)]
Need to cover region with >7-fold redundancy (7X) if you use Sanger technology
Overlap reads and extend to reconstruct the original genomic region
Definition of Coverage

Length of genomic segment: \( L \)
Number of reads: \( n \)
Length of each read: \( l \)

Definition: Coverage \( C = \frac{n \cdot l}{L} \)

How much coverage is enough?

**Lander-Waterman model:**
Assuming uniform distribution of reads, \( C=10 \) results in 1 gapped region /1,000,000 nucleotides
Challenges with Fragment Assembly

- **Sequencing errors**
  - ~0.1% of bases are wrong

- **Repeats**
  - false overlap due to repeat

- **Computation:** $\sim O(N^2)$ where $N = \# \text{ reads}$
Sanger sequencing

**Advantages**
- Longest read lengths possible today (>1000 bp)
- Highest sequence accuracy (error < 0.1%)
- Clone libraries can be used in further processing

**Disadvantages**
- The most expensive technology
  - $1500 per Mb
- Building and storing clone libraries is hard & time consuming
HIGH THROUGHPUT SEQUENCING
Human genome reference

- 1986: Announced (USA+UK)
- 1990: Started
- 1999: Chromosome 22 sequenced
- 2001: First draft
- 2004: Finished

4 human samples, 14 years, 3-10 billion dollars

Current version: hg38


Chromosomes 1-22, X, Y, MT
Alternative haplotypes
HLA haplotypes
WGS revisited

Test genome

Random shearing and Size-selection

Paired-end sequencing

Read mapping

Reference Genome (HGP)

Maps to Forward strand

Maps to Reverse strand
WGS revisited

Test genome

Random shearing and Size-selection

Paired-end sequencing

Read mapping

Reference Genome (HGP)

Maps to Forward strand

Maps to Reverse strand
HTS Technologies

- **Short read:**
  - 454 Life Sciences: the first, acquired by Roche -- *dead*
    - *Pyrosequencing*
  - Illumina (Solexa): *current market leader*
    - *GAIIx, HiSeq2000, MiSeq, HiSeq2500, NovaSeq*
    - *Sequencing by synthesis*
  - Applied Biosystems -- *dead*
    - *SOLiD: “color-space reads”*

- **Long Read:**
  - Pacific Biosciences Single Molecule Real Time
    - *RSII, Sequel*
  - Oxford Nanopore Technologies:
    - *MinION, Flongle, PromethION, GridION*
Fundamental informatics challenges

1. Interpreting machine readouts – base calling, base error estimation

2. Data visualization

3. Data storage & management
Gzip compressed raw data for one human genome > 100 GB (Illumina)
Informatics challenges (cont’d)

4. SNP, indel, and structural variation discovery

5. *De novo* Assembly
### What can we use them for?

<table>
<thead>
<tr>
<th></th>
<th>Sanger</th>
<th>Illumina</th>
<th>PacBio</th>
<th>ONT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>De novo assembly</strong></td>
<td>Fragmented</td>
<td>Heavily Fragmented</td>
<td>Fragmented, needs polishing</td>
<td>Less Fragmented, needs polishing</td>
</tr>
<tr>
<td><strong>SNP Discovery</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Larger events</strong></td>
<td>Yes</td>
<td>Mid-range</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Transcript profiling</strong></td>
<td>No</td>
<td>Yes</td>
<td>Somewhat</td>
<td>Somewhat</td>
</tr>
</tbody>
</table>
CURRENT PLATFORMS
Features of HTS data

- Short sequence reads
  - 150 - 300 bp Illumina

- Long, but error prone sequence reads
  - Average ~50 Kb PacBio - 12% error
  - Up to 1 Mb ONT – 20% error

- Huge amount of sequence per run
  - Up to terabases per run (3 Tbp for Illumina/NovaSeq 6000)

- Huge number of reads per run
  - Up to billions

- Higher error (compared with Sanger)
  - Illumina: mostly substitutions
  - PacBio / ONT: mostly indels
Whole Genome Sequencing

Test genome

Random shearing and Size-selection

Paired-end sequencing (Illumina)

Reference Genome (HGP)

Single-end sequencing (PacBio/ONT)

Long range Sequencing (10x Genomics)
Sequencing technologies

**Short-Read**
- Illumina
  - 100-200bp
  - Paired-end
  - Billions of reads
  - < 0.1% error

**Long Read**
- PacBio and Oxford Nanopore
  - > 10 Kb, up to 1 Mb
  - Single-end
  - Hundreds of millions of reads
  - 12-20% error – indel dominated

**Long Range**
- 10X + Illumina
  - 100-200bp
  - Paired-end
  - Billions of reads
  - < 0.1% error
  - Barcoded: 30-50 Kb molecule range
Illumina

- Current market leader
- Based on *sequencing by synthesis*
- Current read length 150-300bp
- Paired-end easy, longer matepairs harder
- Error ~0.1%
  - Substitution errors dominate
- Throughput: Up to 3 Tbp in one run (2 days)
- Cheapest sequencing technology
  - Cost: ~ $1,000 per human
Illumina

- HiSeq 2000/2500
- MiSeq
- NovaSeq
- HiSeq 2000/2500
Illumina – FASTQ output

Read and Quality (1)

@FC81ET1ABXX:3:1101:1215:2154/1
TTTTTCAATGTGTTGCTATTTTTATATCTTCTTTTTGAGAATTGTCCTGTCNTNCNCTNCNNTNCTANGGAGTTTGT
+HHGHHHHHHGGHHHHDDFHHHHHHHHHHHHHHHHHHHHHHHHHEHHHHHHHHEHHHHEGGDEF2CGDCDFB0>DA#...

Read and Quality (2)

@FC81ET1ABXX:3:1101:1215:2154/2
AAGCCANNTNNNNTNNNNTNNNNTNTGATCCTCATAGCTACCTTATGCAAAAATCAACTCAAGATGGATGAAGGTCTTAACCTAATAC
+HHHBF###.83<9::7DFBFEEFE;EBEBE8C>2D8@BACDFG=E@=CDDHEGDB;.<:19*23?=@####

- Read length and quality string length are the same
- All read/1s are the same length in the same run
- All read/2s are the same length in the same run
Illumina

- **Read mapping:**
  - mrsFAST, BWA-MEM, minimap2, Bowtie2, BFAST, many more

- **De novo assembly:**
  - SPAdes, Velvet, ABysS, SGA, ALLPATHS, ....
Pacific Biosciences

- “Third generation”; single molecule real time sequencing (SMRT)
- No replication with PCR
- Phosphates are labeled. Watches DNA polymerase in real-time while it copies single DNA molecules.
- Premise: long sequence reads in short time (median 1.4 kbp)
- Errors: ~12%; indel dominated
- ~$ 3,000 / human
Pacific Biosciences

- For any DNA polymerase you can read a total of ~60 kb (median) sequence

Two sequencing protocols:
- CLR: single read
- CCS: Make a circle, re-read the same molecule 5-6 times
  - Multiple sequence alignment to correct errors
  - Median length = 60000 / 6 = 10 Kbp
  - > 99% accuracy
Nanopore sequencing

- Up to 2 Mbp reads
  - 15-20% error, indel dominated
- Real-time analysis supported
- RNN-based basecallers
Nanopore sequencing
PacBio & ONT

- **Read mapping:**
  - Minimap2, MashMap, NGM-LR, …

- **De novo assembly:**
  - Canu, Flye, FALCON
HTS: Computational Challenges

- **Data management**
  - Files are very large; compression algorithms needed

- **Read mapping**
  - Finding the location on the reference genome
  - All platforms have different data types and error models
  - Repeats!!!!

- **Variation discovery**
  - Depends on mapping
  - Again, all platforms has strengths and weaknesses

- **De novo assembly**
  - It’s very difficult to assemble short sequences and/or long sequences with high errors