Genetic Variation II
Genomics: Bio5488, Spring 2022

Sheng Chih (Peter) Jin, Ph.D.
Assistant Professor of Genetics and Pediatrics
Washington University School of Medicine
jin810@wustl.edu
(with thanks to Ira Hall, Don Conrad and slides from past years)
Outline

• Organizing principles: the forces that shape genetic variation

• The landscape of genome variation: definitions and numbers

• Genome-wide detection and interpretation of genome variation
Mapping disease genes in the 20\textsuperscript{th} century

(1) Genome wide approaches: low, laborious, inaccurate, low resolution (~10 megabases).

(2) High resolution targeted approaches: low throughput and expensive; one gene at a time (e.g., sequencing “suspicious” genes)

(3) Clever (but laborious and cumbersome) genome-wide screening methods

Germline:
- Positional cloning: Cystic fibrosis (1989)
- Linkage analysis: Huntington’s disease (1994)

Somatic (cancer):
- Cellular transformation to clone RAS (1982)
- Loss of heterozygosity (LOH) mapping to clone RB (1987)
- Representational differential analysis (RDA) to clone PTEN (1997)
2,000-2,003: Two “complete” human genomes

Public Project: (NIH + Universities)
10 years, $3 billion

Private Project (Celera Corp.)
2 years, $300 million

IMPORTANT: We refer to the genome sequence generated by the public project as the “reference genome”. The reference genome has been invaluable for medical research.
DNA hybridization as a crude form of DNA “sequencing”

How does it work?

- Each “spot”, or “feature”, contains millions of identical single-stranded DNA molecules termed “probes”.
- DNA sequences with high nucleotide identity to a given probe on the array will bind to it strongly. Those with low identity will not bind (or weakly).
- By labelling DNA with fluorescent molecules, we can measure how many DNA molecules bind to each probe. This is a bit like DNA sequencing.
- We can directly compare two samples by labelling their DNA with different colors.
RNA expression microarray analysis

(1) Competitive hybridization

- mRNA extraction
- Reverse Transcription, fluorescent labeling
- Combine equal amounts and hybridize
- Scan

(2) Scan: measure fluorescence

(3) Identify differentially expressed genes
Genomic microarray design

Reference Genome

Gene A

Intergenic region

Gene B

RNA Isoforms

3’ RNA expression array

Transcriptome expression array

Genomic Array (Tiling array) (array-CGH) (CMA)

* Not to scale
Array comparative genomic hybridization (array-CGH)  
a.k.a., Chromosomal microarray analysis (CMA)

(1) Competitive hybridization
Control Sample  Experimental Sample
mRNA extraction
Reverse Transcription, fluorescent labeling
Combine equal amounts and hybridize
Scan

(2) Scan: measure fluorescence

(3) Map copy number variants (CNVs) based on deviation of multiple contiguous probes.
Affymetrix SNP arrays

SNP probe design

5' TAGCCATCGGTA N GTACTCAATGATCAGCT 3'

Genomic Sequence

probe = 25 bases

Perfect Match
Mismatch

Allele ‘A’

Perfect Match
Mismatch

Allele ‘B’
SNP genotyping with Illumina bead arrays

- Hybridization is used to capture the correct DNA molecules
- A single base extension is used to genotype the SNP
- This approach produces much higher quality data than hybridization-based genotyping
Where does the content come from?

Publicly Funded Genomics Projects:

- Human Genome Project
- Phase I HapMap Project
- CNV Project
- Phase II HapMap
- 1000 Genomes Project

Number of SNPs on an array:
- 10k
- 500k
- 1M
- 5M
## SNP genotyping platform

<table>
<thead>
<tr>
<th></th>
<th>Chip size in number (SNPs)</th>
<th>Lowest MAF captured</th>
<th>Number (non-synonymous SNPs)</th>
<th>Based on</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Affymetrix</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Axiom Genome-Wide Human EU (Axiom GW EU)</td>
<td>~600 000</td>
<td>1%</td>
<td>10 648</td>
<td>HapMap, Single Nucleotide Polymorphism database (dbSNP), 1000 GP</td>
<td>Targeting European population</td>
</tr>
<tr>
<td>Axiom Genome-Wide Human ASI (Axiom GW ASI)</td>
<td>~600 000</td>
<td>1%</td>
<td>10 346</td>
<td>HapMap, dbSNP, 1000 GP</td>
<td>Targeting Asian population</td>
</tr>
<tr>
<td>Axiom Genome-Wide Human CHB (Axiom GW CHB)</td>
<td>~1 200 000</td>
<td>2%</td>
<td>10 560</td>
<td>HapMap, dbSNP, 1000 GP</td>
<td>Targeting CHB subpopulation</td>
</tr>
<tr>
<td>Axiom Genome-Wide Human PanAFR (Axiom GW PanAFR)</td>
<td>~2 200 000</td>
<td>2%</td>
<td>12 250</td>
<td>HapMap, dbSNP, 1000 GP, Southern African Genomes Project</td>
<td>Targeting African population</td>
</tr>
<tr>
<td><strong>Illumina</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human OmniExpress</td>
<td>~700 000</td>
<td>5%</td>
<td>15 062</td>
<td>HapMap</td>
<td>Optimized tag SNP</td>
</tr>
<tr>
<td>Human Omni1S-8</td>
<td>~1 000 000</td>
<td>5%</td>
<td>5 641</td>
<td>1000GP</td>
<td>Optimized tag SNP</td>
</tr>
<tr>
<td>Human Omni2.5-8</td>
<td>~2 500 000</td>
<td>2.5%</td>
<td>41 900</td>
<td>1000GP</td>
<td>Targeting common and rare variants</td>
</tr>
<tr>
<td>Human Omni2.5S-8</td>
<td>~2 500 000</td>
<td>1%</td>
<td>57 360</td>
<td>1000GP</td>
<td>Targeting rare variants</td>
</tr>
</tbody>
</table>

* Ha et al., EJHG 2014
The genotyping challenge

- **PolyHighResolution**: Good cluster resolution and at least two examples of the minor allele.
- **MonoHighResolution**: Less than two examples of the minor allele usually due to low MAF samples, but possible cluster fusion.
- **Off-Target Variant**: Off-target variant (OTV) can be assayed by OTV genotyping.
- **No Minor Hom**: Two clusters with no examples of the minor homozygous genotypes.
- **CallRate Below Threshold**: SNP call rate is below threshold, but other cluster properties are above threshold.
- **Other**: One or more cluster properties are below threshold. Expect lower quality genotypes.

**Well-genotyped site:**
99.99% of all sites

**Serious batch effects:** 1/10000 sites

(1/10000 x 1M sites typed = 100 hits!)

14
A pretty sad “Manhattan” plot

- Ninety percent of samples typed on 370K array, 10% of samples typed with 610K array
- 33 variants associated with longevity
- Trained model with 150 sites, classified centenarians and non-centenarians with 77% accuracy
- No validation of associated sites

A nice plot for comparison
Strategies for a successful genomic project

Be suspicious! High-throughput genomic data is usually messy. It is very easy to be fooled. Open data exploration and “scrubbing” is the most important part of a project. Replicate results with independent samples and methods when possible. Amazing results are almost always wrong.
The impact of high-throughput genotyping

(1) Haplotype structure

(2) Genome-wide association studies

Patients

Patient DNA

Non-patients

Non-patient DNA

Compare differences to discover SNPs associated with diseases

Disease-specific SNPs

Non-disease SNPs

“Manhattan” Plot
The impact of high-throughput genotyping

(3) Personal disease risk

\[ GRS_i = \sum_{j=1}^{m} x_{ij} \]

(4) Human population history

(5) Personal ancestry determination
Genome sequencing costs: 2001 - 2017

Cost per Genome

NIH National Human Genome Research Institute
genome.gov/sequencingcosts

Moores's Law

A modern genome sequencing experiment

1. Genomic DNA → Shear to desired length (~400bp)
2. DNA fragments → Ligate adapters, size select
3. Sequencing library → Clusters on a flow-cell
4. 500 million paired-end reads ("read-pairs")
Genome assembly is hard!

Ideally, we would “stitch” reads together using the process of whole genome assembly to produce the complete diploid genome of that person, or tumor. This is not possible with the second generation DNA sequencing data.

The raw data: 500 million read-pairs

The genome (3.2 billion bp)

Why is genome assembly so hard? Reads are short, the genome is big and complex.
- The human genome is a nasty beast: it is large, complex, and laden with repetitive elements.
- When repeat size exceeds read-length, it is impossible to assemble through them. This causes errors and gaps.
- Long-range sequence information (10-100 kb) is required to resolve complex regions. Short reads (~150 bp) do not suffice.
**KEY POINT:** We do not sequence human genomes from scratch with second generation technologies. It is too hard.

Instead, we infer genome variation by comparing raw sequencing data from a given individual (or tumor) to the high-quality reference genome produced by the public human genome project in ~2003.

So, we are not really “sequencing genomes”. We are mapping genome variation, indirectly and imperfectly.
Aligning reads to the reference genome

(1) a read-pair (2 x 100 bp)

5’ GGTTACGAATTTTTACACTTTGACCATTAGC -----------------------------------------------/-----------------------------------------------
                           -----------------------------------------------/-----------------------------------------------
GGACTGAAACTTCTCATGTTTATAGATATGCGTGCAAGCGC 5’

(2) The human reference genome (~3 billion bp)

(3) Alignment of the “read-pair” to the reference genome gives coordinates describing where in the human genome the read-pair came from, and whether there are any sequence differences.
To distinguish genome variation from sequencing/alignment errors, we weigh evidence from multiple reads.

**REFERENCE GENOME (HAPLOID)**

```
......GCTATATAAAAATTTTGGGTATCAATGGTTTTTTGGGTATCTCGGATATCGGTATCGGTAC GCACAAACACCGTGATGAC......
```

**REFERENCE GENOME (read length = 50 bp)**

```
C
 TAAAATGGC ATCAATGGTTTTTGCTGCTATCGGC ATCGTATTTCCGT
A
 AAATTTGCTATCACTGGTTTTTCGTGCTATCGGC ATCGTATTTCCGTGCA
T
 ATGGAATTCATACTGGTTTTTCGTGCTATCGGC ATCGTATTTCCGTGCA
G
 TGGTATCAATGTTTTTCGTGCTATCGGC ATCGTATTTCCGTGACA
C
 GTATCAATGCTTTCTGCTATCGGC ATCGTATACCGTGCA GCACA
A
 ATCAATGTTTTTTTGGGTATCGGATATCGGTATCGGTAC GCACAA
C
 CAATGTTTTTTTGGGTATCGGATATCGGTATCGGTAC GCACAAACACCGTGATGAC
A
 ATGGAATTCATACTGGTTTTTCGTGCTATCGGC ATCGTATTTCCGTGCA
G
 GGGGGGATCTCGGATATCGGTATCGGTAC GCACAAACACCGTGATGAC
T
 TTTGATCGGATATCGGTATCGGTAC GCACAAACACCGTGATGAC
```

**10 READS ALIGNED TO**

**REFERENCE GENOME**

**heterozygous single nucleotide variant/polyorphism (SNV/SNP)**

**heterozygous 3 bp deletion**

**homozygous single nucleotide variant/polyorphism (SNV/SNP)**

**sequencing error**

**IMPORTANT:** To do this, we must sequence each base in the genome multiple times. The number of times, on average, that each base in the genome is sequenced is termed **coverage**. Today, the industry standard for Illumina whole genome sequencing is >30X coverage relative to the haploid reference genome. Since our genomes are diploid, this corresponds to >15X coverage for each chromosome.
How do we detect SNVs and small INDELs?

- Examine alignments at each base in the reference genome and assign genotypes with a probabilistic model that takes into account the sequencing error rate.
  - Generic: GATK, FreeBayes, SAMTools, SOAPSNP, etc. We use GATK and FreeBayes.
  - Tailored to somatic mutations: MUTECT, SPELKTRA, SomaticSniper.
  - Variant detection and genotyping improves under the assumption of diploidy.

- Handle INDELs with care.
  - Alignment algorithms do a poor job aligning INDEL-containing reads.
  - Best variant detectors perform realignment around putative INDELs to reduce errors.
  - Newer methods perform local assembly to avoid alignment biases.

- In general, SNV detection is relatively “easy” and robust; INDEL detection is difficult and error prone.
Visualizing alignments with the Integrative Genomics Viewer (IGV)

~90X coverage whole genome sequence data (breast cancer)

Robinson et al., Nature Biotechnology (2011); http://www.broadinstitute.org/igv/
Detecting germline and somatic SNVs

- ~4 million germline SNVs are detected in a typical whole-genome dataset.
- $10^2$-$10^5$ somatic SNVs, depending on tumor and tumor type.
- Loss of heterozygosity can be observed at heterozygous germline SNVs.
Example of a somatic INDEL (13 bp deletion)

- ~350,000 germline INDELs detected in whole genome sequence data, ~3,000 in exome.
- Tens to thousands of somatic INDELs in tumors, depending on tumor and tumor type.
- INDELs are a common source of loss of function mutations.
Example of an INDEL alignment error

Alignment errors make accurate INDEL detection hard, and can cause false SNV calls nearby.
This looks pretty straightforward, so what’s the big deal?

There are a lot of artifacts. The genome is big and complex, the reads are short and error-prone.

Large, repetitive & polymorphic genome + short reads + heuristic aligners = systematic alignment errors.

Systematic alignment errors look like genetic variants.
Identification of poorly assembled regions of the reference genome showing abnormally high read-depth

This identifies 10,701 loci encompassing ~10 Mb (0.34%) of the genome.

Chiang et al., *Nature Methods* 2015
Variant detection sensitivity and accuracy

• False positive = a false variant call. How do we reduce false positives?
  - Document/fix genomic regions prone to artifacts.
  - Simple filters: variant quality scores, mapping quality (uniqueness), read depth, etc.
  - Take intersection of multiple algorithms (but, increases false negatives).

• False negative = a missed variant call. How do we reduce false negatives?
  - Deeper sequencing (>100X).
  - Call mutations with lenient criteria, then perform large-scale validation.
  - Take union of multiple algorithms (but, this increases false positives).

• How do we minimize both at the same time?
  - Requires some knowledge of the truth + Careful testing + Manual labor.
  - Machine learning using a training set of true positives and covariates. GATK variant quality score recalibration does this.
Tuning variant detection performance using receiver operating characteristic (ROC) curves

**Requires:** some relatively unbiased method for distinguishing true and false variant calls. Simulation, or independent data.

**Approach:** try a bunch of algorithms or parameters, measure performance.
Low overlap between different tools, even for the easiest task (100X exome data, normal samples, SNVs & INDELs)

**Key points:**
- 57.4% of SNVs called by all 5 tools.
- 26.8% of INDELs called by all 3 tools.
- **Note:** Alignment methods can also affect overlap among pipelines.
- **Note:** Parameter selection open more important than algorithm.
- **Note:** Overlap is increasing as data improves & strategies converge.

---

**GATK alone (~30,000 variants)**

<table>
<thead>
<tr>
<th></th>
<th>sensitivity</th>
<th>specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SNV</strong></td>
<td>95.34%</td>
<td>99.72%</td>
</tr>
<tr>
<td><strong>INDEL</strong></td>
<td>????</td>
<td>61.8%</td>
</tr>
</tbody>
</table>

O’Rawe et al., *Genome Medicine* (2013)
Genome structural variation

**Structural variation (SV):** Differences in the copy number, orientation, or location of “large” genomic segments (e.g., >50 bp)

**Other terms:**
- Genomic Rearrangements
- Copy Number Variants (CNVs)
- Mobile Element Insertions (MEIs)
Why map structural variation?

• **SV is a common form of genetic variation**
  - 5,000 variants between two humans; SVs are present in all species
  - Large functional potential relative to point mutations
  - Major source of mutation in cancer genomes: gene amplification, gene deletion, gene fusion

• **To map traits**
  - Inherited causal variants may no be well-tagged by SNPs
  - Some traits are caused by spontaneous mutations (e.g., cancer, sporadic human disorders)

• **To understand genome/tumor/species evolution**
  - Gene birth, gene dosage, effects on recombination
  - Evolution of genome architecture
  - Local and global genomic instability
  - Punctuated genome evolution = multiple simultaneous mutations
SV breakpoints defined

Breakpoints are the junctions that define structurally variable genomic segments. SVs are usually detected based on the presence of these junctions in the experimentally sequenced “test” genome, but not the reference genome.

However, “breakpoint” is an ambiguous term because it can simultaneously describe one junction in the test genome, and two junctions in the reference genome.

The VCF file format accounts for this ambiguity by introducing two new terms:
- **Novel adjacency**: the breakpoint in the test genome.
- **Breakends**: the two breakpoints in the reference genome.
Visualizing SV breakpoints

Deletion

Reference: ABCDEFGHIJ

Test: ABCCHIJ

Inversion

Reference: ABCDEFGHIJ

Test: ABCGFEHDIJ

Tandem Duplication

Reference: ABCDEFGHIJ

Test: ABCDFEGDEFGHIJ

Distant Insertion

Reference: ABCDEFGHIJ

Test: ABCDEXFGHIJ

Reciprocal translocation

Reference Chr1: ABCDEFGHIJ

Test chr1/2: ABCDE12345

Test Chr2/1: 12345FGHIJ

Reference Chr2: 12345678910
3 ways to detect a SV

1) depth of sequence coverage
   = “read-depth analysis”
   (copy number variation)

2) “read-pairs” span a breakpoint
   = “paired-end mapping”
   (all classes of SV)

3) read contains a breakpoint
   = “split-read mapping”
   (all classes of SV)
Detecting CNVs with read-depth analysis

**Basic approach:**

1. Count reads in sliding windows (e.g., 1 kb) of uniquely mappable genome sequence.
2. Normalize for GC bias.
3. Use segmentation to define CNVs (similar to array-CGH data).
4. Some methods predict absolute copy number, others compare two samples.
GC Normalization of Illumina data

Raw Data

Normalized Data

Read count per 5kb window

GC Fraction per 5 kb window

GC Normalized Z-score

GC Normalized Z-score

GC Fraction per 5 kb window

All Windows

33–34% GC Windows

43–44% GC Windows
Detecting CNVs with read-depth analysis

Strengths:
1) Fast and simple.
2) Directly measures copy number.
3) Relatively straightforward interpretation: Is gene X duplicated, deleted, or amplified?

Weaknesses:
1) Limited resolution (1-10 kb) = imprecise boundaries.
2) Cannot detect balanced events or reveal variant architecture.
CNV detection in single cells (male trisomy 21 fibroblasts)

McConnell et al., Science (2013)
3 ways to detect a SV

1) depth of sequence coverage
   = “read-depth analysis” (copy number variation)

2) “read-pairs” span a breakpoint
   = “paired-end mapping” (all classes of SV)

3) read contains a breakpoint
   = “split-read mapping” (all classes of SV)
Discordant paired-end alignments suggest breakpoints

Illumina HiSeq

paired-end reads (readpairs)

5' GGTGTAAGATAGTTTTCTTTTACACTCTTGAACCACCATCTACGCT

GGACTGAAACTCATCTGTCTTTATAGATATGCGTGCAAGC 5'

align reads

concordant readpairs (>99%)

discordant read-pairs (<1%)

everted orientation = tandem duplication

too big = deletion

same-strand = inversion

distant rearrangement

size distribution
Paired-end mapping algorithms cluster discordant alignments that “agree” with each other (support the same breakpoint).
Split-read mapping localizes breakpoints to base-pair resolution (more or less)

Traditional split-read mapping
- align to genome
- Cluster reads (+/- a few bp)

PINDEL (Ye et al., 2009)
- align reads to genome
- sensitive split-read alignment at loci with hints of SV

NOTE: BWA-MEM now does joint paired-end and split-read alignment
Multi-signal SV detection using the LUMPY algorithm

By representing SV detection signals in a general way (breakpoint probability distributions), LUMPY can analyze them together, simply.
A typical human genome has several thousand deletions
Several hundred duplications
Several hundred inversions
Several hundred transposon insertions in the reference genome
Several hundred transposon insertions in the test genome
Tens of retrogene insertions

breakpoint calls

92kb
Tens of hundreds of complex variants

Important: we cannot infer variant class based solely upon relative read orientation; e.g., an apparent deletion may really be part of a complex rearrangement.

Note: 5-15% of SVs have complex breakpoints patterns. (Quinlan et al., 2010, Genome Research; Conrad et al., 2010, Nature Genetics)
So, why is structural variation breakpoint mapping so #$%^&#@##% hard?

1) Lots of false positives.

2) Lots of false negatives.

3) Some structural variation breakpoints are virtually impossible to detect with current sequencing technologies.
Three mechanisms for SV formation

(1) Homologous recombination (HR / NAHR)
- Homology > 20bp
- Sturtevant 1925
- 5-10% of SVs*

(2) Non-homologous end joining (NHEJ)
- Homology ~ 0
- Moore & Haber, 1996
- >80% of SVs*

(3) Template switching during DNA replication
- Multiple breakpoints & microhomology (2-5p)
- BIR: Smith et al., 2007
- FoSTeS: Lee et al., 2007
- MMBIR: Hastings & Lupski, 2009
- 5-10% of SVs*

*Ira Hall’s guess; this is an open question
Recombination within a tandem array

A B B C

X

A B B C

↓

readpair

A B C

product 1

A B B B B C

product 2
Recombination between large flanking repeats

A  B  C  B’  D

X

A  B  C  B’  D

↓

product 1

A  B  D

readpair

product 2

A  B  C  B’  C’  B”  D
**Definition:** a variant with multiple clustered breakpoints that most likely arose through a single mutation, but cannot be explained by a simple end-joining or recombination event.

**Implications:**
- multiple simultaneous mutations = punctuated evolution.
- novel mechanisms.
- difficult to detect and interpret.