Genetic Variation II

Ira Hall, Ph.D.
Associate Director, McDonnell Genome Institute
Associate Professor, Department of Medicine
Washington University School of Medicine
ihall@genome.wustl.edu
~4 million germline variants

Environment

Human Diversity

Rare “Mendelian” Disease

Common “Complex” Disease

Cancer

(+ $10^2$ - $10^5$ somatic mutations)
Mapping disease genes in the 20th century

**Mapping disease genes in the 20th century**

**Karyotyping**

**Spectral Karyotyping**

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

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

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

**Germline:**
- 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)
2000-2003: 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
   - Control Sample
   - Experimental Sample
     - mRNA extraction
     - Reverse Transcription, fluorescent labeling
     - Combine equal amounts and hybridize

2. Scan: measure fluorescence

3. Identify differentially expressed genes
   - Expression Level: Control Sample
   - Expression Level: Experimental Sample
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

(2) Scan: measure fluorescence

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

Relative DNA copy number (experimental / control)
SNP genotyping microarray design

Gene A

Intergenic region

Gene B

RNA Isoforms

3’ RNA expression array

Transcriptome expression array

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

SNPs

SNP Genotyping Array

A/C G/T A/G G/C A/T G/C T/G
Affymetrix SNP arrays

SNP probe design

5´ T / G 3´
TAGCCATCGGTA N GTACTCAATGATCAGCT

SNP probe = 25 bases

Allele ‘A’
Perfect Match
Mismatch

Allele ‘B’
Perfect Match
Mismatch
SNP genotyping with Illumina bead arrays

Illumina Bead Array

Genotyping by hybridization + single base extension

- 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 platforms

<table>
<thead>
<tr>
<th>Platform</th>
<th>SNP probes</th>
<th>CNV probes</th>
<th>Cost/sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affymetrix 6.0</td>
<td>900K</td>
<td>950K</td>
<td>$450</td>
</tr>
<tr>
<td>Affymetrix Axiom</td>
<td>600K</td>
<td>11K</td>
<td>?</td>
</tr>
<tr>
<td>Illumina Omni5</td>
<td>5M</td>
<td>0?</td>
<td>$1000</td>
</tr>
<tr>
<td>Illumina Omni1</td>
<td>1M</td>
<td>100K</td>
<td>$450</td>
</tr>
<tr>
<td>Illumina OmniExp</td>
<td>720K</td>
<td>0</td>
<td>$250</td>
</tr>
</tbody>
</table>

Typical Caucasian is polymorphic at 3M sites, Africans a little more, East Asians a little less

*Prices as of 2012*
The genotyping challenge

Well-genotyped site: 99.99% of all sites

Serious batch effects: 1/10000 sites

(1/10000 x 1M sites typed = 100 hits!)
The importance of careful quality control

- 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

Editorial Expression of Concern

Researchers involved in genome-wide association studies have expressed technical concerns about a Report by P. Sebastiani et al., "Genetic signatures of exceptional longevity in humans," published in Science Express on 1 July 2010. In their study (1), Sebastiani et al. used a number of different genotyping platforms and neglected to perform data quality-control steps, which resulted in their reporting several false-positive single-nucleotide polymorphism (SNP) associations. In particular, one of the platforms used in their work, the Illuma 610-Quad array, has been shown in unpublished studies by other investigators to produce artifactual genotype data at a subset of SNPs.

Science and the authors are taking these concerns seriously. Since learning of these potential problems, Sebastiani et al. have been performing a thorough quality-control analysis on the original raw data, as well as generating new data to compare the genotype calls from the 610-Quad array and the other platforms within the same individuals. These steps aim to eliminate biases between platforms. Furthermore, they are undertaking an additional validation measure on several SNPs via the TaqMan<sup>®</sup> assay, a non-microarray-based genotyping method. After ensuring that all data are clean, they will redo the statistical and modeling analyses, which they expect to be completed in December. At that point, Science will reevaluate the paper, determine the extent to which the strength of its original conclusions has been altered by the revised data, and take the appropriate action.

Bruce Alberts
Editor-in-Chief

Reference
Strategies for a successful genomics project

Generate Discovery Data $\rightarrow$ Visualize Data $\rightarrow$ QC, Normalize $\rightarrow$ Make calls $\leftarrow$ Visualize Data $\leftarrow$ QC, Normalize $\rightarrow$ Visualize & QC Data $\rightarrow$ Validate Calls $\rightarrow$ Replicate Biology

**Be suspicious!** High-throughput genomic data is usually messy. It is very easy to be fooled. Often 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

International HapMap Project

(2) Genome Wide Association Studies

Patients → Patient DNA → Compare differences to discover SNPs associated with diseases → Disease-specific SNPs

Non-patients → Non-patient DNA → 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
The $1000 Genome has finally arrived! (sort of)

- Caveat: The X10 cluster costs $10 million, and sequencing costs do not include labor, data storage, electricity, etc.
- **Throughput**: one run: 3 days, 165 genomes = ~20,000 genomes / year
- **Parallelism**: 3 billion molecules per flowcell, 2 flowcells per machine. 60 billion molecules are sequenced simultaneously.
- We (The Genome Institute) purchased one last year. A handful of other US institutions also have one.
A modern genome sequencing experiment

genomic DNA

Shear to desired length (~400bp)

DNA fragments

ligate adapters, size select

sequencing library

Illumina HiSeq

clusters on a flow-cell

500 million paired-end reads (“read-pairs”)

5' GGTGTACGAATGTTTCTCTTTACTCCTGGACCATTGCCTAGC 3'

5' GGACTGAACTTCATCTGCTTTATAGATATGCGTGCAGC 3'
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 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 an 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 (~125 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' GGTACGAATTCTTTTACACTCCCTTCGACCCTTCACTTCGACCAATTC
------------------------------------------
GGACTGAAACTTCATCTGTCTTTATAGATATGCGTGCAGC 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 and alignment errors, we weigh evidence from multiple reads.

**REFERENCE GENOME (HAPLOID)**

```
GCTATATAAAATGGGATCAATGGTTTTGGGTCATCGGCGGTATCGGTATTCCGTGCAGCACAACACCCGTGATGAC
```

10 READS ALIGNED TO REFERENCE GENOME (read length = 50 bp)

- **heterozygous single nucleotide variant/polymorphism (SNV/SNP)**
- **heterozygous 3 bp deletion**
- **homozygous single nucleotide variant/polymorphism (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 single nucleotide variants (SNVs)

- ~4 million germline SNVs are detected in a typical whole-genome dataset.
- $1 \times 10^2$-$10^5$ somatic SNVs, depending on tumor and tumor type.
- Loss of heterozygosity can be observed at heterozygous germline SNVs.
- ~350,000 germline INDELs detected in whole genome sequence data, ~3000 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, 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.

Forcing reads to align where they do not belong leads to false variant calls (a particularly egregious example)
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, SNPs & 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 often more important than algorithm.
- Note: Overlap is increasing as data improves & strategies converge.

<table>
<thead>
<tr>
<th></th>
<th>sensitivity</th>
<th>specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNV</td>
<td>95.34%</td>
<td>99.72%</td>
</tr>
<tr>
<td>INDEL</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)
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
3 ways to detect a structural variant (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

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.
5) Lots of read-depth methods. What do we use? CNVnator + in-house tools.

**Basic approach:**
GC normalization of Illumina Data

**Raw Data**
- Read count per 5 kb window
- GC Fraction per 5 kb window

**Normalized Data**
- GC Normalized Z-score
- GC Fraction per 5 kb window

**Histograms**
- All Windows
- 33–34% GC Windows
- 43–44% GC Windows

- Number reads per 5 kb window
- Number Windows

- Distribution of GC fraction and number of reads per 5 kb window for different GC content ranges.
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 structural variant (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' GGTGTAAGATTTCTTCTACACTCTTTGACCATCTAGC ---------------//--------------- GGACTGAAACTTCATCTGTCTTTATAGATATGCGTGACAGC 5'

aligned reads

concordant readpairs (>99%)

discordant read-pairs (<1%)

everted orientation = tandem duplication

too big = deletion

same-strand = inversion

distant rearrangement

Fragment Length (bp)

size distribution

No. Fragments

0 500 1000

0
Paired-end mapping algorithms cluster discordant alignments that “agree” with each other (support the same breakpoint).

- **too big (+/-)**
  - = deletion
  - Test genome: A B C
  - Ref. genome: A B C

- **everted (-/+)**
  - = tandem duplication
  - Test genome: A B B C
  - Ref. genome: A B C

**A deletion**

**A tandem duplication**

2.5kb
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

3.8kb

breakpoint calls
Several hundred transposon insertions in the reference genome
Several hundred transposon insertions in the test genome
Tens of retrogene insertions
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*

*my guess; this is an open question
Recombination within a tandem array

- product 1: A B B C
- product 2: A B B B C

Readpair connections:

- product 1: A B B C
- product 2: A B B B C
Recombination between large flanking repeats

Product 1: A B D

Product 2: A B C B’ C’ B’’ D

Readpair: A B D
4) Complex rearrangements can produce extremely confusing breakpoint patterns

Complex chromosomal rearrangement (CCR)

**Chromothripsis** *(Stephens et al., 2011)*

**Variant**

- A
- D
- C
- D
- E

**Ref.**

- A
- B
- C
- D
- E

**mildly complex germline SV**

**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

*Rausch et al., 2012*

*Chromothripsis* *(Stephens et al., 2011)*
CNAs are often linked by complex rearrangements

breakpoints
red: direct orientation (deletion)
green: everted orientation (tandem dup.)
blue: inverted orientation (inversion)
grey: inter-chromosomal

Circos visualization tool: Krzywinski et al., 2009
Co-amplification of genes from different chromosomes
Catastrophic chromosome shattering: Chromothripsis
An extreme case of punctuated cancer genome evolution: one complex rearrangement, 5 fusion products
Chromothriptic Cure of WHIM Syndrome

David H. McDermott,1,14 Ji-Liang Gao,1,14 Qian Liu,1 Marie Siwicki,1 Craig Martens,2 Paesjonette Jacobs,1 Daniel Velez,1 Erin Yim,1 Christine R. Bryke,1,12 Nancy Hsu,1,12 Zunyan Dai,1,13 Martha M. Marquesen,2 Elina Stregevsky,3 Nana Kwatemaa,6 Narda Theobald,6 Debra A. Long Priel,6 Stefania Pittaluga,1 Mark A. Raffeld,7 Katherine R. Calvo,8 Irina Marie,9 Ronan Desmond,5,10 Kevin L. Holmes,7 Douglas B. Kuhns,7 Karl Balabanian,11 Françoise Bachelier,11 Steohen F. Porosilla,2 Harv L. Malech,2 and Philip M. Murphy1


Chromothripsy is a catastrophic cellular event recently described in cancer in which chromosomes undergo massive deletion and rearrangement. Here, we report a case in which chromothripsy spontaneously cured a patient with WHIM syndrome, an autosomal dominant combined immunodeficiency disease caused by gain-of-function mutation of the chemokine receptor CXCR4. In this patient, deletion of the disease allele, CXCR4<sup>R334X</sup>, as well as 163 other genes from one copy of chromosome 2 occurred in a hematopoietic stem cell (HSC) that repopulated the myeloid but not the lymphoid lineage. In competitive mouse bone marrow (BM) transplantation experiments, Cxcr4 haploinsufficiency was sufficient to confer a strong long-term engraftment advantage of donor BM over BM from either wild-type or WHIM syndrome model mice, suggesting a potential mechanism for the patient’s cure. Our findings suggest that partial inactivation of CXCR4 may have general utility as a strategy to promote HSC engraftment in transplantation.