Genome Sequencing and Structural Variation

Peter N. Robinson

Institut für Medizinische Genetik und Humangenetik
Charité Universitätsmedizin Berlin

Genomics: Lecture #10
Today

- Structural Variation
  - Deletions
  - Duplications
  - Inversions
  - Other

- Array CGH
  - Algorithms for detecting structural variations from WGS data (Introduction)
    - Read-depth
    - Split reads etc

- Read-depth Algorithm: Detailed Example
Outline

1. Structural variants
2. Array CGH
3. Bioinformatics Approaches for Structural Variant Discovery
4. Poisson
5. GC Content
6. Read depth
7. CNV Calling
**CNVs vs. SNVs**

### Single-nucleotide variants

- Several thousand SNVs in typical exome (1% of the genome)
- ca. 3–4 million SNVs in typical genome

### CNV

- Hundreds/Thousands of CNVs per Genome
- average size 250,000 nt
  - (n.b.: avg. gene is ca. 60,000 nt)
CNVs vs. SNVs

**Single-Nucleotide Variants (SNV)**

- Most missense, nonsense mutations, class also includes synonymous substitutions and intergenic substitutions
- Previously thought to be main source of interindividual genomic variability

**Copy-Number Variants (CNV)**

- Major class of genomic structural variation
- Alteration in normal number of copies of a genomic segment

(Normal: 2 copies; Deletion: 1 copy; Duplication 3 copies.)
Structural Variation: Definition

Structural variations (SV) are Genomic rearrangements that effect more than 1 Kb\(^1\)

- Duplication and Amplification
- Deletion (often called Loss of heterozygosity if deletion occurs somatically, e.g., cancer)
- Translocation and Fusion
- Inversion
- Breakpoints at SV edges

\(^1\)Yes, this definition is arbitrary!
Inversion

- A balanced structural variation (no loss/gain of genomic segment)
- Can be a neutral variation
- Can disrupt a coding sequence
- Can interrupt regulatory interactions
Intrachromosomal translocation

- A balanced structural variation (no loss/gain of genomic segment)
- Can be a neutral variation
- Can disrupt a coding sequence
- Can interrupt regulatory interactions
A balanced structural variation (no loss/gain of genomic segment)

Translocation between two different chromosomes

Like other balanced SVs, can be neutral or disrupt coding sequences or regulatory interactions
An unbalanced structural variation (loss of genomic segment) results in dosage abnormality of genes contained in deletion.

Indirect regulatory imbalances also possible.
An *unbalanced* structural variation (gain of genomic segment)

- results in dosage abnormality of genes contained in deletion
- Indirect regulatory imbalances also possible
Structural Variation: Distribution in Genome

∼ 1000 SVs >2.5kb per Person


Detection of Structural Variants

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Detection</th>
<th>Copy-neutral events</th>
<th>Maximum resolution</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Deletions and duplications</td>
<td>Insertions</td>
<td>Unbalanced translocations</td>
<td>Balanced translocations</td>
</tr>
<tr>
<td>Early 1970s</td>
<td>Karyotyping/G-bANDING</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>FISH-based</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Early 1990s</td>
<td>CGH</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Mid 1990s</td>
<td>M-FISH/SKY/COBRA</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Late 1990s</td>
<td>RxFISH</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Array-based</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Early 2000s</td>
<td>1-Mb BAC array-CGH</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Tiling-path BAC array-CGH</td>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Oligonucleotide array-CGH</td>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Late 2000s</td>
<td>SNP arrays</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>NGS-based</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Still no method to reliably detect all SVs
Array CGH currently the gold standard for CNVs

A small heterozygous deletion in the $\beta$-globin locus.

DNA Hybridization:

- If two DNA strands are separated, they still "recognize" their opposite (reverse complementary) strand.
- Denaturation: Heat DNA until strands separate.
- Renaturation (hybridization): Cool slowly and allow reverse complementary to anneal to one another.
Array-CGH

- Ratio of 2 fluorescent signals indications loss or gain of DNA segment
Array CGH can detect

- Deletions
- Duplications (& and other gains in copy number)
- More complex copy number changes (e.g., mixed)

Array-CGH: Indications in Human Genetics

- Intellectual disability or developmental delay of unknown cause
- Congenital malformation or facial dysmorphism
- Autism or suspicion of a specific chromosomal disorder

Array-CGH is a screening investigation to investigate nearly the entire genome for CNVs in an untargeted fashion. Many findings are “new” and may be difficult to interpret: cause of a disease or neutral polymorphism?
Bioinformatics Approaches for SV Discovery with WGS data

Several characteristics of NGS data can be exploited for identification of different kinds of structural variants

- Read depth
- Read pairs
  1. Orientation of mates
  2. Distance of aligned mates to one another
- Split reads
- Fine mapping of breakpoints by local assembly
Paired sequences are extremely useful for read mapping in whole genome sequencing because we not only have the information about the DNA sequences but also the distance and orientation of the two mapped reads to one another. There are two major classes of paired sequences.

1. **Paired end.** Fragment libraries\(^2\) are sequenced from both ends. The sequencing direction is from the ends towards the middle.

2. **Mate-pair** libraries. We will review this today

\(^2\)As discussed in the very first lecture.
Construction of Illumina mate-pair sequencing libraries.

1. Fragments are end-repaired using biotinylated nucleotides
2. After circularization, the two fragment ends (green and red) become located adjacent to each other
3. The circularized DNA is fragmented, and biotinylated fragments are purified by affinity capture. Sequencing adapters (A1 and A2) are ligated to the ends of the captured fragments
4. The fragments are hybridized to a flow cell, in which they are bridge amplified. The first sequence read is obtained with adapter A2 bound to the flow cell
5. The complementary strand is synthesized and linearized with adapter A1 bound to the flow cell, and the second sequence read is obtained
6. The two sequence reads (arrows) will be directed outwards from the original fragment.

## Paired-end vs. Mate pair

<table>
<thead>
<tr>
<th></th>
<th>Paired-end</th>
<th>Mate pair</th>
</tr>
</thead>
<tbody>
<tr>
<td>insert size</td>
<td>( \approx 250 \text{ bp} )</td>
<td>2–20 kb</td>
</tr>
<tr>
<td>DNA</td>
<td>1.5–5 ( \mu \text{g} )</td>
<td>5–120 ( \mu \text{g} )</td>
</tr>
<tr>
<td>lab work</td>
<td>easier</td>
<td>harder</td>
</tr>
<tr>
<td>Costs</td>
<td>less</td>
<td>more</td>
</tr>
</tbody>
</table>

Note:

|-----75----|------------------------100-----------------|-----75-----|

If we have two 75 bp paired-end reads with a 100bp middle piece, the insert size is calculated as

\[ 2 \times 75 + 100 = 250 \text{ nt}. \]

The fragment size is insert size plus length of both adapters (\( \approx 120 \text{ nt extra} \)).
Analysis of read depth can identify deletion/duplications

Heterozygous Deletion?
Mappability Issue?
Poor "sequencability"?
Read depth

Characteristic signatures of paired-end sequences

<table>
<thead>
<tr>
<th></th>
<th>sequenced</th>
<th>reference /mapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td><img src="normal_diagram" alt="Diagram" /></td>
<td><img src="normal_diagram" alt="Diagram" /></td>
</tr>
<tr>
<td>Inversion</td>
<td><img src="inversion_diagram" alt="Diagram" /></td>
<td><img src="inversion_diagram" alt="Diagram" /></td>
</tr>
<tr>
<td>Tandem duplication</td>
<td><img src="tandem_diagram" alt="Diagram" /></td>
<td><img src="tandem_diagram" alt="Diagram" /></td>
</tr>
<tr>
<td>Insertion</td>
<td><img src="insertion_diagram" alt="Diagram" /></td>
<td><img src="insertion_diagram" alt="Diagram" /></td>
</tr>
<tr>
<td>Deletion</td>
<td><img src="deletion_diagram" alt="Diagram" /></td>
<td><img src="deletion_diagram" alt="Diagram" /></td>
</tr>
<tr>
<td>Translocation</td>
<td><img src="translocation_diagram" alt="Diagram" /></td>
<td><img src="translocation_diagram" alt="Diagram" /></td>
</tr>
</tbody>
</table>

graphic credit: Victor Guryev
Deletions in WGS Data

What are the signals that let us detect a deletion?
Deletions in WGS Data

<table>
<thead>
<tr>
<th>SV classes</th>
<th>Read pair</th>
<th>Read depth</th>
<th>Split read</th>
<th>Assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deletion</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Read pair increased interpair mapping distance
- Read depth fewer reads
- Split read single read is “merged” from two segments surrounding deletion
- Assembly assembled sequence shows “gap”
## Insertions in WGS Data

What are the signals that let us detect an insertion?

<table>
<thead>
<tr>
<th>SV classes</th>
<th>Read pair</th>
<th>Read depth</th>
<th>Split read</th>
<th>Assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novel sequence insertion</td>
<td><img src="image1" alt="Diagram" /></td>
<td>Not applicable</td>
<td><img src="image2" alt="Diagram" /></td>
<td><img src="image3" alt="Diagram" /></td>
</tr>
</tbody>
</table>

**Table:**
- **SV classes**: Various types of structural variants, including insertions.
- **Read pair**: Diagram showing two paired sequences.
- **Read depth**: Indicates whether the depth of sequencing is applicable or not.
- **Split read**: Illustrates the splitting of reads due to insertions.
- **Assembly**: Diagram showing the assembly process with contigs/scaffolds.

**Diagram Notes:**
- Insertions in WGS data are detected through analysis of paired reads, read depths, and assembly processes.
- Novel sequence insertions are characterized by specific read pair and assembly patterns.
Insertions in WGS Data

<table>
<thead>
<tr>
<th>SV classes</th>
<th>Read pair</th>
<th>Read depth</th>
<th>Split read</th>
<th>Assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novel sequence insertion</td>
<td><img src="image1" alt="Read pair illustration" /></td>
<td><img src="image2" alt="Read depth illustration" /></td>
<td><img src="image3" alt="Split read illustration" /></td>
<td><img src="image4" alt="Assembly illustration" /></td>
</tr>
<tr>
<td></td>
<td>decreased interpair mapping distance</td>
<td>Not applicable</td>
<td>single read is split into two segments</td>
<td>Contig/ scaffold Assemble</td>
</tr>
<tr>
<td></td>
<td>not applicable</td>
<td></td>
<td>surrounding novel insertion sequence</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>assembled sequence with inserted novel sequence</td>
<td></td>
</tr>
</tbody>
</table>

3 Novel sequence will not map to genome
Inversions in WGS Data

What are the signals that let us detect an inversion?
## Inversions in WGS Data

<table>
<thead>
<tr>
<th>SV classes</th>
<th>Read pair</th>
<th>Read depth</th>
<th>Split read</th>
<th>Assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inversion</td>
<td><img src="image1" alt="" /></td>
<td>Not applicable</td>
<td><img src="image2" alt="" /></td>
<td><img src="image3" alt="" /></td>
</tr>
</tbody>
</table>

- **Read pair**: aberrant mapping (>---> instead of >---<) and interpair distance.
- **Read depth**: not applicable\(^4\).
- **Split read**: single read is split into two segments, one of which is inverted.
- **Assembly**: assembled sequence with inverted sequence.

\(^4\) Same amount of sequence
Duplications in WGS Data

What are the signals that let us detect a duplication?
Duplications in WGS Data

<table>
<thead>
<tr>
<th>SV classes</th>
<th>Read pair</th>
<th>Read depth</th>
<th>Split read</th>
<th>Assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tandem duplication</td>
<td>Read pair aberrant mapping (&lt;-----&gt; instead of &gt;---&lt; ) and interpair distance</td>
<td>Read depth increase</td>
<td>Split read single read is split into end of one duplicated block followed by beginning of next block</td>
<td>Assembly assembled sequence with duplicated sequence</td>
</tr>
</tbody>
</table>

What are the signals that let us detect a translocation?
In sum: There are many different signals that are used for SV detection. Different read types have distinct attributes.
In the remainder of this lecture, we will examine how read depth analysis can be used to search for CNVs. We will concentrate on three topics.

- Poisson distribution: Review
- G/C dependence
- Simplified version of algorithm in Yoon et al.\(^5\)

A Poisson experiment is a statistical experiment that has the following properties:

1. The experiment results in outcomes that can be classified as successes or failures.
2. The average number of successes ($\mu$) that occurs in a specified region is known.
3. The probability that a success will occur is proportional to the size of the region.
4. The probability that a success will occur in an extremely small region is virtually zero.

The “region” can be a length, an area, a volume, a period of time, etc.

Early use of Poisson distribution: Ladislaus Bortkiewicz (1898): investigation of the number of soldiers in the Prussian army killed accidentally by horse kick.
Poisson

\[ P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \] (1)

- \( k \) = number of occurrences
- \( \lambda \) = average occurrences/time interval

For example, if the average number of soldiers killed by being kicked by a horse each year in each of 14 cavalry corps is 1.7, what is the probability of 4 soldiers being killed in one year?

\[ P(X = 4) = \frac{(1.7)^4 e^{-(1.7)}}{4!} = 0.063 \] (2)

In R,

> dpois(4,1.7)
[1] 0.06357463
For $X \sim \text{Poisson}(\lambda)$, both the mean and the variance are equal to $\lambda$. 
Poisson and Read counts

Many NGS algorithms model read counts as a Poisson distribution

- Segment the genome into Windows (e.g., 1000 bp).
- Count number of reads in each Window
- All else equal, we expect half as many reads as normal in the case of a deletion, and 1.5 times as many reads as normal in the case of a duplication

\[
\lambda = \frac{NW}{G}
\]

where

\[
\begin{align*}
N & \quad \text{Total number of reads} \\
W & \quad \text{size of window} \\
G & \quad \text{Size of genome}
\end{align*}
\] (3)
The Poisson distribution can be derived as a limiting form of the binomial distribution in which n is increased without limit as the product $\lambda = np$ is kept constant.

- This corresponds to conducting a very large number of Bernoulli trials with the probability $p$ of success on any one trial being very small.
- This suggests we can approximate the Poisson distribution by the Normal distribution.

The central limit theorem: the mean of a sufficiently large number of independent random variables, each with finite mean and variance, is approximately normally distributed.
For sufficiently large values of $\lambda$, (say $\lambda > 1,000$), the Normal$(\mu = \lambda, \sigma = \sqrt{\lambda})$ Distribution is an excellent approximation to the Poisson$(\lambda)$ Distribution.

If $\lambda$ is greater than about 10, then the Normal Distribution is a good approximation if an appropriate continuity correction is performed.
Poisson and Normal Approximation

- $X \sim \text{Poisson}(\lambda = 20)$
- $P(X \geq 25) = 1 - P(X < 25) = 1 - \sum_{k=0}^{24} \frac{\lambda^k e^{-\lambda}}{k!}$
**Poisson and Normal Approximation**

- $X \sim \mathcal{N}(\mu = \lambda, \sigma = \sqrt{\lambda})$ for $\lambda = 20$

- $P(X \geq 25) = \int_{x=25}^{\infty} \frac{1}{2\sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{x-\lambda}{\sqrt{\lambda}} \right)^2} \, dx$
Finally, we can check in R that the Normal is a reasonable approximation to the Poisson (it is not an extremely close approximation for $\lambda$ in this range yet)\(^6\).

\[
\begin{align*}
> \text{pnorm}(25, \text{mean}=20, \text{sd} = \sqrt{20}, \text{lower.tail}=FALSE) \\
[1] 0.1317762 \\
> \text{ppois}(25, 20, \text{lower.tail}=FALSE) \\
[1] 0.112185
\end{align*}
\]

For this reason, we will see the Normal distribution (often a z-score) used to calculate read depth statistics.

\(^6\)It would be better for $\lambda = 50$ and better yet for $\lambda = 1000$ or above.
Poisson and Normal Approximation

Poisson

$$z = \frac{x - \mu}{\sigma}$$ (4)

graphic: wikipedia
The GC content \[
\frac{G + C}{A + C + G + T}
\] of a sequence affects many properties, e.g., annealing temperature of PCR primers.
GC Content in Bioinformatics

GC content is correlated with multiple other parameters, and bioinformatics analysis often needs to take this into account.

- ↑ GC content ⇔ ↑ mRNA stability
- Giemsa dark bands (cytogenetics) ⇔ locally GC-poor regions compared with light bands
- Housekeeping (ubiquitously expressed) genes in the mammal genome ⇔ on average slightly GC-richer than tissue-specific genes.
- Silent-site GC content correlates with gene expression efficiency in mammalian cells.

for instance...
GC Content in Genomics

- Dependence between fragment count (read coverage) and GC content found in Illumina sequencing data.
- The GC effect is unimodal: both GC-rich fragments and AT-rich fragments $\Leftrightarrow$ underrepresented.
- RNA-seq: GC-rich and GC-poor fragments tend to be under-represented in RNA-Seq, so that, within a lane, read counts are not directly comparable between genes.
- ChIP-seq: Peaks (profiles) correlate positively with genomic GC content.
- Whole genome sequencing: GC content may correlate positively with read depth.

Read Depth

We can get a simple picture of the distribution of reads across a chromosome by counting how many reads start in a given chromosomal window.

Basic workflow

- Align reads from high or low coverage genome sequencing
- Count the number of reads that begin in each window of size $N^7$
- Plot (eyeball-o-metrics)

There is a tutorial on how to do the next few analysis steps on the website.

---

$^7$The best size for $N$ will depend on the questions, the coverage, and the algorithm, but might be between 1000–100,000.
Read Depth

This is a typical plot showing the raw read depth following genome sequencing.

Thousand genomes project, individual HG00155, chromosome 11, low-coverage
Here, we have plotted read count vs. GC content. The loess-smoothed regression line is shown.

There is a clear, if complicated, relationship between GC-content and read depth in this sample.
With all of this in hand, we now will examine how to call CNVs from whole genome data. We will present a simplified version of Yoon S et al (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res* 9:1586-92.

1. Align whole-genome sequences (high-coverage)
2. Filter out reads with low mapping quality (PHRED < 30)
3. Count read depth in windows (100bp)
4. Adjust read-depth according to GC content of window
5. Calculate z-score for each window
6. Combine neighboring windows to maximize score
CNV Calling via Read Depth

Step 1-3. Alignment and raw read depth

- Align sequence reads to genome
- In the Yoon et al. paper, the MAQ aligner was used with default settings
- Filter out low quality reads (PHRED < 30)
- Segment genome into 100bp windows and count reads (by start position)
Step 4. GC adjustment

\[ \bar{r}_i = r_i \cdot \frac{m}{m_{GC}} \]

- adjusted read count
- raw read count
- median count for GC content
- overall median count per window
Step 5a. Event detection.

- A deletion or duplication is evident as a decrease or increase across multiple consecutive windows.
Step 5a. Event detection. The authors developed a heuristic they call Event-wise testing (EWT).

- Rapidly search across all windows for windows that meet criteria of statistical significance.
- Clusters of small events are grouped into larger events.
- Basic idea: Identify regions of consecutive 100-bp windows with significantly increased or decreased read depth ($\bar{r}_i$).
Step 5b. Convert to Z-Score

- Calculate mean ($\mu$) and standard deviation ($\sigma$) of $\hat{r}_i$ (adjusted read depth) across genome
- Transform the adjusted read depth into the corresponding Z-score

$$z_i = \frac{\bar{r}_i - \mu}{\sigma}$$  \hspace{1cm} (5)
Step 5b. Convert to upper and low tail probabilities

- Convert the Z-score to its upper-tail probability
  \[ P_i^{\text{Upper}} = \Pr(Z > z_i) \]  

- This is simply the probability that the read count in window \( i \) is at least as high as observed

- Analogously, we calculate a lower-tail probability
  \[ P_i^{\text{Lower}} = \Pr(Z < z_i) \]
CNV Calling via Read Depth

Step 5c. Evaluate intervals of consecutive windows

- For an interval $\mathcal{A}$ of $\ell$ consecutive windows, we call it an unusual event if (for duplications)

$$
\max_i \left\{ P_i^{\text{Upper}} \mid i \in \mathcal{A} \right\} < \left( \frac{\ell}{L} \times \text{FPR} \right)^{\frac{1}{\ell}}
$$

(8)

- Here, $L$ is the length in windows of the entire chromosome
- Thus $\frac{\ell}{L}$ is the proportion of the chromosome that is taken up by the candidate CNV
- If all $p$-values for the windows of $\mathcal{A}$ are less (more significant) than the term on the right side, we call the interval an “unusual event”
Step 5c. The EWT score \( \left( \frac{\ell}{L} \times \text{FPR} \right)^{\frac{1}{\ell}} \) increases as the number of windows, \( \ell \), increases.
Step 5c. Iteration

- The initial analysis calculates the p-values for each single window.
- The EWT procedure then searches for two-window intervals (i.e., $\ell = 2$) such that

$$\max_i \left\{ P_i^{\text{Upper}} \mid i \in \mathcal{A} \right\} < \left( \frac{\ell}{L} \times \text{FPR} \right)^{\frac{1}{\ell}}.$$  

- Continue iterating (increasing the size of $\ell$ by 1) as long as this condition is fulfilled.
Step 5d. Deletion

The same procedure is now done separately for deletions, using the formula

$$\max_i \{ P_{i, \text{Lower}} | i \in A \} < \left( \frac{\ell}{L} \times \text{FPR} \right)^{\frac{1}{\ell}}$$  \hspace{1cm} (9)
CNV Calling via Read Depth

Correction for multiple testing

- We are making millions of tests across the genome . . .
The *p*-value is the probability, under the null hypothesis, that the test statistic assumes the observed or a more extreme value. It is important to realize that if we go on testing long enough, we will inevitably find something which is “significant” by chance alone.

- If we test a null hypothesis that is true using a significance level of $\alpha = 0.05$, then there is a probability of $1 - \alpha = 0.95$ of arriving at a correct conclusion of non-significance.
- If we now test two independent true null hypotheses, the probability that neither test is significant is $0.95 \times 0.95 = 0.90$. 
You can see where this is leading...

- If we test 20 independent null hypotheses, the probability that none will be significant is then \((0.95)^{20} = 0.36\).

- This corresponds to a probability of \(1 - 0.36 = 0.64\) of getting at least one spurious significant result, and the expected number of spurious significant results in 20 tests is \(20 \times 0.05 = 1\)