

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

Genome Sequencing and Structural Variation

Peter N. Robinson

Institut für Medizinische Genetik und Humangenetik
Charité Universitätsmedizin Berlin

Genomics: Lecture #10

Today

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

- 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

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

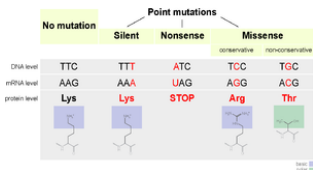
- 1 **Structural variants**
- 2 Array CGH
- 3 Bioinformatics Approaches for Structural Variant Discovery
- 4 Poisson
- 5 GC Content
- 6 Read depth
- 7 CNV Calling

CNVs vs. SNVs

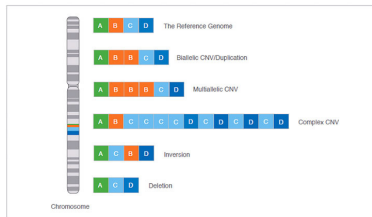
WGS & SVs

Peter N.
Robinson

Single-nucleotide variants



CNV



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

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

CNVs vs. SNVs

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

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

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

Structural variations (SV) are Genomic rearrangements that effect more than 1 Kb¹

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

¹Yes, this definition is arbitrary!

Inversion

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

Allele A



Allele B



- 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

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

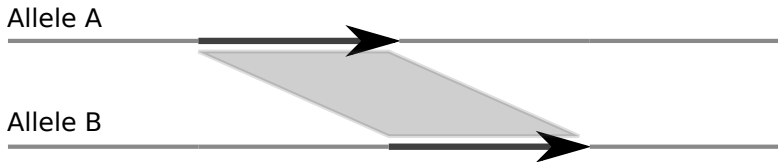
SV Discovery

Poisson

GC Content

Read depth

CNV Calling



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

Interchromosomal translocation

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

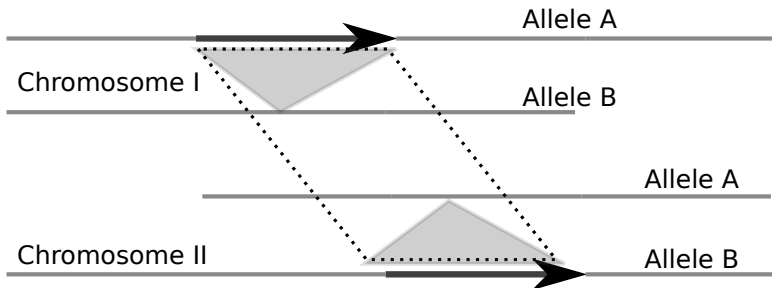
SV Discovery

Poisson

GC Content

Read depth

CNV Calling



- 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

Deletion

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

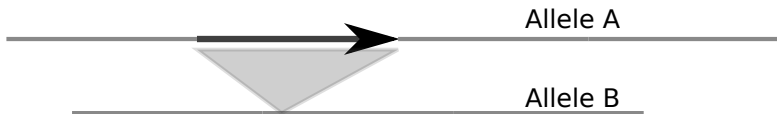
SV Discovery

Poisson

GC Content

Read depth

CNV Calling



- An **unbalanced** structural variation (loss of genomic segment)
- results in dosage abnormality of genes contained in deletion
- Indirect regulatory imbalances also possible

Duplication

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

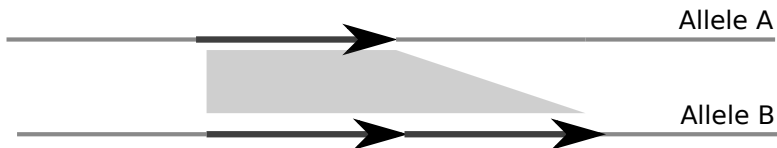
SV Discovery

Poisson

GC Content

Read depth

CNV Calling



- 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

WGS & SVs

Peter N. Robinson

Structural variants

Array CGH

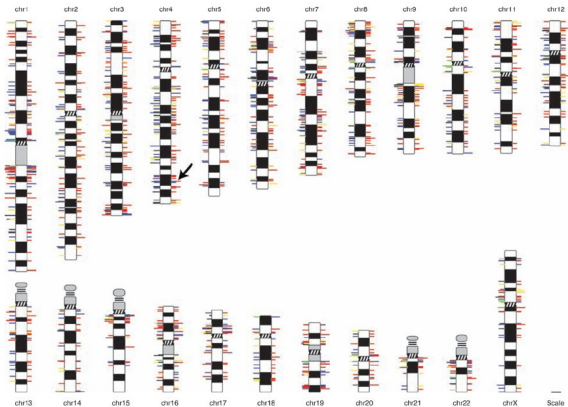
SV Discovery

Poisson

GC Content

Read depth

CNV Calling



~ 1000 SVs >2.5kb per Person

Korbel JO et al (2007) Paired-end mapping reveals extensive structural variation in the human genome.

Science 318:420–6.

Detection of Structural Variants

WGS & SVs

Peter N. Robinson

Structural variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

	Techniques	Detection						Maximum resolution	Sensitivity
		Copy-neutral events							
		Deletions and duplications	Insertions	Unbalanced translocations	Balanced translocations	Inversions	LOH and UPD		
Early 1970s	Karyotyping/G-banding	Yes	Yes	Yes	Yes	Yes	No	Low (>several Mb)	Low
	<i>FISH-based</i>								
Early 1990s	CGH	Yes	No	Yes	No	No	No	Low (>several Mb)	High
Mid 1990s	M-FISH/SKY/COBRA	Yes	Yes	Yes	Yes	No	No	Low (>several Mb)	High
Late 1990s	RxFISH	Yes	Yes	Yes	Yes	Yes	No	Low (>several Mb)	High
	<i>Array-based</i>								
Early 2000s	1-Mb BAC array-CGH	Yes	No	Yes	No	No	No	Average (> 1 Mb)	High
	Tiling-path BAC array-CGH	Yes	No	Yes	No	No	No	High (>50–100 kb)	High
	Oligonucleotide array-CGH	Yes	No	Yes	No	No	No	High (catalogue > 1 kb, custom > 400 bp)	Very high
Late 2000s	SNP arrays	Yes	No	Yes	No	No	Yes	High (> 5–10 kb)	High
	<i>NGS-based</i>	Yes	Yes	Yes	Yes	Yes	Yes	Very high (bp level)	Very high

Abbreviations: BAC, bacterial artificial chromosome; CGH, comparative genomic hybridisation; COBRA, combined binary ratio labelling; FISH, fluorescence *in situ* hybridisation; LOH, loss of heterozygosity; M-FISH, multiplex FISH; NGS, next-generation sequencing; RxFISH, Rainbow cross-species FISH or cross-species colour banding; SNP, single-nucleotide polymorphism; SKY, spectral karyotyping; UPD, uniparental disomy.
Methods in the grey-shaded area are discussed in this review.

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

Le Scouarnec S, Gribble SM (2012) Characterising chromosome rearrangements: recent technical advances in molecular cytogenetics. *Heredity (Edinb)* **108**:75–85.

Array-CGH

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

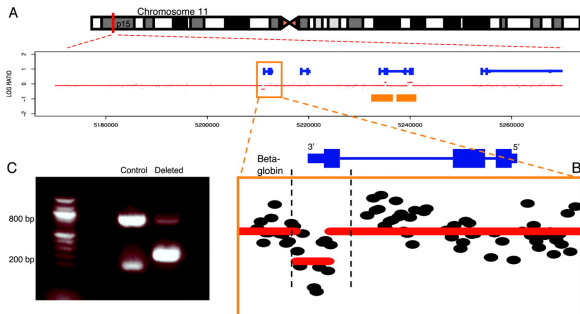
SV Discovery

Poisson

GC Content

Read depth

CNV Calling



A small heterozygous deletion in the β -globin locus.

Urban AE et al. (2006) High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. *Proc Natl Acad Sci U S A*. **103**:4534-9.

DNA Hybridization

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

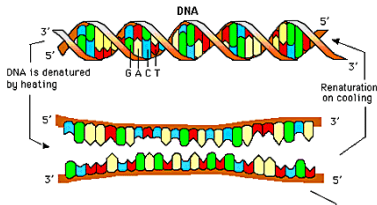
SV Discovery

Poisson

GC Content

Read depth

CNV Calling



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

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

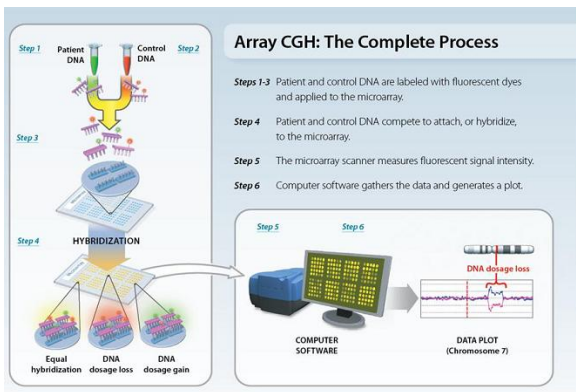
SV Discovery

Poisson

GC Content

Read depth

CNV Calling



- Ratio of 2 fluorescent signals indicates loss or gain of DNA segment

Array-CGH

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

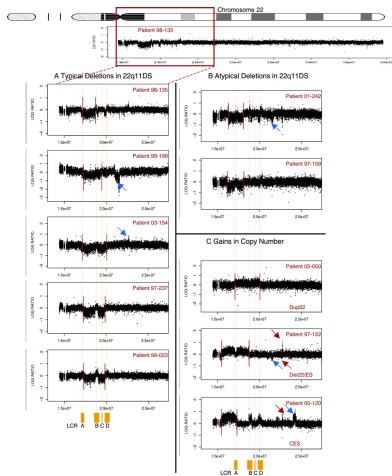
GC Content

Read depth

CNV Calling

Array CGH can detect

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



Urban AE et al. (2006) High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. *Proc Natl Acad Sci U S A.* **103**:4534-9.

Array-CGH: Indications in Human Genetics

WGS & SVs

Peter N.
Robinson

Structural variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

- 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 un targeted fashion. Many findings are “new” and may be difficult to interpret: cause of a disease or neutral polymorphism?

Outline

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

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

Bioinformatics Approaches for SV Discovery with WGS data

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

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 NGS Reads

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

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 libraries² are sequenced from both ends. The sequencing direction is from the ends towards the middle.
- 2 **Mate-pair** libraries. We will review this today

²As discussed in the very first lecture.

Mate pair

WGS & SVs

Peter N.
Robinson

Structural variants

Array CGH

SV Discovery

Poisson

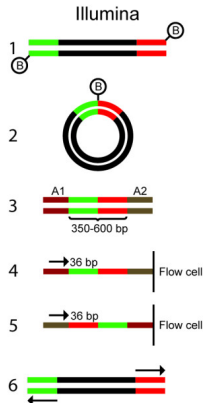
GC Content

Read depth

CNV Calling

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
- 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.



Berglund EC et al. (2011). *Investig Genet* 2:23.

Paired-end vs. Mate pair

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

	Paired-end	Mate pair
insert size	≈ 250 bp	2–20 kb
DNA	1.5–5 μg	5–120 μg
lab work	easier	harder
Costs	less	more

Note:

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

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

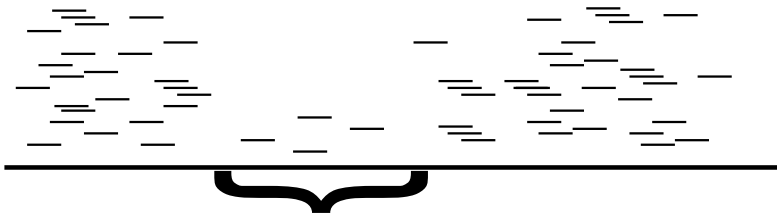
$2 \times 75 + 100 = 250$ nt. The fragment size is insert size plus length of both adapters (≈ 120 nt extra).

Read depth

WGS & SVs

Peter N.
Robinson

Analysis of read depth can identify deletion/duplications



Heterozygous Deletion?
Mappability Issue?
Poor "sequencability"?

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

Read depth

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

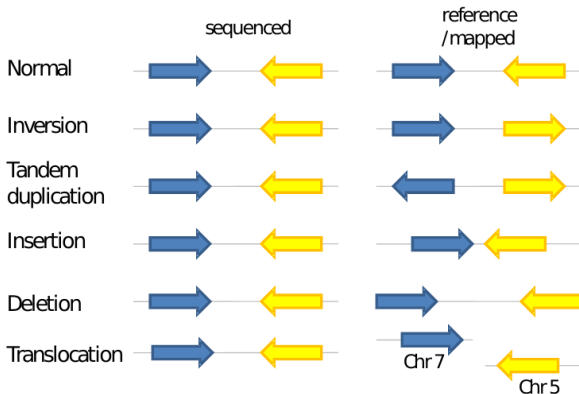
Poisson

GC Content

Read depth

CNV Calling

Characteristic signatures of paired-end sequences



graphic credit: Victor Guryev

Deletions in WGS Data

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

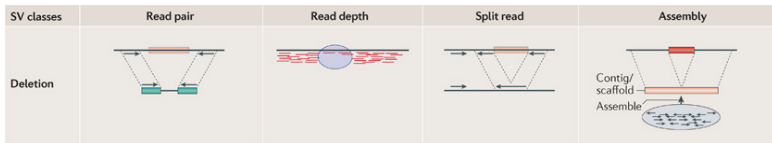
SV Discovery

Poisson

GC Content

Read depth

CNV Calling



What are the signals that let us detect a deletion?

Deletions in WGS Data

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

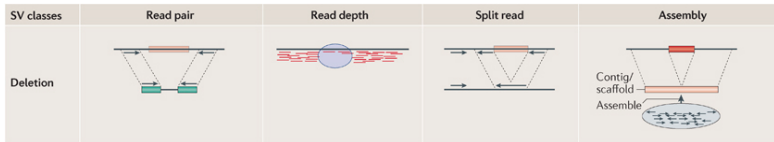
SV Discovery

Poisson

GC Content

Read depth

CNV Calling



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

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH



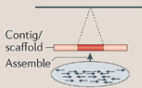
SV Discovery

Poisson

GC Content

Read depth

CNV Calling

SV classes	Read pair	Read depth	Split read	Assembly
Novel sequence insertion		Not applicable		

What are the signals that let us detect a insertion?

Insertions in WGS Data

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH


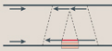
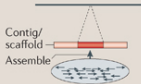
SV Discovery

Poisson

GC Content

Read depth

CNV Calling

SV classes	Read pair	Read depth	Split read	Assembly
Novel sequence insertion		Not applicable		

Read pair decreased interpair mapping distance
Read depth not applicable³
Split read single read is split into two segments
 surrounding novel insertion sequence
Assembly assembled sequence with inserted
 novel sequence

³ Novel sequence will not map to genome

Inversions in WGS Data

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

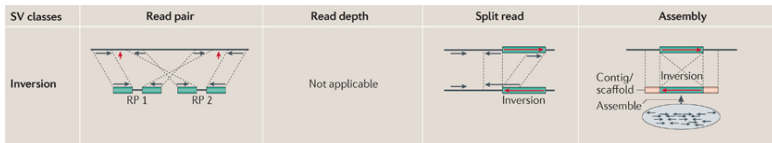
SV Discovery

Poisson

GC Content

Read depth

CNV Calling



What are the signals that let us detect a inversion?

Inversions in WGS Data

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

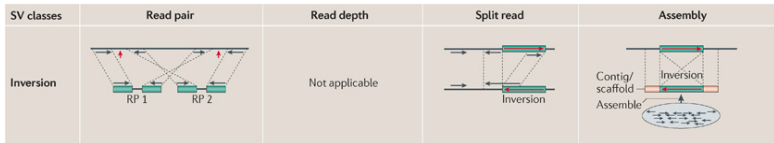
SV Discovery

Poisson

GC Content

Read depth

CNV Calling



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 se- quence

⁴ Same amount of sequence

Duplications in WGS Data

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

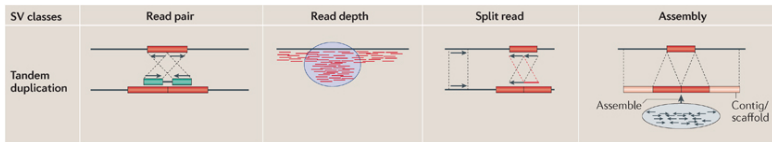
SV Discovery

Poisson

GC Content

Read depth

CNV Calling



What are the signals that let us detect a duplication?

Duplications in WGS Data

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

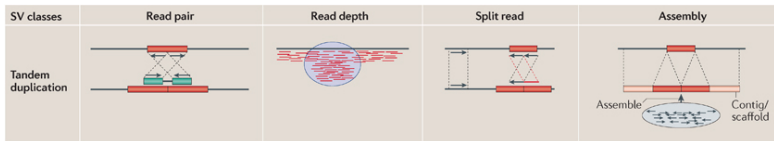
SV Discovery

Poisson

GC Content

Read depth

CNV Calling



Read pair aberrant mapping ($\leftarrow\text{---}\rightarrow$ instead of >---<) and interpair distance

Read depth increased

Split read single read is split into end of one duplicated block followed by beginning of next block

Assembly assembled sequence with duplicated sequence

Graphics credit: Le Scouarnec and Gribble SM *Heredity (Edinb)*. 2012; **108**:75-85.

Translocations in WGS Data

WGS & SVs

Peter N. Robinson

Structural variants

Array CGH

SV Discovery

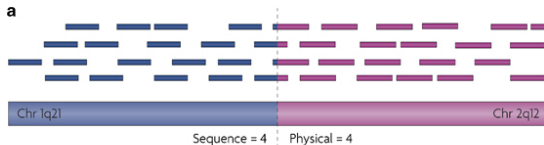
Poisson

GC Content

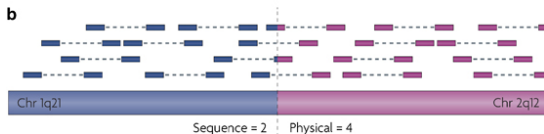
Read depth

CNV Calling

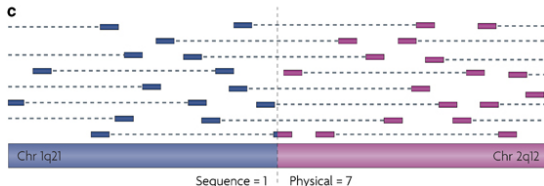
a) single-end sequencing



b) paired end (short insert library)



c) mate-pair (large insert library)



What are the signals that let us detect a translocation?

Signals and Read Types

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

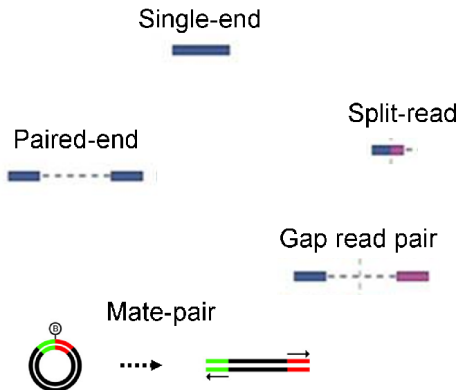
SV Discovery

Poisson

GC Content

Read depth

CNV Calling



- In sum: There are many different signals that are used for SV detection. Different read types have distinct attributes

Read depth

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

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

⁵Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.*

Poisson

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

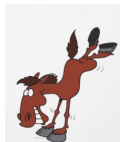
CNV Calling

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 (μ) 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

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

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

- k = number of occurrences
- λ = 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?

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

In R,

```
> dpois(4,1.7)
[1] 0.06357463
```

Poisson

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

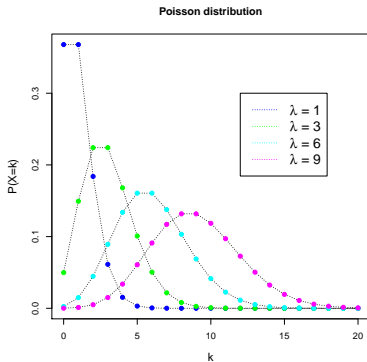
SV Discovery

Poisson

GC Content

Read depth

CNV Calling



- For $X \sim \text{Poisson}(\lambda)$, both the mean and the variance are equal to λ

Poisson and Read counts

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

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

$$\lambda = \frac{NW}{G} \quad \text{where} \quad \begin{cases} N & \text{Total number of reads} \\ W & \text{size of window} \\ G & \text{Size of genome} \end{cases} \quad (3)$$

Poisson and Normal Approximation

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

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

Poisson and Normal Approximation

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

For sufficiently large values of λ , (say $\lambda > 1,000$), the Normal($\mu = \lambda, \sigma = \sqrt{\lambda}$) Distribution is an excellent approximation to the Poisson(λ) Distribution.

If λ is greater than about 10, then the Normal Distribution is a good approximation if an appropriate continuity correction is performed.

Poisson and Normal Approximation

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

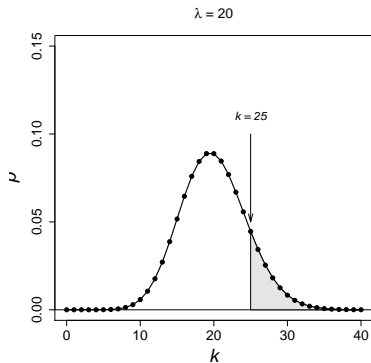
SV Discovery

Poisson

GC Content

Read depth

CNV Calling



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

Poisson and Normal Approximation

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

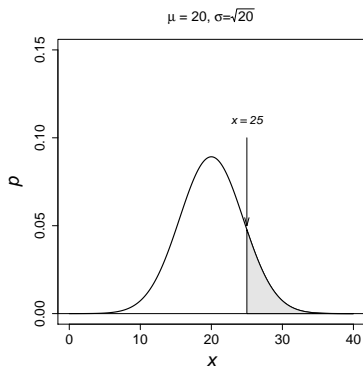
SV Discovery

Poisson

GC Content

Read depth

CNV Calling



- $X \sim \mathcal{N}(\mu = \lambda, \sigma = \sqrt{\lambda})$ for $\lambda = 20$
- $P(X \geq 25) = \int_{x=25}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\lambda}{\sqrt{\lambda}}\right)^2} dx$

Poisson and Normal Approximation

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

- Finally, we can check in R that the Normal is a reasonable approximation to the Poisson (it is not an extremely close approximation for λ in this range yet)⁶.

```
> pnorm(25,mean=20,sd=sqrt(20),lower.tail=FALSE)
[1] 0.1317762
> ppois(25,20,lower.tail=FALSE)
[1] 0.112185
```

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

⁶It would be better for $\lambda = 50$ and better yet for $\lambda = 1000$ or above. ↻ 🔍

Poisson and Normal Approximation

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

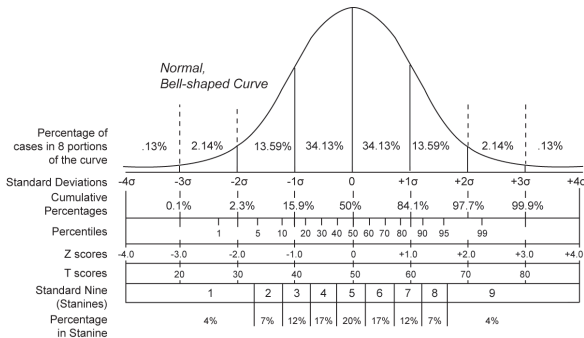
SV Discovery

Poisson

GC Content

Read depth

CNV Calling



z-score

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

graphic: wikipedia

GC Content

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

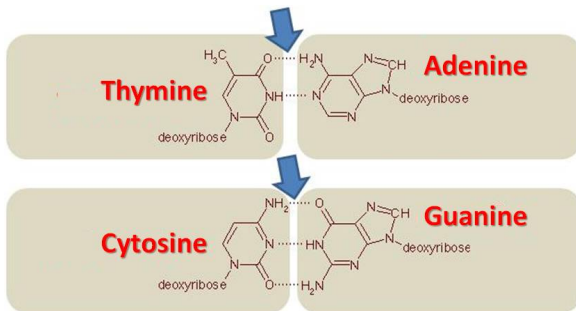
SV Discovery

Poisson

GC Content

Read depth

CNV Calling



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

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

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

- \uparrow GC content \Leftrightarrow \uparrow mRNA stability
- Giemsa dark bands (cytogenetics) \Leftrightarrow locally GC-poor regions compared with light bands
- Housekeeping (ubiquitously expressed) genes in the mammal genome \Leftrightarrow 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

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

GC content is can confound the results of a number of genomics experiments

- 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

See for instance: Benjamini Y, Speed TP (2012) Summarizing and correcting the GC content bias in

high-throughput sequencing. *Nucleic Acids Res* 40:e72.

Read Depth

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth


CNV Calling

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.

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

Read Depth

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

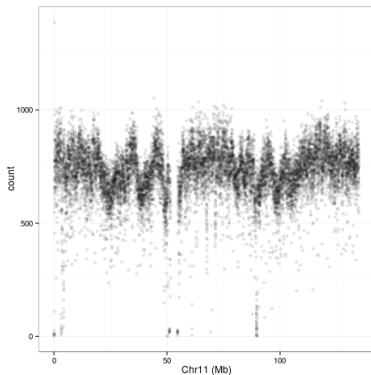
GC Content

Read depth

CNV Calling

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

Thousand genomes project, individual HG00155, chromosome 11, low-coverage



GC content vs. Read Depth

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

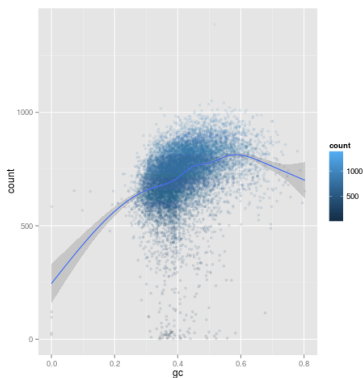
GC Content

Read depth

CNV Calling

Here, we have plotted read count vs. GC content

loess-smoothed regression line is shown



There is a clear, if complicated, relationship between GC-content and read depth in this sample.

CNV Calling via Read Depth

WGS & SVs

Peter N.
Robinson

Structural variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

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

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

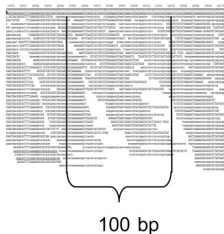
GC Content

Read depth

CNV Calling

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)



CNV Calling via Read Depth

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

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

CNV Calling via Read Depth

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

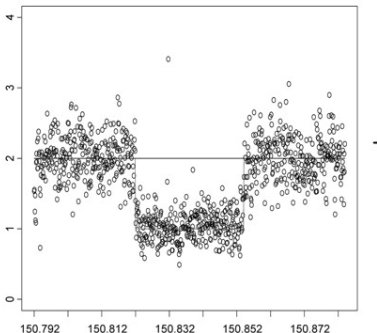
GC Content

Read depth

CNV Calling

Step 5a. Event detection.

- A deletion or duplication is evident as a decrease or increase across multiple consecutive windows



CNV Calling via Read Depth

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

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).

CNV Calling via Read Depth

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

Step 5b. Convert to Z-Score

- Calculate mean (μ) and standard deviation (σ) of \hat{r}_i (adjusted read depth) across genome
- transform the adjusted read depth into the corresponding Z-score

$$z_i = \frac{\bar{r}_i - \mu}{\sigma} \quad (5)$$

CNV Calling via Read Depth

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

Step 5b. Convert to upper and low tail probabilities

- Convert the Z-score to its upper-tail probability

$$P_i^{\text{Upper}} = \Pr(Z > z_i) \quad (6)$$

- 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

$$P_i^{\text{Lower}} = \Pr(Z < z_i) \quad (7)$$

CNV Calling via Read Depth

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

Step 5c. Evaluate intervals of consecutive windows

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

$$\max_i \{P_i^{\text{Upper}} | i \in \mathcal{A}\} < \left(\frac{\ell}{L} \times \text{FPR}\right)^{\frac{1}{\ell}} \quad (8)$$

- 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”

CNV Calling via Read Depth

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

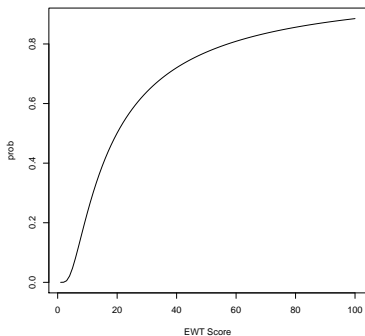
Poisson

GC Content

Read depth

CNV Calling

Step 5c. The EWT score $\left(\frac{\ell}{L} \times \text{FPR}\right)^{\frac{1}{\ell}}$ increases as the number of windows, ℓ , increases



CNV Calling via Read Depth

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

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 ℓ by 1) as long as this condition is fulfilled.

CNV Calling via Read Depth

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

Step 5d. Deletion

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

$$\max_i \{P_i^{\text{Lower}} | i \in \mathcal{A}\} < \left(\frac{\ell}{L} \times \text{FPR}\right)^{\frac{1}{\ell}} \quad (9)$$

CNV Calling via Read Depth

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

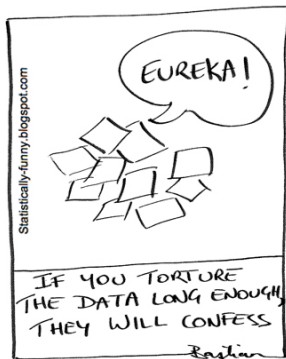
GC Content

Read depth

CNV Calling

Correction for multiple testing

- We are making millions of tests across the genome ...



CNV Calling via Read Depth

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

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 $0.95 \times 0.95 = 0.90$.

CNV Calling via Read Depth

WGS & SVs

Peter N.
Robinson

Structural
variants

Array CGH

SV Discovery

Poisson

GC Content

Read depth

CNV Calling

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 $20 \times 0.05 = 1$