Genome Sequencing and Structural Variation

Peter N. Robinson

Institut für Medizinische Genetik und Humangenetik
Charité Universitätsmedizin Berlin

Genomics: Lecture #10

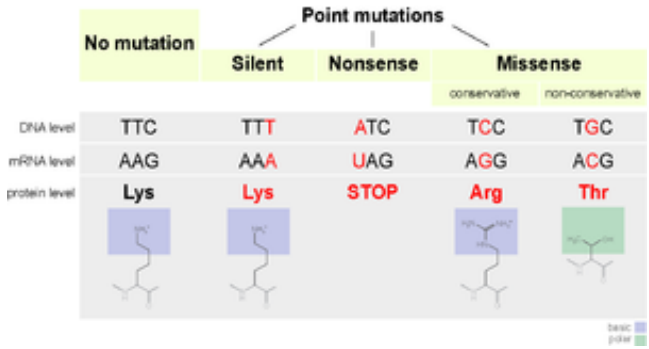
Today

- 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

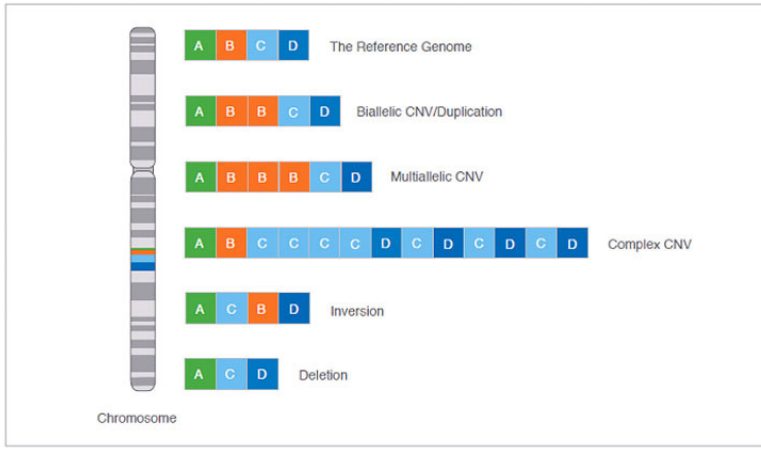
CNVs vs. SNVs

Single-nucleotide variants



- Several thousand SNVs in typical exome (1% des Genoms)
- 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 subsitutions and intergenic subsitutions
- 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.)

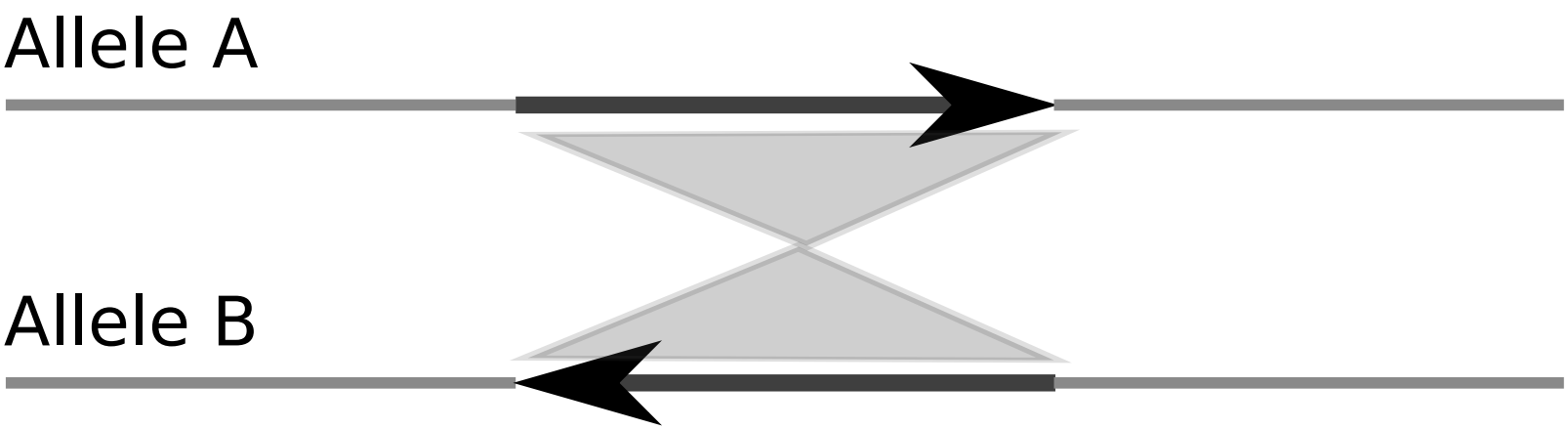
Structural Variation: Definition

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

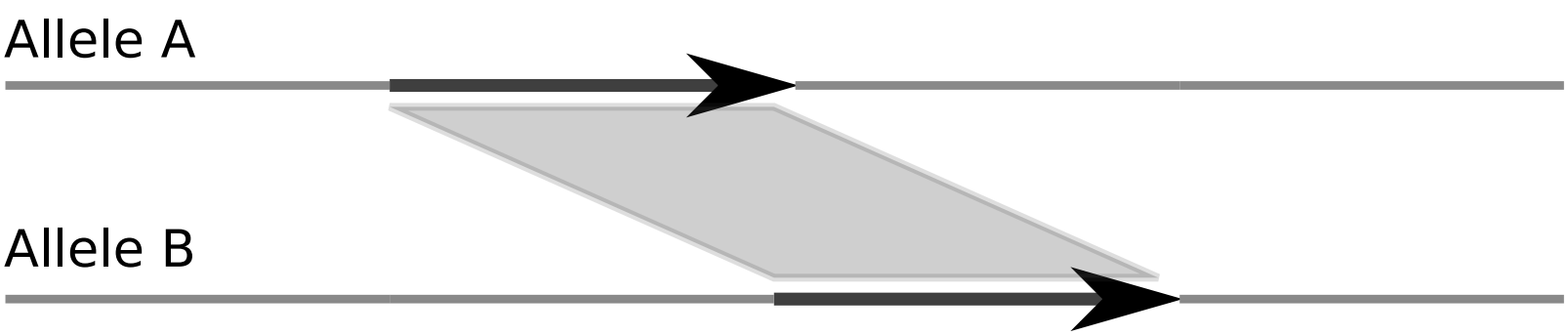
[1] Yes, this definition is arbitrary!

Inversion



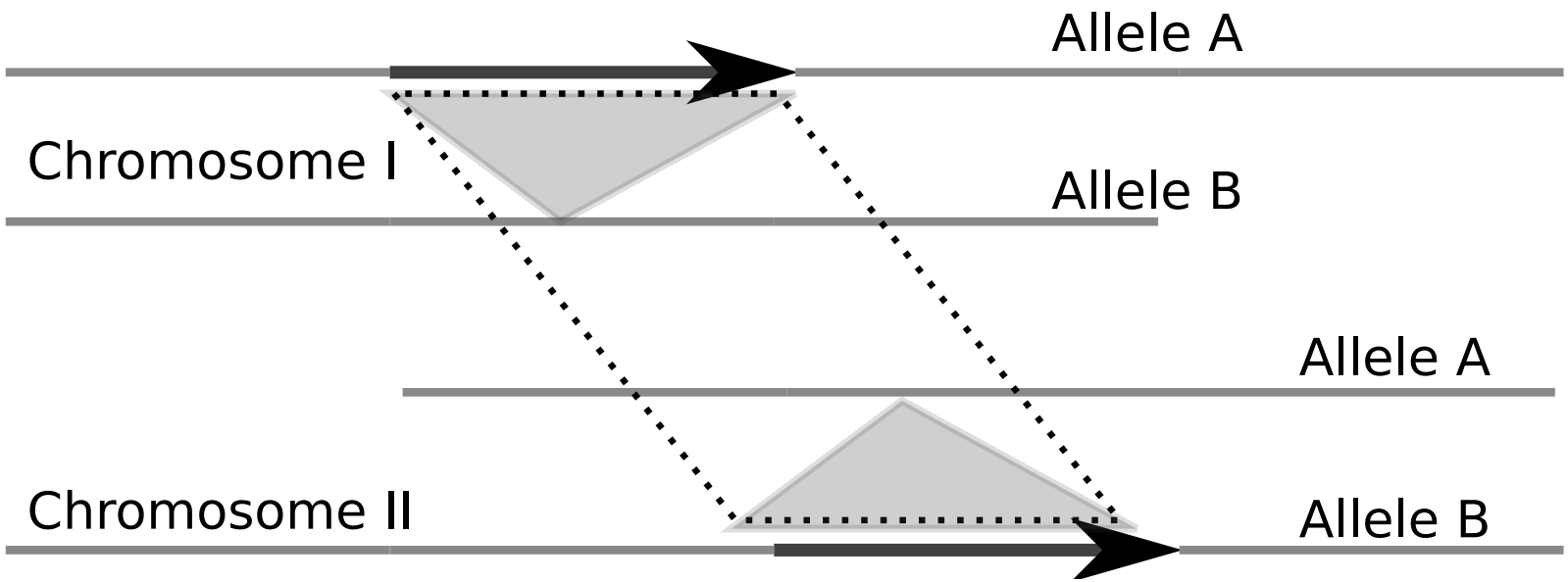
- A balanced structural variation (no loss/gain of genomic segment)
- Can be a neutral variation
- Can disrupt a coding sequence
- Can interrupt regulatory interactions

Intrachromosomal translocation



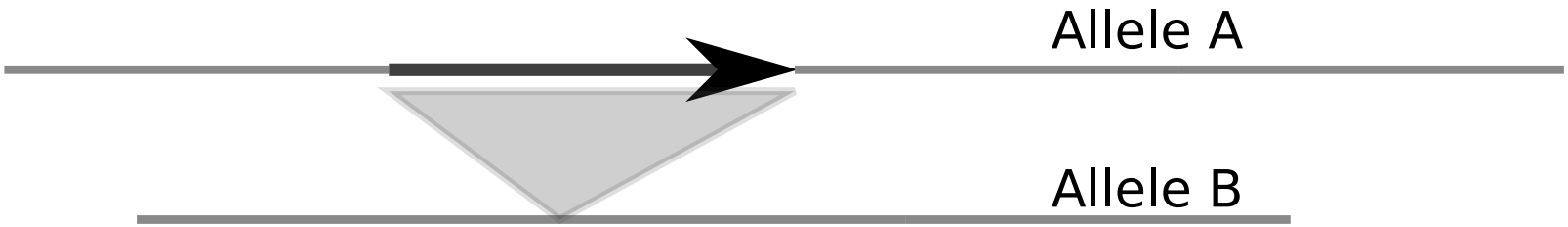
- A balanced structural variation (no loss/gain of genomic segment)
- Can be a neutral variation
- Can disrupt a coding sequence
- Can interrupt regulatory interactions

Interchromosomal translocation



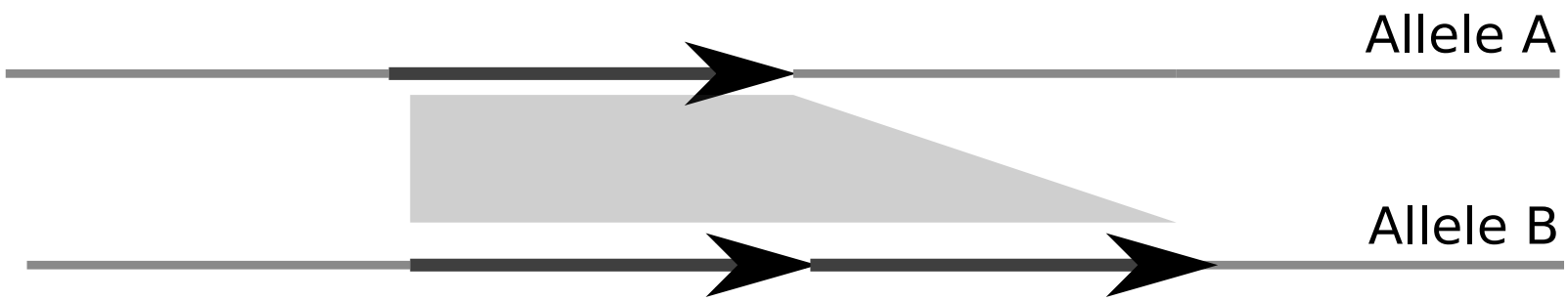
- A balanced structural variation (no loss/gain of genomic segment)
- Translocation between two different chromosomes
- Like other balanced SVs, can be neutral of disrupt coding sequences or regulatory interactions

Deletion



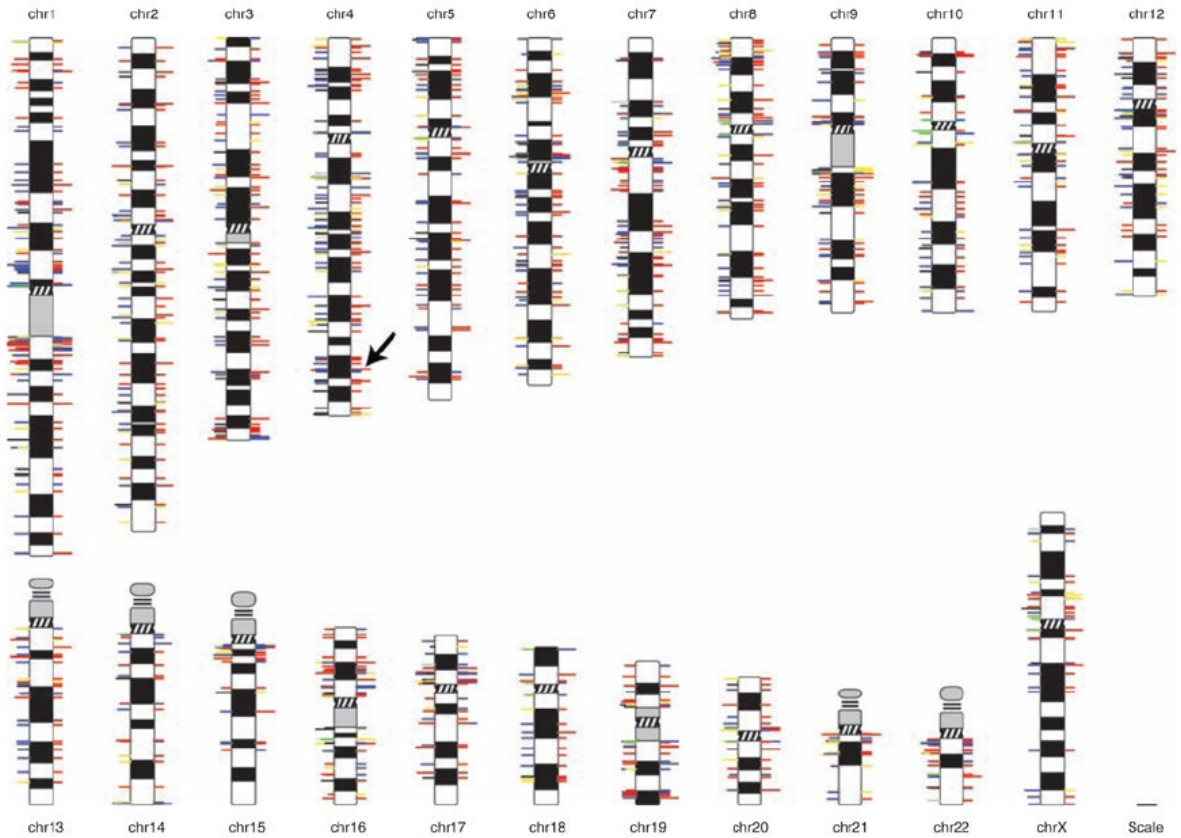
- An **unbalanced** structural variation (loss of genomic segment)
- results in dosage abnormality of genes contained in deletion
- Indirect regulatory imbalances also possible

Duplication



- 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



$\sim$ 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

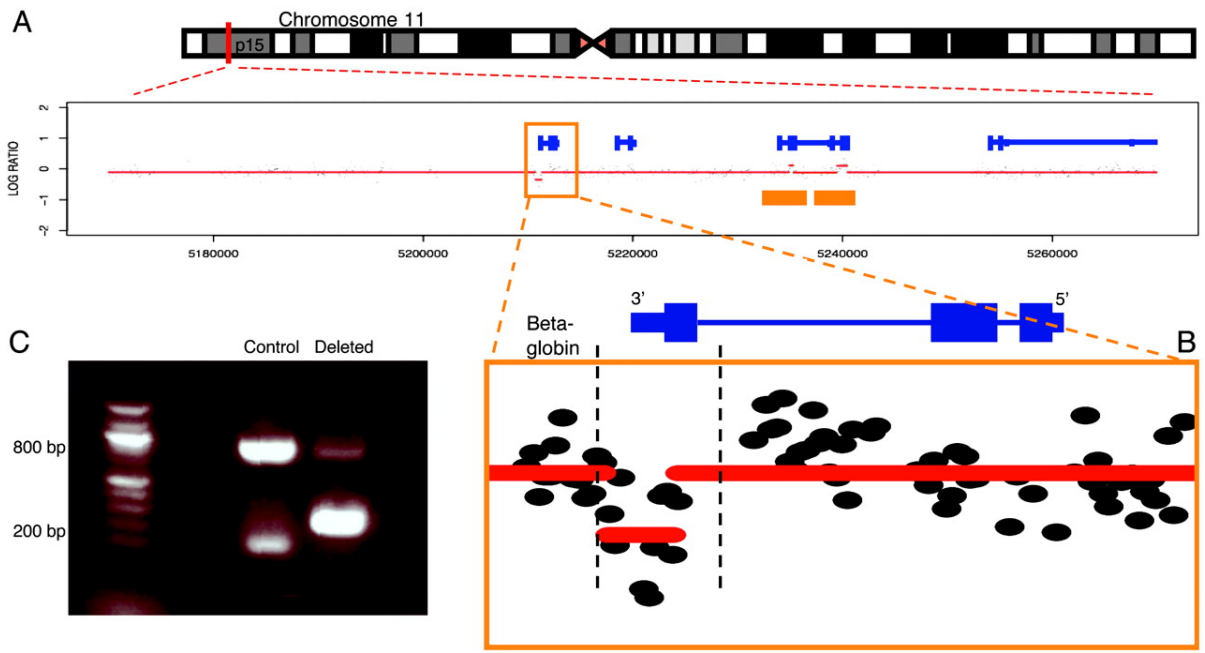
Techniques		Detection						Maximum resolution	Sensitivity
					Copy-neutral events				
		Deletions and duplications	Insertions	Unbalanced translocations	Balanced translocations	Inversions	LOH and UPD		
Early 1970s	Karyotyping/G-banding	Yes	Yes	Yes	Yes	Yes	No	Low (>several Mb)	Low
	FISH-based								
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
	Array-based								
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
	NGS-based	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 heterozyogosity; 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.

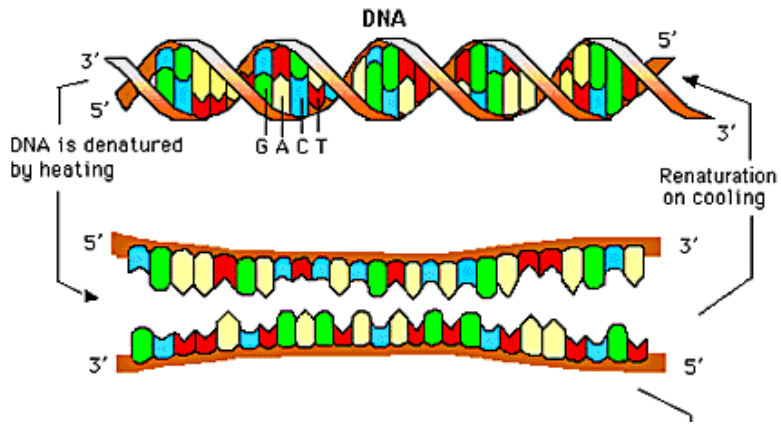
Array-CGH



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.

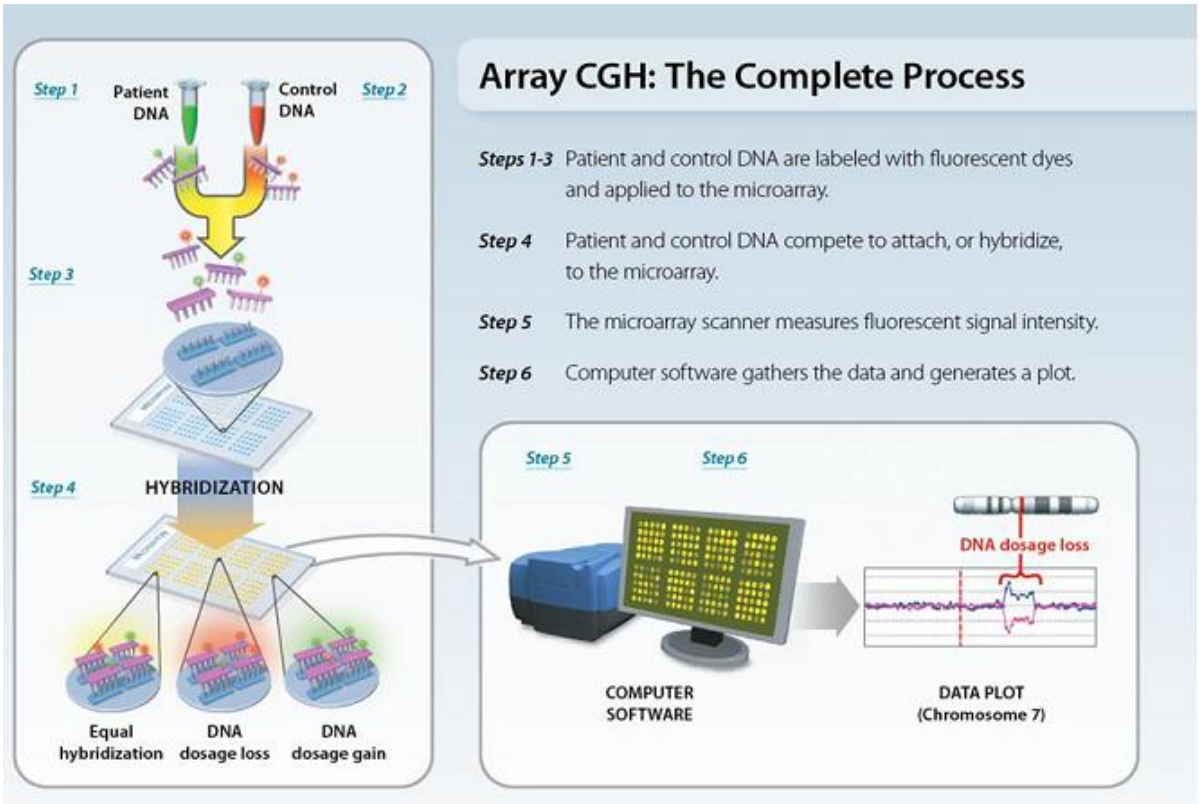
DNA Hybridization



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

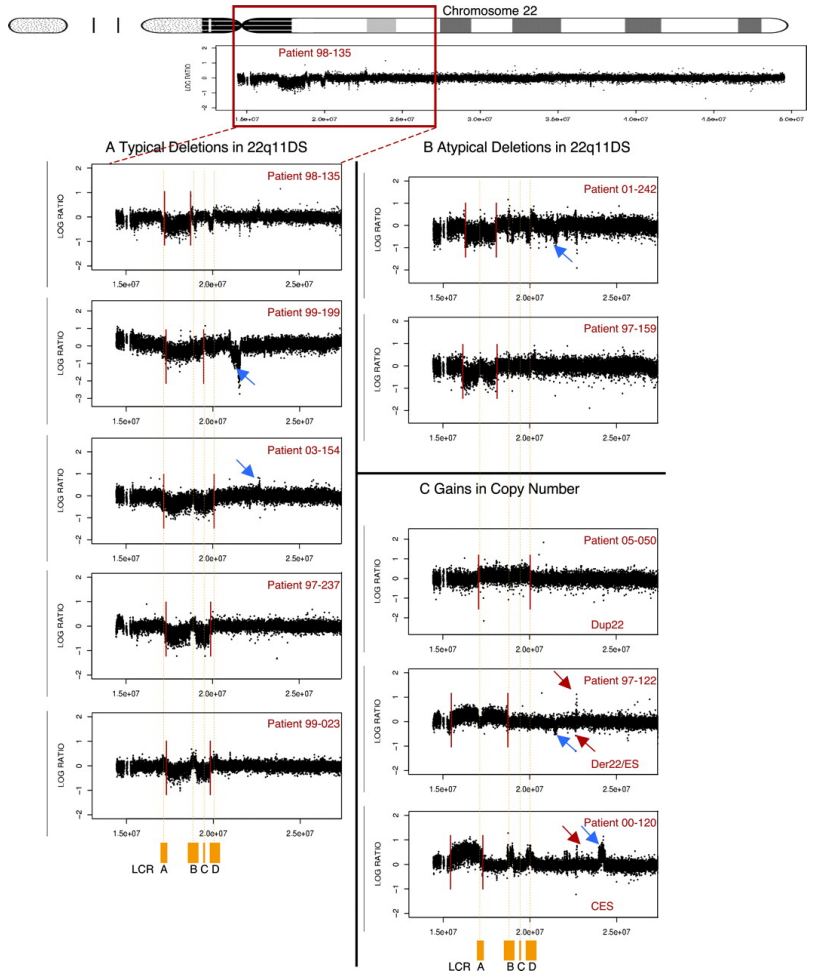


- Ratio of 2 fluorescent signals indications loss or gain of DNA segment

Array-CGH

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

- 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

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 NGS Reads

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.

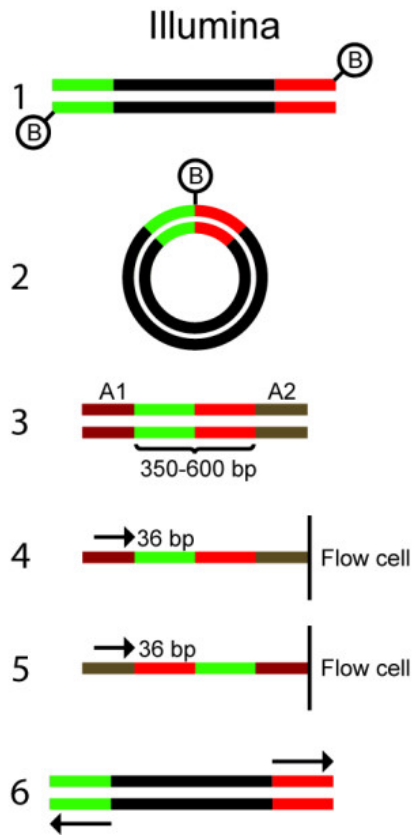
1. Paired end. Fragment libraries[2] are sequenced from both ends. The sequencing direction is from the ends towards the middle.
2. Mate-pair libraries. We will review this today

[2] As discussed in the very first lecture.

Mate pair

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.



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

Paired-end vs. Mate pair

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

Note:

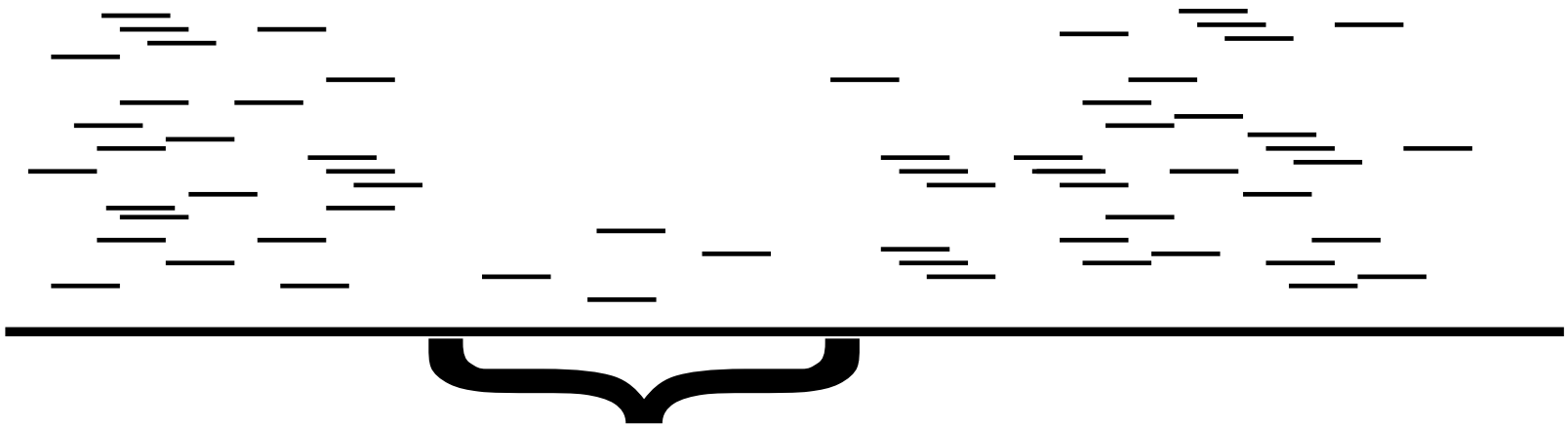
```
|-----75----|----------------------100----------------|-----75-----|
```

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 ($\approx$ 120 nt extra).

Read depth

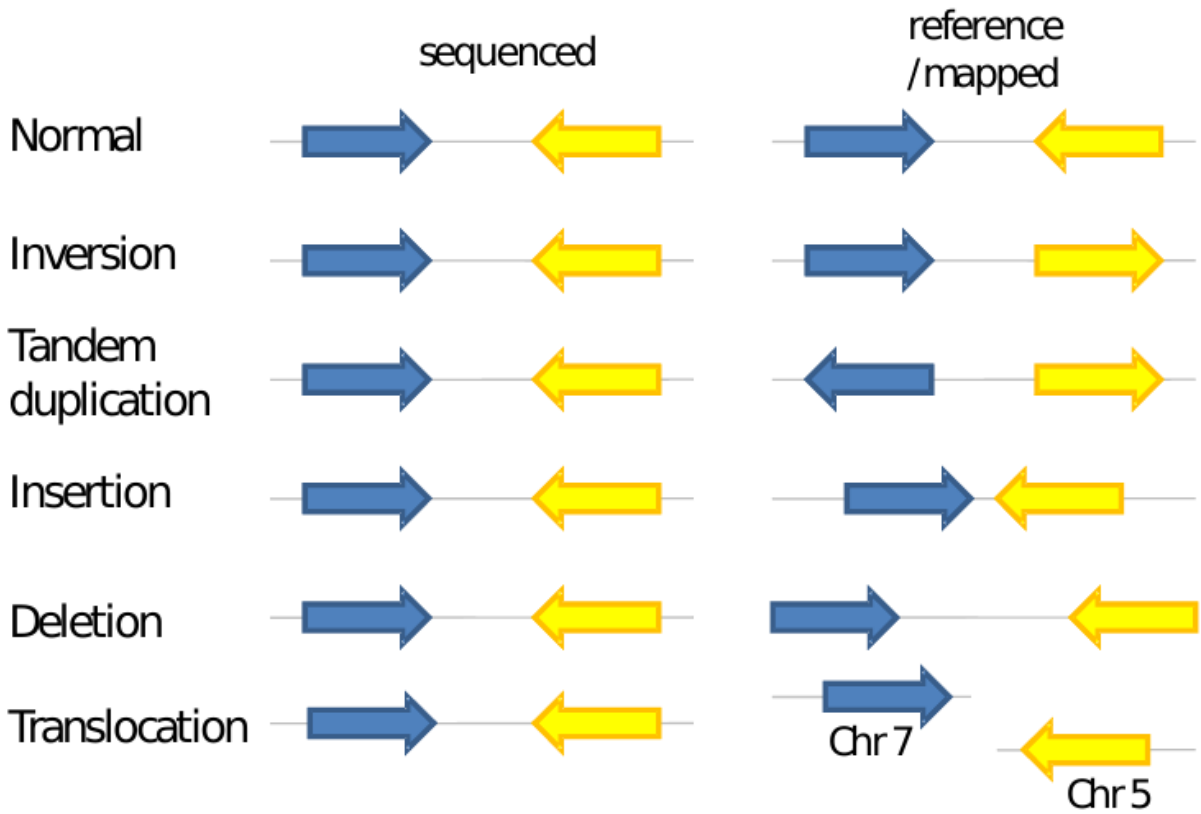
Analysis of read depth can identify deletion/duplications



Heterozygous Deletion?
Mappability Issue?
Poor "sequencability"?

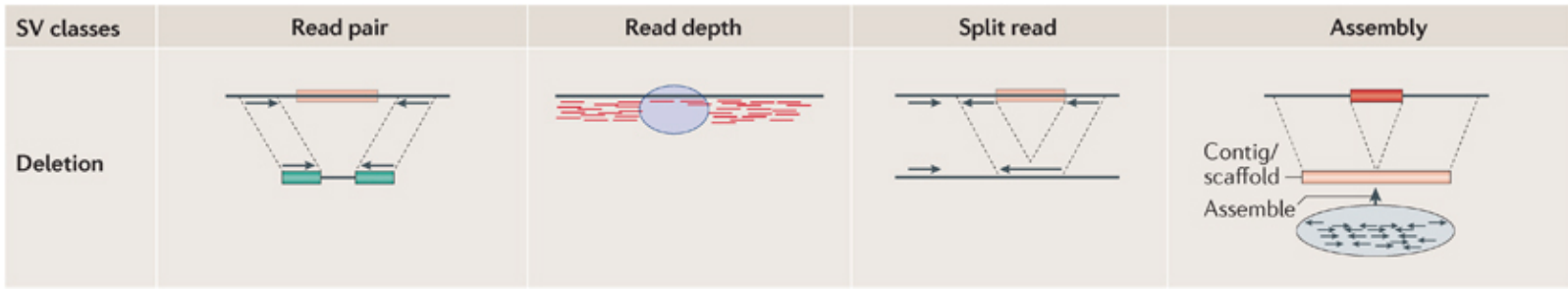
Read depth

Characteristic signatures of paired-end sequences



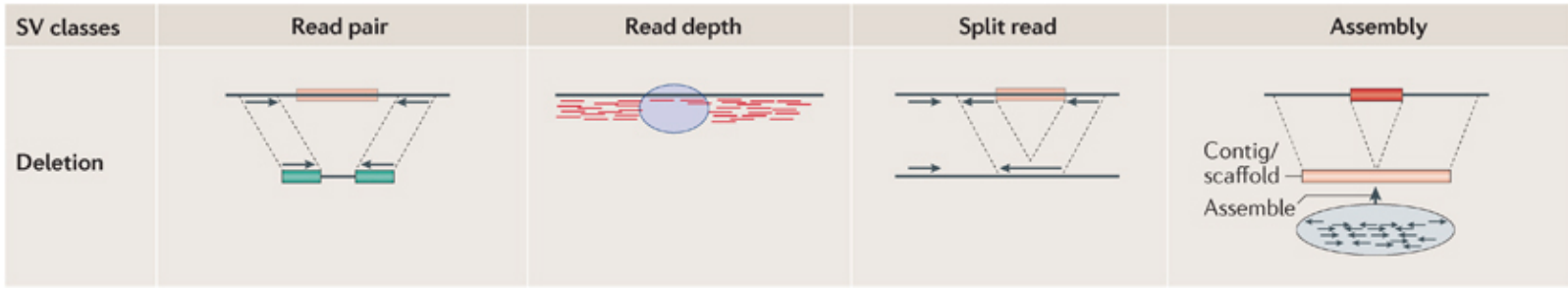
graphic credit: Victor Guryev

Deletions in WGS Data



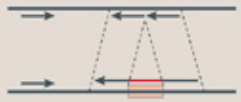
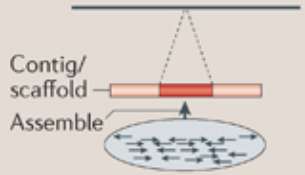
What are the signals that let us detect a deletion?

Deletions in WGS Data



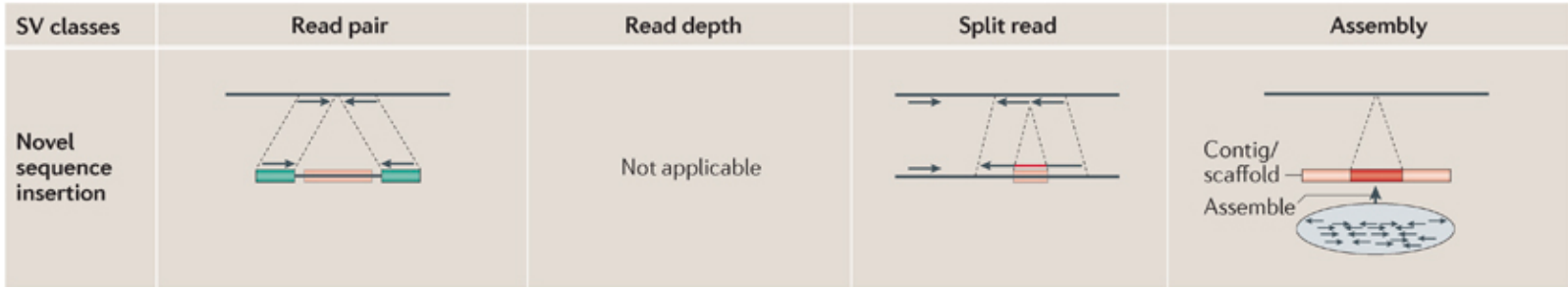
Read pair	increased interpair mapping distance
Read depth	fewer reads
Split read	single read is "merged" from two segments surrounding deletion
Assembly	assembled sequence shows "gap"

Insertions in WGS Data

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

What are the signals that let us detect a insertion?

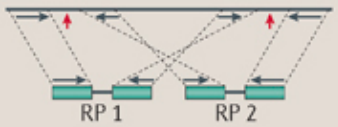
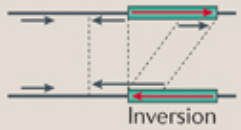
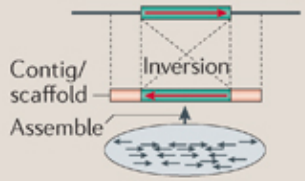
Insertions in WGS Data



Read pair	decreased interpair mapping distance
Read depth	not applicable[3]
Split read	single read is split into two segments surrounding novel insertion sequence
Assembly	assembled sequence with inserted novel sequence

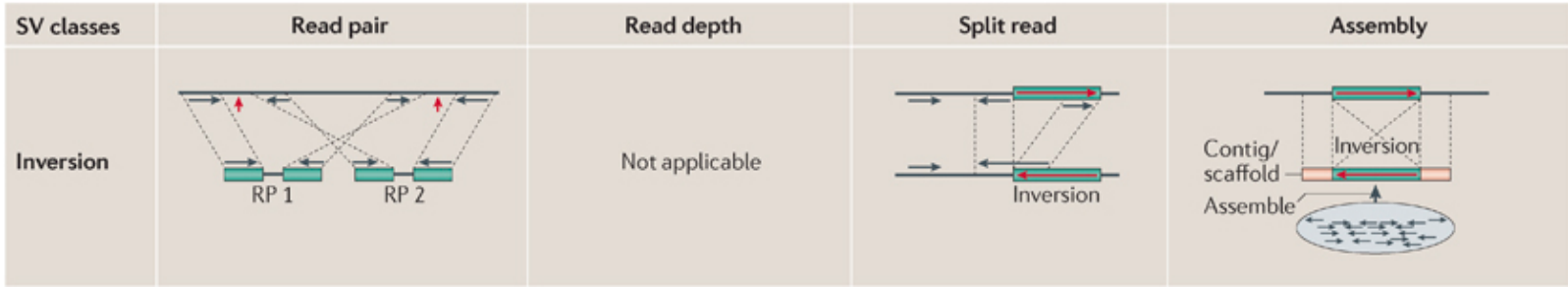
[3] Novel sequence will not map to genome

Inversions in WGS Data

SV classes	Read pair	Read depth	Split read	Assembly
Inversion		Not applicable		

What are the signals that let us detect a inversion?

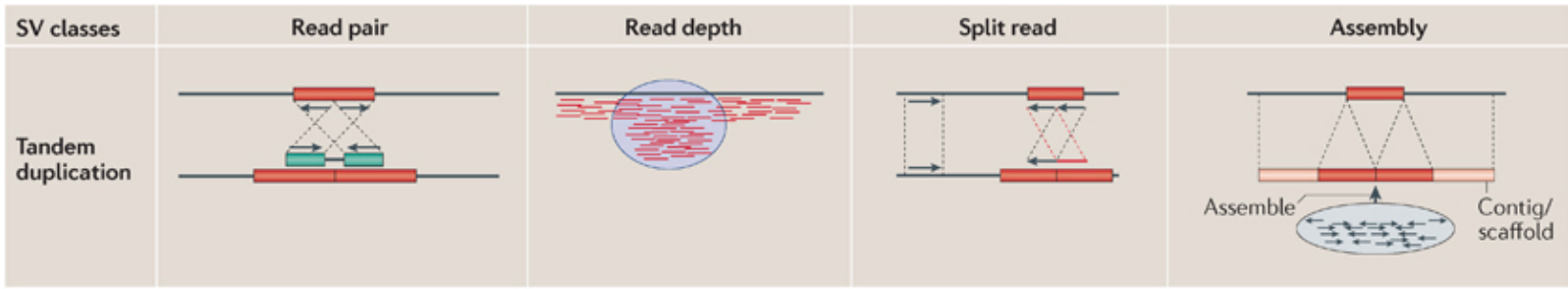
Inversions in WGS Data



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

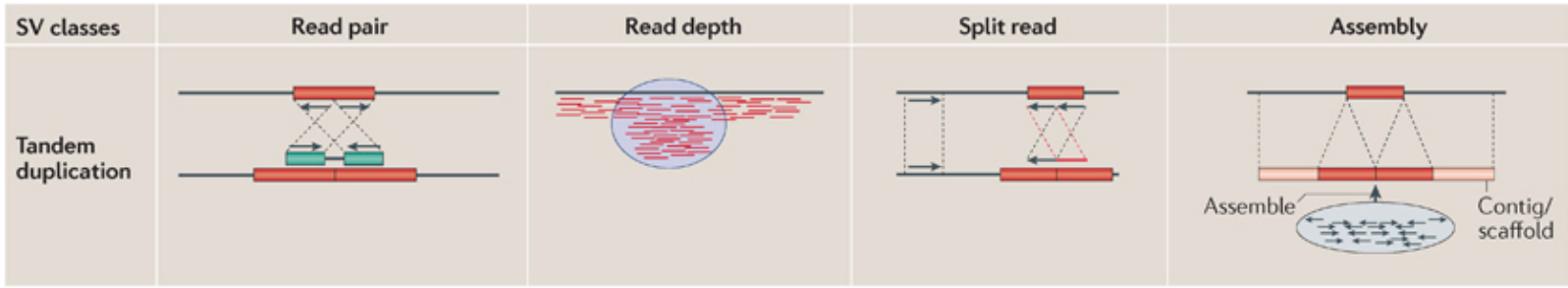
[4] Same amount of sequence

Duplications in WGS Data



What are the signals that let us detect a duplication?

Duplications in WGS Data



Read pair	aberrant mapping (<---> instead of >---<) and interpair distance
Read depth	increasd
Split read	single read is split into end of one duplicated block followed by beginning of next block
Assembly	assembled sequence with duplicated sequence

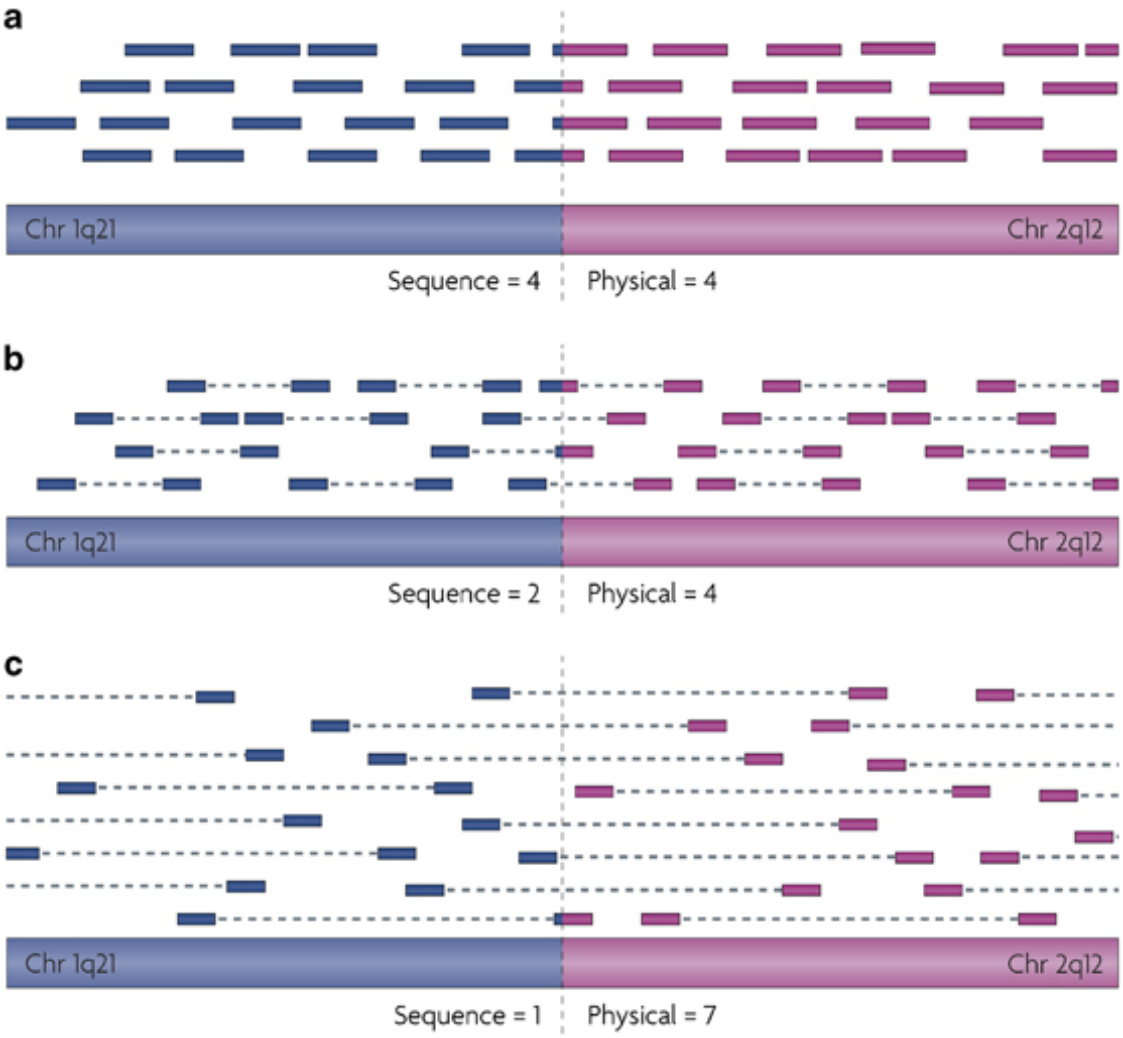
Graphics credit: Le Scouarnec and Gribble SM *Heredity (Edinb)*. 2012; **108**:75-85.

Translocations in WGS Data

a) single-end sequencing

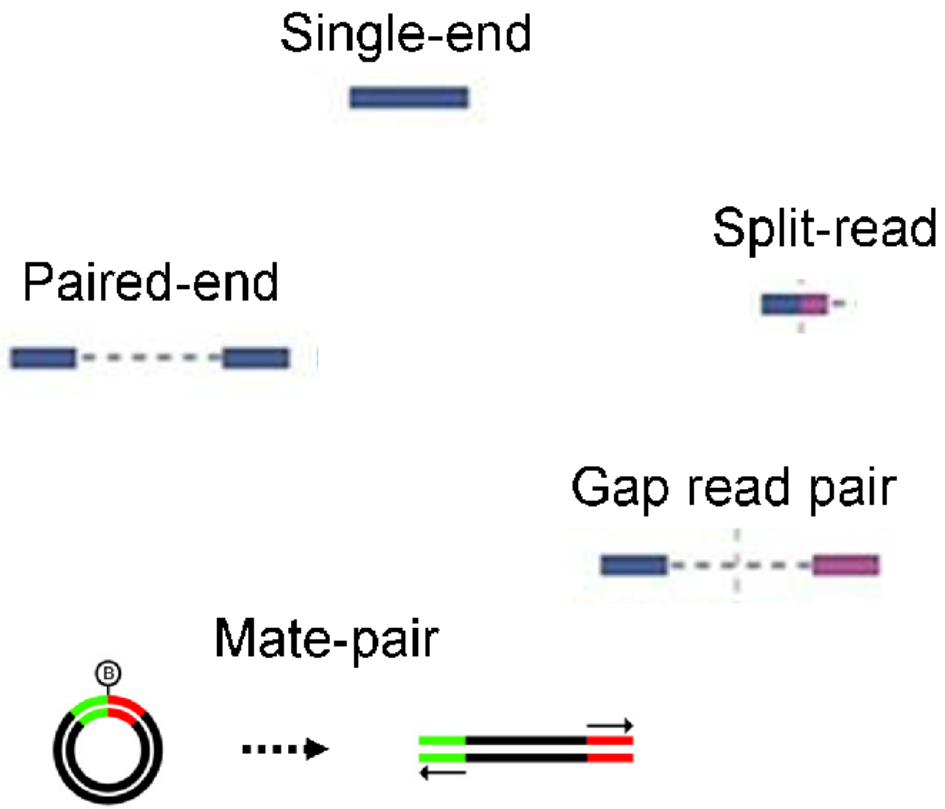
b) paired end (short insert library)

c) mate-pair (large insert library)



What are the signals that let us detect a translocation?

Signals and Read Types



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

Read depth

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]

[5] Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* 2009;**19**:1586–92.

Poisson

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

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

- k = number of occurrences
- λ = average occurrences/time interval

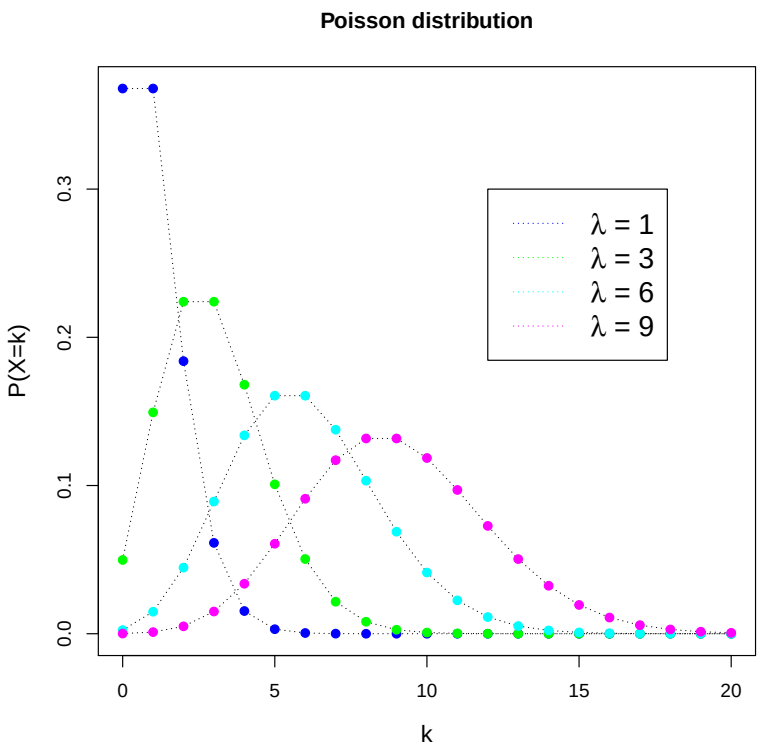
For example, if the average number of soldiers killed by being kicked by a horse each year in each of 14 cavalry corps is 1.7, what is the probability of 4 soldiers being killed in one year?

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

In R,

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

Poisson



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

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

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

Poisson and Normal Approximation

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

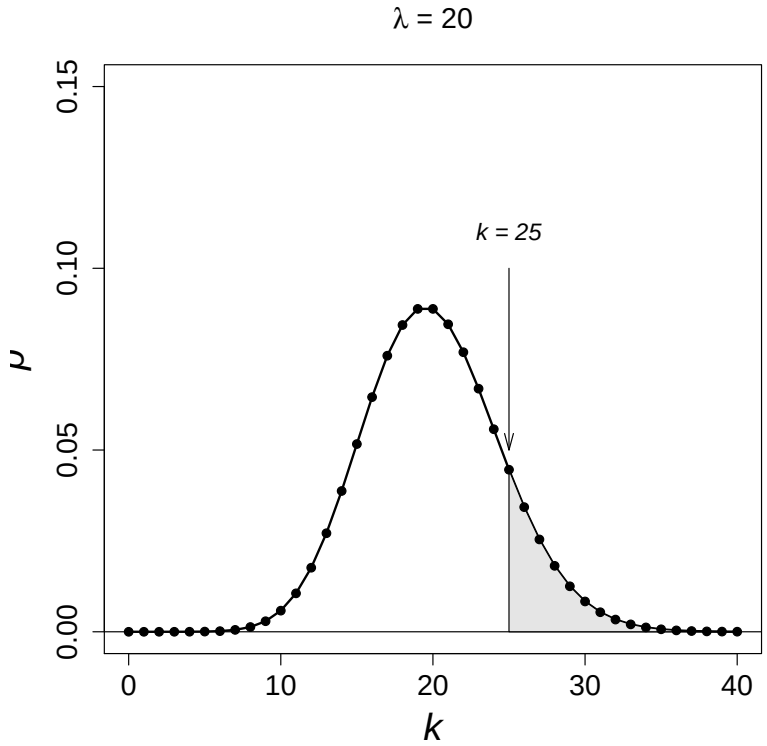
The central limit theorem: the mean of a sufficiently large number of independent random variables, each with finite mean and variance, is approximately normally distributed

Poisson and Normal Approximation

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.

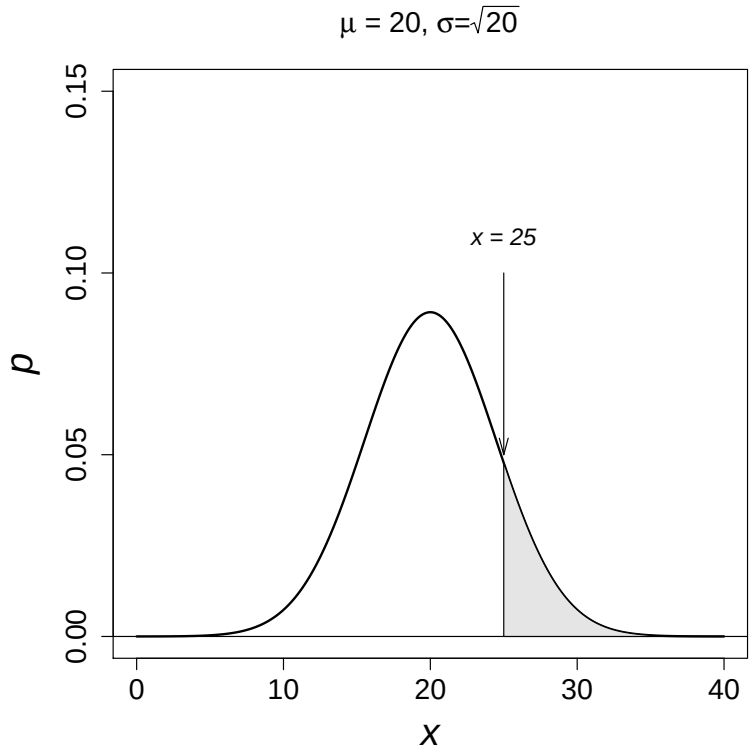
If λ is greater than about 10, then the Normal Distribution is a good approximation if an appropriate continuity correction is performed.

Poisson and Normal Approximation



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

Poisson and Normal Approximation



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

Poisson and Normal Approximation

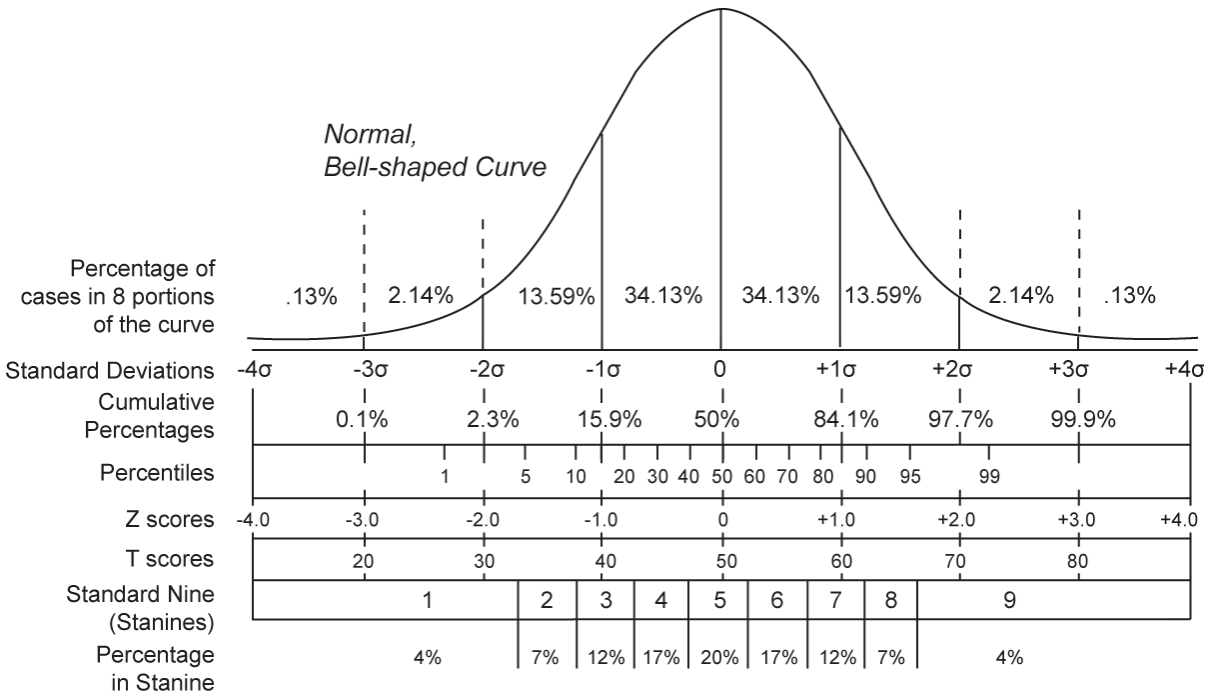
- Finally, we can check in R that the Normal is a reasonable approximation to the Poisson (it is not an extremely close approximation for λ in this range yet)[6].

```
> 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.

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

Poisson and Normal Approximation

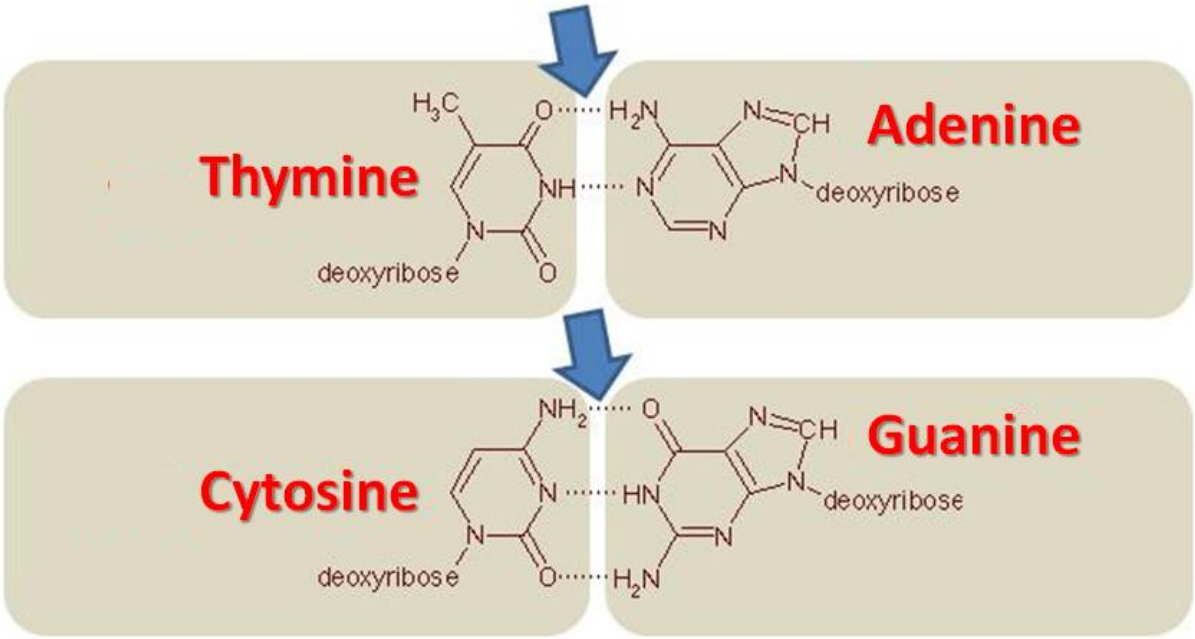


z-score

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

grapic: wikipedia

GC Content



grapic: 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

- $\uparrow$ 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

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

We can get a simple picture of the distribution of reads acrosss 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)

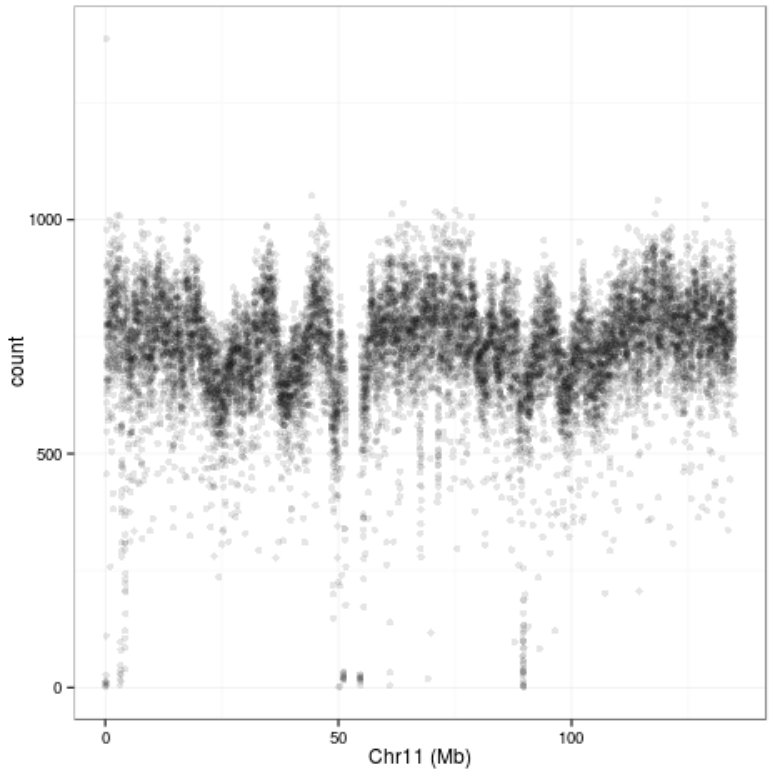
There is a tutorial on how to do the next few analysis steps on the website.

[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.

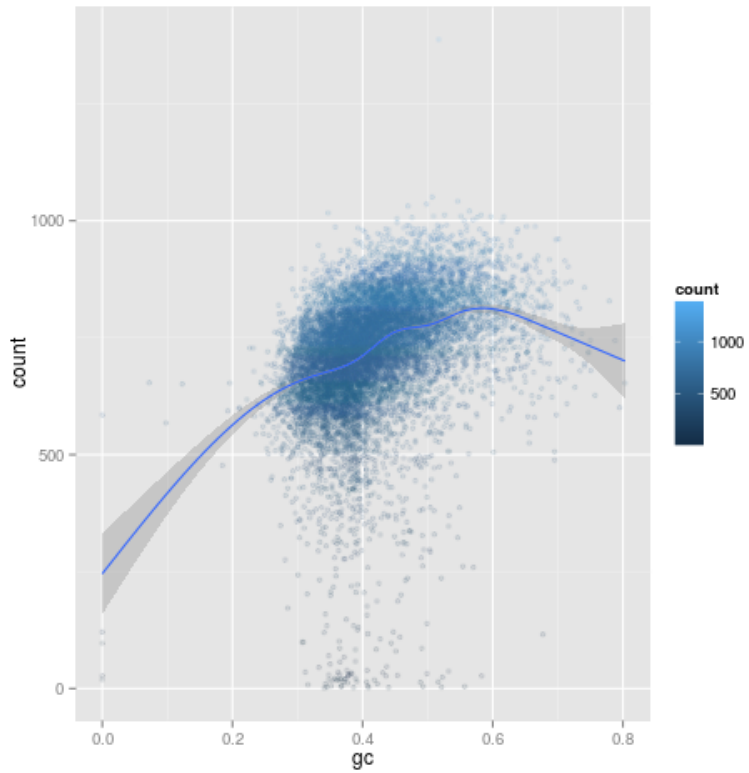
Thousand genomes project, individual HG00155, chromosome 11, low-coverage



GC content vs. Read Depth

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

CNVnator

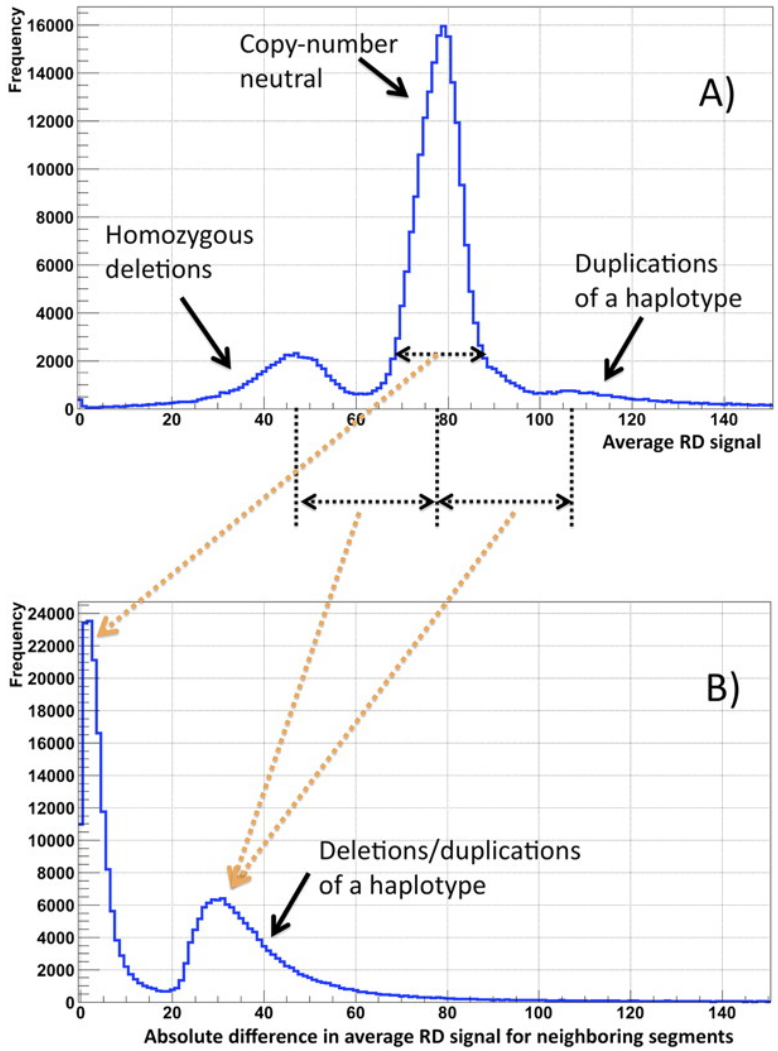
With this information in hand, we will discuss a leading algorithm used to detect CNVs in genomic data, **CNVnator**.

Abyzov A, Urban AE, Snyder M, Gerstein M (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* **21**:974-84.

- CNVnator makes use of a number of nice ideas to provide good CNV calls
  - GC content correction
  - Partitioning of bins with mean shift technique
  - statistical hypothesis testing to call CNVs

CNVnator

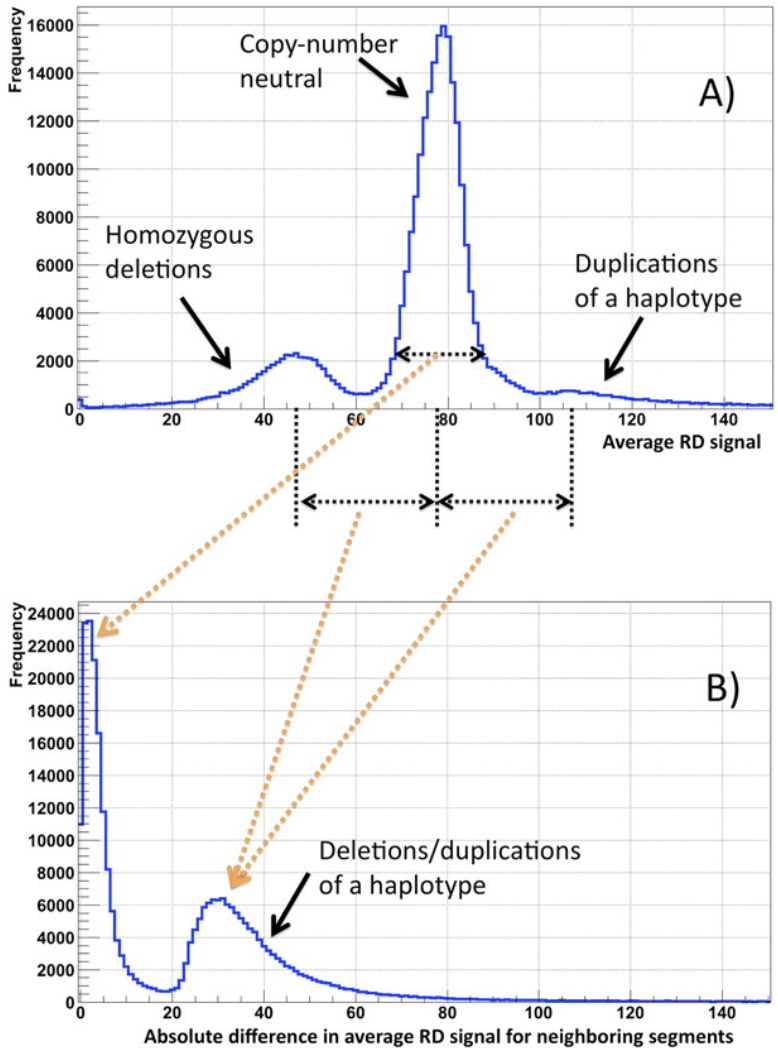
- Average RD signal distribution in produced segments.
- 3 clear peaks
  1. around the genomic RD average (no CNVs)
  2. half of that (heterozygous deletion)
  3. one and one-half of that (duplication of one haplotype).
- The average genomic RD signal is ~ 77 reads.



CNVnator

- Let us examine panel B
- Distribution of the average RD signal difference for **neighboring segments**.

- One cluster of neighboring segments has similar average RD signals (peak around zero).
- The other cluster has an average signal difference of ∼half of the genomic average RD signal.
- ∴ changes in average RD signal at two neighboring segment boundaries cluster, and these clusters can be explained by partitioning that includes deletions and duplications



Partitioning genome into CN segments

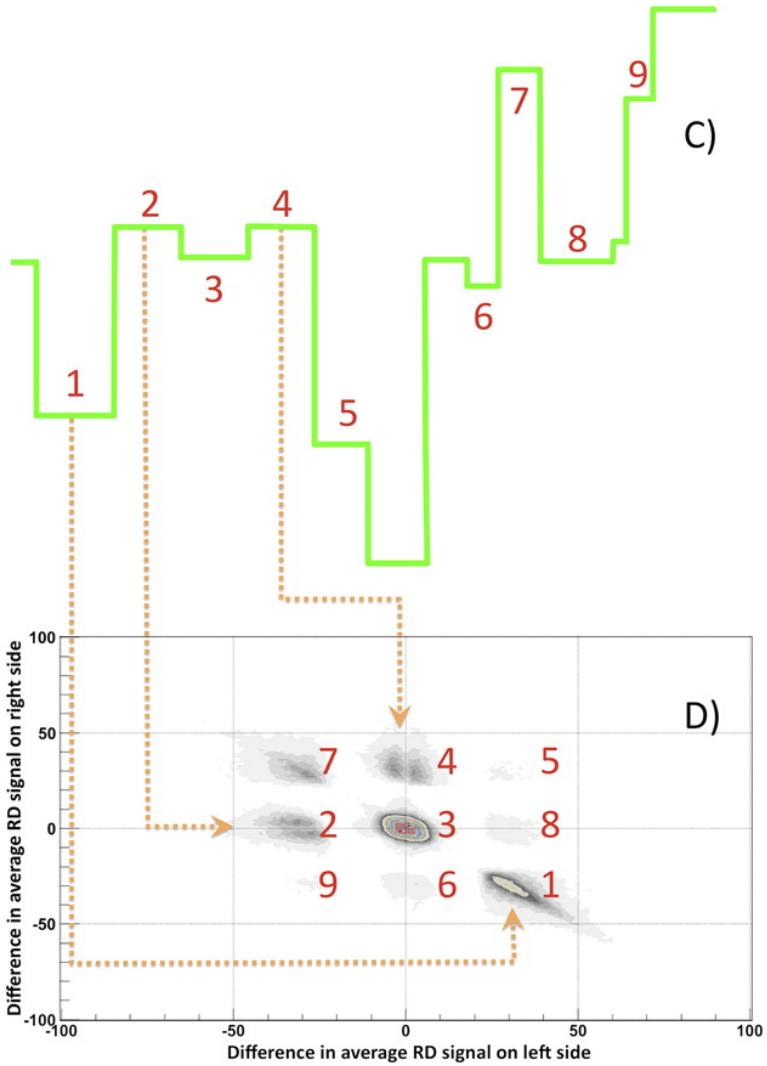
- Distribution of the average RD signal difference at the left and right boundary for each segment.
- The clusters originate due to various combinations of segments with different RD signals.

Cluster 3: CNV neutral.

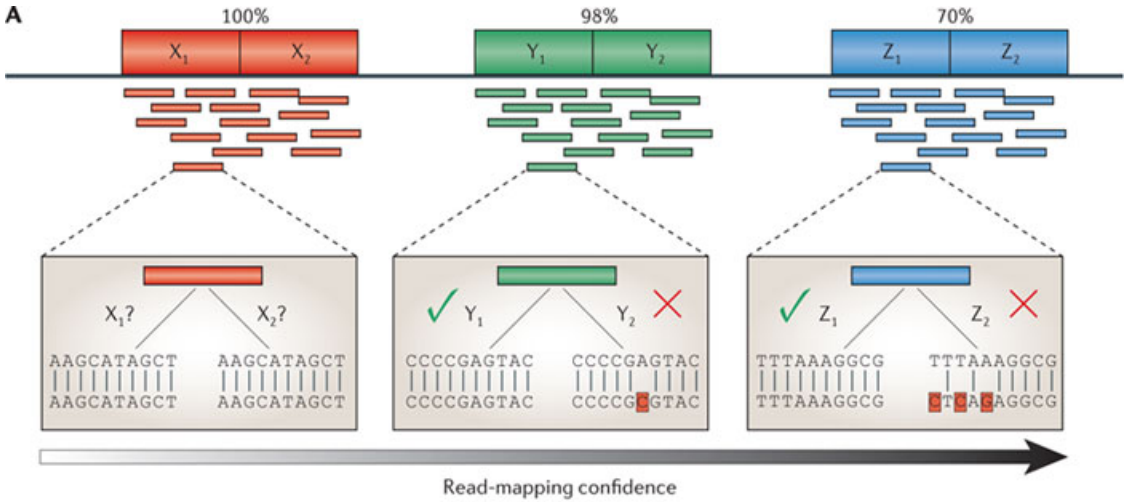
Cluster 2: Deletion begins on left side

Cluster 4: Insertion begins on left side

etc.



CNVnator: Dealing with multiply mappable reads



Treangen TJ and Steven L. Salzberg SL (2012) *Nature Reviews Genetics* **13**:36-46

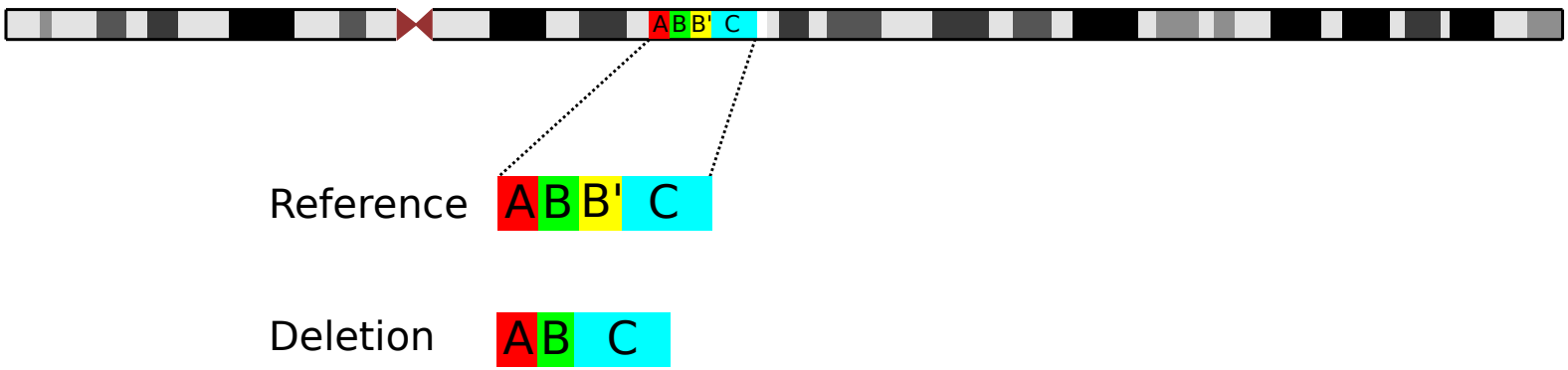
- three different tandem repeats with two copies each
- Left: read aligns equally well to both X_1 and X_2
- Middle: read aligns slightly better to Y_1 than to Y_2
- Right: read aligns perfectly to Z_1, whereas its alignment to Z_2 contains three mismatches

When a read (pair of reads) can map equally well to two or more locations, then one is randomly chosen. In such cases, the associated mapping quality is zero.

CNVnator: Dealing with multiply mappable reads

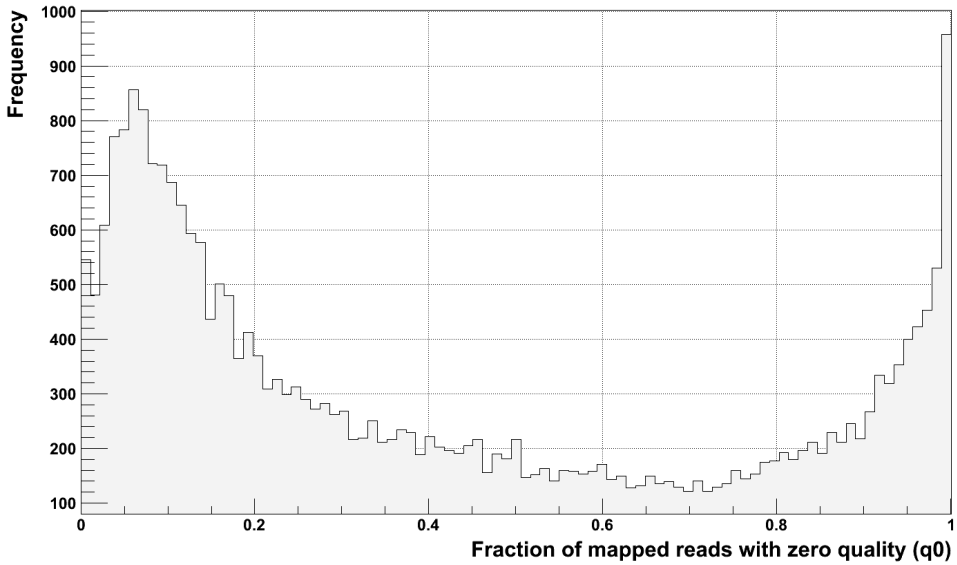
- Calling a CNV in particular regions is confounded by the presence of the same (or very similar) copies of that region in the reference genome.
- The RD signal for a CNV in these regions is effectively smeared (due to random placement of nonuniquely mapped reads) over all copies

Example: Dealing with multiply mappable reads



- Consider the situation where the reference has two nearly identical segmental duplications, B and B'
- The sequenced individual has a homozygous deletion of B'
- Reads that originate from B in the sample will distribute equally between B and B' in the reference
- Thus, both B and B' have half of the average RD (i.e., copy number [CN] $= 1$)
- A naive analysis would identify both B and B' as heterozygous deletions

CNVnator: Dealing with multiply mappable reads



Distribution of fraction of q0 mapped reads in the regions of predicted CNVs

- Quality zero (q0) reads commonly occur in CNV regions
- The distribution of the fraction of q0 reads in the called CNV regions has peaks around 0 and 100%
- CNVnator considers a CNV region redundant if the fraction of q0 reads in the called CNV regions is $> 50\%$.

GC Adjustment

GC adjustment

$$\mathrm{RD}^{i}_{corrected} = \mathrm{RD}^{i}_{raw} \cdot \frac{\overline{\mathrm{RD}}_{global}}{\overline{\mathrm{RD}}_{GC}}$$

- adjusted read count
- raw read count
- median count for GC content
- overall median count per window

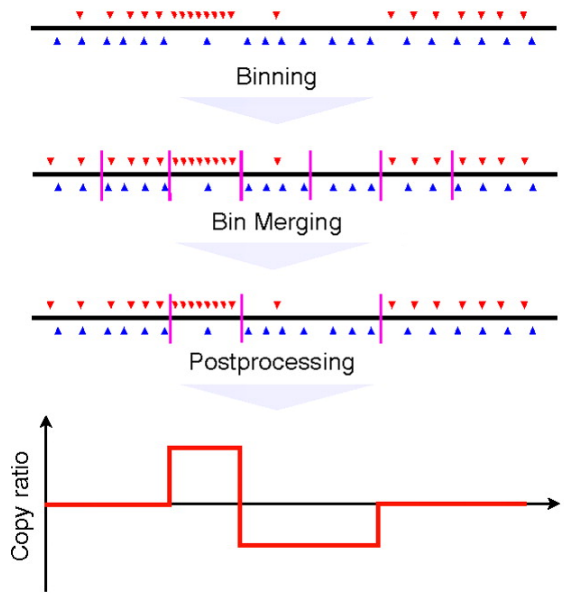
i: bin index, RD^{i}_{raw}: raw RD signal for a bin, $\mathrm{RD}^{i}_{corrected}$: corrected RD signal for the bin, $\overline{\mathrm{RD}}_{global}$: average RD signal over all bins, and $\overline{\mathrm{RD}}_{GC}$: average RD signal over all bins with the same GC content as in the bin.
This correction effectively eliminates correlation of RD signal with GC content

Partitioning

The bin size used by CNVnator is typically 100-300 nucleotides, but many CNVs are much larger than this. We would therefore like to partition a chromosome into segments with the same copy number.

The general flow of the algorithm can be summarized as

- divide genome into bins, count reads
- Use partitioning algorithm to join adjancent bins together
- statistical postprocessing to call deletions/duplications



Xi R et al. (2011) Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *PNAS* **108**:E1128-36.

Mean shift: Kernel density estimation

- Given a random sample $X_1, X_2, \ldots, X_n$ with a continuous univariate density f
- The kernel density estimator with kernel K and **bandwidth** h is

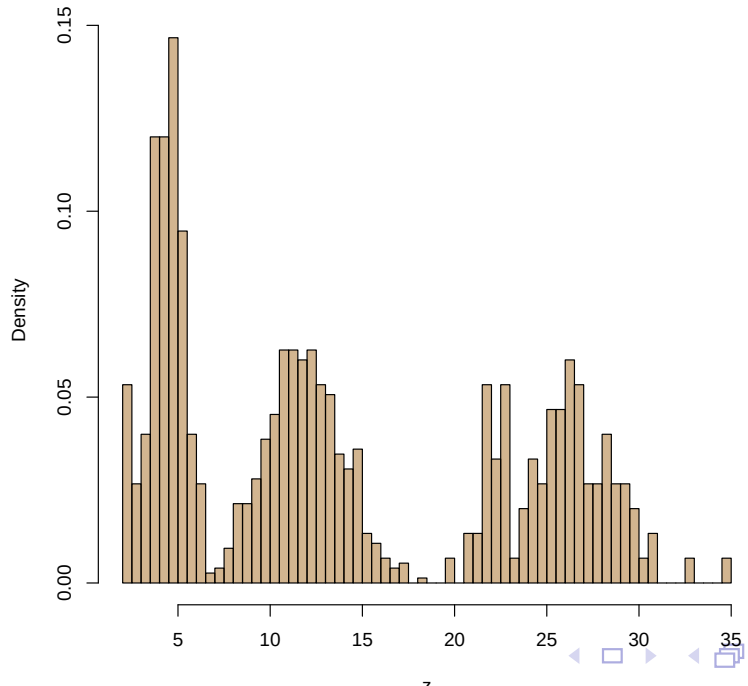
$$\hat{f}(x, h) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{5}$$

- Center of kernel is placed right over each data point.
- Influence of each data point is spread about its neighborhood.
- Contribution from each point is summed to overall estimate

Mean shift: Kernel density estimation

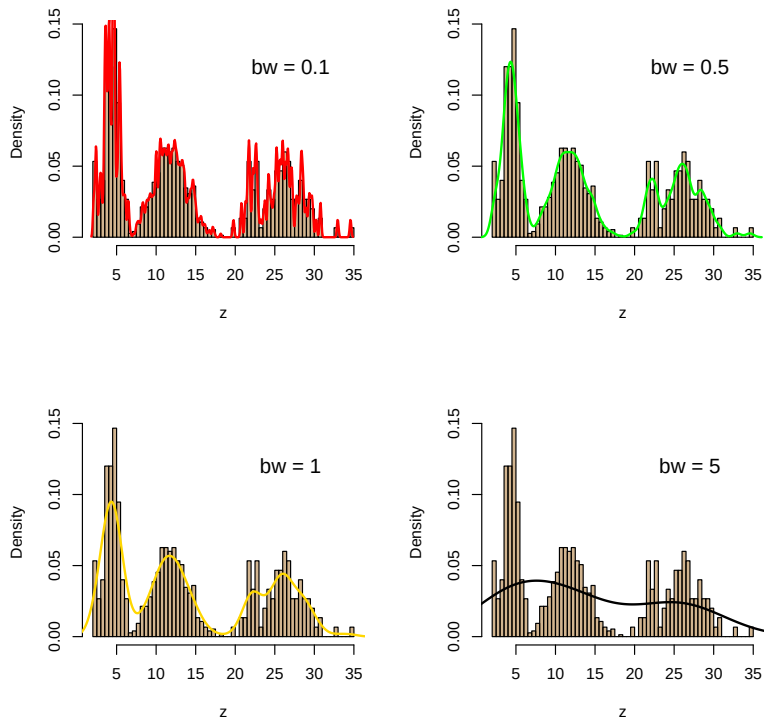
- Consider use of Gaussian kernel to estimate the density of the following data

$$\hat{f}(x,h) = \frac{1}{\sqrt{2\pi}\sigma} \sum_{i=1}^{n} e^{-\left(\frac{x - X_i}{2\sigma^2}\right)} \tag{6}$$



Mean shift: Kernel density estimation

- The bandwidth is a scaling factor that controls how wide the probability mass is spread around a point.
- it controls the smoothness or roughness of a density estimate



Mean shift: Read count bins

In CNVnator, each read count bin is represented as a point in 2 dimensional space $x_i = (i, r_i)$, where r_i is the signal bin index using a two-dimensional Gaussian kernel

$$F(x_i) = norm \sum_{j \neq i}^{n} e^{-\left(\frac{j-i}{2H_b^2}\right)} e^{-\left(\frac{r_j - r_i}{2H_r^2}\right)} \quad (7)$$

j is the index of neighboring bins, H_b and H_r are the bandwidths for the bin index and RD signal accordingly, and *norm* is the normalization factor.

- The mean-shift vector is the gradient of this function

$$\nabla F(x_i) = \begin{pmatrix} \frac{\partial F}{\partial i} \\ \frac{\partial F}{\partial r} \end{pmatrix} \quad (8)$$

Mean shift: Read count bins

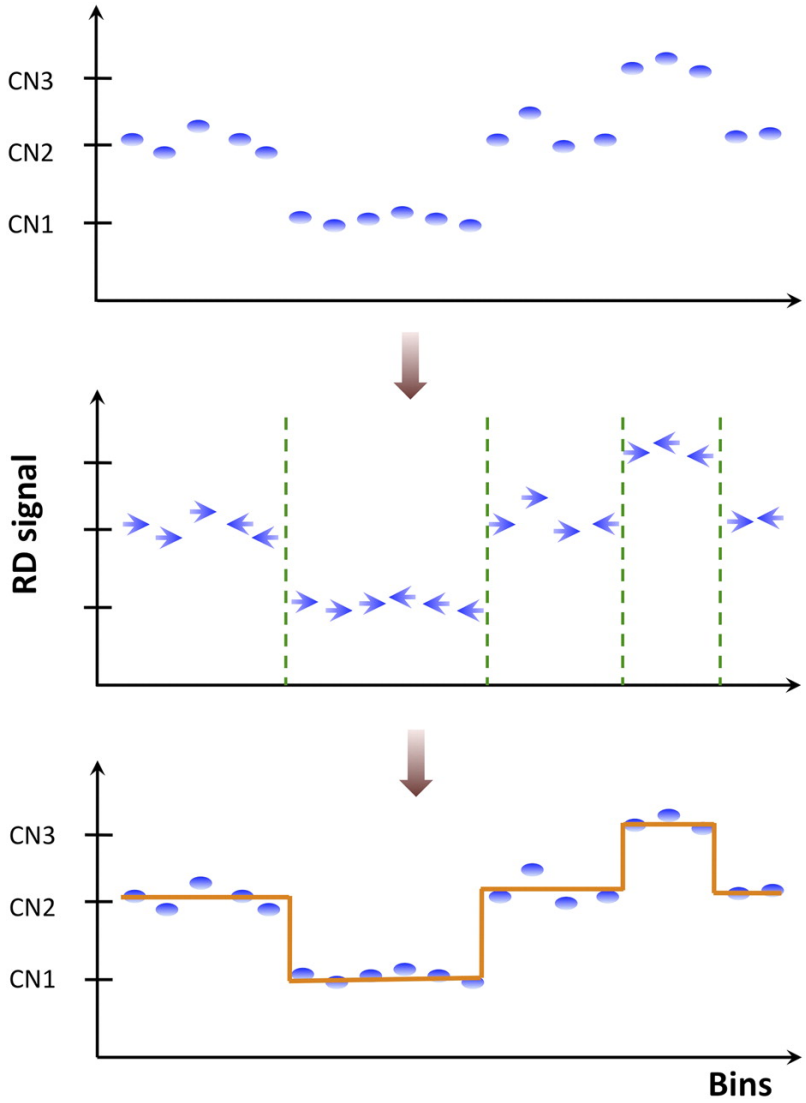
Since we are only interested in the direction of the gradient along the bins, we calculate

$$\begin{aligned}\frac{\partial F}{\partial i} &= \frac{\partial}{\partial i} norm \sum_{j \neq i}^{n} e^{-\left(\frac{j-i}{2H_b^2}\right)} e^{-\left(\frac{r_j - r_i}{2H_r^2}\right)} \\ &= norm \sum_{j \neq i}^{n} -(j-i) e^{-\left(\frac{j-i}{2H_b^2}\right)} e^{-\left(\frac{r_j - r_i}{2H_r^2}\right)}\end{aligned}$$

We are now only interested in the direction of $\frac{\partial F}{\partial i}$, i.e., whether it is pointing to the right ($\frac{\partial F}{\partial i} > 0$ or to the left ($\frac{\partial F}{\partial i} < 0$). Thus we do not need to calculate *norm*, which is always positive.

Mean shift: Read count bins

For each bin, i.e., data point, the mean-shift vector points in the direction of bins with the most similar RD signal. Segment breakpoints are determined where two neighboring vectors have opposite directions but do not point to each other.



CNVnator

CNVnator uses some interesting heuristics to estimate the optimal bandwidth in order to come up with a final partitioning of bins, which we will not go into here. Finally, the analysis is performed with a one-sample t test

Recall: To test the null hypothesis that a population mean is equal to a specified value μ_0, one uses the t statistic

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \tag{9}$$

where $\overline{x}$ is the sample mean, s is the sample standard deviation of the sample and n is the sample size. The degrees of freedom used in this test are $n1$.

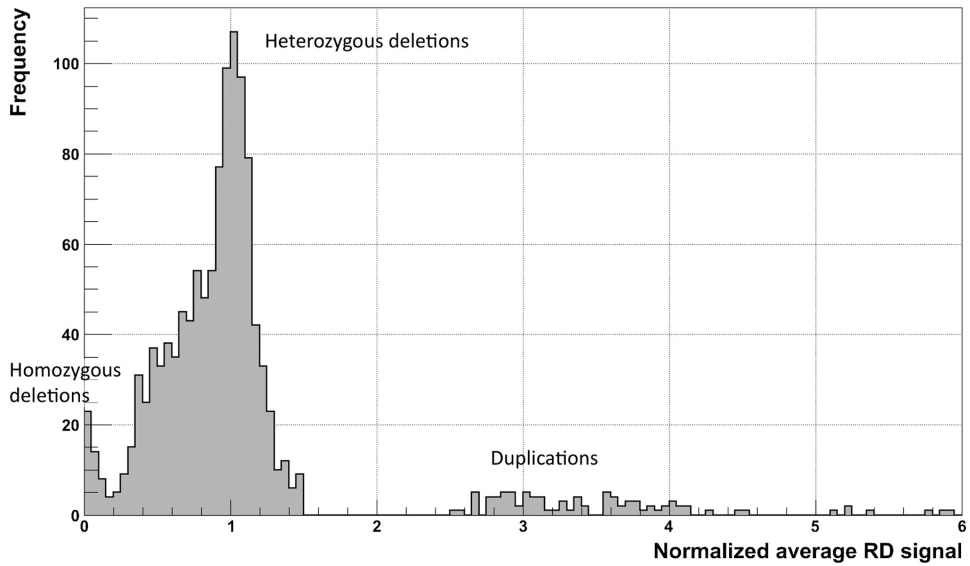
CNVnator

The t test for CNVnator is formulated as

$$t = \frac{\overline{RD}_{global} - \overline{RD}_{segment}}{s_{segment}/\sqrt{n}} \tag{10}$$

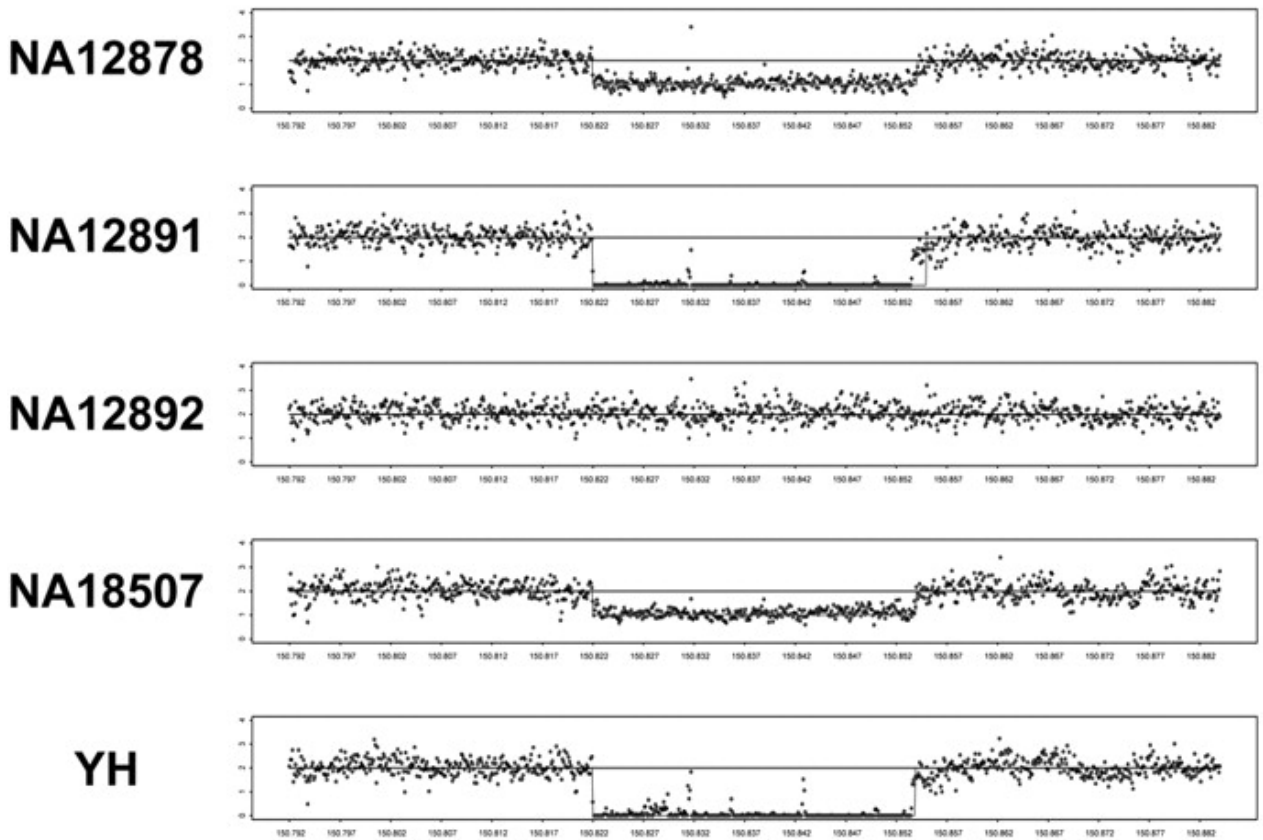
where n is the number of bins within the segment, $\overline{RD}_{segment}$ is its average RD signal, and $s_{segment}$ is the signal standard deviation.

CNVnator



Distribution of normalized average RD signal for predicted CNVs (for a CEPH daughter) that are >1 kb and pass the q0 filter. Abyzov A et al (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* **21**:974-84.

CNV Calling via Read Depth



Yoon et al. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* 2009;**19**:1586–92.

Summary

What you should take away from this lecture

- The various kinds of signals used to detect structural variants (SVs) in NGS data
- The various kinds of SVs and what effects they have on NGS reads
- Basic steps in using read depth to identify copy number variations (CNVs)
- GC bias

peter.robinson@charite.de