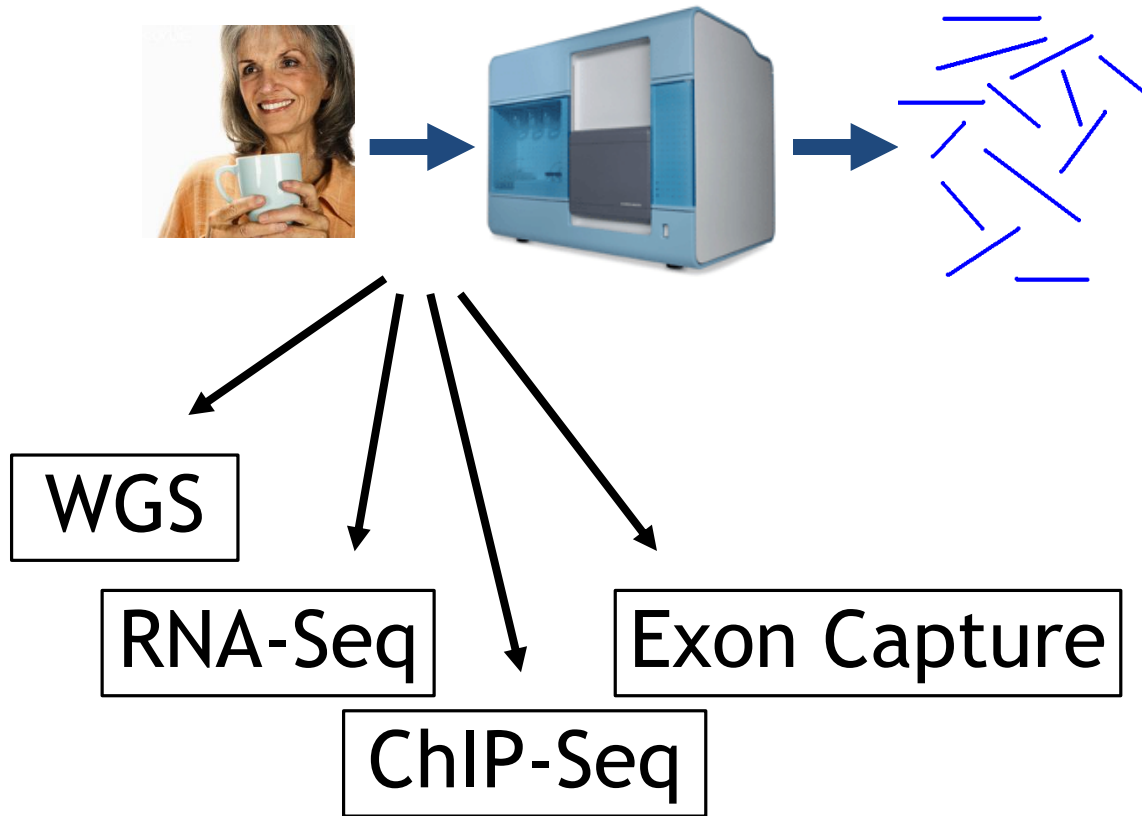


# SNPs, Short InDels and Structural Variant Calling

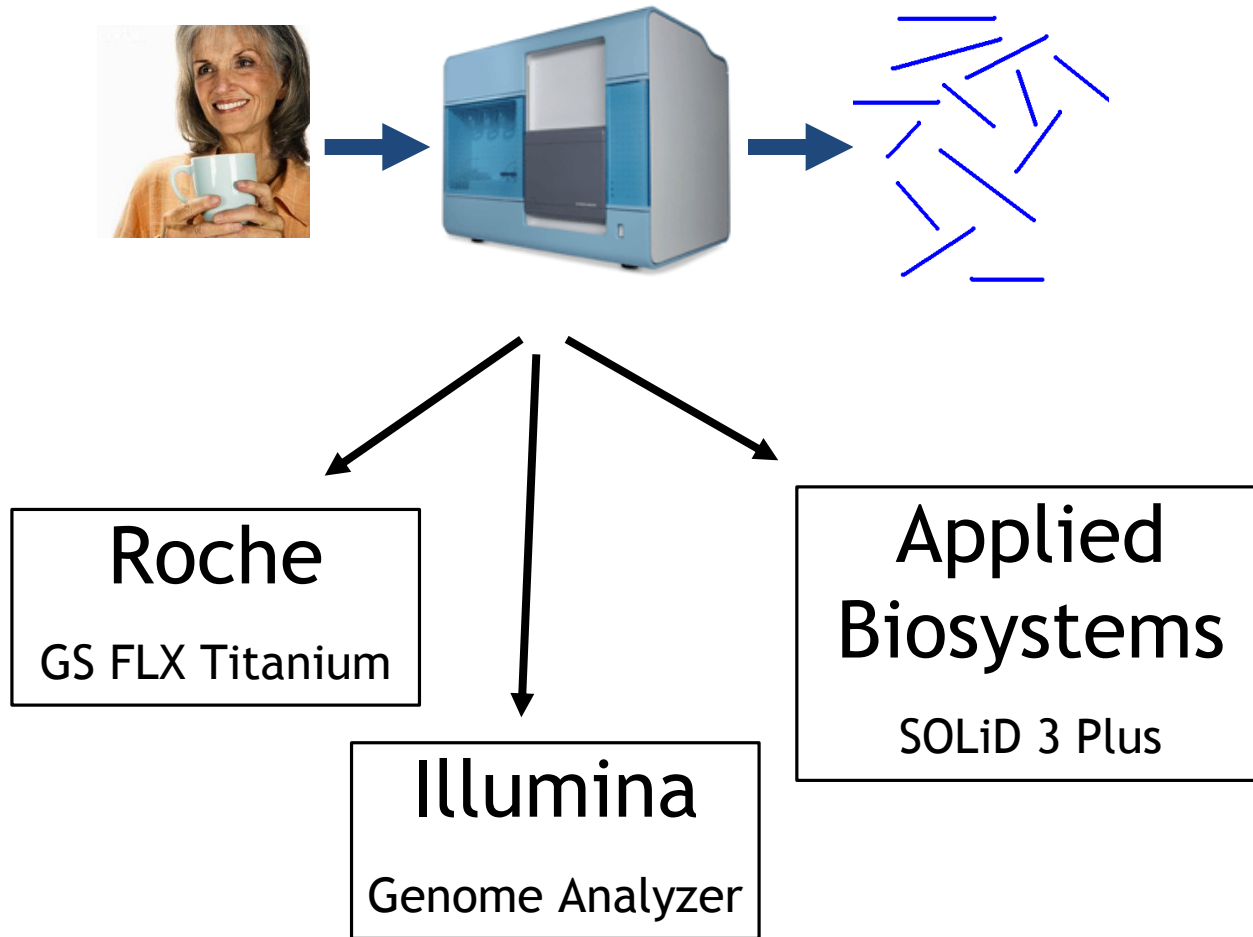
Tobias Rausch

28<sup>th</sup> April 2011

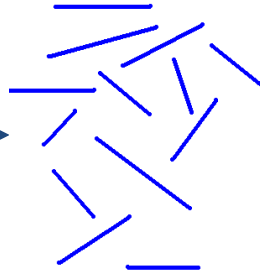
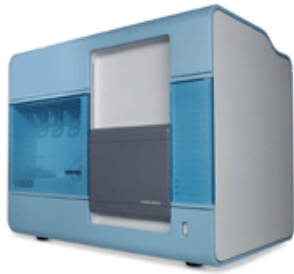
# Sequencing



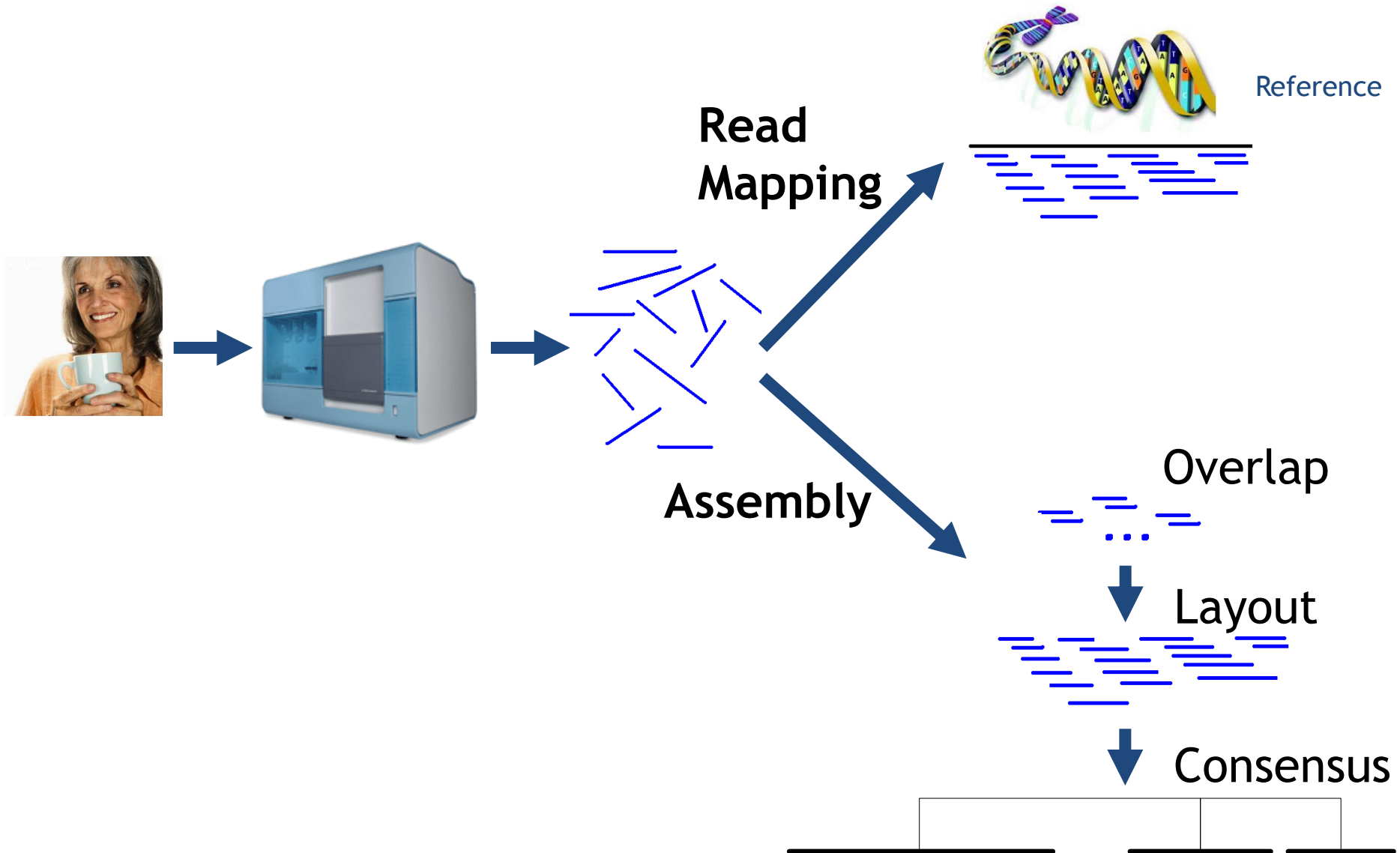
# Sequencing



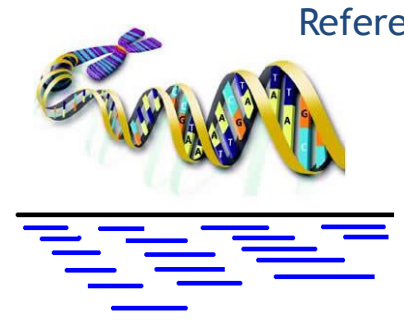
# Data Analysis



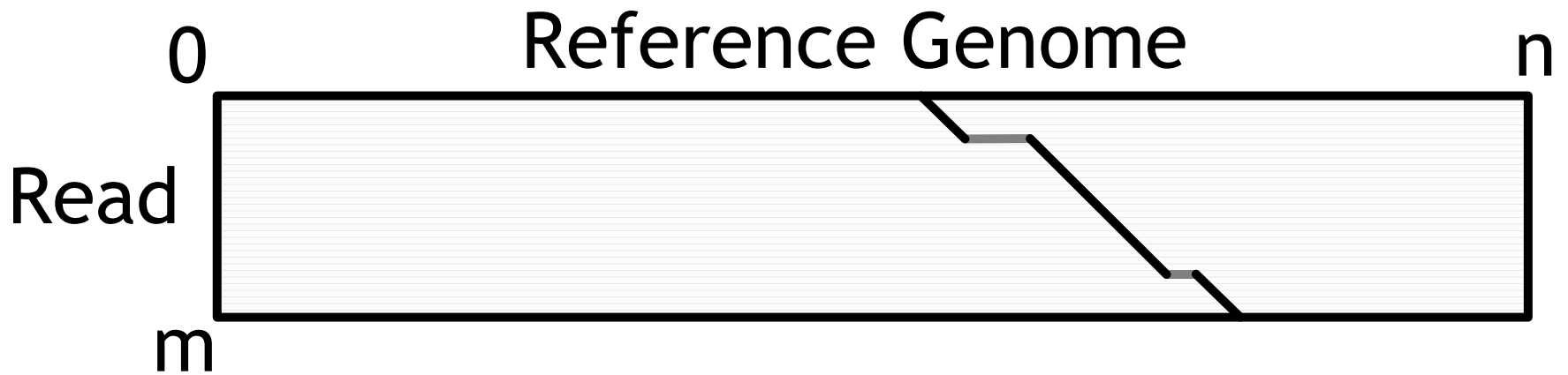
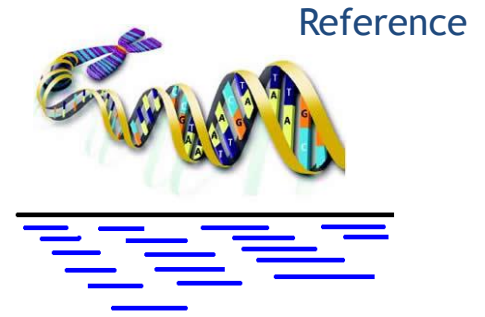
# Data Analysis

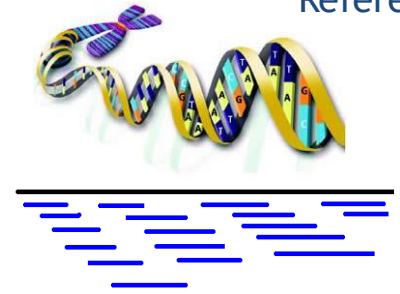


# Read Mapping

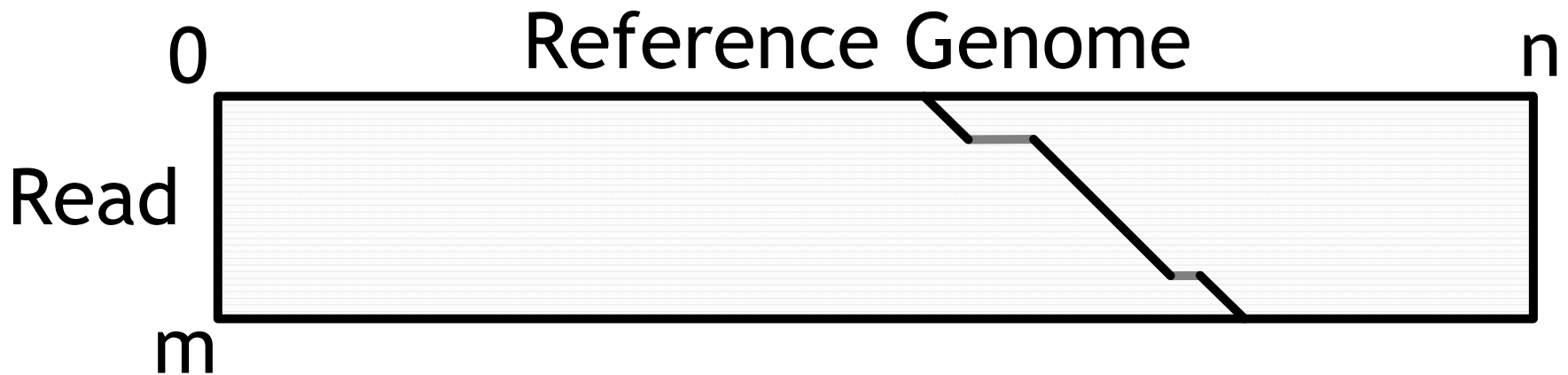


# Read Mapping





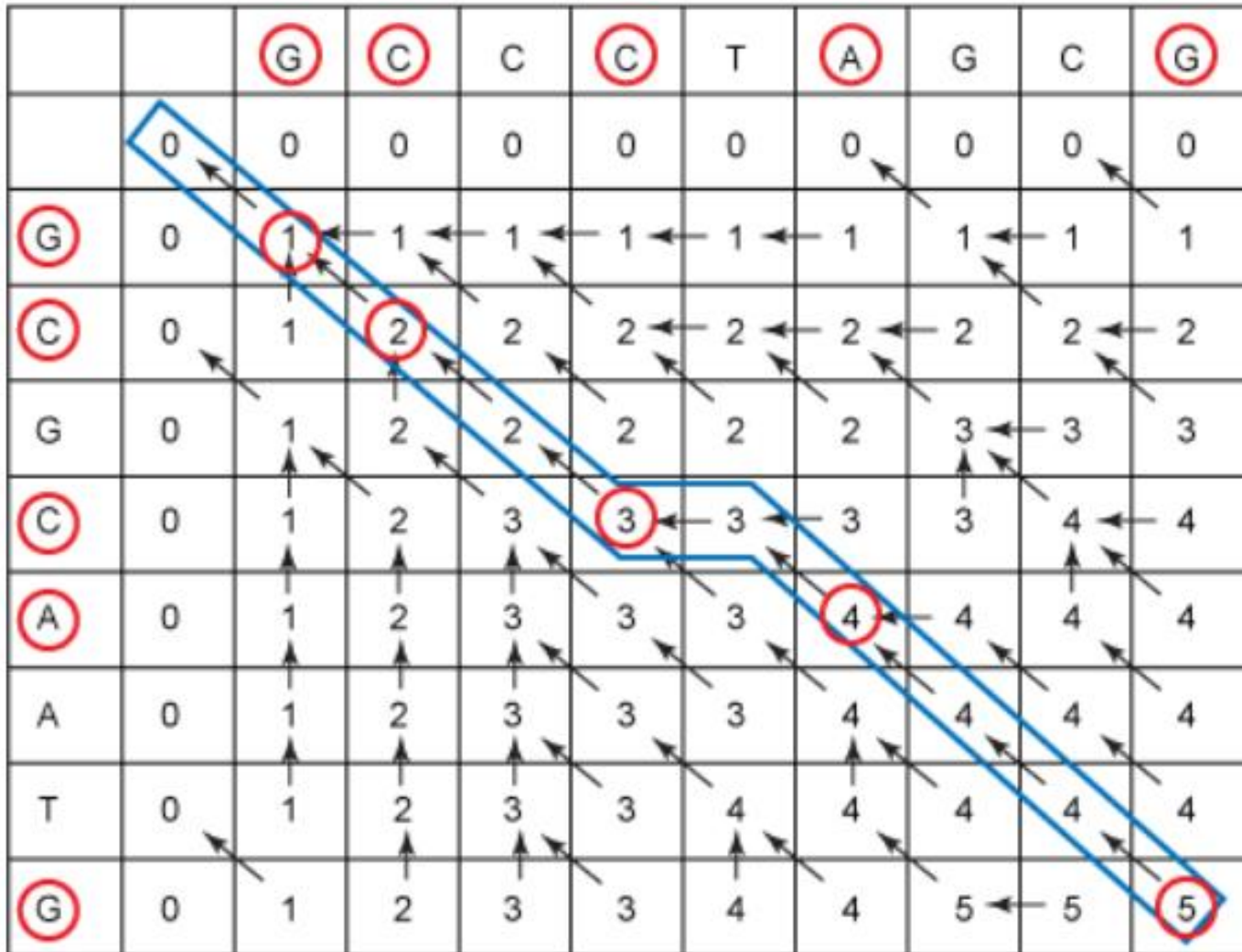
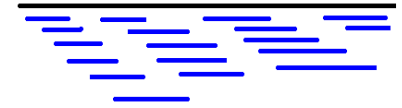
# Read Mapping

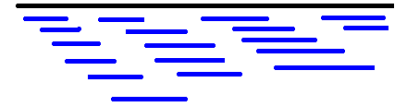


- Dynamic Programming: Quadratic algorithm
  - Requires  $O(m*n)$  time and space
- Infeasible for millions of short reads

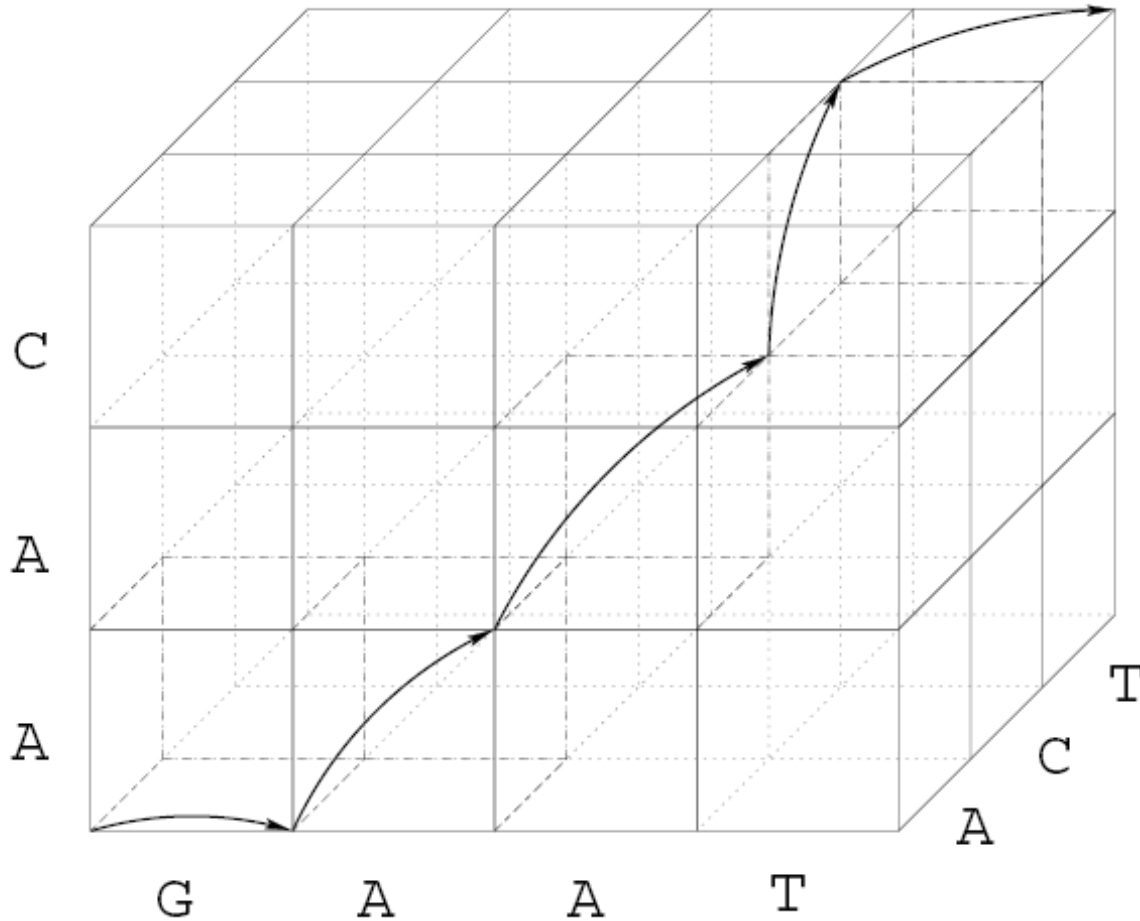


# Read Mapping





# Multiple Alignment



GAA-T  
 -AAC-  
 --ACT

# Filtering

*Genome*



Preprocess



Index

# Filtering

*Genome*



Preprocess

*Read*



Filter  
Algorithm

Index



*Filtration Phase*

*Potential Matches* 

# Filtering

*Genome*



Preprocess

Index

Filter  
Algorithm

*Read*



Exact  
Algorithm



*Filtration Phase*

*Potential Matches* 

*Verification Phase*

*True Matches* 

*False Matches* 

# Simple k-mer Index, k=3

S = ACGAAAAC TCGATTACTCGACC

	Hitlist		Hitlist		Hitlist
AAA		ACC		CGA	
AAC		ACG		...	
AAG		ACT		GAA	
AAT		AGA		...	
ACA		...		TTT	

- Size of that table:  $4^3 = 64$  entries =  $|\Sigma|^k$

# Simple k-mer Index, k=3

S = **ACG**AAAAC TCGATTACTCGACC

	Hitlist		Hitlist		Hitlist
AAA		ACC		CGA	
AAC		<b>ACG</b>	0	...	
AAG		ACT		GAA	
AAT		AGA		...	
ACA		...		TTT	

- Size of that table:  $4^3 = 64$  entries =  $|\Sigma|^k$

# Simple k-mer Index, k=3

S = A**CG**AAA**A**CTCGATTACTCGACC

	Hitlist		Hitlist		Hitlist
AAA		ACC		<b>CGA</b>	1
AAC		ACG	0	...	
AAG		ACT		GAA	
AAT		AGA		...	
ACA		...		TTT	

- Size of that table:  $4^3 = 64$  entries =  $|\Sigma|^k$

# Simple k-mer Index, k=3

S = AC**GAA**AACTCGATTACTCGACC

	Hitlist		Hitlist		Hitlist
AAA		ACC		CGA	1
AAC		ACG	0	...	
AAG		ACT		<b>GAA</b>	2
AAT		AGA		...	
ACA		...		TTT	

- Size of that table:  $4^3 = 64$  entries =  $|\Sigma|^k$

# Simple k-mer Index, k=3

S = ACGAAAAC TCGATTACTCGACC

	Hitlist		Hitlist		Hitlist
AAA	3,4	ACC	19	CGA	1
AAC	5	ACG	0	...	...
AAG	Empty	ACT	6,14	GAA	2
AAT	Empty	AGA	...	...	...
ACA	Empty	...	...	TTT	Empty

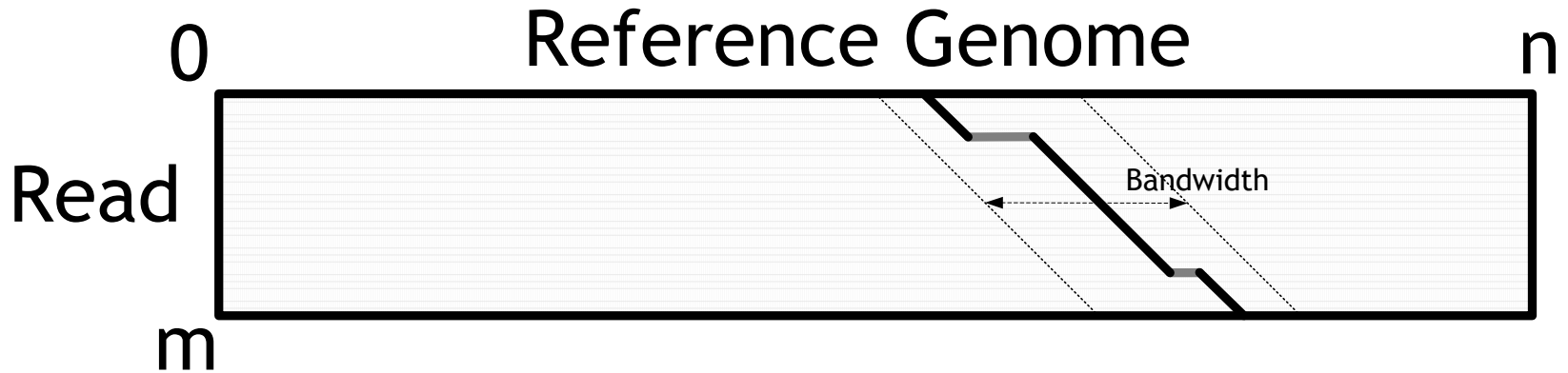
# Searching a Read

	Hitlist		Hitlist		Hitlist
AAA	3,4	ACC	19	CGA	1
AAC	5	ACG	0	...	...
AAG	Empty	<b>ACT</b>	6,14	GAA	2
AAT	Empty	AGA	...	...	...
ACA	Empty	...	...	TTT	Empty

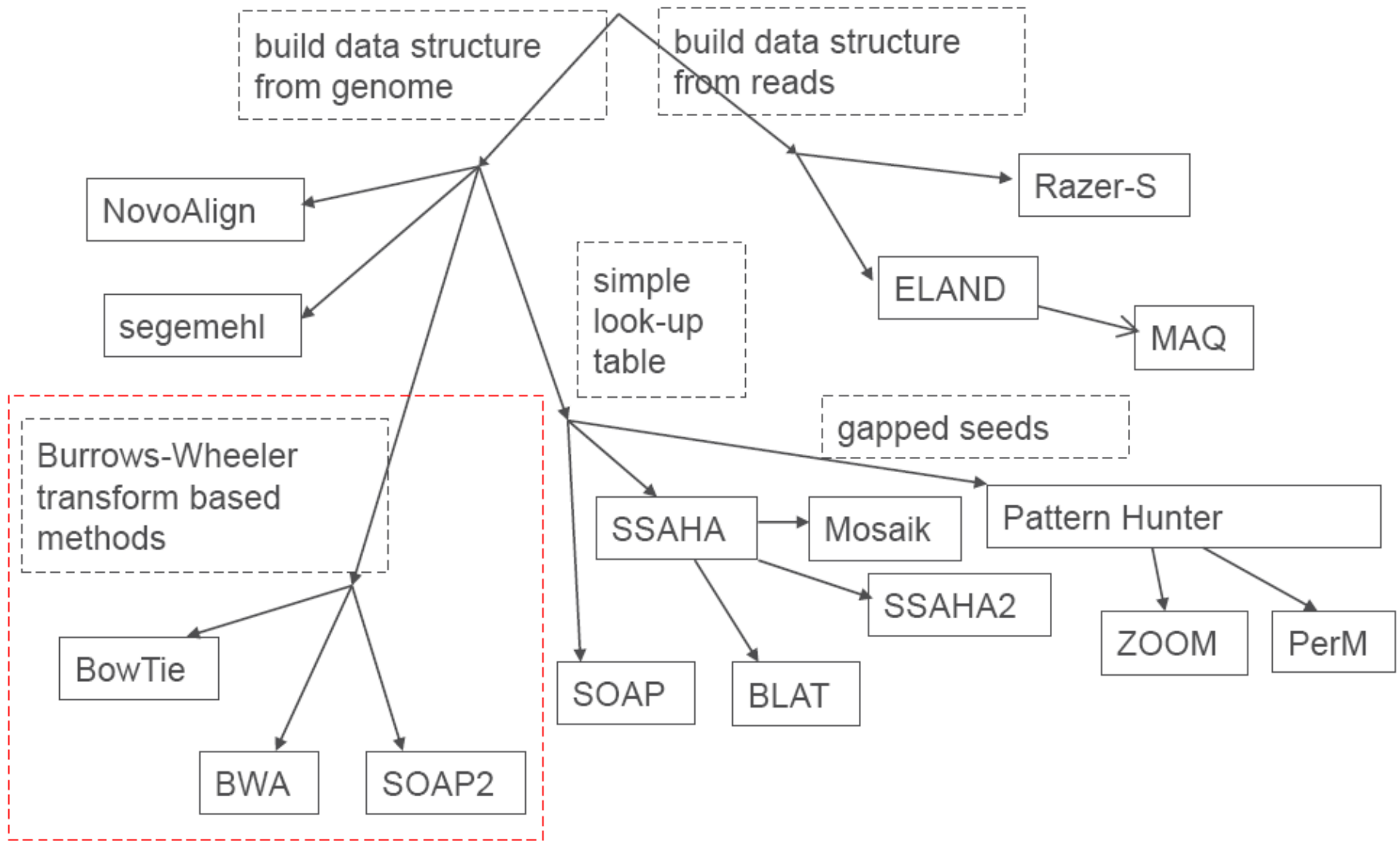
- Read Sequence: **ACTG**
  - Potential match at position 6 and 14

# Verification Algorithm

## Banded Dynamic Programming

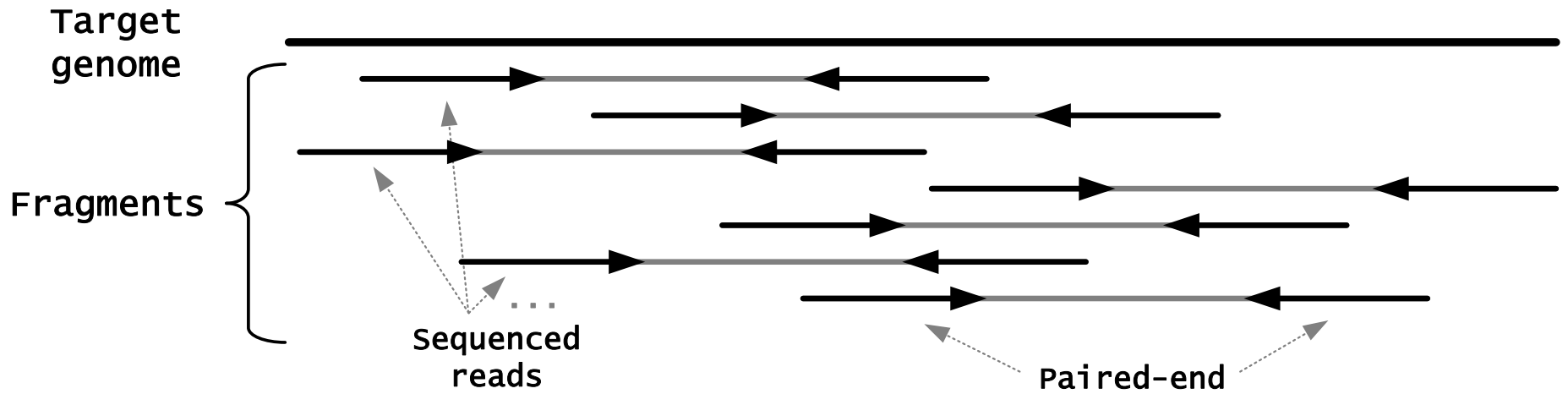


# Read Mappers

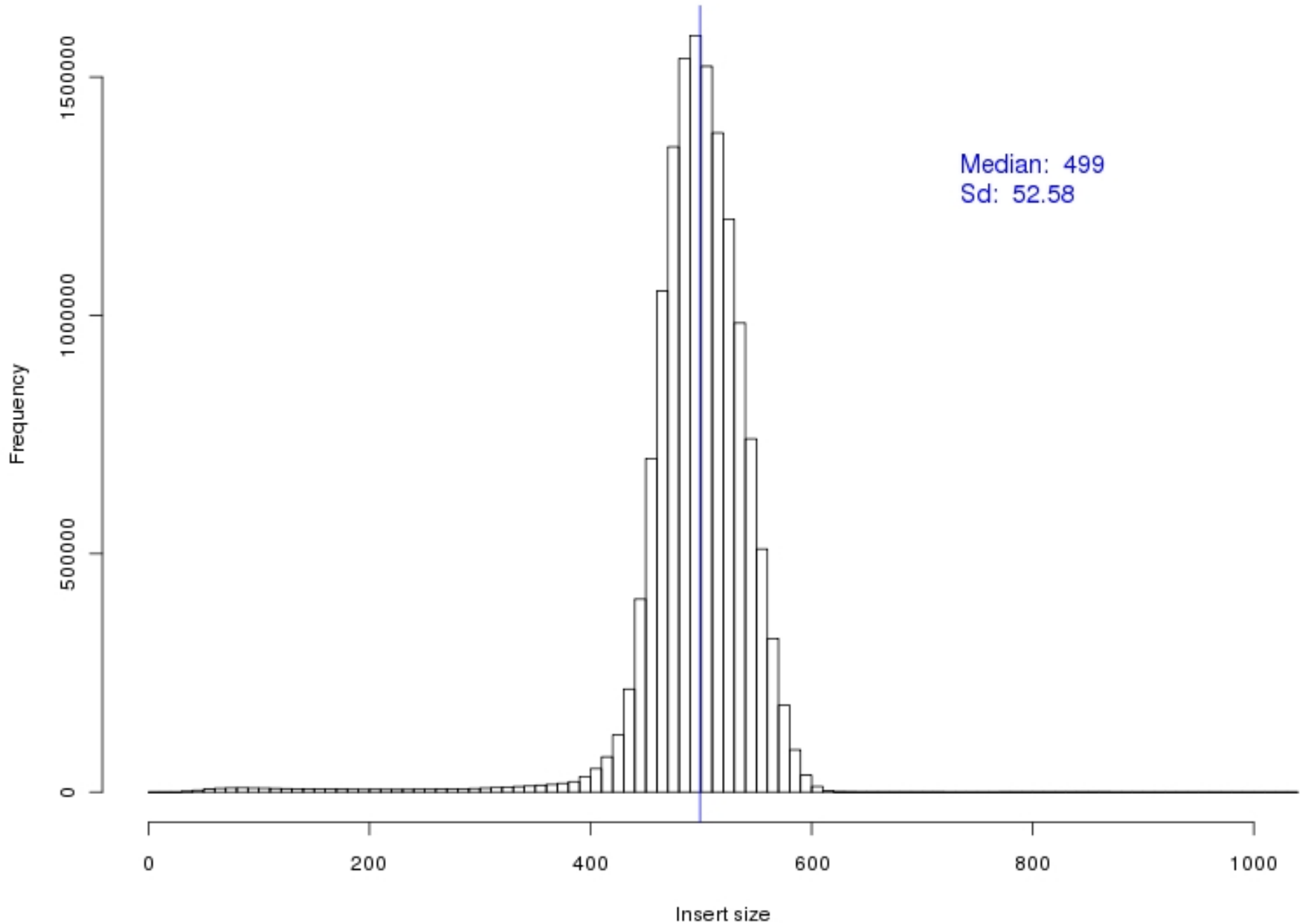


Source: illumina

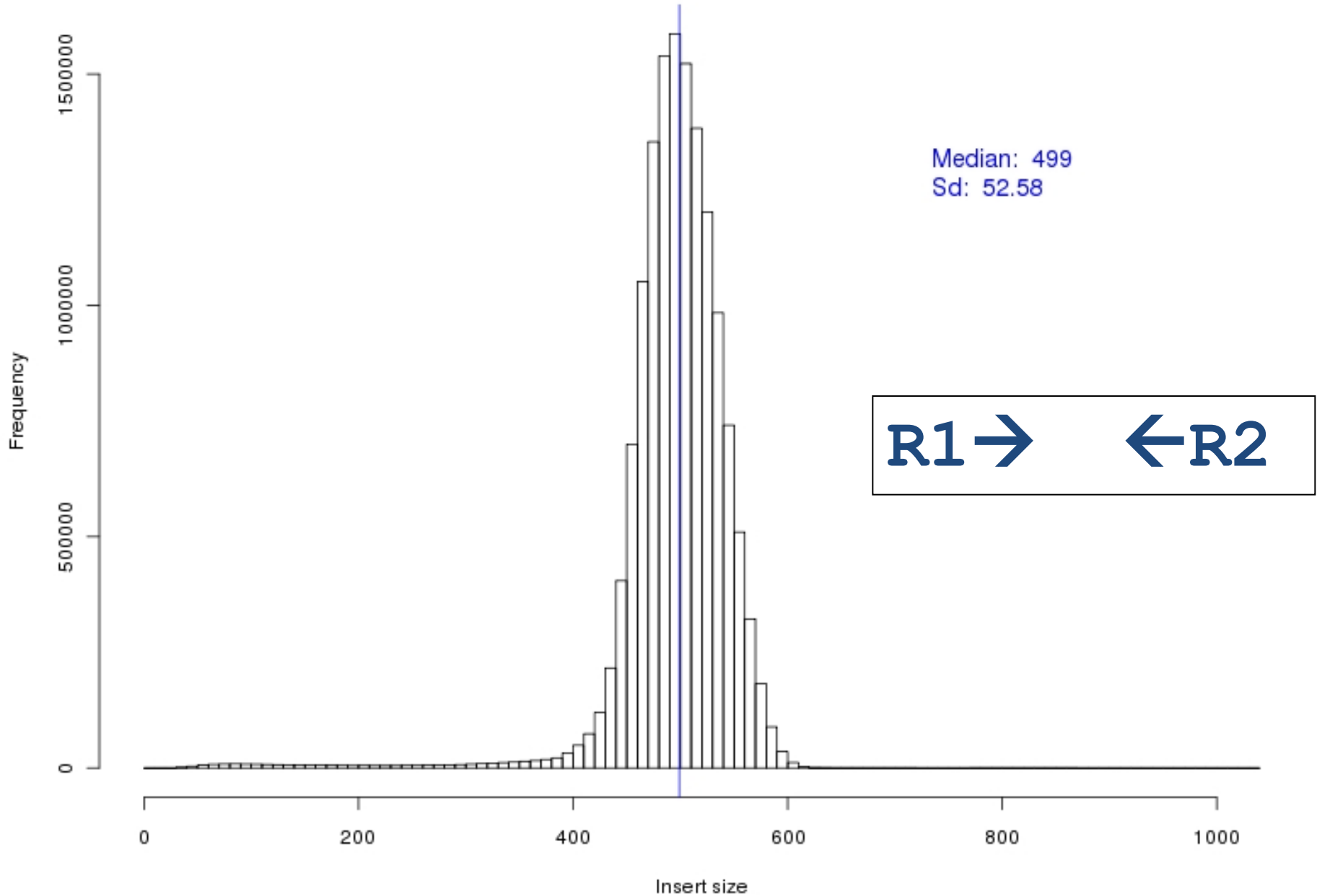
# Paired-End Sequencing



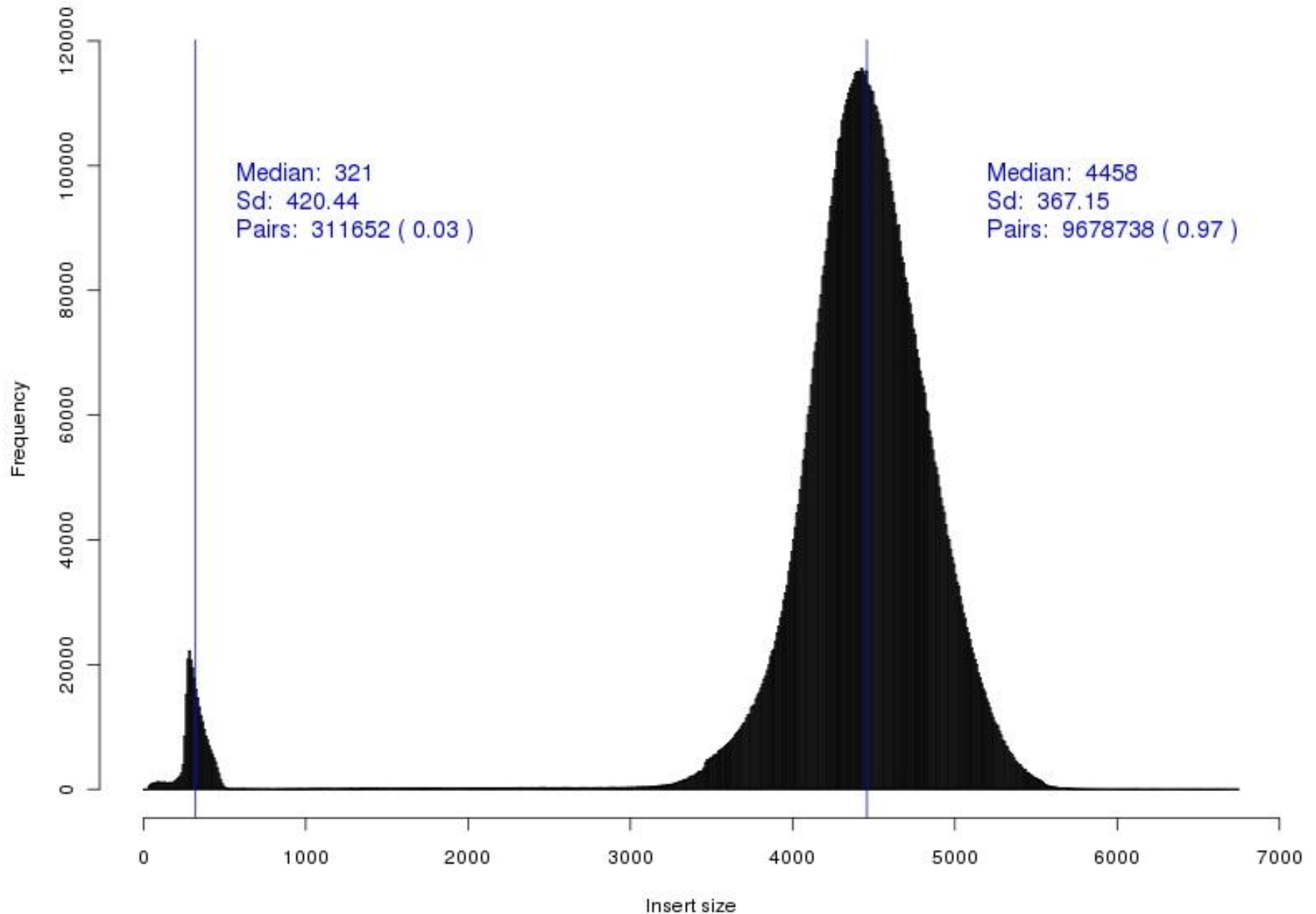
# Paired-End Libraries



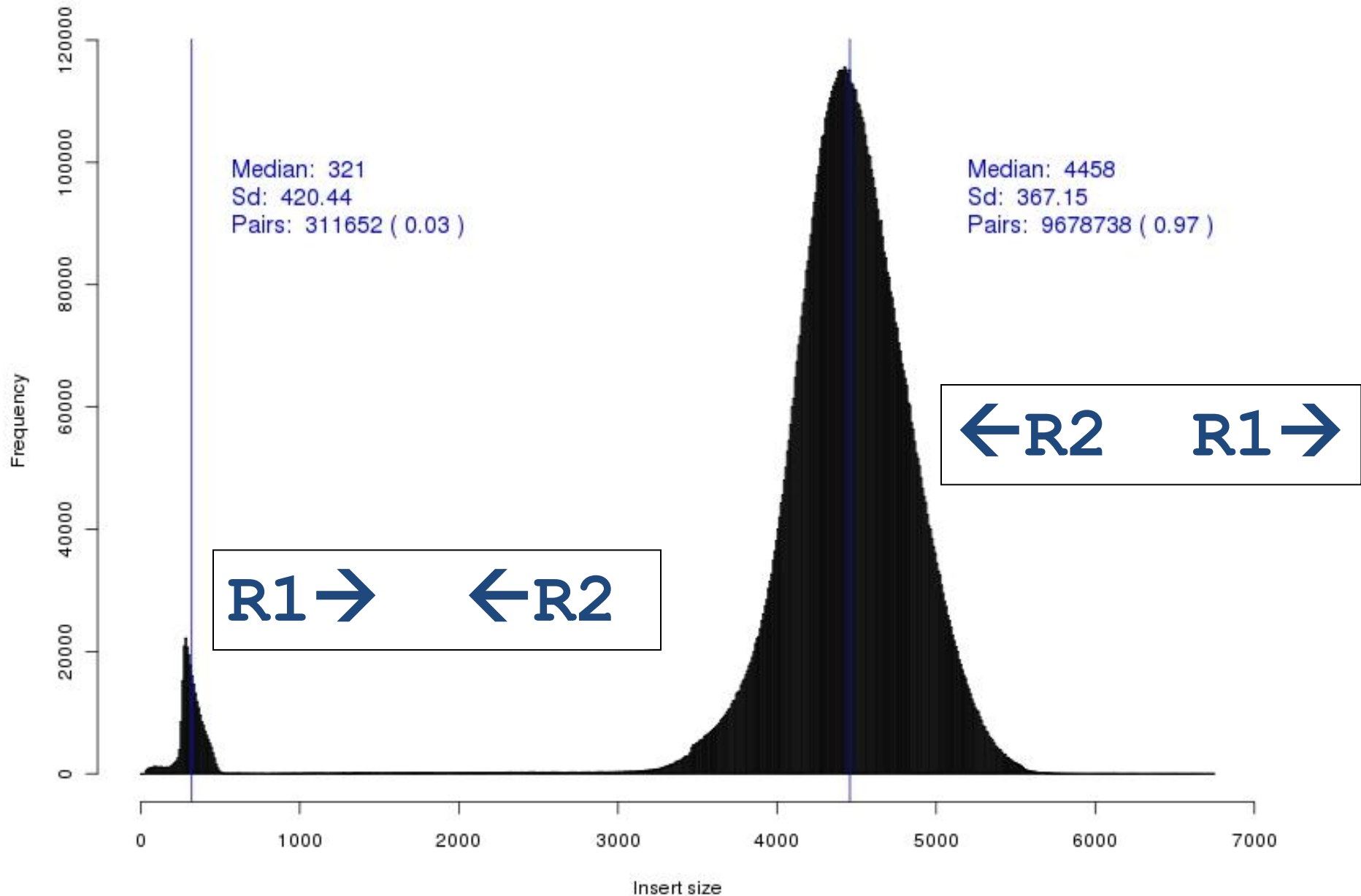
# Paired-End Libraries



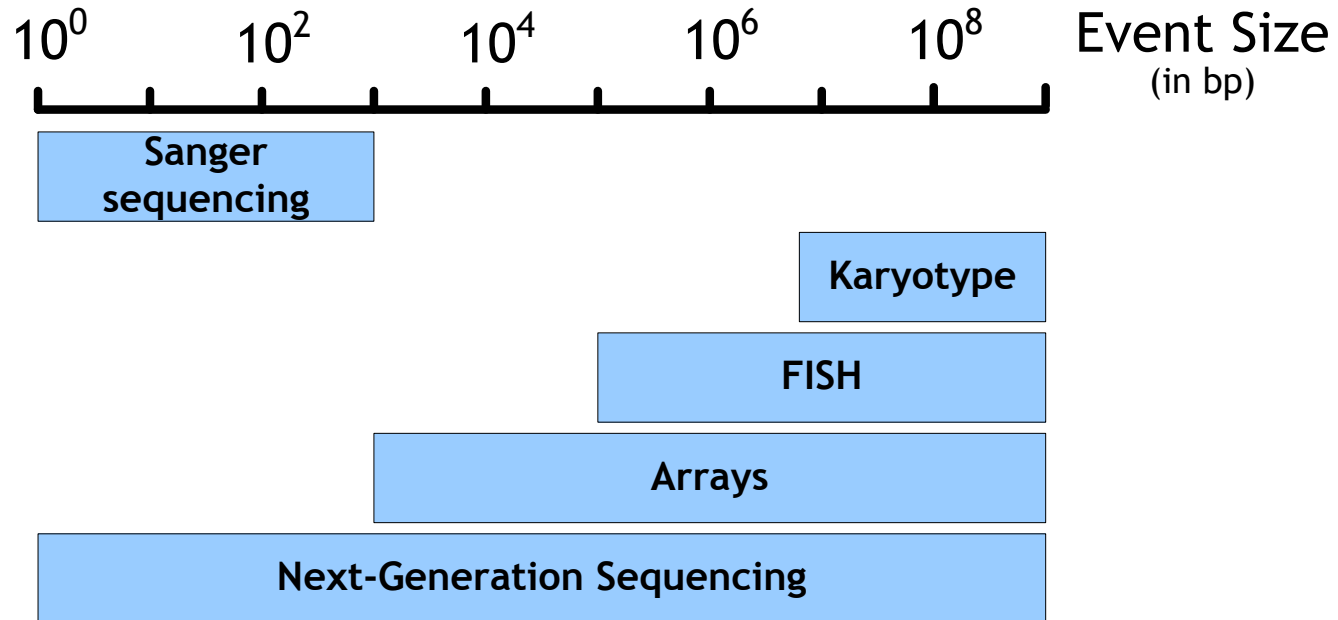
# Mate-Pair Libraries



# Mate-Pair Libraries

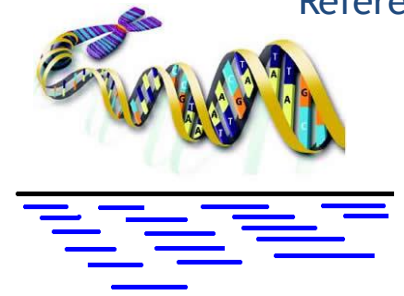


# Next-Generation Sequencing



- Limitations of Arrays
  - Lower resolution for genomic rearrangements
  - Balanced events (e.g., inversions) cannot be detected using signal intensity differences
  - No breakpoint information

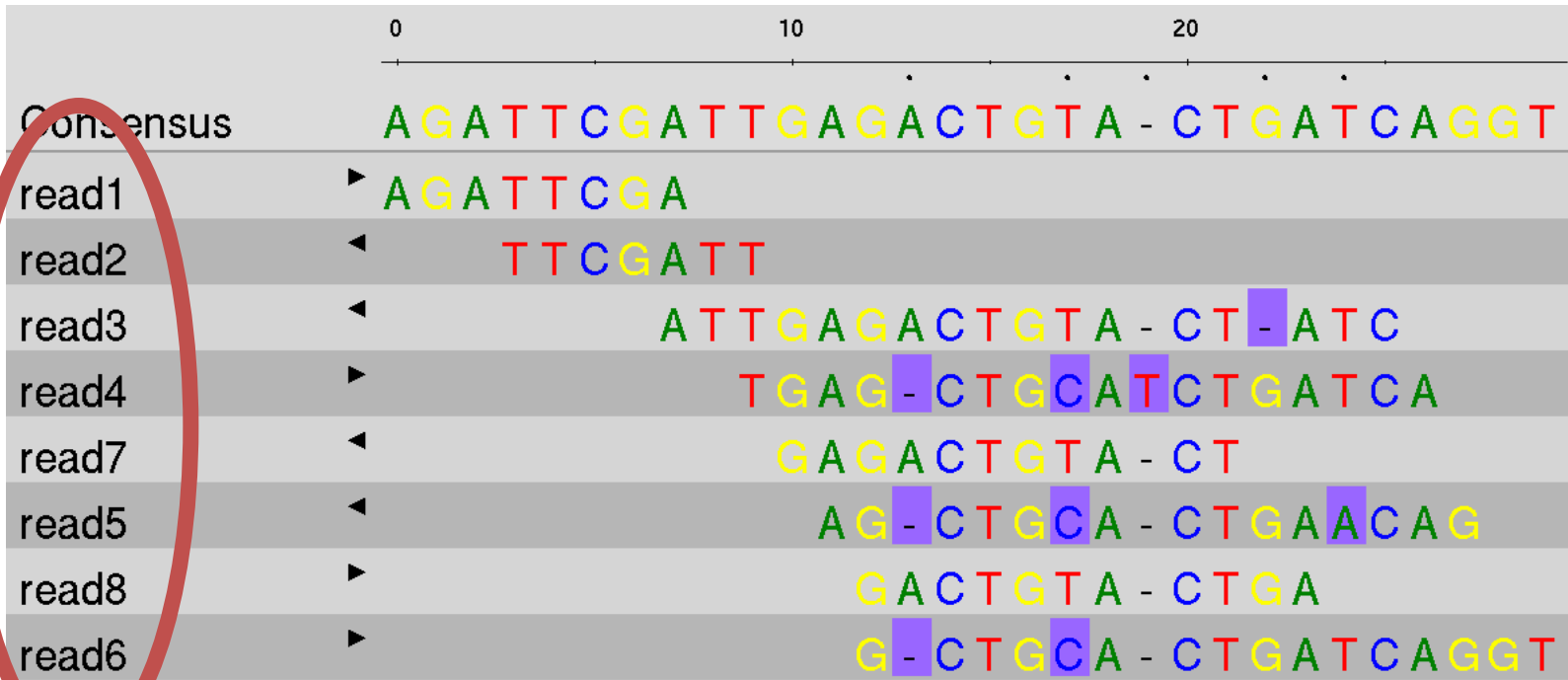
# SNP Calling



# SNP Calling

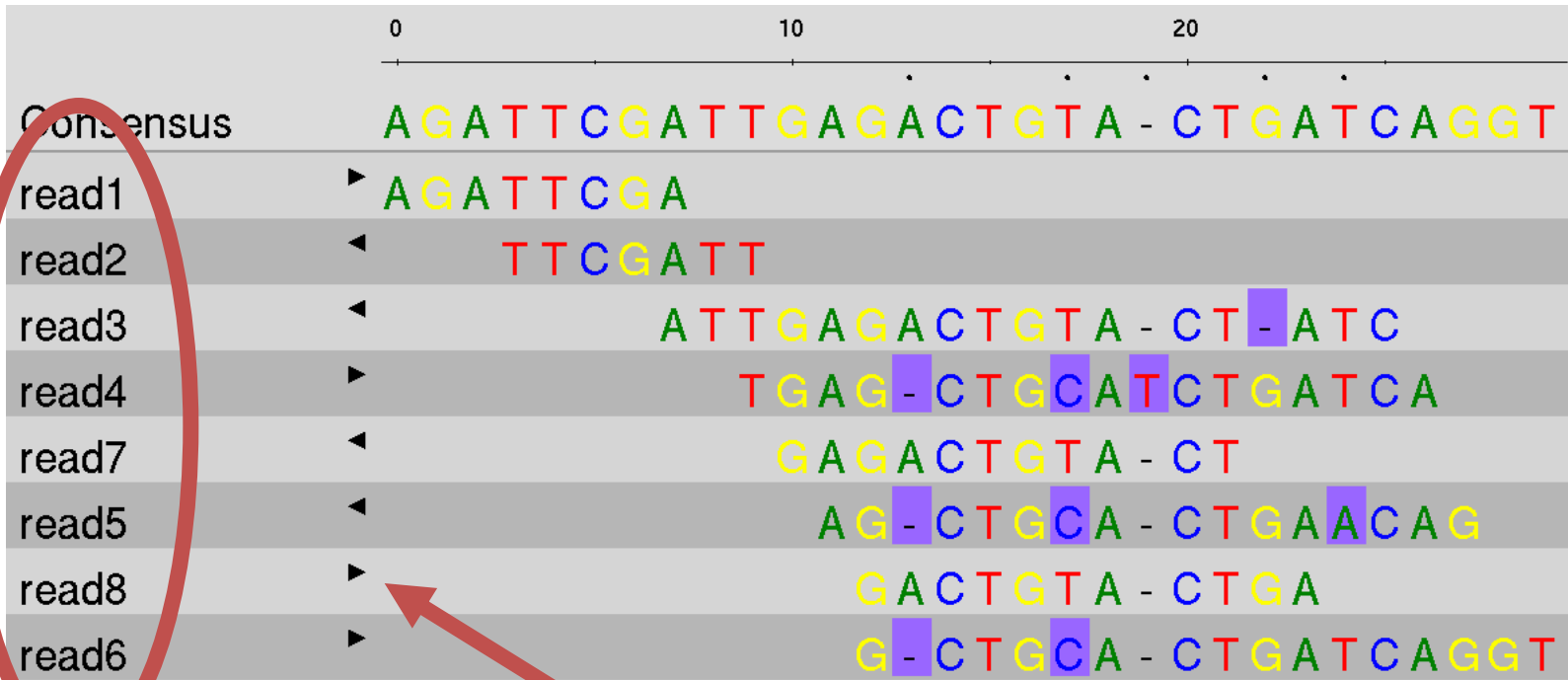
	0	10	20
Consensus	A G A T T C G A T T G A G A C T G T A - C T G A T C A G G T		
read1	▶ A G A T T C G A		
read2	◀ T T C G A T T		
read3	◀ A T T G A G A C T G T A - C T - A T C		
read4	▶ T G A G - C T G C A T C T G A T C A		
read7	◀ G A G A C T G T A - C T		
read5	◀ A G - C T G C A - C T G A A C A G		
read8	▶ G A C T G T A - C T G A		
read6	▶ G - C T G C A - C T G A T C A G G T		

# SNP Calling



Set of reads

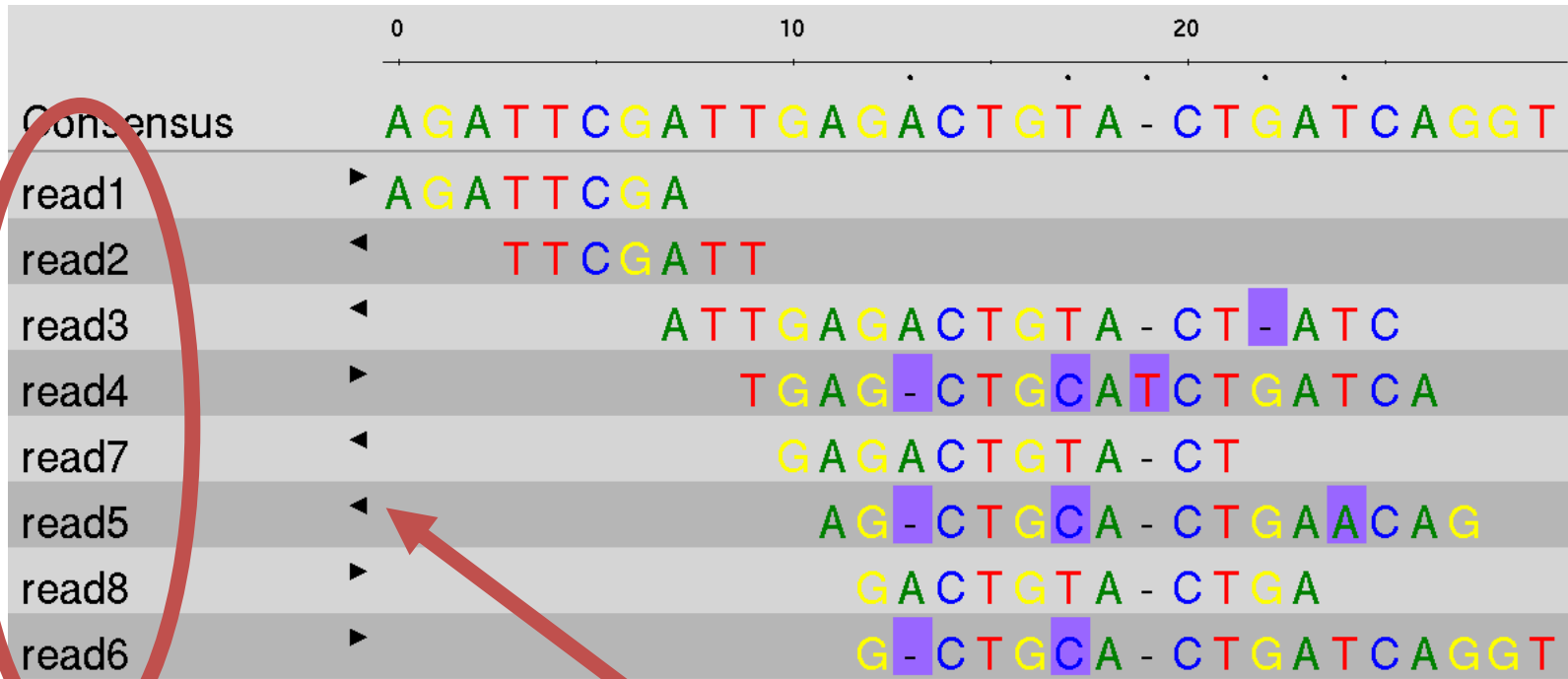
# SNP Calling



Set of reads

Forward

# SNP Calling

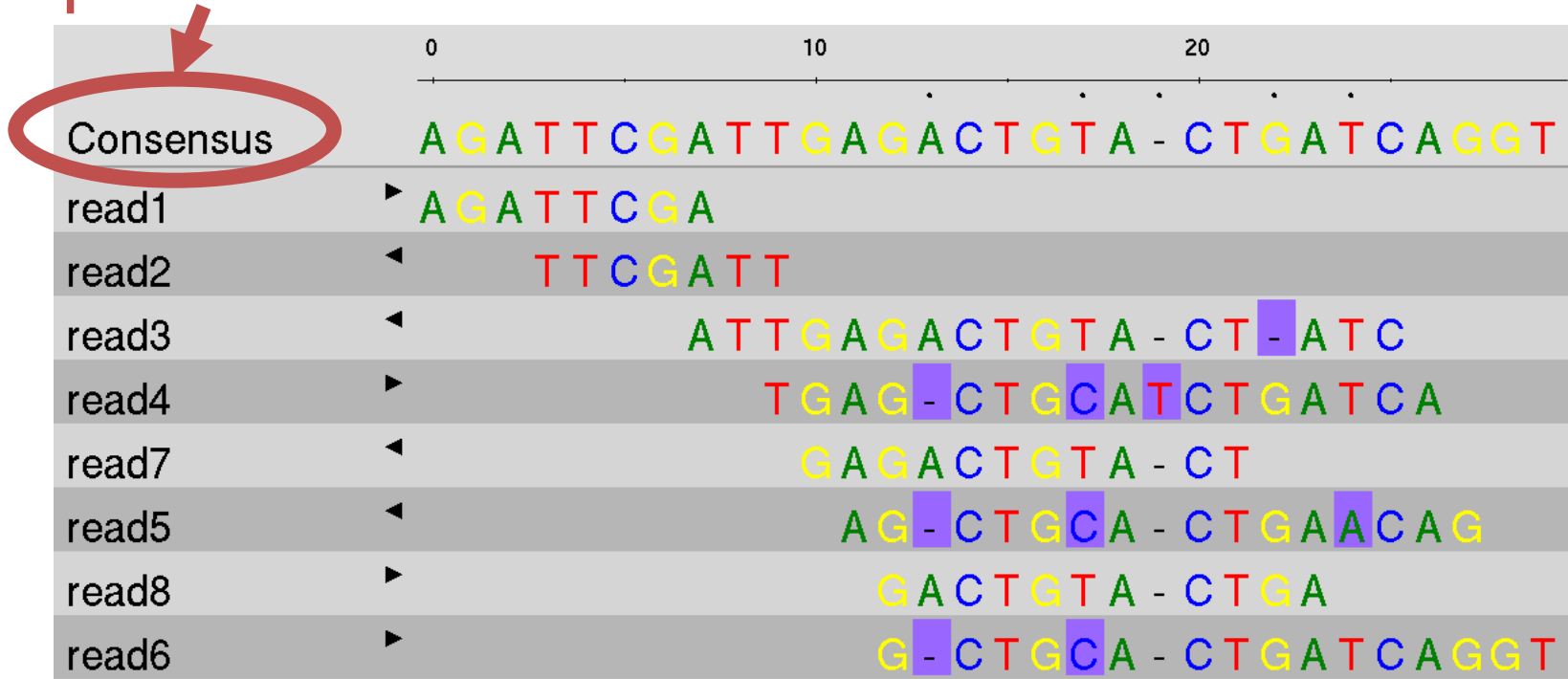


Set of reads

Reverse

# SNP Calling

Consensus  
sequence



# SNP Calling

	0	10	20
Consensus	A G A T T C G A T T G A G A C T G T A - C T G A T C A G G T		
read1	▶ A G A T T C G A		
read2	◀ T T C G A T T		
read3	◀ A T T G A G A C T G T A - C T - A T C		
read4	▶ T G A G - C T G C A T C T G A T C A		
read7	◀ G A G A C T G T A - C T		
read5	◀ A G - C T G C A - C T G A A C A G		
read8	▶ G A C T G T A - C T G A		
read6	▶ G - C T G C A - C T G A T C A G G T		



Variations: Indels & SNPs

# SNP Calling

	0	10	20
Consensus	A G A T T C G A T T G A G A C T G T A - C T G A T C A G G T		
read1	▶ A G A T T C G A		
read2	◀ T T C G A T T		
read3	◀ A T T G A G A C T G T A - C T - A T C		
read4	▶ T G A G - C T G C A T C G A T C A		
read7	◀ G A G A C T G T A - C T		
read5	◀ A G - C T G C A - C T G A A C A G		
read8	▶ G A C T G T A - C T G A		
read6	▶ G - C T C A C A - C T G A T C A G G T		

Sequencing errors: Insertions, deletions & basecalling errors

# SNP Calling

- Tools
  - GATK (Genome Analysis Toolkit)
  - SAMtools mpileup
  - CASAVA SNP Caller
  - Pyrobayes (454)
  - GigaBayes
  - Commercial packages (CLC Bio, Genomatix, etc.)

# SNP Calling

- Tools
  - GATK (Genome Analysis Toolkit)
  - SAMtools mpileup (MAQ SNP Caller)
  - CASAVA SNP Caller
  - Pyrobayes (454)
  - GigaBayes
  - Commercial packages (CLC Bio, Genomatix, etc.)
- My rough guess for the open-source community is that about 80% or 90% use GATK or SAMtools mpileup

# SNP Annotation

- Differentiating
  - Coding/Non-coding SNPs
  - Known/Unknown SNPs (dbSNP, 1kGP)
  - Homozygous/Heterozygous
- Annotating coding SNPs
  - Affected Gene/Transcript names
  - Silent/Non-silent mutations
  - Amino acid change
- Predicting possible impacts of an amino acid substitution

# Annotation Tools

- Annotation of coding SNPs
  - GATK (Genome Analysis Toolkit)
  - ANNOVAR
  - Commercial packages (CLC Bio, Genomatix, etc.)
- Predicting possible impacts of an amino acid substitution
  - PolyPhen2, Mutation Taster, Sift, etc.

# Raw SNP Calls, Human Genome

Sample	#Total	#dbSNPs	#Novel_SNPs	#Missense_SNPs	#Nonsense_SNPs	#Novel_missense_SNPs	#Novel_nonsense_SNPs
Sample1	3994410	3589636	404774	10970	93	1227	25
Sample2	3783797	3400926	382871	10416	90	1142	23

Even focusing only on the novel missense and nonsense SNPs leaves a list of several hundreds of SNPs!

# Raw SNP Calls, Human Genome

Sample	#Total	#dbSNPs	#Novel_SNPs	#Missense_SNPs	#Nonsense_SNPs	#Novel_missense_SNPs	#Novel_nonsense_SNPs
Sample1	3994410	3589636	404774	10970	93	1227	25
Sample2	3783797	3400926	382871	10416	90	1142	23

## Cross-comparisons!

At least Tumor vs. Germline, even better multiple matched tumor-germline pairs showing a similar phenotype.



# Short InDels

- What is obvious for you by looking at the multiple alignment isn't obvious to the read mapper because they have only the local view
  - One read against the reference at a time
- Hence, almost all short InDel callers start with local realignment
  - Time consuming (depending on the number of realignment windows)

# Short InDels - Tools

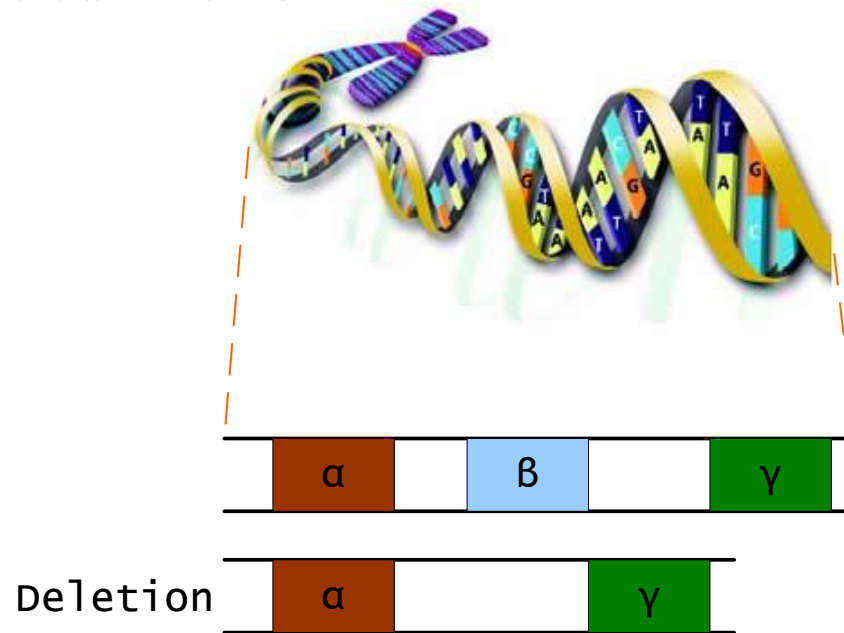
- Open-source tools
  - Dindel
  - Pindel
  - MoDIL
- Commercial packages
  - Maybe CLC Bio and others
- Indel calling has a higher false positive rate than SNP calling

# Structural Variants / Genomic Rearrangements

- Less established methodology
- Presumably a much higher false positive rate than in SNP and Short Indel Calling
- Requires the integration of various different signals (read-depth, abnormal paired-end mapping, split-reads, etc.)

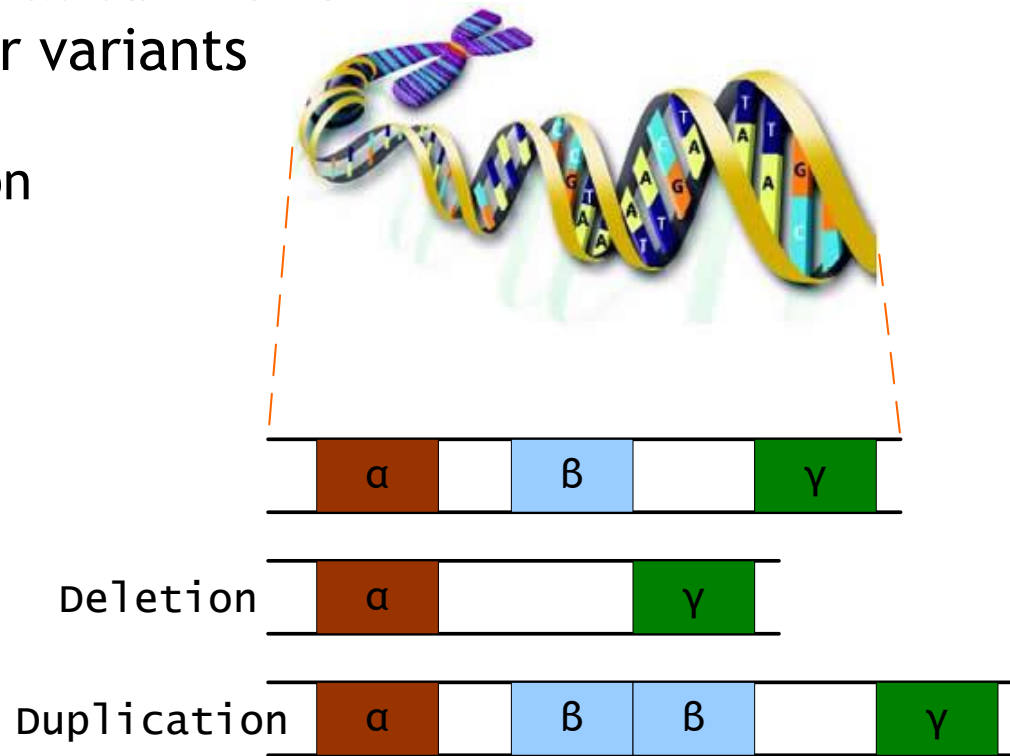
# Genomic Rearrangements

- 1 Kb to several Mb in size



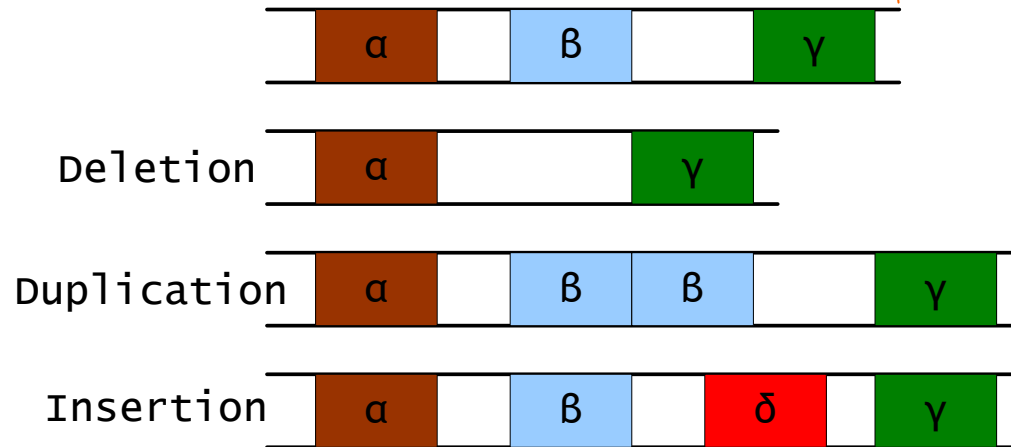
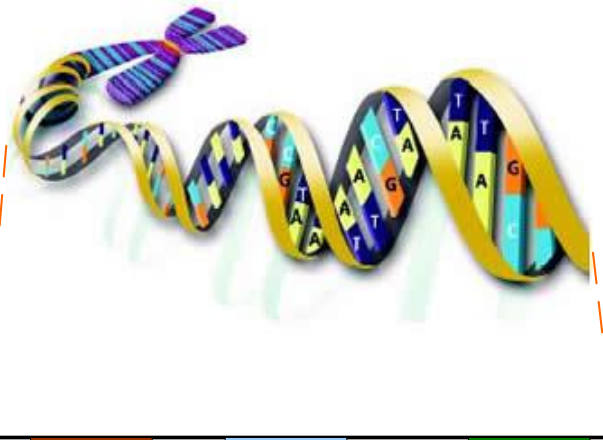
# Genomic Rearrangements

- 1 Kb to several Mb in size
- Copy number variants
  - Deletion
  - Duplication



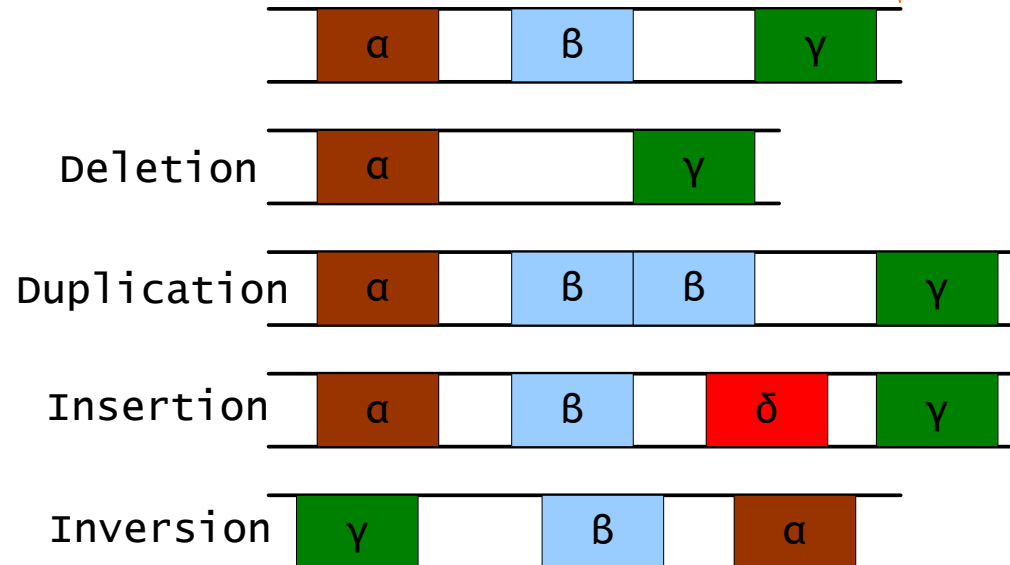
# Genomic Rearrangements

- 1 Kb to several Mb in size
- Copy number variants
  - Deletion
  - Duplication
- Insertion



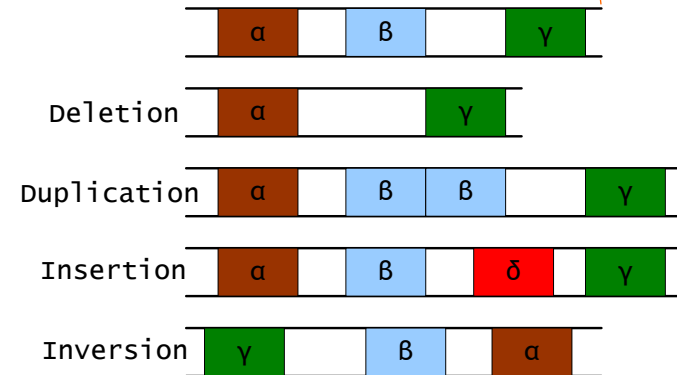
# Genomic Rearrangements

- 1 Kb to several Mb in size
- Copy number variants
  - Deletion
  - Duplication
- Insertion, Inversion



# Genomic Rearrangements

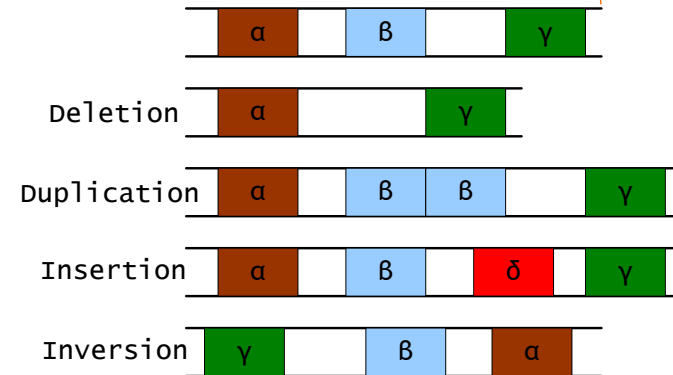
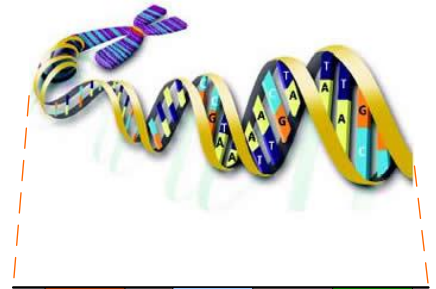
- 1 Kb to several Mb in size
- Copy number variants
  - Deletion
  - Duplication
- Insertion, Inversion, Translocation



# Genomic Rearrangements

- 1 Kb to several Mb in size
- Copy number variants
  - Deletion
  - Duplication
- Insertion, Inversion, Translocation
- More abundant than SNPs

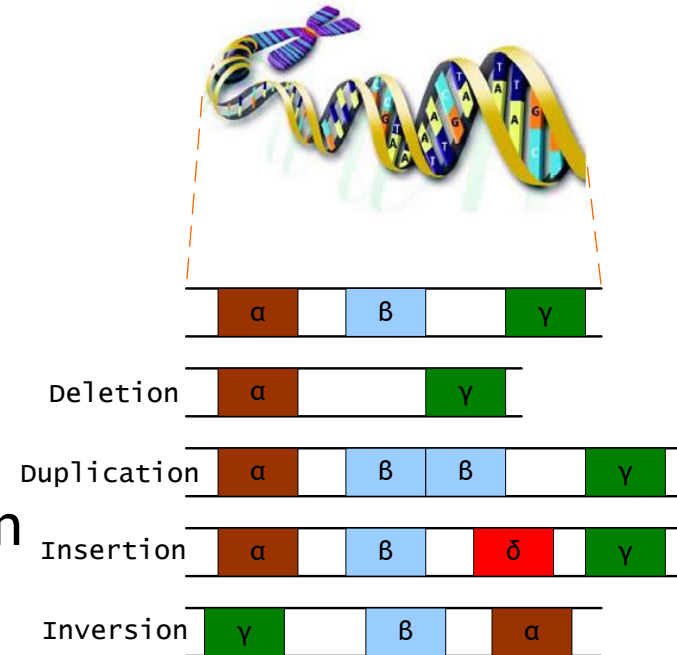
...ACGATACG...  
 ...ACGAGACG...



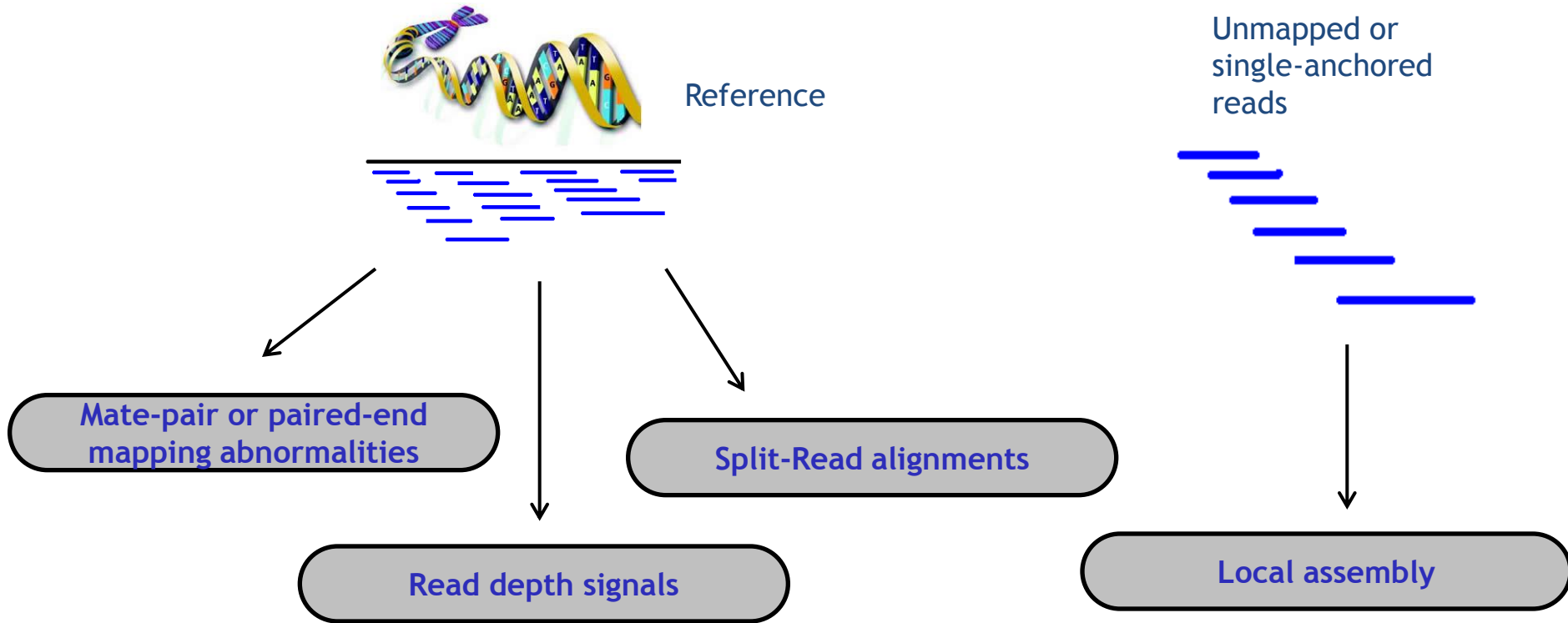
	SNPs	CNVs
Base pairs	2.5 Mb	4 Mb
% genome	0.08%	0.12%

# Genomic Rearrangements

- 1 Kb to several Mb in size
- Copy number variants
  - Deletion
  - Duplication
- Insertion, Inversion, Translocation
- More abundant than SNPs
- Either neutral or non-neutral in function
- Non-neutral mechanisms
  - Disrupting genes
  - Creating fusion genes
  - Copy number changes of dosage-sensitive genes



# Structural Variants



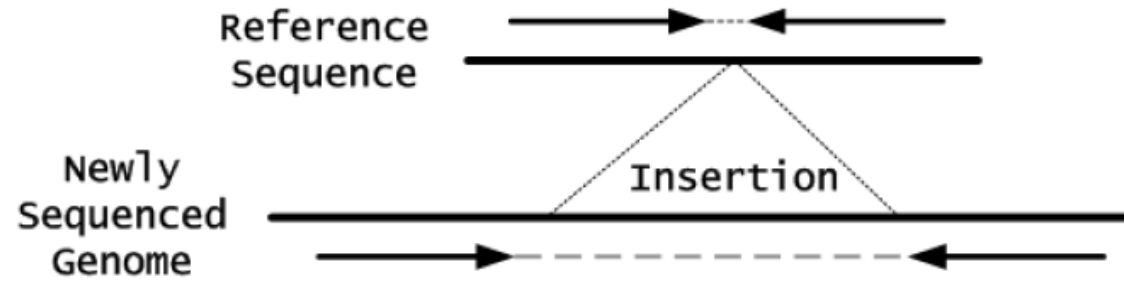
Mate-pair or paired-end mapping abnormalities

Read-depth signals

Split-read alignments

Local assembly

Insertion



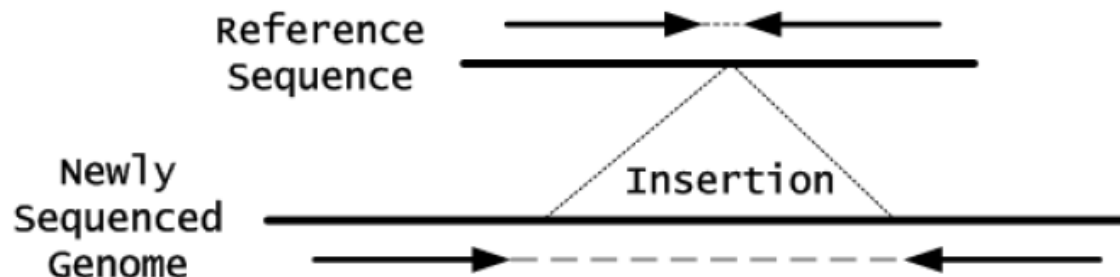
Mate-pair or paired-end mapping abnormalities

Read-depth signals

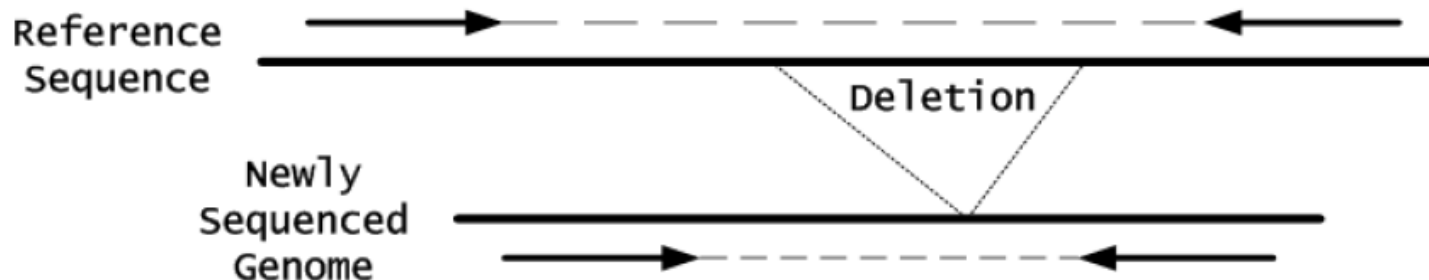
Split-read alignments

Local assembly

Insertion



Deletion



Mate-pair or paired-end mapping abnormalities

Read-depth signals

Split-read alignments

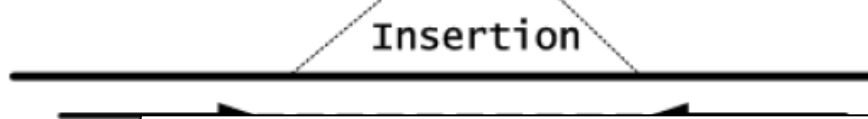
Local assembly

Insertion

Reference Sequence



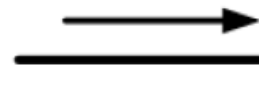
Newly Sequenced Genome



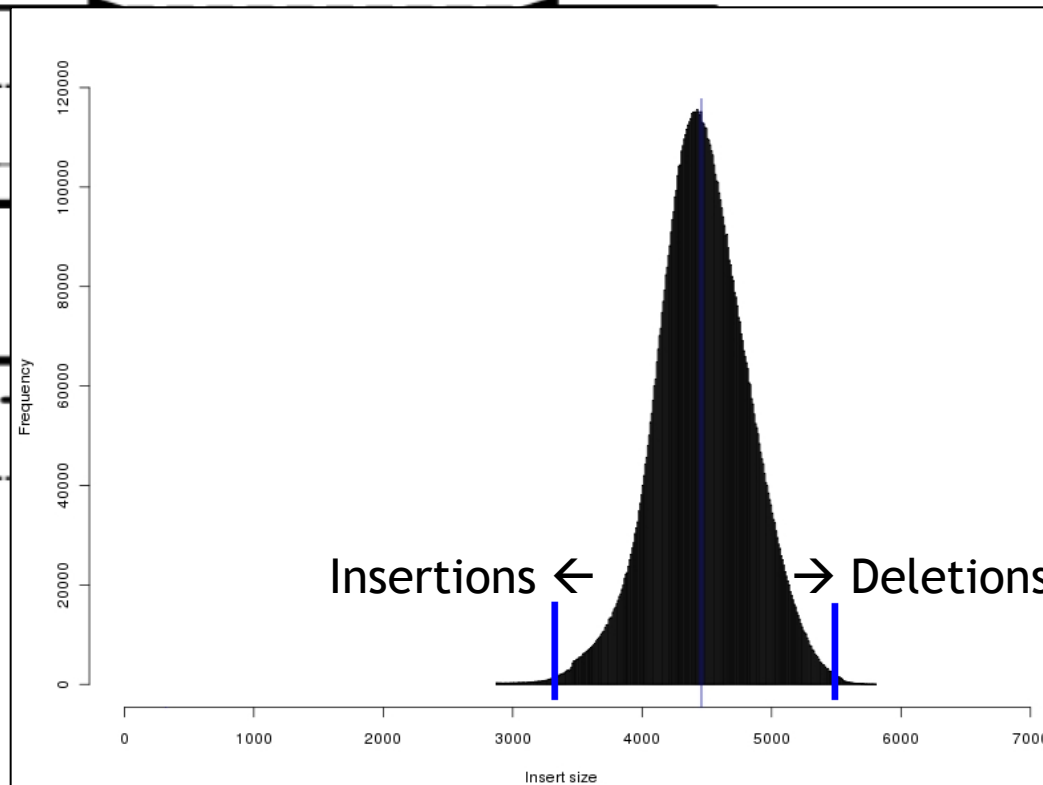
Insertion

Deletion

Reference Sequence



Newly Sequenced Genome



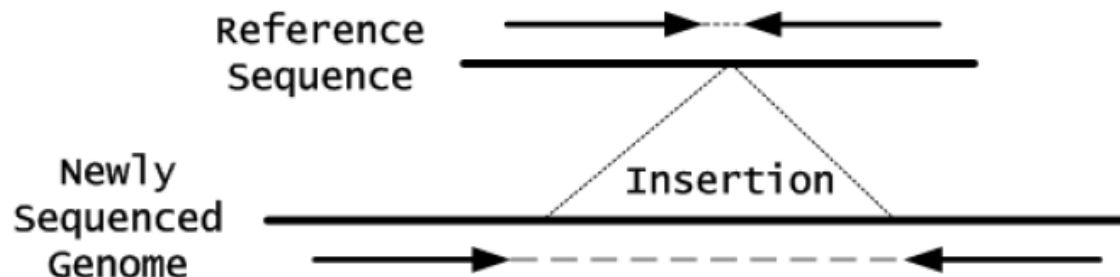
**Mate-pair or paired-end mapping abnormalities**

Read-depth signals

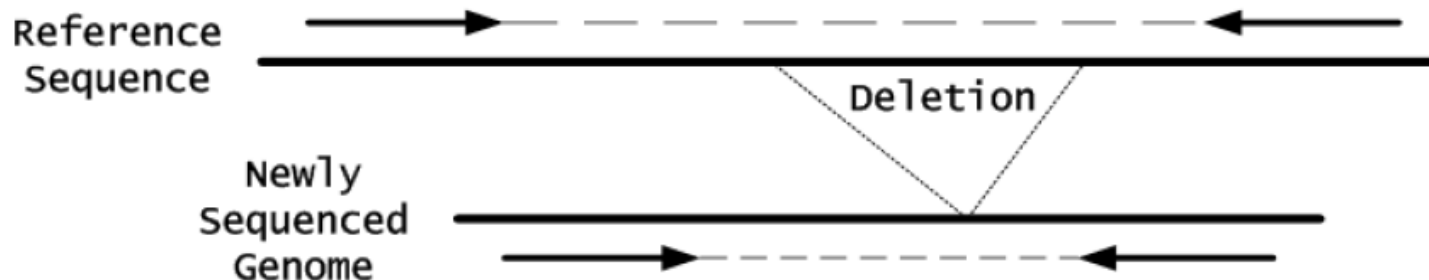
Split-read alignments

Local assembly

Insertion



Deletion



Inversion

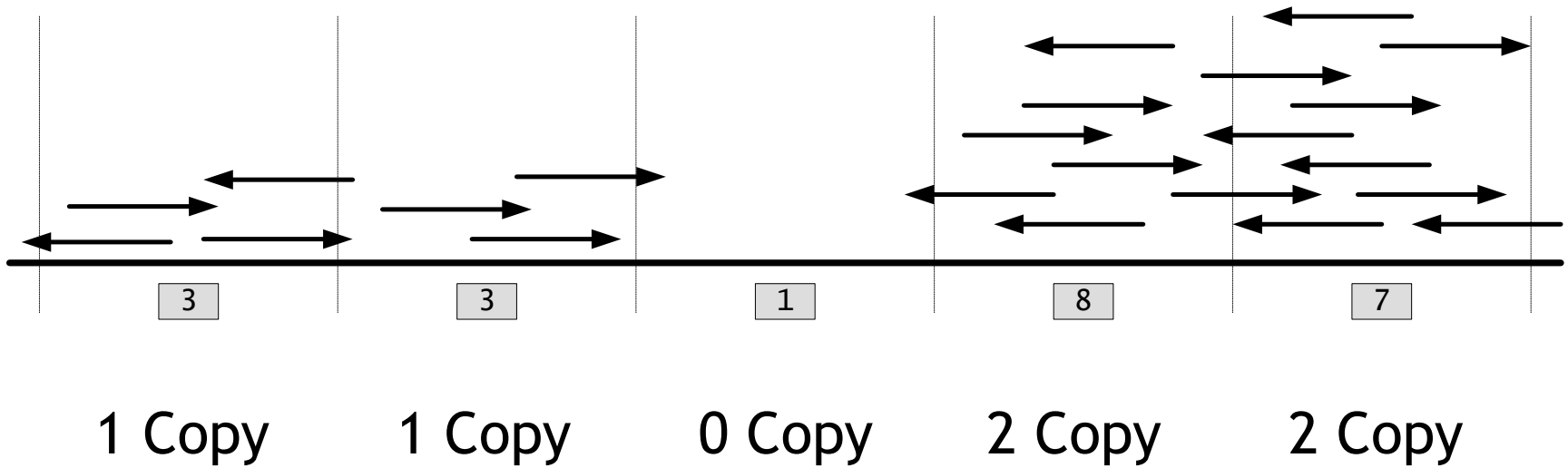


Mate-pair or paired-end mapping abnormalities

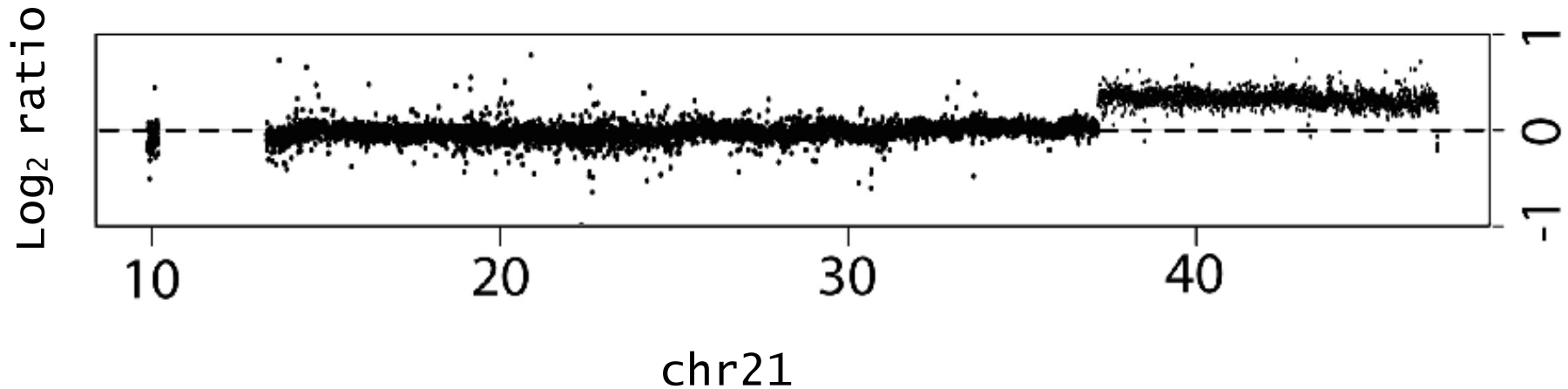
Read-depth signals

Split-read alignments

Local assembly



- Read-depth plots
  - Tumor vs. Germline



$$\log_2 \frac{\# \text{ Reads}_{Disease}}{\# \text{ Reads}_{Normal}}$$

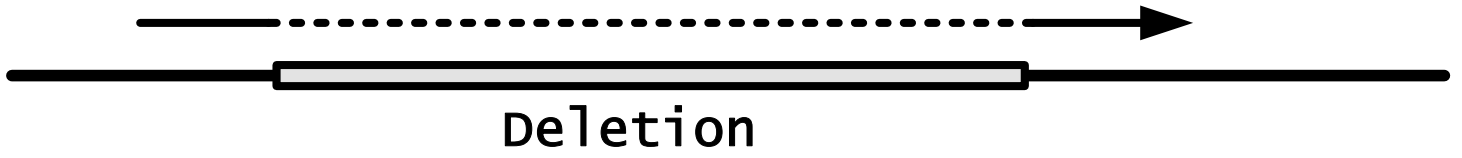
Mate-pair or paired-end mapping abnormalities

Read-depth signals

**Split-read alignments**

Local assembly

Reference  
Sequence

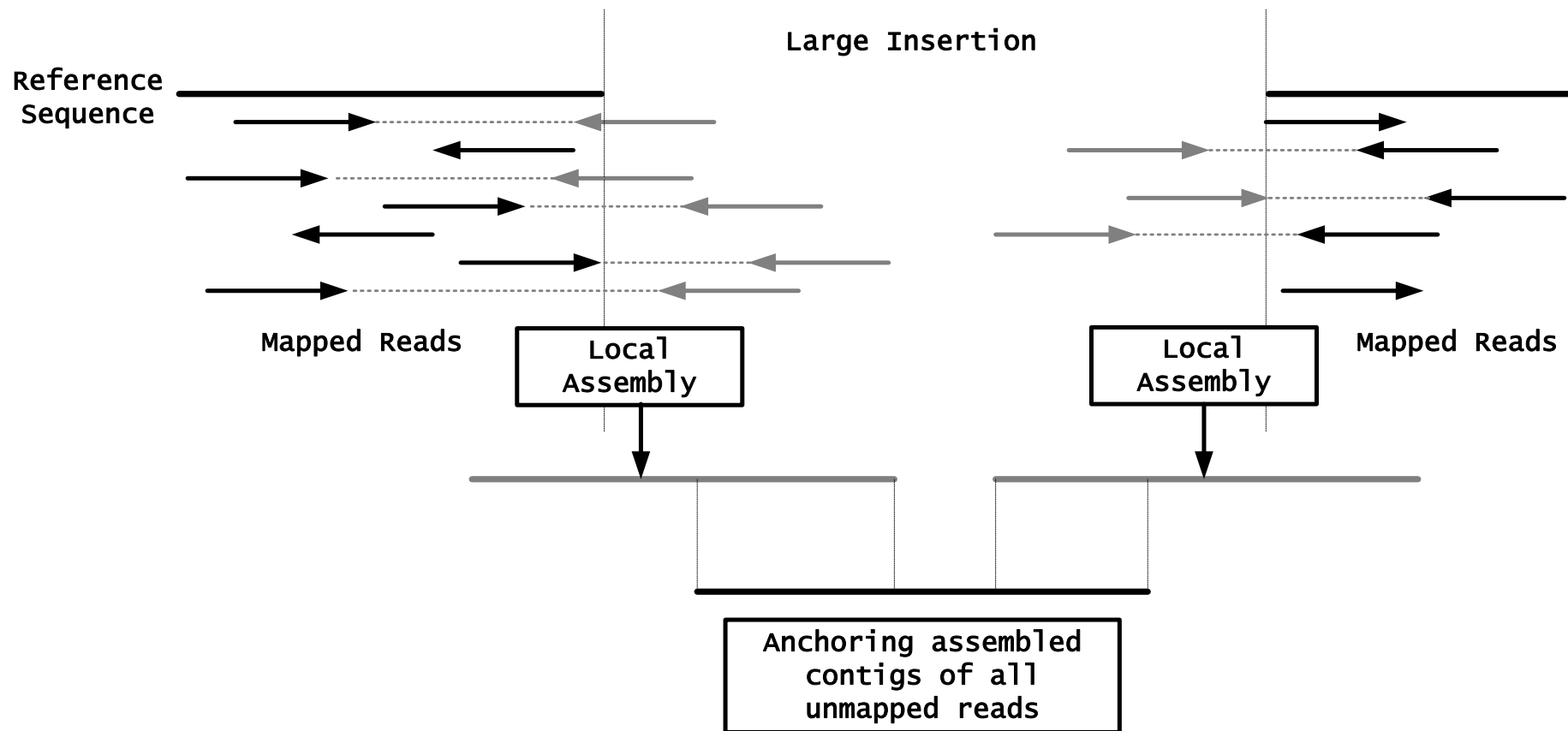


Mate-pair or paired-end mapping abnormalities

Read-depth signals

Split-read alignments

Local assembly



# Structural Variant Detection

	Paired-end mapping	Read-depth	Split-read	Local assembly
Deletion				
Short insertion (< Insert Size)				
Large insertion (> Insert Size)				
Inversion				
Tandem duplication				
Translocation				
Gain/Loss (CNVs)				
Region / Breakpoint	Region	Region	Breakpoint	Breakpoint

# Structural Variant Detection Tools

- Read-depth tools
  - CNVer, CNVnator, etc.
- Paired-end mapping
  - PEMer, Breakdancer
- Split-read
  - Age, Pindel
- We recently developed a paired-end, split-read detection pipeline that also takes into account read-depth

# Acknowledgment

## □ Genecore

Vladimir Benes  
Jonathon Blake  
Bettina Haase  
Dinko Pavlinic  
Jens Stolte  
Jürgen Zimmermann

## □ Korbelt group

Jan Korbelt  
Megumi Onishi-Seebacher  
Andreas Schlattl  
Adrian Stuetz  
Verena Tischler  
Thomas Zichner