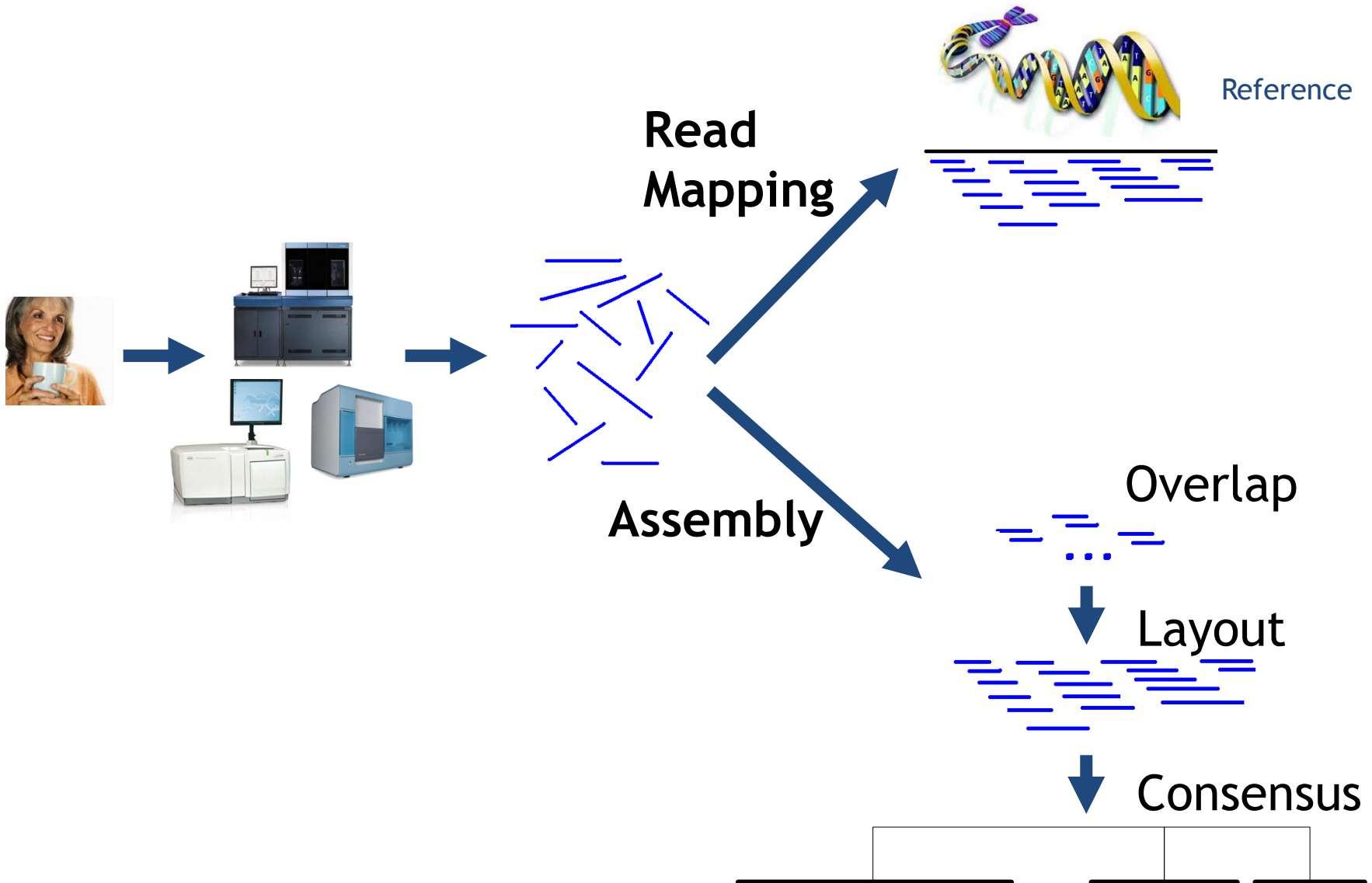


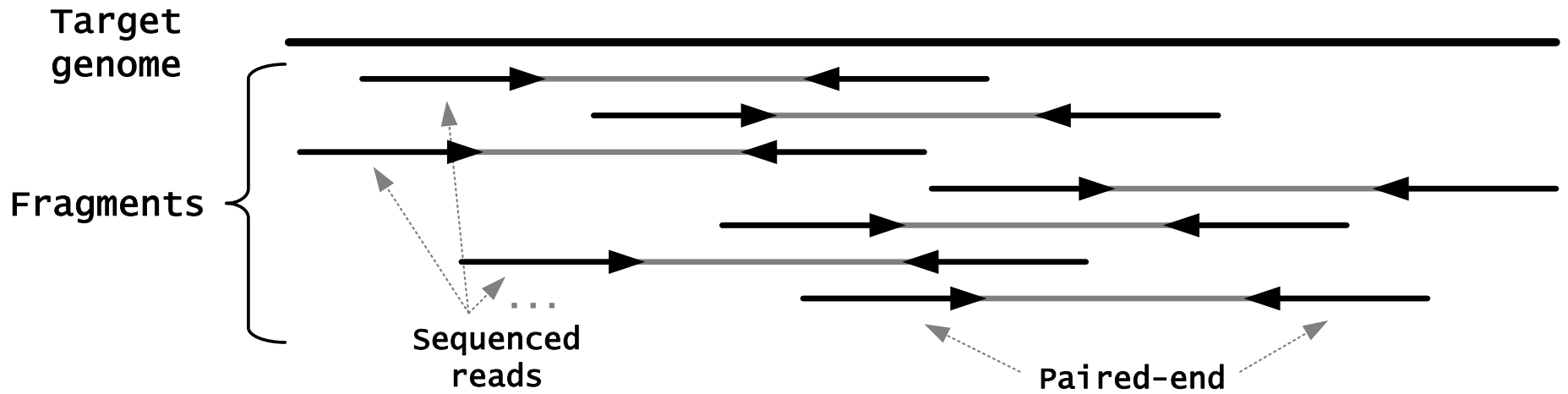
# Next Generation Sequencing Data Analysis

Tobias Rausch  
October 2011

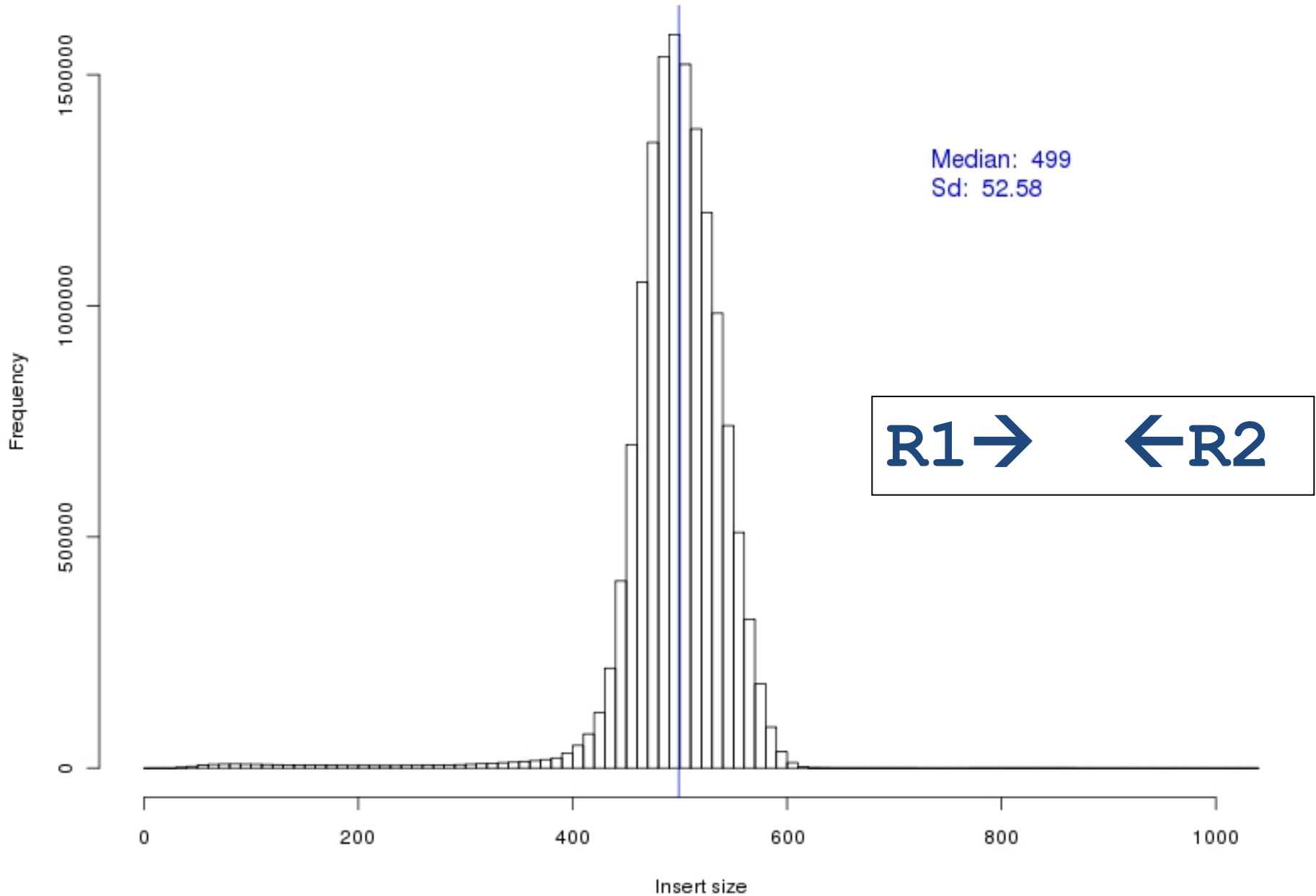
# NGS Data Analysis



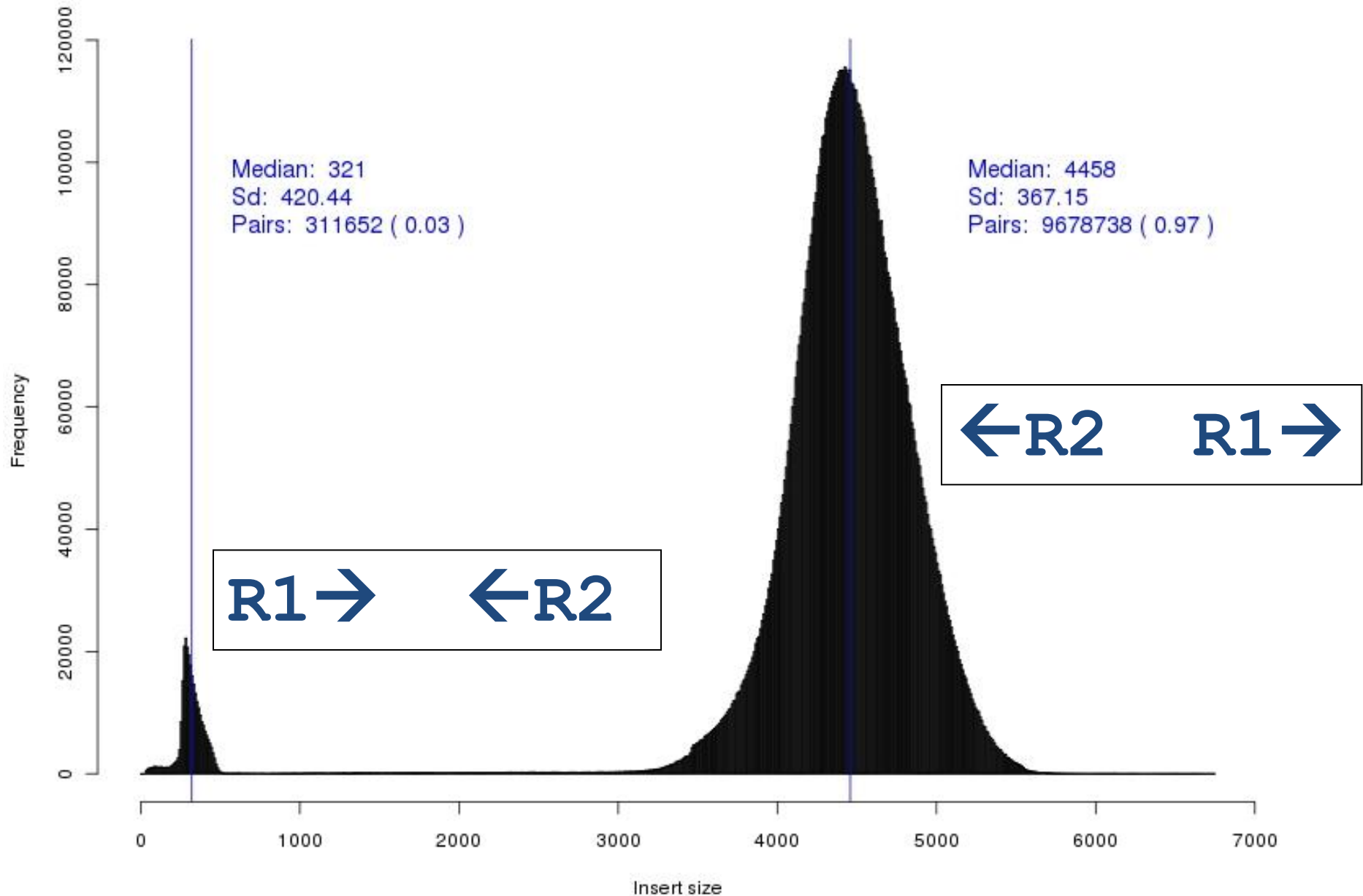
# Paired-End Sequencing



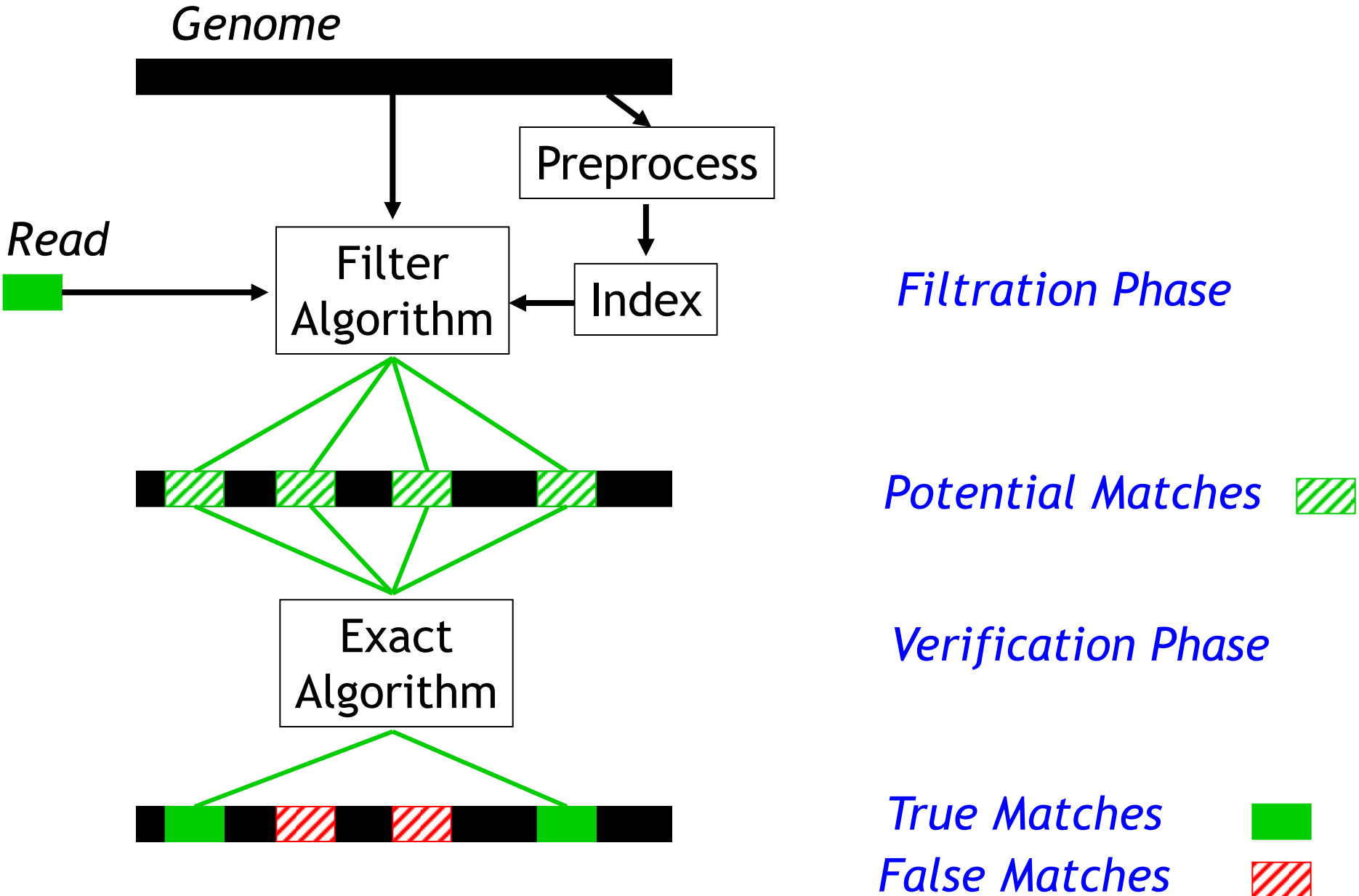
# Paired-End Libraries



# Mate-Pair Libraries



# Read Mapping

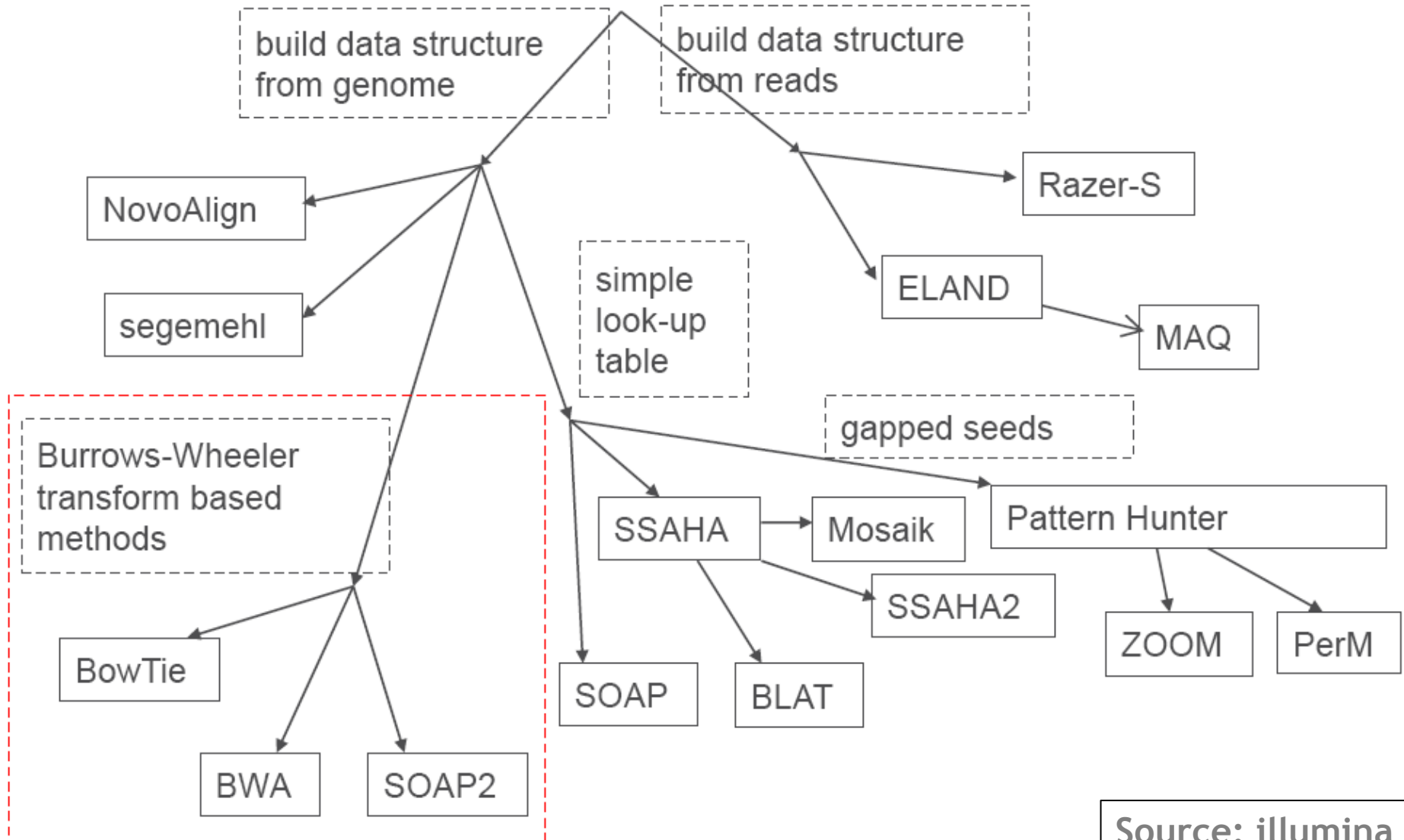


# Techniques

- Index
  - Hash tables, k-mer Index
  - Suffix trees, suffix arrays
  - Burrows-Wheeler-Transformation (BWT) of a suffix array
- Filtering Algorithms
  - Single or multiple seeds
  - Pigeonhole principle
  - q-gram filtering
- Verification
  - Simple seed-and-extend
  - Banded dynamic programming
  - Quality-based dynamic programming

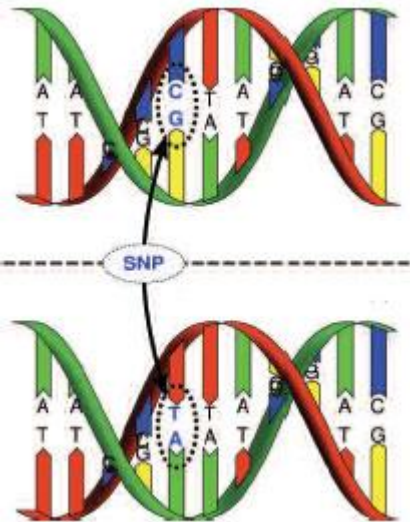


# Read Mappers

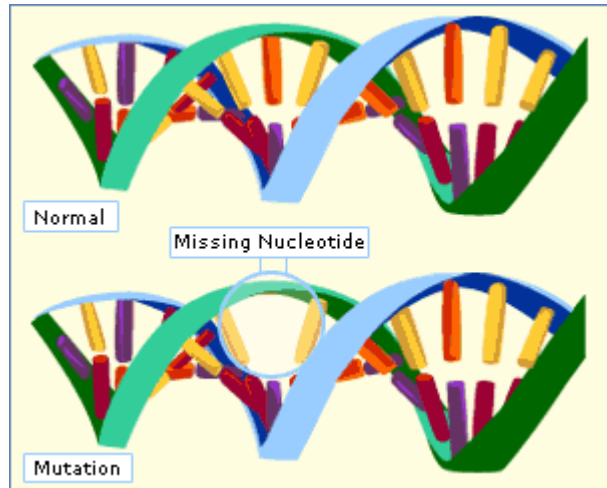


# Variant Calling

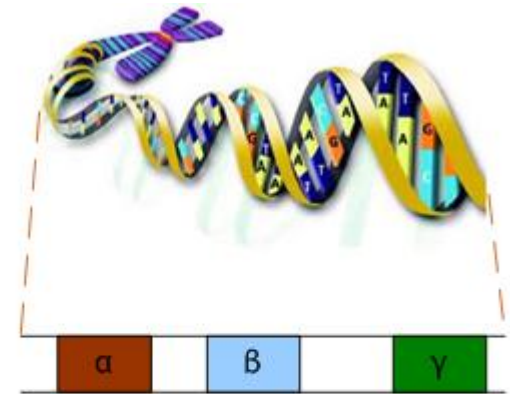
## SNVs



## Short Indels



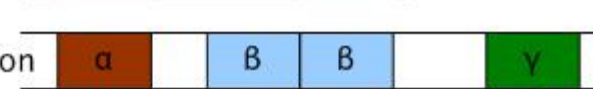
## Structural Variants



Deletion



Duplication



# SNV Calling

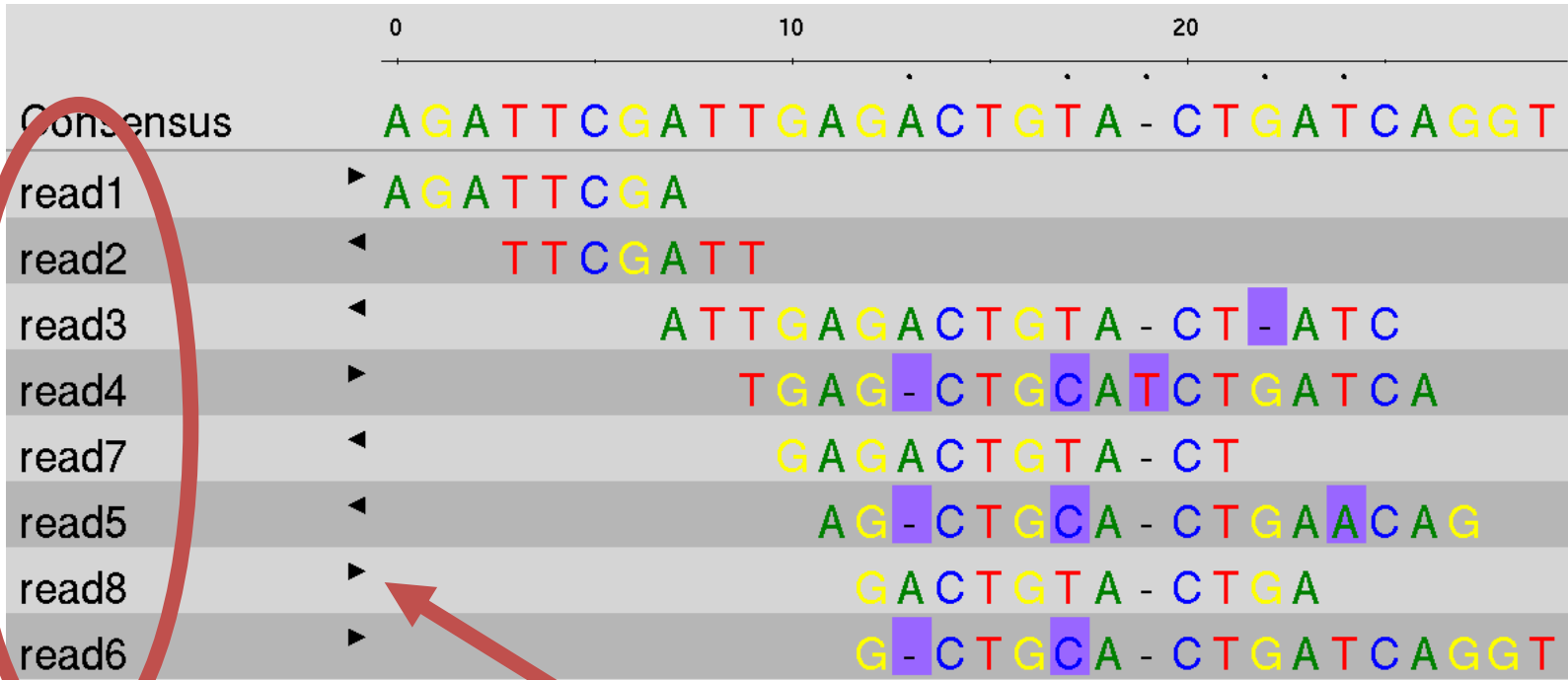
	0	10	20
Consensus	A G A T T C G A T T G A G A C T G T A - C T G A T C A G G T		
read1	▶ A G A T T C G A		
read2	◀ T T C G A T T		
read3	◀ A T T G A G A C T G T A - C T - A T C		
read4	▶ T G A G - C T G C A T C T G A T C A		
read7	◀ G A G A C T G T A - C T		
read5	◀ A G - C T G C A - C T G A A C A G	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

# SNV 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		

Set of reads

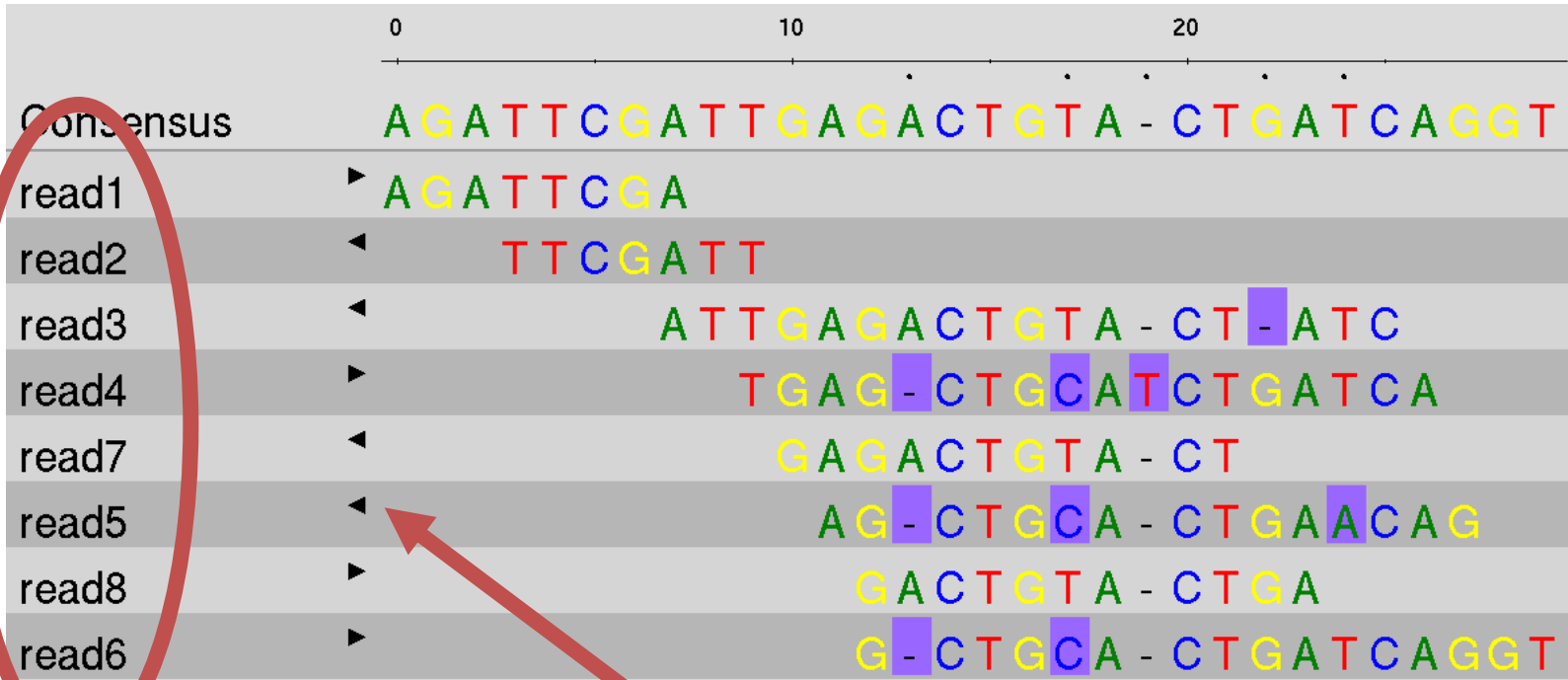
# SNV Calling



Set of reads

Forward

# SNV Calling

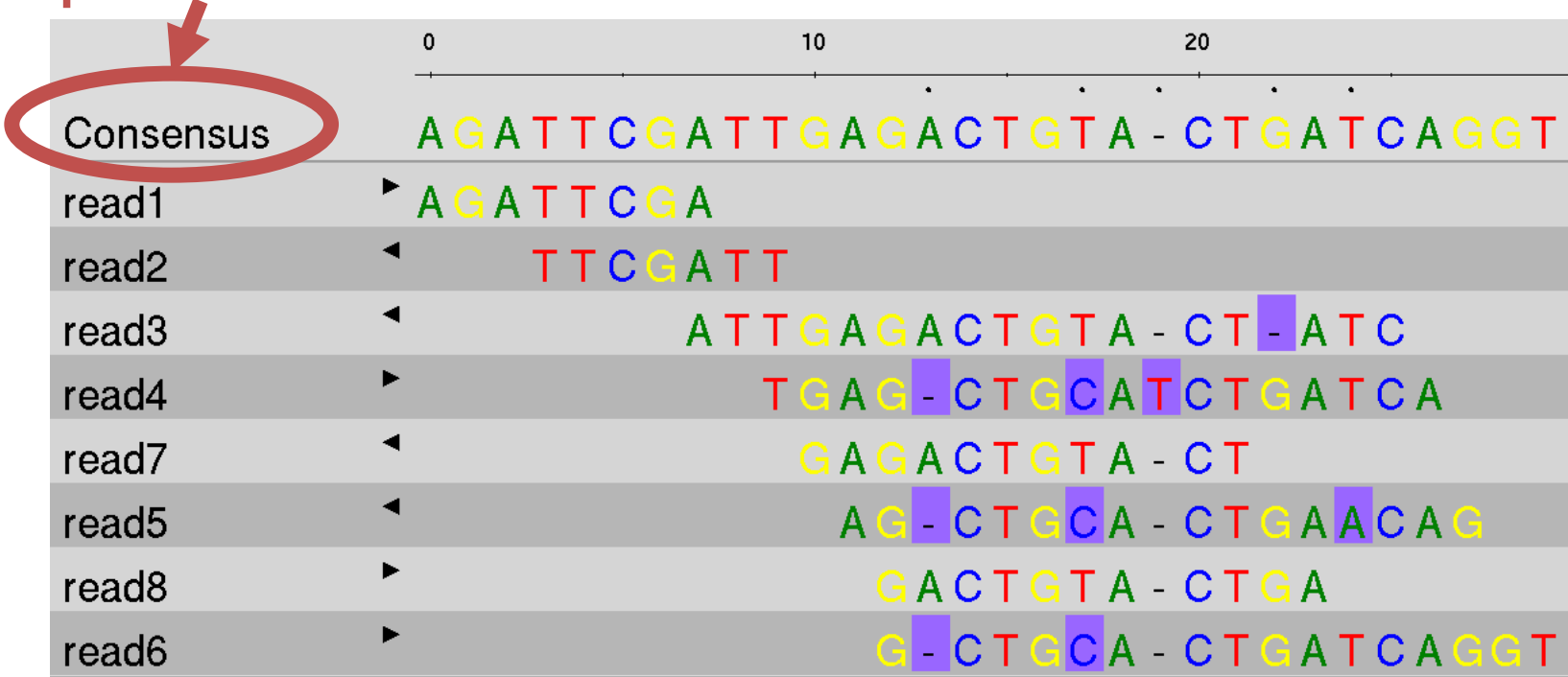


Set of reads

Reverse

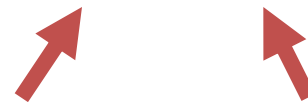
# SNV Calling

Consensus  
sequence



# SNV 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 & SNVs

# SNV 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

# SNV Calling

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

# SNV 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 (Sift, Polyphen, ...)

# Variant Calling by Consensus

## - Human, SNV Calls, chr19 -

### Mpileup/Annovar

Sample	#Total	#dbSNPs	#Novel_SNVs	#Synonymous_SNVs	#Nonsynonymous_SNVs	#Novel_synonymous_SNVs	#Novel_nonsynonymous_SNVs
Sample1	69682	66122	3560	895	880	31	62
Sample2	69092	65590	3502	897	869	28	56

### GATK

Sample	#Total	#dbSNPs	#Novel_SNVs	#Missense_SNVs	#Nonsense_SNVs	#Novel_missense_SNVs	#Novel_nonsense_SNVs
Sample1	80188	71113	9075	879	6	86	1
Sample2	80008	71133	8875	884	5	76	1

# SNV Call Overlap

- Sample1
  - 69682 Mpileup/Annovar Calls
  - 80188 GATK Calls
  - 67149 calls in common (96%, 83%)
- Sample2
  - 69092 Mpileup/Annovar Calls
  - 80008 GATK Calls
  - 66855 calls in common (97%, 84%)

# Raw SNV Calls, Human Genome

Sample	#Total	#dbSNPs	#Novel_SNVs	#Missense_SNVs	#Nonsense_SNVs	#Novel_missense_SNVs	#Novel_nonsense_SNVs
Sample1	3994410	3589636	404774	10970	93	1227	25
Sample2	3783797	3400926	382871	10416	90	1142	23

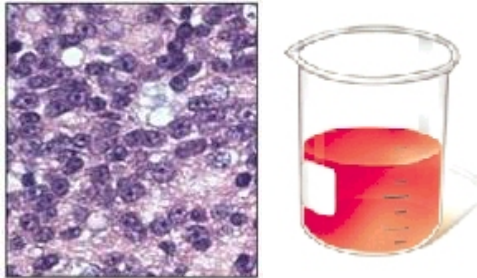
Even focusing only on the novel missense and nonsense SNVs leaves a list of several hundreds of SNVs!

Cross-comparisons!

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

# Comparing / Annotating Variant Calls

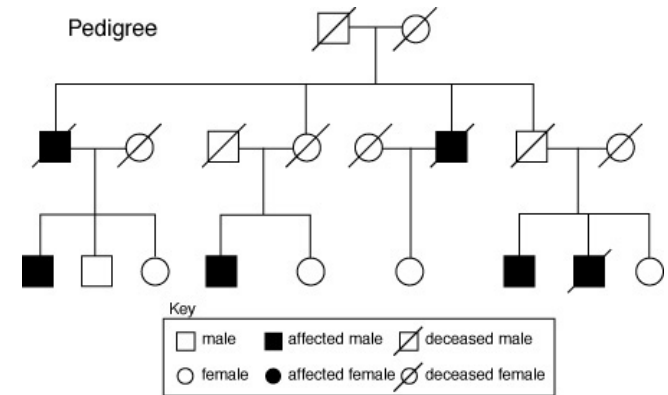
## Matched Tumor – Normal data



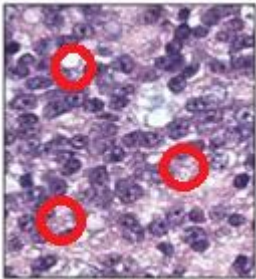
## Time-course data

- 1) Initial Tumor
- 2) Remission
- 3) Relapse
- 4) Remission
- 5) Secondary Tumor

## Pedigree data



## Multiple Tumor Probes



## Population Scale



## Constraints

- Tumor heterogeneity
- Tumor purity
- Aneuploidy
- Sample availability



# Short InDels

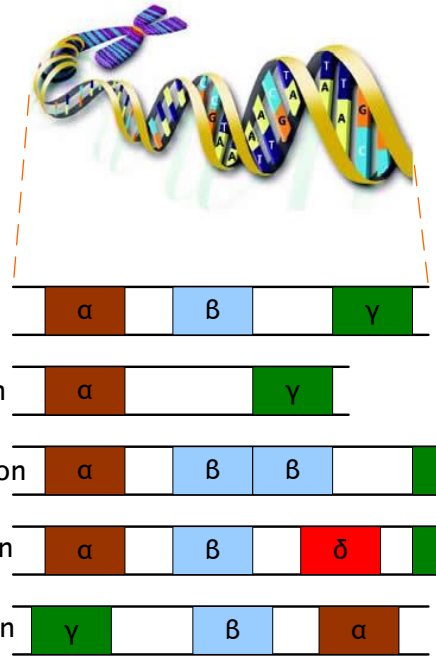
- What is obvious for you by looking at the multiple alignment isn't obvious to the read mapper because they have only the local view
  - One read against the reference at a time
- Hence, almost all short InDel callers start with local realignment
  - Time consuming (depending on the number of realignment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

# Short InDels - Tools

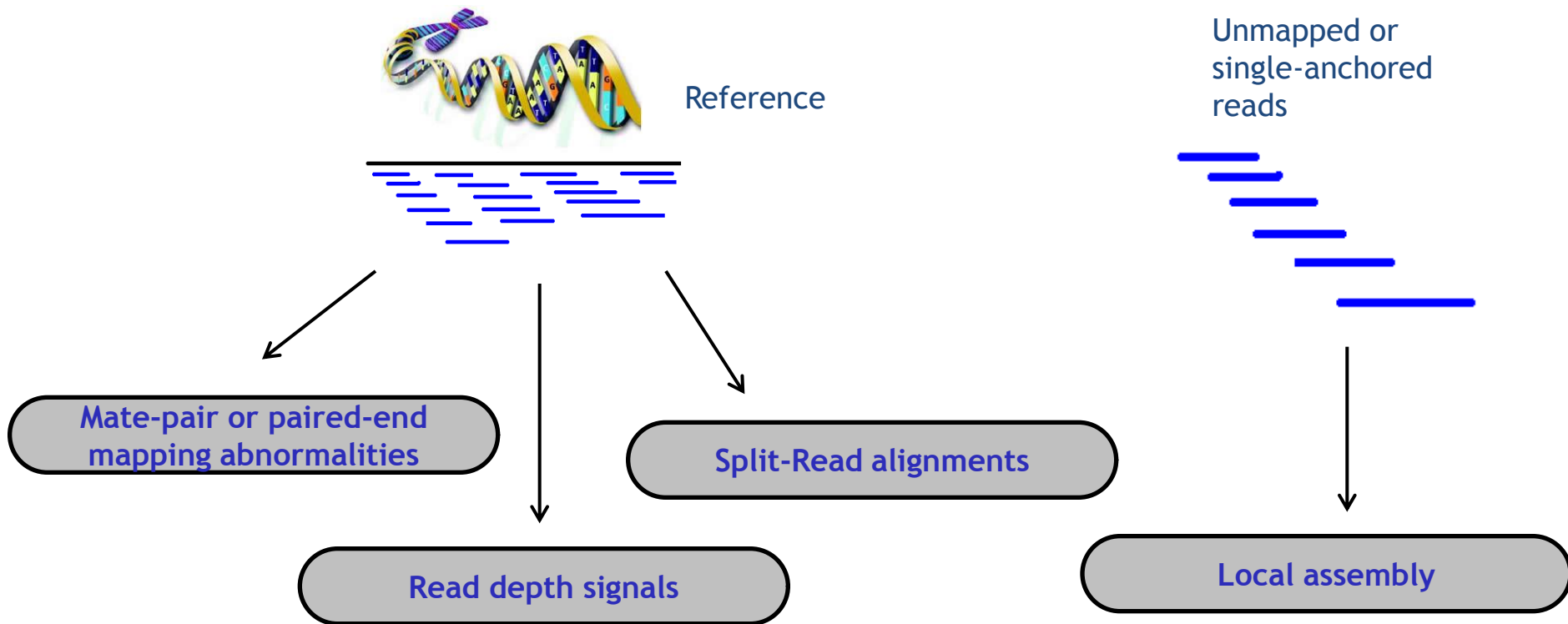
- Open-source tools
  - Dindel
  - Pindel
  - MoDIL
  - GATK
- Commercial packages
  - Maybe CLC Bio and others
- Indel calling has a higher false positive rate than SNV calling

# Genomic Rearrangements / Structural Variants

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



# Detecting Genomic Rearrangements



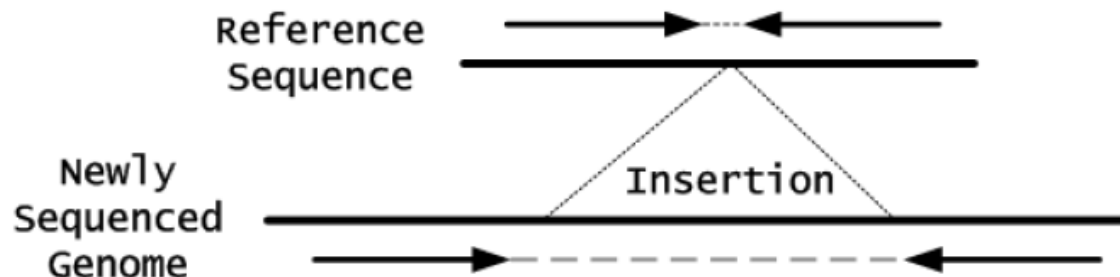
Mate-pair or paired-end mapping abnormalities

Read-depth signals

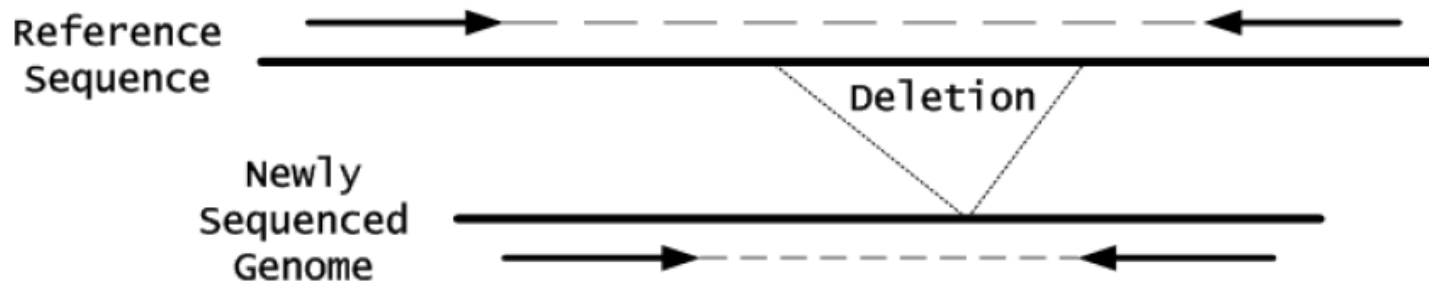
Split-read alignments

Local assembly

Insertion



Deletion



Mate-pair or paired-end mapping abnormalities

Read-depth signals

Split-read alignments

Local assembly

Insertion

Reference Sequence

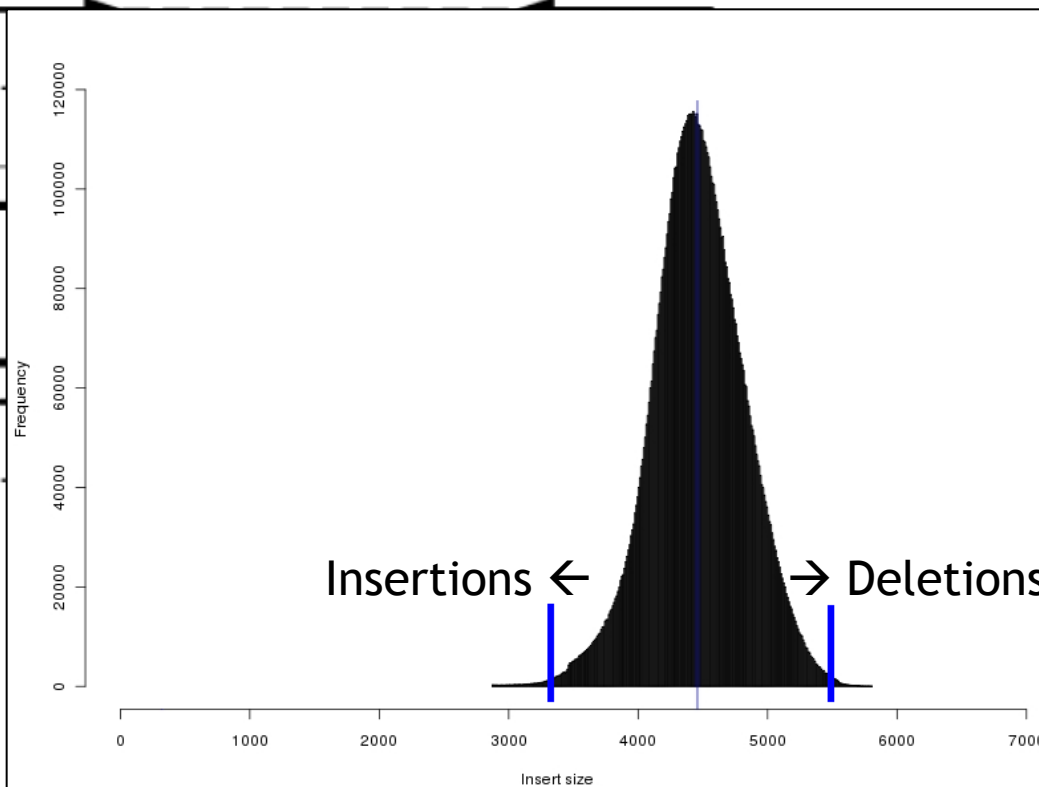
Newly Sequenced Genome

Insertion

Deletion

Reference Sequence

Newly Sequenced Genome



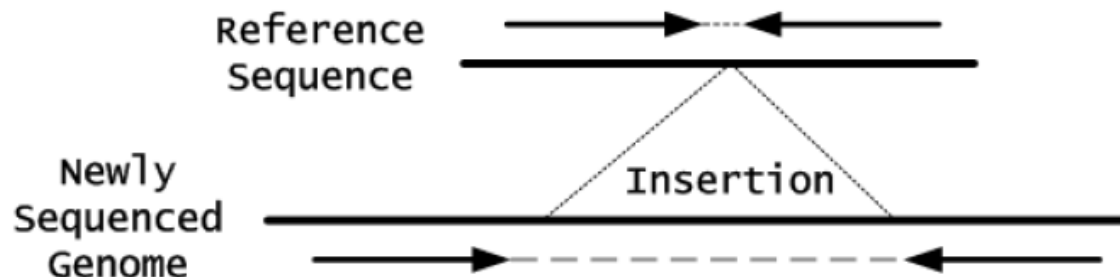
**Mate-pair or paired-end mapping abnormalities**

Read-depth signals

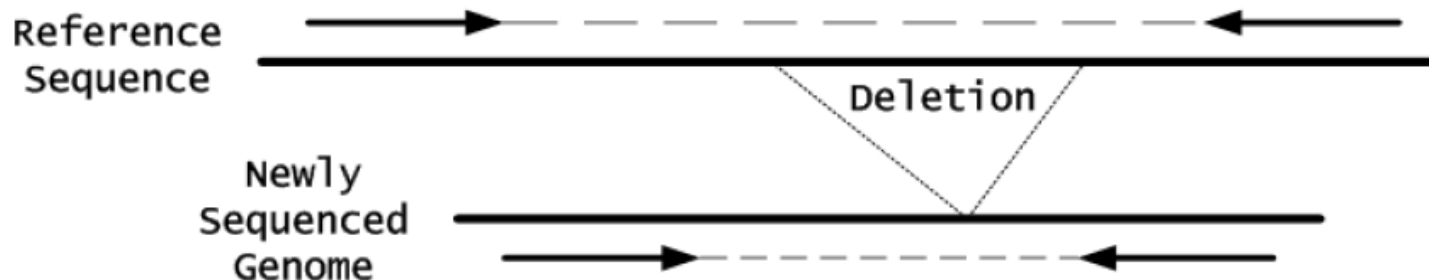
Split-read alignments

Local assembly

Insertion



Deletion



Inversion

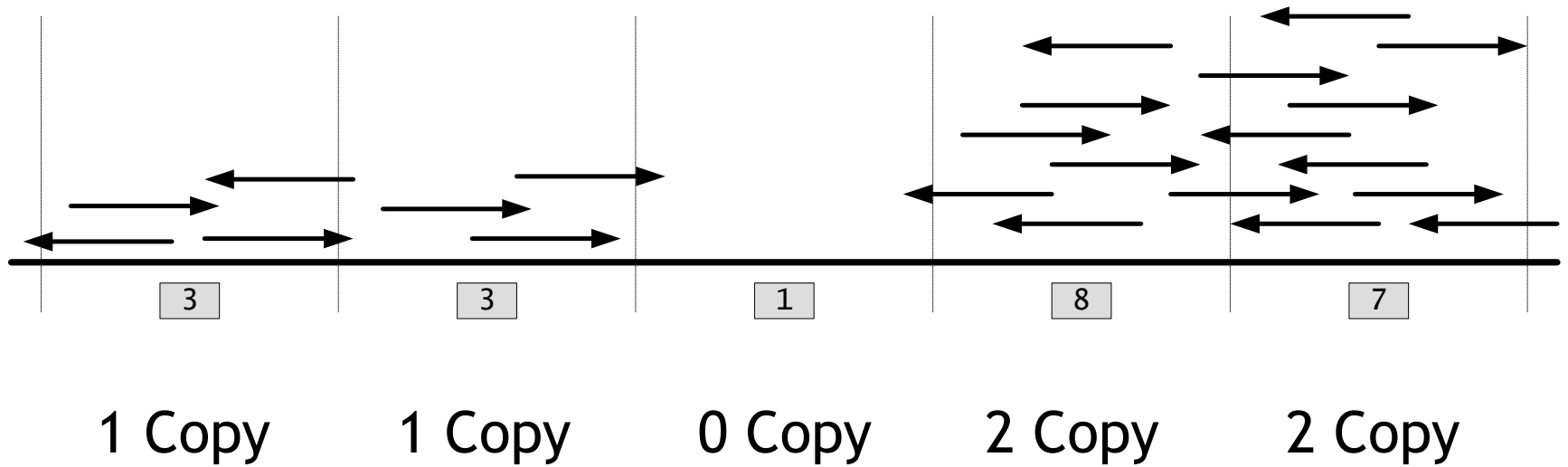


Mate-pair or paired-end mapping abnormalities

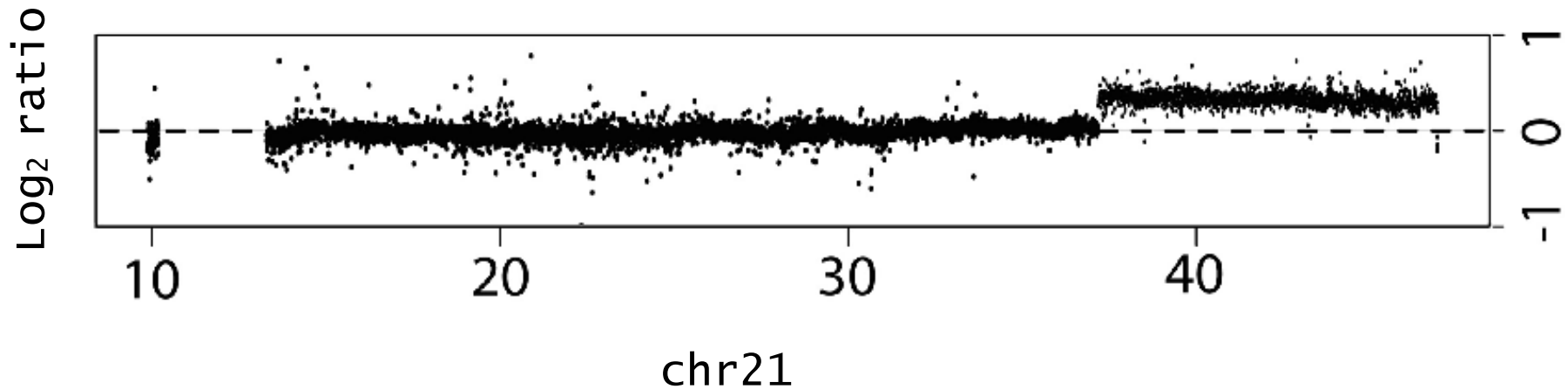
Read-depth signals

Split-read alignments

Local assembly



- Down-Syndrom
  - Partial Trisomie 21



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

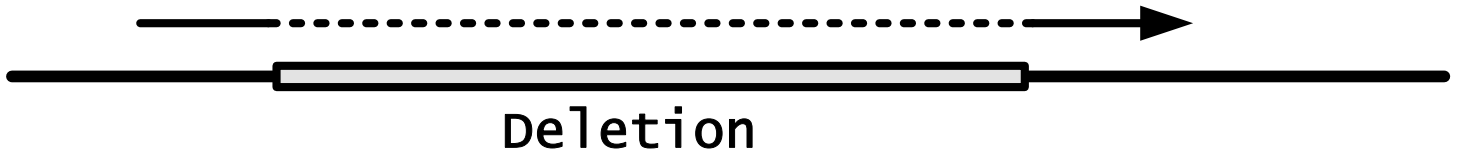
Mate-pair or paired-end mapping abnormalities

Read-depth signals

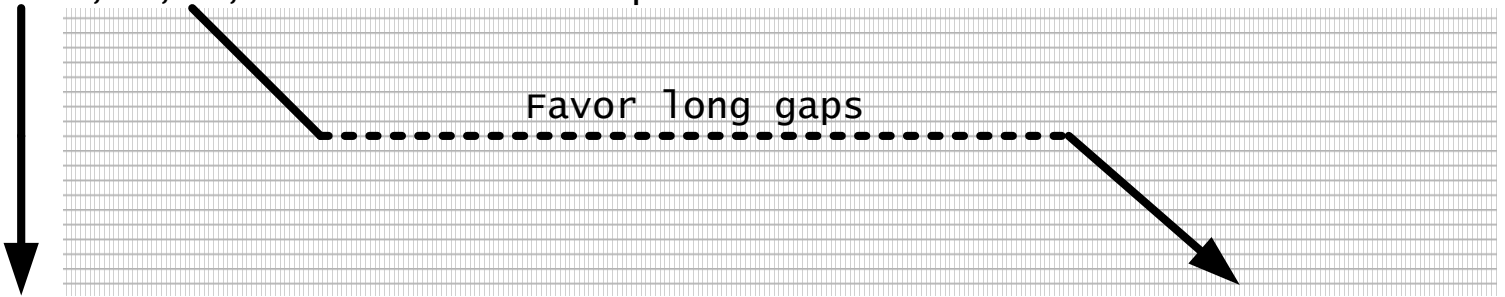
Split-read alignments

Local assembly

Reference Sequence



0, 0, 0, 0... Initialize top row with 0s

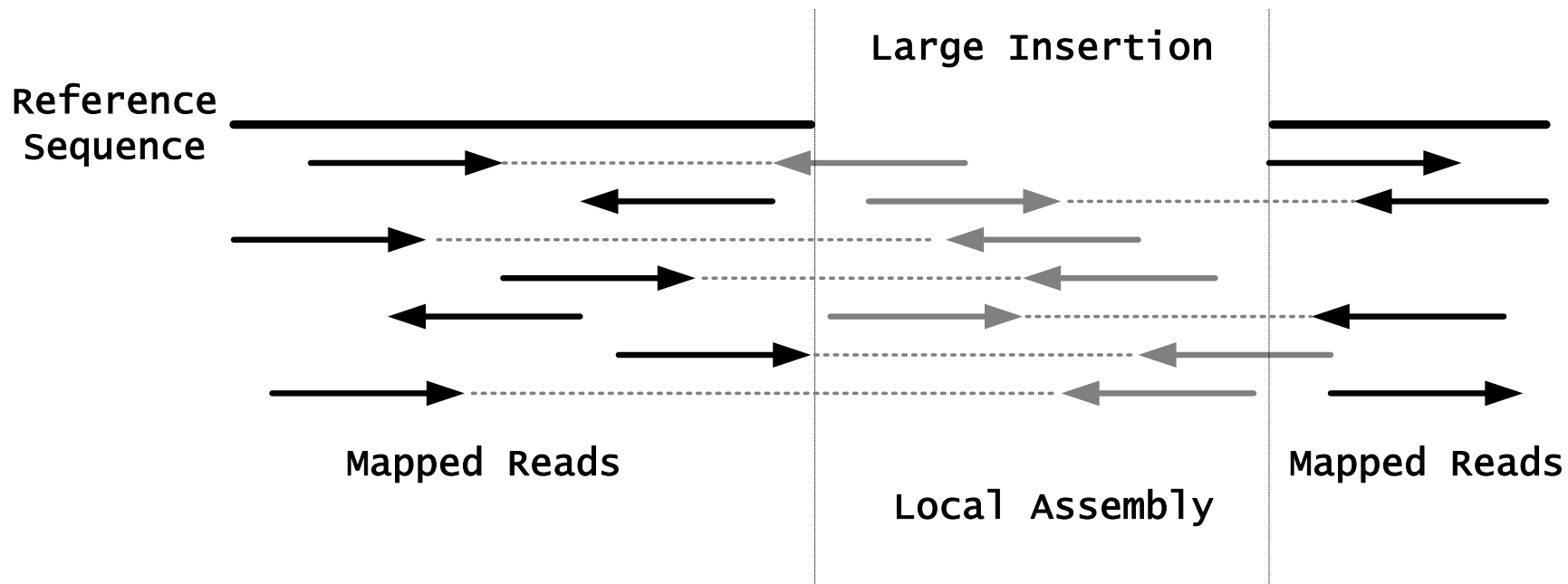


Mate-pair or paired-end mapping abnormalities

Read-depth signals

Split-read alignments

Local assembly

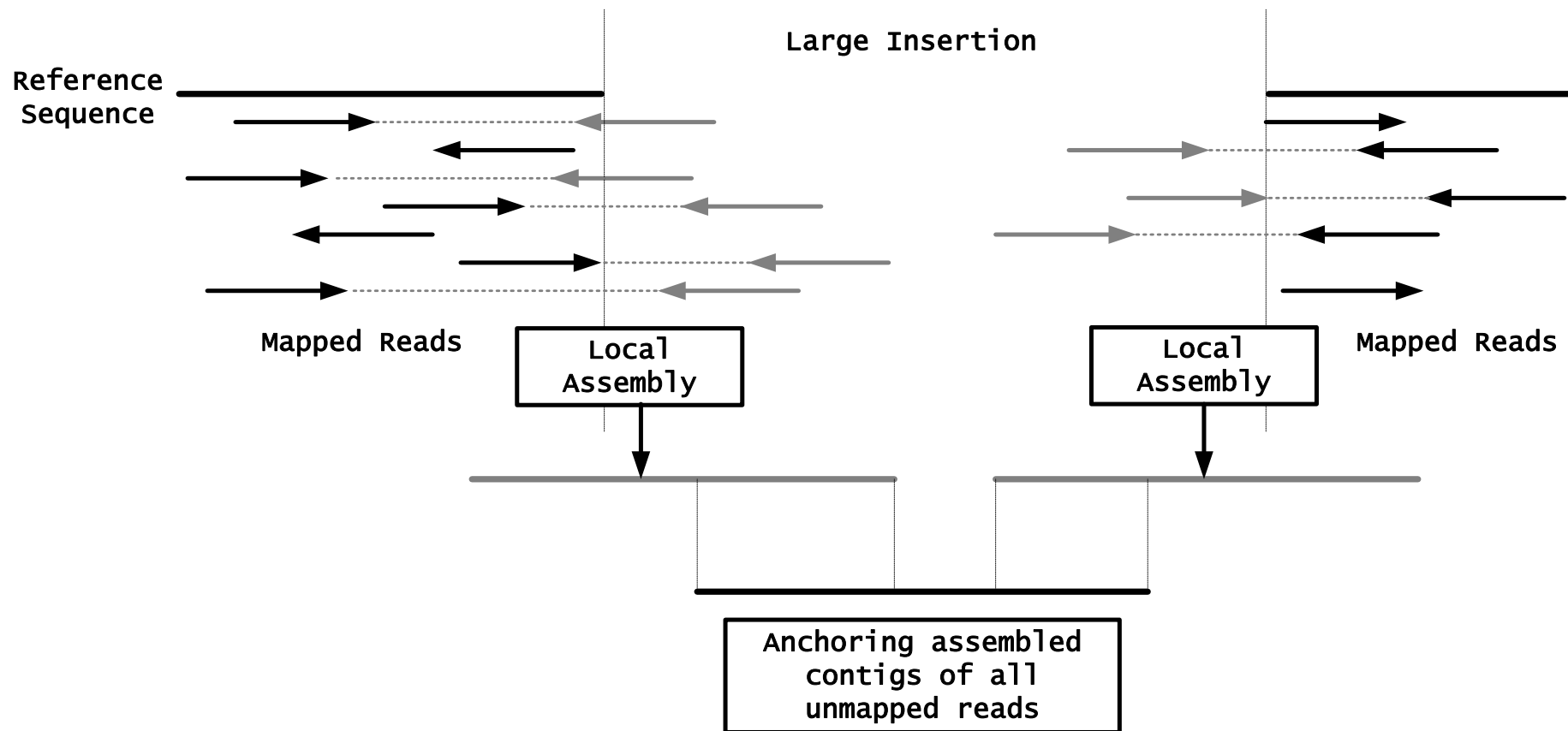


Mate-pair or paired-end mapping abnormalities

Read-depth signals

Split-read alignments

Local assembly



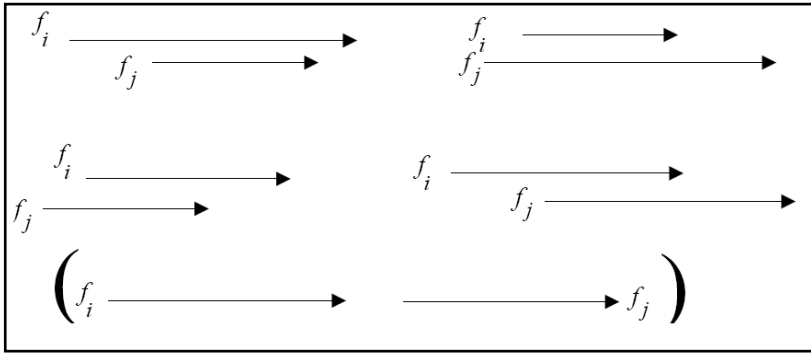
# Computational Methods for De Novo Genomic Rearrangement Detection

	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

# Assembly

- De novo assemblers
  - Classical assemblers
    - Overlap - Layout - Consensus assembler
  - Short-read assemblers
    - De Bruijn graph assembler
- Reference-based assembler
- Different types
  - Whole genome assembly
  - Transcriptome assembly
  - Metagenome assembly

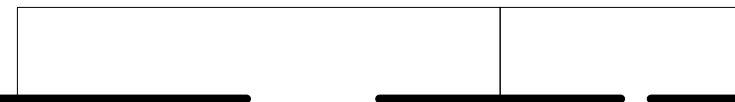
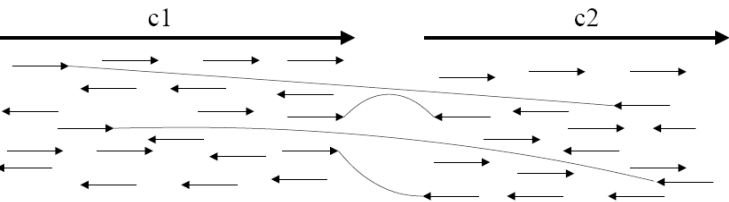
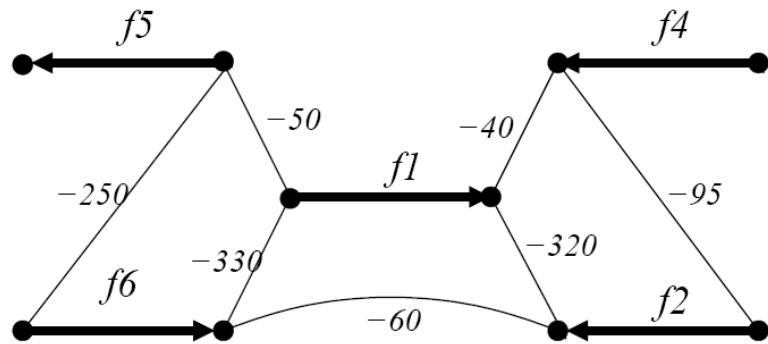
# Overlap - Layout - Consensus Assembler



Overlap



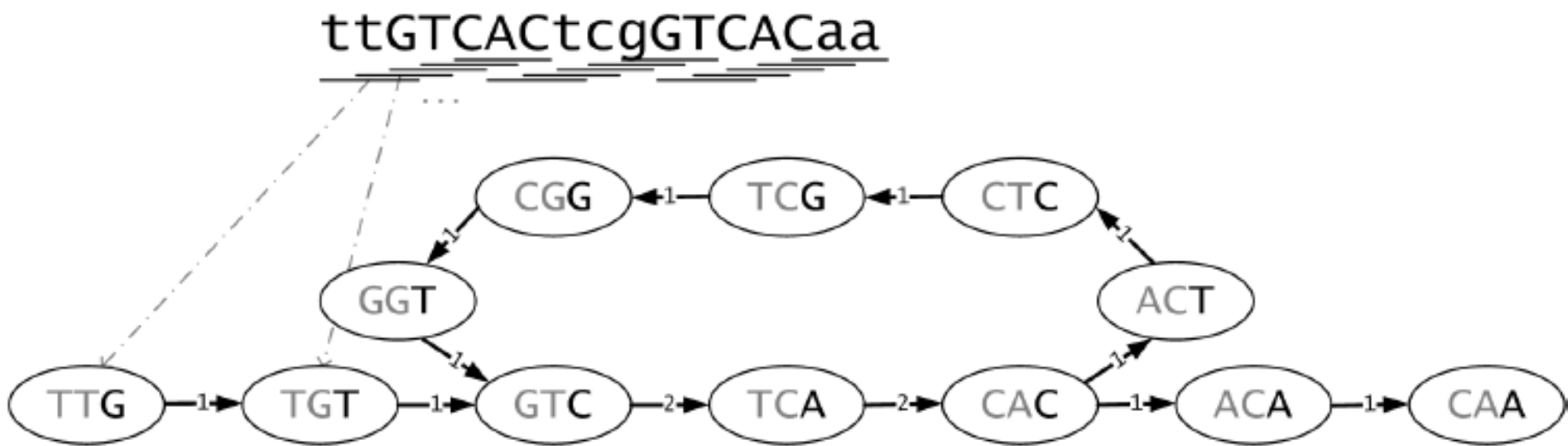
Layout



Consensus

# de Bruijn Graph

- Reads are too short to compute a reliable pairwise overlap
- K-mer dissection



# Thanks!

