Next Generation Sequencing data analysis

Andrey Prjibelski
Algorithmic Biology Lab
SPbAU RAS

From the very beginning

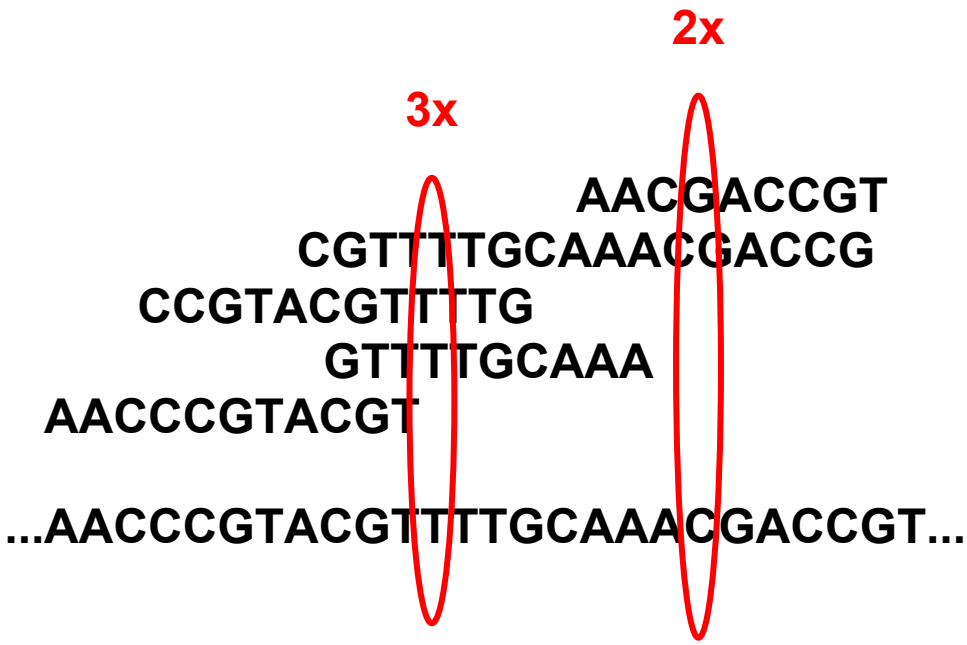
...AACCCGTACGTTTTGCAAACGACCGT...

From the very beginning

- Sequencing

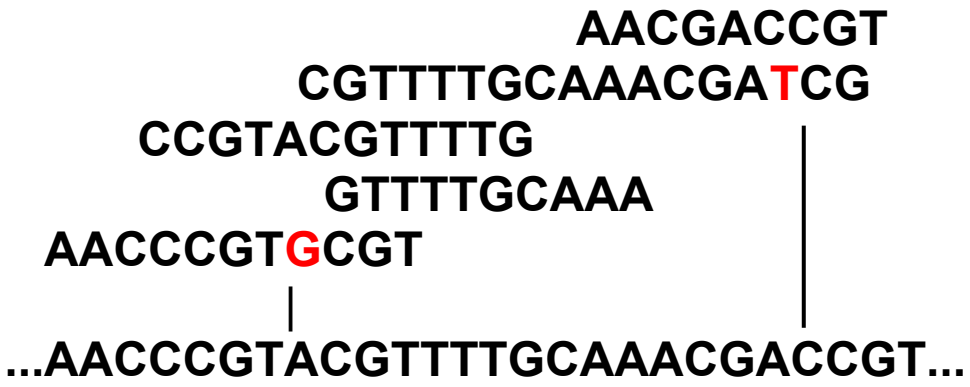
From the very beginning

- Sequencing
- Coverage



From the very beginning

- Sequencing
- Coverage
- Errors
  - Mismatches



From the very beginning

- Sequencing
- Coverage
- Errors
  - Mismatches
  - Indels

Early days

- Sanger sequencing
  - Long reads (~900 bp)
  - Low coverage (< 10x)
  - Extreme cost

- Human genome project
  - 3 Mbp
  - 3 billion USD
  - 10 years

NGS

- Shorter reads (25-500bp)
- High coverage (50-1000x)
- Huge amount of data
- Low cost
- **Required completely new algorithms**

NGS technologies

	illumina	Roche 454 SEQUENCING	ion torrent by life technologies	PACIFIC BIOSCIENCES
Read length, bp	25-250	400-1100	200-400	1000-5000
Error rate	0.01-1%	1%	1-2%	10-13%
Error type	Mismatches	Indels & Mismatches	Indels & Mismatches	Indels & Mismatches
Comments	Error rate grows to the end of the read	Problems with homopolymers	Problems with homopolymers	
Cost per 1 mbp, $	0.05 - 0.5	30	0.5 - 5	2

FASTA/FASTQ

- FASTA

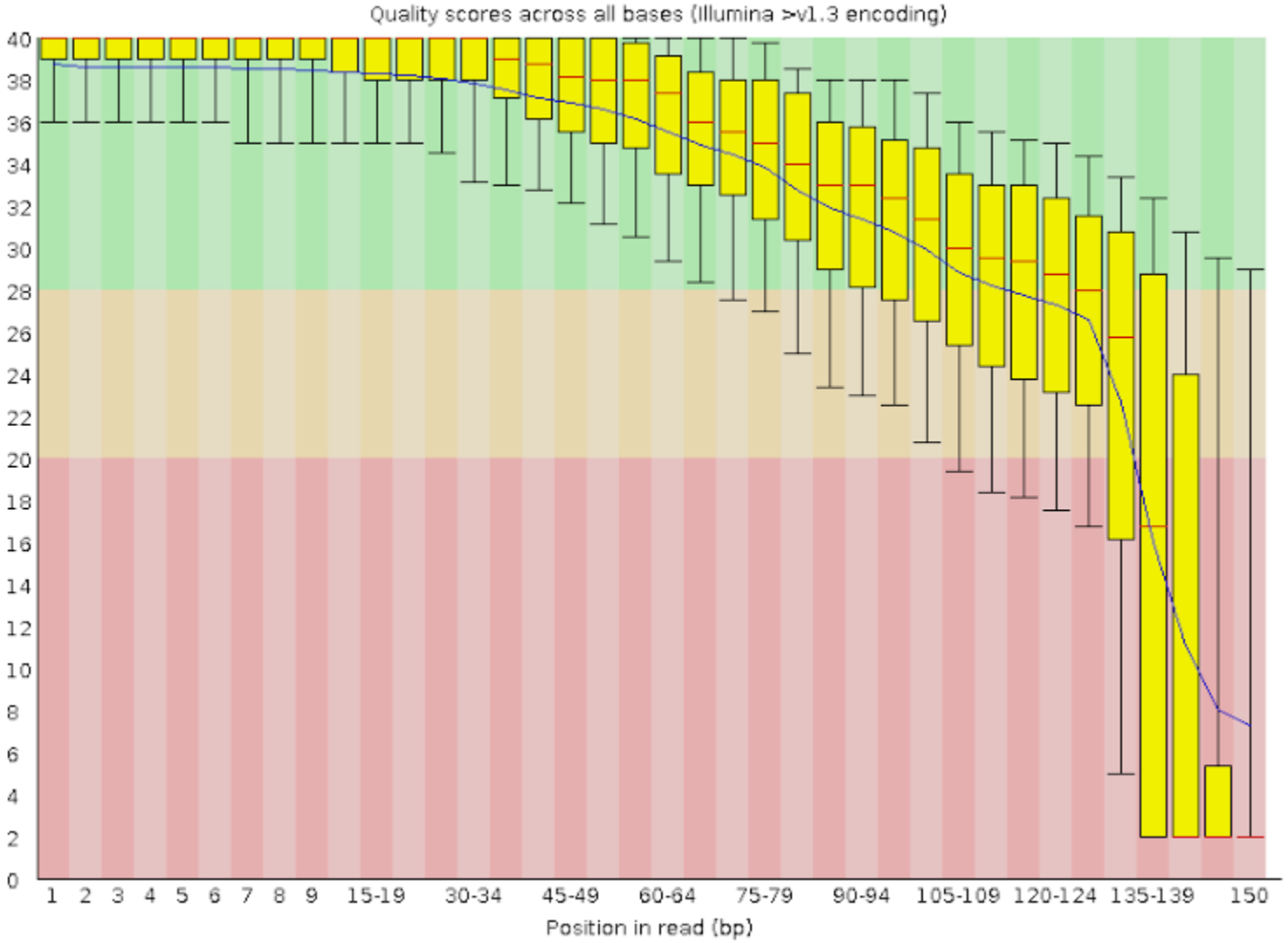
```
>EAS20_8_6_1_9_1972/1
ACCACCATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGCG
T
>EAS20_8_6_1_163_1521/1
GCAGAAAACGTTCTGCATTTGCCACTGATGTACCGCCGAACTTCAACACTCGCAT
G
```

- FASTQ

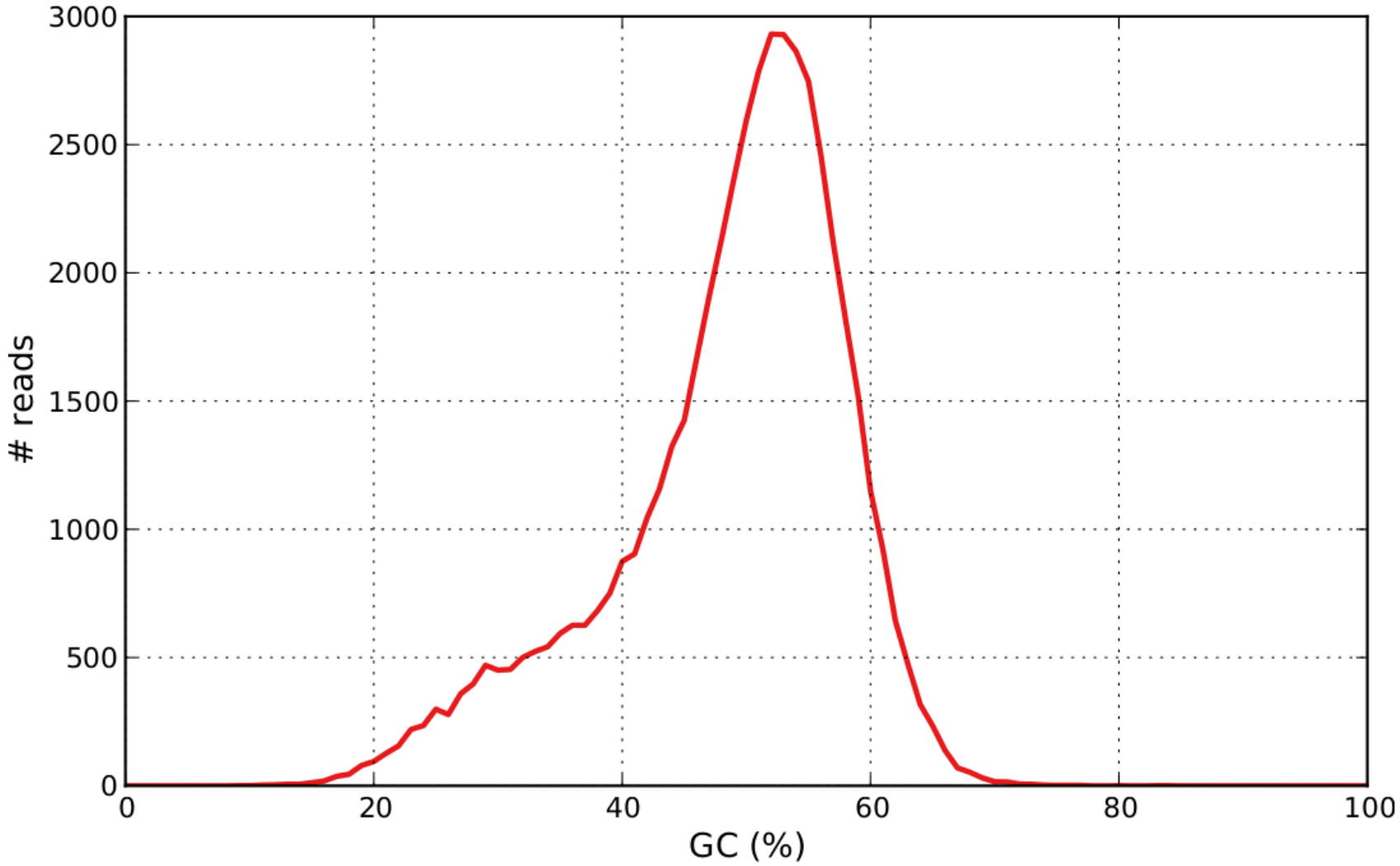
```
@EAS20_8_6_1_1477_92/1
ACCGTTACCTGTGGTAATGGTGATGGTGGTGGTAATGGTGGTGCTAATGCGTTTC
A
+EAS20_8_6_1_1477_92/1
HHGHFHHHHHHHHHGFFHHHBG?GGC8DD9GF??
=FFBCGBAF>FGCFHGHGGGD5
```

- Phred quality

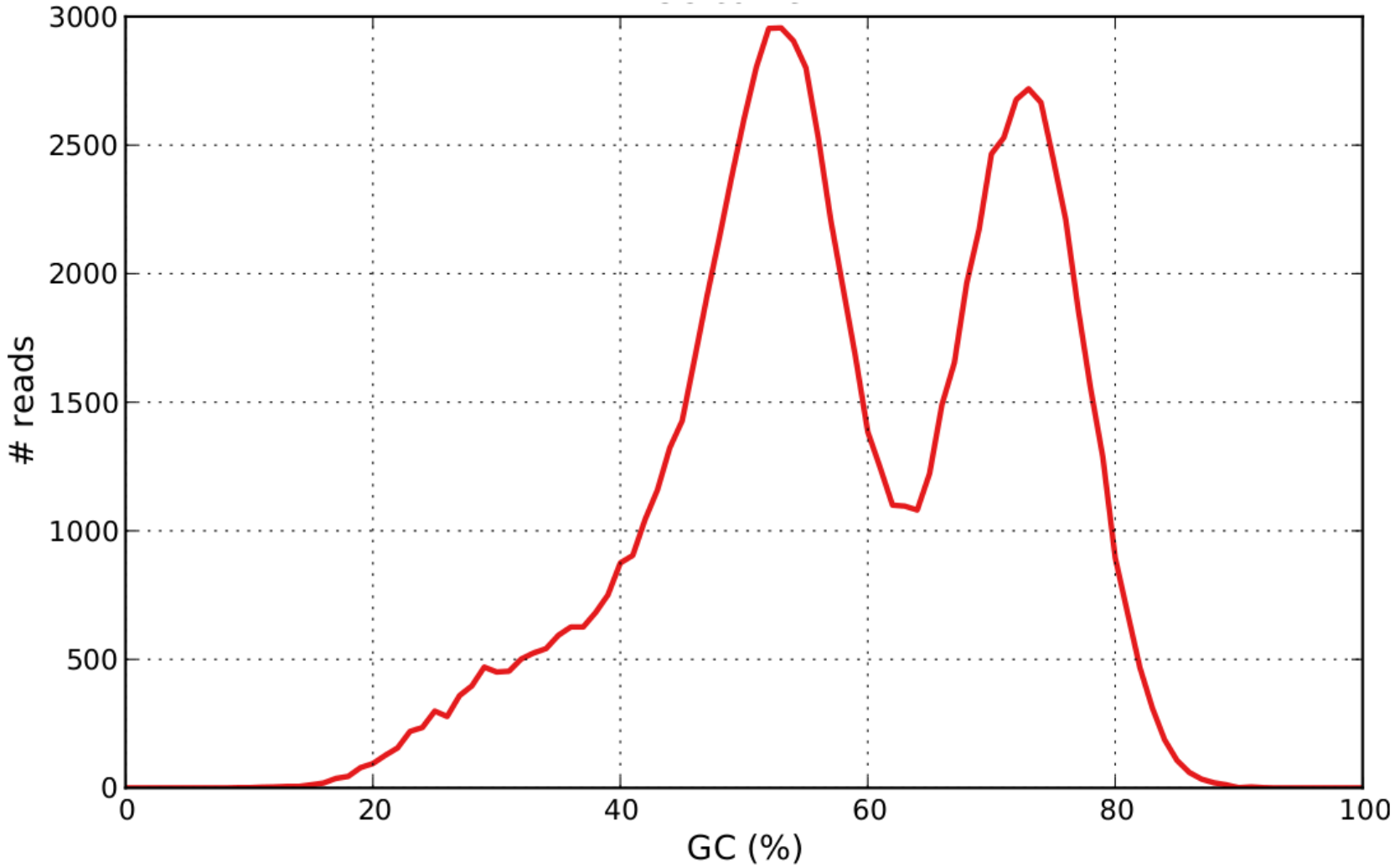
Illumina error rate



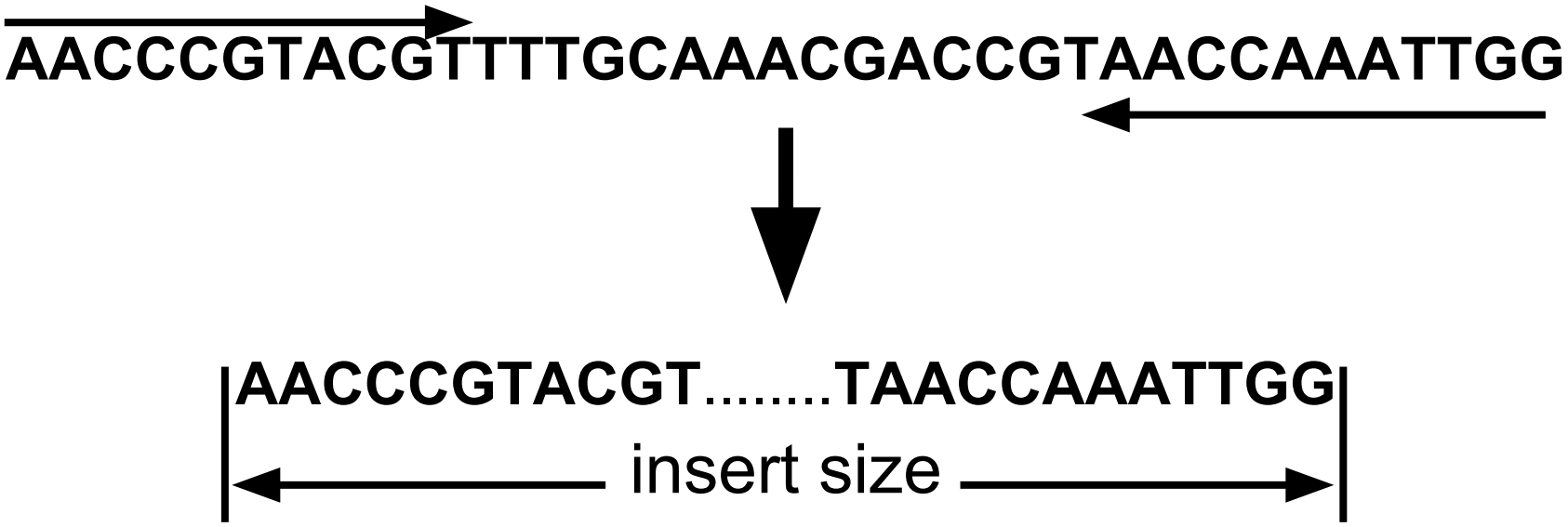
GC content



GC content



Paired reads



- Paired-end (< 1 kbp)
- Mate-pairs (1 - 20 kbp)

Short Read Archive

- http://www.ncbi.nlm.nih.gov/sra/
- SRA toolkit

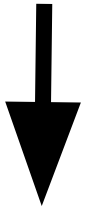
Alignment

AACGCTAACGGTAA
AACCGCGAACTAA

Alignment

AACGCTAACGGTAA

AACCGCGAACTAA



AAC - GCTAACGGTAA

AACCGCGAAC - - TAA

Short read alignment

- Challenges?

Short read alignment

- Challenges
  - Small length
  - Gigabytes of data
  - Sequencing errors
  - Genomic repeats

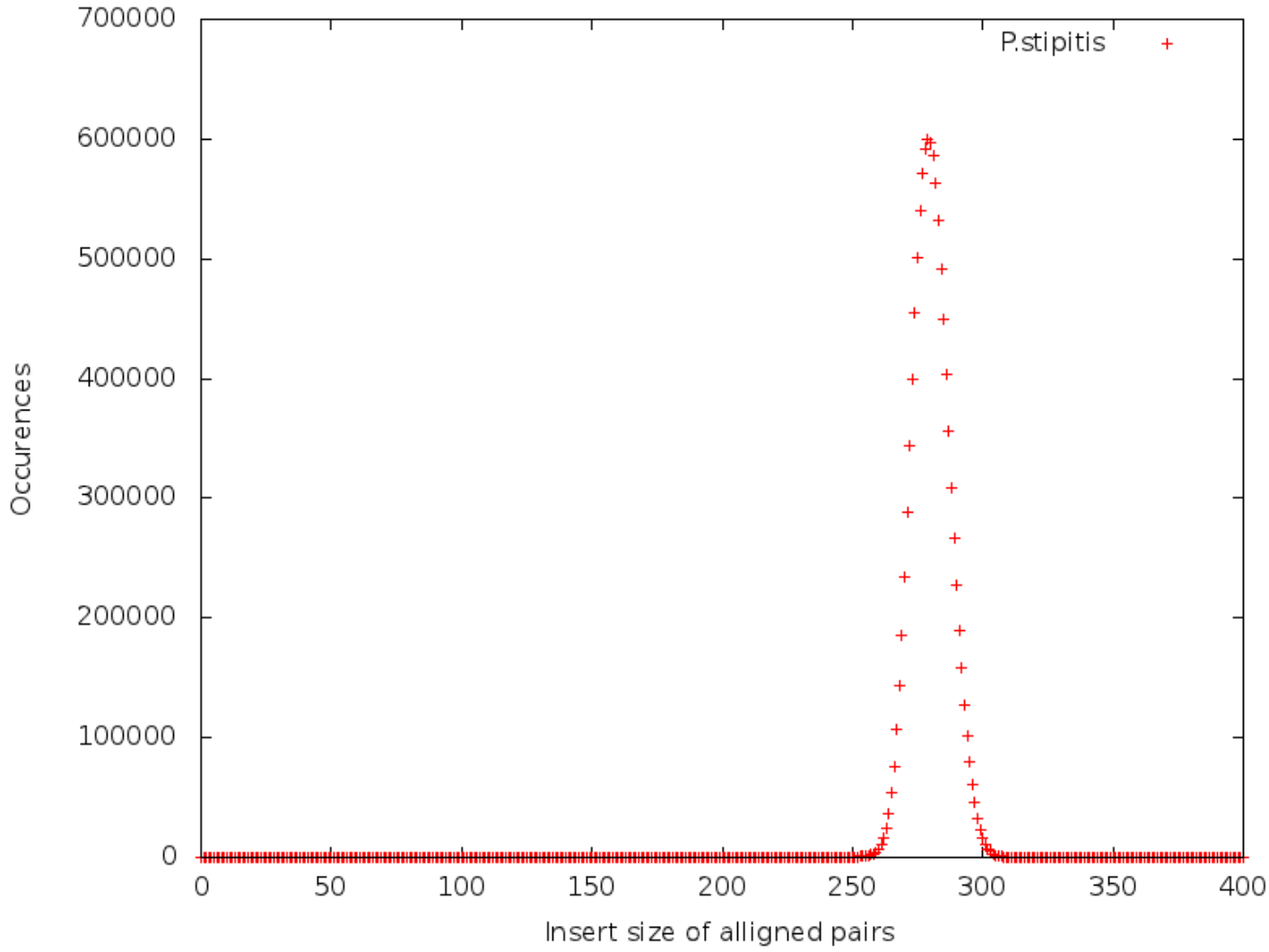
- Tools
  - Bowtie, BWA (Illumina)
  - BWA-SW (454, IonTorrent)
  - and many more

SAM/BAM

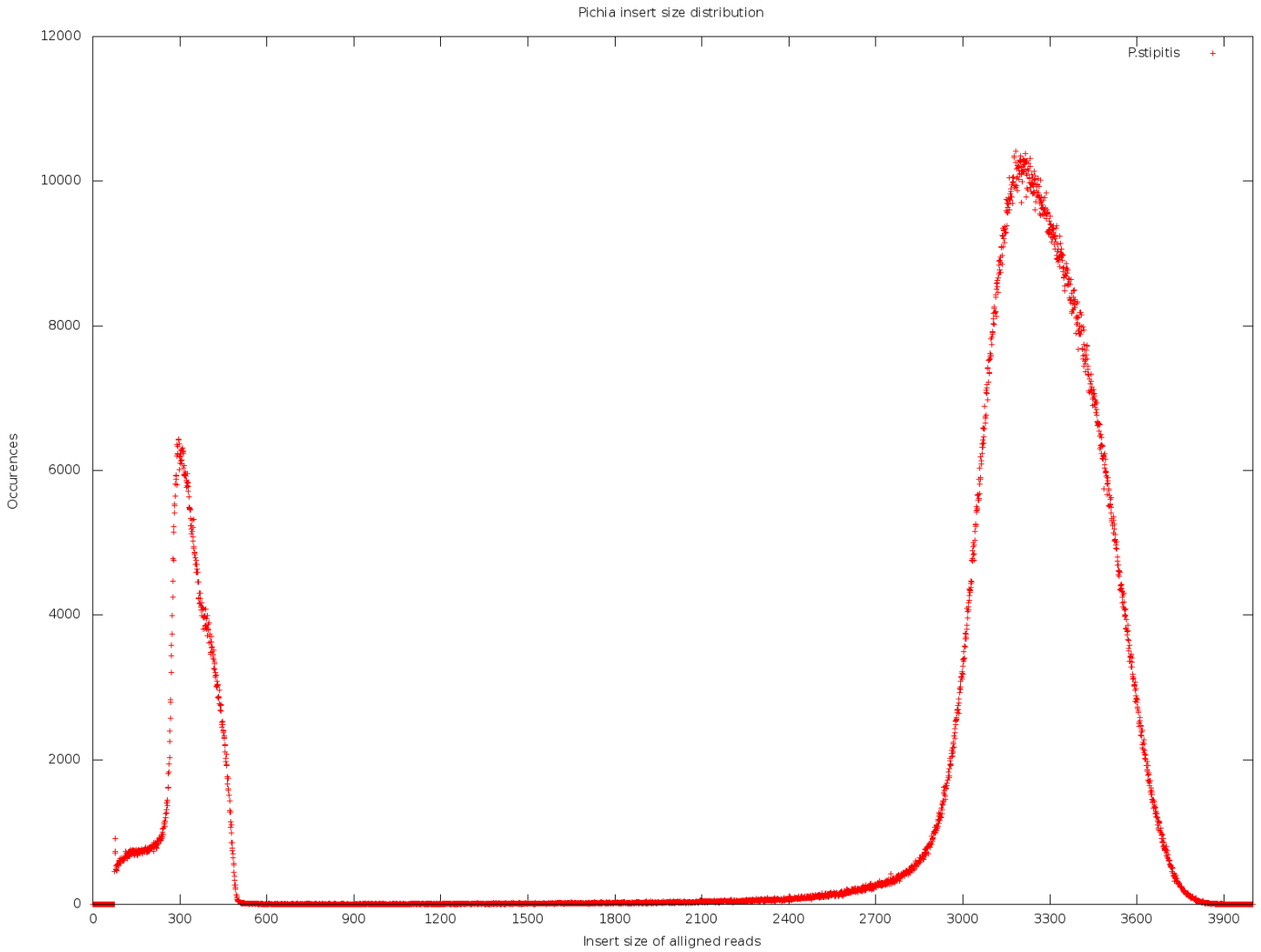
SAM/BAM

- Read ID (QNAME)
- Reference ID (RNAME)
- Mapping position (POS)
- Mate reference ID (RNEXT)
- Mate position (PNEXT)
- Observed insert length (TLEN)
- Read sequence (SEQ)
- Read quality (QUAL)
- CIGAR string
  - 34M 1I 4M 2D 1X 3M

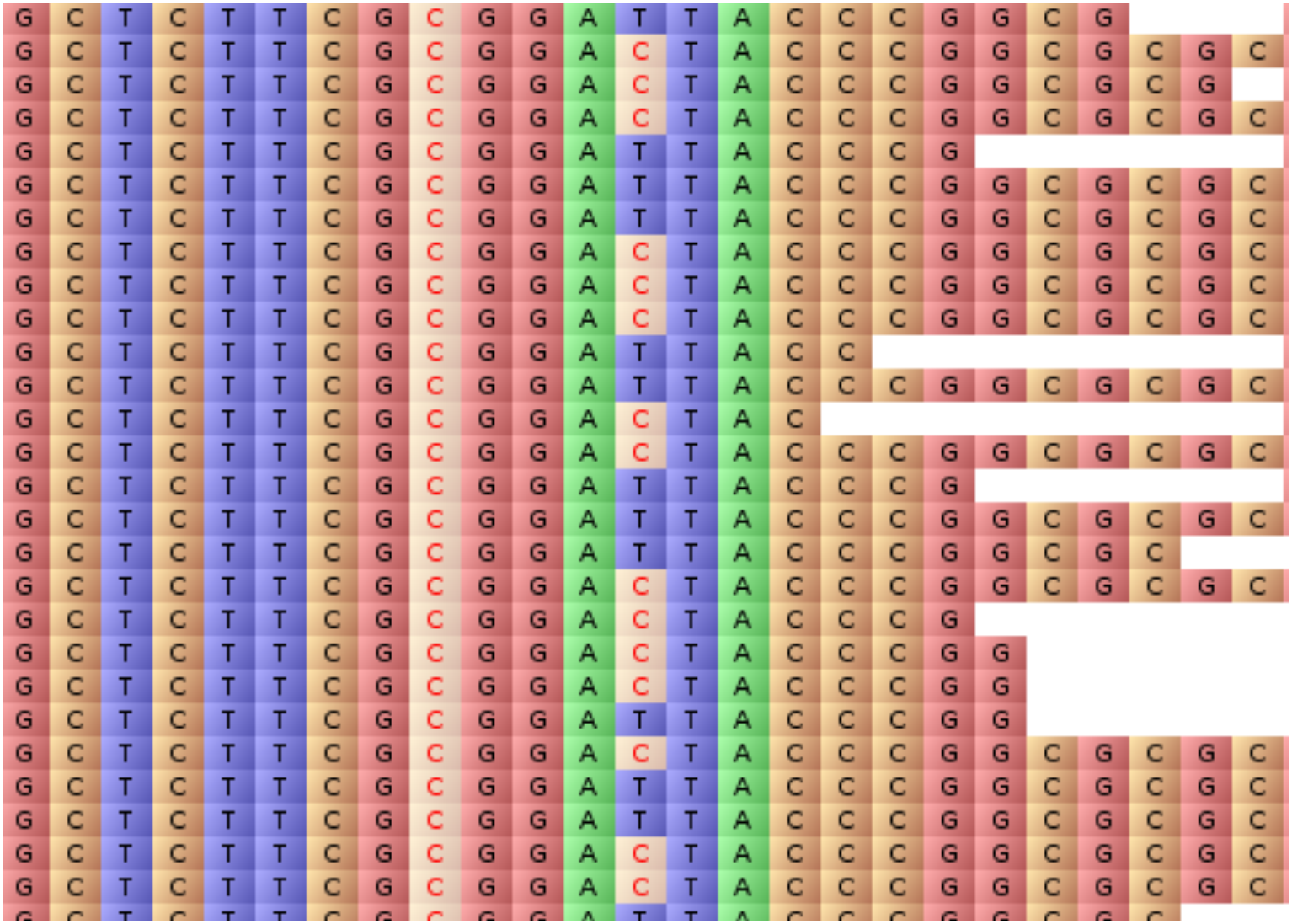
Insert size distribution



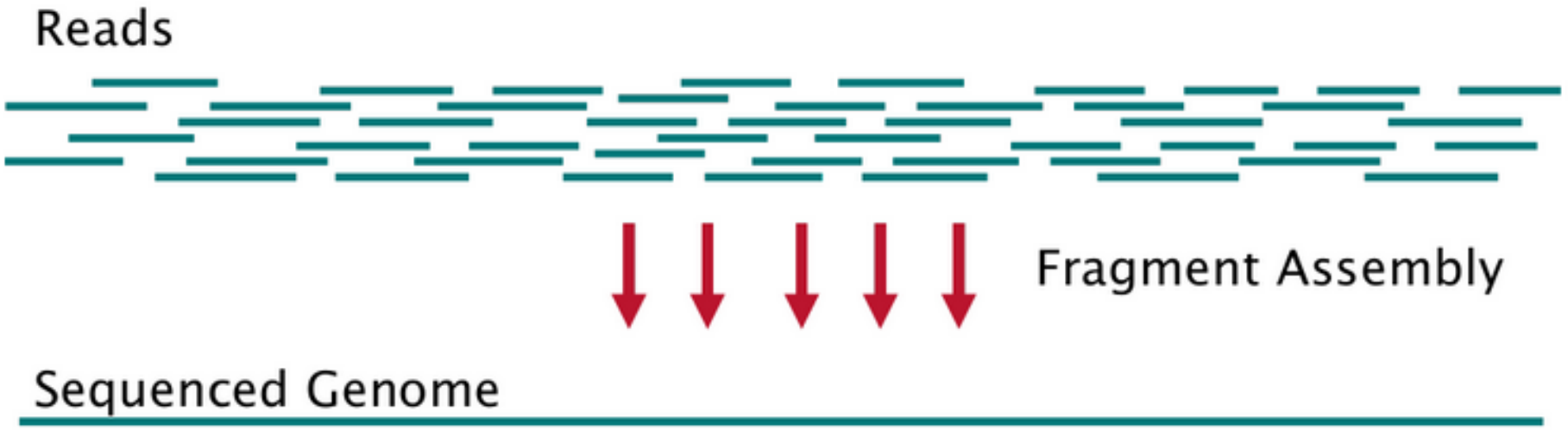
Insert size distribution



SNPs


GCTCTTCGCGGATTACCCGGGCG
GCTCTTCGCGGACTACCCGGGCGCGC
GCTCTTCGCGGACTACCCGGCGCG
GCTCTTCGCGGACTACCCGGGCGCC
GCTCTTCGCGGATTACCCG
GCTCTTCGCGGATTACCCGGCGCGCC
GCTCTTCGCGGATTACCCGGCGCGCC
GCTCTTCGCGGACTACCCGGCGCGCC
GCTCTTCGCGGACTACCCGGCGCGCC
GCTCTTCGCGGACTACCCGGCGCGCC
GCTCTTCGCGGATTACC
GCTCTTCGCGGATTACCCGGCGCGCC
GCTCTTCGCGGACTAC
GCTCTTCGCGGACTACCCGGCGCGCC
GCTCTTCGCGGATTACCCG
GCTCTTCGCGGATTACCCGGCGCGC
GCTCTTCGCGGATTACCCGGCGC
GCTCTTCGCGGACTACCCG
GCTCTTCGCGGACTACCCGG
GCTCTTCGCGGACTACCCGG
GCTCTTCGCGGATTACCCG
GCTCTTCGCGGACTACCCGGCGCGC
GCTCTTCGCGGATTACCCGGCGCGC
GCTCTTCGCGGACTACCCGGCGCGC
GCTCTTCGCGGACTACCCGGCGCGC
GCTCTTCGCGGACTACCCGGCGCGC

De novo whole genome assembly



Assembly pipeline

- Sequencing
- Artifacts & contaminants cleaning
- Draft assembly
  - Error correction
  - Assembly
  - Repeat resolution
  - Scaffolding
- Postprocessing
- Finishing
- Annotation

Why to assemble?

- NGS
  - Billions of short reads
  - Sequencing errors
  - Contaminants

Hard to perform analysis

- Assembly
  - Corrects sequencing errors
  - Much longer sequences
  - Each genomic region is presented only once
  - *May introduce errors*

Which assembler to use?

- ABySS
- ALLPATHS-LG
- CLC
- EULER
- IDBA-UD
- MaSuRCA
- Ray
- SOAPdenovo
- SPAdes
- Velvet
- and many more...

Which assembler to use?

- Assemblathon 1 & 2
  - Simulated and real datasets
  - More than 30 teams competing
- GAGE, GAGE-B papers
- Genome assembly evaluation tools
  - QUAST
  - GAGE

Genome Assembly Gold-Standard Evaluations

There is no best assembler.

Which assembler to use?

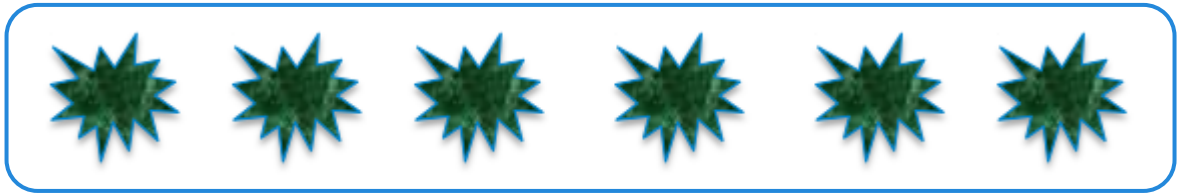
- Different technologies (Illumina, 454, IonTorrent, ...)
- Genome type and size (bacteria, insects, mammals, plants, ...)
- Type of prepared libraries (single reads, paired-end, mate-pairs, combinations)
- Type of data (multicell, metagenomic, single-cell)

Evaluating assemblies

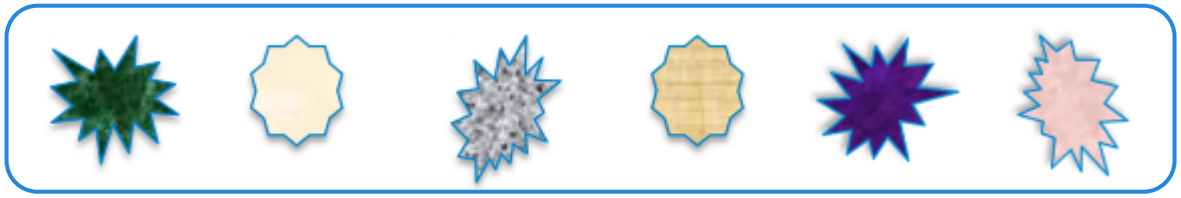
- BLAST
- Assembly statistics
  - Basic statistics (N50, Nx plots etc)
  - Genome fraction
  - Misassemblies
  - Mismatch/indel rates

Why to create new assembler?

- Conventional sequencing

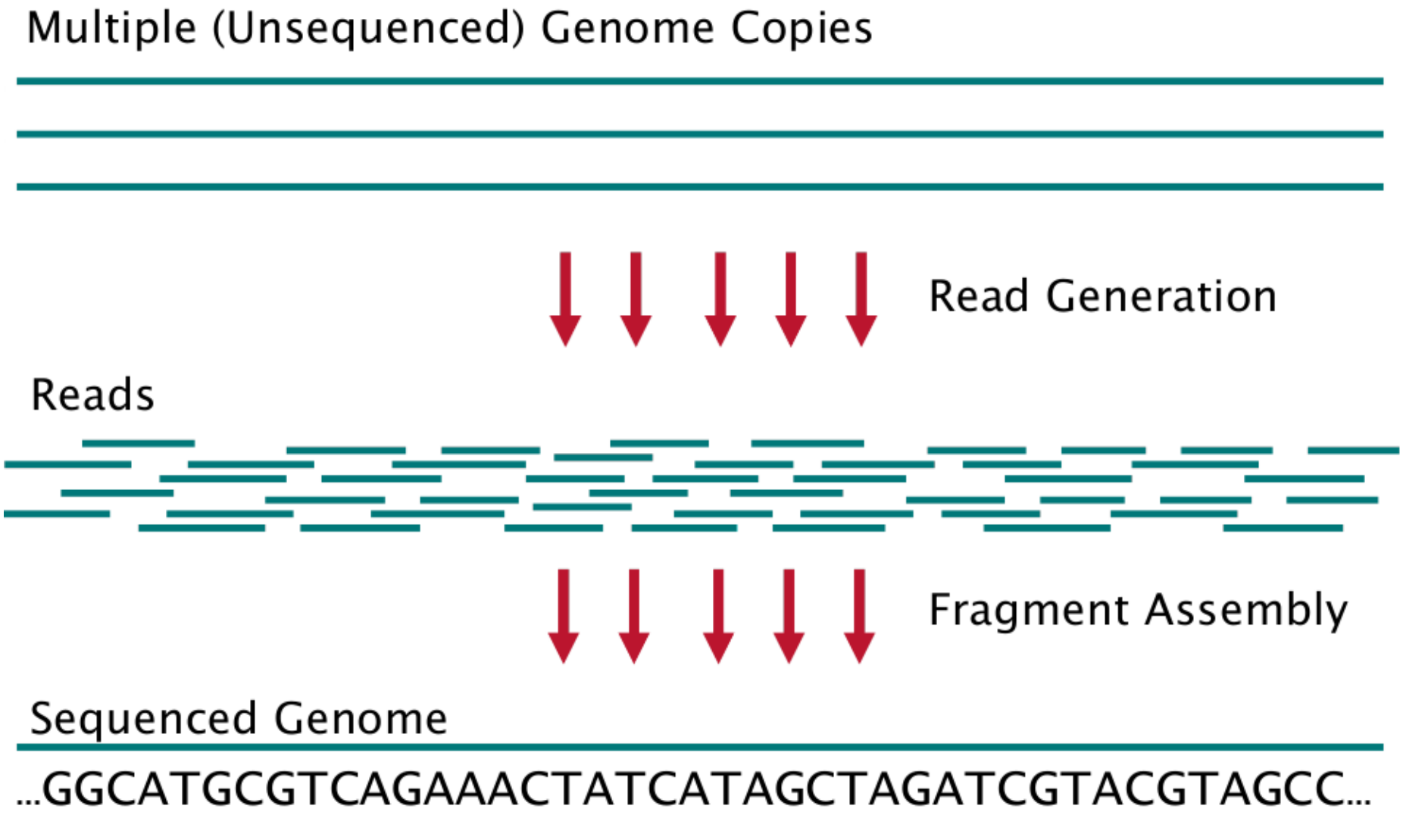


- Metagenomics



- Single cell

Conventional bacterial sequencing



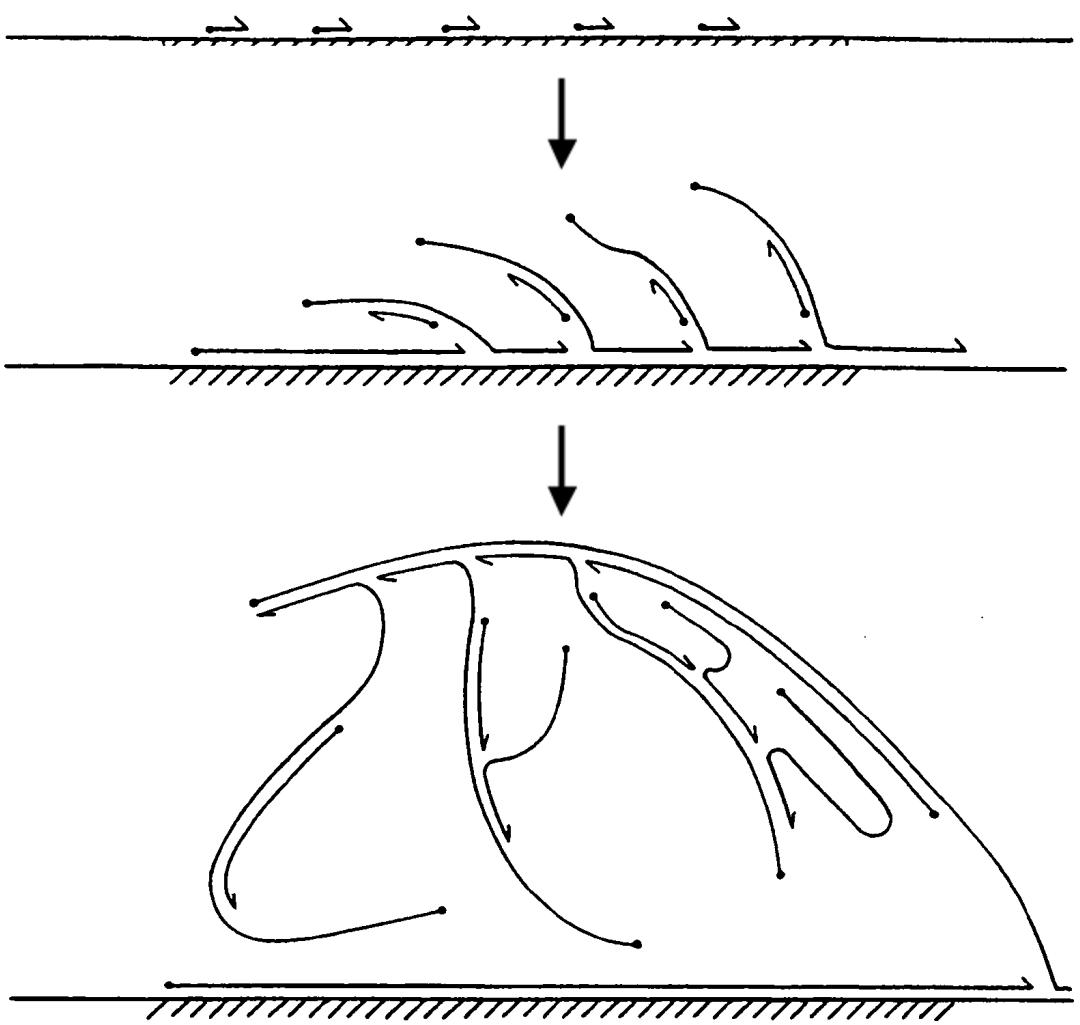
Metagenomics

- >99% bacteria cannot be cultured
- Metagenomics: sequencing of whole bacterial community
  - Reads from dozens of different genomes mixed in one data set
  - Different coverage for different bacteria
  - Presence of different strains
  - Conservative genomic regions
- Hard to assemble and classify resulting sequences
  - Usually allows to identify only a few genes

Single-cell sequencing via MDA

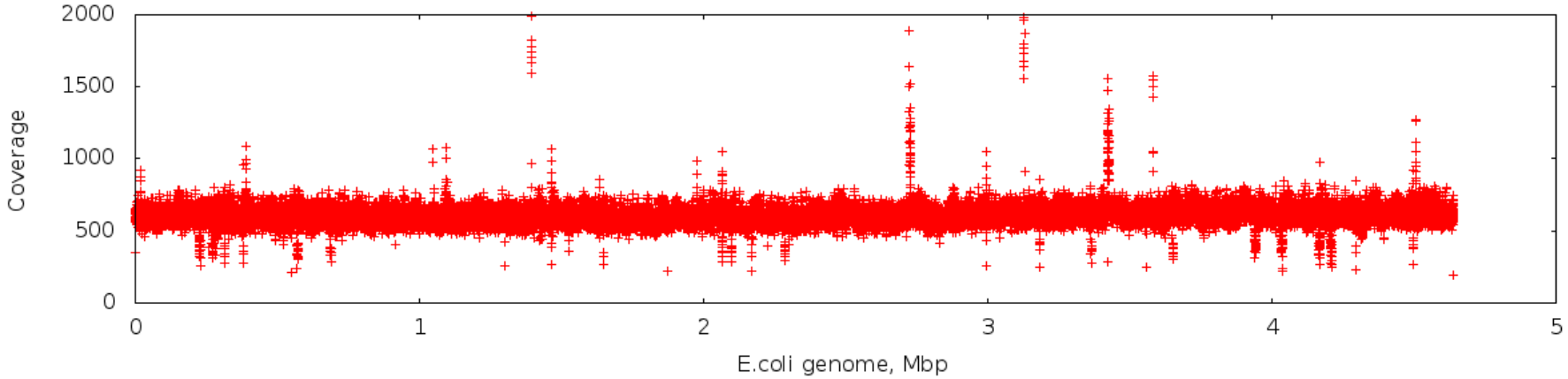
Multiple Displacement Amplification

- Random hexamer primers
- Phi29 DNA polymerase strand displacement

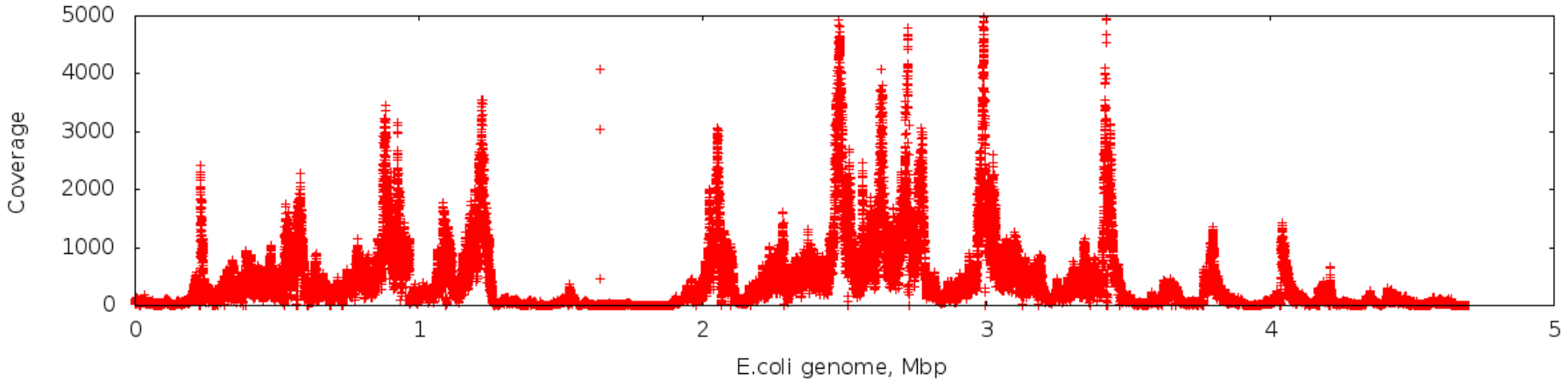


Challenges in single-cell assembly

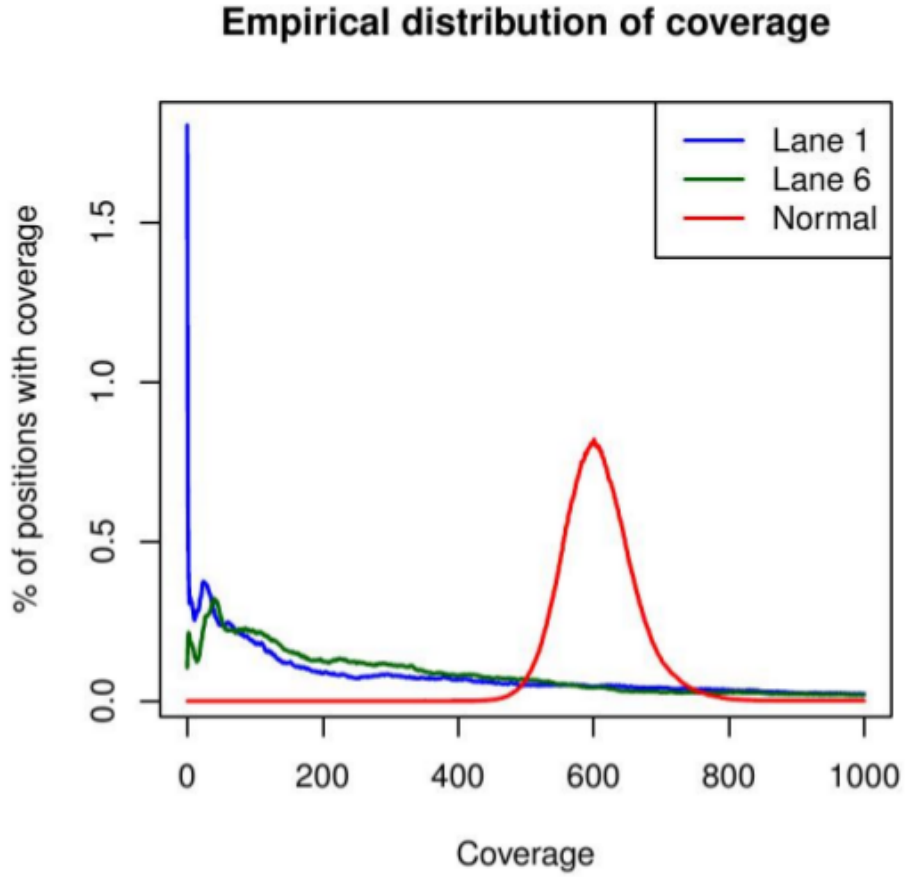
- *E. coli* isolate dataset



- *E.coli* single-cell dataset



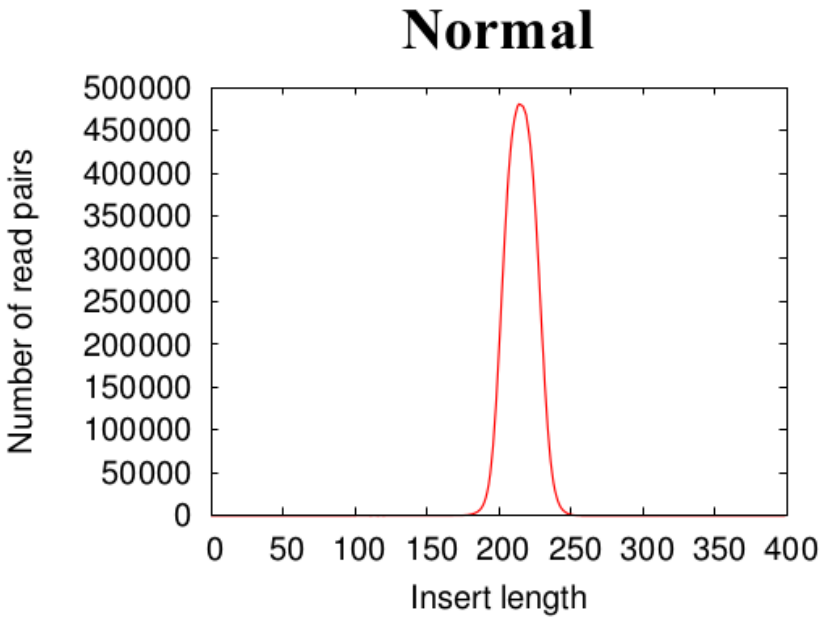
Challenges in single-cell assembly

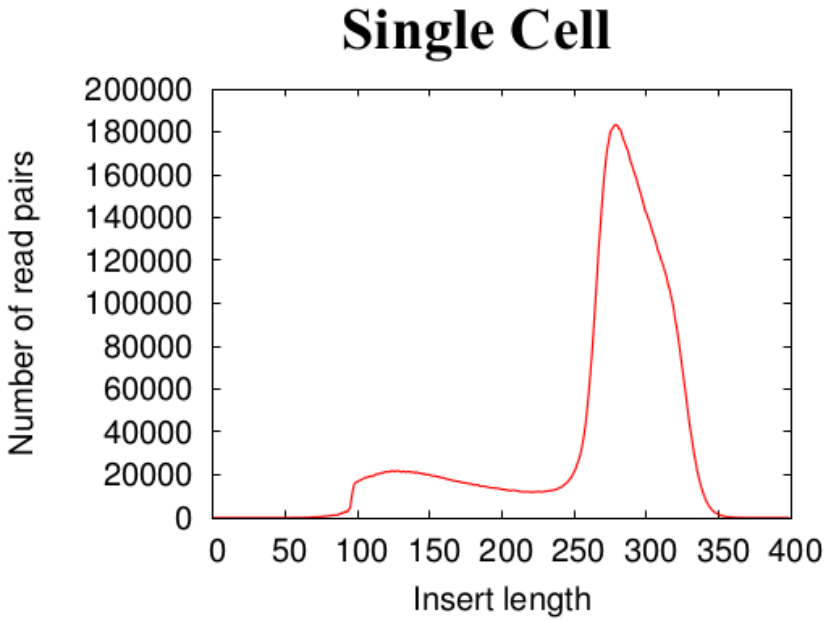


A cutoff threshold will eliminate about 25% of valid data in the single cell case, whereas it eliminates noise in the normal multicell case.

Challenges in single-cell assembly

- Insert size deviation





- Chimeric reads
  - Isolate dataset 0.01%
  - Single-cell dataset ~2%

Thank you!

Questions?