

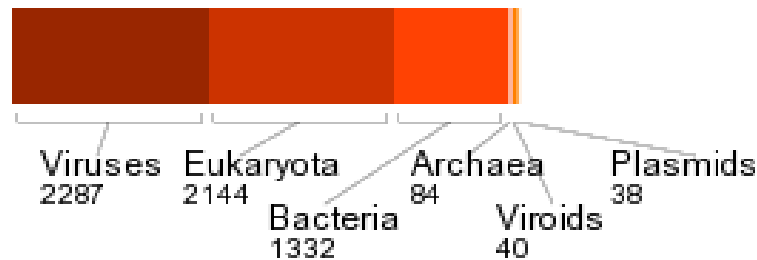
De Novo Assembly

Tobias Rausch
Genomics Core Facility
Korbel Group

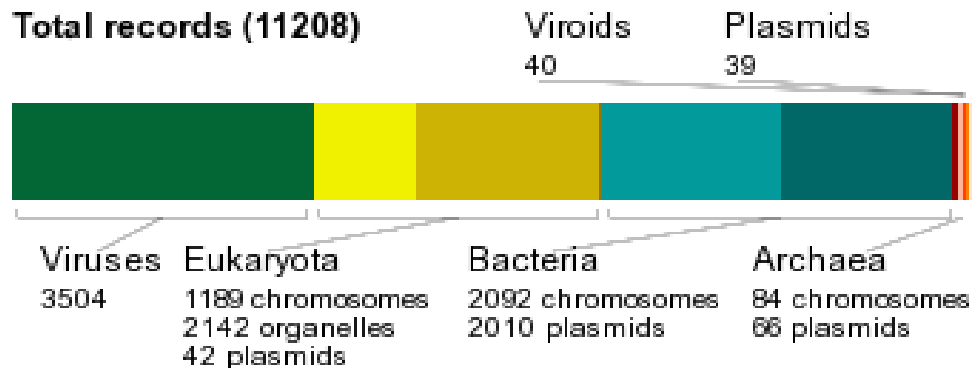
- Classical assemblers
 - Overlap - Layout - Consensus assembler
- Short-read assemblers
 - De Bruijn graph assembler
- Transcriptome assembly

- 1994: Bacteria *H.influenzae*, 1.8Mb genome
- 2000: *Drosophila*, 130Mb genome
- 2001: Human genome, 3Gb genome
- Currently about 6000 sequenced genomes

Total species (5925)



Total records (11208)



- Current sequencing technologies can only determine short consecutive pieces of DNA
 - 50 bp - 1200 bp, depending on the technology
- To sequence a larger piece of DNA
 - Whole genome shotgun sequencing (WGS)

- Source is broken randomly into fragments

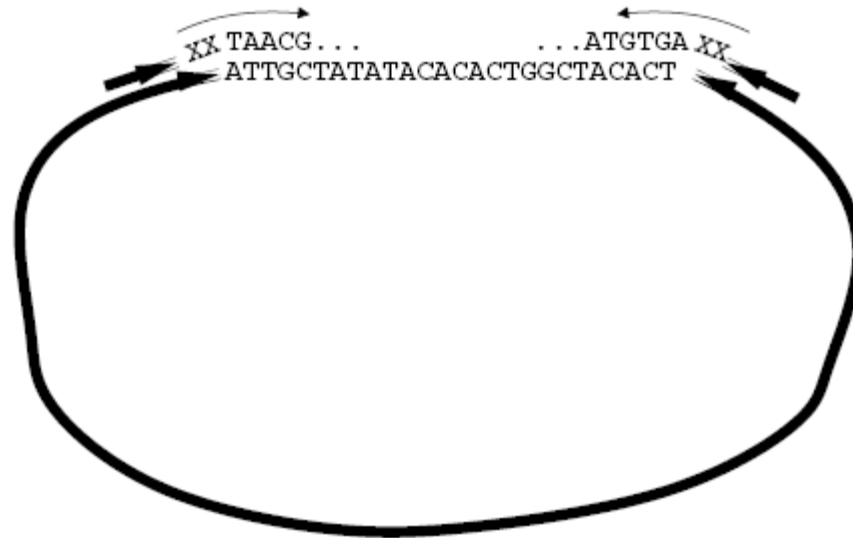
AGCGCGCTATATCGACTACG
TGCACTAGCACAGCGCGCTATATCGACT
ACGACTCAGC
ACGACTCAGC
CGCTATATCGACTACGA
ACTAGCACAGCGCGA
ACTAGCACAGCGCGA
TGCACTAGCACAGCGCGCTATATCGACT
ACGACTCAGC
ACGTTGCACTAGCACAGCGCGCT
CGCTATATCGACTACGA
CGCTATATCGACTACGA
TACGACTACGACTCAGCA

- Fragments are size selected
 - 2kb, 10kb, 50kb or 150kb

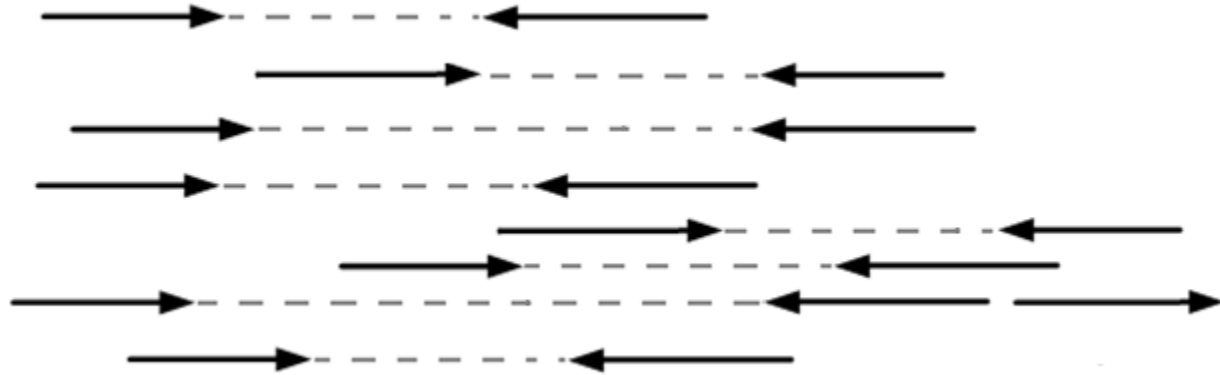
```
ACCGGCAGCAGCAGCACAGA  
AGCAGCAGCGCACAGACGACACG  
ATATATACACACTGGCTACTC  
ATTGCTATATACACACTGGCTACA
```

```
ACCGGCAGC  
AGCAGCAGCG  
ATATATACACTC  
ATTGCTATATACA
```

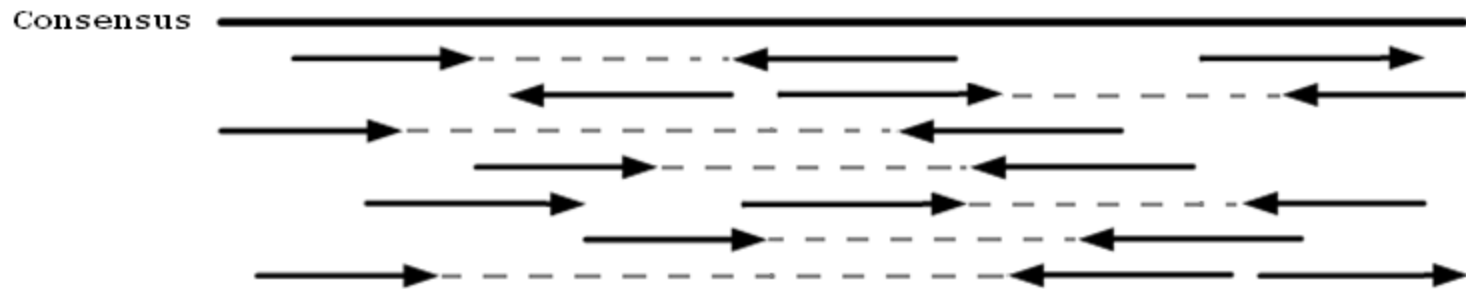
- Insert fragments into cloning vectors
- Sequence each clone from both ends to obtain a mate-pair of reads



- Input: Set of mate-pair reads

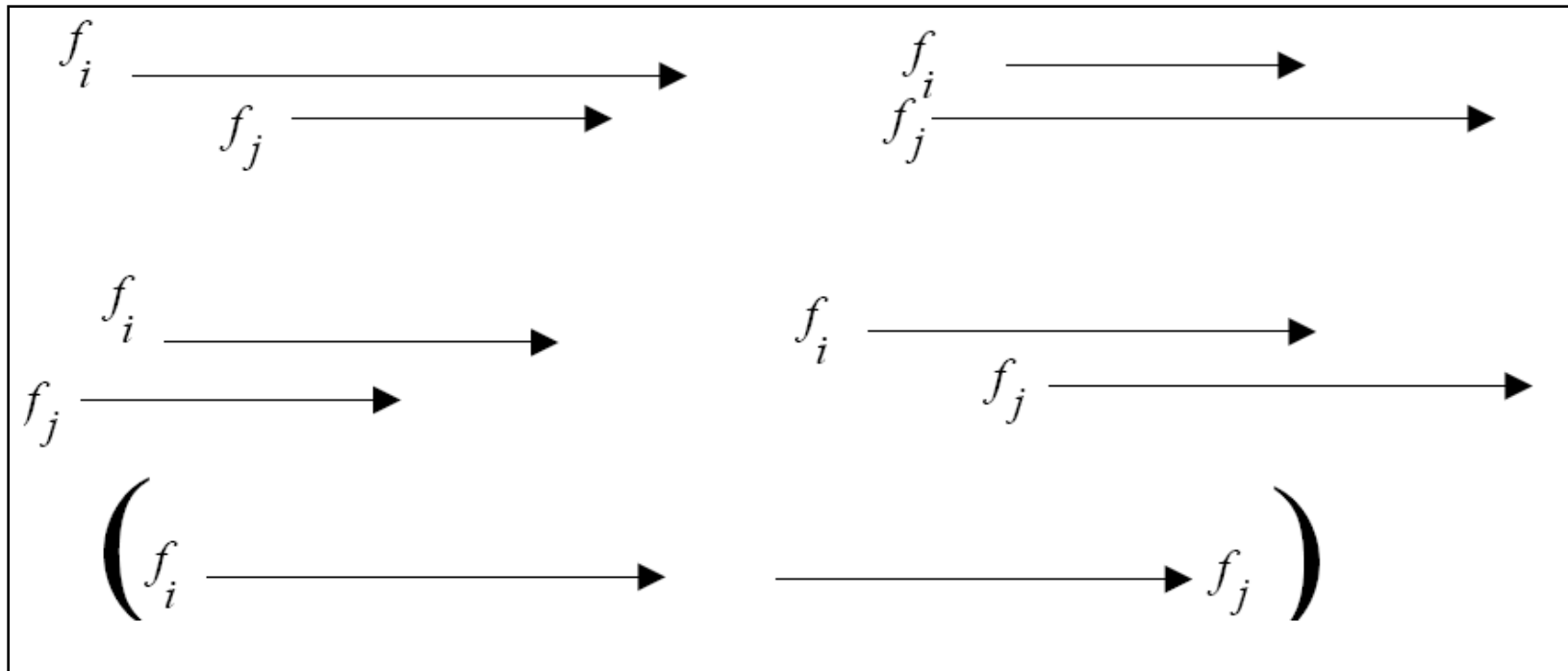


- Output



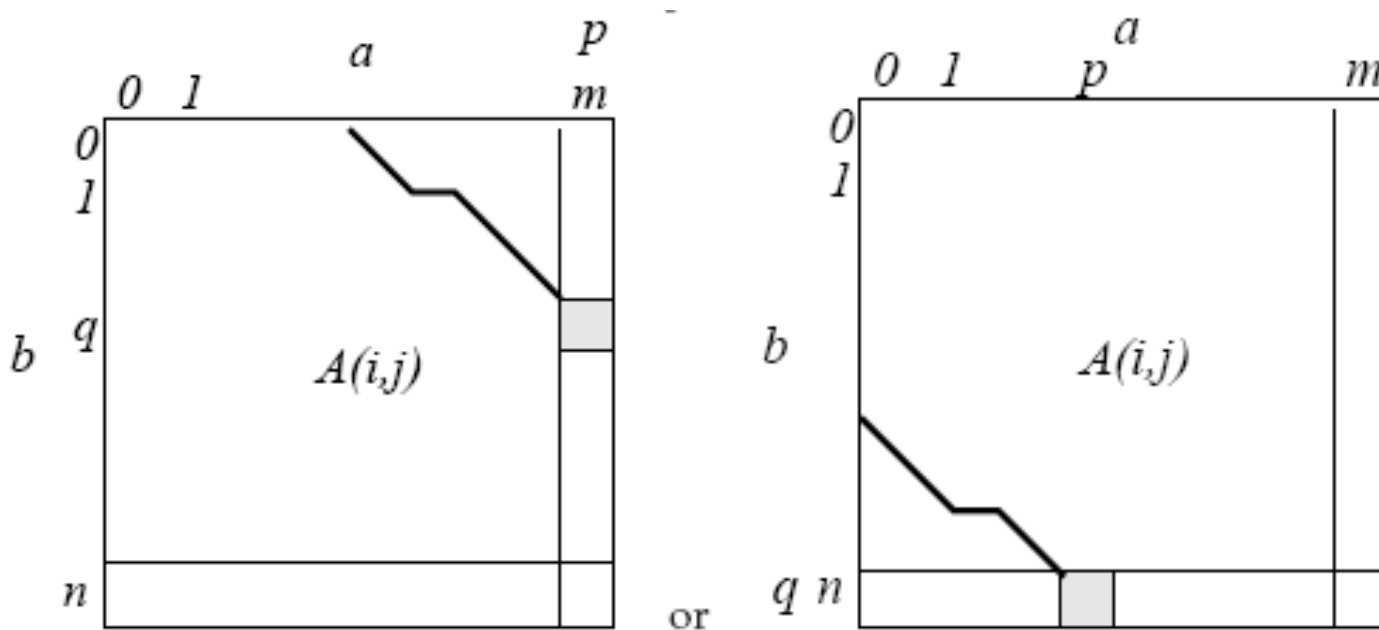
- Overlap phase

- For each pair of reads get a potential overlap

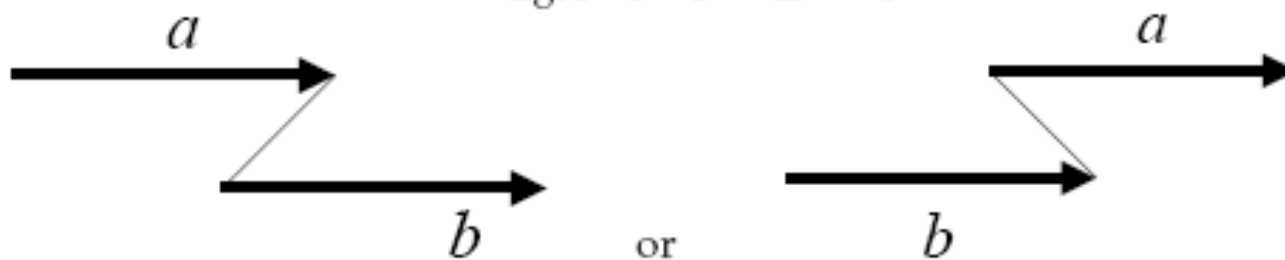


- The number of possible relations doubles, when we also consider the reverse read

Dynamic programming



The alignments look like this:

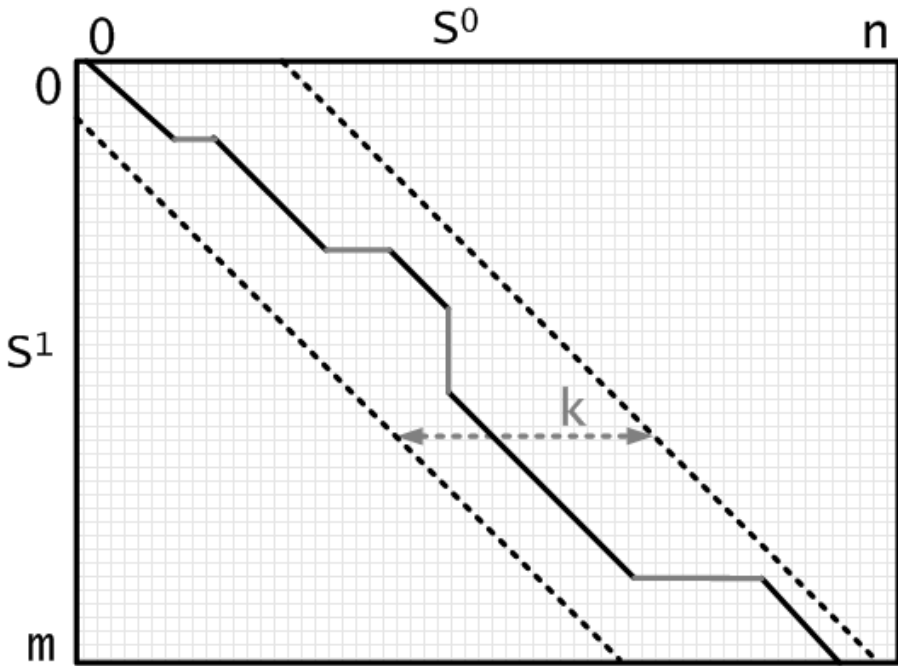
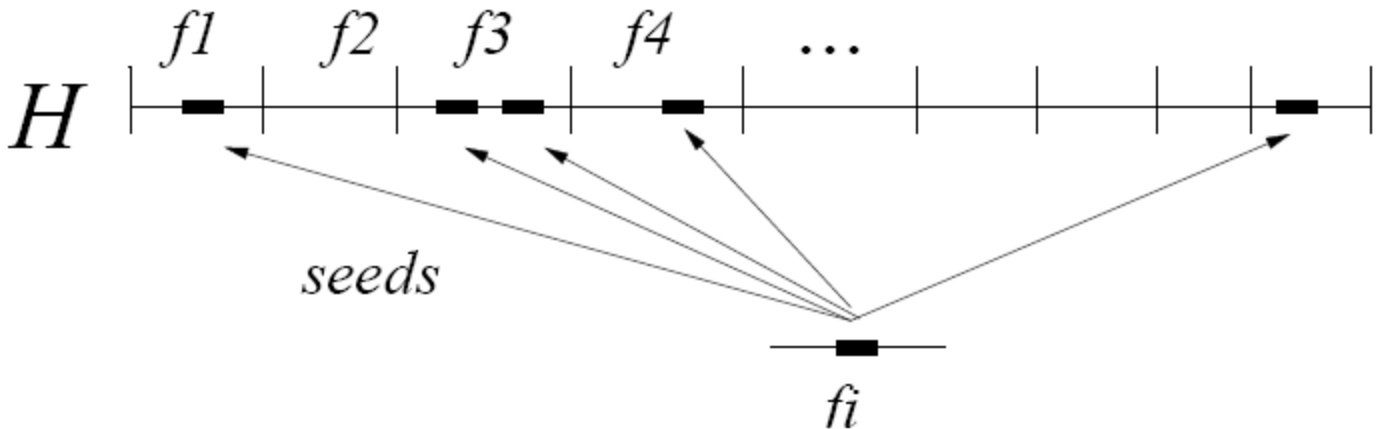


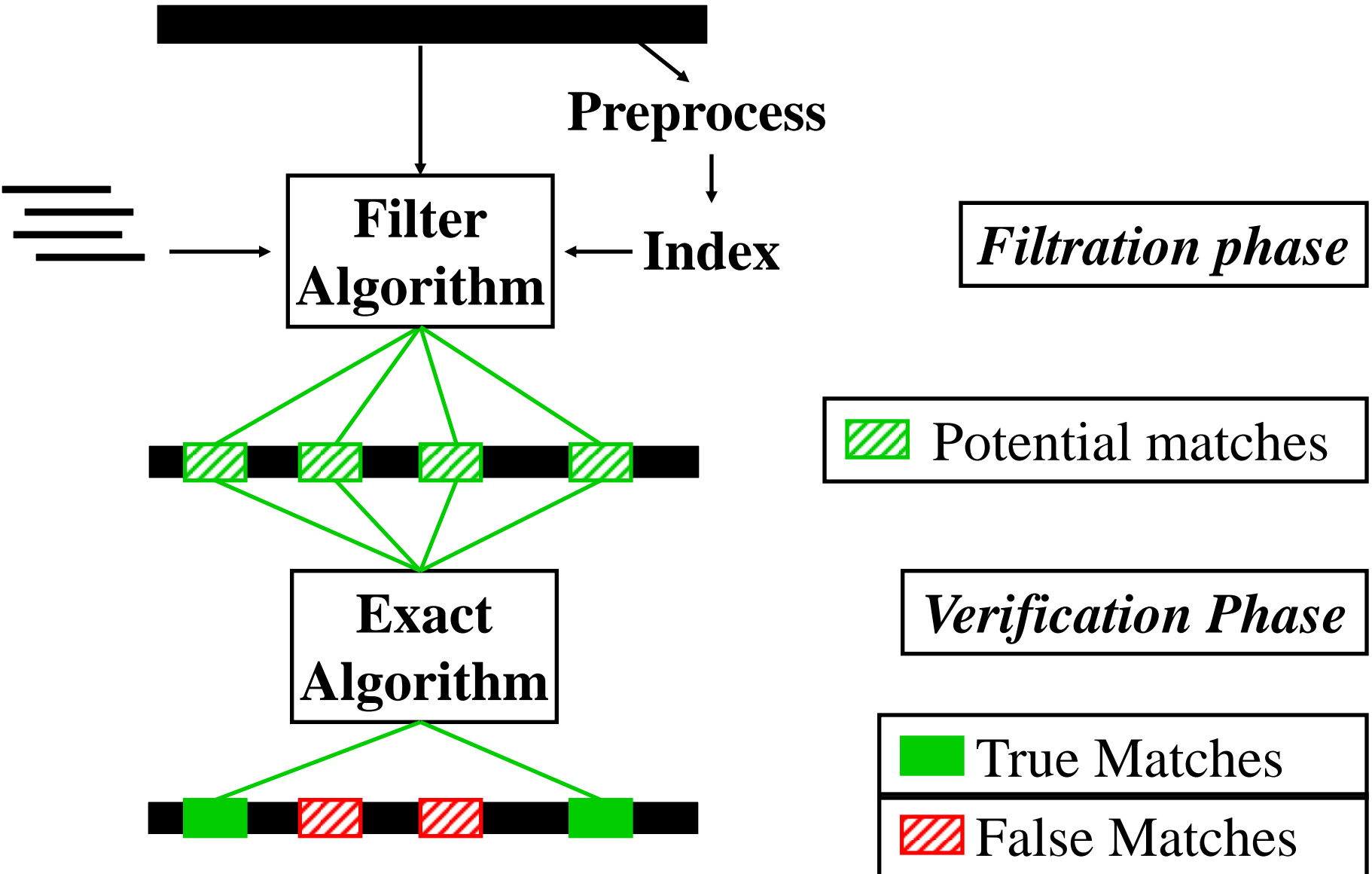
**Human genome assembly:
About 27 million reads**

Number of required comparisons:

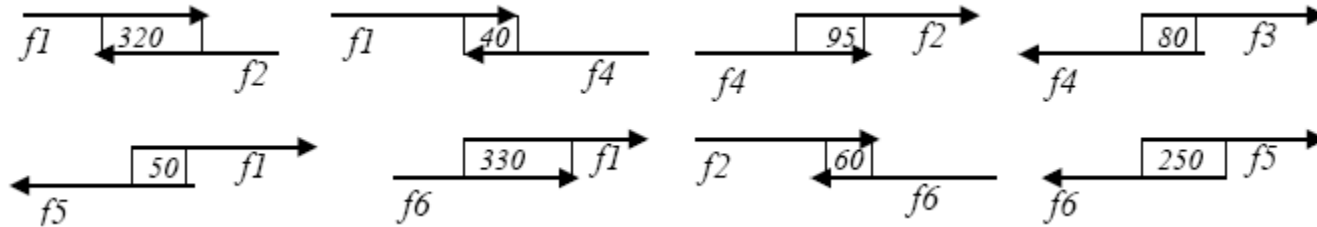
$$2 \cdot \binom{27000000}{2} \approx 1458000000000000 \approx 1.5 \cdot 10^{15}$$

Seed and extend approach

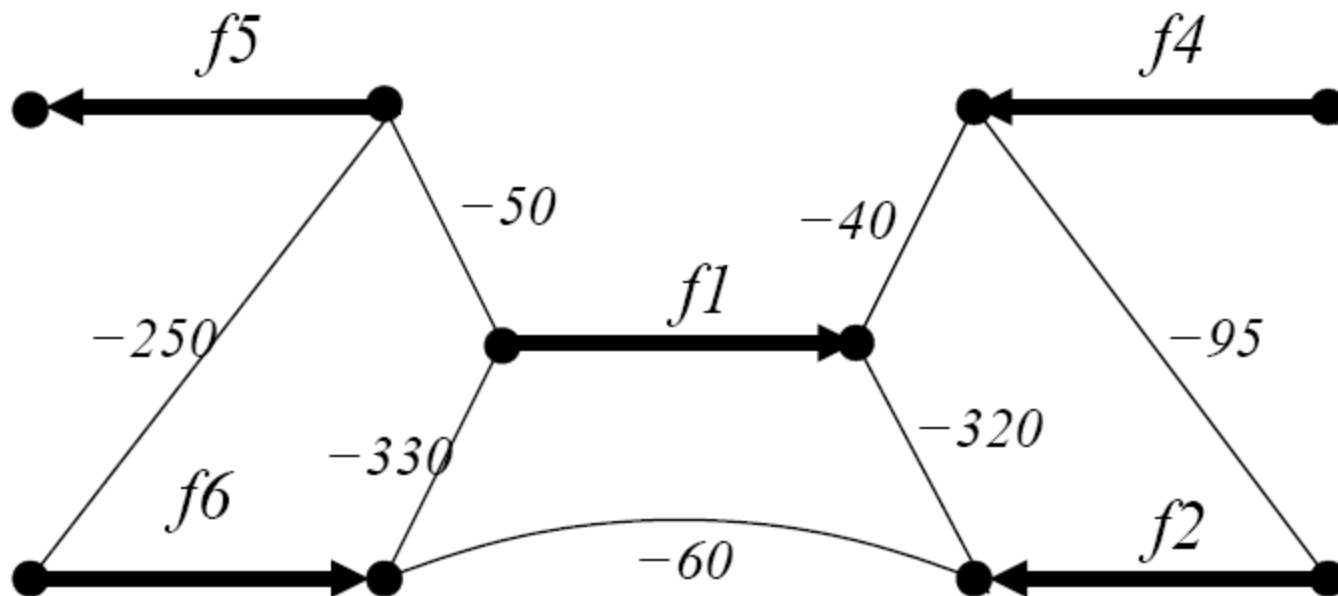




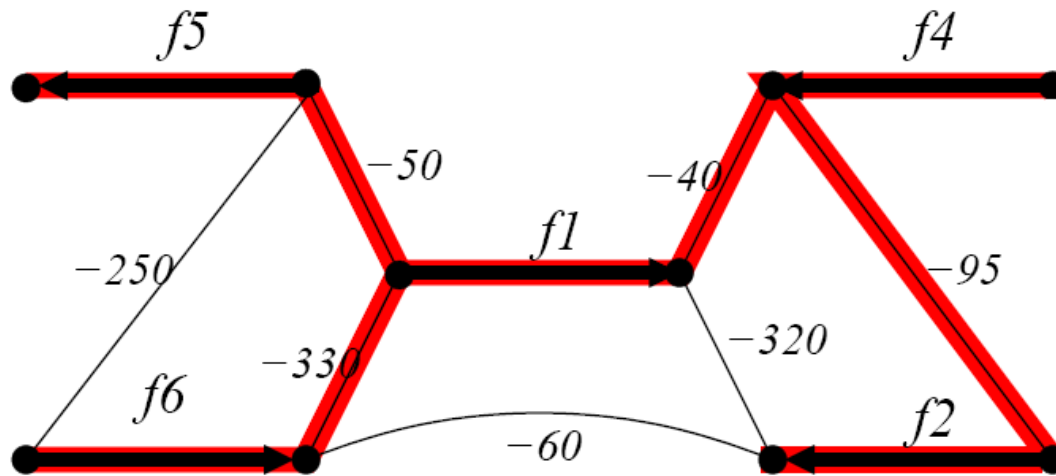
- Potential overlaps



- Are stored in a graph

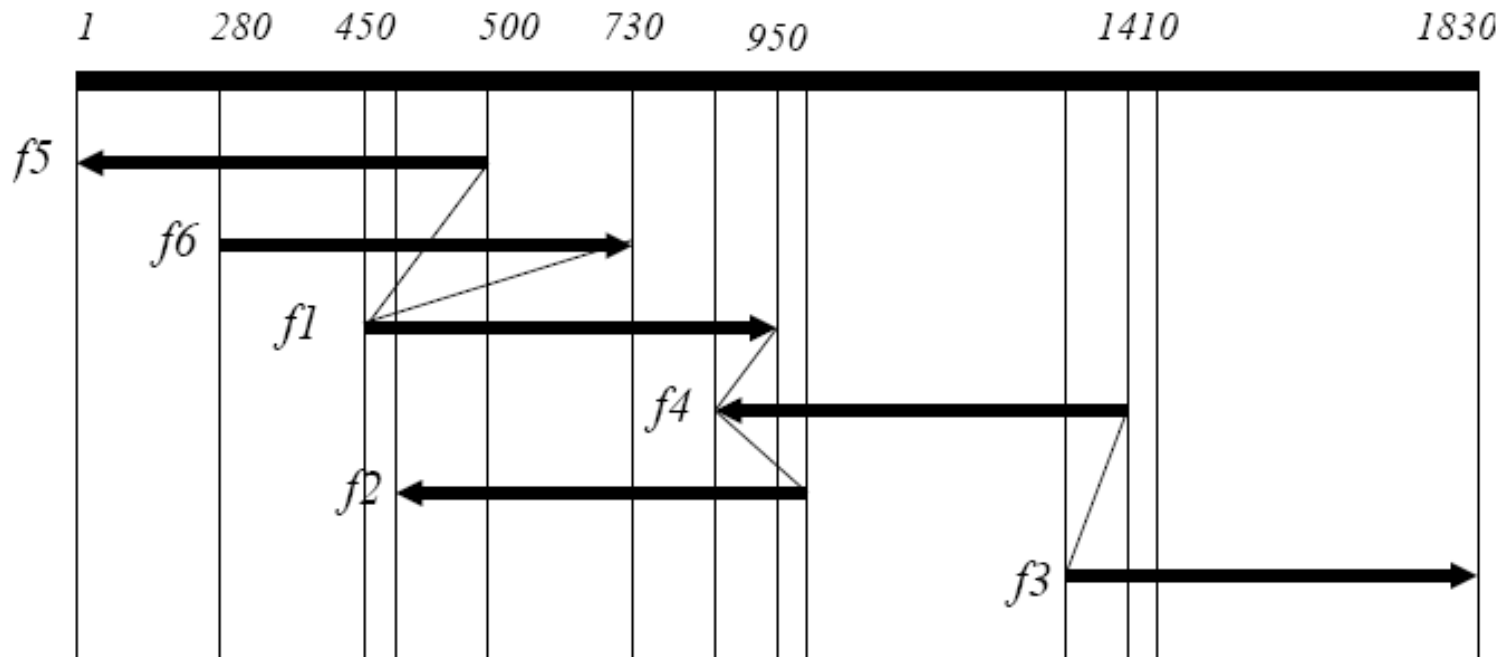
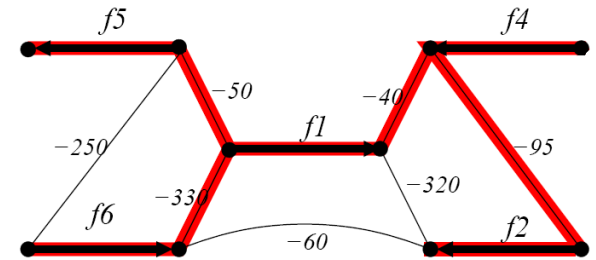


- A simple heuristic: Spanning tree



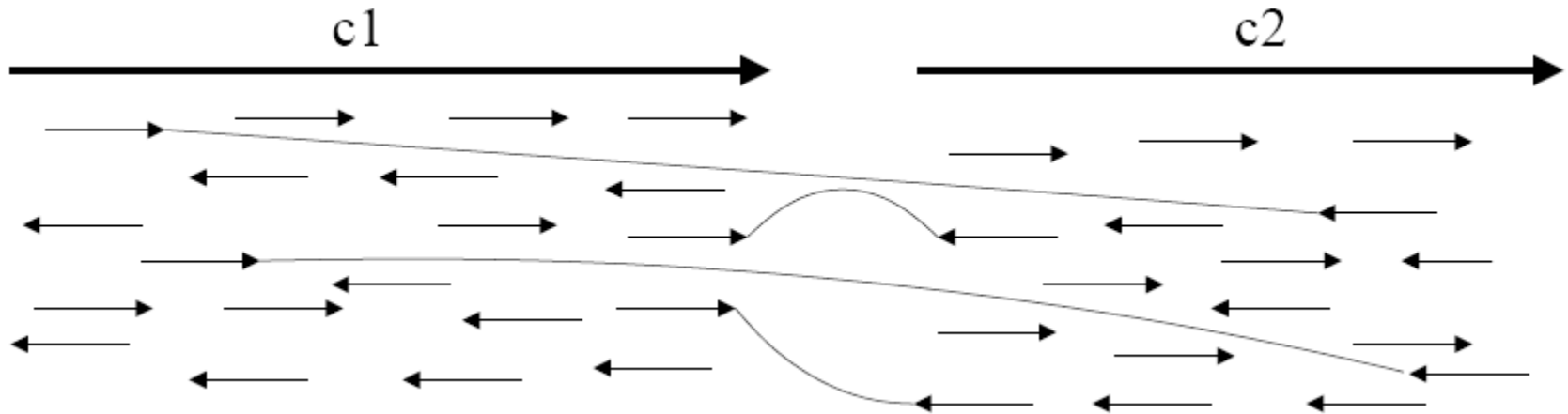
- In fact, it will be a spanning forest due to repeats or low-coverage regions

- Layout the reads

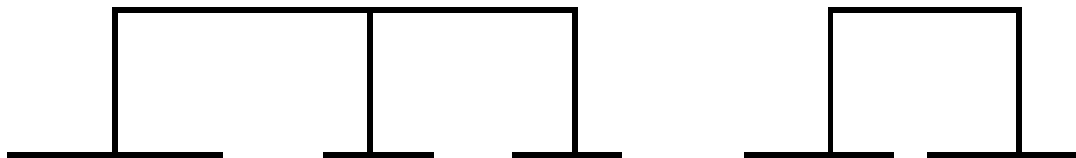


- Our first putative contig

- Scaffold the contigs with the help of the mate-pairs

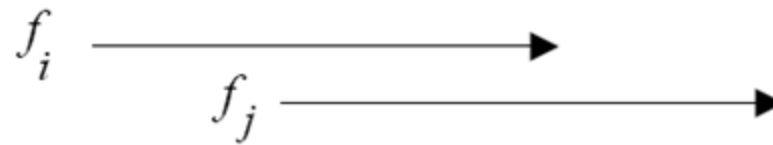


- Result: A set of scaffolds

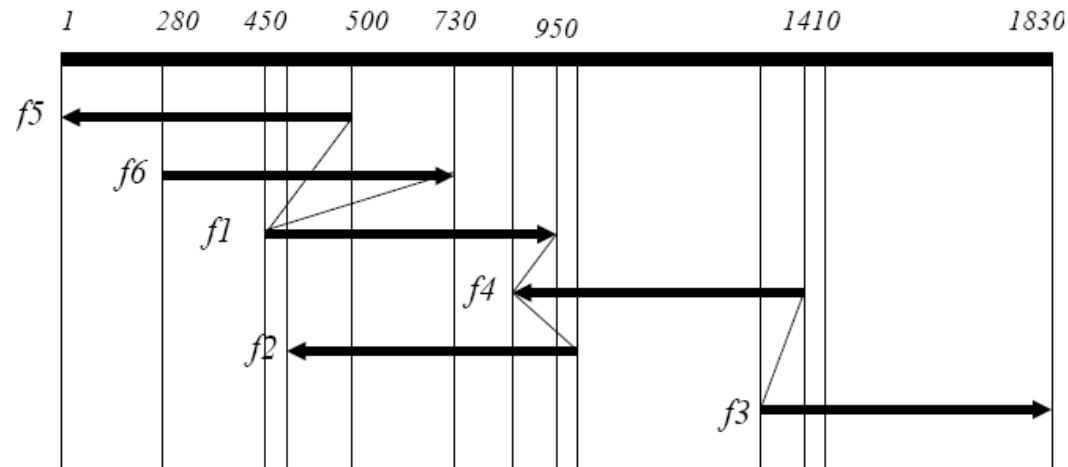


Summary: Classical assembly

- Overlap



- Layout



- Consensus

```
R1          ACGCTCCAACCGCTAATACG
R2                ATCGCTAATCCACGCCCGCCCCGC
R3          AAAC-CTCCAACCG
R4                TGCGCGCCCGCCCCGAAACCGC
Consensus AAAC-CTCCAACCGCTAATGCGCGCCCGCCCCGAAACCGC
```

Input: 27 million fragments of av. length 550bp, 70% paired:

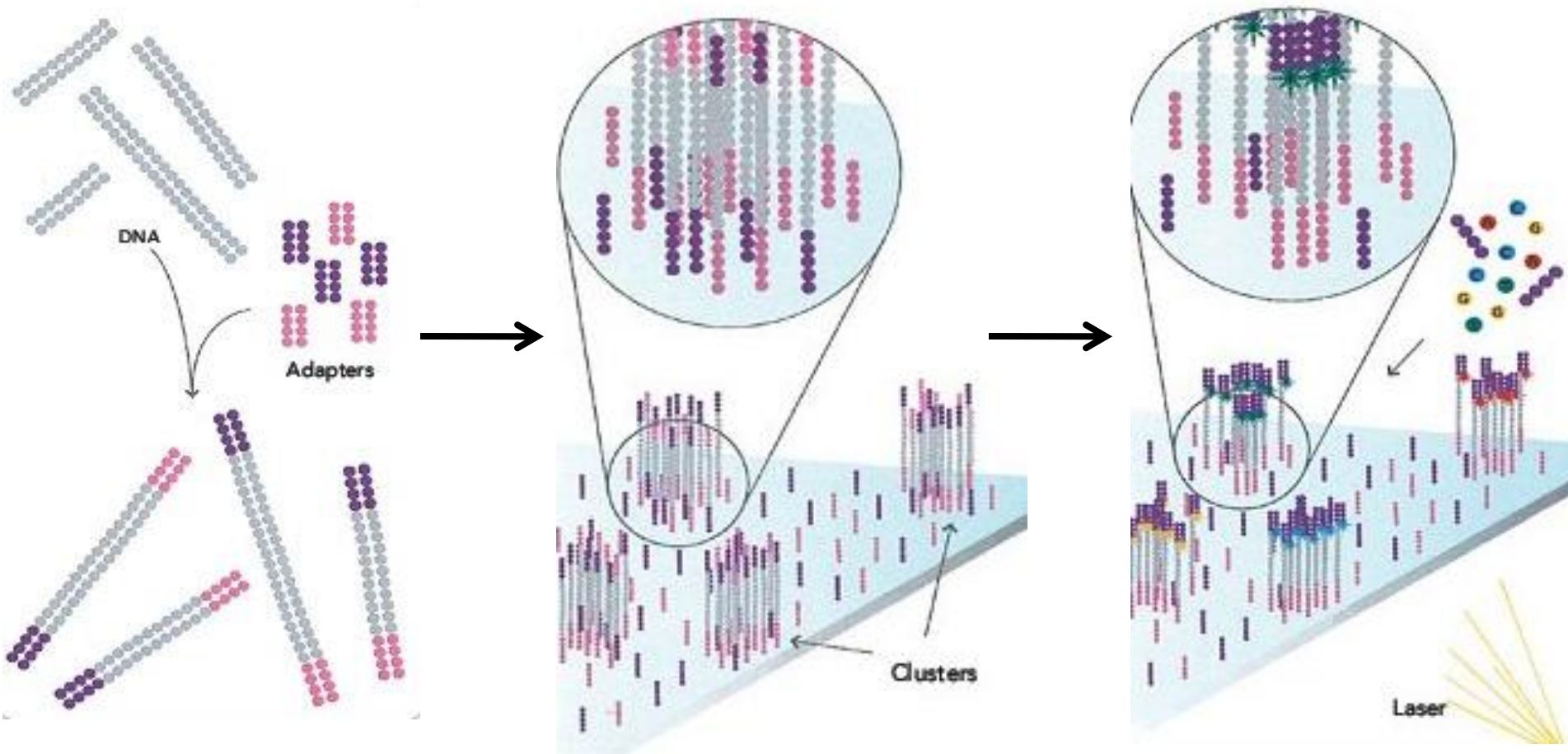
5m	pairs of length 2kb
4m	pairs of length 10kb
0.9m	pairs of length 50kb
0.35m	pairs of length 150kb

Celera's assembler uses approximately the following resources:

Program	CPU hours		Max. memory
Screeener	4800	2-3 days on 10-20 computers	2GB
Overlapper	12000	10 days on 10-20 computers	4GB
Unitigger	120	4-5 days on a single computer	32GB
Scaffolder	120	4-5 days on a single computer	32GB
RepeatRez	50	Two days on a single computer	32GB
Consensus	160	One day on 10-20 computers	2GB

Total: \approx 18000 CPU hours.

- Illumina sequencing process

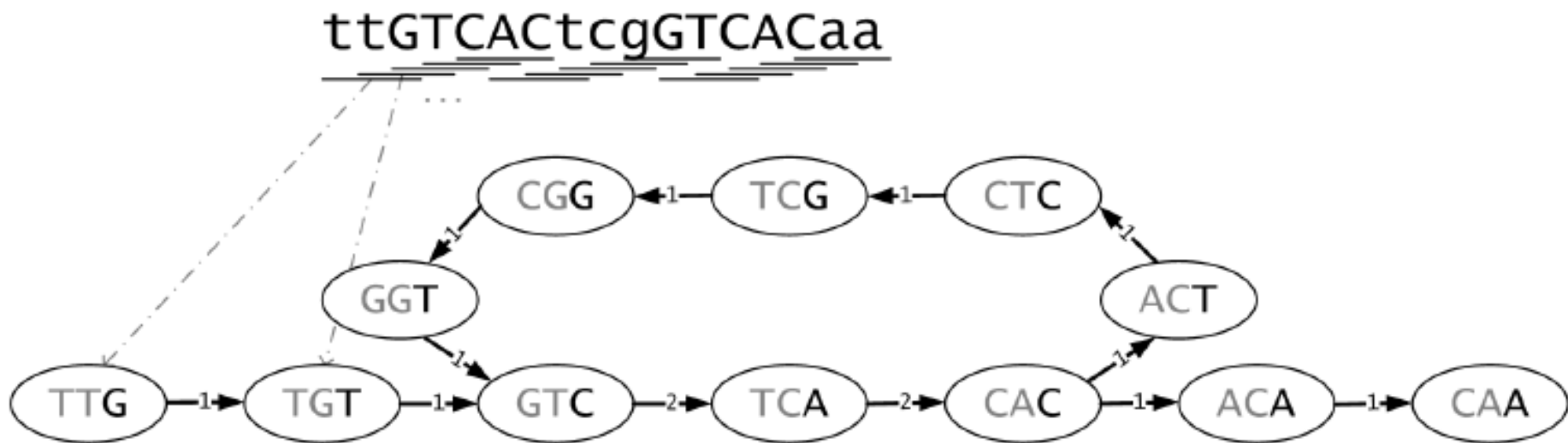


Much shorter reads, much higher throughput

Next-Generation Sequencer	Library Type	No. of hours of days/run	Mappable output/run	Read length
GS FLX Titanium (454 Life Sciences)	Single read	10 hours	0.4-0.6 GB	400 bp average
GS FLX Standard (454 Life Sciences)	Mate-pair	7.5 hours	0.1 GB	250 bp average
Genome Analyzer (Illumina)	Single read	2 days	2-3 GB	35 bp
	Mate-pair	5.5 days	8-10 GB	2x50 bp
SOLiD (Applied Biosystems)	Single read	6-7 days	10-15 GB	50 bp
	Mate-pair	12-14 days	20-30 GB	2x50 bp

Sanger sequencing: 50-100 KB per run, but read lengths from 500 - 1100 bp

- Reads are too short to compute a reliable pairwise overlap

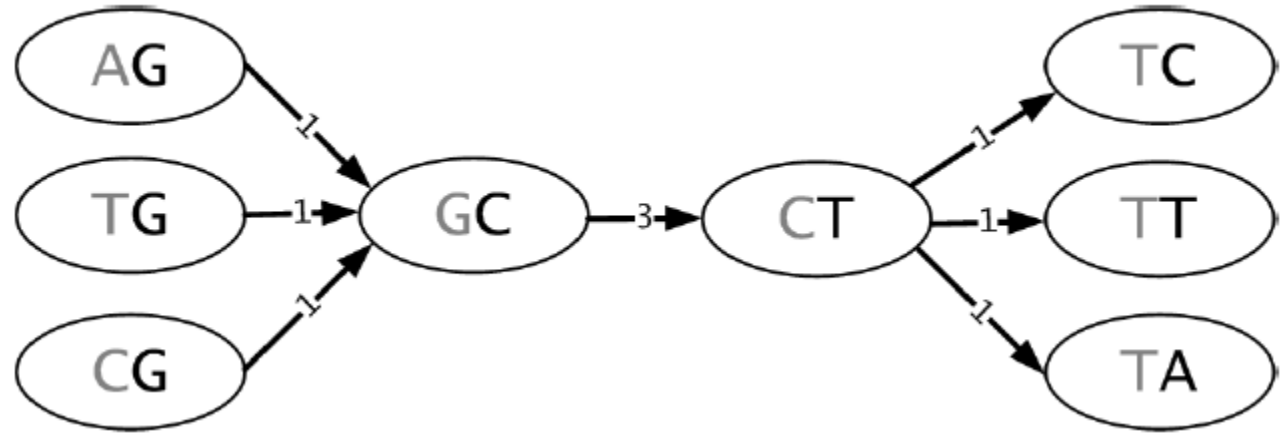


De Bruijn Graph

AGCTC

TGCTT

CGCTA



- Assembly greatly depends on the k-mer size
- Assembly quality measure: N50
 - Sort contigs by length
 - Add the contig's lengths until the summed length exceeds 50% of the total length of all contigs

Contig	A	B	C	D
Size (bp)	100	80	70	50

Assembly of the yeast SK1 strain

k-mer size	N50	longest contig
21	16584	72826
23	30407	133739
25	32967	149035
27	32569	274179
29	23160	134911
31	9763	47081

- Short-read assemblers
 - EULER-SR
 - Velvet
 - ABySS
 - Newbler (454 Data)
 - Others: ALLPATHS, SSAKE, ...
- Overlap-layout-consensus assemblers
 - Celera assembler
 - Arachne
 - Others: Atlas, JAZZ, PCAP, ...

- No ready-made tool available yet
- Possible approach
 - Use Velvet to compute reliable contigs with high coverage
 - Pull out these putative transcript sequences
 - Align all remaining reads against those transcripts with a low tolerance for errors
 - Use Velvet again on the remaining reads with more sensitive settings
 - Pull out those transcripts and so on

- **Genecore**

- Vladimir Benes
- Jonathon Blake
- Tomi Bähr-Ivacevic
- Richard Paul Carmouche
- Jos de Graaf
- David Ibberson
- Sabine Schmidt
- Jens Stolte
- Jürgen Zimmermann

- **Furlong group**

- Nicolas Delhomme

- **Korbel group**

- Jan Korbel
- Megumi Onishi-Seebacher
- Andreas Schlattl
- Adrian Stuetz
- Thomas Zichner

- **Genecore**

- Vladimir Benes
- Jonathon Blake
- Tomi Bähr-Ivacevic
- Richard Paul Carmouche
- Jos de Graaf
- David Ibberson
- Sabine Schmidt
- Jens Stolte
- Jürgen Zimmermann

- **Furlong group**

- Nicolas Delhomme

- **Korbel group**

- Jan Korbel
- Megumi Onishi-Seebacher
- Andreas Schlattl
- Adrian Stuetz
- Thomas Zichner

Thank you!