RSG-Turkey Student Symposium 2021
12.09.2021

Introduction to Genome Assembly

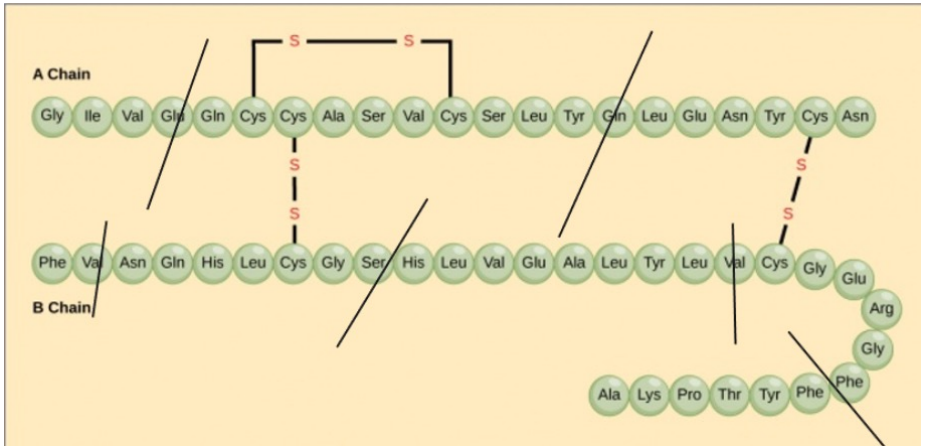
Here is where bioinformatics begins

Yasin Kaya

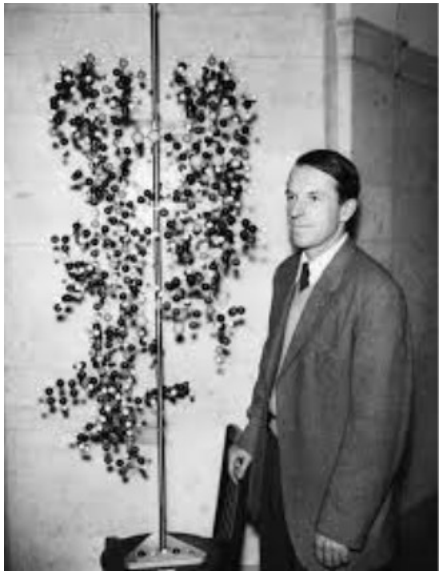
01

A brief history of sequencing

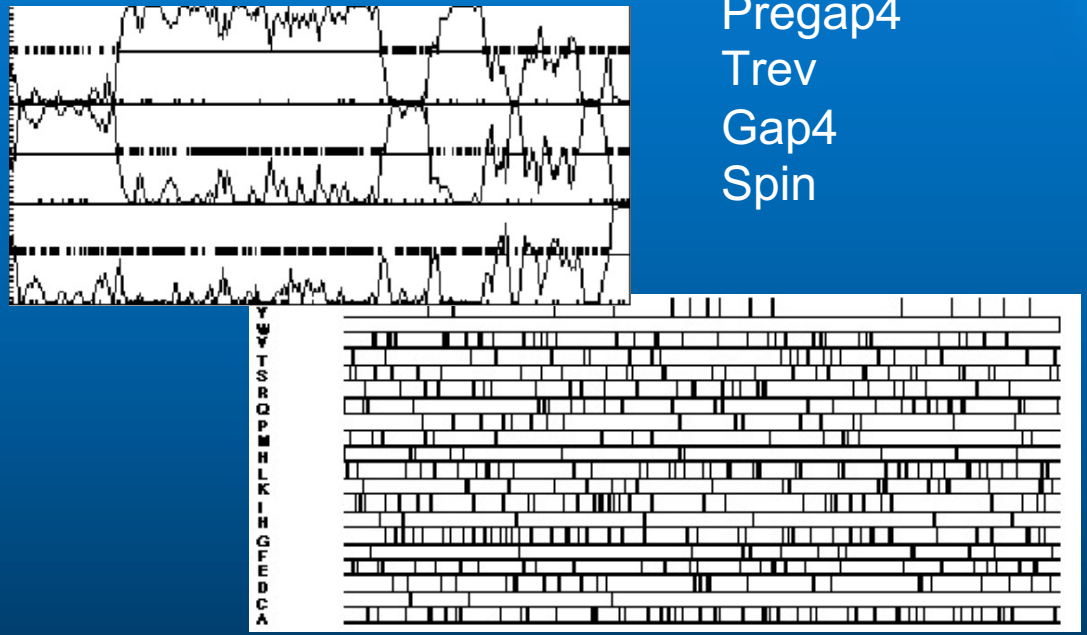
Refined partition chromatography



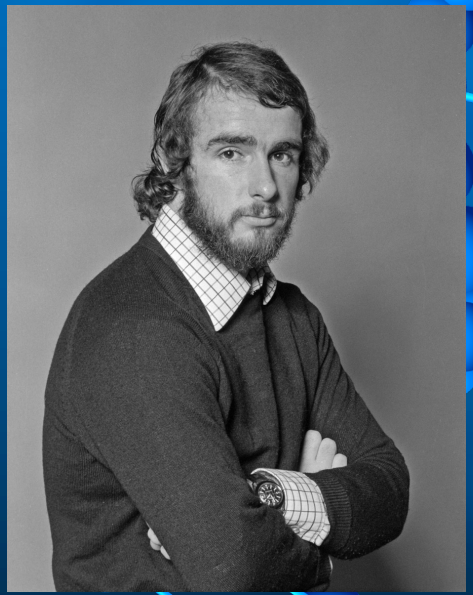
Gly-Ile-Val-Glu-Gln-Cys-Cys-Ala-Ser-Val-Cys-Ser
Gly-Ile-Val-Glu-Gln-Cys-Cys-Ala
Cys-Ala-Ser-Val
Gly-Ile-Val-Glu
Cys-Cys-Ala-Ser-Val-Cys
Val-Glu-Gln-Cys-Cys-Ala-Ser
Cys-Cys-Ala-Ser-Val-Cys-Ser



Rodger Staden : first DNA sequencing ‘software’ 1977



Pregap4
Trev
Gap4
Spin



jmb
Journal of Molecular Biology
Volume 162, Issue 4, 25 December 1982, Pages 729-773

Nucleotide sequence of bacteriophage λ DNA

F. Sanger, A.R. Coulson, G.F. Hong, D.F. Hill †, G.B. Petersen †

Show more

+ Add to Mendeley Cite

https://doi.org/10.1016/0022-2836(82)90546-0

Get rights and content

Abstract

The nucleotide sequence of the DNA of bacteriophage λ has been determined using the dideoxy chain termination method in conjunction with random cloning in M13 vectors. Various methods were studied for sequencing specific regions to complete the sequence, but all were much slower than the random approach. The DNA in its circular form contains 48,502 base-pairs. Open reading frames were identified and, where possible, ascribed to genes by comparing with the previously determined genetic map. The reading frames for 46 genes were clearly identified, though in about 20 the position of the protein initiation site could not be rigorously established. Probable positions for the *kil*, *cIII* and *lom* genes are suggested but remain uncertain. There are about 20 other unidentified reading frames that may code for proteins.

› Nature. 1977 Feb 24;265(5596):687-95. doi: 10.1038/265687a0.

Nucleotide sequence of bacteriophage phi X174 DNA

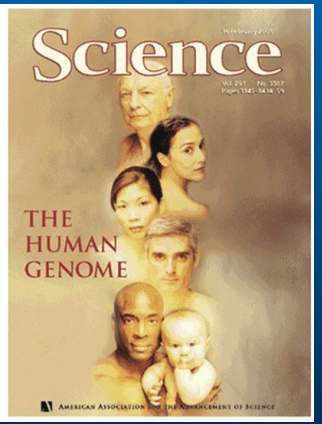
F Sanger, G M Air, B G Barrell, N L Brown, A R Coulson, C A Fiddes, C A Hutchison, P M Slocombe, M Smith

PMID: 870828 DOI: 10.1038/265687a0

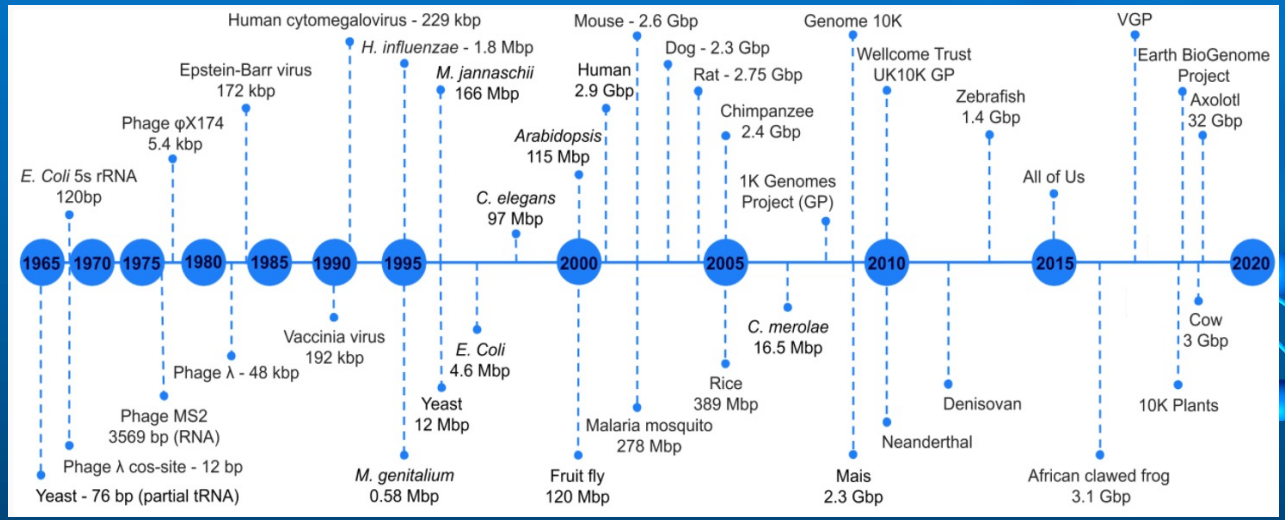
Abstract

A DNA sequence for the genome of bacteriophage phi X174 of approximately 5,375 nucleotides has been determined using the rapid and simple 'plus and minus' method. The sequence identifies many of the features responsible for the production of the proteins of the nine known genes of the organism, including initiation and termination sites for the proteins and RNAs. Two pairs of genes are coded by the same region of DNA using different reading frames.

+Decoding the book of life!
-Really :) ?

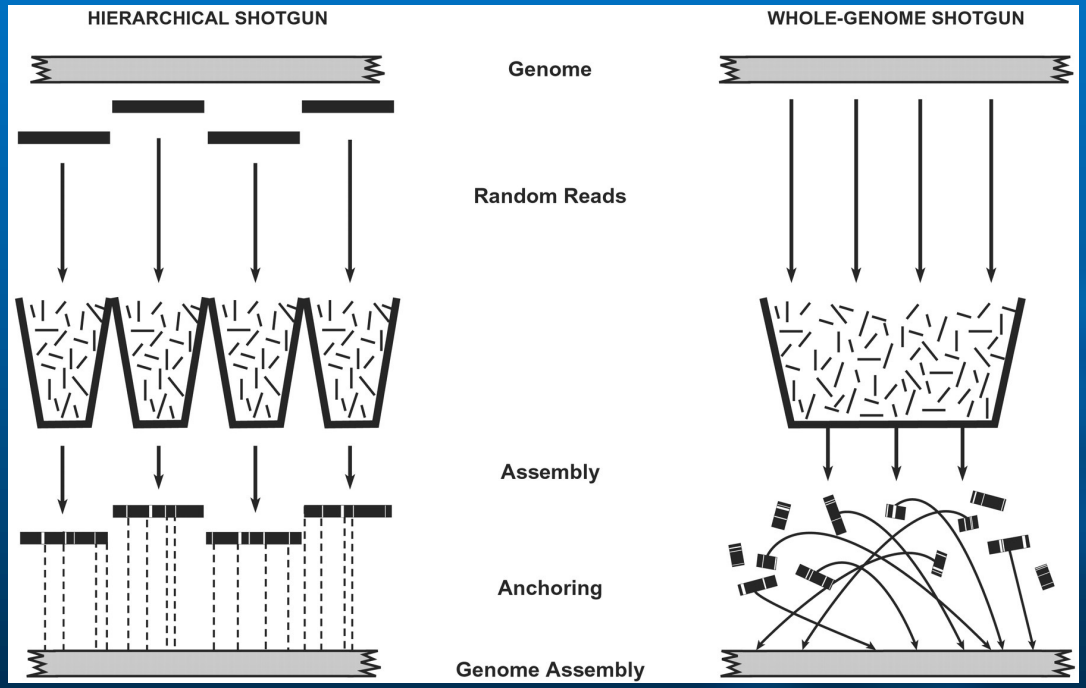


Why are genome projects or WGS necessary?



De novo genome assembly

The process of determining the sequence of an organism without existing reference.



Public human genome project

Private human genome project (Celera)

Cauliflower mosaic virus

Volume 9 Number 12 1981 Nucleic Acids Research

The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing

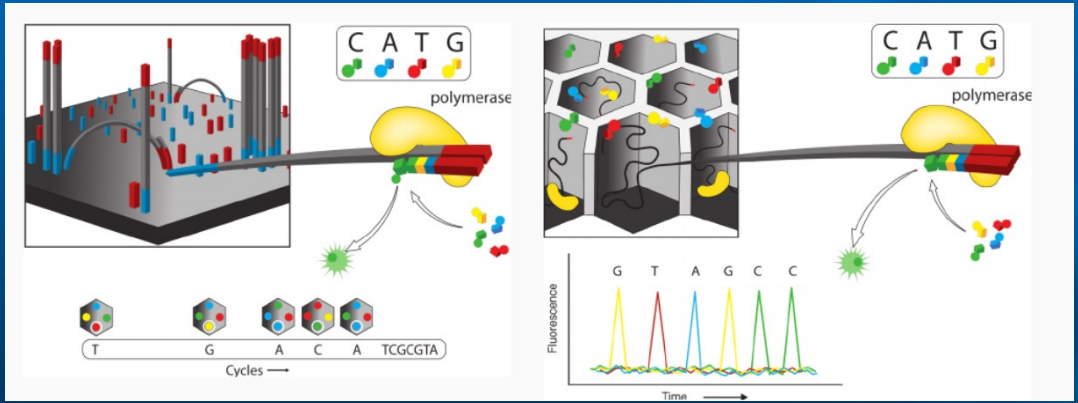
Richard C. Gardner, Alan J. Howarth, Peter Hahn, Marianne Brown-Luedi[2], Robert J. Shepherd[1] and Joachim Messing[2,3]

Department of Plant Pathology, and [2]Department of Bacteriology, University of California, Davis, CA 95616, USA

Received 21 April 1981

NGS based WGS

- Better
- Faster
- Cheaper

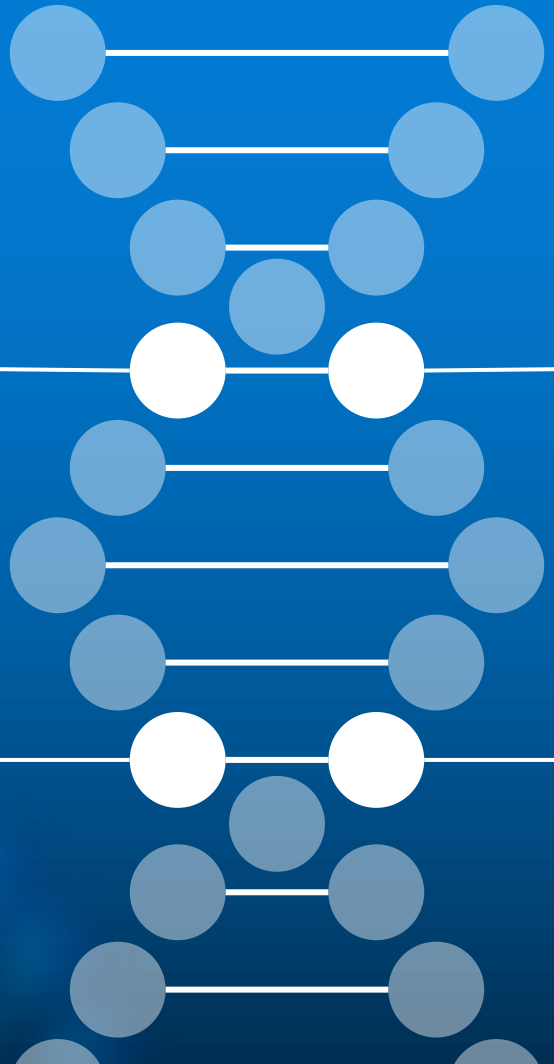


NGS vs Third Gen.

Bridge amplification
Short reads
High throughput
High quality

Single molecule
Long reads
Lower throughput
Lower quality

Challenges



Large genome sizes

Complexity

High chromosome numbers

Data storage/management

everything starts with …?

- Brute force
- Global alignments
- Local alignments
- Mixed global/local alignments
- Alignment free approaches:

Transcript quantification
Genotyping
De novo genome assembly
Metagenomics
Barcoding

Assembly algorithms

Data model

Overlap-Layout-Consensus (OLC)
Eulerian / de Bruijn Graph (DBG)

Search method

Greedy – Non-greedy

Parallelizability

Multithreaded
Distributable

Overlap–Layout–Consensus *vs* de-bruijn-Graph

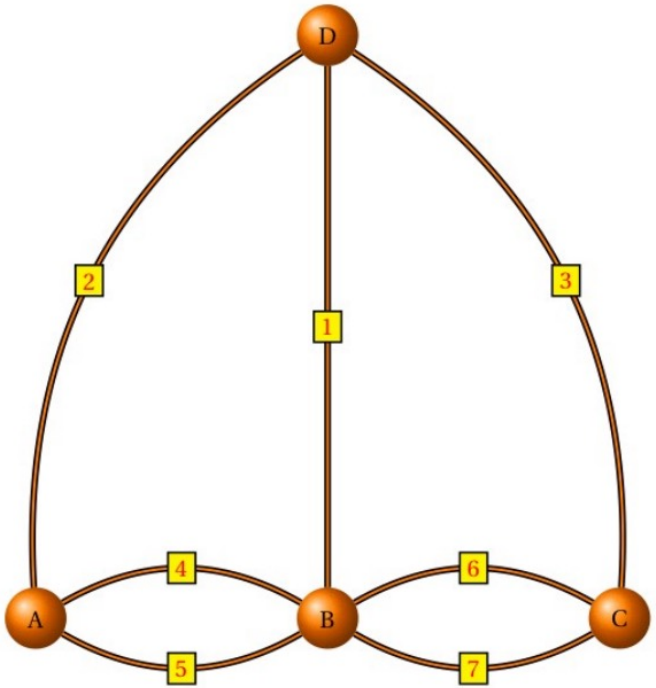
1- Overlap

Reads ⇒ **4 nodes / vertices**

A, B, C, D

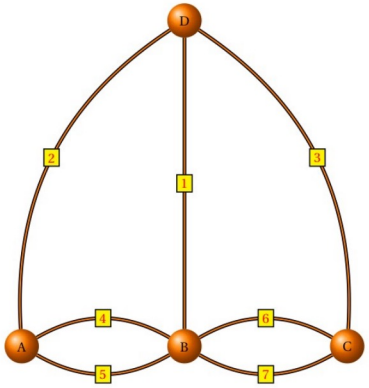
Overlaps ⇒ **7 edges / arcs**

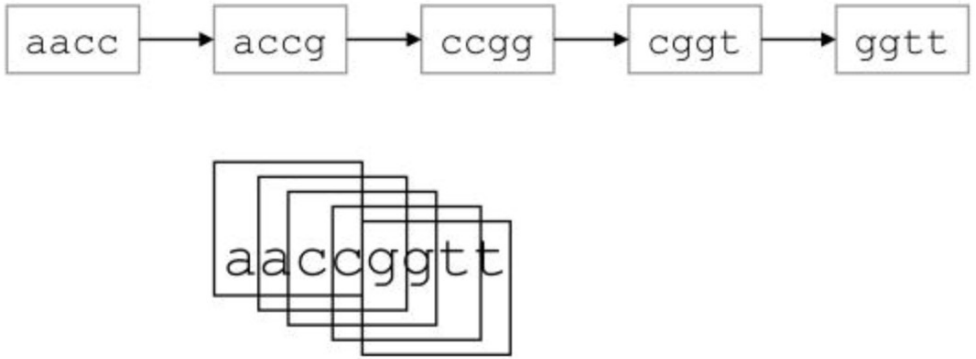
1,2,3,4,5,6,7



Overlap–Layout–Consensus

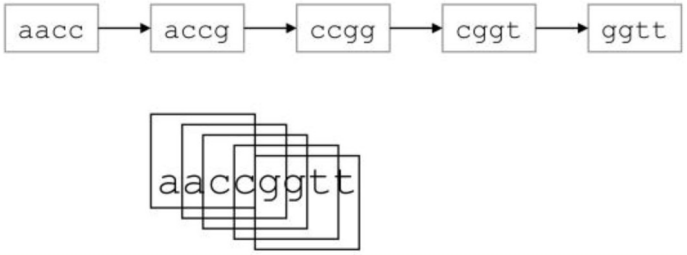
2- Layout

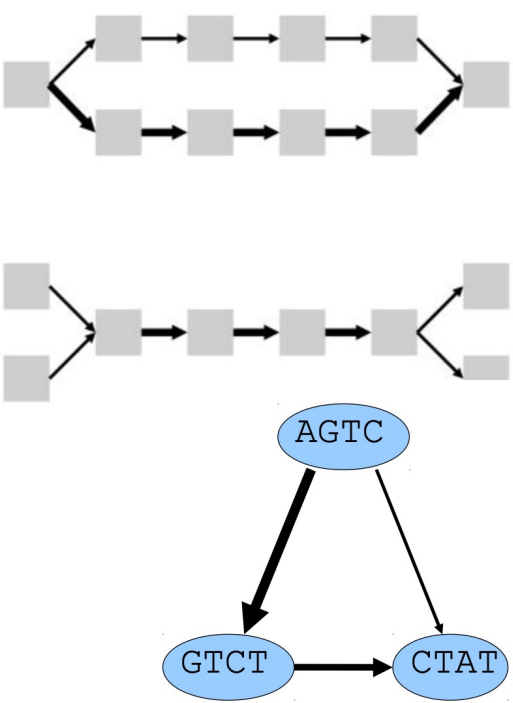




Overlap–Layout–Consensus

3- Consensus





de-bruijn-Graph

AACCGG

AACC ACCG CCGG K-mers: (k=4)

L: **Length of seq**
K: **k-mers**
Final: **L-K+1**

Overlap–Layout–Consensus *vs* de-bruijn-Graph

Less sensitive to repeats and read errors

Graph construction more time consuming

cannot scale to voluminous short reads

More sensitive to repeats and read errors

Graph converges at repeats of length k

One read error introduces k false nodes

Assembling a genome today

Short read assemblers:

SGA *String graph*
ValVel *String graph*
DISCOVAR *DBG*
SOAPdenovo *DBG*
Euler *DBG*
ABySS *DBG*
Velvet *DBG*
SPAdes *DBG*
Edena *OLC*
Ray *Hybrid*
SSAKE *Greedy*
Perga *Greedy*

Long read assemblers:

Hifiasm *OLC*

Canu/HiCanu (ex Celera) *OLC*

Peregrine *HGAP/OLC*

Falcon Unzip *HGAP/OLC*

Flye *Repeat graph*

Genome finishing methods

Fill gaps, join contigs and publish it!!

Close gaps

SCARPA, SSPACE, BESST

Join contigs/scaffolds

Optical map or mate-pair sequences

to obtain good number of scaffolds

Genome Assembly Gold-Standard Evaluations

GENOME RESEARCH
CSHL Press | Journal Home | Subscriptions | eTOC Alerts | BioSupplyNet

Genome Res. 2012 Mar; 22(3): 557–567.
doi: 10.1101/gr.131383.111

PMCID: PMC3290791
PMID: 22147368

GAGE: A critical evaluation of genome assemblies and assembly algorithms

Steven L. Salzberg,[1,7] Adam M. Phillippy,[2] Aleksey Zimin,[3] Daniela Puiu,[1] Tanja Magoc,[1] Sergey Koren,[2,4] Todd J. Treangen,[1] Michael C. Schatz,[5] Arthur L. Delcher,[6] Michael Roberts,[3] Guillaume Marçais,[3] Mihai Pop,[4] and James A. Yorke[3]

1-Genome Integrity

2-Genome continuity

For the sake of the contiguity..

Nb of IUPAC nucleotides **matters**

Nb of contigs/scaffolds **matters**

GC content (%) **matters**

Coverage **matters**

N50 score **matters**
(Corresponds to the length of the scaffold when the summation of total assembly length)

Last step.. File formats

Most reads are supplied as “**fastq**”

```
@HWUSI-EAS100R:6:73:941:1273
AGTCGCTTTAGAGTATCTTAGATTTTTCTCCTATGAGGAG
+
HWUSI-EAS100R:6:73:941:1273 hhhggggfdba[[^_Z_ZYXWWWWPQQQRNOOHGFBBBBB
```

1. '@' and unique sequence identifier (id)
2. Read sequence
3. '+' with optional duplication of id
4. Read quality (ASCII encoded)

Quality check & Trim -→ FastQC - cutadapt

Mismatch correction → Karect or SPAdes –careful step

K-mer selection → SGA-preqc or Kmergenie

Repeat masking → RepeatMasker libs

Assembly → SPAdes, Abyss, Velvet (Short-read) or Canu, FALCON (Long-read)

Mapping quality → Busco, minimap2, Quast etc.

THANKS!

Do you have any questions?
yyasinkkaya@gmail.com

yyasinkkaya

kaya_yasinn


Let’s practise..