Basics about Sequence Assembly for NGS data: Resolving genomic puzzles.

by
Aureliano Bombarely
ab782@cornell.edu

1. A brief history of the sequence assembly.

2. Sequencing, tools and computers.

3. Things that you should know about genomes.

4. What about transcriptomes ? Differences

1. A brief history of the sequence assembly.

2. Sequencing, tools and computers.

3. Things that you should know about genomes.

4. What about transcriptomes ? Differences

1. A brief history of the sequence assembly.

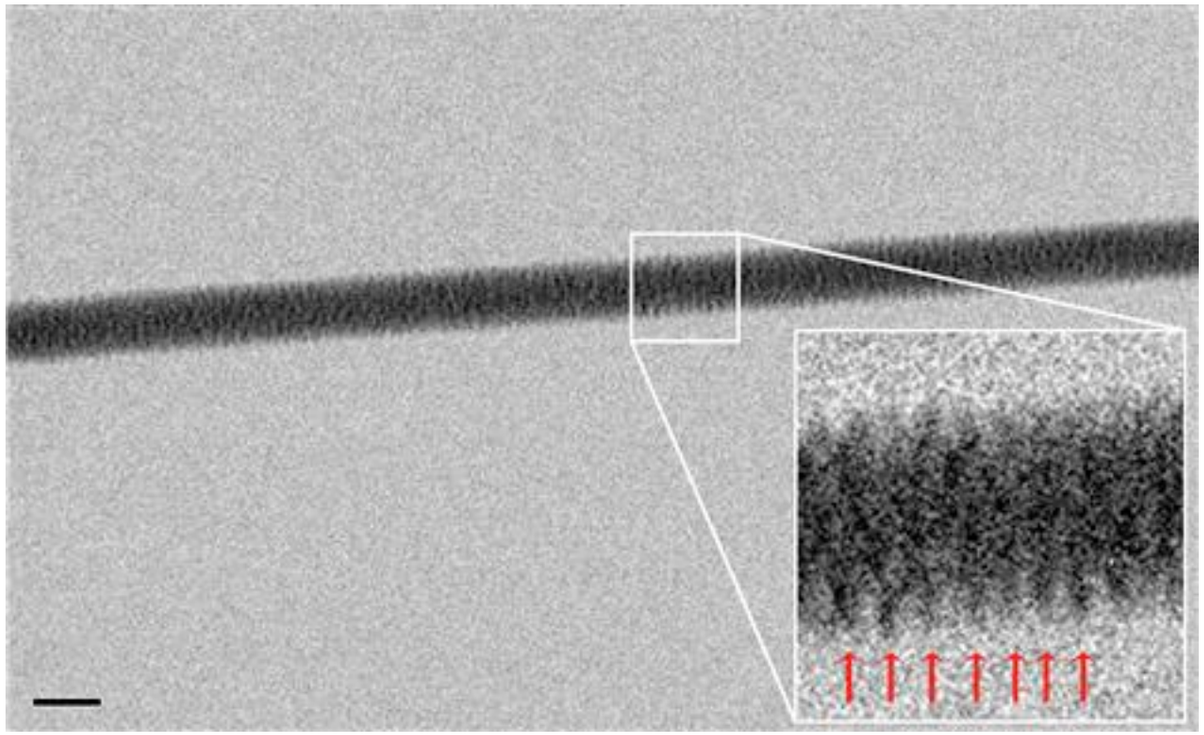
WHAT IS A GENOME?

Life is specified by **genomes**. Every organism, including humans, has a genome that contains all of the biological information needed to build and maintain a living example of that organism. The biological information contained in a genome is encoded in its **deoxyribonucleic acid (DNA)** and is divided into discrete units called **genes**. Genes code for proteins that attach to the genome at the appropriate positions and switch on a series of reactions called gene expression.

In 1909, Danish botanist Wilhelm Johanssen coined the word **gene** for the hereditary unit found on a chromosome. Nearly 50 years earlier, Gregor Mendel had characterized hereditary units as **factors**— observable differences that were passed from parent to offspring. Today we know that a single gene consists of a unique sequence of DNA that provides the complete instructions to make a functional product, called a protein. Genes instruct each cell type—such as skin, brain, and liver—to make discrete sets of proteins at just the right times, and it is through this specificity that unique organisms arise.

http://www.ncbi.nlm.nih.gov/About/primer/genetics_genome.html

Genome = N x Sequence of DNA



Gentile F. et al. *Direct Imaging of DNA Fibers: The Visage of Double Helix* Nano Lett., 2012, 12 (12), pp 6453–6458

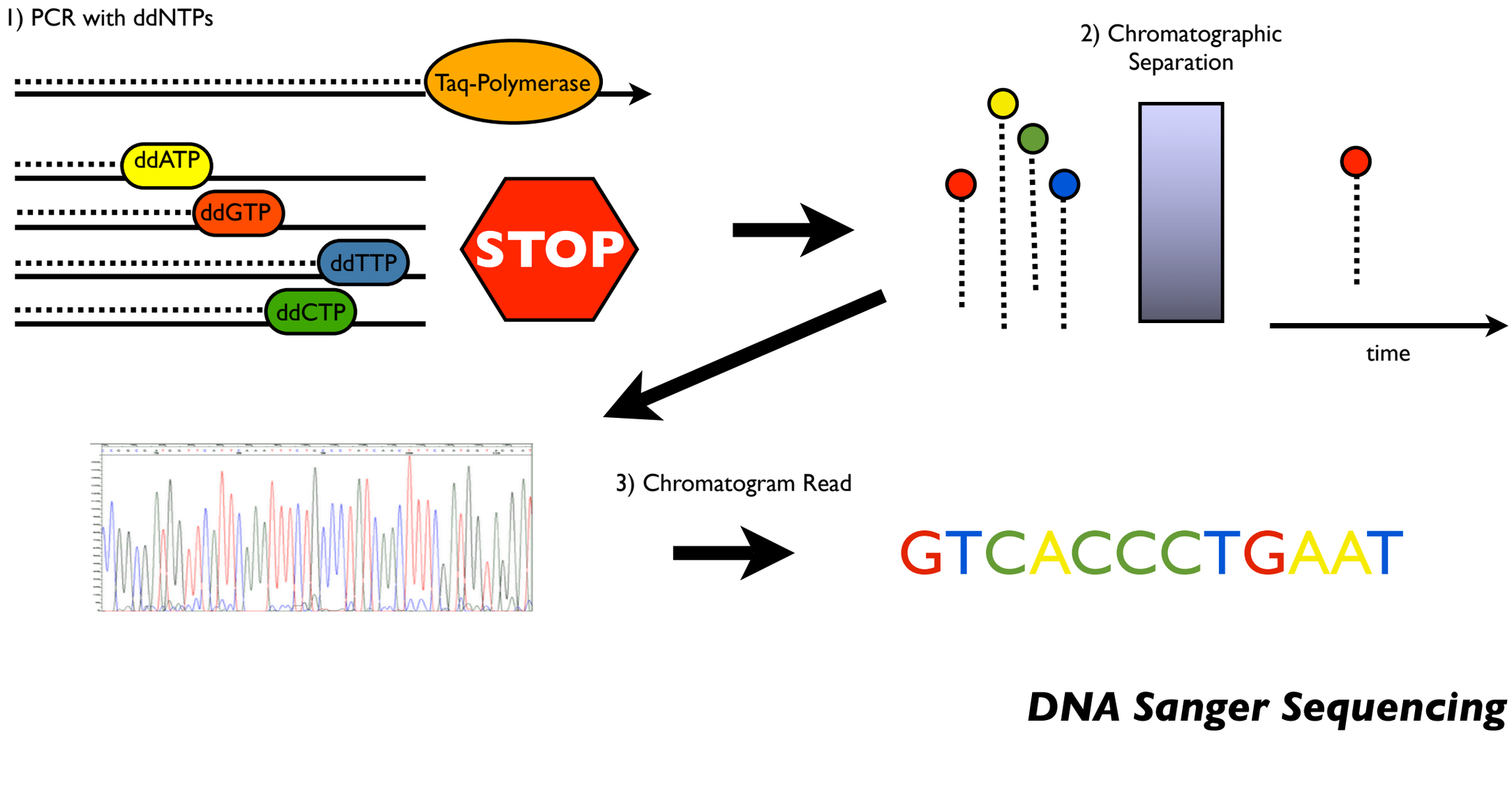
??? → DNA sequencing

1. A brief history of the sequence assembly.

DNA Sequencing:

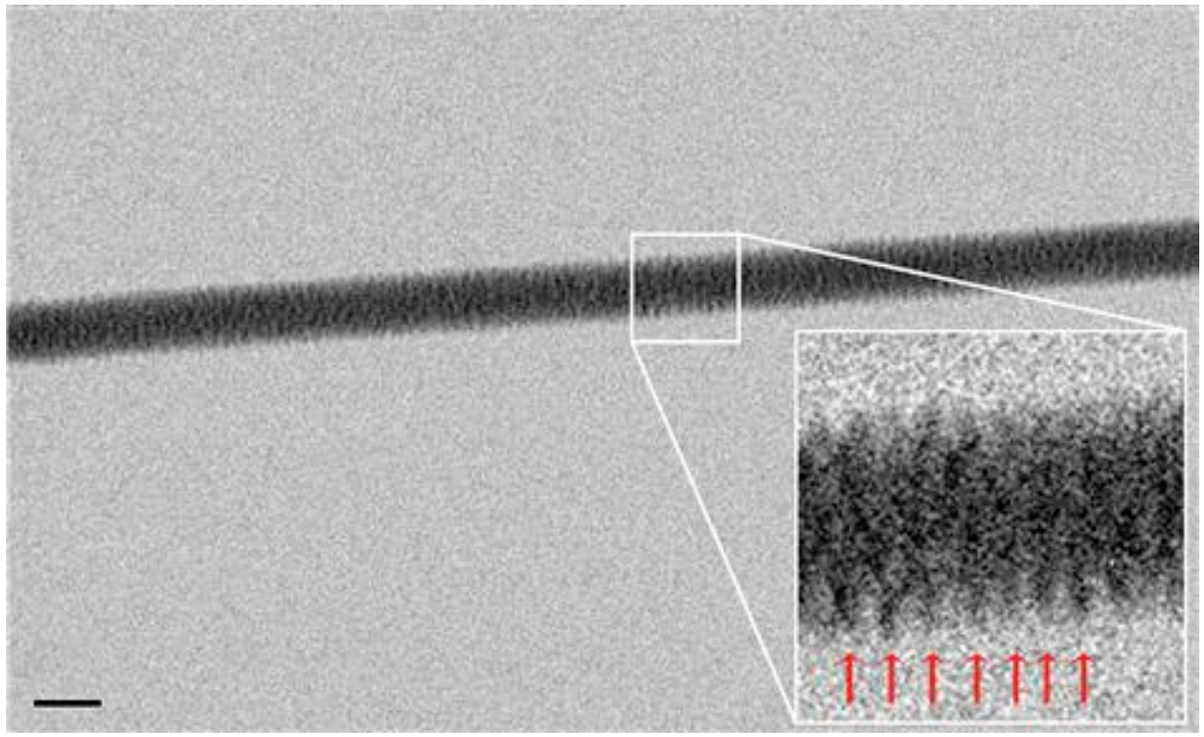
"Process of determining the precise order of nucleotides within a DNA molecule."

-Wikipedia



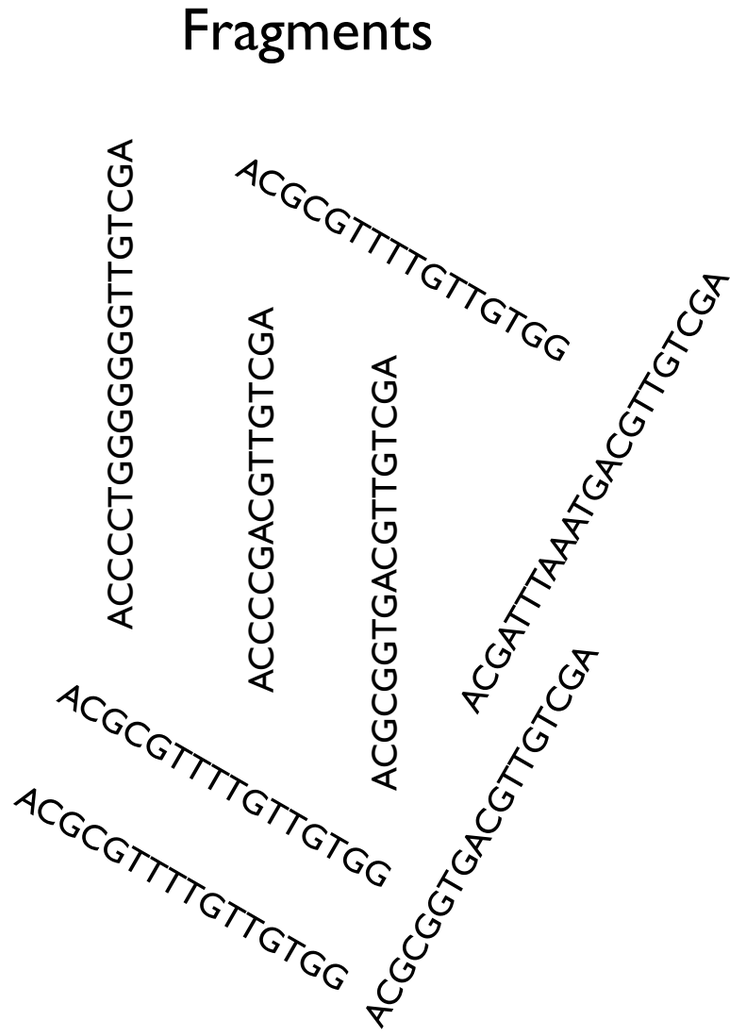
DNA Sanger Sequencing

1. A brief history of the sequence assembly.



Gentile F. et al. *Direct Imaging of DNA Fibers: The Visage of Double Helix*
Nano Lett., 2012, 12 (12), pp 6453–6458

DNA sequencing



Sequence assembly

ACGCGTTTTGTTGTGGTGGCCACACCACGCAGTGACGGAGATAACGGCGAGAGCATGGACGGAGGATGAGGATGG

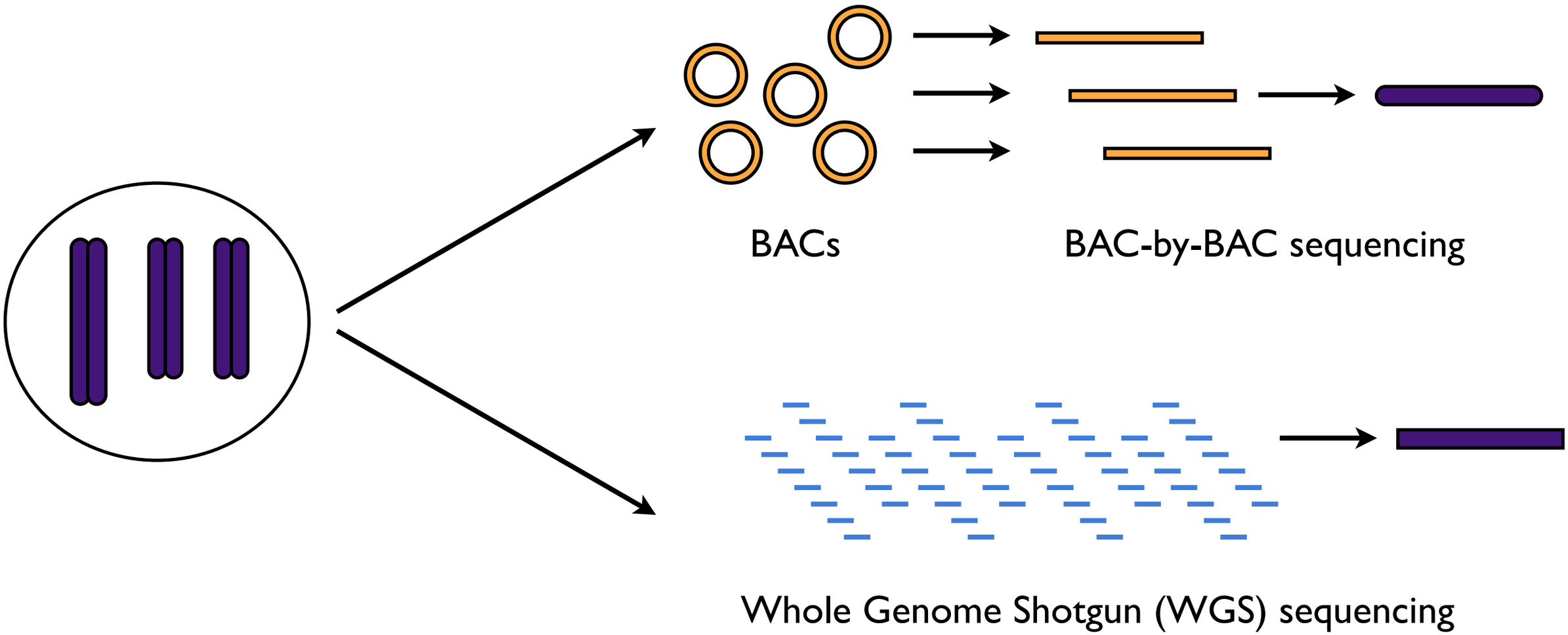
Sequence assembly

=

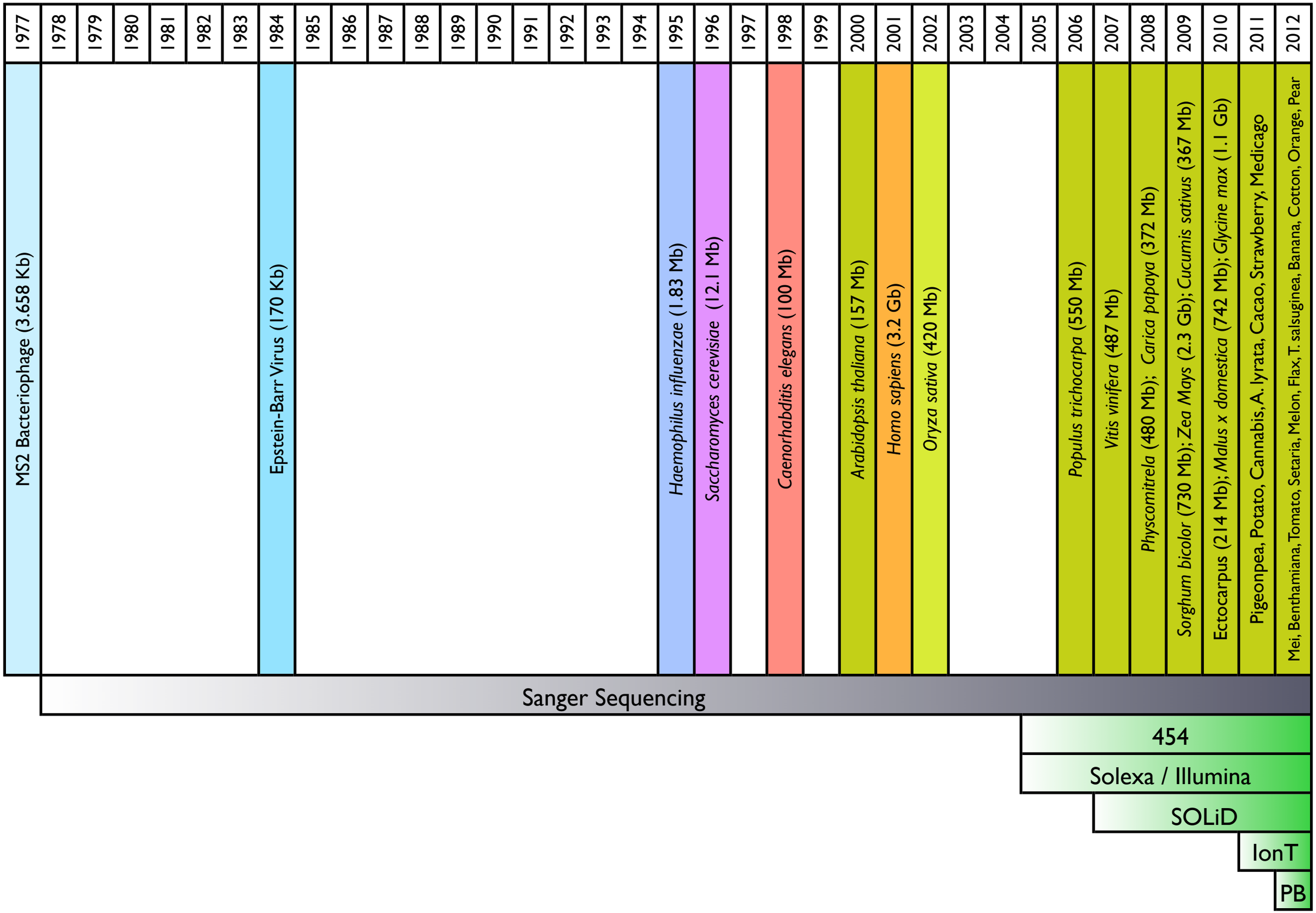
Resolve a puzzle and rebuild a DNA sequence from its pieces (fragments)

Image courtesy of iStock photo

Resolve a puzzle and rebuild a DNA sequence from its pieces (fragments)



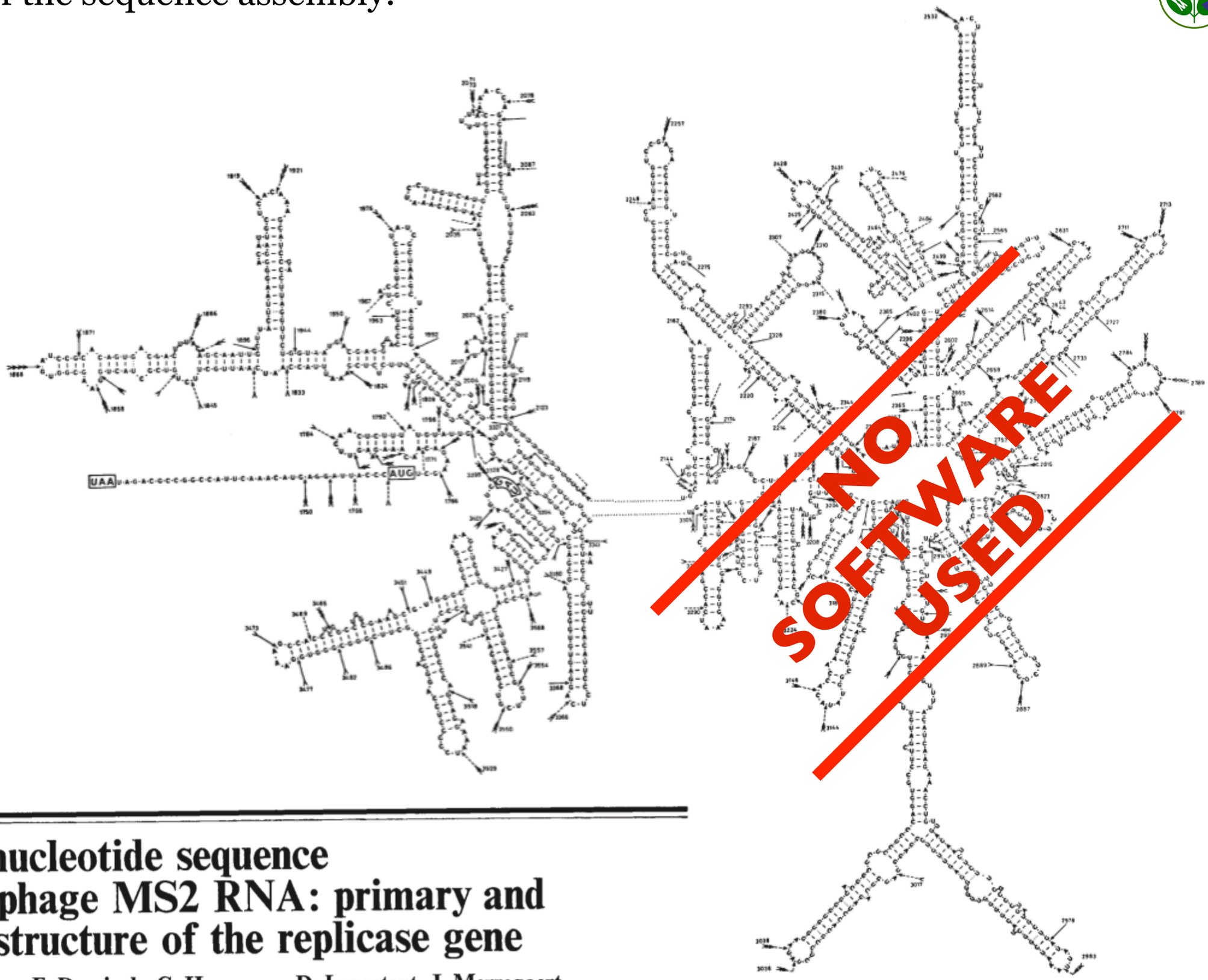
1. A brief history of the sequence assembly.



1. A brief history of the sequence assembly.

1977

MS2 Bacteriophage (3.658 Kb)



Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene

W. Fiers, R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Min Jou, F. Molemans, A. Raeymaekers, A. Van den Berghe, G. Volckaert & M. Ysebaert

Laboratory of Molecular Biology, University of Ghent, 9000 Ghent, Belgium

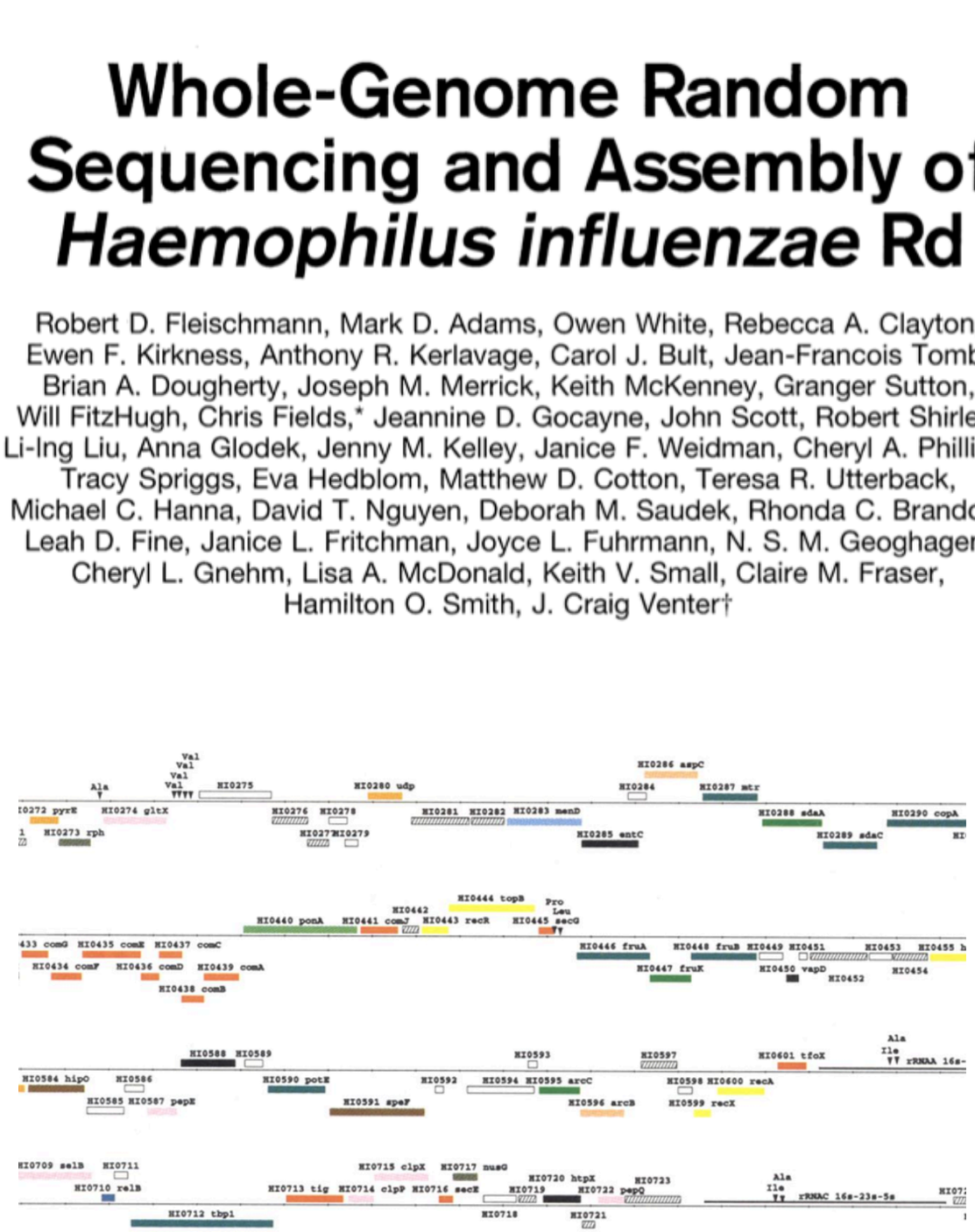
1. A brief history of the sequence assembly.

1995

Haemophilus influenzae (1.83 Mb)

Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd

Robert D. Fleischmann, Mark D. Adams, Owen White, Rebecca A. Clayton, Ewen F. Kirkness, Anthony R. Kerlavage, Carol J. Bult, Jean-Francois Tomb, Brian A. Dougherty, Joseph M. Merrick, Keith McKenney, Granger Sutton, Will FitzHugh, Chris Fields,* Jeannine D. Gocayne, John Scott, Robert Shirley, Li-Ing Liu, Anna Glodek, Jenny M. Kelley, Janice F. Weidman, Cheryl A. Phillips, Tracy Spriggs, Eva Hedblom, Matthew D. Cotton, Teresa R. Utterback, Michael C. Hanna, David T. Nguyen, Deborah M. Saudek, Rhonda C. Brandon, Leah D. Fine, Janice L. Fritchman, Joyce L. Fuhrmann, N. S. M. Geoghagen, Cheryl L. Gnehm, Lisa A. McDonald, Keith V. Small, Claire M. Fraser, Hamilton O. Smith, J. Craig Venter†

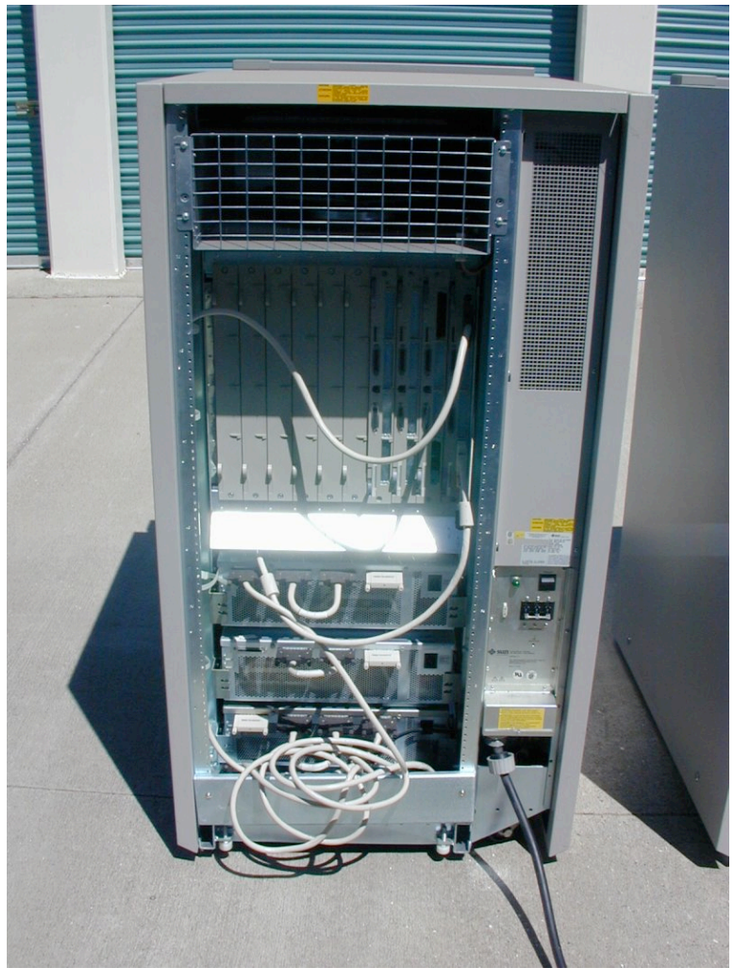


Software:

TIGR ASSEMBLER

Hardware:

SPARCenter 2000 (512 Mb RAM)



1. A brief history of the sequence assembly.

1995

Haemophilus influenzae (1.83 Mb)

Software:

TIGR ASSEMBLER → Smith-Waterman alignments

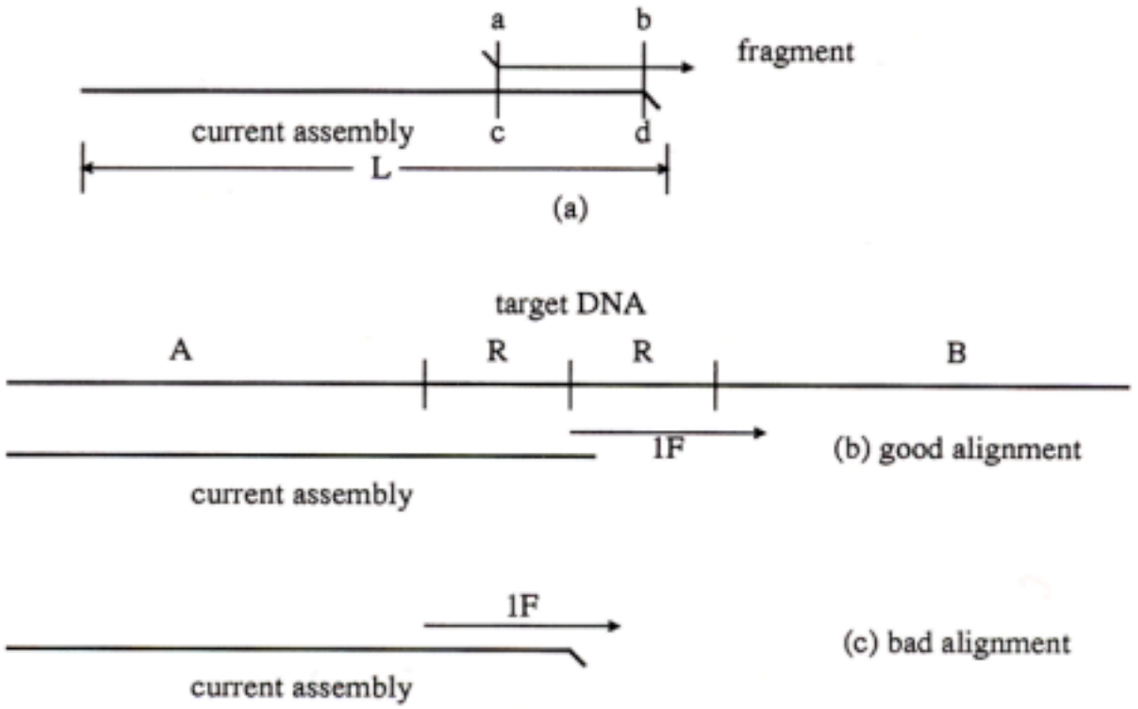
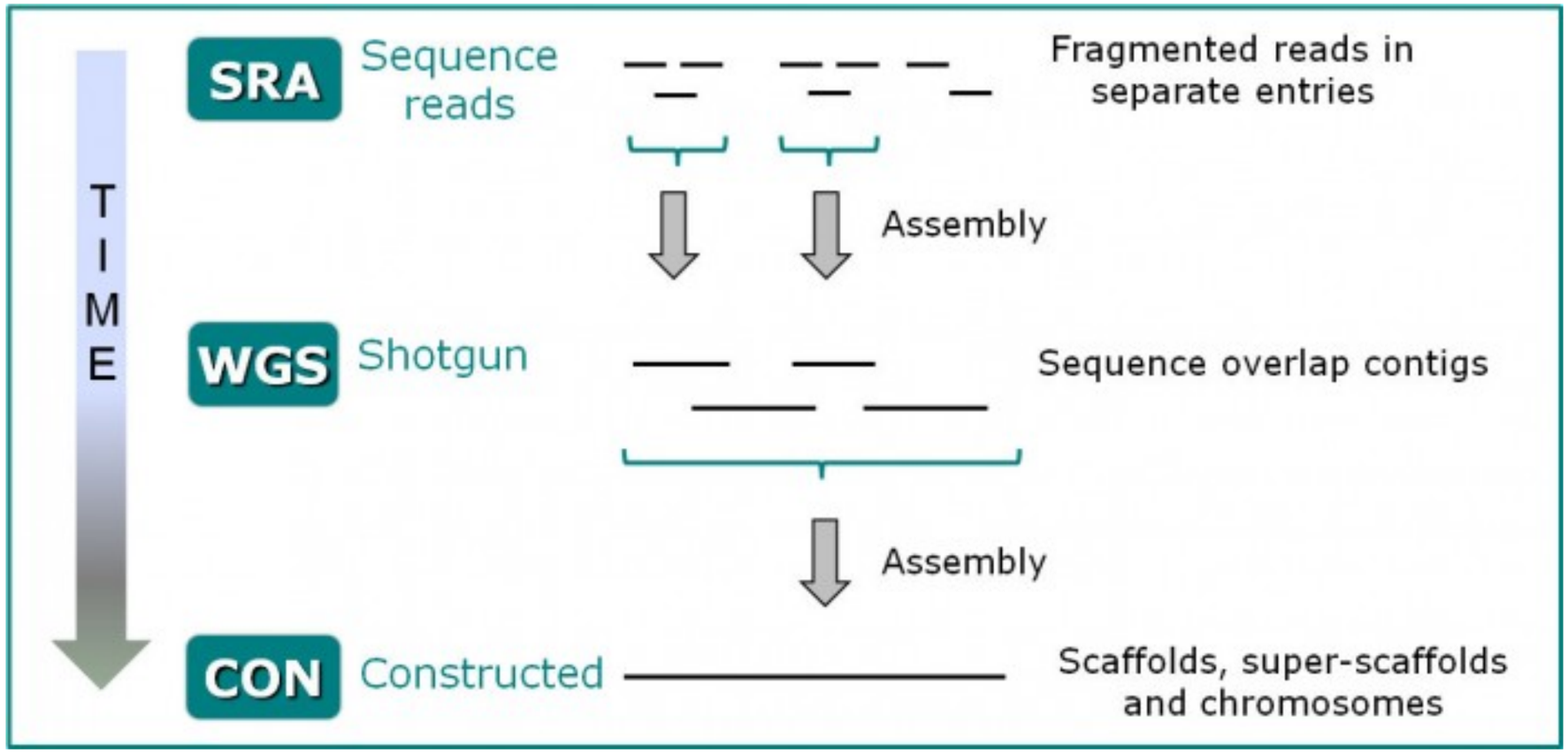


FIG. 1. **a.** A graphic representation of a Smith-Waterman alignment is shown. The overlap is shown between vertical bars, and the diagonal lines represent overhangs where the sequences do not match. The current assembly has a length of L. **b.** A tandem repeat of two copies of repeat region R is shown along with the proper alignment of fragment 1F with the current assembly. **c.** A bad alignment of the current assembly and fragment 1F is produced if the overlap is maximized without regard to the length of the overhang. The bad alignment can result in two outcomes. If the overhang is short, it will be ignored, and the two repeat regions will be compressed into a single region. If the overhang is long, the merge will not be allowed, and the current assembly will not be extended.

OVERLAPPING METHODOLOGY

Sutton GG. et al. TIGR Assembler: A New Tool to Assembly Large Shotgun Sequencing Projects Genome Science and Technology, 1995,1:9-19

1. A brief history of the sequence assembly.



http://www.ebi.ac.uk/training/online/course/nucleotide-sequence-data-resources-ebi/what-ena/how-sequence-assembled

1. A brief history of the sequence assembly.

2001

Homo sapiens (3.2 Gb)

articles

BAC-by-BAC sequencing

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

** A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.*

International Human Genome Sequencing Consortium. Initial Sequencing and Analysis of the Human Genome. Nature. 2001. 409:860-921

OVERLAPPING METHODOLOGY

Whole Genome Shotgun (WGS) sequencing

The Sequence of the Human Genome

J. Craig Venter,[1]* Mark D. Adams,[1] Eugene W. Myers,[1] Peter W. Li,[1] Richard J. Mural,[1] Granger G. Sutton,[1] Hamilton O. Smith,[1] Mark Yandell,[1] Cheryl A. Evans,[1] Robert A. Holt,[1]

Venter JC. et al. The Sequence of the Human Genome. Science. 2001. 291:1304-1351

1. A brief history of the sequence assembly.

2001

Homo sapiens (3.2 Gb)

Whole Genome Shotgun (WGS) sequencing

The Sequence of the Human Genome

J. Craig Venter,[1]* Mark D. Adams,[1] Eugene W. Myers,[1] Peter W. Li,[1] Richard J. Mural,[1]
Granger G. Sutton,[1] Hamilton O. Smith,[1] Mark Yandell,[1] Cheryl A. Evans,[1] Robert A. Holt,[1]

Venter JC. et al. The Sequence of the Human Genome. Science. 2001. 291:1304-1351

Software:

WGA ASSEMBLER (CABOG)

Hardware:

40 machines AlphaSMPs (4 Gb RAM/each and 4 cores/each, total=160 Gb RAM and 160 cores); 5 days.

1. A brief history of the sequence assembly.

2009

Ailuropoda melanoleura (2.3 Gb)

ARTICLES

Whole Genome Shotgun (WGS) sequencing

The sequence and *de novo* assembly of the giant panda genome

Ruiqiang Li[1,2]*, Wei Fan[1]*, Geng Tian[1,3]*, Hongmei Zhu[1]*, Lin He[4,5]*, Jing Cai[3,6]*, Quanfei Huang[1], Qingle Cai[1,7], Bo Li[1], Yinqi Bai[1], Zhihe Zhang[8], Yaping Zhang[6], Wen Wang[6], Jun Li[1], Fuwen Wei[9], Heng Li[10], Min Jian[1], Jianwen Li[1],

Li R. et al. The Sequence and the Novo Assembly of the Giant Panda Genome. Nature. 2009. 463:311-317

Software:

SOAPdenovo

Hardware:

Supercomputer with 32 cores and 512 Gb RAM.

BRUIJN GRAPHS METHODOLOGY

What is a Kmer ?

Specific n-tuple or n-gram of nucleic acid or amino acid sequences.

-Wikipedia

ordered list
of elements

contiguous sequence
of *n* items from a given
sequence of text

ATGCGCAGTGGAGAGAGAGCGATG Sequence A with 25 nt

↓ 5 Kmers of 20-mer

ATGCGCAGTGGAGAGAGAGC
TGCGCAGTGGAGAGAGAGCG
GCGCAGTGGAGAGAGAGCGA
CGCAGTGGAGAGAGAGCGAT
GCAGTGGAGAGAGAGCGATG

N_kmers = L_read - Kmer_size

1. A brief history of the sequence assembly.

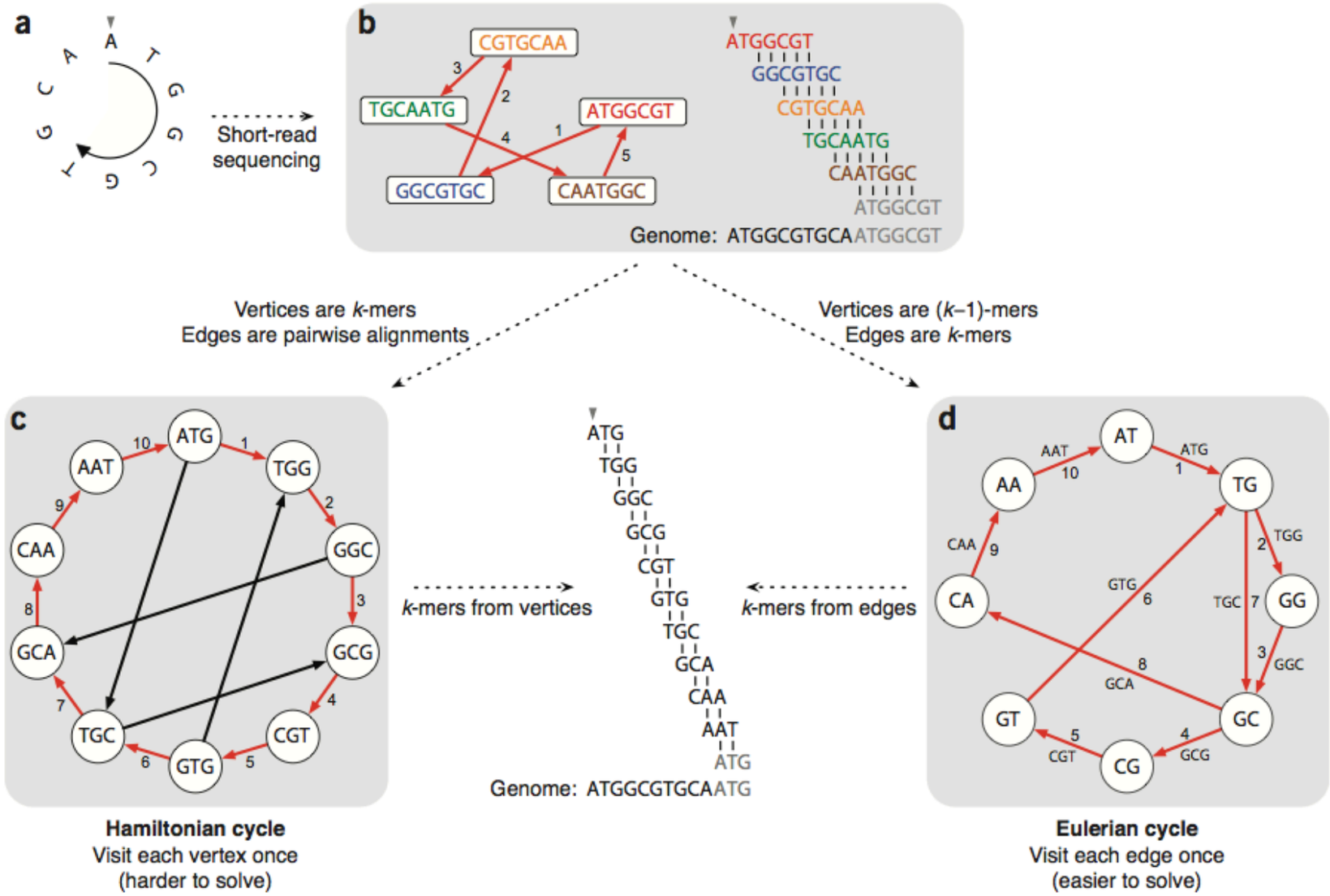


Figure 3 Two strategies for genome assembly: from Hamiltonian cycles to Eulerian cycles. (**a**) An example small circular genome. (**b**) In traditional Sanger sequencing algorithms, reads were represented as nodes in a graph, and edges represented alignments between reads. Walking along a Hamiltonian cycle by following the edges in numerical order allows one to reconstruct the circular genome by combining alignments between successive reads. At the end of the cycle, the sequence wraps around to the start of the genome. The repeated part of the sequence is grayed out in the alignment diagram. (**c**) An alternative assembly technique first splits reads into all possible *k*-mers: with $k = 3$, ATGGCGT comprises ATG, TGG, GGC, GCG and CGT. Following a Hamiltonian cycle (indicated by red edges) allows one to reconstruct the genome by forming an alignment in which each successive *k*-mer (from successive nodes) is shifted by one position. This procedure recovers the genome but does not scale well to large graphs. (**d**) Modern short-read assembly algorithms construct a de Bruijn graph by representing all *k*-mer prefixes and suffixes as nodes and then drawing edges that represent *k*-mers having a particular prefix and suffix. For example, the *k*-mer edge ATG has prefix AT and suffix TG. Finding an Eulerian cycle allows one to reconstruct the genome by forming an alignment in which each successive *k*-mer (from successive edges) is shifted by one position. This generates the same cyclic genome sequence without performing the computationally expensive task of finding a Hamiltonian cycle.

Compeau PEC. et al. How to apply de Bruijn graphs to genome assembly. Nature Biotech. 2011. 29:287-291

1. A brief history of the sequence assembly.

BRIEFINGS IN FUNCTIONAL GENOMICS. VOL 11. NO 1. 25–37 doi:10.1093/bfgp/elr035

Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph

Zhenyu Li, Yanxiang Chen*, Desheng Mu*, Jianying Yuan, Yujian Shi, Hao Zhang, Jun Gan, Nan Li, Xuesong Hu, Binghang Liu, Bicheng Yang and Wei Fan*

Advance Access publication date 19 December 2011

OLC (Overlap-layout-consensus) algorithm is more suitable for the low-coverage long reads, whereas the DBG (De-Bruijn-Graph) algorithm is more suitable for high-coverage short reads and especially for large genome assembly

Key Points

- High-quality genome sequences for many species are still strongly desired by the genomics community. With the rapid development of sequencing technologies and assembly algorithms, we have seen practical improvements and a bright future lies ahead.
- There are two major types of assembly algorithms: OLC and DBG; both of them are in accordance with Lander–Waterman model, but suit the assembly of different read lengths and sequencing depths, and have significant differences in computational efficiency.
- How well a genome can be assembled depends not only on sequencing technologies such as read length and sequencing error rate, but also on the characteristics of the genome, including repeat and the heterozygosity rate of the sequenced sample.

1. A brief history of the sequence assembly.

2. Sequencing, tools and computers.

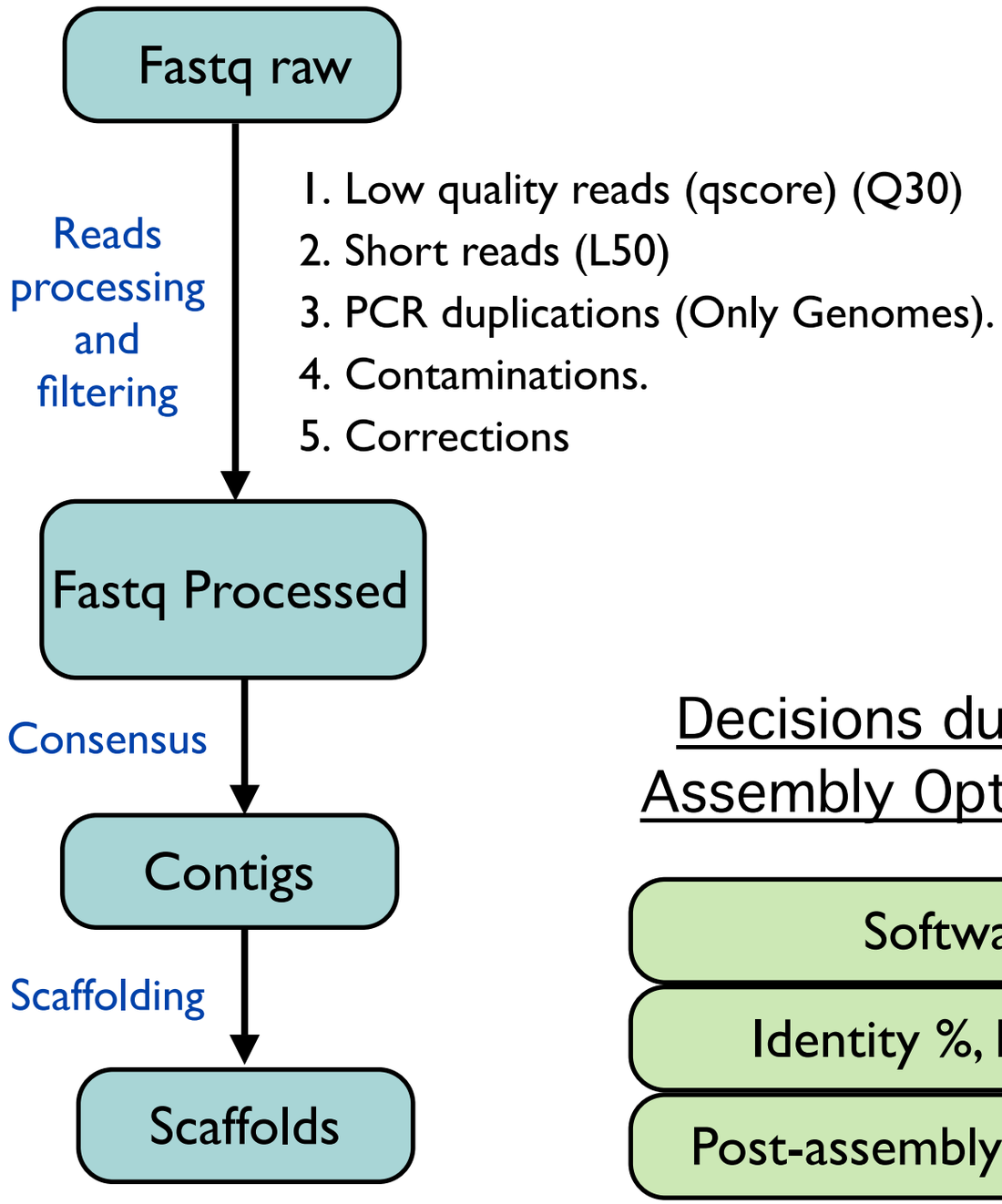
3. Things that you should know about genomes.

4. What about transcriptomes ? Differences

2.0 Overview of a Sequencing Project: Assembly

Decisions during the Experimental Design

- Technology
- Library Preparation
- Sequencing Amount



Decisions during the Assembly Optimization

- Software
- Identity %, Kmer ...
- Post-assembly Filtering

2. Sequencing, tools and computers.

2.1 Technologies

	Run Time	Sequence Length	Reads/Run	Total nucleotides sequenced per run
Capillary Sequencing (ABI37000)	~2.5 h	800 bp	386	0.308 Mb
454 Pyrosequencing (GS FLX Titanium XL+)	~23 h	700 bp	1,000,000	700 Mb (0.7 Gb)
Illumina (HiSeq 2500)	264 h / 27 h (11 days)	2 x 100 bp 2 x 150 bp	2 x 3,000,000,000 2 x 600,000,000	600,000 / 120,000 Mb (600 / 120 Gb)
Illumina (MiSeq)	39 h	2 x 250 bp	2 x 17,000,000	8,500 Mb (8.5 Gb)
SOLID (5500xl system)	48 h (2 days)	75 bp	400,000,000	30,000 Mb (30 Gb)
Ion Torrent (Ion Proton I)	2 h	100 bp	100,000,000	10,000 Mb (10 Gb)
PacBio (PacBioRS)	1.5 h	~3,000 bp	25,000	100 Mb (0.1 Gb)

2. Sequencing, tools and computers.

2.1 Technologies

	Strenghs	Weaknesses
454 Pyrosequencing (GS FLX Titanium XL+)	– Long reads (450/700 bp). – Long insert for mate pair libraries (20Kb). – Low observed raw error rate (0.1%) – Low percentage of PCR duplications for mate pair libraries	– Homopolymer error. – Low sequence yield per run (0.7 Gb). – Preferred assembler (gsAssembler) uses overlapping methodology.
Illumina (HiSeq 2500)	– High sequence yield per run (600 Gb) – Low observed raw error rate (0.26%)	– High percentage of PCR duplications for mate pair libraries. – Long run time (11 days) – High instrument cost (~ $650K)
Illumina (MiSeq)	– Medium read size (250 bp) – Faster run than Illumina HiSeq	– Medium sequence yield per run (8.5 Gb)
SOLID (5500xl system)	– 2-base encoding reduce the observed raw error rate (0.06%)	– 2-base color coding makes difficult the sequence manipulation and assembly. – Short reads (75 bp)
Ion Torrent (Ion Proton I)	– Fast run (2 hours) – Low instrument cost ($80K). – Medium read size (200 bp)	– Medium sequence yield per run (10 Gb) – Medium observed raw error rate (1.7%)
PacBio (PacBioRS)	– Long reads (3000 bp) – Fast run (2 hours)	– Really high observed raw error rate (12.7%) – High instrument cost (~ $700K) – No pair end/mate pair reads

2.2 Libraries

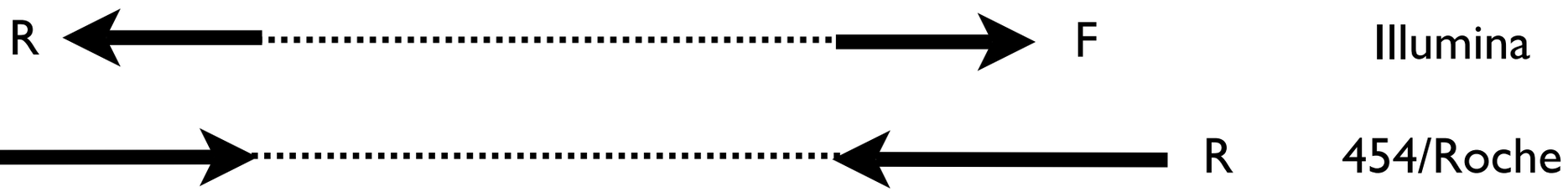
★ Library types (orientations):

- Single reads

F ⟶

- Pair ends (PE) (150-800 bp insert size)

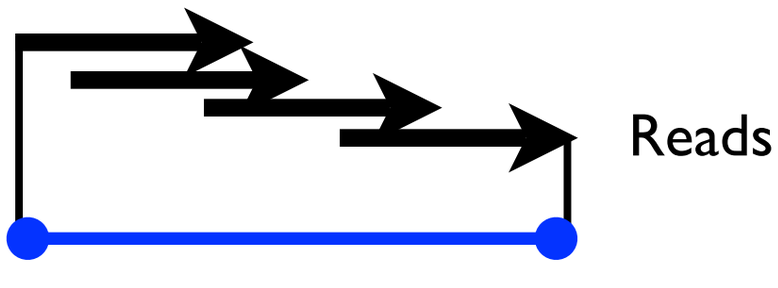
- Mate pairs (MP) (2-40 Kb insert size)

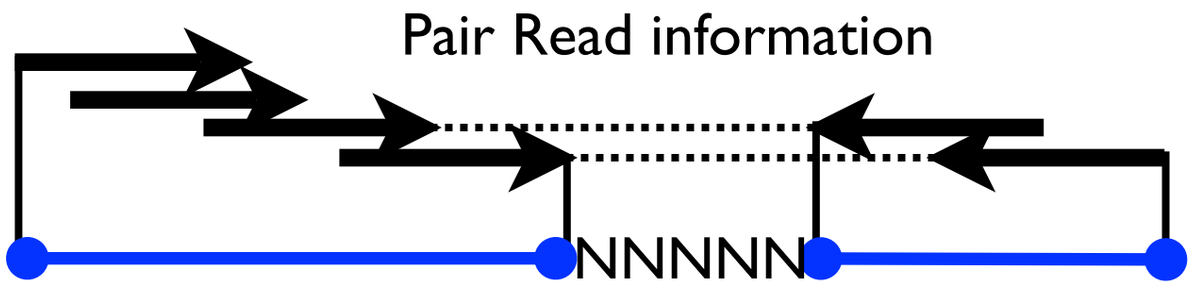


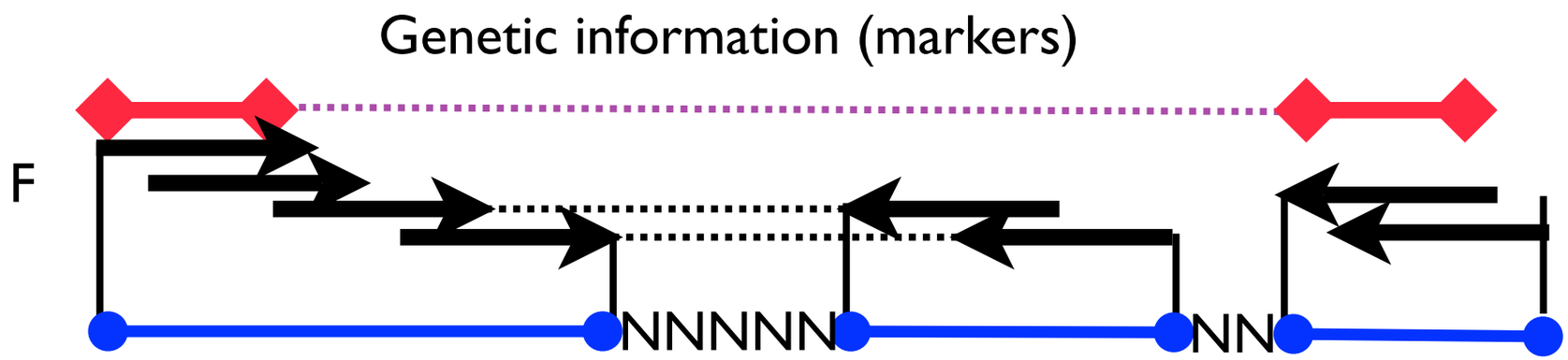
2.2 Libraries

- Why is important the pair information ?

  - *novo* assembly:







2.3 Sequencing Amount

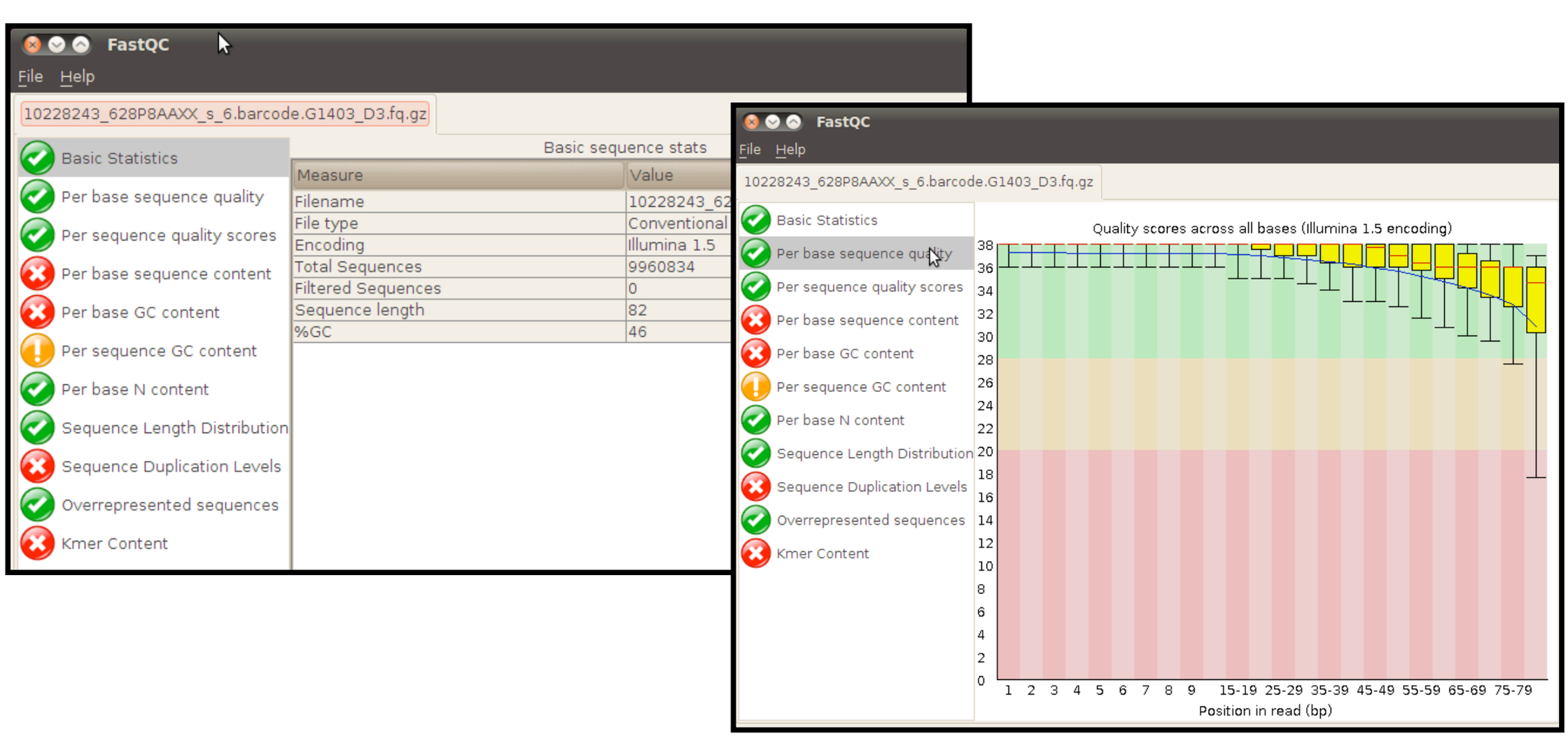
Depending of the genome complexity, technology used and assembler:

- More is better (if you have enough computational resources).
  - Sanger > 10X (less for BACs-by-BACs approaches).
  - 454 > 20X
  - Illumina > 100X
- Polyploidy or high heterogozyty increase the amount of reads needed.
- The use of different library types (pair ends and mate pairs with different insert sizes is essential).
- Longer reads is preferable.

2.4 Tools: Read Quality Evaluation

a) Length of the read.
b) Bases with qscore > 20 or 30.

- FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)



2.4 Tools: Read Trimming and Filtering

1. Adaptor removal.
2. Low quality reads (qscore) (Q30)
3. Short reads (L50)
4. PCR duplications (Only Genomes, Use PrinSeq).

- Fastx-Toolkit (http://hannonlab.cshl.edu)
- Ea-Utils (http://code.google.com/p/ea-utils/)
- PrinSeq (http://edwards.sdsu.edu/cgi-bin/prinseq/prinseq.cgi)

Software	Multiplexing	Trimming/Filtering
Fastx-Toolkit	fastx_barcode_splitter	fastq_quality_filter
Ea-Utils	fastq-multx	fastq-mcf
PrinSeq	PrinSeq	PrinSeq

2.4 Tools: Contaminations

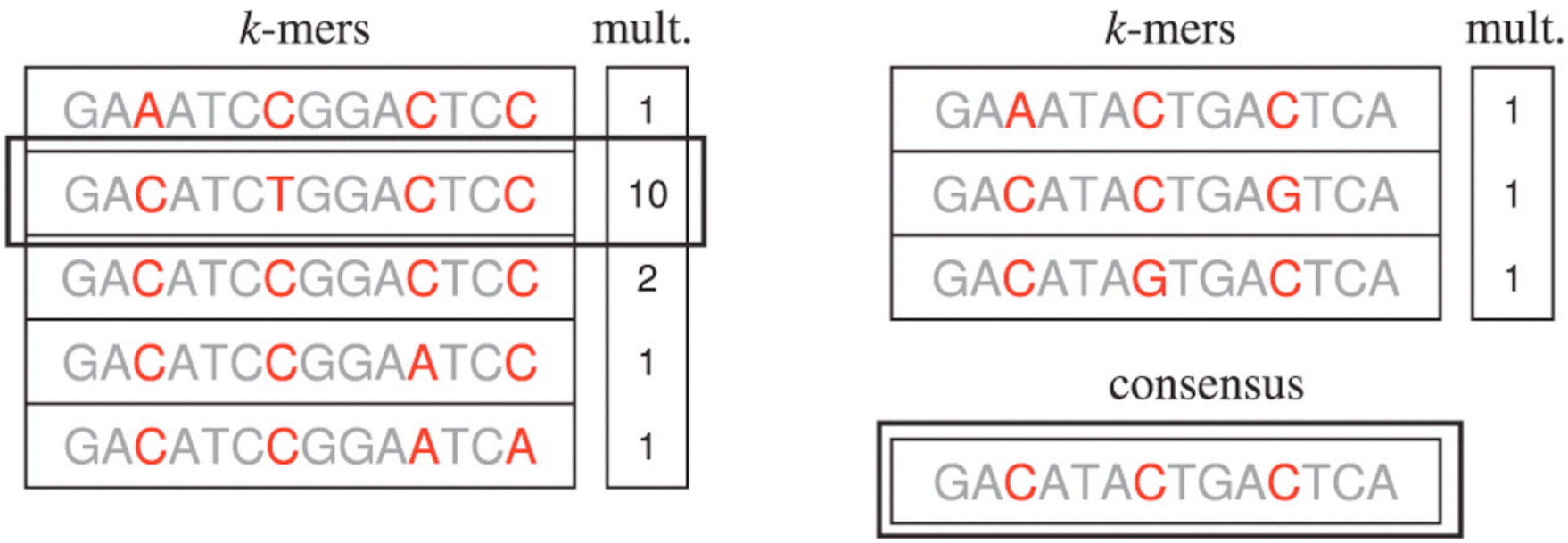
5. Contaminations

Contaminations can be removed mapping the reads against a reference with the contaminants such as E. coli and human genomes. The most common tools are Bowtie or BWA (for short reads) and Blast (for long reads).

2.4 Tools: Read Corrections

6. Read Corrections

Read corrections are generally based in the Kmer analysis.



Medvedev P. et al. Error correction of high-throughput sequencing datasets with non-uniform coverage Bioinformatics. 2011 27 (13):i137-i141

2.4 Tools: Read Corrections

6. Read Corrections

American Journal of Bioinformatics Research
p-ISSN: 2167-6992 e-ISSN: 2167-6976
2013; 3(1): 1-9
doi:10.5923/j.bioinformatics.20130301.01

Review of Genome Sequence Short Read Error Correction Algorithms

M. Tahir[1], M. Sardaraz[1], Ataul Aziz Ikram[1], Hassan Bajwa[2]

Usual Suspects:

- Quake (http://www.cbcb.umd.edu/software/quake/index.html)
- Reptile (http://aluru-sun.ece.iastate.edu/doku.php?id=software)
- ECHO (http://uc-echo.sourceforge.net/)
- Corrector (http://soap.genomics.org.cn/soapdenovo.html)

2.4 Tools: Assemblers

	Type	Technology Used	Features	Link
Arachne	Overlap-layout-consensus	Sanger, 454	Highly configurable	http://www.broadinstitute.org/crd/wiki/index.php/Main_Page
CABOG	Overlap-layout-consensus	Sanger, 454, Illumina	Highly configurable	http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=Main_Page
MIRA	Overlap-layout-consensus	Sanger, 454	Highly configurable	http://sourceforge.net/apps/mediawiki/mira-assembler
gsAssembler	Overlap-layout-consensus	Sanger, 454	Easy to use	http://454.com/products/analysis-software/index.asp
iAssembler	Overlap-layout-consensus	Sanger, 454	Improves MIRA	http://bioinfo.bti.cornell.edu/tool/iAssembler
ABySS	Bruijn graph	454 or Illumina	Easy to use	http://www.bcgsc.ca/platform/bioinfo/software/abyss
ALLPATH-LG	Bruijn graph	454 or Illumina	Good results	http://www.broadinstitute.org/software/allpaths-lg/blog
Ray	Bruijn graph	454 or Illumina	Slow but use less memory	http://denovoassembler.sf.net/
SOAPdenovo	Bruijn graph	454 or Illumina	Fastest	http://soap.genomics.org.cn/soapdenovo.html
Velvet	Bruijn graph	454 or Illumina or SOLiD	SOLiD	http://www.ebi.ac.uk/~zerbino/velvet/

2.4 Tools: Assemblers

... but there are more assemblers and information... Take a look to SeqAnswers

http://seqanswers.com/wiki/Special:BrowseData/Bioinformatics_application?Bioinformatics_method=Assembly&Biological_domain=De-novo_assembly

Also highly recommendable:

Assemblathon 1: A competitive assessment of de novo short read assembly methods

Dent Earl, Keith Bradnam, John St. John, et al.

Genome Res. 2011 21: 2224-2241 originally published online September 16, 2011
Access the most recent version at doi:10.1101/gr.126599.111

GAGE: A critical evaluation of genome assemblies and assembly algorithms

Steven L. Salzberg, Adam M. Phillippy, Aleksey Zimin, et al.

Genome Res. 2012 22: 557-567 originally published online December 6, 2011
Access the most recent version at doi:10.1101/gr.131383.111

2.5 Assembly evaluation

During the assembly optimization will be generated several assemblies. The most used parameters to evaluate the assembly are:

1. Total Assembly Size,

How far is this value from the estimated genome size

2. Total Number of Sequences (Scaffold/Contigs)

How far is this value from the number of chromosomes.

3. Longest scaffold/contig

4. Average scaffold/contig size

5. N50/L50 (or any other N/L)

Number sequence (N) and minimum size of them (L) that represents the 50% of the assembly if the sequences are sorted by size, from bigger to smaller.

2.5 Assembly evaluation

N50/L50

SIZE	ADDITIVE SIZE
240000	240000
200000	440000
120000	560000
80000	640000
20000	660000
10000	670000
4000	674000
3000	677000
2000	679000
1000	680000
1000	681000
100	681100
100	681200
100	681300
100	681400
100	681500

N50 = 2; L50 = 440,000

N90 = 4; L90 = 640,000

90% → 613350

50% → 340740

The 90% of the assembly is represented for 4 sequences with a minimum size of 640,000

2.6 Computers

Bigger is better:

- How much do you need depends of:
  - ➡ how big is your genome ?
    ~ Human size (3Gb) require ~256 Gb to 1 Tb
  - ➡ how many reads do you have, ?
    More reads, more memory and hard disk.
  - ➡ what software are you going to use ?
    OLC uses more memory and time than DBG.
  - ➡ what parameters are you going to use ?
    Bigger Kmers, more memory.

Four options:

1. Buy your own server (512 Gb, 4.5 Tb, 64 cores; ~ $15,000).
2. Rent a server for ~ 1 month (same specs. $1.5/h; ~$1,000).
3. Use a supercomputing center associated with NSF, NIH, USDA... where they offer reduced prices (iPlant, Indiana University....).
4. Collaborate with some group with a big server.

1. A brief history of the sequence assembly.

2. Sequencing, tools and computers.

3. Things that you should know about genomes.

4. What about transcriptomes ? Differences

3. Things that you should know about genomes.

1. They have variable size, for example in angiosperm plants they range from 63 Mb (*Genlisea margaretae*) to 150 Gb (*Paris japonica*) with an average of 5.6 Gb. More data at:
   - ✴ Plants: http://data.kew.org/cvalues/
   - ✴ Animals: http://www.genomesize.com/

2. They can be polyploids. It means that homoeologus regions with highly similar will collapse during the assembly.

3. They can be highly heterozygous and polymorphic. In this case some of the alleles will collapse, some of them not. The effective coverage will lower than expected.

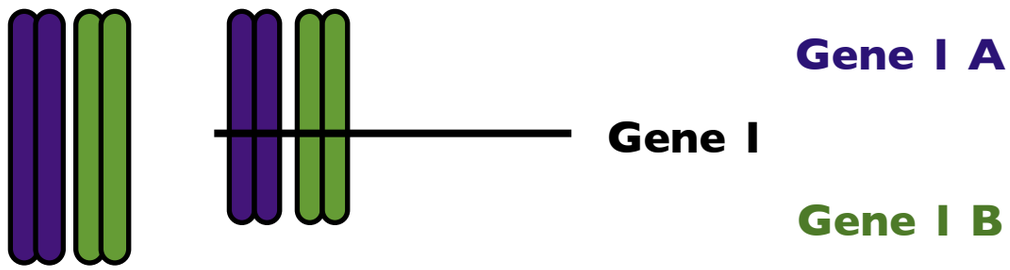
4. They can have recent whole genome duplication (or triplication) events. Paralogous genes may colapse during the assembly.

5. Repeats, most of the genomes are full of repeats and they are difficult to resolve. By default assemblers throw them away based in the kmer histogram.

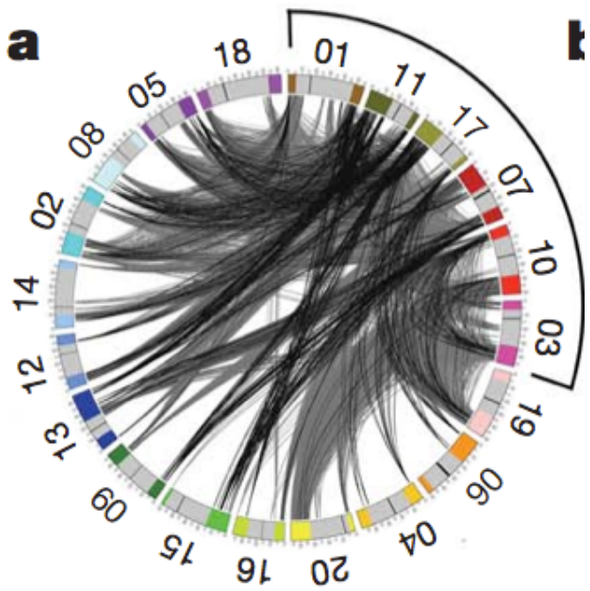
3. Things that you should know about genomes.

3.1 Collapsing problem

✳Polyploidy



✳Whole Genome Duplication



Schmutz J et al. Genome Sequence of the Paleoploid Soybean. Nature 2010 463:178-183

CACTTGACGACATGACG Gene I A

CTTGACGACATGACGAC

CCCTTGACGACATGACG Gene I B

CGCCCTTGACGACATGA

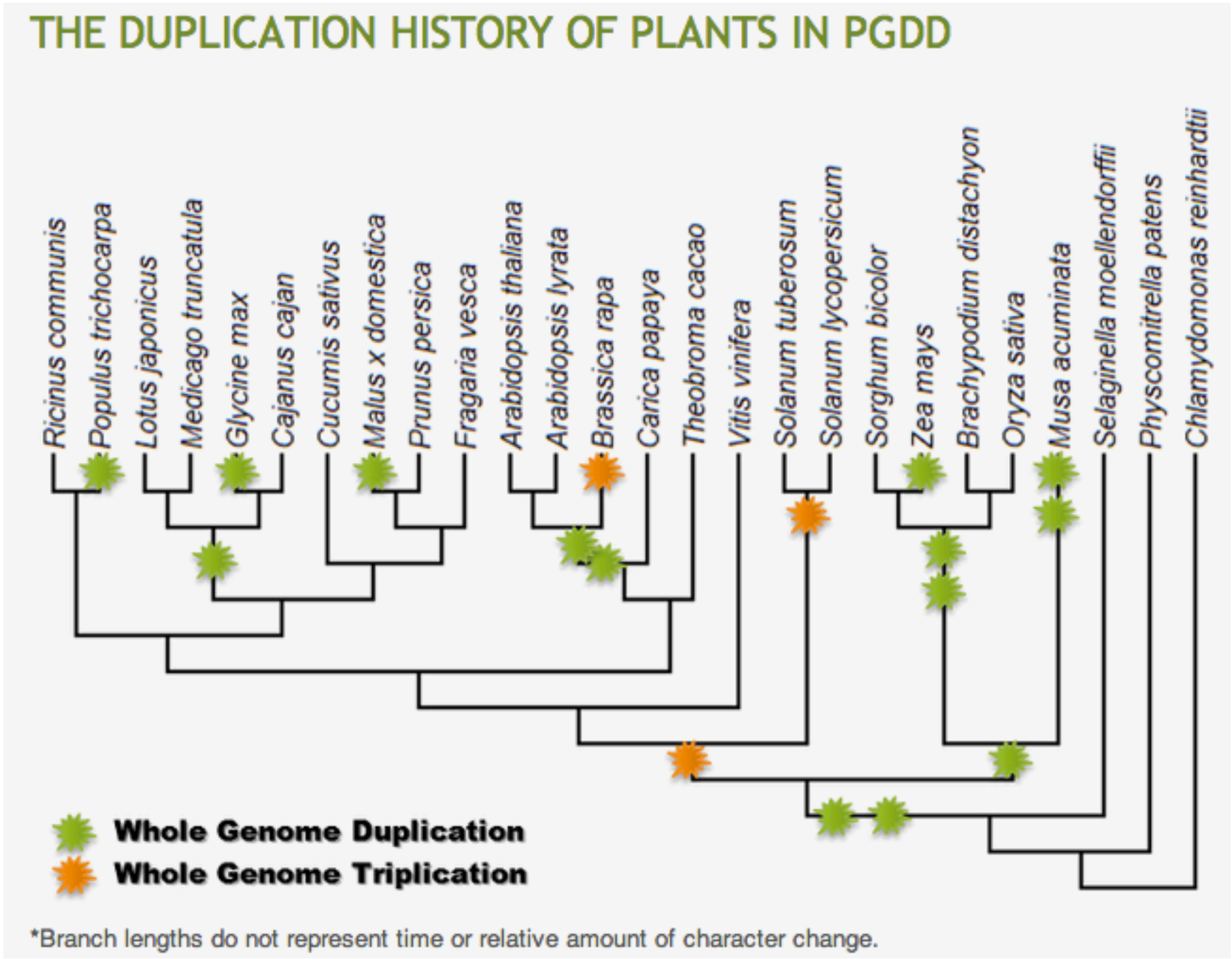
CGCCCTTGACGACATGACGACA

Collapsed consensus Gene I A + Gene I B

3. Things that you should know about genomes.

3.1 Collapsing problem

✳Whole Genome Duplication



http://chibba.agtec.uga.edu/duplication/index/home

3. Things that you should know about genomes.

3.1 Collapsing evaluation

Read collapsing during the assembly process can be evaluated mapping reads back to the consensus sequence and analyzing SNPs.

CGCCCTTGACGACATGACGACA **Consensus**

CACTTGACGACATGA
CACTTGACGACATGACG
CTTGACGACATGACGAC **Reads**
CCCTTGACGACATGACG
CGCCCTTGACGACATGA

↓

SNPs

=

Heterozygous Positions + Homoeologs Collapsing + Paralogs Collapsing

3. Things that you should know about genomes.

3.1 Collapsing solutions

1. Sequence the two progenitors and use them as a reference.

✓ Approach used previously (cotton genome; Patterson A. et. al 2012).

❖ Information not contained in the references will be lost.

❖ Progenitors are not always available.

2. Use single molecule technologies such as PacBio or Moleculo

✓ Best approach because it will give the right phase for long sequence chunks. It exists software to integrate big sequences (minimus2...)

❖ Experimental and probably expensive.

3. Assembly everything with the higher Kmer, evaluate the collapsed regions and rescue the haplotype/region using SNPs and pair end read information.

✓ Cheap, easy to apply for the current technologies.

❖ No software available

3. Things that you should know about genomes.

3.2 Repeats and assemblers

5. Repeats, most of the genomes are full of repeats and they are difficult to resolve. By default **some assemblers throw them away based in the kmer histogram**. For example, SOAPdenovo throw away anything up to 255.

It is convenient to do a **Kmer analysis using Jellyfish (or other Kmer counter)** before do any assembly to analyze contaminations (bacterial contaminations may appears with high Kmers content) and repeats.

Software to analyze repeats based in sequence graphs:
RepeatExplorer

(http://repeatexplorer.umbr.cas.cz/)

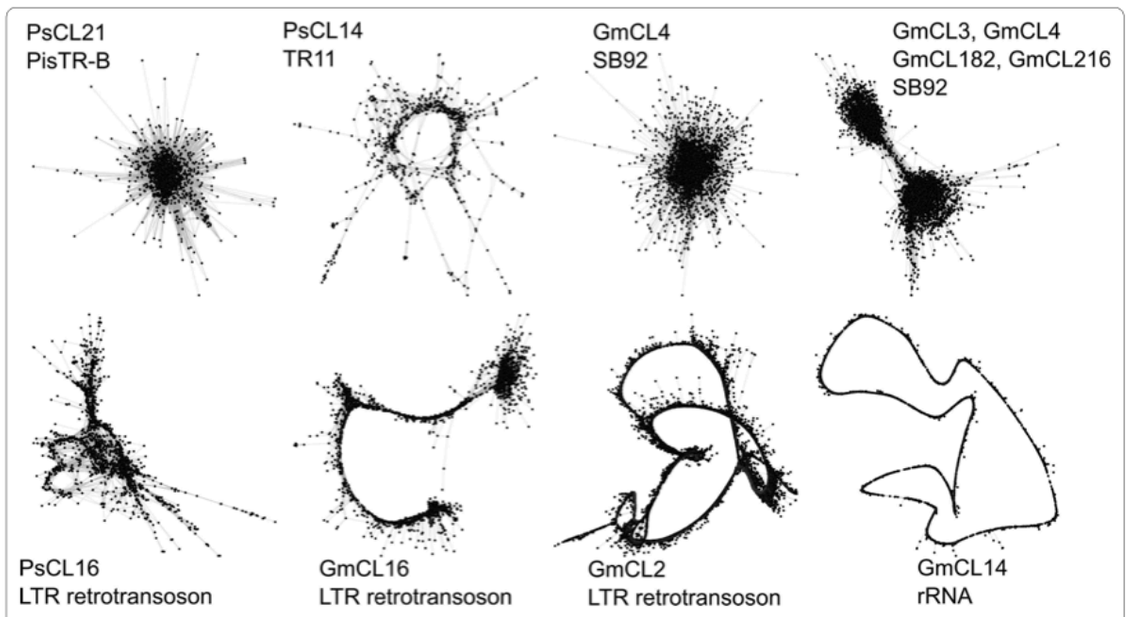


Figure 4 Examples of graph layouts derived from clusters of repetitive sequences. Graph layouts were calculated using the 3D version of Fruchterman and Reingold algorithm from which a 2D projection is shown. Individual reads are represented by vertices and similar reads are connected by edges. Individual clusters are described further in Table 1.

Novak, P., Neumann, P. & Macas, J. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. Bmc Bioinformatics 11, 378 (2010)

1. A brief history of the sequence assembly.

2. Sequencing, tools and computers.

3. Things that you should know about genomes.

4. What about transcriptomes ? Differences

4. What about transcriptomes ? Differences

	Genome	Transcriptome
Expected Assembly Size	Variable (plants from 63 Mb to 150 Gb) Example: *Oryza sativa* 380 Mb	Depends of transcriptome size. Example: *Oryza sativa* 42.25 Mb
Expected Number of Sequences	Few (ideally as many as chromosomes)	Many (ideally as many as expressed genes splicings)
Coverage	Fixed according the number of reads	Variable depending of the gene expression
Polymorphic Diversity for Duplications	High, from genes to repeats	Low, UTRs and CDSs Medium, alternative splicings
Complexity	High, specially in centromeric regions highly repetitive	Low, except in duplicated genes
Computational needs	High, assembly needs at least 256 Gb of RAM for medium size genome by WGS	Medium, ~30,000 genes, one Illumina lane (~60Gb) needs ~64 Gb of RAM

4. What about transcriptomes ? Differences

Software	Sequencing technology	Type	Features	URL
MIRA	Sanger, 454	Overlap-layout-consensus	Highly configurable	http://sourceforge.net/apps/mediawiki/mira-assembler
gsAssembler	Sanger, 454	Overlap-layout-consensus	Splicings	http://454.com/products/analysis-software/index.asp
iAssembler	Sanger, 454	Overlap-layout-consensus	Improves MIRA	http://bioinfo.bti.cornell.edu/tool/iAssembler
Trans-ABySS*	454 or Illumina	Bruijn graph	Splicings, Gene fusions	http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss
SOAPdenovo-trans*	454 or Illumina	Bruijn graph	Fastest	http://soap.genomics.org.cn/SOAPdenovo-Trans.html
Velvet/Oases	454 or Illumina or SOLiD	Bruijn graph	SOLiD	http://www.ebi.ac.uk/~zerbino/oases/
Trinity*	454 or Illumina	Bruijn graph	Downstream expression	http://trinityrnaseq.sourceforge.net/

* Comparisons in the Article: Vijay N. *et al* (2012) *Molecular Ecology* DOI: 10.1111/mec.12014

Keywords that you should remember

BAC-by-BAC

Whole Genome Shotgun

Reads

Mate Pairs

Pair Ends

Bruijn Graph

Overlap-layout-Consensus

Kmer

Scaffold

Contig

Consensus