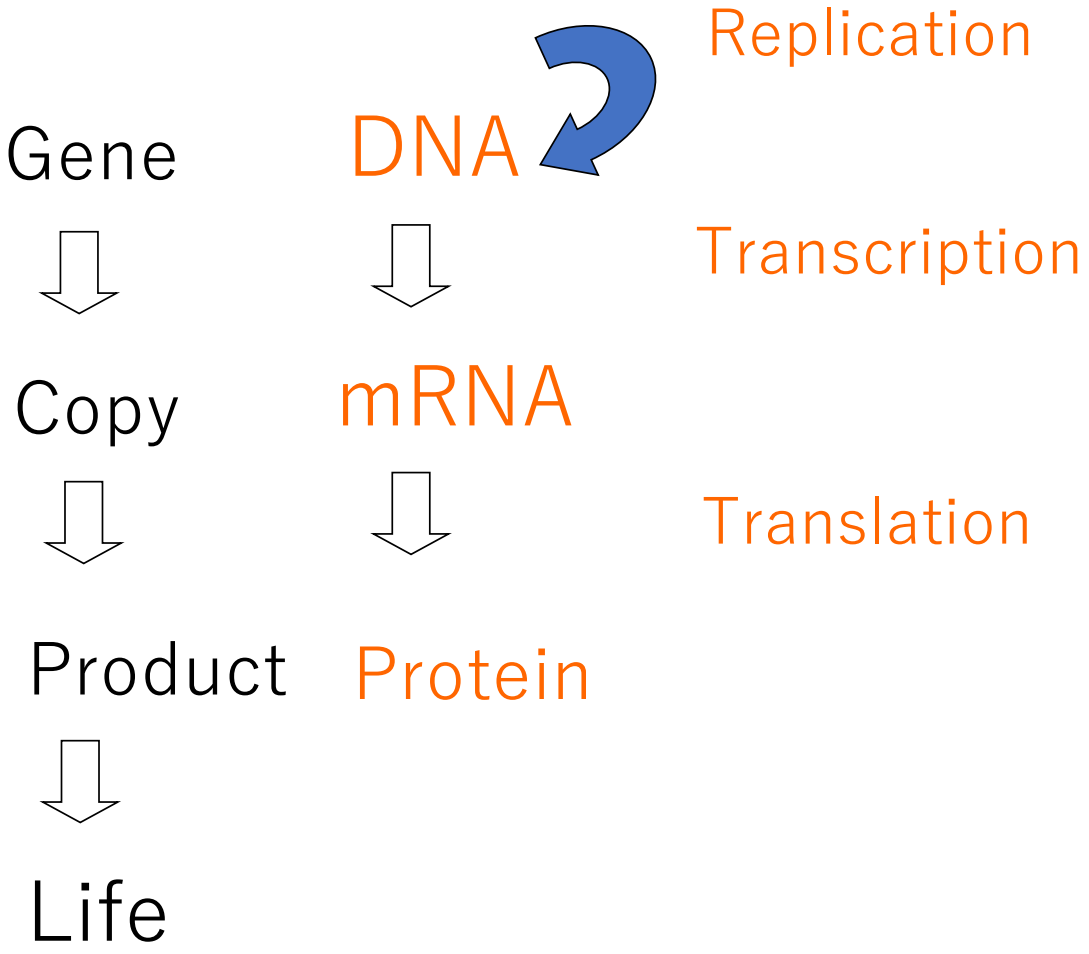
IAEA Training Course RAS5094

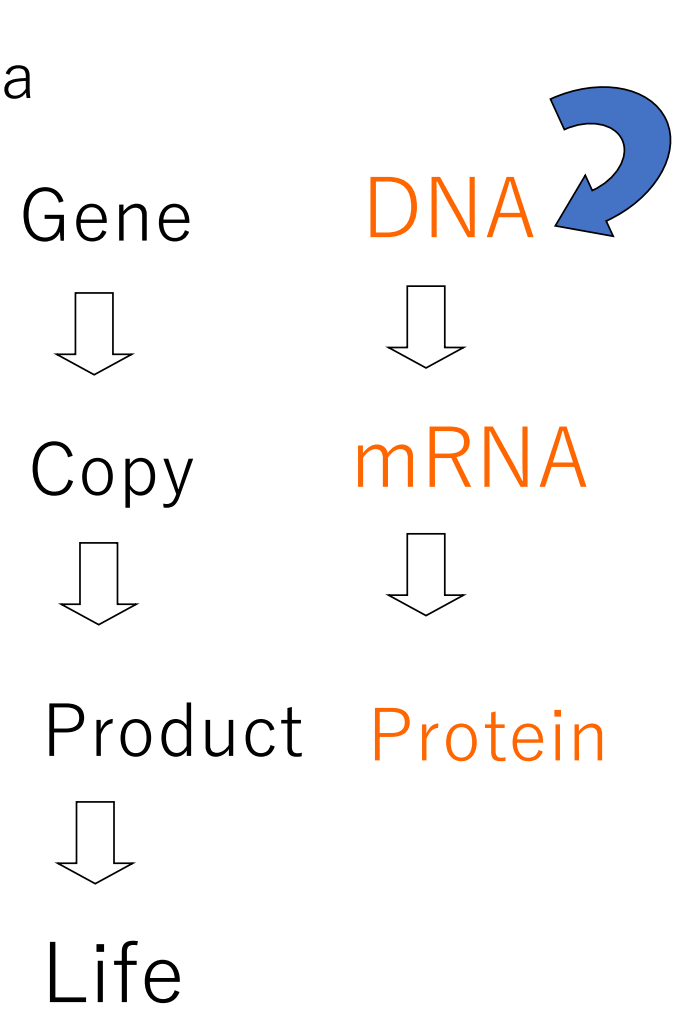
30 April, 2024 DNA-sequencing based technology for crop improvement

- Introduction to DNA sequencing
- Next Generation Sequencing (NGS)
- Genome assembly
- Use of long read sequences

Central Dogma
(Crick 1958)



Central Dogma
(Crick 1958)



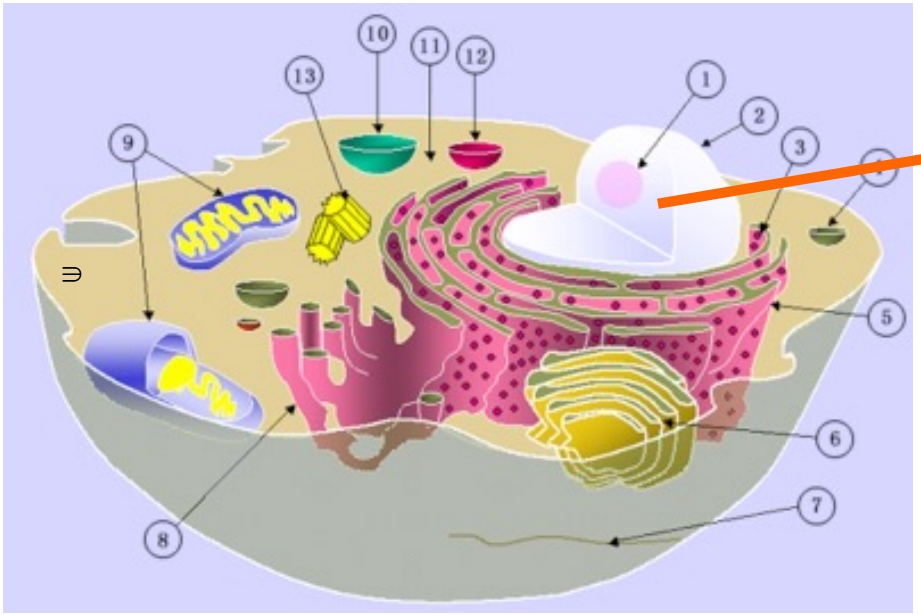
Omics

Genomics

Transcripto-
mics

Proteomics

A typical eukaryotic cell



Nucleus

Gene

Genome

DNA

5’-ATGCTG….-3’

(Wikipedia：Cell)

Cracking whole genome sequences

	Prokayote	Eukatyote	Note
1995	2	0	Influenza bactrium
1997	6	1	Yeast
1998	6	1	Nematode
2000	16	2	Arabidopsis
2002	38	9	Human, Rice (Draft sequences)

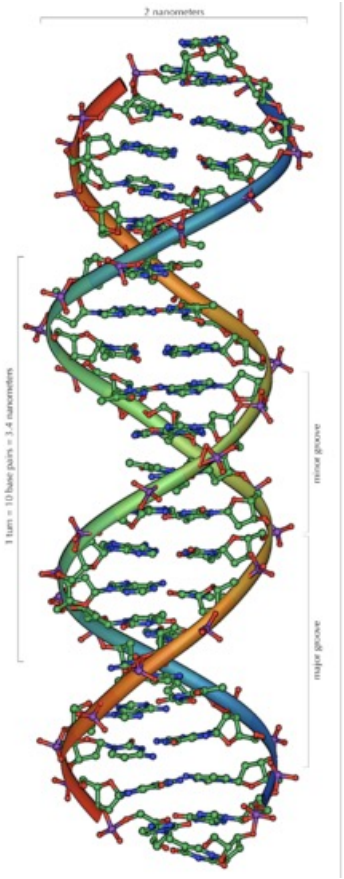
Higgs and Attwood (2005)

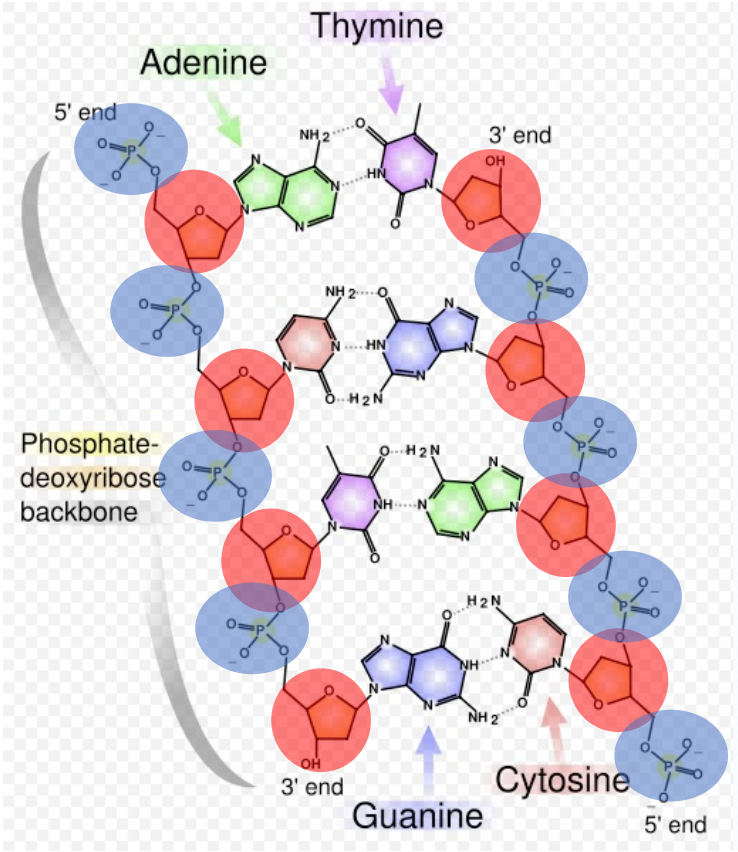
DNA sequencing platforms

1/ Sanger sequencing
ABI capillary

2/ Next-generation sequencing
454 (Roche)
Illumina
PacBio
Oxford Nanopore etc.

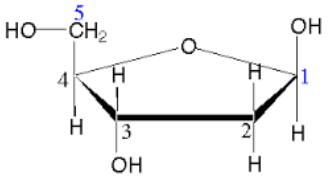
DNA





DNA

-Sugar (pentose)
- Phosphate
- Base



Sugar

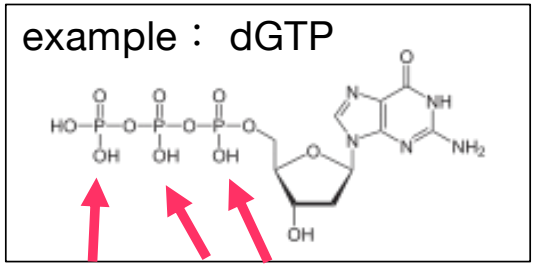
DNA sequencing technology

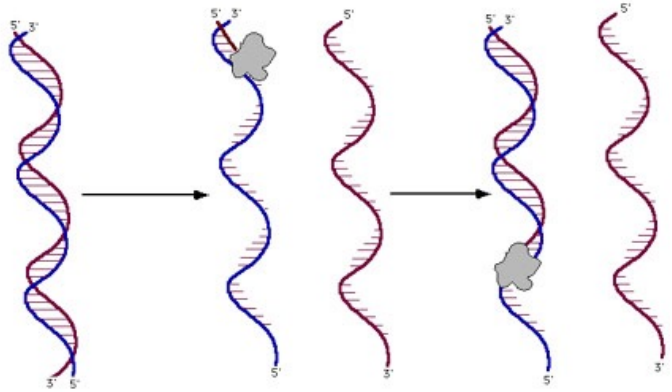
- Sanger method

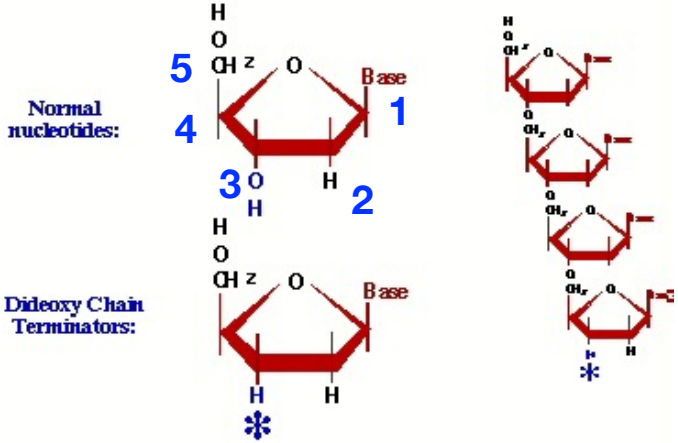
- Massively parallel sequencing

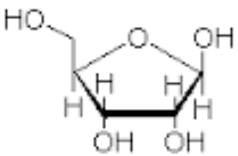
Sanger Sequencing

DNA replication: - template DNA
- primer
- DNA polymerase
- deoxyribonucleotides (dATP, dCTP, dTTP, dGTP)

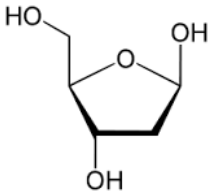




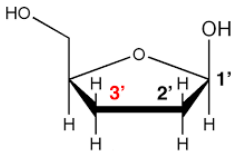




Ribose



Deoxy ribose



Dideoxy ribose

Sanger Sequencing

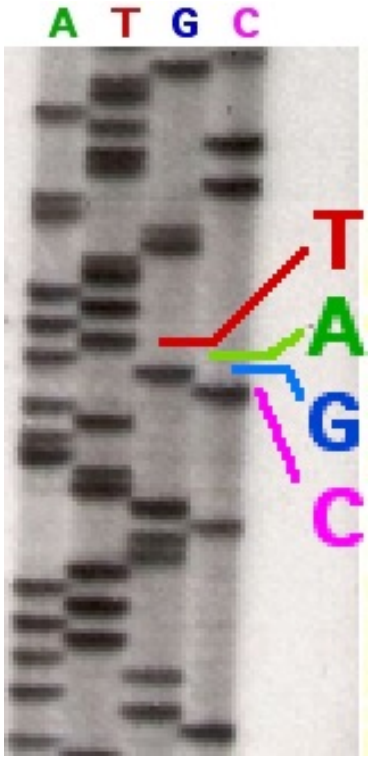
DNA Polymerase reads the template strand and synthesizes a new second strand to match:

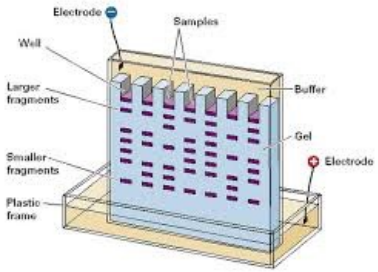
```
5' - TACGCGGTAACGGTATGTTCGACCGTTTAGCTACCGAT
3' - ATGCGCCATTGCCATACAAGCTGGCAAATCGATGGCTAGAGATCCAA - 5'
```

IF 5% of the T nucleotides are actually dideoxy T, then each strand will terminate when it gets a ddT on its growing end:

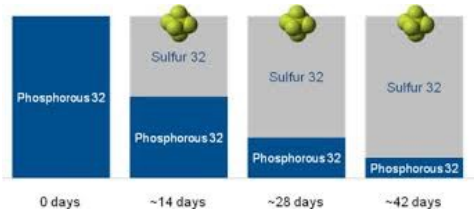
```
5' - TACGCGGTAACGGTATGTTCGACCGTTTAGCTACCGAT•
5' - TACGCGGTAACGGTATGTTCGACCGTTTAGCT•
5' - TACGCGGTAACGGTATGTTCGACCGTTT•
5' - TACGCGGTAACGGTATGTTCGACCGTT•
5' - TACGCGGTAACGGTATGTTCGACCGT•
5' - TACGCGGTAACGGTATGTT•
5' - TACGCGGTAACGGTATGT•
5' - TACGCGGTAACGGTAT•
5' - TACGCGGTAACGGT•
5' - TACGCGGT•
```

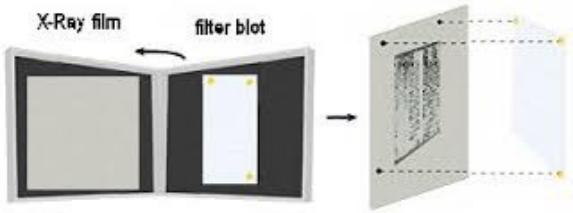
100% dATP
100% dGTP
100% dCTP
95% dTTP 5% ddTTP





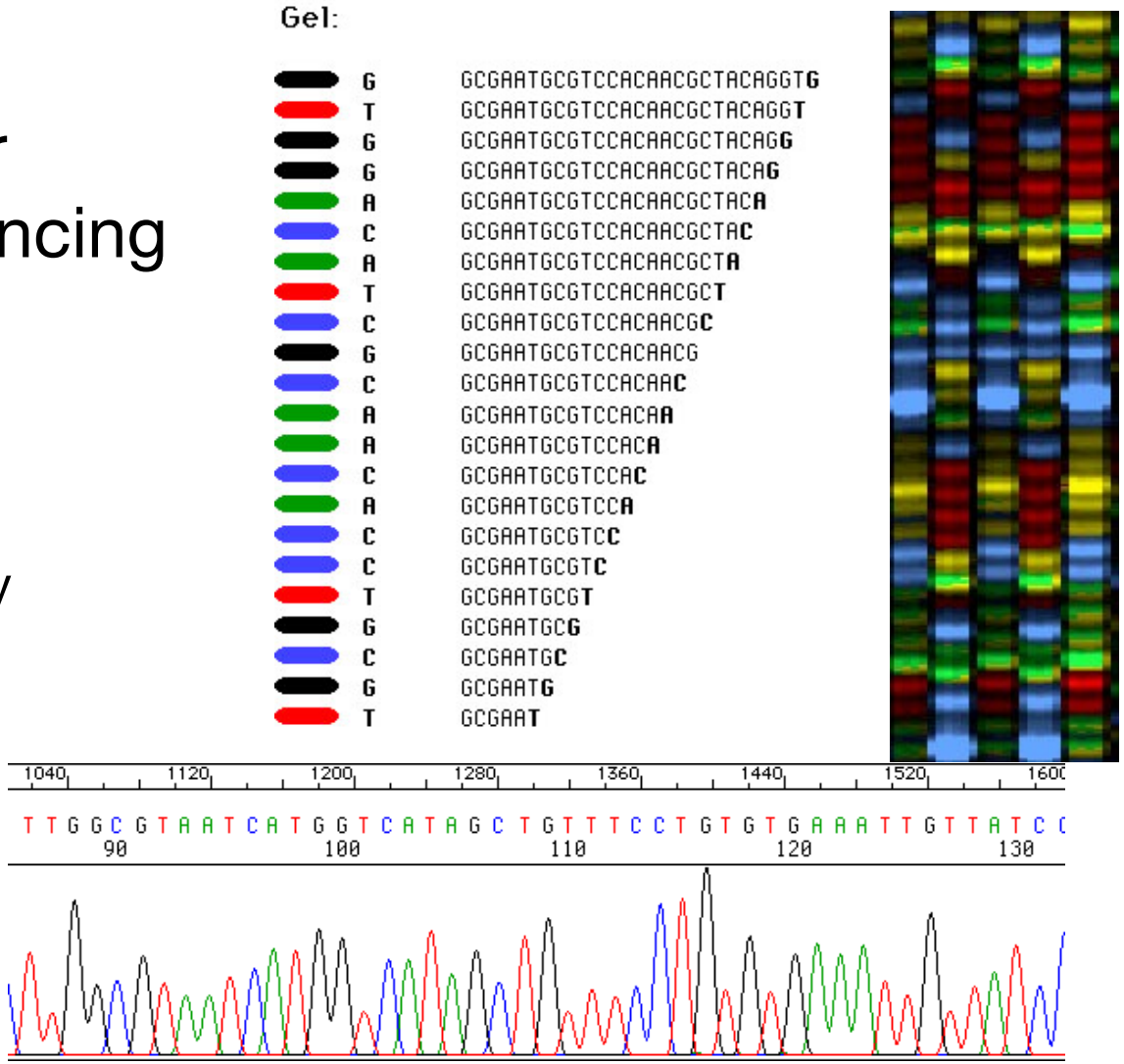
^{32}P Beta ray emitter





Sanger sequencing

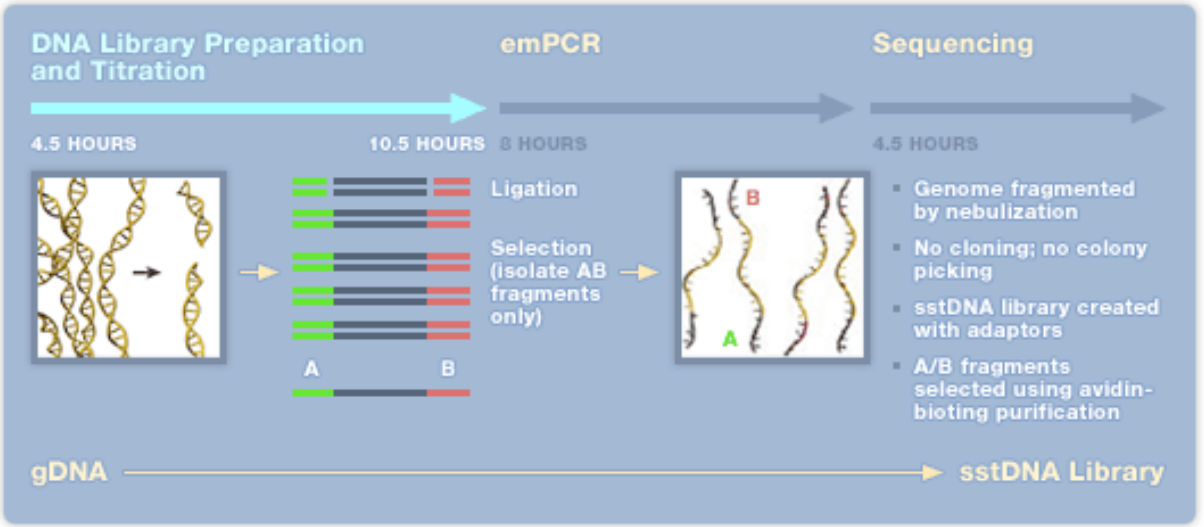
ABI capillary sequencing



http://seqcore.brcf.med.umich.edu/doc/educ/dnapr/sequencing.html

Massively parallel sequencing

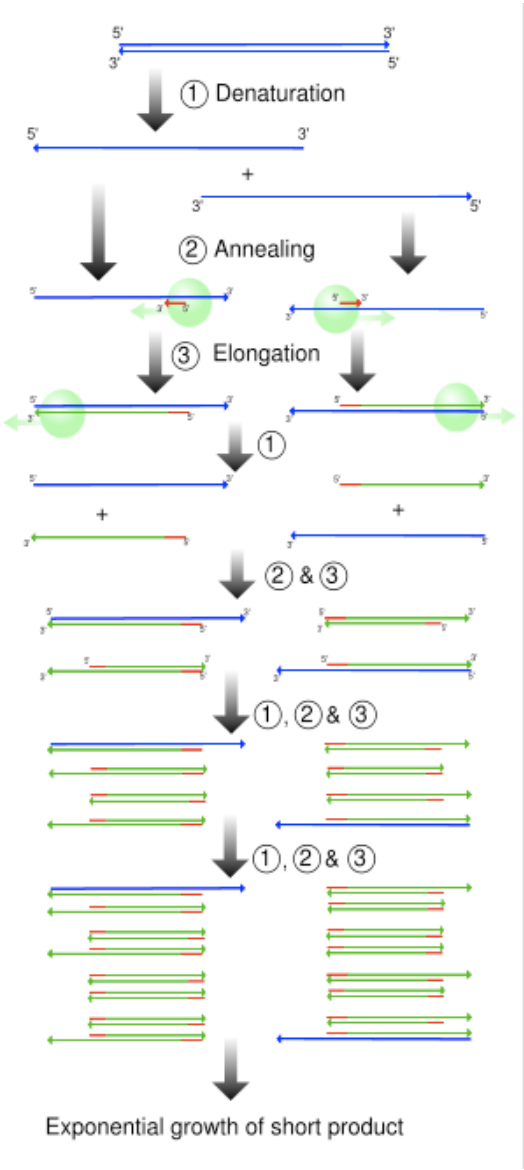
454 pyrosequencing (~2010)



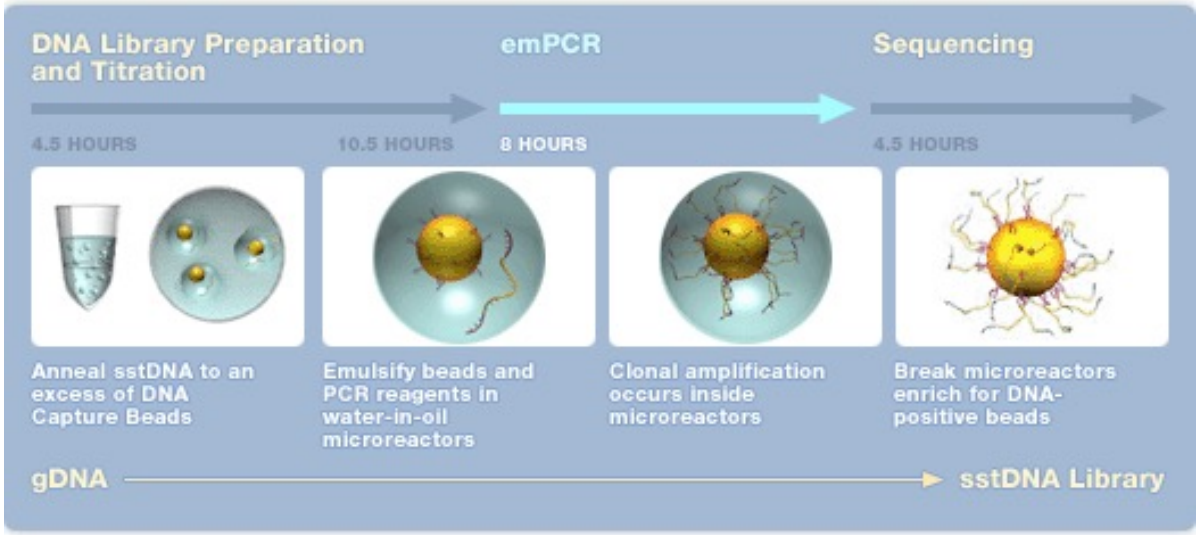
http://www.454.com/sequencing-services/services.asp

PCR

- Template DNA
- 2 primers
- Taq DNA polymerase
- dNTPs

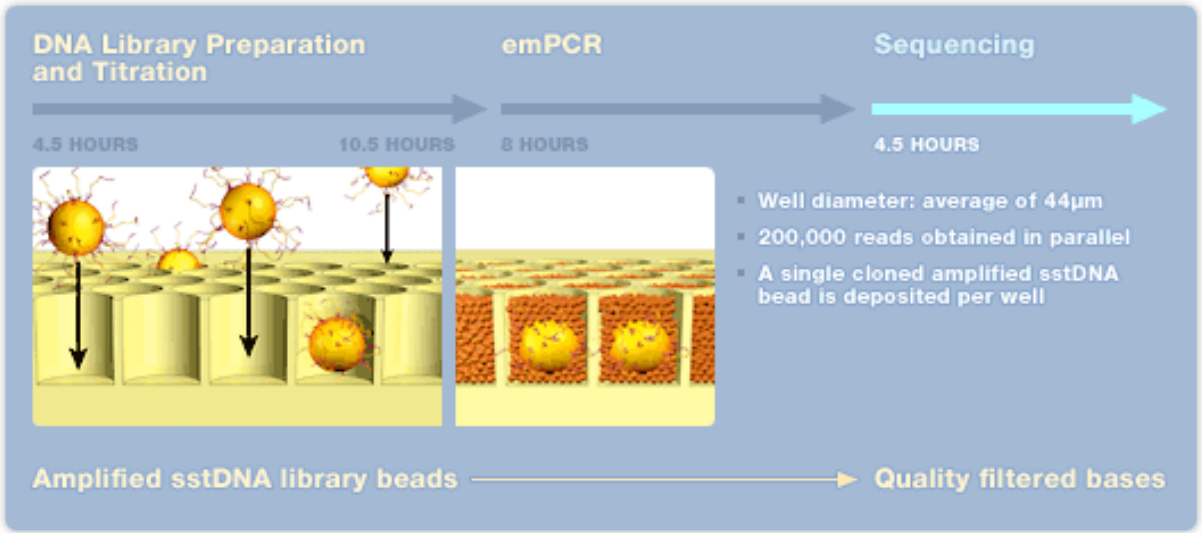


454 pyrosequencing



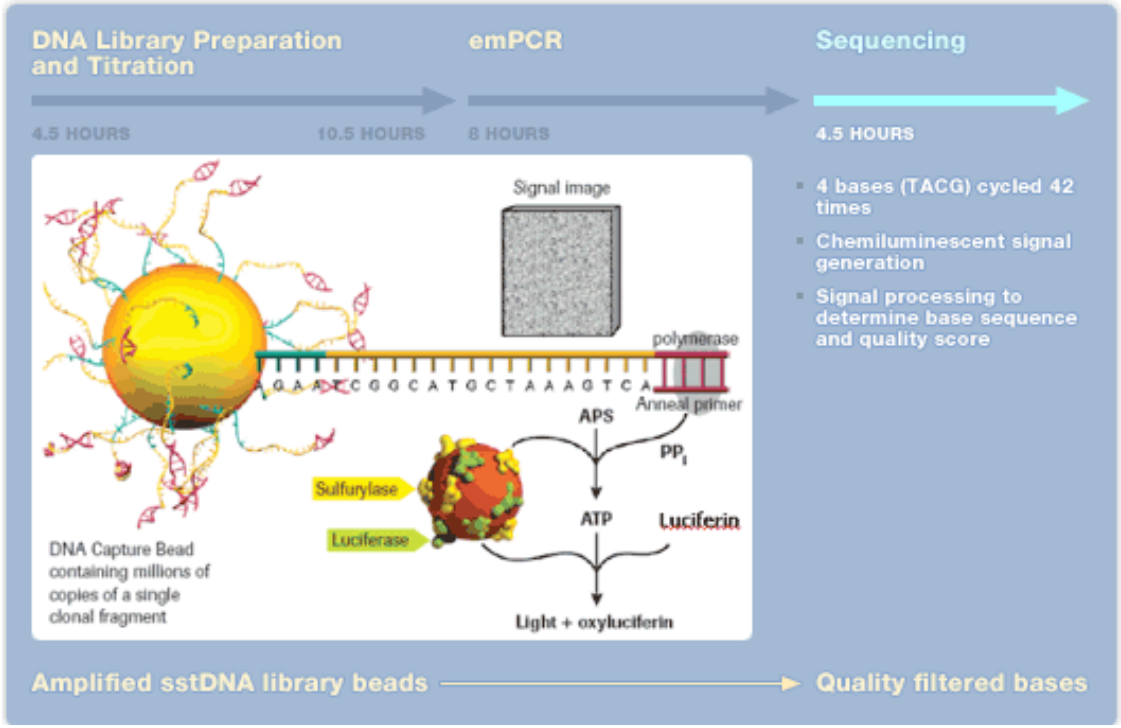
http://www.454.com/sequencing-services/services.asp

454 pyrosequencing



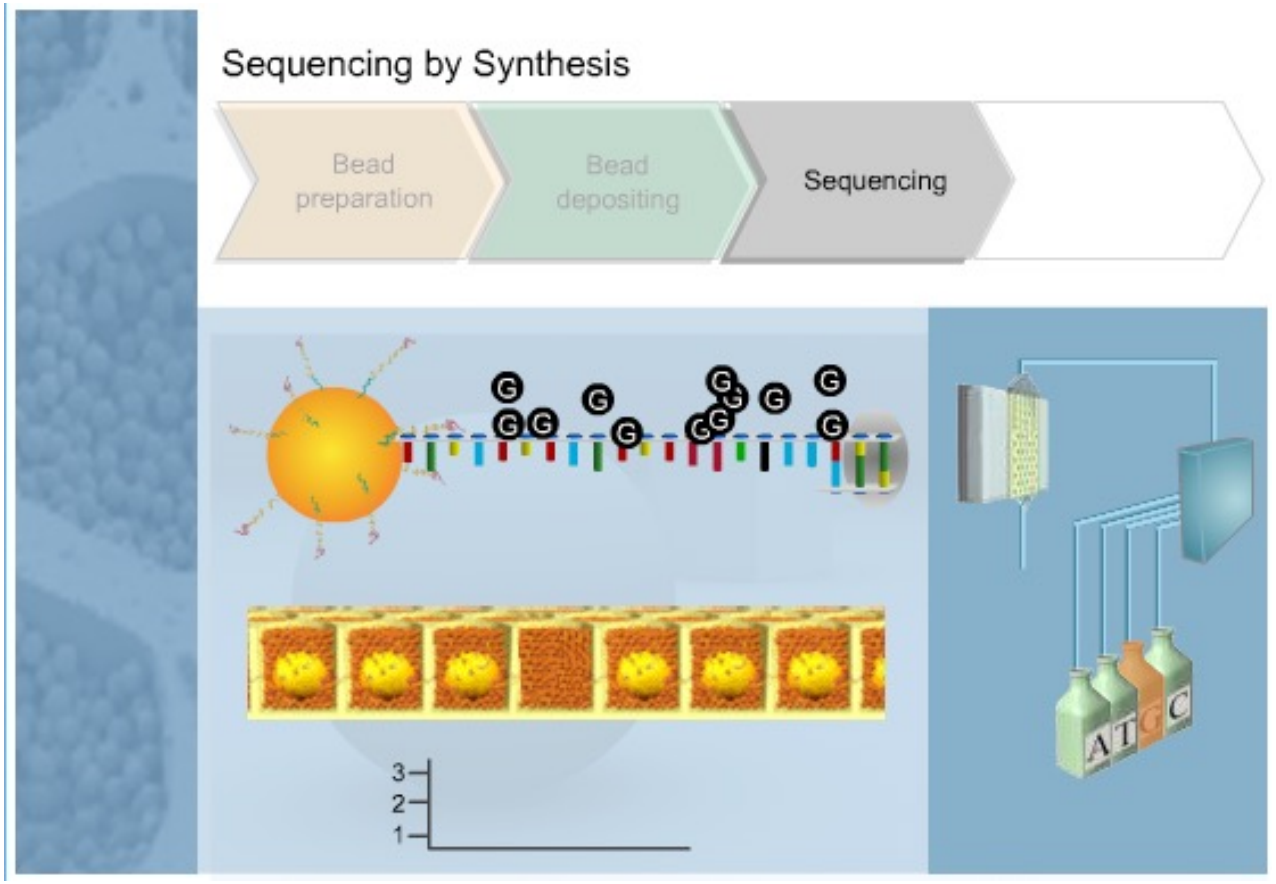
http://www.454.com/sequencing-services/services.asp

454 pyrosequencing

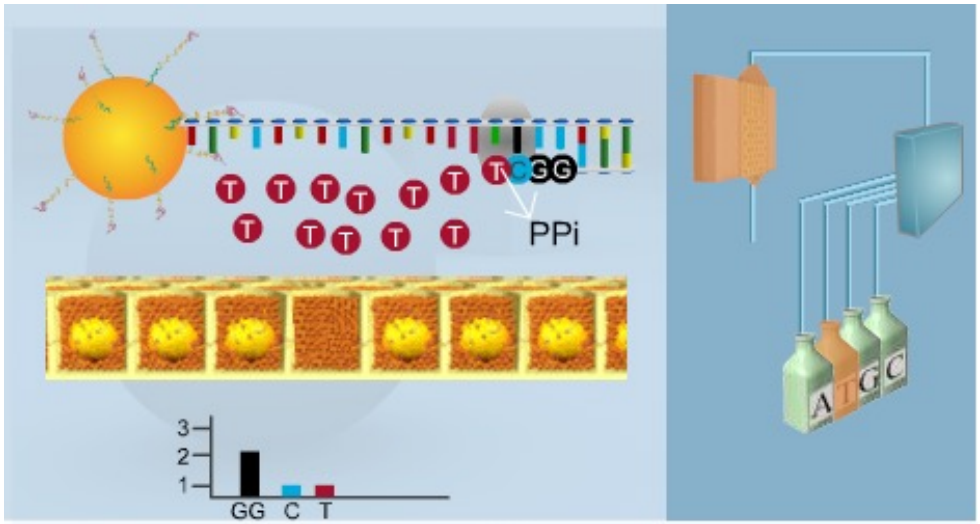


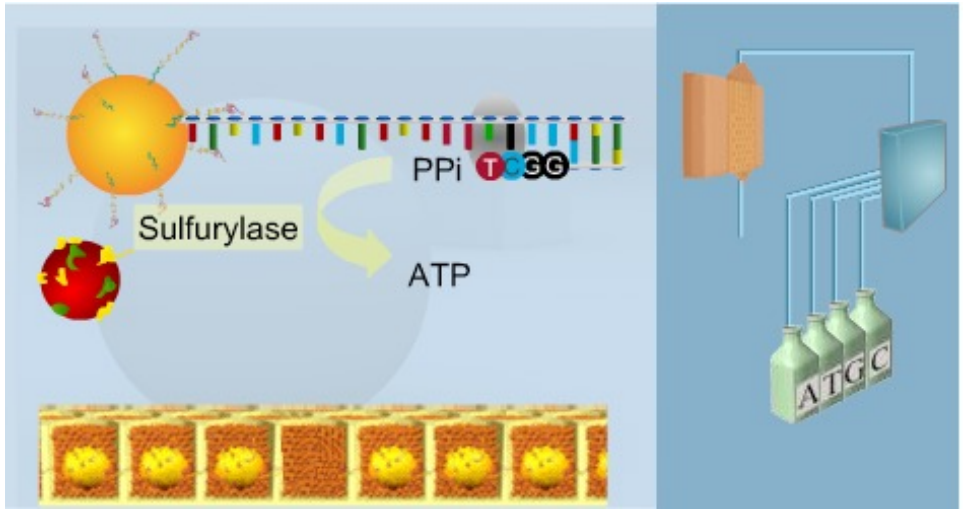
- GTP => GMP + PPi (by DNA polymerase)
- PPi + APS(adenosine 5'-phosphosulfate) => ATP (by Sulfurylase)
- ATP + Luciferin => LIGHT (by Luciferase)

454 pyrosequencing

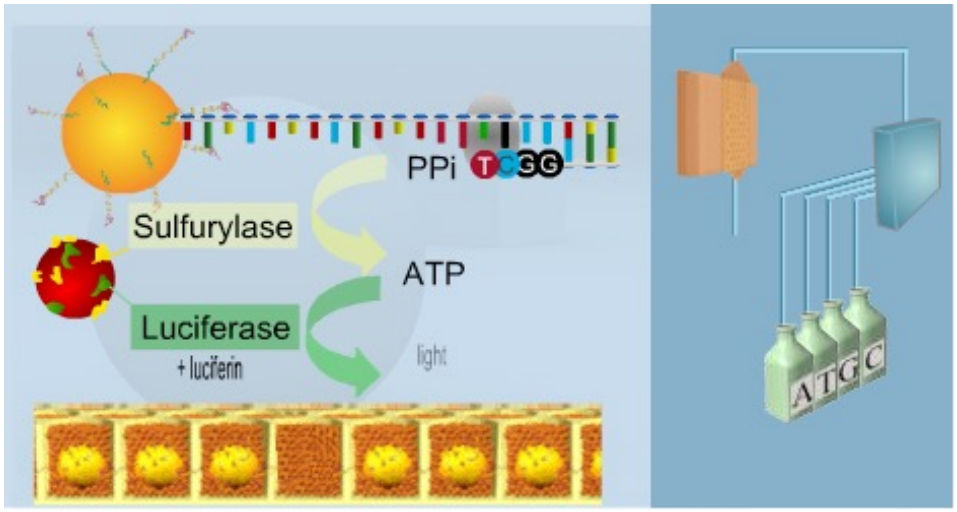


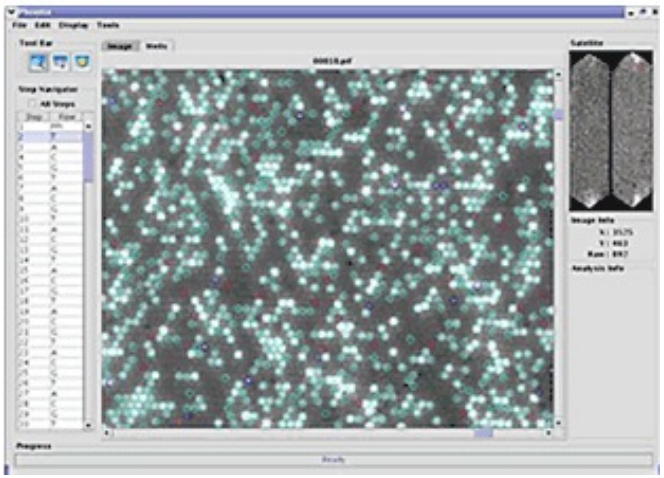
http://www.454.com/sequencing-services/services.asp



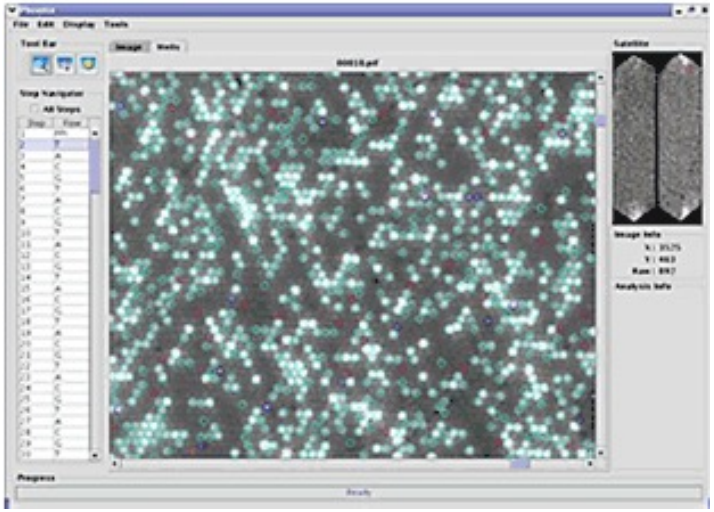


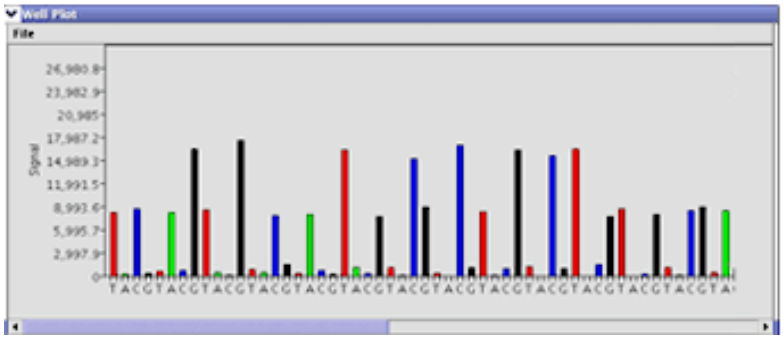
http://www.454.com/sequencing-services/services.asp

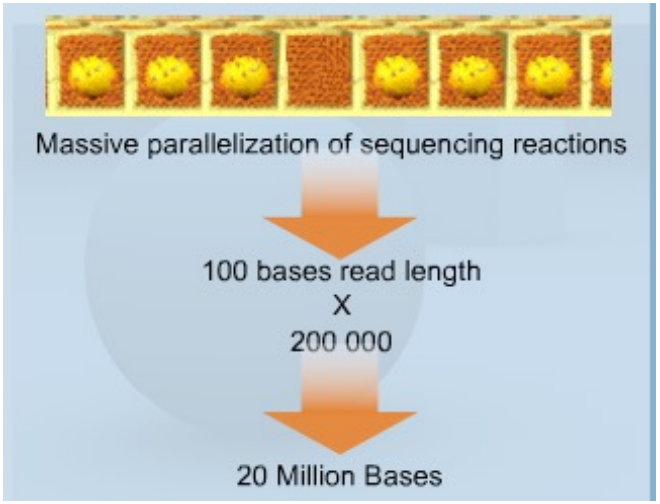




http://www.454.com/sequencing-services/services.asp







http://www.454.com/sequencing-services/services.asp

➔ Illumina seq info

Illumina sequencing

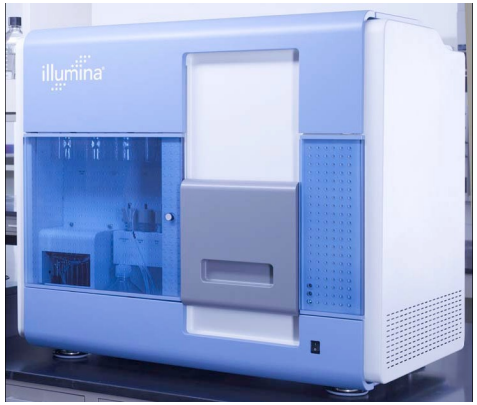
Sequence PCR amplicons

Watch video

https://www.youtube.com/watch?annotation_id=annotation_228575861&feature=iv&src_vid=womKfikWlxM&v=fCd6B5HRaZ8

Paired-End (PE) reads
Mate Pair (MP) reads

Illumuna DNA sequencer



Illumina Genome Analyzer Iix (2010)

50 Gb per run

Illumina HiSeq2500

250 Gb per run

Short read size: 100~200 bp

+ NextSeq500 (2015-)

120 Gb per run

単位

T G M K 1 milli micro nano pico

PacBio sequencing
-single molecule sequencing

Watch video

https://www.youtube.com/watch?v=v8p4ph2MAvI

Oxford Nanopore sequencing
- Single molecule

https://www.youtube.com/watch?v=qzusVw4Dp8w

https://www.youtube.com/watch?v=RcP85JHLmnI

Portable, real-time biological analyses

MinION is the only portable real-time device for DNA and RNA sequencing.

Each consumable flow cell can now generate 10–20 Gb of DNA sequence data. Ultra-long read lengths are possible (hundreds of kb) as you can choose your fragment length. The MinION streams data in real time so that analysis can be performed during the experiment and workflows are fully versatile.

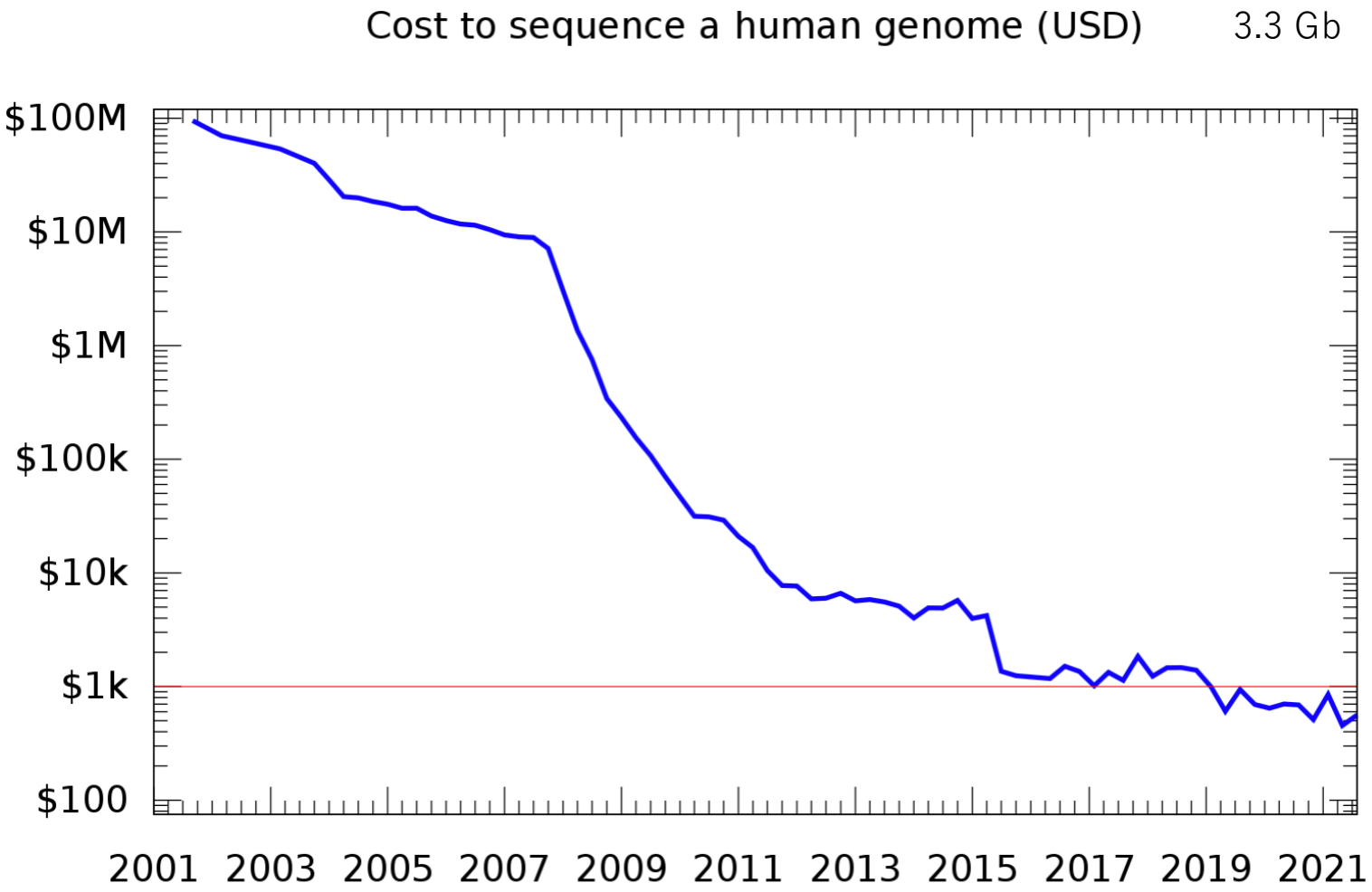
The MinION weighs under 100 g and plugs into a PC or laptop using a high-speed USB 3.0 cable. No additional computing infrastructure is required. Not constrained to a laboratory environment, it has been used up a mountain, in a jungle, in the arctic and on the International Space Station.

The MinION is commercially available, simply by paying a starter-pack fee of $1,000. The MinION starter pack includes materials you need to run initial sequencing experiments, including a MinION device, flow cells and kits, as well as membership of the Nanopore Community.

Commonly used NGS

Library schematic	Output	Typical assembly
Illumina 	~4 × 10^8 × 2 × 150 reads (one lane HiSeq 4,000)	10^3–10^5 contig N50
PacBio 	~5 × 10^5 × ~10 kb reads (PacBio Sequel SMRT cell)	~10^6 contig N50
Oxford Nanopore 	~3.6 × 10^6 × ~10 kb reads (ONT Minion)	~10^6 contig N50

Rice ES, Green RE. 2019.
Annu. Rev. Anim. Biosci. 7:17–40



Source: Creative Commons Originally Ben Moore, rewritten in gnuplot by grendel|khan.

Read amount and Read length

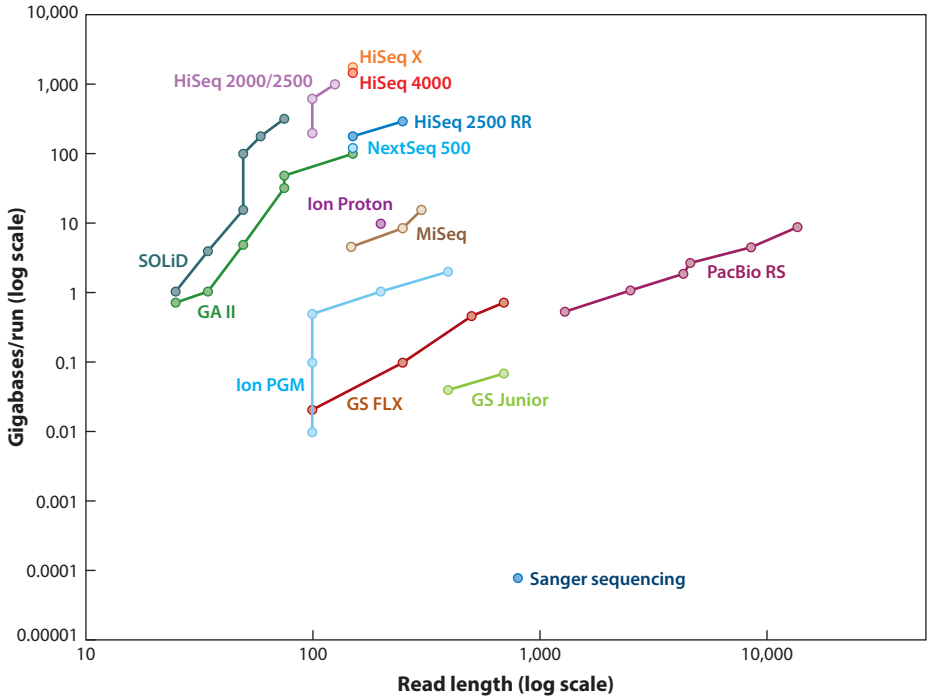


Figure 1

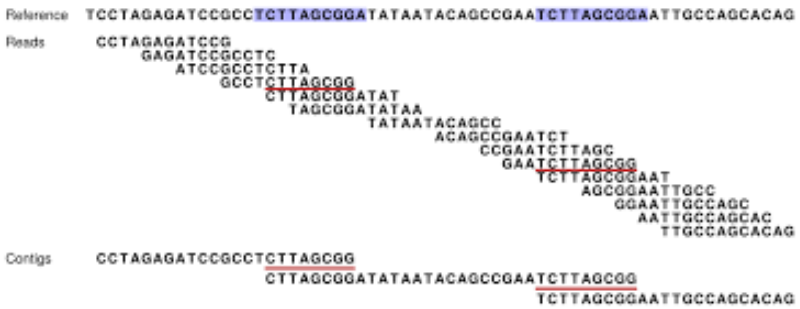
Developments in high-throughput sequencing. SOLiD is an Applied Biosystems platform; Ion PGM and Ion Proton are Ion Torrent platforms; GA II, HiSeq, NextSeq, and MiSeq are Illumina platforms; GS FLX and GS Junior are Roche 454 platforms; and PacBio RS is a Pacific Biosciences platform. Adapted from a figure created by Lex Nederbragt (**http://dx.doi.org/10.6084/m9.figshare.100940**) under the Creative Commons Attribution license (**http://creativecommons.org/licenses/by/4.0**).

Annu. Rev. Genom. Hum. Genet. 2016. 17:95–115

First published online as a Review in Advance on June 9, 2016

The *Annual Review of Genomics and Human Genetics*

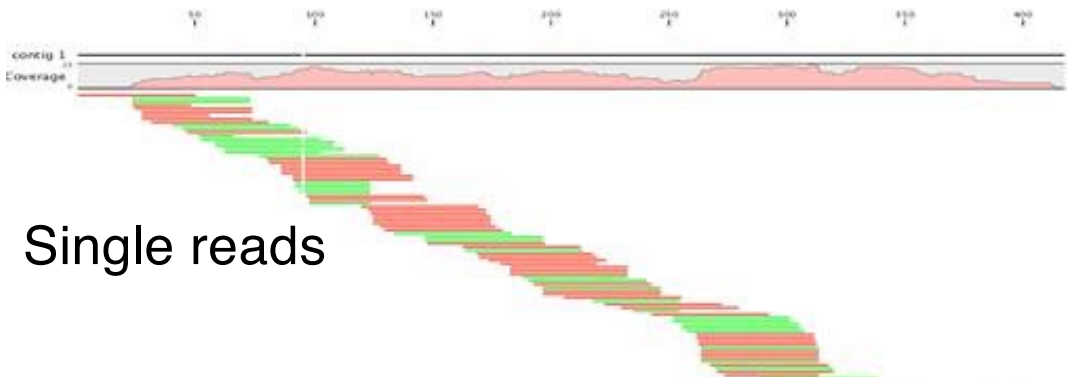
De novo assembly



Alignment (Mapping)

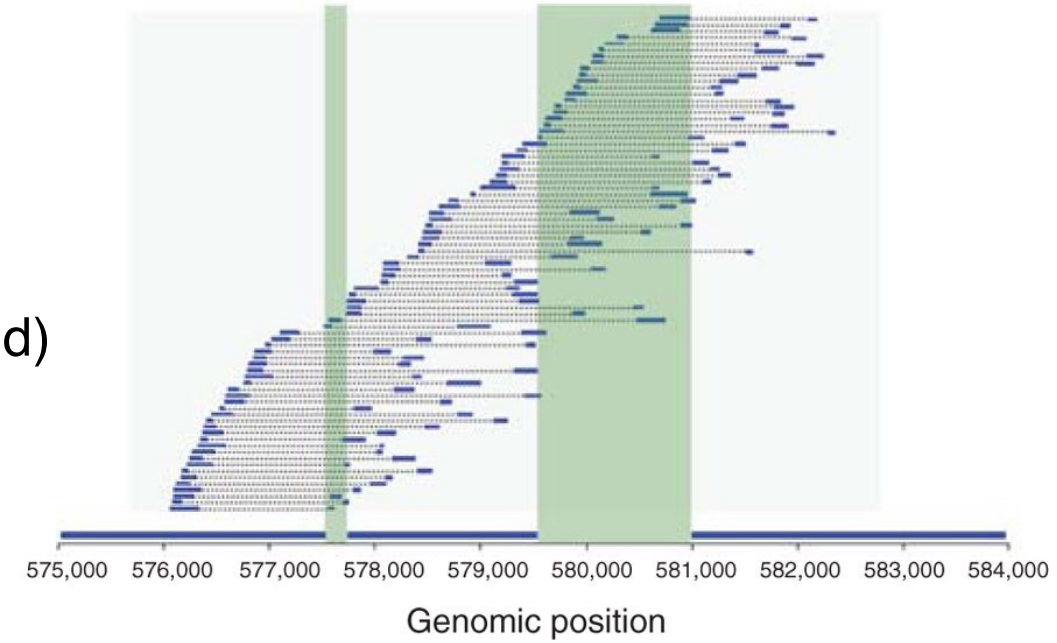
Genome assembly

de novo assembly



Single reads

PE
(Paired-End)
reads



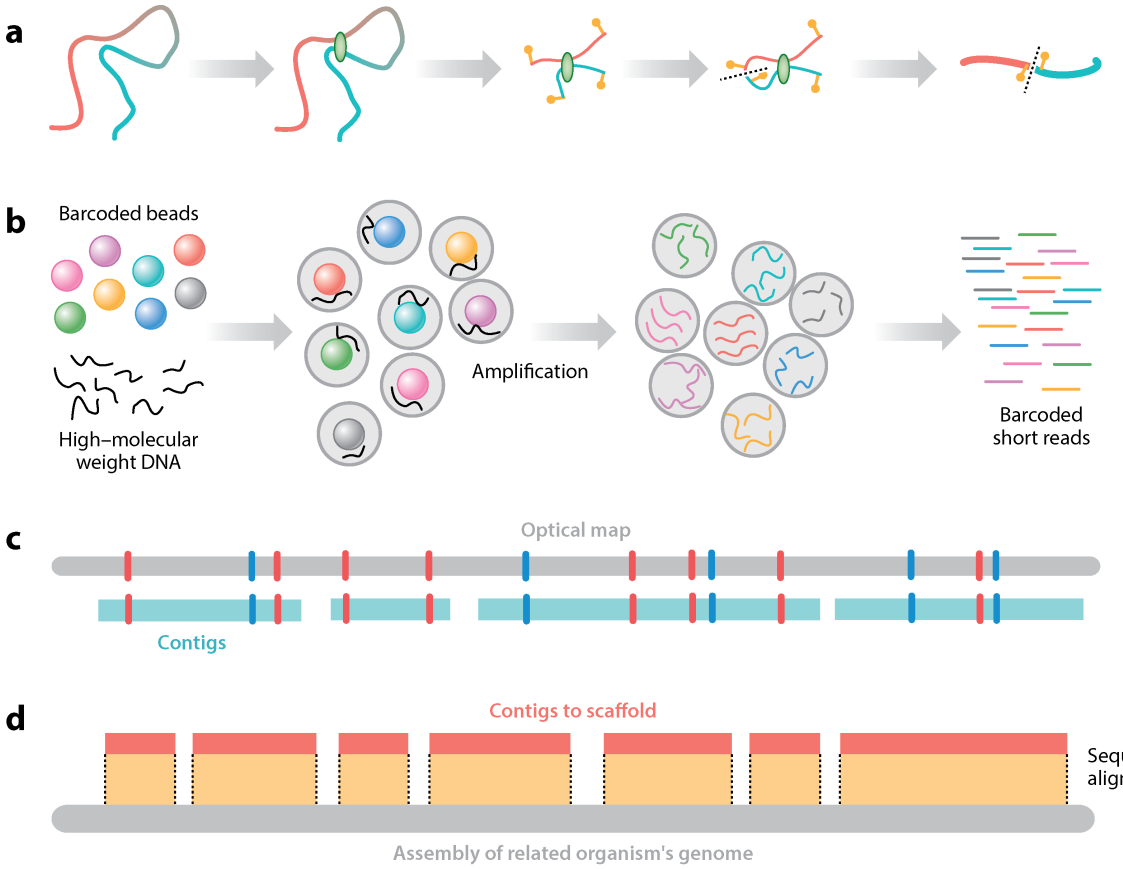
Commonly used assembly software

Table 1 Commonly used assembly software

Software	URL and reference	Description
Short-read assembly software		
Velvet	http://github.com/dzerbino/velvet (**168**)	Original de Bruijn graph assembler
SOAPdenovo	http://soap.genomics.org.cn/ (**169**)	De Bruijn graph assembler with error-correction step
Meraculous	https://jgi.doe.gov/data-and-tools/meraculous/ (**170**)	Hybrid k-mer/read-based
ALLPATHS-LG	http://software.broadinstitute.org/allpaths-lg/blog/ (**171**)	Uses unipath graph to collapse repeats
SGA	https://github.com/jts/sga (**172**)	Uses string graphs
ABySS	https://github.com/bcgsc/abyss (**173**)	Represents de Bruijn graph with a Bloom filter
DISCOVAR de novo	https://software.broadinstitute.org/software/discovar/blog/ (**174**)	Requires 250-bp PCR-free reads
Supernova	https://github.com/10XGenomics/supernova (**149**)	Assembles 10× linked reads
Long-read assembly software		
HGAP	https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/HGAP (**124**)	Error correction, overlap-layout-consensus assembly, and polishing workflow
Canu	https://github.com/marbl/canu (**125**)	K-mer-based overlap computation
FALCON	https://github.com/PacificBiosciences/FALCON (**103**)	Assembles phased diploid genomes
Flye	https://github.com/fenderglass/Flye (**129**)	Uses A-Bruijn graph
Miniasm	https://github.com/lh3/miniasm (**128**)	Fast, but no error correction
Polishing software		
Pilon	https://github.com/broadinstitute/pilon (**133**)	Uses short-read alignments to correct errors
Arrow	https://github.com/PacificBiosciences/GenomicConsensus	Hidden Markov model and long-read alignments
Nanopolish	https://github.com/jts/nanopolish (**115**)	Nanopore only; uses original voltage data to correct errors

Rice and Green, New Approaches for Genome Assembly and Scaffolding
Annual Review of Animal Biosciences
Vol. 7:17-40 (2019)
https://doi.org/10.1146/annurev-animal-020518-115344

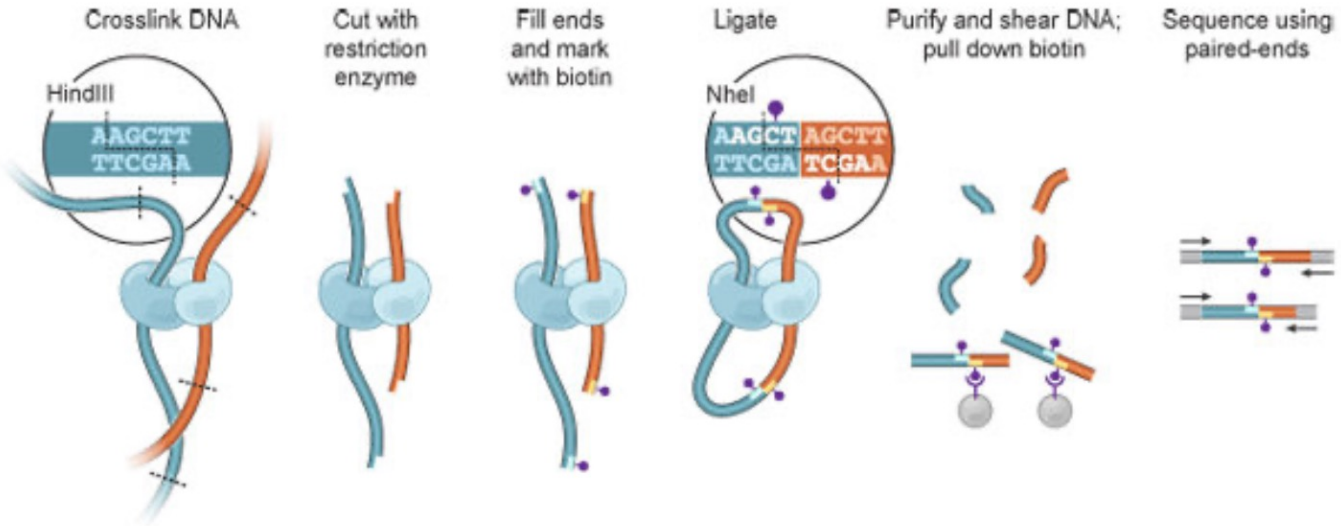
New approaches for long-range genome scaffolding



Hi-C

Optical mapping

Rice ES, Green RE. 2019.
Annu. Rev. Anim. Biosci. 7:17–40



Hi-C: High-throughput chromosome conformation capture

Biology

Hi-C: A Method to Study the Three-dimensional Architecture of Genomes.

Published: May 6, 2010 doi: 10.3791/1869

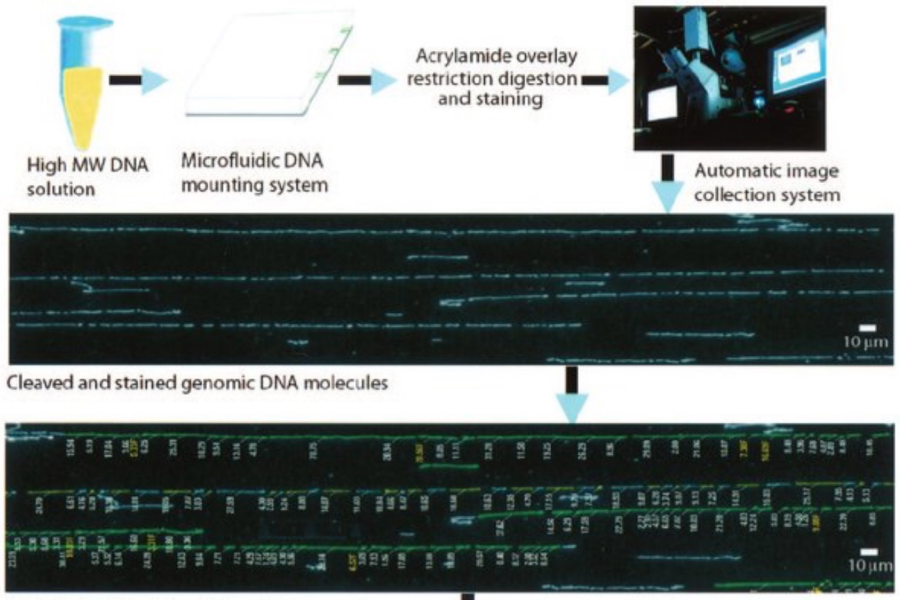
Nynke L. van Berkum*[1], Erez Lieberman-Aiden*[2,3,4,5], Louise Williams*[2], Maxim Imakaev[6], Andreas Gnirke[2], Leonid A. Mirny[3,6], Job Dekker[1], Eric S. Lander[2,7,8]

Figure

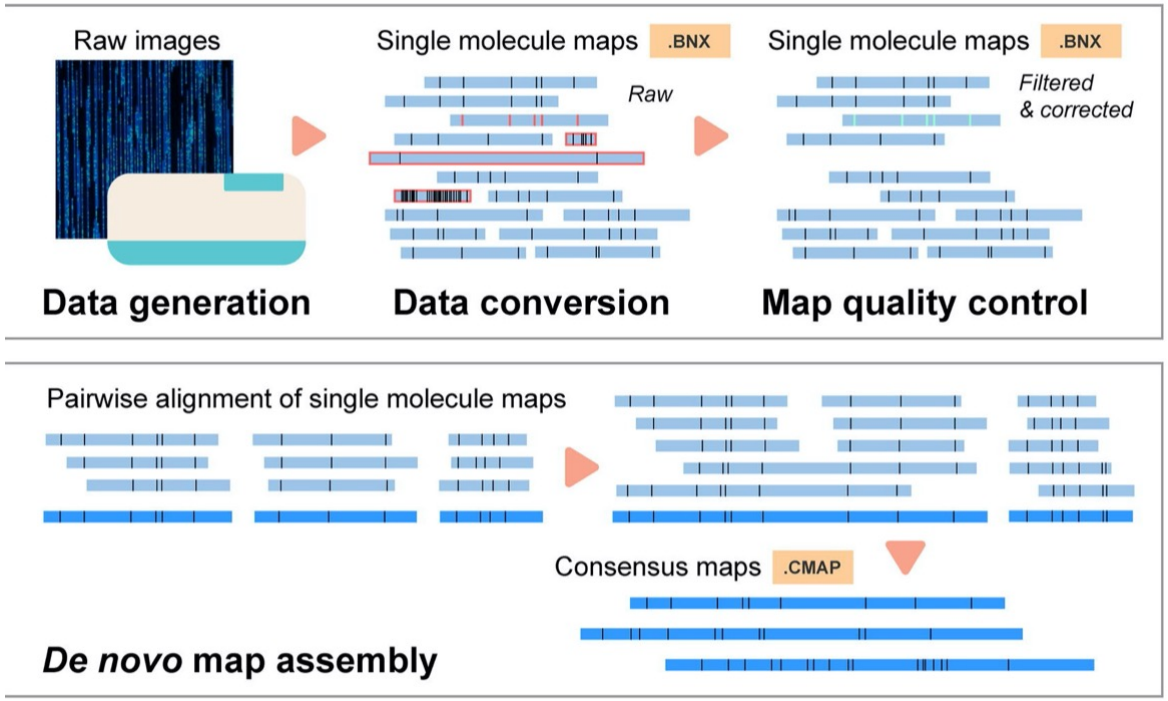
Caption

Figure 1. Hi-C overview. Cells are cross-linked with formaldehyde, resulting in covalent links between spatially adjacent chromatin segments (DNA fragments: dark blue, red; Proteins, which can mediate such interactions, are shown in light blue and cyan). Chromatin is digested with a restriction enzyme (here, HindIII; restriction site: dashed line, see inset). The resulting sticky ends are filled in with nucleotides, one of which is biotinylated (purple dot). Ligation is performed under extremely dilute conditions favoring intramolecular ligation events; the HindIII site is lost and an NheI site is created (inset). DNA is purified and sheared, and biotinylated junctions are isolated using streptavidin beads. Interacting fragments are identified by paired-end sequencing.

Optical Mapping



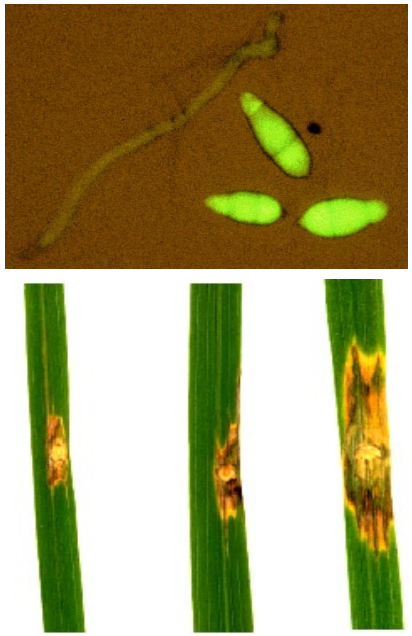
Zhou et al. 2004
Single-Molecule Approach to Bacterial Genomic Comparisons via Optical Mapping
December 2004
Journal of Bacteriology 186(22):7773-82
DOI: 10.1128/JB.186.22.7773-7782.2004



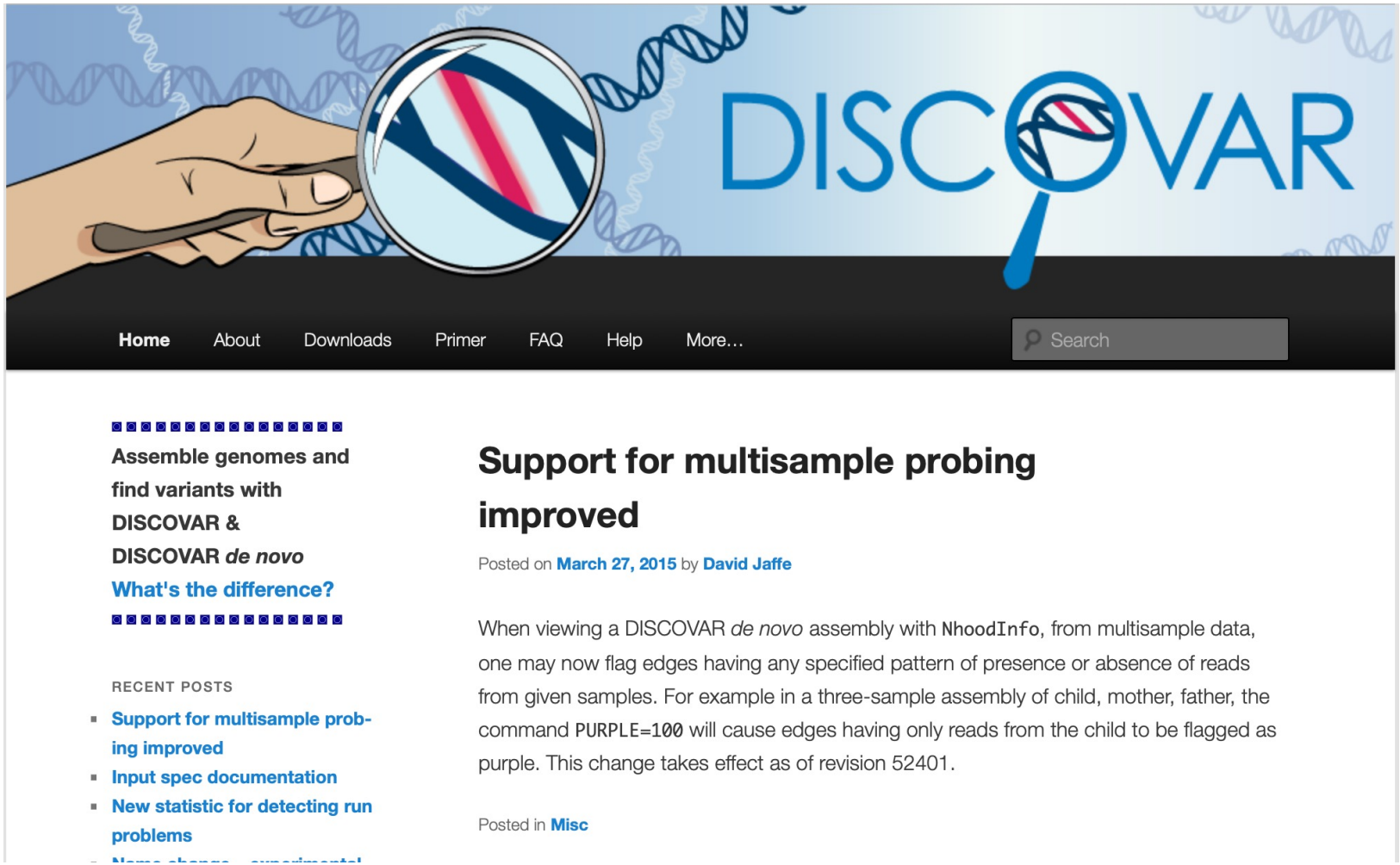
Yuan et al. Advances in optical mapping for genomic research
Computational and Structural Biotechnology Journal Volume 18, 2020, Pages 2051-2062

Example of de novo assembly

Rice blast disease caused by the ascomycete fungus
Magnaporthe oryzae



De novo assembler



Demonstration of *de novo* assembly

010.discovar_denovo

Material: *Magnaporthe oryzae*
isolate “Ina168” (~40 Mb genome)

Sequence reads:
Illumina Myseq PE reads (231 bp)
Total reads: PE1 = PE2 = 28,8730 reads
Total size: PE1= PE2 = 0.19 Gbp

Assembler: *Discovar de novo* (Broad Institute)

Example of de novo assembly of the genome of a fungus, *Magnaporthe oryzae*

Accessing supercomputer at National Institute of Genetics, Mishima, Japan

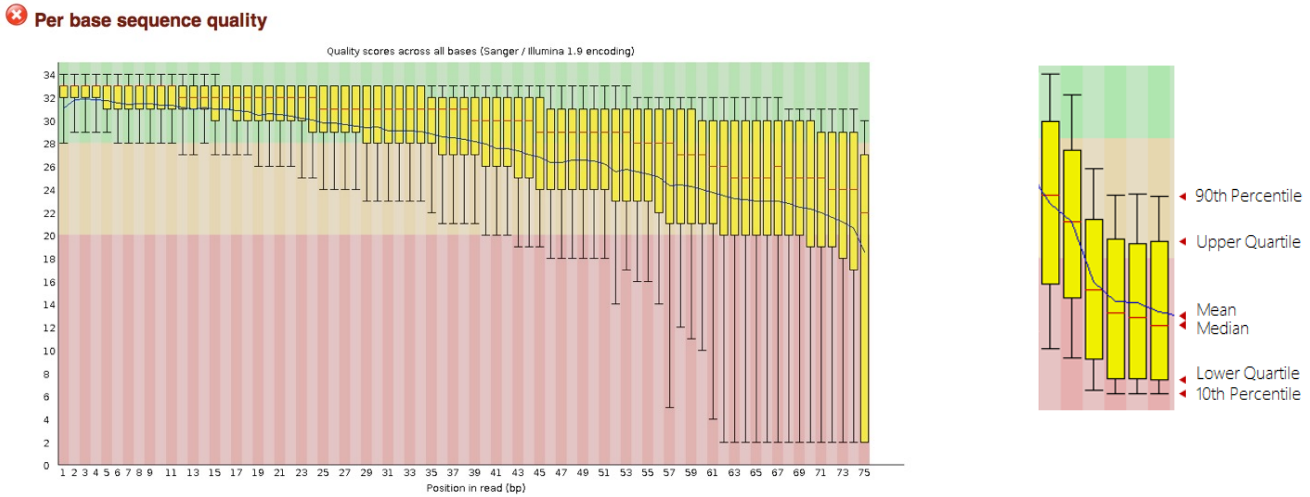
Terminal

$ ssh XXXXXX@gw2.ddbj.nig.ac.jp
login
$ ssh at 139

Fastq file

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Fastqc output



N50

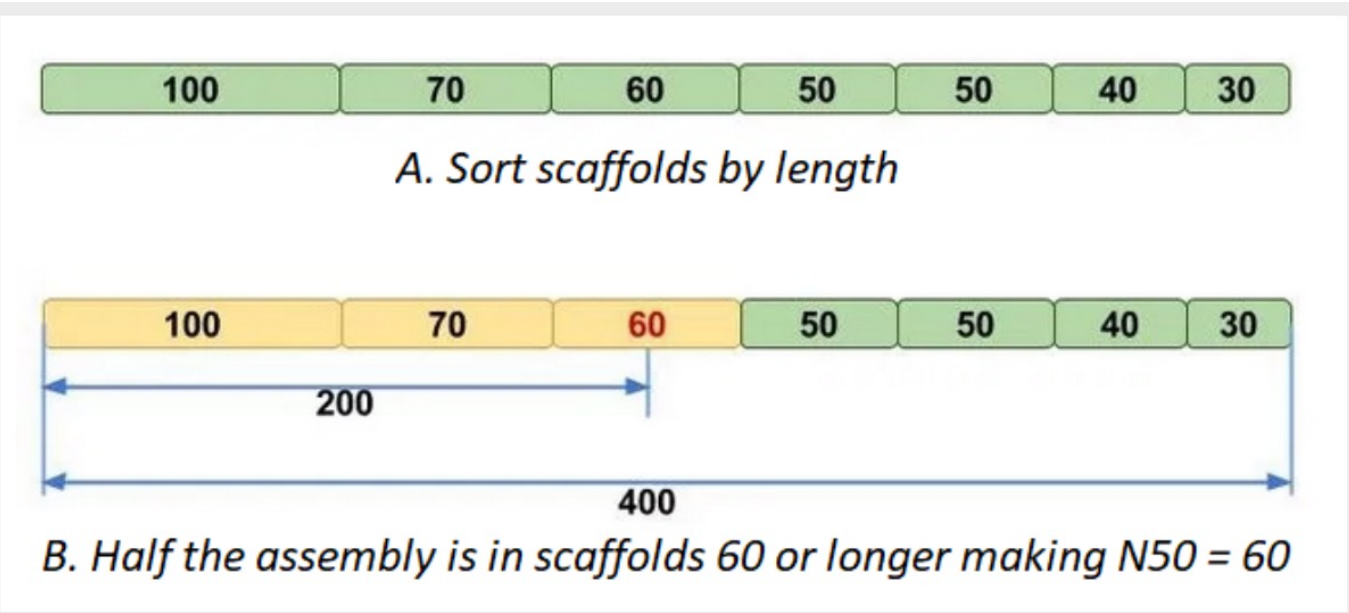
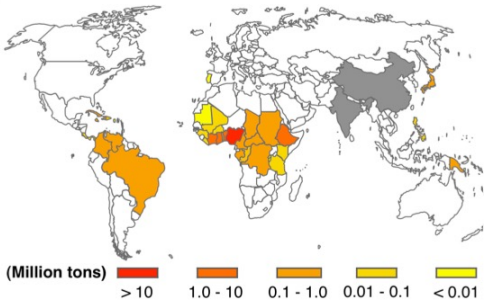


Figure 3.2: Scaffold N50. Source: Molecular Ecologist

Source: https://gwct.github.io/congen/assembly.html

Genome assembly

Example of yam genomes



Tamiru *et al. BMC Biology* (2017) 15:86
DOI 10.1186/s12915-017-0419-x

BMC Biology

RESEARCH Open Access

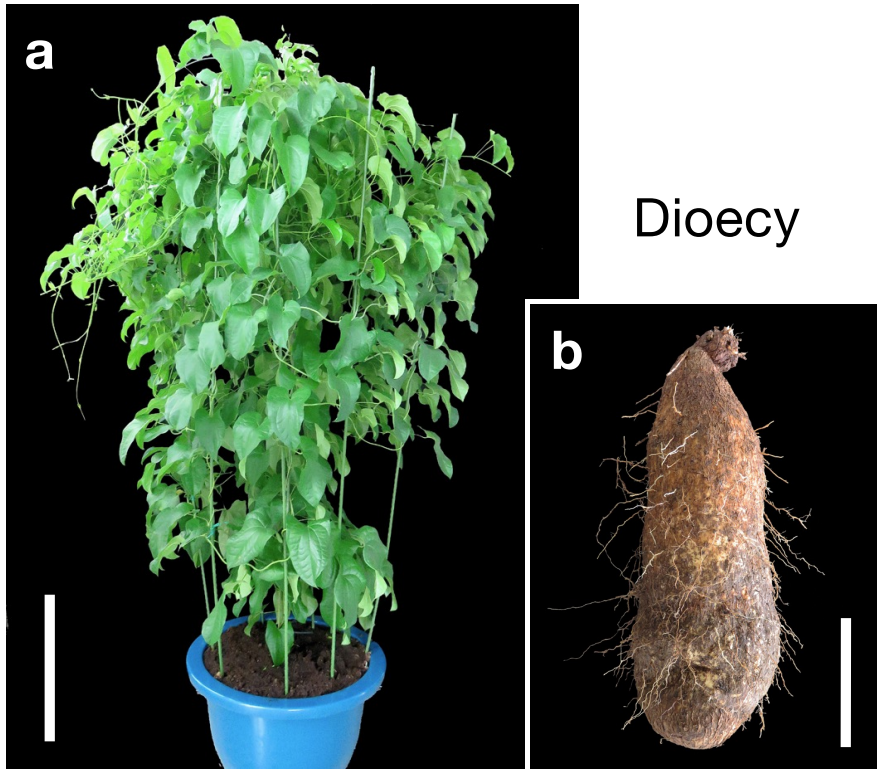
Genome sequencing of the staple food crop white Guinea yam enables the development of a molecular marker for sex determination

Muluneh Tamiru[1†], Satoshi Natsume[1†], Hiroki Takagi[1†], Benjamen White[2†], Hiroki Yaegashi[1†], Motoki Shimizu[1†], Kentaro Yoshida[3], Aiko Uemura[1], Kaori Oikawa[1], Akira Abe[1], Naoya Urasaki[4], Hideo Matsumura[5], Pachakkil Babil[6], Shinsuke Yamanaka[7], Ryo Matsumoto[7], Satoru Muranaka[7], Gezahegn Girma[8], Antonio Lopez-Montes[8], Melaku Gedil[8], Ranjana Bhattacharjee[8], Michael Abberton[8], P. Lava Kumar[8], Ismail Rabbi[8], Mai Tsujimura[9], Toru Terachi[9], Wilfried Haerty[2], Manuel Corpas[2], Sophien Kamoun[10], Günter Kahl[11^], Hiroko Takagi[7*], Robert Asiedu[8*] and Ryohei Terauchi[1,12*]

Tamiru *et al.* 2017 BMC Biology 15:86

26,198 genes predicted

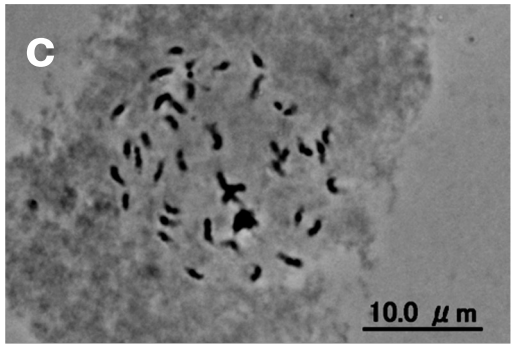
Guinea yam genome assembly



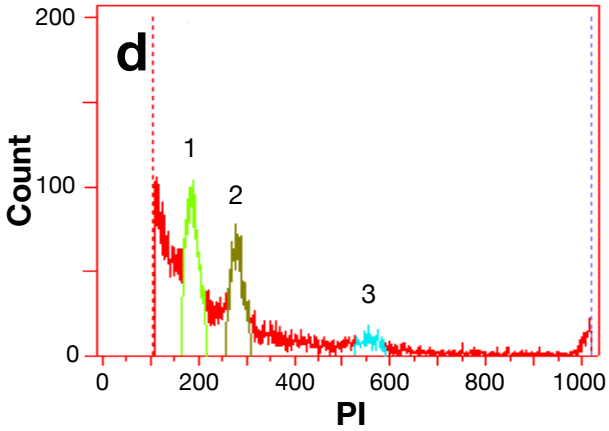
Guinea yam
(*Dioscorea rotundata*)

Dioecy

Tuber



Chromosome number = 40



Flow Cytometry

Genome size = 570 Mb

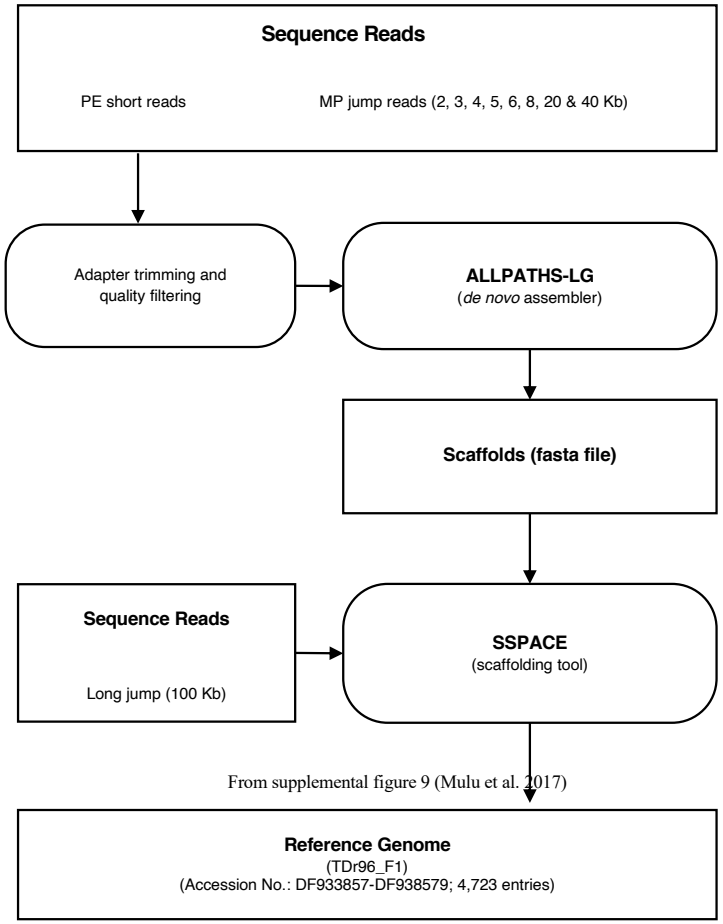
Summary of sequence reads generated to assemble *D. rotundata* genome

Name	Read type	Insert size (bp)	Read length (bp)	Total size (Gb)	Genome coverage	Accession No.
Fragment[a]	Paired-end	330	230	16.77	29.4×	DRX025239
Short jump (2 Kb)[b]	Mate-pair	2,000	100	13.85	24.3×	DRX025240
Short jump (3 Kb)[b]	Mate-pair	3,000	100	10.81	19.0×	DRX025241
Short jump (4 Kb)[b]	Mate-pair	4,000	100	9.98	17.5×	DRX025242
Short jump (5 Kb)[b]	Mate-pair	5,000	100	10.22	17.9×	DRX025243
Short jump (6 Kb)[b]	Mate-pair	6,000	100	7.27	12.8×	DRX025244
Short jump (8 Kb)[b]	Mate-pair	8,000	100	6.79	11.9×	DRX025245
Long jump (20 Kb)[c]	Mate-pair	20,000	100	4.10	7.2×	DRX025246
Long jump (40 Kb)[d]	Mate-pair	40,000	250	4.89	8.6×	DRX025247
Long jump (100 Kb)[e]	Mate-pair	100,000	50	0.46	0.8×	DRX025248
				85.14	**149.4**×	

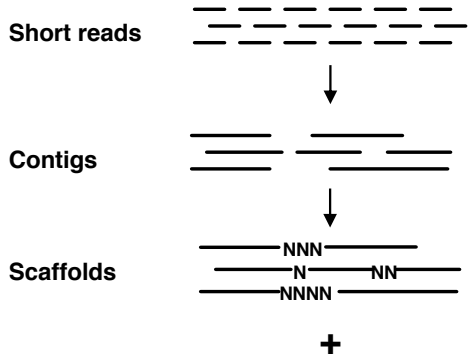
Tamiru et al. BMC Biology, 2017

Improvement of *D. rotundata* genome assembly by MP2 linkage data:

TDr96_F1 reference sequence (Tamiru *et al.* 2017, BMC Biology)



From supplemental figure 9 (Mulu et al. 2017)



\+

RAD-seq (2,598 markers)

TDr96_F1 Pseudo_Chromosome_v1

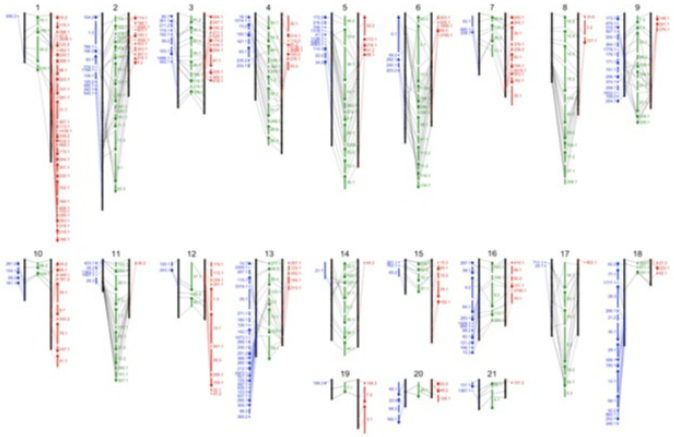
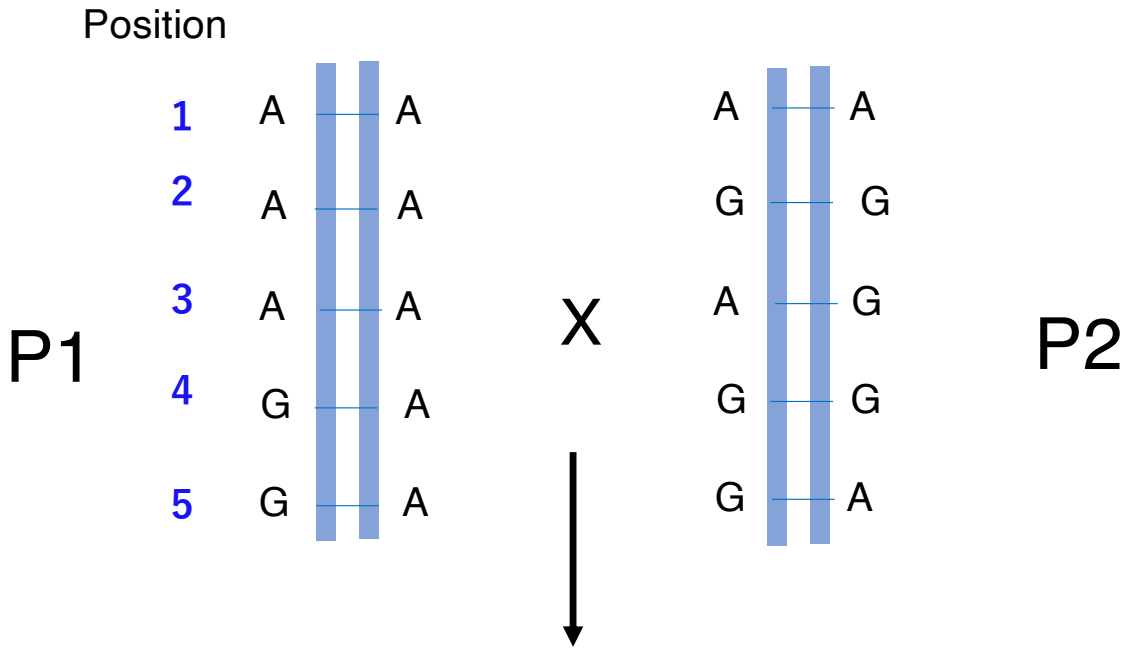


Table 1 Characteristics of nuclear genome sequence in *Dioscorea rotundata* and other angiosperms

Feature	Value			
	D. rotundata (v0.1)	*A. thaliana* (TAIR10)	*B. distachyon* (v3.1)	*O. sativa* (v7_JGI 323)
Total length (Mbp)	594.23	119.67	271.16	374.47
GC (%)	35.83	36.06	46.40	43.57
Number of scaffolds (≥ 0 bp)	4723	7	10	14
Number of scaffolds (≥ 1000 bp)	4704	7	10	14
Largest scaffold (Mbp)	13.61	30.43	75.07	43.27
N50 (Mbp)	2.12	23.46	59.13	29.96
N75 (Mbp)	0.77	19.70	48.59	28.44
Number of Ns per 100 kb	282.45[a]	155.60	155.85	44.13
Ambiguous bases	1,413,029	–	–	–
Number of genes	26,198	27,416	34,310	42,189

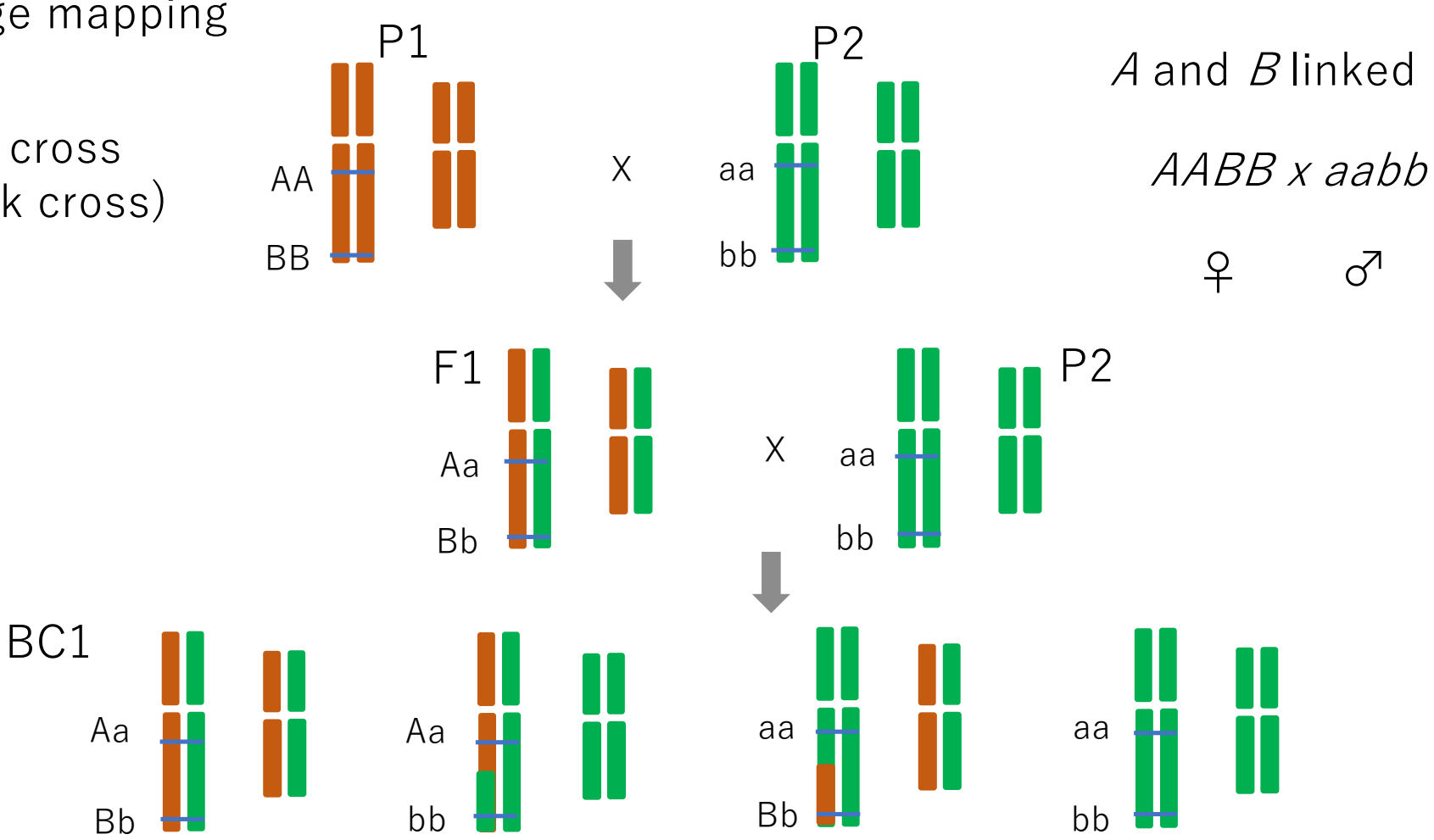
Pseudo-testcross of heterozygous parents allows mapping in F1 generation



F1 individual: 1, 2, 3, 4, 5 ……

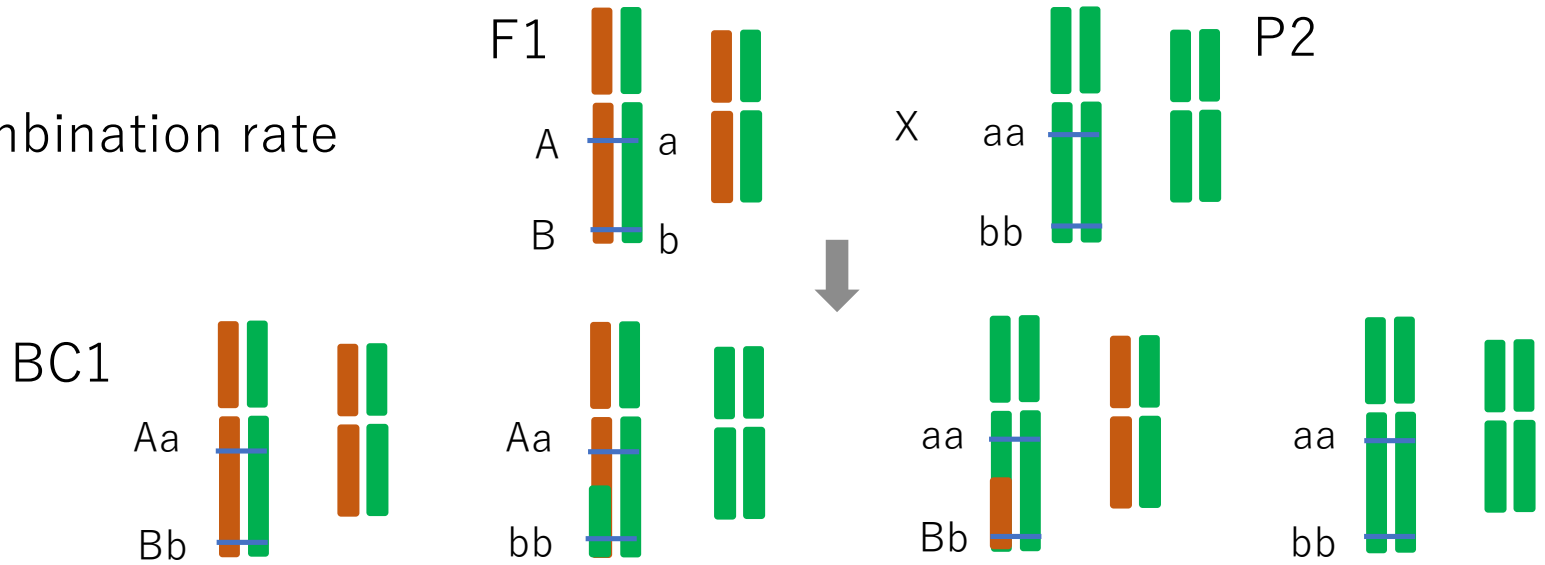
Linkage mapping

Test cross
(Back cross)



AaBb : Aabb: aaBb: aabb → 1 : <1 : <1 : 1

Calculation of recombination rate



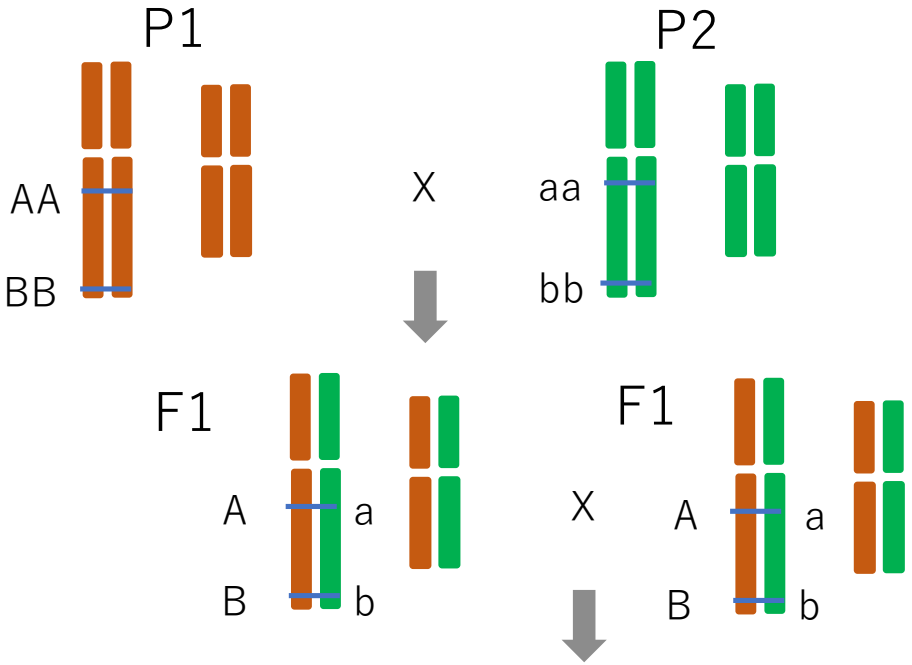
Female gamete		AB (1-r)/2	Ab r/2	aB r/2	aa (1-r)/2
Male gamete	ab	Zygote: AaBb (1-r)/2	Aabb r/2	aaBb r/2	aabb (1-r)/2

r = 0 → Complete linkage

r = 0.5 → Independent segregation

Linkage mapping

F2 population



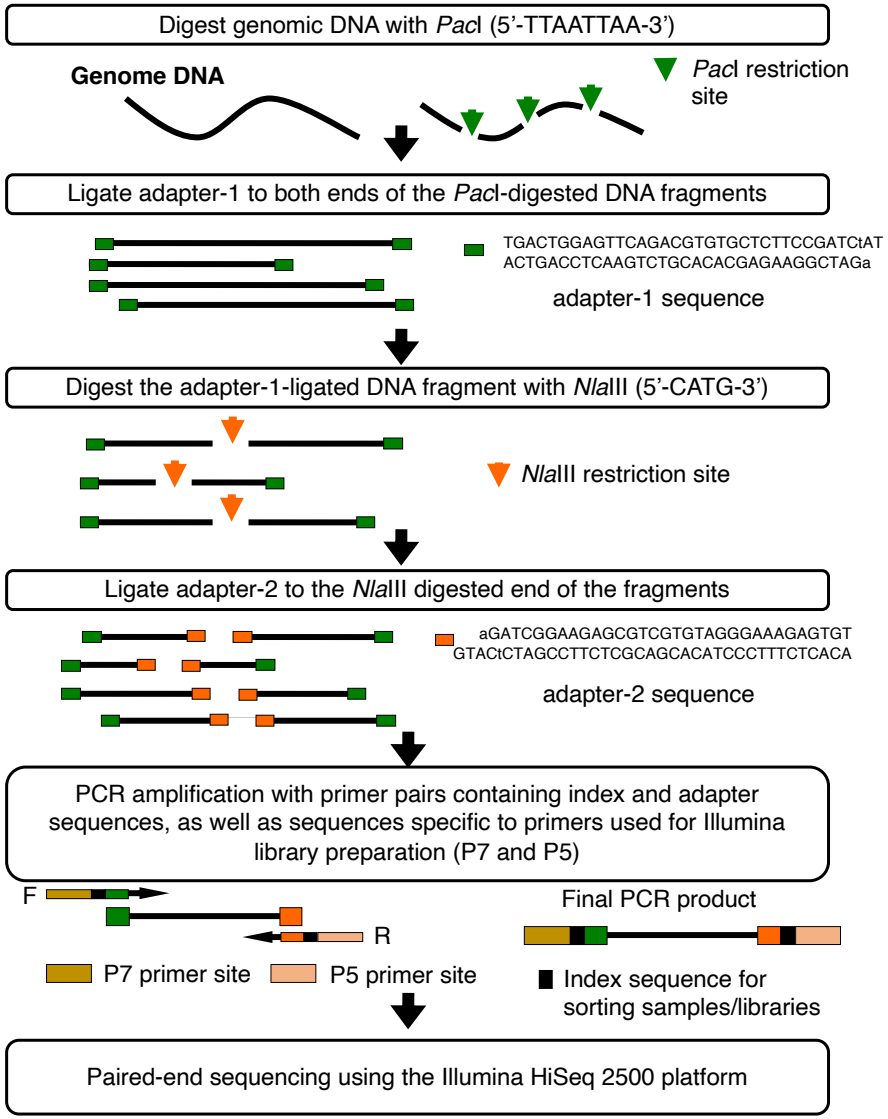
F2

♀ \ ♂	AB	Ab	aB	ab
AB	AABB	AABb	AaBB	AaBb
Ab	AABb	Aabb	AaBb	Aabb
aB	AaBB	AaBb	aaBB	aaBb
ab	AaBb	Aabb	aaBb	aabb

	Ratio in non-linkage
A-B-	9
A-bb	3
aaB-	3
aabb	1

RAD seq

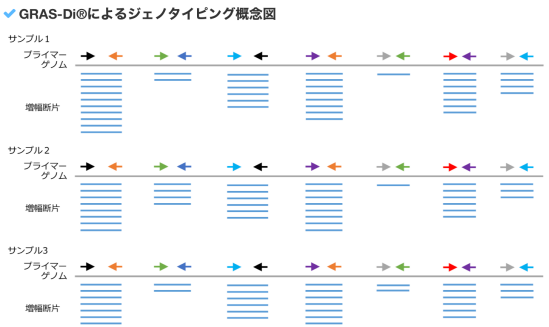
Baird et al. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One. 3:e3376.



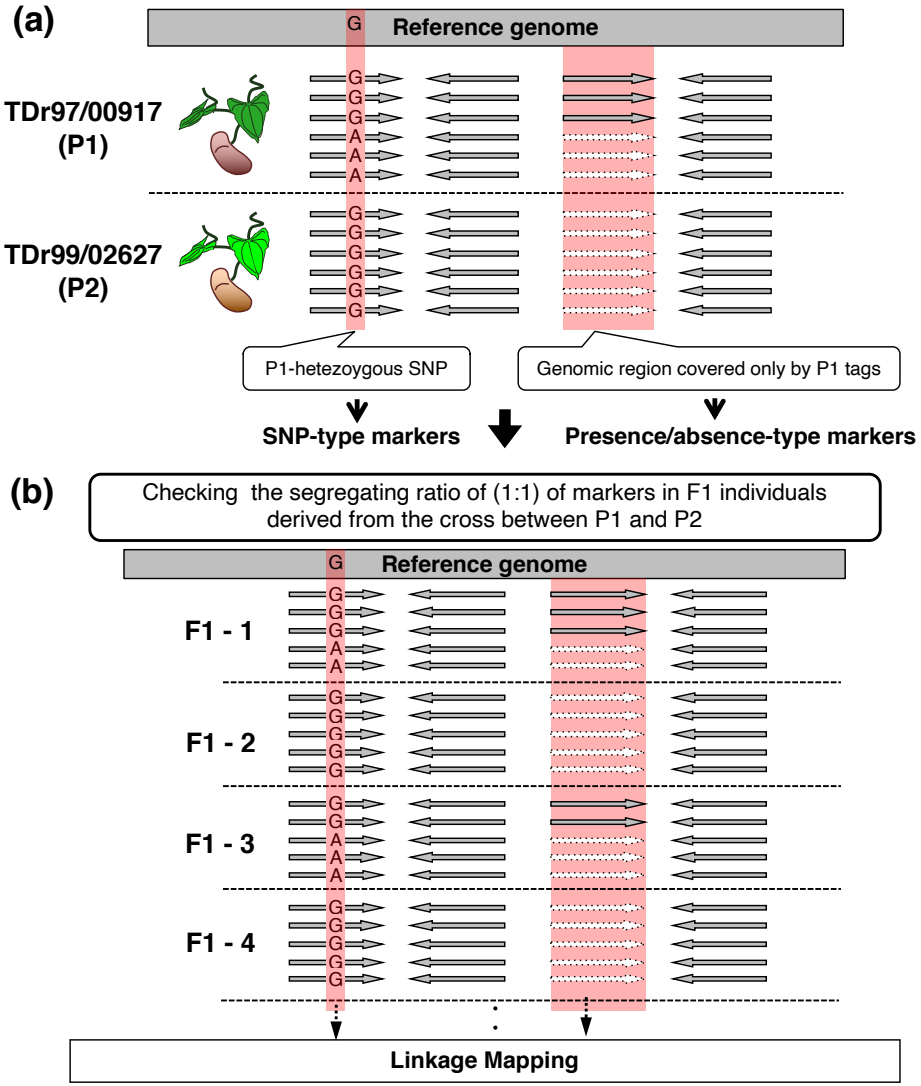
GBS

Elshire et al. (2011-05-04). "A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species". *PLOS ONE*. **6** (5): e19379.

GRAS-Di



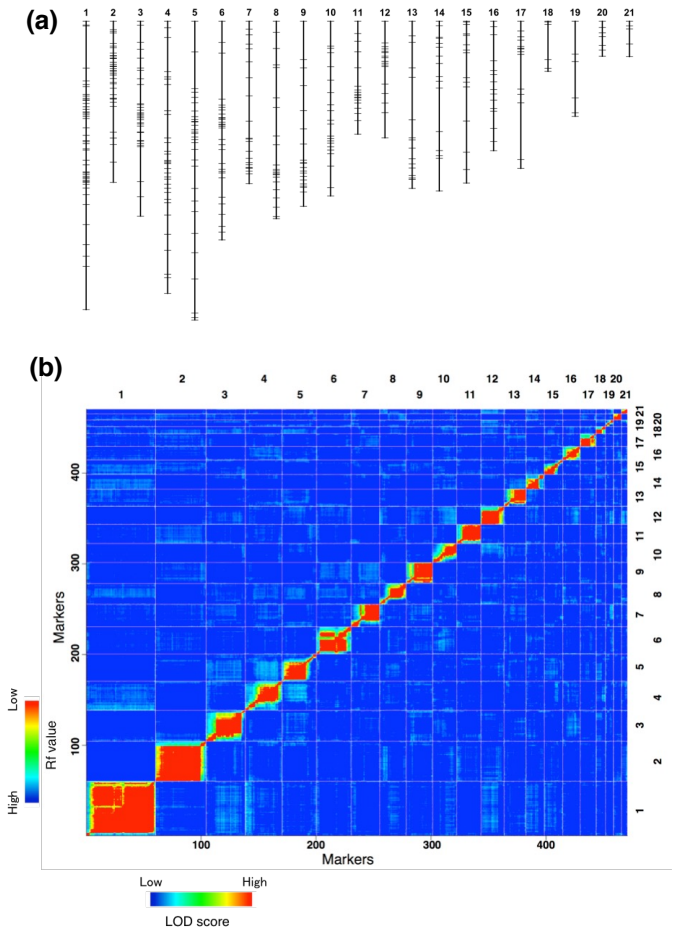
Linkage mapping by pseudo testcross scheme



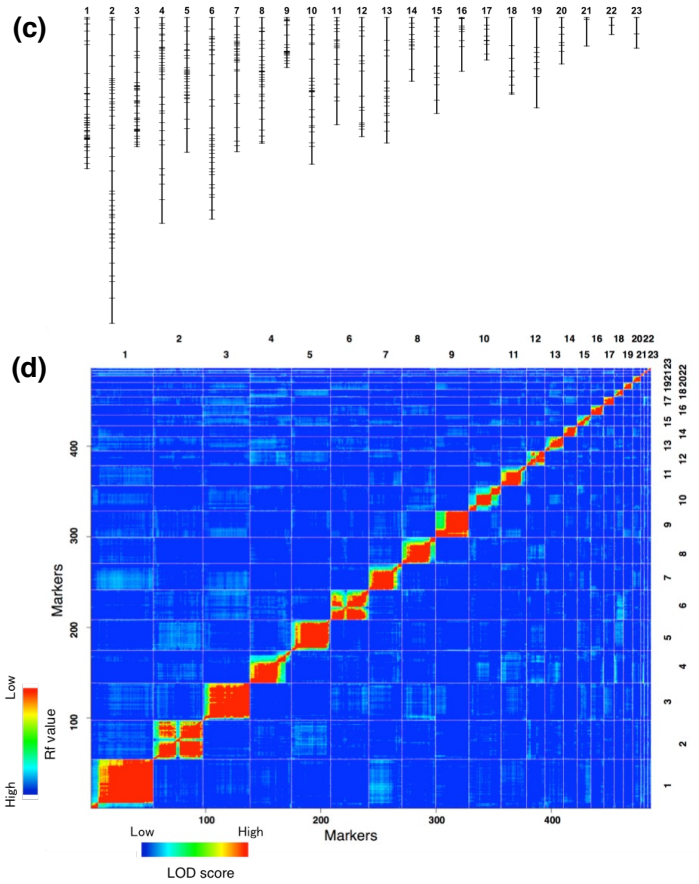
1326 and 1272 markers for P1 and P2 heterozygous sites

150 F1 progeny

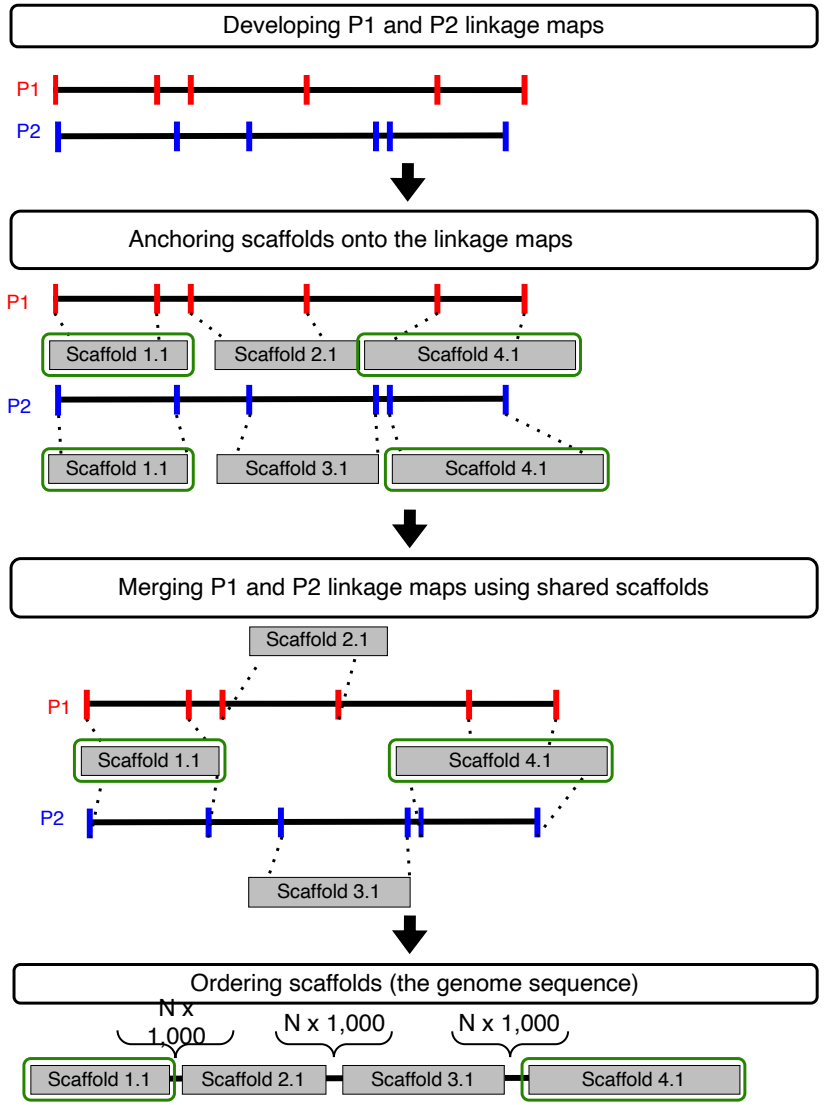
P1 heterozygous markers map



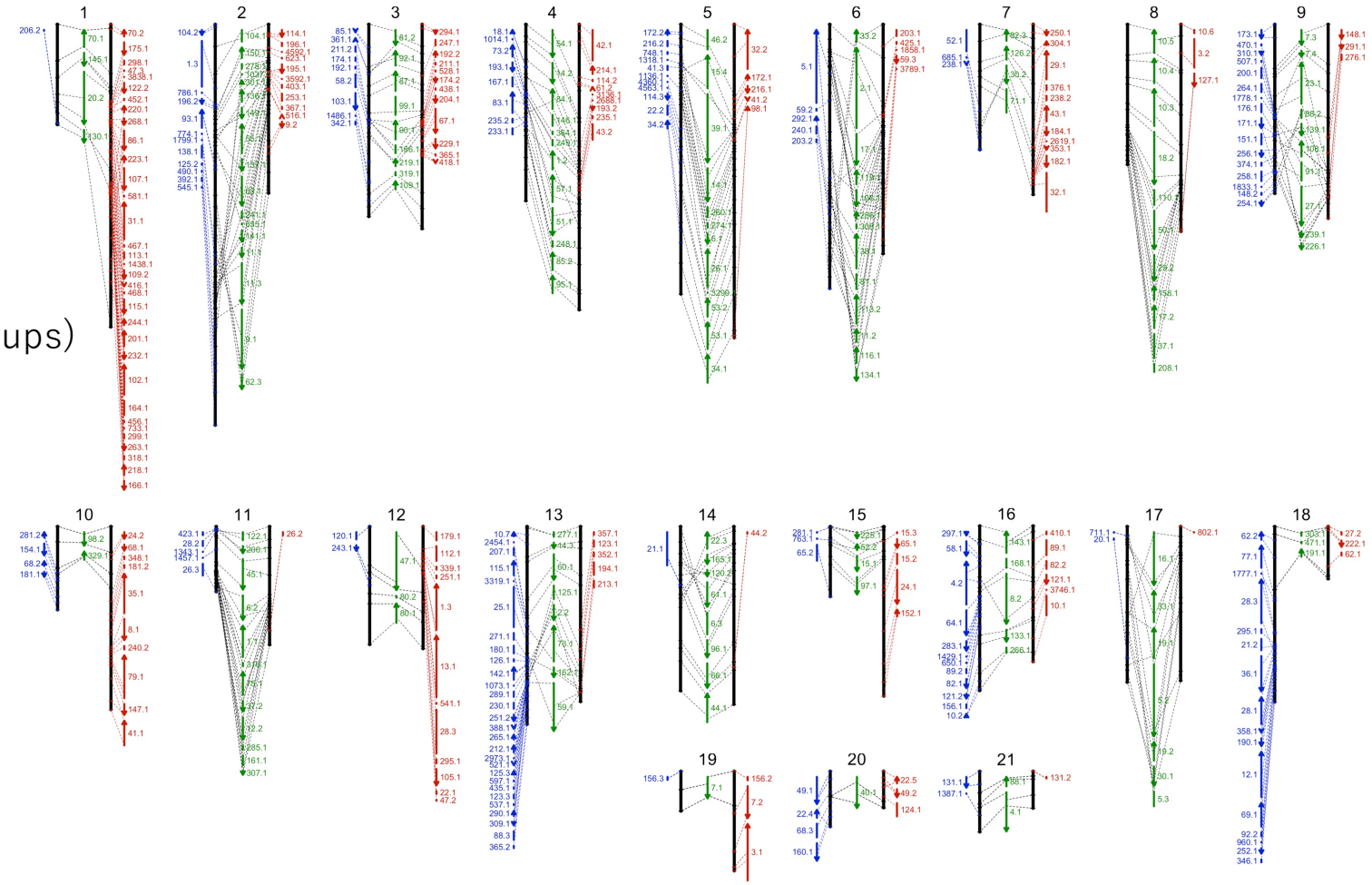
P2 heterozygous markers map



Combining two linkage maps by shared scaffolds



Pseudo
chromosomes
(21 linkage groups)



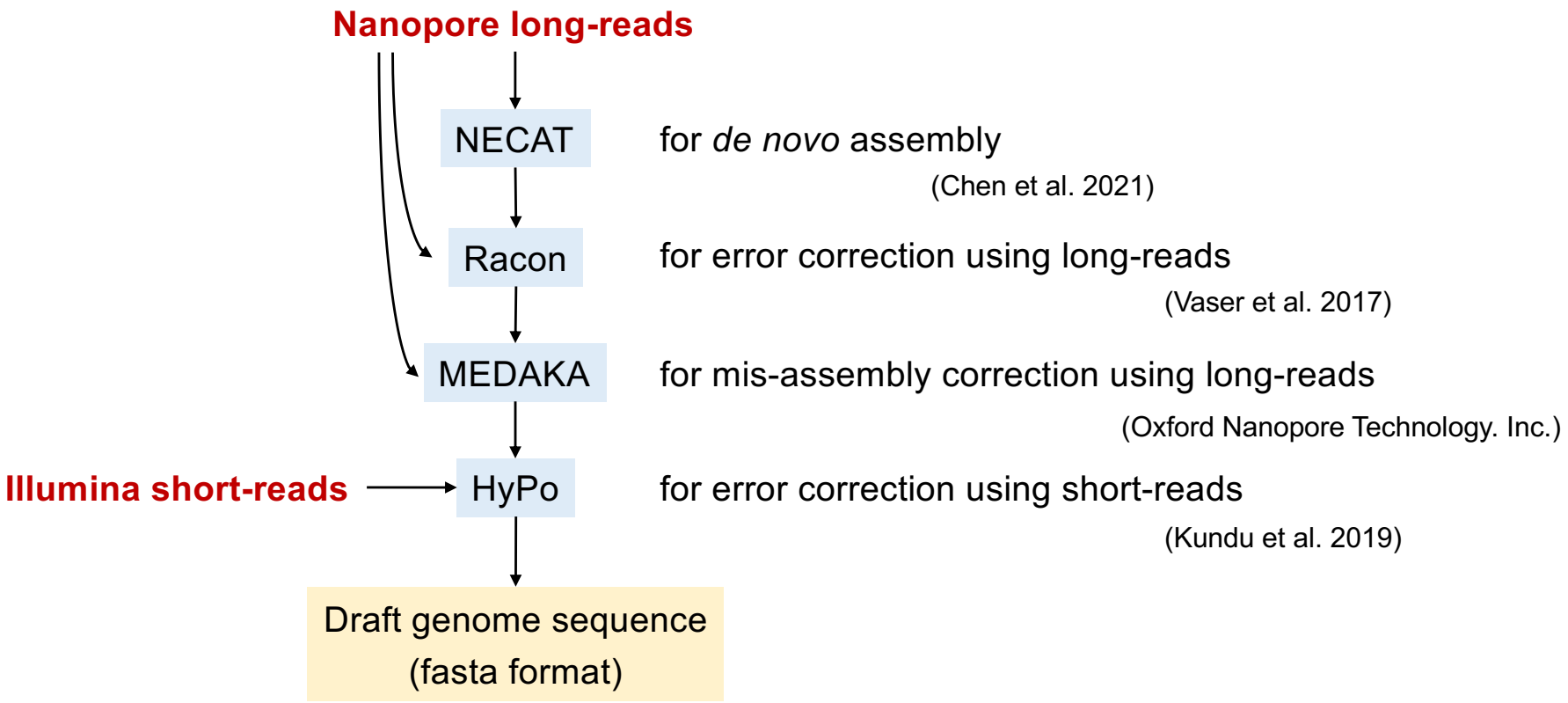
de novo assembly using Nanopore long-reads

MutMap and QTL-seq pipelines require reference genome sequence.

What if there is no reference genome sequence for your crop of interest ?

You can make it by yourself.

A workflow of *de novo* genome assembly



Rice (*Oryza sativa* L. ssp. indica)

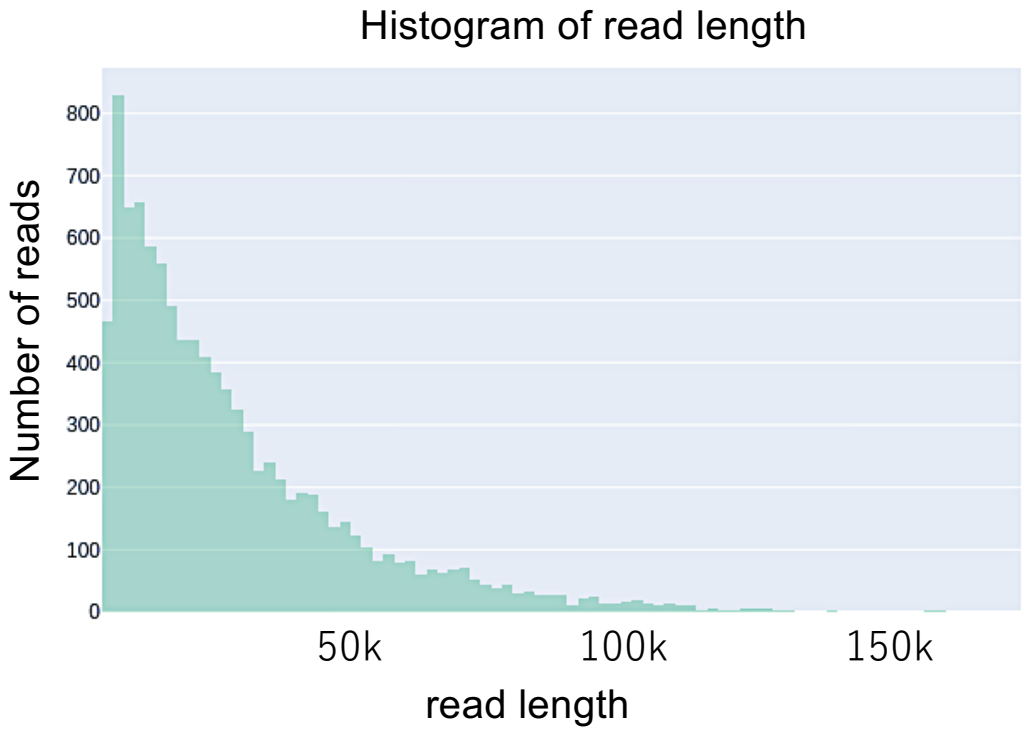
2n=24

Genome size = 380 Mbp

12,555 Mb / 380 Mb = 33

Summary of Nanopore reads

Total bases	12,555,082,810
Number of reads	445,931
Mean read length	28,155
Median read length	21,350
Longest read length	304,488



Rice (*Oryza sativa* L. ssp. indica)

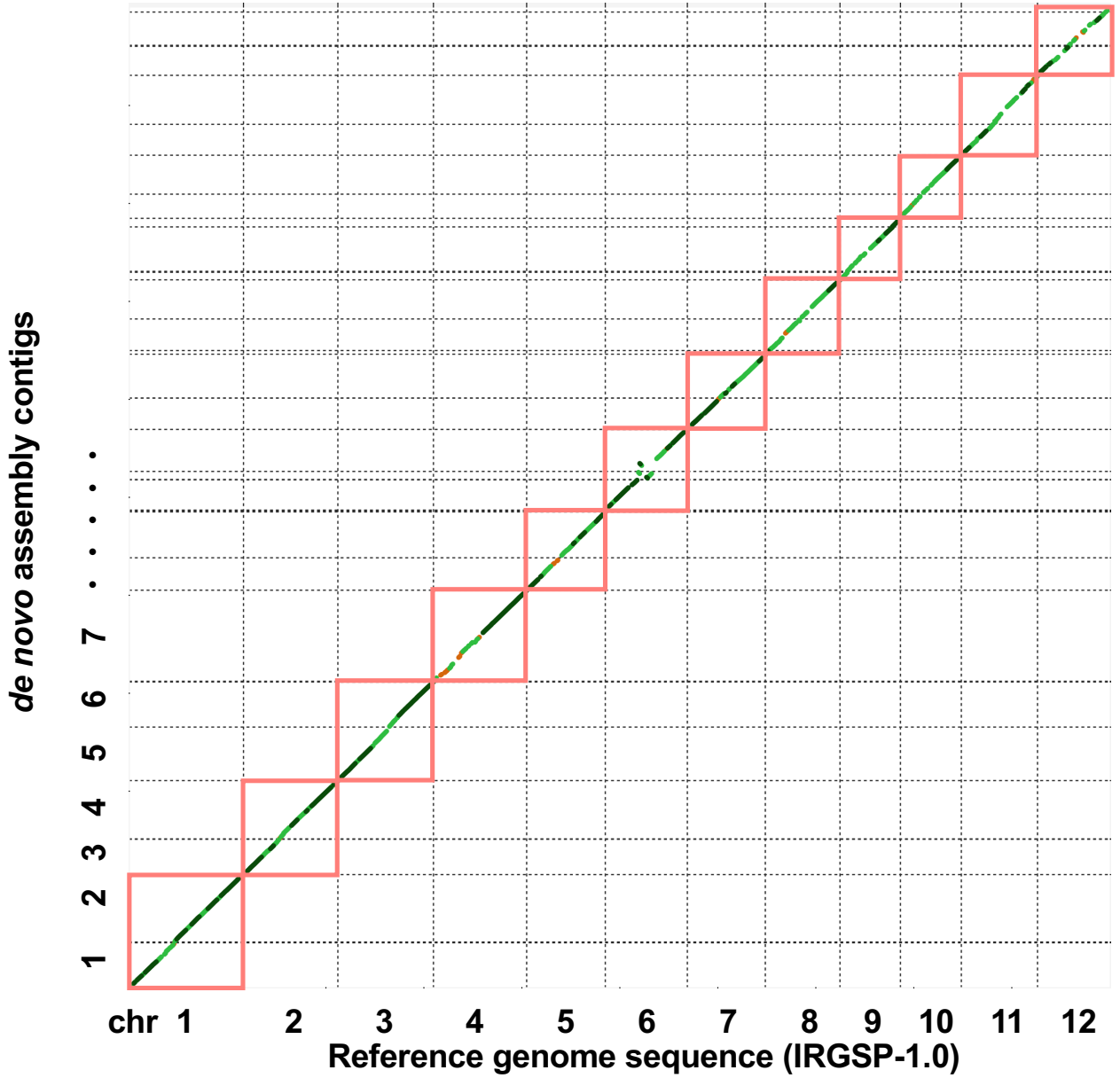
Obtained reads

Nanopore long-reads	12.5 Gb (33x genome size)
Illumina short-reads 2x150 PE	17.4 Gb (47x genome size)

Summary of *de novo* assembly

Number of contig	42
Sum of length	387,333,531
Minimum length	25,233
Maximum length	35,993,077
Average length of contigs	9,222,227

Dot plot de novo assembly contigs against the rice reference genome sequence using D-genies software (Cabanettes and Klopp, 2018)

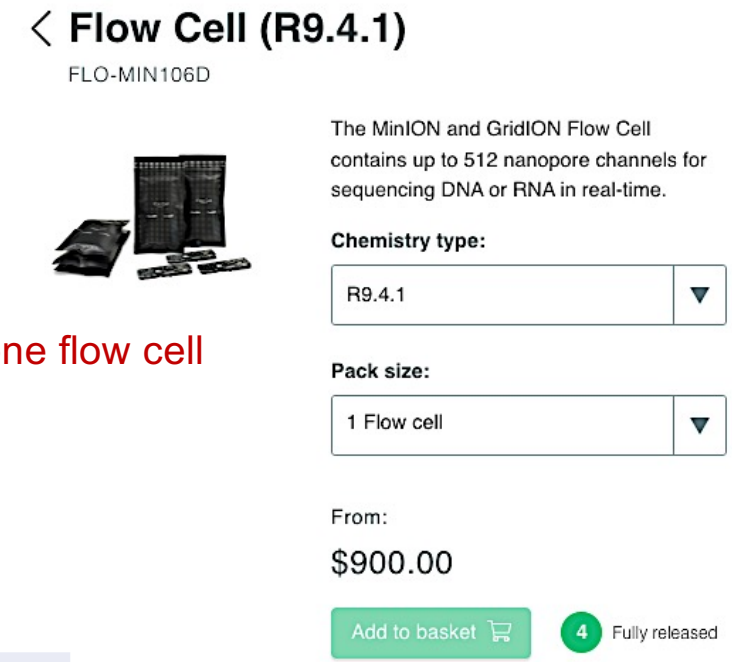


Rice (*Oryza sativa* L. ssp. indica)

Obtained reads

Nanopore long-reads 12.5 Gb (33x genome size) from only one flow cell

Illumina short-reads 17.4 Gb (47x genome size)



（https://nanoporetech.com/）

Summary of *de novo* assembly

Number of contig	42
Sum of length	387,333,531
Minimum length	25,233
Maximum length	35,993,077
Average length of contigs	9,222,227

→ Chromosome 4

Genome analyses reveal the hybrid origin of the staple crop white Guinea yam (*Dioscorea rotundata*)

Yu Sugihara, Kwabena Darkwa, Hiroki Yaegashi, +20, and Ryohei Terauchi Authors Info & Affiliations

Edited by Loren H. Rieseberg, University of British Columbia, Vancouver, BC, Canada, and approved November 2, 2020 (received for review July 27, 2020)

December 2, 2020 | 117 (50) 31987-31992 | https://doi.org/10.1073/pnas.2015830117

Article | Open Access | Published: 07 January 2022

Recombinant inbred lines and next-generation sequencing enable rapid identification of candidate genes involved in morphological and agronomic traits in foxtail millet

Kenji Fukunaga, Akira Abe, Yohei Mukainari, Kaho Komori, Keisuke Tanaka, Akari Fujihara, Hiroki Yaegashi, Michie Kobayashi, Kazue Ito, Takanori Ohsako & Makoto Kawase

Scientific Reports **12**, Article number: 218 (2022) | Cite this article

Genome Analysis Revives a Forgotten Hybrid Crop Edo-dokoro in the Genus *Dioscorea*

Satoshi Natsume, Yu Sugihara, Aoi Kudoh, Kaori Oikawa, Motoki Shimizu, Yuko Ishikawa, Masahiro Nishihara, Akira Abe, Hideki Innan, Ryohei Terauchi

Plant and Cell Physiology, pcac109, https://doi.org/10.1093/pcp/pcac109

Published: 25 July 2022 **Article history ▾**