Molecular markers in plant systematics and population biology

9. Next-generation sequencing (NGS)

Tomáš Fér

tomas.fer@natur.cuni.cz

Next generation sequencing (NGS)

- first generation – Sanger sequencing
- second generation – parallel sequencing of many molecules (PCR amplified)
- third generation (further generations) – single molecule sequencing

General protocol for NGS

- library preparation
  - random shearing of genomic DNA to the fragments
  - adaptor ligation
- spatial separation of individual fragments
- two „basic“sequencing options
  - sequencing of clonally amplified fragments
    - emulsion PCR (emPCR)
    - solid-phase amplification (bridge PCR)
    - rolling circle amplification (RCA)
  - single-molecule real-time sequencing (SMRT)
- immobilization to the surface
- sequencing and data acquisition
  - sequencing by synthesis
    - pyrosequencing (Roche/454)
    - cyclic reversible termination (CRT) (Illumina/Solexa)
    - semiconductor chip (Ion Torrent)
  - sequencing by ligation
    - (SOLiD)
    - combinatorial Probe-Anchor Synthesis (cPAS) (MGI/Complete Genomics)
- data analysis (analysis of image data, quality control, …)

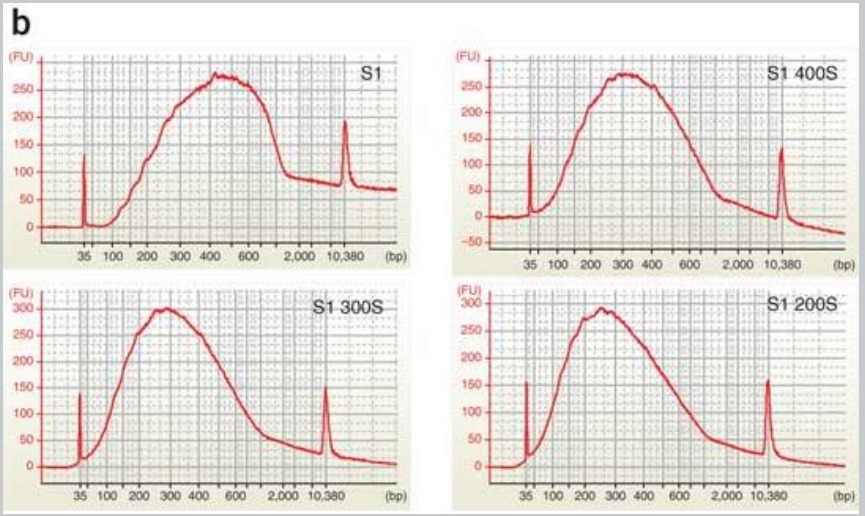
Prevalent NGS platforms

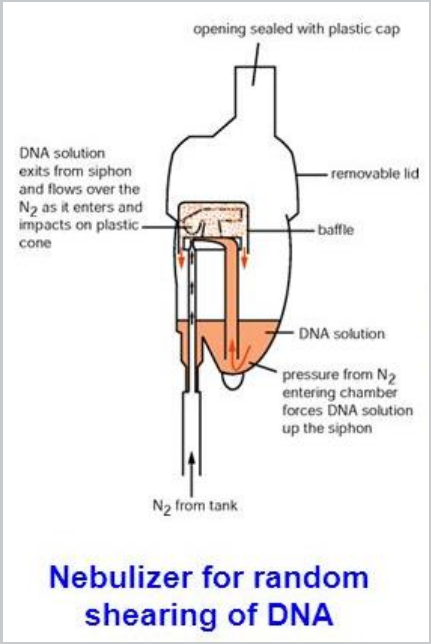
- Roche/454 – emPCR, pyrosequencing
- Illumina/Solexa – solid phase (bridge) PCR, CRT
- Life/APG (SOLiD) – emPCR, ligation
- Pacific Biosciences – single molecule real time (SMRT)
- Ion Torrent – emPCR, semiconductor chips
- Oxford Nanopore – single molecule
- BGI/MGI – nanoball, cPAS

DNA shearing

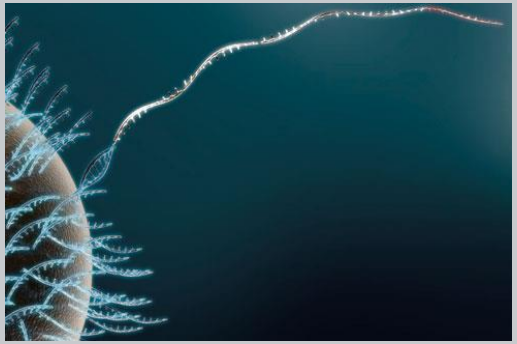
- sonication

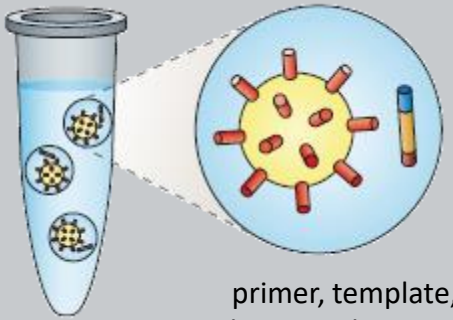
- nebulization



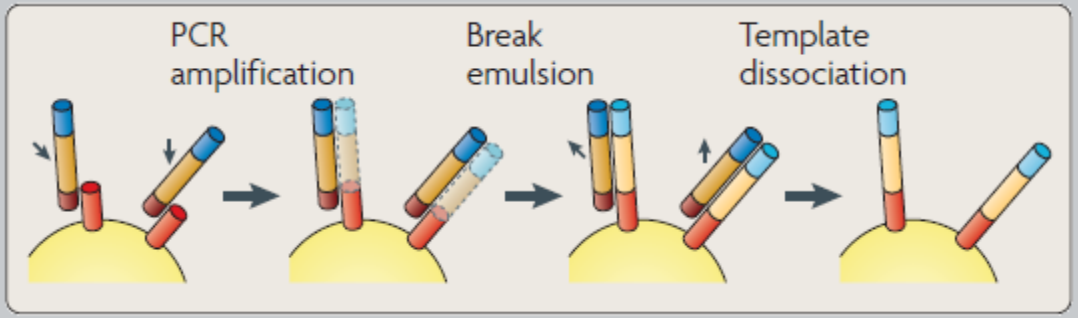


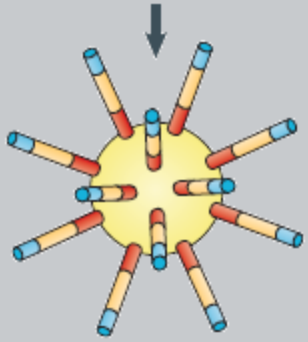
Emulsion PCR

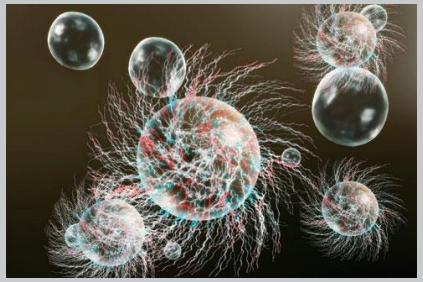


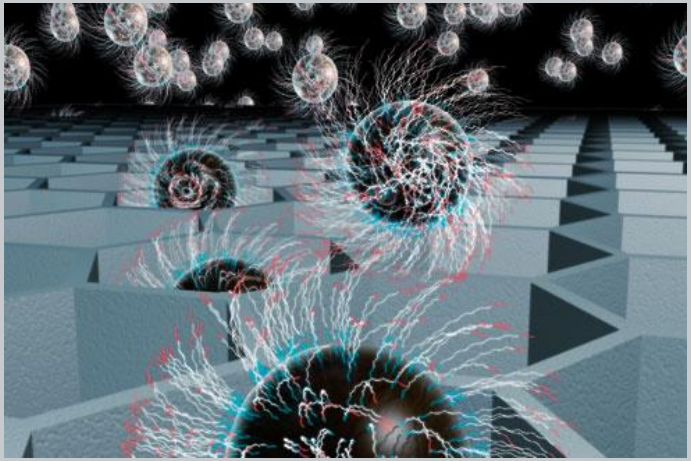


primer, template,
dNTPs, polymerase



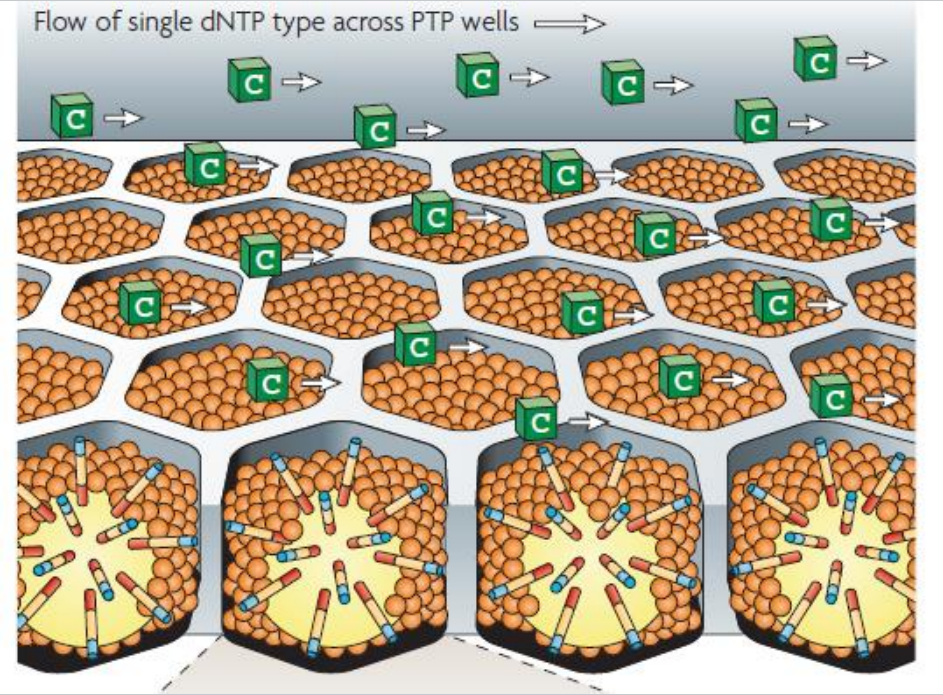


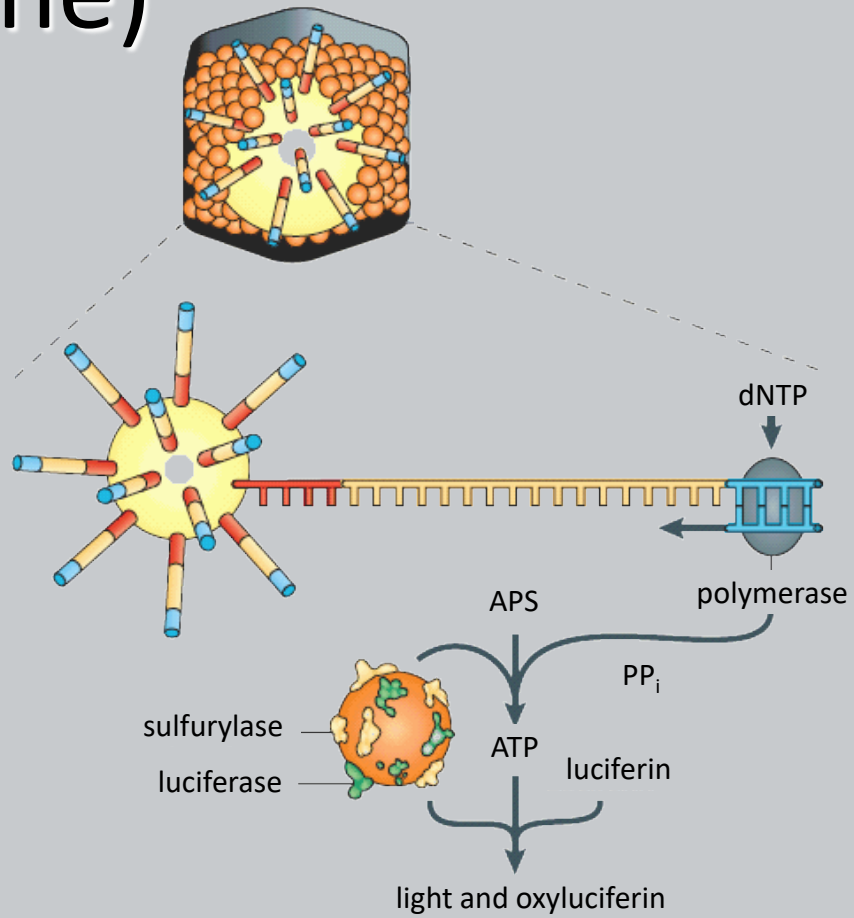




Metzker 2010

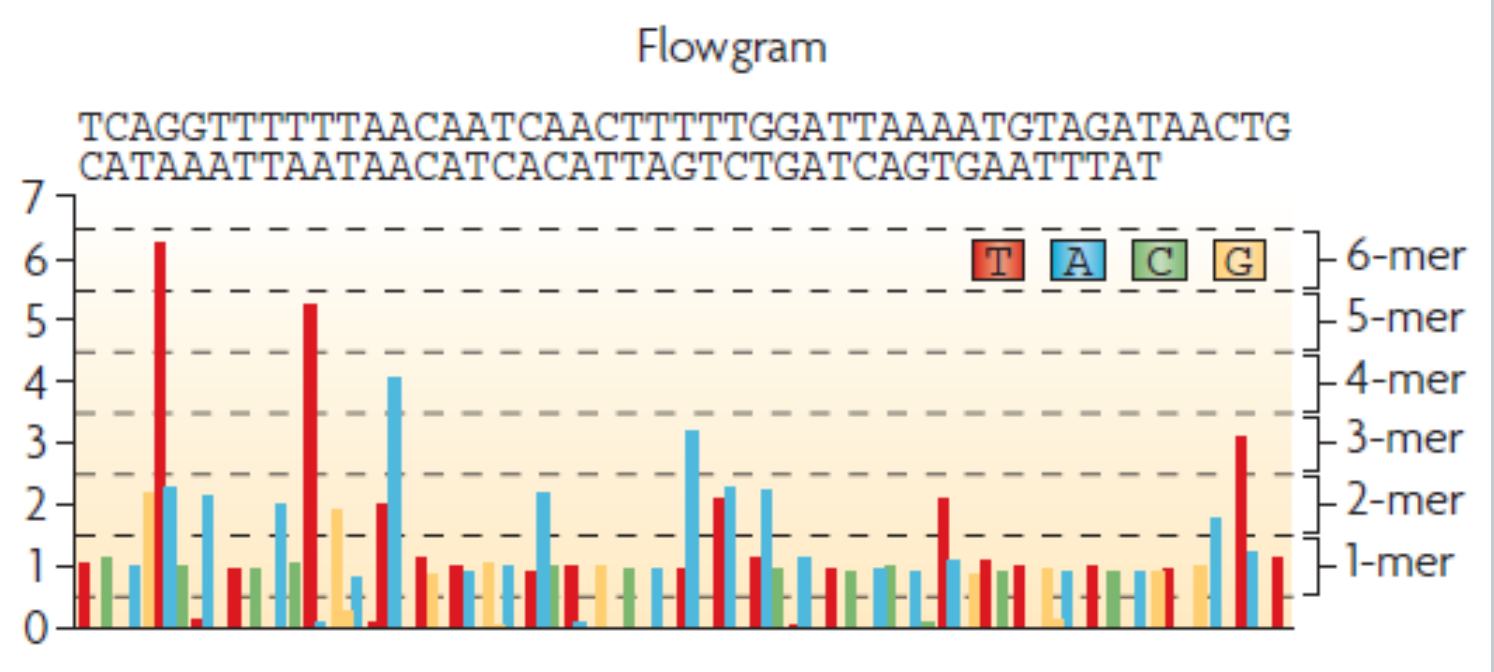
Pyrosequencing (Roche)





Metzker 2010

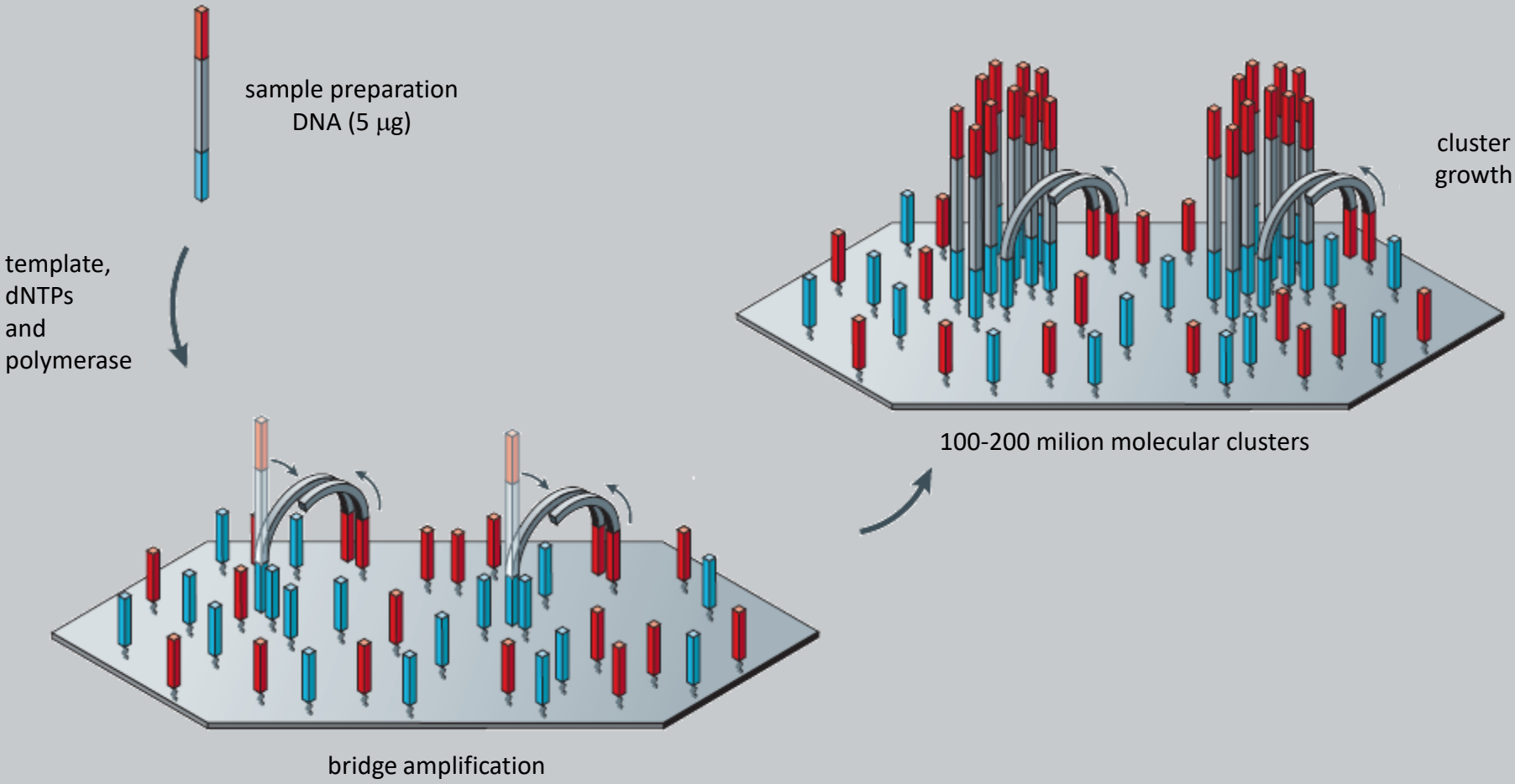
Pyrosequencing (Roche)



instruments – GS Junior, GS FLX Titanium

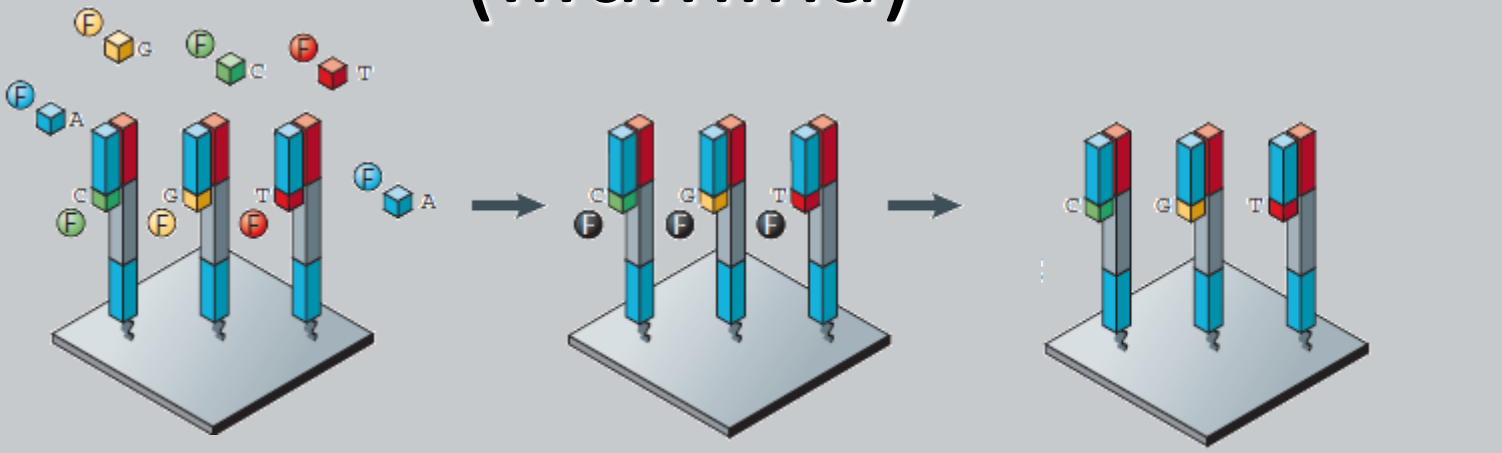
Metzker 2010

Solid-phase amplification (Illumina)



Metzker 2010

Cyclic reversible termination (Illumina)



incorporate all four nucleotides, each label with a different dye

wash, four-colour imaging

cleave dye and terminating groups, wash

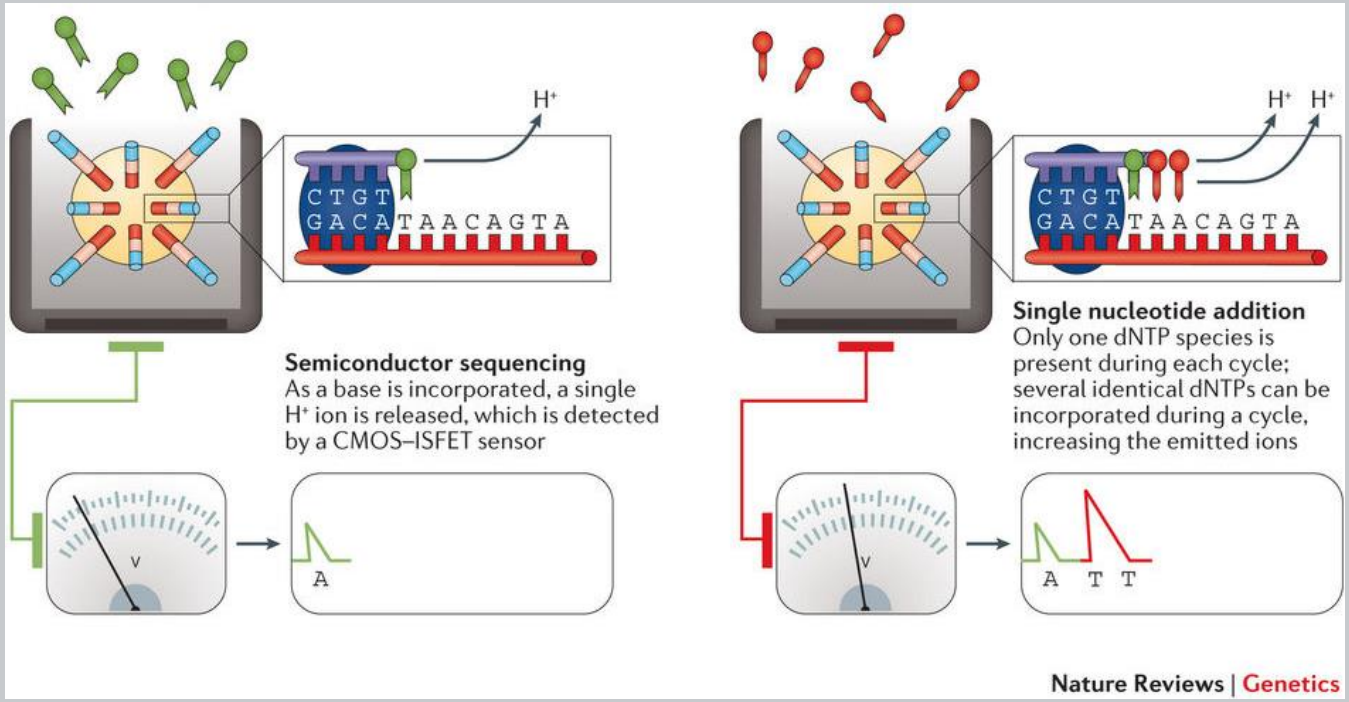
Top: CATCGT
Bottom: CCCCCC

instruments – MiniSeq, MiSeq, NextSeq, HiSeq, NovaSeq

Metzker 2010

Semiconductor sequencing (IonTorrent)

- emulsion PCR
- addition of dNTP releases H^+ which is measured as a change of conductivity

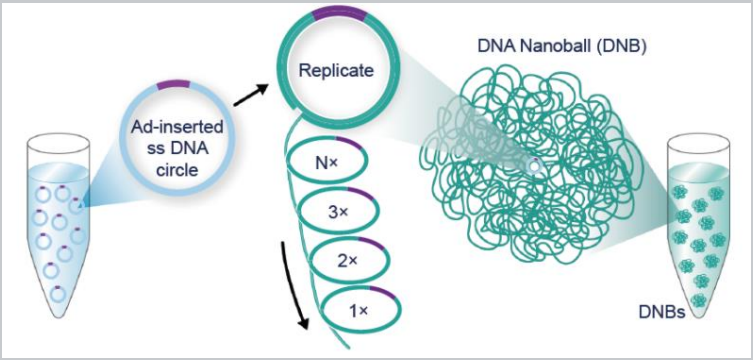


- instruments – Ion PGM, Ion Proton, Ion S5, Ion S5 XL, Genexus GX5

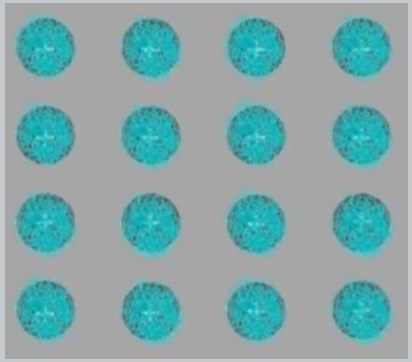
Goodwin et al. 2016

DNA Nanoball (DNB) sequencing (MGI/BGI/Complete Genomics)

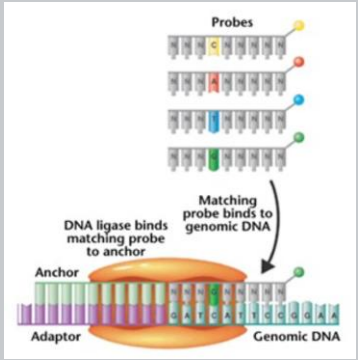
- DNA fragments circularized – ssCirDNA
- rolling circle amplification (RCA) – DNB generation (with Phi 29 DNA polymerase)
- DNB loaded to form patterned array
- sequencing by synthesis – cPAS (combinatorial Probe-Anchor Synthesis)



RCA



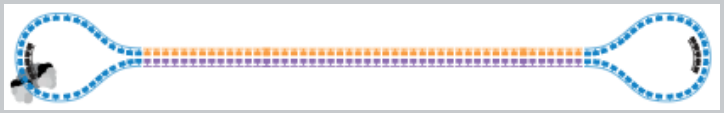
patterned array



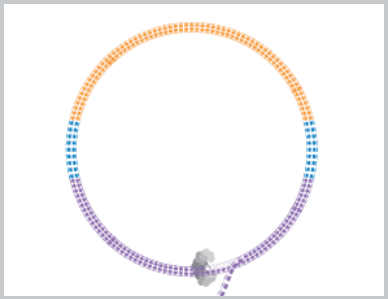
cPAS

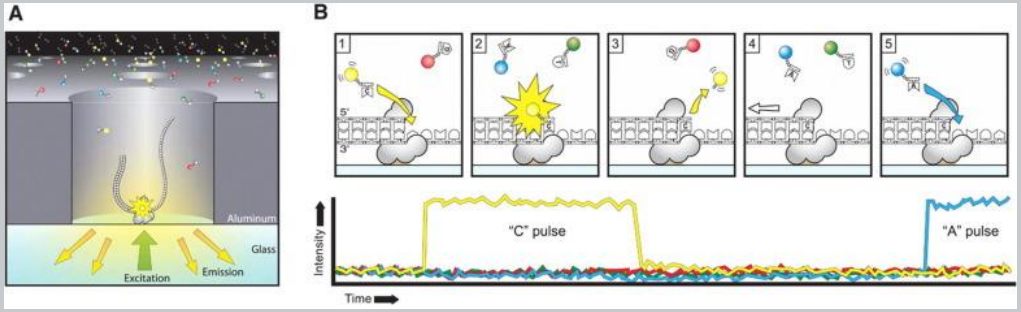
instruments – BGISEQ-500, DNBSEQ-G50, DNBSEQ-G400, DNBSEQ-T7

Single-molecule real-time (SMRT) (PacBio)

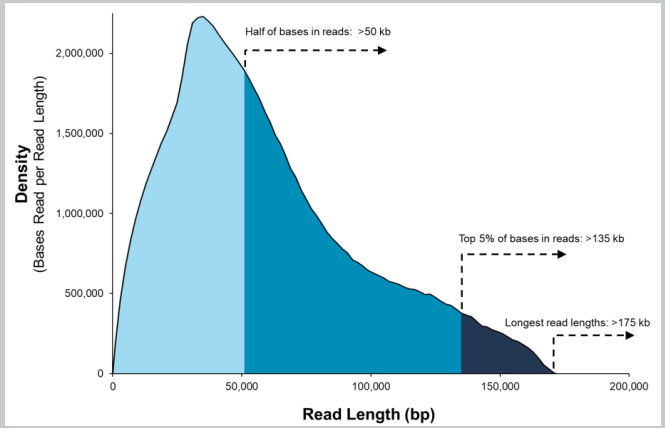


- library preparation – DNA is circularized, no amplification (SMRTbell library)
- zero mode waveguide (ZMW) – DNA polymerase affixed to the bottom of a tiny hole (~70 nm)
- light signal is emitted if a phospholinked nucleotide is incorporated





- high single pass error rate (~10-15%)
- CLR – continuous long read sequencing – >50 kb
- HiFi reads (<0.1%) – consensus of subreads (10-15) CCS – circular consensus sequencing) – 1-20 kb
- long reads – 50% of reads longer than 20,000 bp
- instruments – Sequel System, PacBio RS II, Sequel II, Sequel IIe

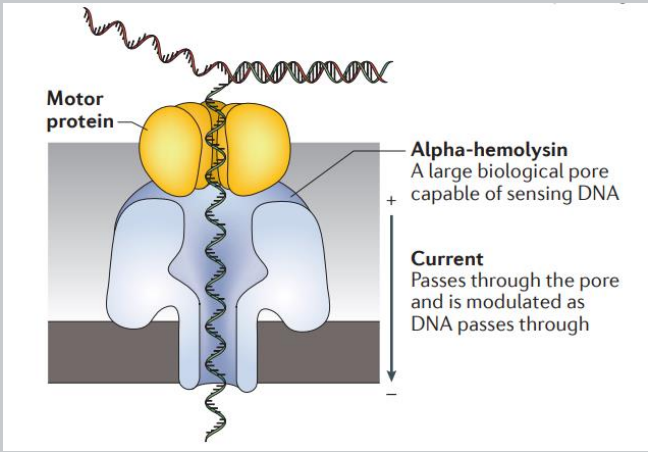


Rhoads & Au 2015

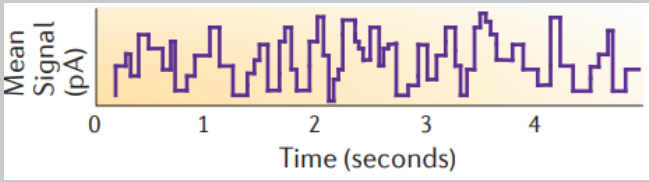
Single molecule sequencing (Oxford Nanopore)

- library preparation – leader-hairpin template: the leader protein interacts with the pore

- DNA is translocating through the (nano)pore



- shifts in electric current (*squiggle space*) corresponds to a particular k-mer (3-6 bases; more than 1,000)



- instruments – MinION, GridIONx5, PromethION

Goodwin et al. 2016

Platform comparison

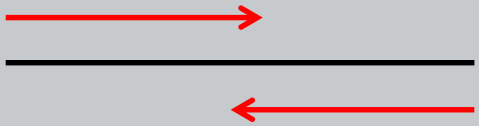
Platform (sequencers)	**Template preparation**	**Chemistry**	**Max read length (bases)**	**Run time (days)**	**Gb per single run / $ per Gbase**	**Error rate (single pass / final)**	**Advantages**	**Disadvantages**
Roche/454 (GS Jr., FLX)	emPCR	sequencing by synthesis (pyrosequencing)	400-650	0.35-0.9	0.05-0.65 9,000-19,000	1-1.7 1-1.7	long reads, quick run	high costs per Mb, high error rate in homopolymers
Illumina/ Solexa (GAII, iSeq, MiSeq, NextSeq, HiSeq, NovaSeq)	solid-phase bridge PCR	sequencing by synthesis (cyclic reversible termination)	75-300 (2x300)	0.8-11	4.5-500 7-220	0.003-1 0.003-1	broadly available, low error rate, cloud data analysis	limited multiplexing level?
Life/APG (SOLiD 5500xl)	emPCR	sequencing by ligation	110	8	155 70	5 0.01-1	high accuracy	relatively short reads, uneven data distribution (A-T bias)
Ion Torrent (PGM, Proton, Ion S5, Genexus GX5)	emPCR	no chemistry (semiconductor sequencing)	200-400	0.1-0.3	0.1-12 80-3,500	1.8 1.8	short runtime	high indel error rate, higher cost per Mb than Illumina
MGI/BGI (DNBSEQ-G50, G400, T7)	Circularization, RCA to prepare nanoballs	sequencing by ligation (cPAS)	50-200 (2x200)	0.5-4.5	75-720 5-360	0.001 0.001	low number of duplicates, low error rate	
PacBio (Sequel, RS II)	Circularization, no PCR (single molecule)	sequencing by synthesis (labelled nucleotides)	15,000 - 60,000	0.2	0.05-0.4 40-200	5-13 <1	long reads, single-molecule, short runtime	high error rate, low throughput, higher cost per Mb than Illumina
Oxford Nanopore (MinION, GridION)	None (single molecule)	no chemistry	100,000 and longer	0.7-3	0.026-0.6 20-160	10-40 ?	long reads, small portable instrument	higher error rate

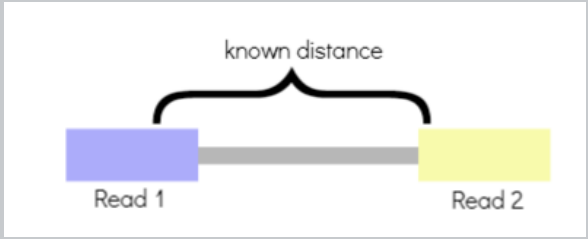
Metzker 2010, Glenn 2011 (NGS Field Guide – http://www.molecularecologist.com/2016/03/2016-ngs-field-guide-preview/)
https://en.wikipedia.org/wiki/List_of_DNA_sequencers

Sequencing libraries

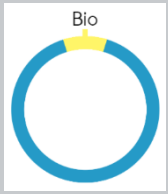
- single end

- pair-end





- mate pair
  - longer fragments (2-5 kbp)
  - circularized, fragmented



What to do with sequences (reads)?

- FASTQ – FASTA + quality scores
- assembling
  - *de novo* assembly
  - *reference-guided assembly*
- applications
  - search for variability (SNP), variant calling
  - search for microsatellites – primer design…
  - identification of suitable single-copy regions for phylogenetic studies
  - phylogenomics – phylogeny based on whole genomes (e.g., cpDNA) or many genes (incl. whole rDNA cistron)
  - …

```
@M01691:49:000000000-AA2TH:1:1101:18780:1973 1:N:0:4
ACTTATTCCATGAGTCGGAAGTGGGGCACGGCCCCTCCTTTTGGCTTGAAGACCCACC
+
>>AA1@DD@3B311BFECEC?F1GHGGGGGGGGGGGHGGHHHHHGHHHGHHHHHHHHGG
```

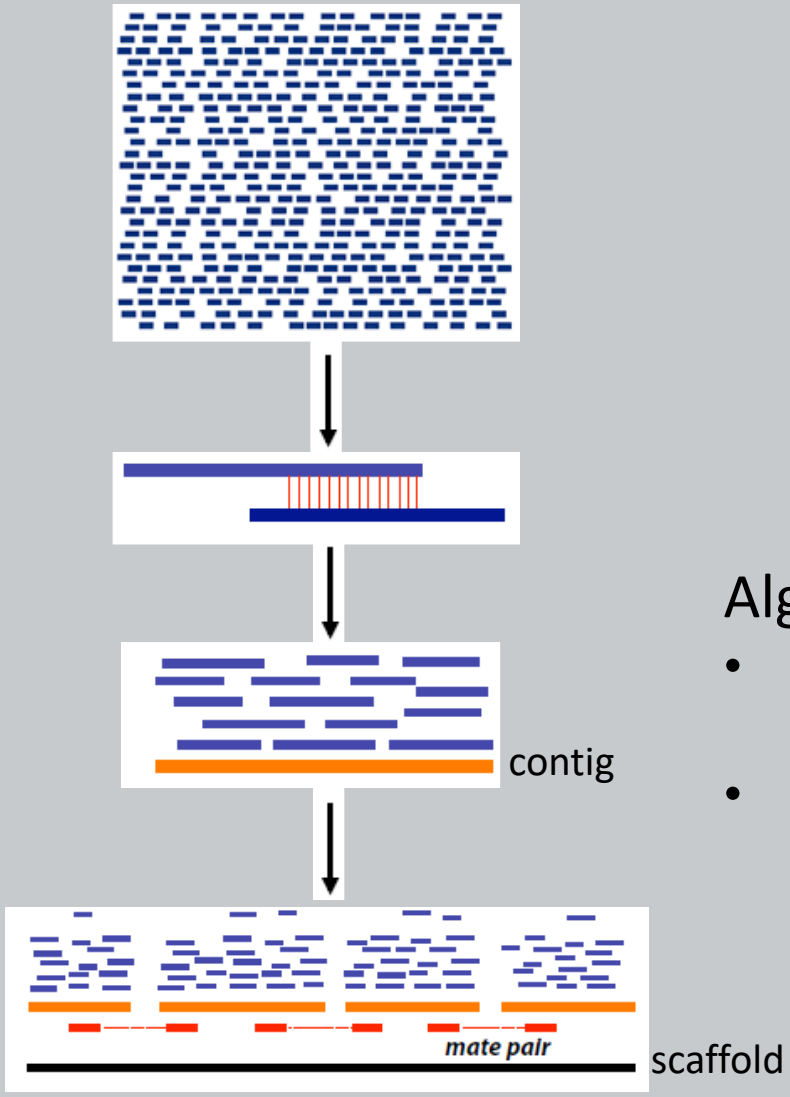
Assembling

generating individual sequences (reads)

search for overlapping reads

assembling reads to contigs

assembling contigs to scaffolds



contig

scaffold

Algorithms

- OLC (overlap/ layout/consensus)
- deBruijn *k*-mer graphs

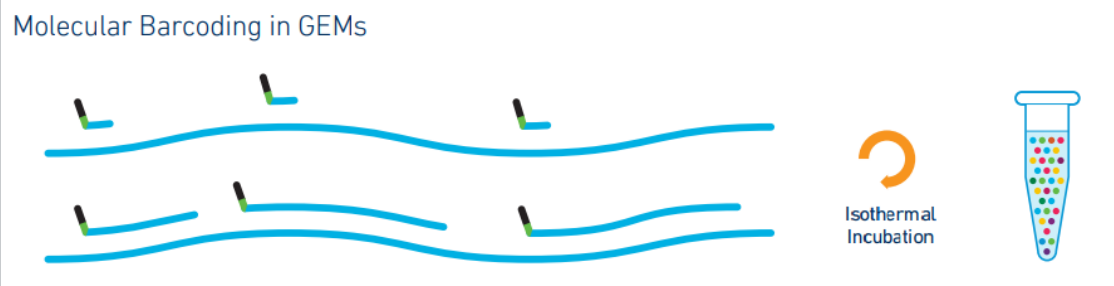
De novo assembly strategies

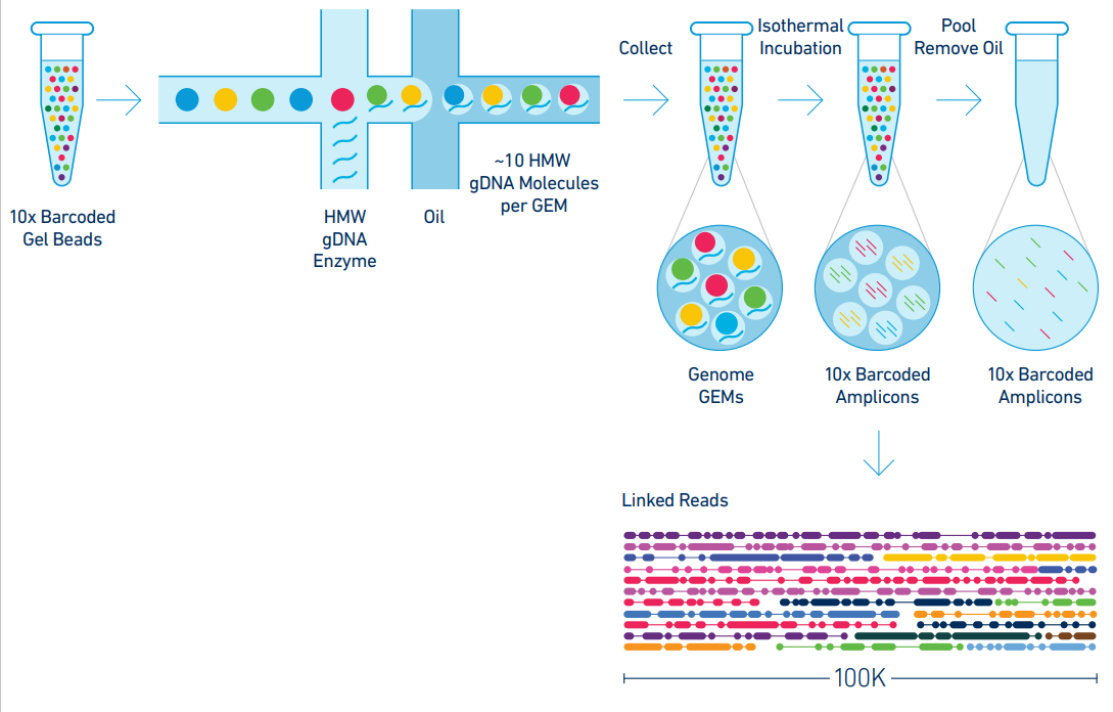
- chromosome-scale assembly hard/impossible for short reads only
- combination of long (PacBio, Nanopore) and short reads – hybrid assembly
- approaches how to obtain long contigs
  - synthetic long reads (10x Genomics)
  - proximity ligation technologies (Dovetail Hi-C)
  - optical mapping (BioNano)

10x Genomics synthetic long reads

- long DNA fragments (up to ~100 kb) are spatially isolated into micelles (GEM droplets – gel beads in emulsion) with a unique barcode (up to 750,000 barcodes available)
- long fragments are amplified (isothermal incubation) – product is a 10x barcoded amplicon ~350 bp
- emulsion is broken, DNA is pooled and sequenced on standard short read platform
- reads sharing the same barcode are derived from the same original large fragment (*linked reads*)
- many long fragments from the same genomic region – generating *read cloud* (stacked linked reads from each fragment)
- microfluidic instrument Chromium – automated preparation of 10x barcoded library

10x Genomics synthetic long reads





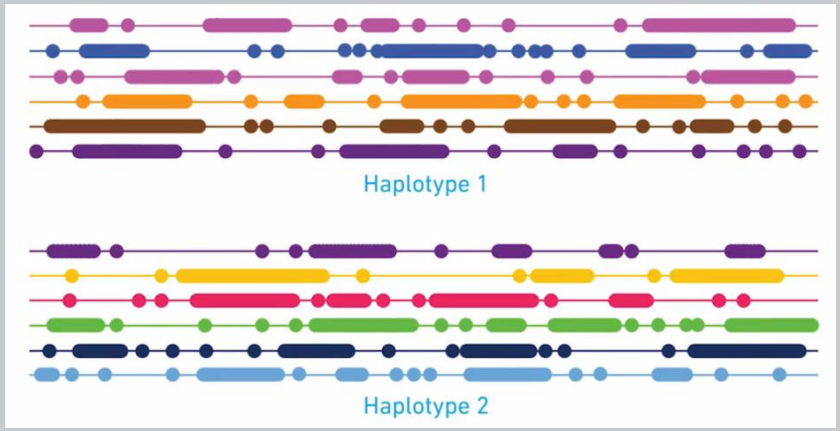
https://www.10xgenomics.com/genome/

10x Genomics synthetic long reads

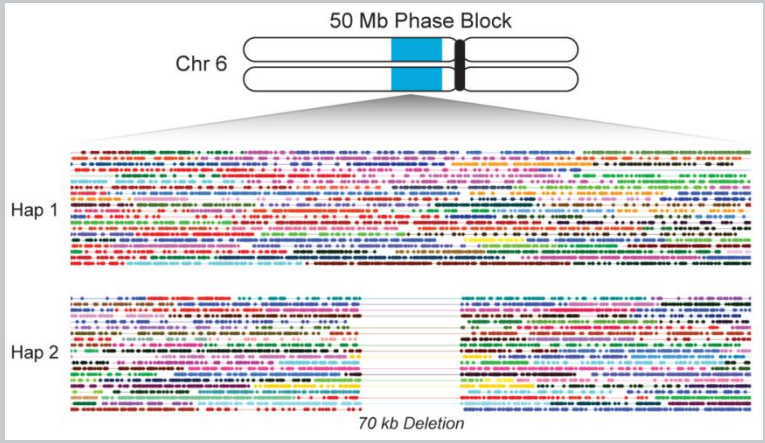
standard short reads cannot place reads correctly in difficult to align regions

linked reads can align reads correctly into paralogous gene loci

haplotype phasing



structural variants

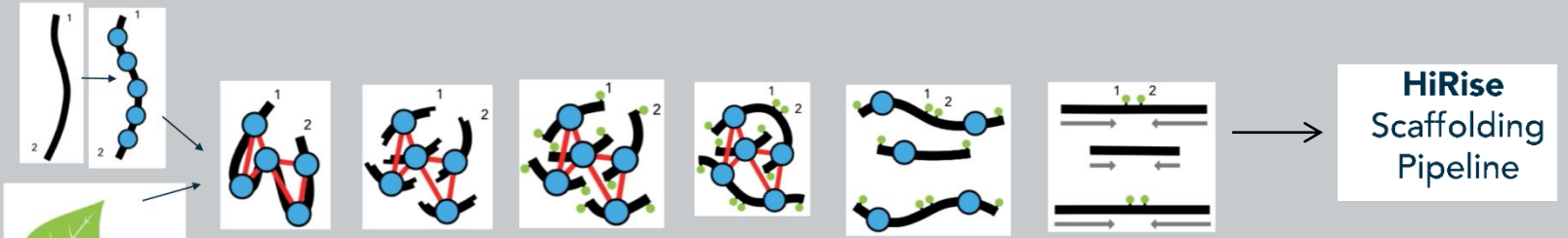


https://www.10xgenomics.com/genome/

Proximity ligation (Dovetail Genomics)

- for chromosome-scale assembly
- chromosome conformation capture sequencing (Hi-C)
- proximity ligation of DNA fragments that are physically close in their natural conformation – ligated in situ before they are cleaved by restriction enzymes and isolated
- Hi-C and Omni-C protocols

Chicago generated libraries start from pure DNA that is reconstituted into chromatin.



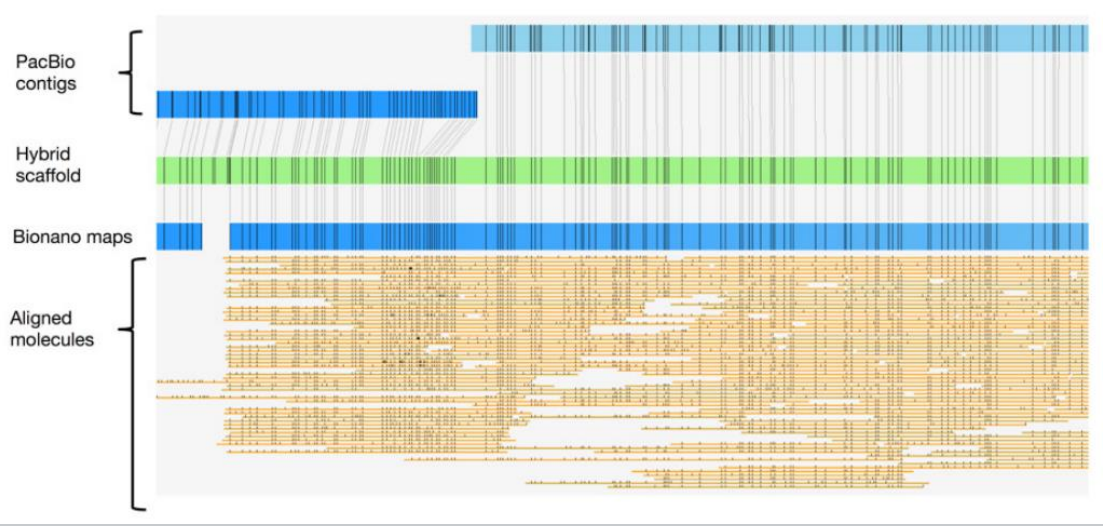
Dovetail Hi-C generated libraries start from tissue or cell culture and endogenous chromatin is extracted after fixation.

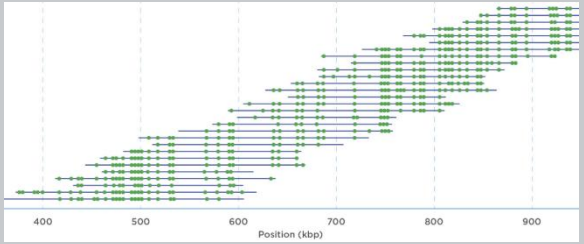
Organism	Assembly size	Starting N50	Final N50	Longest scaffold
Coffee	1,191 Mb	1.85 Mb	82 Mb	122 Mb
Cabernet	683 Mb	1.28 Mb	14.3 Mb	31 Mb
Cashew	377 Mb	1.49 Mb	17 Mb	29 Mb

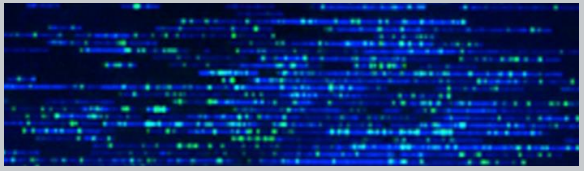
https://dovetailgenomics.com/technology/

Optical mapping (BioNano)

- visualization of long DNA molecules in their native state
- megabase size molecules of genomic DNA are labeled using a *nicking endonuclease*, a specific 6 or 7 basepair sequence is labelled approximately 10 times per 100 kbp
- long labelled molecules are de novo assembled into physical maps using the label patterns
- physical maps are compared to NGS contigs to produce hybrid scaffolds
- instruments – Saphyr, Irys



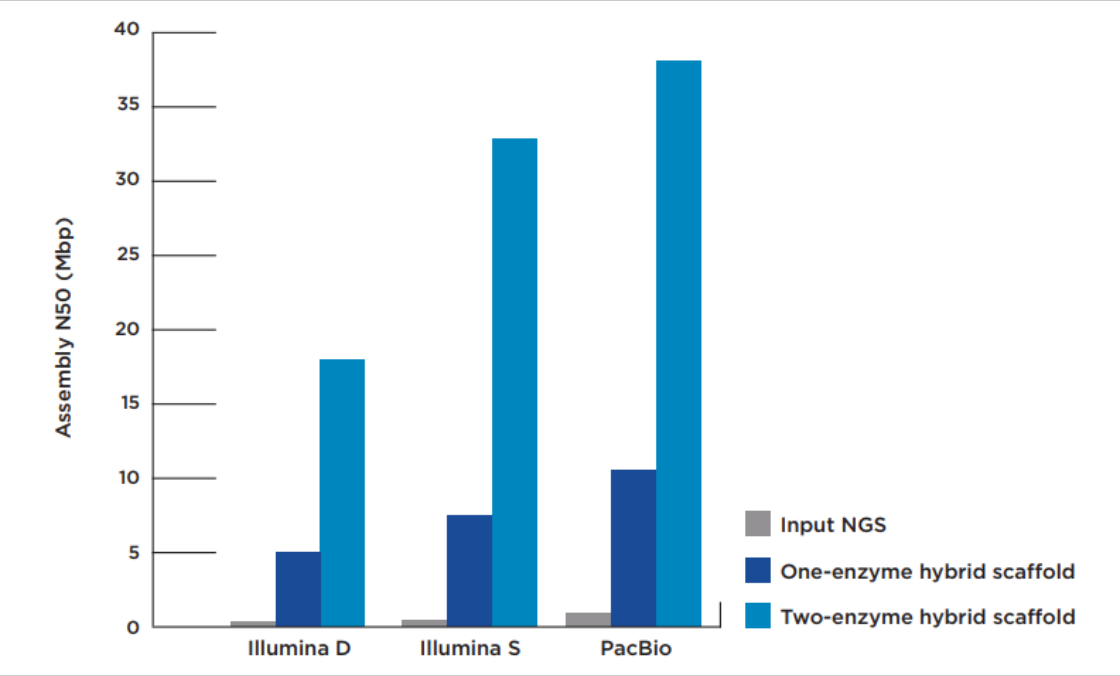




https://bionanogenomics.com/technology/genome-assembly/

Optical mapping (BioNano)

- N50 – assembly quality in terms of contiguity (higher is better)
- the size of the contig which (along with the larger contigs) contain half of sequence of a particular genome



Improvements in assembly contiguity after hybrid scaffold with one-enzyme and two-enzyme genome maps.

Illumina-D: 51x of 250 bp pair-end sequence

Illumina-S: 40x of 101 bp pair-end and 25x of 2.5-2.5 kbp mate-pair sequence

PacBio: 46x with mean read length of 3.6kbp

https://bionanogenomics.com/technology/genome-assembly/

Applications of NGS

- de-novo genome sequencing
  - *targeted enrichment* or *reduction, i.e.,* preferential sequencing of only part of the genome
  - *exome sequencing*, i.e., exons only
- genome re-sequencing
- transcriptome sequencing (RNA-Seq)
- amplicon sequencing
- (environmental) metasequencing
- …

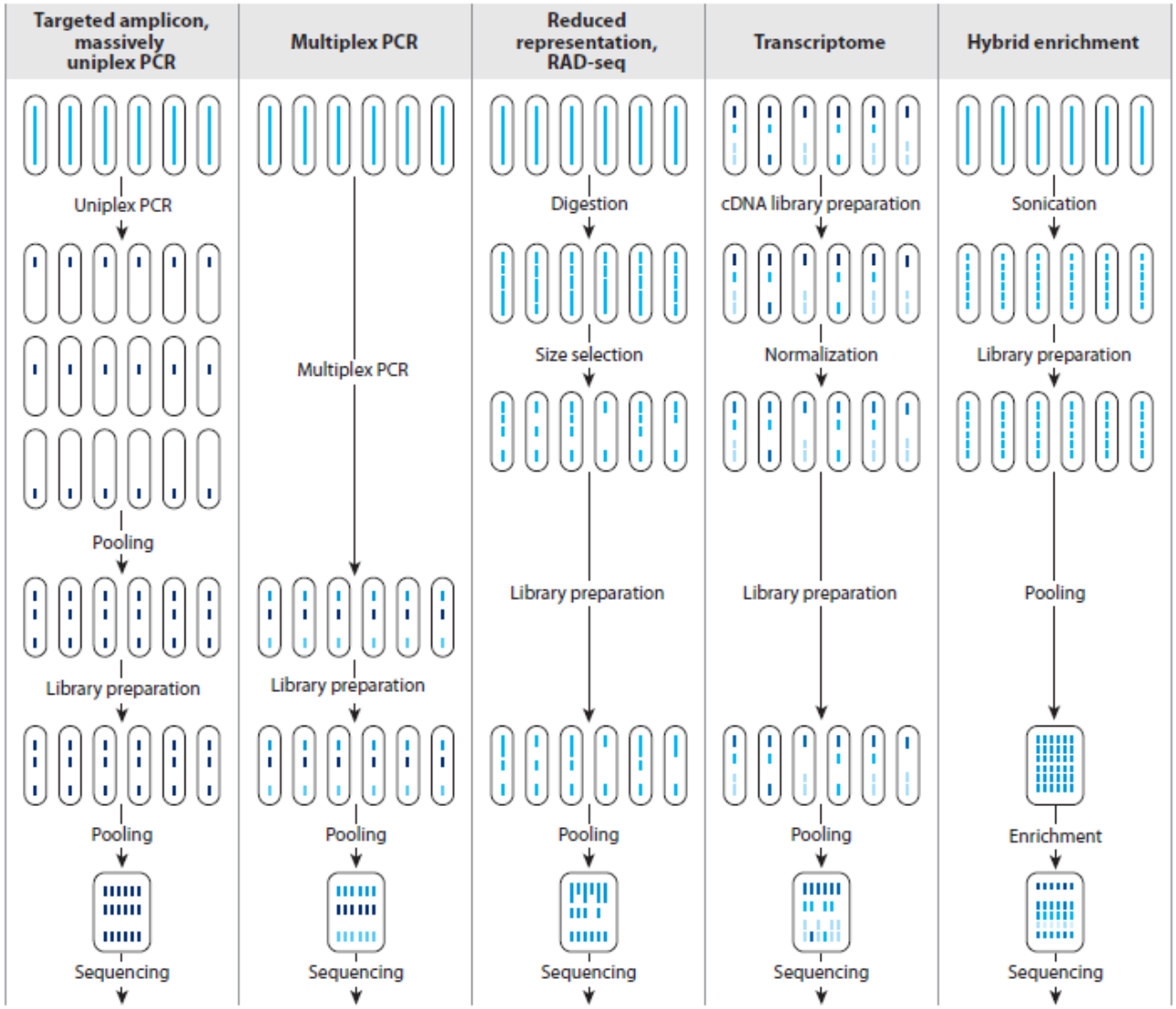
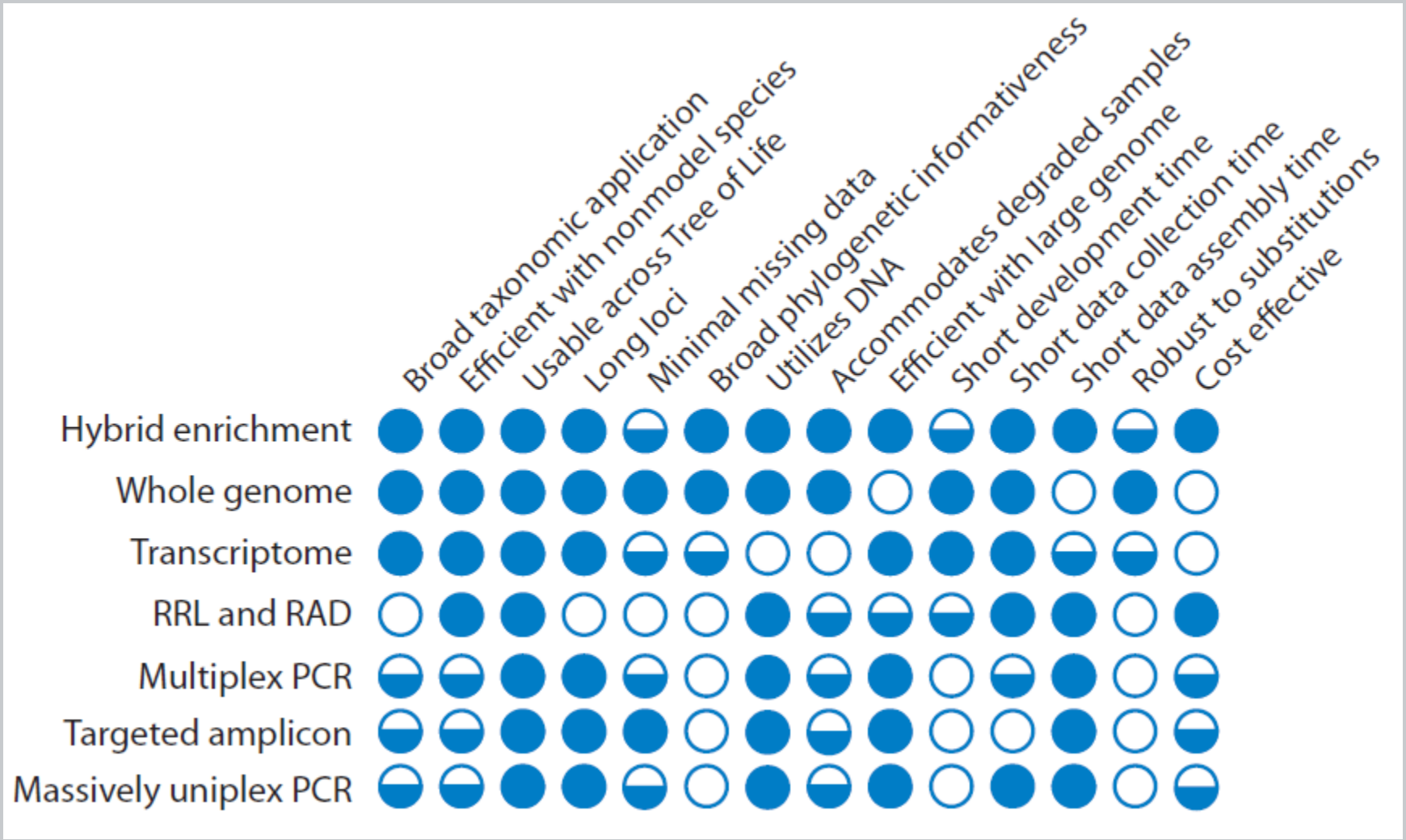


Figure 1

Genomic partitioning workflows for high-throughput phylogenetic data collection. Each oval represents a single sample or pool. Genomic DNA (or RNA, for transcriptome sequencing) is the starting material (*top row*), which undergoes polymerase chain reaction (PCR), enzymatic digestion and size selection, conversion from RNA to cDNA (transcriptome), or shearing (*upper middle rows*), followed by indexed library preparation (*lower middle rows*), pooling across samples (and enrichment in the *rightmost column*), and high-throughput sequencing (*bottom row*). Color intensity (*shades of blue*) indicates relative degree of enrichment of genomic regions during the different stages of each approach. Abbreviations: cDNA, complementary DNA; RAD-seq, restriction-site-associated DNA sequencing.

Lemmon E.M. & Lemmon A.R. (2013): *High-throughput genomic data in systematics and phylogenetics*. Annu. Rev. Ecol. Evol. Syst, 44, 99–121.



Lemmon E.M. & Lemmon A.R. (2013): *High-throughput genomic data in systematics and phylogenetics*. Annu. Rev. Ecol. Evol. Syst, 44, 99–121.

Whole genome sequencing

- sequencing + assembly (+ annotation)
- simple for small genomes
  - bacteria
  - cpDNA
- still challenging for large eukaryotic genomes – data combination from several platforms (long + short reads) – Illumina + PacBio + Hi-C

Plant sequenced genomes

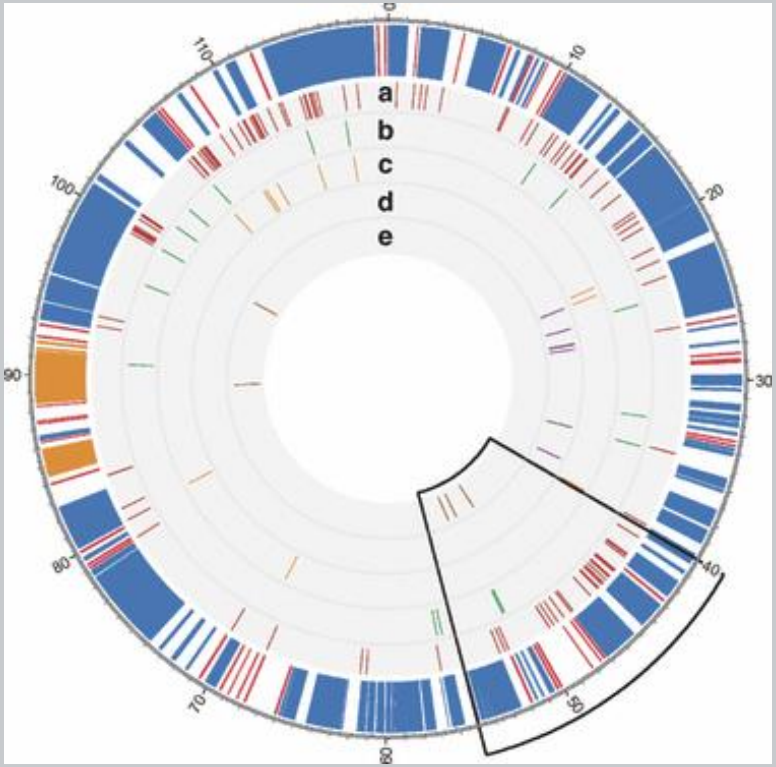
- assembled and published genomes
  - ~1,000 genomes of flowering plants
  - ~100 genomes of non-flowering plants
- https://www.plabipd.de/
  - timeline view
  - cladogram view
- 10KP: 10,000 Plant Genomes Project (https://db.cngb.org/10kp/)

Plastome assembly/annotation

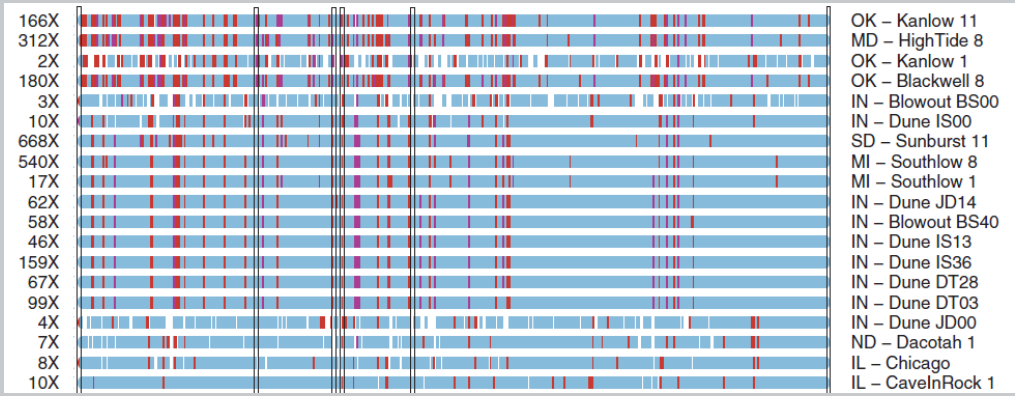
- many ”bioinformatic” pipelines
  - GetOrganelle (https://github.com/Kinggerm/GetOrganelle)
  - FastPlast (https://github.com/mrmckain/Fast-Plast)
  - ORG.asm
  - …
- (semi)automatic annotation
  - DOGMA (https://dogma.ccbb.utexas.edu/)
  - GeSeq (https://chlorobox.mpimp-golm.mpg.de/geseq.html)
  - Plastid Genome Annotator (PGA) (https://github.com/quxiaojian/PGA)

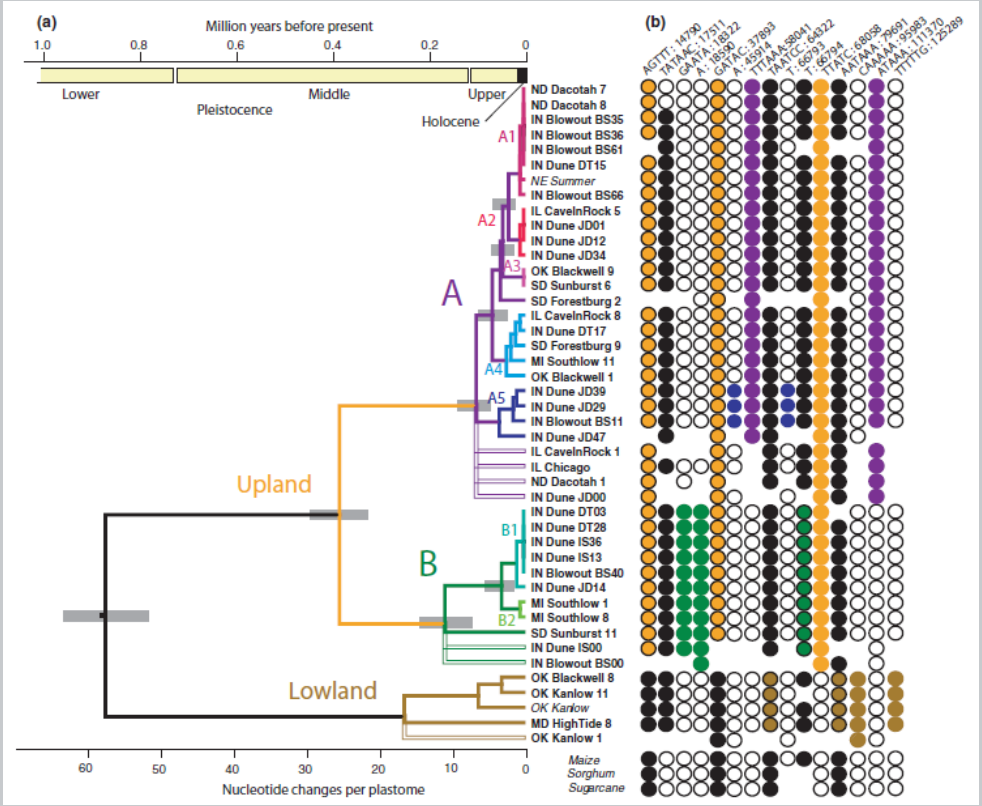
Freudenthal et al. 2020: *A systematic comparison of chloroplast genome assembly tools.* Genome Biology 21: 254.

Whole chloroplast sequencing



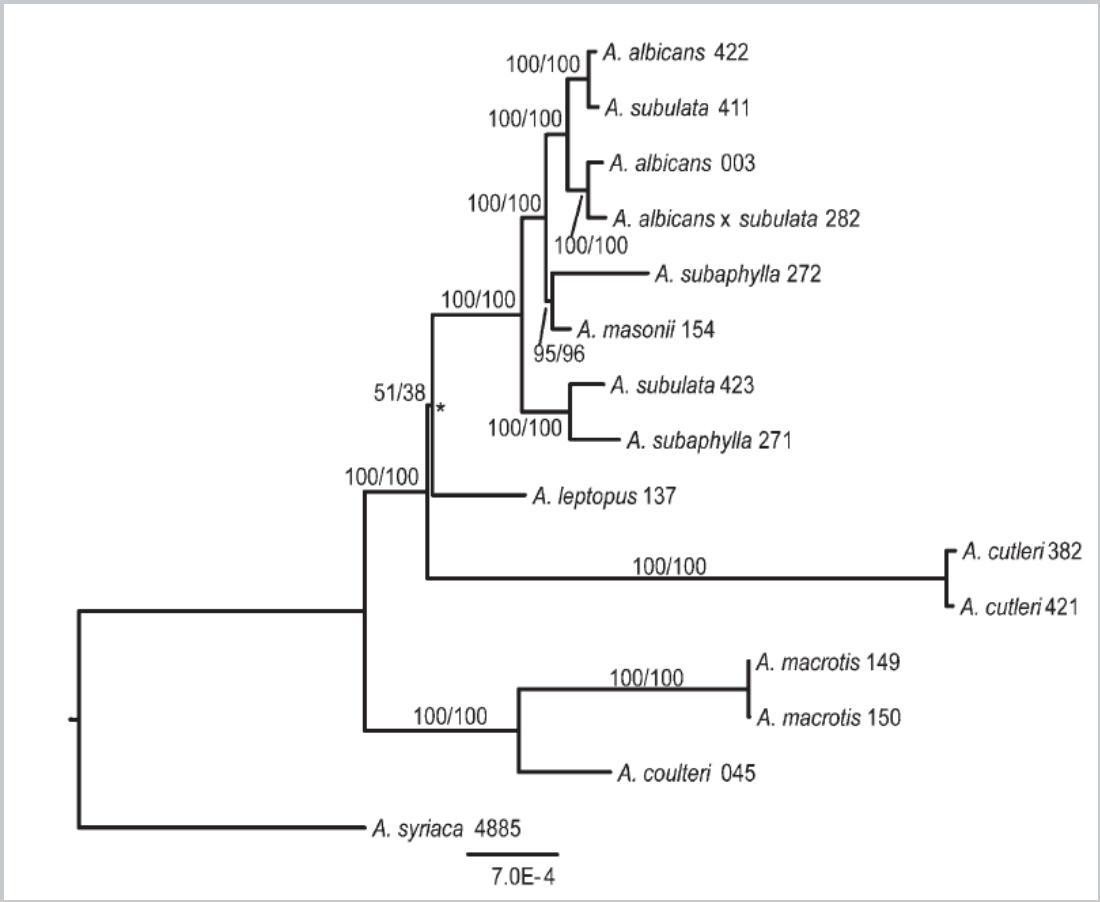
Whittall et al. (2010): Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. *Molecular Ecology* 19:100-114.





Morris et al. (2011): Genomic diversity in switchgrass (Panicum virgatum): from the continental scale to a dune landscape. *Molecular Ecology* 20: 4938–4952

Whole chloroplast sequencing



Asclepias

Straub et al. (2012): *Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics*. American Journal of Botany 99: 349–364.

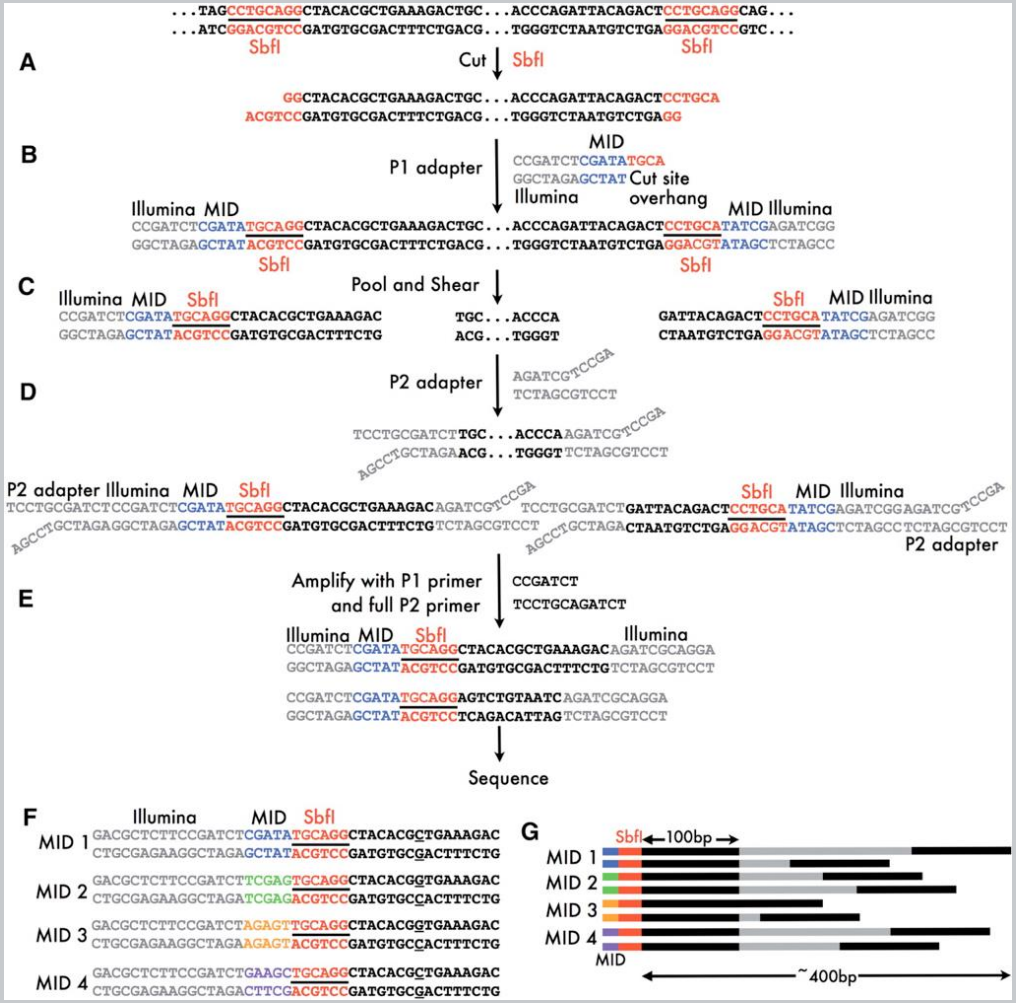
Targeted enrichment

- reduction of the complexity of sequenced parts
- enzyme restriction of the genome
  - sequencing only the part of the genome associated with restriction sites
  - searching for SNPs -> binary data
  - RAD-sequencing
  - GBS (genotyping-by-sequencing)
  - …
- Hyb-Seq
  - hybridization based enrichment
  - selection of specific sequences (thousands of exons)

Cronn et al. (2012): *Targeted enrichment strategies for next-generation plant biology*. American Journal of Botany 99: 291-31.

RAD-sequencing

Restriction-site-associated DNA sequencing

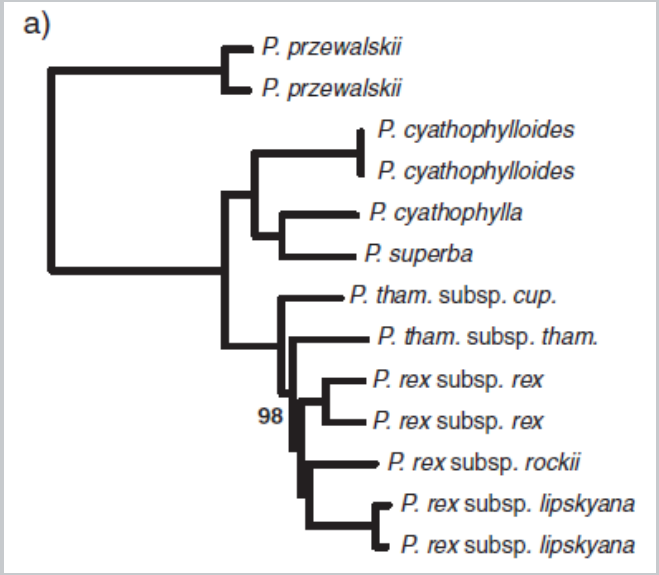


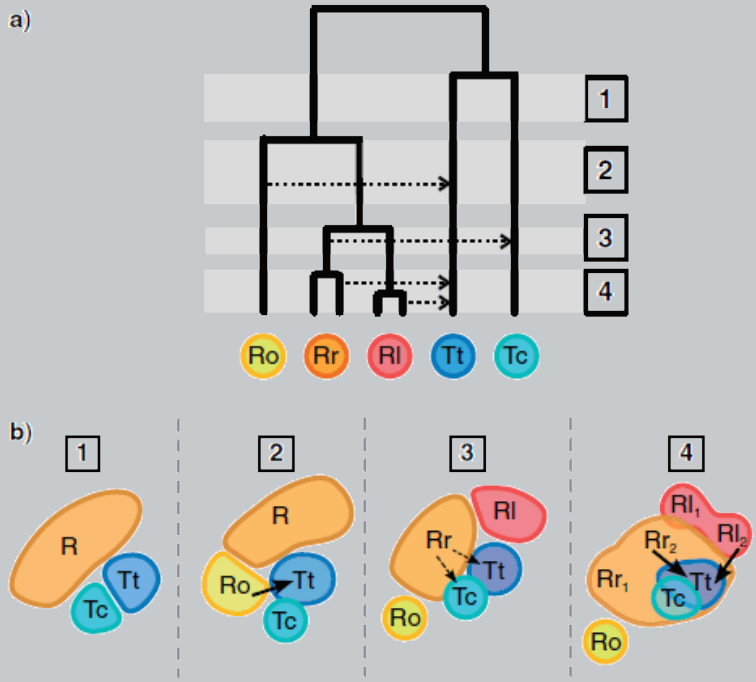
Davey J.W. & Blaxter M.L. (2011): *RADSeq: next-generation population genetics*. Briefings in Functional Genomics 9: 416-423.

Davey J.W. et al. (2011): *Genome-wide genetic marker discovery and genotyping using next-generation sequencing*. Nature Reviews 12: 499-510.

RAD in recently diversified group

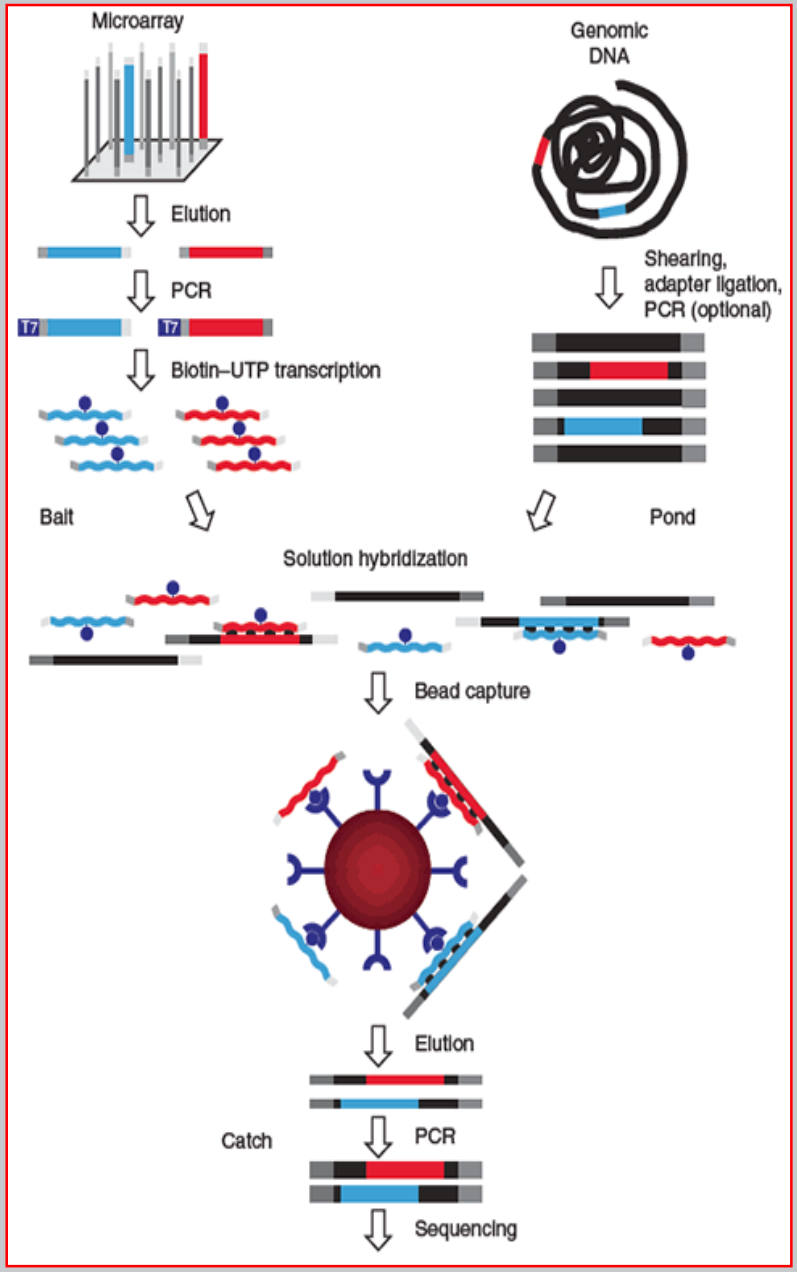
- recently diversified group – closely related species
- reduced representation sequencing (RADSeq)
- phylogeny and detection of ancestral hybridization
- 40,000 loci





Hyb-Seq

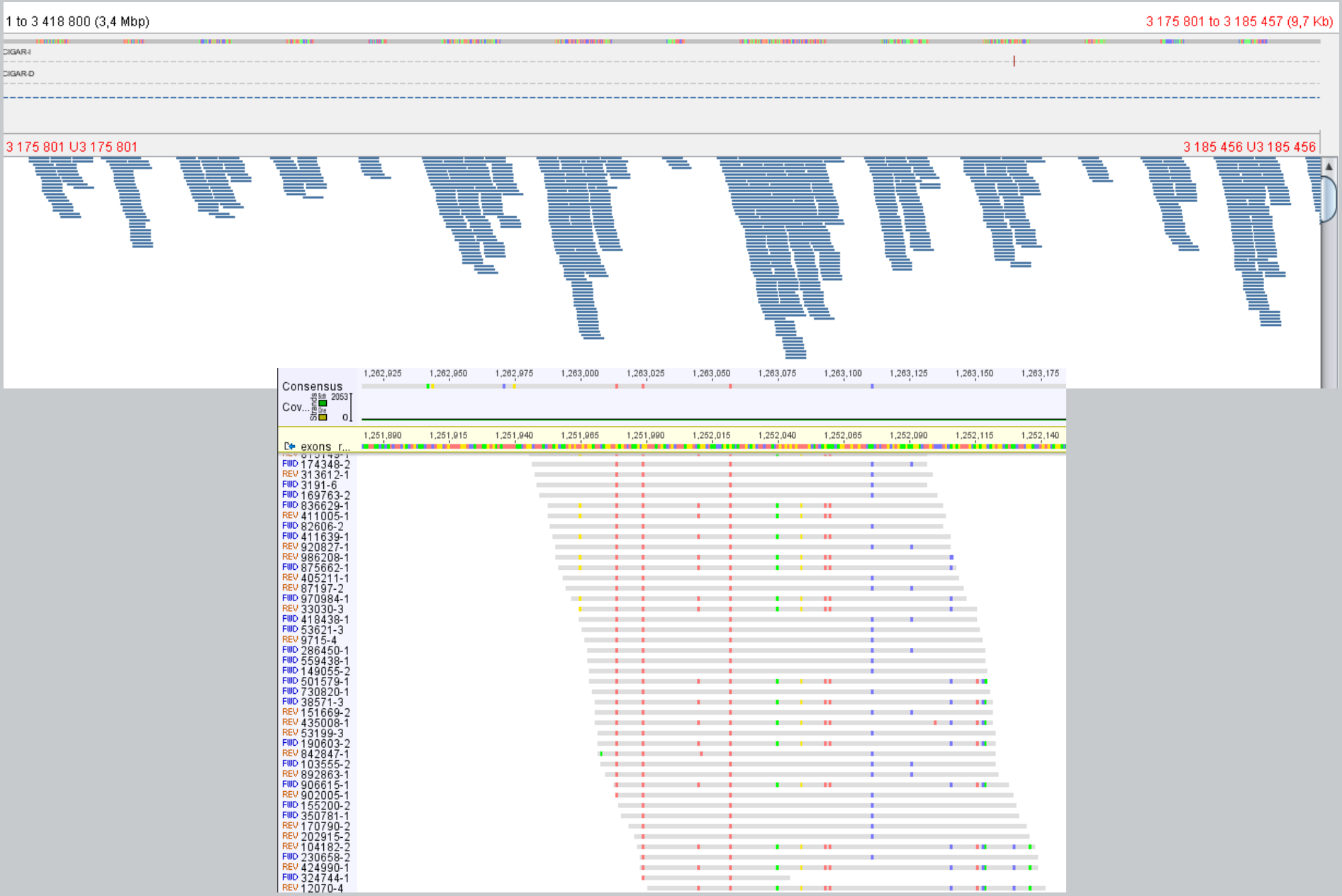
- solution phase hybridization
- ‘baits’ (short RNA fragments) synthetized on arrays
- hybridization in solution
- immobilization via biotin-streptavidine
- enrichment of target sequences



Mycroarray

Weitemier et al. (2014) Appl Plant Sci. 2: apps.1400042
Cronn et al. (2012) Amer. J. Bot 99: 291-311
Lemmon et al. (2012) Syst. Biol.
McCormack et al. (2012) Syst. Biol.
Bi et al. (2012) BMC Genomics

Hyb-Seq – reads mapped to reference

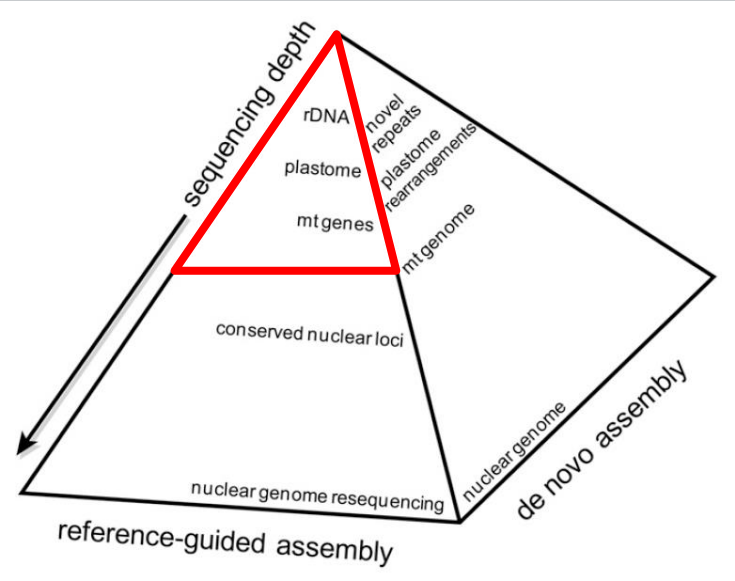


http://www.onekp.com

- transcriptome sequencing for 1,300 plant species (including ca. 750 angiosperms) – free
- informations for robust phylogenetic studies and biotechnology
- usable for selection of suitable regions for phylogeny, e.g., for baits design for enrichment

Genome-skimming

- genome sequencing with low total coverage
- we get enough coverage for assembly
  - whole plastome
  - large portions of mtDNA
  - rDNA cistrone
  - many candidate single-copy genes
  - microsatellite regions



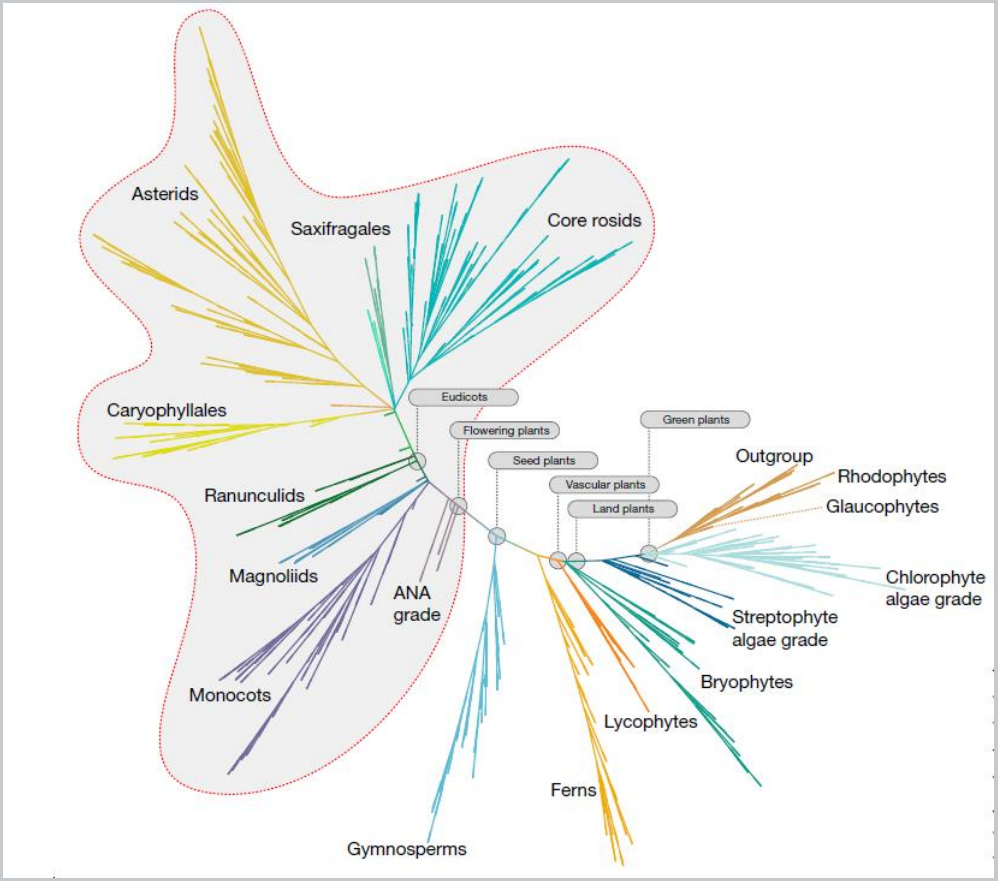
Straub et al. (2012): *Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics*. American Journal of Botany 99: 349–364.
Steel et al. (2012): *Quality and quantity of data recovered from massively parallel sequencing: Examples in Asparagales and Poaceae*. American Journal of Botany 99: 330-348.

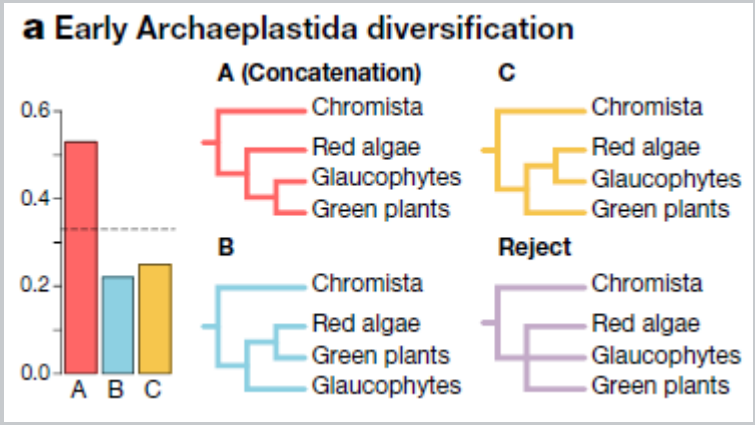
Transcriptome sequencing

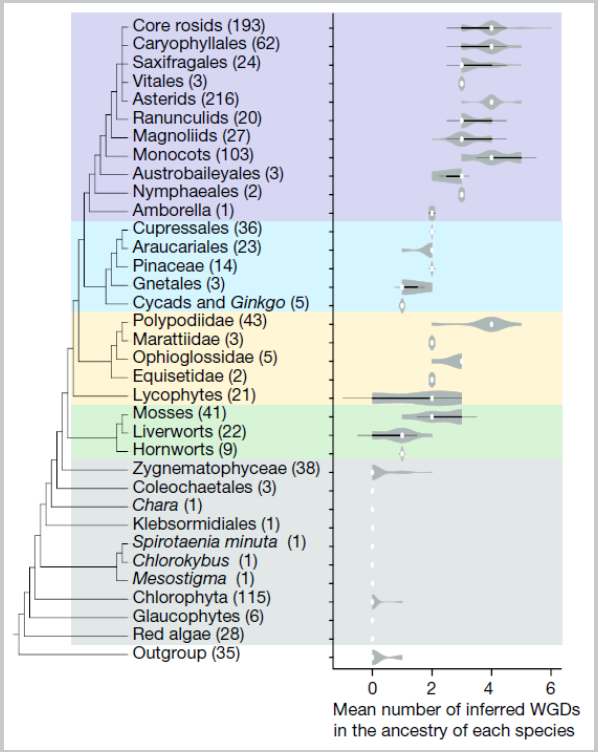
- cDNA sequencing (obtained by reversal transcription of mRNA)
- transcriptome is much smaller than whole genome
- useful for non-model species

- applications
  - transcriptomics – which genes are transcribed, differential expression (DE)…
  - searching for suitable genes for phylogenetic studies (variable regions when comparing information from more individuals/species)
  - microsatellite identification
  - phylotranscriptomics – orthology assessment problém
  - detecting past whole genome duplication (WGD) events
  - …

Phylotranscriptomics



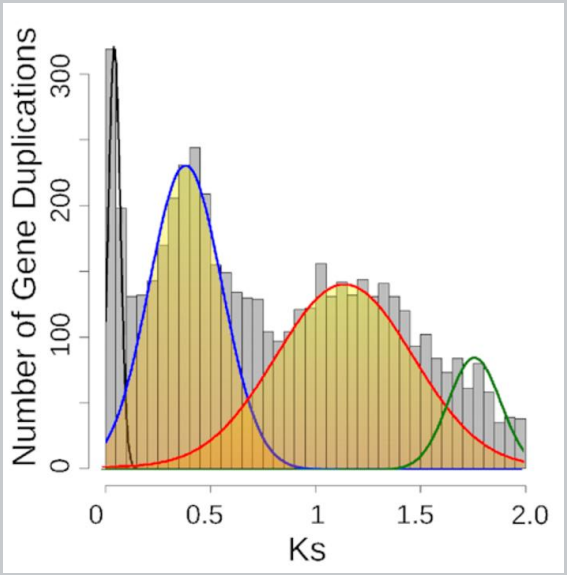




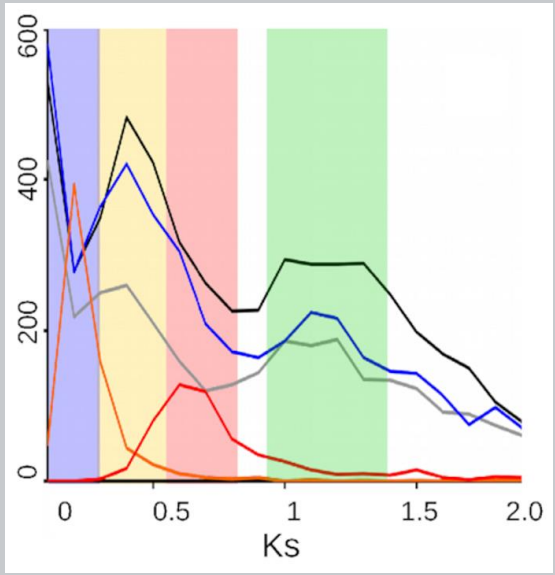
One thousand plant transcriptomes and the phylogenomics of green plants. 2019. Nature 574: 679-685

Ancient WGD detection using K_s

- age distributions of gene duplications
- Ks – estimated number of synonymous mutations between paralogues
- plotting distribution of Ks values
- MultiAxon Paleopolyploidy Search (MAPS) to confirm the placement

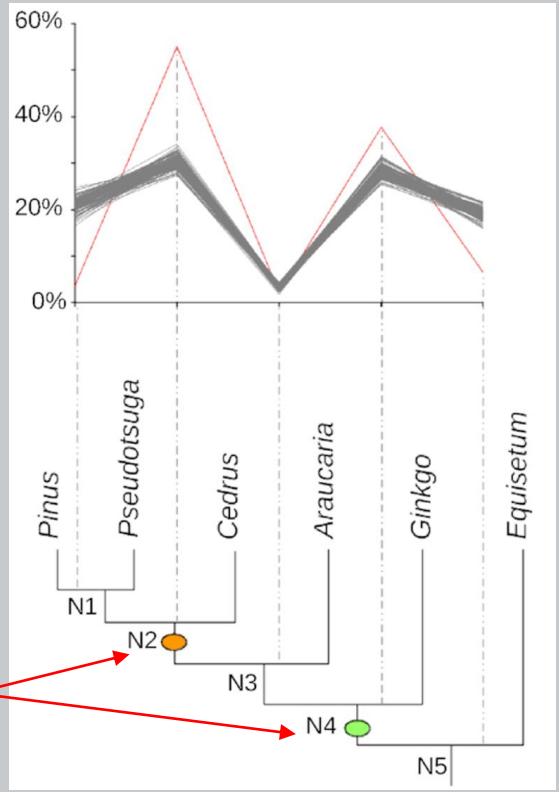


single species



overlay of multiple species

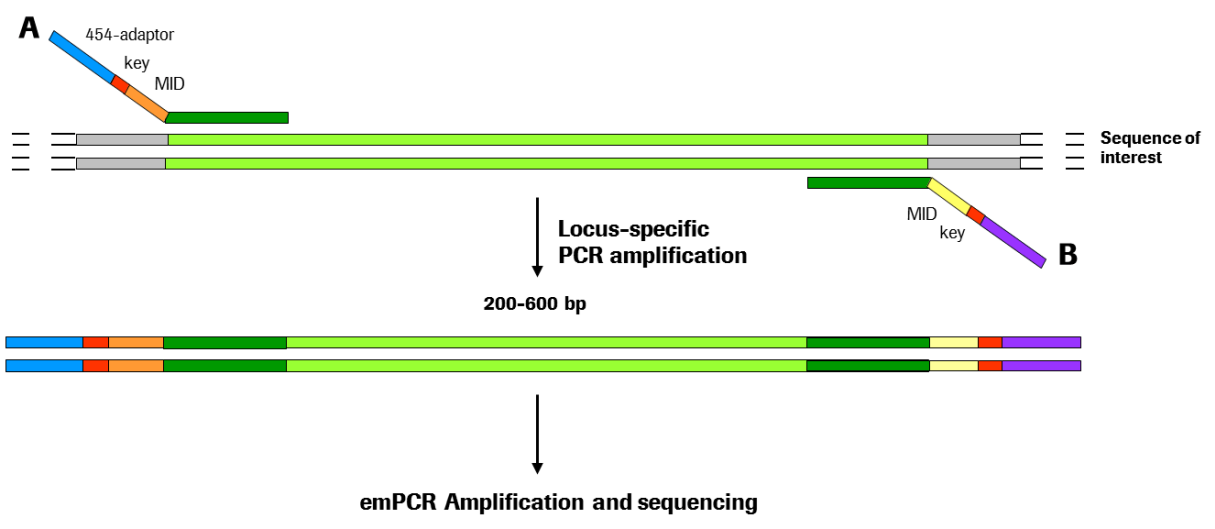
WGD events



Li & Barker (2020). GigaScience 9: giaa004

Amplicon sequencing

- PCR-amplification of target gene or intergenic region
- product labelling with specific sequence (MID)
- parallel sequencing of all PCR reactions
- sequences are bioinformatically separated according to their MID identification



Metasequencing

- PCR amplification of target gene from eDNA (environmental sample – water, soil etc.)
  - 16S rDNA, 18S rDNA
  - mitochondrial 12S (16S) rDNA or cytochrome oxidase I (COI) in protista and animals
- sequencing of all products
- comparison of sequences with database
- species identification and frequency
- data analysis software – OBITools, MOTHUR, QIIME, R…

- application – community composition
  - bacterial or fungal community
  - historical – e.g., DNA from permafrost, sedimantary DNA etc.
  - food preferences of animals

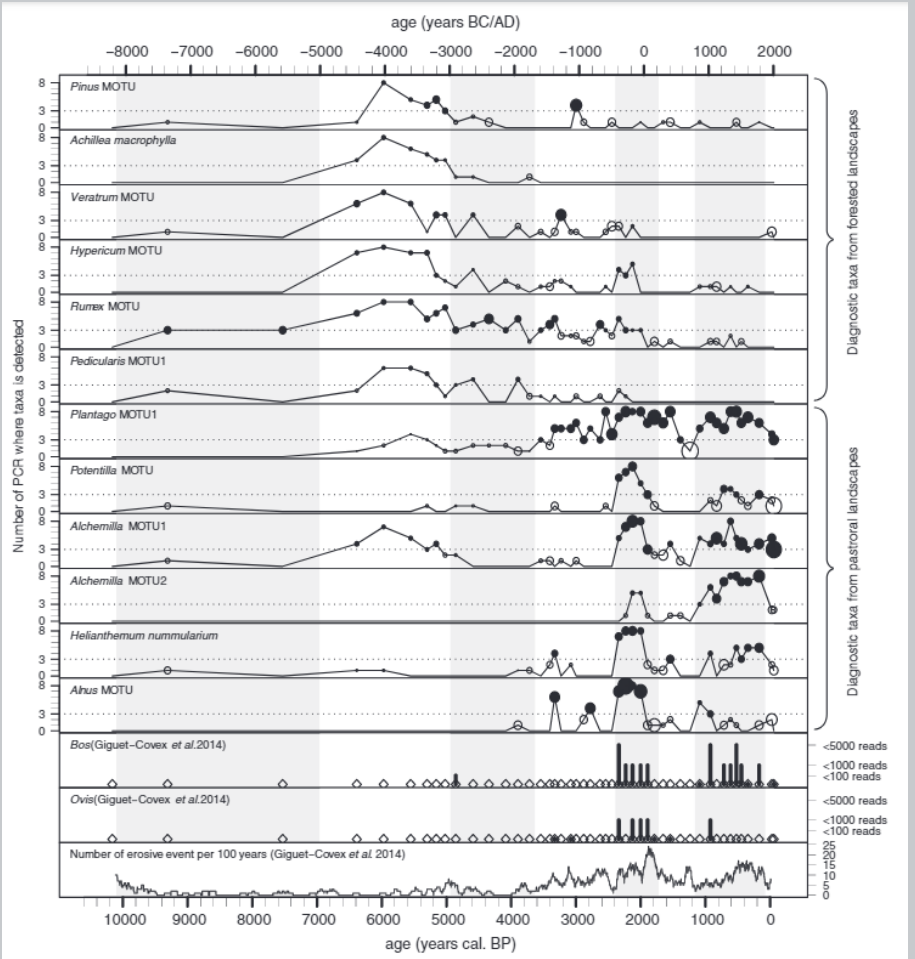
- barcoding – universal short sequence for unequivocal identification
  - plants – *rbc*L, *mat*K, (*trn*H-*psb*A)
  - CBOL – Consortium for the Barcode of Life
  - http://www.barcoding.si.edu/plant_working_group.html

Historical composition of Arctic vegetation

	%
22 960 ± 120 years BP	
Bistorta vivipara	47.25
Equisetum arvense/E. fluviatile/E. sylvaticum	24.31
Salix sp./*Chosenia arbutifolia/Populus balsamifera*	4.74
Armeria scabra	3.03
Thymus oxyodontus	2.77
Lagotis glauca	2.17
Asteraceae 1*	1.87
Avenella flexuosa	1.77
Aconogonon alaskanum/A. ocreatum/A. tripterospermum	1.36
Rumex sp.	1.31
Packera sp./*Senecio* sp.	0.96
Poaceae 1†	0.96
Ranunculus acris/R. subborealis/R. turneri	0.81
Festuca sp.	0.76
Hulteniella integrifolia	0.66
Saxifraga hirculus	0.55
Trientalis europaea	0.45
Asteraceae 2‡	0.40
Valeriana capitata/V. officinalis agg.	0.35

Sønstebø et al. (2010): *Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate*. Molecular Ecology Resources 10: 1009-1018.

Human impact on plant communities

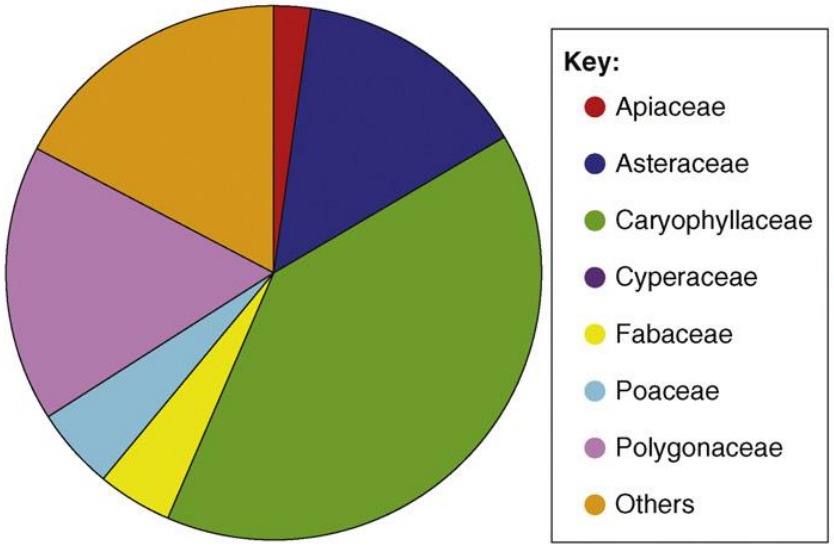


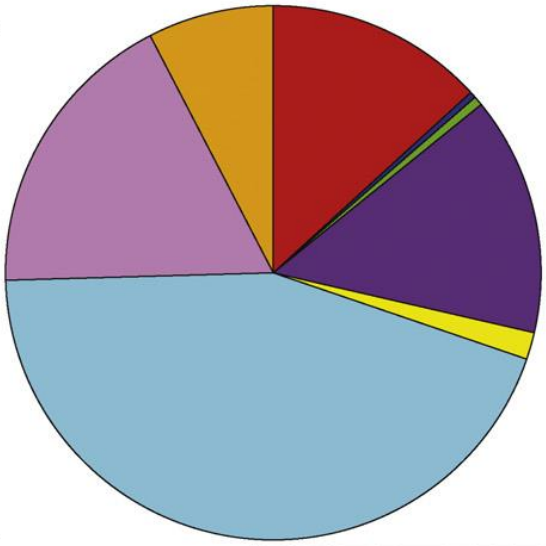
Pansu et al. (2015): *Reconstructing long-term human impacts on plant communities: an ecological approach based on lake sediment DNA*. Molecular Ecology 24: 1485-1498.

Food preferences of animals

Golden marmot

Brown bear





TRENDS in Ecology & Evolution

Valentini A., Pompanon F. & Taberlet P. (2008): *DNA barcoding for ecologists*. TREE 24: 110-117.

Systematic study

Wittall J.B. et al. (2010): Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. *Molecular Ecology* 19, 100-114.

Literature

Metzker M.L. (2010): *Sequencing technologies – the next generation*. Nature Reviews Genetics, 11, 31–46.

Bräutigam A. & Gowik U. (2010): *What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research*. Plant Biology, 12, 831–841.

Ansorge W.J. (2009): *Next-generation DNA sequencing techniques*. New Biotechnology, 25, 195–203.

Glenn T.C. (2011): *Field guide to next-generation DNA sequencers*. Molecular Ecology Resources, 11, 759–769.

Goodwin S. et al. (2016): *Coming of age – ten years of next-generation sequencing technologies*. Nature Reviews Genetics, 17, 333–351.

Jiao W.-B. & Schneeberger K. (2017): *The impact of third generation genomic technologies on plant genome assembly*. Current Opinion in Plant Biology, 36, 64–70.

McCormack J.E. et al. (2011): *Applications of next-generation sequencing to phylogeography and phylogenetics*. Mol. Phylogenet.Evol.

Cronn et al. (2012): *Targeted enrichment strategies for next-generation plant biology*. American Journal of Botany 99: 291-31.

Straub et al. (2012): *Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics*. American Journal of Botany 99: 349–364.

Lemmon E.M. & Lemmon A.R. (2013): *High-throughput genomic data in systematics and phylogenetics*. Annu. Rev. Ecol. Evol. Syst, 44, 99–121.

Taberlet P. et al. (2019): *Environmental DNA for Biodiversity Research and Monitoring*. Oxford University Press.