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

PCR amplification  Break emulsion  Template dissociation

Metzker 2010
Pyrosequencing (Roche)

Flow of single dNTP type across PTP wells

- dNTP
- polymerase
- APS
- PP
- ATP
- luciferin
- light and oxyluciferin

- sulfurylase
- luciferase

Metzker 2010
Pyrosequencing (Roche)

instruments – GS Junior, GS FLX Titanium

Metzker 2010
Solid-phase amplification (Illumina)

Sample preparation DNA (5 μg)

Template, dNTPs and polymerase

Bridge amplification

100-200 million molecular clusters

Cluster growth

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)

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

@M01691:49:000000000-000000000-0000000000-0000000000
ACTTATTCCATGAGTCGGAAGTGGGGCACGGCCCCTCCTTTTGCTTGAAGACCCACC
+>>AA1@DD@3B311BFECEC?F1GHGGGGGGGGGGGGHGGGHHHHHHGHHHGGHGGGHHHHHGGG
Assembling

- generating individual sequences (reads)
- search for overlapping reads
- assembling reads to contigs
- assembling contigs to scaffolds

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

Molecular Barcoding in GEMS

10x Barcoded Gel Beads → HMW gDNA Enzyme → Oil → Collect → Isothermal Incubation → Pool Remove Oil

-10 HMW gDNA Molecules per GEM

Genome GEMs → 10x Barcoded Amplicons → 10x Barcoded Amplicons → Linked Reads

100K

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.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Assembly size</th>
<th>Starting N50</th>
<th>Final N50</th>
<th>Longest scaffold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coffee</td>
<td>1,191 Mb</td>
<td>1.85 Mb</td>
<td>82 Mb</td>
<td>122 Mb</td>
</tr>
<tr>
<td>Cabernet</td>
<td>683 Mb</td>
<td>1.28 Mb</td>
<td>14.3 Mb</td>
<td>31 Mb</td>
</tr>
<tr>
<td>Cashew</td>
<td>377 Mb</td>
<td>1.49 Mb</td>
<td>17 Mb</td>
<td>29 Mb</td>
</tr>
</tbody>
</table>

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
- ...
<table>
<thead>
<tr>
<th>Targeted amplicon, massively uniplex PCR</th>
<th>Multiplex PCR</th>
<th>Reduced representation, RAD-seq</th>
<th>Transcriptome</th>
<th>Hybrid enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniplex PCR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pooling</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Library preparation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pooling</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequencing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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.

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)

Whole chloroplast sequencing


Whole chloroplast sequencing


Asclepias
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)


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

Bi et al. (2012) BMC Genomics
Hyb-Seq – reads mapped to reference
• 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


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

Ancient WGD detection using $K_s$

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

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, sedimentary DNA etc.
  - food preferences of animals

- barcoding – universal short sequence for unequivocal identification
  - plants – *rbcL, matK, (trnH-psbA)*
  - CBOL – Consortium for the Barcode of Life
  - http://www.barcoding.si.edu/plant_working_group.html
### Historical composition of Arctic vegetation

<table>
<thead>
<tr>
<th>Species</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bistorta vivipara</td>
<td>47.25</td>
</tr>
<tr>
<td>Equisetum arvense/E. fluviatile/E. sylvaticum</td>
<td>24.31</td>
</tr>
<tr>
<td>Salix sp./Chosenia arbutijolia/Populus balsamifera</td>
<td>4.74</td>
</tr>
<tr>
<td>Armeria scabra</td>
<td>3.03</td>
</tr>
<tr>
<td>Thymus oxyzendectus</td>
<td>2.77</td>
</tr>
<tr>
<td>Lagotis glauca</td>
<td>2.17</td>
</tr>
<tr>
<td>Asteraceae 1*</td>
<td>1.87</td>
</tr>
<tr>
<td>Avenella flexuosa</td>
<td>1.77</td>
</tr>
<tr>
<td>Aconogonon alaskanum/A. ocreatum/A. tripterospermum</td>
<td>1.36</td>
</tr>
<tr>
<td>Rumex sp.</td>
<td>1.31</td>
</tr>
<tr>
<td>Packera sp./Senecio sp.</td>
<td>0.96</td>
</tr>
<tr>
<td>Poaceae 1†</td>
<td>0.96</td>
</tr>
<tr>
<td>Ranunculus acris/R. subborealis/R. turneri</td>
<td>0.81</td>
</tr>
<tr>
<td>Festuca sp.</td>
<td>0.76</td>
</tr>
<tr>
<td>Hulteniella integrifolia</td>
<td>0.66</td>
</tr>
<tr>
<td>Saxifraga hirculus</td>
<td>0.55</td>
</tr>
<tr>
<td>Trisetalis europaea</td>
<td>0.45</td>
</tr>
<tr>
<td>Asteraceae 2‡</td>
<td>0.40</td>
</tr>
<tr>
<td>Valeriana capitata/V. officinalis agg.</td>
<td>0.35</td>
</tr>
</tbody>
</table>

**22 960 ± 120 years BP**


### Human impact on plant communities

Food preferences of animals

Golden marmot

Brown bear

Key:
- Apiaceae
- Asteraceae
- Caryophyllaceae
- Cyperaceae
- Fabaceae
- Poaceae
- Polygonaceae
- Others

Systematic study


