Next-generation sequencing (NGS)

Sanger sequencing

Primer - F - AAGTCAGTCTAA=o

Primer - F - AAGTCAGTCTA=o

Primer - F - AAGTCAGTCT=o

Primer - F - AAGTCAGTC=o

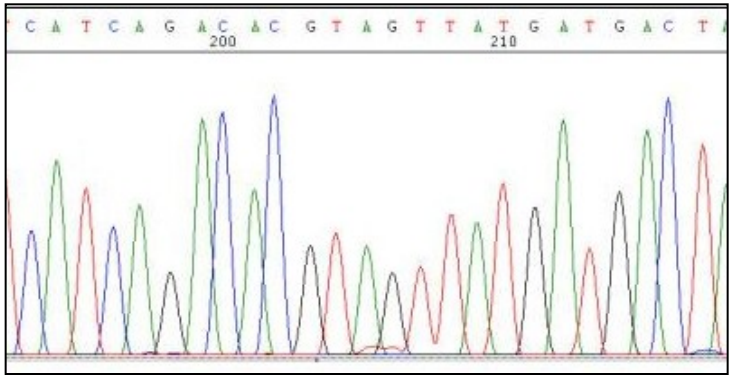
Primer - F - AAGTCAGT=o

Primer - F - AAGTCAG=o

Primer - F - AAGTCA=o

Primer - F - AAGTC=o

-



krátké ---------- dlouhé

(rychlé) ---------- (pomalé)

+

Primer - F AAGTCAGTCTAAATGCGATTGGGA Rev. Primer - R

Rev. Primer - F TTCAGTCAGATTTACGCTAACCCT Primer - R

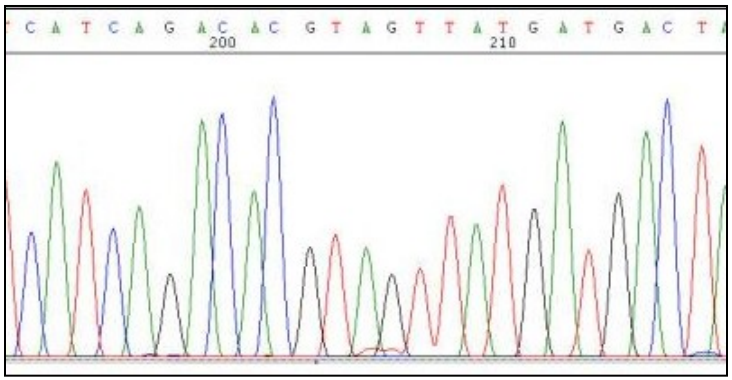
4-kapilární sekvenátor

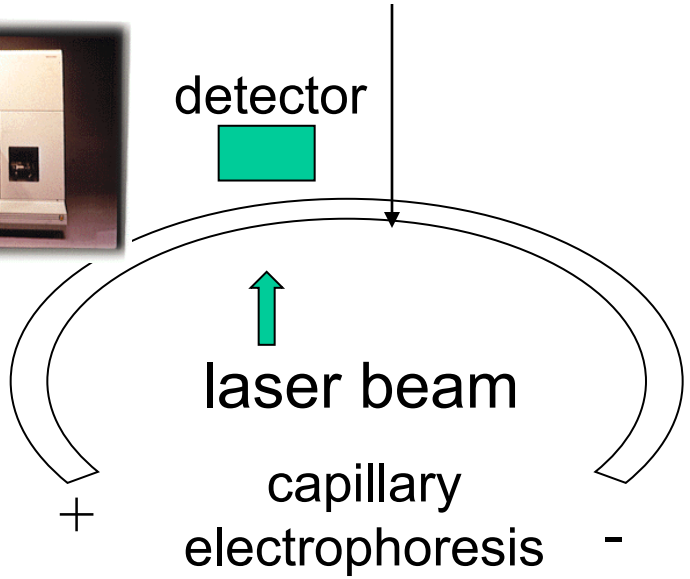
=

96 x 500 bp/12 hodin

=

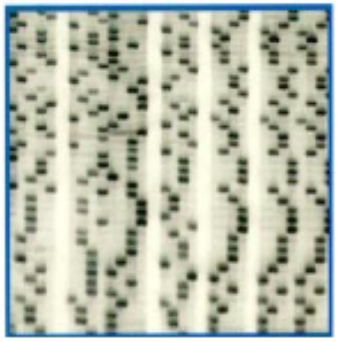
cca 100 000 bp/den





Evoluce Sangerova sekvenování

Pre-1992
"old fashioned way"



S35 ddNTPs
Gels
Manual loading
Manual base calling

1992-1999
ABI 373/377

Fluorescent ddNTPs*
Gels
Manual loading
Automated base calling*

1999
ABI 3700

Fluorescent ddNTPs
Capillaries*
Robotic loading*
Automated base calling
Breaks down frequently

2003
ABI 3730XL

Fluorescent ddNTPs
Capillaries
Robotic loading
Automated base calling
Reliable*

96-kapilární sekvenátor

=

2304 x 500 bp/12 hodin

=

cca 2 400 000 bp/den

NGS (Illumina HiSeqX10)

=

cca 600 000 000 000 bp/den

electrophoresis

Next-generation sequencing (NGS)

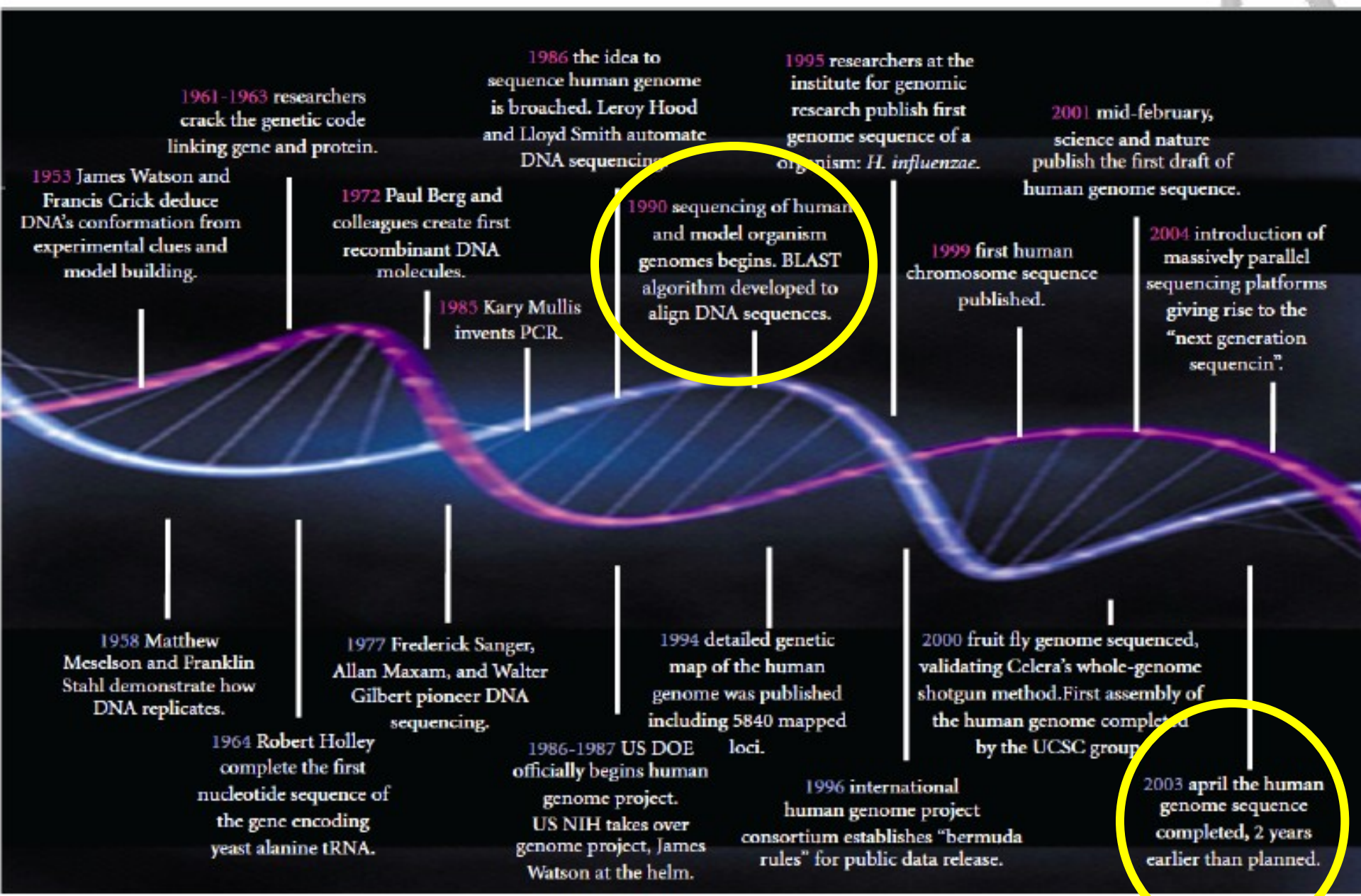
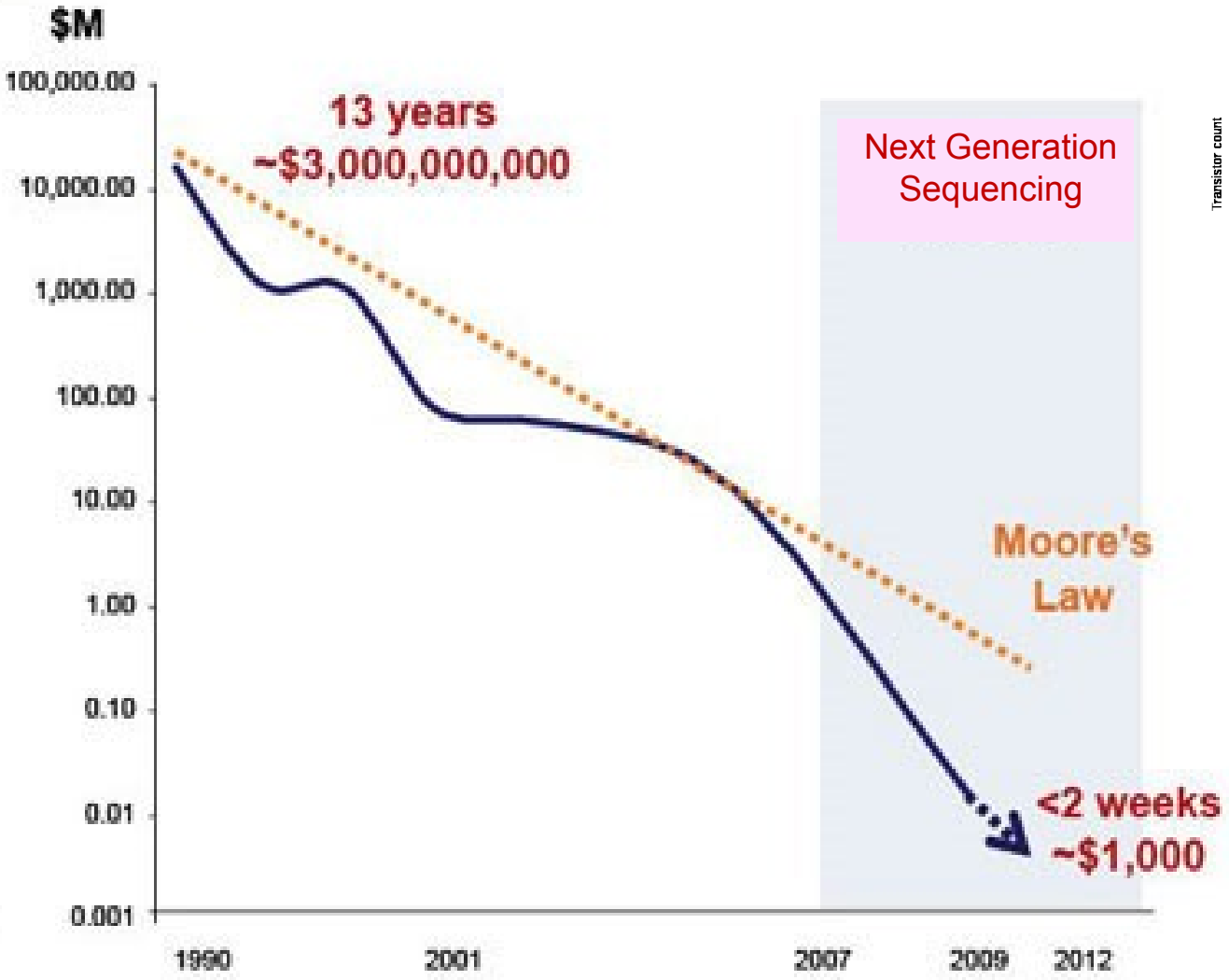
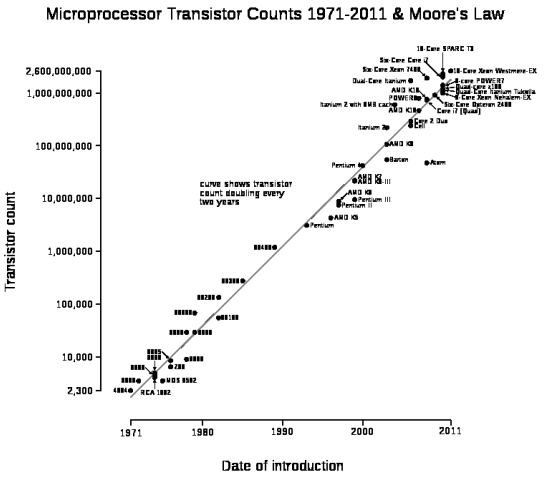


FIGURE 1: Evolution of DNA revolution.

Cost per Human Genome





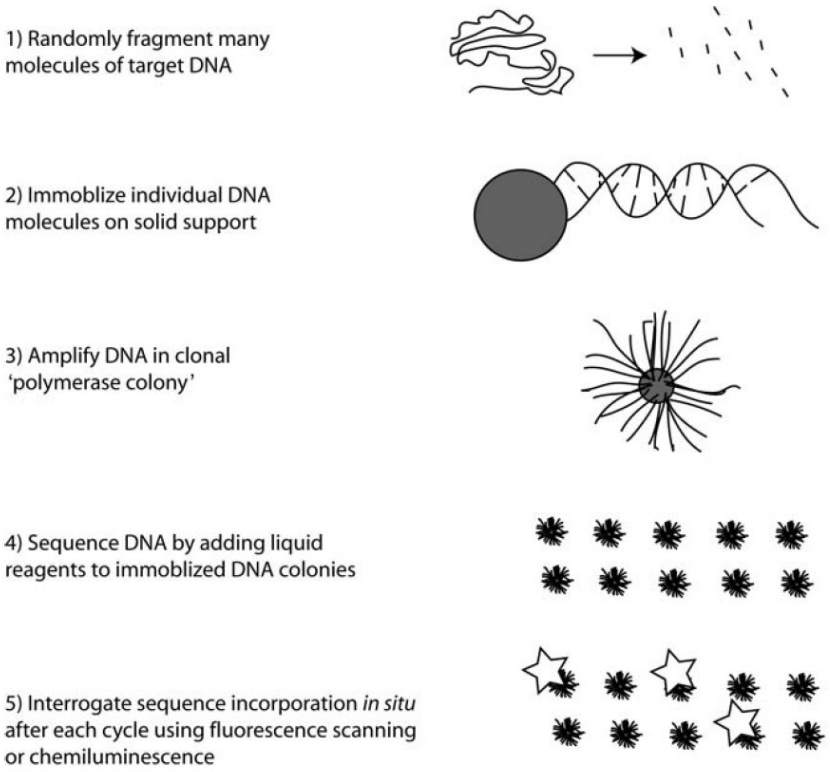
Illumina HiSeqX10

$1 M per machine

1.8 Tbase per machine per 3 days

1800 human genomes per machine per year

Historie „Next generation sequencing“



454 pyrosequencing ... první komerčně dostupná NGS technologie od srpna 2007

2016 – ohlášené stažení z trhu (Roche)

Široké spektrum technologií

Ale jen některé přežijí

Dnes dostupné NGS platformy

- **Roche 454**
- **Illumina HiSeq a MiSeq**
- ABI SOLiD
- IonTorrent (Life Technologies)
- SMRT (Pacific Biosciences)
- Oxford Nannopore
- …

454 pyrosequencing

- emulzní techniky amplifikace
  pikolitrové objemy
- simultánní sekvenování
  na destičce z optických vláken
  detekce pyrofosfátů uvolňovaných při inkorporaci bazí

- První generace GS20
  → 200 000 reakcí najednou
  (zhruba 20 milionů bp)
  FLX systém → 400 000 reakcí najednou =
  eukaryotní genom za týden!!!
- Délka jednotlivých sekvencí 100 – 400 (800 bp)

Molecular Ecology (2008) 17, 1629–1635

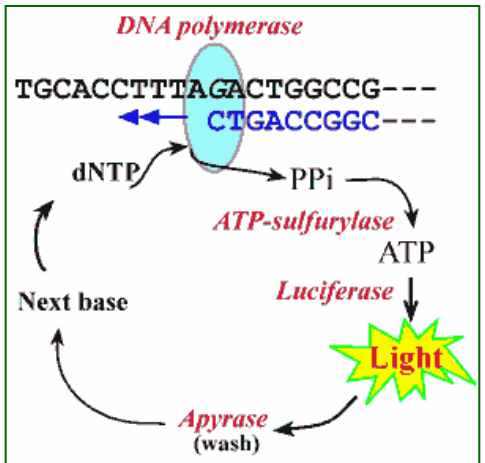
NEWS AND VIEWS

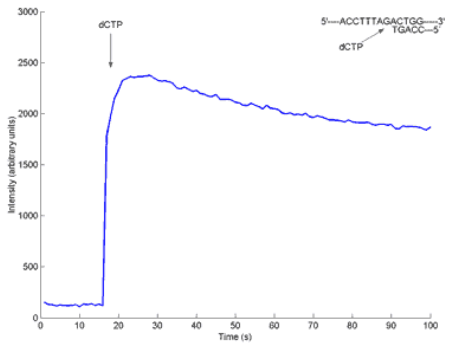
PERSPECTIVE

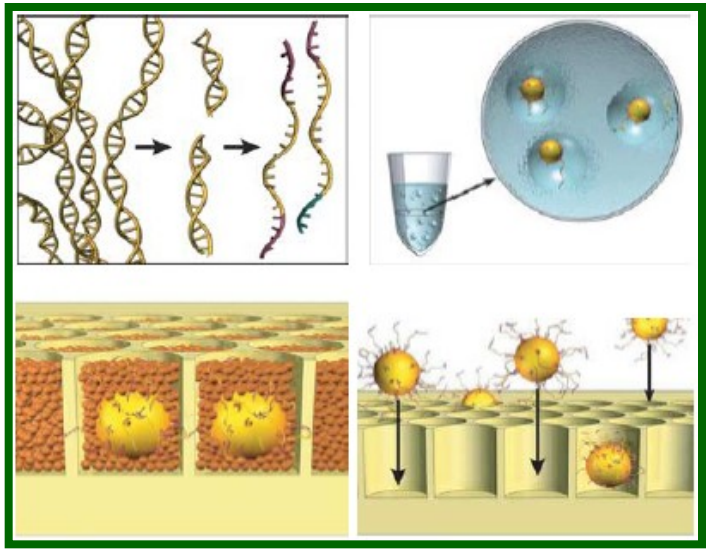
Sequencing goes 454 and takes large-scale genomics into the wild

HANS ELLEGREN
Department of Evolutionary Biology, Uppsala University, Norbyvägen 18D, SE-75236 Uppsala, Sweden

1 600 000 well plate







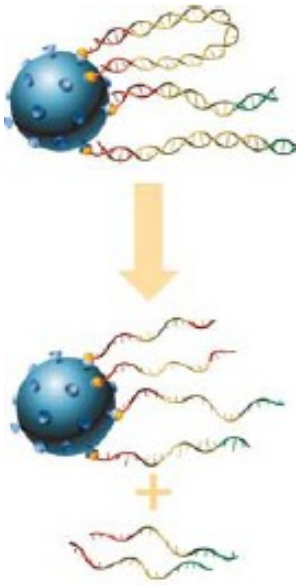
1. Příprava jednořetězcové DNA knihovny (ssDNA library preparation)

1 DNA Fragmentation (Nebulization):

5 Adaptor Ligation:

7 Library Immobilization:

9 ssDNA Library Isolation:



Adaptor A + **Adaptor B**

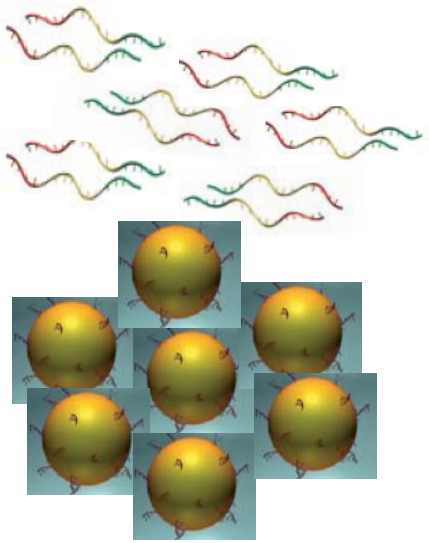
-Slouží jako vazebné místo primerů pro následnou PCR amplifikaci a sekvenování

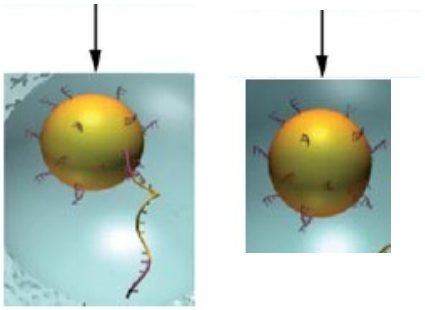
-Slouží k uchycení na kuličky (na adaptor B je připojen biotin)

2. Namnožení každé jednotlivé molekuly pomocí emulzní PCR (emPCR)

1 DNA Library Capture:

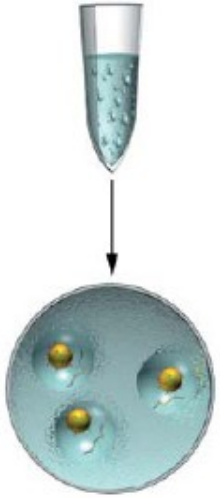
- poměry nastavit tak aby 1kulička ≤ 1 molekula DNA



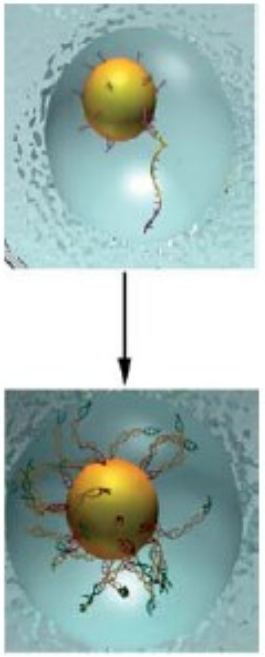


2 Preparation of the Amplific. Mixes

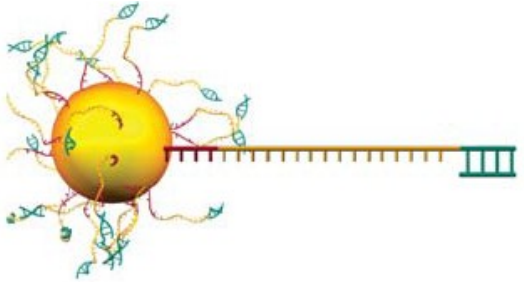
3 Emulsification:



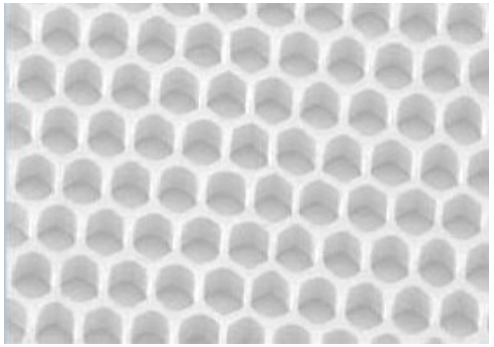
4 emPCR Amplification:



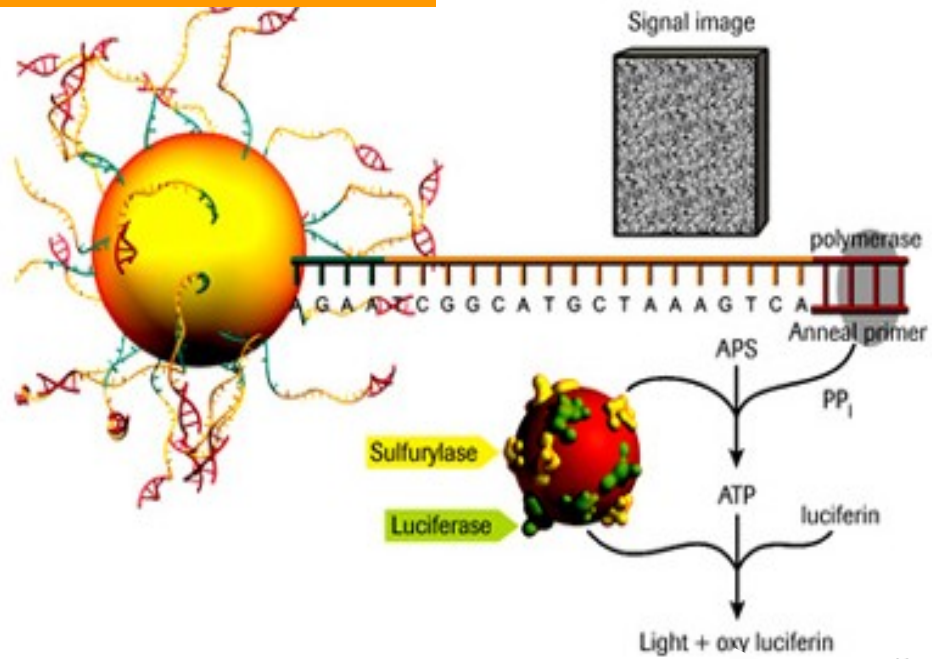
7 Sequencing Primer Annealing:



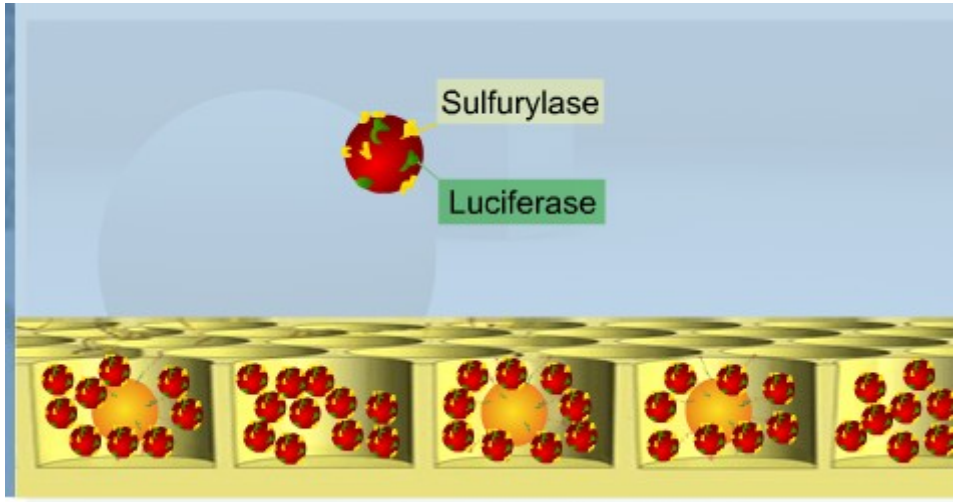
3. Pyrosekvenování („sequencing by synthesis“)

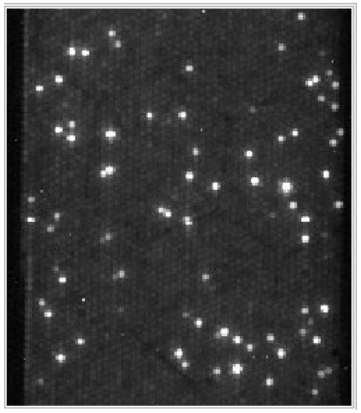


pikotitrační destička



Na jedné desticče 400 000 až 1milión jamek

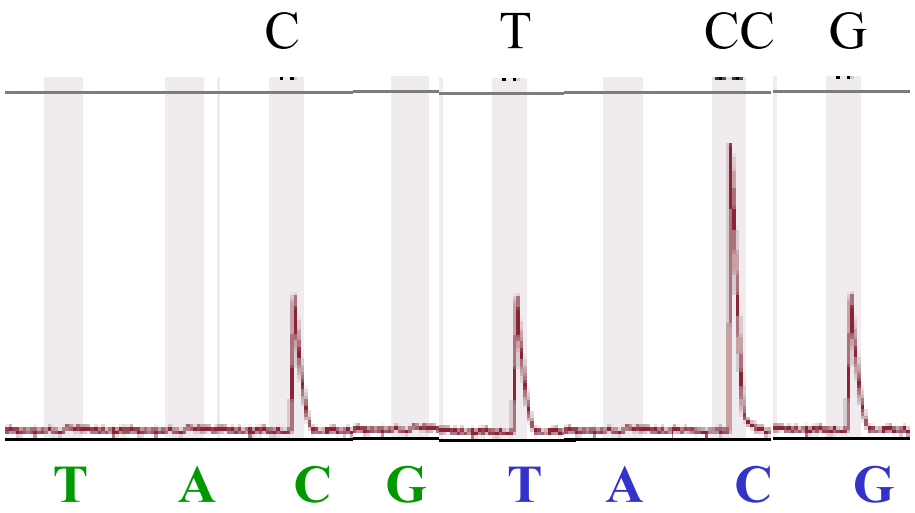




3. Pyrosekvenování – detekce signálu

- postupně se přidávají nukleotidy v definovaném pořadí: např. TACG TACG TACG
- po přidání každého nukleotidu a detekci signálu se nukleotid odmyje a přidá se další

DNA sekvence: **C T C C G**



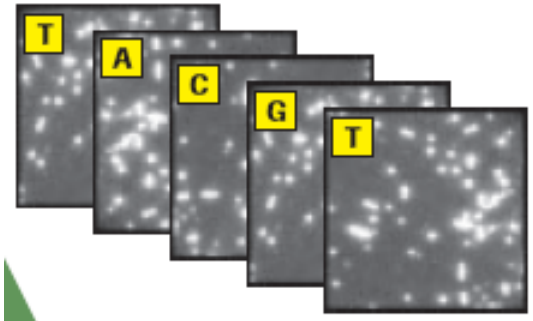


Image Files:
12-15 gigabytes
per run

Problém!!!! Homoplymery např. AAAAAAAAAA

http://www.youtube.com/watch?v=bFNjxKHP8Jc

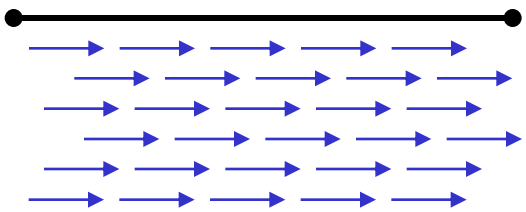
High-throughput – paralelní sekvenování

454 Platform Updates

GS20	• 100bp reads, ~20Mbp / run
GS-FLX	• 250bp reads ~100 Mbp / run (7.5 hrs)
GS-FLX Titanium	• 400bp reads ~400 Mbp / run (10 hrs)
GS-FLX Titanium Plus	• 700 bp reads ~700 Mbp/run (18 hrs)
GS Junior	• 400 bp reads ~ 35Mbp/run (10 hrs)

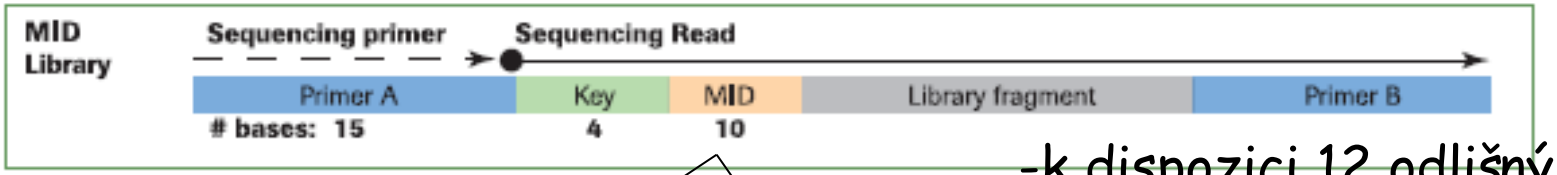
!!! Samozřejmě nestačí mít každou bázi osekvenovanou 1x !!!

- Pospojování (reads assembly) do souvislé sekvence
- Nepřesnosti – pokrytí (coverage)

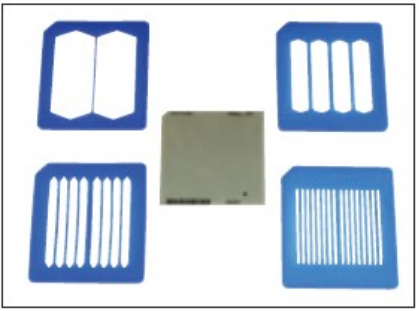


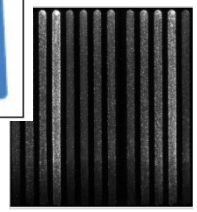
Kapacita destičky 400 Mb (GS FLX Titanium):

Mus:	2700 Mb	→ 7 run 1x coverage
Caenorhabditis:	100 Mb	→ 1 run 4x coverage
E. coli:	5 Mb	→ 1 run 80x coverage
mitoch. Mus:	0.016 Mb	→ 1 run 25000x coverage
HIV:	0.01 Mb	→ 1 run 40000x coverage

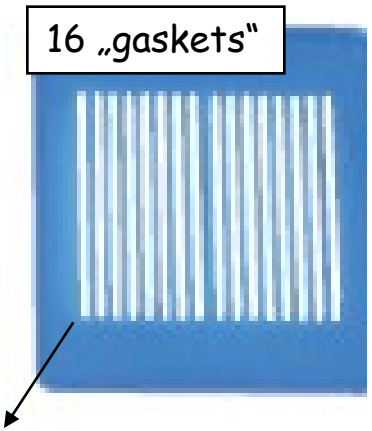


-k dispozici 12 odlišných MID („multiplexing")





1. *CCCCCCCCCC*
2. *GGGGGGGGG*
.
.
.
12. *CCCCCAAAG*



12 MID
X
16 gaskets
=
max. 192 vzorků

V každém max. 12 vzorků
(každý označen svým MID)

Illumina HiSeq/MiSeq

- v současné době nejrozšířenější typ (cca 70%) na trhu
- v horizontu následujících let její používání spíš poroste

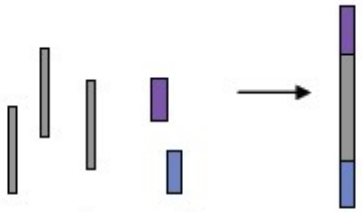
https://www.youtube.com/watch?annotation_id=annotation_228575861&feature=iv&src_vid=womKfikWlxM&v=fCd6B5HRaZ8

Illumina HiSeq

Illumina MiSeq

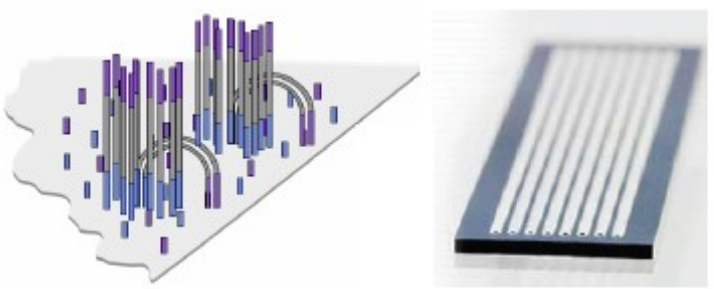
Illumina Sequencing pipeline

1. Sample Prep
(1-5 days)



Ligate adapters

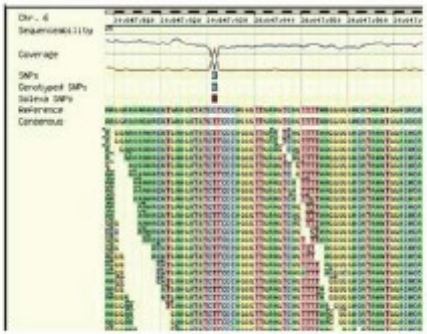
2. Cluster generation on flow cell
(1.5 day)



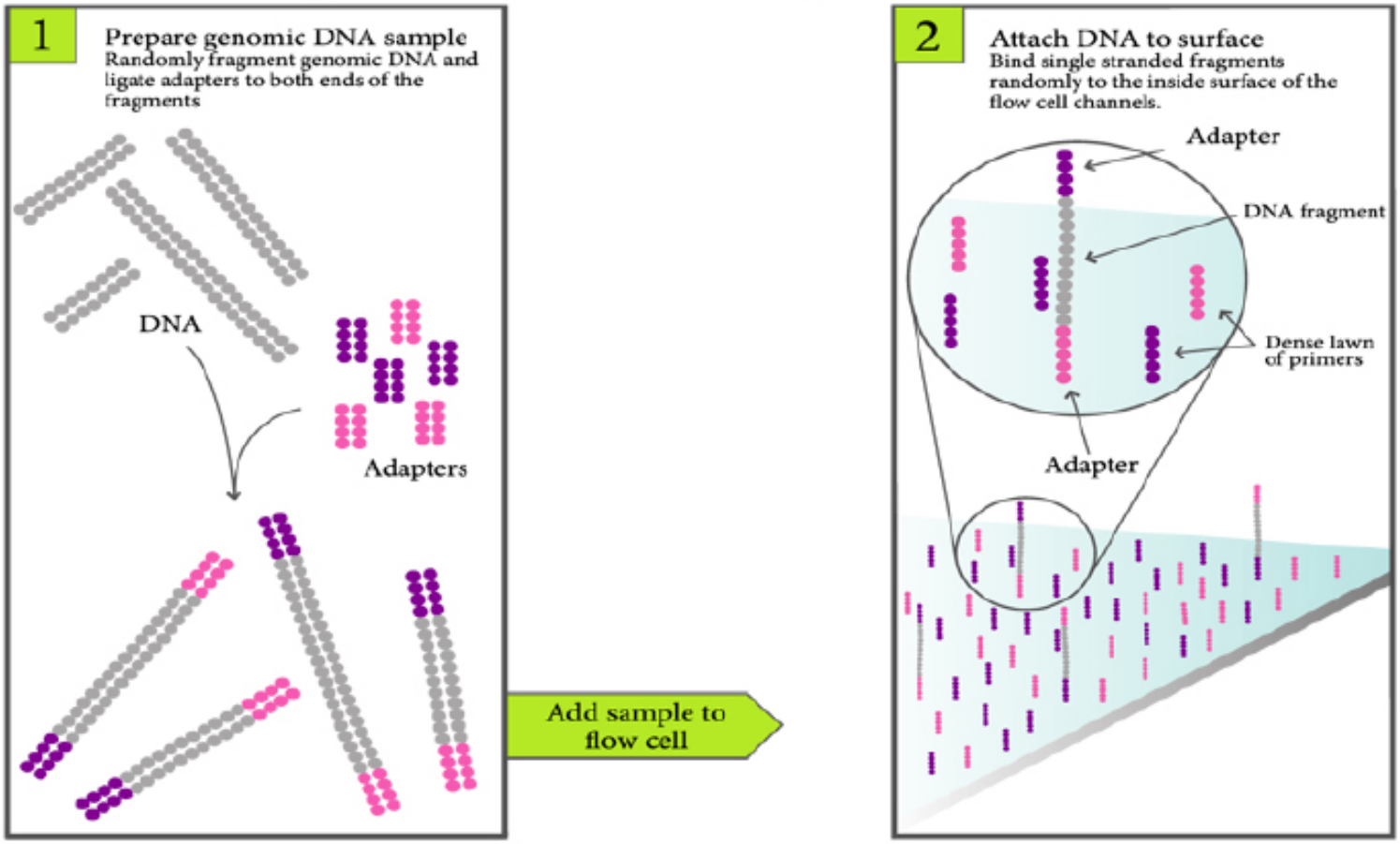
Clonal Single molecular Array

3. Sequencing and imaging
(2-3 days)

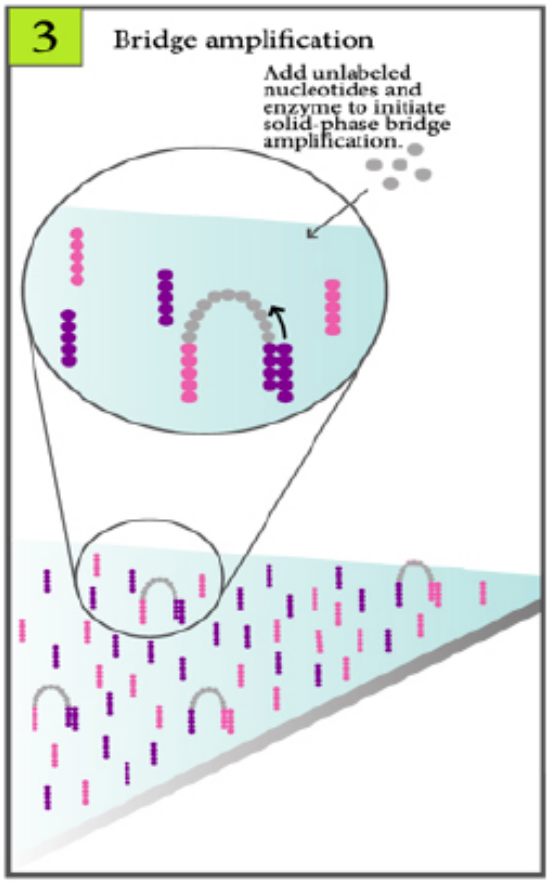
4. Data Analysis
(days-months)

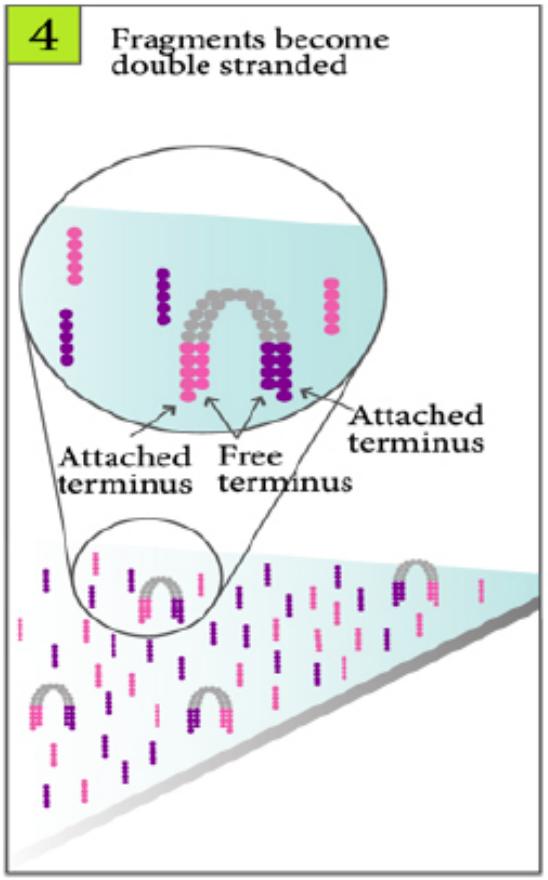


Attach DNA to flow cell

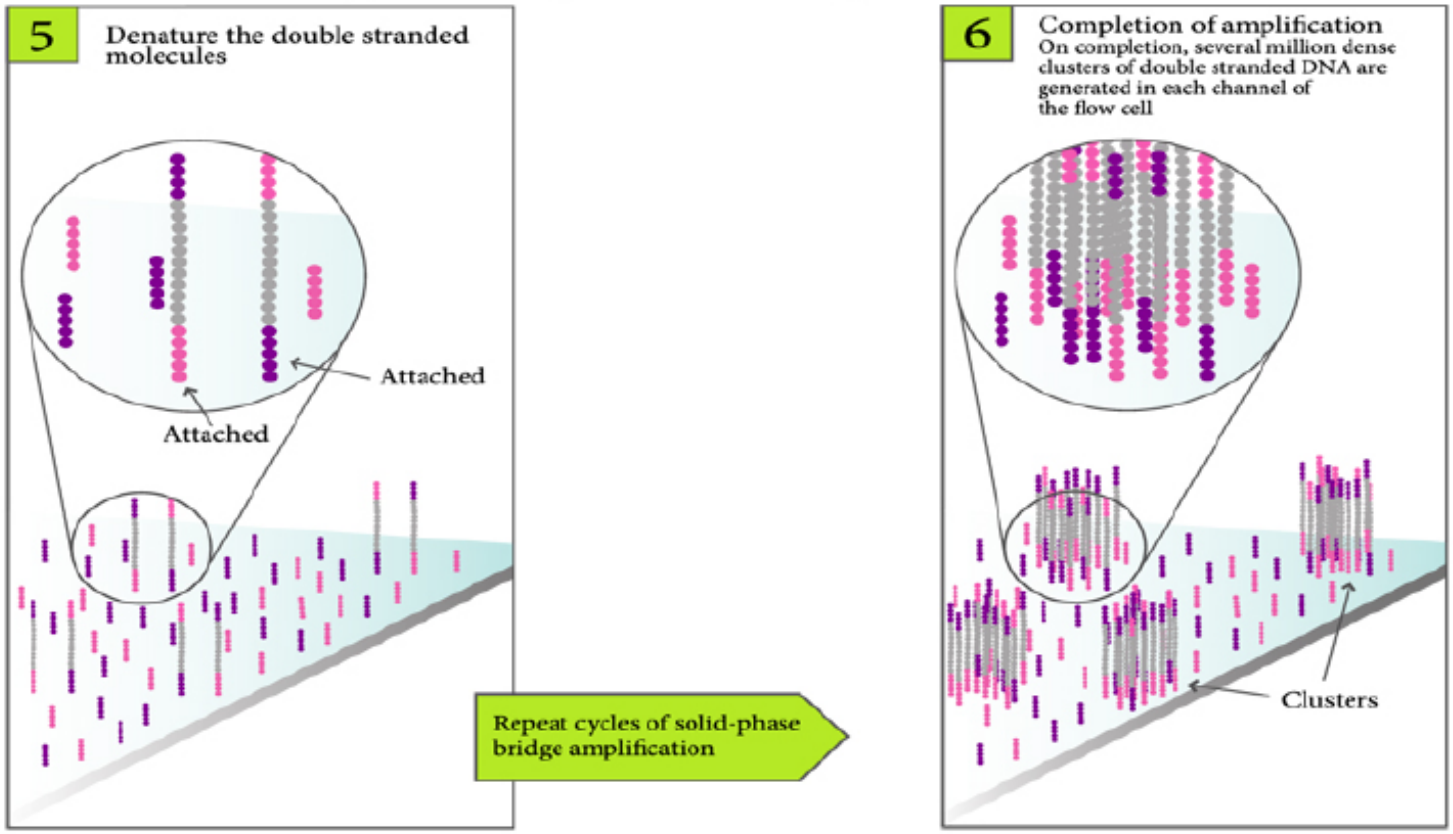


Bridge Amplification



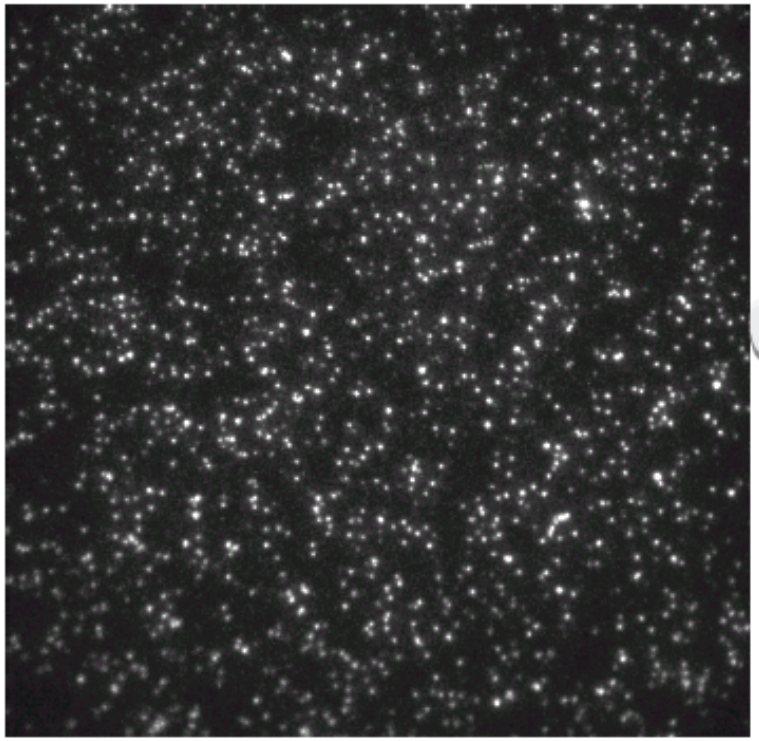


Cluster Generation



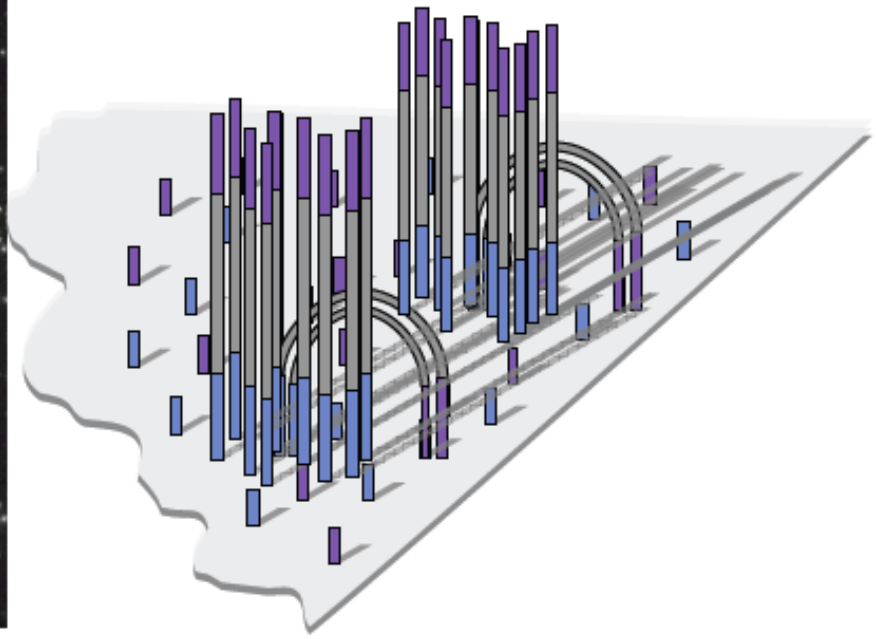
Clonal Single molecular Array

Clonal Single molecule Array



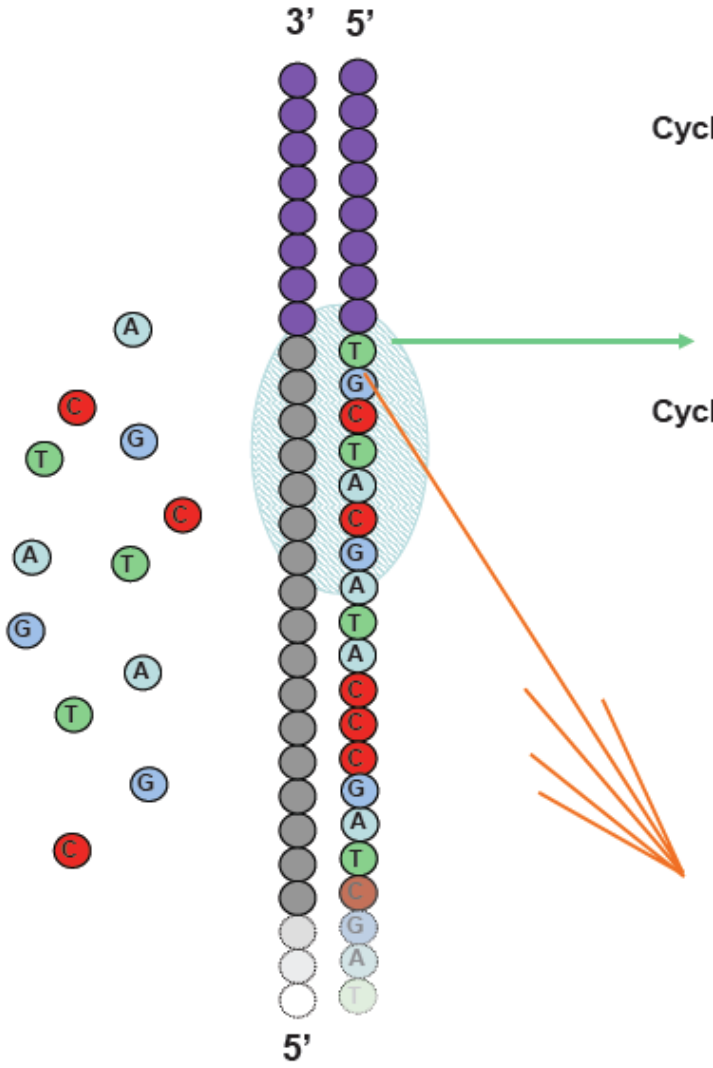
100um

Random array of clusters



~1000 molecules per ~ 1 um cluster
~20-30,000 clusters per tile
~40 M clusters per flowcell

Sequencing By Synthesis (SBS)



Cycle 1: Add sequencing reagents

First base incorporated

Remove unincorporated bases

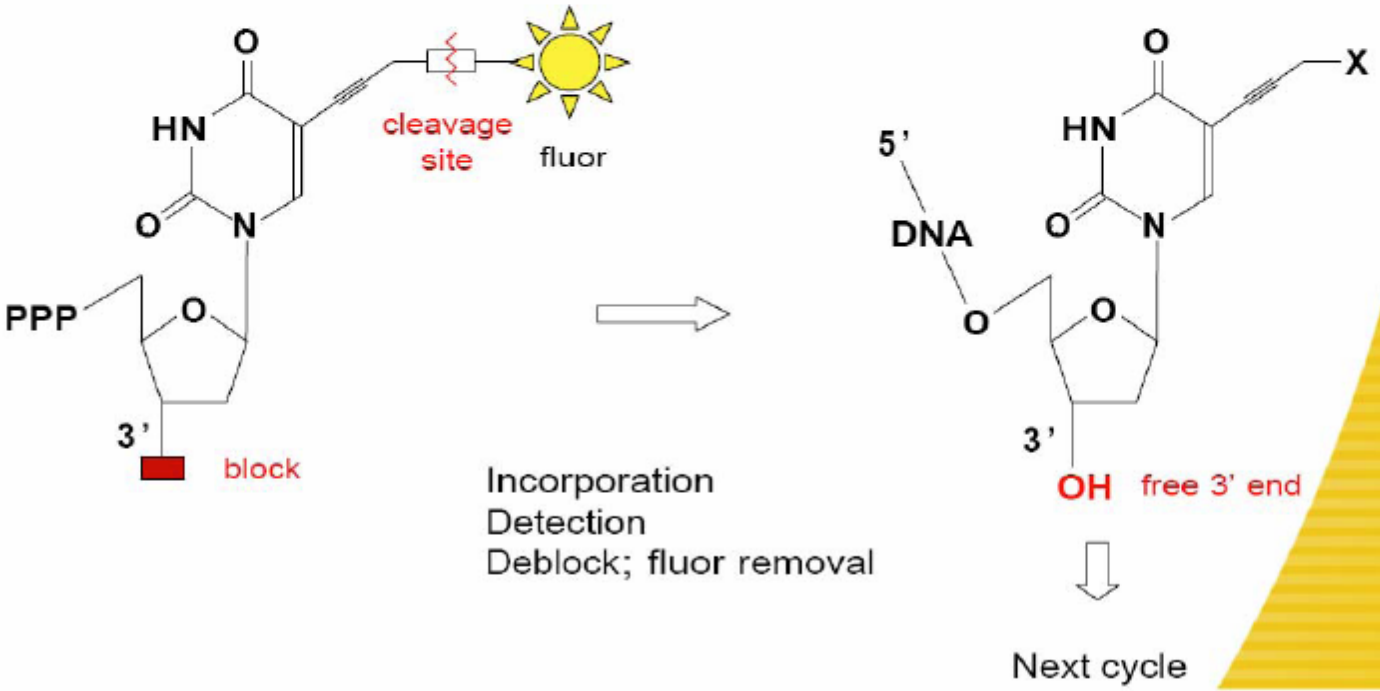
Detect signal

Cycle 2-n: Add sequencing reagents and repeat

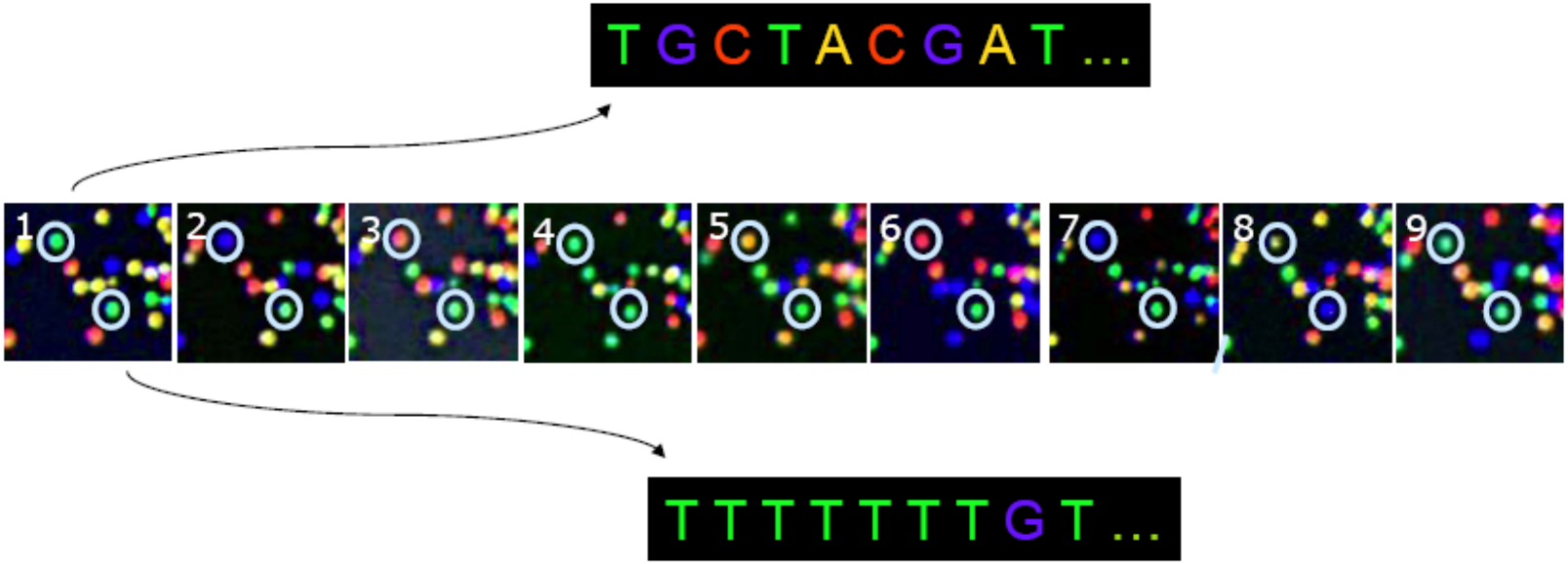
Reversible Terminator Chemistry

Solexa

- All 4 labelled nucleotides in 1 reaction
- Higher accuracy
- No problems with homopolymer repeats

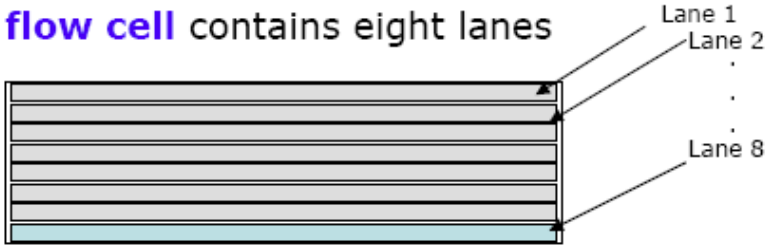


Base Calling From Images

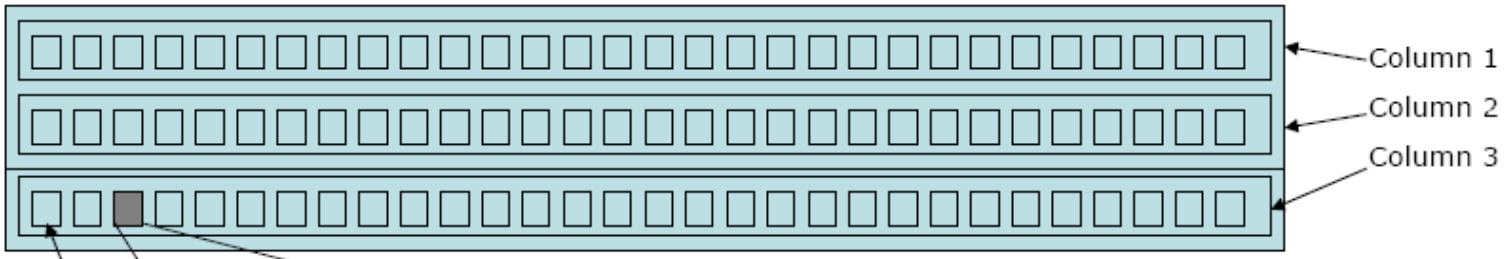


The identity of each base of a cluster is read off from sequential images

A **flow cell** contains eight lanes

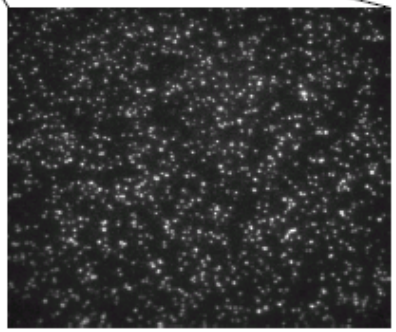


Each **lane/channel** contains **three columns** of tiles



Each **column** contains **100 tiles**

Tile



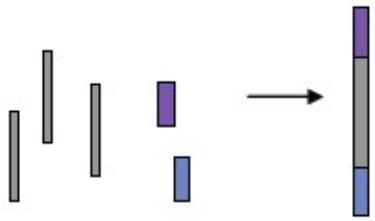
20K-30K Clusters

350 X 350 µm

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

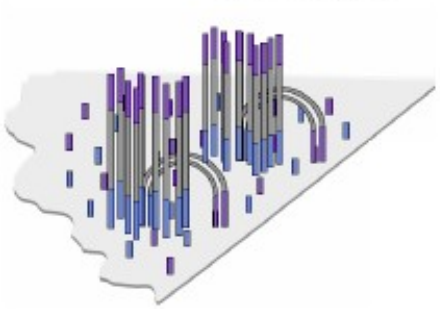
Illumina Sequencing pipeline

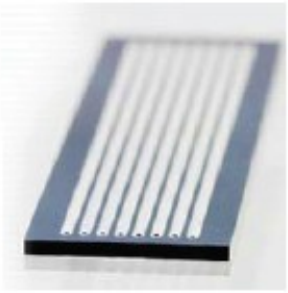
1. Sample Prep
(1-5 days)



Ligate adapters

2. Cluster generation on flow cell
(1.5 day)

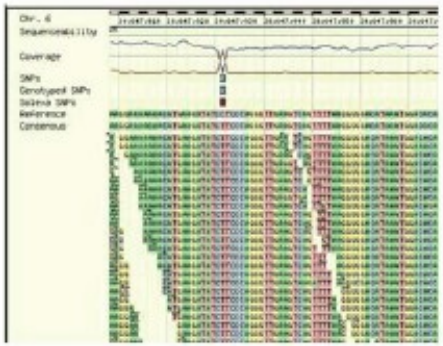




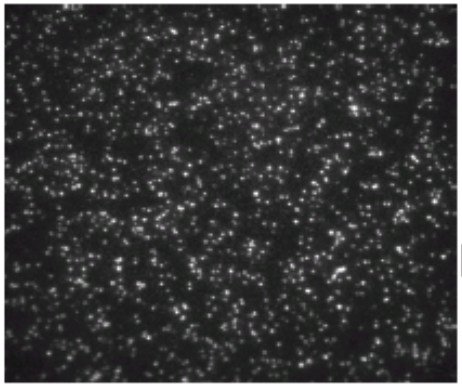
Clonal Single molecular Array

3. Sequencing and imaging
(2-3 days)

4. Data Analysis
(days-months)

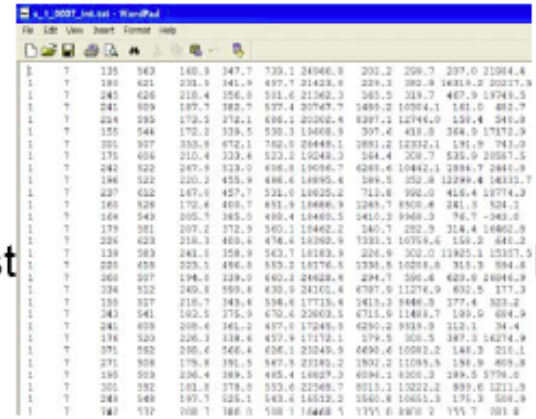


Data Analysis Pipeline



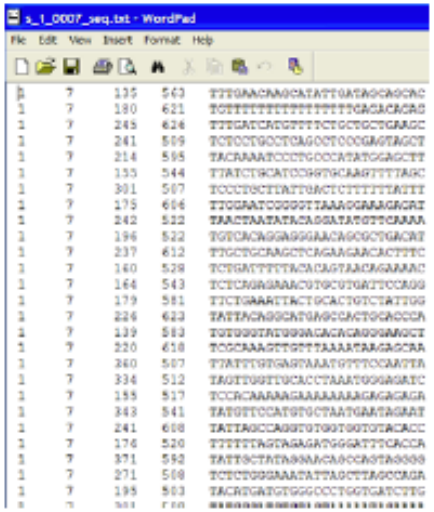
tiff image files
(345,600)

Firecrest



intensity files

Bustard



Sequence files

Eland

Alignment to Genome

Additional
Data Analysis

Illumina fastq

```
      1              2        3     4     5    6 7      8
@HWI-ST226:253:D14WFACXX:2:1101:2743:29814 1:N:0:ATCACG
TGCGGAAGGATCATTGTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTT
GAAAAAAAAAAAAAAAAATTA
+
B@CFFFFFHHFFHJIIGHIHIJJIJIIJJGDCHIIIJJJJJJJGJGIHHEH@)=F@EIGHHEHFFFFDCBBD:@CC@C
:<CDDDD50559<B########
```

1. unique instrument ID and run ID
2. Flow cell ID and lane
3. tile number within the flow cell lane
4. 'x'-coordinate of the cluster within the tile
5. 'y'-coordinate of the cluster within the tile
6. the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)*
7. N if the read passes filter, Y if read fails filter otherwise
8. Index sequence

All this generates a lot of Data!
1.5 TB data/run

- 1 Gig of Space
  - 125,000 pages of text
  - 11 CDs of Music
  - 4000 (1024x768) JPEG images
  - 40,000 pages of PDF
- 1 TB of space
  - 220 Million pages of text
  - 300 hours of video
  - 4,000,000 JPEG images
  - 1,000 copies of the Encyclopedia Britannica
  - 1/10 of the printed Library of Congress

Illumina sequencers

Illumina MiSeq
4 millions reads/run
150 bp/read

Illumina GAIIx
300 millions reads/run
150 bp/read

Illumina HighSeq
1500 – 3000 millions reads/run
100 bp/read

NovaSeq 6000 Sequencing System (2017)

ca. 48 human genomes/run

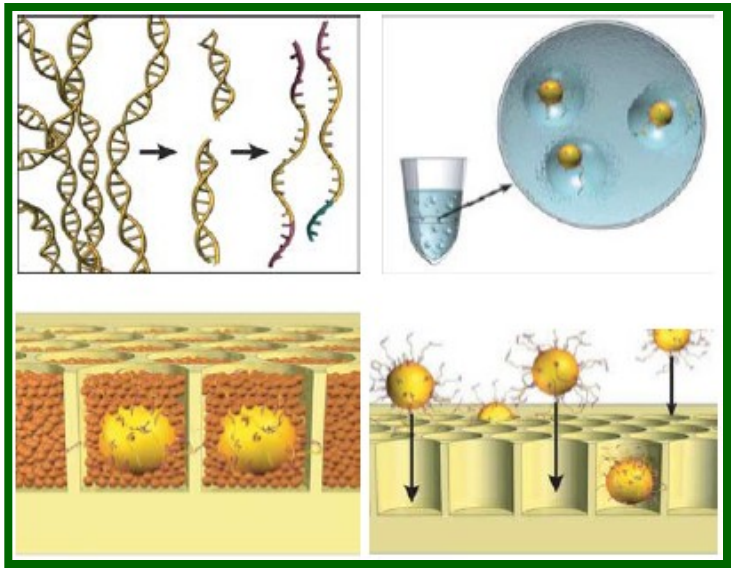
Sequencing Output per Flow Cell

	NovaSeq 6000 System		
Flow Cell Type	S1	S2	S4
2 × 50 bp	134–167 Gb	333–417 Gb	N/A*
2 × 100 bp	266–333 Gb	667–833 Gb	N/A*
2 × 150 bp	400–500 Gb	1000–1250 Gb	2400–3000 Gb

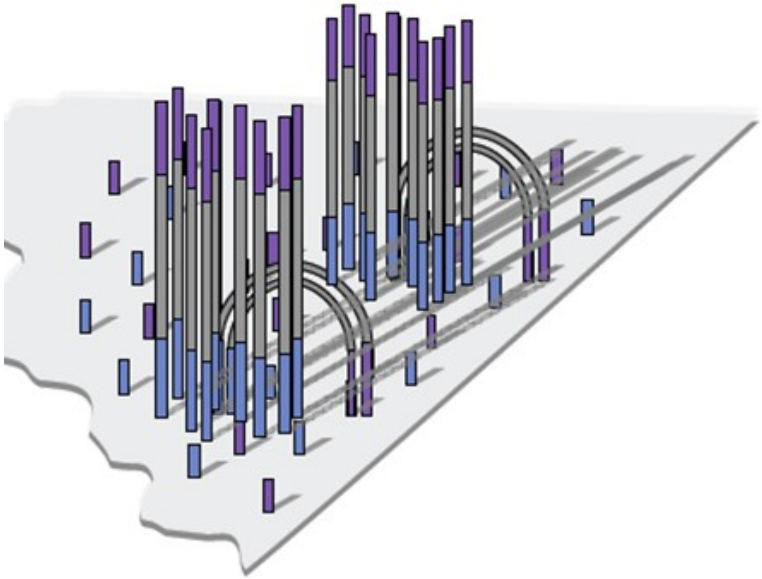
Specifications based on Illumina PhiX control library at supported cluster densities.
* N/A: not applicable

https://emea.illumina.com/systems/sequencing-platforms/comparison-tool.html?langsel=/cz/

NGS technologie



454 pyrosequencing
(Roche)



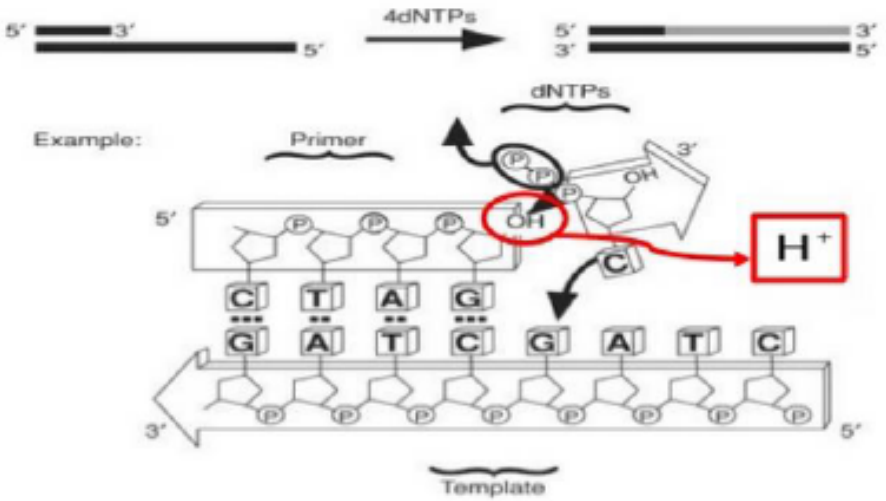
Illumina

Ion Torrent technology

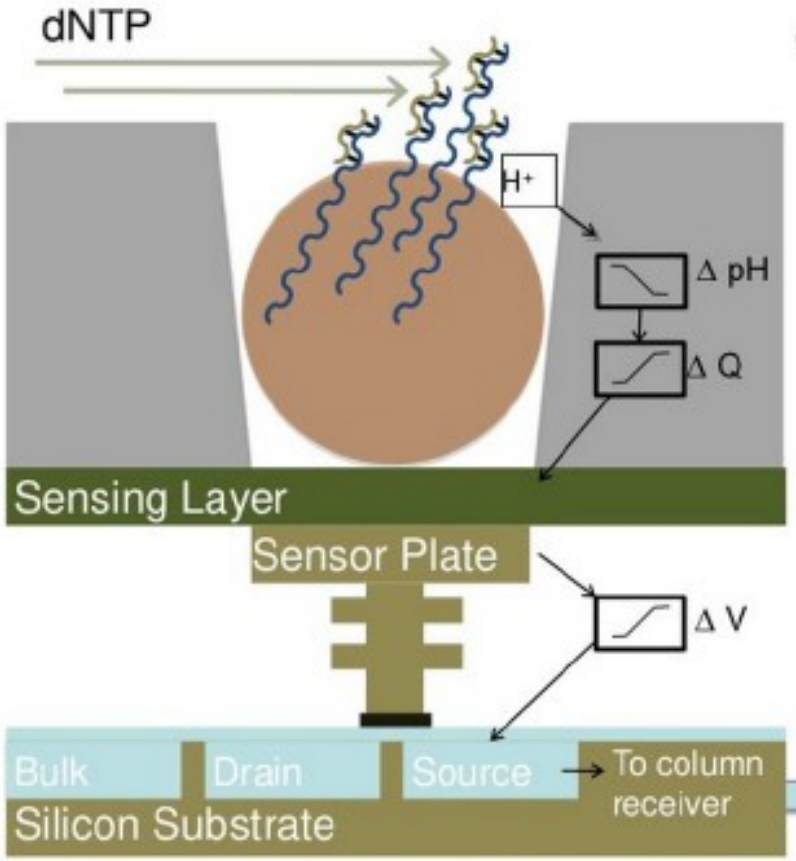
Ion sequencing: Life Technologies

Využívá změny pH při syntéze DNA

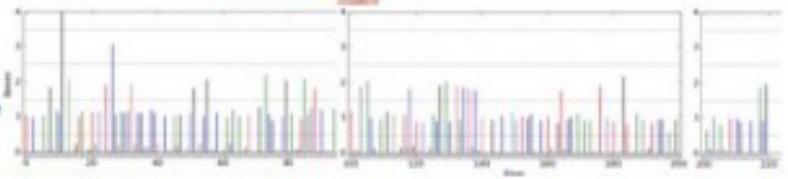
- Ion Semiconductor Sequencing
- Detection of hydrogen ions during the polymerization DNA
- Sequencing occurs in microwells with ion sensors
- No modified nucleotides
- No optics



Ion Torrent



- DNA → Ions → Sequence
  - Nucleotides flow sequentially over Ion semiconductor chip
  - One sensor per well per sequencing reaction
  - Direct detection of natural DNA extension
  - Millions of sequencing reactions per chip
  - Fast cycle time, real time detection

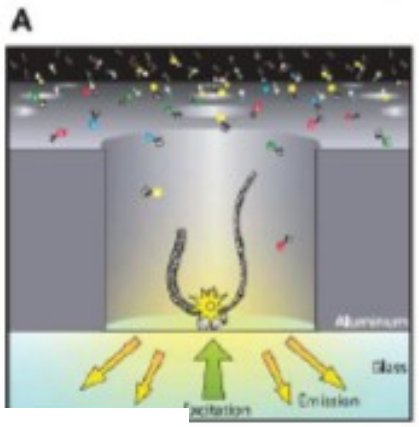


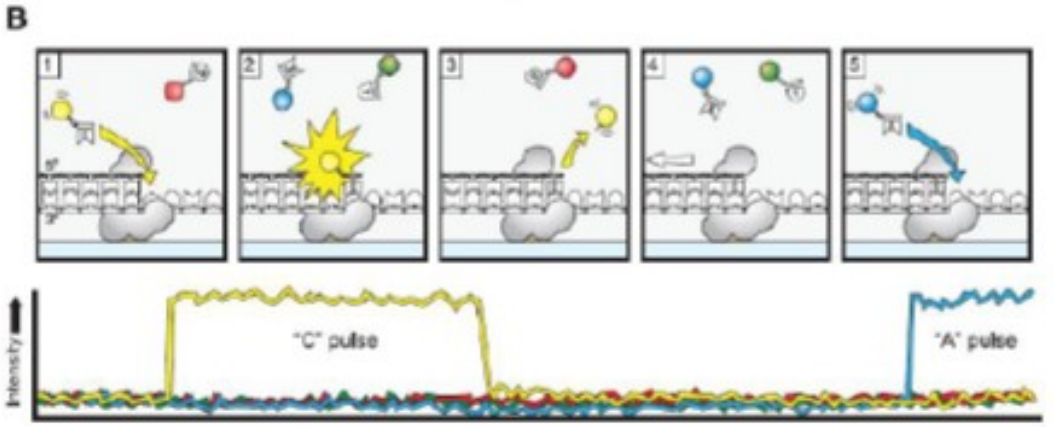
Ion Torrent: System Updates

314 Chip	• 100bp reads ~10 Mb/run (1.5 hrs)
316 Chip	• 100 bp reads ~100 Mbp / run (2 hrs) • 200 bp reads ~200 Mbp/run (3 hrs)
318 Chip	• 200 bp reads ~1 Gbp / run (4.5 hrs)

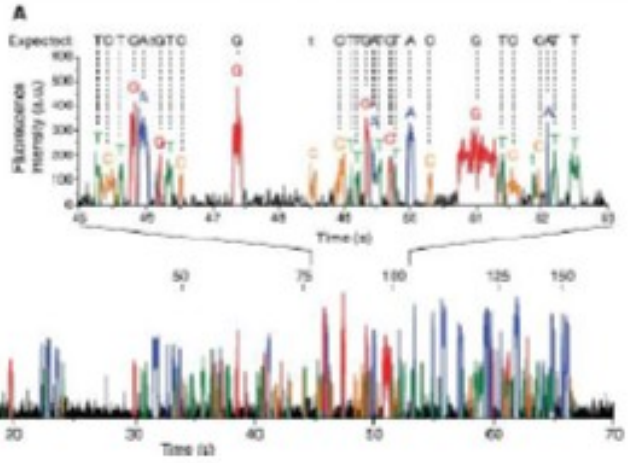
400 bp reads

SMRT („single molecule real-time sequencing") – Pacific Biosciences



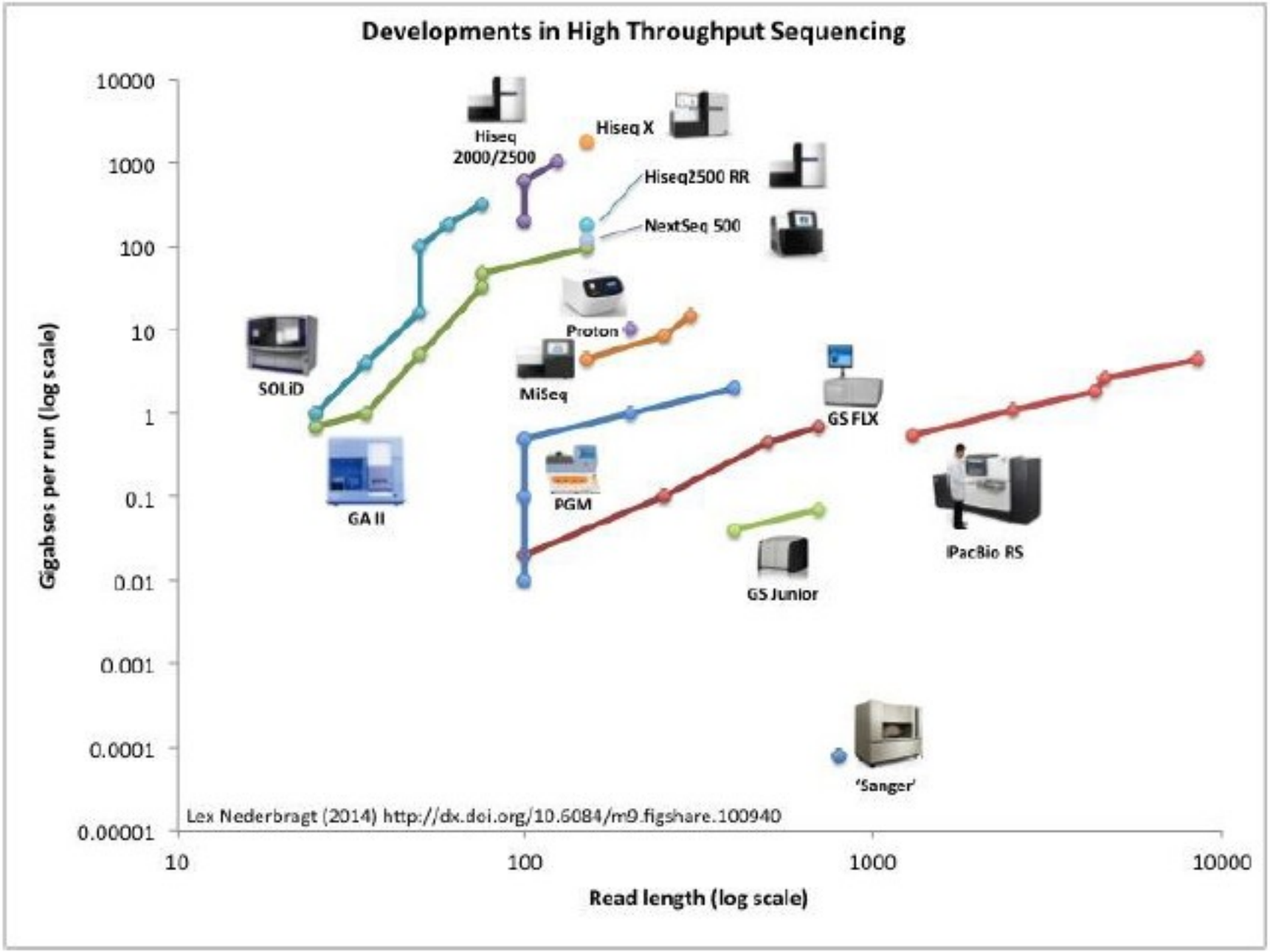


Pacbio RS – raw data



dlouhé čtení (15 kb), hodně chyb

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


Developments in High Throughput Sequencing
Gigabses per run (log scale)
10000
1000
100
10
1
0.1
0.01
0.001
0.0001
0.00001
10
100
1000
10000
Read length (log scale)
Hiseq 2000/2500
Hiseq X
Hiseq2500 RR
NextSeq 500
Proton
SOLiD
MiSeq
GS FLX
GA II
PGM
PacBio RS
GS Junior
'Sanger'
Lex Nederbragt (2014) http://dx.doi.org/10.6084/m9.figshare.100940

3rd generation: Oxford Nannopore

MinION

512 pores

GridION

5 000 pores

Future Sequencing Technologies

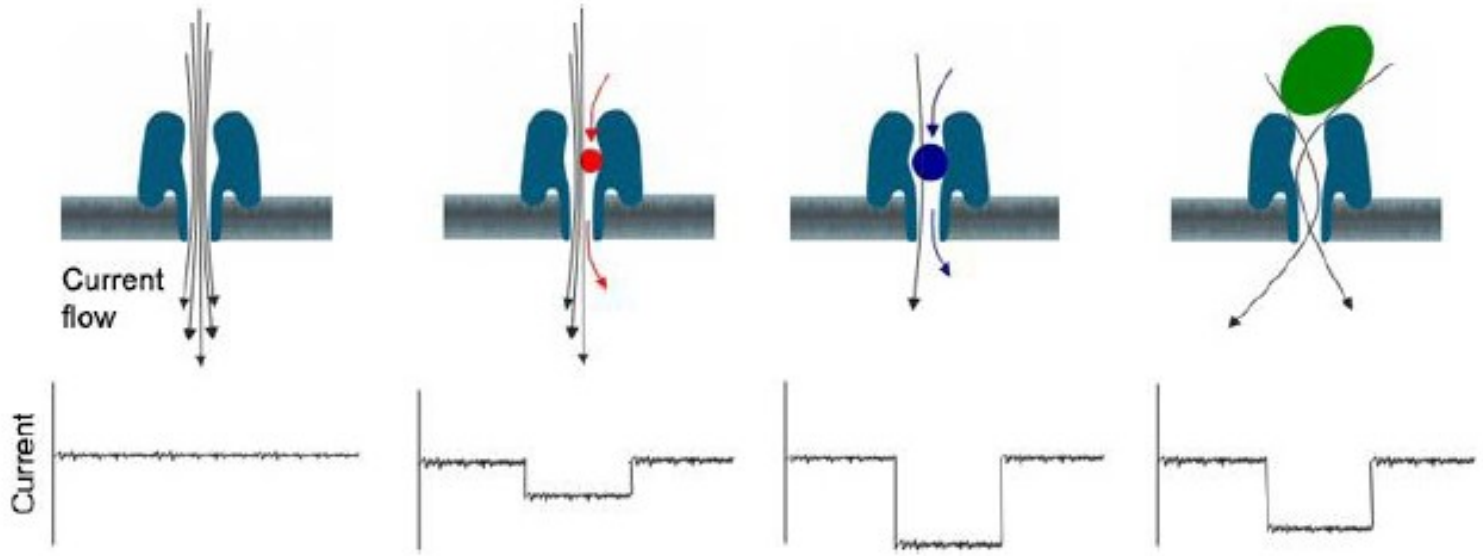
Oxford Nanopore

Nanopore sequencing

up to 50 kb

„Run until sequencing ...“

Princip technologie



http://www.youtube.com/watch?v=3UHw22hBpAk

Sekvenování přímo v terénu (?)

Ebola outbreak

Quick et al., Nature 2016

Přehled současných metod NGS

Platform	Year	Sequencing Method	Amplification	Detection	Features
454	2005	Pyro-sequencing	Emulsion PCR	Light	First NGS
Illumina	2007	Synthesis	Bridge PCR	Light	90% of Market
SOLiD	2008	Ligation	Emulsion PCR	Light	Lowest Error Rate
Ion Torrent	2010	Synthesis	Emulsion PCR	Hydrogen Ion	Semiconductor Chip
Pacific Biosciences	2010	Synthesis	None = Single Molecule	Light	Anchored Polymerases
Oxford Nanopore	2012	Nanopore	None = Single Molecule	Electrical Conductivity	"Run Until" Sequencing

Výkonnost jednotlivých metod

Instrument	Run time	Millions of Reads/run	Bases / read	Yield MB/run
3730xl (capillary)	2 hrs	0.000096	650	0.06
PacBio RS	2 hrs	0.01	860 – 1,500	5-10
454 GS Jr. Titanium	10 hrs	0.1	400	50
Ion Torrent – 314 chip	2.5 hrs	0.25	200	50
454 FLX Titanium	10 hrs	1	400	400
454 FLX+	20 hrs	1	650	650
Ion Torrent – 316 chip	3 hrs	1.6	200	320
Illumina MiSeq	26 hrs	4	150+150	1200
Ion Torrent – 318 chip	4.5 hrs	4	200	800
Illumina GAIIx	14 days	300	150+150	96,000
SOLiD – 5500xl	8 days	>1,410[d]	75+35	155,100
Illumina HiSeq 1000	8.5 days	≤1500	100+100	≤300,000
Illumina HiSeq 2000	11.5 days	≤3000	100+100	≤600,000

2012. NGS Field Guide (www.molecularecologist.com)

Chybovost jednotlivých metod

Platform	Primary Errors	Single-pass Error Rate (%)	Final Error Rate (%)
3730xl (capillary)	Substitution	0.1-1	0.1-1
454	Indel	1	1
Illumina	Substitution	~0.1 (85% of reads)	~0.1 (85% of reads)
SOLiD	A-T bias	~5	≤0.1
Ion Torrent	Indel	~1	~1
PacBio RS	CG deletions	~15	≤15
Oxford Nanopore	Deletions	≥4	4

2012. NGS Field Guide (www.molecularecologist.com)

Traditional Sequencing vs. Next Generation Sequencing: Data Throughput

1 x Illumina GAII

Vs.

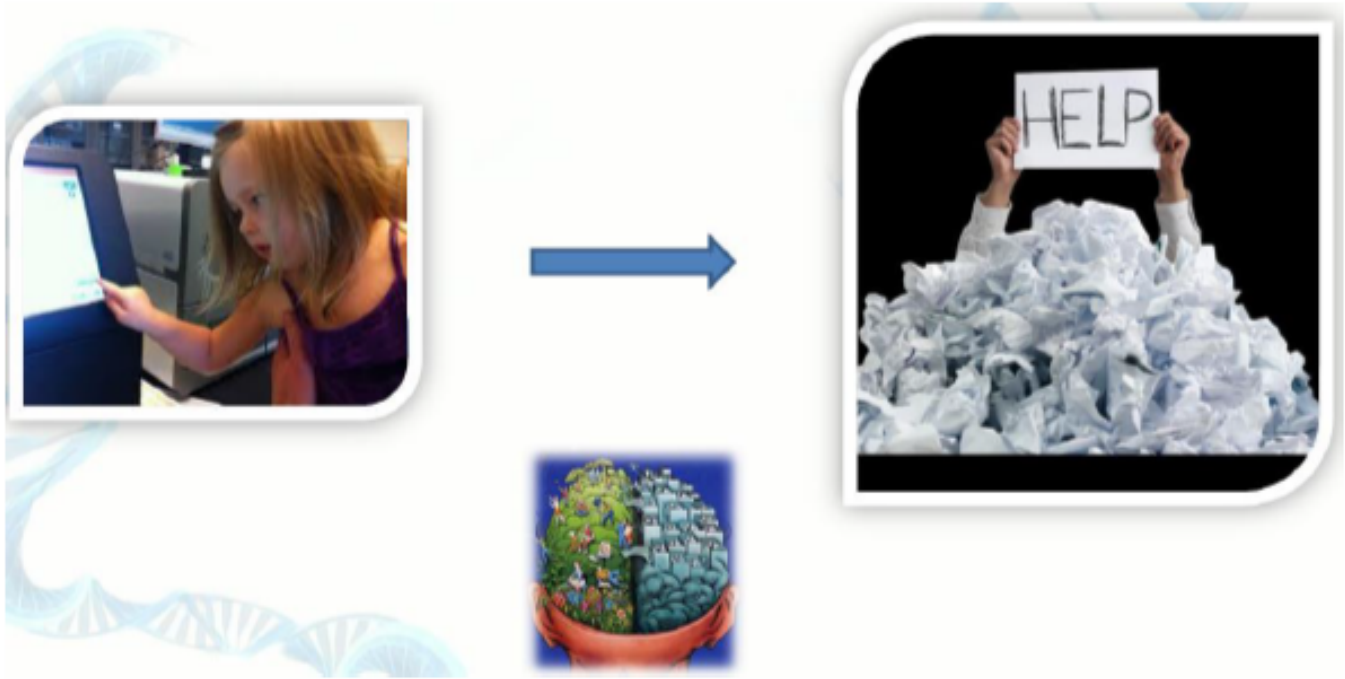
200+ of 3730xl

Days vs. Years

The Sequencing Landscape is Changing

Bioinformatika – největší brzda dalšího rozvoje

Basically, analyzing genomes in interaction with their environment is now feasible and accessible to anyone



Sekvenační strategie

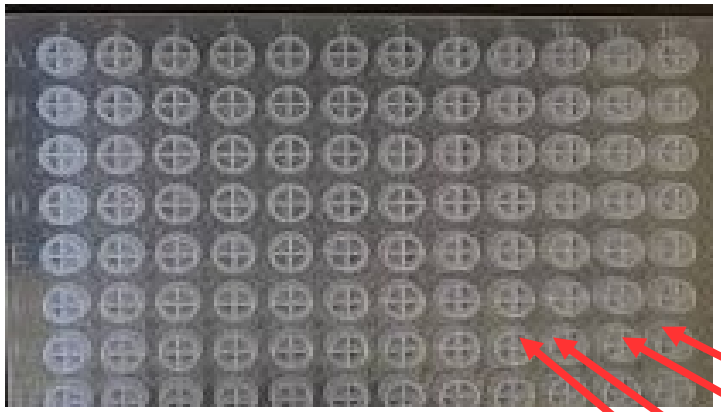
- nutno velmi dobře počítat než se začne sekvenovat
- celkový výtěžek sekvenování = **počet „reads“ * délka „reads“ * coverage**
- zásadně závisí na konkrétním cíli výzkumu a použité technologii

Sekvenační strategie

…JEDEN VZOREK NA RUN JE MÁLO

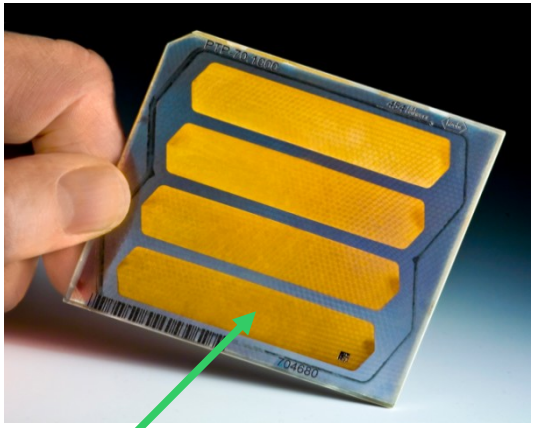
Kapilární sekvenátor

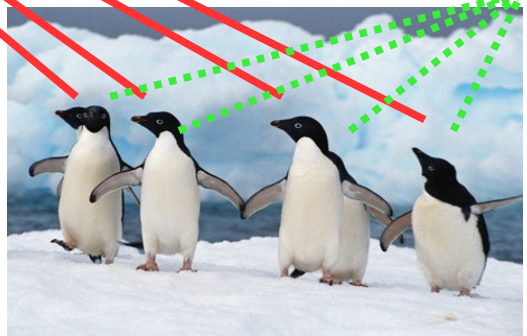
U kapilárních sekvenátorů není problém přiřadit sekvenci k jednotlivým vzorkům na základě pozice na platíčku



Sekvenátor druhé generace

U sekvenátorů druhé generace se najednou sekvenuje pool desítek až stovek vzorků





Sekvenační strategie

…JEDEN VZOREK NA RUN JE MÁLO

Jednotlivé vzorky pro sekvenátory druhé generace se značí tzv. barcody (midy, tagy)

Krátká (obvykle 6-12bp) oligonukleotidová sekvence před primerem (pokud sekvenujeme PCR amplikon), která je specifická pro daný vzorek

Přiřazení identity jednotlých sekvencí k vzorkům probíhá bioinformaticky

BARCODE PRIMER SEQUENCE

AGCGTAGGTCATTTCGATGCGGTCATGCCTGGATTAAAGCT.....
T**T**CGTAGGTCATTTCGATGCGGTCATGCCTGGATTAAAGCT.....
TG**G**GTAGGTCATTTCGATGCGGTCATGCCTGGATTAAAGCT.....
TGC**C**TAGGTCATTTCGATGCGGTCATGCCTGGATTAAAGCT.....
TGCG**C**AGGTCATTTCGATGCGGTCATGCCTGGATTAAAGCT.....
TGCGT**T**GGTCATTTCGATGCGGTCATGCCTGGATTAAAGCT.....

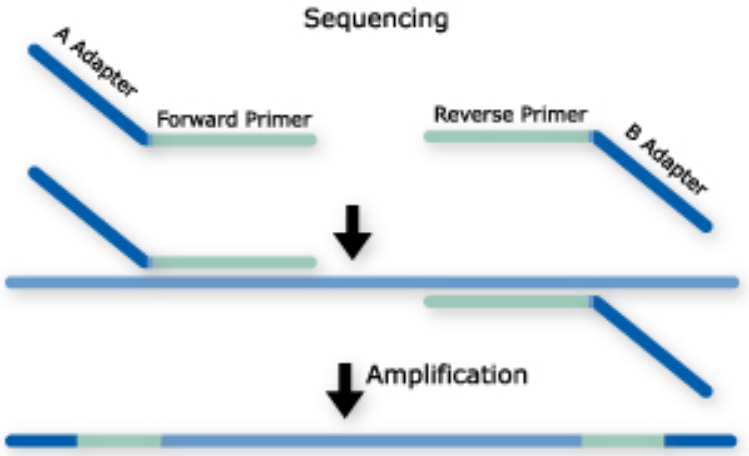
Sekvenační strategie

AMPLIKONOVÉ SEKVENOVÁNÍ

PCR Amplifikace konkrétního úseku daného genomu pomocí specifických primerů (se sekvenačními adaptory)

Následná sekvenace

Taxonomické složení daného vzorku („metabarcoding“), variabilita konkrétních genů apod.



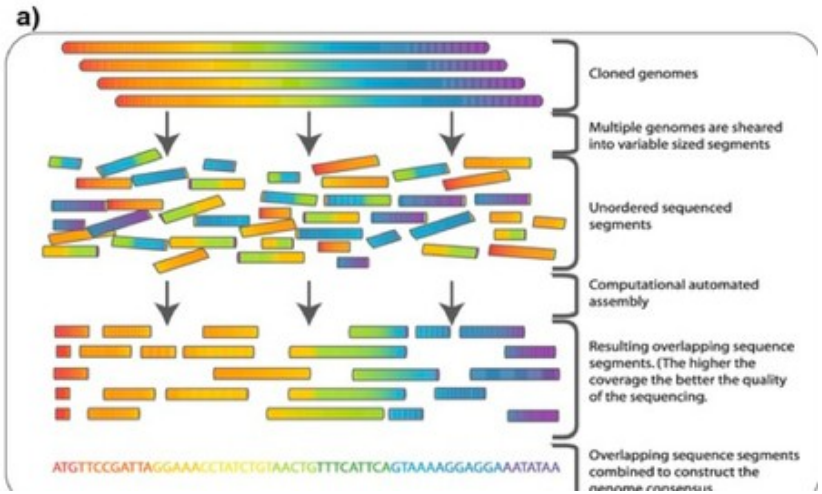
SHOT GUN SEKVENOVÁNÍ

Fragmetace celogenomové DNA

Ligace sekvenačních adaptorů

Následná sekvenace náhodných fragmentů

De novo assembly, resekvenování, transkriptomika, funkční složení daného společenstva



TO NENÍ VŠECHNO.....

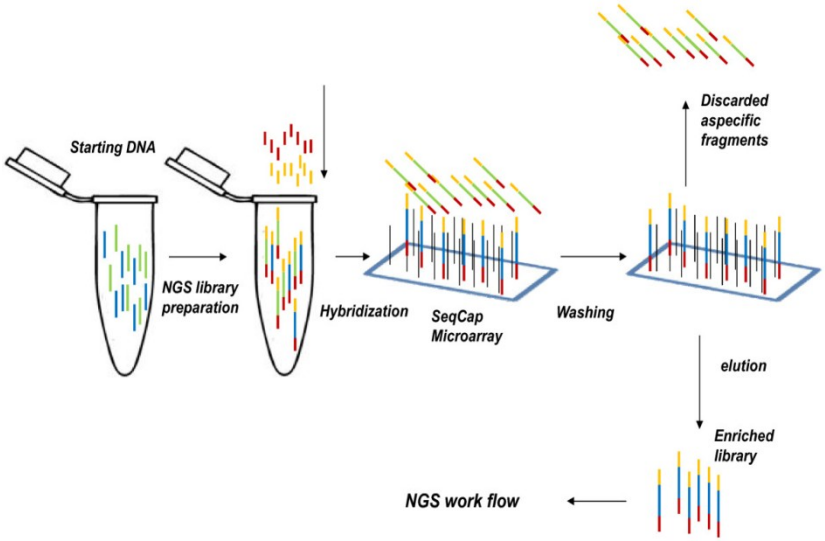
Sekvenační strategie

Sequence capture + shot gun

Separace úseků genomu které nás zajímají na základě jejich hybridizace

Následná sekvenace obohacených knihoven („enrichement“)

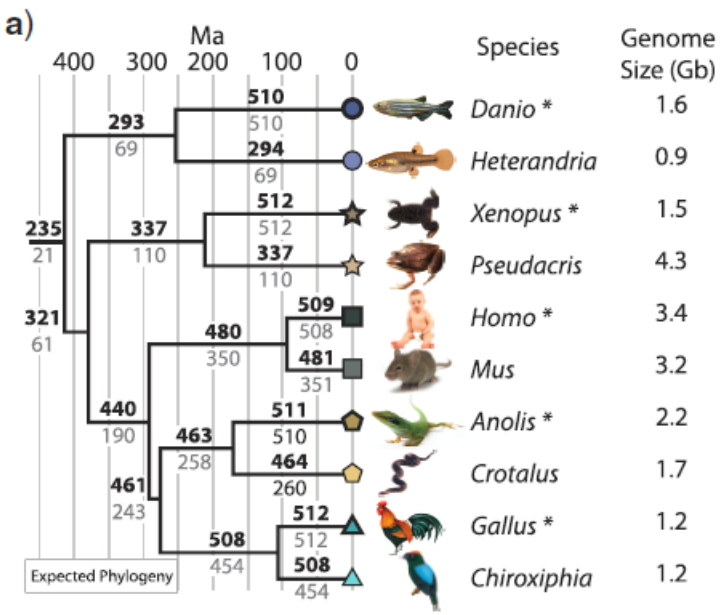
Nové markery (mikrosatelity apod.), kódující oblasti genomu („exom“), „anchored phylogenomics“ apod.

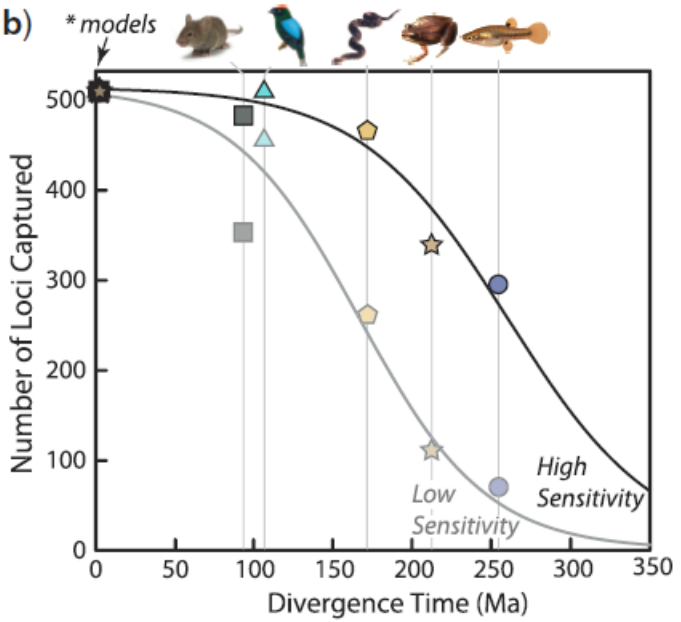


NimbleGen

Anchored phylogenomics

- hundreds of conserved loci
- hybridization enrichement
- u velmi příbuzných taxonů bude málo variability





CENTER FOR ANCHORED PHYLOGENOMICS

ACCELERATING THE RESOLUTION OF LIFE™

Juana, E. Handbook of the Birds of the World Alive (Lynx Edicions, 2015).

LETTER

doi:10.1038/nature15697

A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing

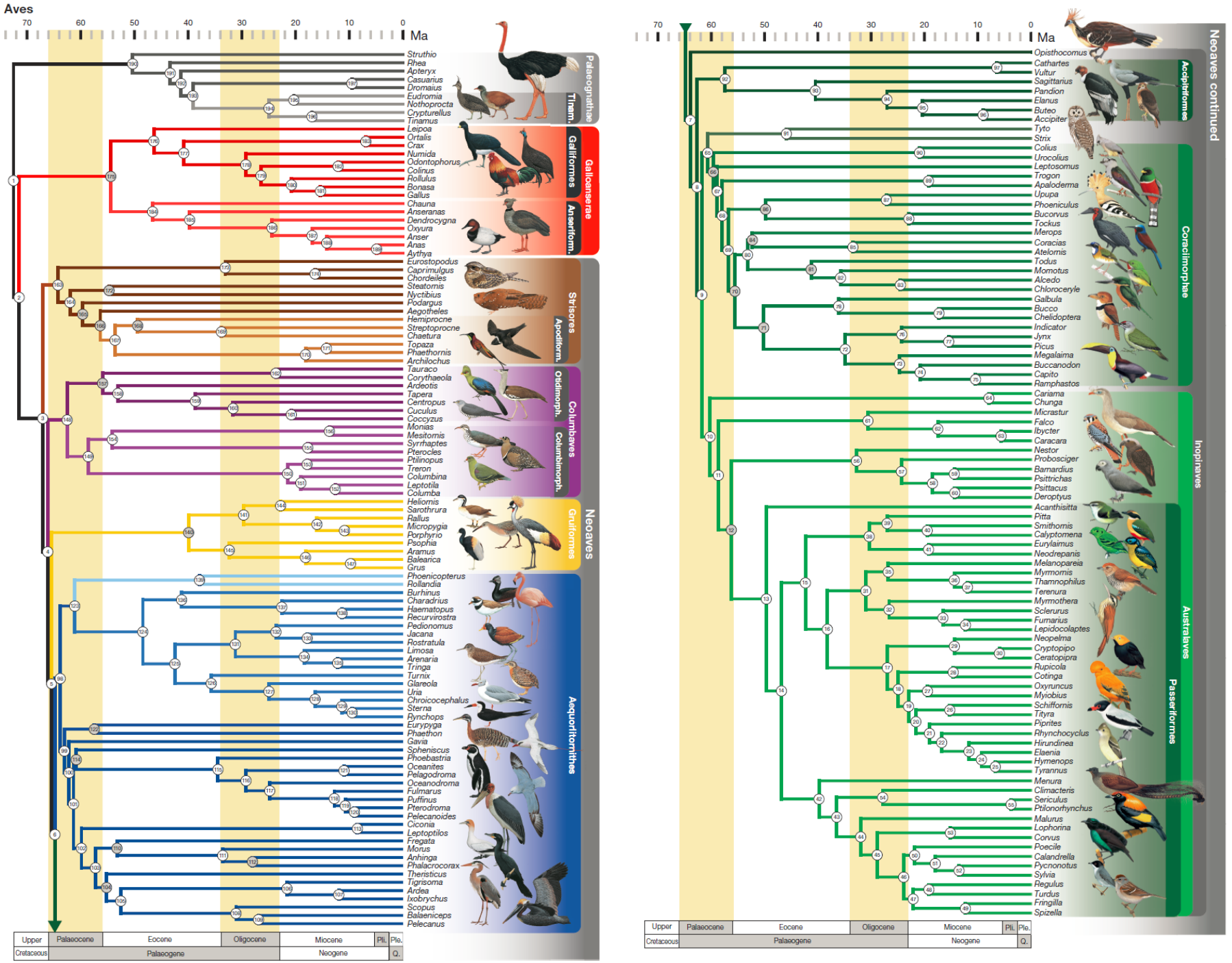
Richard O. Prum[1,2]*, Jacob S. Berv[3]*, Alex Dornburg[1,2,4], Daniel J. Field[2,5], Jeffrey P. Townsend[1,6], Emily Moriarty Lemmon[7] & Alan R. Lemmon[8]

Nature Paper Resolves Bird Tree of Life

Posted on October 6, 2015 by ameer

October 2015

198 species
259 nuclear loci (ca 1500 bp each)
> 390 000 bp


Aves
70
60
50
40
30
20
10
0
Ma
Struthio
Rhea
Apteryx
Casuarius
Dromaius
Eudromia
Nothoprocta
Crypturellus
Tinamus
Palaeognathae
Tinam.
Leipoa
Ortalis
Crax
Numida
Odontophorus
Colinus
Rollulus
Bonasa
Gallus
Galliformes
Galloanserae
Chauna
Anseranas
Dendrocygna
Oxyura
Anser
Anas
Aythya
Anseriform.
Eurostopodus
Caprimulgus
Chordeiles
Steatornis
Nyctibius
Podargus
Aegotheles
Hemiprocne
Streptoprocne
Chaetura
Topaza
Phaethornis
Archilochus
Apodiform.
Strisores
Tauraco
Coruthaeola
Ardeotis
Tapera
Centropus
Cuculus
Coccyzus
Otidimorph.
Monias
Mesitornis
Syrrhaptes
Pterocles
Ptilinopus
Treron
Columbina
Leptotila
Columba
Columbimorph.
Columbaves
Heliornis
Sarothrura
Rallus
Micropygia
Porphyrio
Psophia
Aramus
Balearica
Grus
Gruiformes
Neoaves
Phoenicopterus
Rollandia
Burhinus
Charadrius
Haematopus
Recurvirostra
Pedionomus
Jacana
Rostratula
Limosa
Arenaria
Tringa
Turnix
Glareola
Uria
Chroicocephalus
Sterna
Rynchops
Eurypyga
Phaethon
Gavia
Spheniscus
Phoebastria
Oceanites
Pelagodroma
Oceanodroma
Fulmarus
Puffinus
Pterodroma
Pelecanoides
Ciconia
Leptoptilos
Fregata
Morus
Anhinga
Phalacrocorax
Theristicus
Tigrisoma
Ardea
Ixobrychus
Scopus
Balaeniceps
Pelecanus
Aequorlitornithes
Upper Cretaceous
Palaeocene
Eocene
Oligocene
Miocene
Plio.
Ple.
Palaeogene
Neogene
Q.
Neoaves continued
Opisthocomus
Cathartes
Vultur
Sagittarius
Pandion
Elanus
Buteo
Accipiter
Accipitriformes
Tyto
Strix
Colius
Urocolius
Leptosomus
Trogon
Apaloderma
Upupa
Phoeniculus
Bucorvus
Tockus
Merops
Coracias
Atelornis
Todus
Momotus
Alcedo
Chloroceryle
Galbula
Bucco
Chelidoptera
Indicator
Jynx
Picus
Megalaima
Buccanodon
Capito
Ramphastos
Coraciimorphae
Cariama
Chunga
Micrastur
Falco
Ibycter
Caracara
Nestor
Probosciger
Barnardius
Psittrichas
Psittacus
Deroptyus
Inopinaves
Acanthisitta
Pitta
Smithornis
Calyptomena
Eurylaimus
Neodrepanis
Melanopareia
Myrmornis
Thamnophilus
Terenura
Myrmothera
Sclerurus
Furnarius
Lepidocolaptes
Neopelma
Cryptopipo
Ceratopipra
Rupicola
Cotinga
Oxyruncus
Myiobius
Schiffornis
Tityra
Piprites
Rhynchocyclus
Hirundinea
Elaenia
Hymenops
Tyrannus
Menura
Climacteris
Sericulus
Ptilonorhynchus
Malurus
Lophorina
Corvus
Poecile
Calandrella
Pycnonotus
Sylvia
Regulus
Turdus
Fringilla
Spizella
Australaves
Passeriformes

Sekvenační strategie

Long range PCR + shot gun

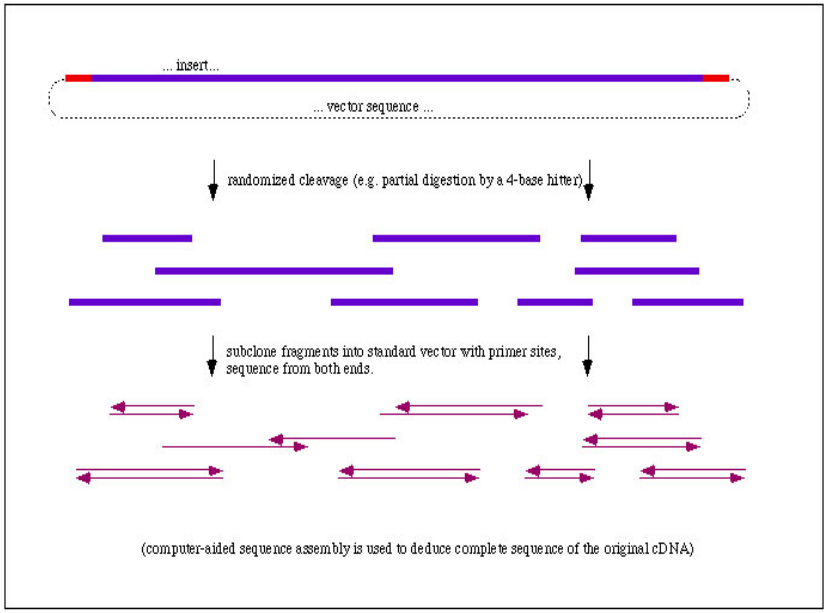
Dlouhé PCR produkty, které nejdou vcelku osekvenovat

Jejich fragmentace

Sekvenování fragmetů

Zpětná rekonstrukce původní sekvence („assembly“)

Použitelné pokud nás zajímá variablita v jednolitém úseku DNA. Např. sekvenace mitochodrální DNA (3 různé PCR produkty).



Sekvenační strategie

Sekvenování podél restrikčních míst

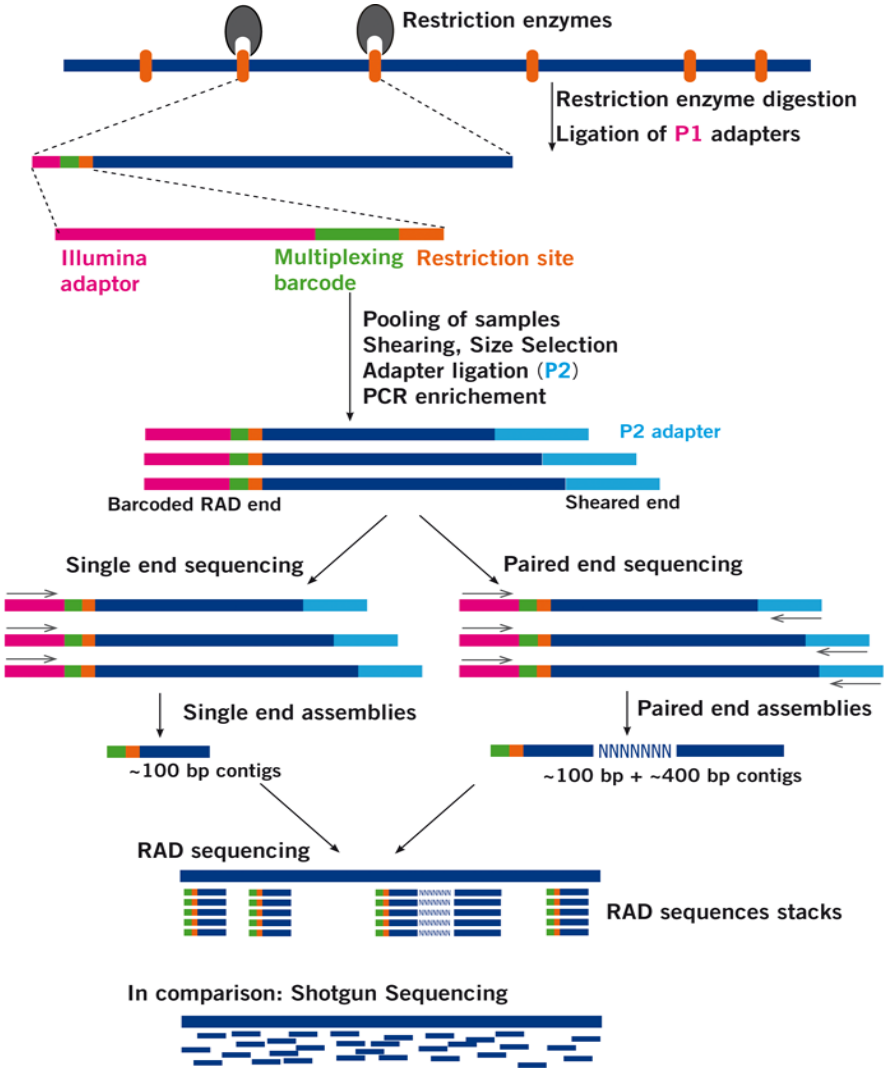
Fragmetace gelogenomové DNA po mocí restrikčních enzymů

Ligace sekvenačních adaptorů na výsledné fragmenty

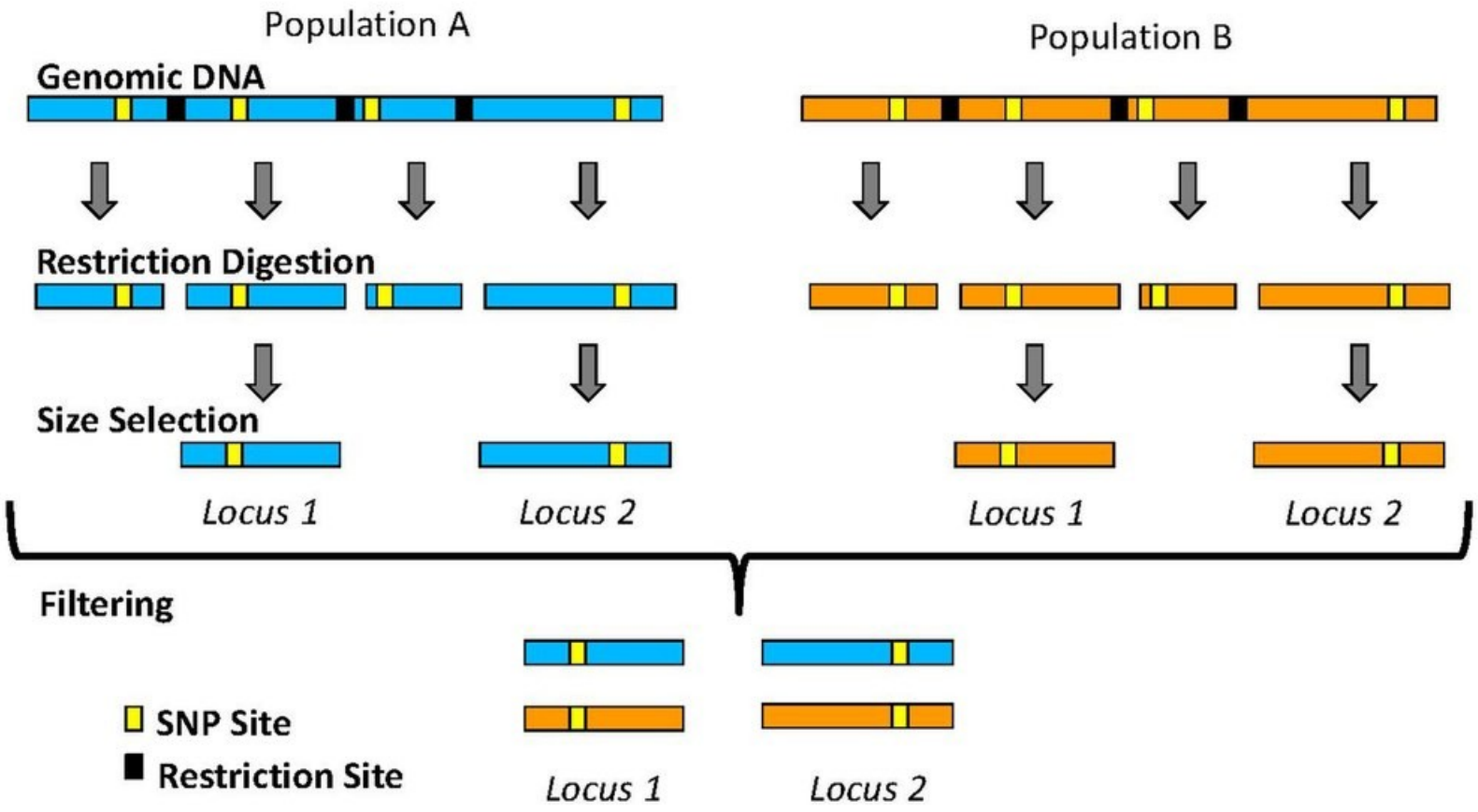
Následná sekvenace podél restrikčních míst

Celogenomové scany genetické variablility

Hledání SNPs, populační genomika (např. RAD-SEQ) apod.



Restriction-site Associate DNA Sequencing (RADseq)



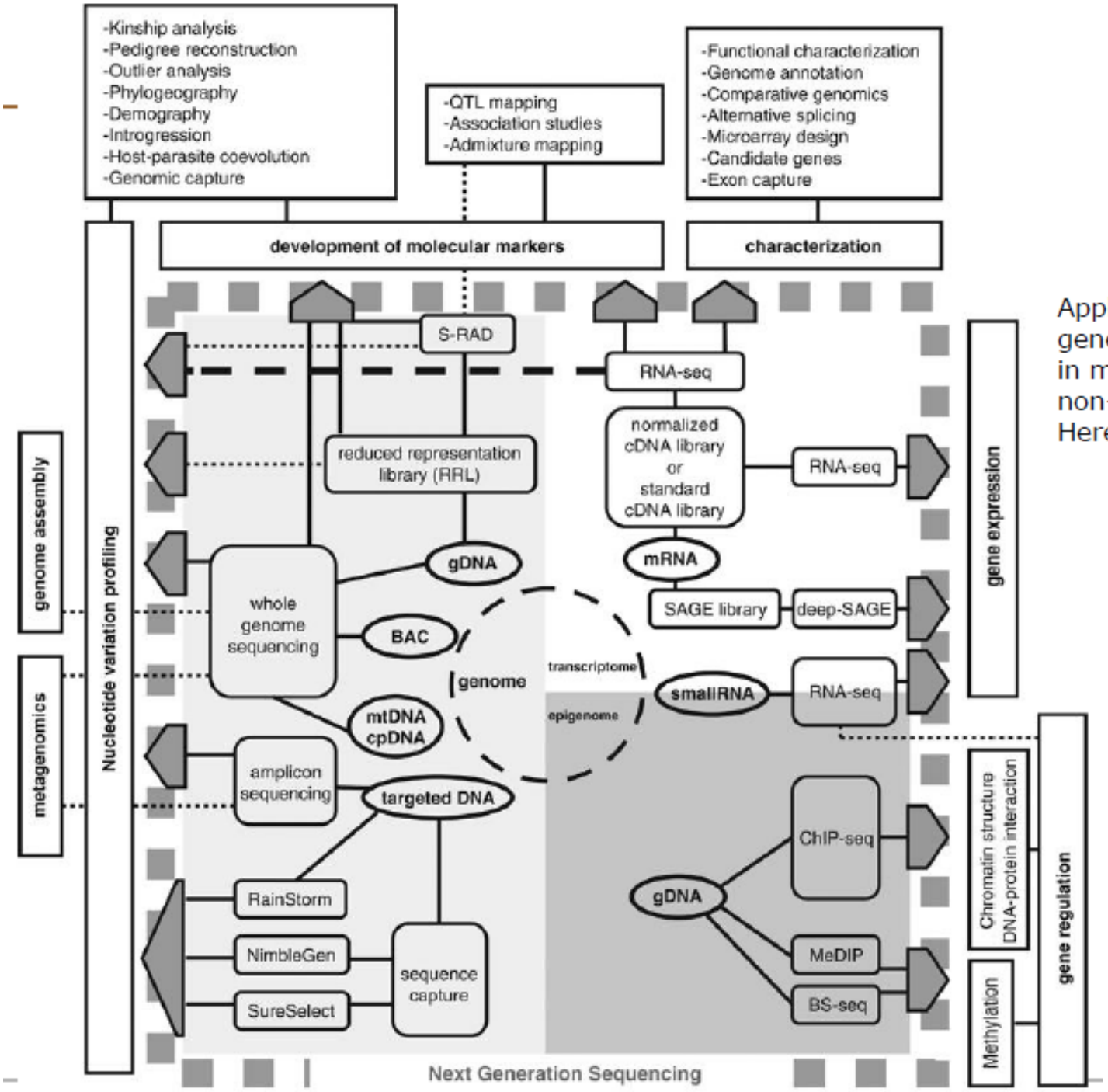
Aplikace

1. Celogenomové sekvenování de novo

2. Celogenomové resekvenování

3. Sekvenování amplikonů (PCR produktů)

4. Další aplikace – např. hledání klasických DNA markerů (mikrosatelity, SNPs)



Applications of next generation sequencing in molecular ecology of non-model organisms, Heredity, 2011

1. Celogenomové sekvenování de novo

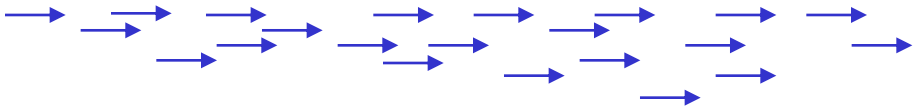
Problém: **KRÁTKÝ READ LENGTH**

- **400bp** 454 FLX Roche (dnes i Illumina), **35-75bp** Solid
  vs **800-1000bp** Sanger
- nové technologie (PacBio, Nannopore) už s tím takový problém nemají

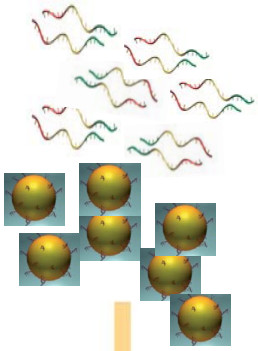
→ Uspořádání (assembly) ještě stále může být problém z hlediska výpočetní kapacity

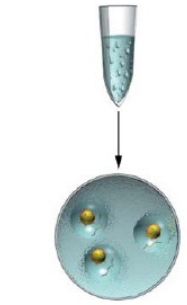
!!!!! REPETITIVNÍ OBLASTI delší než read length !!!!!

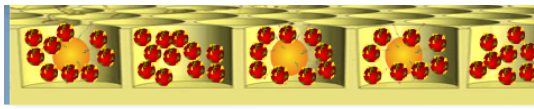
GTAAAAAAAAAAAAAAAAAAAAC

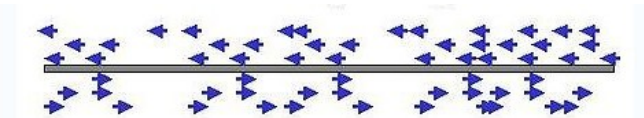


Zvláště komplexní eukaryotické genomy – úseky souvislých oblastí přerušených mezerami









1. Celogenomové sekvenování de novo

- získání kompletní uspořádané sekvence celých velkých eukaryotních genomů pomocí next-generation sequencing de novo je problém (ale to je nakonec i u Sangera)
- viry, prokaryota, malá eukaryota, mitochondrie/plastidy/plasmidy

Genetic Det … s, a
New Hemor … n
Southern Af …

Thomas Briese[1,5]*, Jan … eet[1],
Gustavo Palacios[1], Ma … lm[4],
Stuart T. Nichol[3], W. I …

1 Center for Infection and Immunity … thogens Unit,
National Institute for Communicable … of Viral and
Rickettsial Diseases, Centers for Dise … ited States of
America, 5 Biotechnology Core Faci …

Abstract

Lujo virus (LUJV), a new … m the
Old World discovered in … zed by
nosocomial transmissior … of RNA
extracts from serum ar … zation
within 72 hours of sam … cestral
node of the Old World … t from
that of other Old Worl … V is a
novel, genetically distinct, highly pathogenic arenavirus.

2009

2015

2. Celogenomové resekvenování

- podobné problémy jako u de novo, ale méně (větší strukturální přestavby..)

KOMPARATIVNÍ GENOMIKA
- viry, prokaryota, malá eukaryota
- mitochondrie/plastidy/plasmidy

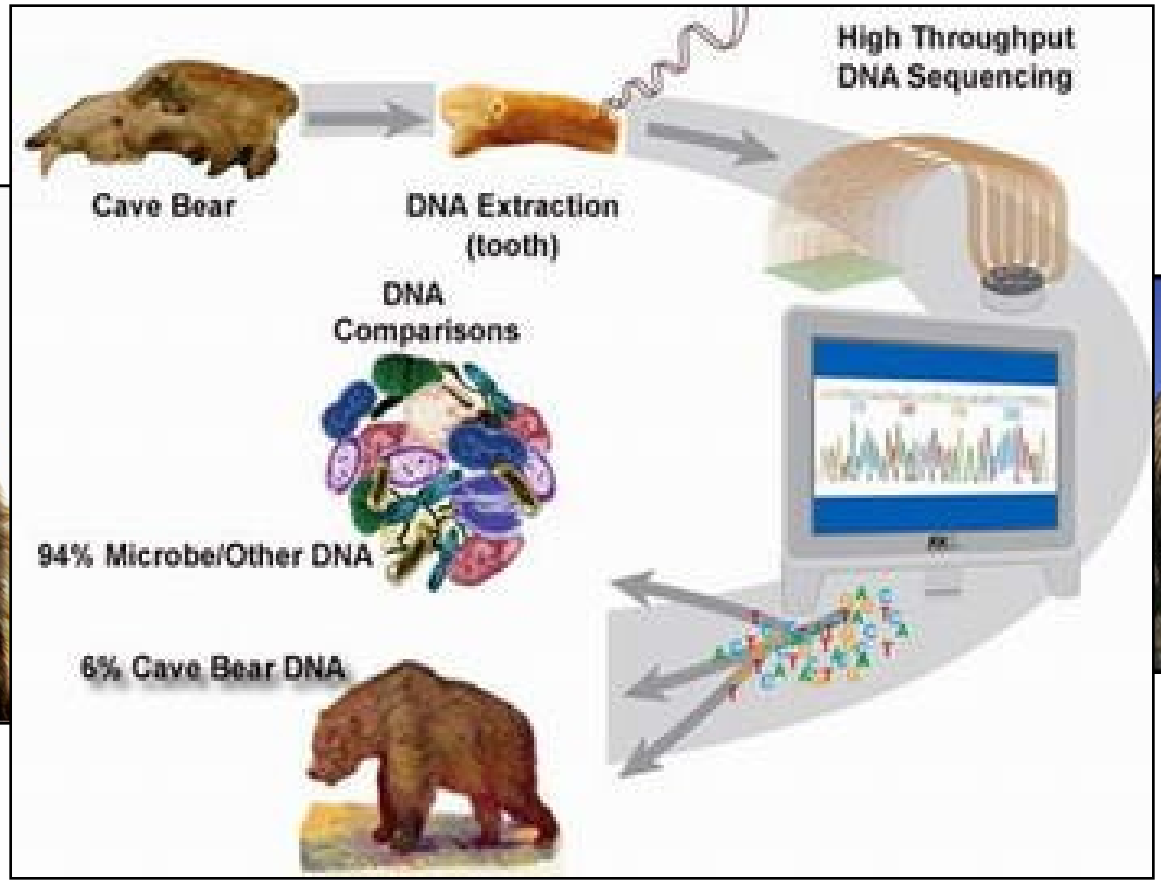
ANCIENT (mt) DNA
- různé směsné, degradované vzorky, např. fosilie

A Complete Neandertal Mitochondrial Genome Sequence Determined by High-Throughput Sequencing

Richard E. Green,[1,*] Anna-Sapfo Malaspinas,[2] Johannes Krause,[1] Adrian W. Briggs,[1] Philip L.F. Johnson,[3] Caroline Uhler,[4] Matthias Meyer,[1] Jeffrey M. Good,[1] Tomislav Maricic,[1] Udo Stenzel,[1] Kay Prüfer,[1] Michael Siebauer,[1] Hernán A. Burbano,[1] Michael Ronan,[5] Jonathan M. Rothberg,[6] Michael Egholm,[5] Pavao Rudan,[7] Dejana Brajković,[8] Željko Kučan,[7] Ivan Gušić,[7] Mårten Wikström,[9] Liisa Laakkonen,[10] Janet Kelso,[1] Montgomery Slatkin,[2] and Svante Pääbo[1]

Ancient Genomes Resurrected

- Degraded state of the sample → mitDNA sequencing
- Nuclear genomes of ancient remains: cave bear, mommoth, Neanderthal (10^6 bp)



Problems: contamination modern humans and coisolation bacterial DNA

3.Sekvenování amplikonů (PCR produktů)

SMĚSNÉ VZORKY – paralelní sekvenování nahrazuje klonování

Metagenomika (= hlavně prokaryota)

- Celé společenstvo půdních, vodních mikroorganismů, střevní mikroflóra - **mikrobiom**
- PCR genu 16S rRNA
- lze i kvantifikovat

Metabarcoding (= hlavně eukaryota, ale dnes používáno jako obecný termín)

- COI gen, příp. jiný barcodingový marker
- složení potravy, monitoring společenstev

Metabarcoding: Taxonomické složení společenstva v environmentální DNA na základě taxonomicky informativního úseku DNA (cyt b, COI, ITS, rRNA...)

Princip

- Směsný vzorek enviromentální DNA
- Amplifikace pomoci primerů specifických pro cílovou skupinu, pokrývající taxonomicky informativní úsek (COI, 16s/18s RNA...)
- Paralelni sekvenování
- Filtrování nekvalitních sekvencí
- Klastrování na základě sekvenční podobnosti do OTUs („operational taxonomic units")
- Jejich taxonomické zařazení na základě referenčních databází

Využití: Analýza druhového vzorků kde lze makroskopicky jednotlivé druhy obtížně odlišit

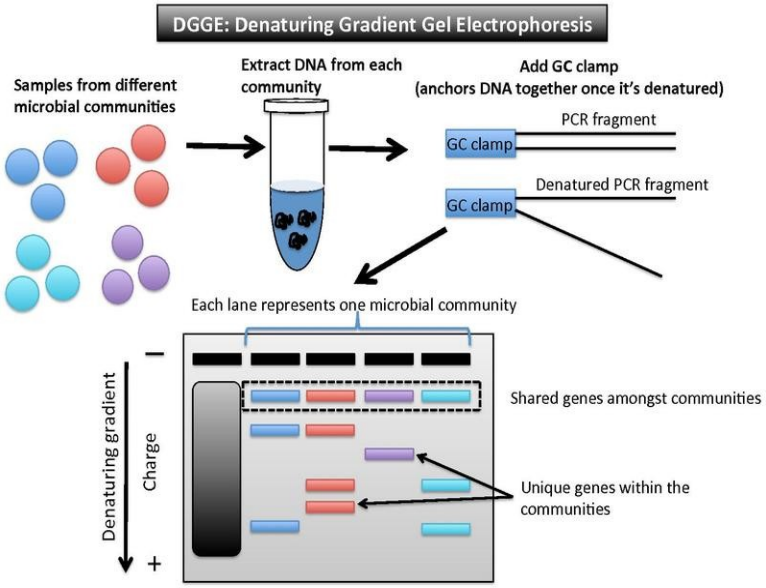
- Potravní analýza z trusu
- Vzorky půdy
- Mikrobiální společenstva
- Permafrost
- Exotická/špatně probádaná společenstva
- Druhově bohatá společenstva („insect traps" v tropech)
- Rutinní analýza velkého množství vzorků

Metabarcoding

Taxonomické složení společenstva na základě taxonomicky informativního úseku DNA

Alternativy:

Klonování amplikonů a sekvenování klonů
Specifické elektroforézy – např. DGGE



Výhody paralelního sekvenování

- Cenově i časově míň nákladné
- Lépe se zachytí vzácné taxony (zlomky promile)

Ale:

- Riziko umělého navýšení diversity díky chybám při procesování dat
- Do jaké míry jsou referenční databáze dostatečné ke klasifikaci vzorků?
- Lze použít tato data kvantitativně a nebo vypovídají jen o přítomnosti/nepřítomnosti?

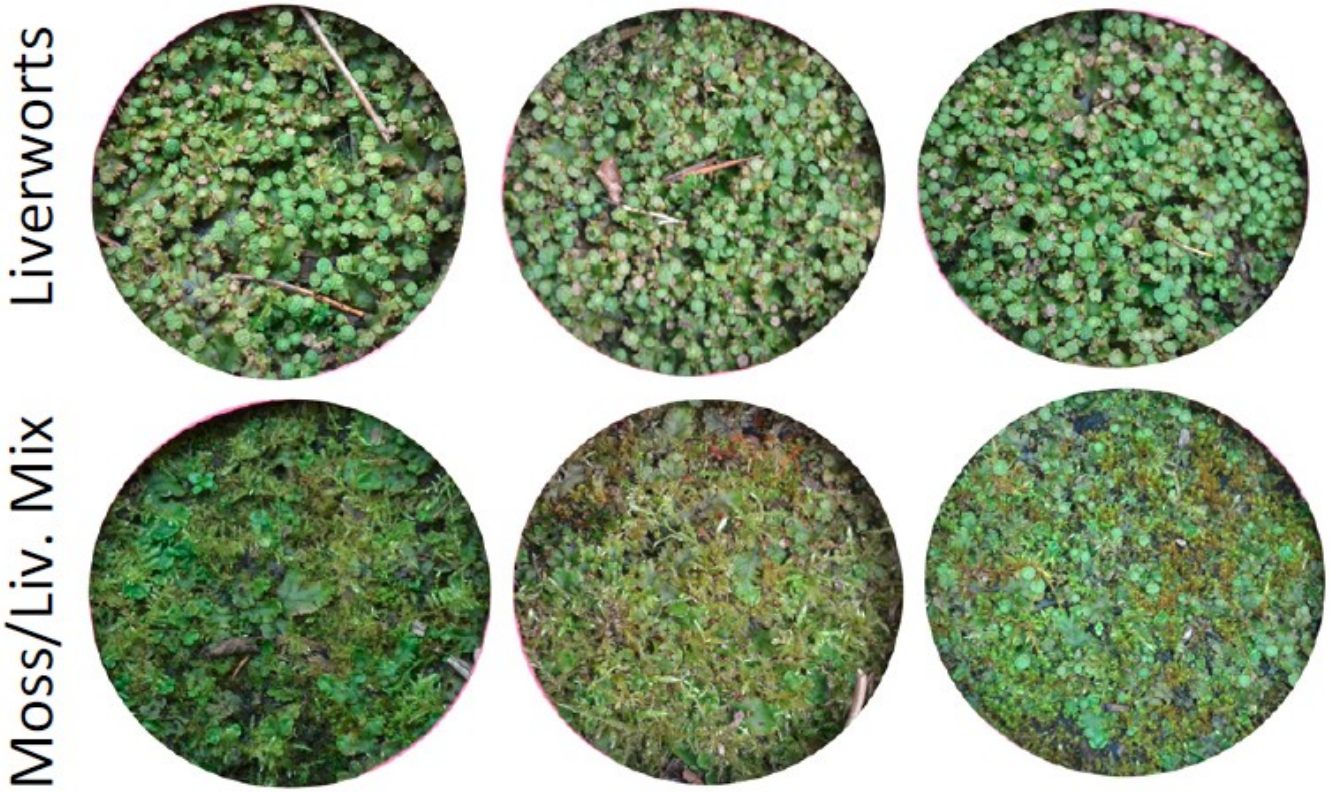
Metabarcoding – příklady využití

- Liverwort only vs. Mixed moss/liverwort (50:50)
- Collected at fixed distance from each other
- 3 replicates each

Společenstvo eukaryot ve vrchní vrstvě půdy

Mark Blaxter, Edinburgh Genomics

Metabarcoding – příklady využití



Společenstvo eukaryot ve vrchní vrstvě půdy

Mark Blaxter, Edinburgh Genomics

Metabarcoding – příklady využití

low diversity PCR products → RTA 1.17.28 → high quality data

Eukaryotic nSSU barcoding

6 samples

3 replicates each of two ecosystems

1200 clusters/mm^2

2% phiX174 spike-in

17 million raw pass filter pairs

Mark Blaxter, Edinburgh Genomics

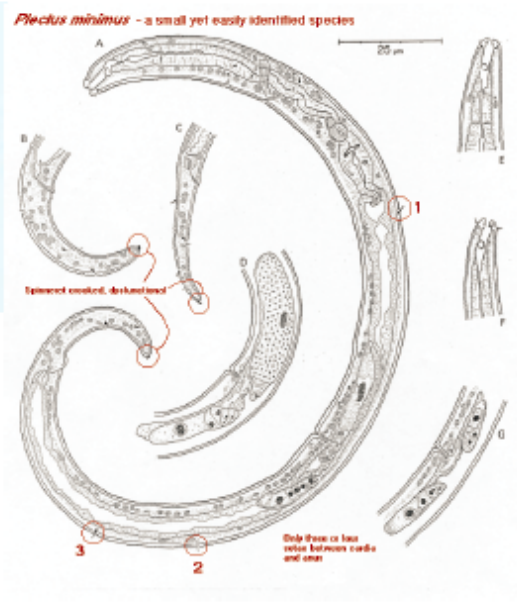
Metabarcoding – příklady využití

Phylum	98% MOTU	proportion of total reads
Nematoda	2862	.3275761886787317
Dikarya	6965	.1894664458485315
Cercozoa	4254	.085598403024025
Annelida	682	.0688691096605889
null	*4867*	*.0579833558234776*
Streptophyta	614	.0579039901203119
Oomycetes	487	.0569565860018453
Bacillariophyta	666	.0250973907279004
Arthropoda	286	.0218196255280743
Fungi_incertae_sedis	417	.0195828162900598
Tardigrada	158	.0169930788889338
Chytridiomycota	473	.0138886146448126
Ciliophora	544	.009428990604366
Chlorophyta	473	.0080418161403385
Synurophyceae	34	.005969106037372
Centramoebida	288	.0053245951821951
Platyhelminthes	94	.0051997954895359
Chrysophyceae	198	.005026302829234
Nucleariidae	61	.0048583696022456
Tubulinea	194	.0025431531840504
Blastocladiomycota	76	.0021085588626O9
Apicomplexa	74	.0016743479503615
Flabellinea	53	.0013524759319671
Dinophyceae	139	.000952196733392
Bicosoecida	46	.0007160166698649
Uncultured_banisveld_eukaryote	17	.0005992685702805
Micronuclearia_podoventralis	20	.0005377313946375
Codonosigidae	30	.0005166438889655
Ichthyophonida	28	.0004928725189351
Px_clade	60	.0004213667042472
Fungal_endophyte_sp._sx01	4	.0003439180470516
Heterophryidae	27	.0002599514335574
Hyphochytriomycetes	9	.0002386722232883
Salpingoecidae	13	.000235604949736
Fungal_sp._gmg_c6	37	.0002262114244821
Eustigmatophyceae	14	.0001882539142724
Capsaspora	10	.0001878705050784
Stramenopile_sp._mast-12_kkts_d3	15	.000173684364899
Raphidophyceae	31	.0000799408169568
Schizoplasmida	16	.0000793657031658
Environmental_samples	8	.0000573196745086
Trimastix_pyriformis_atcc50562	5	.0000548275147474
Labyrinthulida	16	.0000521436503891
Gastrotricha	1	.0000415990975531
Telonema	1	.0000391077377918
Mollusca	26	.0000368072826276
Apusomonadidae	3	.0000325897814932
Soil_amoeba_and16	11	.0000312478493141
Acanthocystidae	4	.0000297142125379
Voromonas	4	.0000291390987468
Leukarachnion_sp._atcc_pra-24	1	.0000264552343886
Rotifera	10	.0000231962562393
Ancyromonadidae	3	.0000090101160599
Perkinsea	3	.0000046009103284
Mycetozoa	3	.0000044092057314
Eccrinaceae	1	.0000038340919404
Glomeromycota	2	.0000030672735523
Eukaryota_marine_clone_mel-24	1	.0000023004551642
Fungal_sp._3cas90	1	.0000021087505672
Cryptomonadales	1	.0000011502275821
Malawimonas	2	.0000011502275821
Phaeothamniophyceae	1	.0000009585323985
Chordata	1	.0000003834039194
Unclassified_alveolata	1	.000000191704597

Eight animal phyla represented

Most frequent are **Nematoda**

Most frequent "98% MOTU" is *Plectus (cf aquatilis)*



Mark Blaxter, Edinburgh Genomics

Metabarcoding – příklady využití

Monitoring vzácných, nedávno popsaných druhů savců na základě sekvenování krve pijavic

Výrazně větši úspěšnost prokázání přítomnosti než za použití klasických technik – fotopasti, terénní pozorování apod.

Correspondences

Screening mammal biodiversity using DNA from leeches

Ida Bærholm Schnell[1,2,†],
Philip Francis Thomsen[2,†],
Nicholas Wilkinson[3],
Morten Rasmussen[2],
Lars R.D. Jensen[1], Eske Willerslev[2]
Mads F. Bertelsen[1],
and M. Thomas P. Gilbert[2,*]

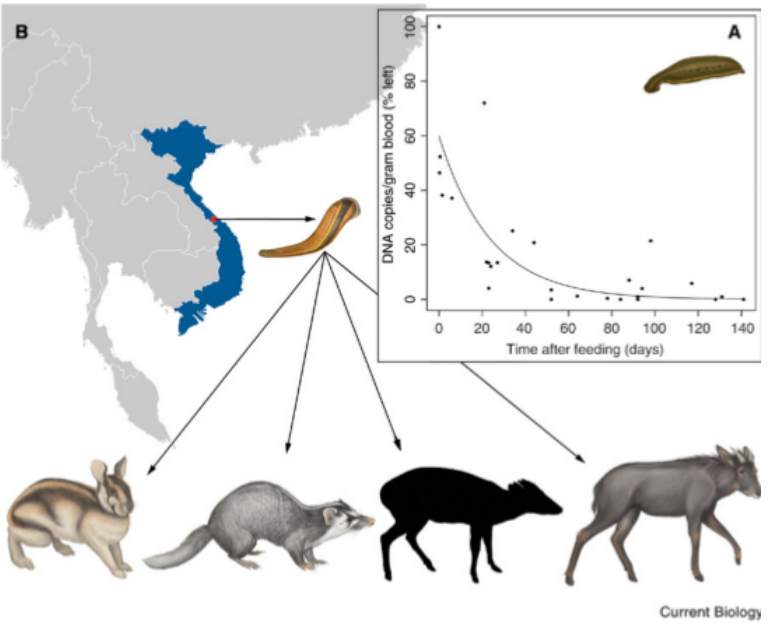
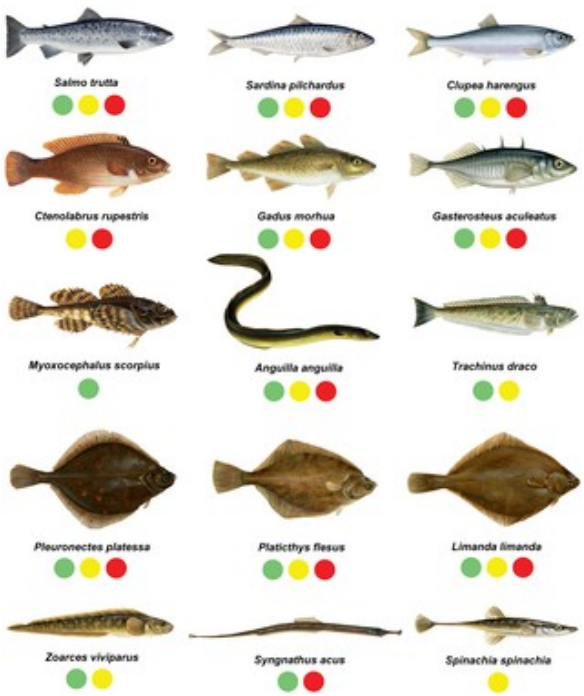
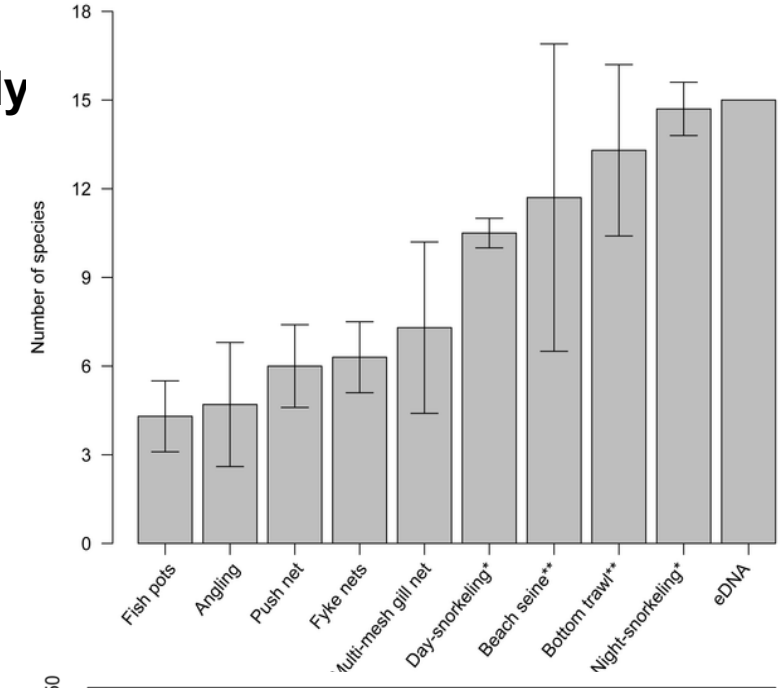


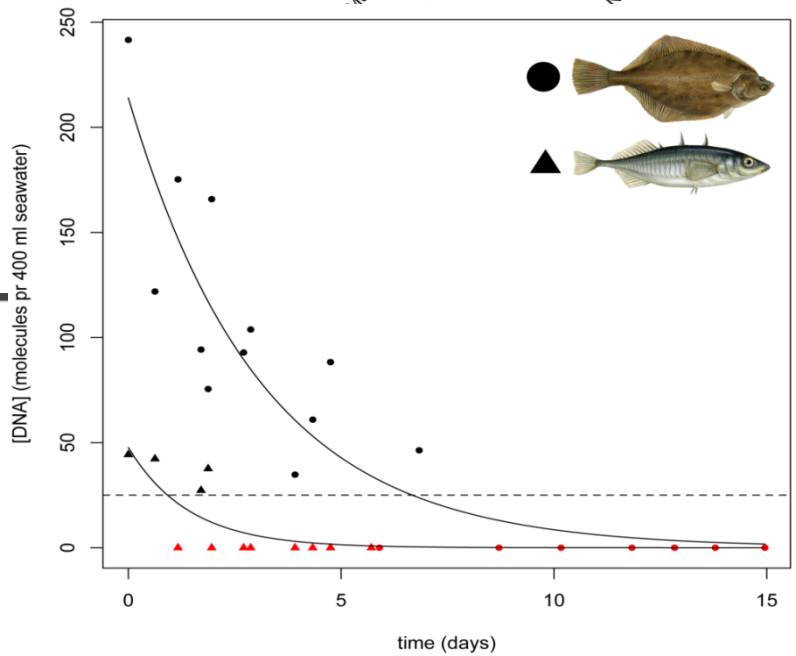
Figure 1. Monitoring mammals with leeches.
(A) Survival of mtDNA in goat blood ingested by *Hirudo medicinalis* over time, relative to freshly drawn sample (100%, ca. 2.4E+09 mtDNA copies/gram blood). Mitochondrial DNA remained detectable in all fed leeches, with a minimum observed level at 1.6E+04 mtDNA/gram blood ingested. The line shows a simple exponential decay model, $p < 0.001$, $R^2 = 0.43$ (Supplemental information). (B) Vietnamese field site location and examples of mammals identified in *Haemadipsa* spp. leeches. From left to right: Annamite striped rabbit, small-toothed ferret-badger, Truong Son munjtac (coat coloration and markings remain unknown), serow. Pictures do not reflect true size proportions. See also Supplemental information.

Metabarcoding – příklady využití

Detekce ryb pomocí izolace eDNA z mořské vody
-taky jedna z nejefektivnějších metod







OPEN ACCESS Freely available online

PLOS ONE

Detection of a Diverse Marine Fish Fauna Using Environmental DNA from Seawater Samples

Philip Francis Thomsen[1]*, Jos Kielgast[1], Lars Lønsmann Iversen[2], Peter Rask Møller[3], Morten Rasmussen[1], Eske Willerslev[1]*

1 Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade, Copenhagen, Denmark, **2** Freshwater Biology Section, Department of Biology, University of Copenhagen, Helsingørgade, Hillerød, Denmark, **3** Vertebrate Department, Natural History Museum of Denmark, University of Copenhagen, Universitetsparken, Copenhagen, Denmark

Metabarcoding – příklady využití

Analýza potravy

Podíl hospodářských zvířat v potravě irbise je minimální

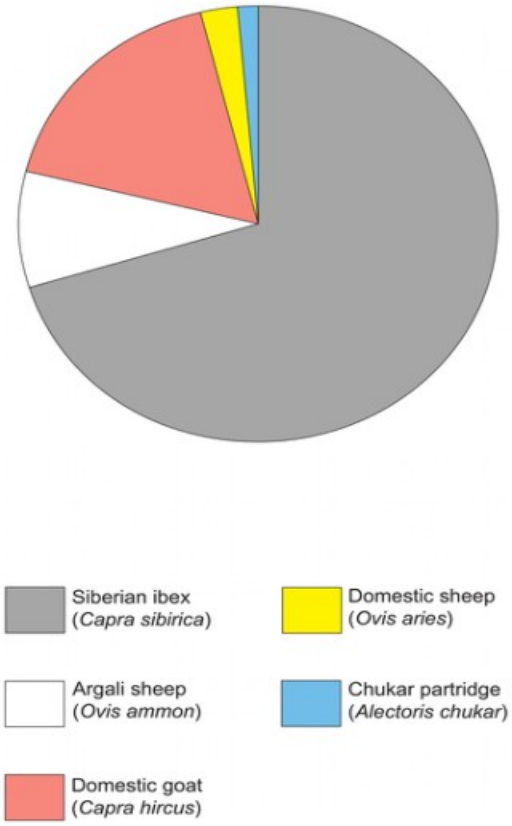
OPEN ACCESS Freely available online

PLoS one

Prey Preference of Snow Leopard (*Panthera uncia*) in South Gobi, Mongolia

Wasim Shehzad[1], Thomas Michael McCarthy[2], Francois Pompanon[1], Lkhagvajav Purevjav[3], Eric Coissac[1], Tiayyba Riaz[1], Pierre Taberlet[1]*

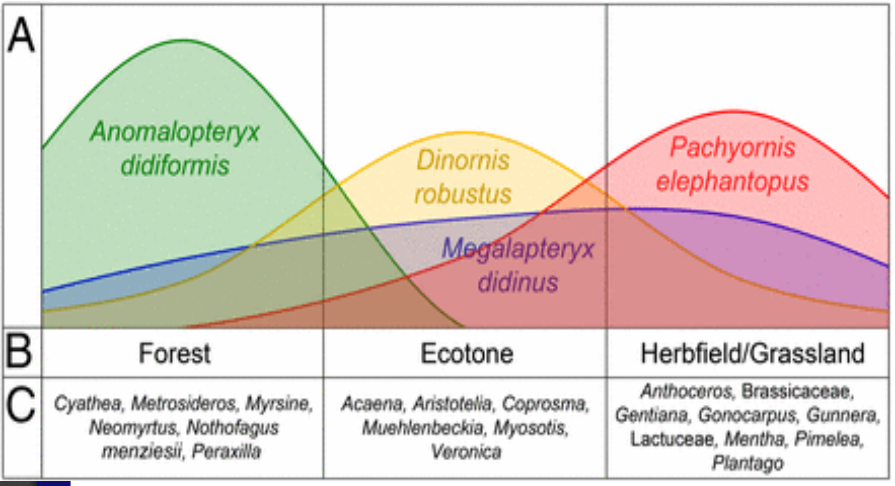
1 Laboratoire d'Ecologie Alpine, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 5553, Université Joseph Fourier, Grenoble, France, **2** Snow Leopard Program, Panthera, New York, New York, United States of America, **3** Snow Leopard Conservation Fund, Ulaanbaatar, Mongolia

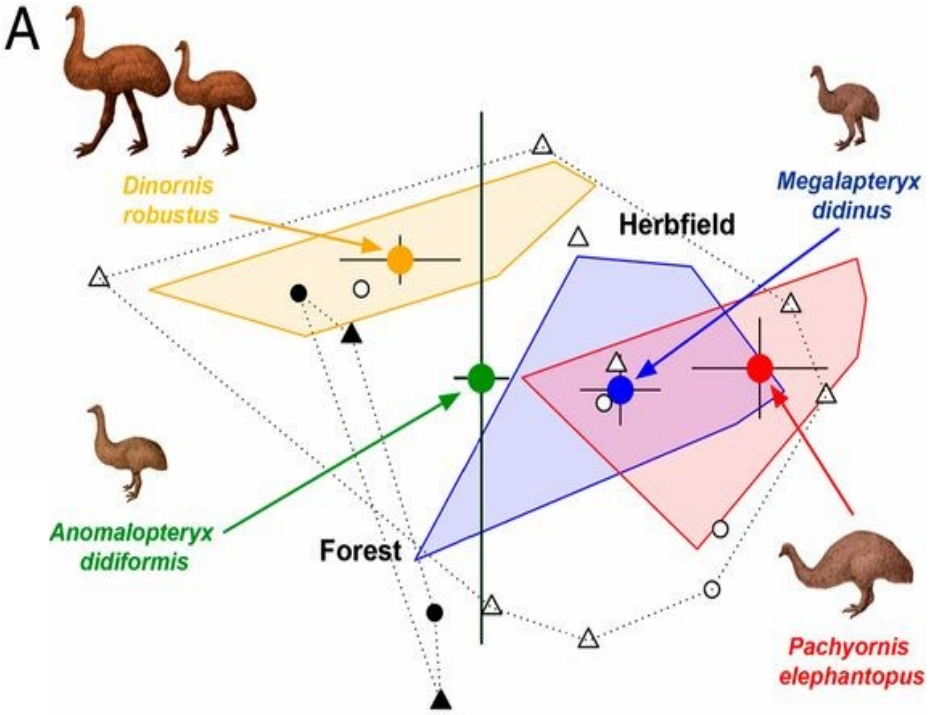


Metabarcoding – příklady využití

Analýza složení společenstva na základě ancient DNA z koprolitů moa (Nový Zéland)

Umožňuje odhadnout typ prostředí které jednotlivé druhy obývaly a separaci ekologických nik





NAS

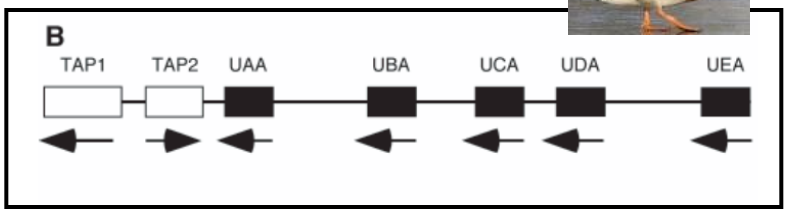
Resolving lost herbivore community structure using coprolites of four sympatric moa species (Aves: Dinornithiformes)

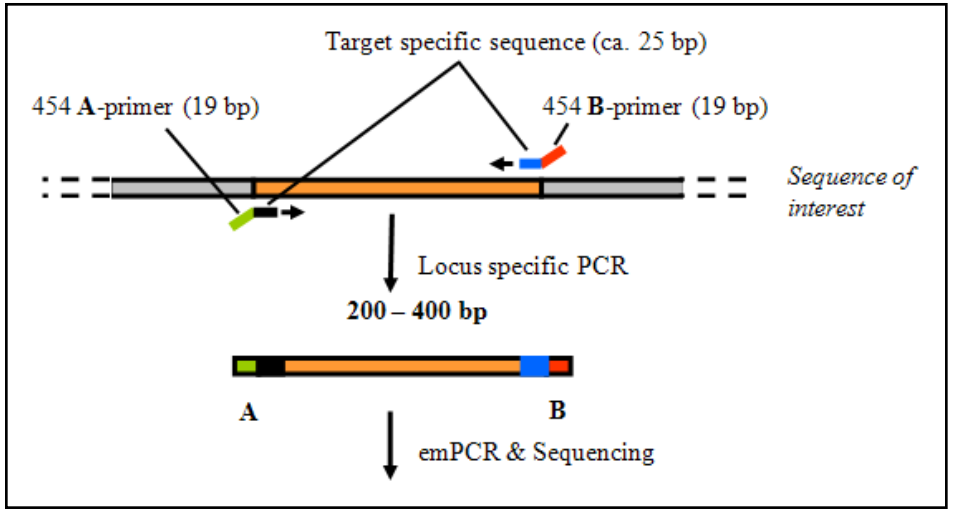
Jamie R. Wood[a,1], Janet M. Wilmshurst[a], Sarah J. Richardson[a], Nicolas J. Rawlence[b,2], Steven J. Wagstaff[a], Trevor H. Worthy[c,3], and Alan Cooper[b]

[a]Landcare Research, Lincoln, Canterbury 7640, New Zealand; [b]Australian Centre for Ancient DNA, University of Adelaide, Adelaide, SA 5005, Australia;

3.Sekvenování amplikonů (PCR produktů)

Genové duplikace



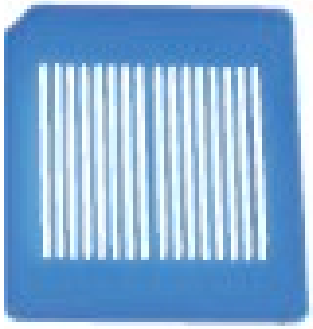


A-adaptor — MID — Target specific

Potřeba k emPCR, sekvenování..

Označí jedince

Amplifikuje všechny kopie MHC genů

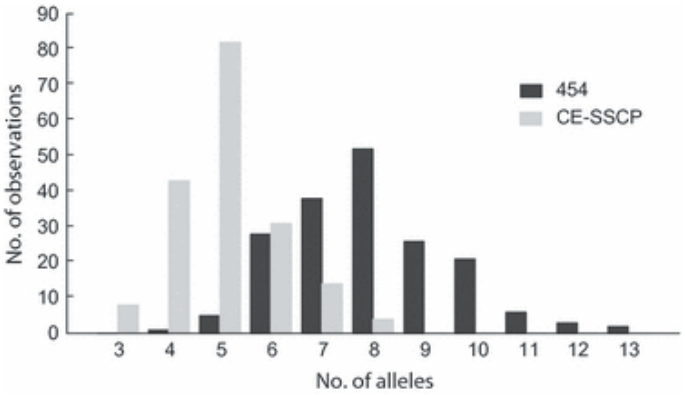


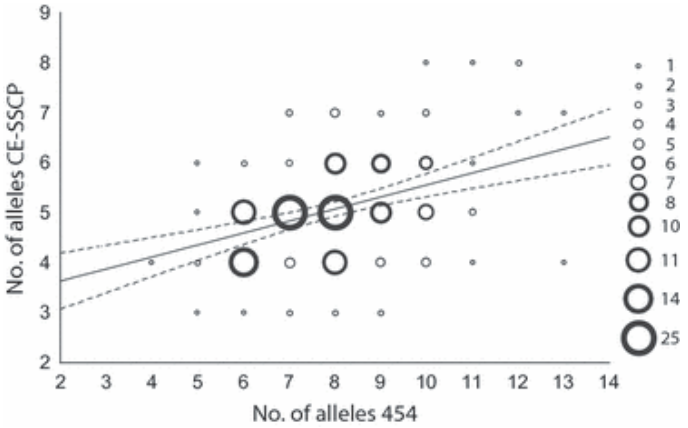
192 jedinců u 454 pyrosekvenování

Amplikonové sekvenování

MHC u hýla rudého

- NGS má větší rozlišovací schopnost než SSCP + klonování





MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2012) **12**, 285–292
doi: 10.1111/j.1755-0998.2011.03082.x

Evaluation of two approaches to genotyping major histocompatibility complex class I in a passerine—CE-SSCP and 454 pyrosequencing

MARTA PROMEROVÁ,* WIESŁAW BABIK,† JOSEF BRYJA,* TOMÁŠ ALBRECHT,*‡ MICHAŁ STUGLIK† and JACEK RADWAN§

4. Další aplikace – hledání nových genetických markerů

Mikrosatelity

- sekvenování obohacených knihoven

SNPs

- kompletní genomické sekvence pro hledání diagnostických SNPs
- např. RAD-sequencing

Hledání nových genetických markerů - mikrosatelity

Obvyklý postup:

-Obohacení genomické knihovy o mikrosatelitové motivy – sequence capture
-Sekvenování obohacených knihoven
-Detekce mikrosatelitů a navržení vhodných primerů

MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2011) 11, 638–644 doi: 10.1111/j.1755-0998.2011.029

High-throughput microsatellite isolation through 454 GS-FLX Titanium pyrosequencing of enriched DNA libraries

THIBAUT MALAUSA,* ANDRÉ GILLES,† EMESE MEGLÉCZ,† HÉLÈNE BLANQUART,‡ STÉPHANIE DUTHOY,‡ CAROLINE COSTEDOAT,† VINCENT DUBUT,† NICOLAS PECH,† PHILIPPE CASTAGNONE-SERENO,* CHRISTOPHE DÉLYE,§ NICOLAS FEAU,¶ PASCAL FREY,** PHILIPPE GAUTHIER,†† THOMAS GUILLEMAUD,* LAURENT HAZARD,‡‡ VALÉRIE LE CORRE,§ BRIGITTE LUNG-ESCARMANT,¶ PIERRE-JEAN G. MALÉ,§§ STÉPHANIE FERREIRA‡ and JEAN-FRANÇOIS MARTIN††

*INRA, UMR 1301 IBSV INRA/UNSA/CNRS, 400 Route des Chappes, BP 167, 06903 Sophia-Antipolis Cedex, France, †Aix-Marseille Université, CNRS, IRD, UMR 6116 – IMEP, Equipe Evolution Génome Environnement, Centre Saint-Charles, Case 3
3 Place Victor Hugo, 13331 Marseille Cedex 3, France, ‡Genoscreen, Genomic Platform and R&D, Campus de l'Institut Pasteur, rue du Professeur Calmette, Bâtiment Guérin, 59000 Lille, France, §INRA, UMR 1210 Biologie et Gestion des Adventices, 17 rue Sully, 21000 Dijon, France, ¶INRA, UMR 1202 BIOGECO, Equipe de Pathologie Forestière, Domaine de Pierroton, 69 route d'Arcachon, 33612 Cestas Cedex, France, **INRA, Nancy-Université, UMR 1136, Interactions Arbres – Microorganismes, IFR 1
54280 Champenoux, France, ††UMR CBGP (INRA/IRD/Cirad/Montpellier SupAgro), Campus International de Baillarguet, C
30016, 34988 Montferrier-sur-Lez Cedex, France, ‡‡INRA – UMR 1248 AGIR, BP 52627, 31326 Castanet-Tolosan Cedex, Franc
§§UMR Evolution et Diversité Biologique (Université Toulouse III; CNRS), 118 Route de Narbonne, 31062 Toulouse, France

allgenetics

HOME COMPANY SERVICES

HOME » SERVICES » Microsatellite Development

Experts in Microsatellite Development

Microsatellites (also known as short tandem repeats) are repetitive DNA elements usually found in non-coding regions of the genome. They have h
mutation rates, and therefore are frequently highly polymorphic. Variations in the number of repetitions generate different alleles. This makes th
appropriate molecular markers for population genetics and molecular ecology projects.

We develop microsatellite markers for your study species

At AllGenetics, we use next-generation sequencing to obtain primer pairs which amplify polymorphic microsatellite loci in your study species. Genor
DNA is used to generate genomic libraries. We usually enrich these libraries with 4 to 6 different microsatellite motifs. However, we can customise
number of motifs to your needs. We obtain thousands of microsatellite-containing reads by using high-throughput sequencing. Our bionformaticians
these reads for primer design. The primers obtained are multiplexed and tested for polymorphism in a number of individuals from different populations

How we work

High quality DNA at a concentration of 100 ng/µL in a minimum volume of 50 µL from a number of individuals is required. Alternatively, we can isol
DNA from your samples. These samples should be adequately preserved to ensure DNA integrity. We will deliver tested primer pairs which amp
polymorphic loci for your study species. A detailed methodological report and all sequencing reads generated will also be provided.

Our microsatellite development projects are divided into four steps. For your convenience, we can carry out the entire project or only the parts you nee

OPEN ACCESS Freely available online PLOS ONE

32 species validation of a new Illumina paired-end approach for the development of microsatellites

Stacey L. Lance[1*], Cara N. Love[1], Schyler O. Nunziata[1], Jason R. O'Bryhim[1], David E. Scott[1], R. Wesley Flynn[1], Kenneth L. Jones[2]

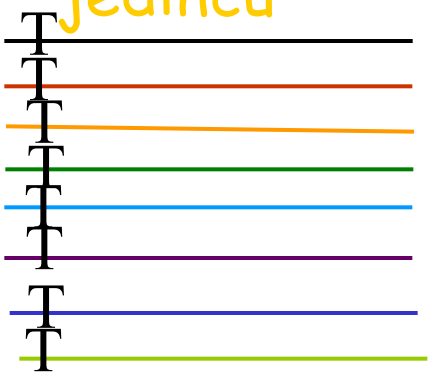
1 Savannah River Ecology Laboratory, University of Georgia, Aiken, South Carolina, United States of America, 2 Department of Biochemistry and Molecular

Hledání diagnostických SNP (např. pro studium hybridizace)

smíchat a osekvenovat

10 jedinců

10 jedinců



Hledání nových SNPs - RAD-sequencing

Sekvenování podél restrikčních míst

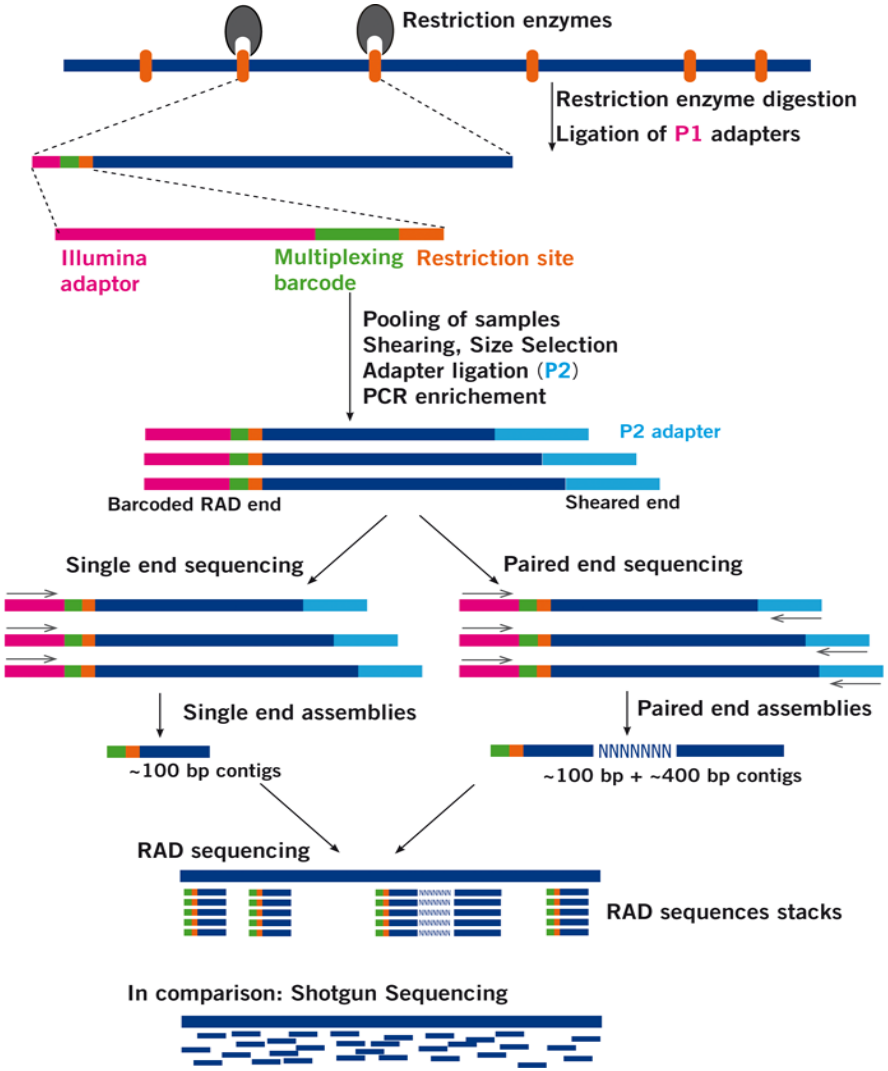
Fragmetace gelogenomové DNA po mocí restrikčních enzymů

Ligace sekvenačních adaptorů na výsledné fragmenty

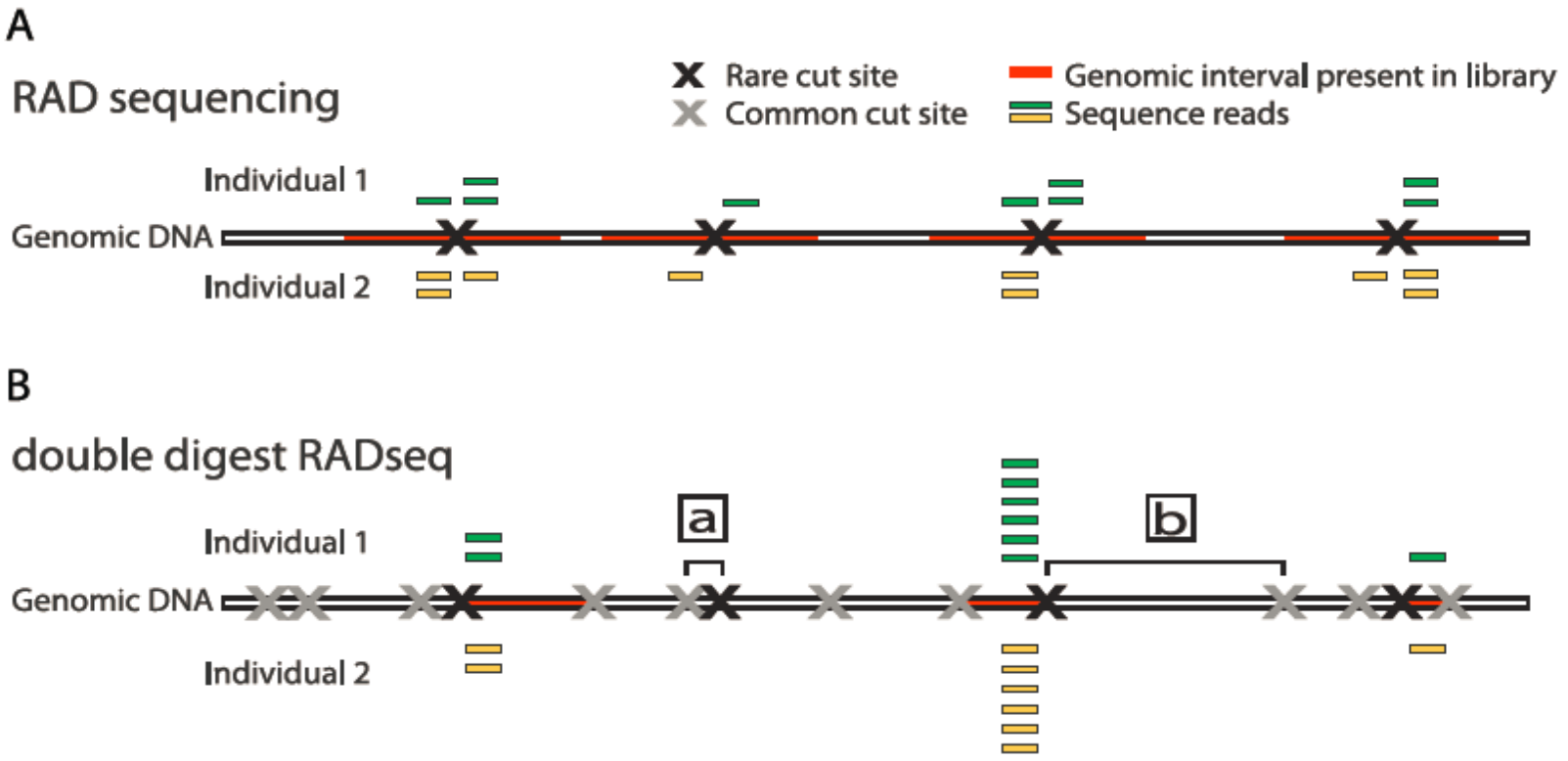
Následná sekvenace podél restrikčních míst

Celogenomové scany genetické variablility

Hledání SNPs, populační genomika (např. RAD-SEQ) apod.

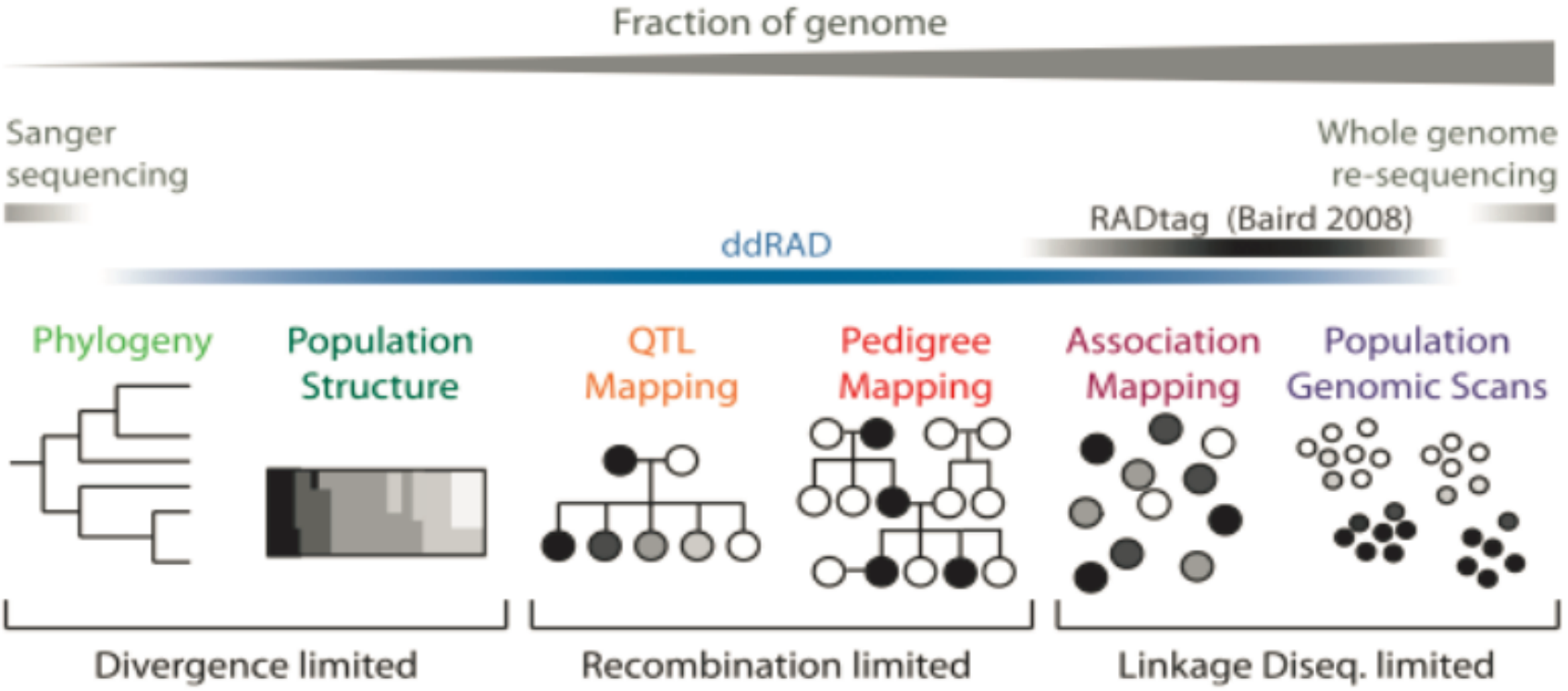


RAD vs. ddRAD



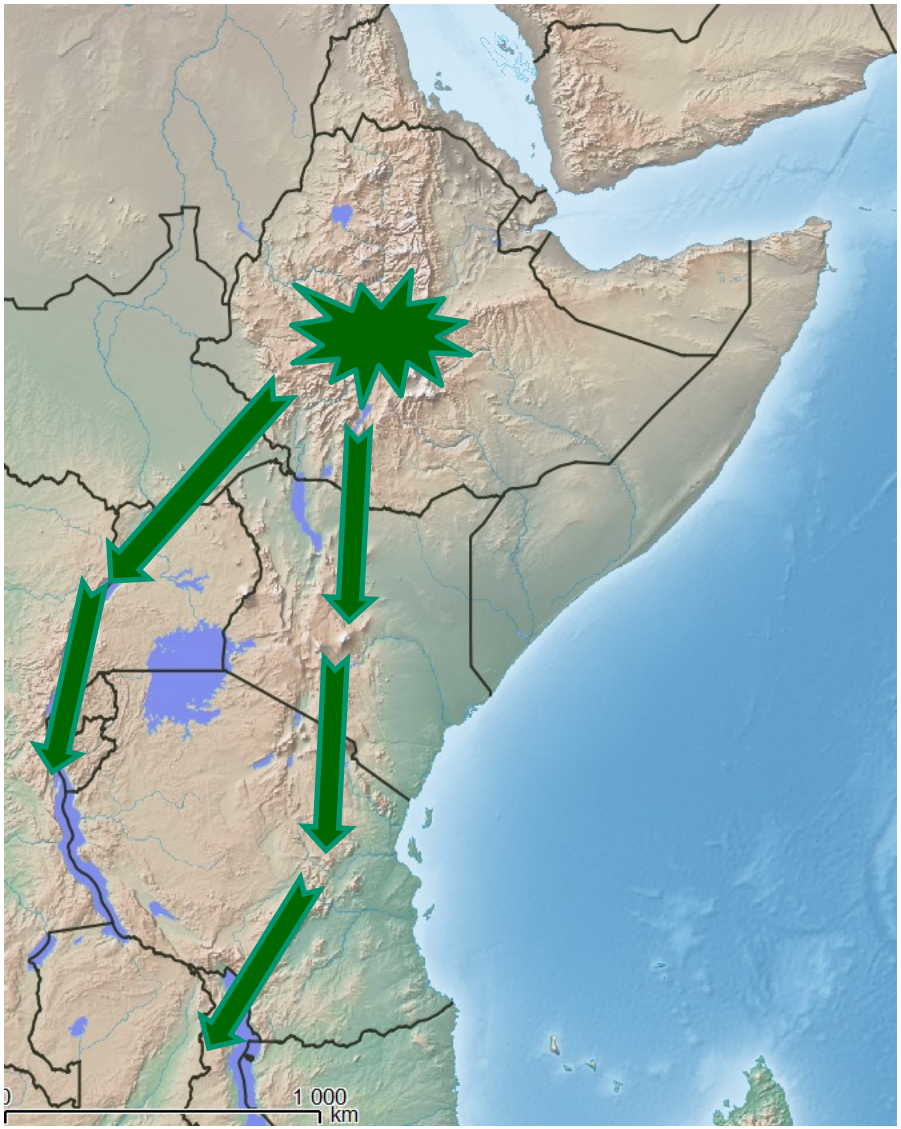
Peterson et al PLOS One 2012

Sekvenování podél restrikčních míst



Peterson et al PLOS One 2012

Phylogenomics of *Lophuromys*



- ancestral lineage „trapped“ in Ethiopian highlands, where diversified and sourced the colonization of other mountains (**mostly in Pleistocene**)

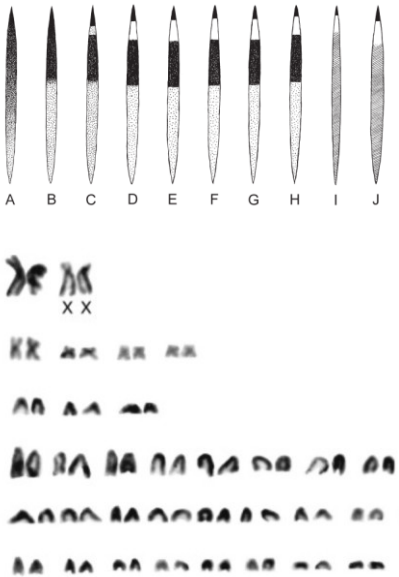
- ***Lophuromys flavopunctatus*** **complex (9 Ethiopian species)**

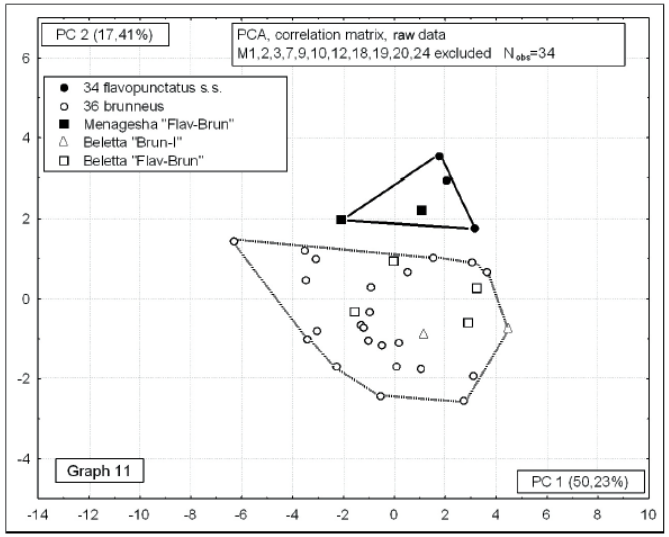
9 endemic species in Ethiopia

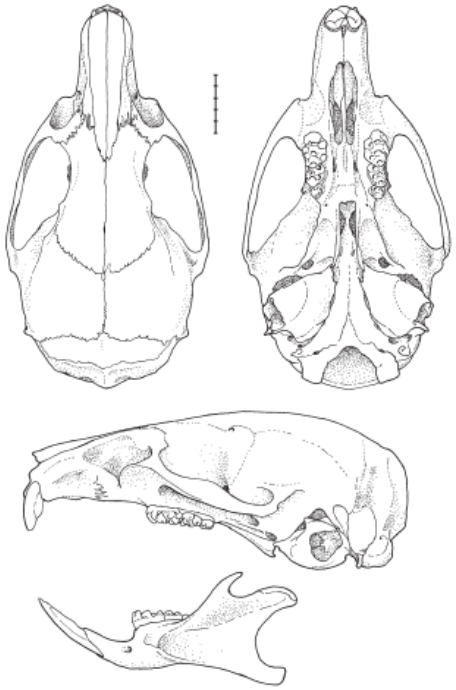
BULLETIN DE L'INSTITUT ROYAL DES SCIENCES NATURELLES DE BELGIQUE
BULLETIN VAN HET KONINKLIJK BELGISCH INSTITUUT VOOR NATUURWETENSCHAPPEN
BIOLOGIE, 77: 77-117, 2007
BIOLOGIE, 77: 77-117, 2007

Morphometric and genetic study of Ethiopian *Lophuromys flavopunctatus* THOMAS, 1888 species complex with description of three new 70-chromosomal species (Muridae, Rodentia)

by Leonid A. LAVRENCHENKO, Walter N. VERHEYEN, Erik VERHEYEN, Jan HULSELMANS & Herwig LEIRS







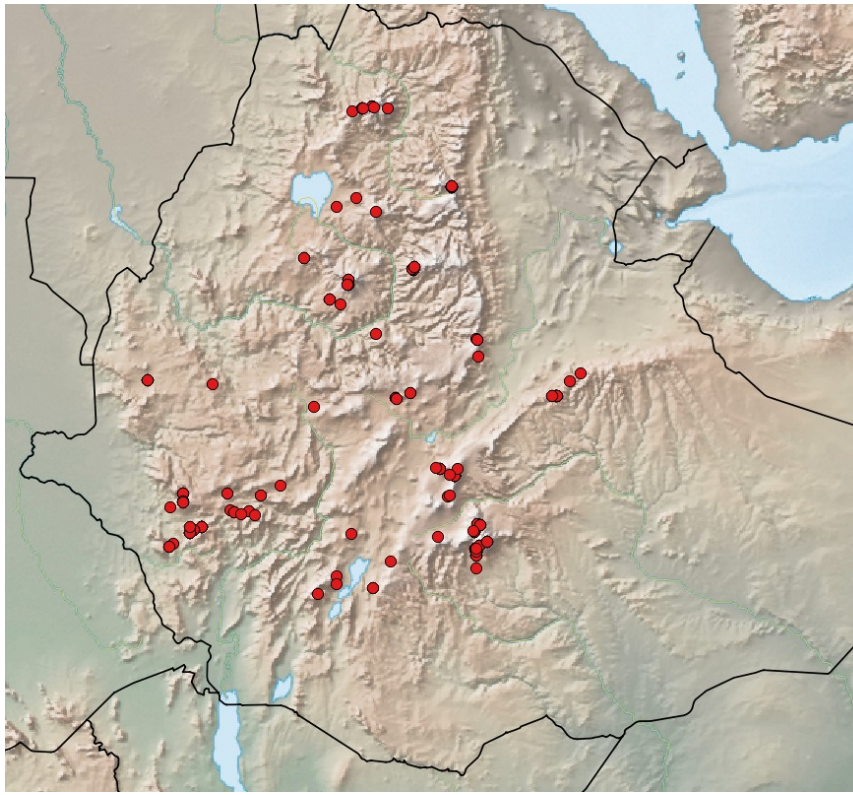
3.2. Views of skull and mandible of *Lophuromys menageshae* n.sp. (ZMMU S-165969, holotype). Scale bar = 5 mm.

Lophuromys - questions

- Are there really 9 well delimited species?
- Are they easily (genetically) recognizable? (e.g. mtDNA-barcoding)
- What is their distribution and ecological requirements? -> IUCN assessment, etc.

Material and Methods

- cca 500 specimens from all major mountain ranges
- mtDNA marker (CYTB)
- 4 nuclear markers (2 introny + 2 exony)
- **genomic approach – ddRAD sequencing**



Retaining well-covered & informative loci

All loci

No. of individuals:	213
No. of loci:	80570
No. of informative loci:	69724
No. of SNPs / PISs per informative locus:	
Min:	1 / 1
25%:	5 / 4
50%:	10 / 9
75%:	20 / 17
Max:	60 / 57
Loci per individual:	
Min:	5178
25%:	9719
50%:	12000
75%:	14607
Max:	23205
Individuals per locus:	
Min:	4
25%:	6
50%:	13
75%:	37
Max:	208
Proportion of missing data:	0.85

HQ loci

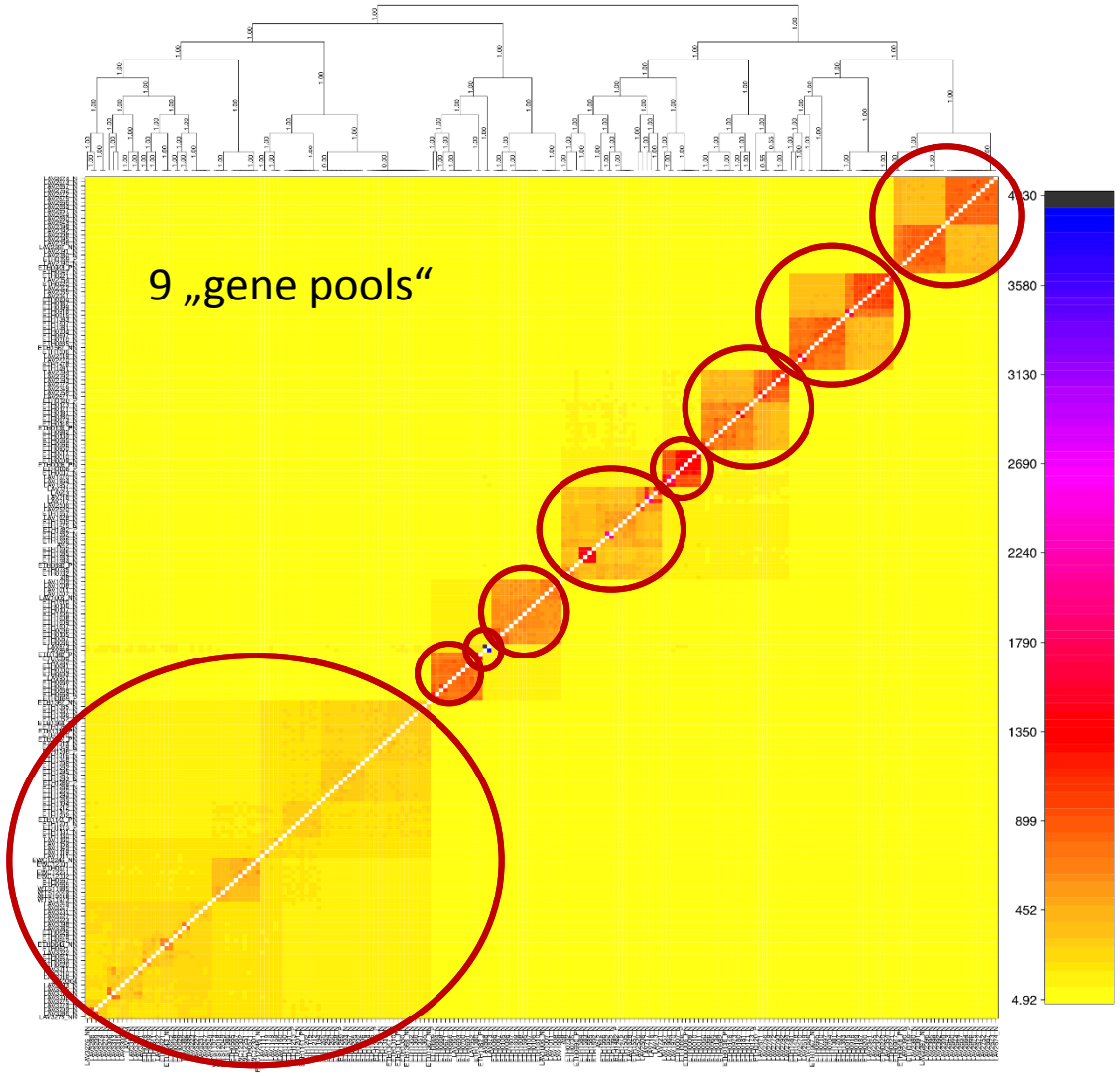
No. of individuals:	213
No. of loci:	15164
No. of informative loci:	15164
No. of SNPs / PISs per informative locus:	
Min:	1 / 1
25%:	17 / 14
50%:	25 / 21
75%:	32 / 28
Max:	57 / 54
Loci per individual:	
Min:	3393
25%:	6912
50%:	8074
75%:	9297
Max:	11912
Individuals per locus:	
Min:	54
25%:	74
50%: ✔	103
75%:	149
Max:	208
Proportion of missing data:	0.47 ✔

80 570 loci → filtering → 15 164 loci

ddRADseq: co-ancestry matrix

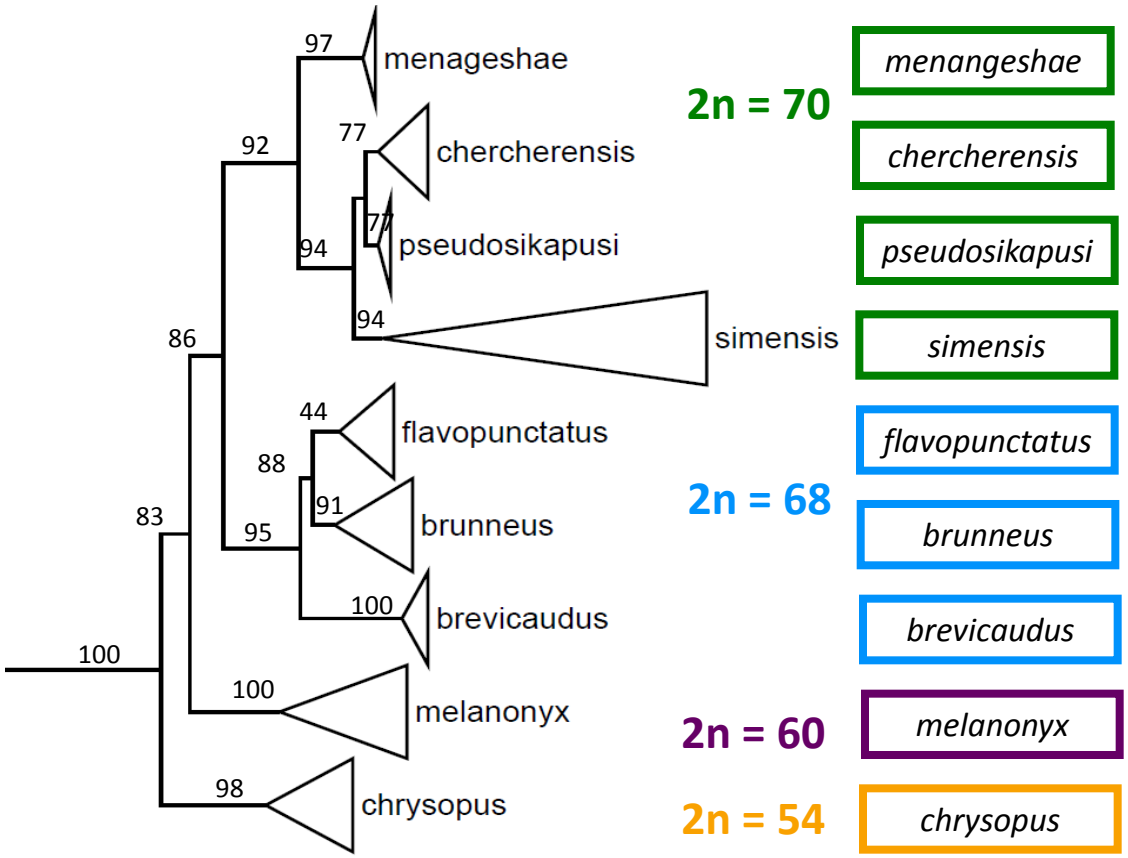
209 individuals

15 623 informative loci



Maximum likelihood analysis of concatenated nuclear dataset

Sanger sequencing



2n = 70

menangeshae

chercherensis

pseudosikapusi

simensis

2n = 68

flavopunctatus

brunneus

brevicaudus

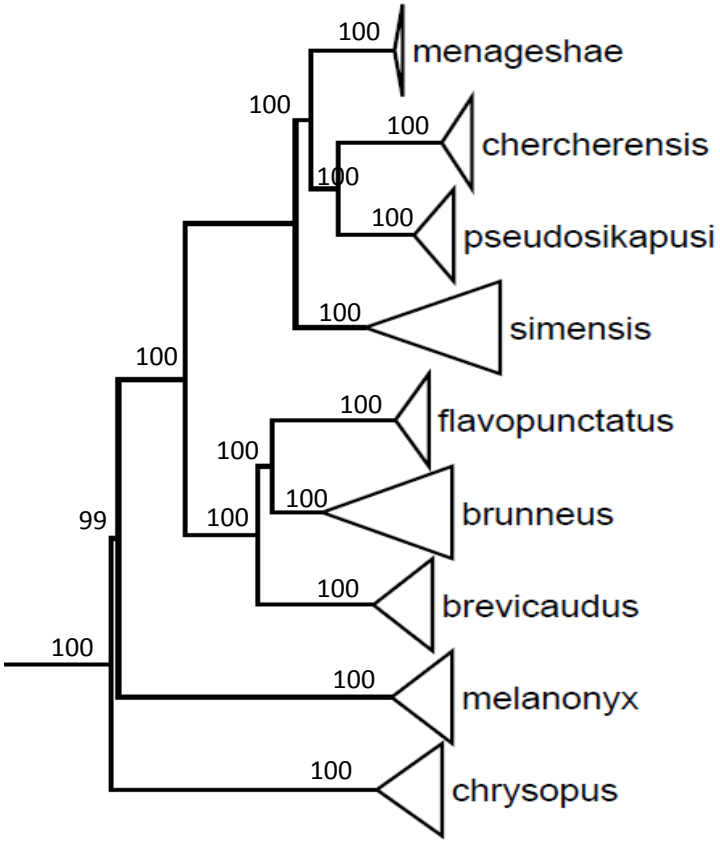
2n = 60

melanonyx

2n = 54

chrysopus

ddRADseq

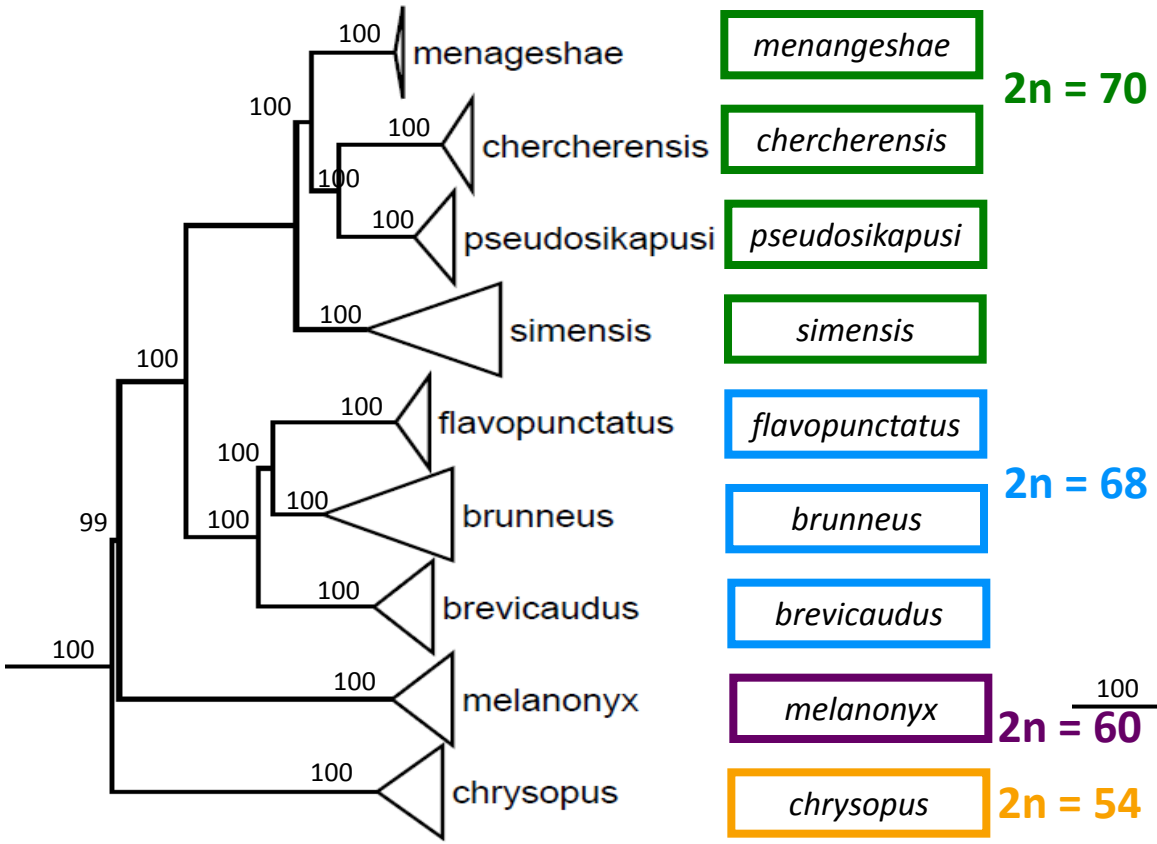


4 nuclear markers (V. Komarova et al.)
(2 604 bp concatenated dataset)

15 623 informative loci

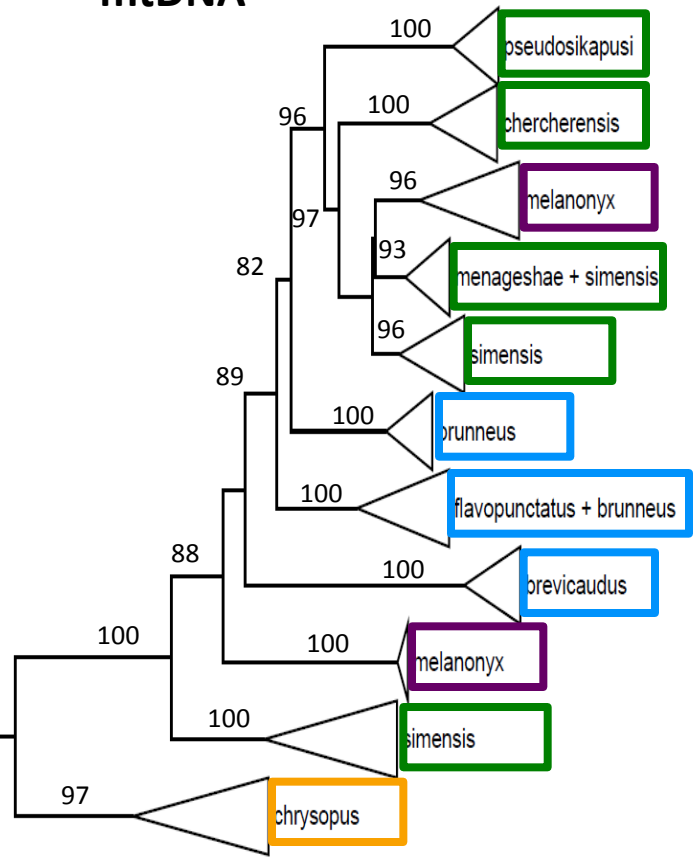
And what about mtDNA?

ddRADseq



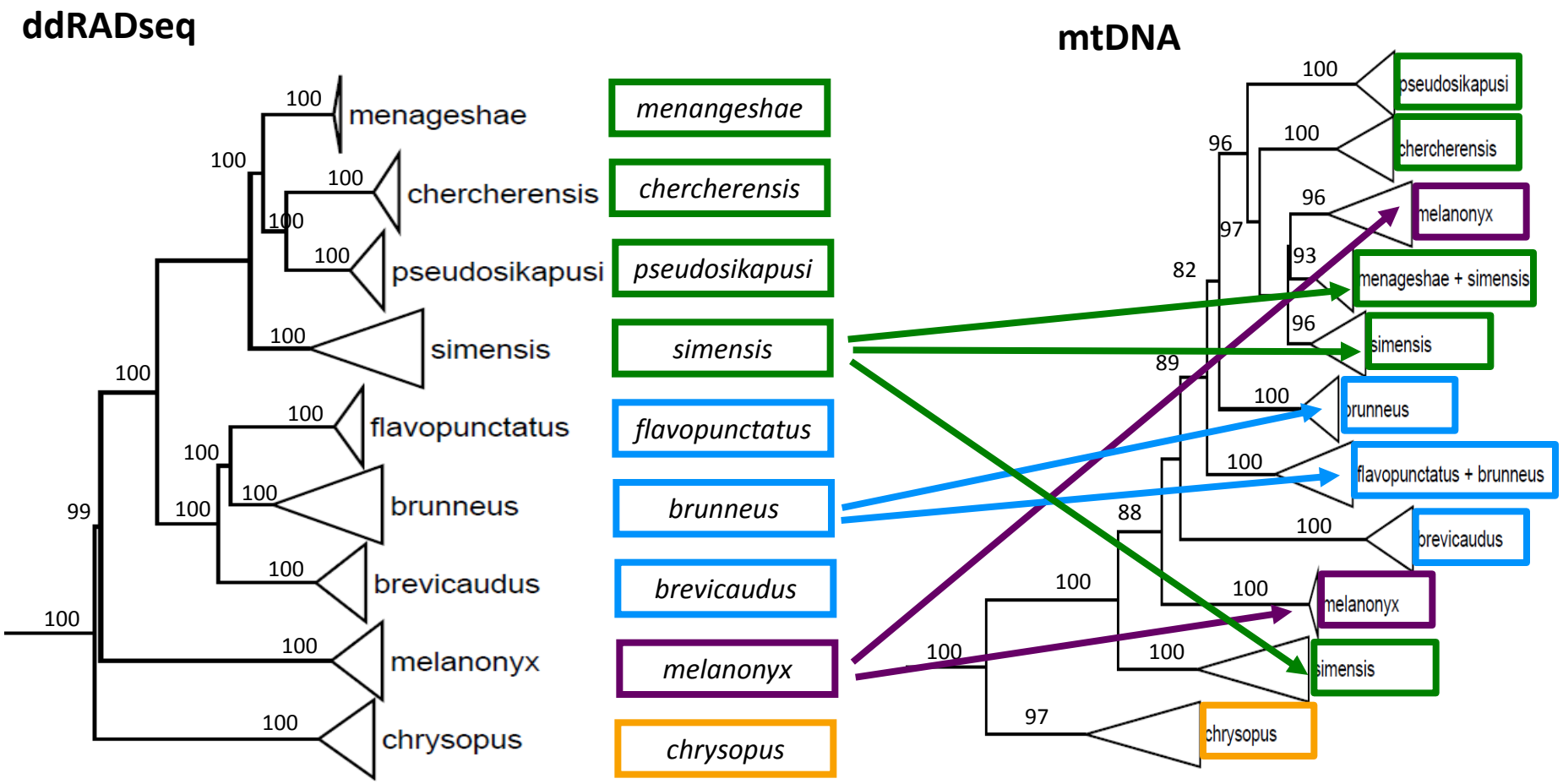
15 623 informative loci

mtDNA



cytochrome *b* (1140 bp)

And what about mtDNA?



15 623 informative loci

cytochrome *b* (1140 bp)

„reticulate evolution“