

Next-generation sequencing (NGS)

Sanger sequencing

Primer - F - AAGTCAGTCTAA**A**=0 -

Primer - F - AAGTCAGTCT**A**=0

Primer - F - AAGTCAGTCT**T**=0

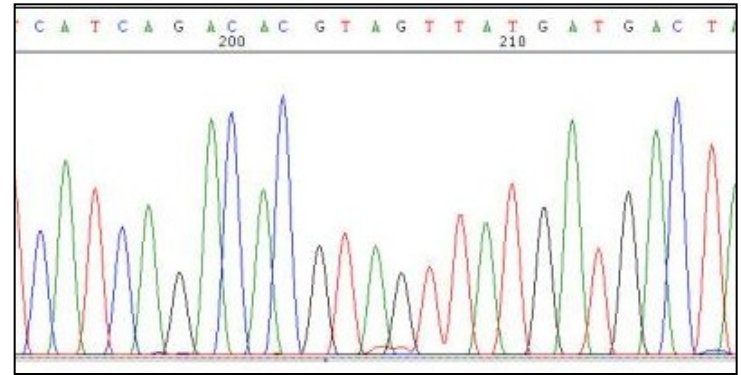
Primer - F - AAGTCAGT**C**=0

Primer - F - AAGTCAG**T**=0

Primer - F - AAGTCAG**G**=0

Primer - F - AAGTC**A**=0

Primer - F - AAGT**C**=0



krátké ----- dlouhé
(rychlé) ----- (pomalé)

+

Primer - F **AAGTCAGTCTAA**ATGCGATTGGGA Rev. Primer - R

Rev. Primer - F **TTCAGTCAGATTACGCTAACCT** Primer - R

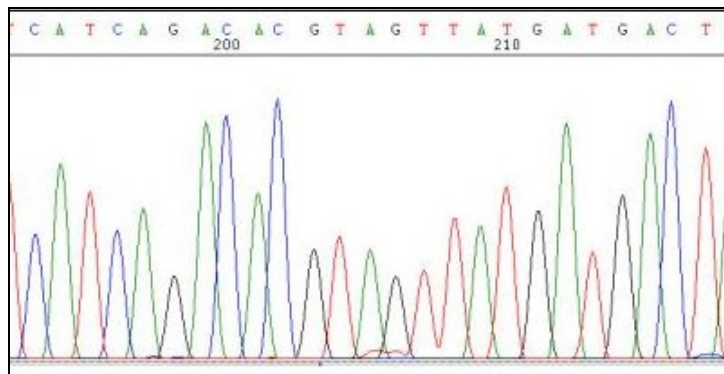
4-kapilární sekvenátor

=

96 x 500 bp/12 hodin

=

cca 100 000 bp/den



detector

↑
laser beam

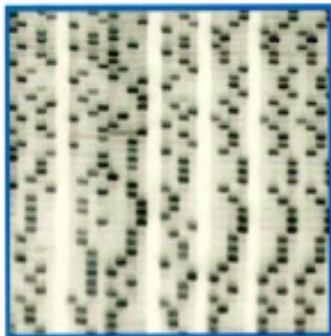
capillary
electrophoresis

+

-

Evolve 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*

4-kapilární sekvenátor

=

96 x 500 bp/12 hodin

=

cca 100 000 bp/den

Next-generation sequencing

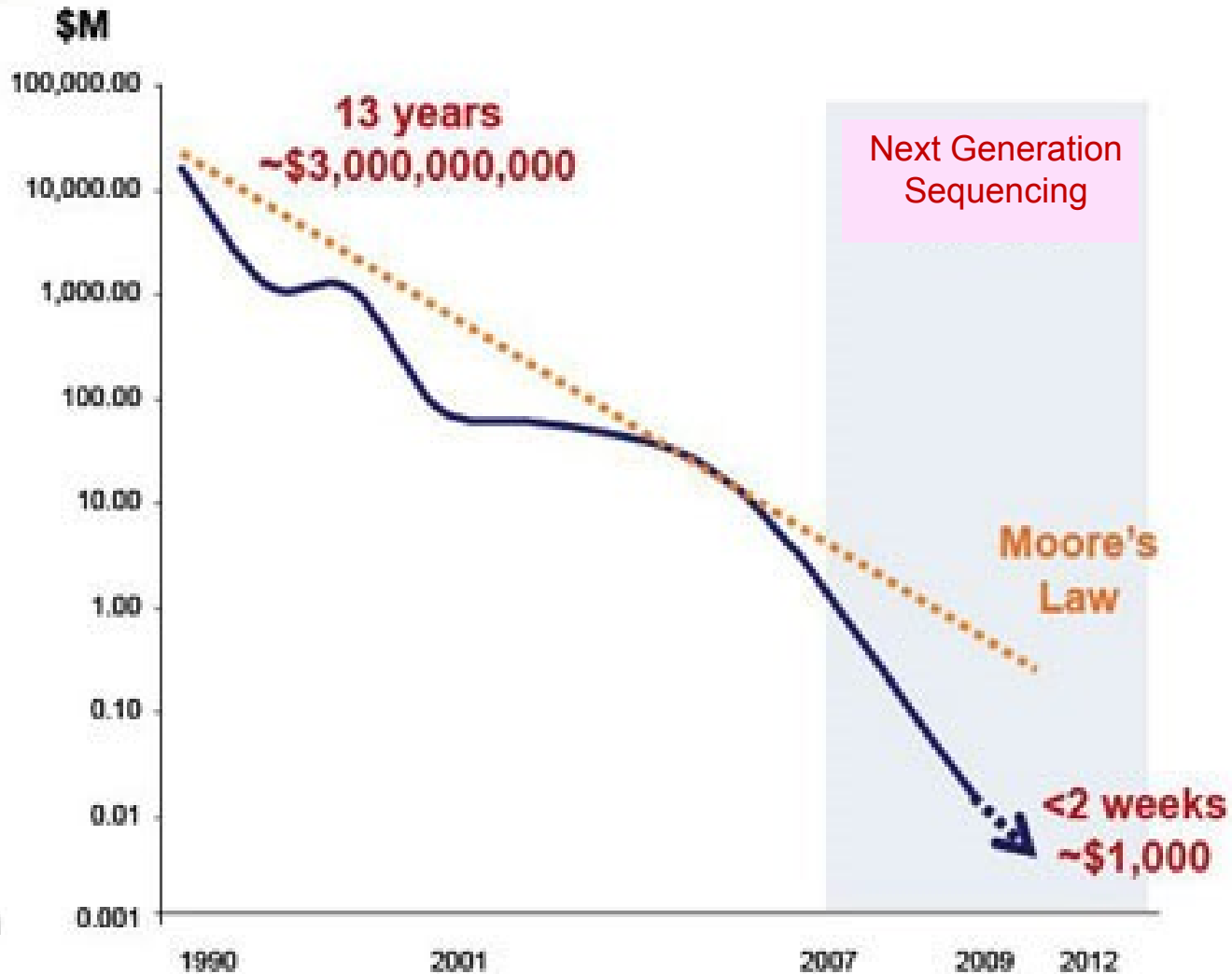
=

cca 5 000 000 000 bp/den

electrophoresis

Next-generation sequencing (NGS)

Cost per Human Genome



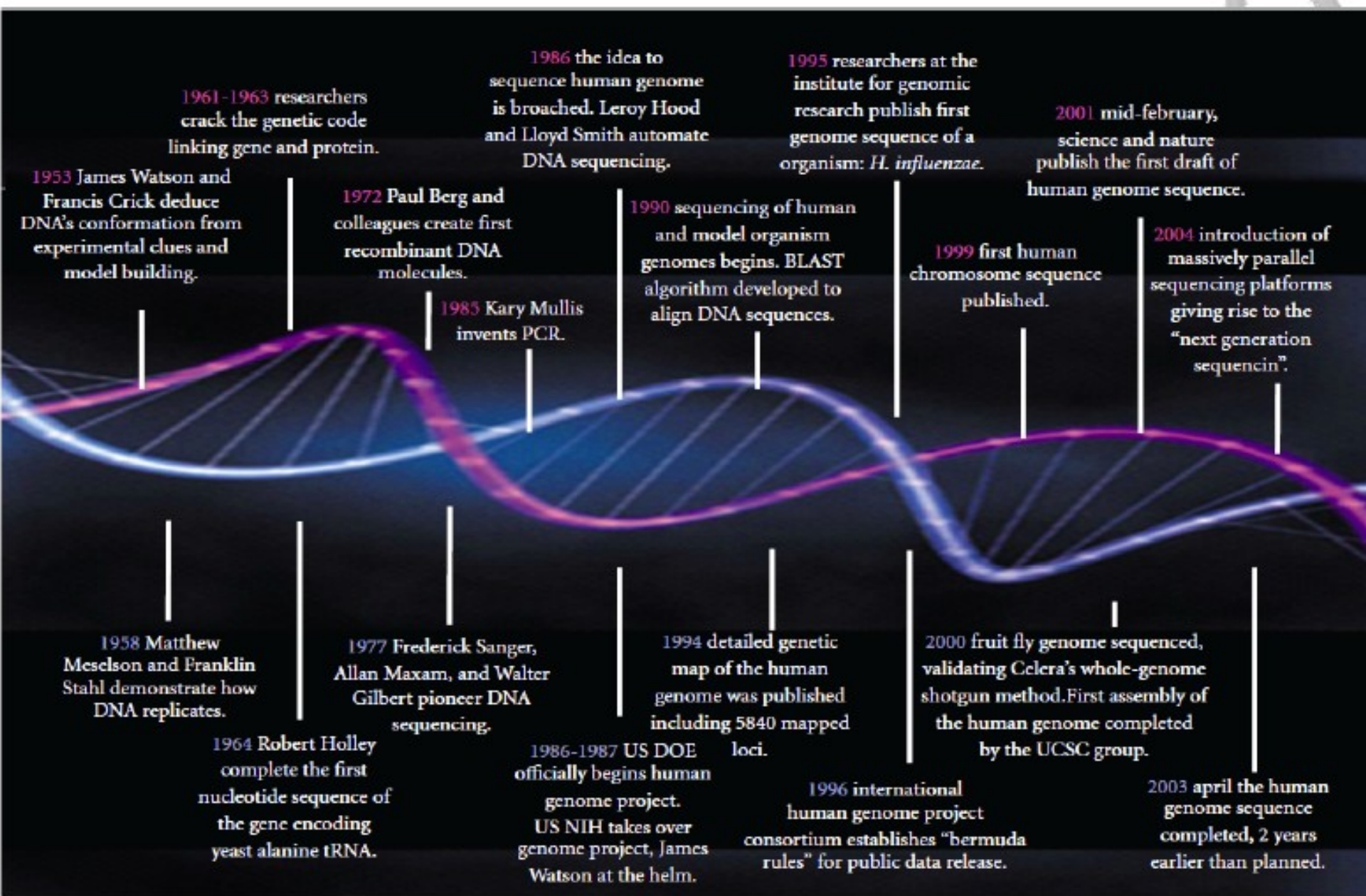
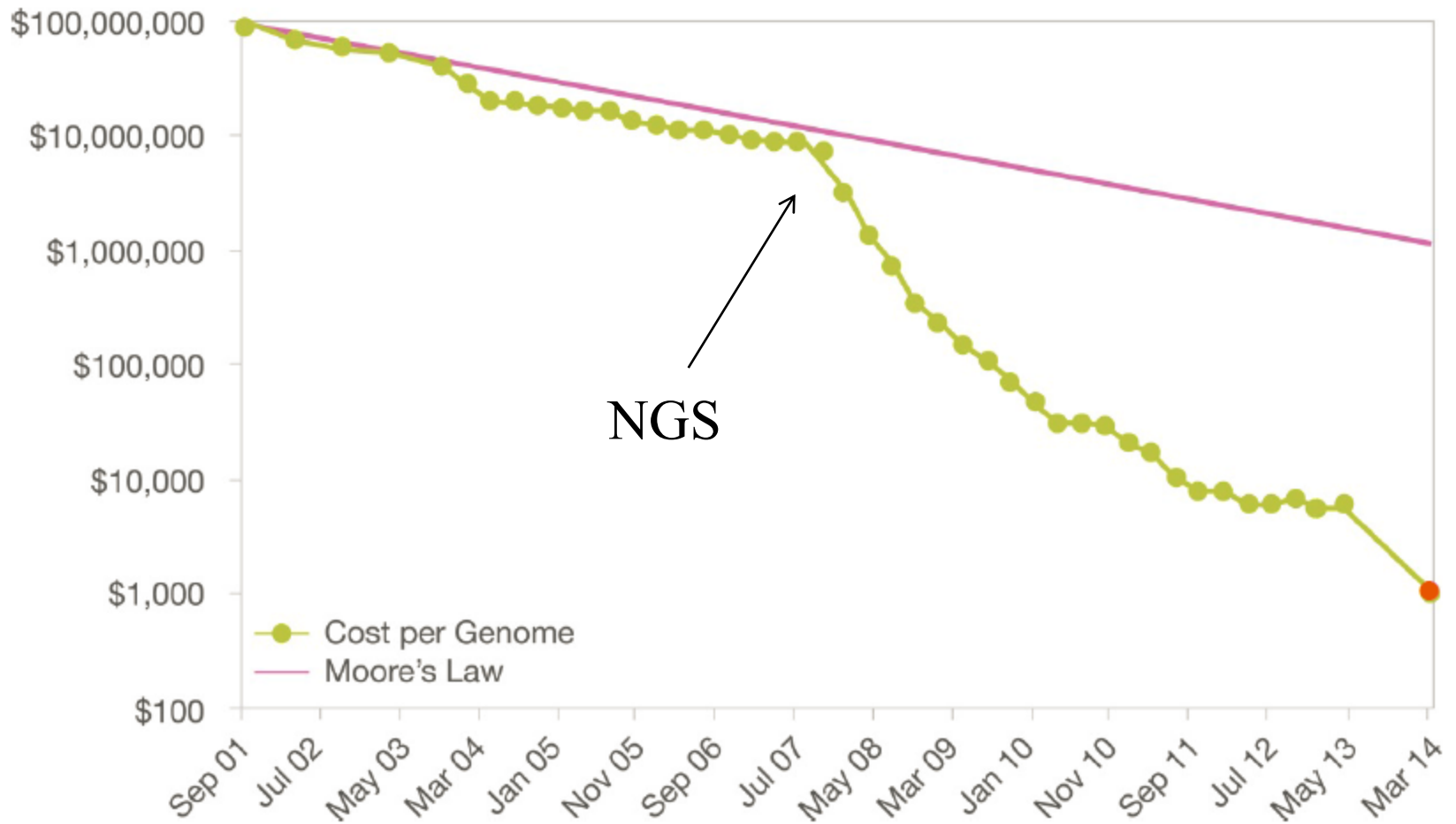


FIGURE 1: Evolution of DNA revolution.



NHGRI "Cost of sequencing a human genome", with one data point ● from Illumina corporate sales

Illumina HiSeqX10



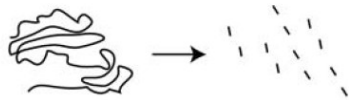
\$1 M per machine

1.8 Tbase per machine per 3 days

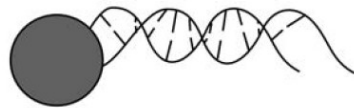
1800 human genomes per machine per year

Historie „Next generation sequencing“

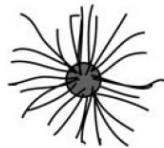
1) Randomly fragment many molecules of target DNA



2) Immobilize individual DNA molecules on solid support



3) Amplify DNA in clonal 'polymerase colony'



4) Sequence DNA by adding liquid reagents to immobilized DNA colonies



5) Interrogate sequence incorporation *in situ* after each cycle using fluorescence scanning or chemiluminescence



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
- SMRT (Pacific Biosciences)
- Oxford Nanopore – stále ve vývoji
- ...

454 pyrosequencing

- emulzní techniky amplifikace pikolitrové objemy
- simultánní sekvenování na destičce z optických vláken detekce pyrofosfátů uvolňovaných při inkorporaci bází
- První generace GS20 → 200 000 reakcí najednou (zhruba 20 milionů bp) dnes FLX → 400 000 reakcí najednou = eukaryotní genom za týden!!!
- Délka jednotlivých sekvencí 100 - 400 (800 bp)



Molecular Ecology (2008) 17, 1629–1635

NEWS AND VIEWS

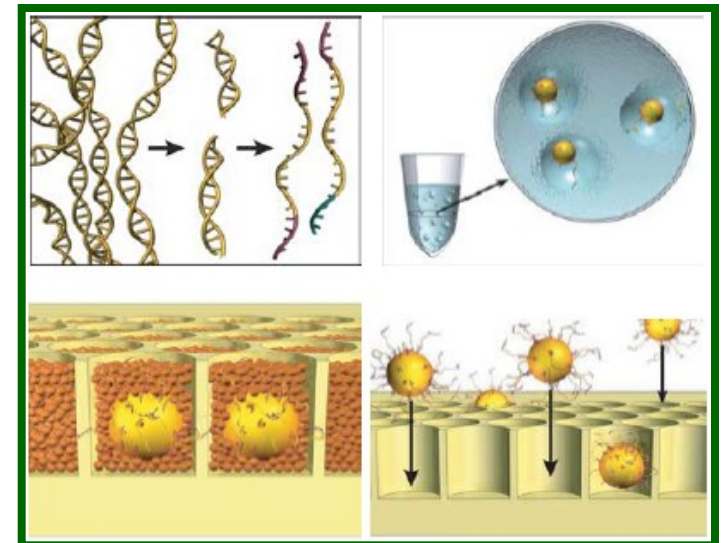
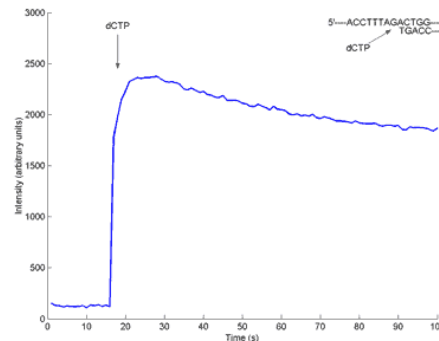
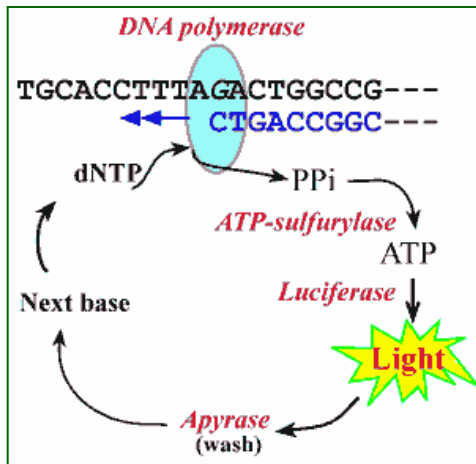
PERSPECTIVE

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



Pracovní postup



1

DNA Library Preparation

1. DNA Fragmentation (Nebulization)
2. DNA Fragment Size Selection
3. DNA Sample Quality Assessment (Nebulized or *LMW* DNA Sample)
4. Fragment End Polishing
5. Adaptor Ligation
6. Small Fragment Removal
7. Library Immobilization
8. Fill-In Reaction
9. Single-Stranded DNA Library Isolation
10. DNA Library Quality Assessment and Quantitation

Time: 11 - 72 h

General Laboratory 1



2

Emulsion-Based Clonal Amplification (emPCR)

1. Preparation of the Live and Mock Amplification Mixes
2. DNA Library Capture
3. Emulsification
4. Amplification
5. Bead Recovery
6. DNA Library Bead Enrichment
7. Sequencing Primer Annealing

Time: 11 - 13 h

Controlled Room

Amplicon Room



3

Sequencing / Genome Sequencer FLX Operation

1. The Pre-Wash
2. PicoTiterPlate Device Preparation
3. The Sequencing Run

Time: 11.5 h

General Laboratory 2



4

Data Processing and Analysis

1. Data Processing
 - a) Image Processing
 - b) Signal Processing
2. Data Analysis
 - a) Assembly
 - b) Mapping
 - c) Amplicon Variant Analysis

Time: variable



1. Příprava jednořetězcové DNA knihovny (ssDNA library preparation)

1 DNA Fragmentation (Nebulization):



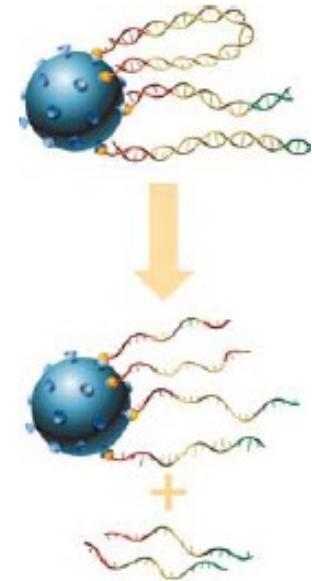
5 Adaptor Ligation:



7 Library Immobilization:



9 ssDNA Library Isolation:



Adaptor A + Adaptor B

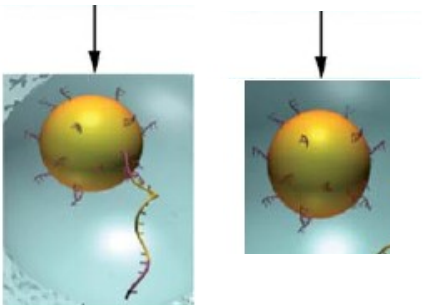
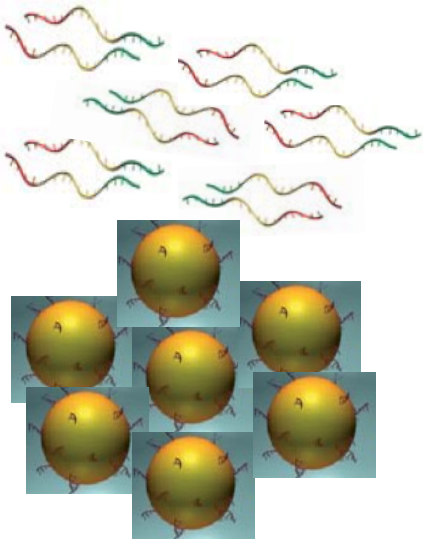
-Slouží jako vazebné místo primerů pro následnou PCR amplifikaci a sekvenování

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

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

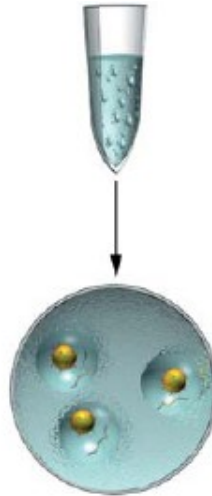
1 DNA Library Capture:

- poměry nastavit tak aby
1 kulička \leq 1 molekula DNA

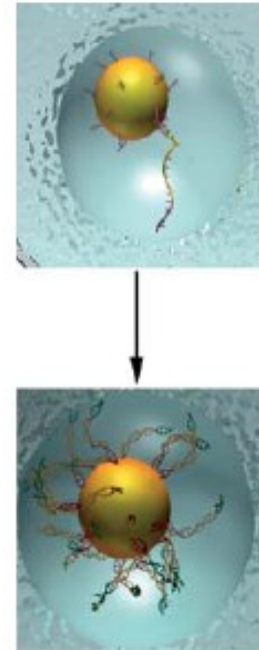


2 Preparation of the Amplific. Mixes

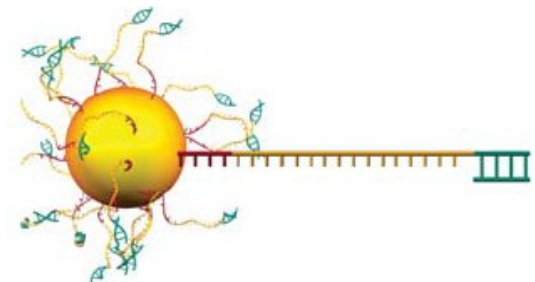
3 Emulsification:



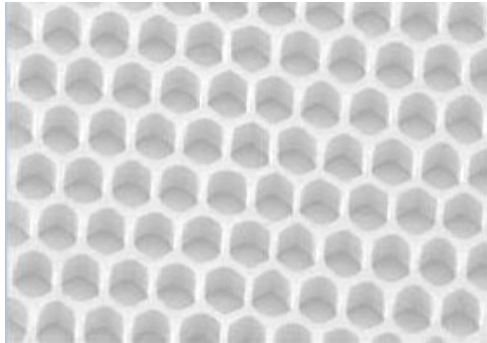
4 emPCR Amplification:



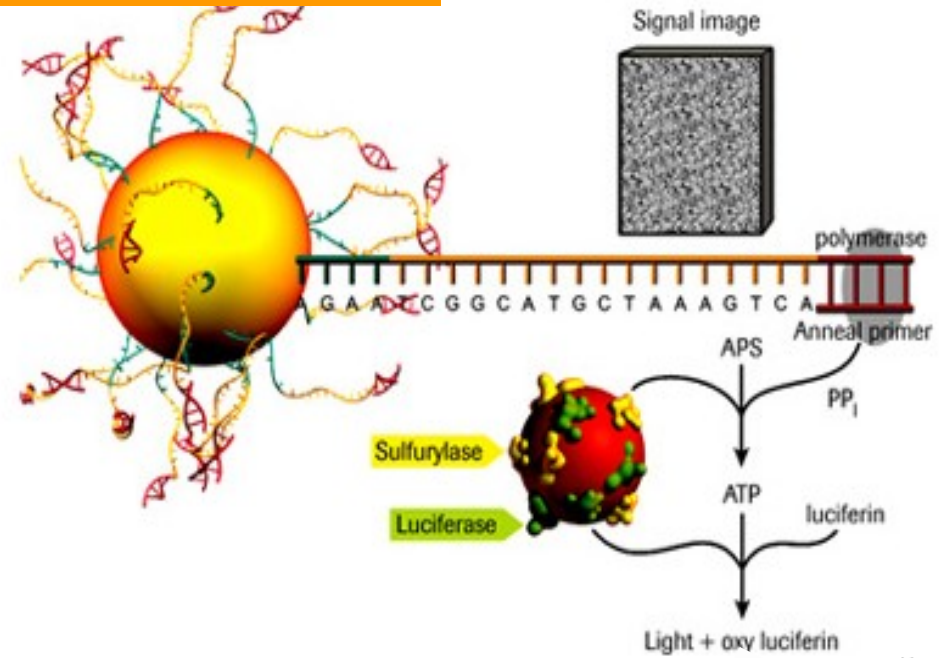
7 Sequencing Primer Annealing:



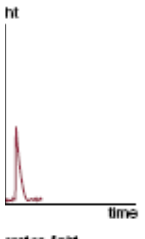
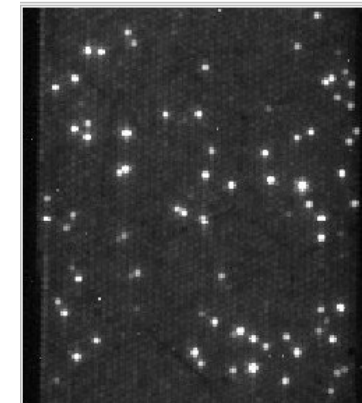
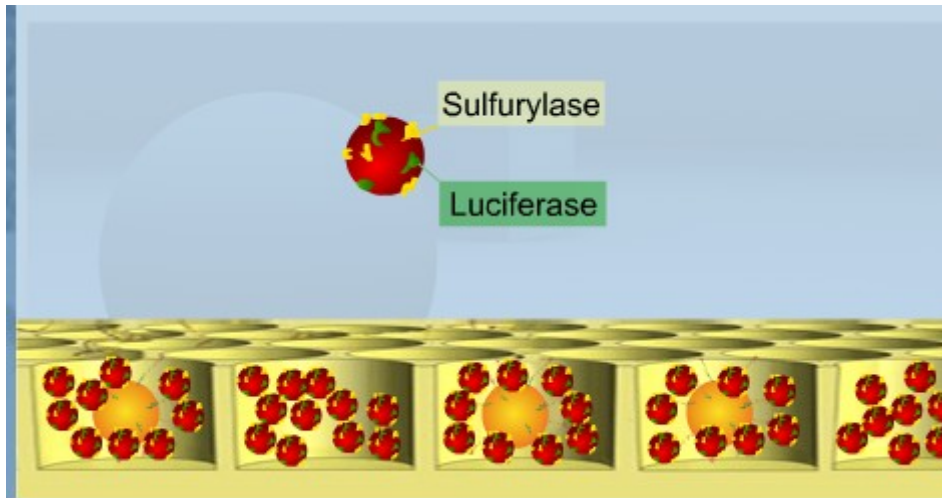
3. Pyrosekvenování („sequencing by synthesis“)



pikotitrační destička



Na jedné destičce 400 000 až 1milión jamek



3. Pyrosekvenovani - 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 odmyje

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

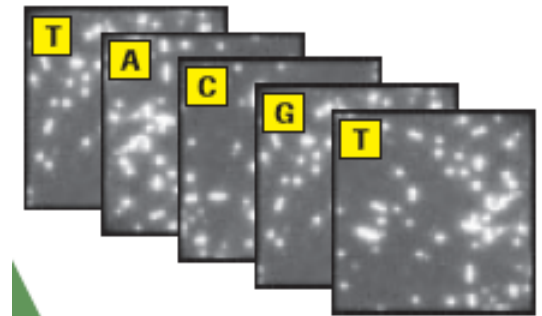
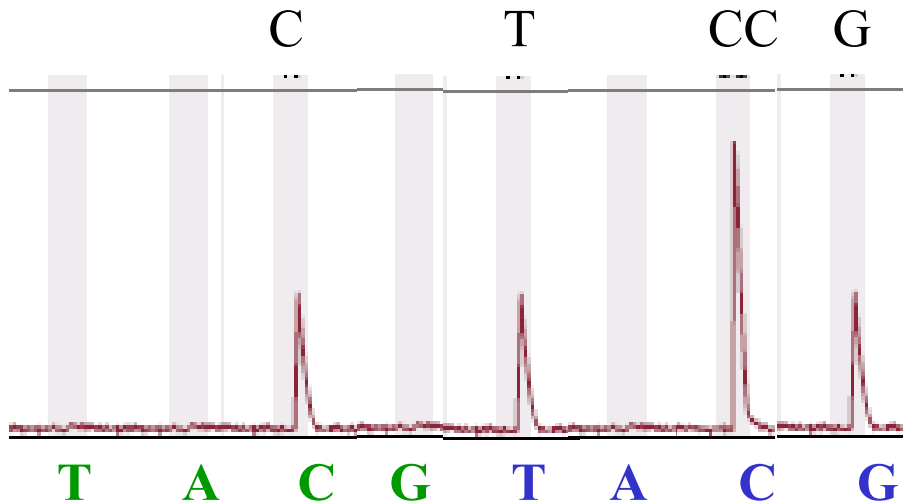


Image Files:
12-15 gigabytes
per run

Problém!!!! Homopolymery 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)

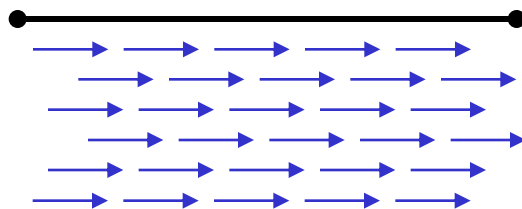
GS Junior

• 400 bp reads ~ 35Mbp/run (10 hrs)



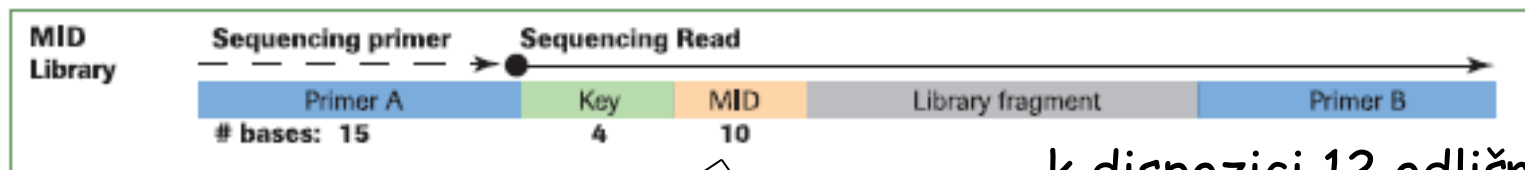
!!! Samozřejmě nestačí mít každou bázi osekvenovanou 1x !!!

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

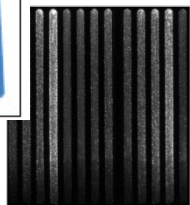
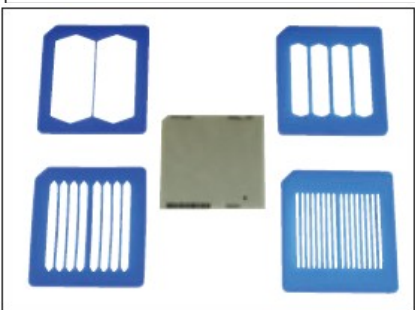


Kapacita destičky **400 Mb (GS FLX Titanium)**:

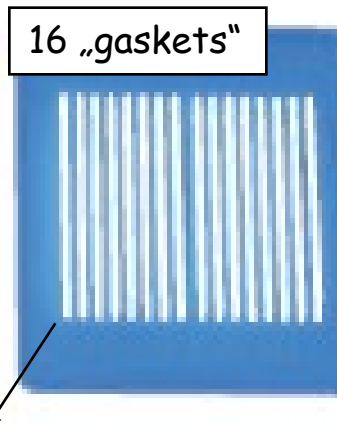
Mus:	2700 Mb	→ 7 run 1x coverage
Caenorhabditis:	100 Mb	→ 1 run 4x coverage
E. coli:	5 Mb	→ 1 run 80x coverage
mitoch. Mus:	0.016 Mb	→ 1 run 25000x coverage
HIV:	0.01 Mb	→ 1 run 40000x coverage



-k dispozici 12 odlišných MID („multiplexing“)



1. CCCCCCCCCC
2. GGGGGGGGGG
- ...
12. CCCCAAAG



$$\begin{array}{r}
 12 \text{ MID} \\
 \times \\
 16 \text{ gaskets} \\
 = \\
 \text{max. 192 vzorků}
 \end{array}$$

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

Illumina HiSeq/MiSeq

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

<http://www.youtube.com/watch?v=womKfikWlxM::>

Illumina HiSeq

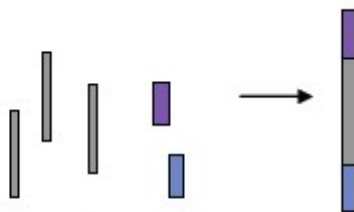


Illumina MiSeq



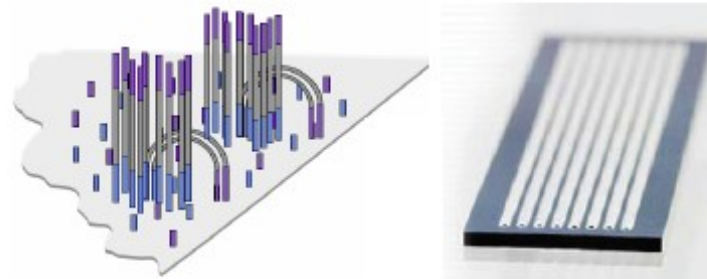
Illumina Sequencing pipeline

1. Sample Prep (1-5 days)



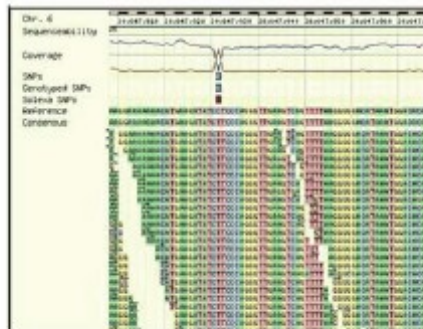
Ligate adapters

2. Cluster generation on flow cell (1.5 day)



Clonal Single molecular Array

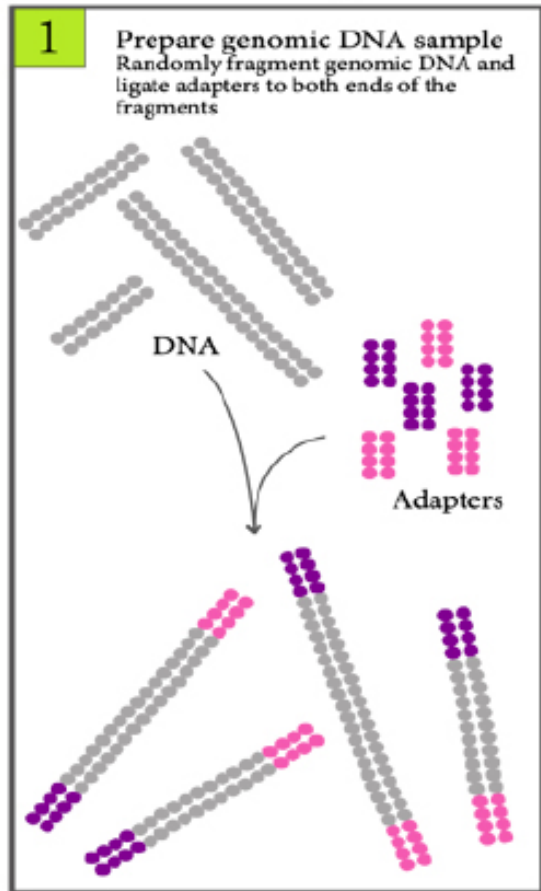
4. Data Analysis (days-months)



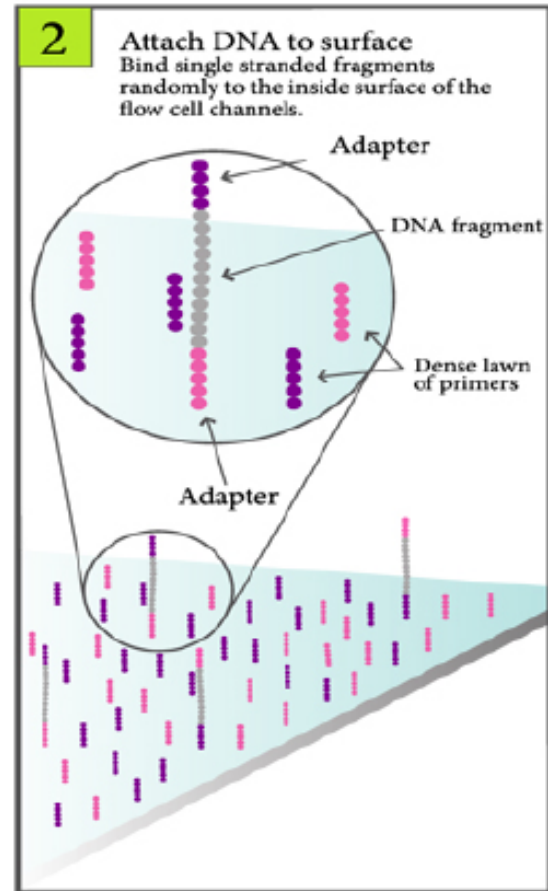
3. Sequencing and imaging (2-3 days)



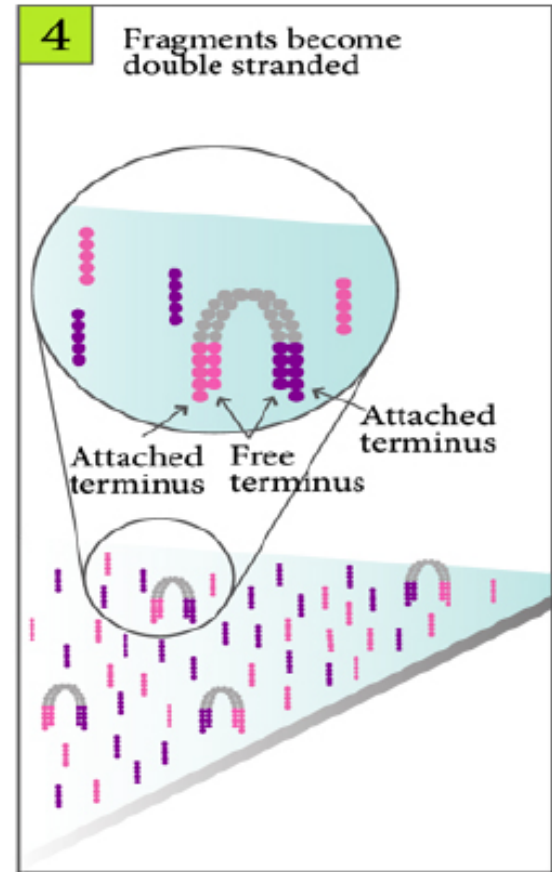
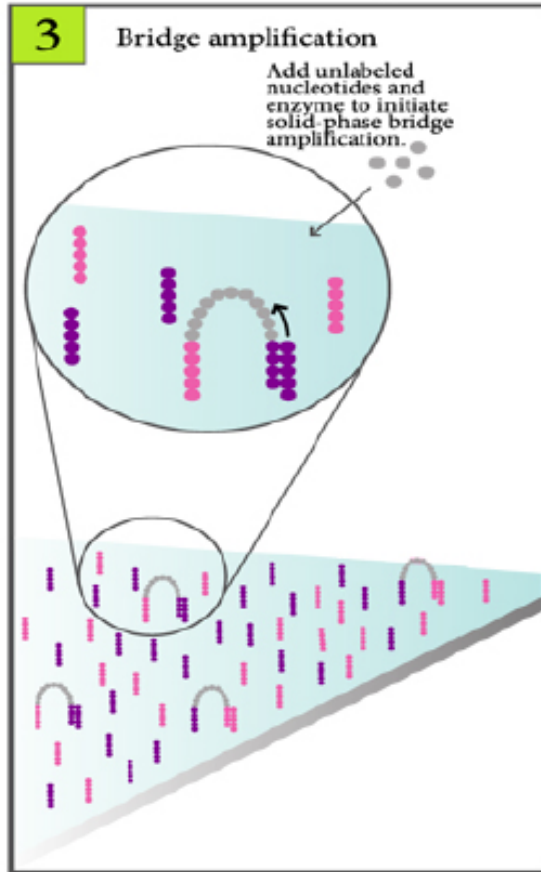
Attach DNA to flow cell



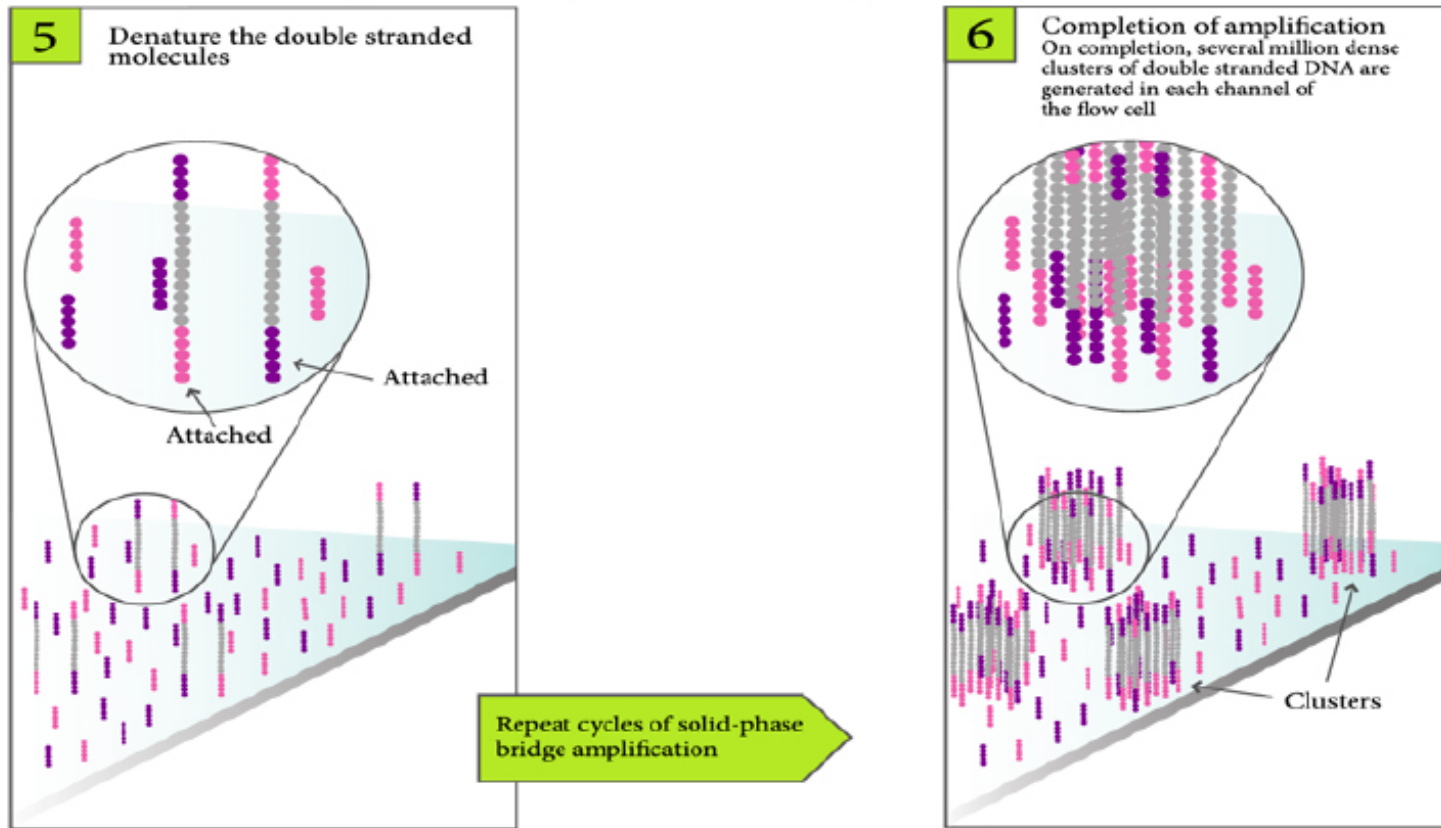
Add sample to flow cell



Bridge Amplification

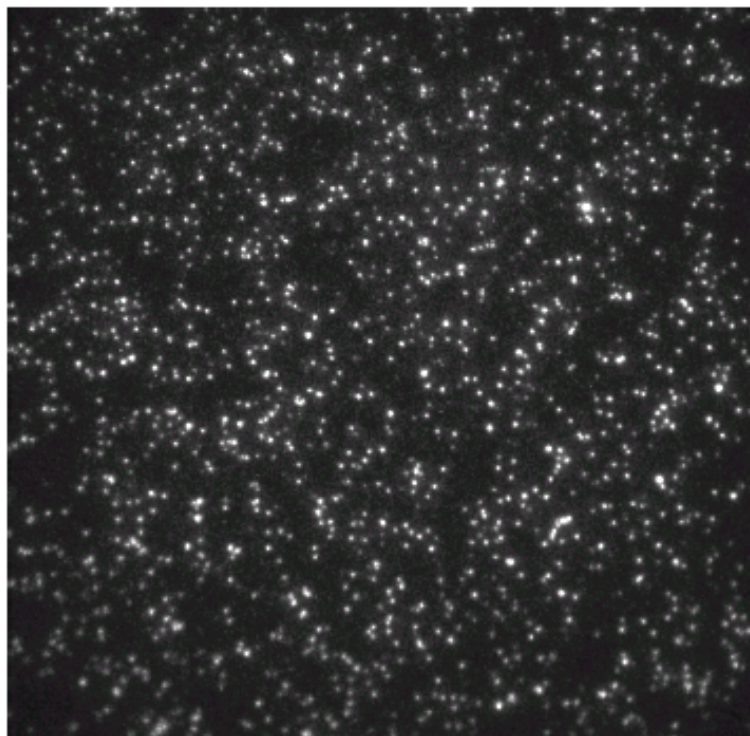


Cluster Generation



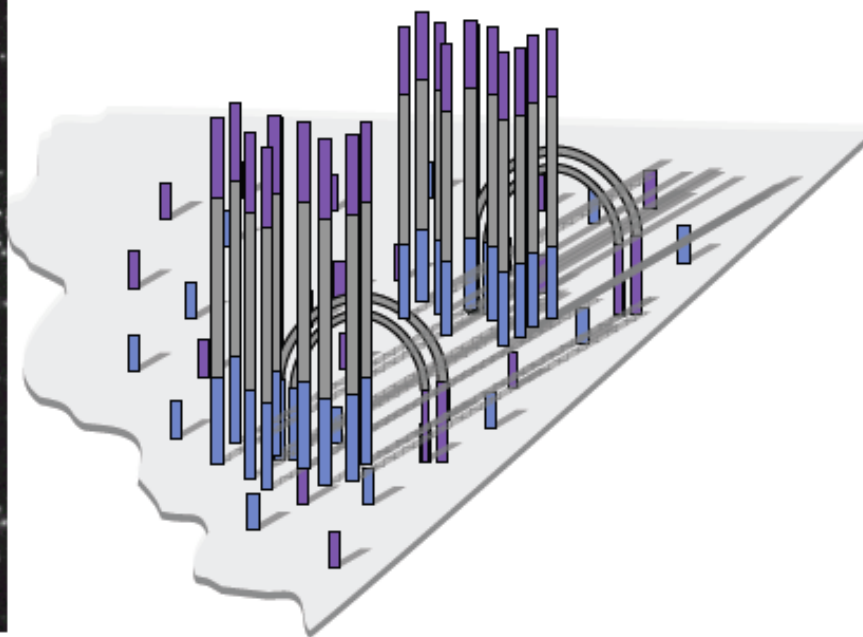
Clonal Single molecular Array

Clonal Single molecule Array



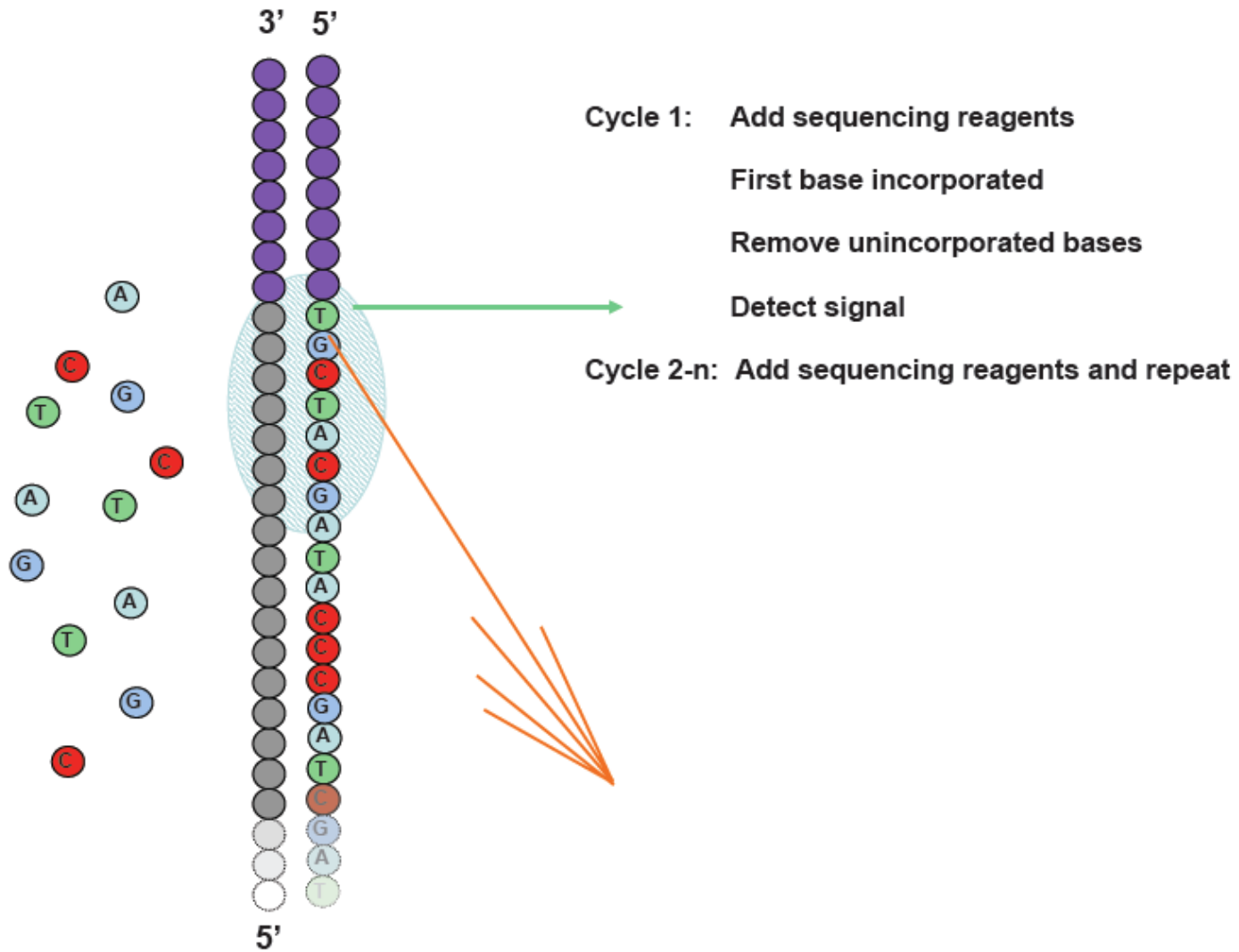
100um

Random array of clusters



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

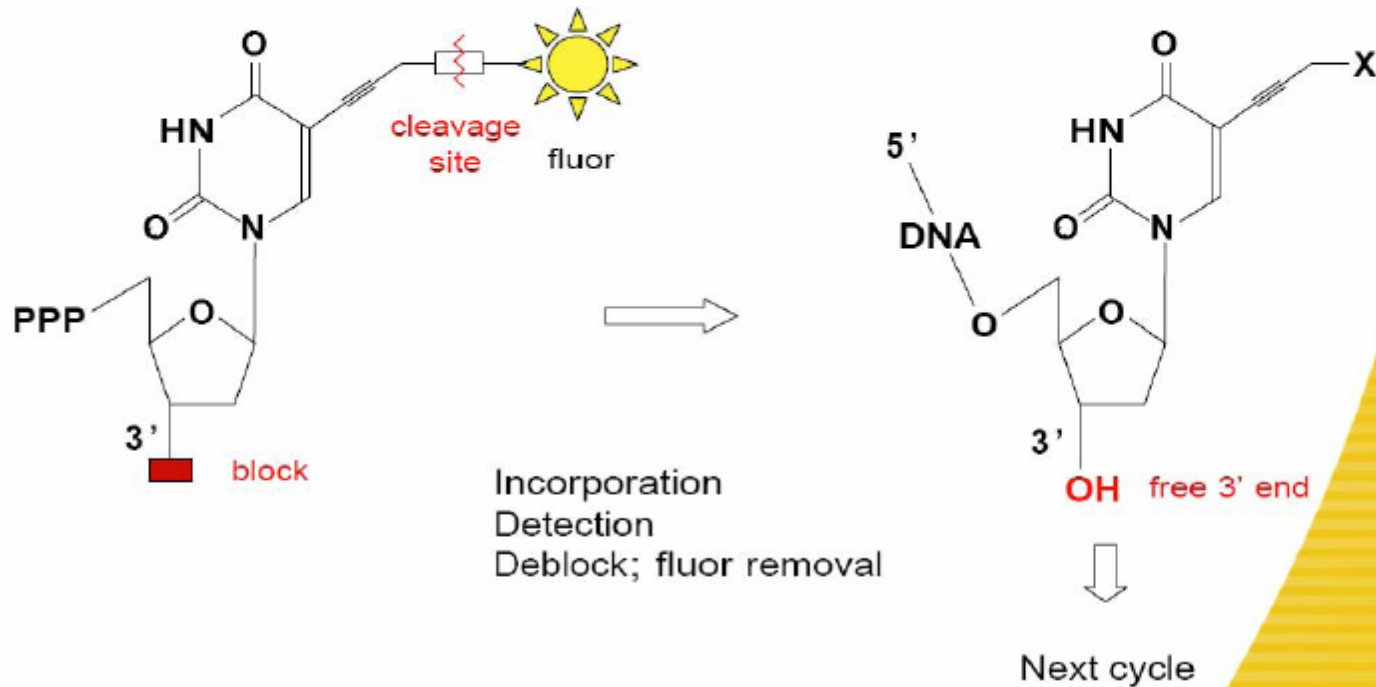
Sequencing By Synthesis (SBS)



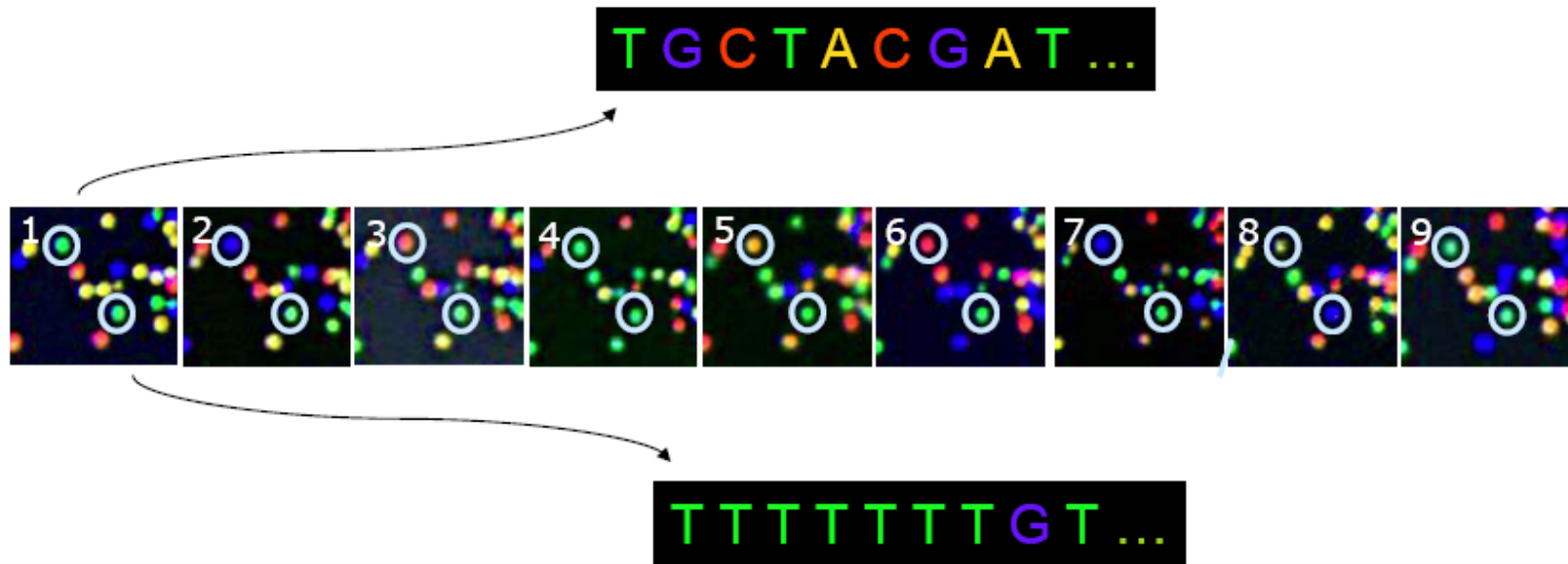
Reversible Terminator Chemistry



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

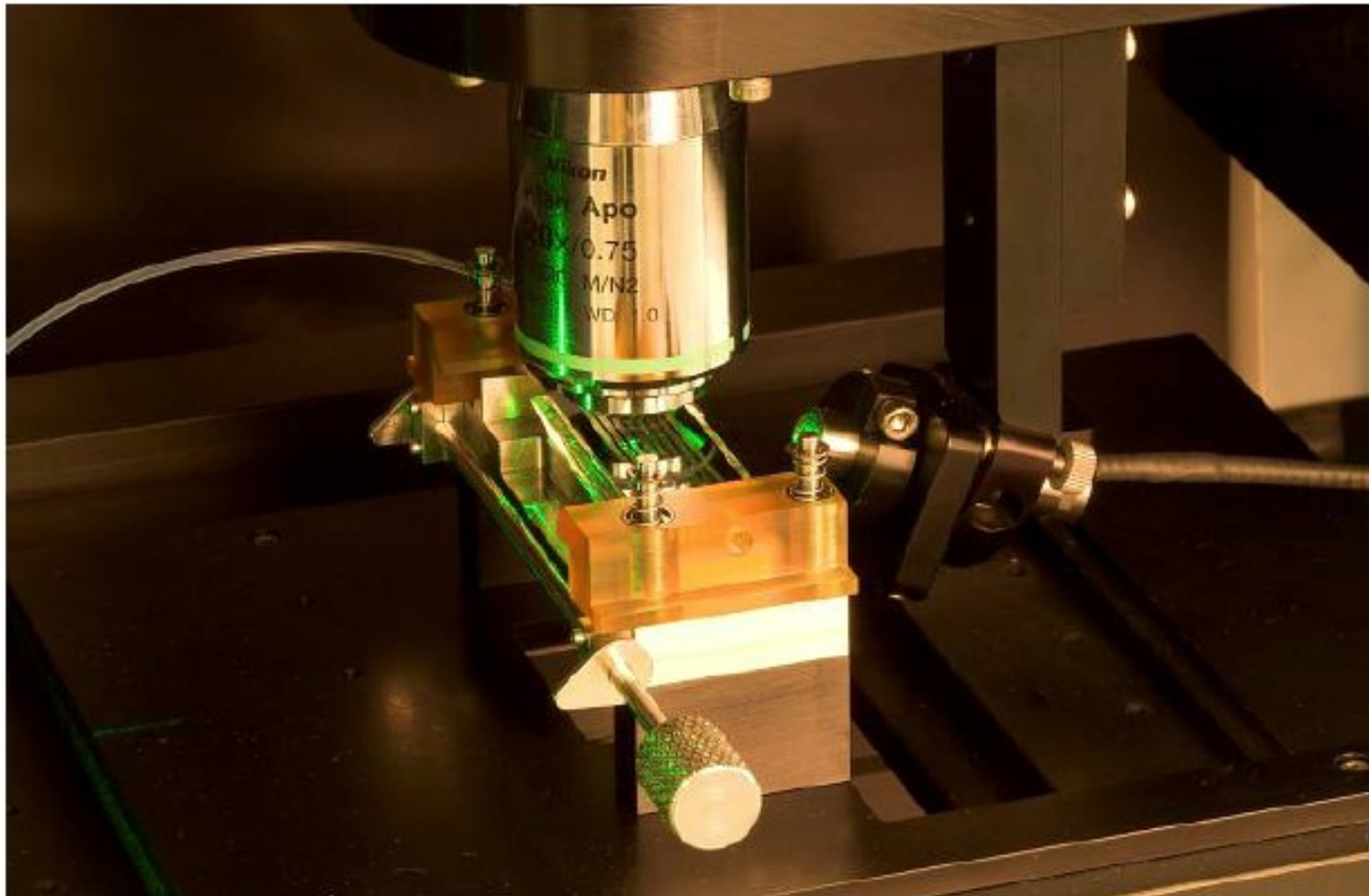


Base Calling From Images



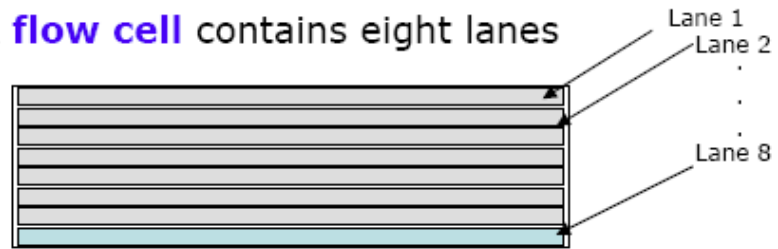
The identity of each base of a cluster is read off from sequential images

Flowcell imaging

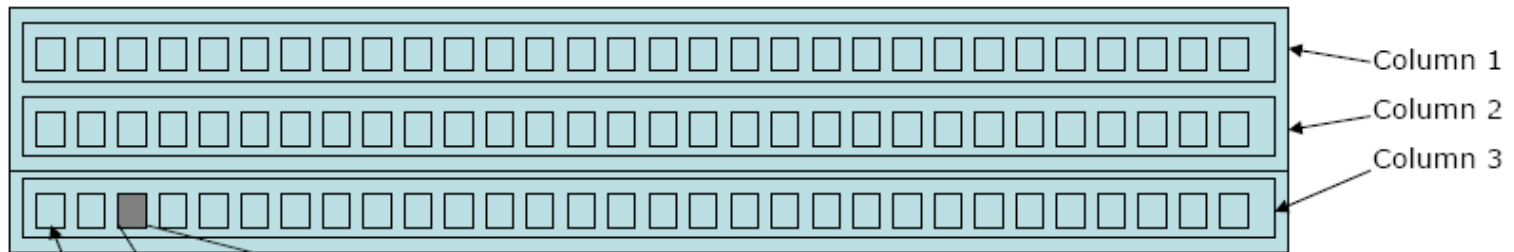




A **flow cell** contains eight lanes



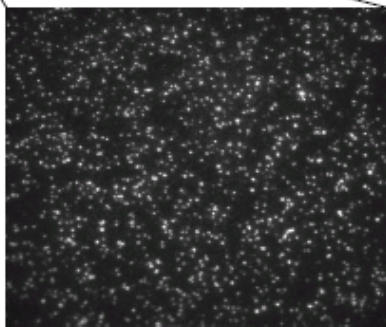
Each **lane/channel** contains **three columns** of tiles



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

Tile

20K-30K
Clusters



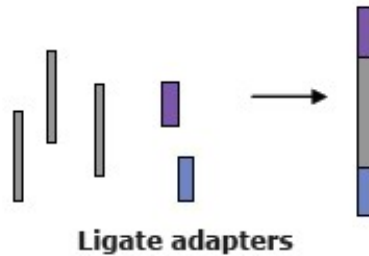
350 X 350 μm

Each tile is imaged four times per cycle – one image per base.

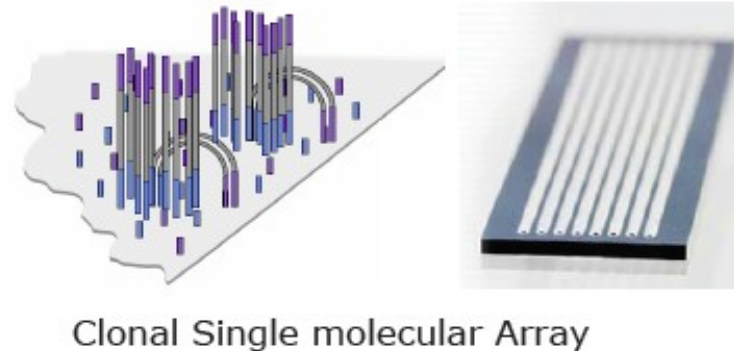
345,600 images for a 36-cycle run

Illumina Sequencing pipeline

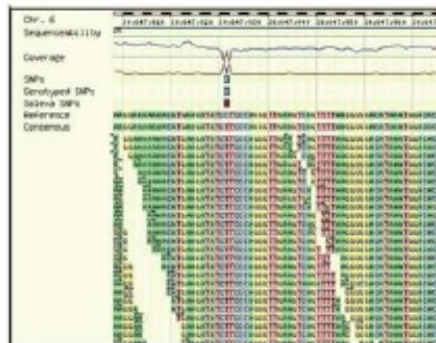
1. Sample Prep (1-5 days)



2. Cluster generation on flow cell (1.5 day)



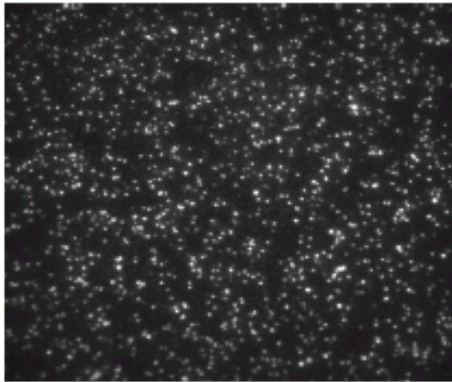
4. Data Analysis (days-months)



3. Sequencing and imaging (2-3 days)



Data Analysis Pipeline



tiff image files
(345,600)

Firecrest

1	T	130	543	140.9	347.7	739.1	24946.9	202.2	209.7	297.0	2194.4
1	T	180	421	231.9	341.9	497.7	21423.8	229.3	382.8	14319.2	20217.9
1	T	240	426	216.4	356.8	581.6	21362.3	345.5	319.7	467.9	19749.5
1	T	241	509	187.7	382.7	597.4	20747.7	3489.2	10394.1	141.0	482.7
1	T	224	285	179.5	272.1	486.5	20302.6	8297.1	12746.0	159.4	296.8
1	T	155	544	170.2	339.5	530.3	18408.9	307.6	418.8	364.9	17172.9
1	T	305	307	355.8	472.1	782.9	28449.1	3891.2	12332.1	191.9	743.0
1	T	175	456	210.4	323.8	522.3	16249.2	544.4	208.7	535.9	20587.5
1	T	240	522	287.9	533.0	456.9	15096.7	4285.6	10442.1	3394.7	2486.9
1	T	194	522	220.2	455.9	486.6	18895.6	289.5	352.8	12299.4	14131.7
1	T	237	432	147.6	457.7	521.0	16025.2	712.8	990.0	416.4	16774.0
1	T	160	526	170.4	409.7	481.9	14486.9	1249.7	4305.8	241.3	524.1
1	T	164	549	205.7	385.0	480.4	13465.5	2410.3	9408.2	76.7	243.0
1	T	179	381	207.2	372.9	560.3	10442.2	240.7	282.9	314.4	16462.8
1	T	224	423	216.3	469.4	474.4	18360.9	3321.1	10764.6	159.2	446.3
1	T	139	583	241.9	358.9	542.7	18183.9	226.9	302.0	13425.1	15107.5
1	T	220	428	225.1	486.8	553.2	15716.8	3338.8	10291.0	311.2	594.4
1	T	300	307	194.0	329.0	460.3	24628.4	294.7	590.4	403.0	16946.9
1	T	334	512	249.8	599.6	430.9	24101.4	4787.9	11274.9	602.5	177.3
1	T	150	327	216.7	349.4	536.6	17715.4	2413.2	9446.9	377.4	523.2
1	T	243	541	182.5	375.9	470.2	22603.1	4711.6	11481.7	199.5	604.9
1	T	243	458	206.4	341.2	497.0	17248.9	4290.2	9319.9	312.1	34.4
1	T	174	509	226.3	328.4	457.9	17172.1	279.5	306.5	387.3	14274.9
1	T	371	582	280.4	546.4	406.1	21245.9	4630.4	10982.2	146.3	216.1
1	T	271	608	176.8	391.5	487.5	21381.2	2832.2	11093.9	191.9	409.8
1	T	195	353	236.4	389.5	485.4	14629.3	4094.2	8305.9	289.5	3794.6
1	T	301	392	181.8	378.0	553.4	22549.7	8013.1	13222.2	899.6	1211.8
1	T	249	549	197.7	525.1	543.4	14512.2	1640.8	16851.3	171.9	504.9
1	T	140	517	108.7	388.0	508.1	14448.1	1755.8	8409.2	155.7	381.8

intensity files

Bustard

1	T	130	543	TTTGAACAGCATATTATAGCGACG
1	T	180	421	TGTTTTTTTTTTTTTTTGAGACAGS
1	T	240	426	TTTGATCATGTTTTCTGCTGCGAGG
1	T	241	509	TCTGCTGCTGCTGCTGCTGCTGCTG
1	T	214	595	TACAAAATCCCTGCCCATATGGAGCT
1	T	135	544	TTATCTGCATCCGATGCAATTTTAG
1	T	301	507	TCCTGCTTATTTGCTCTTTTJATTT
1	T	175	604	TTGGATCCGGGTAAAGGGAGAGRI
1	T	242	522	TACTAATATACAGSATAATGTTGAAA
1	T	196	522	TGTGACGGAGGGACGCGCTGACRI
1	T	237	612	TTGCTGCTGCTGCTGCTGCTGCTG
1	T	140	528	TCTGATTTTTTACAGTAAAGAGAAAC
1	T	144	543	TCTGAGAAACATGCTGATCTCCAGS
1	T	179	581	TCTGAAATCTTGCATGCTCTTTGG
1	T	224	623	TATTAGCGGCTGAGCGCTGCGCCCA
1	T	129	583	TTATGGATGGGAGCGAGGGAGGCT
1	T	220	418	TCCGAAATGTTTTAAATATAGAGGCA
1	T	340	507	TTATTTGAGATAAATGTTTTCAATTA
1	T	334	512	TTATTTGTTTCACCTAATATGGGATC
1	T	155	517	TCCCAAAAGAAAAAGAGAGAGGAG
1	T	343	541	TATTTCCATGCTGCTAATGATAGAT
1	T	241	608	TATTAGCCAGGTGAGTGGTTGACCC
1	T	174	520	TTTTTTAGTAGAGTGGGATTTACCC
1	T	371	592	TATTCCTATAGAACAGCCATAGSBS
1	T	271	508	TCTCTGGAAATATAGCTTAGCCAG
1	T	195	603	TACTGAGTGGGGCCCTGGTATCTTG
1	T	501	710	AAAAAAAAAAAAAAAAAAAAAAAAAAAA

Sequence files

Additional
Data Analysis

Alignment to Genome

Eland

Illumina fastq

```
      1           2           3           4           5           6 7           8
@HWI-ST226:253:D14WFACXX:2:1101:2743:29814 1:N:0:ATCACG
TGC GGAAGGATCATTGTGGAATTCTCGGGTGCCAAGGA ACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTT
GAAAAAAAAAAAAAAAAAATTA
+
B@CFFFFFFHFFHJIIGHIHIJJJIJIIJJGDCHIIJJJJJJGJGIHHEH@)=F@EIGHHEHFFFDCBBD:@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

Platform Updates

Solexa 1G	• 18bp reads, ~1Gbp / run
Illumina GA	• 36bp reads ~3Gbp / run
Illumina GAI	• 75bp paired reads ~10Gbp / run (8 days)
Illumina GAIx	• 75bp paired reads ~40Gbp / run (8 days)
Illumina HiSeq 2000	• 100 bp paired reads ~200 Gbp/ run (10 days)
Illumina HiSeq, v3 SBS	• 100bp paired reads ~600Gbp / run (12 days)
MiSeq	• 150 paired reads ~1.5 Gb/run (27 hrs)

Maximum yield / day 50,Gbp
~16x the human genome

300bp paired reads

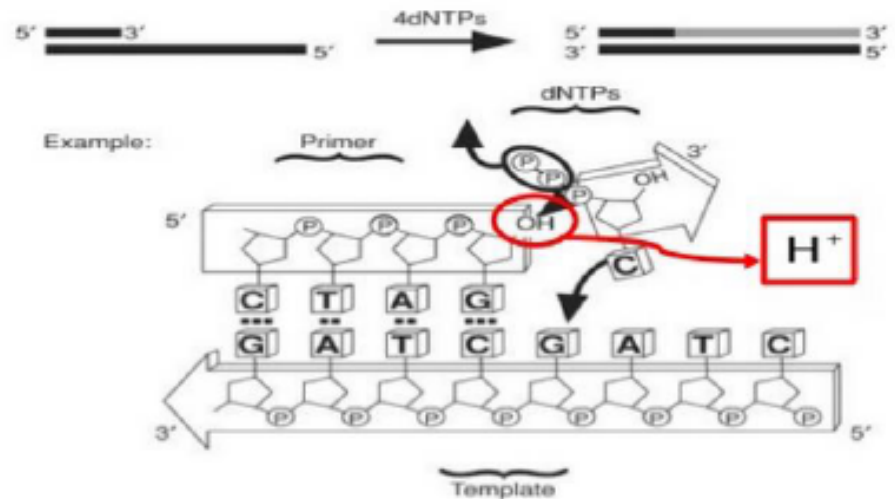


Applied Biosystems: Ion Torrent PGM

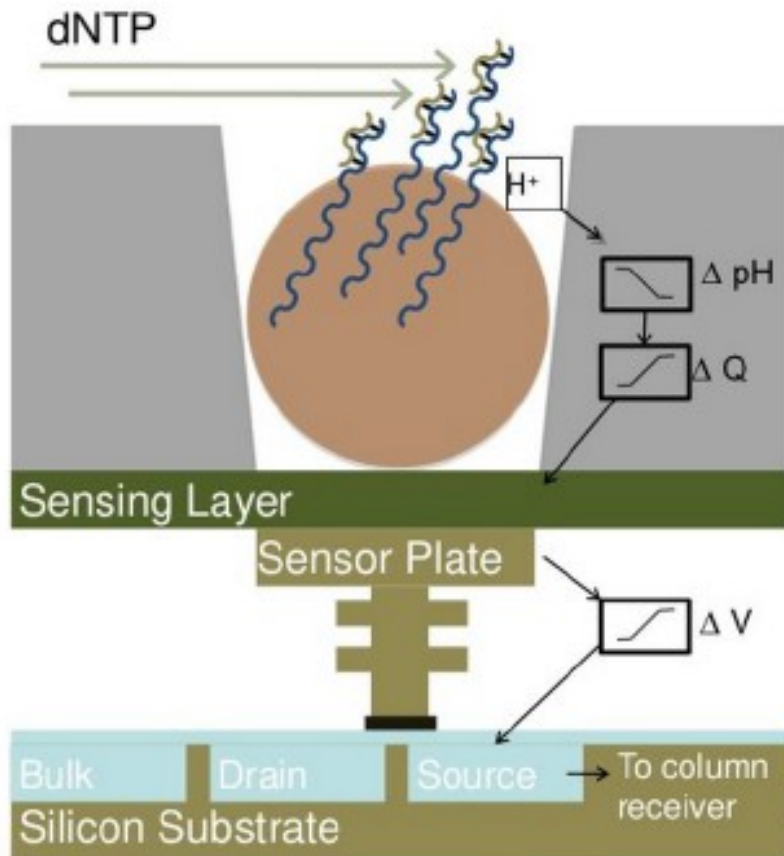


Využívá změny pH při syntéze DNA

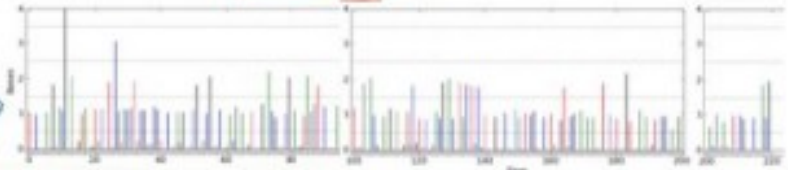
- Ion Semiconductor Sequencing
- Detection of hydrogen ions during the polymerization DNA
- Sequencing occurs in microwells with ion sensors
- No modified nucleotides
- No optics



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



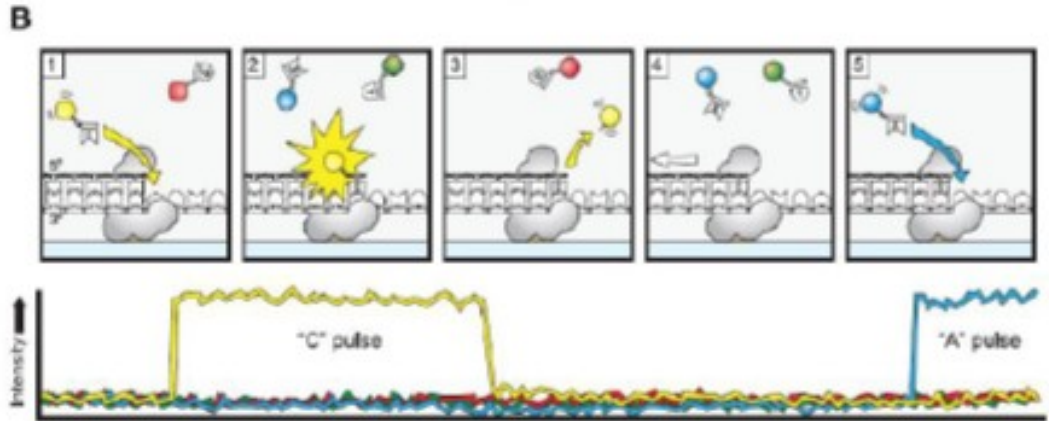
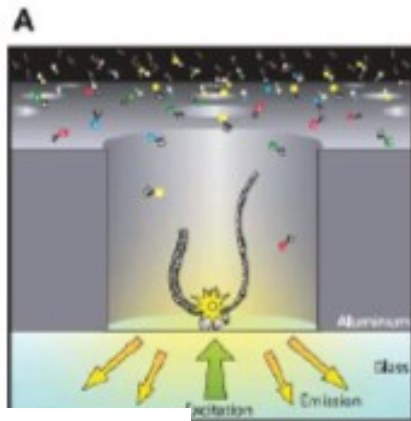
SOLiD

(sequencing by Oligonucleotide Ligation and Detection)

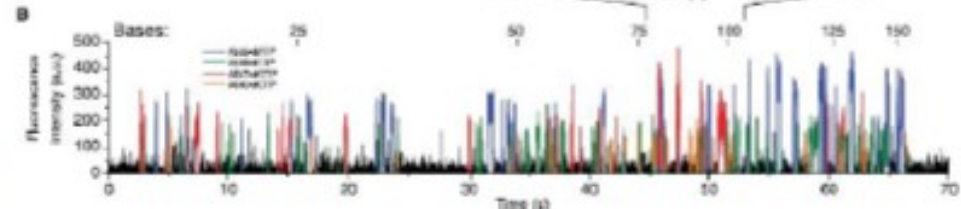
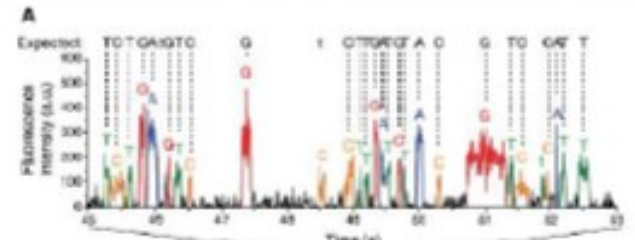


velice přesný, ale jen krátké „reads“

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

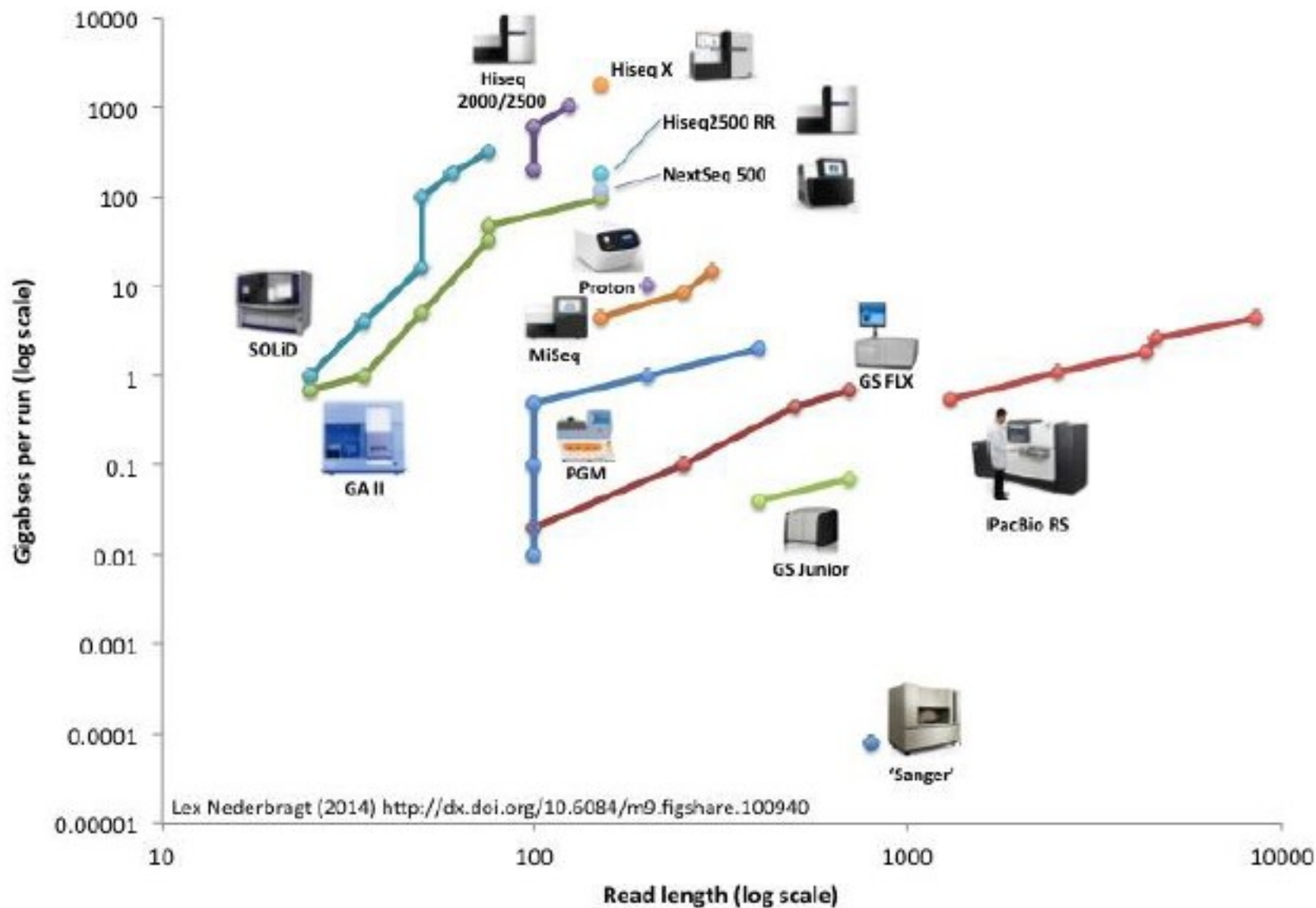


Pacbio RS – raw data



dlouhé čtení (15 kb), hodně chyb

Developments in High Throughput Sequencing



Lex Nederbragt (2014) <http://dx.doi.org/10.6084/m9.figshare.100940>

3rd generation: Oxford Nanopore



MinION
512 pores

GridION
5 000 pores

Future Sequencing Technologies

Oxford Nanopore

Nanopore sequencing

up to 50 kb

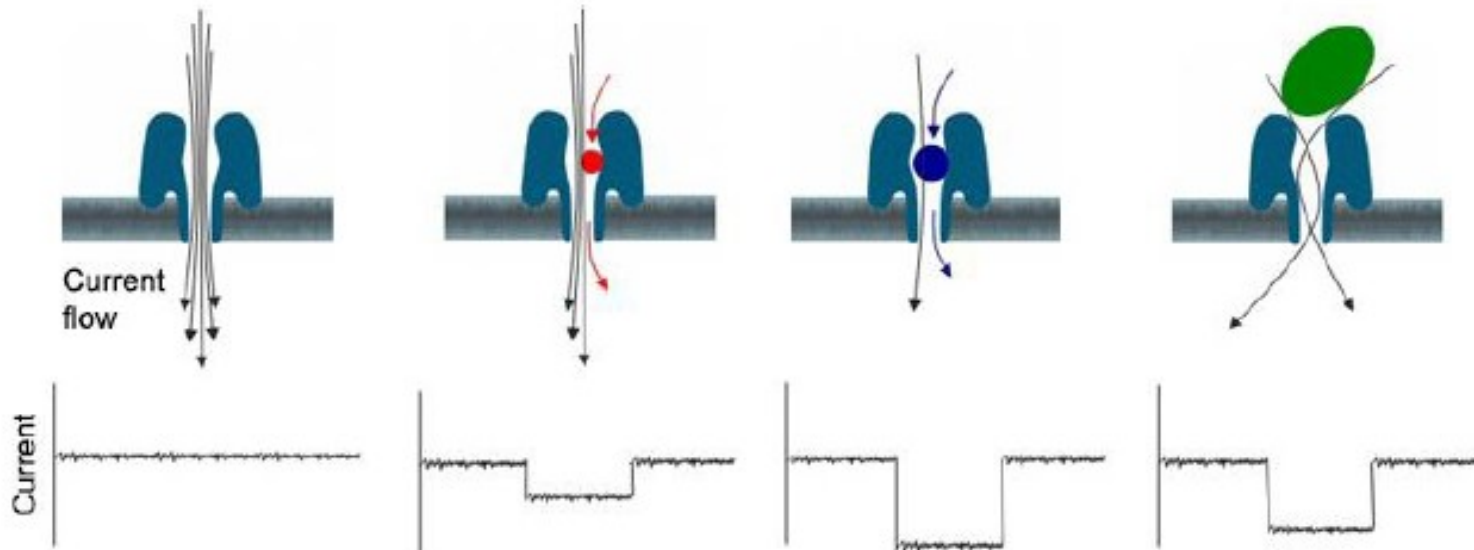
high per-base error rate (>5%)

?? throughput

?? cost (Minlon, Gridlon)



Princip technologie



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



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

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

Traditional Sequencing vs. Next Generation Sequencing: Data Throughput

1 x Illumina GAI



200+ of 3730xl



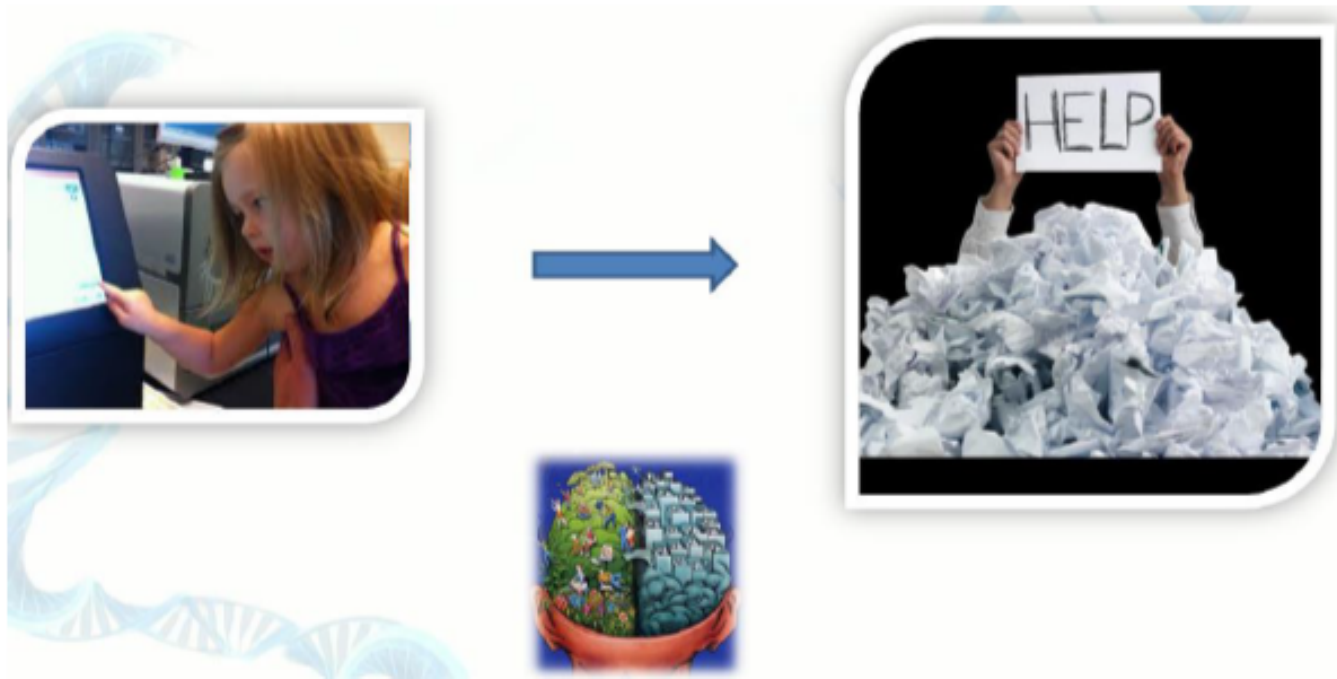
Vs.

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



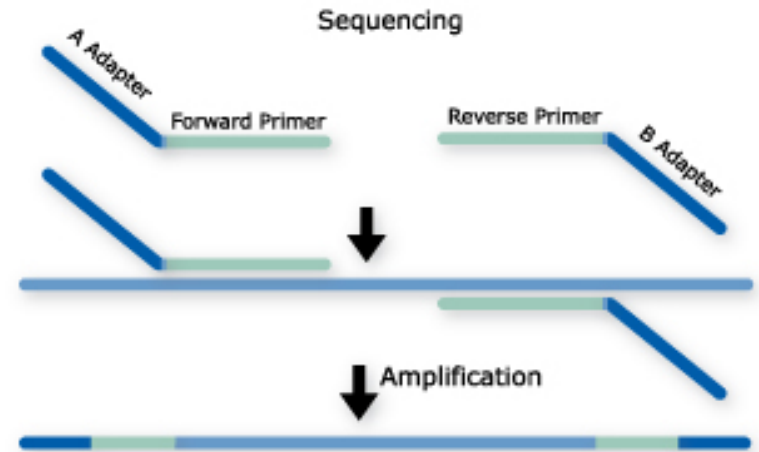
Sekvenční strategie

AMPLIKONOVÉ SEKVENOVÁNÍ

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

Následná sekvenace

Taxonomické složení daného vzorku, sekvenční variability konkrétních genů apod.



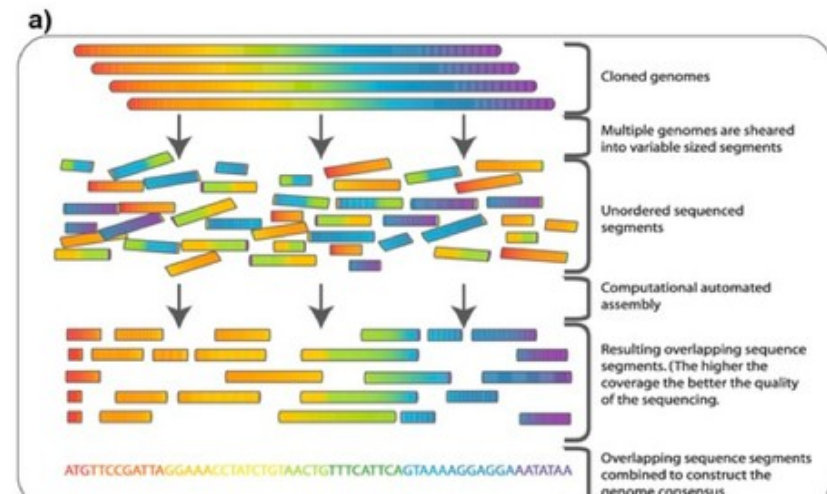
SHOT GUN SEKVENOVÁNÍ

Fragmentace gelogenomové DNA

Ligace sekvenčních adaptorů

Následná sekvenace náhodných fragmentů

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



TO NENÍ VŠECHNO.....

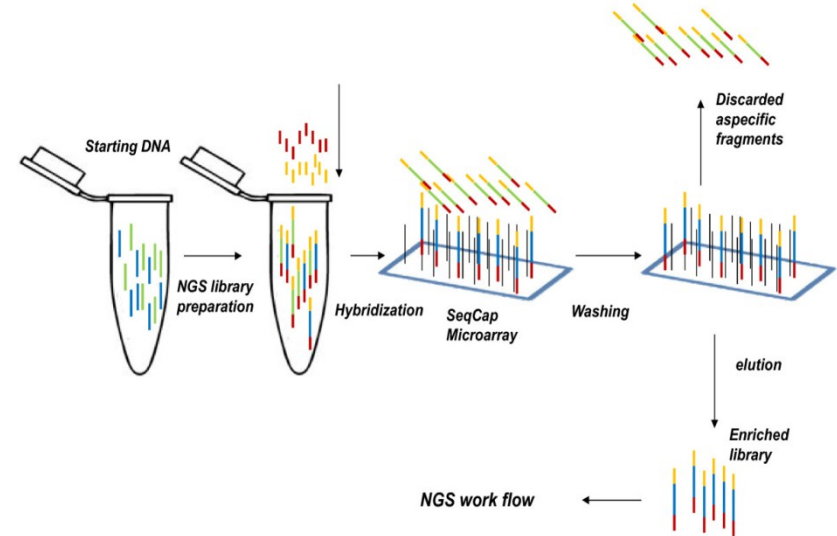
Sekvenční strategie

Sequence capture + shot gun

Separace úseků genomu které nás zajímají na základě jejich hybridizace

Následná sekvenace obohacených knihoven

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



NimbleGen

Long range PCR + shot gun

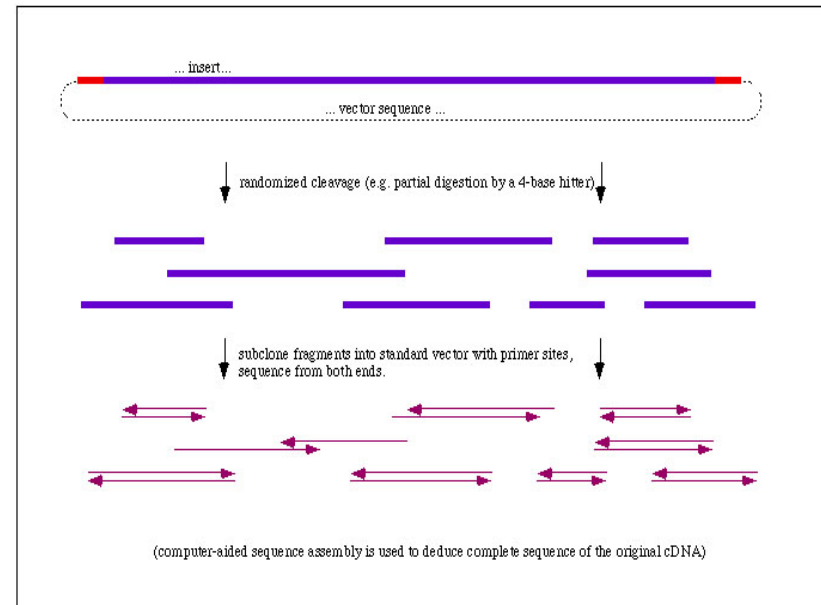
Dlouhé PCR produkty, které nejdou vcelku osekvenovat

Jejich fragmentace

Sekvenování fragmetů

Zpětná rekonstrukce původní sekvence („assembly“)

Použitelné pokud nás zajímá variabilita v jednolitém úseku DNA. Např. sekvenace mitochondriální DNA nebo jejich částí.



Sekvenční strategie

Sekvenování podél restričních míst

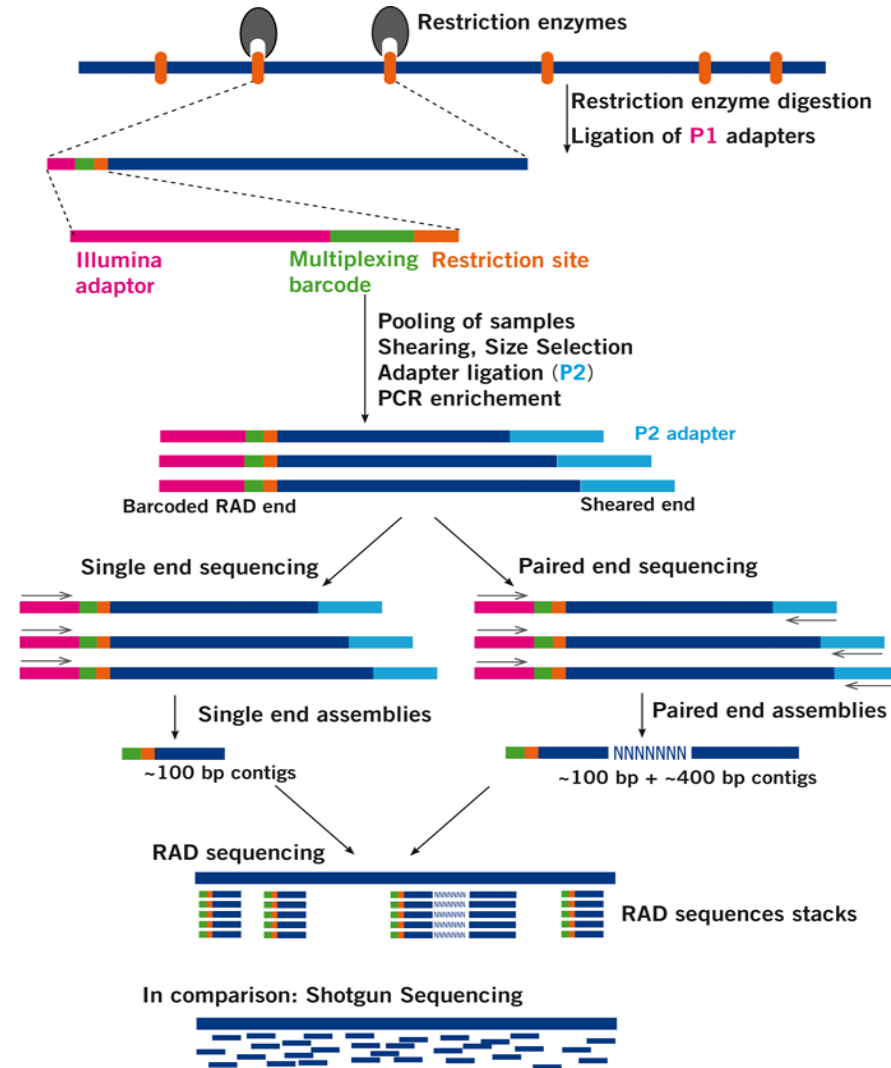
Fragmetace gelogenomové DNA po mocí restričních enzymů

Ligace sekvenačních adaptorů na výsledné fragmenty

Následná sekvenace podél restričních míst

Celogenomové scany genetické variability

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

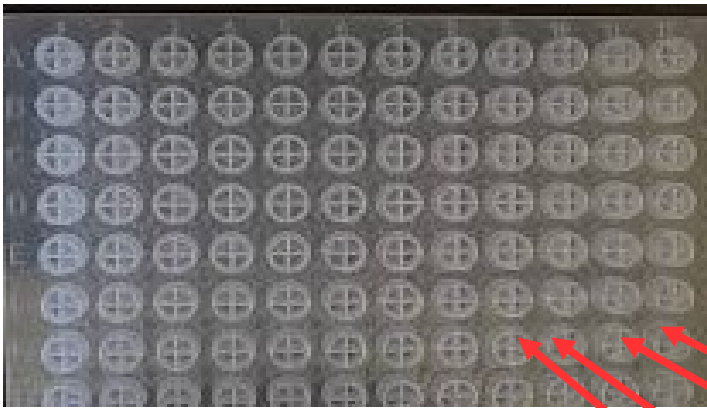


Sekvenční strategie

...JEDEN VZOREK NA RUN JE MÁLO

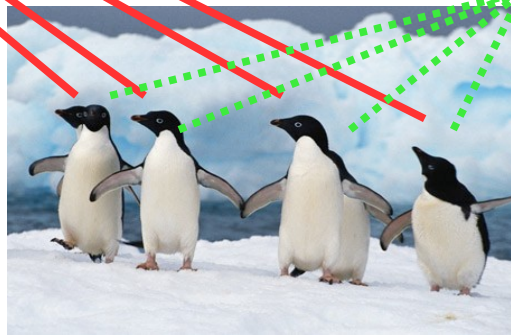
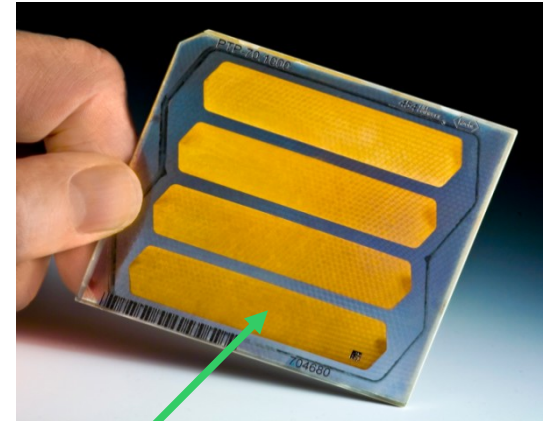
Kapilární sekvenátor

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



Sekvenční strategie

...JEDEN VZOREK NA RUN JE MÁLO

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

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

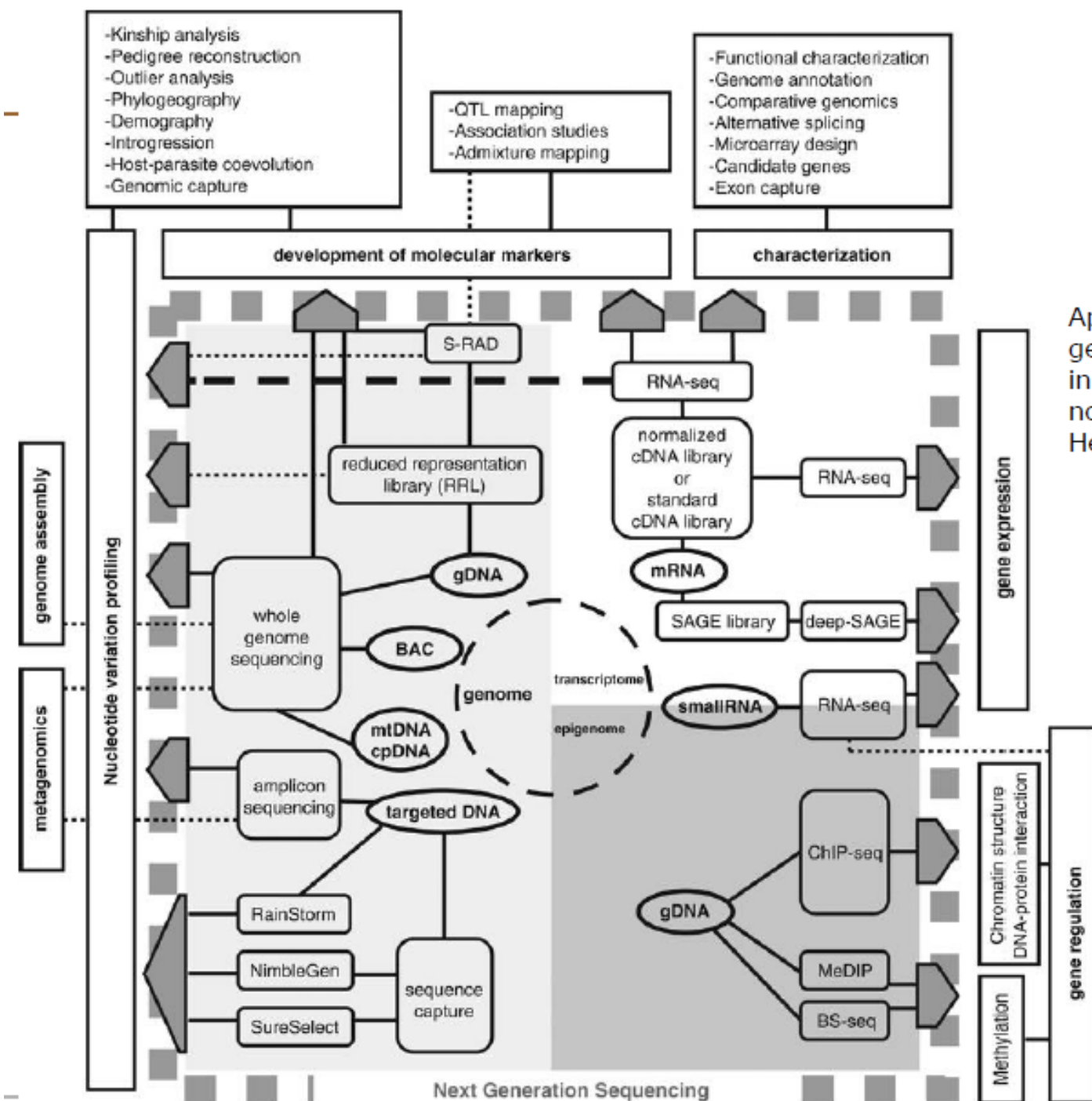
Přiřazení identity jednotlivých sekvencí k vzorkům probíhá bioinformaticky

BARCODE

PRIMER

SEQUENCE

```
AGC GTA GGT CATTTCGATGCGGTCATGCCTGGATTAAAGCT.....  
TTCGTA GGT CATTTCGATGCGGTCATGCCTGGATTAAAGCT.....  
TGGGTA GGT CATTTCGATGCGGTCATGCCTGGATTAAAGCT.....  
TGCCTA GGT CATTTCGATGCGGTCATGCCTGGATTAAAGCT.....  
TGCGCA GGT CATTTCGATGCGGTCATGCCTGGATTAAAGCT.....  
TGCGTI GGT CATTTCGATGCGGTCATGCCTGGATTAAAGCT.....
```



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

Aplikace

1. Celogenomové sekvenování de novo
2. Celogenomové resekvenování
3. Sekvenování amplikonů (PCR produktů)
4. Další aplikace - např. hledání klasických DNA markerů (mikrosatelity, SNPs)

1. Celogenomové sekvenování de novo

Problém: **KRÁTKÝ READ LENGTH**

- **400bp** 454 FLX Roche (dnes i Illumina)

- **35-75bp** Solid

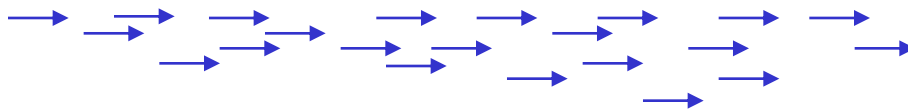
vs **800-1000bp** Sanger



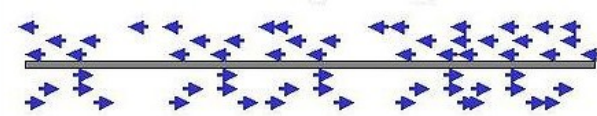
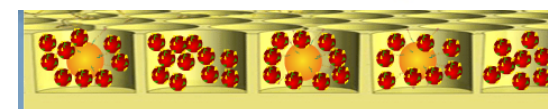
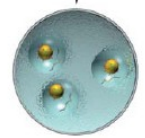
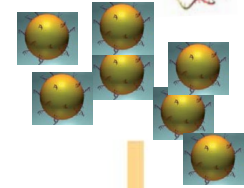
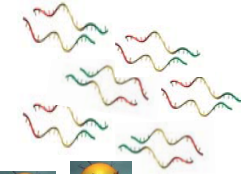
→ Uspořádání (assembly) už není problém z hlediska výpočetní kapacity

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

GTAAAAAAAAAAAAAAAAAAAAAAC



Zvláště komplexní eukaryotické genomy - úseky souvislých oblastí přerušovaných mezerami



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

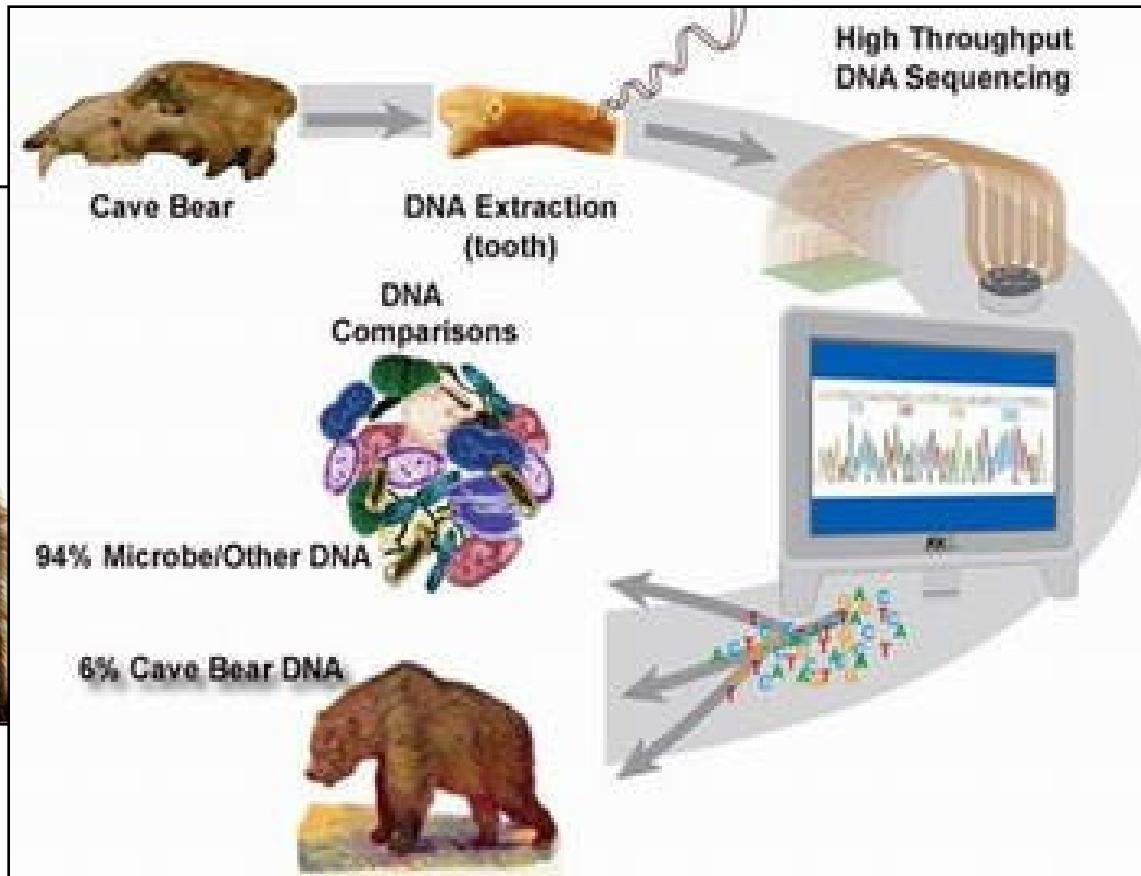
Cell

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 Märten Wikström,⁹ Liisa Laakkonen,¹⁰ Janet Kelso,¹ Montgomery Slatkin,² and Svante Pääbo¹

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 ampliconů (PCR produktů)

SMĚSNÉ VZORKY - paralelní sekvenování nahrazuje klonování

Metagenomika

- Celé společenstvo půdních, vodních mikroorganismů, střevní mikroflóra - **mikrobiom**
- PCR genu 16S (18S) rRNA
- lze i kvantifikovat

Metabarcoding

- 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 environmentální DNA
- Amplifikace pomocí primerů specifických pro cílovou skupinu, pokrývající taxonomicky informativní úsek (COI, 16s/18s RNA...)
- Paralelní 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:

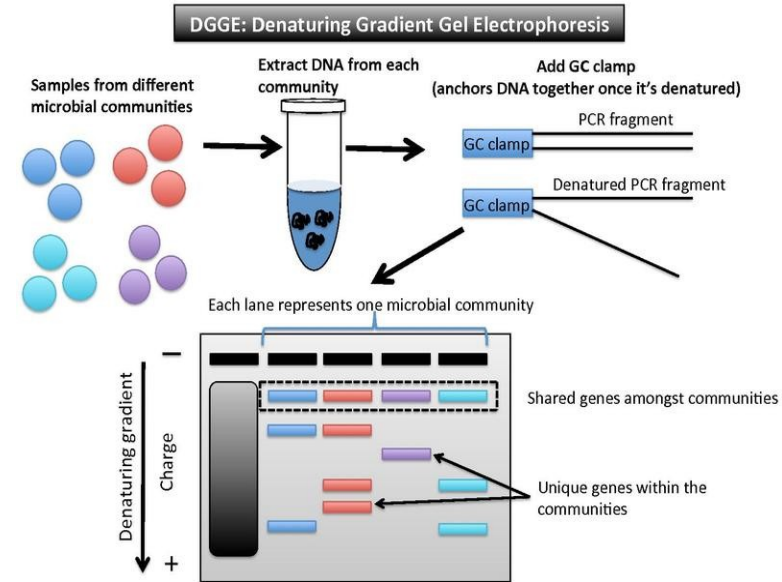
Klonování amplikonů a sekvenování klonů
Specifické elektroforézy - např. DGGE

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?



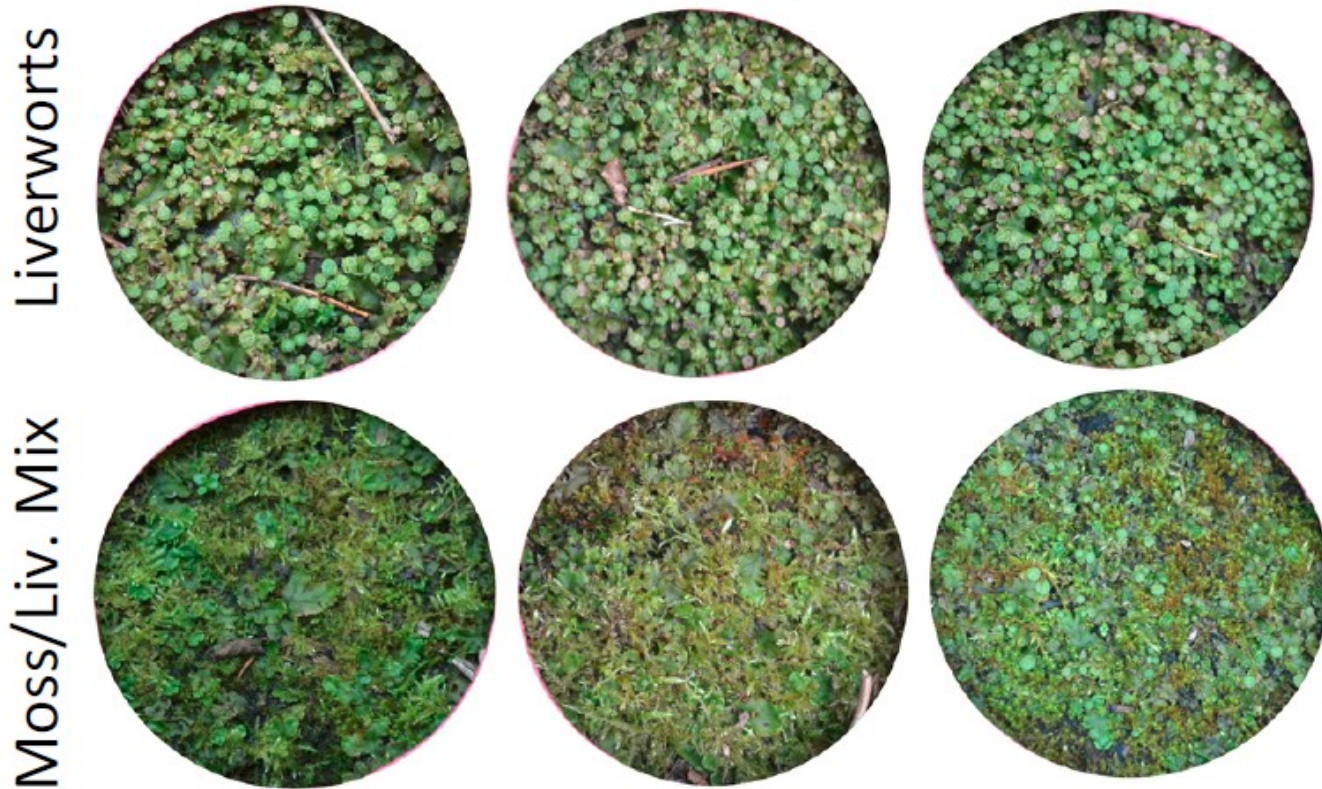
Metabarcoding - příklady využití

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



Společenstvo eukaryot ve vrchní vrstvě půdy

Metabarcoding - příklady využití



Společenstvo eukaryot ve vrchní vrstvě půdy

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/mm²

2% phiX174 spike-in

17 million raw pass filter pairs

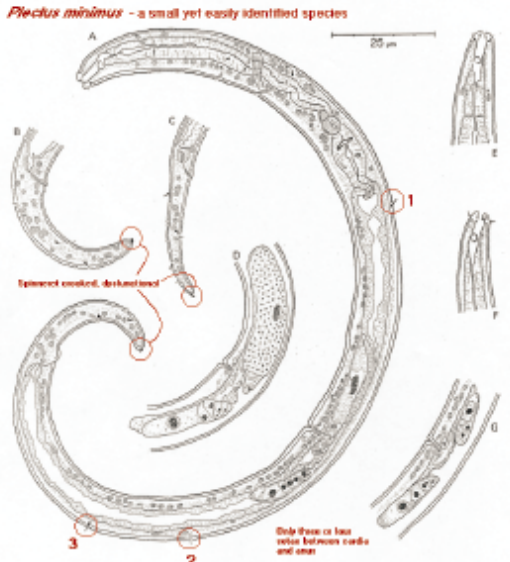
Metabarcoding - příklady využití

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	.00210858862609
Apicomplexa	74	.0016743479503615
Flabellinea	53	.0013524759319671
Dinophyceae	139	.000952196733392
Bicosoecia	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
Hypochytridiomycetes	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	.0001799408169568
Schizopyrenida	16	.0001793457031658
Environirenal_samples	8	.00016573196745086
Trinastix_pyriformis_atcc50562	5	.0001548375140474
Labyrinthella	16	.0001521436503891
Gastrotricha	1	.0001415996975531
Telonema	26	.0001291107737918
Pollicipes	26	.0001268072826276
Ascomoradidae	3	.000125997814932
SoL_amoeba_and16	11	.00012478493141
Acanthocystidae	4	.00012297142125379
Voronomas	4	.00012291390987468
Leukarichnion_sp_atcc_pra-24	1	.0001206455234986
Rodifera	10	.000120231962562993
Ancyromonadidae	3	.00010991011460599
Peritoma	3	.00010446009103284
Nyctozoa	3	.0001044492057314
Ecorticaceae	1	.00010038940919404
Glomeromycota	2	.00010030472735523
Eukaryote_marine_clone_mel-24	1	.00010023004551642
Fungal_sp_fca90	1	.00010021087505672
Cryptomonadales	1	.00010011502275821
Phaeobryozoa	2	.00010011502275821
Phaeocharniphyceae	1	.00010009585229851
Chordata	1	.0001000380409194
Unclassified_alveolata	1	.0001000191704597

Eight animal phyla represented

Most frequent are **Nematoda**

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



Metabarcoding - příklady využití

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

Výrazně větší 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 Wilkinson³,
Morten Rasmussen²,
Lars R.D. Jensen¹, Eske Willerslev²
Mads F. Bertelsen¹,
and M. Thomas P. Gilbert^{2,*}

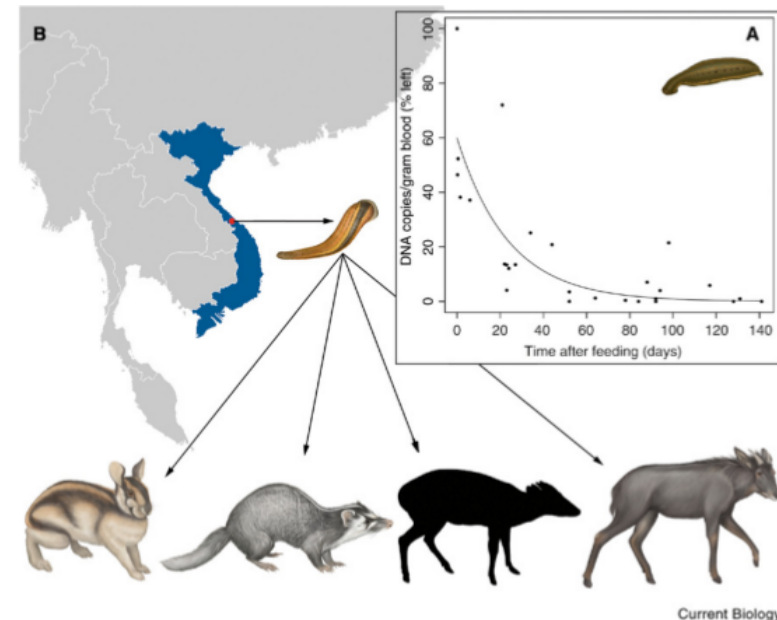
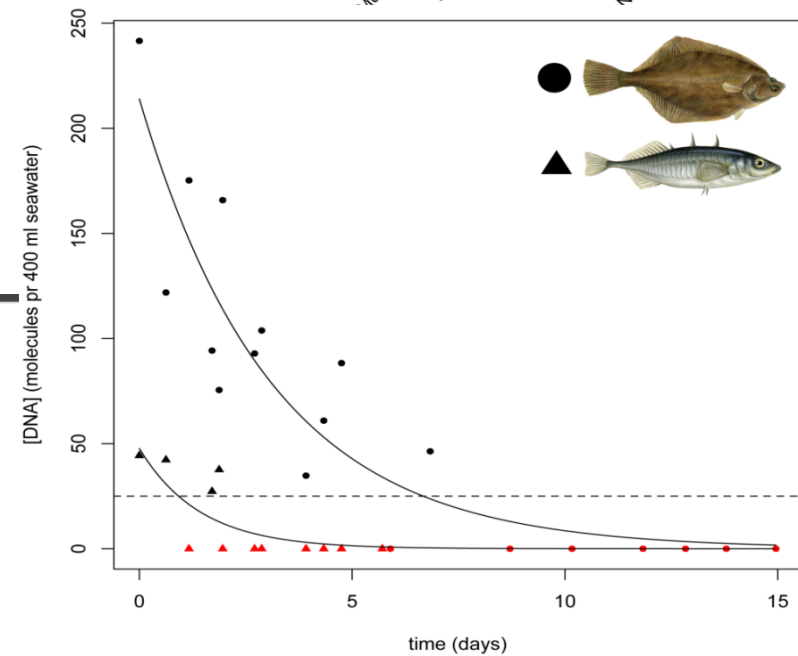
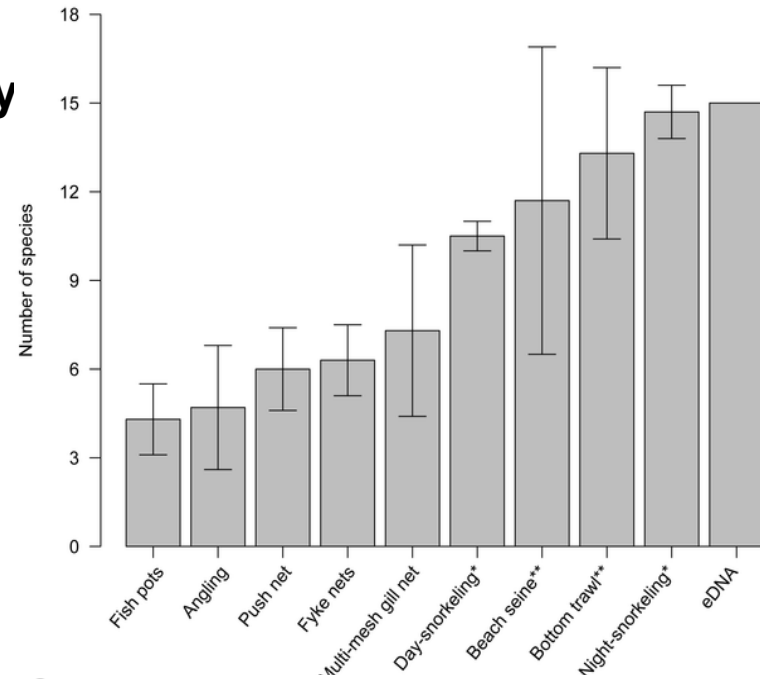
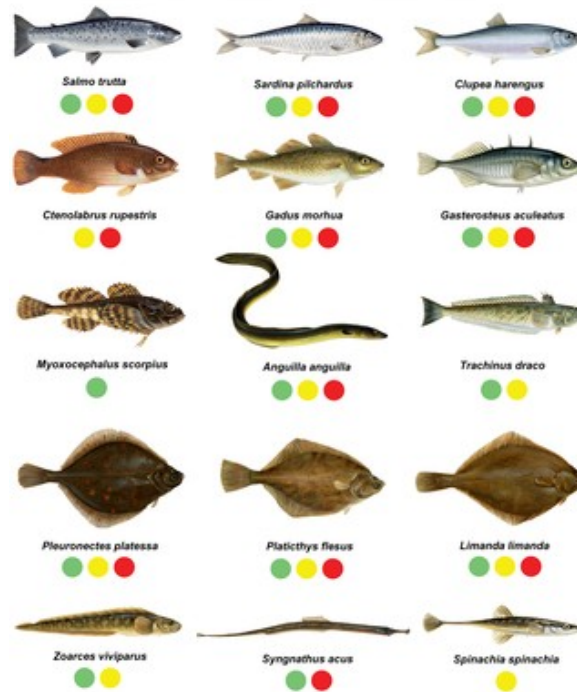


Figure 1. Monitoring mammals with leeches. (A) Survival of mtDNA in goat blood ingested by *Hirudo medicinalis* over time, relative to freshly drawn sample (100%, ca. 2.4E+09 mtDNA copies/gram blood). Mitochondrial DNA remained detectable in all fed leeches, with a minimum observed level at 1.6E+04 mtDNA/gram 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 *Hae madipsa* spp. leeches. From left to right: Annamite striped rabbit, small-toothed ferret-badger Truong Son muntjac (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

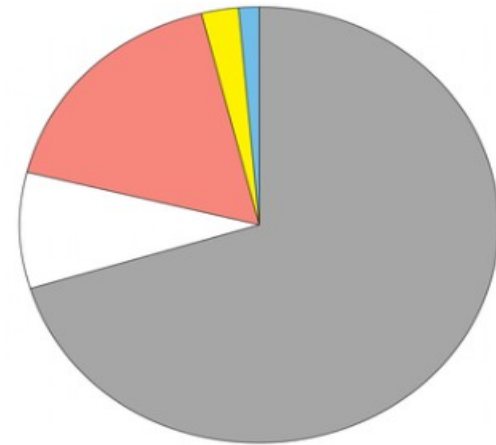
Philip Francis Thomsen^{1*}, Jos Kielgast^{1,3}, Lars Lønsmann Iversen², Peter Rask Møller³, Morten Rasmussen¹, Eske Willerslev^{1*}

¹Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade, Copenhagen, Denmark, ²Freshwater Biology Section, Department of Biology, University of Copenhagen, Helsingørgade, Hillerød, Denmark, ³Vertebrate 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 Shehzad¹, Thomas Michael McCarthy², Francois Pompanon¹, Lkhagvajav Purevjav³, Eric Coissac¹, Tiayyba Riaz¹, Pierre Taberlet^{1*}

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

Siberian ibex
(*Capra sibirica*)

Domestic sheep
(*Ovis aries*)

Argali sheep
(*Ovis ammon*)

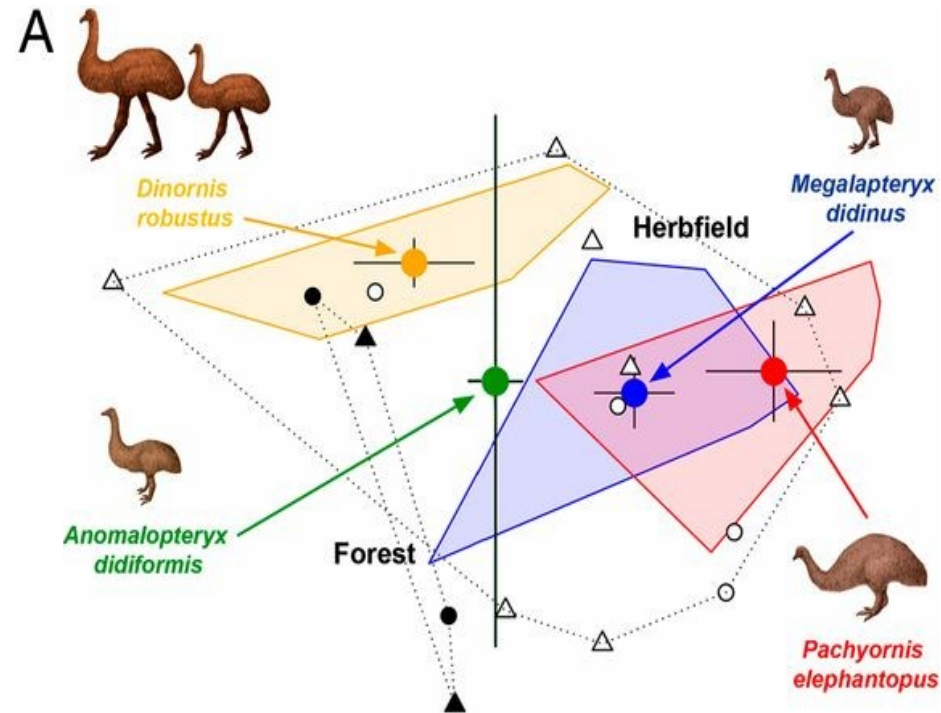
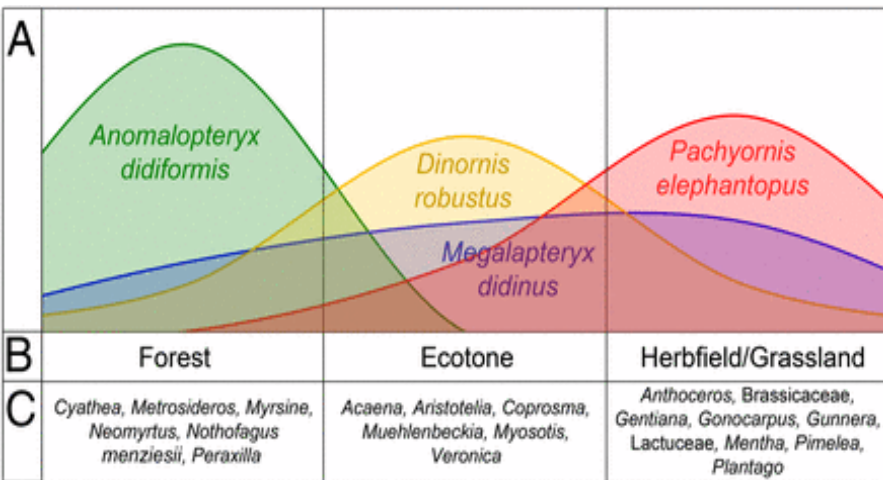
Chukar partridge
(*Alectoris chukar*)

Domestic goat
(*Capra hircus*)

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



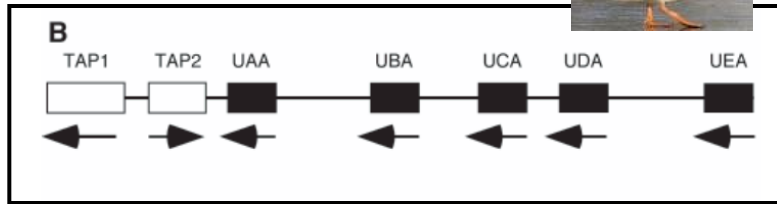
Resolving lost herbivore community structure using coprolites of four sympatric moa species (Aves: Dinornithiformes)

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^{a,3}, and Alan Cooper^b

^aLandcare Research, Lincoln, Canterbury 7640, New Zealand; ^bAustralian Centre for Ancient DNA, University of Adelaide, Adelaide, SA 5005, Australia;

3. Sekvenování amplikonů (PCR produktů)

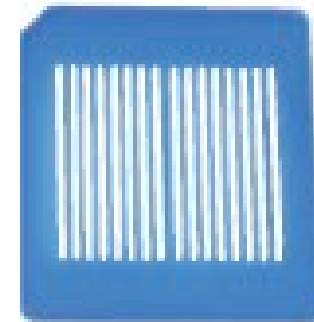
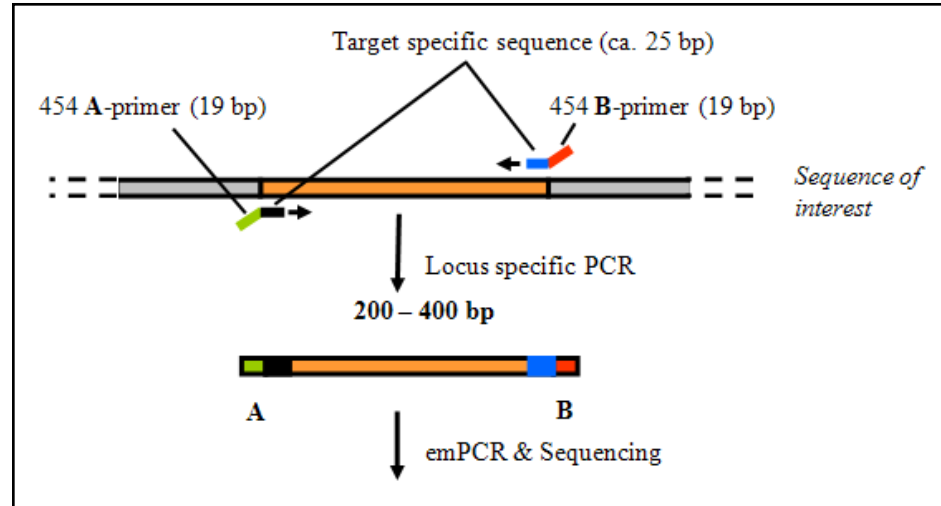
Genové duplikace



Označí jedince

Amplifikuje všechny kopie MHC genů

Potřeba k emPCR, sekvenování..

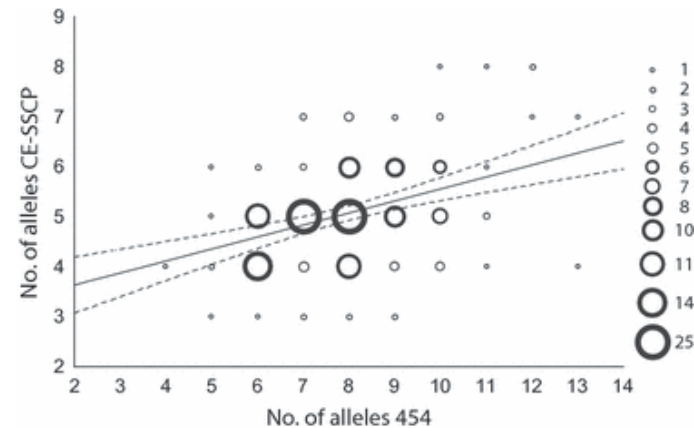
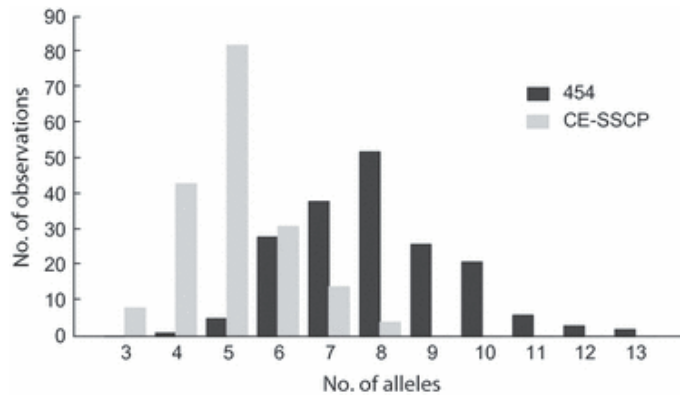


192 jedinců u 454 pyrosekvenování

Amplikonové sekvenování

MHC u hýla rudého

- 454 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 RADWAŃŚ

4. Další aplikace - hledání nových genetických markerů

Mikrosatelity

- sekvenování obohacených knihoven

SNPs

- např. RAD-sequencing

Hledání nových genetických markerů - mikrosatelity

Obvyklý postup:

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

MOLECULAR ECOLOGY RESOURCES

Molecular Ecology Resources (2011) 11, 638–644

doi: 10.1111/j.1755-0998.2011.0295

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-ESCARDANT,¶ PIERRE-JEAN G. MALÉ,§§ STÉPHANIE FERREIRA‡ and JEAN-FRANÇOIS MARTIN††

*INRA, UMR 1301 IBSV INRA/INSA/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 31 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 Montpellier-sur-Lez Cedex, France, ‡‡INRA – UMR 1248 AGIR, BP 52627, 31326 Castanet-Tolosan Cedex, France §§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 high mutation rates, and therefore are frequently highly polymorphic. Variations in the number of repetitions generate different alleles. This makes them 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. Genomic DNA is used to generate genomic libraries. We usually enrich these libraries with 4 to 6 different microsatellite motifs. However, we can customise the number of motifs to your needs. We obtain thousands of microsatellite-containing reads by using high-throughput sequencing. Our bioinformaticians then filter 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 isolate DNA from your samples. These samples should be adequately preserved to ensure DNA integrity. We will deliver tested primer pairs which amplify 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 need.

OPEN ACCESS Freely available online

PLOS ONE

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

Stacey L. Lance¹, Cara N. Love¹, Schyler O. Nunziata¹, Jason R. O'Bryhim¹, David E. Scott¹, R. Wesley Flynn¹, Kenneth L. Jones²

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

Hledání nových SNPs - RAD-sequencing

Sekvenování podél restričních míst

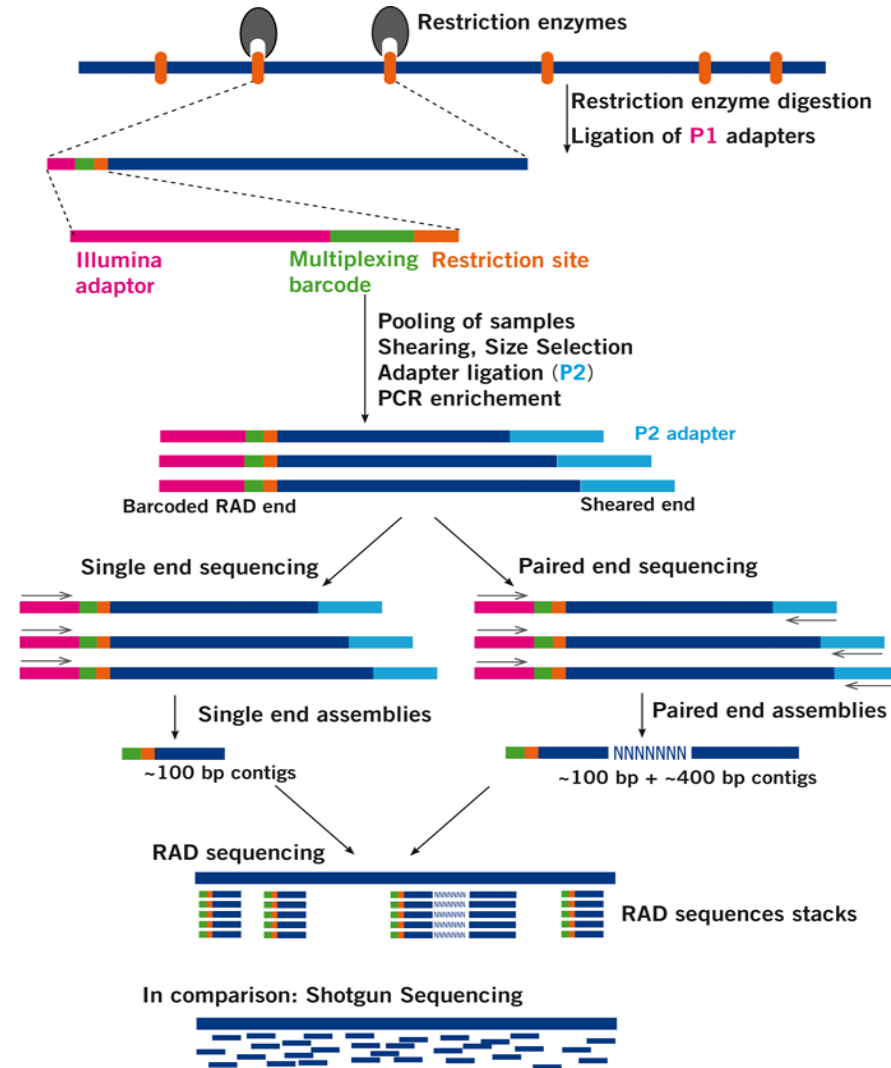
Fragmetace gelogenomové DNA po mocí restričních enzymů

Ligace sekvenačních adaptorů na výsledné fragmenty

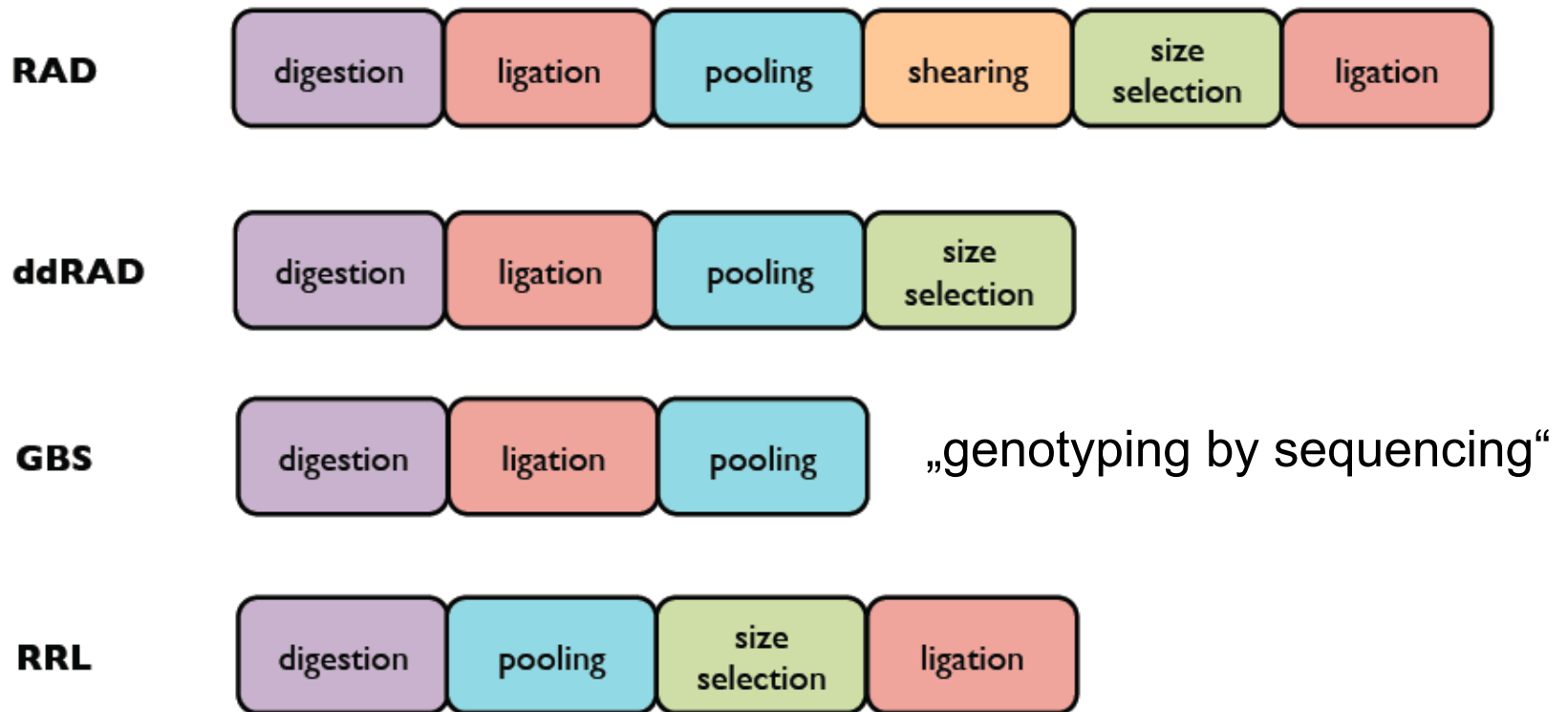
Následná sekvenace podél restričních míst

Celogenomové scany genetické variability

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



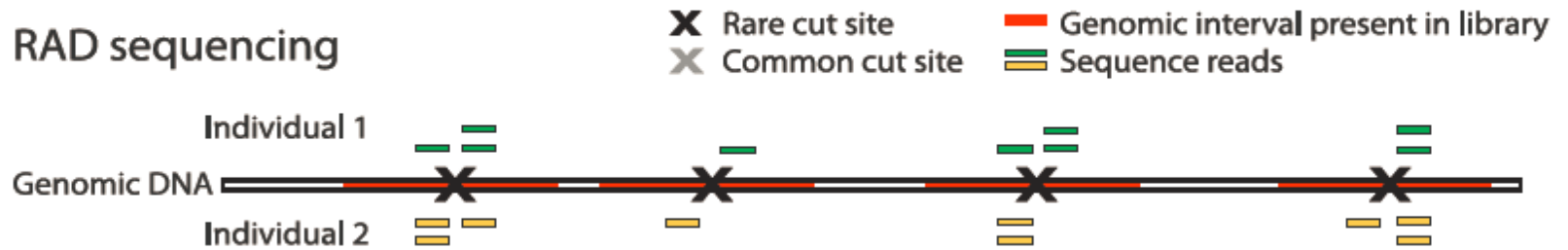
Sekvenování podél restričních míst



RAD vs. ddRAD

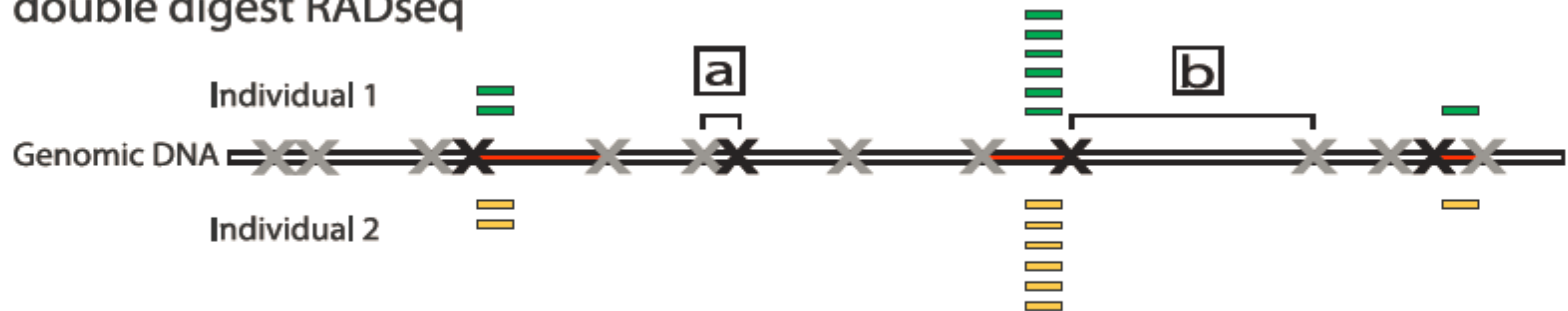
A

RAD sequencing

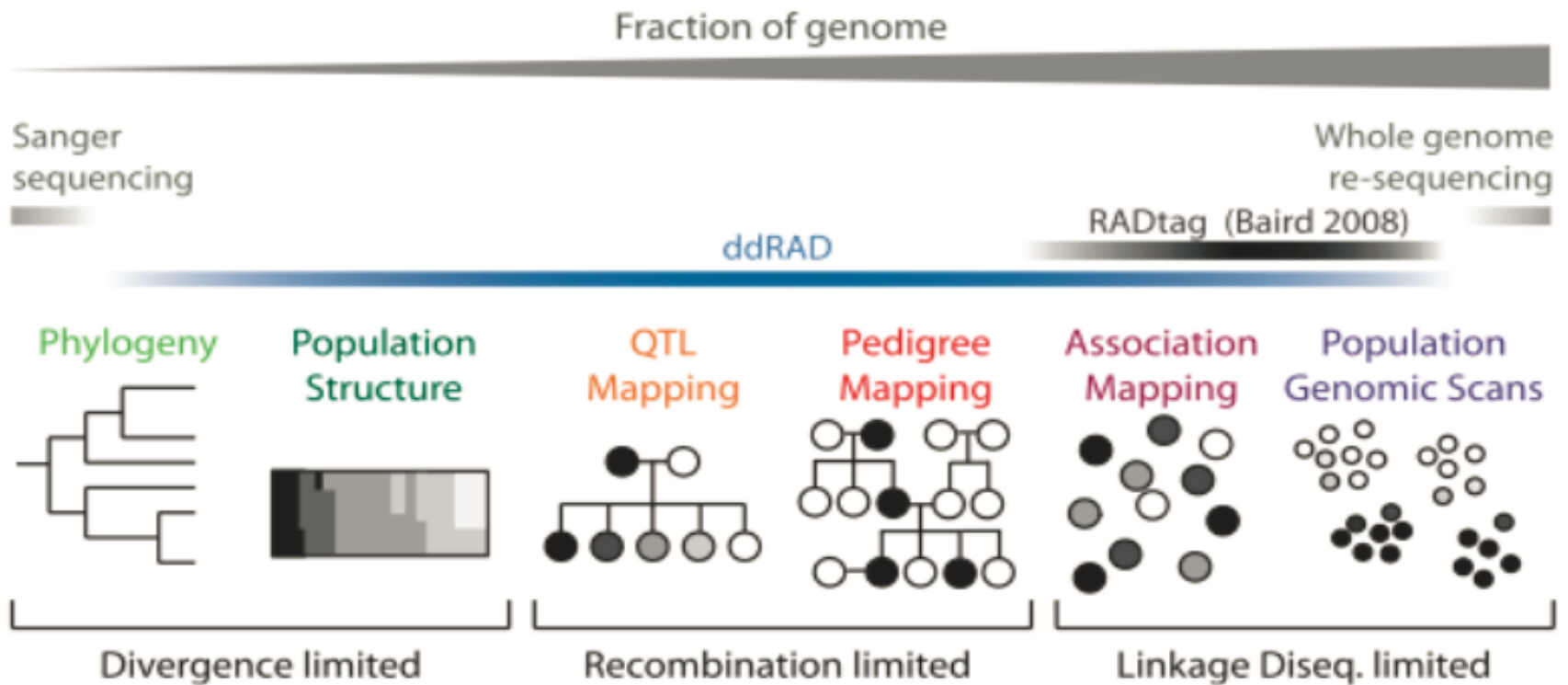


B


double digest RADseq



Sekvenování podél restričních míst



Review a příklady

 STUDY DESIGNS

Genome-wide genetic marker discovery and genotyping using next-generation sequencing

John W. Davey*, Paul A. Hohenlohe†, Paul D. Etter§, Jason Q. Boone||, Julian M. Catchen† and Mark L. Blaxter**



RESEARCH ARTICLE

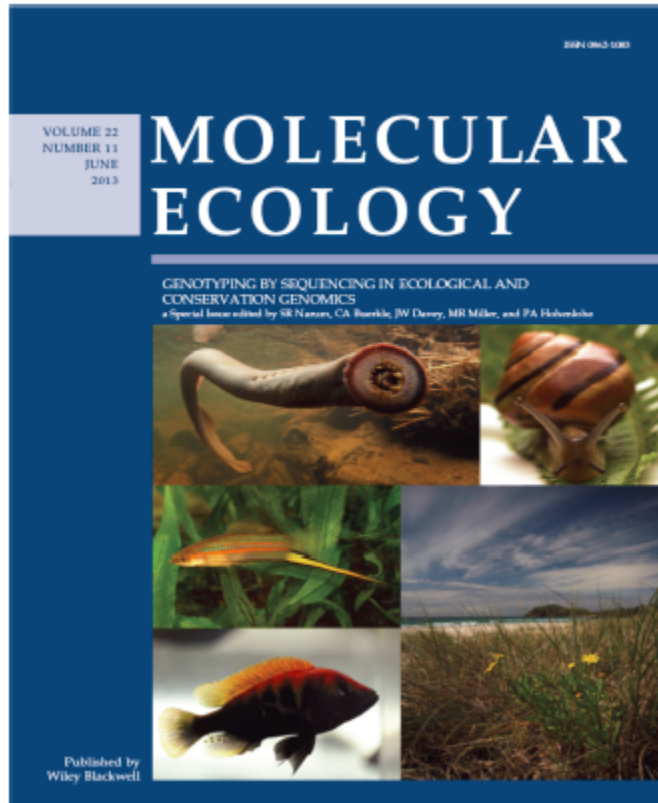
Open Access

Genome wide SNP discovery, analysis and evaluation in mallard (*Anas platyrhynchos*)

Robert HS Kraus^{1††}, Hindrik HD Kerstens^{2†}, Pim Van Hooft¹, Richard PMA Crooijmans², Jan J Van Der Poel², Johan Elmerg³, Alain Vignal⁴, Yinhua Huang⁵, Ning Li⁵, Herbert HT Prins¹, Martien AM Groenen²



Review a příklady



MOLECULAR ECOLOGY

Molecular Ecology (2012)

doi: 10.1111/mec.12084

Special features of RAD Sequencing data: implications for genotyping

JOHN W. DAVEY,* TIMOTHÉE CEZARD,† PABLO FUENTES-UTRILLA,* CATHLENE ELAND,† KARIM GHARBI† and MARK L. BLAXTER*†

*Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, West Mains Road, Edinburgh, EH9 3JT, UK, †The GenePool, Ashworth Laboratories, University of Edinburgh, West Mains Road, Edinburgh, EH9 3JT, UK

Abstract

Restriction site-associated DNA Sequencing (RAD-Seq) is an economical and efficient method for SNP discovery and genotyping. As with other sequencing-by-synthesis methods, RAD-Seq produces stochastic count data and requires sensitive analysis to develop or genotype markers accurately. We show that there are several sources of bias specific to RAD-Seq that are not explicitly addressed by current genotyping tools, namely restriction fragment bias, restriction site heterozygosity and PCR GC content bias. We explore the performance of existing analysis tools given these biases and discuss approaches to limiting or handling biases in RAD-Seq data. While these biases need to be taken seriously, we believe RAD loci affected by them can be excluded or processed with relative ease in most cases and that most RAD loci will be accurately genotyped by existing tools.

Keywords: contig assembly, genotyping by sequencing, population genetics, RAD Sequencing, restriction enzymes

Received 29 June 2012; revision received 7 September 2012; accepted 12 September 2012

Příště: Analýza genové exprese