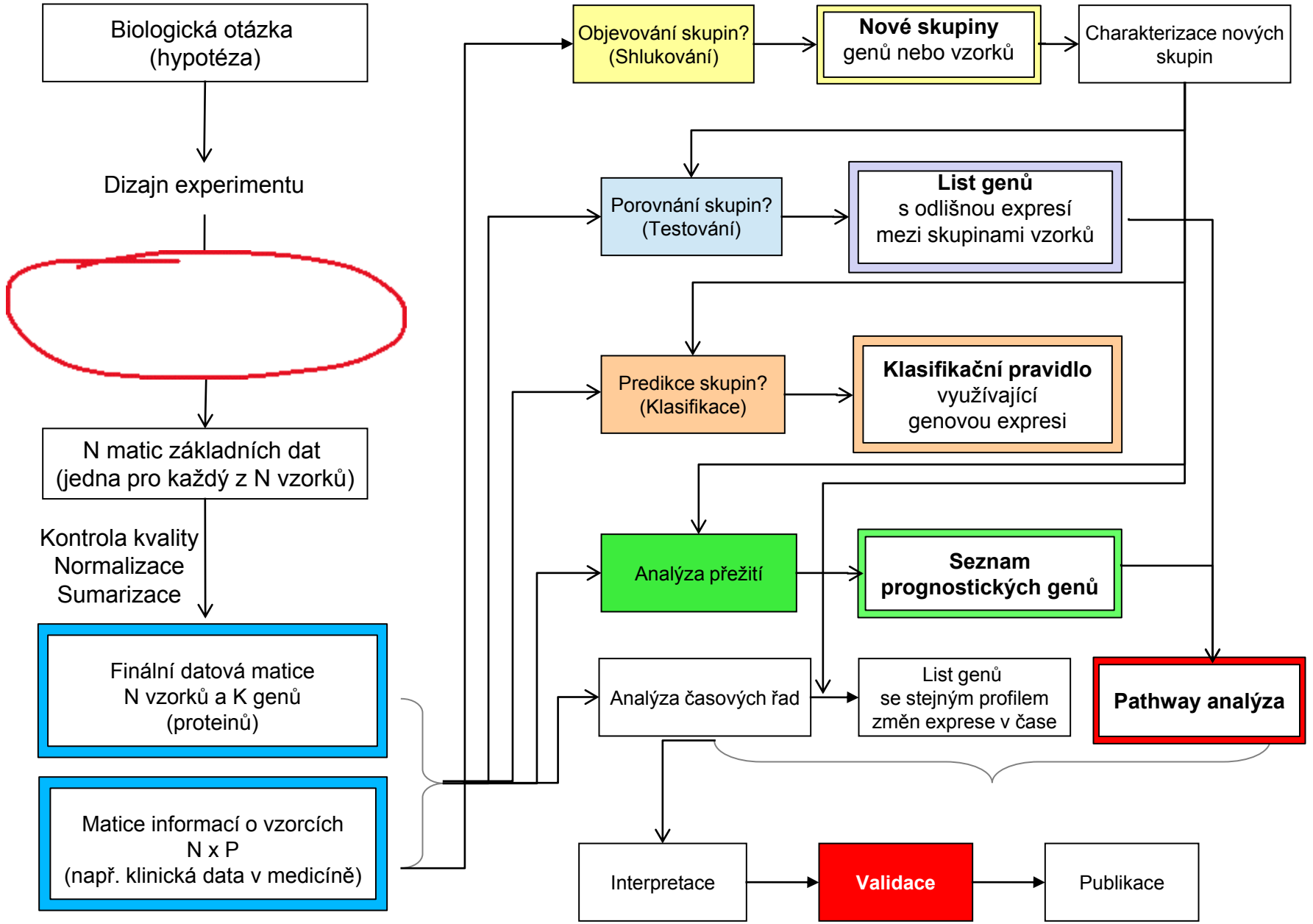
Detekce biomarkerů z omics experimentů

- Mgr. Eva Budinská, PhD
- RECETOX
- eva. budinska@recetox.muni.cz

- Podzim 2023

Jak se hledá potenciální biomarker v omics datech

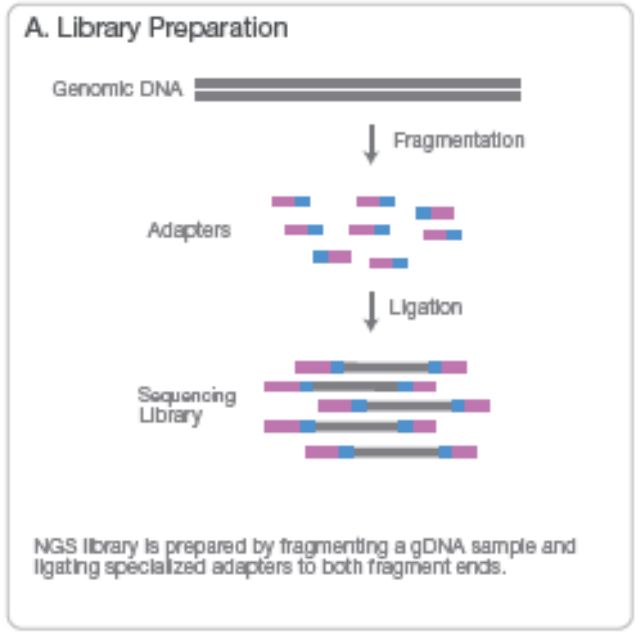


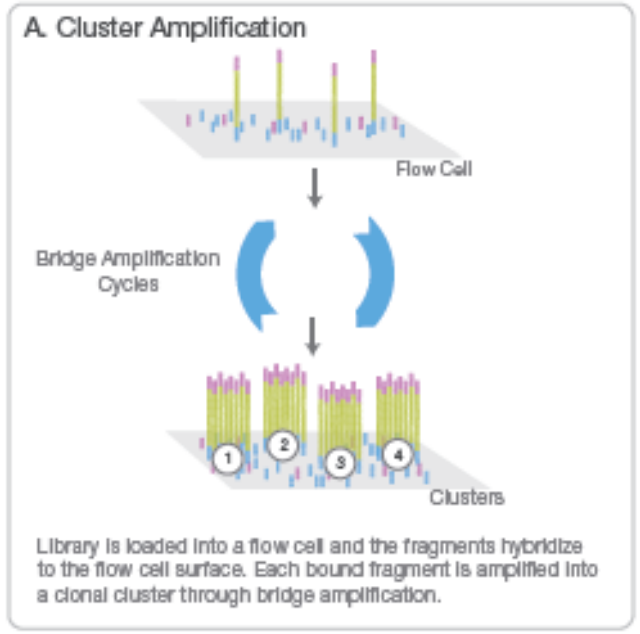
Sekvenování technologií Illumina

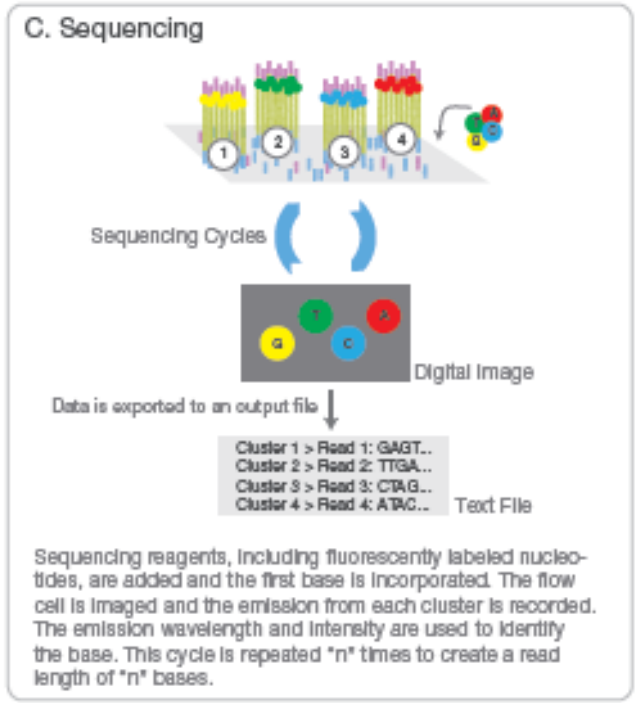
Illumina Sequencing Technology - YouTube

Kroky sekvenování technologií Illumina

1. Fragmentace genomické DNA, např. sonikátorem
2. Ligování adaptérů na oba konce fragmentů
3. PCR amplifikace fragmentů s adaptéry
4. Rozprostření molekul DNA napříč flowcelami. Cílem je získat přesně jednu molekulu DNA na každý potenciální shluk primerů. To závisí čistě na pravděpodobnosti, založené na koncentraci DNA.
5. Použití bridge PCR k amplifikaci jedné molekuly na každém shluku, k získání dostatečně silného signálu pro detekci. Obvykle to vyžaduje několik set nebo málo tisíc molekul.
6. Sekvence syntézou komplementárního vlákna: chemie reverzibilního terminátoru.

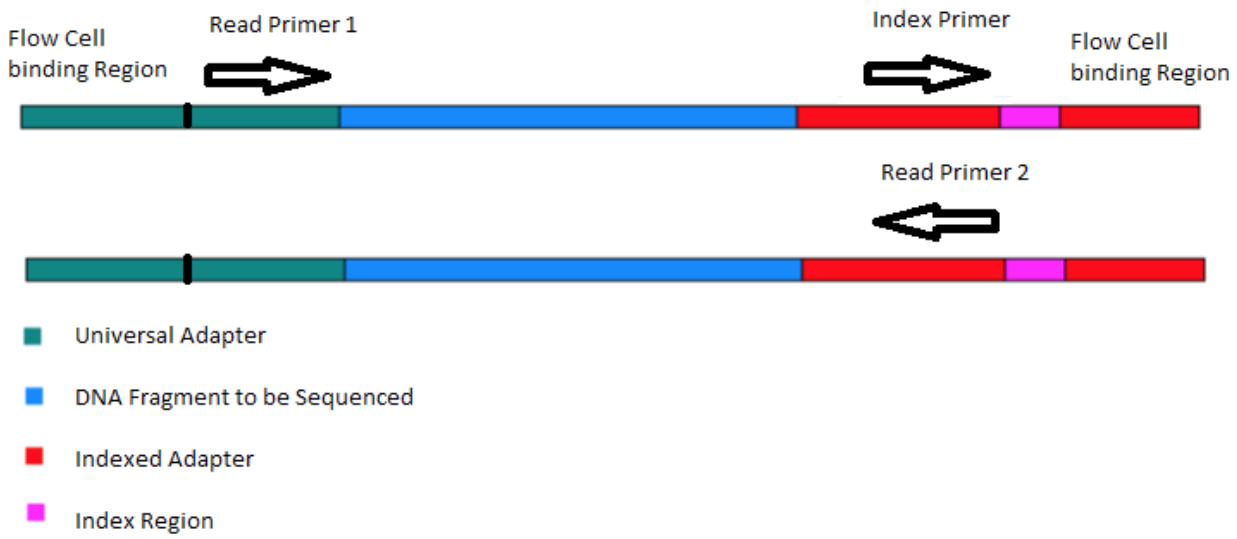






Zdroj chyb: ligování adaptérů

- V kroku 2 jsou adaptéry ligovány na konec fragmentů



Sekvenování náhodných fragmentů DNA je možné přidáním krátkých nukleotidových sekvencí, které slouží k:

1) Navázání fragmentů na NGS flow cell
2) PCR pouze fragmentů DNA ligovaných s a daptérem
3) Indexování nebo „čárové kódování“ vzork ů pro smíchání více knihoven v jednom bě hu (multiplexing)
4) Značení pro zjištění chyb v sekvenaci (PCR duplikátů...)

From:
http://tucf-genomics.tufts.edu/documents/protocols/TUCF_Understanding_Illumina_TruSeq_Adapters.pdf

Co vše se liguje

Adaptery

Primery

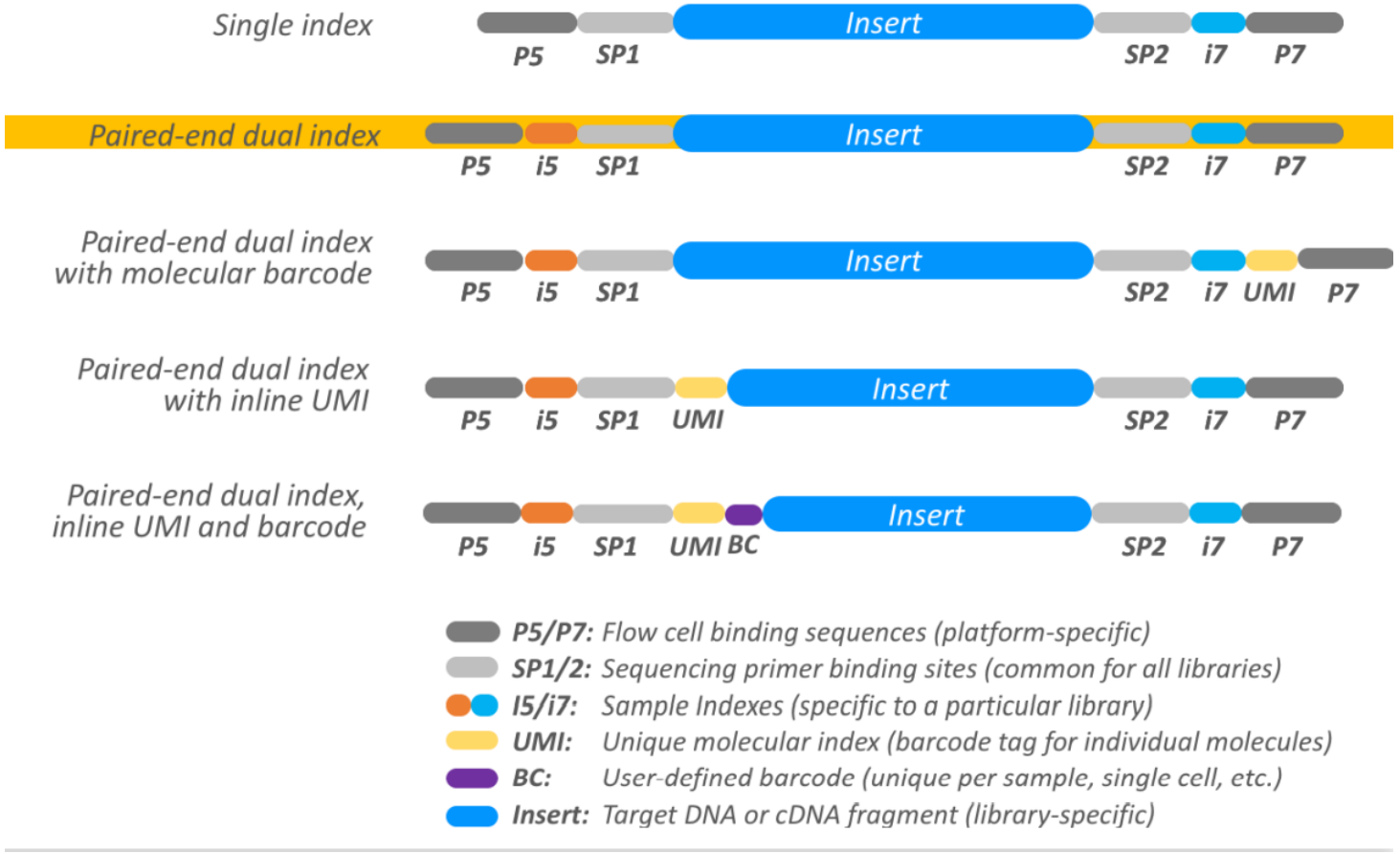
Tagy

Barkódy

UMIs

Spacery

Linkery



Co vše se liguje

Musí být přítomny:

P5/P7 – adaptéry pro vazbu na flow cell

SP1/SP2 – vazebné místo sekvenačního primeru

Volitelné – ale často používané:

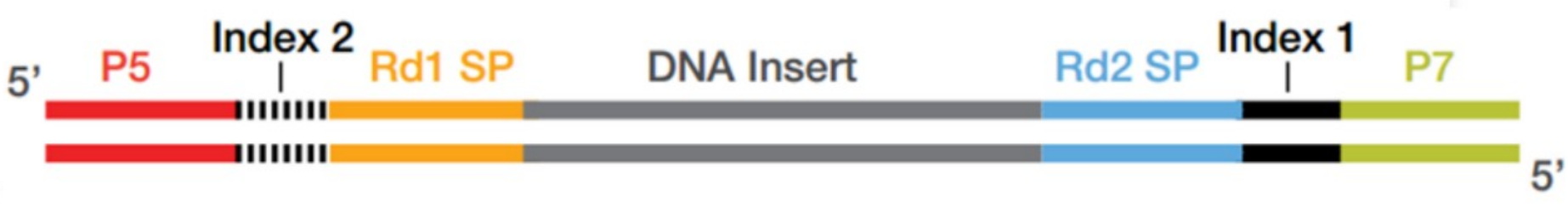
i5/i7 – Index vzorku – k rozpoznání sekvenovaných knihoven

Volitelné:

Barcode - jedinečná sekvence pro rozpoznání vzorku

UMI – Unique Molecular Identifikátor – k identifikaci technických duplikátů

Spacery - Pokud kombinujeme různé délky knihoven
Linkery - pro lepší slučování sekvencí



[Indexed Sequencing Overview Guide (15057455) (ox.ac.uk)]

Odstranění adaptérů z knihovny

- Nutný krok!
- Odstranění neligovaných adaptérů a adaptérových dimerů (dva adaptéry vzájemně ligované) je zásadní pro zlepšení výstupu a kvality dat
- Přebytečné adaptéry často soutěží s fragmenty knihovny ve vazbě na průtokovou buňku, čímž se snižuj e datový výstup.
- Adaptérové dimery mohou také klonálně amplifikovat a generovat sekvenační „šum“, který musí být bě hem analýzy dat odfiltrován.
- Přebytek neligovaných adaptérů činí knihovny náchylnějšími k indexovému přeskakování během sekve nování

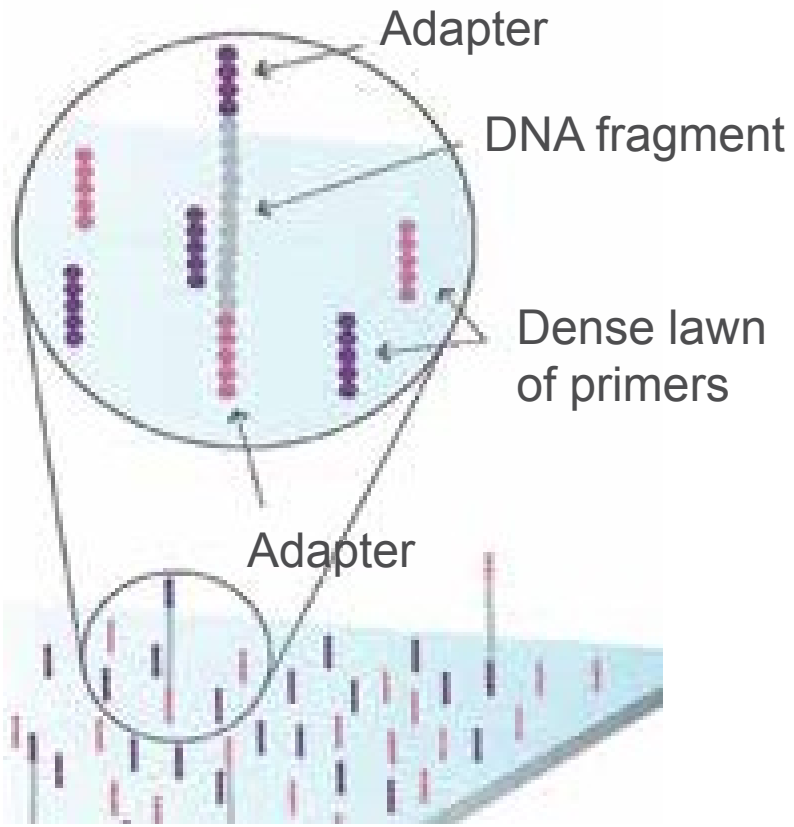
Zdroj chyb: PCR duplikáty

•V kroku 3 záměrně vytváříme více kopií každé původní molekuly genomové DNA, abychom jich měli dostatek.

•K duplikátům PCR dochází, když se dvě kopie stejné původní molekuly dostanou na různé primerové oblasti ve flowcele

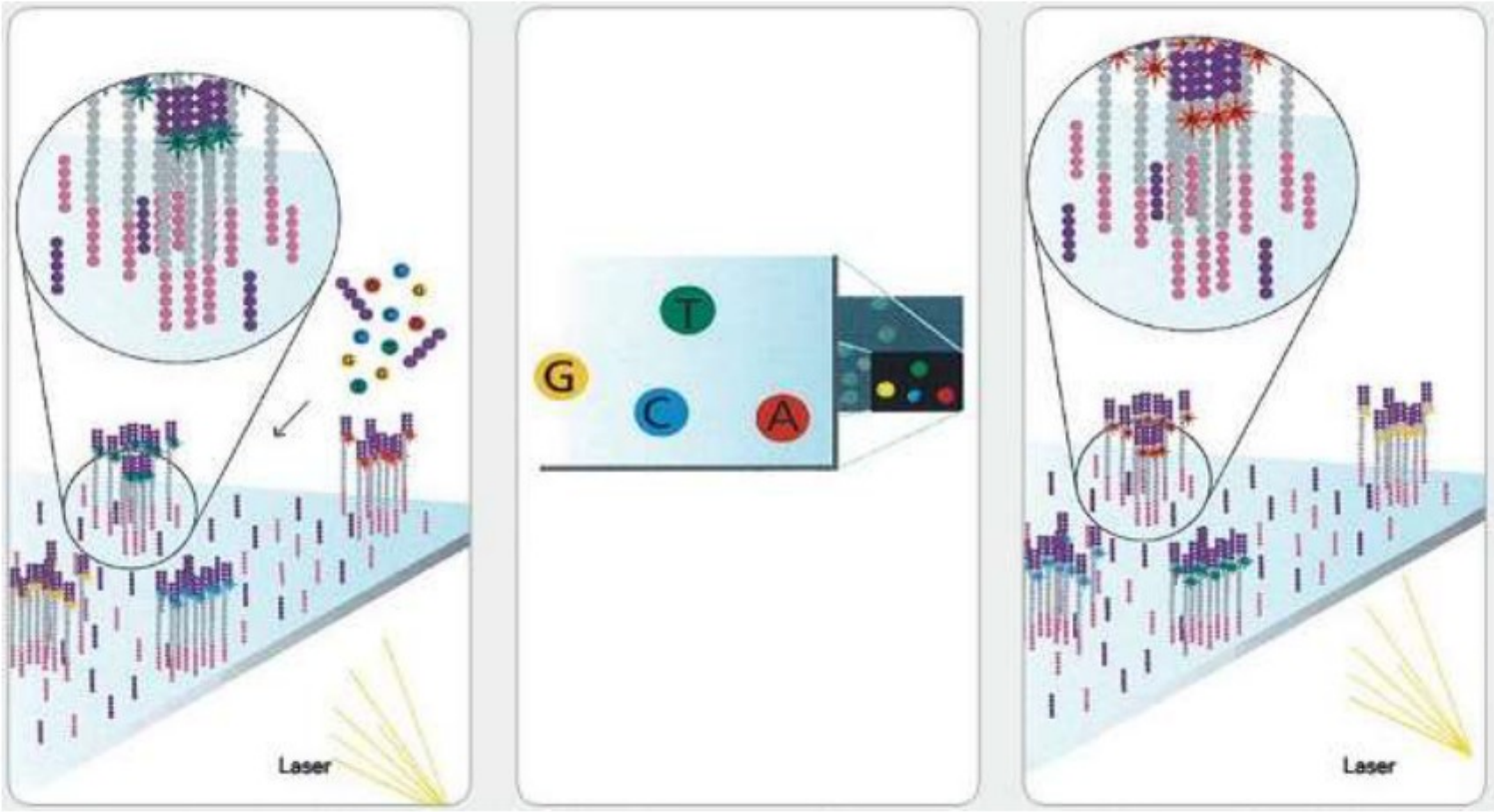
•V důsledku toho čteme stejnou sekvenci dvakrát!

• Vyšší četnosti PCR duplikátů, např. 30 % vznikají, když máte příliš málo výchozího materiálu, takže je potřeba větší amplifikace knihovny v kroku 3, nebo když máte příliš velký rozptyl ve velikosti fragmentu, takže menší fragmenty, které se snadněji amplifikují pomocí PCR jsou overreprezentovány



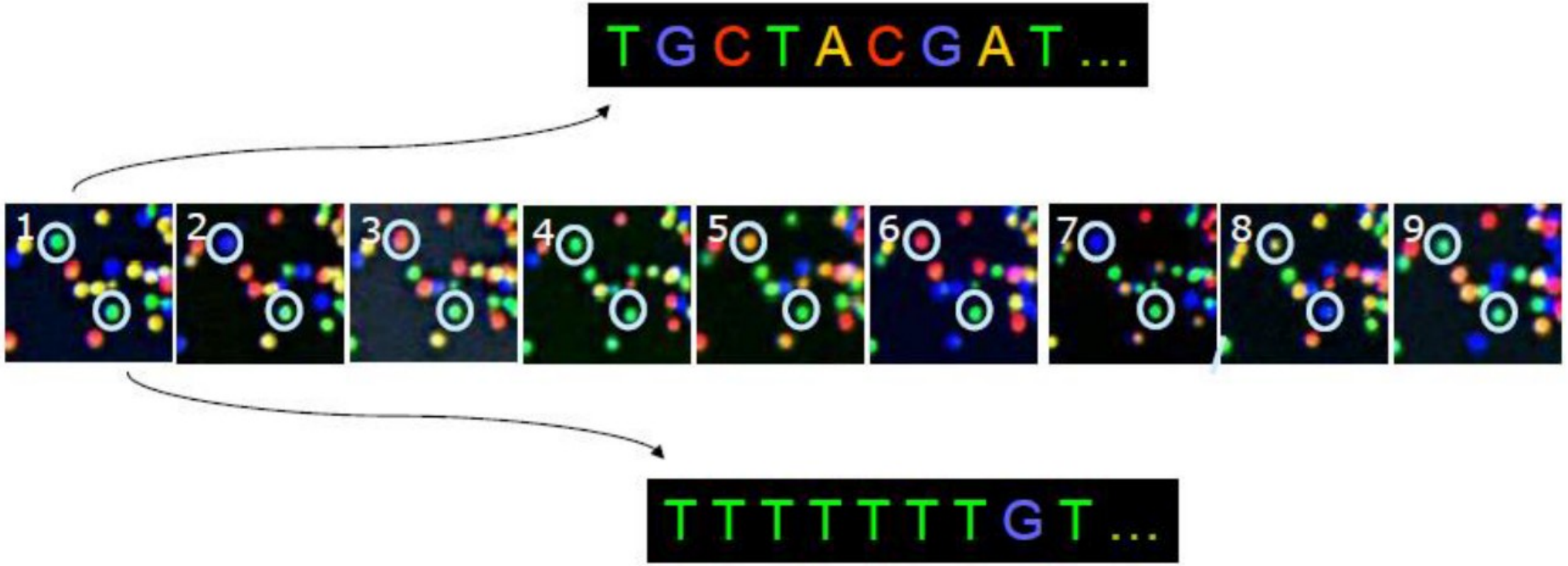
Find beautiful explanation of probabilities and much more at: https://www.cureffi.org/2012/12/11/how-pcr-duplicates-arise-in-next-generation-sequencing/

Během sekvencování se tvoří shluky stejných sekvencí - clusters

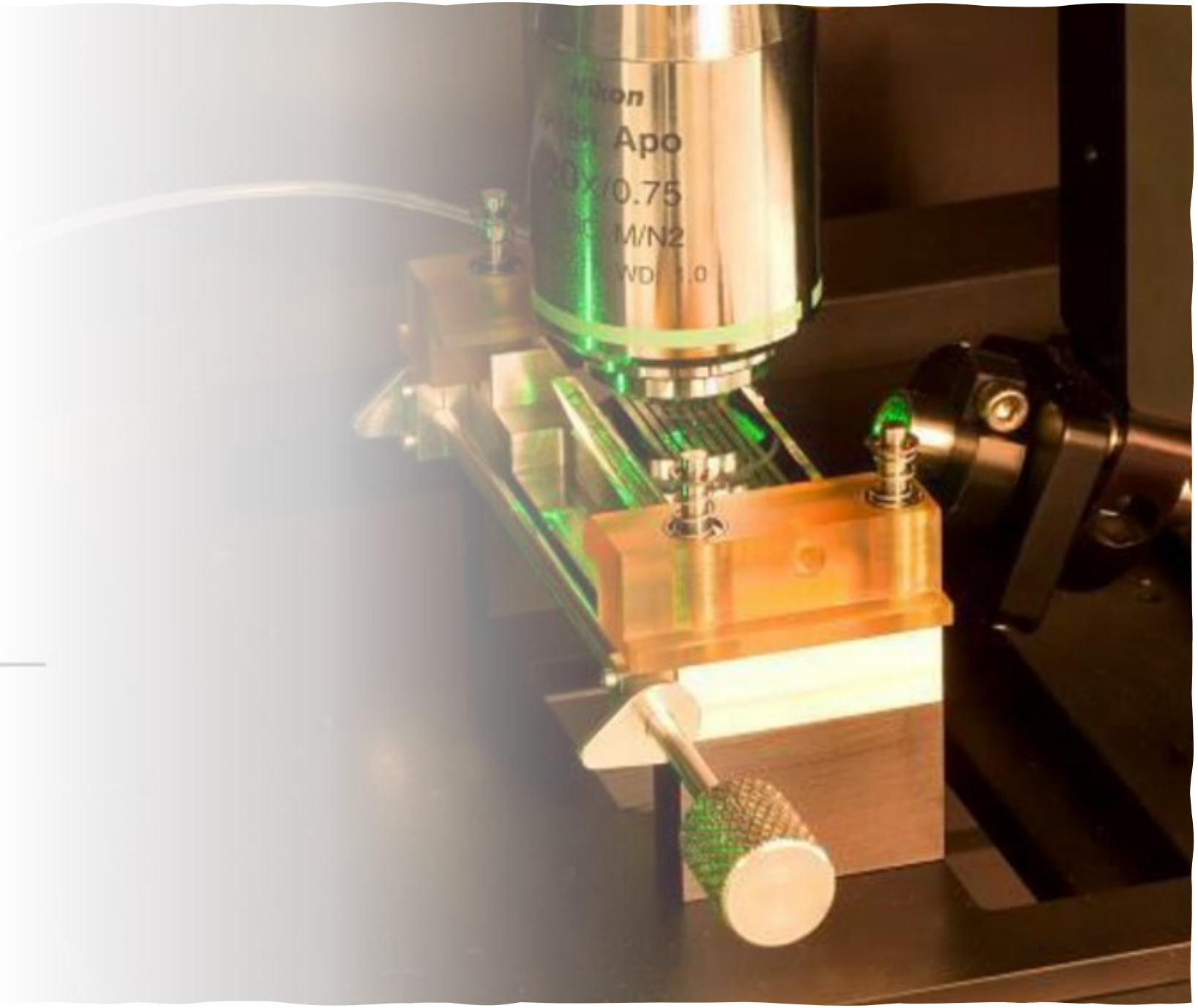


Krok 0 analýzy

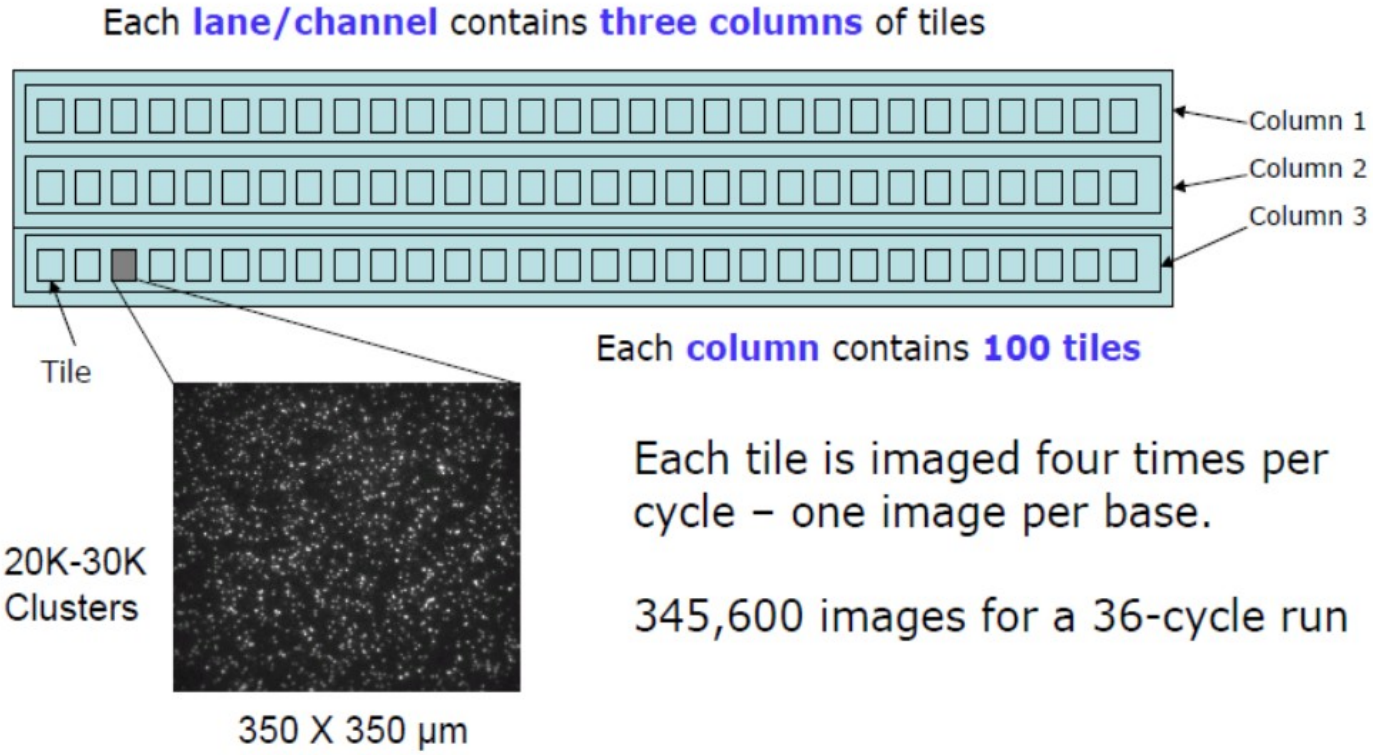
- Identita každé báze ve shluku se odečítá ze sekvenčních obrázků
- Jeden cyklus -> čtyři snímky!



Flow-cell imaging



Jak to probíhá

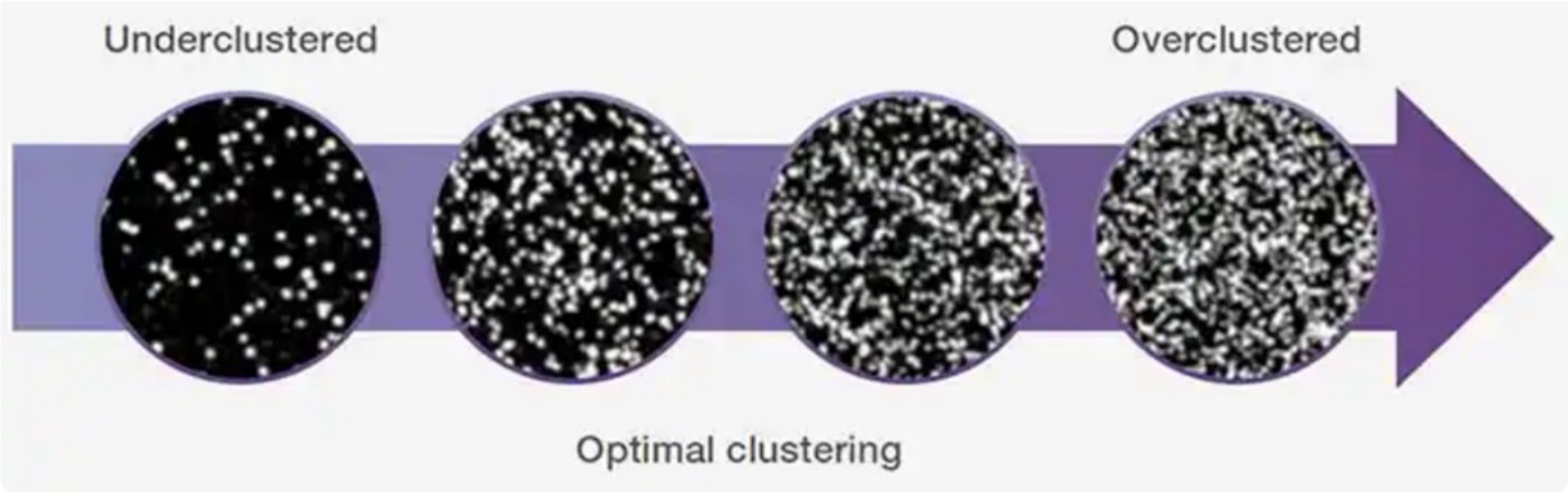


Each tile is imaged four times per cycle – one image per base.

345,600 images for a 36-cycle run

Trochu počítejme

- 100 tiles na lane, 8 lanes na flow cell, 36 cyklů
- 4 obrázky (A,G,C,T) na dlaždici a cyklus = 115 200 obrázků
- Každý obrázek tiff má ~ 7 MB = 806 400 MB dat
- 1,6 TB na 70 nt čtení, 3,2 TB pro 70 nt párové čtení
- Většina technologií při sekvenování vymazává intenzity, a to z důvodu tak velkého množství dat



Zdroj chyb: Koncentrace knihovny

- Koncentrace připravených knihoven NGS se mohou široce lišit kvůli rozdílům v množství a kvalitě vstupu nukleové kyseliny, stejně jako v cílové metodě obohacení, která může být použita.
- **podshlukování** v důsledku **nadhodnocených** koncentrací knihoven může mít za následek snížený počet readů proti kapacitě
- **nadměrné** množství shluků může mít za následek **nízké skóre kvality** a problematickou následnou analýzu - shluky se špatně odlišují programem pro analýzu obrazu!

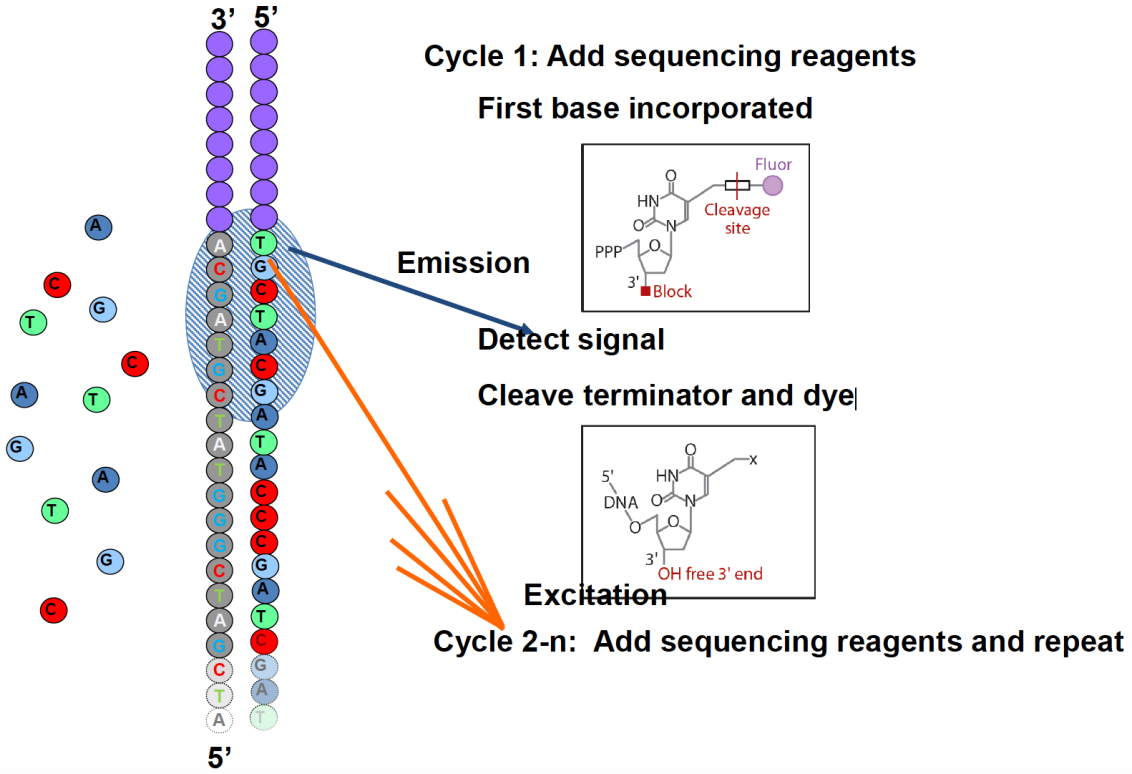
Zdroje chyb: sekvenování syntézou – fluorescence

V kroku 5 zesilujeme signál a detekujeme fluorescenci každé báze

Předpokladem je, že v cyklu je každá molekula na průtokové cele prodloužena o jednu bázi

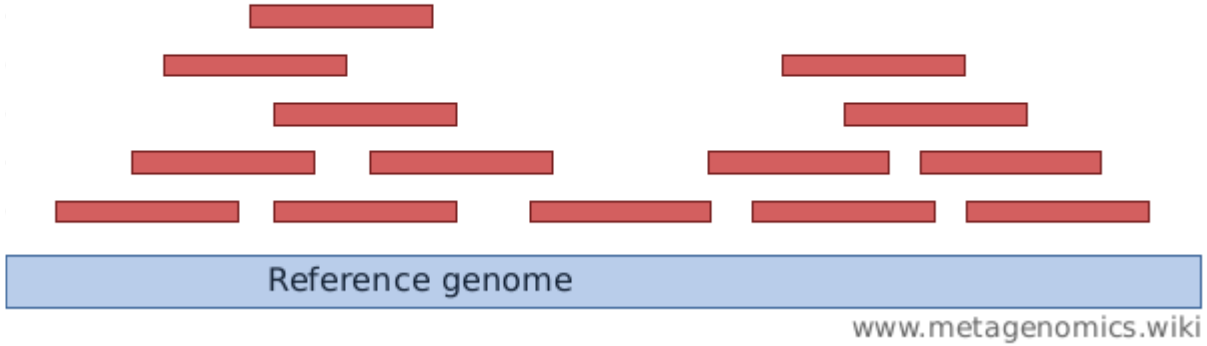
Realita:

- Některé molekuly nejsou prodlouženy nebo jejich báze nemá fluorescenční barvivo
- Předchozí fluorescenční barvivo není štěpeno – signál z klastru po několika cyklech je směsí signálů z předchozích bází



Sekvenační pokrytí (**coverage**)

Pokrytí v sekvenování DNA je **počet jedinečných čtení**, která **zahrnují daný nukleotid** v referenční sekvenci.



Hloubka pokrytí (**coverage depth**)

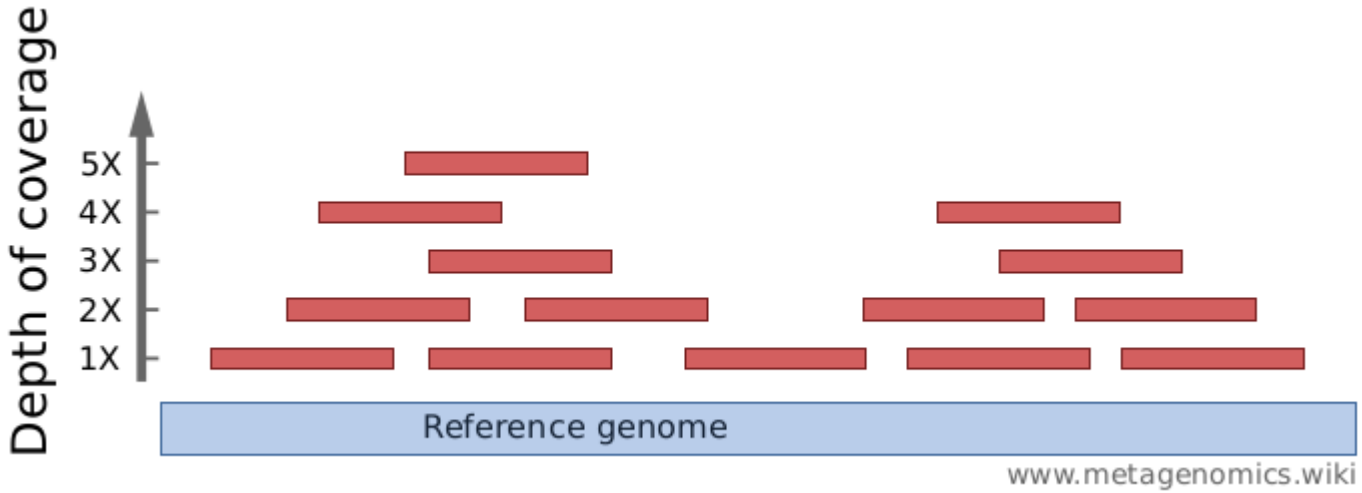
Jak silně je genom „pokryt“ sekvenovanými fragmenty (krátké čtení)?

Pokrytí na bázi (per-base coverage) je průměrný počet, kolikrát byla sekvenována daná báze genomu (jinými slovy, kolik čtení ji pokrývá).

Průměrné pokrytí genomu (Av)

$$Av = (NxL)/G$$

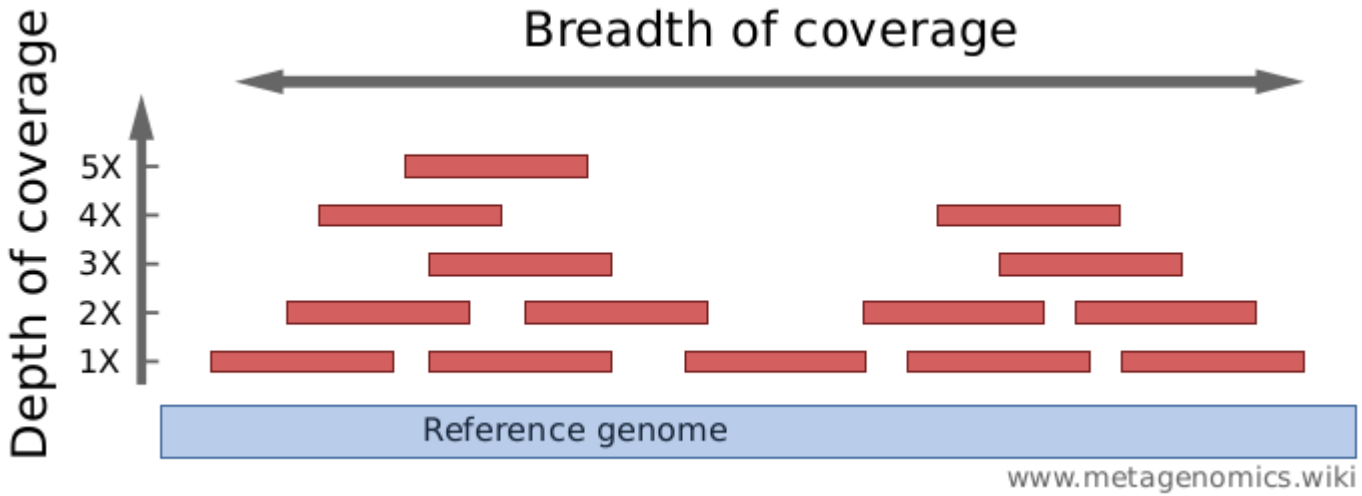
G - délka původního genomu
N - počet čtení
L - průměrná délka čtení



Hloubka pokrytí genomu se vypočítá jako počet bází všech krátkých čtení, která odpovídají genomu, podělené délkou tohoto genomu. Často se vyjadřuje jako 1X, 2X, 3X,... (1, 2 nebo 3násobné pokrytí).

Šířka krytí (**breadth of coverage**).

Jaká část genomu je „pokryta“ krátkým čtením? Existují oblasti, které nejsou pokryty, a to ani jedním přečtením?



Šířka krytí je procento bází referenčního genomu, které jsou **pokryty určitou hloubkou**. Například: "90 % genomu je pokryto v hloubce 1X a stále 70 % je pokryto v hloubce 5X."

Doporučení pr
pokrytí

Určuje se na základě :

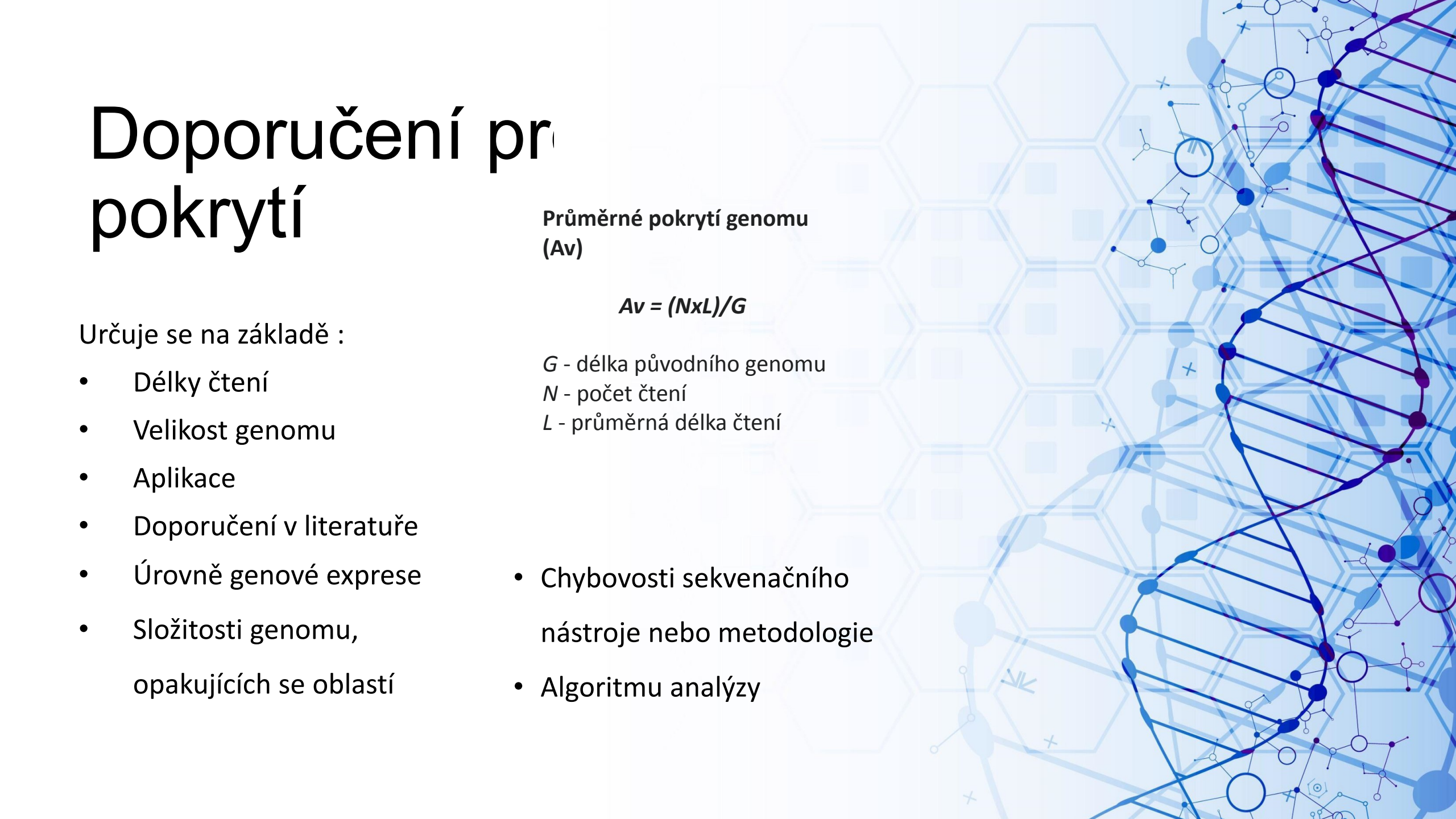
- Délky čtení
- Velikost genomu
- Aplikace
- Doporučení v literatuře
- Úrovně genové exprese
- Složitosti genomu, opakujících se oblastí

Průměrné pokrytí genomu (Av)

$$Av = (NxL)/G$$

G - délka původního genomu
N - počet čtení
L - průměrná délka čtení

- Chybovosti sekvenačního nástroje nebo metodologie
- Algoritmu analýzy



Doporučení pro pokrytí - DNA

Application Type	Coverage
DNA-Seq (Re-Sequencing)	30 - 80X
DNA-Seq (De novo assembly)	100X
SNP Analysis / Rearrangement Detection	10 - 30X
Exome	100 - 200X
ChIP-Seq	10 - 40X

Average coverage of the genome (Av)

$$Av = (NxL)/G$$

G - length of the original genome
N - number of reads
L - average read length

Coverage and Read Depth Recommendations for Next-Generation Sequencing Applications (genohub.com)

Doporučení pro pokrytí - RNA

Sample Type	Reads Needed for Differential Expression (millions)	Reads Needed for Rare Transcript or De Novo Assembly (millions)	Read Length
Small Genomes (i.e. Bacteria / Fungi)	5	30 - 65	50 SR or PE for positional info
Intermediate Genomes (i.e. Drosophila / C. Elegans)	10	70 - 130	50 – 100 SR or PE for positional info
Large Genomes (i.e. Human / Mouse)	15 - 25	100 - 200	>100 SR or PE for positional info

Různé transkripty jsou exprimovány na různých úrovních => více čtení bude zachyceno z vysoce exprimovaných genů
Složitost transkriptomu, alternativní exprese, 3' související zkreslení a distribuce úrovní exprese ztěžují stanovení pokrytí.

PŘI VÝPOČTU POZOR ! Potřebujeme počítat s namapovanými čteními, ne s celkovým počtem čtení.

Coverage and Read Depth Recommendations for Next-Generation Sequencing Applications (genohub.com)

Doporučení pro pokrytí - dle aplikace

Category	Detection or Application	Recommended Coverage (x) or Reads (millions)	References
Whole genome sequencing	Homozygous SNVs	15x	Bentley et al., 2008
	Heterozygous SNVs	33x	Bentley et al., 2008
	INDELs	60x	Feng et al., 2014
	Genotype calls	35x	Ajay et al., 2011
	CNV	1-8x	Xie et al., 2009; Medvedev at al., 2010
Whole exome sequencing	Homozygous SNVs	100x (3x local depth)	Clark et al., 2011; Meynert et al., 2013
	Heterozygous SNVs	100x (13x local depth)	Clark et al., 2011; Meynert et al., 2013
	INDELs	not recommended	Feng et al., 2014
Transcriptome Sequencing	Differential expression profiling	10-25M	Liu Y. et al., 2014; ENCODE 2011 RNA-Seq
	Alternative splicing	50-100M	Liu Y. et al., 2013; ENCODE 2011 RNA-Seq
	Allele specific expression	50-100M	Liu Y. et al., 2013; ENCODE 2011 RNA-Seq
	De novo assembly	>100M	Liu Y. et al., 2013; ENCODE 2011 RNA-Seq

Coverage and Read Depth Recommendations for Next-Generation Sequencing Applications (genohub.com)

Doporučení pro pokrytí - dle aplikace

DNA Target-Based Sequencing	ChIP-Seq	10-14M (sharp peaks); 20-40M (broad marks)	Rozowsky et al., 2009; ENCODE 2011 Genome; Landt et al., 2012
	Hi-C	100M	Belton, J.M et al., 2012
	4C (Circularized Chromosome Confirmation Capture)	1-5M	van de Weken, H.J.G. et al., 2012
	5C (Chromosome Carbon Capture Carbon Copy)	15-25M	Sanyal A. et al., 2012
	ChIA-PET (Chromatin Interaction Analysis by Paired-End Tag Sequencing)	15-20M	Zhang, J. et al., 2012
	FAIRE-Seq	25-55M	ENCODE 2011 Genome; Landt et al., 2012
	DNAse 1-Seq	25-55M	Landt et al., 2012
DNA Methylation Sequencing	CAP-Seq	>20M	Long, H.K. et al., 2013
	MeDIP-Seq	60M	Taiwo, O. et al., 2012
	RRBS (Reduced Representation Bisulfite Sequencing)	10X	ENCODE 2011 Genome
	Bisulfite-Seq	5-15X; 30X	Ziller, M.J et al., 2015; Epigenomics Road Map

Coverage and Read Depth Recommendations for Next-Generation Sequencing Applications (genohub.com)

Doporučení pro pokrytí - dle aplikace

RNA-Target-Based Sequencing	CLIP-Seq	10-40M	Cho J. et al., 2012; Eom T. et al., 2013; Sugimoto Y. et al., 2012
	iCLIP	5-15M	Sugimoto Y. et al., 2012; Rogelj B. et al., 2012
	PAR-CLIP	5-15M	Rogelj B. et al., 2012
	RIP-Seq	5-20M	Lu Z. et al., 2014
Small RNA (microRNA) Sequencing	Differential Expression	~1-2M	Metpally RPR et al., 2013; Campbell et al., 2015
	Discovery	~5-8M	Metpally RPR et al., 2013; Campbell et al., 2015

Coverage and Read Depth Recommendations for Next-Generation Sequencing Applications (genohub.com)

Kolik vzorků na běh?

Závisí od použité platformy a jejího maxima
a požadovaného počtu čtení na vzorek (v milionech)

Numbers of Single Reads by Instrument Manufacturer

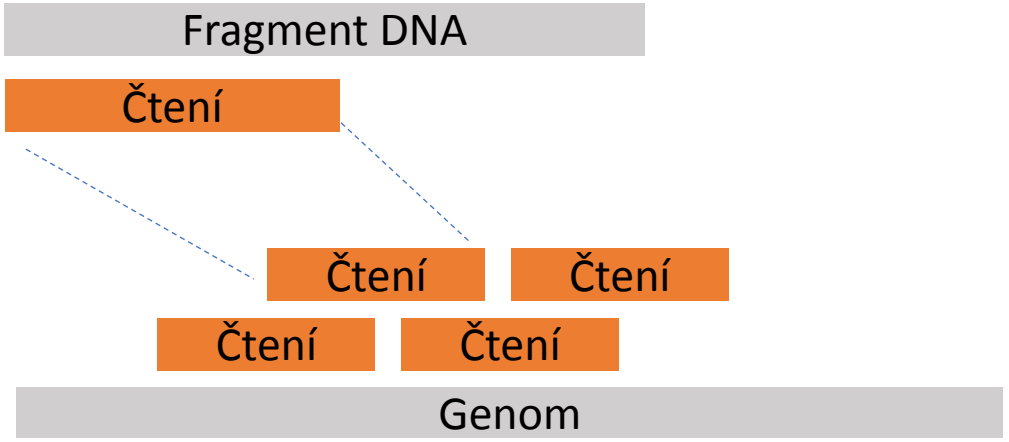
Platform	Instrument	Unit	Reads / Unit	Reference
Illumina	HiSeq X Ten	Lane	375,000,000	1
Illumina	HiSeq 3000/4000	Lane	312,500,000	1
Illumina	HiSeq NextSeq 500 High-Output	Run	400,000,000	2
Illumina	HiSeq NextSeq 500 Mid-Output	Run	130,000,000	2
Illumina	HiSeq High-Output v4	Lane	250,000,000	3
Illumina	HiSeq High-Output v3	Lane	186,048,000	3
Illumina	HiSeq Rapid Run	Lane	150,696,000	3
Illumina	HiScanSQ	Lane	93,024,000	3
Illumina	GAIIx	Lane	42,075,000	3
Illumina	MiSeq v3	Lane	25,000,000	4
Illumina	MiSeq v2	Lane	16,000,000	3
Illumina	MiSeq	Lane	5,000,000	3
Illumina	MiSeq v2 Micro	Lane	4,000,000	5
Illumina	MiSeq v2 Nano	Lane	1,000,000	5

Designing Next-Generation Sequencing Runs (genohub.com)

Single nebo paired- end?

Single-end sekvencování

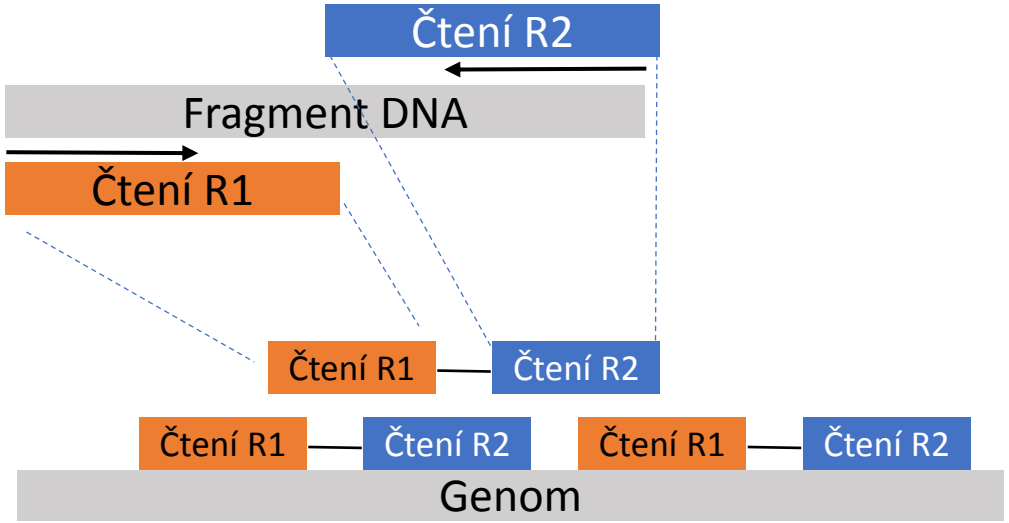
- Výhody: rychlé, levné
- Nevýhody: omezené použití
- Použití: obvykle postačuje pro studie, jejichž cílem je zjistit spíše počet molekul, než jejich typ, jako je RNA-Seq nebo ChIP-Seq



Single nebo paired- end?

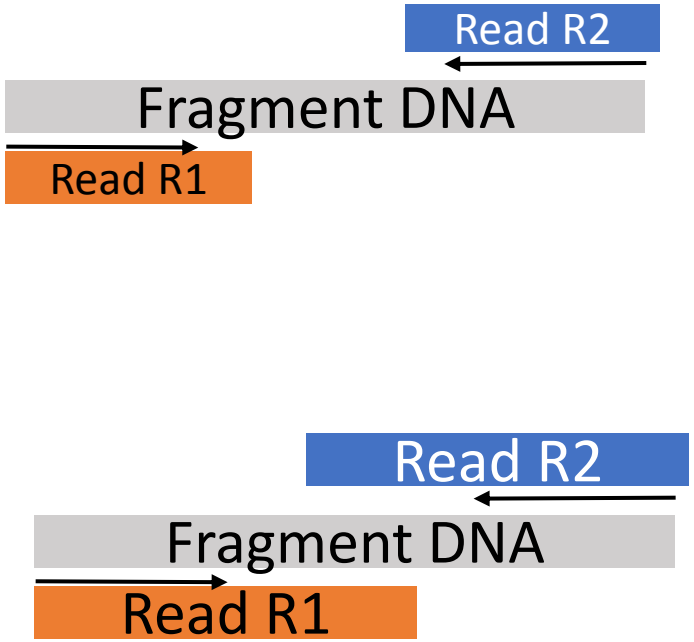
Paired-end sekvencování

- Výhody:
  - větší přesnost, v jednom běhu dvounásobný počet čtení na vzorek (větší kapacita) za méně než cena dvou sekvenačních běhů
- Nevýhody: pomalejší, dražší (relativně)
- Použití:
  - de novo sestavení genomu
  - Analýza strukturálních změn (delece, inzerce, inverze) a SNP
  - Studium sestřihových variant
  - Epigenetické modifikace (metylace)



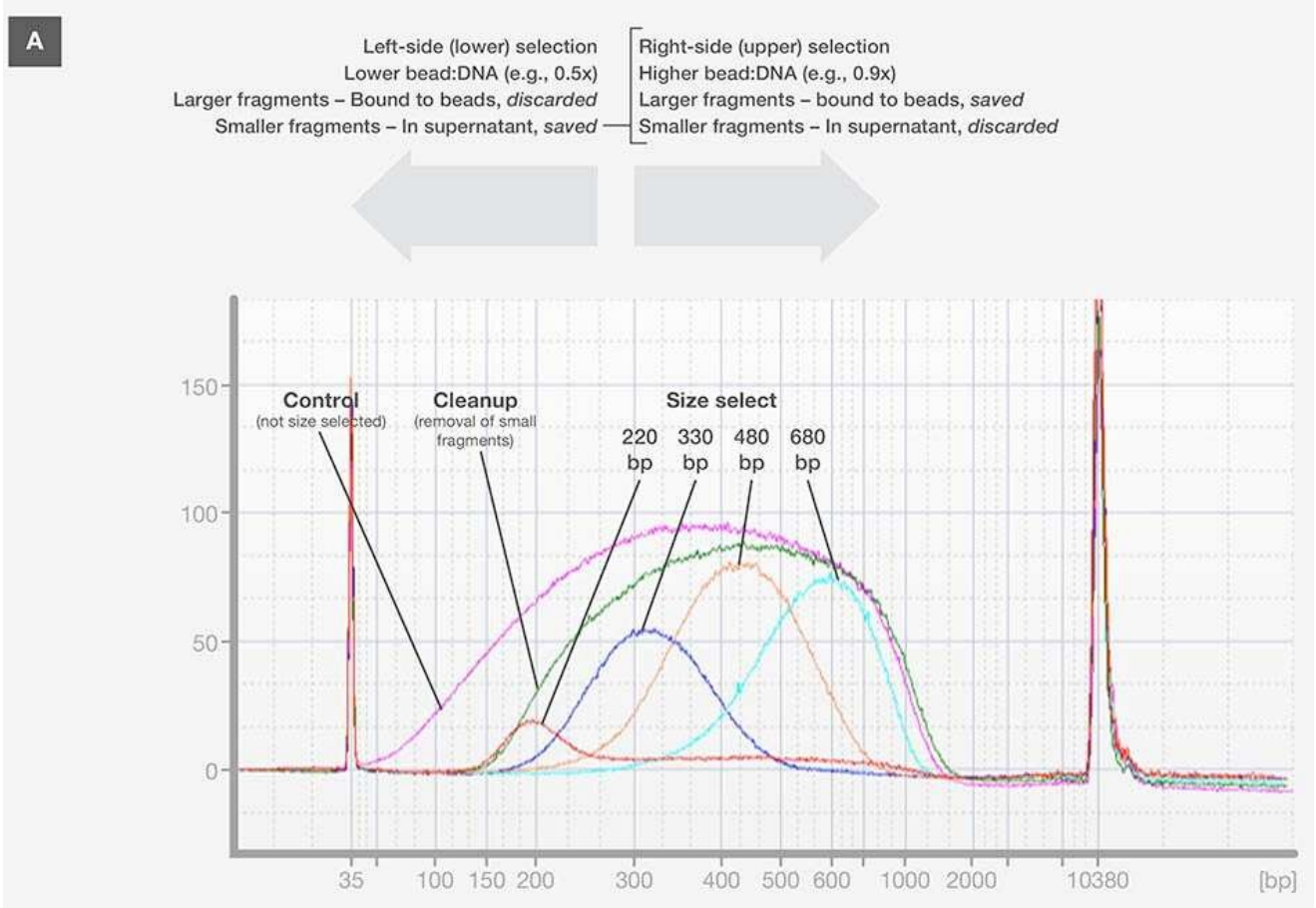
Délka čtení

- Delší délky čtení poskytují přesnější informace o relativních pozicích bazí v genomu, jsou dražší než kratší.
- 50-75 cyklů je typicky dostačujících pro jednoduché mapování čtení do referenčního genomu a experimenty s kvantifikující třeba genovou expresi (RNA-Seq)
- Délky čtení větší nebo rovné 100 se typicky volí pro studie genomu nebo transkriptomu, které vyžadují větší **přesnost**
- **Přesná délka záleží na délce insertů!!!**



Délka čtení a cílové fragmenty!

- Délka fragmentů by měla zhruba odpovídat délce čtení (v případě paired-end readů jejich součtu)
- Uniformita velikostí fragmentů je zásadní, protože délka čtení je omezená
- **Podstatně delší inzerty DNA** => některé části inzertů zůstanou **nesekvenované**.
- **Kratší než doporučené** => **neoptimální** využití sekvenačních činidel a zdrojů.
- Kombinace krátkých a dlouhých insertů => **snižuje efektivitu** sekvenování a představuje problémy při analýze dat.



[Preparation of DNA Sequencing Libraries for Illumina Systems—6 Key Steps in the Workflow | Thermo Fisher Scientific - CZ]

Délka čtení a cílové fragmenty!

Délka čtení je omezena sekvenační platformou a reagenčním kitem

Reagent Type		Reagent Kit Size	Maximum Number of Cycles	Additional Cycles Needed for Dual Index?
iSeq™ 100	i1 (v1 or v2)	300	322	No
MiniSeq™	Rapid Kit	100	128	Yes - 7 cycles
	High Output or Mid Output	75	92	No
		150	168	
		300	318	
MiSeq™	v2 (including Micro and Nano kits)	50	79	Yes - 7 cycles
		300	329	
		500	529	
	v3	150	179	
		600	629	
NextSeq™ 500/550	High Output or Mid Output	75	92	No
		150	168	
		300	318	
HiSeq™ 1000/1500/2000/2500	Rapid SBS v2	50	79	7 cycles required for paired-end flow cells
		200	229	
		500	529	
	TruSeq SBS v3	50	58	
		200	209	
	HiSeq SBS v4	50	79	
		250	279	

[How many cycles of SBS chemistry are in my kit? (illumina.com)]

Sequencing Platform	SBS Kit Version	Maximum Read Length
iSeq 100	v1	2 x 151bp
	v2	2 x 151bp
MiniSeq	MO*	2 x 151bp
	HO*	2 x 151bp
MiSeq	v2	2 x 251bp
	v3	2 x 301bp
NextSeq 500/550	MO*	2 x 151bp
	HO*	2 x 151bp
NextSeq 1000/2000	P1, P2, P3	2 x 151bp
HiSeq 1000/1500/2000/2500	HO* v3	2 x 101bp
	HO* v4	2 x 126bp
	RR** v4	2 x 251bp
HiSeq 3000/4000	N/A	2 x 151bp

[Maximum read length for Illumina sequencing platforms]

Užitečné zdroje

- Praktické laboratorní tipy pro knihovny:
  - Preparation of DNA Sequencing Libraries for Illumina Systems—6 Key Steps in the Workflow | Thermo Fisher Scientific - CZ

- Praktické tipy pro nastavení sekvenačního běhu:
  - Designing Next-Generation Sequencing Runs (genohub.com)
  - Optimizing Cluster Density on Illumina Sequencing Systems

- Indexed sequencing Illumina guide:
  - Indexed Sequencing Overview Guide (15057455) (ox.ac.uk)
- Další zdroje
  - Sequencing depth and coverage: key considerations in genomic analyses | Nature Reviews Genetics