Jak vypadají genomická data (a co se s nimi dělá)

Jaroslav Juráček

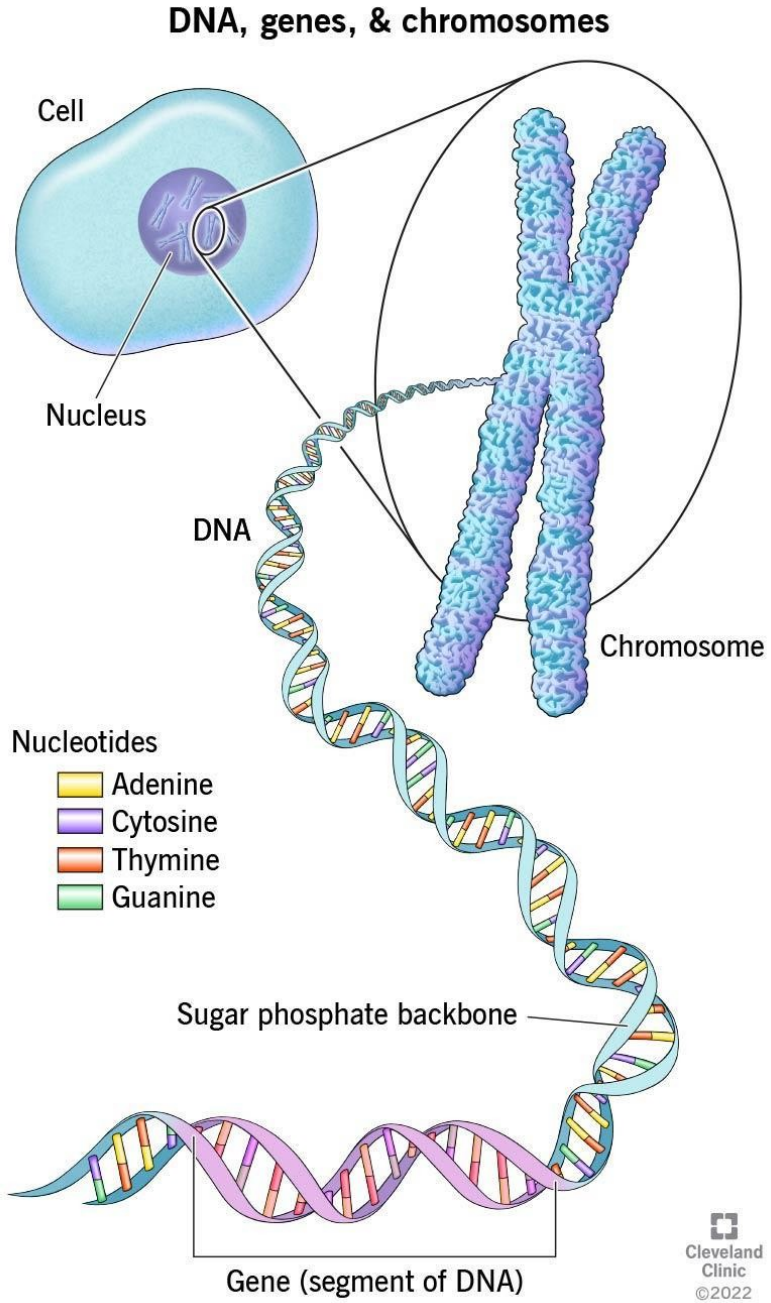
SitSem 2023

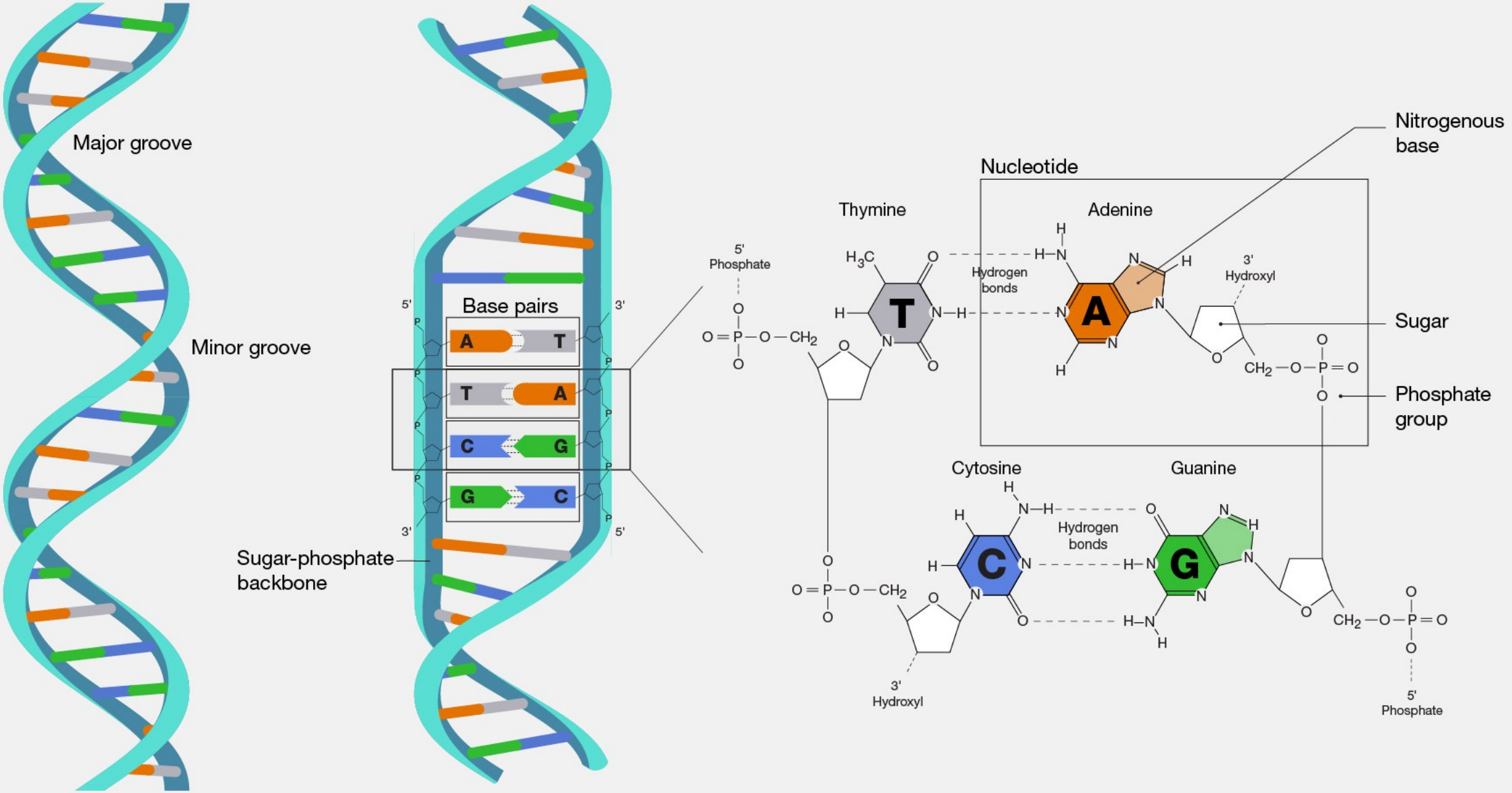
Co je to genom?

Genom – celková genetická informace nebo genetický materiál, který je obsažen v jádře buňky organismu. Genom zahrnuje celkovou DNA a obsahuje informace o všech vlastnostech jedince.

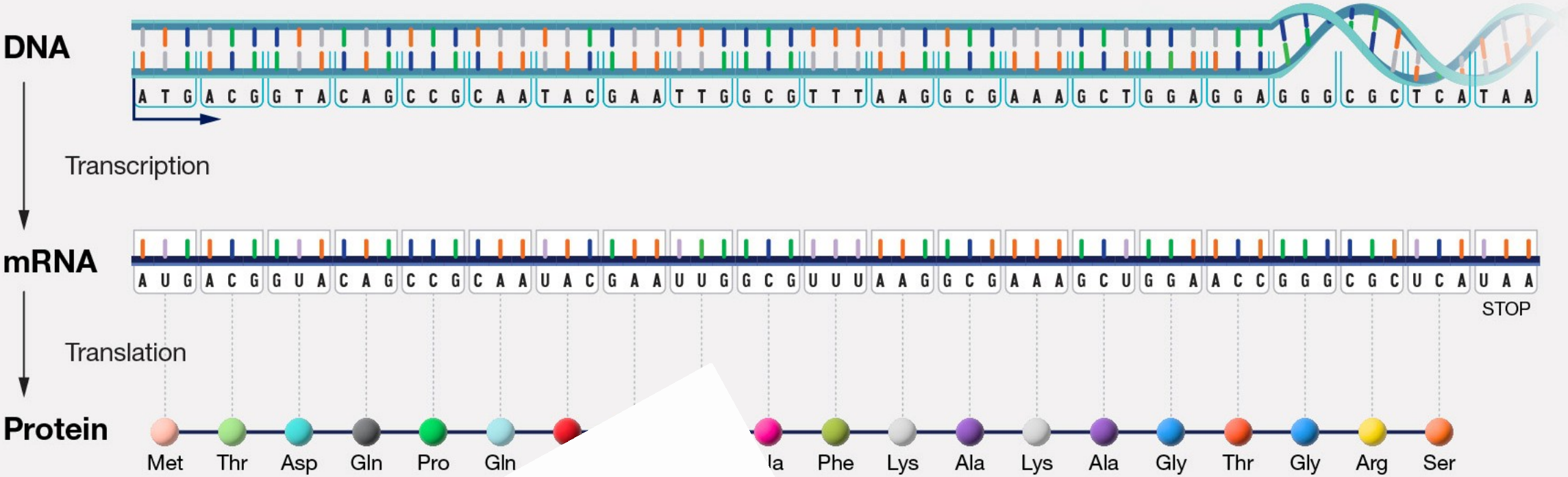
Gen – základní informační a funkční jednotka dědičné informace v živých organismech. Geny obsahují informace pro vytváření proteinů a regulačních molekul, které jsou nezbytné pro fungování organismu. Každý gen má specifickou funkci a přispívá k určitým vlastnostem a charakteristikám jedince.

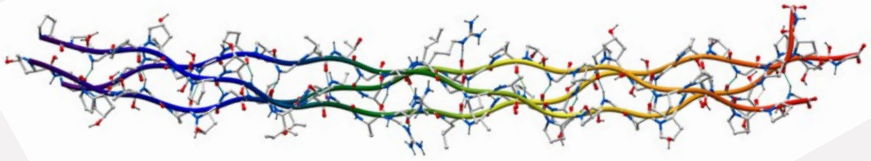
* Lidský genom zahrnuje cca 20 000 protein-kódujících genů
** To představuje jen asi 1-2 % lidského genomu



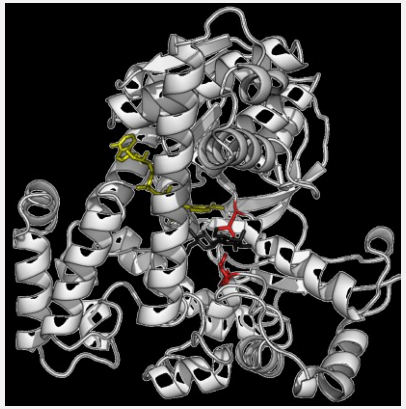


* V lidském genomu se nachází asi 3,2 miliardy párů bází





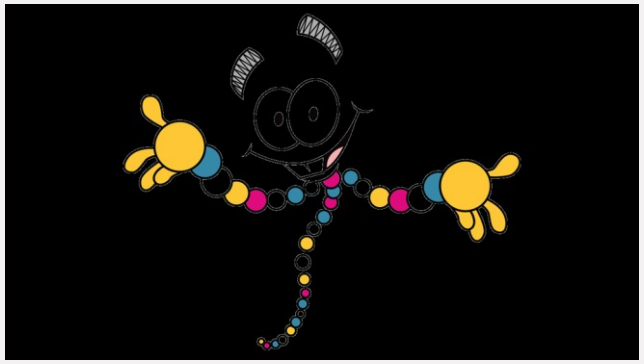
Stavební jednotky



Proč nás zajímá primární struktura DNA?

1 Výjimečnost

Genom každého člověka je **unikátní** (99,9 % genetického materiálu sdílíme, 0,1 % unikátní). Studium DNA a její primární struktury je klíčové pro **forenzní vědu** a **identifikaci** osob.



2 Principy dědičnosti

Studiem primární struktury můžeme lépe porozumět tomu, jak se genetická informace ukládá, replikuje a předává z jednoho pokolení na druhé

3 Medicína

Genetické choroby jsou spojeny s konkrétními změnami v primární struktuře DNA. Studium těchto změn je nezbytné pro **diagnostiku** genetických onemocnění a vývoj **léčebných** postupů.

4 Genové inženýrství

Manipulace s primární strukturou DNA vede k vytvoření nových genetických konstrukcí, transgenních organismů a terapeutických postupů, jako je genová terapie.

5 Genová exprese

6 Studium evoluce/biodiverzity

7 Studium mikrobiomu

8 …

Co jsou tedy genomická data?

Genomická data jsou soubory informací získaných z analýzy genetického materiálu (DNA nebo RNA) organismu. Tato data obsahují podrobné informace o **struktuře, sekvenci a funkci** genů a dalších genetických elementů v genomu daného organismu. Genomická data mohou zahrnovat následující informace:

1.Sekvence DNA: Pořadí nukleotidů (A, C, G, T) v molekule DNA. Sekvence DNA umožňují identifikovat geny, regulační sekvence a další důležité úseky DNA.

2.Geny: Genomická data obsahují informace o umístění a struktuře genů v genomu. Identifikujeme, jaký protein nebo RNA každý gen kóduje, a jakým způsobem může ovlivnit funkci organismu.

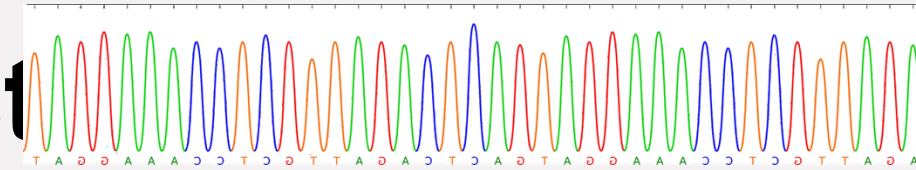
3.Genetické varianty: Změny v sekvenci DNA, jako jsou jednonukleotidové polymorfismy (SNP), delece, inzerce a další mutace. Genomická data obsahují informace o těchto variantách, které mohou mít význam pro dědičnost, vývoj chorob a individuální rozdíly mezi lidmi.

4.Genová exprese: Informace aktuální aktivaci genů a jak se projevují prostřednictvím produkce proteinů nebo RNA molekul. To umožňuje studovat, jaké geny jsou zapojeny do různých biologických procesů a jak mohou být ovlivněny různými podmínkami.

5.Struktura chromozomů: Data mohou obsahovat informace o uspořádání chromozomů v jádře buňky, což je důležité pro studium chromozomálních abnormalit a genetických onemocnění.

6.Mikrobiom: Analýza genetického materiálu mikroorganismů v rámci lidského těla – zastoupení a funkce/vliv těchto mikroorganismů.

The Human Genome Project



Human Genome Project měl za cíl zmapovat a rozluštit celý lidský genom. Projekt byl zahájen v roce **1990** a oficiálně dokončen v roce **2003**. Jednalo se o společný úsilí mezinárodního vědeckého komunity. Klíčové cíle projektu zahrnovaly:

1. Zmapování lidského genomu: Pořadí nukleotidů v lidské DNA v každém z 23 lidských chromozomů.

2. Identifikace genů: Identifikace a anotace všech genů v lidském genomu. To zahrnovalo určení, kde se jednotlivé geny nacházejí, jaké mají funkce a jaké proteiny kódují.

3. Studium genetických variant: Projekt zkoumal genetické varianty, jako jsou SNP (jednonukleotidové polymorfismy).

4. Aplikace v medicíně: Využití genomických znalostí pro zlepšení diagnostiky, prevence a léčby genetických chorob.

Dokončení projektu v roce 2003 bylo historickým milníkem v oblasti biologie a genetiky. Výsledky projektu poskytly základ pro rozvoj genomiky a personalizované medicíny.

Důležitým aspektem projektu bylo, že data z něj byla **veřejně dostupná** pro vědeckou komunitu, což umožnilo mnoha dalším výzkumným projektům a studiím využívat tuto cennou informační základnu pro další pokroky v oblasti genetiky a biomedicíny.

The Human Genome Project

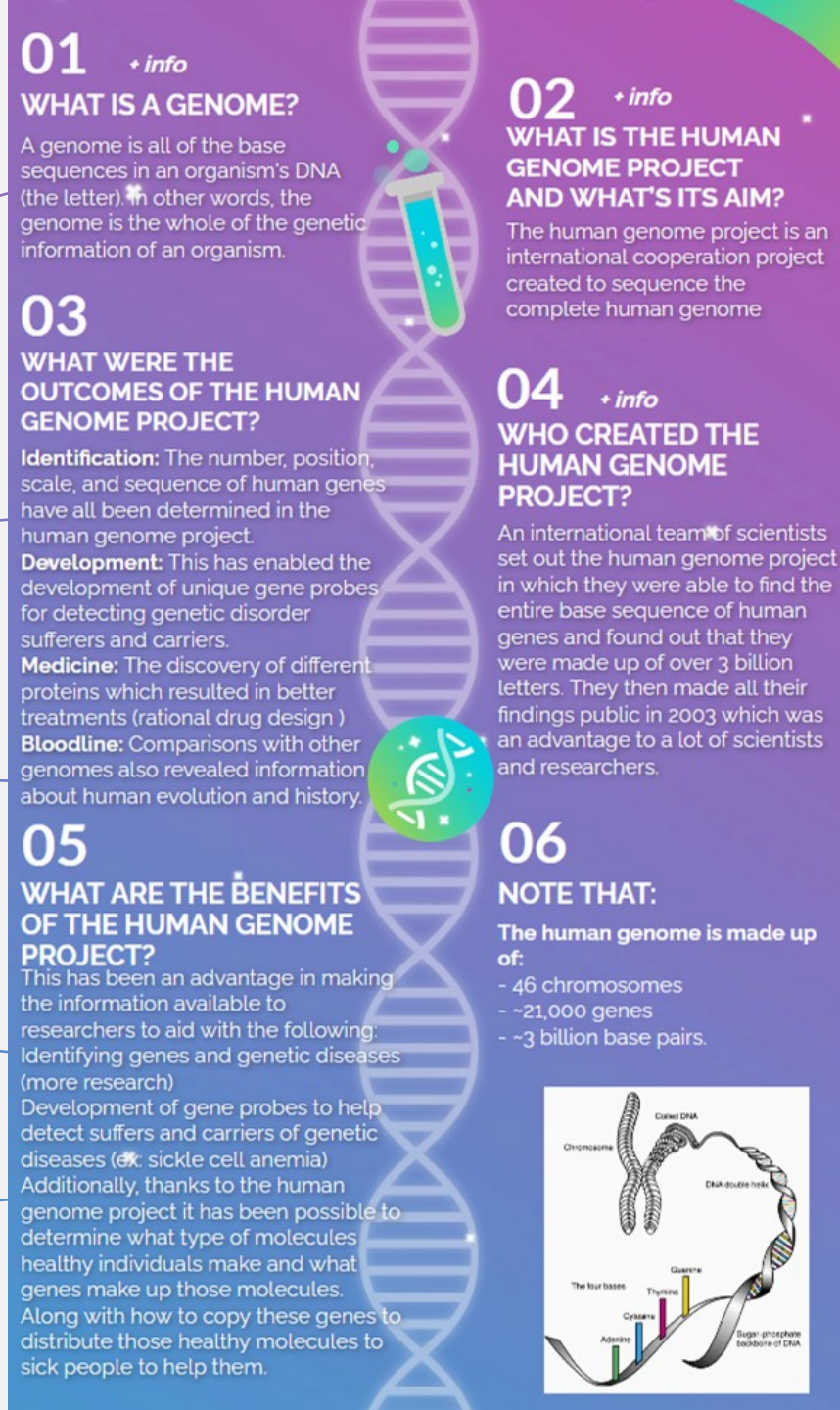
Na sekvenování lidského genomu se podíleli vědci z **20 různých univerzit a výzkumných center** ze Spojených států, Velké Británie, Francie, Německa, Japonska a Číny.

Sekvence lidského genomu **nepochází od jediného člověka**, ale několika lidí, jejichž identita byla záměrně anonymizována, aby bylo chráněno jejich soukromí.

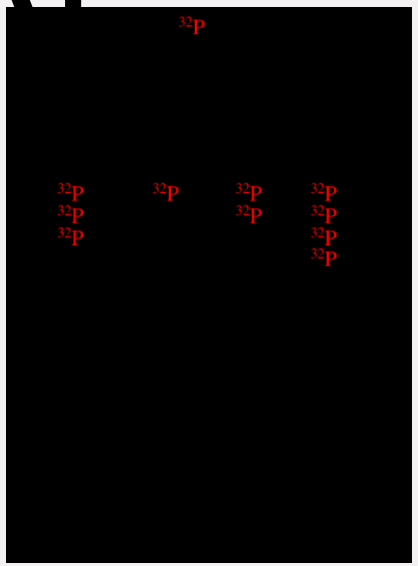
Původně předpokládané náklady činily **3 miliardy dolarů**, přičemž předpokládaná doba trvání projektu byla 15 let. Tato přibližná částka se blíží přesnému číslu.

Projekt ve výsledku **nevygeneroval** kompletní lidský genom. V dubnu 2003 konsorcium oznámilo, že vytvořilo v podstatě kompletní sekvenci lidského genomu. Konkrétně představovala 92 % lidského genomu a obsahovala méně než 400 mezer.

Dne 31. března 2022 konsorcium Telomere-to-Telomere (T2T) oznámilo, že doplnilo zbývající mezery a vytvořilo **první skutečně kompletní sekvenci** lidského genomu.



Jak přečíst primární sekvenci DNA?



Určení sekvence = metoda SEKVENOVÁNÍ

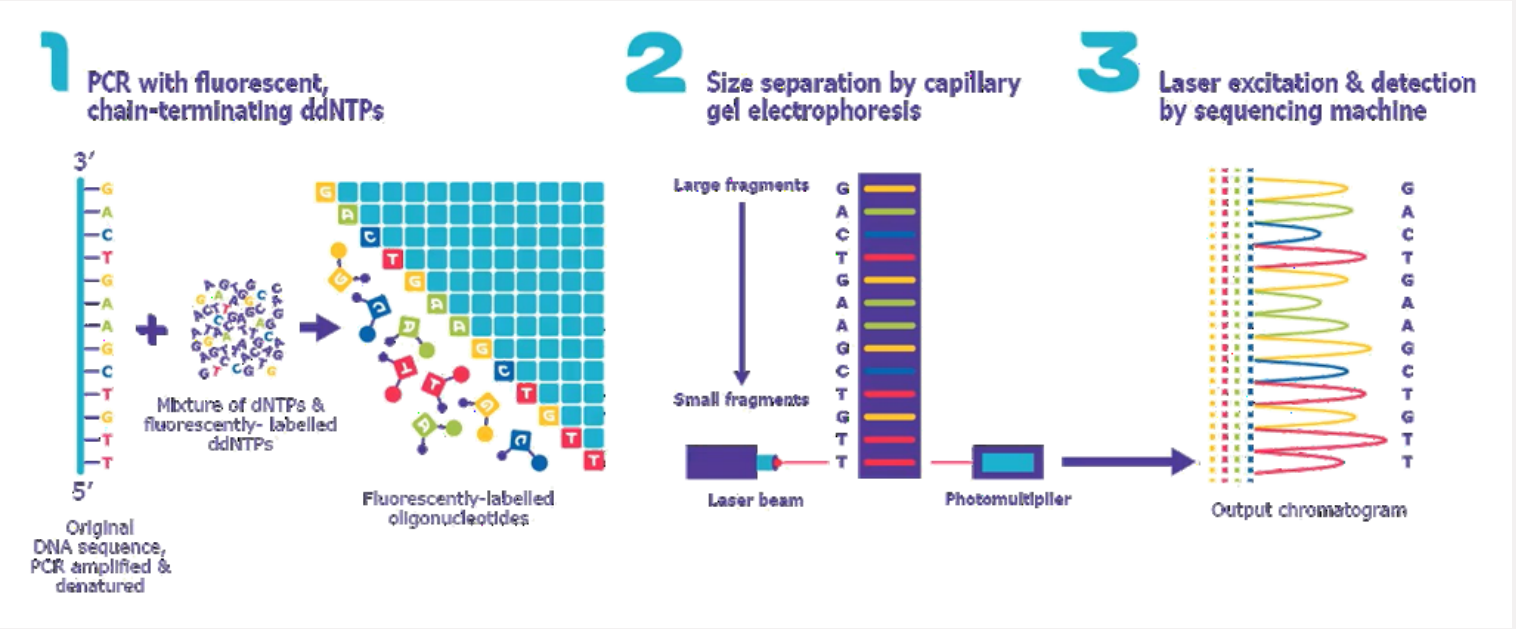
MAXAM-GILBERTOVA METODA

Vysoce toxické chemikálie (radioaktivní značení 5′ konce DNA pomocí ^{32}P), štěpení DNA činidly, gelová elektroforéza, autoradiografie

SANGEROVA METODA

Založená na selektivním začleňování dideoxyribonukleotidů (ddNTP) – přerušení syntézy DNA.

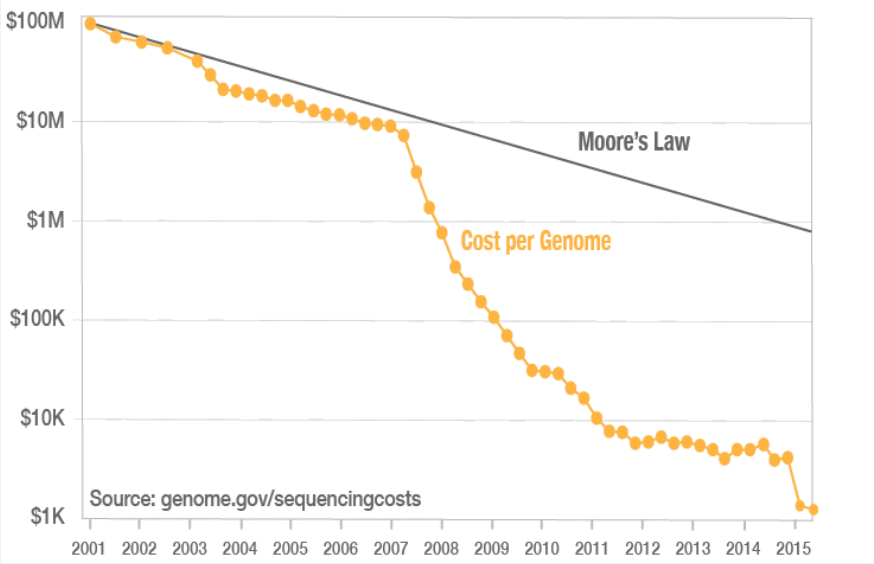
Vzorek DNA je rozdělen do 4 oddělených reakcí, které obsahují všechny standardní deoxynukleotidy. Ke každé reakci je přidán pouze jeden ze čtyř dideoxynukleotidů (ddATP, ddGTP, ddCTP, nebo ddTTP). Dochází k syntéze komplementárního řetězce do začlenění značených ddNTP. Fragmenty jsou separovány pomocí GE a analyzován fluorescenční signál (každý ddNTP jiná barva).



Sekvenování nové generace (NGS)

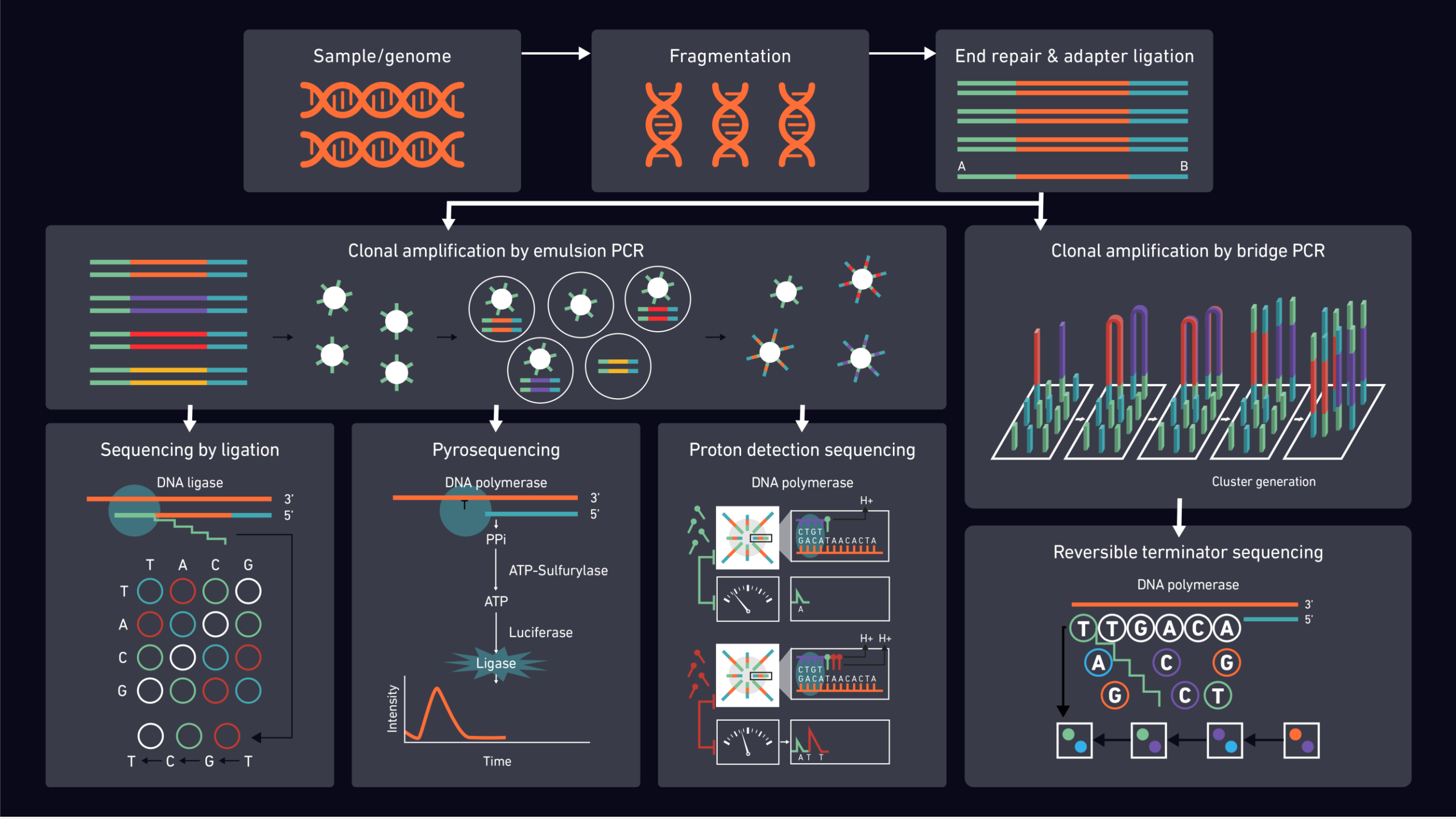
- Miniaturizace a paralelizace sekvenačních technologií
- Analýza mnoha molekul/fragmentů najednou
- Rychlejší než tradiční metody

- Snížené náklady na sekvenování (automatizace)
- Vyšší přesnost sekvenování
- Fragmentace DNA – 50 až 500 bazí
- **Několik technologií (každá má výhody i omezení a je vhodná pro různé typy**



Metoda	Sangerova	Illumina
Kapacita (bp / hod)	76 000	1 800 000 000
Cena (€ / Gbp)	1 250 000	50

Náklady na sekvenování lidského genomu. Odklon křivky nákladů na sekvenování od Moorova zákona se shoduje s nástupem sekvenování nové generace (NGS). Moorův zákon pochází z odvětví počítačového hardwaru, který zahrnuje zdvojnásobení "výpočetního výkonu" každé dva roky. Má se za to, že technologie, které se řídí tímto zákonem, jsou považovány za úspěšné. Představuje tedy užitečný vztah pro porovnávání technologického pokroku.


Sample/genome
Fragmentation
End repair & adapter ligation
A
B
Clonal amplification by emulsion PCR
Clonal amplification by bridge PCR
Cluster generation
Sequencing by ligation
DNA ligase
3'
5'
T A C G
T
A
C
G
T ← C ← G ← T
Pyrosequencing
DNA polymerase
3'
5'
PPi
ATP-Sulfurylase
ATP
Luciferase
Ligase
Intensity
Time
Proton detection sequencing
DNA polymerase
H+
CTGT
GACATAACACTA
A
H+ H+
CTGT
GACATAACACTA
A T T
Reversible terminator sequencing
DNA polymerase
3'
5'
T T G A C A
A C G
G C T

Illumina

- Dříve Solexa
- Momentálně nejrozšířenější technologie masivně paralelního sekvenování
- Princip sekvenování syntézou za pomoci reverzibilních terminátorů
- Klonální amplifikace (můstková PCR)
- 99% přesnost

\+ vysoká přesnost

nejnižší cena za jednu osekvenovanou bázi

množství publikací využívající Illumina technologii

množství komerčně dostupných kitů pro různé aplikace

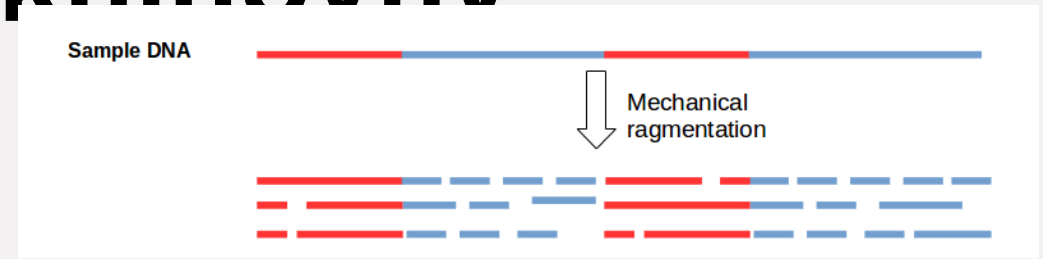
\- vysoké pořizovací náklady

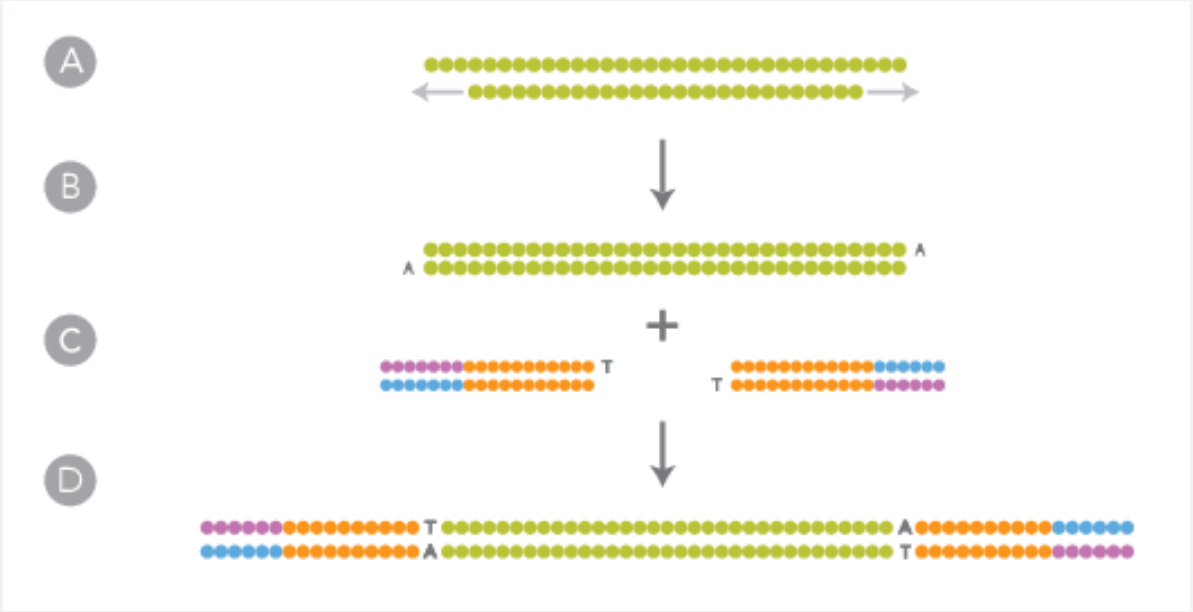
krátká délka čtení – max 300 bazí – pouze některé přístroje

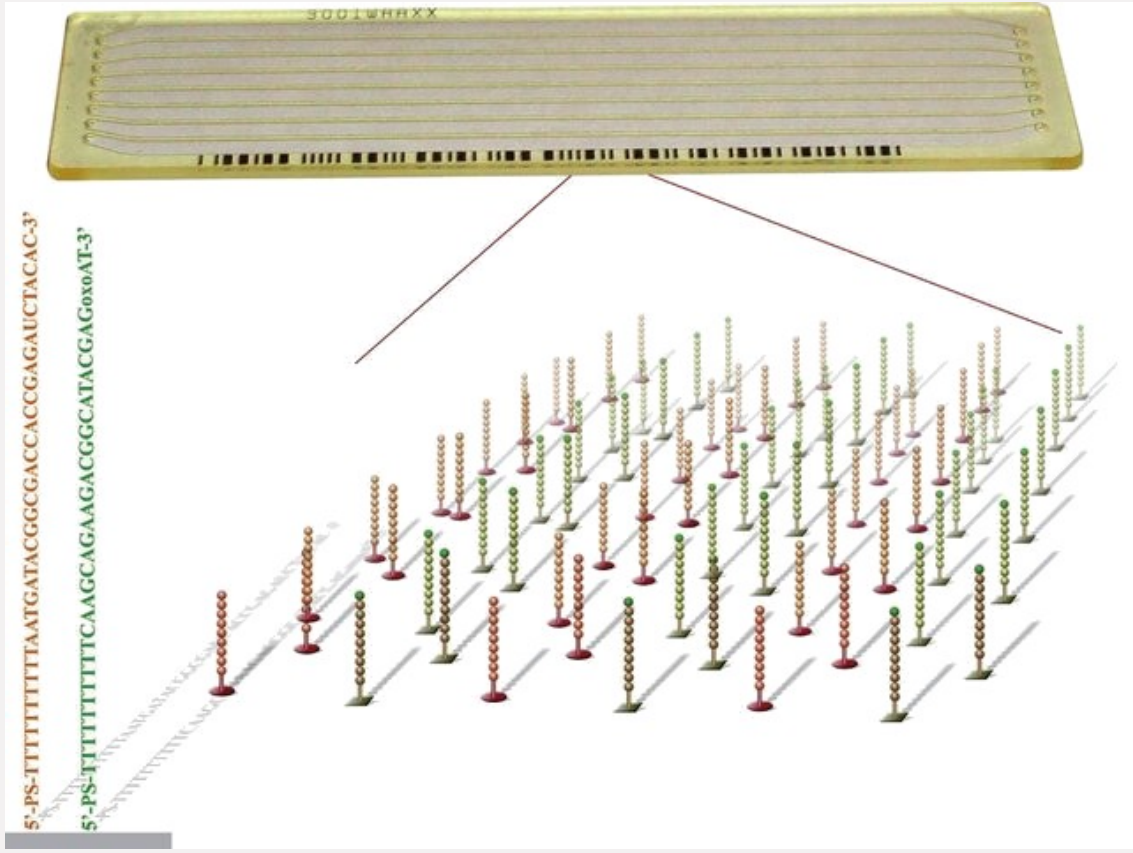
komplikovaná de novo assembly z Illumina NGS dat

dlouhá doba sekvenování 12h až 4 dny

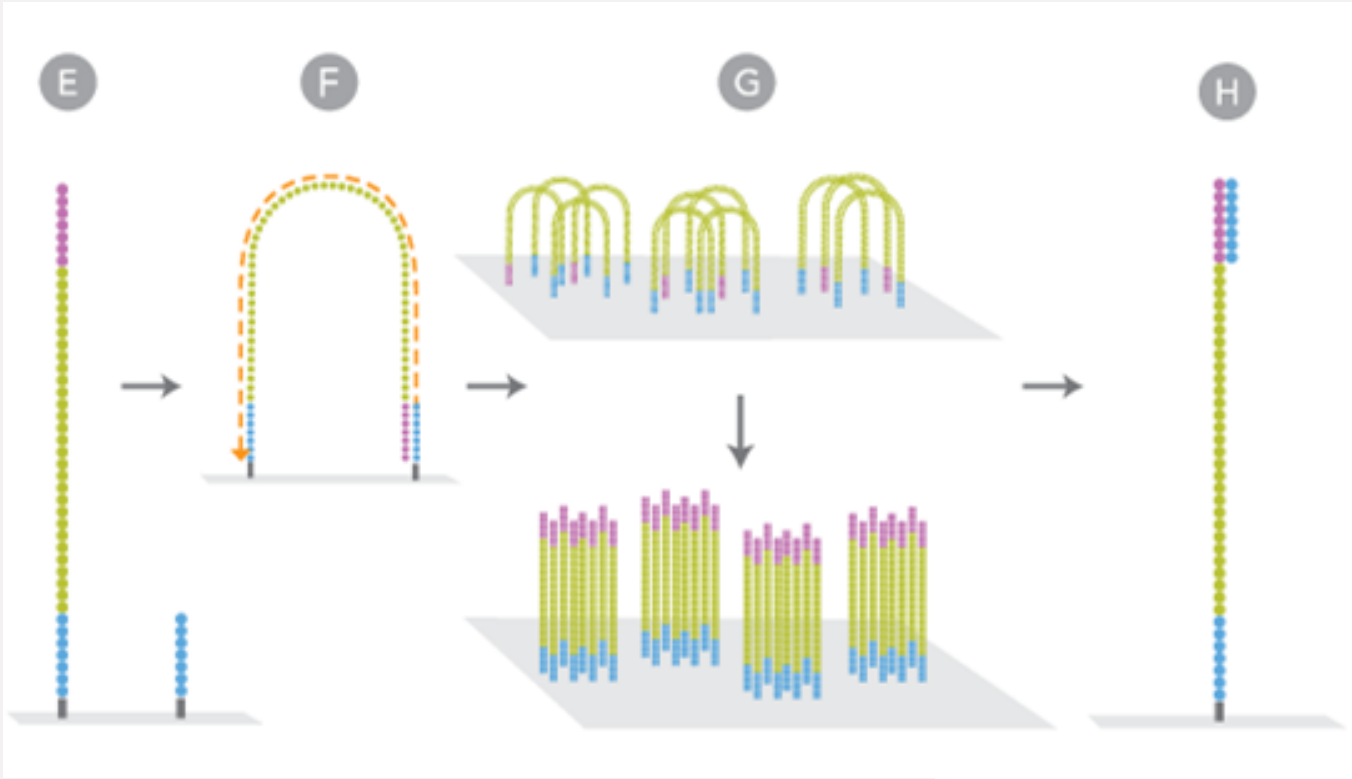
Illumina – příprava sekvenační knihovny

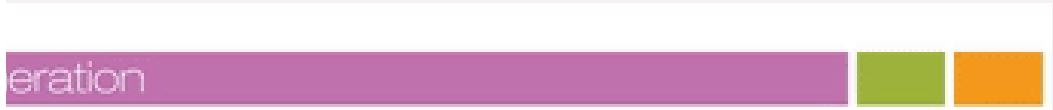


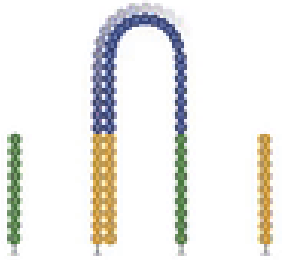




Illumina – amplifikace

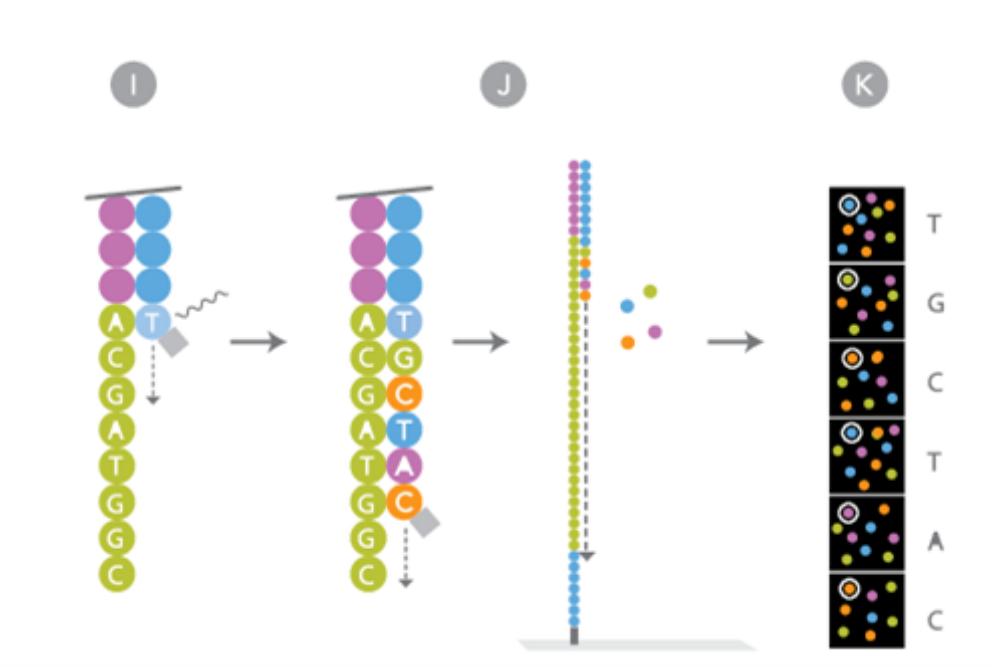


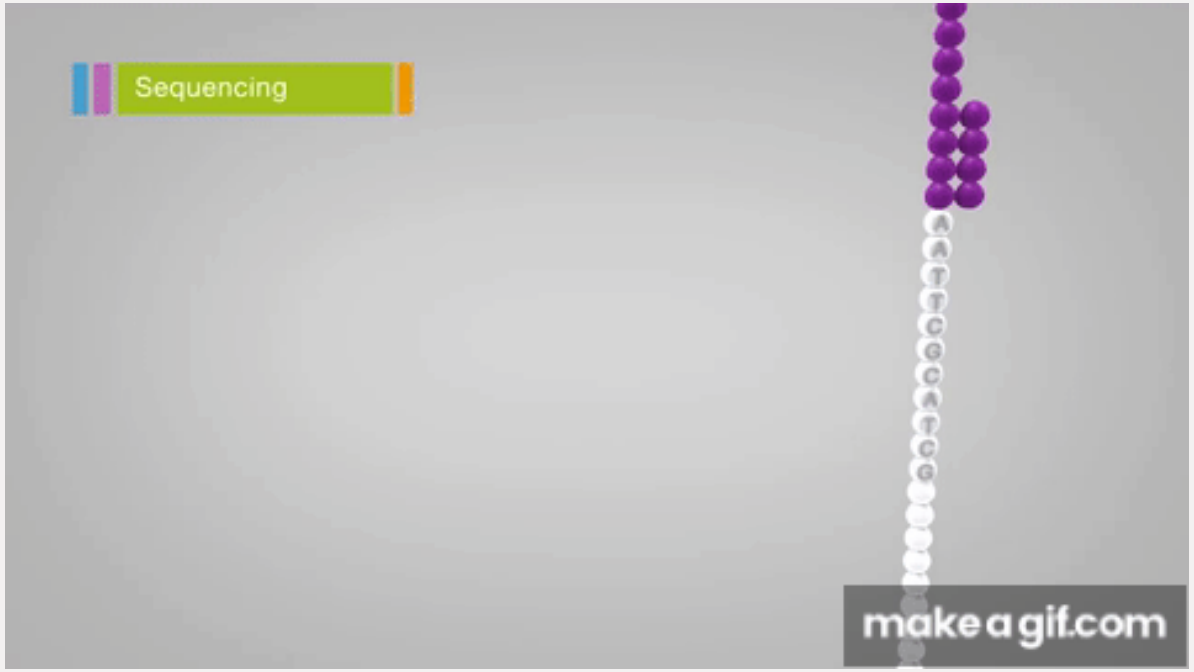




Fragmenty jsou amplifikovány pomocí můstkové PCR, dochází k tvorbě klastrů (jeden fragment = jeden klastr, obsahuje tisíce až miliony kopií)

Illumina – sekvenace





Bioinformatická analýza NGS dat

1 Primární

Konverzi surových přístrojových dat na sekvenční data sestávající z pořadí nukleotidových bází. Často prováděna přímo na sekvenátoru.

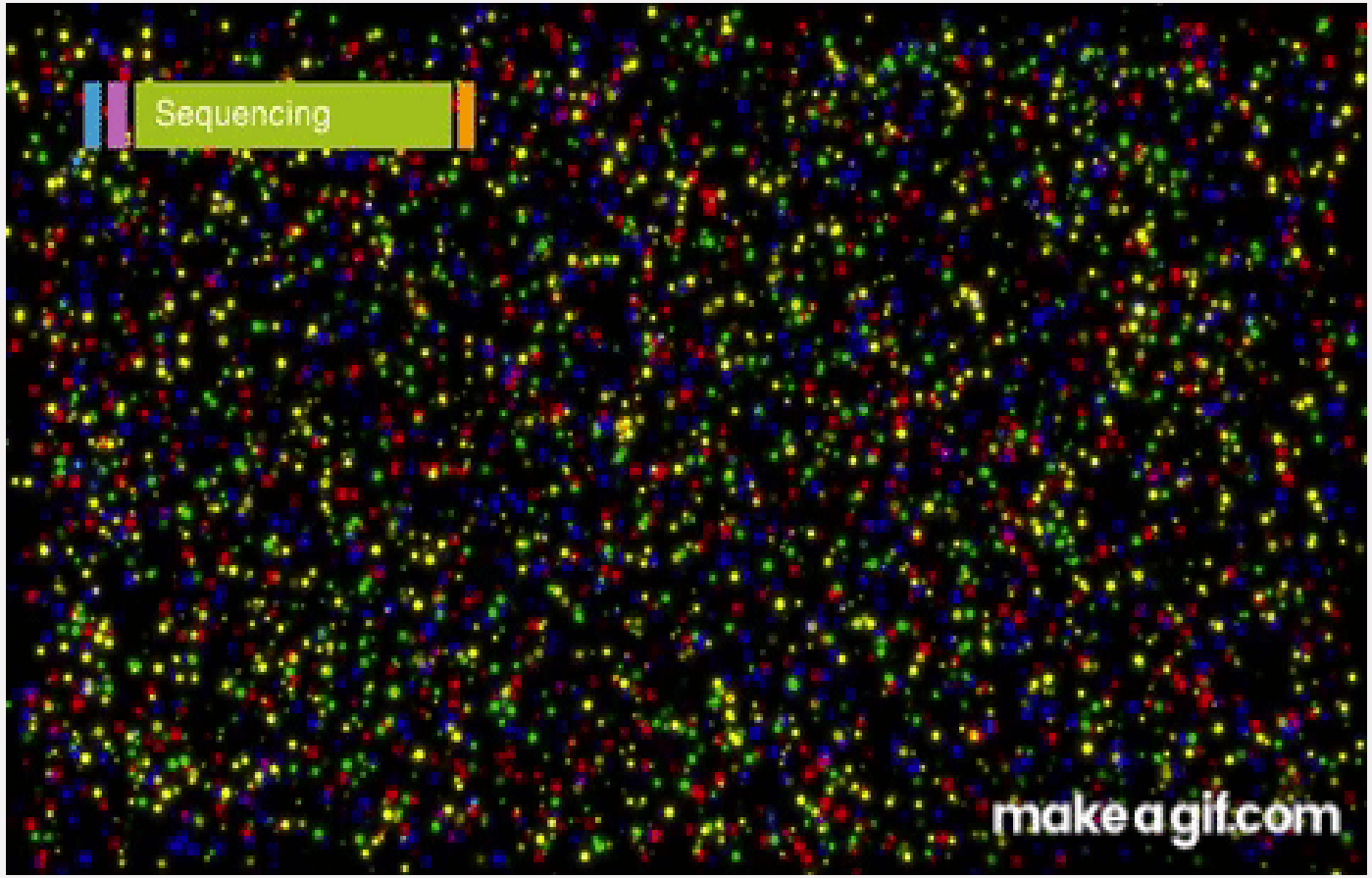
2 Sekundární

Sestavení genomu/ části genomu a detekce variací.

3 Terciární

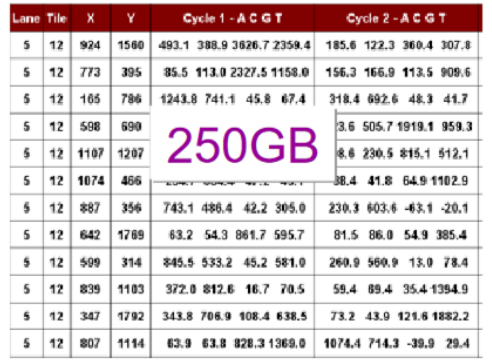
Anotace získaných dat, biologický kontext.

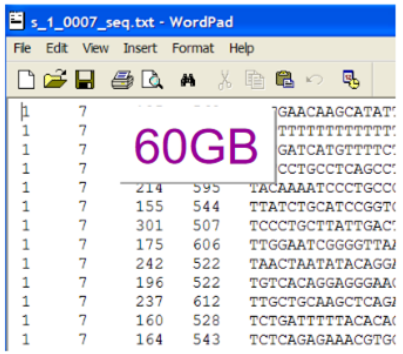
Primární analýza dat



4,8 TB

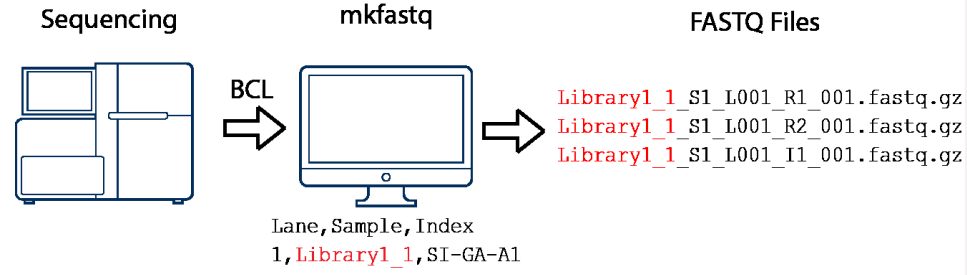
Surová obrazová data se převedou do tabulky intenzity, která zaznamenává umístění každého klastru a hodnoty intenzity barev (jedna barva pro každou ze čtyř bází). Tyto číselné hodnoty se převedou na tzv. base call a použijí se k přiřazení hodnoty kvality pro dané pořadí bazí.

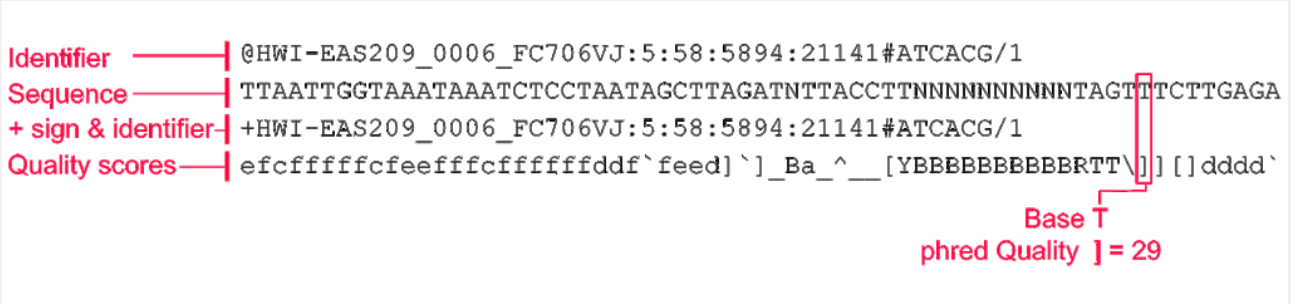




* Každý z miliard klastrů vytvoří 150 bp čtení

Primární analýza dat



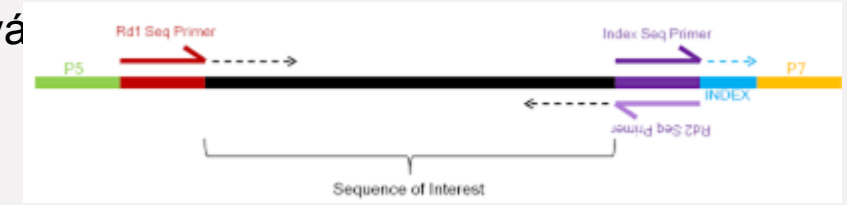


Pre-procesování dat

Filtrování: Čtení jsou z dat filtrována na základě kvality base call a délky čtení. Báze s nízkou intenzitou mohou vést k detekci falešně pozitivních variant, proto je třeba je odstranit. Čtení, která jsou příliš krátká, se pravděpodobně zarovnávají k více oblastem v genomu a způsobují špatné mapovací metriky.

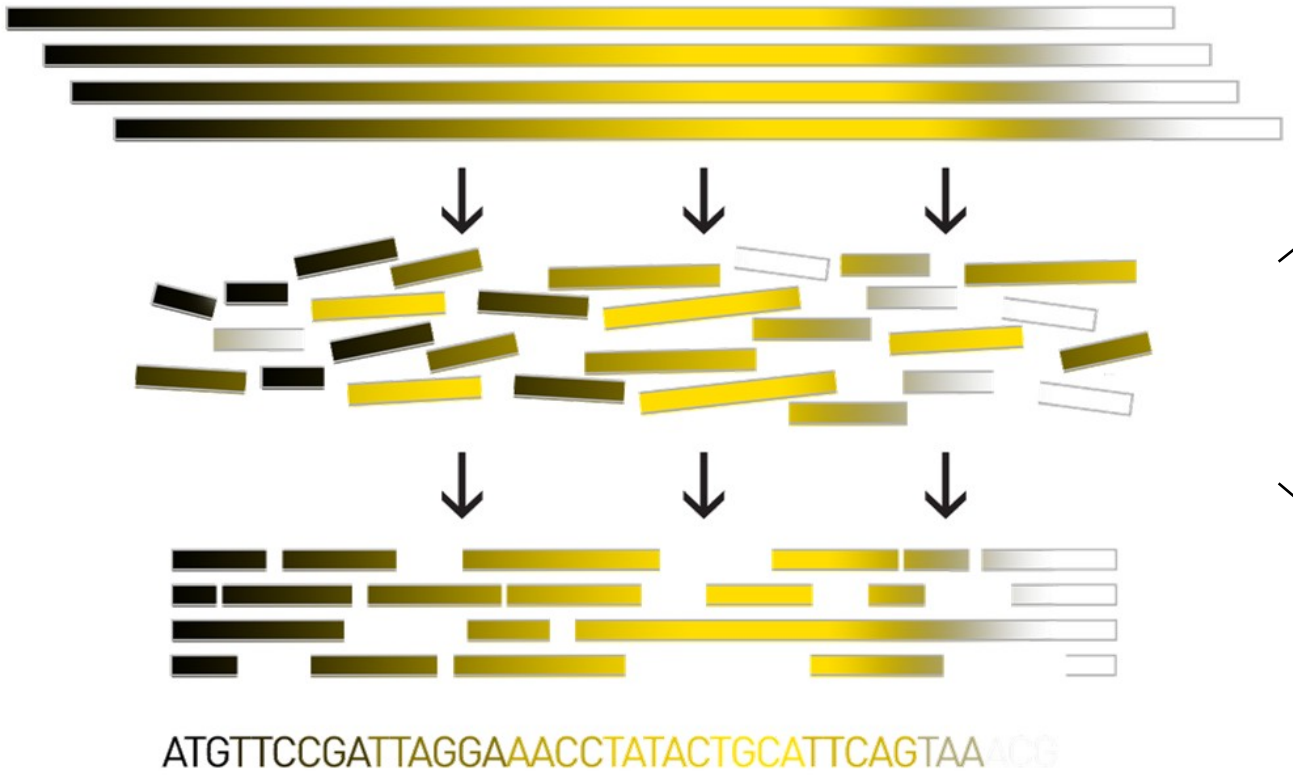
Demultiplexování: Multiplexování v NGS znamená sekvenování více vzorků současně na stejném přístroji. Demultiplexování znamená rozdělení sekvenačních čtení do samostatných souborů podle indexu „čárového kódu“ použitého pro každý vzorek.

Ořezávání: Adaptorové sekvence ligované na konce knihoven během procesu přípravy knihovny je třeba ze sekvenačních čtení odstranit, protože mohou narušovat mapování a sestavová



Sekundární analýza dat

"Imagine a book cut by scissors into 10 million small pieces. Assuming that 1 million pieces are lost and the remaining 9 million are splashed with ink… try to recover the original text!" [P.Pevzner, UCSD]

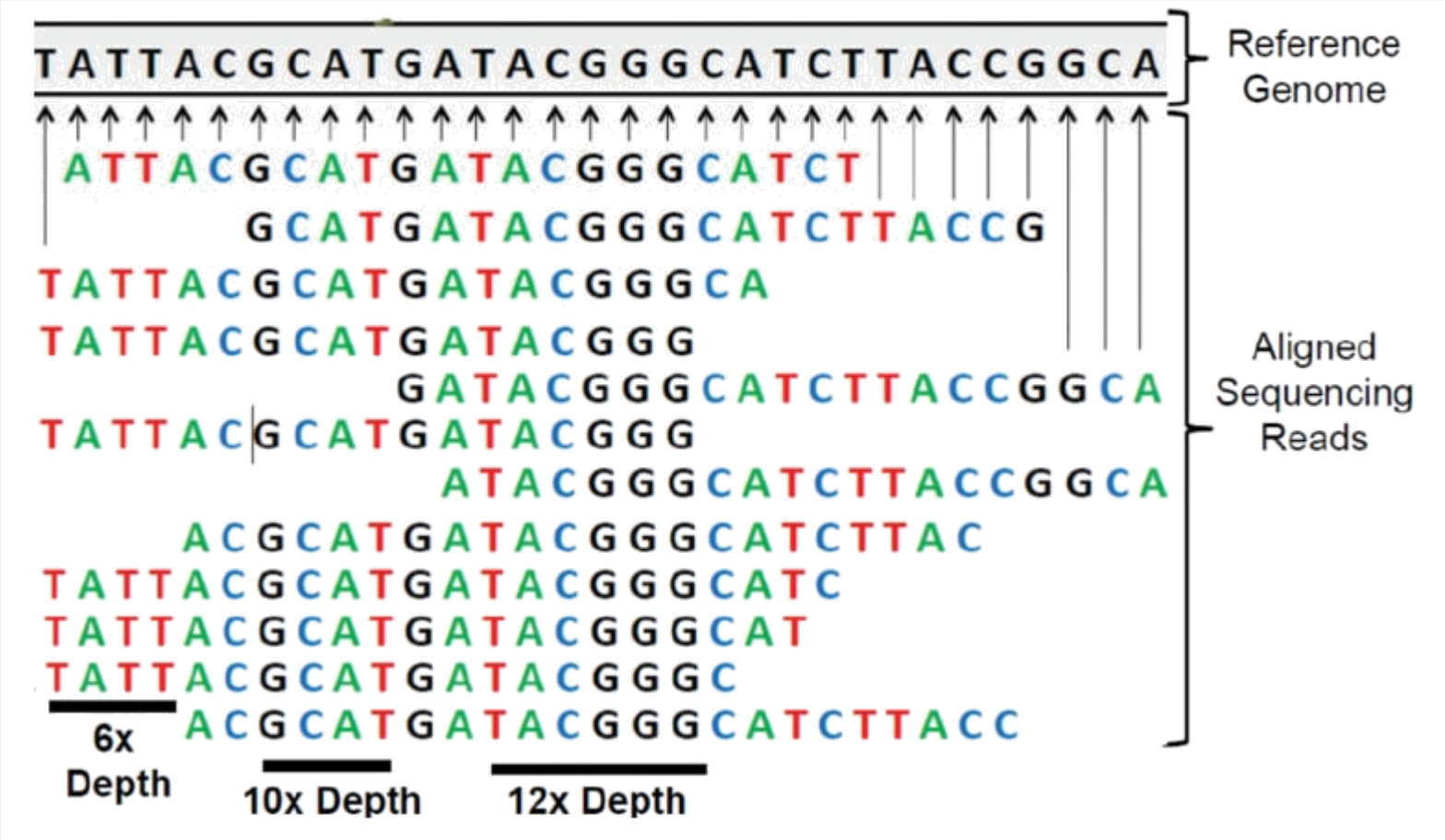


De novo sestavení

Čtení se zarovnávají navzájem na základě jejich sekvenční podobnosti, aby se vytvořila dlouhá konsenzuální sekvence nazývaná kontig.

Sestavení na základě referenčního

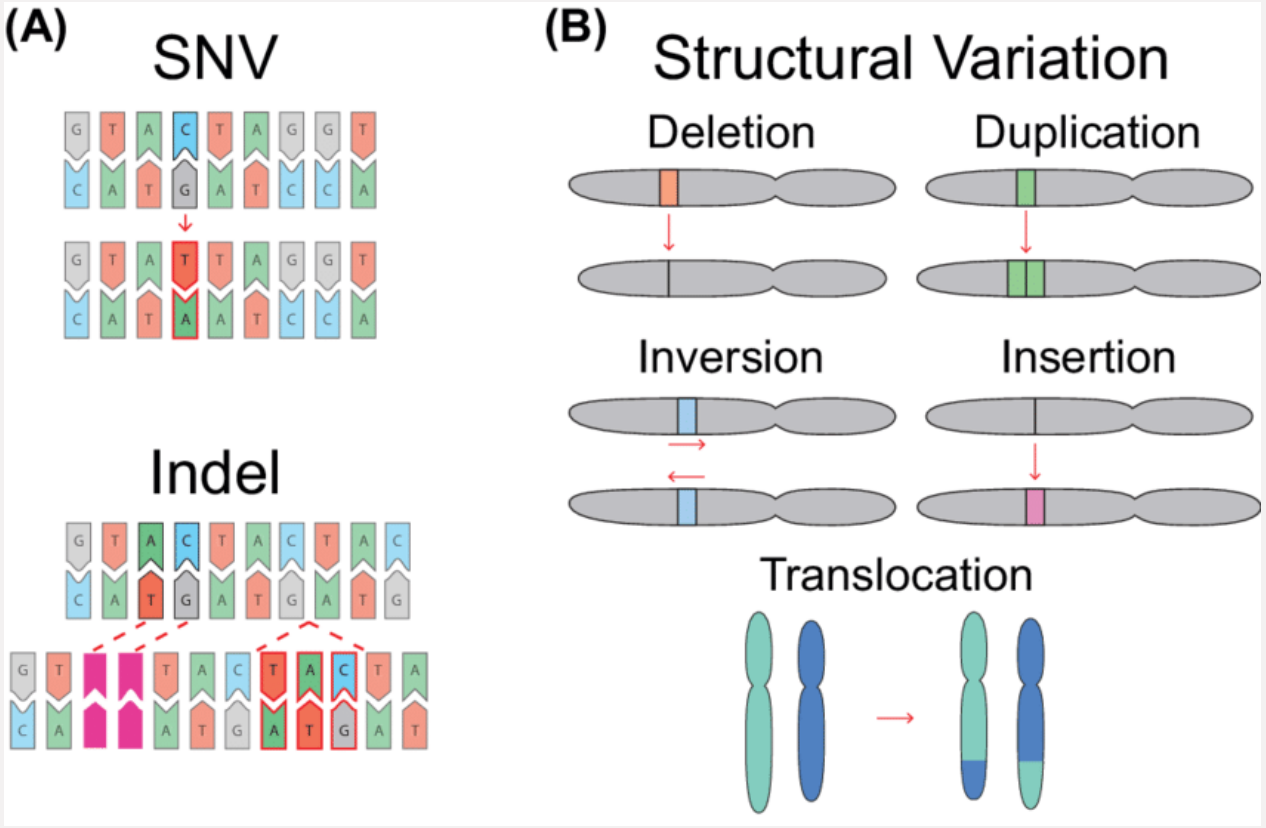
Sekundární analýza dat



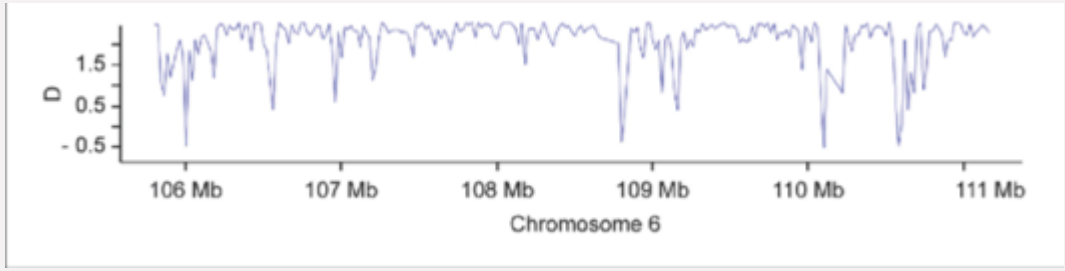
Cílem zarovnání sekvencí je najít místo, odkud čtení pochází, a určit, kolik čtení se k tomuto místu zarovnalo.

Sekundární analýza dat

Identifikace variací



Referenční genom je standard, dokážeme identifikovat odchylky.



Strukturní variace na úrovni chromozomu

Sekundární analýza dat

Používané formáty souborů

FILE TYPE	DESCRIPTION	WHERE IT IS USED
FASTQ	Text-based file format containing raw sequence reads and the associated quality score of each base	Storage of raw sequence data and input into sequence alignment
BED	Browser Extensible Data file is a tab-delimited text file that is used to store genomic regions as coordinates	In variant calling pipelines to direct the analysis to a genomic region
SAM	Sequence Alignment Map file, used to store text-based information for reads aligned to a reference sequence	Store information on read alignment, e.g. position and quality
BAM	Binary Alignment Map file is a compressed binary version of a SAM file. Can be opened in genome browsers to view read alignment	Used for input into variant calling pipelines
VCF	The Variant Call Format is a text file which stores sequence variants, each variant occupies a single row	Generated by variant calling pipelines. Used as input into variant annotation

Terciální analýza dat

Biologický kontext získaných dat

Anotace variant je proces předpovídání biologického vlivu nebo funkce genetických variant. Využívají se anotační nástroje, které pracují s VCF soubory, výstupem je zpráva o anotovaných variantách a jejich biologickém účinku.

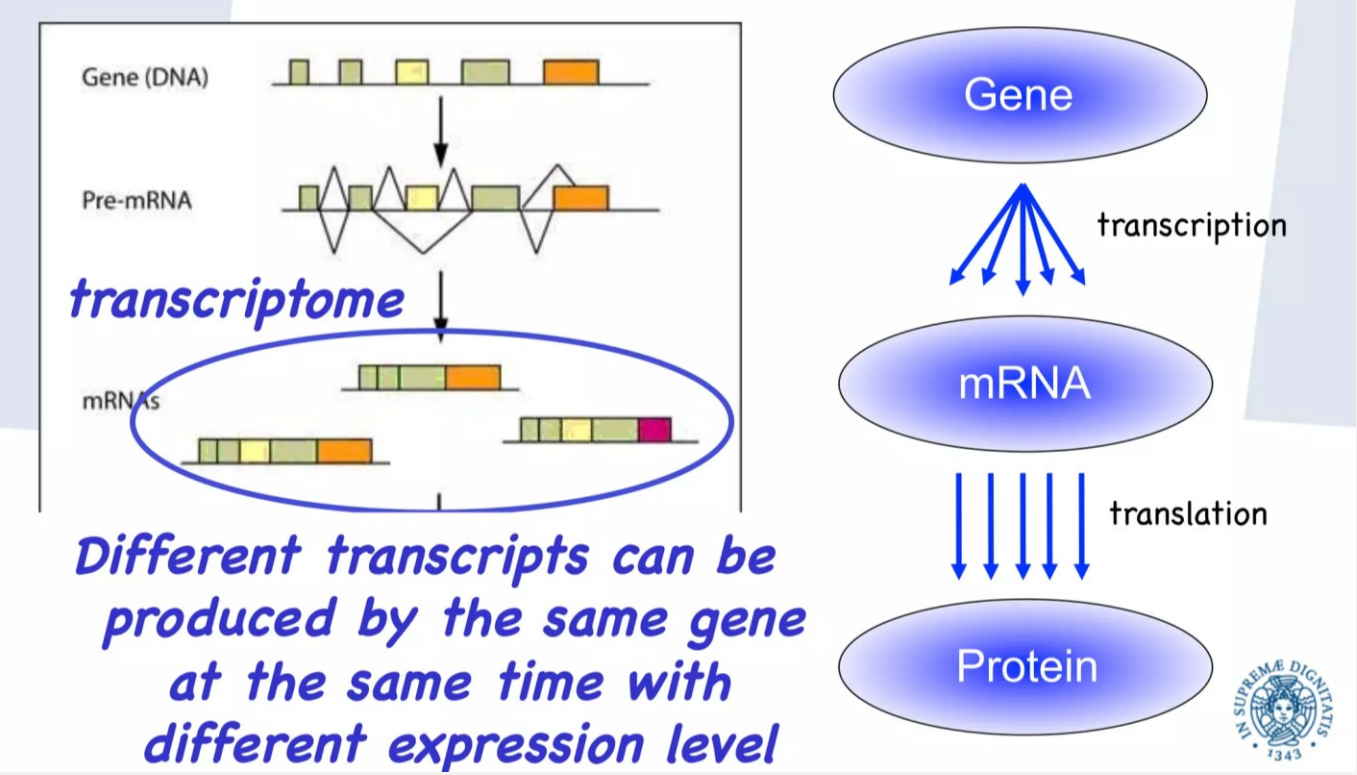
Interpretaci variant obvykle provádí kvalifikovaná osoba, například klinický genetik. Jejich práce zahrnuje shromáždění všech dostupných informací o pacientovi, včetně rodinné anamnézy onemocnění, a porovnání genotypu pacienta s klinickým fenotypem.

Součástí je také **statistická** analýza a **vizualizace** dat.

Další aplikace NGS

Genová exprese

Proces, kterým je v genu uložená informace převedena v reálně existující buněčnou strukturu (RNA) – množství transkriptu je úměrné míře genové exprese genů. Jedná se o reakci organismu na vnější i vnitřní vlivy.



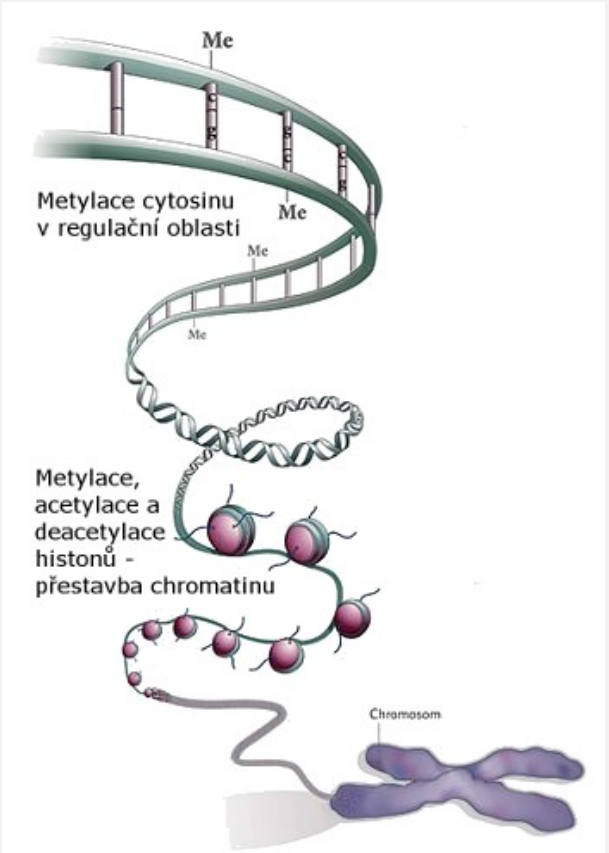
U sekvenace RNA se provádí zpětný přepis do DNA, tzv. reverzní transkripce.

Další aplikace NGS

Epigenetika

Změny v genové expresi, které nejsou způsobeny změnou nukleotidové sekvence DNA.

Metylace DNA, acetylace histonů, **mikroRNA**, …



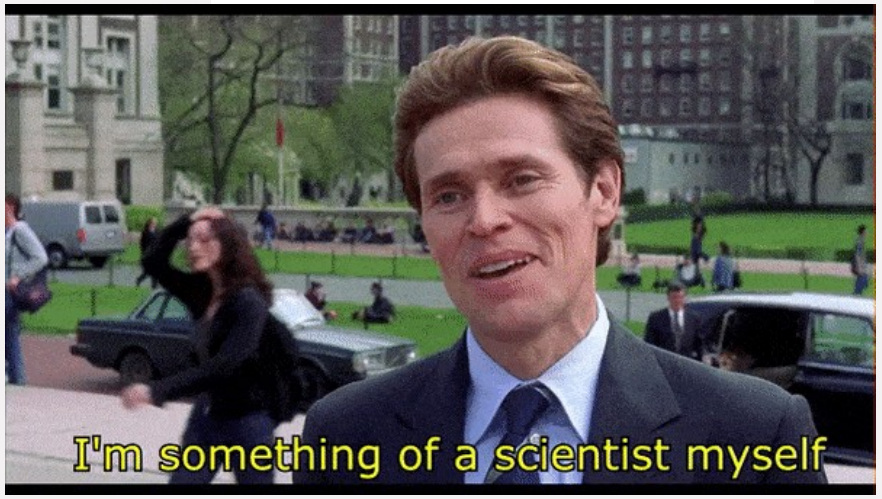
Journal of Cellular and Molecular Medicine
Open Access

Short Communication | Open Access

Genome-wide identification of urinary cell-free microRNAs for non-invasive detection of bladder cancer

Jaroslav Juracek, Barbora Peltanova, Jan Dolezel, Michal Fedorko, Dalibor Pacik, Lenka Radova, Petra Vesela, Marek Svoboda, Ondrej Slaby, Michal Stanik

First published: 24 January 2018 | https://doi.org/10.1111/jcmm.13487 | Citations: 28



Shrnutí na závěr

- Genomická data jsou unikátní a vysoce citlivá data, která do určité míry předpovídají délku života a dispozice k onemocněním jedince
- Znalost primární struktury (pořadí nukleotidů) umožňuje získat informace o umístění, struktuře i funkci genů a dalších oblastí DNA
- Zlatá metoda analýzy primární struktury NK je Sekvenování nové generace (NGS)
- Součástí procesu je jak práce v laboratoři (wet lab), tak bioinformatické zpracování dat
- NGS generuje velké množství dat (TB / genom)

DĚKUJI ZA POZORNOST!

☺

juracek@ics.muni.cz