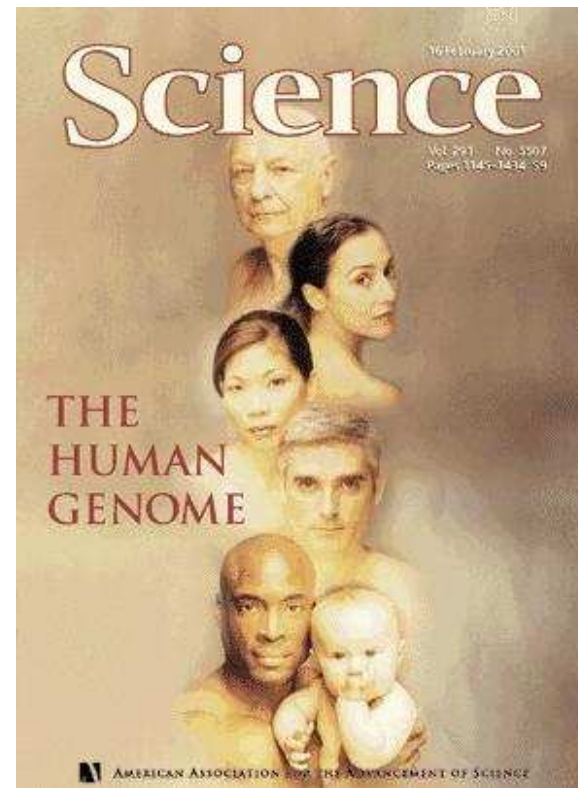
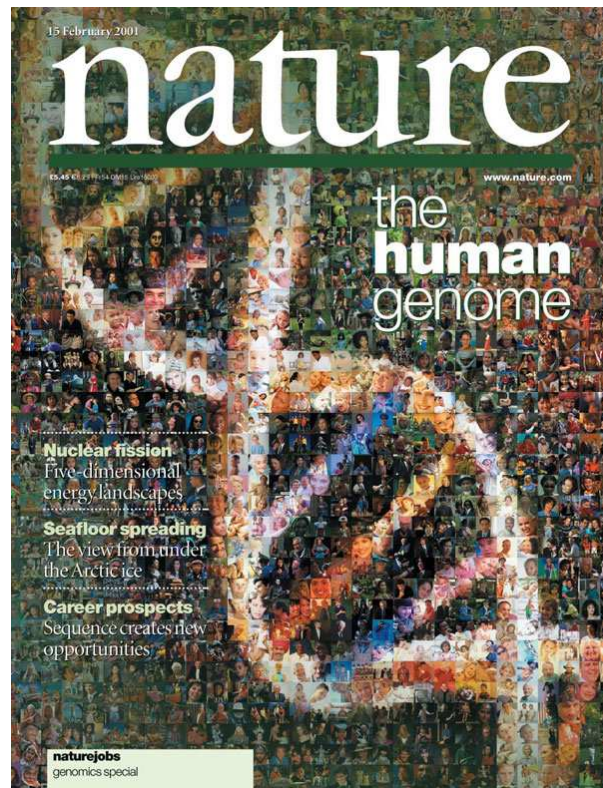


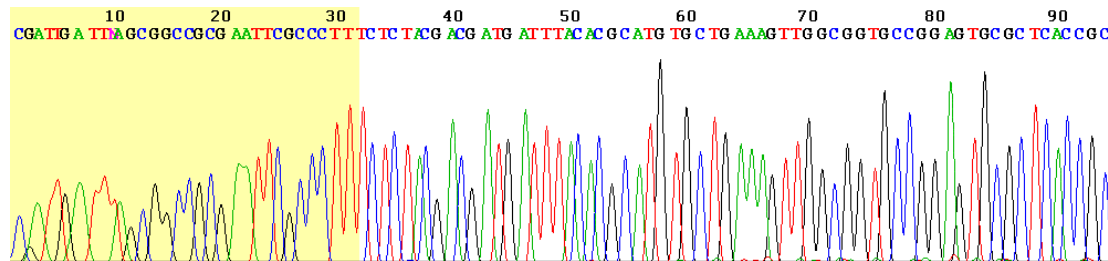
# Sekvenovanie a zostavovanie genómov (genome sequencing and assembly)

Tomáš Vinar̃  
27.9.2018



## Sangerovo sekvenovanie

- Výsledok: sekvenovací profil (trace)



- Ďalej sa spracuje pomocou programu PHRED:
  - Na každej pozícii (kde sa dá) určí bázu (A,C,G,T)
  - Pre každú bázu odhadne kvalitu  $q$   
( $10^{-q/10}$  je pravdepodobnosť chyby,  
t.j. bázy s kvalitou  $q > 40$  sú správne na 99.99%)
- Sangerovo sekvenovanie produkuje čítania (reads)  
dlhé 500-1000 bp
- Ako osekvenovať dlhú DNA sekvenciu?

## Typický priebeh sekvenovania

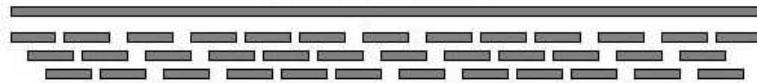
1. Chromozómy náhodne rozsekáme na menšie kúsky  
(napr. pomocou sonikácie)
2. Menšie kúsky namnožíme  
(napr. pomocou PCR, bakteriálneho klonovania a pod.)
3. Konce týchto kúskov osekvenujeme niektorou zo sekvenovacích technológií  $\Rightarrow$  mnoho krátkych reťazcov, ktoré nazývame **čítania**
4. Čítania **výpočtovo zostavíme** späť do chromozómov

## Prehľad sekvenovacích technológií

Technológia	Dĺžka čítania	Chybovosť	Za deň	Cena za MB
<b>1. generácia</b>				
Sanger	do 1000 bp	< 2%	3 MB	\$4000
<b>2. (next) generácia (cca 2004)</b>				
Illumina	2× 150bp	< 1%	30 GB	< \$1
<b>3. generácia (práve nabieha)</b>				
PacBio	6-25 Kbp	10%	1 GB	\$2
Oxford Nanopore	5-100+ kbp	10%	1 GB	\$1

## Bioinformatický problém: Zostavenie genómu (sequence assembly)

- **Vstup:** krátke čítania sekvenovanej DNA
- **Cieľ:** zostaviť pôvodnú DNA
  - riadime sa zhodou v prekrývajúcich častiach čítaní
- Dôležité faktory:
  - **dĺžka genómu**
  - **pokrytie** (coverage) – koľko krát čítania pokrývajú genóm?



## Formulácia problému

### Najkratšie spoločné nadslovo (shortest common superstring)

**Úloha:** Daných je niekoľko reťazcov (čítaní), nájdite **najkratší** reťazec, ktorý obsahuje **všetky** vstupné reťazce ako (súvislé) **podreťazce**.

Motivácia: čo najviac využiť prekryvy medzi čítaniami

### Príklad:

Vstup: GCCAAC, CCTGCC, ACCTTC

Výstup: CCTGCCAACCTTC (najkratšie možné)

## Najkratšie spoločné nadslovo

- **Problém je NP ťažký**

takže nepoznáme rýchly algoritmus, ktorý vždy nájde najlepšie riešenie

- **Jednoduchá heuristika:** opakovane nájdí dva reťazce, ktoré sa prekrývajú najviac a zlúč ich do jedného reťazca

- Príklad: CATATAT, TATATA, ATATATC

Optimum: CATATATATC, dĺžka 10

Heuristika: CATATATCTATATA, dĺžka 14

- V skutočnosti táto heuristika **aproximačný algoritmus:**

Nájsené riešenie je najviac  $3,5\times$  horšie ako optimálne

T.j. je to 3,5-aproximačný algoritmus

(možno aj 2-aproximačný, otvorený problém)

- Existuje aj 2,5-aproximačný algoritmus

## Najkratsie spoločné nadslovo: Čo sme nezahrnuli do formulácie

- V sekvenovaní sa vyskytujú chyby (cca 1 zo 100 báz)
- Polymorfizmus
- Orientácia čítaní (vlákno, strand)
- Kontaminácia cudzou sekvenciou (napr. baktérie, v ktorých sa segmenty DNA klonovali), chiméry
- Viac chromozómov, neúplné pokrytie čítaniami
- Repetitívna sekvencia (sequence repeats, opakovania)

cca 50% ľudského genómu

Príklad: 10xTTAATA, 10xATATTA, 3xTTAGCT

TTAATATTAGCT?

TTAATATTAATATTAATATTAATATTAGCT?

TTAATATTA + ATATTAGCT?



## Najkratšie spoločné nadslovo: Zľahčujúce faktory

**Prídavná informácia:** spárované čítania (Pair-end reads)



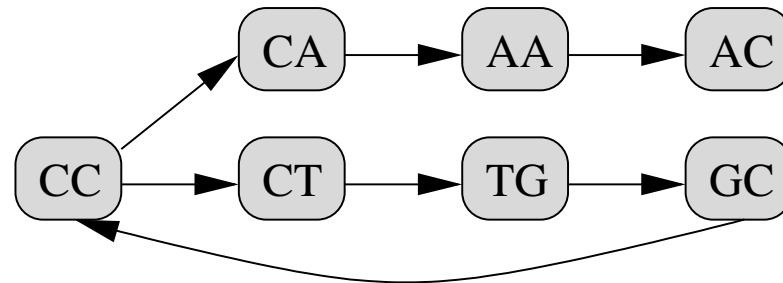
**Zjednodušenie:** nemusíme spojiť všetko do jedného reťazca, spájame len časti spojené viacerými čítaniami  $\Rightarrow$  konzervatívny prístup (radšej menej pospájať, ale nerobiť chyby)

## Najkratšie spoločné nadslovo: Zhrnutie

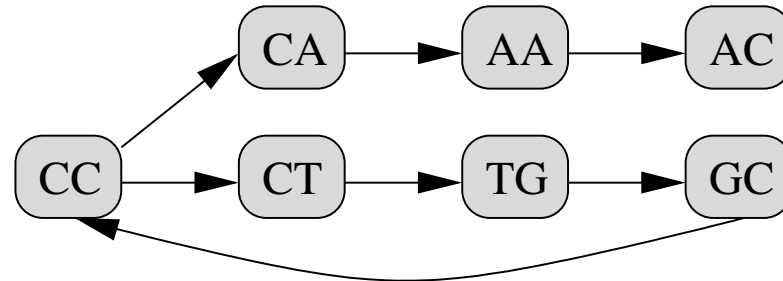
- Nerealistická formulácia, ťažký výpočtový problém
- Ale teoretický problém môže poskytnúť nejaký posun k pochopeniu skutočného problému
- Overlap-Layout-Consensus prístup motivovaný greedy algoritmami pre najkratšie spoločné nadslovo

## Skladanie krátkych čítaní: de Bruijnové grafy

- Nasekajme čítania na (prekrývajúce sa) kúsky dĺžky  $k$
- Zostavme z nich **de Bruijnov graf**
  - **vrcholy**: podreťazce dĺžky  $k$  všetkých čítaní
  - **hrany**: nadväzujúce  $k$ -tice v rámci každého čítania (s prekryvom  $k - 1$ )
  - Graf je orientovaný (hrany majú smer)
- **Príklad**:  $k = 2$ , čítania: CCTGCC, GCCAAC



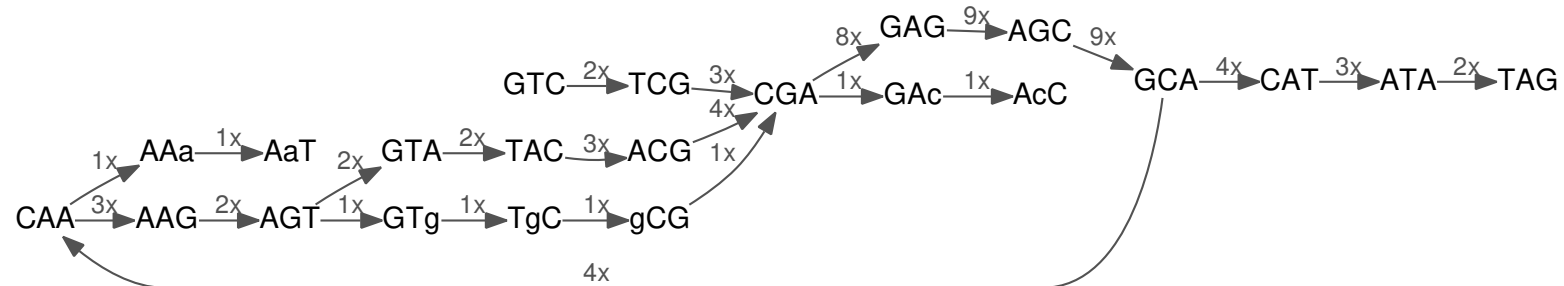
## Ako použiť de Bruijnovu grafy?



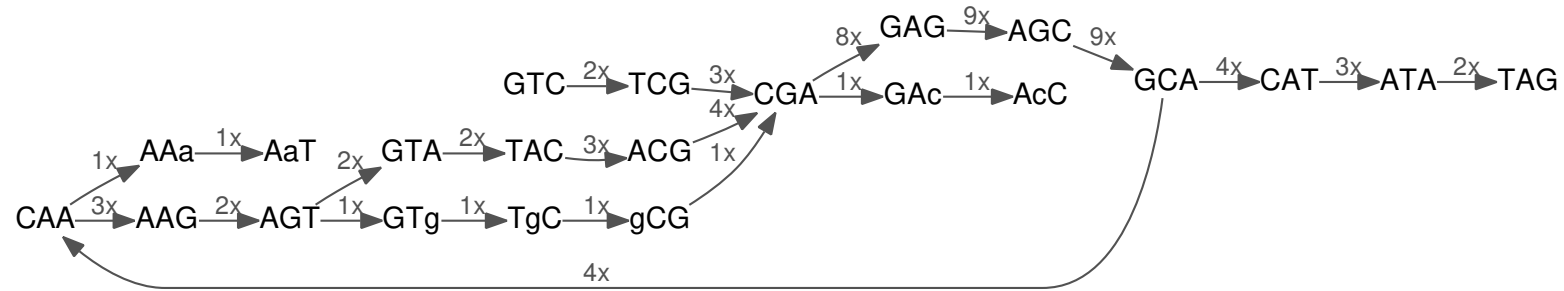
- jediný chromozóm a žiadne “nejednoznačné”  $k$ -tice  
⇒ zostavenie = **Eulerovská cesta**  
(cesta v grafe, ktorá použije každú hranu práve raz)
- Eulerovskú cestu možno nájsť v čase  $O(m + n)$
- v realistickom prípade:  
zostavenie genómu zodpovedá niekoľkým  
**pochôdkam v de Bruijnovom grafe** (nazývame **kontigy**),  
ktoré dohromady pokrývajú veľkú časť hrán

## Príklad: sada čítaní a zodpovedajúci deBruijnov graf

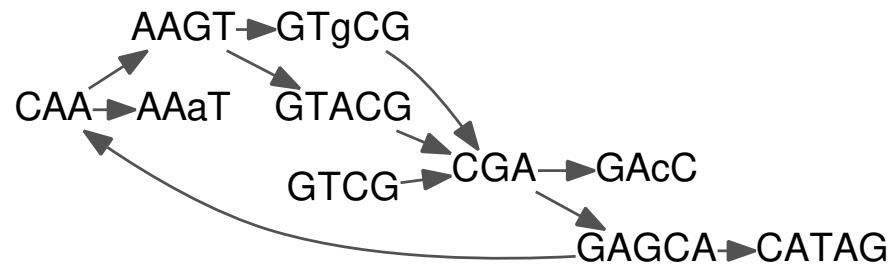
GTCGAGCAAGTACGAGCATAG  
 TCGAGCA AGCATAG  
 AGCAAaT AGCATAG  
 GTCGAcC GTACGAG  
 GTCGAGC TACGAGC  
 CGAGCAA ACGAGCA  
 AGTgCGA  
 CAAGTAC  
 GCAAGTA GAGCAT  
 GAGCAAG GAGCATA  
 TACGAGC



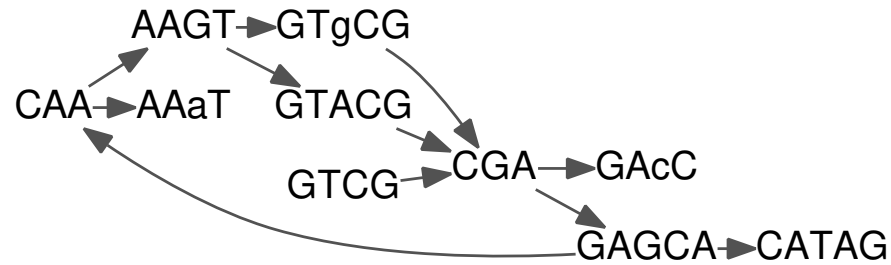
## Príklad: zjednodušovanie de Bruijnovho grafu



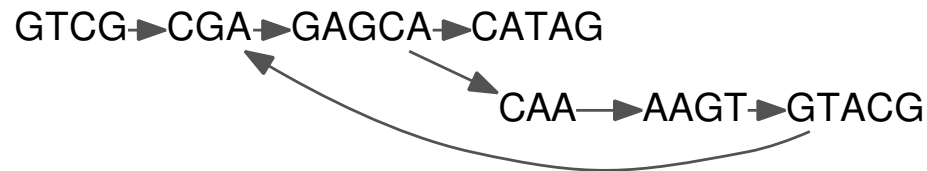
Spojíme jednoznačné cesty do vrcholov



## Príklad: odstraňovanie chýb z de Bruijnovho grafu



Odstránenie chýb (výbežkov a bublín s nízkym pokrytím)



Spájaním dostaneme 4 kontigy (pôv. GTCGAGCAAGTACGAGCATAG)



## Typické výsledky zostavovania genómov

- Veľa **kratších kontigov**, ktoré možno ďalej kombinovať do väčších celkov (**scaffolds**) pomocou ďalšej informácie (napr. spárované čítania, čítania 3. generácie)
- Niektoré časti nemožno jednoznačne zostaviť z dôvodu **dlhých opakujúcich sa sekvencií**

**Príklad:** človek chr14, 88 Mbp, 70× pokrytie

Metóda	Počet kontigov	Chýb	N50 po korekcii
Velvet (základný de Bruijn)	>45000	4910	2.1 kbp
Velvet (scaffolding)	3565	9156	27 kbp
AllPaths-LG	225	45	4.7 Mbp

N50: kontigy s touto alebo dlhšou dĺžkou pokrývajú 50% genómu

korekcia: rozsekne všetky zle spojené kontigy



## História sekvenovania genómov

- 1976 MS2 (RNA vírus) 40 kB
- 1988 projekt sekvenovania ľudského genómu (15 rokov)
- 1995 baktéria H. influenzae 2 MB, shotgun (TIGR)
- 1996 S. cerevisiae 10 MB, BAC-by-BAC (Belgicko, Británia)
- 1998 C. elegans 100 MB, BAC-by-BAC (Wellcome Trust)
- 1998 Celera: ľudský genóm do troch rokov!
- 2000 D. melanogaster 180 MB, shotgun (Celera, Berkeley)
- 2001 2x ľudský genóm 3 GB (NIH, Celera)
- po 2001 Myš, potkan, kura, šimpanz, pes, makak,...
- 2007 Watsonov a Venterov genóm (454)
- po 2012 1000 ľudských genómov
- čoskoro 10k genómov stavovcov, sekvenovanie ako diagnostický nástroj

## Zhrnutie

- Sekvenovanie genómu je zložitý proces, v ktorom hrá bioinformatika dôležitú úlohu
- V súčasnosti nízka cena, krátke čítania nastupuje 3. generácia s extra dlhými ale chybovými čítaniami
- Problém zostavovania genómu, najkratšie spoločné nadslovo
- Praktické riešenie pre krátke čítania: de Bruijnove grafy
- V zostavenej sekvencii môžu byť chyby, medzery, viaceré kontigy
- Na budúce: ako sa vysporiadať s dlhými čítaniami 3. generácie?
- Pokrytie genómu a veľkosť čítania hrajú najdôležitejšiu úlohu pri tom, ako fragmentovaný bude výsledok:
  - pre Sanger: 7-10× pokrytie
  - pre NGS: 40-70× pokrytie
  - pre 3. generáciu: 30× pokrytie

## Použitie NGS: Populačná genetika

- Sekvenujeme krátke čítania z genómu určitého človeka
- Ako sa môj vlastný genóm líši od genómu “priemerného” človeka?
- Ako jednoduché genetické rozdiely ovplyvňujú fenotyp?
- Personalizovaná medicína
- Populačná štruktúra, história ľudstva
- Etické otázky

## Problémy:

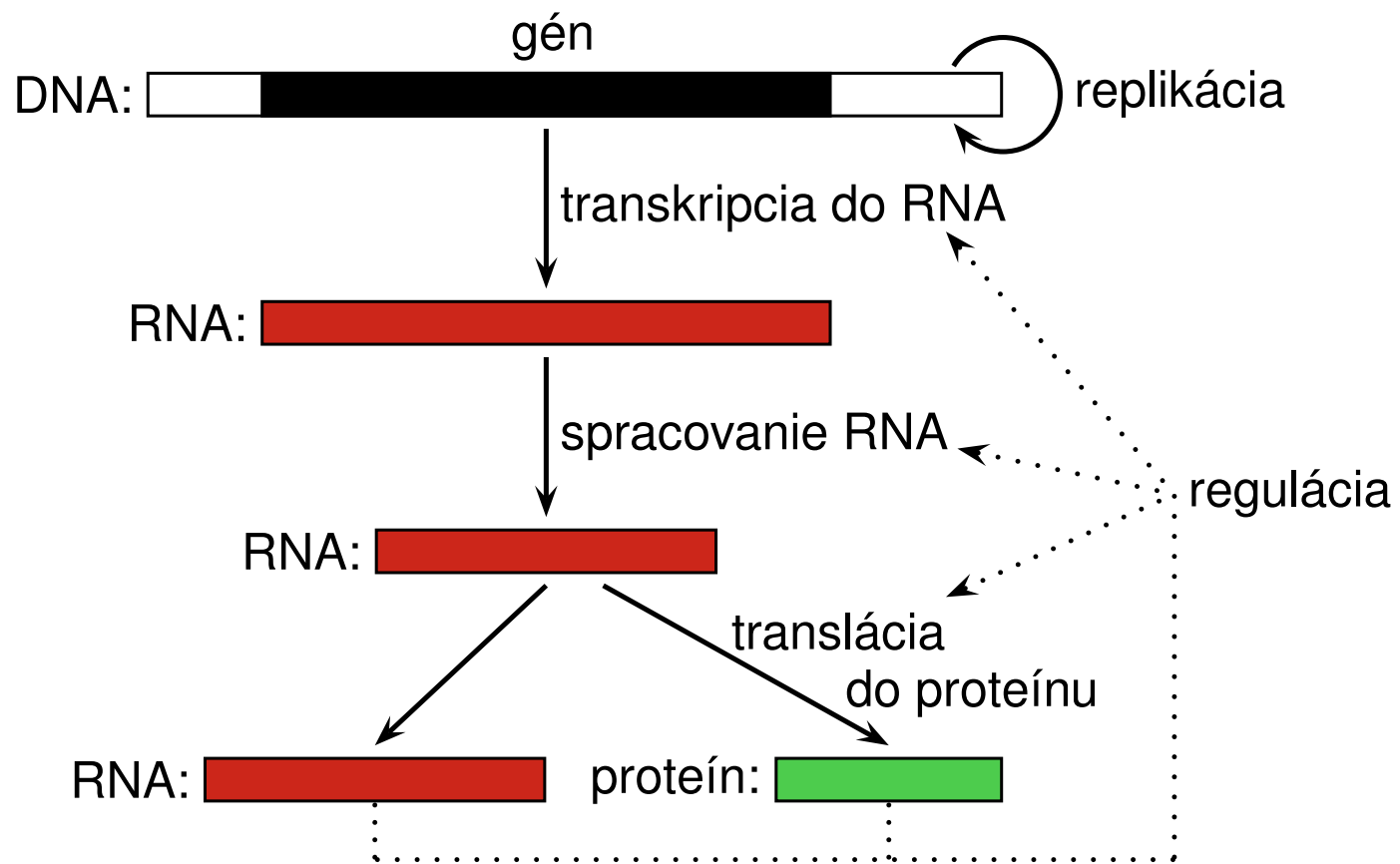
- Mapovanie krátkych čítaní na referenčnú sekvenciu
- Identifikácia rozdielov (malých a väčších)

## Použitie NGS: Environmentálne sekvenovanie – Metagenomika

- Aké mikroorganizmy žijú v našich telách?  
črevná a žalúdočná flóra, ústna dutina, koža, ...
- Diverzita mikroorganizmov v rôznych ekosystémoch
- Ťažké izolovať jednotlivé organizmy
- Sekvenujeme zmes čítania z rôznych genómov
- Snažíme sa zostaviť aspoň krátke kontigy

### Problémy:

- Oddelenie čítaní/kontigov patriacich do rôznych genómov



## Použitie NGS: Hľadanie génov, väzobných miest,...

- Sekvenovať môžeme aj RNA, dostávame gény v genóme
- Chip-Seq: vyfiltrujeme kúsky DNA, na ktoré je naviazaný určitý proteín, sekvenujeme, mapujeme na genóm

### Problémy:

- Identifikácia miest zotrihu
- Identifikácia väzobných miest podľa hĺbky pokrytia