Sekvenovanie génov očami informatika

Peter Glaus

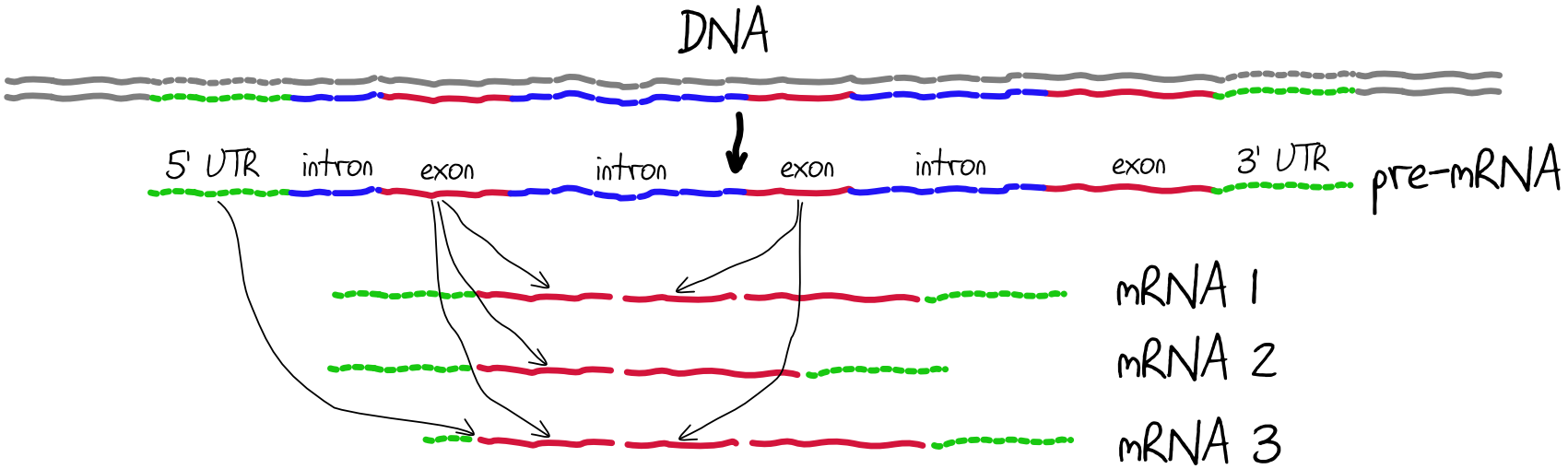
The University of Manchester

DNA:

- sekvencia nukleotidov: Adenín, Cytozín, Guanín, Tymín
- obsahuje zápis génov
  - "kódujúce"gény - zapisujú proteíny
  - "nekódujúce"gény - zapisujú *pomocné* RNA

DNA:

- sekvencia nukleotidov: Adenín, Cytozín, Guanín, Tymín
- obsahuje zápis génov
  - "kódujúce" gény - zapisujú proteíny
  - "nekódujúce" gény - zapisujú *pomocné* RNA



Ľudský genóm:

Ľudský genóm:

- báz: 3,300,551,249

Ľudský genóm:

- báz: 3,300,551,249
- génov: 21,224

Ľudský genóm:

- báz: 3,300,551,249
- génov: 21,224
- prepisov(mRNA): 194,015

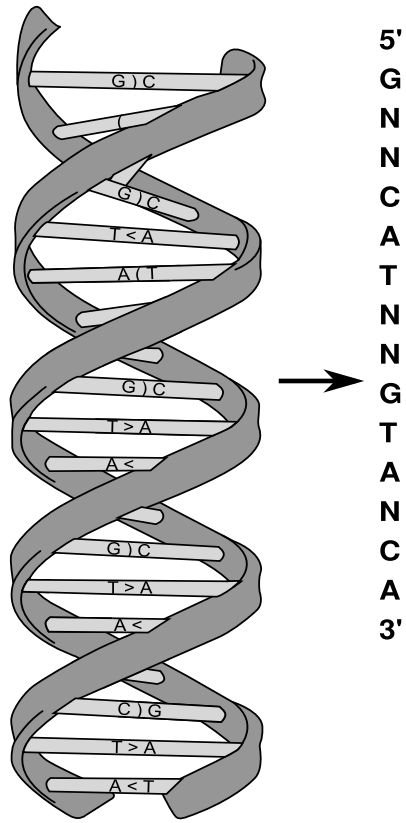
Ľudský genóm:

- báz: 3,300,551,249
- génov: 21,224
- prepisov(mRNA): 194,015
- variabilných báz (SNPs): 187,852,828 (5.7%)

Ľudský genóm:

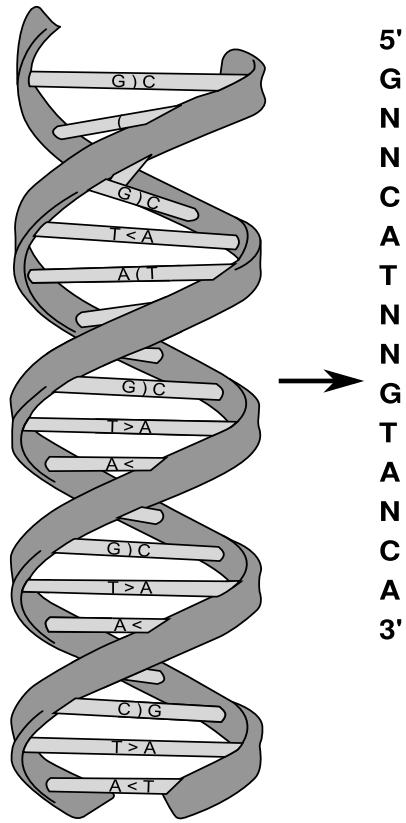
- báz: 3,300,551,249
- génov: 21,224
- prepisov(mRNA): 194,015
- variabilných báz (SNPs): 187,852,828 (5.7%)
- dôležitá je aj sekundárna štruktúra

Sekvenovanie DNA:



Sekvenovanie DNA:

- prvá generácia: *Sangerovo* sekvenovanie
- princíp "chain-termination"s použitím fluorescenčných značiek
- použité pri "Human genome project"
  - trval takmer 15 rokov
  - v roku 2003 dokončili 99%



"Sekvenovanie druhej generácie":

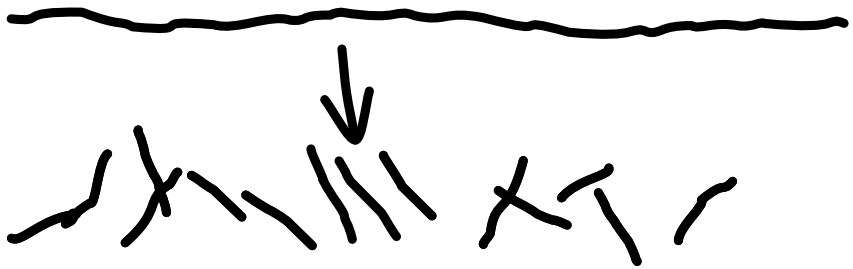
- *Next Generation Sequencing*, *High-throughput sequencing*
- viacero podobných technológí
- využívajú princíp Sangerovho sekvenovania alebo sekvenovania syntézou
- produkujú relatívne krátke reťazce: sub-sekvencie

"Sekvenovanie druhej generácie" príklad:

1. fragmentácia
2. prilepenie
3. namnoženie
4. pozorovanie (fluorescencia)
   - pridanie značiek
   - odfotenie
   - odstránenie značiek

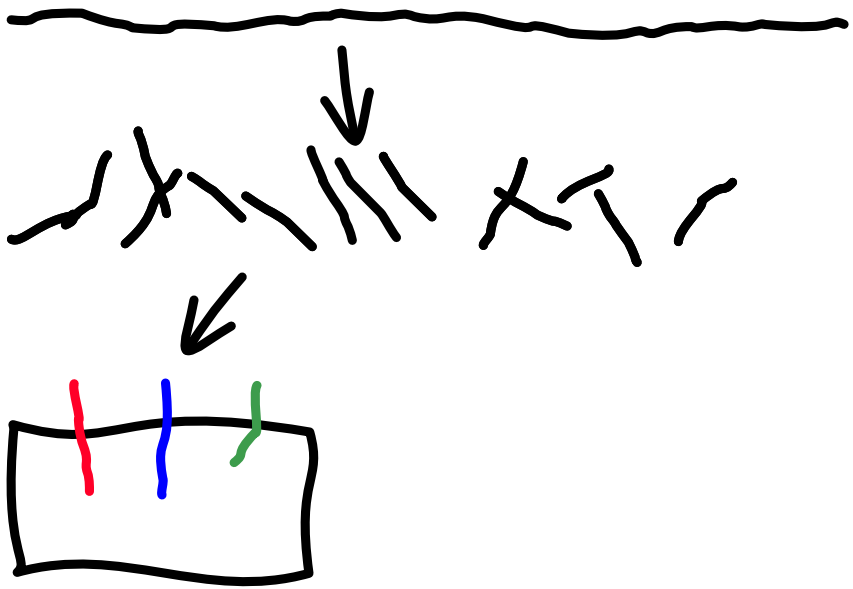
"Sekvenovanie druhej generácie" príklad:

1. fragmentácia
2. prilepenie
3. namnoženie
4. pozorovanie (fluorescencia)
   - pridanie značiek
   - odfotenie
   - odstránenie značiek



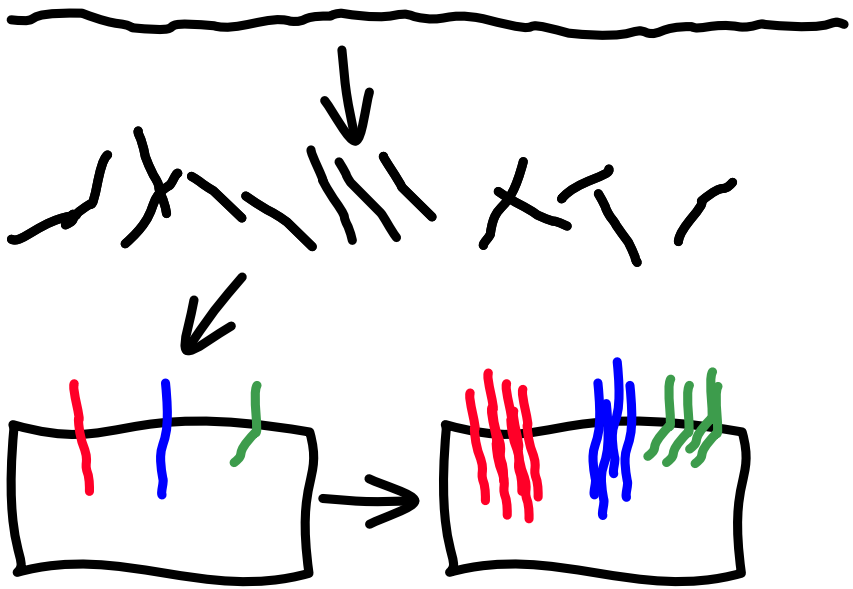
"Sekvenovanie druhej generácie" príklad:

1. fragmentácia
2. prilepenie
3. namnoženie
4. pozorovanie (fluorescencia)
   - pridanie značiek
   - odfotenie
   - odstránenie značiek



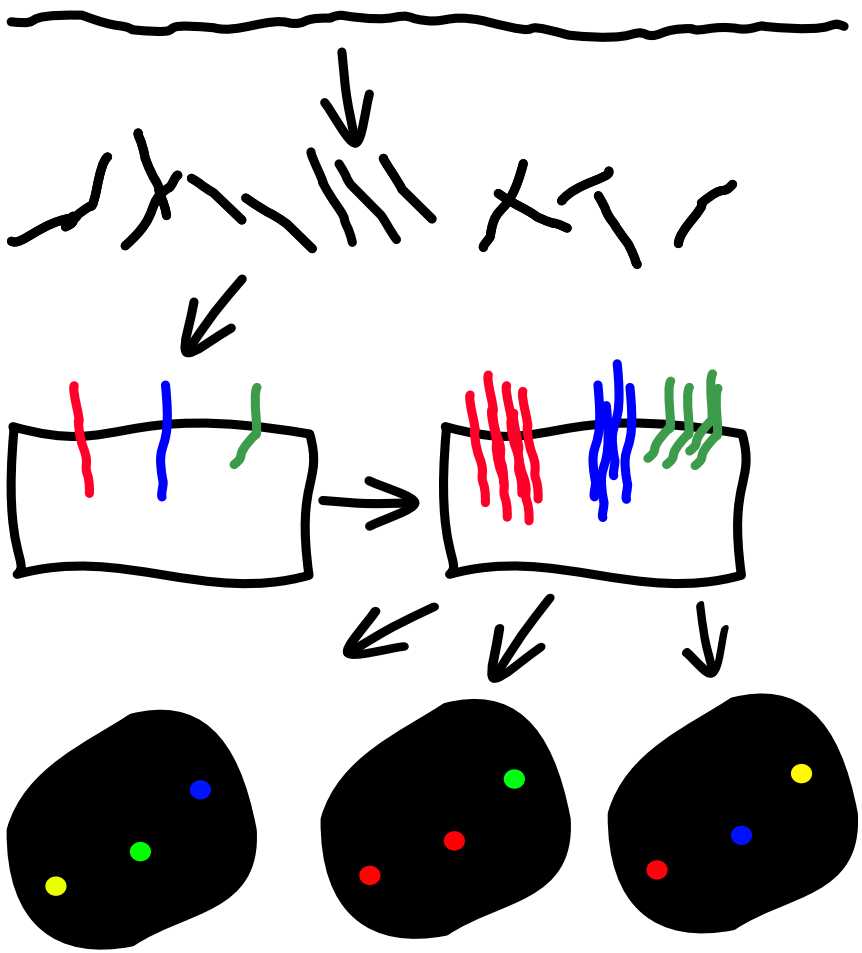
"Sekvenovanie druhej generácie" príklad:

1. fragmentácia
2. prilepenie
3. namnoženie
4. pozorovanie (fluorescencia)
   - pridanie značiek
   - odfotenie
   - odstránenie značiek



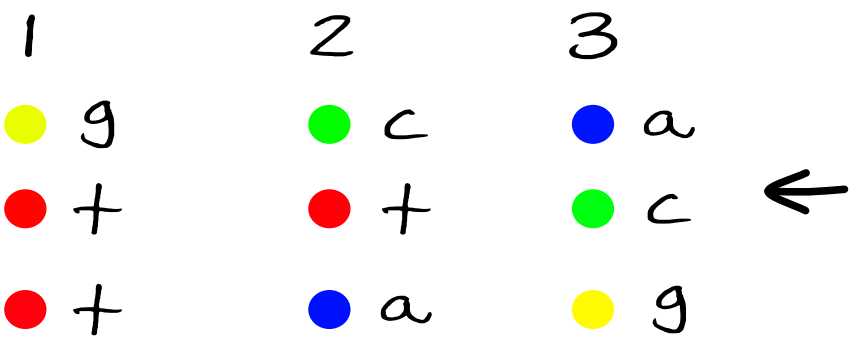
"Sekvenovanie druhej generácie" príklad:

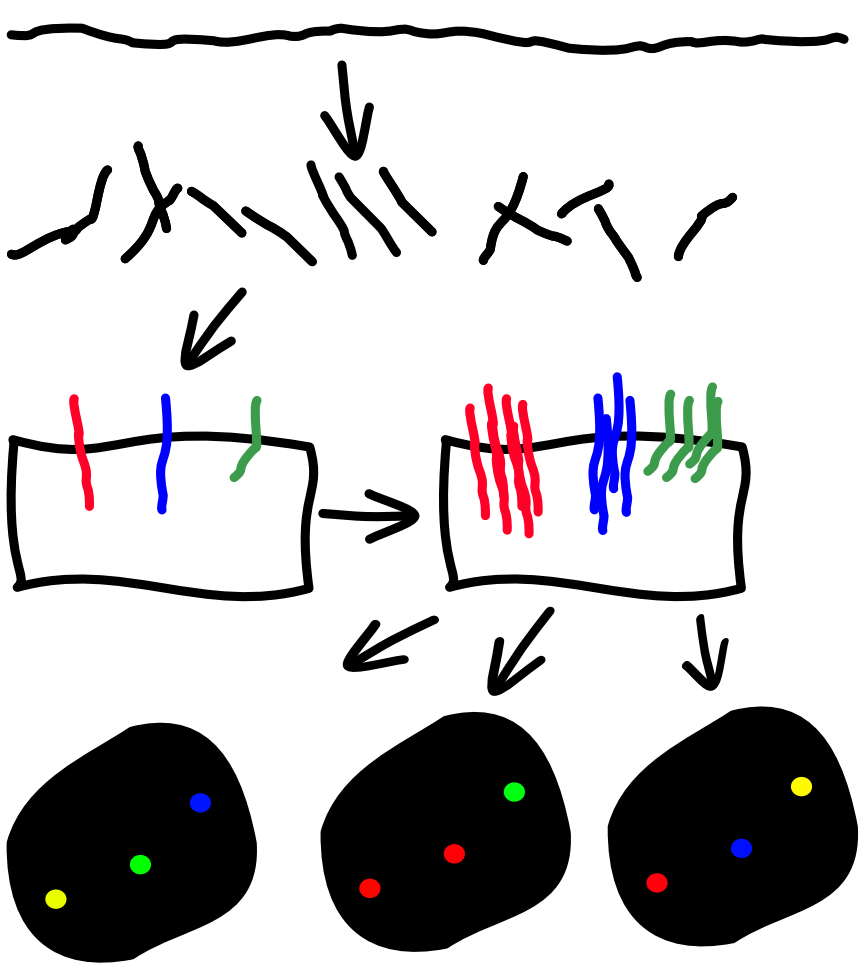
1. fragmentácia
2. prilepenie
3. namnoženie
4. pozorovanie (fluorescencia)
   - pridanie značiek
   - odfotenie
   - odstránenie značiek



"Sekvenovanie druhej generácie" príklad:

1. fragmentácia
2. prilepenie
3. namnoženie
4. pozorovanie (fluorescencia)
   - pridanie značiek
   - odfotenie
   - odstránenie značiek



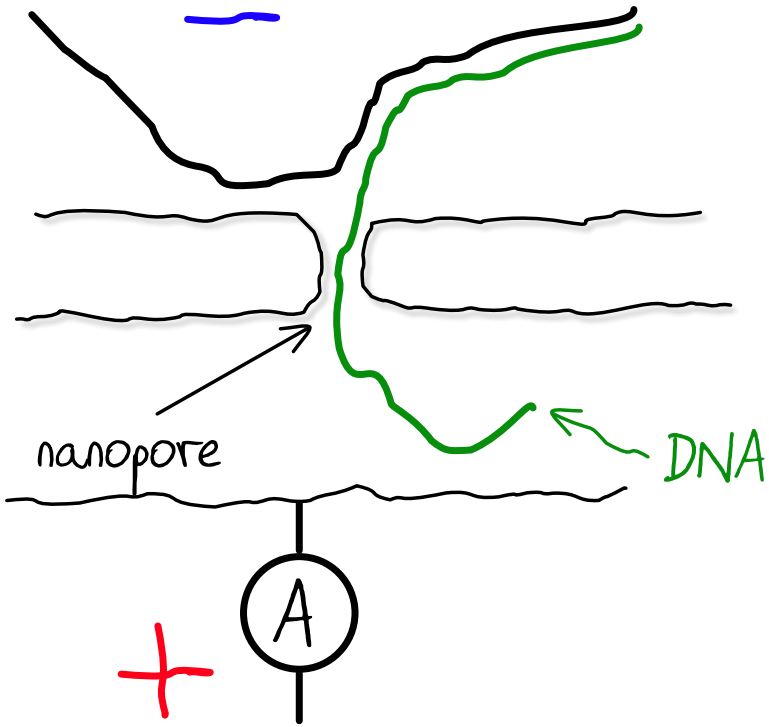


Sekvenovanie tretej generácie

- skevenovanie individuálnych molekúl
- sekvenovanie pomocou nanopórov

Sekvenovanie pomocou nanopórov

- meranie elektrického prúdu prechádzajúceho nanopórom
- molekuly chrakteristicky menia meraný prúd



- Oxford Nanopore `http://vimeo.com/52696609`

Sekvenovanie druhej generácie:

- výhody
  - rýchlosť (pár dní až týždňov)
  - cena
  - pokrytie
- problémy
  - krátke reťazce 30 – 400 báz
  - nutnosť množenia molekúl
  - chyby:
    - zámena báz
    - vynechanie/pridanie báz
  - veľké množstvo dát
    - 100K - 100M reťazcov

Príklad

```
hejge enera druhe anied racie iedru novan uhejg
kveno ovanie enera sekve
```

```
sekve ovanie  uhejg   racie
  kveno   iedru    enera
        anied  hejge
     novan  druhe
```

```
SekvenovanieDruhejGeneracie
```

Príklad

```
hejge enera druhe anied racie iedru novan uhejg
kveno ovanie enera sekve
```

```
sekve ovanie  uhejg   racie
  kveno   iedru    enera
        anied  hejge
     novan  druhe
```

SekvenovanieDruhejGeneracie

Príklad

```
hejge enera druhe anied racie iedru novan uhejg
kveno ovanie enera sekve
```

```
sekve ovanie  uhejg   racie
  kveno   iedru    enera
        anied  hejge
     novan  druhe
```

```
SekvenovanieDruhejGeneracie
```

Príklad 2

sekvenovaniedruhejgeneracie

hejae enera drihe anied racie iedru nofan uhejg

sekvenovaniedruhejgeneracie
uhejg racie
iedri enera
anied hejae
nofan drihe

sekvenovaniedruhejgeneracie
sekvenoFaniedrIhejAeneracie

Príklad 2

```
sekvenovaniedruhejgeneracie

hejae enera drihe anied racie iedru nofan uhejg
```

```
sekvenovaniedruhejgeneracie
            uhejg   racie
         iedri    enera
       anied  hejae
     nofan  drihe
```

```
sekvenovaniedruhejgeneracie
sekvenoFaniedrIhejAeneracie
```

Príklad 2

```
sekvenovaniedruhejgeneracie

hejae enera drihe anied racie iedru nofan uhejg


sekvenovaniedruhejgeneracie
             uhejg   racie
          iedri    enera
        anied  hejae
     nofan  drihe


sekvenovaniedruhejgeneracie
sekvenoFaniedrIhejAeneracie
```

Príklad 2

```
sekvenovaniedruhejgeneracie

hejae enera drihe anied racie iedru nofan uhejg


sekvenovaniedruhejgeneracie
             uhejg   racie
          iedri    enera
        anied  hejae
     nofan  drihe


sekvenovaniedruhejgeneracie
sekvenoFaniedrIhejAeneracie
```

Využitie sekvenovania:

- *de-novo* skladanie genómu
- hľadanie rozdielov (SNPs)
- hľadanie miest viazania proteínov (ChIP-seq)
- meranie expresie génov (RNA-seq)

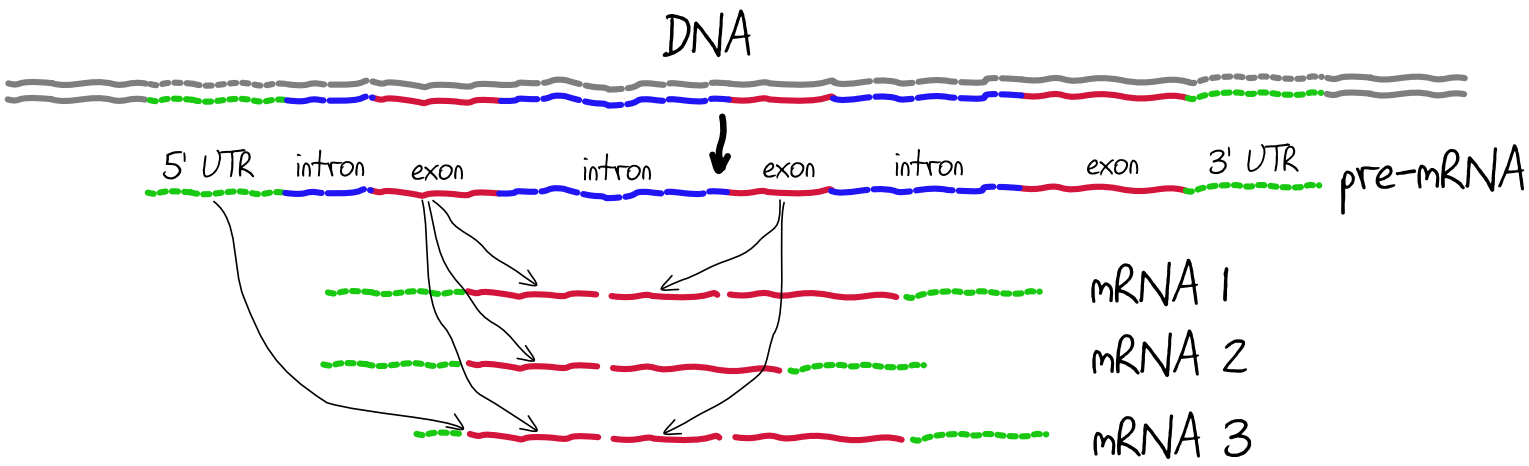
RNA-seq:

- Sekvenovanie cDNA

- počet reťazcov $\approx$ množstvo fragmentov
- počet reťazcov $\approx$ (expresia) x (dĺžka)

RNA-seq:

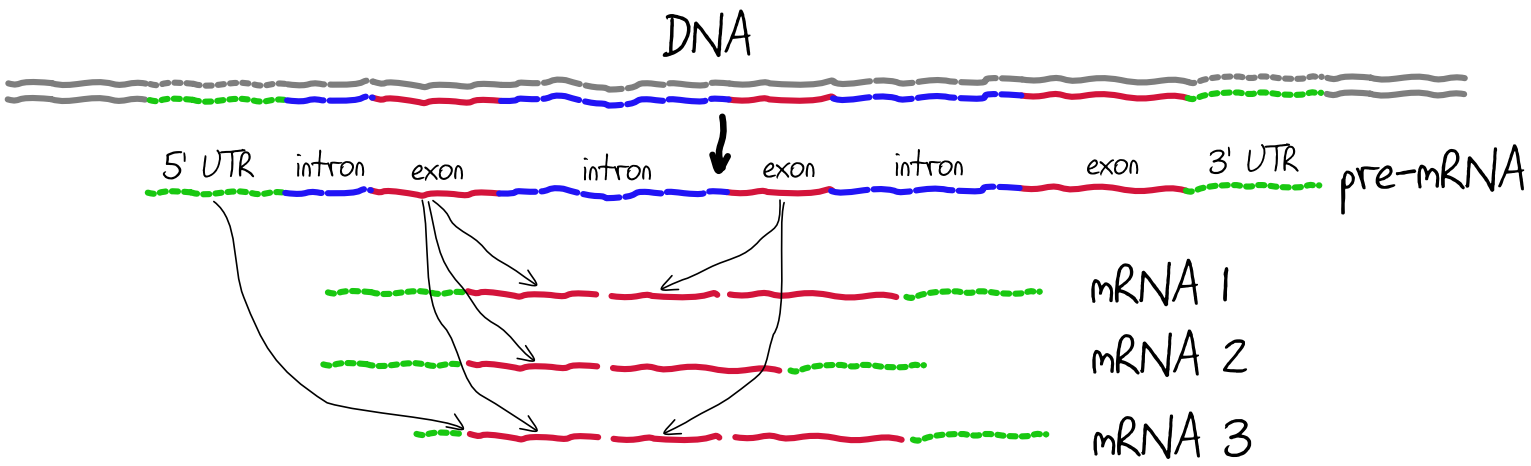
- Sekvenovanie cDNA



- počet reťazcov $\approx$ množstvo fragmentov
- počet reťazcov $\approx$ (expresia) x (dĺžka)

RNA-seq:

- Sekvenovanie cDNA



- počet reťazcov $\approx$ množstvo fragmentov
- počet reťazcov $\approx$ (expresia) x (dĺžka)

Pravdepodobnostné modelovanie + Bayesova rovnica

- Reprezentácia neznámych pomocou pravdepodobnostného rozdelenia
  (namiesto hodnoty a odhadu chyby)

$$\phi \sim N(\mu|\sigma) \qquad \phi \sim \mathbf{S} = \{s_1, \ldots, s_n\}$$

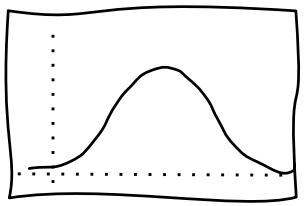
- Používame Bayesovu vetu na manipulovanie pravdepodobností
- Výsledok možno zhrnúť pomocou priemeru a strednej odchýlky

$$\mathrm{E}[\phi] = \mathrm{mean}(S); \sigma_\phi = \mathrm{stdev}(S)$$

Pravdepodobnostné modelovanie + Bayesova rovnica

- Reprezentácia neznámych pomocou pravdepodobnostného rozdelenia
  (namiesto hodnoty a odhadu chyby)

$$\phi \sim N(\mu|\sigma) \qquad \phi \sim \mathbf{S} = \{s_1, \ldots, s_n\}$$



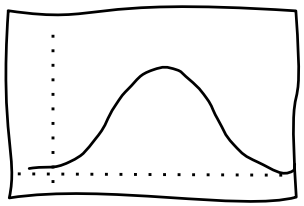
- Používame Bayesovu vetu na manipulovanie pravdepodobností
- Výsledok možno zhrnúť pomocou priemeru a strednej odchýlky

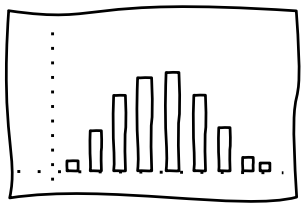
$$\mathrm{E}[\phi] = \mathrm{mean}(S); \sigma_\phi = \mathrm{stdev}(S)$$

Pravdepodobnostné modelovanie + Bayesova rovnica

- Reprezentácia neznámych pomocou pravdepodobnostného rozdelenia
  (namiesto hodnoty a odhadu chyby)

$$\phi \sim N(\mu|\sigma) \qquad \phi \sim \mathbf{S} = \{s_1, \ldots, s_n\}$$





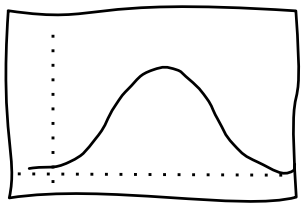
- Používame Bayesovu vetu na manipulovanie pravdepodobností
- Výsledok možno zhrnúť pomocou priemeru a strednej odchýlky

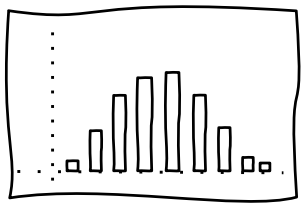
$$\mathrm{E}[\phi] = \mathrm{mean}(S); \sigma_\phi = \mathrm{stdev}(S)$$

Pravdepodobnostné modelovanie + Bayesova rovnica

- Reprezentácia neznámych pomocou pravdepodobnostného rozdelenia
  (namiesto hodnoty a odhadu chyby)

$$\phi \sim N(\mu|\sigma) \qquad \phi \sim \mathbf{S} = \{s_1, \ldots, s_n\}$$





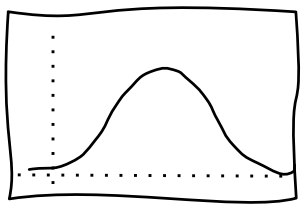
- Používame Bayesovu vetu na manipulovanie pravdepodobností
- Výsledok možno zhrnúť pomocou priemeru a strednej odchýlky

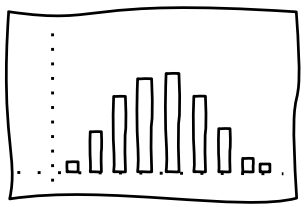
$$\mathrm{E}[\phi] = \mathrm{mean}(S); \sigma_\phi = \mathrm{stdev}(S)$$

Pravdepodobnostné modelovanie + Bayesova rovnica

- Reprezentácia neznámych pomocou pravdepodobnostného rozdelenia
  (namiesto hodnoty a odhadu chyby)

$$\phi \sim N(\mu|\sigma) \qquad \phi \sim \mathbf{S} = \{s_1, \ldots, s_n\}$$





- Používame Bayesovu vetu na manipulovanie pravdepodobností
- Výsledok možno zhrnúť pomocou priemeru a strednej odchýlky

$$\mathrm{E}[\phi] = \mathrm{mean}(S); \sigma_\phi = \mathrm{stdev}(S)$$

RNA-seq alternatívny pohľad:

- Neznáma expresia prepisov θ

$$P(\text{reťazec}|\theta) = P(\text{mRNA}|\theta)P(\text{fragment}|\text{mRNA})P(\text{reťazec}|\text{fragment})$$

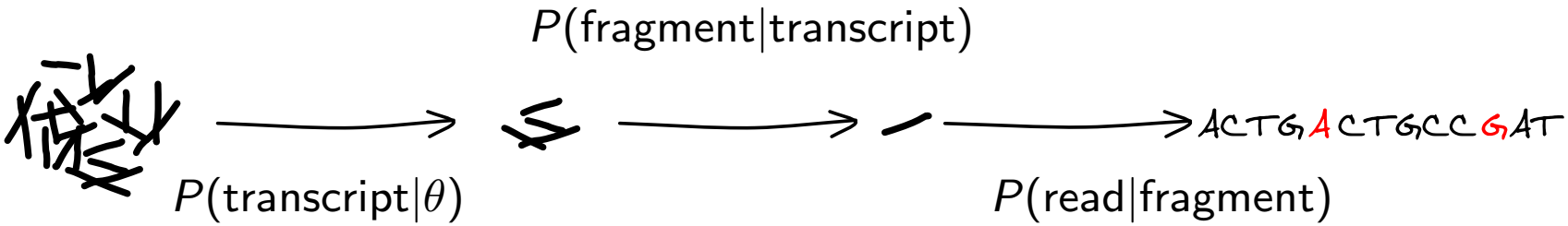
$$P(\text{Dáta}|\theta) = \prod_{i=1}^{n} P(\text{read}_i|\theta)$$

$$P(\theta|\text{Dáta}) = \frac{P(\text{Dáta}|\theta)P(\theta)}{P(\text{Dáta})}$$

- Používame algoritmus "Markov Chain Monte Carlo" (MCMC) na vypočítanie $P(\theta|\text{Dáta})$

RNA-seq alternatívny pohľad:

- Neznáma expresia prepisov θ



$$P(\text{reťazec}|\theta) = P(\text{mRNA}|\theta)P(\text{fragment}|\text{mRNA})P(\text{reťazec}|\text{fragment})$$

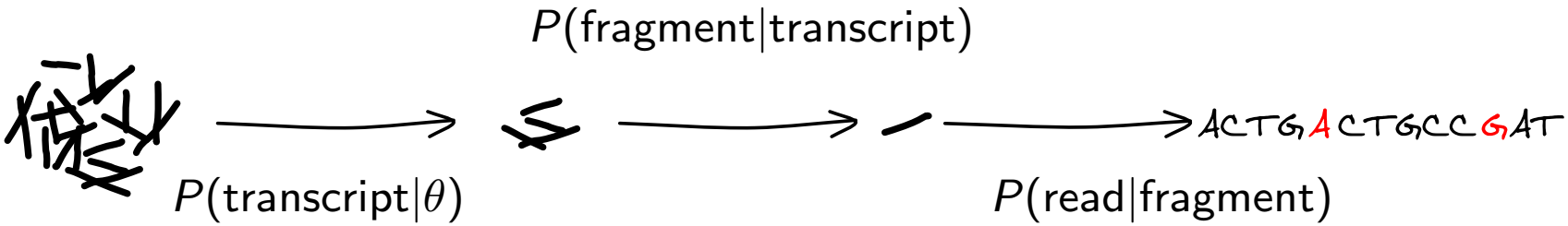
$$P(\text{Dáta}|\theta) = \prod_{i=1}^{n} P(\text{read}_i|\theta)$$

$$P(\theta|\text{Dáta}) = \frac{P(\text{Dáta}|\theta)P(\theta)}{P(\text{Dáta})}$$

- Používame algoritmus "Markov Chain Monte Carlo" (MCMC) na vypočítanie $P(\theta|\text{Dáta})$

RNA-seq alternatívny pohľad:

- Neznáma expresia prepisov θ



$$P(\text{reťazec}|\theta) = P(\text{mRNA}|\theta)P(\text{fragment}|\text{mRNA})P(\text{reťazec}|\text{fragment})$$

$$P(\text{Dáta}|\theta) = \prod_{i=1}^{n} P(\text{read}_i|\theta)$$

$$P(\theta|\text{Dáta}) = \frac{P(\text{Dáta}|\theta)P(\theta)}{P(\text{Dáta})}$$

- Používame algoritmus "Markov Chain Monte Carlo" (MCMC) na vypočítanie $P(\theta|\text{Dáta})$

RNA-seq alternatívny pohľad:

- Neznáma expresia prepisov θ

$$P(\text{reťazec}|\theta) = P(\text{mRNA}|\theta)P(\text{fragment}|\text{mRNA})P(\text{reťazec}|\text{fragment})$$

$$P(\text{Dáta}|\theta) = \prod_{i=1}^{n} P(\text{read}_i|\theta)$$

$$P(\theta|\text{Dáta}) = \frac{P(\text{Dáta}|\theta)P(\theta)}{P(\text{Dáta})}$$

- Používame algoritmus "Markov Chain Monte Carlo" (MCMC) na vypočítanie $P(\theta|\text{Dáta})$

RNA-seq alternatívny pohľad:

- Neznáma expresia prepisov θ

$$P(\text{reťazec}|\theta) = P(\text{mRNA}|\theta)P(\text{fragment}|\text{mRNA})P(\text{reťazec}|\text{fragment})$$

$$P(\text{Dáta}|\theta) = \prod_{i=1}^{n} P(\text{read}_i|\theta)$$

$$P(\theta|\text{Dáta}) = \frac{P(\text{Dáta}|\theta)P(\theta)}{P(\text{Dáta})}$$

- Používame algoritmus "Markov Chain Monte Carlo" (MCMC) na vypočítanie $P(\theta|\text{Dáta})$

RNA-seq alternatívny pohľad:

- Neznáma expresia prepisov θ

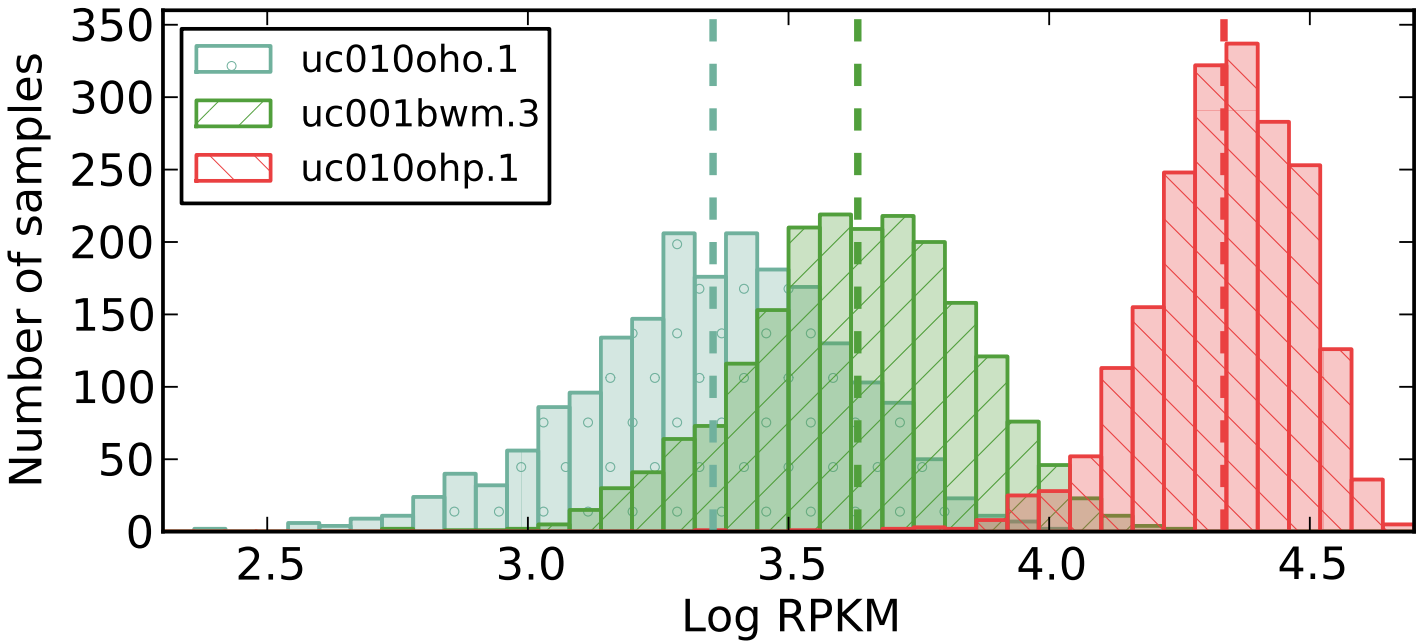
$$P(\text{reťazec}|\theta) = P(\text{mRNA}|\theta)P(\text{fragment}|\text{mRNA})P(\text{reťazec}|\text{fragment})$$

$$P(\text{Dáta}|\theta) = \prod_{i=1}^{n} P(\text{read}_i|\theta)$$

$$P(\theta|\text{Dáta}) = \frac{P(\text{Dáta}|\theta)P(\theta)}{P(\text{Dáta})}$$

- Používame algoritmus "Markov Chain Monte Carlo" (MCMC) na vypočítanie $P(\theta|\text{Dáta})$

Príklad výsledkov:



Histogramy vzoriek z pravdepodobnostného rozdelenia $P(\theta|\text{Dáta})$.

Môj výsledok: BitSeq

Aplikácia má dva ciele

- Určiť množstvo prepisov (mRNA) pomocou dát z RNA-seq
- Vybrať tie gény a prepisy (mRNA) ktoré sú rôzne zastúpené v rôznych podmienkach

Môj výsledok: BitSeq

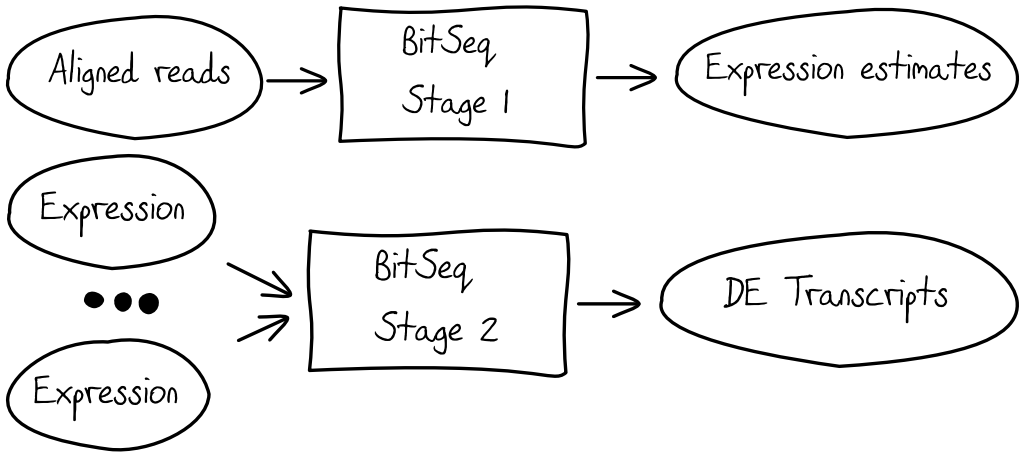
Aplikácia má dva ciele

- Určiť množstvo prepisov (mRNA) pomocou dát z RNA-seq
- Vybrať tie gény a prepisy (mRNA) ktoré sú rôzne zastúpené v rôznych podmienkach

Môj výsledok: BitSeq

Aplikácia má dva ciele

- Určiť množstvo prepisov (mRNA) pomocou dát z RNA-seq
- Vybrať tie gény a prepisy (mRNA) ktoré sú rôzne zastúpené v rôznych podmienkach



Záver:

- technológie sekvenovania prudko napredujú
- produkované dáta vyžadujú efektívne spracovanie
- sekvenovaním vieme:
  - sekvenovať nové organizmy
  - hľadať odlišnosti medzi DNA
  - detekovať množstvo génov v rôzných vzorkách
  - a oveľa viac

Záver:

- technológie sekvenovania prudko napredujú
- produkované dáta vyžadujú efektívne spracovanie
- sekvenovaním vieme:
  - sekvenovať nové organizmy
  - hľadať odlišnosti medzi DNA
  - detekovať množstvo génov v rôzných vzorkách
  - a oveľa viac

Záver:

- technológie sekvenovania prudko napredujú
- produkované dáta vyžadujú efektívne spracovanie
- sekvenovaním vieme:
  - sekvenovať nové organizmy
  - hľadať odlišnosti medzi DNA
  - detekovať množstvo génov v rôznych vzorkách
  - a oveľa viac

Questions!