# Využitie reťazcových grafov pri hľadaní chýb v genómoch

Angelika Fedáková
Školiteľ: doc. Mgr. Bronislava Brejová, PhD.

01.07.2020

# DNA

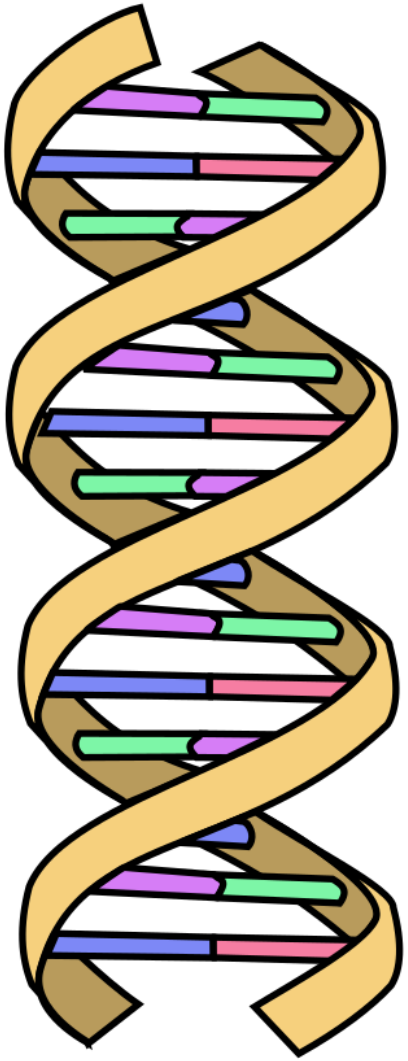
- Reťazec báz A,C,G,T
- Dve komplementárne vlákna

# Chromozóm

- Súvislý úsek DNA

# Genóm

- Súbor chromozómov



DNA

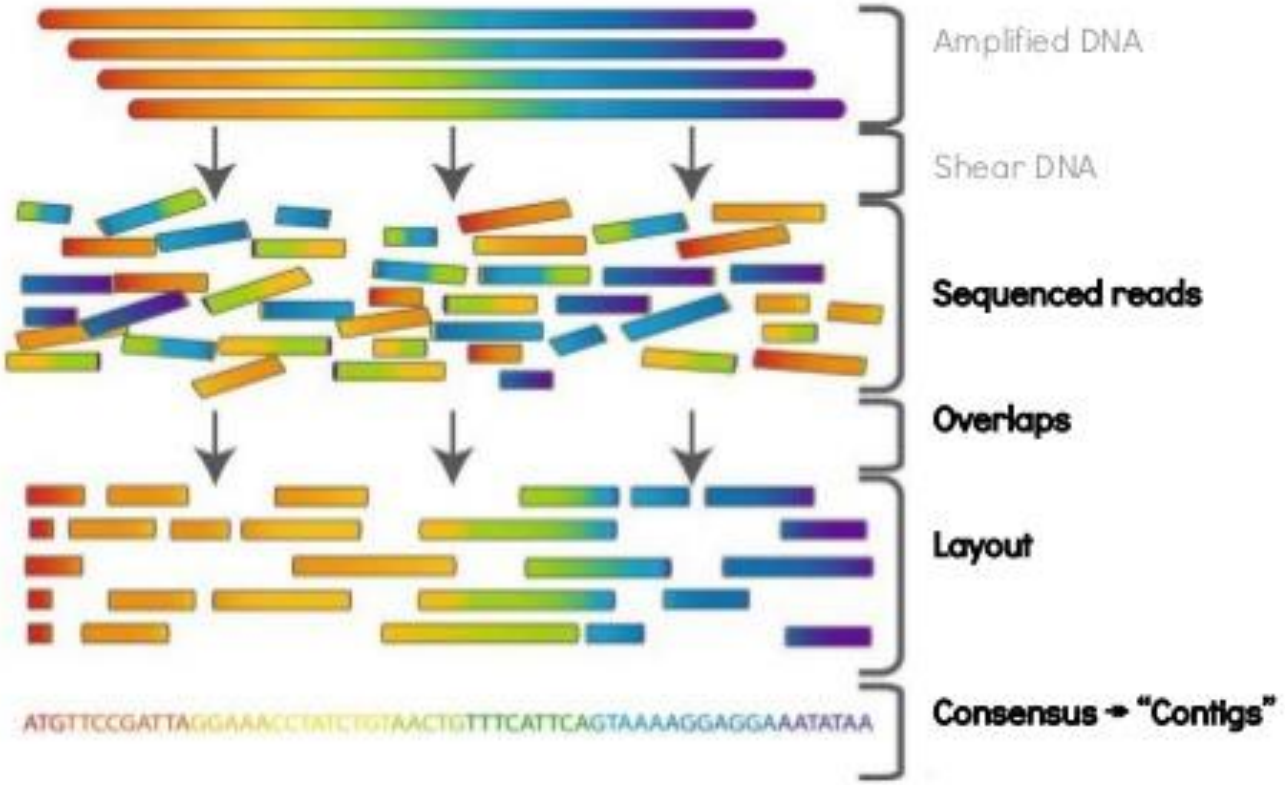
# Sekvenovanie DNA

- Zisťovania báz chromozómov
- Sekvenovanie krátkych kúskov – čítaní
  - Krátke čítania, dlhé čítania

# Zostavovanie genómu

- Proces skladania čítaní do pôvodných chromozómov

# Overlap - Layout - Consensus



# Zarovnávanie sekvencií

- Hľadanie podobností a rozdielov sekvencií
- Mapovanie báz sekvencií
- Mapovanie sekvencie ku grafu

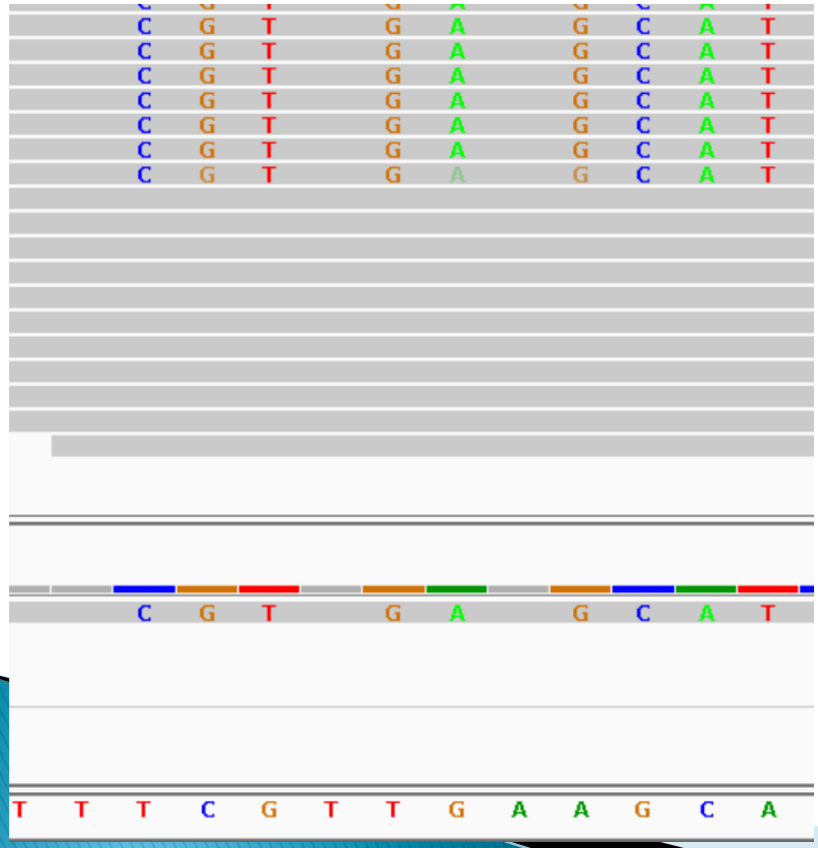
Sekvencia 1: AGCTGGCTT
Sekvencia 2: GCAGGTCTT

A G C T G G - C T T
- G C A G G T C T T
-1 +1 +1 -1 +1 +1 -1 +1 +1 +1

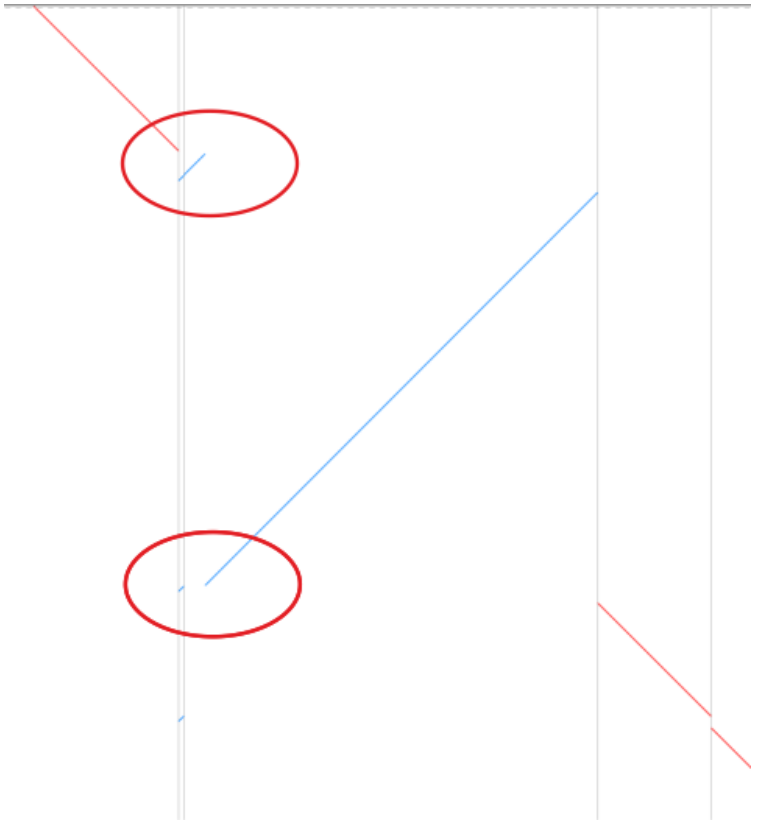
Skóre zarovnania 4

# Chyby DNA genómov

- Menšie chyby
  - Zlé určenie jednotlivých báz




- Rozsiahlejšie chyby
  - Zlé určenie poradia väčších kusov chromozómu



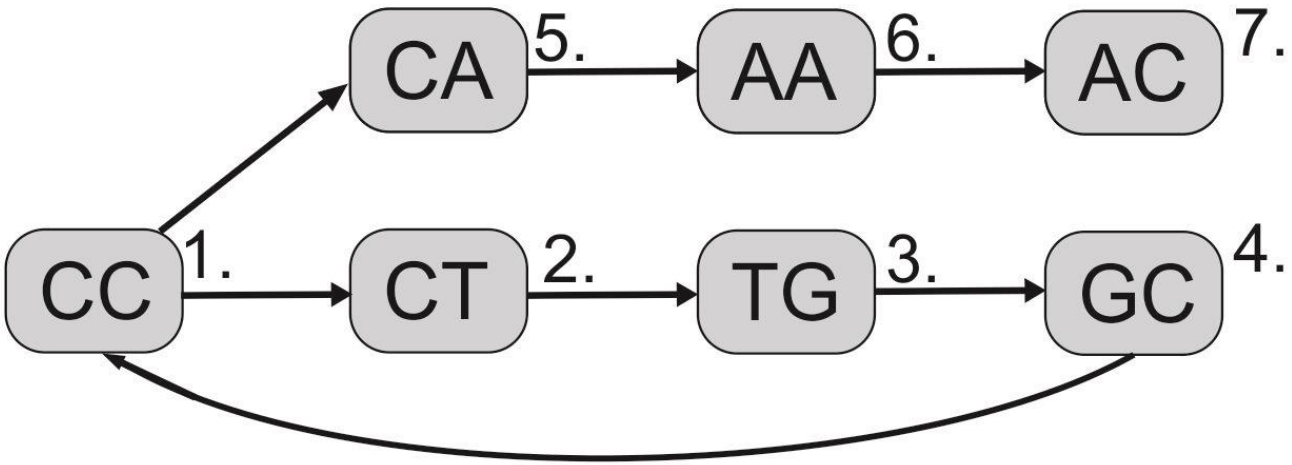
# Ciele práce

- Priniesť nový pohľad na problém hľadania chýb
- Identifikovať miesta s pravdepodobnosťou chyby
- Porovnať výsledky s existujúcim softvérom

# Postup hľadania chýb

- Vytvorenie grafu

k = 2, krátke čítania: CCTG, TGCC, GCCA, CAAC




Sekvencia: TGCCAAC

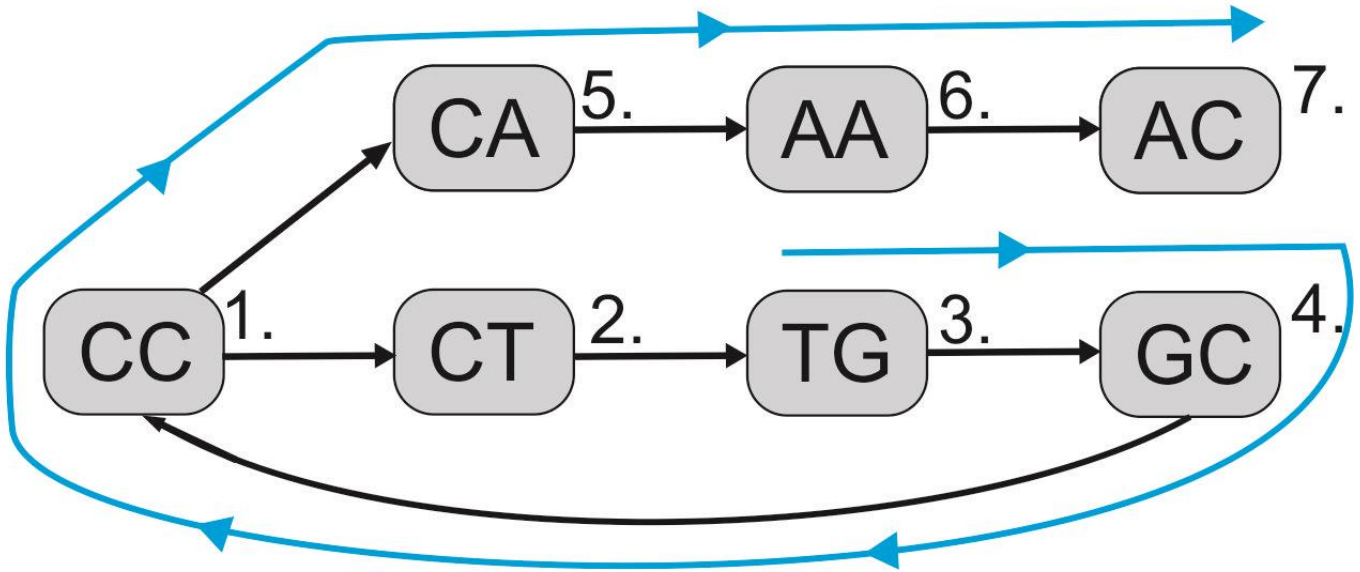
Dlhé čítania: CCTGCC, GCCAAC

# Postup hľadania chýb

- Zarovnanie dlhých čítaní a sekvencie

Sekvencia: TGCCAAC

Dlhé čítania: CCTGCC, GCCAAC




Zarovnanie čítaní: 1. 2. 3. 4. 1. 4. 1. 5. 6. 7.

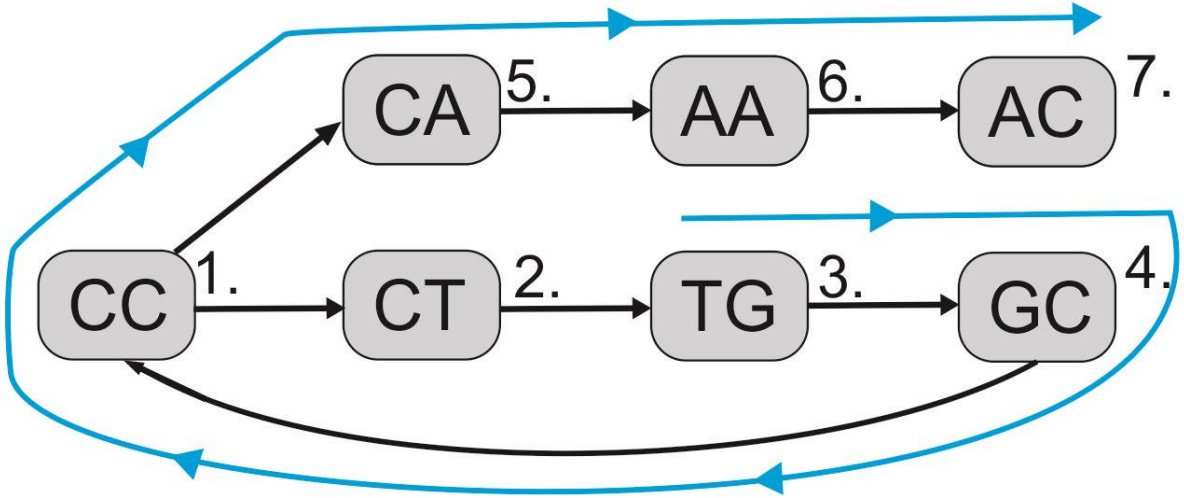
Zarovnanie sekvencie: 3. 4. 1. 5. 6. 7.

# Postup hľadania chýb

- Kontrola sledov

Sekvencia: TGCCAAC

Dlhé čítania: CCTGCC, GCCAAC




Zarovnanie čítaní: 1. 2. 3. 4. 1. 4. 1. 5. 6. 7.

Zarovnanie sekvencie: 3. 4. 1. 5. 6. 7.

Sledy dĺžky 1: — = = — — — LIMIT 2

Chybové miesta: TGCCAAC

# Vstup a výstup programu

- Vstup:
  - Skúmaný genóm/sekvencia (FASTA)
  - Krátke čítania (FASTA, FASTQ)
  - Dlhé čítania (FASTA, FASTQ)
- Výstup
  - Oblasti chybových miest (BED)

# Experimentálne dáta

- Baktéria Escherichia Coli (E.~Coli)
  - Referenčná sekvencia (~500 000bp)
  - 5 zostavených sekvencií
- Kvasinka Saprochaete fungicola
  - Referenčná sekvencia (~20 000 000bp)
  - 2 zostavené sekvencie

# Vyhodnotenie algoritmu

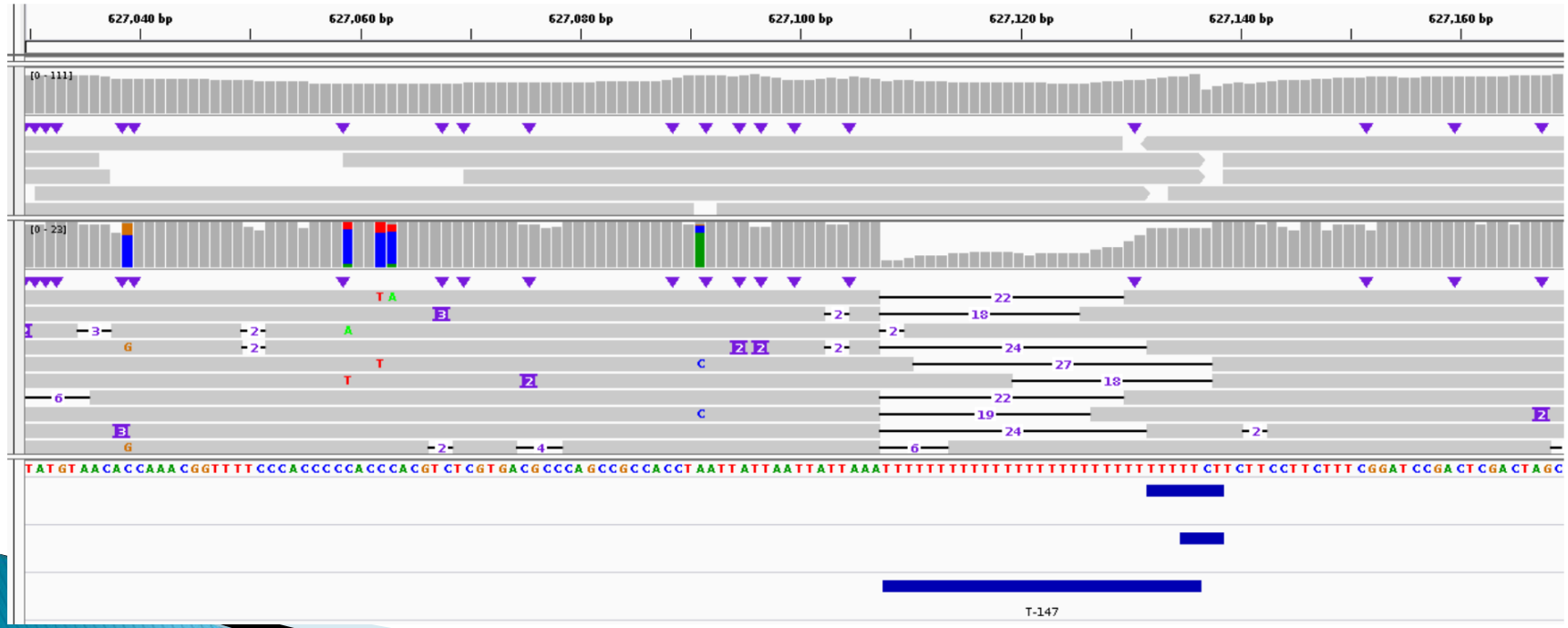
- Referenčné chyby, nájdené chyby, správne chyby
- Presnosť : $\frac{\text{správne nájdené chyby}}{\text{všetky nájdené chyby}}$
- Úplnosť : $\frac{\text{nájdené referenčné chyby}}{\text{všetky referenčné chyby}}$

# Prvé výsledky kvasinky

| | *Spades* | | *Miniasm* | |
|---|---|---|---|---|
| | | ***Presnosť*** | | |
| ***Limit1*** | 543/837 | 64.8% | 325/480 | 67.7% |
| ***Limit3*** | 1209/2591 | 46.6% | 472/816 | 57.8% |
| ***Limit5*** | 1310/3144 | 41.6% | 593/1062 | 55.8% |
| ***Limit7*** | 1379/3902 | 35.3% | 662/1244 | 53.3% |
| | | ***Úplnosť*** | | |
| ***Limit1*** | 599/1599 | 37.4% | 386/2513 | 15.3% |
| ***Limit3*** | 1298/1599 | 81.2% | 567/2513 | 22.5% |
| ***Limit5*** | 1389/1599 | 86.8% | 692/2513 | 27.5% |
| ***Limit7*** | 1459/1599 | 91.2% | 770/2513 | 30.6% |

# Filtrácia nájdených chýb

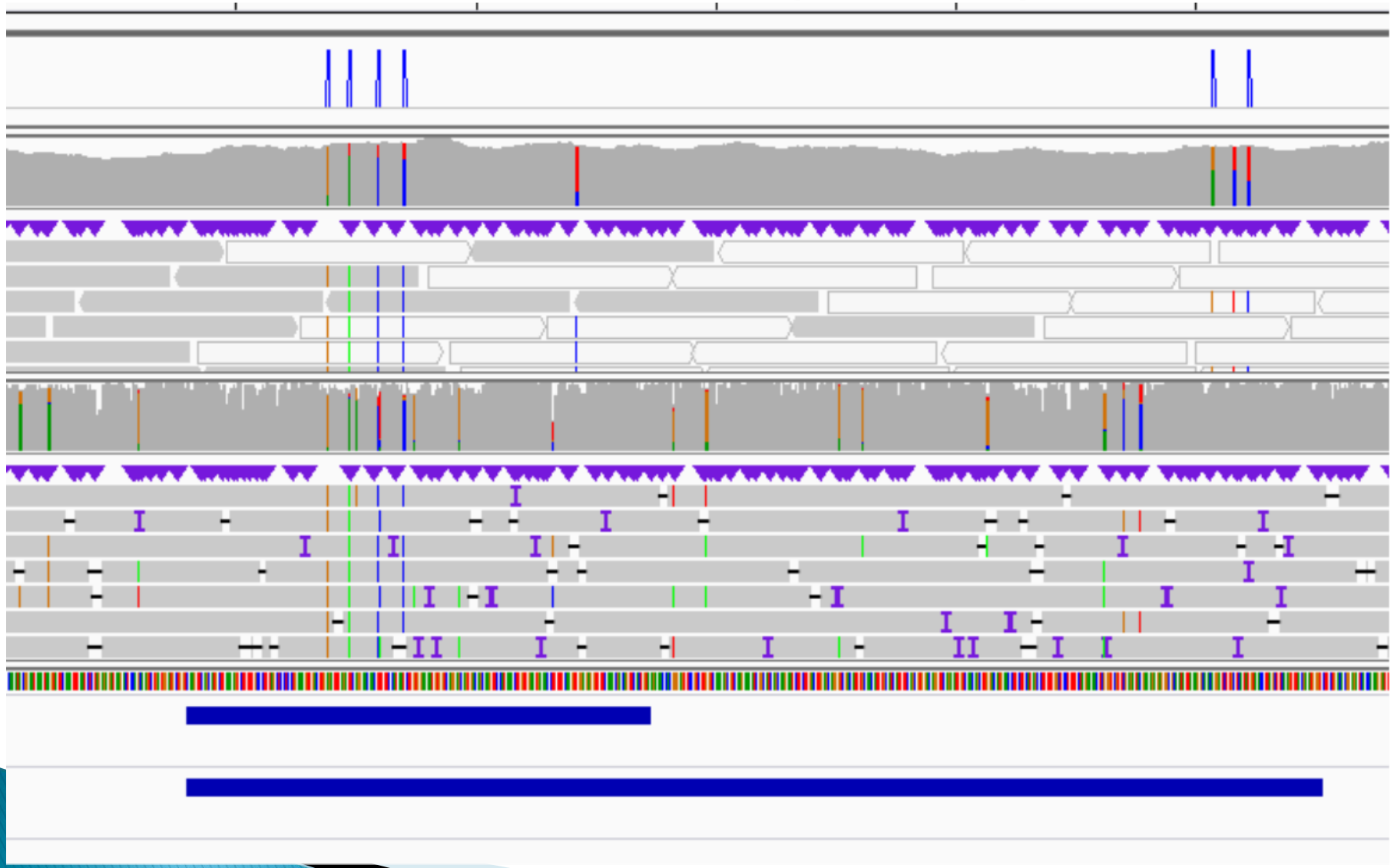
- Odstránenie chýb prekrývajúcich sa s homopolymérmi
- Odstránenie dlhých chýb



# Finálne výsledky po vykonaných filtráciách

| | *Spades* | | *Miniasm* | |
|---|---|---|---|---|
| | | ***Presnosť*** | | |
| ***Limit1*** | 153/198 | 77.3% | 101/136 | 74.3% |
| ***Limit3*** | 439/615 | 71.4% | 124/183 | 67.8% |
| ***Limit5*** | 498/883 | 56.4% | 153/227 | 67.4% |
| ***Limit7*** | 544/1418 | 38.4% | 174/263 | 66.2% |
| | | ***Úplnosť*** | | |
| ***Limit1*** | 185/661 | 28% | 124/1506 | 8.2% |
| ***Limit3*** | 459/661 | 69.4% | 157/1506 | 10.4% |
| ***Limit5*** | 515/661 | 77.9% | 190/1506 | 12.6% |
| ***Limit7*** | 546/661 | 82.5% | 211/1506 | 14% |

# Príklad správne nájdenej chyby



# Porovnanie so softvérom Pilon

| | *Spades* | *Miniasm* |
|---|---|---|
| *Celá sekvencia* | 0.39% | 1% |
| *Referenčné chyby* | 3.7% | 38% |
| *Limit1* | 5% | 19.6% |
| *Limit3* | 2.4% | 8.4% |
| *Limit5* | 2.5% | 7.5% |
| *Limit7* | 2.4% | 8.1% |

# Možnosti ďalšej práce

- Zlepšenie zarovnaní ku grafu
- Objavenie ďalších filtrácií chýb
- Zložitejšia filtrácia homopolymérov
- Testovanie ďalších dát

# Ďakujem za pozornosť

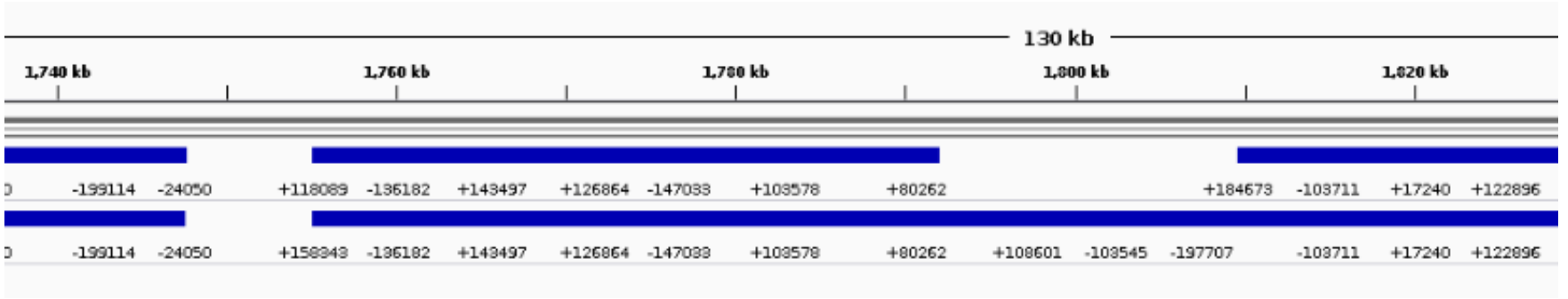
# 1. Ako vel'mi by sa zmenili výsledky, ak by sme uvažovali realistickejšie prekryvy (napr. 50%)?

| | *Spades* | | *Miniasm* | |
|---|---|---|---|---|
| *Presnosť* | | | | |
| *Prekryv* | *1bp* | *50%* | *1bp* | *50%* |
| *Limit1* | 77.3% | 65.2% | 74.3% | 55.1% |
| *Limit3* | 71.4% | 65.5% | 67.8% | 47.5% |
| *Limit5* | 56.4% | 51.9% | 67.4% | 51.5% |
| *Limit7* | 38.4% | 35% | 66.2% | 51% |
| *Úplnosť* | | | | |
| *Prekryv* | *1bp* | *50%* | *1bp* | *50%* |
| *Limit1* | 28% | 24.1% | 8.2% | 6.8% |
| *Limit3* | 69.4% | 65.8% | 10.4% | 9% |
| *Limit5* | 77.9% | 75.2% | 12.6% | 11% |
| *Limit7* | 82.5% | 77.3% | 14% | 12.7% |

# 2. Čím si vysvetľujete nekonzistentnosť v zarovnaní pôvodného a reverzného vlákna programom GraphAligner? Je k tomuto niečo v dostupnej literatúre, popr. kontaktovali ste tvorcov programu?

- Problém s nezarovnávaním rozsiahlejších kusov sekvencie
- Potenciálne súvisiaci problém uvedený na GitHube

# 2. Čím si vysvetľujete nekonzistentnosť v zarovnaní pôvodného a reverzného vlákna programom GraphAligner? Je k tomuto niečo v dostupnej literatúre, popr. kontaktovali ste tvorcov programu?



# 3. Ktoré z uvedených filtrácií hodnotíte ako najperspektívnejšie pre ďalšie použitie?

- Filtrácia homopolymérov

# 4. Aká bola dĺžka behu, bola by rovnaká schéma ako na Obr. 3.1 použiteľná aj pre väčšie genómy, napr. ľudský?

- Nie
- CPU servre
  - 47-142GB pamäte
  - 16-48 procesorov