VYUŽITIE DLHÝCH SEKVENAČNÝCH ČÍTANÍ PRI BIOINFORMATICKEJ ANALÝZE HYBRIDNÝCH GENÓMOV

Bc. DOMINIKA SZABOVÁ

Študijný odbor: Informatika, magisterský

Školiteľ: doc. Mgr. Bronislava Brejová, PhD.

V Bratislave dňa 31.8.2022

ANALÝZA HYBRIDNÉHO GENÓMU

- Hybrid je organizmus, ktorý vznikol krížením dvoch organizmov rôzneho druhu
- Referenčná vzorka nie je jednoznačná

MOTIVÁCIA A CIEĽ

- Nájsť spoľahlivý postup na analýzu a vylepšovanie už zostaveného genómu
- Odhaliť chyby v zostavení

DÁTA

HLAVNÝ GENÓM

- *Loderromyces elongisporus* CBS 5301

SYNTETICKY PRIPRAVENÉ GENÓMY

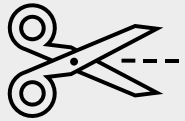
- *Escherichia coli* WTP3B1-WTP2B1 (EC32)
- *Escherichia coli* WTP1.2A-WTP2A (EC212)
- *Escherichia coli* WTP1.3A-WTP2A (EC213)

Vstupné dáta:

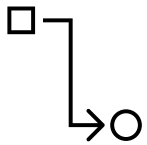
- Dlhé čítania – Oxford Nanopore Technologies (MinION), PacBio
- Kontigy – Illumina + ONT

METODOLÓGIA

HLAVNÁ ČASŤ PRÁCE – NAŠA PIPELINE



Spracovanie čítaní a segmentácia

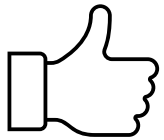


Mapovanie segmentov

Filtrácia

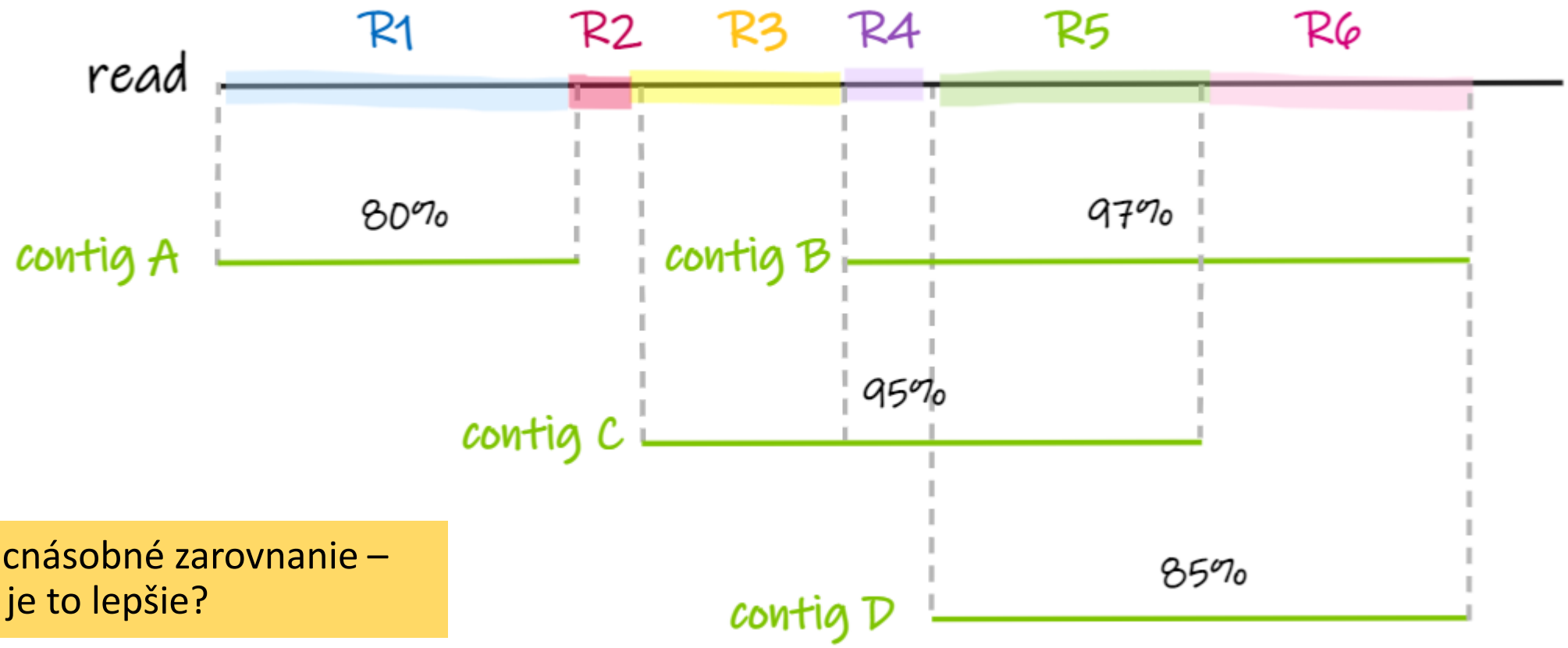
VEDĽAJŠIE ÚLOHY

Výpočet a analýza medzier v kontigoch a v zarovnaniach

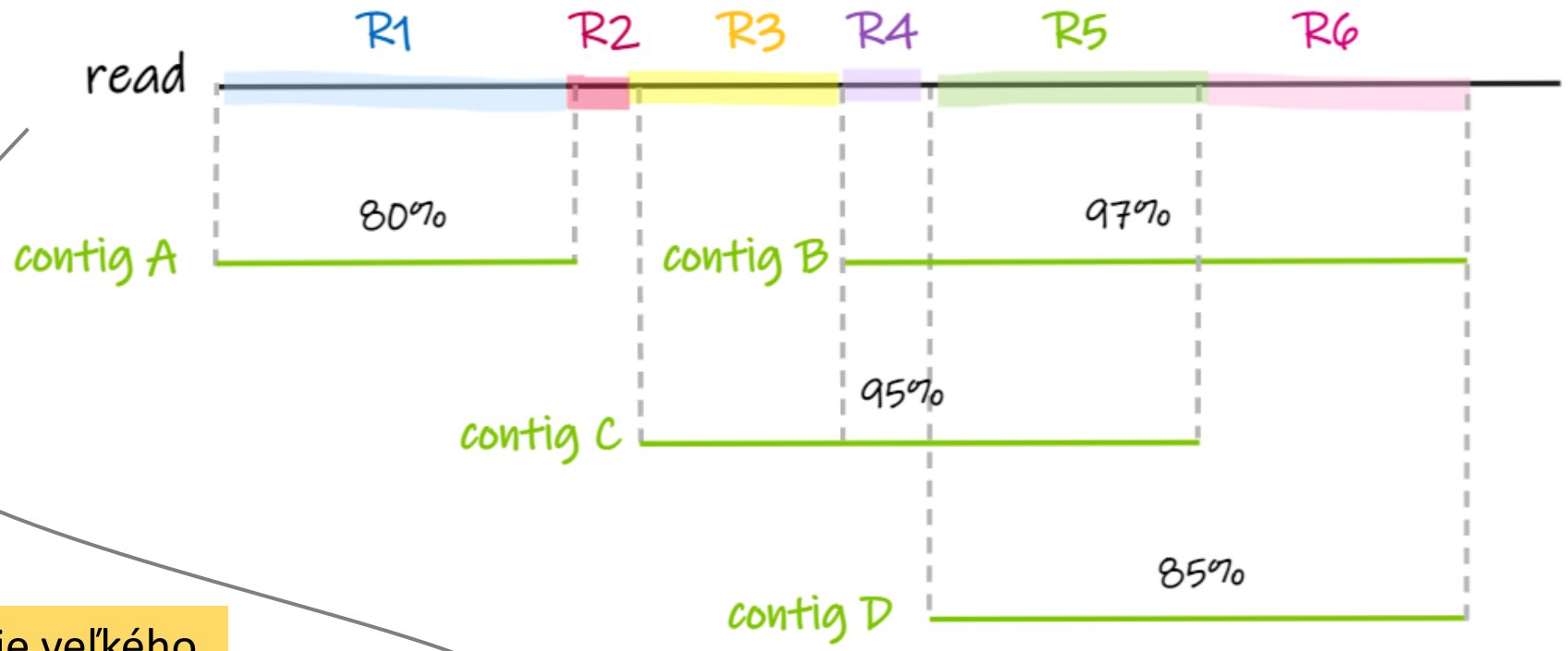


Overenie správnosti s iným SW

NAŠA PIPELINE

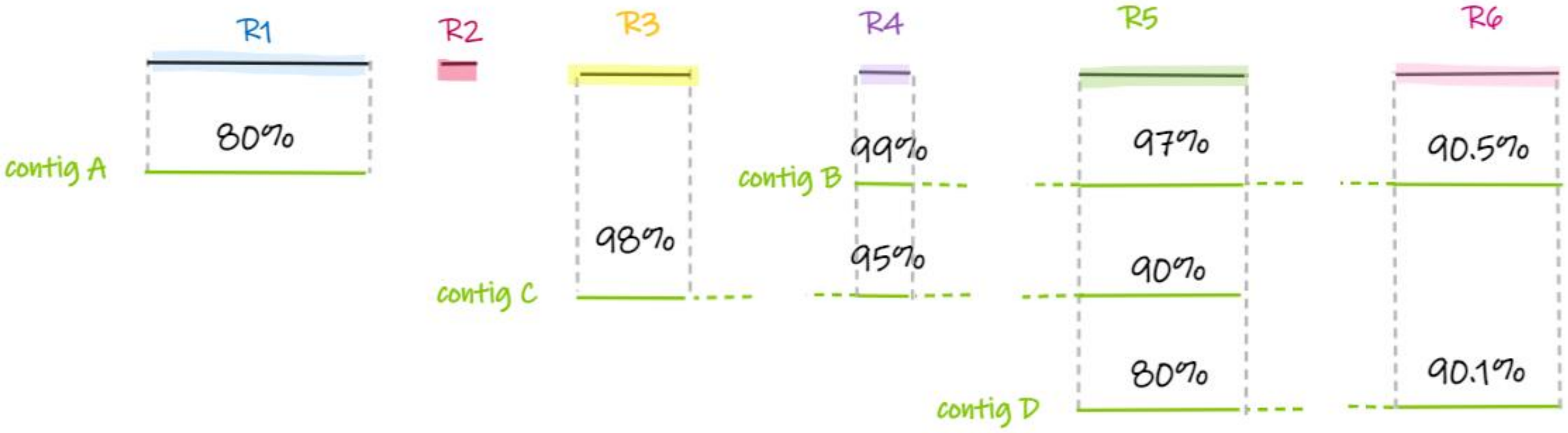


1. Viacnásobné zarovnanie – ktoré je to lepšie?

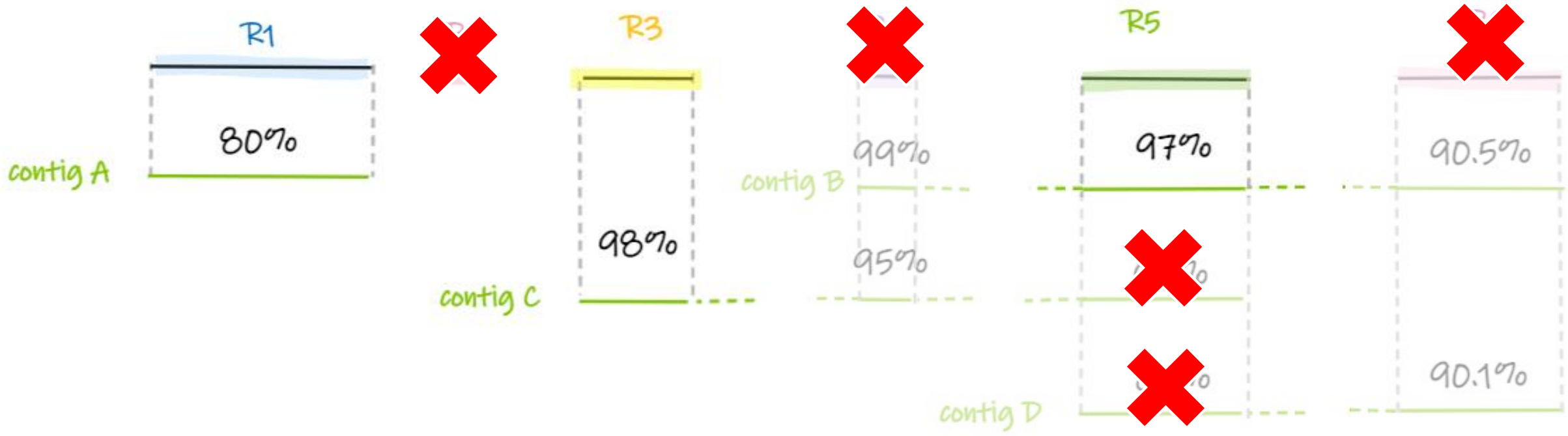


1.2. Redukovanie veľkého problému na menšie = segmentácia

2. Mapovanie segmentov



3. Filtrácia dlhých a jednoznačných mapovaní

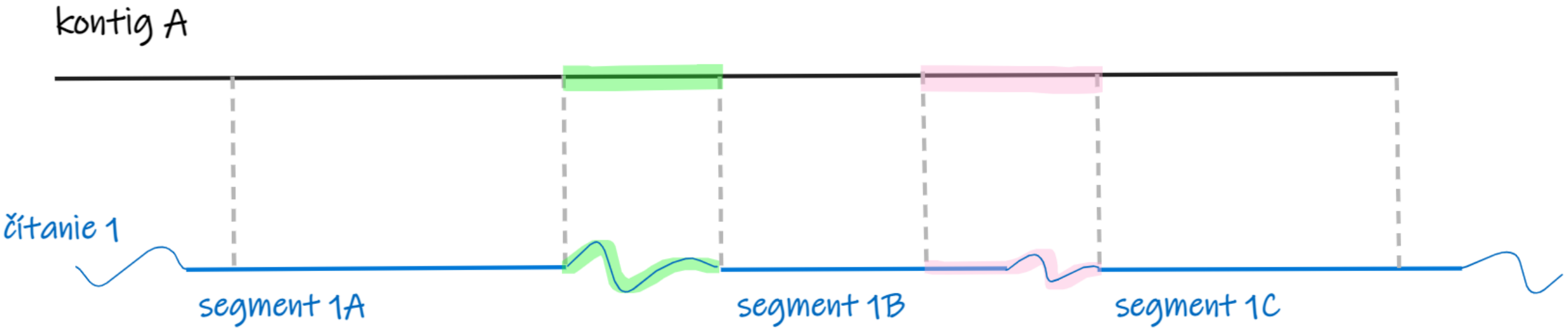


VÝSLEDOK: zoznam akceptovaných mapovaní segmentov (čítaní) a kontigov

Bc. D. SZABOVÁ, 31/08/2022

ANALÝZA MEDZIER

Výpočet a analýza medzier



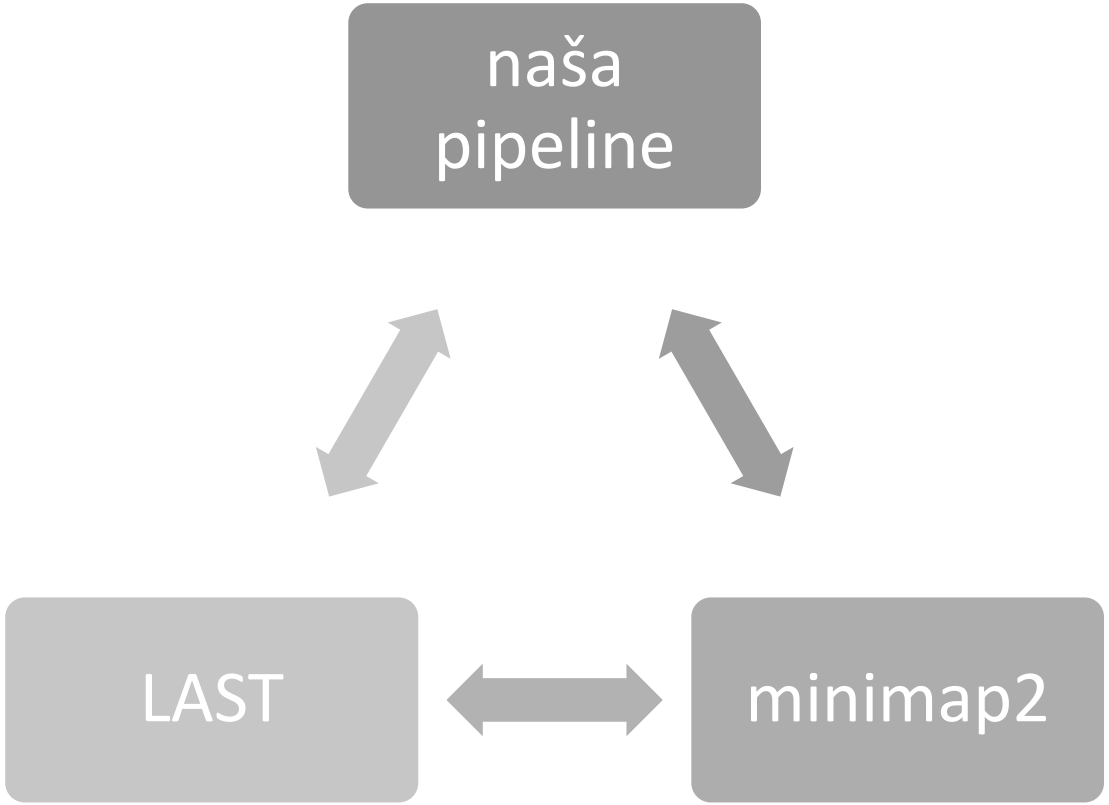
MOŽNÉ SCENÁRE:

- Medzery sú rovnaké (akceptovateľná odýchlka X báz)
- Medzera v kontigu je kratšia
- Medzera v segmente je kratšia
- V kontigu sa táto oblasť cez seba prekrýva
- ...

1. Je pomer dĺžky medzery približne 1:1?
2. Našiel tieto medzery aj iný SW na hľadanie medzier?

VÝSLEDKY

VÝSLEDKY
NAŠA PIPELINE - OVERENIE SPRÁVNOSTI



OVERENIE VOČI BIOINFORMATICKÝM NÁJSTROJOM

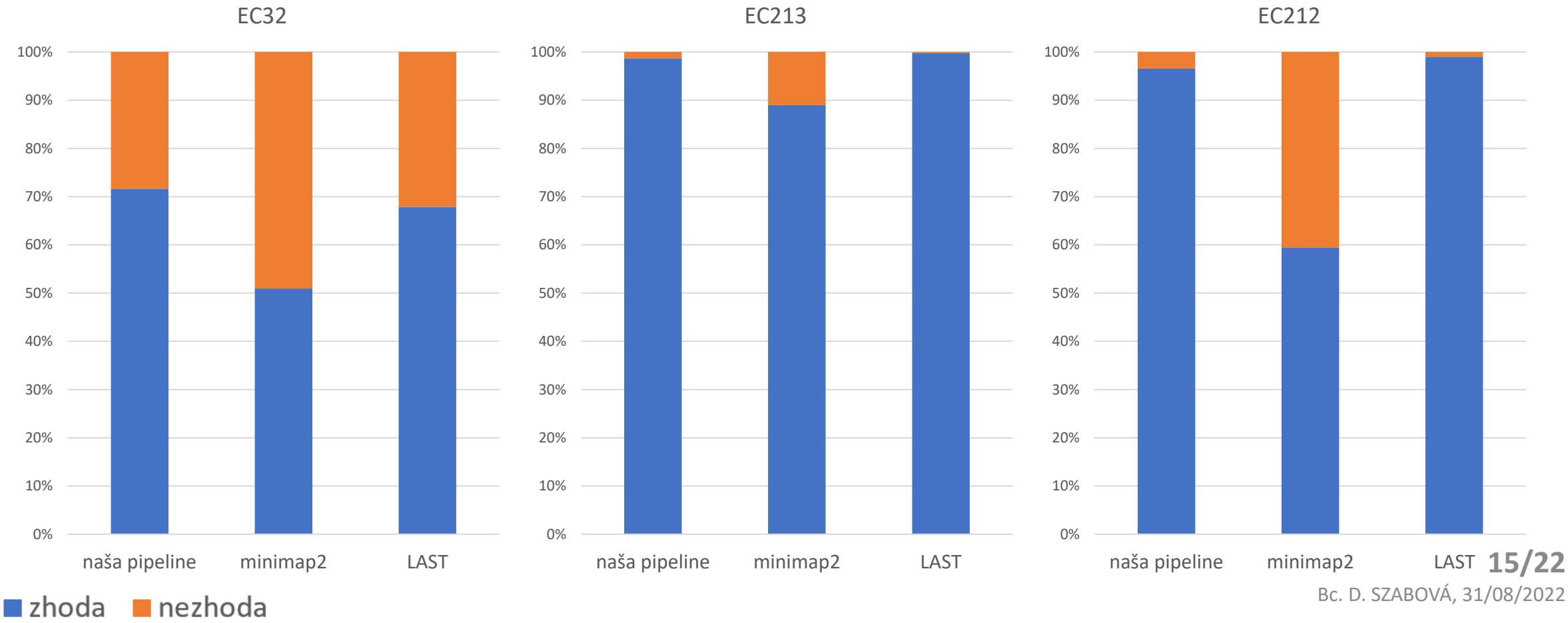
- Štandardné nástroje by mali ukázať podobné výsledky ako tie naše
- Ak nie, kde robíme chybu?
- Ako zmeniť prístup?

Počet zarovnaných čítaní a kontigov z pôvodných dát

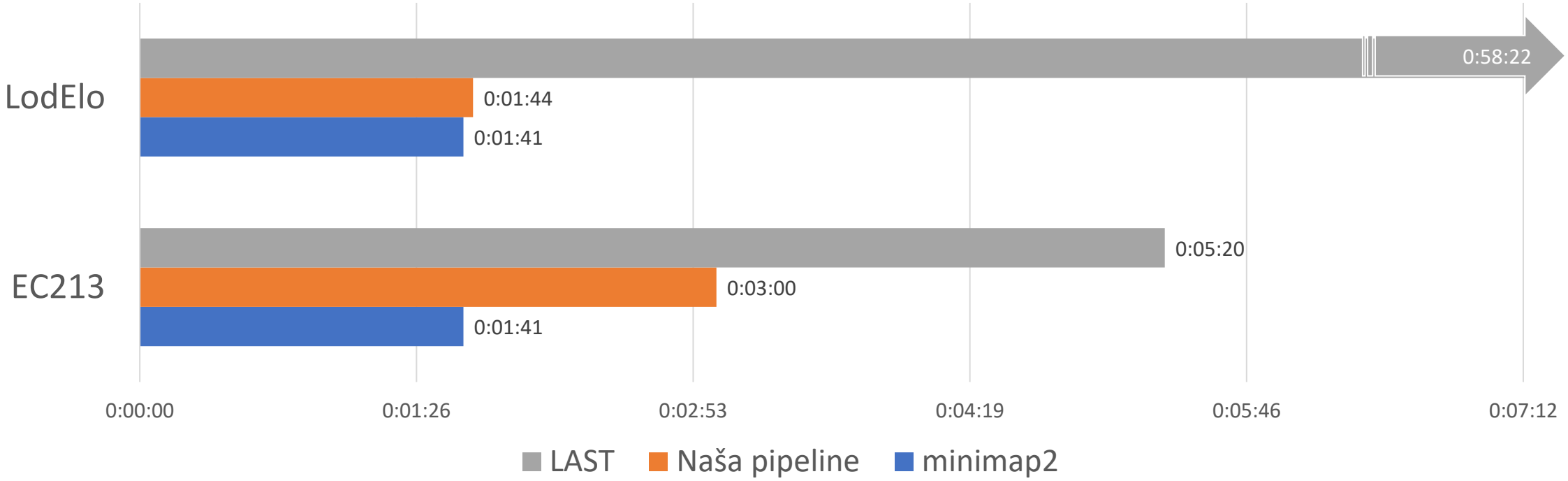
	Počet zarovnaných čítaní	%	Počet zarovnaných kontigov	%
EC32				
naša pipeline	1070	0.72%	10	55.56%
minimap2	62709	42.16%	13	72.22%
LAST	7682	3.86%	7	72.22%
EC212				
naša pipeline	1350	10.92%	13	100.00%
minimap2	5741	38.95%	13	100.00%
LAST	3955	26.84%	13	100.00%
EC213				
naša pipeline	3018	19.51%	10	100.00%
minimap2	4439	28.70%	10	100.00%
LAST	9922	64.15%	10	100.00%
LodElo				
naša pipeline	75955	52.31%	487	7.17%
minimap2	76265	52.52%	344	5.06%
LAST	111745	76.96%	554	8.16%

Overenie pomocou syntetických genómov

lr-WTP1.2A-SRR12298719.1739:3752-6340	2588	0	2445	+	cn-WTP1.2A-CP056412.1	5121078	1931998	1934454	2325	2519
lr-WTP1.2A-SRR12298719.1739:3752-6340	2588	0	2445	-	cn-WTP2A-CP056625.1	4945124	3264531	3266987	2326	2517
lr-WTP2A-SRR12298910.1136:11866-22941	11075	0	11075	+	cn-WTP2A-CP056625.1	4945124	4695848	4706957	10548	11390



Časová náročnosť



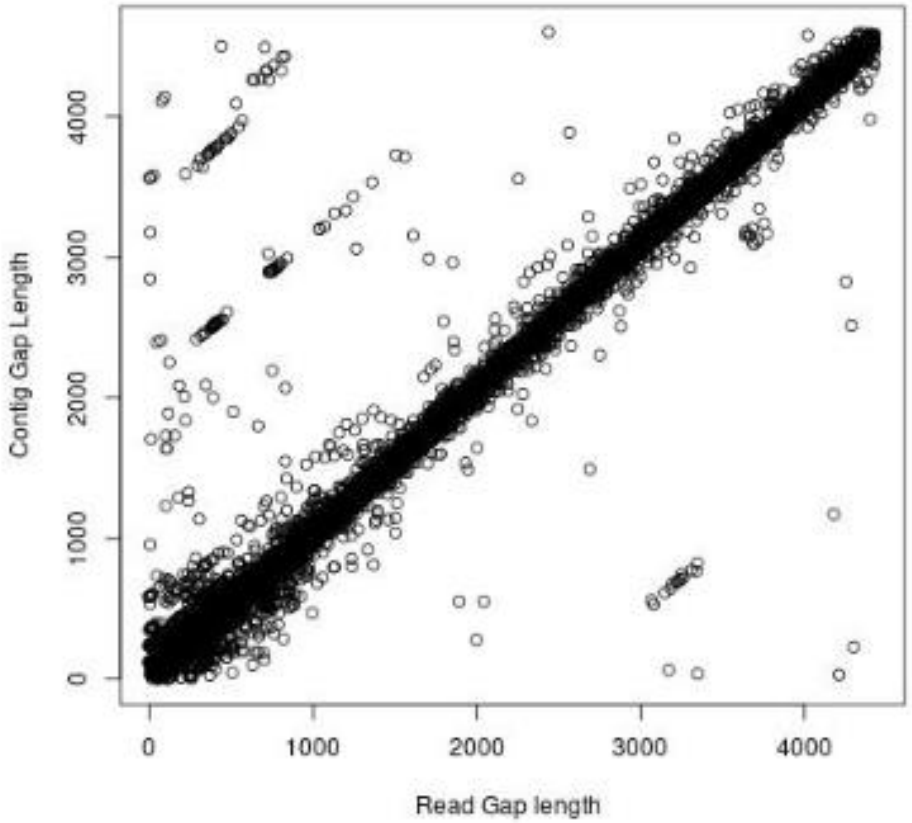
- ✓ Merania s náhodnou podmnožinou čítaní
- ✓ Rovnaké CPU
- ✓ Graf:
  - EC213 = 7.7K čítaní (50%), priemerná dĺžka 3 514.22 báz
  - LodElo = 10K čítaní (6.8%), priemerná dĺžka 13 001.3 báz

VÝSLEDKY
ANALÝZA MEDZIER

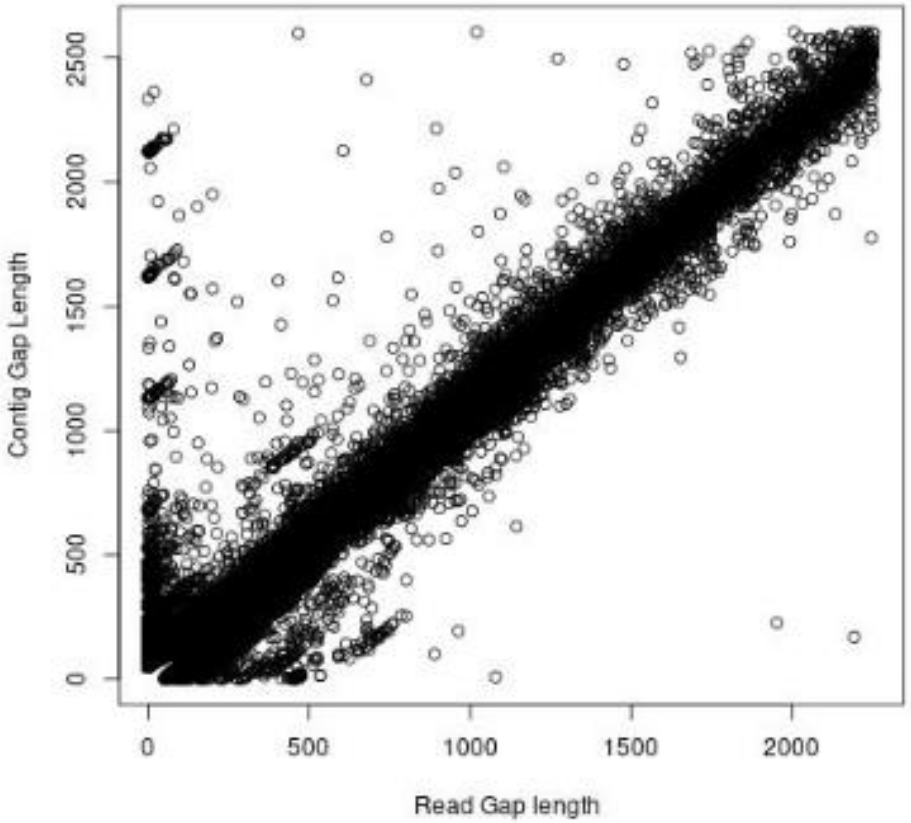
	Počet záznamov	Počet unikátnych čítaní	Počet medzier	Priemerná dĺžka medzery v čítaní	[bedtools] identifikované chyby	[bedtools] identifikované chyby maximálne pokrytie
EC213						
naša pipeline	3782	3018	250	3908.79	7	3.5
LAST	111285	9922	611	457.971	10	3
LodElo						
naša pipeline	164507	75955	27984	2217.75	3204	168
LAST	225333	111745	41020	1128.48	5384	420

Pomer dĺžky identifikovanej medzery v čítaní a kontigu v *LodElo*

naša pipeline



LAST



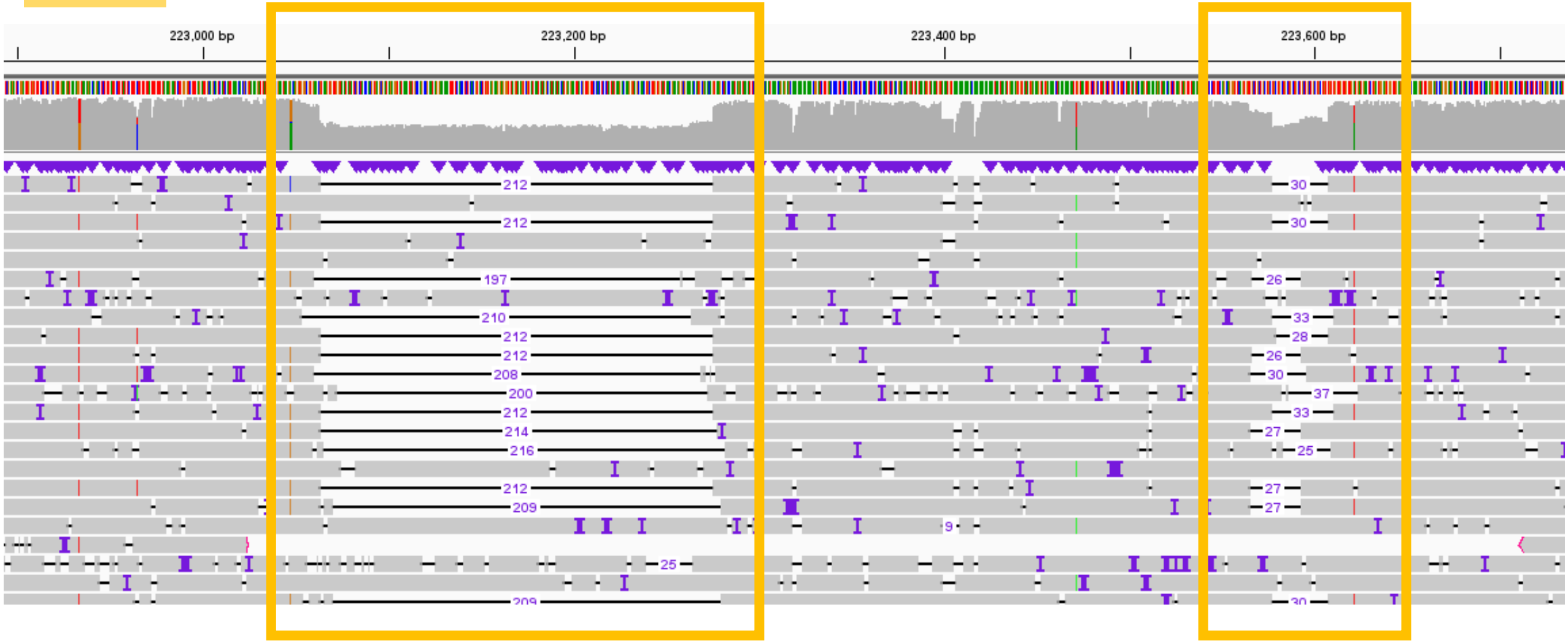
Porovnanie našich objavených medzier s výsledkami iného SW

Chyby identifikované v našom zarovnaní
- Pozície v kontigoch, kde je možná chyba
Inspector
- Výsledok mapovania kontigov a čítaní *Lodelo* a *EC213*

IGV

Porovnanie našich objavených medzier s výsledkami iného SW

LodElo



ZÁVER

- Naša pipeline:
  - Podobné výsledky ako štandardné nástroje na mapovanie čítaní
  - Rýchlejšia než LAST
- LAST mal najlepšie výsledky aj napriek striktnejšiemu základnému algoritmu
- Syntetické genómy *E. coli*
  - Dopomohli overiť správnosti našej pipeline
  - Porovnateľne dobré výsledky nájstroja LAST a našej pipeline
- Zhodné medzery identifikované z dát našej pipeline a nástrojom Inspector

ĎAKUJEM ZA POZORNOSŤ

Vysvetlenie pojmov (dotaz v posudku od školiteľky)

Genome assembly (zostavovanie genómu)	Rekonštrukcia pôvodného genómu, ktorá sa riadi zhodou prekrývajúcich časti čítaní alebo zhody s referenčnou vzorkou.
Sequence alignment (zarovnanie sekvencií)	Metóda, ktorá zarovná dve (alebo viac) sekvencie pod seba tak, aby rovnaké bázy boli pod sebou. Cieľom je určiť presný **rozdiel** medzi dvoma (alebo viac) sekvenciami až na úroveň báz.
Read mapping (mapovanie čítaní)	Zarovnanie sekvencií ku referenčnej vzorke na základe zhody báz. Hľadá približnú **pozíciu** sekvencie v referenčnej vzorke. Sekvencie nemusí byť zarovnaná celá, stačí lokálne.

- Hľadanie pôvodnej pozície čítania/segmentu v kontigu pomocou mapovania
- Úroveň zhody namapovaného čítania/segmentu s jedným alebo viac kontigmi vo fáze 3 našej pipeline
- Vylepšenie už zostaveného genómu za pomoci výsledkov z analýzy medzier

Ďalšie dotazy v posudkoch

- Chýbajúca elektronická príloha s kódom.

  Moje nedopatrenie. Napriek tomu, však v metodológií je pri každom kroku príklad spustenia nástrojov a skriptov, ich funkcia a detial vstupno-výstupných dát.

- Z pohľadu textu je úplne jedno, či daný vstup je typu FASTA, ... Podstatné je, že to je „referencia hybrida“ alebo „referencia predka“ alebo „čítania z hybrida“

  Uvedenie prípony súboru pomáha čitateľovi pripraviť svoje dáta pre použitím daného nástroja.

- Kde v procese využívame, že daný genóm je hybrid? Kebyže vyrobený nástroj spustím na dátach z obyčajnej Ecoli, čo sa stane?

  Naša pipeline je použiteľná aj na štandardný genóm, avšak pridaná hodnota je práve pre hybridy, ktoré majú problém s validáciou genómu z dôvodu nejednoznačnej referenčnej vzorky.