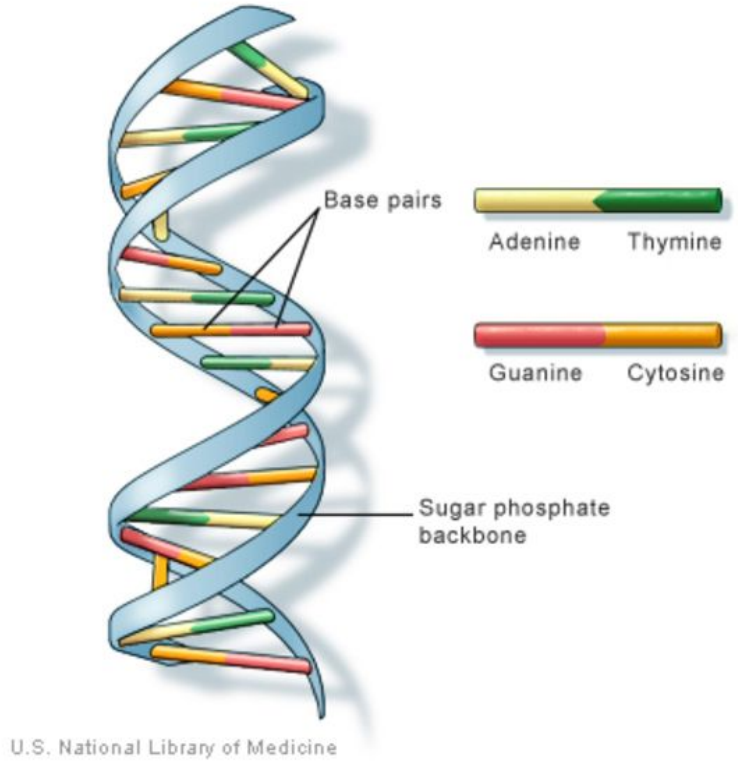
Algoritmy pre dynamické zostavovanie nanopórových čítaní

Jana Černíková
školiteľ: doc. Mgr. Tomáš Vinař, PhD.

Biologické pozadie problému



TTGCCGCGCACTCGATATTGCGCTGCCGGAC
CGAGATTGCGGCCTGTCGCTGGGGTTACCGA
GGCAATGCCGACAGCGGCAATATCGGCCGGC
GCGCGAAAATCTCGCCGACAAAACCAGCGCA
CGTCGCCTTAATCAATGCGCCTGAATCTGGC
GGGATATGCGCAGTCGCCGACAGCGGCAATA
TCGGCCGGCGCGCGAAAATCTCGCCGACAAA
ACCAGCGCACGTCGCCTTAATCAAGTGGAAG
GAGATAGAGGATATACACACCACCACCACTGA
GATTTAATCAATGCGCCTGAATCTGGCGGGAT
ATGCGCAGTCGCCGACAGCGGCAATATCGGC
CGGCGCGCGAAAATCTCGCCGACAAAACCAG
CGCACGTCGCCTTAATCAAGTGGAAGGAGAT
AGAGGATATACG….

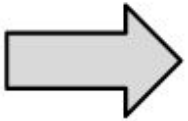
Čítania (reads)

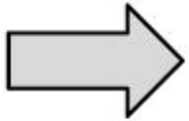
@6bfee659-58a2-4870-ad8f-f37b1f37279f runid=37ed0c95ec68e8222ec9d629d5e1715ef411b250 read=101 ch=394 start_time=2018-10-10T09:41:10Z
CCCGTCCGTTCCATTCATTCCGATCATTCCGGTCCATTCCATTCCATTCCATCCATTCCATTCCAGGAGTCCATTCCATTCCAT
TCCATTCCTCGAGATCCGGTCAGTCCGGTCCATTCCATCCGTCCA

@910d707b-a589-487b-9766-fe40fa349ce0 runid=37ed0c95ec68e8222ec9d629d5e1715ef411b250 read=101 ch=293 start_time=2018-10-10T09:39:04Z
TTGCCGCGCACTCGATATTGCGCTGCCGGACCGAGATTGCGGCCTGTCGCTGGGGTTACCGAGGCAATGCCGACAGCGGC
AATATCGGCCGGCGCGCGAAAATCTCGCCGACAAAACCAGCGCACGTCGCCTTAATCAATGCGCCTGAATCTGGCGGGATAT
GCGCAGTCGATTTGAGGTTCAAGAACCAGAAGAATACCGAGACATCTGGTCAAGGACGCCGGAATGTTTTCGGTCGAGAGAG
AACAGCGTCAGGATATATATTGAAATATTTTATATTACACCAGCCAGCATATTTTTATTGAGAAATTAAGTCTCTCTCTTCCTTCA
TTCGATCTCAAGTACAATATTAACGCGAGAGCGGCGGCGACATCAGAGCGTGGTACTGGTAGTGTGGCACTGAAGTATTATTT
TGTCTTCCTGAAGAAGTGAAGCTGCGTTCGGCTGTTGAAAATAAAGAGCGATGAATGAATGAAATGAATTCAATGGCTGTAAC
AACACCGGCTTTACATTTCACACCGTGACTACATTTTGAAGCCAGCGTATGAGCTGTATGCTGTGCGTAAATGGAAGCGTCTT…

@18ef32ab-3feb-4163-a853-56e58c4c5ae1 runid=37ed0c95ec68e8222ec9d629d5e1715ef411b250 read=100 ch=481 start_time=2018-10-10T09:39:53Z
TTGGGCAGCCTGAATTGGGTAAAAATGACTGAATAGTATTGTATTTTATGCGATGATAATTAAATTTTTATTTTCCATTTATTTAT
TTTTTTCGAAACTCTGTAAAAGATTTTTATTTTATTTGTTATTATTTTTGATTTAGGGCGTTATGACGGCGGCGATAGCGGCAGT
ATTGTAGTAAGCGTTGAAACGATGACGTGATCAAAATTGTAAATCGGCGTCGGTCGAGGCGAGGGCAGTGTCAAAGTATATTC
AAAGTCTGTCGCGATCGTATTGTCAGTATTGATAGTATTGATAAGTATTGTCCCGTGTCCCGTGTGAGTAGAATGAAGTCGAAA
TTGAAAGTATGTGTACTATTACATGGCATTTTTTATTGTTATTTGGCCCTGAAATCGACCGTGAAATGAATAAATGAAGAGAGA
GA…

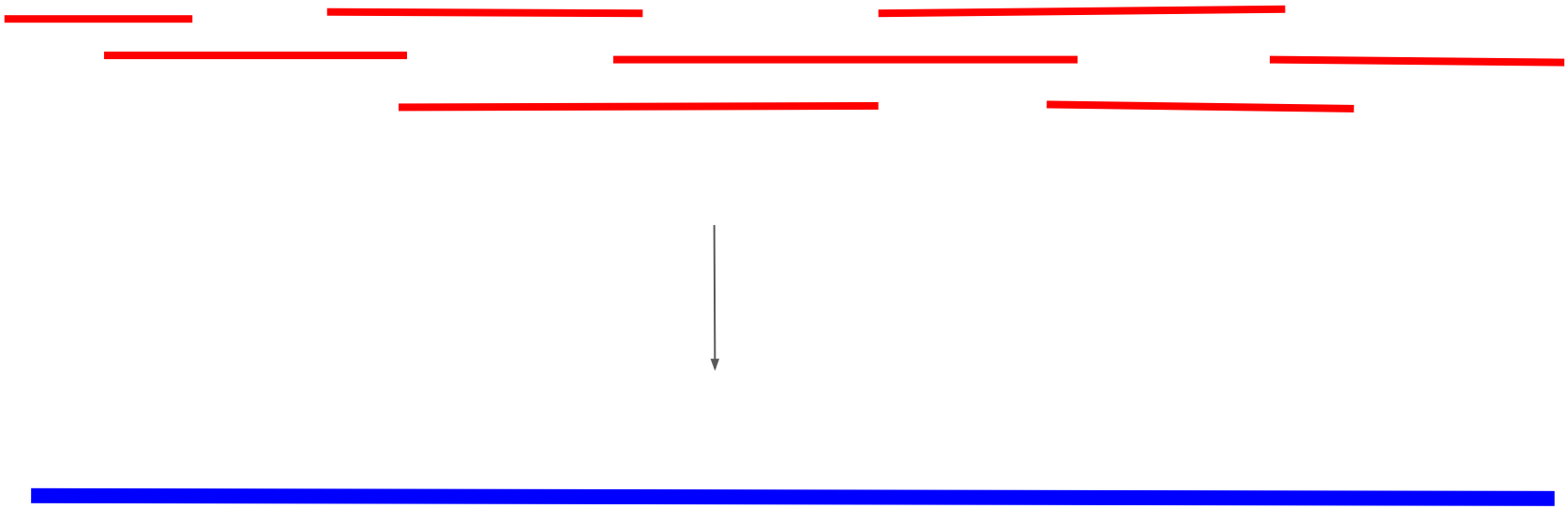
Skladanie genómov





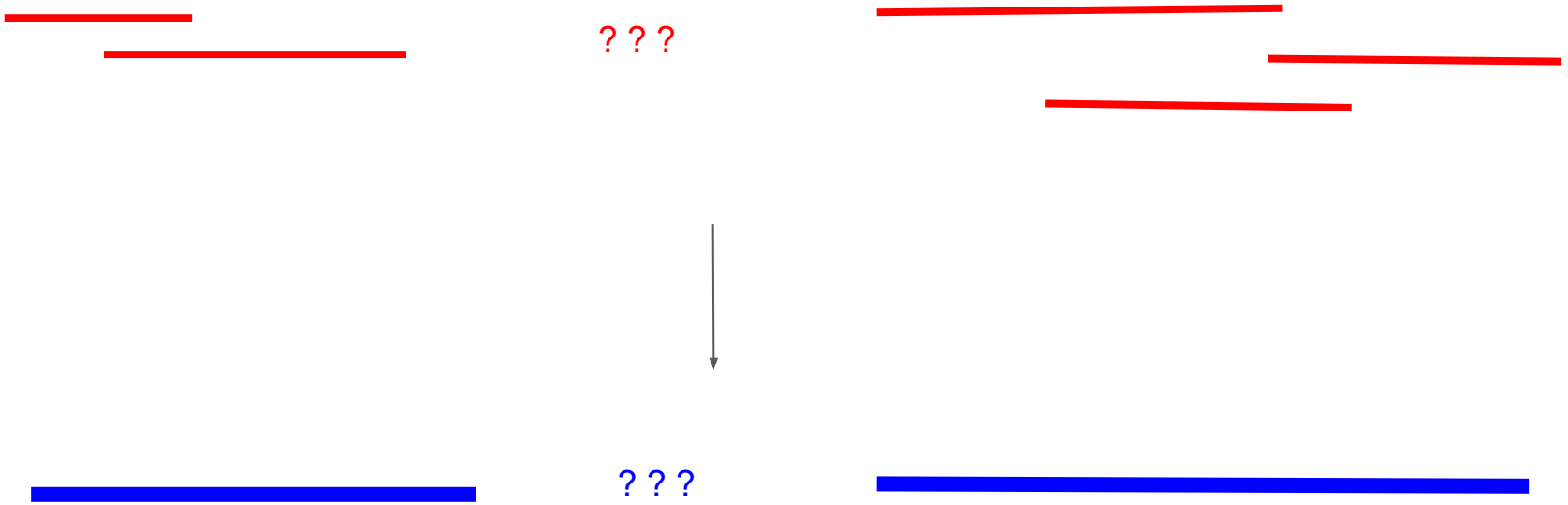
Zostavenie (assembly) genómu

(chceli by sme)



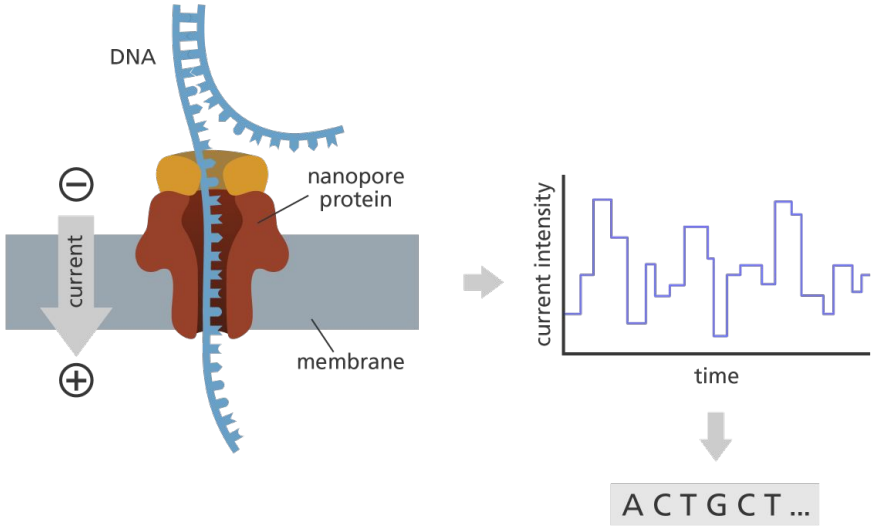
Zostavenie (assembly) genómu

(máme)



Statický vs. dynamický problém

https://www.genengnews.com/insights/first-nanopore-sequencing-of-human-genome/



https://www.yourgenome.org/facts/what-is-oxford-nanopore-technology-ont-sequencing/

“Kedy máme dostatočné množstvo dát?”

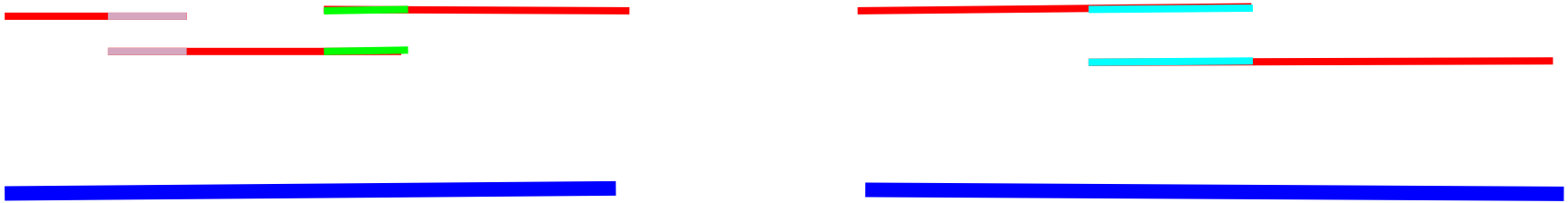
cieľ práce: chceme odpoveď v reálnom čase

Typický nástroj na skladanie genómov

- rieši “statický” problém – všetky dáta dostane naraz
- rôzne heuristiky + dynamické programovanie + grafy
- overlap-layout-consensus prístup

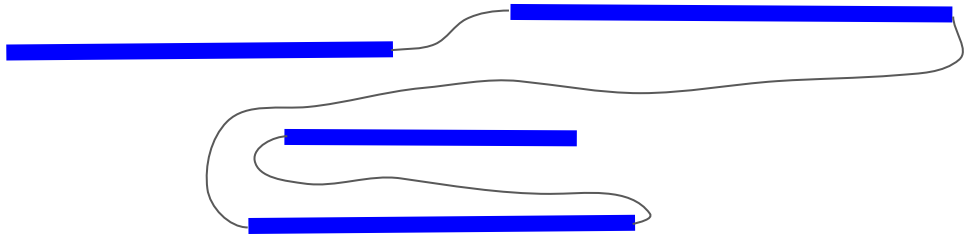
Overlap-layout-consensus

- **overlap:** hľadáme zarovnania (prekryvy), budujeme contigy



Overlap-**layout**-consensus

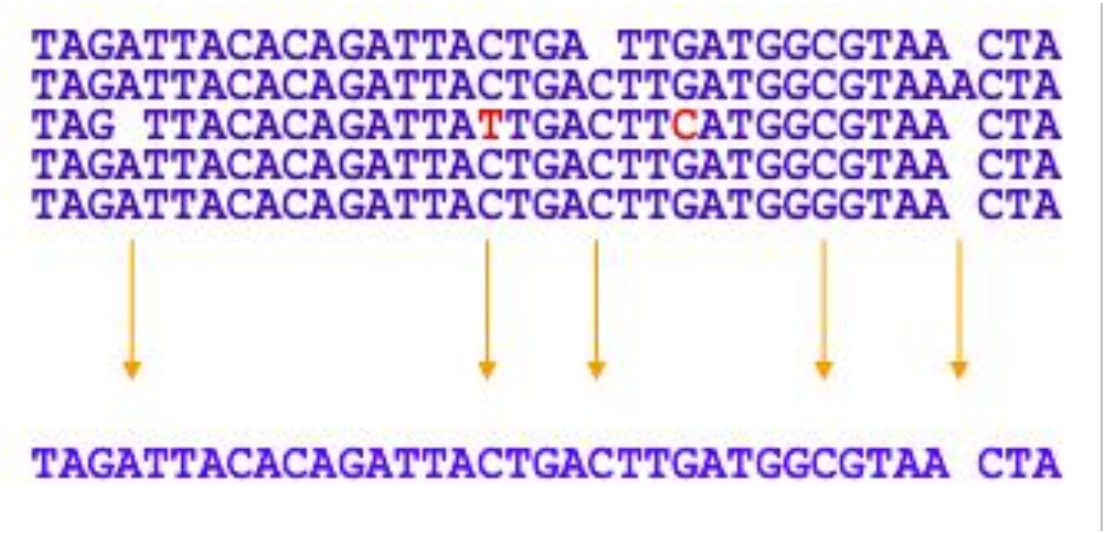
- **layout:** cieľ je zistiť vzájomnú orientáciu contigov, vzdialenosti medzi nimi



supercontig

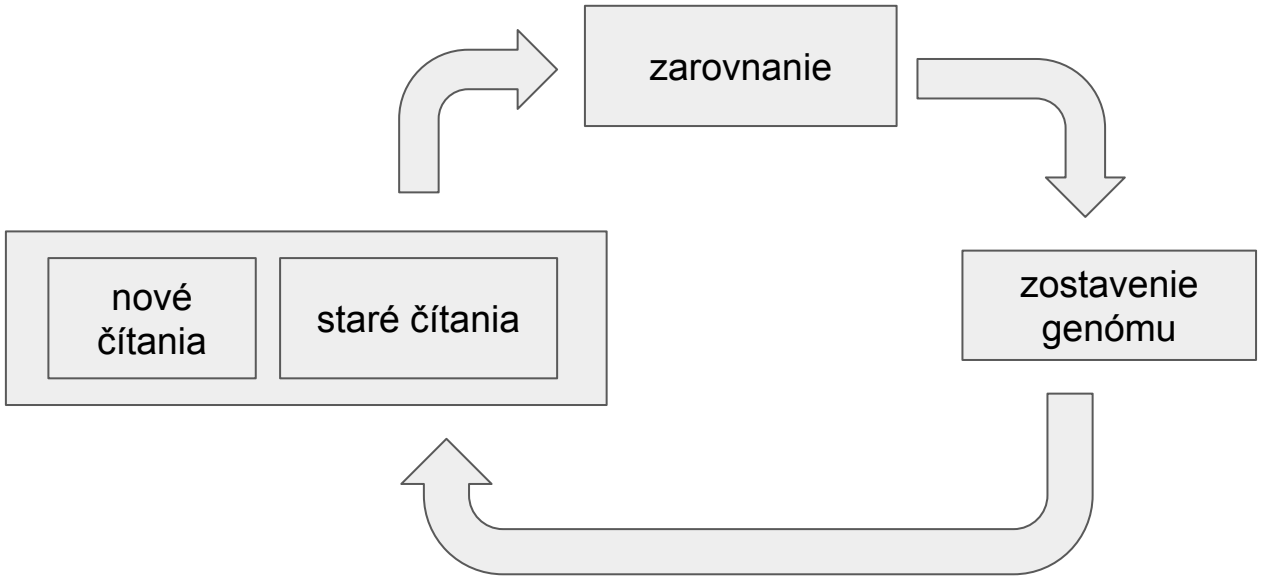
Overlap-layout-**consensus**

- **consensus:** presnejšie určenie báz na základe nejakého "*väčšinového pravidla*"

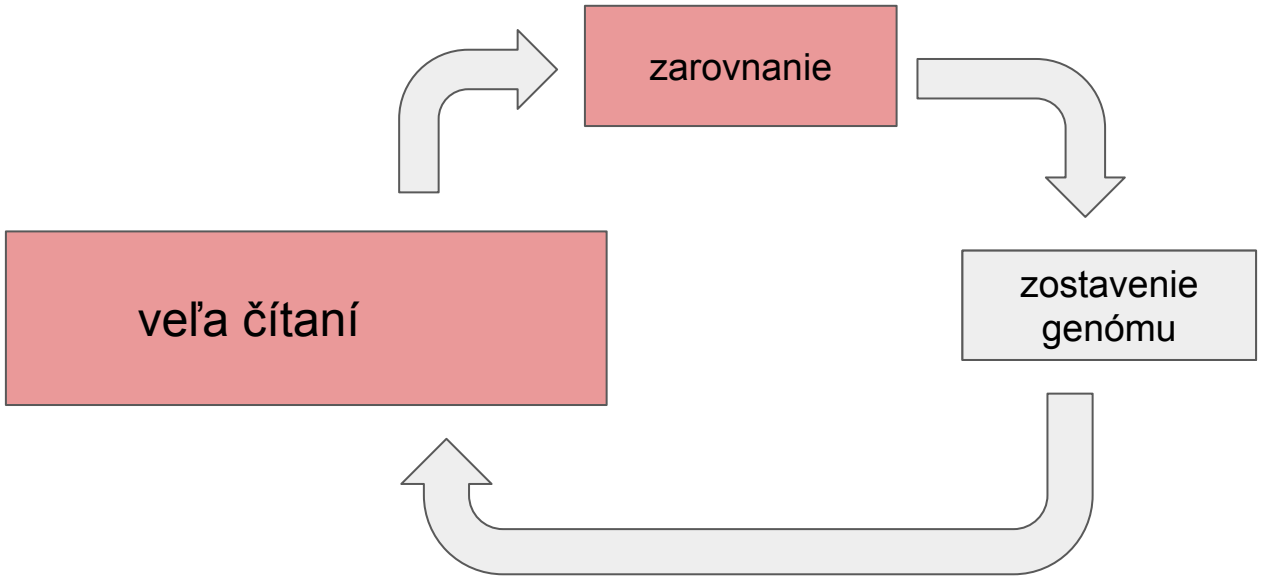


https://bio.libretexts.org/Bookshelves/Computational_Biology/

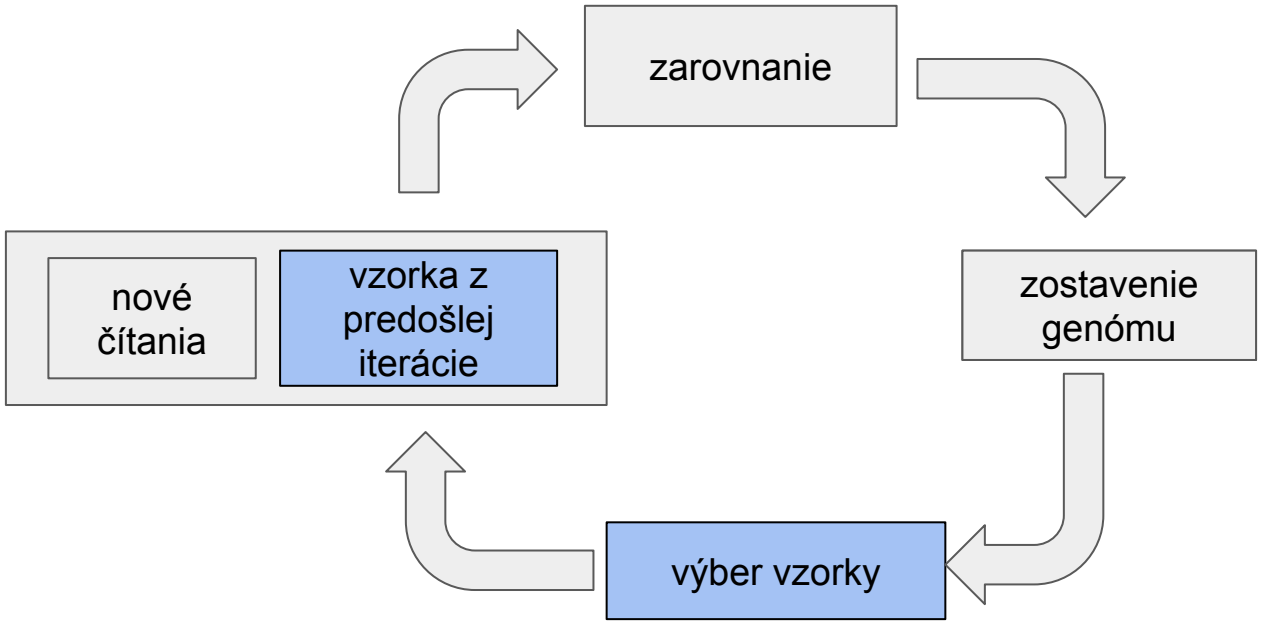
Dynamický prístup (pomalý)



Dynamický prístup (pomalý)



Náš prístup k dynamickému problému



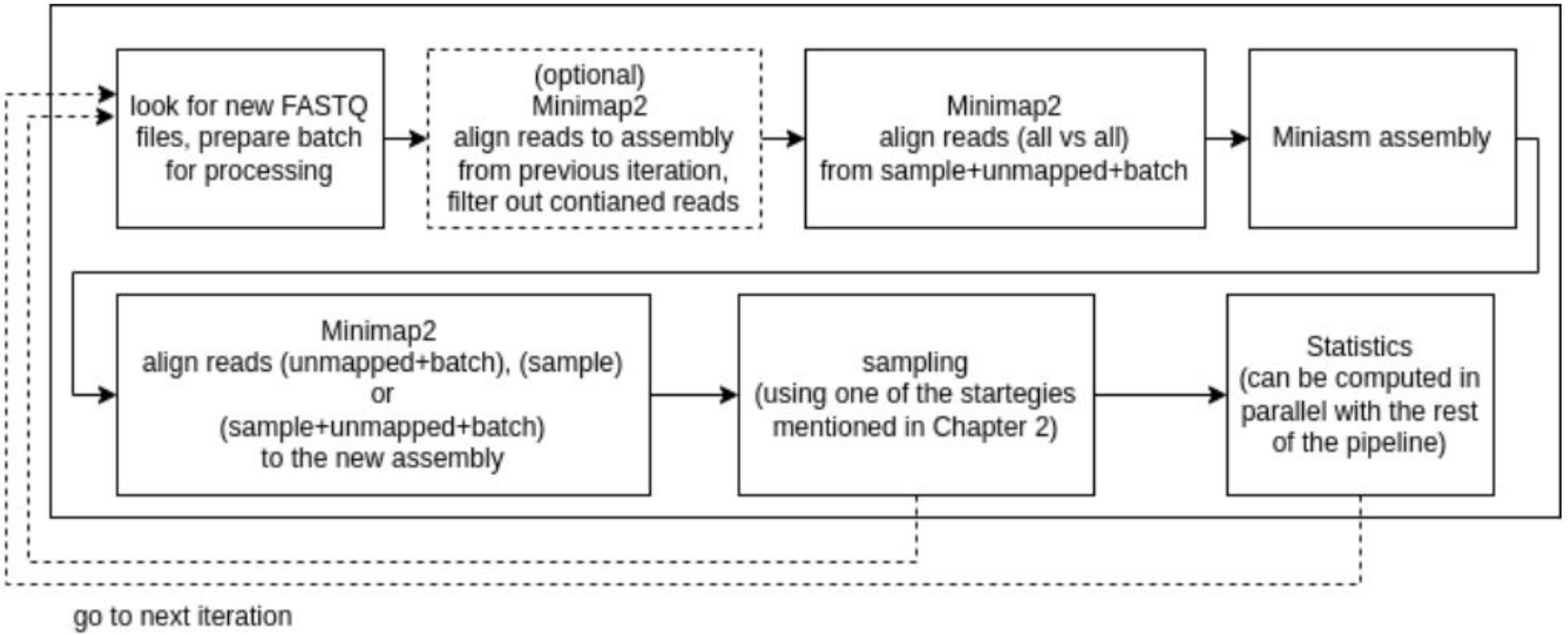
Použité nástroje

- zarovnanie (overlap): Minimap
- zostavenie genómu zo zarovnaní (layout): Miniasm
- consensus: –

- rýchle
- často používané

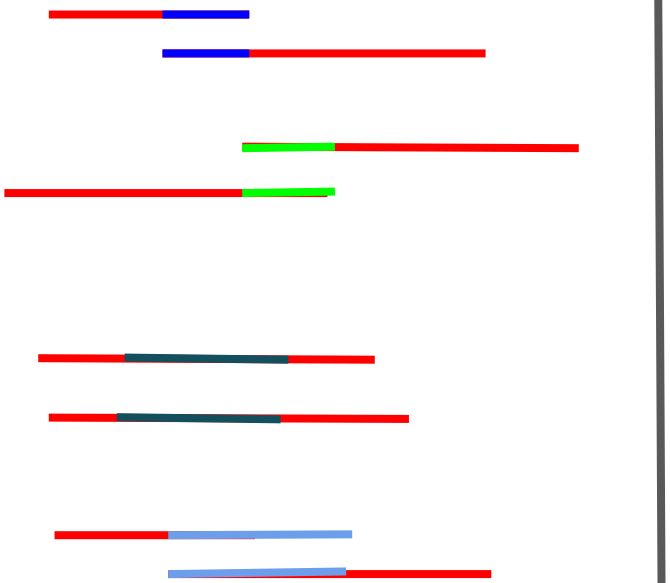
Heng Li, Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences , Bioinformatics, Volume 32, Issue 14, 15 July 2016, Pages 2103–2110, https://doi.org/10.1093/bioinformatics/btw152

Dynamický prístup s využitím Minimap & Miniasm



Zarovnanie (alignment) [Minimap]

čítania k čítaniam



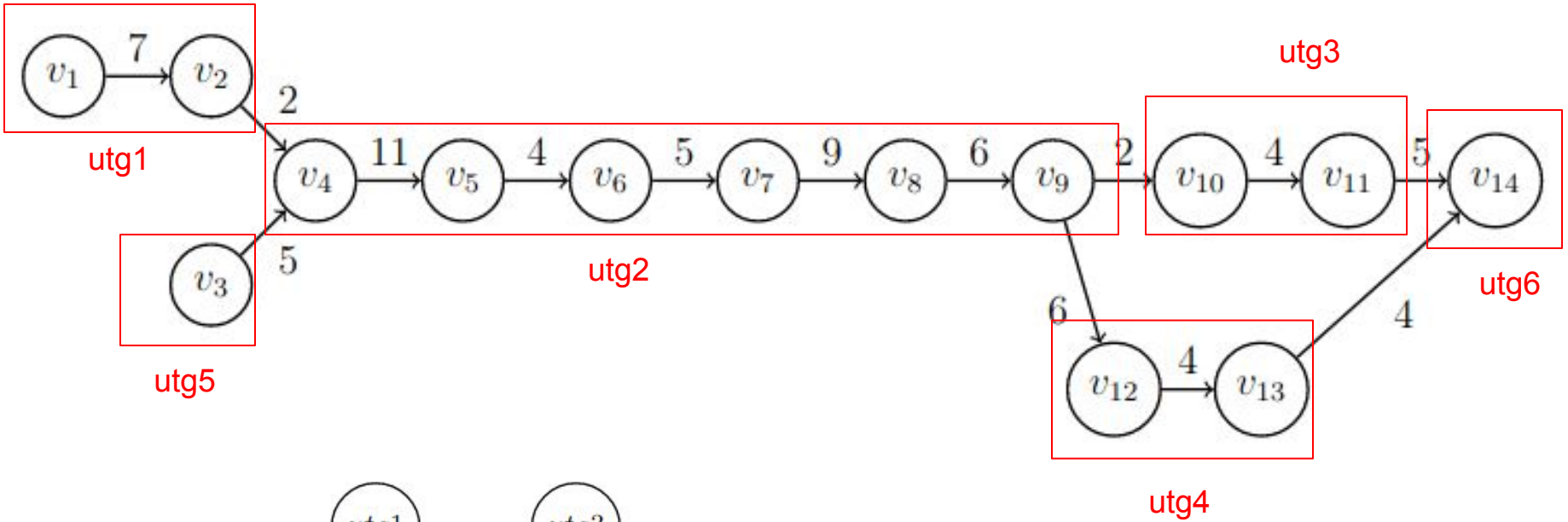
Assembly graph [Miniasm]

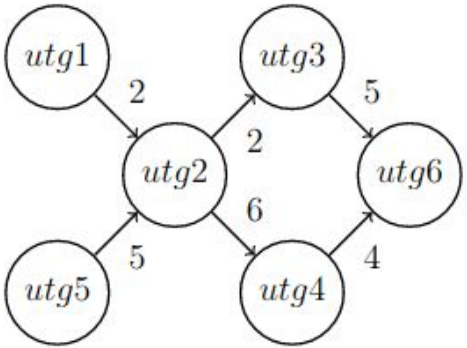
- $G = (V,E,\ell)$
- V …. množina DNA sekvencií
- E …. množina prekryvov (overlaps) medzi sekvenciami z V
  - bez násobných hrán
- ℓ: $E \rightarrow R^+$

- Containment-free - žiadna sekvencia nie je obsiahnutá v inej

overlap

contained

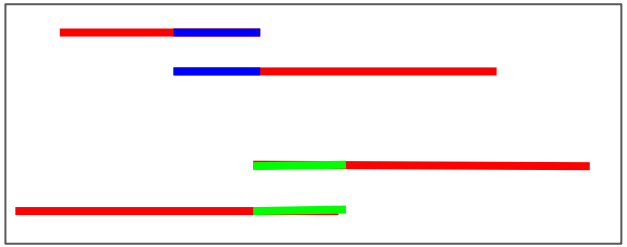

v_1
7
v_2
2
utg1
v_4
11
v_5
4
v_6
5
v_7
9
v_8
6
v_9
2
utg3
v_{10}
4
v_{11}
5
v_{14}
utg6
v_3
5
utg5
utg2
6
v_{12}
4
v_{13}
4
utg4


utg1
2
utg3
5
2
utg2
utg6
6
5
4
utg5
utg4

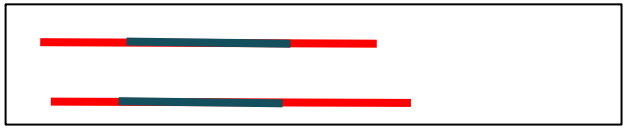
Zarovnanie (alignment) [Minimap]

čítania k čítaniam

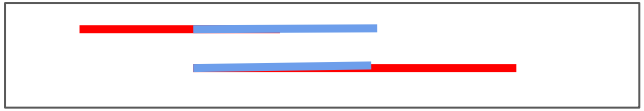
Miniasm používa v assembly grafe



Miniasm používa na počítanie pokrytia



Miniasm používa v assembly grafe

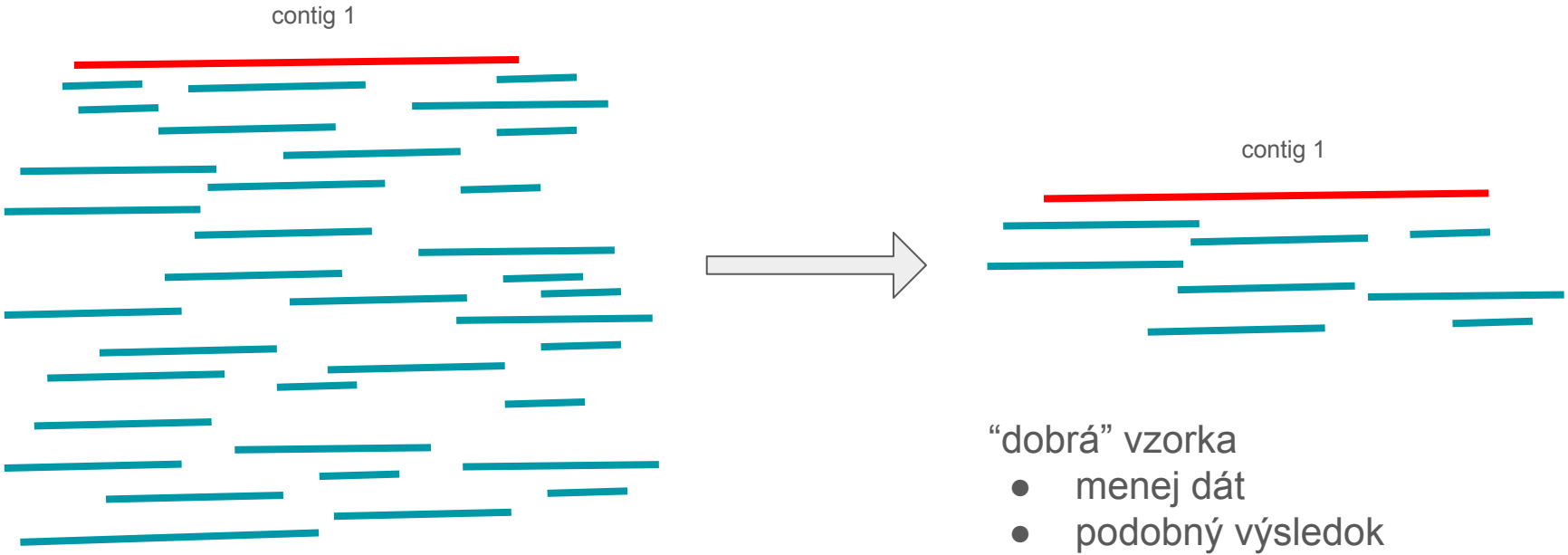


čítania k assembly

čítania, ktoré nám dávajú novú informáciu

niektoré z týchto treba zachovať

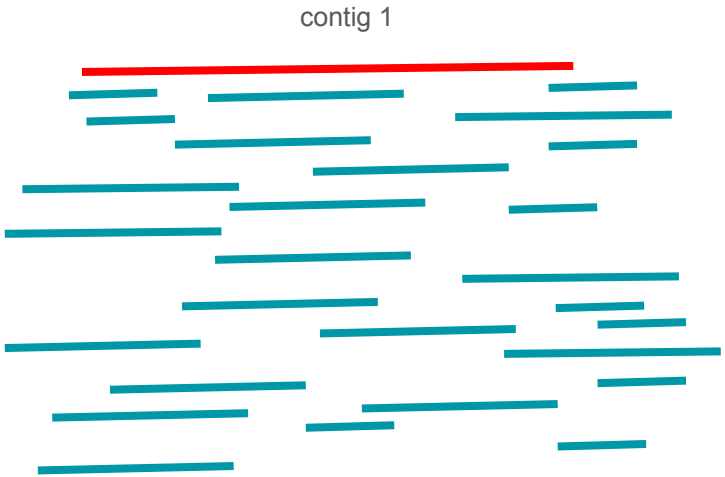
reprezentatívna vzorka



“dobrá” vzorka

- menej dát
- podobný výsledok

rôzne pokrytie kontigov

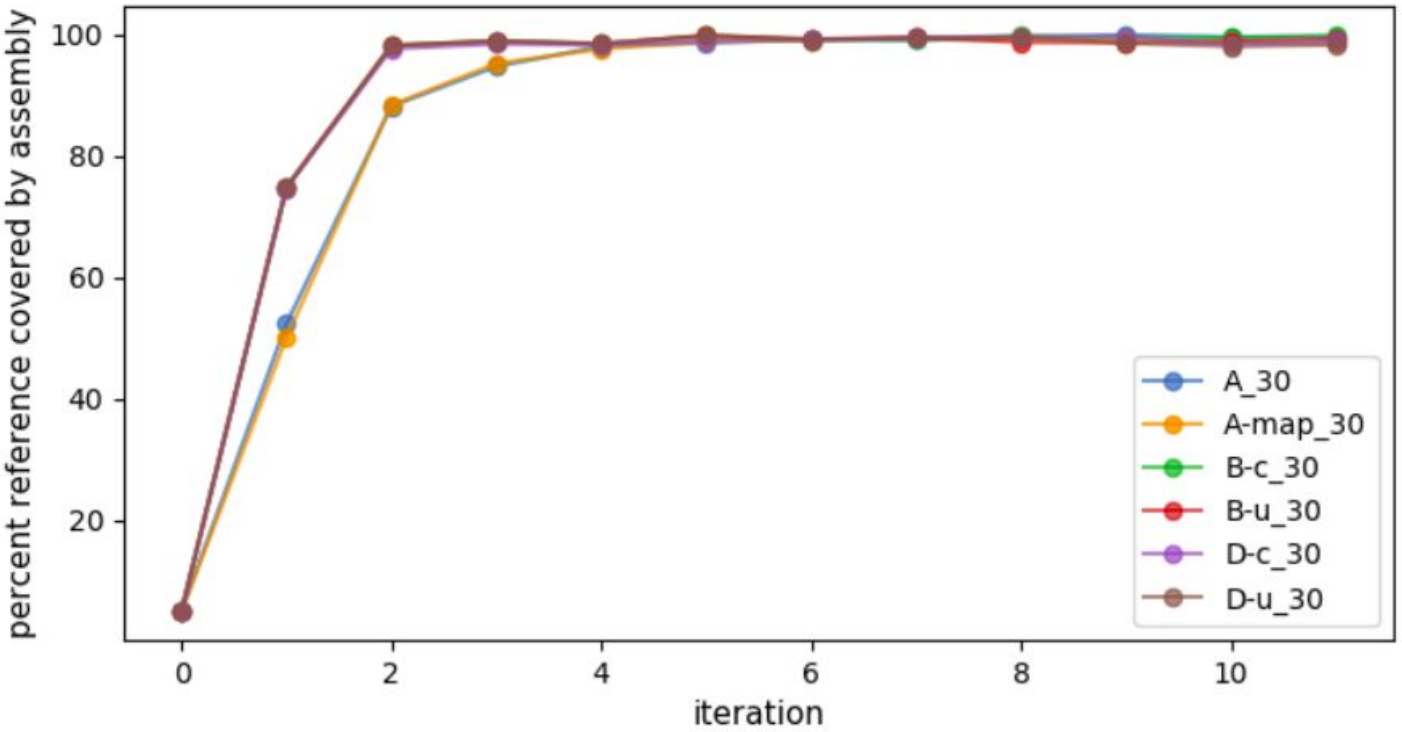


Rôzne stratégie výberu reprezentatívnej vzorky

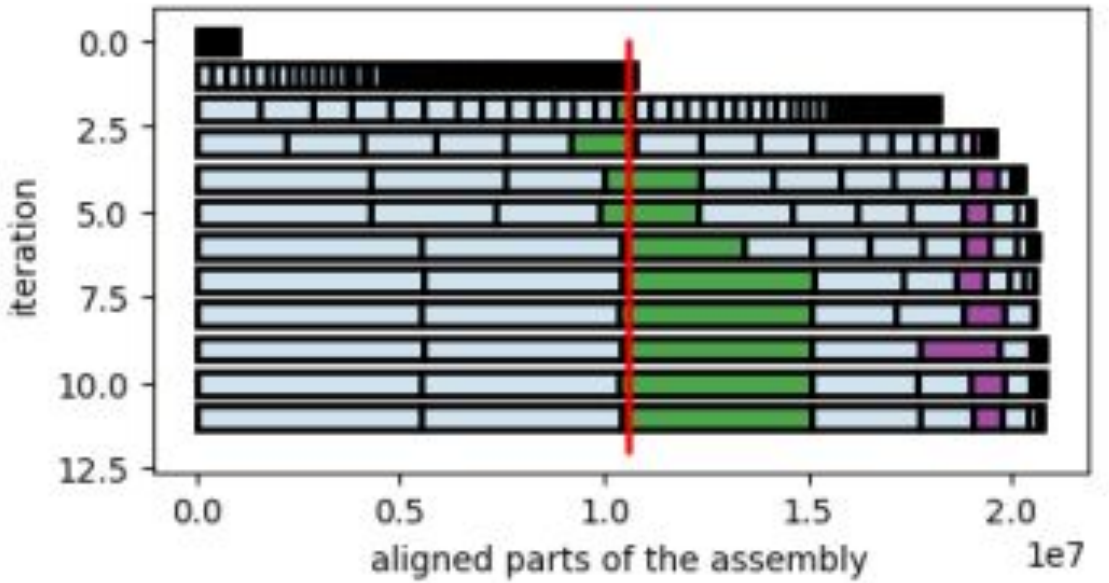
- preferovať dlhšie čítania (dlhšie čítania -> vyššia pravdepodobnosť výberu) [A]
- po dosiahnutí určitého pokrytia pridávať iba čítania na koncoch už poskladaných kontigov (nová informácia) [A-map]
- nechceme príliš preferovať čítania zo začiatku behu pred tými, ktoré vzniknú neskôr [B]
- vyberať najdlhšie čítania z celého behu (deterministická) [D]

Porovnanie zostaveného genómu s referenčným genómom

Výsledky: pokrytie referenčného genómu zostaveným genómom

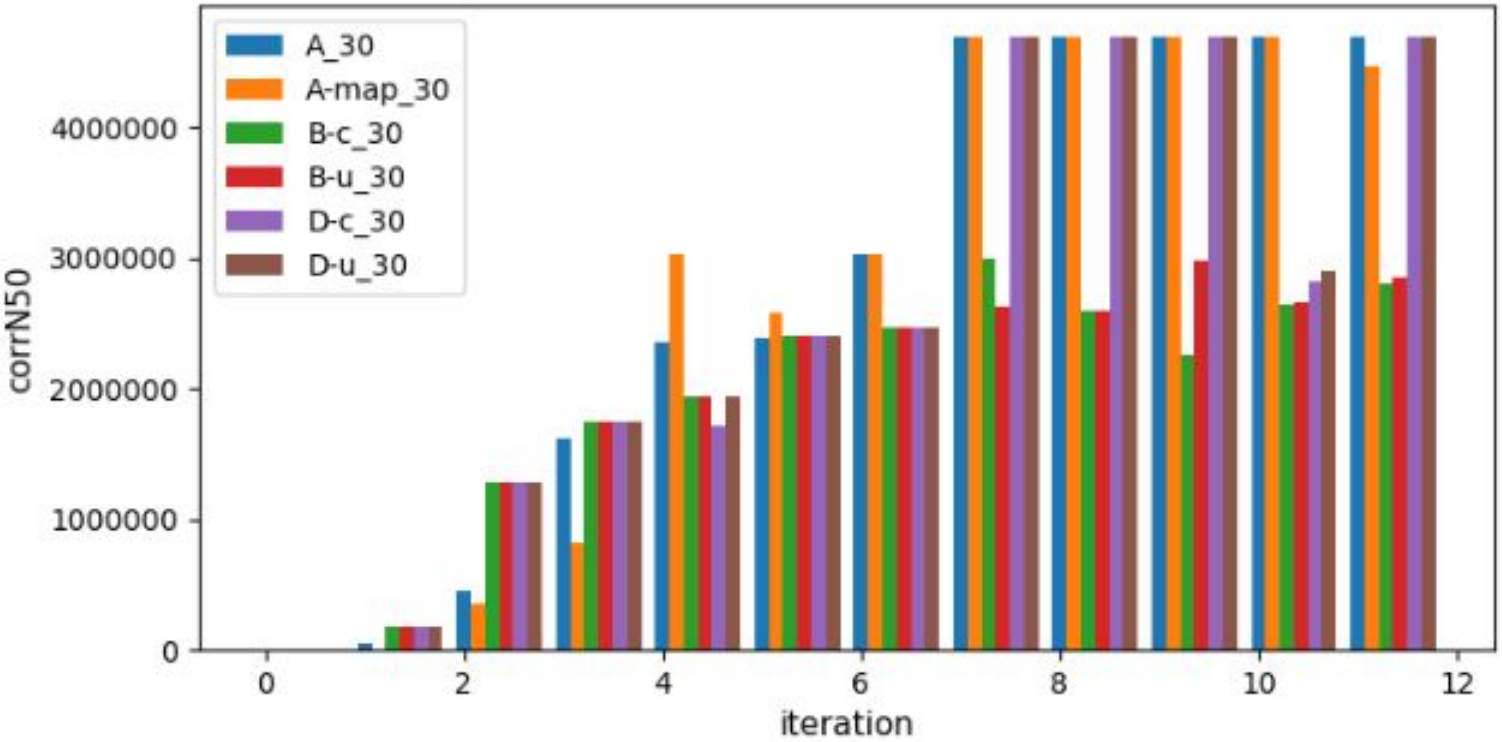


Výsledky: corrN50 score



(a) A

Výsledky: porovnanie stratégií



Čas vs. parametre behu

#čítaní / iterácia

5000

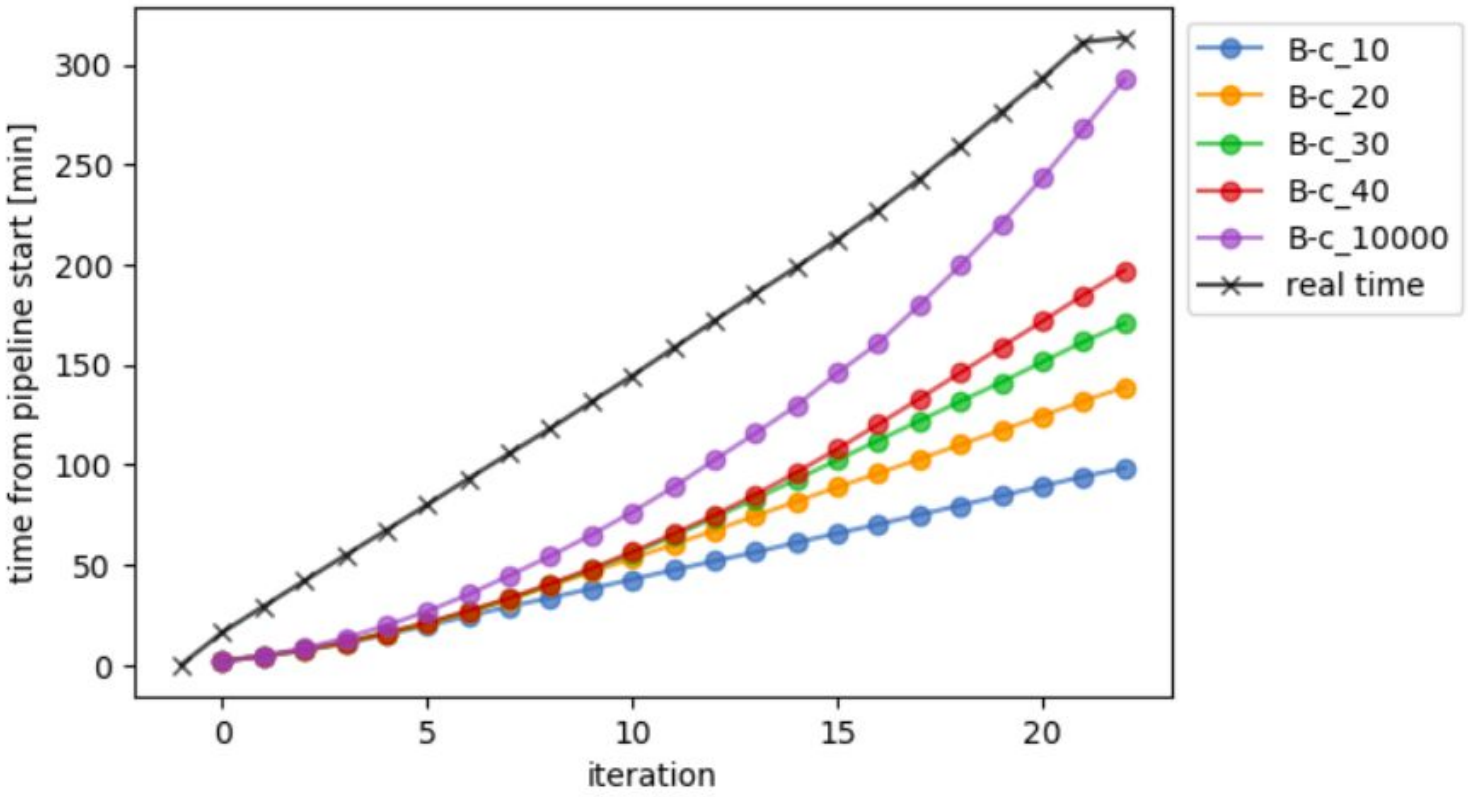
(10 súborov, v každom 500 čítaní)

- čím viac čítaní/iterácia, tým kratší čas

cieľové pokrytie

- čím nižšie pokrytie, tým kratší čas
- po dosiahnutí pokrytia konštantný pre 1 iteráciu

Výsledky: Čas



Zhrnutie

- cieľ: skladanie genómu v reálnom čase počas sekvenovania

- problém sme riešili opakovaným skladaním genómu z priebežne upravovanej reprezentatívnej vzorky dát,
- navrhli a implementovali sme niekoľko rôznych stratégií výberu vzorky

- porovnali sme výsledky (s využitím referenčného genómu)
- dáta stíhame analyzovať v reálnom čase

Ďakujem za pozornosť

1. Jednou z úloh navrhnutého softvéru je určenie, či máme dosť čítaní pre *de novo* zostavovanie genómu (str. 4). Vedeli by ste na príklade dát z kapitoly 4 ukázať, kedy je tých čítaní dosť?

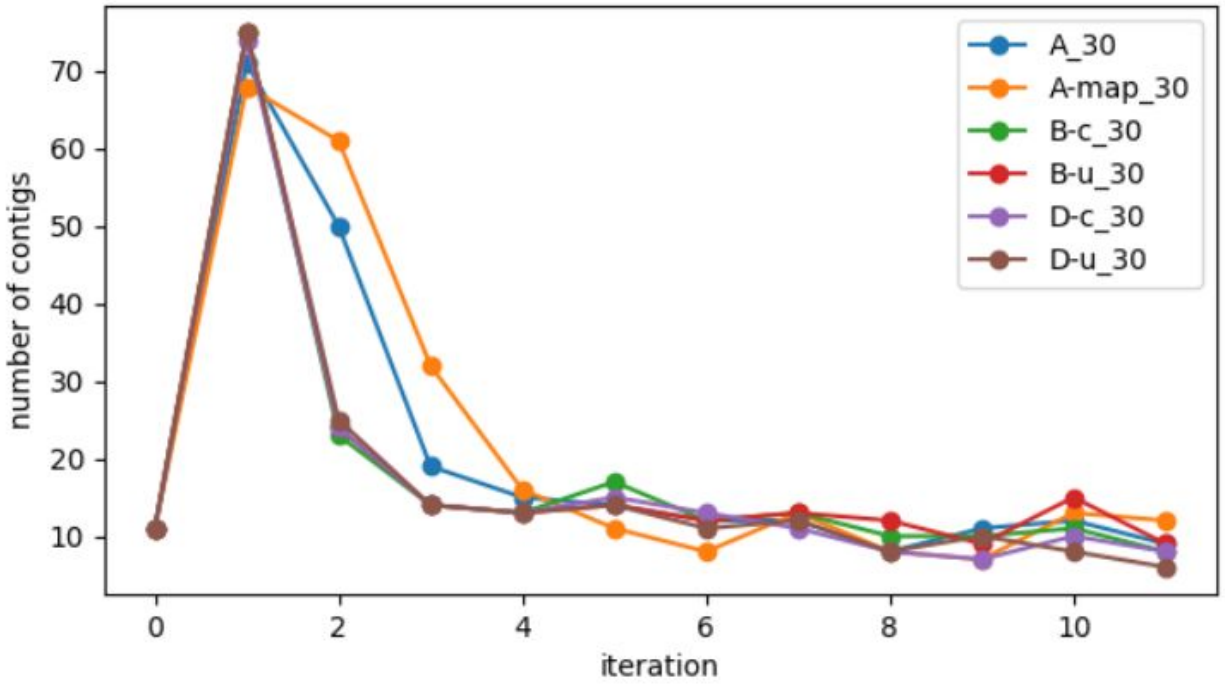


Figure 4.16: Number of contigs per iteration for 30x threshold runs with batch size 20.

1. Jednou z úloh navrhnutého softvéru je určenie, či máme dosť čítaní pre *de novo* zostavovanie genómu (str. 4). Vedeli by ste na príklade dát z kapitoly 4 ukázať, kedy je tých čítaní dosť?

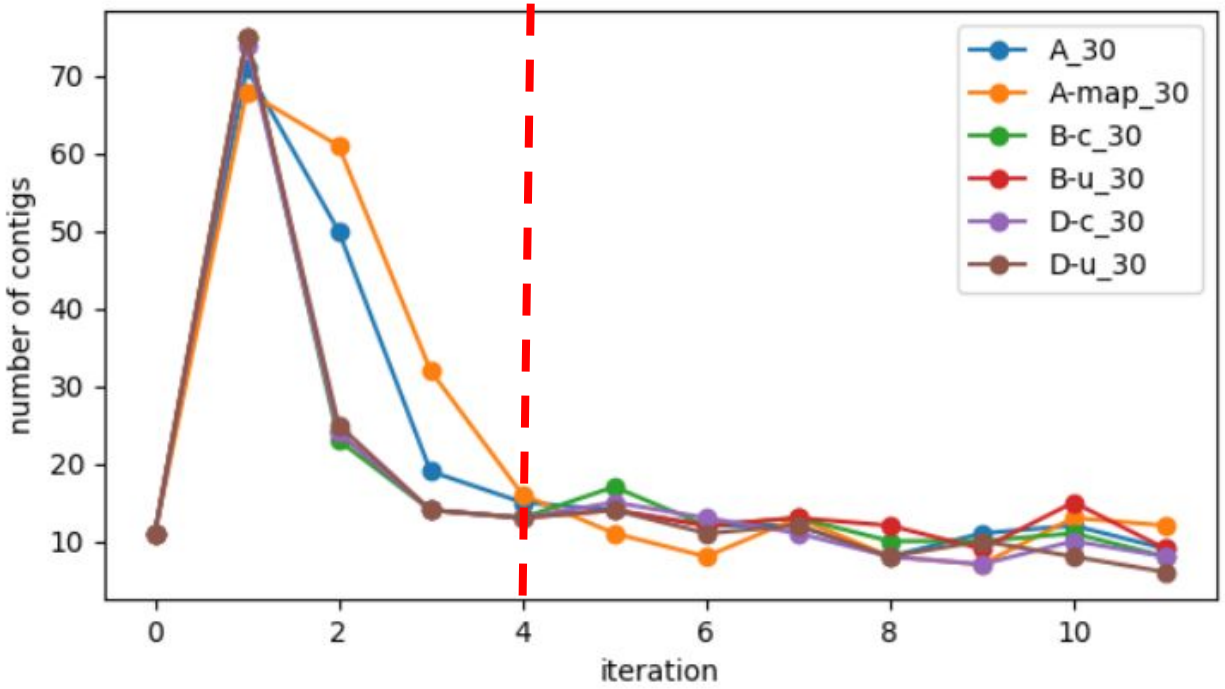


Figure 4.16: Number of contigs per iteration for 30x threshold runs with batch size 20.

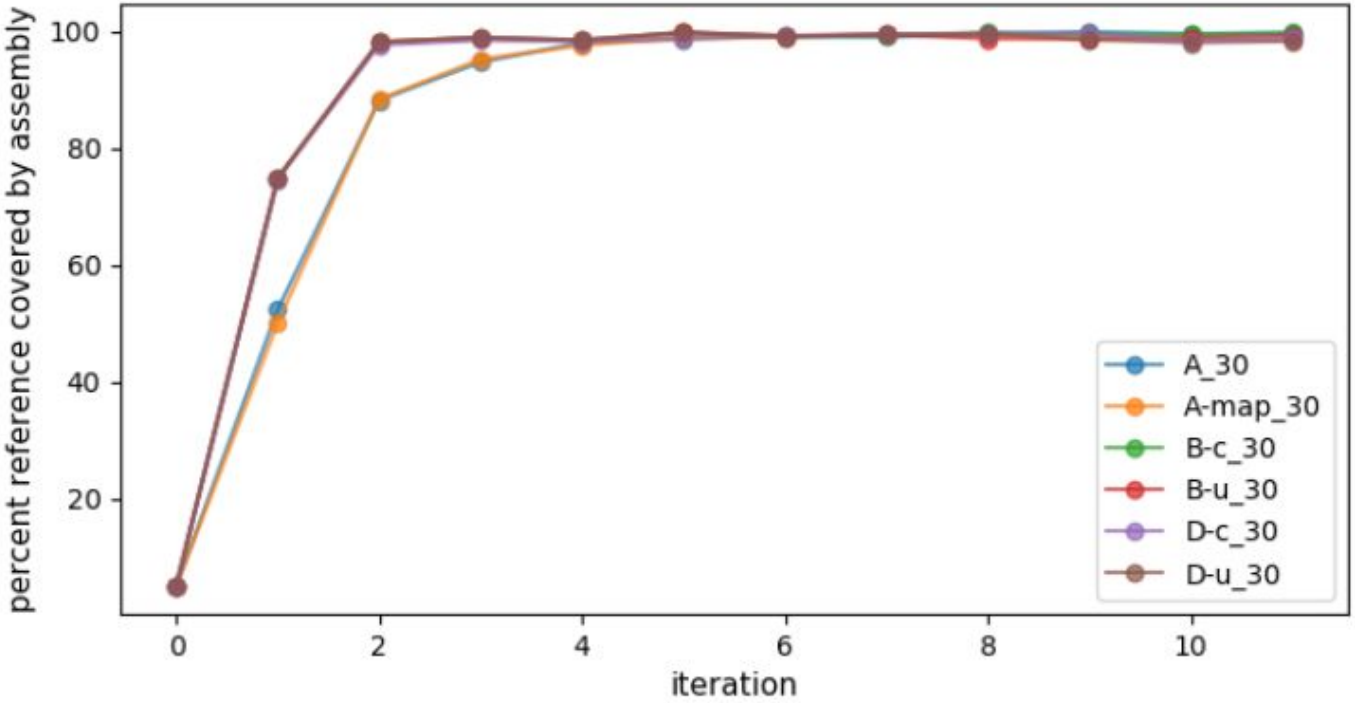


Figure 4.18: Percentages of reference genome length covered by assembly for 30x coverage threshold and batch size 20 (per iteration). We can see that the trend is similar to the trend in runs with batch size 10.

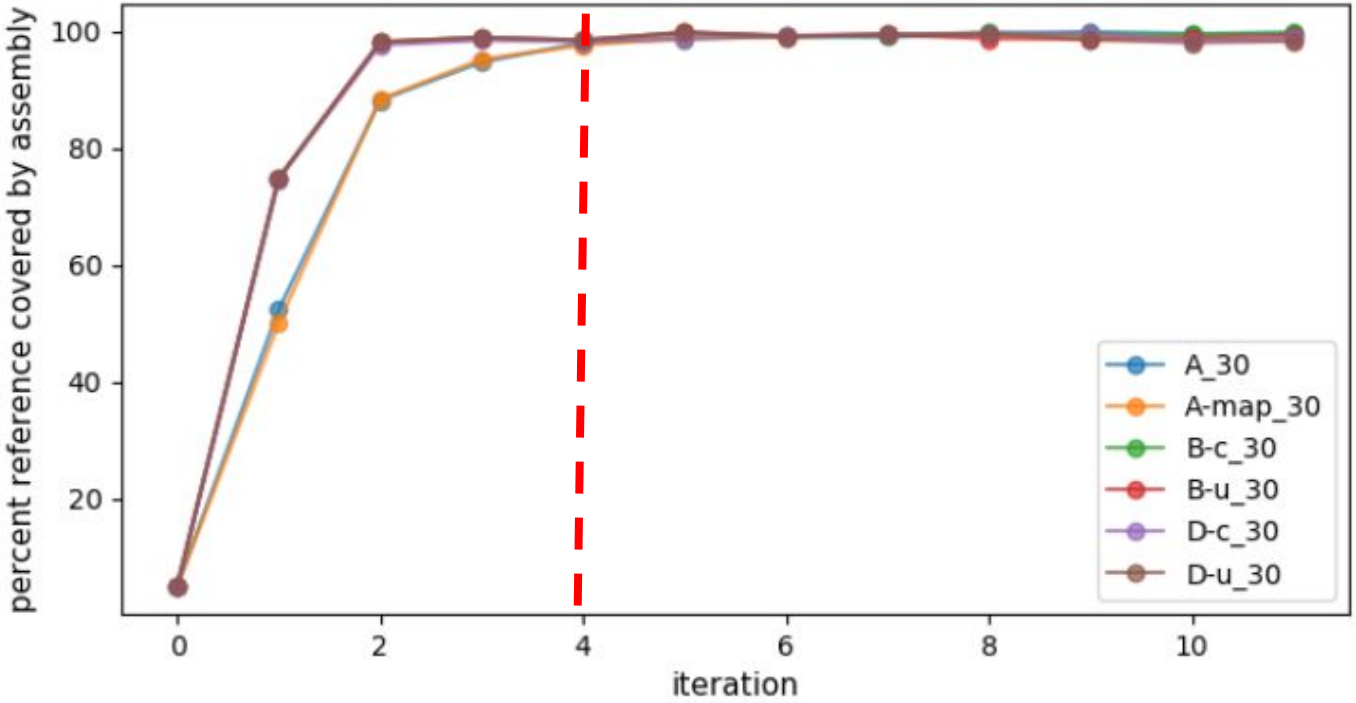


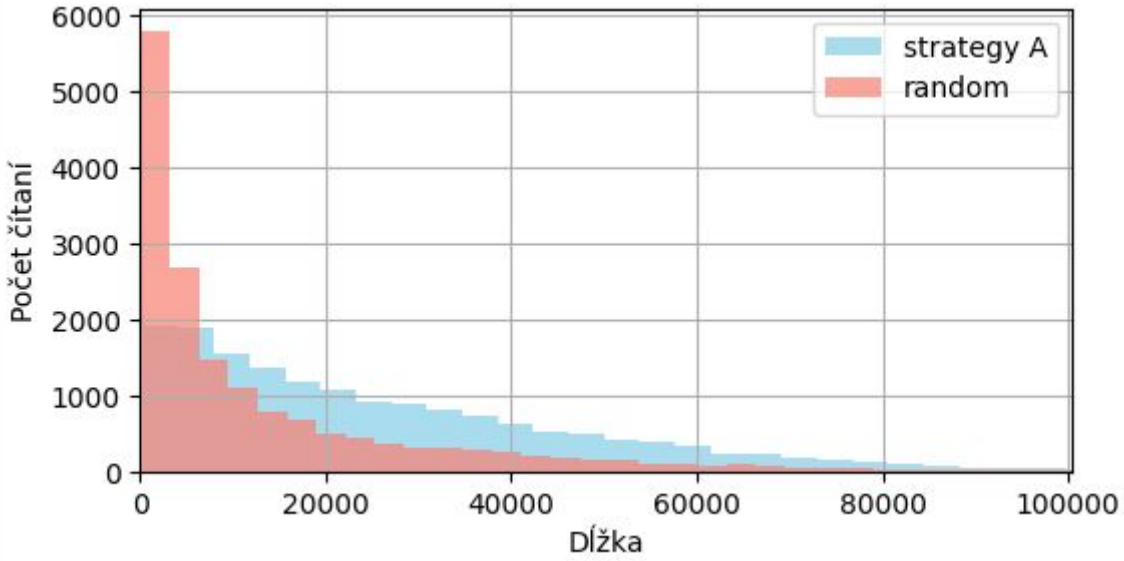
Figure 4.18: Percentages of reference genome length covered by assembly for 30x coverage threshold and batch size 20 (per iteration). We can see that the trend is similar to the trend in runs with batch size 10.

2. V základnej vzorkovacej stratégii (str. 24) pravdepodobnosť výberu čítania je priamo úmerná jeho dĺžke (respektíve dĺžke jeho zarovnania). Účelom bolo prioritizovať dlhšie čítania. Vedeli by ste znázorniť a kvantifikovať dopad tejto stratégie na dĺžky vzorkovaných čítaní (oproti rovnomerne náhodnému výberu)?

priemerná dĺžka čítaní na konci behu (16442 čítaní, pokrytie 20): 26596.10

priemerná dĺžka pre rovnomerne náhodný výber (16442 čítaní): 13551.61

priemerná dĺžka (všetky čítania z behu, 111042 čítaní): 13586.53



3. Vedeli by ste porovnať súčasnú vzorkovaciu stratégiu s jej modifikáciou, kde sa čítania vzorkujú priamo úmerne druhej mocnine ich dĺžok?

"súčasná vzorkovacia stratégia":

$$\sum_{j \in S_c} r_j \approx t \cdot \ell_c$$

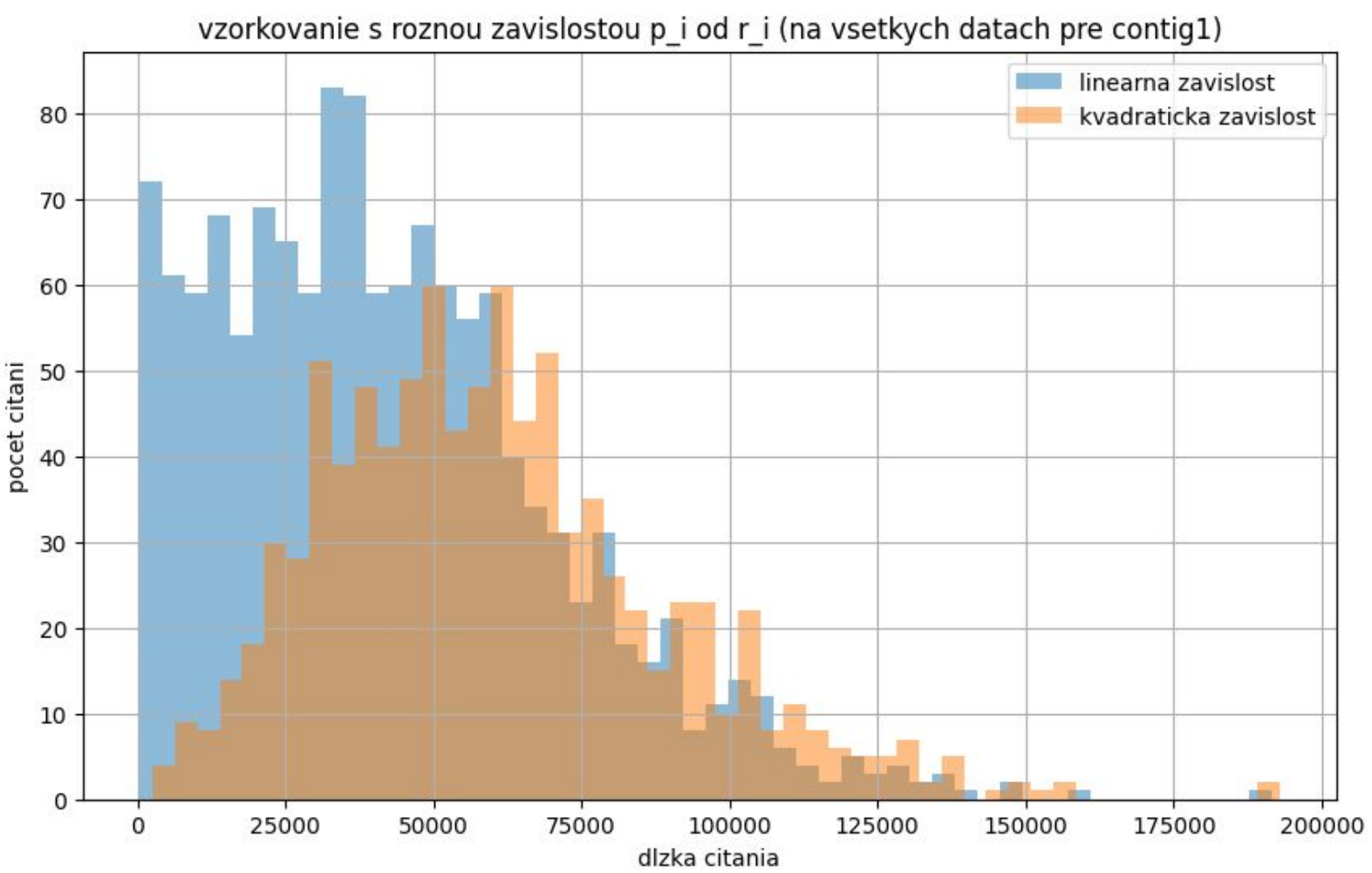
$$p_i = \min\left(\frac{\ell_c \cdot t \cdot r_i}{\sum_{j \in M_c} r_j^2}, 1\right)$$

M_c množina čítaní zarovnaných ku kontigu
S_c vzorka (množina vybraných) čítaní pre kontig
ℓ_c dĺžka kontigu
r_i súčet dĺžok zarovnaní pre i-te čítanie
p_i pravdepodobnosť výberu čítania i
t hranica pokrytia

$$\sum_{i \in M_c} p_i r_i \leq \sum_{i \in M_c} \frac{\ell_c \cdot t \cdot r_i}{\sum_{j \in M_c} r_j^2} r_i = \frac{\ell_c \cdot t}{\sum_{j \in M_c} r_j^2} \sum_{i \in M_c} r_i^2 = \ell_c \cdot t$$

3. Vedeli by ste porovnať súčasnú vzorkovaciu stratégiu s jej modifikáciou, kde sa čítania vzorkujú priamo úmerne druhej mocnine ich dĺžok?

$p_i = \min(\frac{\ell_c \cdot t \cdot r_i}{\sum_{j \in M_c} r_j^2}, 1)$	$p_i = \min(\frac{\ell_c \cdot t \cdot r_i^2}{\sum_{j \in M_c} r_j^3}, 1)$
očakávaný súčet dĺžok zarovnaní $\leq \ell_c \cdot t$	očakávaný súčet dĺžok zarovnaní $\leq \ell_c \cdot t$
menej hodnôt $p_j = 1$	viac hodnôt $p_j = 1$
lineárna závislosť p_i od dĺžky čítania	kvadratická závislosť p_i od dĺžky čítania
menej výrazná preferencia dlhších čítaní	výraznejšia preferencia dlhších čítaní



súčet dĺžok

lin. z.:
56374085

kv.z:
55237099

$\sum_{i \in M_c} p_i r_i$

lin.z:
57145100

kv.z:
57145099
(9 hodnôt
p_i=1)

l_c*t
57145100

4. Vedeli by ste preukázať či Vaša stratégia na vymazanie starých čítaní (str. 26) zachováva pôvodné pravdepodobnostné rozdelenie, t.j. že po spracovaní novej "várky" dát jednotlivé čítania majú rovnakú pravdepodobnosť byť vybrané do vzorky? Ak nie, vedeli by ste navrhnúť stratégiu, ktorá je toho schopná?

"stratégia na vymazanie starých čítaní":

- každému čítaniu je pridelená hodnota *p_i* z intervalu < 0,1 > ktorá sa počas behu nemení
- *pre každý kontig* v každej iterácii hľadáme *hranicu q_c*, pre ktorú súčet dĺžok zarovnaní pre čítania s p_i >= q_c je dostatočný (pokrytie * dĺžka kontigu)
- čítania s p_i < q_c ďalej nepoužívame

4. Vedeli by ste preukázať či Vaša stratégia na vymazanie starých čítaní (str. 26) zachováva pôvodné pravdepodobnostné rozdelenie, t.j. že po spracovaní novej "várky" dát jednotlivé čítania majú rovnakú pravdepodobnosť byť vybrané do vzorky? Ak nie, vedeli by ste navrhnúť stratégiu, ktorá je toho schopná?

- dá sa ukázať za predpokladov:
  - žiadny už zostavený kontig sa nerozpadne (predĺžiť sa môže),
  - už poskladaná časť kontigu sa nezmení (môžu len pribudnúť zarovnané čítania)
- -> hranice q_c pre jednotlivé kontigy sa potom nebudú medzi iteráciami zmenšovať
- "vymazané" čítania majú p_i < q_c, do vzorky by sa už nedostali
- každému čítaniu je p_i pridelené náhodne a nemení sa medzi iteráciami

- prakticky predpoklady nie sú vždy splnené

- Myslíte si, že namiesto vzorkovania by bolo možné vytvoriť nový algoritmus skladania genómov, ktorý by dokázal prijímať sekvenačné čítania postupne a ako medzivýsledok by udržiaval aktuálny zoskladaný genóm, pričom jeho kvalita by sa zlepšovala s ďalšími prichádzajúcimi dátami? Vidíte nejaké možné úskalia pri tvorbe takéhoto algoritmu?
  - postupne pridávať čítania do grafu, udržiavať graf
  - úskalia:
    - graf sa aj po pridaní malého množstva čítaní môže zásadne zmeniť
    - problémy s bežne používanými heuristikami:
      - Miniasm:
        - filtruje čítania na základe pokrytia – necháva si len krátke úseky z nich -> problém ako spracovať “zahodené” časti
        - stále by bolo nutné zarovnávať nové čítania k starým
        - čas zarovnávania >> čas zostavenia grafu
      - Flye: vyrába graf z čítaní, v ktorých *najprv* opravuje chyby – pomalé pre analýzu v reálnom čase
    - nová heuristika?
  - scaffoldery

contig name	length [bp]
contig1	5714510
contig2	4992828
contig3	4821795
contig4	2900717
contig5	2723818
mtDNA	35540

Table 4.1: Lengths of *Saprochaete ingens* reference genome contigs.

čítania vybraté stratégiou B počas behu vs. náhodný výber rovnakého počtu čítaní