**Genomika: analýza a algoritmy - Seminární práce**

## **Assembly a anotace bakteriálního genomu**

**Postup práce a výsledky:**

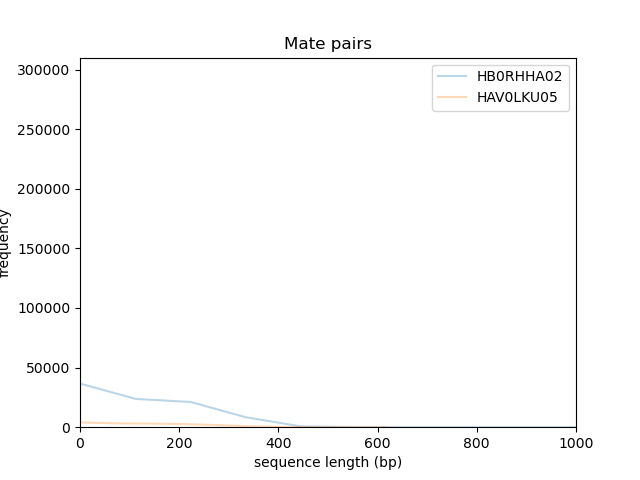
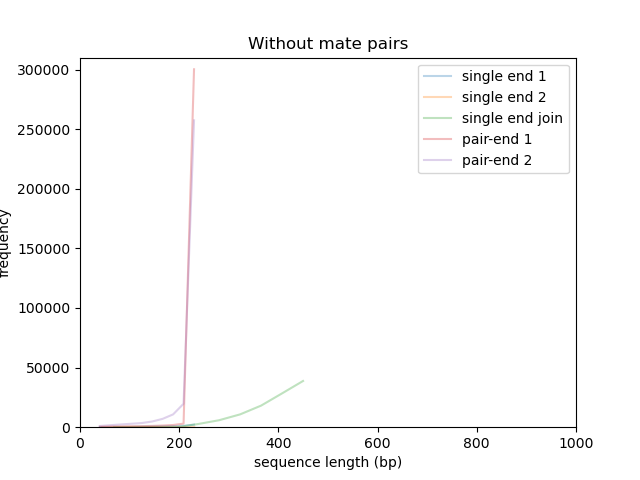
1. Zkontrolovala jsem kvalitu sekvencí pomocí programu fastqc. Žádné sekvence jsem neodstraňovala, zdáli se mi knihovny kvalitní a adaptory ve výstupu u fastqc nebyly nalezeny.

*(Proč? Ověříte si, jak málou část dat tvoří mate-pair knihovny, ale přitom jak výrazně zlepší assembly.)*

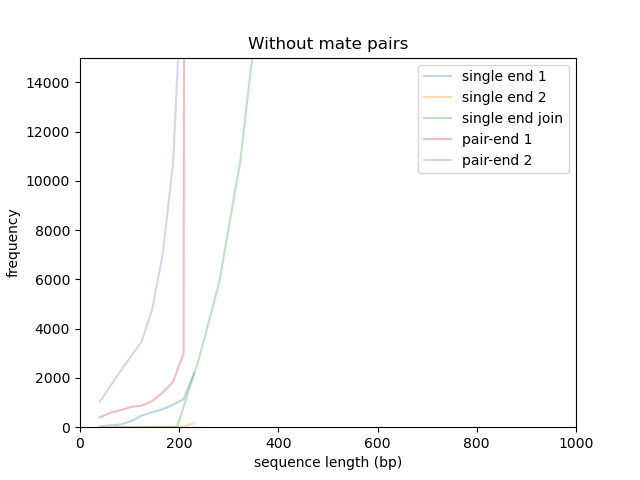
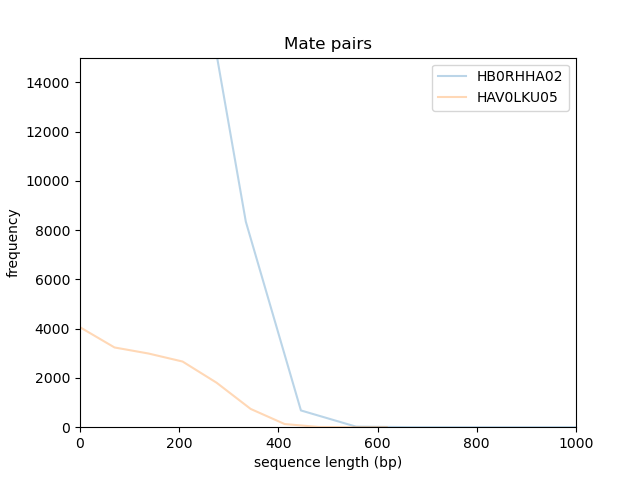
Počty jednotlivých sekvencí v knihovnách + kvalita knihoven:

|  |  |  |
| --- | --- | --- |
|  | Počet sekvencí | kvalita knihoven |
| single end | 6 513 | kvalitní sekvence (průměr 37) |
| 260 | ještě celkem slušná kvalita, vyšší variabilita (průměr 36) |
| single end (join) | 104 543 | velmi kvalitní sekvence (průměr 37) |
| paired-end | 310 894 | kvalitní sekvence (průměr 37) |
| 310 894 | celkem kvalitní sekvence (průměr 36) |
| mate pairs | 15 625 | nejnižší kvalita (průměr 27) vysoká variabilita |
| 90 426 | nejnižší kvalita (průměr 38) - velké množství sekvencí s kvalitou mezi 23 - 34 |

Porovnání knihoven:



Přiblížení :

1. Provedla jsem assembly pomocí OLC algoritmu (newbler) a deBruin grafů (Soap2).

Výsledky z programu quast:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | N50 | L50 | totální délka | GC% |
| Bcc7419\_assembly\_without\_matepairs: | 98 645 | 25 | 7 785 205 | 67.16 |
| Bcc7419\_assembly\_with\_matepairs: | 107 247 | 23 | 7 784 980 | 67.16 |
| bcc7419\_assembly\_soap | 1 710 | 1 211 | 6 447 556 | 67.16 |

*(Proč? Porozumíte metrikám, pomocí kterých hodnotíme kvalitu assembly.)*

**Vybrání nejlepší assembly dle statistiky:**

Jako nejlepší jsem vybrala assembly, která obsahovala knihovnu mate-pairs.

**Vysvětlení(jak se to hodnotí):**

Při srovnání výsledků assembly, tak SOAP-denovo dopadlo nejhůře. Celková délka contigů je nejmenší (total length), prostřední contig je také nejmenší (N50) a je až 1211. (L50 - pořadí prostředního contigu).

Assembly bez knihovny mate-pairs vyšlo o trošku hůře, nejlépe dopadlo assembly s knihovnou mate-pairs.

*(Proč? Ujasníte si rozdíly ve výsledcích obou typů algoritmů.)*

**vysvětlení:**

Psali v (https://www.animalgenome.org/bioinfo/resources/manuals/SOAP.html), že soap denovo je speciálně vytvořen pro krátké ready assembly Illumina GA. Soap denovo se zaměřuje na velké rostlinné a zvířecí genomy, ačkoli by mělo také dobře fungovat na bakteriálních a fungi genomech. Netuším, proč vyšlo assembly tak špatně.

*(Proč? Ověříte si, jak málou část dat tvoří mate-pair knihovny, ale přitom jak výrazně zlepší assembly.)*

**vysvětlení:**

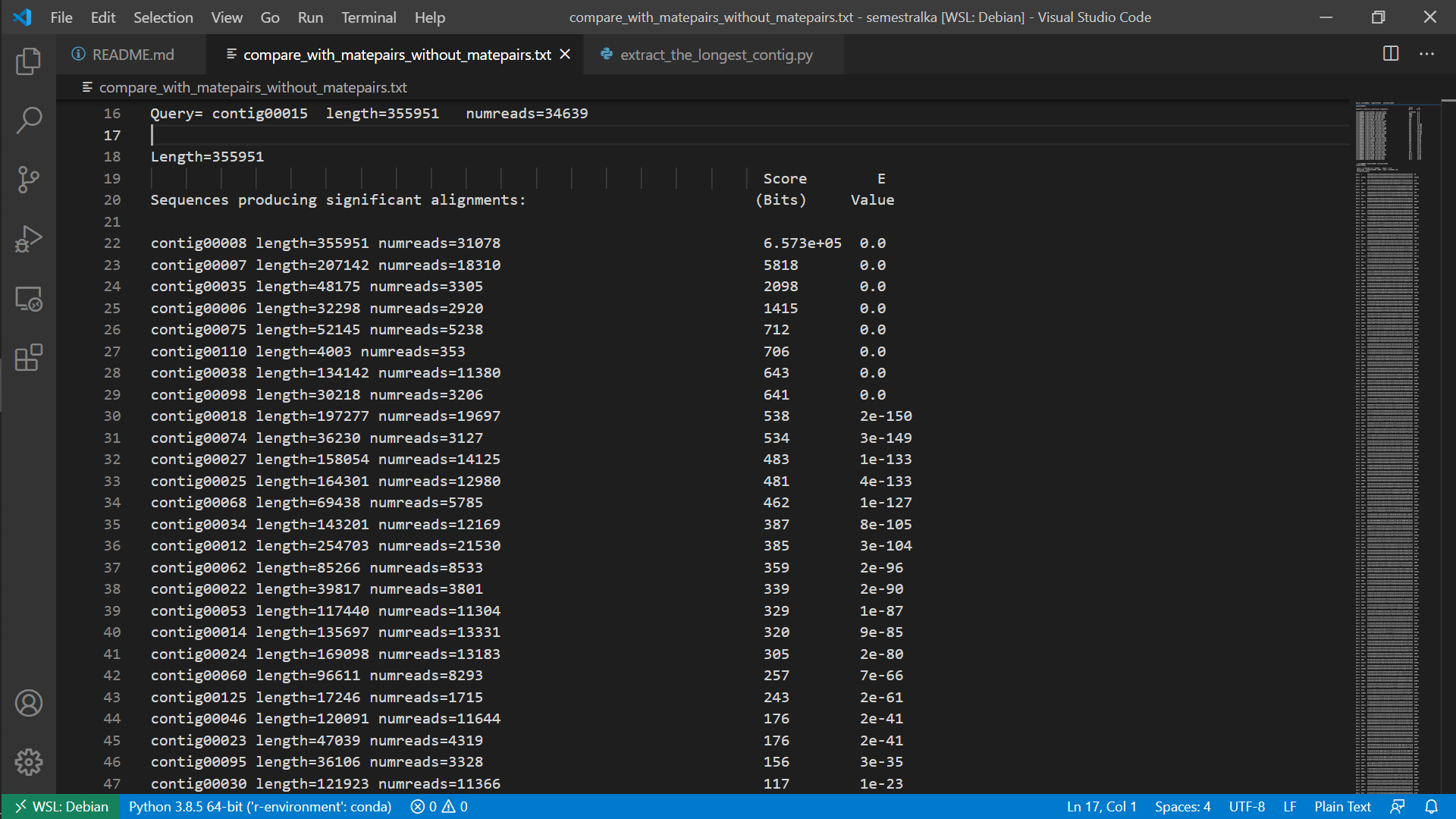
V článku (https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-95) psali, že použitím mate-pairs v prokaryotickém assembly lépe definovali repetické oblasti než knihovny, které byly běžně konstruovány pro aktuální genom. To by tedy mohlo být důvodem lepší assembly při použití mate-pairs.

Vybrala jsem nejdelší scaffold z nejlepšího assembly (with mate-pairs) a pomocí BLAST jsem na něj namapovala contigy z ostatních dvou assembly.

*(Proč? Vizualizace rozdílů mezi jednotlivými přístupy a porozumění významu jednotlivých datasetů.)*

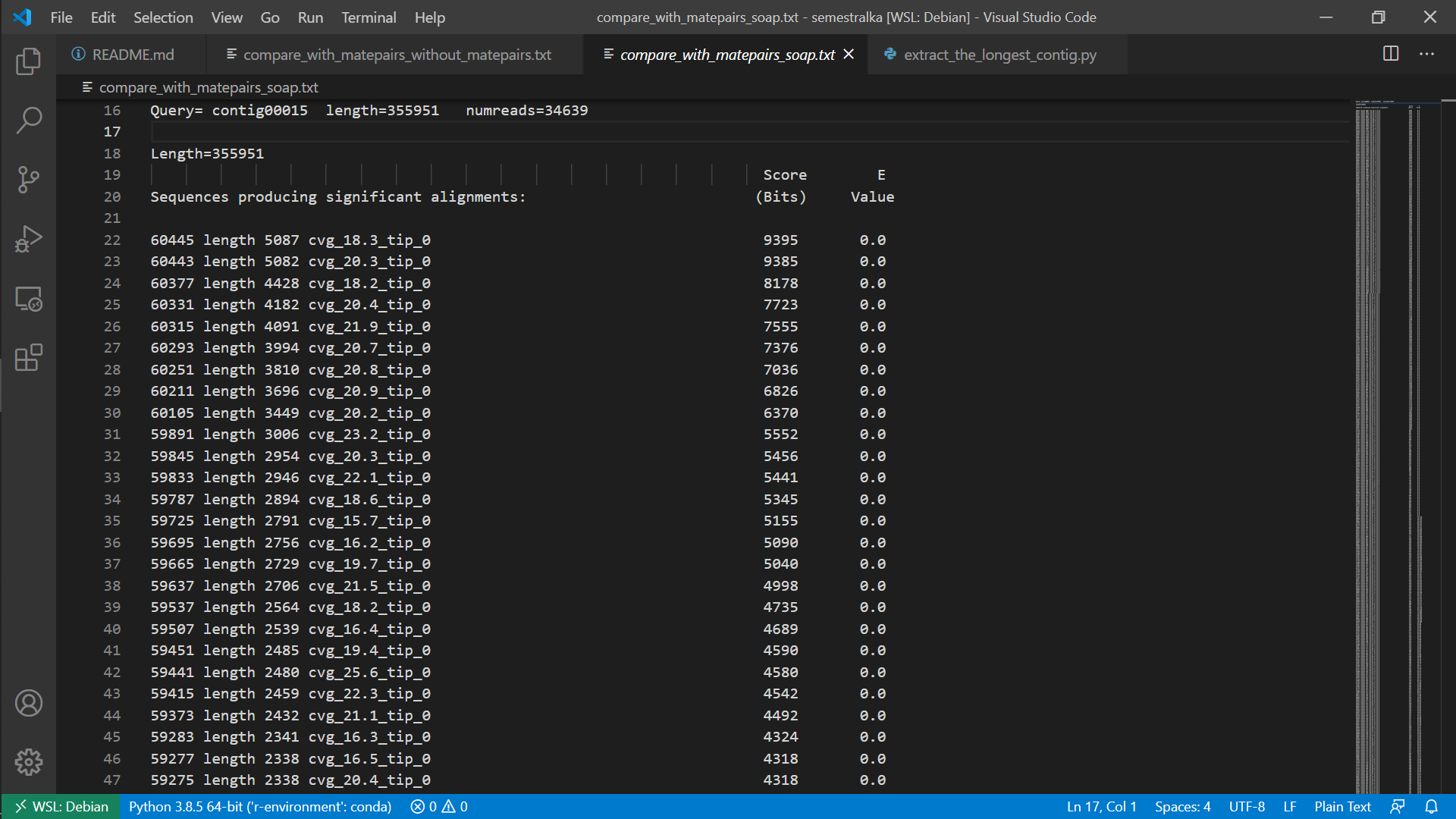
Zarovnani nejdelsiho contigu s assembly bez matepairs

Dohromady zarovnano s 32 contigy



Srovnani nejdelsiho contigu s assembly programu SOAP-denovo

dohromady zarovnano s 500 contigy



**Srovnání:**

Newbler je oproti SOAP denovo souvislejší a při namapování na assembly newbler s matepairs má méně namapování oproti SOAP denovo.

1. Vybrala jsem nejdelší scaffold a na něm jsem predikovala geny pomocí programu glimmer.
2. Pomocí programu BLAST jsem nalezla nejpravděpodobnější názvy jednotlivých genů.

**počet genů**

program Glimmer mi predikoval celkem 345 genů

**anotace a její úspěšnost**

úspěšnost: 88.97% úspěšnost, 11.03 % (39 256 z celkové délky 355 951) bází mi to vyřadilo a nezařadilo ke genům

**co jste se o organizmu z analýzy dozvěděli:** (organismus Burkholderia)

* dělení buňky a sporulace
* buněčná adheze nebo předání genetické informace (vytváření pilusu)
* biosyntéza nukleotidů
* záchrana AMP, adeninu (katalyzuje záchrannou reakci −> tvorba AMP, energeticky méně nákladná než de novo)
* DNA templátová transkripce, iniciace
* DNA - oprava DNA (bázová excize, DNA rekombinace, nukleotidová excize) + regulace opravy DNA
* regulace transkripce (spousta regulátorů transkripce)
* katalýza syntézy RNA (DNA jako templát)
* ligázy pro tRNA
* zpracování rRNA
* regulace translace
* štěpení v proteinu (serine endoproteáza)
* enzymy citrátového cyklu
* syntéza acyl-CoA
* enzymy Beta-oxidace mastných kyselin
* různé metabolické procesy aminokyselin (transport, biosyntéza, katabolické procesy, …)
* modifikace prolinu (cist trans)
* biosyntéza mastných kyselin, lipoproteinů, lipopolysacharidů, biosyntéza fosfolipidů
* spousta procesů sacharidů, glykolýza
* syntéza pyruvátu
* fixace uhlíku (beta podjednotka - CO z CO2, alfa podjednotka kombinuje CO s CoA a methylovou skupinou za vzniku acetyl-CoA)
* tvorba hemu do cytochromu
* Fe-S protein - přenos elektronů, uskladnění železa(ferredoxin, ferritin)
* dýchací řetězec - obsahuje cytochrom c oxidasy
* různé integrální membránové proteiny, transport přes membránu (aquaporiny H2O, sacharidy, sodíkové ionty, pumpy, měďnatý ion, …)
* ligásové, hydrolázové, transferázové, oxidoreduktázové aktivity, negativní regulace katalytické aktivity