Computational analysis of the genome of fungus Jaminaea angkorensis

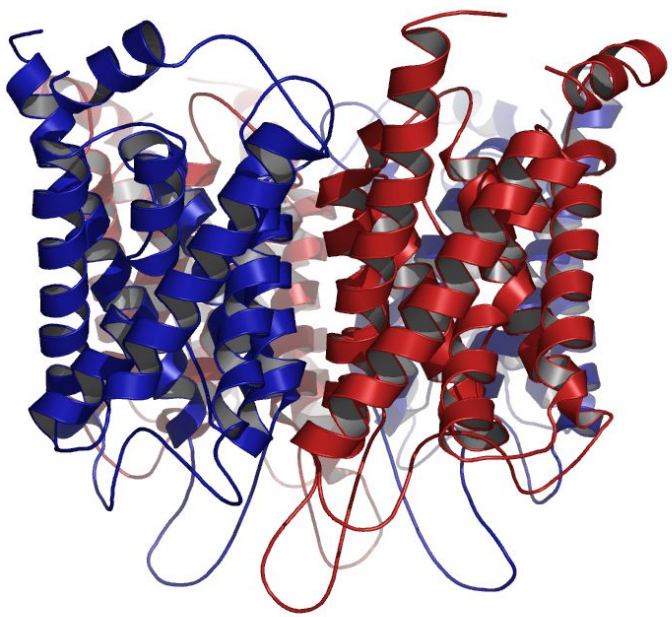
Autor: Dominik Bujna
Školiteľ: doc. Mgr. Bronislava Brejová, PhD

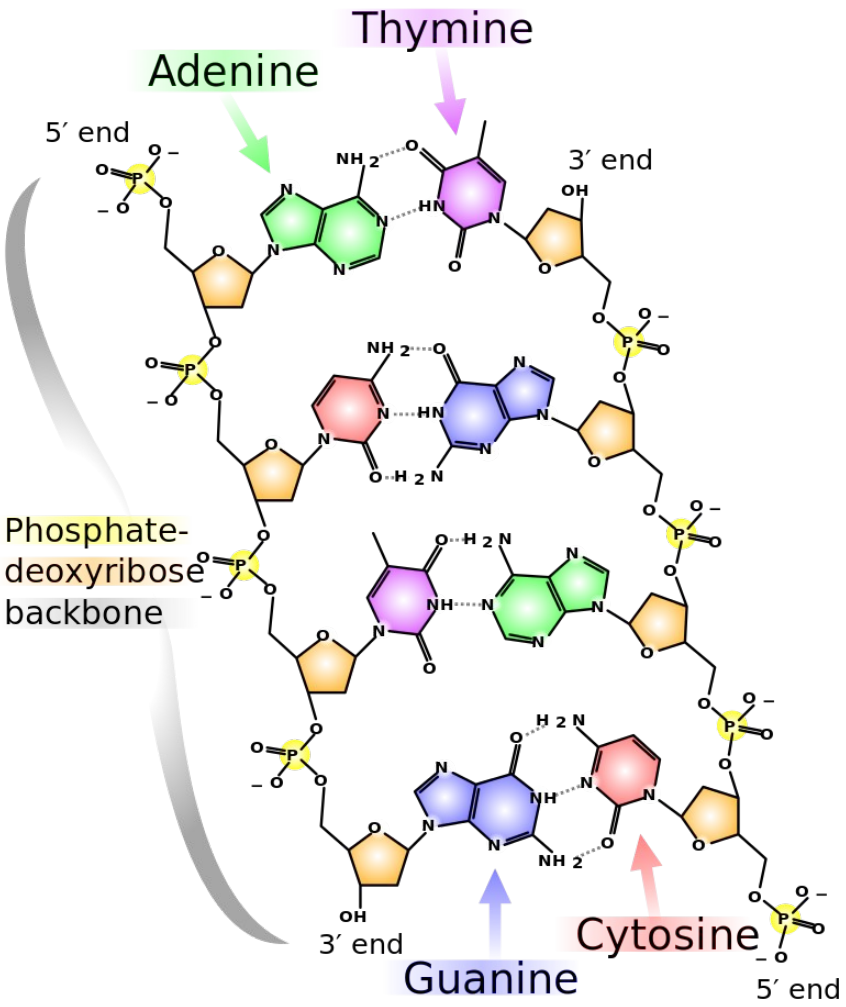
Výpočtová analýza genómu huby Jaminaea angkorensis

Autor: Dominik Bujna
Školiteľ: doc. Mgr. Bronislava Brejová, PhD

Základné pojmy

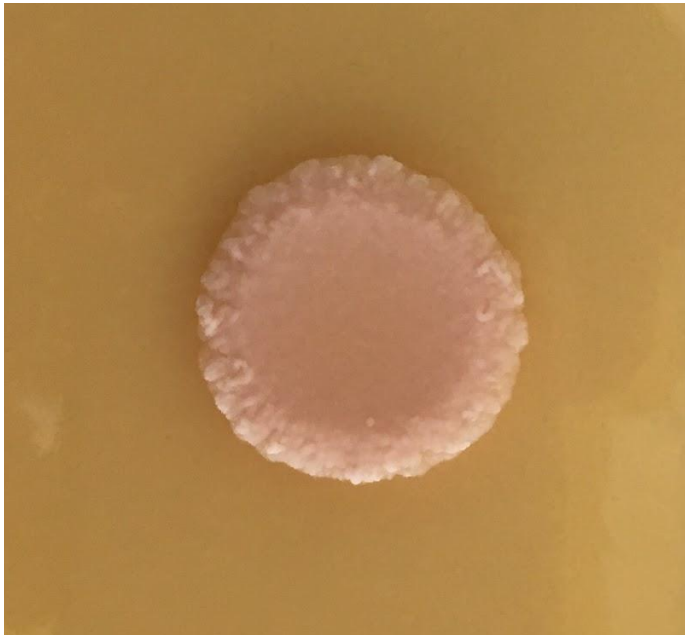
- Genóm
  - DNA
  - RNA
  - Proteíny





Jaminaea angkorensis

- Objavená zo vzoriek z roku 2006, publikovaný nález 2009
- Nájdená v rozkladajúcom sa lístí v Kambodži
- Základom pre vytvorenie nového rodu
  - Medzi časom 4 nové druhy (2 v databáze NCBI *Jaminaea lanaiensis, Jaminaea rosea* a 2 v literatúre - *Jaminaea pallidilutea, Jaminaea phylloscopi*)
- Predpokladané fungicídne vlastnosti nepotvrdené
- Doposiaľ publikovaný iba mitochondriálny genóm

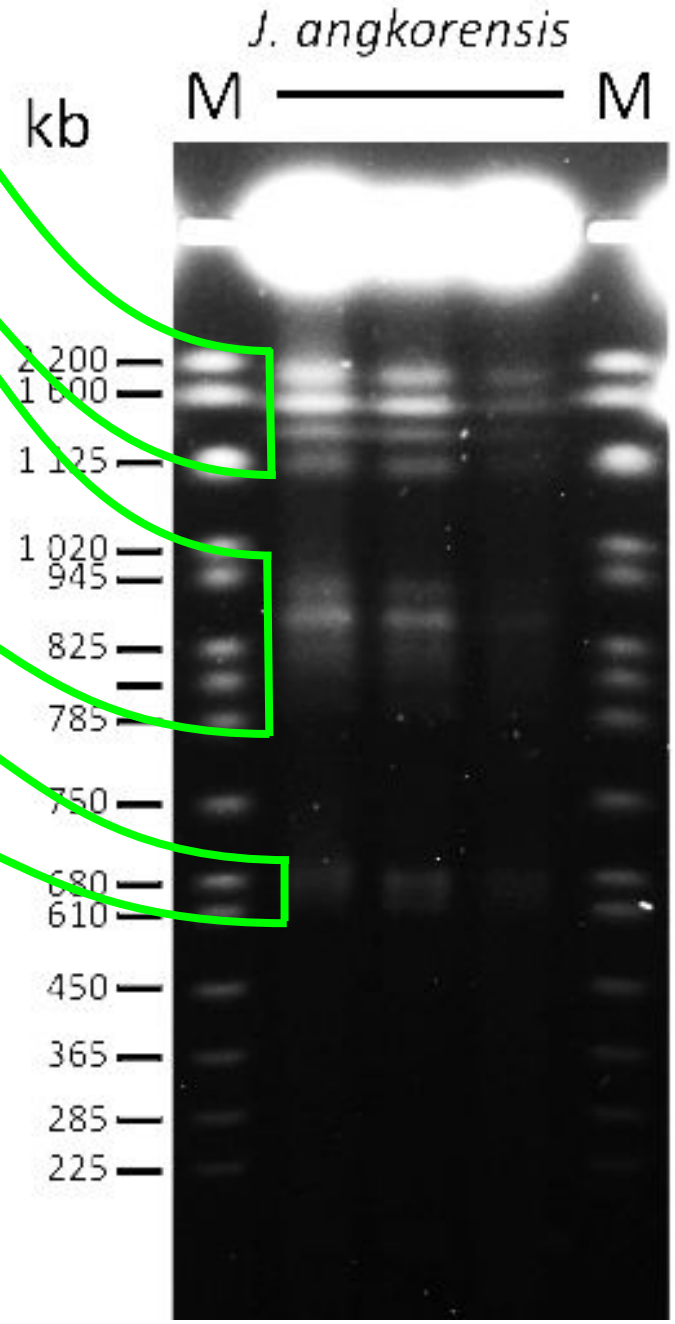


Zhodnotenie výsledkov

- Genómová sekvencia zostavená do 29 kontigov
- Identifikované tRNA gény, porovnanie využitých kodónov a kópií s inými druhmi
- Identifikovaných ďalších 9 RNA génov
- Predikcia 7284 génov v poskladanom genóme - použité na zostavenie fylogenetického stromu
- Porovnanie fylogenetického zaradenia s existujúcou literatúrou

- Preskúmanie existujúcich nástrojov na zostavenie a opravy malých genómov
- Vytvorenie viacerých nástrojov na konverziu a prispôsobenie výstupov existujúceho softvéru

p4tig00000001:0-2286041	2,286,041
p4tig00000017	1,280,186
p4tig00000022	1,112,239
p4tig00000020	1,111,661
p4tig00000066:0-983769	983,769
p4tig00000013:688156-1577203	889,047
p4tig00000028	885,255
p4tig00000025	870,338
p4tig00000030	846,148
p4tig00000037	830,089
p4tig00000042	806,314
p4tig00000031	786,406
p4tig00000033	767,511
p4tig00000040	758,635
p4tig00000069	712,617
p4tig00000013:0-688156	688,156
p4tig00000067	676,145
p4tig00000048	594,234
p4tig00000008:381663-975773	594,110
p4tig00000049	594,075
p4tig00000001:2286041-2859557	573,516
p4tig00000053	556,103
p4tig00000051	542,061
p4tig00000008:0-381663	381,663
p4tig00000066:983769-1282227	298,458
p4tig00000062	135,606
p4tig00000064	111,735
rDNA	81,042
mtDNA	29,999

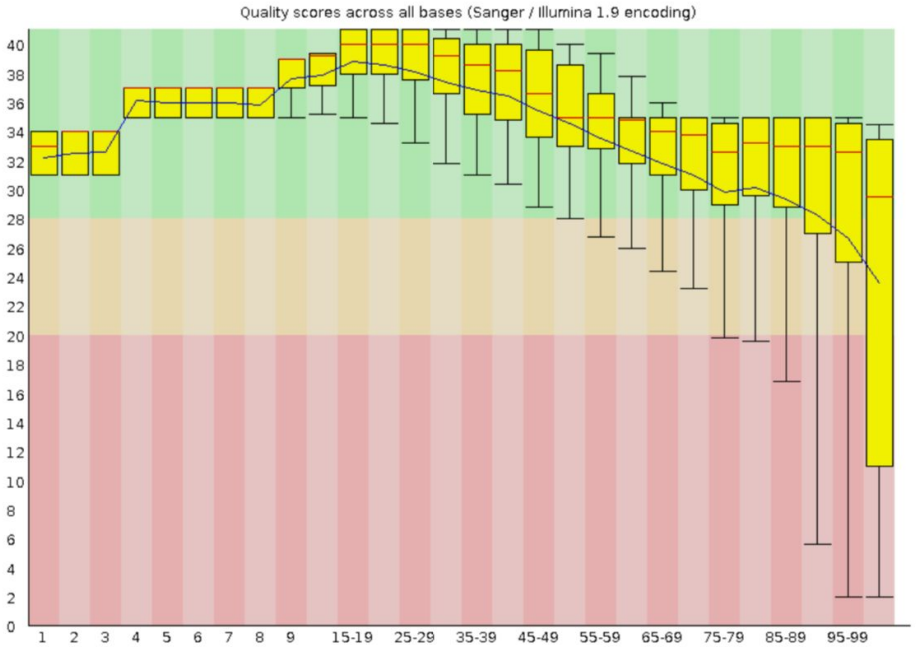


Počiatočné dáta

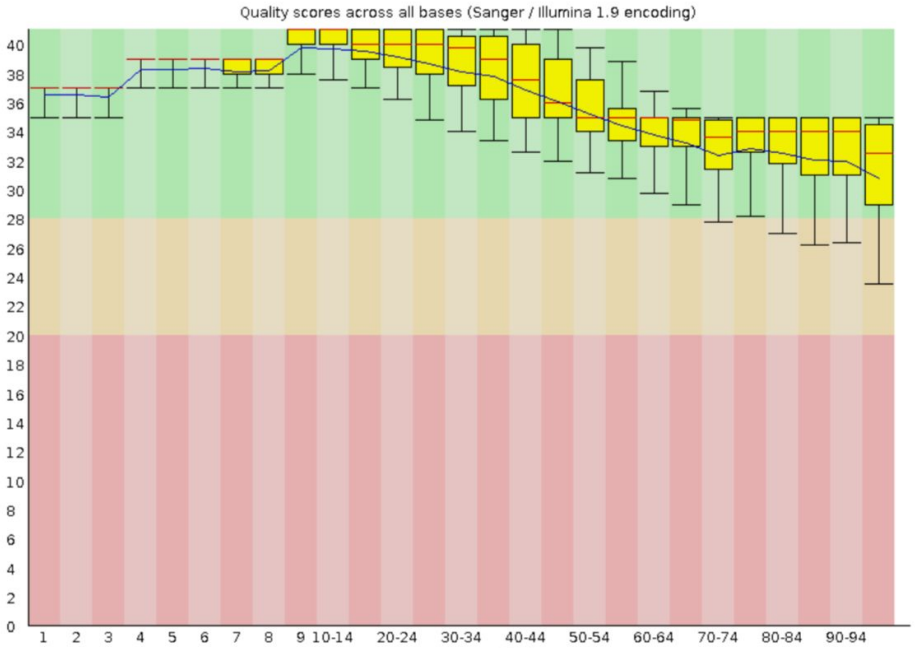
	Illumina PE	MinION1	MinION2	MinION3
Priemerná dĺžka	101 (2x)	2,836	3,864.5	11,305.5
Medián	101	1,841	1,786	8,083
Najdlhšie čítanie	101	163,344	420,011	118,582
Celkový počet báz	6,604,661,286	233,643,658	289,973,425	6,290,579,184
Pokrytie	330x	11x	14x	315x

- Skoršia verzia assembly vytvorená Velvet assemblerom
- Mitochondriálny genóm
- Anotovaná 7980 báz dlhá ribozomálna DNA sekvencia
- Obrázky pulsed-field gel elektroforézy

Úprava Illumina dát pomocou Trimmomatic



Pozícia v čítaní



Pozícia v čítaní

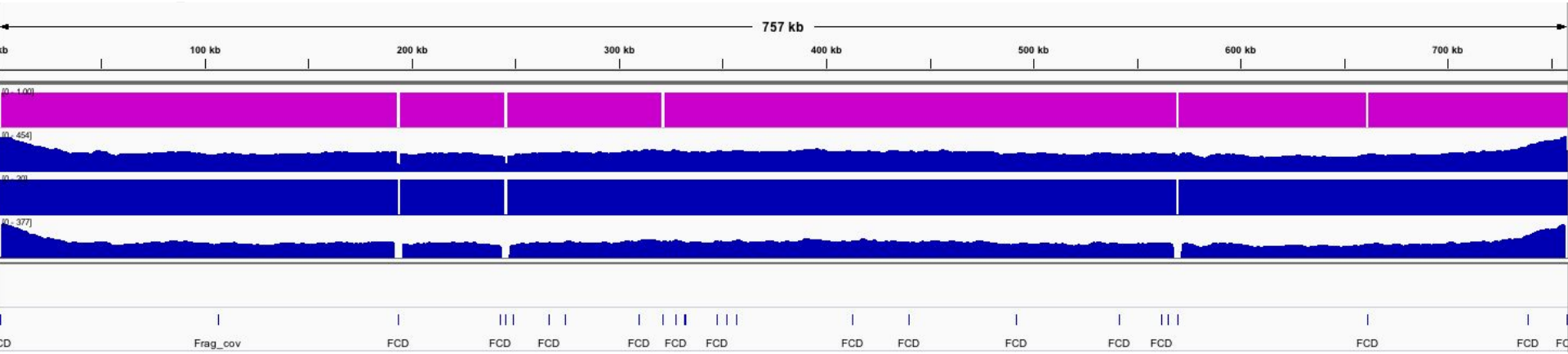
Skladanie genómu

- *Problém prečítať celý genóm - vieme čítať len krátke fragmenty*
- Testované programy:
- SPAdes (de Bruijn)
  - Bez MinION čítaní
  - S MinION čítaniami
- CANU (OLC)
  - MinION
- Miniasm (OLC)
  - MinION

	Velvet I	SPAdes I	SPAdes I+M	Canu	Miniasm
Počet kontigov	150	155	87	25	41
Najdlhší kontig	602,748	968,971	1,183,989	2,859,557	2,816,157
Veľkosť genómu	20,232,047	20,214,748	20,272,948	20,826,098	20,540,498
N50	223,699	286,195	436,155	855,255	813,073

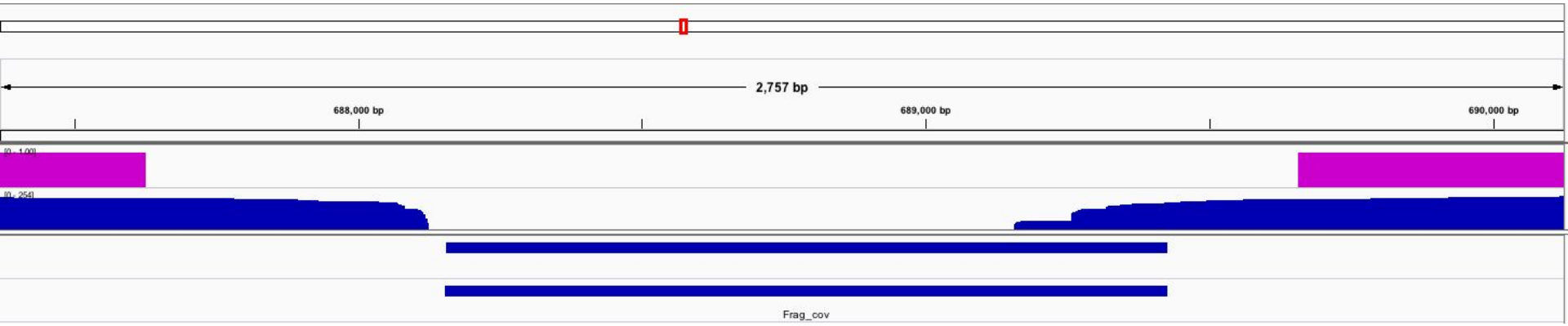
Následné opravy

- Pilon - lokálne zlepšenia kvality
  - Pomocou zarovnania krátkych, kvalitných čítaní opravuje chyby v assembly
  - 79,889 -> 3,228 -> 425 -> 45 zmien
- REAPR - odstránenie zle spojených miest
  - Využíva Illumina čítania na identifikáciu zle spojených miest na základe nerovnomerného alebo nízkeho pokrytia
  - 80 identifikovaných miest, 4 nami vybrané rozdelené
- Hľadanie oblastí so slabou celistvosťou pokrytia pomocou zarovnaní a úprav
  - Zarovnanie dlhých čítaní a SPAdes assembly voči našim kontigom, rozdelenie v nami vybraných miestach identifikovaných pomocou REAPR



Následné opravy

- Pilon - lokálne zlepšenia kvality
  - Pomocou zarovnania krátkych, kvalitných čítaní opravuje chyby v assembly
  - 79,889 -> 3,228 -> 425 -> 45 zmien
- REAPR - odstránenie zle spojených miest
  - Využíva Illumina čítania na identifikáciu zle spojených miest na základe nerovnomerného alebo nízkeho pokrytia
  - 80 identifikovaných miest, 4 nami vybrané rozdelené
- Hľadanie oblastí so slabou celistvosťou pokrytia pomocou zarovnaní a úprav
  - Zarovnanie dlhých čítaní a SPAdes assembly voči našim kontigom, rozdelenie v nami vybraných miestach identifikovaných pomocou REAPR



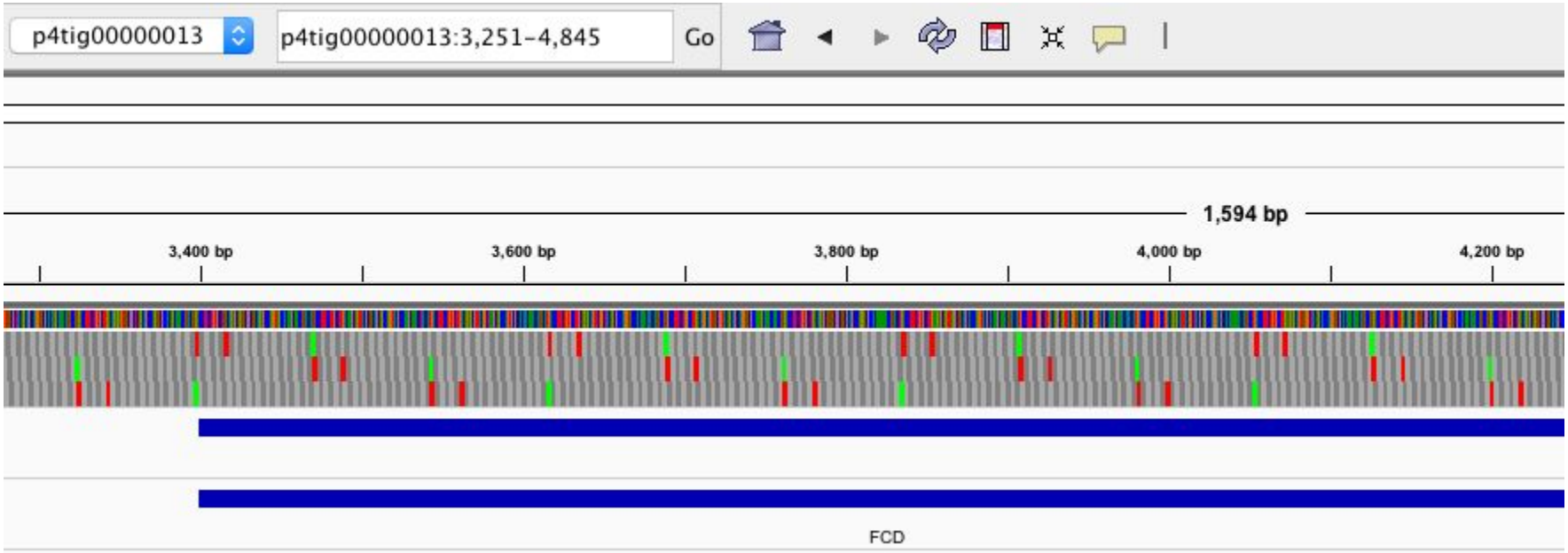
Následné opravy

- Pilon - lokálne zlepšenia kvality
  - Pomocou zarovnania krátkych, kvalitných čítaní opravuje chyby v assembly
  - 79,889 -> 3,228 -> 425 -> 45 zmien
- REAPR - odstránenie zle spojených miest
  - Využíva Illumina čítania na identifikáciu zle spojených miest na základe nerovnomerného alebo nízkeho pokrytia
  - 80 identifikovaných miest, 4 nami vybrané rozdelené
- Hľadanie oblastí so slabou celistvosťou pokrytia pomocou zarovnaní a úprav
  - Zarovnanie dlhých čítaní a SPAdes assembly voči našim kontigom, rozdelenie v nami vybraných miestach identifikovaných pomocou REAPR

- LINKS - scaffolding
  - Žiadne zlepšenie na našom genóme
  - Po nasekaní genómu ho zrekonštruoval naspäť

Hľadanie telomér

- *Teloméry sú repetitívne sekvencie na koncoch chromozómov*
- Zarovnanie koncov kontigov voči sebe
  - Identifikovaných niekoľko dvojíc kontigov
- Tandem Repeat Finder
  - 2869 identifikovaných opakujúcich sa sekvencií
  - Vybrané iba consensus sekvencie s vysokým skóre (veľa nájdených presných opakovaní)
  - Žiadne opakovanie zdieľané viac ako dvojicou kontigov



Interpretovanie sekvencie

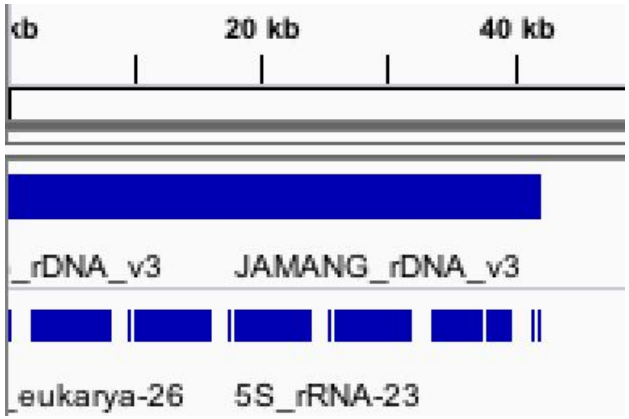
- tRNAscan-SE

Organizmus	J. angkorensis	U. maydis	S. reilianum	S. cerevisiae
Celkový počet	55	110	106	273

- CMscan

RNA type	Rfam ID
5' ureB small RNA	RF02514
Group I catalytic intron	RF00028
RNase MRP	RF00030
TPP riboswitch (THI element)	RF00059
U1 spliceosomal RNA	RF00003
U2 spliceosomal RNA	RF00004
U4 spliceosomal RNA	RF00015
U6 spliceosomal RNA	RF00026
Small nucleolar RNA Z13/snr52	RF00335

- Zarovnanie mtDNA, rDNA



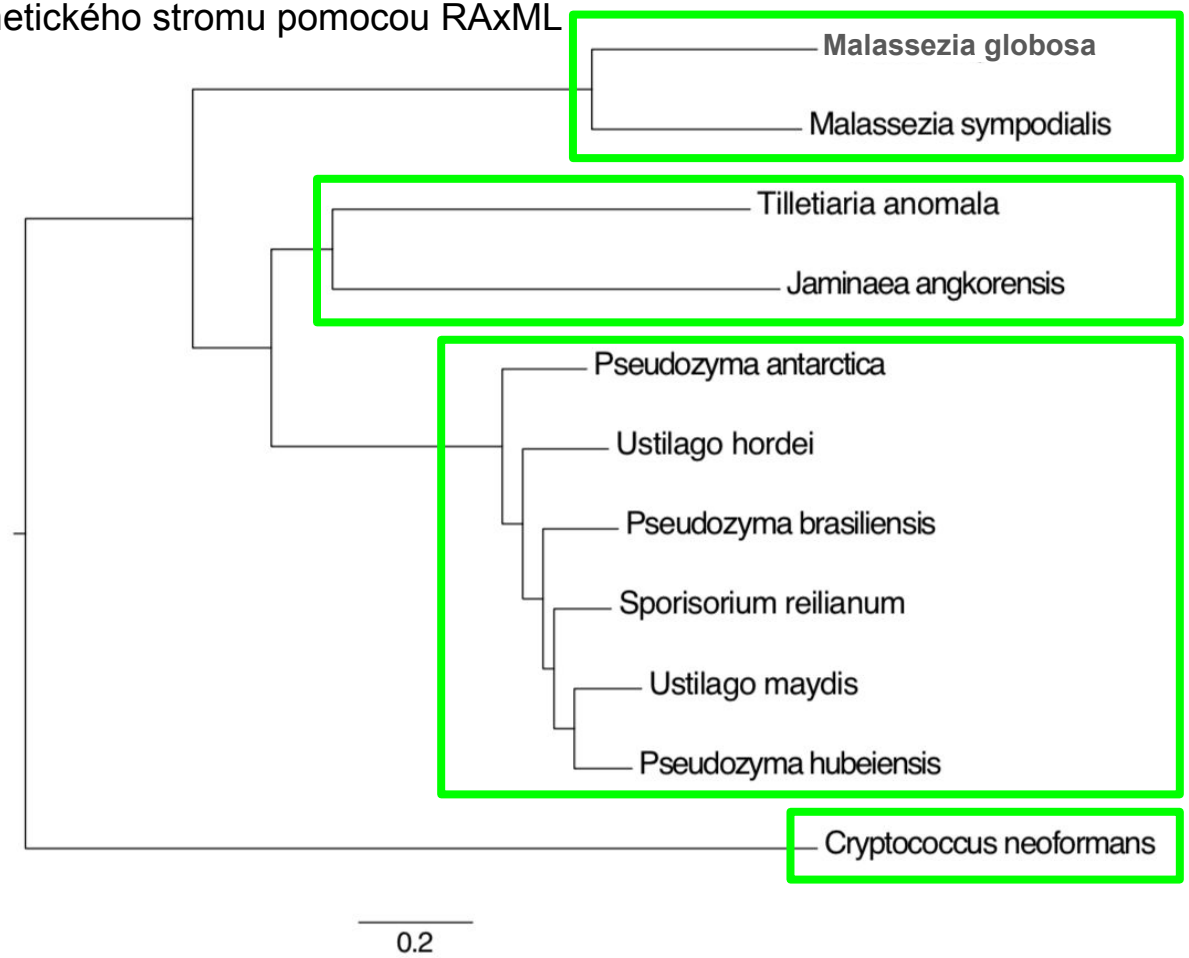
Porovnanie genómu s inými druhmi

- Proteómy získané z databázy Uniprot
- 9 príbuzných húb, jedna zvolená ako outgroup

Species	Number of proteins
Jaminaea angkorensis	7,284
Malassezia globosa	4,274
Malassezia sympodialis	3,318
Pseudozyma antarctica	6,640
Pseudozyma brasiliensis	5,765
Pseudozyma hubeiensis	7,472
Sporisorium reilianum	6,673
Ustilago hordei	7,111
Ustilago maydis	6,806
Tilletiaria anomala	6,806
Cryptococcus neoformans	6,743

Porovnanie s inými druhmi

- Hľadanie ortológov pomocou BLAST a následného výberu najlepších vzájomných zhôd
- Spoločné globálne zarovnanie všetkých ortológov pomocou MUSCLE
- Tvorba spoločného fylogenetického stromu pomocou RAxML



Vytvorené nástroje

Prevody do BED formátu:

- tRNAscan-SEtoBED
- cmscanToBED
- trfToBED

Manipulácia dát:

- trfToConsensusFasta
- reaprMiddleCuts
- tRNAscan-SEcodonsToTex

Ďalšia práca

- Využitie RNA-seq dát na identifikáciu transkribovaných oblastí na násldnú anotáciu
- Doplnenie údajov do genómového prehliadača
- Popis poskladaného genómu a analýza zistení

- Využitie iných programov na úpravu čítaní, spájanie kontigov a iné vylepšenia (Nanocorrect, pyScaf,...)

- Refaktorovanie vytvorených skriptov pre jednoduchšie použitie

Ďakujem za pozornosť

Otázky

Strana 8.: Píšete pokrytie genómu J.A. - odkiaľ vieme dĺžku tohoto genómu?

- Konzistentná dĺžka poskladaného genómu z rôznych dát a algoritmov (rozpätie cca. 20-22 Mbp)

	Velvet I	SPAdes I	SPAdes I+M	Canu	Miniasm
Najdlhší kontig	602,748	968,971	1,183,989	2,859,557	2,816,157
Veľkosť genómu	**20,232,047**	**20,214,748**	**20,272,948**	**20,826,098**	**20,540,498**

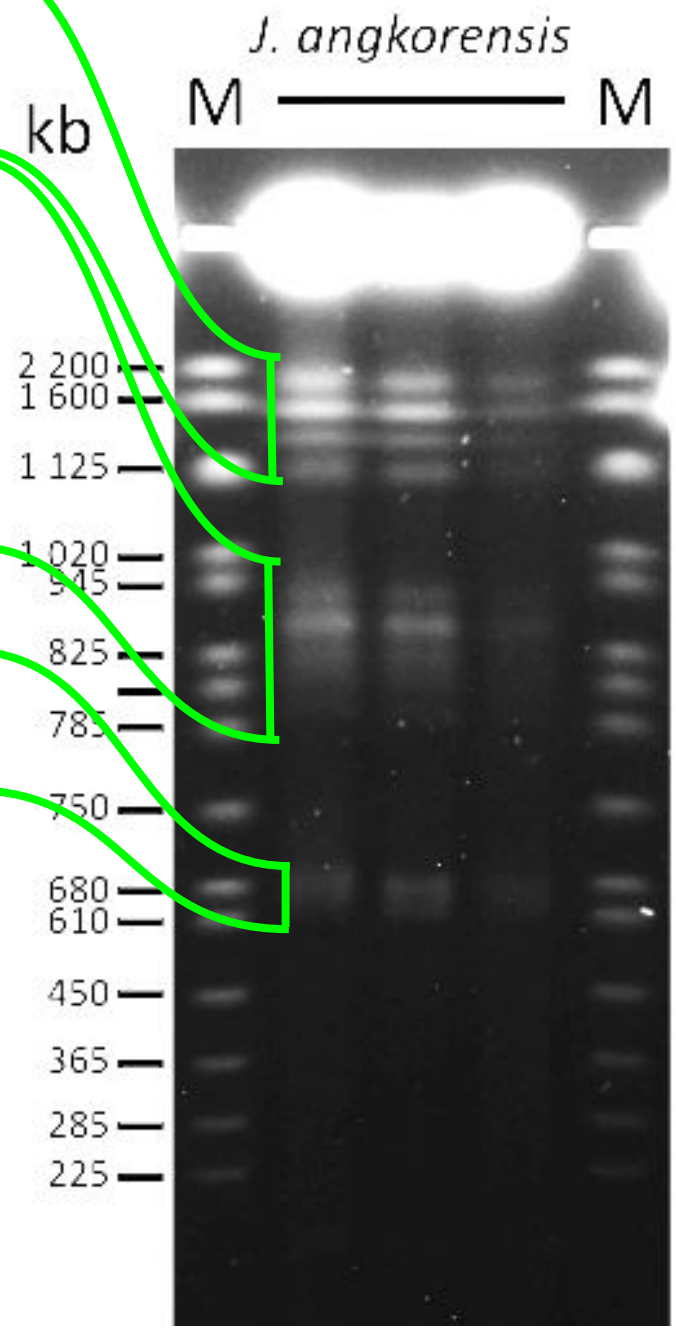
Strana 15.: Obrázok 2.2 je nedostatočne vysvetlený - koľko a aké dlhé chromozómy predpokladáme na základe elektroforézy? Je známy počet chromozómov pre J.A. popr. príbuzné druhy?

- Mohla by mať 8-10 chromozómov rádovo v rozsahu chromozómov S. cerevisiae (4 chromozómy sú medzi 1-2 Mbp (najväčší má cca 2 Mbp) a ostatné sú menšie 0.5-0.9 Mbp).[1]

- Na základe samotnej dĺžky genómu počet nesedí - zrejme je chromozómov viac (20Mb/10 = 2Mbp priemer)

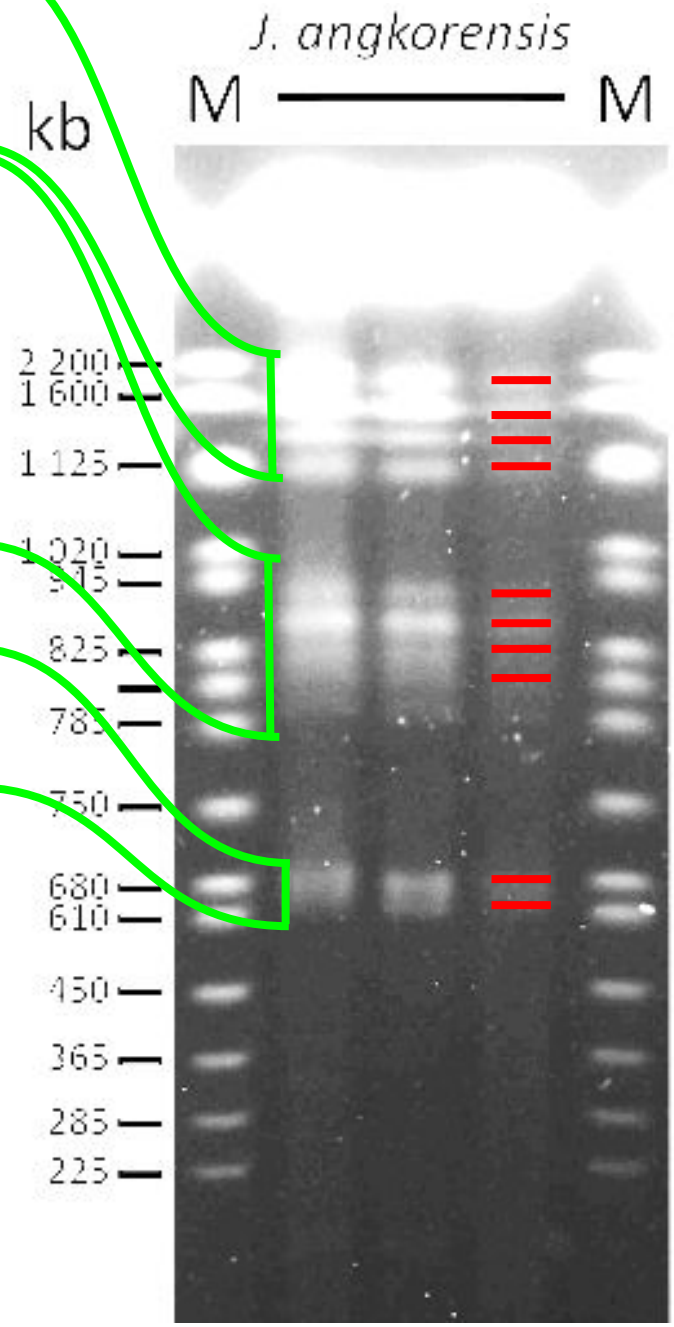
1. Nosek, J., parafrázované

p4tig00000001:0-2286041	2,286,041
p4tig00000017	1,280,186
p4tig00000022	1,112,239
p4tig00000020	1,111,661
p4tig00000066:0-983769	983,769
p4tig00000013:688156-1577203	889,047
p4tig00000028	885,255
p4tig00000025	870,338
p4tig00000030	846,148
p4tig00000037	830,089
p4tig00000042	806,314
p4tig00000031	786,406
p4tig00000033	767,511
p4tig00000040	758,635
p4tig00000069	712,617
p4tig00000013:0-688156	688,156
p4tig00000067	676,145
p4tig00000048	594,234
p4tig00000008:381663-975773	594,110
p4tig00000049	594,075
p4tig00000001:2286041-2859557	573,516
p4tig00000053	556,103
p4tig00000051	542,061
p4tig00000008:0-381663	381,663
p4tig00000066:983769-1282227	298,458
p4tig00000062	135,606
p4tig00000064	111,735
rDNA	81,042
mtDNA	29,999

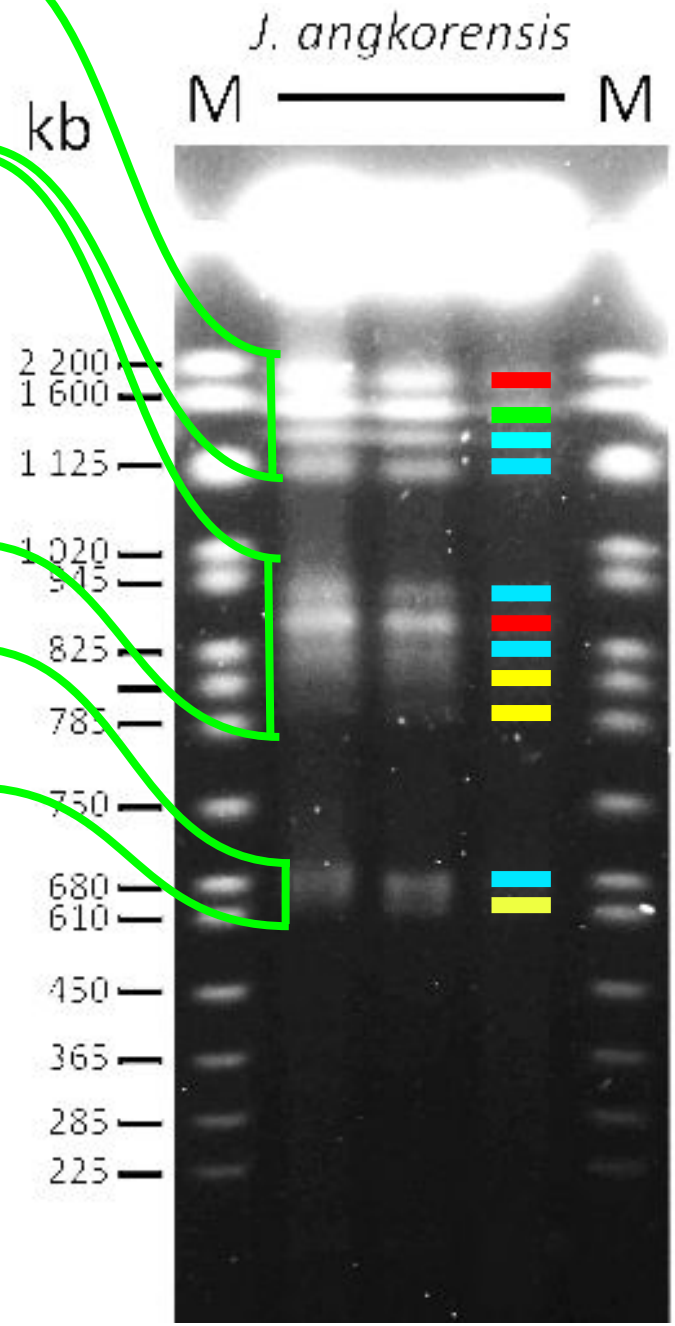


- 15 ~ok
- 10 mimo
- 4 nevidno

p4tig00000001:0-2286041	2,286,041
p4tig00000017	1,280,186
p4tig00000022	1,112,239
p4tig00000020	1,111,661
p4tig00000066:0-983769	983,769
p4tig00000013:688156-1577203	889,047
p4tig00000028	885,255
p4tig00000025	870,338
p4tig00000030	846,148
p4tig00000037	830,089
p4tig00000042	806,314
p4tig00000031	786,406
p4tig00000033	767,511
p4tig00000040	758,635
p4tig00000069	712,617
p4tig00000013:0-688156	688,156
p4tig00000067	676,145
p4tig00000048	594,234
p4tig00000008:381663-975773	594,110
p4tig00000049	594,075
p4tig00000001:2286041-2859557	573,516
p4tig00000053	556,103
p4tig00000051	542,061
p4tig00000008:0-381663	381,663
p4tig00000066:983769-1282227	298,458
p4tig00000062	135,606
p4tig00000064	111,735
rDNA	81,042
mtDNA	29,999



p4tig00000001:0-2286041	2,286,041
p4tig00000017	1,280,186
p4tig00000022	1,112,239
p4tig00000020	1,111,661
p4tig00000066:0-983769	983,769
p4tig00000013:688156-1577203	889,047
p4tig00000028	885,255
p4tig00000025	870,338
p4tig00000030	846,148
p4tig00000037	830,089
p4tig00000042	806,314
p4tig00000031	786,406
p4tig00000033	767,511
p4tig00000040	758,635
p4tig00000069	712,617
p4tig00000013:0-688156	688,156
p4tig00000067	676,145
p4tig00000048	594,234
p4tig00000008:381663-975773	594,110
p4tig00000049	594,075
p4tig00000001:2286041-2859557	573,516
p4tig00000053	556,103
p4tig00000051	542,061
p4tig00000008:0-381663	381,663
p4tig00000066:983769-1282227	298,458
p4tig00000062	135,606
p4tig00000064	111,735
rDNA	81,042
mtDNA	29,999



1x4

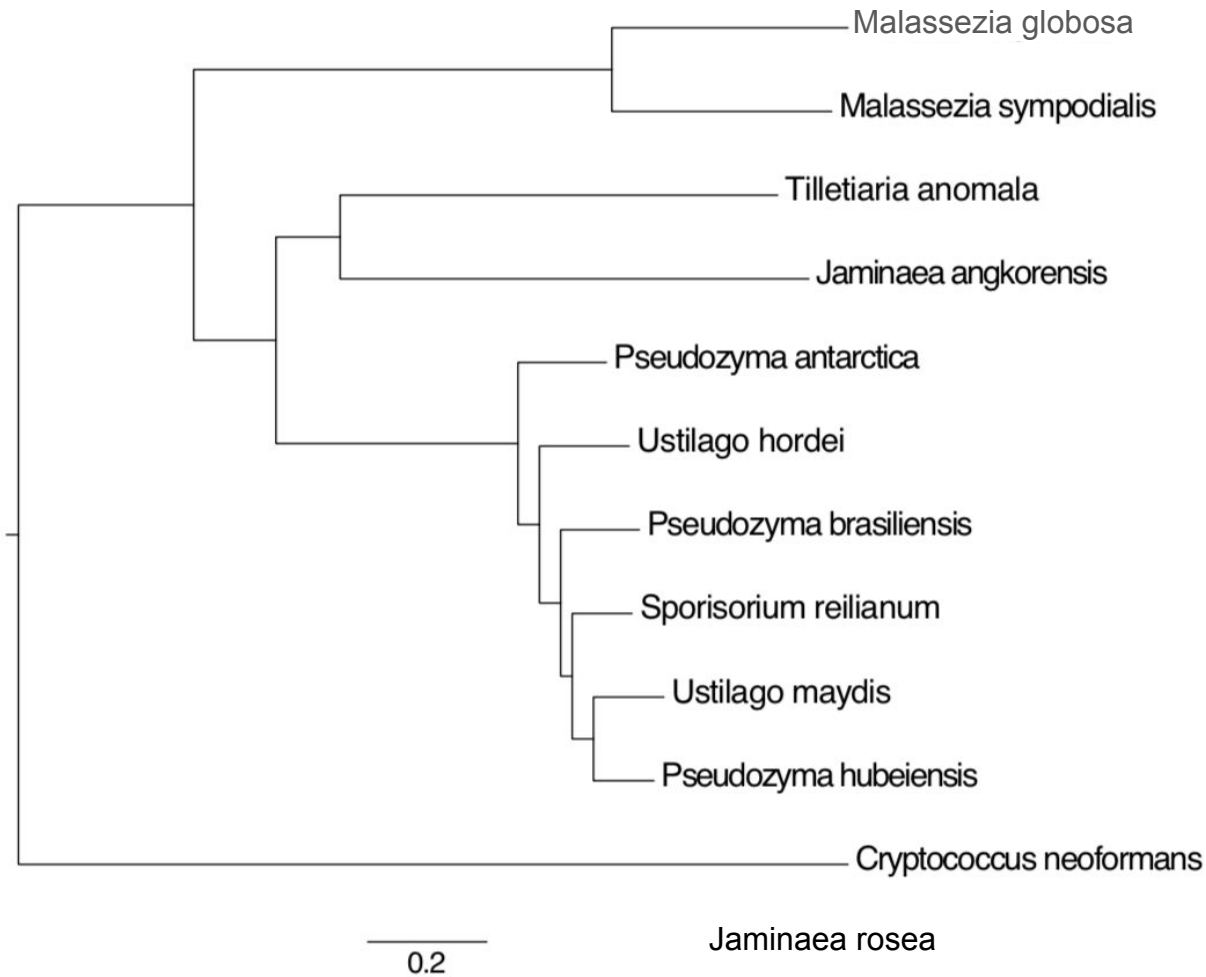
2x3

5x2

3x1

Spolu: 23

Porovnanie s inými druhmi



	Veľkosť genómu	Počet chromozómov
Malassezia globosa	8.96	-
Malassezia sympodialis	7.73 - 7.87	8
Tilletiaria anomala	18.70	-
Jaminaea angkorensis	20.88	? (29)
Pseudozyma antarctica	18.08	- (27)
Ustilago hordei	21.15 - 27.05	-
Pseudozyma brasiliensis	17.32	-
Sporisorium reilianum	18.52	23
Ustilago maydis	19.66 - 20.62	23
Pseudozyma hubeiensis	18.44 - 21.32	-
Cryptococcus neoformans	18.59	14
Jaminaea rosea	16.61	-(21)
Saccharomyces cerevisiae	12.12	16

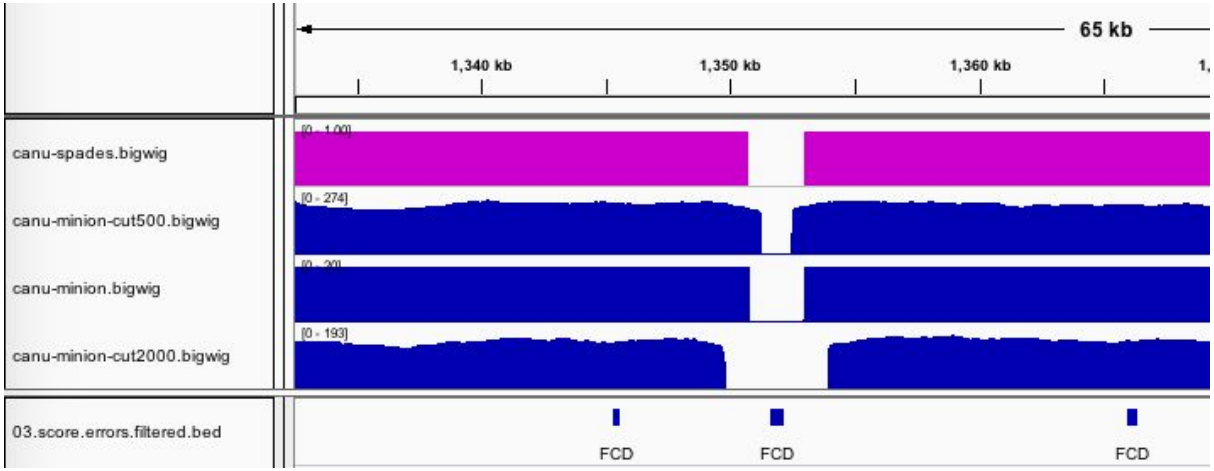
Dĺžka kontigov nie úplne sedí s predpokladanou dĺžkou z elektroforézy. Ako si to vysvetľujete? Mohlo by toto byť vysvetlené iným počtom repetitívnych sekvencií?

- Repetitívne sekvencie sú vo všeobecnosti problémom pre assemblery
  - MinION dáta by mali mať dostatočnú dĺžku na ich pokrytie (v assembly len 4 nájdené repetitívne časti dlhšie ako priemerný read, OLC assembler)
  - REAPR však na základe distribúcie pokrytia párovými readmi považoval repetitívne sekvencie v assembly za problematické
- Nenašli sme telomerickú sekvenciu
  - Chýbajú dáta z krajov chromozómov
  - Kontigy sú fragmentovanejšie, niektorých kombinácia môže zodpovedať chromozómu

https://www.nature.com/articles/nbt.4060

Strana 19.: Čím si vysvetľujete, že sa nepodarilo spojiť kontigy nástrojom LINKS? Bolo by možné použiť iný nástroj?

- Dôvody:
  - Zle použité parametre (vzdialenosť medzi pármi k-merov, % odchýlky vzdialeností)
  - Chyba v assembly znemožňujúca spojiť správny pár kontigov (zlý počet repetícií, chybný spoj)
    - **Potenciálne riešenie:**agresívnejšie sekanie sekvencie REAPRom, väčšia povolená odchýlka



  - Assembly
- Áno, napr. pyScaf
  - Reference-based scaffolding

Otázky