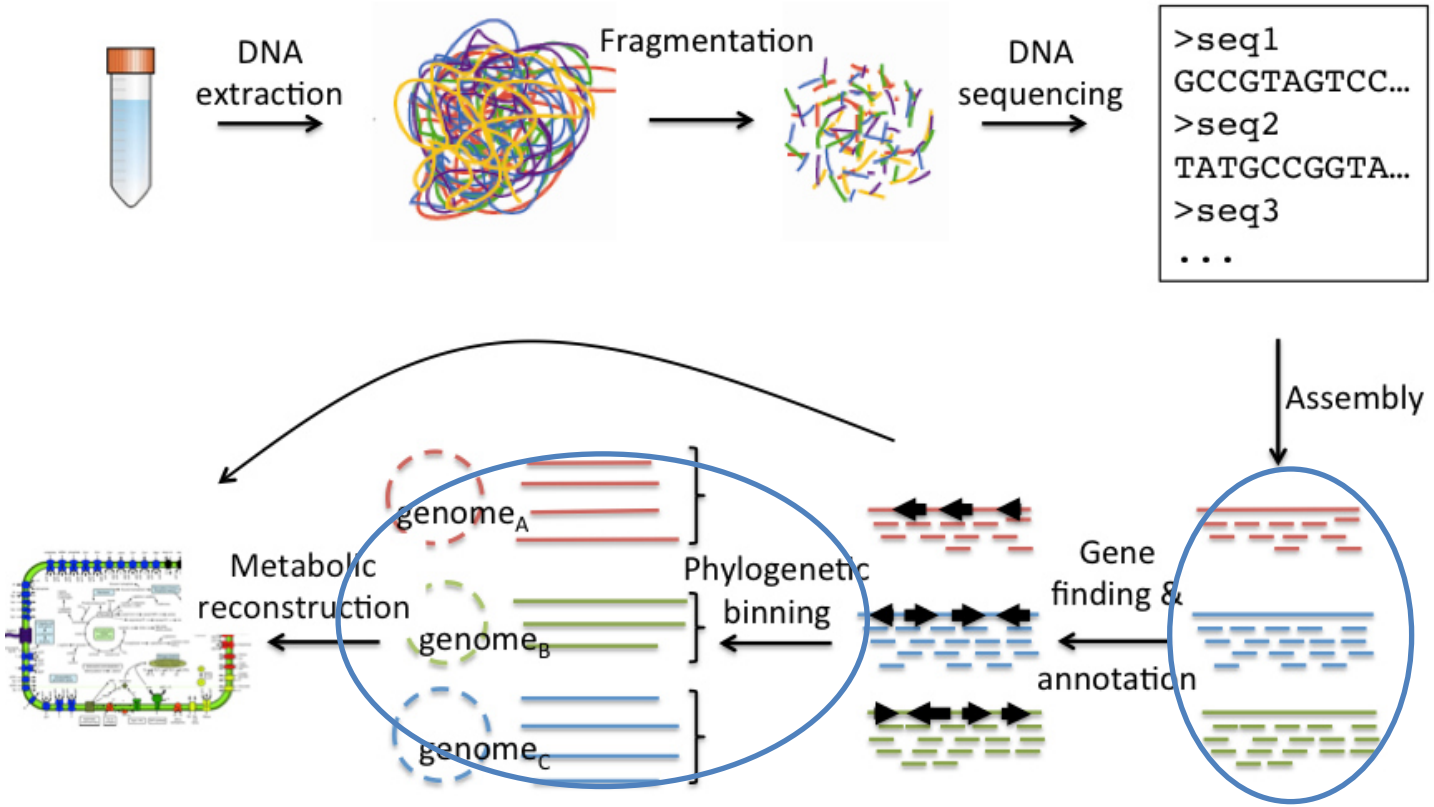
Metagenomika – Zpracování a assembly celometagenomových (shotgun) dat

Mgr. Ing. Karel Sedlář

Celometagenomová sekvenace



Jak můžeme efektivně sestavit kontigy, které je možné fylogeneticky zaškatulkovat?

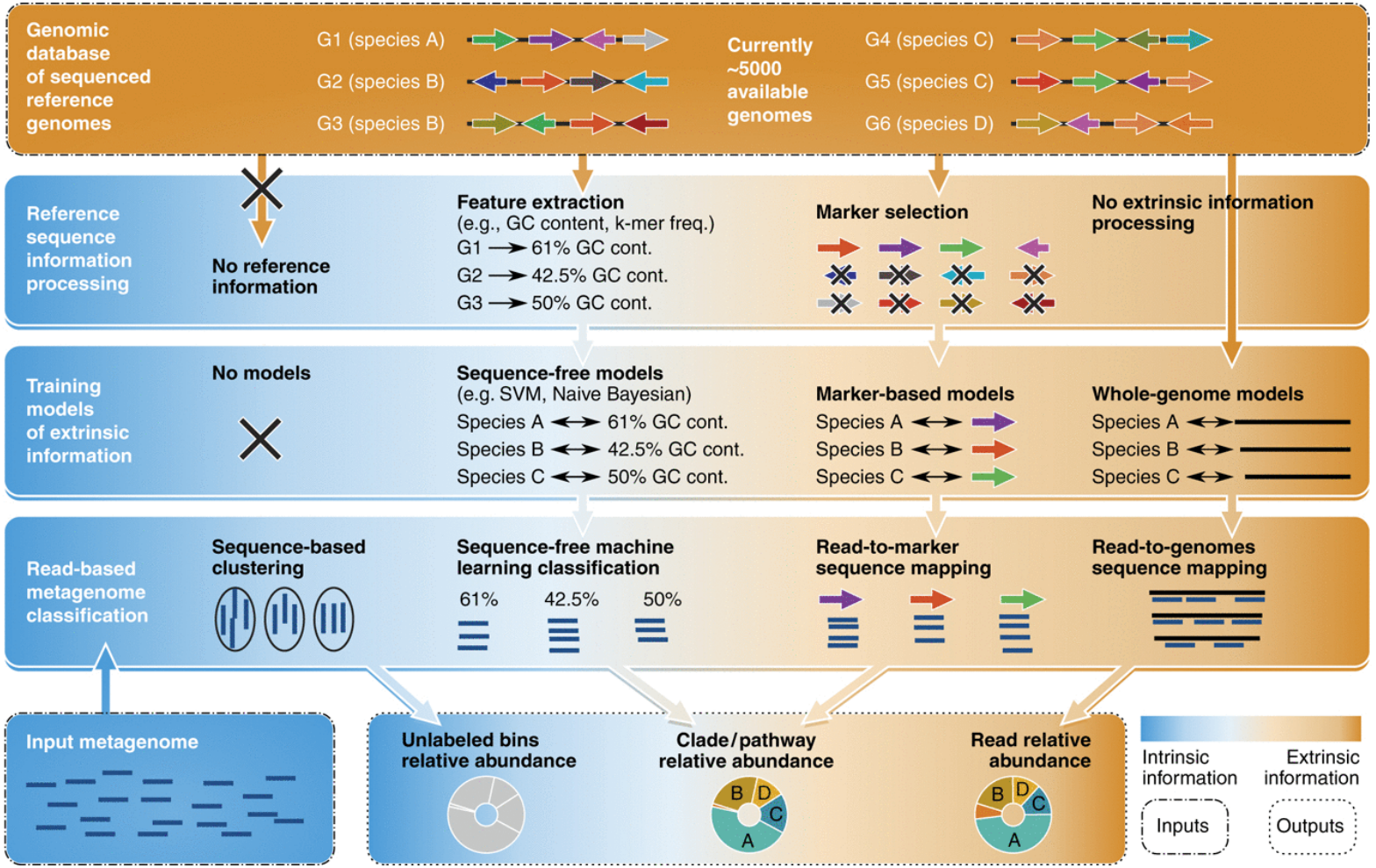
Jak poznáme ke kterému genomu čtení patří?

Celometagenomová sekvenace

- laboratorní metody se vyvíjí velmi rychle, přičemž vyžadují neustálý vývoj nových výpočetních metod, který ale probíhá se zpožděním
- „wet-lab“ metody pro získání metagenomických (MG) a metatranskriptomických dat (MT) jsou formalizované a reprodukovatelné
- „dry-lab“ metody v tomto ohledu zaostávají, formalizace postupů je složitá
- výpočetně velmi náročné problémy, často neřešitelné deterministicky

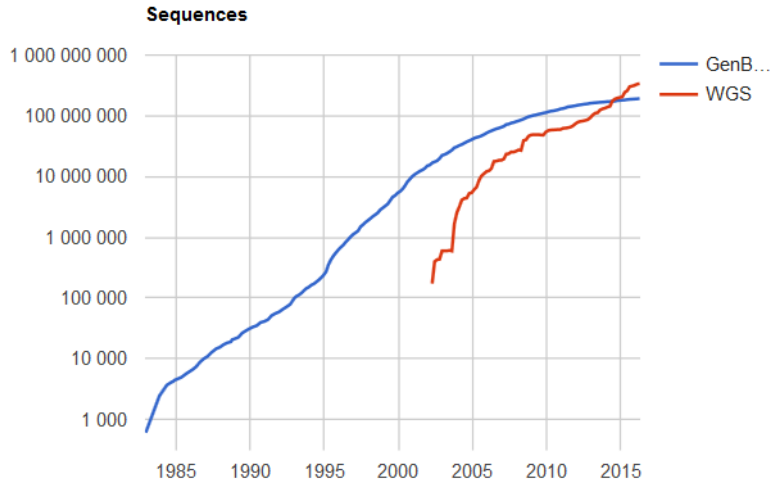
Základní přístupy

- metody pro zpracování surových dat (čtení):
  - závislé na referenci × nezávislé na referenci



Metody závislé na referenci

- založené na přímém mapování či zarovnávání čtení k referenčním databázím
- 4-6 x 10^{30} prokaryot
- GenBank assembly prokaryotních genomů: 68 450
- kompletních jen 5282
- srovnávání s databází je pomalé
- i pro lidský střevní mikrobiom stále chybí reference pro 43 % genomů
- výsledek: relativní abundance čtení v jednotlivých skupinách



GenBank: 193 739 511

WGS: 338 922 537

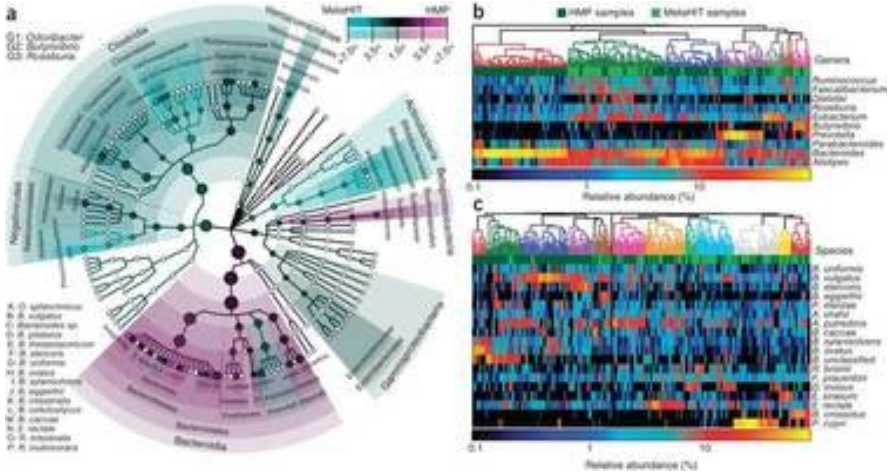
Metody závislé na referenci

- nemusí data nutně porovnávat s celými sekvencemi, ale jen s markerovými subsekvencemi (třeba vybranými geny)
- srovnání je tak časově méně složité
- může probíhat 2 způsoby:
  - porovnává samostatná čtení jako celé sekvence
  - porovnává profilovou informaci ze čtení (př. počty specifických k-merů)
- důležitým parametrem je délka porovnávaných sekvencí → vhodnější pro dlouhá čtení
- vhodné spíše pro MT než pro MG
- výsledek: relativní abundance drah/taxonomických skupin

Metody závislé na referenci

- nástroje pracující s vybranými geny:
  - AMPHORA (31 vybraných genů)
  - MetaPhyler (navíc geny pro doménu Archea)
  - AMPHORA2 (navíc geny pro doménu Archea)
  - MetaPhlAn (různé geny pro různé vývojové větve)

- nástroje pracující s celými sekvencemi:
  - BLASTN
  - BowTie2
  - BWA

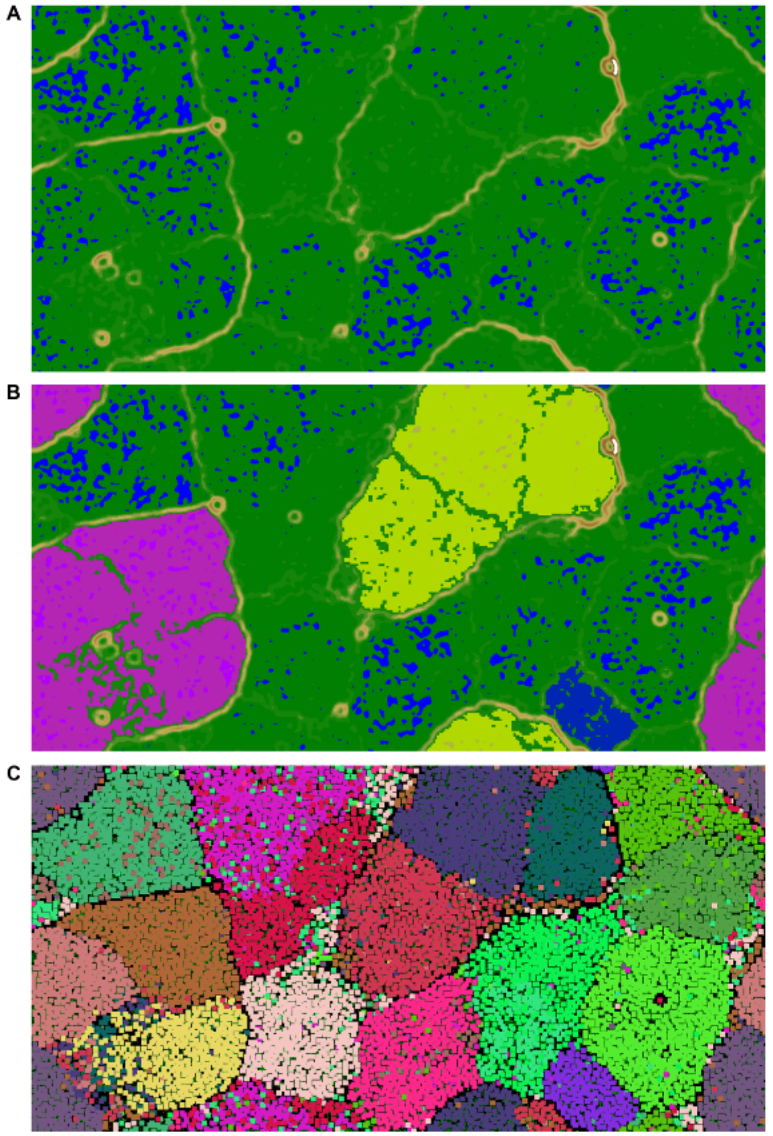


Metody nezávislé na referenci

- nevyžadují apriorní znalost → využijí i čtení, která patří dosud nepopsaným genomům
- může opět pracovat s celými sekvencemi nebo profilovou informací
- (mezi)výsledek: relativní abundance skupin podobných, neidentifikovaných sekvencí
- umožní sestavení delších sekvencí (kontigů), které jsou teprve následně identifikovány
- vyžaduje de novo assembly dat, často několika krokovou s postupnou klasifikací vznikajících kontigů

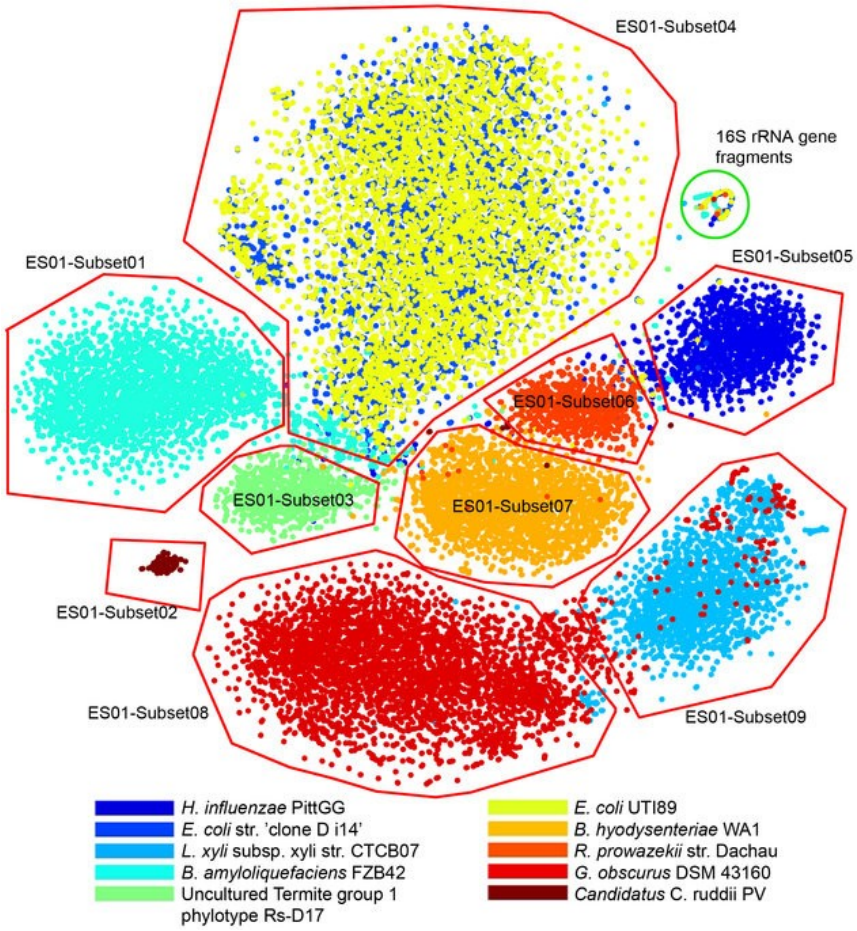
Metody nezávislé na referenci

- SOM (samoorganizační mapy)
  - založené na zpracování počtů k-merů a následné redukci dimenzionality
  - původní dimenzionalita je daná délkou k-meru: 4^k
  - pro k-mery délky 5 nukleotidů je to $4^5 = 1024$



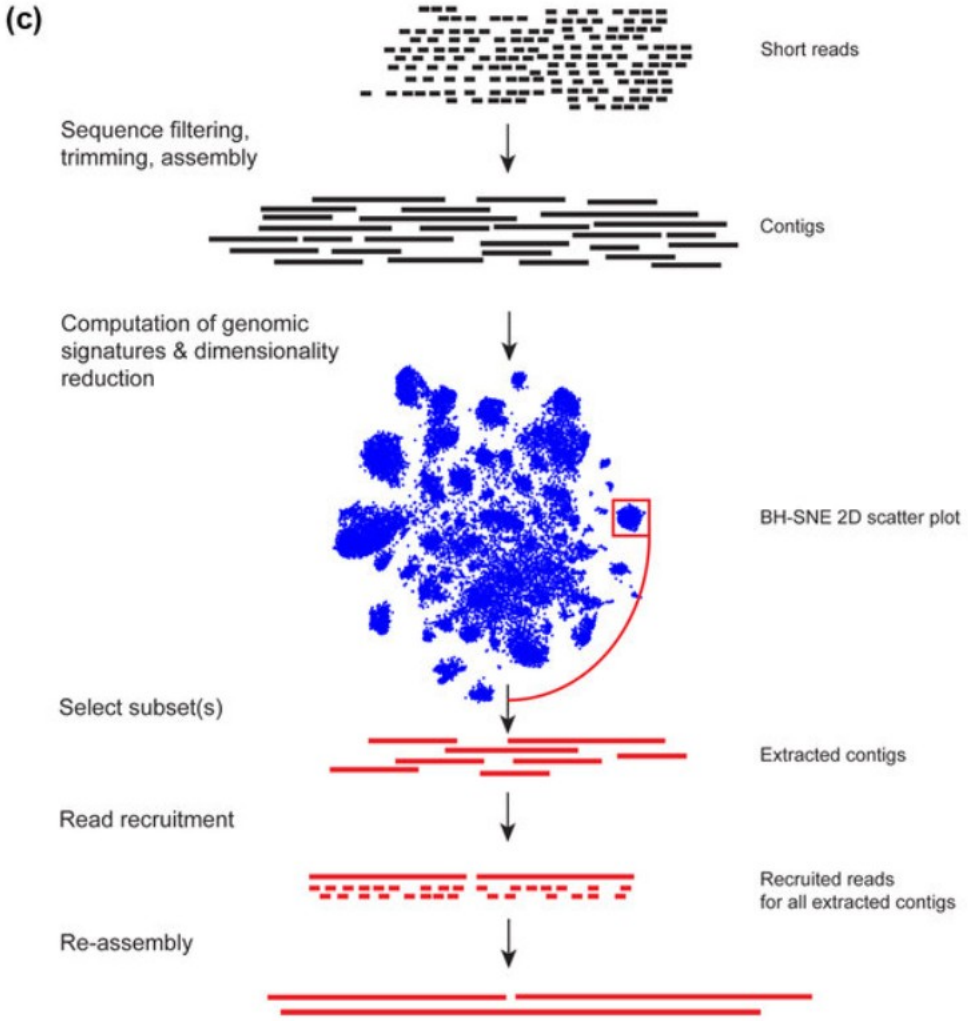
Metody nezávislé na referenci

- VizBin
  - založené také na zpracování počtů k-merů a následné redukci dimenzionality pomocí BH-SNE (Barnes-Hut Stochastic Neighbor Embedding)
  - používá k-mery délky 4 nukleotidů, původní dimenzionalita je to 4^4 = 256
  - redukce do 2D
  - problém je automatické shlukování



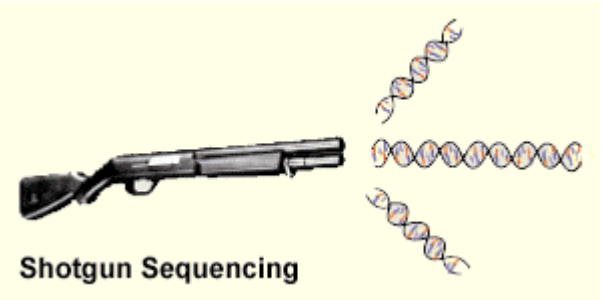
Metody nezávislé na referenci

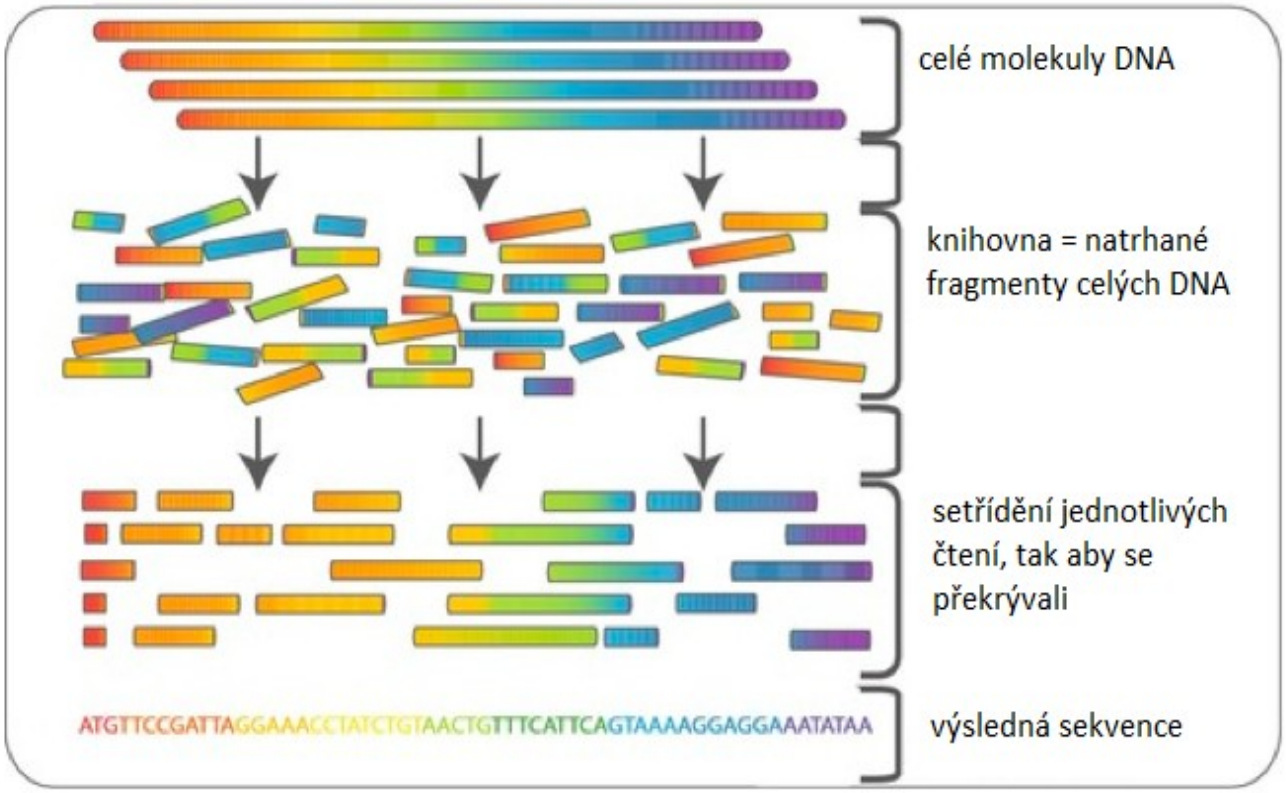
- využití v kombinaci s de novo assembly



de novo assembly

- shotgun data



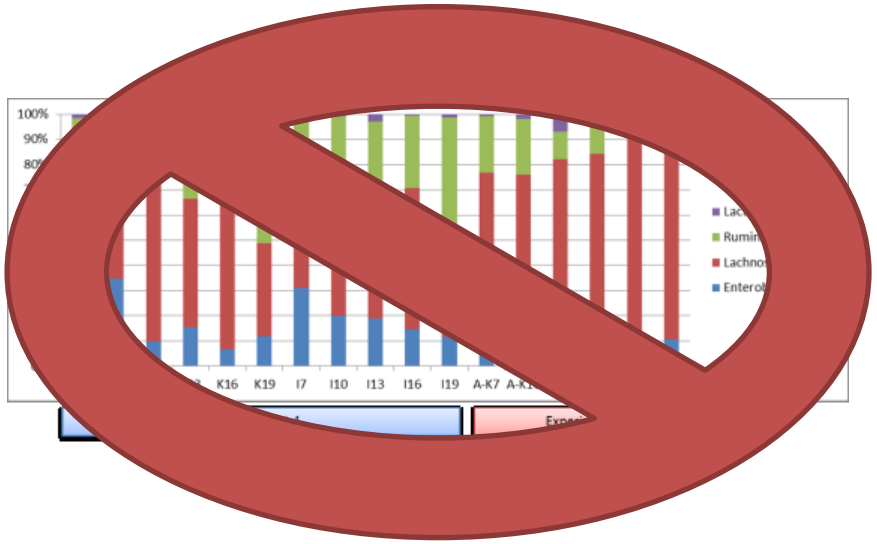


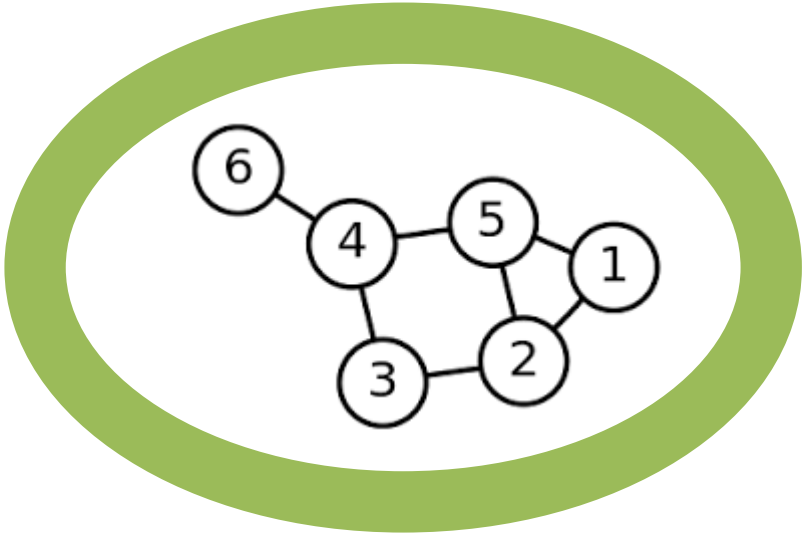
de novo assembly

- de novo assembly je odjakživa jedním z hlavních problémů bioinformatiky
- problematická i pro data jednoho konkrétního genomu, natož při zpracování shotgun metagenomů
- kvůli výpočetní náročnosti není možné použít dynamické programování (nw, sw, clustal)
- velké množství algoritmů, chybí celkové srovnání, celý obor se rychle vyvíjí
- zásadní roli hraje délka čtení → čím delší čtení, tím delší kontigy sestavíme při nižší coverage

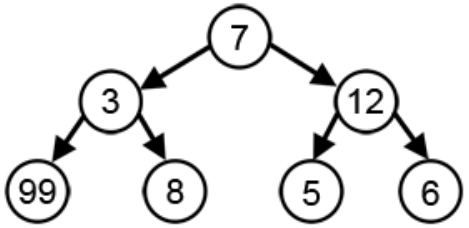
de novo assembly

- znakové metody
- grafové metody
  - OLC grafy
  - de Bruijn grafy





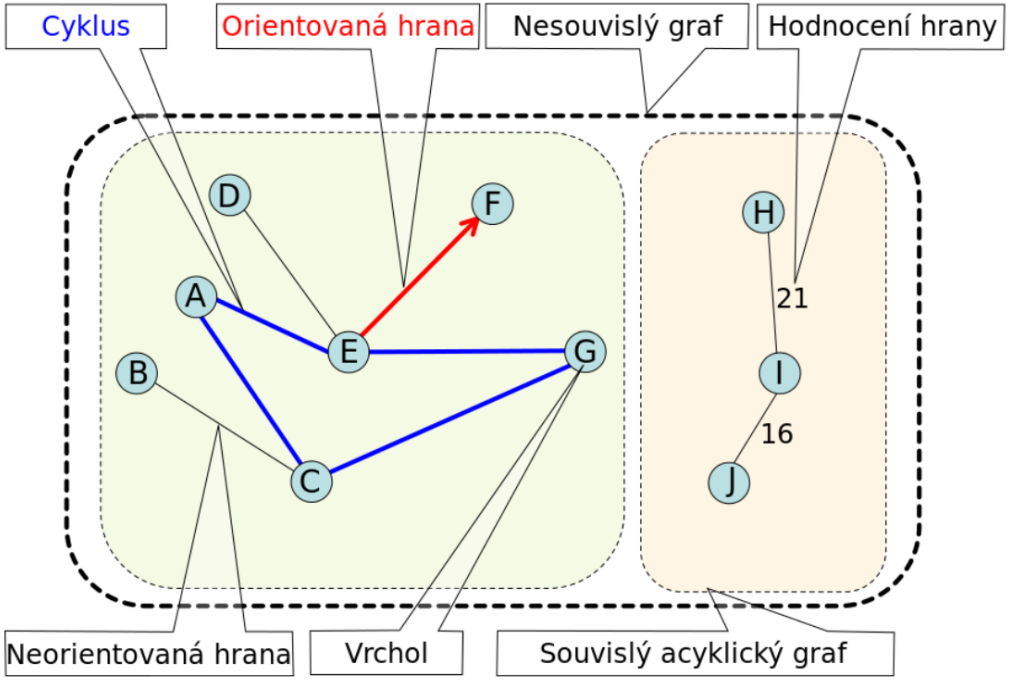
de novo assembly



- znakové metody
  - greedy extension algoritmy (hladové algoritmy)
  - „hladový“ algoritmus spojuje jedno čtení s druhým, tak aby bylo dosaženo co největšího překryvu, skončí když už nelze připojit další
  - největší překryv neznamená vždy nejlepší řešení
  - tendence poskytovat sub-optimální řešení
  - dobré pro malé genomy a krátká čtení
  - vyšší nárok na operační paměť počítače
  - v metagenomice v praxi nepoužitelné
  - nástroje: SSAKE, VCAKE

*Graf

- totožný pojem jako síť
- $G = (V, E)$
- uspořádaní dvojice vrcholů a hran



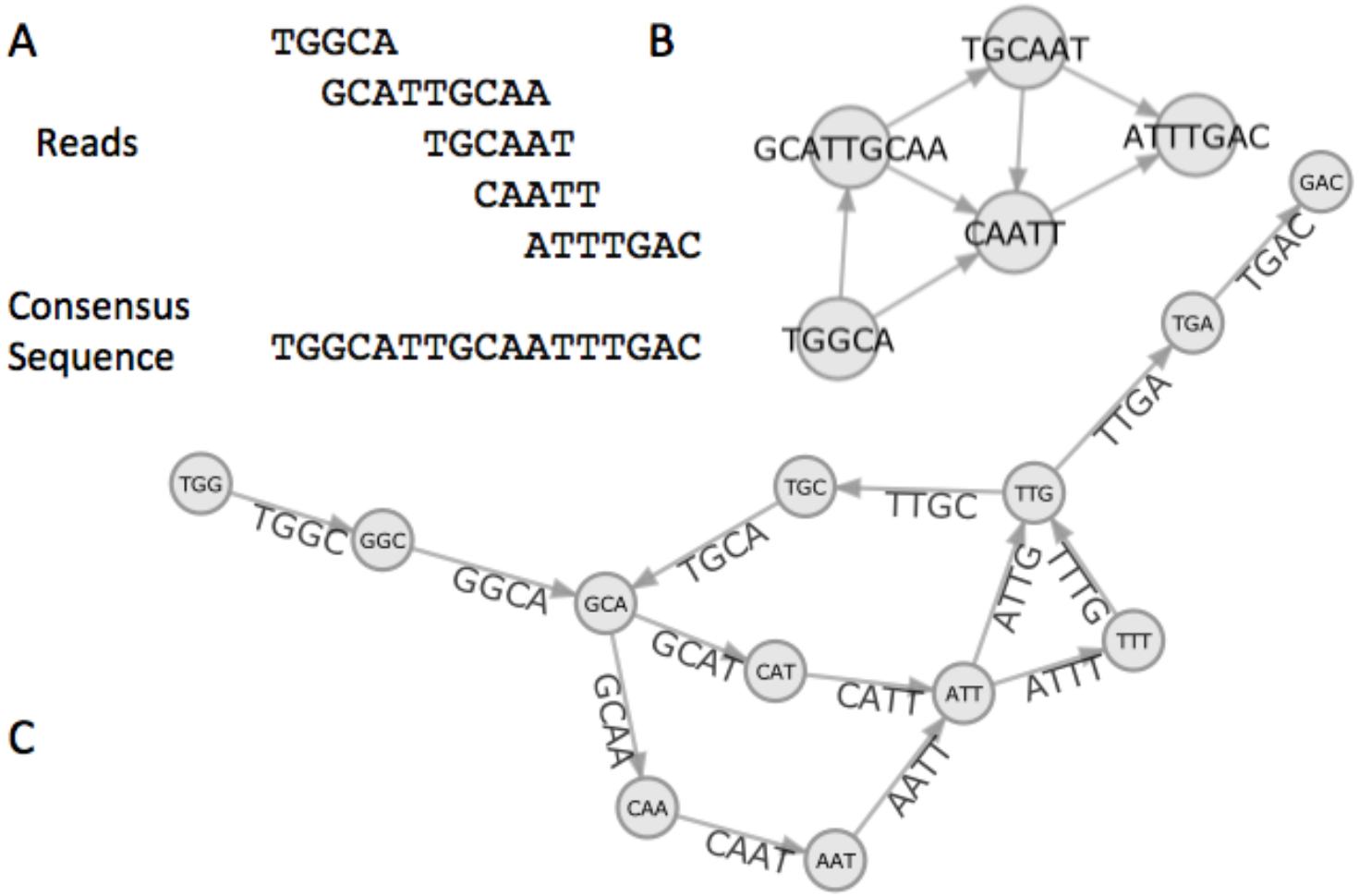
de novo assembly – grafové m.

- OLC (overlap layout concensus)
  - graf je vygenerován s použitím čtení a jejich překryvů
  - vrcholy (uzly) jsou jednotlivá čtení
  - hrany mezi vrcholy reprezentují překryv čtení
  - hledá se Hamiltonovská cesta (každý uzel je navštíven jednou)
  - vhodné především pro delší čtení
  - metagenomická data jsou velmi komplexní, což znesnadňuje výpočet, který je u OLC náročnější
  - nástoje: Edena, Newbler (454 data), SMRT Analysis (PacBio data)

de novo assembly – grafové m.

- de Bruijn graf
  - graf je vygenerován s použitím čtení a jejich překryvů přesně naopak než u OLC
  - vrcholy (uzly) jsou překryvy
  - hrany mezi vrcholy reprezentují unikátní sekvenci každého čtení
  - hledá se Eulerovský tah (každá hrana ke navštívena jednou)
  - vhodné pro krátká čtení a komplexní genomy
  - délka překryvu je jeden z předem volených parametrů → umožňuje efektivnější algoritmus pro výpočet, na druhou stranu může vynechat některé překryvy
  - některé sestavují graf pro více různých délek překryvů
  - nástroje: MetAMOS, SOAPdenovo, MetaVelvet, Meta-IDBA…

de novo assembly



hodnocení kvality assembly

- N50
  - něco jako medián, ale je daná větší váha delším kontigům (~vážený medián)
  - N50 = 100 000 bp znamená, že alespoň polovina bází v assembly je obsažena v kontizích o délce alespoň 100 000bp
- používají se i další obdobné deskriptory, další nejčastější je N90, N75
- L50
  - udává počet kontigů jejichž součet délek splňuje podmínku N50

hodnocení kvality assembly

- **příklad: Máme určit N50 a L 50 pro 10 kontigů o délkách** $\Delta l = \{62,45,56,91,70,4,16,77,69,29\} \cdot 10^3$ **bp**

1) **seřadíme kontigy sestupně: 91,77,70,69,62,56,45,29,16,4**
2) **sečteme délky 519 000 bp**
3) **pro N50 platí, že musí být větší než 519 000 · 0,5 = 259 500 bp**
4) **srovnáváme:**

91 000 < 259 500

91 000 + 77 000 = 168 000 < 259 500

91 000 + 77 000 + 70 000 = 238 000 < 259 500

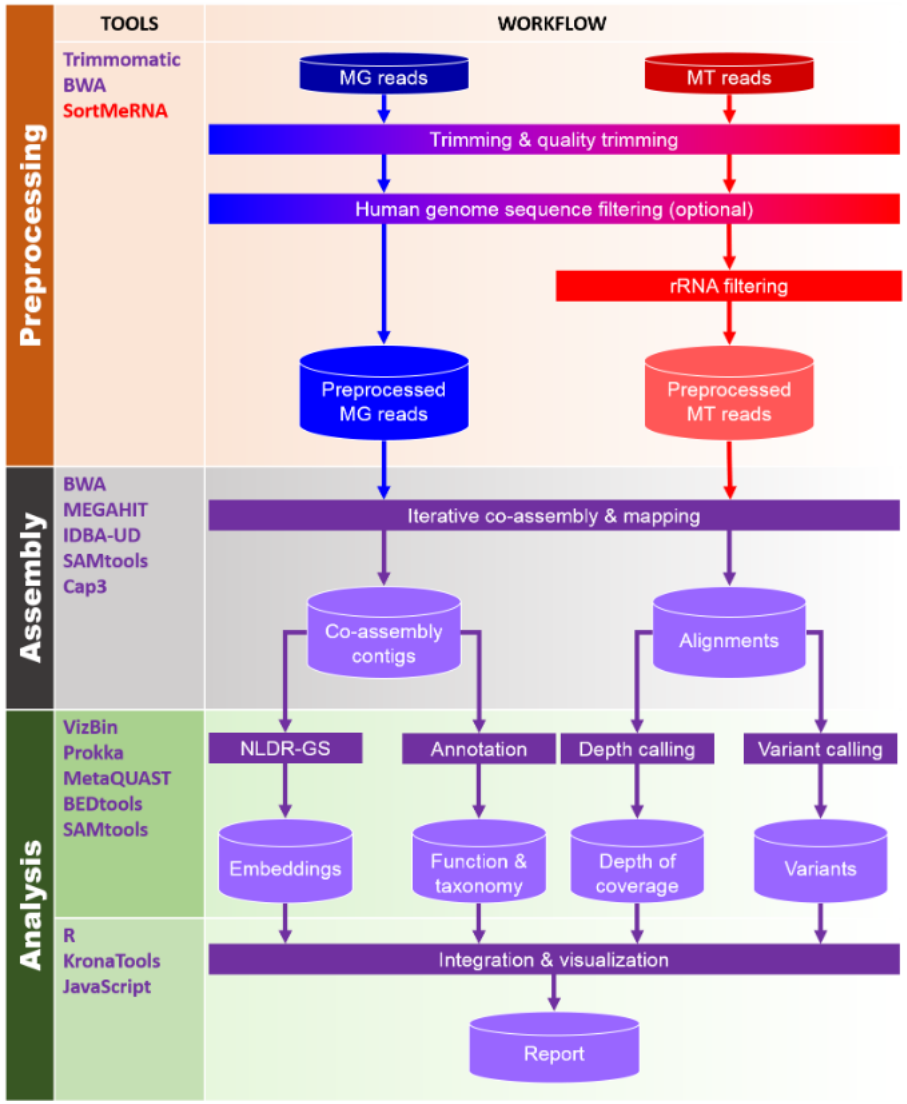
91 000 + 77 000 + 70 000 + 69 000 = 307 000 > 259 500

5) **výsledek: N50 = 69 000 bp, L50 = 4**

co-assembly

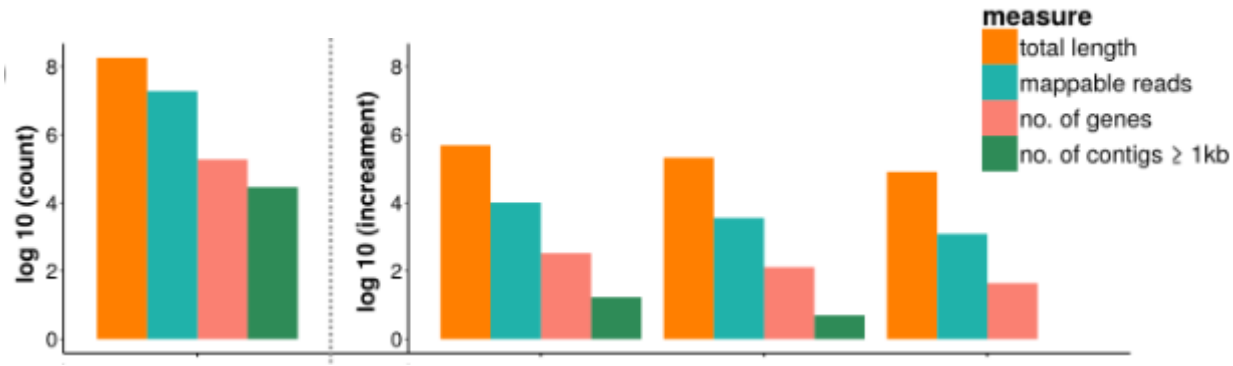
- assembly založená na kombinaci MG a MT dat
- IMP: pipeline pro reprodukovatelnou integrovanou analýzu spojených metagenomických a metatranskriptomických dat
- umožňuje jak odhad abundance populací, tak aktivity celé komunity
- reference-independent → využívá maximum dat

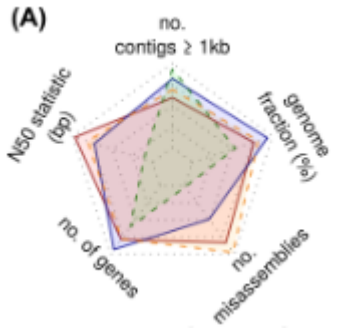
IMP


Preprocessing
Assembly
Analysis
TOOLS
WORKFLOW
Trimmomatic
BWA
SortMeRNA
MG reads
MT reads
Trimming & quality trimming
Human genome sequence filtering (optional)
rRNA filtering
Preprocessed MG reads
Preprocessed MT reads
BWA
MEGAHIT
IDBA-UD
SAMtools
Cap3
Iterative co-assembly & mapping
Co-assembly contigs
Alignments
VizBin
Prokka
MetaQUAST
BEDtools
SAMtools
NLDR-GS
Annotation
Depth calling
Variant calling
Embeddings
Function & taxonomy
Depth of coverage
Variants
R
KronaTools
JavaScript
Integration & visualization
Report

IMP

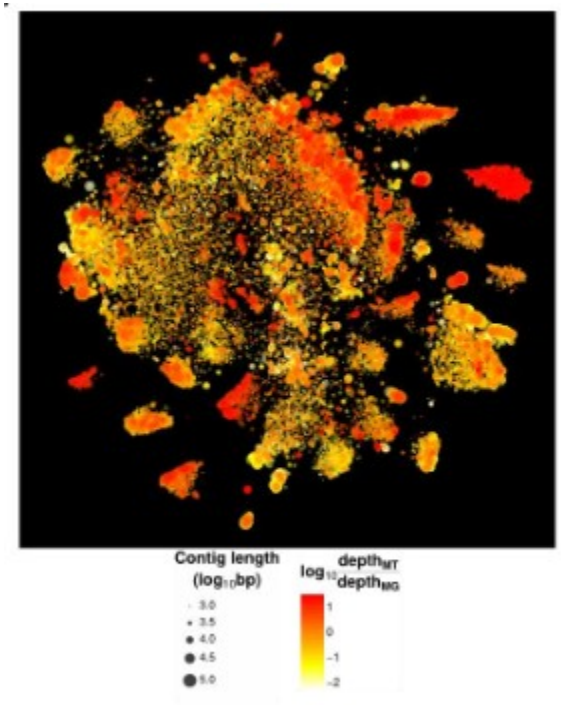
Sample	Assembly	No. of contigs (all)	Total length	N50	No. of predicted genes (unique)	Genome fraction (%)	MG mapped reads (%)	MT mapped reads (%)
SM	IMP	84052	198407791	10154	201480	65.3	97.6	90.2
	IMP-MEGAHIT	108011	208588327	8243	212669	70	98	91.39
	MG-only	74498	183252937	10902	187382	60.8	96.39	80.36
	MT-only	14723	9497250	961	7357	3.5	8.58	29.03

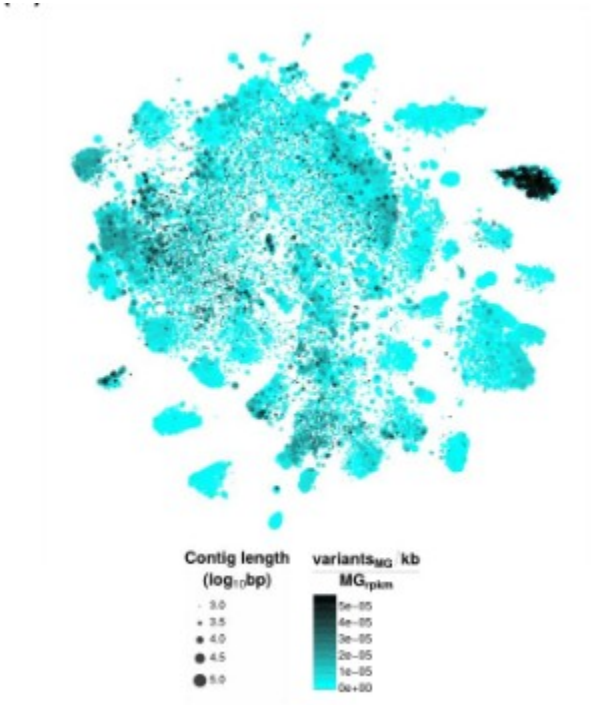


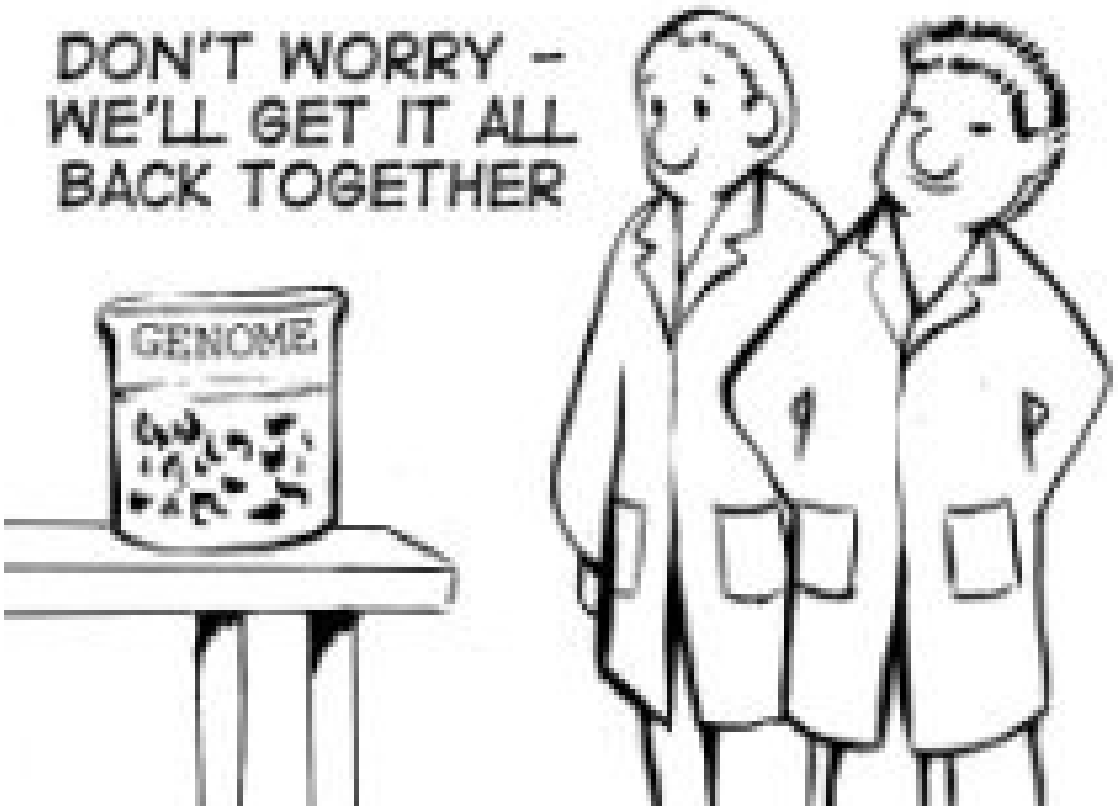


Legend: IMP IMP-MEGAHIT MetAMOS MetAMOS-IDBA_UD

VizBin


Contig length
(log10bp)
3.0
3.5
4.0
4.5
5.0
log10 depthMT
depthMG
1
0
-1
-2


Contig length
(log10bp)
3.0
3.5
4.0
4.5
5.0
variantsMG kb
MGrpkm
5e−05
4e−05
3e−05
2e−05
1e−05
0e+00


DON'T WORRY –
WE'LL GET IT ALL
BACK TOGETHER
GENOME