Minia's entry at Mosaic Strains#1 assembly challenge

Rayan Chikhi

CNRS, CRIStAL, University of Lille, Clarity Genomics

26 June 2018

Mosaic webinar

Slides are available at: `http://rayan.chikhi.name`

metagenomic assembly

- Reconstruct genomes of species, possibly even strains, from short read sequencing data of an environment

Challenges: (adapted from A. Korobeynikov presentation)

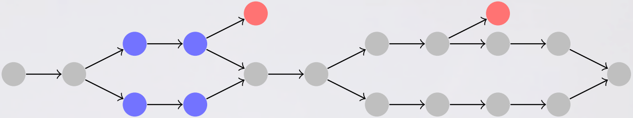
1. closely related strains
2. uneven depths, & low depths
3. inter-species repeats
4. size of datasets
5. lack of long reads

Software

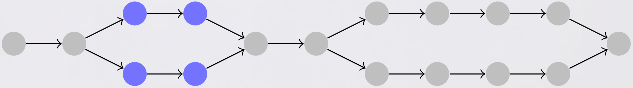
- **metaSPAdes**
- **MEGAHIT**
- **IDBA-UD**
- Minia-pipeline
- Ray-meta
- SOAPdenovo2
- metaVelvet/-SL
- Omega
- InteMAP
- Meraga
- Velour
- A*

How a metagenome assembler generally works

1) de Bruijn **graph** construction



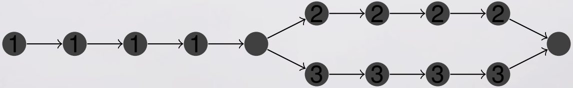
2) Likely sequencing errors are removed.



3) Variations (e.g. SNPs, similar repetitions) are removed.

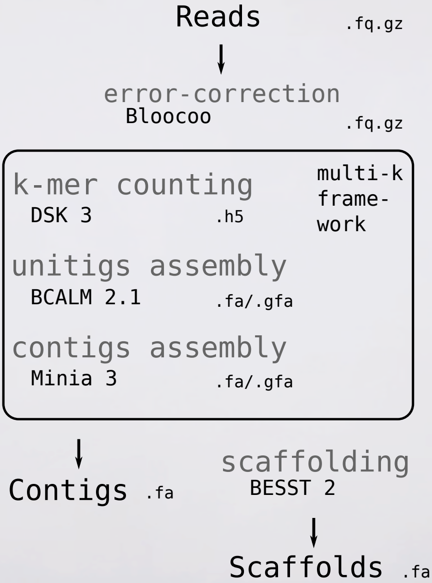
→ **Skipped in Strains #1**

4) **Simple paths** (i.e. contigs) are returned.



5) Extra steps: repeat-resolving, scaffolding (**not done in Minia**)

the Minia pipeline



Reads .fq.gz

error-correction
Bloocoo .fq.gz

k-mer counting
DSK 3 .h5

multi-k frame-work

unitigs assembly
BCALM 2.1 .fa/.gfa

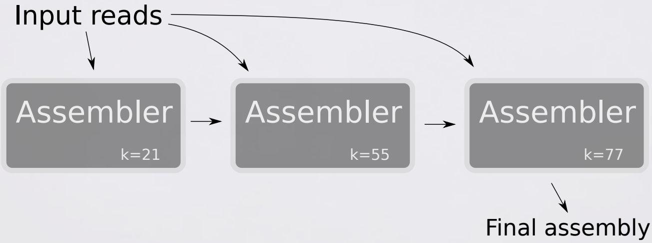
contigs assembly
Minia 3 .fa/.gfa

Contigs .fa

scaffolding
BESST 2

Scaffolds .fa

Multi-k


Input reads
Assembler
k=21
Assembler
k=55
Assembler
k=77
Final assembly

Aftermath

Regular **multi-k** assembly with **conservative** simplifications → high genome fraction, limited number of misassemblies

No bubble removal → larger-than-expected assembly

Forced QUAST to consider **all contigs** by N-padding them → higher reported Genome Fraction than competitors

Low training dataset

Method	N50	Genome Fraction	# misassemblies
Unitigs (BCALM)	106 Kbp	99.6%	2
Minia-pipeline only tip clipping	195 Kbp	99.4 %	8
Minia-pipeline with all simplifications	235 Kbp	99.5 %	14

Remarks:

- QUAST, contigs ≥ 500 bp
- Multi-k up to *k* = 241
- No scaffolding
- merged PE reads

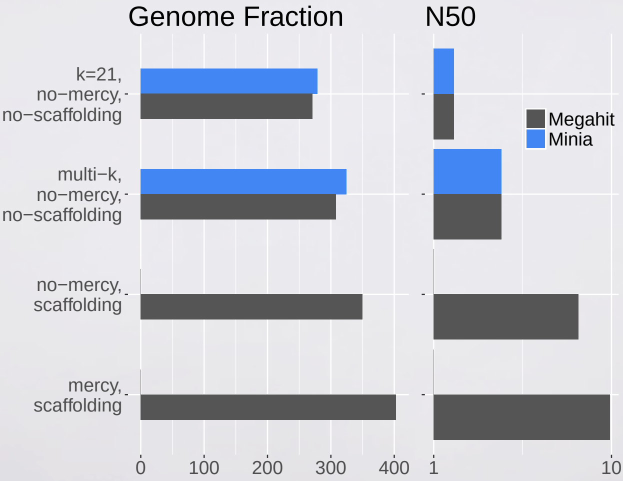
High training dataset

Method	**N50**	**Genome Fraction**	**# misassemblies**
Unitigs (BCALM)	0.5 Kbp	95.3%	23
Minia-pipeline only tip clipping	1.3 Kbp	90.8%	286
Minia-pipeline with all simplifications	7.1 Kbp	84.1%	1998

Remarks:

- QUAST, contigs ≥ 500 bp (w/ 500 bp N-padding)
- Multi-k up to $k = 91$
- No scaffolding
- Merging PE reads didn’t always improve Genome Fraction
- Performance: ≈ 5 GB & ≈ 5 hours per Gbp in assembly.

Minia-pipeline matches MEGAHIT, up to mercy *k*-mers and scaffolding



CAMI, medium dataset, PE data only

Conclusion

- In strains reconstruction, there seems to be a trade-off between contiguity, and genome fraction/misassemblies. Raises questions on how to rank assemblies.

Minia references:

- `https://github.com/GATB/minia-pipeline`
- *Critical Assessment of Metagenome Interpretation - A Benchmark of Metagenomics Software*, 2017
- *On the representation of de Bruijn graphs*, 2014
- *Space-efficient and exact de Bruijn graph representation based on a Bloom filter*, 2012

Slides are available at: `http://rayan.chikhi.name`

Acknowledgments: Sergey Nurk, Chris Quince, Aaron Darling, Guillaume Rizk, Dominique Lavenier, Claire Lemaitre, Pierre Peterlongo, Charles Deltel, Antoine Limasset, Paul Medvedev