Scaffolding and optical map assembly: testing and adaptation of existing software

Svyatoslav Sidorov, St. Petersburg University of the Russian Academy of Sciences
Scientific adviser: **Aleksey Komissarov**, postdoc at Theodosius Dobzhansky Center for Genome Bioinformatics

Outline

1. Problem statement.
2. Scaffolders.
3. Optical mapping (in brief).
4. Results.
5. Conclusion.

Outline

1. Problem statement.
2. Scaffolders.
3. Optical mapping (in brief).
4. Results.
5. Conclusion.

Problem statement: 1st version

Scaffolding of mammalian genome assemblies using mate-pairs and maps

Outline

1. Problem definition.
2. Scaffolders.
3. Optical mapping (in brief).
4. Result 1: OMA tool.
5. Result 2: *Aurelia aurita* optical map assembly.
6. Result 3: *Mus musculus* genome-wide scaffolding.
7. Conclusion.

Scaffolders

Scaffolders using mate-pairs:

1. SSPACE.
2. SOPRA.
3. SOAPdenovo2.
4. ABySS.
5. SGA.

} Scaffolding modules in these assemblers

etc.

Scaffolders

Scaffolders using mate-pairs:

1. SSPACE.
2. SOPRA.
3. SOAPdenovo2.
4. ABySS.
5. SGA.

} Scaffolding modules in these assemblers

etc.

Scaffolders using PacBio reads:

1. AHA.
2. SSPACE-LongRead.

Scaffolders

Scaffolders using mate-pairs:

1. SSPACE.
2. SOPRA.
3. SOAPdenovo2.
4. ABySS.
5. SGA.

} Scaffolding modules in these assemblers

etc.

Scaffolders using PacBio reads:

1. AHA.
2. SSPACE-LongRead.

Scaffolders using optical maps:

1. SOMA2.
2. OMACC.

Scaffolders

Scaffolders using mate-pairs:

1. SSPACE.
2. SOPRA.
3. SOAPdenovo2.
4. ABySS.
5. SGA.

} Scaffolding modules in these assemblers

etc.

Scaffolders using PacBio reads:

1. AHA.
2. SSPACE-LongRead.

Scaffolders using optical maps:

1. SOMA2.
2. OMACC.

No scaffolders using linkage groups and radiation hybrid maps.

Problem statement: 1st version

Scaffolding of mammalian genome assemblies using mate-pairs and maps

Problem statement: 2nd version

Scaffolding of mammalian genome assemblies using mate-pairs and maps

Scaffolding of genome assemblies using maps

Problem

There are 626’030 short single-molecule optical maps of *Aurelia aurita* (optical reads) that we need to assemble into longer maps (contigs) before scaffolding *A. aurita* sequences using this optical mapping data.

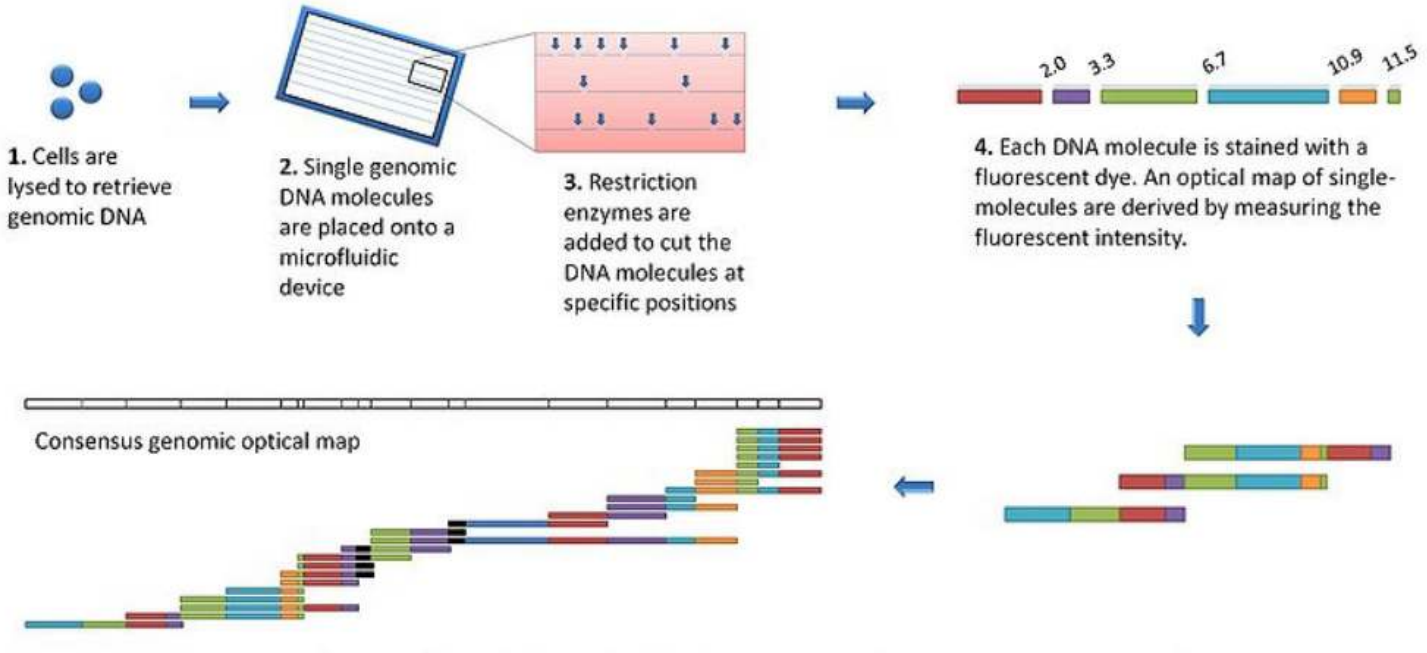
Genome size: 300 Mb

Aurelia aurita
http://bioexpedition.com/moon-jellyfish/

Outline

1. Problem statement.
2. Scaffolders.
3. Optical mapping (in brief).
4. Results.
5. Conclusion.

Optical mapping (in brief)



Problem statement: 2nd version

Scaffolding of mammalian genome assemblies using mate-pairs and maps

Scaffolding of genome assemblies using maps

Problem statement: 3rd version

Scaffolding of mammalian genome assemblies using mate-pairs and maps

Scaffolding of genome assemblies using maps

Scaffolding and optical map assembly: testing and adaptation of existing software

Outline

1. Problem statement.
2. Scaffolders.
3. Optical mapping (in brief).
4. Results.
5. Conclusion.

Result 1: OMA tool

OMA stands for **O**ptical **M**ap **A**ssembler.

Result 1: OMA tool

OMA stands for **O**ptical **M**ap **A**ssembler.

It's based on alignment and assembly software written by Dr. Anton Valouev (PhD at University of Southern California) for the following papers:

1. **Valouev A, Li L, Liu YC, Schwartz DC, Yang Y, Zhang Y, Waterman MS**, "Alignment of optical maps", Journal of Computational Biology, Mar 2006.
2. **Valouev A, Zhang Y, Schwartz DC, Waterman MS**, "Refinement of optical map assemblies", Bioinformatics, May 2006.
3. **Valouev A, Schwartz DC, Zhou S, Waterman MS**, "An algorithm for assembly of ordered restriction maps from single DNA molecules", PNAS, Oct 2006.

Result 1: OMA tool

Software for papers → Tool

Result 1: OMA tool

Software for papers → Tool

1. Minor changes of source code to get it compile.
2. Python wrappers for starting alignment and assembly or assembly only.
3. Parallelization of alignment step.
4. Python scripts for generating single-molecule optical maps *in silico* from complete genome for OMA testing.
5. Makefile with testing feature.
6. *Yersinia pestis* test single-molecule dataset.
7. Some additional scripts (including statistics calculating script).
8. README.

Result 2: *Aurelia aurita* optical map assembly

Single-molecule maps:

There are **626030** optical maps.
Min map length: 765.
Max map length: **978939**.
Avg map length: 66612.
Median map length: **46664**.

Min intervals quantity: 2.
Max intervals quantity: 131.
Avg intervals quantity: 6.
Median intervals quantity: **5**.

Contigs:

There are **40** optical maps.
Min map length: 42466.
Max map length: **2085679**.
Avg map length: 151817.
Median map length: **95372**.
Total length of all maps: 6072695.
Min intervals quantity: 9.
Max intervals quantity: 455.
Avg intervals quantity: 24.
Median intervals quantity: **14**.
Total intervals quantity of all maps: 976.

Result 3: *Mus musculus* genome-wide scaffolding

Scaffolds evaluation in QUAST without reference:

Statistics without reference	scaff_.final.scaffolds
# contigs	37 530
# contigs (>= 0 bp)	37 927
# contigs (>= 1000 bp)	36 798
Largest contig	2 106 024
Total length	2 439 105 199
Total length (>= 0 bp)	2 439 242 989
Total length (>= 1000 bp)	2 438 522 122
N50	256 406
N75	129 046
L50	2697
L75	6035
GC (%)	41.71
Mismatches	
# N's	160 390 580
# N's per 100 kbp	6575.8

Outline

1. Problem statement.
2. Scaffolders.
3. Optical mapping (in brief).
4. Results.
5. Conclusion.

Conclusion

1. Dr. A. Valouev's software for optical maps assembly can be installed, tested and used as a simple tool (OMA, Optical Map Assembler).
2. Single-molecule maps of *A. aurita* were assembled into contigs.
3. Scripts for single-molecule datasets generation were developed.
4. SSPACE was tested of *M. musculus* contigs.

Thank you!

Adapted from:
Zhou S. ***et al.*** Single-Molecule Approach to Bacterial Genomic Comparisons via Optical Mapping // Journal of Bacteriology, November 2004, vol. 186, no. 22, 7773 – 7782.