UNI FR
UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

INTRODUCTION TO *DE NOVO* ASSEMBLY OF GENOMES AND TRANSCRIPTOMES (PART2)

Laurent Falquet
MA/MER @ UniFr
Group Leader @ SIB

SIB
Swiss Institute of Bioinformatics

Overview of a sequencing project

Select DNA and sequencing method
Quality Control
if necessary Trimming/Filtering
De novo assembly with cleaned data
1st step: Assembly
consists mainly in building contigs from short reads
2nd step: Scaffolding (optional)
where contigs are ordered and oriented
3rd step: Finishing or Closing (optional)
where gaps are closed
Genome annotation (draft or final)
Submission to databases & publication (not discussed here)

Scaffolding

Step by which non-overlapping contigs are ordered and oriented along a chromosome

Linking information source

Overlaps

Paired-Ends & Mate Pairs

Reference based links

Physical markers

Gene/operon synteny

Other methods…

Existing methods for scaffolding and closing

More Next Generation Sequencing

- **Illumina protocol (linked reads or synthetic long reads)**
- **PacBio or Nanopore long reads**
- **HiC scaffolding**

Other methods (without sequencing)

- **Optical maps**

Illumina protocol (linked reads or synthetic long reads)

LooSeq synthetic long reads (up to 7Kbp long reads)

Illumina Infinity long read (up to 10Kbp long reads)

10X Genomics Linked reads (discountinued)

But these methods suffers from the same flaws as standard Illumina for long repeats if located on the same DNA fragment.
In addition the size of synthetic long reads is rather limited.

Existing methods for scaffolding and closing

More Next Generation Sequencing

- Illumina protocol (linked reads or synthetic long reads)
- **PacBio or Nanopore long reads**
- **HiC scaffolding**

Other methods (without sequencing)

- **Optical maps**

Long reads sequencing

PacBio

~ 20Kbp to 200Kbp

Oxford Nanopore

Up to 1Mbp

Possible strategies for assembling long and/or short reads

Vasudevan et al, https://doi.org/10.1016/j.ygeno.2019.04.006

Assembly with long reads only

Flye

Kolmogorov, M., Yuan, J., Lin, Y. et al. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 37, 540–546 (2019). https://doi.org/10.1038/s41587-019-0072-8

Canu

Koren S., Walenz B. P., Berlin K., Miller J. R., Bergman N. H., Phillippy A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27, 722–736. 10.1101/gr.215087.116,

Miniasm

Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics 32, 2103–2110 (2016).

FALCON/HGAP

Chin et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nature Methods. 13(12), 1050.

Flye algorithm to improve assembly

Kolmogorov, M., Yuan, J., Lin, Y. et al. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 37, 540–546 (2019). https://doi.org/10.1038/s41587-019-0072-8

Problem: error rate is high!!

Match	83.7%
Mismatch	1.4%
Insertions	11.5%
Deletions	3.4%

Long reads only

No read correction: very fast, error-prone

Applied by: **miniasm, smartdenovo, HINGE, wtdbg2 ...**

http://schatzlab.cshl.edu/presentations/2012-01-17.PAG.SMRTassembly.pdf

Long reads only + polishing (2 options pre and/or post assembly)

Pre-assembly correction: reduces errors, helps in assembly, slow

Post-assembly correction (polishing): removes remaining errors, slow

Applied by: **Canu, HGAP3(Quiver), HGAP4(Arrow), FALCON, MECAT, Racon, Nanopolish,…**

Hierarchical Genome Assembly Process (HGAP) with PacBio

Hybrid assembly (Illumina short reads first)

Usually performed to scaffold the short reads contigs using a low coverage of long reads

Applied by: **HybridSPAdes, UniCycler, SSPACE-LongReads, LINKS**

http://schatzlab.cshl.edu/presentations/2012-01-17.PAG.SMRTassembly.pdf

Hybrid assembly (long reads first)

Usually performed to correct the long reads contigs using a high coverage of Illumina short reads (polishing)

Applied by: **PILON, Nanopolish, Racon, LSC, PacBioToCA, LoRDEC, proovread, CoLoRMap, HECIL**

http://schatzlab.cshl.edu/presentations/2012-01-17.PAG.SMRTassembly.pdf

Error correction improves gene prediction

PacBio and Oxford Nanopore reads mainly contain **INDELs errors**, whereas Illumina reads mainly contains **substitution errors**.

As a result a simple 1 bp indel can have dramatic consequences on the predicted genes (-> frameshifts).

It is thus critical to polish the assemblies to remove as much as possible of these indels.

Watson, M., Warr, A. Errors in long-read assemblies can critically affect protein prediction. Nat Biotechnol 37, 124–126 (2019). https://doi.org/10.1038/s41587-018-0004-z

PacBio HiFi reads avoids error correction!

No read correction: very fast, little errors

Applied by: **hifiasm, hicanu, Flye, FALCON**

http://schatzlab.cshl.edu/presentations/2012-01-17.PAG.SMRTassembly.pdf

Existing methods for scaffolding and closing

More Next Generation Sequencing

- **Illumina protocol (linked reads or synthetic long reads)**
- **PacBio or Nanopore long reads**
- **HiC scaffolding**

Other methods (without sequencing)

- **Optical maps**

HiC Scaffolding: a promising technique

Takes advantage of proximity crosslinking

Then sequence junctions with Illumina, the reads are mapped onto contigs

Connections are identified and ordered into chromosome-like scaffolds

Example with *Aedes aegypti*

Software: instaGRAAL, SALSA2, Juicer
Visualisation: JuiceBox

Existing methods for scaffolding and closing

More Next Generation Sequencing

- **Illumina protocol (linked reads or synthetic long reads)**
- **PacBio or MinION long reads**
- **HiC scaffolding**

Other methods (without sequencing)

- **Optical maps**

Bionano Optical Mapping (no sequencing!)

1) Extraction of ultra HMW DNA

2) Label DNA at specific sequence motifs

3) Linearise DNA in NanoChannel array and image

4) Combine contigs based on optical image

Bionano Genomics Systems

discontinued

Irys system

Approx. 25-50 Gb per flowcell

2 flowcells/1 chip

Only compatible with the NLRS chemistry

Saphyr system

320-480 Gb per flowcell for human samples

Less data for other species

2 flowcells/1 chip

Compatible with NLRS and DLS chemistry

Supported Apps

Large Structural variations up to 1Mb

Genome scaffolding in combination with NGS

Reduced representation optical methylation (R2OM2)

Assembly assessment

QUAST – incorporates several tools and databases

BUSCO – commonly used to assess an assembly's ability to predict genes.
Some use BUSCO results to train gene predictors for genome annotation step

MAUVE – for small genomes

Bandage – assembly graph visualization tool

Example output from QUAST

Report Example

E. coli single-cell assemblies

Aligned to "e.coli_reference" | 4 639 675 bp | 50.79 % G+C
4324 genes | 884 operons

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Worst Median Best ☑ Show heatmap

Statistics without reference	E+V-SC	IDBA-UD	SPAdes-2.4	Velvet-SC
# contigs	344	250	277	519
Largest contig	132 865	224 018	269 177	121 367
Total length	4 540 286	4 791 744	4 877 521	4 526 656
N50	33 616	96 947	106 927	20 445
Misassemblies				
# misassemblies	2	9	2	2
Misassembled contigs length	23 485	66 335	26 551	22 359
Mismatches				
# mismatches per 100 kbp	2.26	3.65	5.06	1.77
# indels per 100 kbp	0.7	0.2	0.7	0.92
# N's per 100 kbp	0	0	4.86	0
Genome statistics				
Genome fraction (%)	91.727	94.943	95.759	91.43
Duplication ratio	1.001	1.001	1.004	1.002
# genes	3767 + 160 part	4026 + 80 part	4046 + 102 part	3630 + 288 part
# operons	723 + 87 part	802 + 40 part	809 + 48 part	650 + 158 part
NGA50	32 051	96 947	110 539	19 791
Predicted genes				
# predicted genes (unique)	4258	4394	4417	4331
# predicted genes (>= 0 bp)	4258	4394	4490	4331
# predicted genes (>= 300 bp)	3643	3736	3784	3666
# predicted genes (>= 1500 bp)	524	559	559	515
# predicted genes (>= 3000 bp)	44	49	48	39

Extended report

Example output from QUAST

BUSCO: validate assembly by gene content

https://busco.ezlab.org

BUSCO searches for unique genes

BUSCO attempts to provide a quantitative assessment of the completeness in terms of expected gene content of a genome assembly, transcriptome, or annotated gene set. The results are simplified into categories of Complete and single-copy, Complete and duplicated, Fragmented, or Missing BUSCOs.

However, a high busco score in multiple domains might help you identify possible contaminations.

Mosè Manni et al. Molecular Biology and Evolution, Volume 38, Issue 10, October 2021, Pages 4647–4654, https://doi.org/10.1093/molbev/msab199

Bandage examples

SPAdes assembly

UniCycler assembly

Bandage examples: a metagenomics assembly

Recommended pipeline of the Vertebrate Genome Project

VERTEBRATE GENOMES PROJECT

A PROJECT OF THE G10K CONSORTIUM

The VGP standard assembly pipeline. Adapted from: Towards complete and error-free genome assemblies of all vertebrate species.

Summary of the *de novo* assembly pipeline

Jung H, Ventura T, Chung JS, Kim W-J, Nam B-H, Kong HJ, et al. (2020) Twelve quick steps for genome assembly and annotation in the classroom. PLoS Comput Biol 16(11): e1008325. https://doi.org/10.1371/journal.pcbi.1008325

Genome *vs* Transcriptome assembly, some differences

Genome is far more repetitive than the transcriptome
de Bruijn graph of a genome <-> transcriptome
K-mer distribution of reads
Transcriptome is not a “single” molecule
Transcriptome is redundant (splice variants)
Transcriptome coverage is highly variable
On a reference, difficult to remap exon-intron junctions and all splice variants are grouped

-> We need to use different or modified tools

Algorithms for *de novo* assembly

3 main algorithm classes

Greedy and other
- **MIRA3, Ray, MaSuRCA, HGAP…**

Overlap Layout Consensus (OLC) assemblers
- **Newbler, Edena, WGS, CABOG, miniasm…**

deBruijn graph assemblers
- **ABySS, ALLPATH-LG, SOAP-denovo, SPAdes**

Eulerian path (deBruijn) for a transcriptome
- **rnaSPAdes, Trans-ABySS, SOAPdenovo-Trans, Trinity…**

Trinity a different *de novo* method

https://github.com/trinityrnaseq/trinityrnaseq/wiki

some slides borrowed from Brian Haas, Broad Institute

Trinity aggregates isolated graphs

How Trinity works

Inchworm constructs contigs by extending kmers by abundance

Chrysalis re-group Inchworm contigs

Chrysalis uses (k-1) overlaps and read support to link related Inchworm contigs

Chrysalis builds de Bruijn graphs

Butterly reconstructs isoforms

How to evaluate a transcriptome assembly?

N50 is not ideal for transcriptome

Number of transcripts and size distribution might help

Use other data

- **Compare with the closest known organism (genome or transcriptome)**
- **Blast transcripts vs UniProtKB**
- **Use reference databases like BUSCO (http://busco.ezlab.org)**
- **Use rnaQUAST (with reference genome and gene database)**

Genome vs. Transcriptome assembly

Genome	Transcriptome
Large molecules	Small molecules
Standard PE + long reads	Standard PE (long reads)
Linear sequences	Multiple solutions (alternative splicing)
Even coverage	Uneven coverage (expression levels)

In common: k-mer based approach

Summary

Current methods for short reads and long reads genome assembly are based on deBruijn graphs and OLC (overlap-layout-consensus)

Hybrid assemblies with HiC or HiFi assemblies are the most successful

Issues and limitations are mainly due to repeat sequences

Specific methods are needed for assembling transcriptomes

QC of the reads and validation of assemblies are essential steps