Developments on genome assembly

Andreea Dréau & Matthias Zytnicki

April 2nd, 2021
Table of Contents

Introduction

10X assembly

Data integration
Why genome assembly?

Keywords to get your assembly founding accepted

▶ For SARS-CoV-2: Find putative drugs, understand its mode of action.
▶ For human: cure/prevent genetic diseases.
▶ For cattle: improve production, informed selection.
▶ For plant: hydric stress resistance, pathogen resistance.

In general, crucial for:

▶ genetics studies,
▶ molecular studies.
Genome sequencing

Definition

- Extracting DNA.

GATTACA
GATTACA
GATTACA

Several flavors of sequencers
- Illumina: correct, cheap, high-throughput, short (150).
- ONT: noisy, very long (15k-100k).
- HiFi: correct, long (15k).
Genome sequencing

Definition

- Extracting DNA.
- Possibly cleave it.

Several flavors of sequencers

- Illumina: correct, cheap, high-throughput, short (150).
- ONT: noisy, very long (15k-100k).
- HiFi: correct, long (15k).

Rem.: The machine can be unfaithful.

GATTACA
GATTACA
GATTACA
ATT
ACA
GAT
TAC
TTA
Genome sequencing

Definition

- Extracting DNA.
- Possibly cleave it.
- Run it through a machine, which gives lots of A, C, G, T.
Genome sequencing

Definition

- Extracting DNA.
- Possibly cleave it.
- Run it through a machine, which gives lots of A, C, G, T.

Rem. The machine can be unfaithful.

Several flavors of sequencers

- Illumina: correct, cheap, high-throughput, short (150).
- ONT: noisy, very long (15k-100k).
- HiFi: correct, long (15k).
Genome sequencing

Definition

- Extracting DNA.
- Possibly cleave it.
- Run it through a machine, which gives lots of A, C, G, T.

Rem. The machine can be unfaithful.

Several flavors of sequencers

- **Illumina**: correct, cheap, high-throughput, short (150).
- **ONT**: noisy, very long (15k-100k).
- **HiFi**: correct, long (15k).
Definition
Reads can be assembled into contigs if they merge.

Problem: the repetitions
- A targeted coverage can be 60X.
- If the puzzle were perfect, we could assemble the genome.
- But repetitions make the assembly impossible.
Why assembly fails?

- Consider the genome:
  \[ \text{ACGATTTTGACATTTTCCGGTTTTAAGG} \]

- Cut it into 4-letters long reads:
  ACGA, CGAT, GATT...

- Shuffle them.

- You know that:
  ACGA, GACA, CCGG are before TTTT
  GACA, CCGG, AAGG are after TTTT

- You can draw the following graph, but cannot solve it!
  \[
  \begin{array}{c}
  \text{GACA} \\
  \uparrow \\
  \text{ACGA} \quad \rightarrow \quad \text{TTTT} \quad \rightarrow \quad \text{AAGG} \\
  \uparrow \\
  \text{CCGG}
  \end{array}
  \]
Genome scaffolding

Workaround

- Fragmented genomes are suboptimal for many analyses.
- If we could stitch the contigs in the right order, it would help.
- It is OK if the content of the glued positions is unknown, if the distance between the contigs is approximate.
- We can use long-range interactions: information which indicate that contig $X$ is “close to” contig $Y$.

GTCAC---?---GCTAGCA
Interaction matrix
Long-range data: 10X Genomics

Linked-Reads

10x Barcoded Gel Beads

HMW gDNA Enzyme

Oil

1-5 HMW gDNA Molecules per GEM

Collect

Isothermal Incubation

Pool Remove Oil

Genome GEMs

10x Barcoded Fragments

10x Barcoded Fragments

INRAE

Developments on genome assembly

04/02/2021 / Andreea Dréau & Matthias Zytnicki
### Long-range data: Hi-C

**Figure 1. Dovetail™ Hi-C proximity ligation workflow**

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Dovetail™ Hi-C libraries start with endogenous mammalian chromatin.</td>
</tr>
<tr>
<td>B</td>
<td>Crosslinking (red lines) the chromatin creates a stabilized nucleosome (blue circles) scaffold.</td>
</tr>
<tr>
<td>C</td>
<td>Restriction endonucleases digest the cross-linked chromatin.</td>
</tr>
<tr>
<td>D</td>
<td>Biotin (green dots) mark digested DNA ends (black lines).</td>
</tr>
<tr>
<td>E</td>
<td>Proximity ligation creates chimeric molecules (ex. 1 and 2).</td>
</tr>
<tr>
<td>F</td>
<td>The crosslinks are reversed.</td>
</tr>
<tr>
<td>G</td>
<td>DNA is purified, and enriched for ligation-containing chimeric molecules. An Illumina® library is prepared and sequenced.</td>
</tr>
</tbody>
</table>
Long-range data: BioNano

Customer Sample

- Blood
- Cells
- Tissue
- Microbes

Isolate High Molecular Weight DNA

Label Specific Sequences Across the Entire Genome

Transfer Labeled DNA into Cartridge for Scanning

Load, Linearize & Image Labeled DNA in Repeated Cycling to Scan Whole Genome

High-Throughput, High-Resolution Imaging Gives Contiguous Reads up to Mb Length

Algorithms Convert Images into Molecules

Assembly Algorithms Align Molecules de novo for Constructing Consensus Genome Maps

Cross-Mapping Across Multiple Samples or to a Reference

INRAE
Developments on genome assembly
04/02/2021 / Andreea Dréau & Matthias Zytnicki
Table of Contents

Introduction

10X assembly

Data integration
Contig assembly graphs are messy

- node: A,C,G,T sequence
- arc: if significant overlap between nodes
Contig assembly graphs are messy

- node: A,C,G,T sequence
- arc: if significant overlap between nodes
Scaffolding methods using Hi-C reads

- align reads on contigs
  
<table>
<thead>
<tr>
<th>Read</th>
<th>Contig</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATTAGTTACTGATATG</td>
<td></td>
</tr>
<tr>
<td>....ACTACTAGATTACGGATCATGCCTACGT....</td>
<td></td>
</tr>
</tbody>
</table>

- split contigs based on coverage drop

- connect contigs using contact information
Hi-C heat map: contact information for contigs
Hi-C heat map (zoom): coverage is not uniform
Split contigs with linked reads

Developments on genome assembly

04/02/2021 / Andreea Dréau & Matthias Zytnicki
Split contigs with linked reads

- Align linked-reads on contigs
- Identify molecules (barcode, beginning and ending position, number of reads)
- Compute molecule profiles per interval (10kb)
  - Number of starting molecules
  - Number of ending molecules
  - Molecule coverage
  - Mean read density/molecule
  - Mean molecule length
- Identify outliers intervals and split contigs
- Re-connect contigs with Hi-C scaffolding methods
Scaffold split contigs with Hi-C reads

Bovine male

- Scaffolds from contigs
- Scaffolds from split contigs
Challenges

- False positive splits
- Contig splits too short for Hi-C scaffolding
- Inversed contigs

Solution: Scaffold first with linked reads
1. For each contig extremity create a representative barcode set

<table>
<thead>
<tr>
<th>Contig</th>
<th>10x molecules</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>W</td>
</tr>
<tr>
<td>W</td>
<td>W</td>
</tr>
</tbody>
</table>

2. Connect two contigs if their representative barcode sets share enough barcodes

3. Scaffold contigs from unbranched paths
Scaffolding graph
Scaffolding graph
Scaffolding graph
Scaffolding graph
Challenges

- Definition of "enough" shared barcodes for very short contigs
- Read alignment in contig extremity containing repeat sequences
- Risk of introducing new connection errors
Table of Contents

Introduction

10X assembly

Data integration
Aim

- Current methods proceed step-wise: short range first, long range last.
- However, some choices made at step $n$ are not consistent with data at step $n + 1$.
- It would be best to integrate the data.
- Including long reads.
Contact map

- We suppose that we “merge” all contigs.
- The contigs are chunked into *bins* of given since (such as 1k).
- A contact map is a symmetric matrix, where each cell \((i, j)\) stores the number of times long-ranged data saw bins \(i\) and \(j\) together.
Main scaffolding steps

A split

- A split occurs when the contig step joined 2 sequences erroneously.
- They can be detected when counts are low around the diagonal.

A join

- A join occurs when the contig step failed to join 2 sequences.
- They can be detected when counts are high in a corner.

It is thus crucial to clean noisy data!
Cleaning

Discarding low counts row/cols

▶ Poisson, negative binomial, logistic regressions does not work.
▶ Just remove all those lines with count less than mean $- 3$ standard deviations.

Downsizing high counts row/cols

▶ Matrix balancing on a such big matrix does not work (+ some matrices are limited to the diagonal).
▶ Decrease the counts on the lines with count greater than mean $- 3$ standard deviations to the average count.
A key parameter: the “molecule” size

Definition
The max range/size/distance where you expect to see an interaction. Call it $m_s$.

- When looking for splits: the thickness of the diagonal.
- When looking for joins: the size of the corner.
- The name comes from 10X.

The parameter can be estimated from the raw data (for long reads and 10X) or the matrix (for Hi-C).
Splitting: finding hollow triangles

Principle

▶ In a split at bin $i$, there should be no interaction between bins $[0, i - 1]$ and $[i + 1, +\infty]$.

▶ Since there is no interaction after $m_s$, you just look for triangular “holes”.

▶ Splits are thus triangles such that the sum of the counts is low.

▶ Triangles with “too many” missing values are discarded.
Splitting: finding hollow triangles

Accounting for the diagonal strength

- Most of the counts are on the diagonal.
- Summing is more or less 1 point: the diagonal.
- Each point $c_{ij}$ is thus transformed as: $\log_2 \frac{c_{ij}}{\text{mean}(c_{i'j':i'-j'=i-j})}$. 
Finding a suitable threshold

- The distribution of triangles should be centered around 0.
- Positive triangles are noise, and supposed to be the background distribution.
- A kind of p-value can be given to negative triangles, comparing the negative with the positive distributions.
Splitting: merging results

- Splits from one dataset are compared with other datasets.
- If the corresponding triangles are positive, splits are discarded.
Joining: finding full triangles

Principle

- Count distribution of one corners are compared to the count distribution of the other corners, the “interior”.
- The min p-value is kept.
Joining: merging results

- Joins are merged, and sorted by p-value.
- Contigs are joined greedily.
Conclusion

▶ Benchmarking.
▶ Benchmarking.
▶ Benchmarking.

Developments on genome assembly

04/02/2021 / Andreea Dréau & Matthias Zytnicki