

# ヒトの Telomere-to-telomere にAssembleだけでトライしてみました

「Telomere-to-telomere consortium HG002 “Q100” project」  
で使用されたデータを使って  
2種類のAssemble pipe-line(hifiasm,verkko)で  
アセンブルを実施してみました。

株式会社メイズ

2024年5月22日

# 参考文献と利用したデータ

- 参考文献
  - Telomere-to-Telomere consortium primates project  
<https://github.com/marbl/Primates?tab=readme-ov-file>
  - Telomere-to-telomere consortium HG002 "Q100" project  
<https://github.com/marbl/HG002>
  - Telomere-to-telomere assembly of diploid chromosomes with Verkko  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10427740>
  - VGP assembly pipeline: Step by Step  
[https://training.galaxyproject.org/training-material/topics/assembly/tutorials/vgp\\_genome\\_assembly/tutorial.html#purging-the-primary-and-alternate-assemblies](https://training.galaxyproject.org/training-material/topics/assembly/tutorials/vgp_genome_assembly/tutorial.html#purging-the-primary-and-alternate-assemblies)
  - Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7961889>
  - Haplotype-resolved assembly of diploid genomes without parental data  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9464699>
- 使用Data HG002(HG003,HG004) of human-pangenomics
  - [https://github.com/marbl/HG002/blob/main/Sequencing\\_data.md](https://github.com/marbl/HG002/blob/main/Sequencing_data.md)
  - <https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=T2T/scratch/HG002/sequencing/>
    - PacBio Hifi read  
/hifirevio/  
m84005\_220827\_014912\_s1.hifi\_reads.fastq.gz
    - ONT Ultra long read  
/ont/03\_08\_22\_R941\_HG002\_rebasecalling-guppy-6.3.7/  
03\_08\_22\_R941\_HG002\_1.fq.gz  
03\_08\_22\_R941\_HG002\_2.fq.gz  
03\_08\_22\_R941\_HG002\_3.fq.gz
    - Trio Insert size 100
      - HG002 (Child)  
/element/trio/HG002/ins1000/  
ASHG-C063\_R1.fastq.gz  
ASHG-C063\_R2.fastq.gz
      - HG003 (paternal)  
/element/trio/HG003/ins1000/  
GAT-LI-C038\_R1.fastq.gz  
GAT-LI-C038\_R2.fastq.gz
      - HG004 (maternal)  
/element/trio/HG004/ins1000/  
GAT-LI-C042\_R1.fastq.gz  
GAT-LI-C042\_R2.fastq.gz
  - Hi-C  
[https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC\\_PLUS/HG002/raw\\_data/hic/downsampled/](https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC_PLUS/HG002/raw_data/hic/downsampled/)  
HG002.HiC\_2\_NovaSeq\_rep1\_run2\_S1\_L001\_R1\_001.fastq.gz  
HG002.HiC\_2\_NovaSeq\_rep1\_run2\_S1\_L001\_R2\_001.fastq.gz

# ソースデータのサイズ

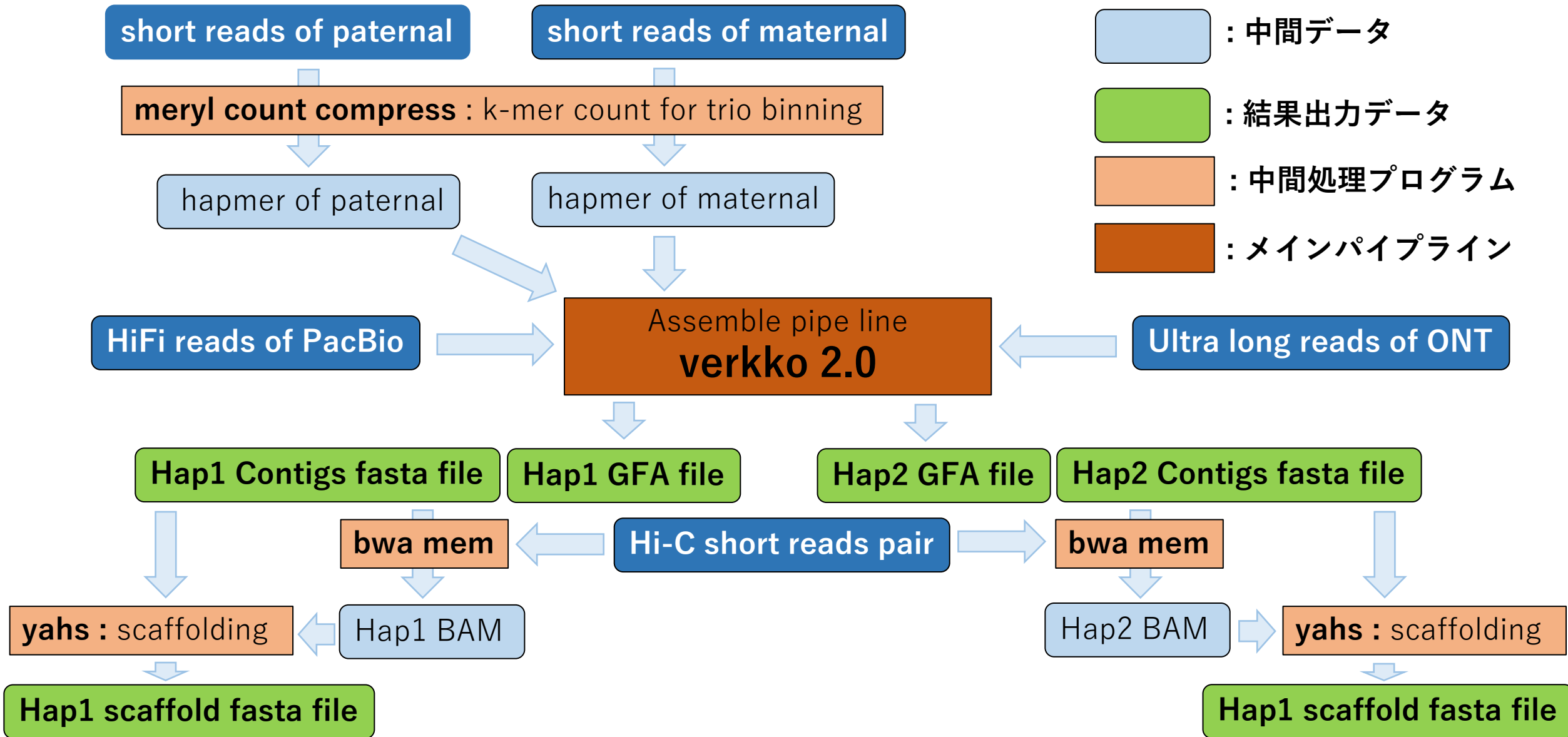
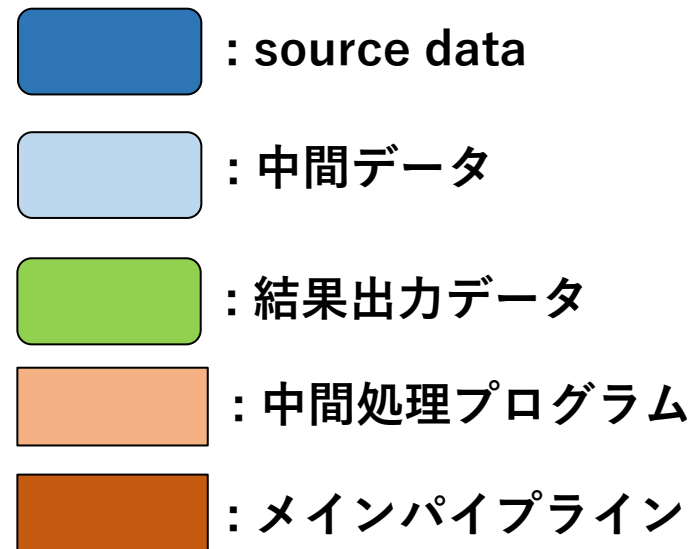
|                      | depth | Reads       | Bases           | MAX       | MIN   | AVG.    | Genome Size |
|----------------------|-------|-------------|-----------------|-----------|-------|---------|-------------|
| Pacbio HIFI read     | 37x   | 6,094,226   | 116,133,646,455 | 67,082    | 110   | 19,056  |             |
| ONT Ultra long read  | 68x   | 4,969,537   | 212,294,987,586 | 1,511,889 | 13    | 42,719  |             |
| Trio child paired    | 39x   | 385,889,841 | 115,604,433,850 | 150-150   | 16-16 | 150-150 |             |
| Trio paternal paired | 39x   | 394,074,835 | 117,967,687,966 | 150-150   | 16-16 | 150-150 |             |
| Trio maternal paired | 40x   | 398,182,798 | 119,345,388,750 | 150-150   | 16-16 | 150-150 |             |
| HI-C                 | 36x   | 441683225   | 221,066,628,565 | 250-250   | 35-35 | 250-250 |             |

※Trioのshort read paired-endは、Insert 1000でライブラリ調整を行う

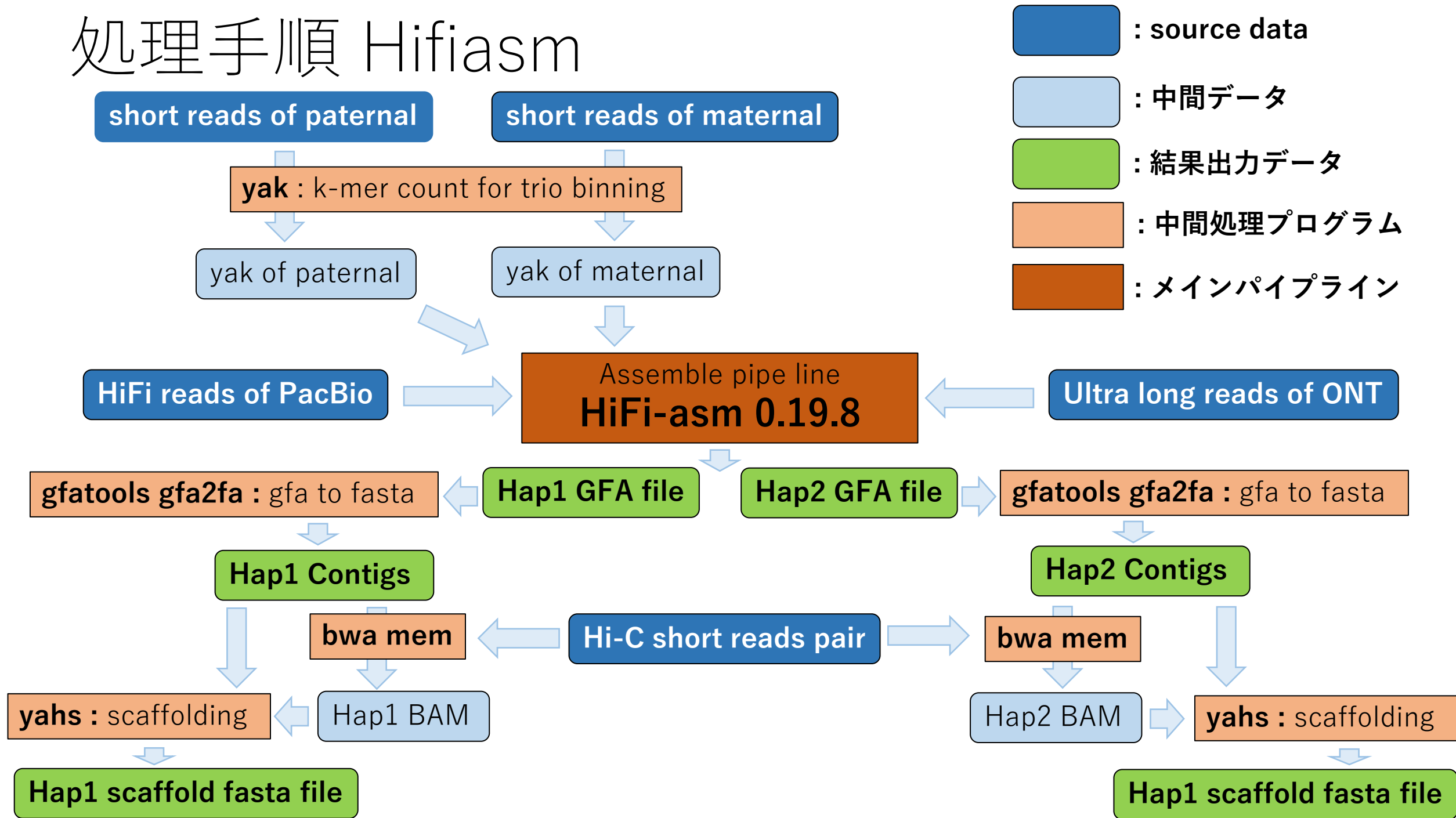
Genome size 0.8G の生物種の目標read量

|                      | depth | Reads       | Bases                 | MAX     | MIN   | AVG.    | Genome Size |
|----------------------|-------|-------------|-----------------------|---------|-------|---------|-------------|
| Pacbio HIFI read     | 37x   | 1,553,317   | <b>29,600,000,000</b> |         |       | 19,056  | 800,000,000 |
| ONT Ultra long read  | 68x   | 1,273,438   | <b>54,400,000,000</b> |         |       | 42,719  |             |
| Trio child paired    | 39x   | 106,666,667 | <b>32,000,000,000</b> |         |       | 300     |             |
| Trio paternal paired | 39x   | 106,666,667 | <b>32,000,000,000</b> |         |       | 300     |             |
| Trio maternal paired | 40x   | 106,666,667 | <b>32,000,000,000</b> |         |       | 300     |             |
| HI-C                 | 36x   | 57,600,000  | <b>28,800,000,000</b> | 250-250 | 35-35 | 250-250 |             |

# 処理手順 verkko



# 処理手順 Hifiasm



# Assemble 結果

アレルごとに分離されたContigの総塩基数は、それぞれほぼヒトのゲノムサイズに達している。  
Contig数も 3 Gのゲノムサイズの生物としては、絞り込まれた数になった。

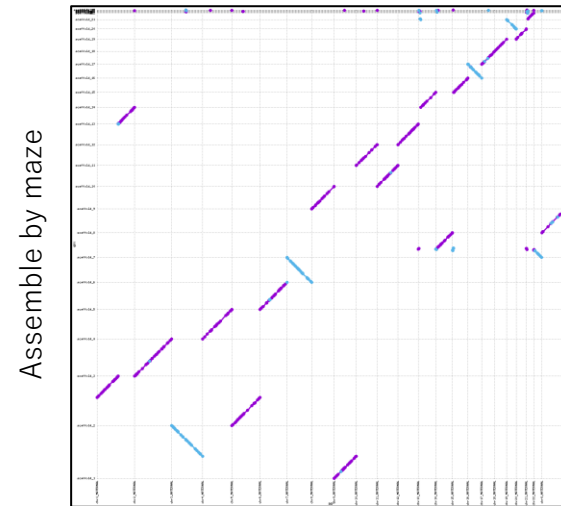
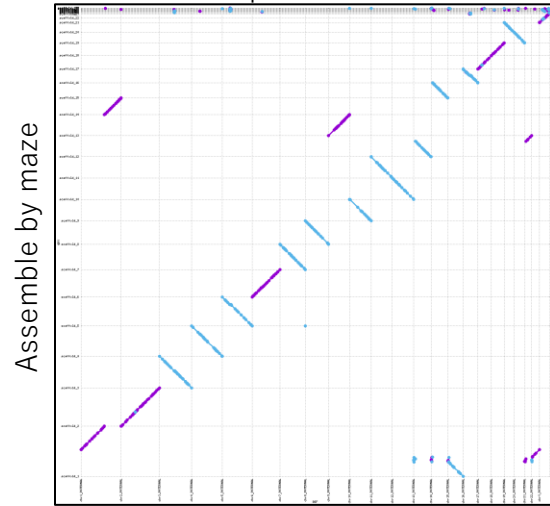
| verkko      | HiFi reads (PacBio) + Ultra-Long reads (ONT) + trio |                |                         |                         | HiFi+UL ONT   |
|-------------|-----------------------------------------------------|----------------|-------------------------|-------------------------|---------------|
|             | Hap1(paternal)                                      | Hap2(maternal) | Hap1(paternal)<br>+Hi-C | Hap2(maternal)<br>+Hi-C |               |
| Count       | 260                                                 | 248            | 239                     | 177                     | 2,042         |
| Total bases | 3,032,889,715                                       | 2,946,478,000  | 3,032,895,315           | 2,946,495,800           | 5,968,416,529 |
| MAX         | 201,096,426                                         | 201,509,204    | 242,895,120             | 329,916,864             | 114,325,491   |
| MIN         | 14,383                                              | 16,617         | 1,000                   | 4,000                   | 10,015        |
| AVG         | 11,664,960                                          | 11,880,960     | 12,689,939              | 16,646,869              | 2,922,829     |
| NG50        | 132,826,252                                         | 132,607,644    | 160,968,550             | 160,138,791             | 22,433,084    |
| NG90        | 132,663,905                                         | 113,749,518    | 154,353,810             | 158,243,678             | 13,252,115    |
| N50         | 61,302,799                                          | 47,443,367     | 83,741,811              | 80,786,575              | 12,667,762    |
| N90         | 50,403,022                                          | 39,318,803     | 78,825,830              | 61,684,081              | 3,274,990     |

| hifiasm     | HiFi reads (PacBio) + Ultra-Long reads (ONT) + trio |                |                         |                         | HG38          |
|-------------|-----------------------------------------------------|----------------|-------------------------|-------------------------|---------------|
|             | Hap1(paternal)                                      | Hap2(maternal) | Hap1(paternal)<br>+Hi-C | Hap2(maternal)<br>+Hi-C | Refseq Only   |
| Count       | 278                                                 | 195            | 195                     | 125                     | 25            |
| Total bases | 2,942,055,091                                       | 3,036,716,663  | 2,942,073,691           | 3,036,733,663           | 3,088,286,401 |
| MAX         | 190,348,608                                         | 200,790,271    | 314,859,149             | 342,454,084             | 248,956,422   |
| MIN         | 15,648                                              | 14,953         | 13,101                  | 1,000                   | 16,569        |
| AVG         | 10,582,932                                          | 15,572,906     | 15,087,557              | 24,293,869              | 123,531,456   |
| NG50        | 92,109,671                                          | 89,982,861     | 146,585,631             | 160,432,402             | 156,040,895   |
| NG90        | 86,824,164                                          | 85,123,516     | 135,604,170             | 154,393,854             | 145,138,636   |
| N50         | 24,603,582                                          | 37,584,915     | 80,102,668              | 83,605,520              | 80,373,285    |
| N90         | 13,092,539                                          | 35,025,270     | 61,100,867              | 78,884,356              | 64,444,167    |

# ゲノム配列比較 T2T project assembly との比較

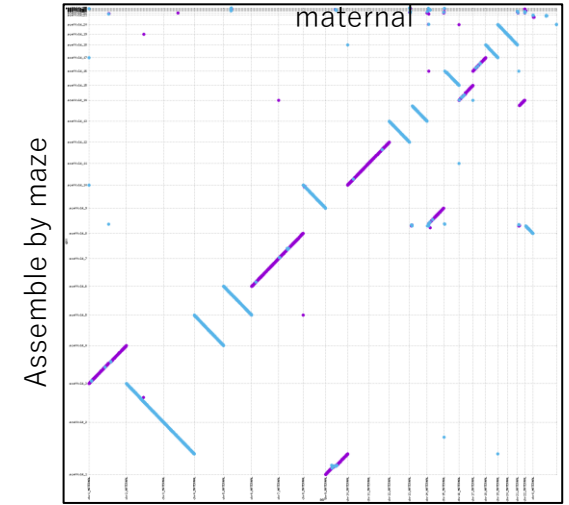
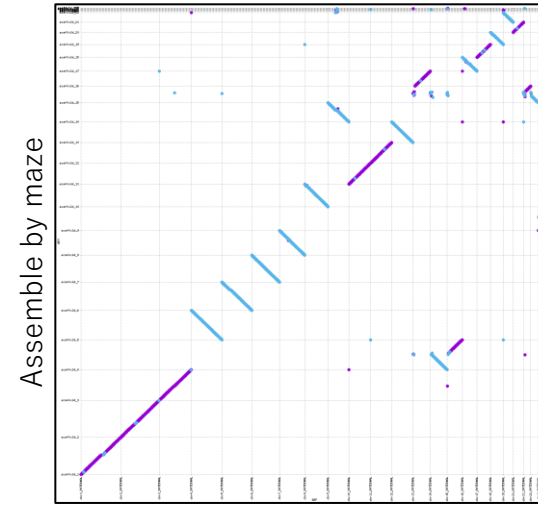
Assembled by hifiasm

HiFi reads (PacBio) + Ultra-Long reads (ONT) + trio + Hi-C  
paternal maternal

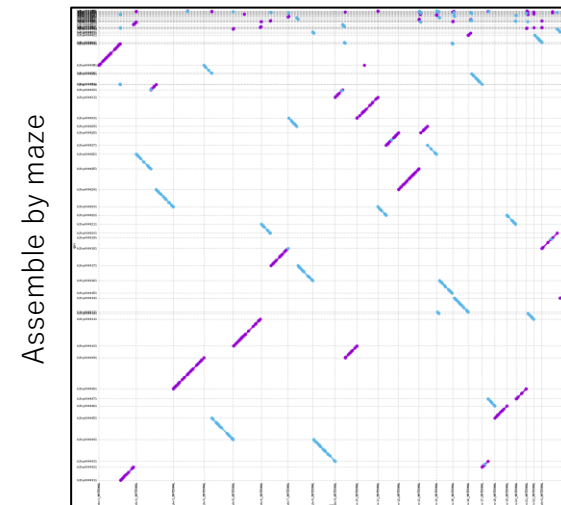
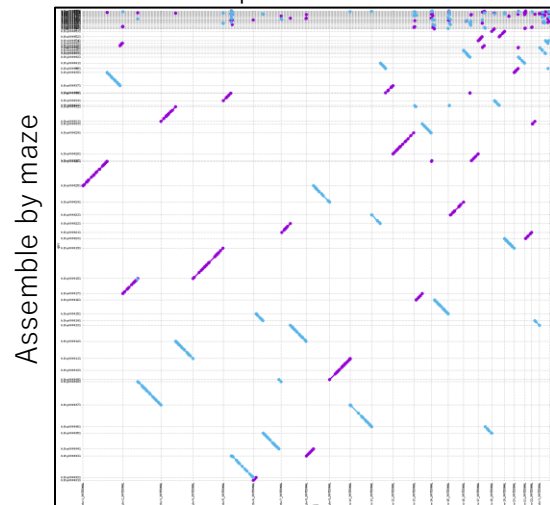


Assembled by verrko

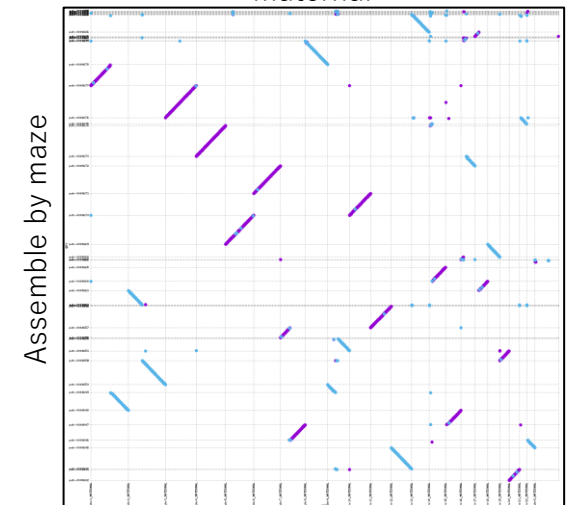
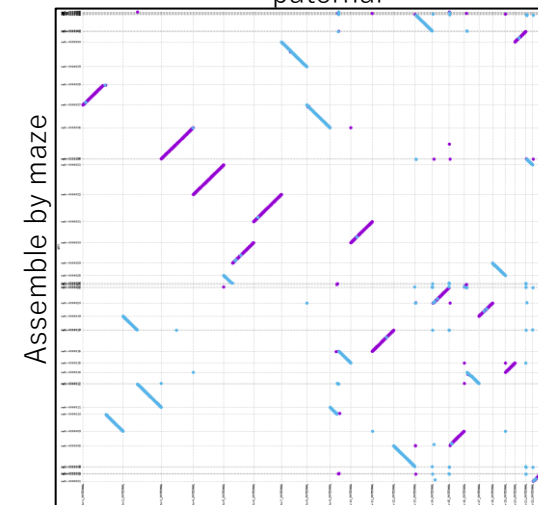
HiFi reads (PacBio) + Ultra-Long reads (ONT) + trio + Hi-C  
paternal maternal



HiFi reads (PacBio) + Ultra-Long reads (ONT) + trio  
paternal maternal



HiFi reads (PacBio) + Ultra-Long reads (ONT) + trio  
paternal maternal

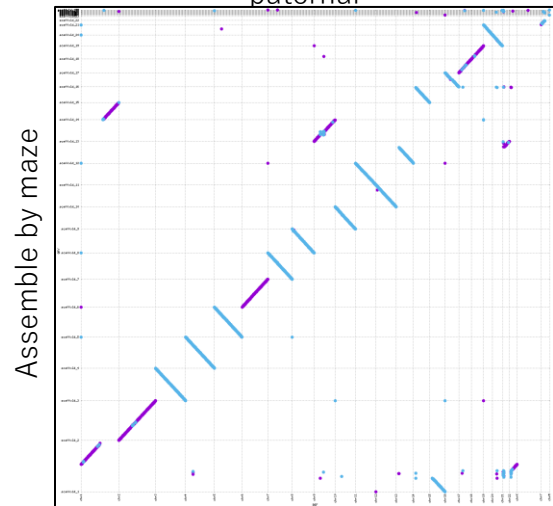


# ゲノム配列比較 hg38との比較

Assembled by hifiasm

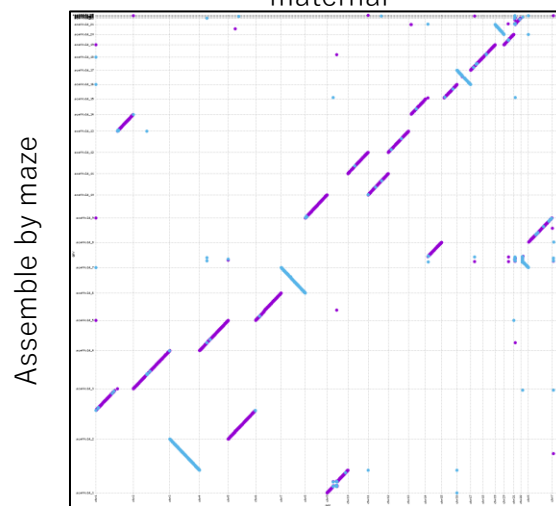
HiFi reads (PacBio) + Ultra-Long reads (ONT) + trio + Hi-C

paternal



HG38

maternal

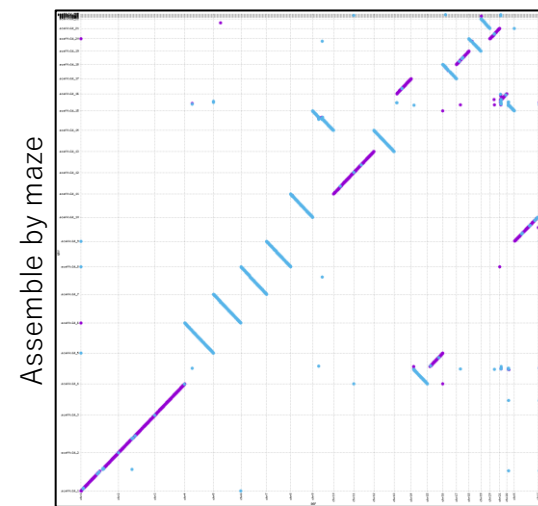


HG38

Assembled by verrko

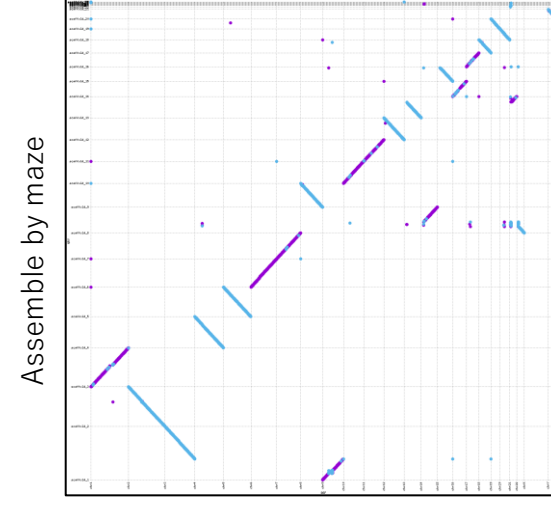
HiFi reads (PacBio) + Ultra-Long reads (ONT) + trio + Hi-C

paternal



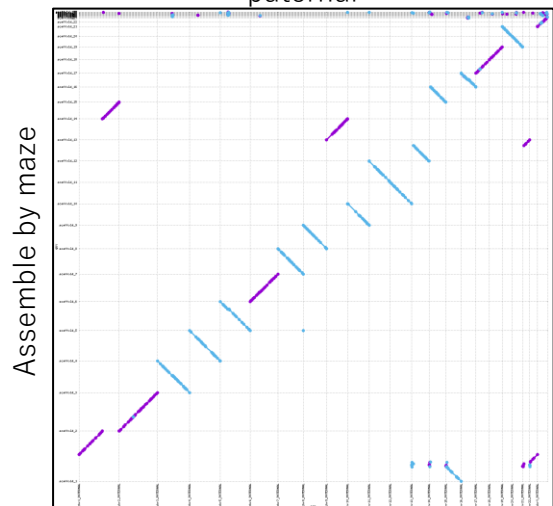
HG38

maternal



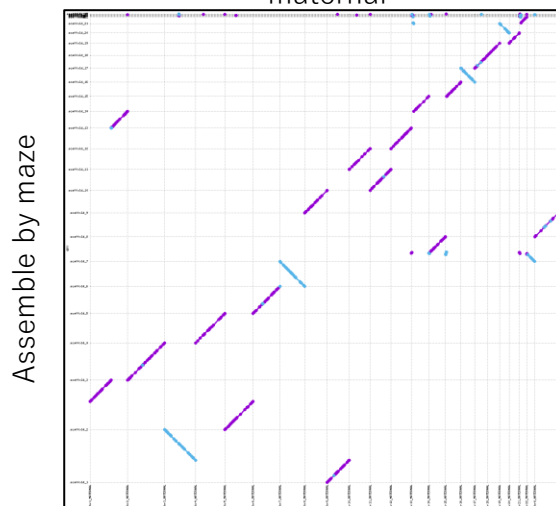
HG38

paternal



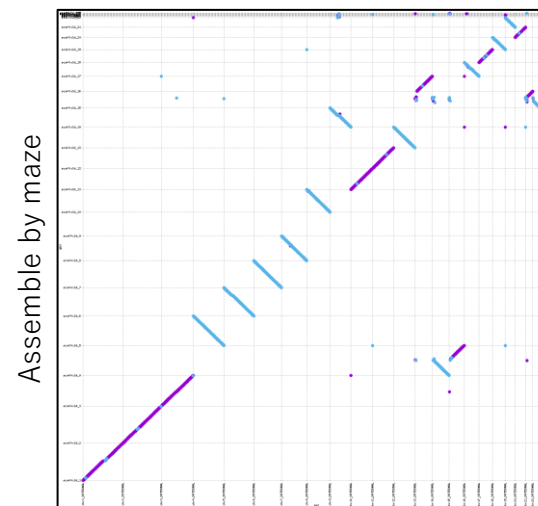
T2T project

maternal



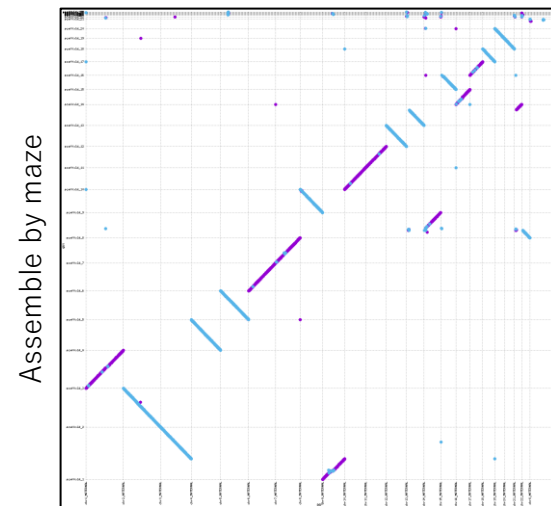
T2T project

paternal



T2T project

maternal



T2T project