De Novo Viral Quasispecies Assembly using Overlap Graphs

Alexander Schönhuth
joint with Jasmijn Baaijens, Amal Zine El Aabidine, Eric Rivals

Milano
18th of November 2016
Viral Quasispecies Assembly: “HaploClique”

Reference

- A. Töpfer, T. Marschall, R.A. Bull, F. Luciani, A. Schönhuth*, N. Beerenwinkel*
  Viral Quasispecies Assembly via Maximal Clique Enumeration
  PLoS Computational Biology, 10(3), e1003515, 2014
  *Joint last authorship
QUASISPECIES ASSEMBLY WORKFLOW

plasma sample

Next Generation Sequencing

alignment

reconstruction

reference genome
**Viral vs. Human Haplotyping**

<table>
<thead>
<tr>
<th></th>
<th>Human</th>
<th>Virus</th>
</tr>
</thead>
<tbody>
<tr>
<td># Haplotypes</td>
<td>2</td>
<td>? (1–1000)</td>
</tr>
<tr>
<td>Polymorphic locus</td>
<td>biallelic</td>
<td><em>up to all 4 nucleotides</em></td>
</tr>
<tr>
<td>Low-frequency variants</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Genome size</td>
<td>3 Giga</td>
<td>1 kilo – 1 Mega</td>
</tr>
<tr>
<td>Diversity</td>
<td>0.1 %</td>
<td>? (1–10 %)</td>
</tr>
<tr>
<td>Coverage</td>
<td>30-50x</td>
<td>≥ 20 000x</td>
</tr>
</tbody>
</table>
Viral quasispecies assembly: challenges

- Strains show at different frequencies
- Distinguish between closely related strains
- Low-frequency variants $\Leftrightarrow$ sequencing errors
- Exploit deep coverage
**Overlap Graph**

- **Nodes**: Probability distribution over sequences
- **Edges**: Connect two nodes $P, Q$, iff

$$\sum_{S \in \{A, C, G, T\}^*} P(S) \cdot Q(S) > \delta$$
**Overlap Graph**

- **Edges**: nodes (highly likely) reflect same haplotype

- **Max-Clique**: A (locally) maximal group of reads from the same haplotype
Haploclique Workflow (Part I)

1. Construct overlap graph (edges as just shown)
2. Enumerate maximal cliques
3. Transform cliques into contigs (implicit error correction!)
Iteration

1. Graph construction to contig formation as just shown
2. Contigs are new nodes in next iteration

*Result*: Contigs grow longer along iterations
## Full-Length Haplotype Reconstruction

### Setting:

- **HXB2**
- **JRCSF**
- **NL4-3**
- **YU2**
- **89.6**

- Gene-wise Hamming distance: 1-17%
- 600x data set

### Method

<table>
<thead>
<tr>
<th>Method</th>
<th>Error rate per bp</th>
<th>Number of haplotypes</th>
<th>Error-free</th>
</tr>
</thead>
<tbody>
<tr>
<td>HaploClique</td>
<td>0.012 %</td>
<td>49</td>
<td>99.7 %</td>
</tr>
<tr>
<td>ShoRAH</td>
<td>1.99 %</td>
<td>196</td>
<td>0 %</td>
</tr>
<tr>
<td>PredictHaplo</td>
<td>1.65 %</td>
<td>95</td>
<td>0 %</td>
</tr>
<tr>
<td>QuRe</td>
<td>2.95 %</td>
<td>15</td>
<td>0 %</td>
</tr>
</tbody>
</table>
Metagenome Gene Assembly: “Snowball”

Reference

I. Gregor, A. Schönhuth*, A.C. McHardy*

Snowball: Strain Aware Gene Assembly of Metagenomes

Bioinformatics, 32(17), i649-i657, 2016

*Joint last authorship
Metagenome Gene Assembly

- 3 genes (red, green, blue) for 3 strains
- Gene assembly reconstructs gene sequence from metagenome short read data
Conclusion

Mutatis mutandis (different clustering algorithm): analogous framework yields

*first strain-aware (and not only species-aware)*

gene assembler
ASSEMBLING POLYPLOID GENOMES

MANTRA

1. Construct an overlap graph:
   ▶ edges ↔ identical haplotypes
2. Identify suitable collections of cliques
3. Transform cliques into contigs
4. Back to 1.: contigs become nodes

Outcome

▶ Contigs grow along iterations
▶ Contigs reflect individual haplotypes
Viruses evolve rapidly.
Viruses evolve rapidly
Motivation summary

▶ Virus infection through a cloud of mutant strains
▶ Viral quasispecies assembly $\leadsto$ virus pan-genome
▶ Mixed sample; need to construct individual haplotypes first
▶ Existing methods depend on high-quality reference

What to do when there is a virus outbreak and there is no reference genome or the virus has diverged too far from the reference?
De Novo Viral Quasispecies Assembly: “SAVAGE”

Reference

- J. Baaijens, A.Z. El Aabidine, E. Rivals*, A. Schönhuth*
  De novo viral quasispecies assembly using overlap graphs
  bioRxiv:080341
  * Joint last authors
OVERLAP GRAPH BASED ASSEMBLY

1. Sequencing reads

   ▲ true mutation
   X sequencing error

2. Overlap graph
   - full length reads
   - strong edge constraints

3. Contigs

   Repeat steps 2 and 3 until convergence!
Challenge 1:

De novo overlap graph construction on deep sequencing data *without reference alignments*?

(\(\sim 10^6\) reads \(\sim 10^{12}\) potential overlaps...)
OVERLAP GRAPH CONSTRUCTION

SAVAGE-de-novo

Use FM-index + suffix filters

[N. Välimäki, S. Ladra, V. Mäkinen. Approximate all-pairs suffix-prefix overlaps.]

SAVAGE-ref

Construct ad-hoc reference (VICUNA)

[Yang et al. De novo assembly of highly diverse viral populations.]

1. compute read-to-reference alignments
2. get induced read-to-read alignments

⇒ For all overlap candidates, also check overlap quality!
Challenge 2:

Now we have the overlap graph, but it’s still huge...
FROM OVERLAP GRAPH TO SUPER-READS

**Step 1:** Read orientations

**Step 2:** Transitive edges

**Step 3:** Read clustering

**Step 4:** Super-read construction & integrated error correction

now iterate, back to step 1
Algorithm overview

Algorithm stages:

**Reads**
- Local assembly & error correction

**Contigs**
- Global assembly

**Maximized contigs**
- Master strain assembly

**Master strains**

Overlap graph construction:

1. Pairwise overlaps
2. Overlap quality check
3. Read orientations
4. Transitive edge removal
5. Read clustering
6. Graph updating until convergence
7. Output contigs

- Illumina reads
- Undirected overlap graph
- Directed overlap graph
- Simplified overlap graph
- Super-reads & overlaps
- Contigs

Local assembly & error correction
Global assembly
Master strain assembly
Master strains

Maximized contigs
## Benchmarking Results

**Simulated Data: HIV 5-virus-mix (Di Giallonardo et al)**

<table>
<thead>
<tr>
<th>Method</th>
<th>HIV-1 ref</th>
<th>VICUNA ref</th>
<th>HIV-1 ref</th>
<th>VICUNA ref</th>
<th>HIV-1 ref</th>
<th>VICUNA ref</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># contigs ≥ 500bp</td>
<td>largest contig</td>
<td>MAC length</td>
<td>ref. cov.</td>
<td>N-rate</td>
<td>mismatches</td>
</tr>
<tr>
<td>PredictHaplo</td>
<td>4</td>
<td>9710</td>
<td>0</td>
<td>79.5</td>
<td>0.659</td>
<td>1.374</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>9800</td>
<td>100</td>
<td>59.1</td>
<td>0.289</td>
<td>1.628</td>
</tr>
<tr>
<td>ShoRAH</td>
<td>39</td>
<td>9526</td>
<td>0</td>
<td>98.0</td>
<td>0.318</td>
<td>0.394</td>
</tr>
<tr>
<td></td>
<td>63</td>
<td>9657</td>
<td>1.6</td>
<td>97.2</td>
<td>0.307</td>
<td>1.356</td>
</tr>
<tr>
<td>MLEHaplo</td>
<td>185</td>
<td>9104</td>
<td>0</td>
<td>56.0</td>
<td>0</td>
<td>3.396</td>
</tr>
<tr>
<td>SAVAGE</td>
<td>de novo</td>
<td>6</td>
<td>9663</td>
<td>0</td>
<td>97.1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>HIV-1 ref</td>
<td>7</td>
<td>9633</td>
<td>0</td>
<td>97.0</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>VICUNA ref</td>
<td>9</td>
<td>9638</td>
<td>0</td>
<td>97.4</td>
<td>0</td>
</tr>
</tbody>
</table>

MAC = misassembled contigs; columns 3-7 = %
**BENCHMARKING RESULTS**

**Real Data: HIV 5-virus-mix (Di Gallionardo et al.)**

<table>
<thead>
<tr>
<th></th>
<th># contigs ≥ 500bp</th>
<th>largest contig</th>
<th>MAC length</th>
<th>ref. cov.</th>
<th>N-rate</th>
<th>mismatches</th>
<th>indels</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PredictHaplo</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HIV-1 ref</td>
<td>5</td>
<td>9642</td>
<td>0</td>
<td>99.2</td>
<td>0.259</td>
<td>0.615</td>
<td>0.104</td>
</tr>
<tr>
<td>VICUNA ref</td>
<td>5</td>
<td>11000</td>
<td>100</td>
<td>94.5</td>
<td>0.425</td>
<td>0.011</td>
<td>0.136</td>
</tr>
<tr>
<td><strong>ShoRAH</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HIV-1 ref</td>
<td>160</td>
<td>9581</td>
<td>0</td>
<td>98.9</td>
<td>0.378</td>
<td>3.203</td>
<td>0.113</td>
</tr>
<tr>
<td>VICUNA ref</td>
<td>169</td>
<td>10854</td>
<td>89.3</td>
<td>99.0</td>
<td>0.770</td>
<td>0.911</td>
<td>0.165</td>
</tr>
<tr>
<td><strong>SAVAGE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>de novo</td>
<td>1937</td>
<td>3320</td>
<td>0</td>
<td>90.7</td>
<td>0.017</td>
<td>0.039</td>
<td>0.043</td>
</tr>
<tr>
<td>HIV-1 ref</td>
<td>1525</td>
<td>2493</td>
<td>0</td>
<td>90.1</td>
<td>0.013</td>
<td>0.145</td>
<td>0.041</td>
</tr>
<tr>
<td>VICUNA ref</td>
<td>1508</td>
<td>2849</td>
<td>0</td>
<td>89.7</td>
<td>0.014</td>
<td>0.136</td>
<td>0.037</td>
</tr>
</tbody>
</table>

MAC = misassembled contigs; columns 3-7 = %
EVALUATING ALGORITHM STAGES

Reads

Local assembly & error correction

Contigs

Global assembly

Maximized contigs

Master strain assembly

Master strains

Largest contig

Mismatch rate

492 bp

1.276%

2078 bp

0.045%

3320 bp

0.039%

3320 bp

0.085%

Data set: 5-virus-mix (Di Giallonardo et al) -
https://github.com/cbg-ethz/5-virus-mix
CONCLUSIONS AND FUTURE WORK

▶ Challenges: unknown ploidy, low-frequency mutations and sequencing errors.

▶ Overlap graphs: full-length read information, detection of co-occurring mutations

▶ SAVAGE needs no prior information or reference genome

Future work

▶ More efficient de novo overlap computations

▶ Extend the method to long reads (model indels!)

▶ Constructing virus pan-genomes
**Strain divergence: limits**

**SAVAGE-de-novo**

<table>
<thead>
<tr>
<th>Two-strain mixture ratio</th>
<th>1:1</th>
<th>1:2</th>
<th>1:5</th>
<th>1:10</th>
<th>1:50</th>
<th>1:100</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>99.7</td>
<td>99.6</td>
<td>99.5</td>
<td>99.1</td>
<td>88.9</td>
<td>53.7</td>
</tr>
<tr>
<td>5%</td>
<td>99.7</td>
<td>99.6</td>
<td>99.3</td>
<td>99.2</td>
<td>77.7</td>
<td>56.9</td>
</tr>
<tr>
<td>2.5%</td>
<td>99.7</td>
<td>94.3</td>
<td>95.9</td>
<td>99.8</td>
<td>85.8</td>
<td>49.9</td>
</tr>
<tr>
<td>1%</td>
<td>76.3</td>
<td>99.6</td>
<td>95.2</td>
<td>82.7</td>
<td>78.6</td>
<td>53.3</td>
</tr>
<tr>
<td>0.75%</td>
<td>99.6</td>
<td>94.0</td>
<td>87.2</td>
<td>90.7</td>
<td>76.0</td>
<td>52.8</td>
</tr>
<tr>
<td>0.5%</td>
<td>93.4</td>
<td>91.0</td>
<td>90.9</td>
<td>49.0</td>
<td>68.5</td>
<td>49.6</td>
</tr>
</tbody>
</table>

**Total genome fraction (%)**

<table>
<thead>
<tr>
<th>Two-strain mixture ratio</th>
<th>1:1</th>
<th>1:2</th>
<th>1:5</th>
<th>1:10</th>
<th>1:50</th>
<th>1:100</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2.5%</td>
<td>0.19</td>
<td>0.41</td>
<td>0.40</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1%</td>
<td>0.25</td>
<td>0.06</td>
<td>0</td>
<td>0.28</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.75%</td>
<td>0.10</td>
<td>0.21</td>
<td>0.12</td>
<td>0.06</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.5%</td>
<td>0.08</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
</tr>
</tbody>
</table>

**Overall mismatch rate (%)**
Thanks for your attention!

Reference

J. Baaijens, A.Z. El Aabidine, E. Rivals*, A. Schönhuth*
* Joint last authors

De novo viral quasispecies assembly using overlap graphs

bioRxiv:080341