On DeBruijn Graphs, Consensus, and Variation

A new algorithm does it all (rapidly)

Gene Myers
Chair of Systems Biology
MPI for Cell Biology and Genetics
Dresden, DE
1. Framework & “What’s New”
   a. Sequel Preprocessing
   b. GFA2 Format
   c. VGP Goals & Protocol Proposal
   d. Scrubber 1.0 Release

2. A New Consensus Method
Reads with no chimeras, no adaptamers, no lowQ dropouts, capped error
Detecting and Masking Repeats

The alignments was specifically designed to find local alignments (LAs) between reads as opposed to overlaps where in each end of the alignment reaches an end of one or the
REPmask [-v] [-m<track(rep)>] -c<int> <subject:db> <overlaps:las> ...

This command takes as input a database <source> and a sequence of sorted local alignments blocks, <overlaps>, produced by a daligner run for said database. Note carefully that <source> must always refer to the entire DB, only <overlaps> can involve a block number.

DAMASKER commands

The latest source code can be found on GitHub here.
Module to determine where repeats are and make soft-masks of said

- **8 commits**
- **1 branch**
- **0 releases**
- **1 contributor**

- **thegenemyers** Small bug fix in DB.c
  - Latest commit 4c86fb0 4 days ago
- **DB.c** Small bug fix in DB.c
  - 4 days ago
- **DB.h** Kernel patch (not directly relevant)
  - 16 days ago
- **HPC.REPmask.c** New HPC.REPmask
  - A month ago
- **HPC.TANmask.c** Initial commit of masking programs
  - 2 months ago
- **Makefile** Makefile update
  - 2 months ago
- **QV.c** Kernel patch (not directly relevant)
  - 16 days ago
- **QV.h** Kernel patch (not directly relevant)
  - 16 days ago
1. Framework & “What’s New”
   a. Sequel Preprocessing
   b. GFA2 Format
   c. VGP Goals & Protocol Proposal
   d. Scrubber 1.0 Release

2. A New Consensus Method
New subsetting/filtering option, e.g. ‘-e ln >= 500 && rq >= 750’

zm - well number
ln - length of the subread
rq - quality value of the subread (normalized to [0,1000])
bc1 - # of the first barcode
bc2 - # of the second barcode
bq - quality of the barcode detection (normalized to [0,100])
np - number of passes producing subread
qs - start pulse of subread
BAM/SAM Input, Arrow Data Output

- **F.fasta**
  - 1N
- **F.fasta**
  - 5N
- **F.quiva**
  - 5N
- **F.example**
  - 1N

### .fasta

```
gaagagaatgttcagtaagcgcgaatcctgacataagtgaacgataaaggcggtgagttaaagcccctgccgggttcctgt
```

Sequence, 80bp’s per line

- **.quiva**
  - 111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111
  - DeletionQV, DeletionTag, InsertQV, MergeQV, SubstitutionQV, each one line

- **.arrow**
  - Pulse Widths, 80 per line
dex2DB: One-Step PreProcessing

S-DB: just contains sequence (.bps)
A-DB: contains sequence and arrow (.arw)
Q-DB: contains sequence and quiver (.qvs)
DB State Transition Diagram

- **S-DB**: just contains sequence (.bps)
- **A-DB**: contains sequence and arrow (.arw)
- **Q-DB**: contains sequence and quiver (.qvs)
dex2DB Relative Performance

90min for 100X Human Data Set
Thread or HPC acceleration?
1. Framework & “What’s New”
   a. Sequel Preprocessing
   b. GFA2 Format
   c. VGP Goals & Protocol Proposal
   d. Scrubber 1.0 Release

2. A New Consensus Method
Facilitate exchange between application developers (or application modules)

1. 5 line types: **Segment (S)**, **Edge (E)**, **Gap (G)**, **Path (O)**, **Subset (U)**
2. Simple tab delimited format
3. SAM-tag extensible
4. Can encode:
   - string graph,
   - local alignments,
   - consensus multi-alignment,
   - path in graph
   - subset of graph
   - scaffolds

[github.com/GFA-spec/GFA-spec](http://github.com/GFA-spec/GFA-spec)
1. Framework & “What’s New”
   a. Sequel Preprocessing
   b. GFA2 Format
   c. VGP Goals & Protocol Proposal
   d. Scrubber 1.0 Release

2. A New Consensus Method
Vertebrate Genome Project (VGP) Goals

1. The G10K is morph’ing into the VGP: bylaws, council, etc.

2. There are $\geq 66K$ vertebrate species.

3. (Revised) Goal is to do them all at Reference Quality ($3.4.2Q40$)
   a. Contig N50 $\geq 10^3$ Kbp (1Mbp)
   b. Scaffold N50 $\geq 10^4$ Kbp (10Mbp)
   c. $>90\%$ assigned to chromosomes with doubly confirmed O&O
   d. Consensus quality is $\geq Q40$

All agree on PacBio 60X as basis for reconstruction but also need
**short range scaffolding:** Bionano, 10X read clouds, Dovetail
**long-range scaffolding:** Hi-C, genetic maps

2017 goal: sequence 1 representation from 220 orders of vertebrates
My Protocol Proposal

50X Pacbio
5X Illumina in Hi-C read pairs (~ 30 mega-pairs)
15X Illumina in 10X read clouds

~ 20K EU today
100K genomes ~ 1 B2 bomber

1. Long read assembly
2. Link contigs and (further) separate haplotypes with read clouds
3. Search for missed spanning reads
4. Chromosome separation and “weak” scaffolding with Hi-C pairs
5. Polish consensus with 20X Illumina

6 bats with Bat1K and 6 fish with Axel Weber in 2017
BTW Schmidtae & Axolotl genomes submitted.
1. Framework & “What’s New”
   a. Sequel Preprocessing
   b. GFA2 Format
   c. VGP Goals & Protocol Proposal
   d. Scrubber 1.0 Release

2. A New Consensus Method
What is a “Scrubber”?

With as little loss of data as possible, “edit” reads so that:

(a) every edited read is a contiguous segment of the underlying genome (e.g. no chimeras or adapters)

(b) every part of an edited read is of a reasonable quality (e.g. < 20% error)

Guarantees a nice string graph!
DASqv: Intrinsic Quality Values:

DAStrim: Identify Contiguous Segments

Identify HQ segments (begin & end with Good, contain no Bad)
Classify all gaps as: Low-Q / Adaptamer / Chimer

Select longest stretch between ends or adaptamer

Chimera break to make new reads

DASpatch: Find patches for LowQ segments
**Scrubby 1.0 Pipeline**

- **DASqv**: compute intrinsic quality values
- **DASttrim**: find all HQ segments, remove chimers & adaptemers,
- **DASpatch**: find best B-read patch for each lowQ segment
- **DASedit**: produce new DB of trimmed, patched reads
- **DASrealign**: Adjust and extend all LAs in read piles
Effect of Scrubbing

100x D.Mel P5/C3 data set:

Before
Effect of Scrubbing

100x D.Mel P5/C3 data set:

After

21.9% fewer bases, 17.8% fewer reads

N50 read length: 14.3Kbp to 13.6Kbp
Consider DB’ your starting data set!

Compute anew

Use DASrealign

DASrealign: 40X faster on *D. mel* dataset but ~.1% of overlaps missing
Scrubber Futures …

Falcon

daligner → Correction → daligner → String Graph → Arrow

repmasker → daligner → scrubber → corrector → string graph → Arrow

courtesy J. Chin

Inverse cumulative read length distribution

- Raw
- Raw-FALCON EC
- Scrubbed-RAW
- Scrubbed-FALCON EC

Number of bases (Gbp)

0 10000 20000 30000 40000 50000 60000 70000

Read Length (bp)
1. Framework & “What’s New”
   a. Sequel Preprocessing
   b. GFA2 Format
   c. VGP Goals & Protocol Proposal
   d. Scrubber 1.0 Release

2. A New Consensus Method
A Review of DeBruijn Graphs

Consider set of all k-mers of reads for fixed k.

Every k-mer is a vertex

Every (k-1)bp overlap is an edge.

Every k-mer unique in target genome
Every read is perfect

A significant fraction \( m \in [0,1] \) must remain correct

\[ (1-\epsilon)^k \geq m \]
\[ k \leq \log m / \log (1-\epsilon) \]

<table>
<thead>
<tr>
<th>( \epsilon )</th>
<th>( m = .5 )</th>
<th>( m = .2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>.5%</td>
<td>138</td>
<td>321</td>
</tr>
<tr>
<td>1%</td>
<td>69</td>
<td>160</td>
</tr>
<tr>
<td>2%</td>
<td>34</td>
<td>80</td>
</tr>
<tr>
<td>3%</td>
<td>23</td>
<td>53</td>
</tr>
<tr>
<td>5%</td>
<td>14</td>
<td>31</td>
</tr>
<tr>
<td>10%</td>
<td>7</td>
<td>15</td>
</tr>
</tbody>
</table>
A Review of DeBruijn Graphs

Every k-mer unique in target genome

Error rate $\varepsilon$

<table>
<thead>
<tr>
<th>$\varepsilon$</th>
<th>m = .5</th>
<th>m = .2</th>
</tr>
</thead>
<tbody>
<tr>
<td>.5%</td>
<td>138</td>
<td>321</td>
</tr>
<tr>
<td>1%</td>
<td>69</td>
<td>160</td>
</tr>
<tr>
<td>2%</td>
<td>34</td>
<td>80</td>
</tr>
<tr>
<td>3%</td>
<td>23</td>
<td>53</td>
</tr>
<tr>
<td>5%</td>
<td>14</td>
<td>31</td>
</tr>
<tr>
<td>10%</td>
<td>7</td>
<td>15</td>
</tr>
</tbody>
</table>

k-mers not unique in target genome
Every read is perfect

No random dup’s requires $k \geq 2 \log_4 G$

E.Coli 22
Fly 27
Human 32

Sanger: $c = 10X$, $\varepsilon = 2-3\%$: only bacteria
Illumina: $c = 100X$, $\varepsilon = .5\%$: anything goes
PacBio: $c = 60X$, $\varepsilon = 10\%$: nothing is possible

However: $c = 60X$, $\varepsilon = 10\%$: good for $G = 100bp$!
Trace Points Save The Day

Partition A into 100bp panels
For each LA, length of B-read aligned to each panel recorded

Consensus / Correction of A-read?
Trace Points Save The Day

A sequence of \(|A|/100 \sim 100\text{bp}\) consensus problems !!

DeBruijn graph approach applies !!
Performance

Operates from 3X to 10,000X (no limit actually)

Converges more quickly to Q50 than other methods

Time is linear in # of bp’s

<table>
<thead>
<tr>
<th>N</th>
<th>Time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>4.2</td>
</tr>
<tr>
<td>10</td>
<td>9.2</td>
</tr>
<tr>
<td>20</td>
<td>18.9</td>
</tr>
<tr>
<td>30</td>
<td>24.6</td>
</tr>
<tr>
<td>40</td>
<td>35.7</td>
</tr>
<tr>
<td>50</td>
<td>43.7</td>
</tr>
<tr>
<td>60</td>
<td>48.5</td>
</tr>
<tr>
<td>70</td>
<td>55.3</td>
</tr>
<tr>
<td>80</td>
<td>60.2</td>
</tr>
<tr>
<td>90</td>
<td>68.5</td>
</tr>
<tr>
<td>100</td>
<td>74.7</td>
</tr>
</tbody>
</table>

For 10Mbp

⇒ 3.7h for 50X Human Consensus

⇒ 182h for Correcting 50X Human Data Set

At least 30X FalconSense & PBDagCon
Extreme Cases & Haplotypes

(a^{10}c^{10})^*  

Hap. 1  
Hap. 2

Hap. 1
Hap. 2
To Conclude …
Reported:

Sequel and Bam Preprocessing

An assembly specification format GFA 2

Announced VGP and plans

A complete, released Scrubber

Preliminary results on a fast and accurate correction/consensus algorithm

Forthcoming:

A complete assembly pipeline with Arrow consensus (this year)