De novo assembly

Brian J. Knaus
USDA Forest Service
Pacific Northwest Research Station

Kelly Vining and Katherine Pillman
Oregon State University
Calendar

- December meeting cancelled.
- January meeting organizational.
  - Introductions.
  - Projects we’re involved with.
  - What we want to get out of this group?
- February meeting: Cufflinks (pending)
Overview

- What are de Bruijn graphs?
- Software:
  - Velvet
  - ABysS
  - Trinity
Genomic vs. transcriptomic

- Genomic expectations:
  - Long contigs (i.e., chromosomes).
  - Uniform coverage.

- Transcriptomic expectations:
  - Short contigs (i.e., genes or operons).
  - Non-uniform coverage.
Two types of computer error

- Errors which crash the program and generate an error message.
- Errors where the software appears to exit successfully but has actually done something bad.
Essential bioinformatic software

- Database/query applications
  - Blast
  - Blat
  - MAQ
  - Bowtie
  - BWA

- De novo assembly
  - Velvet
  - ABySS

- Soap
De novo assembly

- Compiled code (usually C; Dennis Ritchie R.I.P).
- Require LOTS of memory!
- Implement de Bruijn graphs.
ACTGATTG

\( k\)-mer = 5
$\text{ACTGA}$

$\text{ACTGATTG}$

$k$-mer = 5
$k$-mer = 5
ACTGATTG

k-mer = 5
$k$-mer = 5
Schematic representation of our implementation of the de Bruijn graph.

Zerbino D R, Birney E. Genome Res. 2008;18:821-829
Schematic representation of our implementation of the de Bruijn graph.

Example of Tour Bus error correction.

Example of Tour Bus error correction.

Issues:
- Bubbles.
- Tips.
- Errors.

**Velvet**

- Current version: 1.1.06 (1.1.05 at CGRB cluster)


Basic Velvet de novo assembly

- **Run velveth**
  Velveth takes in a number of sequence files, produces a hashtable, then outputs two files in an output directory (creating it if necessary), Sequences and Roadmaps, which are necessary to velvetg.

- **Run velvetg**
  Velvetg is the core of Velvet where the de Bruijn graph is built then manipulated.
  
  Note that although velvetg saves some files during the process to avoid useless recalculations, the parameters are not saved from one run to the next.
On the command line...

/local/cluster/velvet
[viningk@needleman velvet]$ ll

```
total 32

drwxr-xr-x  2 root  root  34 Sep 24 2007 VELVET/
drwxr-xr-x  8 root  root  4096 Jul  7  2009 velvet_0.7.42/
drwxr-xr-x  8 root  root  4096 Dec 23 2009 velvet_0.7.55/
drwxr-xr-x  8 root  root  4096 Feb 16  2010 velvet_0.7.58/
drwxr-sr-x  8  4205  1304 4096 Oct  4  2010 velvet_1.0.12/
drwxr-xr-x  8 root  root  4096 May 18 10:29 velvet_1.1.02/
drwxr-xr-x  8 boyda spatafojlab 4096 Sep  7 14:39 velvet_1.1.05/
lrwxrwxrwx  1 root  root  13 Sep 22 14:15 velvet -> velvet_1.1.05/
drwxr-xr-x  9 root  root  152 Sep 22 14:15 ./
drwxrwxr-x 123 root  cgrb  4096 Nov  1 13:34 ../```
total 884

```bash
[viningk@needleman velvet_1.1.05]$ ll

```
```text
total 884

```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
```text
```
```bash
```text
```
Velveth

- Designate file type (fastq, fasta, etc)
- Designate read category
- Designate hash size

**Supported file formats are:**
- fasta (default)
- fastq
- fasta.gz
- fastq.gz
- sam
- bam
- eland
- gerald

**Read categories are:**
- short (default)
- shortPaired**
- short2 (same as short, but for a separate insert-size library)
- shortPaired2** (see above)
- long (for Sanger, 454 or even reference sequences)
- longPaired**

**Paired-end reads must be in a single file (ShuffleSequences.pl)**
Sample commands: velveth

> ./velveth output_directory hash_length
[[[-file_format]][[-read_type]] filename]

> /deepspace/strauss/viningk/velvet_1.1.02/velveth velvet_041411 29
-fastq -shortPaired s_5_seqsTrim6.txt

> /local/cluster/velvet/velvet_1.1.02/velveth velvet_091410b 31
-fastq -short s_1_seqsMergedTrim6.txt -fasta -long contigs_09.fa
**Velvetg**

- Designate library insert size
- Designate minimum contig length
- Designate coverage cutoff

Need to optimize according to genome size, expected genome coverage depth, etc.

**Examples:**
> ./velvetg output_directory/ -min_contig_lgth 100

> /deepspace/strauss/viningk/velvet_1.1.02//velvetg velvet_041411/ -ins_length 350 -min_contig_lgth 500

> /deepspace/strauss/viningk/velvet_1.1.02/velvetg velvet_051811b -ins_length 350  -exp_cov 20 -cov_cutoff 7.9 -min_contig_lgth 500
**Velvet Optimiser**

Tries a range of k-mer lengths with velveth, then runs velvetg with range of coverage cutoffs

Lives here on the CGRB cluster:
/local/cluster/velvet/velvet_1.1.05/contrib/VelvetOptimiser-2.1.7

Sample command:
/deepspace/strauss/viningk/velvet_1.1.02/contrib/VelvetOptimiser-2.1.7/VelvetOptimiser.pl -s 27 -e 31 -f '-shortPaired -fastq s_5_seqsTrim6.txt s_1_seqsMergedTrim6.txt' > s_15_VOtest3.txt
Sample VelvetOptimiser output

********************************************************
Assembly id: 1
Velveth timestamp: May 20 2011 13:47:08
Velveth version: 1.0.12
Readfile(s): -shortPaired -fastq s_5_seqsTrim6.txt s_1_seqsMergedTrim6.txt
Velveth parameter string: auto_data_31 31 -shortPaired -fastq s_5_seqsTrim6.txt s_1_seqsMergedTrim6.txt
Assembly directory: /deepspace/strauss/viningk/auto_data_31
Velvet hash value: 31
Roadmap file size: 12143769440
Total number of sequences: 150746258
********************************************************
May 20 13:47:08

Beginning vanilla velvetg runs.

********************************************************
Assembly id: 1
Assembly score: 140
Velveth timestamp: May 20 2011 13:47:08
Velvetg timestamp: May 20 2011 18:59:34
Velvetg version: 1.0.12
Velvetg version: 1.0.12
Readfile(s): -shortPaired -fastq s_5_seqsTrim6.txt s_1_seqsMergedTrim6.txt
Velvetg parameter string: auto_data_31 31 -shortPaired -fastq s_5_seqsTrim6.txt s_1_seqsMergedTrim6.txt
Velvetg parameter string: auto_data_31
Assembly directory: /deepspace/strauss/viningk/auto_data_31
Velvet hash value: 31
Roadmap file size: 12143769440
Total number of sequences: 150746258
Total number of contigs: 2635009
n50: 140
length of longest contig: 6491
Total bases in contigs: 343232524
Number of contigs > 1k: 11471
Total bases in contigs > 1k: 15400103
Sample VelvetOptimiser output
(continued)

**********************************************************
May 21 06:11:49 Setting cov_cutoff to 5.195.
********************************************************
Assembly id: 1
Assembly score: 93102316
Velveth timestamp: May 20 2011 13:47:08
Velvetg timestamp: May 21 2011 18:22:55
Velveth version: 1.0.12
Velvetg version: 1.0.12
Readfile(s): -shortPaired -fastq s_5_seqsTrim6.txt
             s_1_seqsMergedTrim6.txt
Velveth parameter string: auto_data_31 31 -shortPaired -fastq
                          s_5_seqsTrim6.txt s_1_seqsMergedTrim6.txt
Velvetg parameter string: auto_data_31 -exp_cov 17 -
cov_cutoff 5.1952
Assembly directory: /deepspace/strauss/viningk/auto_data_31
Velvet hash value: 31
Roadmap file size: 12143769440
Total number of sequences: 150746258
Total number of contigs: 1080880
n50: 602
length of longest contig: 12710
Total bases in contigs: 261325501
Number of contigs > 1k: 52009
Total bases in contigs > 1k: 93102316
Paired Library insert stats:
  Paired-end library 1 has length: 154, sample standard deviation: 57
  Paired-end library 1 has length: 160, sample standard deviation: 63
**********************************************************
ABySS

Katherine Pillman
Oregon State University
Horticulture
