Faster and more reliable genome assembly using Shovill

A/Prof Torsten Seemann
@torstenseemann
Introduction
Doherty Institute

:: Public health lab
   : Pathogen surveillance
   : Outbreak response

:: Dept Microbiology (Uni. Melb)
   : Antimicrobial resistance
   : Virulence mechanisms

:: Royal Melbourne Hospital
Lots of genomics & many pathogens
De novo genome assembly
In an ideal world ...

Sample DNA ➔ iSequencer™ ➔ All chromosomal and plasmid sequences

AGTCTAGGATTCGCTATAG
ATTCAGGCTCTGATATATT
TCGCGGCATTAGCTAGAGA
TCTCGAGATTCGTCCCAGT
CTAGGATTCGCTAT
AAGTCTAAGATTC...
De novo genome assembly

“From scratch”
De novo genome assembly

Usually sequencing a *population* of cells
De novo genome assembly
Velvet

:: The first short read assembler
:: Assemble reads into contigs
  : Single k-mer value
:: Scaffold contigs
:: No longer optimized for > 100 bp reads
SPAdes

:: Correct errors in reads
:: Assemble reads into contigs
    : Multiple k-mer values
:: Merge contigs
:: Scaffold contigs
:: Correct mistakes in contigs
Issues with SPAdes

- Read error correction is slow
- Concern about over-correcting close repeats
- Problems assembling overlapping PE reads
- Default k-mer values not great
- Mysterious --cov-cutoff option changes
- Scaffolding sometimes duplicates contigs
- Non-auditable assembly correction step
Improving on SPAdes
Shovill pipeline

1. Estimate genome size - *Mash*
2. Downsample reads - *seqtk*
3. Trim reads for adapters/quality - *Trimmomatic*
4. Correct errors in reads - *Lighter*
5. Stitch overlapping PE reads - *FLASH*
6. Choose k-mer values from reads - *my logic*
7. Assemble reads into contigs - *SPAdes*
8. Scaffold contigs
9. Correct mistakes in contigs - *BWA-MEM + Pilon*
1. Estimate genome size

:: Count unique k-mers occurring >= 3 times
:: Eg. *Listeria monocytogenes* (2.9 Mbp)

```
No. of k-mers below min. threshold : 22384165
No. of k-mers above max. threshold : 0
No. of unique k-mers : 25471659
No. of unique counted k-mers : 3087494
Total no. of k-mers : 220909116
Total no. of reads : 1780791
Total no. of super-k-mers : 25824881
```
2. Subsample reads

:: Divide read yield by estimated genome size
:: Spades works best with ~100x

Usage: seqtk sample [-2] [-s seed=11] <in.fa> <frac> | <number>

Options: -s INT RNG seed [11]
-2 2-pass mode: twice as slow but with much reduced memory
3. Trim reads

Optional trimming

:: Illumina adapter sequences provided
:: Q10 clipping

```bash
--trimopt XXX Trimmomatic options (default: 'ILLUMINACLIP:/home/tseemann/git/shovill/bin/.../db/trimmomatic.fa:1:30:11 LEADING:3 TRAILING:3 MINLEN:30 TOPHRED33')
```
4. Correct reads

Lighter - fast, scales to CPUs

Start

Scanning the input files to infer alpha(sampling rate)
Average coverage is 91.157 and alpha is 0.077
Bad quality threshold is "\
Finish sampling kmers
Bloom filter A's false positive rate: 0.010863
Finish storing trusted kmers
Finish error correction

Processed 1828582 reads:
1343655 are error-free
Corrected 804943 bases (1.659926 corrections for reads with errors)
Trimmed 0 reads with average trimmed bases 0.000000
Discard 0 reads
5. Overlap PE reads

Sequencing library fragment

Paired-end read R1

Paired-end read R2

Overlapping sequence

Virtual reconstructed single-end read

[FLASH] Read combination statistics:
Total pairs: 914291
Combined pairs: 538800
Uncombined pairs: 375491
Percent combined: 58.93%
6. Choose k-mer values

:: Default SPAdes k-mer values not ideal

Used k-mer sizes: 21, 33, 55, 77

:: Sample the first 10,000 reads

  : Examine read length distribution
  : Skipping k=21 speeds things up a lot
  : Going to k > 77 improves assemblies
7. *De novo* assembly

:: Use `--only-assembler`
   : As we corrected reads with Lighter

:: Use `--pe1-1 / --pe1-2` and `--s2`
   : As we have stitched PE and long SE reads

:: Don’t use `--careful`
   : As we will correct the contigs ourselves
9. Correct contigs

:: PILON
   : Align original PE reads back to contigs
   : Call variants and apply them
   : SNPs, MNPs, can optionally break contigs
   : Some local de novo re-assembly

:: Typically fixes 1 - 50 mistakes
Results
Source code

https://github.com/tseemann/shovill
SYNOPSIS
De novo assembly pipeline for Illumina paired reads

USAGE
shovill [options] --outdir DIR --R1 R1.fq.gz --R2 R2.fq.gz

GENERAL
--help This help
--version Print version and exit
--check Check dependencies are installed

INPUT
--R1 XXX Read 1 FASTQ (default: '')
--R2 XXX Read 2 FASTQ (default: '')
--depth N Sub-sample --R1/--R2 to this depth. Disable with --depth 0
--gsize XXX Estimated genome size eg. 3.2M <blank=AUTODETECT> (default

OUTPUT
--outdir XXX Output folder (default: '')
--force Force overwite of existing output folder (default: OFF)
--minlen N Minimum contig length <0=AUTO> (default: 0)
--mincov n.nn Minimum contig coverage <0=AUTO> (default: 2)
--namefmt XXX Format of contig FASTA IDs in 'printf' style (default:
--keepfiles Keep intermediate files (default: OFF)
Results

:: ~ 2 - 5x faster (but Spades improving)
:: Succeeds where Spades fails
:: More contigs, lower N50, similar Max, more bp
  : Could be a good thing!
:: Still evaluating accuracy
  : Need good finished genomes where SRA reads available
Installing Shovill

conda install
-c conda-forge -c bioconda
-c defaults shovill

brew install
brewsci/bio/shovill
github.com/tseemann/shovill
Contact

tseemann.github.io

torsten.seemann@gmail.com

@torstenseemann
The End