NGS Data Techniques: Reference-Based Mapping and de Novo Assembly

Matt Gitzendanner magitz@ufl.edu

Oleksandr Moskalenko om@hpc.ufl.edu

Reference-based mapping

- Map NGS reads onto a reference genome
  - Identify SNPs
  - RNA-seq
  - ChIP-seq
  - Etc.

Bowtie (Langmead *et al.* 2009)

- Pre-built reference genome index
  - Burrows-Wheeler transform
  - Index needs to be computed prior to mapping
    - Either build your own: bowtie-build
    - Or ask for index to be installed for you
- Important parameters
  - -v vs. –n
    - Two mapping modes

Bowtie (Langmead *et al.* 2009)

- Mapping mode
  - -v: map reads that have less than v mismatches
    - Ignores quality scores
    - -v can be 0-3

```
Reference ATGCGTAGTACGTCAACGTGTCACGTGACAGACAGT
Read        CGAAGTACGACAACGGGTCAC
```

If number of mismatches <= v, read maps

Bowtie (Langmead *et al.* 2009)

- Mapping mode
  - -n: map using quality scores
    - -n: Mismatches in seed (0-3), ignores quality
    - -l: seed length (default 28bp)
    - -e: max quality score of mismatches across read (default 70)
      - Quality scores range from 0-40

```
Reference ATGCGTAGTACGTCAACGTGTCACGTGACAGACAGT
Read        CGAAGTACGACAACGGGTCAC
          ----------
```

Seed: -l 7 -n 1

If sum of quality scores on the mismatches is <=e, read maps here, otherwise not

Bowtie (Langmead *et al.* 2009)

- Dealing with multiple mappings
  - -k: report up to *k* good alignments per read (1)
  - -a: report all alignments for a read (slow!)
  - -m: don't report if more than m alignments exist
  - -M: like –m, but report 1 random alignment
  - --best: guarantees alignment is in best stratum
  - --strata: don't report suboptimal strata

Bowtie (Langmead *et al.* 2009)

- Keeping unmapped/mapped reads
  - --un <filename> unmapped reads
  - --al <filename> mapped reads
  - Can be helpful for downstream analyses
- Use –S for SAM output
  - Most likely will process output using SAM anyway
- -p: Bowtie is threaded, can run using multiple cores on ***one*** node
  - E.g.: nodes=1:ppn=8

Bowtie2 (Langmead & Salzberg 2012)

- Adds gapped read alignment (indels)
- Faster than Bowtie for reads longer than 50bp
- Supports local alignment
  - Can trim ends that don't map
- Can map reads over Ns in reference
- No colorspace option

Bowtie2 (Langmead & Salzberg 2012)

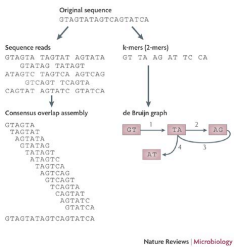
- Presets for both global and local
  - --very-fast(-local)
  - --fast(-local)
  - **--sensitive(-local) Defaults**
  - --very-sensitive(-local)

Other mapping applications

- BFAST
- BWA
- Maq
  - Bowtie is generally faster
- Mosaik
  - Handles gapped alignments relative to reference

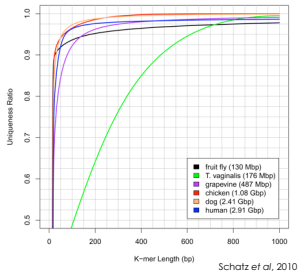
de Novo Assembly

- No reference genome
- Assemble contigs from reads
  - Assemble scaffolds using paired-end data
- Most short-read assemblers are de Buijn graph-based



kmers

- A kmer is a sequence of length *k*
  - Longer kmer
    - More unique
    - Fewer reads/kmer
  - Shorter kmer
    - Less unique
    - More reads/kmer
- The kmer you use does matter!
  - Try different kmers



Schatz *et al*, 2010

Velvet (Zerbino & Birney 2008)

- Two stages
  - velveth
    - Creates the hash table of kmers
  - velvetg
    - Uses the de Bruijn graph to create contigs & scaffolds
- kmer is critical
  - Default maximum value is 31
  - Need to change at compile time
    - velveth_max99 and
    - velveth_max99_OMP (threaded)
      - E.g.: nodes=1:ppn=8
    - velvetg_de: SOLiD colorspace versions

Velvet (Zerbino & Birney 2008)

- Can use multiple types of sequencing inputs
  - Short, long
  - Paired, single
  - Different insert sizes
  - Reference
- A mix of library types is typically needed for de novo genome assembly
- Many helpful scripts distributed with Velvet
  - VelvetOptimiser—helps pick best kmer
- Temporarily not available in Galaxy

Other de novo assembly applications

- Abyss
- ALLPATHS-LG
  - Has very specific requirements for library types and coverage
- Metavelvet
  - Modified version of Velvet for metagenomics
- Newbler
  - Provided by Roche (454), but can use Illumina data
- SOAPdenovo
- For RNA-seq
  - Oases (builds on after Velvet)
  - SOAPdenovo-TRANS
  - Trinity

Training Schedule

- ✓ Aug 28: Intro to UFHPC, getting started
- ✓ Sept 10: Modules, RHEL6 Transition, User Q&A
- ✓ Sept 17: The Linux/Unix Shell - An Introduction
- ✓ Sept 24: Running Jobs, Submission Scripts, Modules
- ✓ Oct 1: Galaxy Overview, The Basics
- ✓ Oct 8: NGS Data Techniques: General Methods and Tools
- ✓ Oct 15: NGS Data Techniques: Reference Based Mapping and de Novo Assembly
- Oct 22: Phylogenetic Analyses
- ◆ Oct 29: Research Computing Day: Moving Big Data
- Nov 5: Multiprocessing at the HPC Center
- Nov 12: Using Git and CMake to Organize and Drive Data Analysis Pipelines
- Nov 19: Introduction to GPU Nodes
- Nov 29: NGS Data Techniques: RNA-Seq
- Dec 3: NGS Data Techniques: Alternative Splicing Analysis

UF Research Computing

- Help and Support (Continued)
  - http://wiki.hpc.ufl.edu
    - Documents on hardware and software resources
    - Various user guides
    - Many sample submission scripts
  - http://hpc.ufl.edu/support
    - Frequently Asked Questions
    - Account set up and maintenance