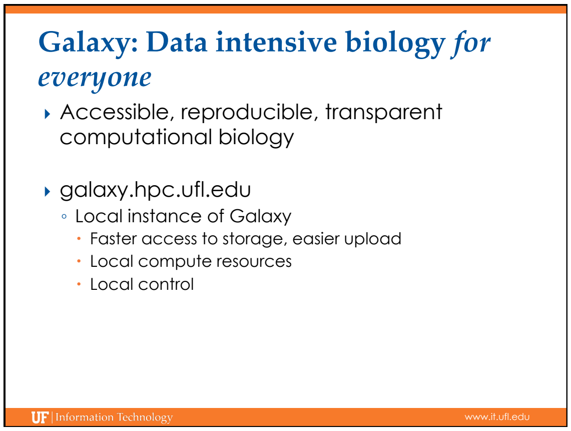
Next Generation Sequencing Data Techniques: Reference-Based Mapping and de Novo Assembly

Matt Gitzendanner
magitz@ufl.edu

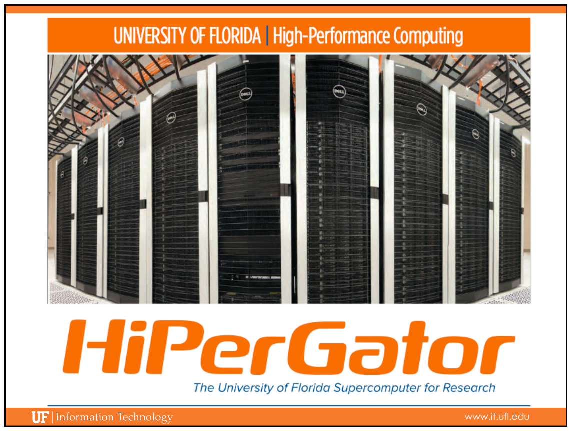
UF UNIVERSITY of FLORIDA Information Technology
RESEARCH COMPUTING
www.rc.ufl.edu

UF Research Computing
Information Technology
Home of High-Performance Computing and HiPerGator

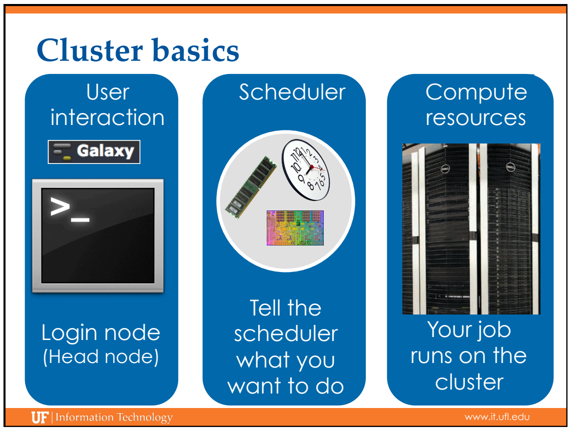


Galaxy: Data intensive biology *for everyone*

- Accessible, reproducible, transparent computational biology
- galaxy.hpc.ufl.edu
  - Local instance of Galaxy
    - Faster access to storage, easier upload
    - Local compute resources
    - Local control



The University of Florida Supercomputer for Research



Cluster basics

User interaction

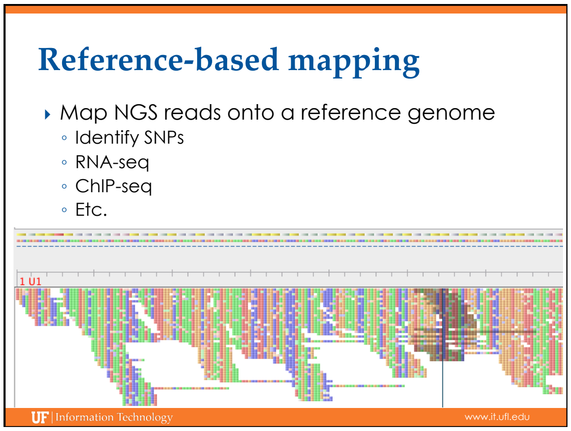
Login node
(Head node)

Scheduler

Tell the scheduler what you want to do

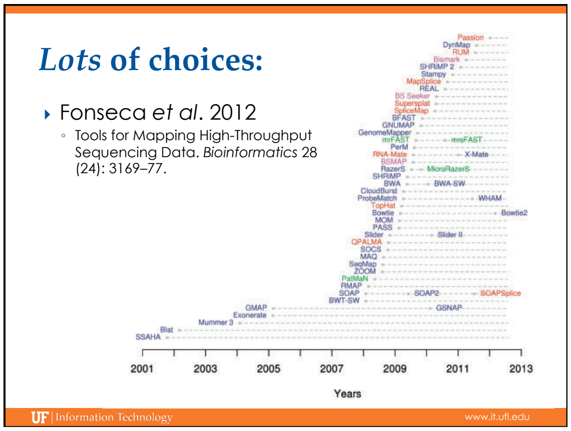
Compute resources

Your job runs on the cluster



Reference-based mapping

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



Lots of choices:

- Fonseca *et al*. 2012
  - Tools for Mapping High-Throughput Sequencing Data. *Bioinformatics* 28 (24): 3169–77.

BOW TIE **Bowtie 2**
Fast and sensitive read alignment
JOHNS HOPKINS UNIVERSITY

- **Tools that use Bowtie (or Bowtie 2):**
- mRNA sequencing:
  - Myrna (paper)
  - TopHat, TopHat 2 (paper)
  - Cufflinks (paper)
  - Trinity (paper)
  - MapSplice (paper)
  - SpliceSeq (paper)
  - RNASEQR (paper)
  - RSEM (paper)
  - GENE-counter (paper)
  - RobiNA (paper)
  - ExpressionPlot (paper)
  - RUM (paper)
  - BowStrap (paper)
  - ContextMap (paper)
  - eXpress (paper)
  - ReXpress (paper)
  - TrueSight (paper)
  - PASTA (paper)
  - QRNAseq
- Bisulfite-seq:
  - BSmooth (paper)
  - Bismark (paper)
  - MethylCoder (paper)
  - BS-Seeker (paper)
  - B-SOLANA (paper)
  - Bicycle
  - Bison
- Re-seqeuncing:
  - Crossbow (paper)
- Assembly and scaffolding:
  - Zorro
  - SSPACE (paper)
  - TIGER (paper)
- Workflow:
  - Galaxy (paper)
  - Chipster (paper)
  - GAPSS
  - Unipro UGENE (paper)
  - Biopieces
  - InSilico DB (paper)
  - LONI Pipeline (paper)
  - KBase
  - bioKepler
- Structural variants:
  - nFuse (paper)
- Gene fusion:
  - TopHat Fusion (paper)
  - deFuse (paper)
  - Comrad (paper)
  - ShortFuse (paper)
  - FusionHunter (paper)
  - SnowShoes-FD (paper)
  - FusionFinder (paper)
  - Chimerascan (paper)
  - FusionQ (paper)
- Metagenomics:
  - Genometa (paper)
  - MetaPhlan (paper)
  - MG-RAST (paper)
  - metAMOS (paper)
- Small RNA:
  - isomiRID (paper)
  - Deqor v3 (paper)
  - NEXT-RNAi (design of genome-wide RNAi libraries)
  - miREvo (paper)
  - miRanalyzer (paper)
  - MapMi (paper)
  - shortran (paper)
  - miRDeep2 (paper)
  - YM500 (paper)
  - starBase (paper)
- Other:
  - PADUA (alignment of DNA to protein databases; paper)
  - SeqTrimMap (sequential trimming and mapping; paper)
  - mPrime (primer design)
  - EMIGRE (ribosomal genes reconstruction; paper)
  - TAPDANCE (identifies CISs in transposon junction fragment insertion data; paper)
  - Paparazzi (reconstruction of viral genomes from small RNA reads; paper)
  - CaPSID (pathogen detection; paper)
  - EuPathDB (eukaryotic pathogen detection; paper)

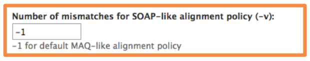
Ben Langmead

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


Number of mismatches for SOAP–like alignment policy (–v):
–1
–1 for default MAQ–like alignment policy

```
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

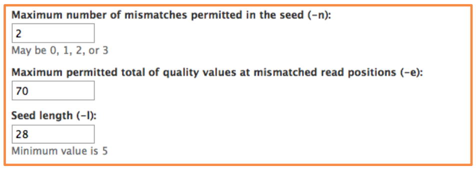
```
Reference ATGCGTAGTACGTCAACGTGTCACGTGACAGACAGT
     Read     CGAAGTACGACAACGGGTCAC
```

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)

- 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)


Maximum number of mismatches permitted in the seed (–n):
2
May be 0, 1, 2, or 3

Maximum permitted total of quality values at mismatched read positions (–e):
70

Seed length (–l):
28
Minimum value is 5

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

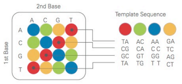
Bowtie2 (Langmead & Salzberg 2012)

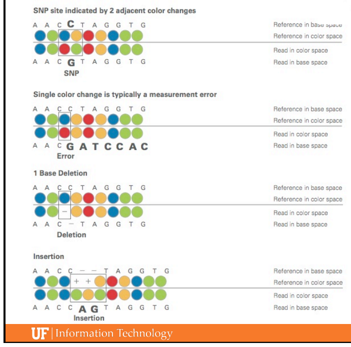
- Adds gapped read alignment (indels)
- Faster than Bowtie for reads longer than 50bp
- Supports local alignment
  - Can trim ends that don't map
- Can map reads over Ns in reference
- No colorspace option

Bowtie2 (Langmead & Salzberg 2012)

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

SOLiD data





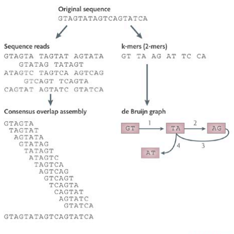
Use colorspace where possible

Other mapping applications

- BWA
- Lastz
- Maq
  - Bowtie is generally faster
- Mosaik
  - Handles gapped alignments relative to reference
- PerM
- SRMA

de Novo Assembly

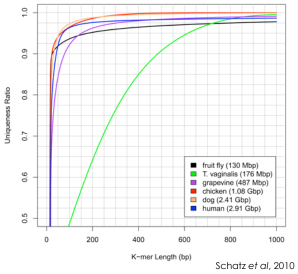
- No reference genome
- Assemble contigs from reads
  - Assemble scaffolds using paired-end data
- Most short-read assemblers are de Buijn graph-based



Nature Reviews | Microbiology

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
  - 11-31: Default for Velvet, most memory efficient
  - Up to 249 available.

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

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

UF Research Computing

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