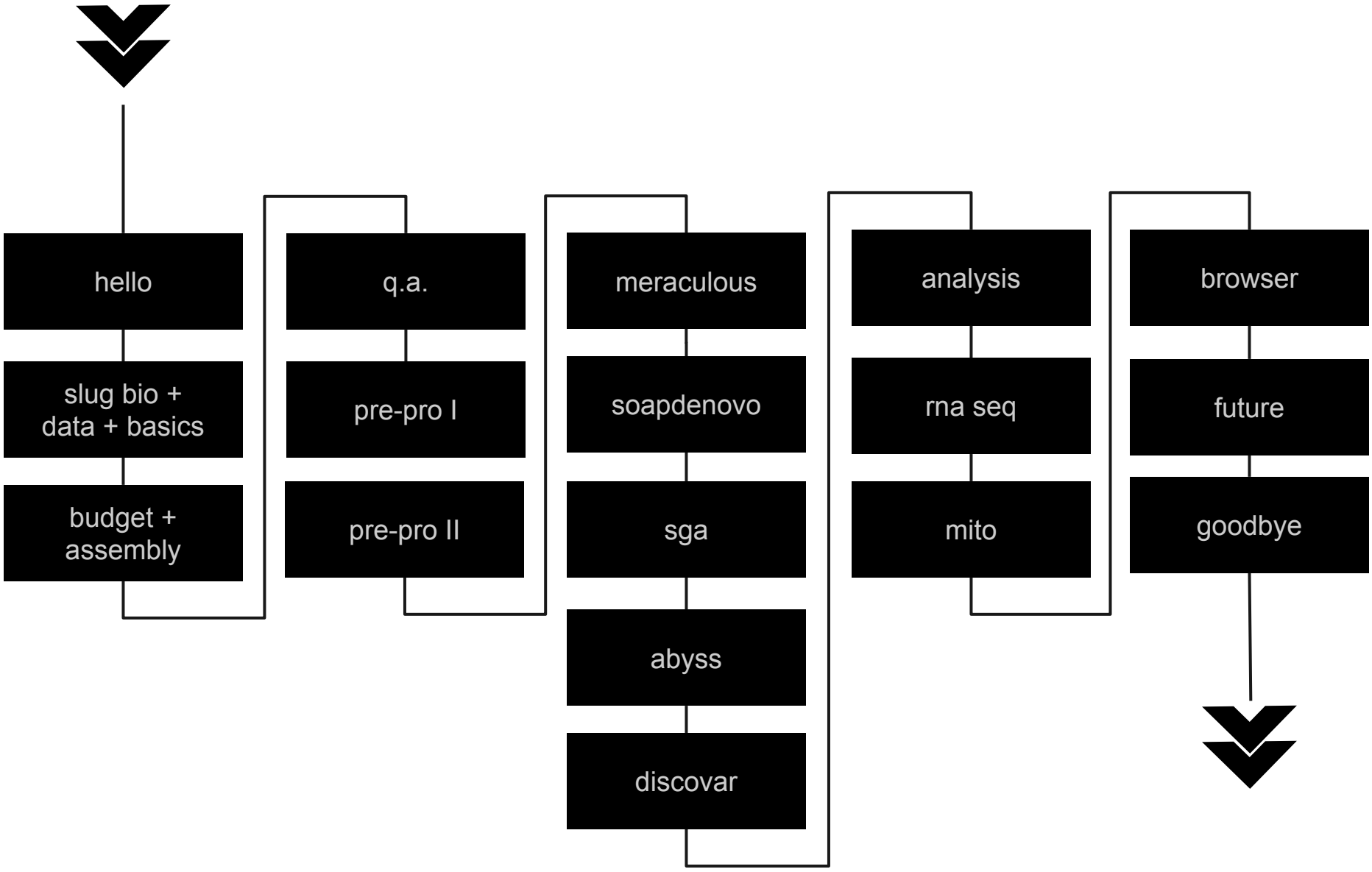
De Novo Assembly of the *Ariolimax dolichophallus* Genome

University of California, Santa Cruz

welcome

thanks


hello
slug bio +
data + basics
budget +
assembly
q.a.
pre-pro I
pre-pro II
meraculous
soapdenovo
sga
abyss
discovar
analysis
rna seq
mito
browser
future
goodbye

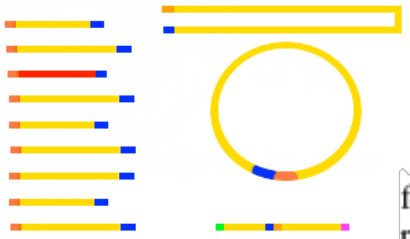
Banana Slug Biology

Banana Slugs are simultaneous hermaphrodites and most of the previous studies on *A. dolichophallus* were primarily concerned with their hermaphrodite anatomy and peculiar mating habits. A common belief is that land slugs evolved from land snails and separately from sea slugs. So slug does not refer to a taxonomic group but rather a body shape. On the land and in the sea some gastropods might have survived better without a bulky shell.

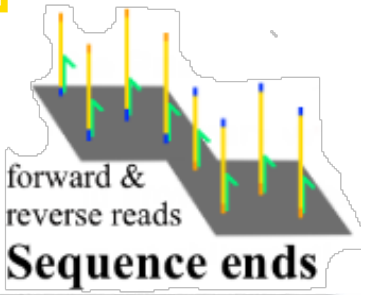

12

DNA extract

Shear & Select



Library prep



Sequence ends

Meyer-Kircher & Tagmentation library prep protocols

SW018 (597 bp insert), SW019 (374 bp insert)

- 2x100 and 2x250 bp reads on Hiseq and NxSeq

BS-MK (450-650 bp insert)

- 2x250 bp reads on NxSeq

BS-Tag (375-575 bp insert)

- 2x250 bp reads on NxSeq

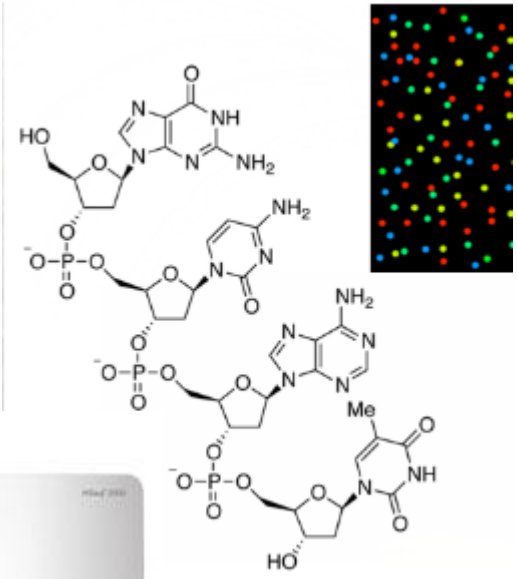
SW041(3-4 kb insert), SW042 (5-6 kb insert)

- 2x76 bp reads on HiSeq

Lucigen mate-pair (apx. 1kb insert)

- 2x300 bp reads on MiSeq

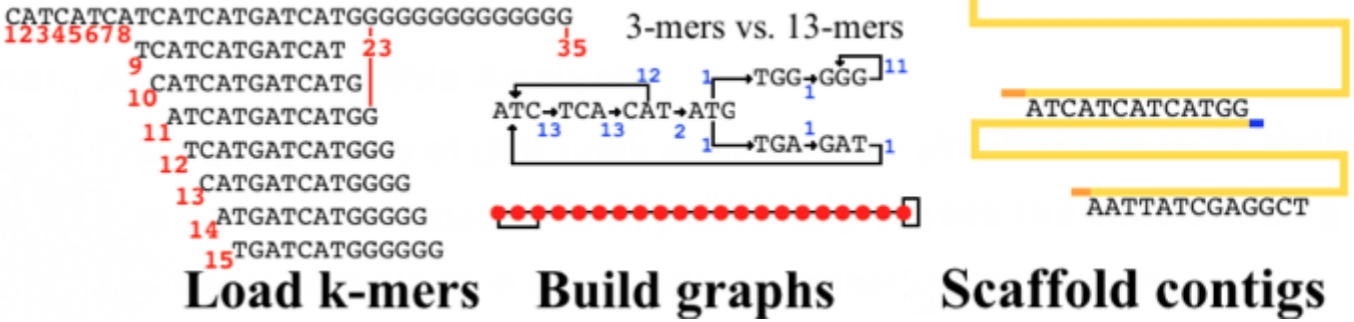
cDNA libraries from RNA



GCATAATTATCGATCG
ATCGATCGATTATCGA
TCTATAGAAATCTCGC
TTATCGCTTATTTCGA
AATCTCGATTTCGATT
CGGCTATTTTCAATAT
TATCGTATCTTTCAAT

Basics of Genome Assembly

A problem of combinatorics, or the art of combining many small structures into one large structure. The assemblers that this class used attempt to build contiguous sequences from paths on de Bruijn graphs built from our data. Reads are mapped back to these “contigs” and when a read maps to two contigs it will “scaffold” them together.




```
nnnACGTGTGCCC
CCnnnnnnnnnnn
nnnnnnnnnnnnn
nnnnnnnnnnnnn
nnnnnnnATCATC
ATCATGGnnnnnn
nnnnnnnnnnAGC
CTGCATAATTnnn
```

Budget

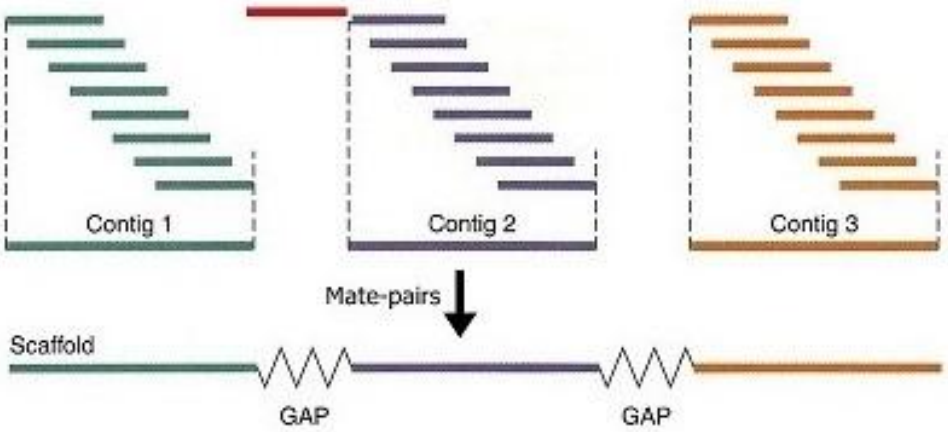
Item	Description	Cost (USD)
Donor Gift Fees	UCSC Fees for crowdfunding	$1,287.00
DNA extraction OMEGA mollusc DNA kit	Technician time for generating 3 DNA extracts	$109.00
Shotgun library Preps	Reagents and technician time for generating two shotgun Illumina libraries	$397.00
MiSeq sequencing	small scale MiSeq sequencing for lib QC	$57.00
MiSeq sequencing	2x300 full MiSeq run w/ 374 bp insert	$2,117.00
HiSeq sequencing	2x100 lib on single HiSeq lane w/ 374 bp insert	$2,000.00
HiSeq sequencing	2x100 lib on single HiSeq lane w/ 597 bp insert	$2,000.00
HiSeq sequencing	2x250 lib pools on two lanes at UCSF	$5,600.00
Mate-pair library Lucigen kit	User-provided reagents; technician time	$3,647.66
RNA extraction OMEGA mollusc RNA kit	Technician time for generating 30 RNA extracts	$445.00
Starting Budget		**$21,443.00**
Total Remaining		**$3,783.34**

Budget Goal 1

- Genome Assembly
  - de novo genome assembly
    - contig assembly
      - DNA extractions
      - Shotgun library preps
      - MiSeq, HiSeq runs
    - scaffolding
      - mate-pairs

No Known Predators

No Known Sequences



Budget Goal 2

- Genome Annotation (Partial)
  - extracting meaningful information from the assembled genome
  - Genes!
  - RNA-Seq
    - provides both support and sequence for genes
    - RNA Extraction
    - HiSeq (Pending)

Quality Assessment - FastQC

- High throughput sequencers generate millions of reads.
- One should check the raw data to ensure it is good enough to draw biological conclusions
- An overrepresented sequence identified by FastQC maybe one of three things:
  - biologically relevant
  - contamination
  - or mean the library is not as diverse as should be

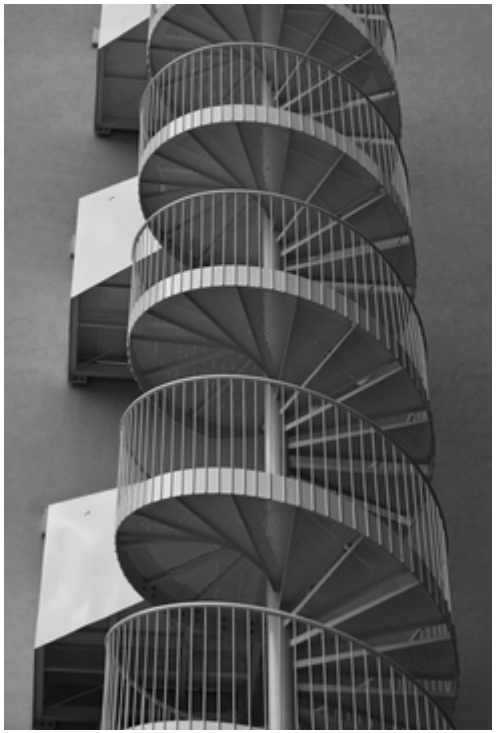
Source: http://www.bioinformatics.nl/courses/RNAseq/FastQC_Manual.pdf

Quality Assessment - Preqc

- Preqc is a module for QC from SGA assembler
- Preqc measures how easy the assembly is going to be by estimating how repetitive the genome is and also the sequencing error rate
- Estimates the genome size based on statistics of k-mer count
- Estimates proportion of read pairs that are PCR-duplicates

Source: https://groups.google.com/forum/#!msg/sga-users/95dTwpJCARU/oKoq54EZqKwJ

Quality Assessment - Preqc



- Preqc suggested a high repeat content in the slug's genome
- Repeat regions represent a challenge to any assembler
- Jim Kent's analogy of prose and poetry:

Mary had a little lamb, little lamb, little lamb, Mary had a little lamb whose fleece was white as snow

http://www.dnalc.org/view/15305-Compiling-the-data-from-the-Human-Genome-Project-Jim-Kent.html

Quality Assessment - fastUniq

- PCR is the main source of duplicates in NGS, affecting paired read mapping
- This has impact on scaffolding, introducing false positive and negative connections between contigs
- fastUniq is specific for one of the assemblers (Discovar *de novo*) pipeline and removes duplicates in three steps:
  - importing reads
  - sorting them
  - identifying duplicates

Source: Xu et al., 2012

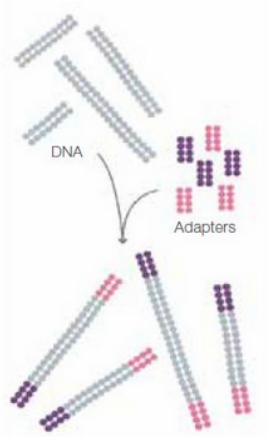
Quality Assessment - fastUniq

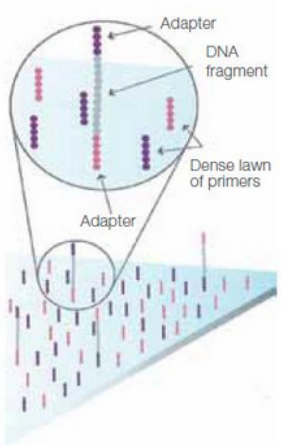
- fastUniq files were subjected to FastQC
- Read counts were compared before and after duplicate removal, allowing estimation of the duplicate content:

Dataset	# Reads before fastUniq	# Reads after fastUniq	% duplicates
UCSF_SW018	61,132,697	56,931,153	6.87%
UCSF_SW019	79,215,987	75,081,065	5.21%

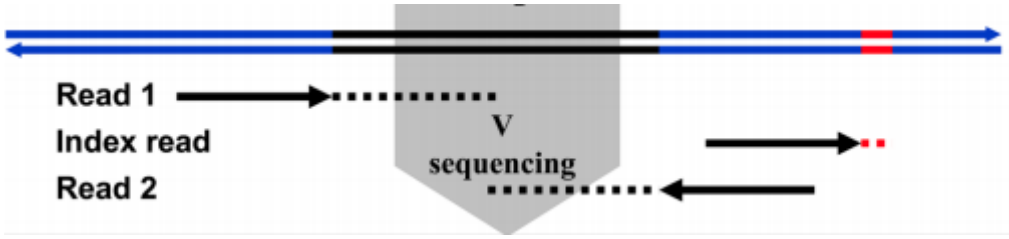
Pre-processing - Adapter Removal

- Adapters are ligated to the ends of fragmented DNA during library preparation





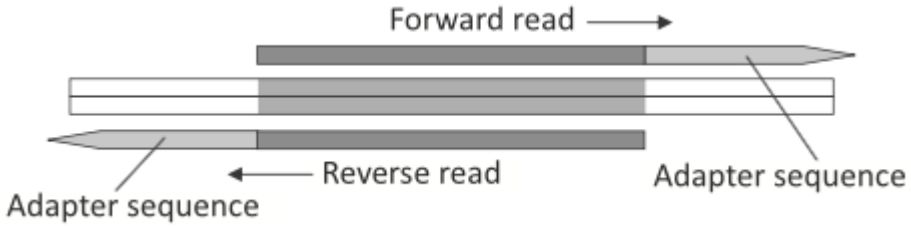
- Must be removed in order to prevent incorrectly aligning the reads that are obtained from sequencing



Figures taken from Mardis, 2008 and Meyer & Kircher, 2010

Pre-processing - Adapter Removal

- 2 tools were used for adapter removal
  - **Skewer**
    - Works on single-end, paired-end, and mate-pair reads
    - Quick run-time



  - **Seqprep**
    - Dedicated paired-end read adapter trimmer and merger
    - Slow run-time

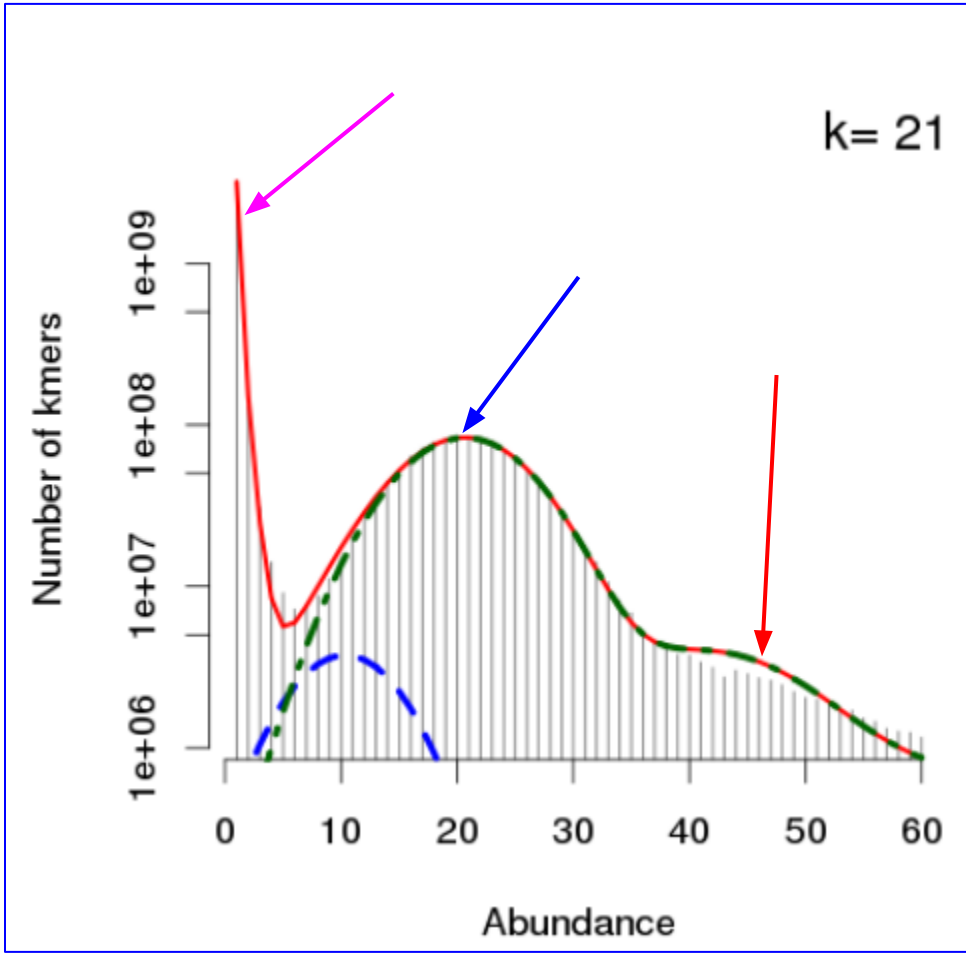
kmers

- kmer examples:

  3-mer:ATG

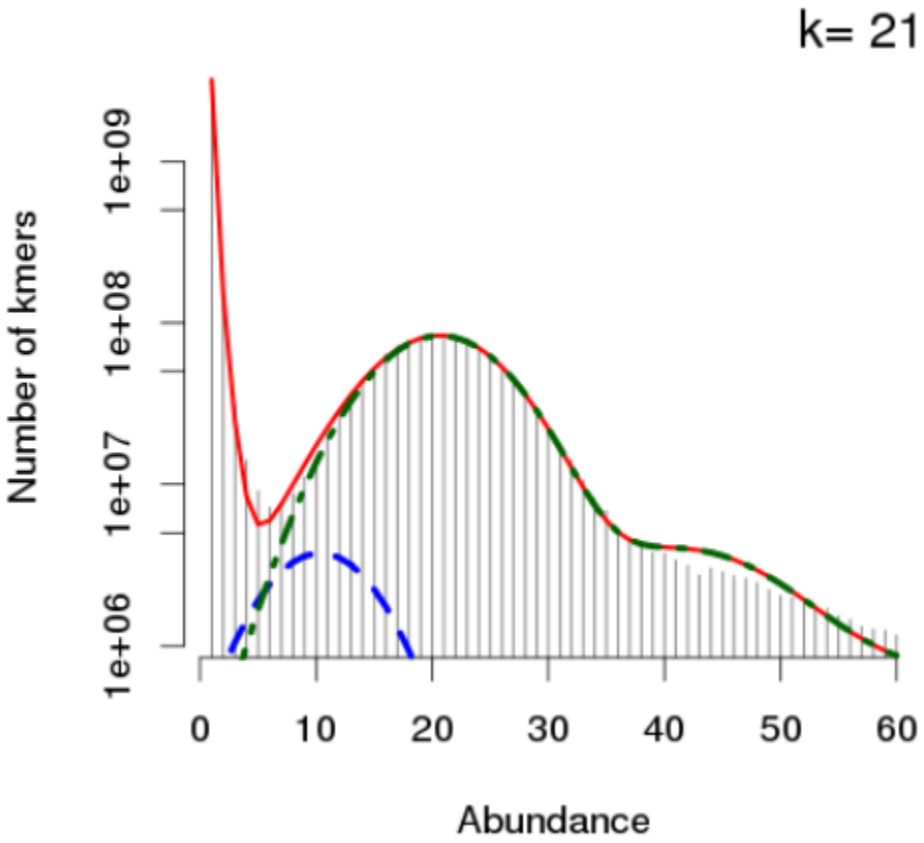
  5-mer:TTATT

- We can learn a lot from how many times a kmer is seen in the raw reads.

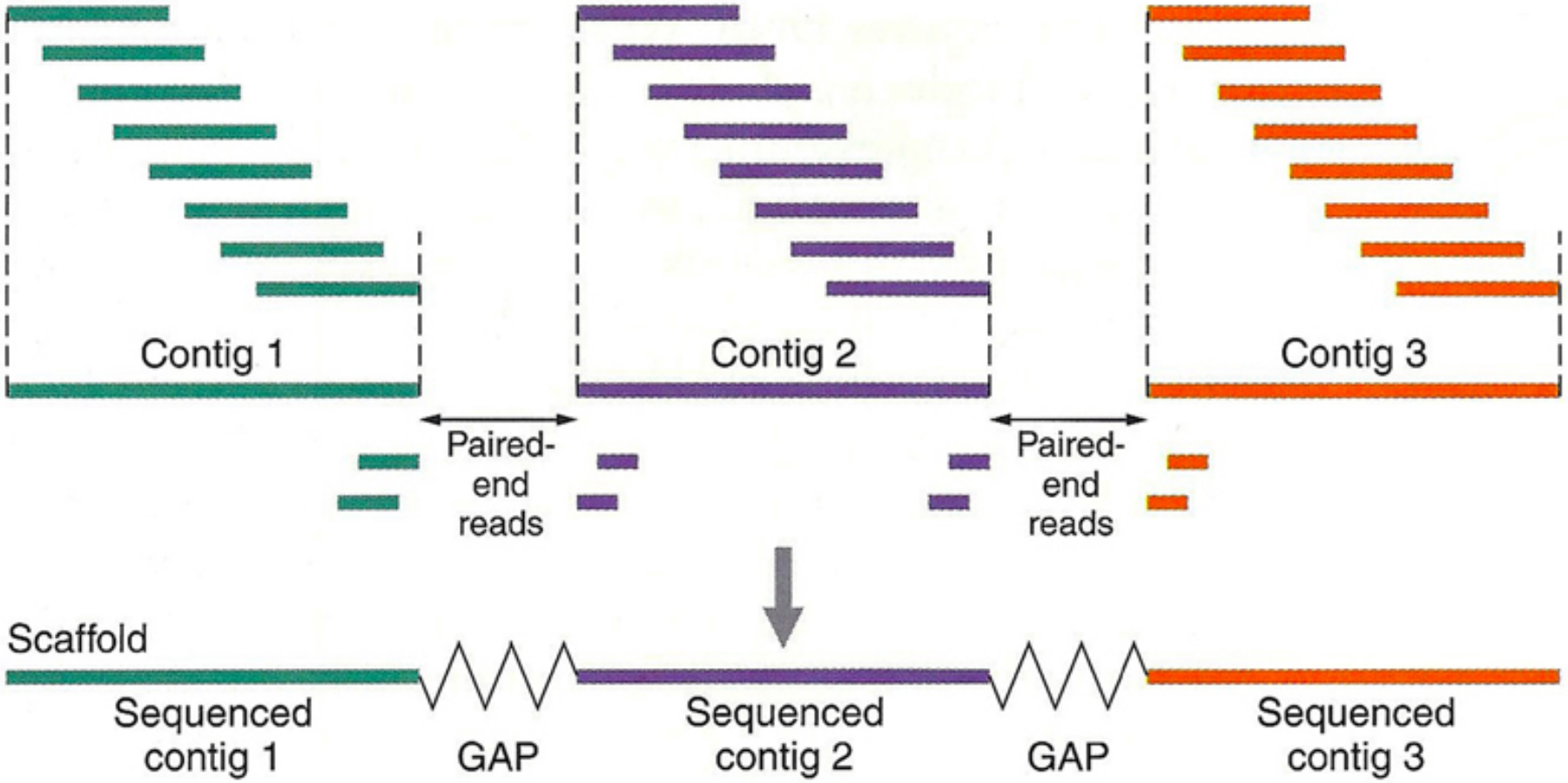


Pre-processing - Error Correction

- Musket - Kmer spectrum based correction
- EC helps trim down size of De Bruijn Graph - Less RAM used!
- EC improves assembly quality - better assembly!



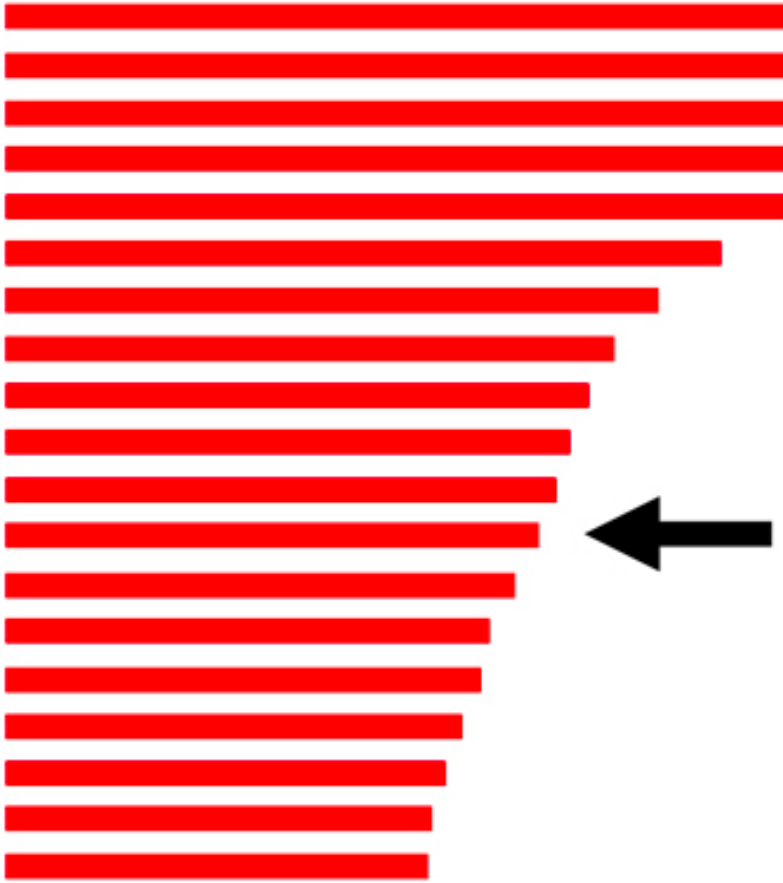
Analysis | (Super)contigs



Analysis | Assembly Quality

- **N50**
  - Statistical measurement of assembly quality
  - Measures average length of a set of sequences
  - Defined as length of the contig for which 50% of the entire genome are in contigs this size or greater
  - i) Order X > x ; ii) Cumulative sum ; iii) Contig length for which 50% of genome is present
  - Contigs & Scaffolds

Analysis | N50



Human genome = 3.2 Gbp

N50 = 46,395,641 bp

Sum = 1.6 Gbp

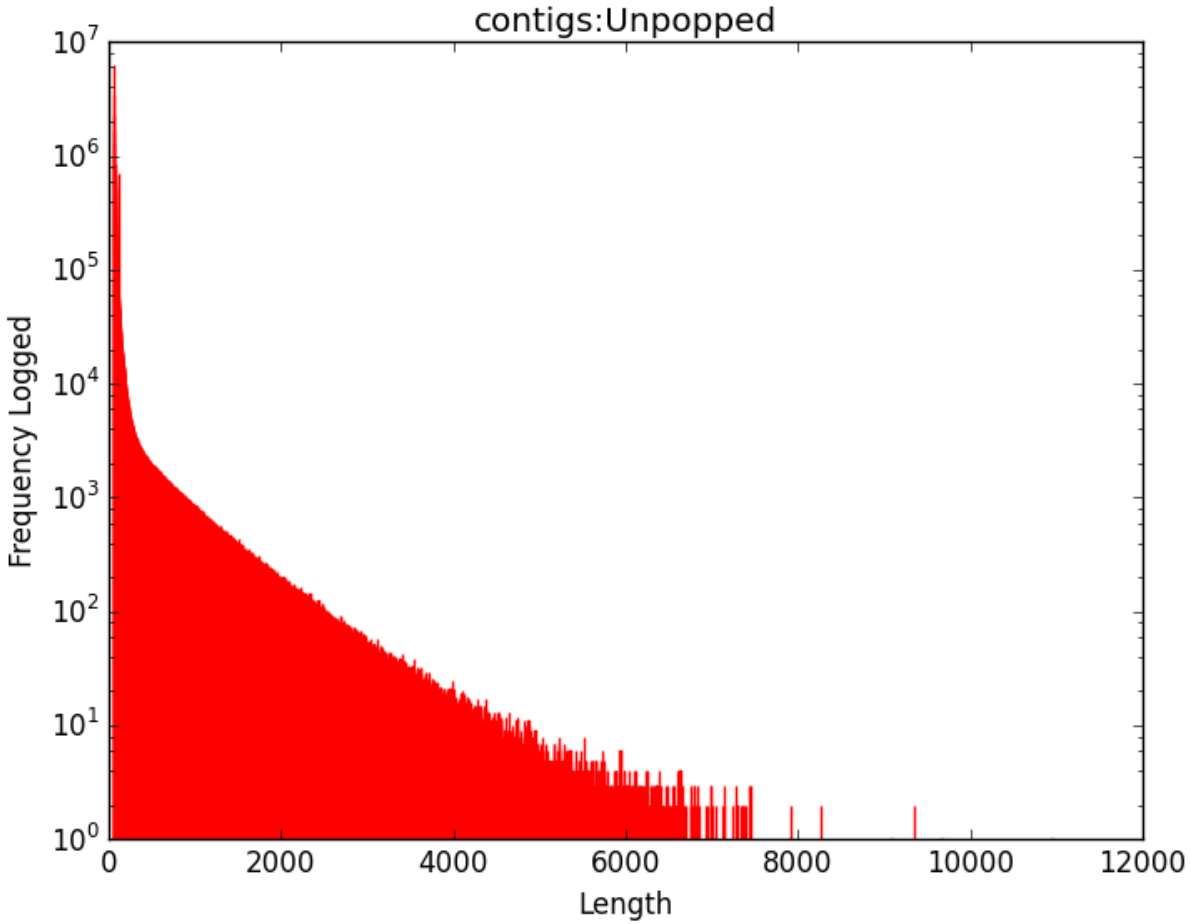
Meraculous

- Published by the US Department of Energy Joint Genome Institute
- Originally designed exclusively for haploid assembly, has since been modified
- Relies on an abundance of data, eschewing error correction
- Uses very conservative assembly methods

Meraculous

- Installation and usage was difficult
- Majority of time working with assembler was spent on debugging
- One of the critical steps in the pipeline consistently failed
- Results were not promising regardless of errors
- Assembler was eventually abandoned in favor of the team spending their time elsewhere

Meraculous


contigs:Unpopped
Frequency Logged
10^7
10^6
10^5
10^4
10^3
10^2
10^1
10^0
0
2000
4000
6000
8000
10000
12000
Length

SOAP denovo

- Short Oligonucleotide Analysis Package (SOAP) *de novo*
- Collection of alignment and assembly programs developed at Beijing Genomics Institute (BGI)
- Multi-threaded
- Adaptable, can use multiple kmer lengths:
  - Large kmer benefit (resolves large repeat regions)
  - Small kmer benefit (resolve errors, low coverage)
- Memory efficient

SOAP denovo

- Version 2 (2012) updates
- Reduces memory consumption
- Increases coverage and length in scaffold construction
- Improves gap closing
- Optimized for larger genomes and longer read datasets

SOAP denovo

- Final assembly using all paired-end and mate-pair libraries along with gap closure:
- Total bases: 2,293,889,058
- Contig N50: 8,033
- Scaffold N50: 11,000
- Wall time: ~20 hours running SOAPdenovo2 assembly and ~42 hours for running GapCloser
- Mem Usage: ~160 Gb max

Image taken from Luo, R. et al

SGA

- Memory Efficient
- CPU intensive in the preprocessing steps
- Modular
- Parallelizable
- Minimum Coverage 20-30X
- Recommended 40X
- Poorly documented

SGA

- Initial assembly completed
- Single dataset

- Poor assembly:
- Contig N50 ~ 180
- Cores - 32
- Wall Time - 12.3 Hours
- CPU Time - 10.96 Hours
- Mem Usage - 34.7 G

SGA

- Full assembly in progress
- Preprocessing steps completed! (majority of CPU time)
- Remainder of pipeline needs to be completed
- No foreseeable obstacles (other than available compute time)

ABySS

- Developed at Canada's Michael Smith Genome Sciences Centre
- Response to memory demands of conventional de Bruijn Graph assembly methods
- Very parallelizable
- Illumina recommended assembler for large genomes

ABySS

Pros:

- Good for large genomes
- Parallelizable -> faster
- Easy to run: encapsulated in Makefile
- Supposedly works well on a cluster

Cons:

- Hard to install and configure on our cluster
- Needs way too much memory

ABySS

- Partial success: assembly with subset of data
- >10 million contigs
- N50 contig:
  - size: 2,669
  - index: 174,507
- max contig size: 31,605
- assembly size: 1.6G
- Used 141GB of RAM
- Ran in 2 days

ABySS

- Run with all data: failure
  - Huge issues getting program running on cluster
  - Never enough free memory
- Different server, more memory
  - Reasonable kmer size -> not enough memory
  - Much smaller kmer size -> enough memory, failed anyway

Discovar *de novo*

- Developed by the Broad Institute (MA)
  - http://www.broadinstitute.org/software/discovar/blog/
- Relatively easy to install
- Easy to use - set wet and dry lab protocols
- Set wet lab protocol involving PCR-free amplification, 2x250 reads on an ~450bp insert, and ~60X coverage.
- Only supports Illumina sequencing platforms

Team members:

Natasha Dudek, Gepoliano Chaves, Chris Eisenhart, Robert Calef

Discovar *de novo*

User experience:

- Relatively easy to install and use
- No error correction required
- Assembly started with a single command:

```
DiscovarDeNovo READS = sample : H19 :: UCSF_SW019_noAdap_noDup.bam +
sample : M19 :: SW019_MiSeq_adapterTrimmed_dupRemoved.bam + sample:tag
:: BS_tag_noAdap_noDup.bam NUM_THREADS=22 MAX_MEM_GB=260
MEMORY_CHECK=True OUT_DIR=fullMergableAssembly/
```

Discovar *de novo*

Results:

- Our current best assembly results:

Contig N50: 9.5 kb

Scaffold N50: 10.6 kb

Total bp present: 2.24 Gb

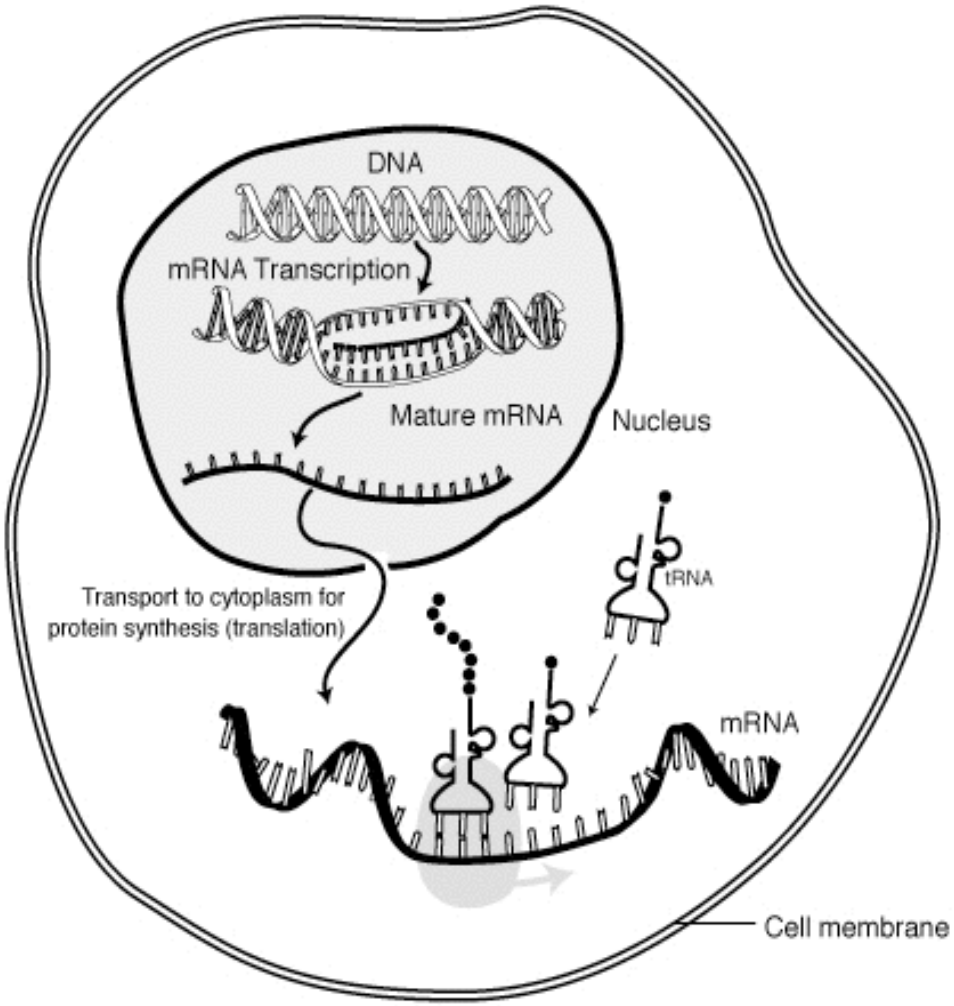
Bases in 1 kb+ scaffolds: 1.85 Gb

Libraries used: UCSF SW019, MiSeq SW019, BS-tag

- Final assembly pending

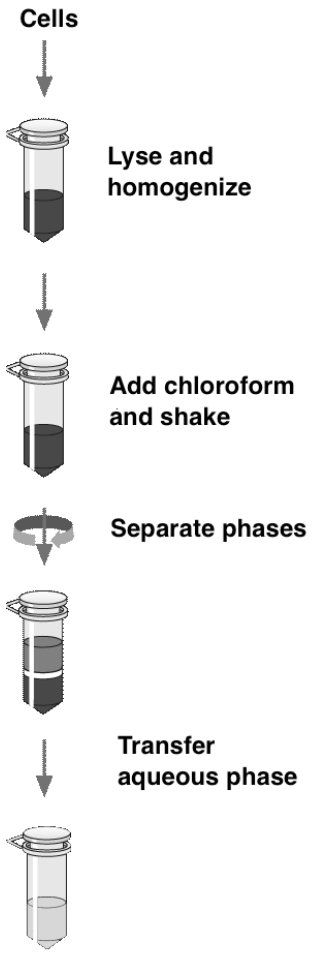
RNA-Seq

- The flow of information is from DNA genes to mRNA to protein.
- RNA-seq allows you to locate genes in the genome.

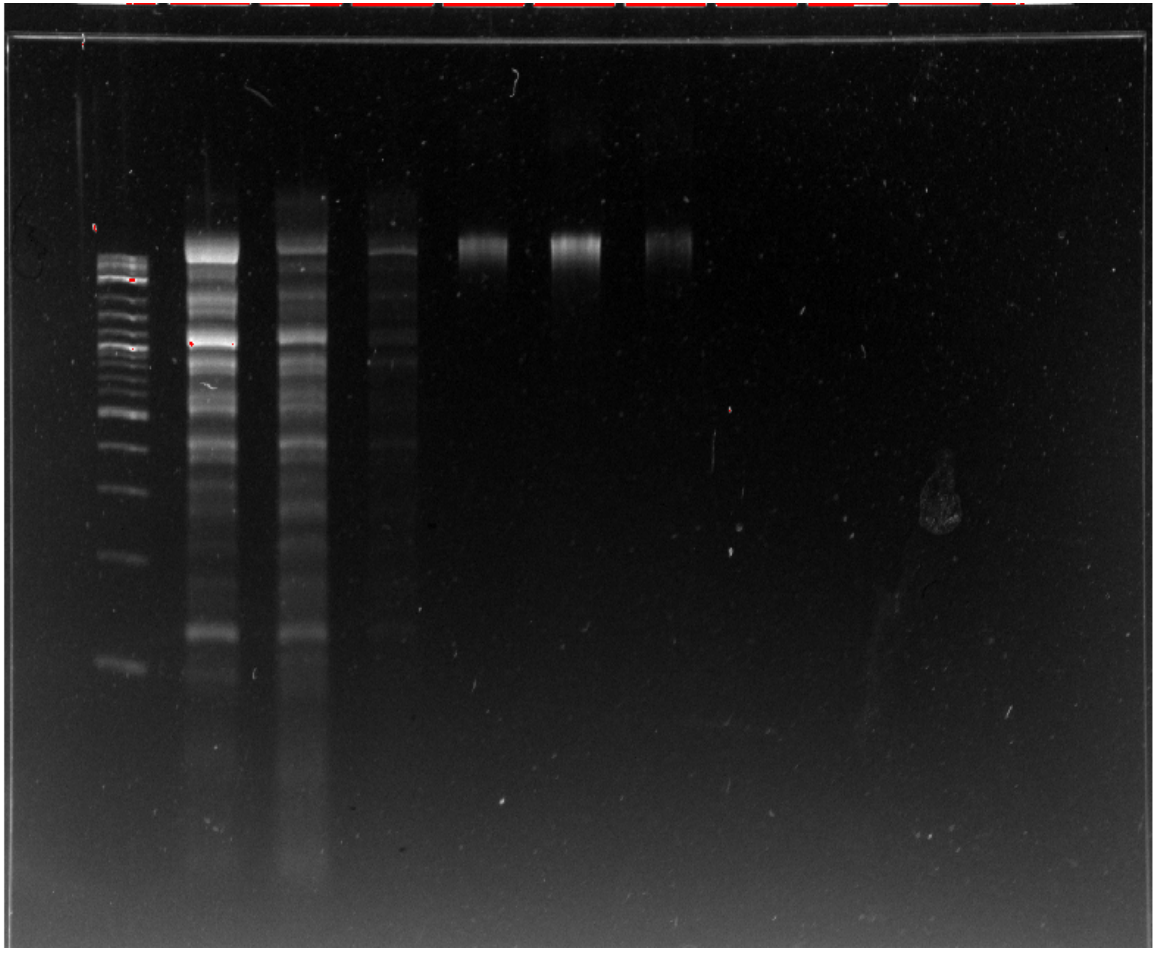


RNA Extraction

- Tissue samples are flash frozen and ground to a powder.
- The powder is mixed with a phenol solution which dissolve the tissues.
- The mixture is mixed with chloroform which causes the protein/carbohydrates and DNA to be separated from the aqueous phase containing the RNA.

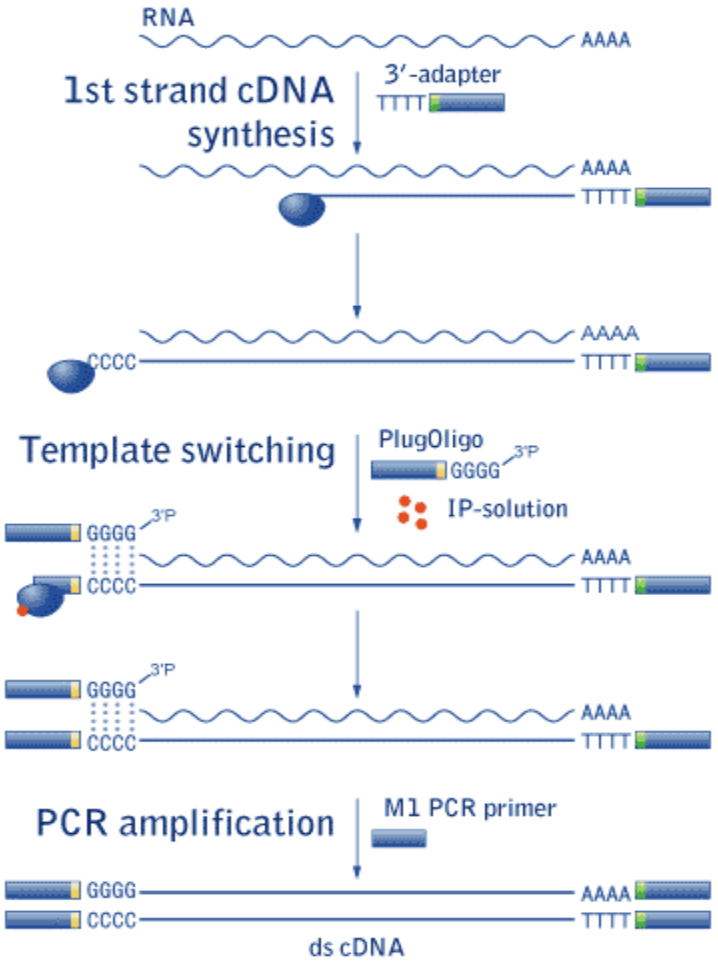


RNA Extraction

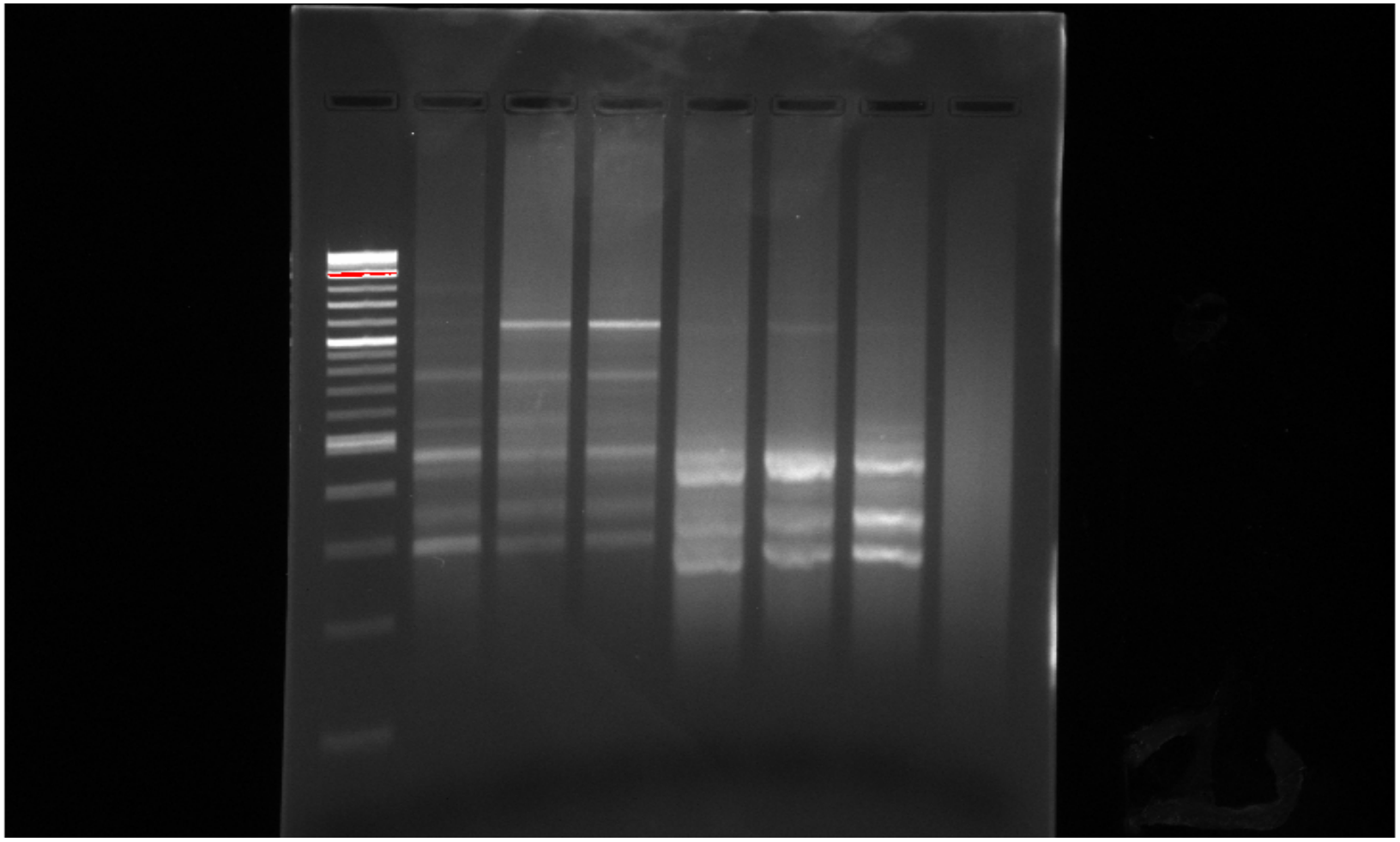


cDNA synthesis

- RNA is too fragile to sequence directly, so it must be reverse-transcribed into DNA.
- mRNA is enriched by using poly-T primers.

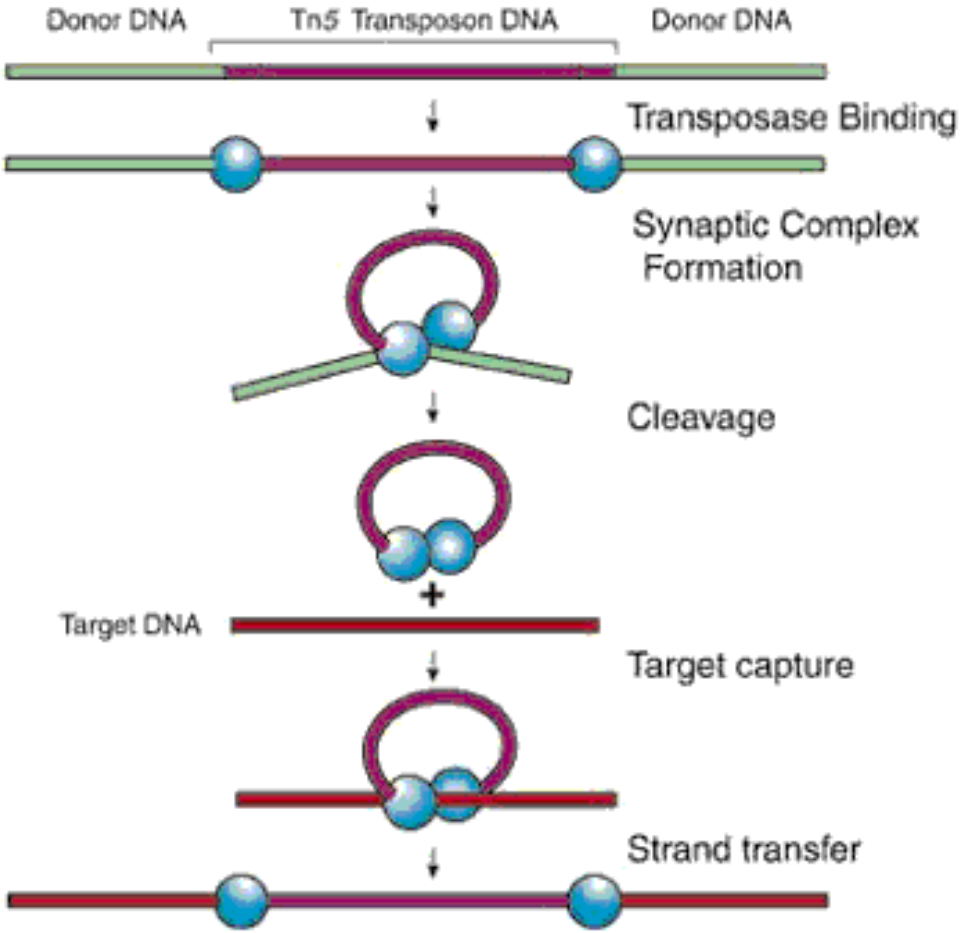


cDNA Synthesis



Adapter Ligation

- Sequencing adapter are added to the cDNA with the Tn5 enzyme.
- tagmentase also shears the DNA in the process.



Libraries

- 6 RNA-Seq libraries for three separate tissues were constructed using a tagmentase protocol to attach linkers to cDNA.
- Three libraries use a 5’ capture protocol which enriches for the 5’ UTR and were size selected to be >= 500 bp
- Three libraries were created using a general tagmentase protocol and were size selected to be >= 300 bp

Repeats

- Repark is being used to assemble the repetitive elements of the genome
- Repark uses Jellyfish to count kmers, then selects all kmers above a certain frequency to be assembled into contigs using velvet.
- the frequency is selected so that it excludes the peak containing non-repetitive kmers.

Mitochondrion

Assembly	Total bases	# scaffolds	Longest scaffold	Scaffold N50	Coverage
2012	23,642	1	-	-	Ranges from 20-2300X
2015	46,248	110	12,883	12,883	Avg of 411X

- The majority of the mitochondrion appears to be in two scaffolds separated by repeats
- Next steps: PCR amplify gaps and send products for Sanger sequencing
- COX1 gene sequence has been assembled and extracted

UCSC Genome Browser

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Ariolimax dolichophallus (*Ariolimax dolichophallus*) Genome Browser Gateway

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.

group	genome	assembly	position	search term
Mollusca	Ariolimax dolichophallus	Discovar de novo	flattened_line_4:1-107,068	enter position or search terms (submit)

Click here to reset the browser user interface settings to their defaults. ***More on-site workshops available!***

add custom tracks | track hubs | configure tracks and display

Ariolimax dolichophallus Genome Browser – DiscovarDeNovo assembly (sequences)

This assembly was generated by the Discovar de novo team in the UCSC BME235 class. Contact Chris Eisenhart with questions. If you are interested in how this assembly was made please see the Discovar de novo team page

This hub is for viewing the assemblies of the banana slug, Ariolimax dolichophallus. This hub was generated for the UCSC BME235 class, contact Chris Eisenhart to have new information put on the hub (ceisenhart@soe.ucsc.edu). More information can be found on the class webpage, BME235 class webiste

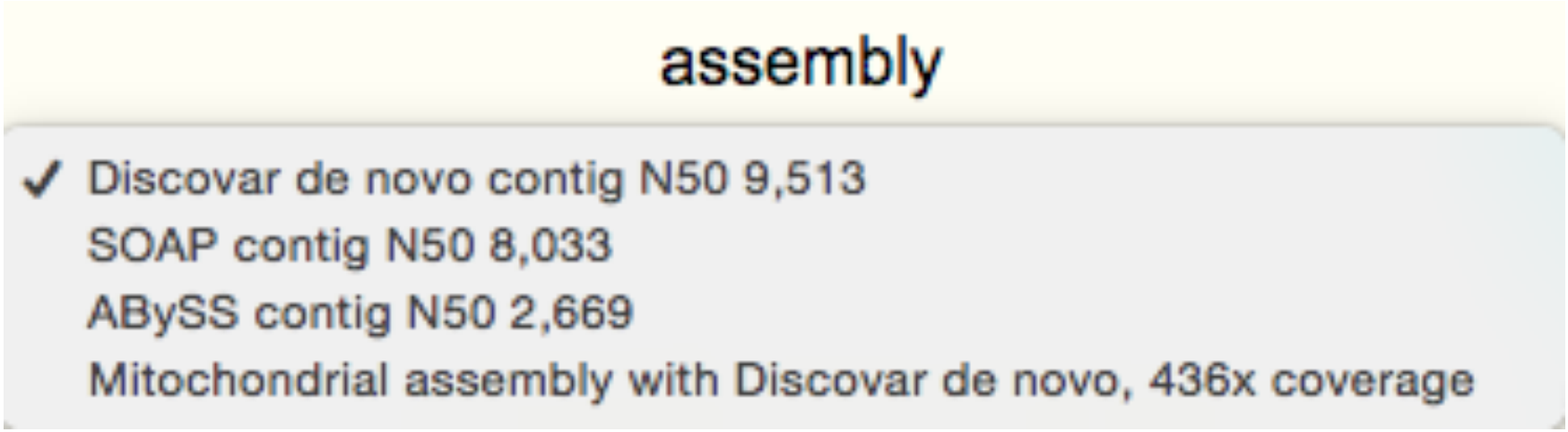
Ariolimax dolichophallus

UCSC Genome Browser

-Four assemblies are currently supported

-Best assembly for each group determined by N50

-Mitochondrial assembly with the highest coverage



UCSC Genome Browser

Ariolimax dolichophallus Genome Browser – SOAP assembly (sequences)

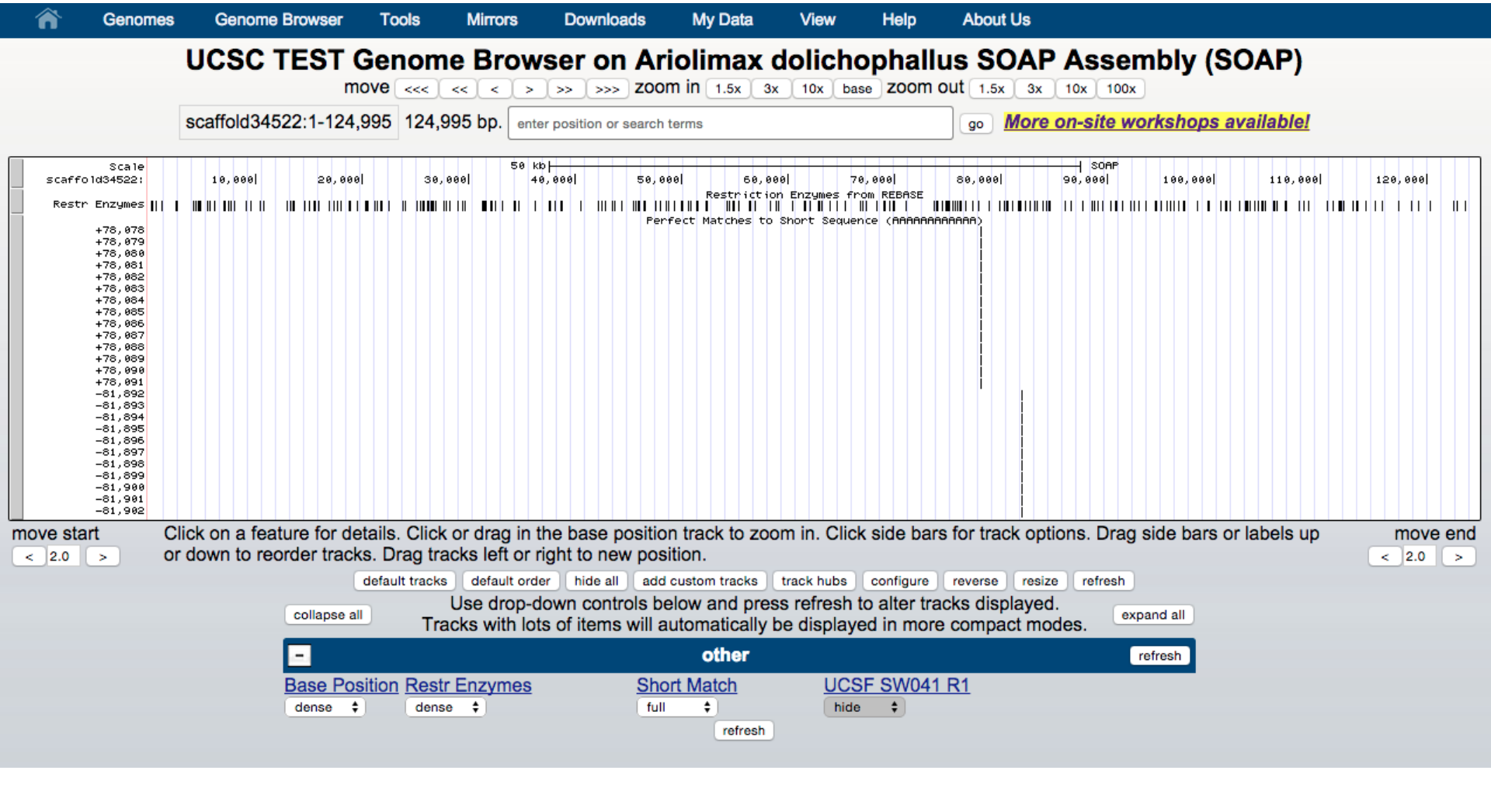
This assembly was generated by the SOAP team in the UCSC BME235 class. Contact Charly Markello (cmarkell@ucsc.edu) with questions. If you are interested in how this assembly was made please see the SOAP team page

This hub is for viewing the assemblies of the banana slug, Ariolimax dolichophallus. This hub was generated for the UCSC BME235 class, contact Chris Eisenhart to have new information put on the hub (ceisenhart@soe.ucsc.edu). More information can be found on the class webpage, BME235 class webiste

Links to the wiki

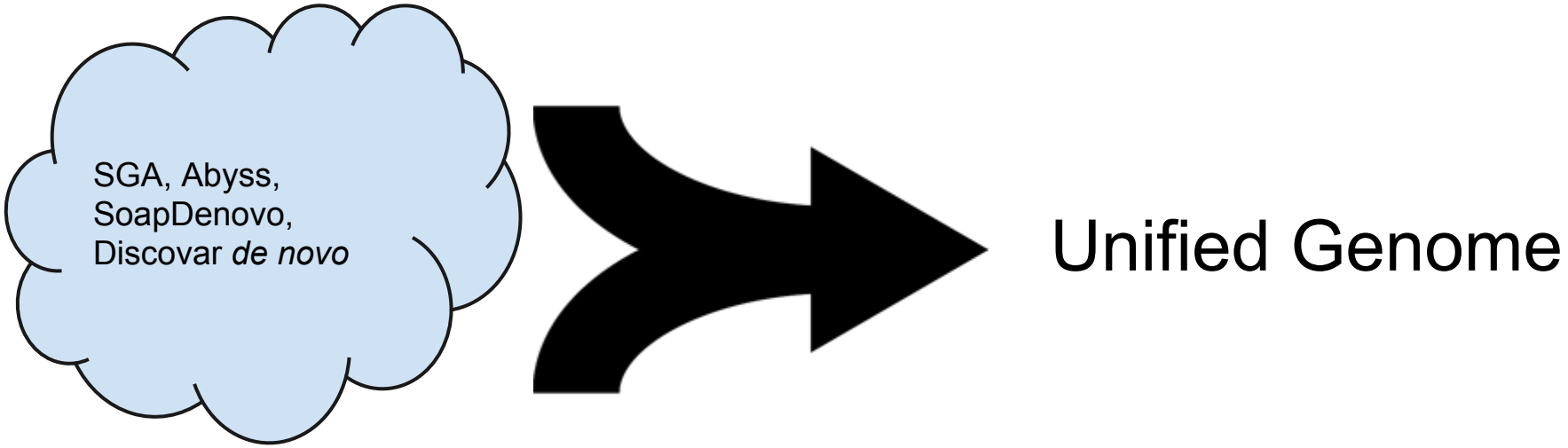
Ariolimax dolichophallus

UCSC Genome Browser



Next Steps - Merging

- Picking the best assemblies for merging
- Cleaning up merged assemblies



Next Steps - Usability

- Making the genome useful for further research
  - Annotation of genes
    - RNA-SEQ
    - *de novo* Prediction
  - Finding repeat regions
- Publishing our findings
  - Novel discoveries
  - Highlighting the state of the genome