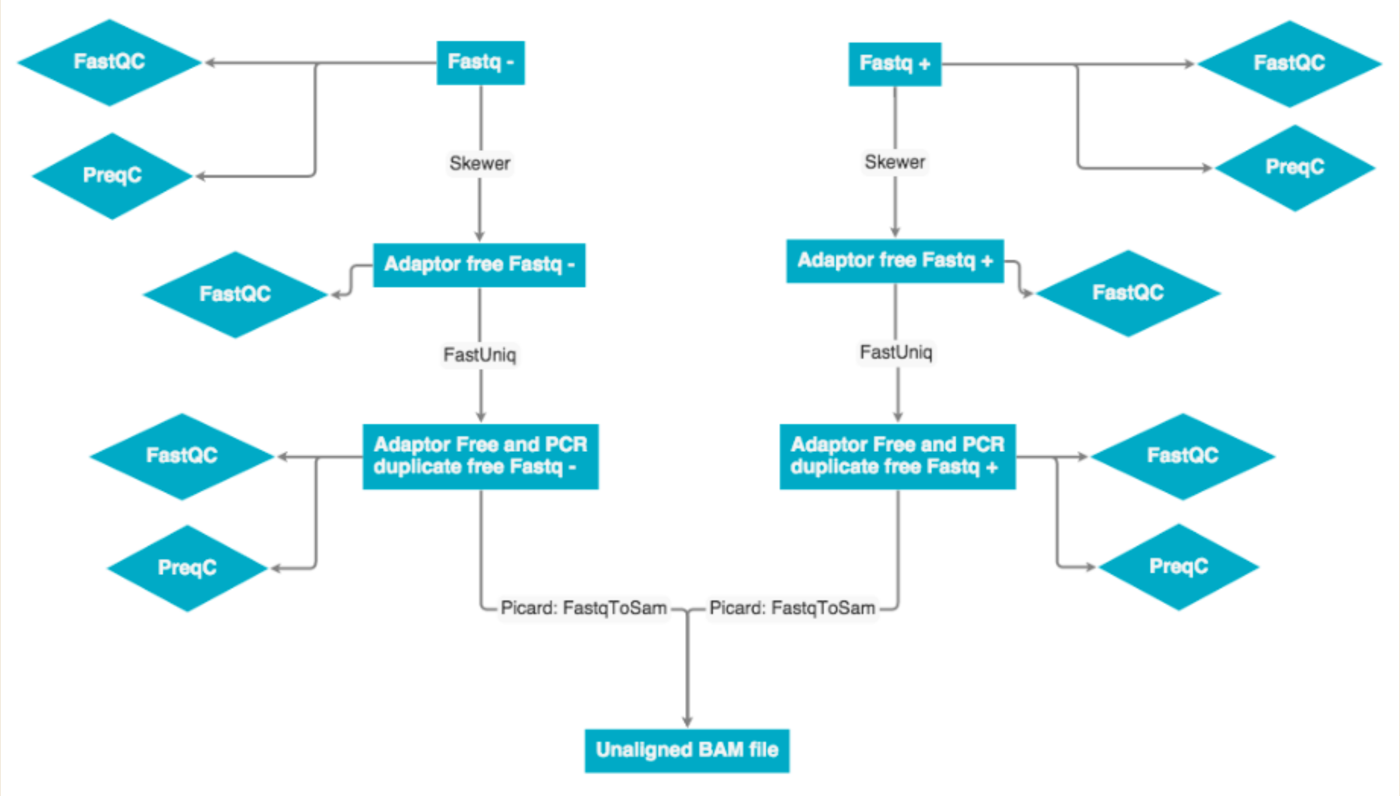
Assembly of the *Ariolimax dolicophallus* genome with Discovar *de novo*

Chris Eisenhart, Robert Calef, Natasha Dudek, Gepoliano Chaves

Discovar *de novo*

- Developed by the Broad Institute in 2014
- Specific wetlab workflow expected
- Illumina libraries only
  - PCR free
  - Insert size ~450bp
  - Read length >250
  - 60X coverage

Preprocessing reads

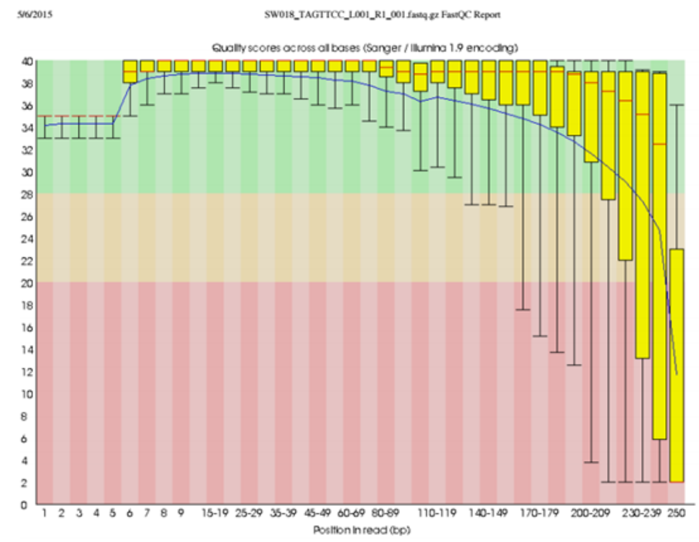


Preprocessing: adapter removal

Quality scores across samples

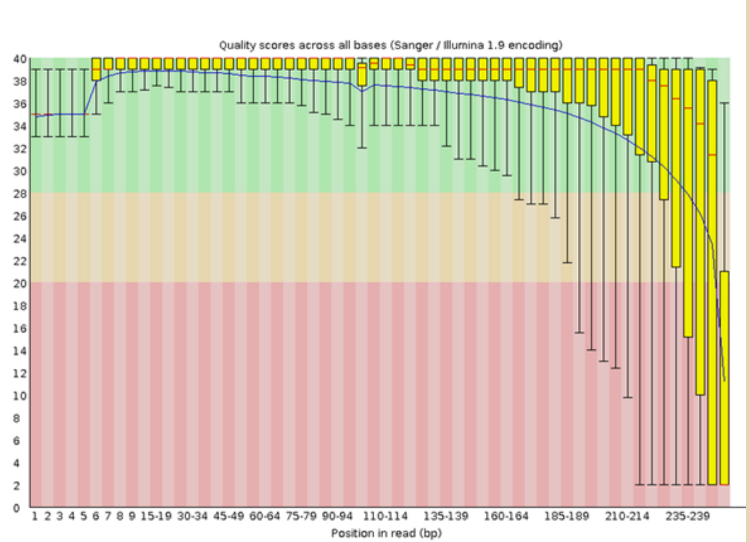
SW018_TAGTTCC_L001_R1_001.fastq FastQC Report

a) Pre-skewer



SW018_noAdap_R1.fastq

b) Post-skewer

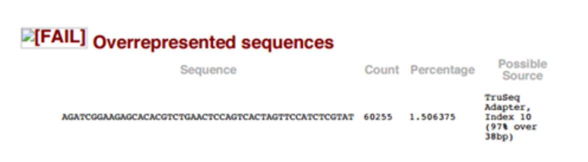


Preprocessing: adapter removal

Quality scores across samples

SW018_TAGTTCC_L001_R1_001.fastq.gz FastQC Report

a) Pre-skewer

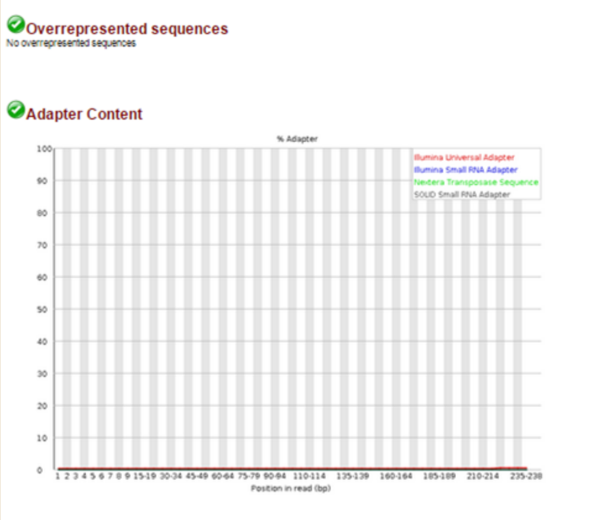


[FAIL] Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACTAGTTCCATCTCGTAT	60255	1.506375	TruSeq Adapter, Index 10 (97% over 38bp)

SW018_noAdap_R1.fastq

b) Post-skewer



✓ Overrepresented sequences

No overrepresented sequences

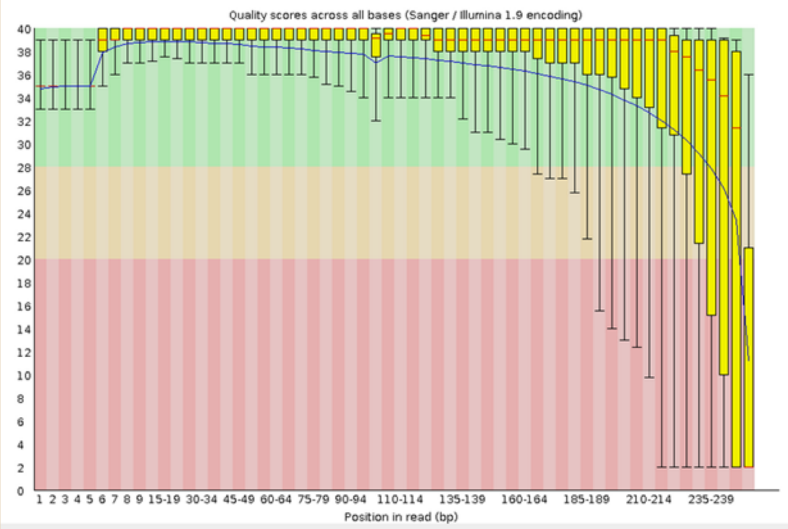
✓ Adapter Content

Preprocessing: PCR-duplicate removal

Quality scores across samples

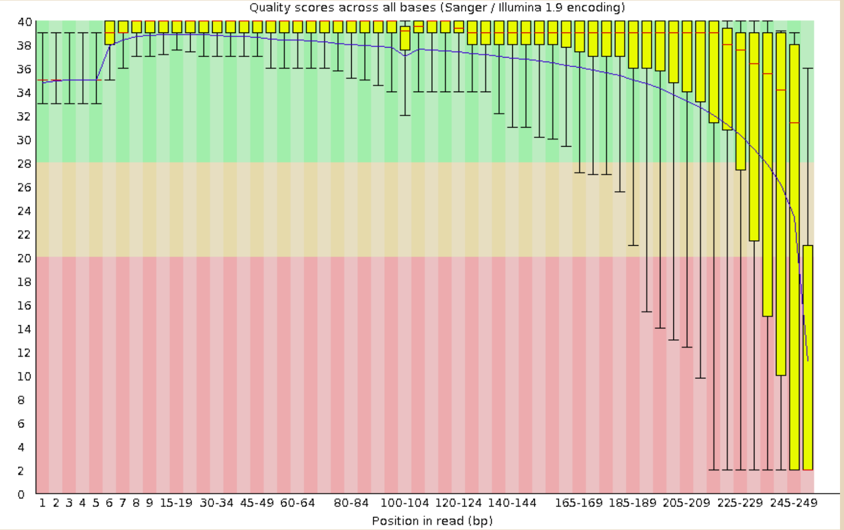
SW018_noAdap_R1.fastq

a) Pre-fastUniq



SW018_noAdap_noDupR1.fastq

b) Post-fastUniq



Preprocessing: PCR-duplicate removal

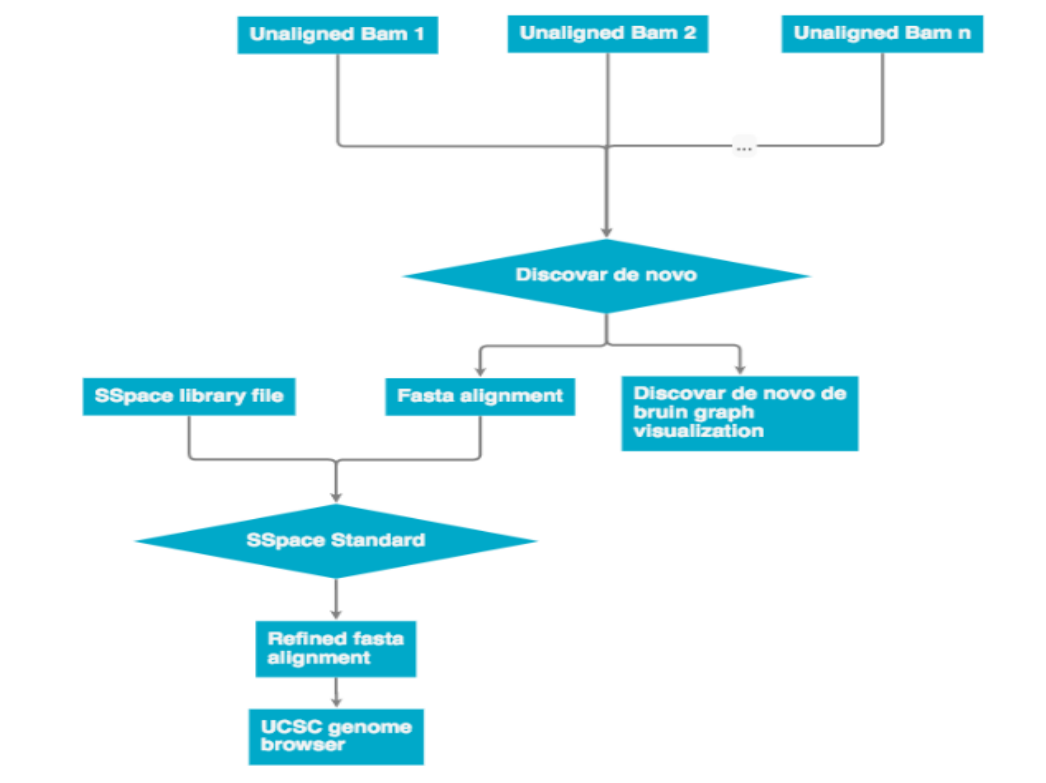
Dataset	# Reads before fastUniq	# Reads after fastUniq	% duplicates
SW018	61,132,697	56,931,153	6.87%
SW019	79,215,987	75,081,065	5.21%

Picard tools: convert to BAM/SAM

After adapter and duplicate removal, reads are converted to BAM files

- Overlapping and non-overlapping reads in BAM format are used as input for Discovar de novo

Discovar *de novo*



Unaligned Bam 1

Unaligned Bam 2

Unaligned Bam n

...

Discovar de novo

SSpace library file

Fasta alignment

Discovar de novo de bruin graph visualization

SSpace Standard

Refined fasta alignment

UCSC genome browser

Running Discovar *de novo*

Input syntax:

-Last time: frac option for downsampling, no white space

-This time: threads and memory

```
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/
```

Running Discovar *de novo*

Thread control: NUM_THREADS

-Available since October 2014

-Hyperthreading: Disable, or set threads to # of physical cores

```
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/
```

Running Discovar *de novo*

Memory control: MAX_MEM_GB

-Available since December 2014

-Throttles maximum memory usage, not airtight

```
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/
```

Running Discovar *de novo*

Memory control: MEMORY_CHECK

-Available since February 2015

-Sequential malloc’s to determine available memory

```
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/
```

Running Discovar *de novo*

Memory control: malloc parallelization

-To avoid blocking, have multiple malloc heaps

-Not always default:

`export MALLOC_PER_THREAD=1`	bash
`setenv MALLOC_PER_THREAD 1`	csh

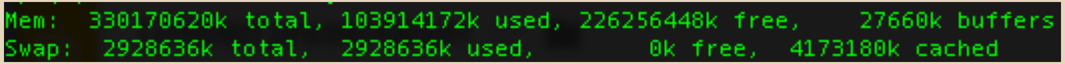
```
DiscovarDeNovo READS = sample : H19 :: UCSF_SW019_noAdap_noDup.bam +
sample : M19 :: SW019_MiSeq_adapterTrimmed_dupRemoved.bam + sample:tag ::
BS_tag_noAdap_noDup.bam NUM_THREADS=22 MAX_MEM_GB=260 MEMORY_CHECK=True
OUT_DIR=fullMergableAssembly/
```

Running Discovar *de novo*

Memory control: top vs htop

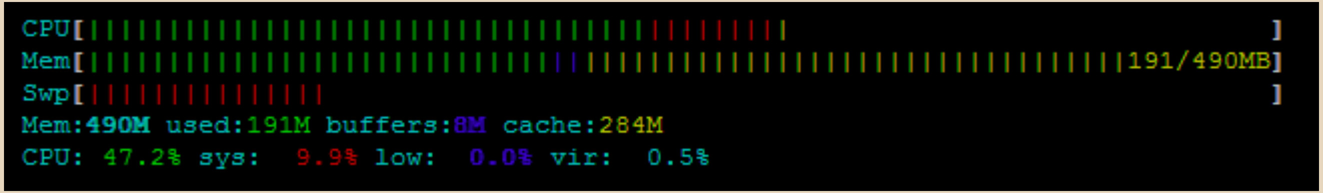
-Linux task managers, must download or compile htop

top



```
Mem:   330170620k total, 103914172k used, 226256448k free,     27660k buffers
Swap:   2928636k total,   2928636k used,         0k free,   4173180k cached
```

htop



```
CPU[||||||||||||||||||||||||||||||||||||||||||||                              ]
Mem[|||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||191/490MB]
Swp[|||||||||||||||                                                           ]
Mem:490M used:191M buffers:8M cache:284M
CPU: 47.2% sys:  9.9% low:  0.0% vir:  0.5%
```

-paging cache shown in yellow

Assembly runs

	50% Run	50% UCSF Run
Input bases	31.6 Gb	31.9 Gb
Input reads	263,835,400	132,012,218
Avg read length	120	242
Avg base quality	36.7	33.0

Downsampled runs:

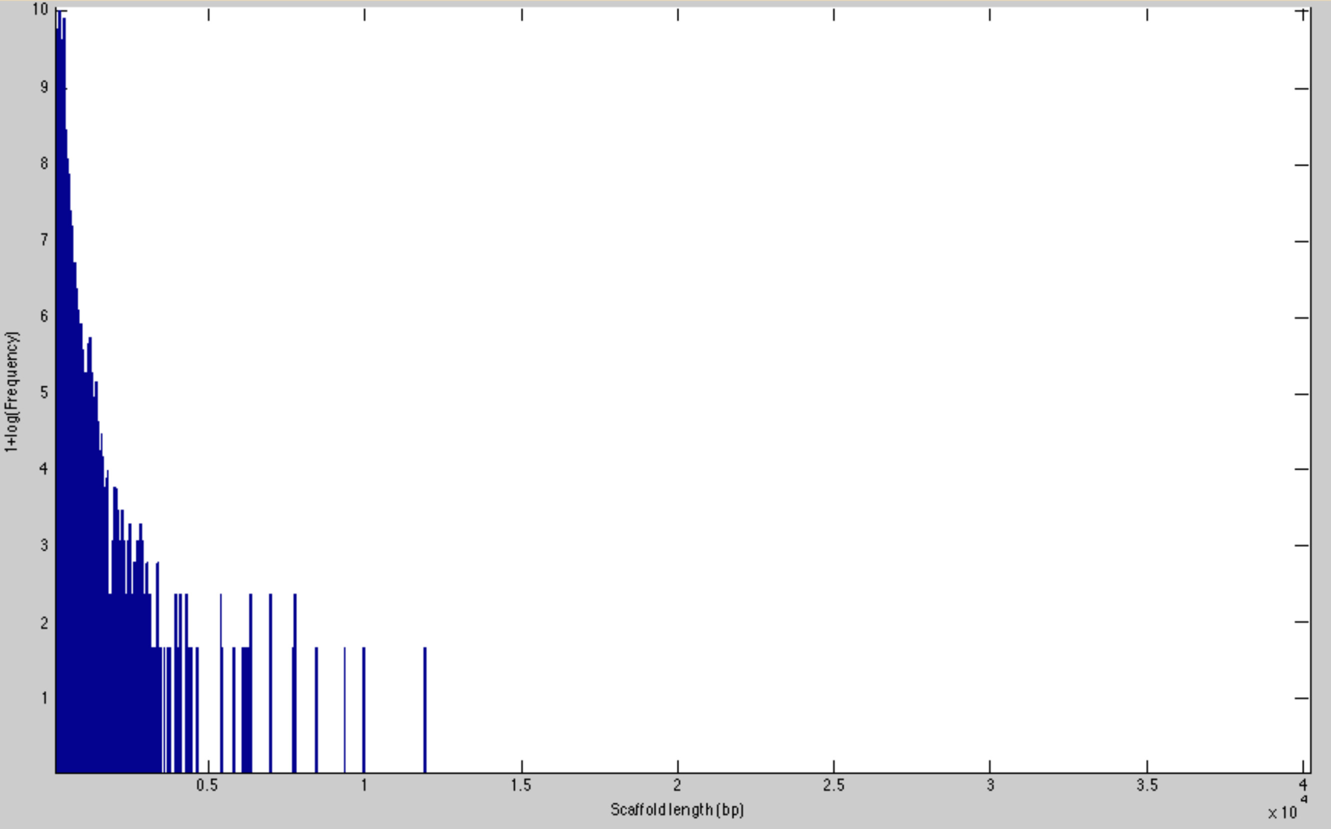
1. 50% of 2x100bp SW018/19, and MiSeq SW019 data
2. 50% of 2x250bp SW018 and SW019 data (UCSF)

Assembly results

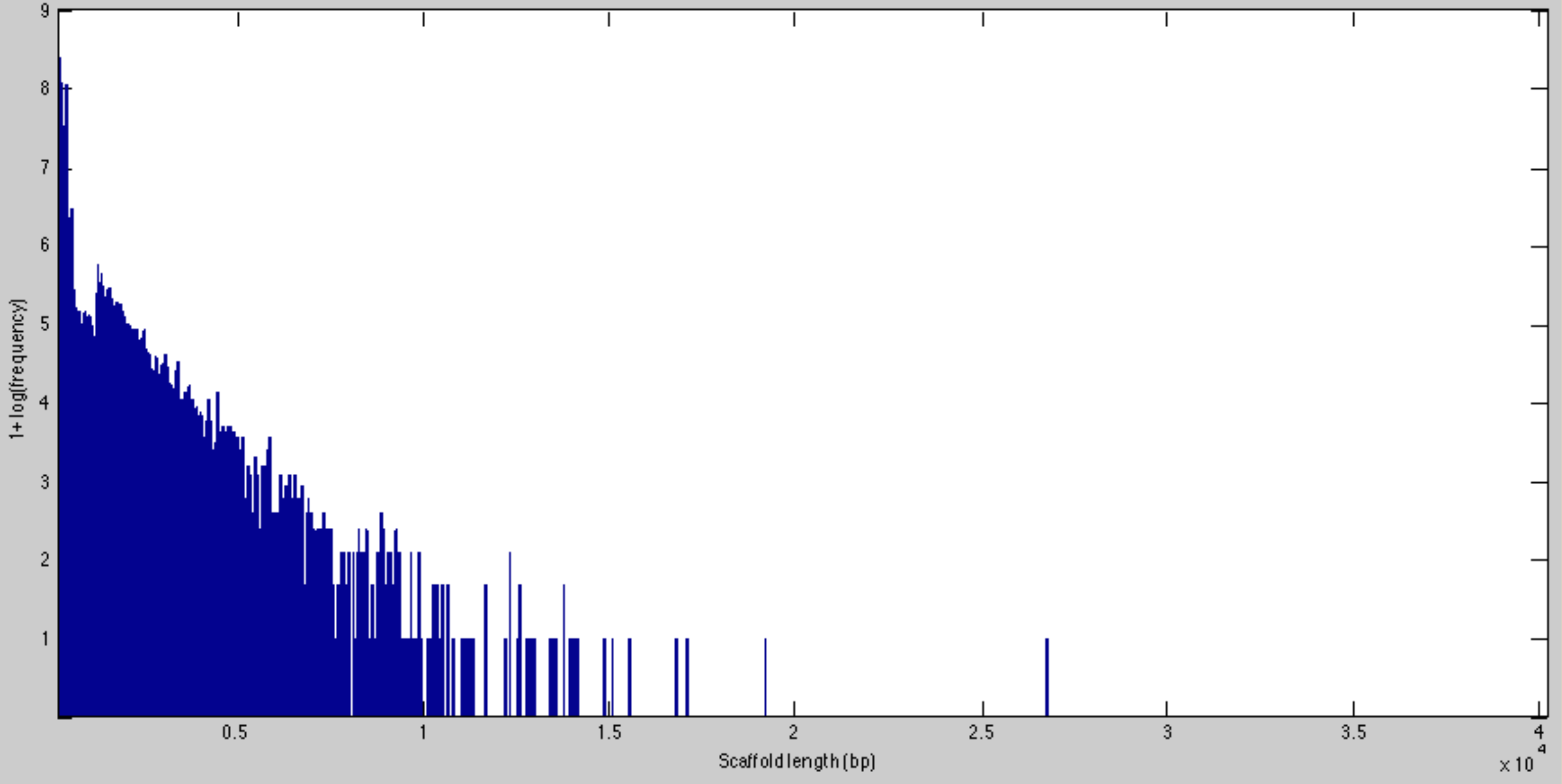
	50% Run	50% UCSF Run
Input bases	31.6 Gb	31.9 Gb
Input reads	263,835,400	132,012,218
Avg read length	120	242

	50% Run	50% UCSF Run
Runtime (hrs)	8.53	14.9
Peak memory usage (GB)	220.11	184.09
Bases in 1+ kb scafs	101,397,871	1,528,625,509
Bases in 10+ kb scafs	151,417	137,959,107
Mean position of first error	7	156
Contig N50	1,489	3,979
Scaffold N50	1,489	3,979

Scaffold histogram with 2x100 reads



Scaffold histogram with 2x250 reads



Assembly results: BLAST, 1st longest scaffold

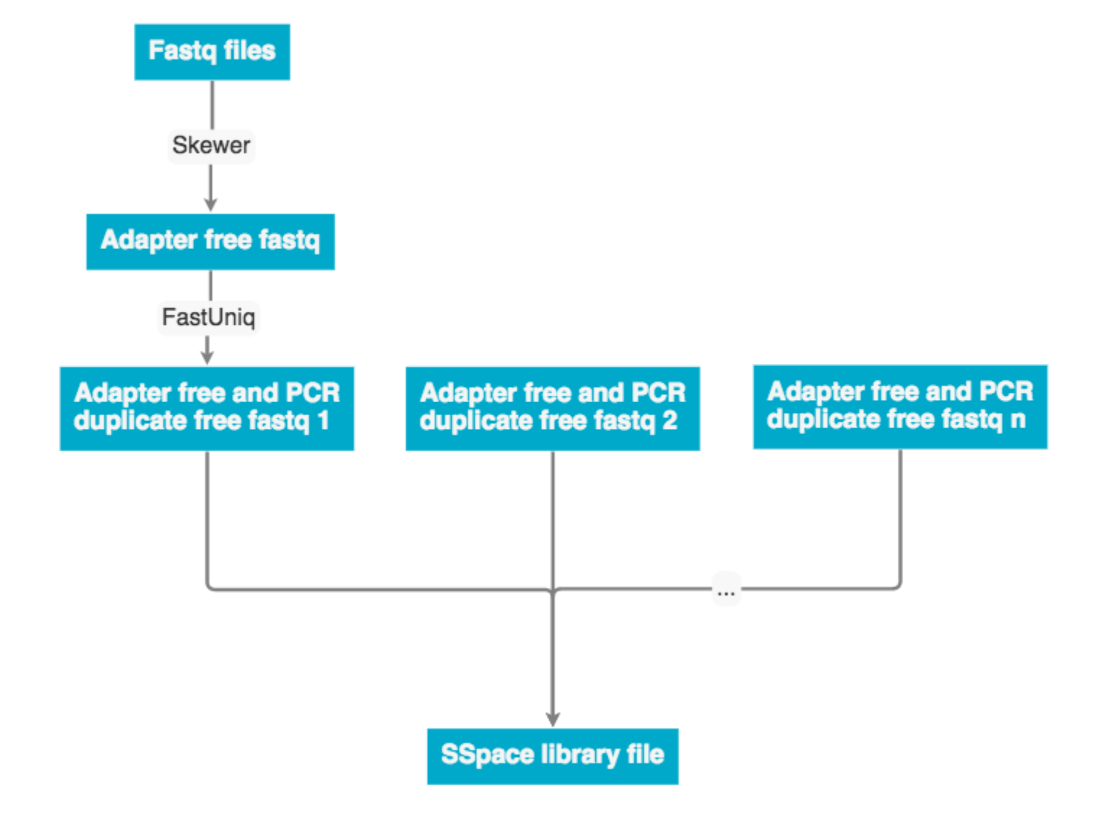
34857nt

Species	**Common name/ known feature**	**E-value**	**Identity**	**Notes**
Cicer arietinum	Chickpea	0.11	100%	Tryptophan synthase
Strongyloides papillosus	Parasitic nematode	0.4	100%	genome assembly S_papillosus_LIN, scaffold

Assembly results: BLAST, 8th longest scaffold

Species	**Common name/known feature**	**E-value**	**Identity**	**Notes**
Pred – Aplysia californica transcript variant X2	Sea hare	6e-19	83%	Transcript variant
Pred – Aplysia californica transcript variant X2	Sea hare	6e-19	83%	Transcript variant
Helix pomatia	Roman snail	4e-15	81%	Metallothionein gene

SSpace scaffolding



Fastq files

Skewer

Adapter free fastq

FastUniq

Adapter free and PCR duplicate free fastq 1

Adapter free and PCR duplicate free fastq 2

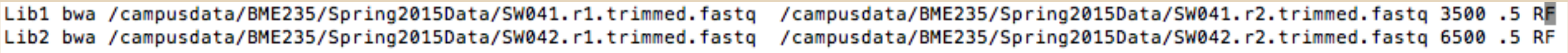
Adapter free and PCR duplicate free fastq n

...

SSpace library file

SSpace library file

- Lib file has a simple format
- One line for each mate pair library



```
Lib1 bwa /campusdata/BME235/Spring2015Data/SW041.r1.trimmed.fastq  /campusdata/BME235/Spring2015Data/SW041.r2.trimmed.fastq 3500 .5 RF
Lib2 bwa /campusdata/BME235/Spring2015Data/SW042.r1.trimmed.fastq  /campusdata/BME235/Spring2015Data/SW042.r2.trimmed.fastq 6500 .5 RF
```

- For future runs consider adding the Lucigen library

SSpace results

-No change in scaffold N50

-Change in number of ‘N’s

- 51,400 before scaffolding
- 169,013 after Lib 1
- 227,375 after Lib 2

2012 draft mitochondrion assembly

- Draft assembly by Kevin Karplus
- General method: iteratively mapping reads, reassembling
- May contain extra repeat regions
- Difficulty with high vs. low coverage regions

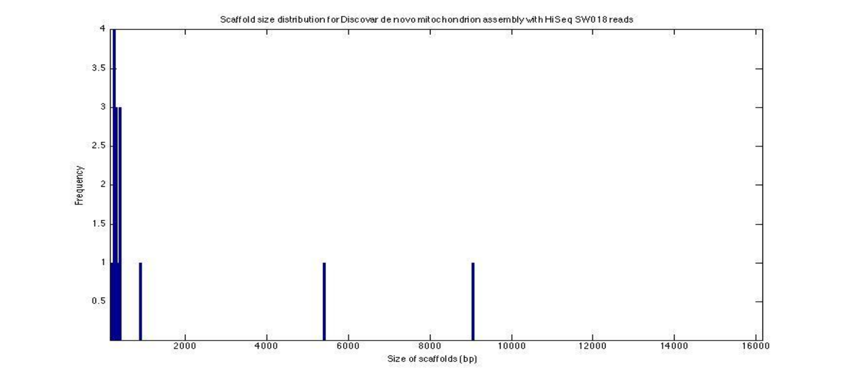
2015 Methods

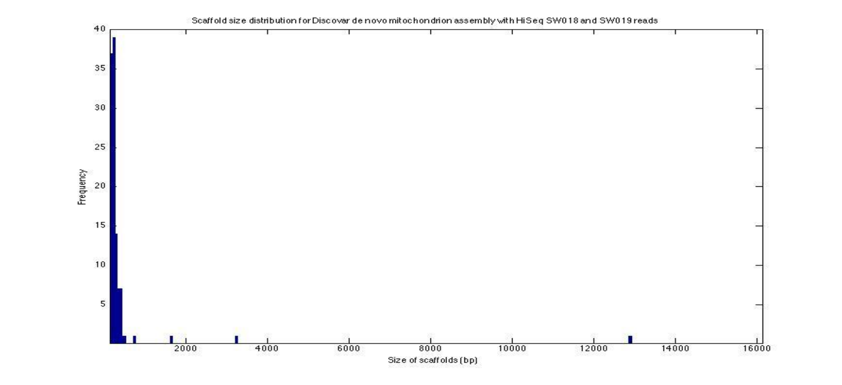
- Collected reads that mapped to the 2012 mitochondrion assembly
- Tried Discovar *de novo* using SW018 and SW019 HiSeq reads
- Currently trying Price using HiSeq SW018

Results

Assembly	Total bases	# contigs	# scaffolds	Longest scaffold	Scaffold N50	Scaffold N50	Bases in 1kb+ scaffolds	Bases in 10kb+ scaffolds	Coverage
Cepacea	14,100	-	-	-		-	-	-	-
Albanaria	14,130	-	-	-		-	-	-	-
Draft assembly	23,642	-	1	-		-	-	-	Ranges from 20-2300X
Discovar *de novo* SW018	18,983	29	29	9,041	9,041	9,041	14,048	0	Avg of 60X
Discovar *de novo* SW018 & SW019	46,248	110	110	12,884	12,883	12,883	17,173	12,684	Avg of 410.9X
Price SW018	20,106	25	-	2806	919	-	TBA	TBA	TBA

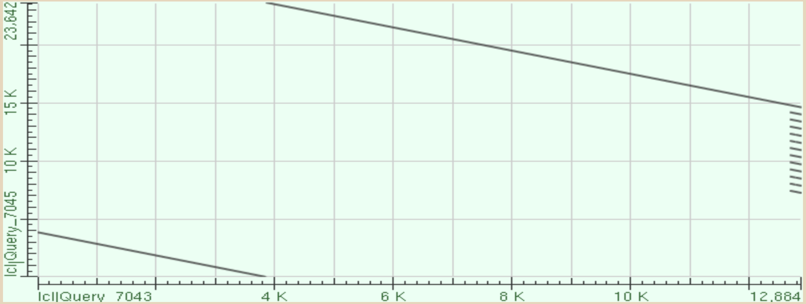
Distribution of scaffold sizes



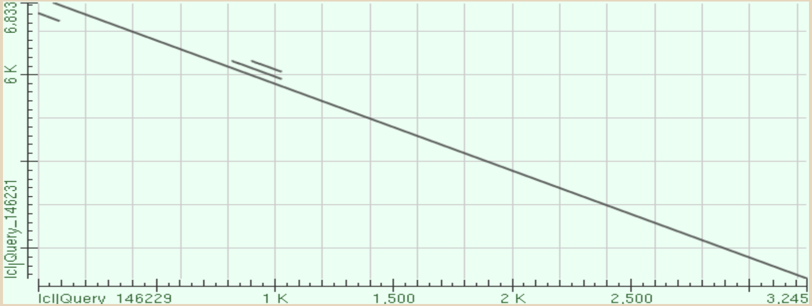


Discovar *de novo* vs. 2012 assembly

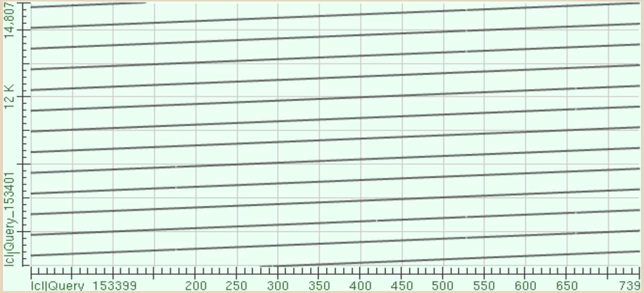
Query: biggest scaffold (12,884bp)



Query: second biggest scaffold (3,245bp)

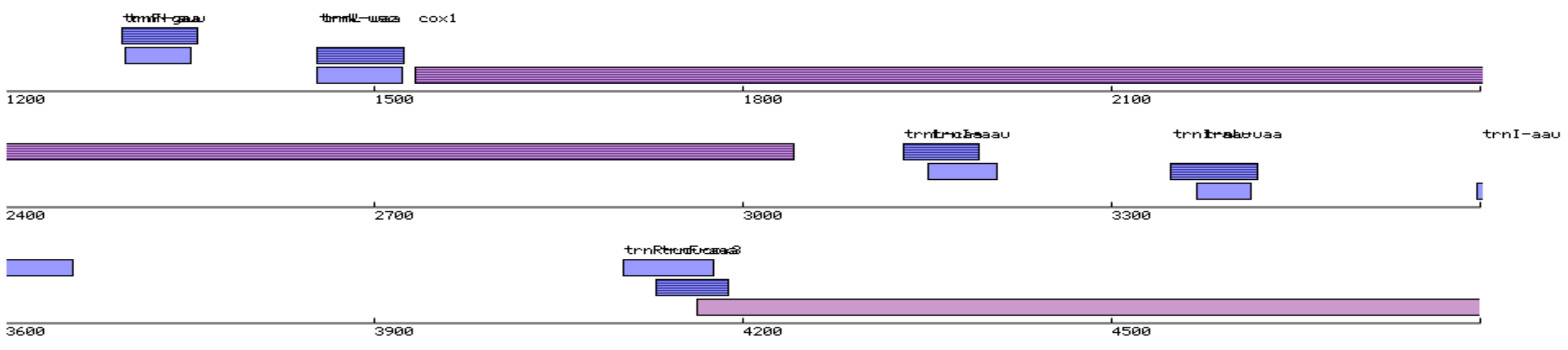


Query: short scaffolds (200-300bp)



COX1 gene sequence

- Standard barcoding gene
- Used blastn & DOGMA

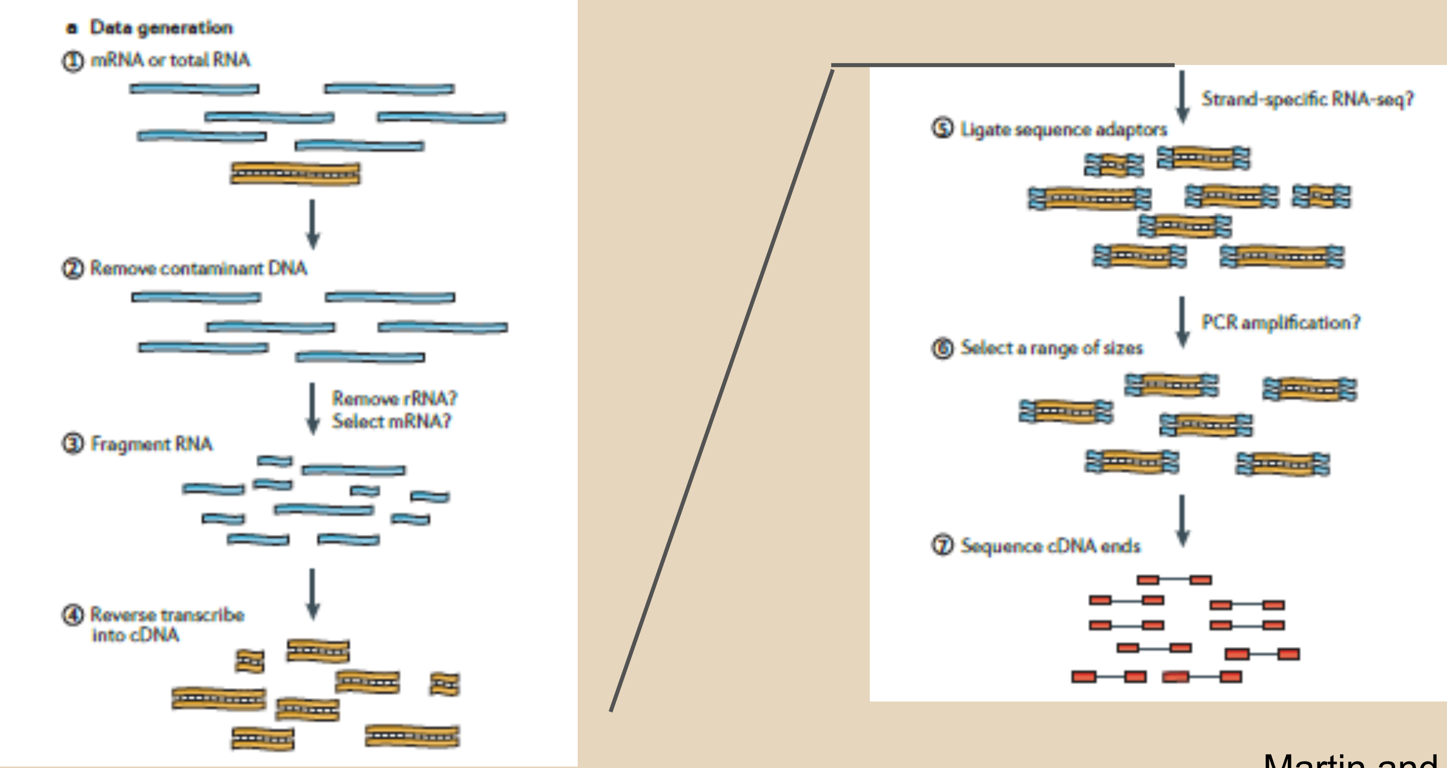


Did we really sequence *A. dollicophalus*?

No published mitochondrion sequences are available.

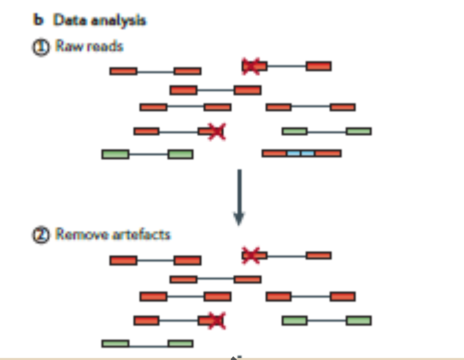
But we have sequenced the same organism as was sequenced in previous years.

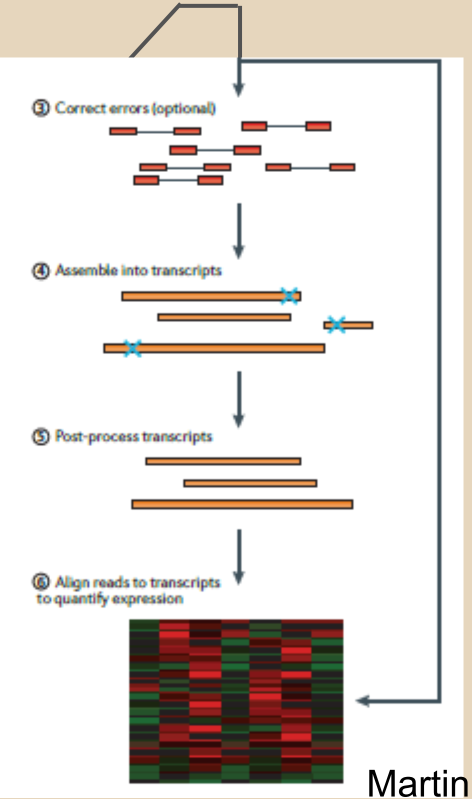
Next steps: RNA-seq recap



Martin and Wang, 2011

Next steps




③ Correct errors (optional)
④ Assemble into transcripts
⑤ Post-process transcripts
⑥ Align reads to transcripts to quantify expression

Martin and Wang, 2011

Next steps: de novo transcriptome assembly?

In Martin and Wang, 2011 the strategies to assembly the RNA-seq data include:

1) Generate all the substrings of length k
2) Generate the de Bruijn graph
3) Collapse the De Bruijn graph
4) Traverse the graph
5) assemble isoforms

Similar to assembling a whole genome

They mention that TransAbyss, Rnnotator and Multiple-k use the De Bruijn strategy to reconstruct transcripts

Martin and Wang, 2011

Next steps: de novo transcriptome assembly?

Bowtie, BMA and TopHat are programs used to align reads against a reference. Maybe we could:

- Compare gene expression tissue-specific (albumen, proximal albumen and penis)
- Compare slugs’ genes levels with other molluscs, using the scaffolds we generated.

Martin and Wang, 2011