Assembly of *Ariolimax dolichophallus* using SOAPdenovo2

Charles Markello, Thomas Matthew, and Nedda Saremi

Image taken by Troy A. Scott

Adapter Trimming 2nd Pass

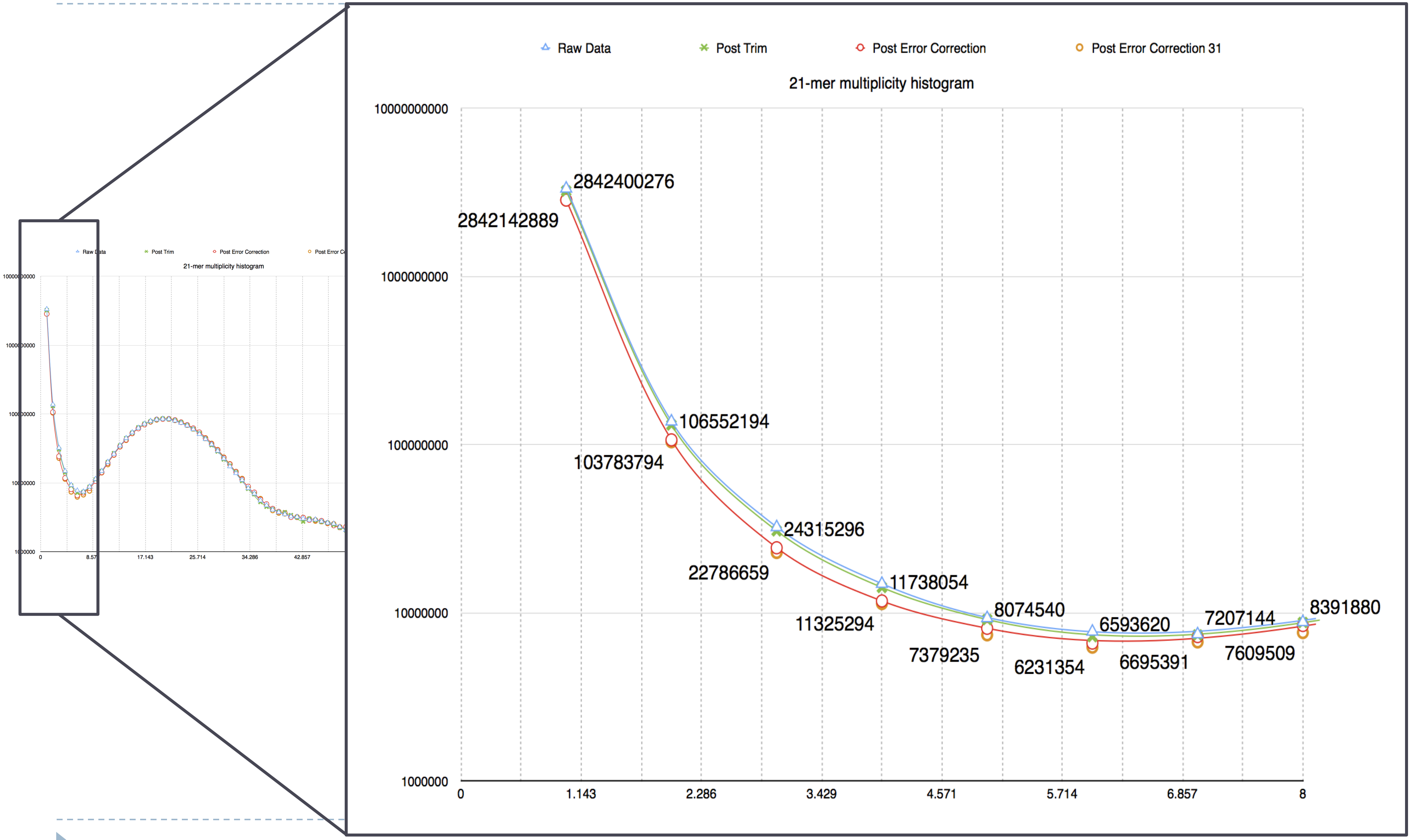
- Took another look at SW019_S1+2 and SW018_S1 reads.

- Confirmed presence of specified adapters primarily in the SW018 reads.
    - Overrepresented sequence matches adapter sequence.

- Ran fastqc before and after trimming to confirm if detected overrepresented sequence was removed.

- Did the same analysis with Team 4 run of SeqPrep and found their results to be virtually the same as those produced with skewer with the same parameters.

Another Attempt at Musket EC



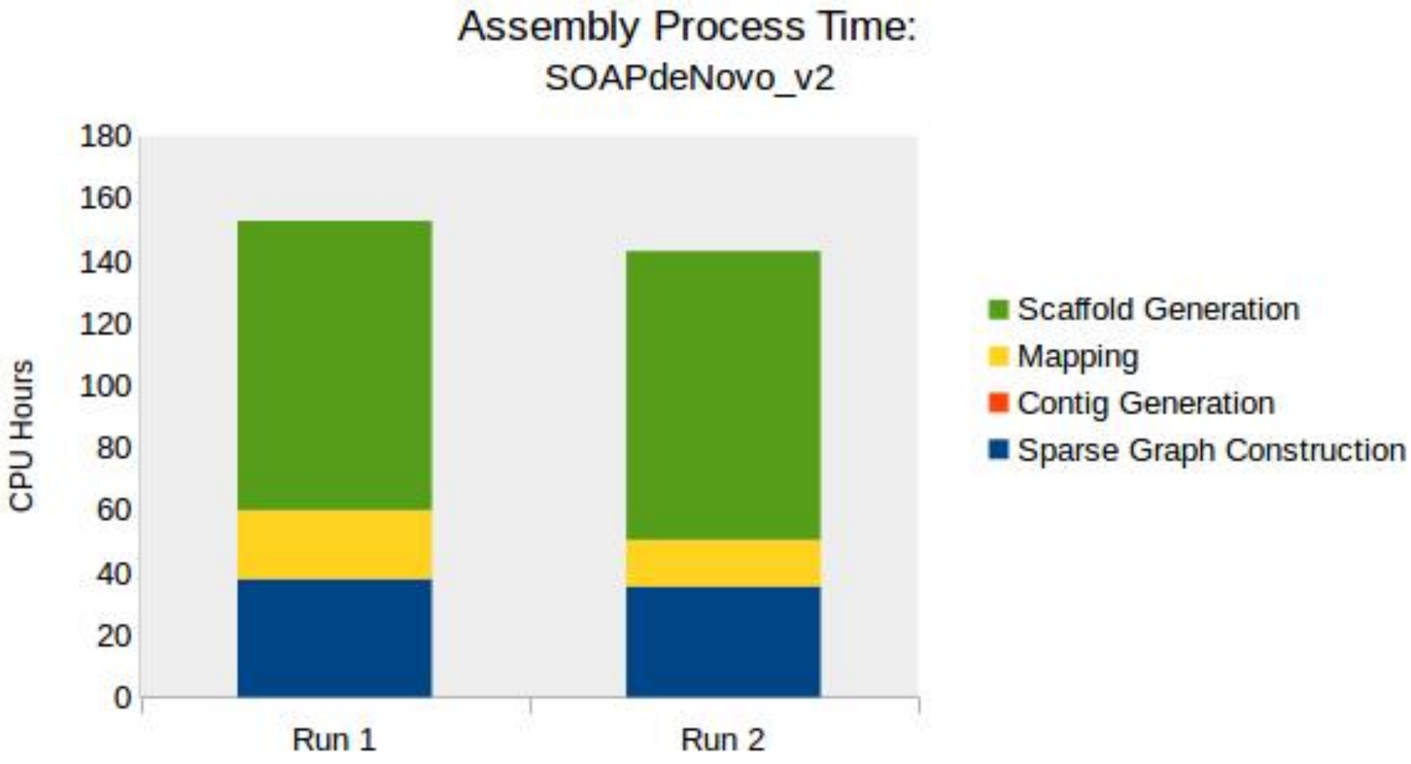
Assembly Run Performance

- Sparse Pregraph
  - 1st Run took about 9 hours and 28 minutes on 20 cores with 50 gb memory.
    - Used 136080.099 CPU seconds (~37.8 CPU hours) and 59.986 Gb max virtual memory.
  - 2nd Run took about 4 hours and 45 minutes on 12 cores with 60 gb memory.
    - Used 127383.665 CPU seconds (~35.4 CPU hours) and 58.941 Gb max virtual memory
- Contig generation
  - 1st Run took about 21 minutes on 20 cores with 50 gb memory.
    - Used 799.740 CPU seconds (~13 CPU minutes) and 5.780 Gb of max virtual memory.
  - 2nd Run took about 18 minutes on 12 cores with 10 gb memory.
    - Used 856.445 CPU seconds (~14.3 CPU minutes) and 5.781 Gb max virtual memory.

Assembly Run Performance

- Mapping
  - 1st Run took about 2 hours and 6 minutes on 20 cores with 50 gb memory.
    - Used 78902.546 CPU seconds (~21.9 CPU hours) and 66.103 Gb max virtual memory.
  - 2nd Run took about 2 hours and 55 minutes on 12 cores with 30 gb memory
    - Used 53220.443 CPU seconds (~14.78 CPU hours) and 78.048 Gb max virtual memory.
- Scaffold generation
  - 1st Run took about 24 hours and 50 minutes running on 20 cores with 50 gb memory.
    - Used 333894.594 CPU seconds (~92.7 CPU hours) and 20.366 Gb max virtual memory.
  - TBA

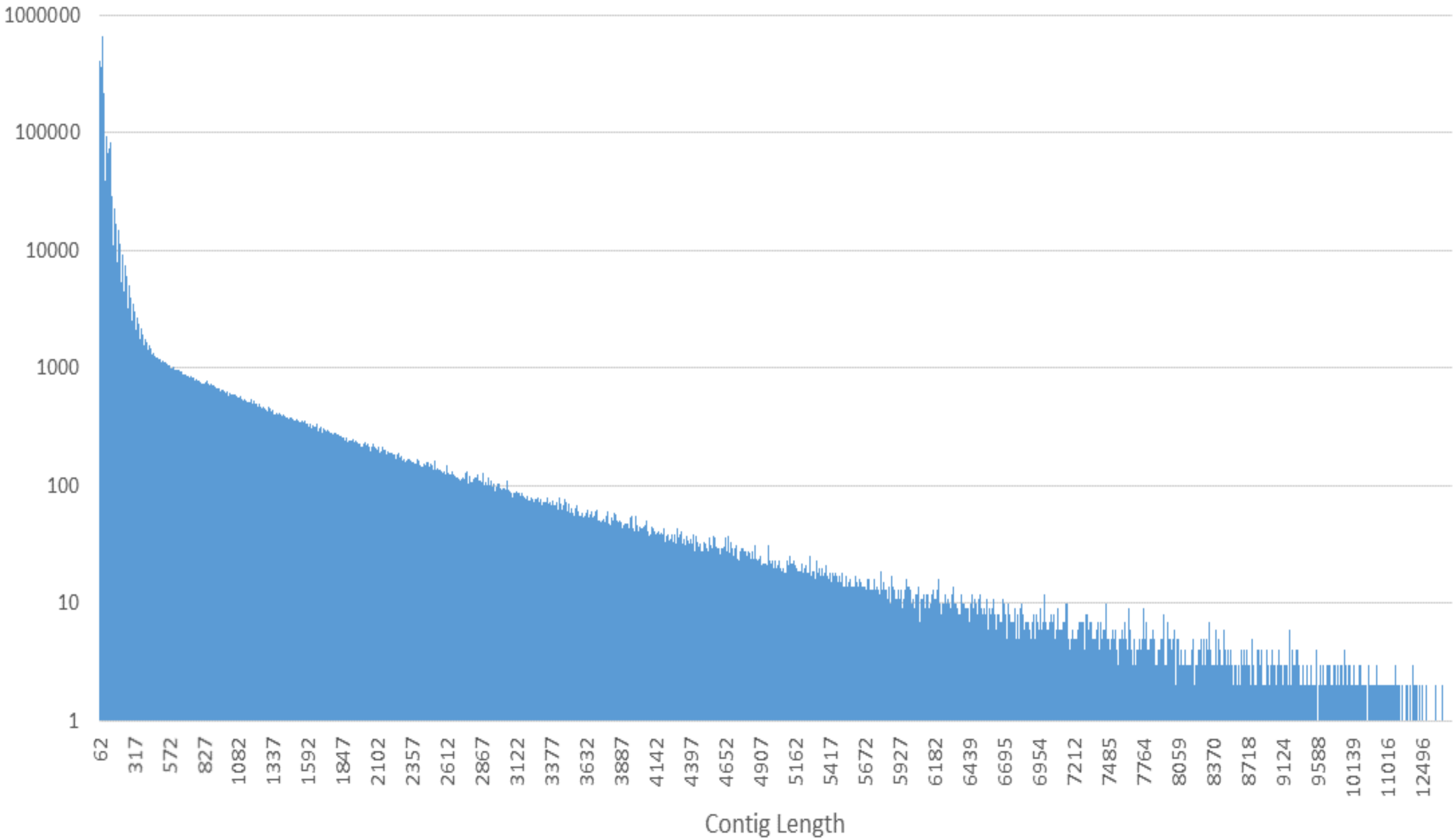
Assembly Run Performance



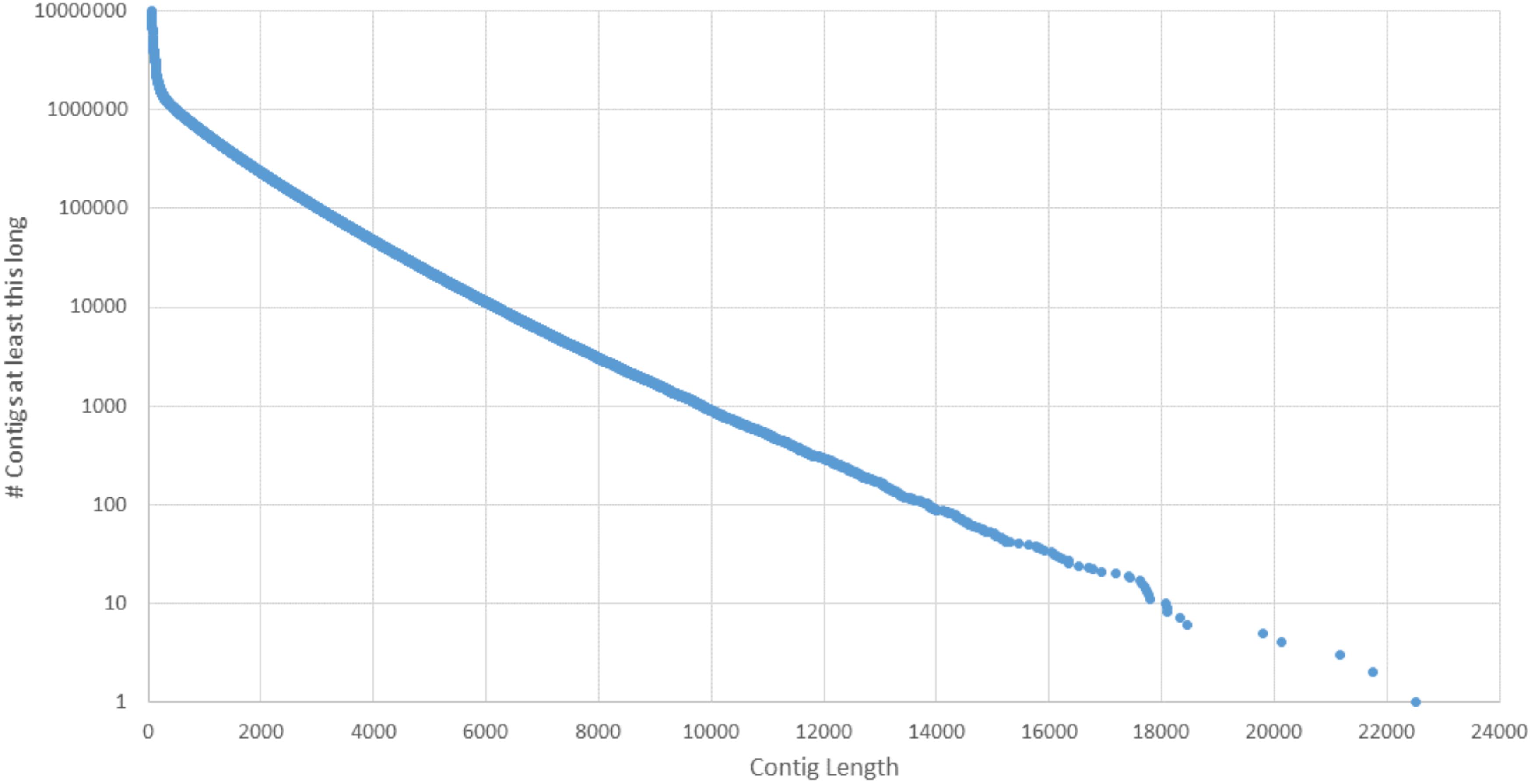
Assembly Run 1 Results (SOAP stat file)

- Contigs
  - Total contig sequence size : 2,051,251,797
  - Contig count : 3,854,379
  - Mean length : 532
  - Longest sequence : 22,512
  - N50 : 1,425 ; count : 389,550
  - length > 1K : 583,671 (15.14%)
  - length > 10K : 891 (0.02%)
- Scaffolds
  - Total assembly size (including 'N's) : 2,064,665,199
  - Total assembly size (without 'N's) : 1,974,393,478
  - Scaffold count : 2,030,303
  - Mean length : 1,016
  - Longest sequence : 60,333
  - N50 : 5,554 ; count : 105,217
  - length > 1K : 381,668 (18.80%)
  - length > 10K : 35,884 (1.77%)

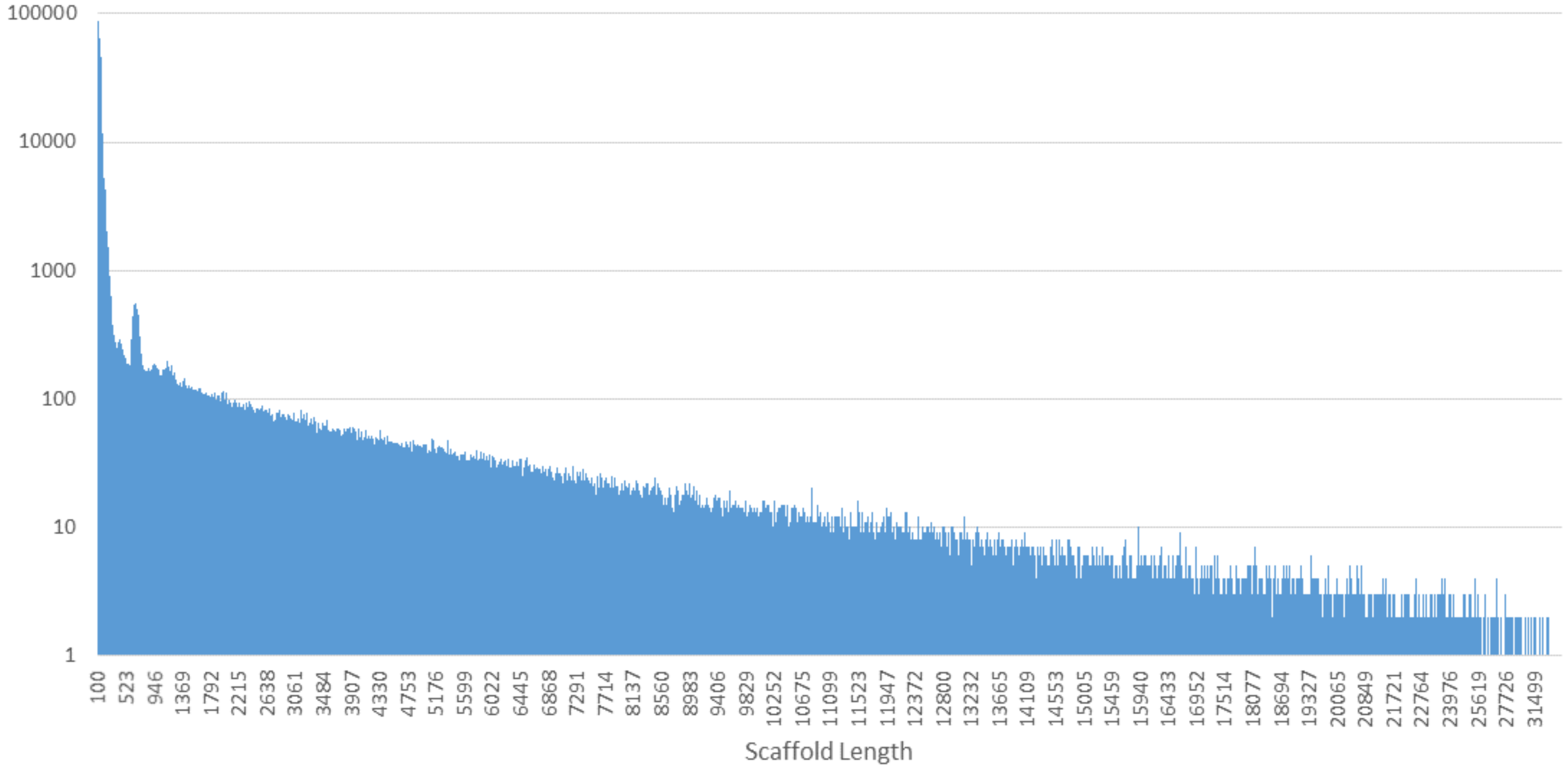
1st Run Contig Histogram



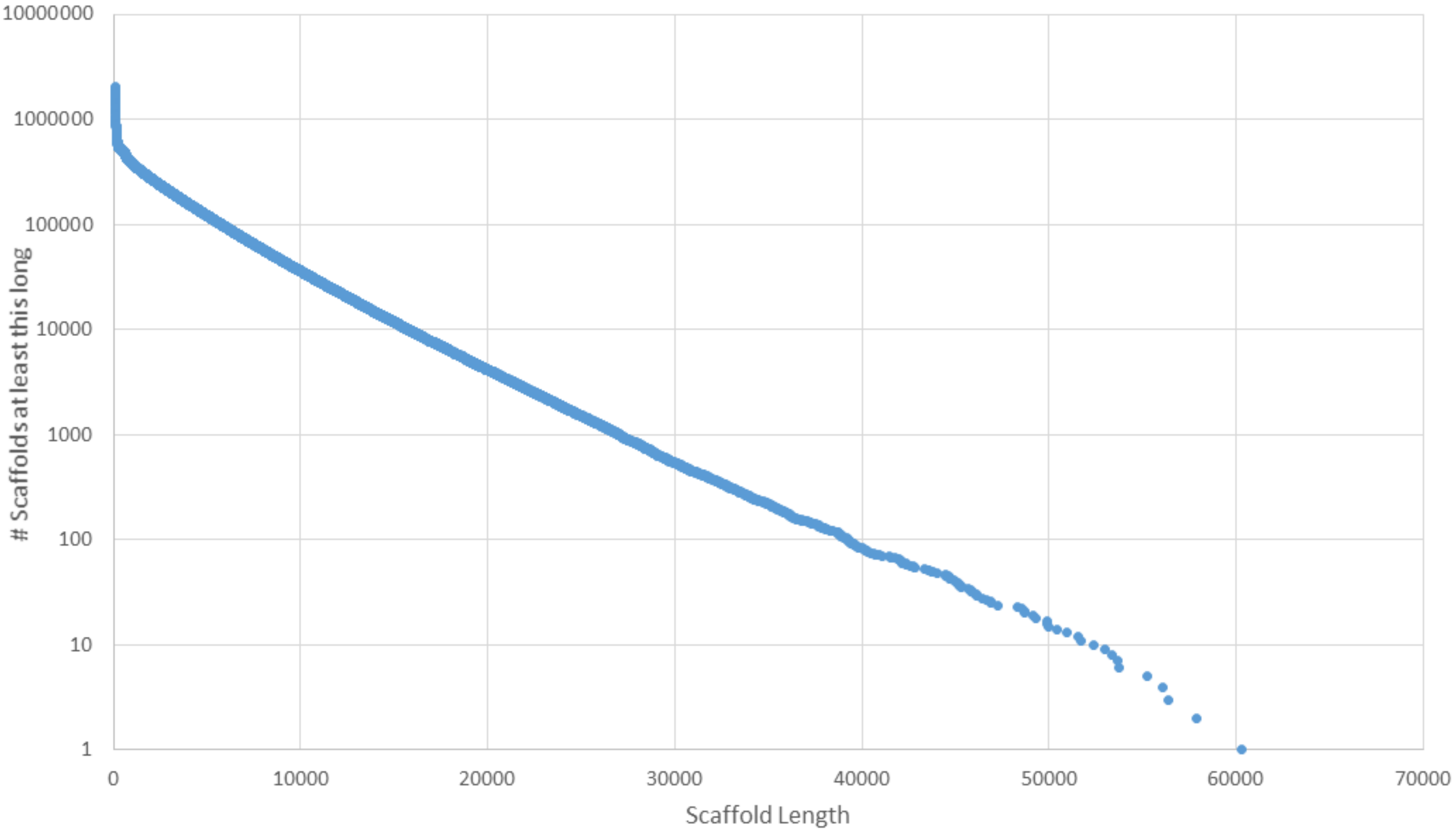
1st Run Contig Cumulative Histogram



1st Run Scaffold Histogram



1st Run Scaffold Cumulative Histogram



Scaffold Run Results (Rough Look)

Run 1

- Library 1 (SW019_S1+SW019_S2)
  - Scaffold number : 458,960
  - Average length : 3,721
  - Longest scaffold : 59,455
  - N50 : 5,728
  - N90 : 902

- Library 2 (SW018_S1)
  - Scaffold number : 412,707
  - Average length : 4,089
  - Longest scaffold : 59,455
  - N50 : 5,804
  - N90 : 932

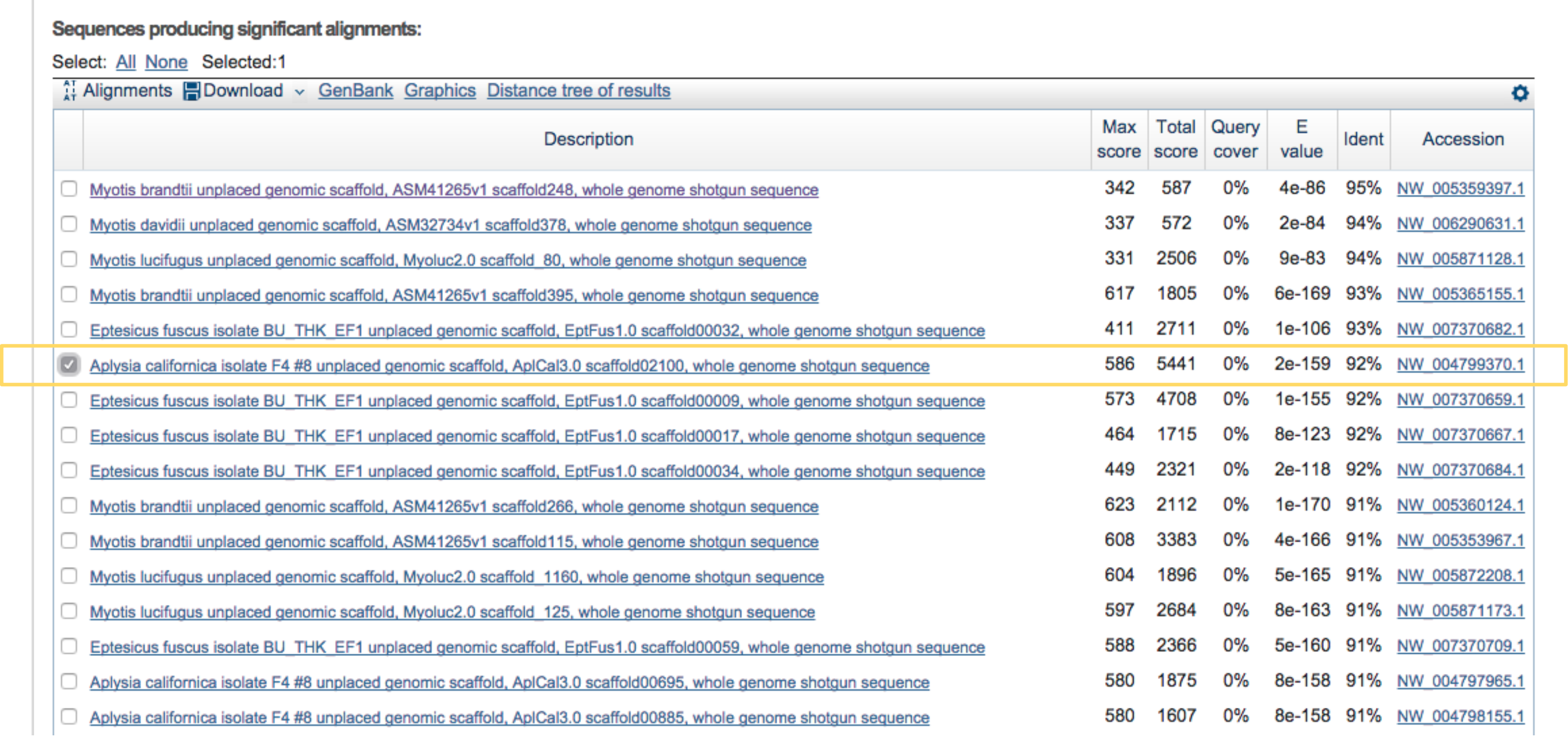
Run 2 (After different trimming and EC)

- Library 1 (SW019_S1+SW019_S2)
  - Scaffold number : 458,922
  - Average length : 3,721
  - Longest scaffold : 59,388
  - N50 : 5,726
  - N90 : 902

- Library 2 (SW018_S1)
  - Scaffold number : 354,147
  - Average length : 6,120
  - Longest scaffold : 103,900
  - N50 : 9,919
  - N90 : 2,018

BLAST 1st assembly results

- Scaffold fasta file first 10k lines
- Highly similar reference genome sequences

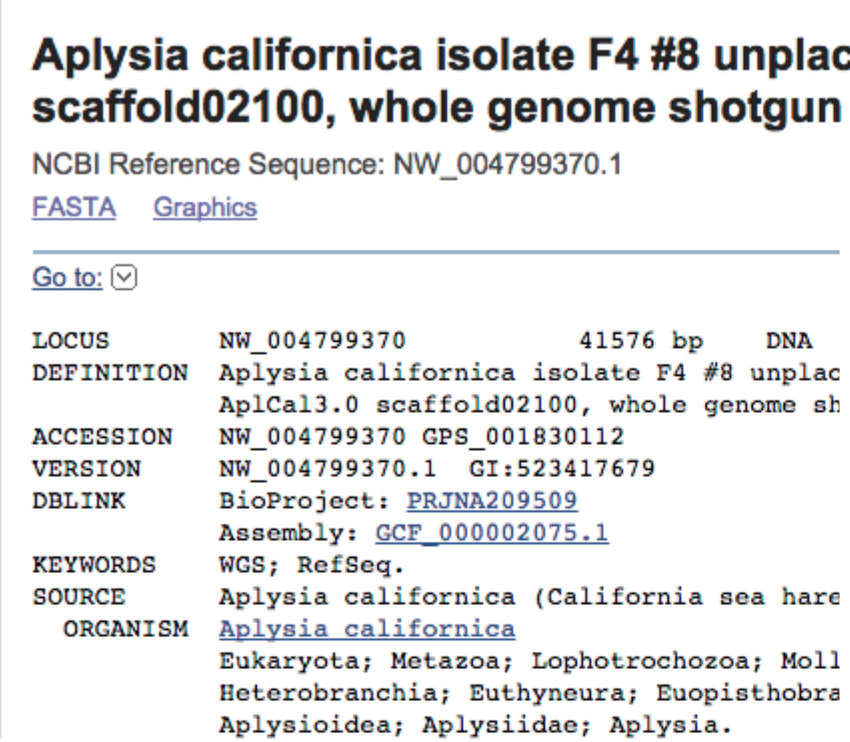


Sequences producing significant alignments:

Select: All None Selected:1

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	Myotis brandtii unplaced genomic scaffold, ASM41265v1 scaffold248, whole genome shotgun sequence	342	587	0%	4e-86	95%	NW_005359397.1
☐	Myotis davidii unplaced genomic scaffold, ASM32734v1 scaffold378, whole genome shotgun sequence	337	572	0%	2e-84	94%	NW_006290631.1
☐	Myotis lucifugus unplaced genomic scaffold, Myoluc2.0 scaffold_80, whole genome shotgun sequence	331	2506	0%	9e-83	94%	NW_005871128.1
☐	Myotis brandtii unplaced genomic scaffold, ASM41265v1 scaffold395, whole genome shotgun sequence	617	1805	0%	6e-169	93%	NW_005365155.1
☐	Eptesicus fuscus isolate BU_THK_EF1 unplaced genomic scaffold, EptFus1.0 scaffold00032, whole genome shotgun sequence	411	2711	0%	1e-106	93%	NW_007370682.1
☑	Aplysia californica isolate F4 #8 unplaced genomic scaffold, AplCal3.0 scaffold02100, whole genome shotgun sequence	586	5441	0%	2e-159	92%	NW_004799370.1
☐	Eptesicus fuscus isolate BU_THK_EF1 unplaced genomic scaffold, EptFus1.0 scaffold00009, whole genome shotgun sequence	573	4708	0%	1e-155	92%	NW_007370659.1
☐	Eptesicus fuscus isolate BU_THK_EF1 unplaced genomic scaffold, EptFus1.0 scaffold00017, whole genome shotgun sequence	464	1715	0%	8e-123	92%	NW_007370667.1
☐	Eptesicus fuscus isolate BU_THK_EF1 unplaced genomic scaffold, EptFus1.0 scaffold00034, whole genome shotgun sequence	449	2321	0%	2e-118	92%	NW_007370684.1
☐	Myotis brandtii unplaced genomic scaffold, ASM41265v1 scaffold266, whole genome shotgun sequence	623	2112	0%	1e-170	91%	NW_005360124.1
☐	Myotis brandtii unplaced genomic scaffold, ASM41265v1 scaffold115, whole genome shotgun sequence	608	3383	0%	4e-166	91%	NW_005353967.1
☐	Myotis lucifugus unplaced genomic scaffold, Myoluc2.0 scaffold_1160, whole genome shotgun sequence	604	1896	0%	5e-165	91%	NW_005872208.1
☐	Myotis lucifugus unplaced genomic scaffold, Myoluc2.0 scaffold_125, whole genome shotgun sequence	597	2684	0%	8e-163	91%	NW_005871173.1
☐	Eptesicus fuscus isolate BU_THK_EF1 unplaced genomic scaffold, EptFus1.0 scaffold00059, whole genome shotgun sequence	588	2366	0%	5e-160	91%	NW_007370709.1
☐	Aplysia californica isolate F4 #8 unplaced genomic scaffold, AplCal3.0 scaffold00695, whole genome shotgun sequence	580	1875	0%	8e-158	91%	NW_004797965.1
☐	Aplysia californica isolate F4 #8 unplaced genomic scaffold, AplCal3.0 scaffold00885, whole genome shotgun sequence	580	1607	0%	8e-158	91%	NW_004798155.1

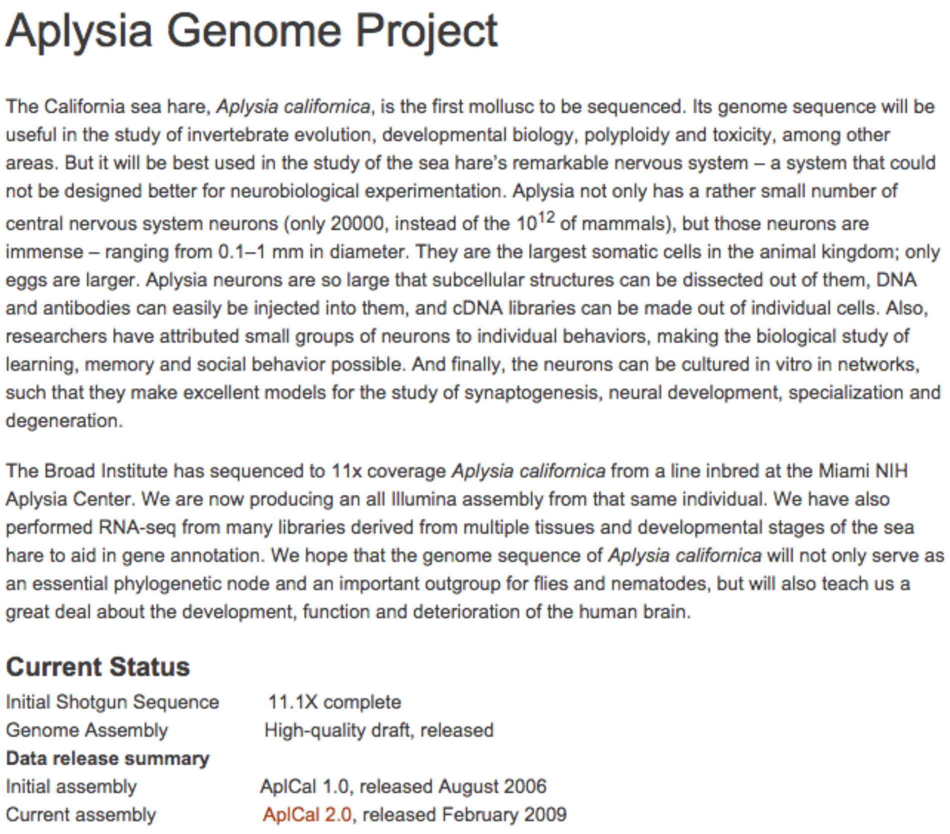
Aplysia californica: California sea hare



Aplysia californica isolate F4 #8 unplaced genomic scaffold, AplCal3.0 scaffold02100, whole genome shotgun sequence

NCBI Reference Sequence: NW_004799370.1

```
LOCUS       NW_004799370           41576 bp    DNA
DEFINITION  Aplysia californica isolate F4 #8 unplac
            AplCal3.0 scaffold02100, whole genome sh
ACCESSION   NW_004799370 GPS_001830112
VERSION     NW_004799370.1  GI:523417679
DBLINK      BioProject: PRJNA209509
            Assembly: GCF_000002075.1
KEYWORDS    WGS; RefSeq.
SOURCE      Aplysia californica (California sea hare
  ORGANISM  Aplysia californica
            Eukaryota; Metazoa; Lophotrochozoa; Moll
            Heterobranchia; Euthyneura; Euopisthobra
            Aplysioidea; Aplysiidae; Aplysia.
```



Aplysia Genome Project

The California sea hare, *Aplysia californica*, is the first mollusc to be sequenced. Its genome sequence will be useful in the study of invertebrate evolution, developmental biology, polyploidy and toxicity, among other areas. But it will be best used in the study of the sea hare's remarkable nervous system – a system that could not be designed better for neurobiological experimentation. Aplysia not only has a rather small number of central nervous system neurons (only 20000, instead of the 10^{12} of mammals), but those neurons are immense – ranging from 0.1–1 mm in diameter. They are the largest somatic cells in the animal kingdom; only eggs are larger. Aplysia neurons are so large that subcellular structures can be dissected out of them, DNA and antibodies can easily be injected into them, and cDNA libraries can be made out of individual cells. Also, researchers have attributed small groups of neurons to individual behaviors, making the biological study of learning, memory and social behavior possible. And finally, the neurons can be cultured in vitro in networks, such that they make excellent models for the study of synaptogenesis, neural development, specialization and degeneration.

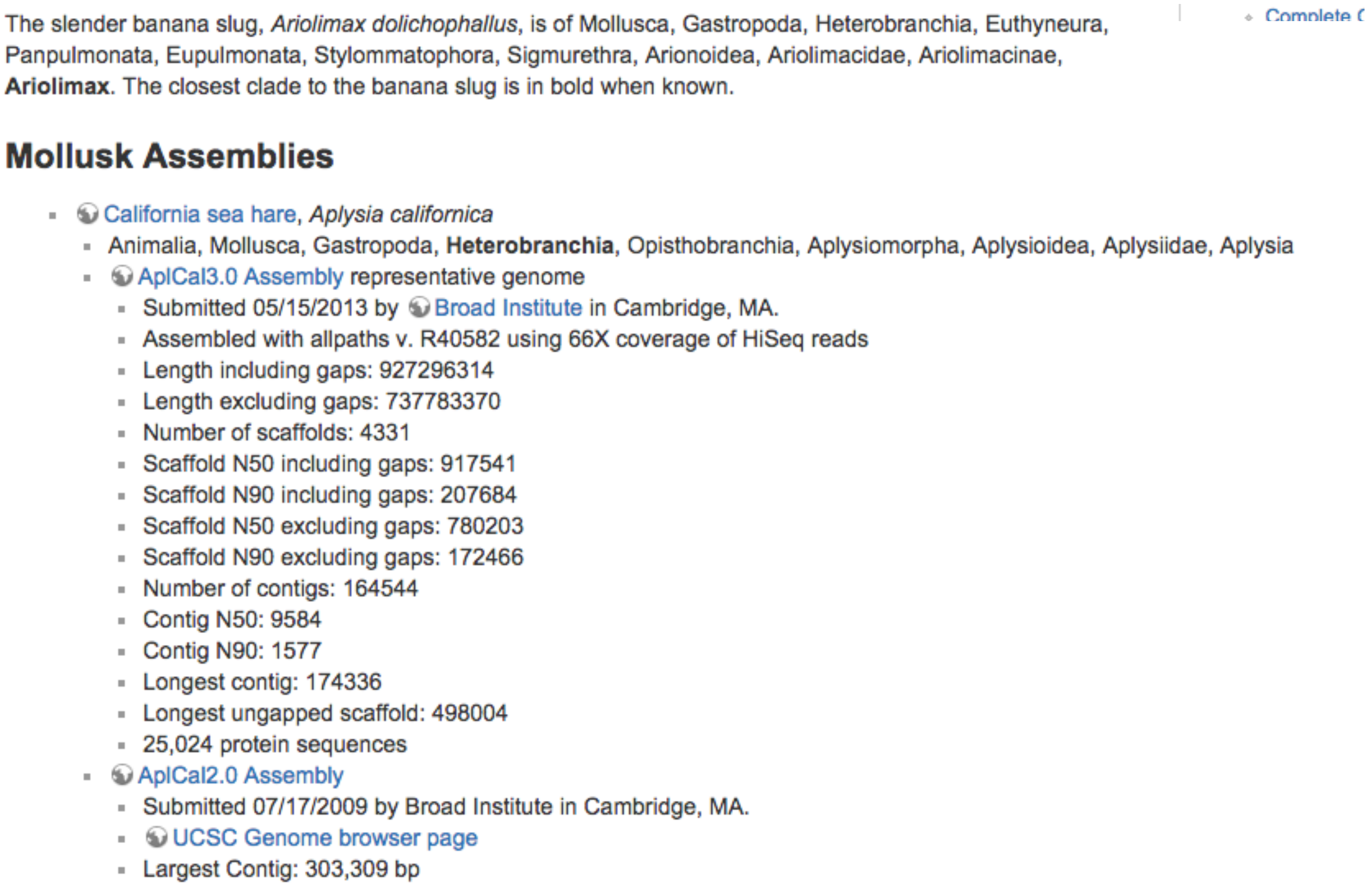
The Broad Institute has sequenced to 11x coverage *Aplysia californica* from a line inbred at the Miami NIH Aplysia Center. We are now producing an all Illumina assembly from that same individual. We have also performed RNA-seq from many libraries derived from multiple tissues and developmental stages of the sea hare to aid in gene annotation. We hope that the genome sequence of *Aplysia californica* will not only serve as an essential phylogenetic node and an important outgroup for flies and nematodes, but will also teach us a great deal about the development, function and deterioration of the human brain.

Current Status

Initial Shotgun Sequence	11.1X complete
Genome Assembly	High-quality draft, released
Data release summary	
Initial assembly	AplCal 1.0, released August 2006
Current assembly	AplCal 2.0, released February 2009

https://www.broadinstitute.org/science/projects/mammals-models/vertebrates-invertebrates/aplysia/aplysia-genome-sequencing-project

More on the wiki…



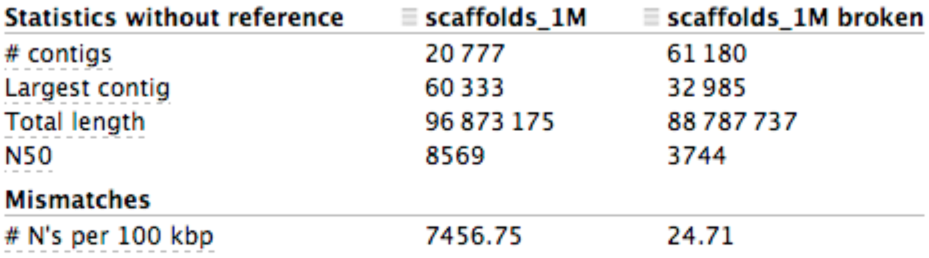
The slender banana slug, *Ariolimax dolichophallus*, is of Mollusca, Gastropoda, Heterobranchia, Euthyneura, Panpulmonata, Eupulmonata, Stylommatophora, Sigmurethra, Arionoidea, Ariolimacidae, Ariolimacinae, **Ariolimax**. The closest clade to the banana slug is in bold when known.

Mollusk Assemblies

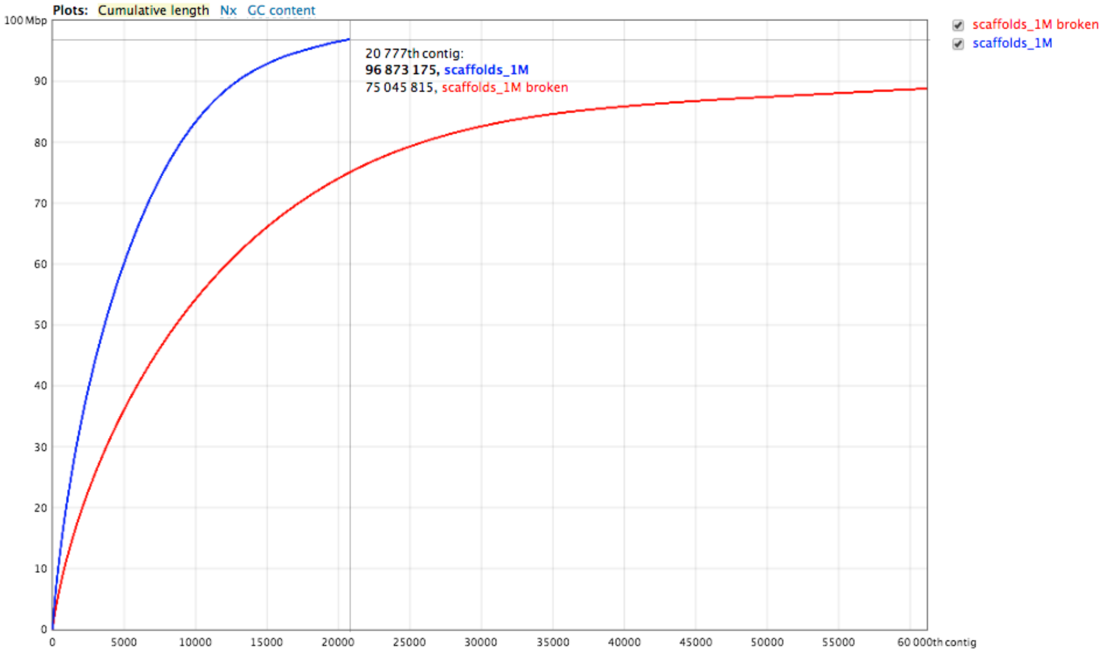
- California sea hare, *Aplysia californica*
  - Animalia, Mollusca, Gastropoda, **Heterobranchia**, Opisthobranchia, Aplysiomorpha, Aplysioidea, Aplysiidae, Aplysia
  - AplCal3.0 Assembly representative genome
    - Submitted 05/15/2013 by Broad Institute in Cambridge, MA.
    - Assembled with allpaths v. R40582 using 66X coverage of HiSeq reads
    - Length including gaps: 927296314
    - Length excluding gaps: 737783370
    - Number of scaffolds: 4331
    - Scaffold N50 including gaps: 917541
    - Scaffold N90 including gaps: 207684
    - Scaffold N50 excluding gaps: 780203
    - Scaffold N90 excluding gaps: 172466
    - Number of contigs: 164544
    - Contig N50: 9584
    - Contig N90: 1577
    - Longest contig: 174336
    - Longest ungapped scaffold: 498004
    - 25,024 protein sequences
  - AplCal2.0 Assembly
    - Submitted 07/17/2009 by Broad Institute in Cambridge, MA.
    - UCSC Genome browser page
    - Largest Contig: 303,309 bp

https://banana-slug.soe.ucsc.edu/other_mollusk_genomes

Partial Scaffold Analysis with Quast



Statistics without reference	scaffolds_1M	scaffolds_1M broken
# contigs	20 777	61 180
Largest contig	60 333	32 985
Total length	96 873 175	88 787 737
N50	8569	3744
Mismatches		
# N's per 100 kbp	7456.75	24.71



Contigs are ordered from largest (contig #1) to smallest.

QUAST with full scaffold file

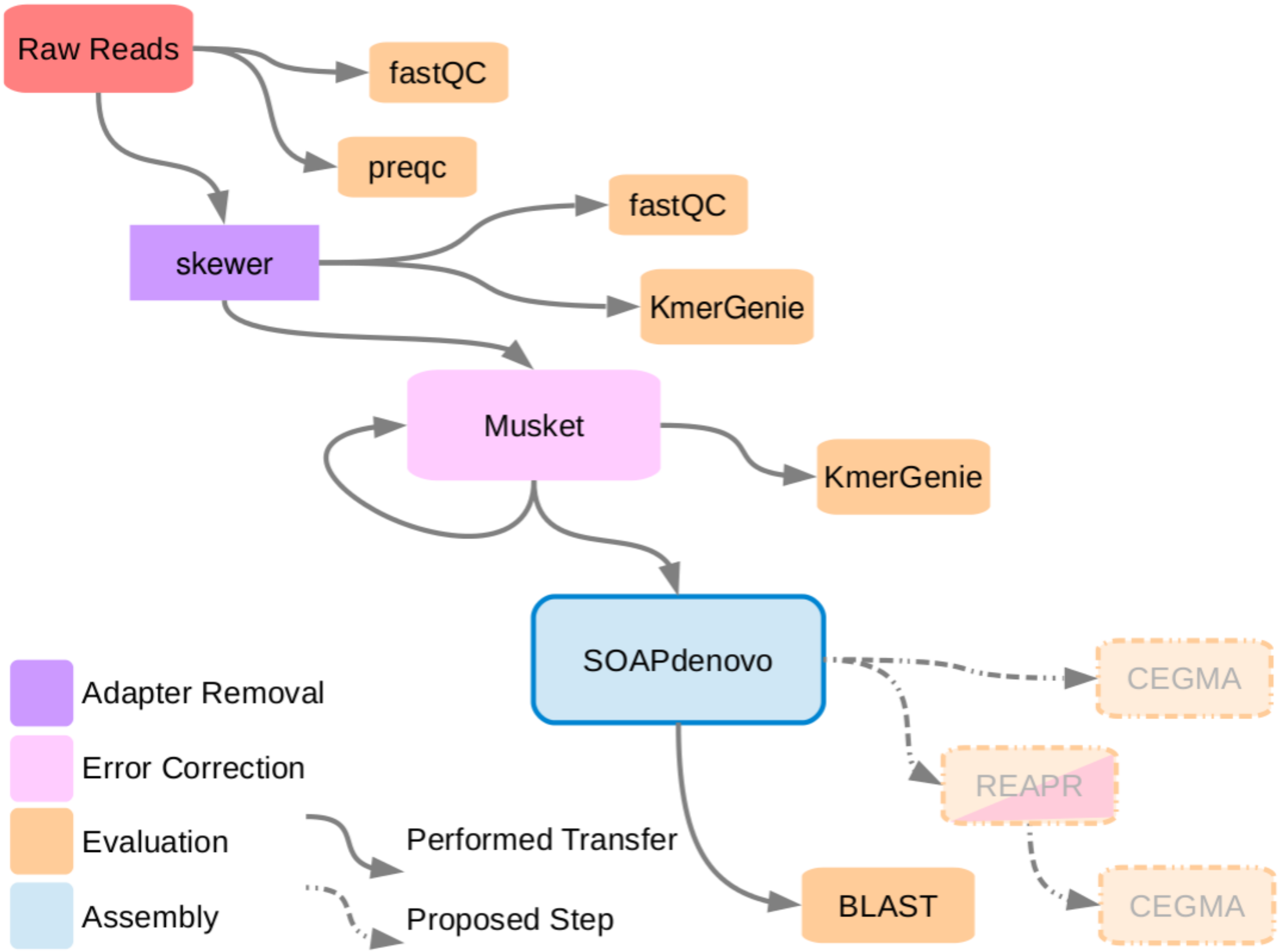
```
Assembly                soapdenovo2_sparseGraph.scafSeq
# contigs (>= 20 bp)     2030303
Total length (>= 20 bp)  2064665199
# contigs                2030303
Largest contig           60333
Total length             2064665199
GC (%)                   41.20
N50                      5554
N75                      2394
L50                      105217
L75                      243761
# N's per 100 kbp        4372.22
```

Quast: 1st vs 2nd assembly contigs

Assembly	soapdenovo2_sparseGraph_2.contig
# contigs (>= 10 bp)	9985423
Total length (>= 10 bp)	2499249369
# contigs	3854447
Largest contig	22512
Total length	2051284322
GC (%)	41.31
N50	1425
N75	513
L50	389531
L75	967321
# N's per 100 kbp	0.00

Assembly	soapdenovo2_sparseGraph.contig
# contigs (>= 10 bp)	9984798
Total length (>= 10 bp)	2499182380
# contigs	3854379
Largest contig	22512
Total length	2051251797
GC (%)	41.31
N50	1425
N75	513
L50	389550
L75	967304
# N's per 100 kbp	0.00

Current SOAPdenovo Pipeline



Raw Reads

fastQC

preqc

skewer

fastQC

KmerGenie

Musket

KmerGenie

SOAPdenovo

CEGMA

REAPR

CEGMA

BLAST

Adapter Removal

Error Correction

Evaluation

Assembly

Performed Transfer

Proposed Step

Next Steps

CEGMA -- Completeness test, pre-post-processing

REAPR -- Evaluate assembly accuracy + Correction

CEGMA -- Completeness test, pre-post-processing

SOAPdenovo -- Meta-assembly

BWA-MEM -- Re-map all read data to merged assembly

BME205 HW7 -- ORF analysis

BLAST -- Just because why not check against more refs?

Post-Mortem

Recommendation for next BME235 offering:

- Establish tentative milestones
- Establish public budget, with earmarks
- Break teams into workflow phase
  - Incentivize communication
  - Reduce redundancy
- Involve us in the library prep process
- Involve us in slug hunting?? jkjk
- Have Stefan do an assembly and we can compare our contig N50 to his