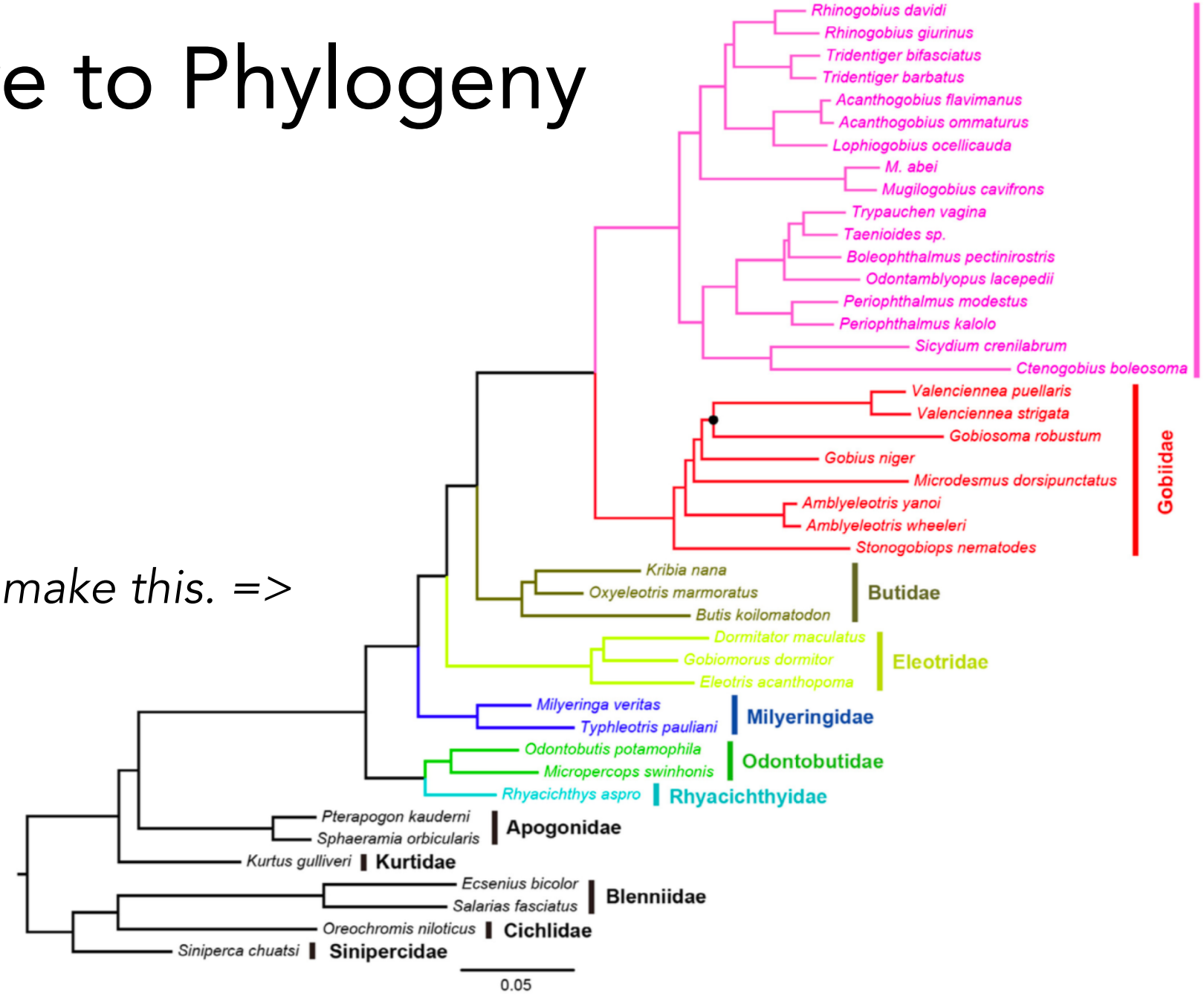
Exon-capture to Phylogeny

FISH 546 – Bioinformatics
Calder Atta

Disclaimer:

I did not make this. =>



Molecular Phylogenetics and Evolution xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Phylogenomic analysis on the exceptionally diverse fish clade Gobioidei (Actinopterygii: Gobiiformes) and data-filtering based on molecular clocklikeness

Ting Kuang[a,b,c,1], Luke Tornabene[d,1], Jingyan Li[a,b,c], Jiamei Jiang[a,b,c], Prosanta Chakrabarty[e], John S. Sparks[f], Gavin J.P. Naylor[g], Chenhong Li[a,b,c,*]

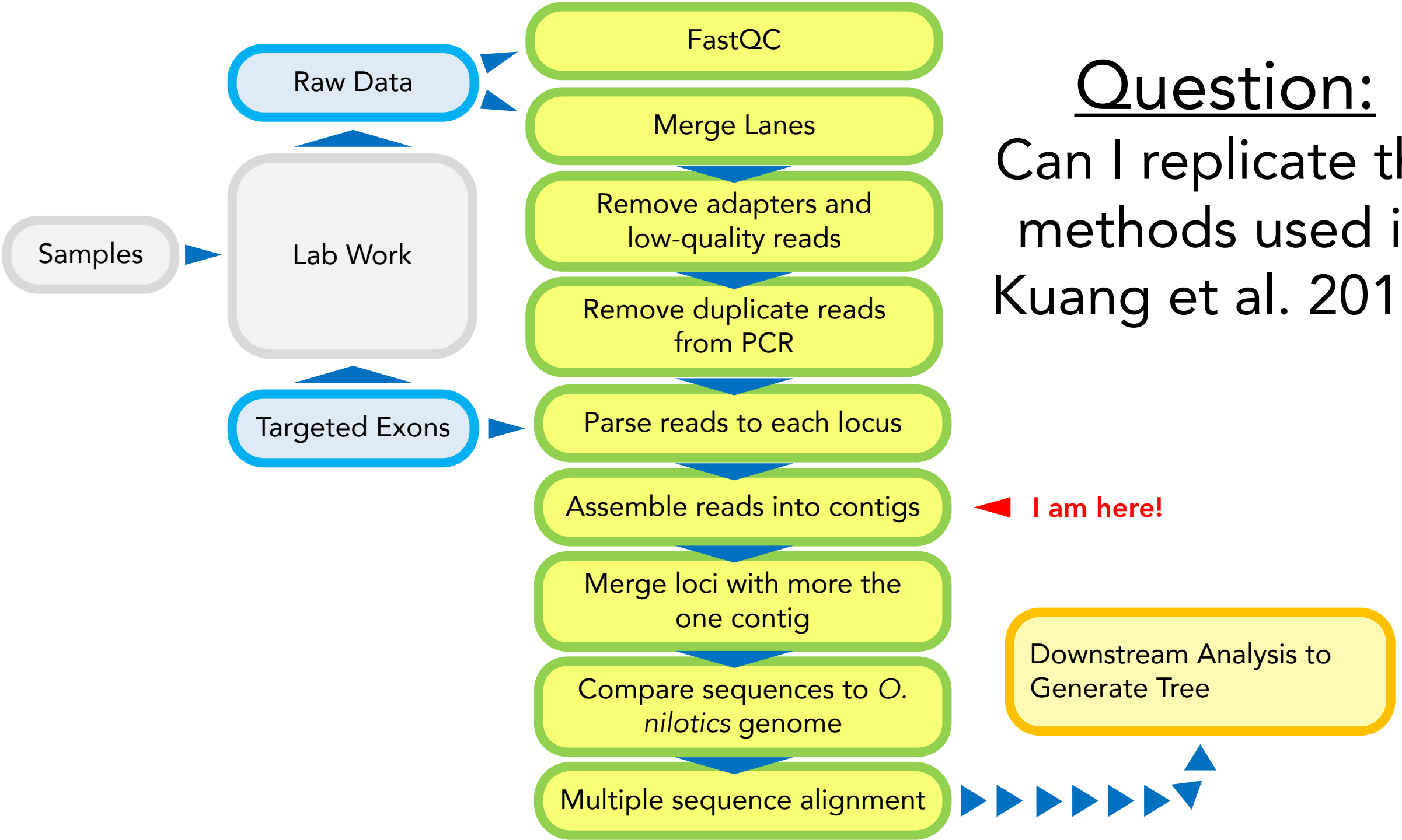
[a] Shanghai Universities Key Laboratory of Marine Animal Taxonomy and Evolution, Shanghai, China
[b] Shanghai Collaborative Innovation for Aquatic Animal Genetics and Breeding, Shanghai, China
[c] National Demonstration Center for Experimental Fisheries Science Education (Shanghai Ocean University), China
[d] School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA 98105, USA
[e] Louisiana State University, Museum of Natural Science, Department of Biological Sciences, Baton Rouge, LA 70803, USA
[f] American Museum of Natural History, Central Park West at 79th Street, NY, NY 10024, USA
[g] University of Florida, Gainesville, FL 3261, USA

ARTICLE INFO

Keywords:
Phylogenomics
Phylogenetics
Gobioidei
Target-gene enrichment
Data filtering
Molecular clocklikeness

ABSTRACT

The use of genome-scale data to infer phylogenetic relationships has gained in popularity in recent years due to the progress made in target-gene capture and sequencing techniques. Data filtering, the approach of excluding data inconsistent with the model from analyses, presumably could alleviate problems caused by systematic errors in phylogenetic inference. Different data filtering criteria, such as those based on evolutionary rate and molecular clocklikeness as well as others have been proposed for selecting useful phylogenetic markers, yet few studies have tested these criteria using phylogenomic data. We developed a novel set of single-copy nuclear

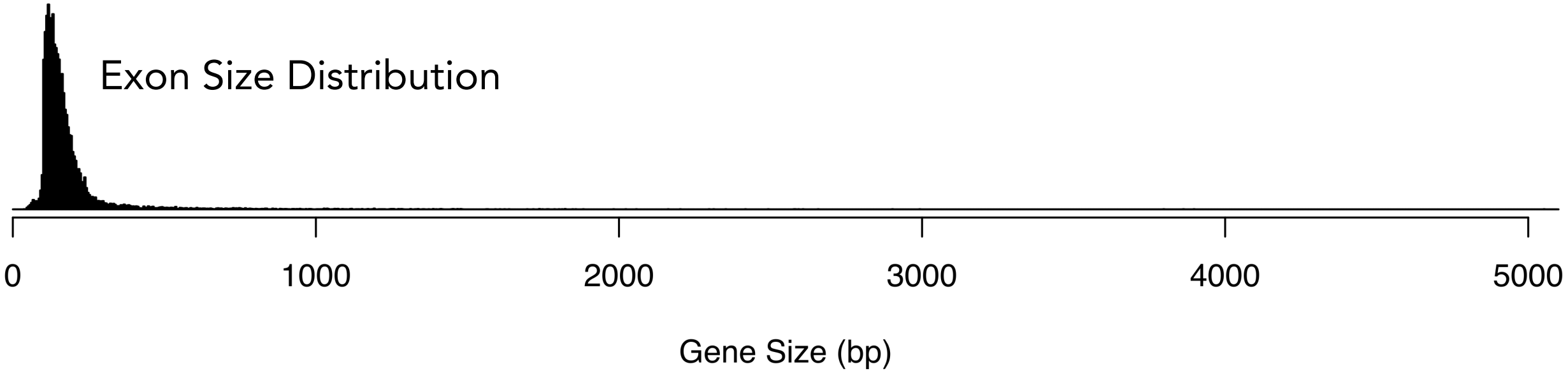


Question:

Can I replicate the methods used in Kuang et al. 2014?

Targeted Exons

- 18,400 exons
- 3,302,187 bp in total
- Derived from 8 species of fish for use on all bony-fishes



Raw Data

- Testing protocol using 7 samples from the original study
- 2 lanes x 2 forward/reverse = 4 files

1 Sample

TORN_Pool_4_S4_L006_R1_001.fastq	1.1 GB	TORN_Pool_7_S7_L008_R1_001.fastq	1.14 GB
TORN_Pool_4_S4_L006_R2_001.fastq	1.1 GB	TORN_Pool_7_S7_L008_R2_001.fastq	1.14 GB
TORN_Pool_4_S4_L008_R1_001.fastq	898.7 MB	TORN_Pool_8_S8_L006_R1_001.fastq	747 MB
TORN_Pool_4_S4_L008_R2_001.fastq	898.7 MB	TORN_Pool_8_S8_L006_R2_001.fastq	747 MB
TORN_Pool_5_S5_L006_R1_001.fastq	447.5 MB	TORN_Pool_8_S8_L008_R1_001.fastq	815.5 MB
TORN_Pool_5_S5_L006_R2_001.fastq	447.5 MB	TORN_Pool_8_S8_L008_R2_001.fastq	815.5 MB
TORN_Pool_5_S5_L008_R1_001.fastq	312 MB	TORN_Pool_9_S9_L006_R1_001.fastq	1.05 GB
TORN_Pool_5_S5_L008_R2_001.fastq	312 MB	TORN_Pool_9_S9_L006_R2_001.fastq	1.05 GB
TORN_Pool_6_S6_L006_R1_001.fastq	1.22 GB	TORN_Pool_9_S9_L008_R1_001.fastq	758.3 MB
TORN_Pool_6_S6_L006_R2_001.fastq	1.22 GB	TORN_Pool_9_S9_L008_R2_001.fastq	758.3 MB
TORN_Pool_6_S6_L008_R1_001.fastq	1.11 GB	TORN_Pool_10_S10_L006_R1_001.fastq	867.6 MB
TORN_Pool_6_S6_L008_R2_001.fastq	1.11 GB	TORN_Pool_10_S10_L006_R2_001.fastq	867.6 MB
TORN_Pool_7_S7_L006_R1_001.fastq	919.7 MB	TORN_Pool_10_S10_L008_R1_001.fastq	536.9 MB
TORN_Pool_7_S7_L006_R2_001.fastq	919.7 MB	TORN_Pool_10_S10_L008_R2_001.fastq	536.9 MB

FastQC

Merge Lanes

Remove adapters and low-quality reads

Remove duplicate reads from PCR

Parse reads to each locus

Assemble reads into contigs

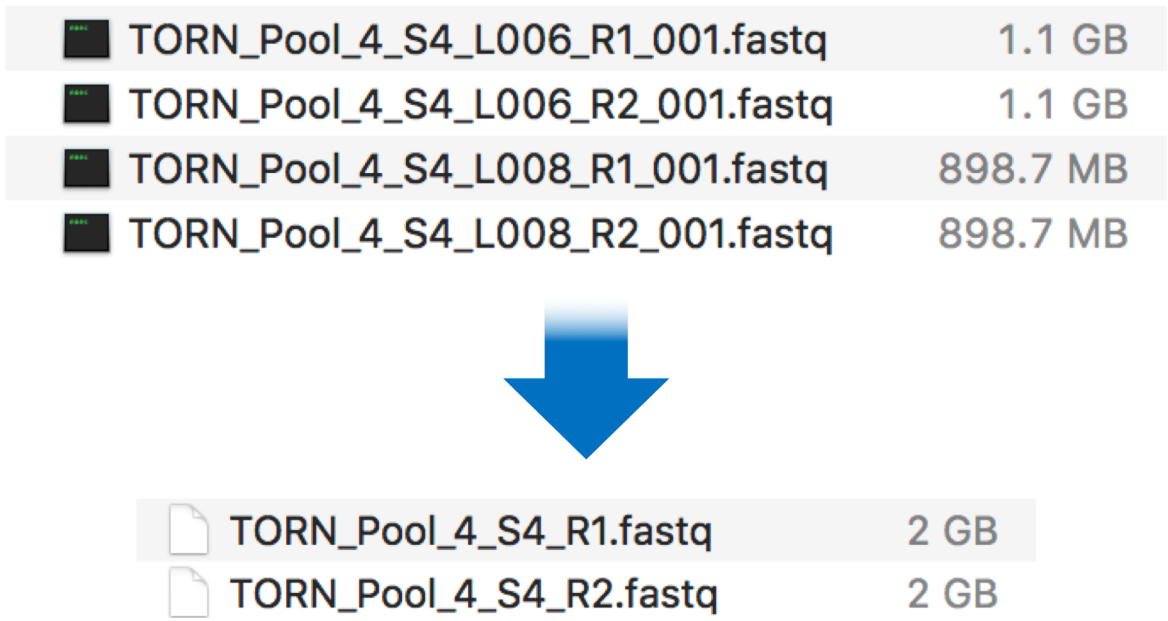
Merge loci with more the one contig

Compare sequences to *O. nilotics* genome

Multiple sequence alignment

Merge Lanes

cat



FastQC

Merge Lanes

Remove adapters and low-quality reads

Remove duplicate reads from PCR

Parse reads to each locus

Assemble reads into contigs

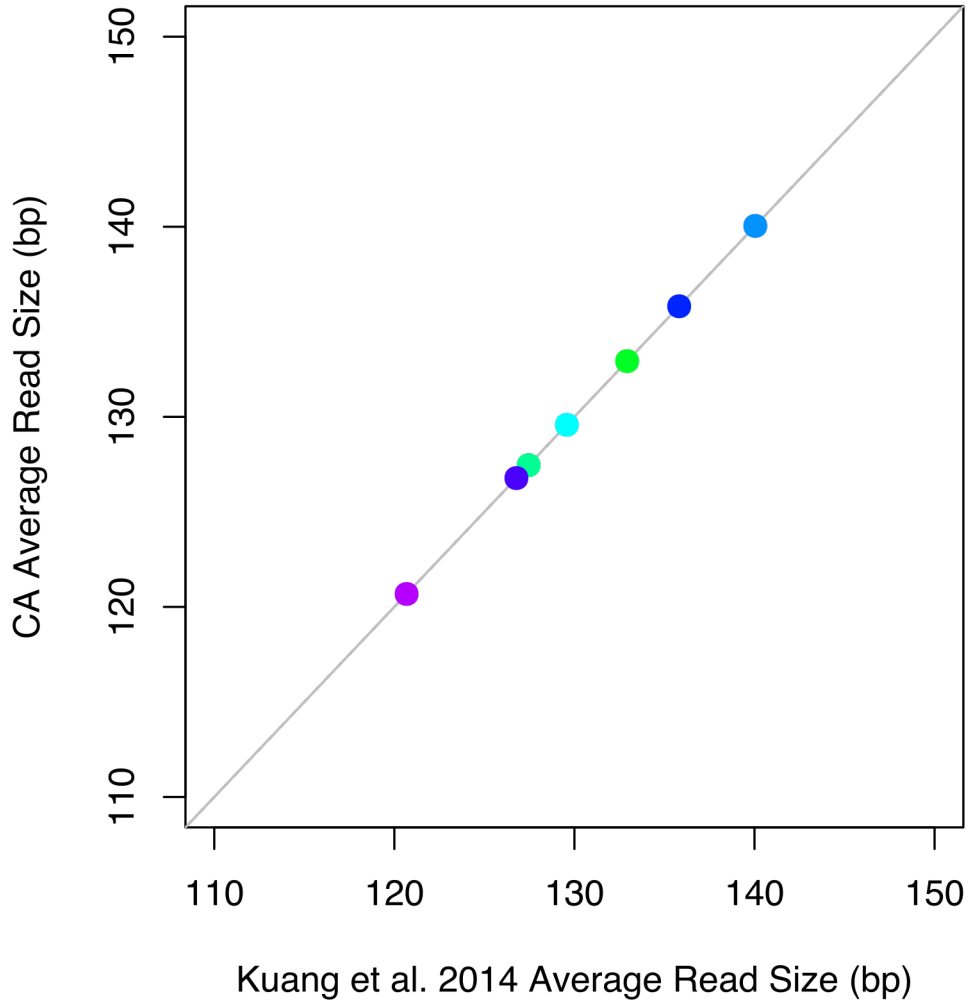
Merge loci with more the one contig

Compare sequences to *O. nilotics* genome

Multiple sequence alignment

Remove adapters & low-quality reads

Trim Galore



FastQC

Merge Lanes

Remove adapters and low-quality reads

Remove duplicate reads from PCR

Parse reads to each locus

Assemble reads into contigs

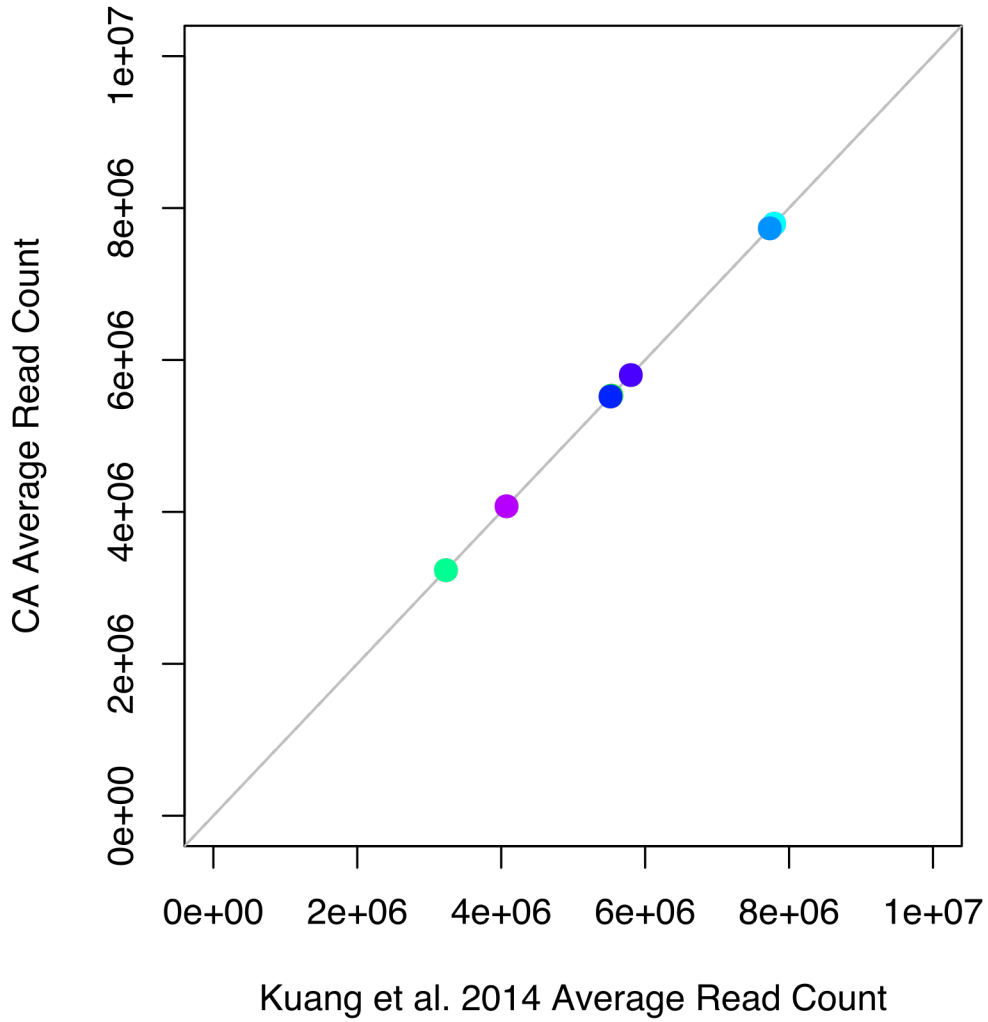
Merge loci with more the one contig

Compare sequences to *O. nilotics* genome

Multiple sequence alignment

Remove duplicate reads

Custom Perl Script



FastQC

Merge Lanes

Remove adapters and low-quality reads

Remove duplicate reads from PCR

Parse reads to each locus

Assemble reads into contigs

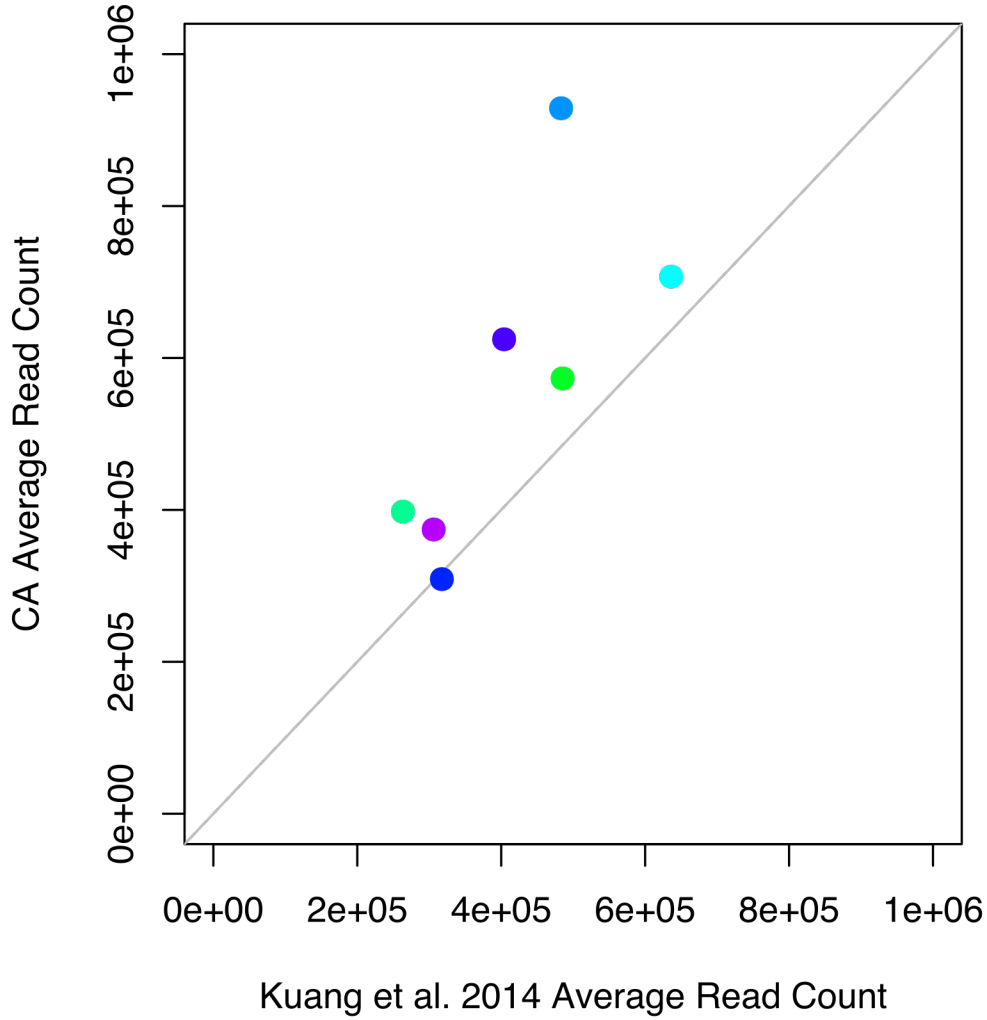
Merge loci with more the one contig

Compare sequences to *O. nilotics* genome

Multiple sequence alignment

Parse reads to each locus

Custom Perl Script



FastQC

Merge Lanes

Remove adapters and low-quality reads

Remove duplicate reads from PCR

Parse reads to each locus

Assemble reads into contigs

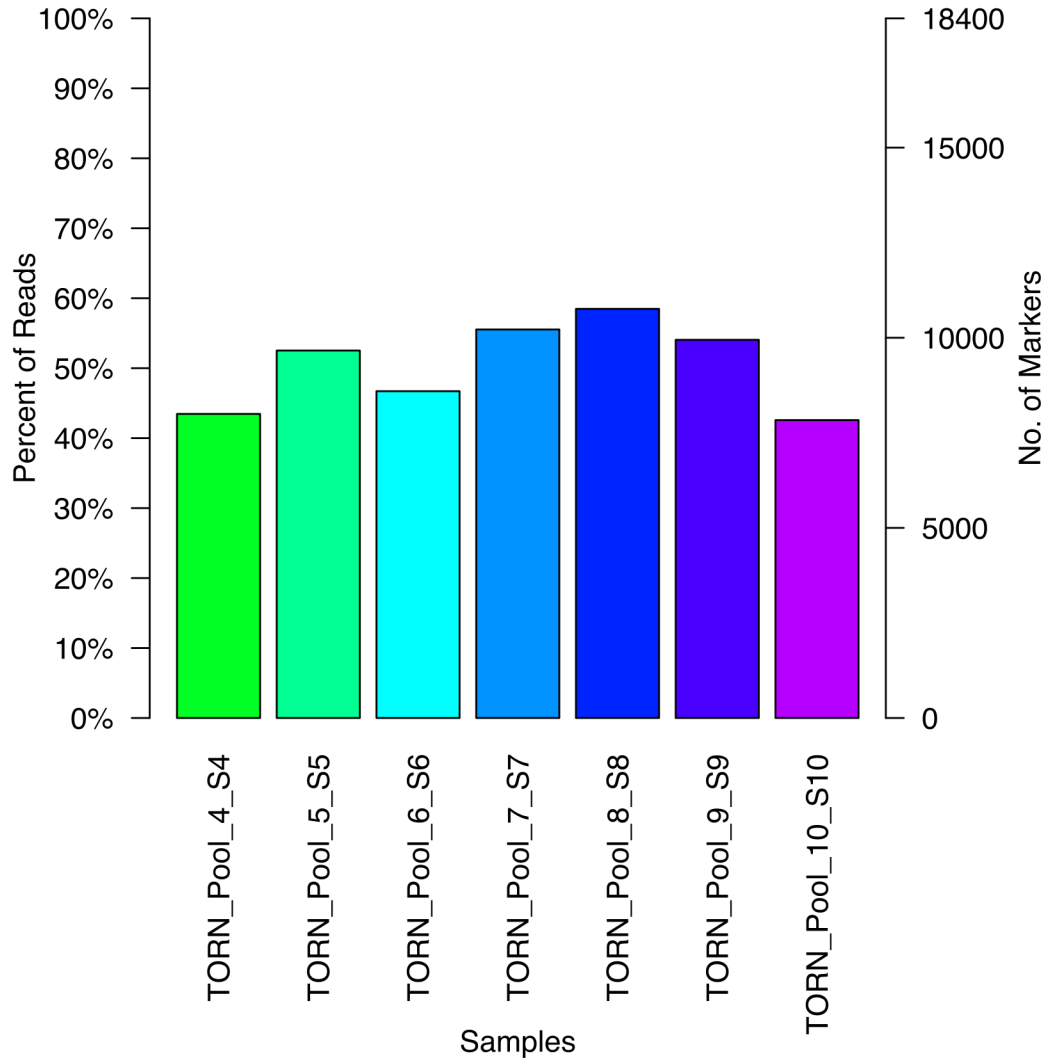
Merge loci with more the one contig

Compare sequences to *O. nilotics* genome

Multiple sequence alignment

Parse reads to each locus

Custom Perl Script



FastQC

Merge Lanes

Remove adapters and low-quality reads

Remove duplicate reads from PCR

Parse reads to each locus

Assemble reads into contigs

Merge loci with more the one contig

Compare sequences to *O. nilotics* genome

Multiple sequence alignment

Assemble reads into contigs

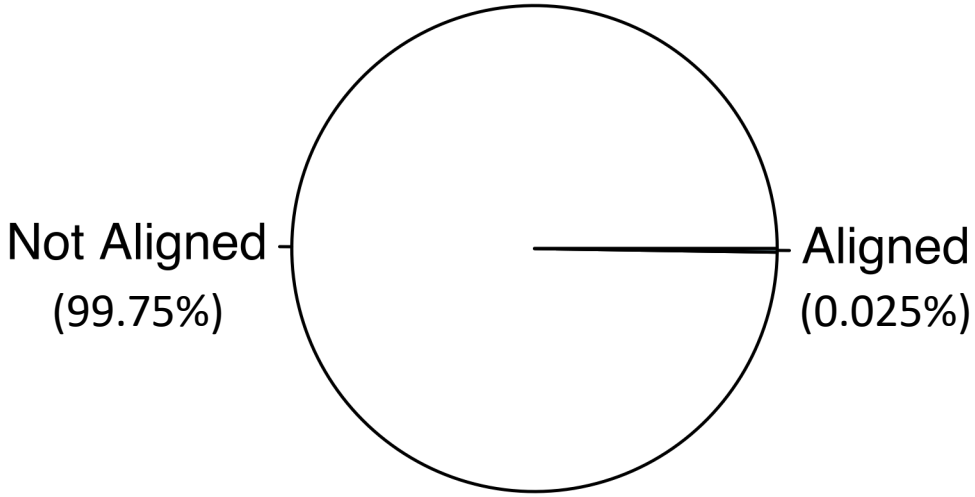
Trinity

Still not able to run Trinity…

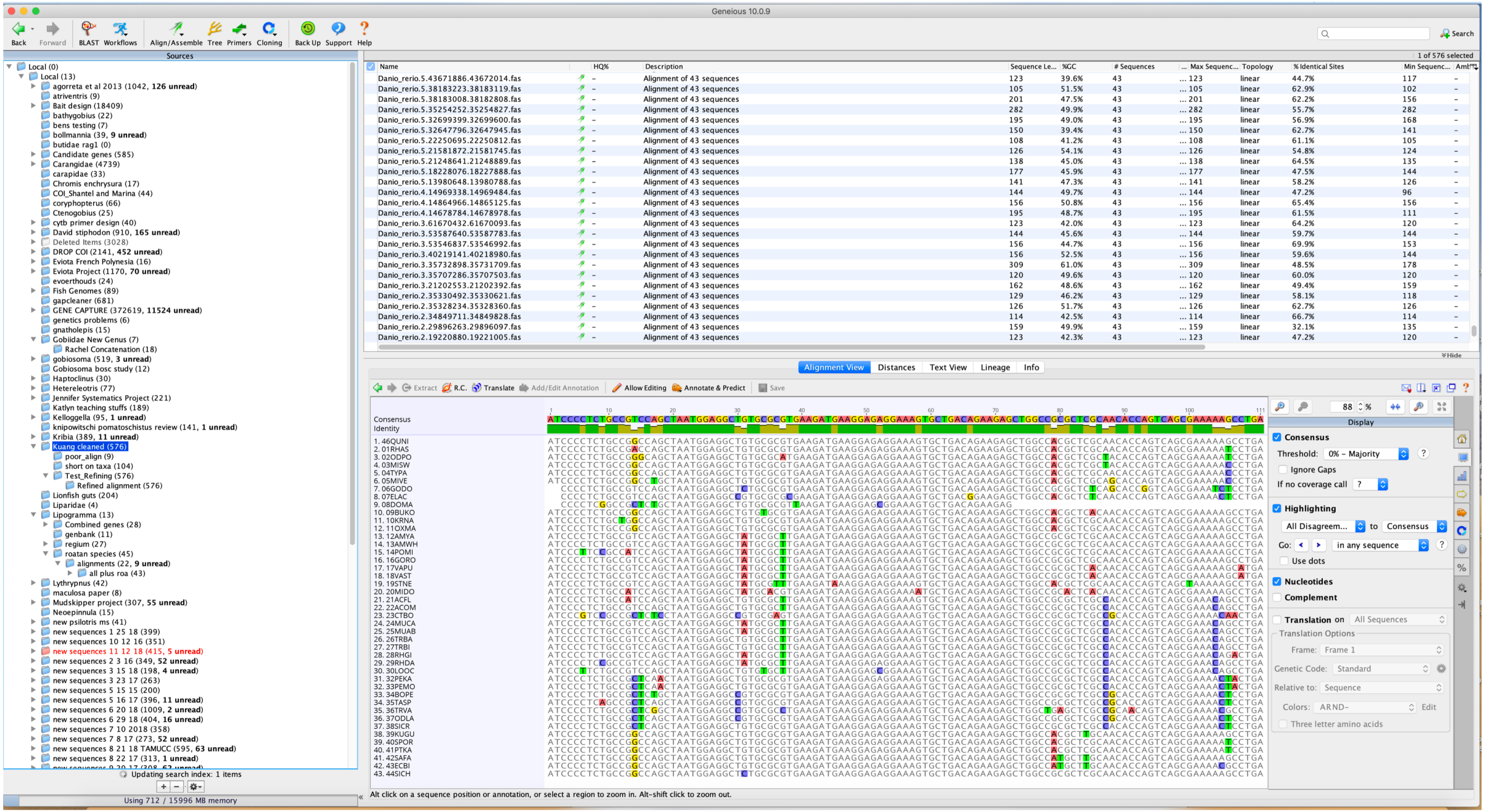
Many Issues:

- Problems installing on GitBash
- Many other required programs
- Input file naming convention contains unusual characters
- Running from custom Perl script
- Trinity must be an environmental variable when run from Perl script
- Perl script was out of date
- Memory issue?

Trouble-shooting with Bowtie2




```
2764977 reads; of these:
  2764977 (100.00%) were unpaired; of these:
    2758315 (99.76%) aligned 0 times
    4028 (0.15%) aligned exactly 1 time
    2634 (0.10%) aligned >1 times
0.24% overall alignment rate
```


Geneious 10.0.9
Back
Forward
BLAST
Workflows
Align/Assemble
Tree
Primers
Cloning
Back Up
Support
Help
Search
Sources
Local (0)
Local (13)
agorreta et al 2013 (1042, 126 unread)
atriventris (9)
Bait design (18409)
bathygobius (22)
bens testing (7)
bollmannia (39, 9 unread)
butidae rag1 (0)
Candidate genes (585)
Carangidae (4739)
carapidae (33)
Chromis enchrysura (17)
COI_Shantel and Marina (44)
coryphopterus (66)
Ctenogobius (25)
cytb primer design (40)
David stiphodon (910, 165 unread)
Deleted Items (3028)
DROP COI (2141, 452 unread)
Eviota French Polynesia (16)
Eviota Project (1170, 70 unread)
evoerthouds (24)
Fish Genomes (89)
gapcleaner (681)
GENE CAPTURE (372619, 11524 unread)
genetics problems (6)
gnatholepis (15)
Gobiidae New Genus (7)
Rachel Concatenation (18)
gobiosoma (519, 3 unread)
Gobiosoma bosc study (12)
Haptoclinus (30)
Hetereleotris (77)
Jennifer Systematics Project (221)
Katlyn teaching stuffs (189)
Kelloggella (95, 1 unread)
knipowitschi pomatoschistus review (141, 1 unread)
Kribia (389, 11 unread)
Kuang cleaned (576)
poor_align (9)
short on taxa (104)
Test_Refining (576)
Refined alignment (576)
Lionfish guts (204)
Liparidae (4)
Lipogramma (13)
Combined genes (28)
genbank (11)
regium (27)
roatan species (45)
alignments (22, 9 unread)
all plus roa (43)
Lythrypnus (42)
maculosa paper (8)
Mudskipper project (307, 55 unread)
Neoepinnula (15)
new psilotris ms (41)
new sequences 1 25 18 (399)
new sequences 10 12 16 (351)
new sequences 11 12 18 (415, 5 unread)
new sequences 2 3 16 (349, 52 unread)
new sequences 3 15 18 (198, 4 unread)
new sequences 3 23 17 (263)
new sequences 5 15 15 (200)
new sequences 5 16 17 (396, 11 unread)
new sequences 6 20 18 (1009, 2 unread)
new sequences 6 29 18 (404, 16 unread)
new sequences 7 10 2018 (358)
new sequences 7 8 17 (273, 52 unread)
new sequences 8 21 18 TAMUCC (595, 63 unread)
new sequences 8 22 17 (313, 1 unread)
Updating search index: 1 items
Using 712 / 15996 MB memory
1 of 576 selected
Name
HQ%
Description
Sequence Le...
%GC
Sequences
... Max Sequenc...
Topology
% Identical Sites
Min Sequenc...
Amb
Danio_rerio.5.43671886.43672014.fas – Alignment of 43 sequences 123 39.6% 43 ... 123 linear 44.7% 117 –
Danio_rerio.5.38183223.38183119.fas – Alignment of 43 sequences 105 51.5% 43 ... 105 linear 62.9% 102 –
Danio_rerio.5.38183008.38182808.fas – Alignment of 43 sequences 201 47.5% 43 ... 201 linear 62.2% 156 –
Danio_rerio.5.35254252.35254827.fas – Alignment of 43 sequences 282 49.9% 43 ... 282 linear 55.7% 282 –
Danio_rerio.5.32699399.32699600.fas – Alignment of 43 sequences 195 49.0% 43 ... 195 linear 56.9% 168 –
Danio_rerio.5.32647796.32647945.fas – Alignment of 43 sequences 150 39.4% 43 ... 150 linear 62.7% 141 –
Danio_rerio.5.22250695.22250812.fas – Alignment of 43 sequences 108 41.2% 43 ... 108 linear 61.1% 105 –
Danio_rerio.5.21581872.21581745.fas – Alignment of 43 sequences 126 54.1% 43 ... 126 linear 54.8% 124 –
Danio_rerio.5.21248641.21248889.fas – Alignment of 43 sequences 138 45.0% 43 ... 138 linear 64.5% 135 –
Danio_rerio.5.18228076.18227888.fas – Alignment of 43 sequences 177 45.9% 43 ... 177 linear 47.5% 144 –
Danio_rerio.5.13980648.13980788.fas – Alignment of 43 sequences 141 47.3% 43 ... 141 linear 58.2% 126 –
Danio_rerio.4.14969338.14969484.fas – Alignment of 43 sequences 144 49.7% 43 ... 144 linear 47.2% 96 –
Danio_rerio.4.14864966.14865125.fas – Alignment of 43 sequences 156 50.8% 43 ... 156 linear 65.4% 156 –
Danio_rerio.4.14678784.14678978.fas – Alignment of 43 sequences 195 48.7% 43 ... 195 linear 61.5% 111 –
Danio_rerio.3.61670432.61670093.fas – Alignment of 43 sequences 123 42.0% 43 ... 123 linear 64.2% 120 –
Danio_rerio.3.53587640.53587783.fas – Alignment of 43 sequences 144 45.6% 43 ... 144 linear 59.7% 144 –
Danio_rerio.3.53546837.53546992.fas – Alignment of 43 sequences 156 44.7% 43 ... 156 linear 69.9% 153 –
Danio_rerio.3.40219141.40218980.fas – Alignment of 43 sequences 156 52.5% 43 ... 156 linear 59.6% 144 –
Danio_rerio.3.35732898.35731709.fas – Alignment of 43 sequences 309 61.0% 43 ... 309 linear 48.5% 178 –
Danio_rerio.3.35707286.35707503.fas – Alignment of 43 sequences 120 49.6% 43 ... 120 linear 60.0% 120 –
Danio_rerio.3.21202553.21202392.fas – Alignment of 43 sequences 162 48.6% 43 ... 162 linear 49.4% 159 –
Danio_rerio.2.35330492.35330621.fas – Alignment of 43 sequences 129 46.2% 43 ... 129 linear 58.1% 118 –
Danio_rerio.2.35328234.35328360.fas – Alignment of 43 sequences 126 51.7% 43 ... 126 linear 62.7% 126 –
Danio_rerio.2.34849711.34849828.fas – Alignment of 43 sequences 114 42.5% 43 ... 114 linear 66.7% 114 –
Danio_rerio.2.29896263.29896097.fas – Alignment of 43 sequences 159 49.9% 43 ... 159 linear 32.1% 135 –
Danio_rerio.2.19220880.19221005.fas – Alignment of 43 sequences 123 42.3% 43 ... 123 linear 47.2% 120 –
Hide
Alignment View
Distances
Text View
Lineage
Info
Extract
R.C.
Translate
Add/Edit Annotation
Allow Editing
Annotate & Predict
Save
Consensus
Identity
1. 46QUNI
2. 01RHAS
3. 02ODPO
4. 03MISW
5. 04TYPA
6. 05MIVE
7. 06GODO
8. 07ELAC
9. 08DOMA
10. 09BUKO
11. 10KRNA
12. 11OXMA
13. 12AMYA
14. 13AMWH
15. 14POMI
16. 16GORO
17. 17VAPU
18. 18VAST
19. 19STNE
20. 20MIDO
21. 21ACFL
22. 22ACOM
23. 23CTBO
24. 24MUCA
25. 25MUAB
26. 26TRBA
27. 27TRBI
28. 28RHGI
29. 29RHDA
30. 30LOOC
31. 32PEKA
32. 33PEMO
33. 34BOPE
34. 35TASP
35. 36TRVA
36. 37ODLA
37. 38SICR
38. 39KUGU
39. 40SPOR
40. 41PTKA
41. 42SAFA
42. 43ECBI
43. 44SICH
88 %
Display
Consensus
Threshold: 0% – Majority
Ignore Gaps
If no coverage call ?
Highlighting
All Disagreem... to Consensus
Go: in any sequence
Use dots
Nucleotides
Complement
Translation on All Sequences
Translation Options
Frame: Frame 1
Genetic Code: Standard
Relative to: Sequence
Colors: ARND~
Edit
Three letter amino acids
Alt click on a sequence position or annotation, or select a region to zoom in. Alt-shift click to zoom out.

Next Steps

- Continue exploring alignment using Bowtie
- Complete the rest of the protocol (ongoing)
- Run protocol on flatfish data (Dec. 2018)
- Workshop on methods in Shanghai (Dec. 2019)
- Get results for flatfish data (Jan. 2019)