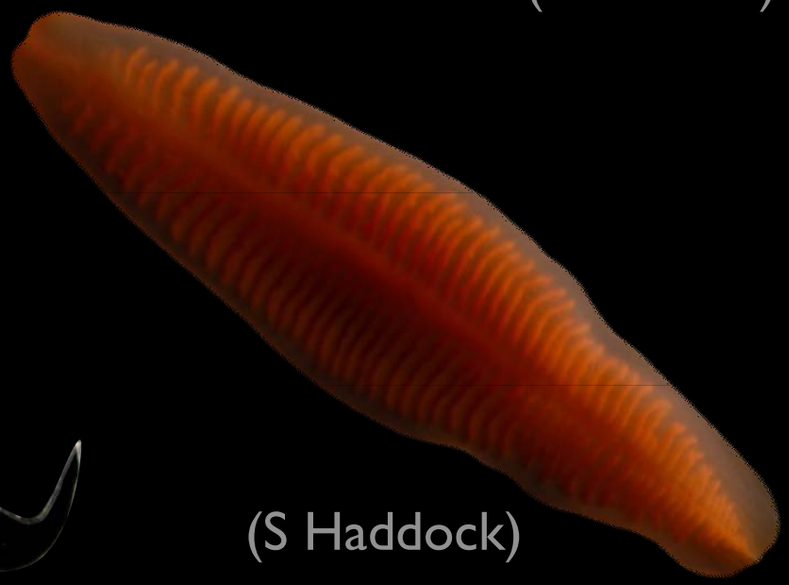
Phylogenomics

Casey Dunn
Assistant Professor
Ecology and Evolutionary Biology

(G Giribet)



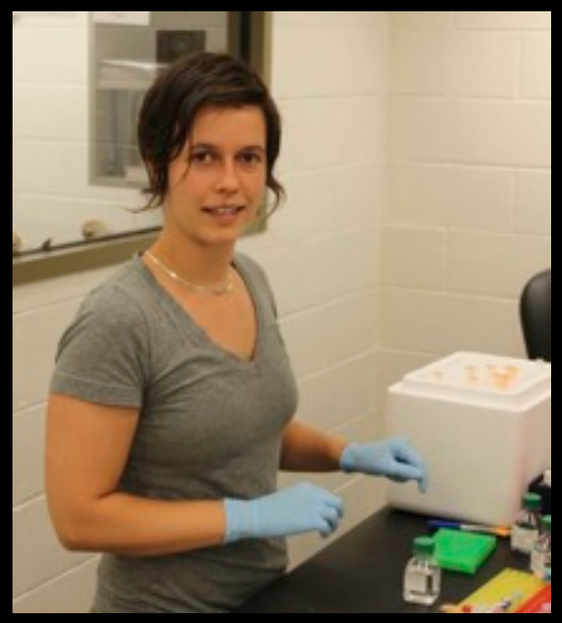
(S Haddock)

Casey Dunn

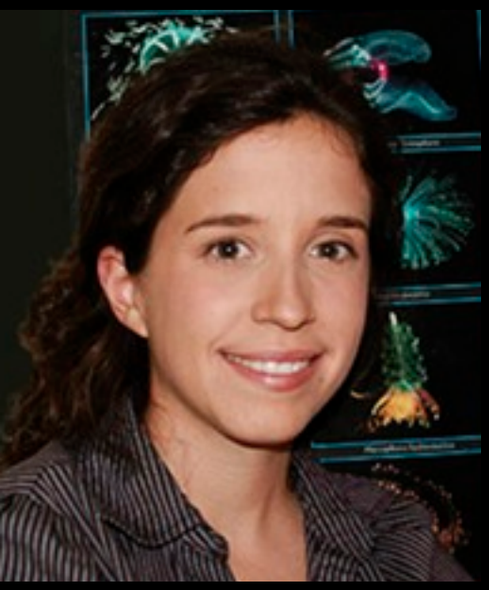
Stefan Siebert
(postdoc)

Sophia Tintori
(tech)

Stephen Smith
(postdoc)



Freya Goetz
(tech)



Rebecca Helm
(grad student)

Collaborators

Steve Haddock

Gonzalo Giribet, A. Stamatakis, Protostome AToL team, Cnidarian AToL team, Mark Robinson

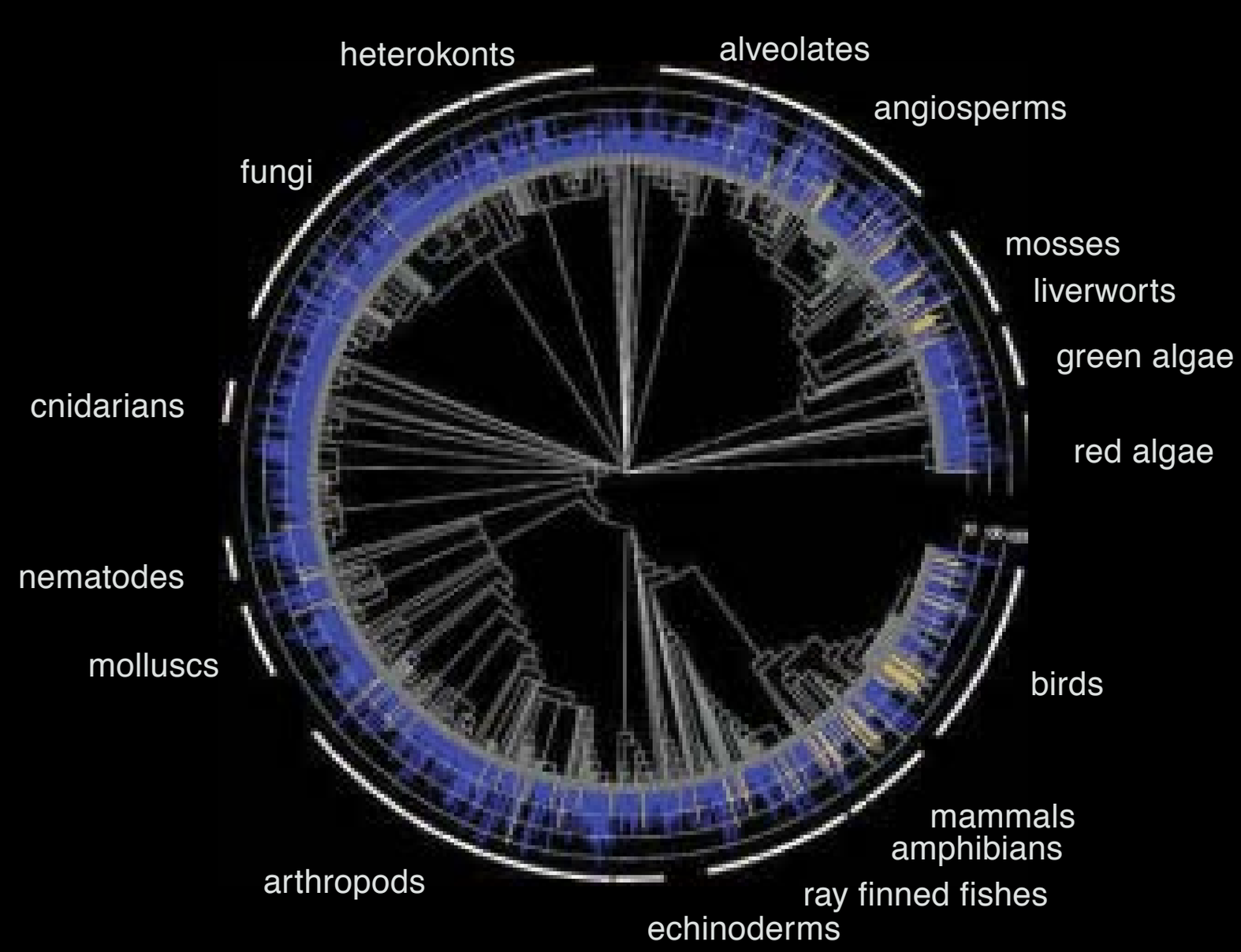
What does “phylogenomics” mean?

1. The study of genome evolution in a phylogenetic context

2. The inference of species phylogenies with genome data

3. The inference of species phylogenies with data from lots of genes

"...stronger sampling effort aimed at genomic depth, in addition to taxonomic breadth, will be required to build high-resolution phylogenetic trees at [a broad] scale."



Sanderson, 2008
doi:10.1126/science.1154449

2.6 million sequences
1127 taxa

Why collect data from lots of genes?

- Many hard problems will require lots of data

- Lots of data makes some aspects of inference easier

- These data are useful for things besides building trees

- It can be much cheaper to collect a lot of data than a little bit of data

DNA sequencing

Helicos Roche Illumina

Current Illumina costs:

$2,095 for one lane (HiSeq)

Paired-end 100bp

~150 million clusters

30 gigabases of data

Current Illumina costs:

- ~$120 per sample to prepare a library

Current Illumina costs:

Samples per lane	Cost per sample	Clusters per sample (millions)	Gigabases per sample
1	$2,215	150	30
4	$644	37.5	7.5
8	$382	18.75	3.75
12	$295	12.5	2.5

Current Illumina costs:

~$100 a gigabase

$0.0000001 per base

Will cheap sequence data allow us to answer all our questions?

Of course not.

Should we approach problems with more data or improved analysis methods?

This is a false dichotomy.

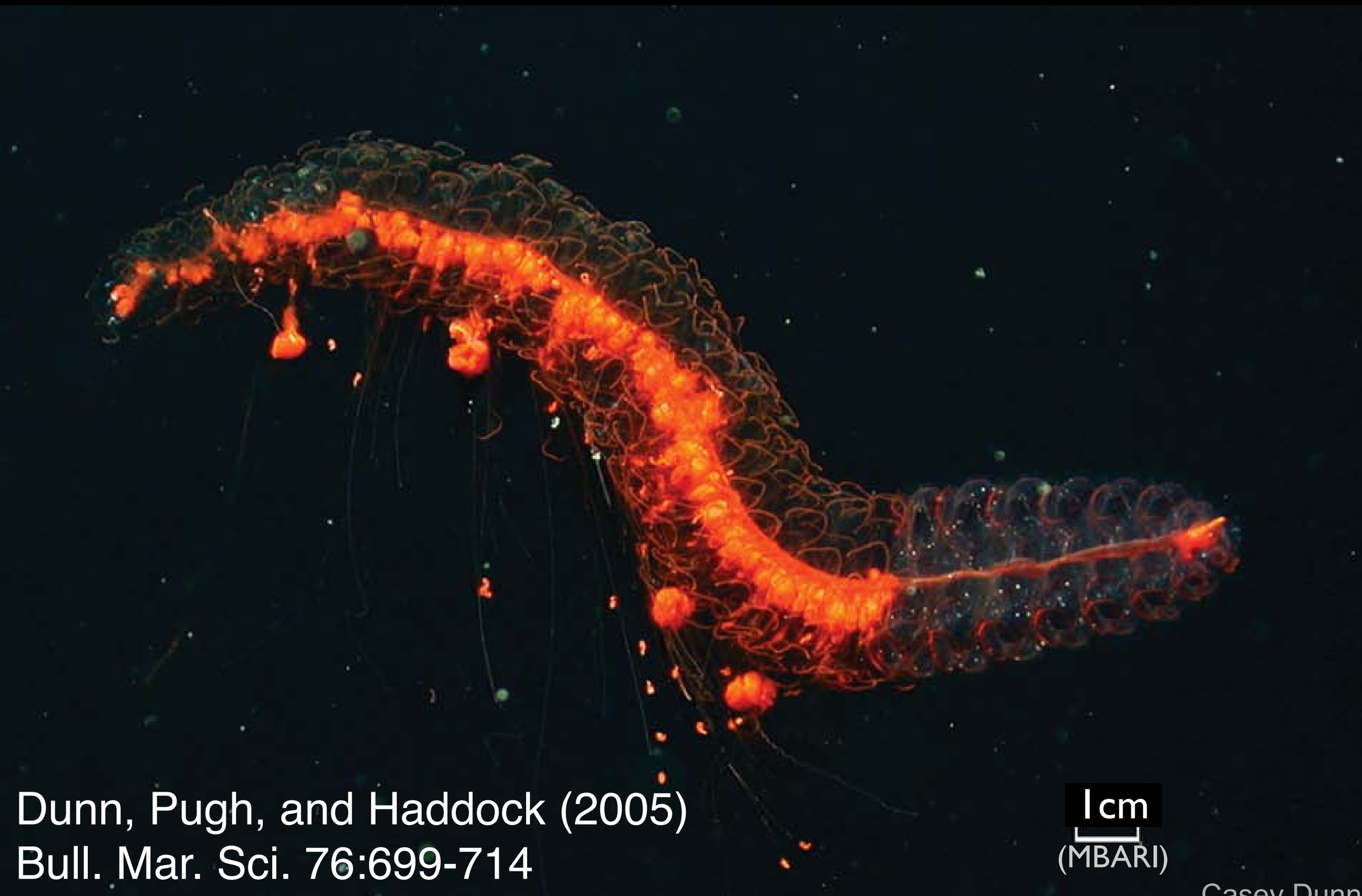
We need both!

Are other types of data now obsolete?

No!

These data open entirely new opportunities for integrating genomic, morphological, and functional perspectives

Marrus claudanielis



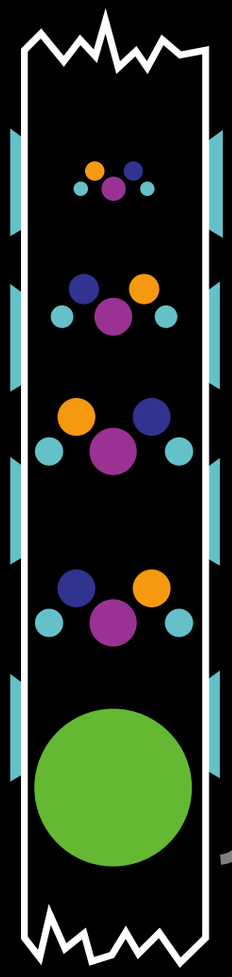
Dunn, Pugh, and Haddock (2005)
Bull. Mar. Sci. 76:699-714

(MBARI)

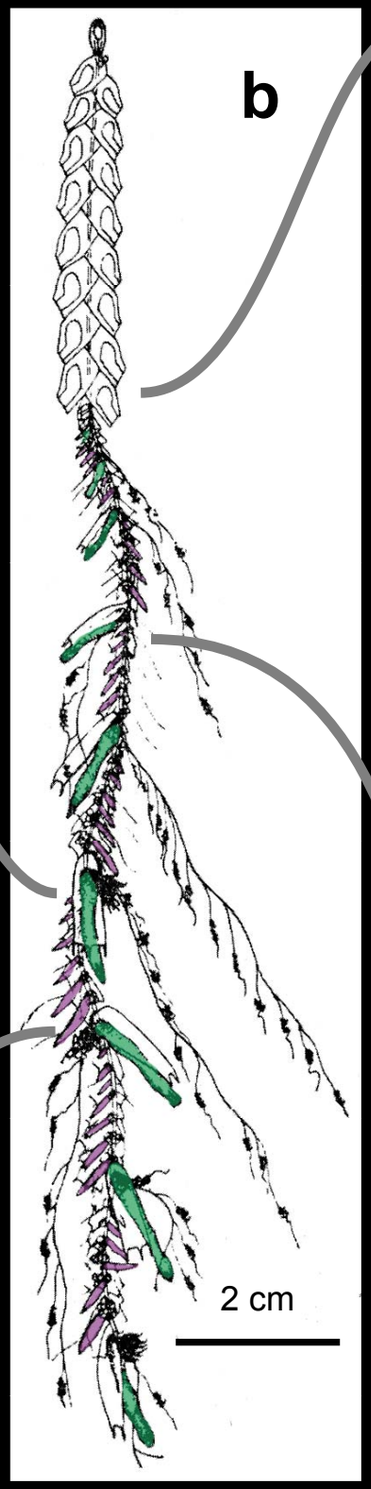
Casey Dunn

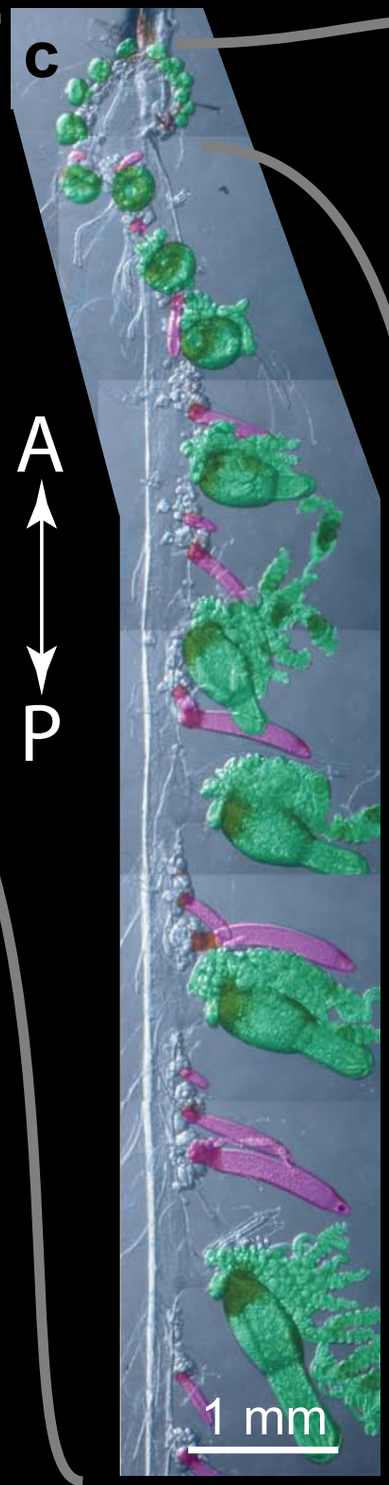
“Physonects”: *Nanomia bijuga*

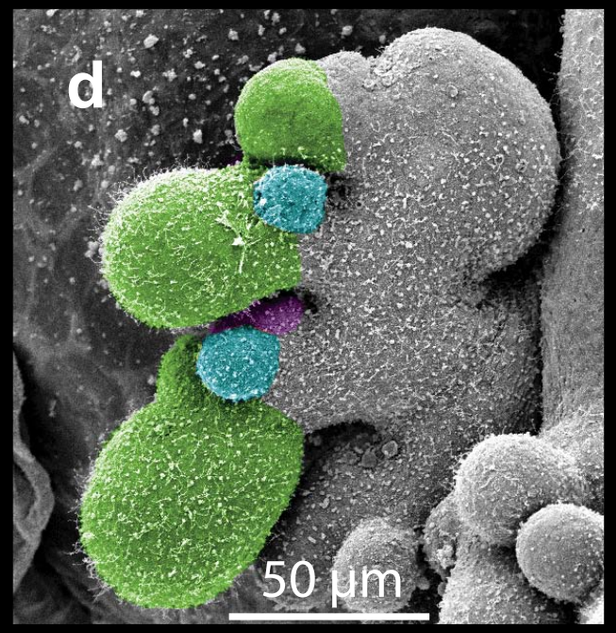
Zooids:
- Bract
- Female
- Male
- Excretory
- Feeding

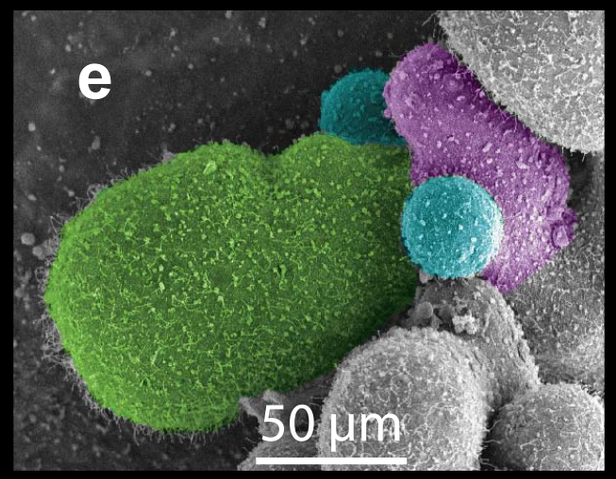


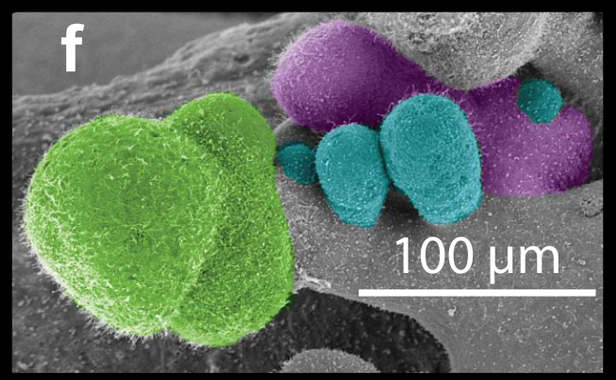
R ⟷ L



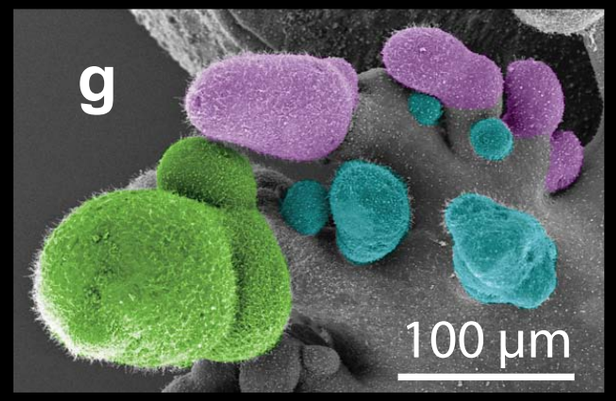


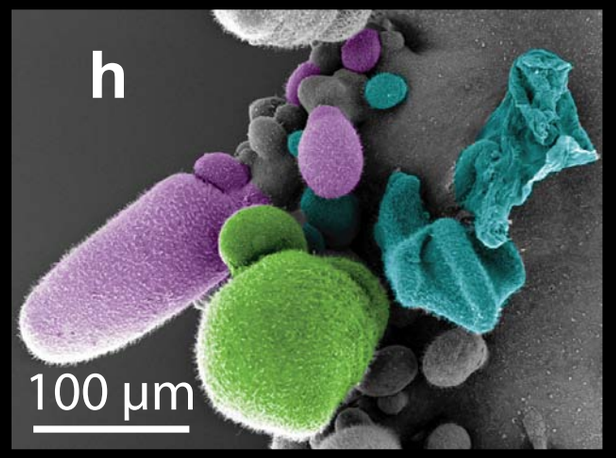


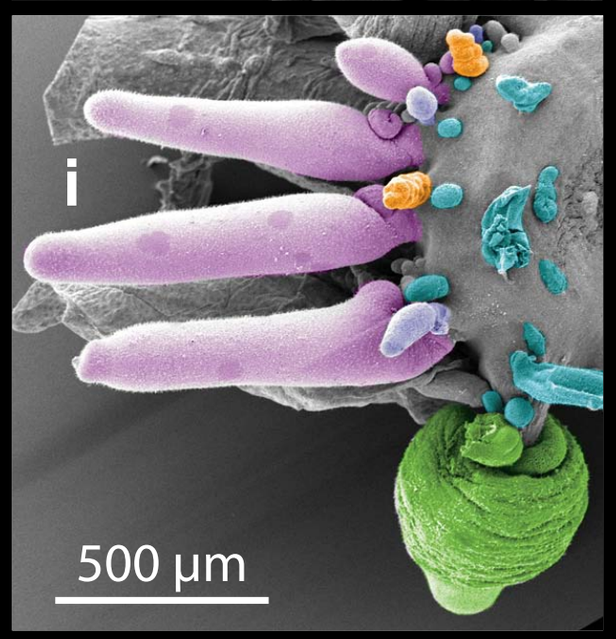




V ⟷ D







V ⟷ D

Dunn and Wagner (2006) Development, Genes, and Evolution 216:743-754

Casey Dunn

“More Isn't Just More—More Is Different”

Wired,June 23, 2008

Gene selection as part of project design (Directed PCR)

- Select genes
- Amplify and sequence selected genes
- Assemble matrix from all sequenced genes
- Phylogenetic inference

Gene selection as part of data analysis (ESTs, shotgun genomes)

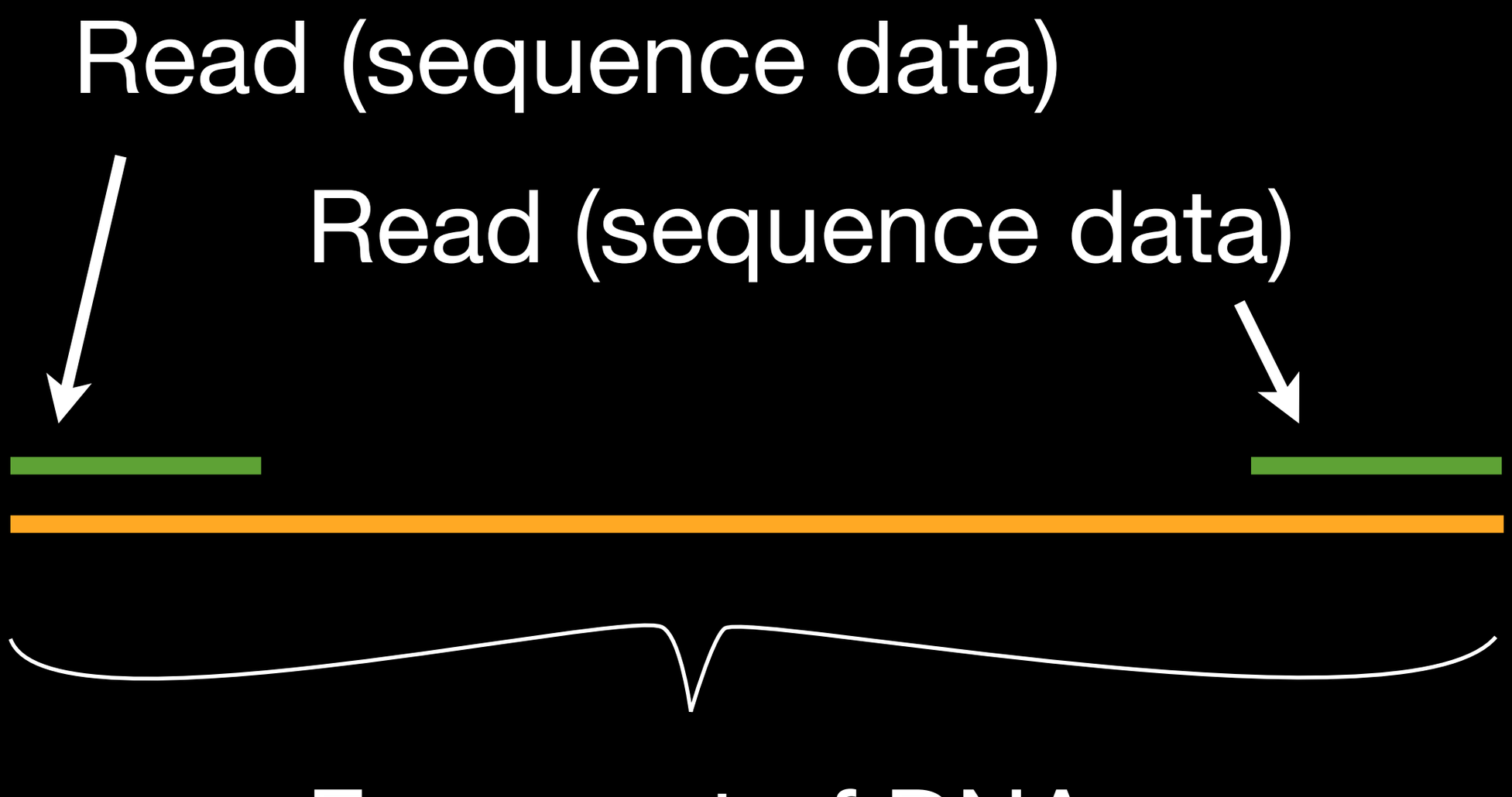
- Sequence at random
- Identify homologous sequences and evaluate paralogy
- Select genes
- Assemble matrix
- Phylogenetic inference

Casey Dunn

Getting data...

DNA sequencing

Helicos Roche Illumina



Fragment of DNA

DNA Fragments can be:

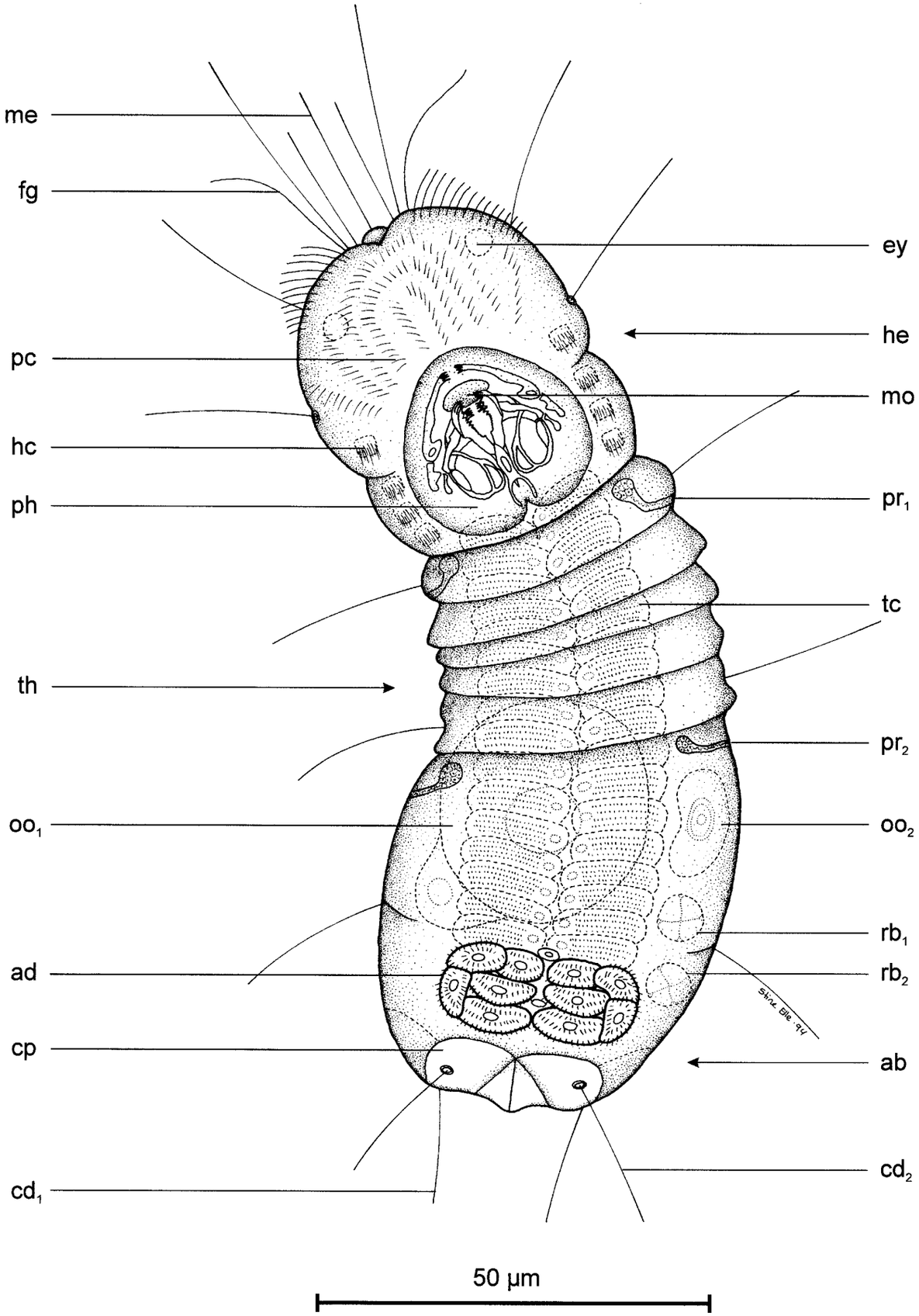
- Amplified/ enriched gene regions
- Genomic DNA
- cDNA (Transcriptomes)

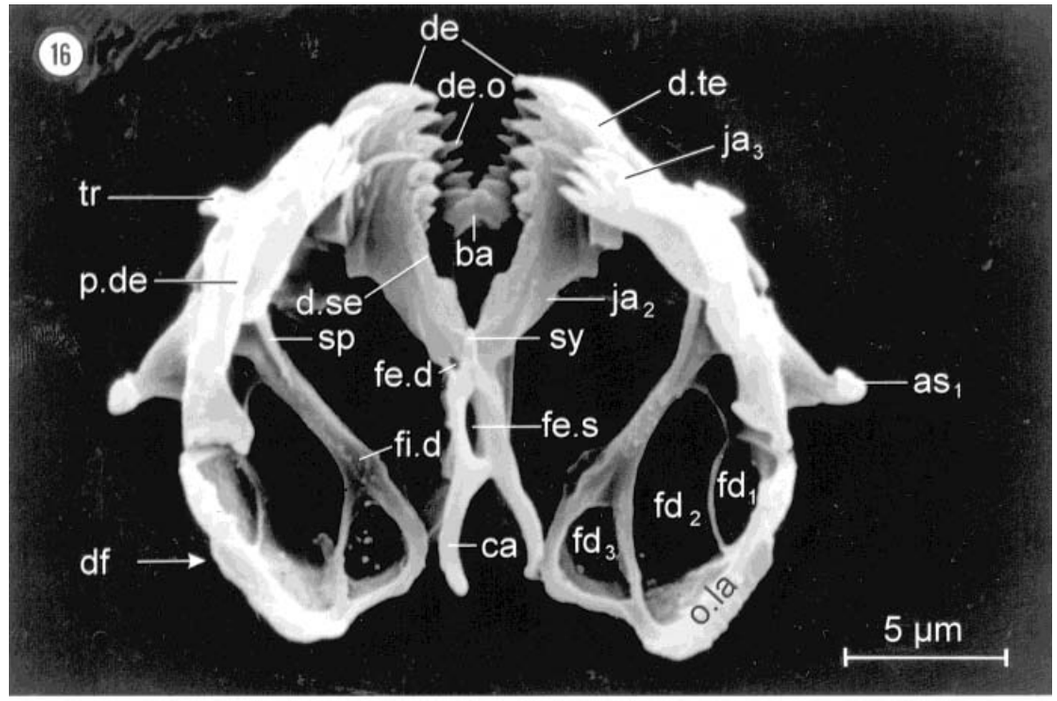
Genome	Transcriptome
Start with DNA	Start with mRNA
Get all genes, regulatory regions, etc	Get a snapshot of active genes
Genomes can be really big	Almost all data is from coding regions
Can be hard to identify genes	Handling RNA is tricky

Overview of sequencing:

- Get DNA or RNA
- Make a library (chunks of DNA with adapters)
- Prepare library for sequencing
- Sequence
- Process data into raw reads


Casey Dunn





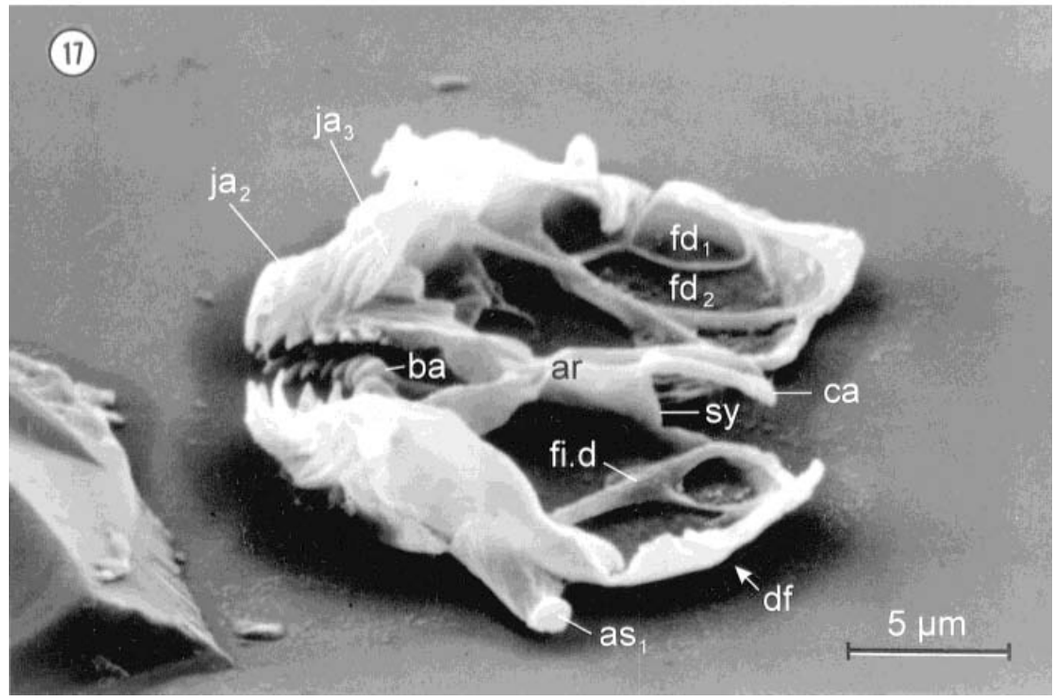
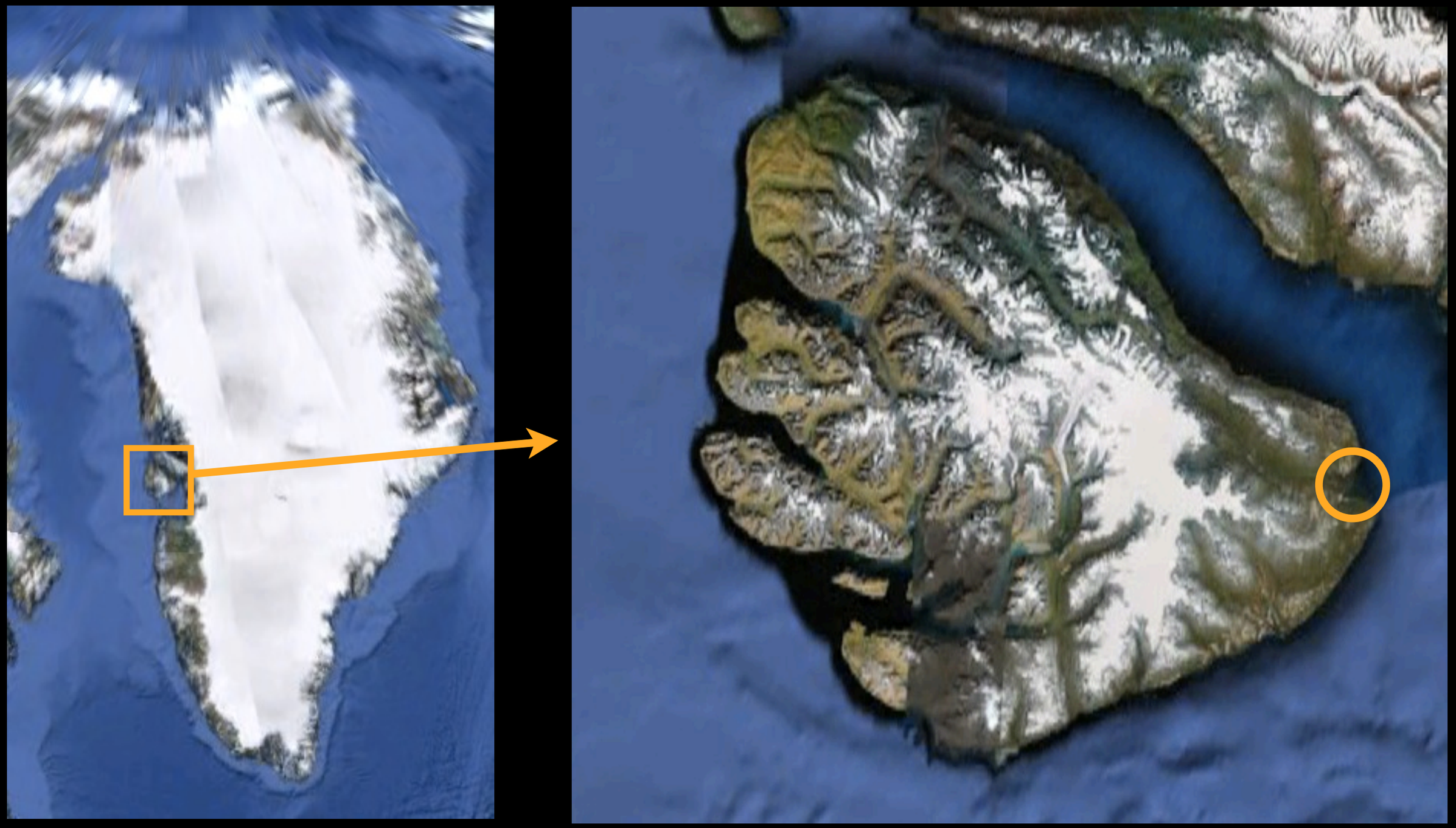


Figure 2

(Kristensen and Funch, 2000)



Casey Dunn

Casey Dunn


Casey Dunn

Some options for preservation

Freeze tissue (-80C or colder)

RNALater (Ambion), kept cold

Extract RNA in the field

Homogonize in Trizol, keep cold

Casey Dunn

mRNA isolation - Lots of tissue

Isolate Total RNA with Trizol

Digest DNA

Isolate mRNA (eg NEB S1550S, Dynabeads mRNA Kit)

mRNA isolation - Small amount of tissue

- mRNA straight from tissue (eg NEB S1550S, Dynabeads mRNA DIRECT Kit)

mRNA isolation - Minute amount of tissue

- Linear cDNA amplification (eg NuGen Ovation)

RNA quality is (almost) Everything!

- Avoid contamination

Reduced sample size requirements have improved this

RNA quality is (almost) Everything!

Quantity matters - be cautious working at the bottom range of sample requirements

RNA quality is (almost) Everything!

Amount of ribosomal RNA matters

There are tradeoffs between rRNA fraction and yield. If material is limiting, purify less and sequence more

RNA quality is (almost) Everything!

If you have enough RNA, look at its size distribution with a BioAnalyzer or similar tool.

What do you do once you have high quality DNA or RNA?

Break it into little pieces!


Read
Read
DNA

Starting material

Fragment

Prepare library, sequence

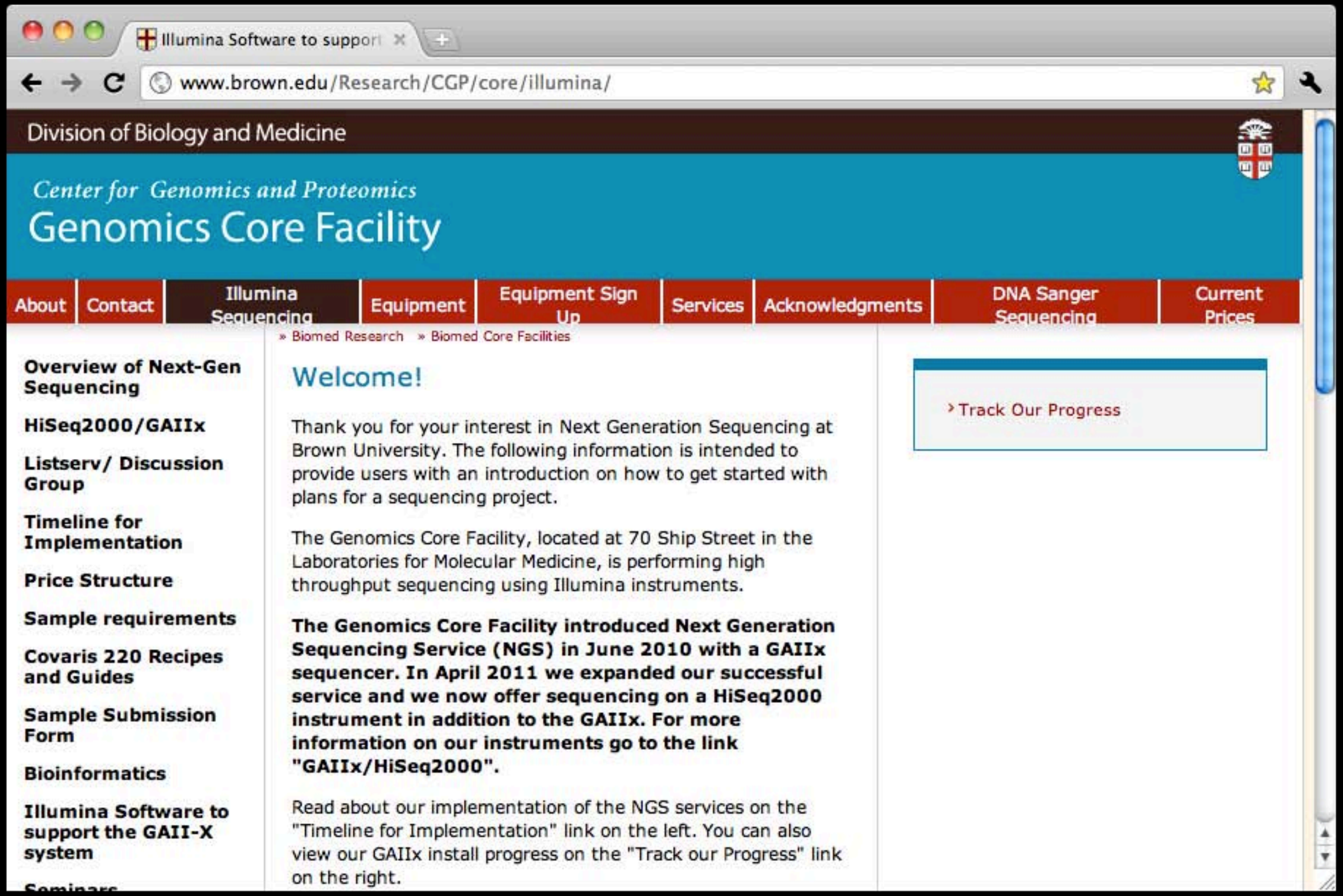
Library preparation options

- Get a library preparation kit from the sequencer vender
- Get a third party library preparation kit
- Make the library from scratch

These days, in my lab we use:

- TruSeq RNA kit (Illumina)

- NEBNext (NEB)



Prices at:
www.brown.edu/Research/CGP/core/illumina/price

Data are usually delivered in fastq or qseq format.

fastq example:

```
@HWI-ST625:51:C02UNACXX:7:1101:1179:1962 1:N:0:TTAGGC
CTAGNTGTTGAAGAGAAGGTTCAAGAACCAAAAGAAAGCTCACAACAACATATGGT
+
=AAA#DFDDDHHFDGHEHIAFHHIIIIGICDGAGDHGGIHG@A@BFIHIIIGC@@8

@HWI-ST625:51:C02UNACXX:7:1101:1242:1983 1:N:0:TTAGGC
ATAATTTCAATGACTGGAGTAGTGAAAATGAACATAGATATGAGAATAACCGTAGA
+
ACCCFFFFFGHHHHJJJIJEHIFHIJJJJIJJJJJIIJIJJIIJJJJJJJJIIJJJ
```

These instruments generate a lot of data.

The data files from lane of an Illumina HiSeq are more than 70 GB.

A HiSeq has 16 lanes.

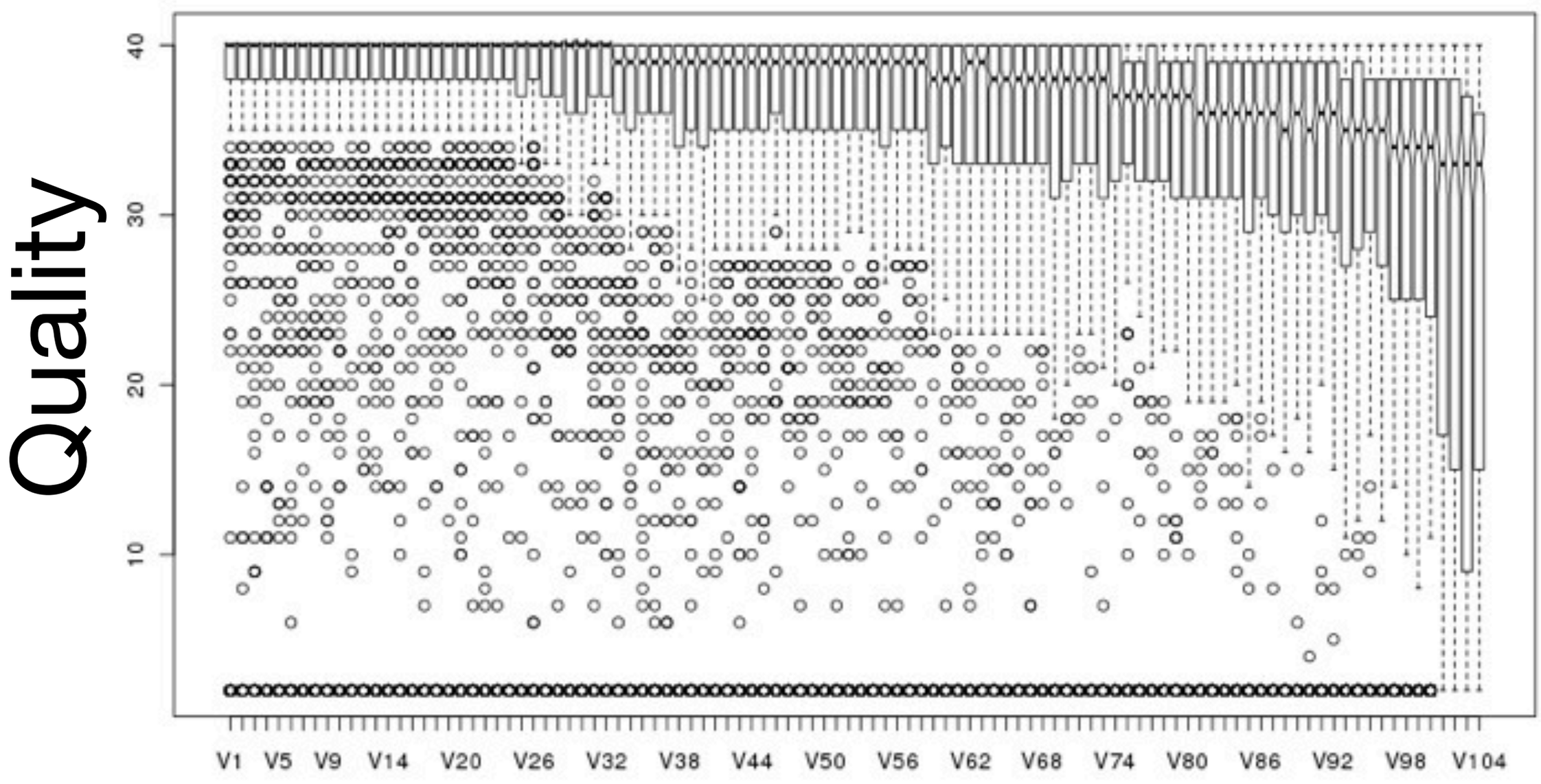
(Sam Fulcomer)

Casey Dunn

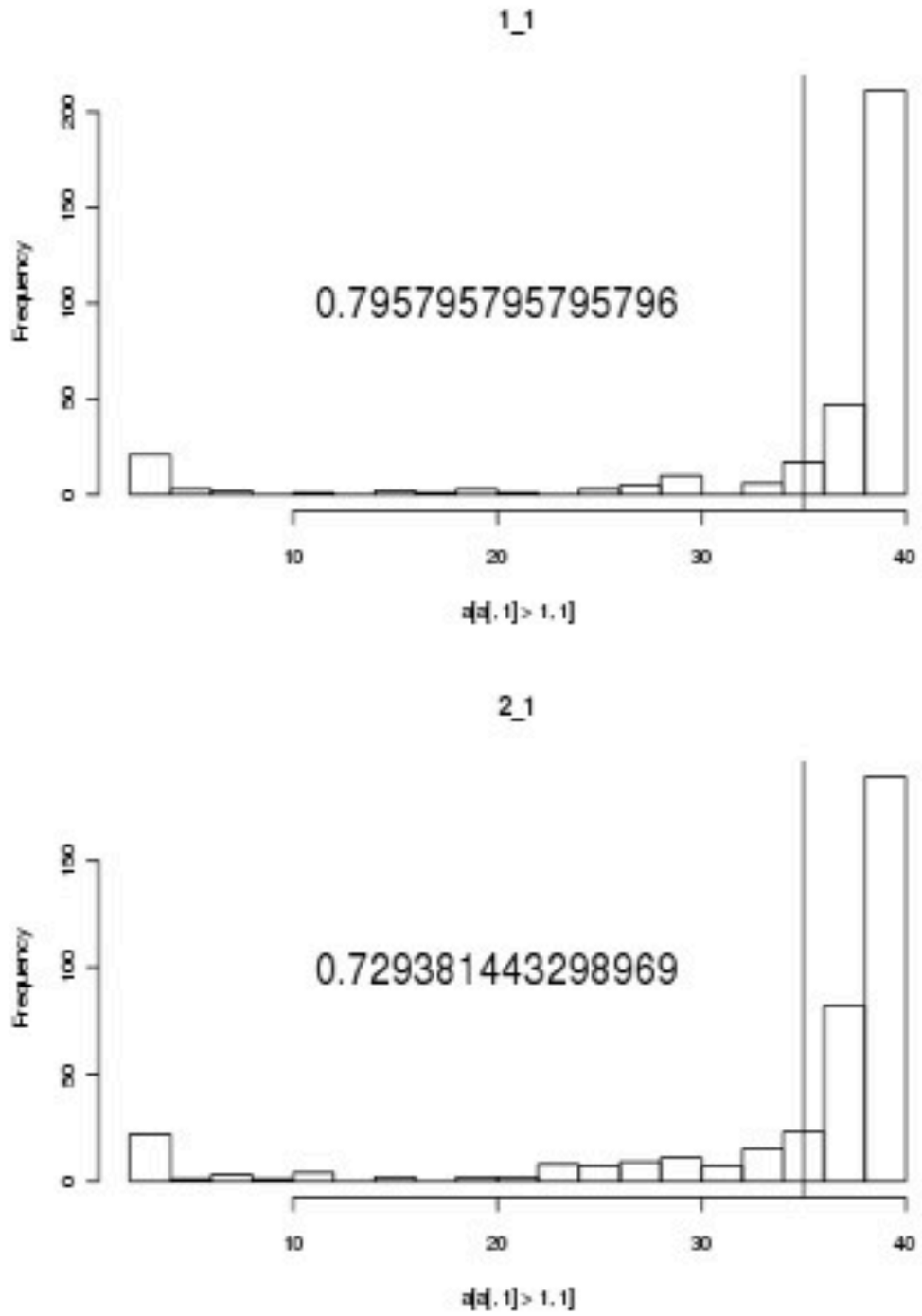
The first thing you'll want to do is to examine the quality profiles of your data.

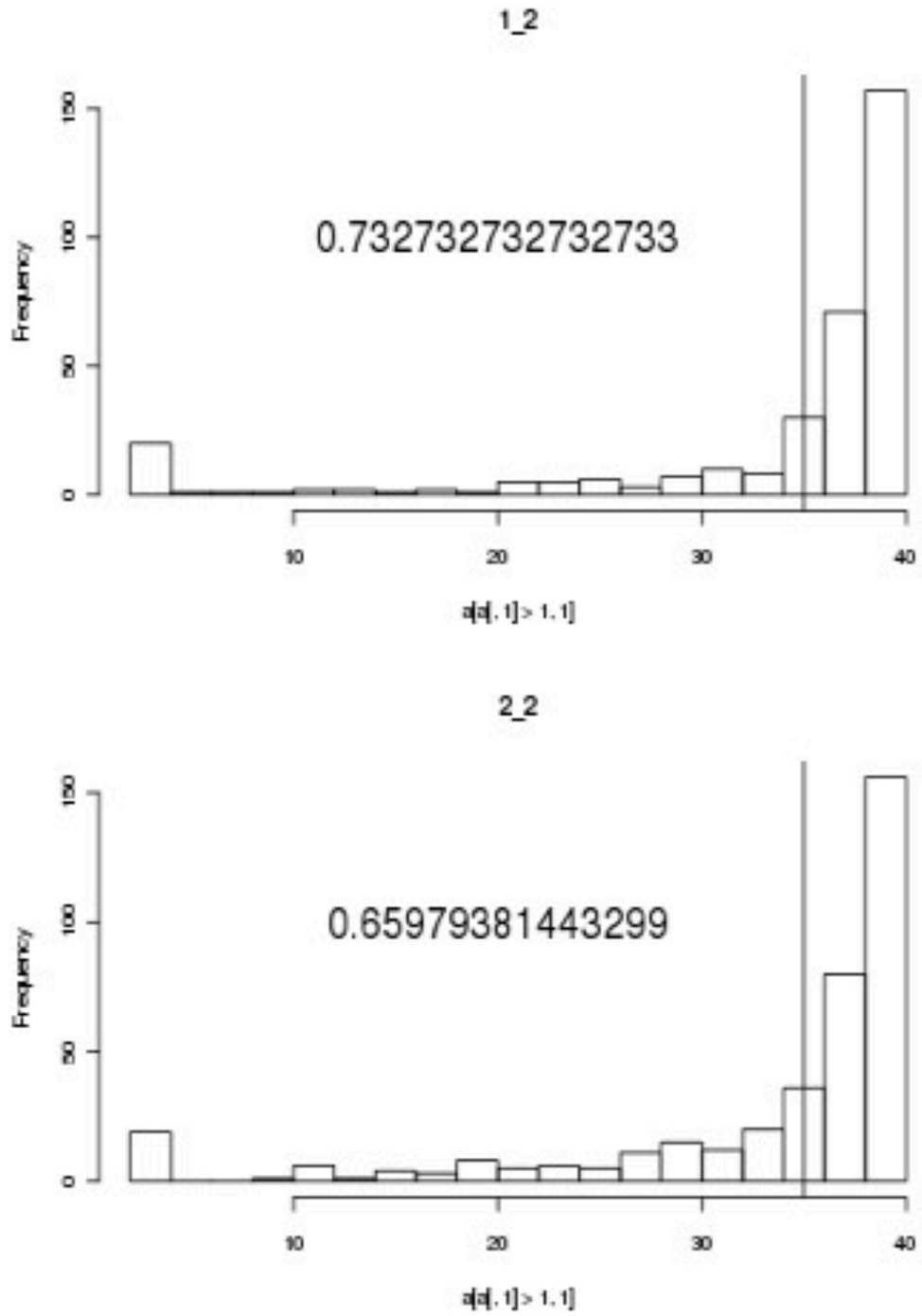
There are many tools that do this, we use python and R.

Plot the quality of each base across reads:



Make a histogram of the mean quality of each read:





Use these plots to decide if you need to trim off the end of reads or discard low quality reads.

Assembly

Starting material

Fragment

Prepare library, sequence

Assembly

Final product

Overlap assemblers that work fine on large sanger datasets don't scale to these very large data sets

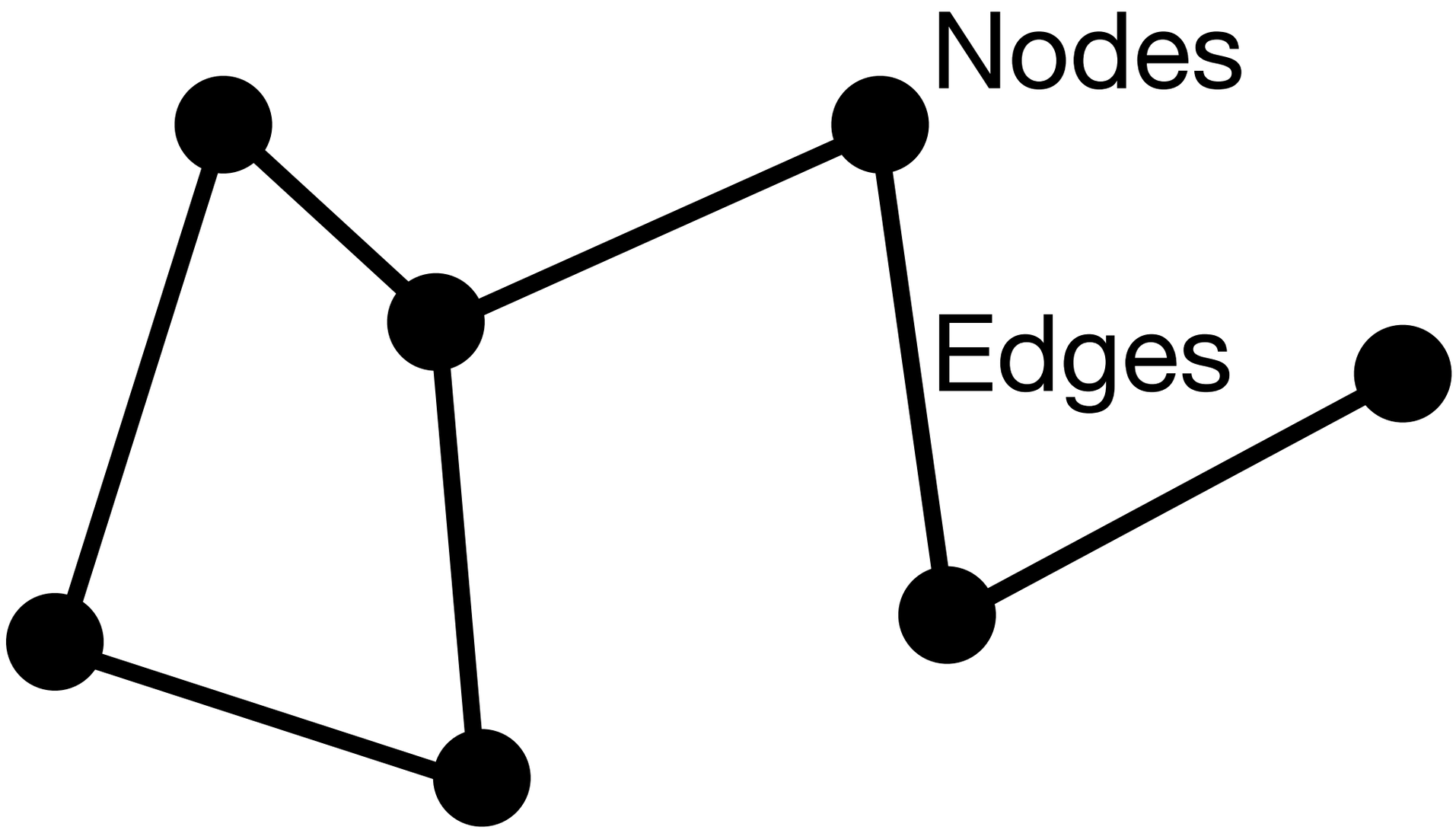
The number of pairwise comparisons that are needed to detect overlap become intractable

A new generation of de Bruijn graph assemblers have been developed to meet these challenges

Better defined memory footprint

Simpler comparisons between sequences

What is a graph?



The first step in de Bruijn graph assembly is breaking each read down into all sequences of k length

actgtcat ⟶

actg
ctgt
tgtc
gtca
tcat

There are 4^k possible k-mers

In practice, k is often in the 25-70 range

The k-mers are loaded into a hash table:

```
actg 1
ctgt 1
tgtc 1
gtca 1
tcat 1
```

A de Bruijn graph is constructed from the has table

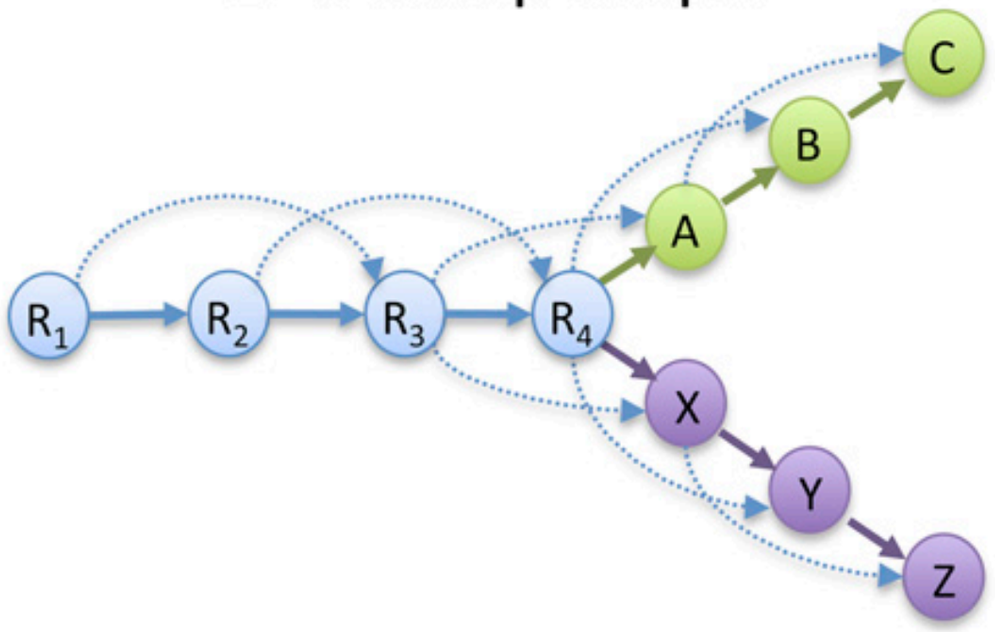
Each node corresponds to a k-mer sequence from the hash table

An edge unites each node that extends another node by one base pair

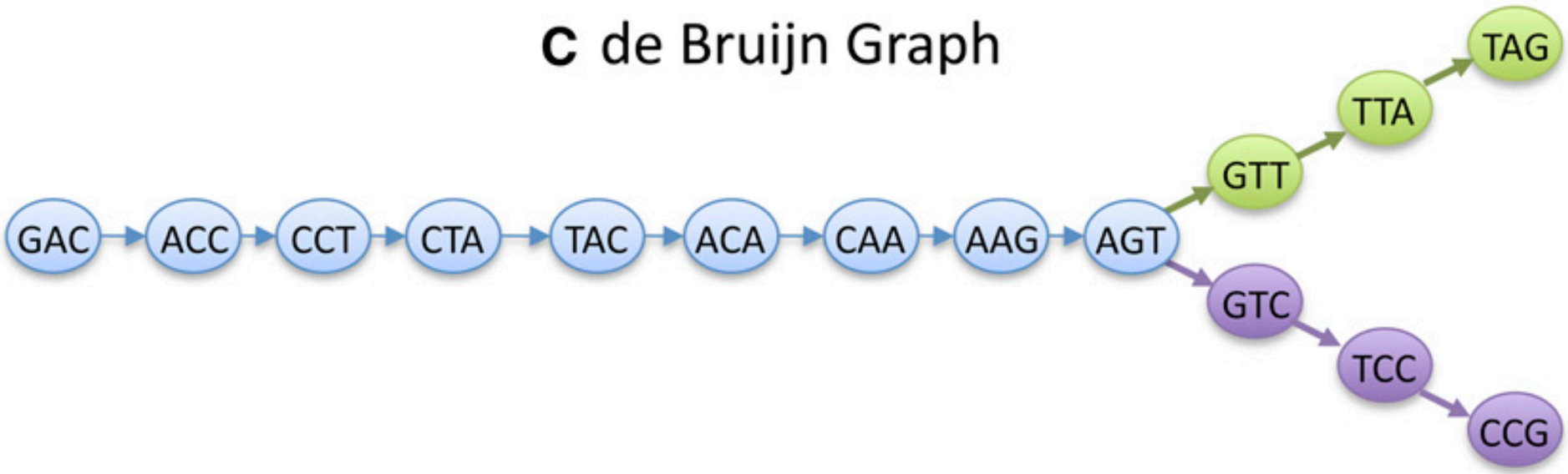
A Read Layout

```
R1: GACCTACA
R2:    ACCTACAA
R3:      CCTACAAG
R4:       CTACAAGT
A:          TACAAGTT
B:           ACAAGTTA
C:             CAAGTTAG
X:          TACAAGTC
Y:           ACAAGTCC
Z:             CAAGTCCG
```

B Overlap Graph



C de Bruijn Graph

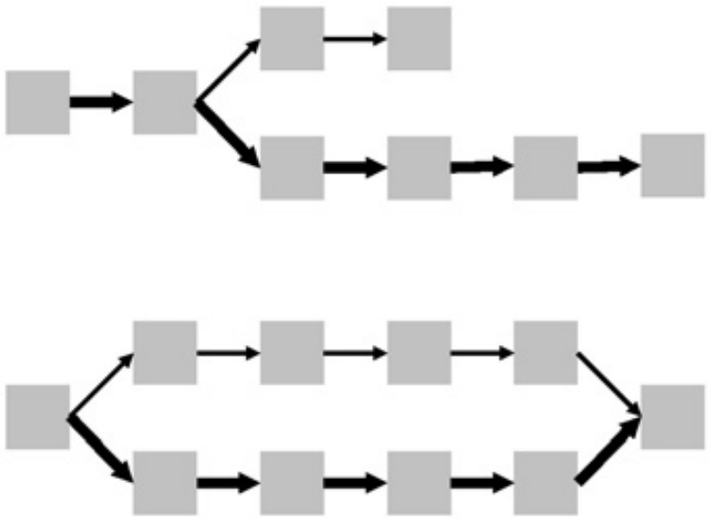


Schatz et al 2010 (dx.doi.org/10.1101/gr.101360.109)

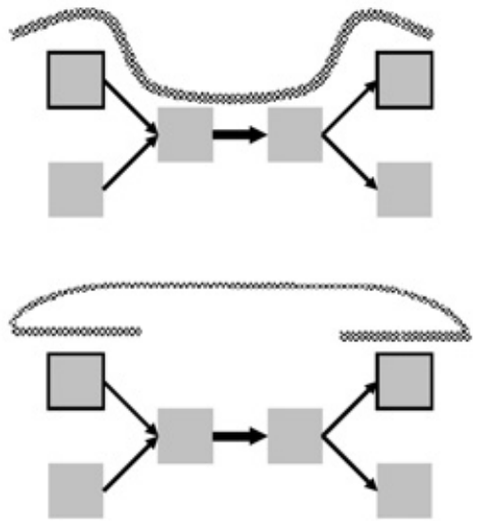
Paths through the de Bruijn graph are assembled sequences

These paths can be very complicated due to sequencing error, snp's, splicing variants, repeats, etc

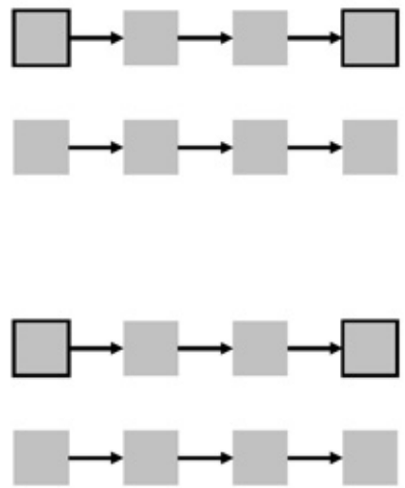
The graphs require considerable post-processing to simplify them (pop bubbles, trim dead ends, etc)

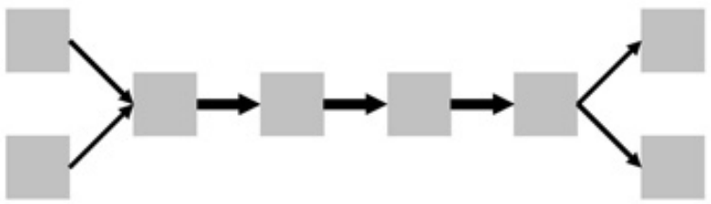


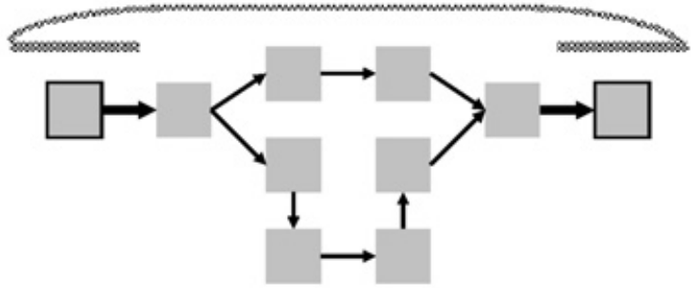
(before)

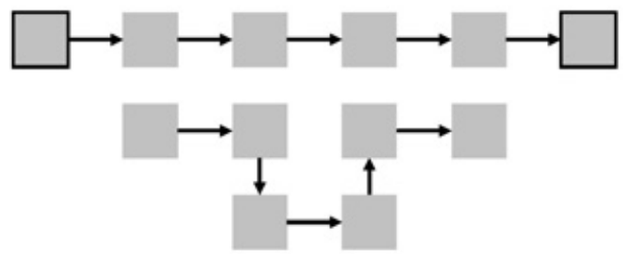


(after)









Miller et al 2010 (dx.doi.org/10.1016/j.ygeno.2010.03.001)

Assembly takes a lot of RAM!

One lane of Illumina HiSeq data can require hundreds of gigabytes of RAM to assemble

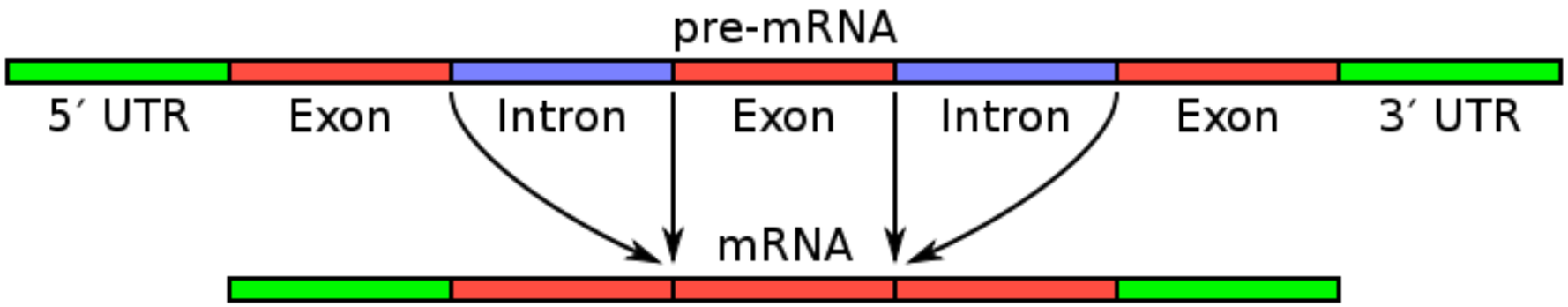
This is one of the largest challenges for using next-generation sequencing data to build trees

Eliminating low-quality data can greatly reduce RAM requirements

Genome	Transcriptome
Start with DNA	Start with mRNA
Get all genes, regulatory regions, etc	Get a snapshot of active genes
Genomes can be really big	Almost all data is from coding regions
Can be hard to identify genes	Handling RNA is tricky

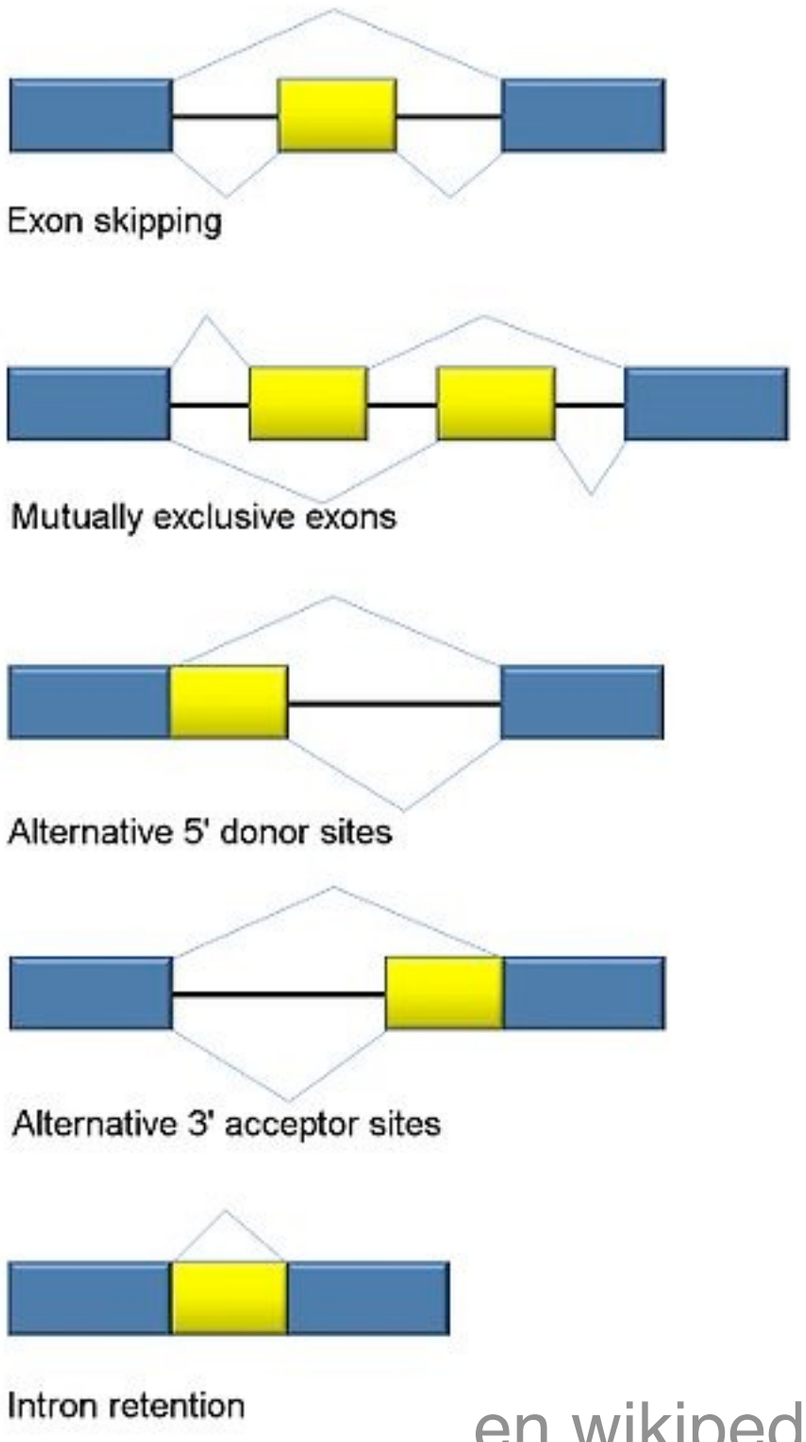
Transcript splicing

mRNA's are spliced before leaving the nucleus



en.wikipedia.org/wiki/File:Pre-mRNA_to_mRNA.svg

Transcript splicing



With deep sequencing, many splice variants are sequenced for each gene

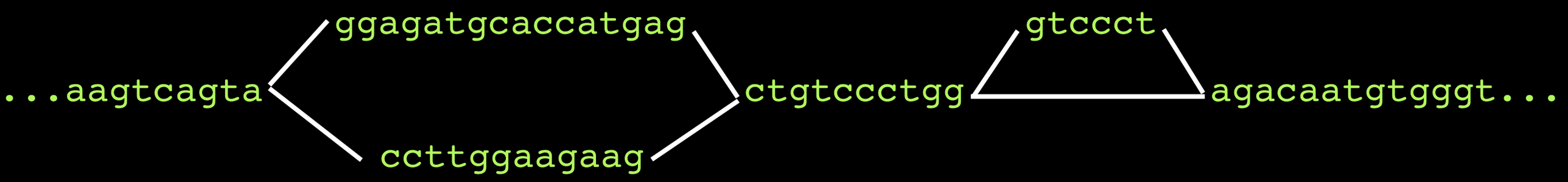
en.wikipedia.org/wiki/File:Alt_splicing_bestiary2.jpg

Assembly results...

Genome

```
...aagtcagtggagatgcaccatgagaccttggaagaagctgtccctggagacaatgtgggt...
```

Transcript



Splice variants

-Different splice variants for a given gene can vary widely in abundance

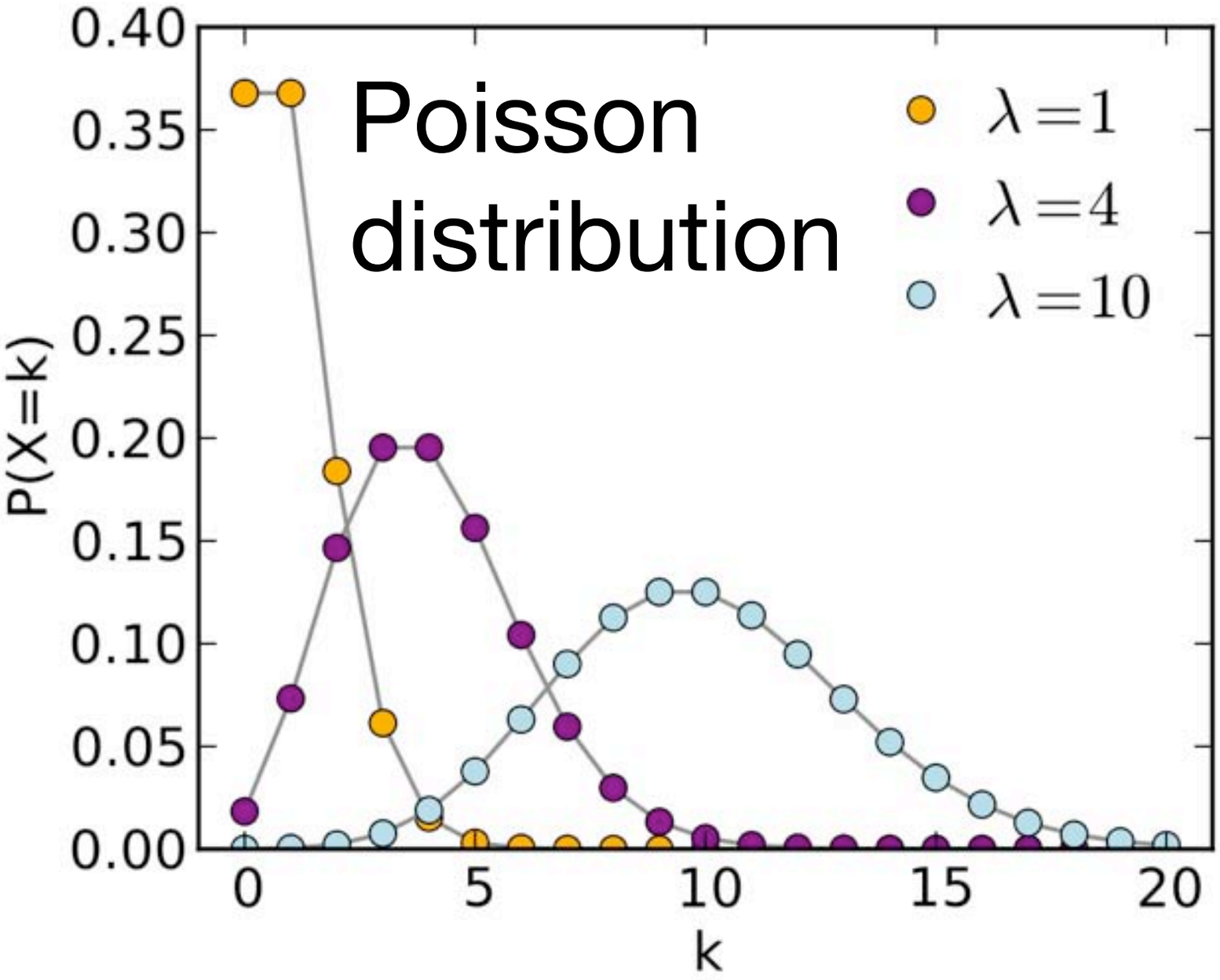
-Deep sequencing captures some "intermediate splice variants", molecules in the process of being spliced

-Sequencing and assembly errors can be misinterpreted as splice variants

-Data may be insufficient to predict splice variants

It gets worse...

Genomes are uniform depth



en.wikipedia.org/wiki/File:Poisson_pmf.svg

Assemblers can make assumptions about uniform distribution of sequencing effort

Expression differences mean:

- Can’t assume that the expected frequency of sequences is uniform across or even within genes

- Low copy number doesn’t necessarily indicate an error

- High copy number doesn’t necessarily indicate a repeat

- Sequencing error is hard to accomodate in transcriptomes

When assembling transcriptomes, it is essential to use an assembler that can explicitly accommodate splice variants and expression differences!!!!!

Transcriptome assemblers include:

Newbler (Roche)

Oases (www.ebi.ac.uk/~zerbino/oases)

Trinity (trinityrnaseq.sourceforge.net)

TransAbyss (www.bcgsc.ca/platform/bioinfo/software/trans-abyss)

Biological concept	Newbler term	Oases term
Gene	isogroup	locus
Splice variant	isotig	transcript
exon	contig	contig

Post-assembly annotation:

- Selection of exemplar transcripts for each gene
- Blastx to a taxon restricted subset of the NCBI nr database
- Translation with prot4est

Identifying homologs

Phylogenetic tools build trees from homologous characters

Most phylogenetic tools assume character homology, they can't evaluate it

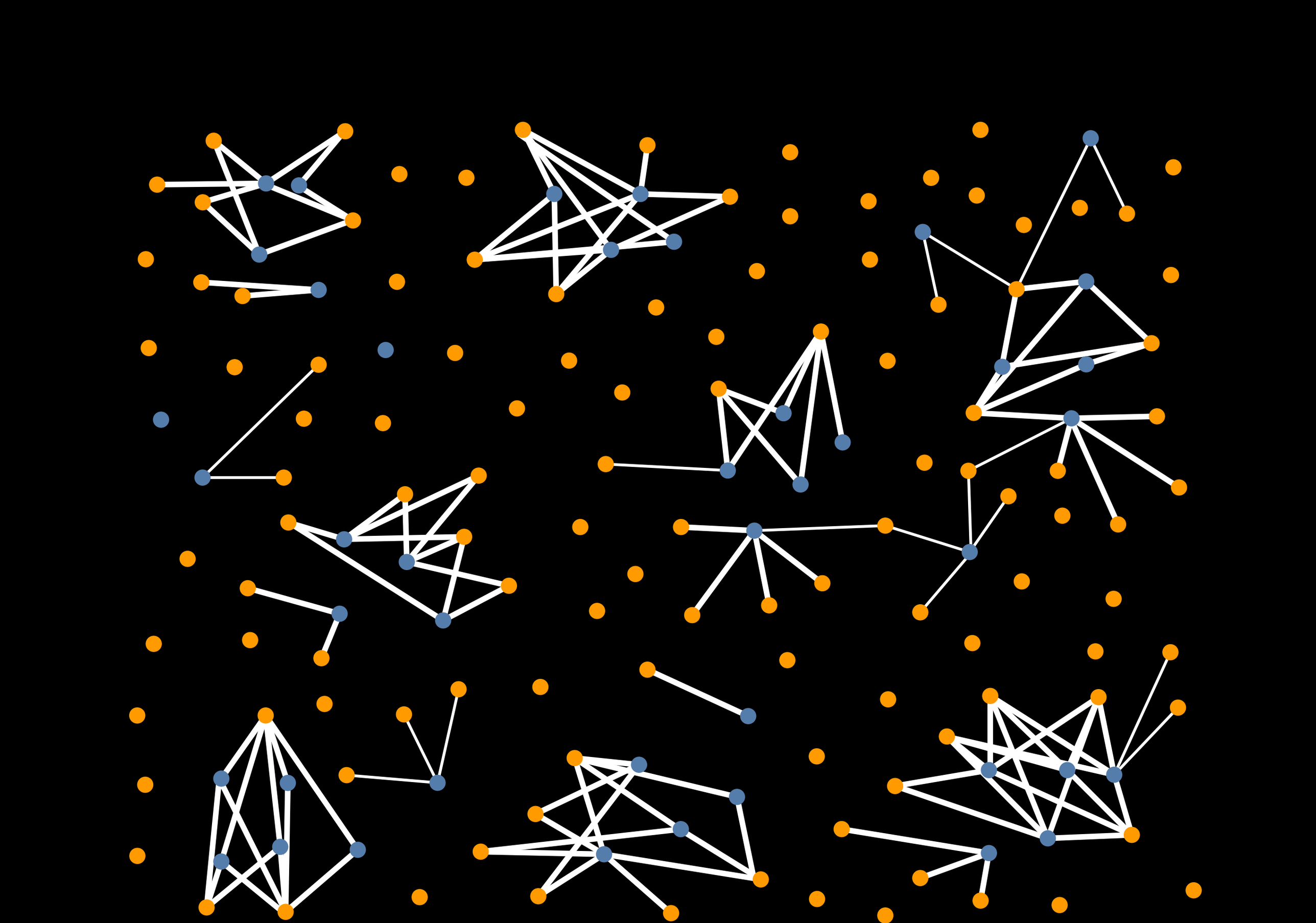
We need to make a first pass with phenetic tools

Throw all sequences for all taxa in a study into a hat

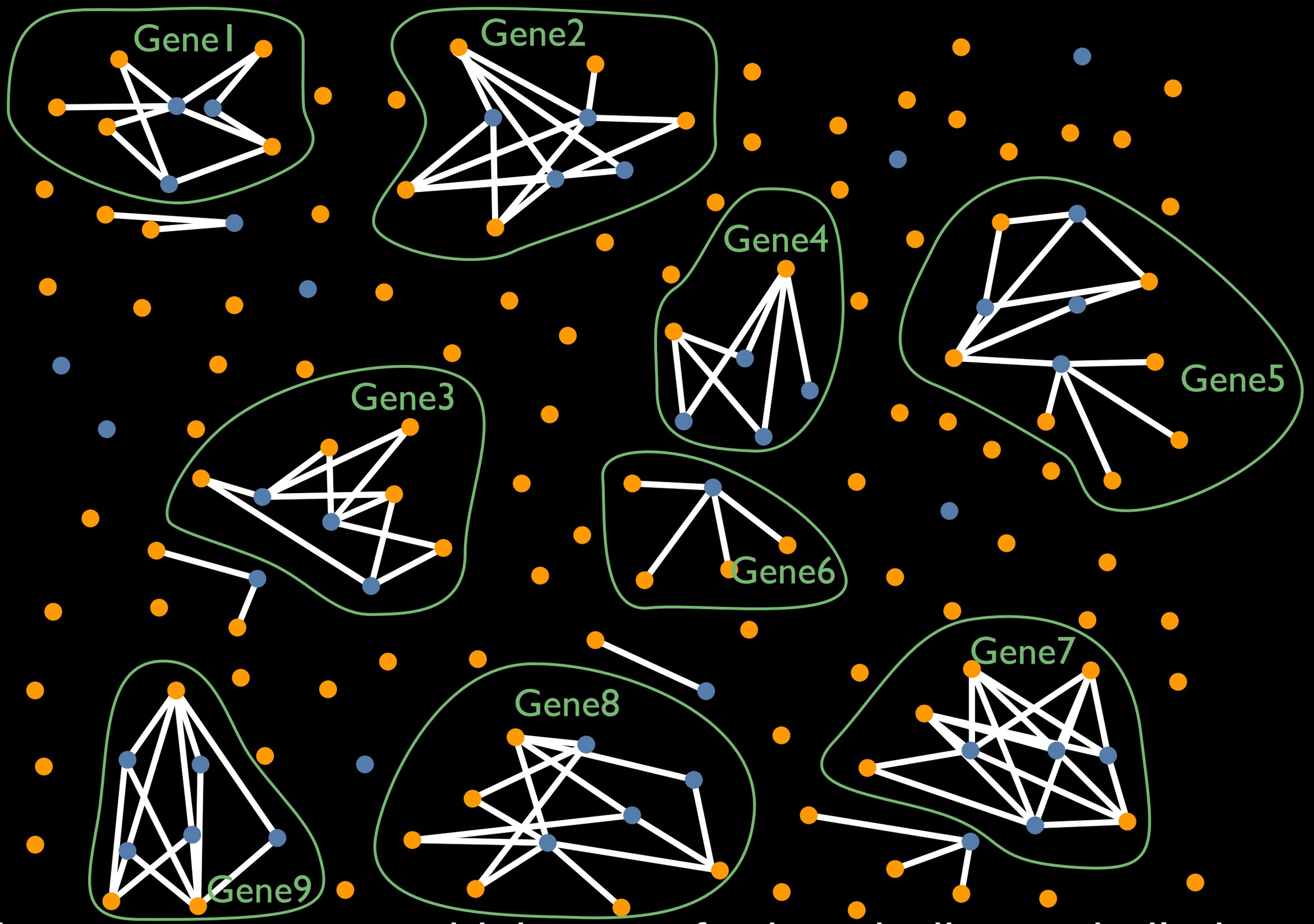
Make all pairwise sequence comparisons (eg blast)

Construct a graph where nodes are sequences and edges indicate similarity

Nodes are sequences, thickness of edges indicate similarity



Nodes are sequences, thickness of edges indicate similarity



Nodes are sequences, thickness of edges indicate similarity

Casey Dunn

We use:

Blastp on a cuda array, assign $-\log_{10}$(e-value) for edge weight

Throw away edges < 20

mcl for clustering (inflation ~2.1)

Apply taxon sampling criteria

BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 3 2011, pages 326–333
doi:10.1093/bioinformatics/btq655

Sequence analysis

Advance Access publication November 29, 2010

Improving the quality of protein similarity network clustering algorithms using the network edge weight distribution

Leonard Apeltsin[1], John H. Morris[1], Patricia C. Babbitt[1,2] and Thomas E. Ferrin[1,2,*]

[1]Department of Pharmaceutical Chemistry and [2]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA, USA

Associate Editor: Burkhard Rost

Identifying orthologs

Casey Dunn

“The paralogy problem”

But paralogs aren’t inherently a problem

The problem is misascribing paralogs as orthologs

Species A

Species B

Species C

Gene divergence due to duplication

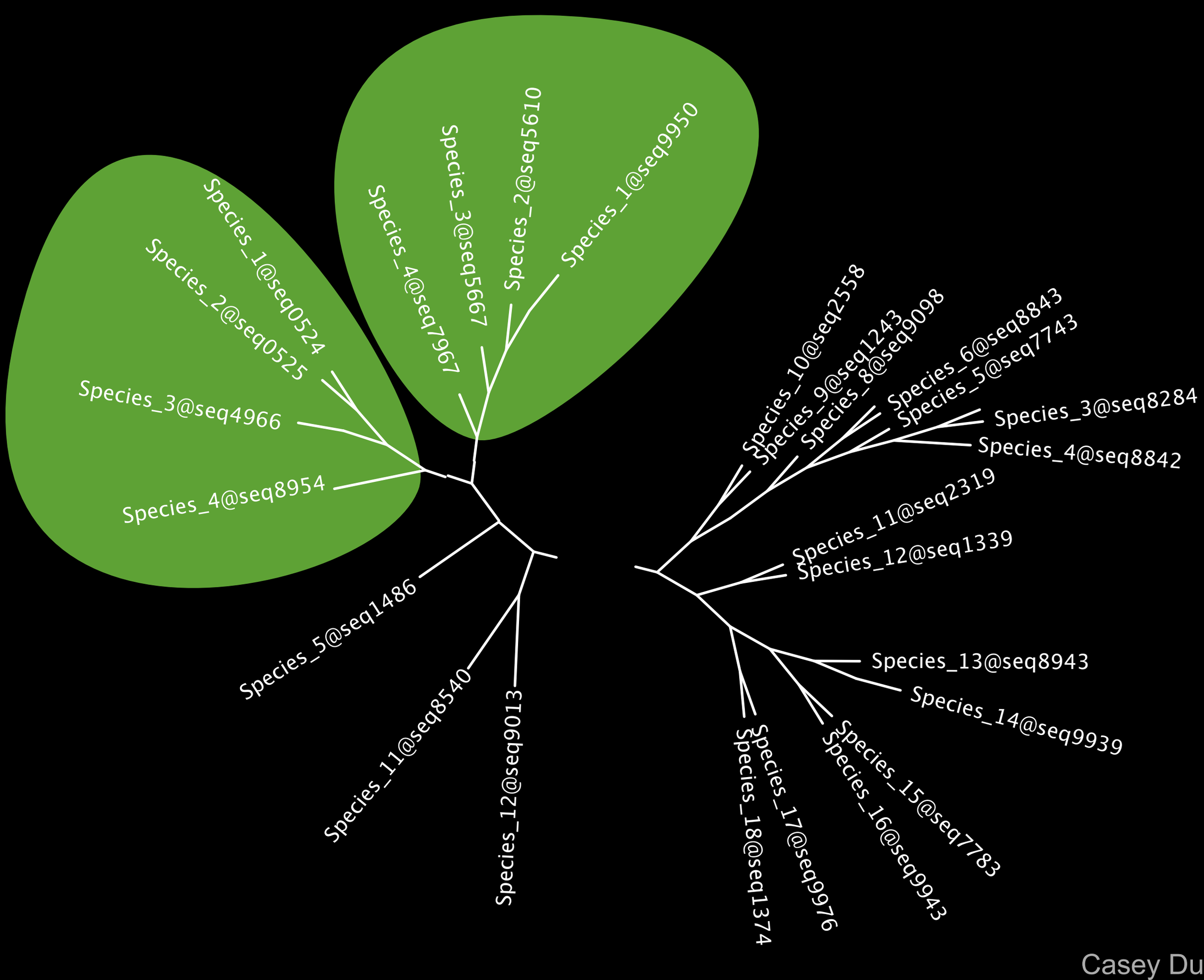
Species A

Species B

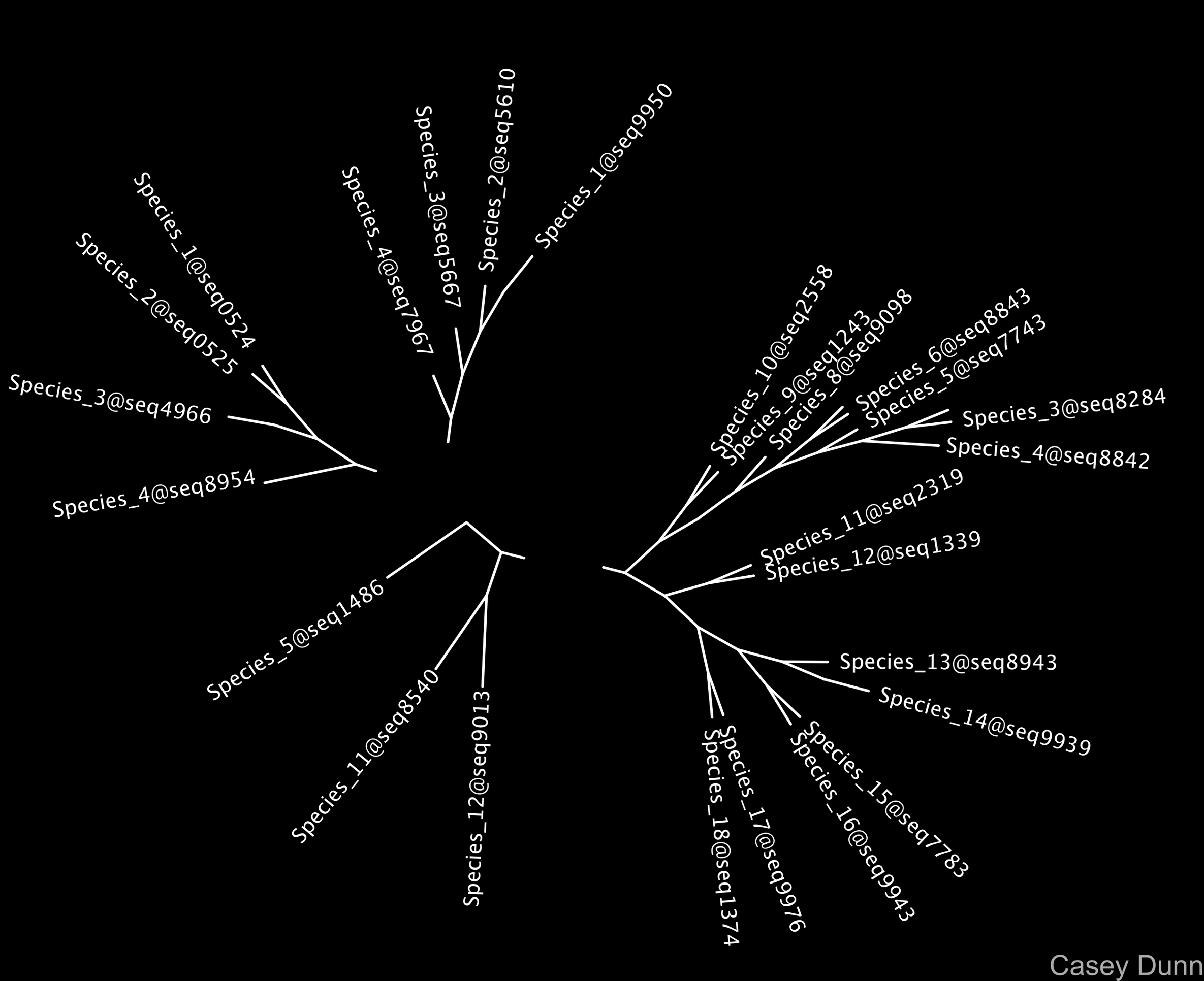
Species C

Gene divergence
due to duplication


Species_1@seq0524
Species_2@seq0525
Species_3@seq4966
Species_4@seq8954
Species_4@seq7967
Species_3@seq5667
Species_2@seq5610
Species_1@seq9950
Species_10@seq2558
Species_9@seq1243
Species_8@seq9098
Species_6@seq8843
Species_5@seq7743
Species_3@seq8284
Species_4@seq8842
Species_11@seq2319
Species_12@seq1339
Species_5@seq1486
Species_11@seq8540
Species_12@seq9013
Species_13@seq8943
Species_14@seq9939
Species_15@seq7783
Species_16@seq9943
Species_17@seq9976
Species_18@seq1374
Casey Dunn


Species_1@seq0524
Species_2@seq0525
Species_3@seq4966
Species_4@seq8954
Species_4@seq7967
Species_3@seq5667
Species_2@seq5610
Species_1@seq9950
Species_10@seq2558
Species_9@seq1243
Species_8@seq9098
Species_6@seq8843
Species_5@seq7743
Species_3@seq8284
Species_4@seq8842
Species_11@seq2319
Species_12@seq1339
Species_5@seq1486
Species_11@seq8540
Species_12@seq9013
Species_13@seq8943
Species_14@seq9939
Species_18@seq1374
Species_17@seq9976
Species_16@seq9943
Species_15@seq7783

Casey Dunn


Species_1@seq0524
Species_2@seq0525
Species_3@seq4966
Species_4@seq8954
Species_4@seq7967
Species_3@seq5667
Species_2@seq5610
Species_1@seq9950
Species_10@seq2558
Species_9@seq1243
Species_8@seq9098
Species_6@seq8843
Species_5@seq7743
Species_3@seq8284
Species_4@seq8842
Species_11@seq2319
Species_12@seq1339
Species_5@seq1486
Species_11@seq8540
Species_12@seq9013
Species_13@seq8943
Species_14@seq9939
Species_18@seq1374
Species_17@seq9976
Species_16@seq9943
Species_15@seq7783

Isolation of Homologs

Isolation of Orthologs

Evaluation of Orthology

Phenetic

Phylogenetic

Phenetic

Phylogenetic

Casey Dunn

Once we have subtrees of orthologs...

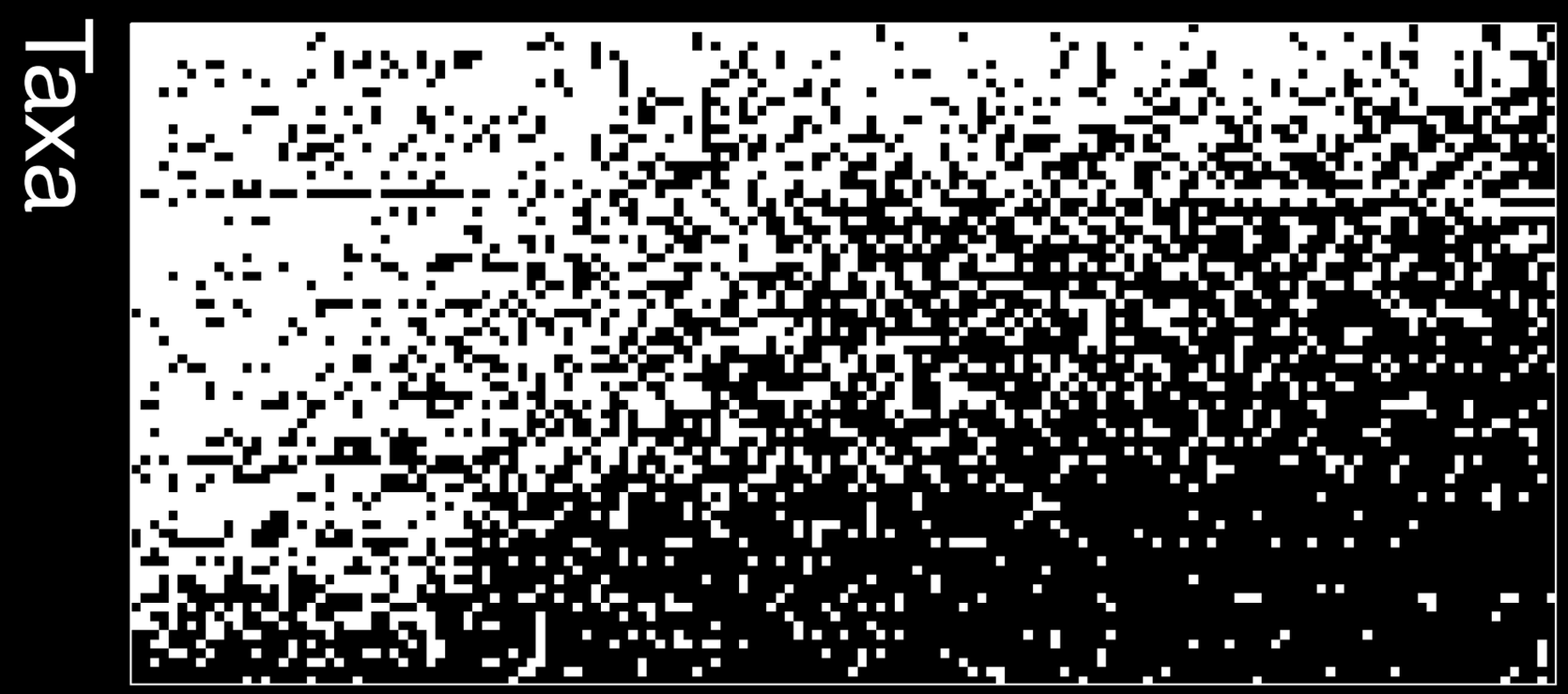
Align each ortholog

Concatenate!

There are many exciting alternatives to concatenation

As these become more computationally efficient, robust to missing data, etc they will be exciting to apply to these datasets

77 taxa, 150 Genes, >20k aa



White cells indicates sampled gene

50.9% gene sampling

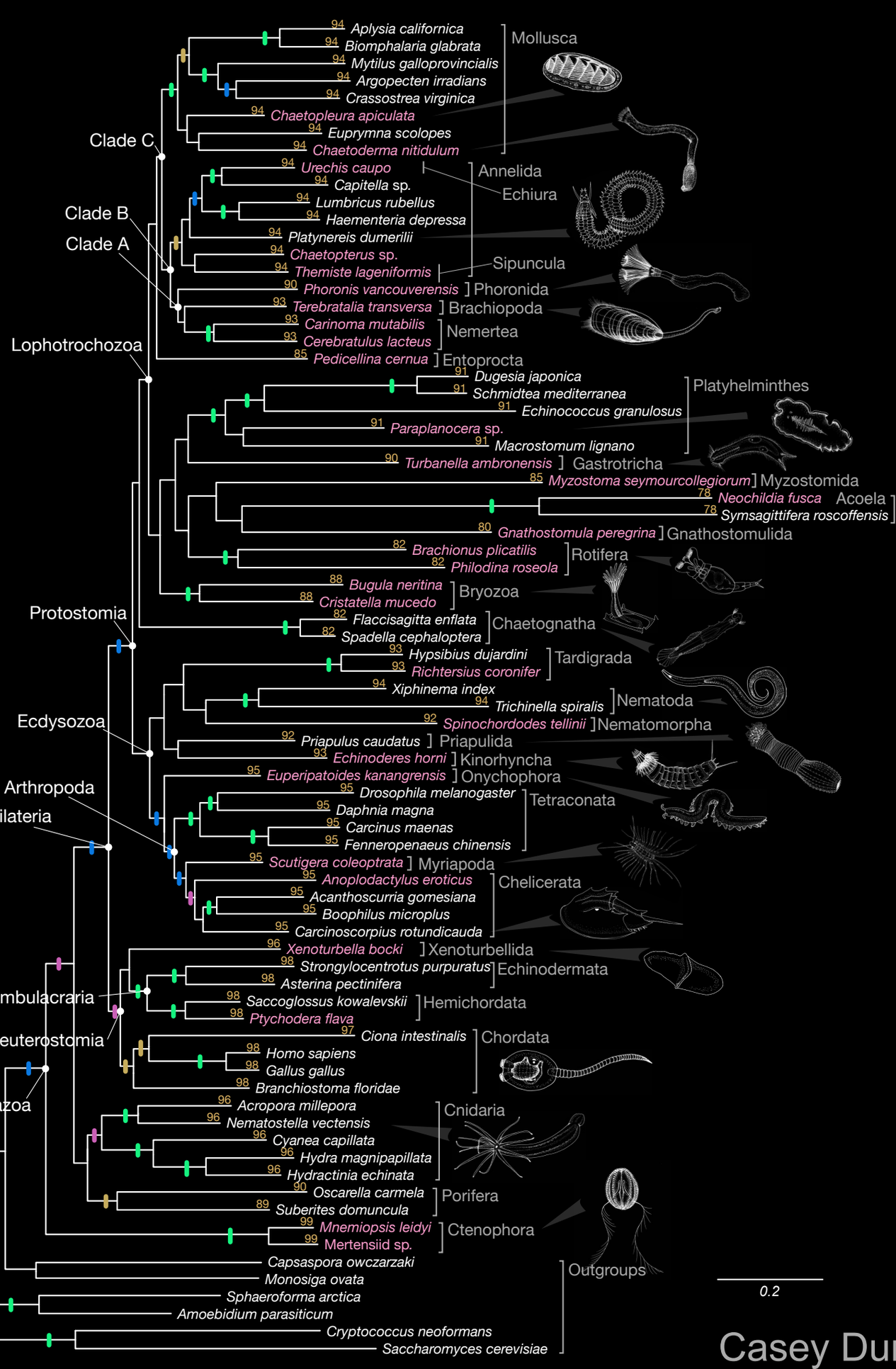
Dunn *et al.*, 2008
doi:10.1038/nature06614

Dunn *et al.*, 2008
doi:10.1038/
nature06614

150 Genes, >20k aa

Bootstrap support
- >98%
- >90%
- >80%
- >70%

raxML
1,000 BS replicates
WAG+Γ



Casey Dunn

Extracting more information from analyses


Casey Dunn

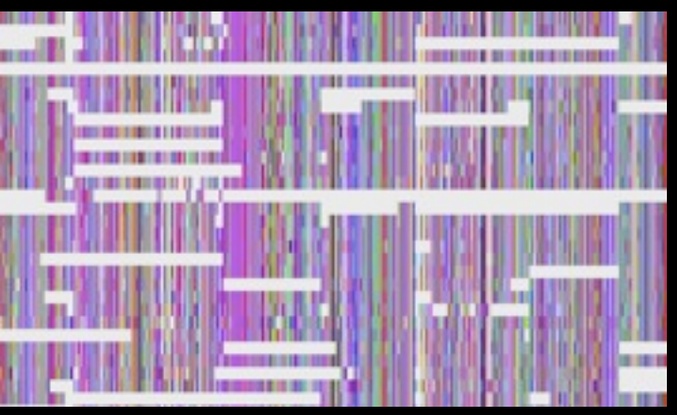
Mess with your dataset and analysis settings

Rerun your analyses

But this doesn’t work well for large analyses

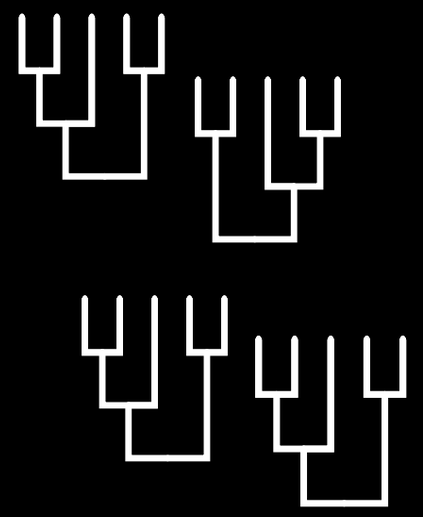
Information attrition

Matrix



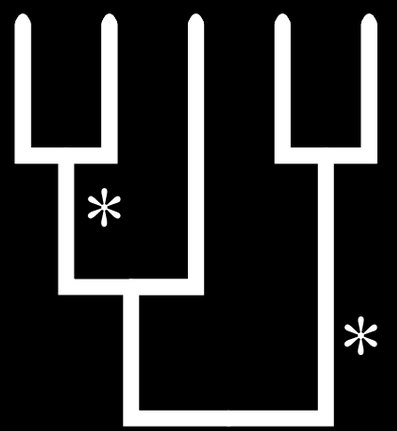
Inference

Tree set



Summary

"Final Product"



We throw away:

Explicit information on distribution of support and conflict

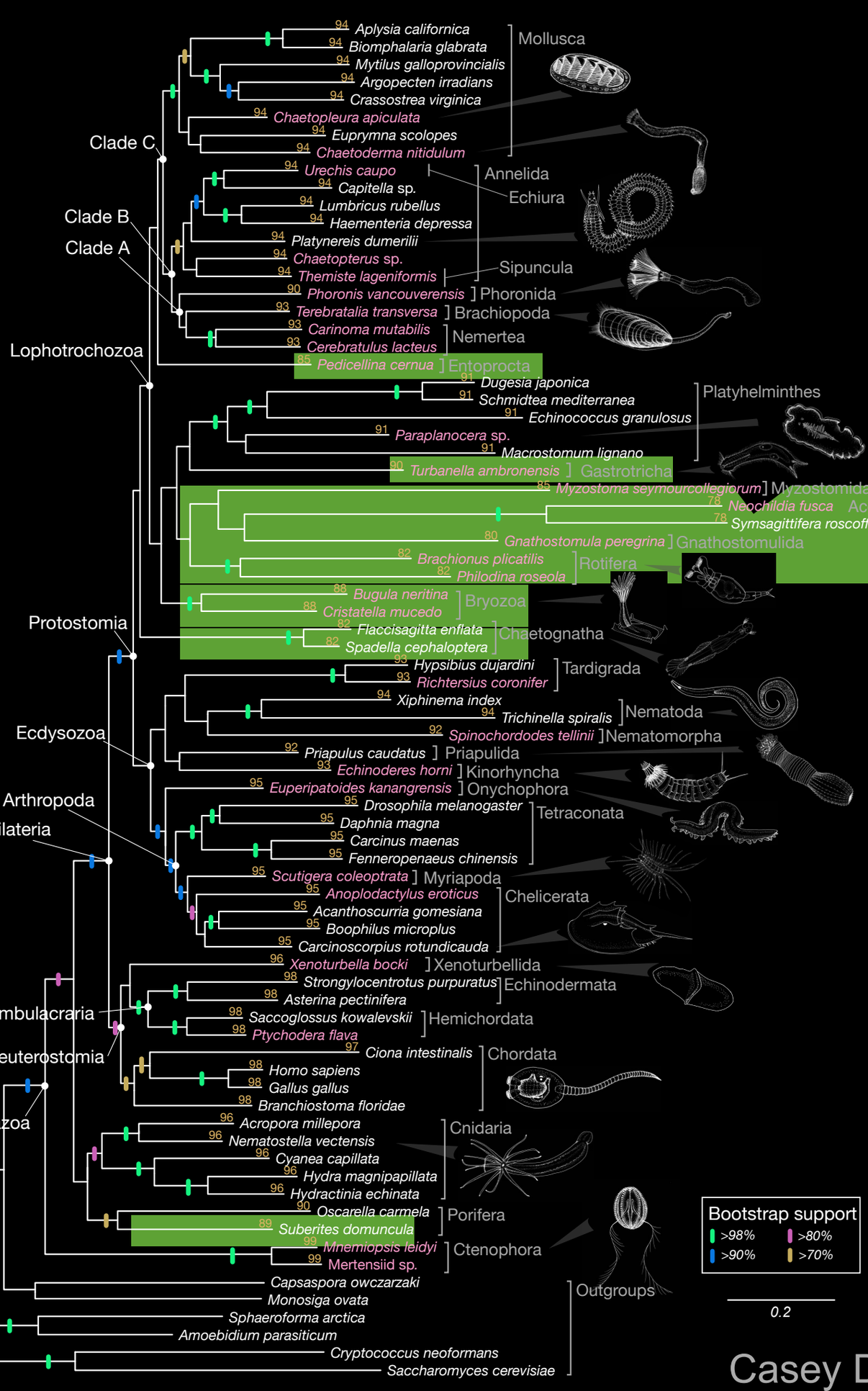
Non-consensus topological variation, information on variation in taxon placement

Casey Dunn

Unstable taxa can obscure support for relationships between stable taxa.

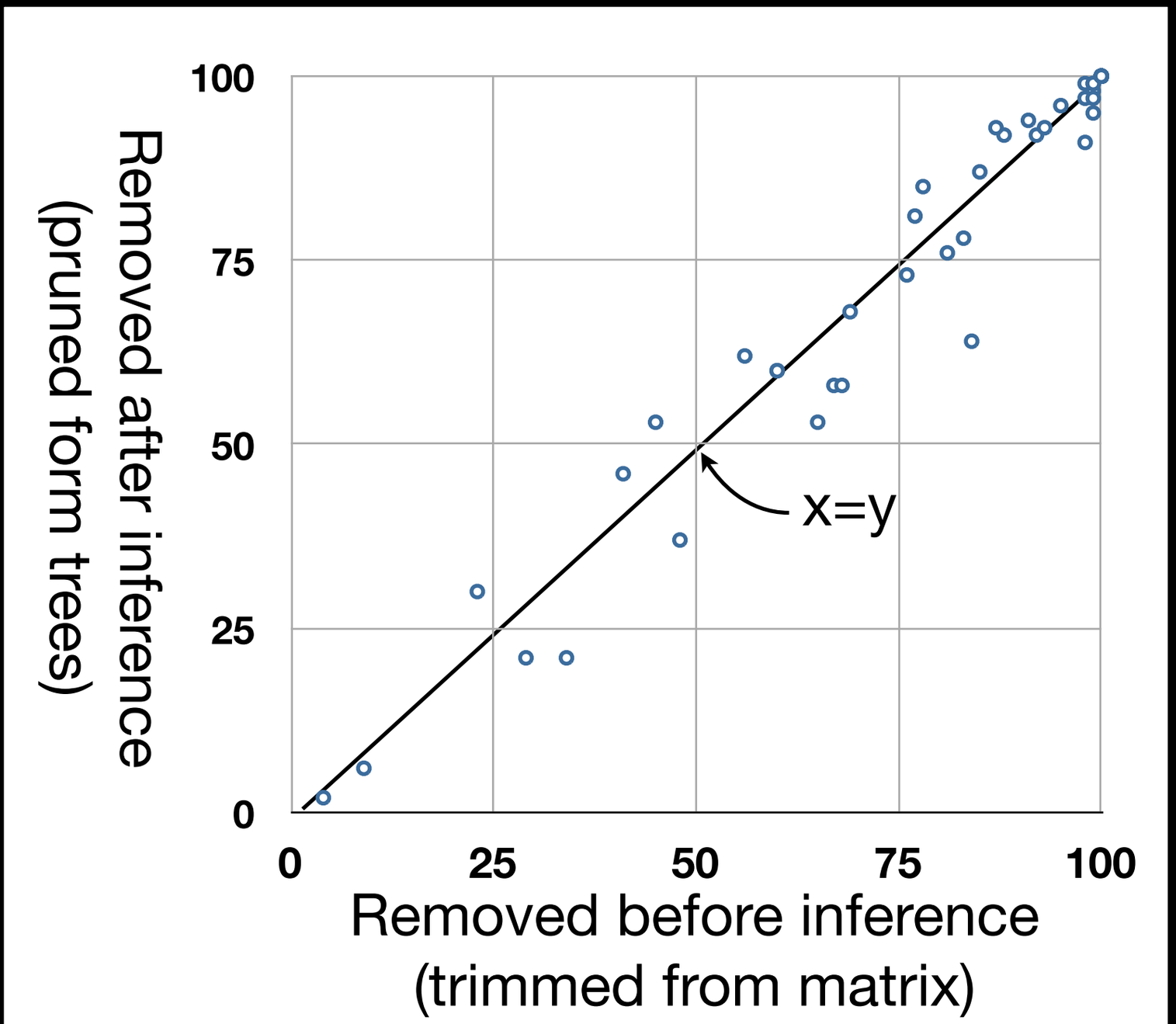
Leaf stability indeces (Thorley & Wilkinson, 1999) quantify the stability of each taxon.

Leaf Stability < 90%



Casey Dunn

Split frequencies following removal of unstable taxa



Unstable taxa have little or no influence on relationships between stable taxa

1,000 BS replicates

Treesets as onions

Least stable taxa

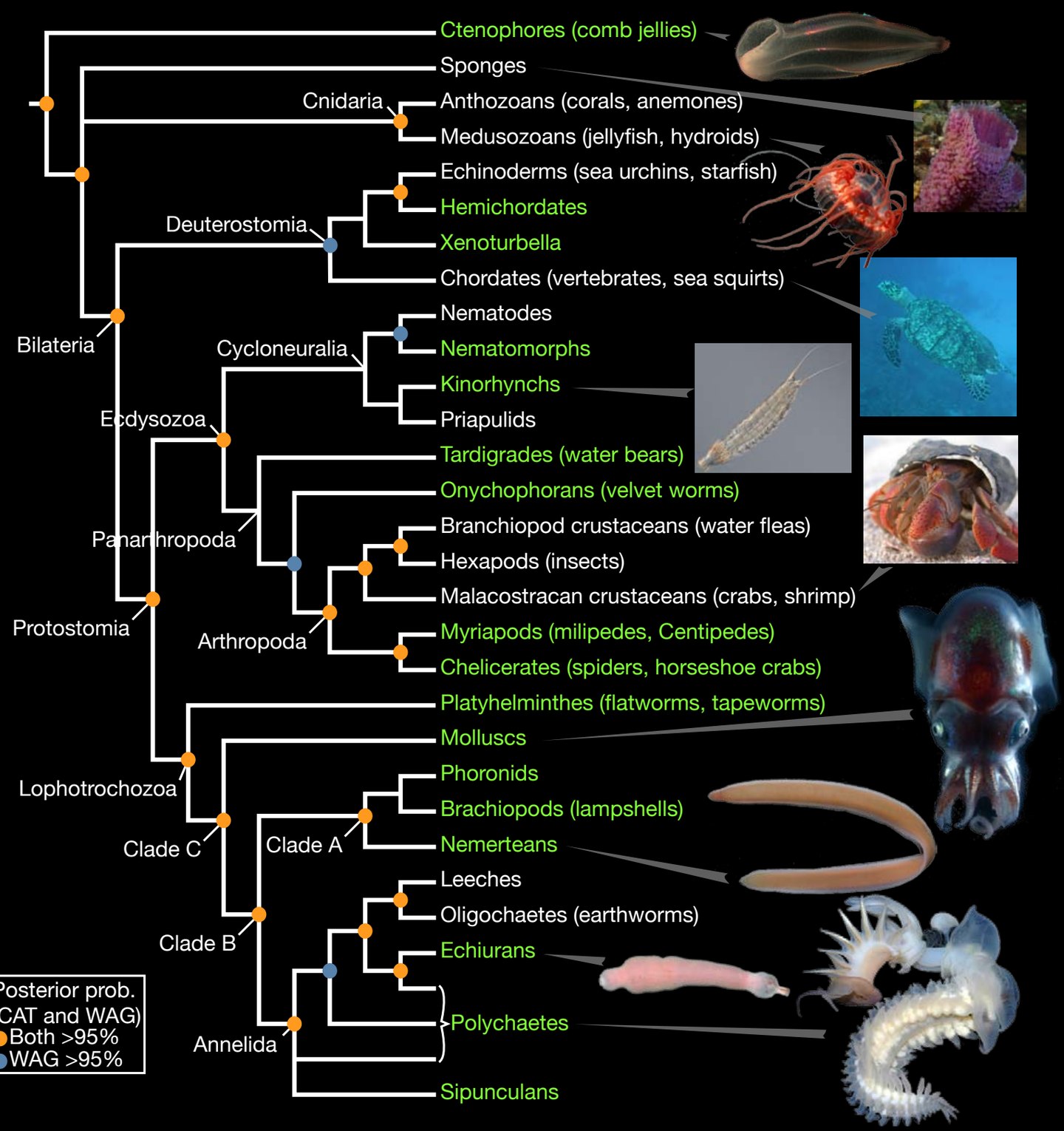
Most stable taxa

Casey Dunn

Investigation of stable taxa

Dunn *et al.*, 2008
doi:10.1038/
nature06614

150 Genes
>20k aa



Casey Dunn

Tools for identifying and visualizing relationships between stable taxa

phyutility

code.google.com/p/phyutility/

Smith & Dunn (2008), Bioinformatics

doi:10.1093/bioinformatics/btm619

- Calculate leaf stabilities
- Prune taxa from treesets
- Explore the positions of unstable taxa

Casey Dunn

Beyond species trees

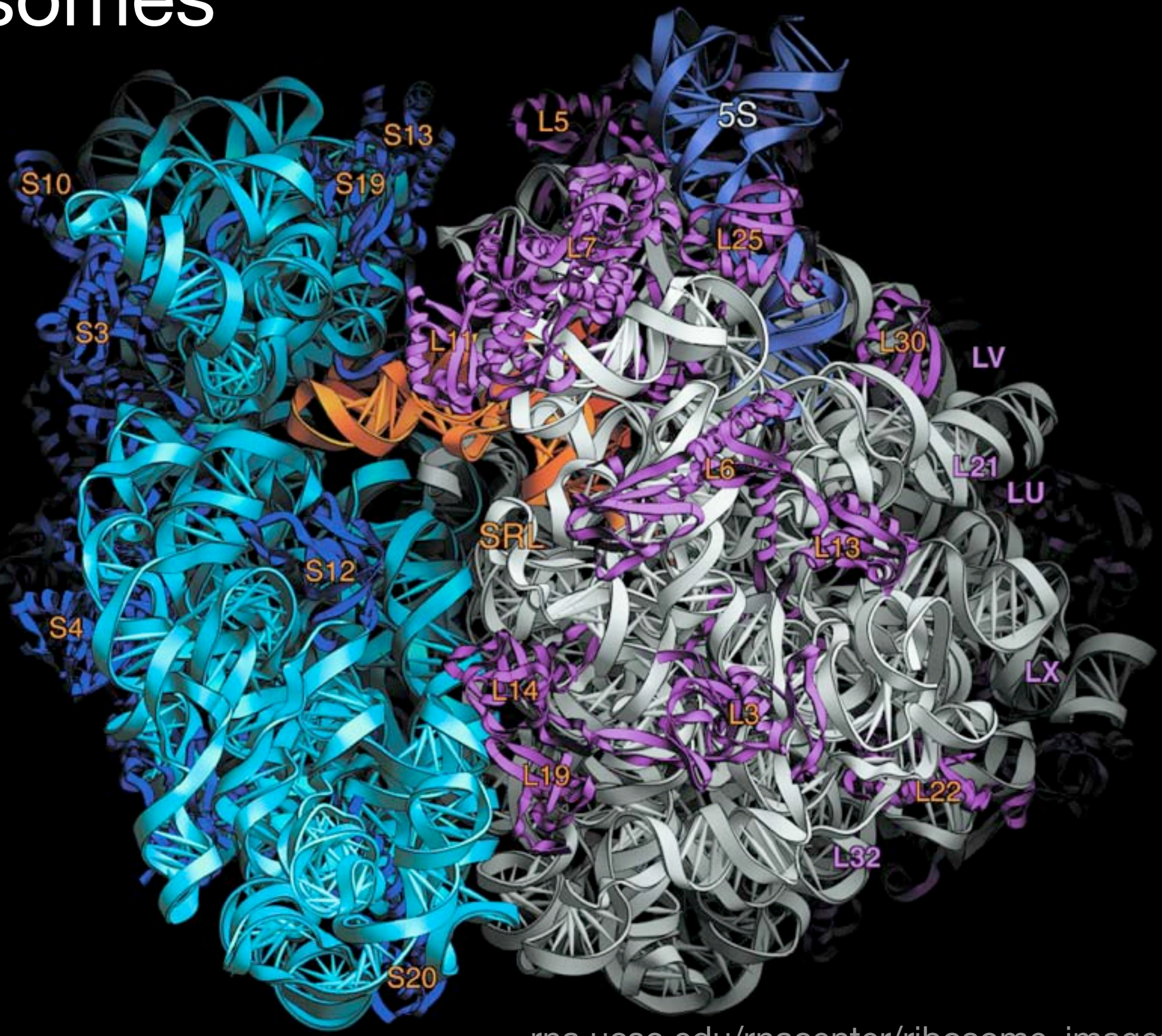
Phylogenies now generate:

- Species trees

- Extensive gene sequence data

- Well sampled gene trees

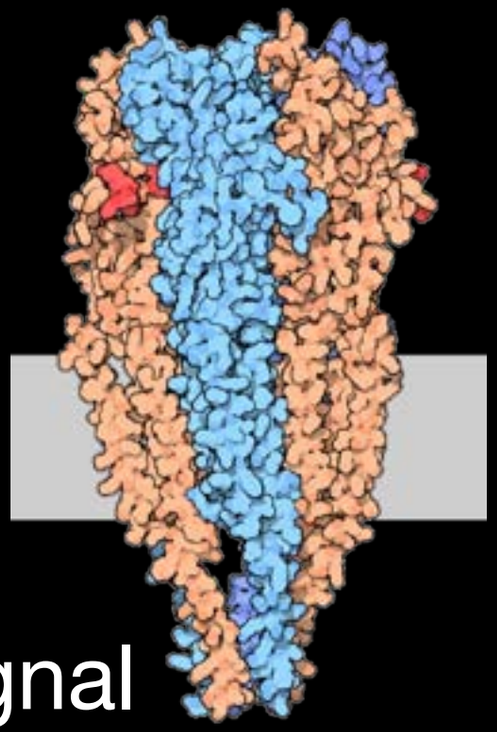
Ribosomes



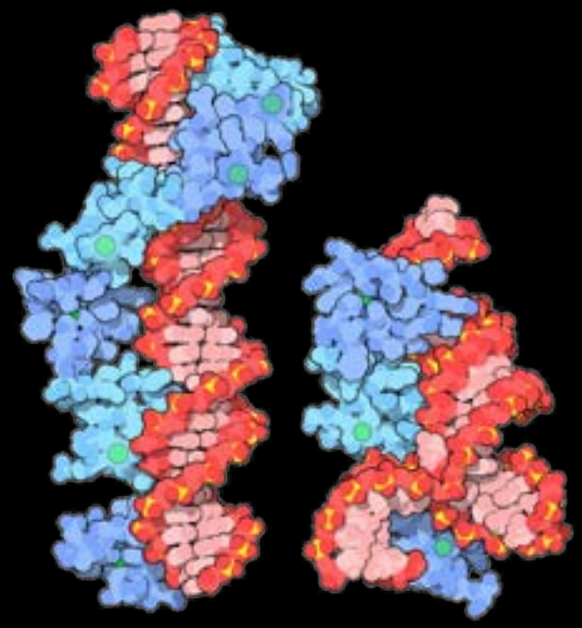
rna.ucsc.edu/rnacenter/ribosome_images.html

Casey Dunn

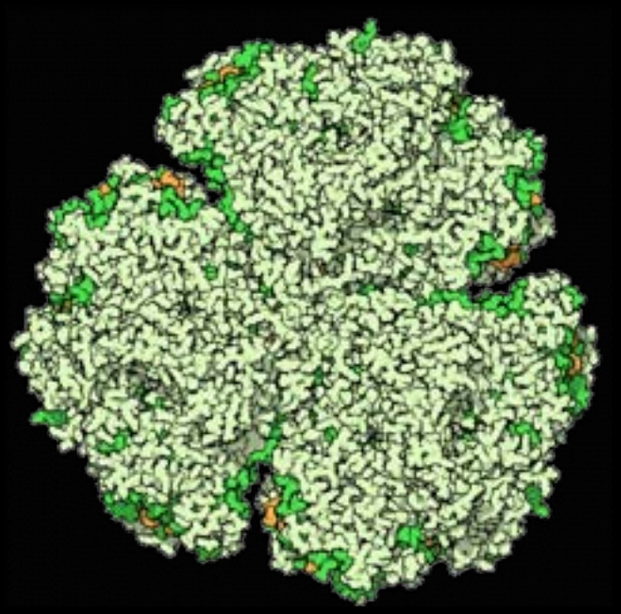
Sequences relevant to focal phenotypes



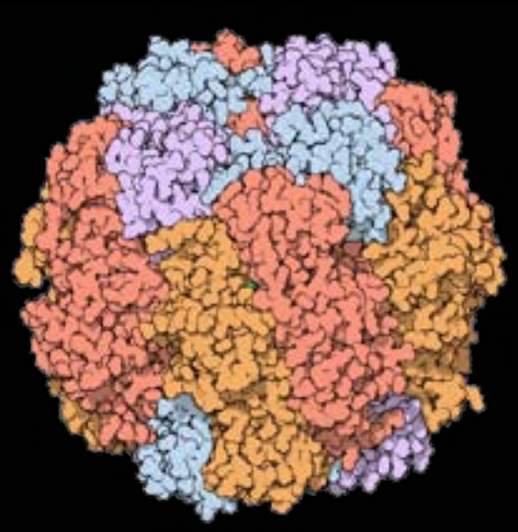
signal transduction



morphogenesis



photosynthesis



carbon sequestration

www.rcsb.org/

Casey Dunn

But we don’t know which genes are relevant to which phenotypes

A small glimpse of a much greater schism

Genomes

NCBI

Trace Archive

Sequence Read Archive

```
>FZTBY7Y04I0Z6U rank=0418088 x=3584.5 y=3492.0 length=457
CAAGGTCTTGAACCAACAGTTGGATACAATTCAGAATGACGGAATGAAAGAAATCCATTT
CACTGGTCATGTTGACTTTTTCCACGTGTTCTGGATCGATCTTCTCCTTTAACTTCTCCT
GAAGCTGTTGTGTTGTCTCCTGGCAGTATTCGCATGTCGTCAGTACACAACATATCAACC
TCCGTTCGTCTTCGTTGAACTTCACGTCGTTGTCCTTCAAGATGCTCGAAATACCGCCCG
TCGAAGAAACCTTGGGTAAGGCTTCCCAGCAGAAGGCGATTAGTGTATTCCACCAGGTAT
CTACTTGAACGTATCTGTAGAAAGGGAAGAAGCAGAGGTTGCCACTGCTCAACTGAACGC
ATTGCACCATACTCTTCTTGAAGAAGACAAACAACTCGCCAGCACTCGGGAGAACTGTCC
CTTCGGAATCCTCGCCGAGTTTTGGAACACCTTGTTC
>FZTBY7Y04IQ5F0 rank=0418094 x=3472.5 y=2494.5 length=288
AATGAAATATGCTGAGCAGTTCAAGTTTCTATACTCACGAAGAAACAACATTGTAGATGG
TTCATACGAACCCAACAATGAAGAGGCGGTTTGGGTTGATCCTTTAGAAGAATTGGTTGA
ACAGTTGAATAAGGGTGGTGAAGAAAAGCTGAATCTGAGAAAACTGAAGAAGAGAAATTG
GCTGGATGGTGTGAAAACTTTATCATTTGGTGAAGAACACAAAAGGTATTCCTGAATTTT
GGCTCACTGCAATGAAGAACGTTGAAATACTTGAAGATATGATTCAGG
>FZTBY7Y04IQ7J9 rank=0418096 x=3473.0 y=1143.0 length=421
AGGCCGGGGCCTTTCGATTAAGATATCTAAAAGAGTTTGGTTCTCCACGGAGCTAAGGCT
AACAAATCTACGTAAATCTTGCATTTGTTGCAACCTTCTCTATTAAAAATGTCTGACACA
TCTGTATCCGAACTTGCCTGTGTATACAGTGCCCTTATTTTATACGATGATGATATCGAT
ATCACAGGAGAAAAAAATGGCTAAAATCATCGCTGCTGCCAACGTCAACGTAGAAACCTT
CTGGCCTGGACTCTTCGCCAAGGCTCTCCAAGGACGTAACATCGGTGACCTTATCTGCAA
TGTAGGATCCTCCGCAGCCGCTGCTCCAGCCGCCGCTGCTGCTGCTGGTGATGCTCCAGC
TGCTGCTGAAGAGAAGAAGAAGAAAAGAAGGTCAGTTCAGATGAGGATCAGATGATGATA
```

Evolutionary functional genomics

Morphology, function, ecology, development

GBIF

Casey Dunn

Measuring expression

Casey Dunn

(MBARI)

(S Haddock)

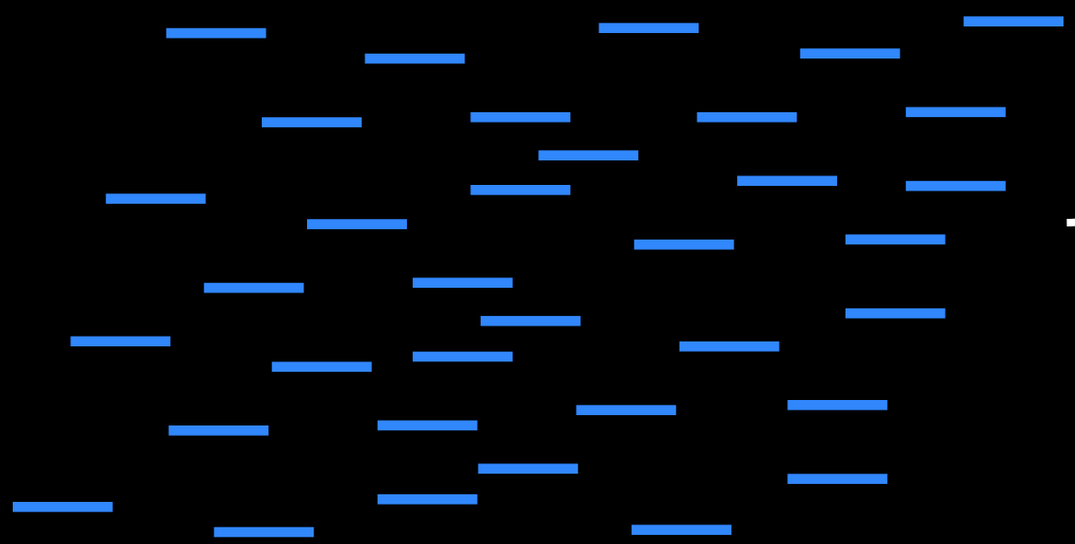
Casey Dunn

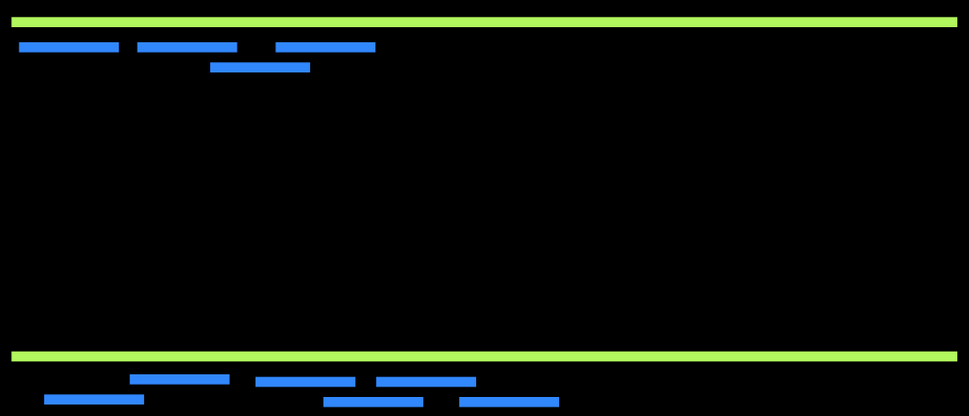
Which genes are differentially expressed between bodies in a siphonophore colony?

Reference

Map

Reads





Gene	Count
Gene001	4
Gene002	6
Gene003	22
Gene004	1
Gene005	2

Casey Dunn

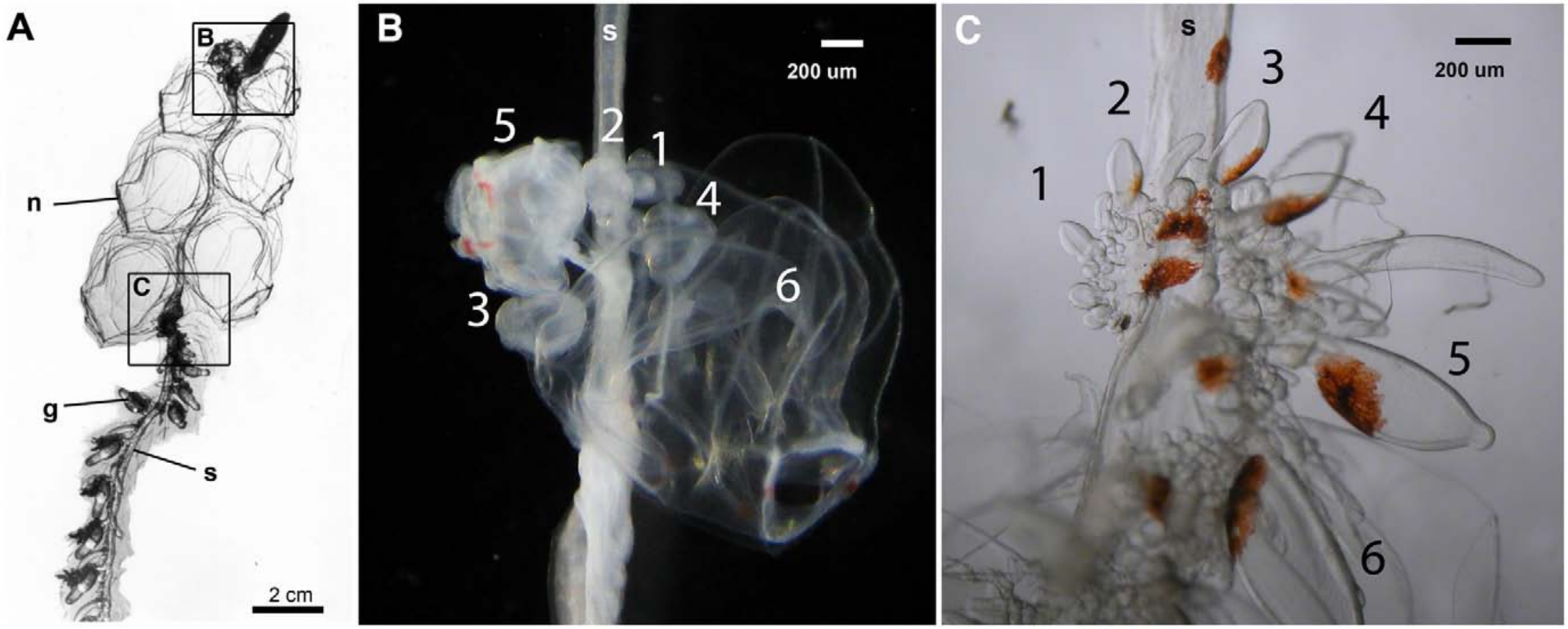
OPEN ACCESS Freely available online

PLoS one

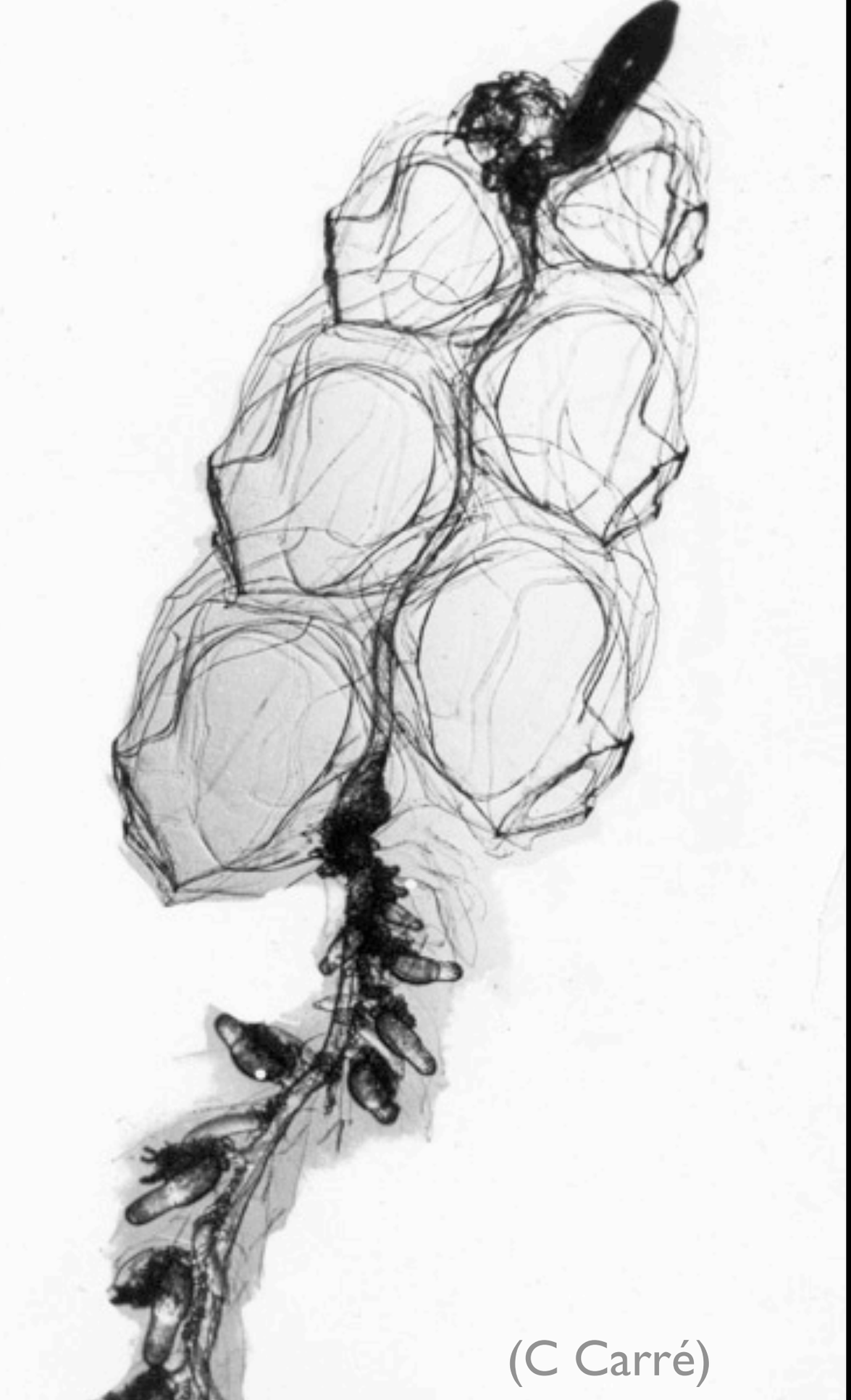
Differential Gene Expression in the Siphonophore *Nanomia bijuga* (Cnidaria) Assessed with Multiple Next-Generation Sequencing Workflows

Stefan Siebert[1]*, Mark D. Robinson[2,3], Sophia C. Tintori[1], Freya Goetz[1], Rebecca R. Helm[1], Stephen A. Smith[1,4], Nathan Shaner[5], Steven H. D. Haddock[5], Casey W. Dunn[1]*

1 Department of Ecology and Evolutionary Biology, Brown University, Providence, Rhode Island, United States of America, **2** Epigenetics Laboratory, Cancer Research Program, Garvan Institute of Medical Research, Sydney, New South Wales, Australia, **3** Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia, **4** Heidelberg Institute for Theoretical Studies, Heidelberg, Germany, **5** Monterey Bay Aquarium Research Institute, Moss Landing, California, United States of America



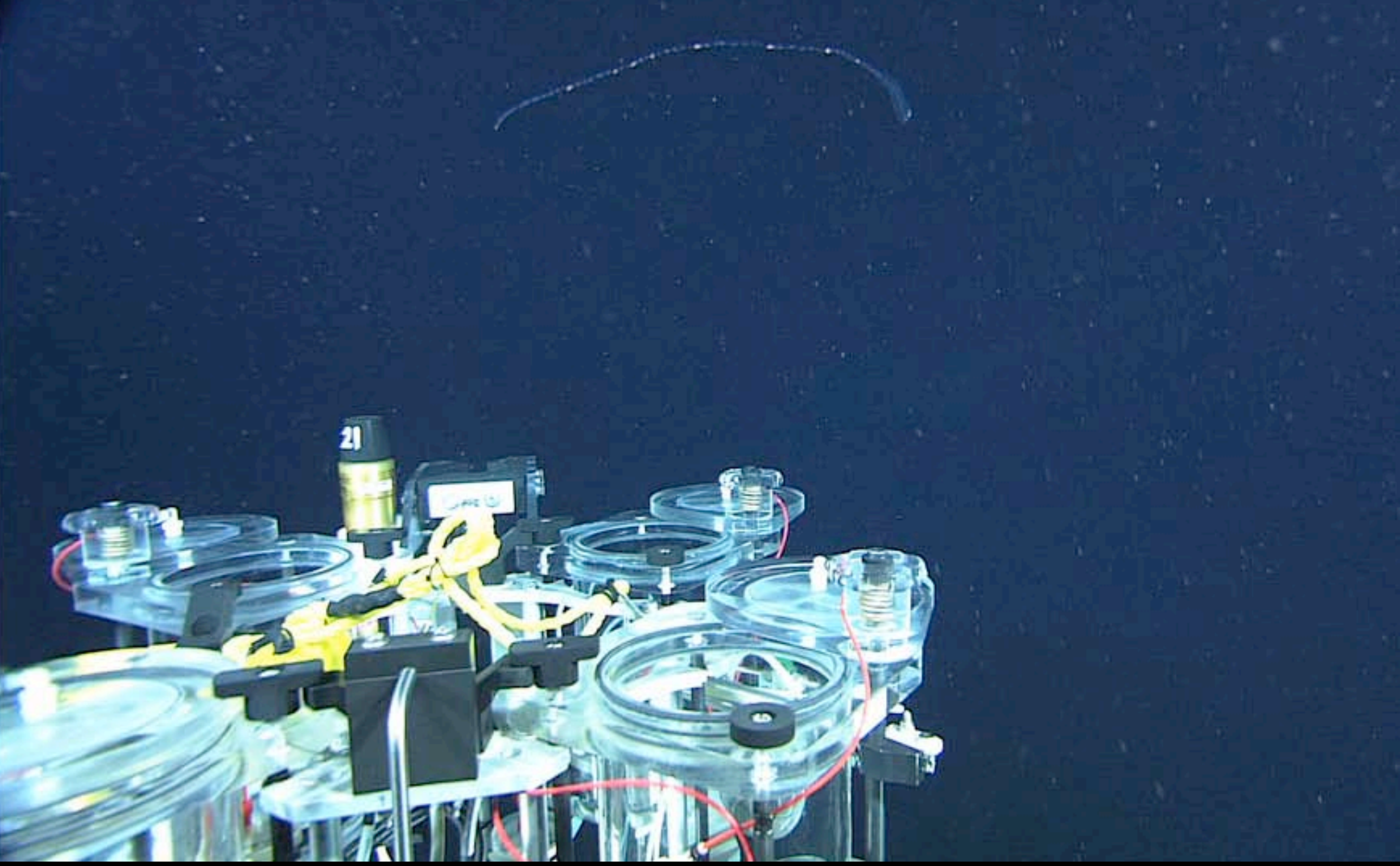
(dx.doi.org/10.1371/journal.pone.0022953)



(C Carré)

Nanomia bijuga

Casey Dunn



(MBARI)

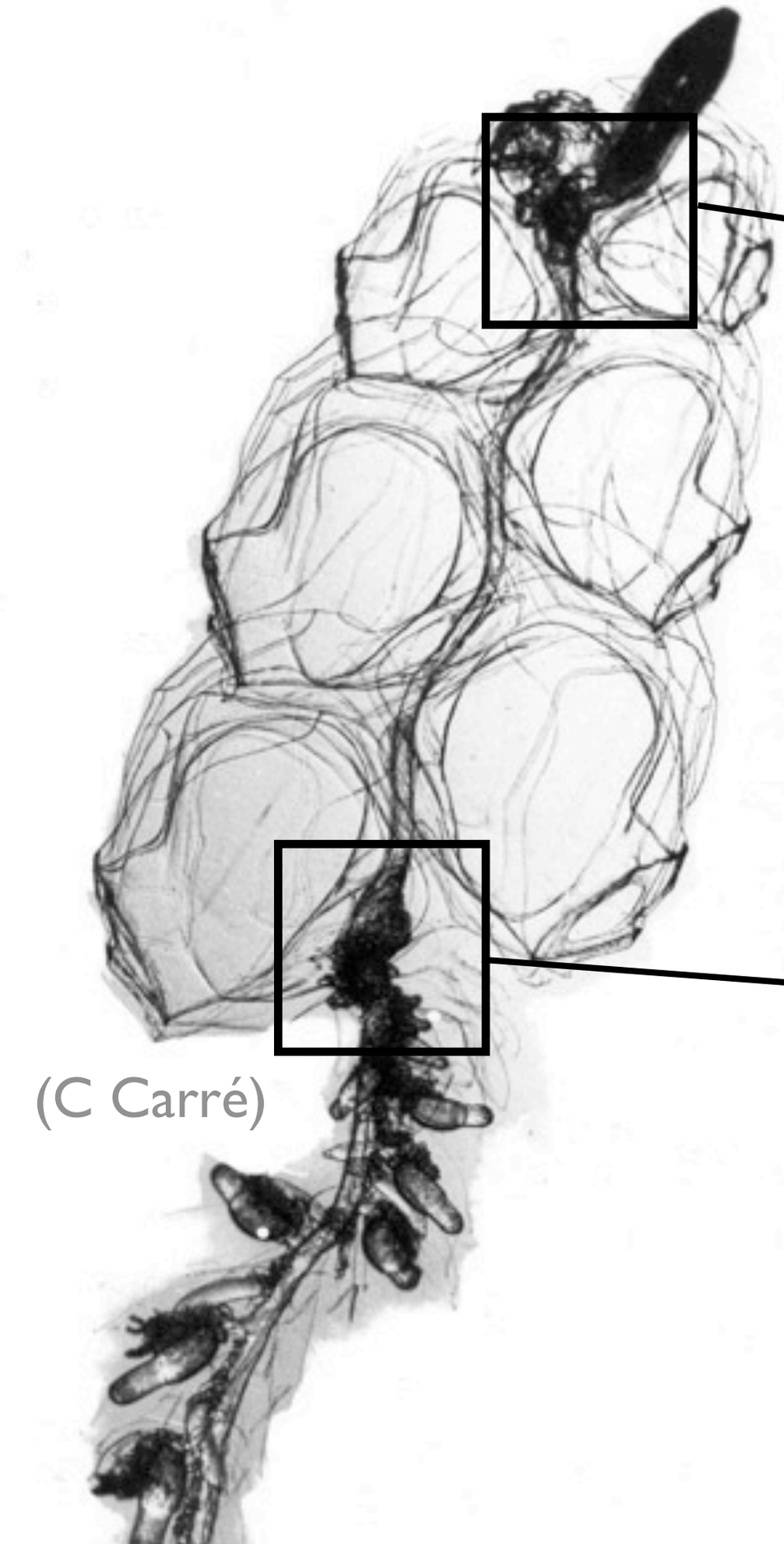
Casey Dunn

Nanomia 454 sequencing

589k reads sequenced (454Titanium)

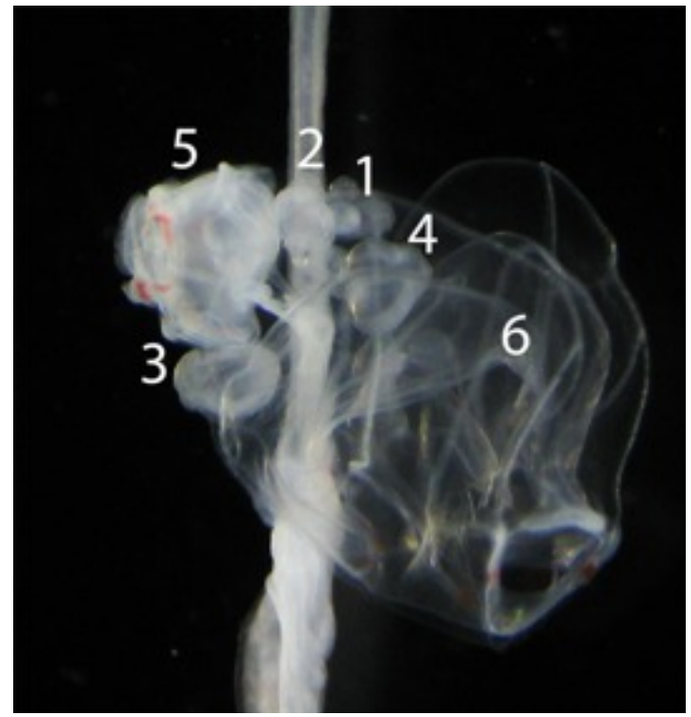
19,925 “genes” in reference (Newbler, cap3)

Paired samples, 3 specimens



(C Carré)

Swimming



Feeding

Casey Dunn

Replicated design

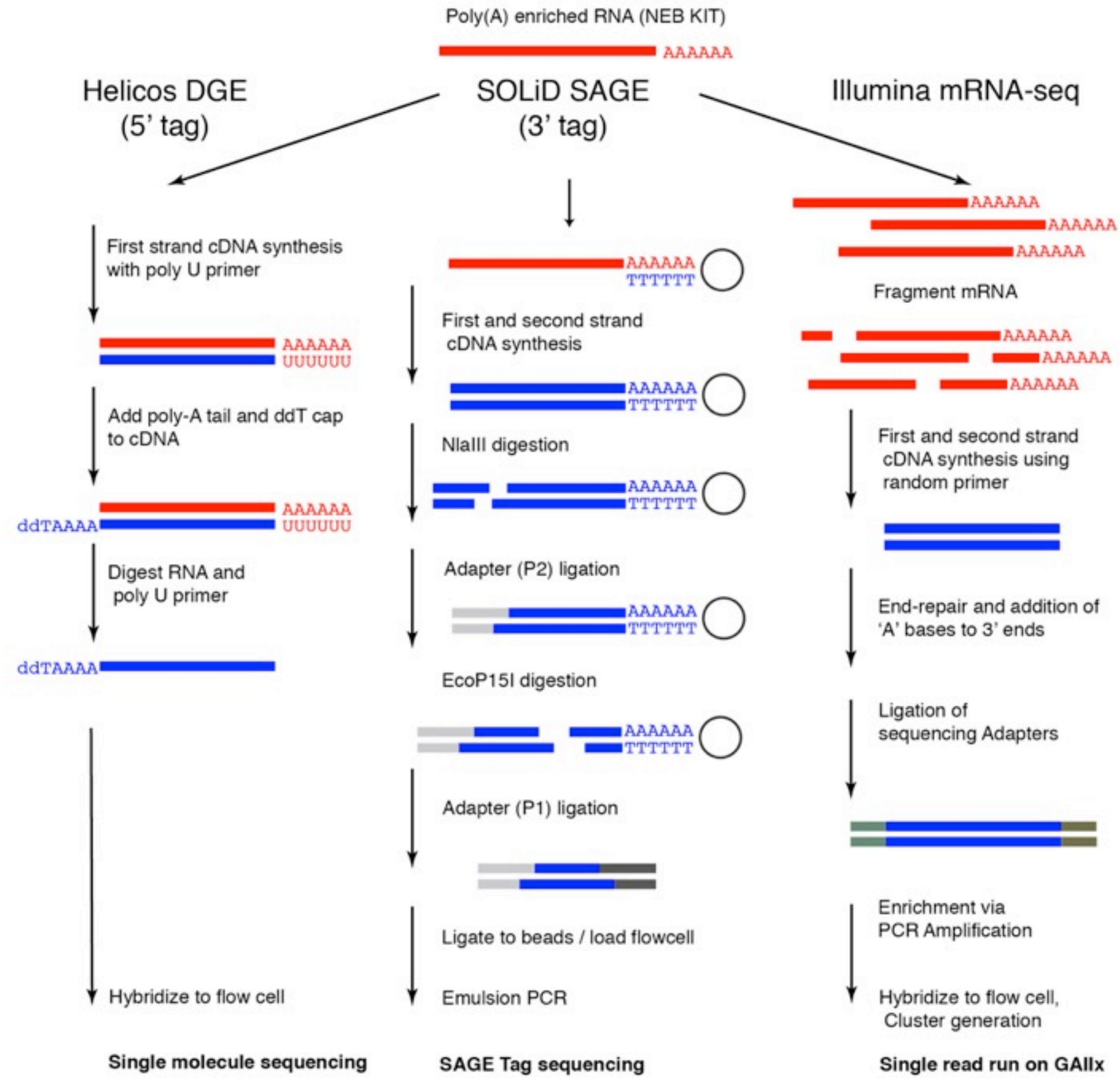
	Tissue A	Tissue B
Specimen 1	X Reads	X Reads
Specimen 2	X Reads	X Reads
Specimen 3	X Reads	X Reads

Casey Dunn

Casey Dunn

Helicos SOLiD Illumina

Casey Dunn



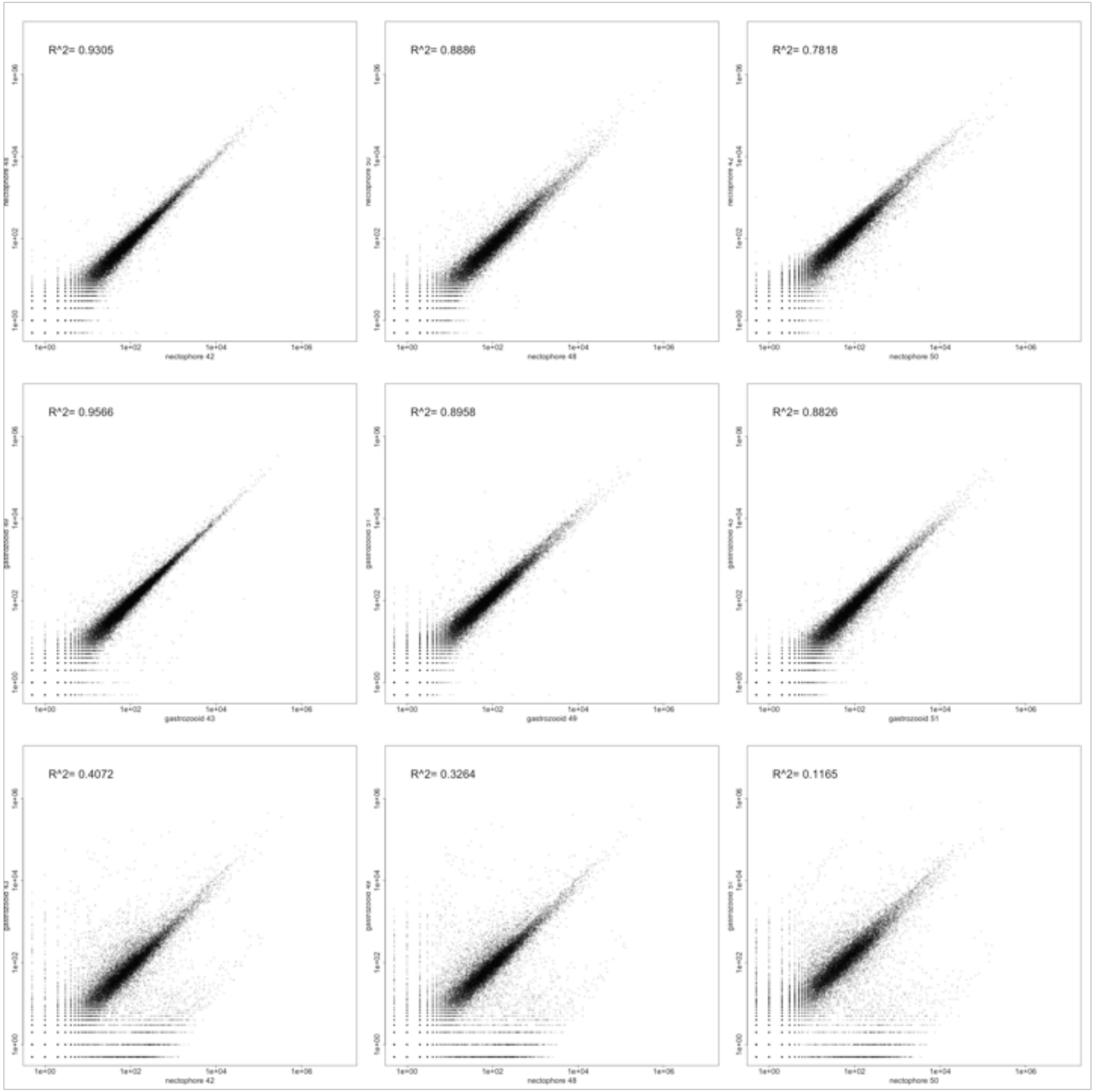
(dx.doi.org/10.1371/journal.pone.0022953)

Illumina

Swimming

Feeding

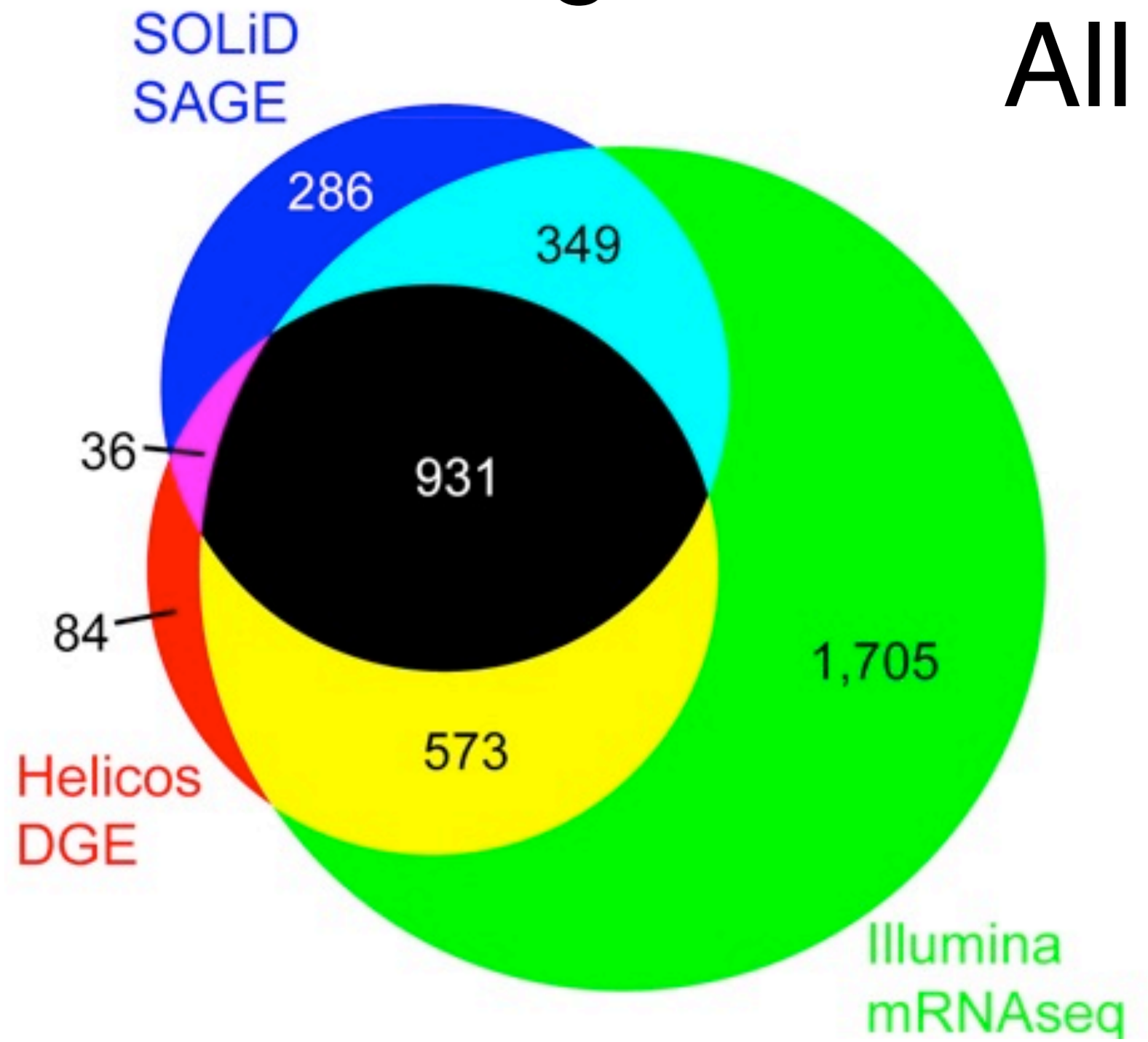
Swimming
v.
Feeding



Casey Dunn

Genes with significant DE

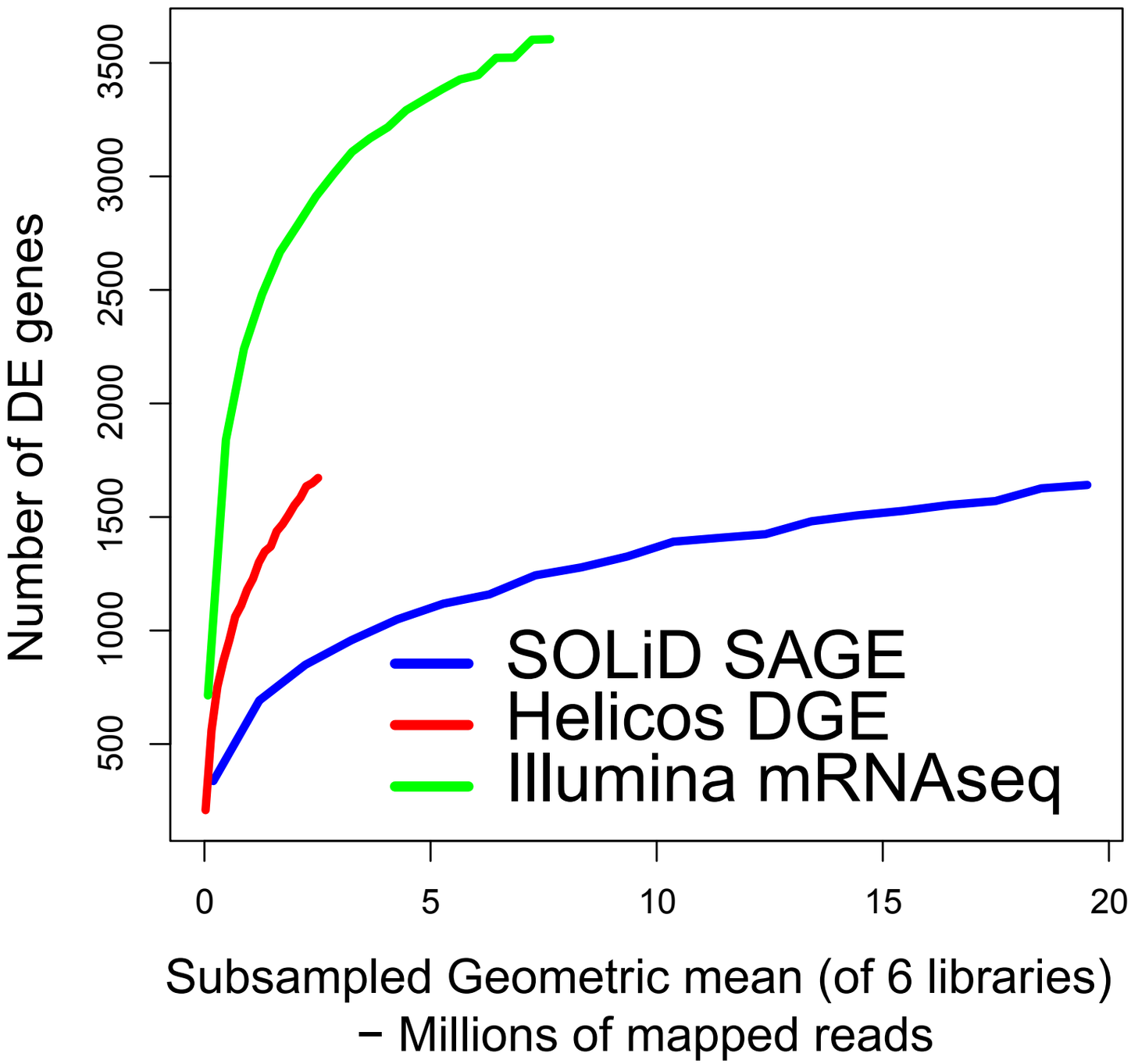
All genes



EdgeR, Bonferroni corrected $p < 0.05$

(dx.doi.org/10.1371/journal.pone.0022953)

Are these differences due to differences in read numbers across workflows?

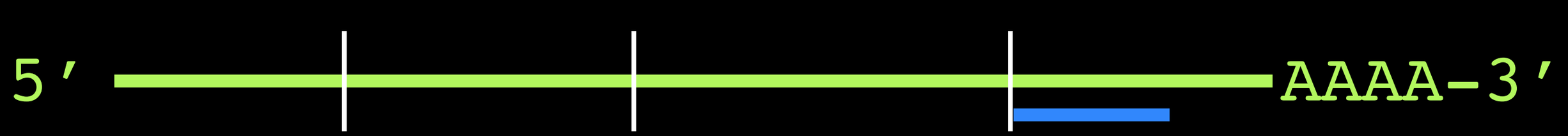


No.

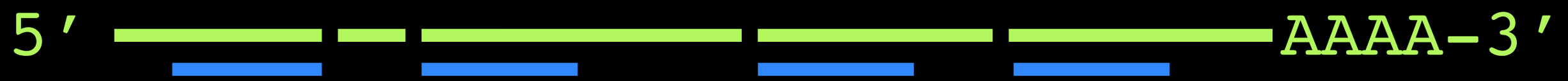
Helicos (Tag-based)

SOLiD (Tag-based)

NlaIII NlaIII NlaIII



Illumina (RNAseq)



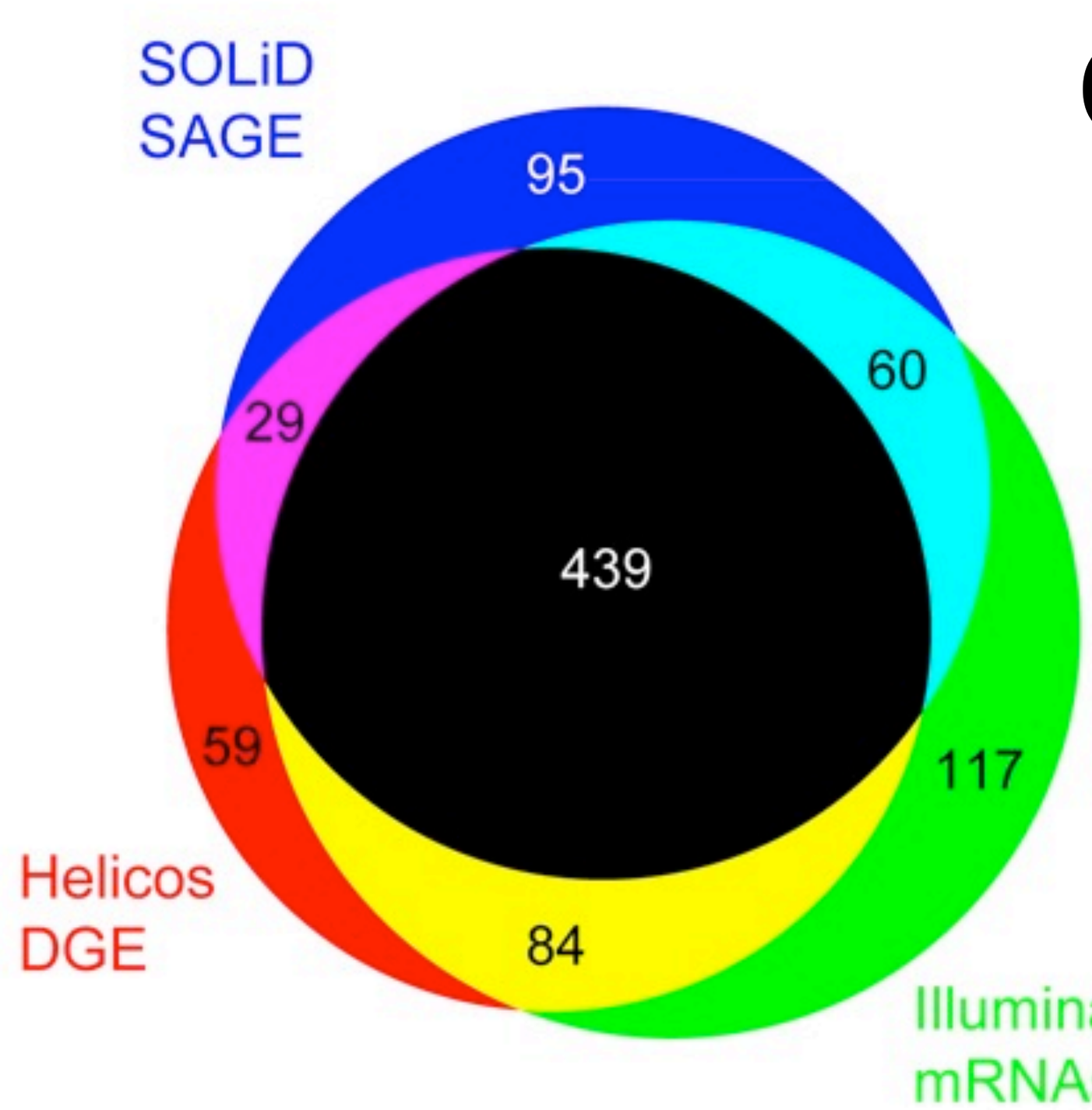
Original 454-derived reference:

19,925 genes

Reference sequences that unambiguously have tag sites:

4,255 genes

Genes with significant DE

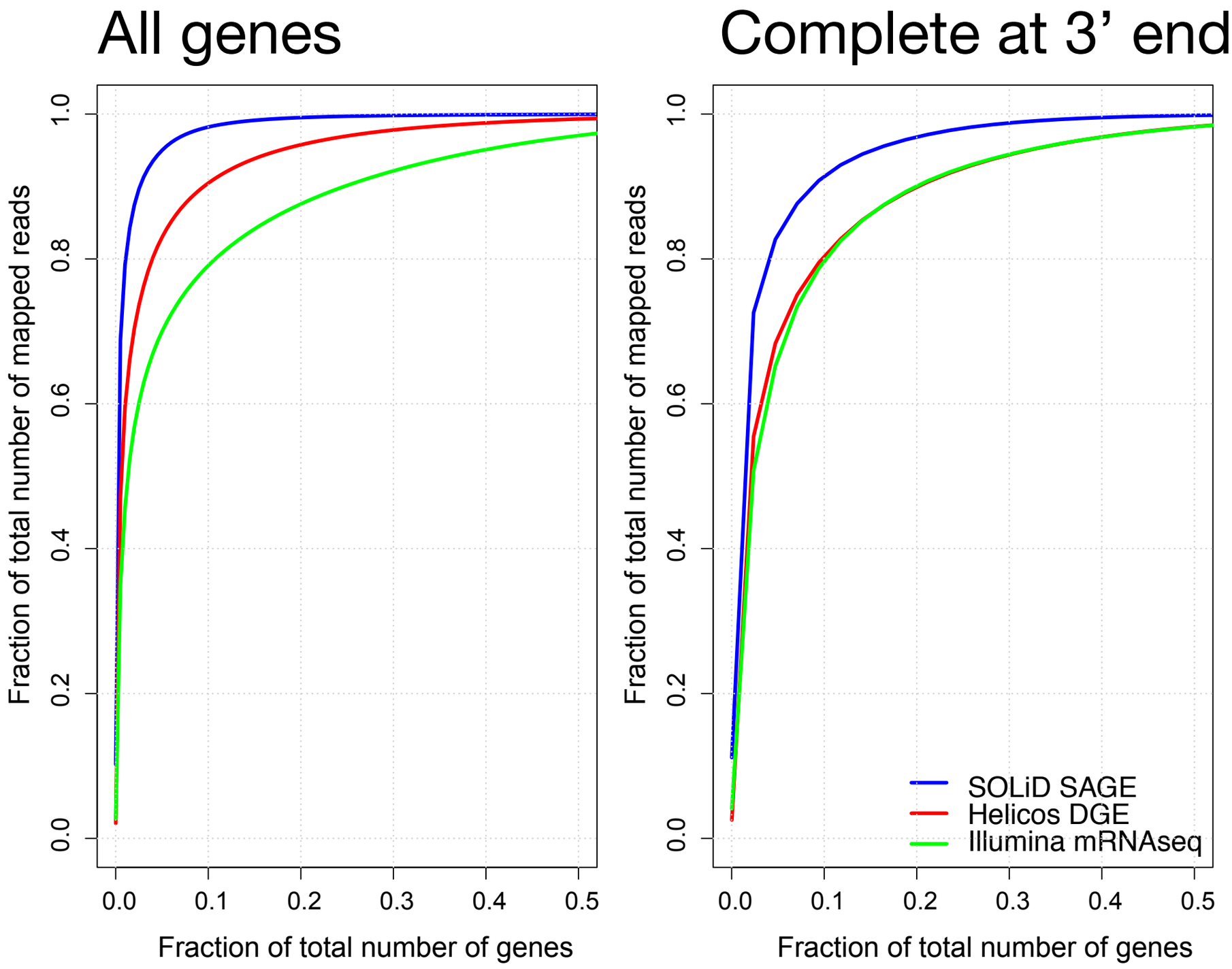


Genes with complete 3’ end

EdgeR, Bonferroni corrected $p < 0.05$

(dx.doi.org/10.1371/journal.pone.0022953)

Distribution of sequencing effort



(dx.doi.org/10.1371/journal.pone.0022953)

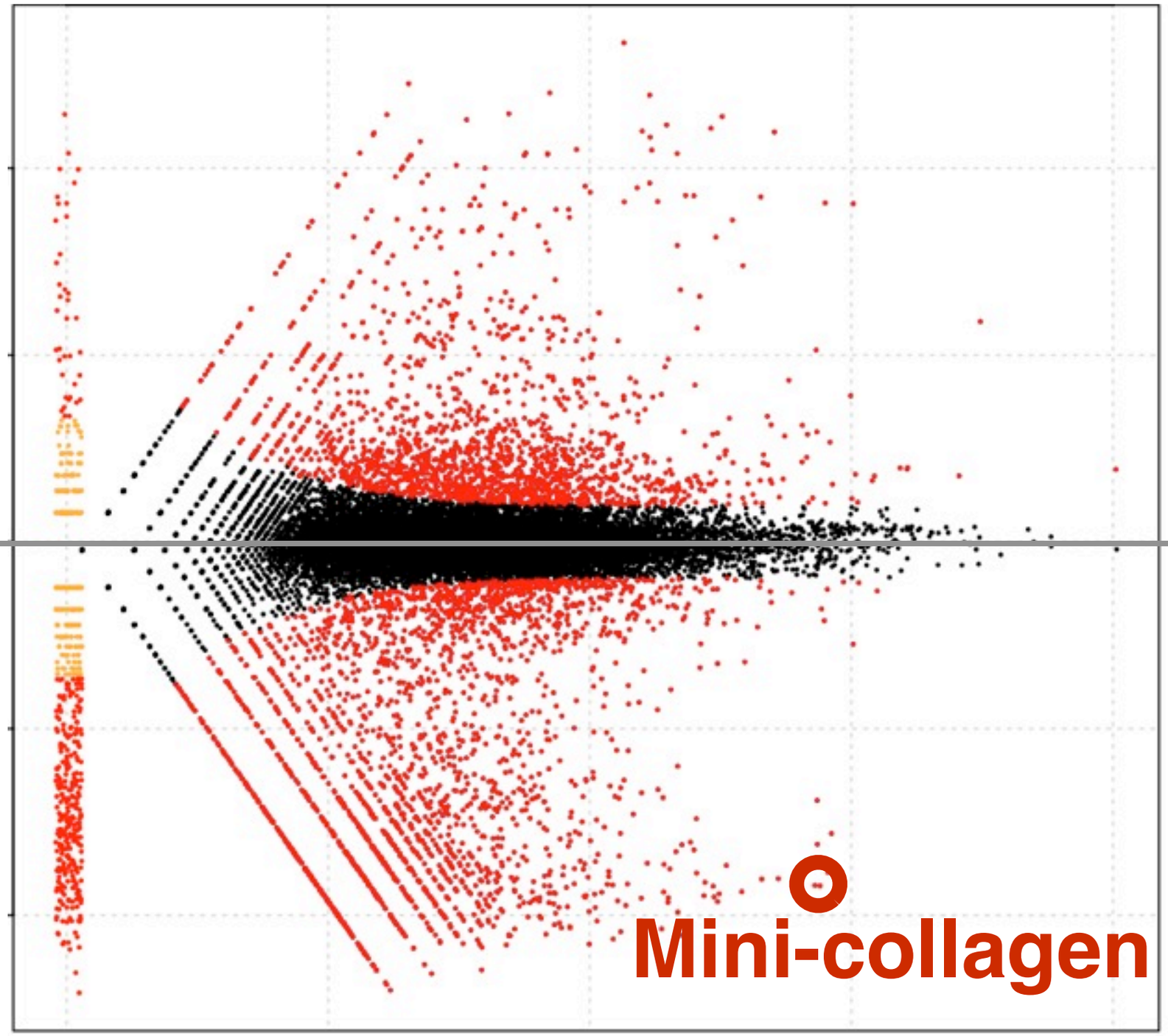
Where to next?

- Characterization of genes with significant differential expression

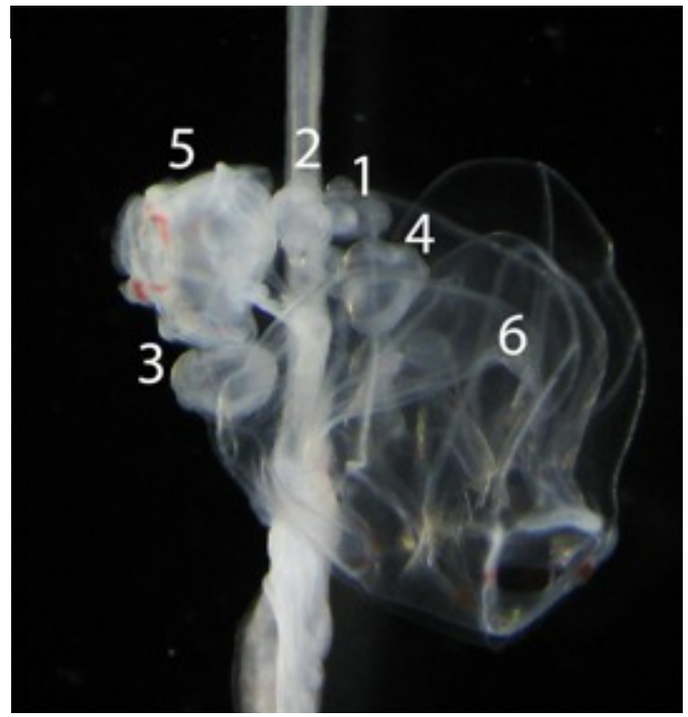
Red genes have significant differential expression

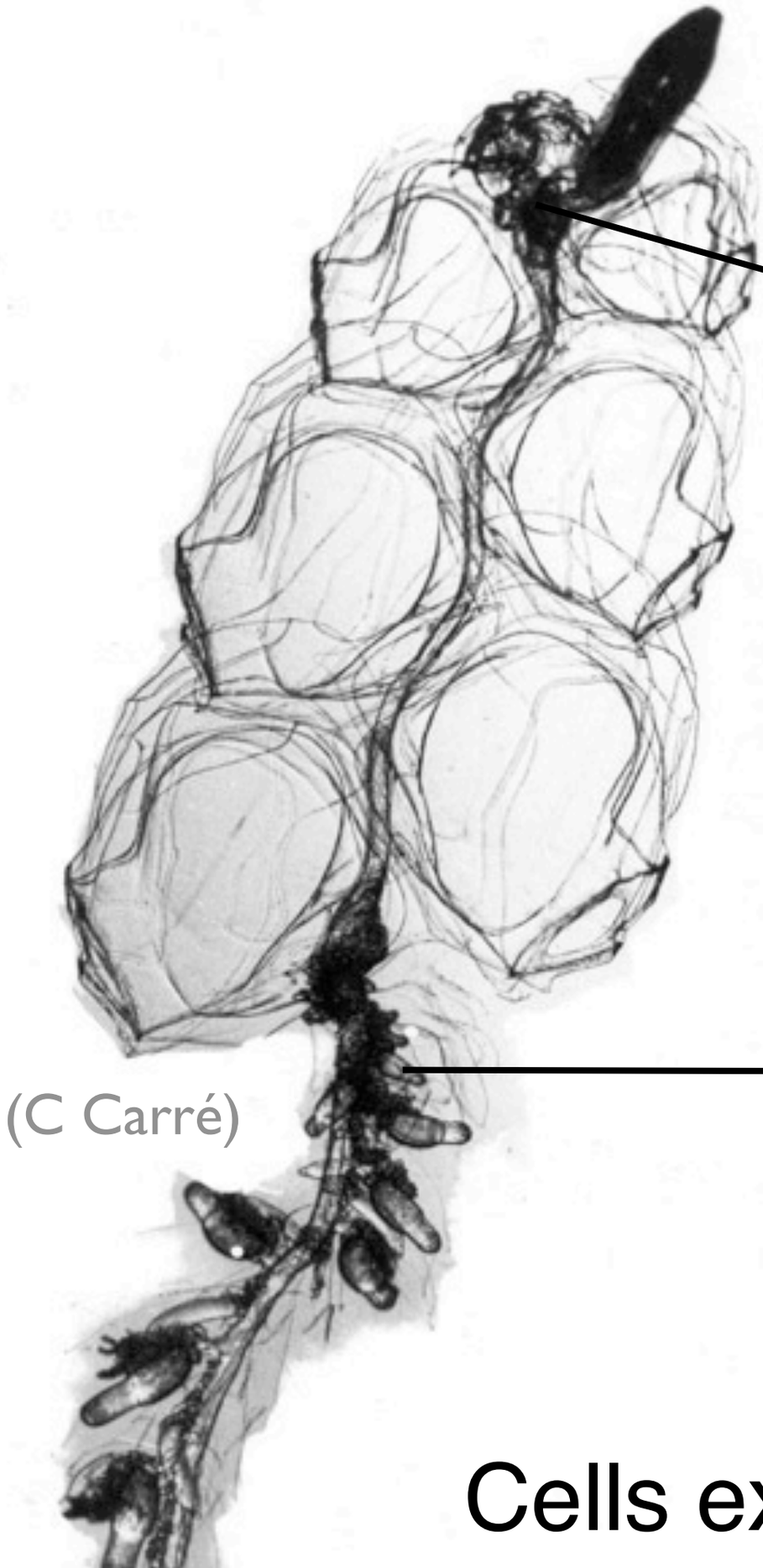
Swimming

Feeding

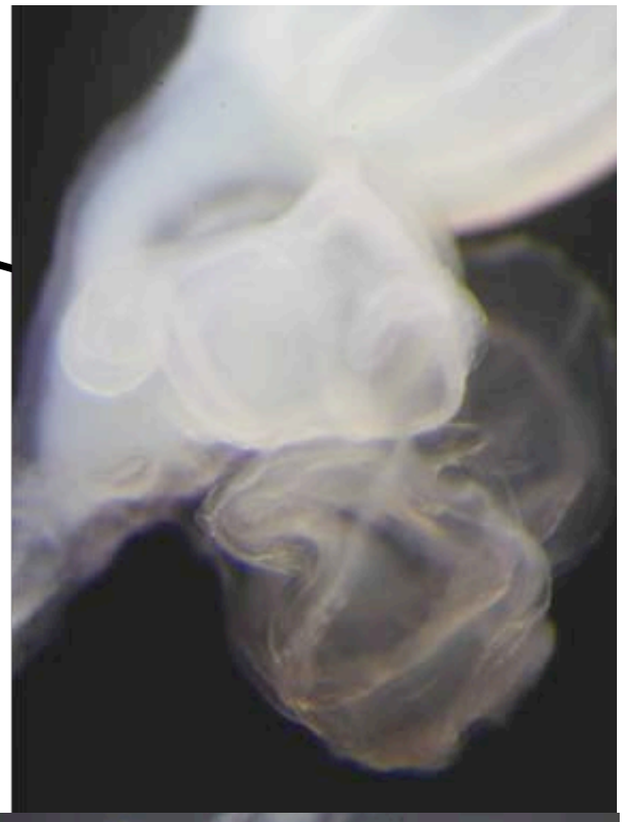


Overall expression

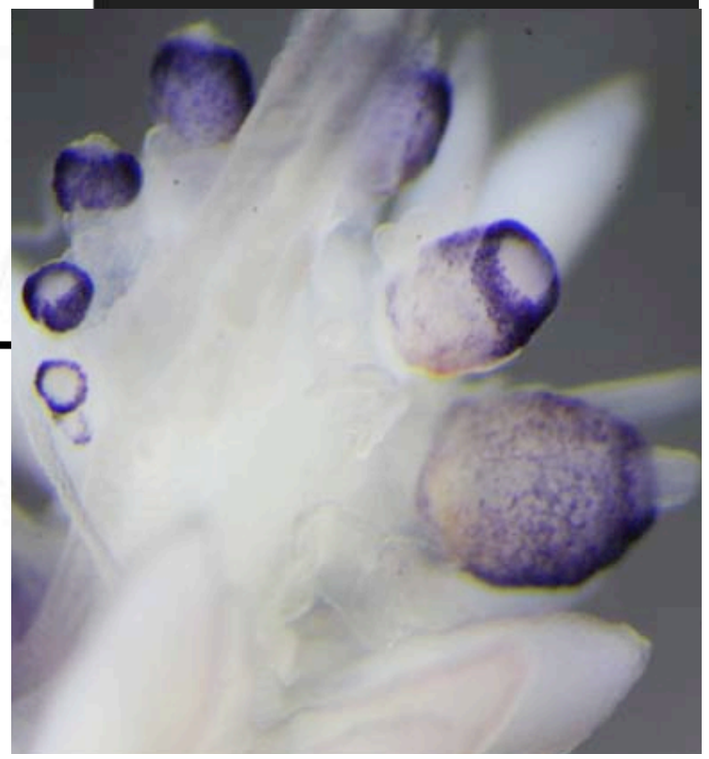




(C Carré)



swimming bodies



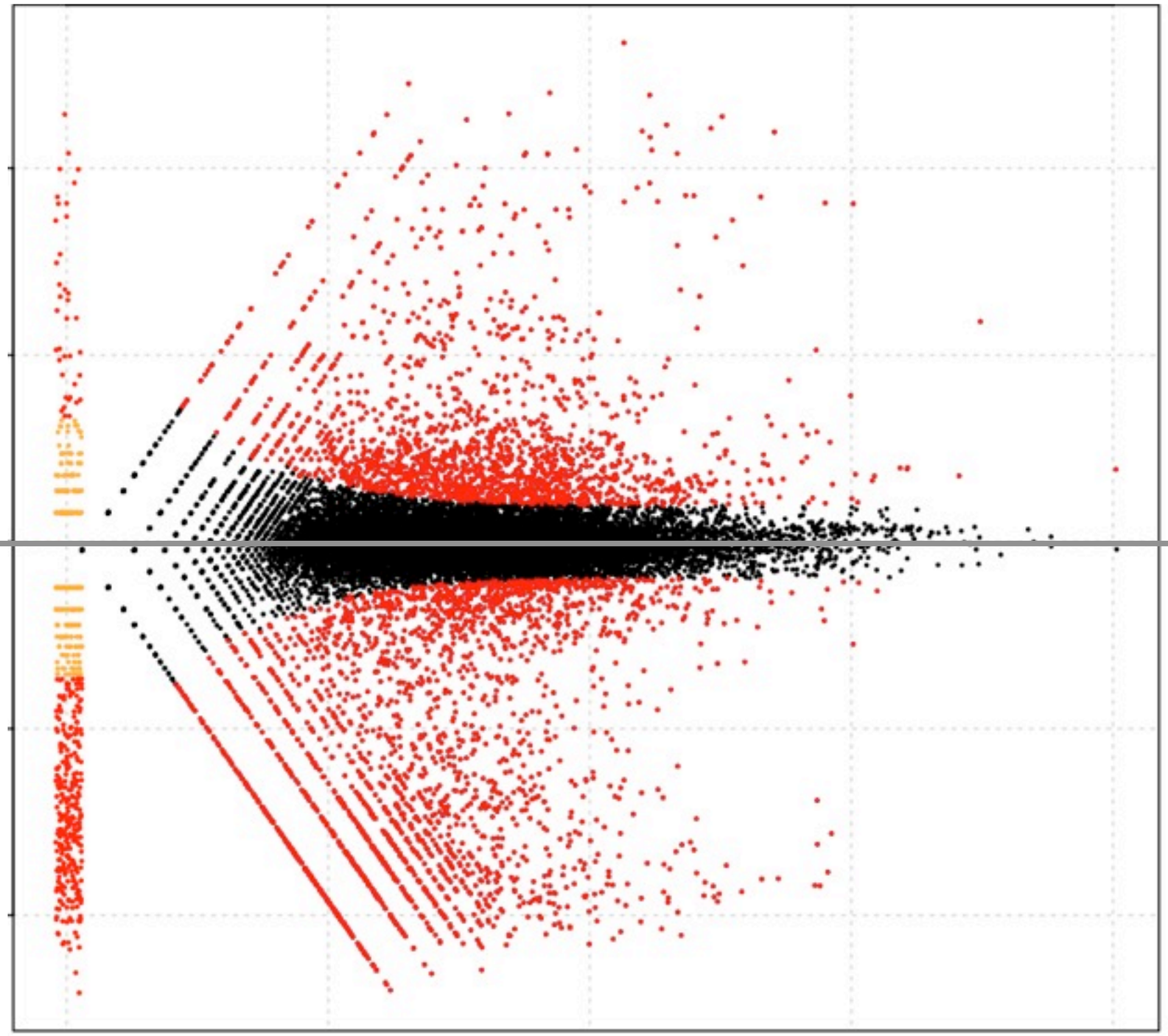
feeding bodies

Cells expressing mini-collagen are **blue**

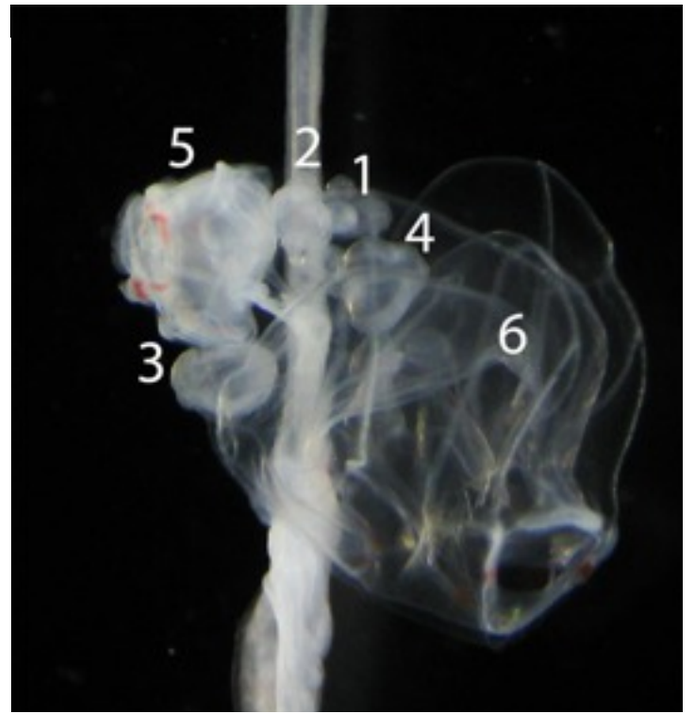
Red genes have significant differential expression

Swimming

Feeding



Overall expression



Uh oh.

"Data deluge"

"Firehose of data"

"I'm drowning in data."

"Data overload"

The problem isn’t too much data.

We need more data that tell us about our data

What other data do we need?

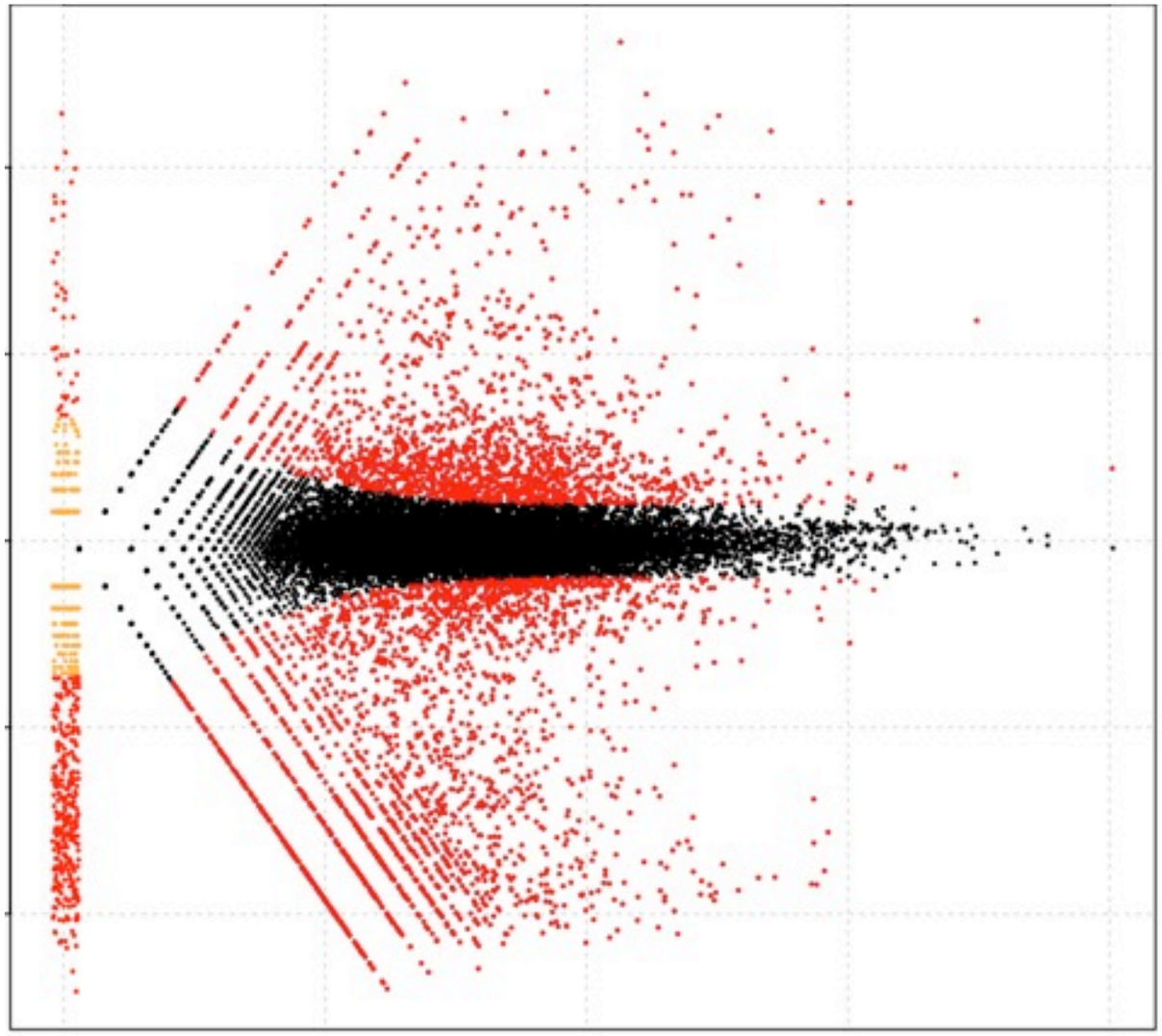
- Comparative data - we need to be looking at a lot more than one species at a time.

Current approach:
Which genes have expression correlated with my phenotype of interest?

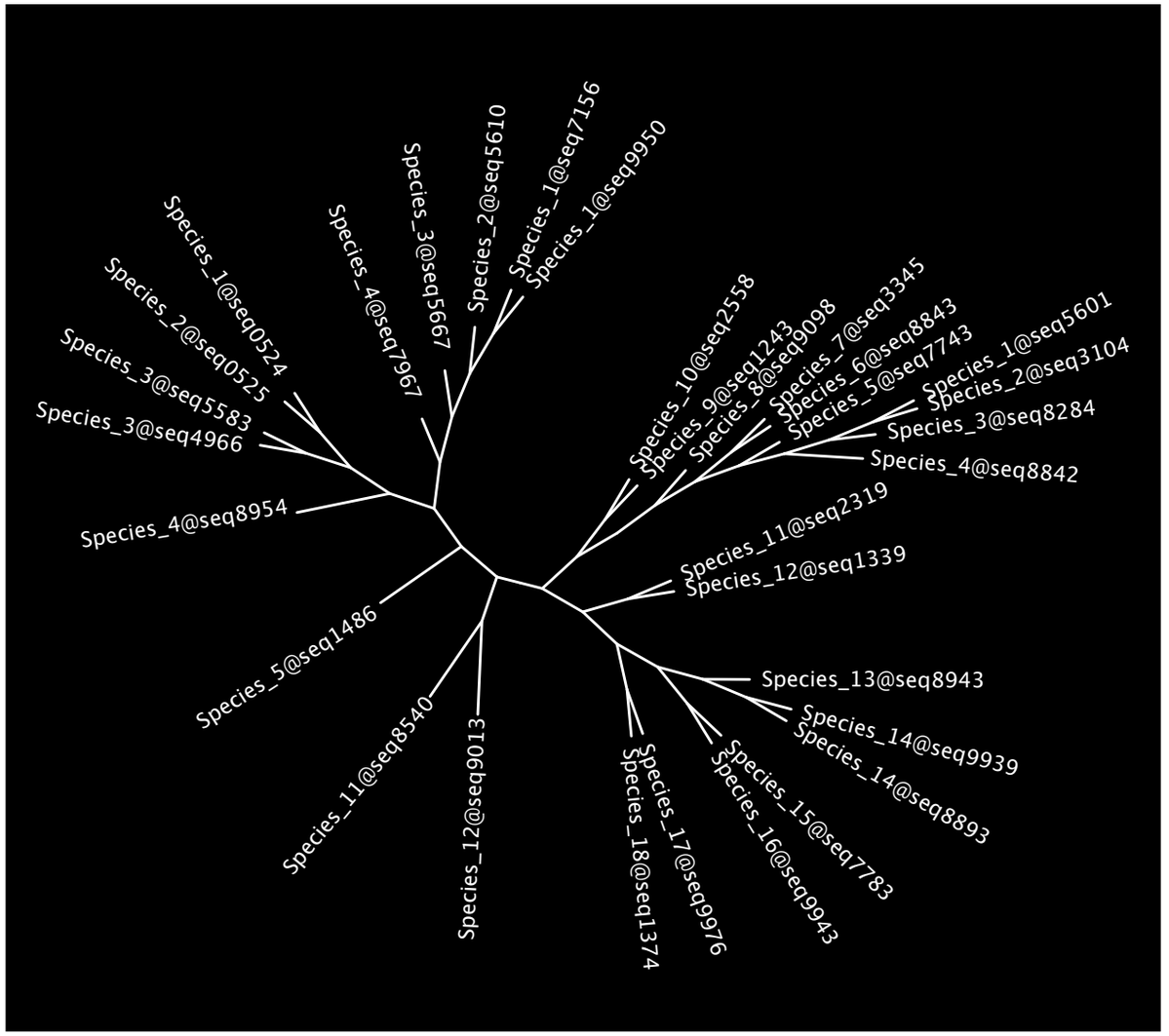
New approach:
Which genes have evolutionary changes in expression that are coincident with changes in my phenotype of interest?

Analyze expression data on phylogenies

Expression data



Gene trees



The state of comparative biology

Casey Dunn

New tools are going to transform comparative biology

But the biggest impact will be how they enable comparative biology to transform the rest of biology

Mechanisms and diversity

20th Century

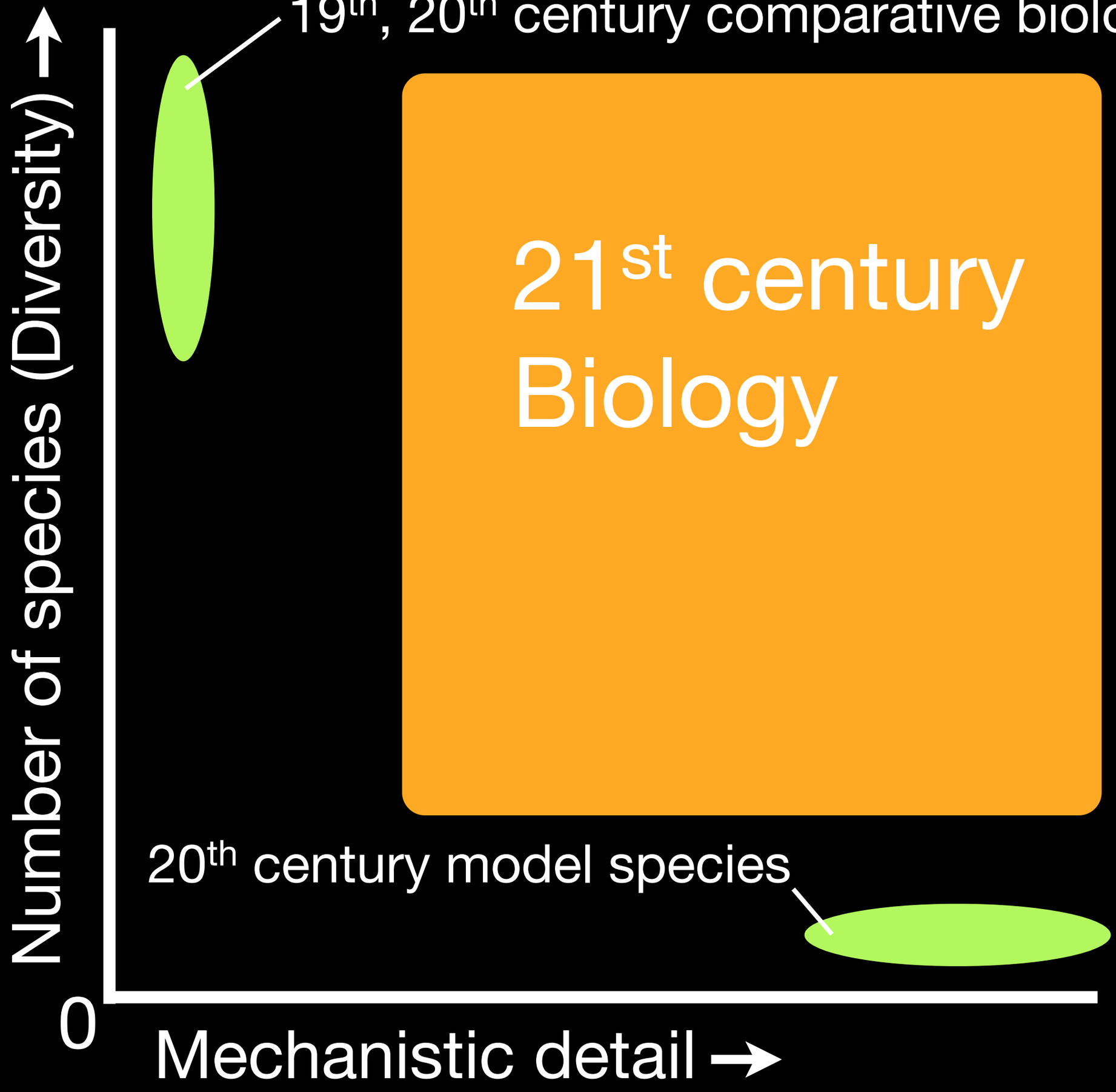
Experimental work in model organism

Comparative work in nonmodel organisms

21st Century

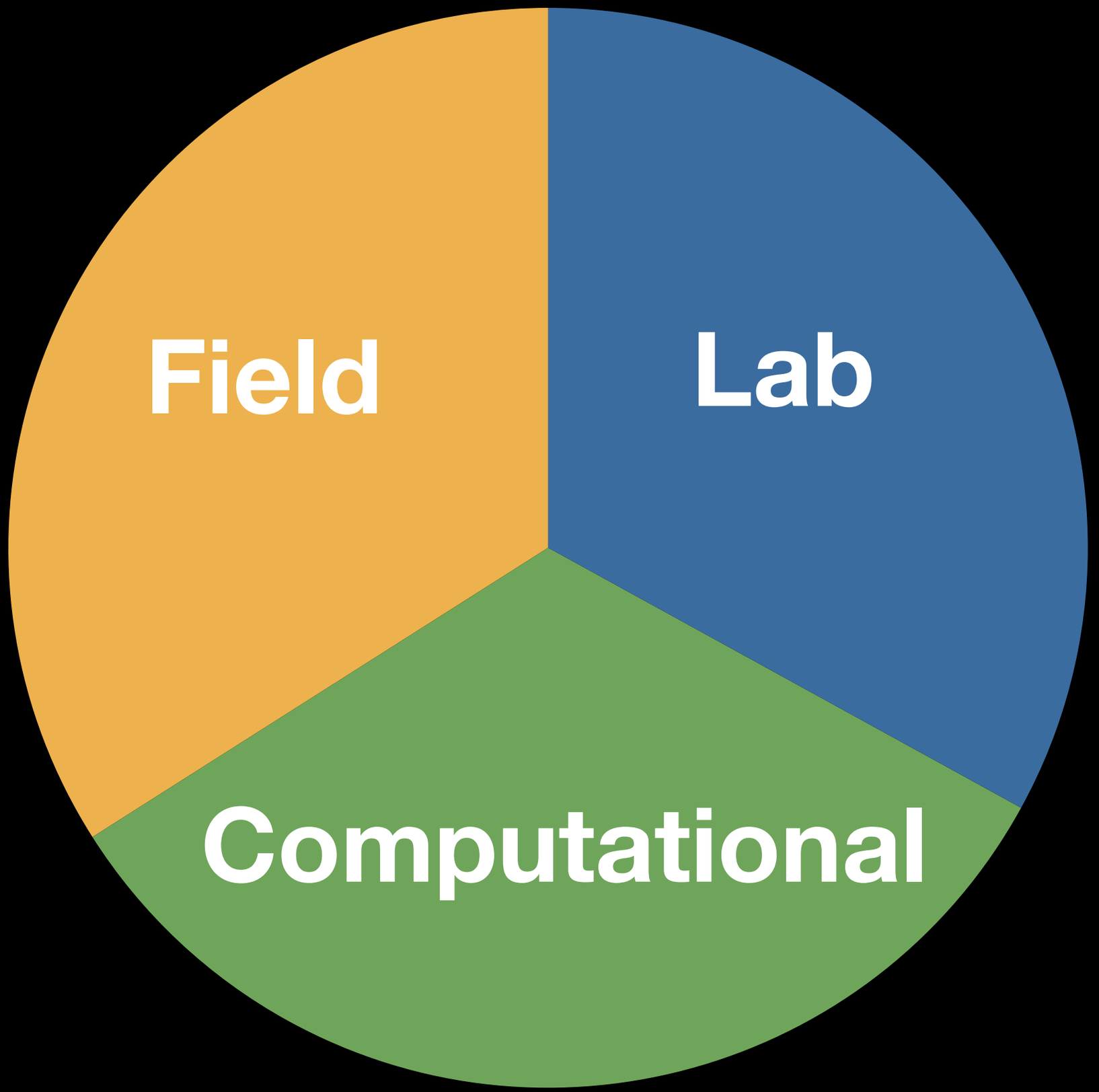
Comparative work in nonmodel organisms

Experimental work in model organism


19th, 20th century comparative biology
21st century Biology
20th century model species
Number of species (Diversity) →
0
Mechanistic detail →

Casey Dunn

Computation


Field
Lab
Computational

practical computing for biologists

Steven H. D. Haddock
The Monterey Bay Aquarium Research Institute, and University of California, Santa Cruz

Casey W. Dunn
Department of Ecology and Evolutionary Biology, Brown University

Sinauer Associates, Inc.

goals

To show you how to use general tools to address the day-to-day computational challenges faced by biologists.

goals

We focus on the entire computer as a general analysis environment, rather than focus on one particular type of analysis or analysis tool.

The material we present can be thought of as glue to hold together and automate your existing analysis tools, and as a general purpose workbench for creating new analysis tools.

goals

Will we show you how to convert file formats so that you can use the output of one program as input to another?

Yes.

Will we show you how to use command-line tools to automate a series of analyses?

Yes.

Will we show you how to use a remote cluster to run programs?

Yes.

goals

Will we show you how to optimize an algorithm to speed it up ten fold?

No.

Will we explain maximum likelihood?

No.

Will we walk you through microarray data analysis?

No.

contents

sample content

```
host:~ lucy$ cd ~/pcfb/sandbox
host:sandbox lucy$ ls
host:sandbox lucy$ cp ../examples/*.txt ./  ← Very different from /  by itself
host:sandbox lucy$ ls
reflist.txt   ← You will see others too...
```

```
#!/usr/bin/env python

DNASeq = 'ATGAAC'
print 'Sequence:', DNASeq

SeqLength = float(len(DNASeq))

print 'Sequence Length:', SeqLength

NumberA = DNASeq.count('A')
NumberC = DNASeq.count('C')
NumberG = DNASeq.count('G')
NumberT = DNASeq.count('T')

print 'A:', NumberA/SeqLength
print 'C:', NumberC/SeqLength
print 'G:', NumberG/SeqLength
print 'T:', NumberT/SeqLength
```

sample content

```
grep ATOM
   REMARK   3  NUMBER OF NON-HYDROGEN ATOMS USED IN REFINEMENT.
   REMARK   3   PROTEIN ATOMS            : 3650
   REMARK 470   M RES CSSEQI  ATOMS
   REMARK 500 RMS DISTANCE OF ALL ATOMS FROM THE BEST-FIT PLANE
   REMARK 500 RMSD 0.02 ANGSTROMS, OR AT LEAST ONE ATOM HAS
grep -v REMARK
   ATOM      1  N   ALA A   1      -14.093  60.494  -9.249  1.00 42.10
   ATOM      2  CA  ALA A   1      -14.989  61.651  -8.981  1.00 41.80
   ATOM      3  C   ALA A   1      -14.809  62.769 -10.006  1.00 41.60
   ATOM      9  O   SER A   2      -11.264  62.734 -12.155  1.00 39.50
   ATOM     10  CB  SER A   2      -13.236  65.292 -11.216  1.00 39.90
   ATOM     11  OG  SER A   2      -12.004  65.880 -11.497  1.00 39.90
   ATOM     12  N   LYS A   3      -12.516  63.462 -13.894  1.00 38.90
   ATOM     13  CA  LYS A   3      -11.712  62.828 -14.936  1.00 38.10
   ...                 ...            ...
   ATOM   3644  CD1 ILE B 229       37.302  62.306   9.573  1.00 42.30
   ATOM   3645  N   THR B 230       39.340  65.048   4.879  1.00 48.70
   ATOM   3646  CA  THR B 230       39.969  64.839   3.566  1.00 50.40
   ATOM   3647  C   THR B 230       41.207  63.924   3.637  1.00 51.30
```

```
cut -c 18-21,
       24-26
   ALA   1
   ALA   1
   ALA   1
   ALA   1
   ALA   1
   SER   2
   SER   2
   SER   2
   SER   2
   SER   2
   ...
   ILE 229
   ILE 229
   ILE 229
   THR 230
   THR 230
   THR 230
   THR 230
   THR 230
   THR 230
   THR 230
```

```
sort | uniq
   ALA   1
   ALA  37
   ALA  87
   ALA 110
   ALA 154
   ALA 179
   ALA 206
   ALA 226
   ALA 227
   ARG  73
   ...
   VAL  68
   VAL  93
   VAL 112
   VAL 120
   VAL 150
   VAL 163
   VAL 176
   VAL 193
   VAL 219
   VAL 224
```

```
cut -f 1
   -d " "
   ALA
   ALA
   ALA
   ALA
   ALA
   ALA
   ALA
   ALA
   ALA
   ARG
   ...
   VAL
   VAL
   VAL
   VAL
   VAL
   VAL
   VAL
   VAL
   VAL
   VAL
```

```
uniq -c
   9 ALA
   7 ARG
  13 ASN
  17 ASP
   2 CYS
   7 GLN
  15 GLU
  21 GLY
   9 HIS
  12 ILE
  18 LEU
  19 LYS
   4 MET
  13 PHE
  10 PRO
  11 SER
  15 THR
   1 TRP
  10 TYR
  17 VAL
```

```
sort -nr
  21 GLY
  19 LYS
  18 LEU
  17 VAL
  17 ASP
  15 THR
  15 GLU
  13 PHE
  13 ASN
  12 ILE
  11 SER
  10 TYR
  10 PRO
   9 HIS
   9 ALA
   7 GLN
   7 ARG
   4 MET
   2 CYS
   1 TRP
```

FIGURE 16.1 The successive extractions and modifications made by each command in the example pipeline Orange boxes show `bash` commands and other boxes show the output once those commands have been added to the pipeline.

sample content

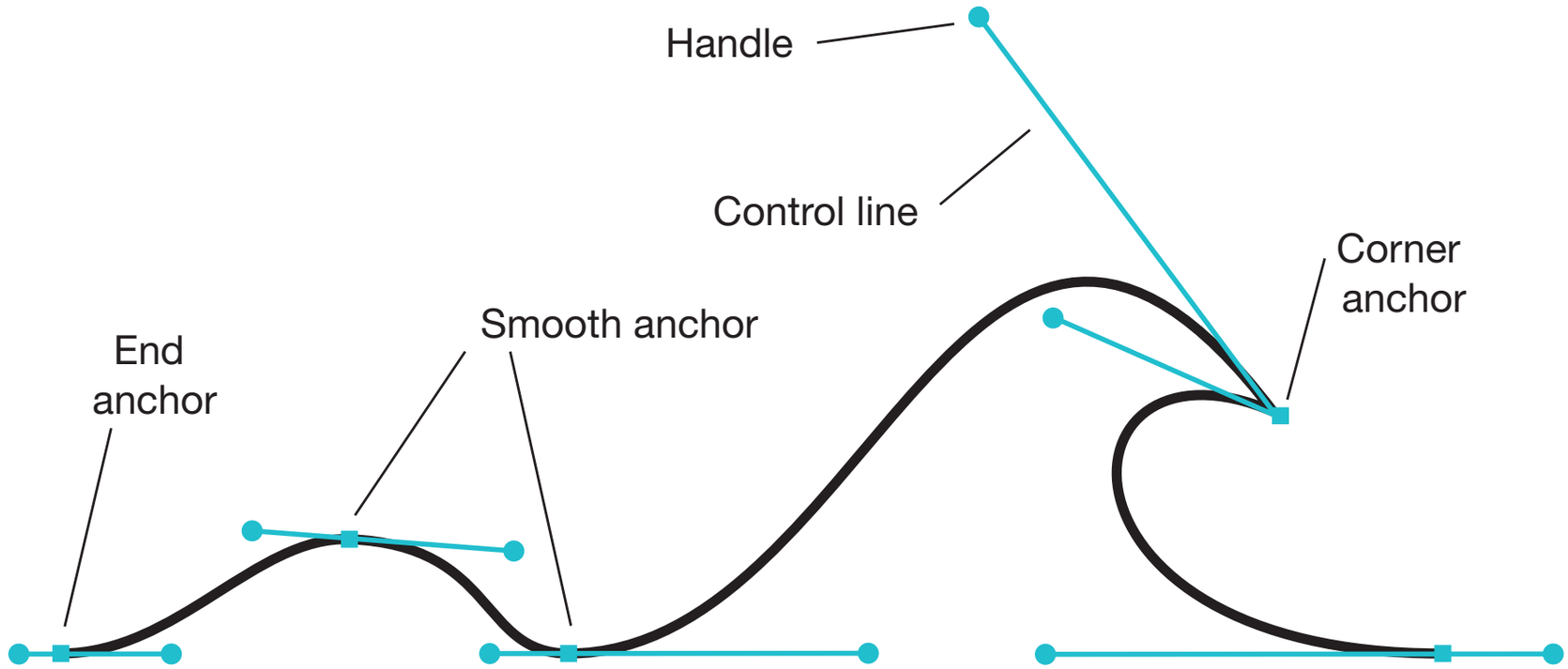


FIGURE 18.2 Bézier curve showing anchor points, handles, and control lines

Casey Dunn