Phylogenomics

Casey Dunn
Associate Professor
Ecology and Evolutionary Biology
@caseywdunn
http://dunnlab.org

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

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

So you want to study molecular evolution in organism X…

1. Design experiment
2. Collect raw data
3. Analysis - Preprocess data
4. Analysis - Molecular evolution
5. Interpret results

In contrast to most other talks, I'm going to focus on these first three steps

1. Design experiment
2. Collect raw data
3. Analysis - Preprocess data
4. Analysis - Molecular evolution
5. Interpret results

As sequencing methods become more sophisticated, preprocessing data becomes a bigger and bigger part of molecular evolution projects

Preprocessing includes:

- Filtering
- Data wrangling (eg formatting)
- Assembly
- Mapping
- Annotation
- Homology evaluation

Understanding sequencing and preprocessing is essential to:

- Implement empirical projects
- Understand errors and ascertainment bias in data
- Design methods that address contemporary challenges

Part I:

Collecting and preprocessing sequence data

Phylodiversity

The Future...

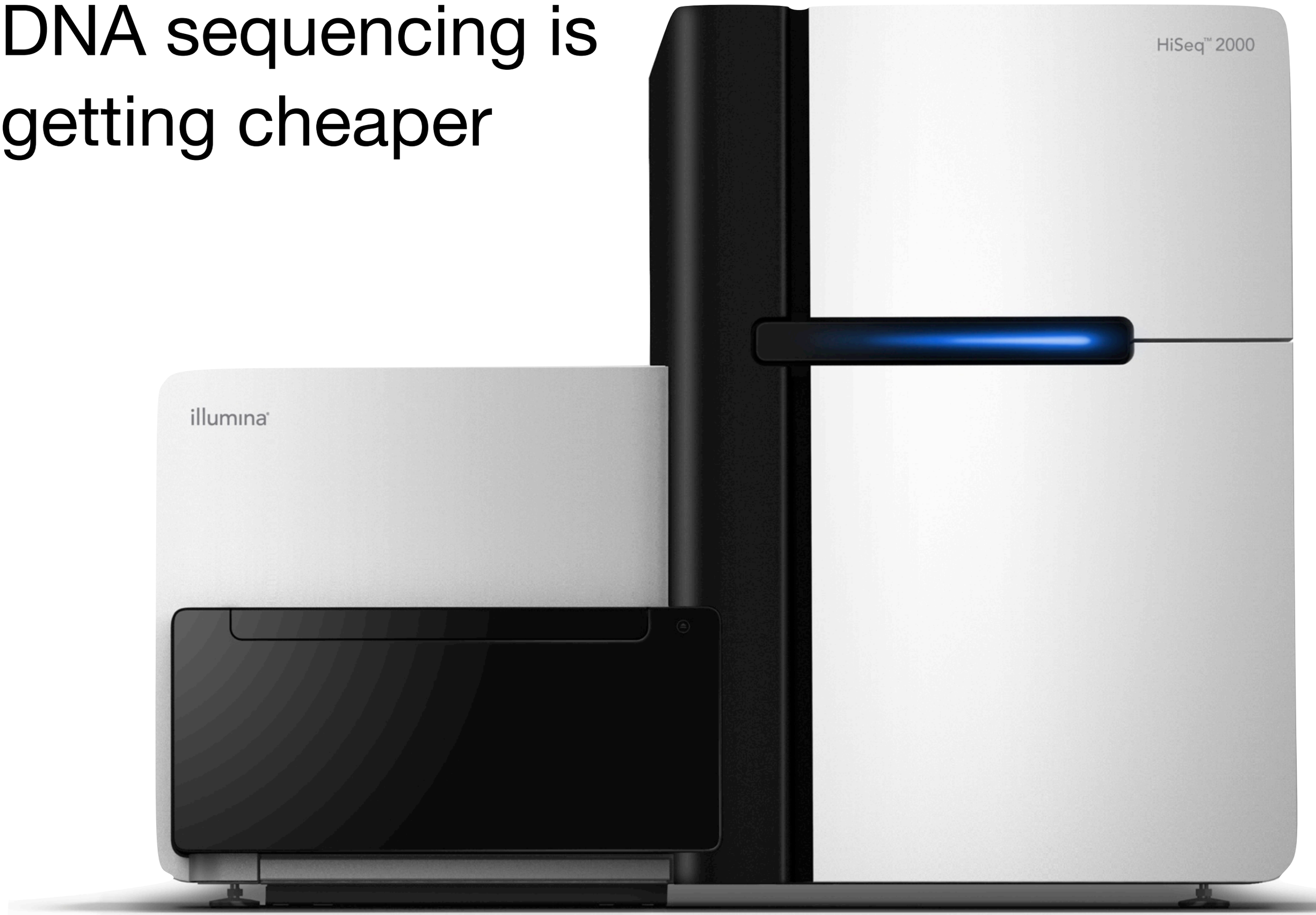
"classical" molecular phylogenetics

Phylogenomics

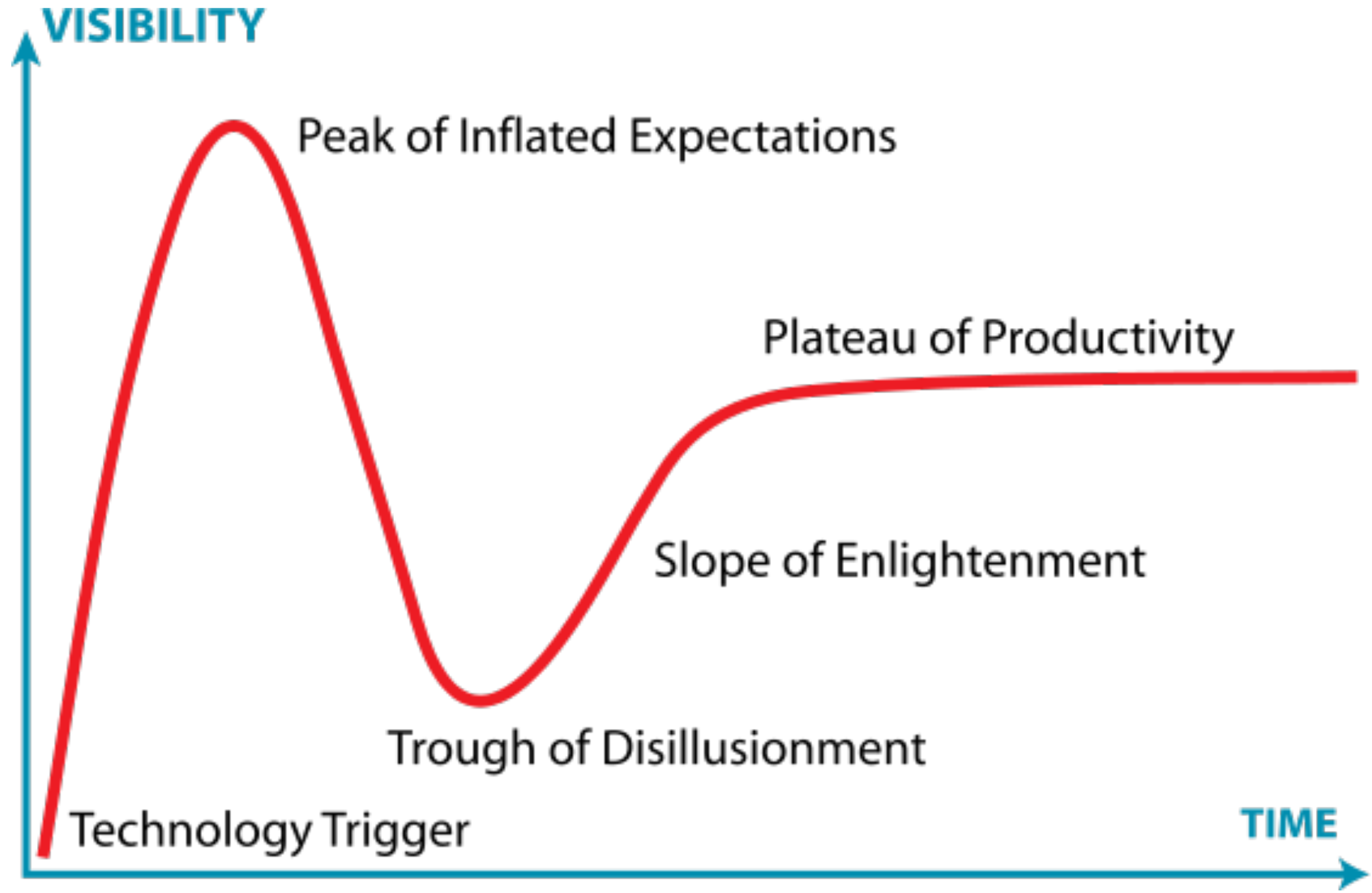
Number of taxa

Number of genes

DNA sequencing is getting cheaper



The Gartner Hype Cycle*



* Not really a cycle

http://en.wikipedia.org/wiki/Hype_cycle#mediaviewer/File:Gartner_Hype_Cycle.svg

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!

We have entirely new opportunities for integrating genomic, morphological, and functional perspectives

Why collect data from lots of genes?

- Gives broad perspective
- 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

Design decisions

There aren’t just more sequences in each molecular evolution analysis...

There are more ways to collect and analyze molecular evolution data.

Which approach is right for you?

Framing questions:

What do you want to know?

What do you already know?

What material will you have available (DNA, RNA, or both)?

Central technical question:

Will you enrich your sample for particular genome regions prior to sequencing?

Enrichment reduces the amount of sequence data you need to collect.

It allows you to sequence homologous genome regions across multiple individuals and species.

Enrichment spectrum

Directed PCR

Whole genome

Targeted enrichment

"Whole" transcriptome

RAD tag

Increasing enrichment

Directed PCR

Whole genome

Targeted enrichment

"Whole" transcriptome

RAD tag

Increasing enrichment

Whole genome

- No enrichment.

- In a phylogenetic context, currently only cost effective for small genomes.

- Often need transcriptome data to annotate genes.

Directed PCR

Whole genome

Targeted enrichment

"Whole" transcriptome

RAD tag

Increasing enrichment

Whole transcriptome

- Enriched for expressed protein coding genes

- There is no One True Transcriptome

Directed PCR

Whole genome

Targeted enrichment

"Whole" transcriptome

RAD tag

Increasing enrichment

Targeted enrichment

- Use hybridization to enrich particular regions
- Works well even on degraded DNA
- Need to synthesize probes specific to each region

Directed PCR

Whole genome

Targeted enrichment

"Whole" transcriptome

RAD tag

Increasing enrichment

RAD tag

- Enriched for randomly distributed, but consistent, genome regions

- No need for specific probes

Directed PCR

Whole genome

Targeted enrichment

"Whole" transcriptome

RAD tag

Increasing enrichment

Directed PCR

- Simple and cheap for a small number of genes

- Doesn’t scale so well to many genes

As prices fall, the best approach tends to move to the left.

Directed PCR

Whole genome

Targeted enrichment

"Whole" transcriptome

RAD tag

Increasing enrichment

Back to the big question...

Is directed PCR, targeted enrichment, transcriptome, or genome sequencing better for phylogenetics?

Nonsensical question!

We used to have a small number of tools for enrichment and sequencing.

We used them for everything.

(Smithsonian)

Nonsensical question!

Now we have an amazing set of specialized tools.

Can fit the tool to the project.

Many features of enrichment strategies are an advantage for some projects and a disadvantage for other projects.

eg, sometimes ascertainment bias is good and sometimes it is bad

The major conceptual difference between these methods is whether genes are selected before or after sequencing

Select genes before sequencing

Directed PCR

Whole genome

Targeted enrichment

"Whole" transcriptome

RAD tag

Increasing enrichment

Select genes after sequencing

Directed PCR

Whole genome

Targeted enrichment

"Whole" transcriptome

RAD tag

Increasing enrichment

Before

Select genes

↓

Amplify and sequence selected genes

↓

Assemble matrix from all sequenced genes

↓

Phylogenetic inference

After

Sequence at random

↓

Identify homologous sequences and evaluate paralogy

↓

Select genes

↓

Assemble matrix

↓

Phylogenetic inference

Selecting after sequencing is a pain if you already knew what you wanted before you started…

But a huge advantage if you don’t know ahead of time.

Identifying and selecting homologs

Species A

Species B

Species C

Species

Gene

Species A

Species B

Species C

Gene divergence due to duplication

Clearest Orthology

Most Informative

Clearest Homology

Available Data

Phylogenetic tools build trees from homologous characters

Most phylogenetic tools assume character homology, they can’t evaluate homology

We need to make a first pass with phenetic tools

Some tools evaluate both homology and orthology with phenetic methods

Use phenetic tools to add new sequences into an existing matrix of pre-selected orthologs

HamStR
dx.doi.org/10.1186/1471-2148-9-157

Some tools evaluate both homology and orthology with phenetic methods

Use phenetic tools to identify orthologs *de novo*

Nice review by Chen et al 2007
dx.doi.org/10.1371/journal.pone.0000383

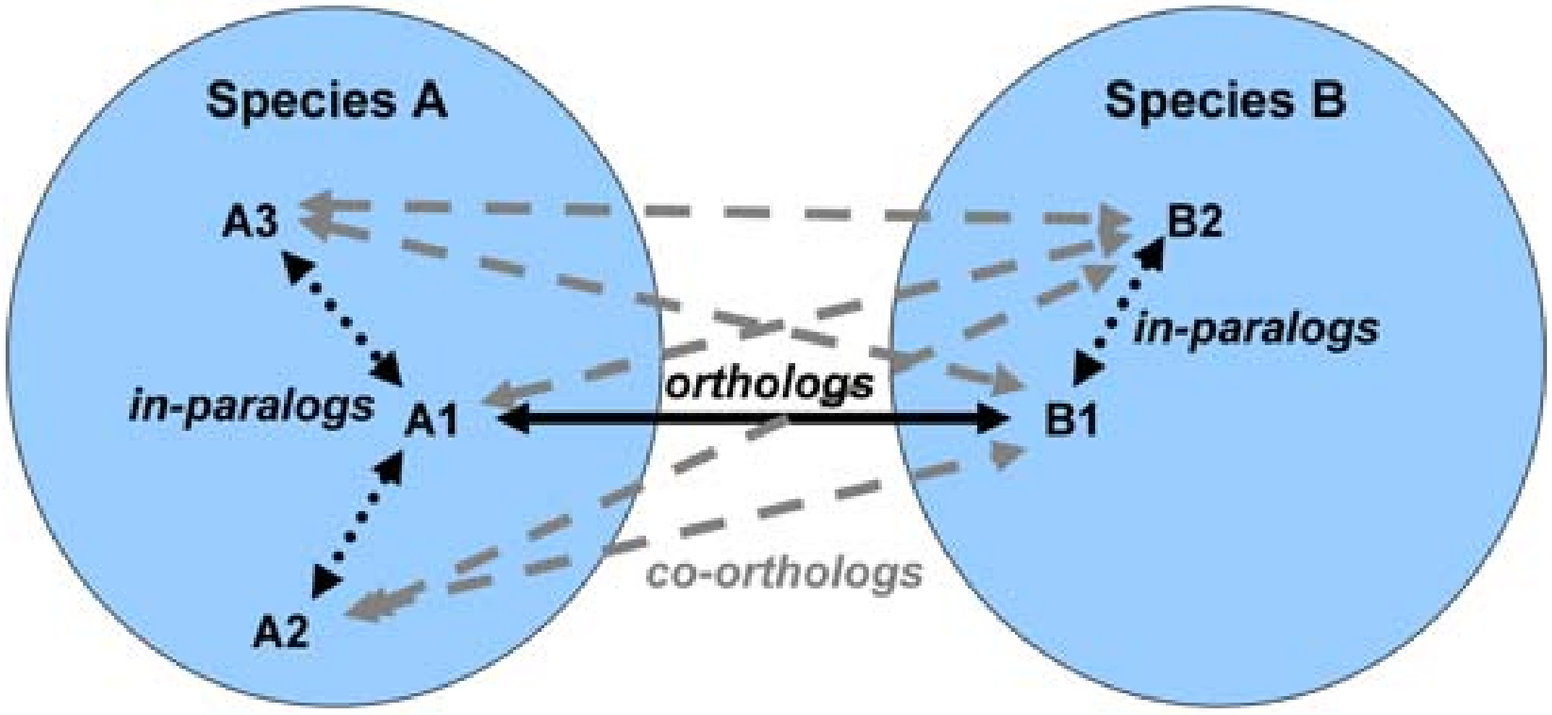


Figure 1. OrthoMCL graph construction between two species, including the establishment of co-ortholog relationships. Solid lines connecting A1 and B1 represent putative ortholog relationships identified by the 'reciprocal best hit' (RBH) rule. Dotted lines (e.g. those connecting A1 with A2 and A3, or B1 with B2) represent putative in-paralog relationships within each species, identified using the 'reciprocal better hit' rule. Putative co-ortholog relationships, indicated by dashed gray lines, connect in-paralogs across species boundaries (e.g. A3 and B2).
doi:10.1371/journal.pone.0000383.g001

Some tools evaluate homology with phenetic methods and orthology with phylogenetic methods

Species A

Species B

Species C

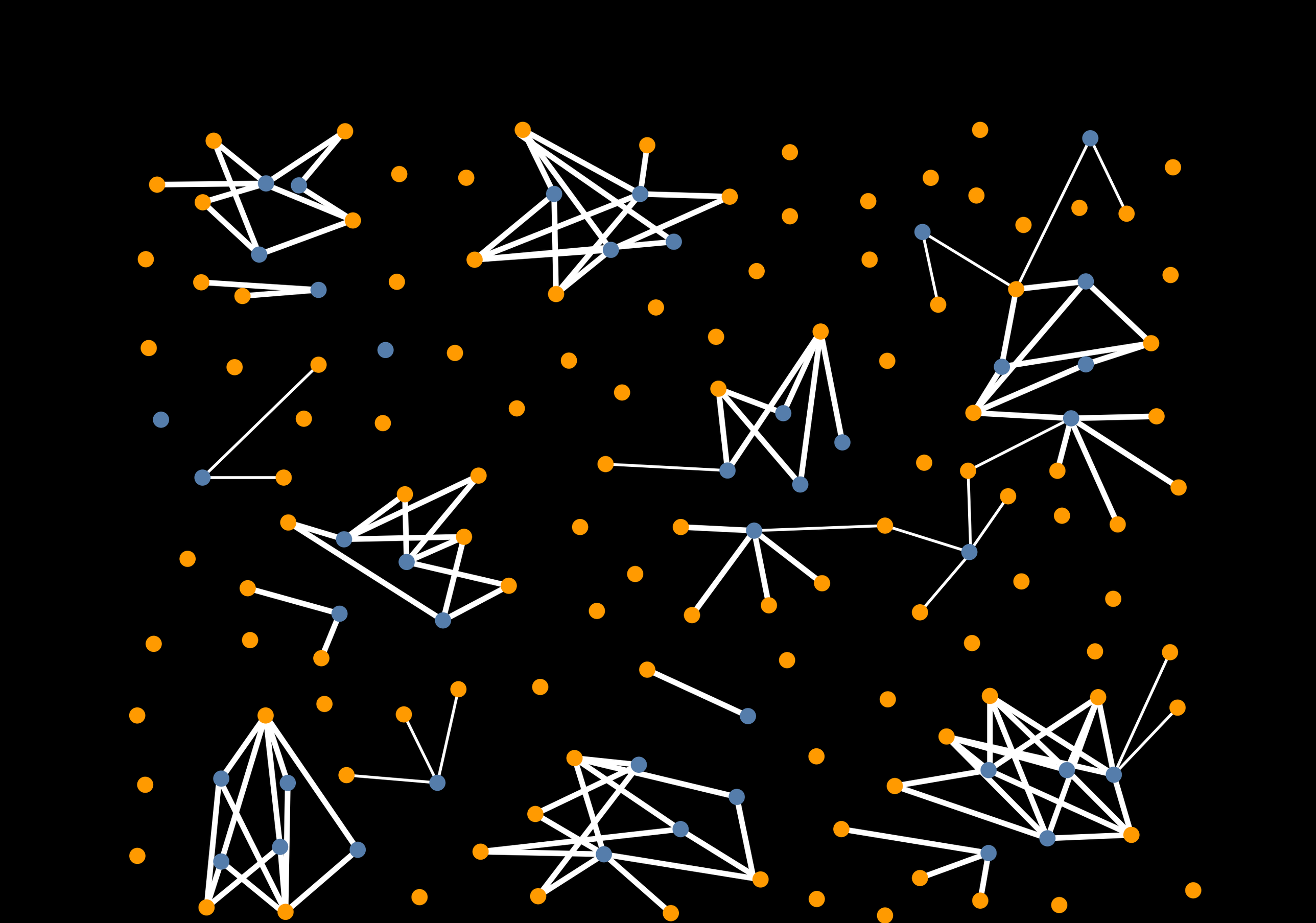
This is our approach...

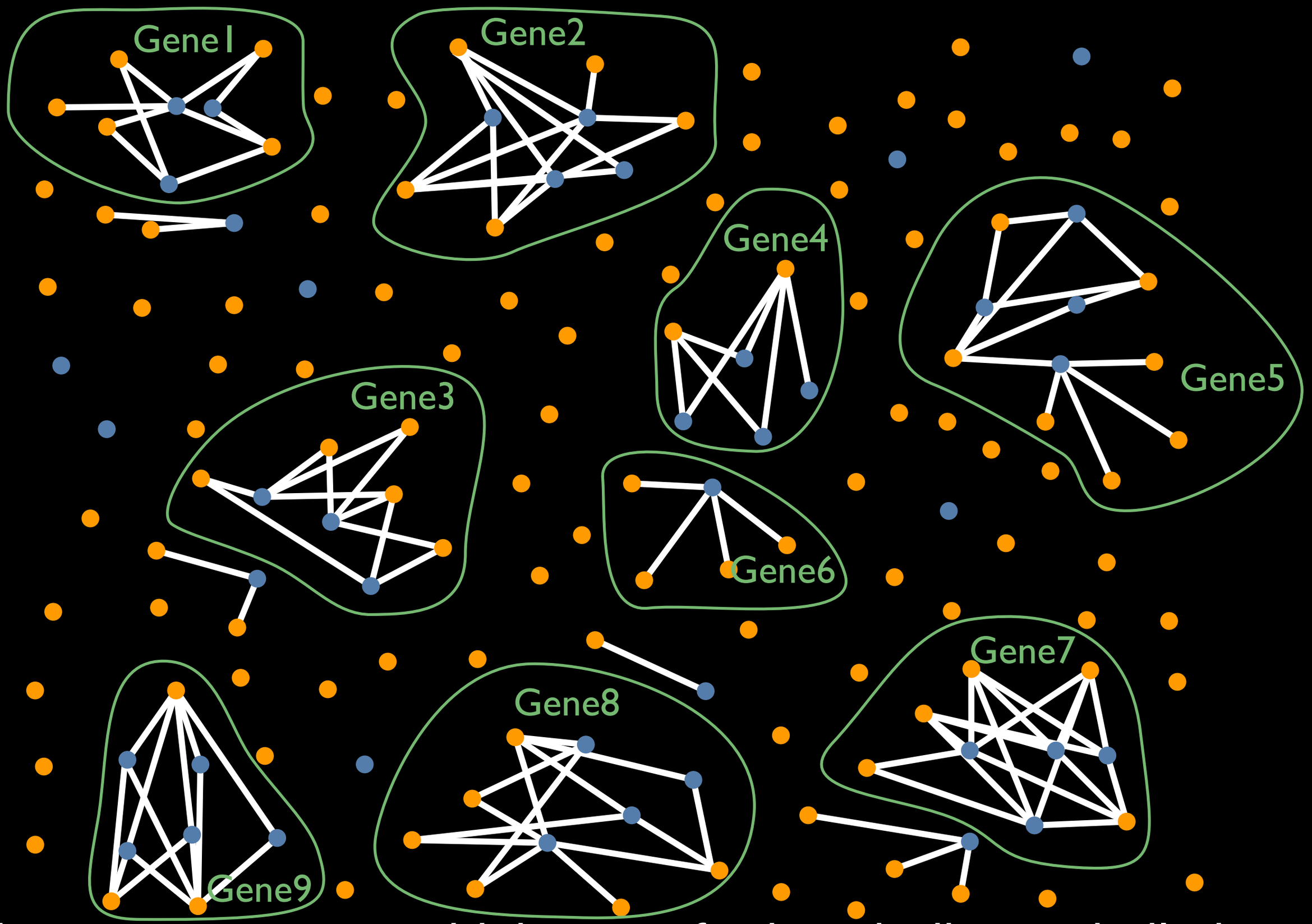
Put all sequences for all taxa in a study into a hat

Make all pairwise sequence comparisons

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

“The paralogy problem”

But paralogs aren’t inherently a problem

The problem is misascribing paralogs as orthologs

Species A

Species B

Species C

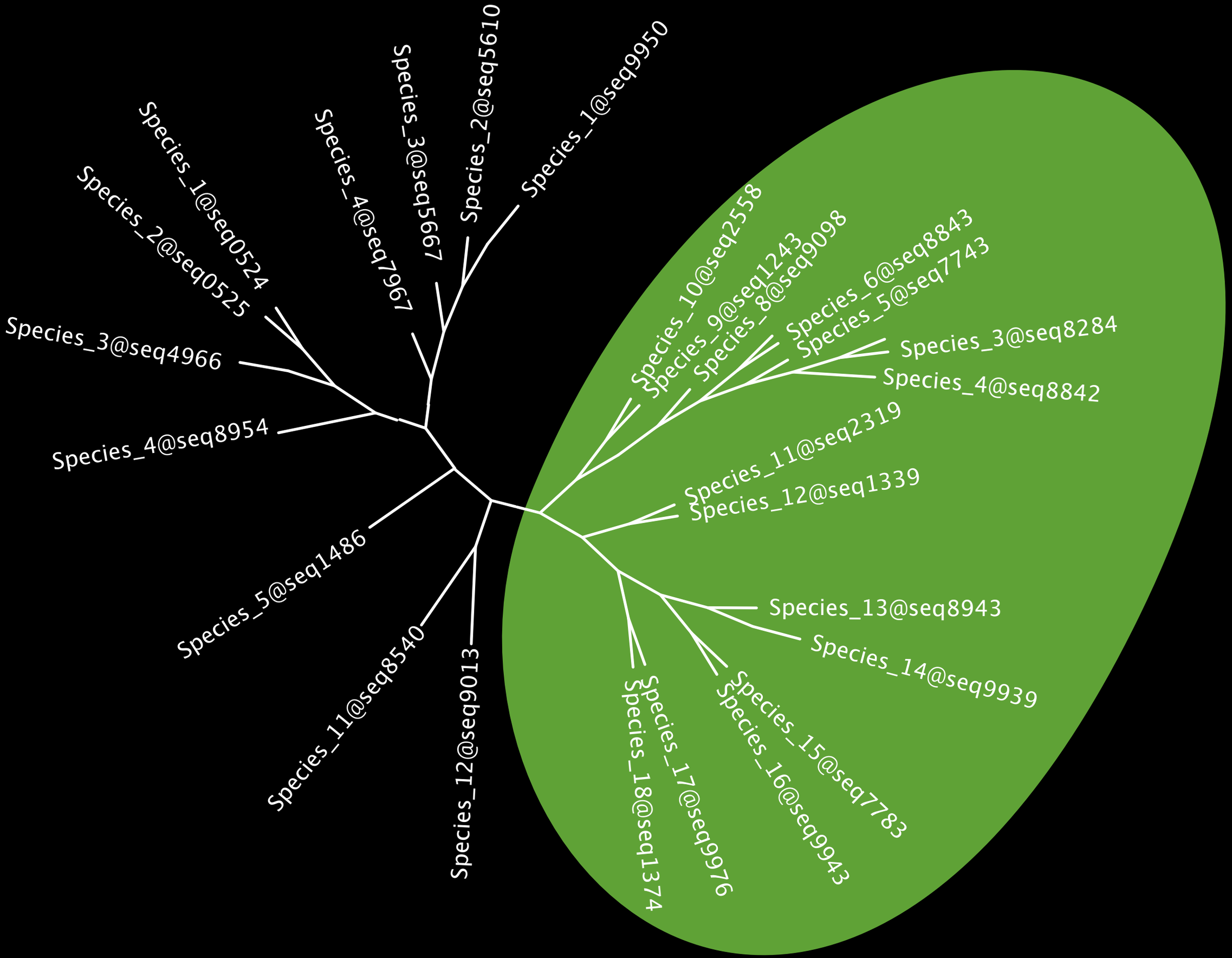
Gene divergence due to duplication

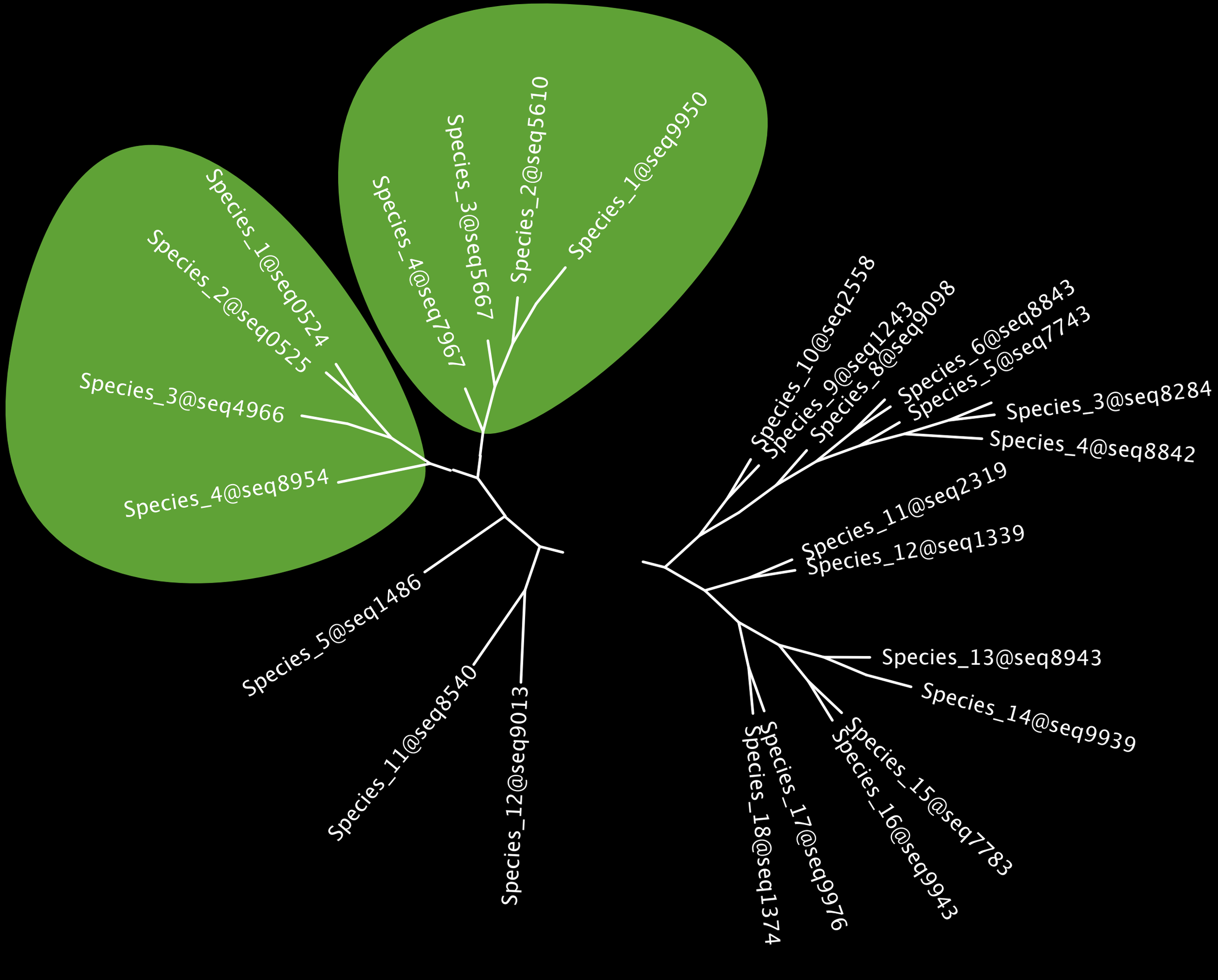
Species A

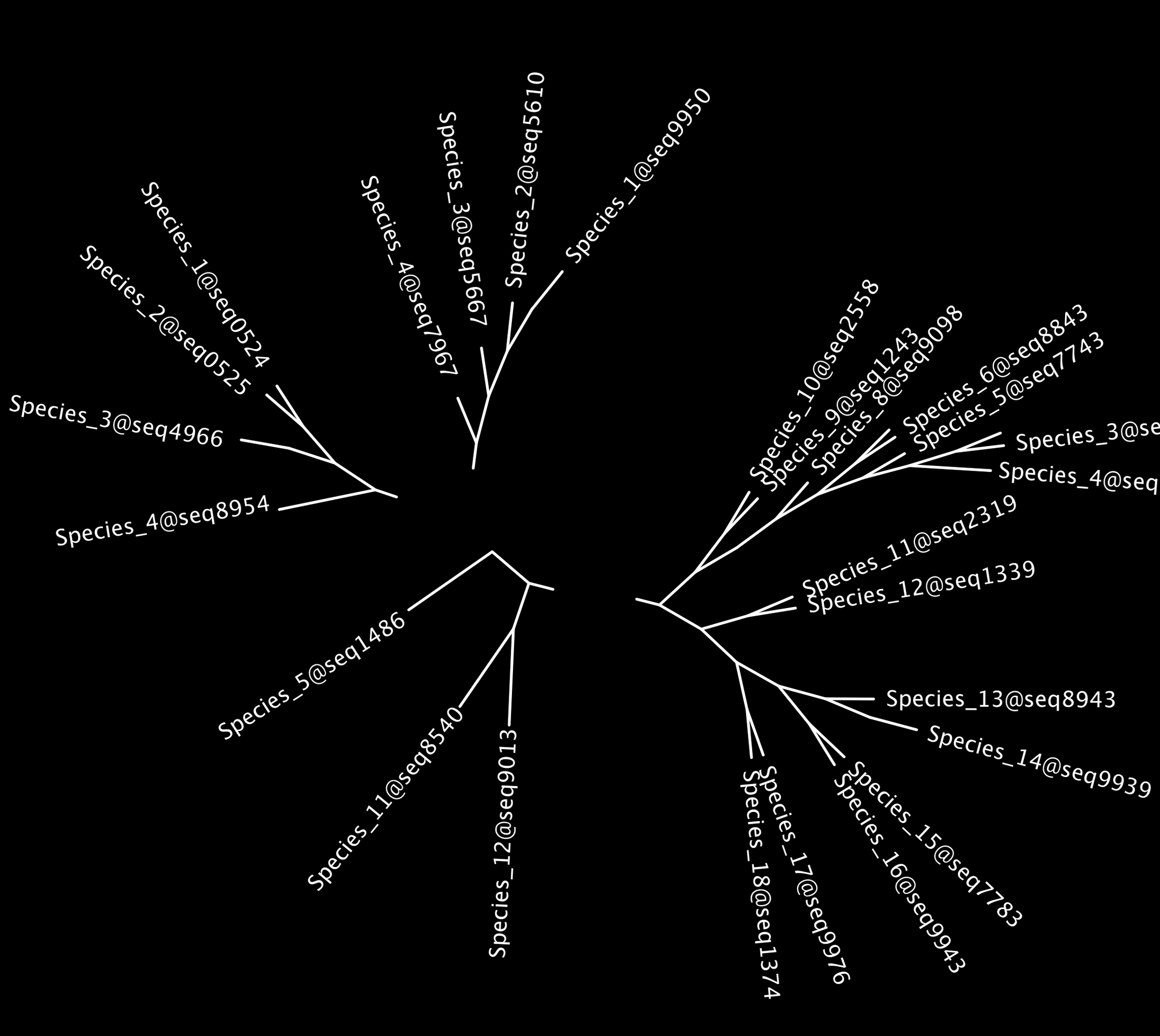
Species B

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_13@seq8943
Species_14@seq9939
Species_15@seq7783
Species_16@seq9943
Species_17@seq9976
Species_18@seq1374
Species_12@seq9013
Species_11@seq8540
Species_5@seq1486


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


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

Once we have subtrees of orthologs...

Align each ortholog

Build trees

77 taxa, 150 Genes, >20k aa

White cells indicates sampled gene

50.9% gene sampling

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

Can do this with:

https://bitbucket.org/caseywdunn/agalma

Bitbucket Repositories Create owner/repository

agalma

caseywdunn Following Share

Clone Fork Compare Pull request

Overview Source Commits Pull requests Issues 1 Downloads 1

Agalma is developed by the Dunn Lab at Brown University.

See TUTORIAL for an example of how to use Agalma with a sample dataset.

SSH git@bitbucket.org:caseywdunn/aga

1 Branch 0 Tags 0 Forks 3 Followers

Owner Casey Dunn
Access level Public
Type Git
Last updated 2013-02-19
Created 2013-02-07
Size 1.9 MB (download)

Overview of Agalma

Agalma is a set of analysis pipelines for transcriptome assembly (paired-end Illumina data) and phylogenetic analysis. It can import gene predictions from other sources (eg, assembled non-Illumina transcriptomes or gene models from annotated genomes), enabling broadly-sampled "phylogenomic" analyses.

Agalma provides a completely automated analysis workflow that filters and assembles the data under default parameters, and records rich diagnostics. The same goes for alignment, translation, and phylogenetic analysis. You can then evaluate these diagnostics to spot problems and examine the success of your analyses, the quality of the orignal data, and the appropriateness of the default parameters. You can then rerun subsets of the pipelines with optimized parameters as needed.

The workflow is highly optimized to reduce the RAM and computational requirements, as well as the disk space used. It logs detailed stats about computer resource utilization to help you understand what type of computational resources you need to analyze your data and to further optimize your resource utilization.

The main functionality of this workflow is to:

- assess read quality with the FastQC package
- remove clusters in which one or both reads have Illumina adapters (resulting from small inserts)
- remove clusters where one or both reads is of low mean quality
- randomize the sequences in the same order in both pairs to make obtaining random subsets easy
- assemble and annotate rRNA sequences based on a subassembly of the data
- remove clusters in which one or both reads map to rRNA sequences

Homology evaluation is poised to undergo a radical transition in the next few years…

Rather than:

1) Use phonetic tools to identify homologous sequences
2) Use phylogenetic tools to identify orthologs
3) Use phylogenetic tools to infer species relationships

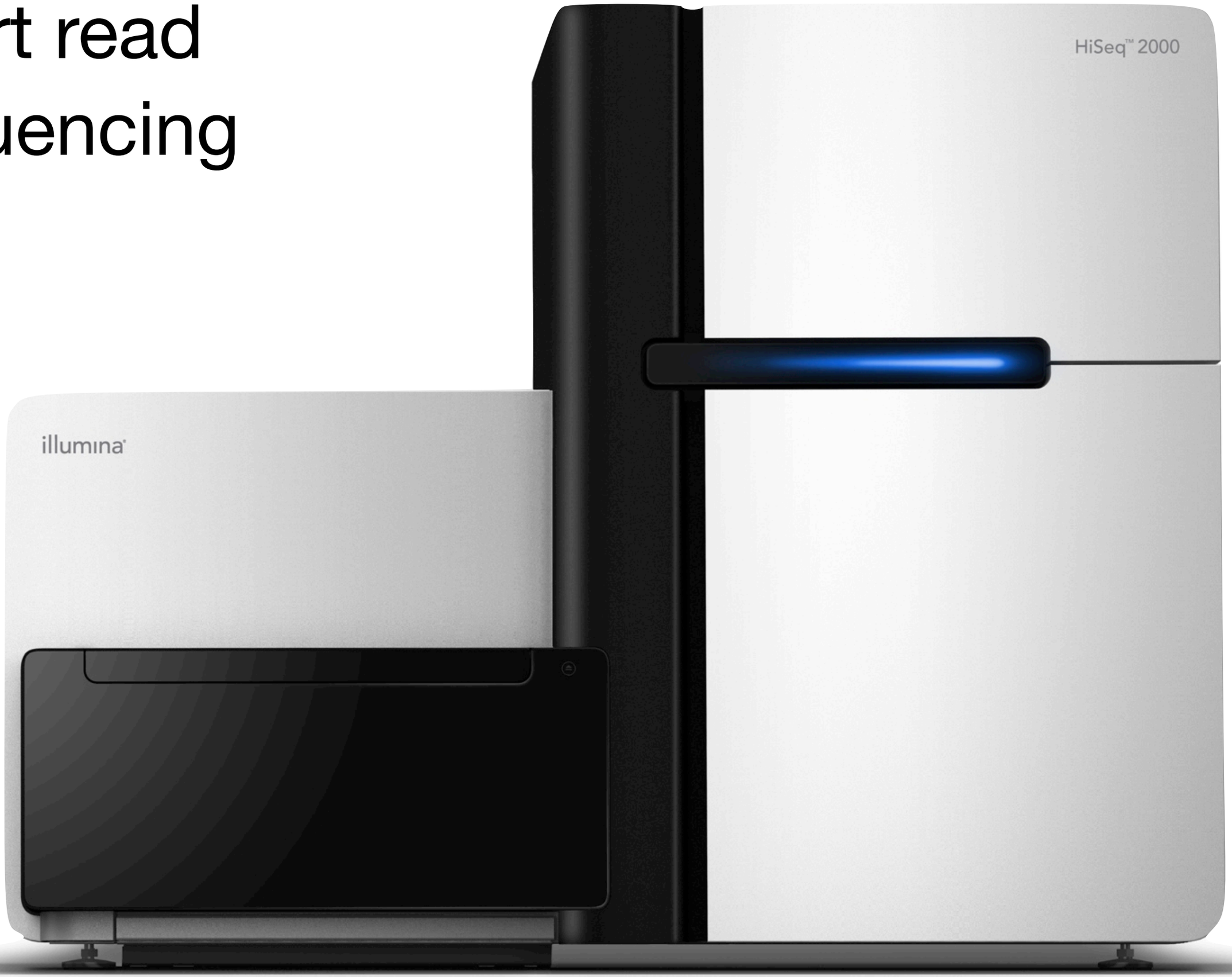
We will:

1) Use phenetic tools to identify homologous sequences
2) Use phylogenetic tools to simultaneously infer gene trees and species trees by modeling gene gain/ loss

A closer look at each enrichment strategy

Whole genome (de novo assembly)

Short read sequencing



Sample preparation

Library preparation usually includes:

- Fragmentation
- Size selection
- Adapter integration
- Amplification

Why fragment?

1. Most sequencers require the input material to have a particular size range

2. To make sequencing coverage more uniform

Starting material

Fragment

Fragments

Prepare library, sequence

Reads

Library preparation options:

Get a library preparation kit from the sequencer vender

Get a third party library preparation kit

Make the library from scratch

The most common library preparation problems:

Poor input material

Over-amplification

Poor size selection

Sequencing

For many studies, sample prep is already more expensive than sequencing.

We are approaching a point where sequencing costs are negligible.

Data are usually delivered in fastq 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
```

Data Preprocessing:

Assembly

Annotation

Assembly

Assembly undoes fragmentation (and reduces redundancy).

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

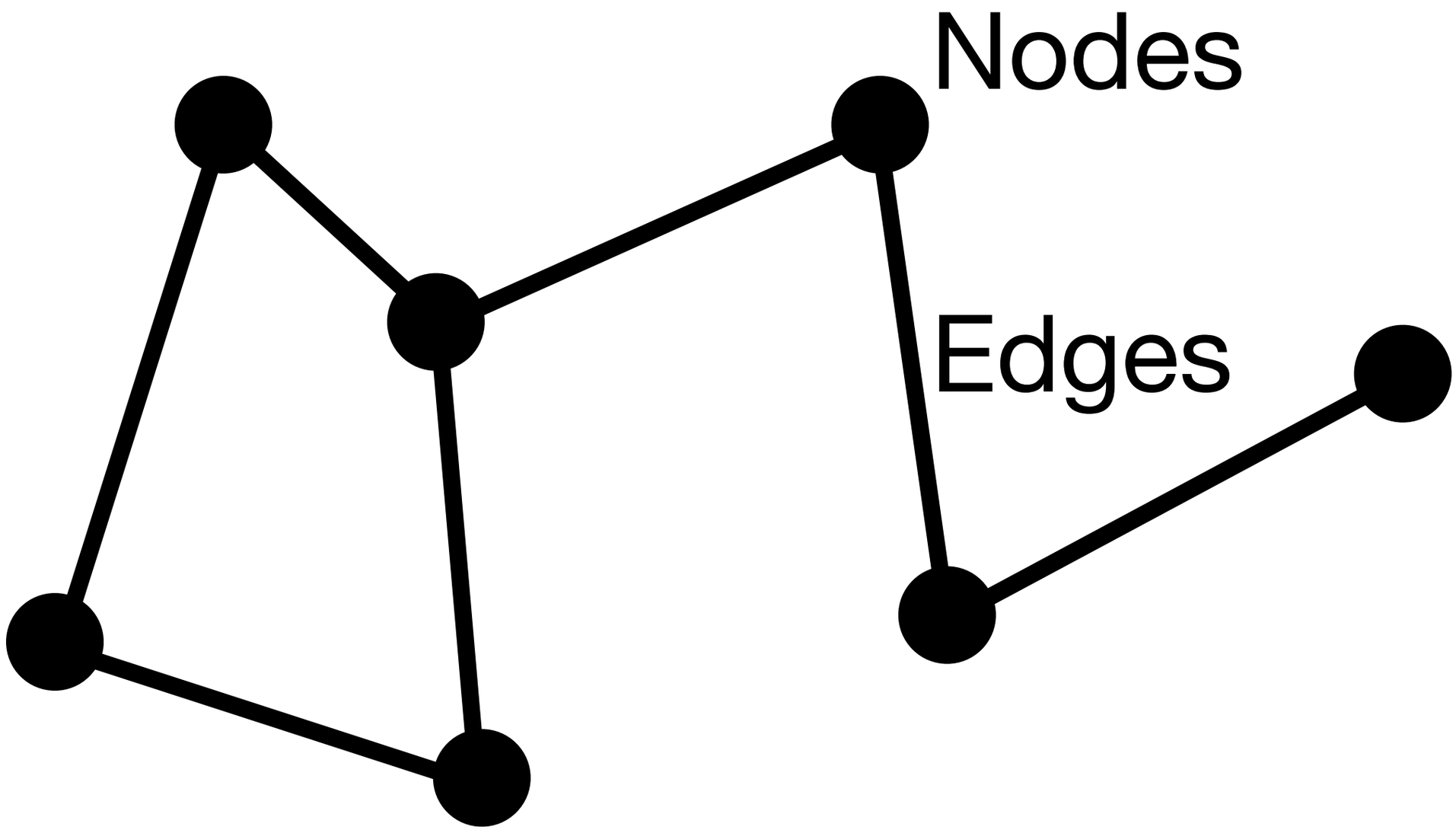
The number of pairwise comparisons that are needed to detect overlap become intractable

de Bruijn graph assemblers have been developed to meet these challenges

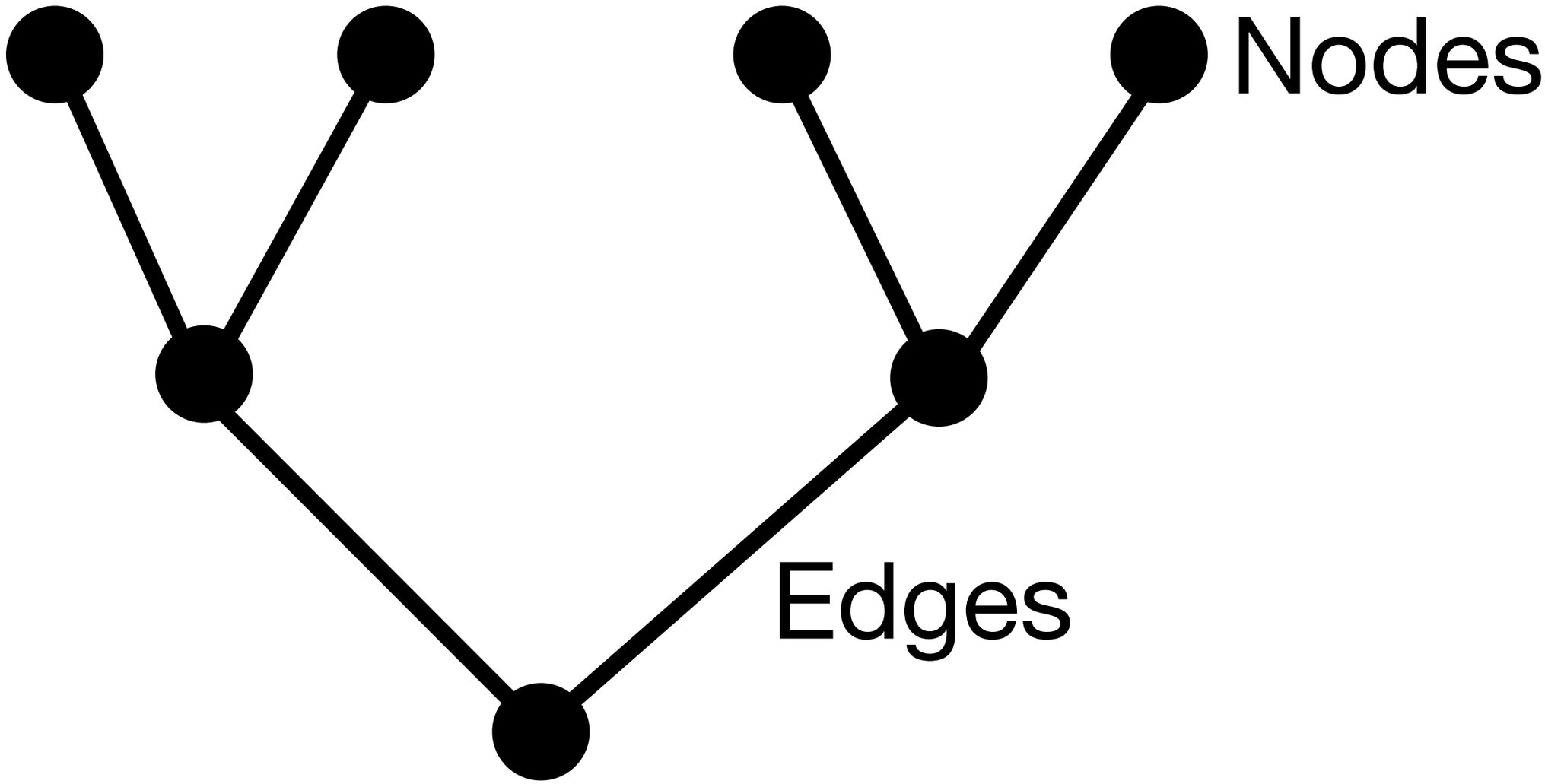
Better defined memory footprint

Simpler comparisons between sequences

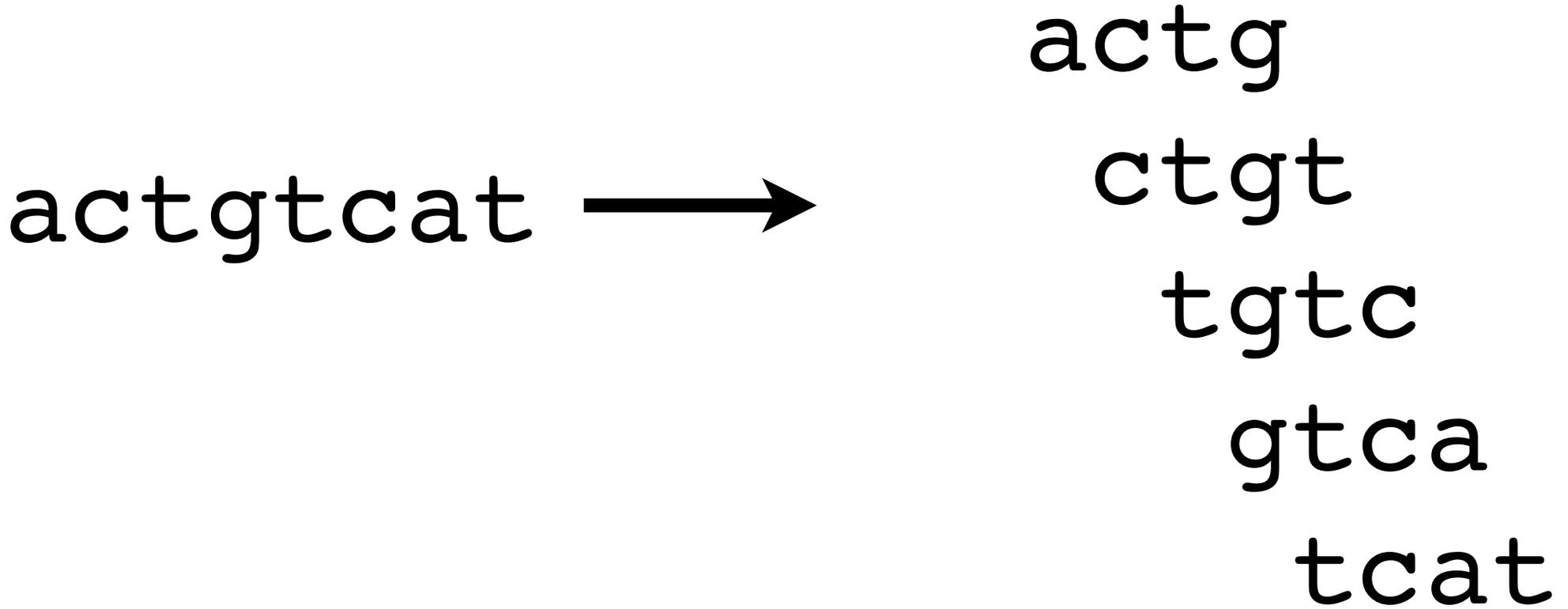
What is a graph?



What is a graph?



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



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 hash table

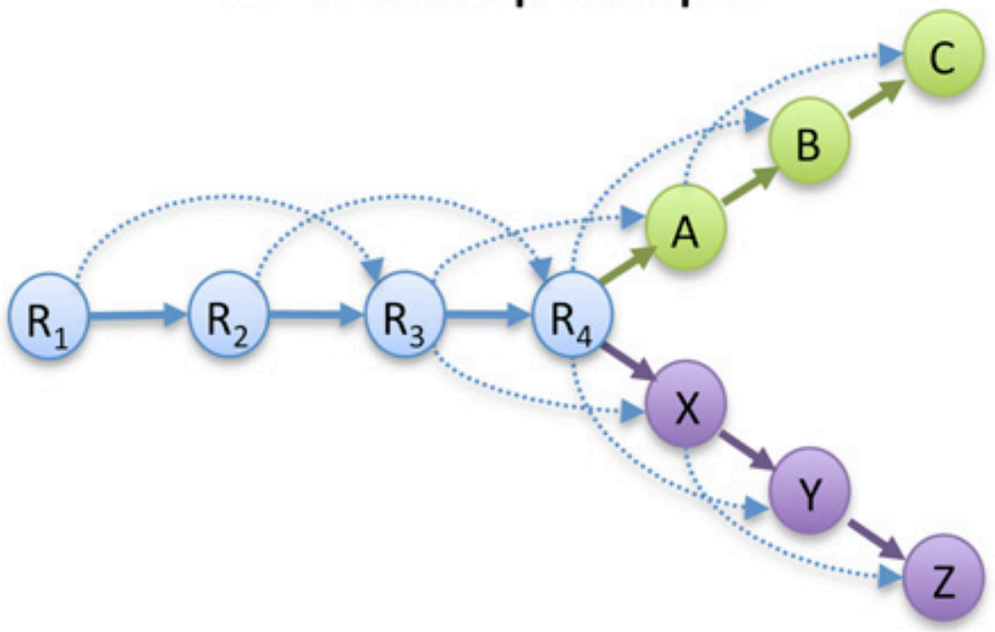
Each node corresponds to a k-mer sequence from the hash table

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

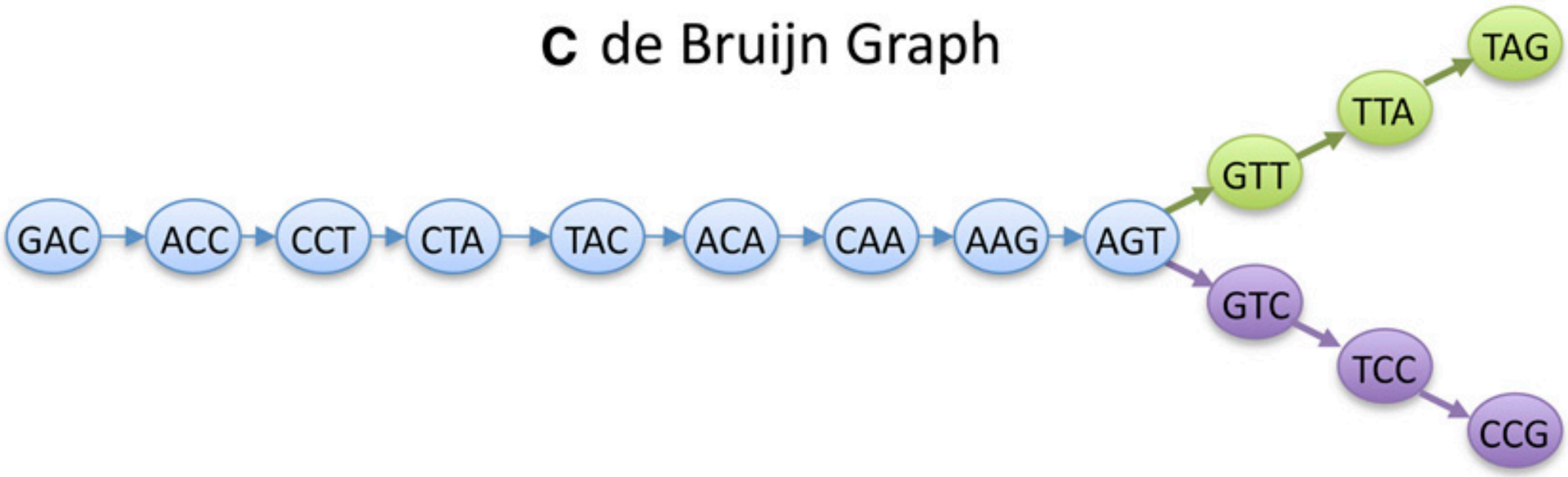
A Read Layout

```
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

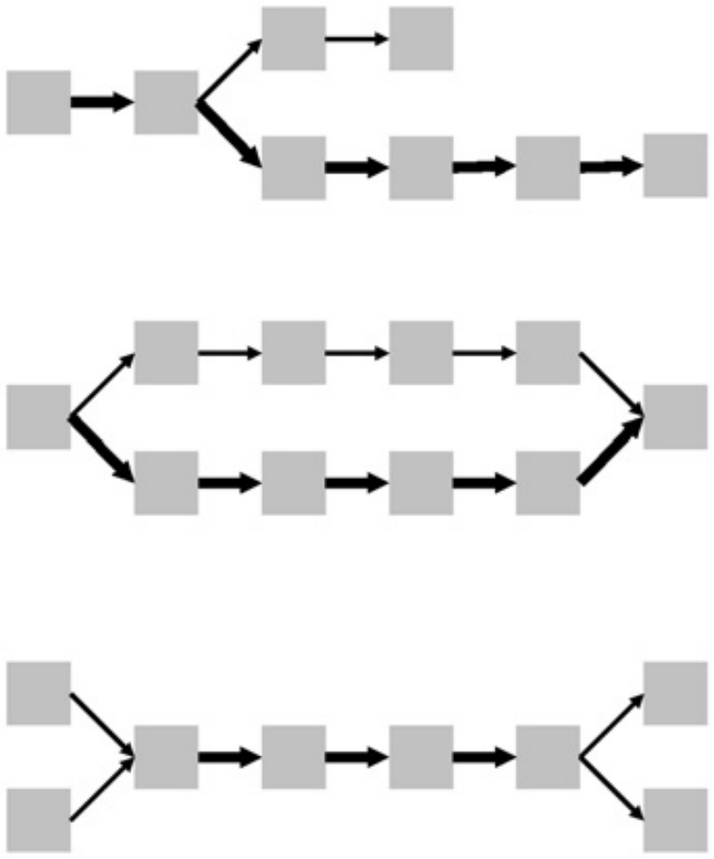


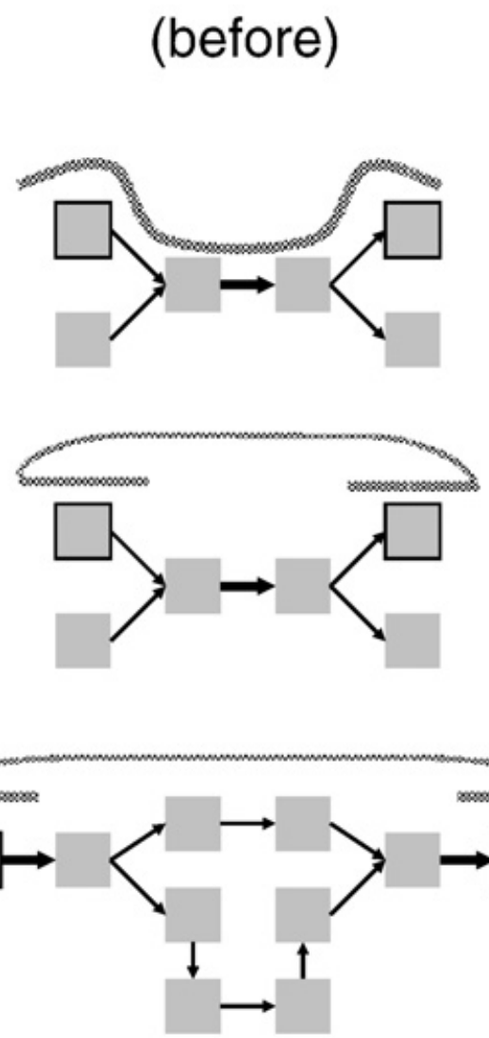
Schatz et al 2010 (dx.doi.org/10.1101/gr.101360.109)

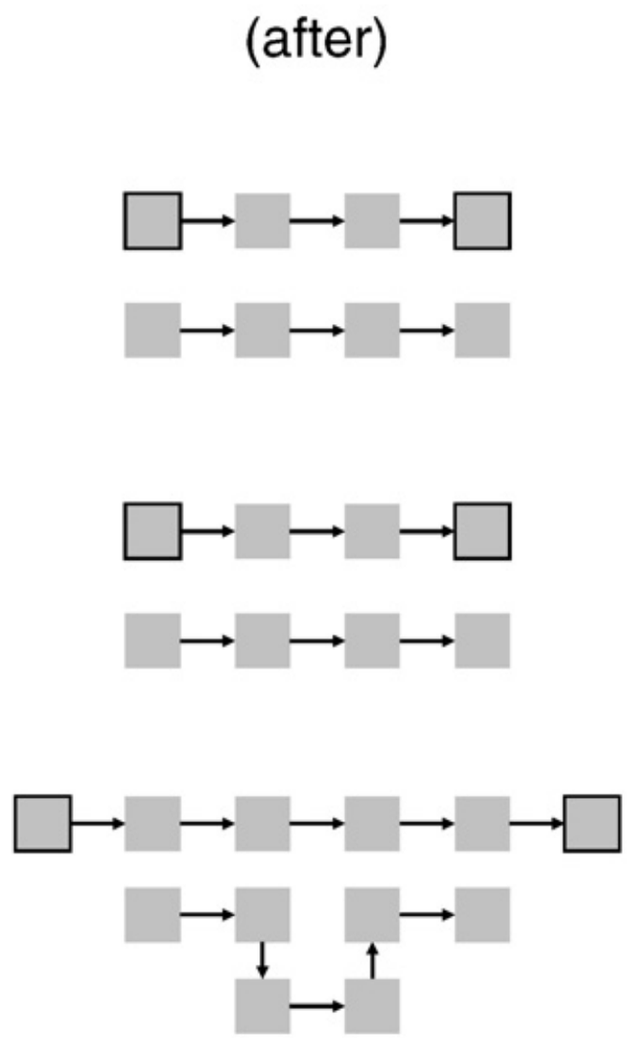
Paths through the de Bruijn graph are assembled sequences

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)







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

de novo sequencing and de Bruijn graph assembly requires very deep sequencing

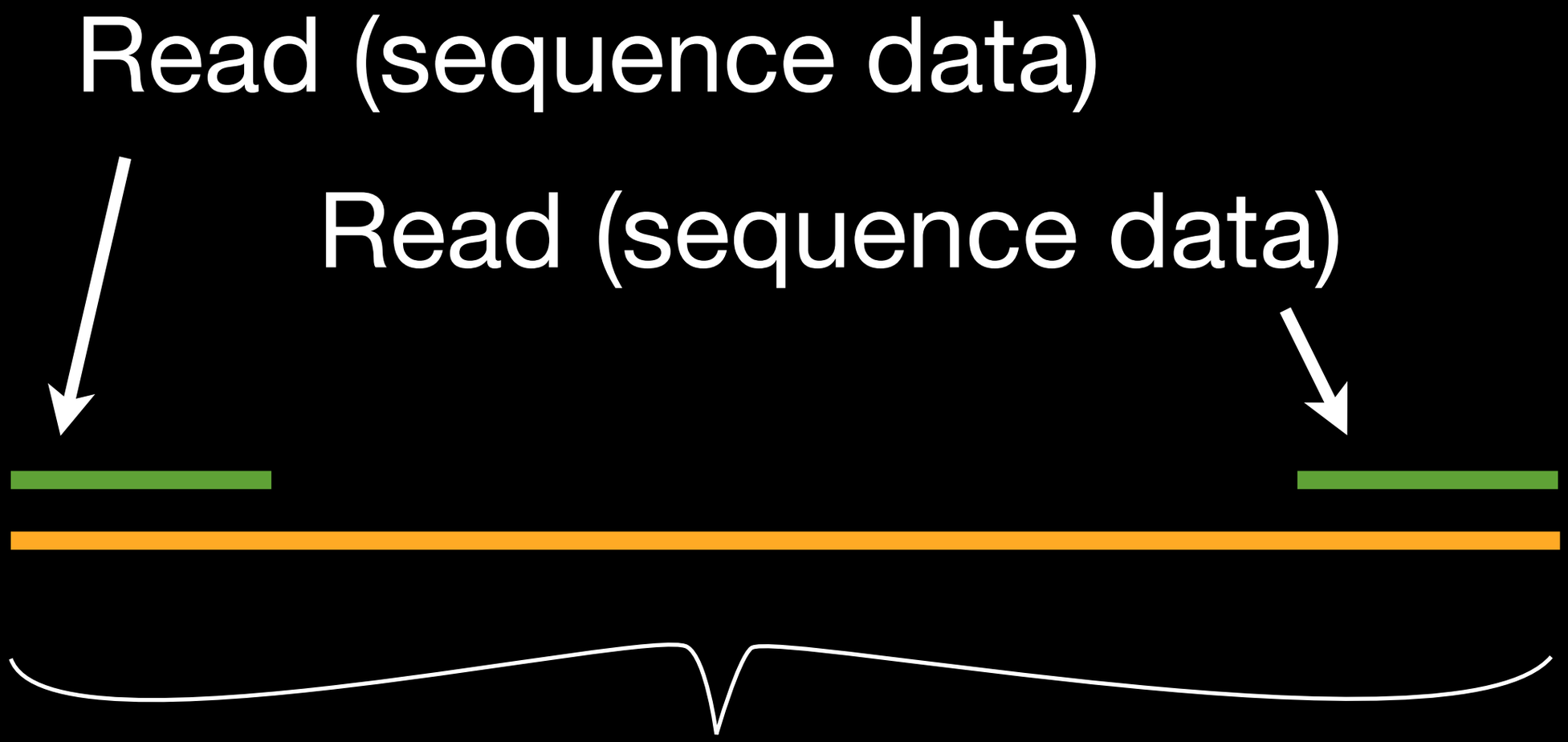
Typically >100 fold coverage

Even then, assemblies are quite fragmented

Can’t resolve repeats longer than the DNA fragments that are sequenced

Paired end sequencing helps by providing structural information longer than read length

Most short read sequencers generate reads from the ends of the DNA molecules



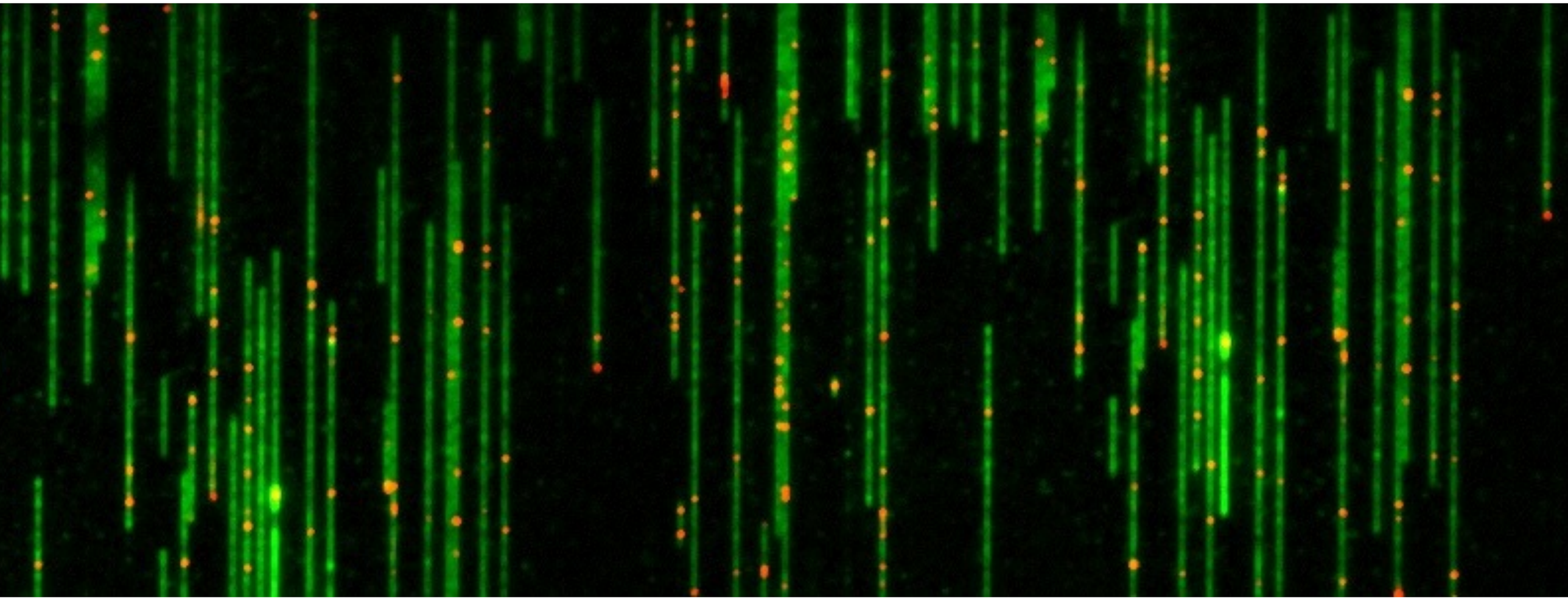
Other tools provide longer range structural information, e.g.:

- Mate pair sequencing provides read pairs that are several kb apart

- Moleculo generates virtual long (~10 kb) reads by preserving information on which reads come from the same fragments

- Restriction site mapping

BioNano and Nabsys both map restriction sites at very large scale



http://www.bionanogenomics.com/technology/irys-technology/

Can be used to stitch together assembly fragments

Annotation

A genome sequence on its own usually isn’t very interesting

You also want to have data about the genome sequence that tells you where genes, regulatory elements, and other features are

A genome sequence on its own usually isn’t very interesting

You also want to have data about the genome sequence that tells you where genes, regulatory elements, and other features are

Annotation based on sequence alone usually has mixed success

Transcriptome and other external data greatly facilitate annotation

Next next generation:
Long reads

Longer reads:

- Make assembly easier

- Have more information (eg improved knowledge of phasing, repeat structure, etc)

Illumina now produces high quality “short” reads on the order of 300 bp

Short read error rates

Table 1 Insertion/deletion and substitution errors on read level for benchtop NGS platforms

Platform	Sequencing kit	Library	Strain	Date of sequencing	Indels per 100 bp	Indels per read	Substitutions per 100 bp	Substitutions per read
GSJ	GSJ Titanium	Nebulization / AMPure XP	Sakai	June 2012	0.4011	1.8351	0.0543	0.2484
MiSeq	2 × 150-bp PE	Nextera	Sakai	June 2012	0.0009	0.0013	0.0921	0.1318
MiSeq	2 × 250-bp PE	Nextera	Sakai	September 2012	0.0009	0.0018	0.0940	0.2033
PGM	100 bp	Bioruptor / Ion Fragment Library	Sakai	July 2011	0.3520	0.3878	0.0929	0.1024
PGM	200 bp	Ion Xpress Plus Fragment	Sakai	July 2012	0.3955	0.6811	0.0303	0.0521
PGM	300 bp	Ion Xpress Plus Fragment	Sakai	August 2012	0.7054	1.4457	0.0861	0.1765
PGM	400 bp[a]	Ion Xpress Plus Fragment	Sakai	November 2012	0.6722	1.8726	0.0790	0.2202

Error rates were calculated by counting indels and substitutions in the mapping against the EHEC Sakai reference sequence for each uniquely mapped read.
[a]Kit was not officially available during time of study.

http://www.nature.com/nbt/journal/v31/n4/pdf/nbt.2522.pdf

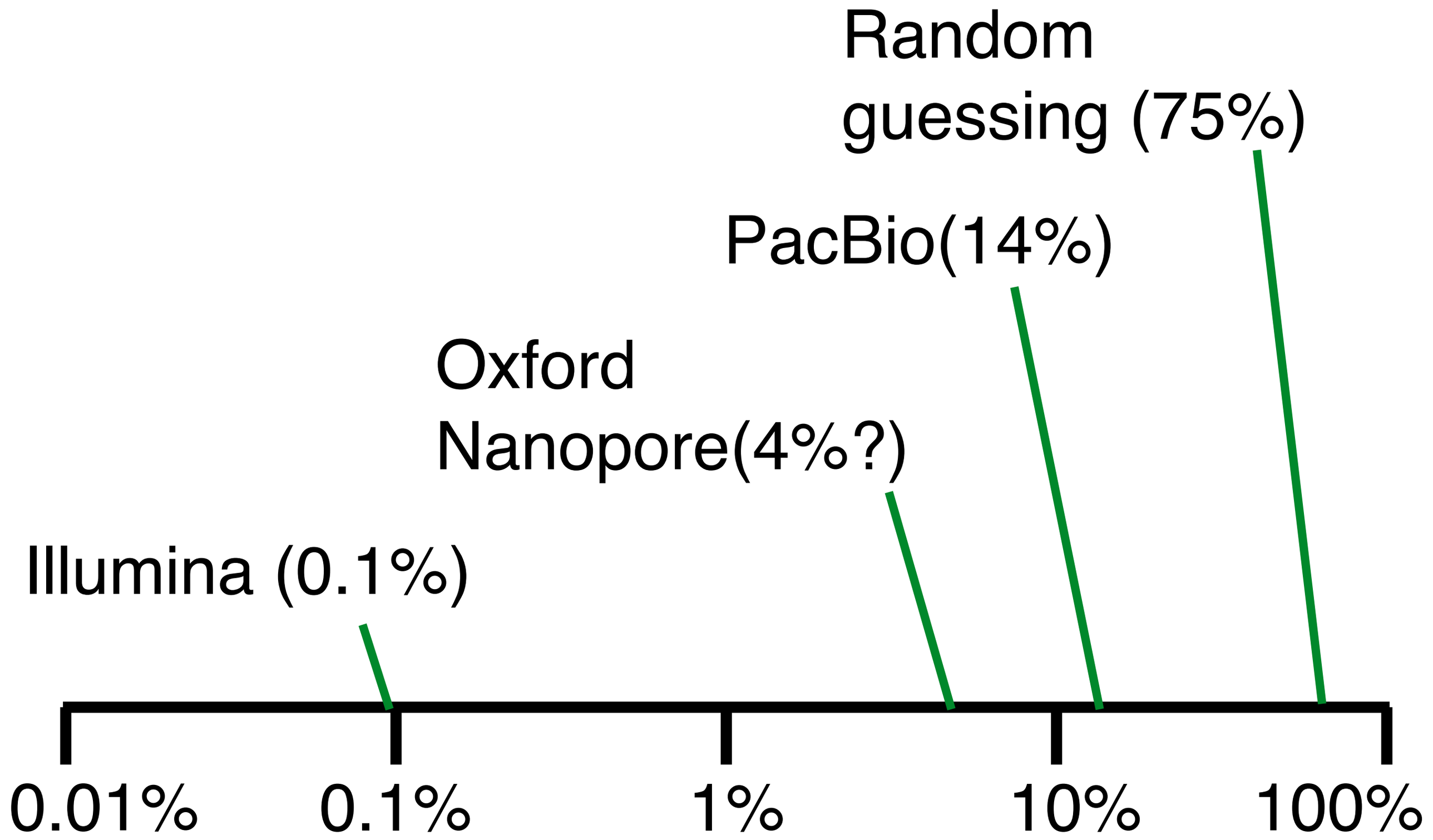
Indel error rates 0.001% to 0.7%

Substitution error rates < 0.1%

Long read platforms now generate reads >10 kb

But the error rate is quite high

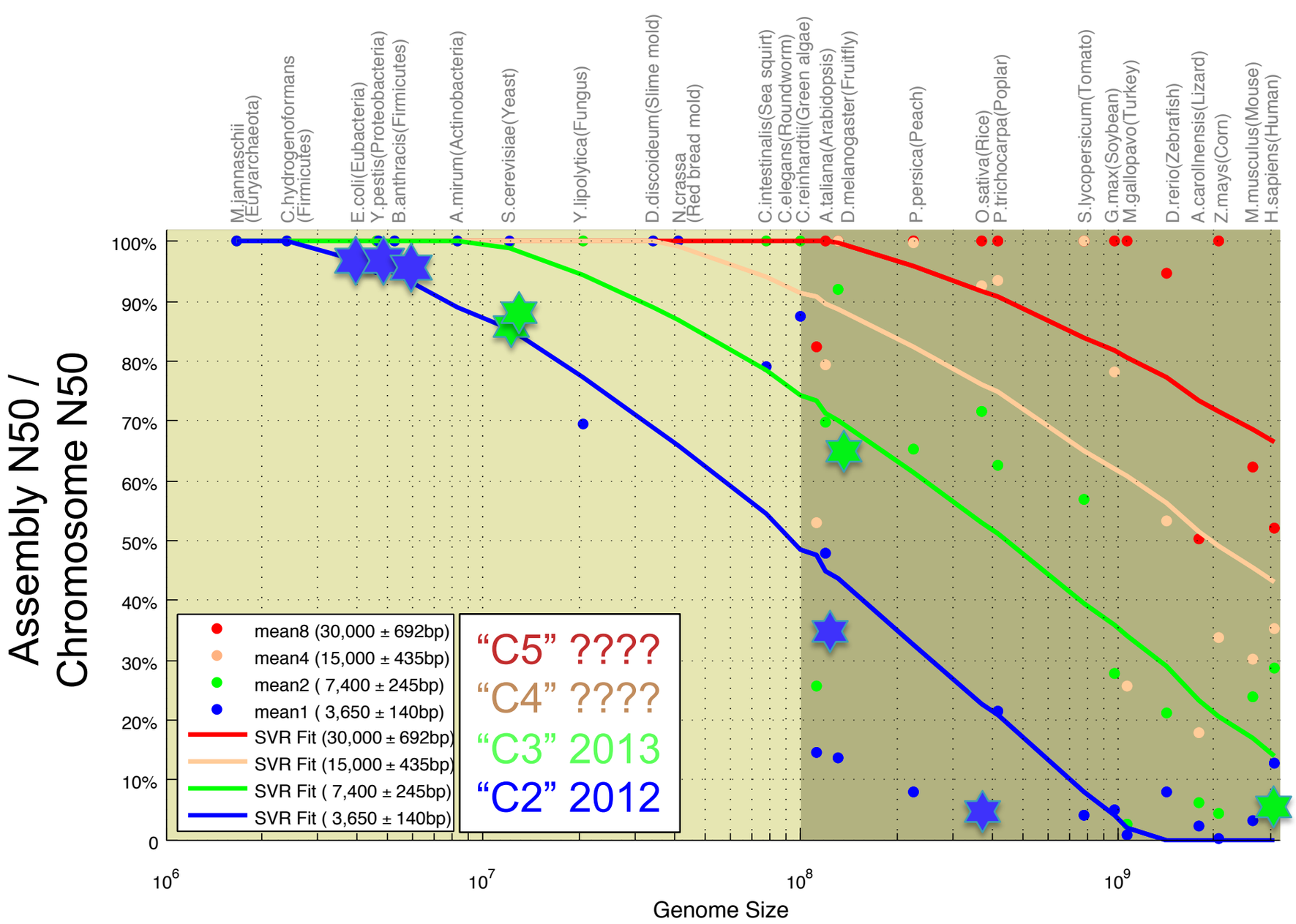
Error rate comparison



How can we use sequence data with such a high error rate?

Use high quality short reads to “fix” low quality long reads prior to assembly (e.g. https://github.com/jgurtowski/ectools)

Assembly Complexity of Long Reads

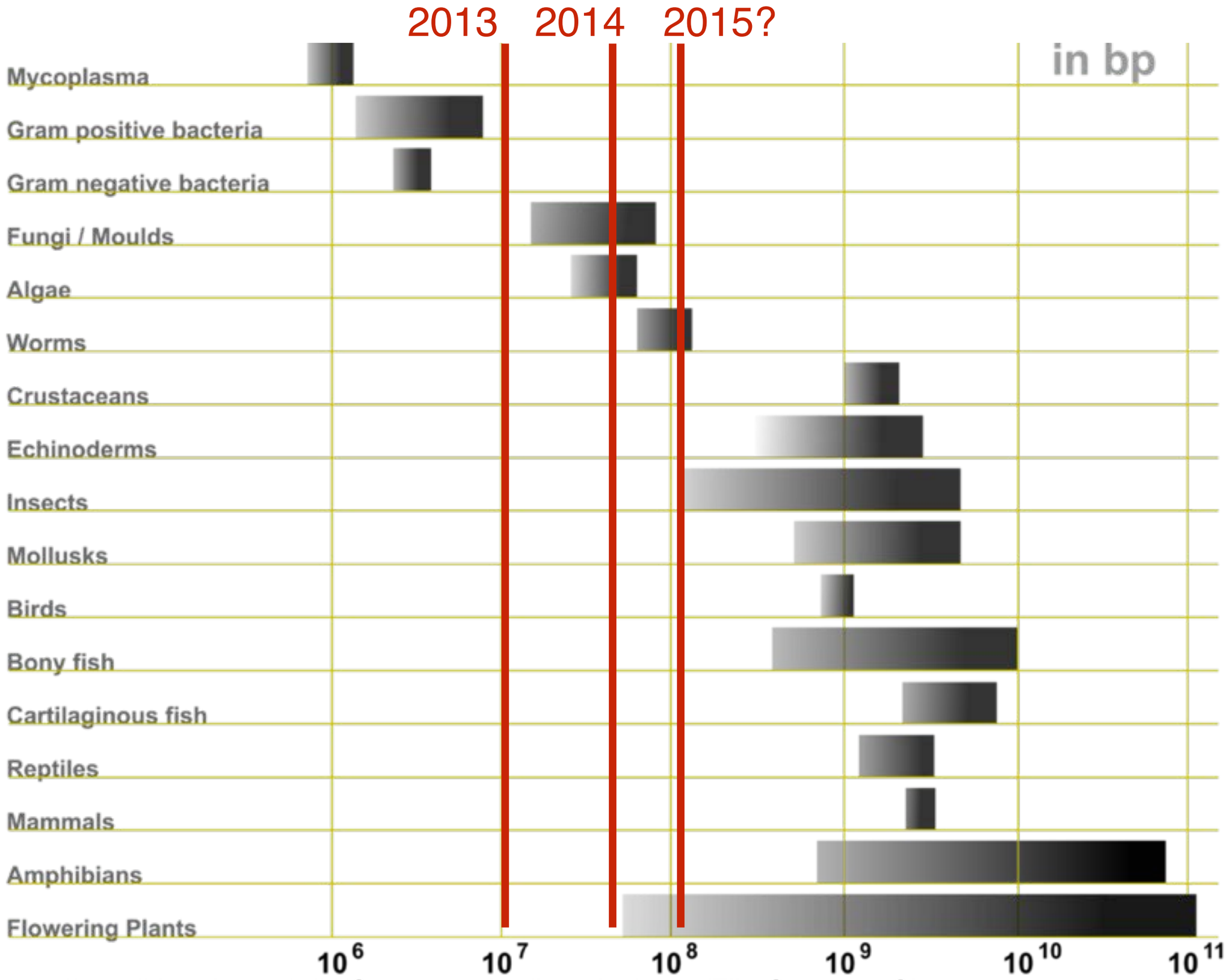


Assembly complexity of long read sequencing

Lee, H*, Gurtowski, J*, Yoo, S, Marcus, S, McCombie, WR, Schatz MC *et al.* (2014) *In preparation*

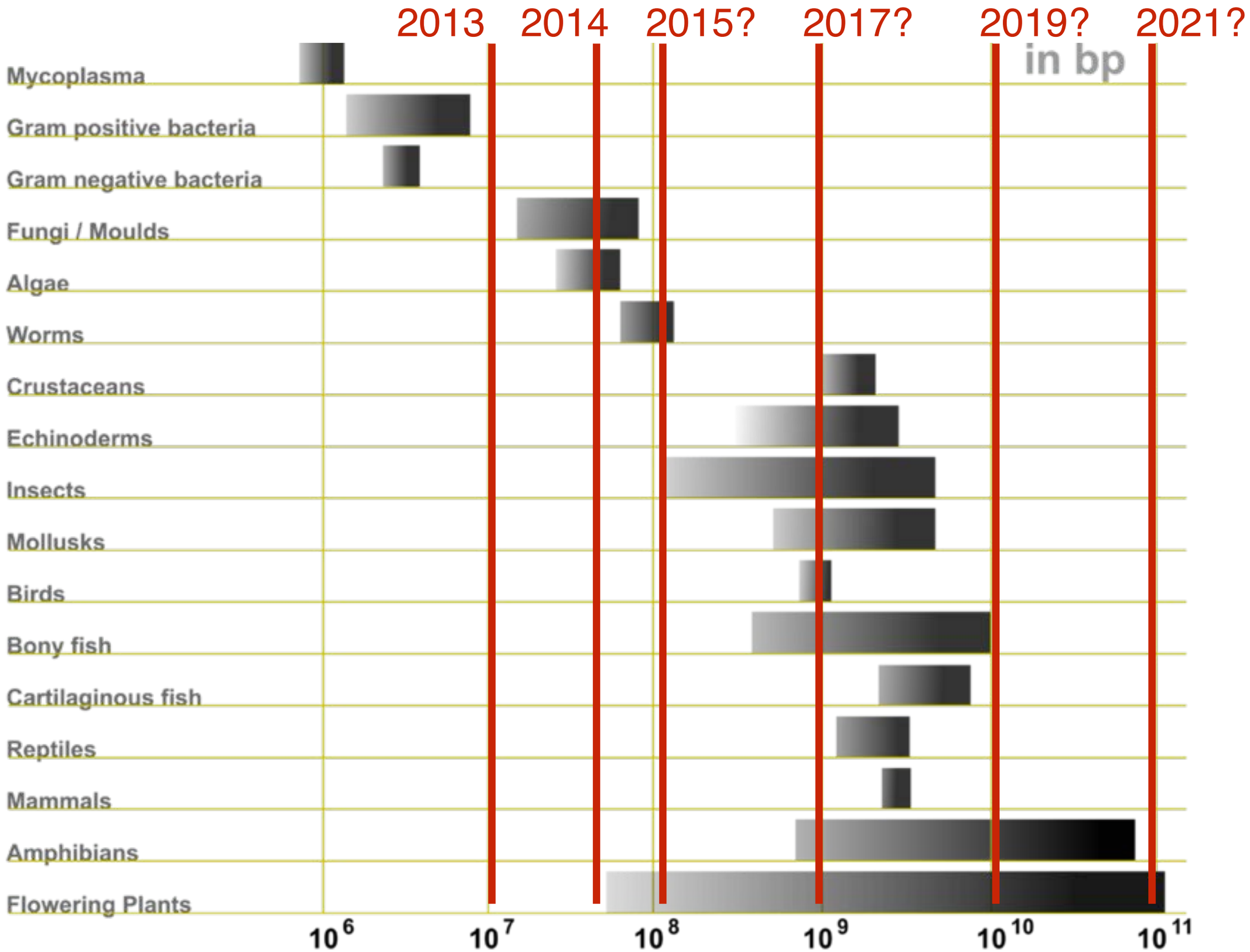
http://schatzlab.cshl.edu/presentations/2014-02-19.Brown.Assembly%20and%20Disease%20Analytics.pdf

Assembly N50 = Chromosome N50



https://en.wikipedia.org/wiki/Genome_size#mediaviewer/File:Genome_Sizes.png

Speculative extrapolation



https://en.wikipedia.org/wiki/Genome_size#mediaviewer/File:Genome_Sizes.png

Summary:

Whole genome de novo assembly

Advantages

Extensive biological information

Low ascertainment bias

Can use in combination with all other enrichment methods

Challenges

Not yet tractable for large genomes

Still expensive for medium-sized genomes

Assembly and annotation still very labor intensive

Typical use case

Now widely used to study molecular evolution of microbes

Targeted application to small numbers of medium-sized genomes

Background reading:

Schatz, M. C., Delcher, A. L. & Salzberg, S. L. Assembly of large genomes using second-generation sequencing. Genome Research 20, 1165–1173 (2010). http://dx.doi.org/10.1101/gr.101360.109

Whole genome (reference mapping)

Mapping is an alternative to assembly

New data are mapped to an existing reference sequence

Requires far less data than *de novo* assembly

Data Preprocessing:

Map to reference

Consensus construction

Annotation

Reference Sequence

Sample Reads

Map

Sample Sequence

Many mapping tools, eg bowtie

Many tools for processing mapped reads, eg samtools

Advantages

Inexpensive

Preprocessing is simpler than for *de novo* assembly

Challenges

Requires a reference sequence from a very closely related taxon

Can be biased by reference (e.g., miss structural differences)

Typical use case

Human and model system resequencing

Background reading:

Consortium, T. 1. G. P. et al. A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073 (2010). http://dx.doi.org/10.1038/nature09534

Transcriptomes

Sample preparation

Some options for preservation

Freeze tissue (-80°C or colder)

RNALater (Ambion), kept cold

Extract RNA in the field

Homogonize in Trizol, keep cold

mRNA isolation - Lots of tissue

Isolate Total RNA with Trizol

Digest DNA

Isolate mRNA

mRNA isolation - Small amount of tissue

- mRNA straight from tissue
  (eg Dynabeads mRNA DIRECT Kit)

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

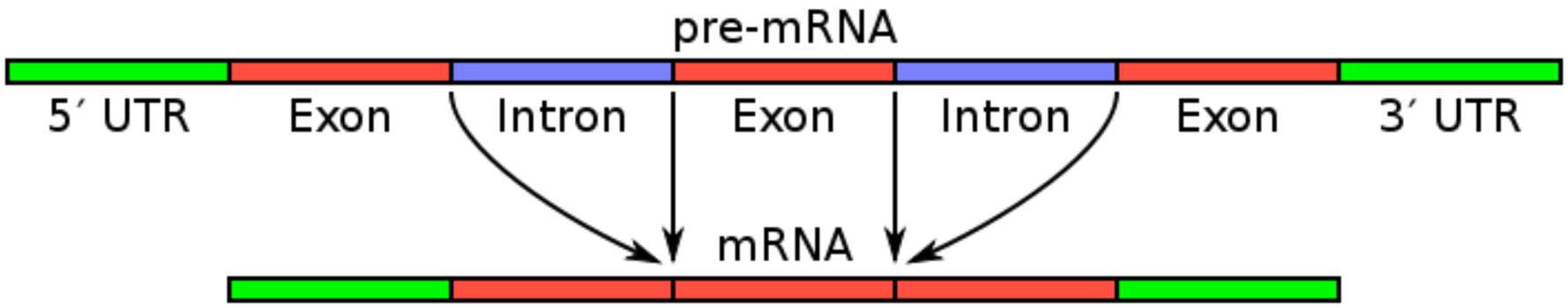
Transcriptome Assembly

Transcriptome assembly has the same challenges as genome assembly…

… and then some.

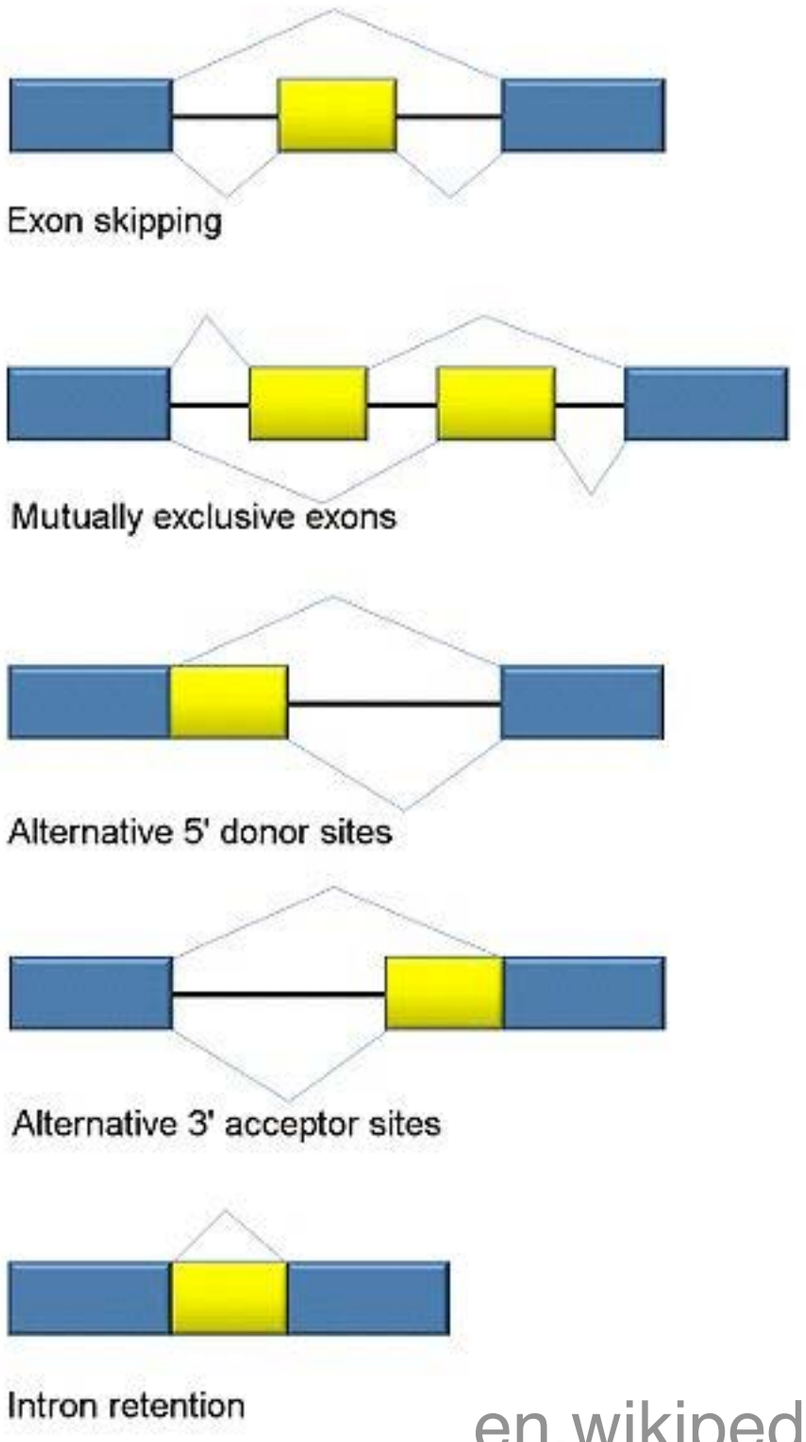
Transcript splicing

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

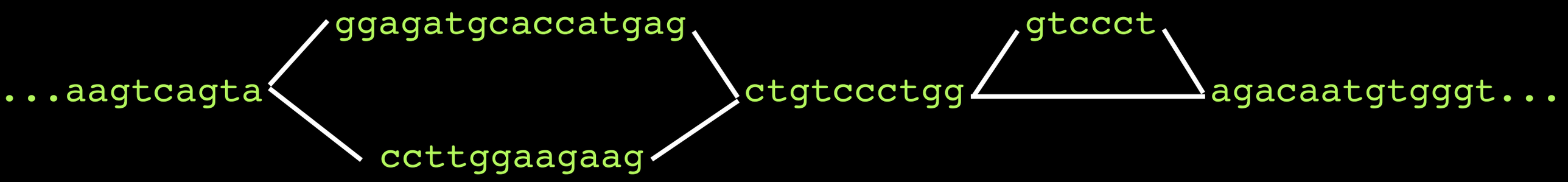
en.wikipedia.org/wiki/File:Alt_splicing_bestiary2.jpg

Assembly results...

Genome

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

Transcript



Splice variants

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

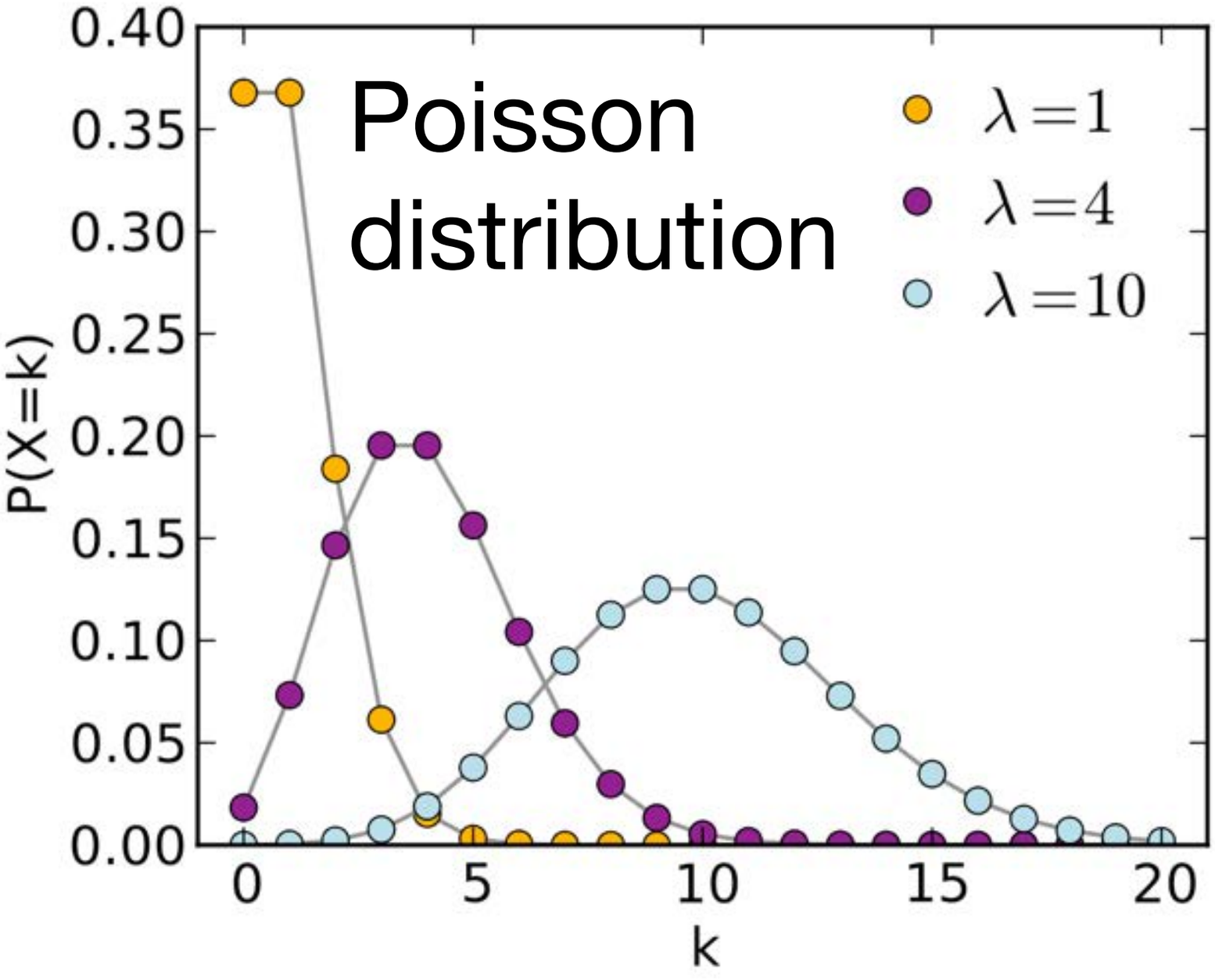
-Deep sequencing captures some “intermediate splice variants”, molecules in the process of being spliced

-Sequencing and assembly errors can be misinterpreted as splice variants

-Data may be insufficient to predict splice variants

It gets worse...

Genomes have uniform depth



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

Assemblers can make assumptions about uniform distribution of sequencing effort

But transcriptomes have non-uniform depth

- Different expression across genes

- Different splice variants within genes

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 accommodate in transcriptomes

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

Agalma

Our automated transcriptome workflow

Why automate?

So that results are reproducible.

Why automate?

So that results can be easily explored and extended.

Why automate?

So that methods can be compared in a controlled setting.

Why automate?

To facilitate methods development by enabling people to focus on particular steps without reinventing everything.

Why reproducing studies is hard:

- Reconstituting the raw data can take weeks
- Methods descriptions are often incomplete
- Manual steps are often subjective
- Code is often not provided

The tool

https://bitbucket.org/caseywdunn/agalma

Bitbucket Repositories Create owner/repository

agalma

caseywdunn Following Share

Clone Fork Compare Pull request

Overview Source Commits Pull requests Issues 1 Downloads 1

Agalma is developed by the Dunn Lab at Brown University.

See TUTORIAL for an example of how to use Agalma with a sample dataset.

Overview of Agalma

Agalma is a set of analysis pipelines for transcriptome assembly (paired-end Illumina data) and phylogenetic analysis. It can import gene predictions from other sources (eg, assembled non-Illumina transcriptomes or gene models from annotated genomes), enabling broadly-sampled "phylogenomic" analyses.

Agalma provides a completely automated analysis workflow that filters and assembles the data under default parameters, and records rich diagnostics. The same goes for alignment, translation, and phylogenetic analysis. You can then evaluate these diagnostics to spot problems and examine the success of your analyses, the quality of the orignal data, and the appropriateness of the default parameters. You can then rerun subsets of the pipelines with optimized parameters as needed.

The workflow is highly optimized to reduce the RAM and computational requirements, as well as the disk space used. It logs detailed stats about computer resource utilization to help you understand what type of computational resources you need to analyze your data and to further optimize your resource utilization.

The main functionality of this workflow is to:

- assess read quality with the FastQC package
- remove clusters in which one or both reads have Illumina adapters (resulting from small inserts)
- remove clusters where one or both reads is of low mean quality
- randomize the sequences in the same order in both pairs to make obtaining random subsets easy
- assemble and annotate rRNA sequences based on a subassembly of the data
- remove clusters in which one or both reads map to rRNA sequences

SSH git@bitbucket.org:caseywdunn/aga

1 Branch	0 Tags	0 Forks	3 Followers

Owner	Casey Dunn
Access level	Public
Type	Git
Last updated	2013-02-19
Created	2013-02-07
Size	1.9 MB (download)

Example analyses

Five siphonophores

https://bitbucket.org/caseywdunn/dunnhowisonzapata2013/

https://bitbucket.org/caseywdunn/dunnhowisonzapata2013/downloads

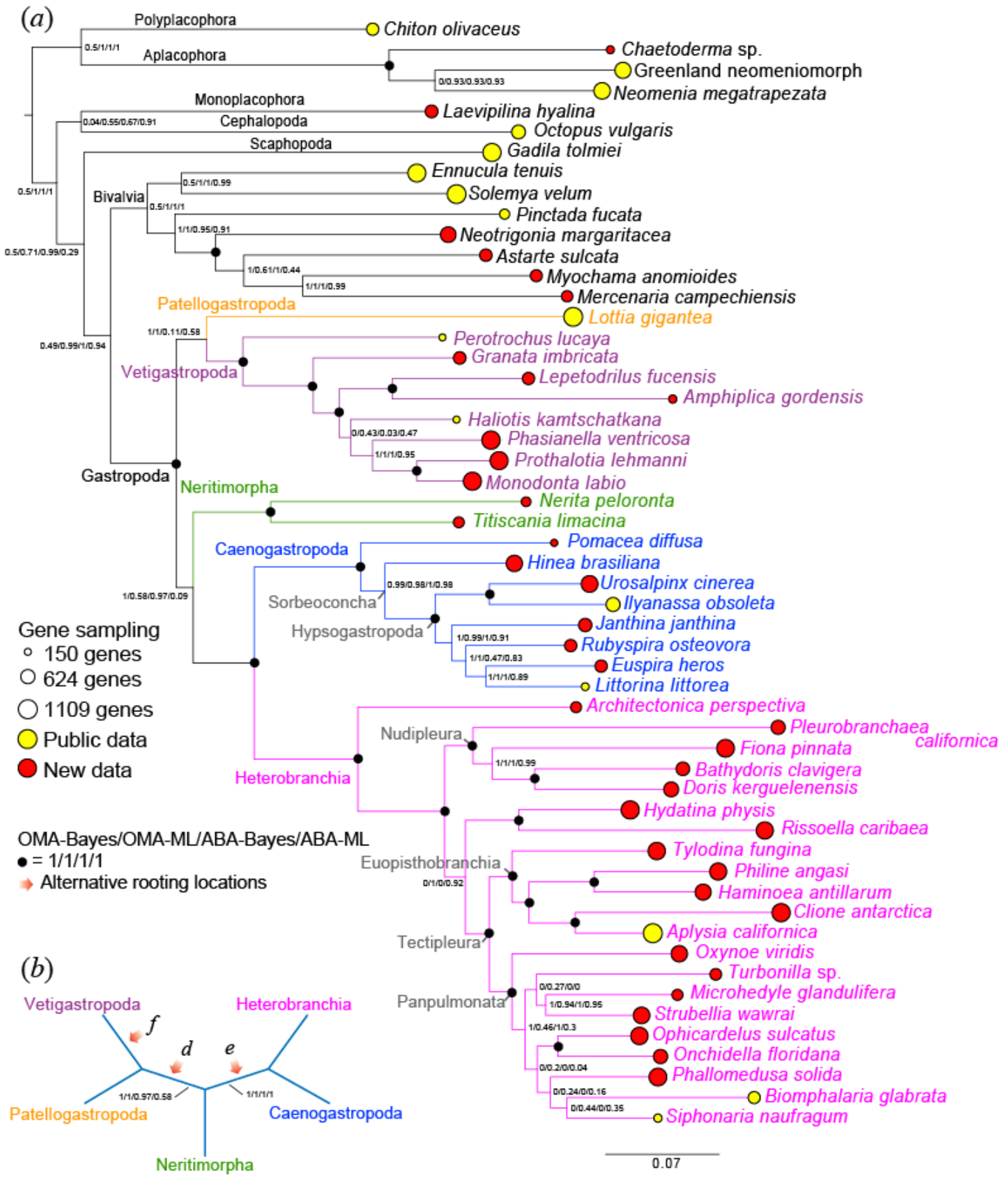
Phylogenomic analyses of deep gastropod relationships reject Orthogastropoda

Felipe Zapata, Nerida G Wilson, Mark Howison, Sónia CS Andrade, Katharina M Jörger, Michael Schrödl, Freya E Goetz, Gonzalo Giribet, Casey W Dunn

bioRxiv beta
THE PREPRINT SERVER FOR BIOLOGY

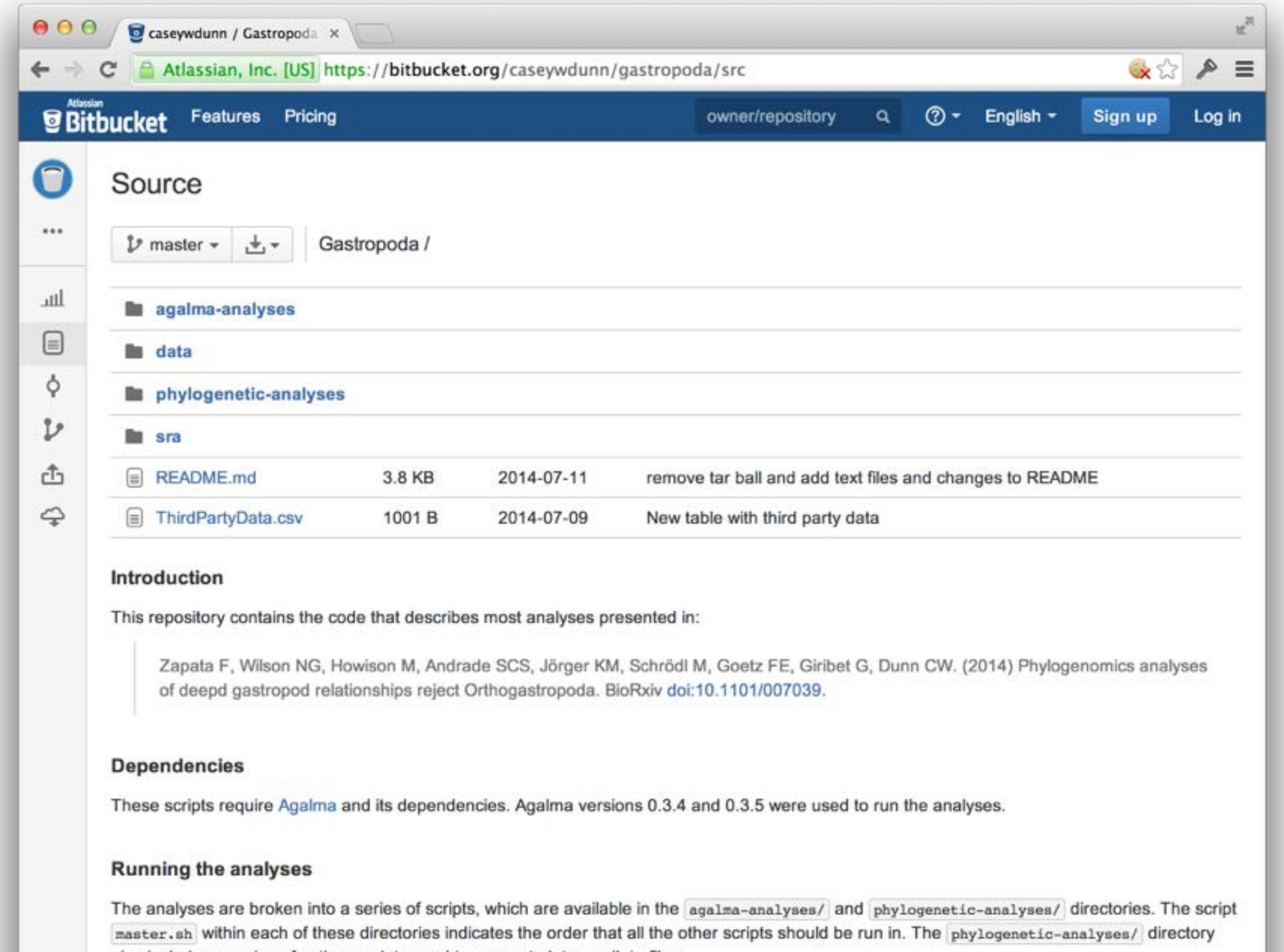
http://dx.doi.org/10.1101/007039

Includes a tree…



Includes a repository of all agalma commands

https://bitbucket.org/caseywdunn/gastropoda/src



Agalma

For each transcriptome:

- Filter adapters/ low quality reads
- Assemble ribosomal RNA
- Remove all ribosomal RNA reads
- Assemble full dataset
- Put assemblies in database

Agalma can also:

- Import reads directly from SRA
- Process externally produced assemblies

Agalma

Across transcriptomes:

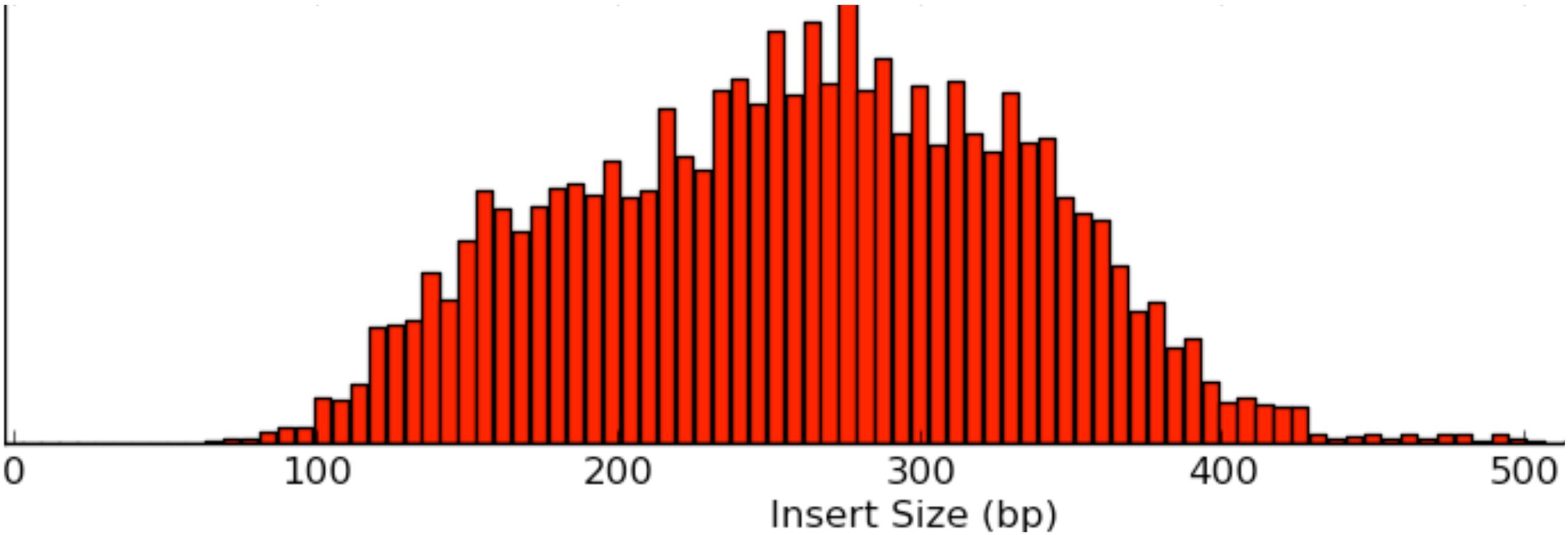
- Identify homologs (all-by-all blastp, mcl)
- Build gene trees (raxml)
- Identify orthologs (based on tree topologies)
- Build preliminary species trees (raxml)

Agalma

- Built on our BioLite framework
- (Relatively) easy to install
- Catalogs specimen data
- Records detailed diagnostics
- Detailed provenance
- Checkpoints (can be restarted)
- Modular
- Generates html reports bundled with output files

Agalma Report

Distribution of library insert sizes



Agalma Report

remove_rrna (Run 8)

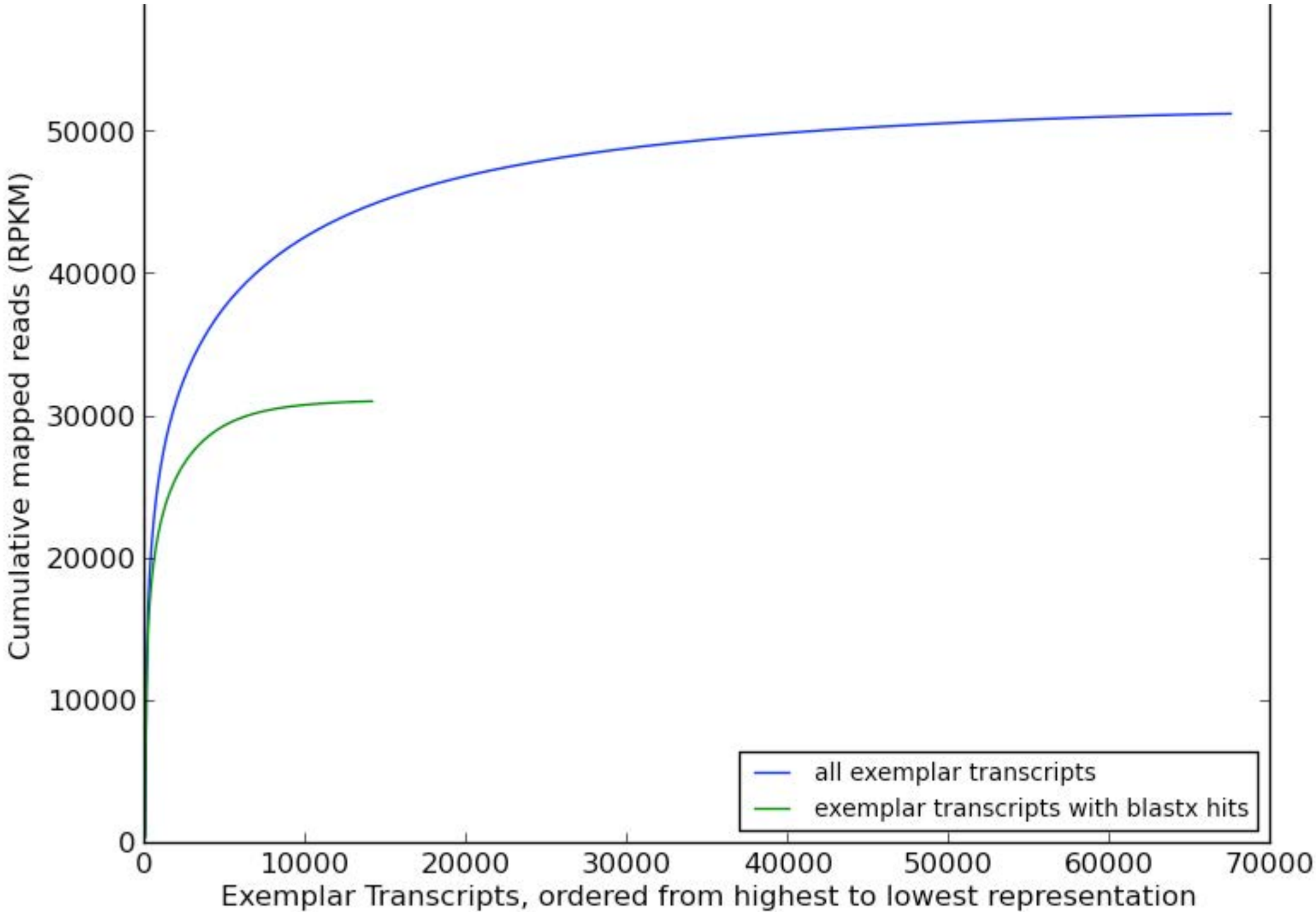
> Assembles and identifies ribosomal RNA (rRNA) sequences, removes read pairs that map to these rRNA sequences, and provides a variety of diagnostics about rRNA. A single exemplar sequence is presented for each type of rRNA that is found, but rRNA read pairs are excluded by mapping to a large set of rRNA transcripts that are derived from multiple assemblies over a range of data subset sizes.

Read pairs examined	49,584,637
Read pairs kept	49,389,302
Percent kept	99.6%

```
>large-nuclear-rRNA|Locus_1000000.230_Transcript_1/1_Confidence_1.000_Length_3648|Run8|HWI-ST625-73-C0JUVACXX-7-AGALMA
TCTCCTTCGACTGATCTCAGTCAGTCGAAAAGTTTTTATTTTGACCTCAGATCAGACAAGACTACCCGCTGAATTTAAGC
ATATTAATAAGCGGAGGAAAAGAAACTAACAAGGATTCCCCTAGTAACGGCGAGTGAAGCGGGAACAGCTCAAACTTAAA
ATCTCCGTTGCTTGCAACGGCGAATTGTAGTCTCGAGAAGCGTTTTCAAGGCGAATGCGCAGTACTTAAGTTGCTTGGAA
CGGCACATCGTAGAGGGTGACAATCCCGTACGTGGTACTGTGCATCGTTCACGATGCGCTTTCTATGAGTCGGGTTGCTT
GGTAATGCAGCCCAAATTGGGAGGTAAACTCCTTCTAAAGCTAAATATTGGCACGAGACCGATAGCGAACAAGTACCGTG
AGGGAAAGATGAAAAGCACTTTGAAAAGAAAGTTAATAGTACGTGAAACCGTTAGGAGGGAAGCGCATGGAATTAGCAAT
GCACTGTCGAGATTCAGACGATCGGTGCTCAGTACGGGCGTCGTACGGATCCGAATGGACCGTTGGCATTCGTCACTTAG
TACTGGTTGTCGCATTTCCCGTCAGTGTGCGTCAACAGGTGTTGGAATCGGGTGATACGCCTCGCAAGAAGGTGGCTGGT
TTCGATCAGTGTTATAGCTTGCGATGTGCTAGCTCGGATCCGACAGAGGTGTCGCAGCACATGCCCTCACGGGCTGGCTT
CTGTTTCCTCAGTCTTGCGTGACCATAGTGGACTGCGTGCAGTGCGCTTGAACTTCGTCGGGCTGTCGGAGGCATGAATG
CACACTATGTGCTTAGGTTGTTGGCGGTCATATGGTTTCATGCGACCCGTCTTGTAACACGGACCAAGGAGTCTAACATG
```

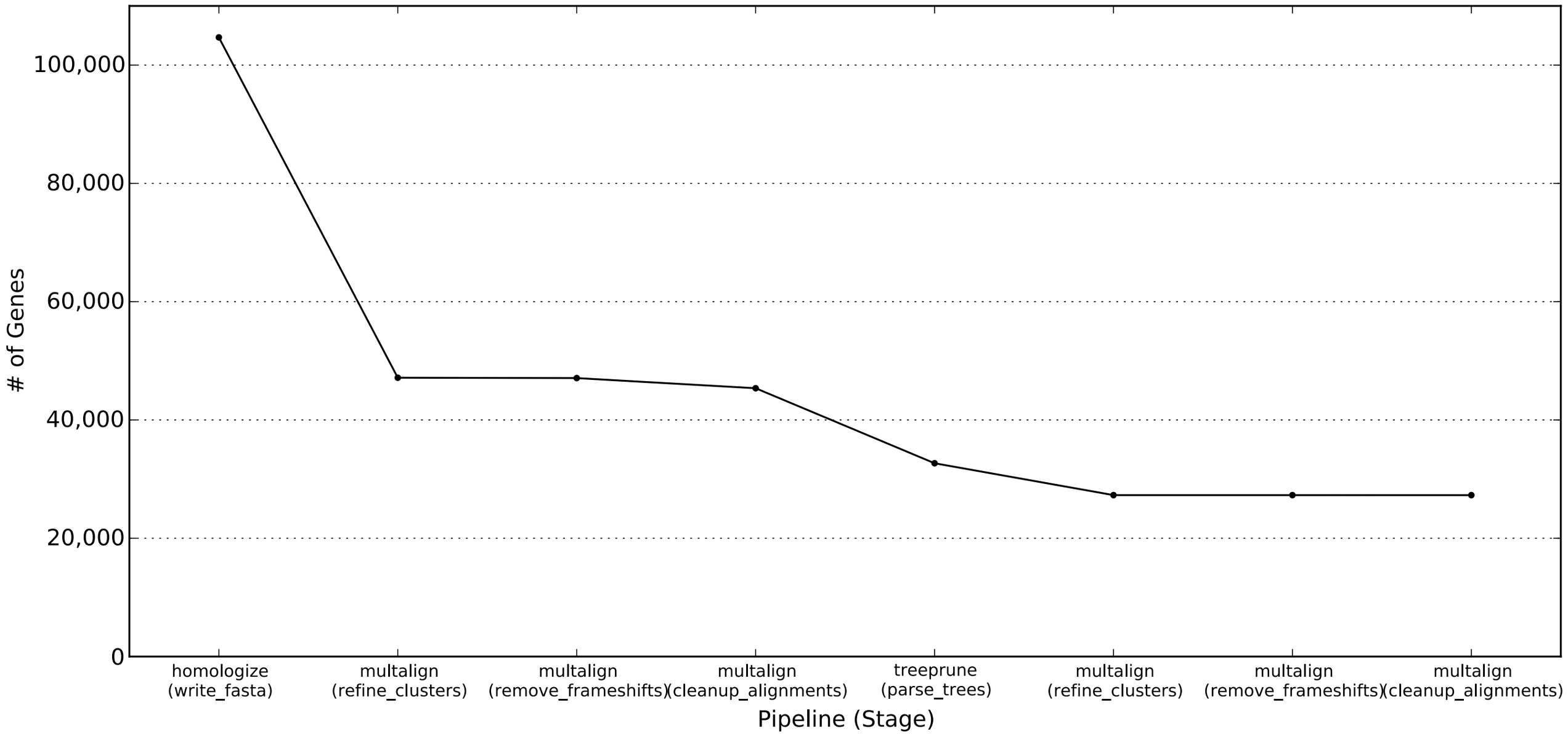
Agalma Report

Distribution of sequencing effort across genes

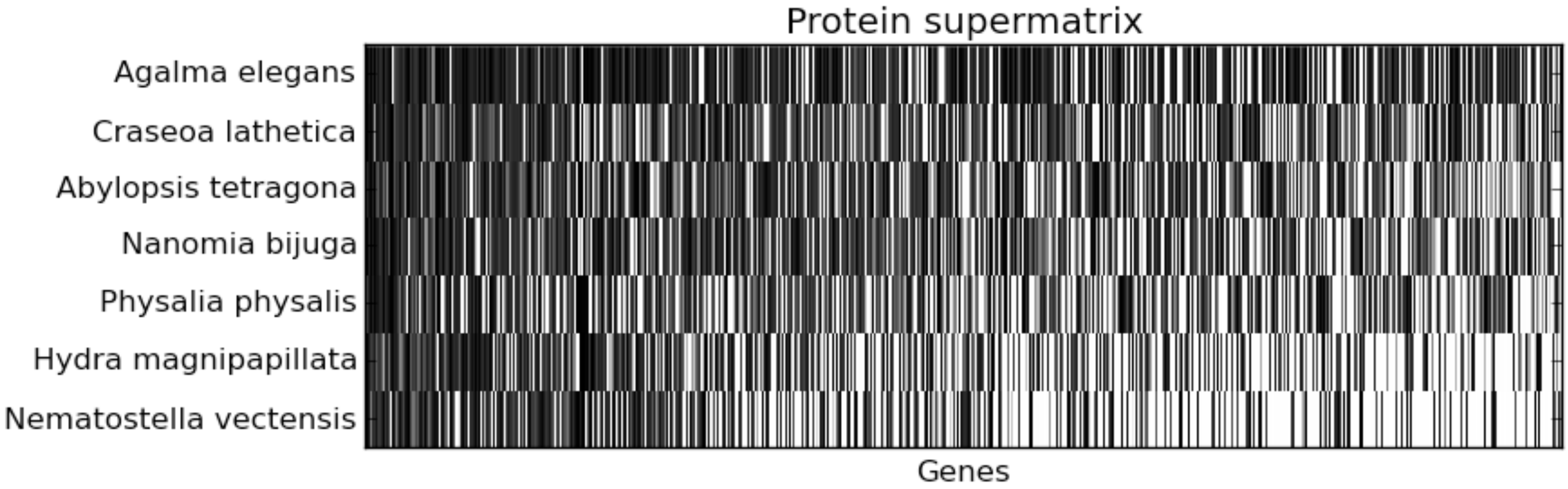


Agalma Report

Reduction in number of genes at each step of matrix construction

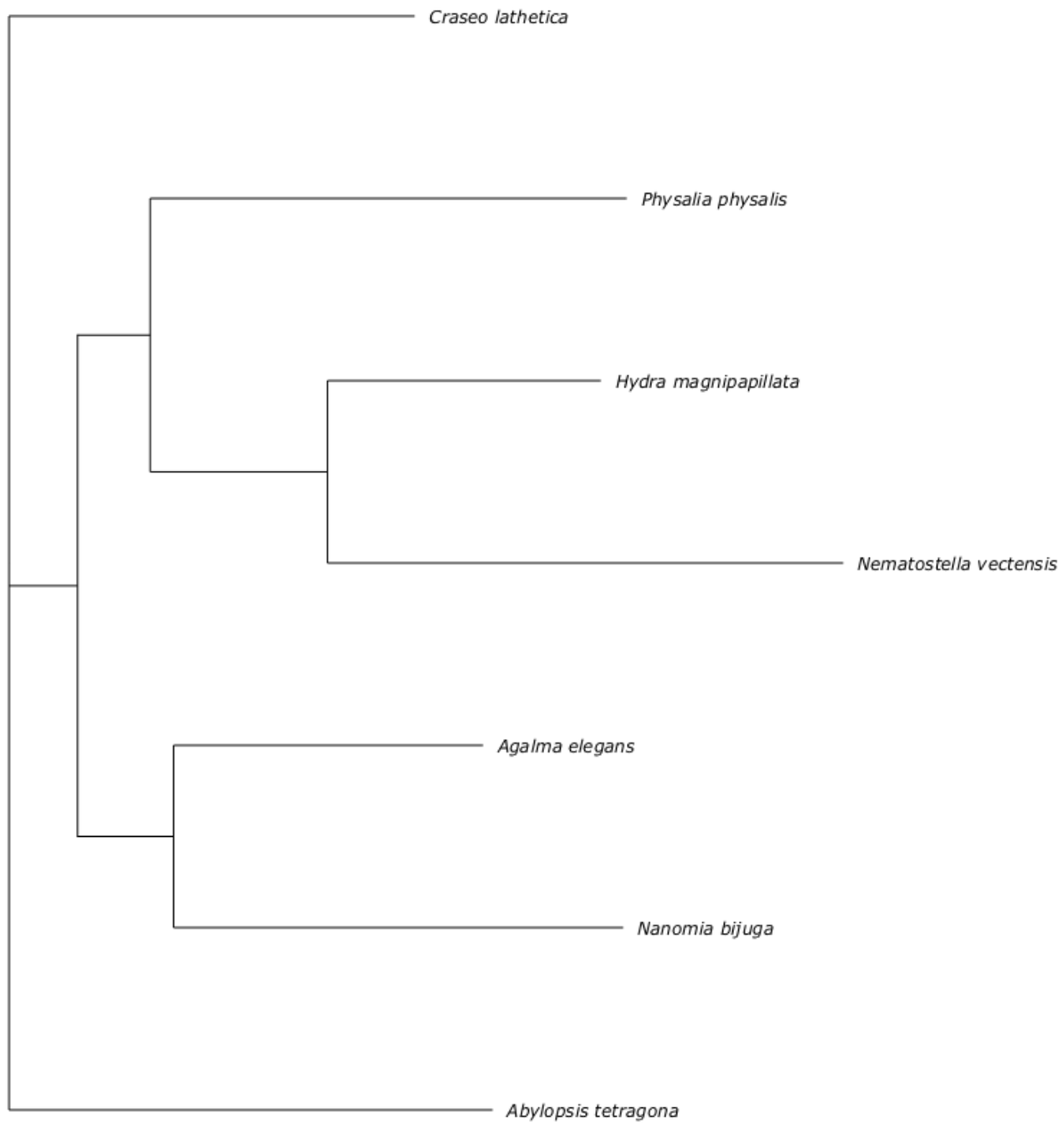


Agalma Report



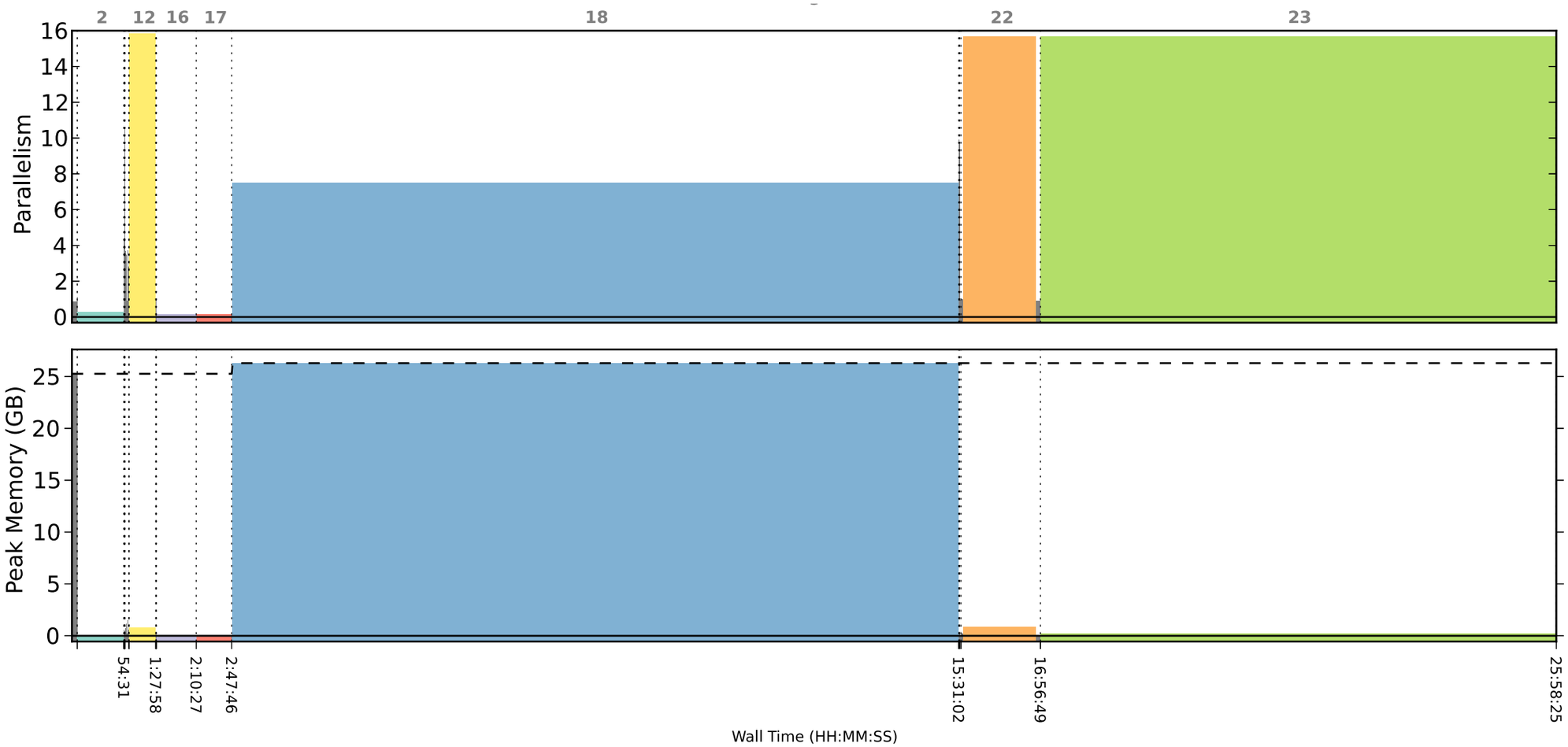
Agalma Report

And preliminary trees...



Agalma Report

Resource utilization



Calls longer than 1% of total runtime

#	Stage / Call	Runtime	User CPU%	System CPU%	Peak Memory
2	sanitize.sanitize.filter_illumina	48:53	26%	72%	1.3 MB
12	remove_rrna.bowtie.bowtie2	27:56	1583%	14%	781.4 MB
16	remove_rrna.exclude_ids.exclude	42:01	15%	84%	96.8 MB
17	assemble.quality_filter.filter_illumina	37:18	16%	83%	1.3 MB
18	assemble.trinity.butterfly	12:42:27	750%	118%	26.3 GB
22	postassemble.coverage.bowtie2	1:16:11	1569%	10%	898.0 MB
23	postassemble.nr_annotate.blastx	9:01:35	1569%	3%	222.7 MB

Downstream from Agalma

Think of Agalma as a tool for generating alignments of homologous genes. It is up to you to figure out the appropriate phylogenetic analyses to resolve the relationships between species.

Summary:

Transcriptomes

Advantages

Can be readily applied across a broad diversity of species

Very cost effective way to collect protein coding regions

Very effective for gene discovery

Select genes after sequencing

Challenges

Requires high quality RNA

Assembly can be tricky

Ascertainment bias - only gives expressed genes

Typical use case

Phylogenetic analyses with broad taxon sampling

Evolutionary development, physiology, ecology studies

Background reading:

Dunn, C. W., Howison, M. & Zapata, F. Agalma: an automated phylogenomics workflow. BMC Bioinformatics 14, 330 (2013). http://dx.doi.org/10.1186/1471-2105-14-330

Felipe Zapata, Nerida G Wilson, Mark Howison, Sónia CS Andrade, Katharina M Jörger, Michael Schrödl, Freya E Goetz, Gonzalo Giribet, Casey W Dunn. Phylogenomic analyses of deep gastropod relationships reject Orthogastropoda. Biorxiv. http://dx.doi.org/10.1101/007039

RADseq

Data acquisition

Digest genomic DNA with one or more restriction enzymes

Size select restriction fragments

Sequence fragments

Data preprocessing

Consolidate redundant reads

Identify homologous reads across samples

Advantages

Inexpensive

Sequence tags are broadly sampled across the genome

Relatively simple preprocessing

Challenges

Can only compare data across closely related taxa

Little control over which particular regions are sequenced

Size selection can be tricky

Typical use case

Population genetics within species

Background reading:

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S. & Hoekstra, H. E. Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. PLoS ONE 7, e37135 (2012). http://dx.doi.org/10.1371/journal.pone.0037135

Targeted enrichment

Data acquisition

Select genes

Design capture probes that hybridize to genes

Use probes to pull out selected genes from fragmented DNA

Data preprocessing

(Select genes)

Assemble reads into gene sequences

Annotate selected genes

Advantages

Inexpensive

Strong control over which regions are sequenced

Greatly simplified assembly and annotation

Works great on poorly preserved specimens

Challenges

Need to know what genes to sequence before you start

Ascertainment biases

Difficult to integrate data across studies with different genes

Need to optimize for different clades

Typical use case

Phylogenetic analyses with broad taxon sampling

Background reading:

Lemmon, A. R., Emme, S. A. & Lemmon, E. M. Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics. Syst. Biol. 61, 727–744 (2012). http://dx.doi.org/110.1093/sysbio/sys049

Directed PCR

Data acquisition

Select genes

Design primer pairs that hybridize to genes

Amplify and sequence genes

Data preprocessing

(Select genes)

Assemble reads into gene sequences

Advantages

Easy to integrate with existing data

Strong control over which regions are sequenced

Greatly simplified assembly and annotation

Challenges

Need to know what genes to sequence before you start

Very labor intensive for more than a few genes

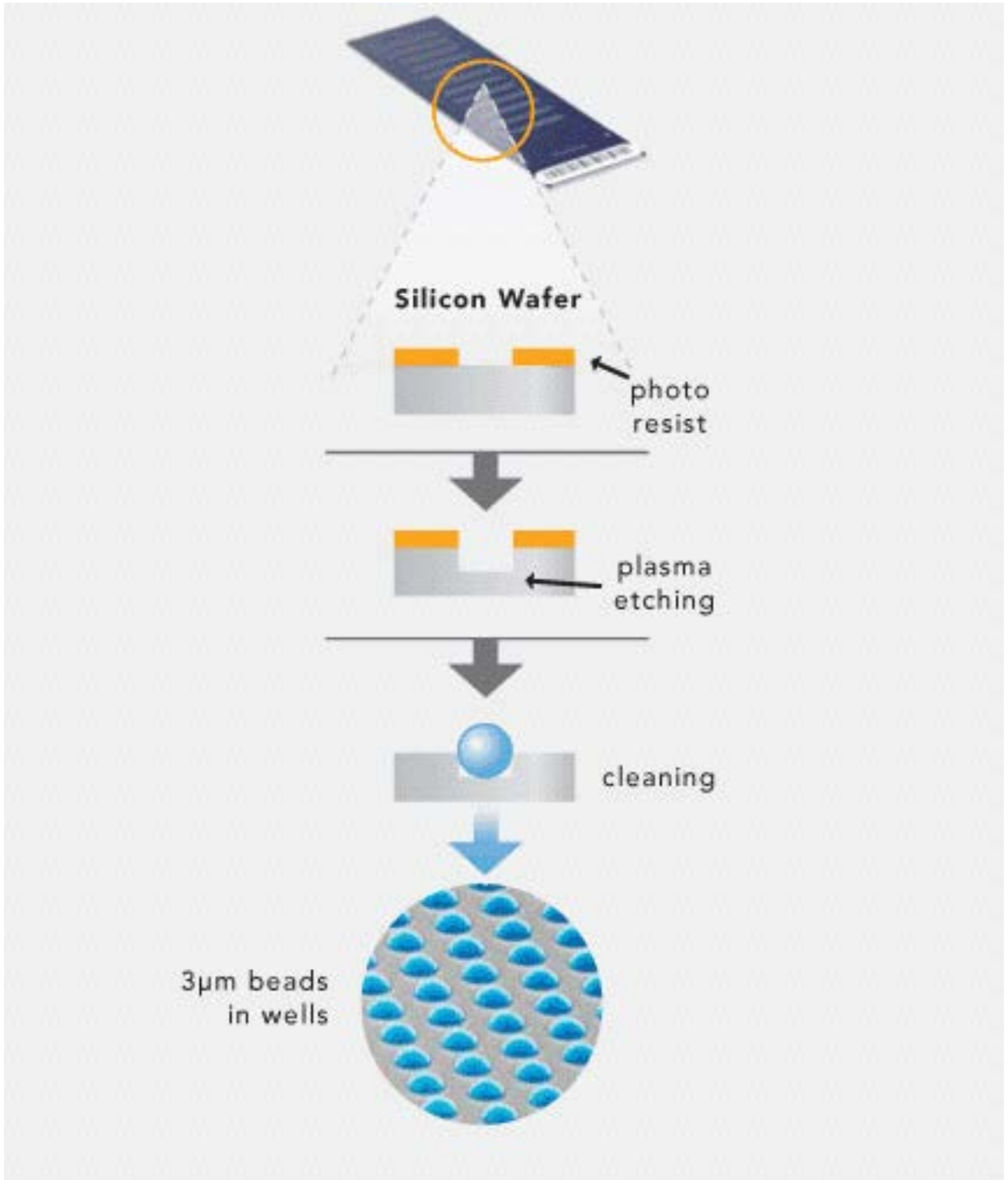
Need to optimize for different clades

Typical use case

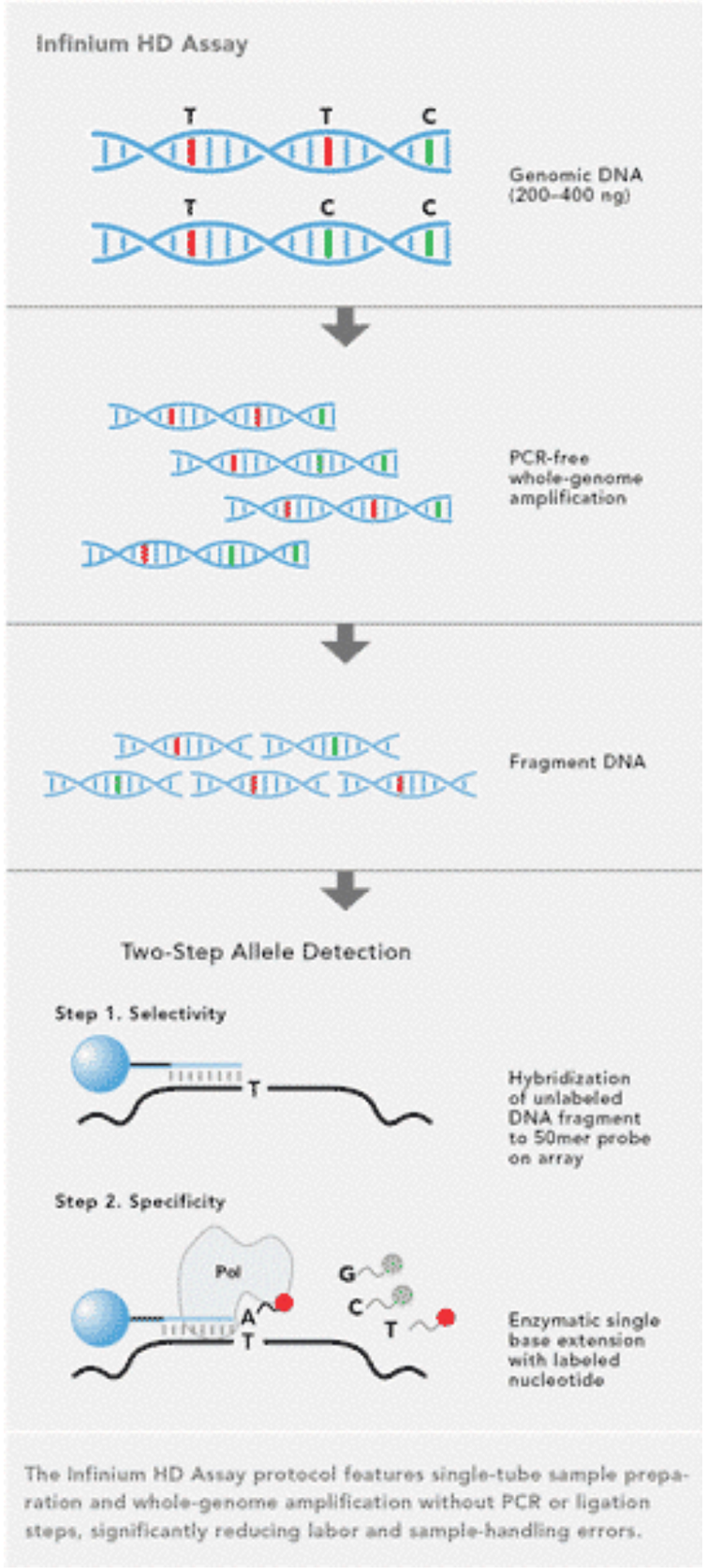
"Phylodiversity" studies, i.e. small number of genes from many taxa

Other enrichment tools...

SNP chips



Illumina multi-sample array formats



The Infinium HD Assay protocol features single-tube sample preparation and whole-genome amplification without PCR or ligation steps, significantly reducing labor and sample-handling errors.

http://www.illumina.com/technology/beadarray-technology/infinium-hd-assay.ilmn

23andMe – Genetic Testin

https://www.23andme.com

The largest DNA ancestry service in the world

sign in | register kit | 0

23andMe

welcome | ancestry | how it works | buy | search | help

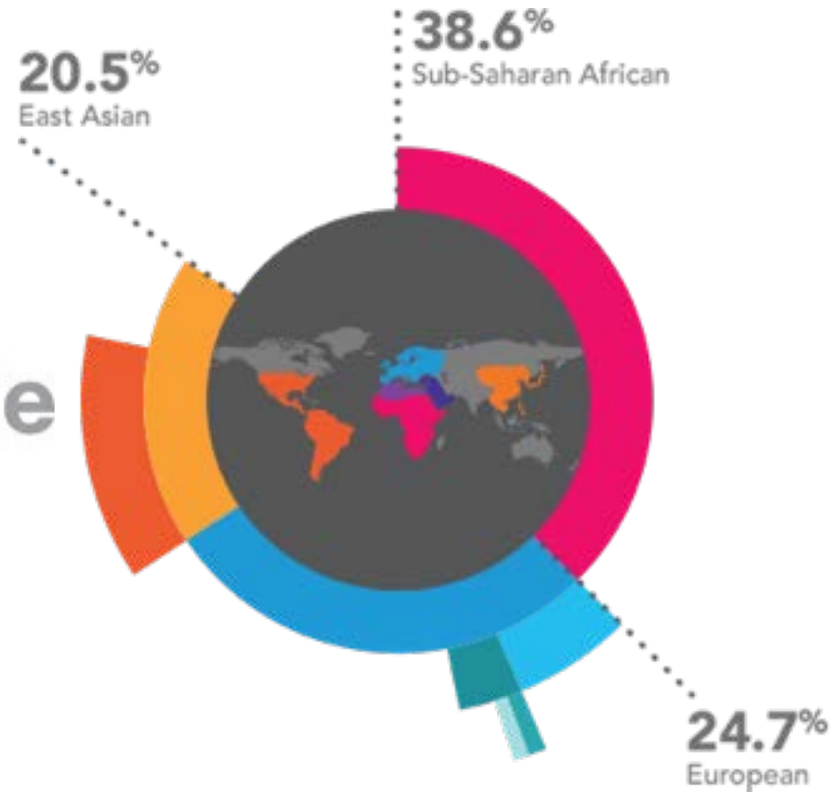
23andMe provides ancestry-related genetic reports and uninterpreted raw genetic data. We no longer offer our health-related genetic reports. If you are a current customer please go to the health page for more information. Close alert.

Find out what your DNA says about you and your family.

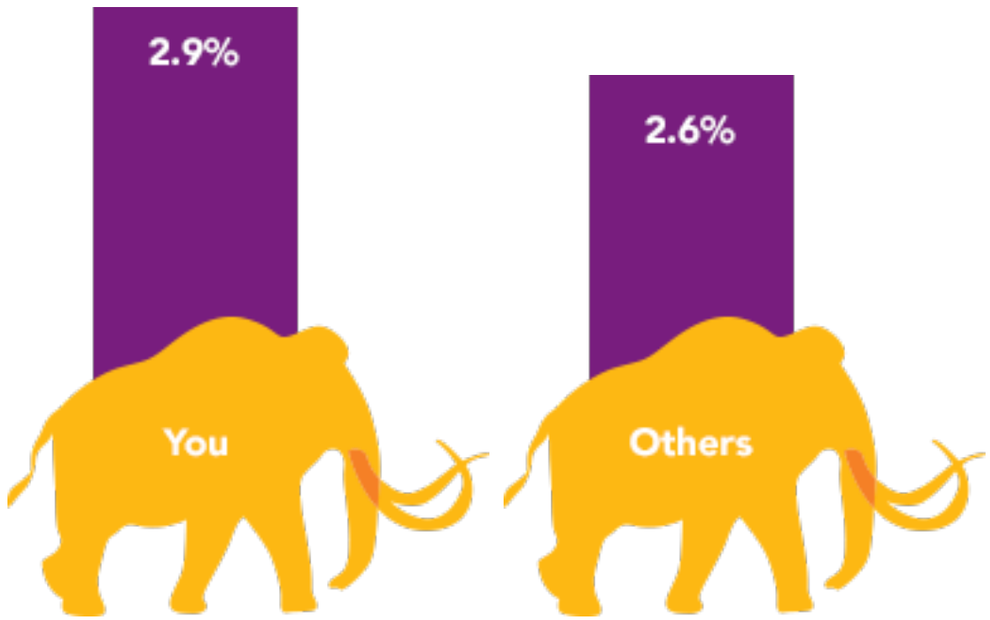
- Learn what percent of your DNA is from populations around the world
- Contact your DNA relatives across continents or across the street
- Build your family tree and enhance your experience with relatives

order now $99

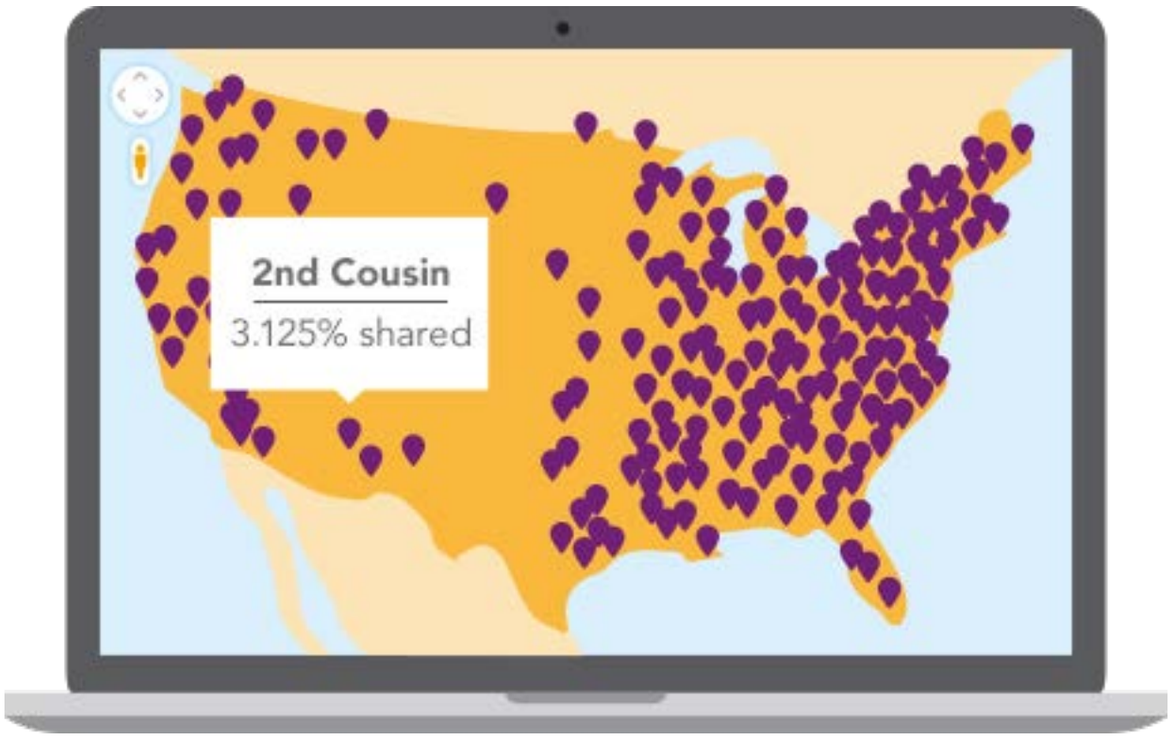
23andMe



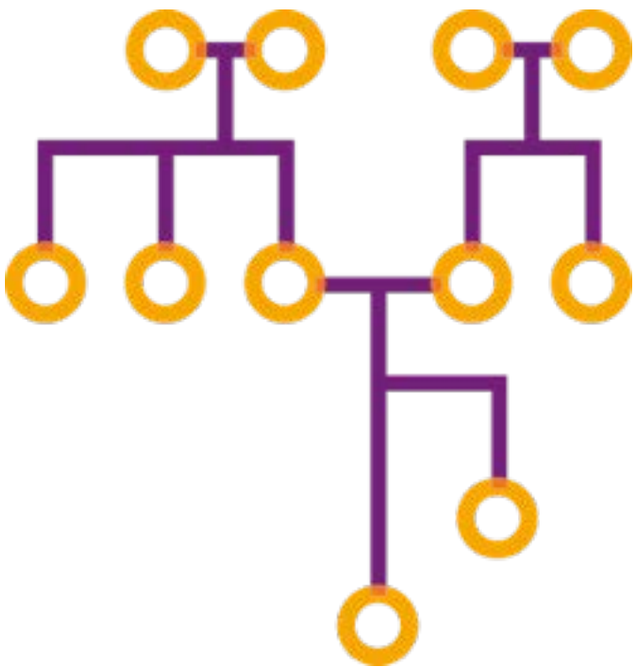
Neanderthal DNA lives on in us.



Find relatives



Build your family tree and enhance your experience.



https://www.23andme.com/ancestry/

Advantages

Very inexpensive

Simple data preprocessing

Challenges

Extremely expensive initial investment

Only works for very closely related taxa

Typical use case

Human and model systems

(an inexpensive alternative to reference mapping)

Part II:

Using trees to study genome function

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

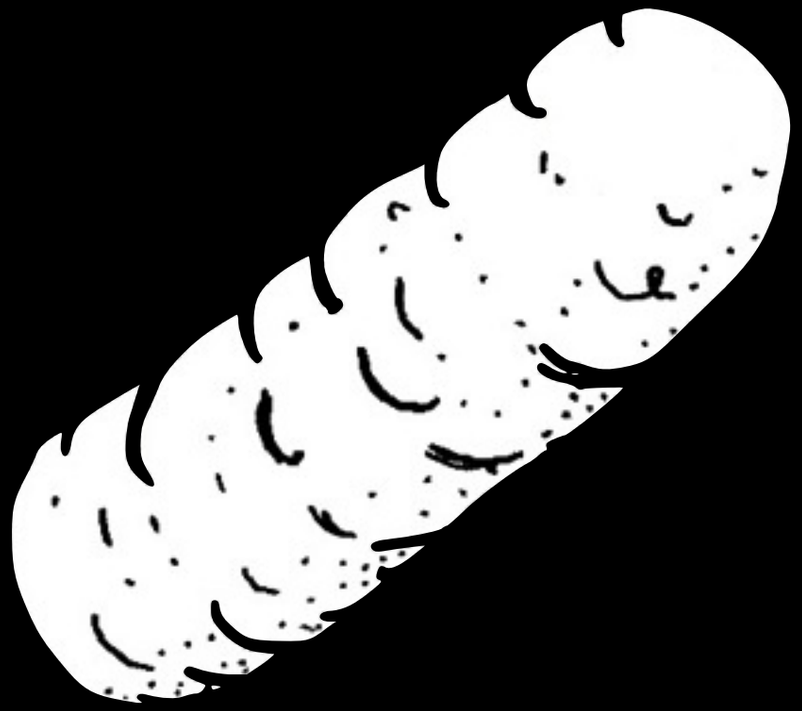
What does “phylogenomics” mean?

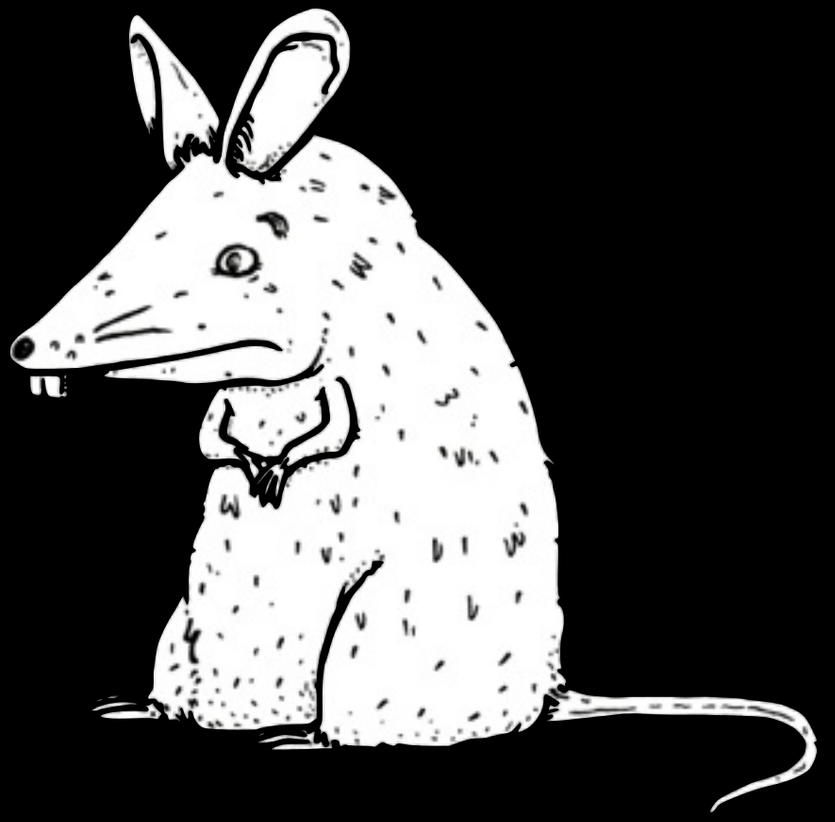
1. The study of genome evolution in a phylogenetic context

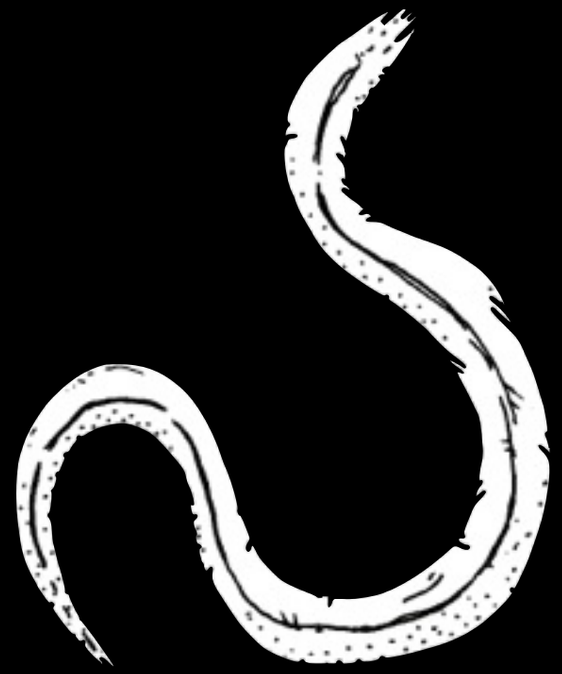
2. The inference of species phylogenies with genome data

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

How do we make links between genes and phenotypes when we can’t do genetics?







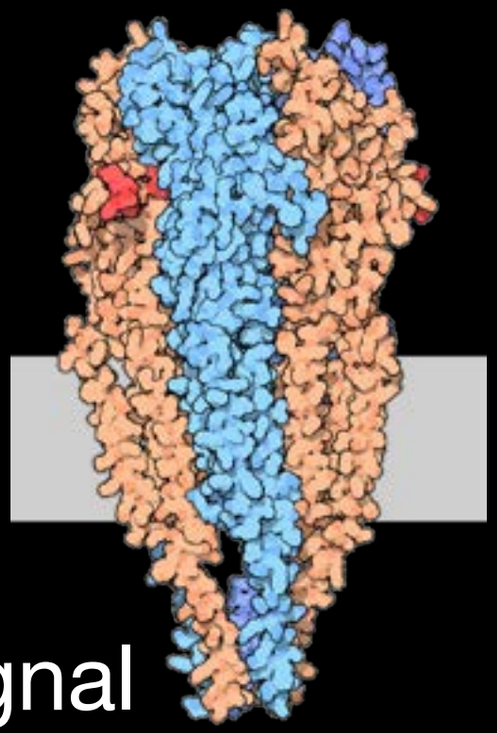
Phylogenetic studies now generate:

- Species trees

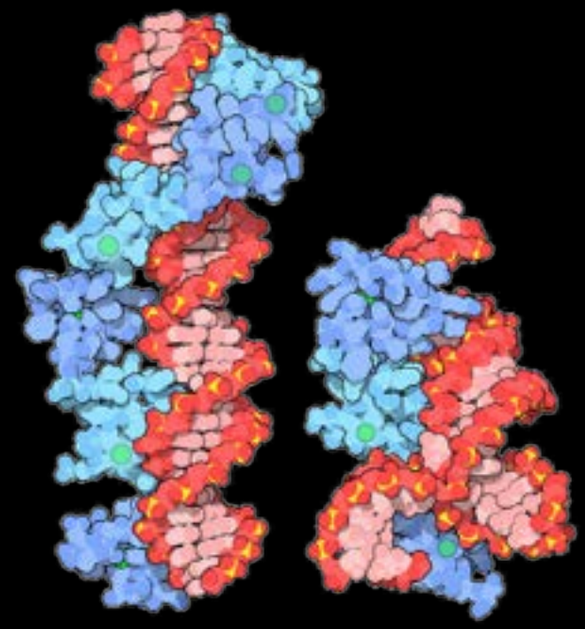
- Extensive gene sequence data

- Well sampled gene trees

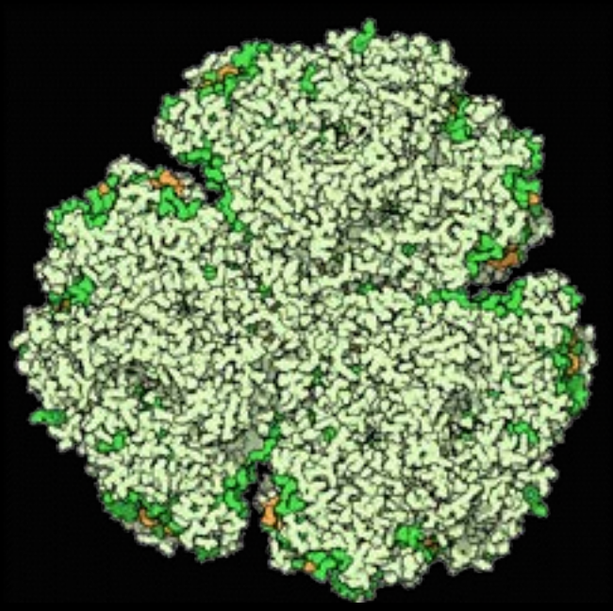
Sequences relevant to focal phenotypes



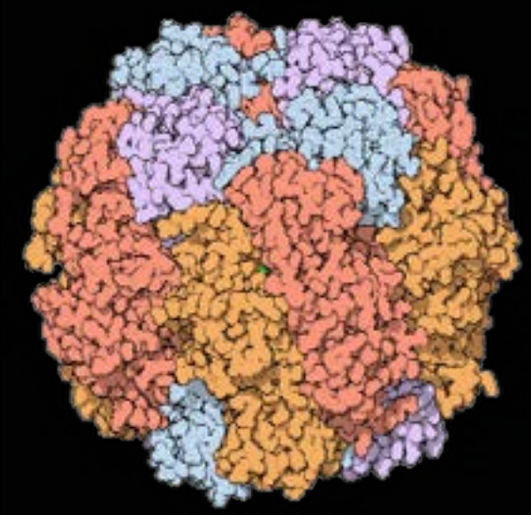
signal
transduction



morphogenesis



photosynthesis



carbon
sequestration

www.rcsb.org/

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

Measuring expression

(MBARI)

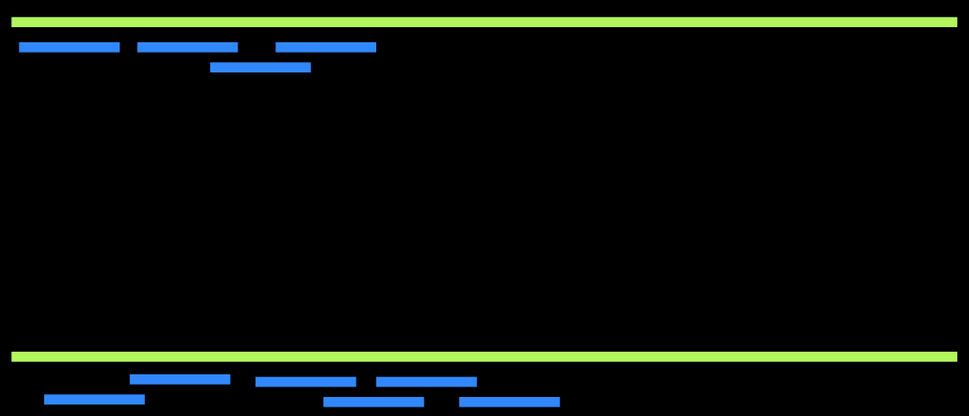
(S Haddock)

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

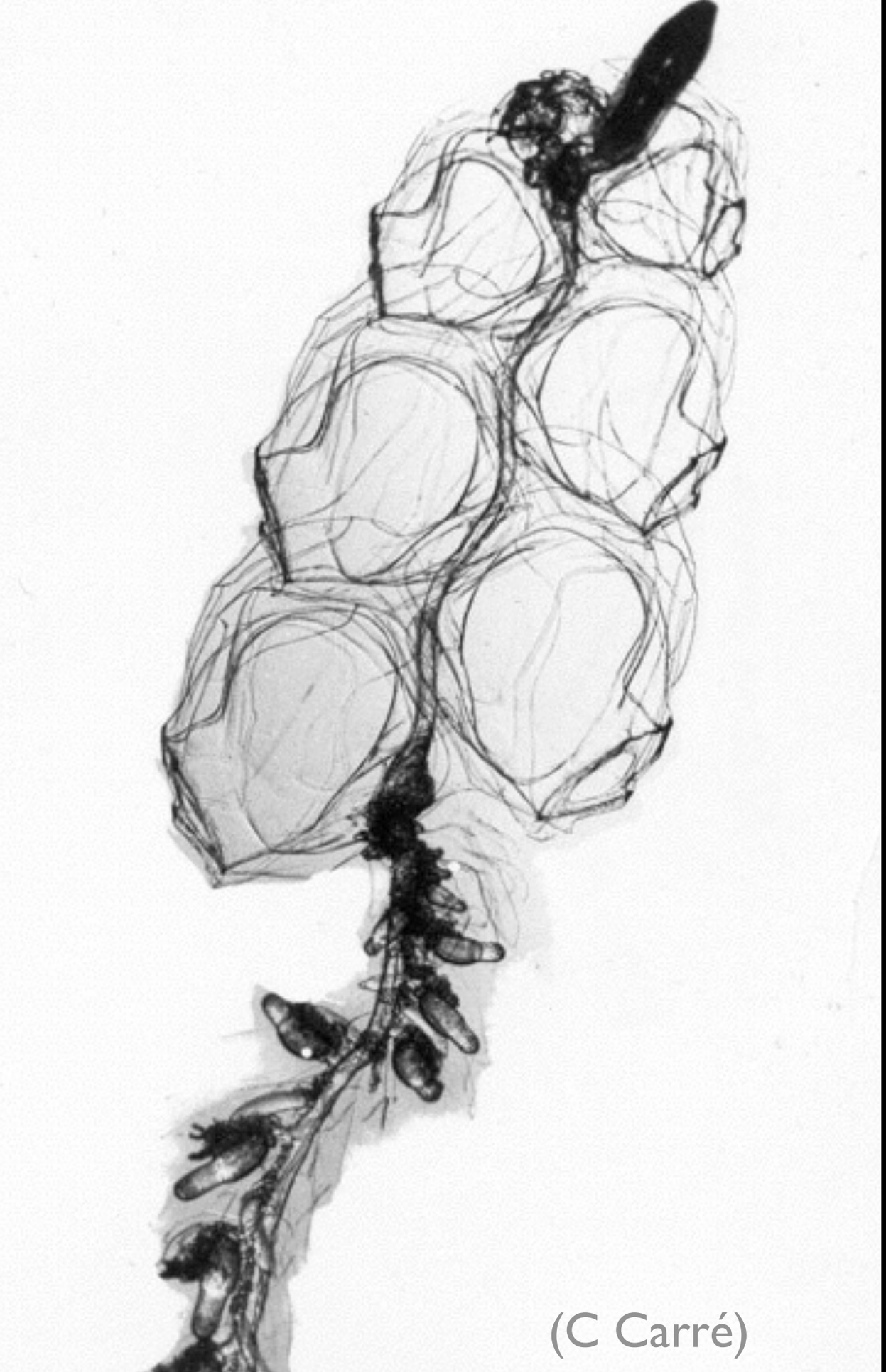
Reference

Reads

Map

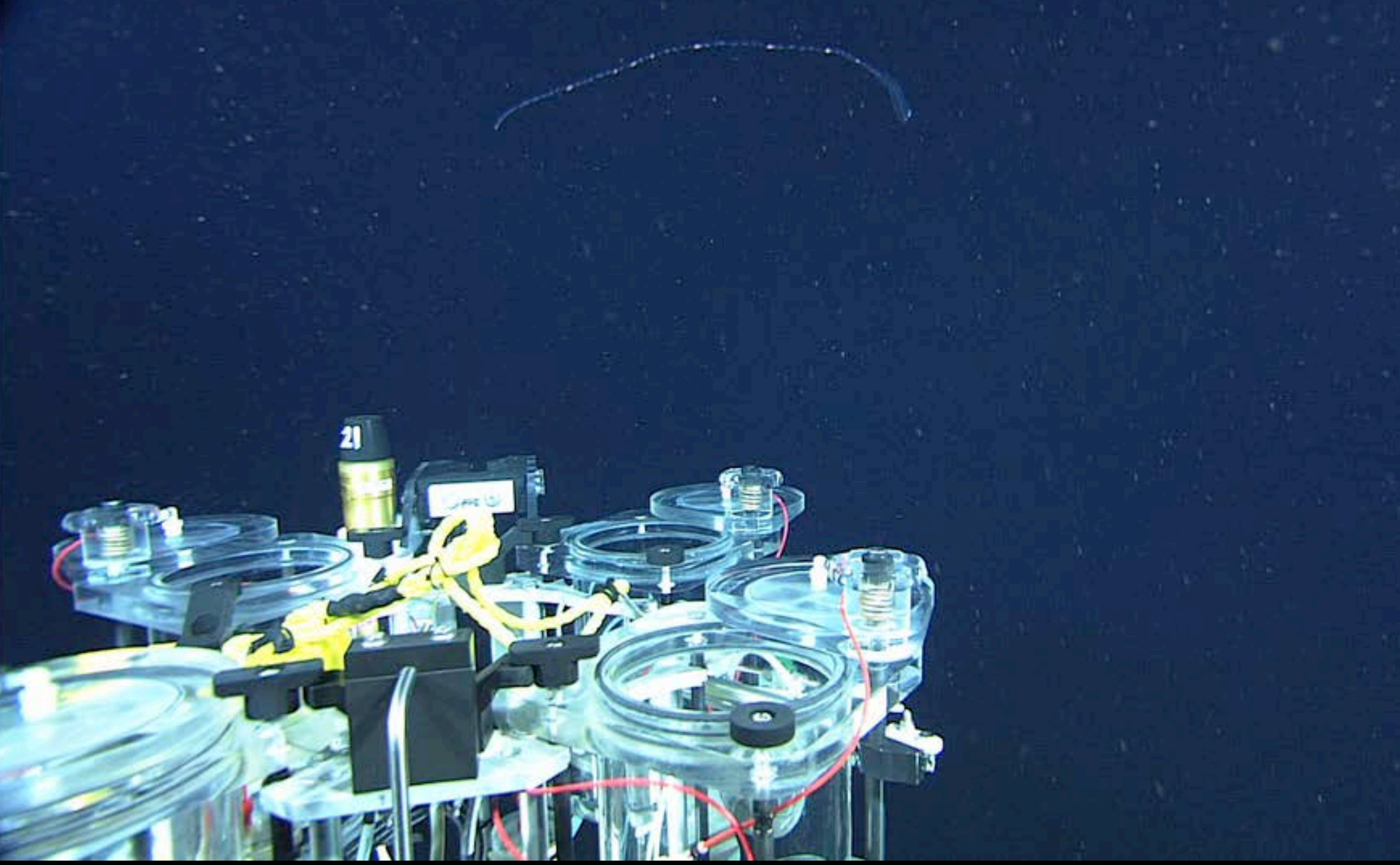


Gene	Count
Gene001	4
Gene002	6
Gene003	22
Gene004	1
Gene005	2

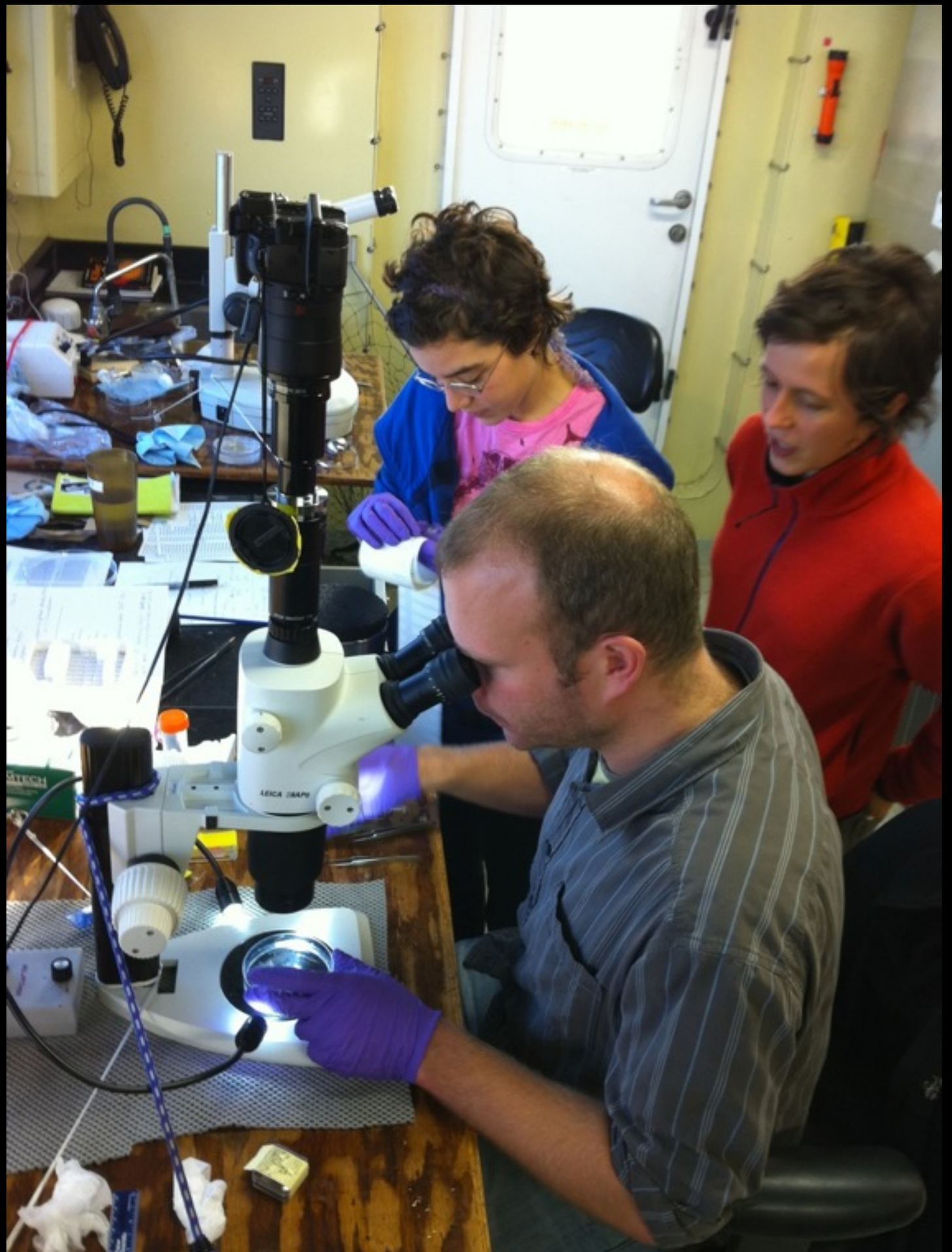


(C Carré)

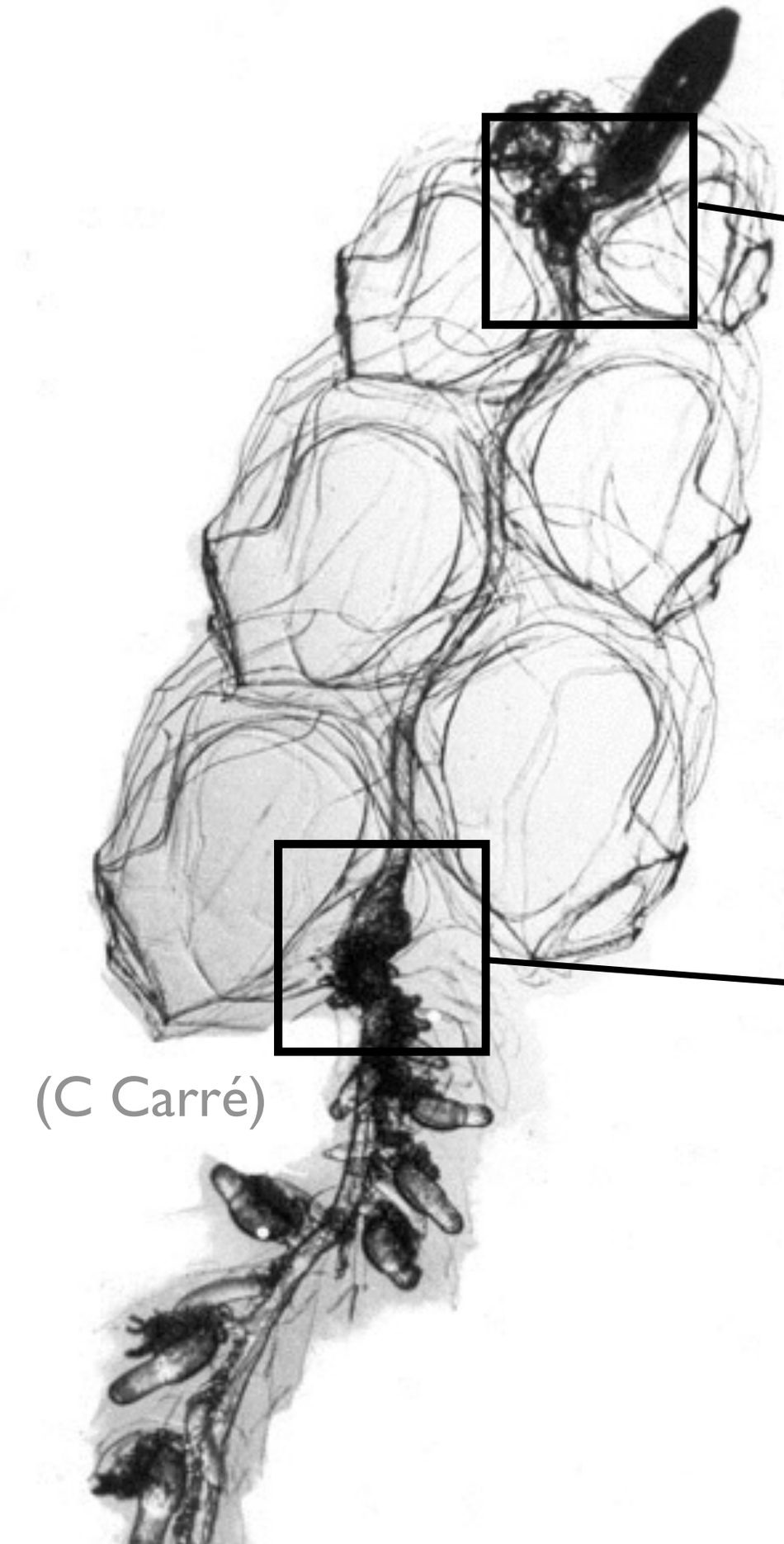
Nanomia bijuga



(MBARI)

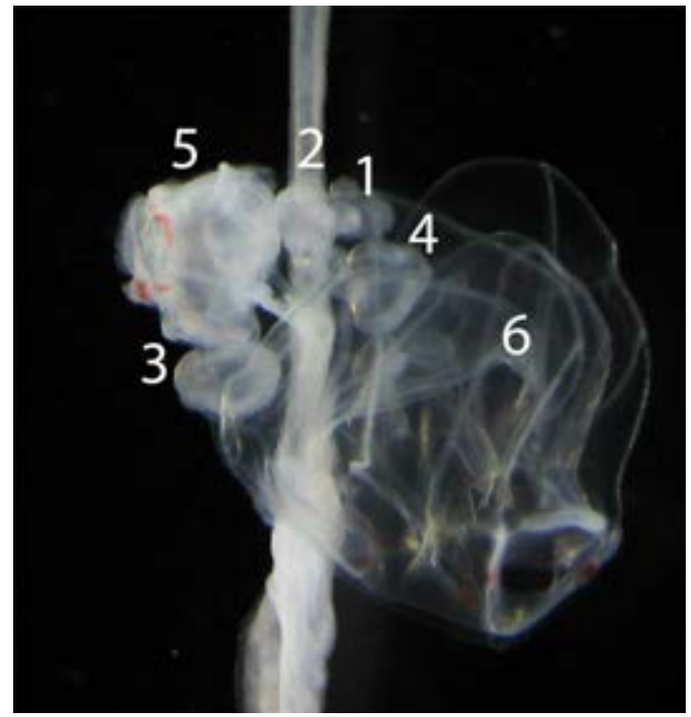


Paired samples, 3 specimens



(C Carré)

Swimming



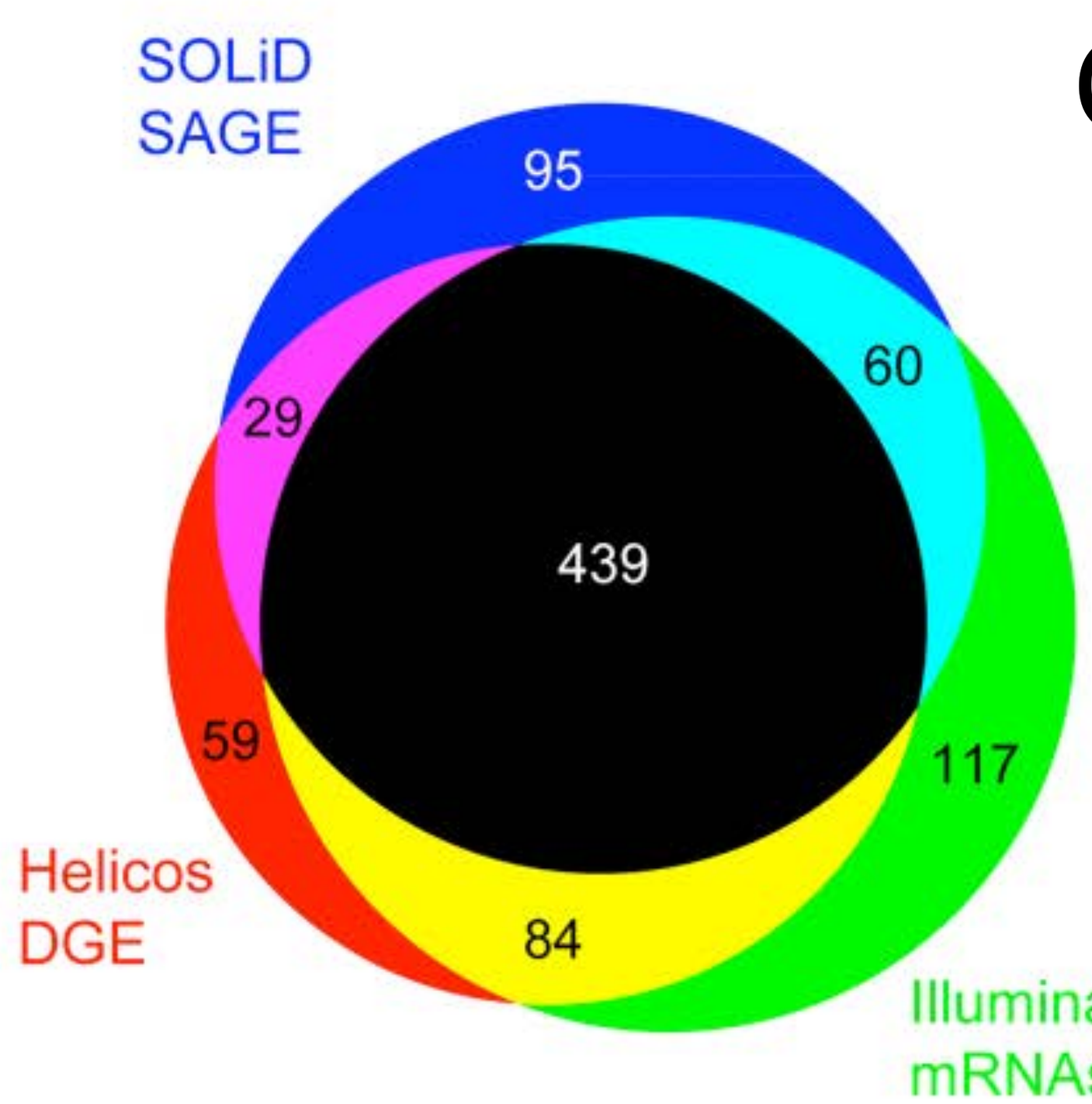
Feeding

Replicated design

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

Helicos SOLiD Illumina

Genes with significant DE



Genes with complete 3’ end

EdgeR, Bonferroni corrected $p < 0.05$

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

Where to next?

- Characterization of genes with significant differential expression

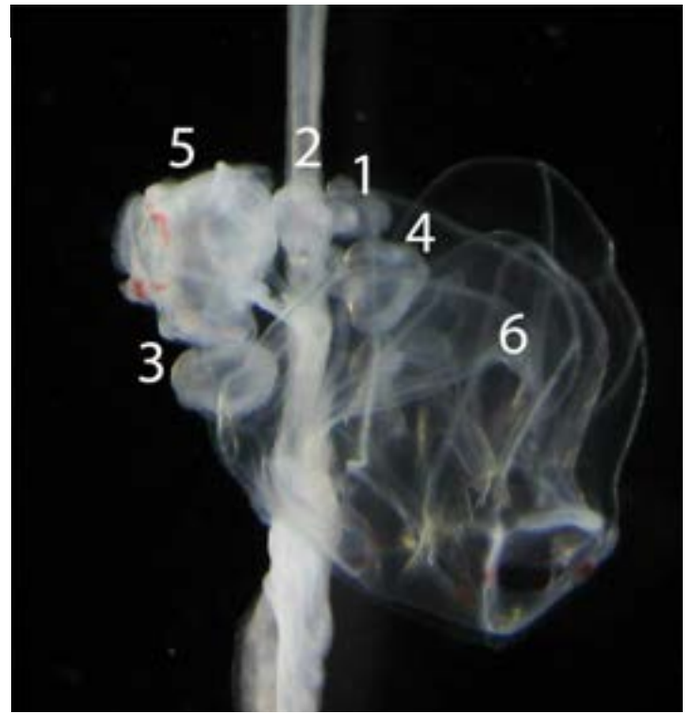
Red genes have significant differential expression

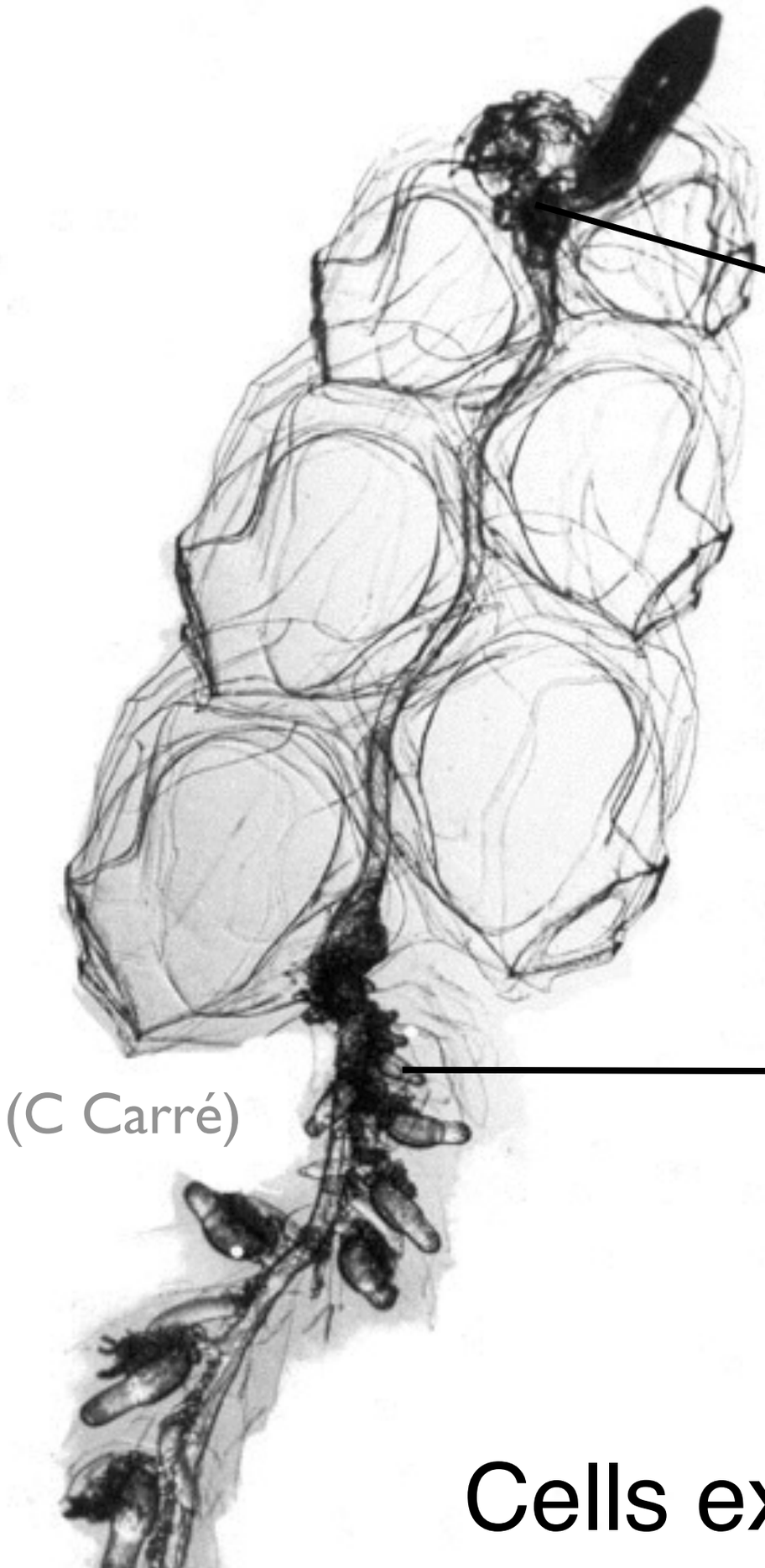
Swimming

Feeding

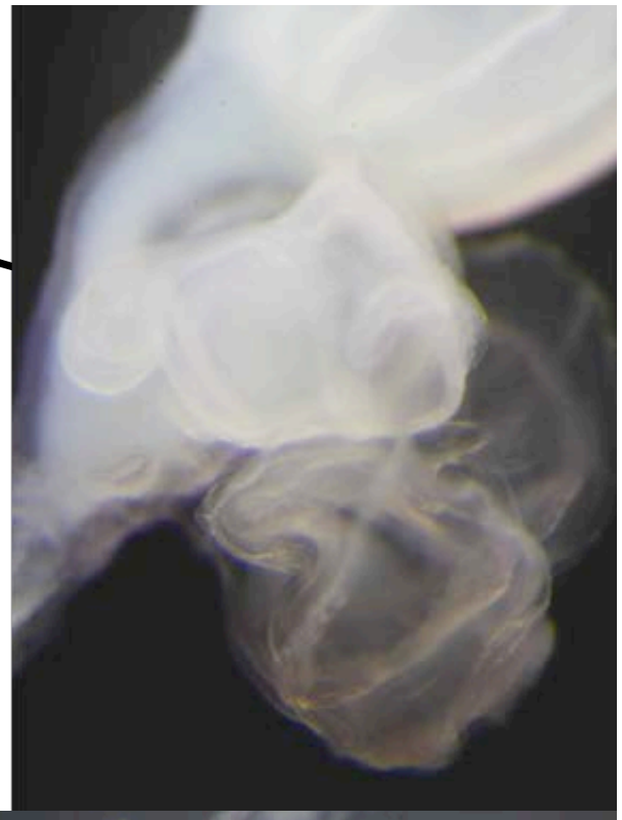
Overall expression →

Mini-collagen

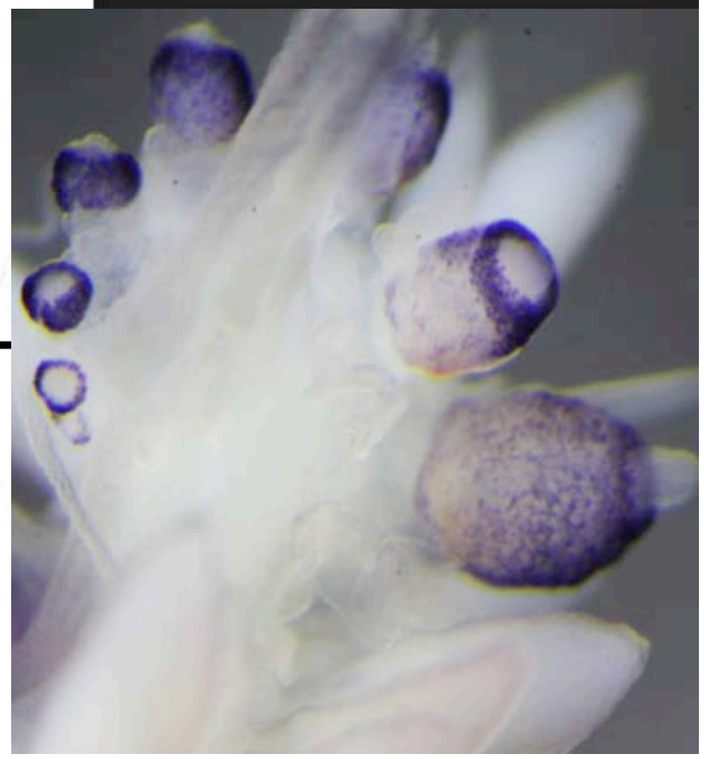




(C Carré)



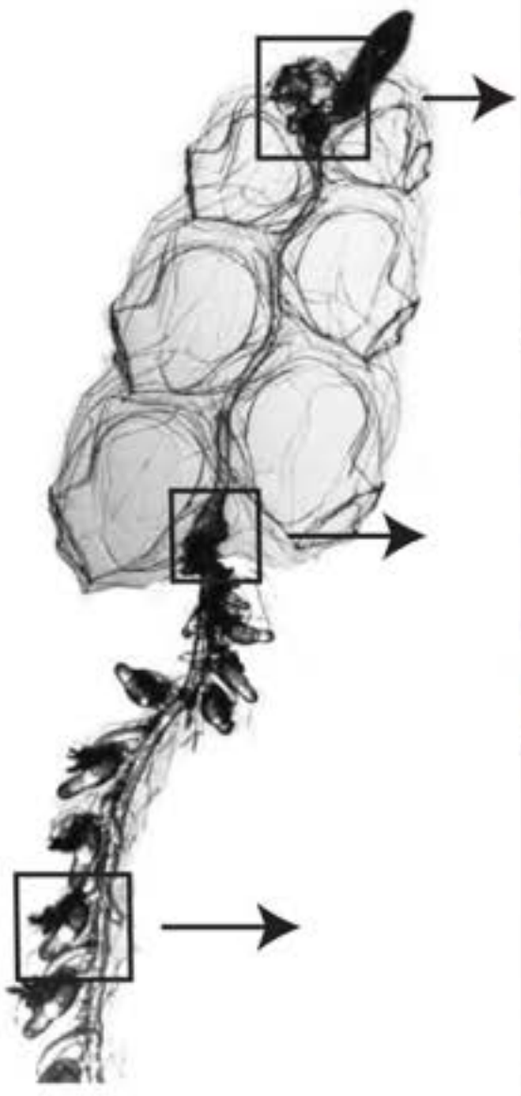
swimming bodies



feeding bodies

Cells expressing mini-collagen are **blue**

Mature Colony



03256
(mini-collagen)

06123
(elav-like hu)

00108
(dickkopf3)

06398
(collagen)

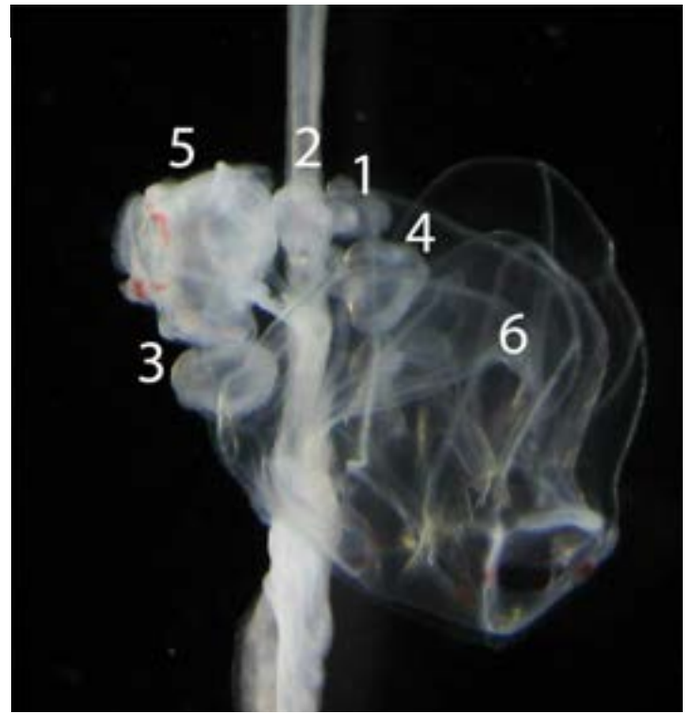
07187
(dapk)

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?

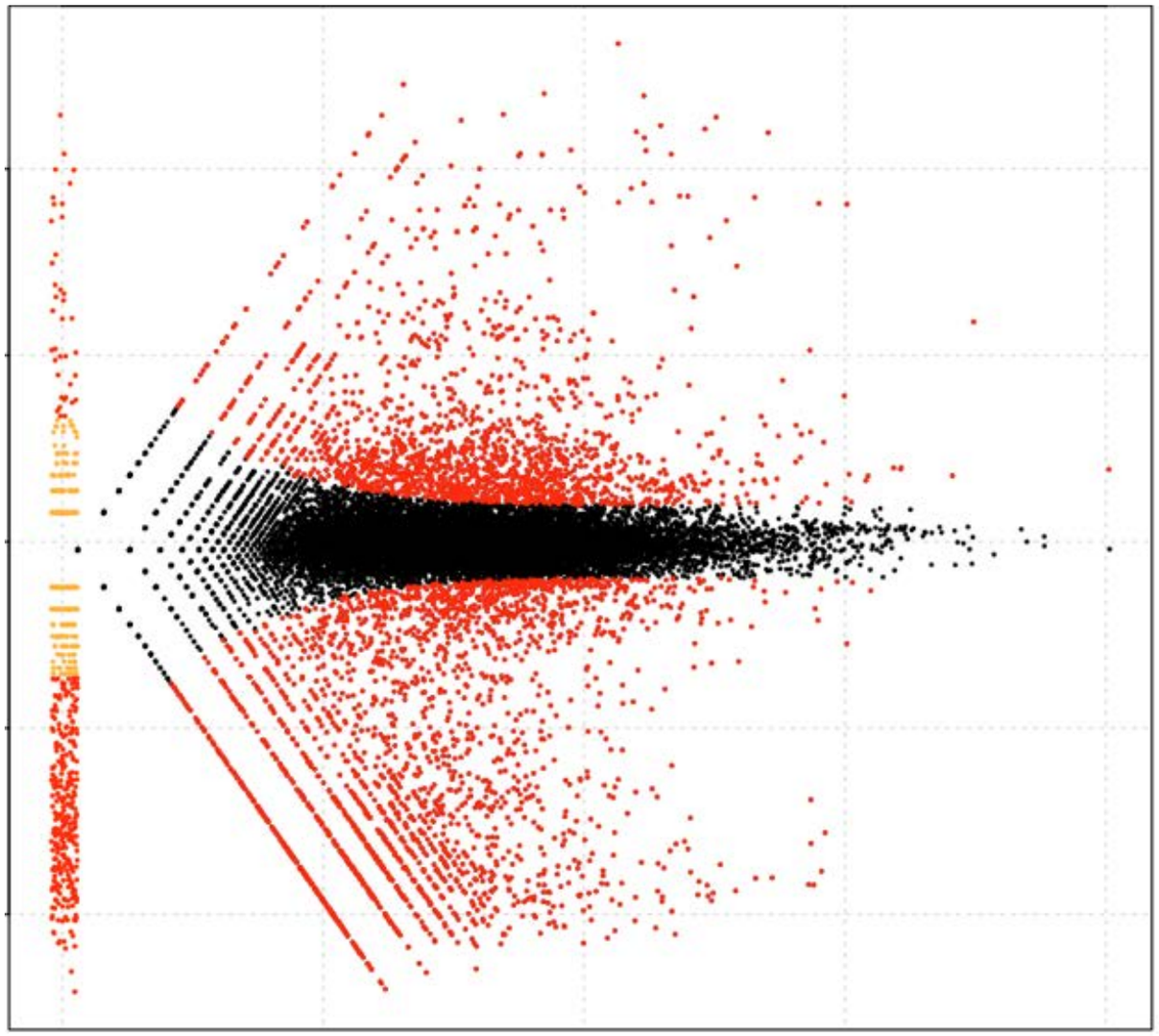
- Comparative data - we need to be looking at a lot more than one species at a time.

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

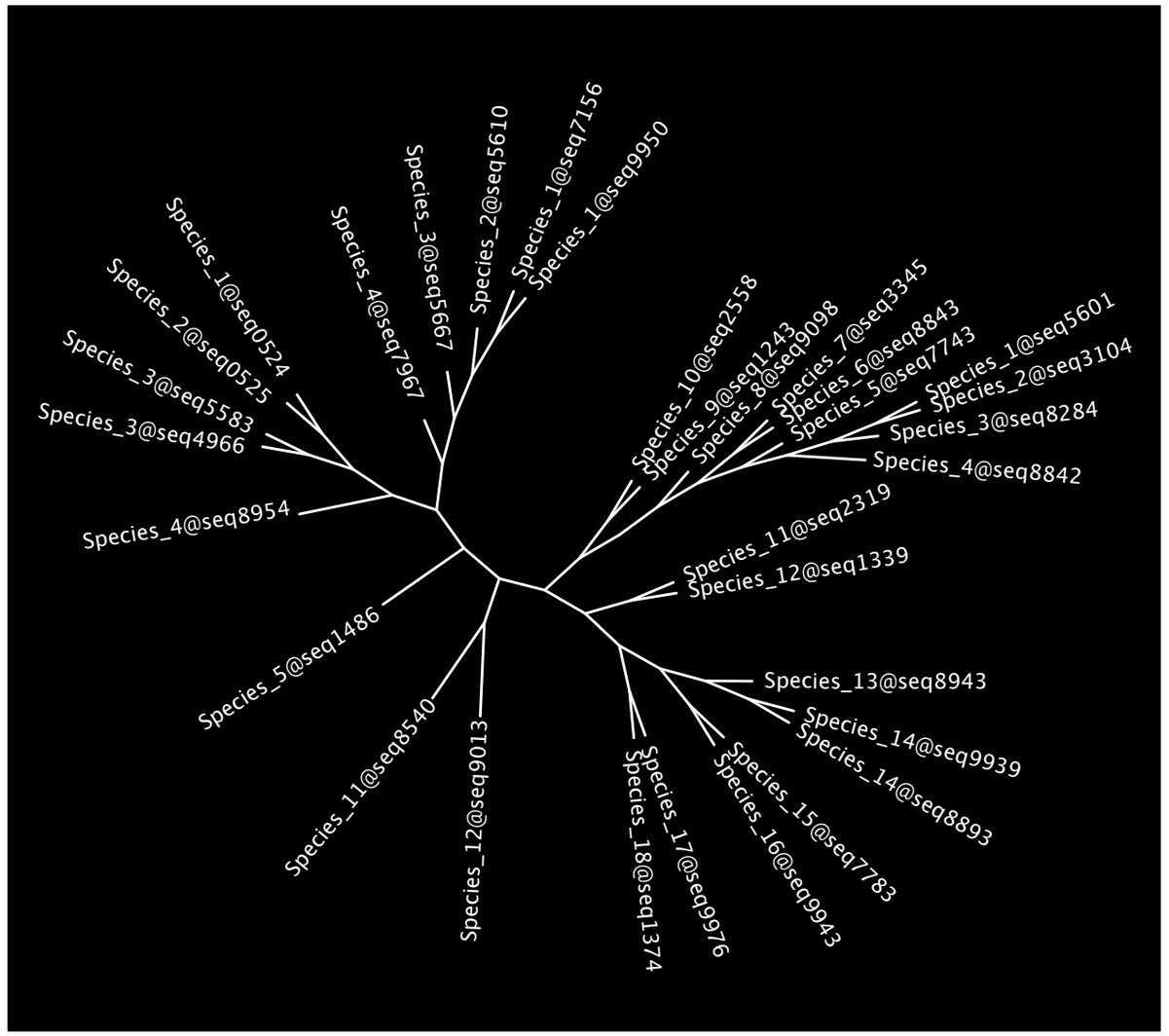
New approach:
Which genes have evolutionary changes in expression that are coincident with changes in my phenotype of interest?

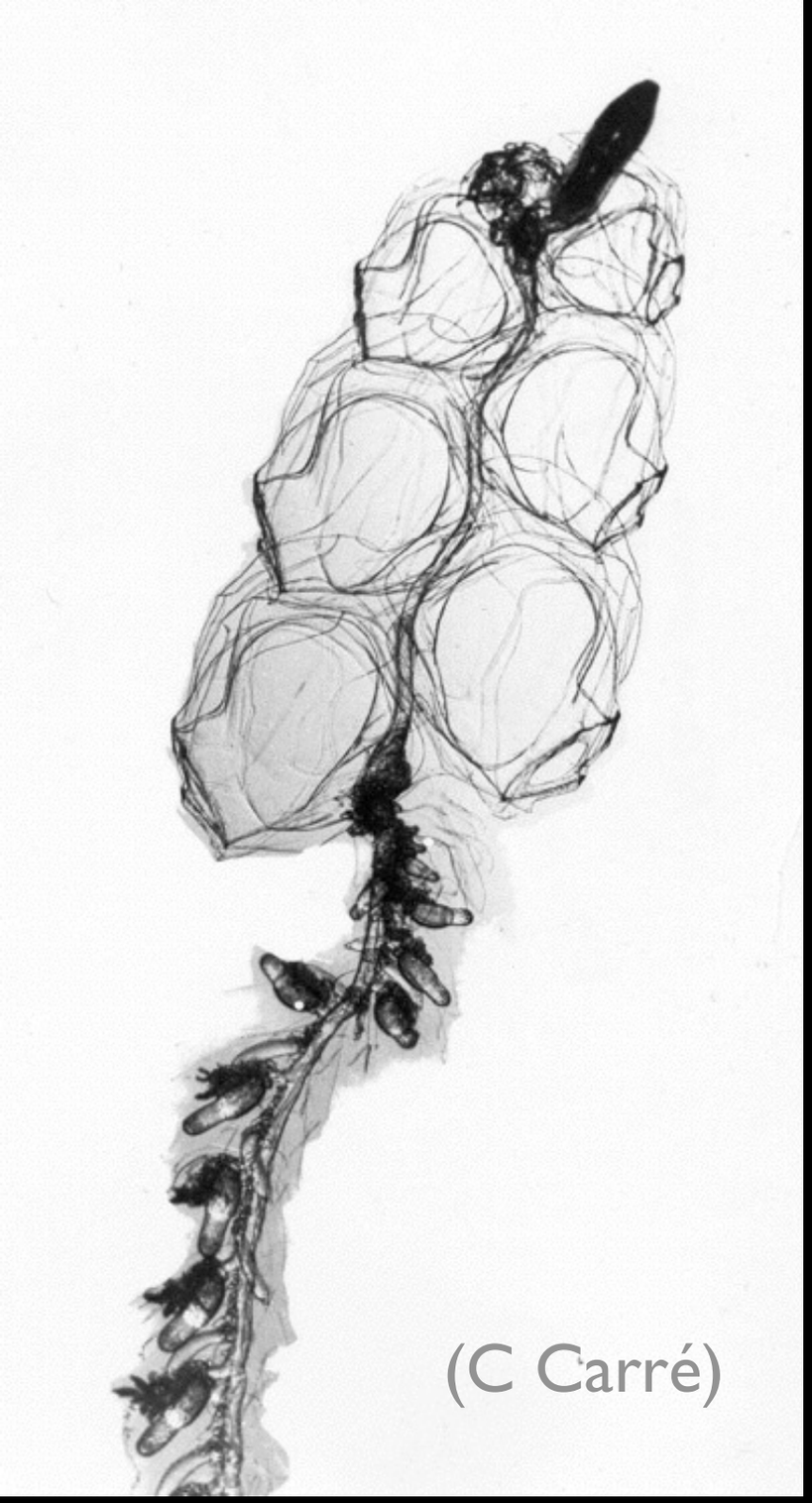
Analyze expression data on phylogenies

Expression data



Gene trees




(C Carré)

Nanomia bijuga

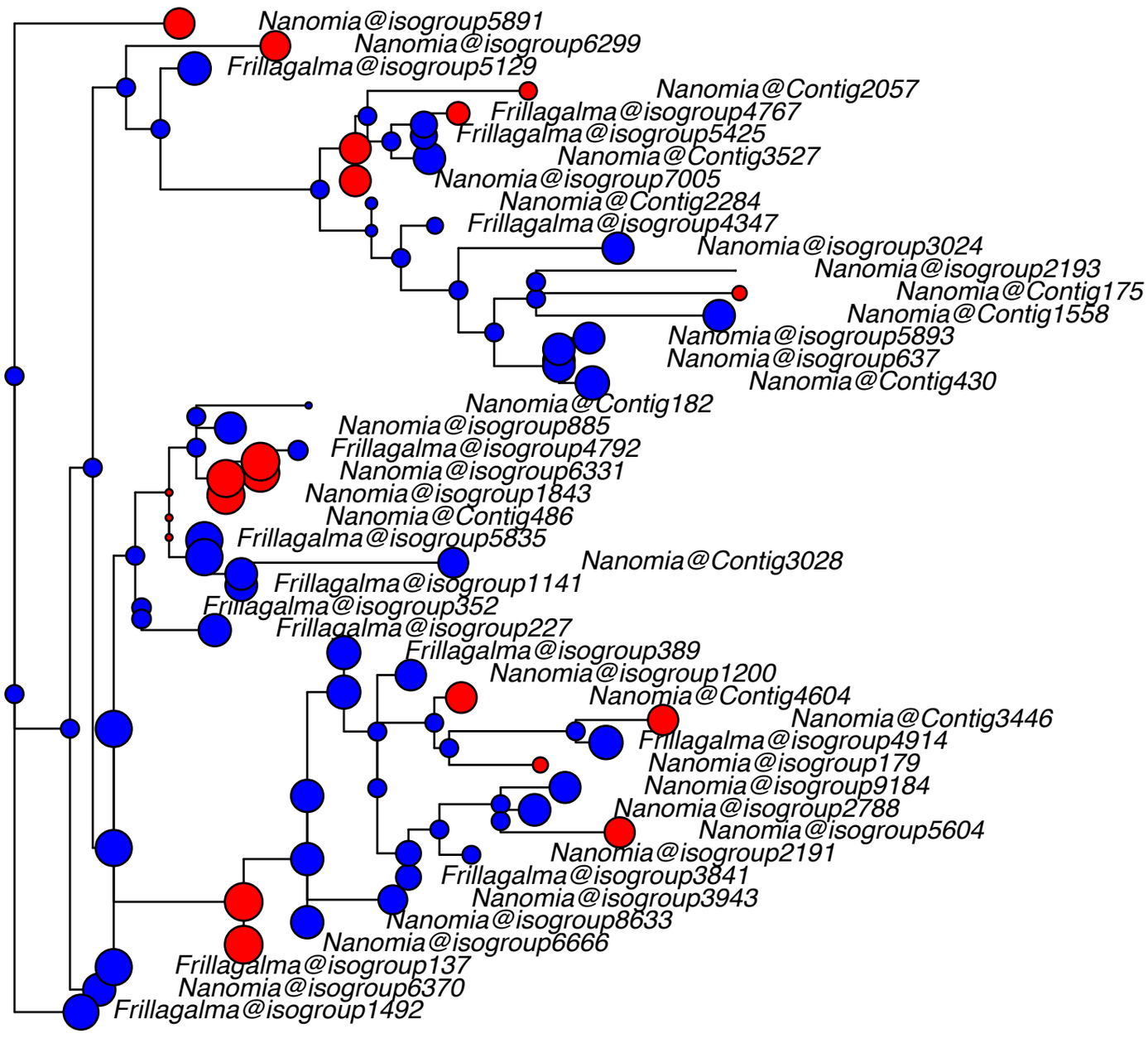
Frillagalma vityazi


(S Haddock)

Bargmannia elongata

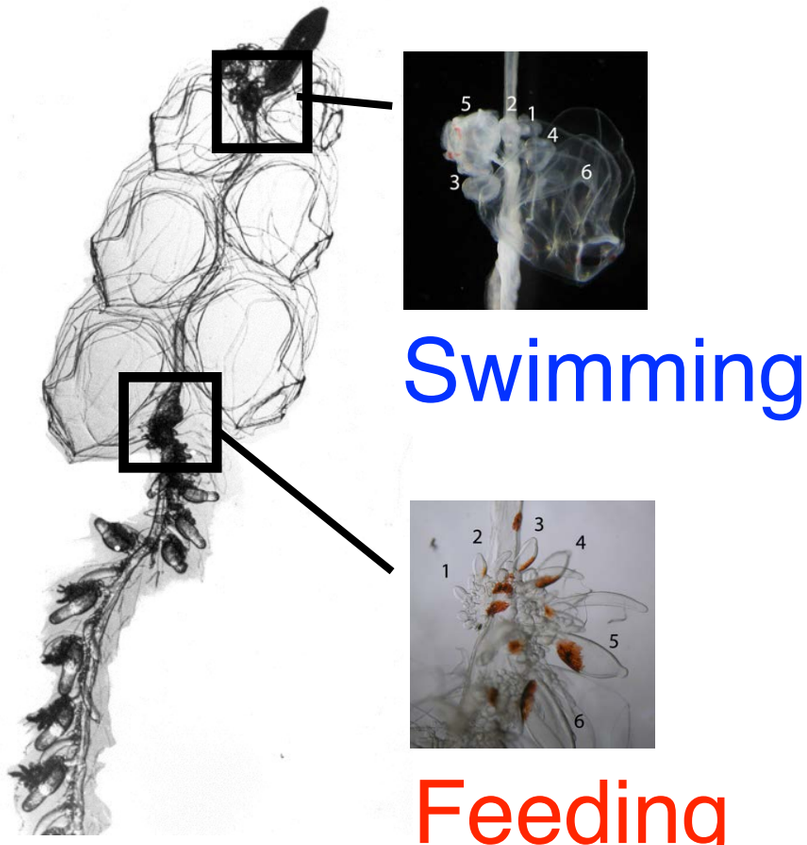
Overview

Gene tree

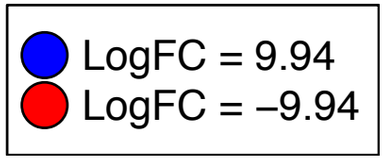


hemicentin

Color indicates direction

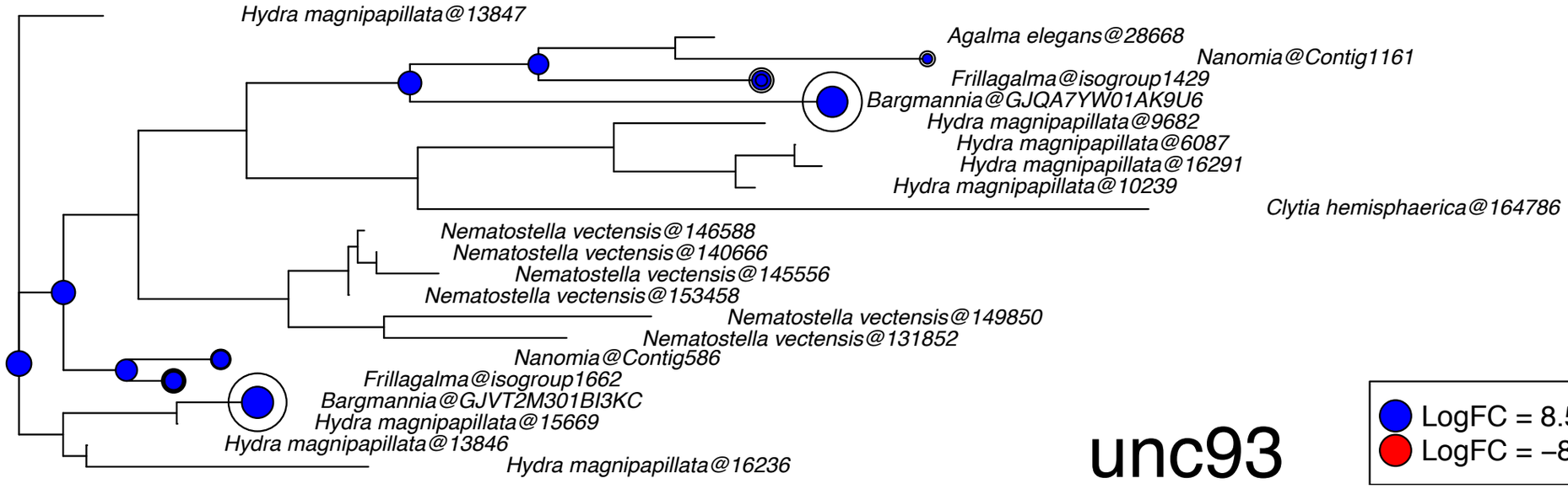


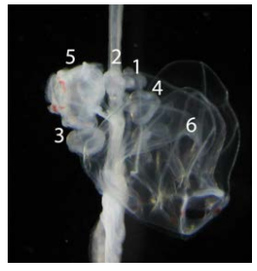
Size indicates magnitude



LogFC - the log base 2 of expression in swimming/ feeding bodies

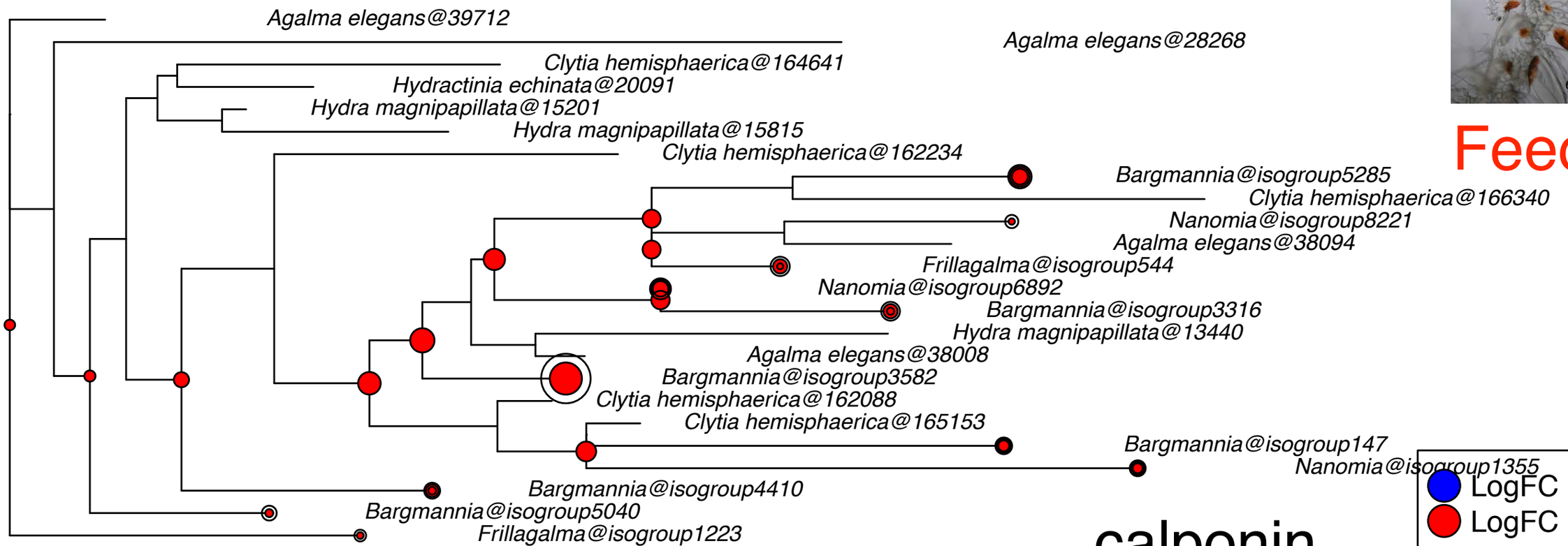
Find gene families that always have differential expression in the same direction





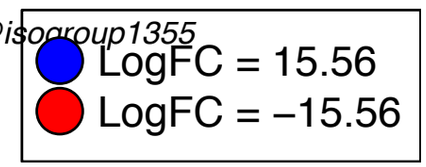
Swimming

unc93



Feeding

calponin



This approach can be used to

- Identify genes that have shifts in expression associated with shifts in other phenotypes of interest

- Genes that have evolutionary covariance in expression

Collaborators

Joe Felsenstein (UW)

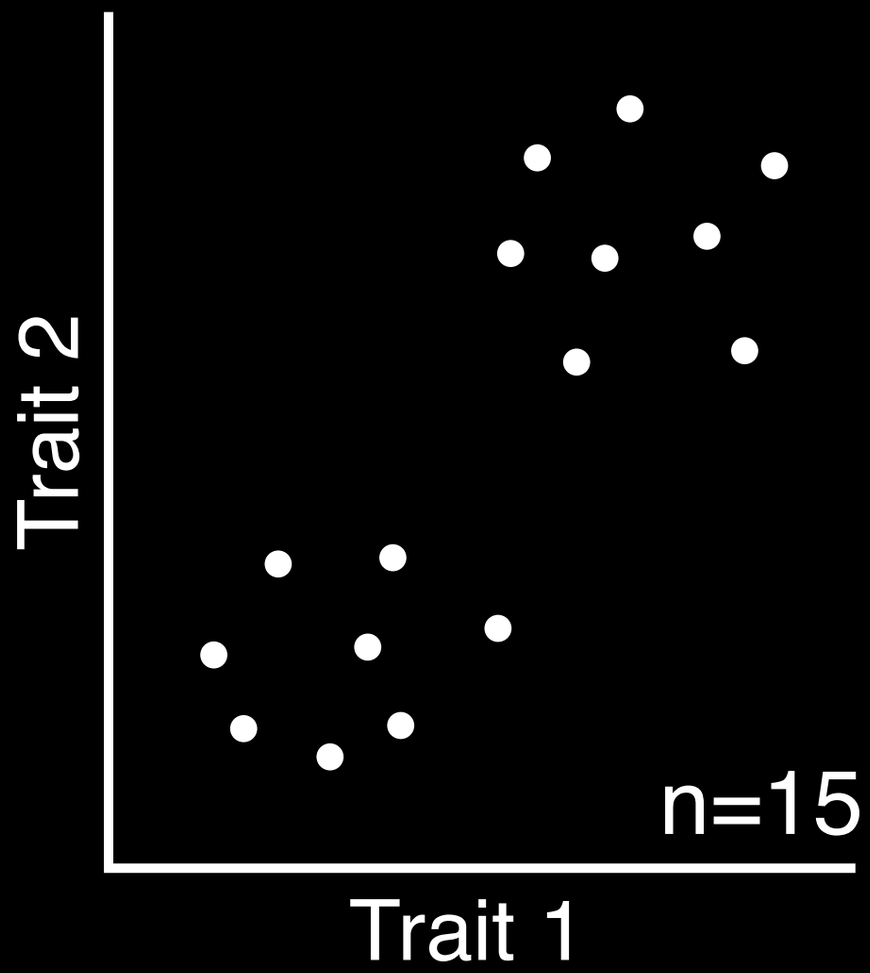
Xi Luo (Brown)

Zhijin Wu (Brown)

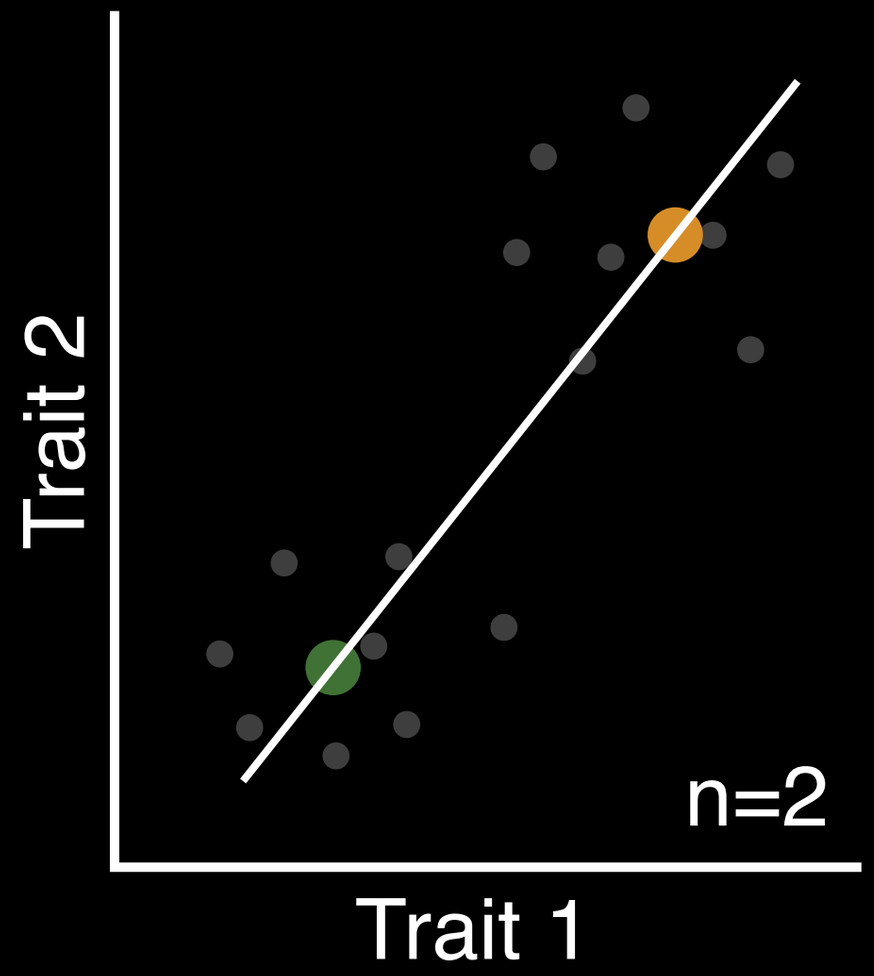
Support

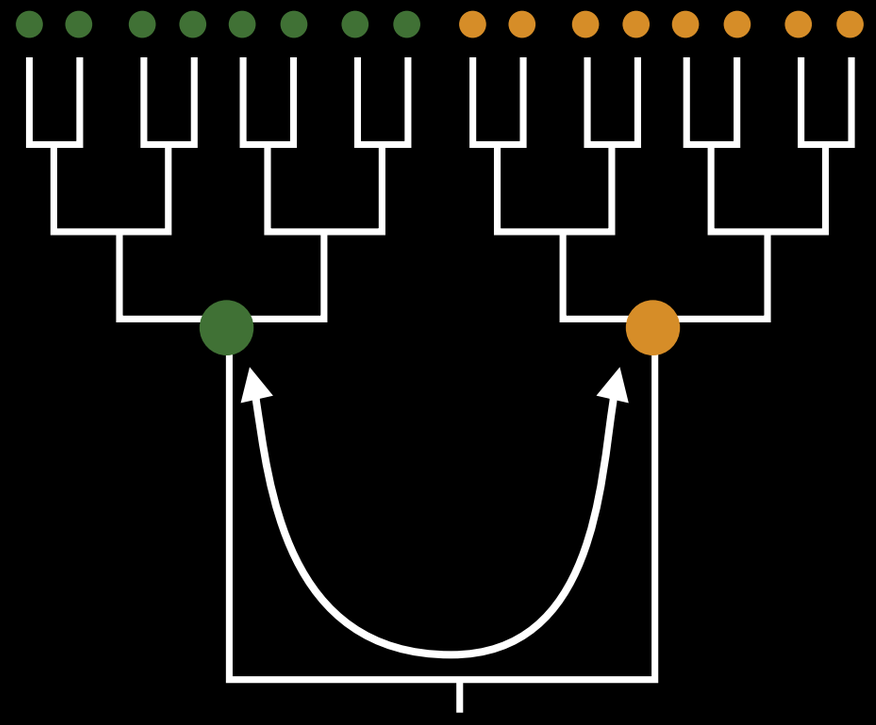
NSF- DEB, Waterman Award

Why use phylogenies to analyze expression across species?



Why use phylogenies to analyze data across species?





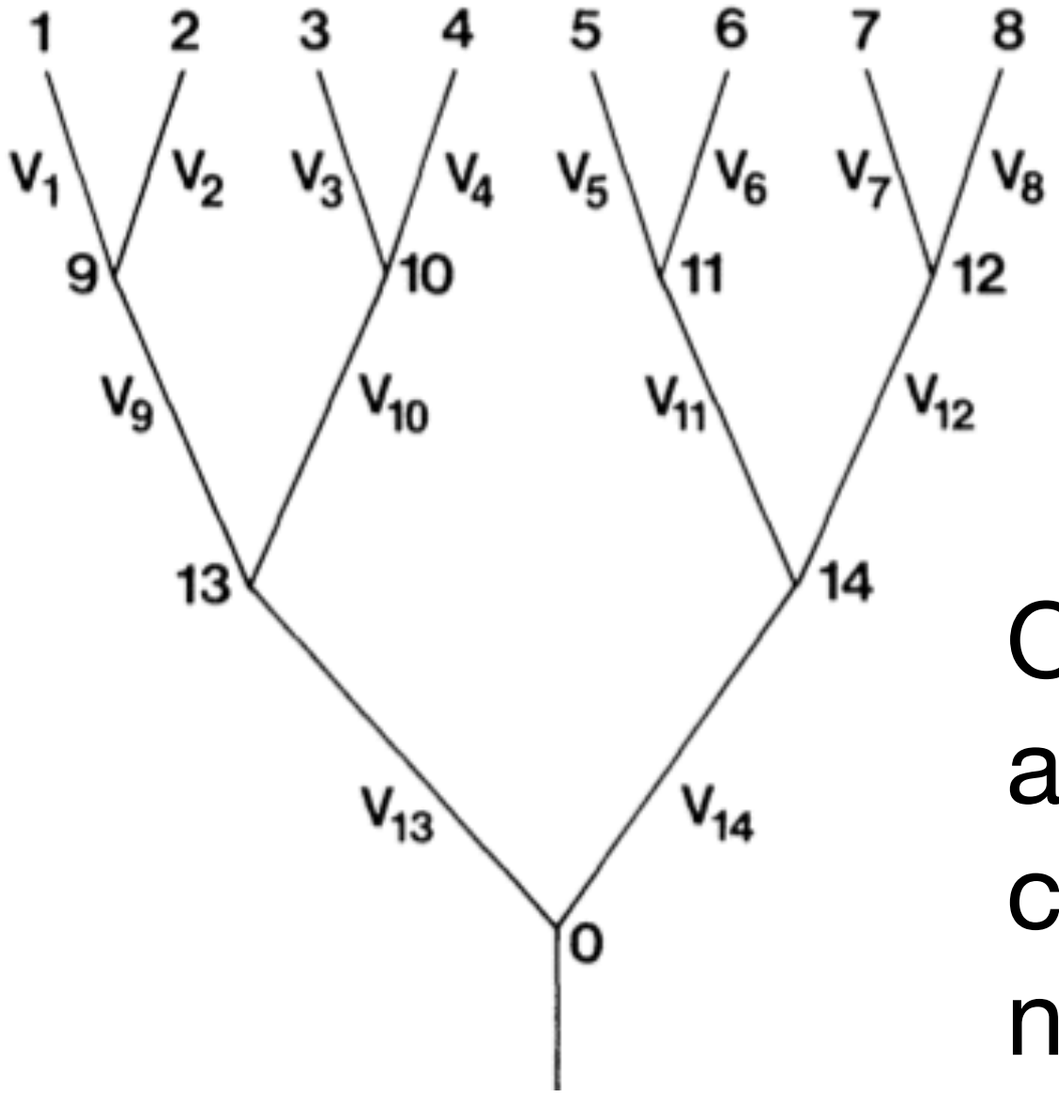
Vol. 125, No. 1 The American Naturalist January 1985

PHYLOGENIES AND THE COMPARATIVE METHOD

JOSEPH FELSENSTEIN

Department of Genetics SK-50, University of Washington, Seattle, Washington 98195

Submitted November 30, 1983; Accepted May 23, 1984



Observations across species are not independent, but contrasts across internal nodes are

Integrative and Comparative Biology

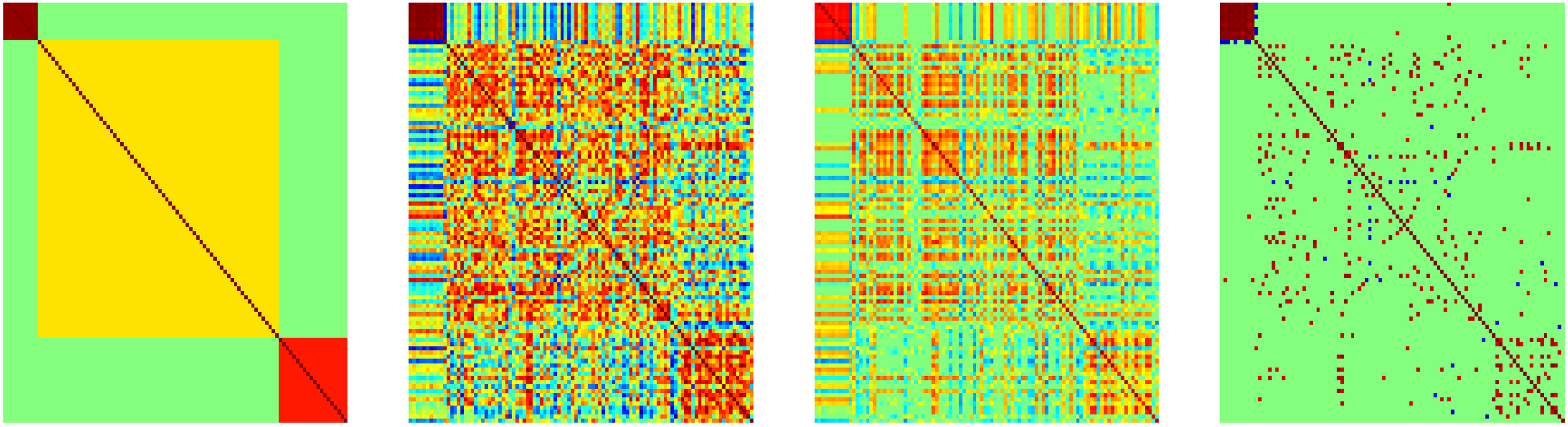
Integrative and Comparative Biology, pp. 1–10
doi:10.1093/icb/ict068

Society for Integrative and Comparative Biology

Phylogenetic Analysis of Gene Expression

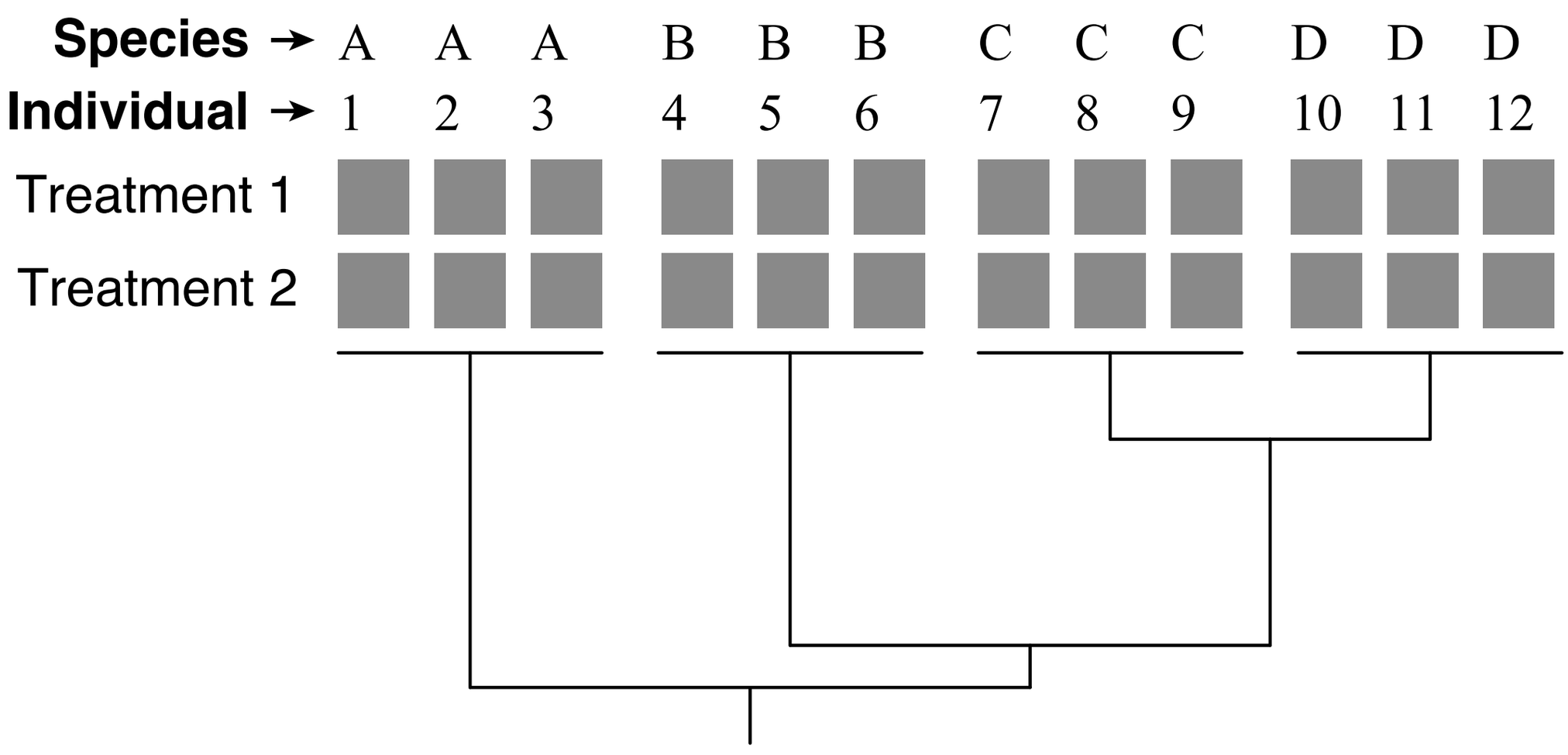
Casey W. Dunn,[1,*] Xi Luo[†] and Zhijin Wu[†]

*Department of Ecology and Evolutionary Biology, Brown University, Providence, RI, USA; [†]Department of Biostatistics and Center for Statistical Sciences, Brown University, Providence, RI 02903, USA



http://dx.doi.org/10.1093/icb/ict068

A typical project design:



Each grey box is a sample

Dunn et al 2013 (http://dx.doi.org/10.1093/icb/ict068)

Three major challenges:

1. Measuring expression so that it can be compared across species.

2. Interpreting covariance when the number of genes greatly exceeds the number of species.

3. Accommodating incongruence between gene and species trees.

Three major challenges:

1. Measuring expression so that it can be compared across species.

2. Interpreting covariance when the number of genes greatly exceeds the number of species.

3. Accommodating incongruence between gene and species trees.

II. Interpreting covariance

II. Interpreting covariance

We want to understand the relationship of expression across genes and relative to other phenotypes

II. Interpreting covariance

In most comparative analyses:

$$n > p$$

n number of observations
(eg contrasts)
p *number* of variables

II. Interpreting covariance

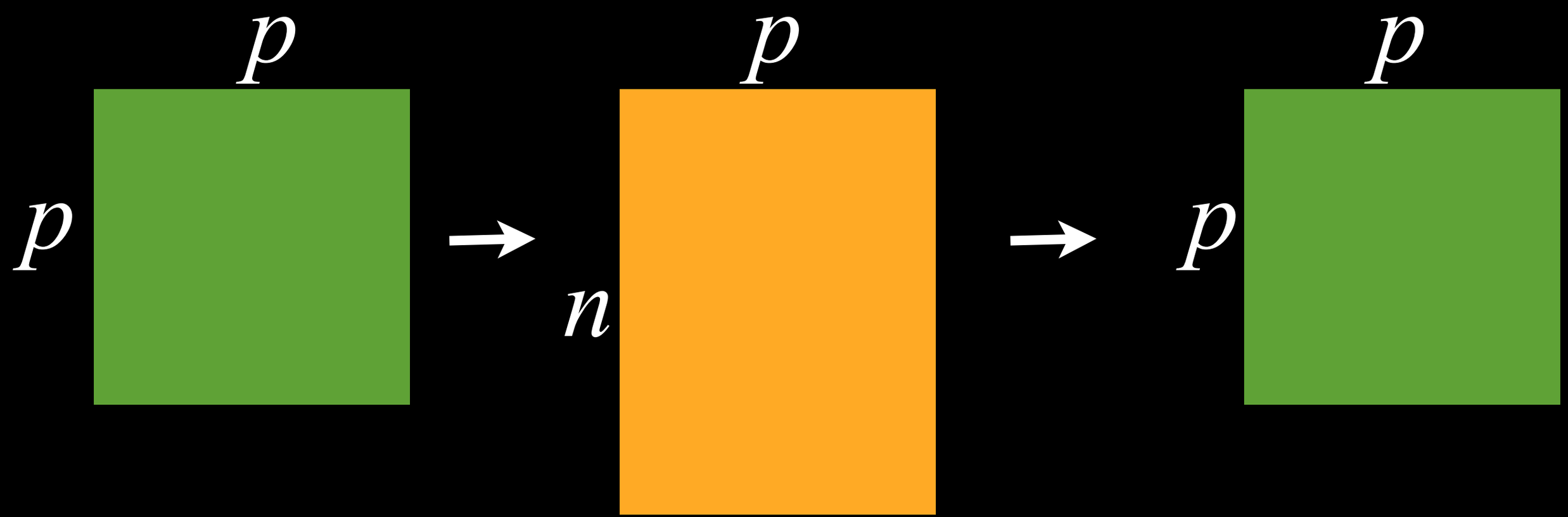
In comparative analyses of gene expression:

$$n << p$$

n number of observations
(eg contrasts)
p number of variables

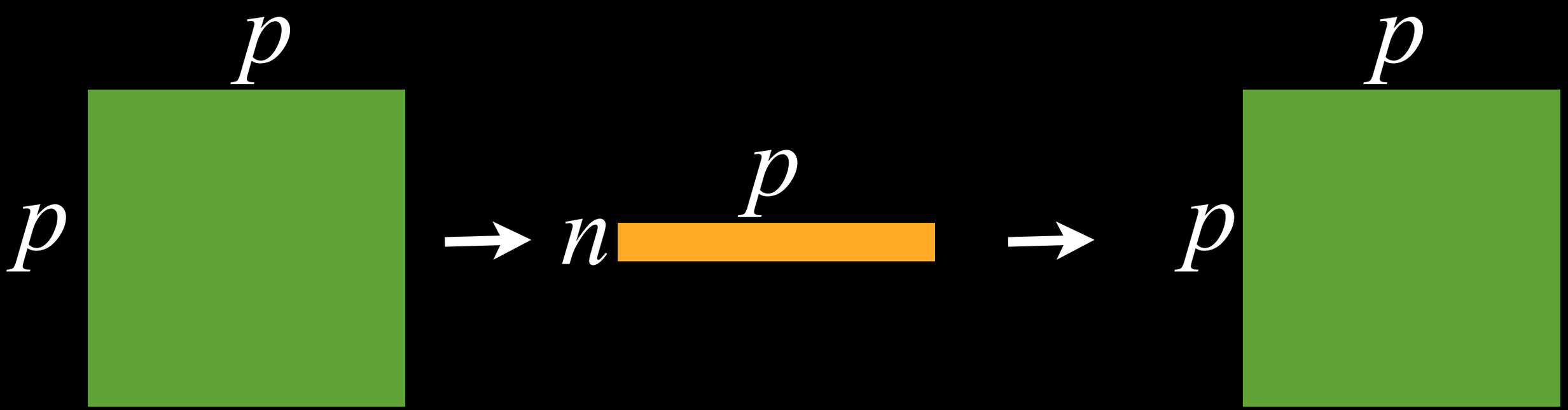
II. Interpreting covariance

When $n > p$



II. Interpreting covariance

When $n << p$



True Covariance

Contrasts

Observed Covariance

II. Interpreting covariance

The covariance matrix is well behaved when $n > p$

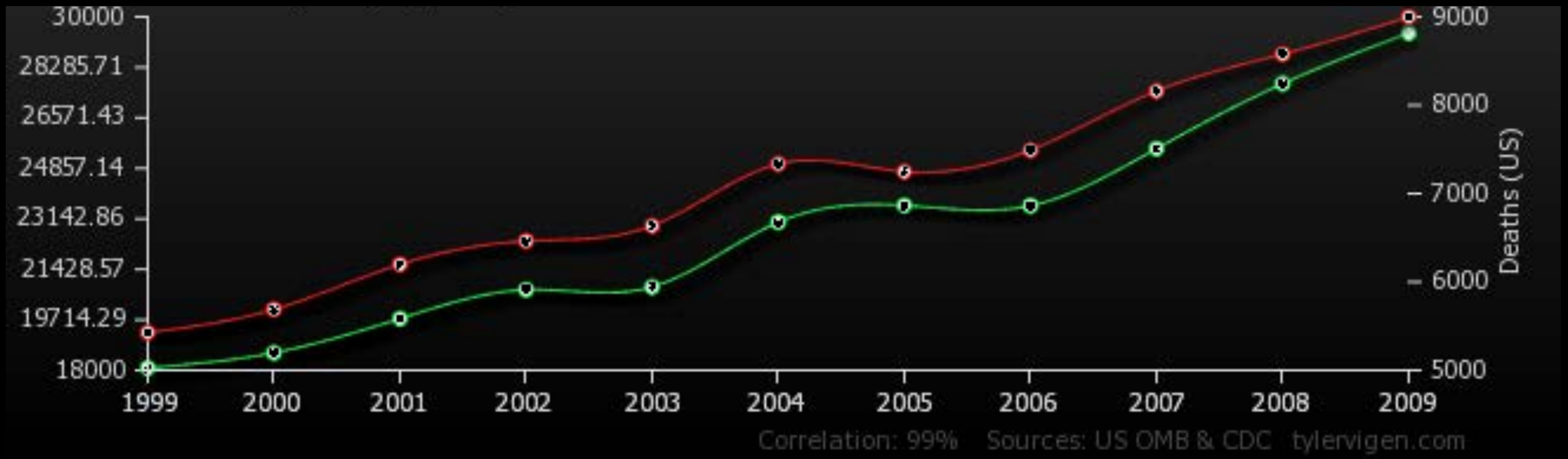
It is difficult to use and potentially misleading when $n << p$

II. Interpreting covariance

Challenges of working with matrices when $n << p$:

- Matrices are singular (non-invertible)
- Many spurious non-zero covariances

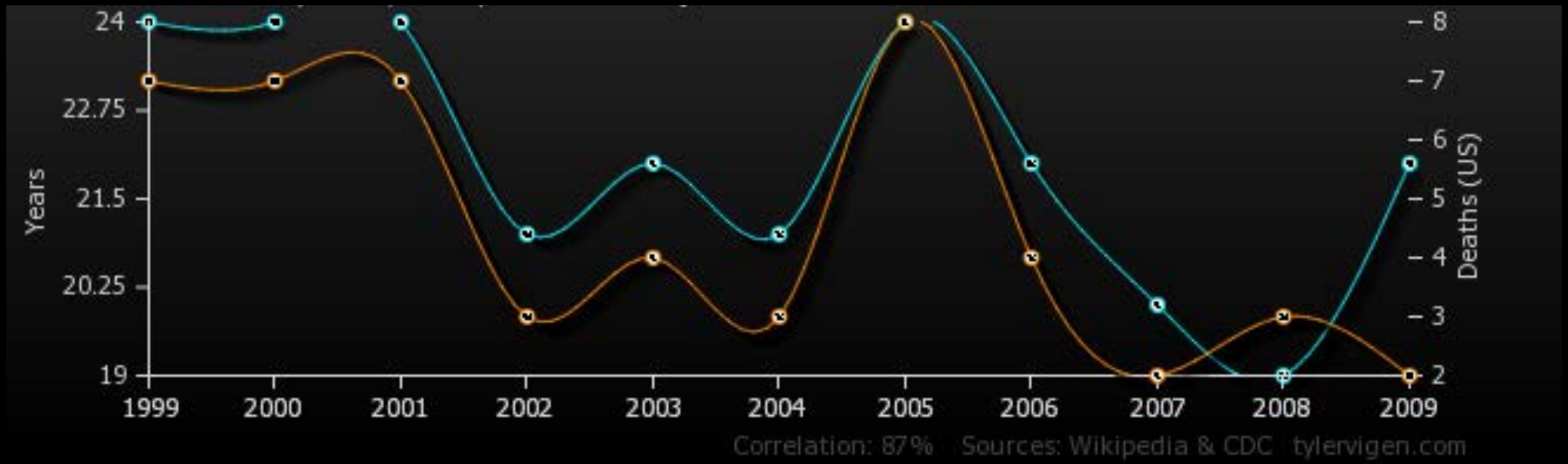
If you are looking at many variables in a small number of observations, you will find many spurious correlations



US spending on science, space, and technology

Suicides by hanging, strangulation, and suffocation

http://www.tylervigen.com

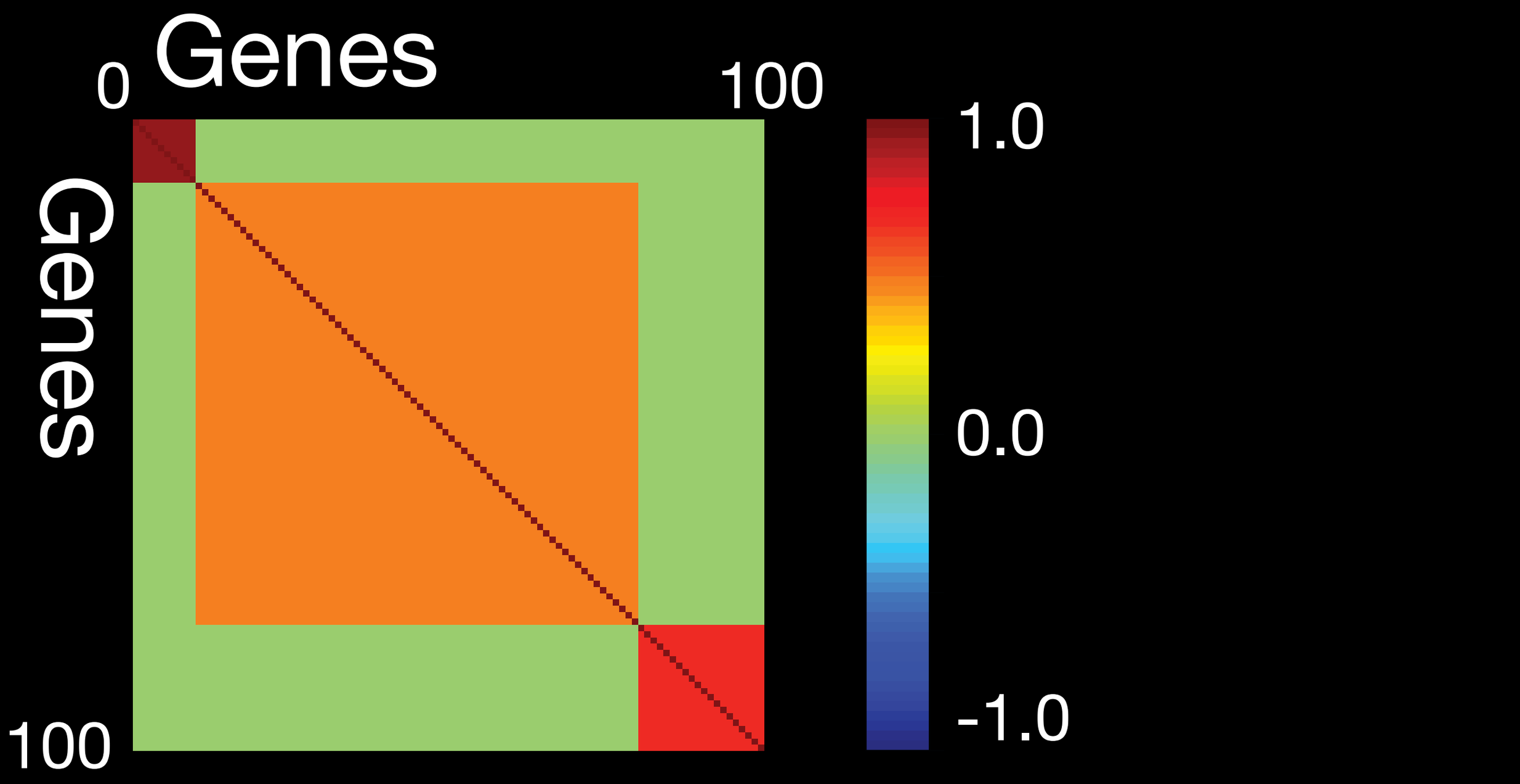


Age of Miss America

Murders by steam, hot vapors, and hot objects

http://www.tylervigen.com

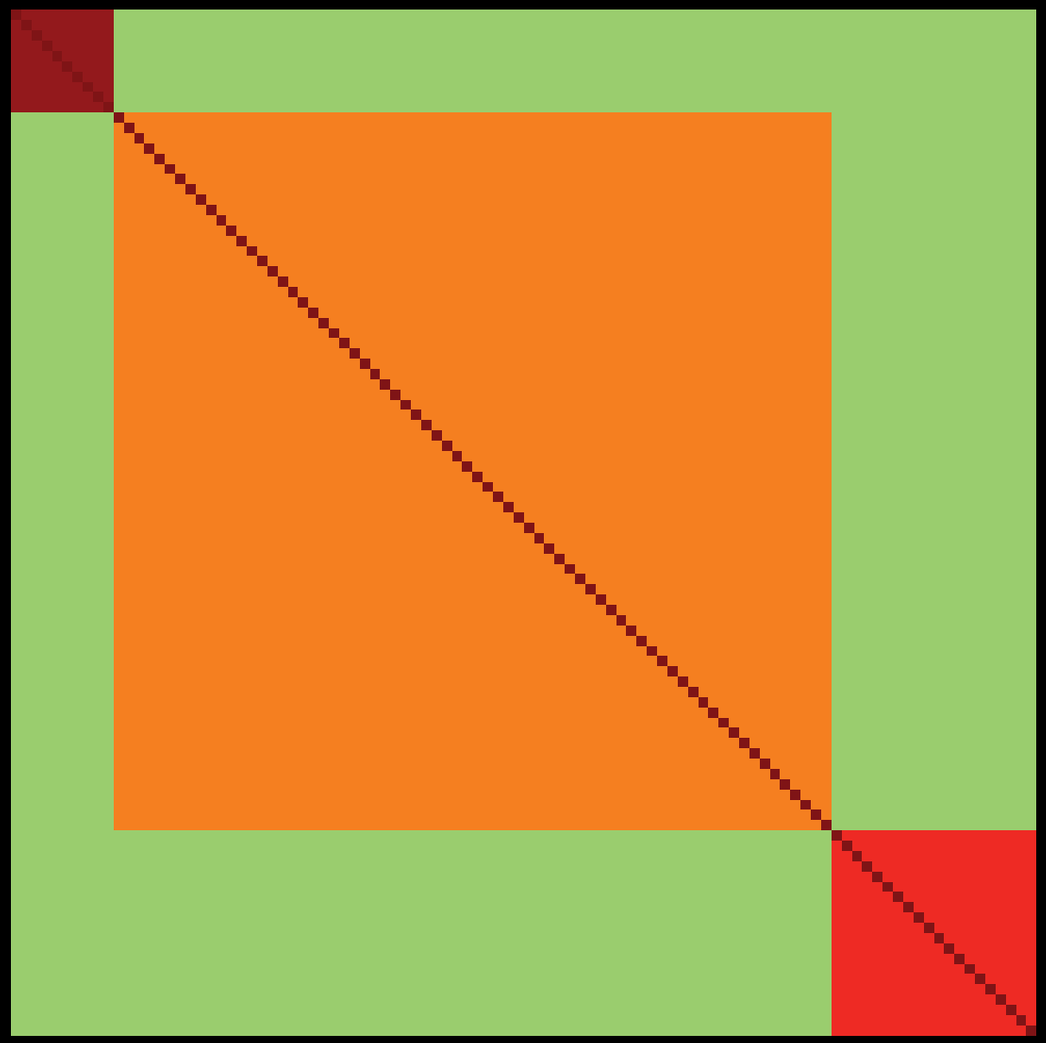
II. Interpreting covariance

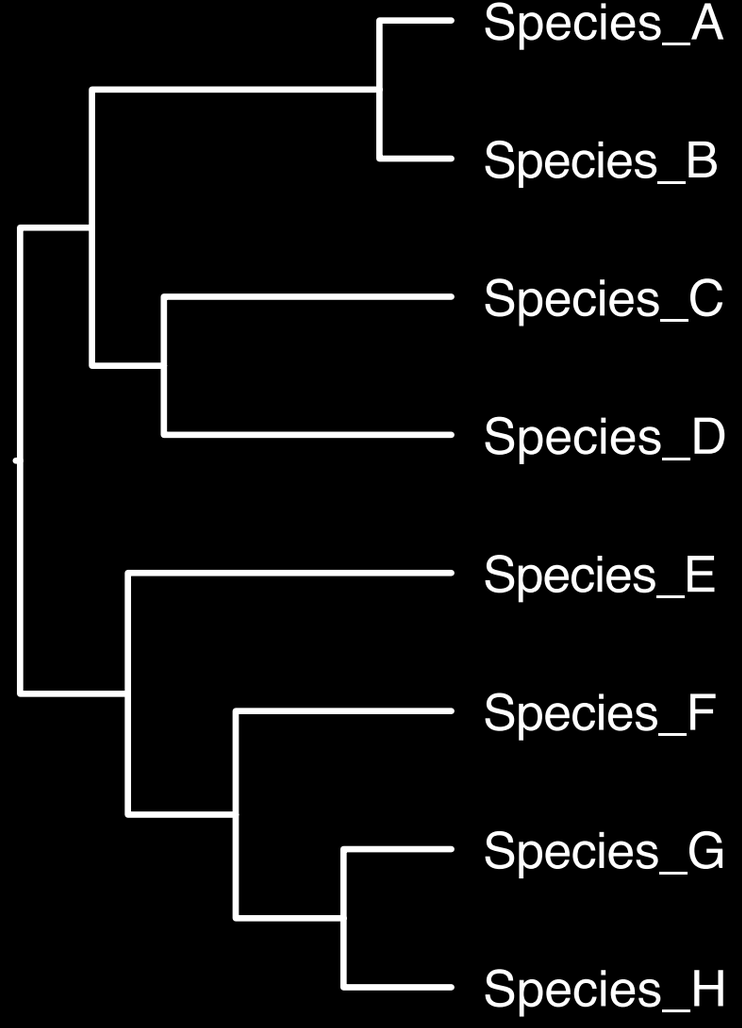


Covariance matrix Scale

II. Interpreting covariance

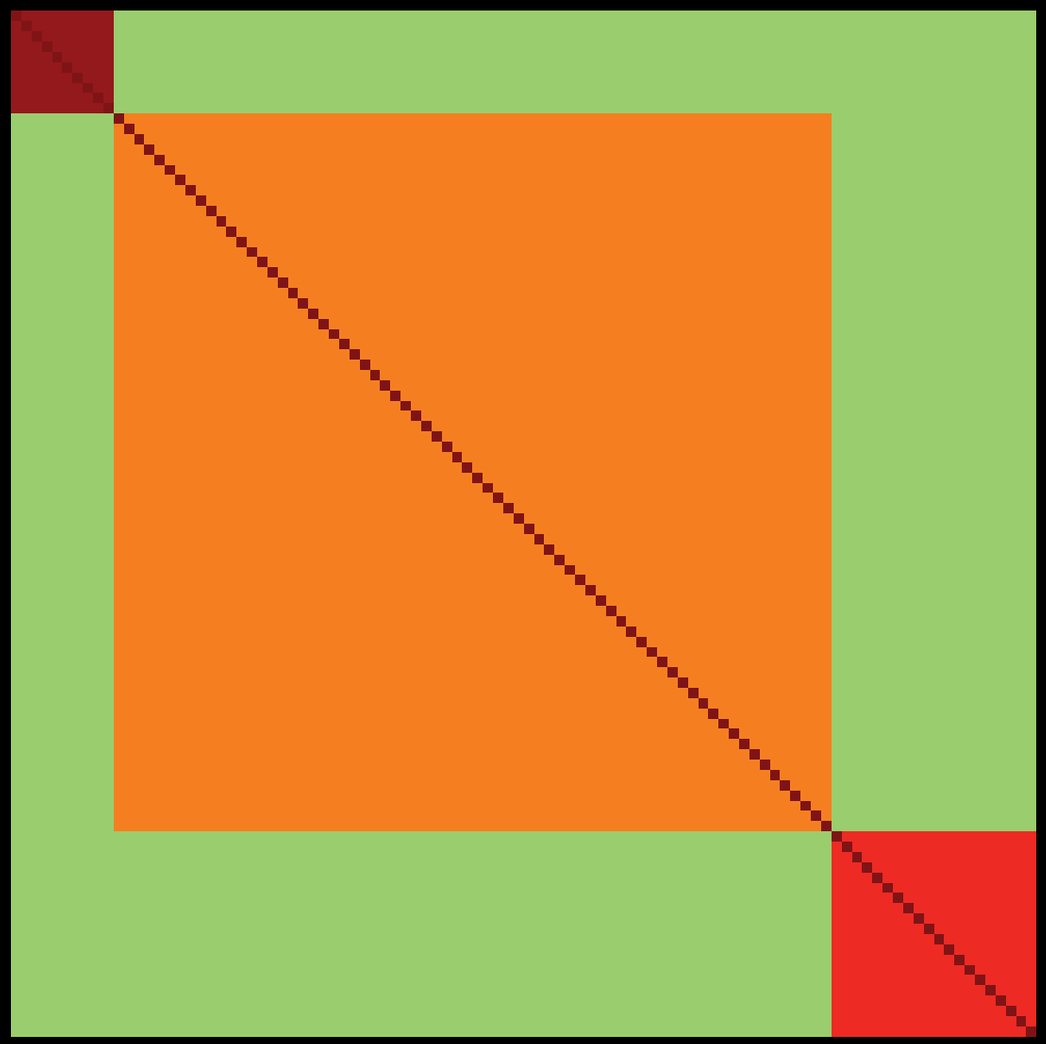
Simulate evolution of these 100 genes on a tree of 8 species



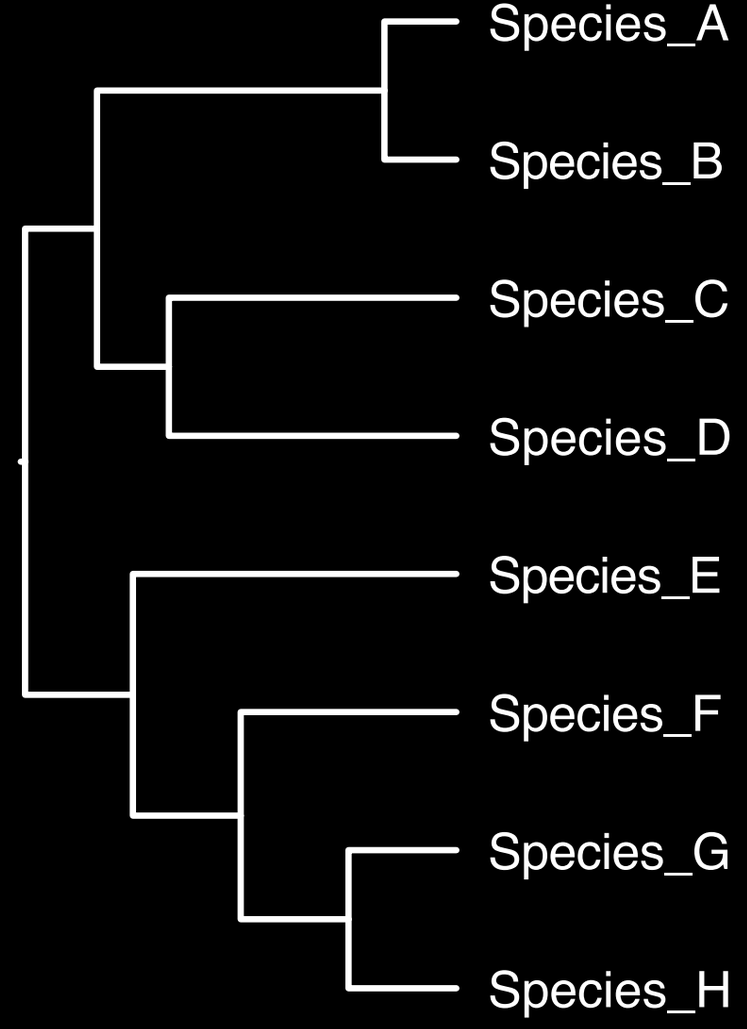


II. Interpreting covariance

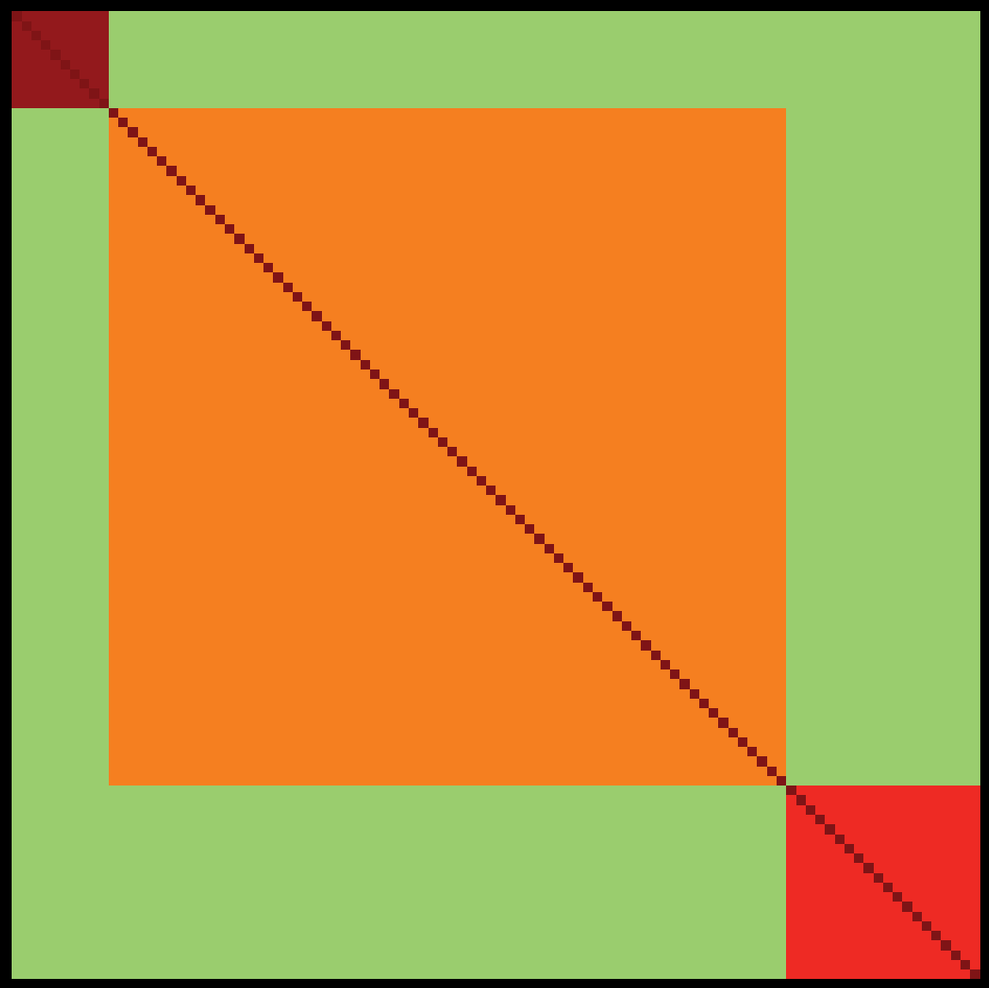
p=100



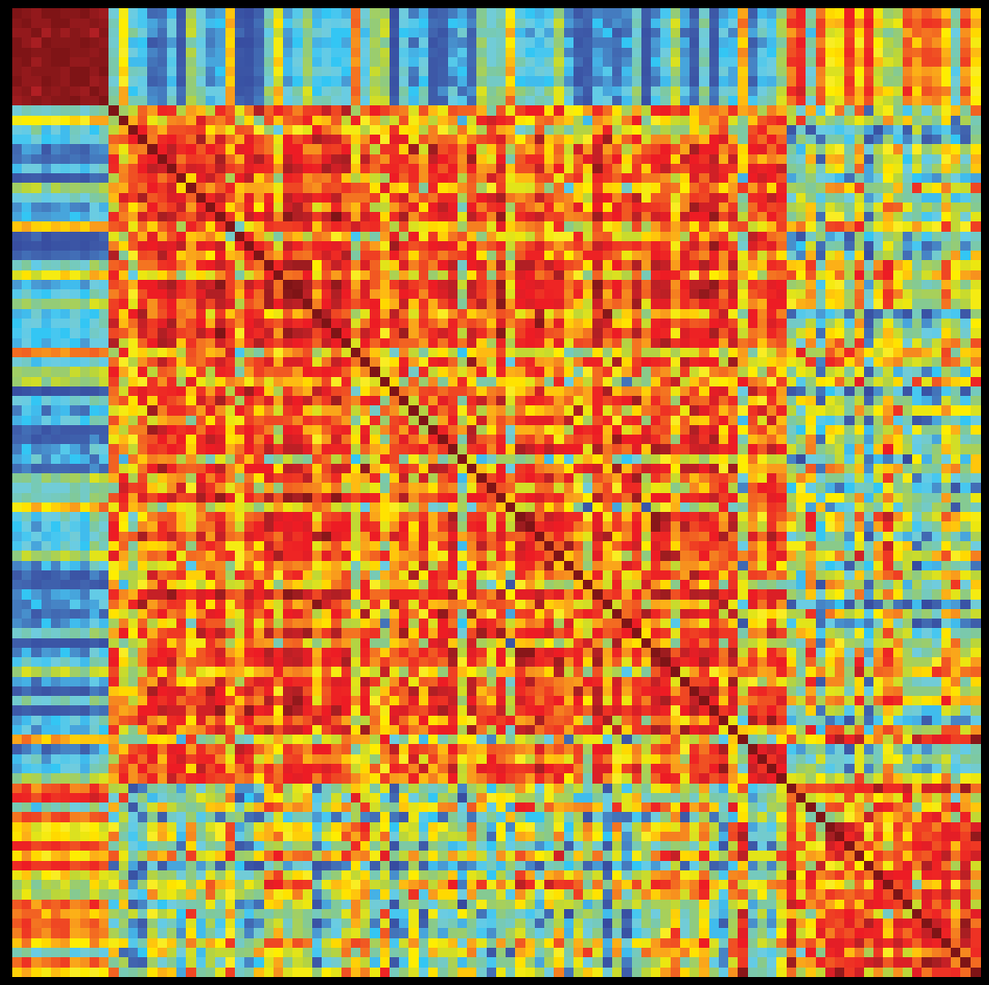
n=7

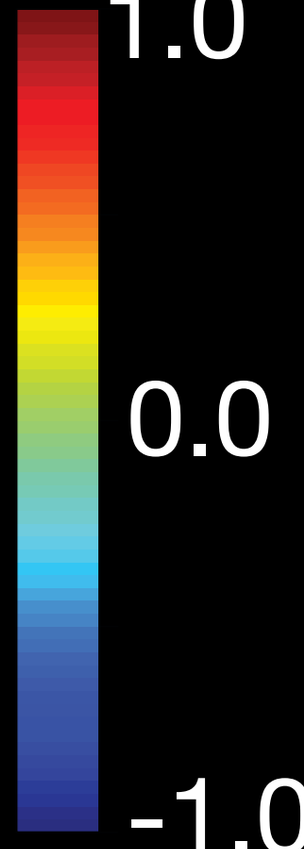


II. Interpreting covariance



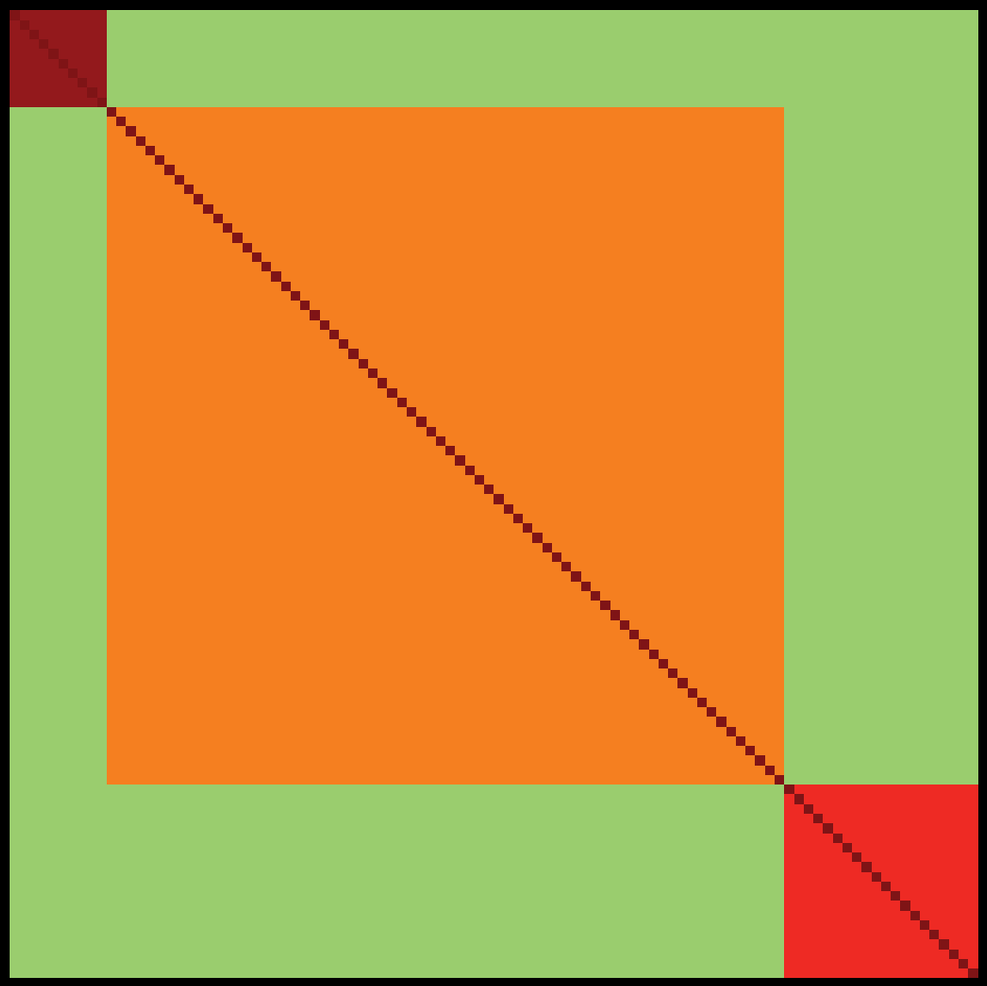
"True"



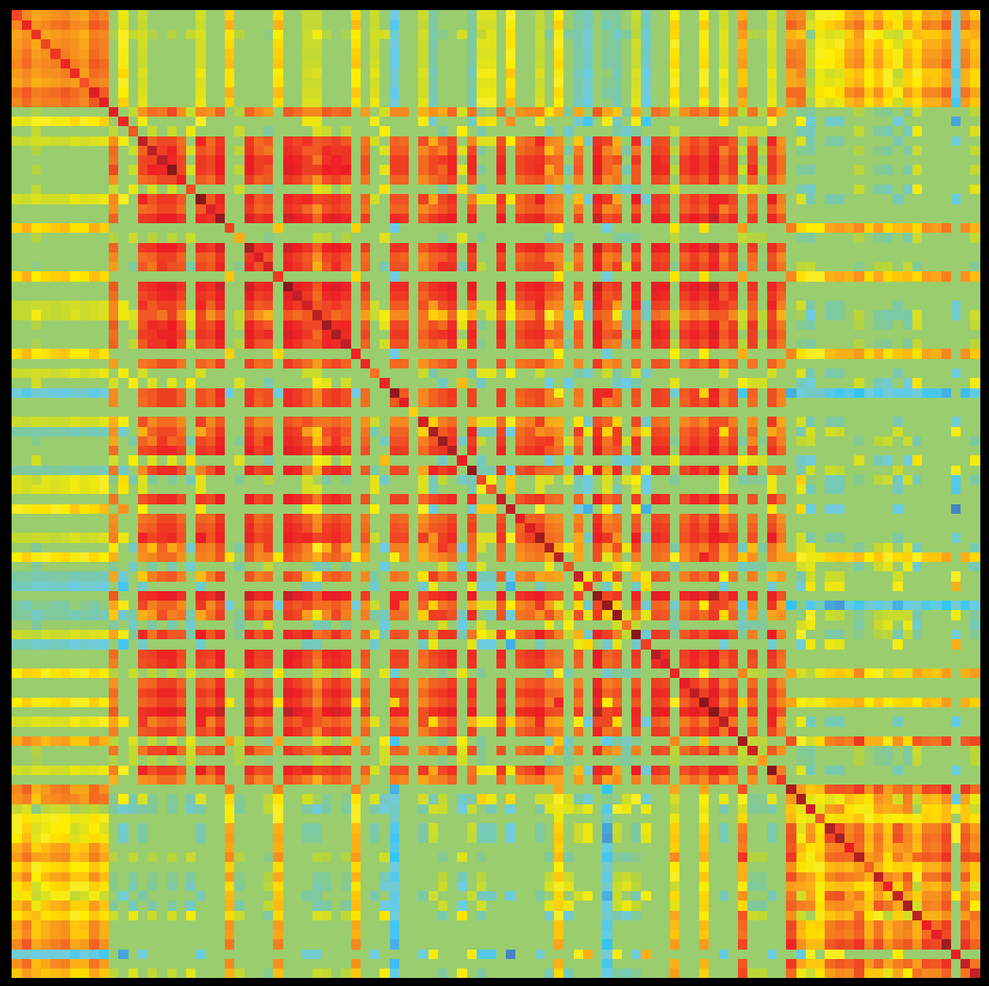


Independent contrasts only

II. Interpreting covariance

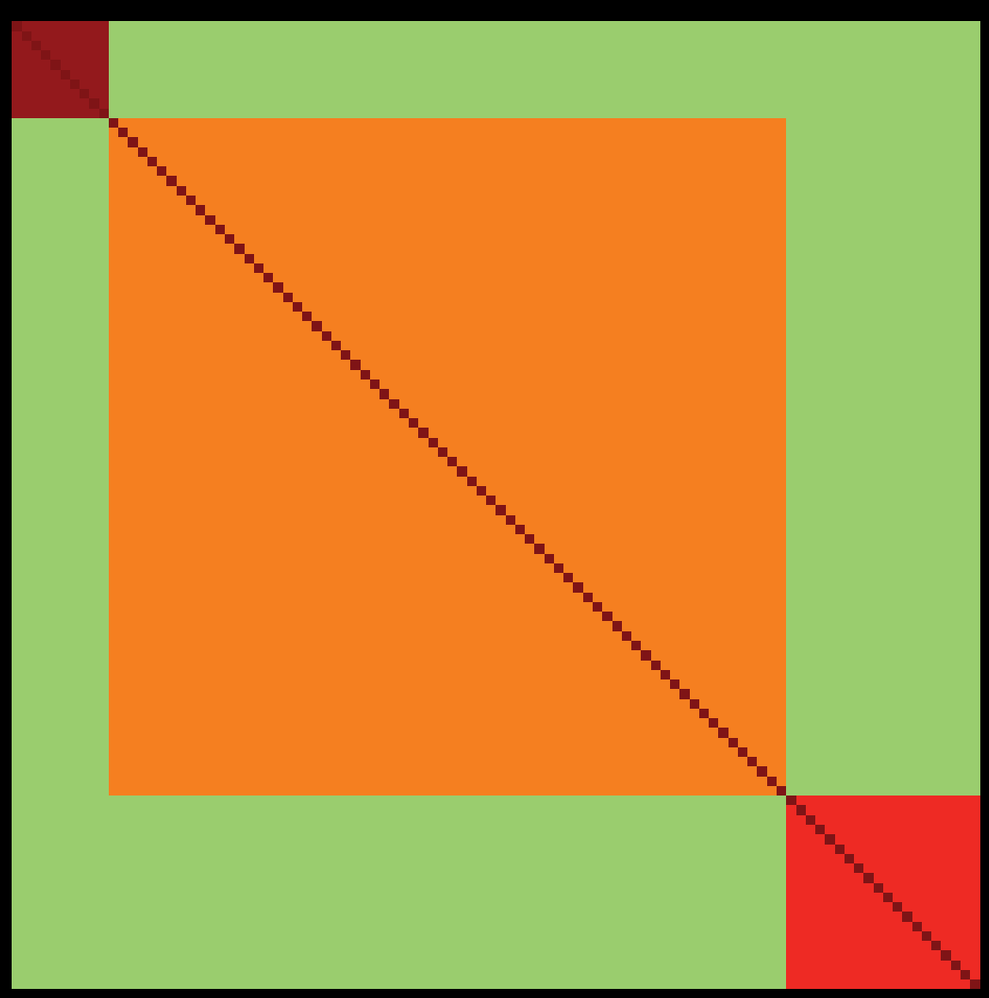


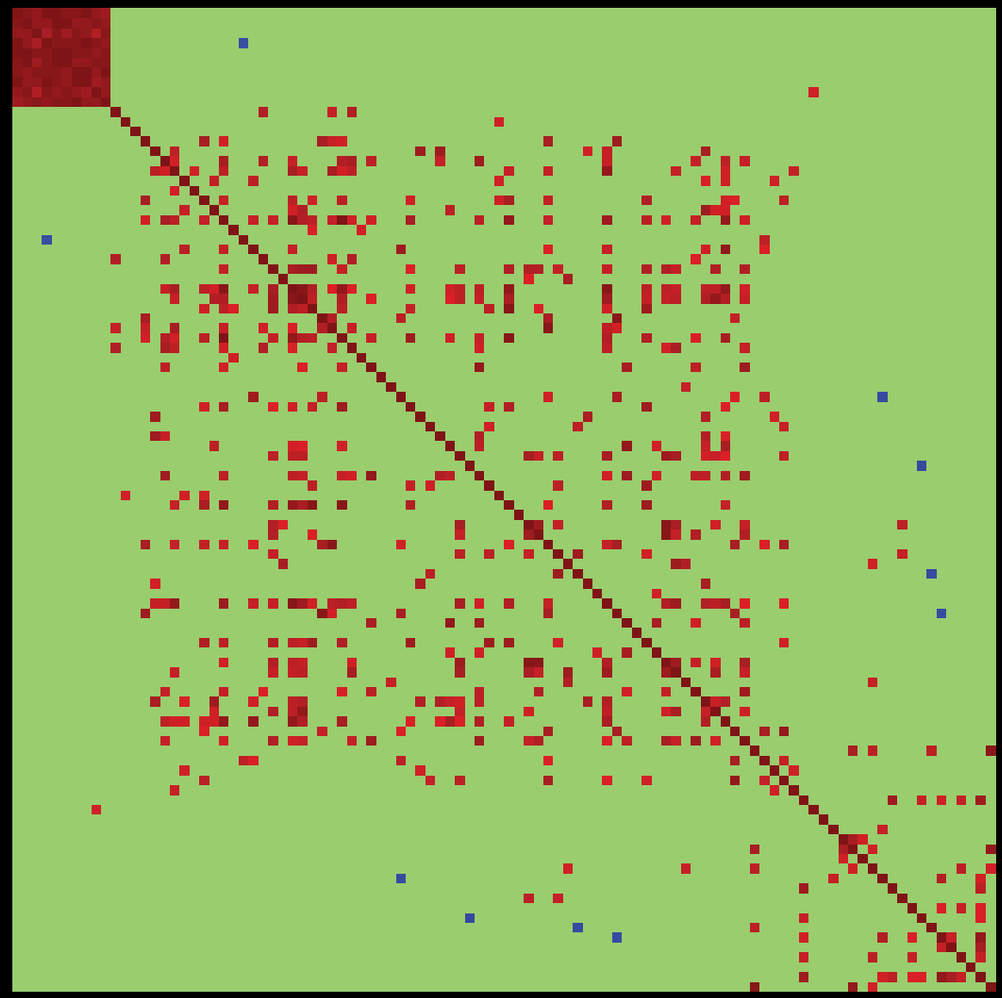
“True”



Regularization
(Luo, 2012)

II. Interpreting covariance





"True"

Regularization
(Bickel & Levina 2008)

II. Interpreting covariance

Take home:

There is no getting around missing information when $n << p$ but false positives can be mitigated through regularization

Visualization

An interactive tree

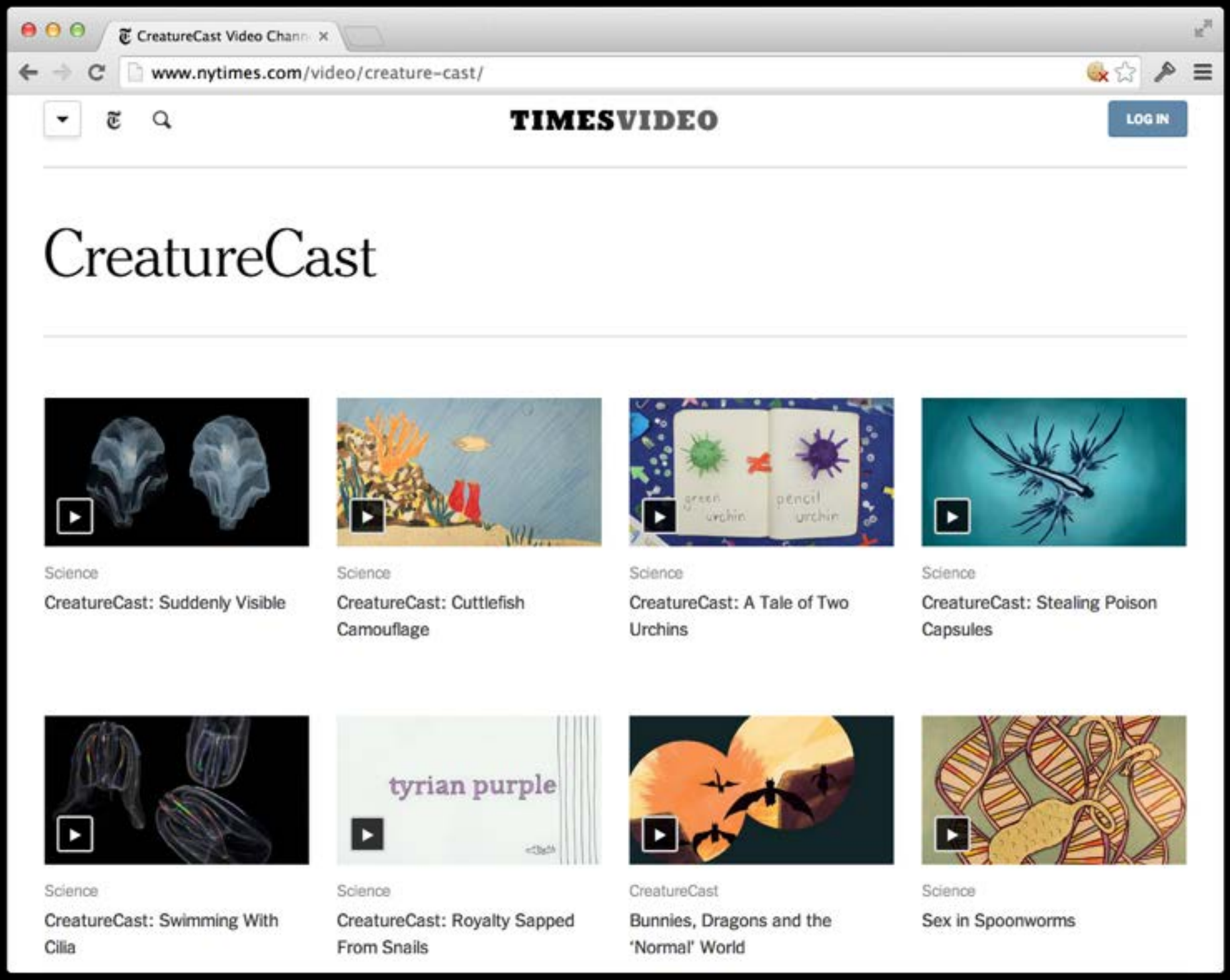
Use it:
http://dunnlab.org/phylotree

Watch a demo:
https://vimeo.com/67665449

Play with the code:
https://github.com/vhsiao/phylotree

We make cartoons

http://nytimes.com/creaturecast



Building skills

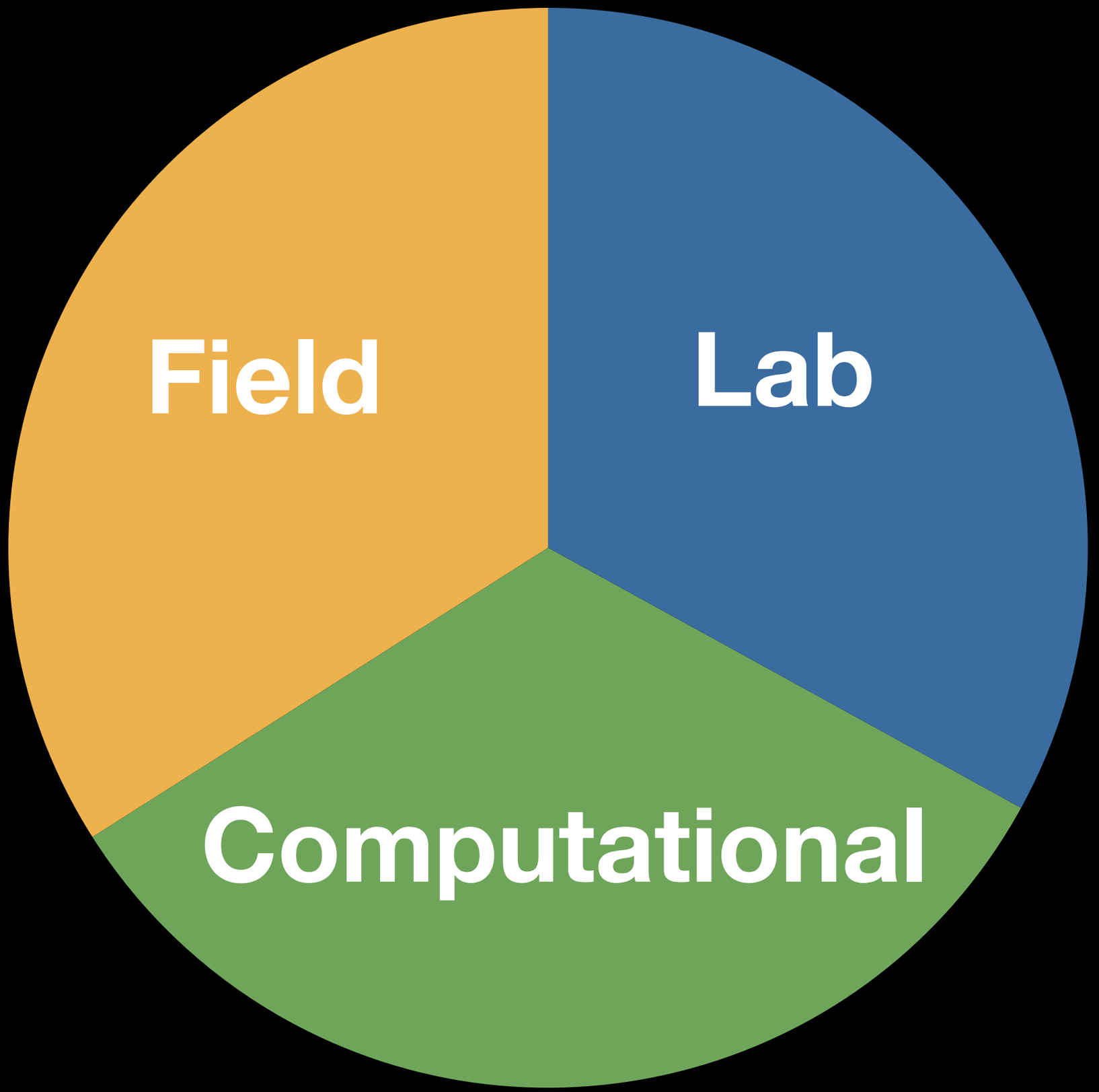
“Routine” phylogenetic analyses now require many skills that biologists are rarely trained in.

High throughput sample preparation

Programming

High performance computing

Stats beyond Sokal and Rohlf


Field
Lab
Computational

Computation

To use the command line.

Efficient text handling.

At least one programming language.

How to work on remote computers.

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.

Biology

Statistics

I posted my own handout/ cheat sheet:

https://bitbucket.org/caseywdunn/statistics/

Math

A little bit of linear algebra and graph theory will take you far in phylogenetics

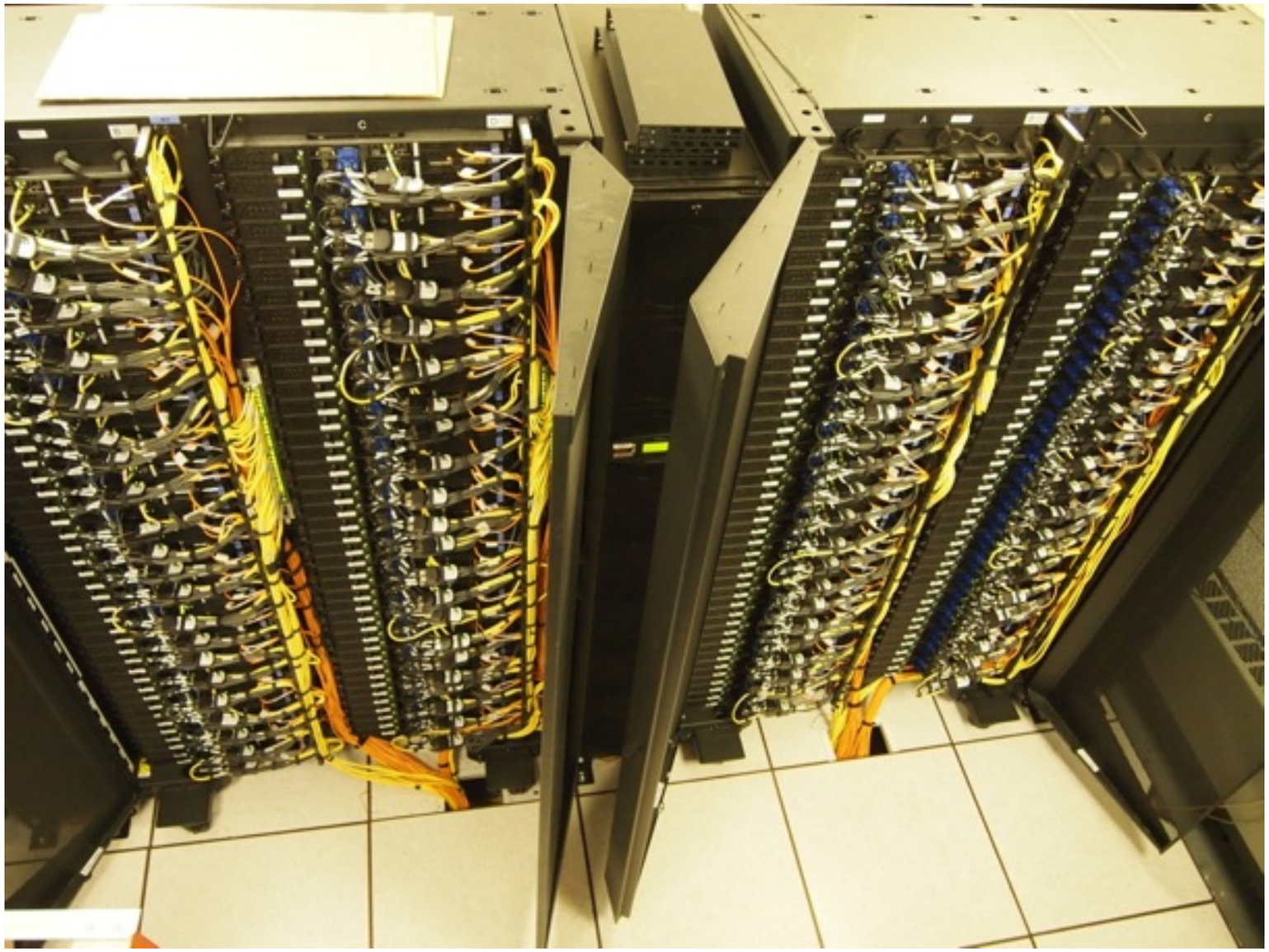
Managing your analyses

Organization is part of the analyses, rather than something that comes after

The data analysis ecosystem in my lab

- Central cluster
- Google docs
- git

Analyses and storage on cluster



git is a:

- Distributed software revision control system
- Allows you to organize all lab software in a single central repository
- Can write and use software in the repository on any computer

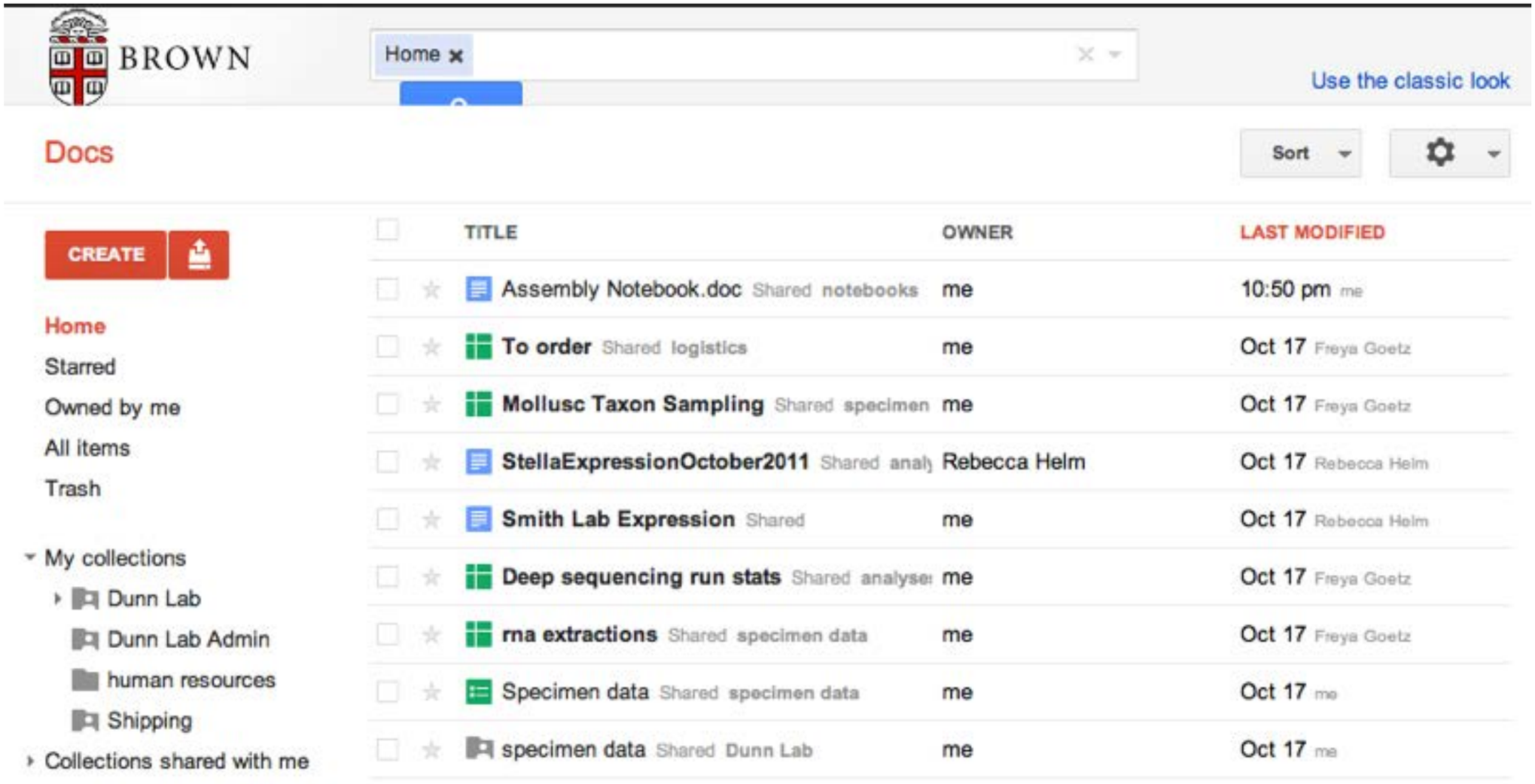
Documentation

Data and analyses are a liability rather than an asset if they aren't well documented

Documentation should be realtime, not something that is done after analyses

Good documentation is a powerful teaching and learning tool

Documentation on Google Docs



Or... literate code that serves as analysis tool and documentation in one.

See:

https://bitbucket.org/caseywdunn/phylogeneticbiology/src/master/analyses/good_programming