# Phylogenetic Biology
# Week 12

Biology 1425

Professor: Casey Dunn, dunnlab.org
Brown University

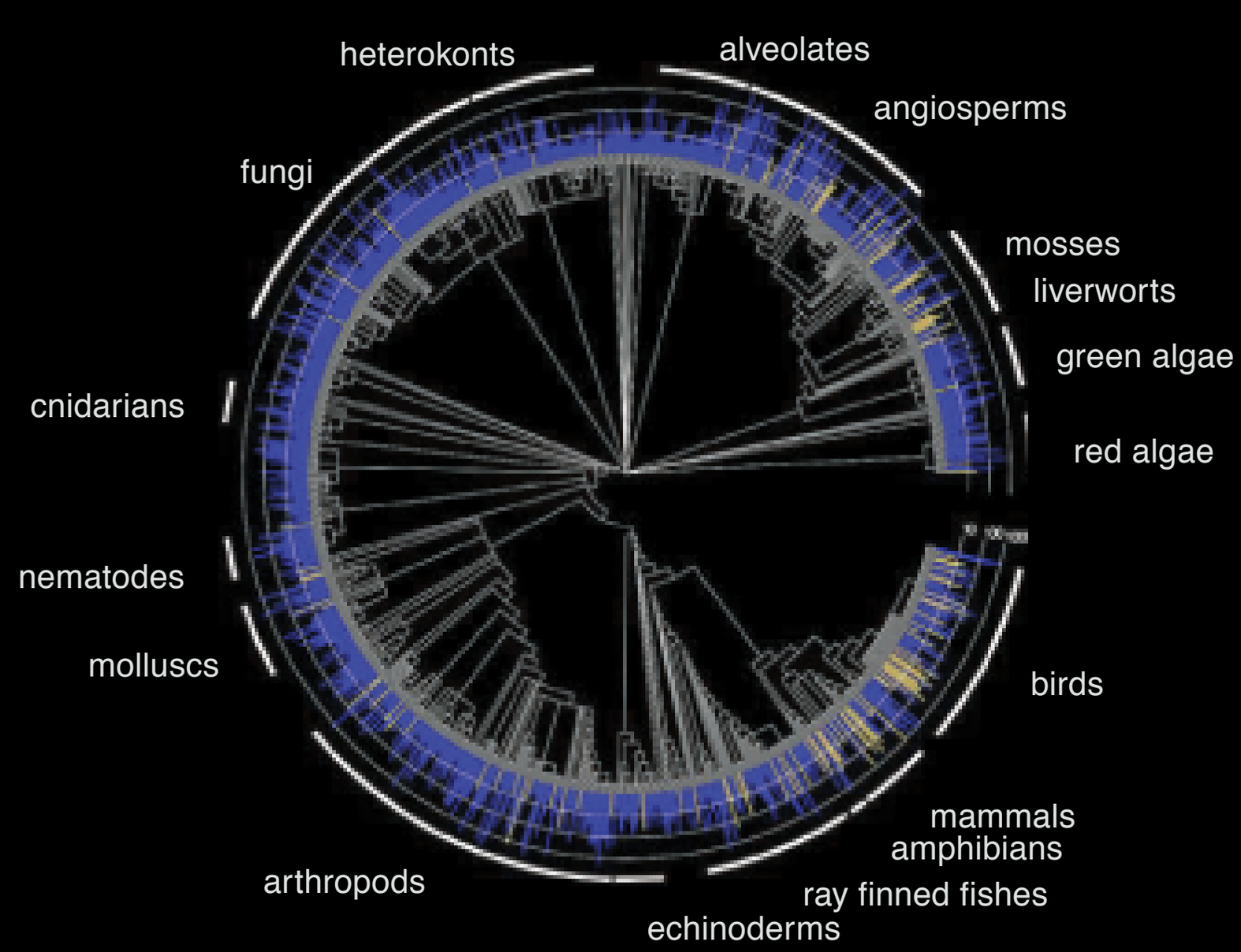
2013

# Front matter...

# Sources

Other non-original content is referenced by url or doi.

"...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

# Why collect data from lots of genes?

- Many hard problems will require lots of data
- Lots of data makes some aspects of inference easier

# 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

# 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

# What does “phylogenomics” mean?

# What does “phylogenomics” mean?

1. The study of genome evolution in a phylogenetic context

# What does “phylogenomics” mean?

1. The study of genome evolution in a phylogenetic context

2. The inference of species phylogenies with genome data

# What does “phylogenomics” mean?

1. The study of genome evolution in a phylogenetic context

2. The inference of species phylogenies with genome data

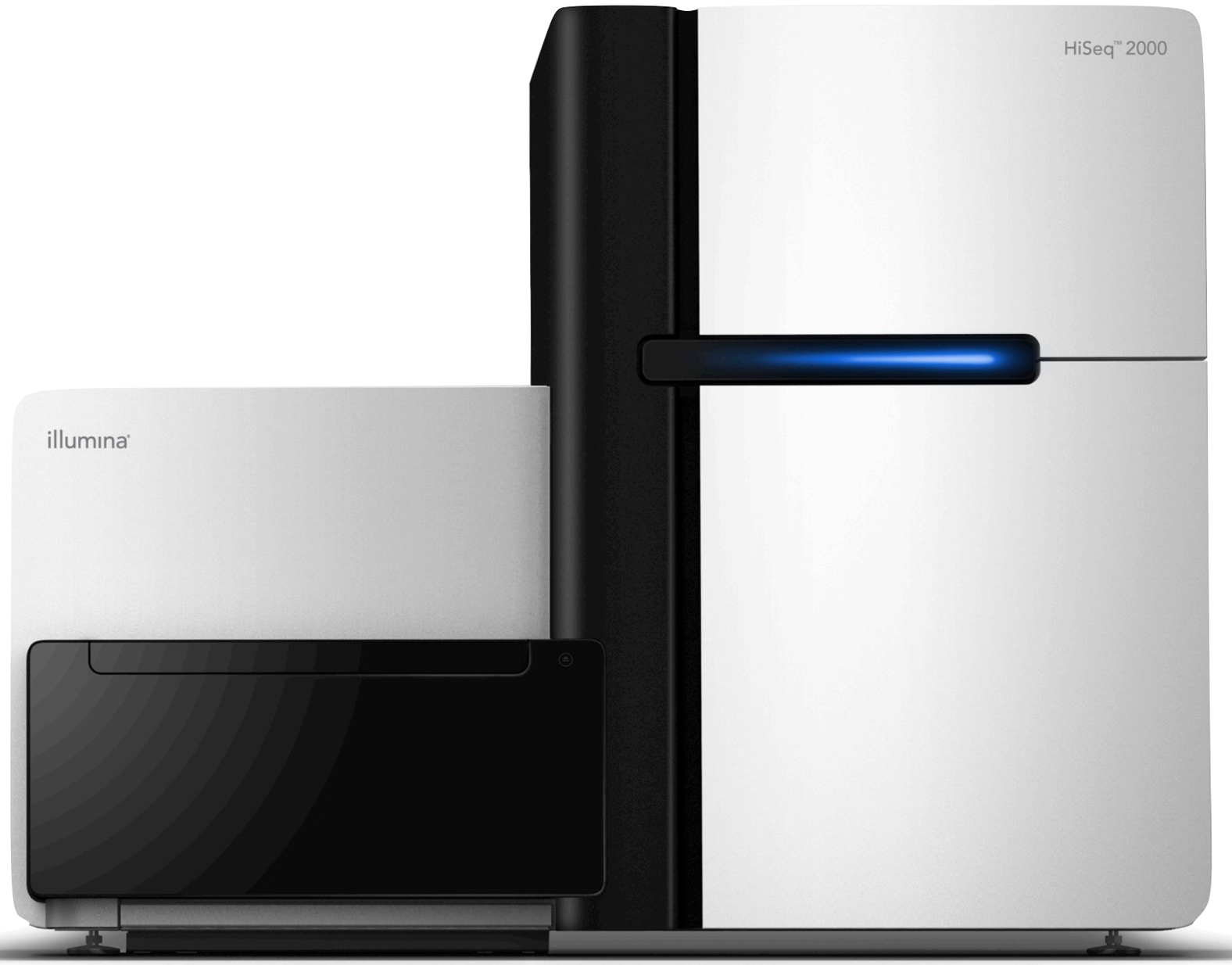
3. The inference of species phylogenies with data from lots of genes

# Inferring phylogenies

1. Collect raw data
2. Process into character data
3. Identify homologous characters
4. Identify characters relevant to problem
5. Build trees
6. Evaluate trees
7. Use trees

# Collecting sequence data

# DNA sequencing

# Current Illumina costs:

~ $2,000 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 |

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

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?

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

This is a false dichotomy.

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?

Are other types of data now obsolete?

No!

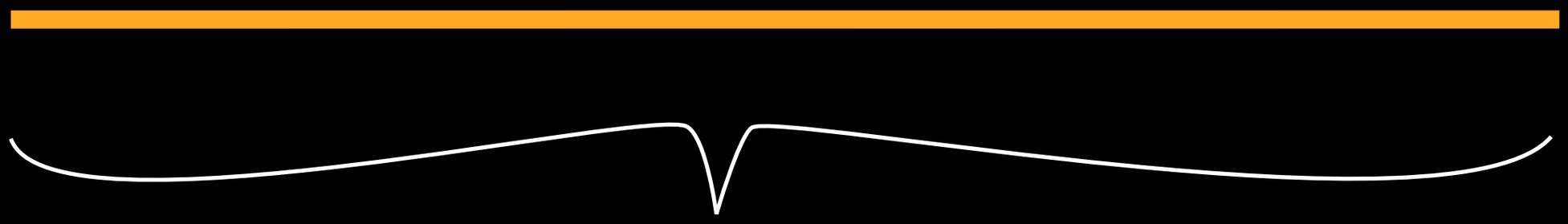
# Are other types of data now obsolete?

No!

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

# DNA sequencing


Illumina



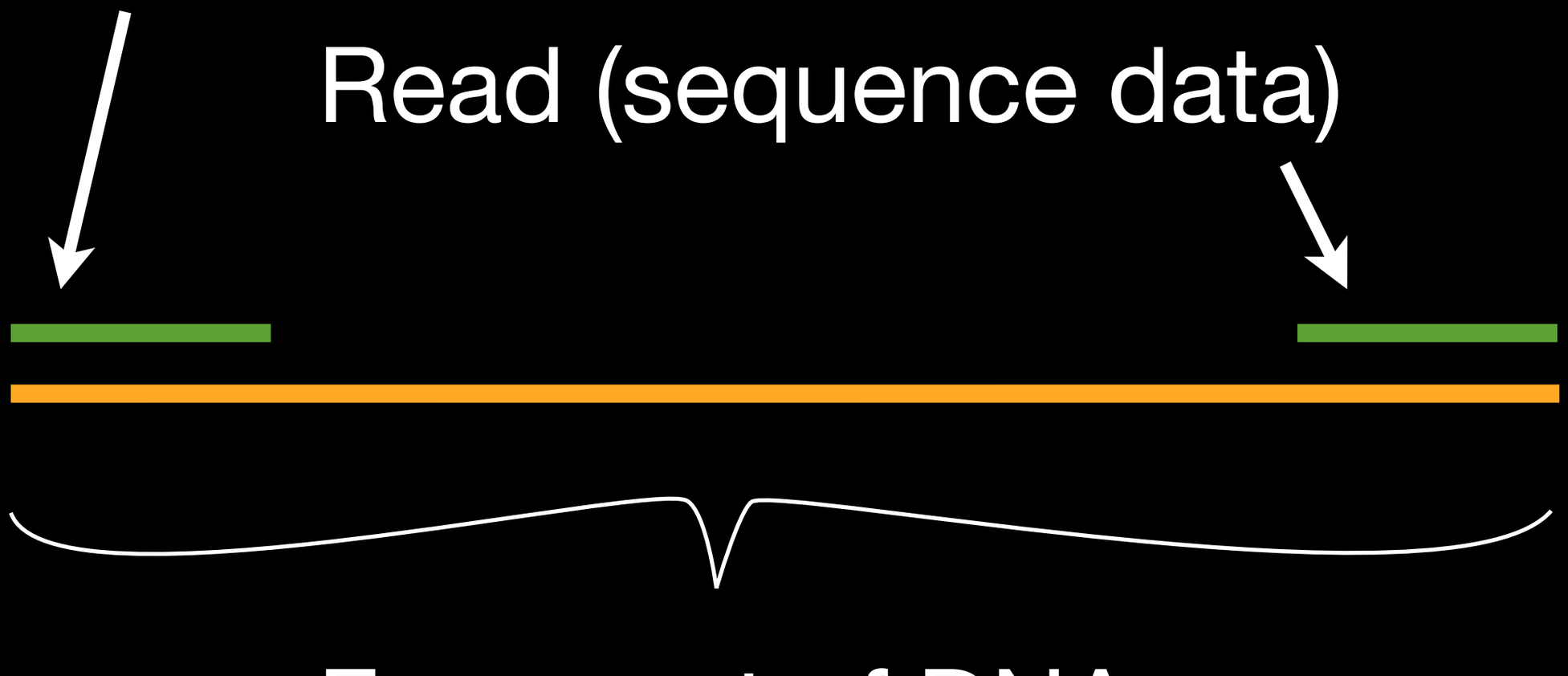
# Fragment of DNA

Read (sequence data)

Fragment of DNA

Read (sequence data)

Read (sequence data)



Fragment of DNA

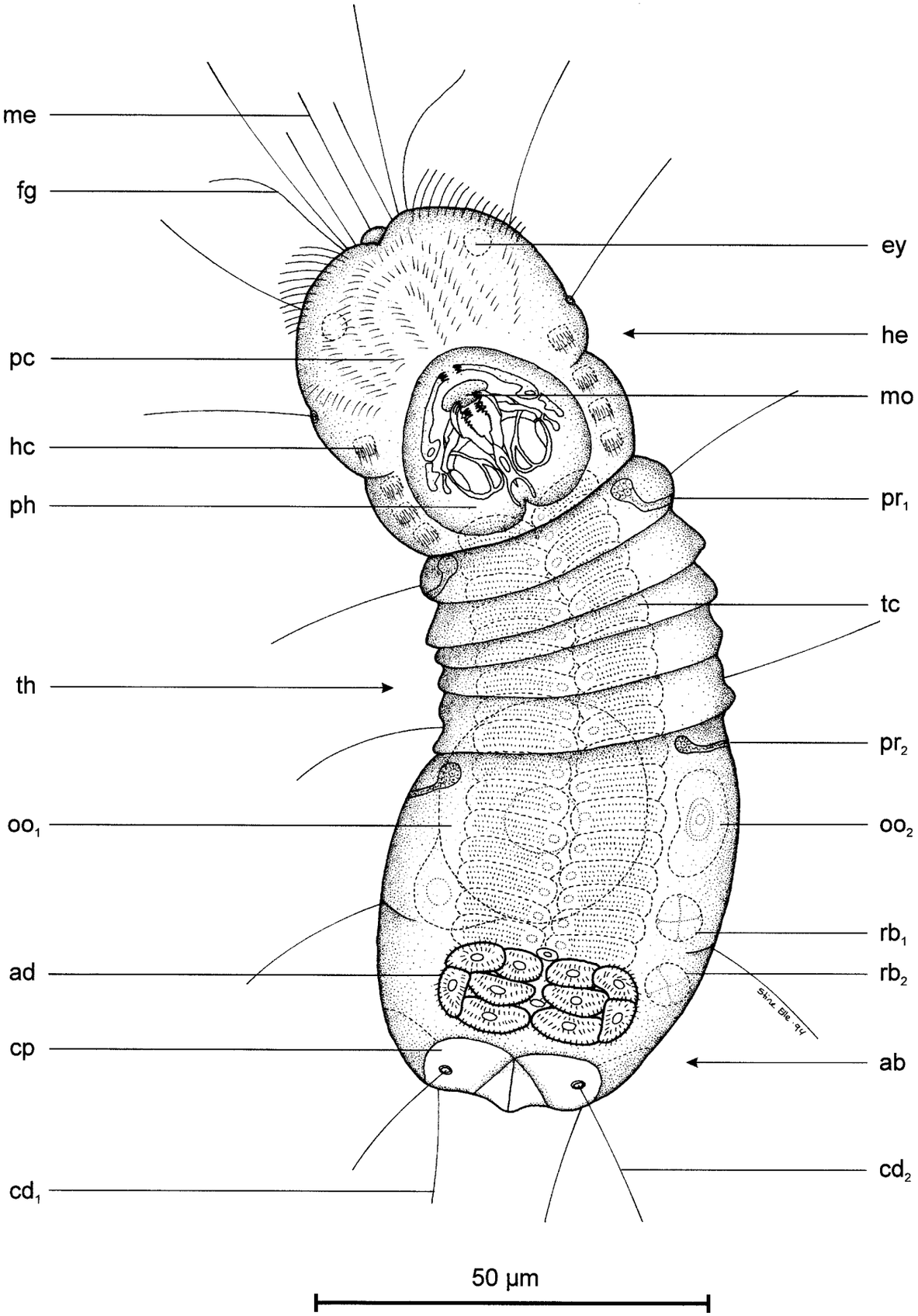
# DNA Fragments can be:

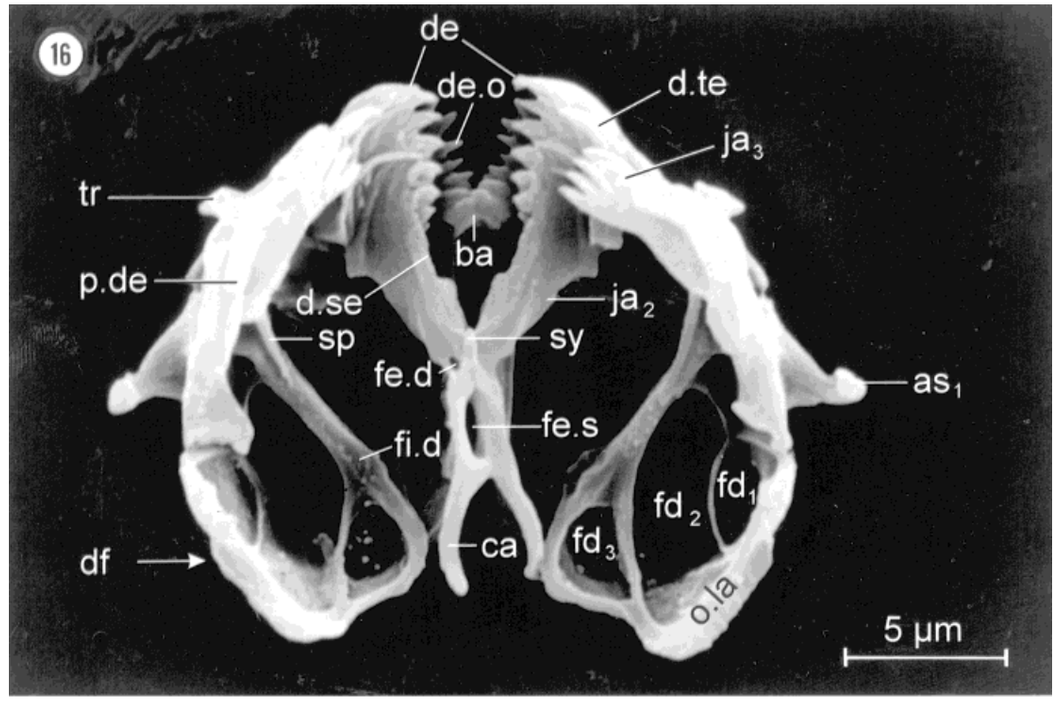
- Amplified/ enriched gene regions
- Genomic DNA
- cDNA (Transcriptomes)

| 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



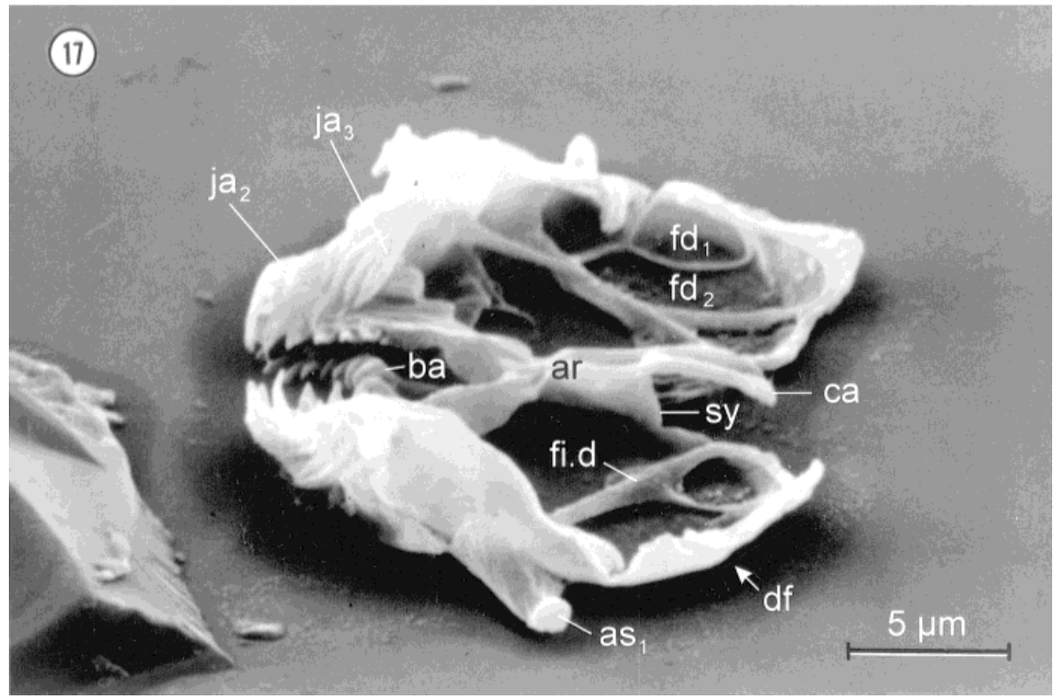
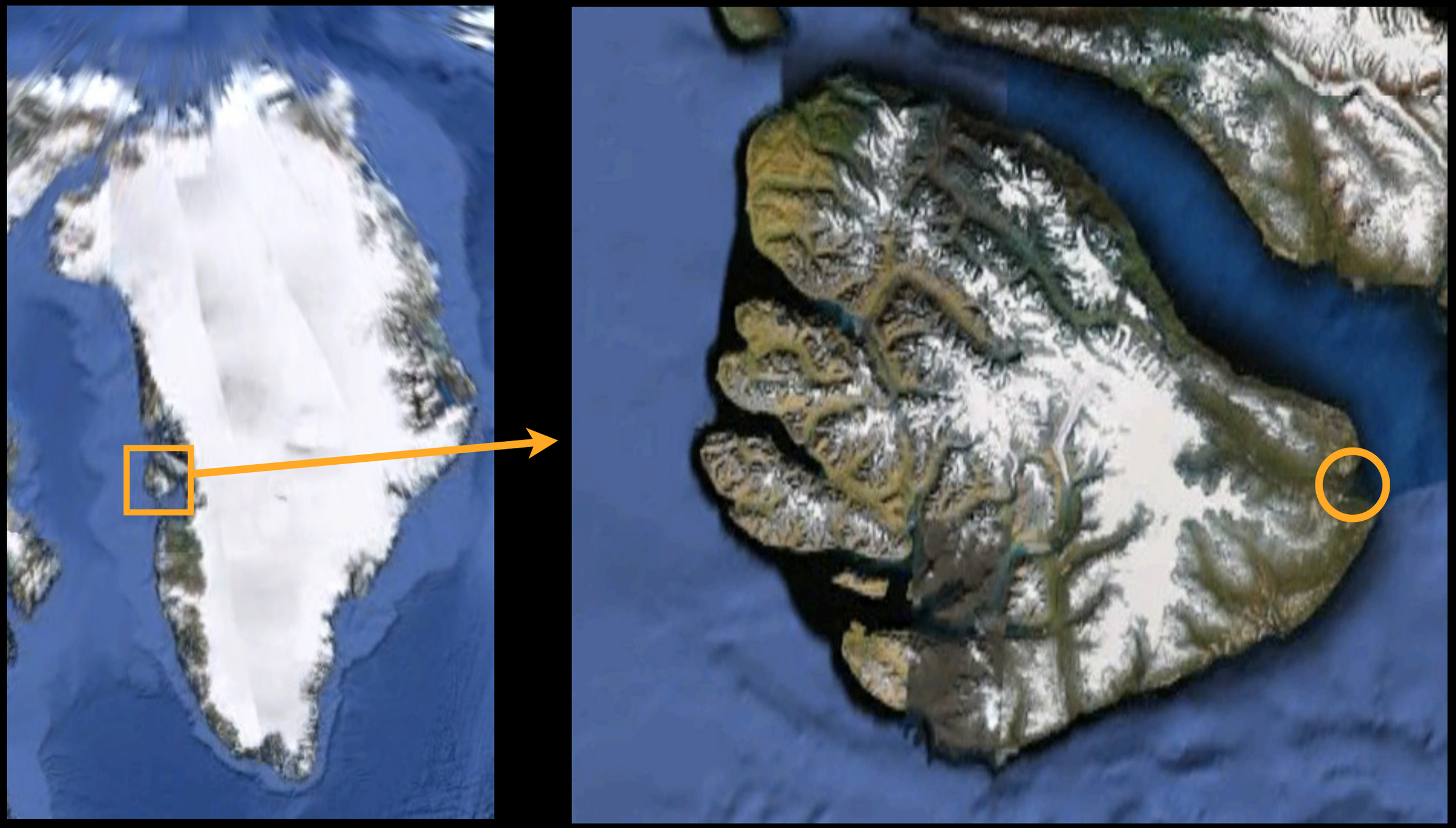




Figure 2

(Kristensen and Funch, 2000)




SIGMA
Nikon
prefectionist.

# 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

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

# 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

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

Break it into little pieces!

DNA


Read
Read
DNA


Read
Read
DNA

Starting material

Fragment

Prepare library, sequence

# Library preparation options

# Library preparation options

- Get a library preparation kit from the sequencer vender

# Library preparation options

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

# Library preparation options

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

Illumina Software to support

www.brown.edu/Research/CGP/core/illumina/

Division of Biology and Medicine

*Center for Genomics and Proteomics*

# Genomics Core Facility

About | Contact | Illumina Sequencing | Equipment | Equipment Sign Up | Services | Acknowledgments | DNA Sanger Sequencing | Current Prices

» Biomed Research » Biomed Core Facilities

**Overview of Next-Gen Sequencing**

**HiSeq2000/GAIIx**

**Listserv/ Discussion Group**

**Timeline for Implementation**

**Price Structure**

**Sample requirements**

**Covaris 220 Recipes and Guides**

**Sample Submission Form**

**Bioinformatics**

**Illumina Software to support the GAII-X system**

**Seminars**

## Welcome!

Thank you for your interest in Next Generation Sequencing at Brown University. The following information is intended to provide users with an introduction on how to get started with plans for a sequencing project.

The Genomics Core Facility, located at 70 Ship Street in the Laboratories for Molecular Medicine, is performing high throughput sequencing using Illumina instruments.

**The Genomics Core Facility introduced Next Generation Sequencing Service (NGS) in June 2010 with a GAIIx sequencer. In April 2011 we expanded our successful service and we now offer sequencing on a HiSeq2000 instrument in addition to the GAIIx. For more information on our instruments go to the link "GAIIx/HiSeq2000".**

Read about our implementation of the NGS services on the "Timeline for Implementation" link on the left. You can also view our GAIIx install progress on the "Track our Progress" link on the right.

› Track Our Progress

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

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
```

# Processing sequence data: Assembly

Starting material

Fragment

Prepare library, sequence

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

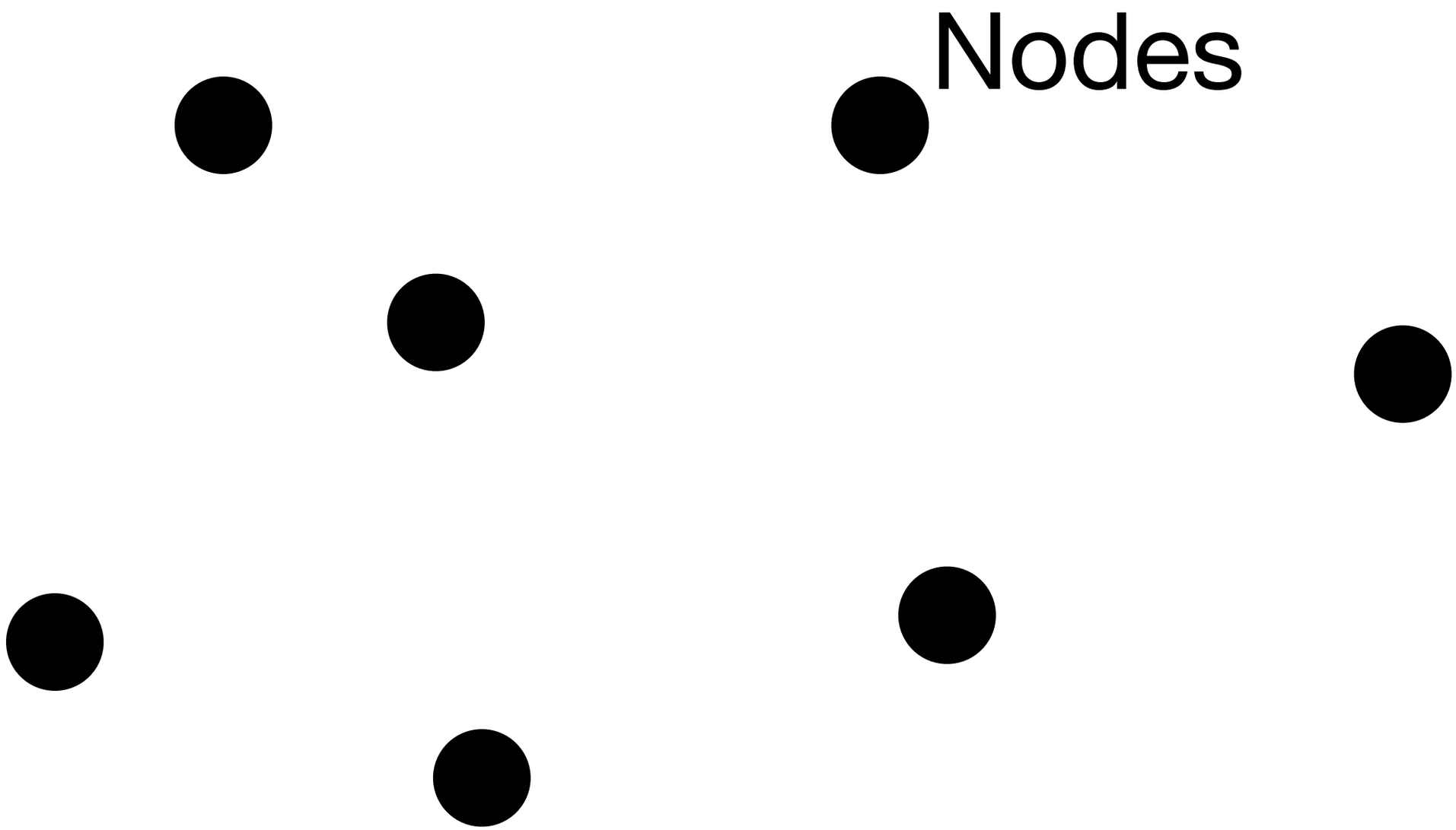
A new generation of de Bruijn graph assemblers have been developed to meet these challenges

Better defined memory footprint

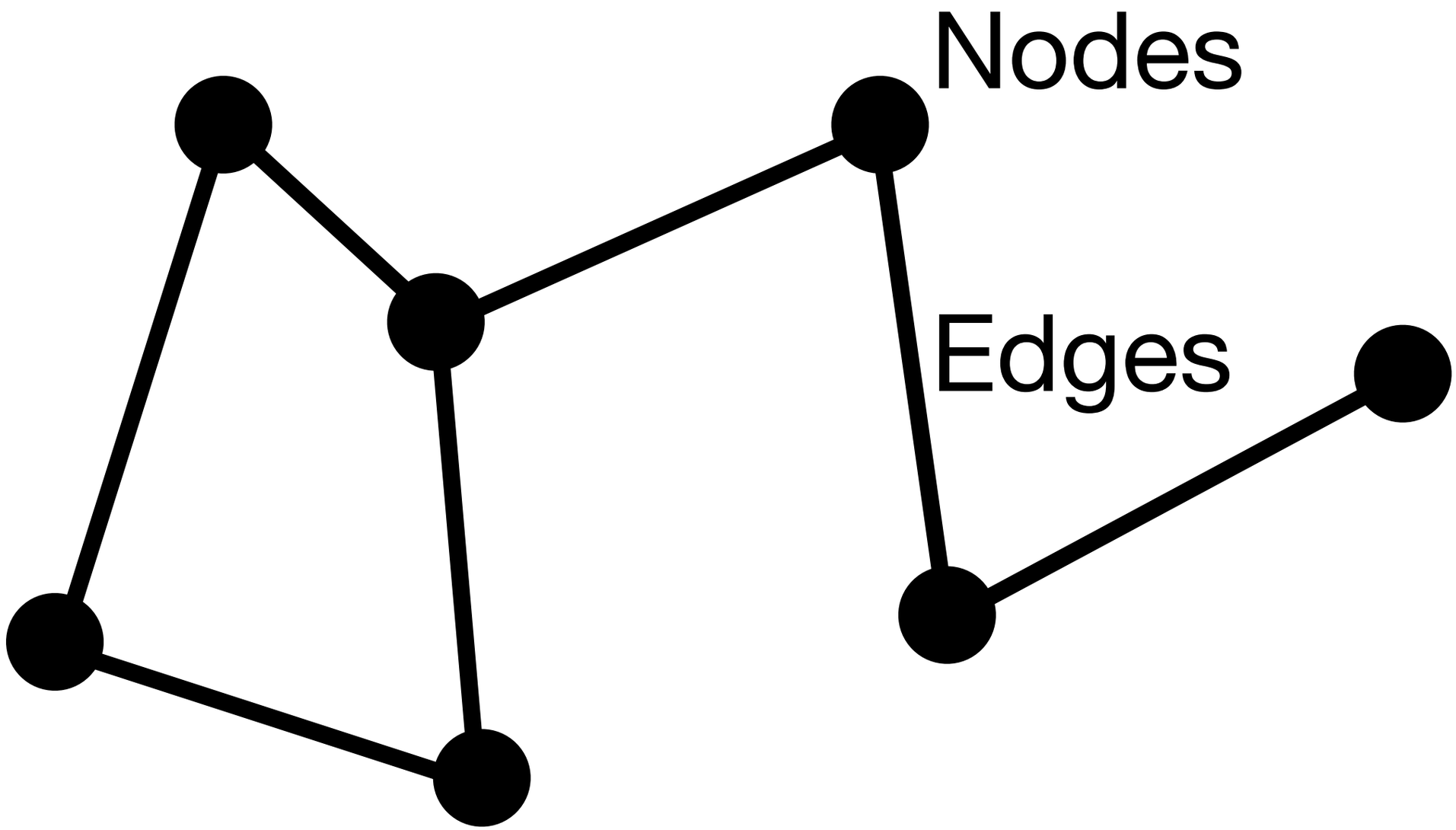
Simpler comparisons between sequences

# What is a graph?

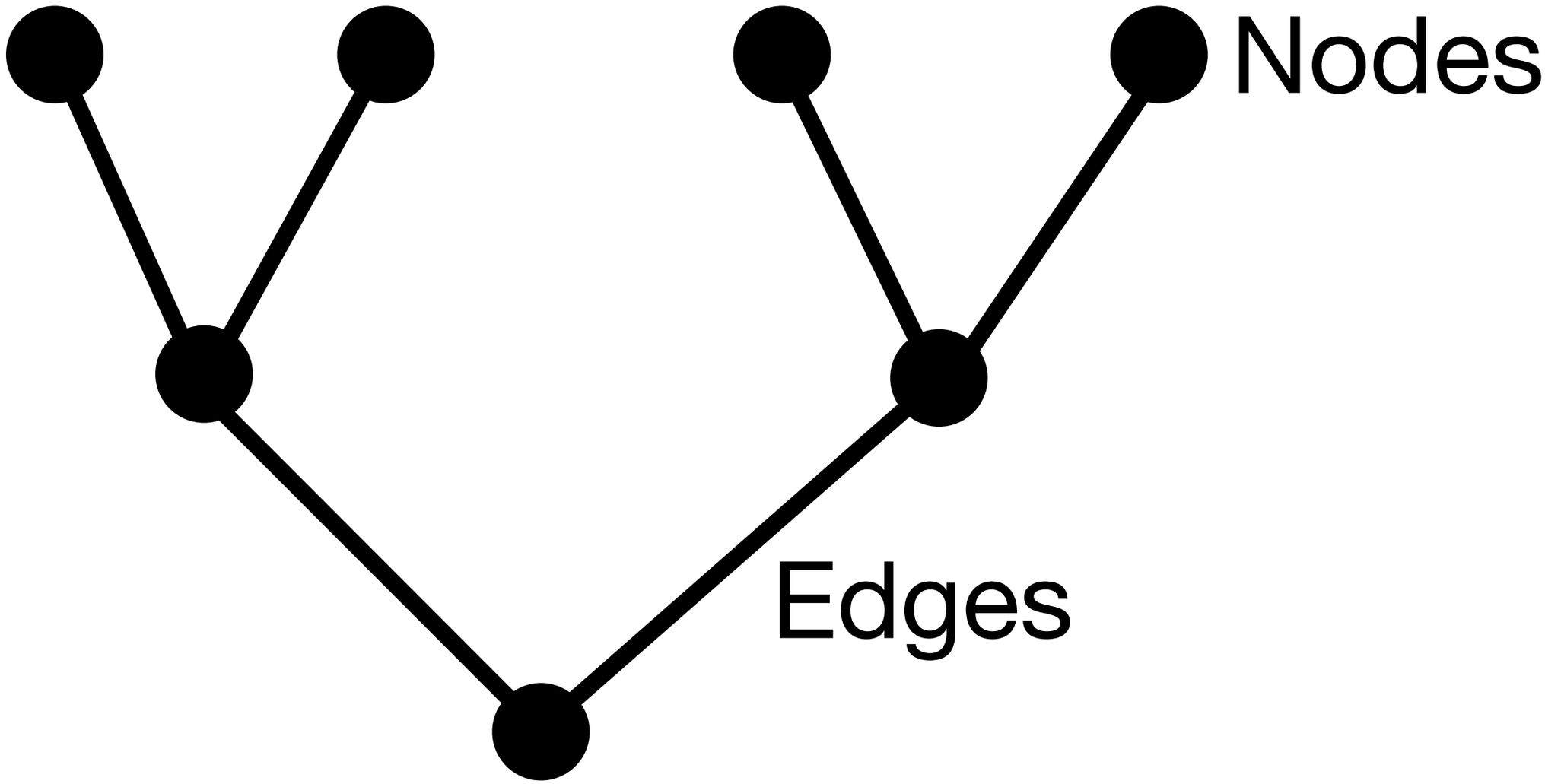
# What is a graph?



# 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

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

A de Bruijn graph is constructed from the has table

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

A de Bruijn graph is constructed from the has table

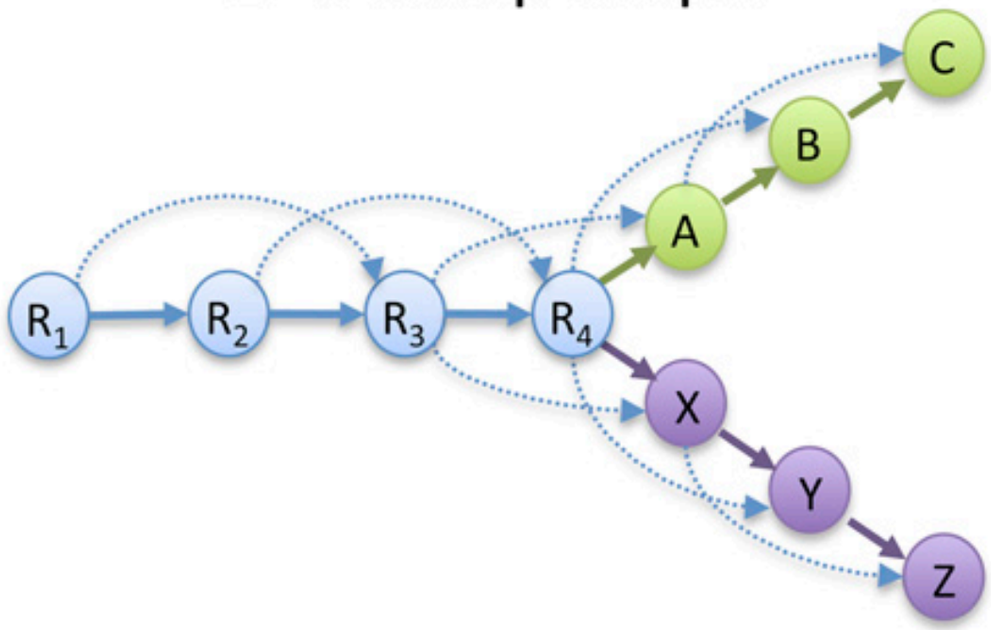
Each node corresponds to a k-mer sequence from the hash table

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

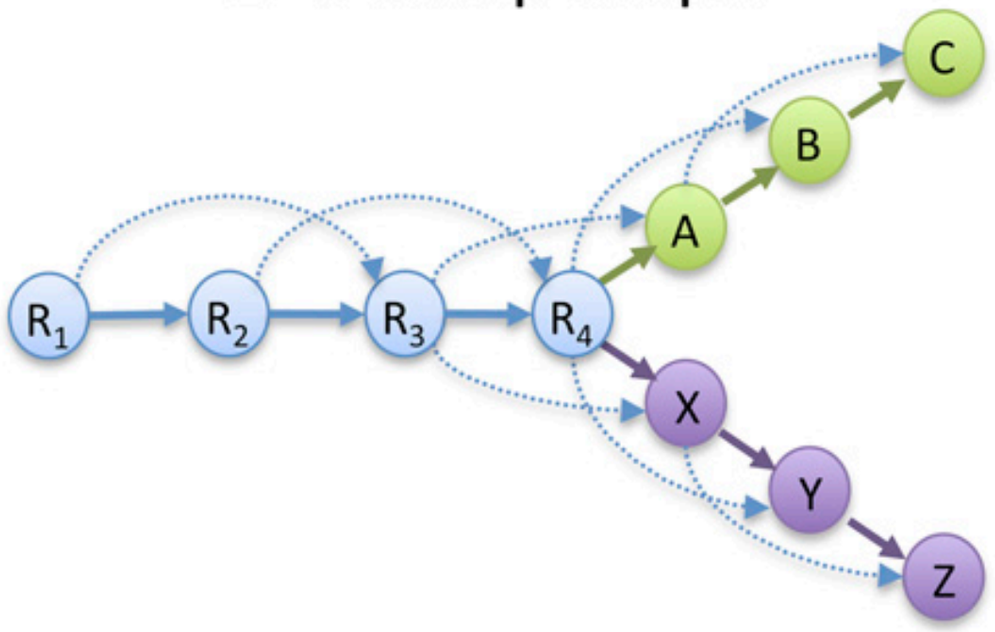



Schatz et al 2010 (dx.doi.org/10.1101/gr.101360.109)

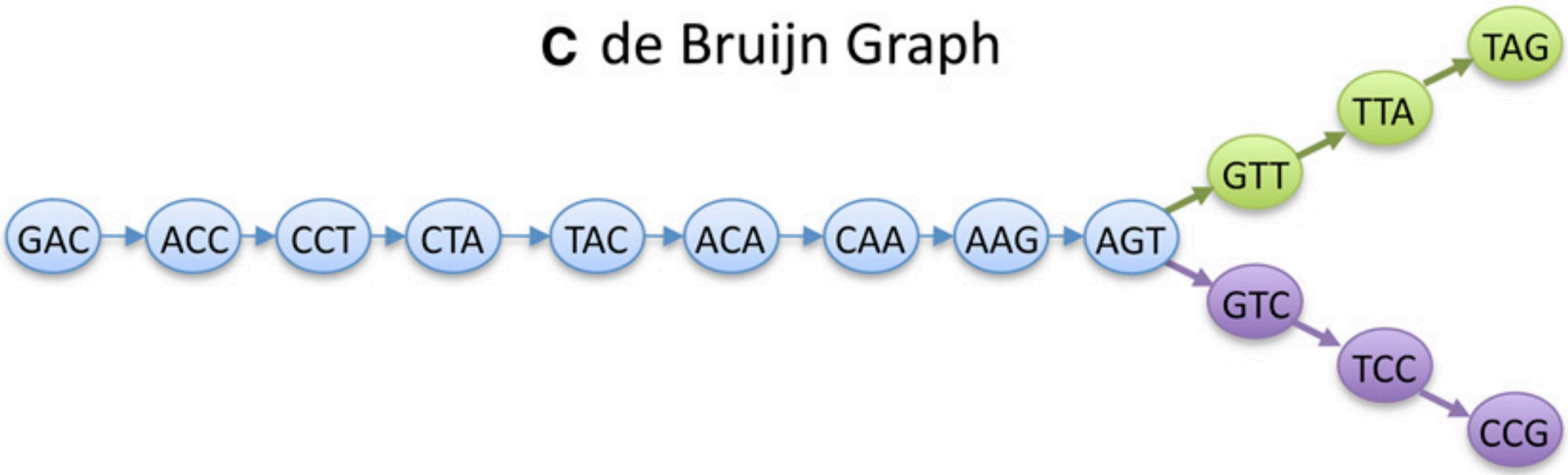
**A** Read Layout

```
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

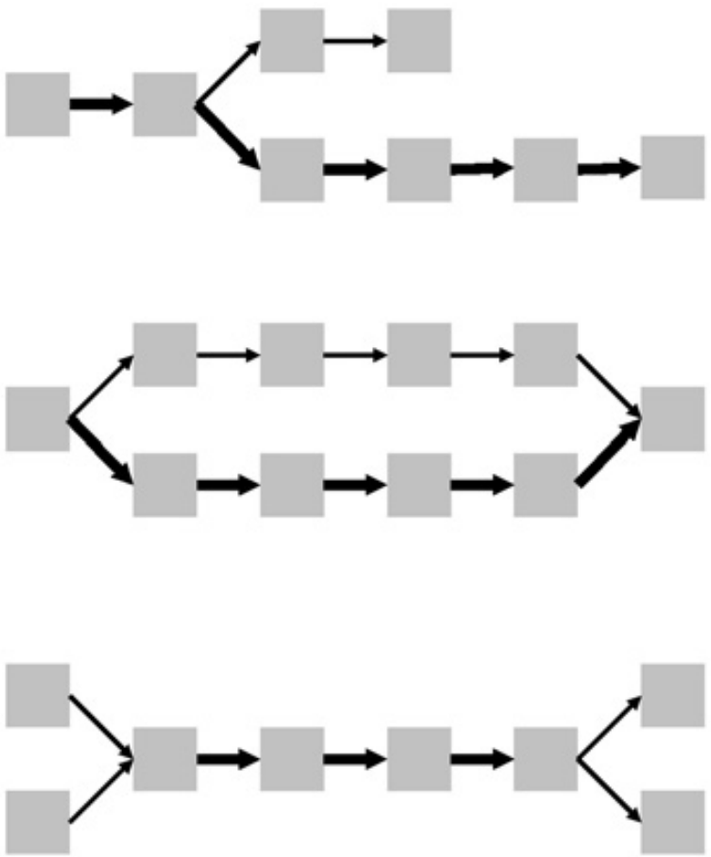
Paths through the de Bruijn graph are assembled sequences

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

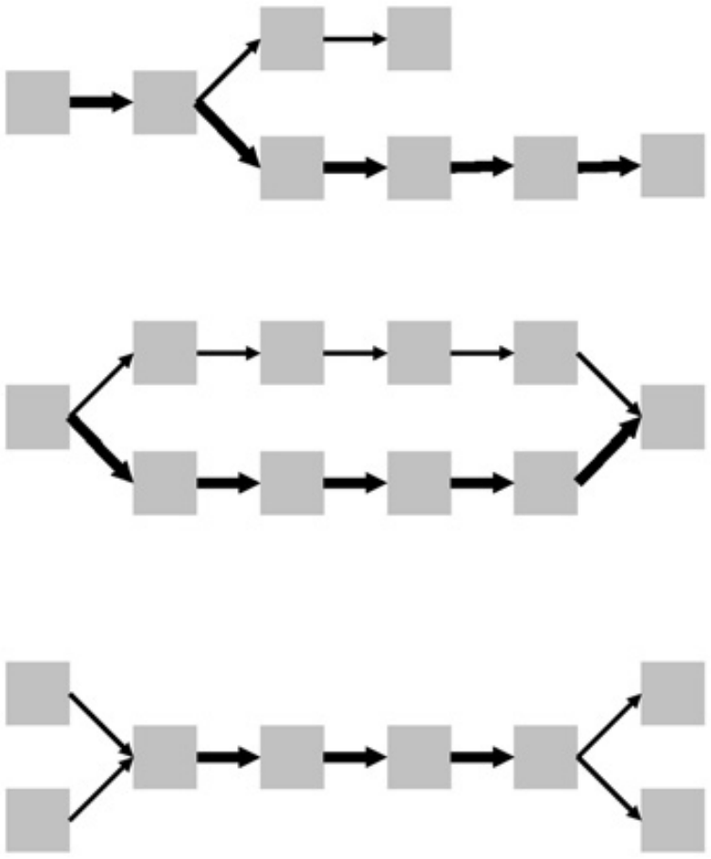
Paths through the de Bruijn graph are assembled sequences

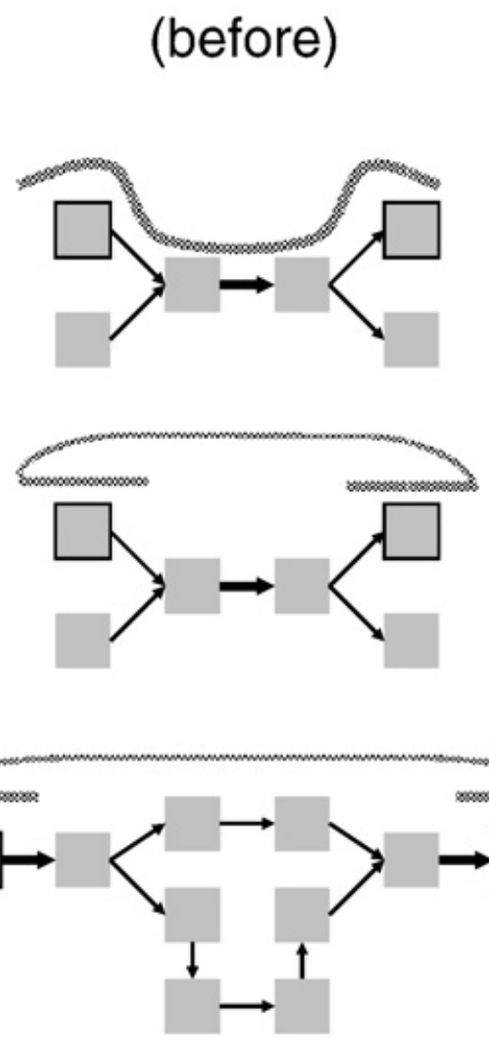
These paths can be very complicated due to sequencing error, snp's, splicing variants, repeats, etc

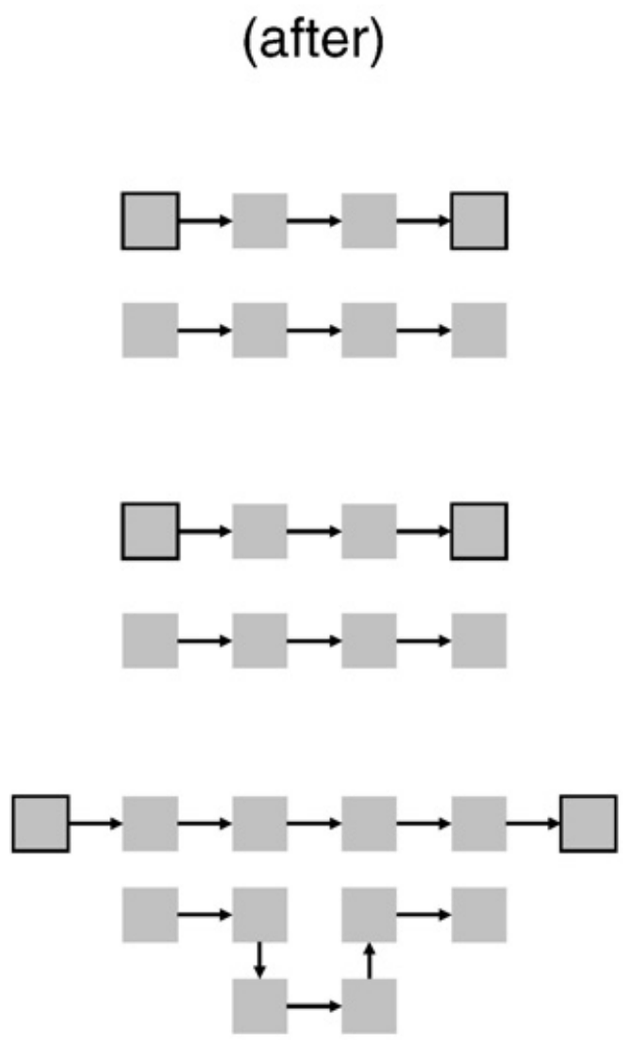
The graphs require considerable post-processing to simplify them (pop bubbles, trim dead ends, etc)



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






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

# Assembly takes a lot of RAM

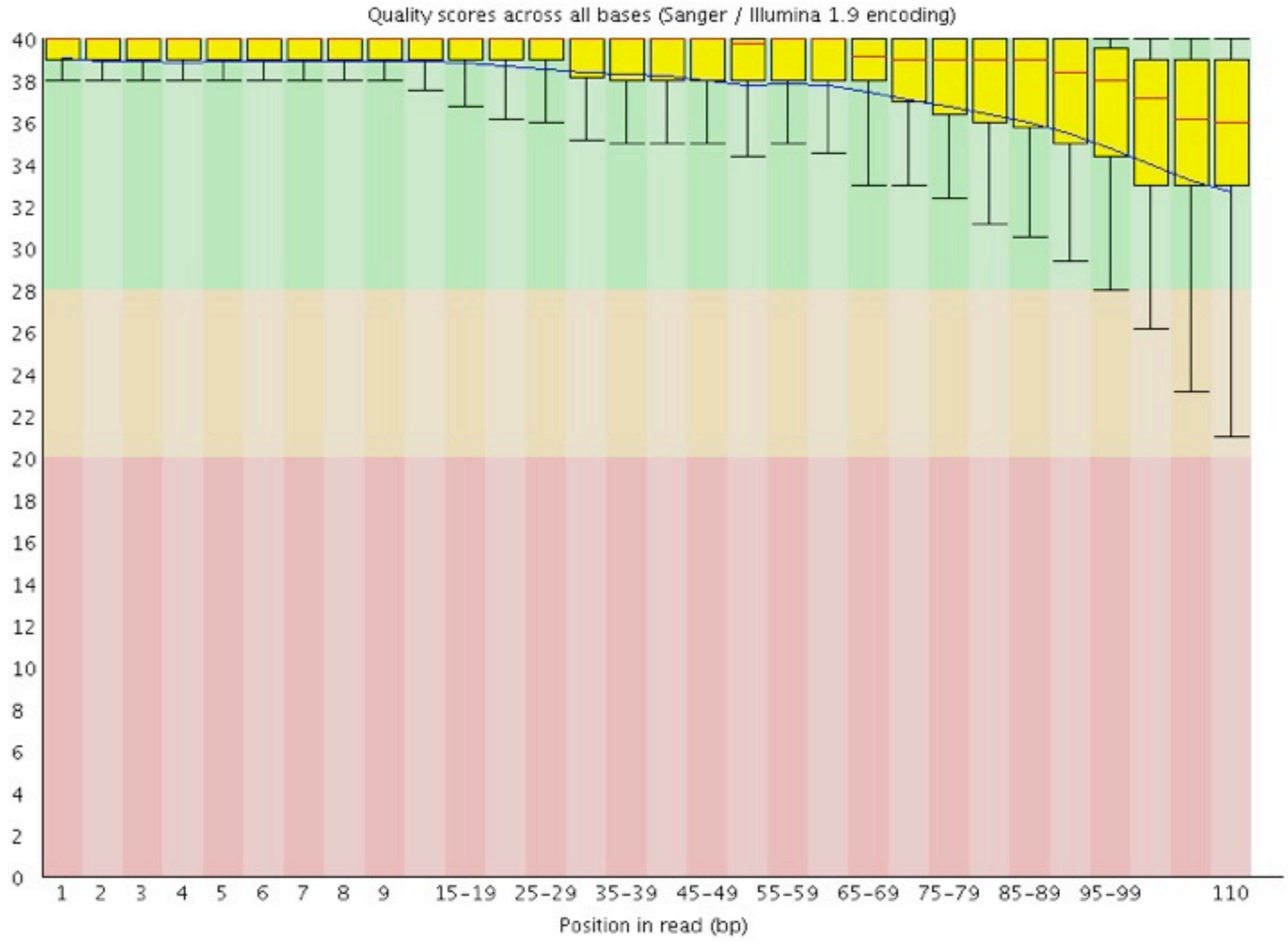
One lane of Illumina HiSeq data can require hundreds of gigabytes of RAM to assemble

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

# Fastqc for quality profiling

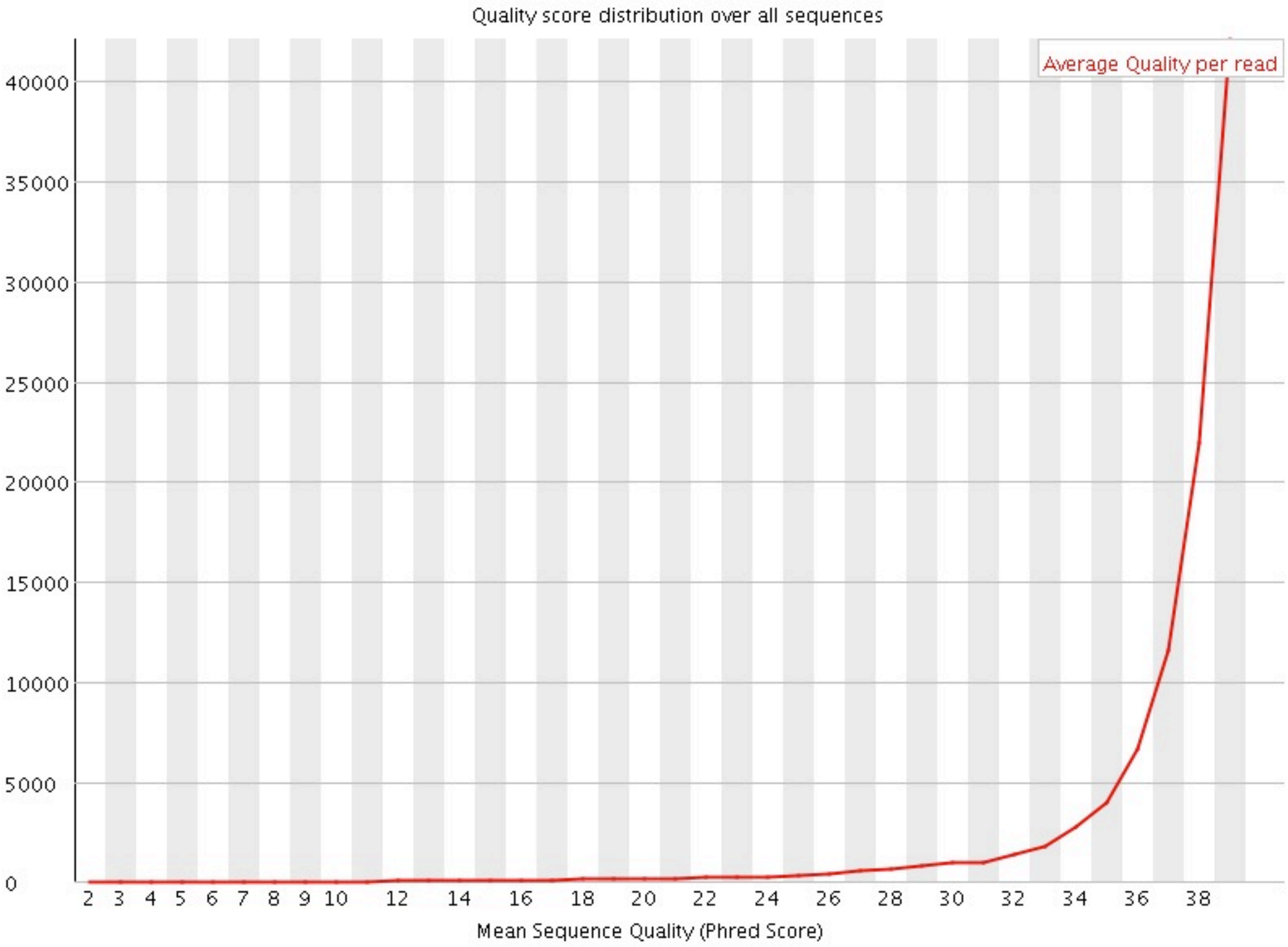
## Per base sequence quality




http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

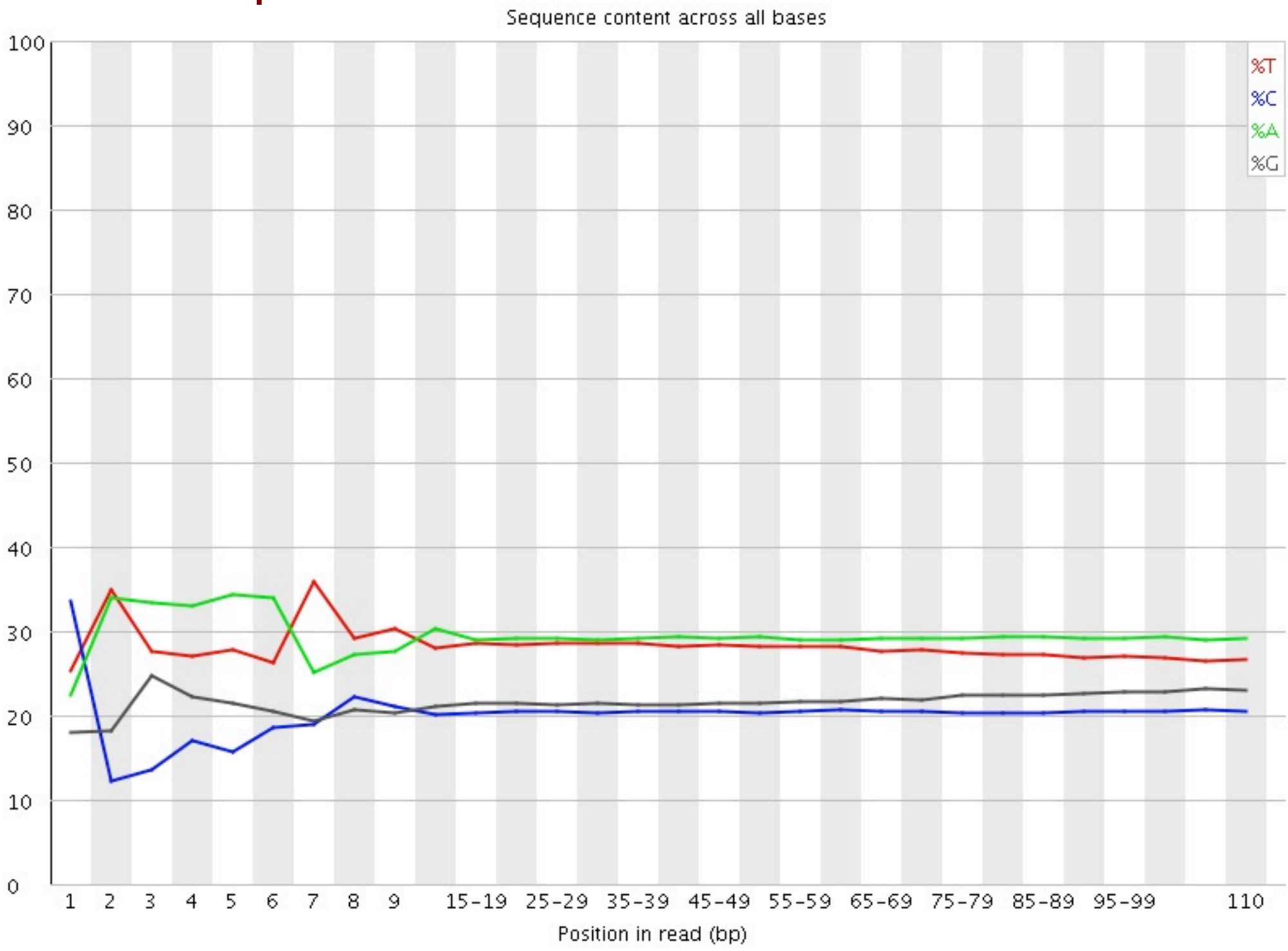
# Fastqc for quality profiling

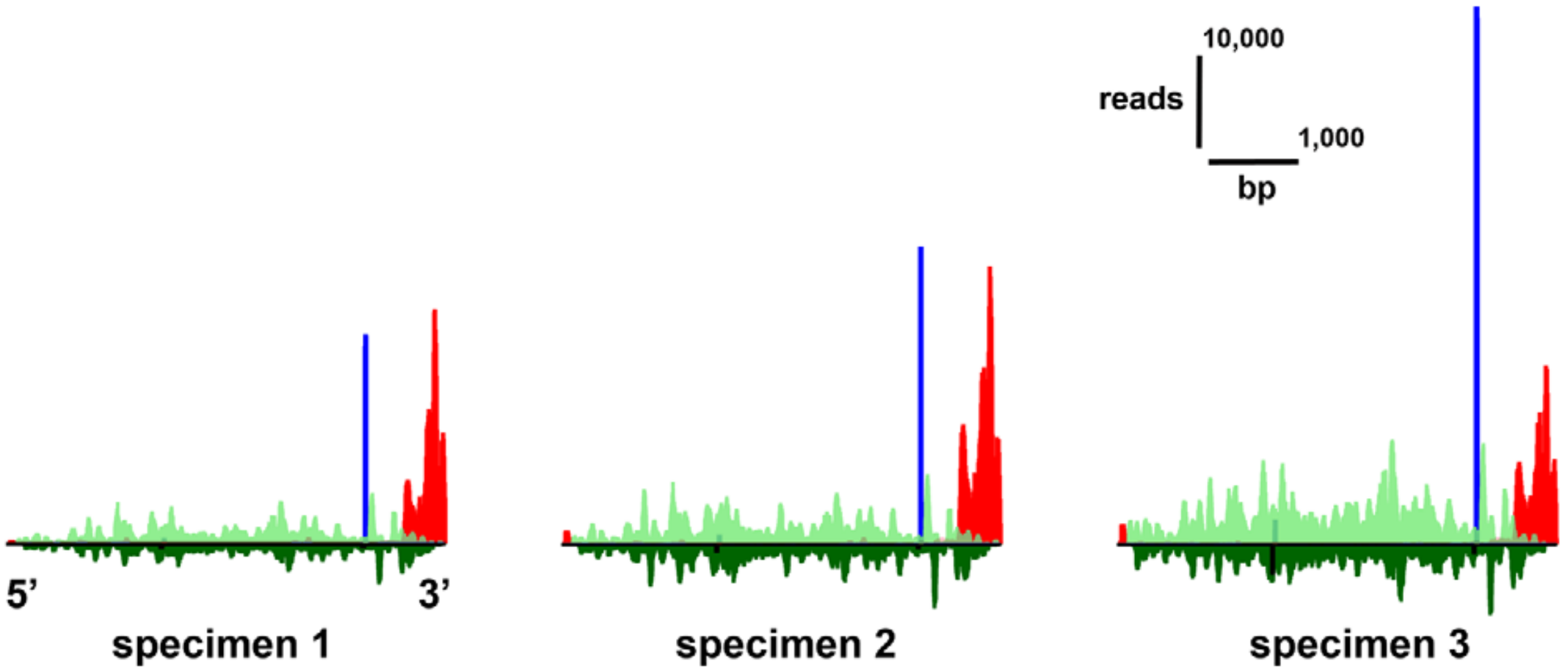
## Per sequence quality scores



# Fastqc for quality profiling

**Per base sequence content**






Illumina mRNA-seq

SOLiD DGE

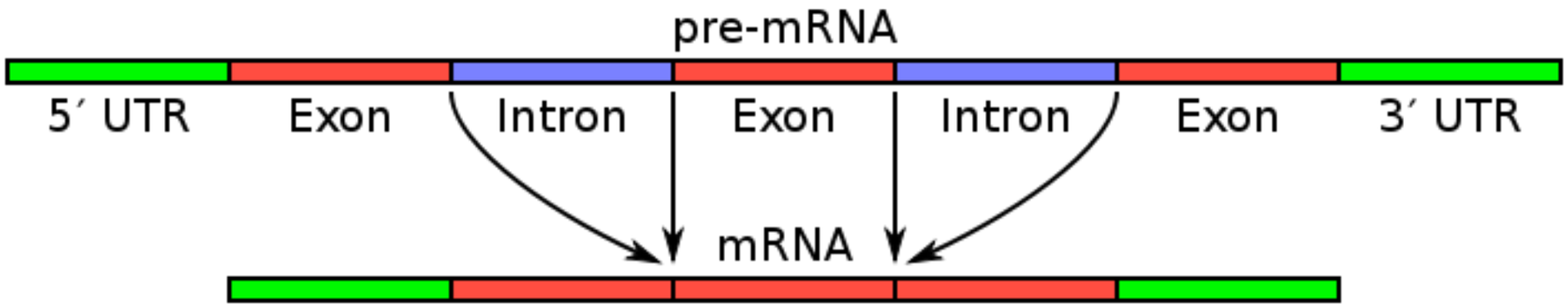
Helicos DGE

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

| 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 |

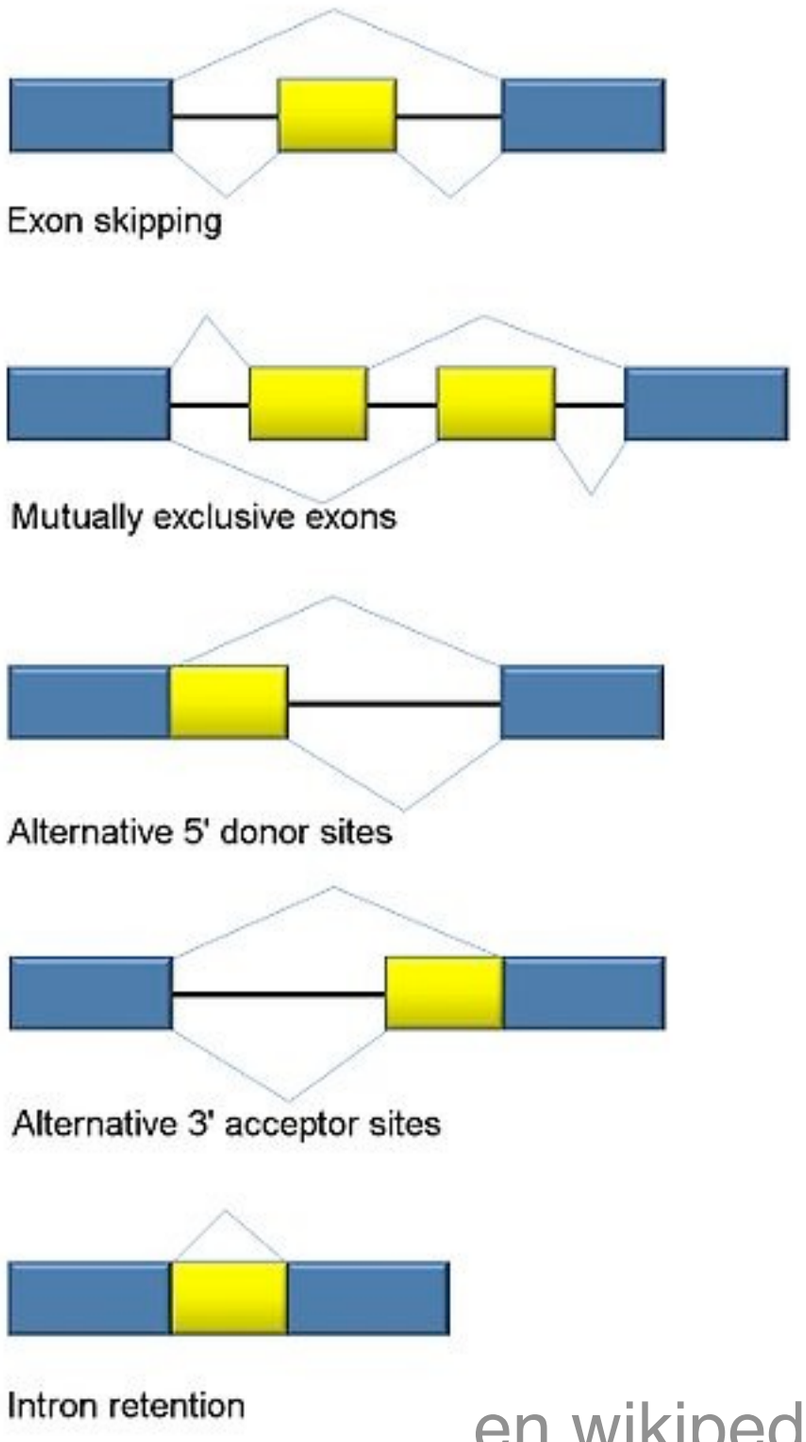
# Transcript splicing

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

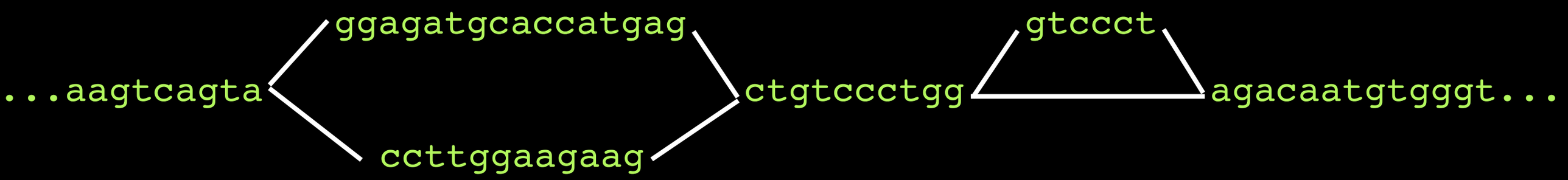
```
...aagtcagtggagatgcaccatgagaccttggaagaagctgtccctggagacaatgtgggt...
```

# Assembly results...

## Genome

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

## Transcript



# Splice variants

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

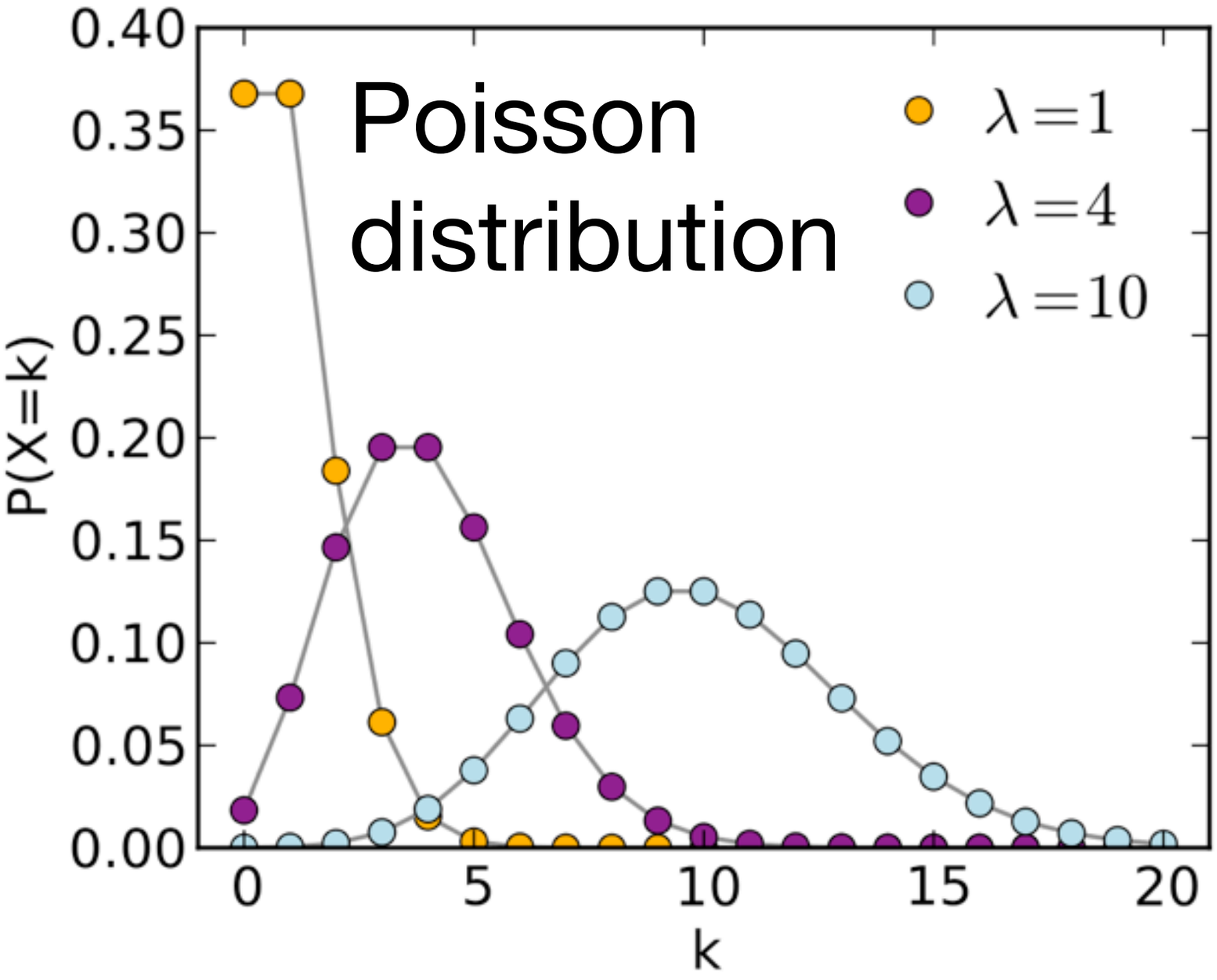
-Deep sequencing captures some “intermediate splice variants”, molecules in the process of being spliced

-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:

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

Trinity (trinityrnaseq.sourceforge.net)

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

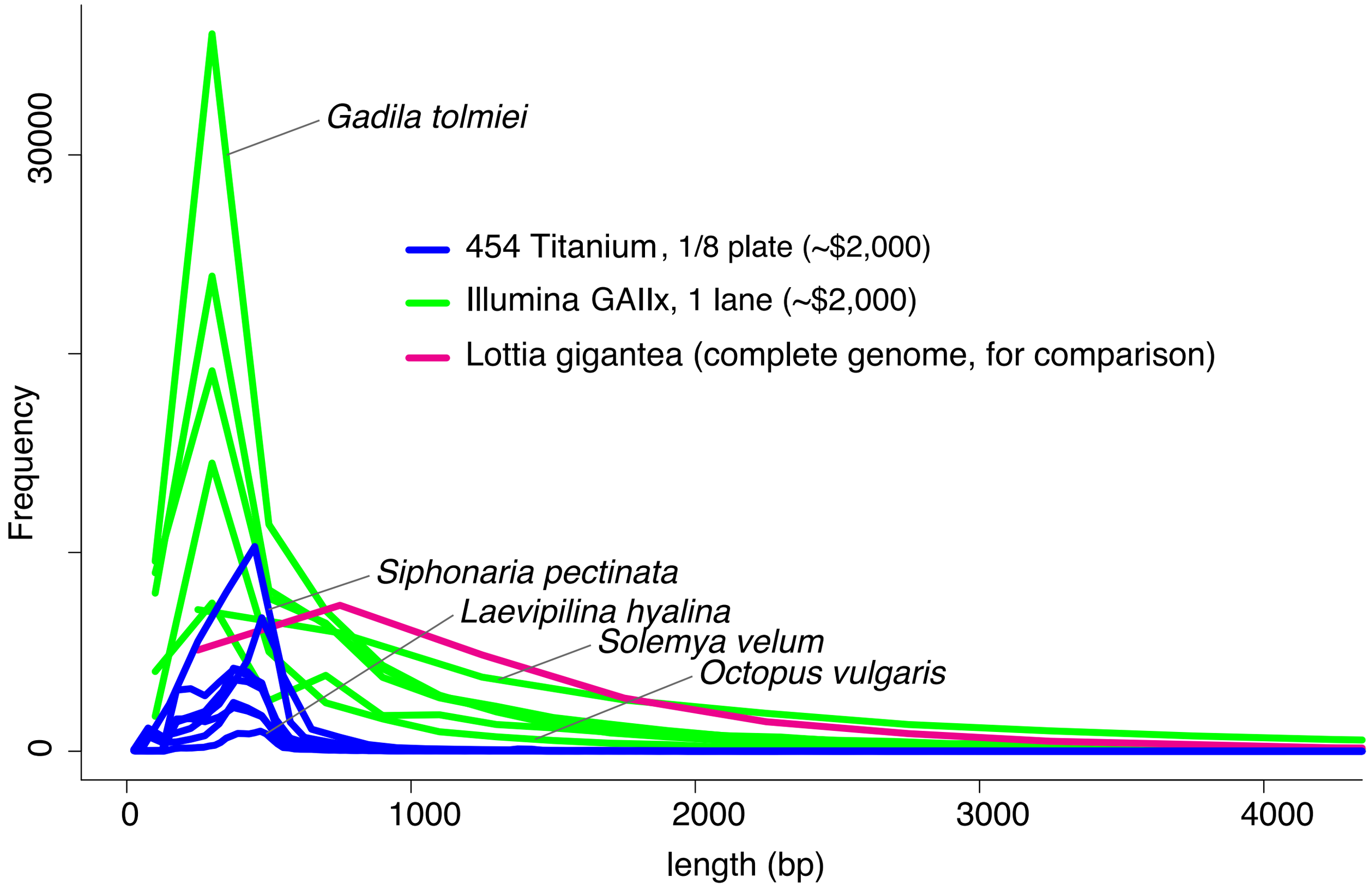
Newbler (Roche)

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

# Mollusca

(NSF funded project with Giribet and Wilson)

# Comparison of 454 and Illumina cDNA Assemblies



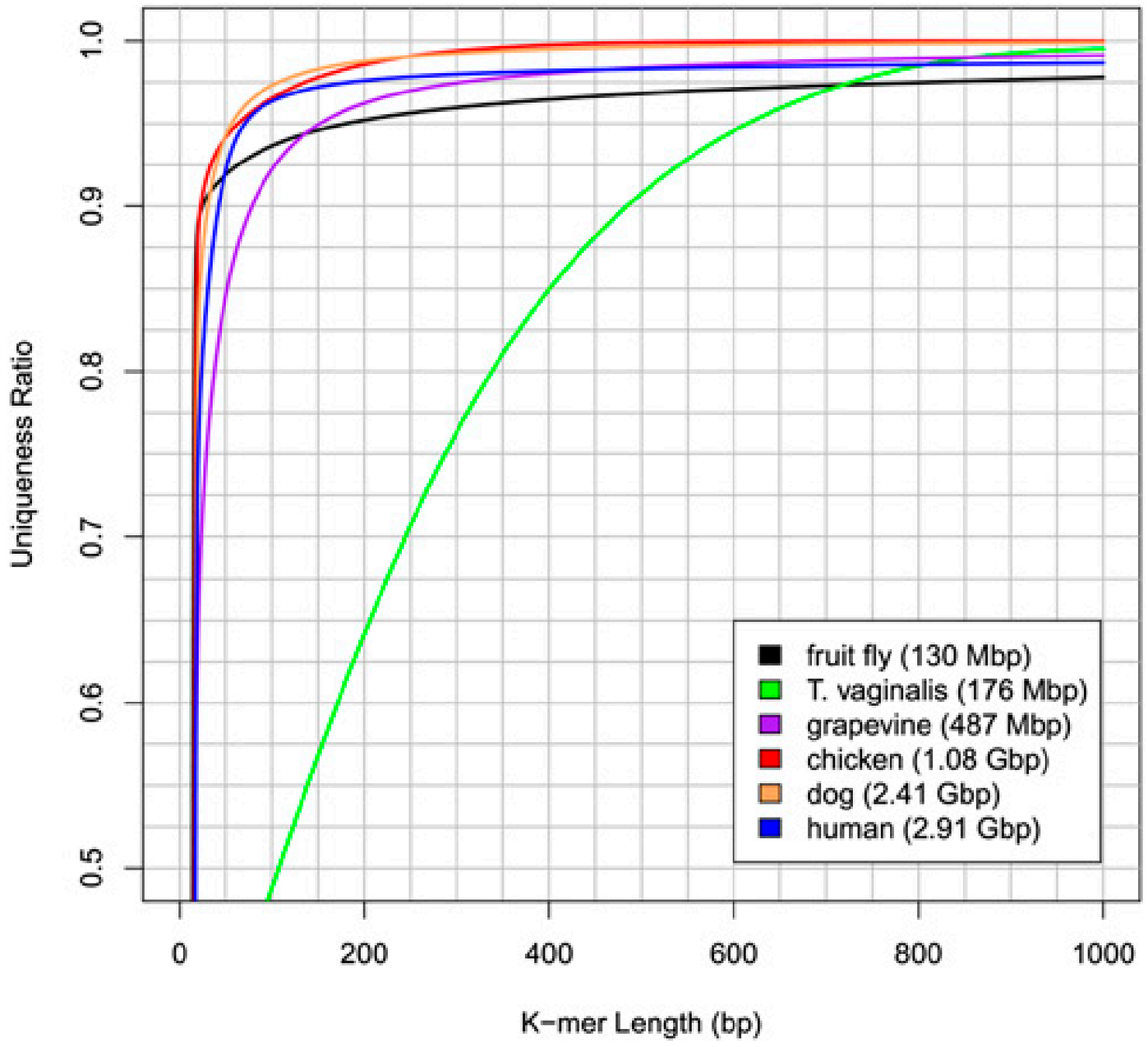



**Figure 1.** The *k*-mer uniqueness ratio for five well-known organisms and one single-celled human parasite. The ratio is defined here as the percentage of the genome that is covered by unique sequences of length *k* or longer. The horizontal axis shows the length in base pairs of the sequences. For example, ~92.5% of the grapevine genome is contained in unique sequences of 100 bp or longer.

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

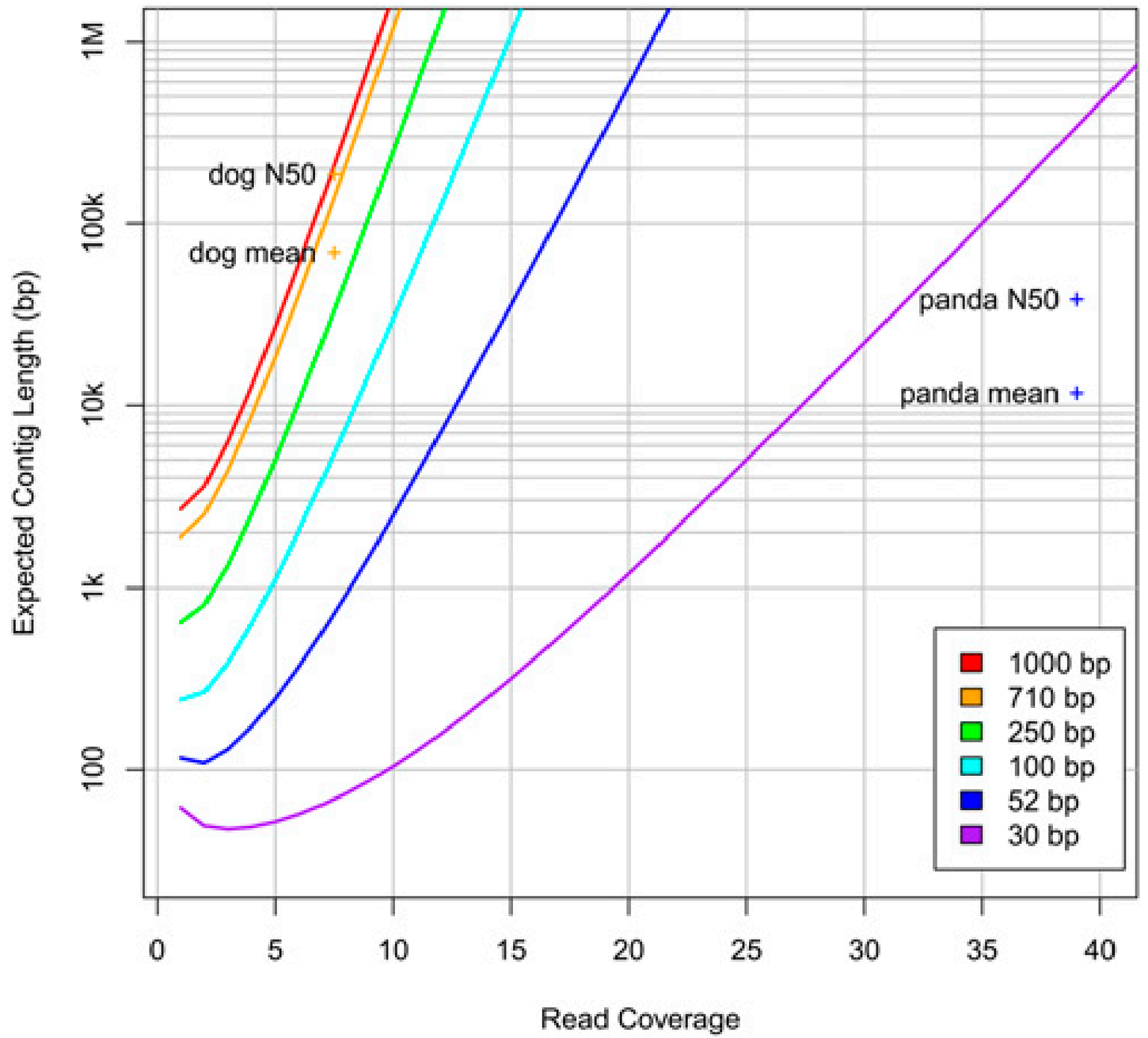



**Figure 3.** Expected average contig length for a range of different read lengths and coverage values. Also shown are the average contig lengths and N50 lengths for the dog genome, assembled with 710-bp reads, and the panda genome, assembled with reads averaging 52 bp in length.

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

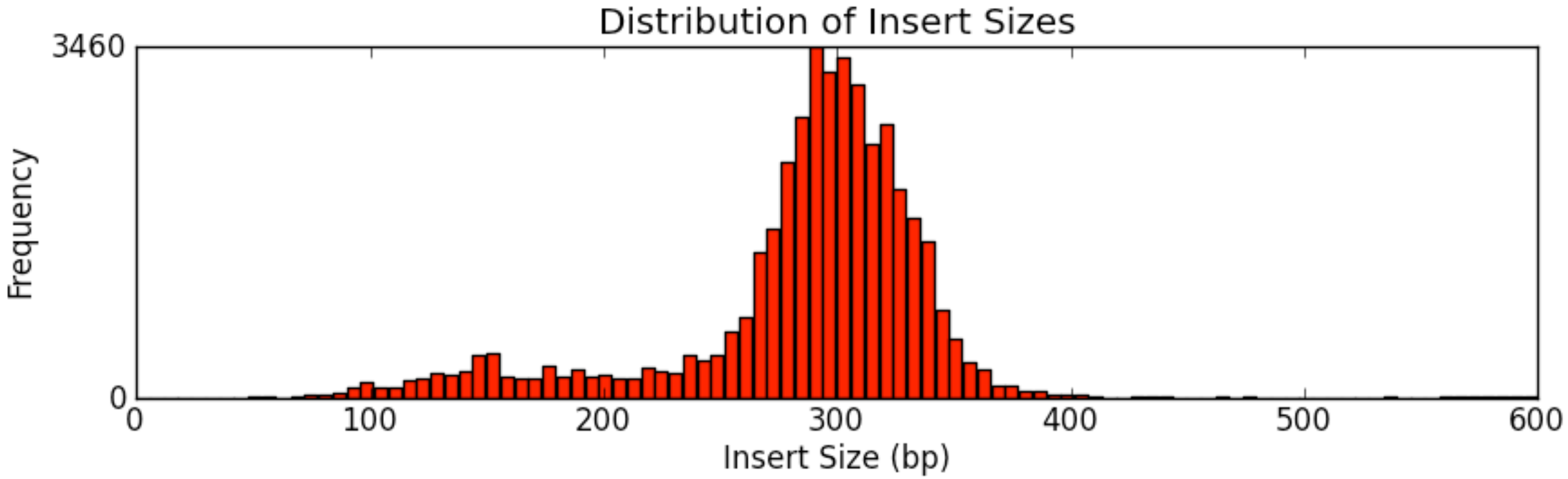
# Post-assembly annotation:

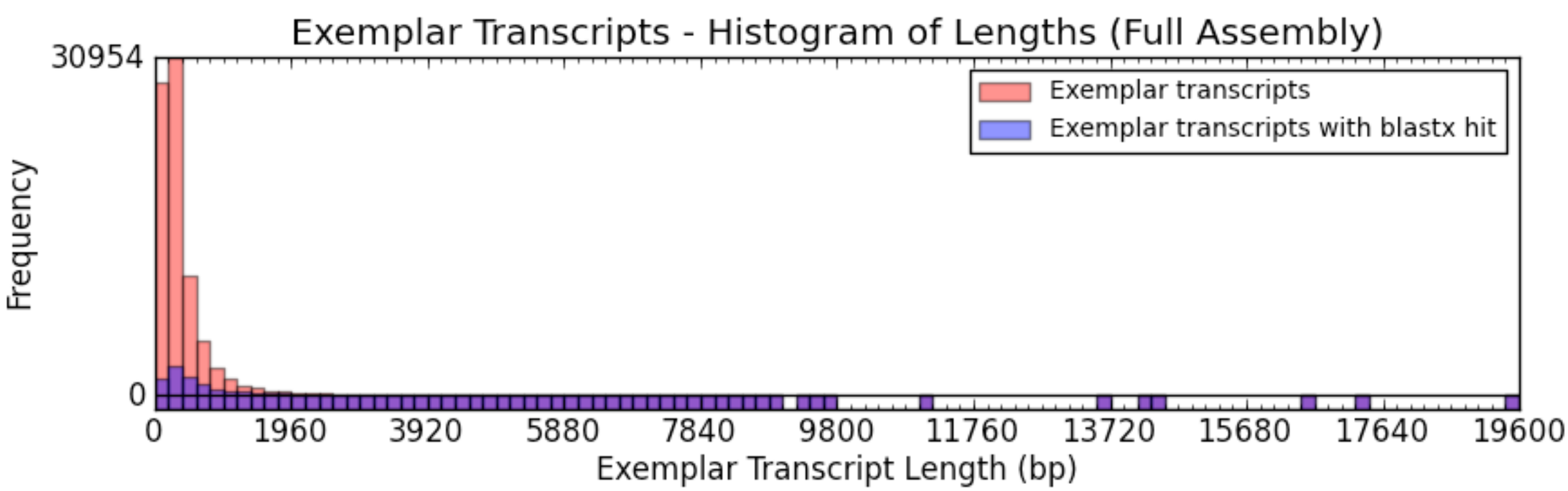
- Selection of exemplar transcripts for each gene
- Blastx to a taxon restricted subset of the NCBI nr database
- Translation

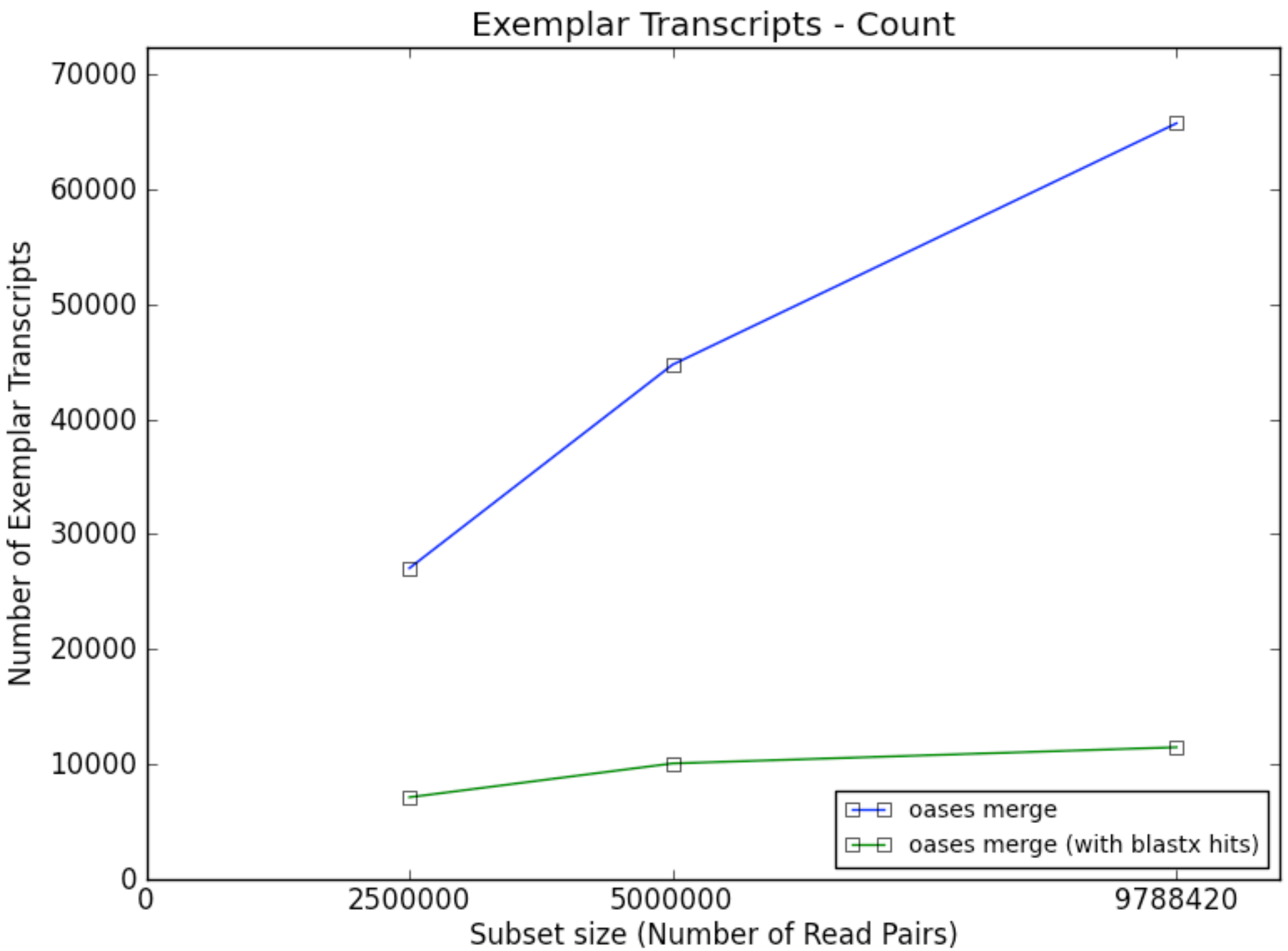
# Our tool for processing assemblies: *Agalma*

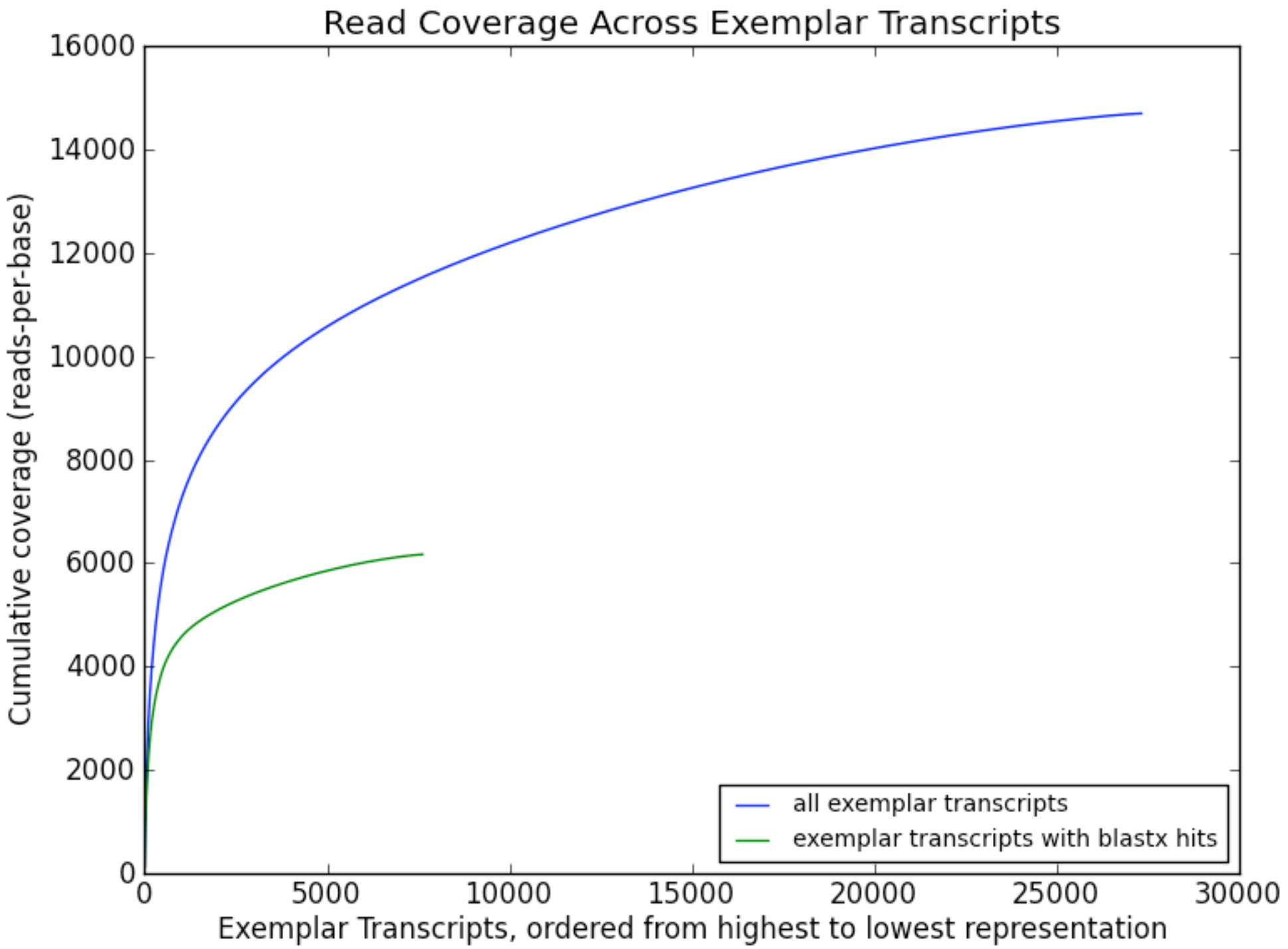
https://bitbucket.org/caseywdunn/agalma

Mark Howison, Felipe Zapata, and Casey Dunn

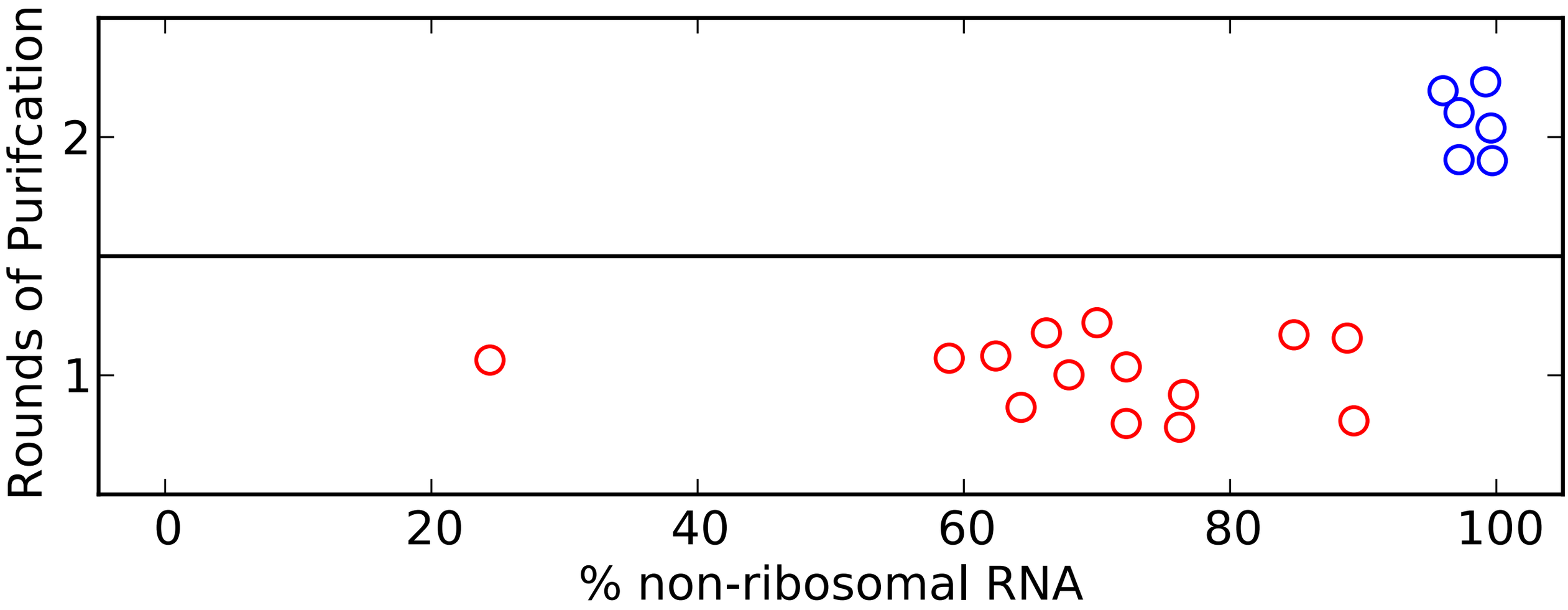

Distribution of Insert Sizes
3460
0
Frequency
0
100
200
300
400
500
600
Insert Size (bp)


Exemplar Transcripts - Histogram of Lengths (Full Assembly)
30954
0
Frequency
Exemplar transcripts
Exemplar transcripts with blastx hit
0
1960
3920
5880
7840
9800
11760
13720
15680
17640
19600
Exemplar Transcript Length (bp)

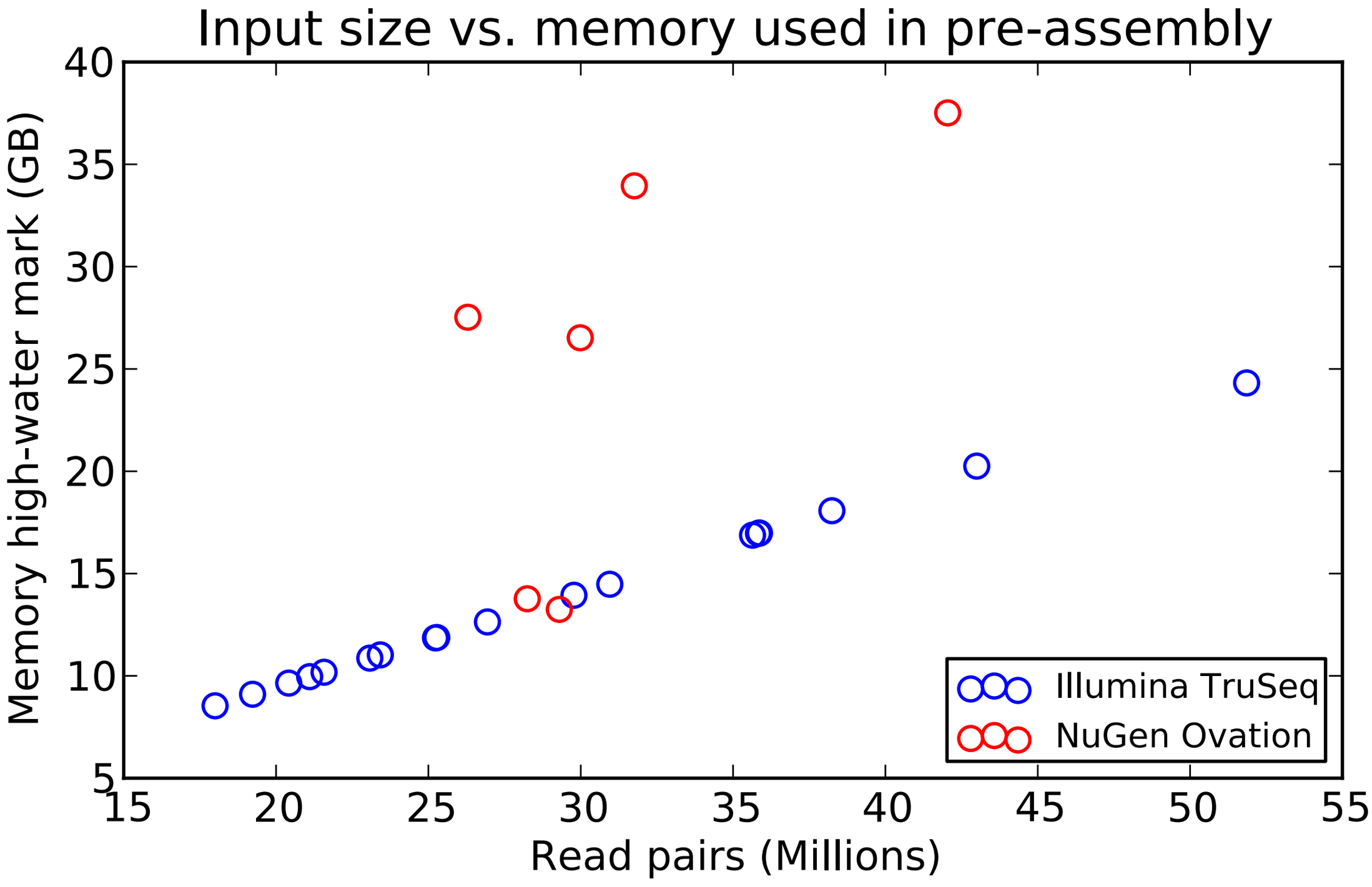

Exemplar Transcripts - Count
Number of Exemplar Transcripts
70000
60000
50000
40000
30000
20000
10000
0
0
2500000
5000000
9788420
Subset size (Number of Read Pairs)
oases merge
oases merge (with blastx hits)


Read Coverage Across Exemplar Transcripts
Cumulative coverage (reads-per-base)
16000
14000
12000
10000
8000
6000
4000
2000
0
0
5000
10000
15000
20000
25000
30000
Exemplar Transcripts, ordered from highest to lowest representation
all exemplar transcripts
exemplar transcripts with blastx hits

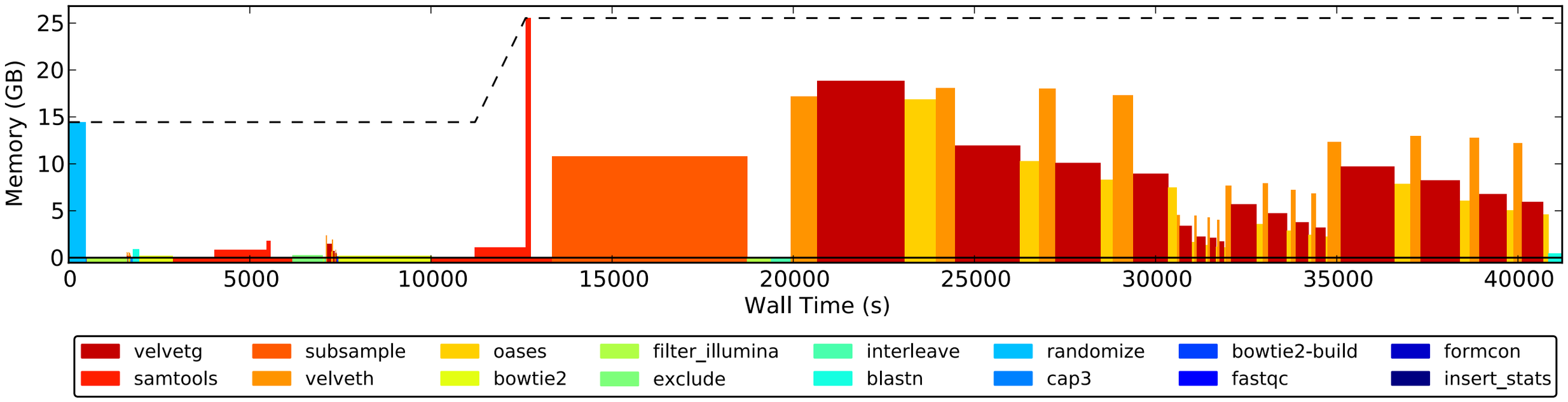
# Tabular reports summarize results across samples



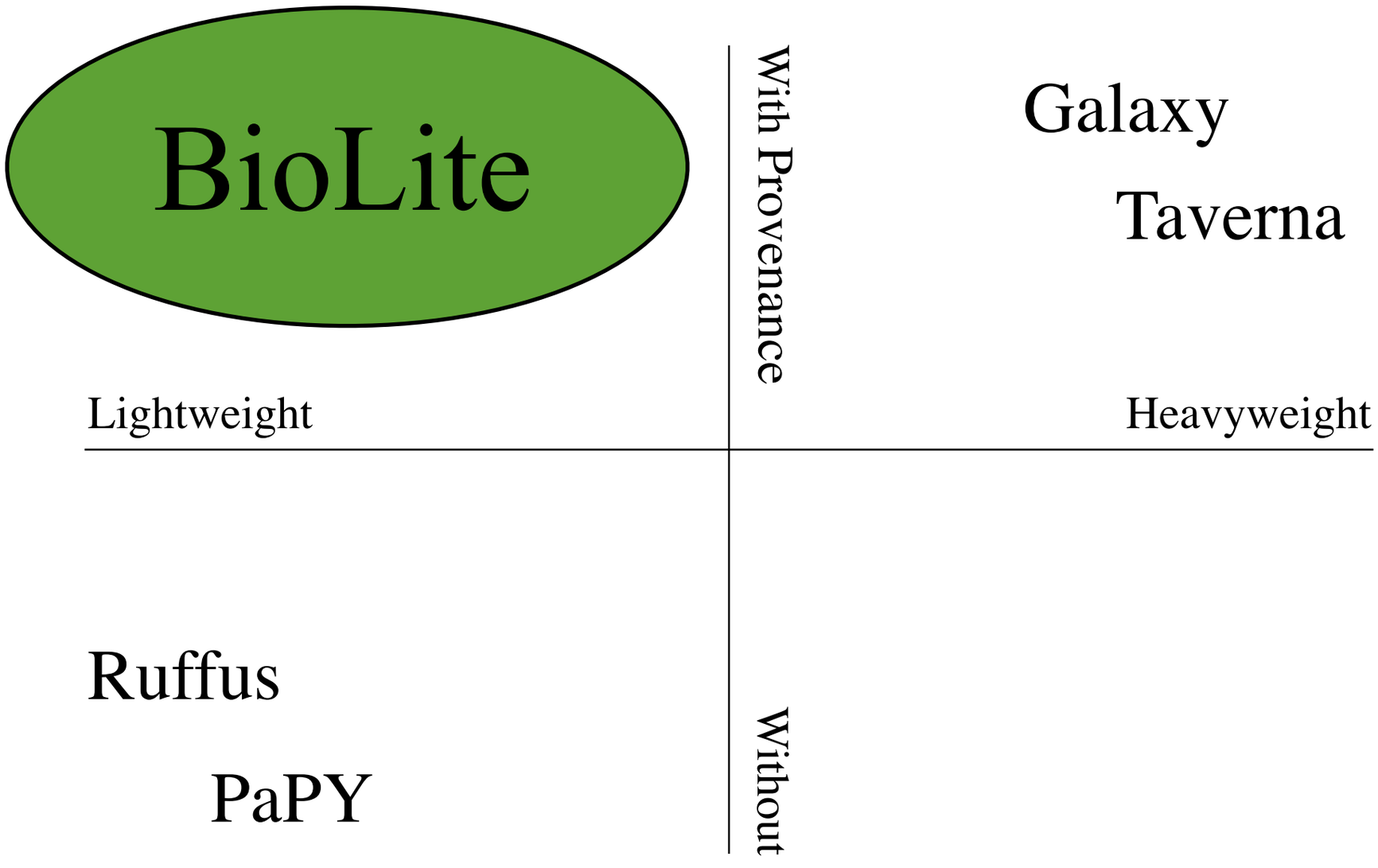
# Tabular reports summarize results across samples



# Fine-scale performance profiling



# Agalma is built with BioLite, our general-purpose data analysis platform




(M Howison)

BioLite is a Python framework and set of C++ tools for:

- building out customized analysis **pipelines**
- fault-tolerance, through built-in **checkpointing**
- automating the collection/reporting of **diagnostics**
- tracking the **provenance** of analyses:
  - resource usage
  - paths and parameters
  - program versioning
  - statistics

(M Howison)

# Identifying and selecting homologs

Species A

Species B

Species C

Species

Gene

Species A

Species B

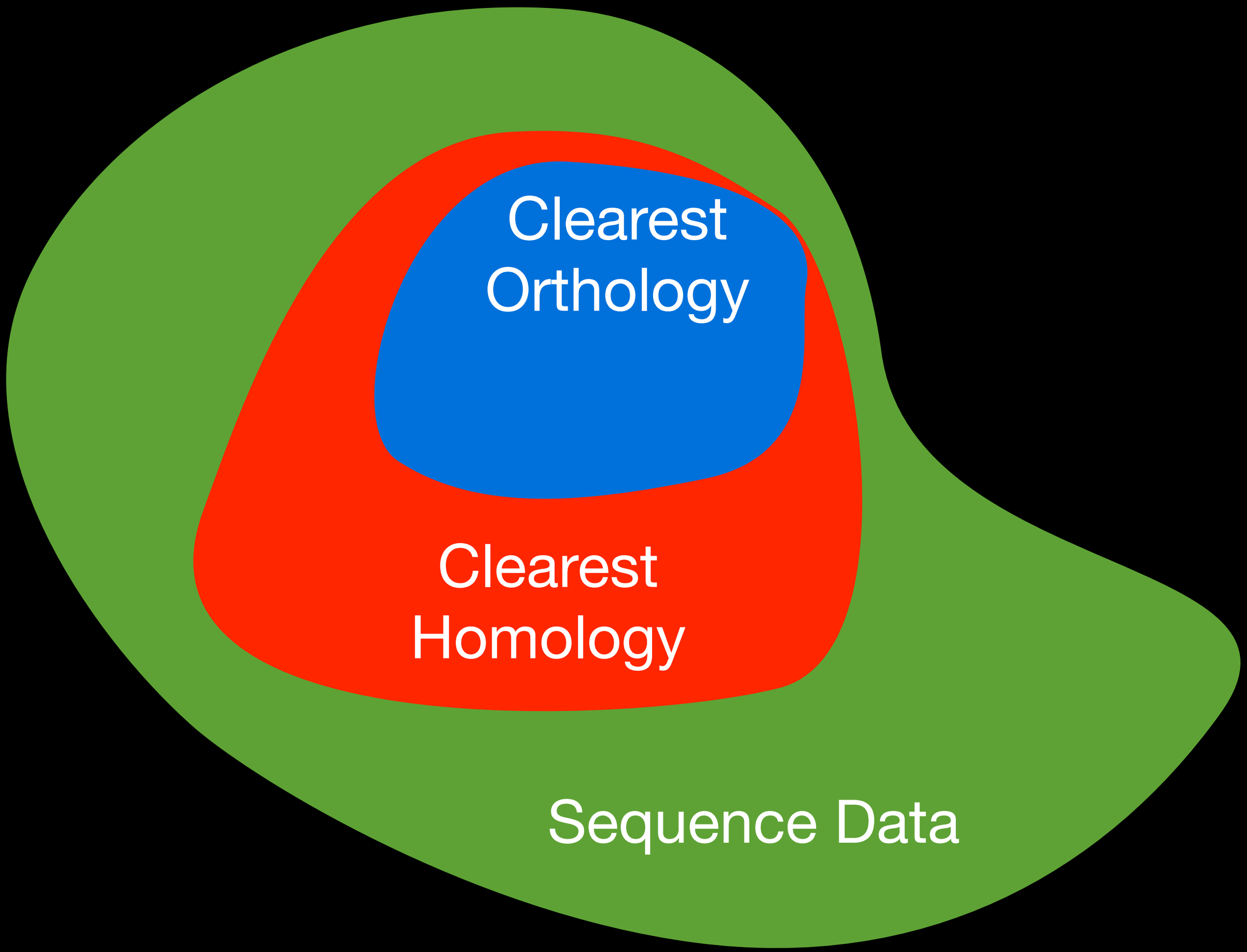
Species C

Gene divergence due to duplication

Sequence Data

Clearest
Homology

Sequence Data


Clearest
Orthology
Clearest
Homology
Sequence Data

Clearest Orthology

Most Informative

Clearest Homology

Sequence Data

Phylogenetic tools build trees from homologous characters

Phylogenetic tools build trees from homologous characters

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

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

Raw sequence reads

*Assembly*

A large set of sequences

Subsets of homologous sequences

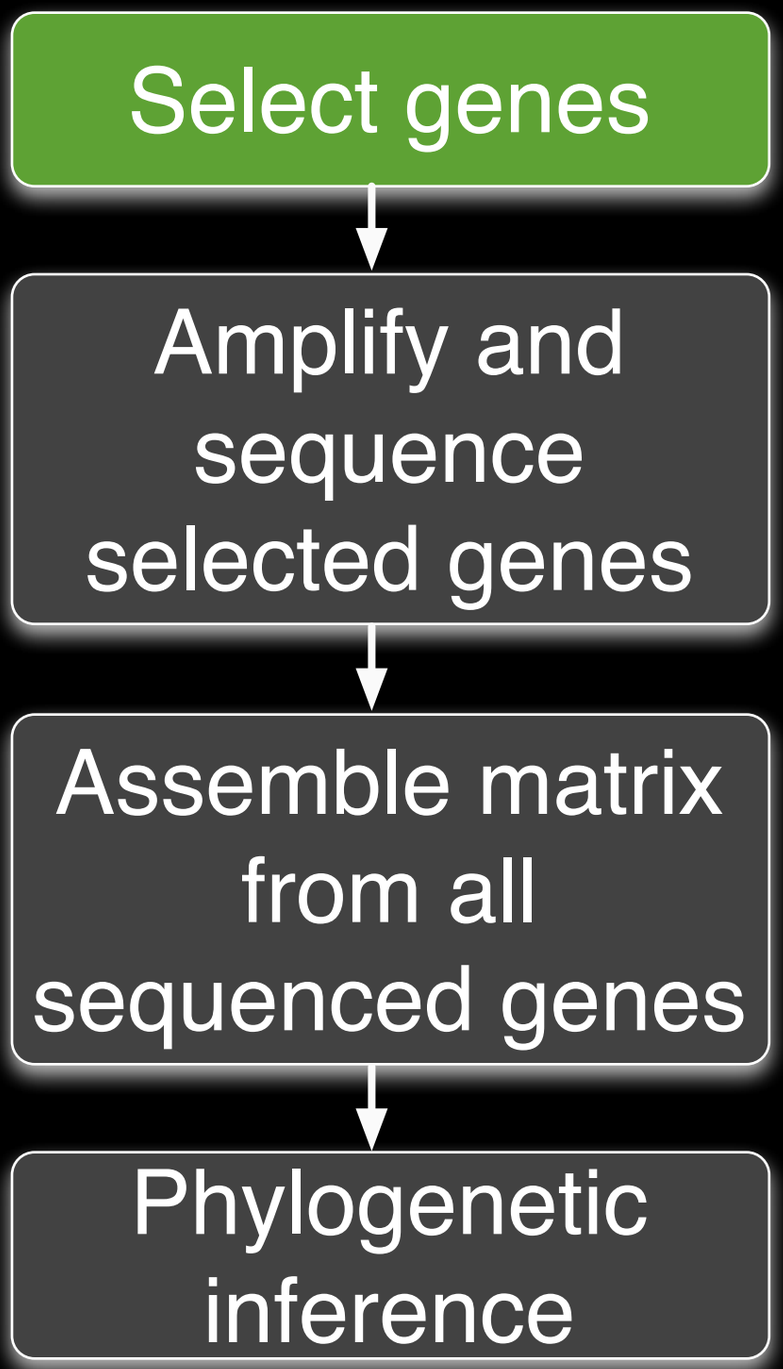
*Alignment*

Columns with homologous characters

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

*Wired,June 23, 2008*

# Gene selection as part of project design (Directed PCR)



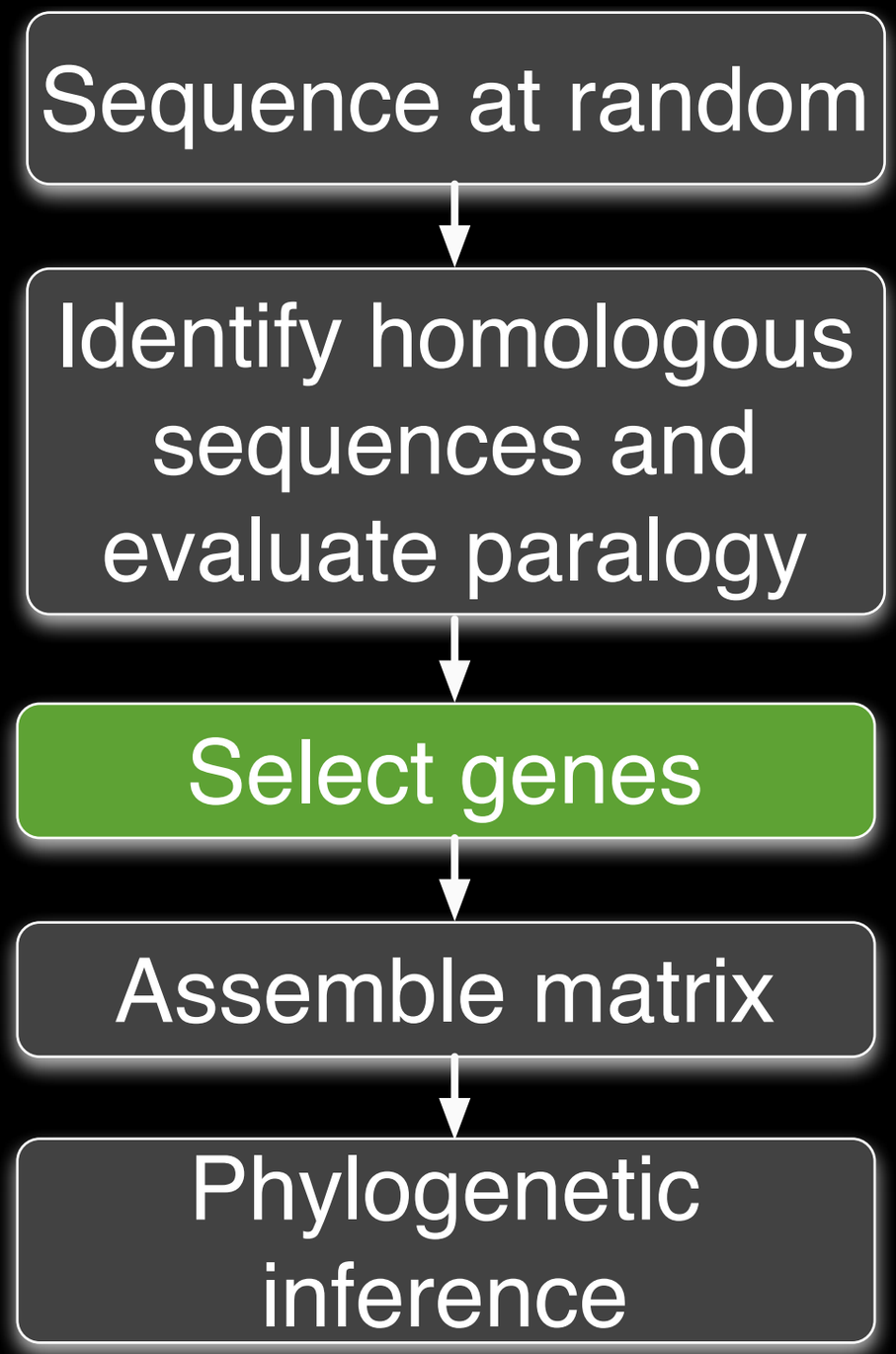
## Gene selection as part of project design (Directed PCR)

- 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

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



# Approaches and tools

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

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

# Approaches and tools

Use phenetic tools to identify orthologs *de novo*

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

OrthoMCL
dx.doi.org/10.1101/gr.1224503

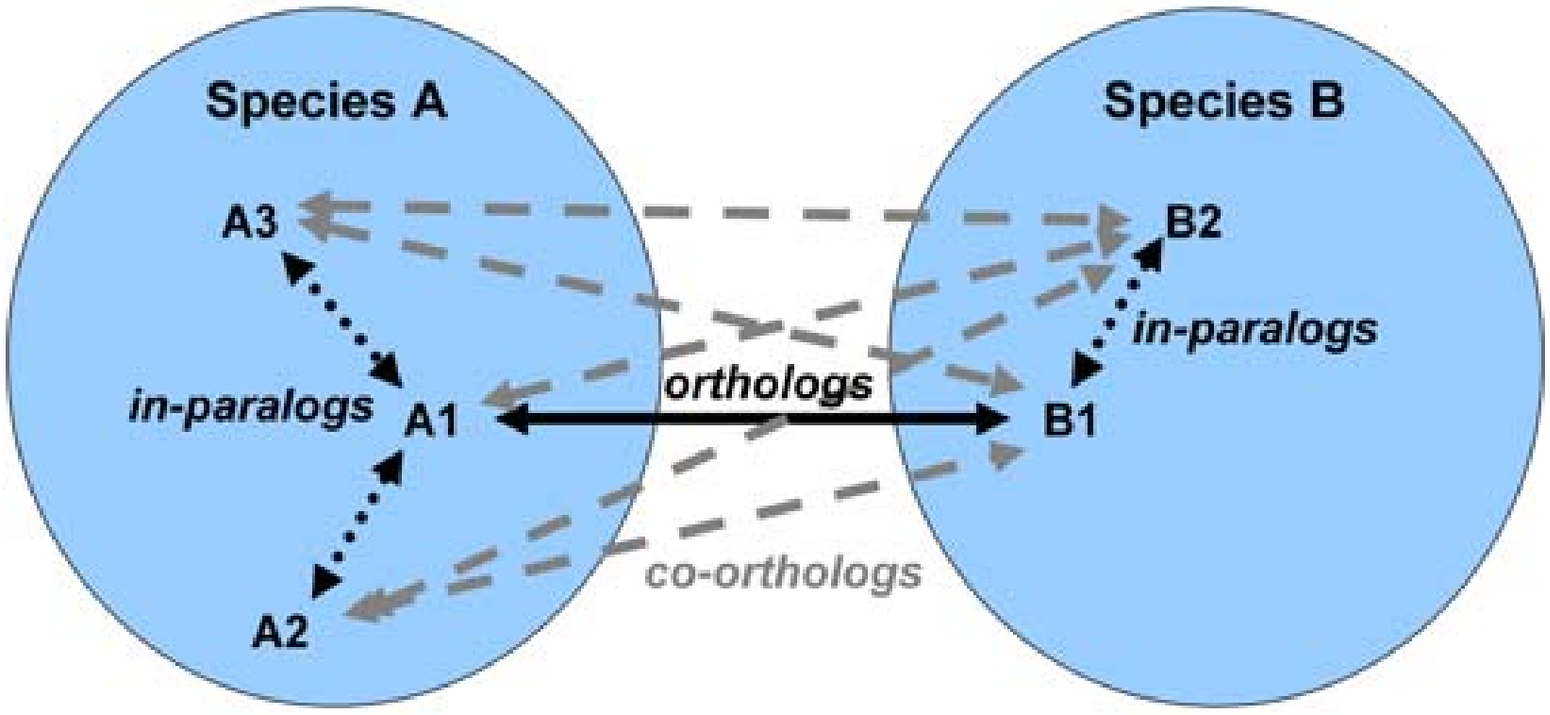



**Figure 1. OrthoMCL graph construction between two species, including the establishment of co-ortholog relationships.** Solid lines connecting A1 and B1 represent putative ortholog relationships identified by the 'reciprocal best hit' (RBH) rule. Dotted lines (e.g. those connecting A1 with A2 and A3, 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

# Approaches and tools

Use phenetic tools to identify homologs *de novo*

Blast followed by MCL (van Dongen, 2000)
http://micans.org/mcl/

Then use phylogenetic tools to identify paralogs...

| Isolation of Homologs | Isolation of Orthologs | Evaluation of Orthology |
|---|---|---|
| Phenetic | | |
| Phenetic | | Phylogenetic |
| Phenetic | | Phylogenetic |

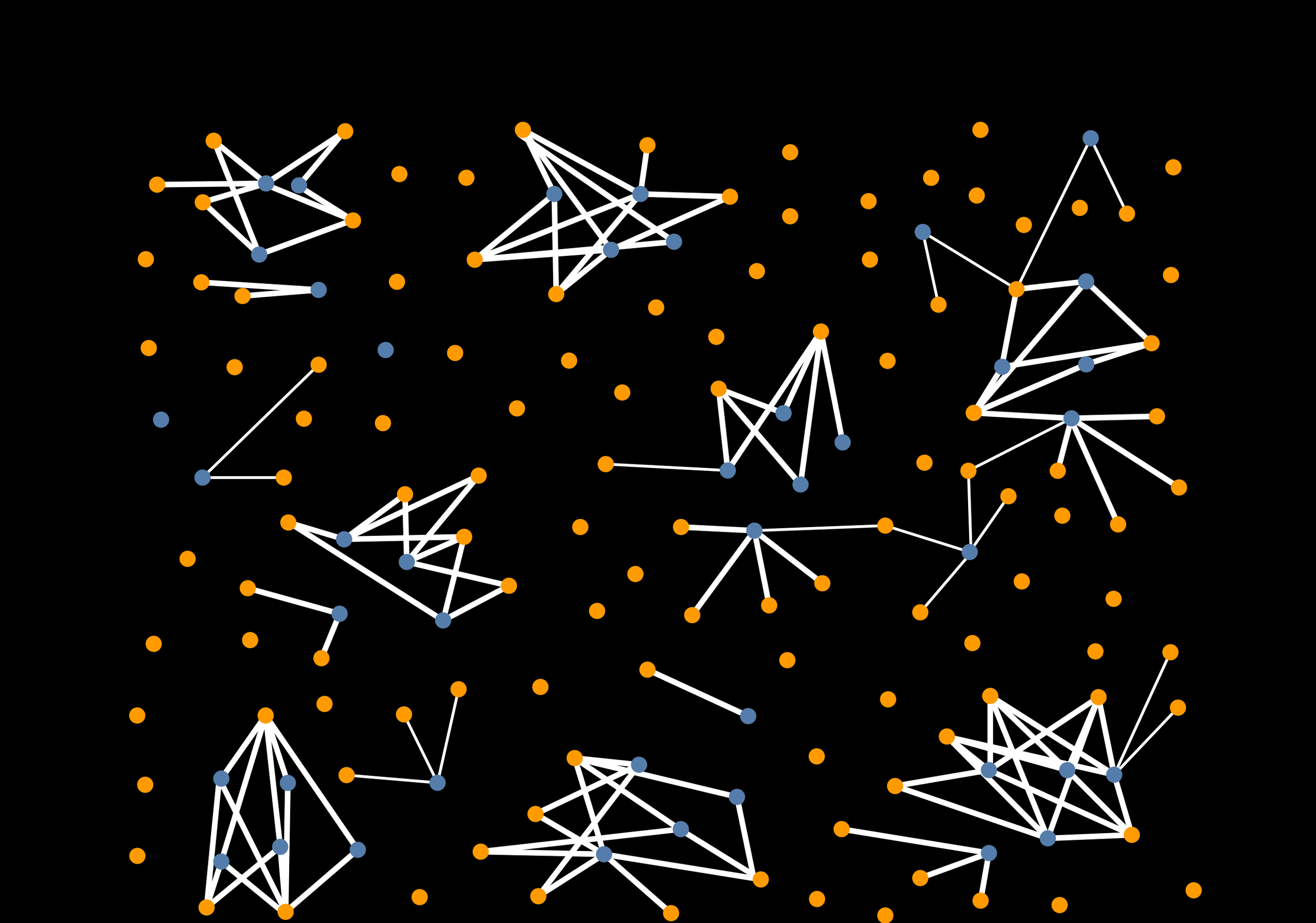
# Our general approach...

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

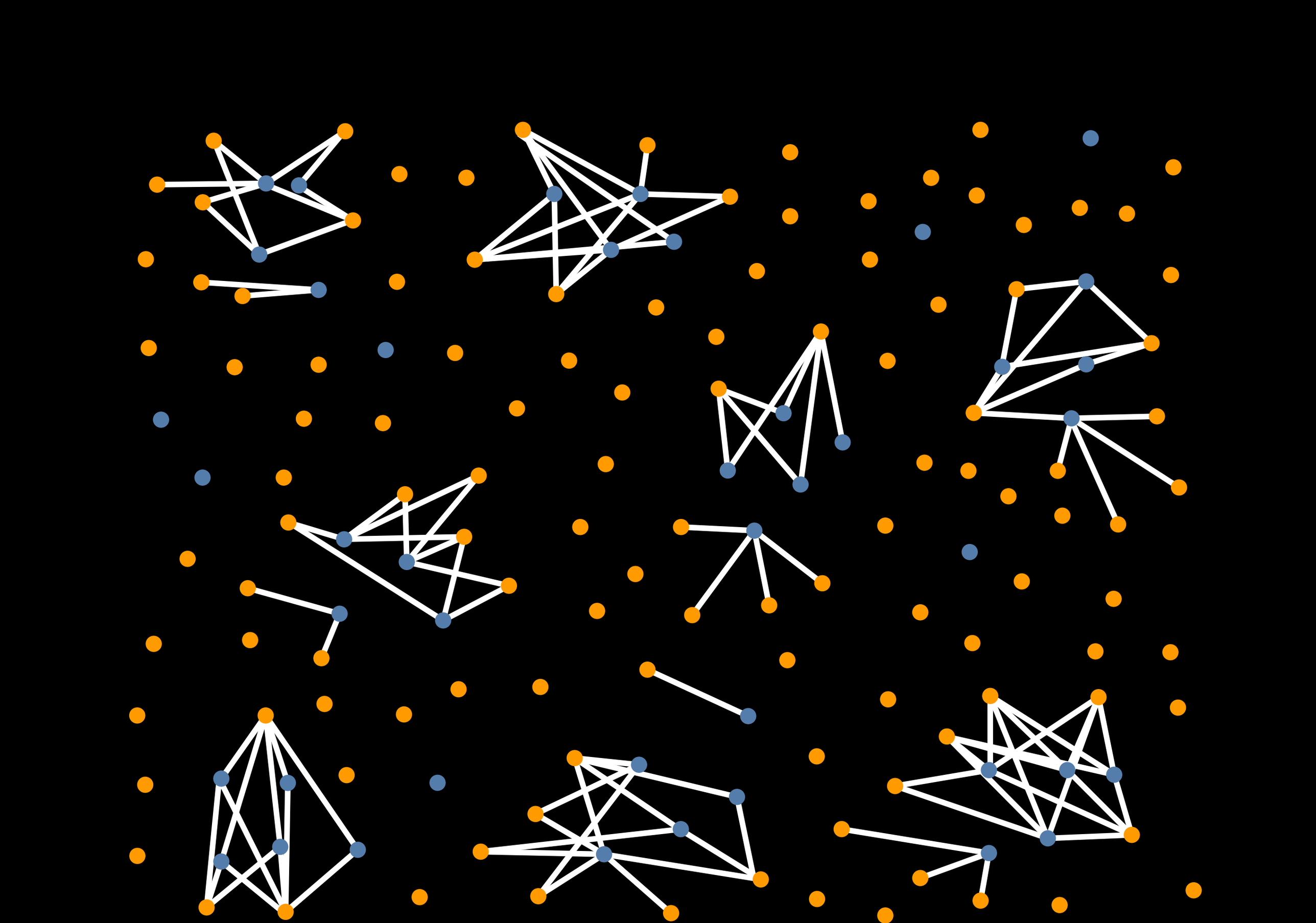
Make all pairwise sequence comparisons

Construct a graph where nodes are sequences and edges indicate similarity

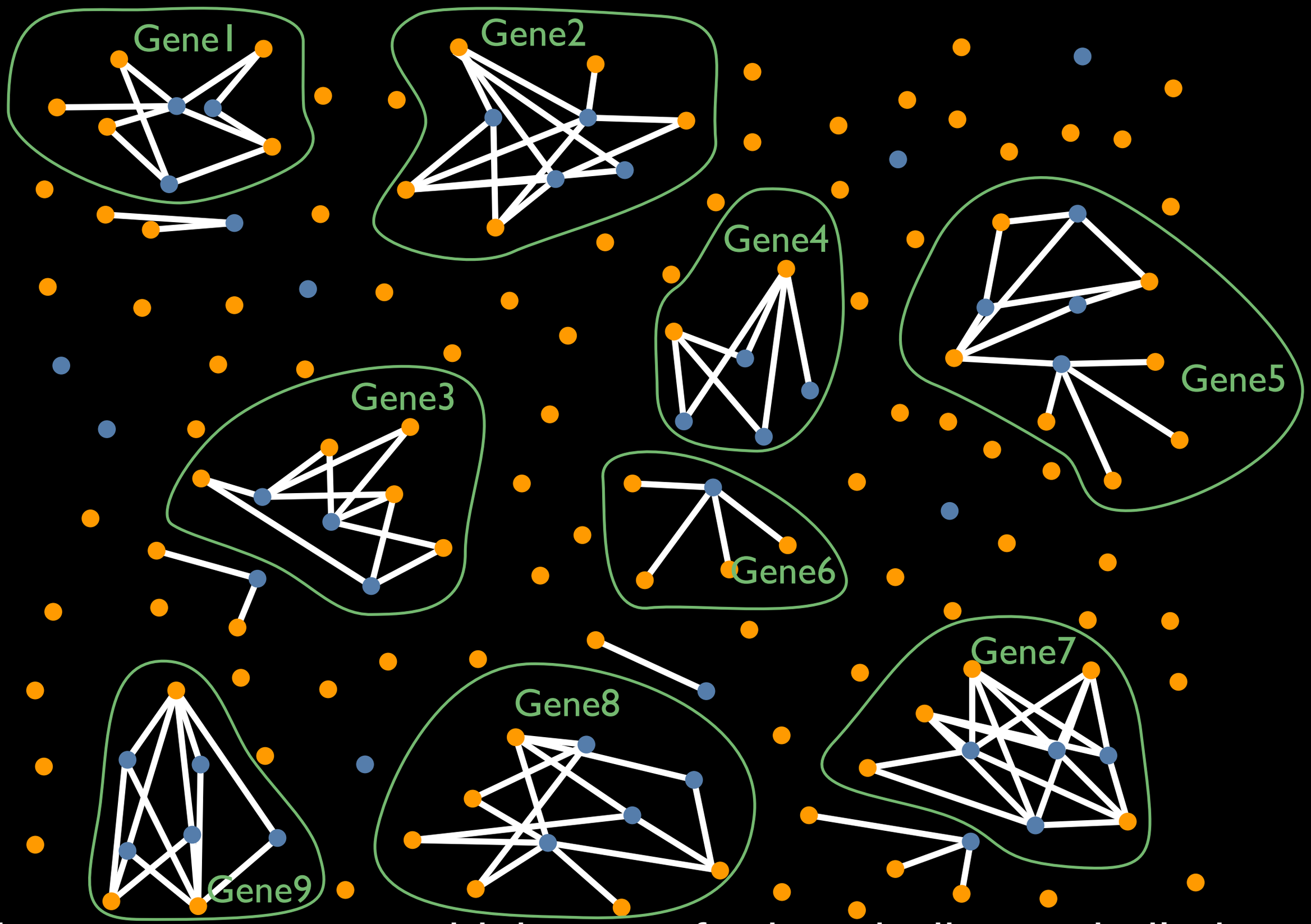
Nodes are sequences, thickness of edges indicate similarity



Nodes are sequences, thickness of edges indicate similarity



Nodes are sequences, thickness of edges indicate similarity




Nodes are sequences, thickness of edges indicate similarity

In the past, we have used:

- Blastp, assign -$\log_{10}$(e-value) for edge weight
- Throw away edges < 20
- mcl for clustering (inflation ~2.1)
- Apply taxon sampling criteria

# “The paralogy problem”

# “The paralogy problem”

But paralogs aren’t inherently a problem

# “The paralogy problem”

But paralogs aren’t inherently a problem

The problem is misascribing paralogs as orthologs

Species A

Species B

Species C

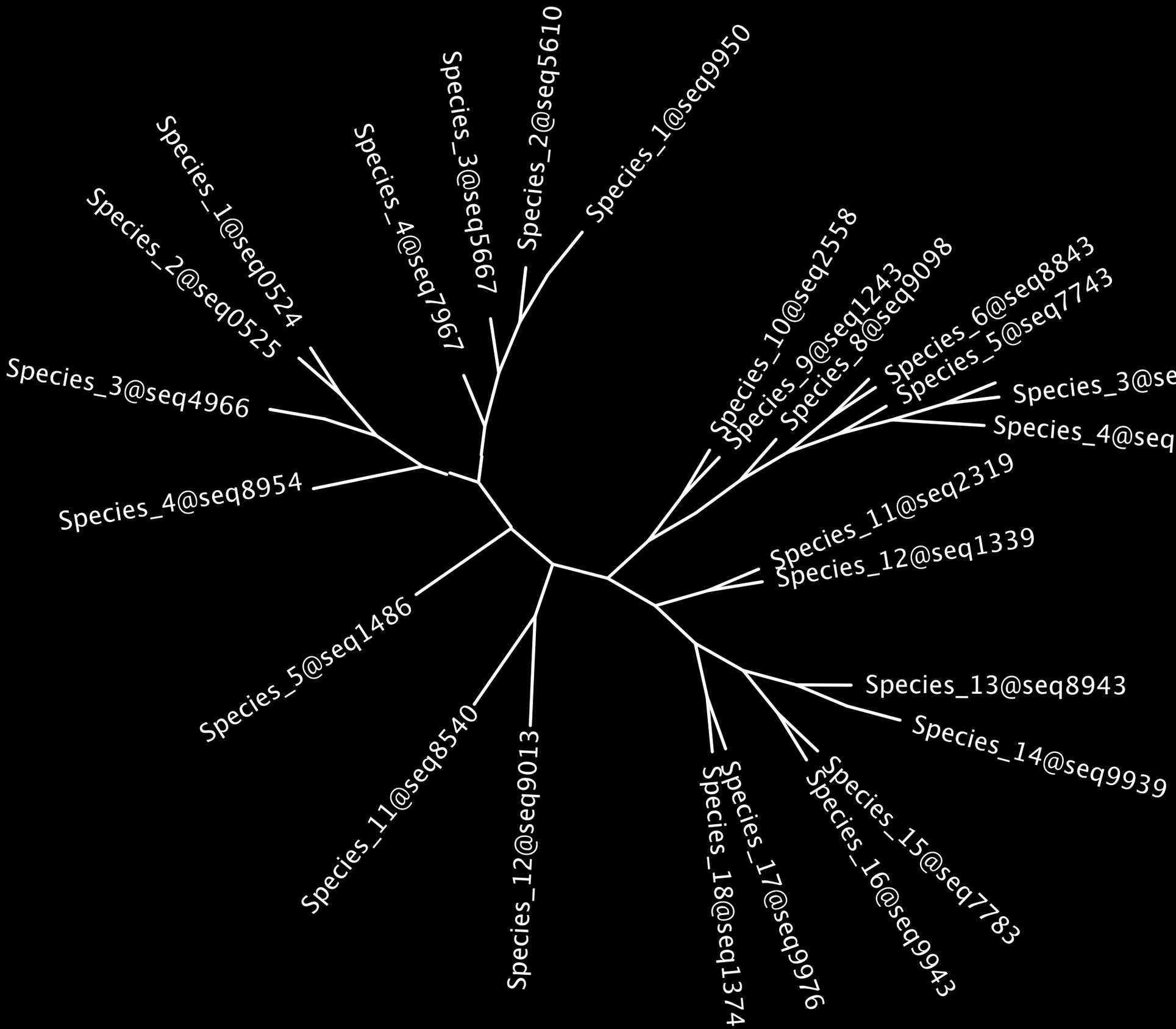
Gene divergence due to duplication

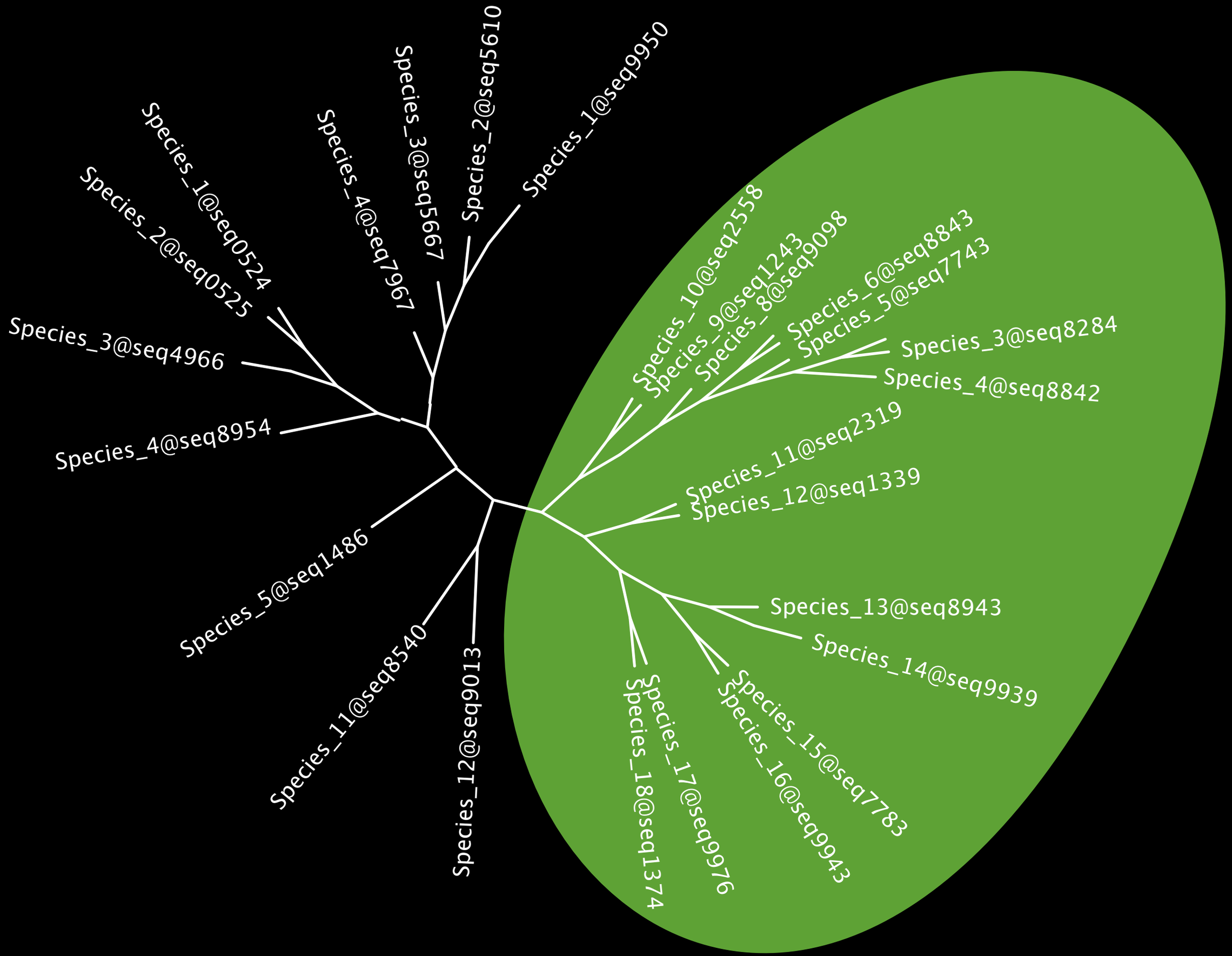
Species A

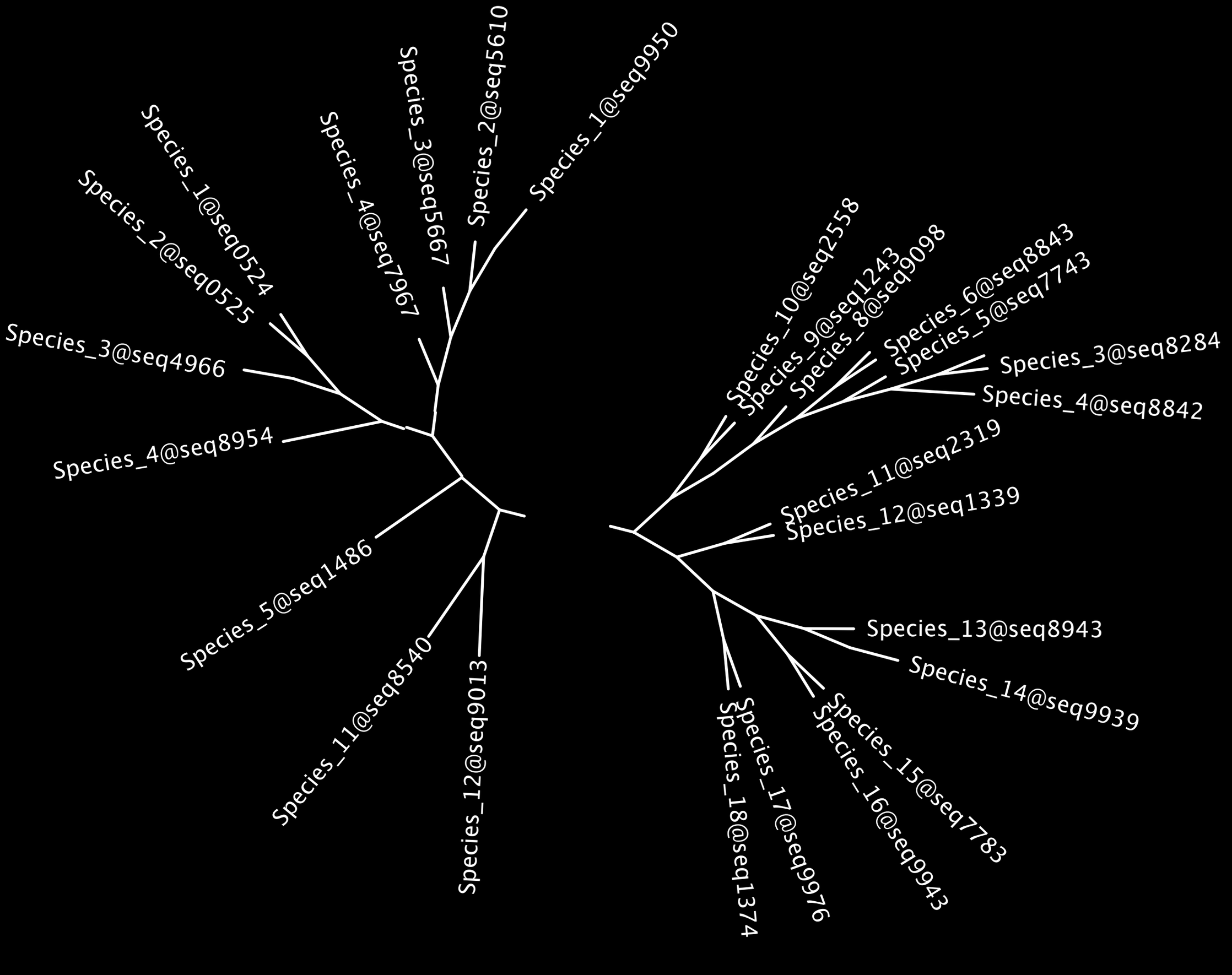
Species B

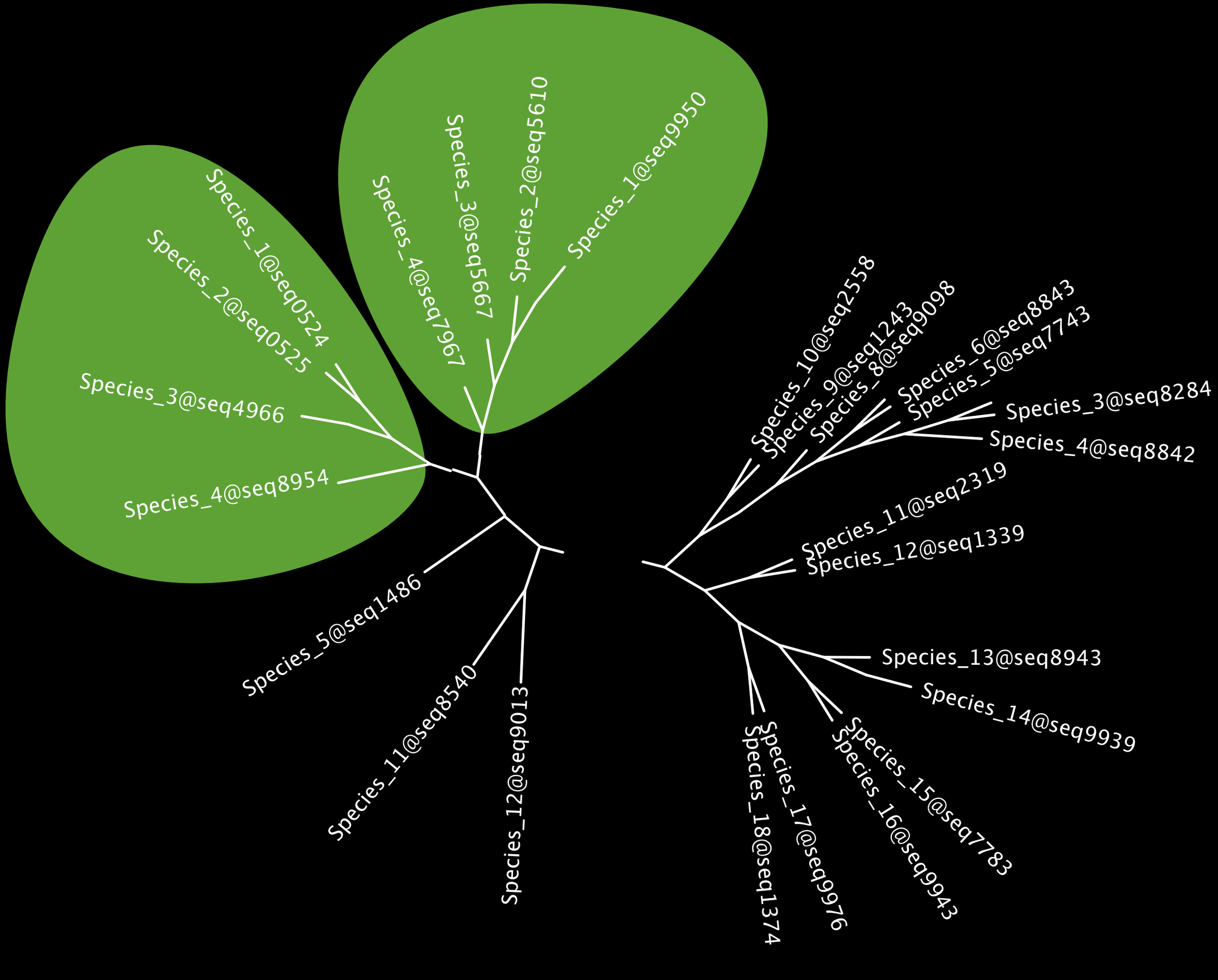
Species C

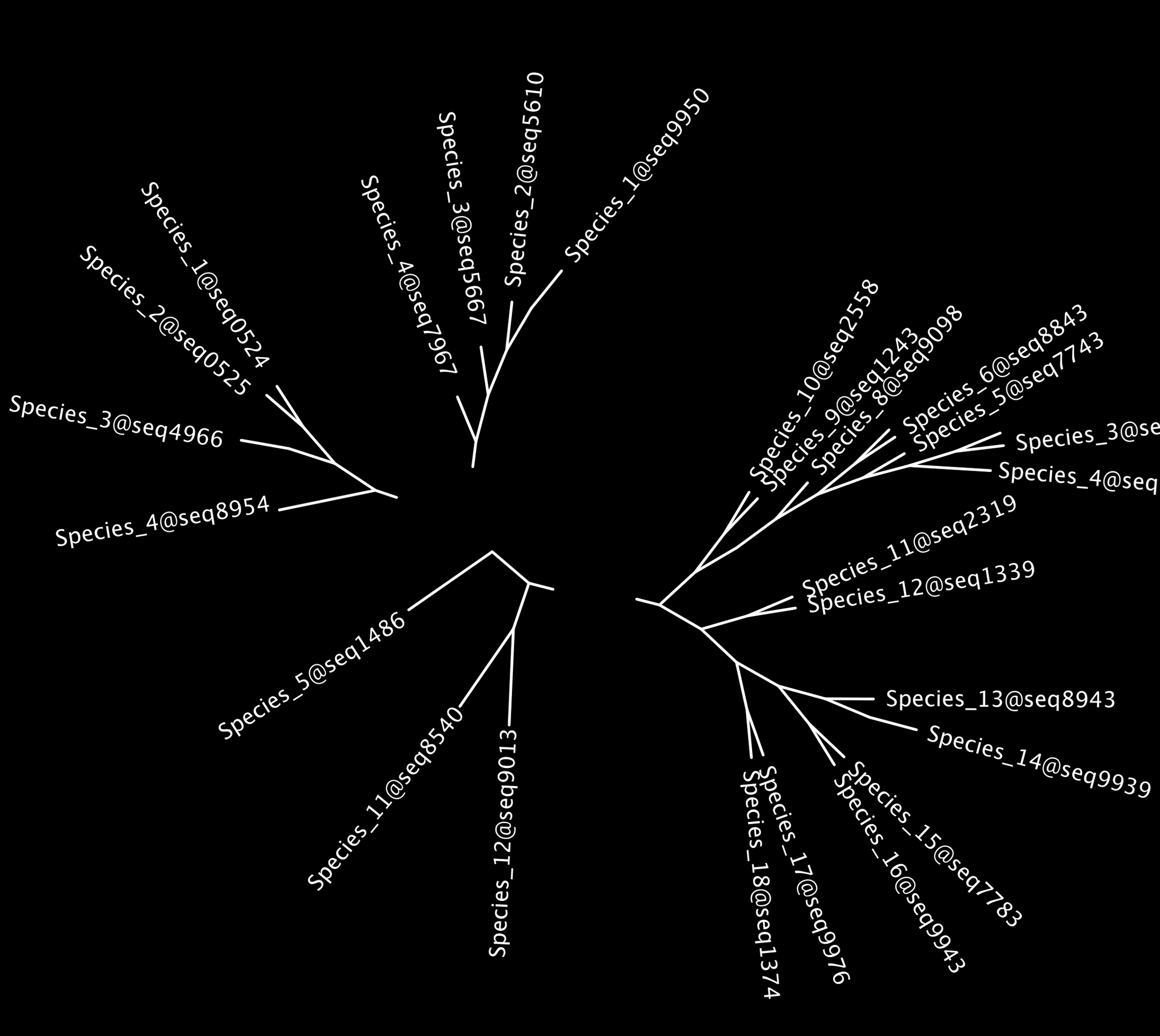
Gene divergence
due to duplication


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


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_5@seq1486
Species_11@seq8540
Species_12@seq9013
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_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...

Once we have subtrees of orthologs...

Align each ortholog

Once we have subtrees of orthologs...

Align each ortholog

Build trees

Once we have subtrees of orthologs...

Align each ortholog

Build trees

Concatenate into supermatrix

Once we have subtrees of orthologs...

- Align each ortholog
- Build trees
- Concatenate into supermatrix
- Other multilocus tools (Scott)

# 77 taxa, 150 Genes, >20k aa


White cells indicates sampled gene

50.9% gene sampling

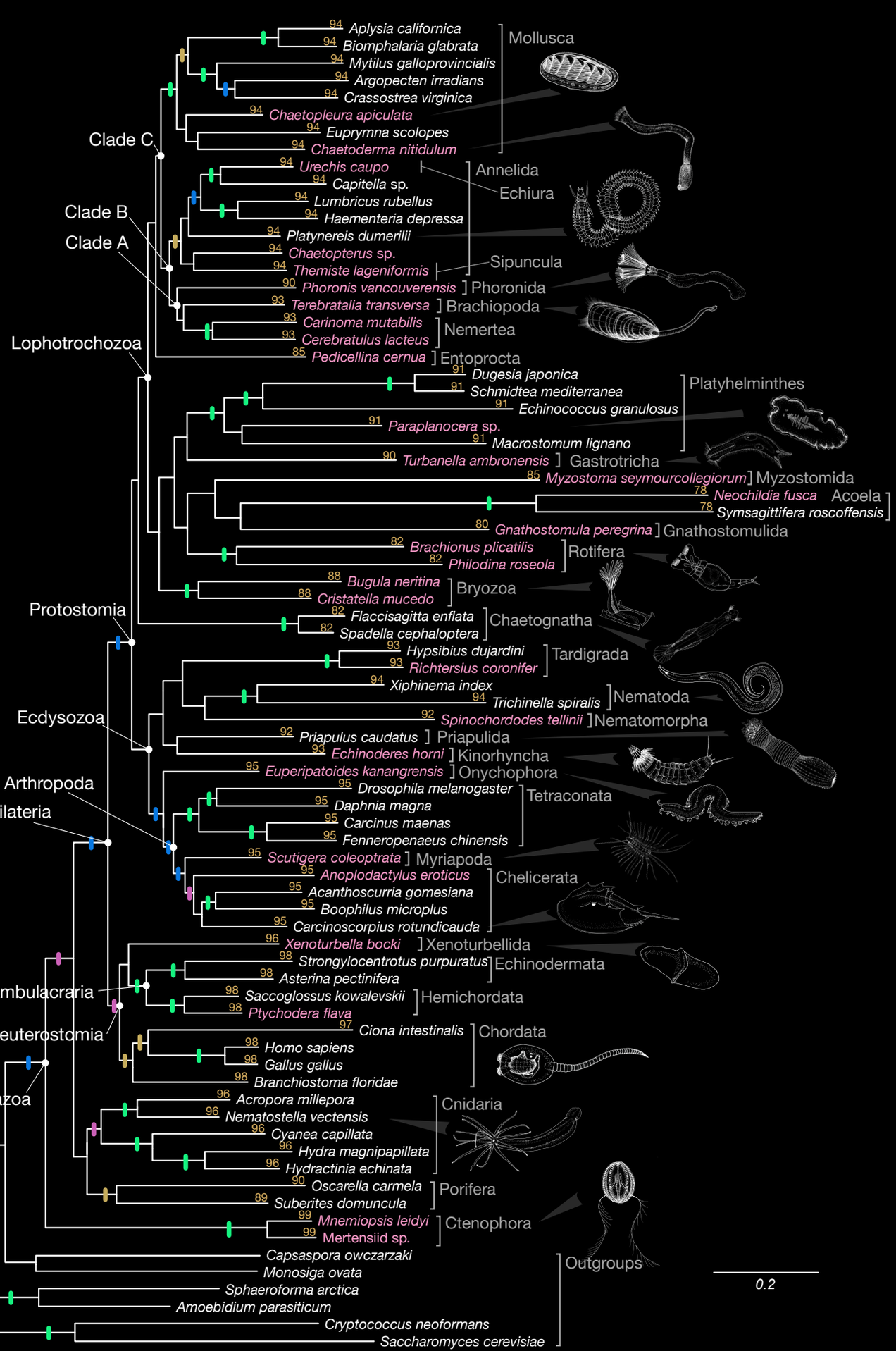
Dunn *et al.*, 2008
doi:10.1038/nature06614

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



Many exciting areas of active research and opportunities for major improvements

Raw sequence reads

*Assembly*

A large set of sequences

Subsets of homologous sequences

*Alignment*

Columns with homologous characters

# Unified modeling of gene duplication, loss, and coalescence using a locus tree

Matthew D. Rasmussen[1] and Manolis Kellis[1]

*Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; Broad Institute, Cambridge, Massachusetts 02139, USA*

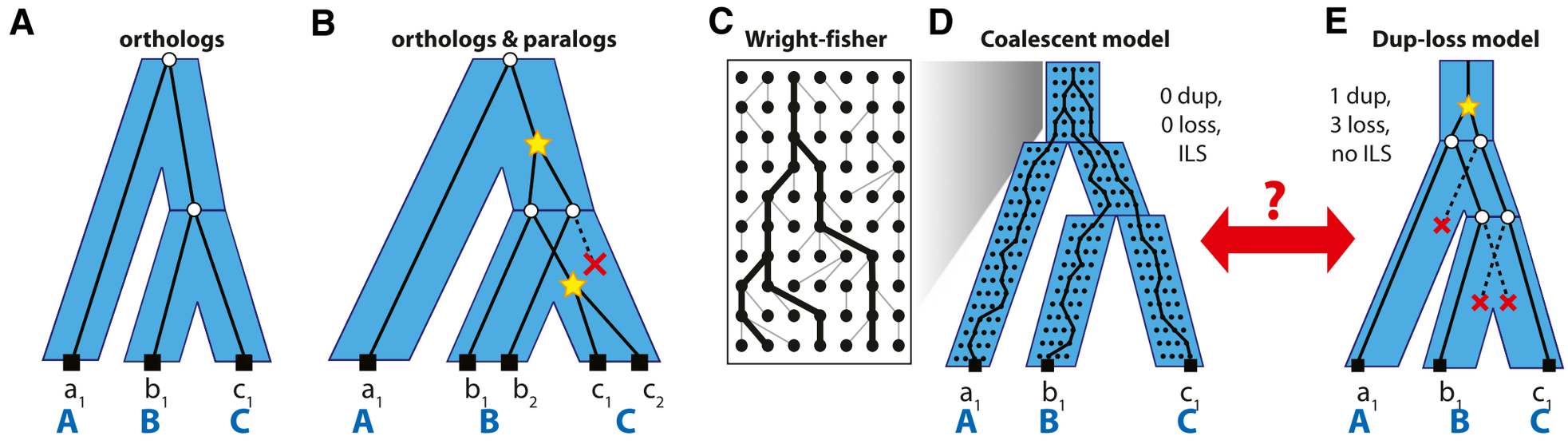



**Figure 1.** Different views of gene trees and species trees. (*A*) In the dup-loss model, a congruent gene tree and species tree indicates that all genes are orthologs. (*B*) Incongruence indicates the presence of gene duplications (stars) and gene losses (red "X"). (*C*) An example of the Wright-Fisher (WF) process and the coalescence of three lineages within the population. (*D*) A multispecies coalescent is a combination of WF processes for each branch of the species tree. In this model, no duplications or losses are allowed, but a gene tree can be incongruent due to a phenomenon known as incomplete lineage sorting (ILS). (*E*) In the dup-loss model, the same gene tree in panel *D* can be explained using one gene duplication and at least three gene losses. ILS cannot be modeled in the dup-loss model.

doi:10.1101/gr.123901.111