

Phylogenomics

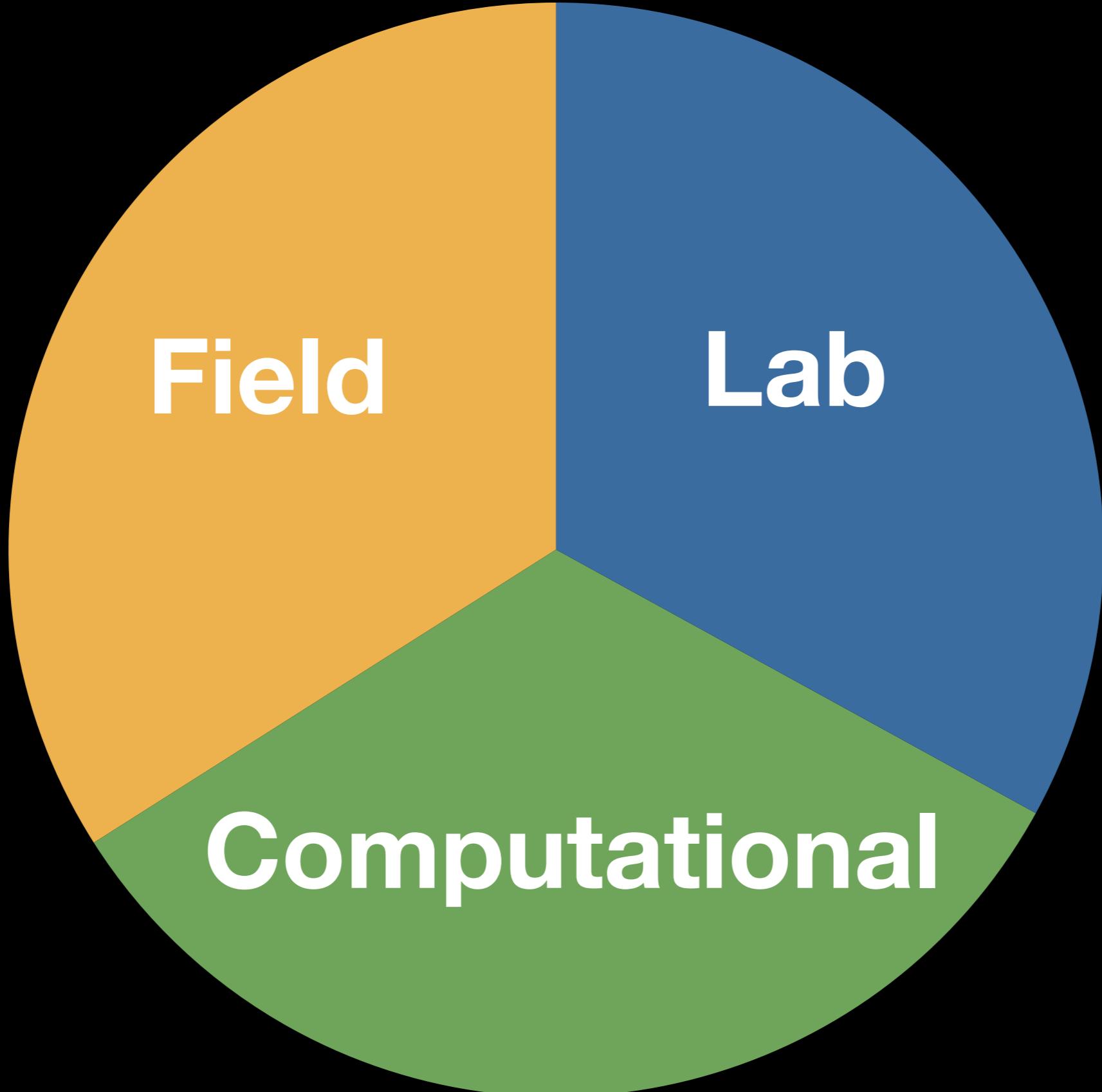
Casey Dunn

Professor, Yale University

Ecology and Evolutionary Biology

Twitter - @caseywdunn

<http://dunnlab.org>







Siebert et al, Zootaxa 2013

Re-evaluation of characters in Apolemidae (Siphonophora), with description of two new species from Monterey Bay, California

STEFAN SIEBERT¹, PHIL R. PUGH², STEVEN H. D. HADDOCK³ & CASEY W. DUNN¹





<https://vimeo.com/103529382>



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*



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:

Sample preparation and sequencing

Number of taxa

Phylogenetic diversity

The Future...

“classical” molecular phylogenetics

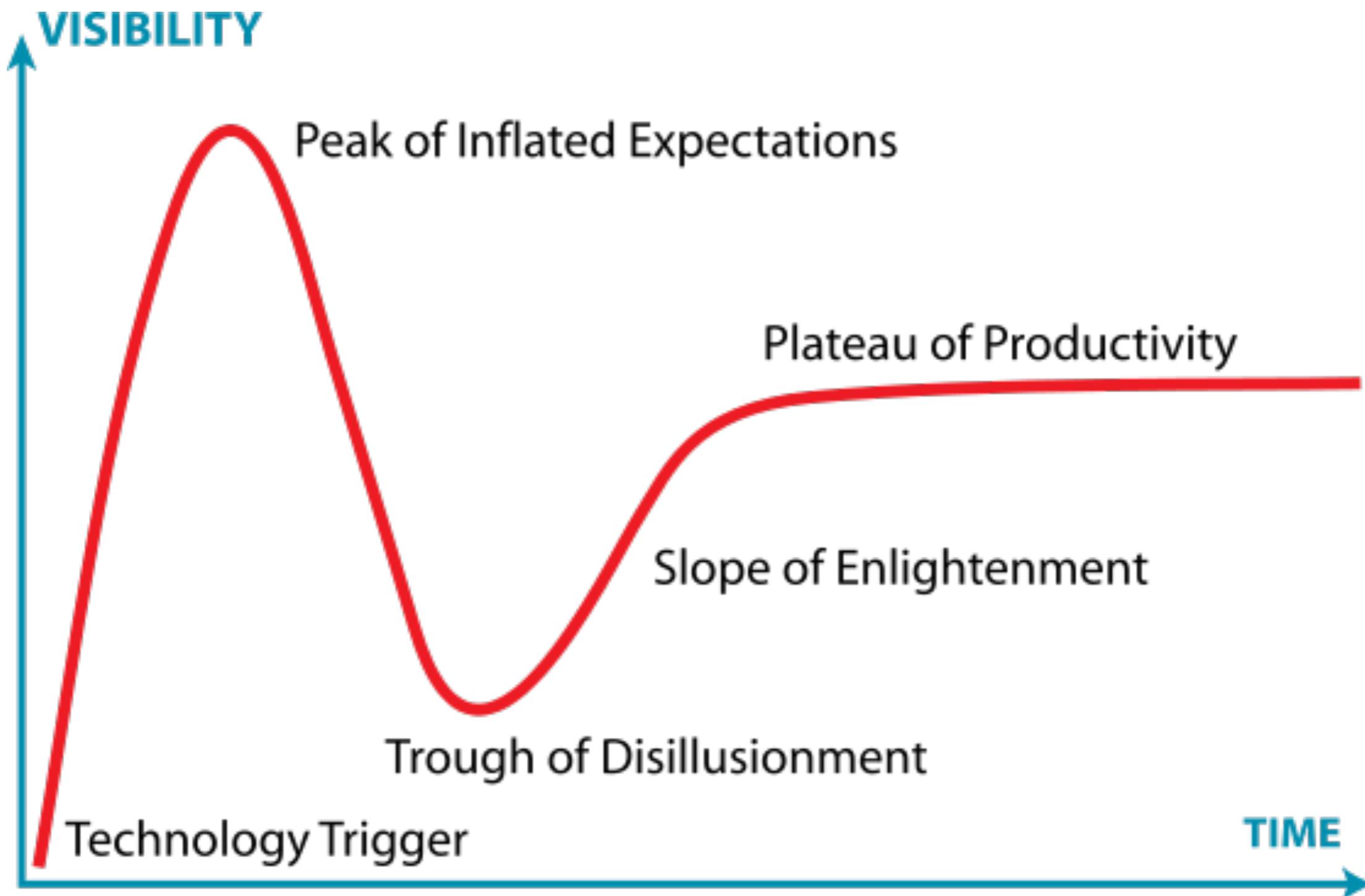
Phylogenomics

Number of genes



DNA sequencing is
getting cheaper

The Gartner Hype Cycle*



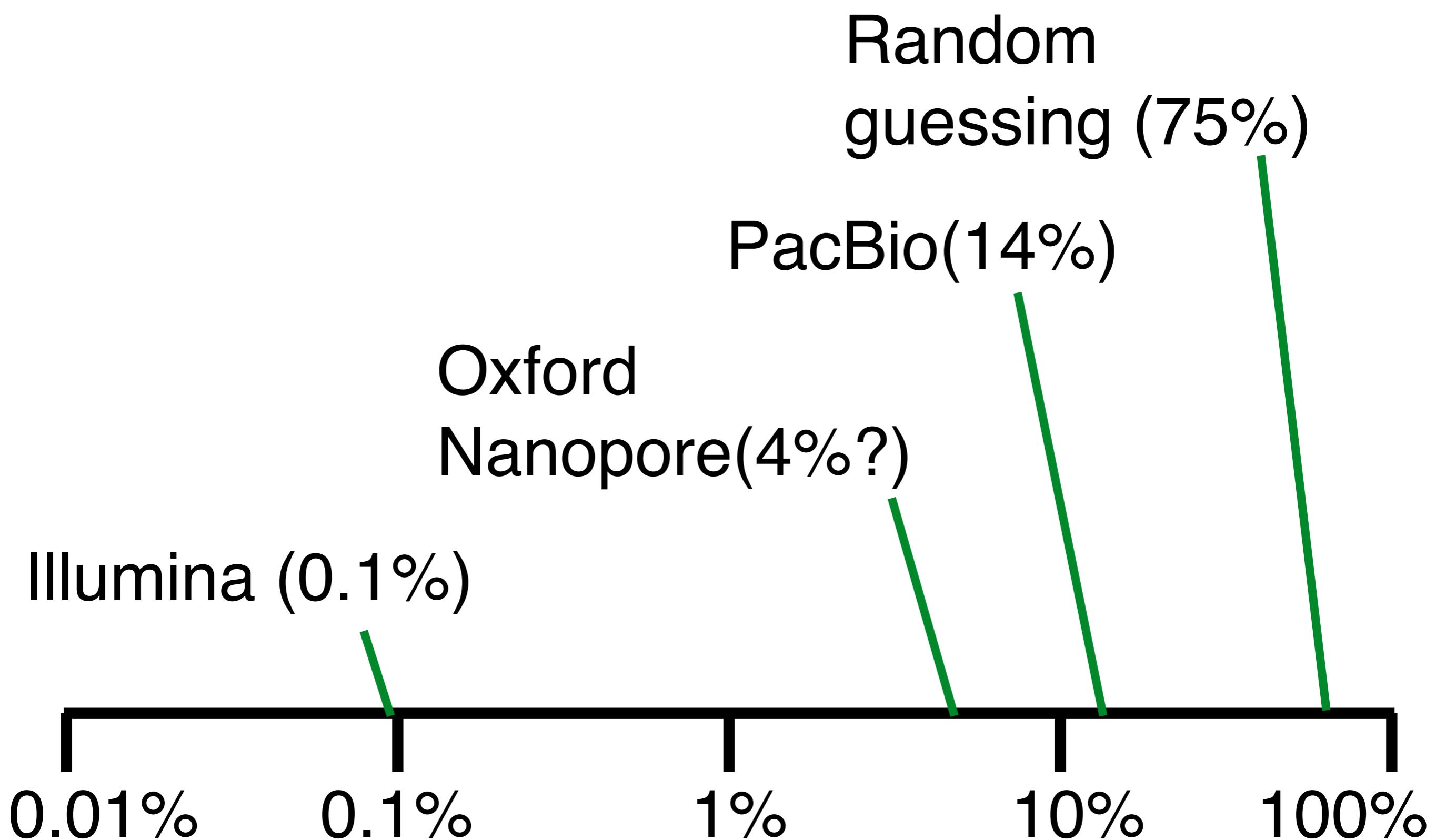
* Not really a cycle

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

Many different sequencers
now, with tradeoffs in:

- Error rate
- Read length
- Per-base run cost
- Sequencer cost

Error rate comparison (2015)



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!

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)?

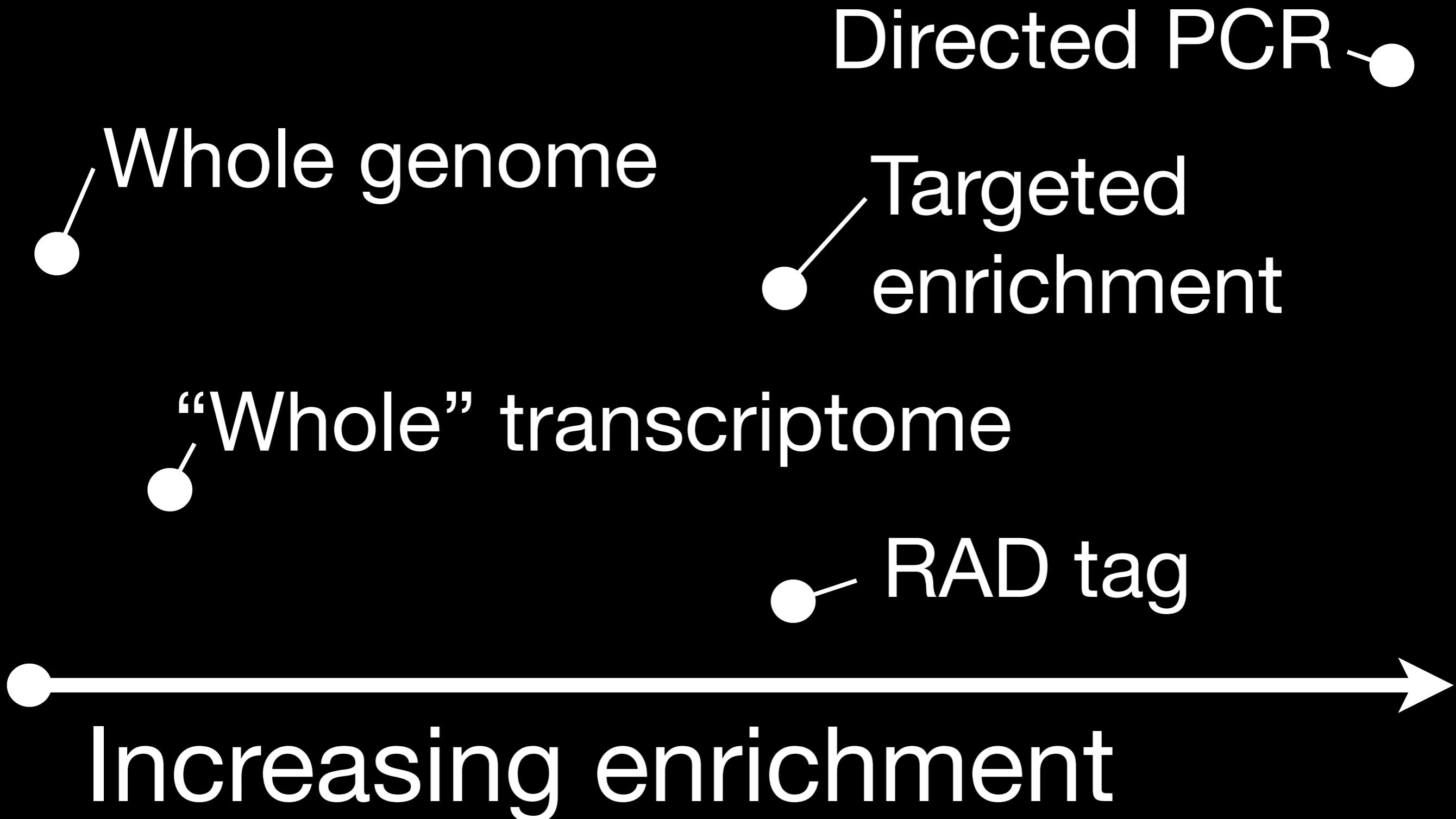
Central technical question:

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 a subset of homologous genome regions across multiple individuals and species.

Enrichment spectrum



Whole genome

Directed PCR

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.

Whole genome

Directed PCR

Targeted enrichment

“Whole” transcriptome

RAD tag

Increasing enrichment

Whole transcriptome

- Enriched for expressed protein coding genes
- There is no One True Transcriptome

Whole genome

Directed PCR

Targeted
enrichment

“Whole” transcriptome

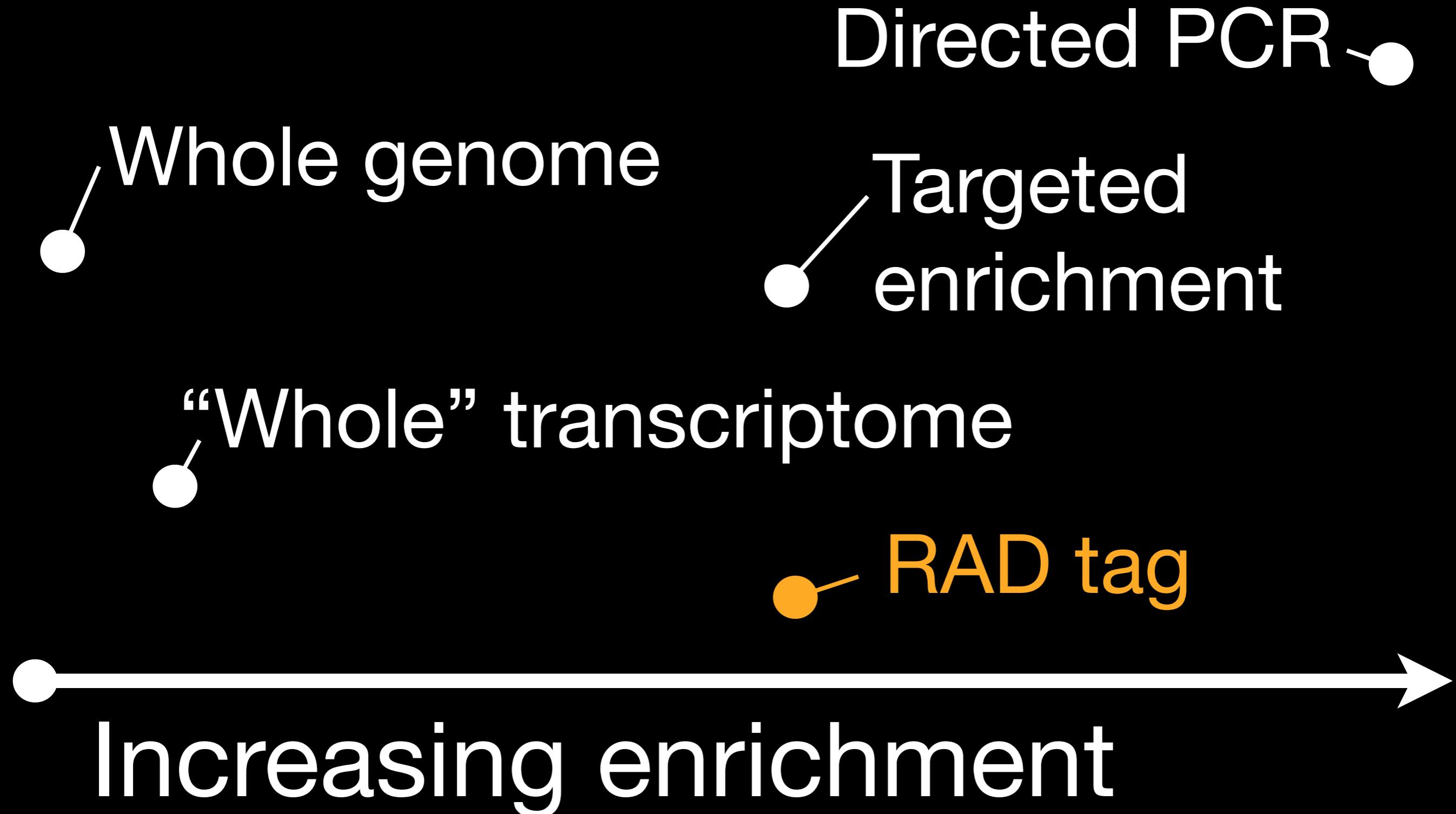
RAD tag

Increasing enrichment



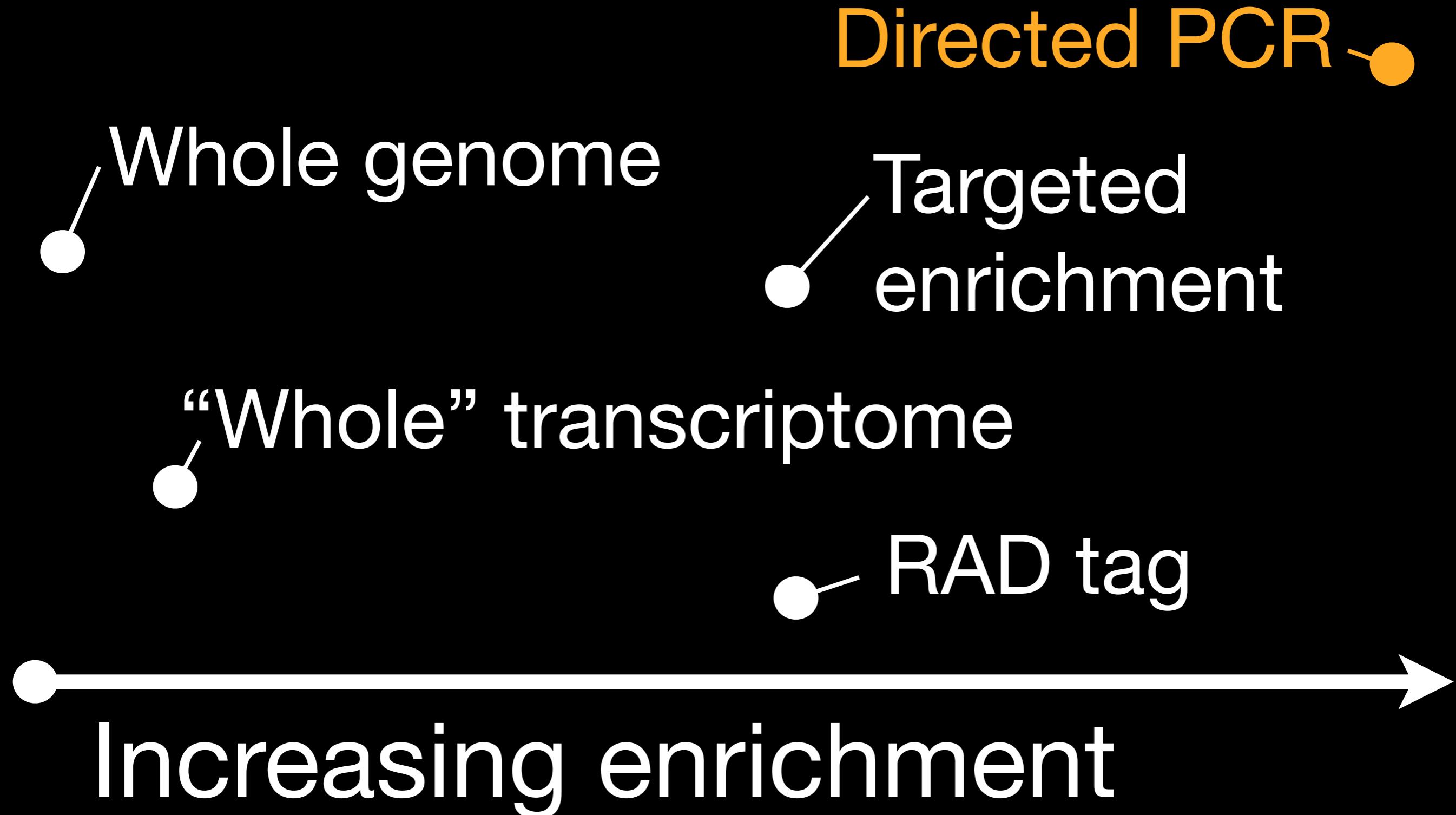
Targeted enrichment

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



RAD tag

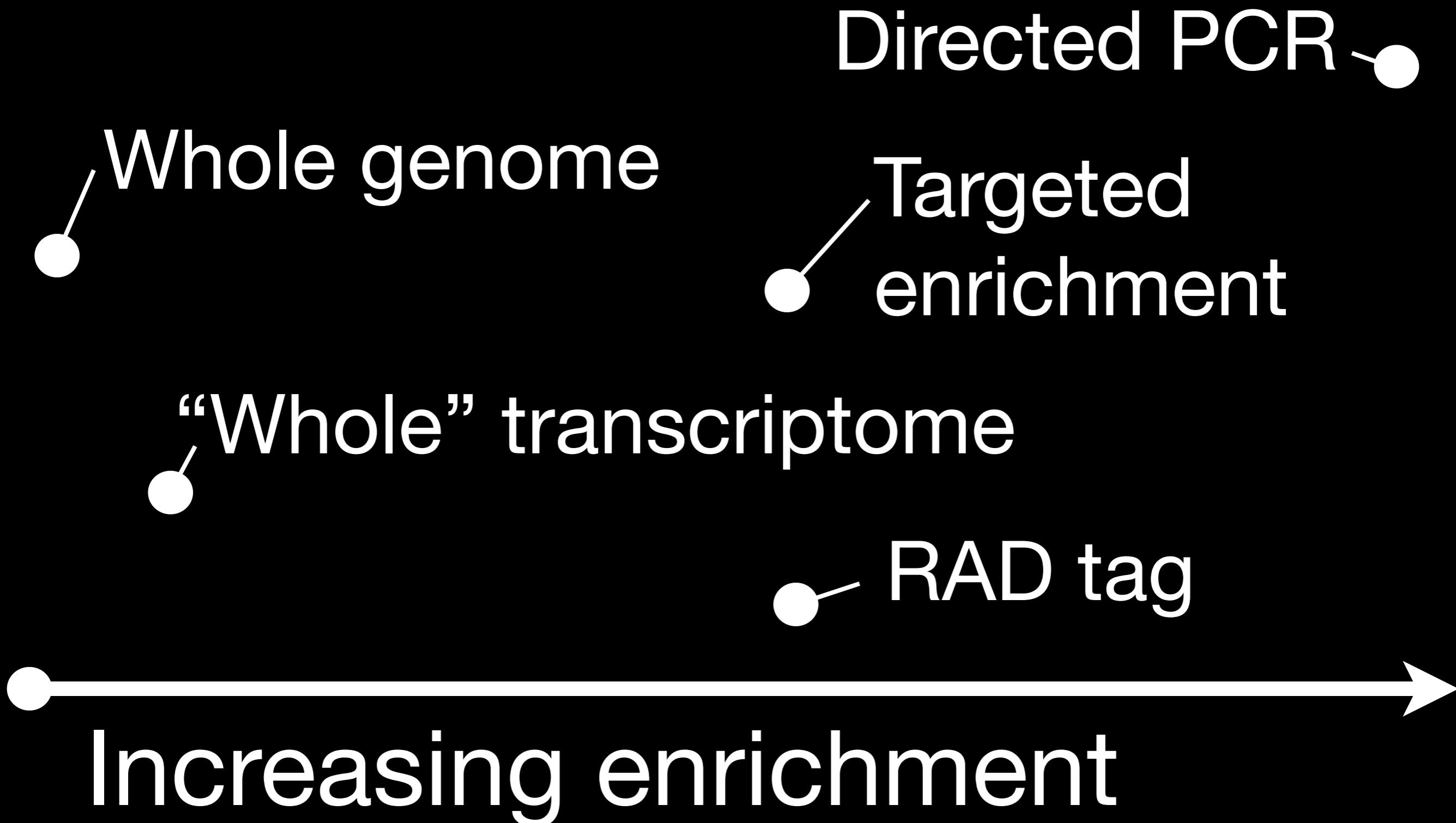
- Enriched for randomly distributed, but consistent, genome regions
- No need for specific probes



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.



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

**Whole genome
(de novo assembly)**

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

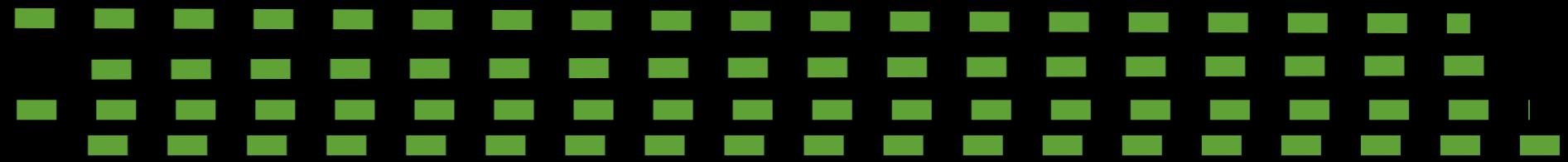
Starting
material

Fragments

Reads

Fragment ↓

Prepare library,
sequence ↓



The most common library preparation problems:

Poor input material

Over-amplification

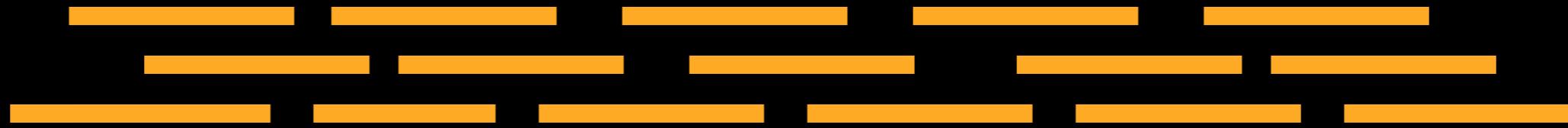
Poor size selection

Assembly

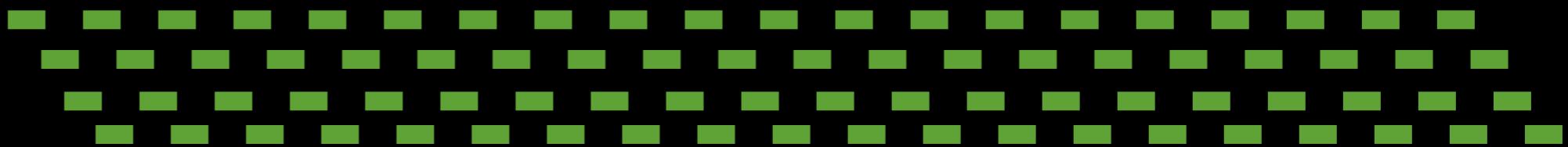
Assembly undoes
fragmentation (and
reduces redundancy).

Starting
material

Fragment ↓



Prepare library,
sequence ↓



Assembly ↓

Final
product



Overlap assemblers - puzzling together of long sequences based on overlap

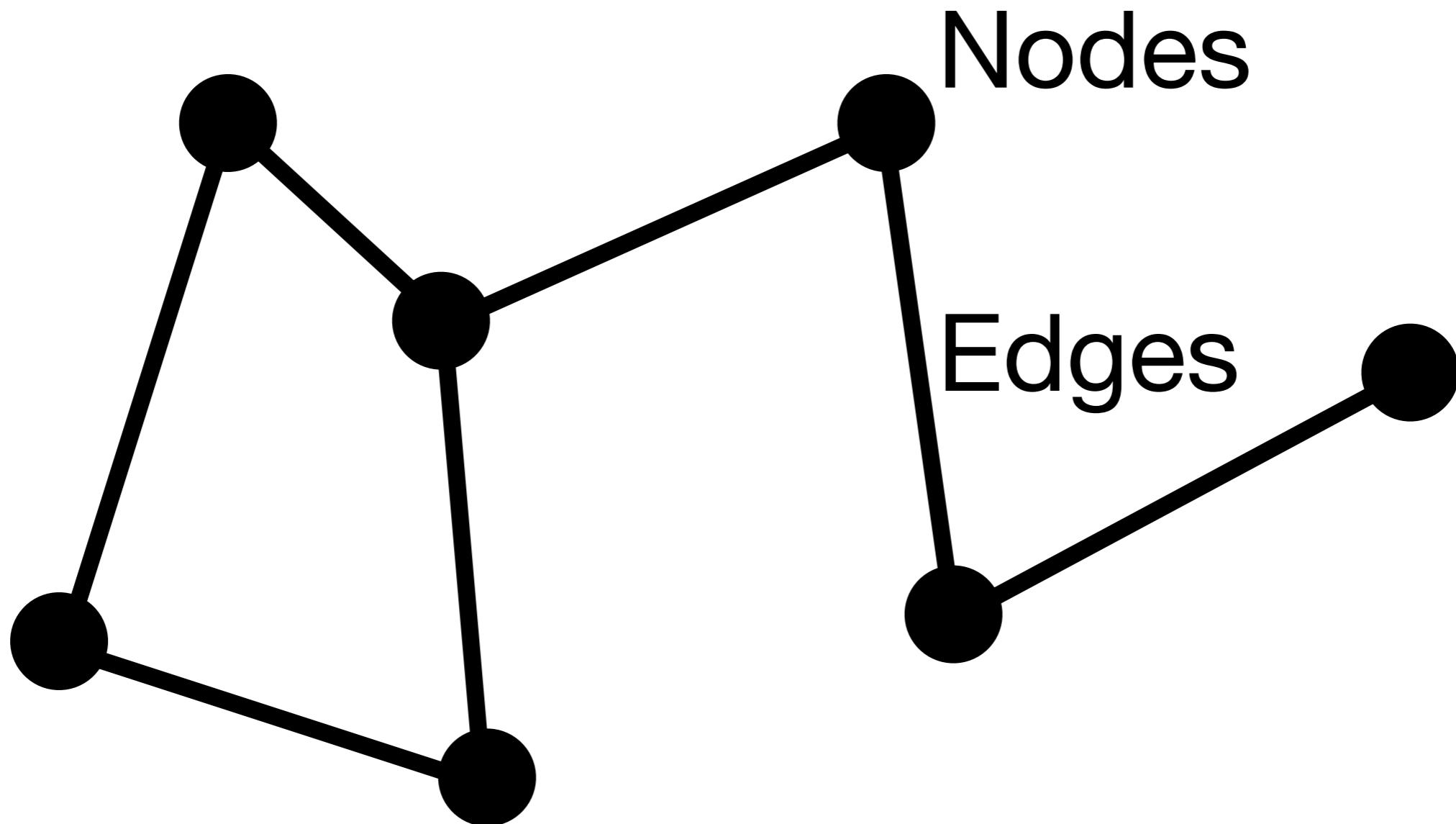
The number of pairwise
comparisons that are needed to
detect overlap become intractable

de Bruijn graph assemblers have been developed to meet these challenges

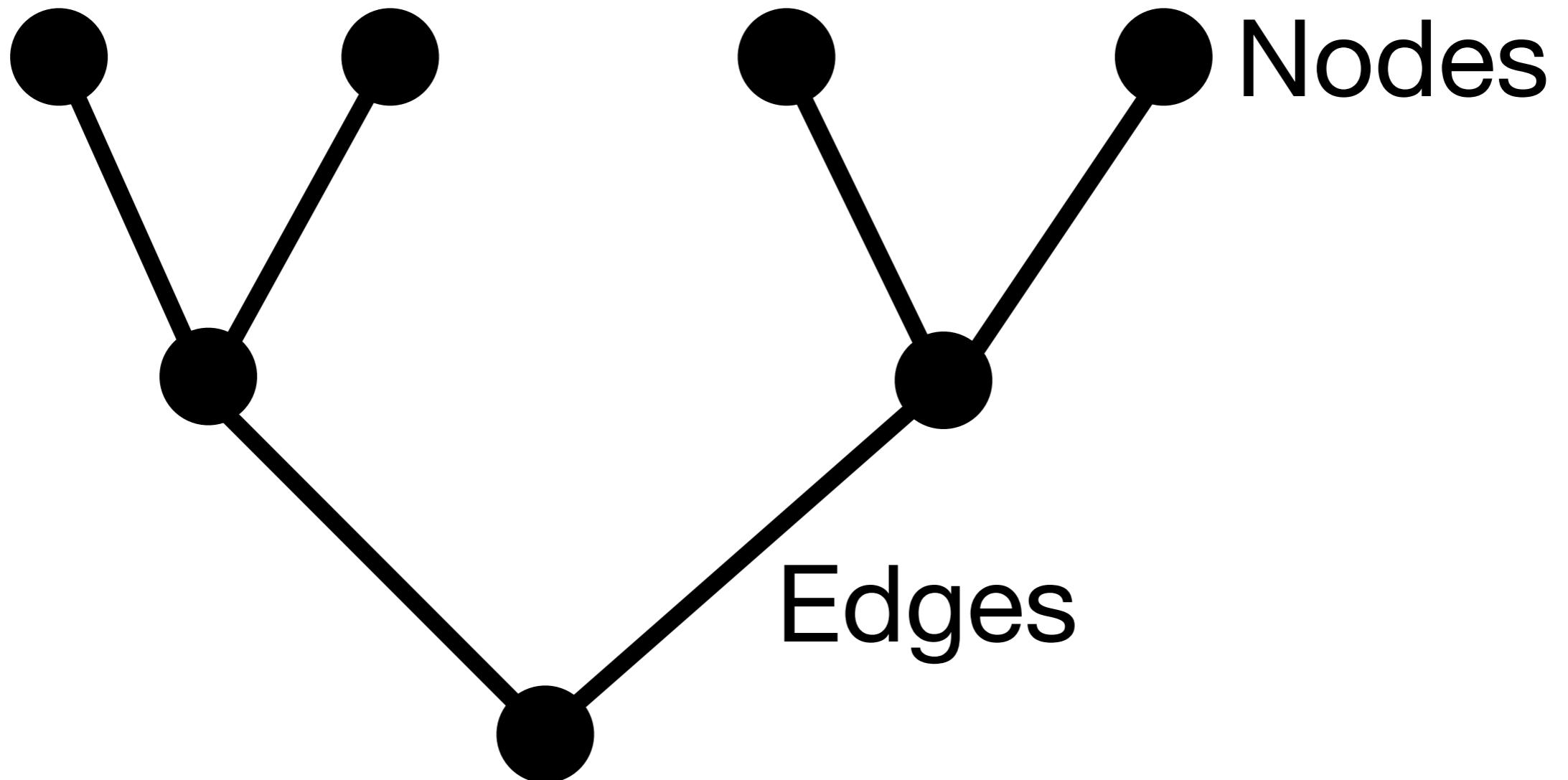
Better defined memory footprint

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

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 hash table

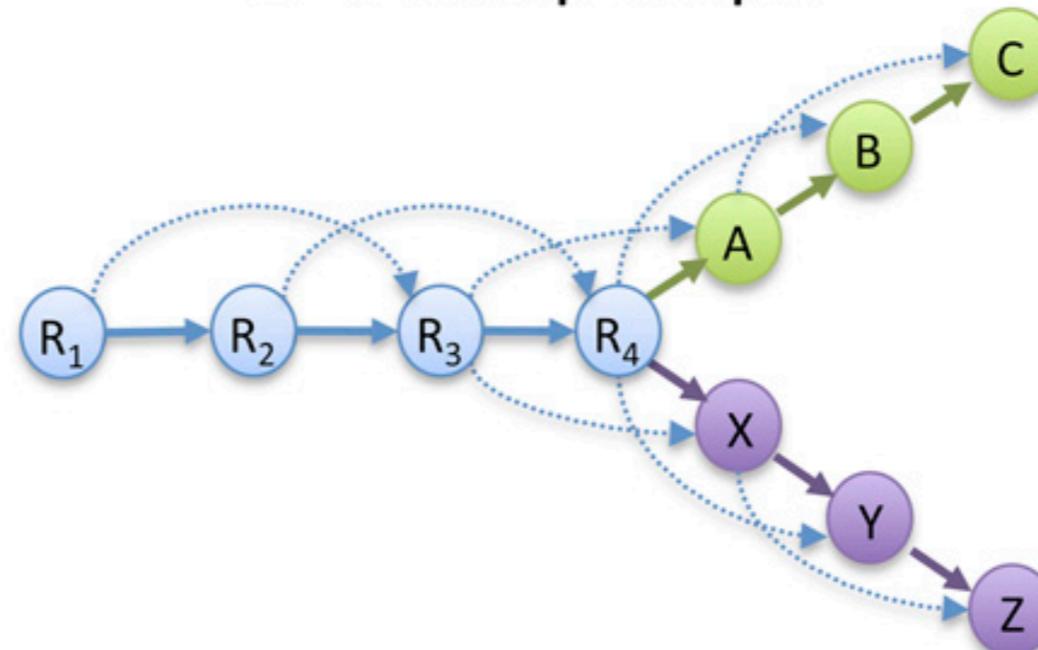
Each node corresponds to a k-mer sequence from the hash table

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

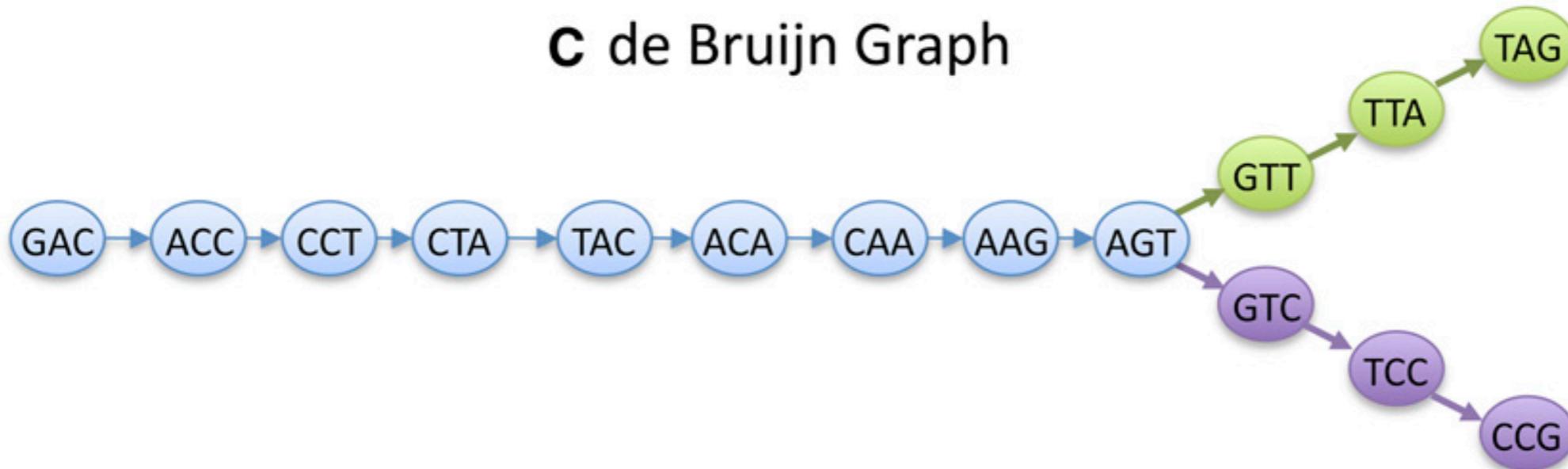
A Read Layout

$R_1:$	GACCTACA
$R_2:$	ACCTACAA
$R_3:$	CCTACAAG
$R_4:$	CTACAAGT
A:	TACAAGTT
B:	ACAAGTTA
C:	CAAGTTAG
X:	TACAAGTC
Y:	ACAAGTCC
Z:	CAAGTCCG

B Overlap Graph



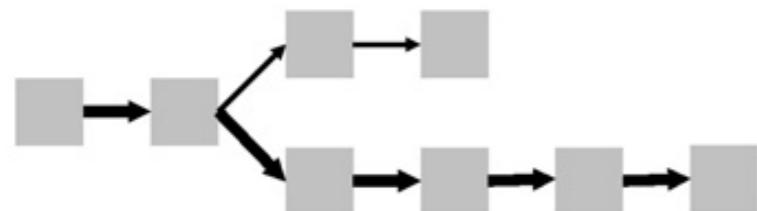
C de Bruijn Graph



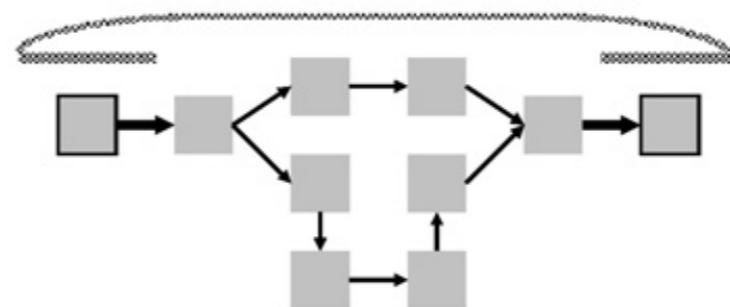
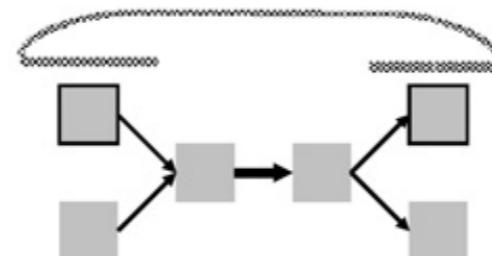
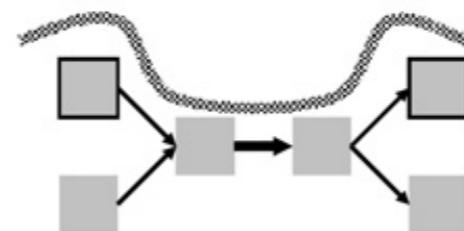
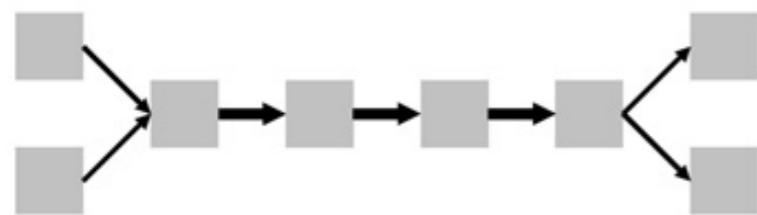
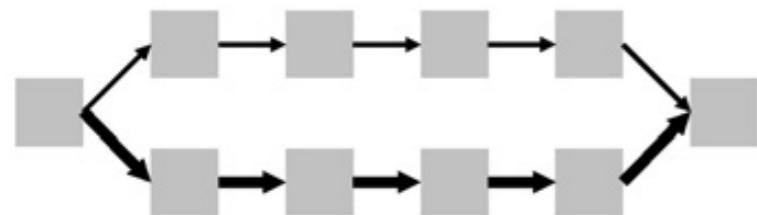
Paths through the de Bruijn graph are assembled sequences

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

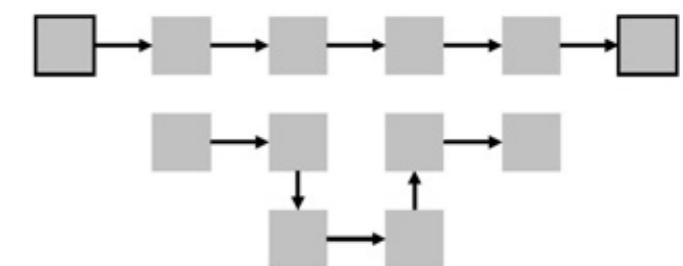
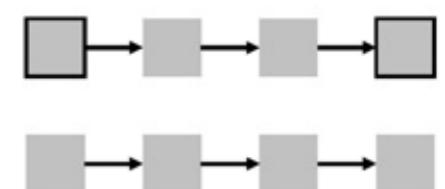
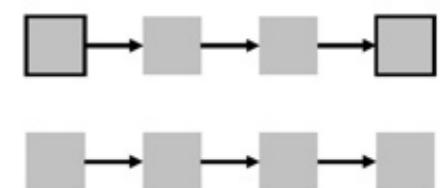
The graphs require considerable post-processing to simplify them (pop bubbles, trim dead ends, etc)



(before)



(after)



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

Read (sequence data)

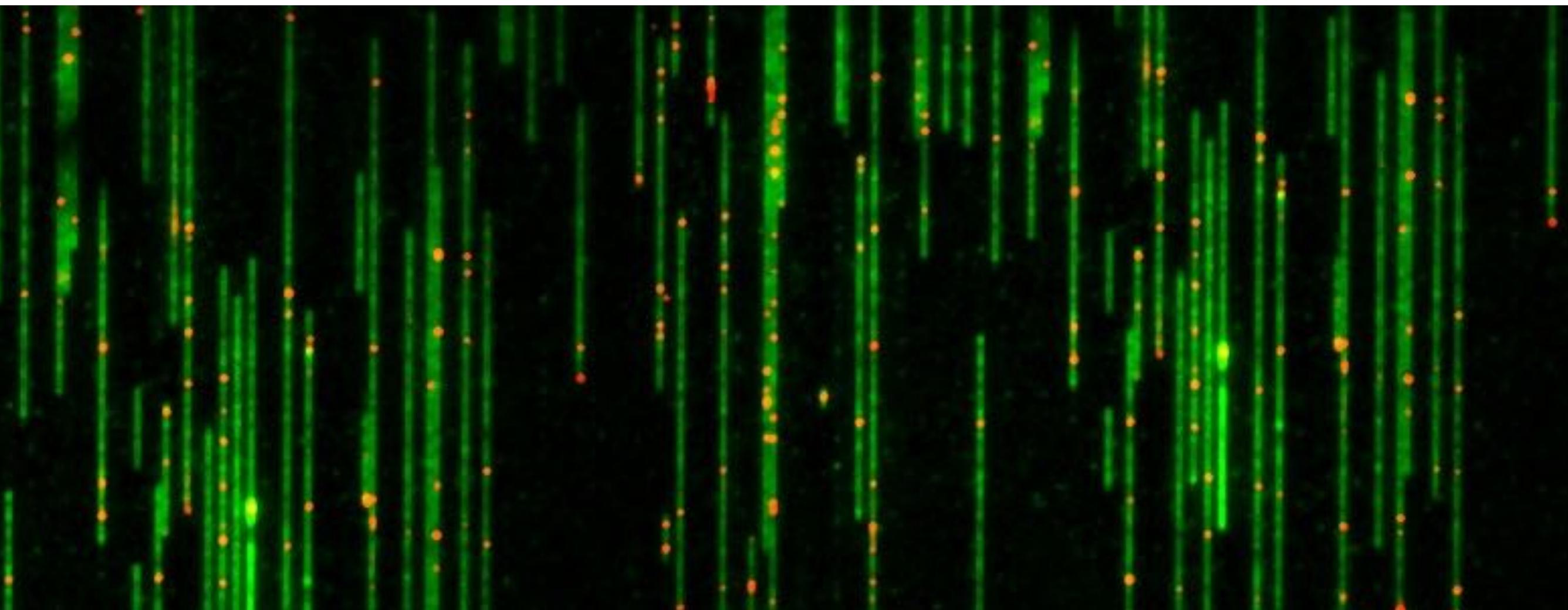


DNA molecule

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
- Proximity ligation

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



Sequence read pairs that are variable distances apart, up to chromosome length



All reads from the same large fragment have the same barcode

<https://vimeo.com/120429438>

Can be used to stitch together assembly fragments

Recent advances

Key improvements relevant to
de novo genome sequencing in
last couple years:

- Improved length and quality
of long read sequencers
- Improved structural data
- Improved assemblers

Longer reads:

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



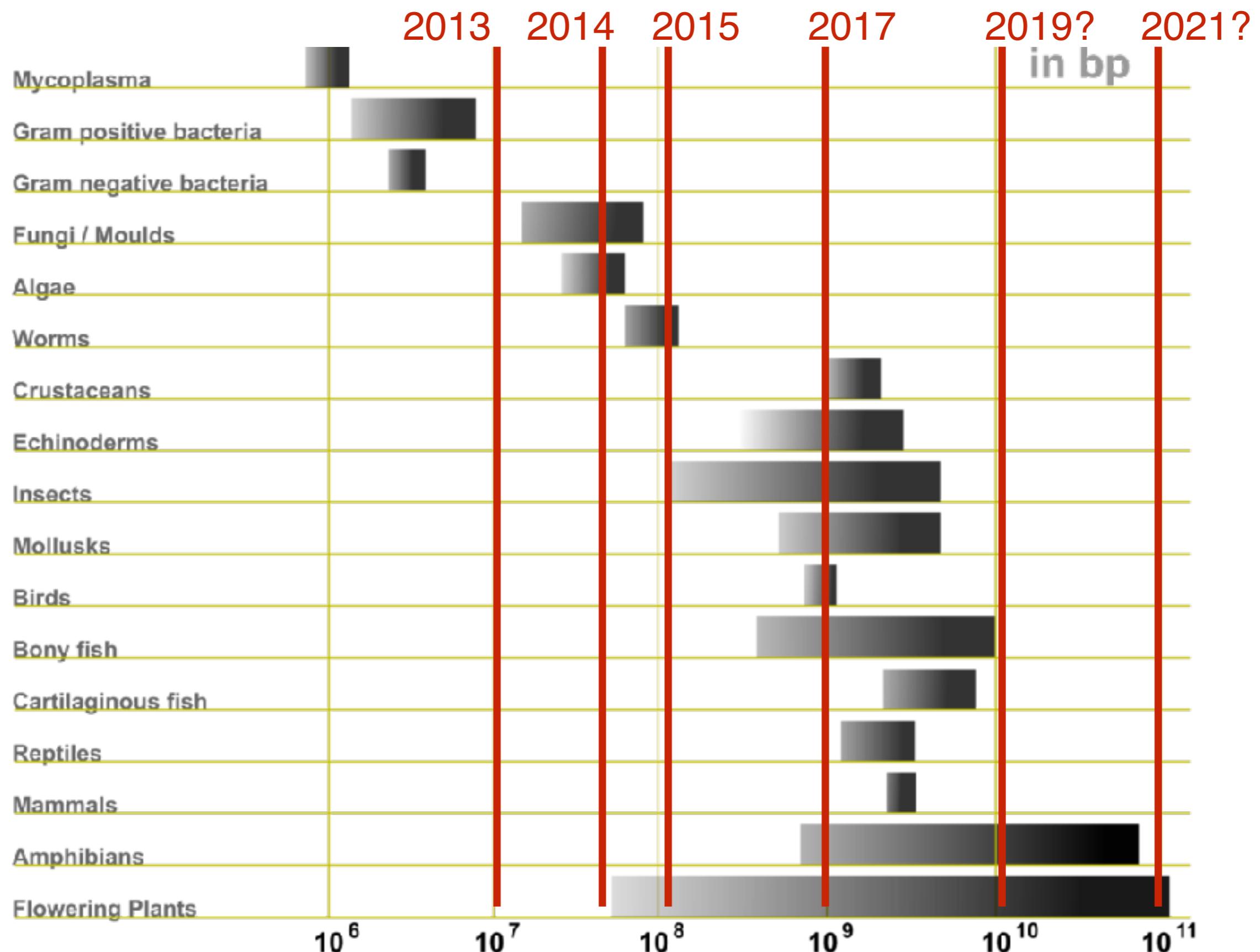
New Results

De Novo PacBio long-read and phased avian genome assemblies correct and add to genes important in neuroscience research

Jonas Korlach, Gregory Gedman, Sarah Kingan, Jason Chin, Jason Howard, Lindsey Cantin, Erich D. Jarvis

doi: <https://doi.org/10.1101/103911>

Tractable genome size on an NSF grant...



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

Medium-sized genomes can be tricky

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

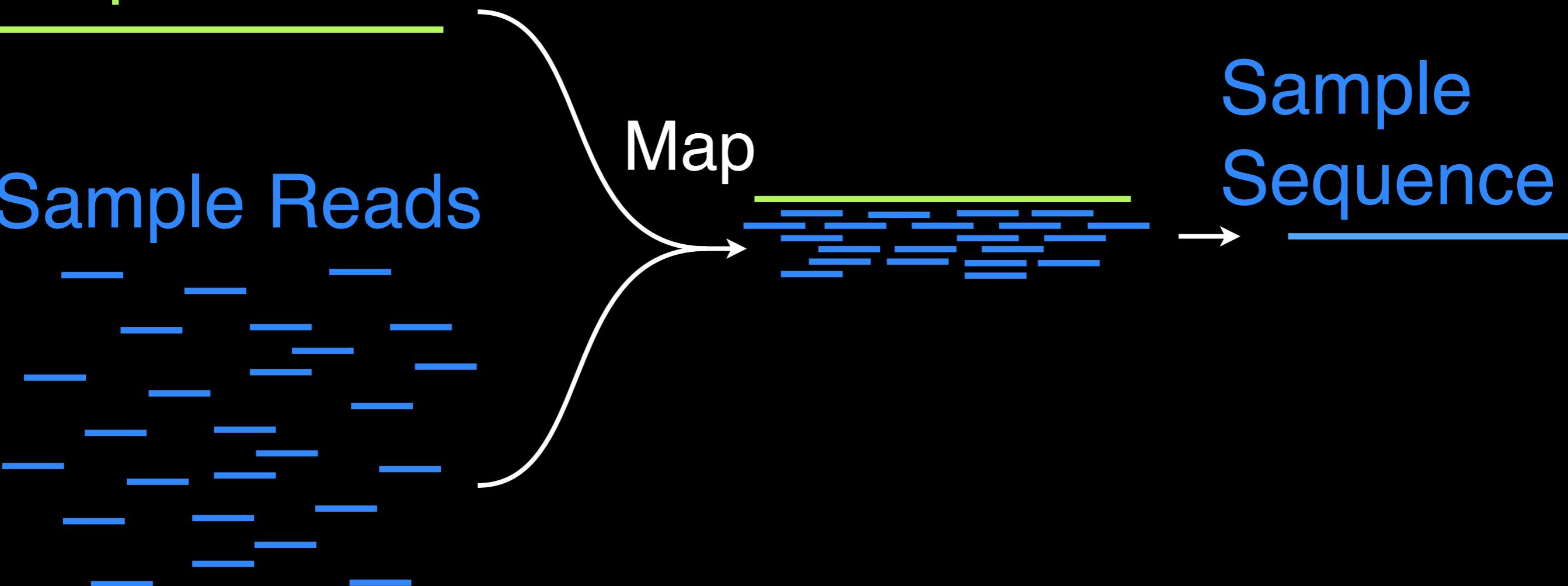
Annotation

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

RNA quality is (almost) Everything!

Avoid contamination

Reduced sample size requirements
have improved this

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

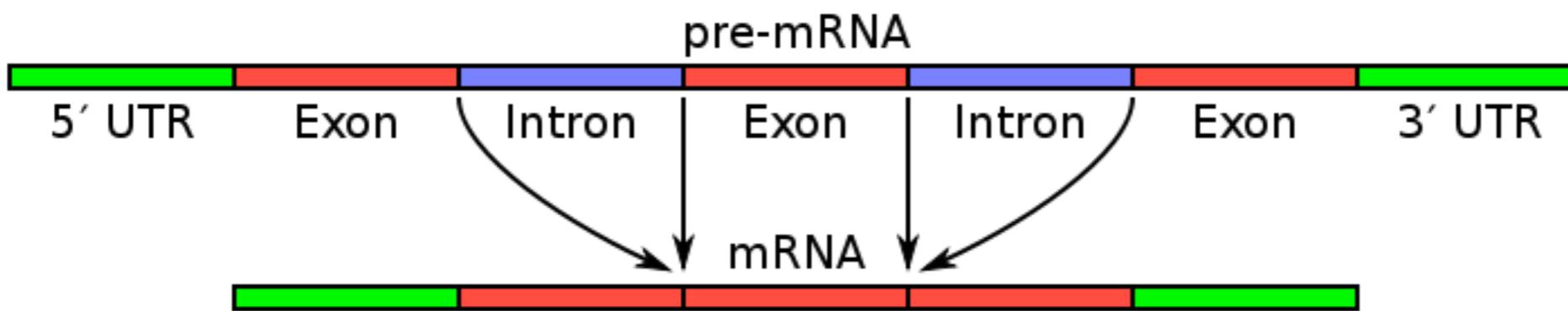
Transcriptome Assembly

Transcriptome assembly has
the same challenges as
genome assembly...

... and then some.

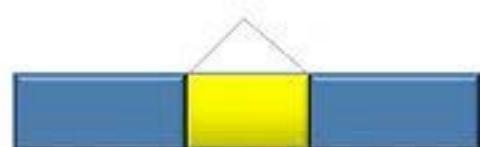
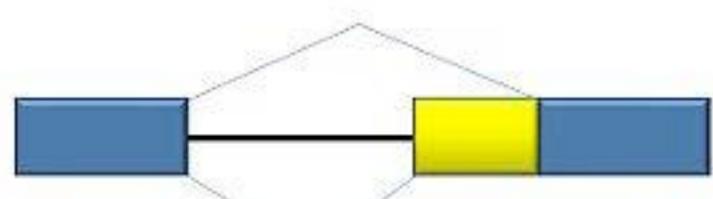
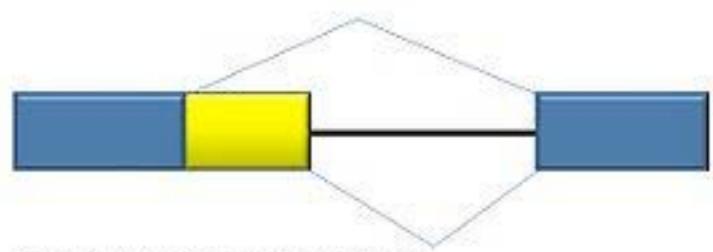
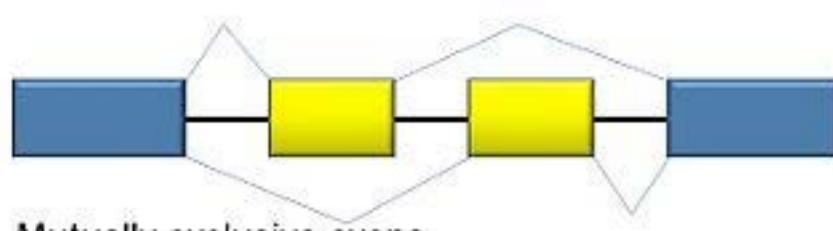
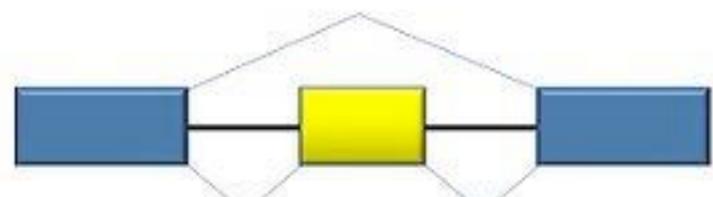
Transcript splicing

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

Intron retention

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

Assembly results...

Genome

...aagtca~~gtgg~~gagatgcaccatgagac~~cctt~~ggaagaagctgtccctggagacaatgtggg...

Transcript

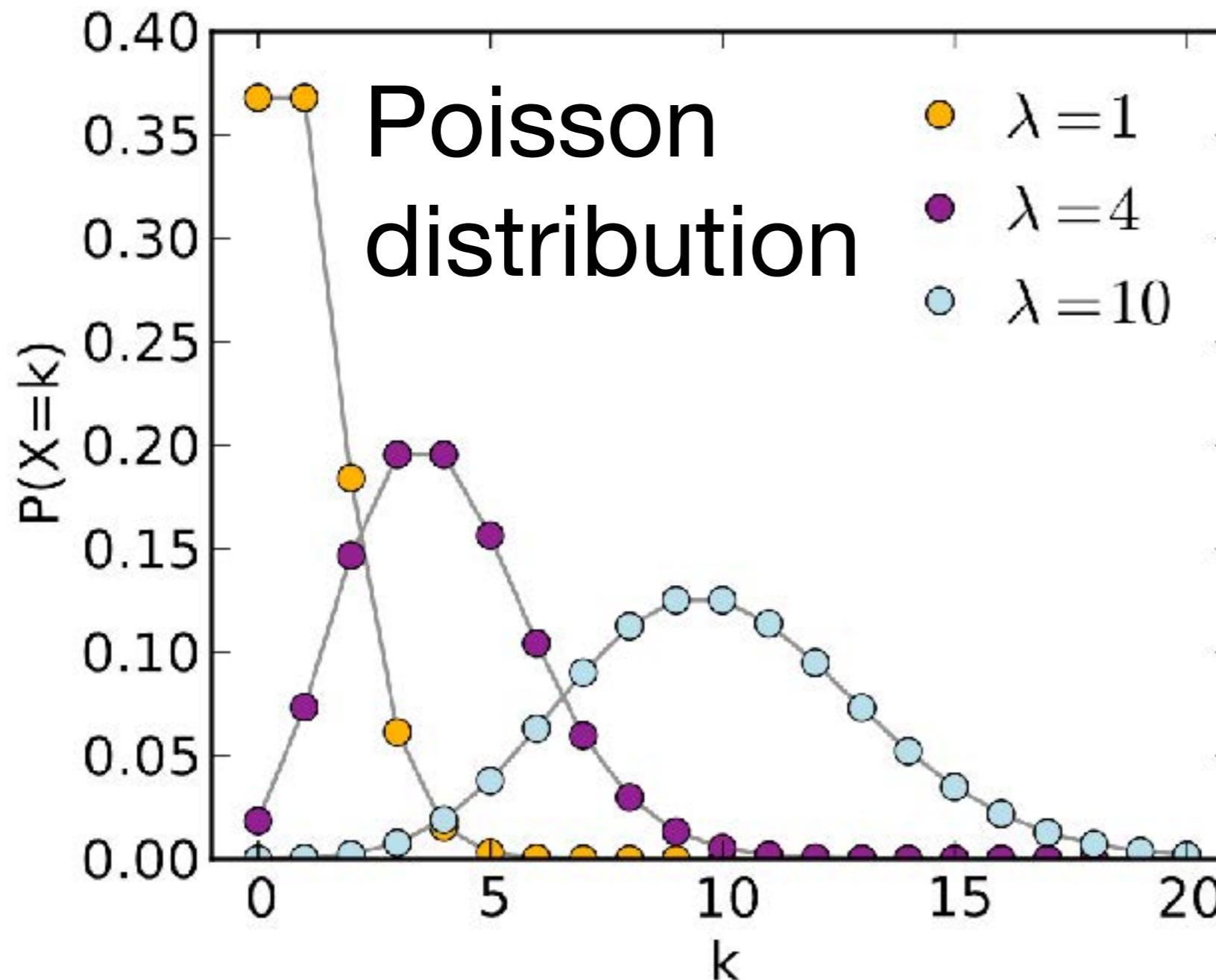
...aagtca~~gtgg~~gagatgcaccatgag
ccttggaaagaag ctgtccctgg gtc
agacaatgtggg...

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](https://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

The tool

<https://bitbucket.org/caseywdunn/agalma>

The screenshot shows the Bitbucket interface for the 'agalma' repository. At the top, there's a navigation bar with 'Bitbucket', 'Repositories', 'Create', and a search bar. Below the header, the repository name 'agalma' is displayed with a blue circular icon containing 'Ag'. It shows the owner 'caseywdunn', a 'Following' button, and a 'Share' button. To the right are buttons for 'Clone', 'Fork', 'Compare', and 'Pull request'. Below this, a navigation menu includes 'Overview', 'Source', 'Commits', 'Pull requests', 'Issues (1)', 'Downloads (0)', and a gear icon for settings.

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

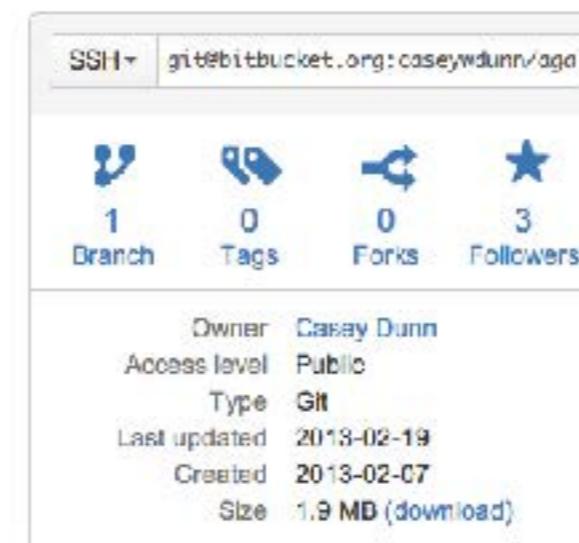
Agalma is a set of analysis pipelines for transcriptome assembly (paired-end Illumina data) and phylogenetic analysis. It can import gene predictions from other sources (eg, assembled non-Illumina transcriptomes or gene models from annotated genomes), enabling broadly-sampled "phylogenomic" analyses.

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 original 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 adaptors (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



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/10.1093/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

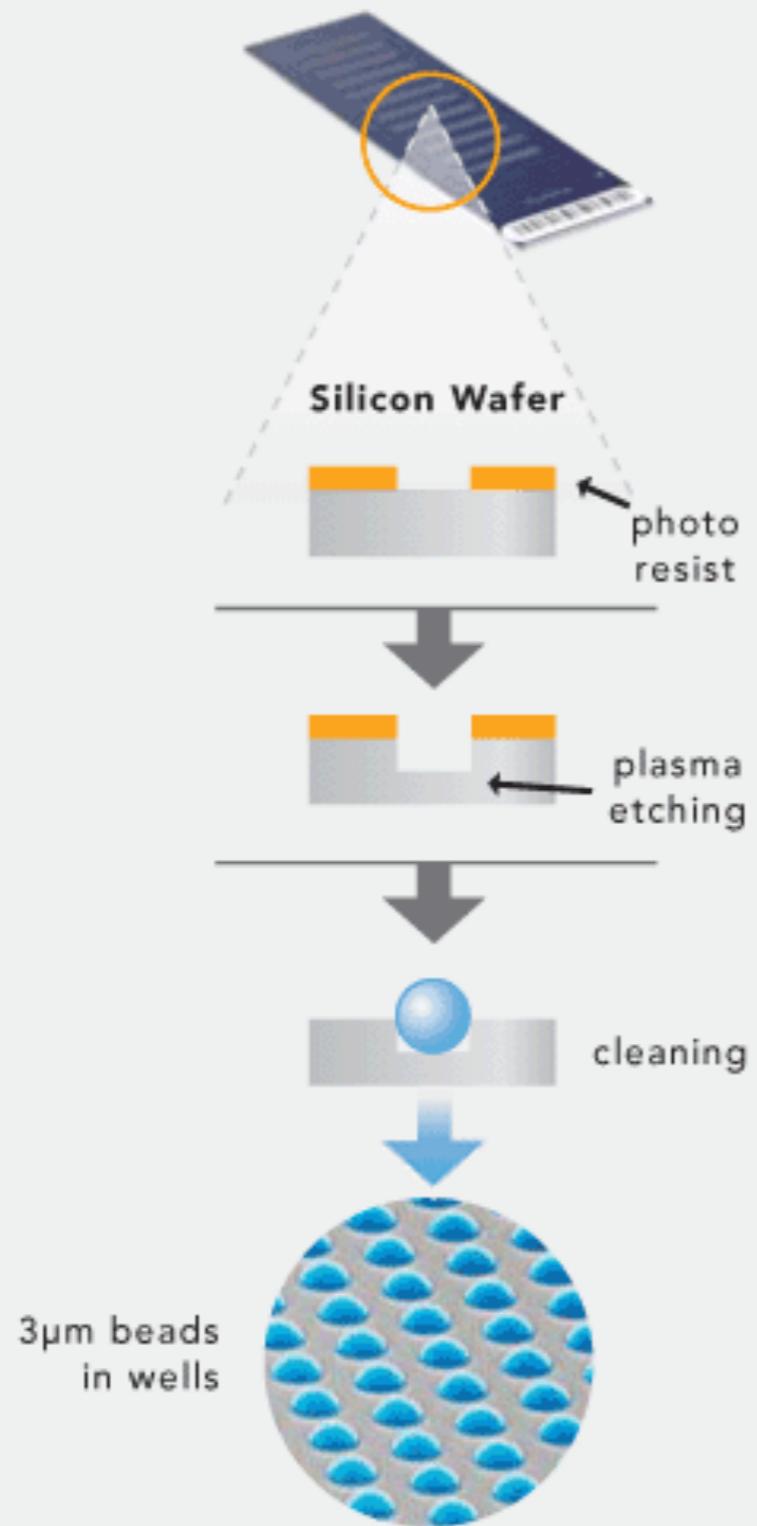
Expensive at scale

Typical use case

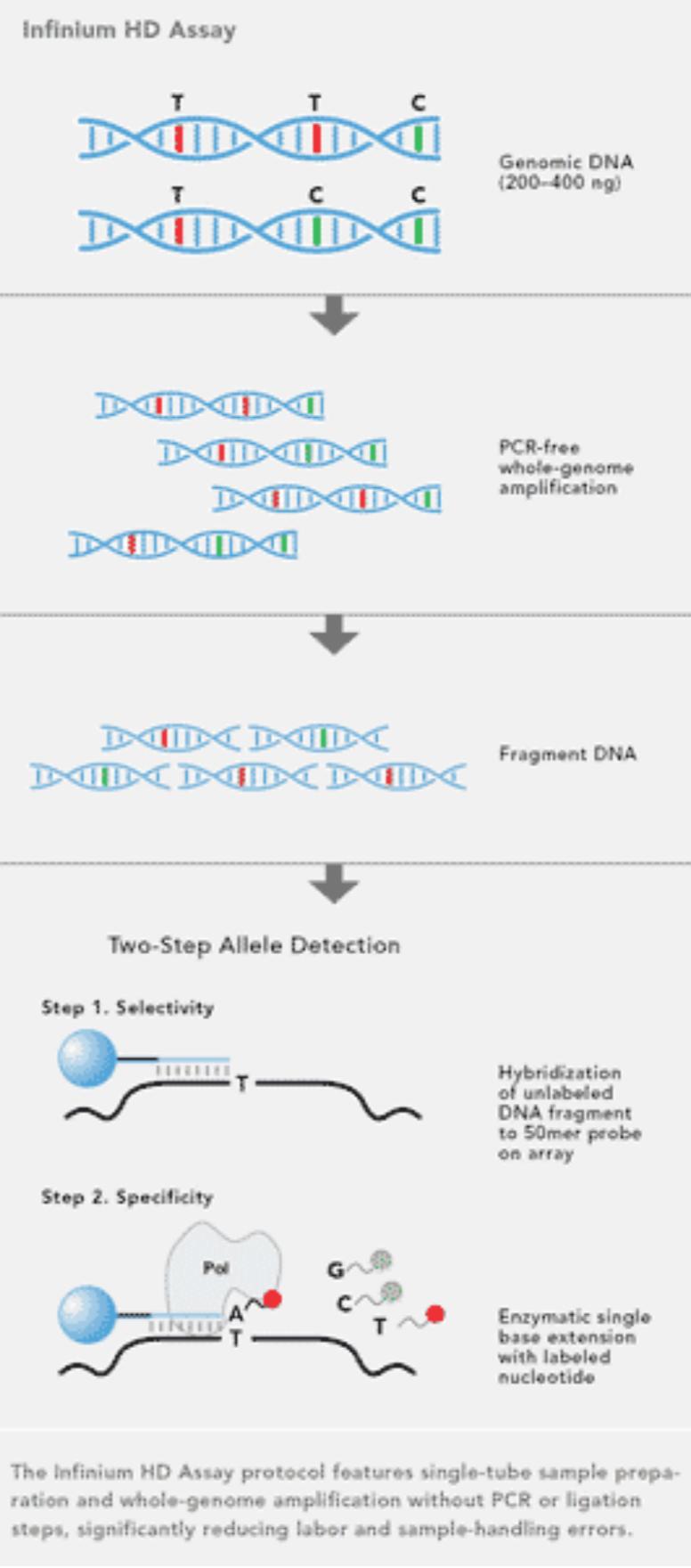
“Phylogenetic diversity” studies, i.e.
small number of genes from
many taxa

**Other
enrichment
tools. . .**

SNP chips



Illumina multi-sample array formats



23andMe - Genetic Testin... X

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

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)

**What's next for
sequencing?**

Space.

The simplest summary of genome composition is base composition.

If I tell you a genome is 46% GC and 3 billion bases long, since G=C and T=A:

The number of bases are:

G 690 million

C 690 million

A 810 million

T 810 million

This is a genome sequence, just a bad one where the size of each contig is 1bp long.

The number of bases are:

G 690 million

C 690 million

A 810 million

T 810 million

A sequencer is a highly specialized microscope that allows you to observe some features of the position of nucleotides in space.

fastq example:

```
@HWI-ST625:51:C02UNACXX:7:1101:1179:1962 1:N:0:TTAGGC
CTAGNTGTTGAAGAGAAGGTTCAAGAACCAAAAGAAAGCTCACAAACACATATGGT
+
=AAA#DFDDDHHFDGHEHIAFHIIIIIGICDGAGDHGGIHG@A@BFIFIHIIIGC@@8

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

Nucleic acids are 3D molecules that exist in space and time.

Sequencing is the molecular equivalent of anatomy - you kill the object of study and take it out of its original context

This is very useful, but doesn't give the full picture of its biology.

Physics has the Heisenburg uncertainty principle - you can't precisely know a particle's location and momentum.

Biology has the blender uncertainty principle- to get sequence data we need to grind things up, and this destroys important spatial information.

The Heisenburg uncertainty principle is a law of physics.

The blender uncertainty principle is a limitation of our current instruments.

To get the whole picture we need to improve our sequencing microscopes to show us both the spatial organization of nucleotides in the molecules and something about their spatial context.

**One of the most common
RNAseq applications is to
identify spatial differences in
gene expression**

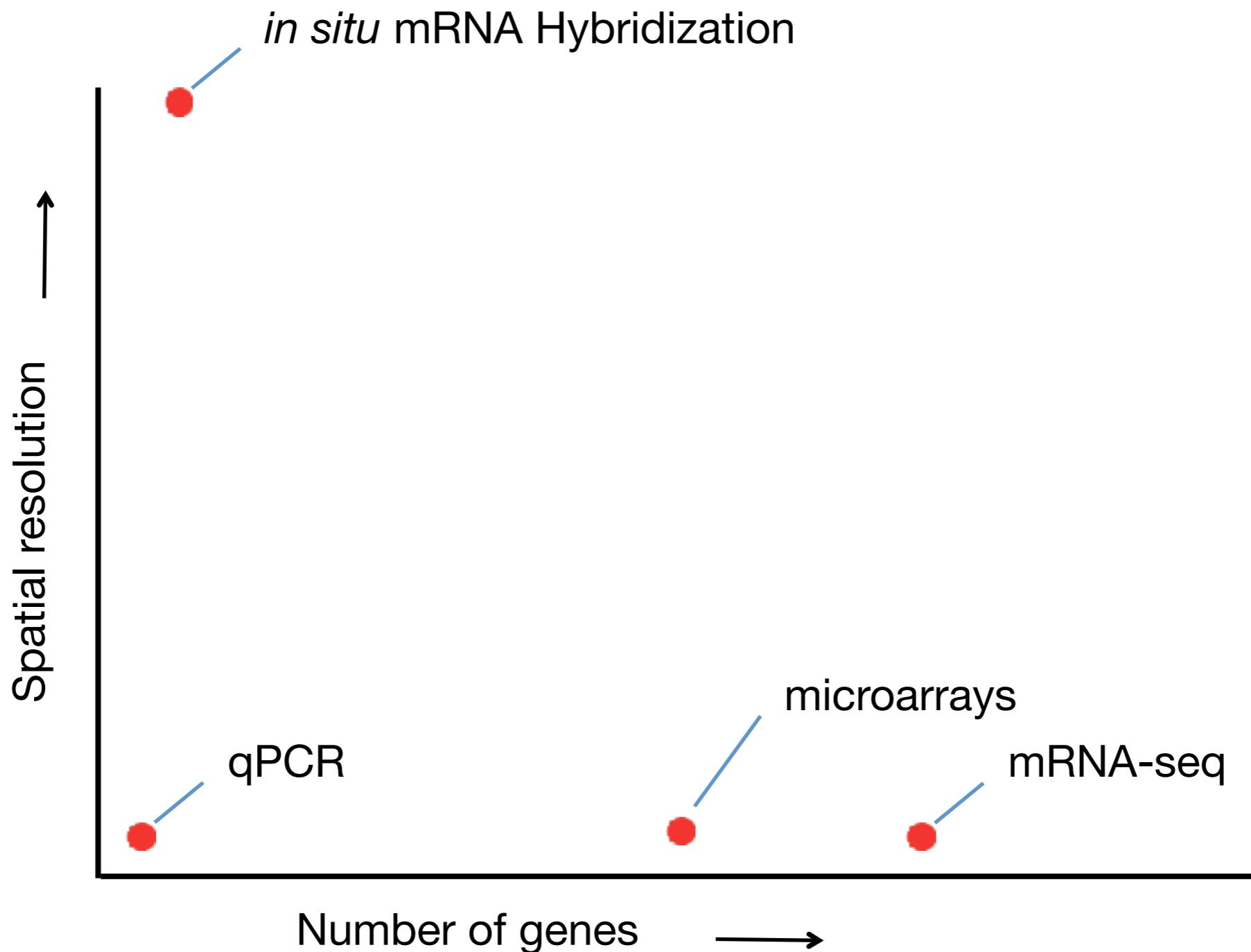
Current approach

1. Isolate cells, tissues, or organs
2. Prep and sequence samples
3. Identify genes with differential expression
4. Characterize these genes, e.g.
in situ hybridization

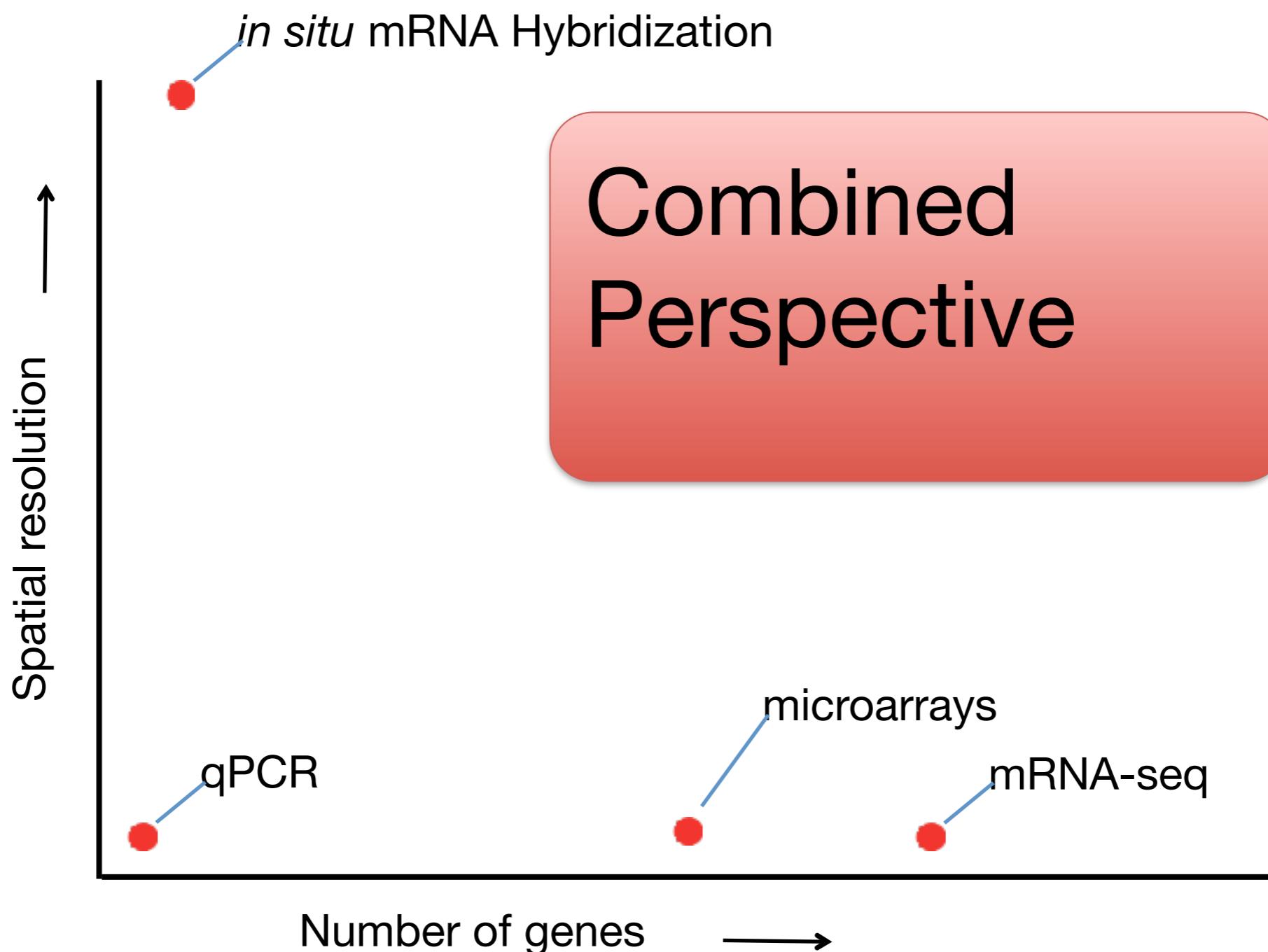
People do it, but it hurts:

- Many steps
- Sample isolation low resolution, difficult, and requires prior knowledge of relevant regions
- Characterization requires many specimens and many repeated tasks

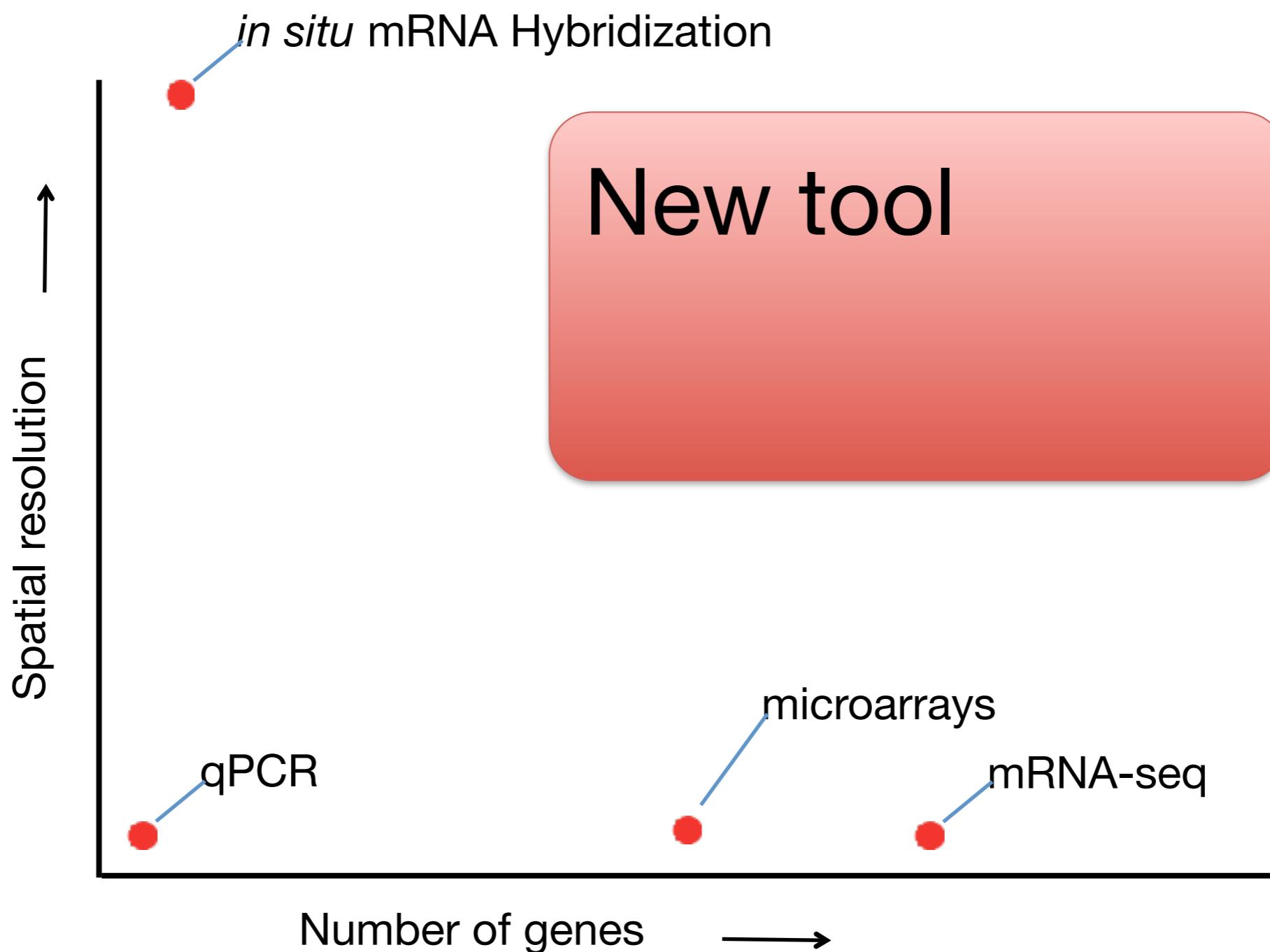
Existing tools have different strengths



Existing tools have different strengths



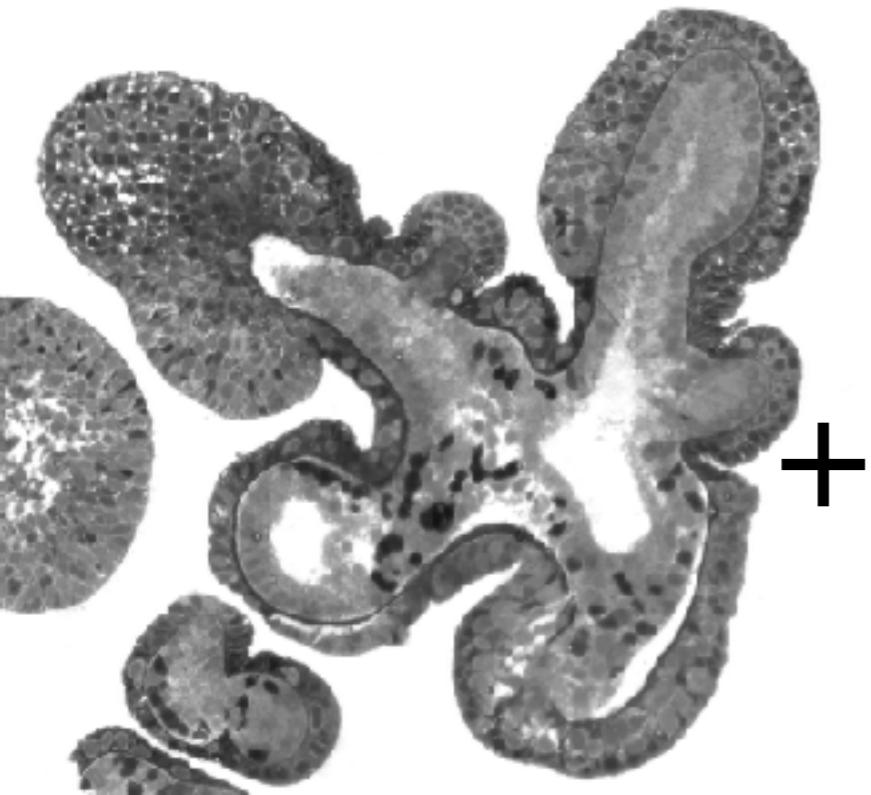
We need a new tool that does it all



This new tool is a gene expression camera

Measures the expression of all genes at high resolution in a 2-dimensional tissue slice, and superimposes the data on anatomical images of the same tissue.

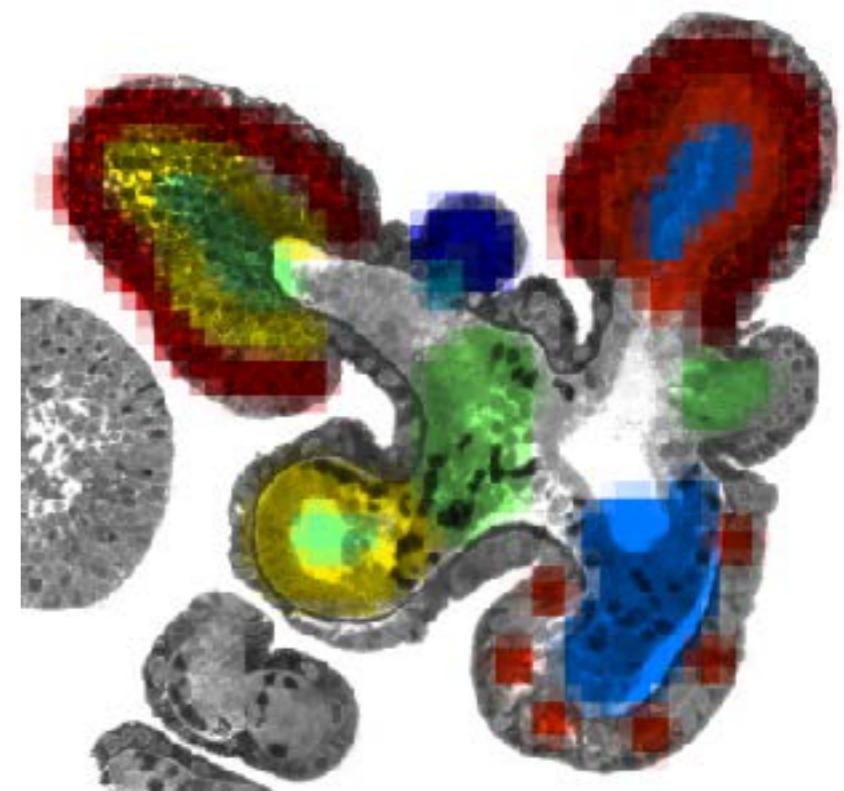
Anatomical image



Gene expression camera

[each pixel has data on
the abundance of mRNA
for each gene]

Integrated view of gene
expression in space



=

Gene 1...

Expression Level

10 1000

The investigator can:

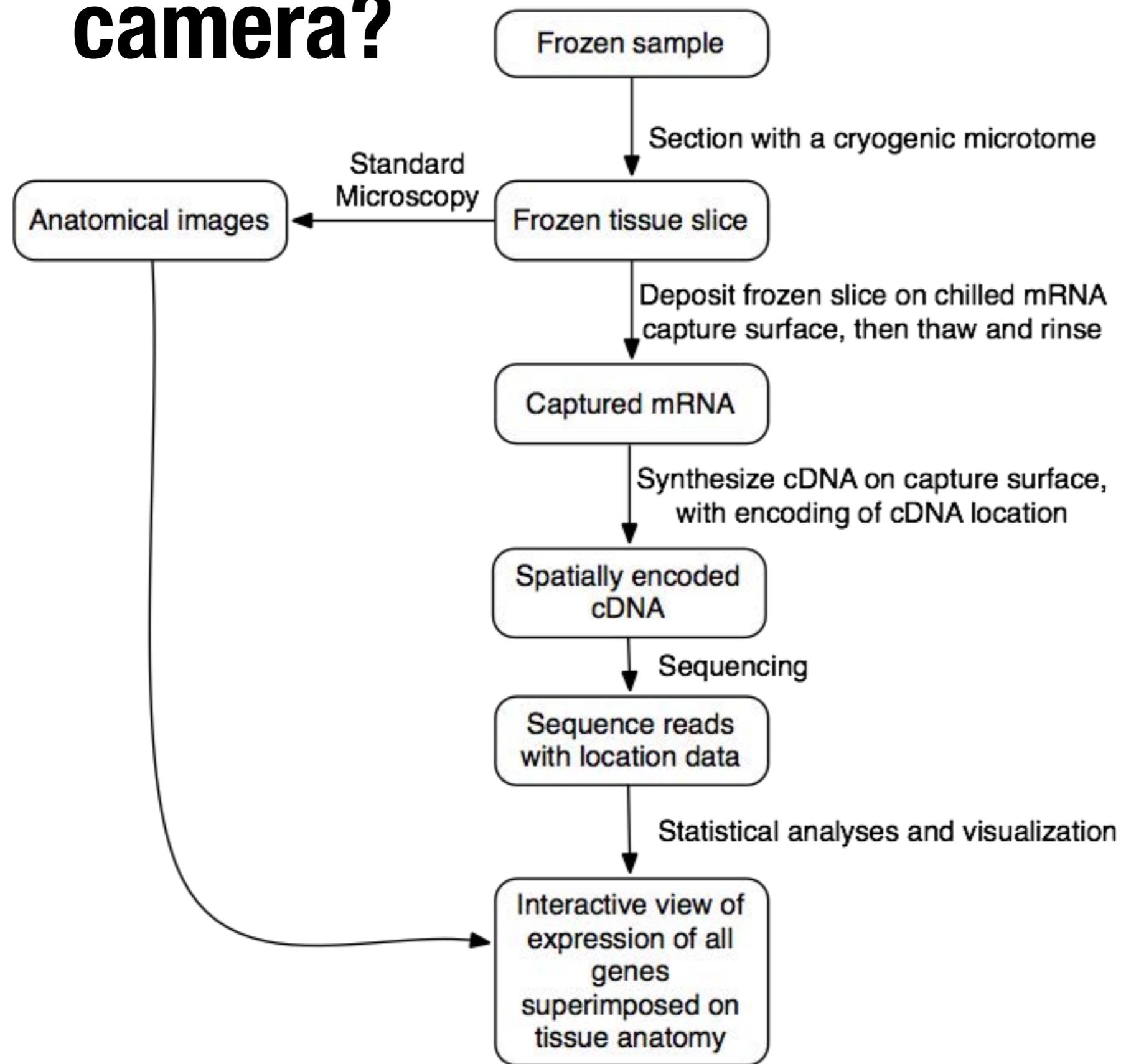
- Identify which genes have differential expression in regions of particular interest
- Look at known genes of interest
- Identify sets of genes that have spatial covariance in expression
- Identify anatomical regions with particularly unique expression

How to do it?

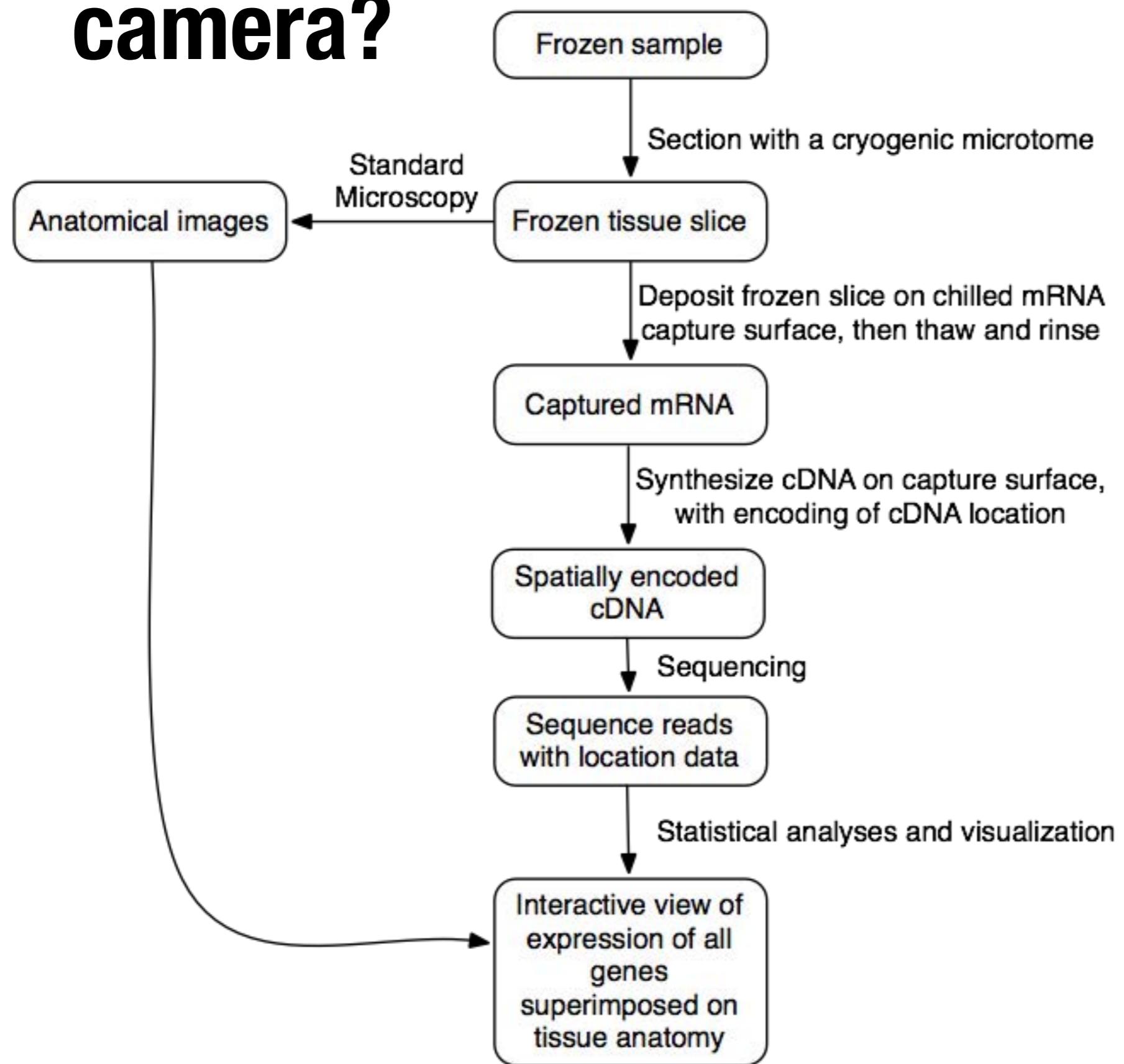
Camera is a sample prep tool

Capture surface is a
patterned microarray

How to build a gene expression camera?



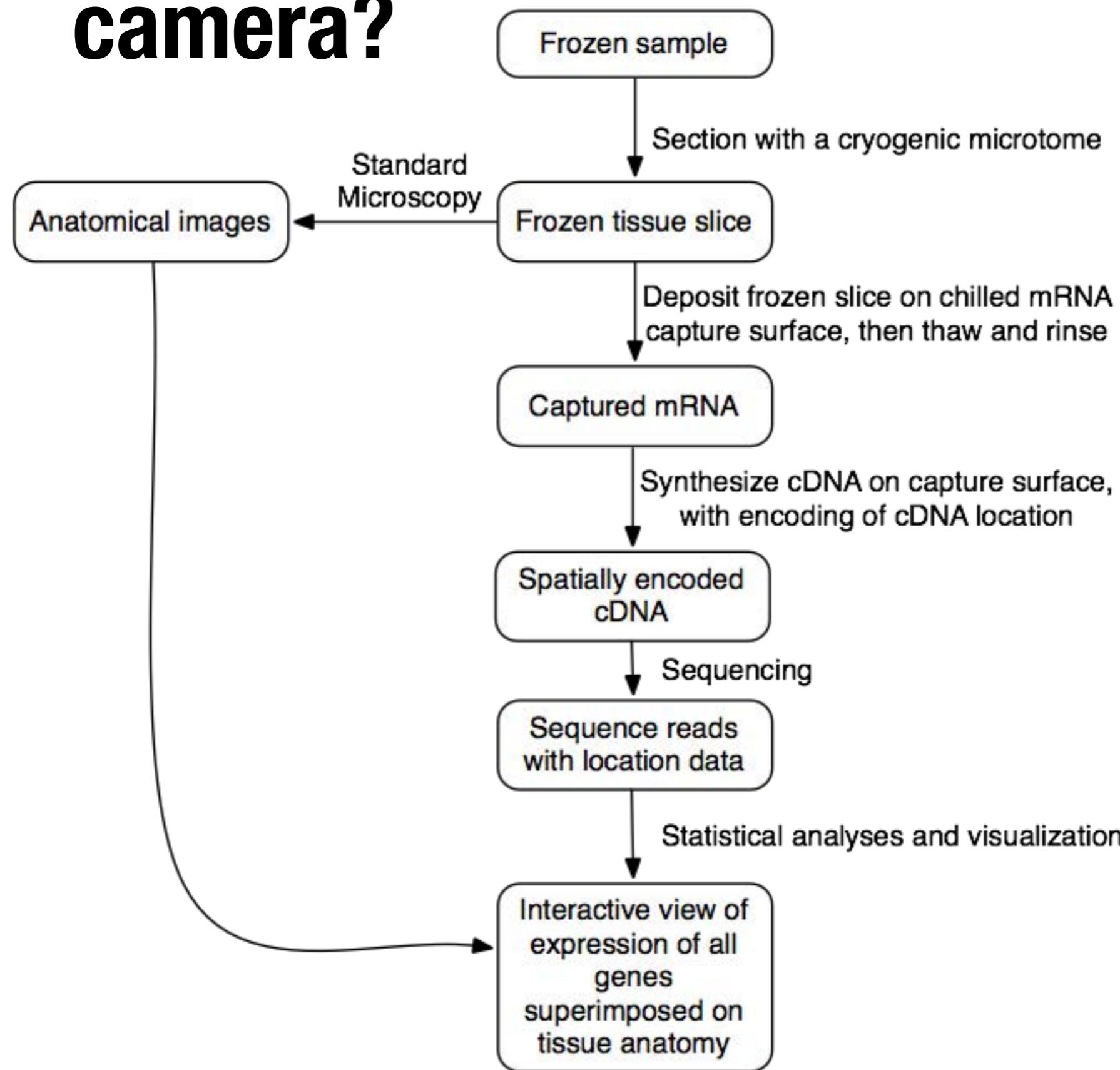
How to build a gene expression camera?



1. Sample prep

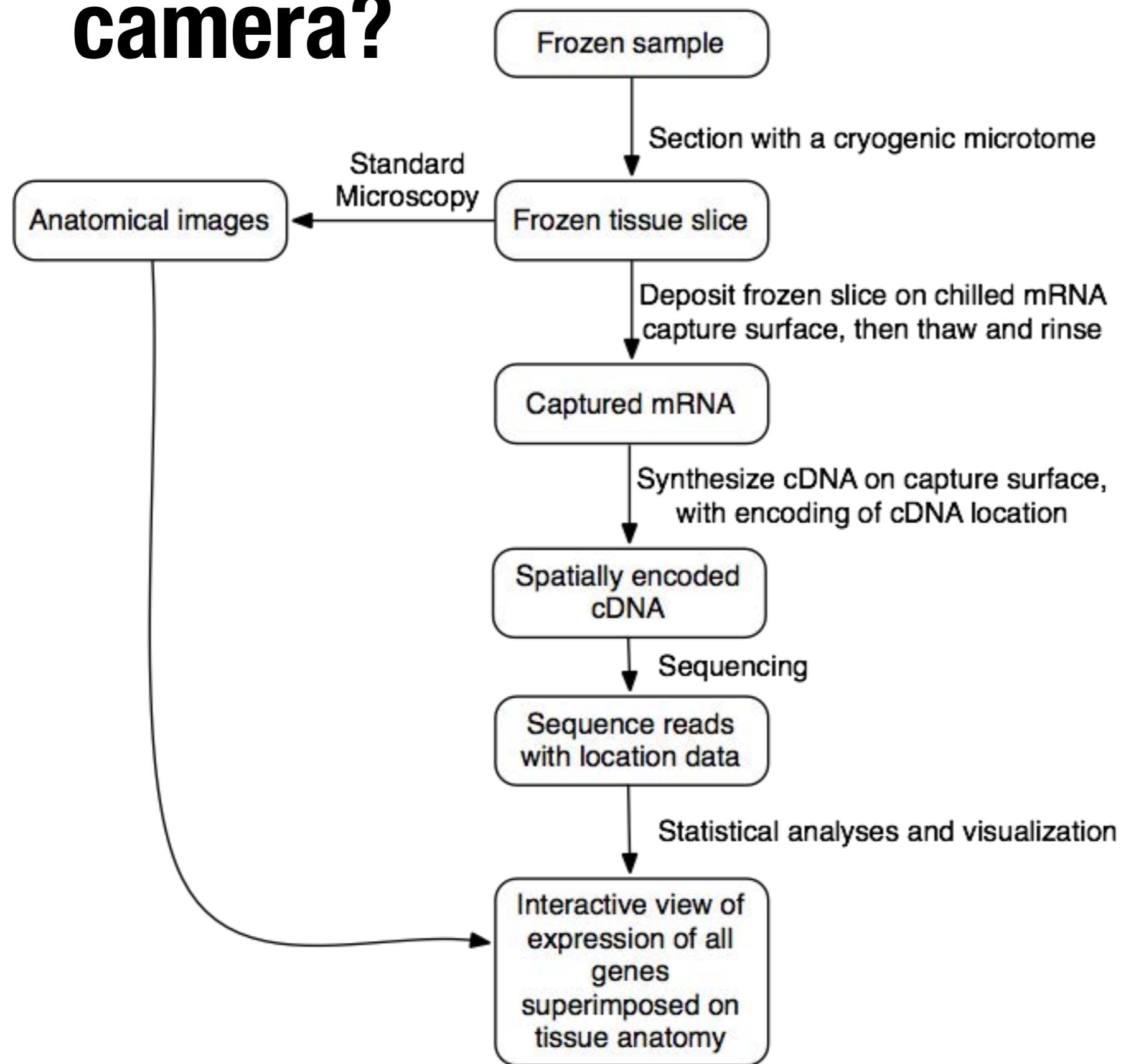
Tissue is frozen and cryosliced to expose mRNA while preserving spatial location

How to build a gene expression camera?

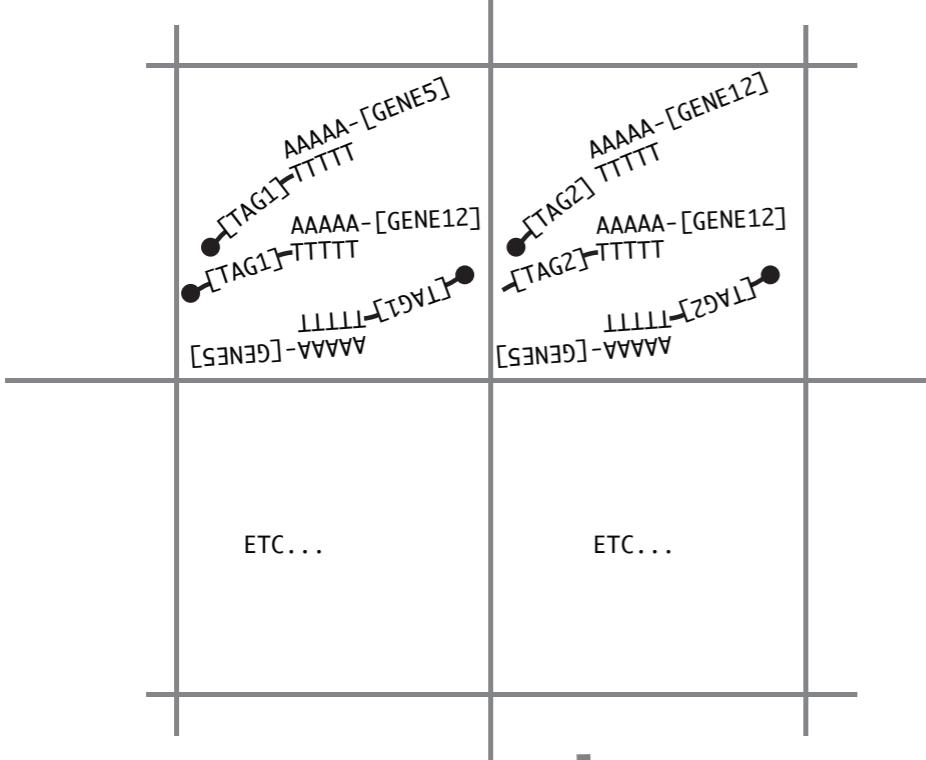
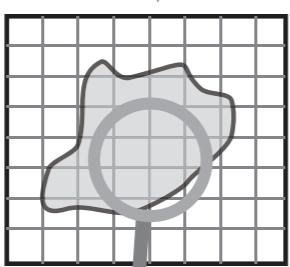
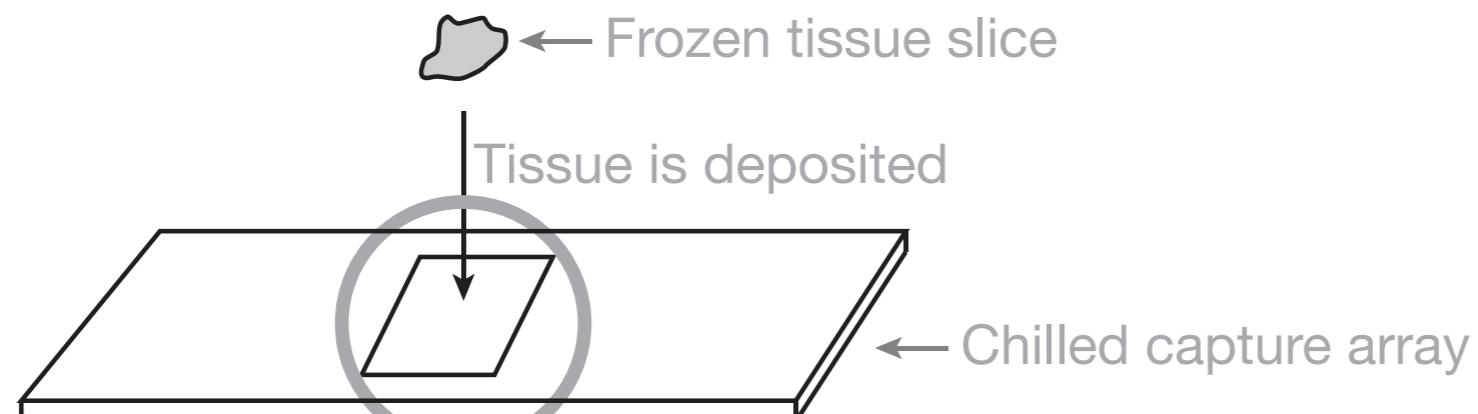


2. Expression Imaging
mRNA is bound to a 2D surface, converted to cDNA, and sequenced in such a way that the location where the mRNA bound is recorded

How to build a gene expression camera?



3. Analysis
Integration of data between genes and anatomy, image generation



mRNA (which have a poly-A tail) bind the poly-T tails of the bound capture adapters

The capture adapter of each adapter has a tag sequence that is specific to each region of the microarray ("pixel")

[Home](#)

[Quick](#)

[Advanced](#)

[Pat Num](#)

[Help](#)

[Bottom](#)

[View Cart](#)

[Add to Cart](#)

[Images](#)

(1 of 1

United States Patent

9,330,295

Dunn

May 3, 2016

Spatial sequencing/gene expression camera

Abstract

Methods, articles and systems that provide imagewise mapping or display of gene expression of a biosample, by contacting the biosample, such as a tissue slice or metacommunity, to a detector which captures material from the biosample and processes the captured material. In one embodiment the detector has an array of one or more capture sites at defined positions on the detector, each site carrying an immobilized capture oligonucleotide and a site-indexing oligonucleotide. The array captures mRNA from the biosample contacted thereto, and the captured mRNA is processed to form a sequenceable amount of amplified captured material which includes the site-indexing oligonucleotide, so that when sequenced, detection of the site-indexing oligonucleotide indicates the original capture location on the array, thereby mapping the sequenced material to its capture location and imaging display of gene expression distribution in the original biosample. In some embodiments the site-encoding sequence is integrated with the capture oligonucleotide. In other embodiments, the detector is a modified sequencing flow cell, which is opened to allow the biospecimen to be contacted to a capture surface; processing is performed while the material remains on the capture surface and locations of the resulting sequences correspond to the location of origin of the templates of the biomolecules in the sample. The spatially resolved sequencing, gene expression camera and technology in various embodiments are applied to genome sequences and DNA fragments present in the biosample, for example to study or diagnose developmental, disease, and tumor conditions.

Inventors: Dunn; Casey (Providence, RI)

Applicant: Name City State Country Type

BROWN UNIVERSITY Providence RI US

Assignee: BROWN UNIVERSITY (Providence, RI)

Family ID: 1000001817265

Appl. No.: 14/215,748

Filed: March 17, 2014

After we started this
project...

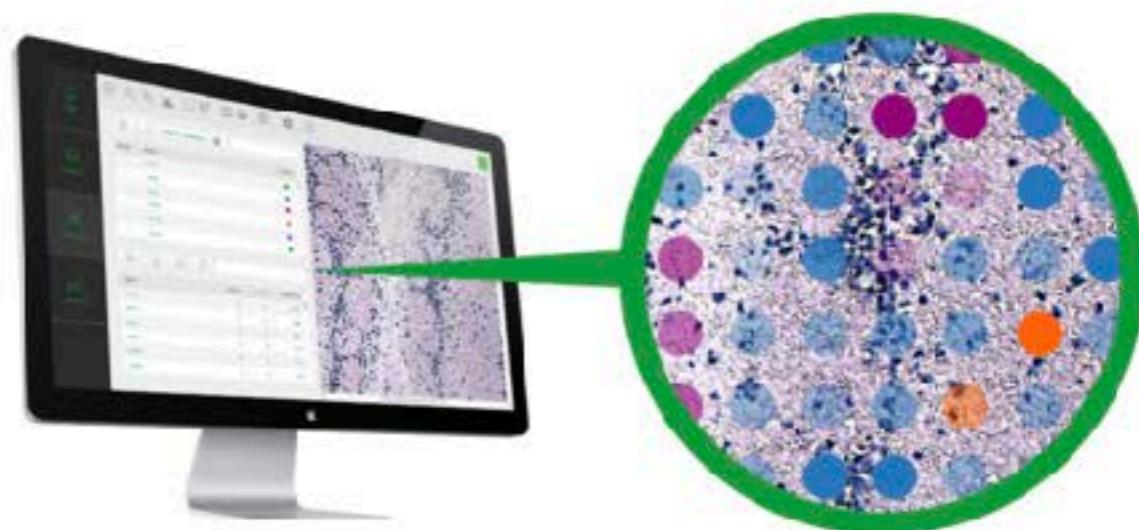


Visualize and quantify gene expression in the histological context

A high resolution image of your tissue section gives an overview of the morphology. At the same time quantitative gene expression data can be presented directly in the image providing the spatial context. This unites two different sources of information opening up a new field of research: Spatial Transcriptomics.

Compare gene expression patterns based on morphology

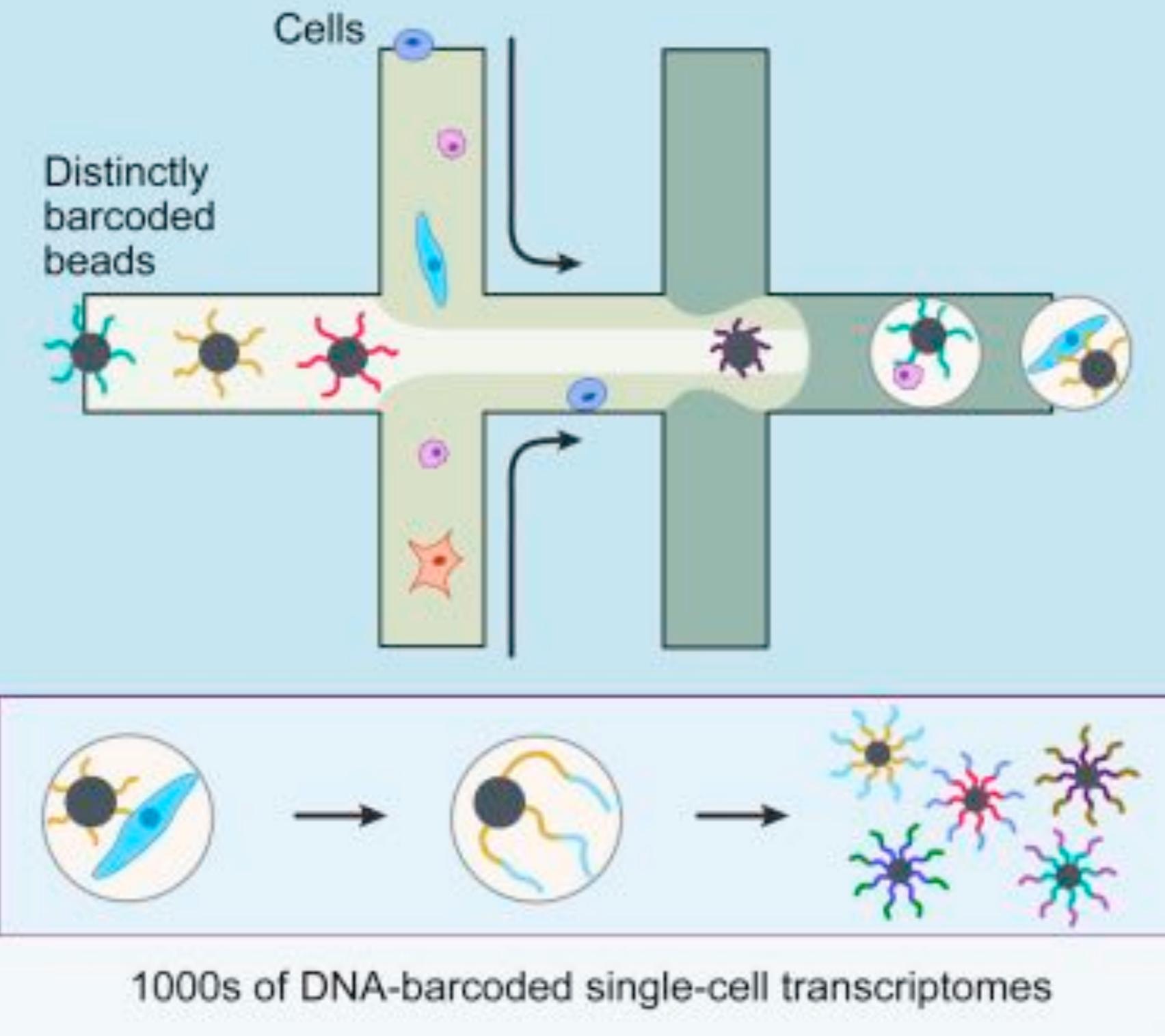
Select different areas of interest in your tissue section image and look at their different expression patterns in detail. Additionally the data allows for a variety of innovative analyses.



Another approach:

Single cell transcriptome sequencing provides information on which transcripts are in same cells (but not where the cells are).

Drop-seq single cell analysis



Macosco et al. 2015

<http://dx.doi.org/10.1016/j.cell.2015.05.002>

Efficient and Scalable Cell Capture



Complete workflow in one day

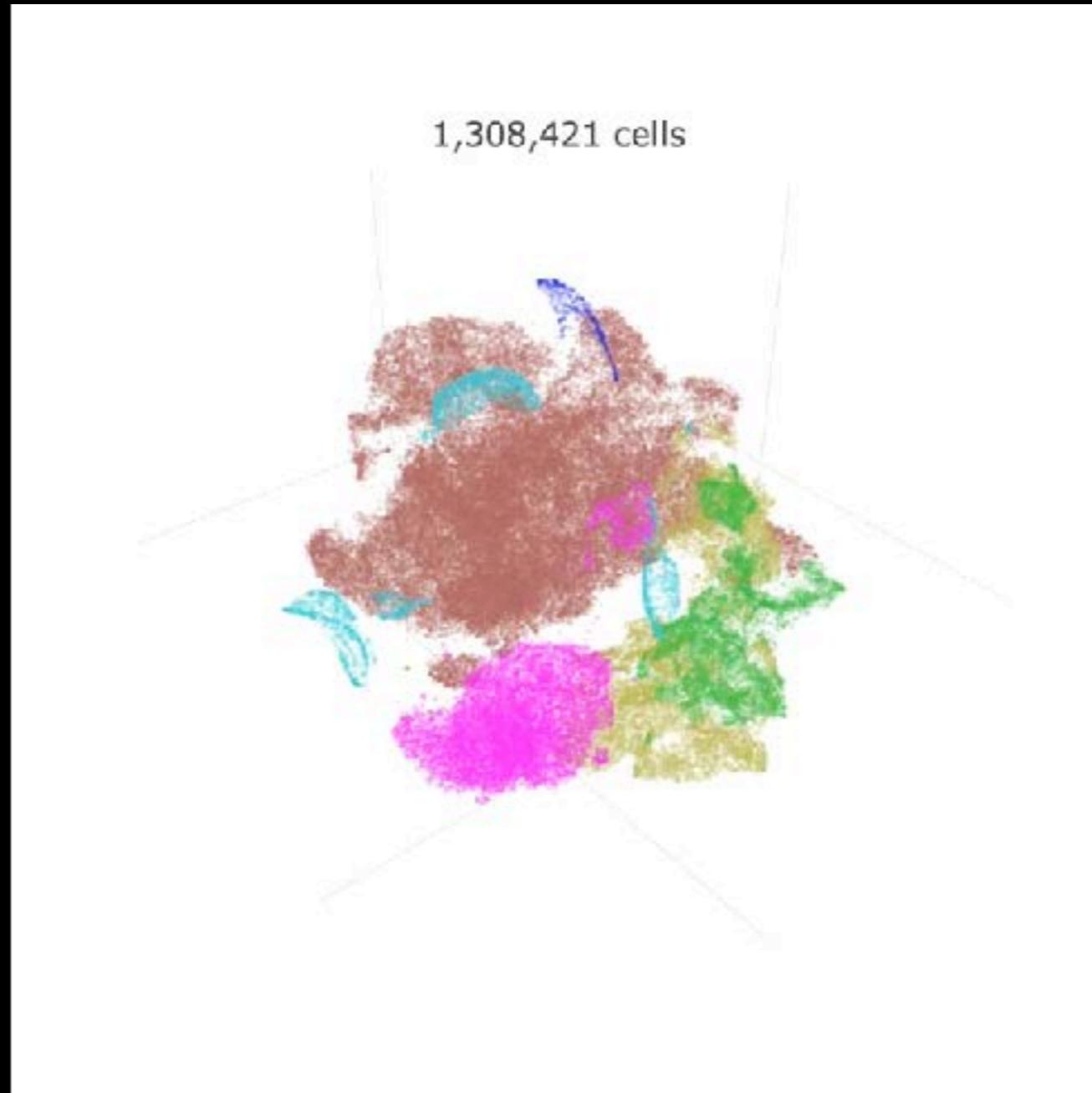
Throughput flexibility

Cell capture efficiency up to 65%

Compatible with Illumina® HiSeq®
4000/2500/NextSeq®/MiSeq® sequencers

Mouse brain expression

Can plot each cell by expression similarity



<https://www.10xgenomics.com/single-cell/>

<http://fast.wistia.net/embed/iframe/z54e2lemhd>

What about spatial genomics?

Mouse chromosome packing

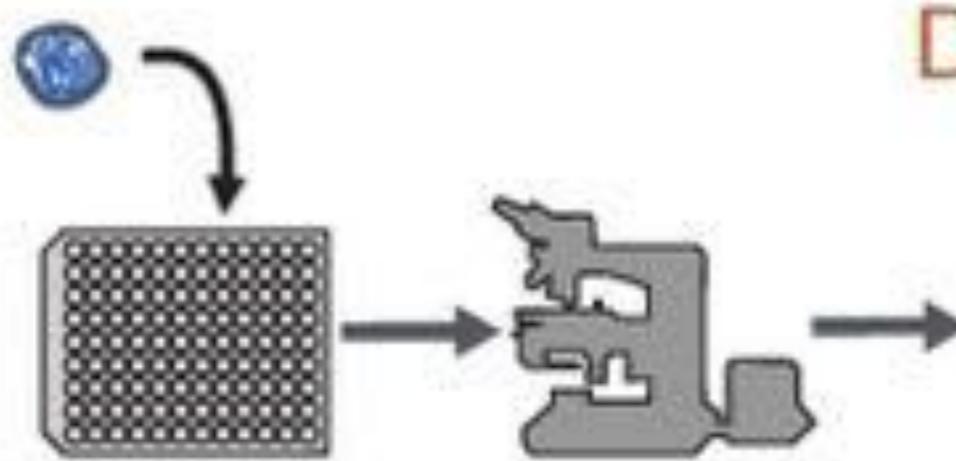


Stevens et al. 2017

http://www.nature.com/nature/journal/v544/n7648/fig_tab/nature21429_SV1.html

a

FACS Imaging

In-nucleus Hi-C
Biotin

Digestion end-fill Ligation



Identify contacts

Cut and
purify

Add

adaptors

Compute structure

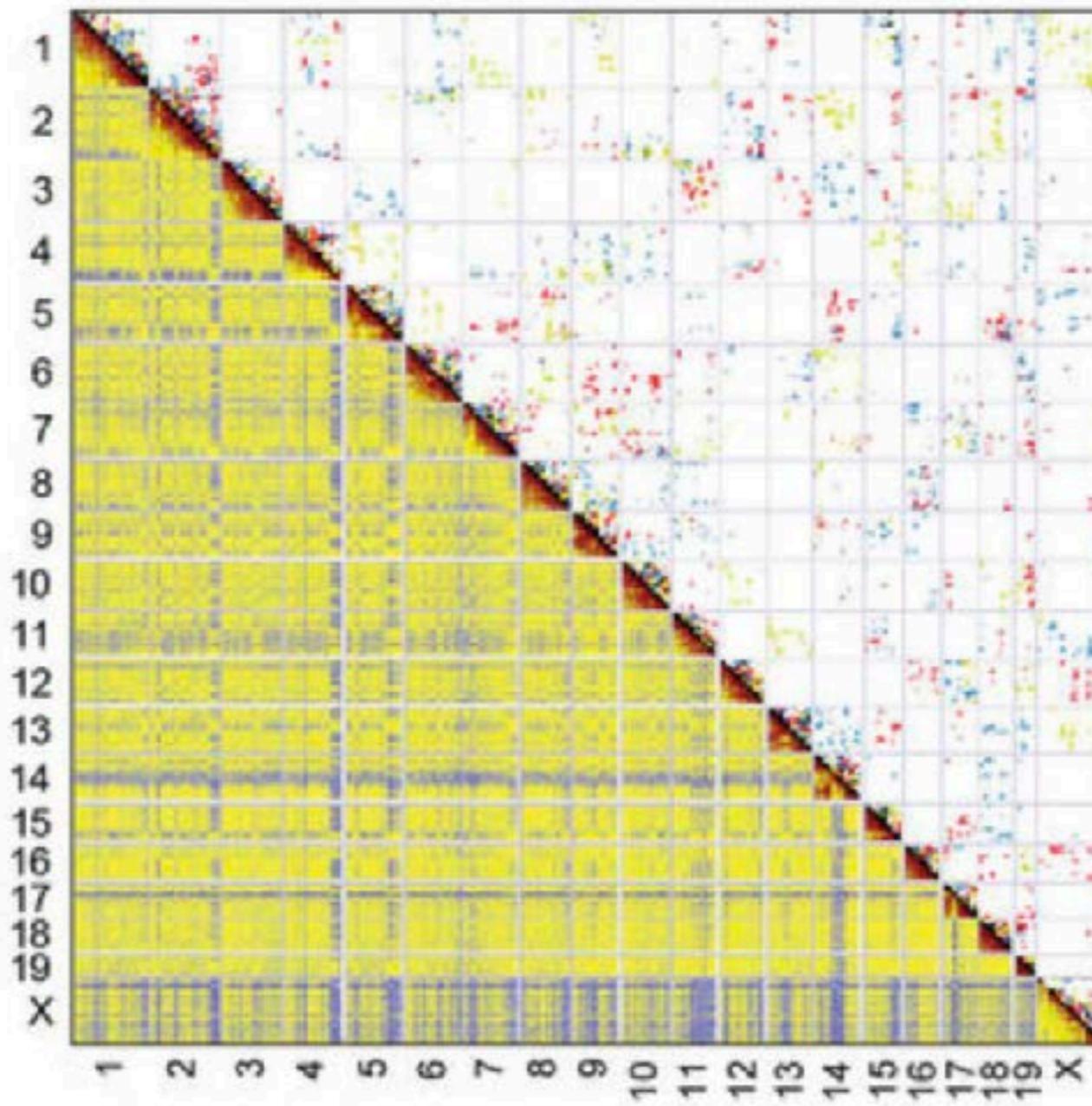
Amplify and
sequence

Map ends

to genome

Use contacts
as restraints

Mouse chromosome contact map



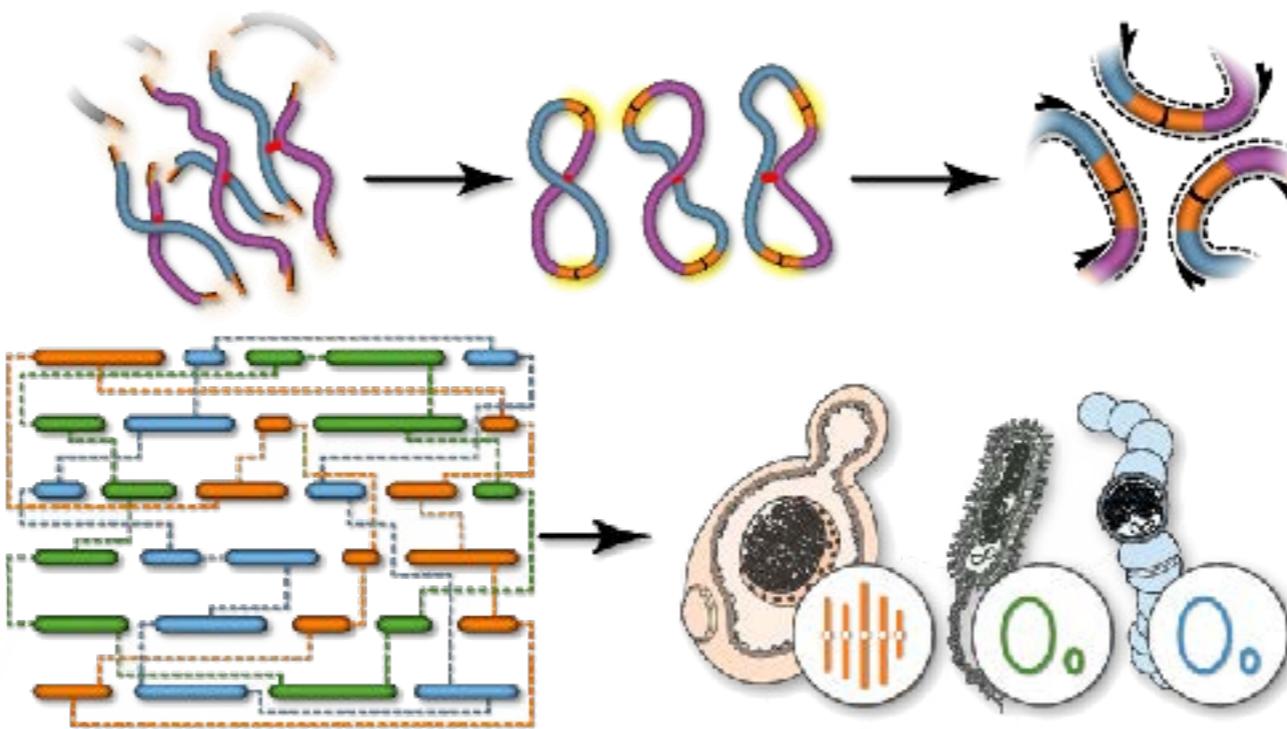
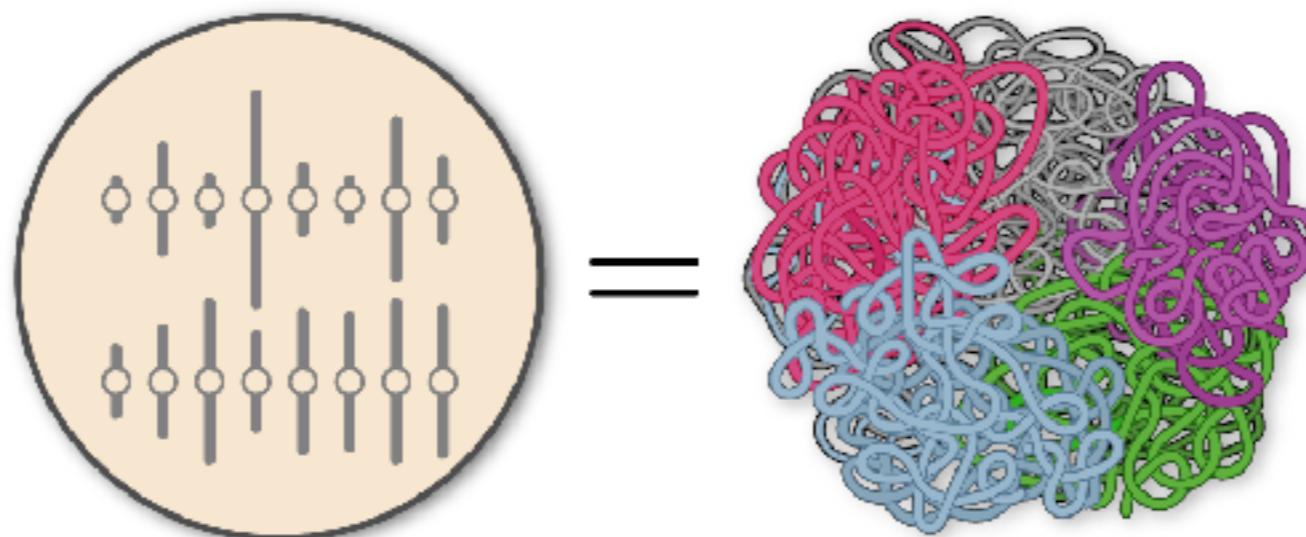
Stevens et al 2017

http://www.nature.com/nature/journal/v544/n7648/fig_tab/nature21429_SF1.html



PHASE GENOMICS

"GET CHROMOSOME-SCALE SCAFFOLDS FOR VIRTUALLY ANY GENOME"



New and improved

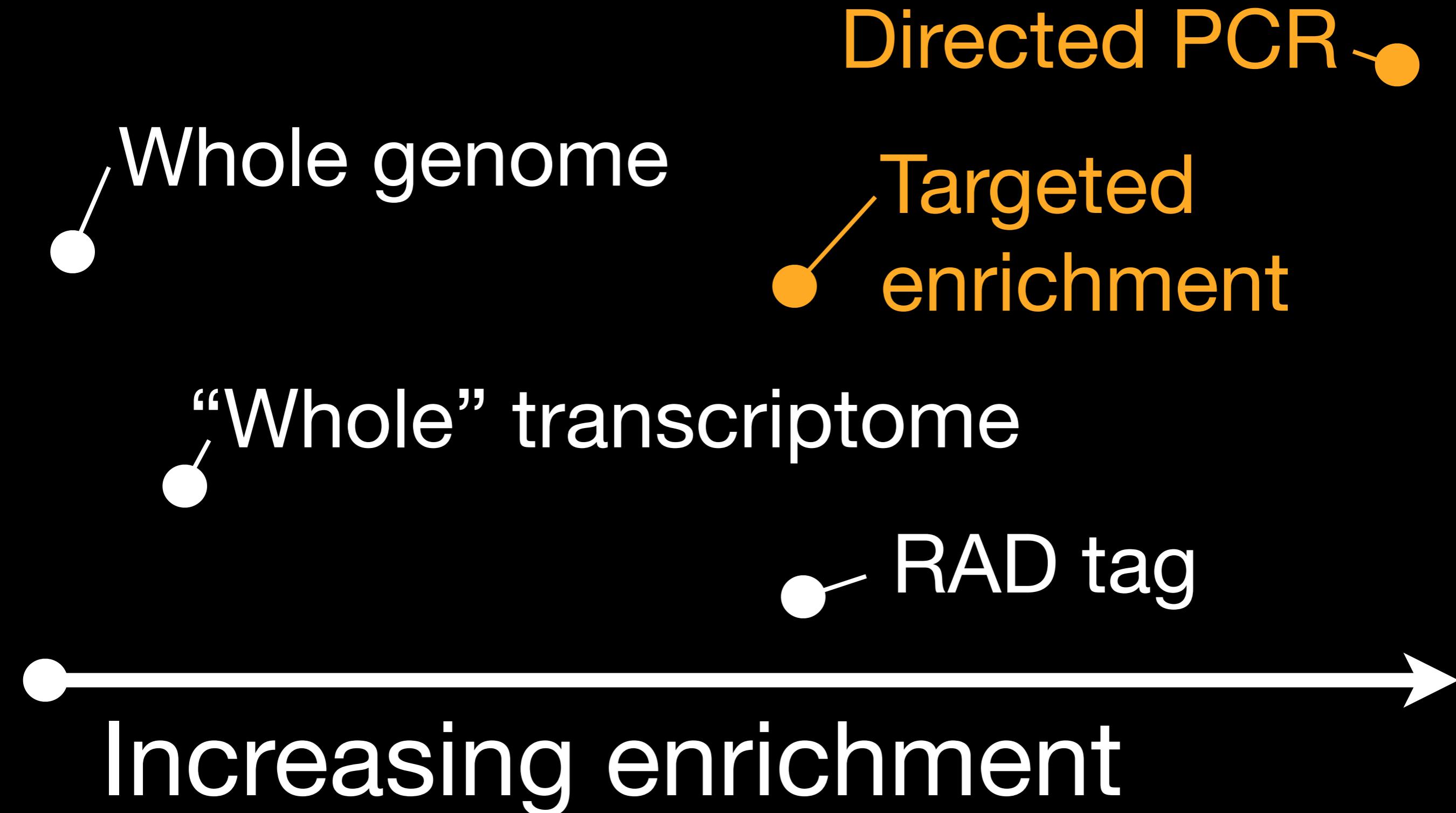
Genomes are assembled from DNA sequences stitched into pieces known as contigs and scaffolds. Better genomes have fewer but longer contigs and scaffolds. (N50 represents a size measurement similar to median length.) Newer genomes also have fewer gaps.

YEAR	HUMMINGBIRD (<i>CALypte anna</i>)		GOAT (<i>Capra hircus</i>)		MAIZE (<i>Zea mays</i>)	
	2014	2017	2012	2017	2010	2017
Contig number	124,820	5971	337,495	30,399	125,200	2789
N50 contig size (bases)	26,738	2,049,416	18,934	26,244,591	40,001	1,279,870
Scaffold number	54,736	N/A	77,432	29,907	2685	598
N50 scaffold size (bases)	4,052,191	1,067,027,607	14,391,519	87,277,232	1,260,849	10,525,104
Number of gaps	70,084	1300	260,474	492	125,191	2522

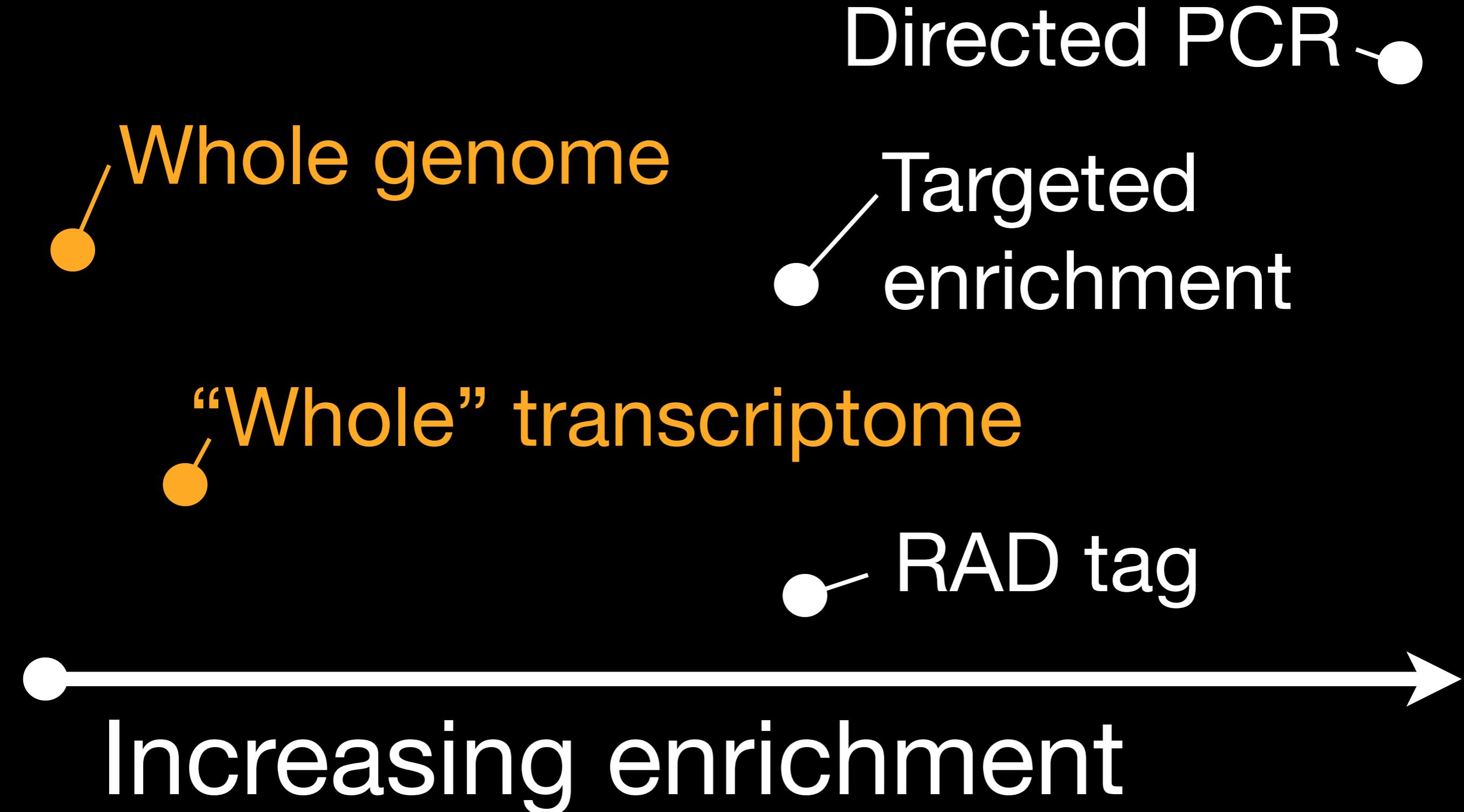
Part 2: Homology evaluation

A major conceptual difference between enrichment methods is whether genes are selected before or after sequencing

Select genes before sequencing



Select genes after sequencing



Before

After

Select genes

Amplify and
sequence
selected genes

Assemble matrix
from all
sequenced genes

Phylogenetic
inference

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

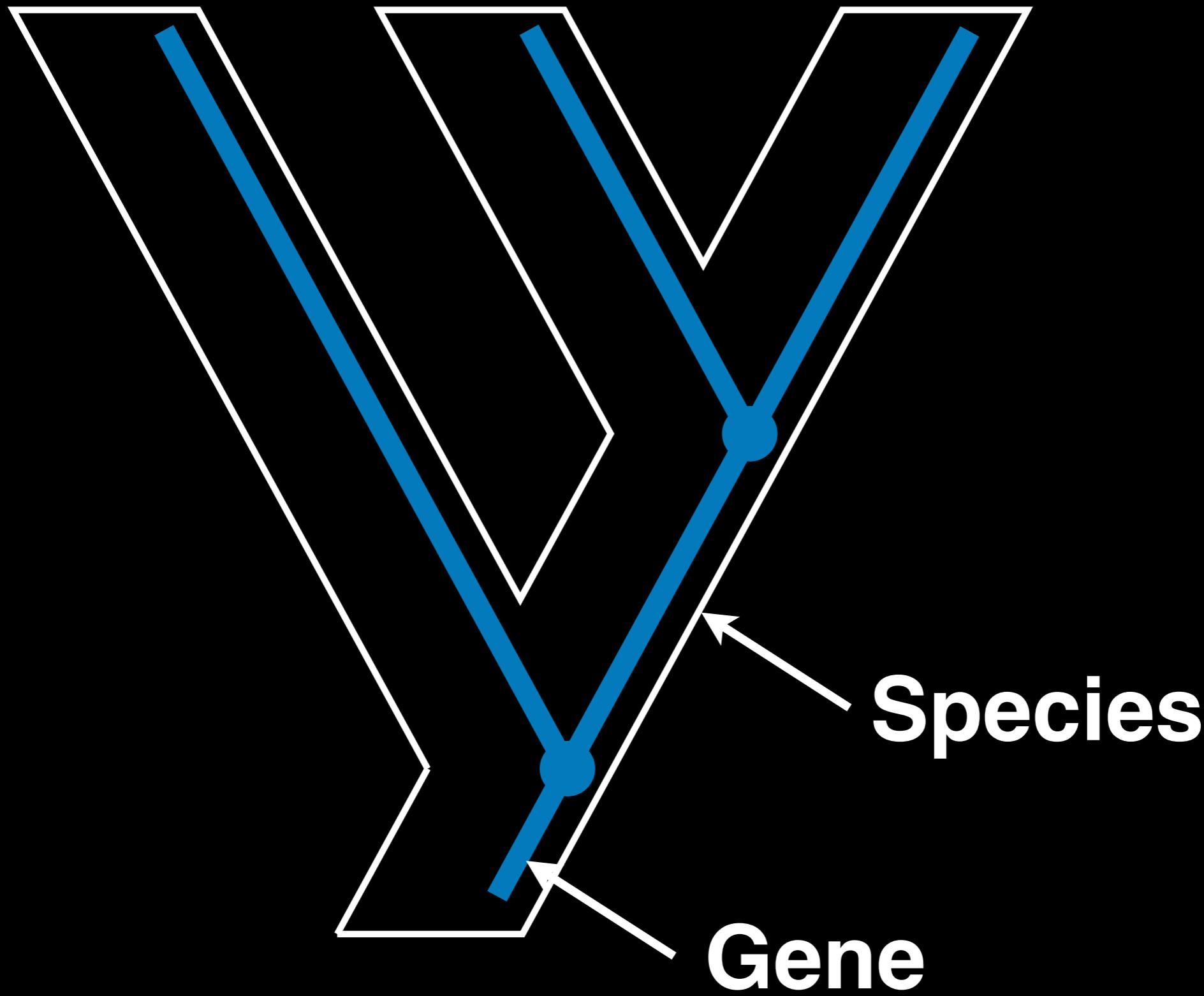
Species A



Species B



Species C



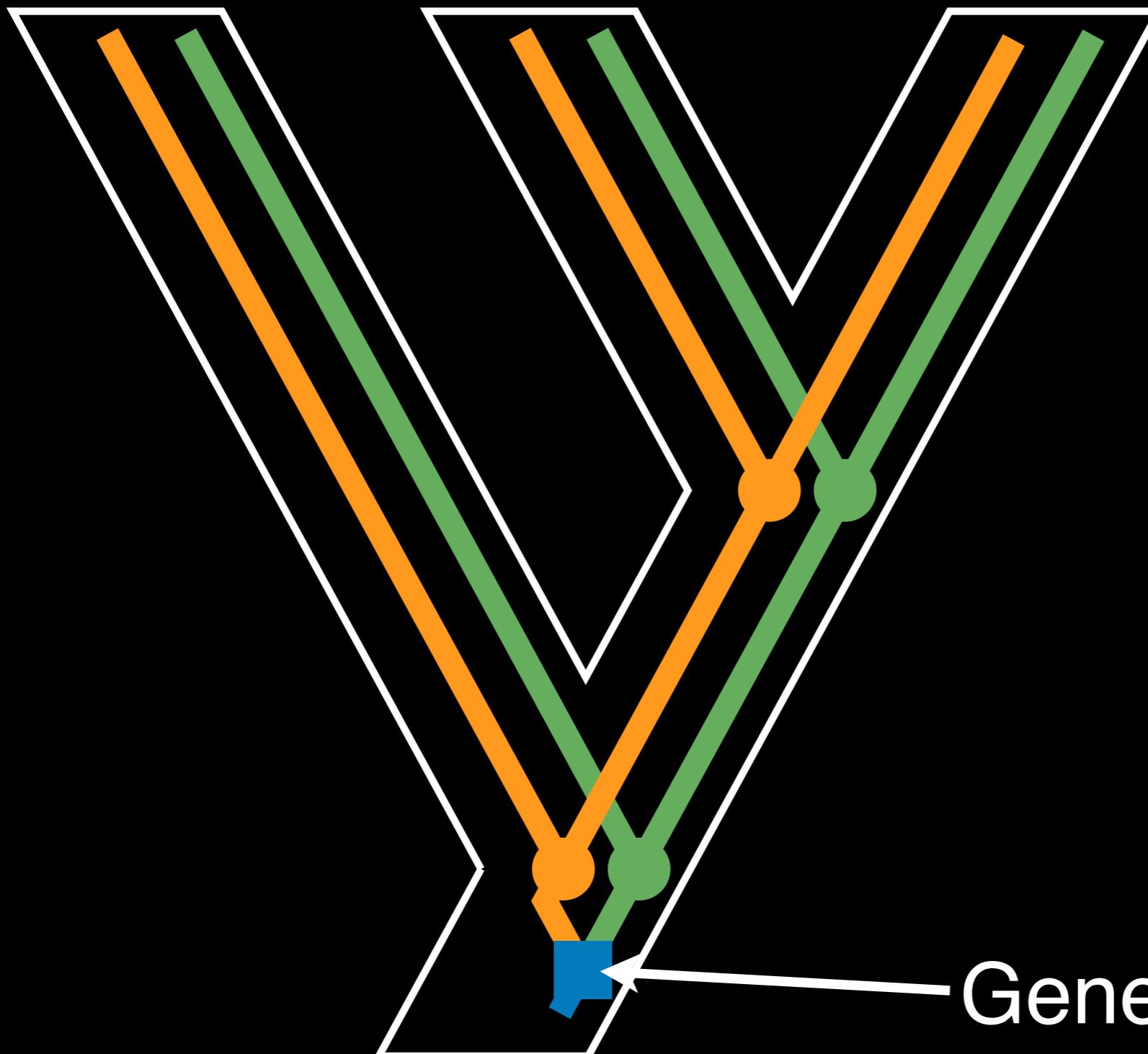
Species A



Species B



Species C



Gene divergence
due to duplication

Clearest
Orthology

Clearest
Homology

Available Data

Most
Informative

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](https://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](https://doi.org/10.1371/journal.pone.0000383)

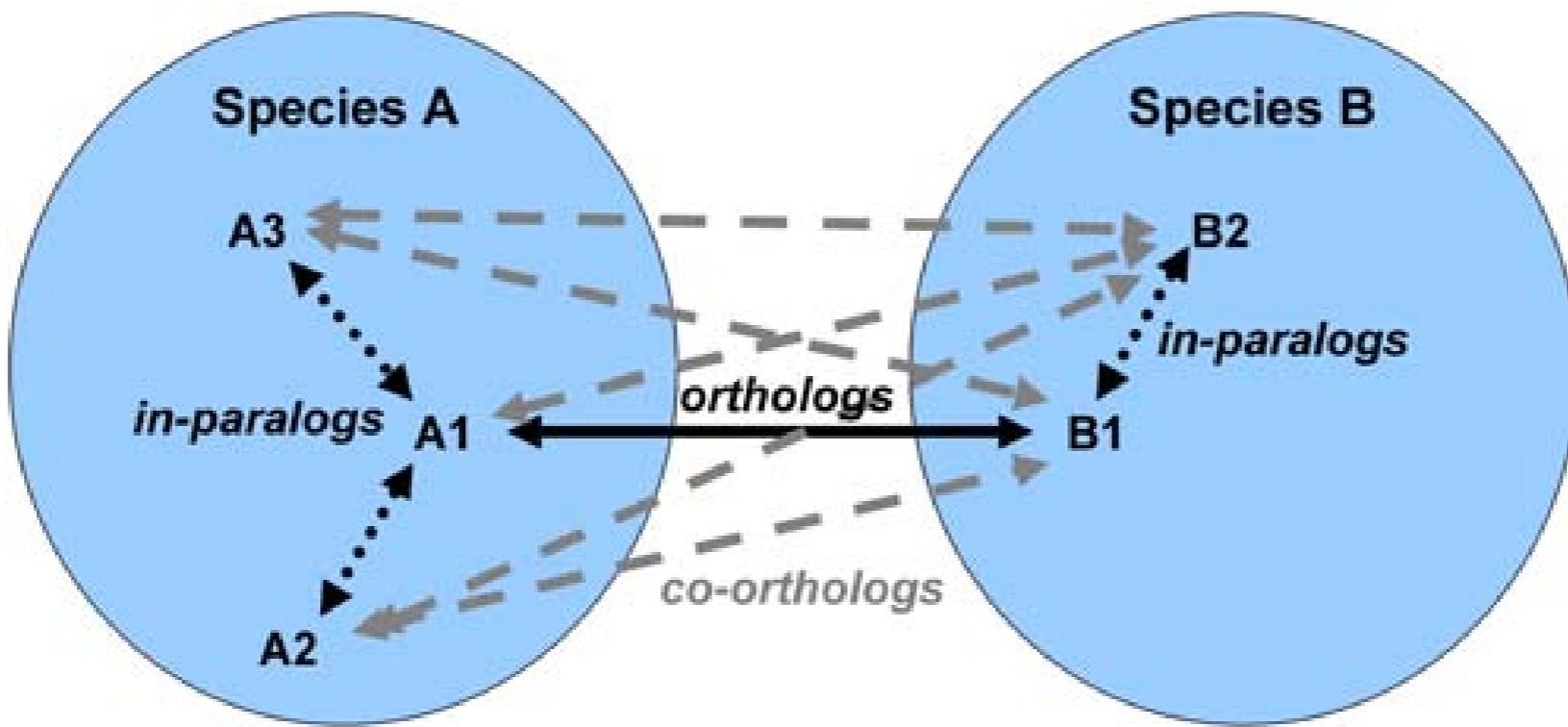


Figure 1. OrthoMCL graph construction between two species, including the establishment of co-ortholog relationships. Solid lines connecting A1 and B1 represent putative ortholog relationships identified by the 'reciprocal best hit' (RBH) rule. Dotted lines (e.g. those connecting A1 with A2 and A3, or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

This is our approach...

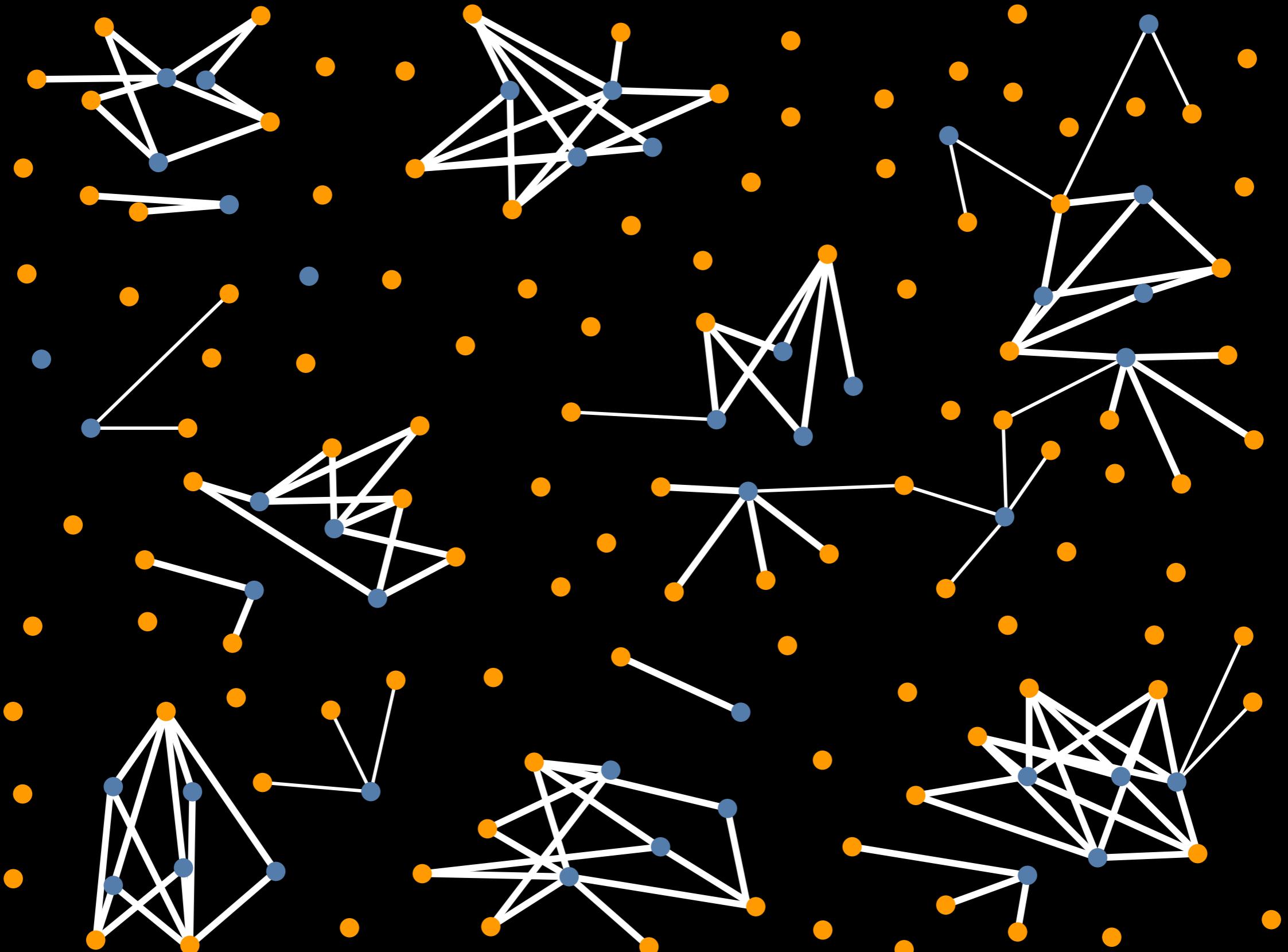
Put all sequences for all taxa in a study into a hat

Make all pairwise sequence comparisons

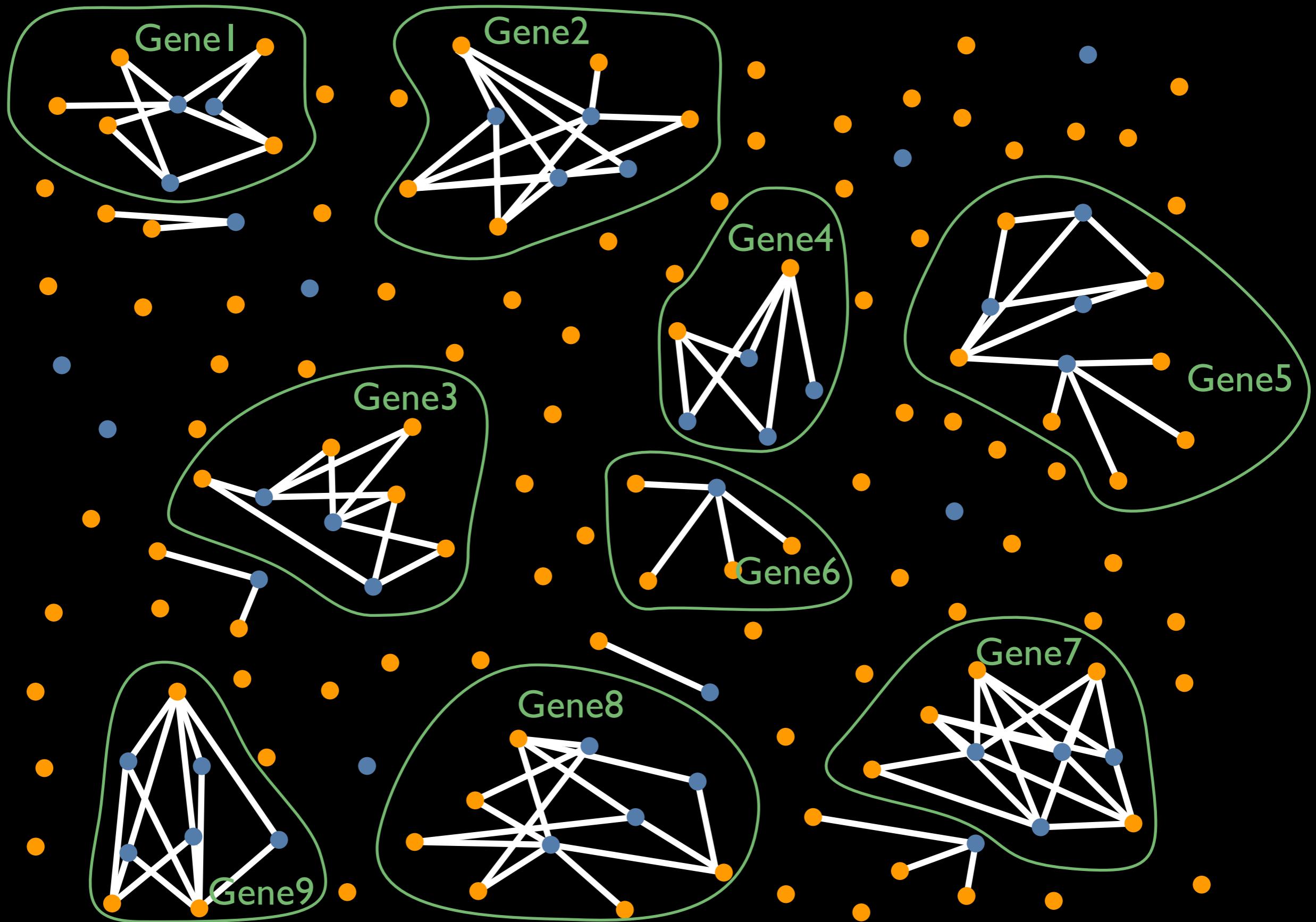
Construct a graph where nodes are sequences and edges indicate similarity



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 miscribing
paralogs as orthologs

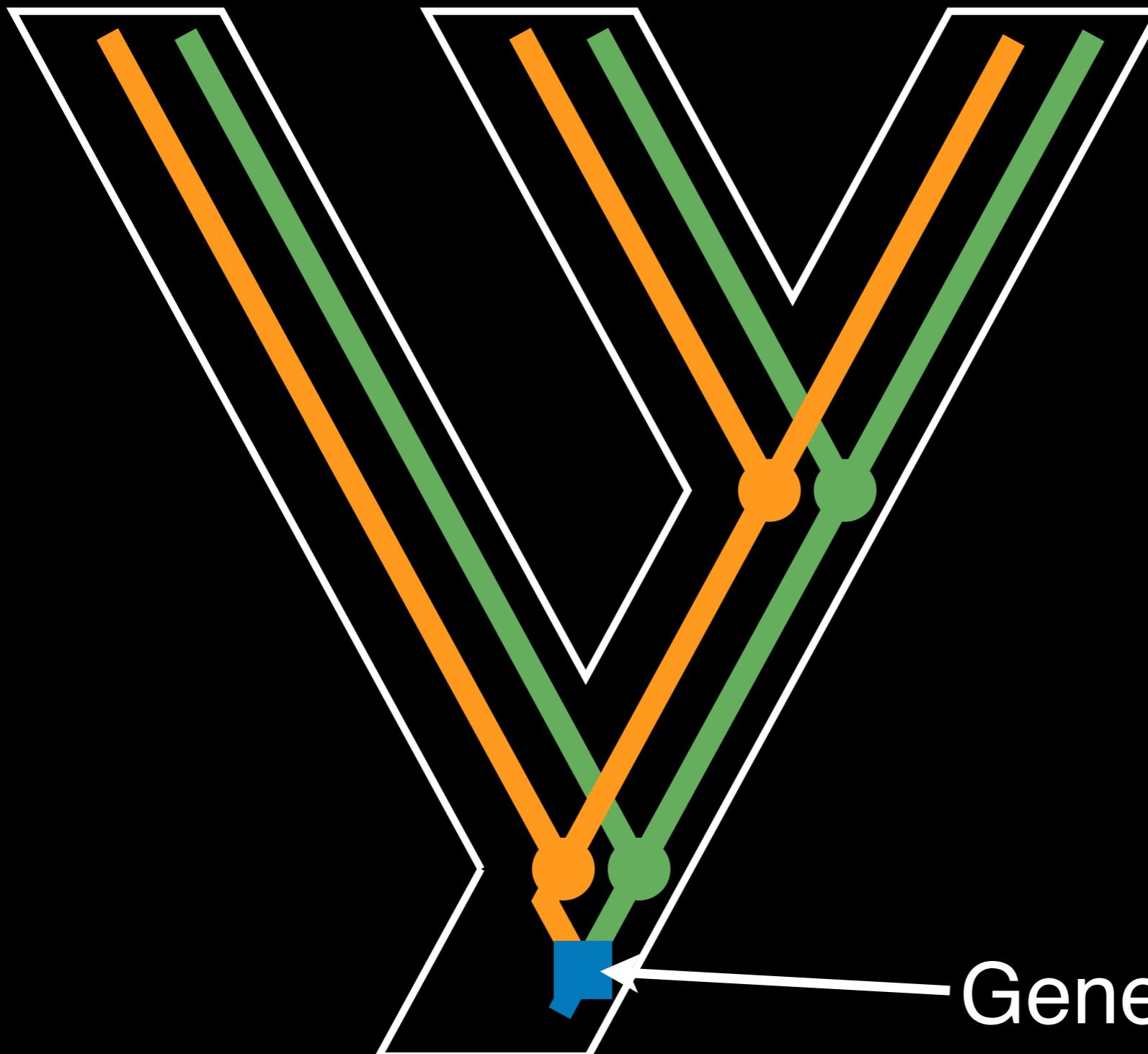
Species A



Species B



Species C



Gene divergence
due to duplication

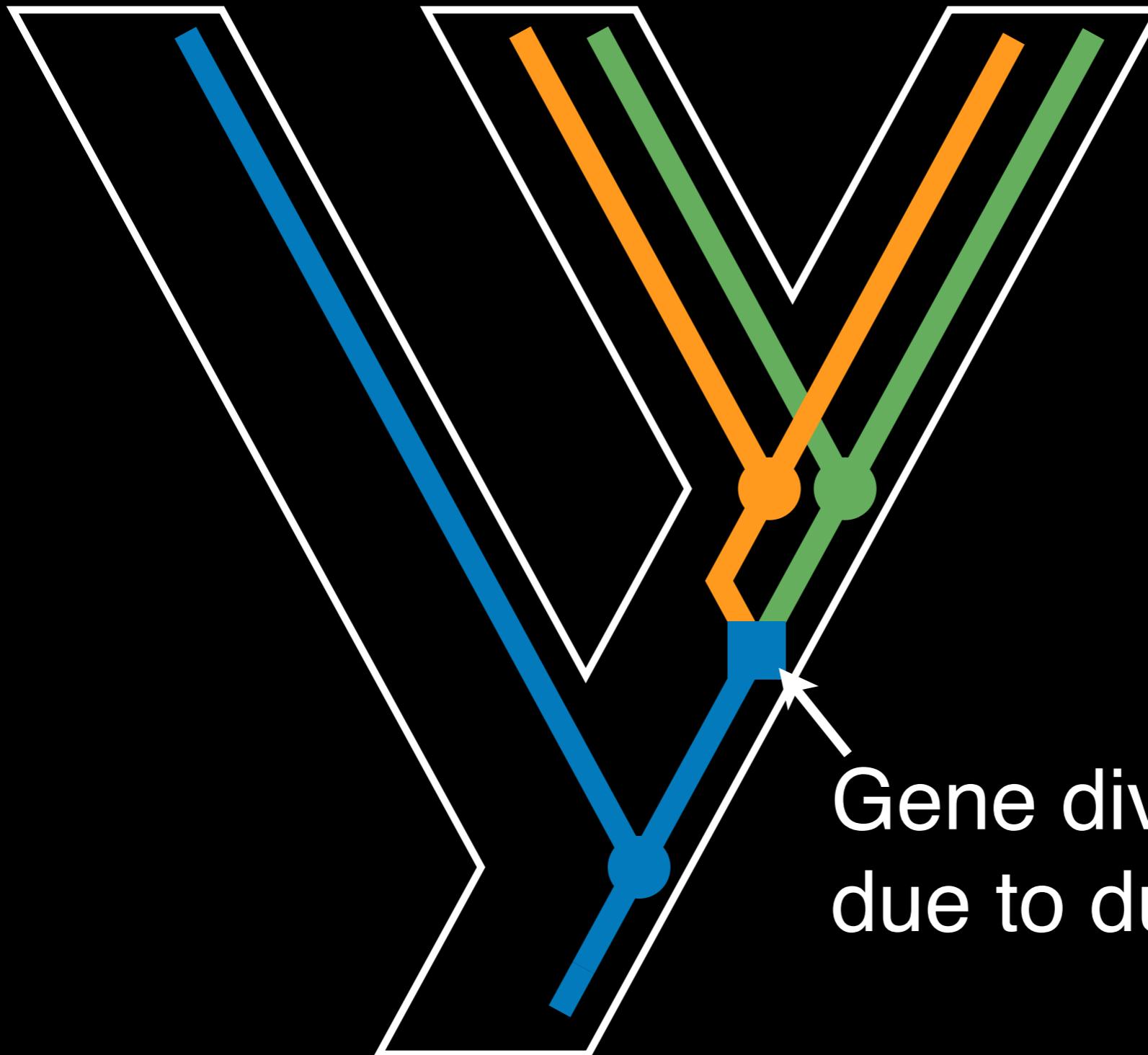
Species A

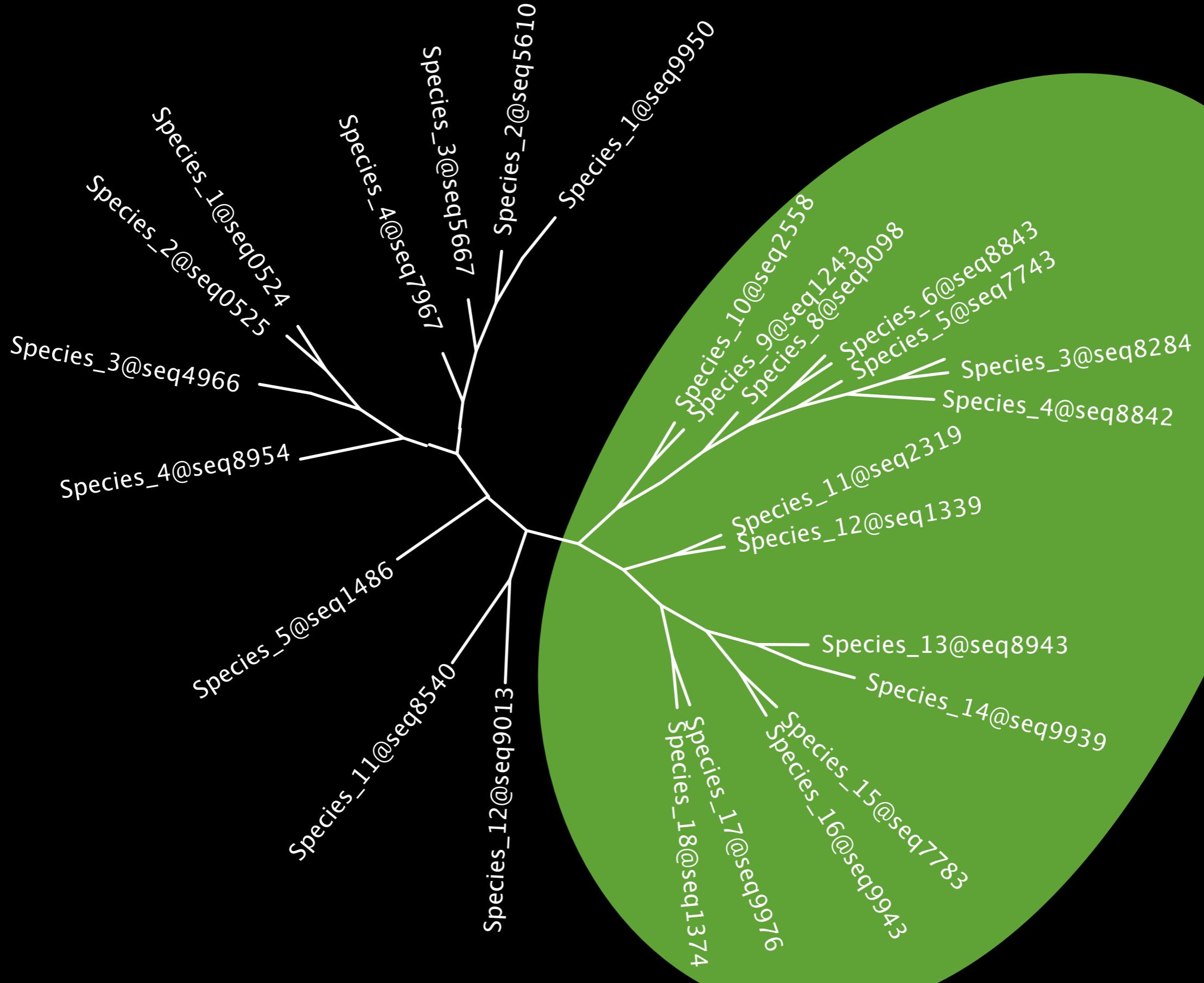


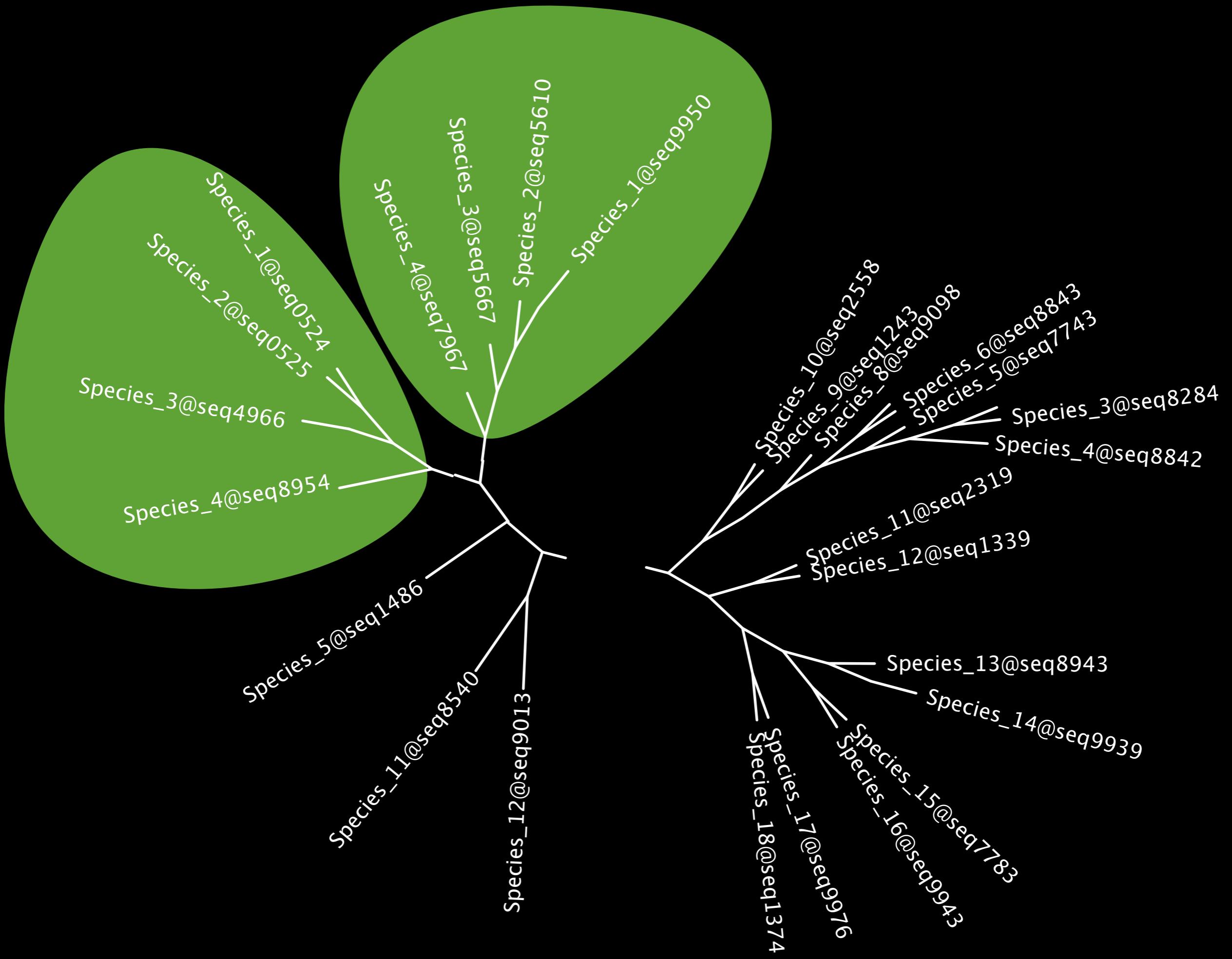
Species B

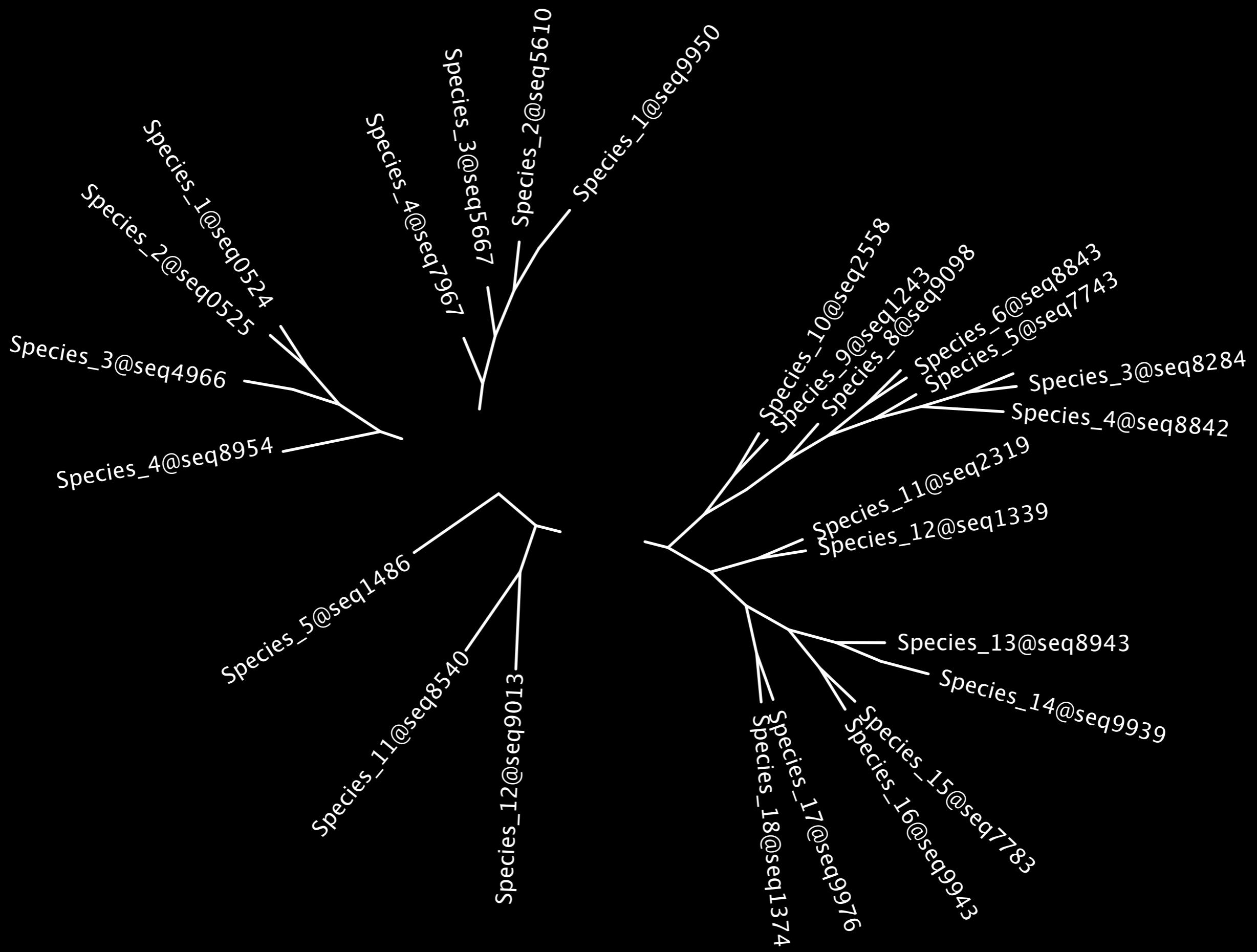


Species C







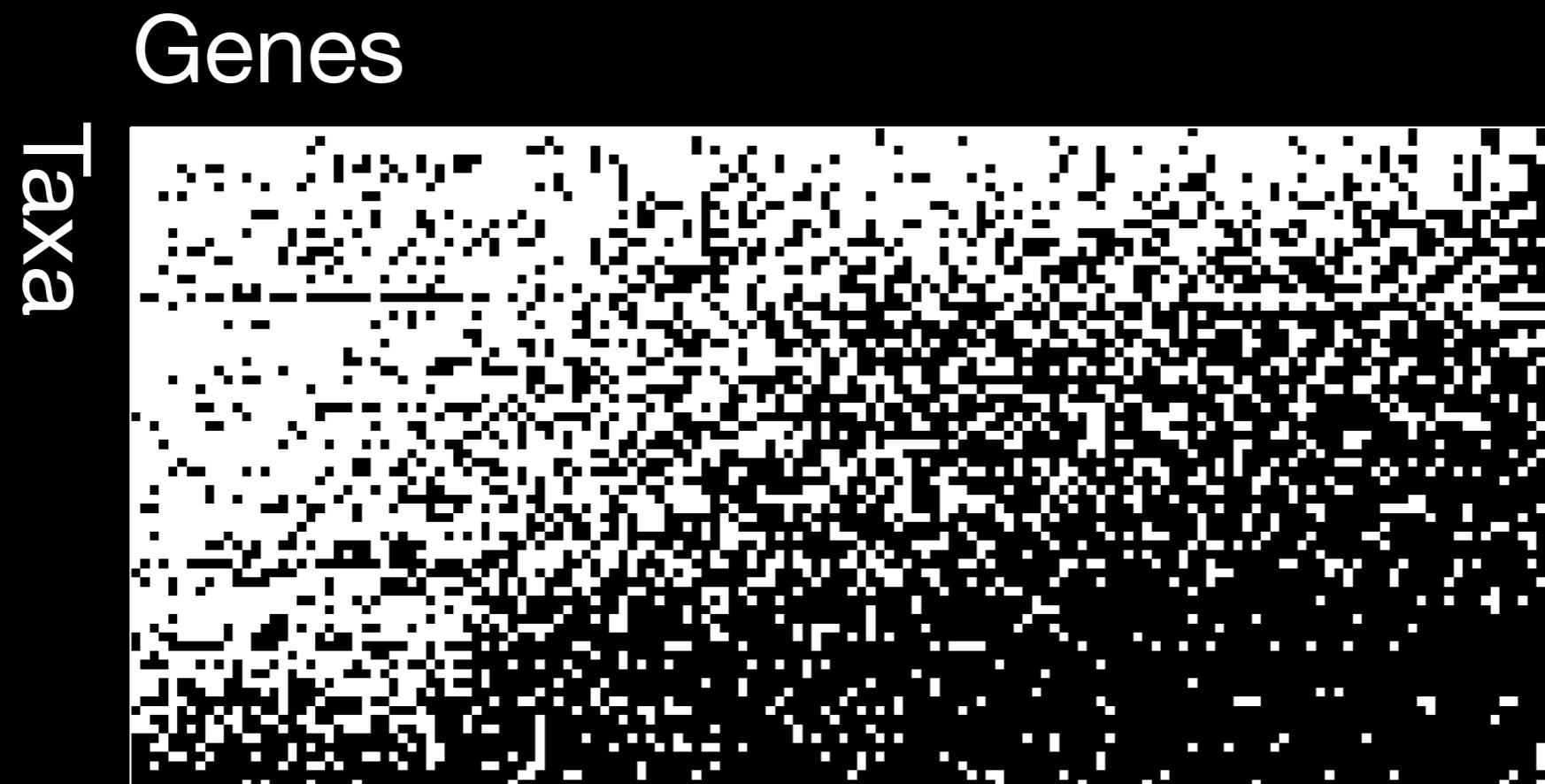


Once we have subtrees of orthologs...

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>

The screenshot shows the Bitbucket interface for the 'agalma' repository. At the top, there's a navigation bar with 'Bitbucket', 'Repositories', 'Create', and a search bar. Below the header, the repository name 'agalma' is displayed with a blue circular icon containing 'Ag'. It shows the owner 'caseywdunn', a 'Following' button, and a 'Share' button. To the right are buttons for 'Clone', 'Fork', 'Compare', and 'Pull request'. Below this, a navigation menu includes 'Overview', 'Source', 'Commits', 'Pull requests', 'Issues (1)', 'Downloads (0)', and a gear icon for settings.

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

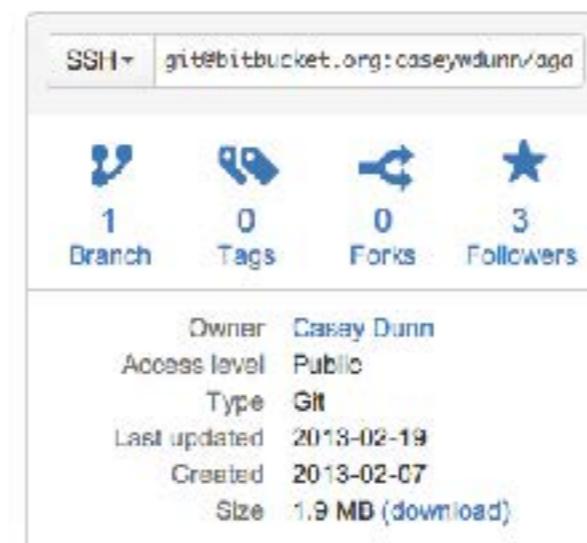
Agalma is a set of analysis pipelines for transcriptome assembly (paired-end Illumina data) and phylogenetic analysis. It can import gene predictions from other sources (eg, assembled non-Illumina transcriptomes or gene models from annotated genomes), enabling broadly-sampled "phylogenomic" analyses.

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 original 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 adaptors (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.

The need to isolate orthologs to study species relationships is a technical quirk of our tools.

Current tools don't model gene gain, duplication, and loss. If we want to know species relationships, we need to give current tools orthologs.

It's analogous to Multiple Sequence Alignment...

We align sequences because our tools don't model insertion and deletion events.

Conceptually, there is no
need to isolate orthologs.
Instead, duplication and loss
could be part of the model.

New approach:

- 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 Bayesian Supertree Model for Genome-Wide Species Tree Reconstruction

LEONARDO DE OLIVEIRA MARTINS*, DIEGO MALLO, AND DAVID POSADA

Department of Biochemistry, Genetics and Immunology, University of Vigo, Vigo, 36310, Spain

*Correspondence to be sent to: *Department of Biochemistry, Genetics and Immunology, University of Vigo, Vigo, 36310, Spain; E-mail: leomrtns@uvigo.es.*

Received 5 February 2014; reviews returned 4 June 2014; accepted 30 September 2014
Associate Editor: Laura Kubatko

parameters. These measures of gene tree/species tree disagreement can be the number of DL or the number of deep coalescences, and the penalty parameters describe how strictly we penalize dissimilar gene/species tree pairs. Note that in our model the species tree becomes then a hyperparameter, that we furthermore assume to come from a fixed uniform hyperprior over all possible species trees with the same number of taxa. The penalty parameters are also hyperparameters, but whose hyperpriors are not fixed. This is because if several gene families are available, then it is natural to partition them since they can obviously have distinct gene trees, as well as their own substitution parameters.

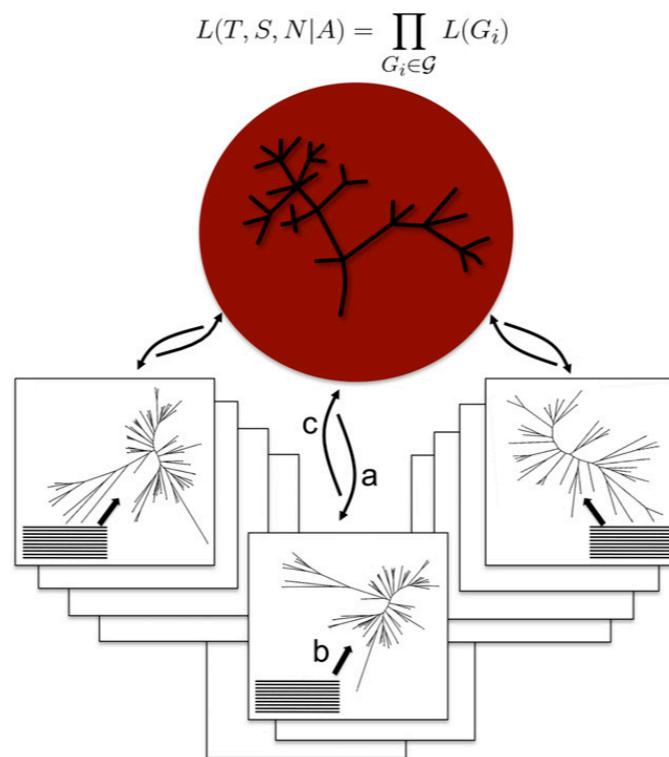
Genome-scale coestimation of species and gene trees

Bastien Boussau,^{1,2,4} Gergely J. Szöllősi,¹ Laurent Duret,¹ Manolo Gouy,¹
Eric Tannier,^{1,3} and Vincent Daubin¹

¹Université de Lyon, Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Villeurbanne

F-69622, France; ²Department of Integrative Biology, UC Berkeley, Berkeley, California 94720-3140, USA; ³INRIA Rhône-Alpes, Montbonnot F-38322, France

Genome Res. 2013 23: 323-330 originally published online November 6, 2012
Access the most recent version at doi:[10.1101/gr.141978.112](https://doi.org/10.1101/gr.141978.112)



Assessing Approaches for Inferring Species Trees from Multi-Copy Genes

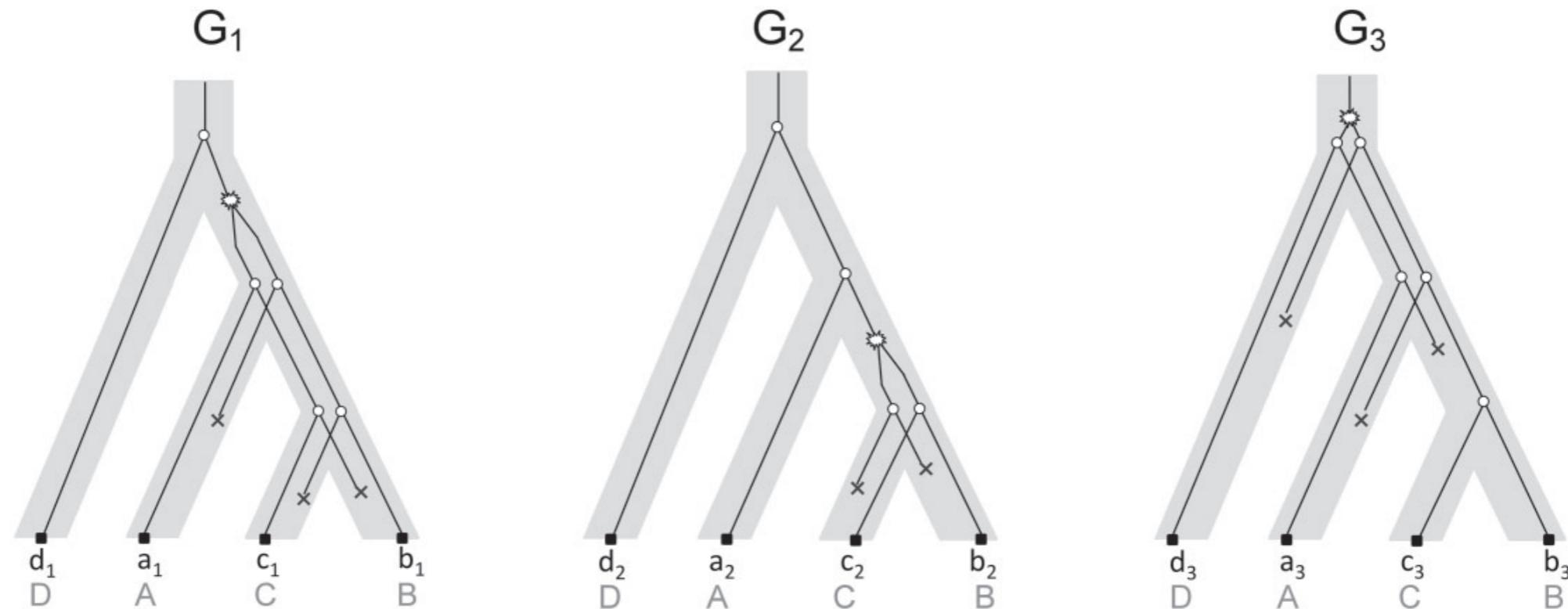
RUCHI CHAUDHARY^{1,2,*}, BASTIEN BOUSSAU³, J. GORDON BURLEIGH², AND DAVID FERNÁNDEZ-BACA¹

¹Department of Computer Science, Iowa State University, Ames, IA 50011, USA; ²Department of Biology, University of Florida, Gainesville, FL 32611, USA; and ³Université de Lyon, Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Villeurbanne F-69622, France

*Correspondence to be sent to: Ruchi Chaudhary, Department of Computer Science, Iowa State University, Ames, IA 50011, USA; E-mail: ruchic@ufl.edu.

Received 30 May 2014; reviews returned 21 August 2014; accepted 18 December 2014

Associate Editor: Laura Kubato



In duplication and loss simulation experiments, MulRF is more accurate than the other methods when the duplication and loss rates are low, and Dup-loss is generally the most accurate when the duplication and loss rates are high. PHYLOG performs well in 10-taxon duplication and loss simulations, but its run time is prohibitively long on larger data sets. In the larger duplication and loss simulation experiments, MulRF outperforms all other methods in experiments with at most 100

Summary of approaches

1. Only examine genes without paralogs (“strict orthologs”)
2. Isolate orthologs according to phenetic criteria (eg orthoMCL, OMA)
3. Isolate orthologs according to phylogenetic criteria (eg Agalma)
4. Simultaneously infer gene trees, species trees, and orthology

Thoughts on orthology/ paralogy

There are many ways to accommodate paralogy when inferring species relationships

We should move toward model-based approaches as they become computationally tractable

The history of gene duplication/loss, incomplete lineage sorting, and horizontal transfer may be non identifiable

**How bad are
the current
limitations?**

Potentially
quite bad.

Dunn and Munro 2016

<http://dx.doi.org/10.1111/zsc.12211>

1. The seductive simplicity of single copy genes

Why do phylogenetic studies tend to focus on genes that occur in single copy?

Don't have to deal with technical challenges of making sure you sequence and correctly assign each paralog

Don't have to deal with the complicated and potentially misleading biological processes that affect multi-copy gene family evolution.

**Why do phylogenetic studies tend to focus
on genes that occur in single copy?**

It is a simpler problem
than multi-copy genes.

Or is it?

Let's talk about mutations for a second...

Mutation rate - The frequency of genetic changes between parents and offspring

Substitution rate - The rate at which genetic changes become fixed in the population

The substitution rate depends
on both the mutation rate and
the rate at which new
mutations are lost

Gene duplication rate

Duplicate origin rate - The frequency at which genes duplicate between parents and offspring

Duplicate fixation rate - The rate at which gene duplications become fixed in the population

The duplicate fixation rate depends on both the duplicate origin rate and the duplicate loss rate.

Single copy genes have a low
duplicate fixation rate.

Do they have a low duplicate
origin rate?

Or a high duplicate loss rate?

For most genes, there isn't a
reason to expect different
duplicate origin rates.

There is growing evidence that
there are big differences in
duplicate loss rates across
genes.

Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants

Riet De Smet^{a,b}, Keith L. Adams^{a,c}, Klaas Vandepoele^{a,b}, Marc C. E. Van Montagu^{a,b,1}, Steven Maere^{a,b}, and Yves Van de Peer^{a,b,1}

2898–2903 | PNAS | February 19, 2013 | vol. 110 | no. 8

www.pnas.org/cgi/doi/10.1073/pnas.1300127110

The importance of gene gain through duplication has long been appreciated. In contrast, the importance of gene loss has only recently attracted attention. Indeed, studies in organisms ranging from plants to worms and humans suggest that duplication of some genes might be better tolerated than that of others. Here we have undertaken a large-scale study to investigate the existence of duplication-resistant genes in the sequenced genomes of 20 flowering plants. We demonstrate that there is a large set of genes that is convergently restored to single-copy status following multiple genome-wide and smaller scale duplication events. We rule out the possibility that such a pattern could be explained by random gene loss only and therefore propose that there is selection pressure to preserve such genes as singletons. This is further substantiated by the observation that angiosperm single-copy genes do not comprise a random fraction of the genome, but instead are often involved in essential housekeeping functions that are highly conserved across all eukaryotes. Furthermore, single-copy genes are generally expressed more highly and in more tissues than non-single-copy genes, and they exhibit higher sequence conservation. Finally, we propose different hypotheses to explain their resistance against duplication.

Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants

Riet De Smet^{a,b}, Keith L. Adams^{a,c}, Klaas Vandepoele^{a,b}, Marc C. E. Van Montagu^{a,b,1}, Steven Maere^{a,b}, and Yves Van de Peer^{a,b,1}

2898–2903 | PNAS | February 19, 2013 | vol. 110 | no. 8

www.pnas.org/cgi/doi/10.1073/pnas.1300127110

Dominant-negative effect hypothesis:

Some genes have strong dominant negative interactions.

When there are multiple copies, there is a greater chance that there will be a mutation in one that will compromise the function of all.

Selection rapidly restores these genes to single copy following duplication.

Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants

Riet De Smet^{a,b}, Keith L. Adams^{a,c}, Klaas Vandepoele^{a,b}, Marc C. E. Van Montagu^{a,b,1}, Steven Maere^{a,b}, and Yves Van de Peer^{a,b,1}

2898–2903 | PNAS | February 19, 2013 | vol. 110 | no. 8

www.pnas.org/cgi/doi/10.1073/pnas.1300127110

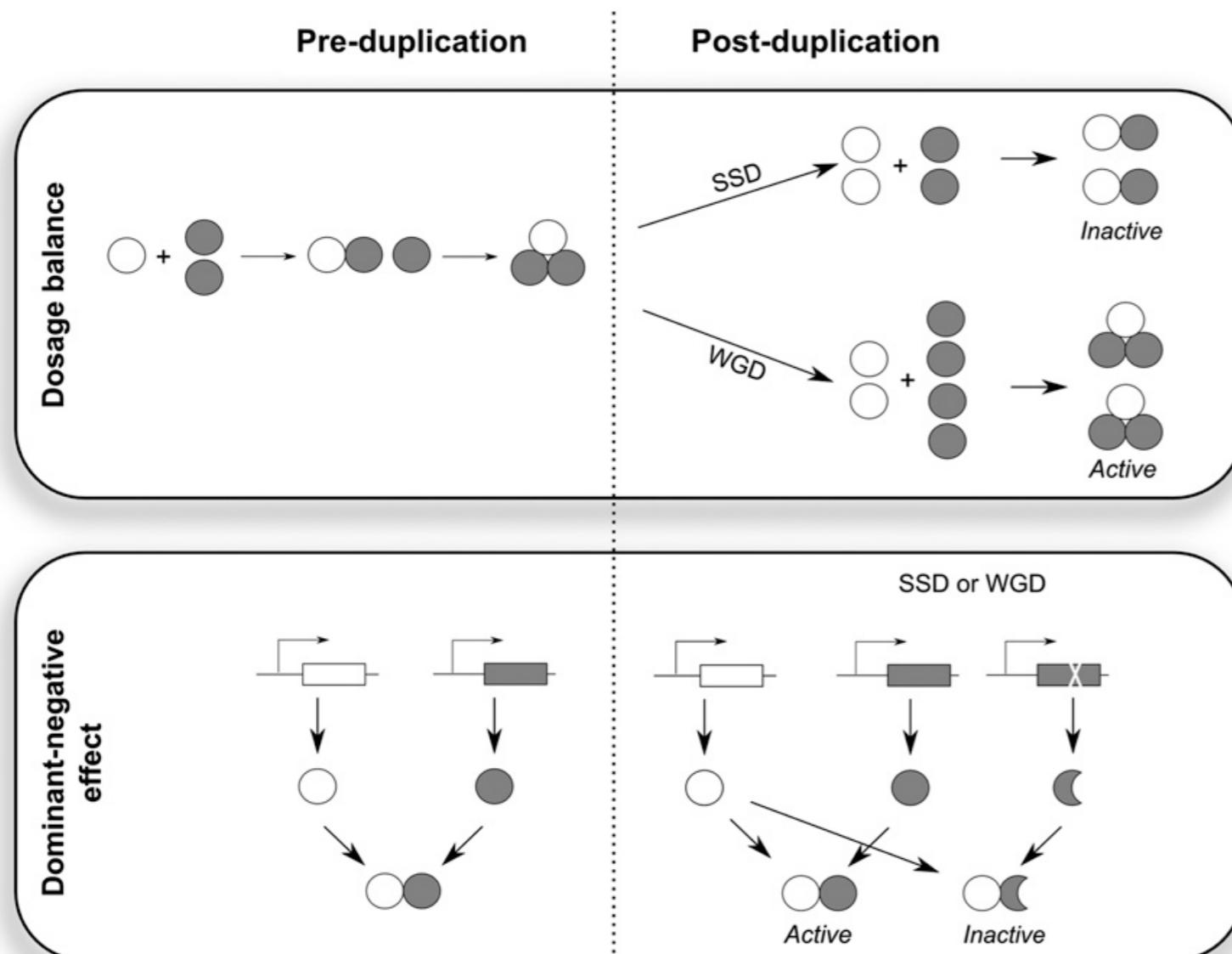


Fig. 4. Two hypotheses to explain single-copy status. (*Upper*) The dosage balance hypothesis, which predicts that stoichiometric imbalance among protein complex subunits is harmful. WGD ensures that relative ratios among subunits are maintained, whereas this is not necessarily the case for SSD (e.g., when the white subunit is duplicated). (*Lower*) The dominant-negative mutation hypothesis, which can explain single-copy status under both scenarios of SSD and WGD. In this hypothesis, gene duplication can result in an extra mutational target, in which mutations can occur that interfere with wild-type function. This is, for instance, possible if a mutation occurs in a protein complex subunit, in such a sense that protein interaction capabilities are maintained and hence the mutant protein competes with the wild-type protein for forming complexes.

Traditional perspective

Single copy genes follow
“standard” evolutionary
processes

Multi-copy genes happen when
you have “standard” evolution
plus an elevated duplication
rate

Emerging perspective

Multi-copy genes follow
“standard” evolutionary
processes

Single-copy genes happen
when you have “standard”
evolution plus an elevated
duplication loss rate

What does this mean?

At a minimum, focusing on single-copy genes excludes particular patterns of molecular evolution (eg, favors slower genes)

At worse, it may lead to biases that negatively impact your analyses

What does this mean?

Focusing on single copy genes
doesn't necessarily reduce impacts of
gene duplication, it just hides them.

2. The limitations of “orthology” and “paralogy”

Orthology and paralogy are central concepts in the way we describe genes and the evolutionary processes that gave rise to them.

These terms are applied many ways, but are only unambiguous when talking about pairs.

Orthologs - two genes whose most recent common ancestor precede a **speciation** event.

Paralogs - two genes whose most recent common ancestor precede a gene **duplication** event.

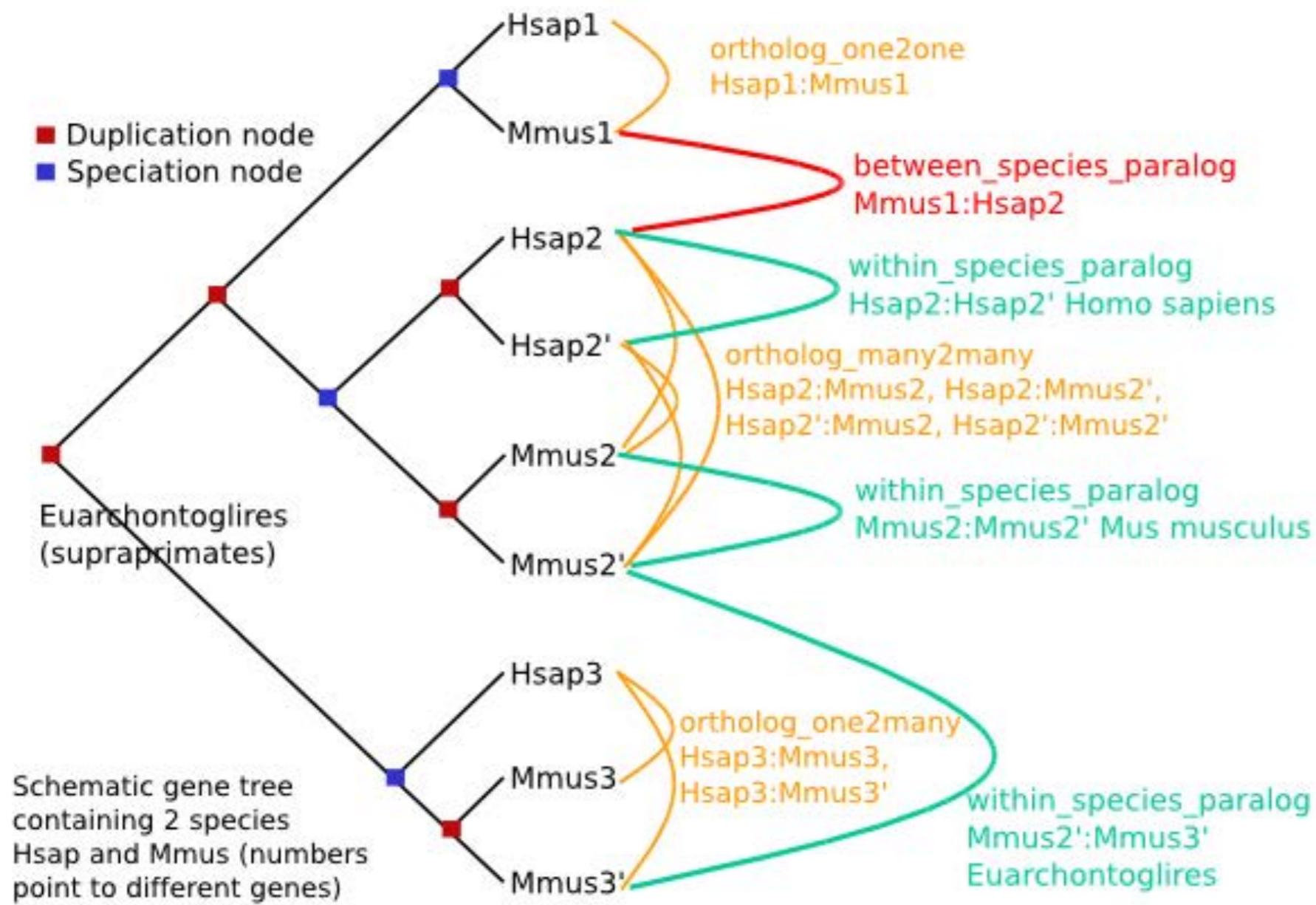
Paralogy and orthology can have clear interpretations on small simple gene trees.

Most of us don't have small simple gene trees anymore.

The fact that the terminology breaks down for common problem is a strong indication that the utility of the concepts are breaking down.

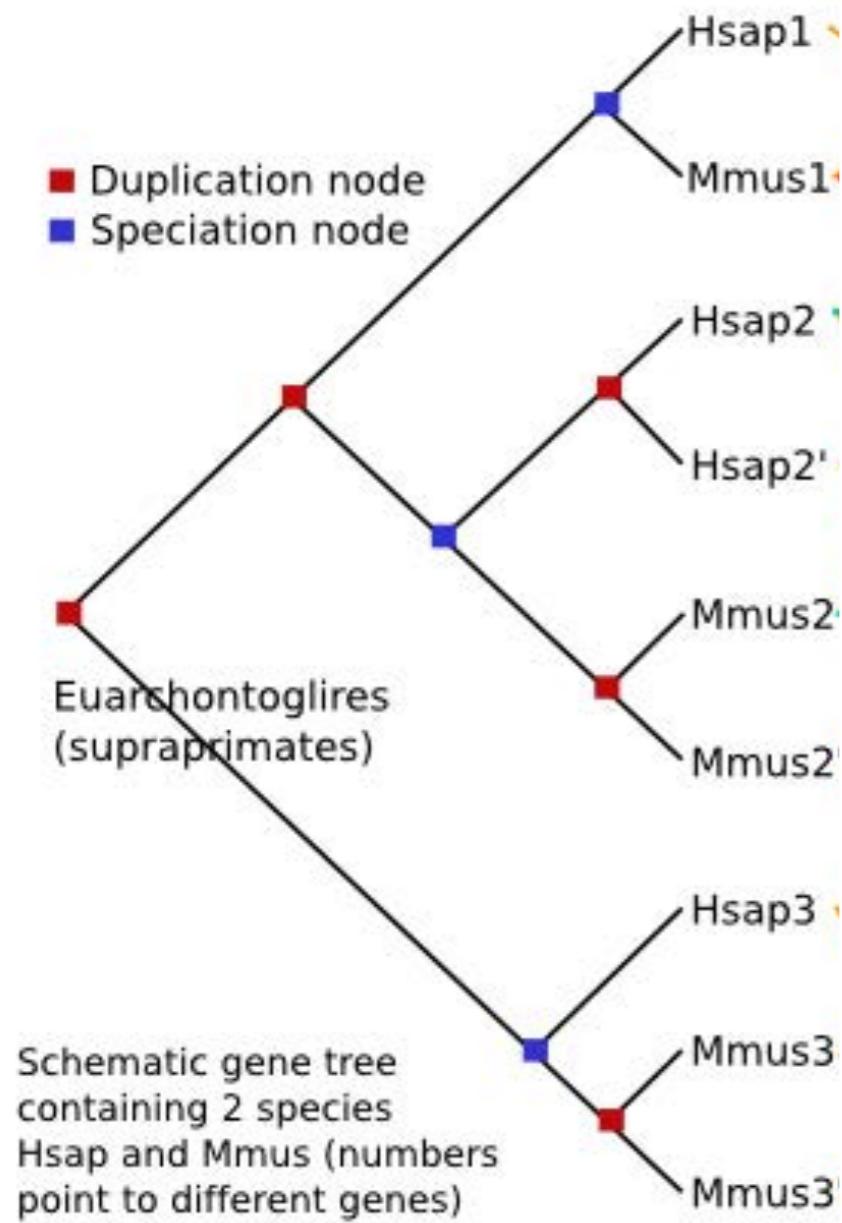
Dunn and Munro 2016
doi:10.1111/zsc.12211

Orthology and paralogy are unnecessarily complicated and overrated...



one to one ortholog
between species paralog
within species paralog
inparalog
outparalog

Orthology and paralogy are unnecessarily complicated and overrated...



To analyze past evolutionary processes, we should look within the tree - not at the tips.

A suggestion:

Whenever you are about to use the words orthology or paralogy, instead explain your idea in terms of speciation and duplication.

You will feel a little weight lifted.

Part 3: Comparative genomics

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

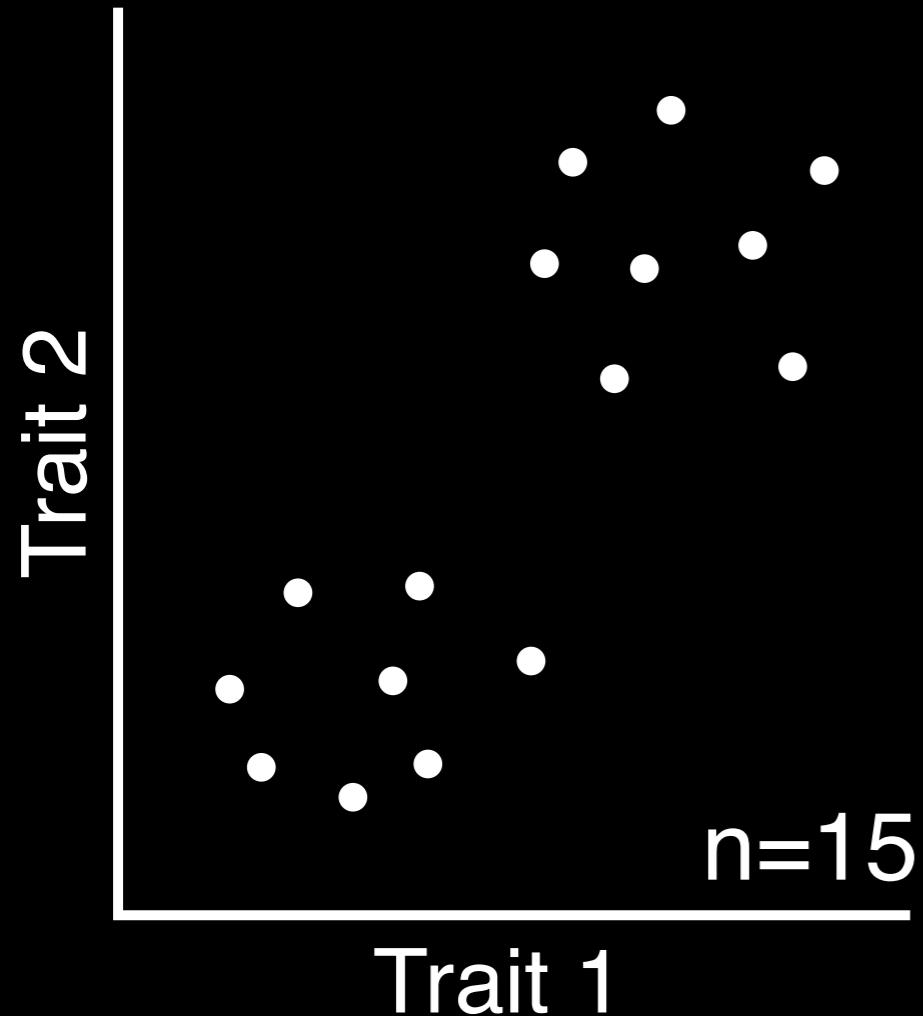
Comparative biology

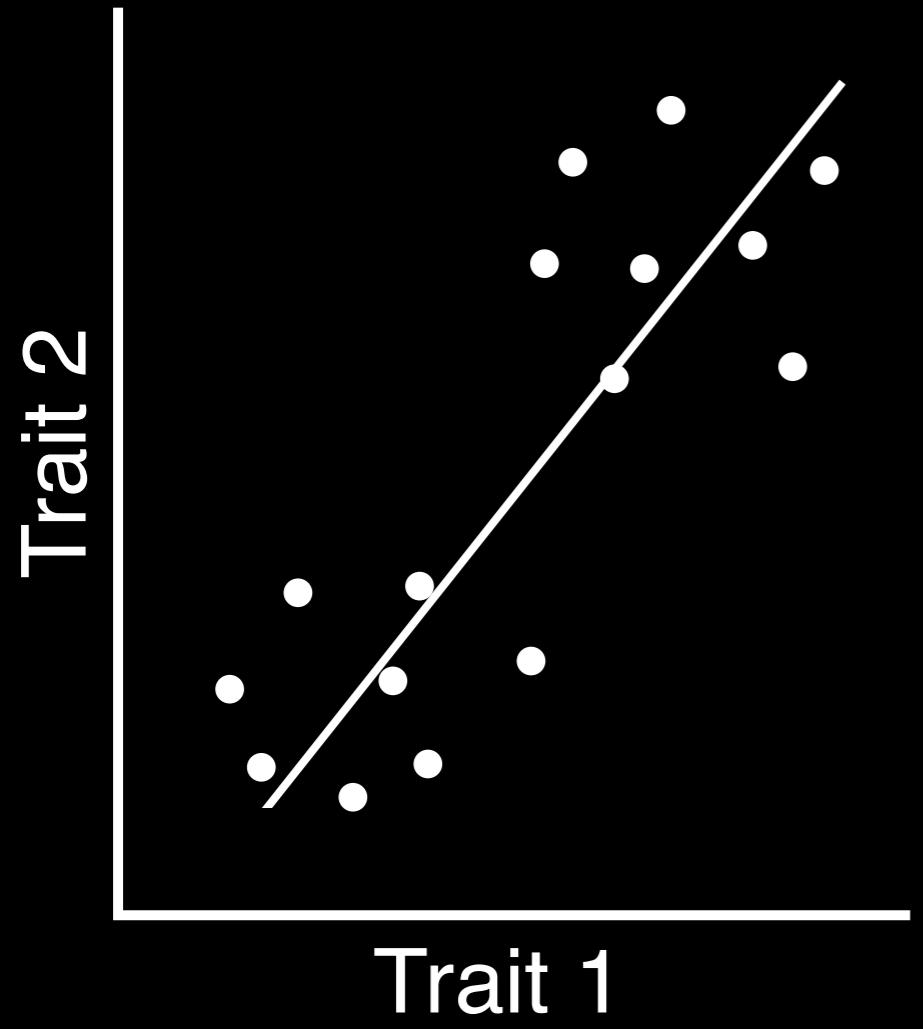
The comparison of traits
across individuals,
populations, and species.

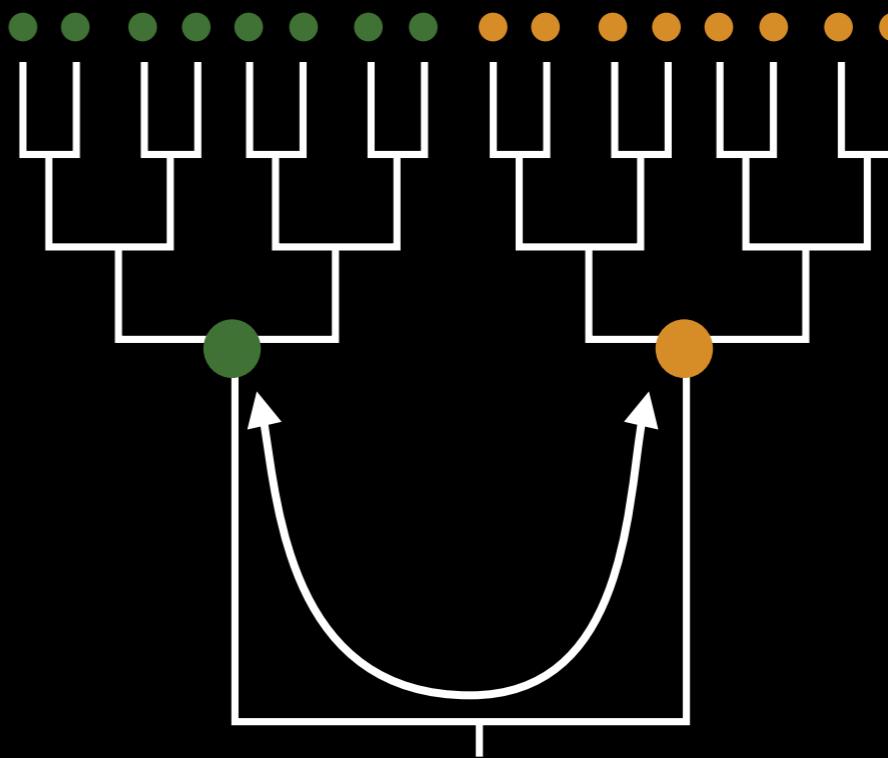
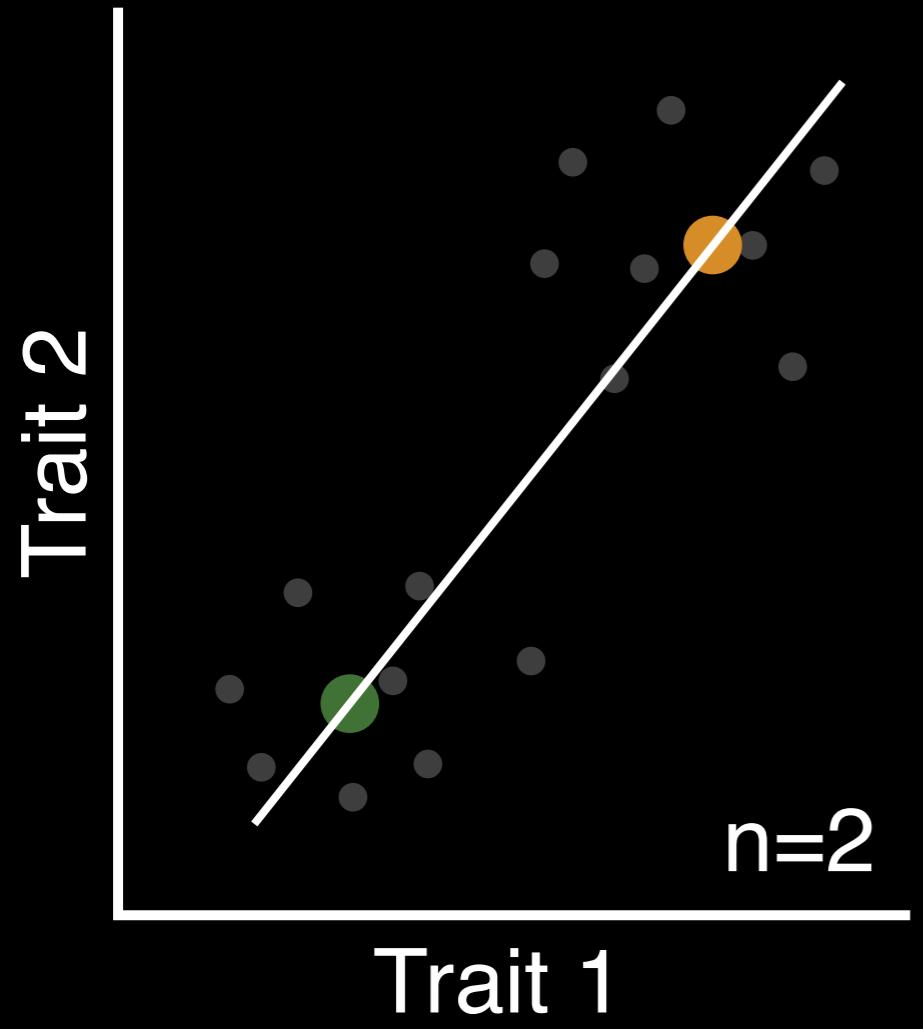
Simple case:

Two traits (x,y axes)

Fifteen individuals (points)







Vol. 125, No. 1

The American Naturalist

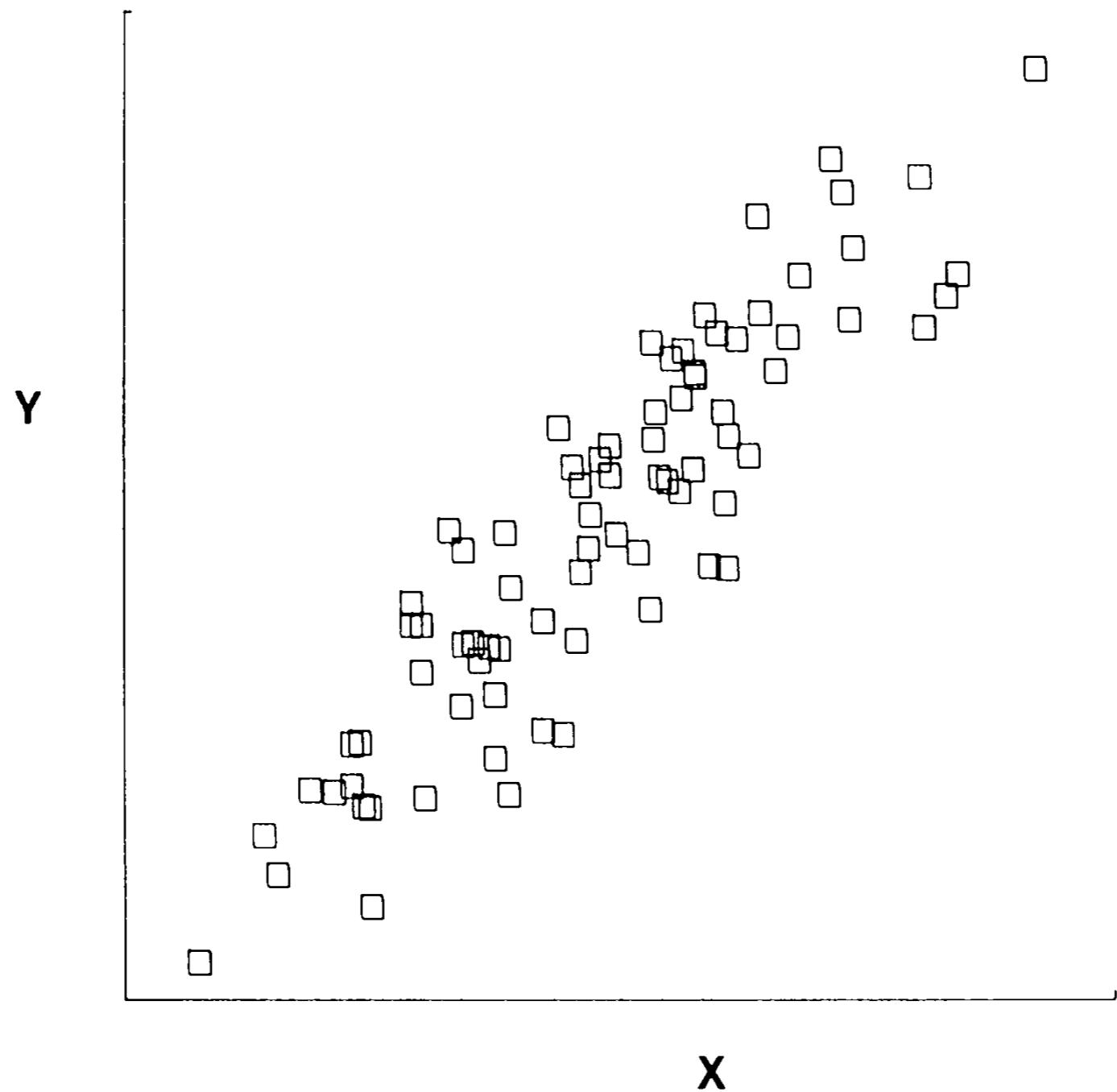
January 1985

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



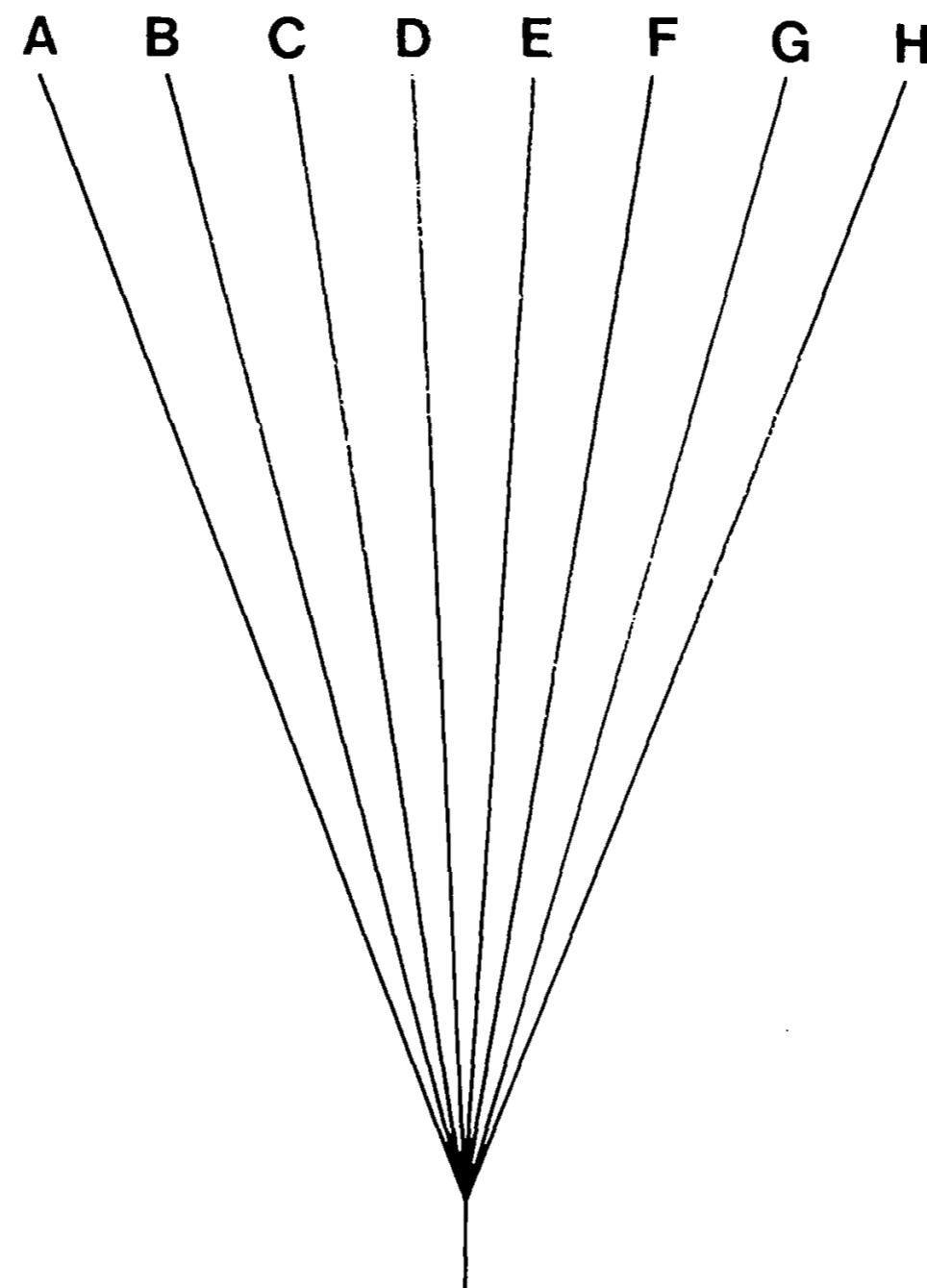


FIG. 2.—One phylogeny for the 8 species, showing a burst of adaptive radiation with each lineage evolving independently from a common starting point.

Felsenstein, 1985

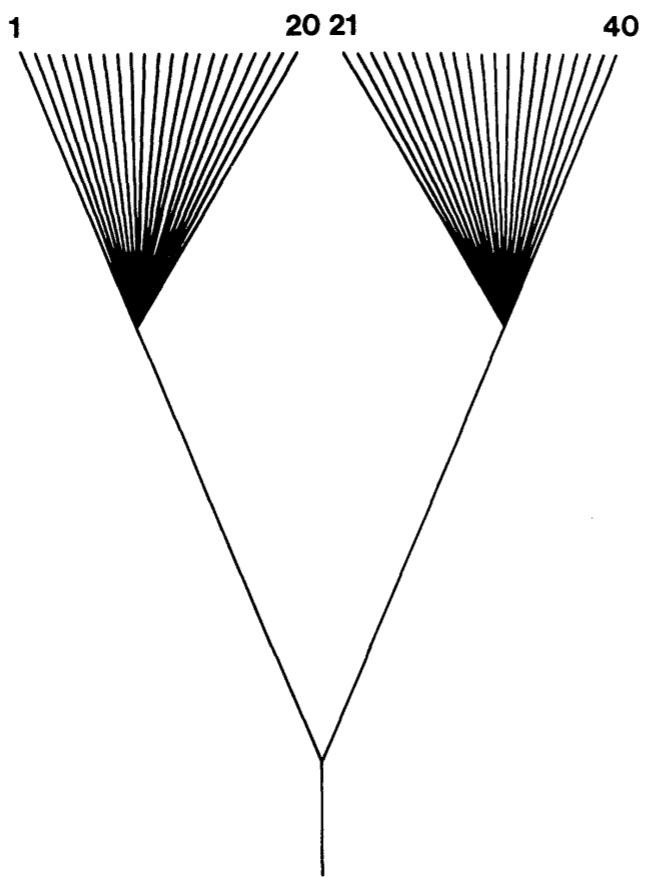


FIG. 5.—A “worst case” phylogeny for 40 species, in which there prove to be 2 groups each of 20 close relatives.

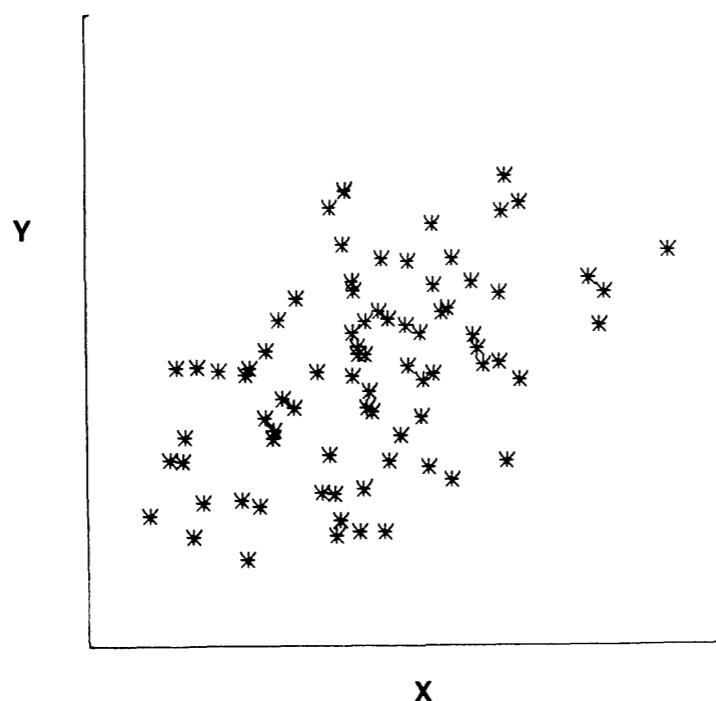


FIG. 6.—A typical data set that might be generated for the phylogeny in fig. 5 using the model of independent Brownian motion (normal increments) in each character.

Felsenstein, 1985

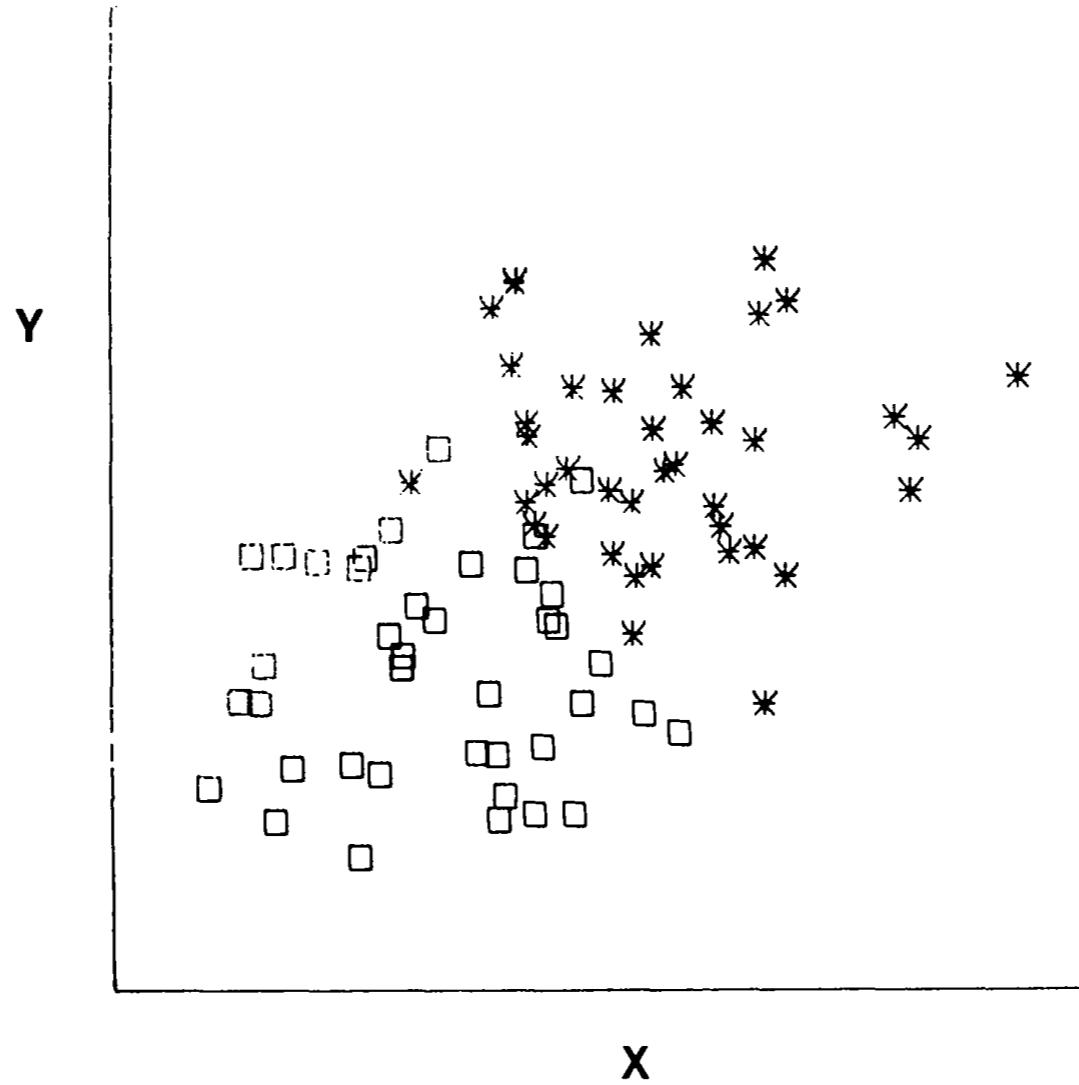


FIG. 7.—The same data set, with the points distinguished to show the members of the 2 monophyletic taxa. It can immediately be seen that the apparently significant relationship of fig. 6 is illusory.

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

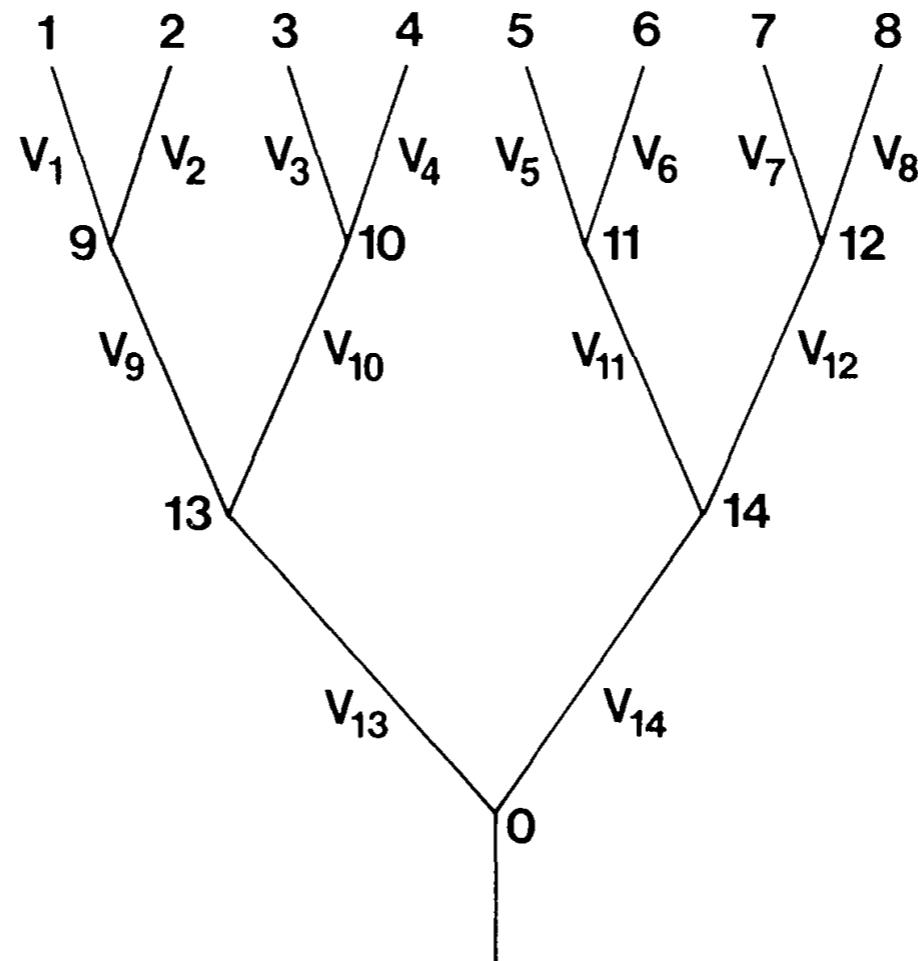


FIG. 8.—An example of a phylogeny, assumed known, from which we can define independent contrasts between taxa. This tree is highly symmetric, so that $v_1 = v_2 = v_3 = v_4 = v_5 = v_6 = v_7 = v_8$, $v_9 = v_{10} = v_{11} = v_{12}$, and $v_{13} = v_{14}$.

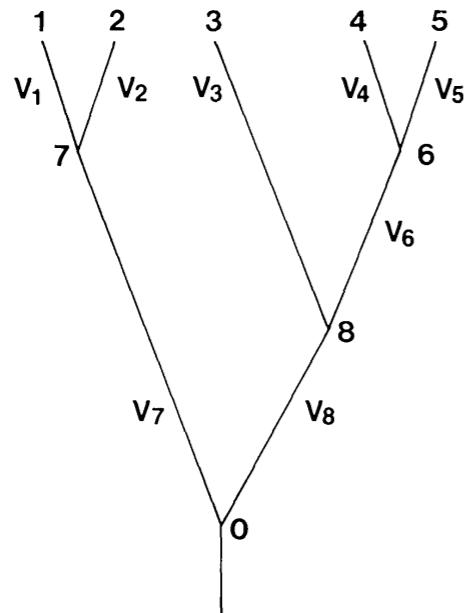


FIG. 9.—A less symmetrical phylogeny. The independent contrasts for this phylogeny are given in table 1.

TABLE 1
THE FOUR CONTRASTS EXTRACTED FROM THE PHYLOGENY SHOWN IN FIGURE 9, EACH WITH ITS VARIANCE, ALL COMPUTED USING STEPS 1–4 IN THE TEXT

CONTRAST	VARIANCE
$X_1 - X_2$	$v_1 + v_2$
$X_4 - X_5$	$v_4 + v_5$
$X_3 - X_6$	$v_3 + v'_6$
$X_7 - X_8$	$v'_7 + v'_8$

where

$$X_6 = \frac{v_4 X_5 + v_5 X_4}{v_4 + v_5}$$

$$v'_6 = v_6 + v_4 v_5 / (v_4 + v_5)$$

$$X_7 = \frac{v_2 X_1 + v_1 X_2}{v_1 + v_2}$$

$$v'_7 = v_7 + v_1 v_2 / (v_1 + v_2)$$

$$X_8 = \frac{v'_6 X_3 + v_3 X_6}{v_3 + v_6}$$

$$v'_8 = v'_7 + v_3 v'_6 / (v_3 + v'_6)$$

What if We Do Not Take the Phylogeny into Consideration?

Some reviewers of this paper felt that the message was “rather nihilistic,” and suggested that it would be much improved if I could present a simple and robust method that obviated the need to have an accurate knowledge of the phylogeny. I entirely sympathize, but do not have a method that solves the problem. The best we can do is perhaps to use pairs of close relatives as suggested above, although this discards at least half of the data. Comparative biologists may understandably feel frustrated upon being told that they need to know the phylogenies of their groups in great detail, when this is not something they had much interest in knowing. Nevertheless, efforts to cope with the effects of the phylogeny will have to be made. Phylogenies are fundamental to comparative biology; there is no doing it without taking them into account.

Comparative genomics

Comparative functional
genomics

**Are current cross species comparisons
of genome function using phylogenetic
comparative methods?**

Unfortunately, not so much

Functional genomics applies tools of molecular genetics at genome scale.

Methods include RNA-seq, ChIP-seq, etc.

A new field of comparative functional genomics is emerging that integrates observations across species to investigate the evolution of genome function, and to associate genes and phenotypes.

Link to notes: <https://git.io/evol2017>

This enthusiasm is apparent in a growing number of comparative functional genomics papers:

ARTICLE nature

doi:10.1038/nature10532

The evolution of gene expression levels in mammalian organs

David Brawand^{1,2*}, Magali Soumillon^{1,2*}, Anamaria Necsulea^{1,2*}, Philippe Julien^{1,2}, Gábor Csárdi^{2,3}, Patrick Harrigan⁴, Manuela Weier¹, Angélica Liechti¹, Ayinuer Aximu-Petri⁵, Martin Kircher⁵, Frank W. Albert^{5†}, Ulrich Zeller⁶, Philipp Khaitovich⁷, Frank Grützner⁸, Sven Bergmann^{2,3}, Rasmus Nielsen^{4,9}, Svante Pääbo⁵ & Henrik Kaessmann^{1,2}



Busby et al. BMC Genomics 2011, 12:635
http://www.biomedcentral.com/1471-2164/12/635

Expression divergence measured by transcriptome sequencing of four yeast species

Michele A Busby¹, Jesse M Gray^{2,4}, Allen M Costa², Chip Stewart^{1,5}, Michael P Stromberg¹, Derek Barnett¹, Jeffrey H Chuang¹, Michael Springer³ and Gabor T Marth^{1*}

LETTER nature

doi:10.1038/nature16994

The mid-developmental transition and the evolution of animal body plans

Michal Levin^{1†*}, Leon Anavy^{1*}, Alison G. Cole¹, Eitan Winter¹, Natalia Mostov¹, Sally Khair¹, Naftalie Senderovich¹, Ekaterina Kovalev¹, David H. Silver¹, Martin Feder¹, Selene L. Fernandez-Valverde^{2†}, Nagayasu Nakanishi^{2†}, David Simmons³, Oleg Simakov⁴, Tomas Larsson⁴, Shang-Yun Liu⁵, Ayelet Jerafi-Vider⁶, Karina Yaniv⁶, Joseph F. Ryan³, Mark Q. Martindale³, Jochen C. Rink⁵, Detlev Arendt⁴, Sandie M. Degnan², Bernard M. Degnan², Tamar Hashimshony¹ & Itai Yanai¹



doi:10.1371/journal.pcbi.1005274

Tissue-Specificity of Gene Expression Diverges Slowly between Orthologs, and Rapidly between Paralogs

Nadezda Kryuchkova-Mostacci^{1,2}, Marc Robinson-Rechavi^{1,2*}

CSH PRESS GENOME RESEARCH

doi:10.1101/gr.163014.113

Tempo and mode of regulatory evolution in *Drosophila*

Joseph D. Coolon,¹ C. Joel McManus,^{2,3} Kraig R. Stevenson,⁴ Brenton R. Graveley,³ and Patricia J. Wittkopp^{1,4,5,6}

ARTICLE nature

doi:10.1038/nature12943

The evolution of lncRNA repertoires and expression patterns in tetrapods

Anamaria Necsulea^{1,2†}, Magali Soumillon^{1,2†}, Maria Warnefors^{1,2}, Angélica Liechti^{1,2}, Tasman Daish³, Ulrich Zeller⁴, Julie C. Baker⁵, Frank Grützner³ & Henrik Kaessmann^{1,2}

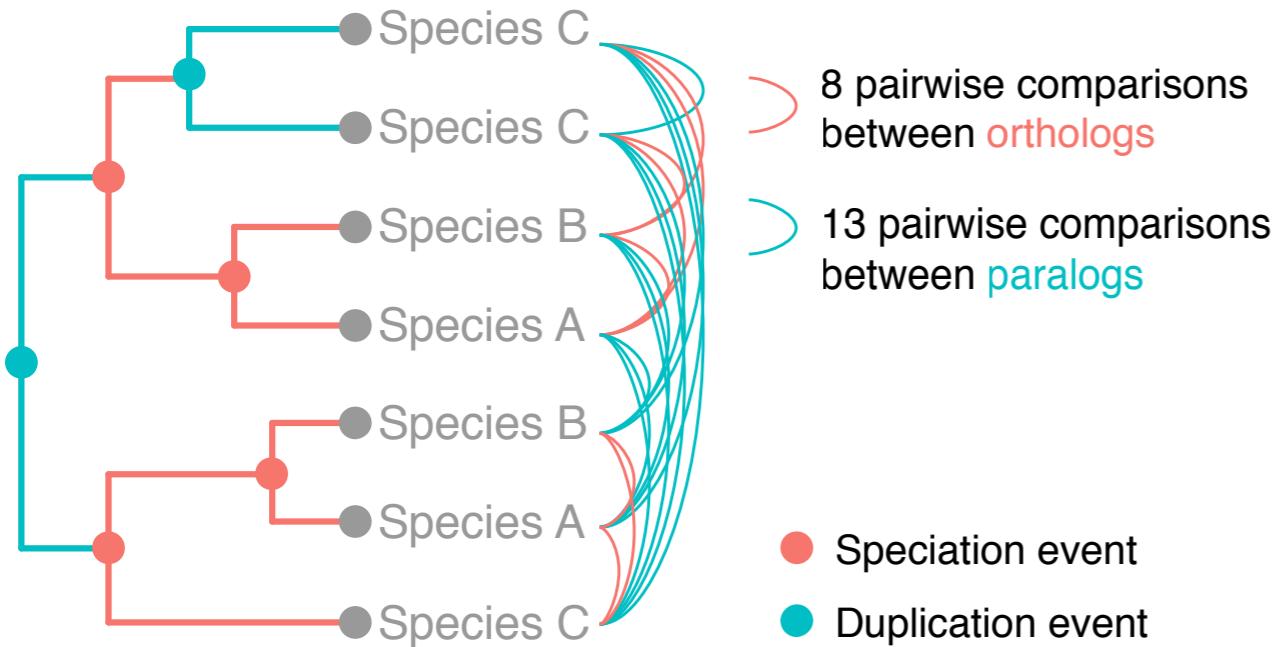
Link to notes: <https://git.io/evol2017>

But there is a
potential problem....

Most of these studies use Pairwise Comparisons

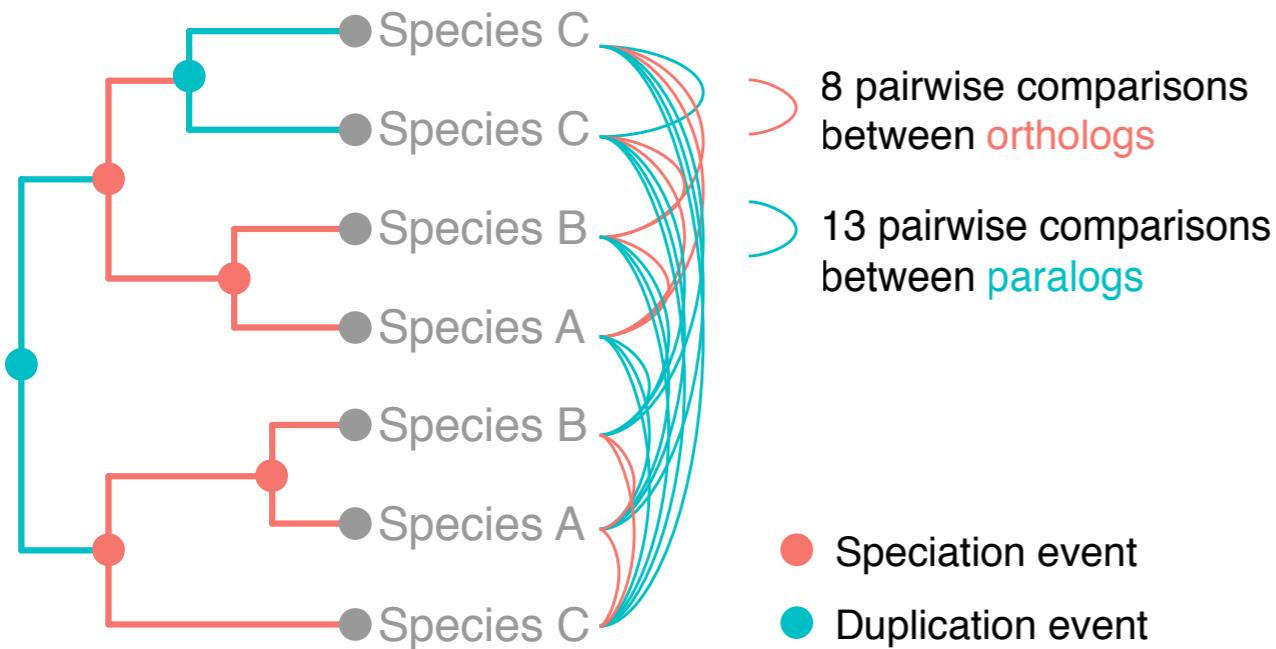


Pairwise Comparisons



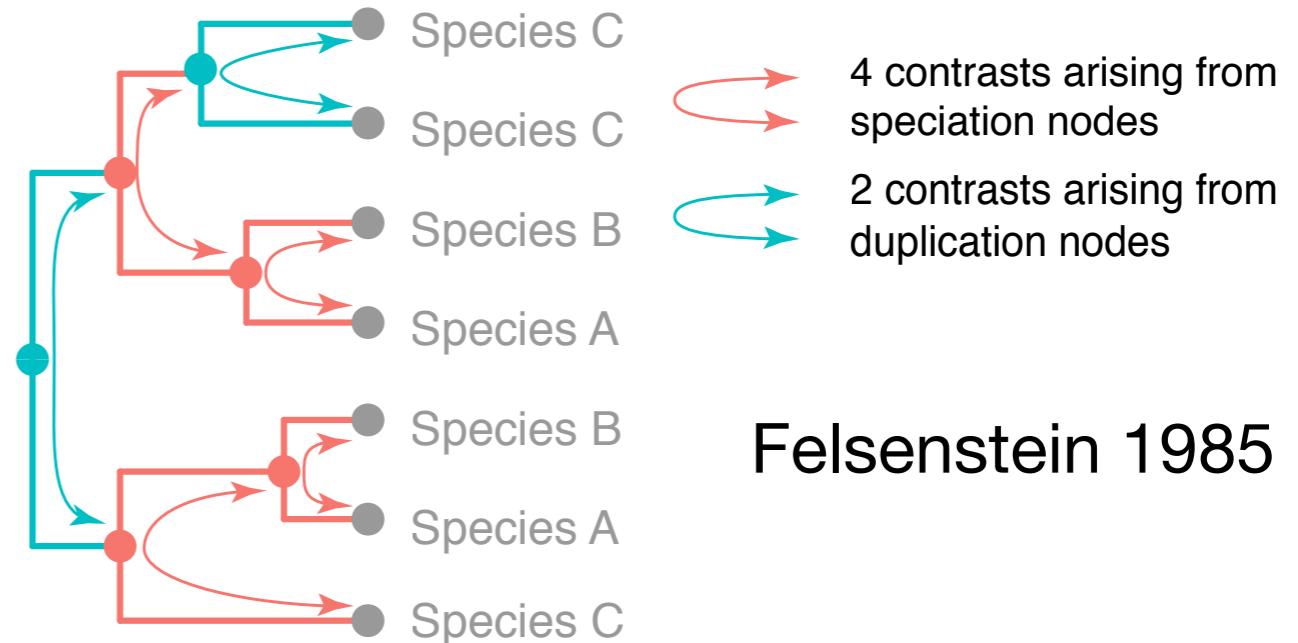
- Don't take evolutionary relationships into account
- Are not independent
- Each branch is included in multiple comparisons.
- A comparison can span many branches and nodes

Pairwise Comparisons



- Don't take evolutionary relationships into account
- Are not independent
- Each branch is included in multiple comparisons.
- A comparison can span many branches and nodes

Phylogenetic Independent Contrasts



Felsenstein 1985

- Do take evolutionary relationships into account
- Are independent
- Each branch is included in only one comparison.
- Isolates changes to well defined regions of the tree

Are the findings of
comparative functional
genomics studies
compromised by the well
known problems with
pairwise comparisons?

Tissue-Specificity of Gene Expression Diverges Slowly between Orthologs, and Rapidly between Paralogs

Nadezda Kryuchkova-Mostacci^{1,2}, Marc Robinson-Rechavi^{1,2*}

LETTER

nature

doi:10.1038/nature16994

The mid-developmental transition and the evolution of animal body plans

Michal Levin^{1†*}, Leon Anavy^{1*}, Alison G. Cole¹, Eitan Winter¹, Natalia Mostov¹, Sally Khair¹, Naftalie Senderovich¹, Ekaterina Kovalev¹, David H. Silver¹, Martin Feder¹, Selene L. Fernandez-Valverde^{2†}, Nagayasu Nakanishi^{2†}, David Simmons³, Oleg Simakov⁴, Tomas Larsson⁴, Shang-Yun Liu⁵, Ayelet Jerafi-Vider⁶, Karina Yaniv⁶, Joseph F. Ryan³, Mark Q. Martindale³, Jochen C. Rink⁵, Detlev Arendt⁴, Sandie M. Degnan², Bernard M. Degnan², Tamar Hashimshony¹ & Itai Yanai¹

Pairwise comparisons are problematic when analyzing functional genomic data across species

Casey W. Dunn^{1*}, Felipe Zapata^{1,2}, Catriona Munro¹, Stefan Siebert^{1,3}, Andreas Hejnol⁴

doi: <https://doi.org/10.1101/107177>

https://github.com/caseywdunn/comparative_expression_2017

The screenshot shows the GitHub repository page for 'caseywdunn/comparative_expression_2017'. The repository has 104 commits, 2 branches, and 4 contributors. The latest commit was made 10 days ago. The repository description states: "This repository contains files associated with our reanalysis of two previously published comparative gene expression studies". A link to notes is provided at the bottom right.

No description, website, or topics provided.

104 commits 2 branches 0 releases 4 contributors

Branch: master New pull request Create new file Upload files Find file Clone or download

caseywdunn/figure scaling	Latest commit 79255d2 10 days ago	
kmrr	ran kmrr/Rscript.R, added files relevant to downstream analyses	a month ago
levin_etal	reformatted with new bib, cherry picked bm model parameters to reduce...	16 days ago
tests/testthat	added testthat framework	10 days ago
.gitignore	reorganized a bit, added to pairwise analyses	a month ago
Figure_overview.pdf	figure formatting, typos	16 days ago
functions.R	modified tree parsing so nearly all are now time calibrated. Moved f...	11 days ago
manuscript.bio	added p values to figures	14 days ago
manuscript.pdf	added shortened url	14 minutes ago
manuscript.rmd	added shortened url	14 minutes ago
nature.csv	added Levin material	a month ago
readme.md	added shortened url	14 minutes ago
readme.md		

Introduction

This repository contains files associated with our reanalysis of two previously published comparative gene expression studies:

Link to notes: <https://git.io/evol2017>

Tissue-Specificity of Gene Expression Diverges Slowly between Orthologs, and Rapidly between Paralogs

Nadezda Kryuchkova-Mostacci^{1,2}, Marc Robinson-Rechavi^{1,2*}

Tissue-Specificity of Gene Expression Diverges Slowly between Orthologs, and Rapidly between Paralogs

Nadezda Kryuchkova-Mostacci^{1,2}, Marc Robinson-Rechavi^{1,2*}

This paper is a test of the
ortholog conjecture.

Two genes are **orthologs** if they diverged due to a speciation event.

Two genes are **paralogs** if they diverged due to a gene duplication event.

Understanding how the evolution of gene function differs between **orthologs** and **paralogs** has been a central focus of understanding the association between genes and phenotypes.

Link to notes: <https://git.io/evol2017>

The ortholog conjecture is the hypothesis that:

- Orthologs are more similar to each other than paralogs are
- Gene evolution is more rapid after duplication than speciation

The ortholog conjecture has received mixed support.

Tissue-Specificity of Gene Expression Diverges Slowly between Orthologs, and Rapidly between Paralogs

Nadezda Kryuchkova-Mostacci^{1,2}, Marc Robinson-Rechavi^{1,2*}

This paper found strong support for the ortholog conjecture with regard to gene expression.

Tissue-Specificity of Gene Expression Diverges Slowly between Orthologs, and Rapidly between Paralogs

Nadezda Kryuchkova-Mostacci^{1,2}, Marc Robinson-Rechavi^{1,2*}

They examined:

- Tissue specificity of gene expression as quantified by τ (where $\tau=0$ is uniform expression, $\tau=1$ is expression entirely specific to one tissue).
- Across multiple public datasets of mammal organ gene expression

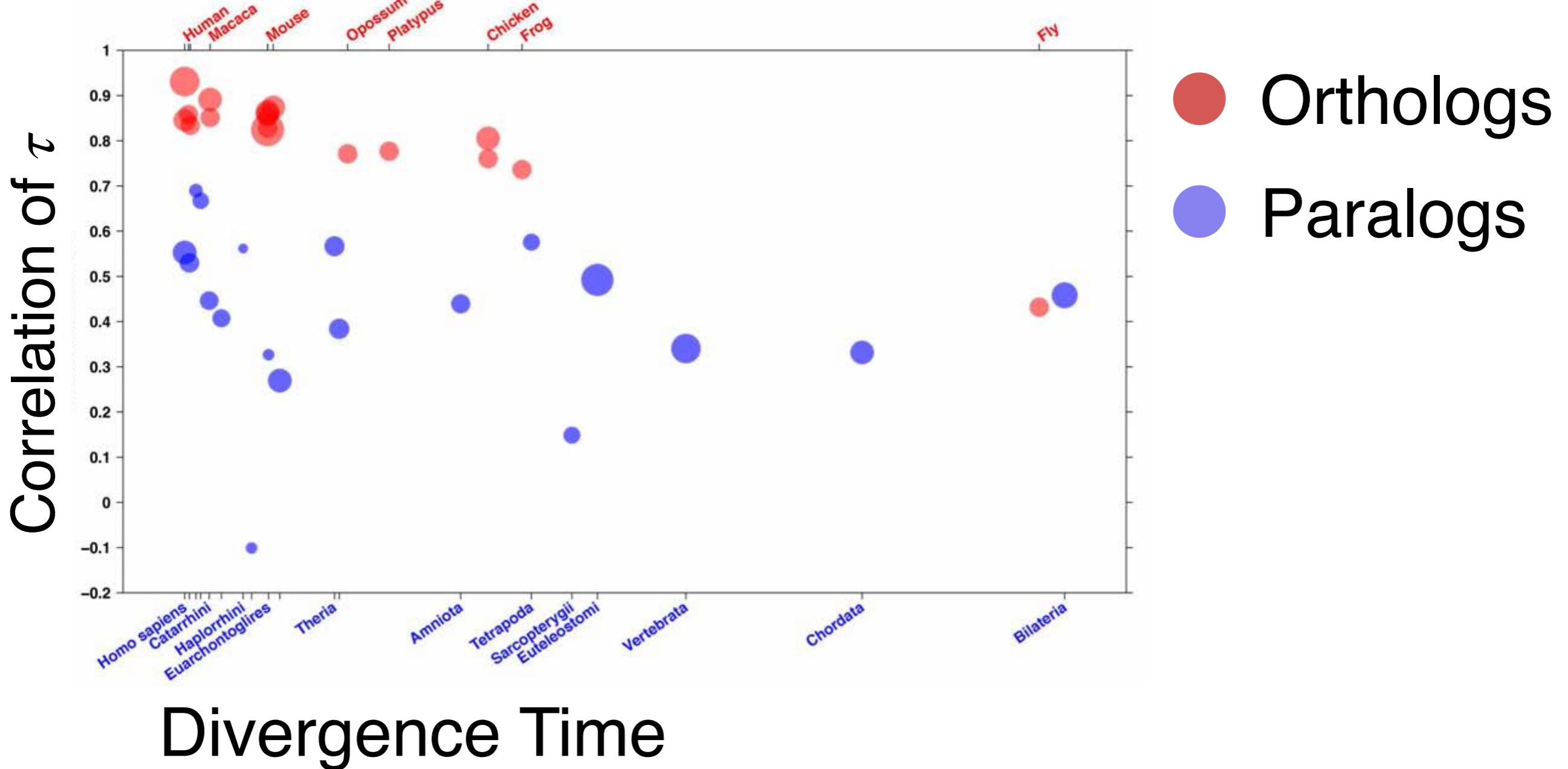
Tissue-Specificity of Gene Expression Diverges Slowly between Orthologs, and Rapidly between Paralogs

Nadezda Kryuchkova-Mostacci^{1,2}, Marc Robinson-Rechavi^{1,2*}

They used pairwise comparisons
across genes and species.

Tissue-Specificity of Gene Expression Diverges Slowly between Orthologs, and Rapidly between Paralogs

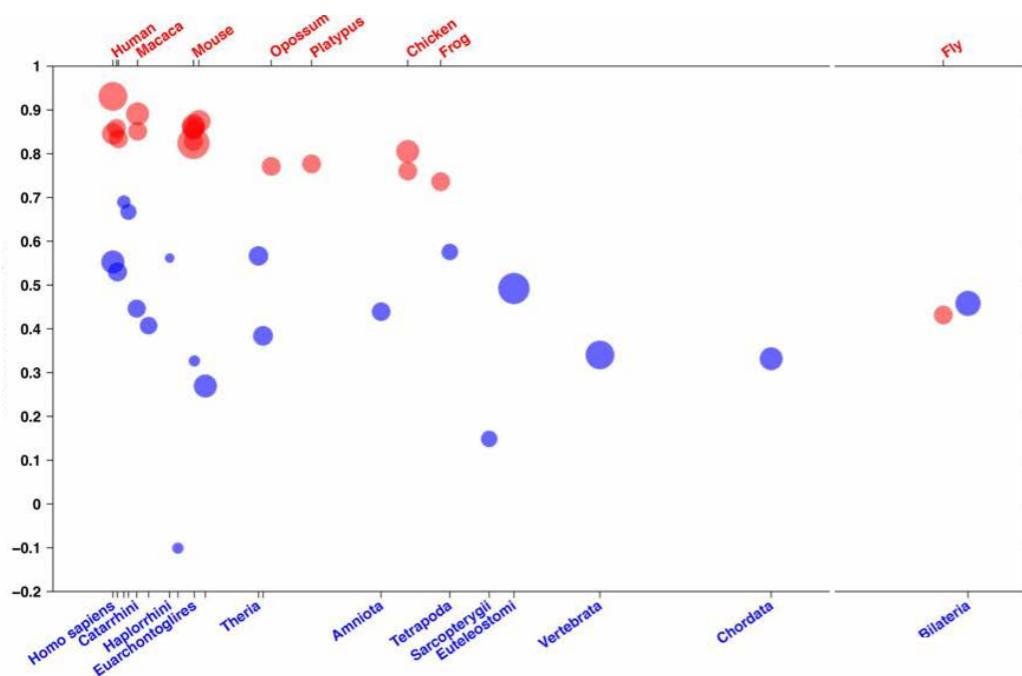
Nadezda Kryuchkova-Mostacci^{1,2}, Marc Robinson-Rechavi^{1,2*}



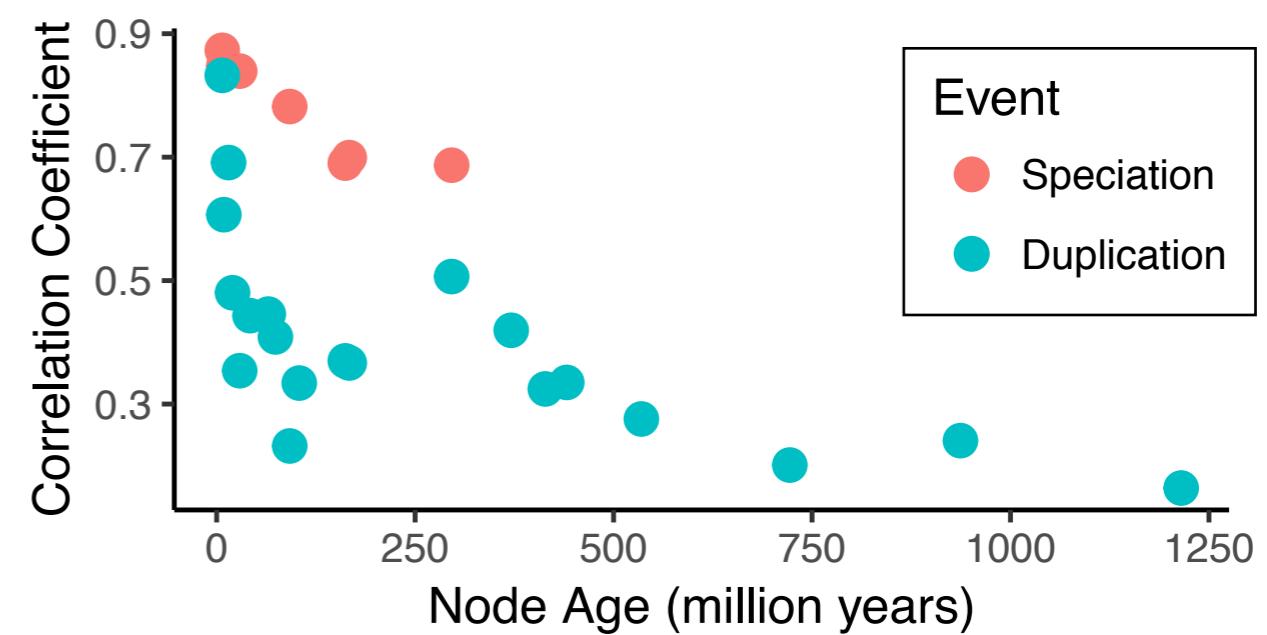
Divergence Time

Link to notes: <https://git.io/evol2017>

We first verified that we can reproduce their result.



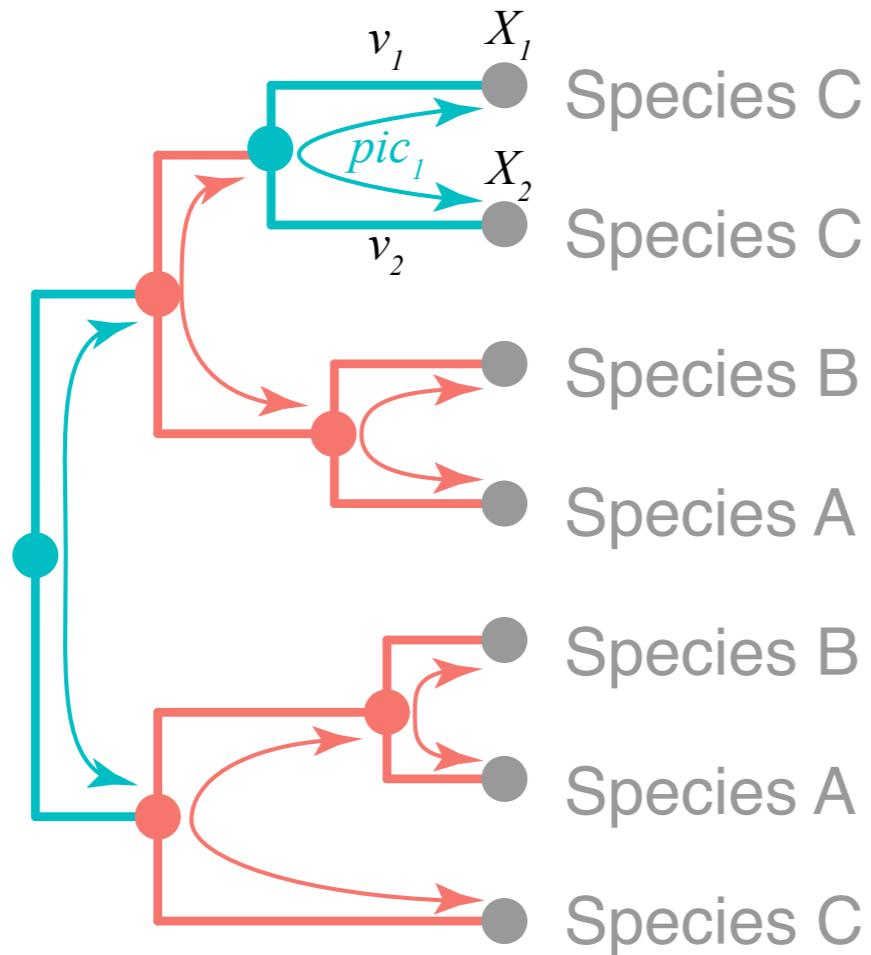
Their plot



Our plot

We reanalyzed the data with phylogenetic independent contrasts.

Phylogenetic Independent Contrasts



4 contrasts arising from speciation nodes

2 contrasts arising from duplication nodes

Felsenstein 1985

The value of the top contrast is:

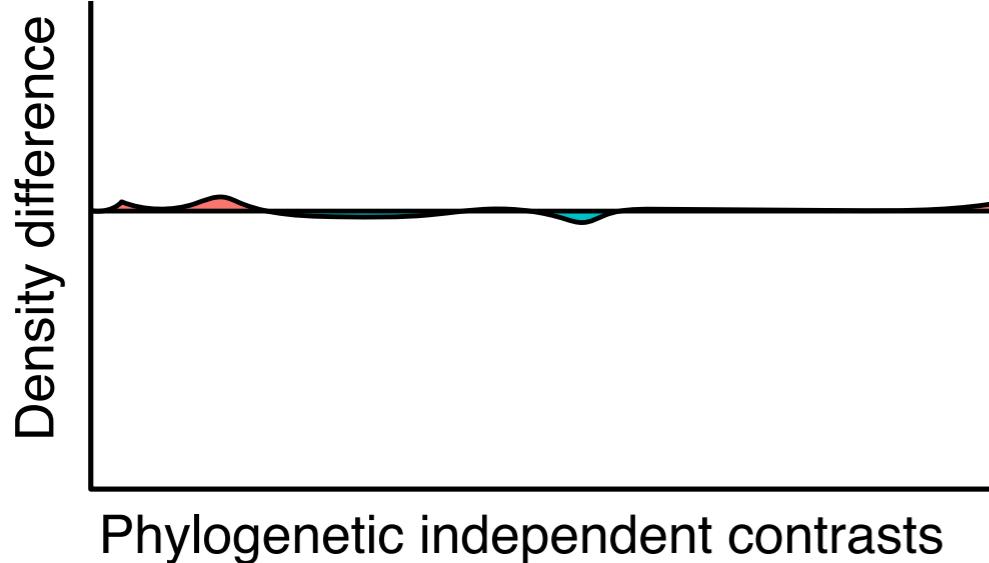
$$pic_1 = \frac{X_1 - X_2}{v_1 + v_2}$$

Observed trait change
(difference between descendant values)

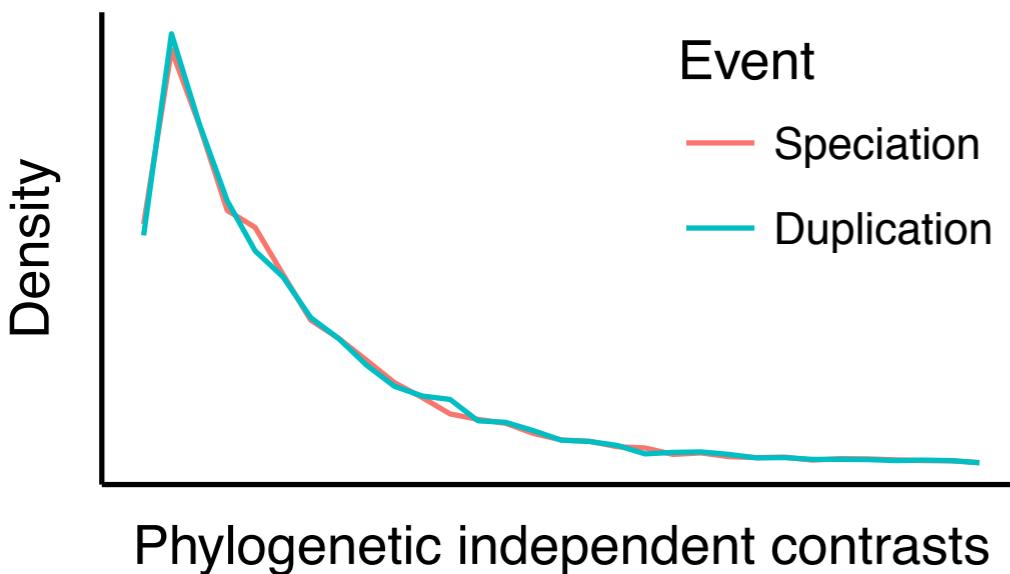
Expected variance
(total branch length)

The larger the contrast, the more change per unit branch length there has been.

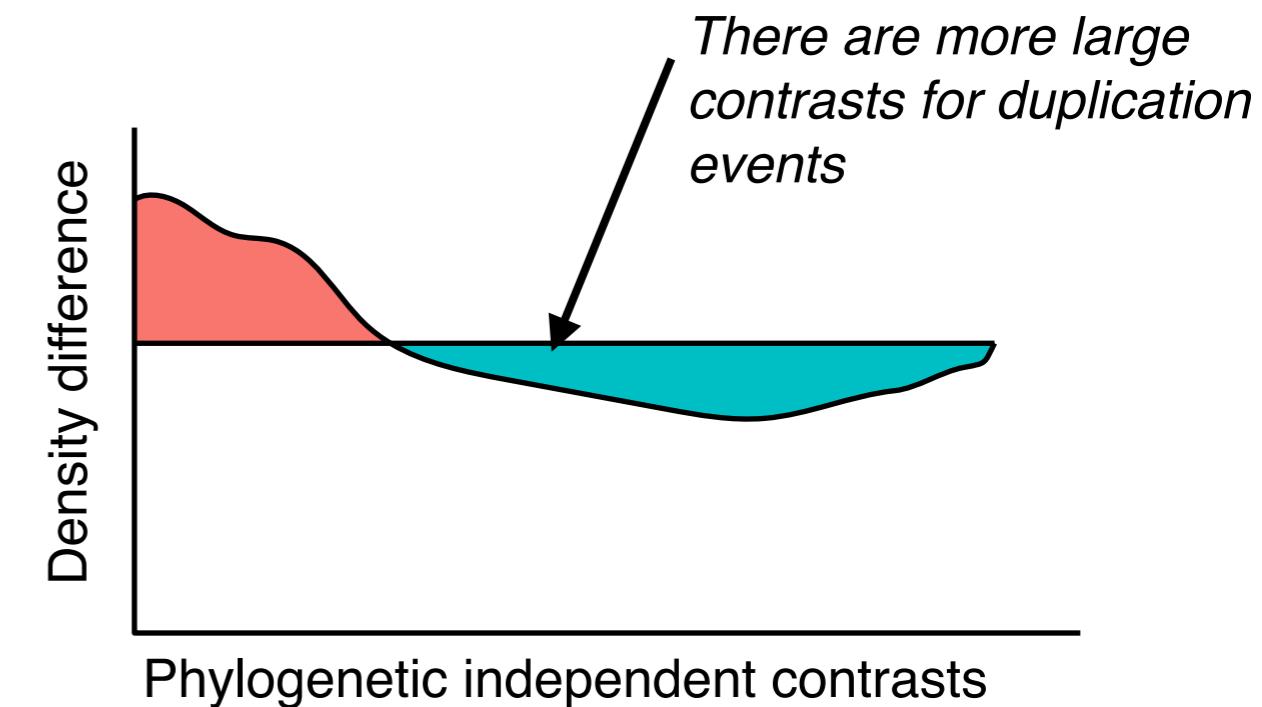
Prediction



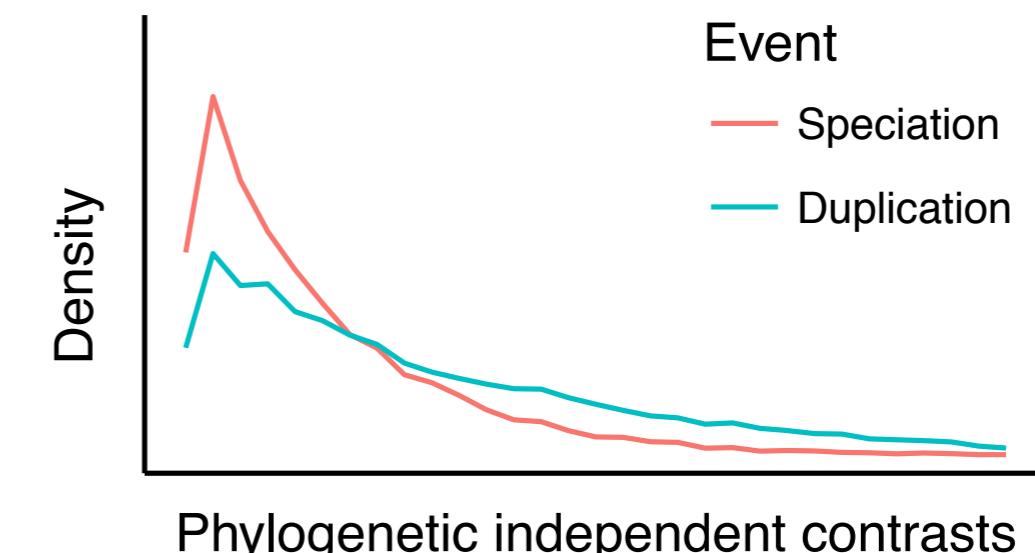
Null expectation



The distribution of contrasts for duplication events is **not** skewed toward larger values



Ortholog Conjecture expectation

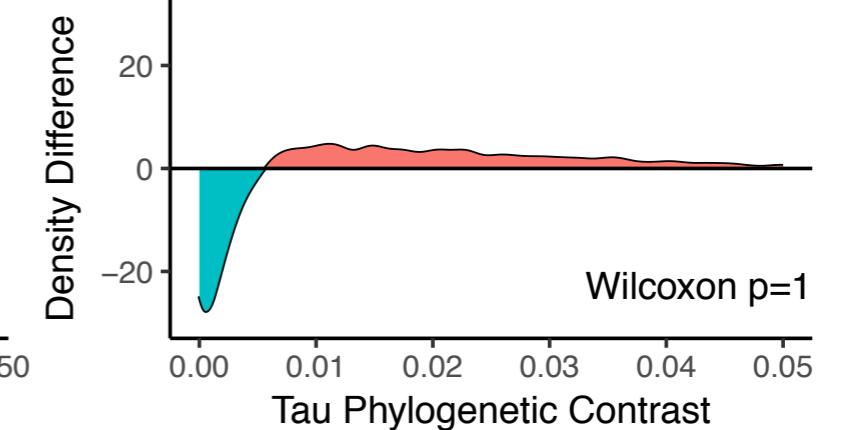
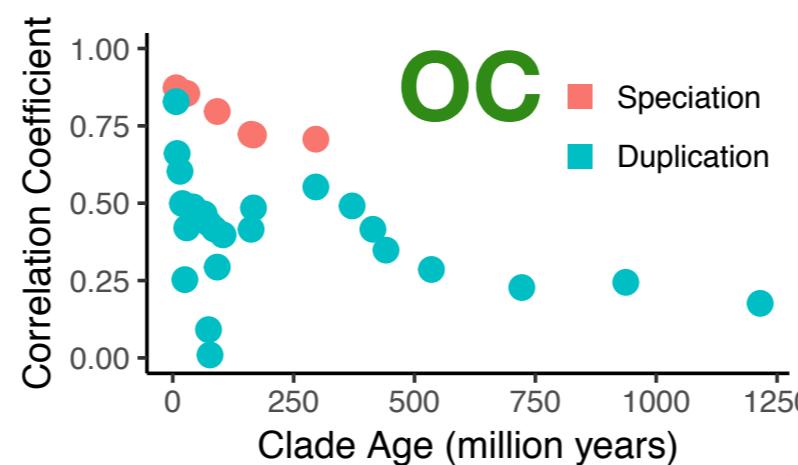


The distribution of contrasts for duplication events is skewed toward larger values

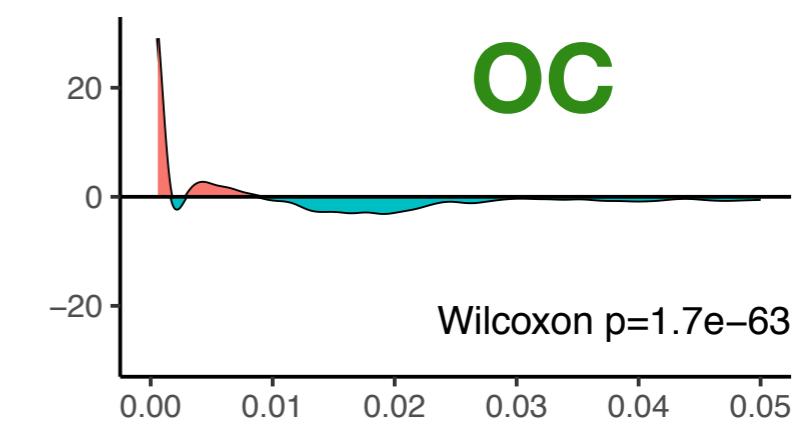
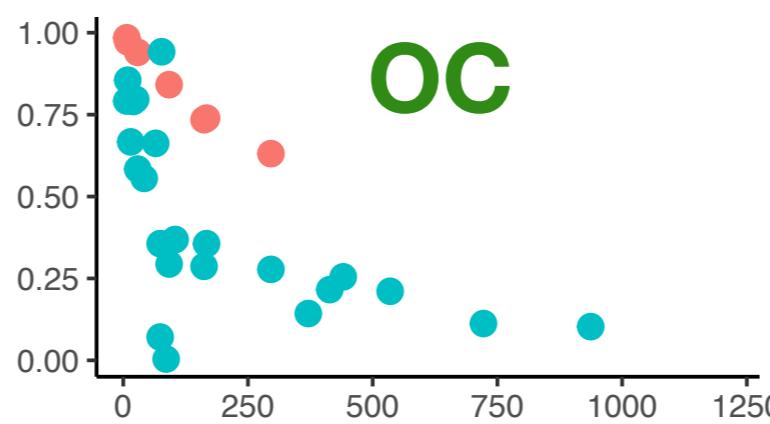
Empirical

Pairwise

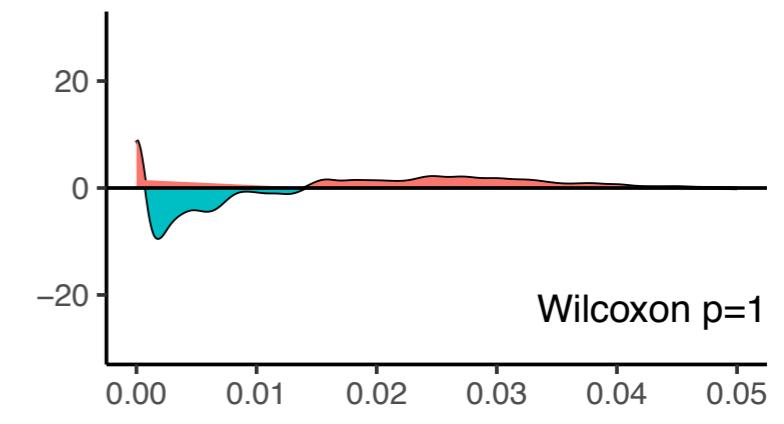
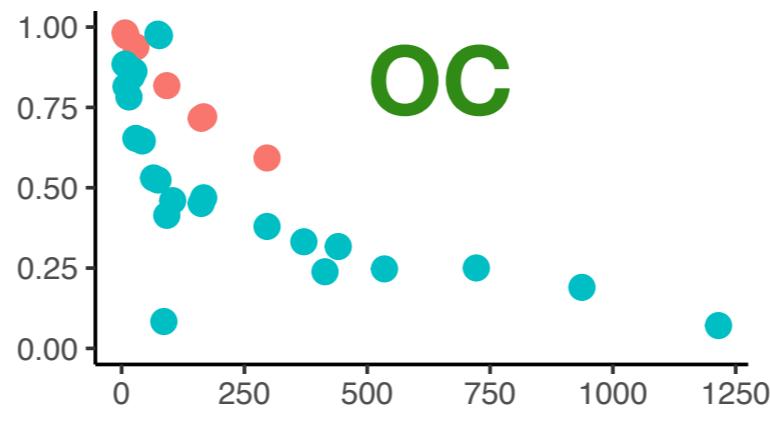
Phylogenetic



Simulation under
ortholog conjecture



Simulation under
null hypothesis



OC - Result consistent with Ortholog conjecture

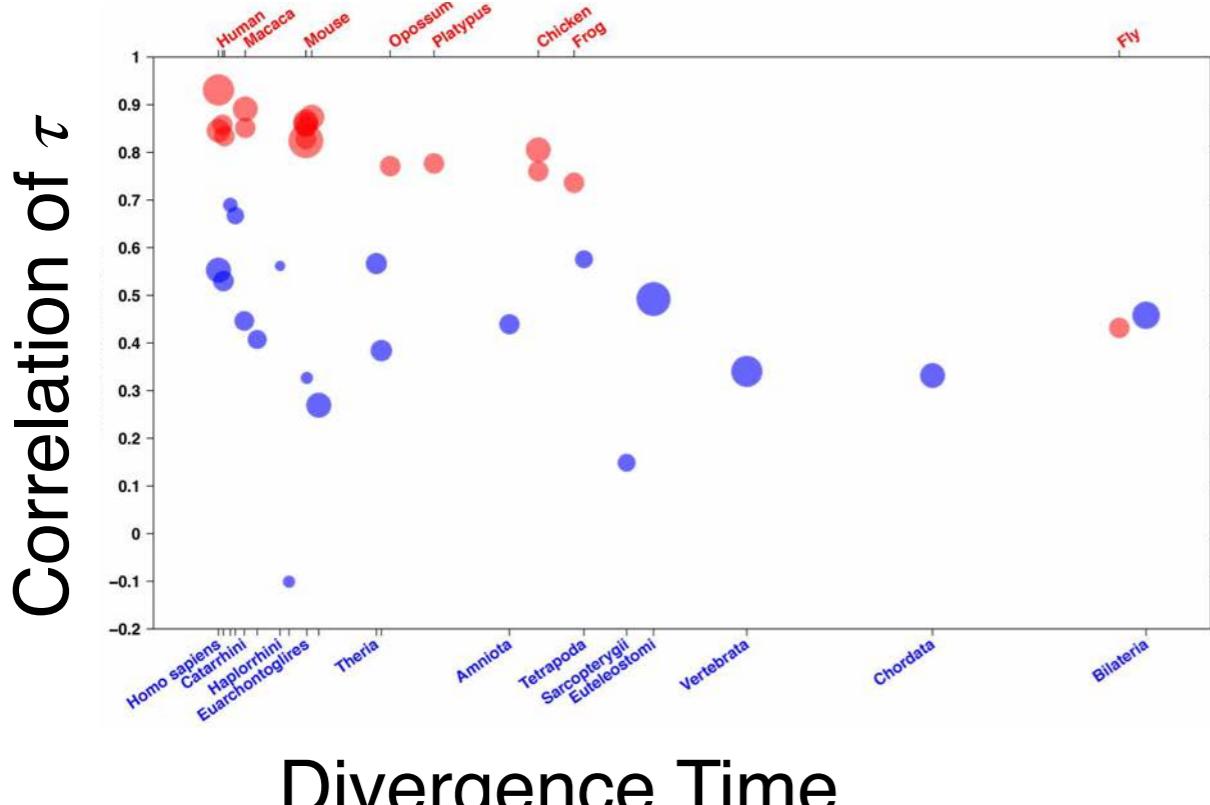
Phylogenetic methods provide distinct testable predictions under different hypotheses

The pairwise methods actually predict the same result under different hypotheses

What is this striking pattern?

Tissue-Specificity of Gene Expression Diverges Slowly between Orthologs, and Rapidly between Paralogs

Nadezda Kryuchkova-Mostacci^{1,2}, Marc Robinson-Rechavi^{1,2*}



Phil. Trans. R. Soc. Lond. B 326, 119–157 (1989) [119

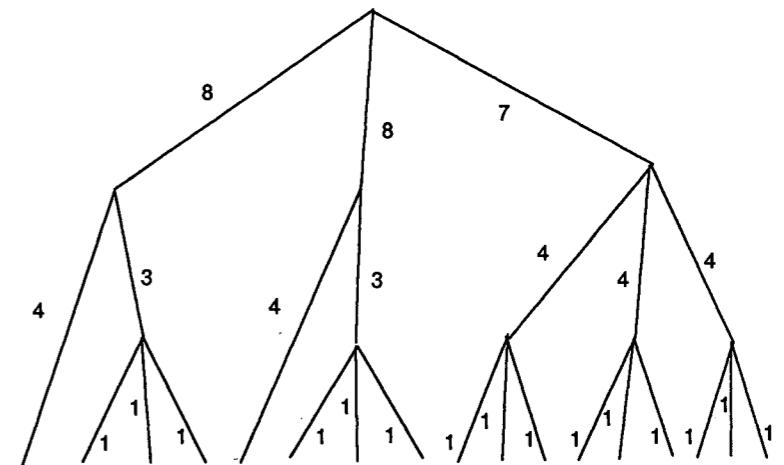
Printed in Great Britain

THE PHYLOGENETIC REGRESSION

BY A. GRAFENT†

*Animal Behaviour Research Group, Department of Zoology, University of Oxford, South Parks Road,
Oxford OX1 3PS, U.K.*

(Communicated by W. D. Hamilton, F.R.S. – Received 13 February 1989)



A phylogenetic tree describes the expected covariance structure of traits.

The pattern reflects the structure of the gene trees, not evolutionary processes on the tree.

Understanding how the evolution of gene function differs between **orthologs** and **paralogs** ~~is fundamental to understanding~~ ~~the~~ does not explain much about the association between genes and phenotypes.

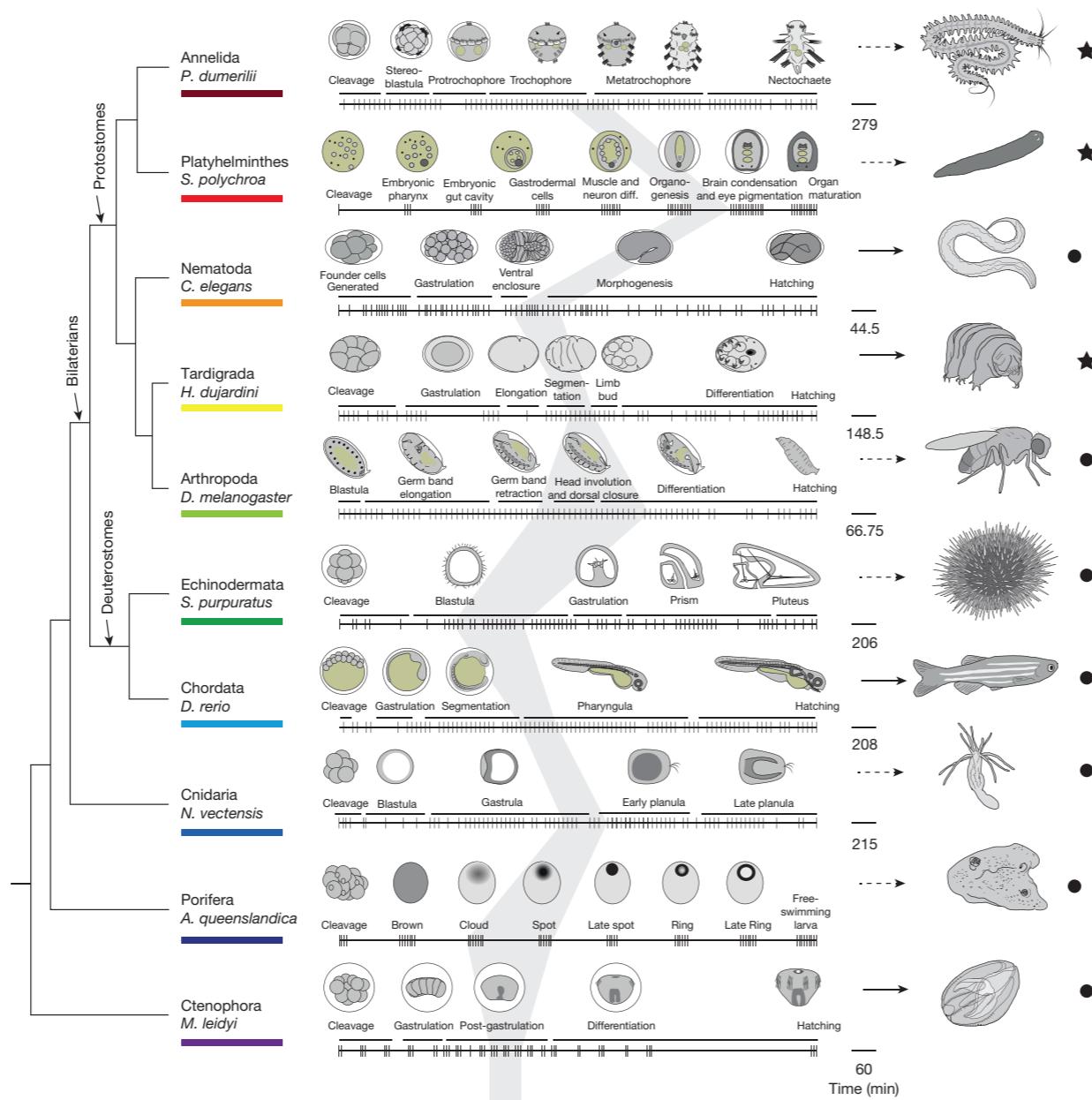
Our results indicate that phylogenetic relationships are very useful for understanding gene function, and information about orthology and paralogy provide little additional information.

For example...

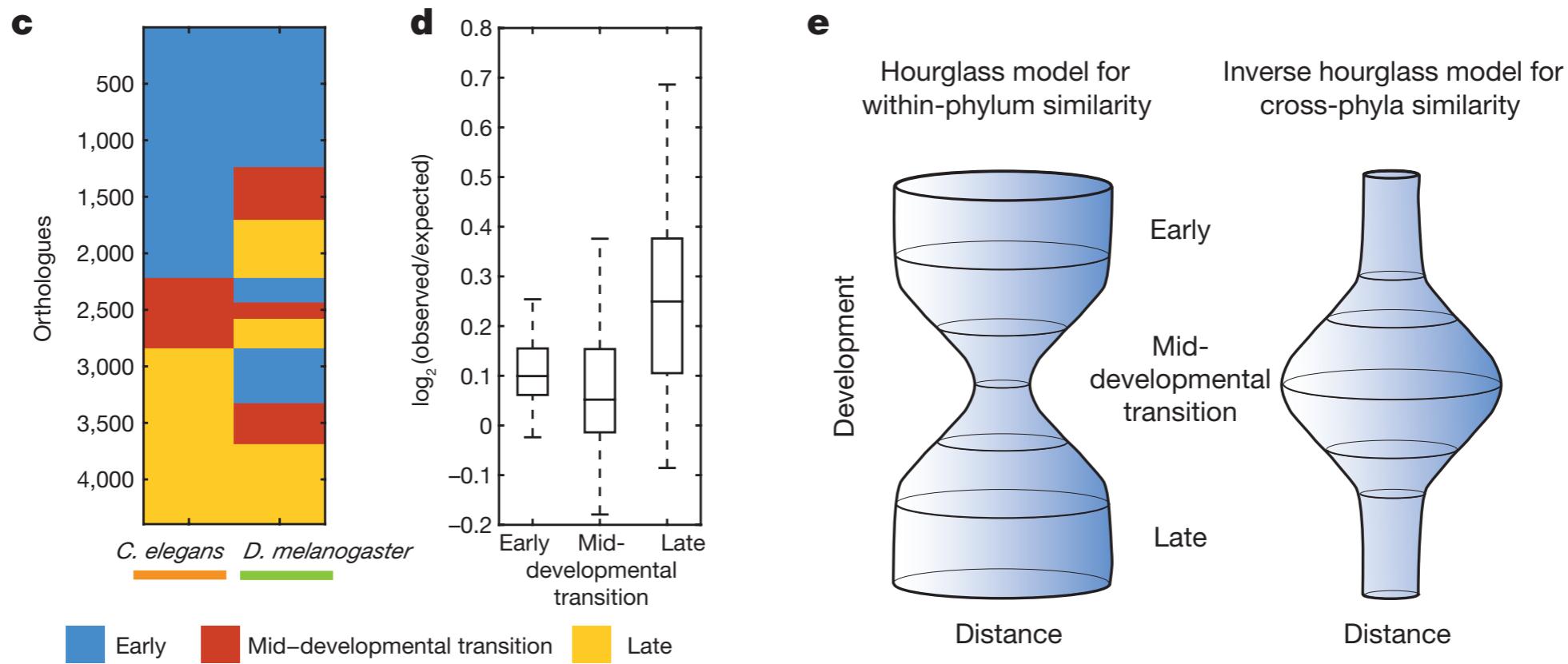
We should expect closely related paralogs to be more similar in function than distantly related orthologs.

The mid-developmental transition and the evolution of animal body plans

Michal Levin^{1†*}, Leon Anavy^{1*}, Alison G. Cole¹, Eitan Winter¹, Natalia Mostov¹, Sally Khair¹, Naftalie Senderovich¹, Ekaterina Kovalev¹, David H. Silver¹, Martin Feder¹, Selene L. Fernandez-Valverde^{2†}, Nagayasu Nakanishi^{2†}, David Simmons³, Oleg Simakov⁴, Tomas Larsson⁴, Shang-Yun Liu⁵, Ayelet Jerafi-Vider⁶, Karina Yaniv⁶, Joseph F. Ryan³, Mark Q. Martindale³, Jochen C. Rink⁵, Detlev Arendt⁴, Sandie M. Degnan², Bernard M. Degnan², Tamar Hashimshony¹ & Itai Yanai¹



Levin et al, 2016



“Our results are consistent with an inverse hourglass model for metazoan body plans (Fig. 4e) in which the molecular components that comprise early and late embryogenesis are more conserved”

“we propose that a phylum may be defined as a collection of species whose gene expression at the mid- developmental transition is both highly conserved among them, yet divergent relative to other species.”

Levin et al, 2016

The mid-developmental transition and the evolution of animal body plans

Michal Levin^{1†*}, Leon Anavy^{1*}, Alison G. Cole¹, Eitan Winter¹, Natalia Mostov¹, Sally Khair¹, Naftalie Senderovich¹, Ekaterina Kovalev¹, David H. Silver¹, Martin Feder¹, Selene L. Fernandez-Valverde^{2†}, Nagayasu Nakanishi^{2†}, David Simmons³, Oleg Simakov⁴, Tomas Larsson⁴, Shang-Yun Liu⁵, Ayelet Jerafi-Vider⁶, Karina Yaniv⁶, Joseph F. Ryan³, Mark Q. Martindale³, Jochen C. Rink⁵, Detlev Arendt⁴, Sandie M. Degnan², Bernard M. Degnan², Tamar Hashimshony¹ & Itai Yanai¹

There is a consistently defined mid-point transition in expression

Global patterns in the evolution of this transition provide biological justification for phyla

But...

Their project design can't even identify changes that are specific to phyla

Animal Evolution: Are Phyla Real?

Current Biology
Dispatches

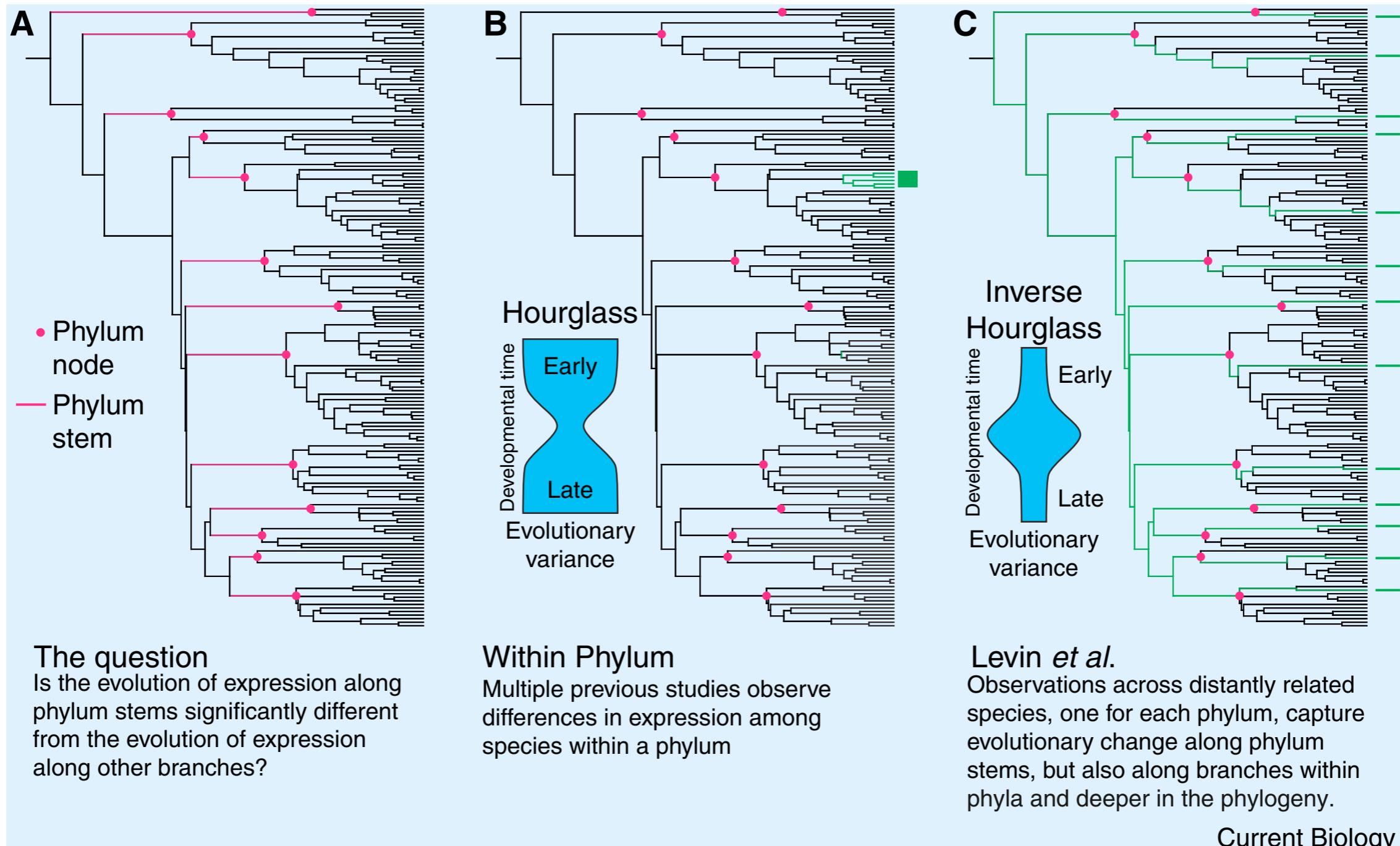
Andreas Hejnol¹ and Casey W. Dunn²

¹Sars International Centre for Marine Molecular Biology, University of Bergen, 5008 Bergen, Norway

²Department of Ecology and Evolutionary Biology, Brown University, Providence, RI 02912, USA

Correspondence: andreas.hejnol@uib.no (A.H.), casey_dunn@brown.edu (C.W.D.)

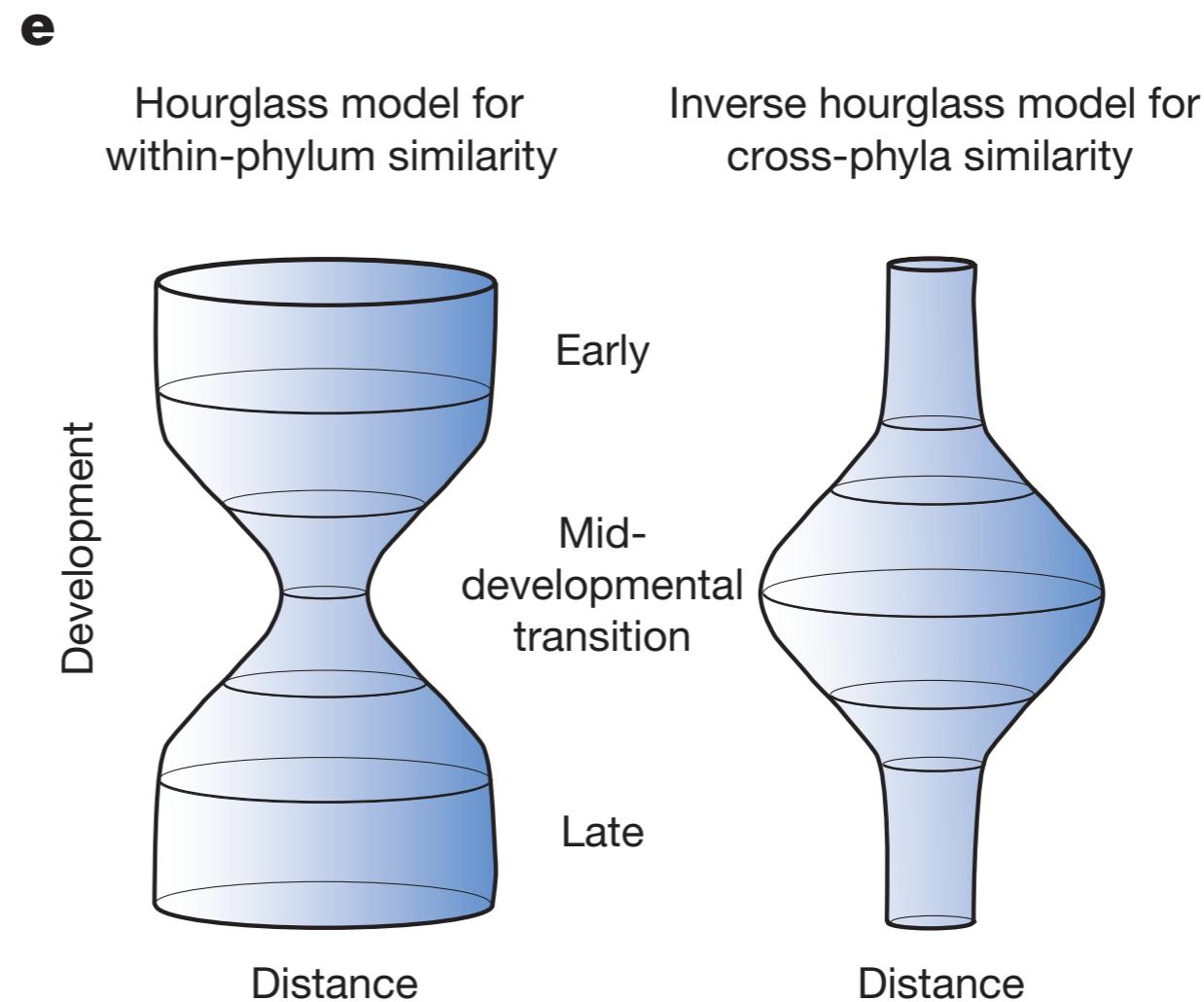
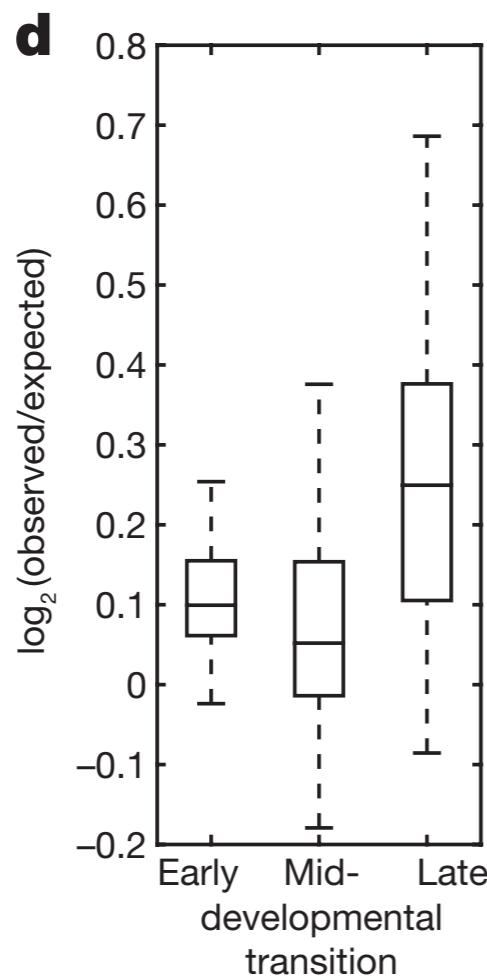
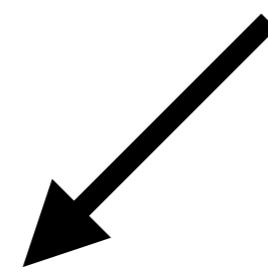
<http://dx.doi.org/10.1016/j.cub.2016.03.058>



**Even so, are their data consistent with
the inverse hourglass model?**

No. They didn't use trees, they used 45 pairwise comparisons between species that didn't take into account the relationships between species.

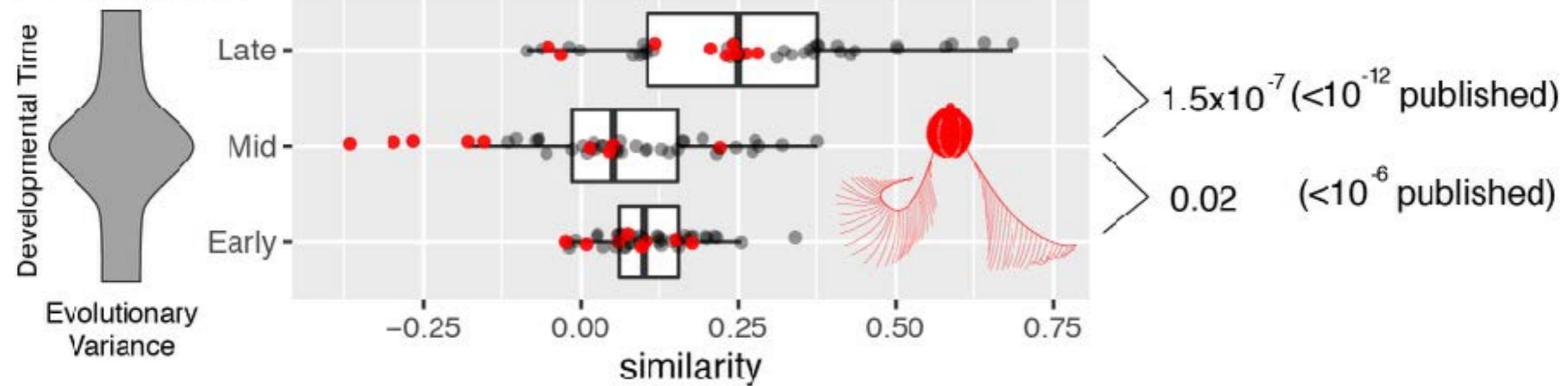
Plots of conservation based on 45 pairwise comparisons



Inverse Hourglass

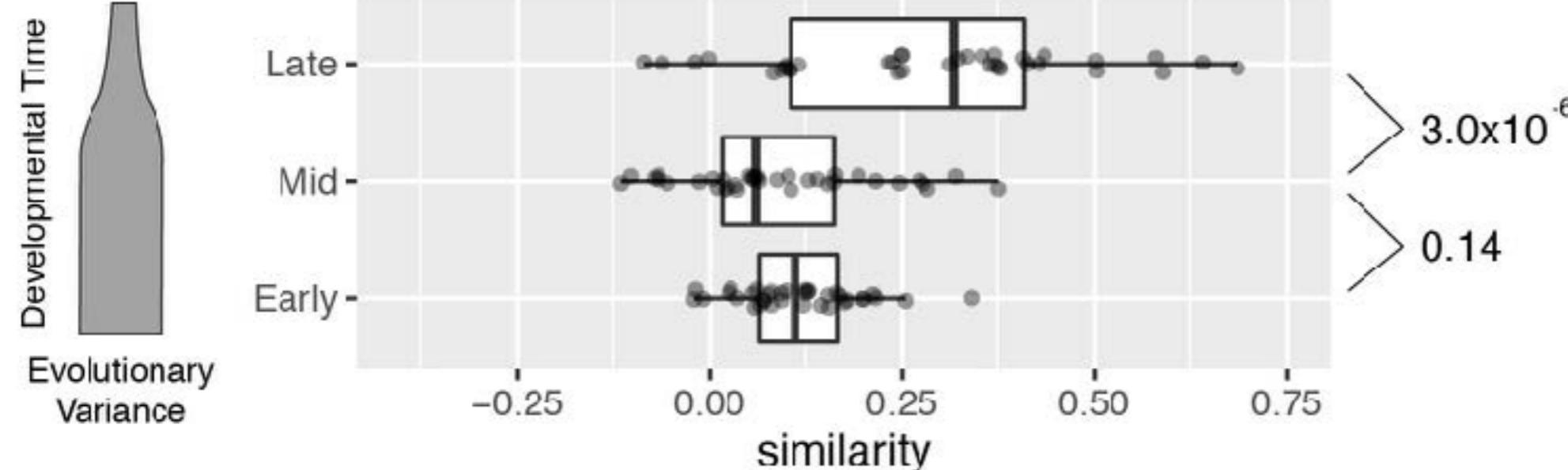
(a) Original, with ctenophore in red

p-values



Bottle

(b) New, without ctenophore



<https://rawgit.com/caseywdunn/levin2016/master/reanalyses.html>

https://github.com/caseywdunn/levin2016/blob/master/communication_arising.md

The “inverse hourglass” isn’t a global pattern.

It is the effect of a single taxon that was counted 9 times because pairwise comparisons that don’t account for phylogenetic structure were used.

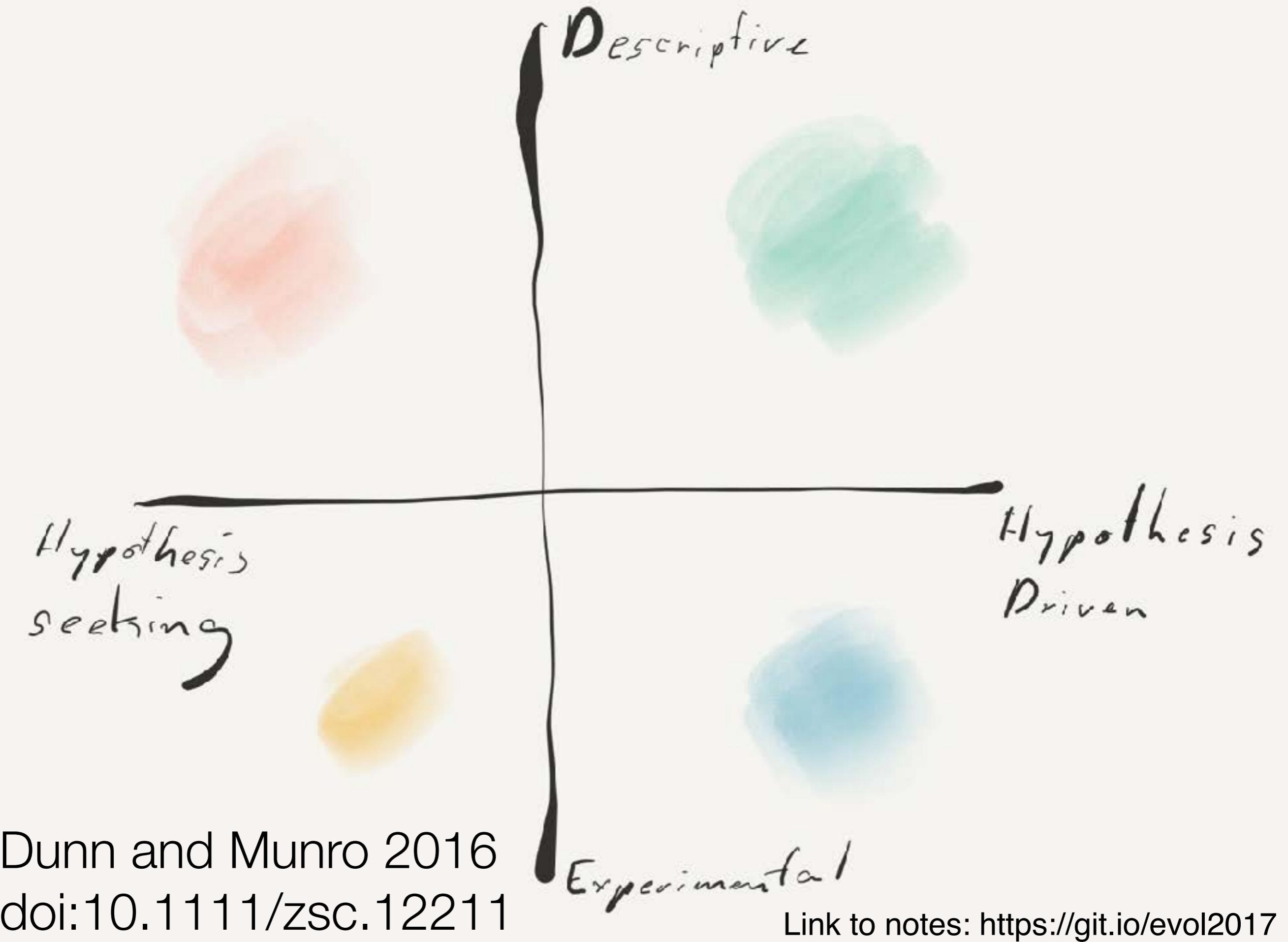
What are the limitations of phylogenetic comparative functional genomics?

“It is just descriptive,
not hypothesis driven”



Descriptive

Hypothesis
Driven



~~“It is just descriptive,
not hypothesis driven”~~

It is descriptive and
hypothesis driven.

“There won’t be enough statistical power to associate functional genomic changes with changes in other phenotypes of interest”

Some ways forward

- Embrace homoplasy
- Improve taxon sampling
- Combine with other evidence

Part 4:

General themes in

project design and

interpretation

There are many biases,
logical fallacies, and
statistical properties
that often lead to
misinterpretation of
analyses.

It is critical to be aware
of these to avoid these
problems yourselves
interpret the work of
others.

Flaws in reasoning are called logical fallacies.

Most fall into a few patterns - get to know them.

<https://yourlogicalfallacyis.com/>



strawman

Misrepresenting someone's argument to make it easier to attack.

After Will said that we should put more money into health and education, Warren responded by saying that he was surprised that Will hates our country so much that he wants to leave it defenceless by cutting military spending.



slippery slope

Asserting that if we allow A to happen, then Z will consequently happen too, therefore A should not happen.

Colin Closet asserts that if we allow same-sex couples to marry, then the next thing we know we'll be allowing people to marry their parents, their cars and even monkeys.



special pleading

Moving the goalposts to create exceptions when a claim is shown to be false.

Edward Johns claimed to be psychic, but when his 'abilities' were tested under proper scientific conditions, they magically disappeared. Edward explained this saying that one had to have faith in his abilities for them to work.



the gambler's fallacy

Believing that 'runs' occur to statistically independent phenomena such as roulette wheel spins.

Red had come up six times in a row on the roulette wheel, so Greg knew that it was close to certain that black would be next up. Suffering an economic form of natural selection with this thinking, he soon lost all of his savings.



black-or-white

Where two alternative states are presented as the only possibilities, when in fact more possibilities exist.

Whilst rallying support for his plan to fundamentally undermine citizens' rights, the Supreme Leader told the people they were either on his side, or on the side of the enemy.



false cause

Presuming that a real or perceived relationship between things means that one is the cause of the other.

Pointing to a fancy chart, Roger shows how temperatures have been rising over the past few centuries, whilst at the same time the numbers of pirates have been decreasing; thus pirates cool the world and global warming is a hoax.



ad hominem

Attacking your opponent's character or personal traits in an attempt to undermine their argument.

After Sally presents an eloquent and compelling case for a more equitable taxation system, Sam asks the audience whether we should believe anything from a woman who isn't married, was once arrested, and smells a bit weird.



loaded question

Asking a question that has an assumption built into it so that it can't be answered without appearing guilty.

Grace and Helen were both romantically interested in Brad. One day, with Brad sitting within earshot, Grace asked in an inquisitive tone whether Helen was having any problems with a fungal infection.



appeal to emotion

Manipulating an emotional response in place of a valid or compelling argument.

Luke didn't want to eat his sheep's brains with chopped liver and brussels sprouts, but his father told him to think about the poor, starving children in a third world country who weren't fortunate enough to have any food at all.



tu quoque

Avoiding having to engage with criticism by turning it back on the accuser - answering criticism with criticism.

The blue candidate accused the red candidate of committing the tu quoque fallacy. The red candidate responded by accusing the blue candidate of the same, after which ensued an hour of back and forth criticism with not much progress.



burden of proof

Saying that the burden of proof lies not with the person making the claim, but with someone else to disprove.

Bertrand declares that a teapot is, at this very moment, in orbit around the Sun between the Earth and Mars, and that because no one can prove him wrong his claim is therefore a valid one.



the fallacy fallacy

Presuming that because a claim has been poorly argued, or a fallacy has been made, that it is necessarily wrong.

Recognising that Amanda had committed a fallacy in arguing that we should eat healthy food because a nutritionist said it was popular, Alyse said we should therefore eat bacon double cheeseburgers every day.



personal incredulity

Saying that because one finds something difficult to understand that it's therefore not true.

Kirk drew a picture of a fish and a human and with effusive disdain asked Richard if he really thought we were stupid enough to believe that a fish somehow turned into a human through just, like, random things happening over time.



ambiguity

Using double meanings or ambiguities of language to mislead or misrepresent the truth.

When the judge asked the defendant why he hadn't paid his parking fines, he said that he shouldn't have to pay them because the sign said 'Fine for parking here' and so he naturally presumed that it would be fine to park there.



genetic

Judging something good or bad on the basis of where it comes from, or from whom it comes.

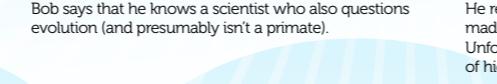
Accused on the 6 o'clock news of corruption and taking bribes, the senator said that we should all be very wary of the things we hear in the media, because we all know how very unreliable the media can be.



begging the question

A circular argument in which the conclusion is included in the premise.

The word of Zorbo the Great is flawless and perfect. We know this because it says so in The Great and Infallible Book of Zorbo's Best and Most Truest Things that are Definitely True and Should Not Ever Be Questioned.



appeal to authority

Using the opinion or position of an authority figure, or institution of authority, in place of an actual argument.

Not able to defend his position that evolution 'isn't true' Bob says that he knows a scientist who also questions evolution (and presumably isn't a primate).



appeal to nature

Making the argument that because something is 'natural' it is therefore valid, justified, inevitable, good, or ideal.

The medicine man rolled into town on his bandwagon offering various natural remedies, such as very special plain water. He said that it was only natural that people should be wary of 'artificial' medicines such as antibiotics.



anecdotal

Using personal experience or an isolated example instead of a valid argument, especially to dismiss statistics.

Jason said that was all cool and everything, but his grandfather smoked, like, 30 cigarettes a day and lived until 97 - so don't believe everything you read about meta analyses of sound studies showing proven causal relationships.



the texas sharpshooter

Cherry-picking data clusters to suit an argument, or finding a pattern to fit a presumption.

The makers of Sugarette Candy Drinks point to research showing that of the five countries where Sugarette drinks sell the most units, three of them are in the top ten healthiest countries on Earth, therefore Sugarette drinks are healthy.



middle ground

Saying that a compromise, or middle point, between two extremes is the truth.

Holly said that vaccinations caused autism in children, but her scientifically well-read friend Caleb said that this claim had been debunked and proven false. Their friend Alice offered a compromise that vaccinations cause some autism.

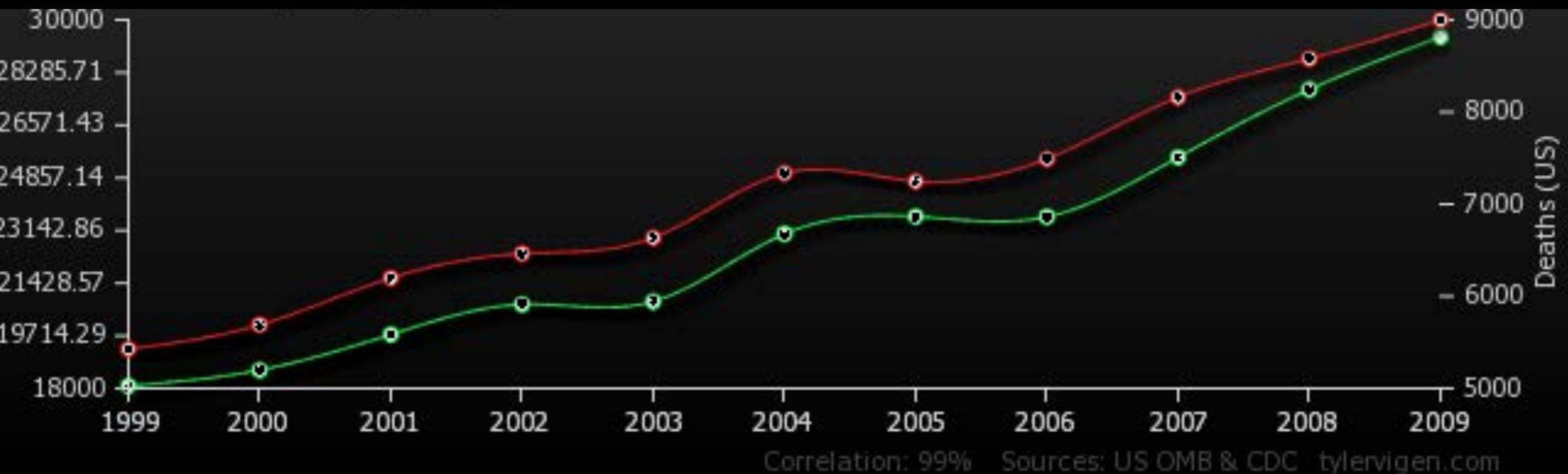
Texas sharp shooter
fallacy.



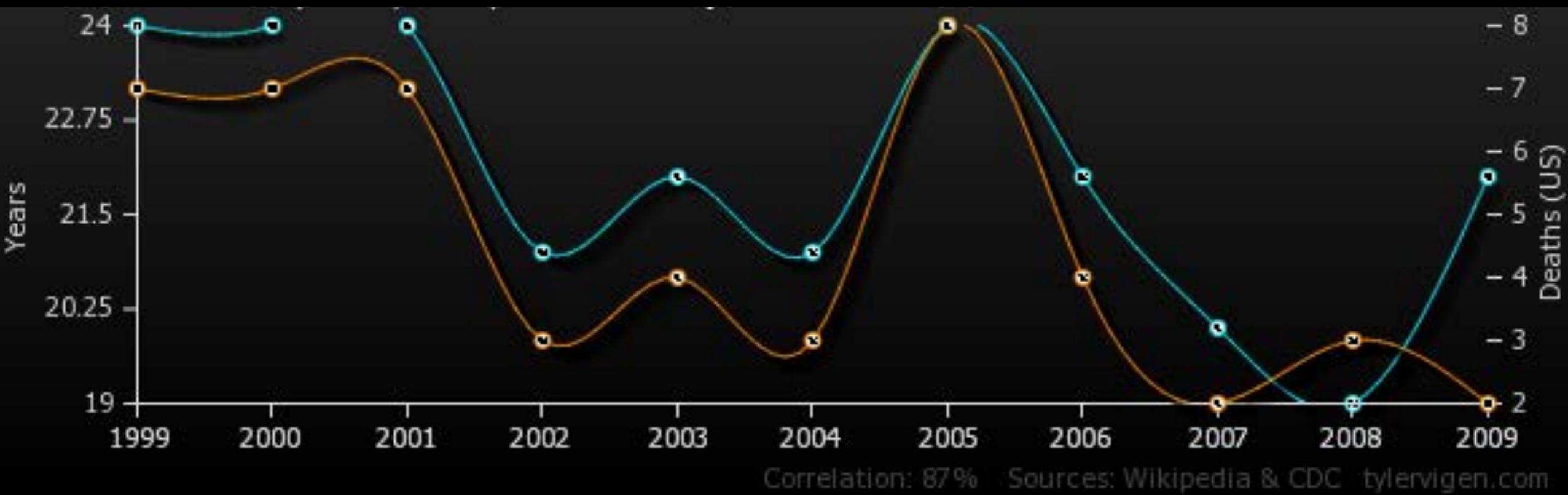
Texas sharp shooter fallacy often reads something like:

“We analyzed expression of 20000 genes through development and were surprised to find a cluster of genes with highly correlated expression.”

But if you are making a small number of observations (eg measuring expression at 8 time points) on a large number of variables (eg gene expression) it is *expected* that there will be many correlations by chance.



- US spending on science, space, and technology
- Suicides by hanging, strangulation, and suffocation



■ Age of Miss America

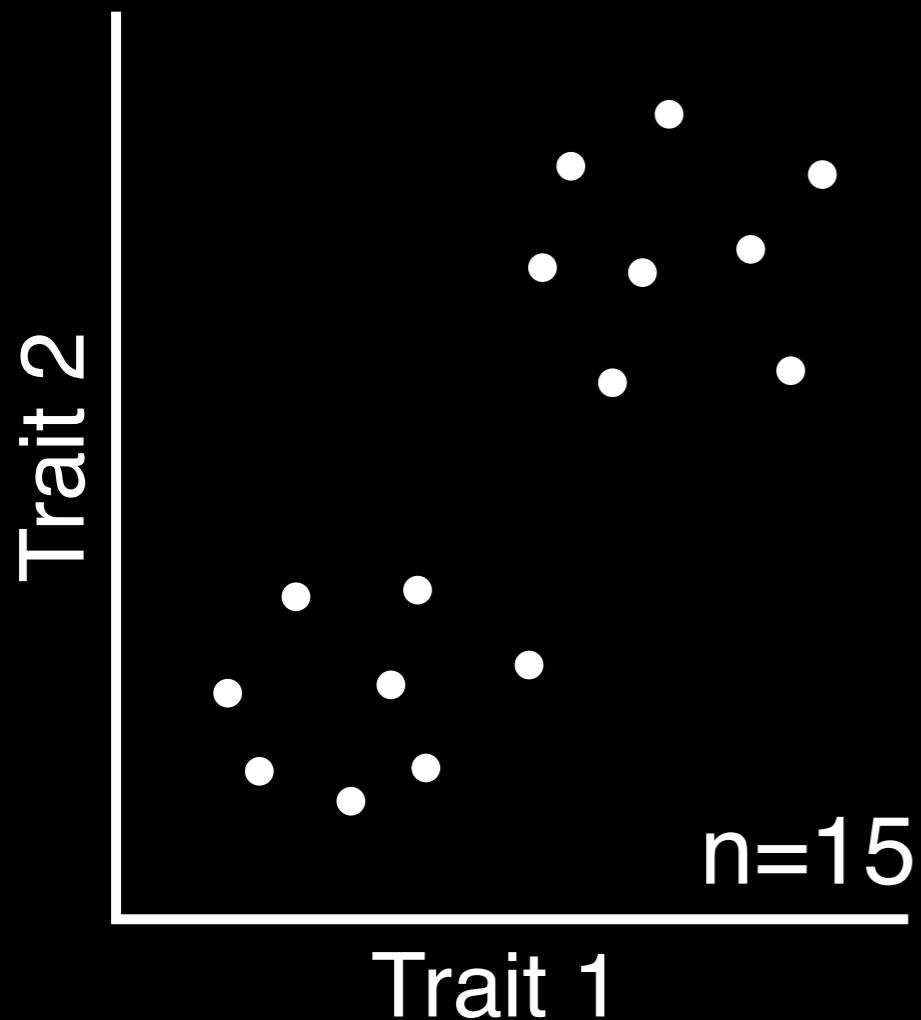
■ Murders by steam, hot vapors, and hot objects

Pseudo-replication.

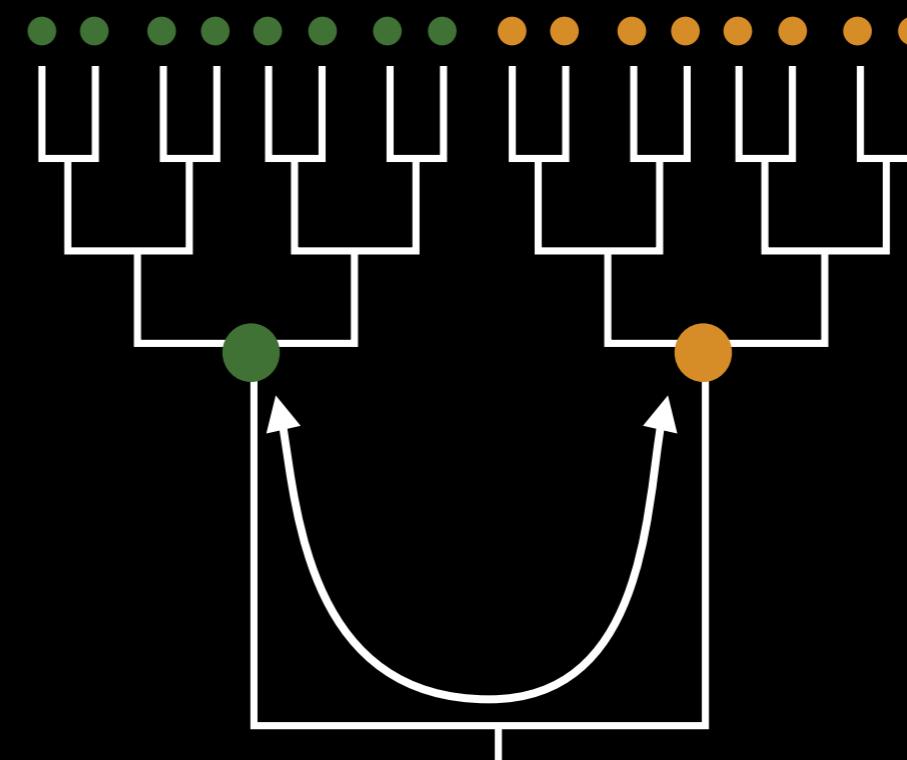
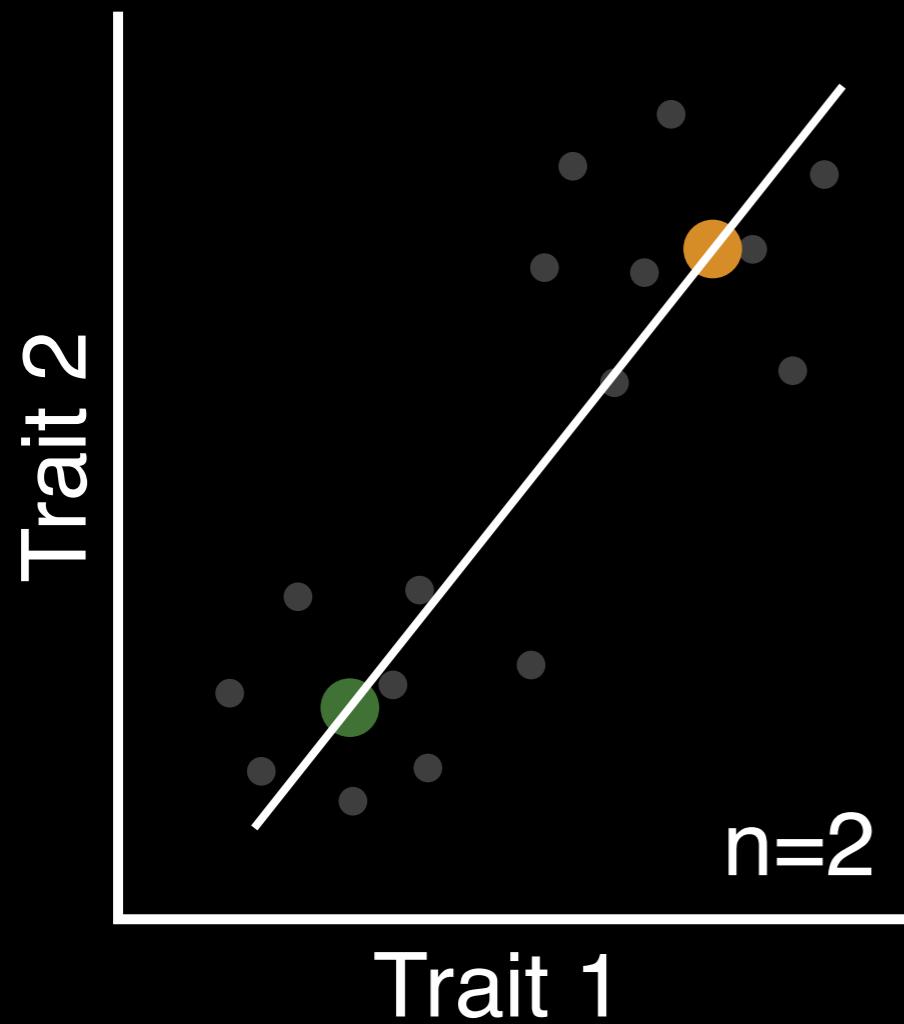
Data are treated as
independent observations
when they are not.

Observations across species.

Are these traits correlated?



Why use phylogenies to analyze data across species?



Pseudo-replication.

Leads to overestimating the degrees of freedom, which always makes results seem more significant.

Always clearly define your **null hypothesis**.

Always ask what the data would look like under the **null hypothesis**. If the prediction is no different than the predictions under other hypotheses, the study can't be used to test your hypothesis.

Hypothesis:

Evolution has selected for an increase in complexity in animals.

Prediction:

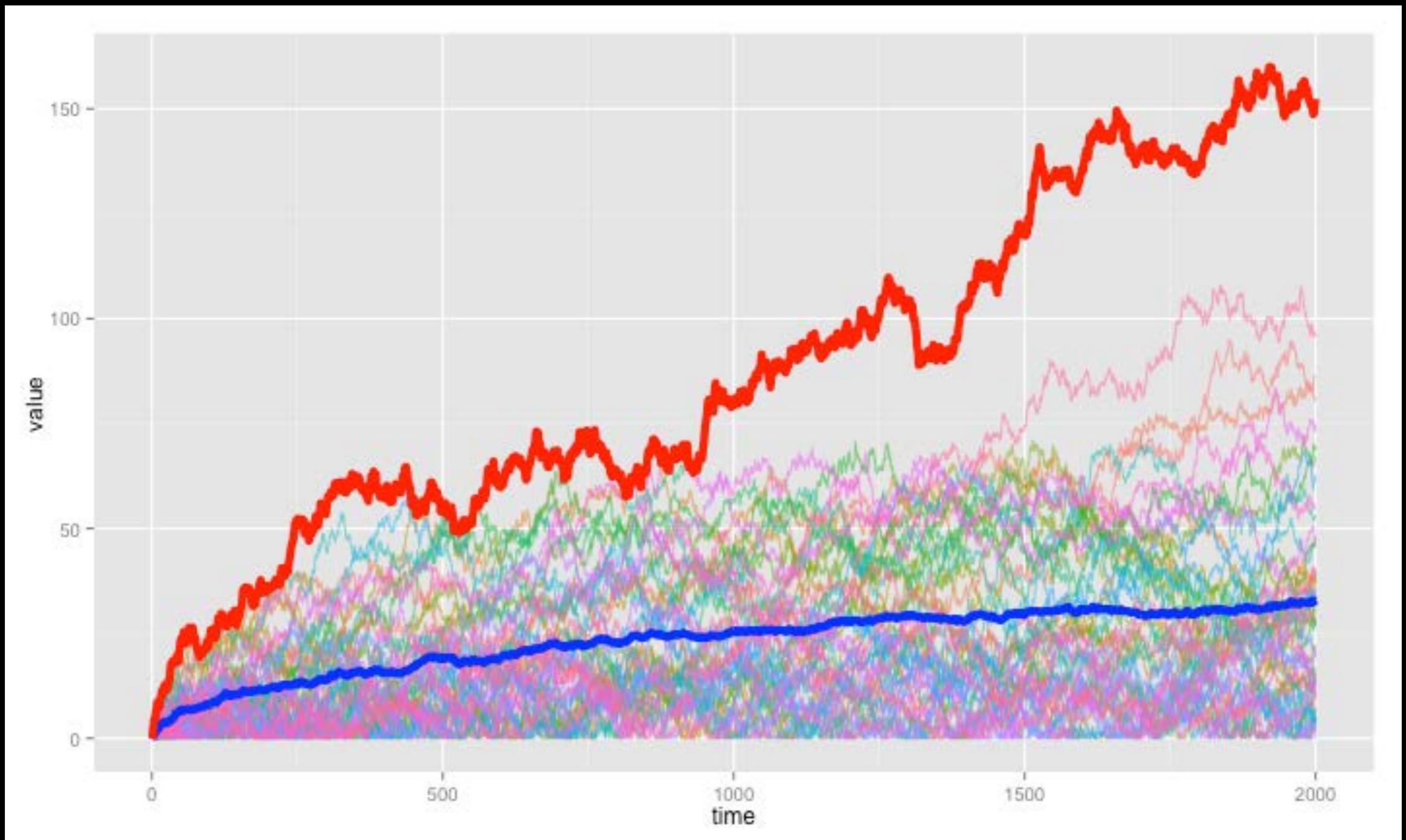
Complexity increases with time in animals.

Null hypothesis:

Evolution has not selected for an increase in complexity in animals.

Prediction:

Complexity increases with time in animals.



http://htmlpreview.github.io/?https://github.com/caseywdunn/random_walk/blob/master/randomwalk.html

Ascertainment biases

There is a preference for collecting some data more than others, and this is mistaken for a pattern in nature.

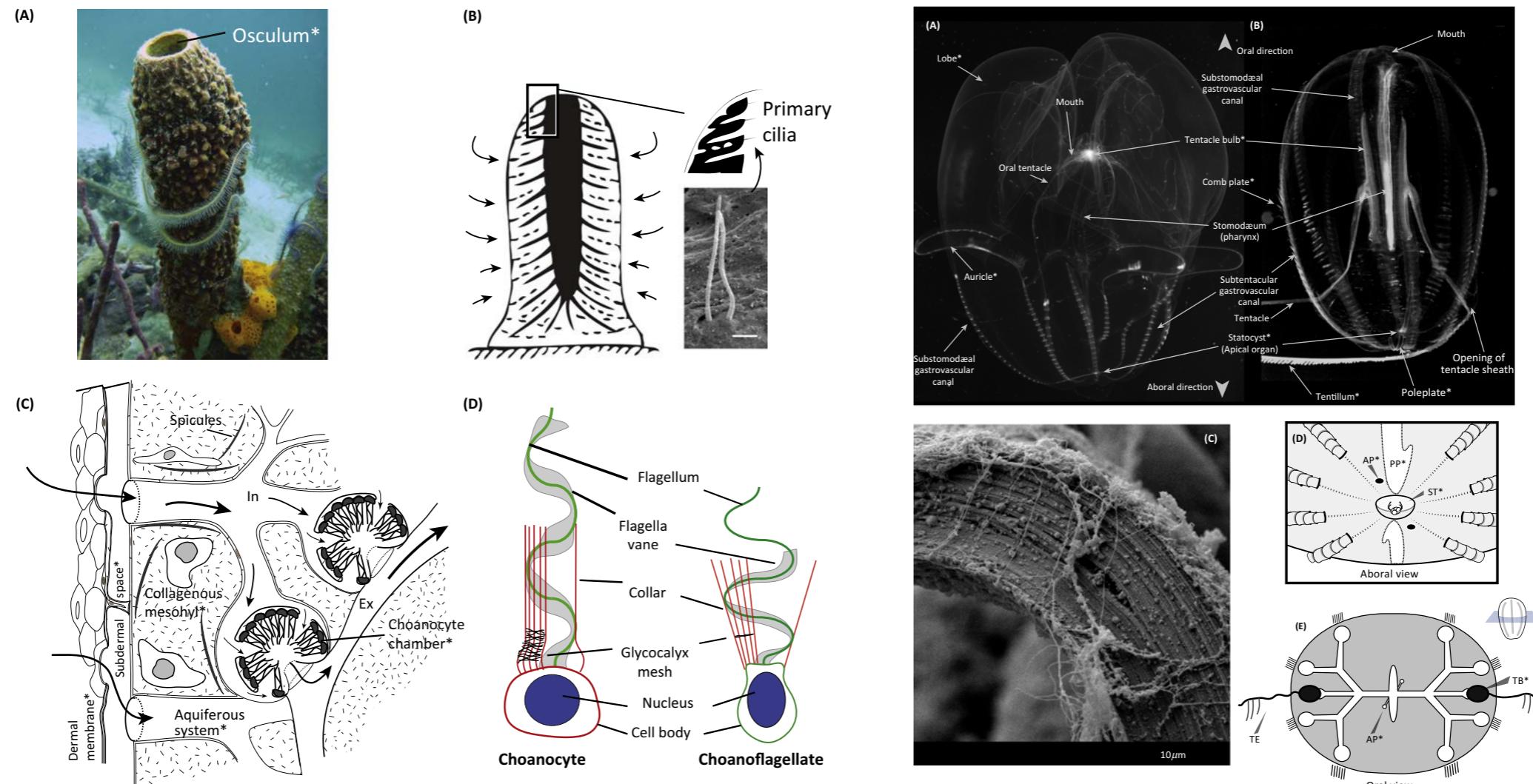
The hidden biology of sponges and ctenophores

Casey W. Dunn¹, Sally P. Leys², and Steven H.D. Haddock³

¹ Department of Ecology and Evolutionary Biology, Brown University, 80 Waterman St, Providence, RI 02906, USA

² Department of Biological Sciences, University of Alberta, Edmonton, AB, T6G 2E9, Canada

³ Monterey Bay Aquarium Research Institute, 7700 Sandholdt Rd, Moss Landing, CA 95039, USA



We have deep
biases in how we
view diversity

Consider:

Ctenophore



Monkey



“lower”

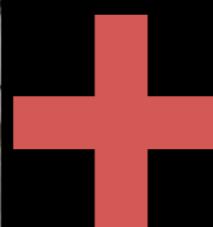
“higher”

Tendency to think of “lower”
animals as the basic model



And “higher” animals as the deluxe package







Ctenophore

Vertebrae

Monkey

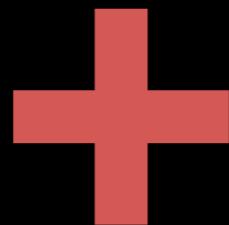
Monkeys have
many traits that
ctenophores don't.

Vertebrae.

Brain.

Ctenophores have
many traits that
monkeys don't.

Colloblasts (Glue cells).
Bioluminescence.



Livingstone, G. RICORDAC



Monkey

Glue Cell

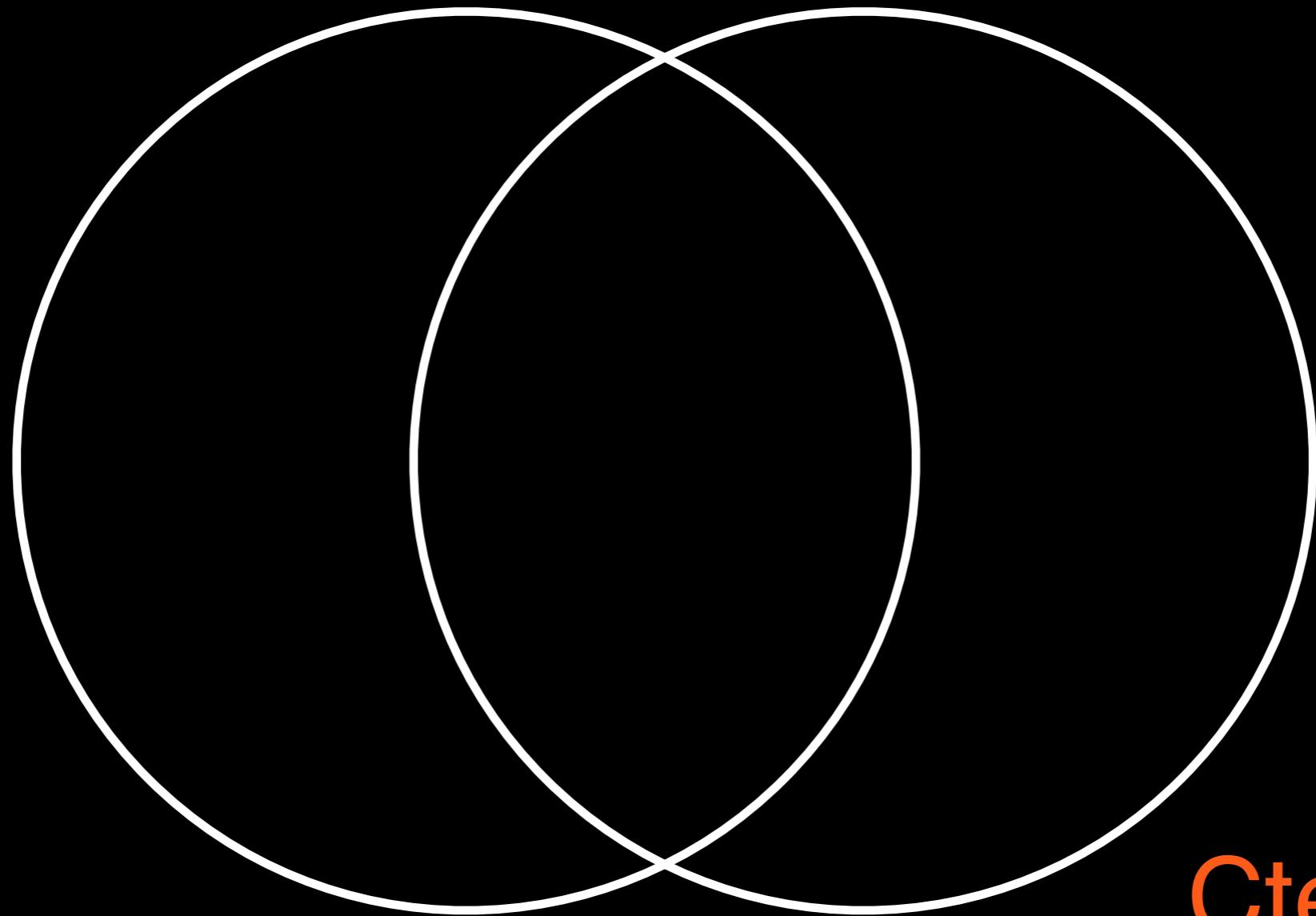
Ctenophore

94/95



Primate

Ctenophore



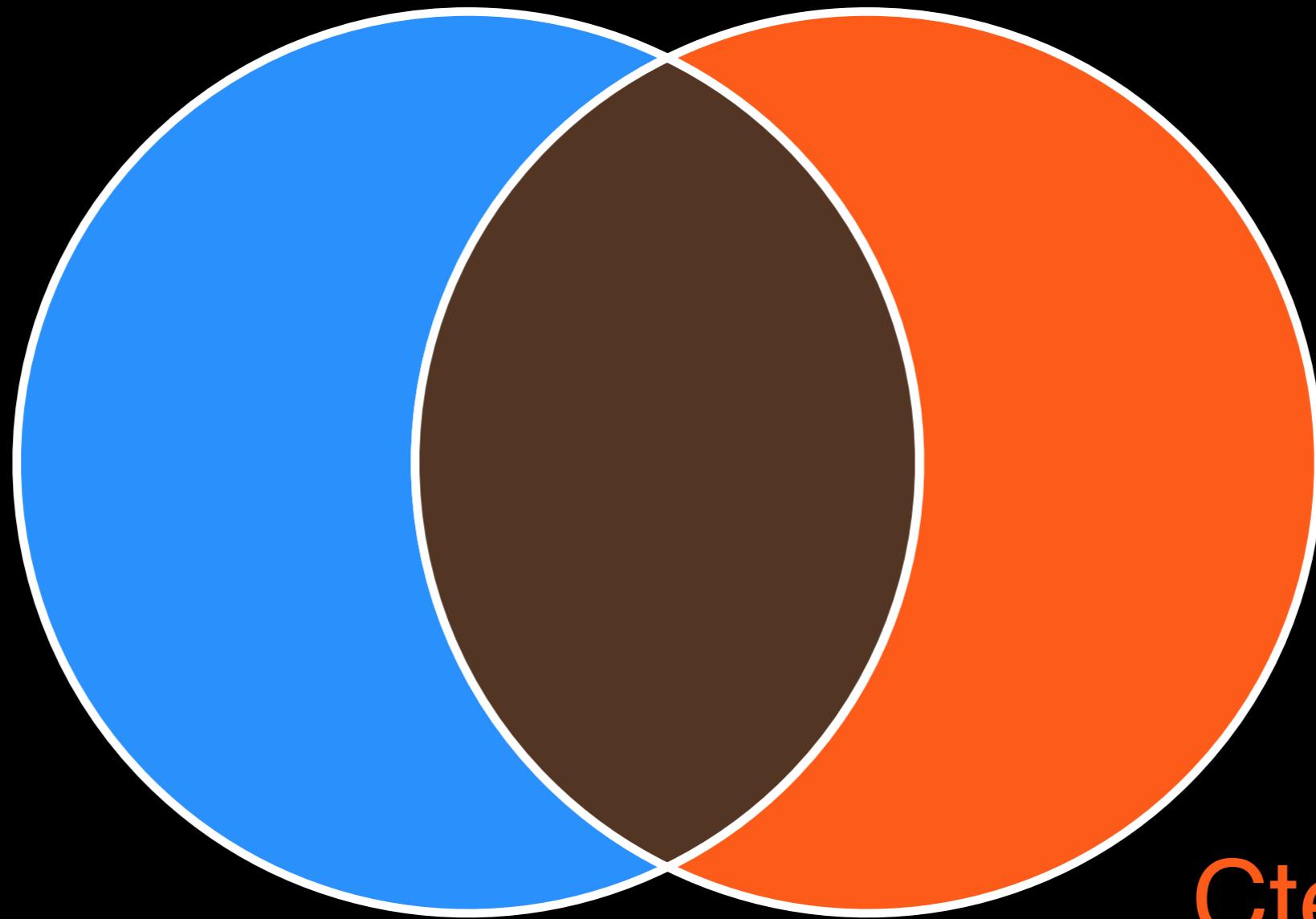
Primate
Specific

Ctenophore
Specific

Shared



What Exists



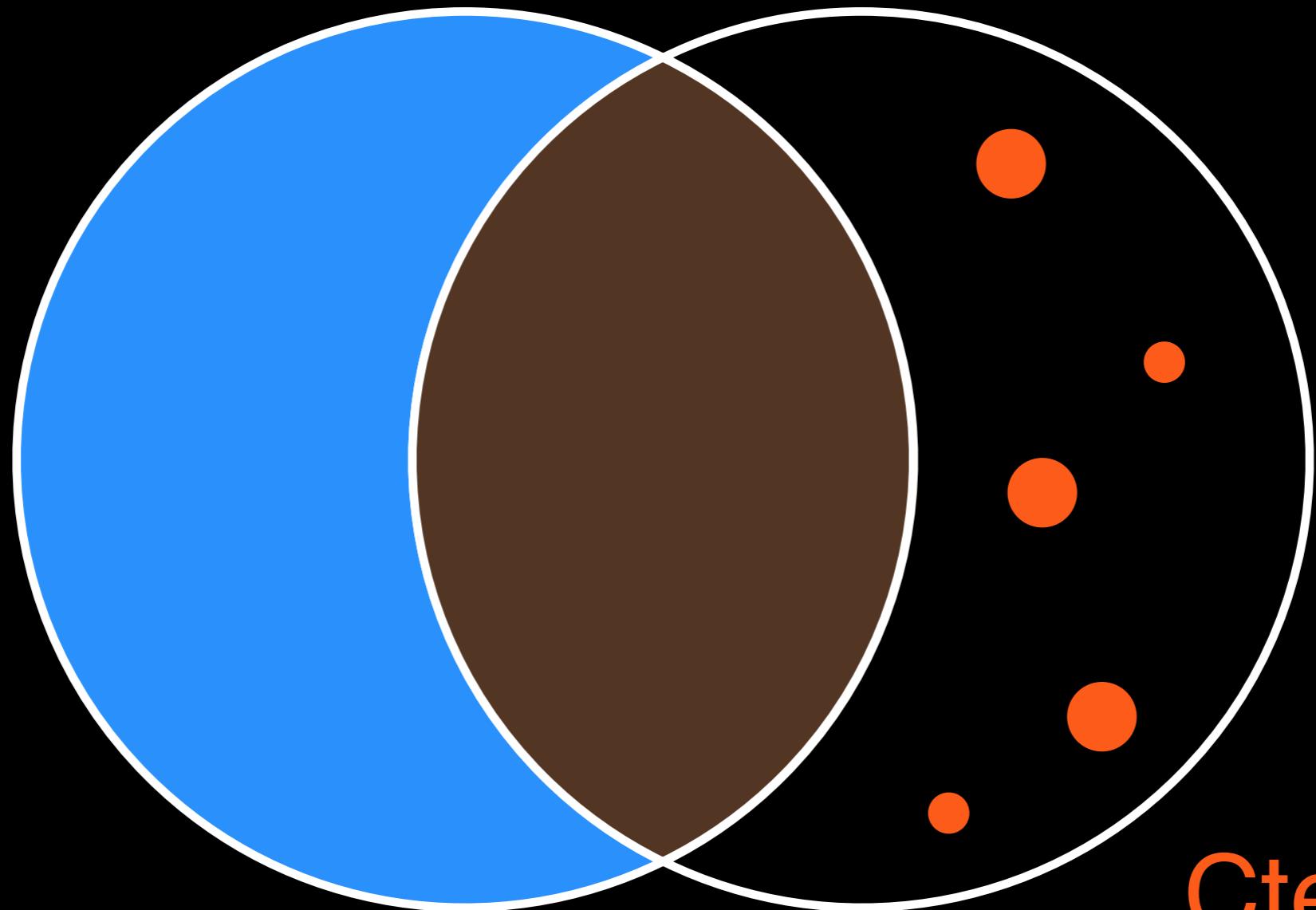
Primate
Specific

Shared

Ctenophore
Specific



What We See



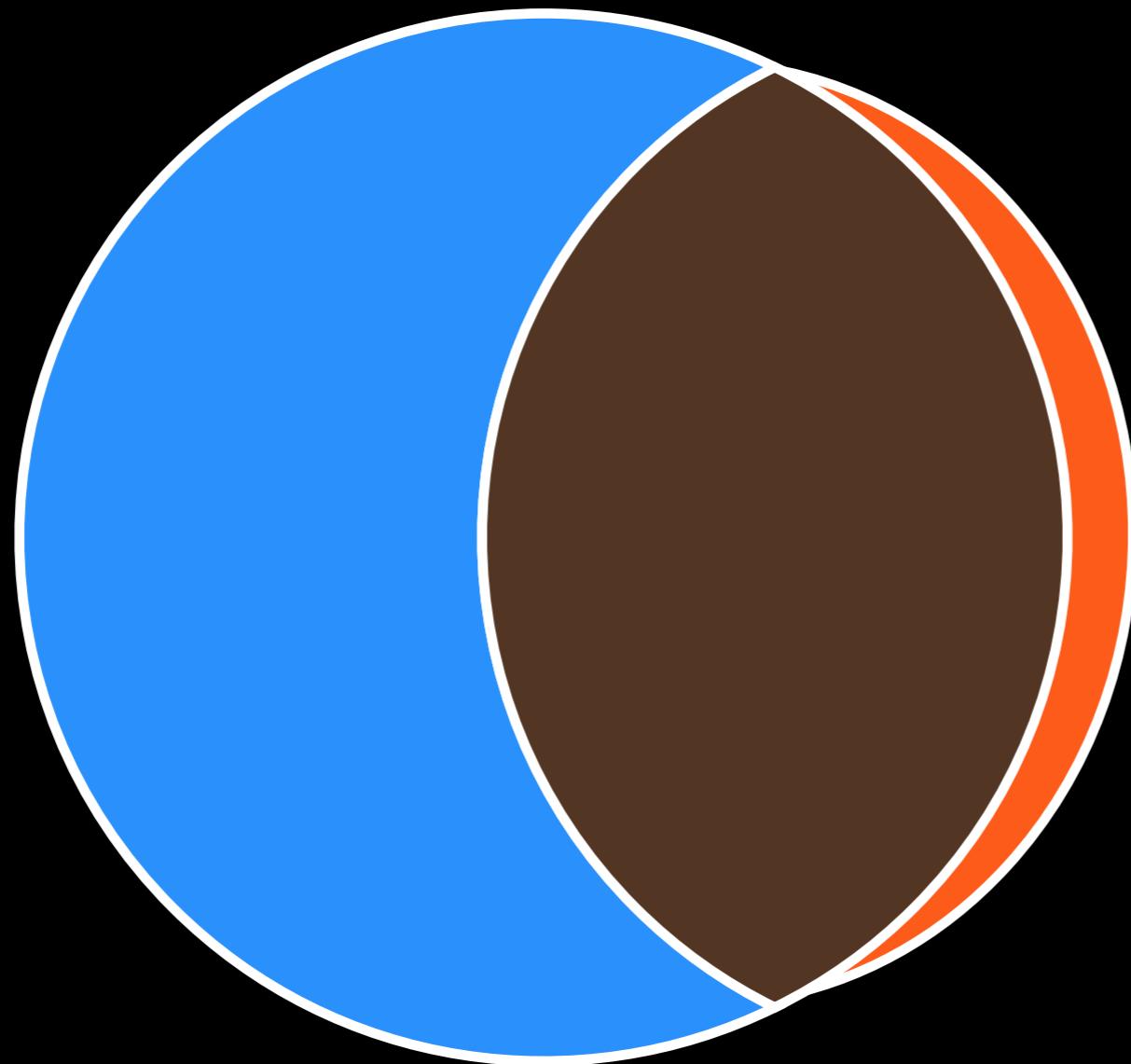
Primate
Specific

Shared

Ctenophore
Specific



What We Infer



Primate
Specific

Shared

Ctenophore
Specific

We have mistaken patterns in
our sampling preferences for
patterns in animal evolution

This is a clear ascertainment
bias



“With no brains, no heart, and no blood,
it's amazing jellyfish have existed for 650
million years!”

<http://news.nationalgeographic.com/2016/02/160302-jellyfish-immortal-science-animals-oceans-deadpool/>



Part 5: The evolution of expression



Integrative and Comparative Biology

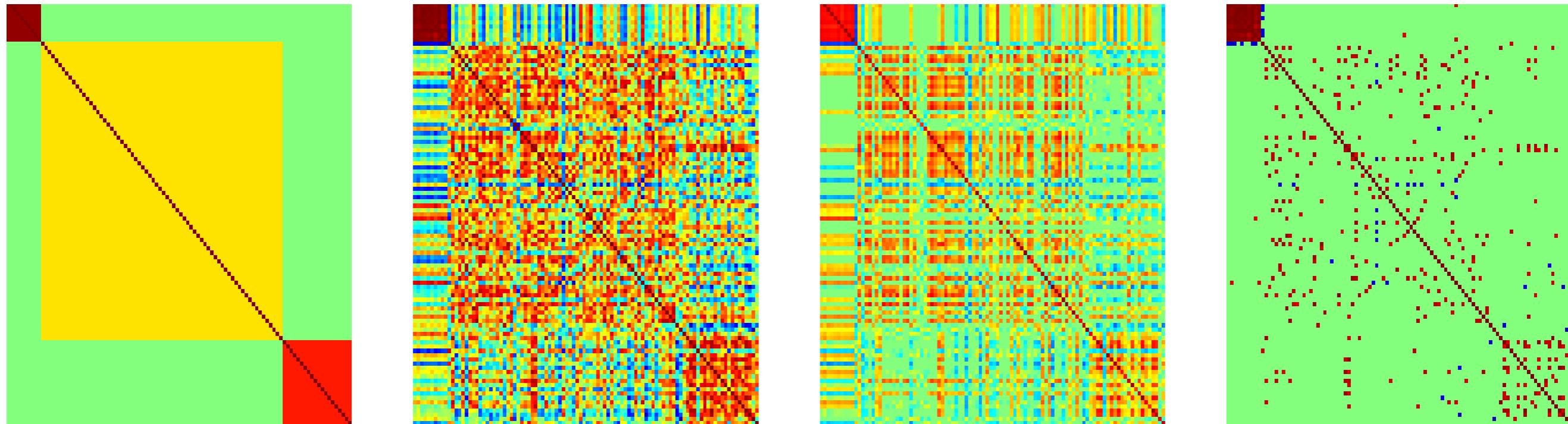
Integrative and Comparative Biology, pp. 1–10
doi:10.1093/icb/ict068

Society for Integrative and Comparative Biology

Phylogenetic Analysis of Gene Expression

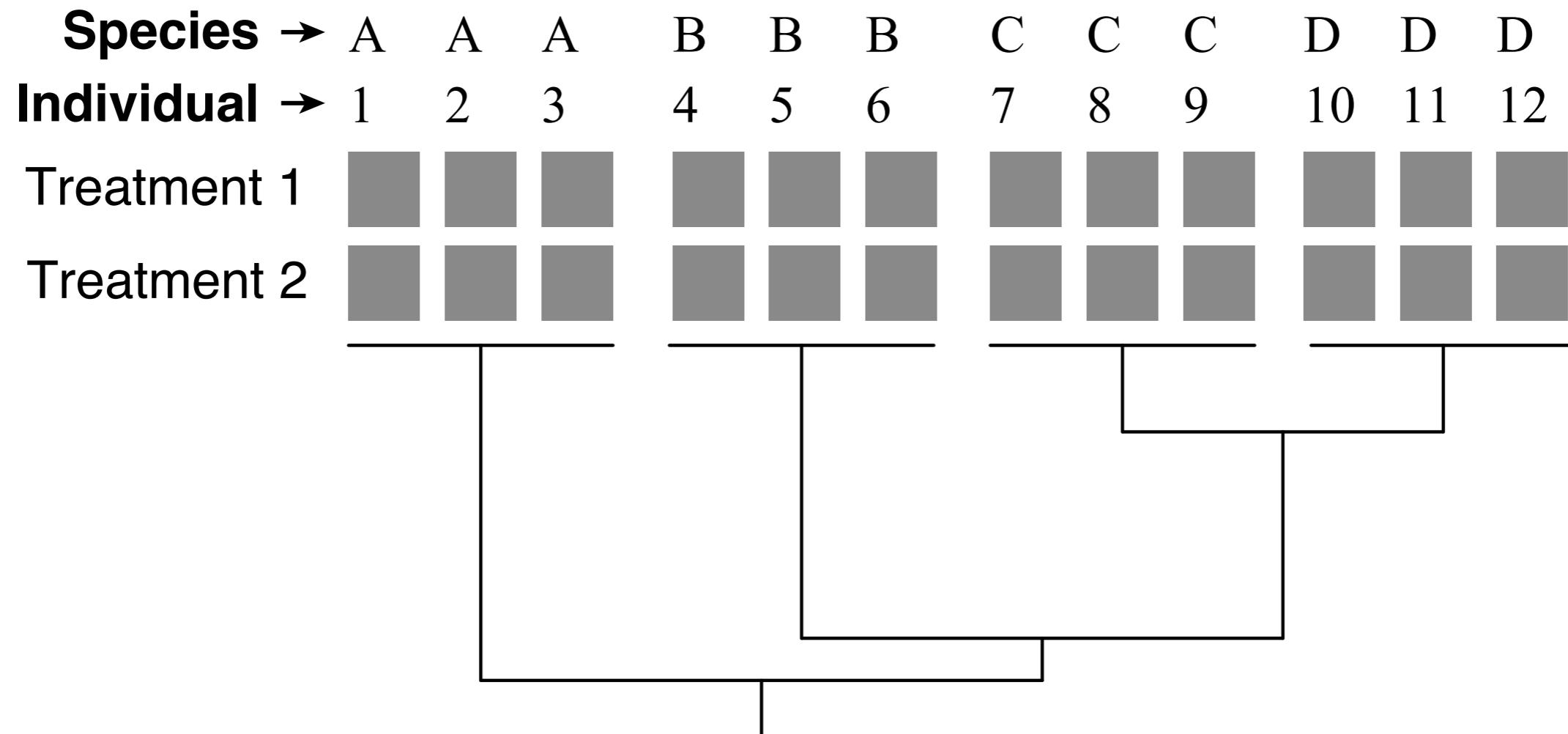
Casey W. Dunn,^{1,*} Xi Luo[†] and Zhijin Wu[†]

^{*}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.

I. Measuring expression

Current tools don't directly measure expression

I. Measuring expression



I. Measuring expression

What we measure

$$E[C_{ist}] = k_{is} E_{ist}$$

Where:

$E[C_{ist}]$ is the expected count of gene i in species s in treatment t

k_{is} is the counting efficiency of gene i in species s

E_{ist} is the expression of gene i in species s in tissue t

I. Measuring expression

What we want to know

$$E[C_{ist}] = k_{is} E_{ist}$$

Where:

$E[C_{ist}]$ is the expected count of gene i in species s in treatment t

k_{is} is the counting efficiency of gene i in species s

E_{ist} is the expression of gene i in species s in tissue t

I. Measuring expression

A nuisance

$$E[C_{ist}] = k_{is} E_{ist}$$

Where:

$E[C_{ist}]$ is the expected count of gene i in species s in treatment t

k_{is} is the counting efficiency of gene i in species s

E_{ist} is the expression of gene i in species s in tissue t

I. Measuring expression

k_{is} is the counting efficiency of gene i in species s

Influenced by:

Sequence length

Sequence composition

I. Measuring expression

So we can't map gene counts onto a phylogeny and derive independent contrasts in the standard way...

What, then, are we to do?

I. Measuring expression

Get rid of k_{is} !

$$E_{ist} = \frac{C_{ist}}{k_{is}}$$

Expression →

Count ←

nuisance ←

I. Measuring expression

Get rid of k_{is} !

Make comparisons between treatments before comparisons between species:

$$\frac{E_{is1}}{E_{is2}} = \frac{C_{is1}}{k_{is}} \frac{k_{is}}{C_{is2}} = \frac{C_{is1}}{C_{is2}}$$

I. Measuring expression

Take home:

Counts can't be directly compared across species, but some ratios of counts can be.

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:

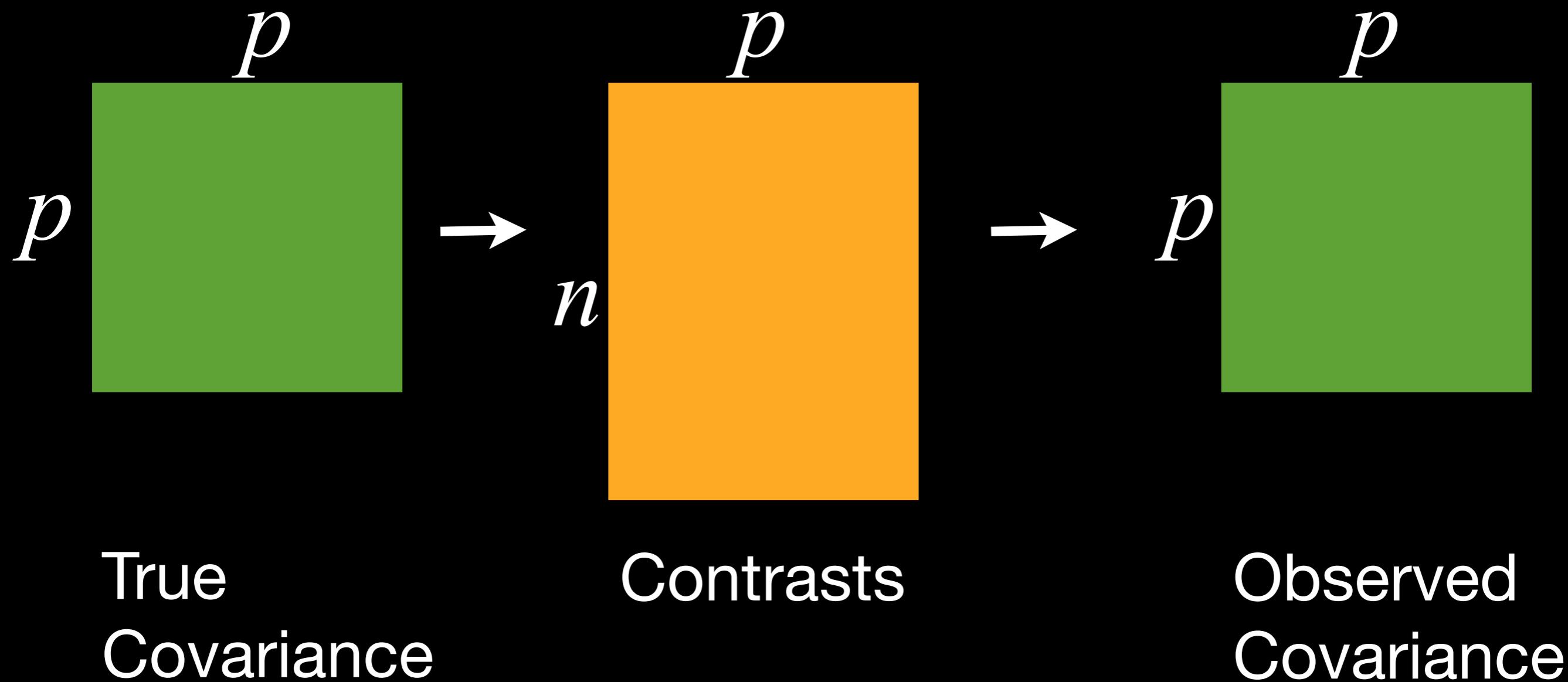
$$n \ll p$$

n number of observations
(eg contrasts)

p number of variables

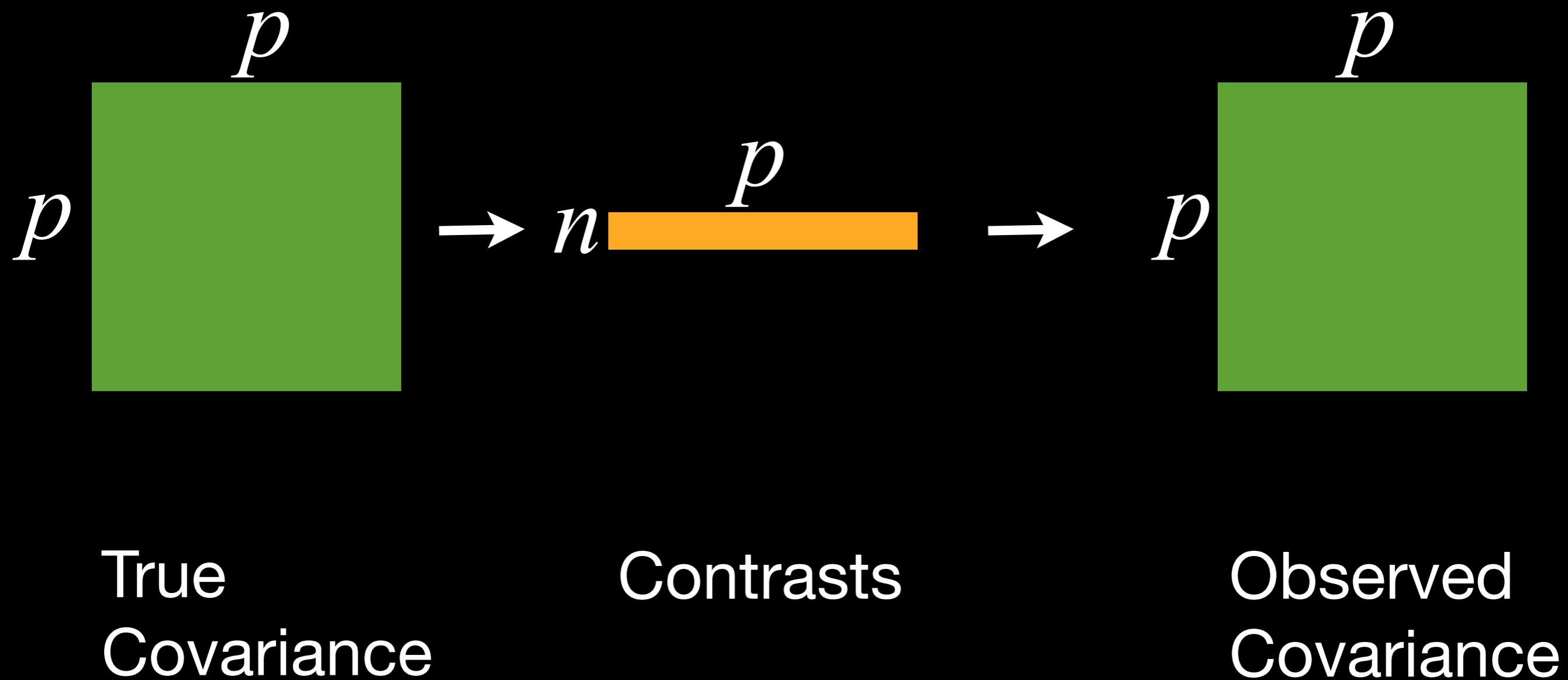
II. Interpreting covariance

When $n > p$



II. Interpreting covariance

When $n \ll p$



II. Interpreting covariance

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

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

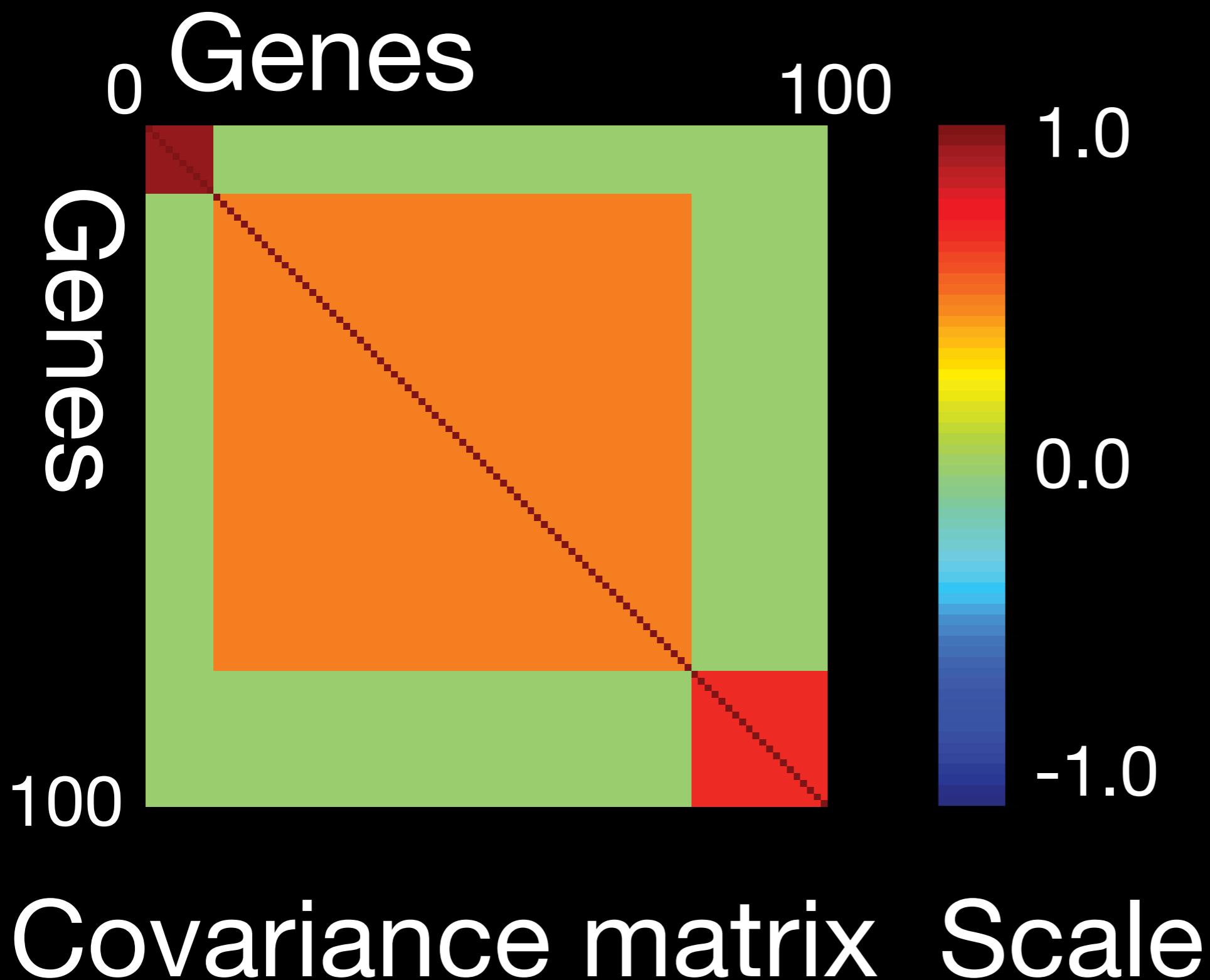
II. Interpreting covariance

Challenges of working with
matrices when $n \ll p$:

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

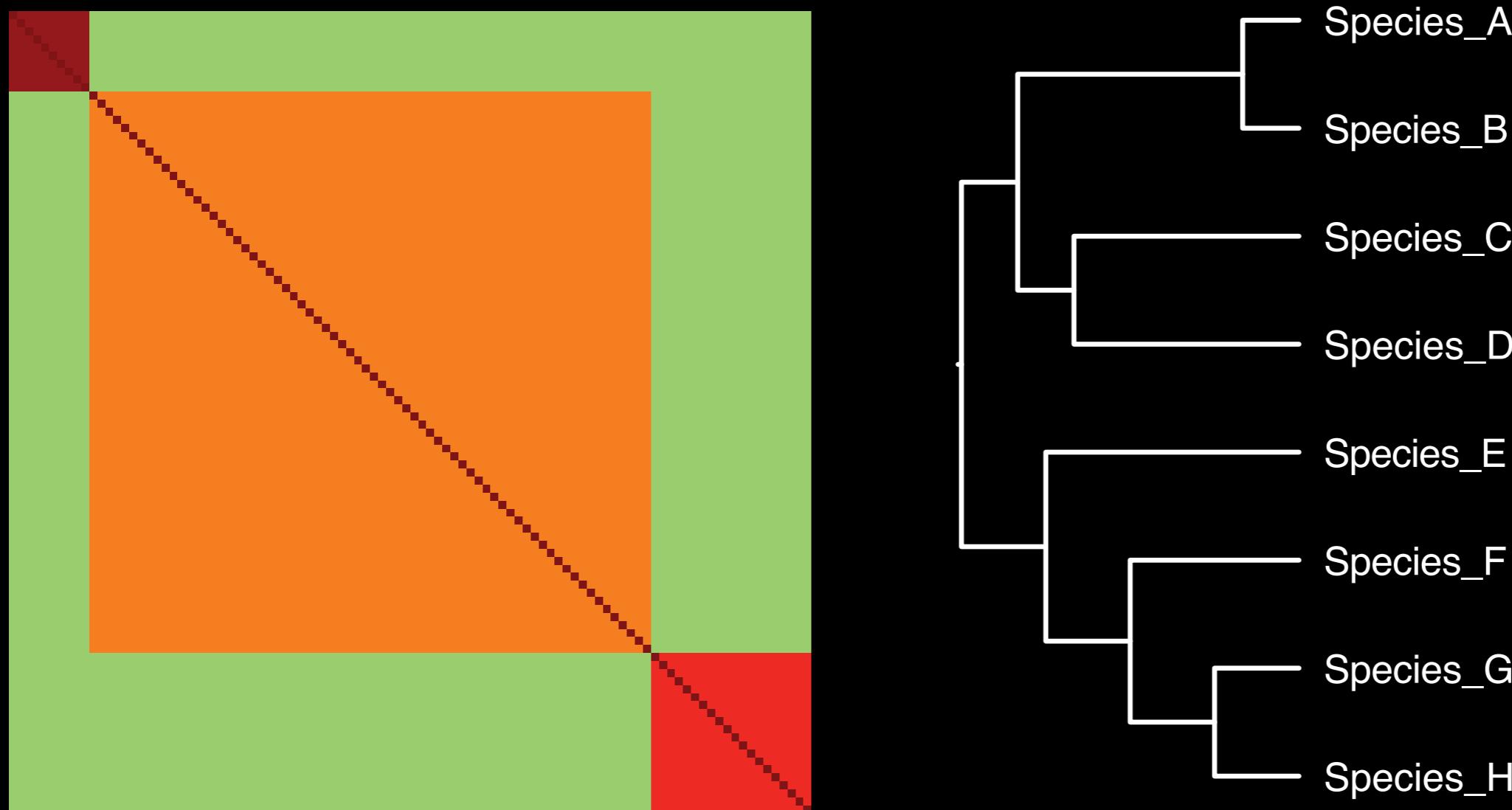
If you are looking at many variables in a small number of observations, you will find many spurious correlations

II. Interpreting covariance



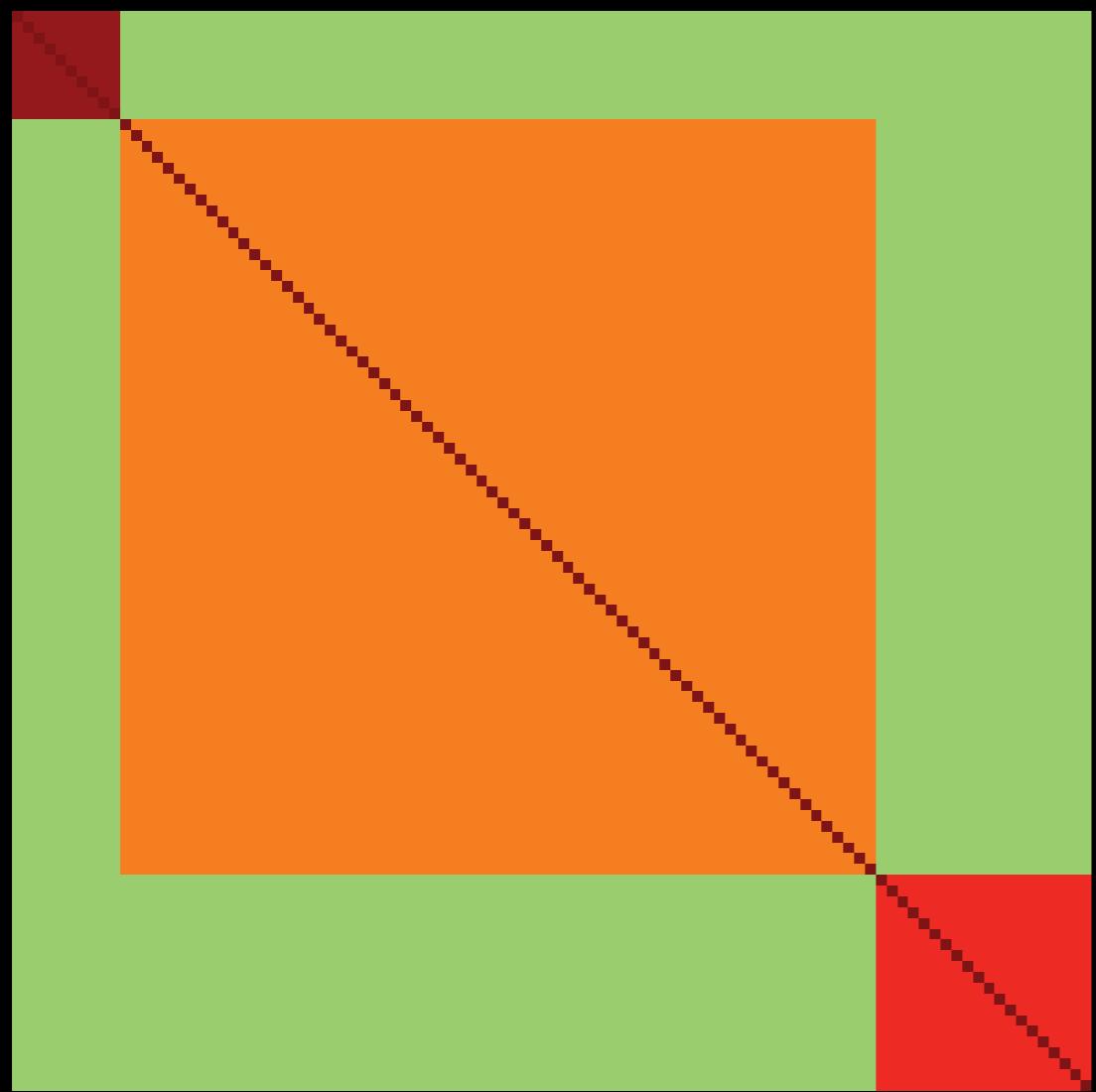
II. Interpreting covariance

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

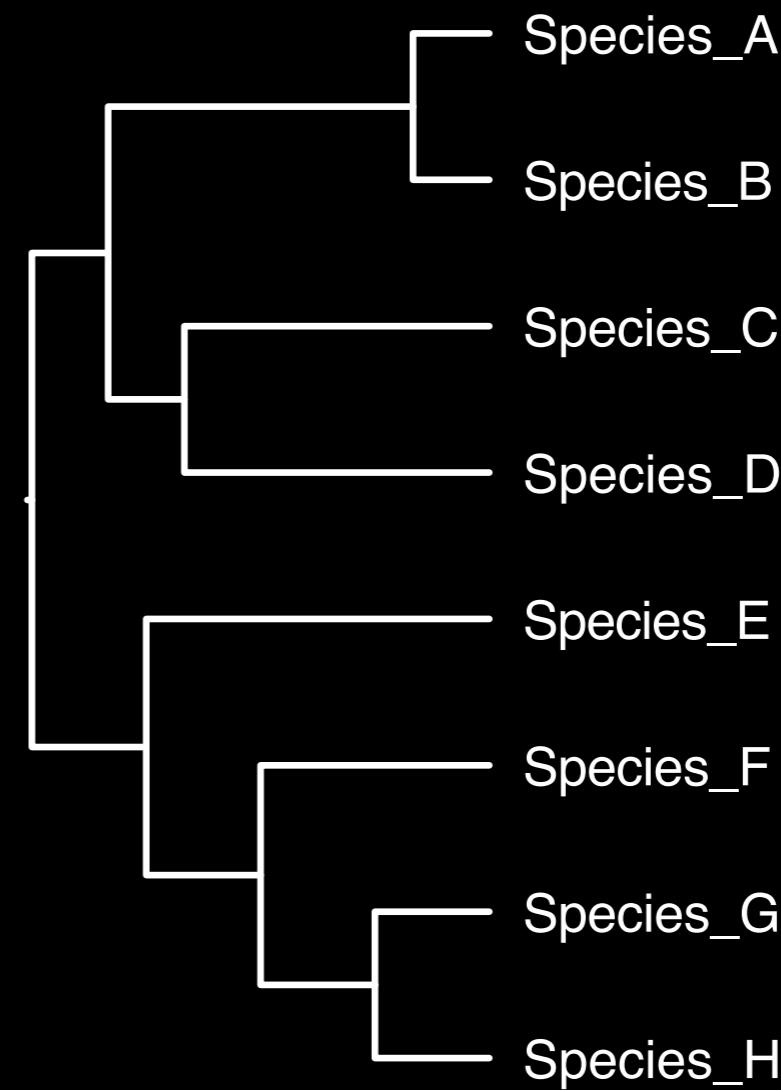


II. Interpreting covariance

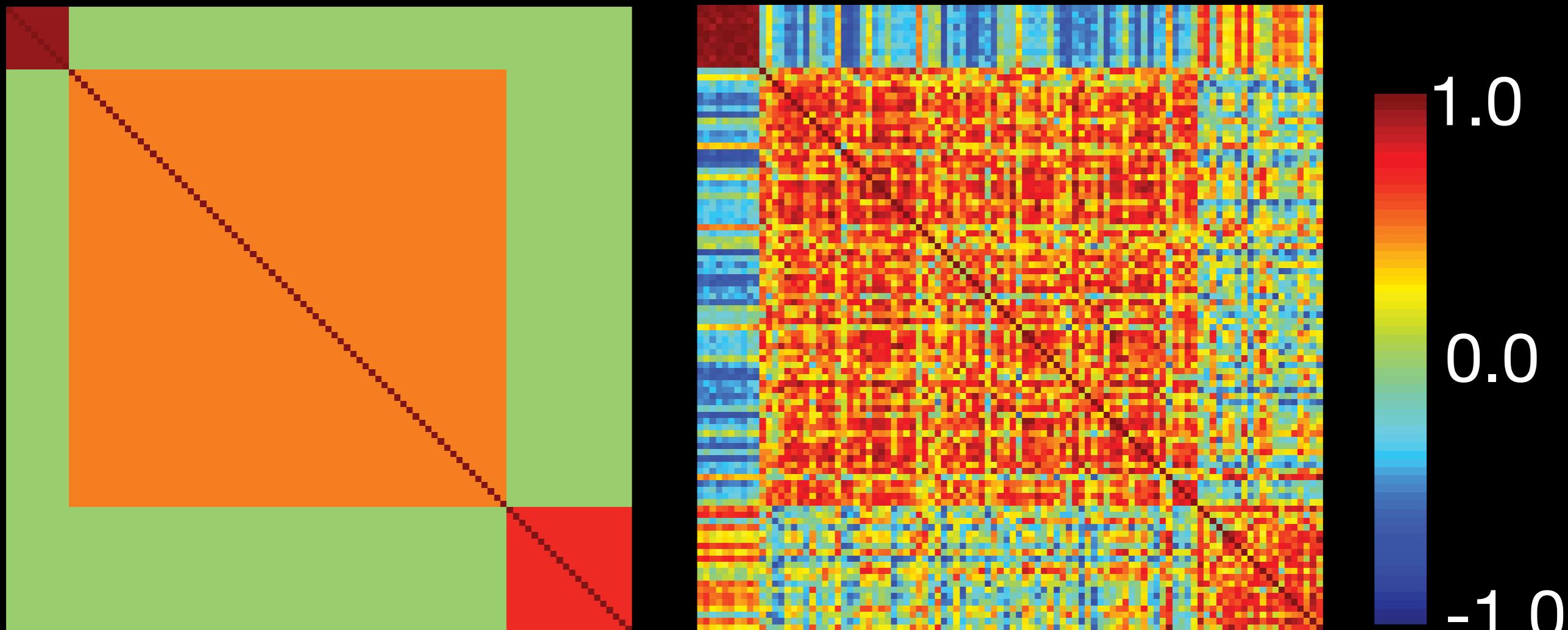
$p=100$



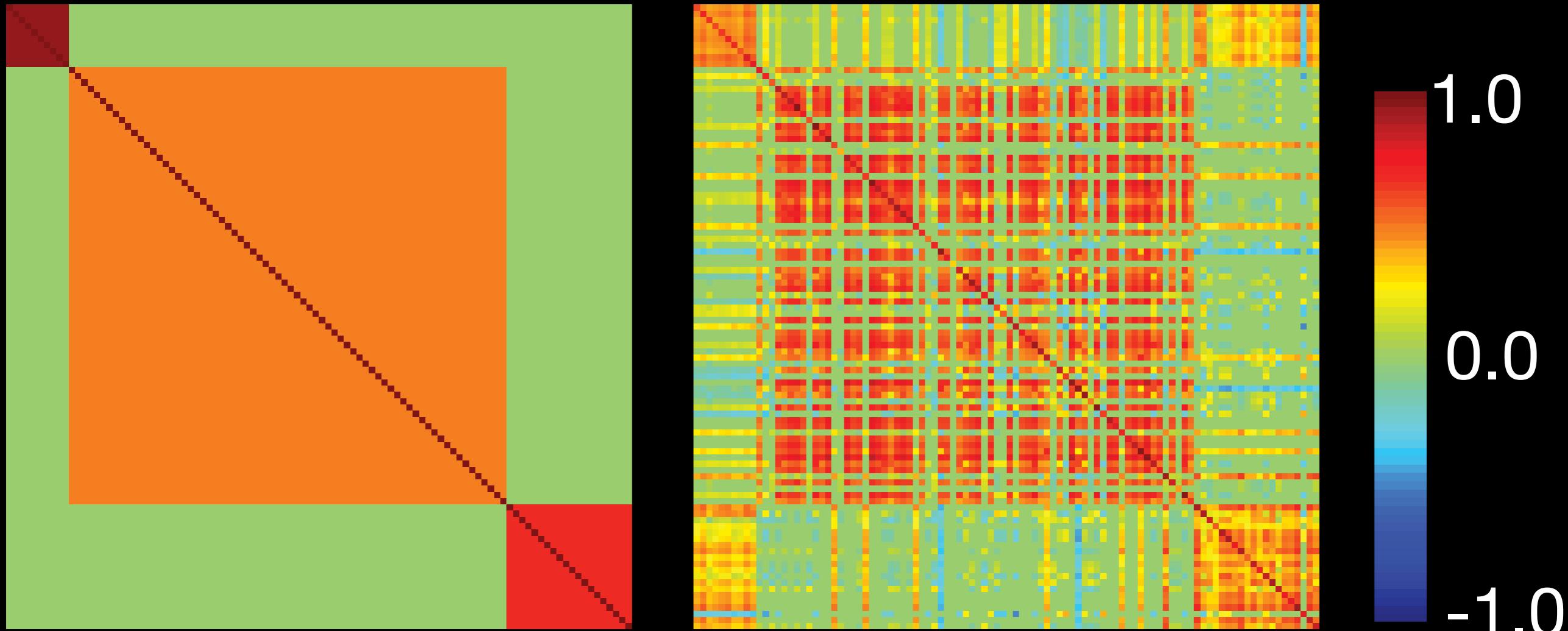
$n=7$



II. Interpreting covariance



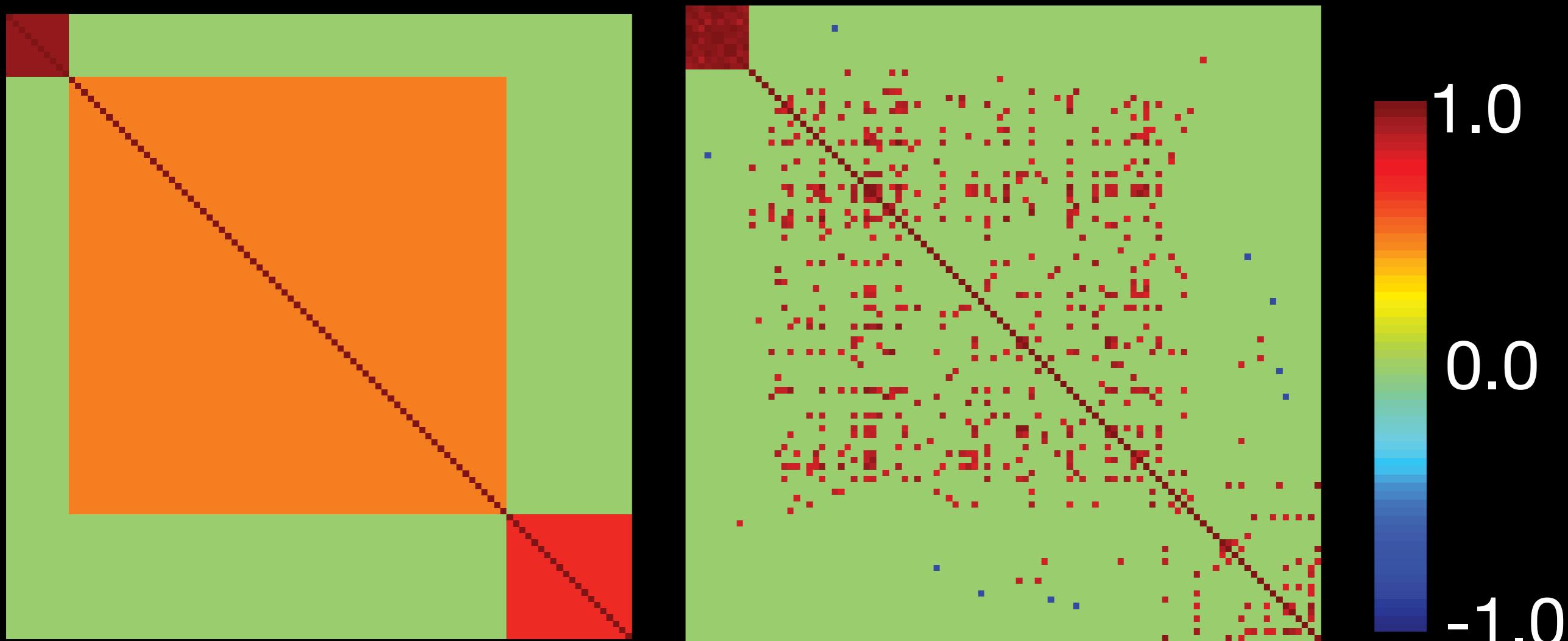
II. Interpreting covariance



“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 $n \ll p$
but false positives can be mitigated through regularization

And now for a
temporary digression:

Cartoons



<http://creaturecast.org/archives/2820-creaturecast-phylogenies>

Support



Waterman Award, Division
of Environmental Biology



iPlant Collaborative

The Manning Chair

RI EPSCoR

The Lab...

