Phylogenomics

Casey Dunn
Yale University
Professor, Dept. of Ecology and Evolutionary Biology
Curator of Invertebrate Zoology, Peabody Museum
Twitter - @caseywdunn

So you want to study molecular evolution…

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 analysis for molecular evolution projects

Preprocessing includes:

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

practical computing for biologists

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

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

Sinauer Associates, Inc.

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

Phylodiversity
Number of taxa
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

There are now many different sequencers, with tradeoffs in:

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

Error rate comparison (2015)

Which approach to sample preparation, sequencing, and data preparation 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

Directed PCR
Whole genome
Targeted enrichment
“Whole” transcriptome
RAD tag
Increasing enrichment

Whole genome

- No enrichment.

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

- Often need transcriptome data to annotate genes.

Directed PCR
Whole genome
Targeted enrichment
“Whole” transcriptome
RAD tag
Increasing enrichment

Whole transcriptome

- Enriched for expressed protein coding genes

- There is no One True Transcriptome

Directed PCR
Whole genome
Targeted enrichment
“Whole” transcriptome
RAD tag
Increasing enrichment

Targeted enrichment

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

Directed PCR
Whole genome
Targeted enrichment
“Whole” transcriptome
RAD tag
Increasing enrichment

RAD tag

- Enriched for randomly distributed, but consistent, genome regions

- No need for specific probes

Directed PCR
Whole genome
Targeted enrichment
“Whole” transcriptome
RAD tag
Increasing enrichment

Directed PCR

- Simple and cheap for a small number of genes
- Doesn’t scale so well to many genes

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

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

Whole genome (de novo assembly)

Sample preparation

DNA Isolation

Critical to:

- Avoid contamination by other DNA (eg symbionts)
- Avoid contamination by compounds that can inhibit downstream steps (eg metabolites)
- Minimize number of unique genomes*
- Maintain DNA integrity

Library preparation usually includes:

Fragmentation*

Size selection

Adapter integration

Amplification*

Why fragment?

1. Short-read sequencing needs DNA that is shorter than most preps provide

2. To make sequencing coverage more uniform

Starting material
Fragment
Fragments
Prepare library, sequence
Reads

Assembly

Assembly:

- Puzzles reads together to undo fragmentation
- Identifies and fixes sequencing errors
- Collapses redundancy

Starting material
Fragment
Prepare library, sequence
Assembly
Final product

Overlap assemblers - puzzling together of long sequences based on overlap

Used for sequencing technologies that produce a smaller number of large molecules

eg, the Canu assembler for PacBio

Overlap assembly isn’t computationally tractable for a large number of short sequences (eg Illumina)

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

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

B Overlap Graph

C de Bruijn Graph

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

Paths through the de Bruijn graph are assembled sequences

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

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

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

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

Typically >100 fold coverage

Even then, assemblies are quite fragmented

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

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

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

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

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

- 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

Dovetail GENOMICS, LLC

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

10X GENOMICS™

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)

Tractable genome size on an NSF grant…

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

Cold Spring Harbor Laboratory

bioRχiv beta

THE PREPRINT SERVER FOR BIOLOGY

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

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

As a tool to design and interpret other sequencing approaches

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

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

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

Assembly results...

Genome

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

Transcript

Splice variants

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

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

-Sequencing and assembly errors can be misinterpreted as splice variants

-Data may be insufficient to predict splice variants

It gets worse...

Genomes have uniform depth

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

Assemblers can make assumptions about uniform distribution of sequencing effort

But transcriptomes have non-uniform depth

- Different expression across genes

- Different splice variants within genes

Expression differences mean:

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

- Low copy number doesn’t necessarily indicate an error

- High copy number doesn’t necessarily indicate a repeat

- Sequencing error is hard to accommodate in transcriptomes

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

Agalma

Our automated transcriptome workflow

The tool

https://bitbucket.org/caseywdunn/agalma

Bitbucket Repositories Create owner/repository

Ag agalma

caseywdunn Following Share

Clone Fork Compare Pull request

Overview Source Commits Pull requests Issues 1 Downloads

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

SSH git@bitbucket.org:caseywdunn/aga

1	0	0	3
Branch	Tags	Forks	Followers

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

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

New developments

Multiple sequencing platforms can now sequence full length transcripts, greatly improving confidence in splicing.

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

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

Can work well on poorly preserved specimens

Challenges

Need to know what genes to sequence before you start

Ascertainment biases

Difficult to integrate data across studies with different genes

Need to optimize for different clades

Typical use case

Phylogenetic analyses with broad taxon sampling

Background reading:

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

Directed PCR

Data acquisition

Select genes

Design primer pairs that hybridize to genes

Amplify and sequence genes

Data preprocessing

(Select genes)

Assemble reads into gene sequences

Advantages

Easy to integrate with existing data

Strong control over which regions are sequenced

Greatly simplified assembly and annotation

Challenges

Need to know what genes to sequence before you start

Very labor intensive for more than a few genes

Need to optimize for different clades

Expensive at scale

Typical use case

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

Other enrichment tools...

SNP chips

Illumina multi-sample array formats

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

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

23andMe – Genetic Testing

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

Can also count frequency of:

- dinucleotides (2-mers)
- 3-mers
- 10-mers
- 10 million-mers

A sequencer is a highly specialized microscope that allows you to observe linear adjacency of nucleotides in space.

fastq example:

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

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

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

Sequencing is a type of molecular anatomy

Most current sequencing methods remove spatial context

This simplifies things, 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 linear proximity of nucleotides to each other and their broader 3D context in the organism.

For example…

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

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

http://fast.wistia.net/embed/iframe/f75ht43w1q

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

Hi-C

Stevens et al 2017
http://www.nature.com/nature/journal/v544/n7648/fig_tab/nature21429_F1.html

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”

http://phasegenomics.com/hic-technology/

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.

	HUMMINGBIRD (*CALYPTE ANNA*)		GOAT (*CAPRA HIRCUS*)		MAIZE (*ZEA MAYS*)	
YEAR	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

http://science.sciencemag.org/content/357/6346/10.full

Part 2: Homology evaluation

Homology is the correspondence of traits due to common descent from a shared ancestral trait.

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
Select genes
Amplify and sequence selected genes
Assemble matrix from all sequenced genes
Phylogenetic inference
After
Sequence at random
Identify homologous sequences and evaluate paralogy
Select genes
Assemble matrix
Phylogenetic inference

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

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

Identifying and selecting homologs

In molecular evolution, we often talk about two different types of homology…

Species A
Species B
Species C
Species
Gene

Species A
Species B
Species C
Gene divergence
due to duplication

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.

Sometimes we are interested in orthologs, sometimes in paralogs, and sometimes both.

For example, to study phylogenetic relationships between species we usually want to focus on orthologs.

Clearest
Orthology
Most
Informative
Clearest
Homology
Available Data

When comparing sequences in a molecular evolution study, we typically ask at some point:

1 - Are the sequences homologous?

2 - If so, are they orthologs or paralogs?

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

Phenetics is the investigation of biological similarity

Similarity can be due to shared ancestry, or convergence

Things that share ancestry can be very different

Some tools evaluate both homology and orthology with phenetic methods

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

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

Some tools evaluate both homology and orthology with phenetic methods

Use phenetic tools to identify orthologs *de novo*

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

Figure 1. OrthoMCL graph construction between two species, including the establishment of co-ortholog relationships. Solid lines connecting A1 and B1 represent putative ortholog relationships identified by the 'reciprocal best hit' (RBH) rule. Dotted lines (e.g. those connecting A1 with A2 and A3, or B1 with B2) represent putative in-paralog relationships within each species, identified using the 're-ciprocal 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

Chen et al 2007 (dx.doi.org/10.1371/journal.pone.0000383)

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 misascribing paralogs as orthologs

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

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

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

Once we have subtrees of orthologs...

Align each ortholog

Build trees

Can do this with:

https://bitbucket.org/caseywdunn/agalma

Bitbucket Repositories Create owner/repository

Ag agalma

caseywdunn Following Share

Clone Fork Compare Pull request

Overview Source Commits Pull requests Issues 1 Downloads 1

Agalma is developed by the Dunn Lab at Brown University.

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

Overview of Agalma

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

Agalma provides a completely automated analysis workflow that filters and assembles the data under default parameters, and records rich diagnostics. The same goes for alignment, translation, and phylogenetic analysis. You can then evaluate these diagnostics to spot problems and examine the success of your analyses, the quality of the 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 adapters (resulting from small inserts)
- remove clusters where one or both reads is of low mean quality
- randomize the sequences in the same order in both pairs to make obtaining random subsets easy
- assemble and annotate rRNA sequences based on a subassembly of the data
- remove clusters in which one or both reads map to rRNA sequences

SSH git@bitbucket.org:caseywdunn/aga

1	0	0	3
Branch	Tags	Forks	Followers

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

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

Syst. Biol. 0(0):1–20, 2014
© The Author(s) 2014. Published by Oxford University Press on behalf of the Society of Systematic Biologists.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
DOI:10.1093/sysbio/syu082

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 Februay 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 RESEARCH

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 Daubin[1]

[1]*Université de Lyon, Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Villeurbanne F-69622, France;* [2]*Department of Integrative Biology, UC Berkeley, Berkeley, California 94720-3140, USA;* [3]*INRIA 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

Syst. Biol. 0(0):1–15, 2014
© The Author(s) 2014. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved.
For Permissions, please email: journals.permissions@oup.com
DOI:10.1093/sysbio/syu128

Assessing Approaches for Inferring Species Trees from Multi-Copy Genes

RUCHI CHAUDHARY[1,2,*], BASTIEN BOUSSAU[3], J. GORDON BURLEIGH[2], AND DAVID FERNÁNDEZ-BACA[1]

[1]*Department of Computer Science, Iowa State University, Ames, IA 50011, USA;* [2]*Department of Biology, University of Florida, Gainesville, FL 32611, USA; and* [3]*Université 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

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. PHYLDOG 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 alternative 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.

PNAS

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.

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

http://useast.ensembl.org/info/genome/compara/homology_method.html

Orthology and paralogy are unnecessarily complicated and overrated…

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

http://useast.ensembl.org/info/genome/compara/homology_method.html

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

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.

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

Part 3:
General themes in project design and interpretation

What are the limitations of describing patterns and processes of molecular evolution?

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

“It is just descriptive, not hypothesis driven”

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

Dunn and Munro 2016
doi:10.1111/zsc.12211

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

Dunn and Munro 2016
doi:10.1111/zsc.12211

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

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

It is descriptive and hypothesis driven.

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

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

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

Some ways forward

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

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

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

bandwagon

Appealing to popularity or the fact that many people do something as an attempted form of validation.

Shamus pointed a drunken finger at Sean and asked him to explain how so many people could believe in leprechauns if they're only a silly old superstition. Sean, however, had had a few too many Guinness himself and fell off his chair.

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).

composition /division

Assuming that what's true about one part of something has to be applied to all, or other, parts of it.

Daniel was a precocious child and had a liking for logic. He reasoned that atoms are invisible, and that he was made of atoms and therefore invisible too. Unfortunately, despite his thinky skills, he lost the game of hide and go seek.

no true scotsman

Making what could be called an appeal to purity as a way to dismiss relevant criticisms or flaws of an argument.

Angus declares that Scotsmen do not put sugar on their porridge, to which Lachlan points out that he is a Scotsman and puts sugar on his porridge. Furious, like a true Scot, Angus yells that no **true** Scotsman sugars his porridge.

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.

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.

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 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 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

http://www.tylervigen.com

- Age of Miss America
- Murders by steam, hot vapors, and hot objects

http://www.tylervigen.com

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

Part 4:
Some of the things we are thinking about in my lab...

Field
Lab
Computational

Siphonophores

Siebert et al, Zootaxa 2013

https://vimeo.com/103529382

Article ZOOTAXA

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

STEFAN SIEBERT[1], PHIL R. PUGH[2], STEVEN H. D. HADDOCK[3] & CASEY W. DUNN[1]

Siebert et al, Zootaxa 2013
http://dx.doi.org/10.11646/zootaxa.3702.3.1

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Improved phylogenetic resolution within Siphonophora (Cnidaria) with implications for trait evolution

Catriona Munro[a,*,1], Stefan Siebert[a,b,1], Felipe Zapata[c], Mark Howison[d,e], Alejandro Damian-Serrano[a,j], Samuel H. Church[a,f], Freya E. Goetz[a,g], Philip R. Pugh[h], Steven H.D. Haddock[i], Casey W. Dunn[j]

https://doi.org/10.1016/j.ympev.2018.06.030

Deep animal phylogeny

Vol 452|10 April 2008|doi:10.1038/nature06614

nature

Broad phylogenomic sampling improves resolution of the animal tree of life

Casey W. Dunn[1]†, Andreas Hejnol[1], David Q. Matus[1], Kevin Pang[1], William E. Browne[1], Stephen A. Smith[2], Elaine Seaver[1], Greg W. Rouse[3], Matthias Obst[4], Gregory D. Edgecombe[5], Martin V. Sørensen[6], Steven H. D. Haddock[7], Andreas Schmidt-Rhaesa[8], Akiko Okusu[9], Reinhardt Møbjerg Kristensen[10], Ward C. Wheeler[11], Mark Q. Martindale[1] & Gonzalo Giribet[12,13]

Dunn et al, 2008

http://dx.doi.org/10.1038/nature06614

The phylogenetic placement of sponges and ctenophores

Ctenophores

Dunn et al 2008

Traditional

Sponges

Few cell types

Most structures not epithelial

Sessile filter feeders

Thought to branch off from other animals before origin of many complex structures (nervous system, musculature, etc)

Ctenophores

Three tissue layers

Active predators

Nervous system, musculature

Largest organisms to move by ciliary action

Capture prey with colloblasts (glue cells)

Very unique embryonic development

(Steve Haddock)

Our results, and most subsequent analyses, reject the two prevailing previous hypotheses for the placement of ctenophores in the animal phylogeny.

Acrosomata

Sponges

Placozoa

Cnidarians

Ctenophores

Bilaterians

Acrosomata

Sponges
Placozoa
Cnidarians
Nervous system
Ctenophores
Bilaterians

Acrosomata

Sponges
Placozoa
Cnidarians
Ctenophores
Bilaterians
Nervous system
Third cell layer (mesoderm)

Coelenterata

Coelenterata

The placement of ctenophores as the sister group to all other animals has been quite controversial…

Multiple studies show that it is sensitive to:

- Which genes you analyze
- Which species you include in the analysis
- Which model you use

The skeptics…

Current Biology *19*, 706–712, April 28, 2009 ©2009 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2009.02.052

Phylogenomics Revives Traditional Views on Deep Animal Relationships

Hervé Philippe,[1,11] Romain Derelle,[2,11] Philippe Lopez,[2] Kerstin Pick,[3,5] Carole Borchiellini,[6] Nicole Boury-Esnault,[6] Jean Vacelet,[6] Emmanuelle Renard,[6] Evelyn Houliston,[7] Eric Quéinnec,[2] Corinne Da Silva,[8] Patrick Wincker,[8] Hervé Le Guyader,[2] Sally Leys,[9] Daniel J. Jackson,[3,10] Fabian Schreiber,[4] Dirk Erpenbeck,[5] Burkhard Morgenstern,[3,4] Gert Wörheide,[5,*] and Michaël Manuel[2,*]

2009

Improved Phylogenomic Taxon Sampling Noticeably Affects Nonbilaterian Relationships

K.S. Pick,†,[1] H. Philippe,†,[3] F. Schreiber,[4] D. Erpenbeck,[1] D.J. Jackson,[2] P. Wrede,[5] M. Wiens,[5] A. Alié,[6] B. Morgenstern,[4] M. Manuel,[6] and G. Wörheide*,[1]

Mol. Biol. Evol. 27(9):1983–1987. 2010 doi:10.1093/molbev/msq089

2010

Integrative and Comparative Biology

Integrative and Comparative Biology, pp. 1–9
doi:10.1093/icb/ict008

Society for Integrative and Comparative Biolog

SYMPOSIUM

Novel Scenarios of Early Animal Evolution—Is It Time to Rewrite Textbooks?

Martin Dohrmann* and Gert Wörheide[1,†,‡,§]

2013

Other studies also find ctenophores sister…

Science

2013

The Genome of the Ctenophore *Mnemiopsis leidyi* and Its Implications for Cell Type Evolution

Joseph F. Ryan, Kevin Pang, Christine E. Schnitzler, Anh-Dao Nguyen, R. Travis Moreland, David K. Simmons, Bernard J. Koch, Warren R. Francis, Paul Havlak, NISC Comparative Sequencing Program, Stephen A. Smith, Nicholas H. Putnam, Steven H. D. Haddock, Casey W. Dunn, Tyra G. Wolfsberg, James C. Mullikin, Mark Q. Martindale, Andreas D. Baxevanis*

nature

2014

The ctenophore genome and the evolutionary origins of neural systems

Leonid L. Moroz[1,2,3], Kevin M. Kocot[4], Mathew R. Citarella[1], Sohn Dosung[1], Tigran P. Norekian[1,3], Inna S. Povolotskaya[5,6], Anastasia P. Grigorenko[7,8], Christopher Dailey[9], Eugene Berezikov[10], Katherine M. Buckley[11], Andrey Ptitsyn[1], Denis Reshetov[8], Krishanu Mukherjee[1], Tatiana P. Moroz[1], Yelena Bobkova[1], Fahong Yu[2], Vladimir V. Kapitonov[12], Jerzy Jurka[12], Yuri V. Bobkov[1], Joshua J. Swore[1,3], David O. Girardo[1,3], Alexander Fodor[1], Fedor Gusev[7,8], Rachel Sanford[1], Rebecca Bruders[1,3], Ellen Kittler[13], Claudia E. Mills[3], Jonathan P. Rast[11], Romain Derelle[5,6], Victor V. Solovyev[14], Fyodor A. Kondrashov[5,6,15], Billie J. Swalla[3], Jonathan V. Sweedler[9], Evgeny I. Rogaev[7,8,16,17], Kenneth M. Halanych[4] & Andrea B. Kohn[1]

PNAS

2015

Error, signal, and the placement of Ctenophora sister to all other animals

Nathan V. Whelan[a,1], Kevin M. Kocot[b], Leonid L. Moroz[c], and Kenneth M. Halanych[a]

[a]Molette Biology Laboratory for Environmental and Climate Change Studies, Department of Biological Sciences, Auburn University, Auburn, AL 36849; [b]School of Biological Sciences, University of Queensland, St. Lucia, QLD 4101, Australia; and [c]Department of Neuroscience and McKnight Brain Institute, University of Florida, Gainesville, FL 32610

Two papers published in PNAS within two weeks in 2015 find “maximal” support for conflicting sponge-sister and ctenophore-sister hypotheses.

Pisani et al.

Genomic data do not support comb jellies as the sister group to all other animals

Davide Pisani[a,b,1], Walker Pett[c], Martin Dohrmann[d], Roberto Feuda[e], Omar Rota-Stabelli[f], Hervé Philippe[g,h], Nicolas Lartillot[c], and Gert Wörheide[d,i,1]

www.pnas.org/cgi/doi/10.1073/pnas.1518127112

Chang et al.

Genomic insights into the evolutionary origin of Myxozoa within Cnidaria

E. Sally Chang[a], Moran Neuhof[b,c], Nimrod D. Rubinstein[d], Arik Diamant[e], Hervé Philippe[f,g], Dorothée Huchon[b,1], and Paulyn Cartwright[a,1]

www.pnas.org/cgi/doi/10.1073/pnas.1511468112

These studies show that placing the root of the animal tree is sensitive to multiple factors, including:

- Taxon sampling
- Gene sampling
- Model of molecular evolution

The Guardian
US edition

Evolution row ends as scientists declare sponges to be sister of all other animals

Longstanding question of whether sponges or comb jellies were first to branch off the evolutionary tree from the common ancestor of all animals may be settled

▲ A diver looks on at a giant barrel sponge, *Xestospongia testudinaria*. Photograph: Alamy Stock Photo

A longstanding row in animal evolution has come to a head, with a team of scientists claiming they have ended the debate over which type of creature is the sister of all other animals.

Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals

Roberto Feuda,[1,11] Martin Dohrmann,[2,11] Walker Pett,[3] Hervé Philippe,[4,5] Omar Rota-Stabelli,[6] Nicolas Lartillot,[7] Gert Wörheide,[2,8,9,*] and Davide Pisani[10,12,*]

Current Biology (2017), https://doi.org/10.1016/j.cub.2017.11.008

This paper addresses compositional heterogeneity, in this case the occurrence of different amino acids at different frequencies in different clades.

Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals

Roberto Feuda,[1,11] Martin Dohrmann,[2,11] Walker Pett,[3] Hervé Philippe,[4,5] Omar Rota-Stabelli,[6] Nicolas Lartillot,[7] Gert Wörheide,[2,8,9,*] and Davide Pisani[10,12,*]

Current Biology (2017), https://doi.org/10.1016/j.cub.2017.11.008

"Sponges are the sister group of the remaining animals"

"Trees showing comb jellies as the sister of all the other animals are artifactual"

The central analyses of the paper are based on amino acid recoding.

This method groups amino acids that are frequently exchanged in the course of evolution.

It discards information about changes within groups, and looks exclusively at changes between groups.

The intent is to discard compositional heterogeneity between clades, which is expected to consist largely of favoring different amino acids within each of these groups.

For example…
SR recoding is based on the JTT matrix

https://www.researchgate.net/publication/221916498_Evolutionary_Bioinformatics_with_a_Scientific_Computing_Environment/figures?lo=1

SR recoding is based on the JTT matrix

```
SR-1   ADEGKNPQRSTCFHILMVWY
SR-2  ADEGKNPQRST CFHILMVWY
SR-3   ADEGNPST CHKQRW FILMVY
SR-4   AGNPST CHWY DEKQR FILMV
SR-5   AGPST CFWY DEN HKQR ILMV
SR-6   APST CW DEGN FHY ILMV KQR
SR-7   AGST CW DEN FY HP ILMV KQR
SR-8   AST CG DEN FY HP ILV KQR MW
SR-9   AST CW DE FY GN HQ ILV KR MP
SR-10  AST CW DE FY GN HQ IV KR LM P
SR-11  AST C DE FY GN HQ IV KR LM P W
SR-12  AST C DE FY G HQ IV KR LM N P W
SR-13  AST C DE FY G H IV KR LM N P Q W
SR-14  AST C DE FL G H IV KR M N P Q W Y
SR-15  AST C DE F G H IV KR L M N P Q W Y
SR-16  AT C DE F G H IV KR L M N P Q S W Y
SR-17  AT C DE F G H IV K L M N P Q R S W Y
SR-18  A C DE F G H IV K L M N P Q R S T W Y
SR-19  A C D E F G H IV K L M N P Q R S T W Y
SR-20  A C D E F G H I V K L M N P Q R S T W Y
```

Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals

Roberto Feuda,[1,11] Martin Dohrmann,[2,11] Walker Pett,[3] Hervé Philippe,[4,5] Omar Rota-Stabelli,[6] Nicolas Lartillot,[7] Gert Wörheide,[2,8,9,*] and Davide Pisani[10,12,*]

The considered:

Four different matrices
(Chang and three versions of Whelan)

Two different models of molecular evolution
(GTR+G and CAT+GTR+G)

Four different recoding strategies
(None, Dayhoff, S&R, KGB)

For a total of 32 focal analyses

Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals

Roberto Feuda,[1,11] Martin Dohrmann,[2,11] Walker Pett,[3] Hervé Philippe,[4,5] Omar Rota-Stabelli,[6] Nicolas Lartillot,[7] Gert Wörheide,[2,8,9,*] and Davide Pisani[10,12,*]

Current Biology (2017), https://doi.org/10.1016/j.cub.2017.11.008

Table 3. Support for Porifera-Sister and Ctenophora-Sister from the WhelanD20-Opistho, WhelanD20-Holo, WhelanD20-Choano, and Chang Datasets

		Dataset			
Recoding	Hypothesis	WhelanD20-Opistho	WhelanD20-Holo	WhelanD20-Choano	Chang
Model: GTR+G					
None	Porifera-sister	≈0	≈0	≈0	≈0
None	Ctenophora-sister	≈1	≈1	≈1	≈1
Dayhoff-6	Porifera-sister	≈0	≈0	≈0	≈0
Dayhoff-6	Ctenophora-sister	≈1	≈1	≈1	≈1
S&R-6	Porifera-sister	≈0	≈0	≈0	≈0
S&R-6	Ctenophora-sister	≈1	≈1	≈1	≈1
KGB-6	Porifera-sister	≈0	≈0	≈0	≈0
KGB-6	Ctenophora-sister	≈1	≈1	≈1	≈1
Model: CAT-GTR+G					
None	Porifera-sister	≈0	0.68	0.77	≈0
None	Ctenophora-sister	≈1	0.31	0.2	≈1
Dayhoff-6	Porifera-sister	0.89	0.99	0.95	0.74
Dayhoff-6	Ctenophora-sister	0.09	≈0	0.01	0.25
S&R-6	Porifera-sister	0.98	0.99	0.94	0.62
S&R-6	Ctenophora-sister	≈0	≈0	0.02	0.22
KGB-6	Porifera-sister	0.42	0.68	0.37	0.78
KGB-6	Ctenophora-sister	0.17	0.05	0.29	0.22

Support values are PPs. All analyses but the CAT-GTR+G analysis of the amino acid version of Chang converged well, and all convergence statistics are detailed in the captions of Figures S1–S4. Support values for Porifera-sister and Ctenophora-sister do not always add up to 1 because alternative topologies, e.g., monophyletic Porifera+Ctenophora sister to the remaining animals, are sometime recovered.

Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals

Roberto Feuda,[1,11] Martin Dohrmann,[2,11] Walker Pett,[3] Hervé Philippe,[4,5] Omar Rota-Stabelli,[6] Nicolas Lartillot,[7] Gert Wörheide,[2,8,9,*] and Davide Pisani[10,12,*]

Current Biology (2017), https://doi.org/10.1016/j.cub.2017.11.008

"As modeling improves, sponges emerge in this position instead, indicating that trees placing the comb jellies at the root of the animals are artifactual."

Table 3. Support for Porifera-Sister and Ctenophora-Sister from the WhelanD20-Opistho, WhelanD20-Holo, WhelanD20-Choano, and Chang Datasets

		Dataset			
Recoding	Hypothesis	WhelanD20-Opistho	WhelanD20-Holo	WhelanD20-Choano	Chang
Model: GTR+G					
None	Porifera-sister	≈0	≈0	≈0	≈0
None	Ctenophora-sister	≈1	≈1	≈1	≈1
Dayhoff-6	Porifera-sister	≈0	≈0	≈0	≈0
Dayhoff-6	Ctenophora-sister	≈1	≈1	≈1	≈1
S&R-6	Porifera-sister	≈0	≈0	≈0	≈0
S&R-6	Ctenophora-sister	≈1	≈1	≈1	≈1
KGB-6	Porifera-sister	≈0	≈0	≈0	≈0
KGB-6	Ctenophora-sister	≈1	≈1	≈1	≈1
Model: CAT-GTR+G					
None	Porifera-sister	≈0	0.68	0.77	≈0
None	Ctenophora-sister	≈1	0.31	0.2	≈1
Dayhoff-6	Porifera-sister	0.89	0.99	0.95	0.74
Dayhoff-6	Ctenophora-sister	0.09	≈0	0.01	0.25
S&R-6	Porifera-sister	0.98	0.99	0.94	0.62
S&R-6	Ctenophora-sister	≈0	≈0	0.02	0.22
KGB-6	Porifera-sister	0.42	0.68	0.37	0.78
KGB-6	Ctenophora-sister	0.17	0.05	0.29	0.22

Support values are PPs. All analyses but the CAT-GTR+G analysis of the amino acid version of Chang converged well, and all convergence statistics are detailed in the captions of Figures S1–S4. Support values for Porifera-sister and Ctenophora-sister do not always add up to 1 because alternative topologies, e.g., monophyletic Porifera+Ctenophora sister to the remaining animals, are sometime recovered.

"Porifera-sister was favored under all recoding strategies"

https://github.com/caseywdunn/feuda_2017

This repository Search Pull requests Issues Marketplace Explore

caseywdunn / feuda_2017

Unwatch 1 Star 1 Fork 1

Code Issues 0 Pull requests 0 Projects 0 Wiki Insights Settings

Branch: master feuda_2017 / readme.md Find file Copy path

caseywdunn adjusted figure legend e656544 on Dec 13, 2017

1 contributor

174 lines (91 sloc) 24.4 KB Raw Blame History

How are ctenophores and sponges related to other animals?

Casey W. Dunn, Department of Ecology and Evolutionary Biology, Yale University, http://dunnlab.org

There has been great interest in resolving the deepest evolutionary relationships in the animal tree of life. This comes down to resolving the relationships between five clades - Cnidaria, Bilateria, Placozoa, Porifera (sponges), and Ctenophora (comb jellies). The deepest split in the animal tree has been of particular interest, with most recent studies finding either Ctenophora or Porifera to be the sister group to all other animals. These are referred to as the Ctenophora-sister and Porifera-sister hypotheses (Figure 1).

Table 3. Support for Porifera-Sister and Ctenophora-Sister from the WhelanD20-Opistho, WhelanD20-Holo, WhelanD20-Choano, and Chang Datasets

		Dataset			
Recoding	Hypothesis	WhelanD20-Opistho	WhelanD20-Holo	WhelanD20-Choano	Chang
Model: GTR+G					
None	Porifera-sister	≈0	≈0	≈0	≈0
None	Ctenophora-sister	≈1	≈1	≈1	≈1
Dayhoff-6	Porifera-sister	≈0	≈0	≈0	≈0
Dayhoff-6	Ctenophora-sister	≈1	≈1	≈1	≈1
S&R-6	Porifera-sister	≈0	≈0	≈0	≈0
S&R-6	Ctenophora-sister	≈1	≈1	≈1	≈1
KGB-6	Porifera-sister	≈0	≈0	≈0	≈0
KGB-6	Ctenophora-sister	≈1	≈1	≈1	≈1
Model: CAT-GTR+G					
None	Porifera-sister	≈0	0.68	0.77	≈0
None	Ctenophora-sister	≈1	0.31	0.2	≈1
Dayhoff-6	Porifera-sister	0.89	0.99	0.95	0.74
Dayhoff-6	Ctenophora-sister	0.09	≈0	0.01	0.25
S&R-6	Porifera-sister	0.98	0.99	0.94	0.62
S&R-6	Ctenophora-sister	≈0	≈0	0.02	0.22
KGB-6	Porifera-sister	0.42	0.68	0.37	0.78
KGB-6	Ctenophora-sister	0.17	0.05	0.29	0.22

Support values are PPs. All analyses but the CAT-GTR+G analysis of the amino acid version of Chang converged well, and all convergence statistics are detailed in the captions of Figures S1–S4. Support values for Porifera-sister and Ctenophora-sister do not always add up to 1 because alternative topologies, e.g., monophyletic Porifera+Ctenophora sister to the remaining animals, are sometime recovered.

Significance of posterior probabilities for Porifera-sister assessed as posterior probability ≥ 0.95.

“Porifera-sister was favored under all recoding strategies.”

“As modeling improves, sponges emerge in this position instead, indicating that trees placing the comb jellies at the root of the animals are artifactual.”

"CAT-GTR+G… describes site-specific amino acid preferences much better than any other model."

Animal Root Support
CAT-GTR+G
Chang-Holo
Whelan-Opis
Whelan-Holo
Whelan-Cho
Dataset
None
K
D
S
Recoding
Result
Ctenophora-sister
Porifera-sister
Unresolved

Posterior predictive analyses…

Posterior predictive analyses

Assess the adequacy of the model to predict the data.

- ppa_div Ability to predict the mean number of states at each site.

- ppa_conv Ability to predict the probability of convergence toward the same state at a given site in different taxa.

- ppa_var Ability to predict compositional heterogeneity of states across sites.

- ppa_max Ability to predict compositional heterogeneity of states across taxa.

- ppa_mean Ability to predict compositional heterogeneity of states across taxa.

Table 1. Comparing Model Adequacy

Posterior Predictive Test

Recoding	Model	Site-Specific Amino Acid Preferences (PPA-DIV)	Across-Taxa Compositional Heterogeneity (PPA-MAX)
Dataset: WhelanD20-Opistho			
none	WAG	127.35	79.99
none	LG	116.17	78.09
none	GTR	92.62	68.99
none	PF-scheme	104.71	44.75
none	CAT-GTR	6.21	34.78
Dayhoff-6	GTR	59.01	33.21
Dayhoff-6	CAT-GTR	−0.94	25.48
S&R-6	GTR	60.52	31.10
S&R-6	CAT-GTR	−1.09	7.09
KGB-6	GTR	56.82	33.13
KGB-6	CAT-GTR	−0.24	13.54
Dataset: Chang			
none	WAG	175.61	24.18
none	LG	161.28	19.00
none	GTR	126.33	18.84
none	PF-scheme	163.06	14.99
none	CAT-GTR	6.17	12.15
Dayhoff-6	GTR	78.06	26.99
Dayhoff-6	CAT-GTR	−1.71	3.64
S&R-6	GTR	87.37	28.48
S&R-6	CAT-GTR	−1.54	3.12
KGB-6	GTR	70.91	8.68
KGB-6	CAT-GTR	−0.71	1.21

“Increased model adequacy is associated with significant topological changes in support of Porifera-sister. “

Scores closer to zero indicate better adequacy

In no case does the only analysis with significant support for Porifera-sister have the best PPA score

Two of the five PPA scores are better for the dataset that hasn't been recoded.

How is recoding affecting the analyses?

The title of the paper…

“Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals”

…suggests that they model compositional heterogeneity.

The goal of recoding is to accommodate variation in composition across branches in the tree.

Rather than apply a model to describe this variation, recoding instead discards this variation.

The resulting models are far simpler, not more complex.

There are two possible reasons recoding can change phylogenetic results:

It successfully accommodates variation

It throws away information, and by removing useful information it changes the result

How can we test this?

Apply new recoding schemes that also reduce twenty states down to six, but are based on random grouping rather than empirical frequencies of amino acid exchange (like JTT).

How can we test this?

Empirical and random recodings both discard the same amount of information, but only empirical recoding reduce the impact of amino-acid frequency as intended.

(B) Analyses

SR6–published

Porifera Sister

PP site hetero

PP convergence

PP diversity

PP taxon hetero max

PP taxon hetero mean

(A) Alternative relationships
Ctenophora
Porifera
Placozoa
Bilateria
Cnidaria
Ctenophora-sister
Porifera
Ctenophora
Placozoa
Bilateria
Cnidaria
Porifera-sister
(B) Analyses
SR6–published
SR6–reproduced
Porifera Sister
posterior
1.00
0.75
0.50
0.25
0.00
20
states
6
PP site hetero
score
12
8
4
0
20
states
6
PP convergence
score
12
9
6
3
0
20
states
6
PP diversity
score
6
4
2
0
20
states
6
PP taxon hetero max
score
30
20
10
0
20
states
6
PP taxon hetero mean
score
100
50
0
20
states
6

Expectation if recoding impacts analyses by accommodating compositional heterogeneity

Expectation of recoding impacts analyses by discarding information

(A) Alternative relationships
Ctenophora
Porifera
Placozoa
Bilateria
Cnidaria
Ctenophora-sister
Porifera
Ctenophora
Placozoa
Bilateria
Cnidaria
Porifera-sister
(B) Analyses
SR6–published
SR6–reproduced
random–00
random–01
random–02
random–03
Porifera Sister
posterior
1.00
0.75
0.50
0.25
0.00
20
states
6
PP site hetero
score
12
8
4
0
20
states
6
PP convergence
score
12
9
6
3
0
20
states
6
PP diversity
score
6
4
2
0
20
states
6
PP taxon hetero max
score
30
20
10
0
20
states
6
PP taxon hetero mean
score
100
50
0
20
states
6

The results are very similar if you analyze the data with the SR6 recoding (based on the JTT recoding) or completely random recoding.

This suggests that recoding impacts their results by discarding information, not because they accommodate compositional variation across branches.

- The large amount of information discarded by recoding are a big cost

- It isn’t clear if any benefit for accommodating variation is worth the cost

What do we do to accommodate model variation across the tree?

We build models that ave parameters for it.

Where to now?

1. Improved analysis methods that model heterogeneity, not discard it.

2. Improved data acquisition.

Regardless of this remaining controversy, the rejection of traditional placements for ctenophores requires new hypotheses for evolution of some traits of broad interest.

Given the placement of Placozoa, there has been multiple gains or loss of the nervous system

Trends in Ecology & Evolution

Opinion

CellPress

Complex Homology and the Evolution of Nervous Systems

Benjamin J. Liebeskind,[1,2,3,*] David M. Hillis,[2,3,4]
Harold H. Zakon,[2,3,4,5,6,7] and Hans A. Hofmann[2,3,4,5,6]

http://dx.doi.org/10.1016/j.tree.2015.12.005

Current Biology

Evolution: Ctenophore Genomes and the Origin of Neurons

Heather Marlow and Detlev Arendt

http://dx.doi.org/10.1016/j.cub.2014.06.057

REVIEW

Journal of Experimental Biology

Convergent evolution of neural systems in ctenophores

Leonid L. Moroz*

http://jeb.biologists.org/cgi/doi/10.1242/jeb.110692

A shrinking number of better defined questions

Dunn et al, 2014

doi:10.1146/annurev-ecolsys-120213-091627

How our biases shape the way we study diversity

Ascertainment biases

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

Review

CellPress

The hidden biology of sponges and ctenophores

Casey W. Dunn[1], Sally P. Leys[2], and Steven H.D. Haddock[3]

[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.

+

≠

Monkey

Glue Cell

Ctenophore

Primate

Ctenophore

Primate Specific

Shared

Ctenophore Specific

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

NATIONAL GEOGRAPHIC

“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/

The evolution of gene expression

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)

Trait 2
Trait 1

Trait 2
n=2
Trait 1

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. 1.—Scatter diagram of hypothetical data from 8 species, showing the relationship between Y and X.

Felsenstein, 1985

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.

Felsenstein, 1985

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 $v_{13} = v_{14}$.

Felsenstein, 1985

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

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

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

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

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

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

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

Felsenstein, 1985

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.

Felsenstein, 1985

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.

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

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 Harrigan[4], Manuela Weier[1], Angélica Liechti[1], Ayinuer Aximu-Petri[5], Martin Kircher[5], Frank W. Albert[5]†, Ulrich Zeller[6], Philipp Khaitovich[7], Frank Grützner[8], Sven Bergmann[2,3], Rasmus Nielsen[4,9], Svante Pääbo[5] & Henrik Kaessmann[1,2]

BMC Genomics

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 Busby[1], Jesse M Gray[2,4], Allen M Costa[2], Chip Stewart[1,5], Michael P Stromberg[1], Derek Barnett[1], Jeffrey H Chuang[1], Michael Springer[3] 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. Cole[1], Eitan Winter[1], Natalia Mostov[1], Sally Khair[1], Naftalie Senderovich[1], Ekaterina Kovalev[1], David H. Silver[1], Martin Feder[1], Selene L. Fernandez-Valverde[2]†, Nagayasu Nakanishi[2]†, David Simmons[3], Oleg Simakov[4], Tomas Larsson[4], Shang-Yun Liu[5], Ayelet Jerafi-Vider[6], Karina Yaniv[6], Joseph F. Ryan[3], Mark Q. Martindale[3], Jochen C. Rink[5], Detlev Arendt[4], Sandie M. Degnan[2], Bernard M. Degnan[2], Tamar Hashimshony[1] & Itai Yanai[1]

PLOS | COMPUTATIONAL BIOLOGY

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 Daish[3], Ulrich Zeller[4], Julie C. Baker[5], Frank Grützner[3] & Henrik Kaessmann[1,2]

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

But there is a potential problem…

Most of these studies use Pairwise Comparisons

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

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

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

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

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

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

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

PLOS | COMPUTATIONAL BIOLOGY
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]*

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. Cole[1], Eitan Winter[1], Natalia Mostov[1], Sally Khair[1], Naftalie Senderovich[1], Ekaterina Kovalev[1], David H. Silver[1], Martin Feder[1], Selene L. Fernandez-Valverde[2]†, Nagayasu Nakanishi[2]†, David Simmons[3], Oleg Simakov[4], Tomas Larsson[4], Shang-Yun Liu[5], Ayelet Jerafi-Vider[6], Karina Yaniv[6], Joseph F. Ryan[3], Mark Q. Martindale[3], Jochen C. Rink[5], Detlev Arendt[4], Sandie M. Degnan[2], Bernard M. Degnan[2], Tamar Hashimshony[1] & Itai Yanai[1]

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

Cold Spring Harbor Laboratory
bioRχiv beta
THE PREPRINT SERVER FOR BIOLOGY

Pairwise comparisons are problematic when analyzing functional genomic data across species

Casey W. Dunn[1]*, Felipe Zapata[1,2], Catriona Munro[1], Stefan Siebert[1,3], Andreas Hejnol[4]

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

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

https://github.com/caseywdunn/comparative_expression_2017

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

PLOS | COMPUTATIONAL BIOLOGY
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]*

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

PLOS | COMPUTATIONAL BIOLOGY 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]*

This paper is a test of the ortholog conjecture.

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

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.

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

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

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

The ortholog conjecture has received mixed support.

PLOS | COMPUTATIONAL BIOLOGY
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]*

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

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

PLOS | COMPUTATIONAL BIOLOGY 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]*

They examined:

- Tissue specificity of gene expression as quantified by τ (where τ=0 is uniform expression, τ=1is expression entirely specific to one tissue).

- Across multiple public datasets of mammal organ gene expression

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

PLOS | COMPUTATIONAL BIOLOGY 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]*

They used pairwise comparisons across genes and species.

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

PLOS | COMPUTATIONAL BIOLOGY 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]*

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

We first verified that we can reproduce their result.

Their plot

Our plot

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

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.

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

Prediction

Null expectation

Ortholog Conjecture expectation

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

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

Pairwise

Phylogenetic

Empirical

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

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

What is this striking pattern?

PLOS | COMPUTATIONAL BIOLOGY 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]*

Phil. Trans. R. Soc. Lond. B **326**, 119–157 (1989) [119]
Printed in Great Britain

THE PHYLOGENETIC REGRESSION

By A. GRAFEN†
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.

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

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

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

For example…

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

Cartoons

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

Support

Waterman Award, Division of Environmental Biology

iPlant Collaborative

The Manning Chair

RI EPSCoR

The Lab…