Big Data in Phylogenetics
Dealing with the deluge

January 2003
Growth of GenBank
(1982 - 2005)

Sequences (millions)

Base Pairs of DNA (billions)

Growth of Sequences & Databases

- Base pairs in GenBank
- Databases in NAR

Plot showing the increase in base pairs and databases over time.
The era of big data in biology

Stein, Genome Biology, 2010
Growth of dbSNP, 2002-2009

- Submissions
- Ref. Clusters
- Validated

Number of SNPs (millions)

Timeline:
- TSC / BAC Overlap
- HapMap
- Phase I
- Phase II
- 1000 Genomes
- NGS
NextGen Sequencing a Game-Changer

- NGS (bp/$) Doubling time 5 months
- Hard disk storage (MB/$) Doubling time 14 months
- Pre-NGS (bp/$) Doubling time 19 months
Growth of dbSNP (2003-2009)

- Ref SNPs
- Validated

Number of SNPs vs dbSNP Release

- X-axis: dbSNP Release
- Y-axis: Number of SNPs

Data points:
- Jun 03, Aug 03, Oct 03, Nov 03, Jan 04, Mar 04, Jun 04, Aug 04, Nov 04, Jan 05, Oct 05, May 05, Mar 06, Oct 06, Apr 07, May 07, Apr 08, May 09
Productivity in DNA Synthesis and Sequencing Using Commercially Available Instruments

Rob Carlson, February 2013, www.synthesis.cc

- Number of transistors per chip
- Productivity: Reading DNA
- Productivity: Writing DNA


Synthesis and Sequences Productivity [bases/person/day]: 1.0E+11, 1.0E+10, 1.0E+9, 1.0E+8, 1.0E+7, 1.0E+6, 1.0E+5, 1.0E+4, 1.0E+3, 1.0E+2, 1.0E+1

Number of transistors per chip: 1.0E+11, 1.0E+10, 1.0E+9, 1.0E+8, 1.0E+7, 1.0E+6, 1.0E+5, 1.0E+4, 1.0E+3, 1.0E+2, 1.0E+1
The graph shows the productivity of Transistors on Chip, Reading DNA, and Writing DNA over the years from 1970 to 2010. The productivity is represented on a logarithmic scale along the y-axis, with Transistors on Chip depicted in blue dashes, Reading DNA in green, and Writing DNA in red. The graph indicates a steady increase in productivity for all three categories, with Transistors on Chip showing the most rapid increase, followed by Reading DNA and then Writing DNA.
The graph illustrates the increase in kilobases per day per machine over the years from 1980 to 2010 and into the future. Key stages in the development of sequencing technologies are highlighted:

- **Gel-based systems**: Manual slab gel, Automated slab gel
- **Capillary sequencing**: First-generation capillary
- **Massively parallel sequencing**
- **Short-read sequencers**: Microwell pyrosequencing
- **Second-generation capillary sequencer**

The timeline extends into the future, suggesting continuous improvement and innovation in sequencing technologies.
Cost to sequence a human genome (USD)
a. Vertebrate sequences in GenBank (millions)

\[ R^2 = 0.968 \]

b. Phylogenetic papers in Web of Science (thousands)

\[ R^2 = 0.999 \]

c. Phylogenetic resolution

\[ R^2 = 0.998 \]
As the cost of DNA sequencing falls, the growth of human genome data becomes exponential.
Number of Whole Genome Sequencing Projects in GenBank Database Over Time
Big Data: Astronomical or Genomical?

Growth of DNA Sequencing

- Recorded growth
- Double every 7 months (Historical growth rate)
- Double every 12 months (Illumina Estimate)
- Double every 18 months (Moore’s Law)

1e+09
1e+06
1e+03
1e+00

Cumulative Number of Human Genomes

Worldwide Annual Sequencing Capacity


1 Pbp
1 Ebp
1 Zbp

1st Sanger
IHGSC et al.
Venter et al.

1st Personal Genome
Levy et al.

1st 454
Wheeler et al.

1st 1000 Genomes

1st Illumina
Bentley et al.
Wang et al.
Ley et al.

1st PacBio
Chaisson et al.

Current Capacity

Stephens et al. 2015
To what end?

“We think big data is what everyone cares about. It’s not. It’s stories.”

- Dr. Jessica Utts
  President, American Statistical Association

The goal is to gather ‘sufficient’ data in order to answer a question ‘robustly.’
To what end?

“We think big data is what everyone cares about. It’s not. It’s stories.”

- Dr. Jessica Utts
  President, American Statistical Association

The goal is to gather ‘sufficient’ data in order to answer a question ‘robustly.’ The question is what is interesting.
To what end?

“We think big data is what everyone cares about. It’s not. It’s stories.”

- Dr. Jessica Utts
  President, American Statistical Association

The goal is to gather ‘sufficient’ data in order to answer a question ‘robustly.’ The question is what is interesting. This is no different than it’s always been.
A case study

- A very specific question:
  - What are the phylogenetic affinities of turtles?

- Brings up more general issues:
  - How do we approach difficult phylogenetic problems?
  - How *should* we approach difficult phylogenetic problems?
Turtle Phylogenetics

- Overarching problem:
  - Where do turtles sit in the amniote tree?

![Phylogenetic tree with mammals, archosaurs, and lepidosaurs]
Osteology

- Early approaches relied on osteology (primarily of the skull)

Anapsid

Diapsid

Synapsid
Osteology

- Early approaches relied on osteology (primarily of the skull).
  
  Anapsid
  
  Diapsid
  
  Synapsid
Early approaches relied on osteology (primarily of the skull)

Günther 1867, Gaffney 1980
Osteology

- Primary issue with this hypothesis
More osteology

Gaffney 1980

Lee 2001 Contrib. to Zool


Reippel and deBraga 1996 Nature
Molecular Information

- Mitochondrial data

  - Zardoya and Meyer 1998 PNAS
  - Strimmer and Von Haeseler 1996 MBE
  - Cao et al. 1998 MBE
  - Hedges 1994 PNAS
Molecular Information

- Nuclear data

Iwabe et al. 2004 MBE

Hugall et al. 2007 Syst Biol

Hedges and Polling 1998 Science
Summary
Turtle Genomics

- 3 genome consortia
- Several more independent studies
Phylogenomics

Shaffer et al. 2013 Genome Biol
Phylogenomics

- All analyses agree!
MicroRNAs support a turtle + lizard clade
Summary

- Ugh... so what do we do?
Summary

- Ugh...so what do we do?
Data in Phylogenetics

- Let’s take a step back.
- How have we been approaching this (and most other) phylogenetic questions?

4 nuclear genes

- Turtles
  - Crocodilians
  - Birds
  - Tuataras
  - Squamates
  - Primates
  - Rodents

11 nuclear genes

- Turtles
  - Crocodilians
  - Birds
  - Squamates
  - Primates
  - Rodents

Hedges and Polling 1998 Science
Let's take a step back.

How have we been approaching this (and most other) phylogenetic questions?

A data centric view

Hedges and Polling 1998 Science
Phylogenomics

- Inferences result from both data and the model

\[
\text{Prob}(\text{TTGGT} | \text{Frequencies} = (A, C, T, G))
\]
Why does this matter?

- In developing a statistical model for a problem, we inevitably make a tradeoff.
Why does this matter?

- In developing a statistical model for a problem, we inevitably make a tradeoff

![Graph showing the tradeoff between bias and variance with increasing number of parameters and amount of data.](image)
Why does this matter?

- In developing a statistical model for a problem, we inevitably make a tradeoff
Why does this matter?

- In developing a statistical model for a problem, we inevitably make a tradeoff 

![Graph showing the relationship between Bias, Variance, Number of Parameters, and Power with increasingAmount of Data.]

- 1 gene
- 10 genes
- 1000 genes
Why does this matter?

- In developing a statistical model for a problem, we inevitably make a tradeoff.

![Graph showing the tradeoff between Bias and Variance with Number of Parameters and Power vs Amount of Data.]

- 1 gene
- 10 genes
- 1000 genes
Why does this matter?

- The point.

Our data centric view focuses on this.
Why does this matter?

- The point.

But this is our bigger problem.
How do we know it’s a bigger problem?

- Osteology
- mtDNA
- nuDNA
- Phylogenomic
How do we know it’s a bigger problem?

Where’s the disagreement coming from?
How do we know it’s a bigger problem?
‘Big data’ turtle studies

- Chiari et al. (2012)
  - 248 transcriptomic loci
  - 12 taxa

- Crawford et al. (2012)
  - 1,145 UCEs
  - 10 taxa

- Fong et al. (2012)
  - 75 Sanger-sequenced loci
  - 129 taxa

- Lu et al. (2013)
  - 1,638 transcriptomic and genomic loci
  - 11 taxa

- Shaffer et al. (2013)
  - 1,955 genomic loci
  - 8 taxa

- Wang et al. (2013)
  - 1,113 genomic loci
  - 12 taxa
Bipartition Bayes Factors

Marginal likelihood with \( AB \mid CDE \)

Marginal likelihood without \( AB \mid CDE \)

Bayes Factor
A Note on Extreme Probabilities

Archosaur + Turtle Monophyly

Shaffer et al.

26% of genes

Brown and Thomson 2017
A Note on Extreme Probabilities

Archosaur + Turtle Monophyly

Shaffer et al.

The probability of observing these sequences is 100 fold lower if monophyly is true.
A Note on Extreme Probabilities

Archosaur + Turtle Monophyly

Shaffer et al.

The probability of observing these sequences is 1,000,000,000,000,000,000,000,000,000 fold lower.
A Note on Extreme Probabilities

1/1,000,000,000,000,000,000,000,000,000,000

That’s 27 zeroes!

If you played a lottery every minute with that chance of winning, you still probably wouldn’t win, unless you played for...

the age of the universe \* 190,258,751,903
A Note on Extreme Probabilities

Archosaur + Turtle Monophyly

Shaffer et al.

The probability of observing these sequences is $1,000,000,000,000,000,000,000,000,000$ fold lower.
Support varies across branches of the tree
Support varies across branches of the tree
Support varies across branches of the tree

All of these strongly supported by this dataset.

PP = 1.0

Brown and Thomson 2017
Support varies across branches of the tree

Equivocation about turtle placement across genes
Support varies across branches of the tree

This dataset supports turtles as sister to crocodilians. But what’s up with these outliers? How influential are they?

Brown and Thomson 2017
Both look like paralogs

Probable Gene Duplications

Clade 1

Clade 2

Clade 3

Original sequences from the Chiari et al. alignments are denoted with "_query". Note that two or three hits were returned from each reference genome for each query sequence. Sequences form roughly three clades with no more than one sequence per species in each clade, but we did not recover a representative sequence for each species in each clade in our preliminary search. Regardless, the structure of the tree clearly shows that query sequences are not monophyletic and are unlikely to be orthologous. Note that searches using the Emys, Chelonoidis, and Phrynops query sequences were all conducted on the Chrysemys picta reference genome, so many of the hits they returned were identical.

Brown and Thomson 2017
Strong influence

Uncertainties in the vertebrate phylogeny examined in this study. (A) The five alternative hypotheses for the placement of turtles within amniotes 1) turtles as basal amniotes, 2) turtles as basal sauropsids, 3) turtle-lepidosaur sister group, 4) turtle-archosaur sister group, and 5) turtle-crocodilian sister group. (B) monophyletic and (C) paraphyletic alternative hypotheses for lissamphibian (extant amphibians) relationships.

Chiari et al.

Brown and Thomson 2017
Figure 1. Alternative hypotheses in the vertebrate phylogeny. (A) The five alternative hypotheses for the placement of turtles within amniotes: 1) turtles as basal amniotes, 2) turtles as basal sauropsids, 3) turtle-lepidosaur sister group, 4) turtle-archosaur sister group, and 5) turtle-crocodilian sister group. (B) Monophyletic and (C) paraphyletic alternative hypotheses for lissamphibian (extant amphibians) relationships.

Chiari et al. without 2 outliers 246/248 genes remain

Strong influence

Phylogenomics of the Vertebrate Phylogeny

PLOS ONE | www.plosone.org 2 November 2012 | Volume 7 | Issue 11 | e48990

Chiari et al.
without 2 outliers
246/248 genes remain

Brown and Thomson 2017
Strong influence

Chiari et al. without 2 outliers
246/248 genes remain

An identical switch occurs for the dataset of Wang et al. when removing ~1% of genes.
A troubling, but common, result

- More recent papers build on this result and find similar patterns:

New Results

**Site and gene-wise likelihoods unmask influential outliers in phylogenomic analyses**

Joseph F. Walker, Joseph W. Brown, Stephen A. Smith
doi: https://doi.org/10.1101/115774

Contentious relationships in phylogenomic studies can be driven by a handful of genes

Xing-Xing Shen, Chris Todd Hittinger & Antonis Rokas

Inadvertent Paralog Inclusion Drives Artifactual Topologies and Timetree Estimates in Phylogenomics

Karen Siu-Ting*,1,2,3 María Torres-Sánchez‡,4 Diego San Mauro,4 David Wilcockson,5 Mark Wilkinson,6 Davide Pisani, Mary J. O’Connell,8,9 and Christopher J. Creevey*1
Take homes

- More data does not necessarily lead to more accuracy, or to consensus.
- A lot of phylogenomic progress is actually about figuring out how to model data well, not collecting more data per se.
Some Possible Ways Forward

- *Embrace* the computational **challenge**
Embrace the computation

- Analyses need not finish quickly
- Advances in computation help a lot here
  - parallel architectures and code
  - fast computation libraries
  - availability of compute resources
  - continual methodological improvement
Embrace the computation
**Embrace the computation**

---

### Compute Resources

<table>
<thead>
<tr>
<th>Name</th>
<th>Status</th>
<th>CPUs</th>
<th>Peak TFlops</th>
<th>Utilization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stampede2</td>
<td>Healthy</td>
<td>368280</td>
<td>12800.0</td>
<td>93%</td>
</tr>
<tr>
<td>Comet</td>
<td>Healthy</td>
<td>46752</td>
<td>2000.0</td>
<td>52%</td>
</tr>
<tr>
<td>SuperMIC</td>
<td>Healthy</td>
<td>7200</td>
<td>925.0</td>
<td>10%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Running Jobs</th>
<th>Queued Jobs</th>
<th>Other Jobs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stampede2</td>
<td>1006</td>
<td>900</td>
<td>381</td>
</tr>
<tr>
<td>Comet</td>
<td>1866</td>
<td>16</td>
<td>82</td>
</tr>
<tr>
<td>SuperMIC</td>
<td>56</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

[xsede.org]
Embrace the computation
XSEDE users by field for 2016

Number of Users: Active: by Field of Science

1. Materials Research: 379
2. Biophysics: 453
3. Advanced Scientific Computing: 665
4. Chemistry: 453
5. Computer and Computation Research: 263
6. Training: 239
7. Biochemistry and Molecular Structure: 215
8. Computer and Information Science and Applications: 201
10. Physical Chemistry: 155
11. Avg of 110 others: 85

Systematic and Population Biology
New Tools on the Horizon

- More complex models
- More efficient sampling. e.g., Hamiltonian Monte Carlo
- More efficient implementations of existing methods
Some Possible Ways Forward

- **Embrace** the computational **challenge**
- **Get very picky** about our data. Careful and detailed data exploration is your friend.
Some Possible Ways Forward

- **Embrace** the computational challenge
- **Get very picky** about our data. Careful and detailed data exploration is your friend.
- **Carefully consider tradeoffs** between speed and approximation

---

Leaché and Rannala 2010