'Open' as a strategy for durability, reproducibility, and scalability

Jonathan A. Rees
National Evolutionary Synthesis Center

BOSC
12 July 2014

Outline

- The challenge
- Open Tree of Life
- Threats
- 'Open' as defense
- General remarks

- The challenge

Lineage (full): root; cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Sauropsida; Sauria; Testudines + Archosauria group; Archosauria; Dinosauria; Saurischia; Theropoda; Coelurosauria; Aves; Neognathae; Galliformes; Phasianidae

---

- **Phasianinae** *Click on organism name to get more information.*
  - **Afropavo**
    - **Afropavo congensis** (Congo peafowl)
  - **Argusianus**
    - **Argusianus argus** (great argus)
      - **Argusianus argus grayi**
  - **Catreus**
    - **Catreus wallichii** (cheer pheasant)
  - **Chrysolophus**
    - **Chrysolophus amherstiae** (Lady Amherst's pheasant)
    - **Chrysolophus pictus** (golden pheasant)
      - **Chrysolophus pictus infuscatus**
      - **Chrysolophus pictus obscurus**
  - **Crossoptilon**
    - **Crossoptilon auritum** (blue eared-pheasant)
    - **Crossoptilon crossoptilon** (white-eared pheasant)
    - **Crossoptilon drouyni**
    - **Crossoptilon harmani** (Tibetan eared-pheasant)
    - **Crossoptilon mantchuricum** (brown eared-pheasant)
  - **Gallus**
    - **Gallus gallus** (chicken)
      - **Gallus gallus bankiva**
      - **Gallus gallus gallus**
      - **Gallus gallus jabouillei**
      - **Gallus gallus murghi**
      - **Gallus gallus spadiceus**
    - **Gallus lafayetii** (Ceylon junglefowl)
    - **Gallus sonneratii** (gray junglefowl)
    - **Gallus varius** (green junglefowl)
    - **Gallus sp.**

Problems

- Information about tree of life is hard to find
- Technical incompatibilities make integration hard (e.g. names / identifiers)
- Scientific differences are rampant but hard to see
- No single view on what is known

-
- Open Tree of Life

Editing study Wang, 2013

http://dx.doi.org/10.1371/journal.pone.0064312

Study quality 100% (show details)

Publication reference	**Wang, Ning, Rebecca T. Kimball, Edward L. Braun, Bin Liang, Zhengwang Zhang. 2013. Assessing phylogenetic relationships among Galliformes: a multigene phylogeny with expanded taxon sampling in Phasianidae. PLoS ONE 8 (5): e64312.**
Publication DOI (or URL)	**http://dx.doi.org/10.1371/journal.pone.0064312**
Study Year	**2013**
Focal clade	**Galliformes**
Tags	**ingroup added;**
Submitted by	**Joseph Brown**

This study should contribute to synthesis.

This study has changed since building the last synthetic tree.

Curator notes ***None***

Hide

This tool is a friendly visual editor for curating NexSON files. Like the NeXML and NEXUS formats from which it evolved, each NexSON file contains information about a single published study, with a particular focus on its trees and OTUs.

Each tab in this curation tool manages a different aspect of the study data. Watch for prompts—they look like 2 *—in each tab that list areas ripe for improvement.*

*Here in the **Metadata** tab, we're primarily concerned with clearly identifying this study and making it easy to find using standard publication references, DOIs, and free-form tags.*

Working with study data in other tools

You can **download this study** in different forms:

NexSON | NeXML | NEXUS | Trees as Newick

Editing study Wang, 2013

http://dx.doi.org/10.1371/journal.pone.0064312

Study quality 100% (show detai

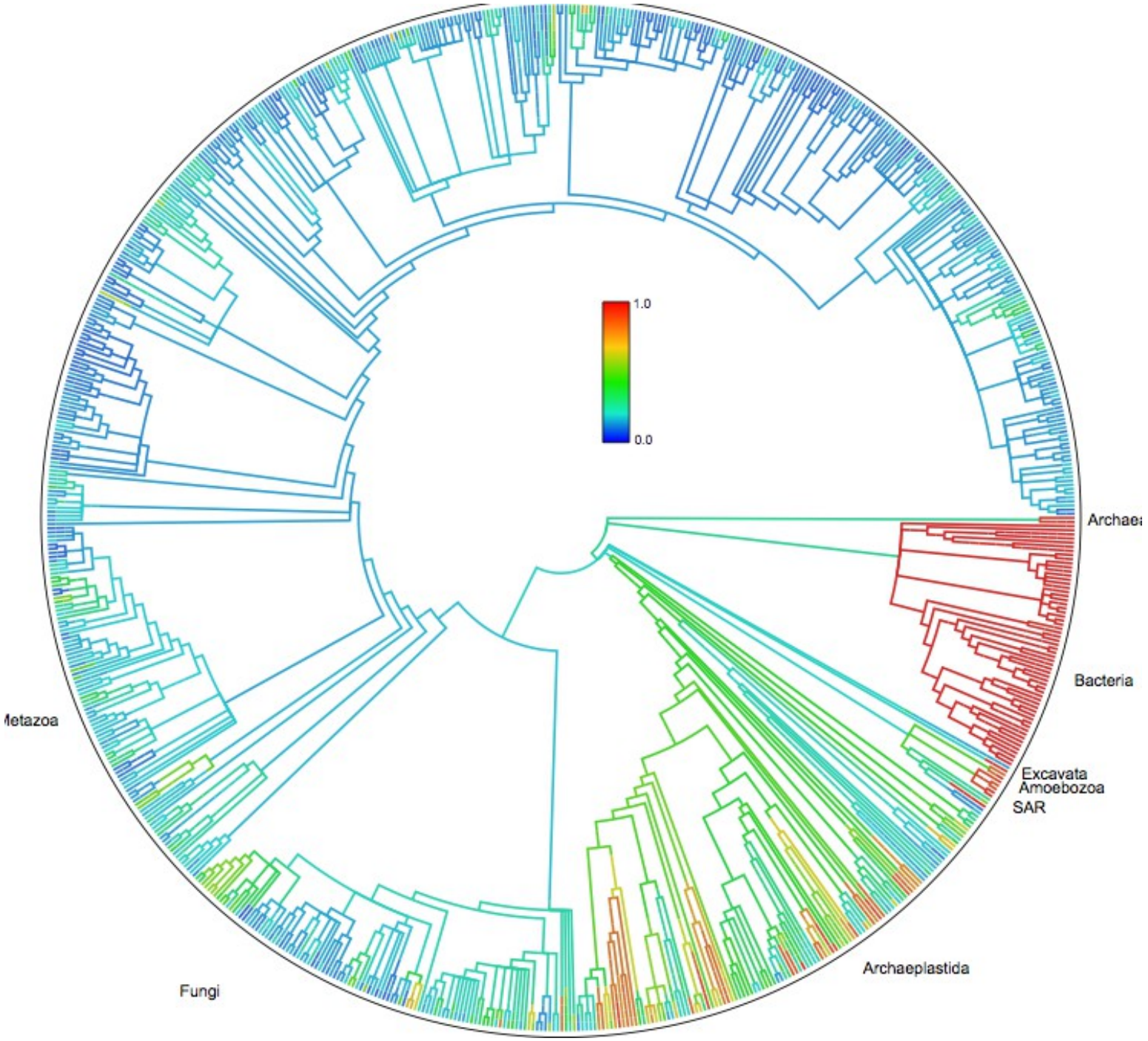
Filter by original or mapped label | In all trees | Unmapped OTUs first

Original label	Taxon label in OTT
Alectura lathami	**Alectura lathami**
Leipoa ocellata	**Leipoa ocellata**
Megapodius layardi	**Megapodius layardi**
Megapodius eremita	**Megapodius eremita**
Crax rubra	**Crax rubra**
Crax alector	**Crax alector**
Ortalis vetula	**Ortalis vetula**
Guttera pucherani	**Guttera pucherani**
Acryllium vulturinum	**Acryllium vulturinum**
Numida meleagris	**Numida meleagris**
Colinus virginianus	**Colinus virginianus**
Colinus cristatus	**Colinus cristatus**
Oreortyx pictus	**Oreortyx pictus**
Cyrtonyx montezumae	**Cyrtonyx montezumae**

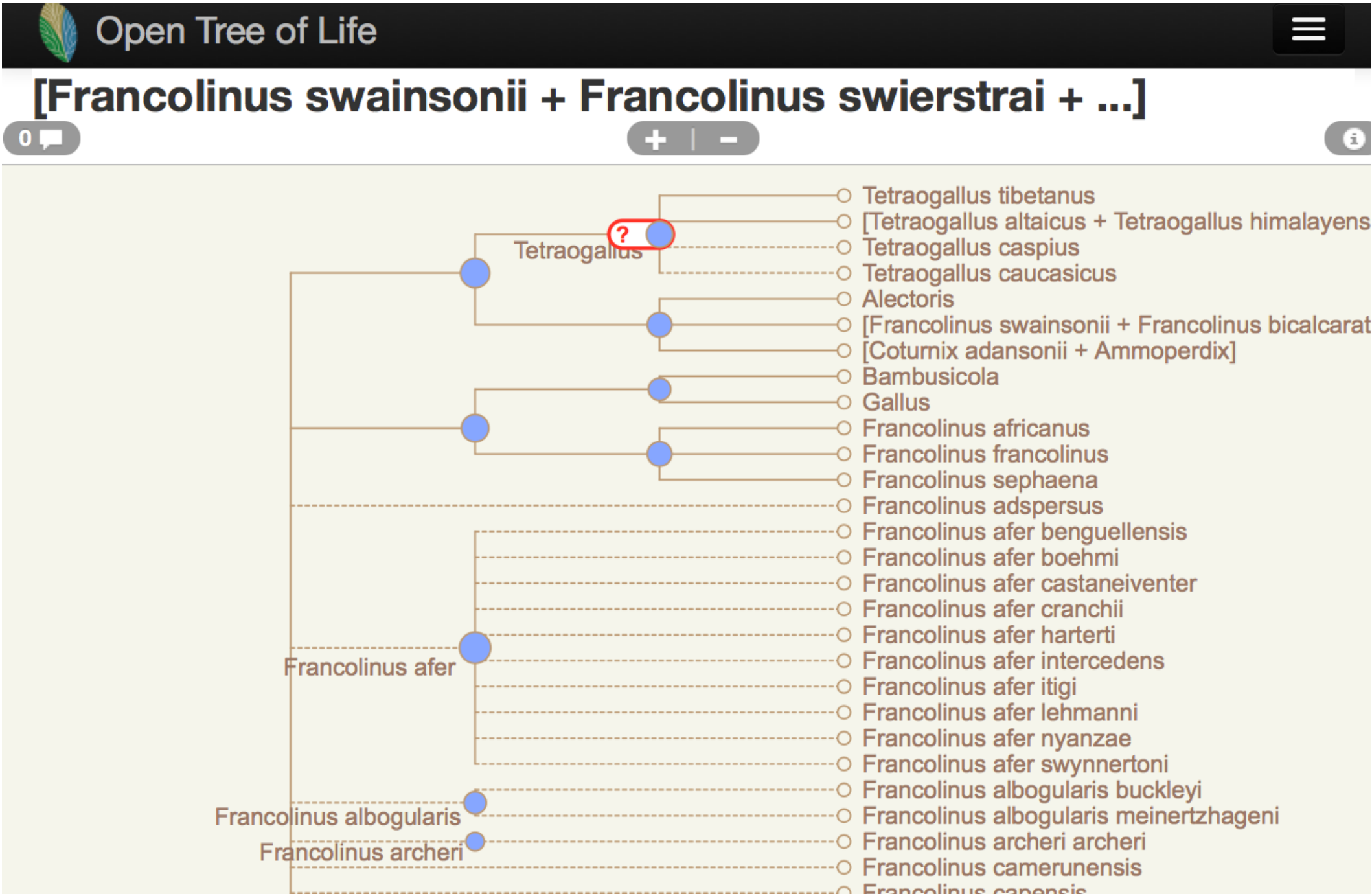
Hide

For a tree in your study to contribute to synthesis in the Open Tree of Life project, its leaf nodes should correspond to taxa in the OTT (Open Tree Taxonomy). This list includes all the OTUs (operating taxonomic units) in your study. Each OTU is a distinct label used in one or more of your study's trees.

We use a TNRS (taxonomic name resolution service) to automate this "mapping" process where possible. If you've used standard taxonomic names for the nodes in your trees, mapping is fast and easy. But most trees submitted have been labeled with institutional shorthand, unrecognized taxa, or simple misspellings. The tools on this page will help you adjust labels for faster mapping.



Stephen Smith


Open Tree of Life
[Francolinus swainsonii + Francolinus swierstrai + ...]
0
Tetraogallus tibetanus
[Tetraogallus altaicus + Tetraogallus himalayens
Tetraogallus caspius
Tetraogallus caucasicus
Alectoris
[Francolinus swainsonii + Francolinus bicalcarat
[Coturnix adansonii + Ammoperdix]
Bambusicola
Gallus
Francolinus africanus
Francolinus francolinus
Francolinus sephaena
Francolinus adspersus
Francolinus afer benguellensis
Francolinus afer boehmi
Francolinus afer castaneiventer
Francolinus afer cranchii
Francolinus afer harterti
Francolinus afer intercedens
Francolinus afer itigi
Francolinus afer lehmanni
Francolinus afer nyanzae
Francolinus afer swynnertoni
Francolinus albogularis buckleyi
Francolinus albogularis meinertzhageni
Francolinus archeri archeri
Francolinus camerunensis
Tetraogallus
Francolinus afer
Francolinus albogularis
Francolinus archeri

- Name gathering (taxonomy)
- Study upload
- Curation interface (OTU matching etc.)
- Deposit to study repository
- Synthesis
- Access (browser & API)

But…

- Trees are hard to obtain (Drew et al. 2013)
  - ~4% relatively easily available (Treebase, Dryad)
  - Another ~12% available in response to email
  - Others available only as JPEGs (cf. Mounce)
- Lots of manual labor (curation)
  - Obtaining trees
  - Ingroup and outgroup
  - OTU matching

-
-
- Threats

Threats (1)

- Under- and non-funding
- Software and/or data could be lost or captured
  - volunteers may not want to invest if they think work might be lost or captured
- Software could become brittle and unimprovable
  - volunteers may not want to invest if they think the data is 'captured' by the software

Threats (2)

- Scientific decision making could be opaque
  - volunteers could turn away if system doesn't make sense or if its claims are hard to confirm

-
-
-
- ‘Open’ as defense

Legal 'open'

- Hybrid CC0 + 'facts are free' (open data)
  - aids persistence (link to Dryad)
  - makes corpus more valuable
- Free software (including marked Javascript)
  - aids persistence
  - encourages experiments
- Open access publications (CC-BY)

Technical 'open'

- Data as JSON
  - 'open' to larger part of ecosystem
- Trees as NeXML (in JSON syntax)
- Data on github ('phylesystem' repo + index)
  - what you see is everything - not locked up in a database
  - helps assure community that data won't be lost
- Scripting

- tree.opentreeoflife.org
- opentreeoflife.org
- github.com/opentreeoflife
- twitter.com/opentreeoflife
- freenode #opentreeoflife
- opentreeoflife google group
- opentreeoflife-software google group

Process 'open'

- Open process ('open science')
  - public issue trackers
  - public discussion groups
  - routing user feedback to github issue tracker
- Reproducibility is another kind of 'open'
  - hard to reproduce → not so open
  - links to source material
  - scripting

-
-
-
-
- General remarks

copy

interpret

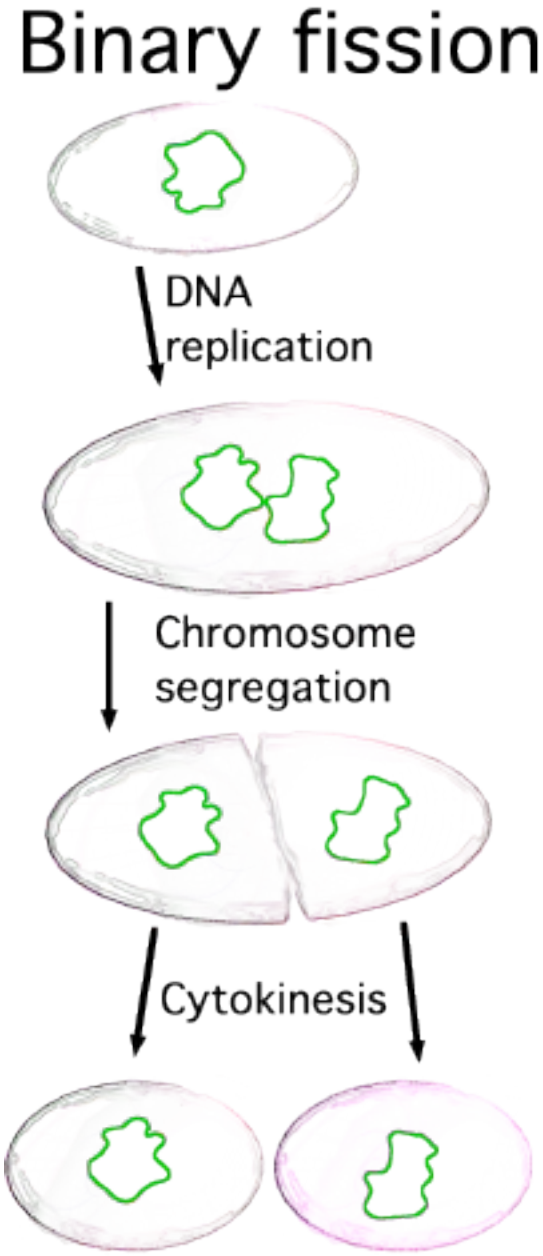
disseminate

compete

change

merge

life = free software ?



- I hereby acknowledge technical debt

Funding: NSF

Software:
Jim Allman
Joseph Brown
Karen Cranston
Cody Hinchliff
Mark Holder
Jonathan Leto
Emily McTavish
Peter Midford
Rick Ree
Jonathan Rees
Stephen Smith

PIs:
Gordon Burleigh
Keith Crandall
Karen Cranston
Karl Gude
David Hibbett
Mark Holder
Laura Katz
Rick Ree
Stephen Smith
Doug Soltis
Tiffani Williams

Curators:
Joseph Brown
Bryan Drew
Romina Gazis
Jessica Grant
Cody Hinchliff
Dail Laughinghouse
Chris Owen