Phylogeny-based methods for analysing genomes and metagenomes

Presented by Aaron Darling

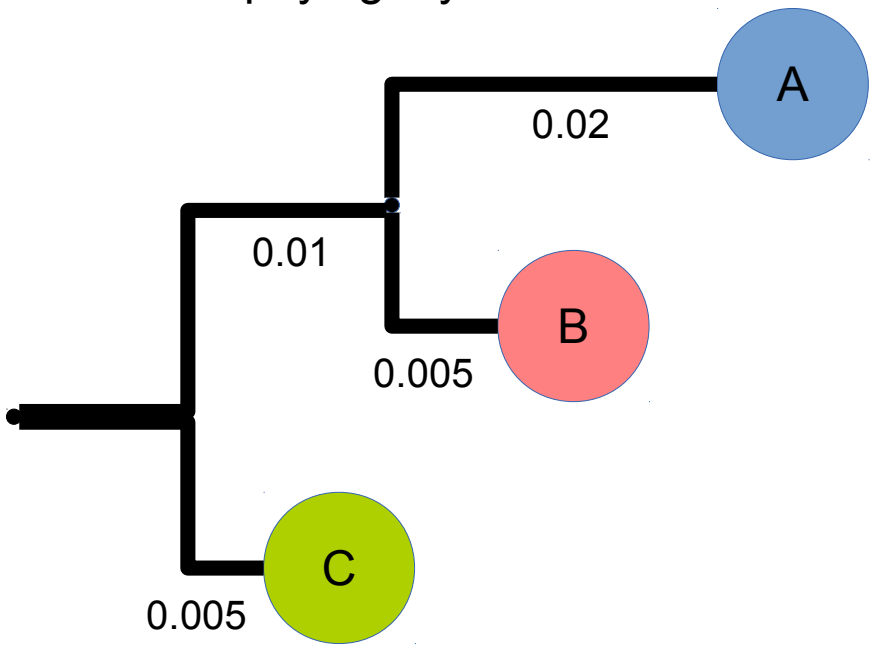
OTUs for ecology

Operational Taxonomic Unit: a grouping of similar sequences that can be treated as a single “species”

- Strengths
  - Conceptually simple
  - Mask effect of poor quality data
    - Sequencing error
    - *in vitro* recombination
- Weaknesses
  - Limited resolution
  - Logically inconsistent definition

Logical inconsistency: OTUs at 97% ID

Assume the true phylogeny:

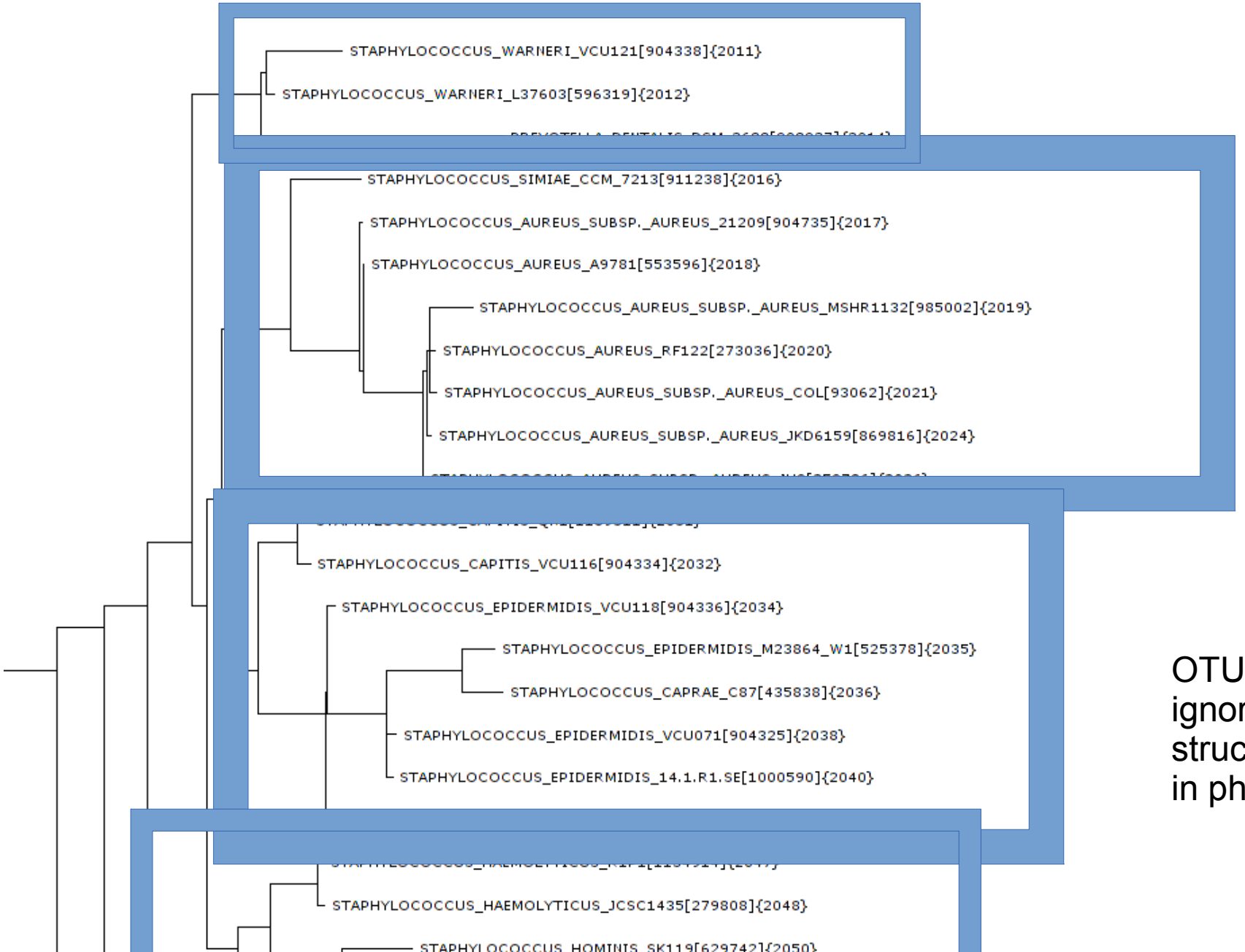


A, B > 97% identity
B, C > 97% identity
A and C not > 97% ID

Possible valid OTUs:
AB, C (with A & C centroids)
A, BC (with A & C centroids)
ABC (with B centroid)

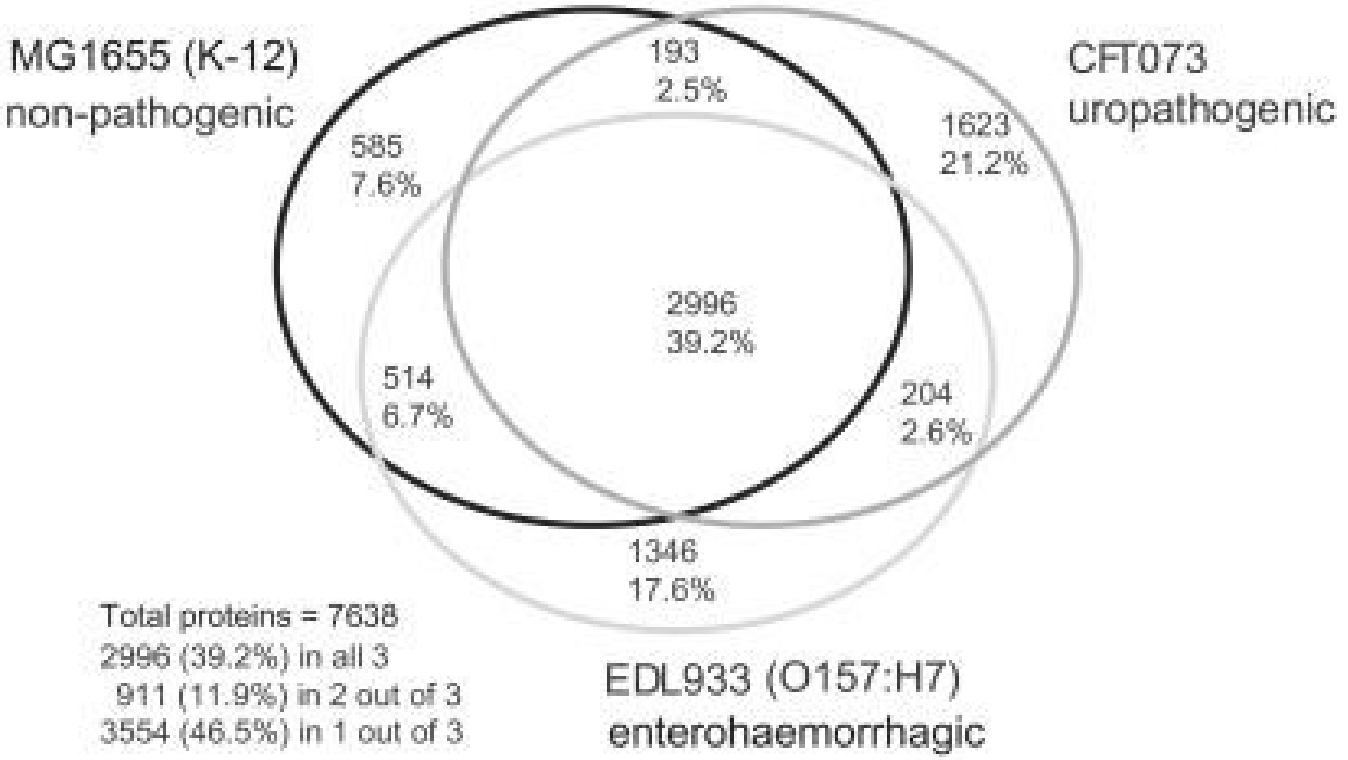
OTU pipelines will arbitrarily pick one of the three solutions.
Is this actually a problem??

Limited resolution



OTU groupings ignore the fine structure present in phylogeny

Same species, different genomes



Perna *et al* 2001 *Nature,* Welch *et al* 2002 *PNAS*

Three genomes, same species only 40% genes in common

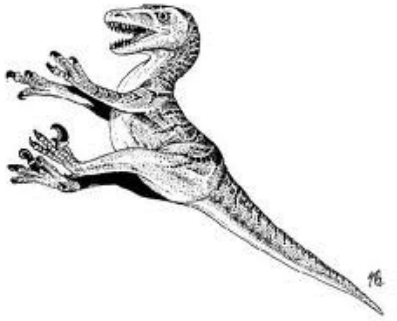
Phylogeny: an alternative path

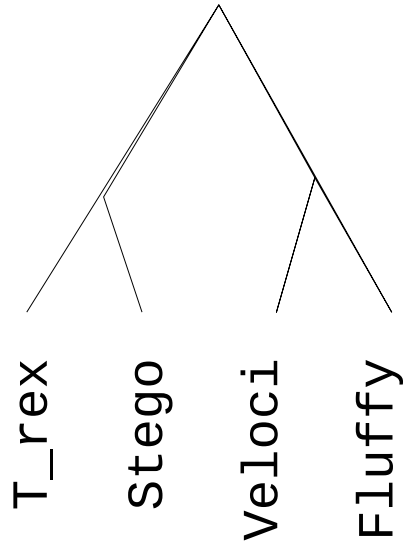
Many ecological analyses can be based on phylogeny:

- Alpha diversity (e.g. species diversity)
- Beta diversity (e.g. comparison of species across samples)
- Community assembly

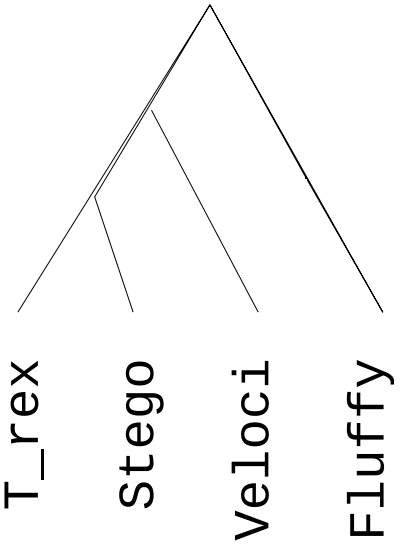
So... what is a phylogeny, anyway?

Imagine you are dating a paleontologist...



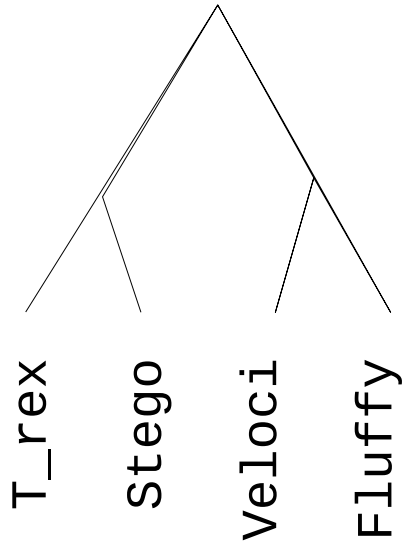


vs.

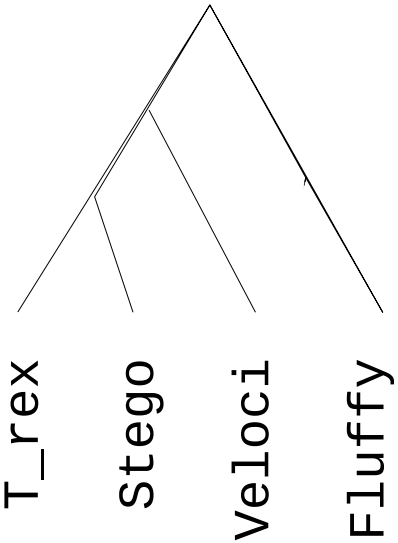


Now imagine you've got dino DNA...

Let's try to reject the reviewer's phylogeny using DNA evidence!



vs.



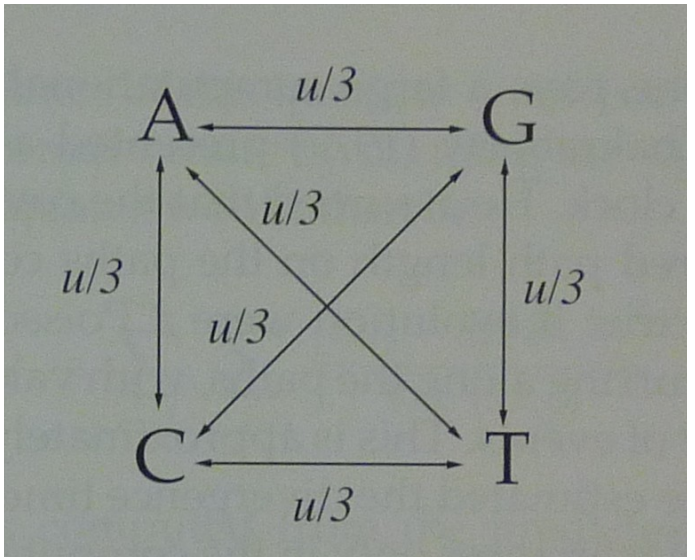

```
>T_rex
ACC
>Stego
TCC
>Veloci
ACG
>Fluffy
ATCG
```

Multiple alignment
(MUSCLE, FSA, etc)

```
T_rex     A-CC
Stego     -TCC
Veloci    A-CG
Fluffy    ATCG
```

How does DNA evolve?

- Simplest model: all nucleotides are equally common, all changes from one to another equally likely (Jukes and Cantor, 1969)



Rate of substitution is $u/3$ per unit time

Expected number of changes on branch of length t is $(4/3)ut$

Prob. of no change: $e^{-(4/3)ut}$
Prob. of at least one change: $1 - e^{-(4/3)ut}$

Prob. of e.g. A to C is $\text{Prob}(C|A,u,t) = (1/4)(1 - e^{-(4/3)ut})$

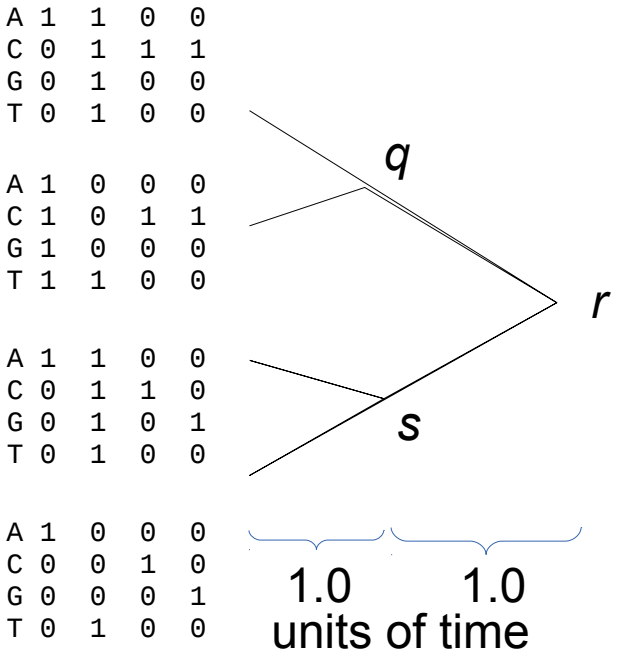
Calculating the likelihood of data given a tree

$$P(X|Y,u,t) = (1/4)(1 - e^{-(4/3)ut})$$

P(X|Y,0.1,1.0) = 0.0312
P(X|X,0.1,1.0) = 0.9064

P(X|Y,0.1,2.0) = 0.0585
P(X|X,0.1,2.0) = 0.8244

```
T_rex A-CC
Stego -TCC
Veloci A-CG
Fluffy ATCG
```



u=0.1, *t*=1.0:
Finite time transition matrix

	A	C	G	T
A	0.91	0.03	0.03	0.03
C	0.03	0.91	0.03	0.03
G	0.03	0.03	0.91	0.03
T	0.03	0.03	0.03	0.91

Steps:

1) Branch lengths
2) Finite-time transition probabilities
3) Leaf node partial probabilities

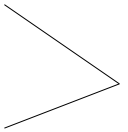
Calculating the likelihood of data given a tree

Sites evolve independently. Calculate site likelihoods one-at-a-time

```
T_rex A-CC
Stego -TCC
```

```
A 1  1  0  0
C 0  1  1  1
G 0  1  0  0
T 0  1  0  0

A 1  0  0  0
C 1  0  1  1
G 1  0  0  0
T 1  1  0  0
```



q

```
A .91  .03  .01  .01
C .03  .03  .82  .82
G .03  .03  .01  .01
T .03  .91  .01  .01
```

Matrix multiply site T_rex 1:

	A	C	G	T
A	0.91	0.03	0.03	0.03
C	0.03	0.91	0.03	0.03
G	0.03	0.03	0.91	0.03
T	0.03	0.03	0.03	0.91

$$\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.91*1 + 0*0.03 + 0*0.03 + 0*0.03 \\ 0.03*1 + 0*0.91 + 0*0.03 + 0*0.03 \\ 0.03*1 + 0*0.03 + 0*0.91 + 0*0.03 \\ 0.03*1 + 0*0.03 + 0*0.03 + 0*0.91 \end{pmatrix}$$

Matrix multiply site Stego 1:

	A	C	G	T
A	0.91	0.03	0.03	0.03
C	0.03	0.91	0.03	0.03
G	0.03	0.03	0.91	0.03
T	0.03	0.03	0.03	0.91

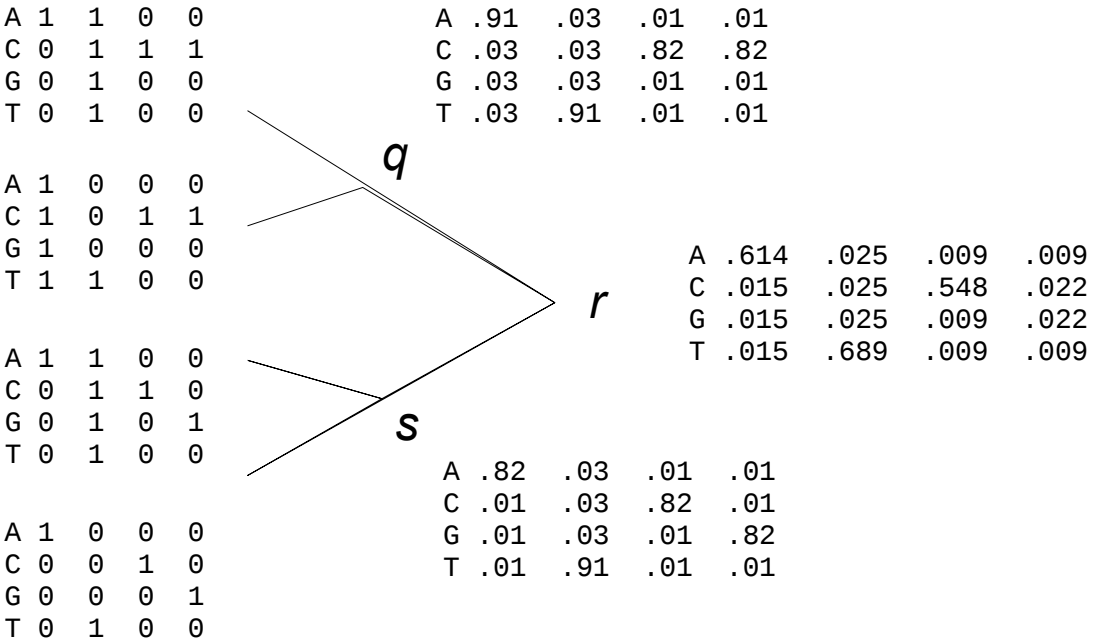
$$\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{matrix} 1\ A \\ 1\ C \\ 1\ G \\ 1\ T \end{matrix}$$

Joint prob T_rex & Stego:

```
0.91*1 A
0.03*1 C
0.03*1 G
0.03*1 T
```

Calculating the likelihood of data given a tree

```
T_rex   A-CC
Stego   -TCC
Veloci  A-CG
Fluffy  ATCG
```



Multiply by background nt probabilities:
A=0.25,C=0.25,G=0.25,T=0.25

```
A .614*.25  .025*.25  .009*.25  .009*.25
C .015*.25  .025*.25  .548*.25  .022*.25
G .015*.25  .025*.25  .009*.25  .022*.25
T .015*.25  .689*.25  .009*.25  .009*.25

A .154      .006      .002      .002
C .004      .006      .137      .006
G .004      .006      .002      .006
T .004      .172      .002      .002
```

Site likelihoods by adding prob. of each nt:

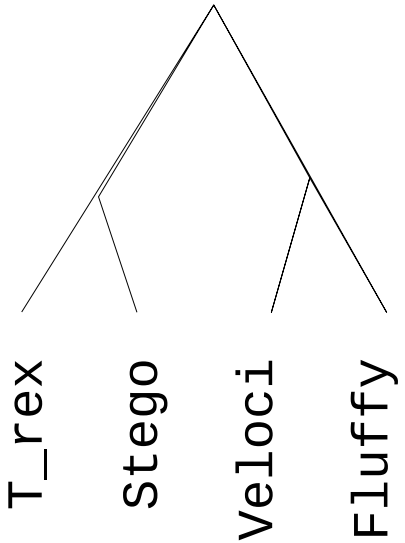
```
.166  .190  .143  .016
```

Tree likelihood is product of sites:

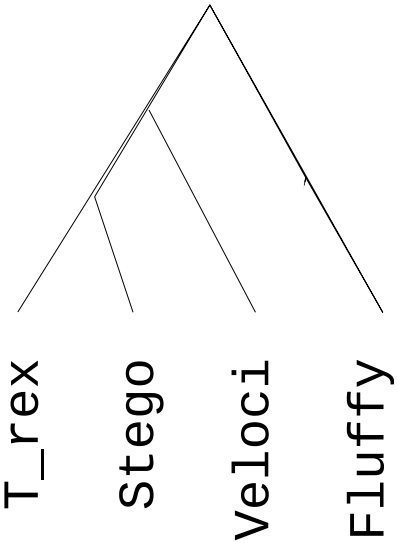
L = .00007216
log(L) = -9.536

Hypothesis testing with tree likelihoods

The likelihood ratio test



vs.



L = .00007216

L = 0.000010348

Take the ratio of likelihoods: $\frac{0.00007216}{0.000010348}$ = 6.973328

Reviewer's tree ~7 times less likely

What if you don't know the tree?

Many methods for tree inference

- Parsimony, Distance, **Maximum Likelihood, Bayesian**
- Maximum Likelihood
  - FastTree, RAxML, GARLI, PHYML, etc.
- Bayesian
  - MrBayes, BEAST, PhyloBayes
  - All based on Markov chain Monte Carlo (MCMC) algorithms

Number of unrooted tree topologies with *n* tips:

$$(2n-3)!! = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

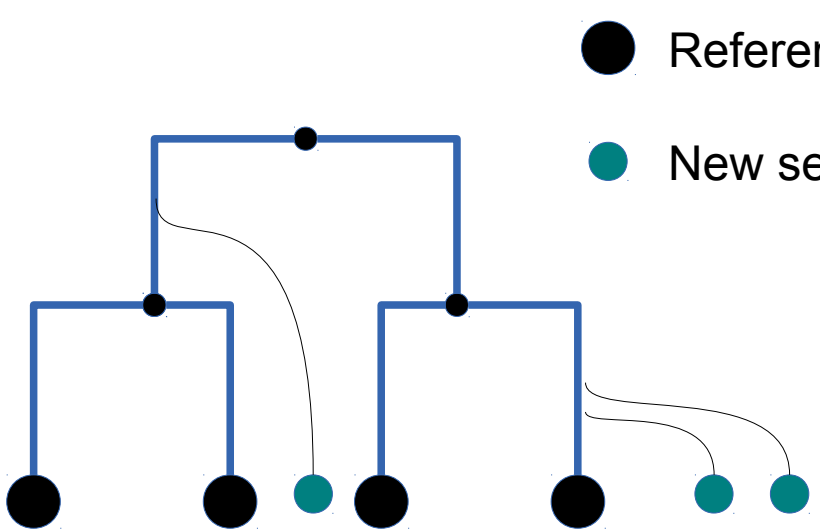
trees with:	
4 tips	3
6 tips	105
8 tips	20,395
10 tips	2,027,025
50 tips	2.84×10^{74}

Bottom line: tree inference is *hard*

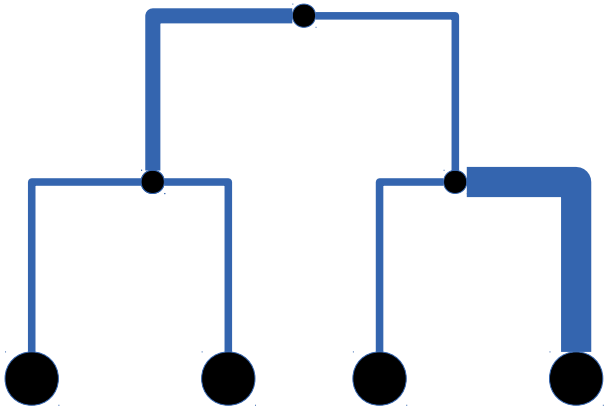
Estimated number of atoms in observable universe: **~10^{80}**

Using phylogenies for microbial ecology

- Building phylogeny from >1M sequences: **impossible**
- Alternative: place new sequences on reference tree
  - RAxML-EPA: Berger *et al* 2011 *Systematic biology*
  - pplacer: Matsen *et al* 2010 *BMC Bioinformatics*
  - SEPP: Mirarab *et al* 2012 *Pac. Symp. Biocomput.*

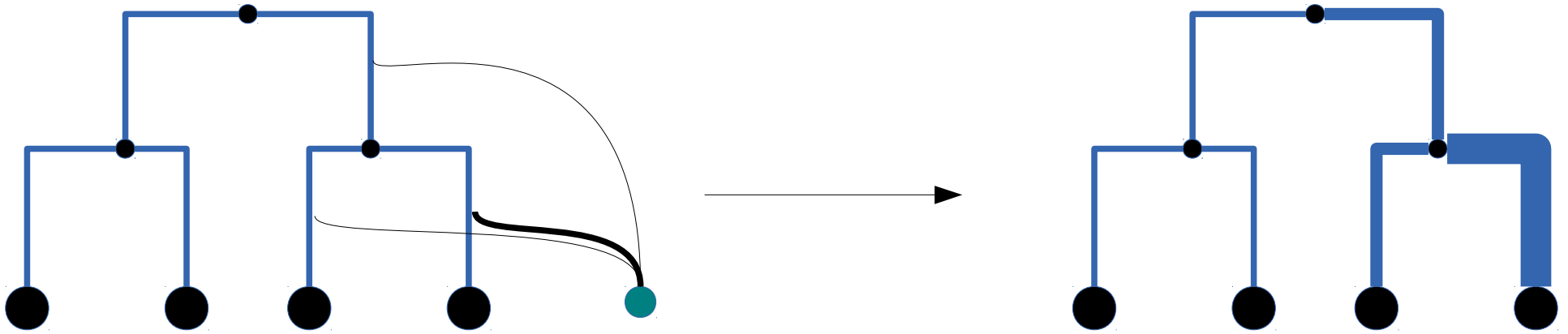


A “fat” tree: showing number of placements on each branch



Handling uncertainty

- Bayesian placement (pplacer)
  - Calculate probability of new sequence on each branch
  - pplacer can do this quickly, analytically (no MCMC)



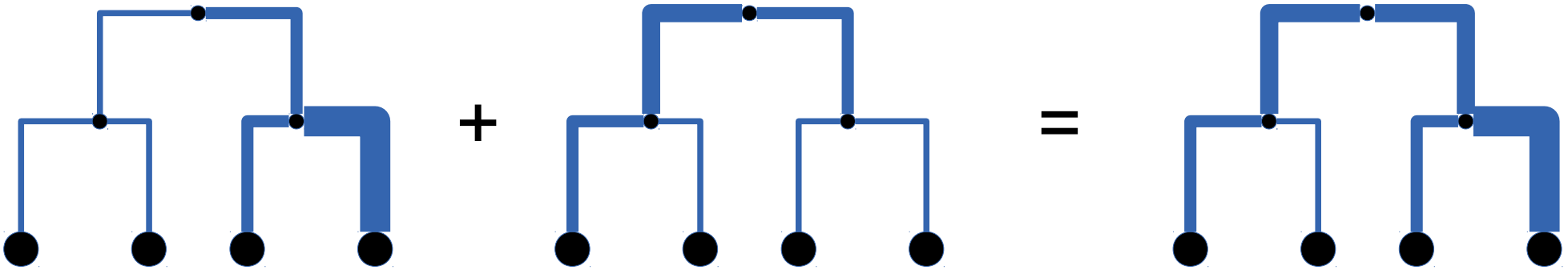
A single sequence with uncertain placement

placement distribution viewed as a fat tree

Placement is starting to look better than OTUs

Uncertainty in many sequences

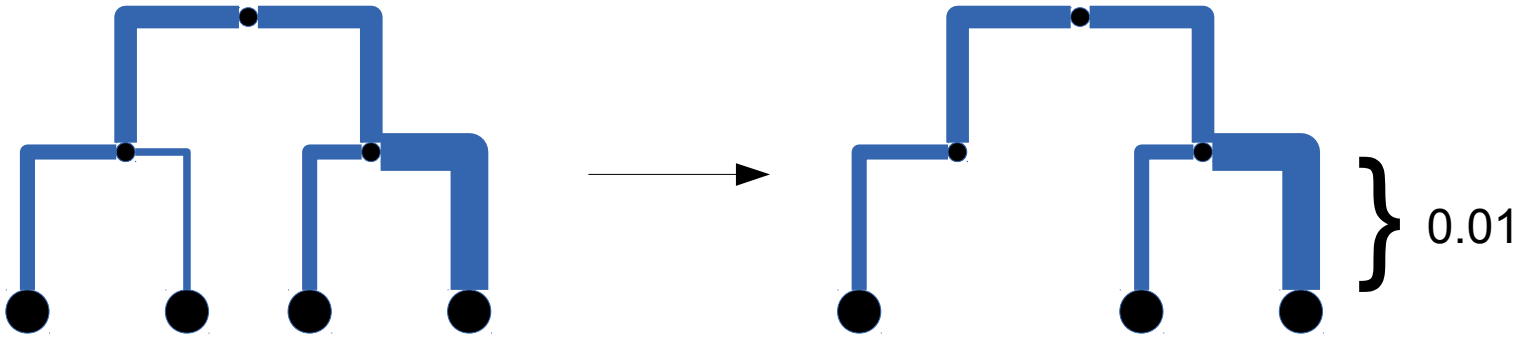
- Combine placement distributions from all seqs in sample



Sequence 1 Sequence 2 Whole sample

Using a placement distrib.: alpha diversity

- Phylogenetic diversity is sum length of branches covered



Sample PD is 0.01 + 0.01 + 0.01 + 0.01 + 0.01 = 0.05

- BWPD: Balance-weighted phylogenetic diversity (Barker 2002)
  - Intuition: weight the contribution each lineage makes to PD by its relative abundance
  - Weights can reflect *placement uncertainty*

$BWPD_\theta$: partial weighting for PD

- A 1-parameter function interpolates between PD and BWPD (Matsen & McCoy 2013, *PeerJ*)
- When $\theta = 0$ it is simply PD. $\theta = 1$ it is BWPD.
- Matsen & McCoy compare:
  - OTU-based diversity metrics
  - Phylogenetic diversity (Faith 1992)
  - Phylogenetic entropy (Rao 1982, Warwick & Clarke 1995)
  - Phylogenetic quadratic entropy (Allen, Kon & Bar-Yam 2009)
  - ${}^{q}D(T)$ (Chao, Chiu, Jost 2010)
  - BWPD (Barker 2002)
  - $BWPD_\theta$

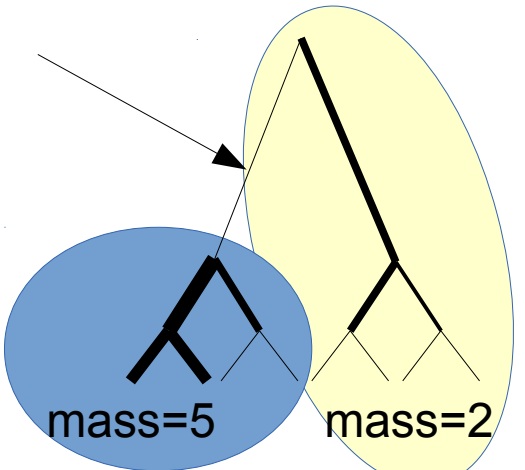
  on 3 different microbial communities, measuring correlation of diversity & phenotype
  - Vaginal, oral, & skin microbiomes

- θ=0.25 & θ=0.5 have highest correlation with microbial community phenotypes
- OTU based diversity metrics have least correlation with phenotype

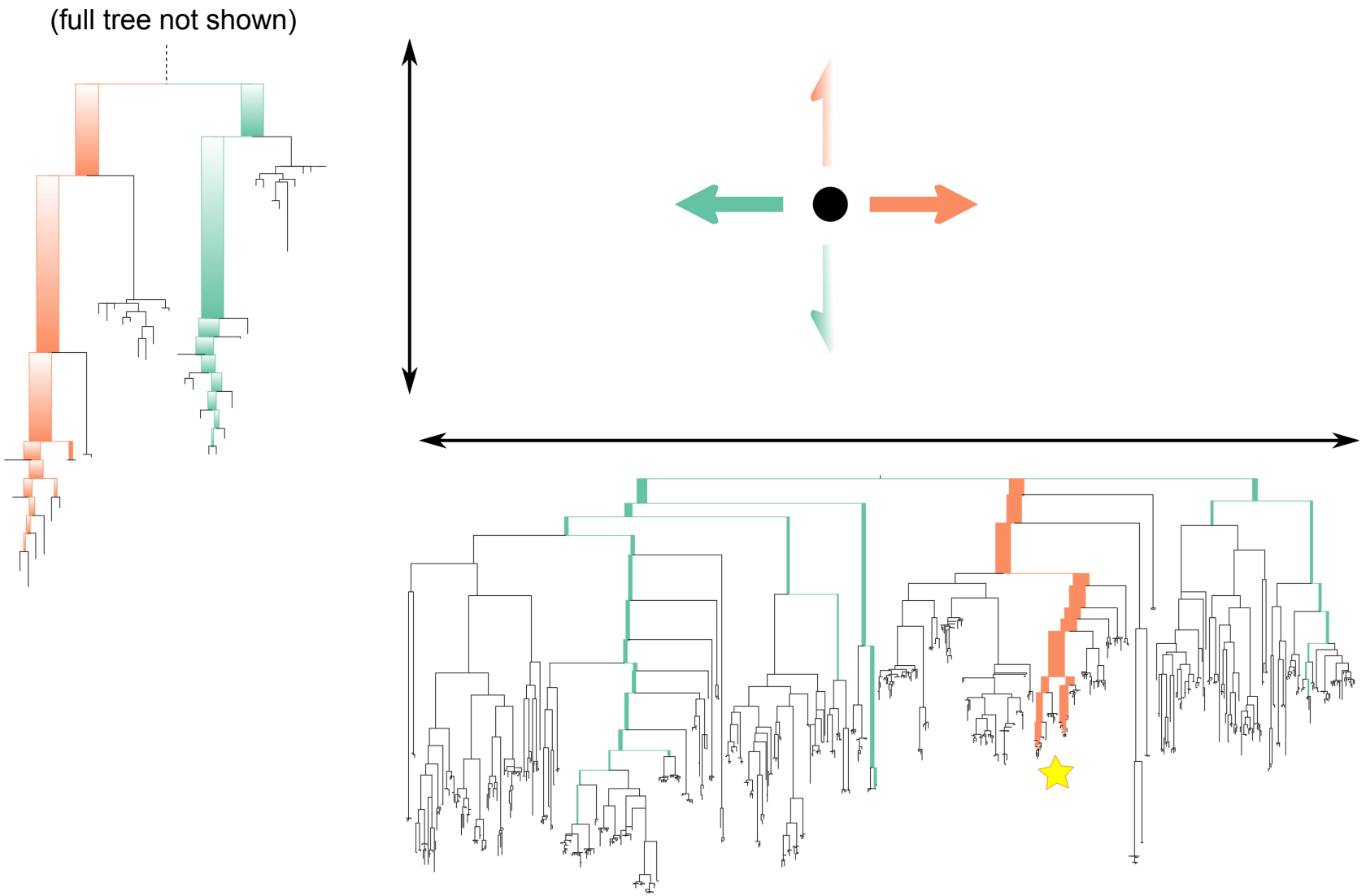
Beta diversity: Edge Principal Component Analysis

- Edge PCA for exploratory data analysis (Matsen and Evans 2013)
- Given *E* edges and *S* samples:
  - For each edge, calculate difference in placement mass on either side of edge
  - Results in *E* x *S* matrix
  - Calculate *E* x *E* covariance matrix
  - Calculate eigenvectors, eigenvalues of covariance matrix
- Eigenvector: each value indicates how “important” an edge is in explaining differences among the *S* samples

Example calculating a matrix entry for an edge:
This edge gets 5-2=3



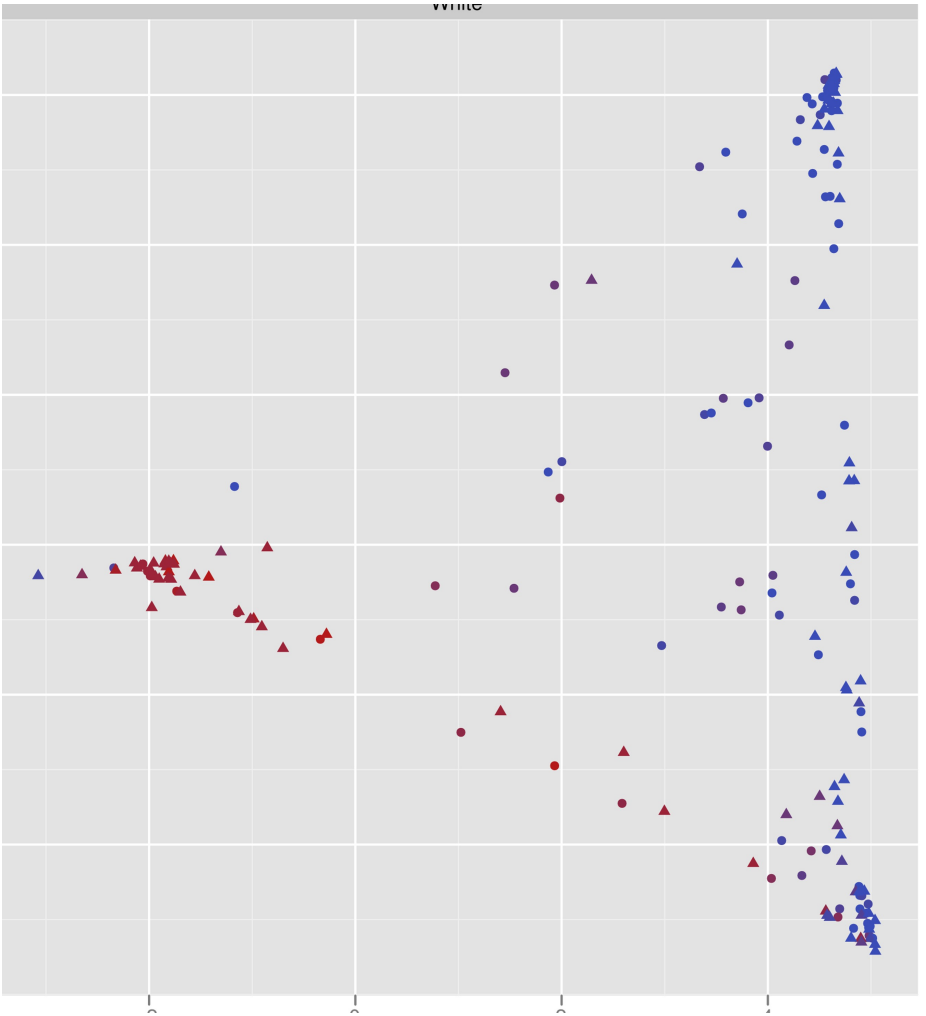
Edge PCA: visually



Branches are thickened & colored according to the amount they shift the sample along an axis

Matsen & Evans 2012 *PLoS ONE*

Edge PCA and the vagina



- Samples colored according to Nugent score of bacterial vaginosis: blue → healthy, red → sick (Matsen & Evans 2012)

How to do it?

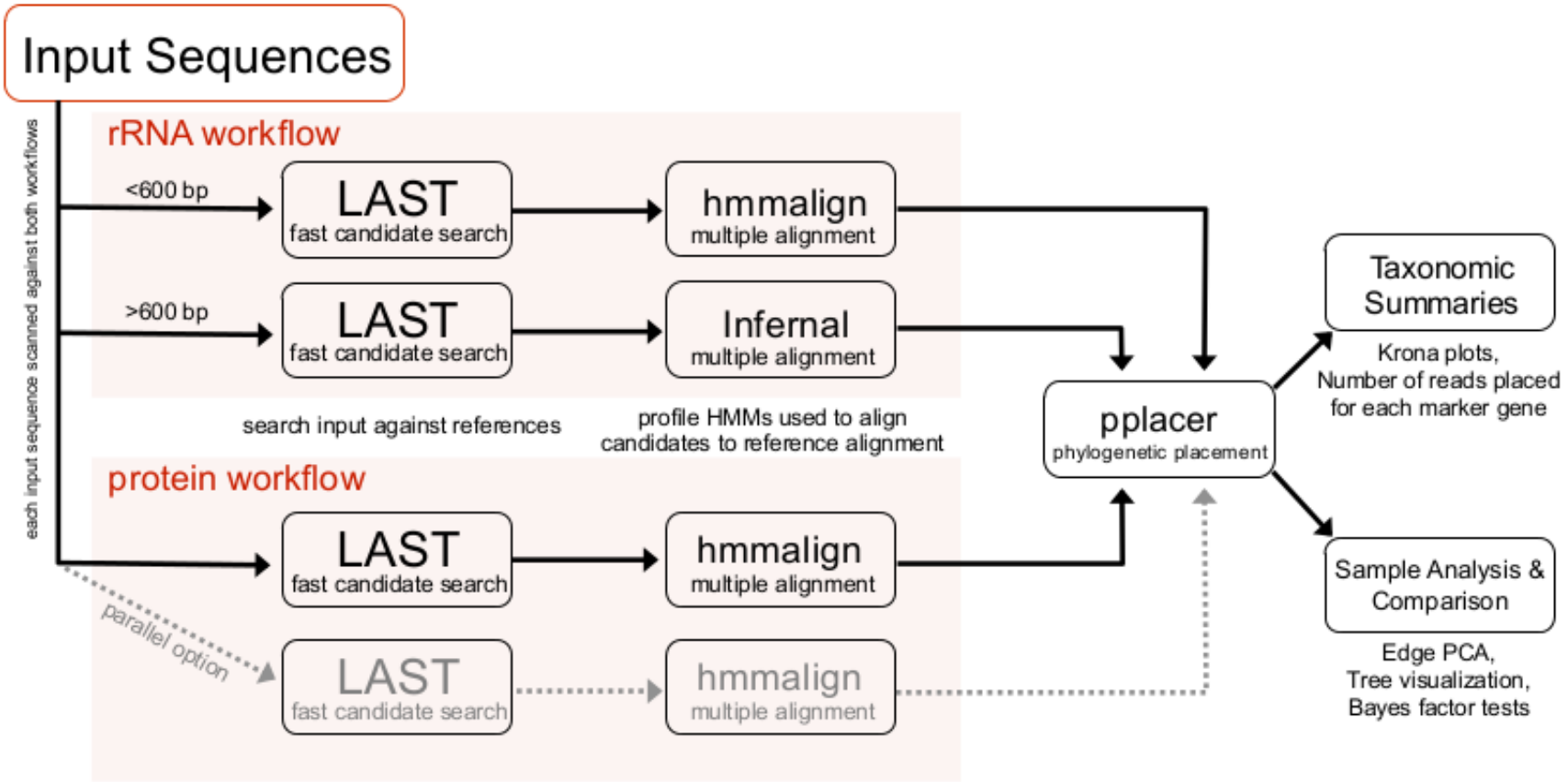
1. Find reference sequences
2. Align reference sequences
3. Infer reference phylogeny
4. For each sample:
    4.1. Add sequences to alignment
    4.2. Place sequences on tree
5. Alpha & Beta diversity analysis

Each step is a unix command

PhyloSift: genome and metagenome phylogeny

:metAMOS:

Treangen *et al* 2013.



Illumina reads placed onto reference gene family trees

- 40 “elite” families: universal among ~4000 Bact, Arch, Euk genomes (Lang *et al* 2013, Wu *et al* 2013)
- 350,000 “extended” families: SFAMs (Sharpton *et al* 2012)
- Amino-acid and nucleotide alignments+phylogenies

Darling *et al* 2014 *PeerJ*.

Using phylosift

Download phylosift: phylosift.wordpress.org

```
bin/phylosift all --output=hmp tutorial_data/HMP_1.fastq.gz
```

Raw illumina data

```
open hmp/HMP_1.fastq.gz.html
```

Shows taxonomic plot (Mac)

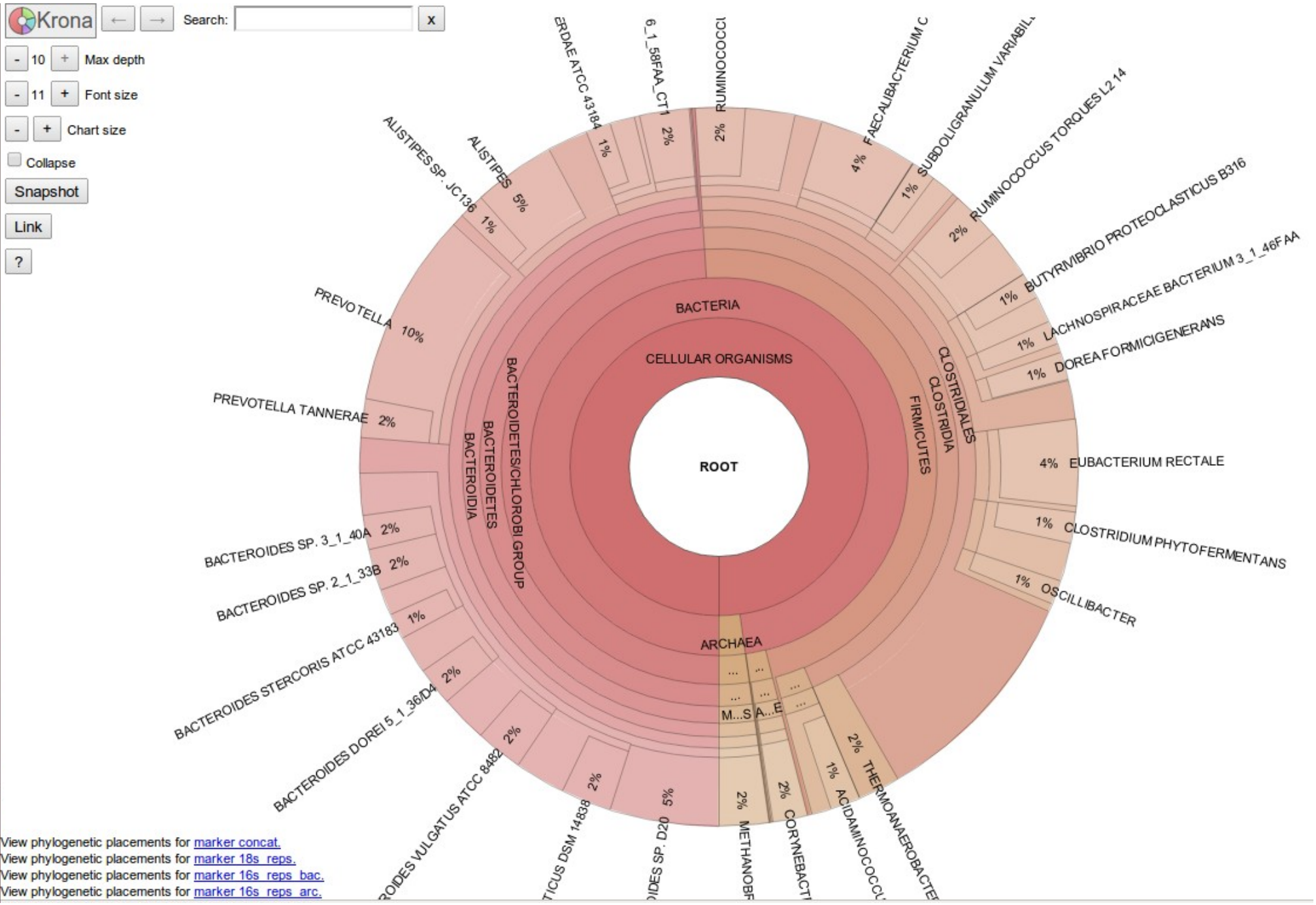
```
bin/guppy fpd --theta 0.25,0.5 hmp/*.gz.jplace
```

Alpha diversity

```
bin/guppy epca --prefix pca hmp/*.gz.jplace
```

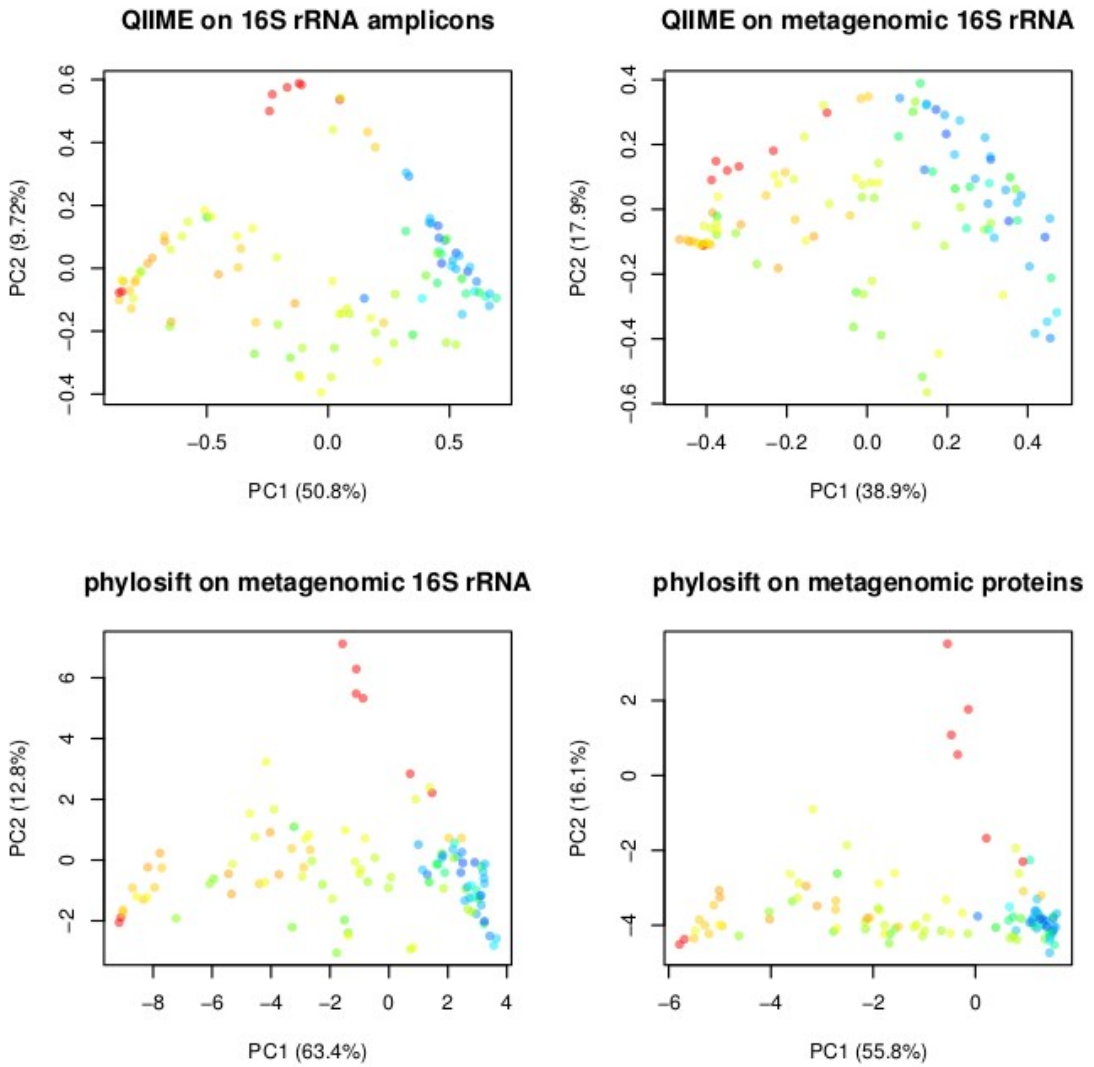
Beta diversity (min 3 samples)

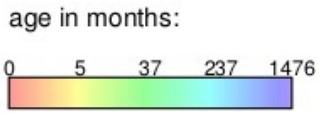
More examples at: phylosift.wordpress.org



View phylogenetic placements for marker concat.
View phylogenetic placements for marker 18s reps.
View phylogenetic placements for marker 16s reps bac.
View phylogenetic placements for marker 16s reps arc.

QIIME vs. PhyloSift





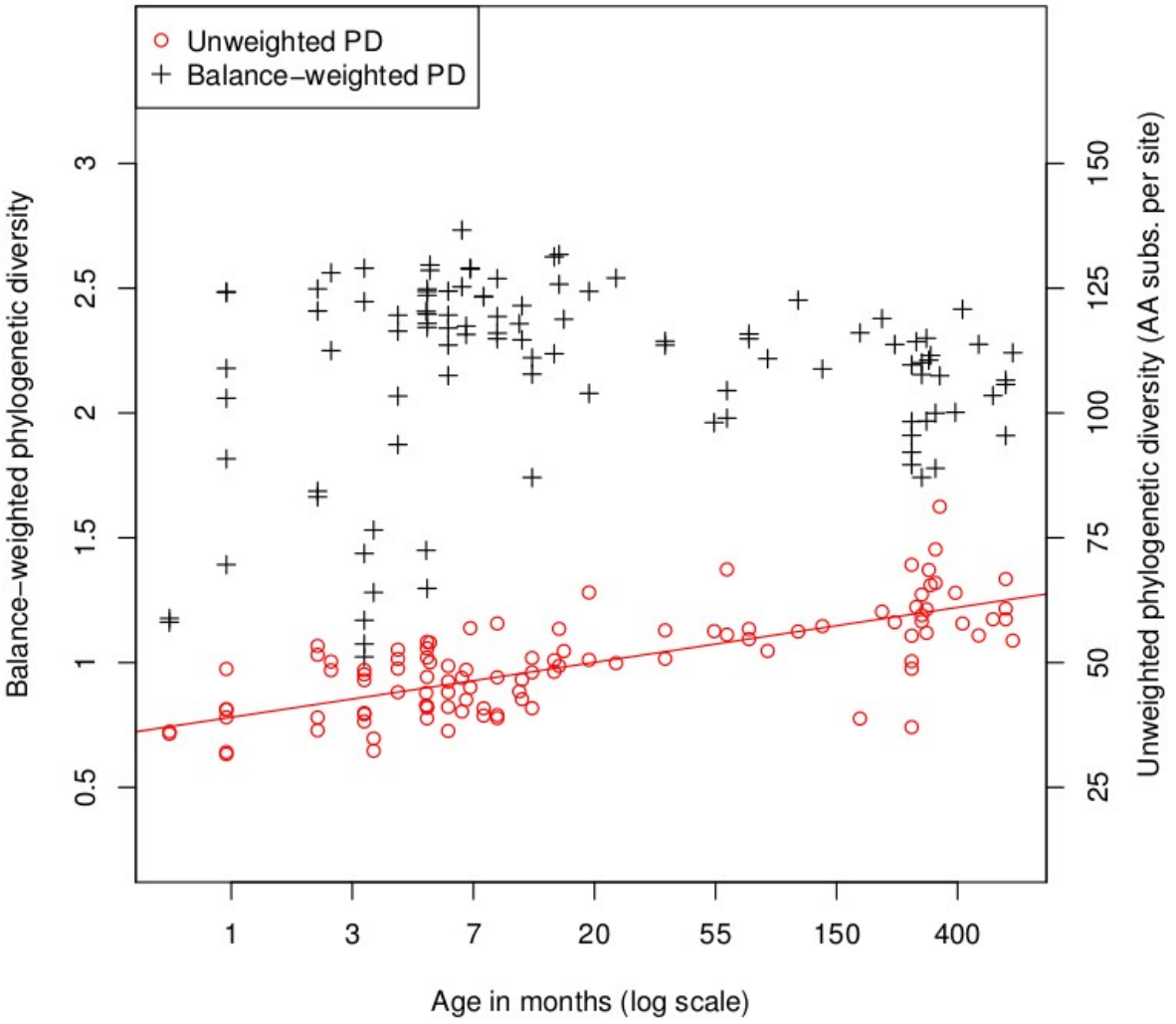
Phylosift on proteins & 16S produces similar results to QIIME on amplicon data

Data from Yatsunenko *et al* 2012. 16S amplicon & metagenomes from same samples

Phylogenetic alpha diversity

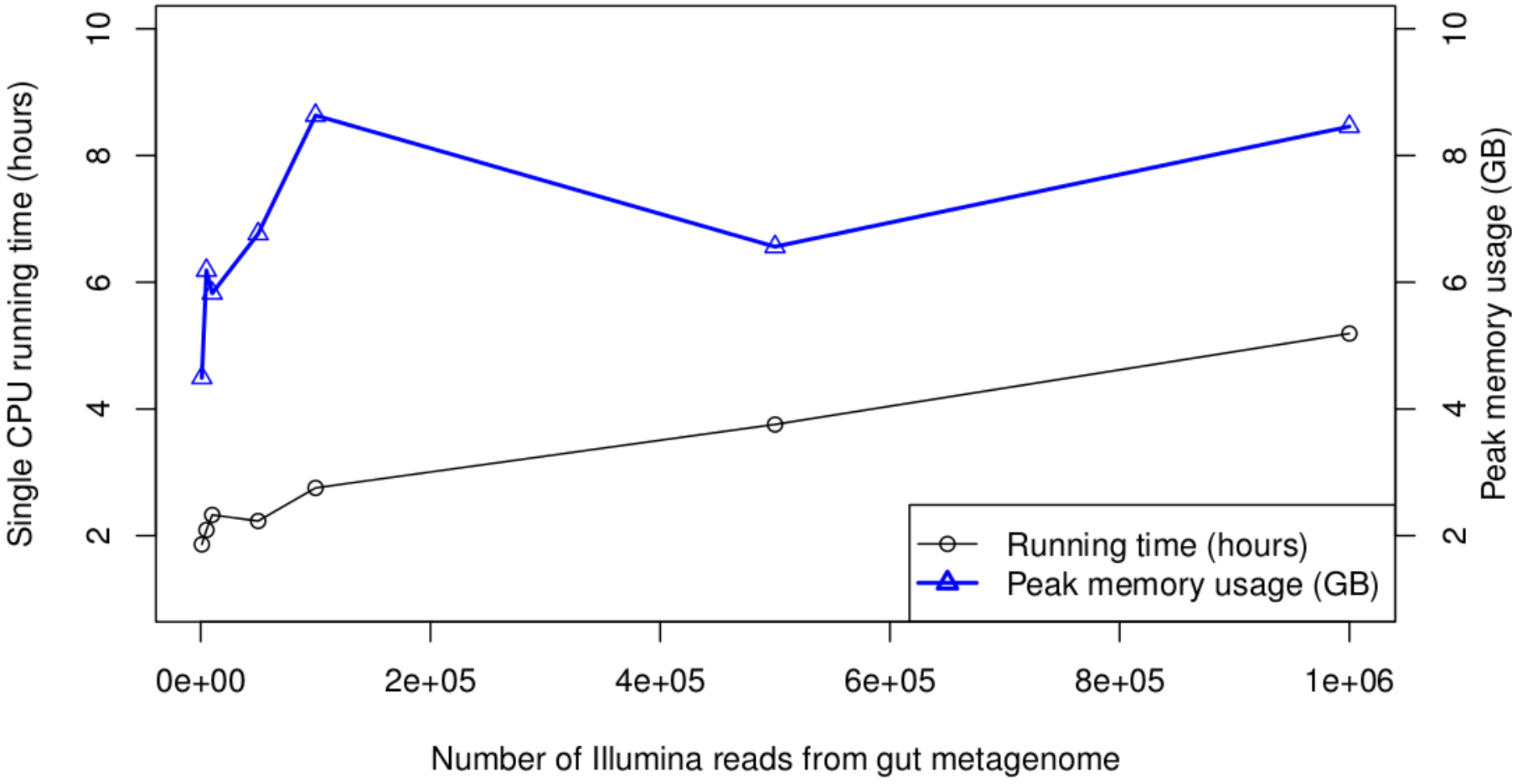
Data from Yatsunenko et al 2012

- Growth in PD over life
- BWPD is biphasic



PhyloSift compute requirements

- You don't need a huge computer to run PhyloSift

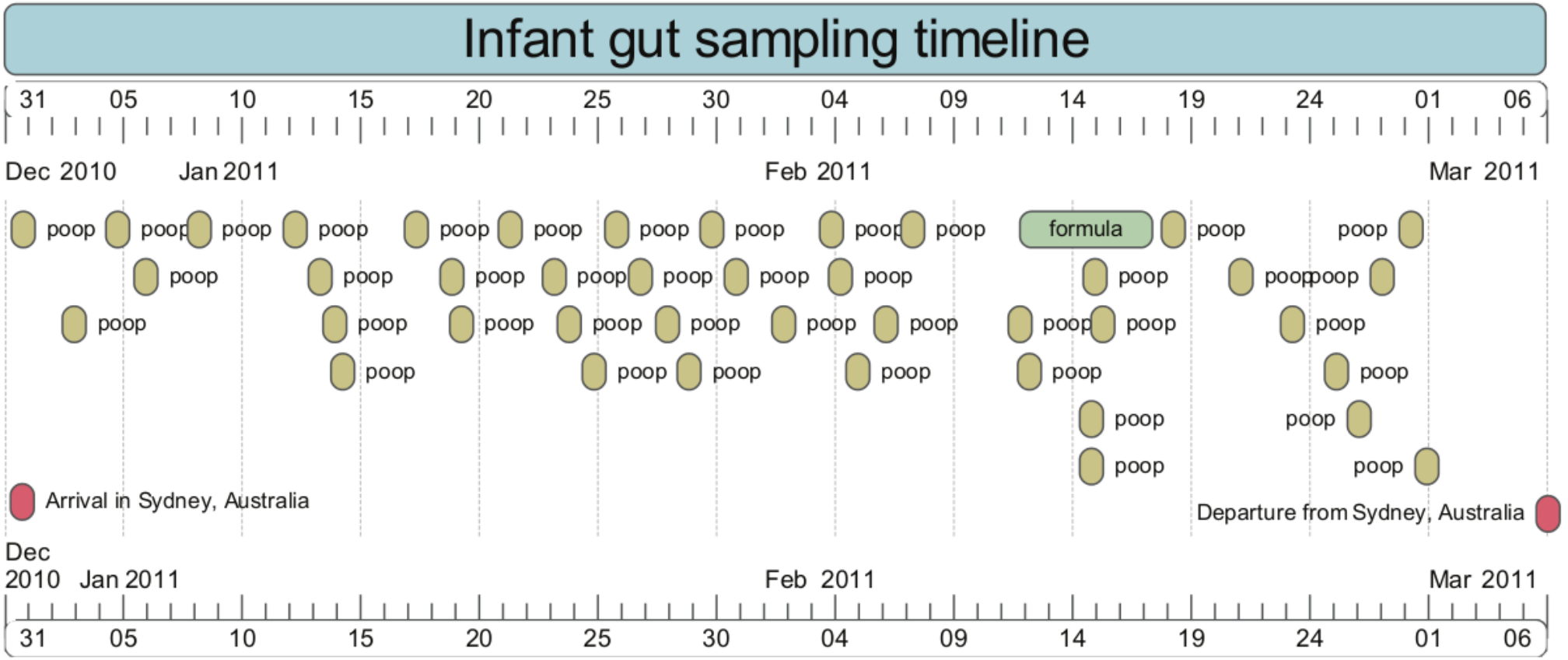


phylosift and major life events

On December 3rd 2010, Kai and his microbiome were born

Lots of nappies, lots of sampling

Kai Darling born 3rd Dec. 2010 in California, flew to Sydney 3.5 weeks later

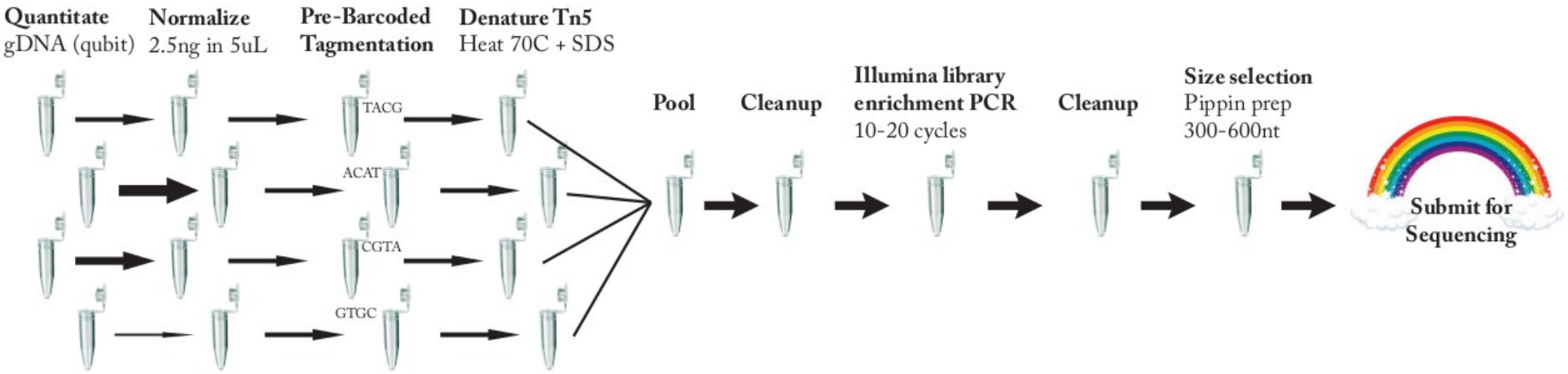


March 1st 2011: a lot of poop in tubes and no idea how to get it through USA quarantine

Tiffanie Nelson at UNSW:
Extracted DNA with PowerSoil kits, mailed to USA

Metagenomics on a shoestring budget*

"Homebrew" Illumina Nextera library prep protocol:



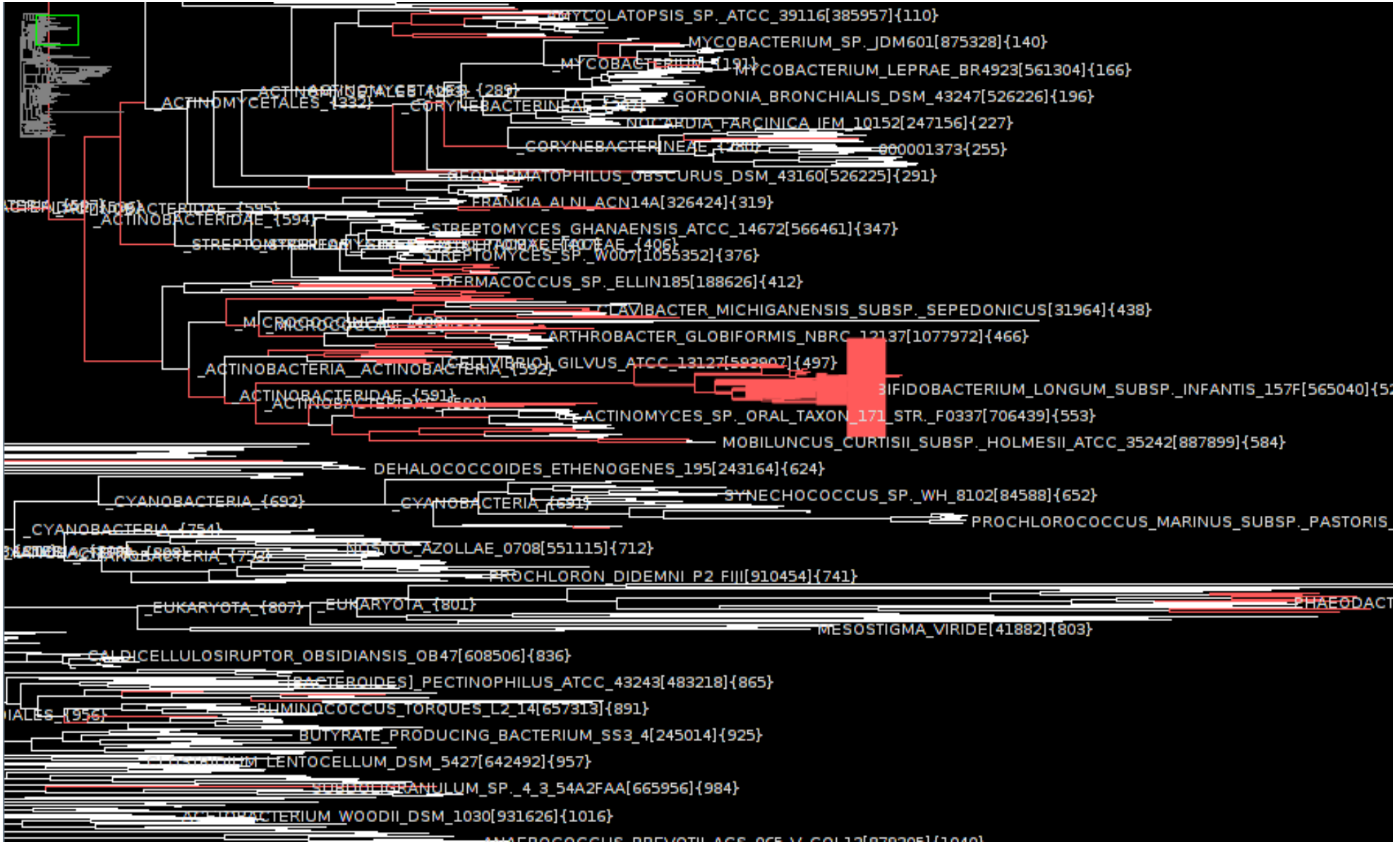
Goal: metagenomics as easy as 16S amplicon studies

Strategy: Transposon-catalyzed library prep.
Express & purify Tn5 from pWH1891. Custom adapters. 2.5ng input
Pool samples as early as possible.

Results: Sequenced 45 time points in HiSeq 2000 lane
~ $1 / library reagent costs, 100s of libraries in a day, NO ROBOTS

*Infant microbiome sequencing sponsored by private funds

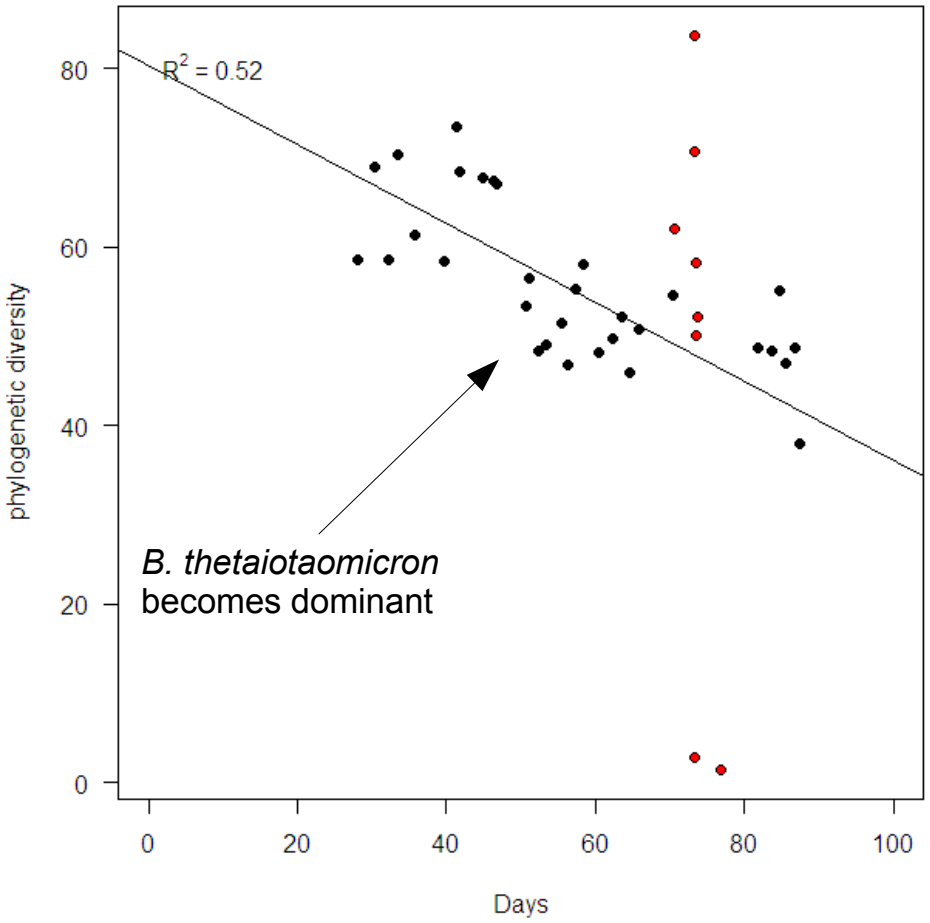
PhyloSift view of fecal microbiome at three weeks age

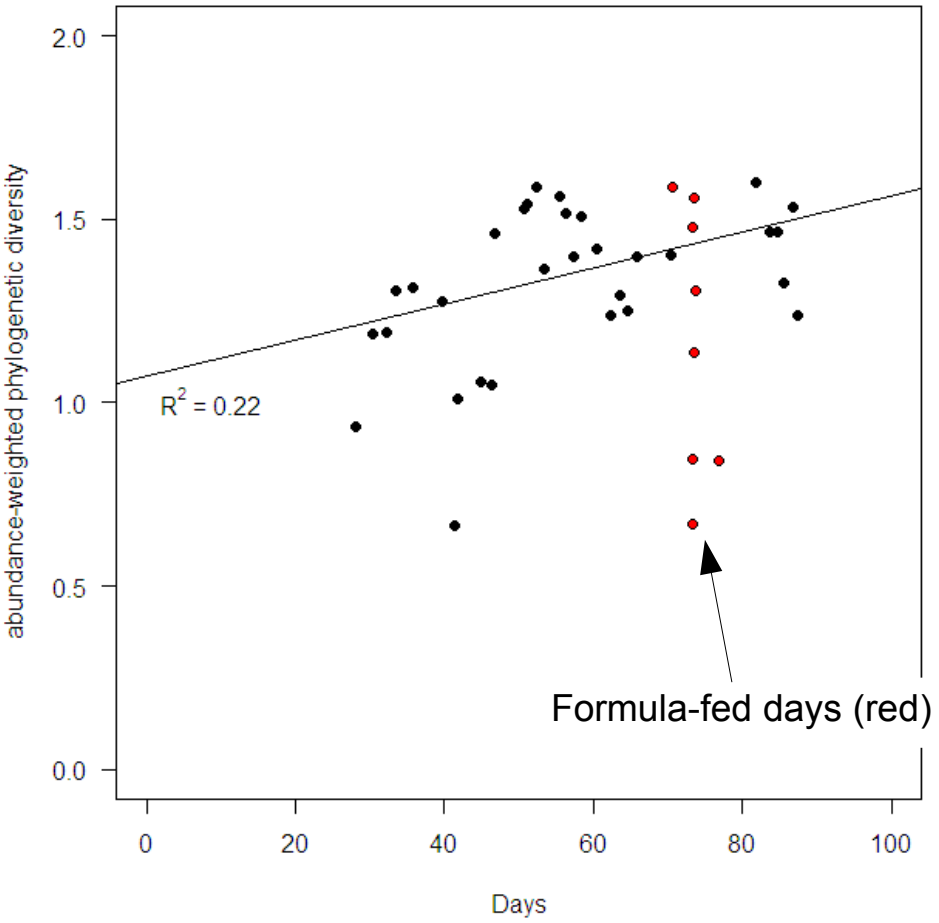


- Tree-browsing of read placement mass (via archaeopteryx)
- Taxonomic summary plots in Krona (Ondov *et al* 2011)

Alpha diversity of gut communities vs. time

- Standard & balance-weighted PD (McCoy & Matsen, 2013)
- Phylogenetic diversity (PD) decreases?!





Pearson's cor: -0.44, p = 0.005
($p < 10^{-6}$ without formula samples)

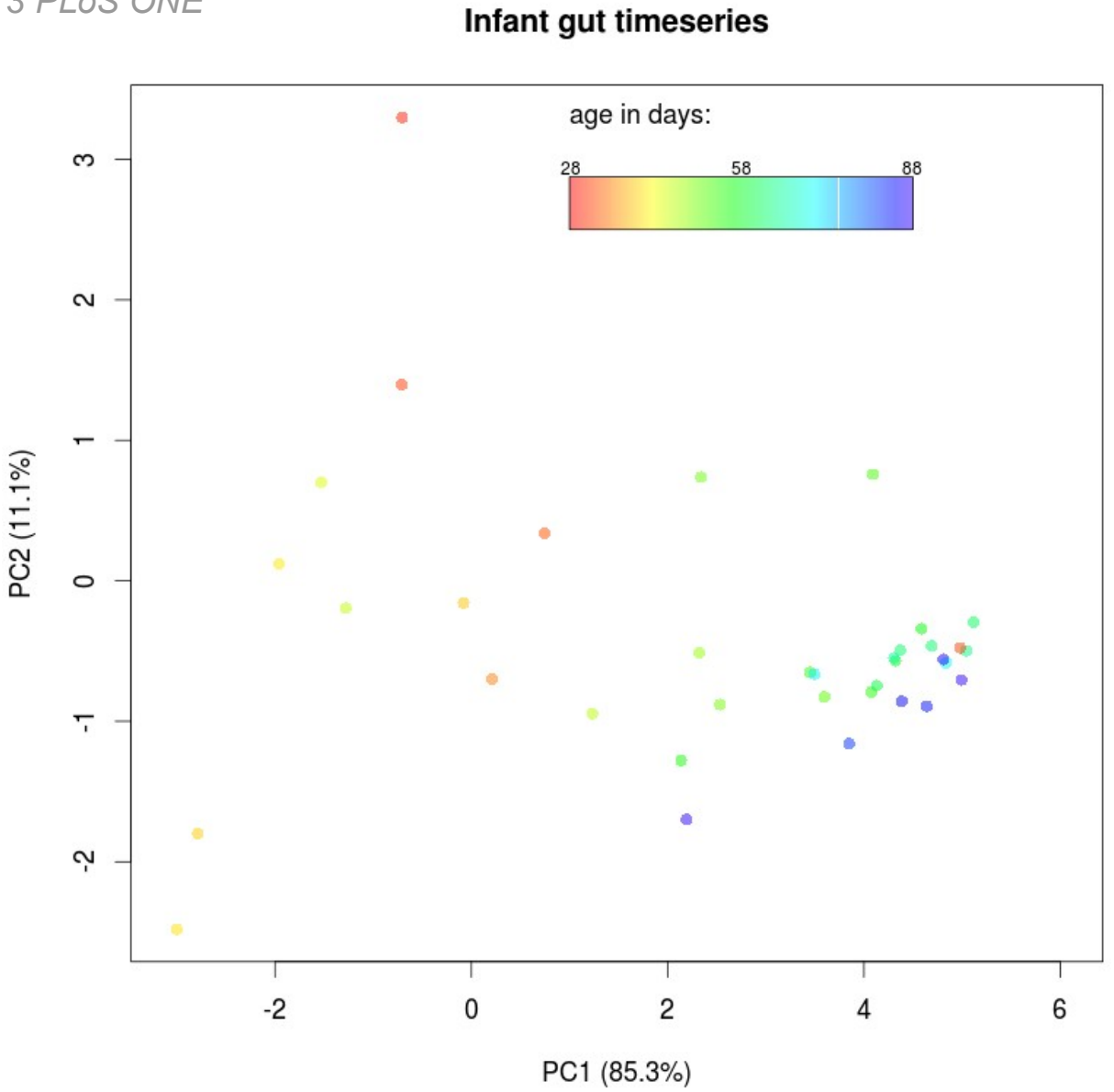
Pearson's cor: 0.21, p = 0.18
(p = 0.0071 w/o formula)

Phylogenetic “Edge PCA” on infant fecal microbiome

Edge PCA: explain variation in community structure among many samples

Matsen & Evans 2013 *PLoS ONE*

Up: *Bifidobacterium longum*



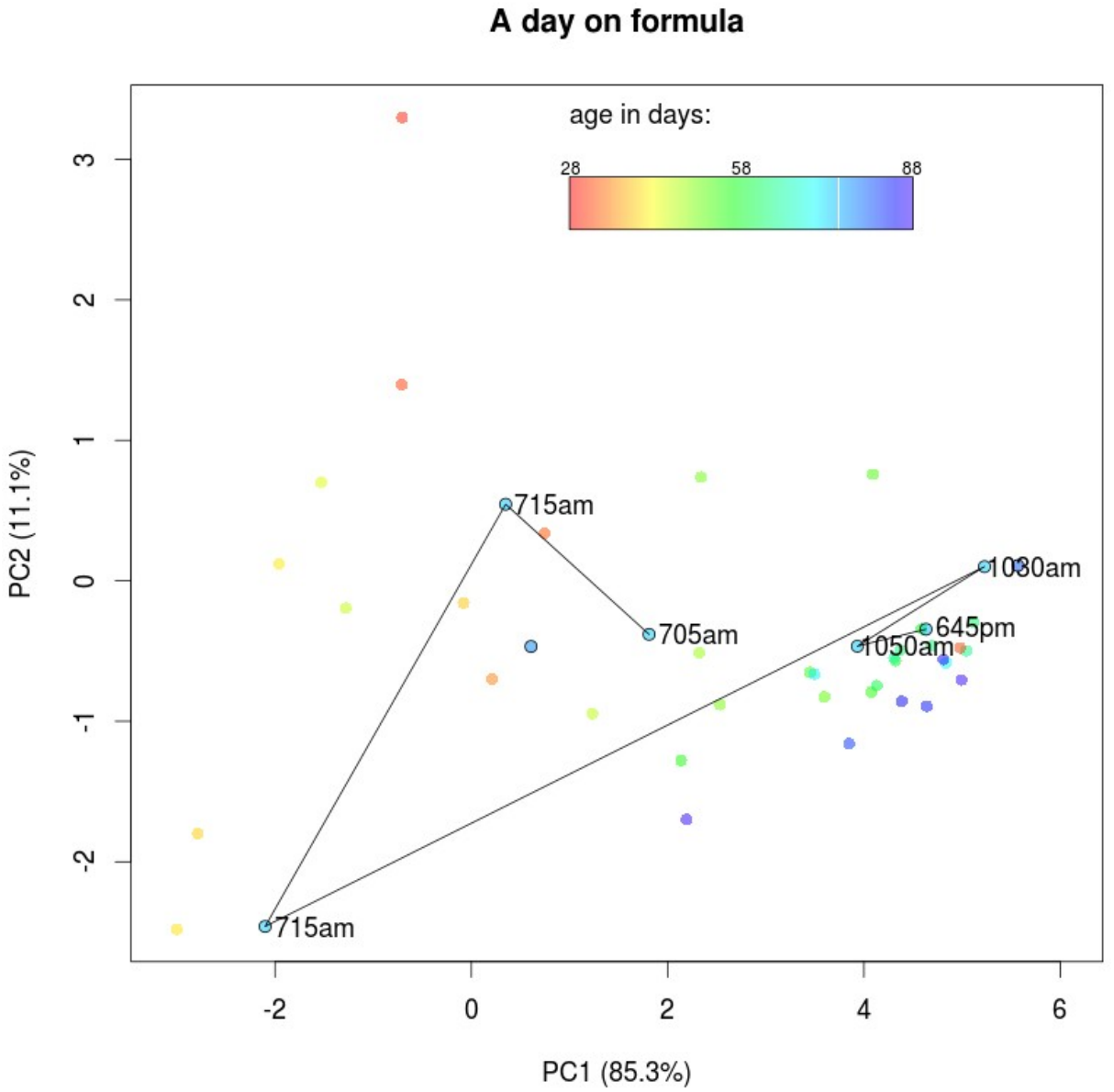
3rd PC (1%):
Staphylococcus
Veillonella

Up: *Bacteroides*, Down: *Bifidobacterium*

Chan *et al* In prep

Formula-fed samples within one day

Up: *Bifidobacterium longum*



One week on formula,

Six poops in one day.

Up: *Bacteroides*, Down: *Bifidobacterium*

Chan *et al* In prep

Thanks!