CAMI: will it be dead on arrival?

Presented by Aaron Darling

Thanks to C. Titus Brown for discussion that inspired this presentation

Three likely ways CAMI will fail

1) Becoming obsolete before it is published

2) Lack of openness & reproducibility in code and data

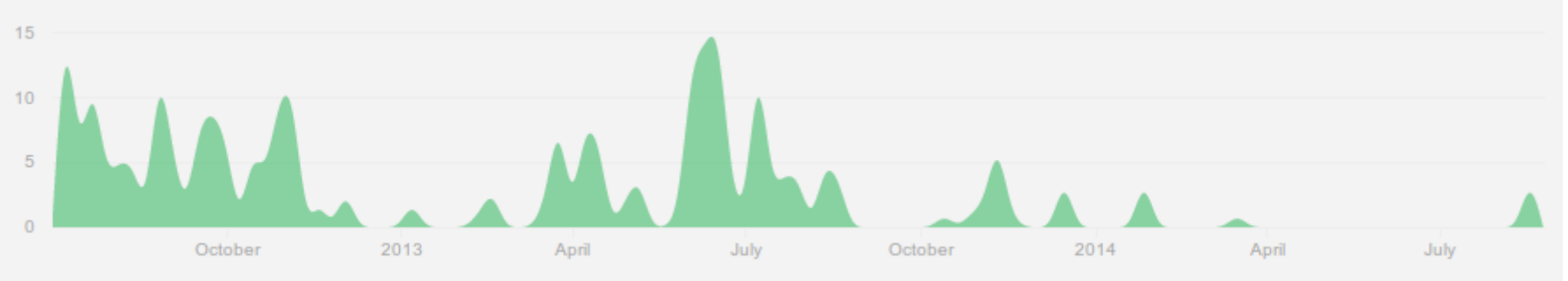
3) Failure to be forward-thinking about future analysis challenges

Will CAMI be obsolete prior to publication?

Analysis software constantly evolves – or it dies

Genome & metagenome databases continuously growing

github commit activity for metAMOS metagenomics pipeline, prior 12 months



commits for GroopM, all time

Genome assembly benchmarks: previous attempts

THE ASSEMBLATHON

- A static assessment of then-current genome assembly
- Generated a simulated dataset, teams independently assembled & submitted results
- Only assemblies submitted, no code, no parameters, **no way to reproduce a team's results**

Genome Assembly Gold-Standard Evaluations

- Also a static assessment
- Real datasets
- Assemblies run by expert team, *slightly* more reproducible

---

An alternative: `http:// nucleotid.es` by Michael Barton

docker

- Weekly evaluation of assembly algorithms
- Uses Docker images for each assembler
- QUAST to measure assembly accuracy

Advantages

Open: Anyone can contribute a docker image
Reproducible: Exact parameters & software available
Current: major improvements (or flaws) identified in days

Read Set 0002

Microbe - approx. size: 4MBp, approx. GC: 40%

Assembler Image	Procfile	NG50	LG50	#contigs	UN	IB
nucleotides/spades	default	109246	15	97	2.62	4.24
nucleotides/spades	careful	109246	15	96	2.62	5.12
nucleotides/spades	single-cell	109202	15	105	2.72	10.38
nucleotides/spades	single-cell-careful	109202	15	100	2.73	3.75
nucleotides/velvet-optimiser	default	91011	18	108	3.29	65.98
nucleotides/idba	single-cell	84779	19	120	2.83	1.16
nucleotides/idba	default	81960	21	134	2.85	0.81
nucleotides/abyss	k-64	78957	19	116	1.27	12.15
nucleotides/abyss	default	38850	41	230	2.56	82.03
nucleotides/idba	idba	16161	94	450	4.4	0.84
nucleotides/soap-denovo	kmer-10-cutoff	12032	118	571	5.96	0.19
nucleotides/soap-denovo	default	11909	118	572	6.01	0.17
nucleotides/velvet	default	2987	504	1469	15.17	7.67

Read Set 0005

Microbe - approx. size: 2MBp, approx. GC: 40%

Assembler Image	Procfile	NG50	LG50	#contigs	UN	IB
nucleotides/spades	single-cell	89875	9	53	2.0	7.49

CAMI: hamstrung by taxonomy?

- Very little of the known diversity is named
- Taxonomy is a human-limited process – naming, renaming, fistfights etc.

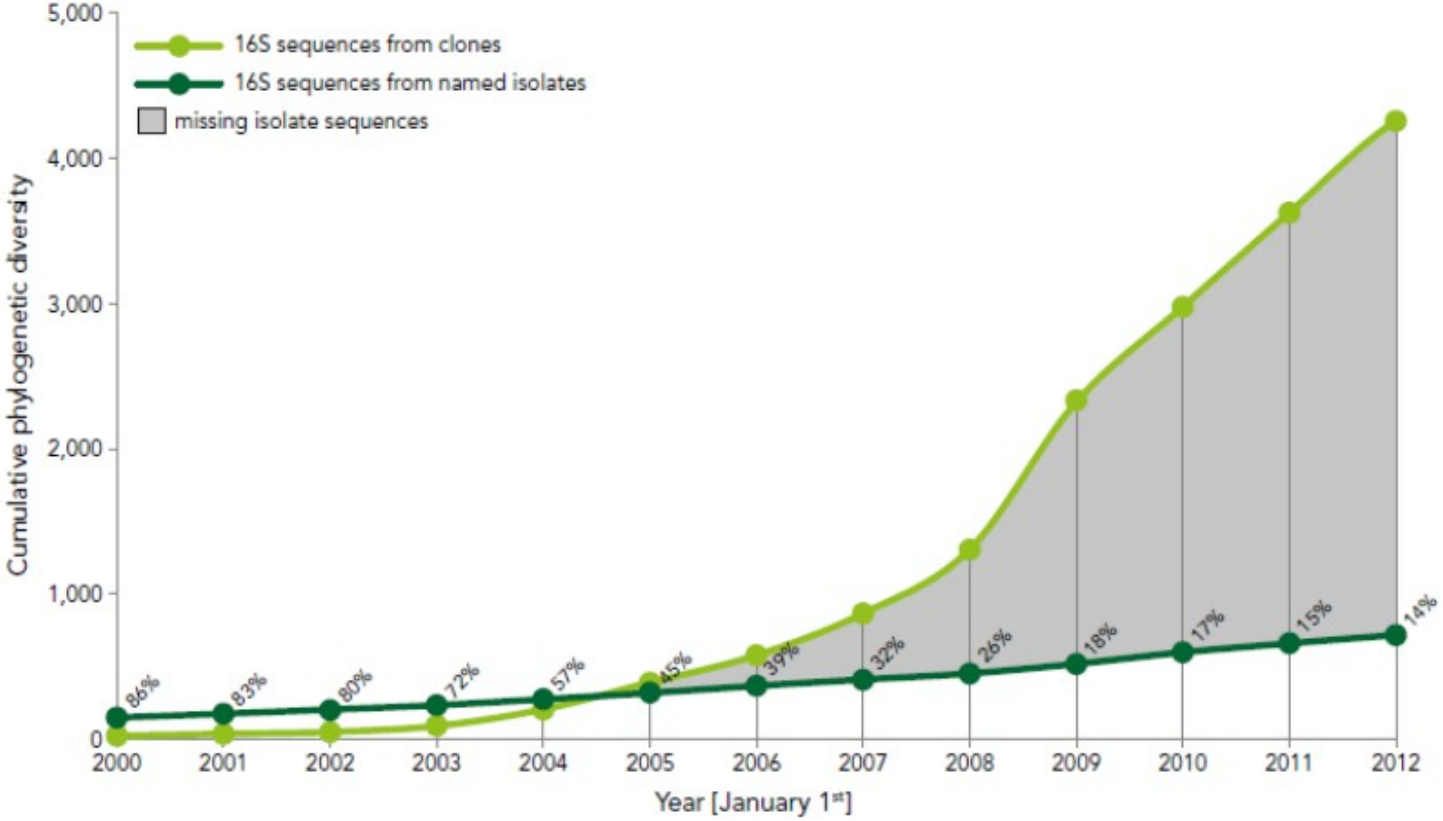


Figure from Rinke *et al* 2013 *Nature* Supplementary material

Is binning to taxonomic marker genes (e.g. 16S or ribosomal proteins) a viable solution?

CAMI: how do we future-proof it?

Recent methods to reconstruct individual genomes from within metagenomes:

- Differential coverage binning:
  - GroopM (Imelfort et al 2014), CONCOCT (Alneberg *et al* 2013), mmgenome (Albertsen 2014)
- High throughput single cell genome sequencing
  - Rinke *et al* 2013 *Nature*, Kashtan *et al* 2014 *Science*
- Metagenomic Hi-C
  - Beitel *et al* 2014 *PeerJ*, Burton *et al* 2014 *G3*

Novel technologies on the horizon?

A solution: benchmark at the limits of science?

- Define a model of a microbial ecosystem's metagenome that is as complete as we could ever hope to reconstruct. What is this limit?
- The complete genome of every cell, with every cell assigned a taxonomic identifier? This is possible via simulation.

Problems, problems, problems:

- What is a meaningful summary?
  - Average fraction of genome reconstructed?
  - Numbers of genes incorrectly binned together?
  - Number of SNPs identified? Correctly linked together?

- Can we can define some of these later?
  - Can individuals contribute benchmark modules for their stats of interest?