Metagenome Assembly

C. Titus Brown

ctb@msu.edu

Aug 14 2014

whoami?

Asst Prof, Microbiology & Computer Science, Michigan State University

Some definitions

- Bioinformaticians write the software that takes your perfectly good data and produces bad clusters and assemblies.

Some definitions

- Bioinformaticians write the software that takes your perfectly good data and produces bad clusters and assemblies.

- Statisticians are the people who take your perfectly good clusters and tell you that your results are not statistically significant.

whoami?

- Primary scientific interest:

  Effective analysis and generation of high-quality biological hypotheses from sequencing data.

- Now entirely *dry* lab.

whoami?

- Cross-cutting interests:
  - Good (efficient, accurate, remixable) software development, on the open source model.
  - Reproducibility.
  - Open science; preprints, blogging, Twitter.

@ctitusbrown on Twitter

http://ivory.idyll.org/blog/

Outline

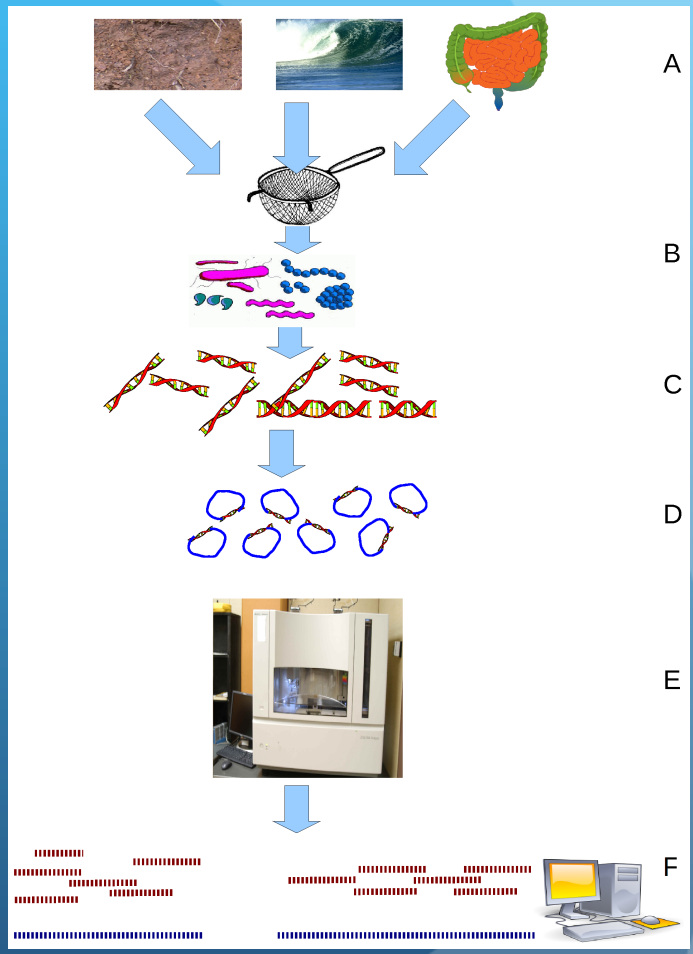
1) Dealing with shotgun metagenomics
2) Short-read annotation & why/why not
3) Annotating soil reads – a story
4) Assembly stories!

Outline

1) Dealing with shotgun metagenomics
2) Short-read annotation & why/why not
3) Annotating soil reads – a story
4) Assembly stories!

Shotgun metagenomics

- Collect samples;
- Extract DNA;
- Feed into sequencer;
- Computationally analyze.



Wikipedia: Environmental shotgun sequencing.png

FASTQ etc.

Name

```
@SRR606249.17/1
GAGTATGTTCTCATAGAGGTTGGTANNNNT
+
B@BDDFFFHHHHHJIJJJJGHIJHJ####1
@SRR606249.17/2
CGAANNNNNNNNNNNNNNNNCCTGGCTCA
+
CCCF##################22@GHIJJJ
```

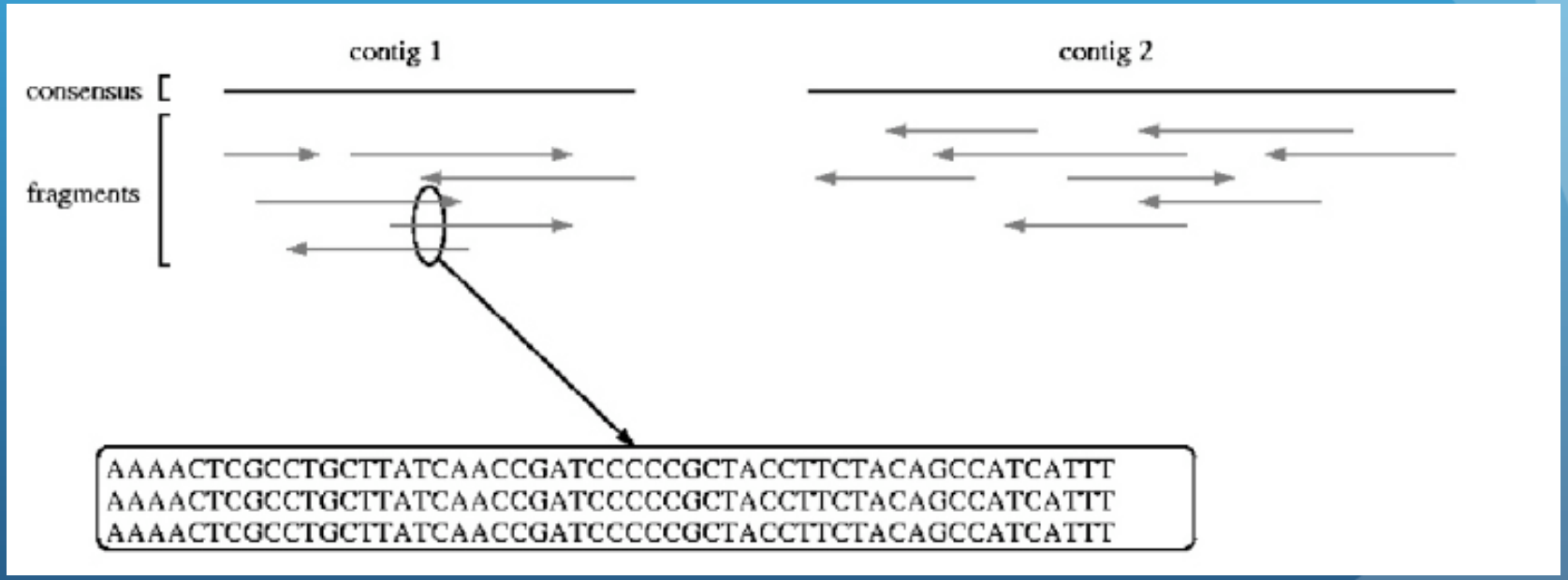
Quality score

Note: /1 and /2 => interleaved

Shotgun sequencing & assembly

Randomly fragment & sequence from DNA;

reassemble computationally.



UMD assembly primer (cbcb.umd.edu)

Dealing with shotgun metagenome reads.

1) Annotate/analyze individual reads.
2) Assemble reads into contigs, genomes, etc.
3) A middle ground (I'll mention on Friday)

Outline

1) Dealing with shotgun metagenomics
2) Short-read annotation & why/why not
3) Annotating soil reads – a story
4) Assembly stories!

Annotating individual reads

- Works really well when you have EITHER
  - (a) evolutionarily close references
  - (b) rather long sequences

(This is obvious, right?)

Annotating individual reads #2

- We have found that this does not work well with Illumina samples from unexplored environments (e.g. **soil**).
- *Sensitivity* is fine (correct match is usually there)
- *Specificity* is bad (correct match may be drowned out by incorrect matches)

Recommendation:

For reads < 200-300 bp,

- Annotate individual reads for human-associated samples, or exploration of well-studied systems.
- For everything else, look to assembly.

So, why assemble?

- Increase your ability to assign homology/ orthology correctly!!

Essentially all functional annotation systems depend on sequence similarity to assign homology. This is why you want to assemble your data.

Why else would you want to assemble?

- Assemble new “reference”.
- Look for large-scale variation from reference – pathogenicity islands, etc.
- Discriminate between different members of gene families.
- Discover operon assemblages & annotate on co-incidence of genes.
- Reduce size of data!!

Why *don't* you want to assemble??

- Abundance threshold – low-coverage filter.
- Strain variation
- Chimerism

Outline

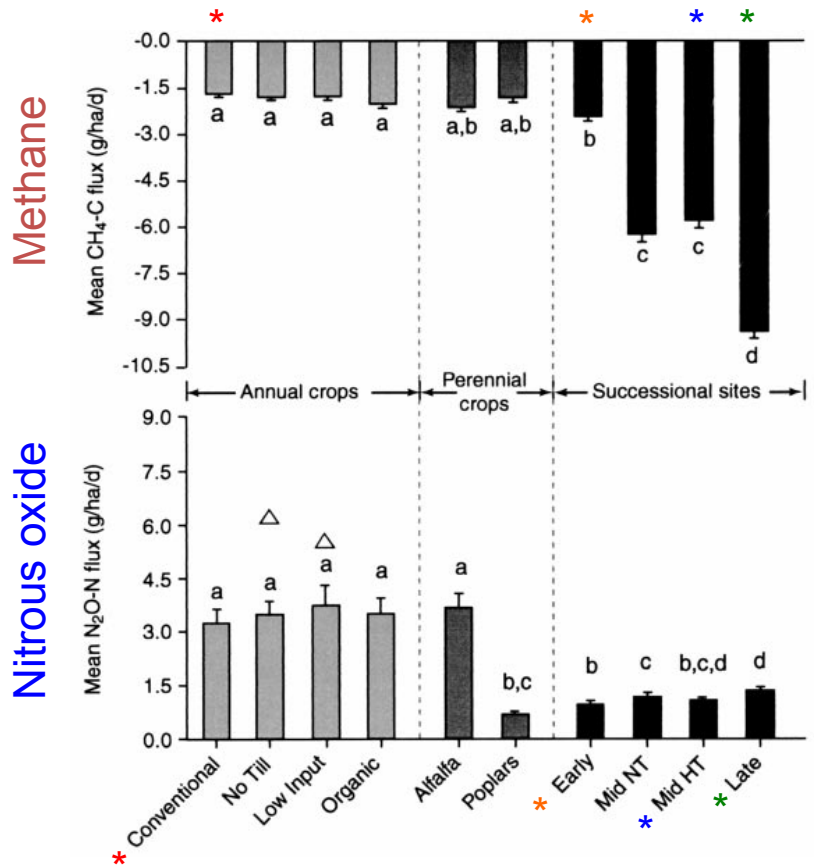
1) Dealing with shotgun metagenomics
2) Short-read annotation & why/why not
3) Annotating soil reads – a story
4) Assembly stories!

A story: looking at land management with 454 shotgun

- Tracy Teal, Vicente Gomez-Alvarado, & Tom Schmidt
- Ask detailed questions of @tracykteal on Twitter, please :)

How do microbial communities change with land management?

Kellogg Biological Station LTER



- AG Conventional Agriculture
- ES Early Successional
- SF Successional Forest
- DF Deciduous Forest

Teal TK, Gomez-Alvarez V, Schmidt TM

Annotating soil reads - thoughts

- Possible to find well-known genes using long (454) reads.
- Normalize for organism abundance!
- Primer independence can be important!
- Note, replicates give you error bars…

A few *assembly* stories

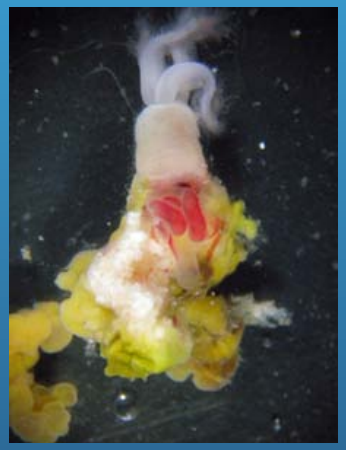
- Low complexity/Osedax symbionts
- High complexity/soil.

Goffredi et al., submitted.

Osedax symbionts

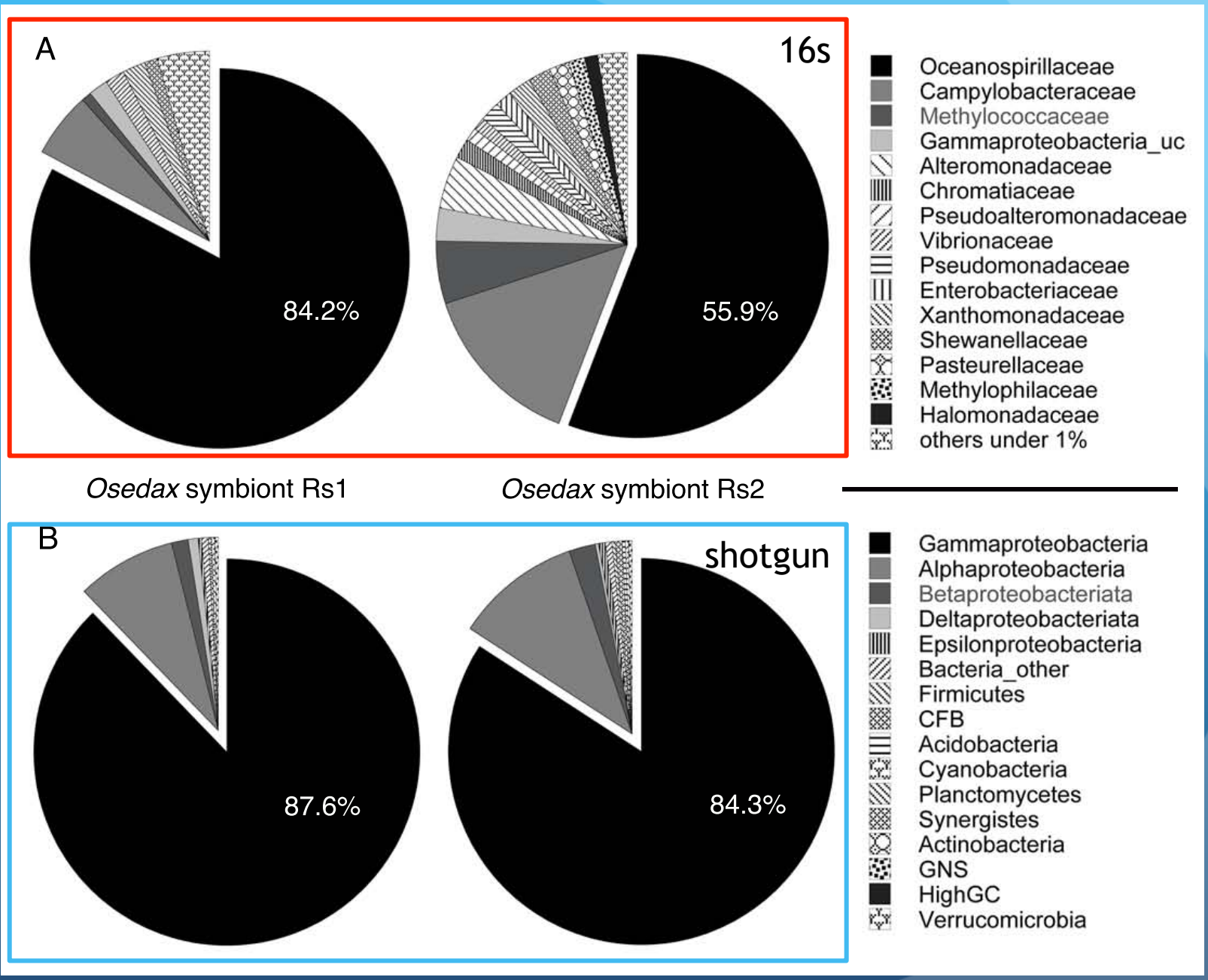
Table S1 | Specimens, and collection sites, used in this study

Site	Dive[1]	Date	Time Zone (months)	# of specimens	*Osedax frankpressi* with Rs1 symbiont
2890m	T486	Oct 2002	8	2	
	T610	Aug 2003	18	3	
	T1069	Jan 2007	59	2	
	DR010	Mar 2009	85	3	
	DR098	Nov 2009	93	4	
	DR204	Oct 2010	104	1	
	DR234	Jun 2011	112	2	
1820m	T1048	Oct 2006	7	3	
	T1071	Jan 2007	10	3	
	DR012	Mar 2009	36	4	
	DR236[2]	Jun 2011	63	2 →	

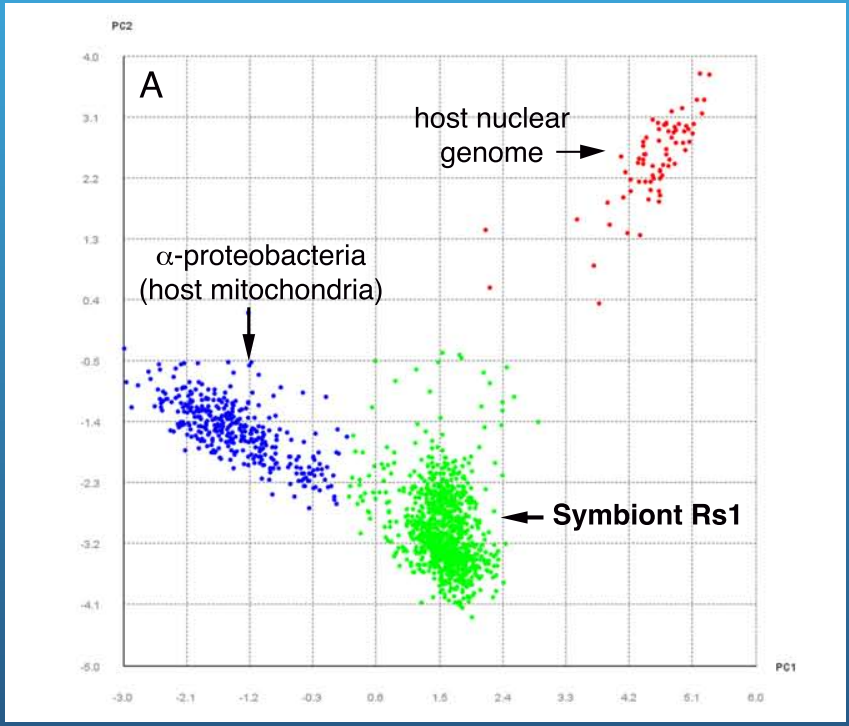


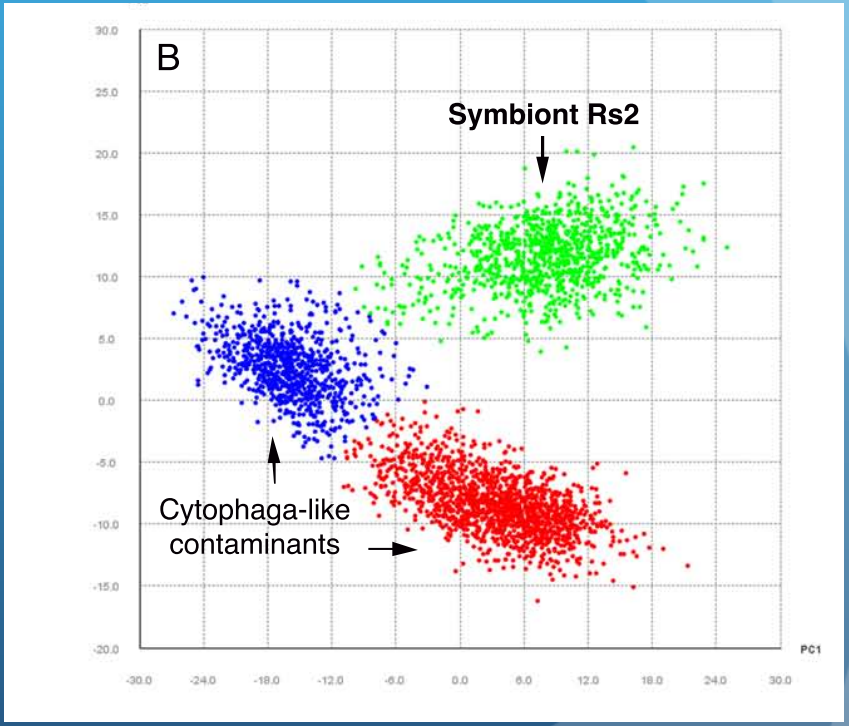
O. frankpressi from DR236

[1]Dive numbers begin with the remotely operated vehicle name; T= *Tiburon*, DR = *Doc Ricketts* (both owned and operated by the Monterey Bay Aquarium Research Institute).
[2]Samples used for genomic analysis were collected during dive DR236.

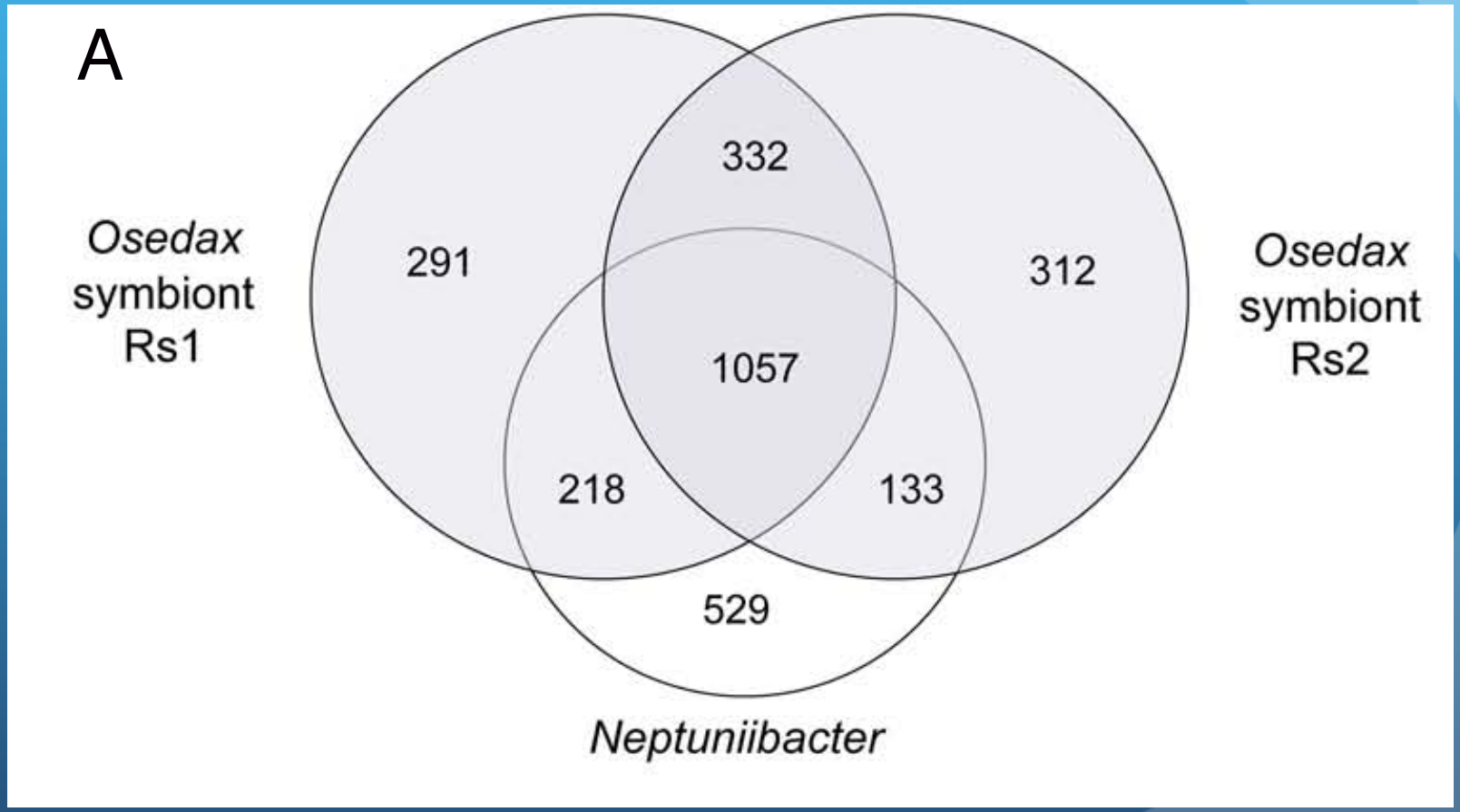

A
16s
84.2%
55.9%
Oceanospirillaceae
Campylobacteraceae
Methylococcaceae
Gammaproteobacteria_uc
Alteromonadaceae
Chromatiaceae
Pseudoalteromonadaceae
Vibrionaceae
Pseudomonadaceae
Enterobacteriaceae
Xanthomonadaceae
Shewanellaceae
Pasteurellaceae
Methylophilaceae
Halomonadaceae
others under 1%
Osedax symbiont Rs1
Osedax symbiont Rs2
B
shotgun
87.6%
84.3%
Gammaproteobacteria
Alphaproteobacteria
Betaproteobacteriata
Deltaproteobacteriata
Epsilonproteobacteria
Bacteria_other
Firmicutes
CFB
Acidobacteria
Cyanobacteria
Planctomycetes
Synergistes
Actinobacteria
GNS
HighGC
Verrucomicrobia

Metagenomic assembly followed by binning enabled isolation of fairly complete genomes





Assembly allowed genomic content comparisons to nearest cultured relative



Osedax assembly story

Conclusions include:

- Osedax symbionts have genes needed for free living stage;
- metabolic versatility in carbon, phosphate, and iron uptake strategies;
- Genome includes mechanisms for intracellular survival, and numerous potential virulence capabilities.

Osedax assembly story

- Low diversity metagenome!
- Physical isolation => MDA => sequencing => diginorm => binning =>
  - 94% complete Rs1
  - 66-89% complete Rs2
- Note: many interesting critters are hard to isolate => so, basically, metagenomes.

Human-associated communities

- "Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization." Sharon et al. (Banfield lab); Genome Res. 2013 23: 111-120

Setup

- Collected 11 fecal samples from premature female, days 15-24.
- 260m 100-bp paired-end Illumina HiSeq reads; 400 and 900 base fragments.
- Assembled > 96% of reads into contigs > 500bp; 8 complete genomes; reconstructed genes down to 0.05% of population abundance.

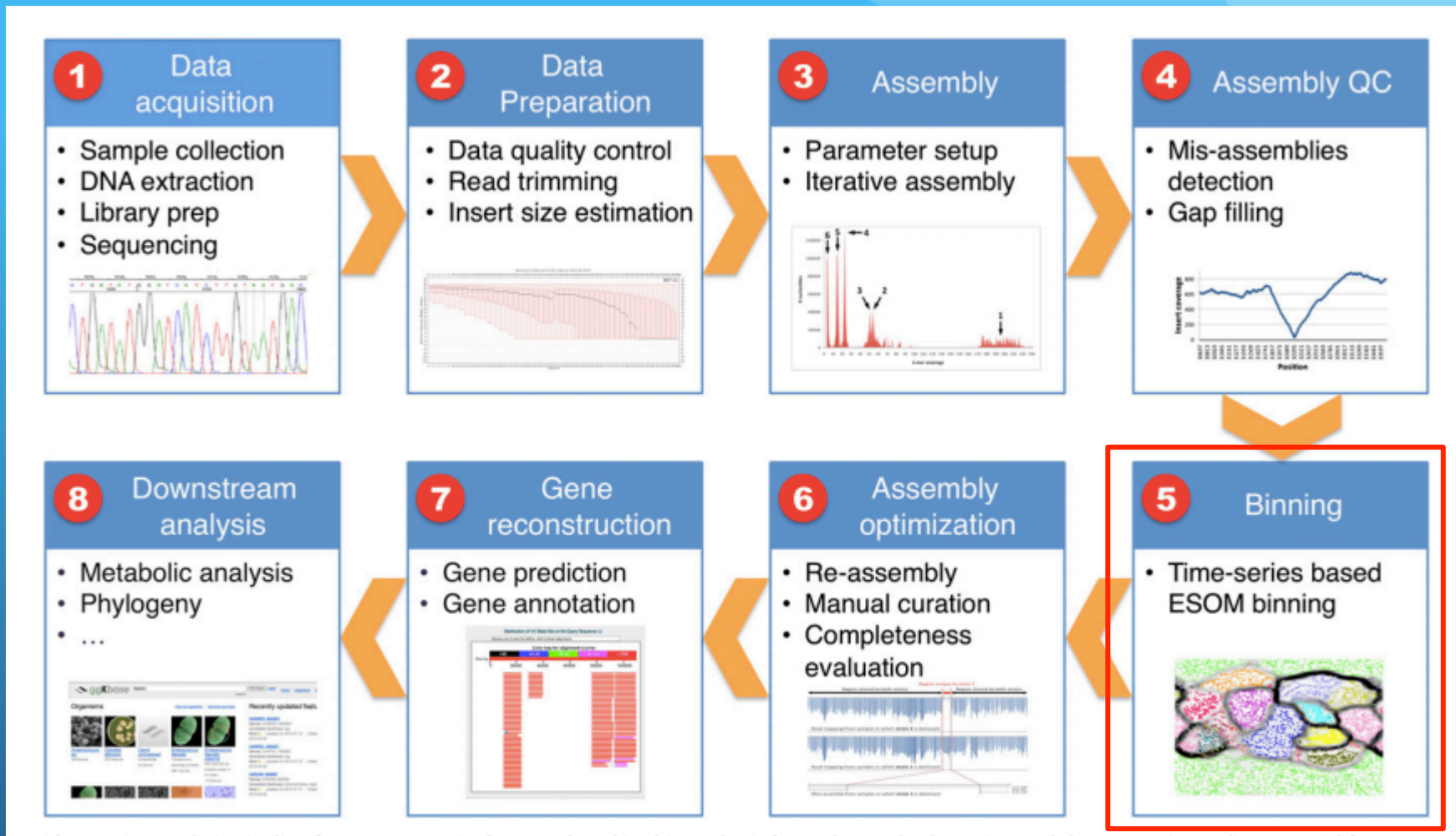
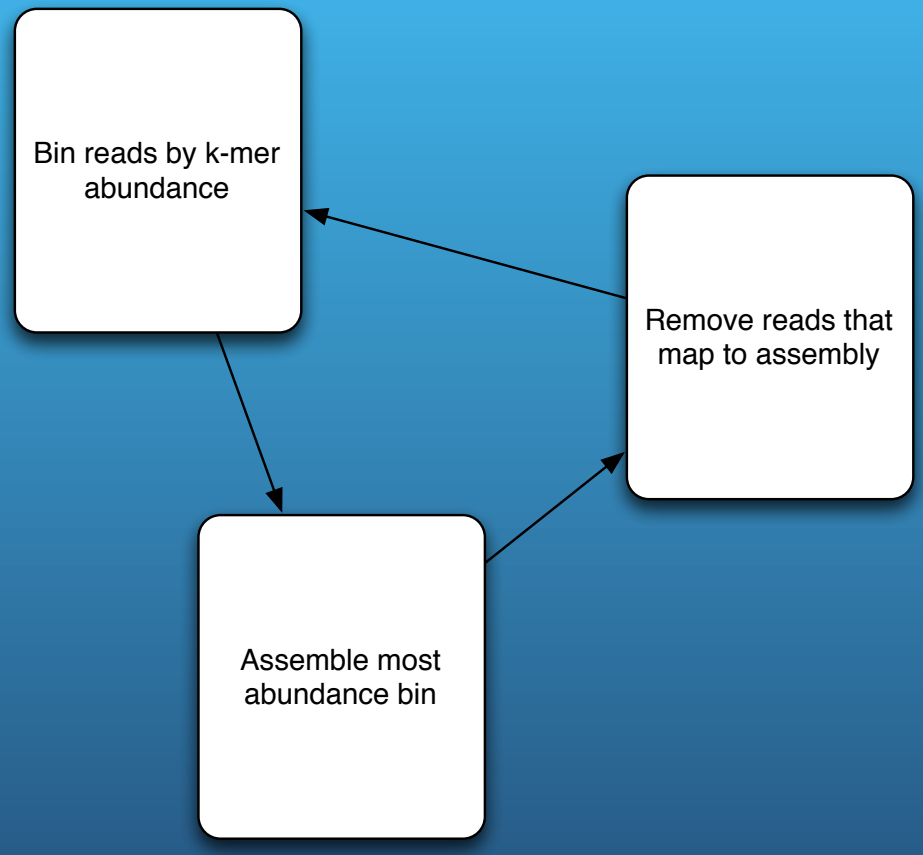


Figure 6. Analysis pipeline for metagenomic data employed in this study. Refer to the Methods section and the supporting online material for more details.

Sharon et al., 2013; pmid 22936250

Key strategy: abundance binning



Sharon et al., 2013; pmid 22936250

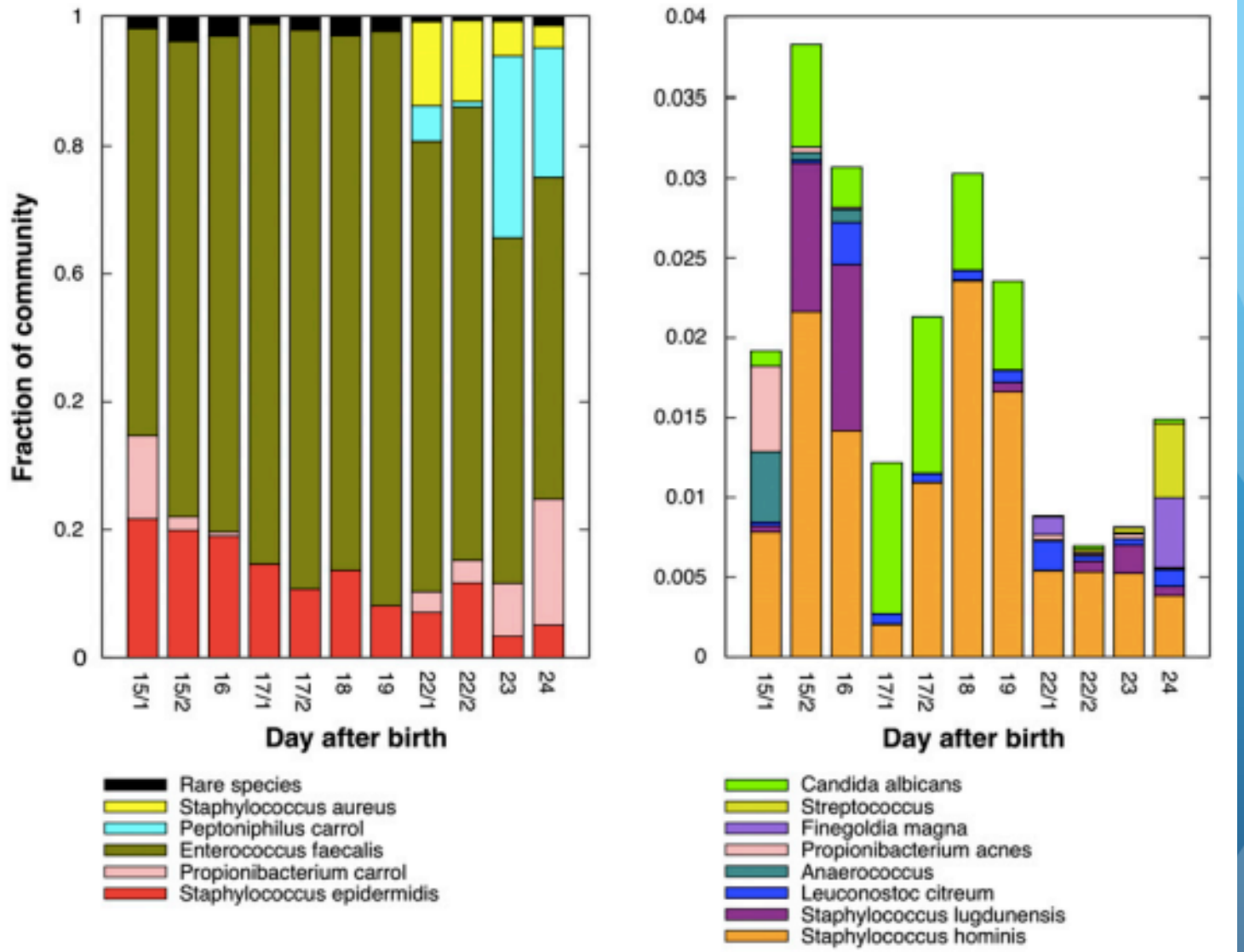


Figure 2. Relative abundance in the community of abundant (*left*) and rare (*right*) species. Abundance was computed based on read mapping to unique regions on the assembled genomes.

Sharon et al., 2013; pmid 22936250

Tracking abundance

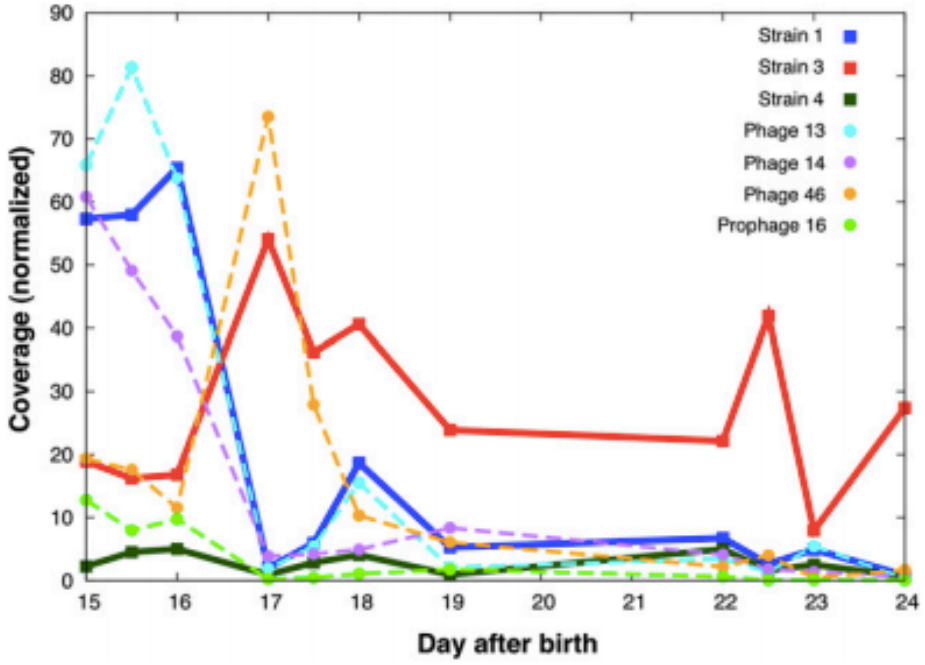


Figure 3. Abundance patterns of the *Staphylococcus epidermidis* strains (thick lines) and their infecting phage (thin dashed lines). All three phages exist primarily as prophages but also were identified as free-existing phages. Insertion positions are in regions that are shared by both abundant strains (and probably also by strain 4), but abundance patterns suggest that phage 13 and 14 infect strain 1, while phage 46 infects strain 3.

Sharon et al., 2013; pmid 22936250

Conclusions

- Recovered strain variation, phage variation, abundance variation, lateral gene transfer.
- Claim that “recovered genomes are superior to draft genomes generated in most isolate genome sequencing projects.”

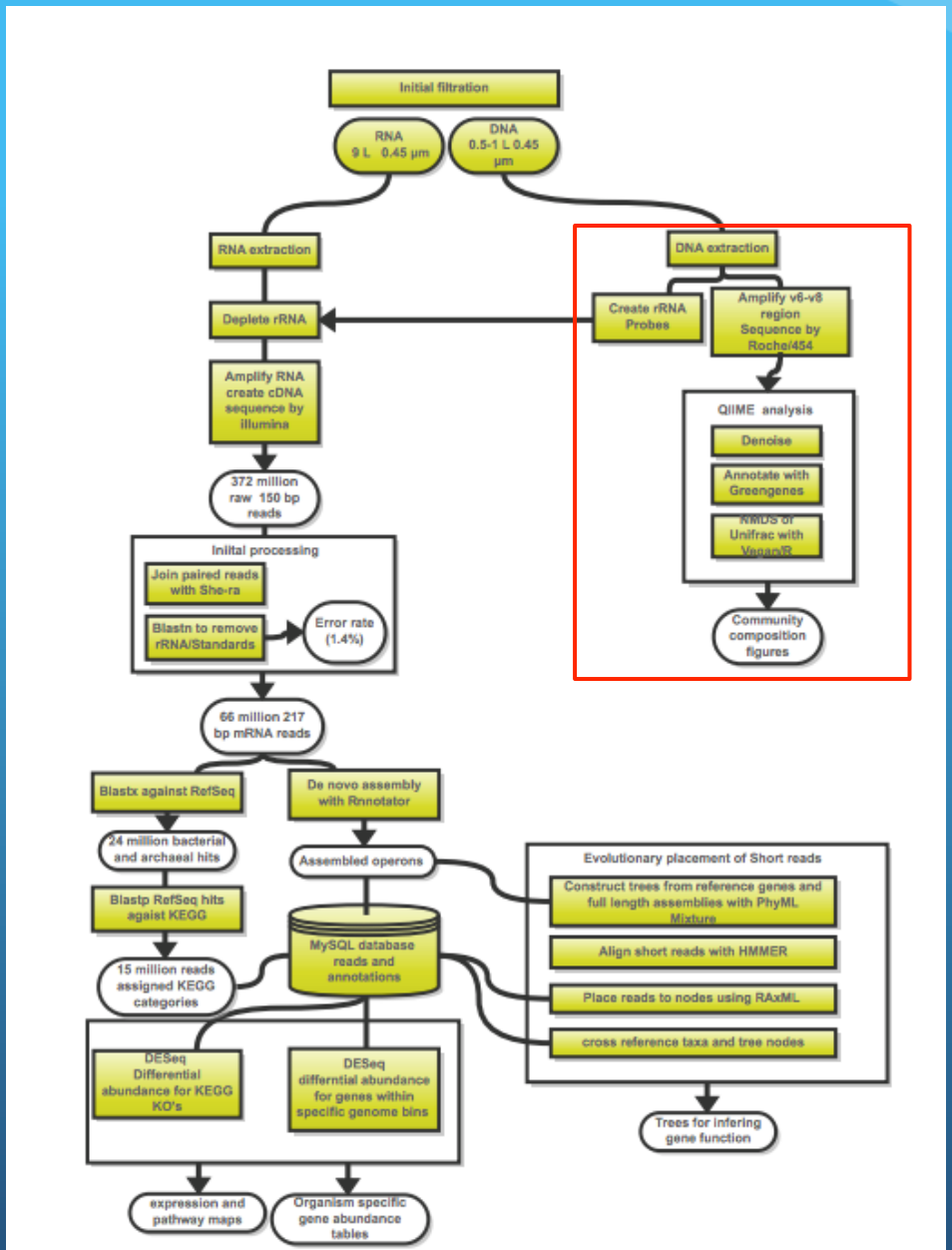
Environmental metagenomics: Deepwater Horizon spill

- "Transcriptional response of bathypelagic marine bacterioplankton to the Deepwater Horizon oil spill." Rivers et al., 2013, Moran lab. Pmid 23902988.

Sequencing strategy

Table 1 Characteristics of metatranscriptome samples

	P16	*P52*	*IP16*	*NP52*
Physical				
Depth (m)	1116	1198	1240	1286
Hydrocarbon plume	Yes	Yes	No	No
Oxygen anomaly	Yes	Yes	Yes	No
Chemical†				
CDOM (relative fluorescence units)	2.14	47.44	0.44	0.44
Methane (nmol l^{-1})	210 171	235 980	4517	106
Nitrate (μmol l^{-1})	21.26	15.14	26.21	27.08
Nitrite (μmol l^{-1})	0.28	0.89	0.13	0.05
Phosphate (μmol l^{-1})	1.46	0.96	1.78	1.72
Biological				
Cell density (cells ml^{-1})	4.95E + 05	1.39E + 05	3.30E + 04	1.74E + 04
Total RNA (μg per l seawater)	1.90E − 04	4.20E − 05	7.12E − 07	3.37E − 07
Sequencing				
Median read length (bases)	217	208	207	233
Total joined reads	28 599 525	23 871 295	22 614 967	29 171 472
rRNA reads (%)	1 780 537 (6%)	9 664 134 (43%)	8 538 576 (36%)	6 104 284 (21%)
Internal standard reads (%)	NA	3 619 342 (16%)	3 918 756 (16%)	5 007 057 (17%)
Possible proteins (%)	26 792 869 (94%)	9 331 491 (39%)	11 413 963 (50%)	18 060 131 (62%)
Archaeal and bacterial hits	10 054 034 (35%)	6 075 577 (25%)	5 611 064 (35%)	1 945 004 (35%)
RefSeq genes	341 983	421 562	428 614	242 768
RefSeq taxa	3627	3628	3742	3485
KEGG assignments	6 141 041	3 781 411	3 680 128	1 097 381



Supplementary Figure S2. Schematic representation of sampling and data analysis.

Protocol for messenger RNA extraction, assembly, downstream analysis, including differential expression.

Note: used overlapping paired-end reads.

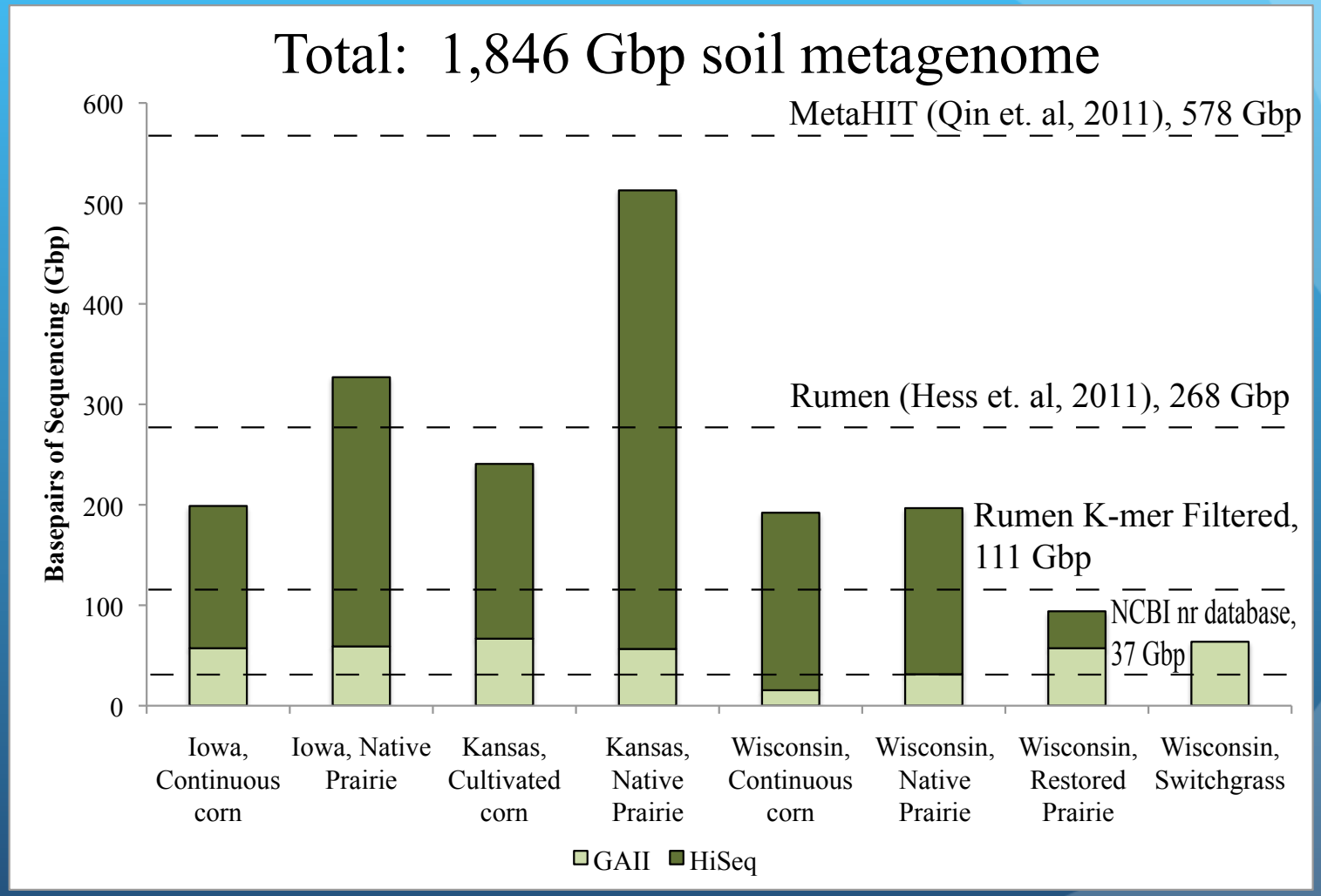
Great Prairie Grand Challenge - *soil*

- Together with Janet Jansson, Jim Tiedje, et al.
- “Can we make sense of soil with deep Illumina sequencing?”

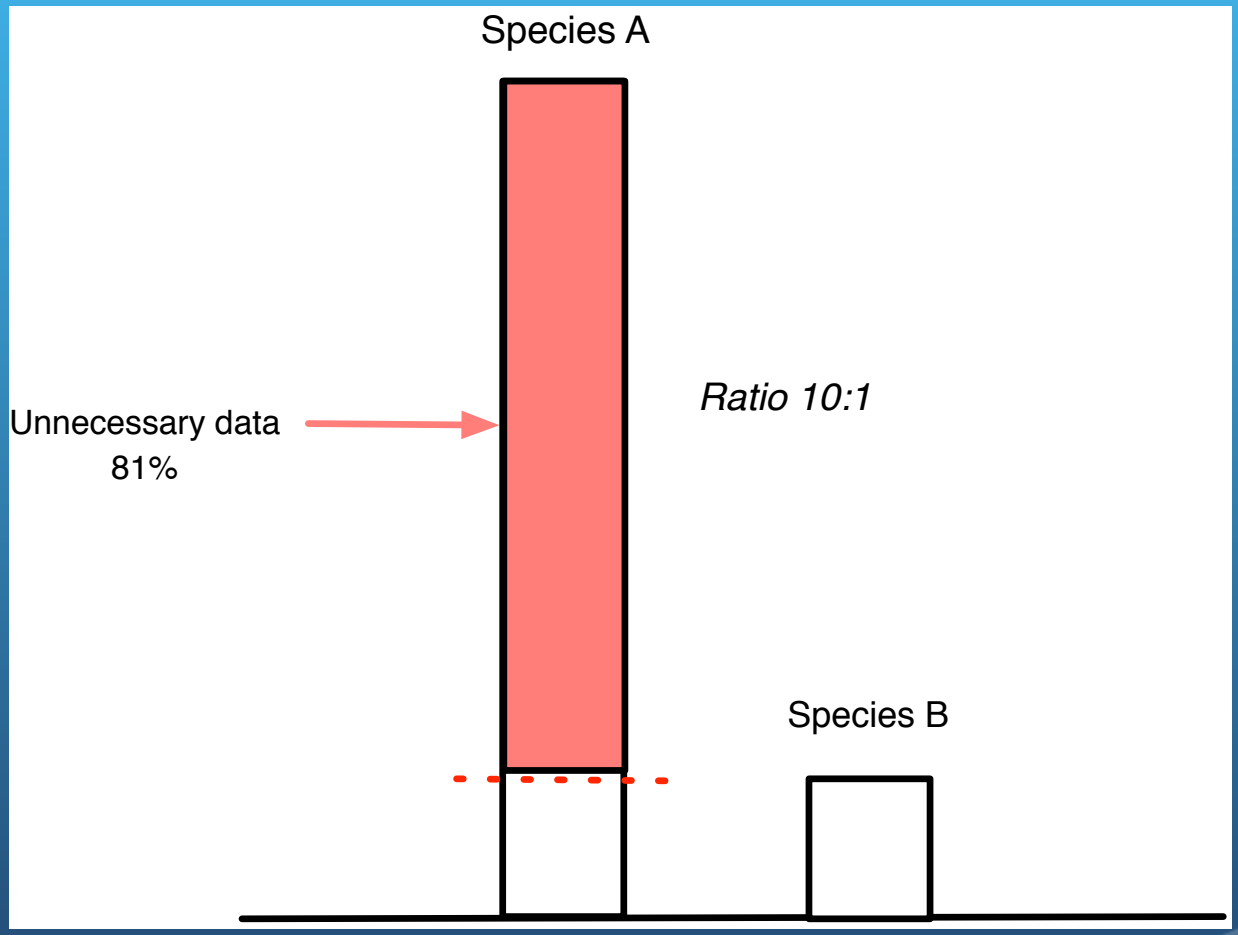
Great Prairie Grand Challenge - *soil*

- What ecosystem level functions are present, and how do microbes do them?
- How does agricultural soil differ from native soil?
- How does soil respond to climate perturbation?
- Questions that are not easy to answer without shotgun sequencing:
  - What kind of strain-level heterogeneity is present in the population?
  - What does the phage and viral population look like?
  - What species are where?

A "Grand Challenge" dataset (DOE/JGI)



Approach 1: Digital normalization
(a computational version of library normalization)



Suppose you have a dilution factor of A (10) to B(1). To get 10x of B you need to get 100x of A! Overkill!!

This 100x will consume disk space and, because of errors, **memory**.

We can discard it for you…

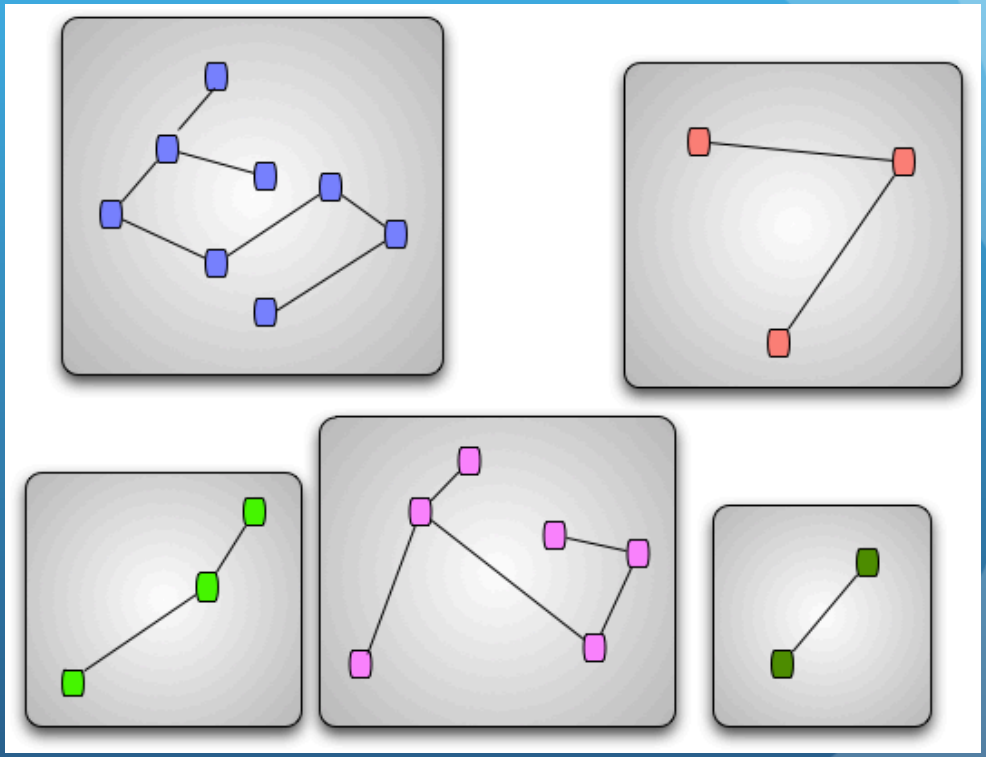
Approach 2: Data partitioning
(a computational version of cell sorting)

Split reads into “bins” belonging to different source species.

Can do this based almost entirely on *connectivity* of sequences.

“Divide and conquer”

Memory-efficient implementation helps to scale assembly.



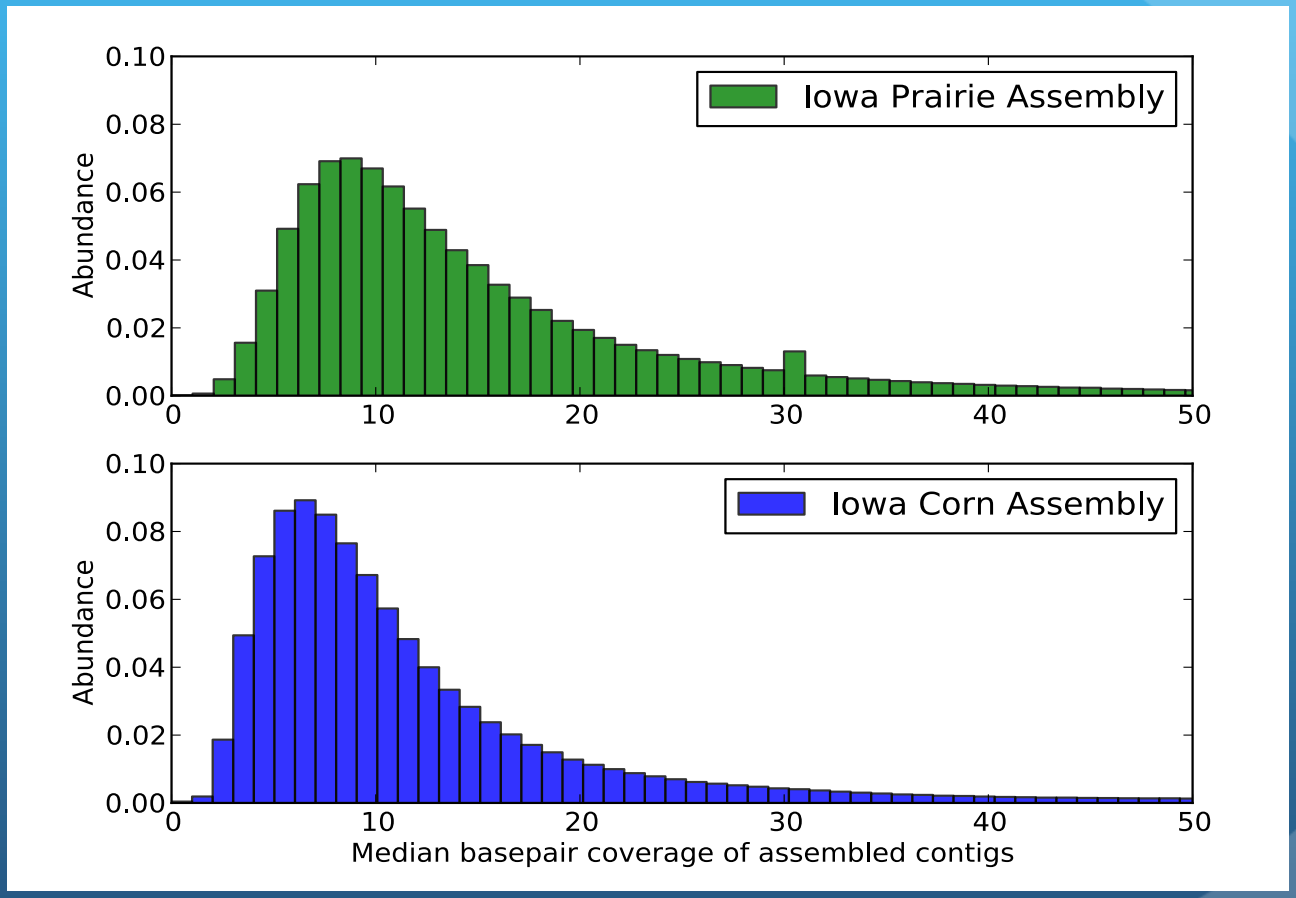
Assembly results for Iowa corn and prairie
(2x ~300 Gbp soil metagenomes)

Total Assembly	Total Contigs (> 300 bp)	% Reads Assembled	Predicted protein coding
2.5 bill	4.5 mill	19%	5.3 mill
3.5 bill	5.9 mill	22%	6.8 mill

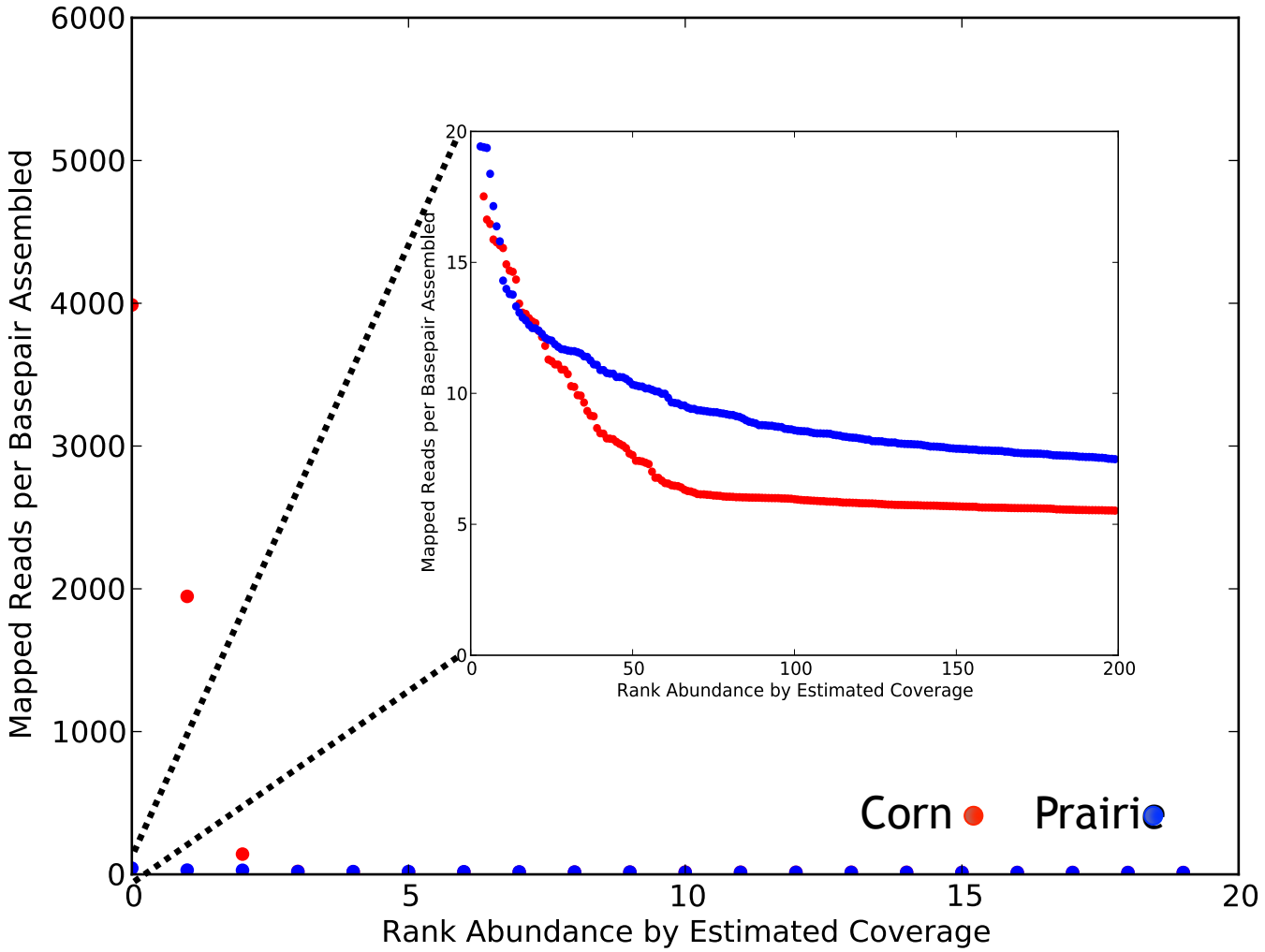
Putting it in perspective:
Total equivalent of ~1200 bacterial genomes
Human genome ~3 billion bp

Adina Howe

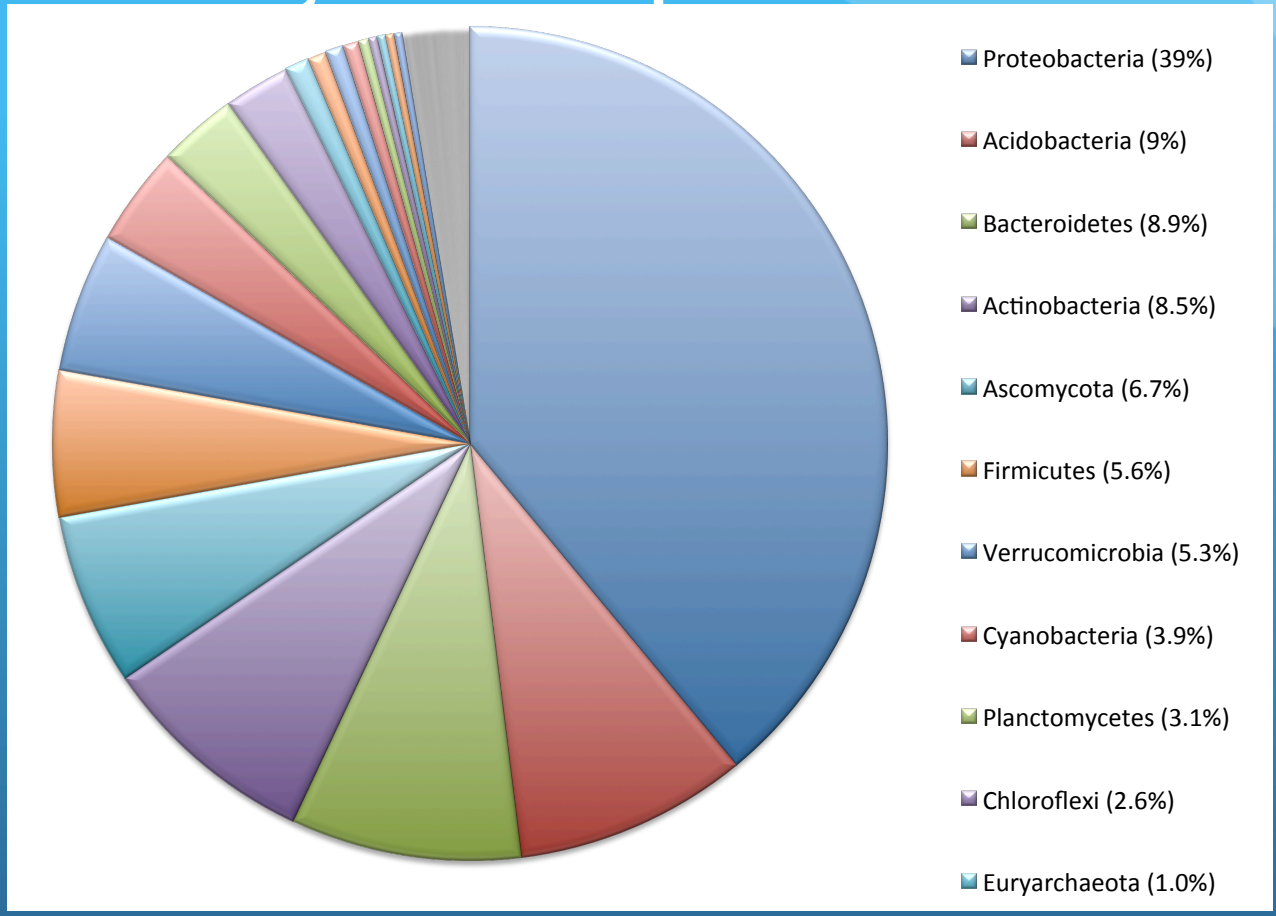
Resulting contigs are low coverage.



Iowa prairie & corn - very even.

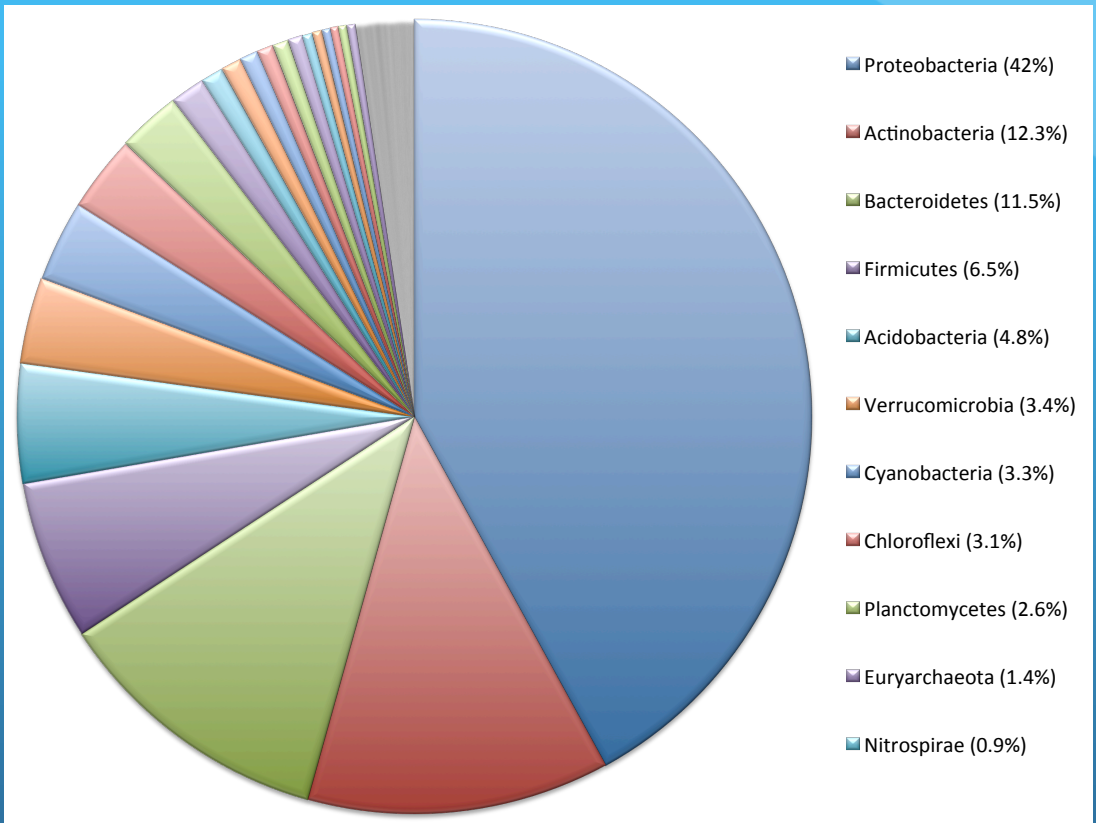


Taxonomy- Iowa prairie



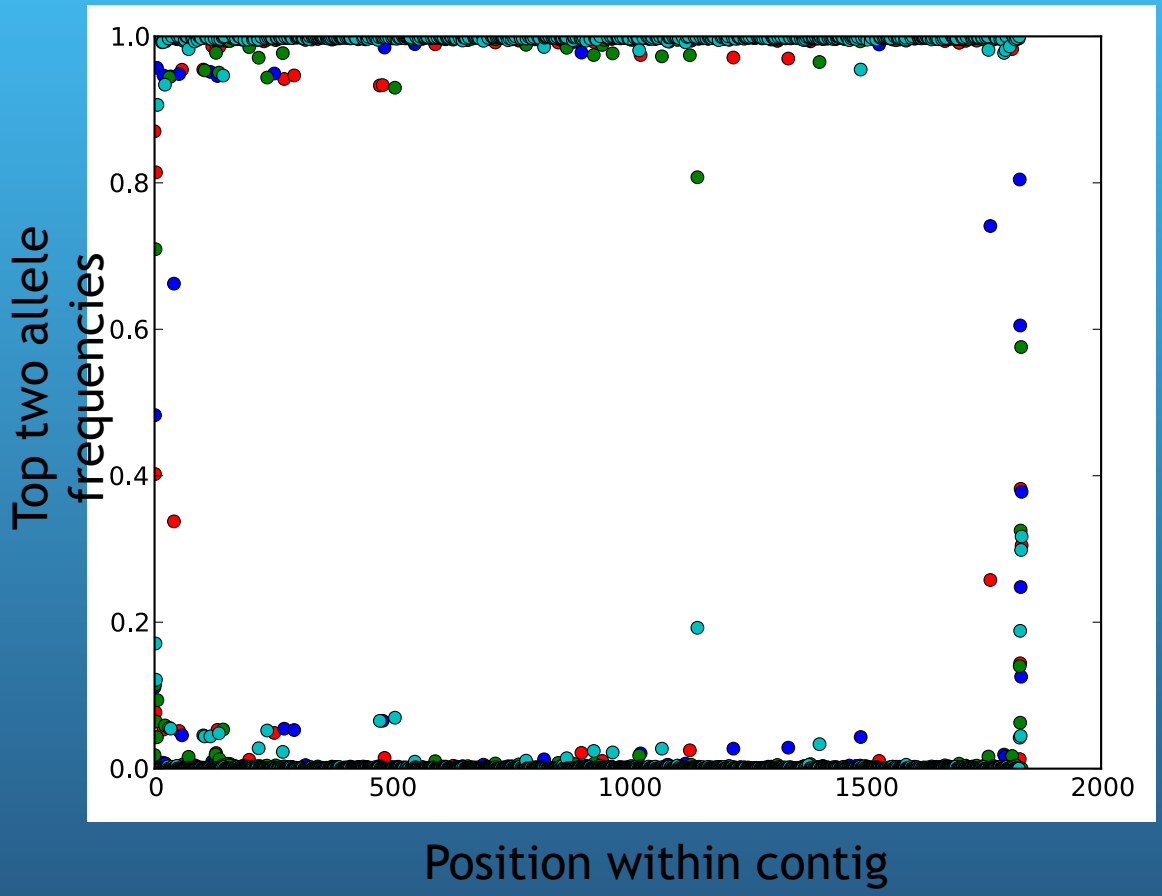
Note: this is predicted taxonomy of contigs w/o considering abundance or length. (MG-RAST)

Taxonomy – Iowa corn



Note: this is predicted taxonomy of contigs w/o considering abundance or length. (MG-RAST)

Strain variation?



Can measure by read mapping.

Of 5000 most abundant contigs, only 1 has a polymorphism rate > 5%

Concluding thoughts on assembly (I)

- There's no standard approach yet; almost every paper uses a specialized pipeline of some sort.
  - More like genome assembly
  - But unlike transcriptomics…

Concluding thoughts on assembly (II)

- Anecdotally, everyone worries about strain variation.
  - Some groups (e.g. Banfield, us) have found that this is not a problem in their system so far.
  - Others (viral metagenomes! HMP!) have found this to be a big concern.

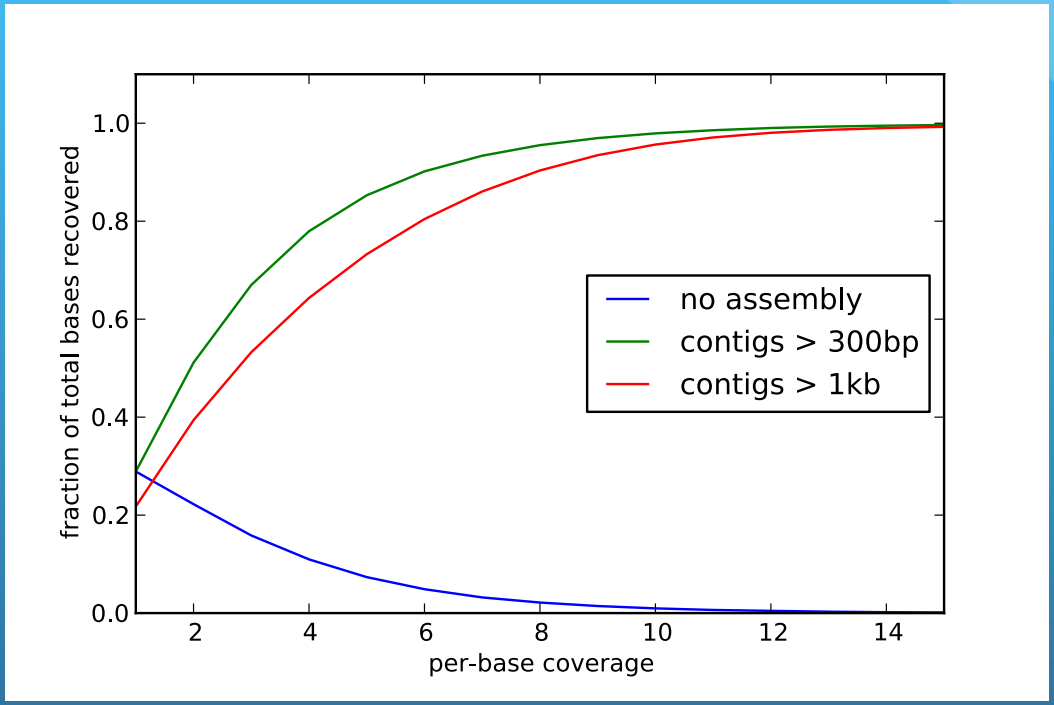
Concluding thoughts on assembly (III)

- Some groups have found metagenome assembly to be very useful.
- Others (us! soil!) have not yet proven its utility.

Questions that you'll have to ask me about later… over gin…

- How much sequencing should I do?
- How do I evaluate metagenome assemblies?
- Which assembler is best?

High coverage is critical.



Low coverage is *the dominant problem* blocking assembly of metagenomes

Materials

- In addition to materials for this class, see:

http://software-carpentry.org/v4/

http://ged.msu.edu/angus/