Shotgun metagenome assembly: how well does it work?

(A moral.)

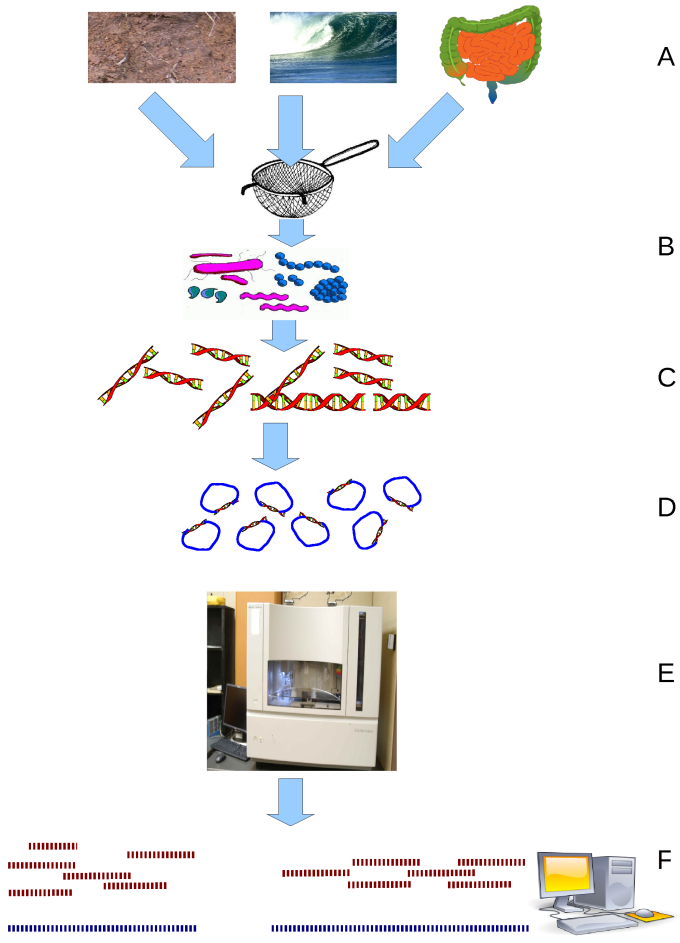
C. Titus Brown

ctbrown@ucdavis.edu

Shotgun metagenomics

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

"Sequence it all and let the bioinformaticians sort it out"



Wikipedia: Environmental shotgun sequencing.png

Goals of shotgun metagenomics

Expand beyond taxonomic/community structure characterization possible with 16s;

Analyze virus, plasmid, strain-level content;

Evaluate metabolic capacity (e.g. “is nirK present?”)

Reconstruct ***genomes*** from metagenomes, if possible.

Shotgun sequencing & *de novo* assembly:

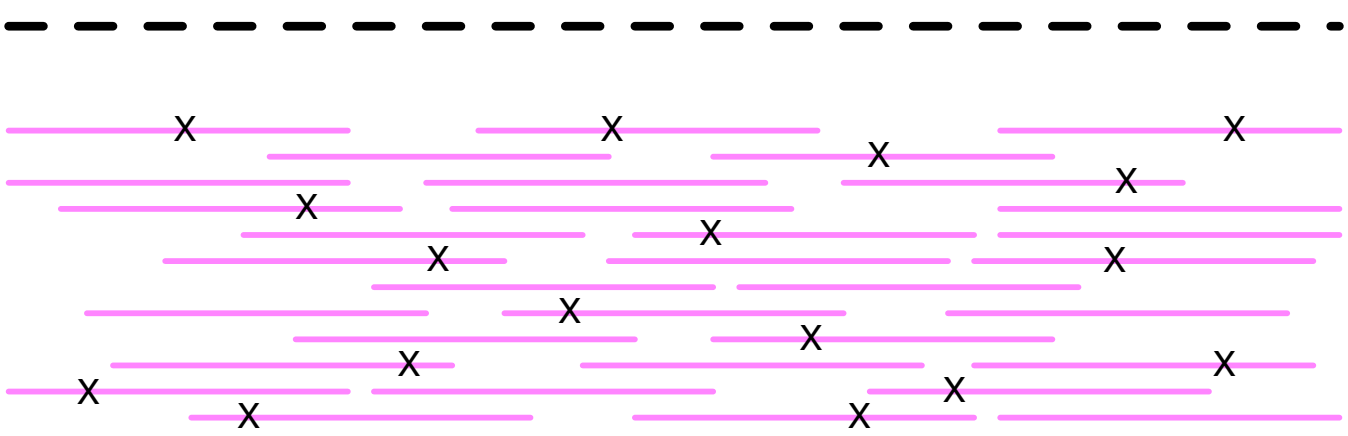
It was the Gest of times, it was the wor
, it was the worst of timZs, it was the
isdom, it was the age of foolisXness
, it was the worVt of times, it was the
mes, it was Ahe age of wisdom, it was th
It was the best of times, it Gas the wor
mes, it was the age of witdom, it was th
isdom, it was tIe age of foolishness

t was the best of times, it was the worst of times, it was the age of wisdom, it was the ag
of foolishness

Shotgun sequencing analogy:
feeding books into a paper shredder, digitizing the shreds, and reconstructing the book.

Although for books, we often know the language and not just the alphabet ☺

Shotgun sequencing

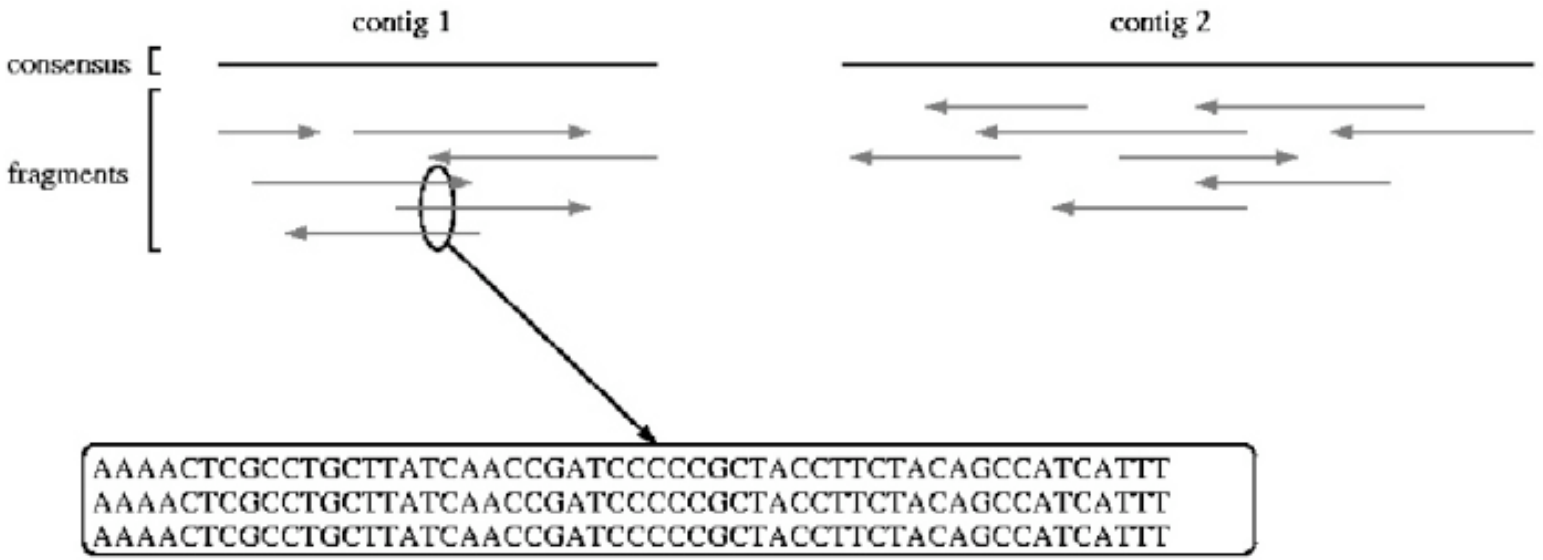


"Coverage" is the average number of reads that overlap each true base in (meta)genome.

Here, the coverage is ~10 – just draw a line straight down from the top through all of the reads.

Shotgun metagenome assembly: reconstruct original genome by finding overlaps in data

Randomly sequencing DNA, then finding overlaps and inferring true sequence:



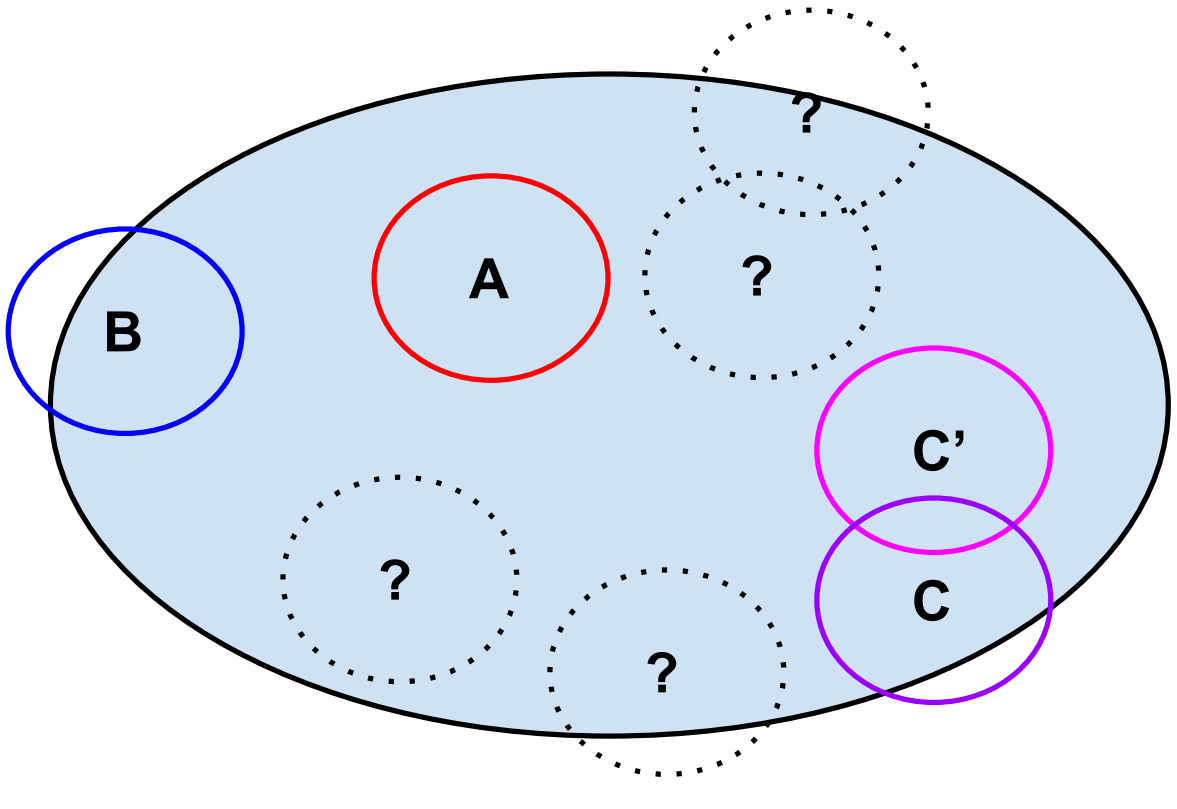
UMD assembly primer (cbcb.umd.ed

Shotgun sequencing & *de novo* assembly:

It was the Gest of times, it was the wor
, it was the worst of timZs, it was the
isdom, it was the age of foolisXness
, it was the worVt of times, it was the
mes, it was Ahe age of wisdom, it was th
It was the best of times, it Gas the wor
mes, it was the age of witdom, it was th
isdom, it was tIe age of foolishness

t was the best of times, it was the worst of times, it was the age of wisdom, it was the ag
of foolishness

Note: Shotgun metagenome data is often incomplete.

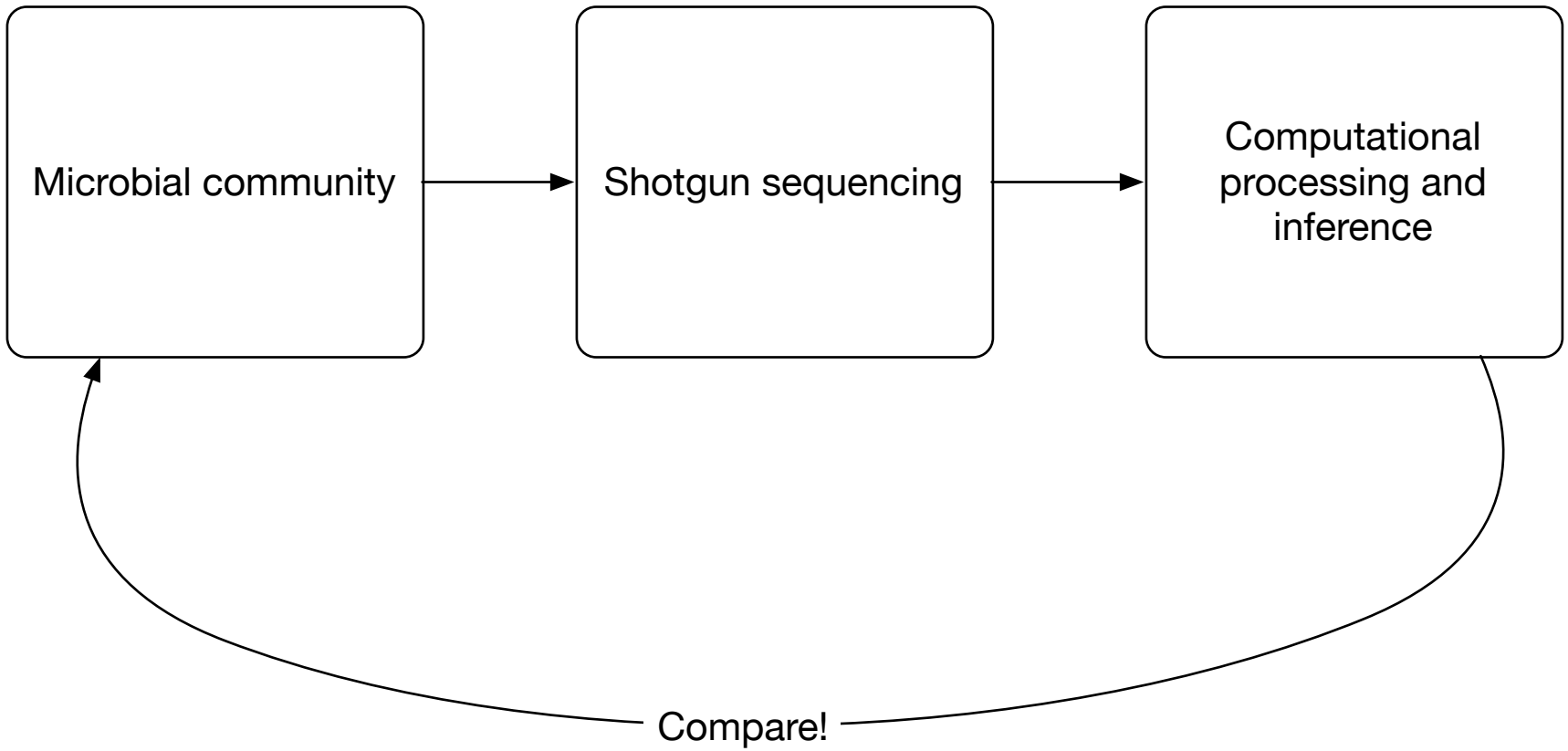


Smaller circles represent (some o
your actual community.

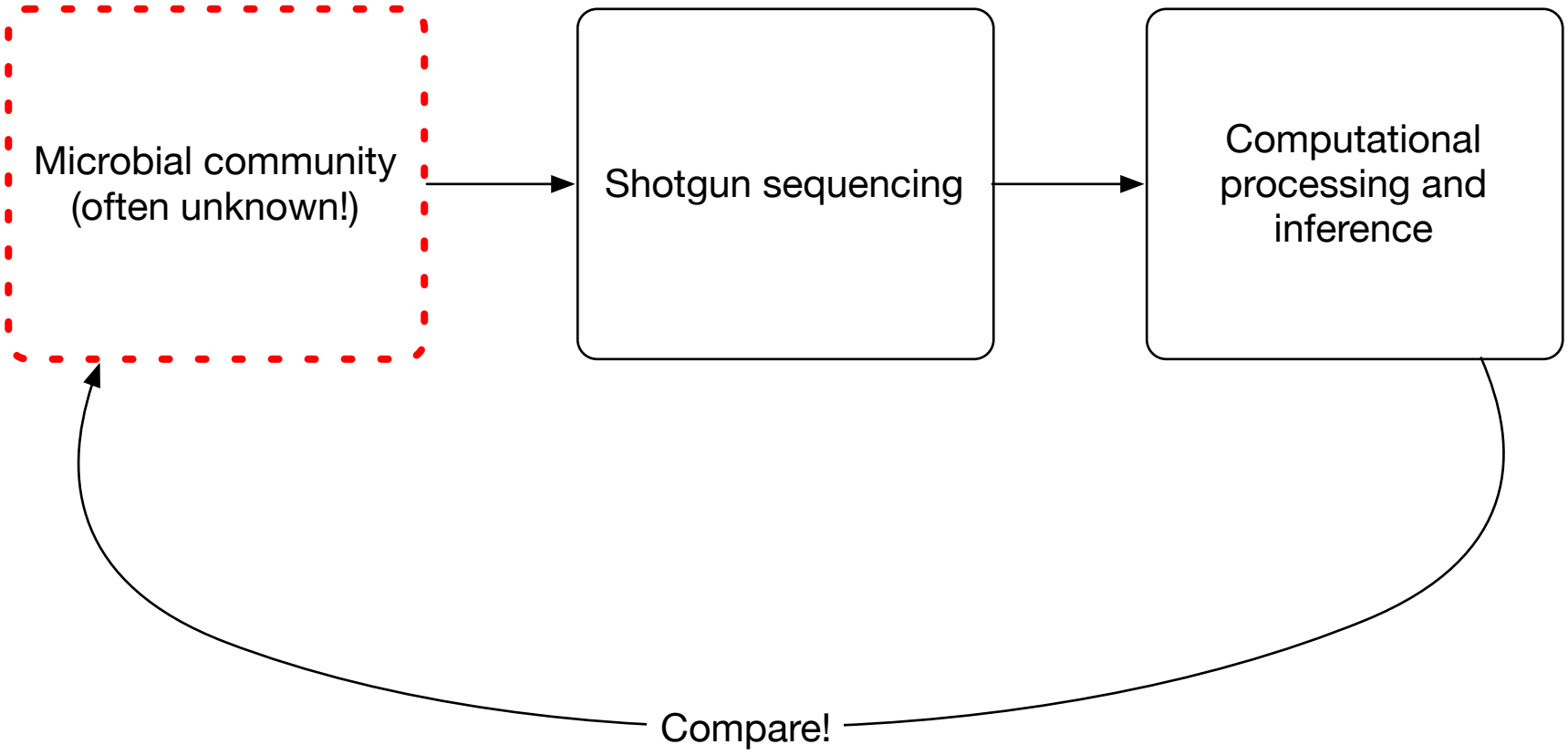
Blue circle represents what's in y
sequencing data set.

Shotgun metagenome data may
contain everything in your
community; may contain strain
variants; may contain "unknown
microbes.

Evaluating metagenome assembly --



Evaluating metagenome assembly --



Goals of shotgun metagenomics

Expand beyond taxonomic/community structure characterization possible with 16s;

Analyze virus, plasmid, strain-level content;

Evaluate metabolic capacity (e.g. “is nirK present?”)

Reconstruct ***genomes*** from metagenomes, if possible.

Questions for evaluation

1. How much of the mock community is in the sequencing data?
2. How well did the assembly recover the reference (big picture)?
3. Did the assemblers mix up sequence ***between*** species?
4. Did the assemblers have problems with ***particular*** species?
5. Did the assemblers recover content ***not*** in the mock community?

Digression: ”preprints”

There is an increasingly broad awareness that scientific publishing is broken in a variety of ways.

- Closed access journals are a blight unto the land;
- Peer review has significant limitations;
- “Journal Impact Factor” concept is fundamentally flawed;
- Focus on novelty over rigor;

One problem with publishing is that it delays broad sharing of work.

- Peer review and publication takes 3 mo - 2 years! Especially if you have to submit to multiple journals!
- In fast-moving fields this is a real problem for progress of field, junior scientists, etc.

Posting papers to bioRxiv and other sites: "preprinting"

Physics has long had a practice of posting papers prior to their submission to a journal; see arxiv.org.

This "preprinting" is standard in some fields −

- Preprinting is considered private scholarly communication;
- Preprints are citable => can gather citations well before pub.
- In some cases, receive comments or exposure; e.g. makes computational tools available (and citable) well before pub.
- Establishes a form of *priority*.

Most biology journals (excluding a few medical journals) now explicitly allow preprints.

See biorxiv.org, PeerJ, etc.

BioRxiv.org:

New Results

Evaluating Metagenome Assembly on a Simple Defined Community with Many Strain Variants

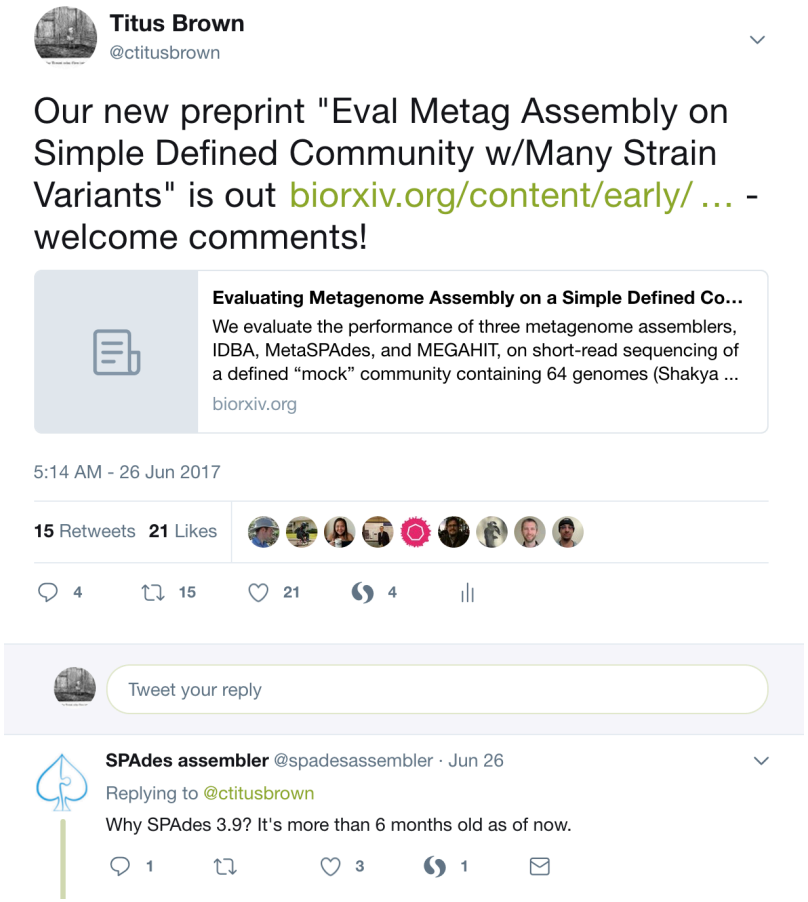
Sherine Awad, Luiz Irber, C. Titus Brown

doi: https://doi.org/10.1101/155358

Abstract

We evaluate the performance of three metagenome assemblers, IDBA, MetaSPAdes, and

=> pre-submission comments, extra refs.



Questions for evaluation

1. How much of the mock community is in the sequencing data?
2. How well did the assembly recover the reference (big picture)?
3. Did the assemblers mix up sequence ***between*** species?
4. Did the assemblers have problems with ***particular*** species?
5. Did the assemblers recover content ***not*** in the mock community?

Questions for evaluation

1. **How much of the mock community is in the sequencing data?**
2. How well did the assembly recover the reference (big picture)?
3. Did the assemblers mix up sequence ***between*** species?
4. Did the assemblers have problems with ***particular*** species?
5. Did the assemblers recover content ***not*** in the mock community?

How much of the known metagenome is *theoretically* reconstructable?

Table 1: Jaccard containment of the reference in the reads

k-mer size	% reference in reads
21	96.8%
31	95.9%
41	94.9%
51	94.1%

At least one genome is *mostly* missing…

Table 2: Top uncovered genomes

Genome	**Read coverage**
Desulfovibrio vulgaris DP4	93.2%
Thermus thermophilus HB27	91.1%
Enterococcus faecalis V583	74.6%
Fusobacterium nucleatum	47.6%

Many genomes are missing > 1% of content.

Table 3: Genomes removed from reference for low 51-mer presence

51-mers in reads	**Genome**
98.7	*Leptothrix cholodnii*
98.7	*Haloferax volcanii* DS2
98.6	*Salinispora tropica* CNB-440
97.4	*Deinococcus radiodurans*
97.2	*Zymomonas mobilis*
97.1	*Ruegeria pomeroyi*
96.8	*Shewanella baltica* OS223
95.5	*B. bronchiseptica* D989
94.5	*Burkholderia xenovorans*
72.0	*Desulfovibrio vulgaris* DP4
65.0	*Thermus thermophilus* HB27
53.4	*Enterococcus faecalis*
4.7	*Fusobacterium nucleatum* ATCC 25586

=> Remove from further consideration for accuracy and completeness metrics.

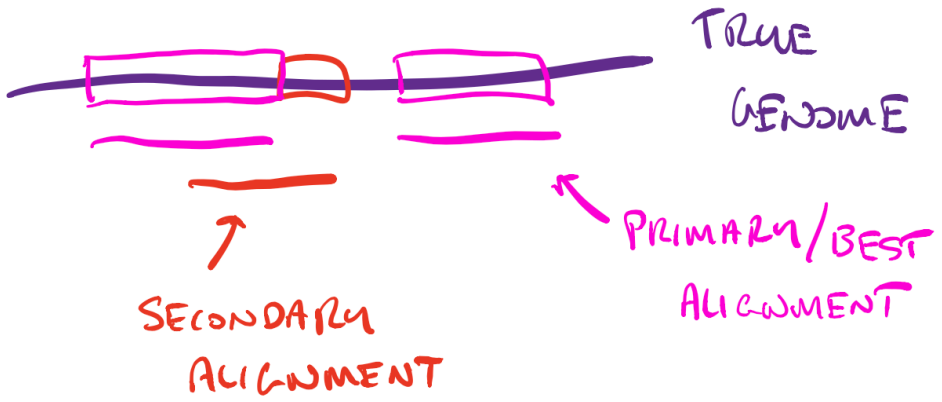
Questions for evaluation

1. How much of the mock community is in the sequencing data?
2. **How well did the assembly recover the reference (big picture)?**
3. Did the assemblers mix up sequence ***between*** species?
4. Did the assemblers have problems with ***particular*** species?
5. Did the assemblers recover content ***not*** in the mock community?

How much pink + red?

Table 6: Contig coverage of reference with loose alignment conditions.

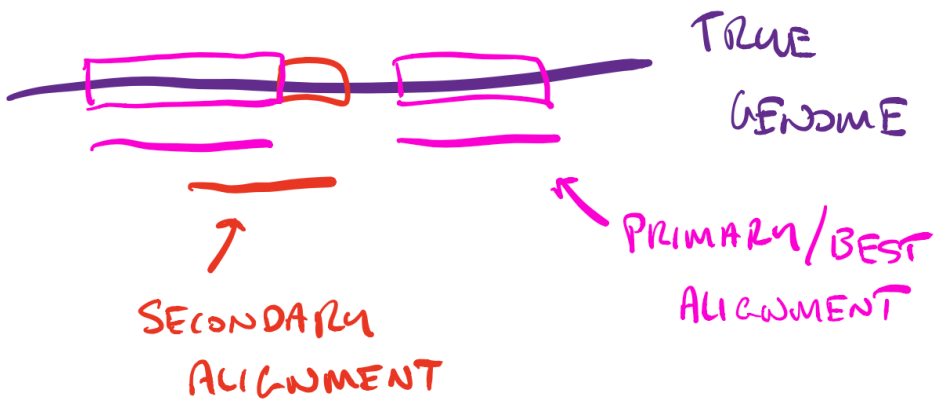
Assembly	bases aligned	duplication	51-mers
MEGAHIT	94.8%	1.0%	96.7%
MetaSPAdes	93.1%	1.1%	96.2%
IDBA	93.6%	0.98%	97.2%



How much just pink?

Table 7: Contig accuracy measured by reference coverage with strict alignment.

Assembly	% covered
MEGAHIT	89.3%
IDBA	87.7%
MetaSPAdes	83.4%



The “truth” lies between, somewhere.

Table 6: Contig coverage of reference with loose alignment conditions.

Assembly	**bases aligned**	**duplication**	**51-mers**
MEGAHIT	94.8%	1.0%	96.7%
MetaSPAdes	93.1%	1.1%	96.2%
IDBA	93.6%	0.98%	97.2%

Table 7: Contig accuracy measured by reference coverage with strict alignment.

Assembly	**% covered**
MEGAHIT	89.3%
IDBA	87.7%
MetaSPAdes	83.4%

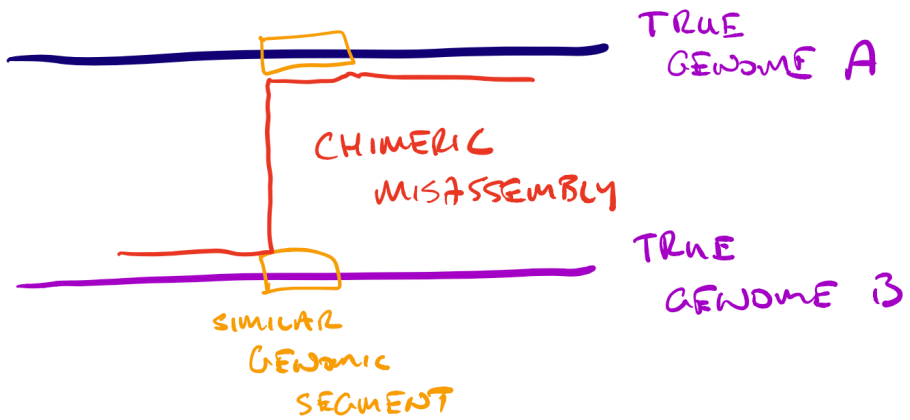
Questions for evaluation

1. How much of the mock community is in the sequencing data?
2. How well did the assembly recover the reference (big picture)?
3. Did the assemblers mix up sequence ***between*** species?
4. Did the assemblers have problems with ***particular*** species?
5. Did the assemblers recover content ***not*** in the mock community?

Assembly rarely makes cross-species mistakes (chimeric contigs)

Table 8: Chimeric contigs by contig length.

Assembly	**> 50kb**	**> 5kb**	**> 500 bp**
IDBA	0	1	7 (0.06%)
MEGAHIT	1	4	14 (0.13%)
MetaSPAdes	0	3	30 (0.48%)



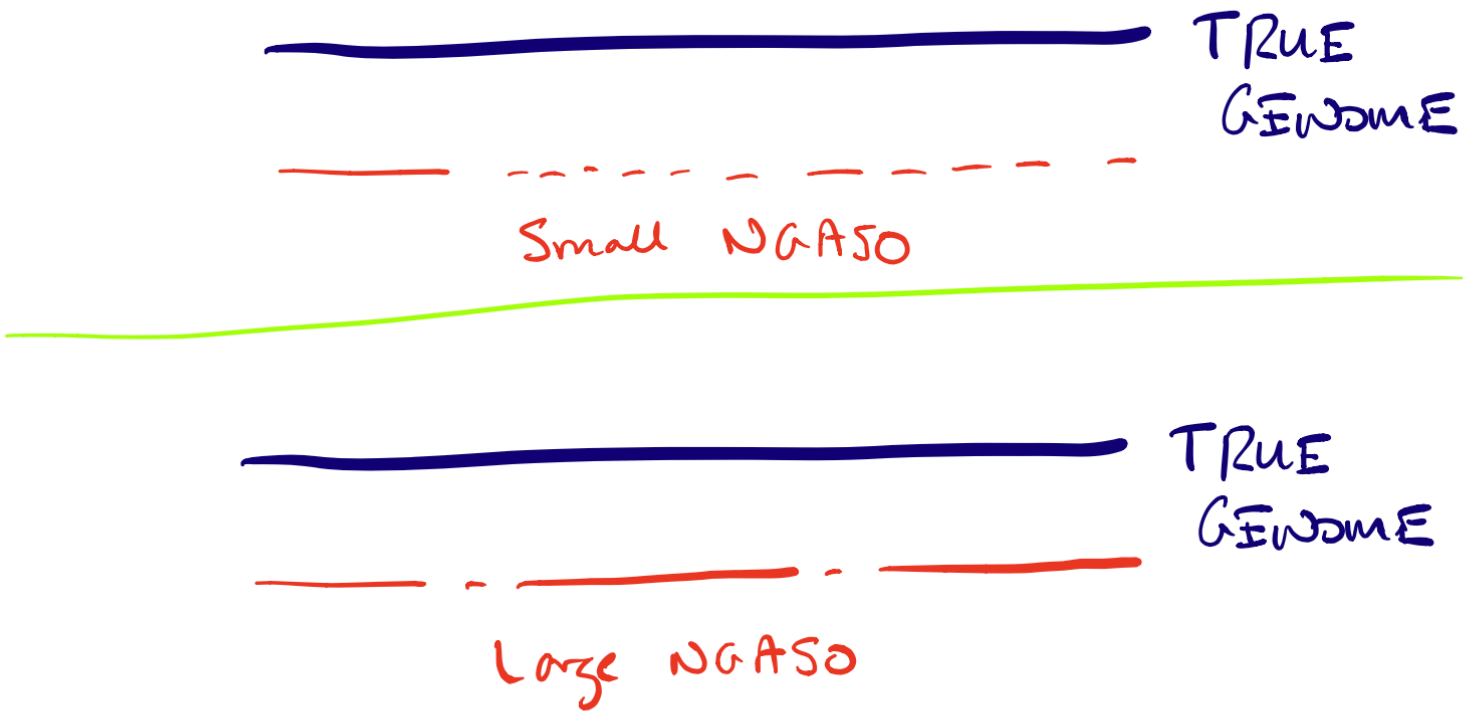
Questions for evaluation

1. How much of the mock community is in the sequencing data?
2. How well did the assembly recover the reference (big picture)?
3. Did the assemblers mix up sequence ***between*** species?
4. **Did the assemblers have problems with *particular* species?**
5. Did the assemblers recover content ***not*** in the mock community?

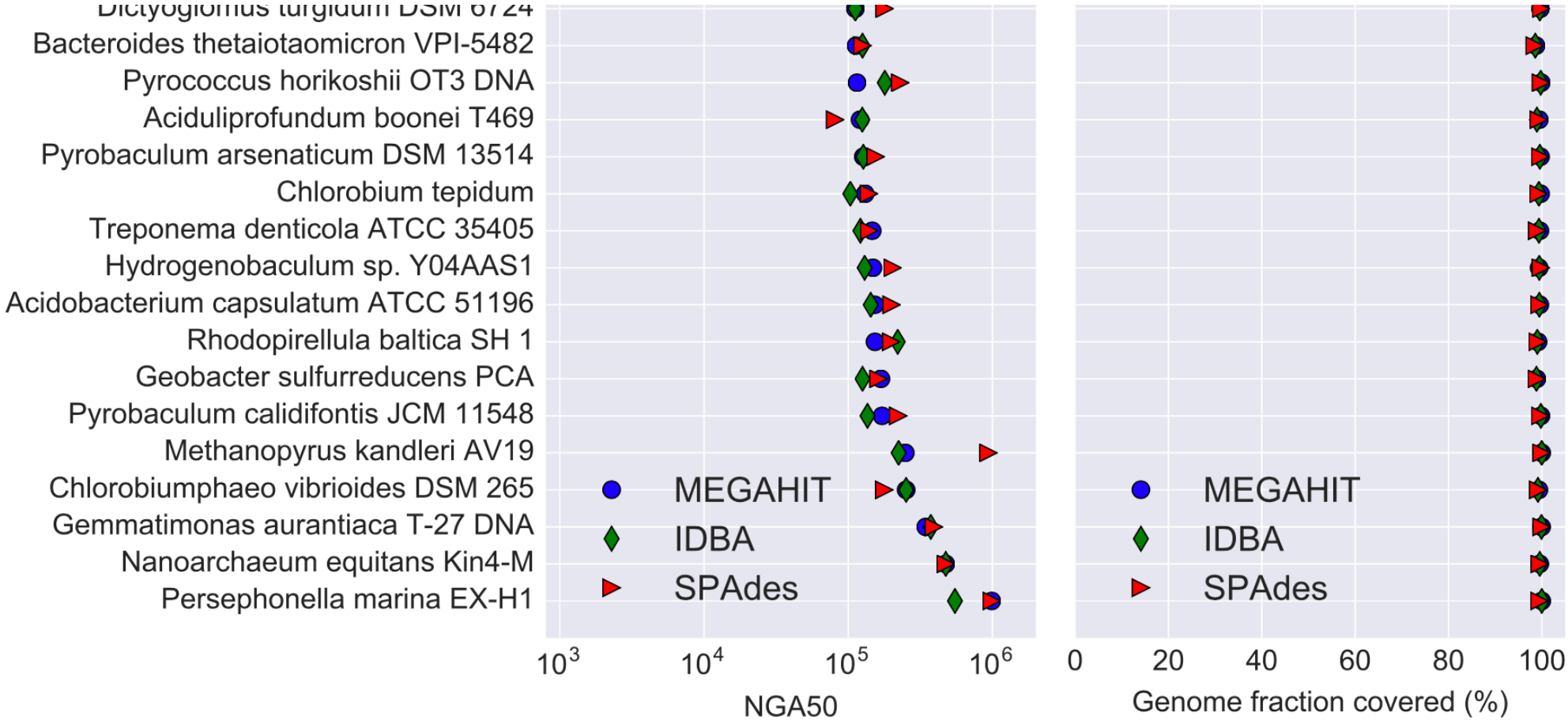
Per-genome measurements of recovery vs assembler –

NGA50 (left)
% recovered (right)

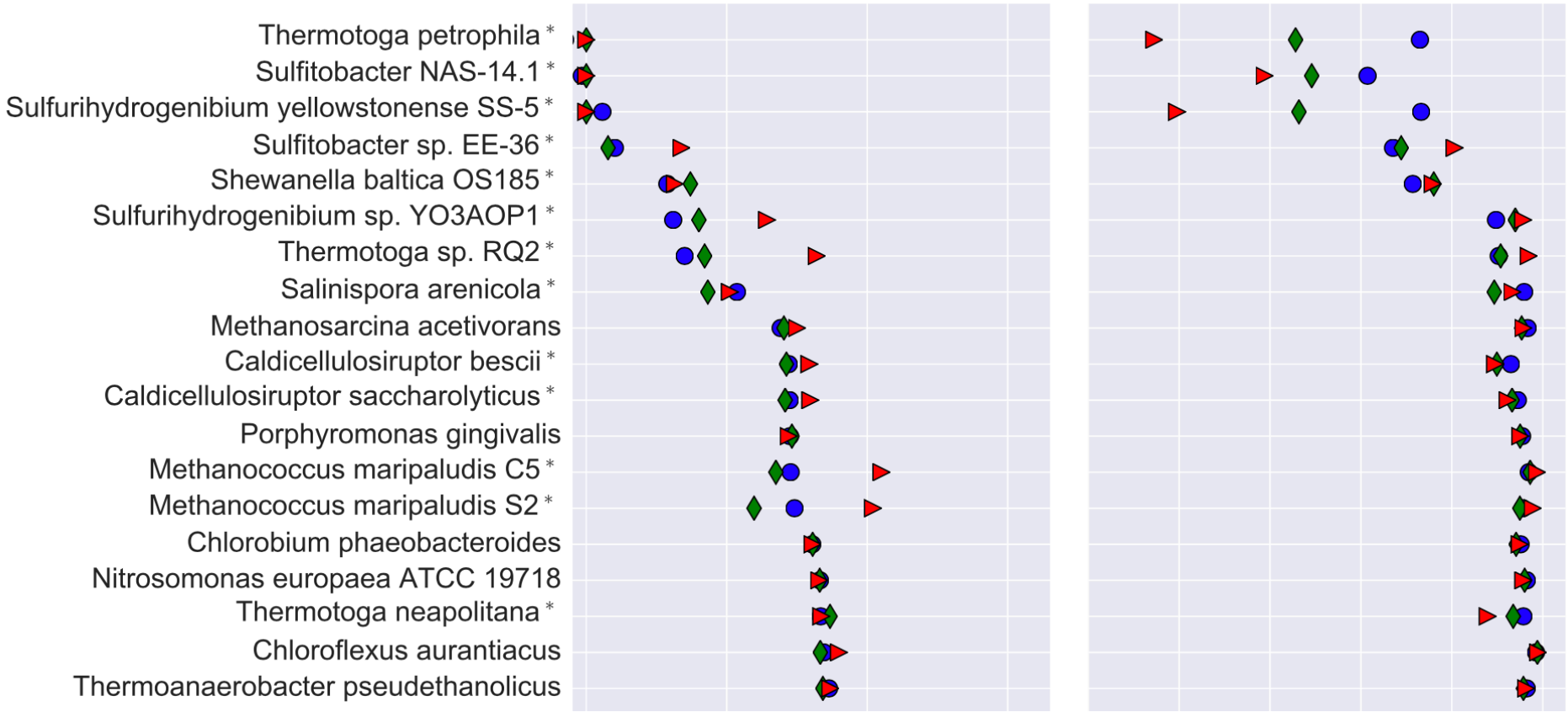
NGA50 and % recovered --



Many genomes are really well recovered.



..some genomes are NOT recovered well.



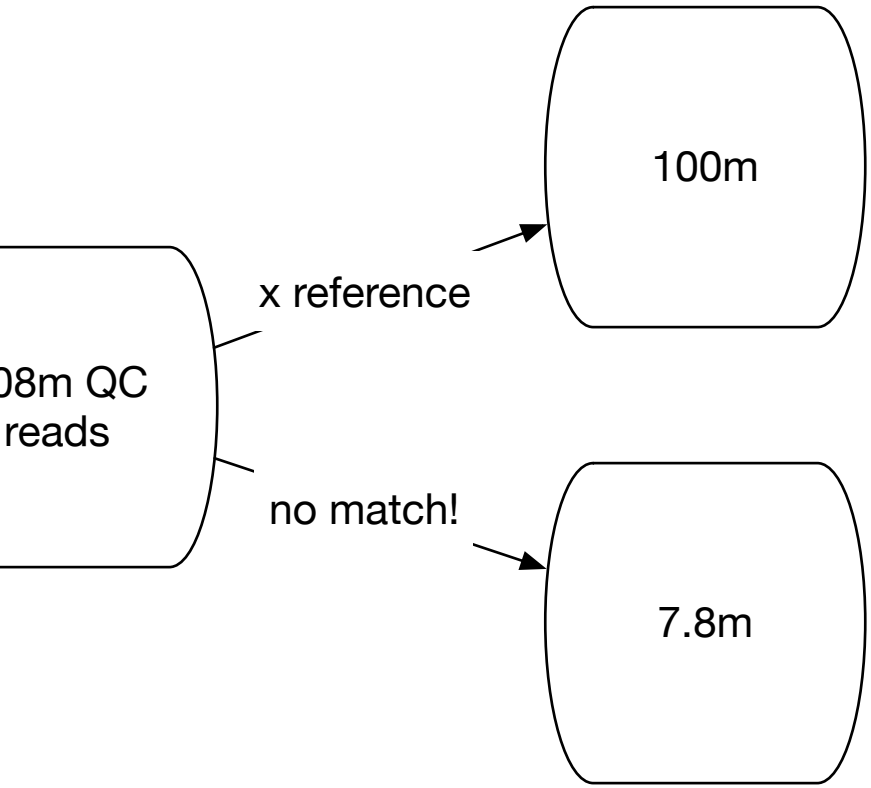
NGA50 (bigger better)

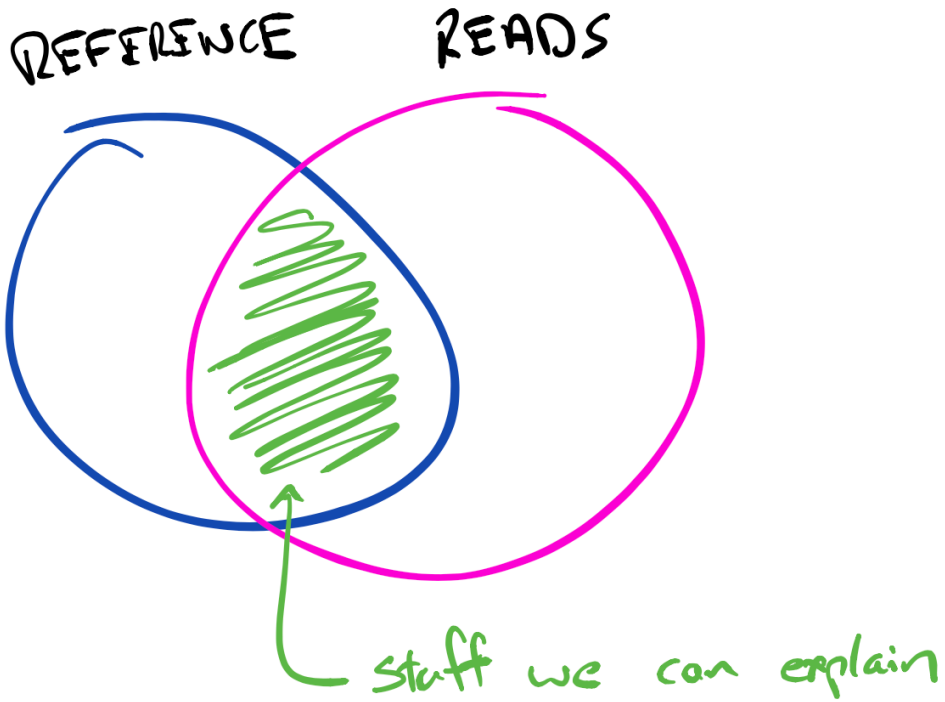
% genome recovered

Questions for evaluation

1. How much of the mock community is in the sequencing data?
2. How well did the assembly recover the reference (big picture)?
3. Did the assemblers mix up sequence ***between*** species?
4. Did the assemblers have problems with ***particular*** species?
5. **Did the assemblers recover content *not* in the mock community?**

Something else I didn't tell you yet:
7.8m reads *didn't* map to the full reference (!?)





(not to scale)

What's in these unmapped reads??

Table 9: GenBank genomes detected in assembly of unmapped reads

match	**GenBank genome**
44.1%	*Fusobacterium sp.* OBRC1
23.0%	*P. ruminis strain* ML2
18.2%	*Thermus thermophilus* HB8
7.7%	*P. ruminis strain* CGMCC
8.2%	*Enterococcus faecalis* M7
7.3%	*F. nucleatum* 13_3C
3.7%	*F. nucleatum subsp. polymorphum*
2.9%	*Fusobacterium hwasookii*
1.0%	*E. coli isolate* YS
1.7%	*F. nucleatum subsp. polymorphum*, alt.
1.9%	*F. nucleatum subsp. vincentii*

Three of the top four uncovered genomes are n the unmapped reads

nk genomes detected in assembly of unmapped reads

ch	GenBank genome
%	*Fusobacterium sp.* OBRC1
%	*P. ruminis strain* ML2
%	*Thermus thermophilus* HB8
%	*P. ruminis strain* CGMCC
%	*Enterococcus faecalis* M7
%	*F. nucleatum* 13_3C
%	*F. nucleatum subsp. polymorphum*
%	*Fusobacterium hwasookii*
%	*E. coli isolate* YS
%	*F. nucleatum subsp. polymorphum*, alt.
%	*F. nucleatum subsp. vincentii*

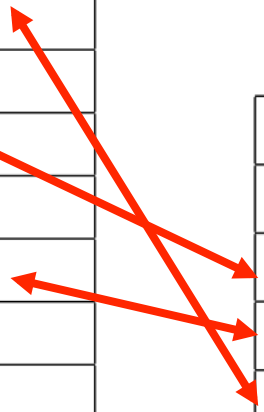
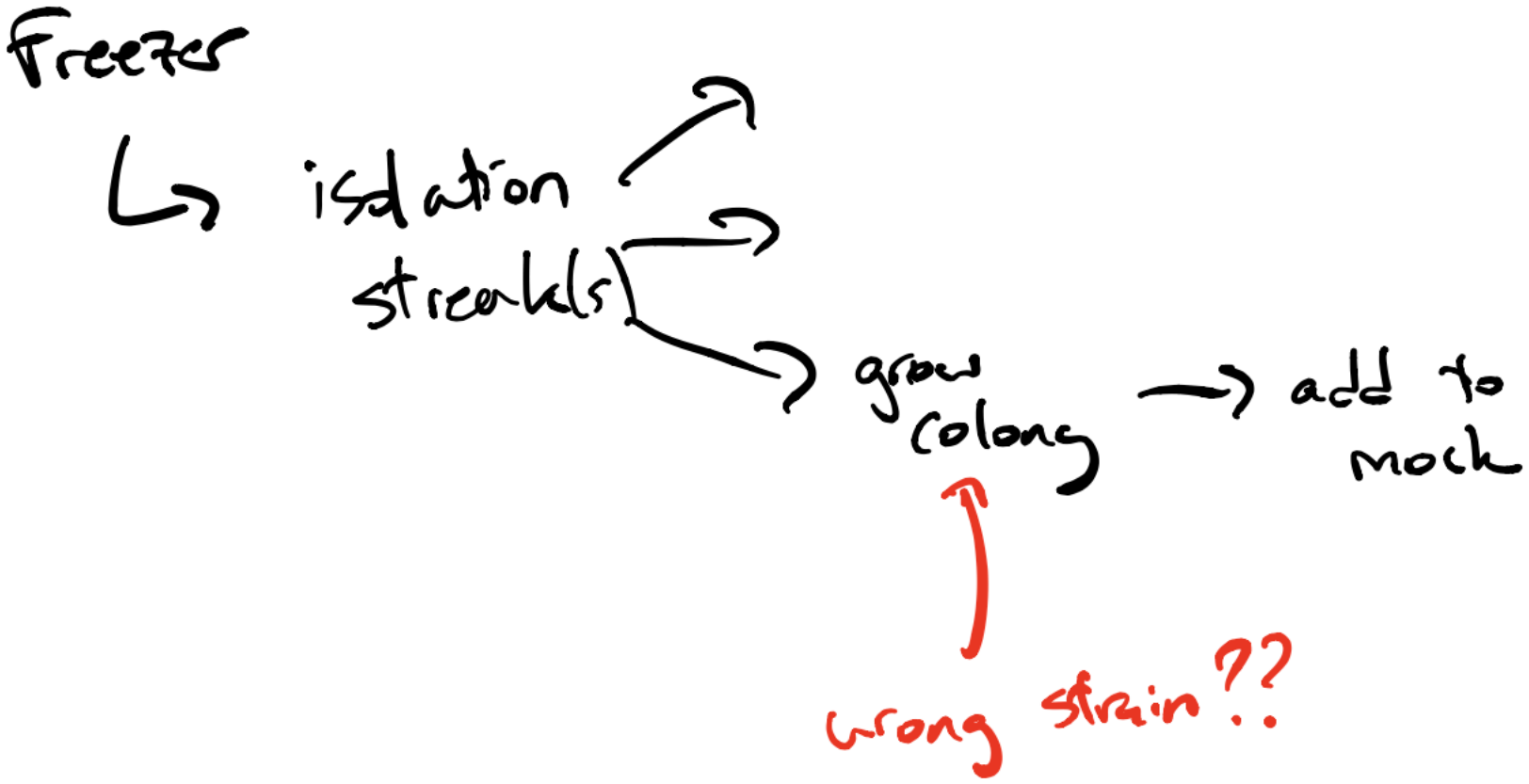


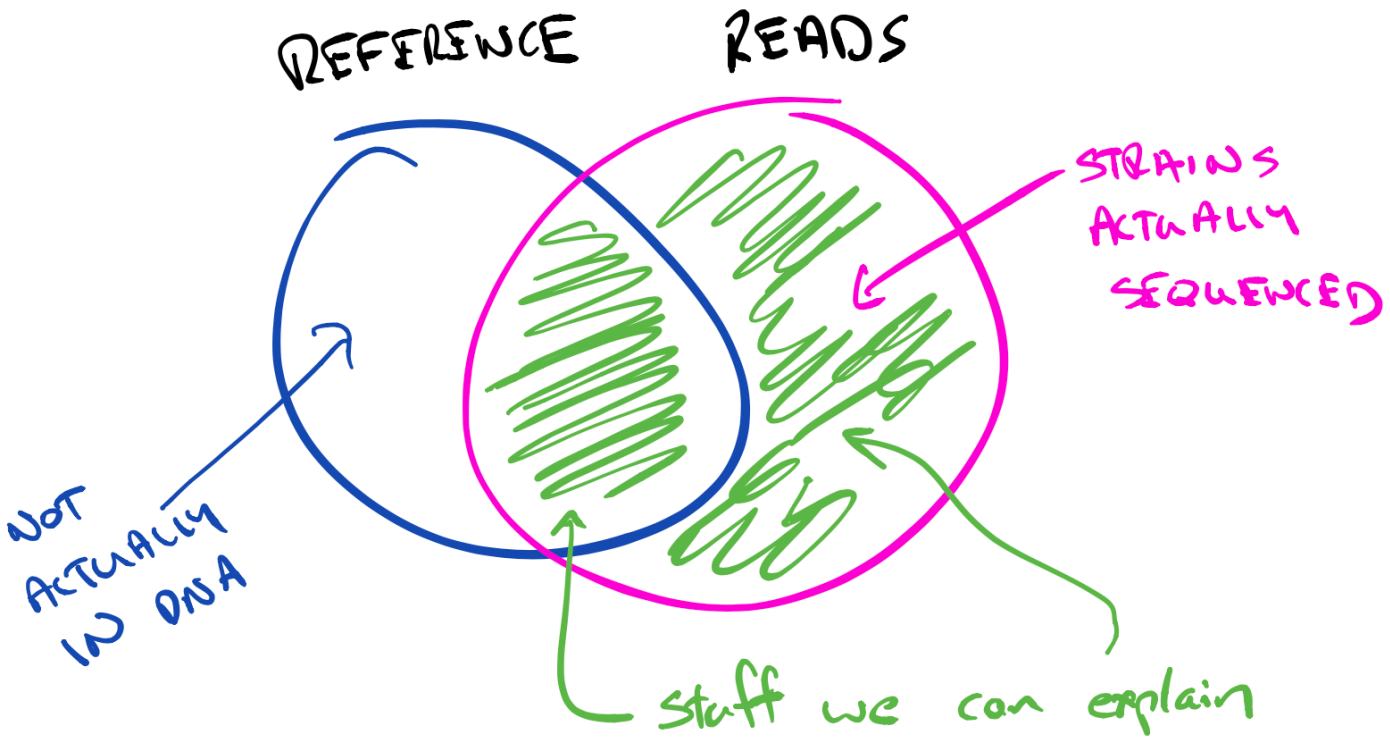
Table 2: Top uncovered genomes

Genome	Read cove
Desulfovibrio vulgaris DP4	93.2%
Thermus thermophilus HB27	91.1%
Enterococcus faecalis V583	74.6%
Fusobacterium nucleatum	47.6%

What happened? Our guess -



What we see:


REFERENCE
READS
STRAINS ACTUALLY SEQUENCED
NOT ACTUALLY IN DNA
stuff we can explain

Concluding thoughts

Other than strain variation, assembly worked really well!

- Recovered majority of genomes when confounding strains not present;
- Picked up “true” strain variants present in the population, we think;
- Assembled a significant part of an unknown Proteiniclasticum genome (contaminant in original data set);

Strain confusion is a major potential problem.

- If you are assembling data from a mixture of closely related strains, you are probably losing at least 20% of the “true” genomes;
- Right now, we have ***no good way*** to detect the presence of strain variation, or measure the extent of it, in shotgun metagenome sequencing;
- (I have some tools and ideas. :)

Some final points

Genome reconstruction from metagenomes may not be a biologically sensible goal: most true communities will contain a mixture of strains / pangenomes of organisms.

- (Is this an example of bioinformatics being misaligned with biology?)

We need to do a better job of characterizing our bioinformatics processes from end-to-end, and this must include generating good test data sets.

Fast “forensic” bioinformatics tools are important.

Bonus slide! What's a species??

Two microbes are the same species if they cannot be assembled out o
a mixture of short reads (to > 95% completeness).

This paper suggests that the metric for different species is < 1% Jaccard
similarity at k=31.

(We'll talk more about this at tomorrow's tutorial :)

Some final points

Genome reconstruction from metagenomes may not be a biologically sensible goal: most true communities will contain a mixture of strains / pangenomes of organisms.

- (Is this an example of bioinformatics being misaligned with biology?)

We need to do a better job of characterizing our bioinformatics processes from end-to-end, and this must include generating good test data sets.

Fast “forensic” bioinformatics tools are important.

Thanks for listening!

Please contact me at ctbrown@ucdavis.edu!

e: everything I talked about today is
nly available; ask if you can't find it.

USDA

NIH
National Institutes of Health

NSF

GORDON AND BETTY MOORE FOUNDATION