Novel Computational Approaches to Investigate Microbial Diversity

Qingpeng Zhang
Department of Computer Science and Engineering
Michigan State University
Advisor: Dr. Titus Brown

Outline

- Background and motivation
- An efficient k-mer counting approach
- Novel method to investigate microbial diversity
  - Concept of IGS (informative genomic segment)
  - Testing IGS method on simulated data sets
  - Testing IGS method on real metagenome data
- Summary

Microbial diversity: Sorcerer II Global Ocean Sampling Expedition

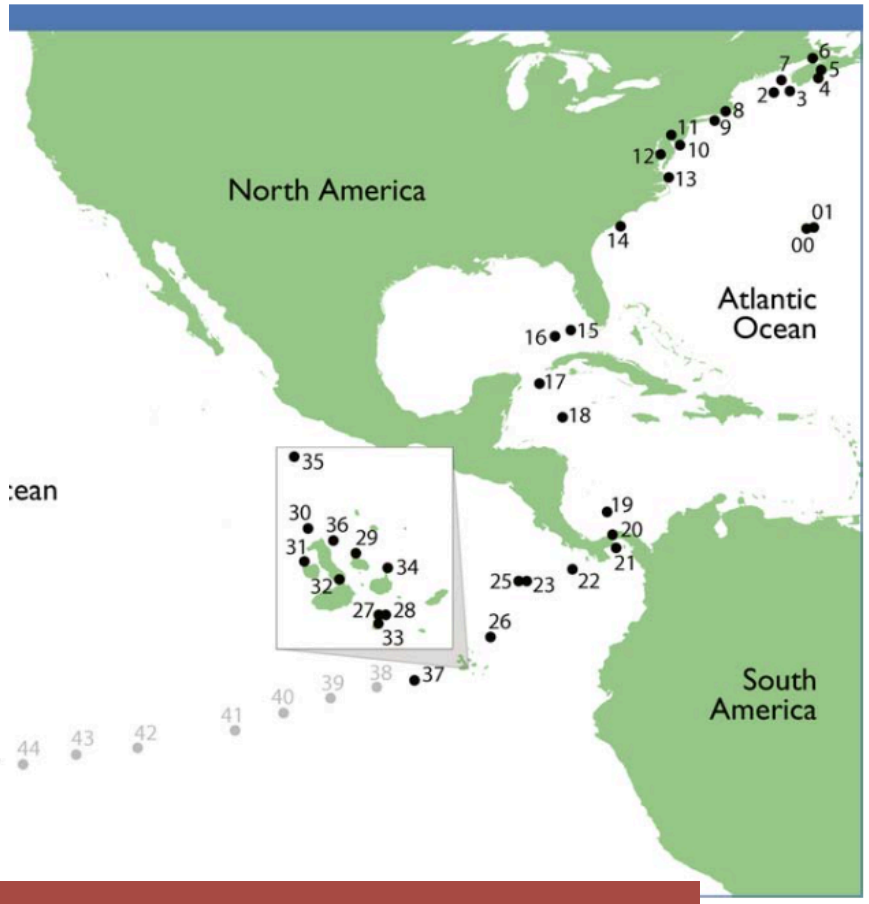
OPEN ACCESS Freely available online PLoS BIOLOGY

The *Sorcerer II* Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific

Douglas B. Rusch[1*], Aaron L. Halpern[1], Granger Sutton[1], Karla B. Heidelberg[1,2], Shannon Williamson[1], Shibu Yooseph[1], Dongying Wu[1,3], Jonathan A. Eisen[1,3], Jeff M. Hoffman[1], Karin Remington[1,4], Karen Beeson[1], Bao Tran[1], Hamilton Smith[1], Holly Baden-Tillson[1], Clare Stewart[1], Joyce Thorpe[1], Jason Freeman[1], Cynthia Andrews-Pfannkoch[1], Joseph E. Venter[1], Kelvin Li[1], Saul Kravitz[1], John F. Heidelberg[1,2], Terry Utterback[1], Yu-Hui Rogers[1], Luisa I. Falcón[5], Valeria Souza[5], Germán Bonilla-Rosso[5], Luis E. Eguiarte[5], David M. Karl[6], Shubha Sathyendranath[7], Trevor Platt[7], Eldredge Bermingham[8], Victor Gallardo[9], Giselle Tamayo-Castillo[10], Michael R. Ferrari[11], Robert L. Strausberg[1],

uthern California, Avalon, California, United
our World, Rockville, Maryland, United States
ty, Mexico, **6** Department of Oceanography,
a, Canada, **8** Smithsonian Tropical Research
10 Escuela de Química, Universidad de Costa
States of America, **12** Department of Earth

ost part, uncharacterized both lanktonic microbiota in which l Ocean Sampling expedition. hrough the Panama Canal and uencing reads (6.3 billion bp). t contains great diversity with 98% sequence identity cutoff. new comparative genomic and ent," addressed questions of iochemical diversity of genes ssembly and reconstruction of pulations analyzed, we found



- How many different species in a sea water sample? What is their abundance distribution? (alpha-diversity) [Hint: ~10,000 different types of microorganisms]
- Are samples from tropical area and temperate area different?(beta-diversity) [Hint: Yes, but how different?]

Metagenomics and "big data"

Most of the microorganisms can not be cultured and isolated and sequenced in way.

*"...functional analysis of the **collective genomes** of soil microflora, which we term the metagenome of the soil."*

- J. Handelsman et al., 1998

In a gram of soil, there are approximately a billion microbial cells, containing an estimated **4 petabase pairs** of DNA (Jack A. Gilbert,2013)

+

Improvement of next generation sequencing

+

Decreasing cost of sequencing

= **data deluge**

Metagenomics and "big data"

"...functional analysis of the collective genomes of soil microflora, which we term the metagenome of the soil."

most of t
cultured
traditiona

In a gram of soil, there are approximate
containing an estimated **4 petabase pai**

+

Improvement of next generation sequer

+

Decreasing cost of sequencing

Metagenomics and "big data"

"…functional analysis of the collective genomes

In a gran
containir
+
Improver
+
Decreasing cost of sequencing

http://www.e-janco.com/

Metagenomics and "big data"

sample	# of reads	size of .gz file	# of bps	ave. length
iowa corn	1,514,290,825	46G	144,202,427,079	95.2
iowa prairie	2,597,093,273	74G	226,815,059,143	87.3
kansas_corn	2,029,883,371	66G	206,933,829,048	101.9
kansas_prairie	4,987,358,734	145G	499,387,223,498	100.3
wisconsin_corn	1,616,440,116	51G	162,257,698,471	100.4
wisconsin_prairie	1,653,557,590	53G	166,467,901,724	100.7
wisconsin_restored	226,830,595	11G	34,241,520,930	151.0
wisconsin_switchgrass	310,966,735	13G	40,259,619,921	129.5

Bacterial genome size: 139,000 bps to 13,000,000 bps
Human genome size: 3,000,000,000 bps

Tools friendly to big data are required.

Sorcerer II Global Ocean Sampling Expedition

OPEN ACCESS Freely available online

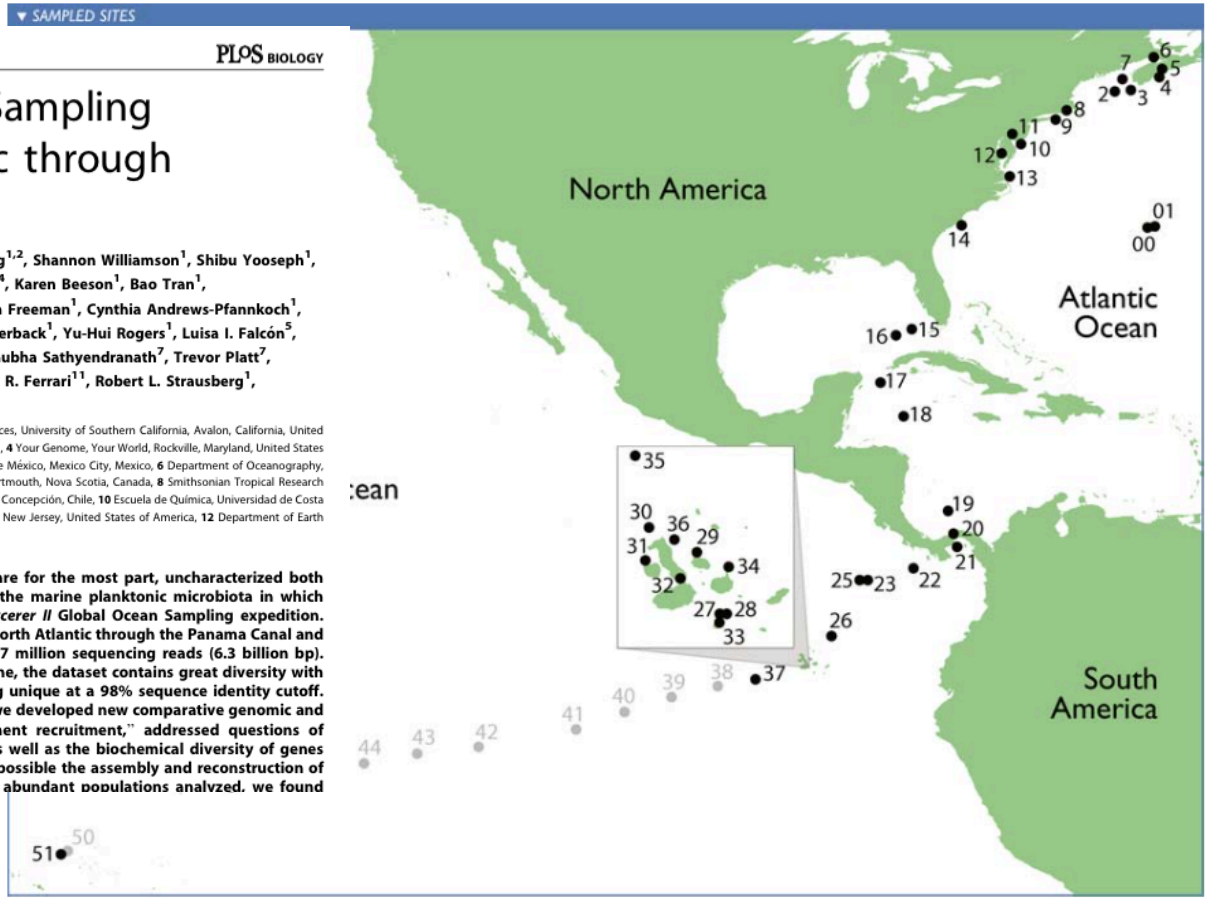
PLoS BIOLOGY

The *Sorcerer II* Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific

Douglas B. Rusch[1*], Aaron L. Halpern[1], Granger Sutton[1], Karla B. Heidelberg[1,2], Shannon Williamson[1], Shibu Yooseph[1], Dongying Wu[1,3], Jonathan A. Eisen[1,3], Jeff M. Hoffman[1], Karin Remington[1,4], Karen Beeson[1], Bao Tran[1], Hamilton Smith[1], Holly Baden-Tillson[1], Clare Stewart[1], Joyce Thorpe[1], Jason Freeman[1], Cynthia Andrews-Pfannkoch[1], Joseph E. Venter[1], Kelvin Li[1], Saul Kravitz[1], John F. Heidelberg[1,2], Terry Utterback[1], Yu-Hui Rogers[1], Luisa I. Falcón[5], Valeria Souza[5], Germán Bonilla-Rosso[5], Luis E. Eguiarte[5], David M. Karl[6], Shubha Sathyendranath[7], Trevor Platt[7], Eldredge Bermingham[8], Victor Gallardo[9], Giselle Tamayo-Castillo[10], Michael R. Ferrari[11], Robert L. Strausberg[1], Kenneth Nealson[1,12], Robert Friedman[1], Marvin Frazier[1], J. Craig Venter[1]

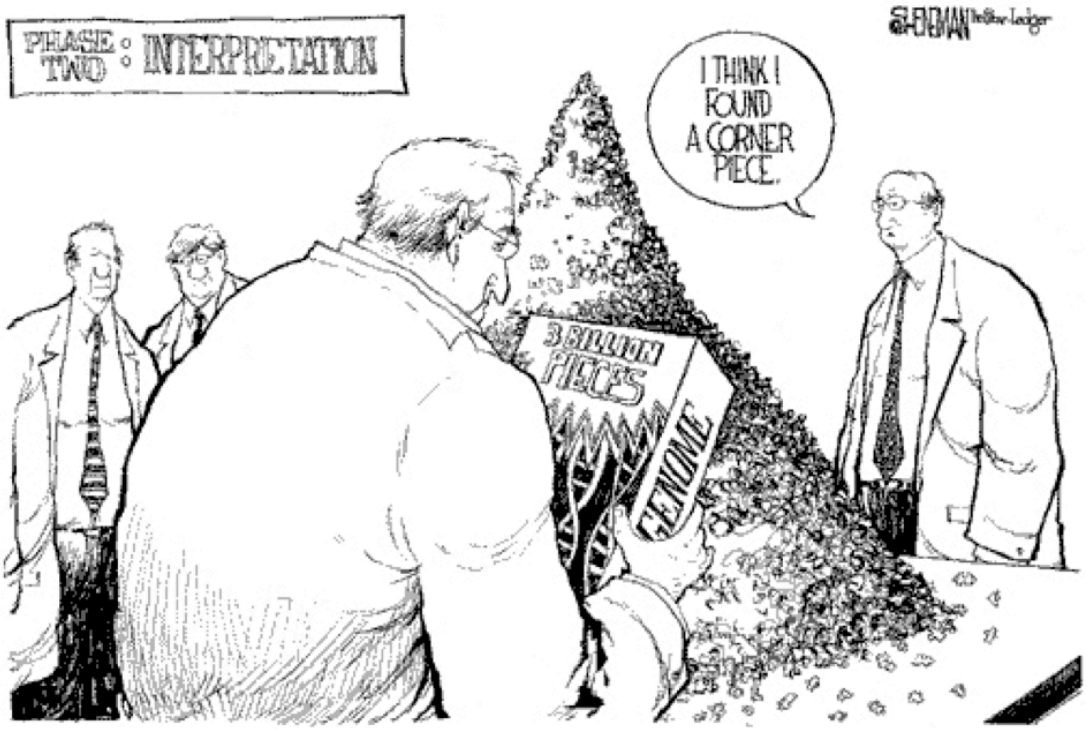
1 J. Craig Venter Institute, Rockville, Maryland, United States of America, **2** Department of Biological Sciences, University of Southern California, Avalon, California, United States of America, **3** Genome Center, University of California Davis, Davis, California, United States of America, **4** Your Genome, Your World, Rockville, Maryland, United States of America, **5** Departmento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, Mexico City, Mexico, **6** Department of Oceanography, University of Hawaii, Honolulu, Hawaii, United States of America, **7** Bedford Institute of Oceanography, Dartmouth, Nova Scotia, Canada, **8** Smithsonian Tropical Research Institute, Balboa, Ancon, Republic of Panama, **9** Departamento de Oceanografía, Universidad de Concepción, Concepción, Chile, **10** Escuela de Química, Universidad de Costa Rica, San Pedro, Costa Rica, **11** Department of Environmental Sciences, Rutgers University, New Brunswick, New Jersey, United States of America, **12** Department of Earth Sciences, University of Southern California, Los Angles, California, United States of America

The world's oceans contain a complex mixture of micro-organisms that are for the most part, uncharacterized both genetically and biochemically. We report here a metagenomic study of the marine planktonic microbiota in which surface (mostly marine) water samples were analyzed as part of the *Sorcerer II* Global Ocean Sampling expedition. These samples, collected across a several-thousand km transect from the North Atlantic through the Panama Canal and ending in the South Pacific yielded an extensive dataset consisting of 7.7 million sequencing reads (6.3 billion bp). Though a few major microbial clades dominate the planktonic marine niche, the dataset contains great diversity with 85% of the assembled sequence and 57% of the unassembled data being unique at a 98% sequence identity cutoff. Using the metadata associated with each sample and sequencing library, we developed new comparative genomic and assembly methods. One comparative genomic method, termed "fragment recruitment," addressed questions of genome structure, evolution, and taxonomic or phylogenetic diversity, as well as the biochemical diversity of genes and gene families. A second method, termed "extreme assembly," made possible the assembly and reconstruction of large segments of abundant but clearly nonclonal organisms. Within all abundant populations analyzed, we found

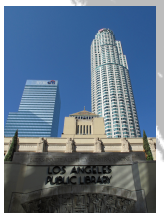


- How many different species in a sea water sample? What is their abundance distribution? (alpha-diversity) [Hint: ~10,000 different types of microorganisms, 10 million viruses, one million bacteria in a drop.]
- Are samples from tropical area and temperate area different?(beta-diversity) [Hint: Yes, but how different?]

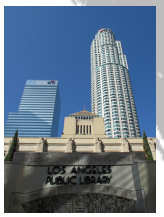
Genome and Book



Sample and Library

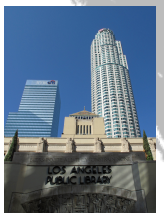


Which library has more titles?



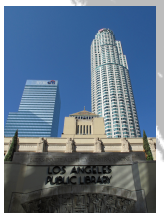
Which library has more titles?

With only shredded pieces of pages
with (partial) sentences on it to read.



How similar are the libraries by their collection?

With only shredded pieces of pages
with (partial) sentences on it to read.



A typical pipeline of metagenome diversity analysis

Sample1

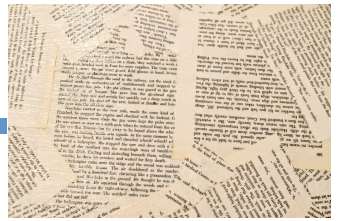
Sequencing

Reads

Assembly

Contigs

1000+ bp



Binning

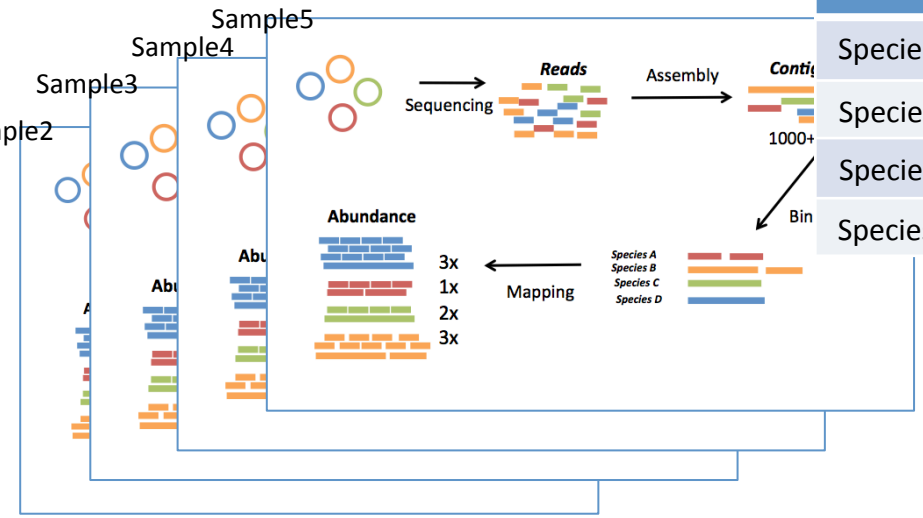
Species

Species A
Species B
Species C
Species D

Mapping

Abundance

3x
1x
1x
2x



	Sample1	Sample2	Sample3	Sample4	Sample5
Species A	3	3	4	4	0
Species B	1	3	0	2	2
Species C	1	2	2	2	3
Species D	2	1	2	1	1

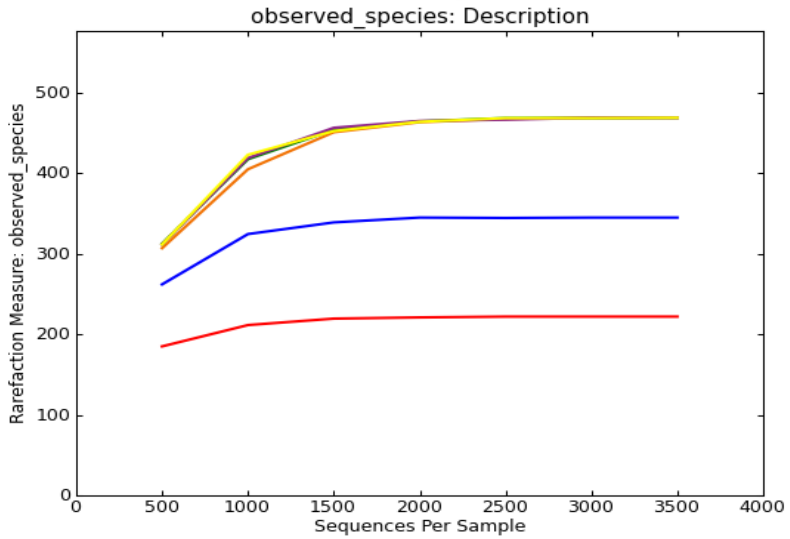
Loosely based on slide by Mads Albertsen

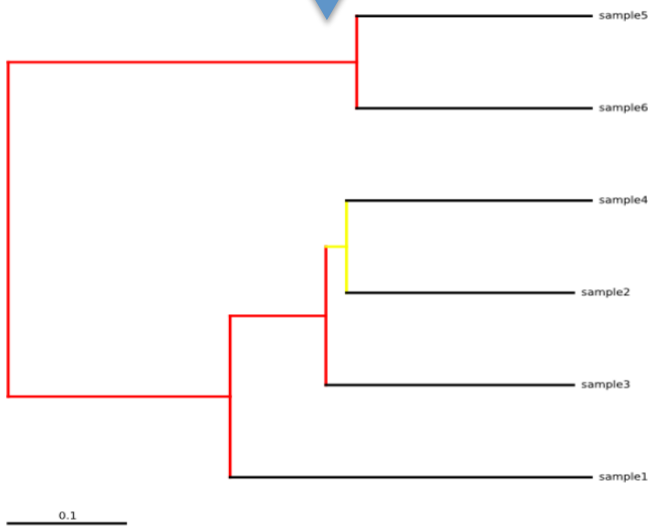
	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6
SpeciesA	3	3	4	4	0	0
SpeciesB	1				2	2
SpeciesC	1				3	3
SpeciesD	2	1	2	1	1	1
SpeciesE	4	1	3	0	2	5
SpeciesF	5	1	2	2	2	5
SpeciesG	2	2	1	2	1	3

We need this!

	Sample1	Sample1	Sample1	Sample1	Sample1	Sample1
Sample1	0.0	0.25	0.5	0.5	0.75	1.0
Sample2	0.25	0.0	0.25	0.25	0.75	1.0
Sample3	0.5	0.25	0.0	0.25	0.75	1.0
Sample4	0.5	0.25	0.25	0.0	0.75	1.0
Sample5	0.75	0.75	0.75	0.75	0.0	0.25
Sample6	1.0	1.0	1.0	1.0	0.25	0.0

Tools: QIIME, mothur, etc.



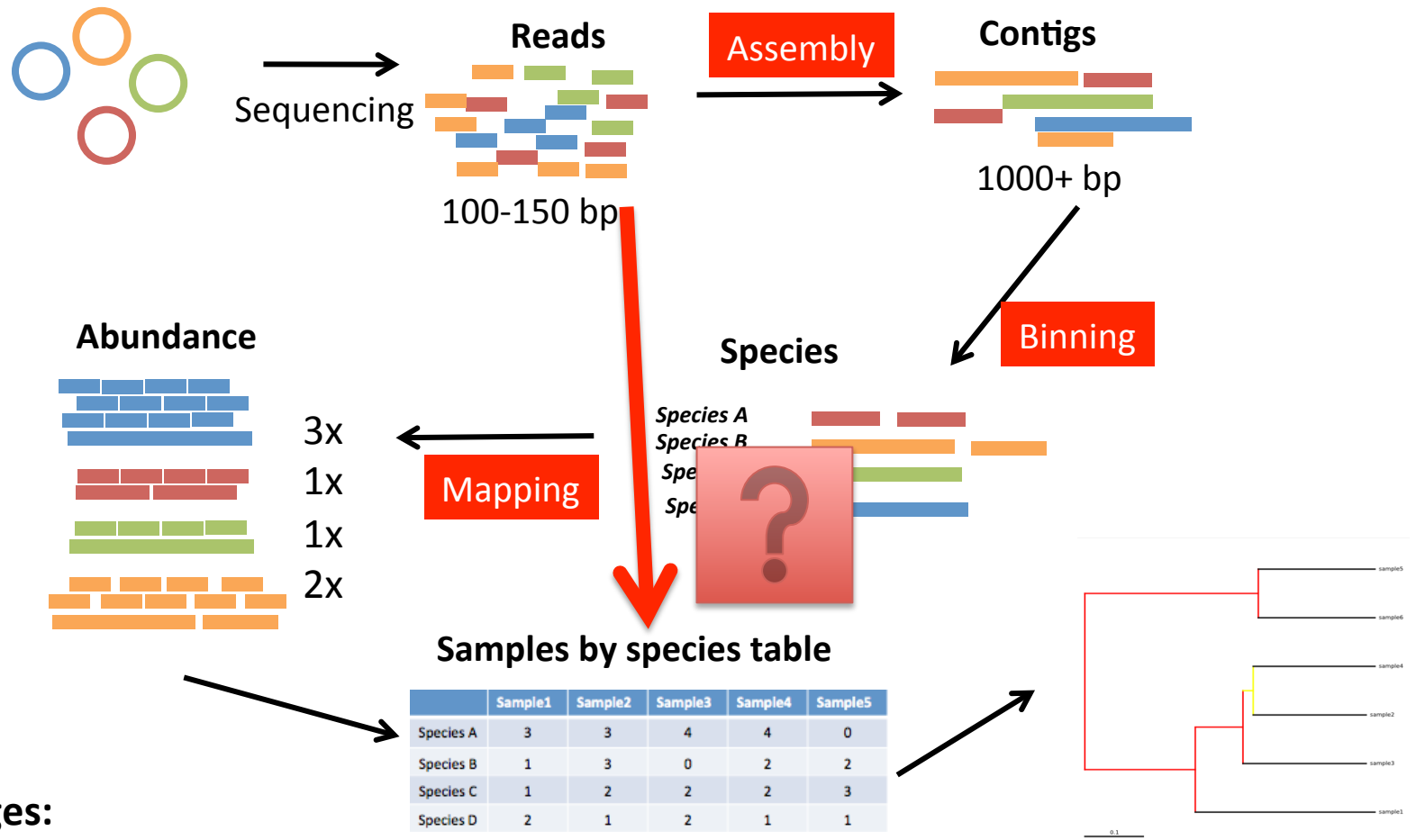


Alpha diversity:
"How many different species in a seawater sample?"

Beta diversity:
"How different are the seawater samples?"

A typical pipeline of metagenome diversity analysis

Sample1

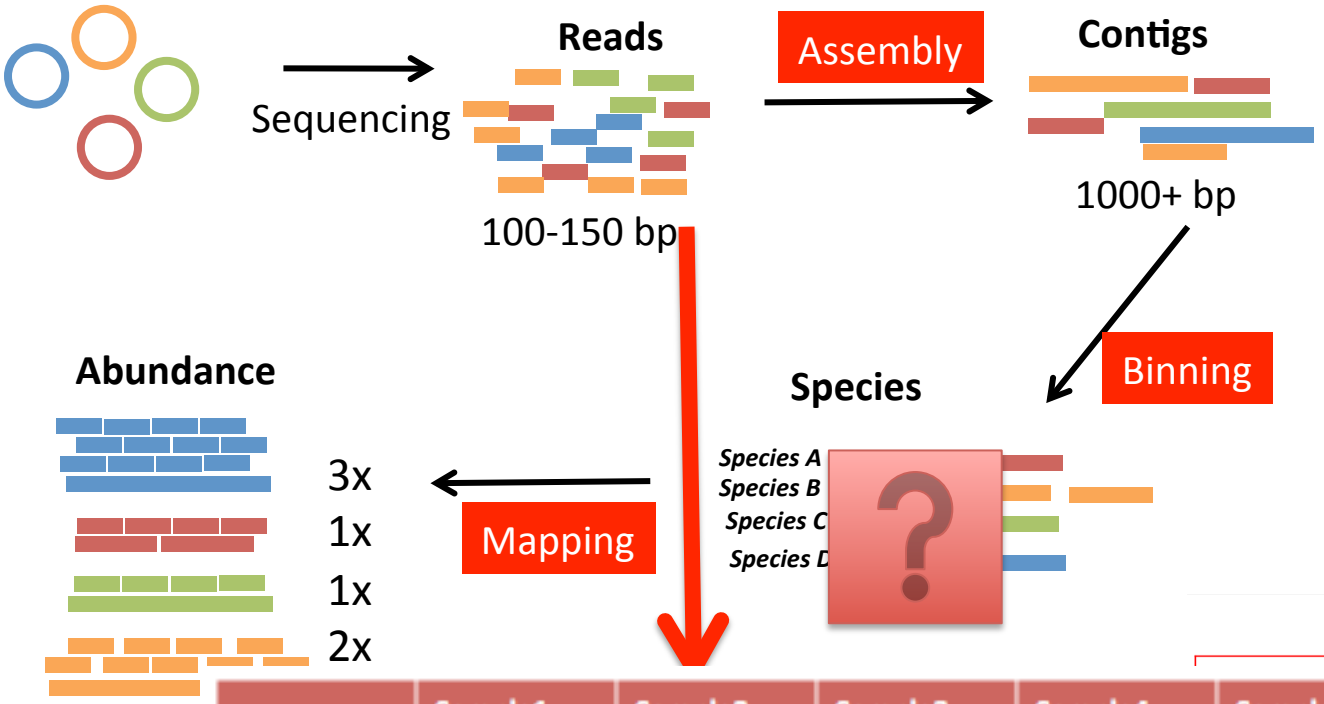


Challenges:

- **Assembly?** - difficult (eg. Only small portions of reads can be assembled)
- **Binning?** – difficult (eg. may need reference genomes, computationally expensive)
- **Mapping?** – difficult (eg. accuracy, computationally expensive)
- **4 petabase** pairs of DNA in a gram of soil(Jack A. Gilbert,2013)

A typical pipeline of metagenome diversity analysis

Sample1



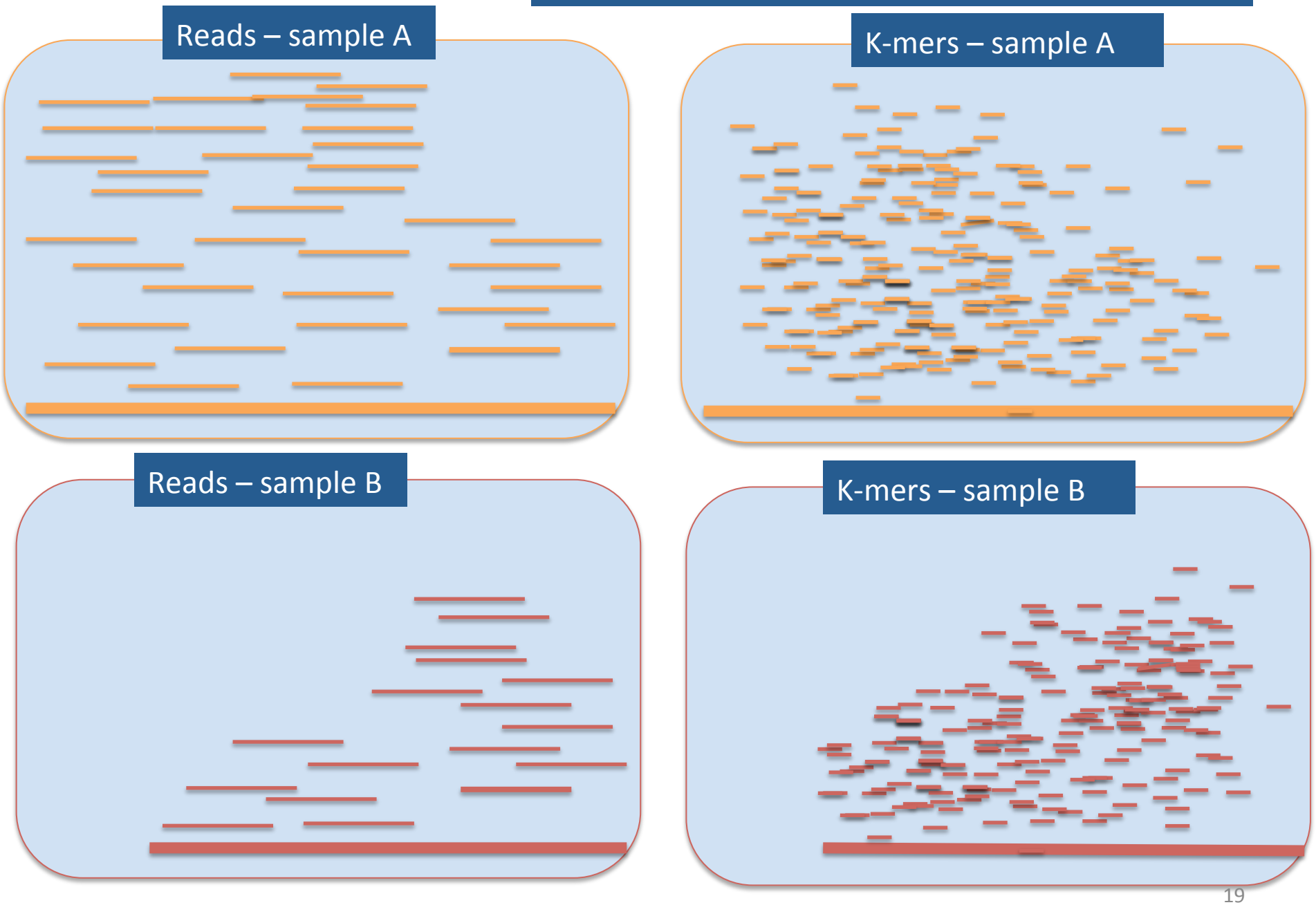
	Sample1	Sample2	Sample3	Sample4	Sample5
Unit_A	3	3	4	4	0
Unit_B	1	3	0	2	2
Unit_C	1	2	2	2	3
Unit_D	2	1	2	1	1

Reads -> k-mers

>24288227
TTTGCCTGTTGAAAACCTGTAAACGGGATGTTTCCGGGGTTTTCCGCCCCGGTAGGGTTGAGA
CGTGCCCCGCGGTCGCGAACAGCTCGCCGCTTACCGGCGTAAGAGCGATCCGCAACACACCC
CGGAGCCTCTACCCCCCCCTCACGC

K-mer : DNA segment with length of k.

Counting (unique) k-mers to evaluate diversity



Counting (unique) words to evaluate diversity

Reads – sample A

K-mers – sample A

Reads – sample B

K-mers – sample B

Challenges in counting k-mers

- handle large number of k-mers. (simple hash table will not work)

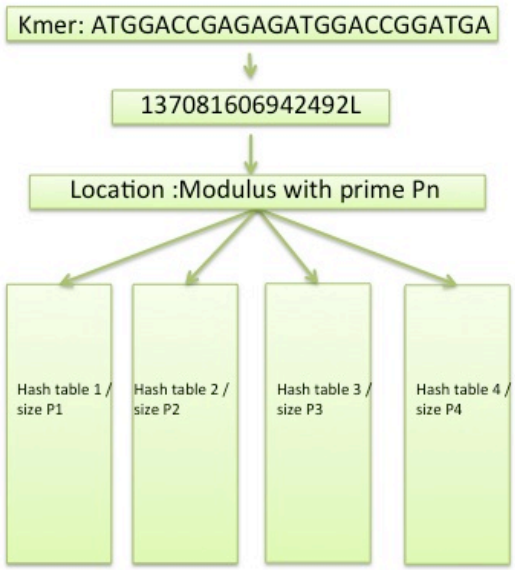
Soil metagenome sample	Number of reads	Number of k-mers k=29	Number of unique k-mers
IOHH.2301.7.1859	273,021,397	32,762,567,744	22,333,571,795

- retrieve count of k-mer randomly (in-memory).
  - To check the occurrence of a k-mer in multiple samples to find shared k-mers

Outline

- Background and motivation
- An efficient k-mer counting approach
- Novel method to investigate microbial diversity
  - Concept of IGS (informative genomic segment)
  - Testing IGS method on simulated data sets
  - Testing IGS method on real metagenome data
- Summary

Khmer - An efficient k-mer counting approach

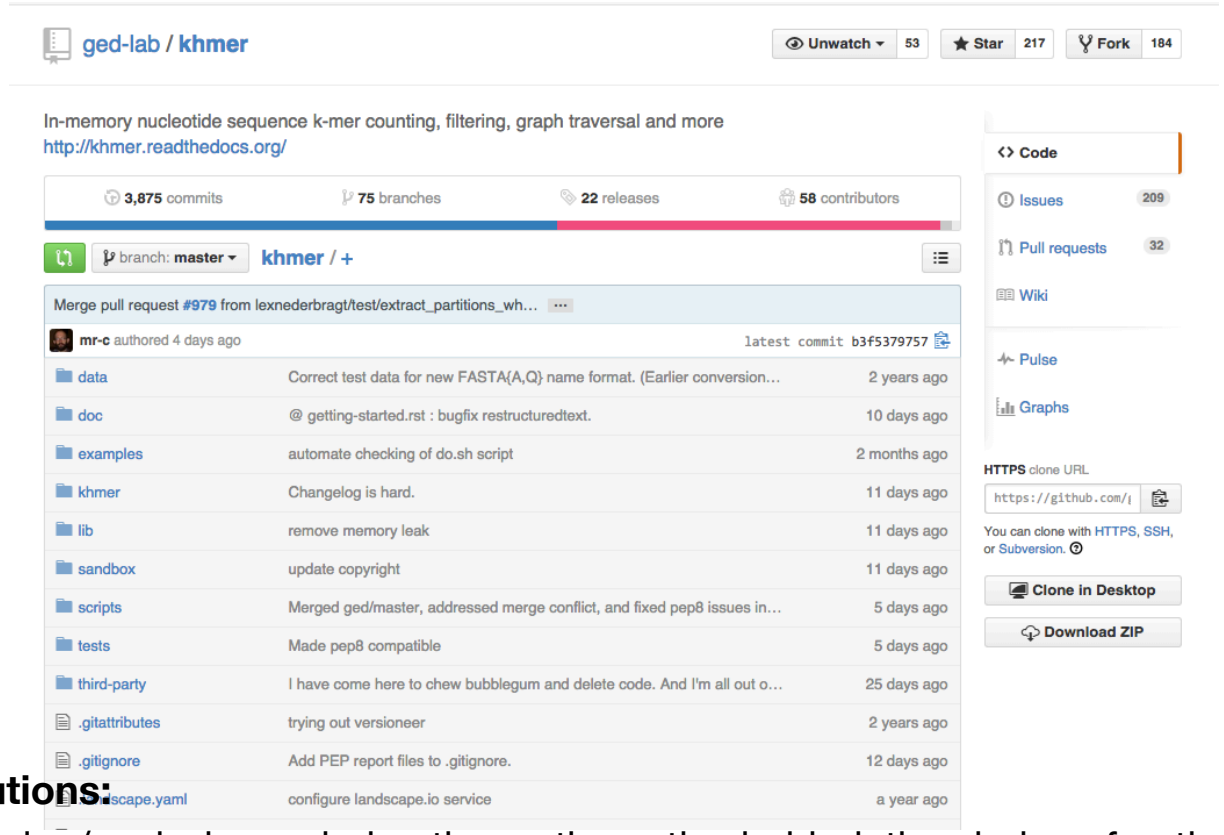


Update entries in the location in each hash table

- The **Count-min Sketch** consists of one or more hash tables of different size
- Each entry in the hash tables is a counter representing the number of k-mers that hash to that location
- To retrieve the count for a given k-mer we select the minimum count across all of the hash entries.

- **Highly scalable**: Constant memory consuming, independent of k and dataset size
- Probabilistic properties well suited to next generation sequencing datasets
  - Memory usage is fixed and low, with certain counting false positive rate as tradeoff because of collision
  - The one-sided counting error is low and predictable.

An efficient k-mer counting approach



- **My contributions:**
- algorithm design/analysis, exploring the mathematics behind, the choice of optimal parameters
- contributing codes, including unique k-mers counting, overlap k-mer counting, optimal parameter choice, others related to my specific research project.
- benchmarking, testing
- exploration of applications like error trimming, filter low abundance reads, digital normalization, etc. suggestion on features
- work on the khmer manuscript

Sequencing errors cause serious problem

>24288227
TTTGCCTGTTGAAAACCTGAAAACGGGATGTTTCC
GGGGTTTTCCGCCCCGGTAGGGTTGAGACGTGCCC
CGCGGTCGCGAACAGCTCGCCGCTTACCGGCGTAA
GAGCGATCCGCAACACACCCCGGAGCCTCTACCCC
CCCCTCACGC

In high coverage reads data set, most of the unique k-mers are erroneous caused by sequencing errors.

GAAAACCTGA
AAAACCTGAA
AAACCTGAAA

Counting k-mer to measure diversity directly?

Sequencing errors cause serious problem

>24288227
TTTGCCTGTTGAAAACCTGAAAACGGGATGTTTCC
GGGGTTTTCCGCCCCGGTAGGGTTGAGACGTGCCC
CGCGGTCGCGAACAGCTCGCCGCTTACCGGCGTAA
GAGCGATCCGCAACACACCCCGGAGCCTCTACCCC
CCCCTCACGC

In high coverage reads data set, most of the unique k-mers are erroneous caused by sequencing errors.

GAAAACCTGA
AAAACCTGAA
AAACCTGAAA

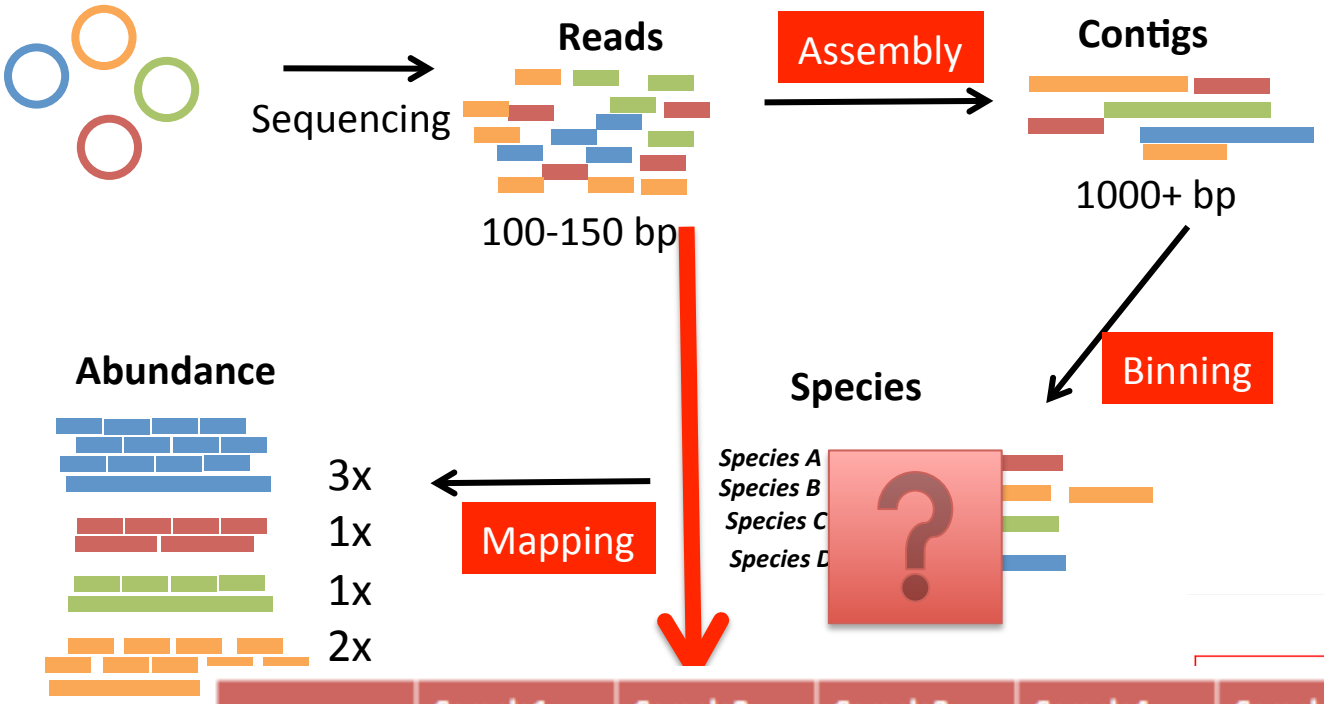
Countin... measure div...tly?

Outline

- Background and motivation
- An efficient k-mer counting approach
- Novel method to investigate microbial diversity
  - Concept of IGS (informative genomic segment)
  - Testing IGS method on simulated data sets
  - Testing IGS method on real metagenome data
- Summary

A typical pipeline of metagenome diversity analysis

Sample1



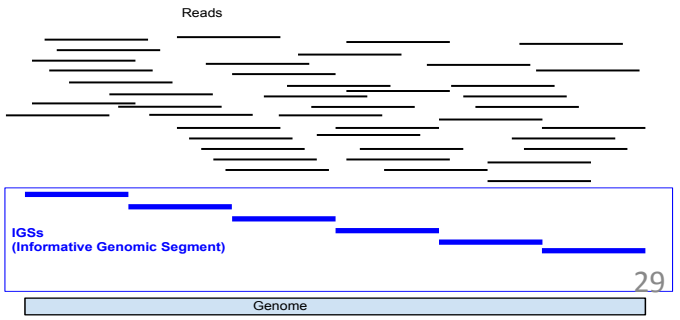
	Sample1	Sample2	Sample3	Sample4	Sample5
Unit_A	3	3	4	4	0
Unit_B	1	3	0	2	2
Unit_C	1	2	2	2	3
Unit_D	2	1	2	1	1

	Sample1	Sample2	Sample3	Sample4	Sample5
Species A	3	3	4	4	0
Species B	1	3	0	2	2
Species C	1	2	2	2	3
Species D	2	1	2	1	1

Replace samples-by-species table with samples-by-IGSs table

	Sample1	Sample2	Sample3	Sample4	Sample5
IGS1	3	3	4	4	0
IGS2	1	3	0	2	2
IGS3	1	2	2	2	3
IGS4	2	1	2	1	1

- IGS (informative genomic segment)
  - can represent the novel information of a genome
  - can be used as a new unit to do diversity analysis to replace species
  - Assembly-free
  - Reference-free
  - Efficient and scalable



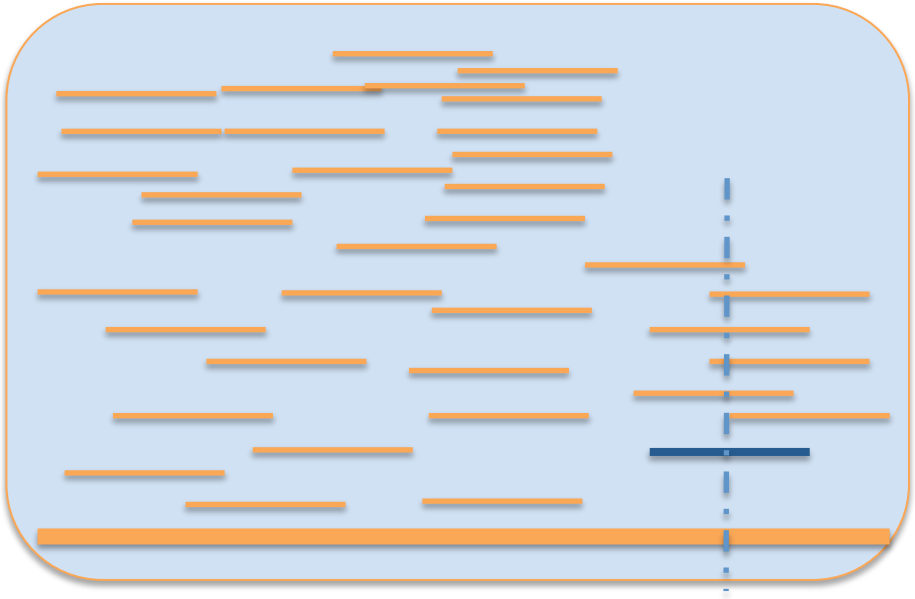
From reads to IGSs

1. We can use median k-mer abundance to estimate the sequencing coverage of a read without a reference assembly.
2. We can retrieve the coverage of a read across different samples.
3. If a read has a coverage as C, there will be approximately other C-1 reads with the same coverage covering the same DNA region.
4. We can estimate the size of covered genomic region by the number of reads and the corresponding coverage.

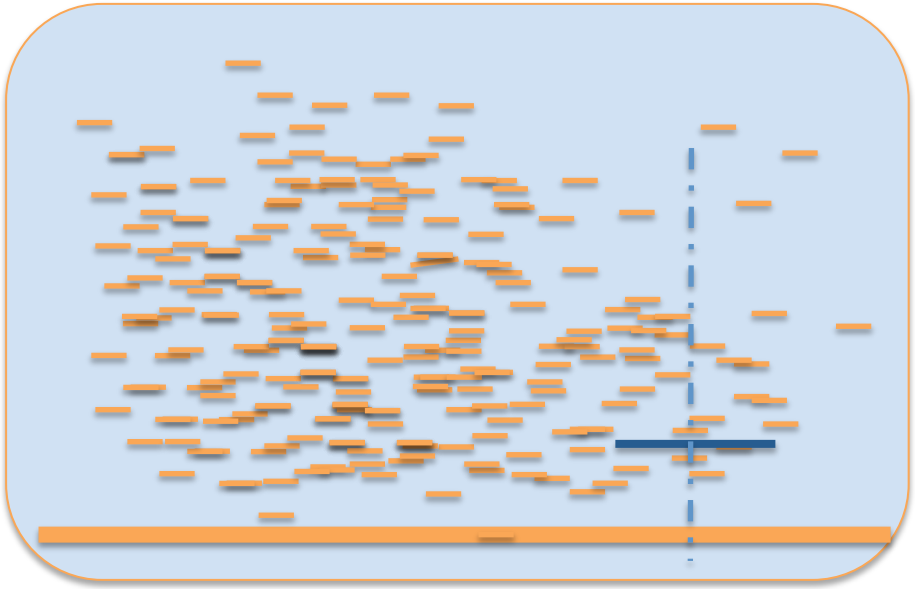
	Sample1	Sample2	Sample3	Sample4	Sample5
IGS1	3	3	4	4	0
IGS2	1	3	0	2	2
IGS3	1	2	2	2	3
IGS4	2	1	2	1	1

Four tricks!

Reads – sample A



K-mers – sample A

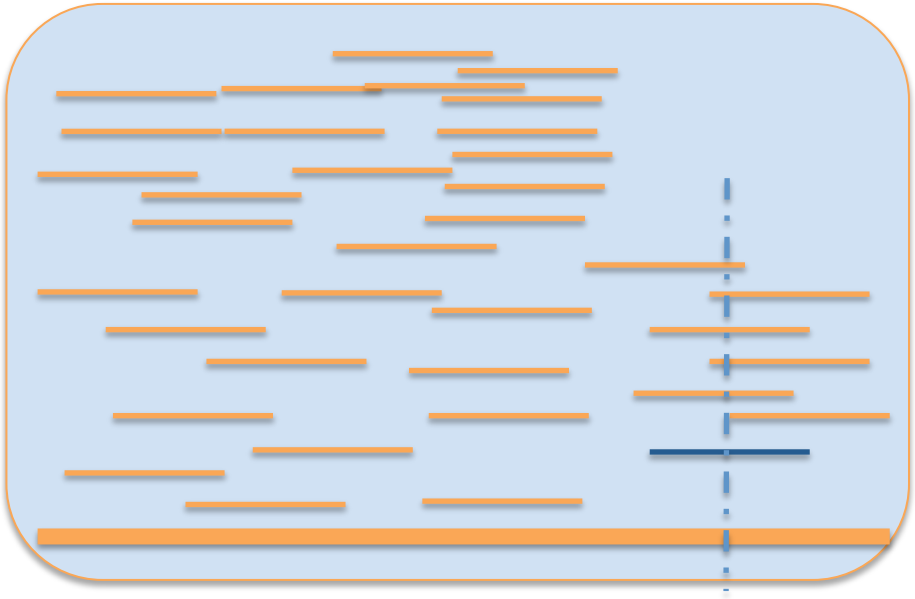


Coverage of read: the sequencing depth of the DNA region that read originates from

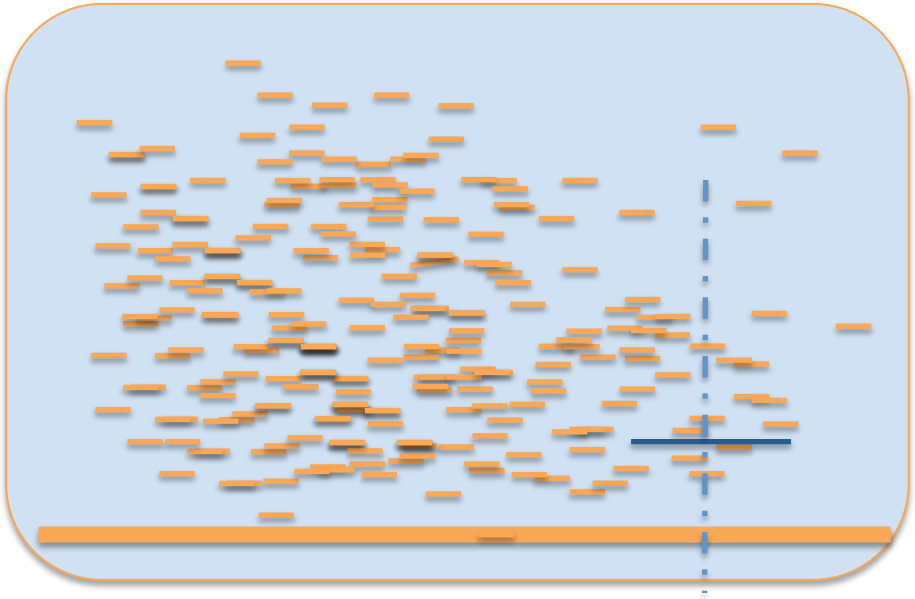
Coverage of a read and how to get it without a reference assembly

The abundance of the k-mers in a read tends to be similar, which is close to the coverage of the read.

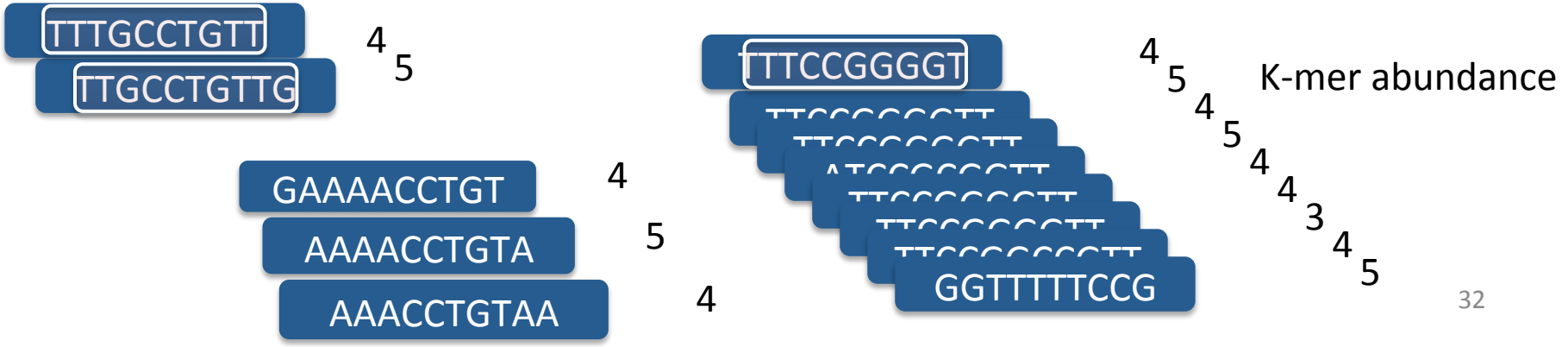
Reads – sample A



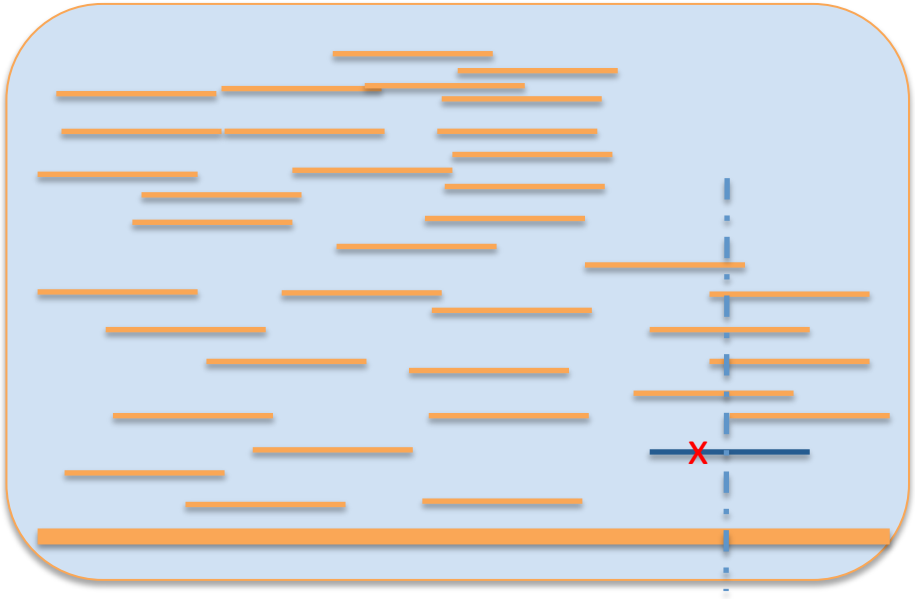
K-mers – sample A



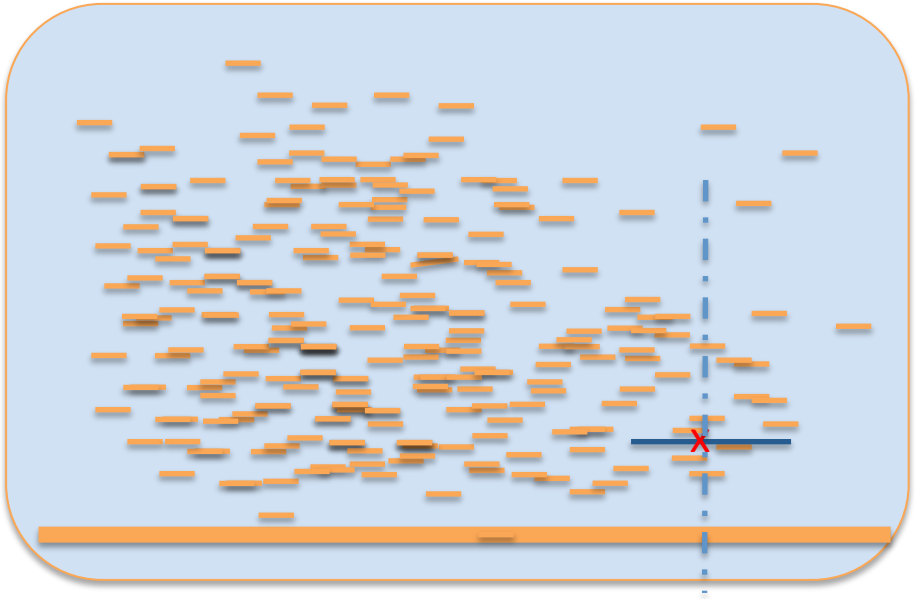
>24288227
TTTGCCTGTTGAAAACCTGTAAACGGGATGTTCCGGGGTTTTCCGCCCCGGTAGGGTTGAGA
CGTGCCCCGCGGTCGCGAACAGCTCGCCGCTTACCGGCGTAAGAGCGATCCGCAACACACCC
CGGAGCCTCTACCCCCCCCTCACGC



Reads – sample A



K-mers – sample A



>24288227
TTTGCCTGTTGAAAACCTGAAAACGGGATGTTTCCGGGGTTTTCCGCCCCGGTAGGGTTGAGA
CGTGCCCCGCGGTCGCGAACAGCTCGCCGCTTACCGGCGTAAGAGCGATCCGCAACACACCC
CGGAGCCTCTACCCCCCCCTCACGC

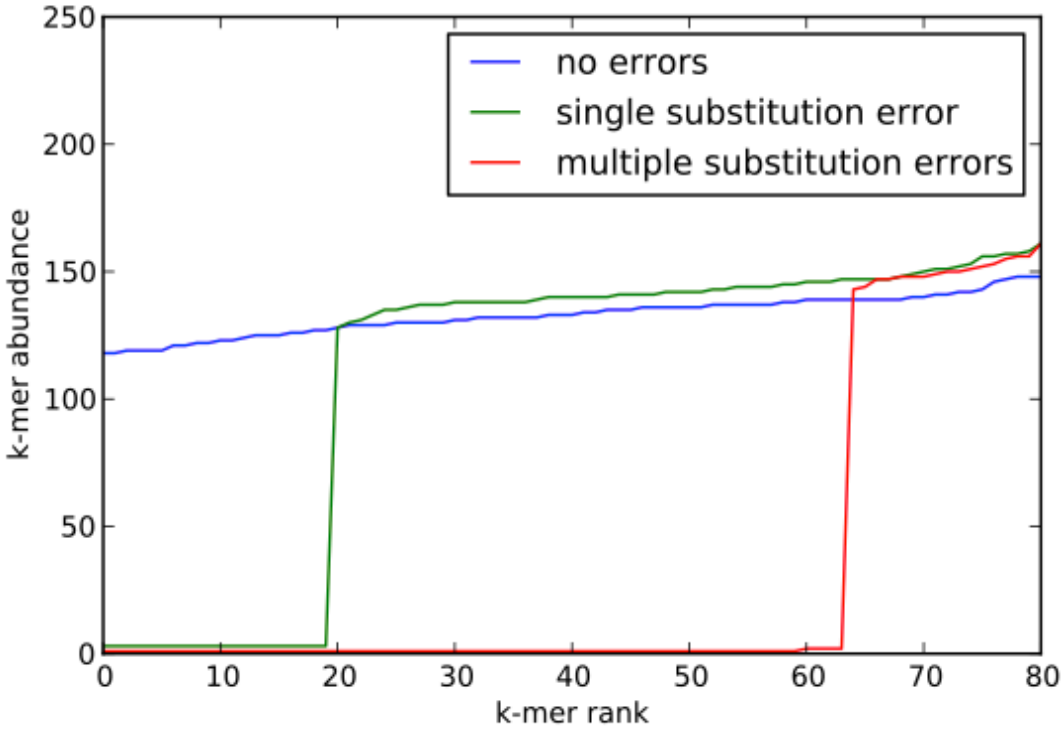
TTTGCCTGTT 4
TTGCCTGTTG 5

GAAAACCTGA 1
AAAACCTGAA 1
AAACCTGAAA 1

TTTCCGGGGT 4
5
4
5
4
4
3
4
5
GGTTTTTCCG

K-mer abundance

median k-mer abundance to represent the sequencing coverage of the read



>24288227
TTTGCCTGTT
CGTGCCCCG
CGGAGCCTC

GGGTTGAGA
AACACACCC

TTTGCCTGT

TTGCCTGTTG

GAAAACCTGA

AAAACCTGAA

AAACCTGAAA

1

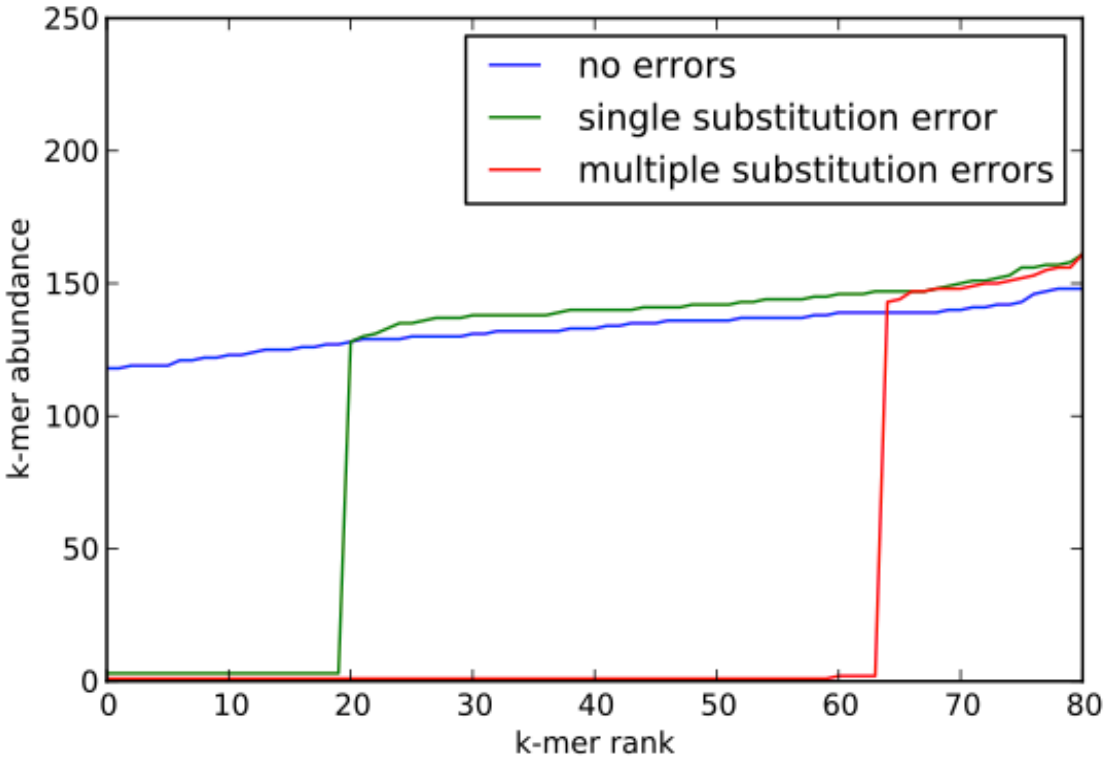
1

1

GGTTTTTCCG

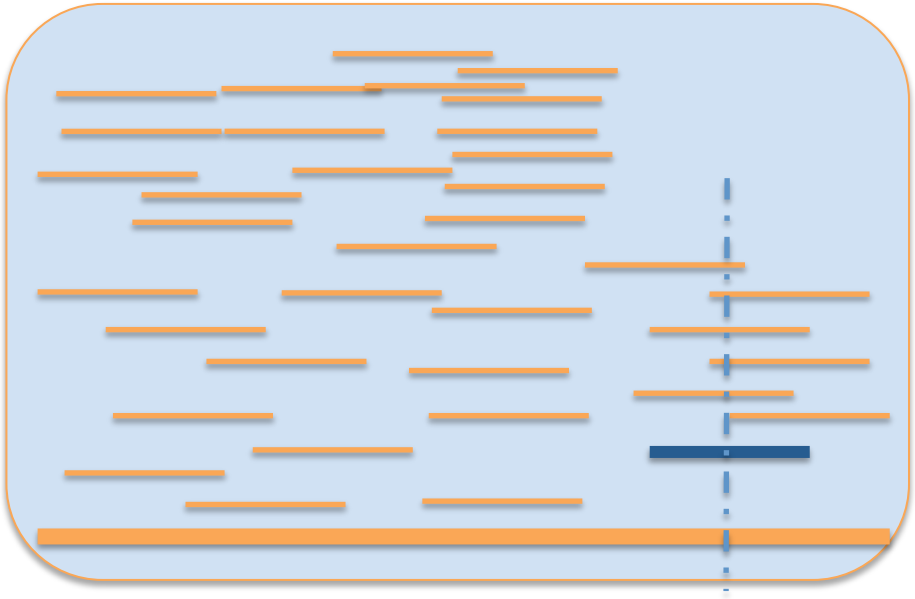
4 5 4 4 3 4 5

median k-mer abundance to represent the sequencing coverage of the read

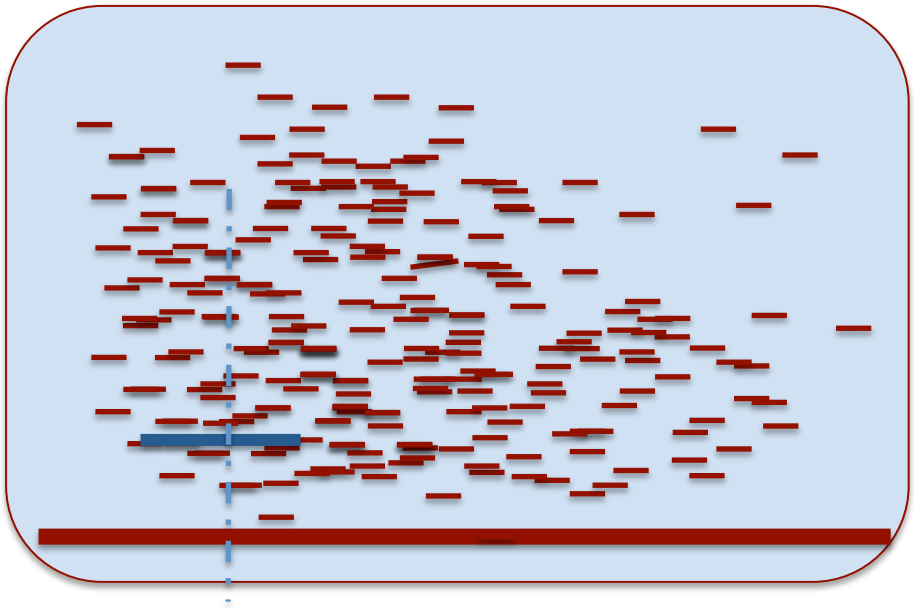


Trick 1: *We can use median k-mer abundance to estimate the sequencing coverage of a read without a reference assembly.*

Reads – sample A

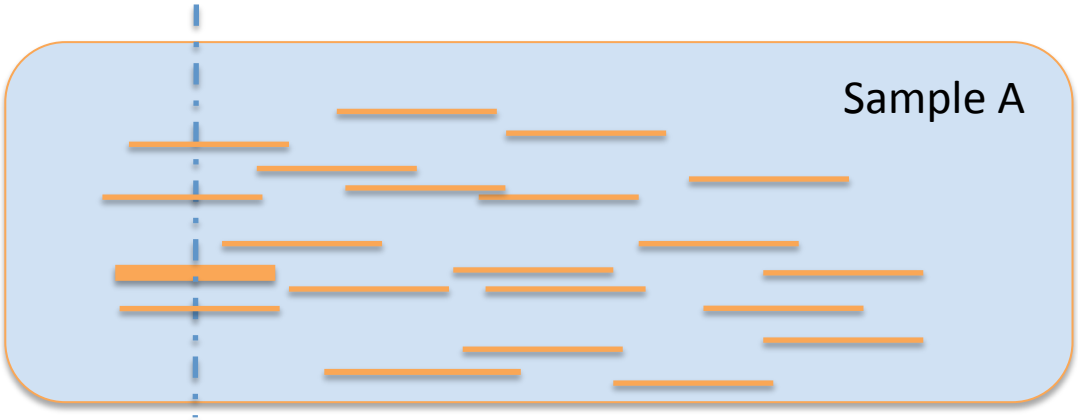


K-mers – sample B

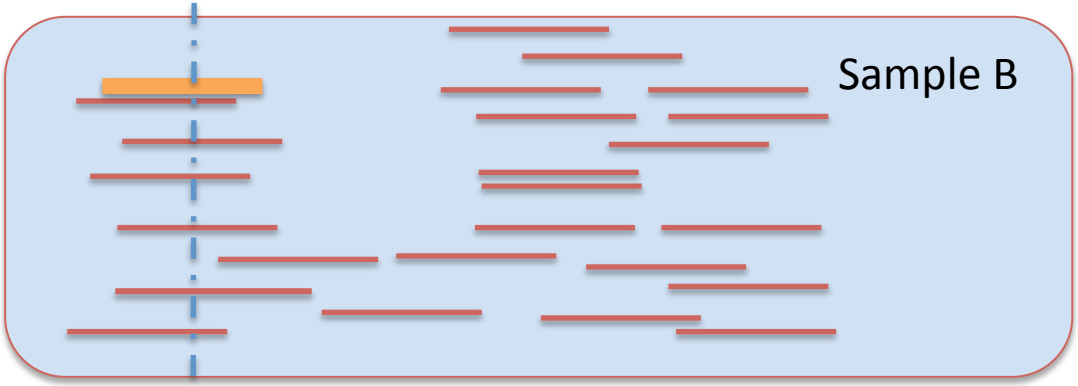


Trick 1: *We can use median k-mer abundance to estimate the sequencing coverage of a read without a reference assembly.*

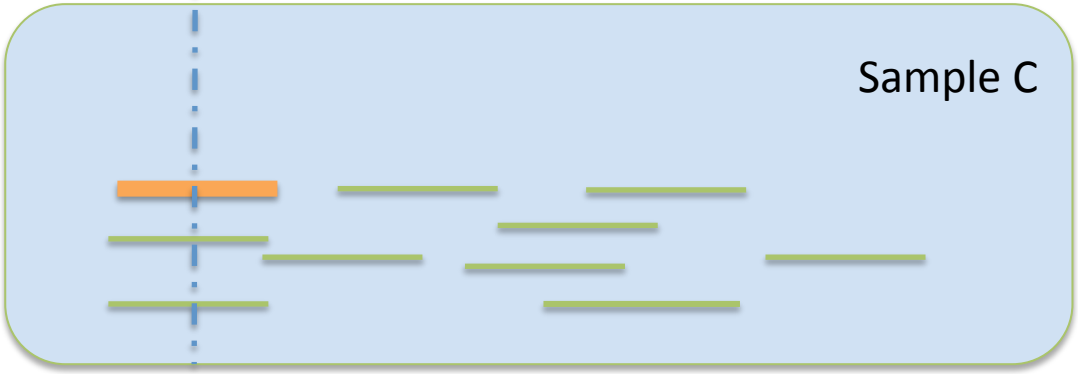
Trick 2: *We can estimate the sequencing coverage of a read across different samples*



Coverage = 4



Coverage = 6



Coverage = 2

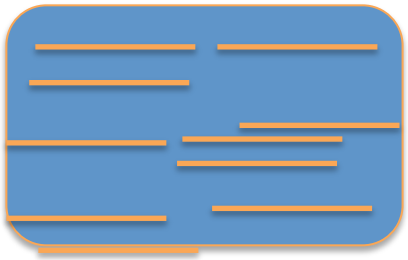
Read Coverage Profile:

A:B:C

4:6:2

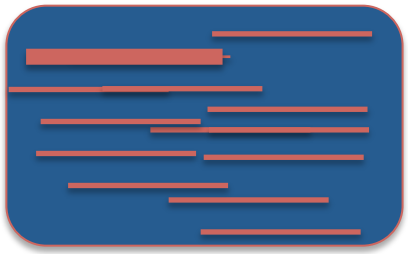
3:2:2
2:4:3
3:2:2
2:0:3
3:2:2
3:2:2
3:2:2
2:0:3
3:2:2
2:0:3
2:0:3
3:2:2
2:4:3

Sample A



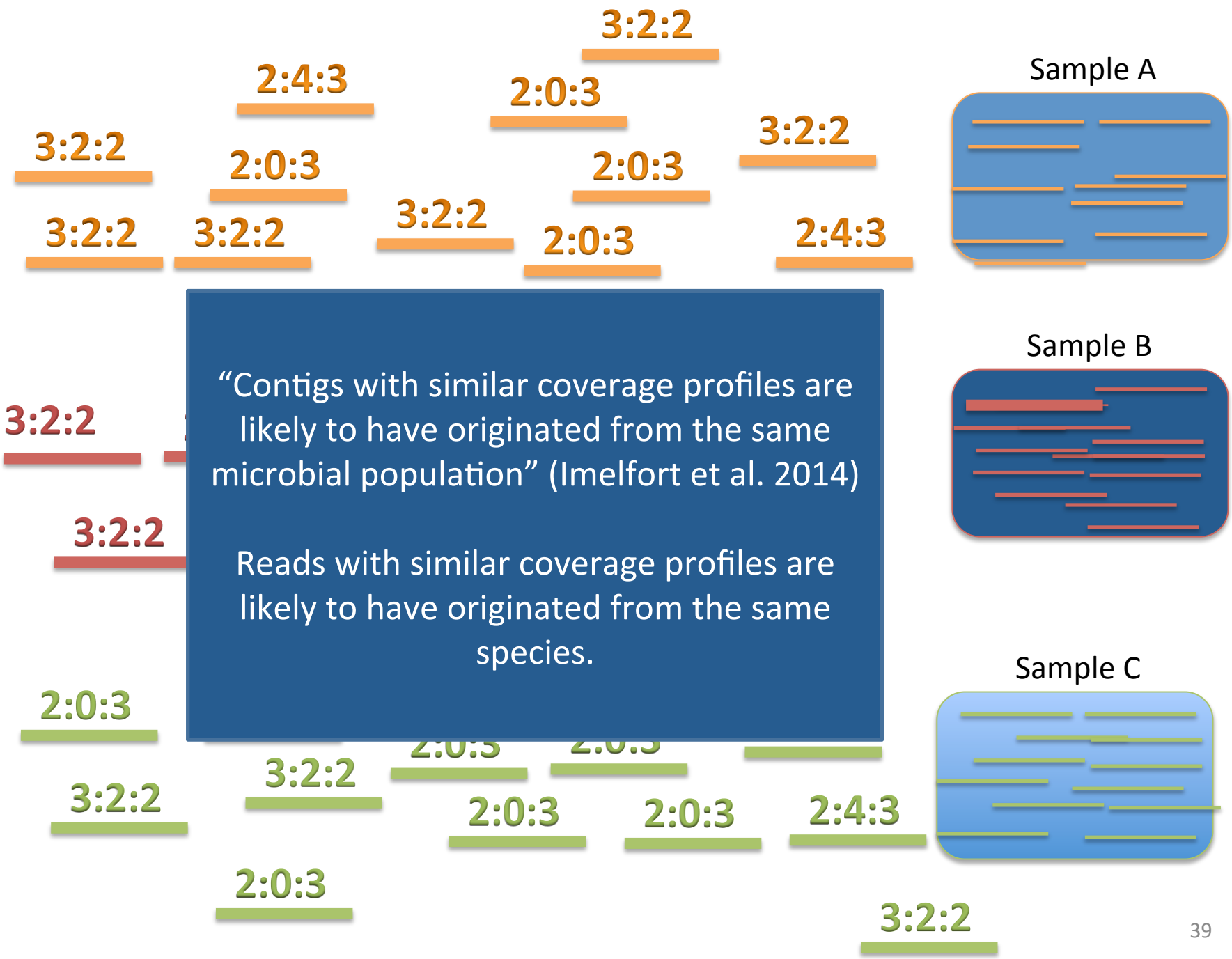
3:2:2
2:4:3
3:2:2
2:4:3
2:4:3
3:2:2
2:4:3
3:2:2

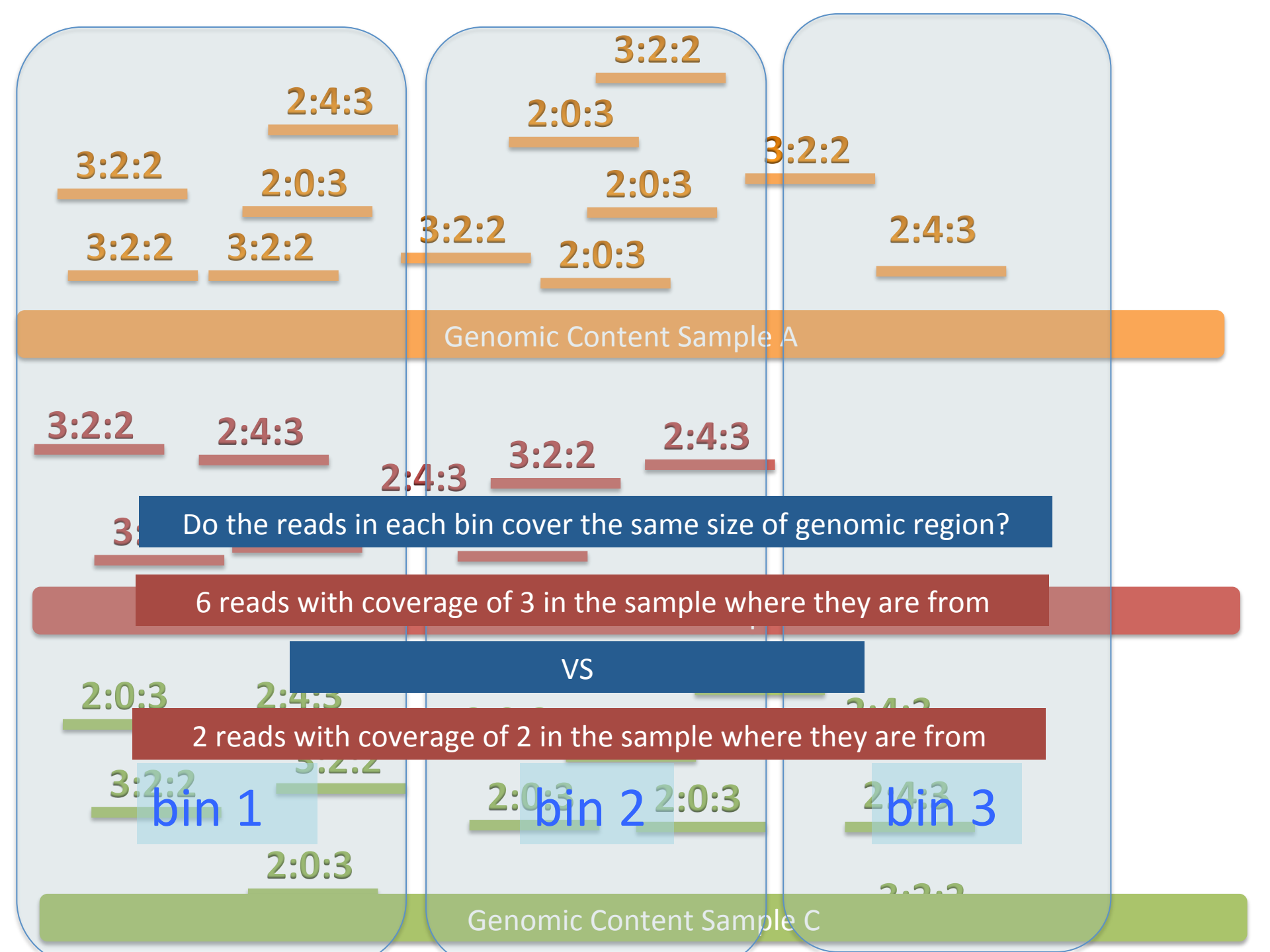
Sample B

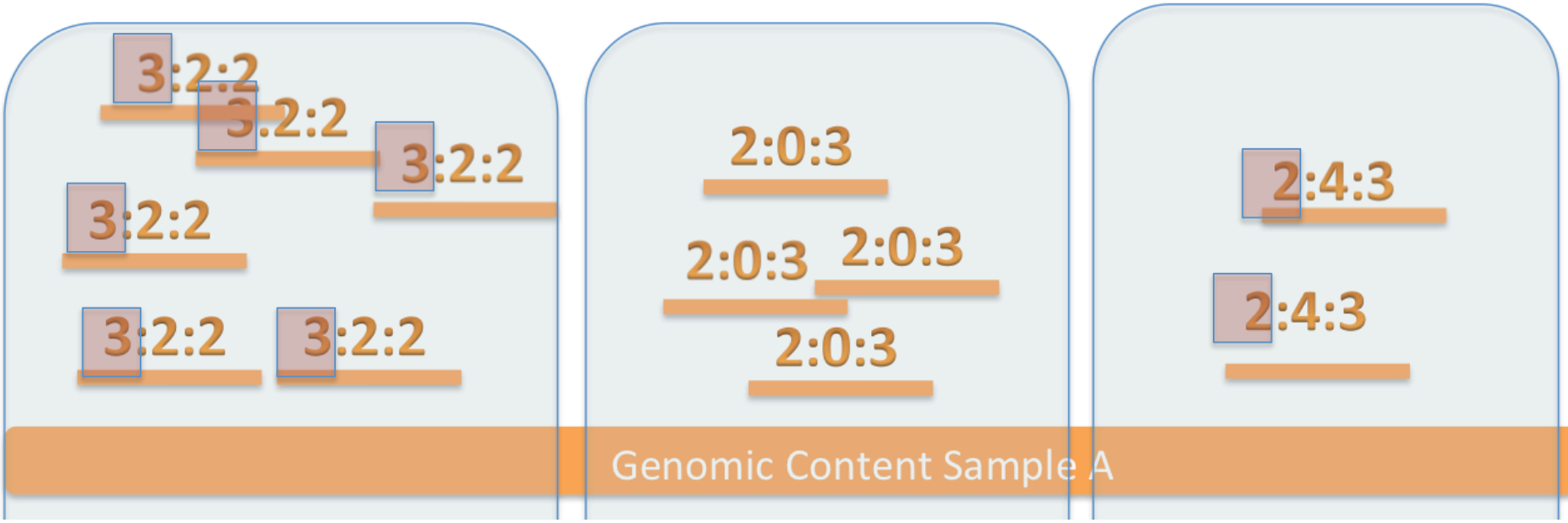


2:0:3
2:4:3
3:2:2
3:2:2
2:0:3
2:0:3
3:2:2
2:4:3
2:0:3
2:0:3
2:4:3
2:0:3
3:2:2

Sample C


3:2:2
2:4:3
2:0:3
3:2:2
2:0:3
3:2:2
2:0:3
3:2:2
3:2:2
3:2:2
2:0:3
2:4:3
Sample A
"Contigs with similar coverage profiles are likely to have originated from the same microbial population" (Imelfort et al. 2014)
Reads with similar coverage profiles are likely to have originated from the same species.
Sample B
3:2:2
3:2:2
Sample C
2:0:3
3:2:2
3:2:2
2:0:3
2:0:3
2:4:3
2:0:3
3:2:2

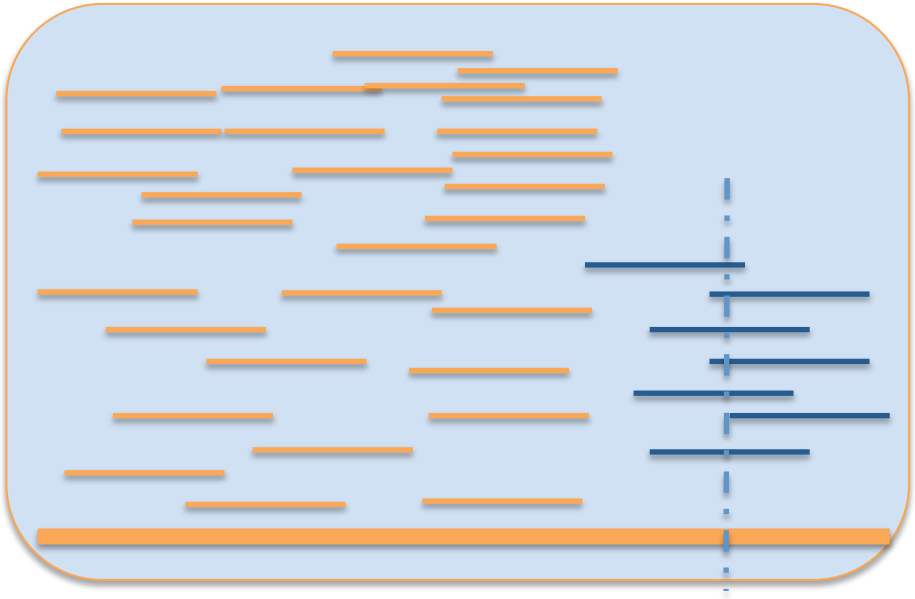

3:2:2
2:4:3
2:0:3
3:2:2
2:0:3
3:2:2
2:0:3
3:2:2
3:2:2
2:0:3
2:4:3
Genomic Content Sample A
3:2:2
2:4:3
2:4:3
3:2:2
2:4:3
Do the reads in each bin cover the same size of genomic region?
6 reads with coverage of 3 in the sample where they are from
VS
2:0:3
2 reads with coverage of 2 in the sample where they are from
bin 1
bin 2
bin 3
2:0:3
Genomic Content Sample C



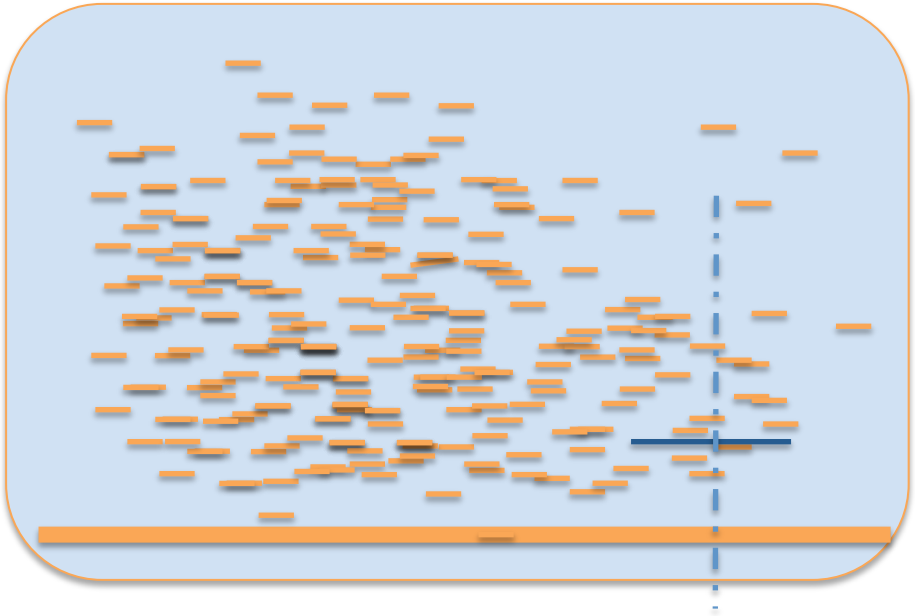
What is the size of genomic region covered by the reads in each bin?

What we know:
C: coverage of the read in the sample in the bin
N: number of reads in the bin

Reads – sample A

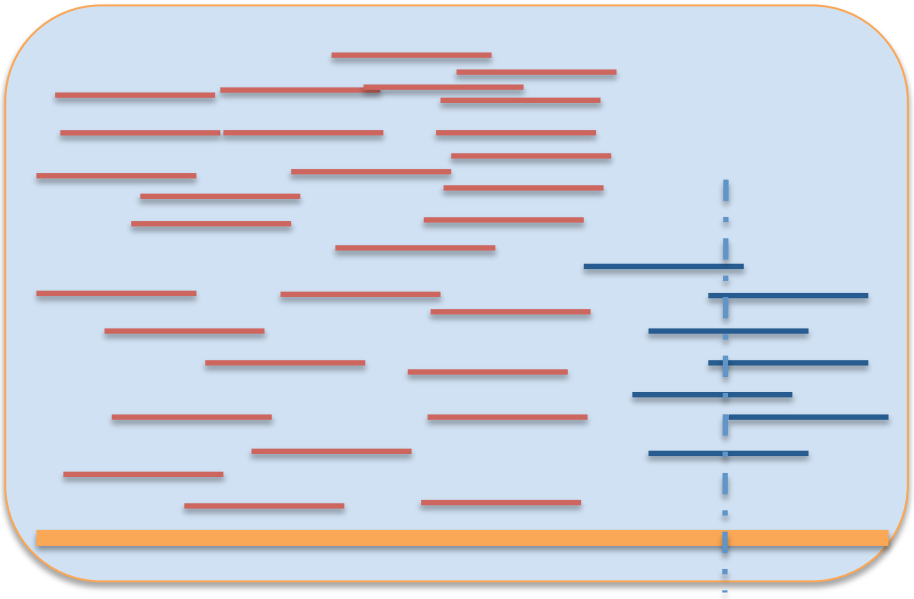


K-mers – sample A

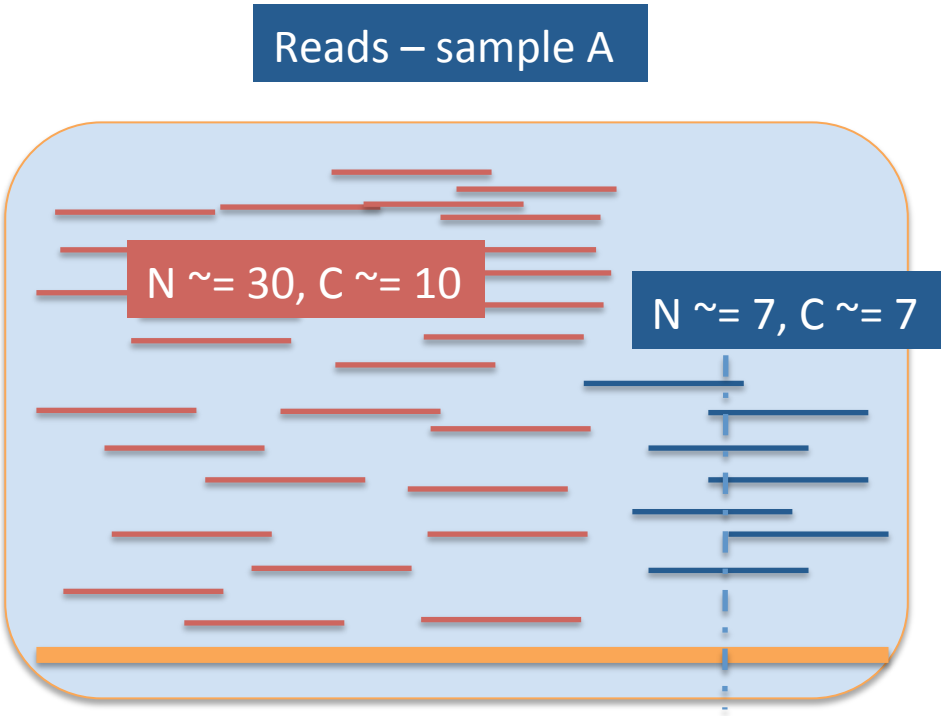


Trick 3: *If a read has a coverage as C, there will be* ***approximately*** *other C-1 reads with the same coverage covering the same DNA region.*

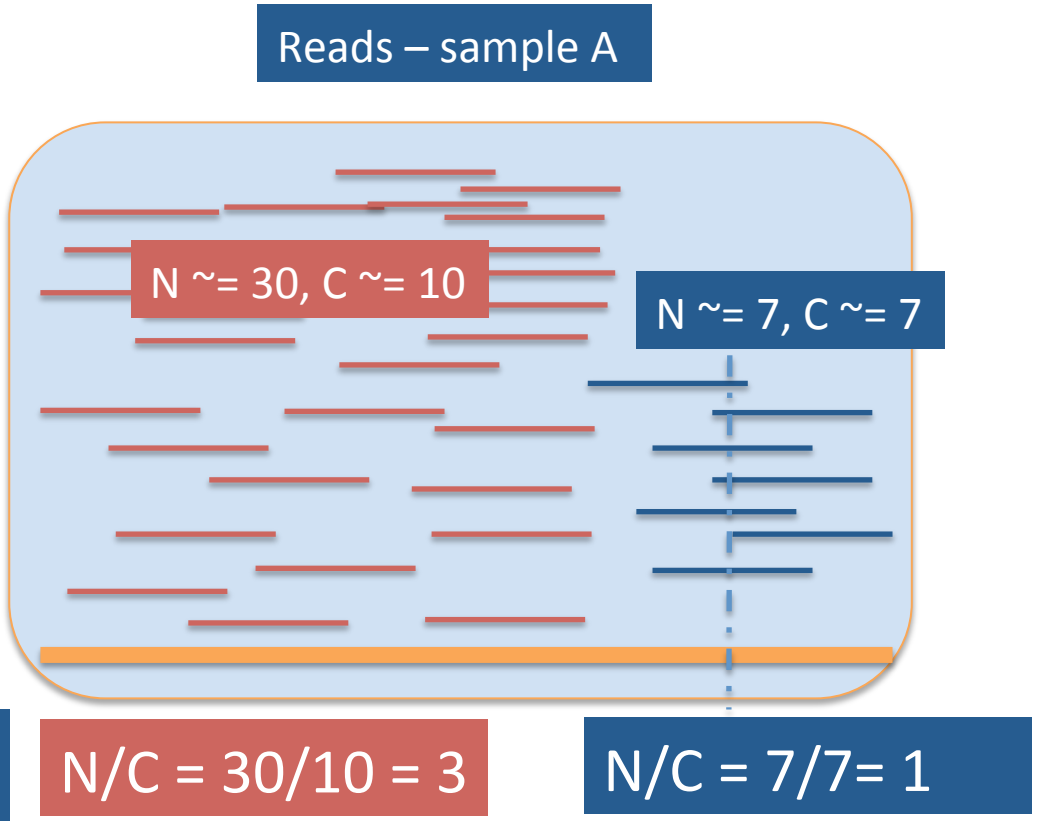
Reads – sample A



***Trick** 4: We can estimate the size of covered genomic region by the number of reads(N) and the corresponding coverage(C).*

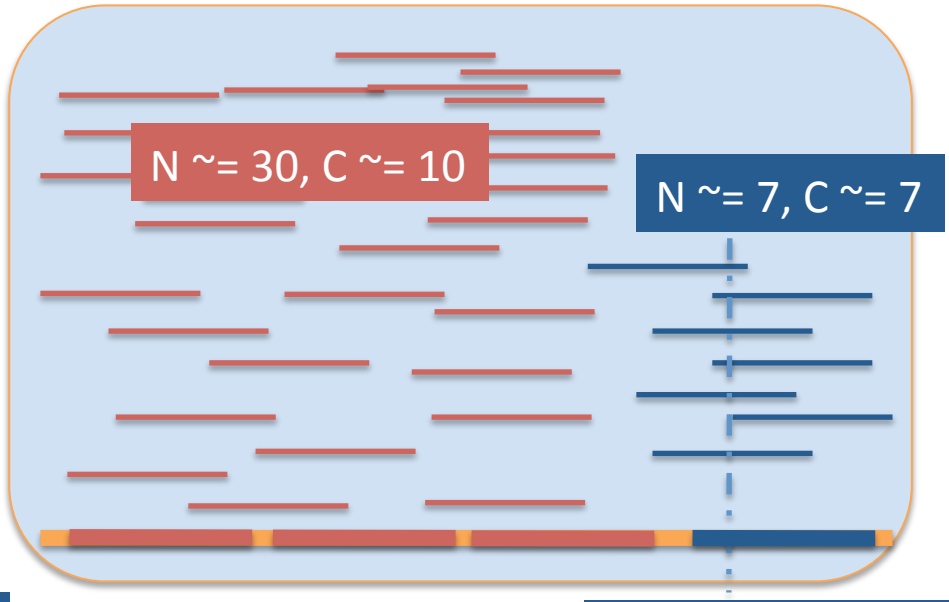


***Trick** 4: We can estimate the size of covered genomic region by the number of reads(N) and the corresponding coverage(C).*



***Trick** 4: We can estimate the size of covered genomic region by the number of reads(N) and the corresponding coverage(C).*

Reads – sample A



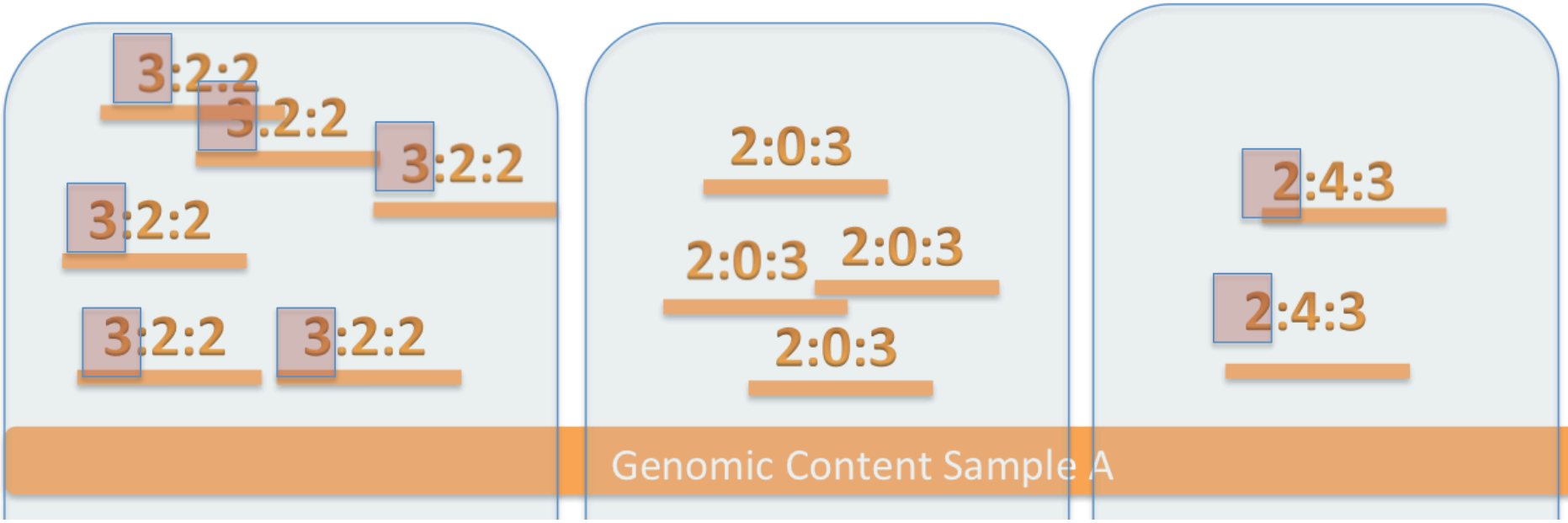
Size of covered DNA region

N/C = 30/10 = 3

N/C = 7/7= 1

Unit of size of covered DNA region:
IGS - IGS(informative genomic segment)

***Trick** 4: We can estimate the size of covered genomic region by the number of reads(N) and the corresponding coverage(C).*

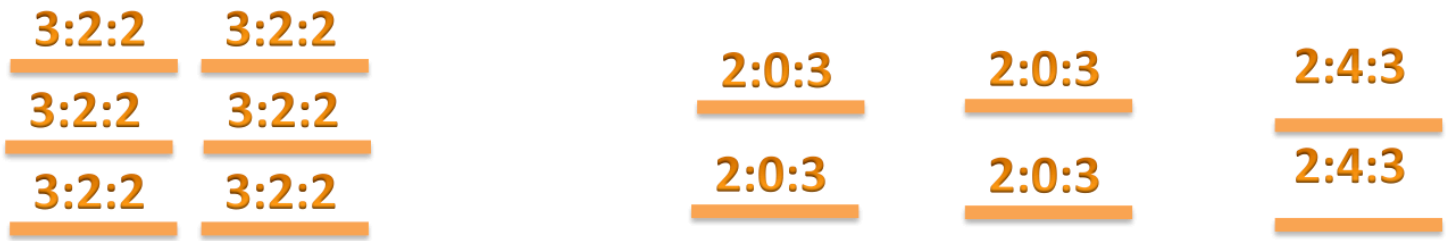


Do the reads in each bin cover the same size of genomic region?

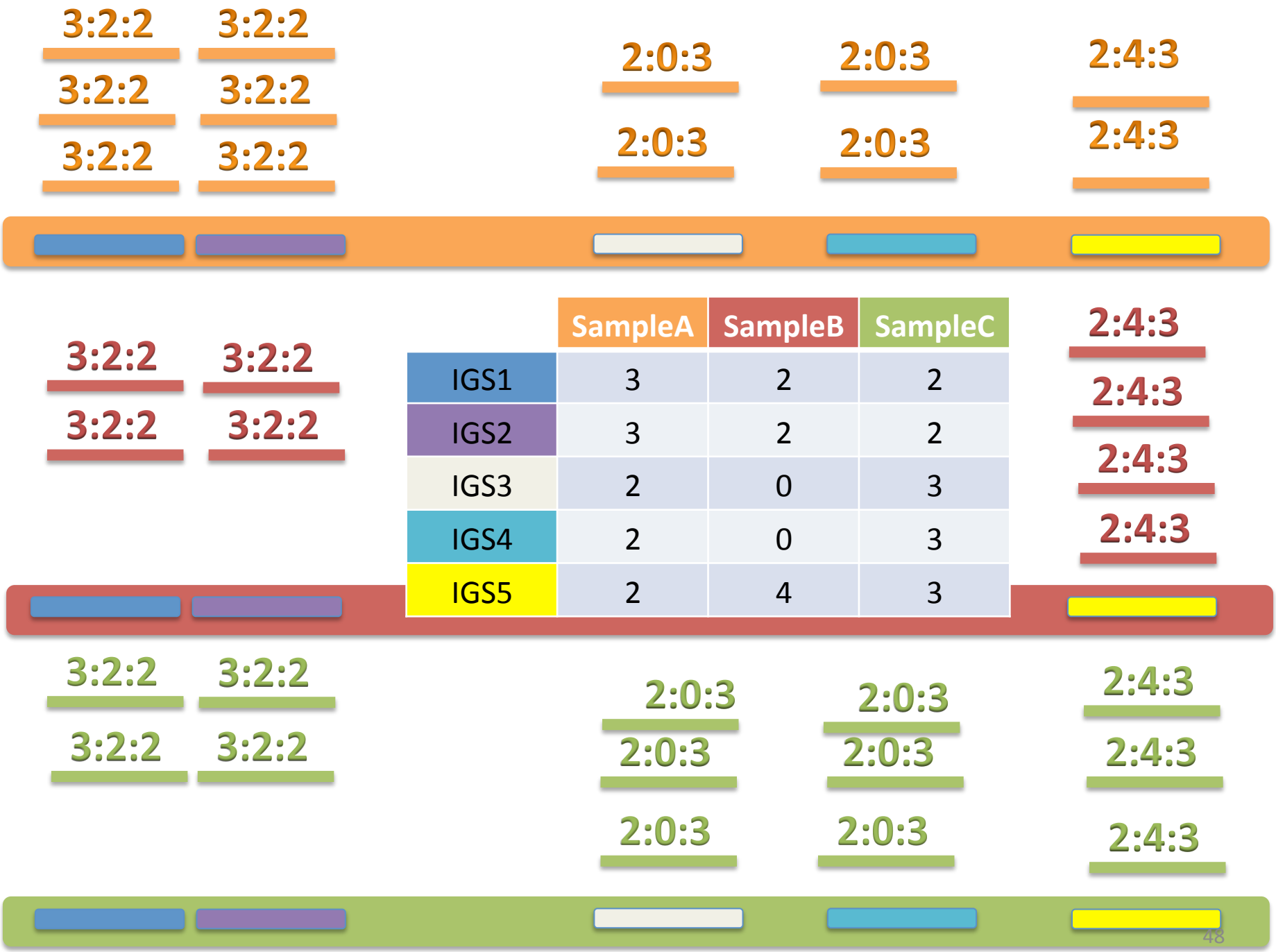
6 reads with coverage of 3 in the sample where they are from

VS

2 reads with coverage of 2 in the sample where they are from

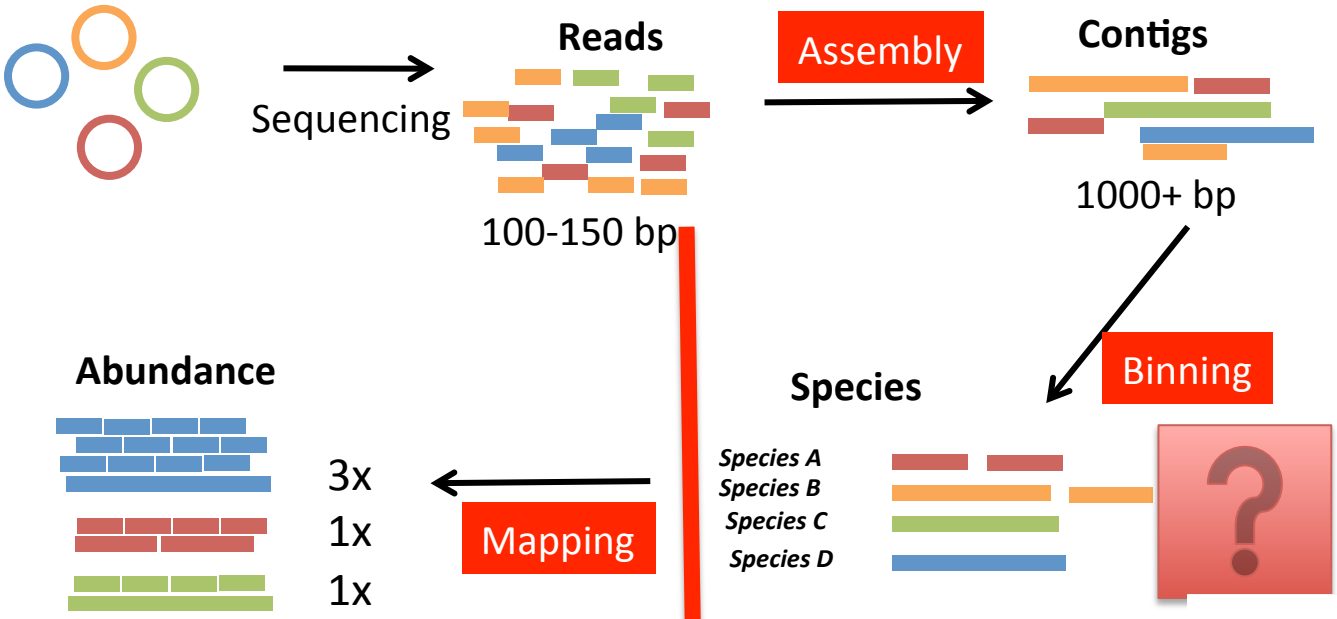


6/3 = 2 | 4/2 = 2 | 2/2 = 1

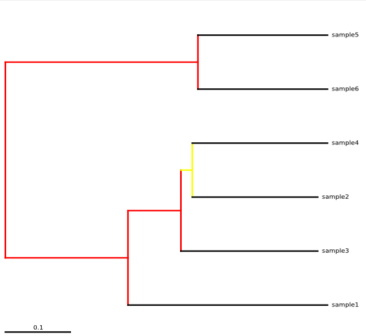


	SampleA	SampleB	SampleC
IGS1	3	2	2
IGS2	3	2	2
IGS3	2	0	3
IGS4	2	0	3
IGS5	2	4	3

A typical pipeline of metagenome diversity analysis



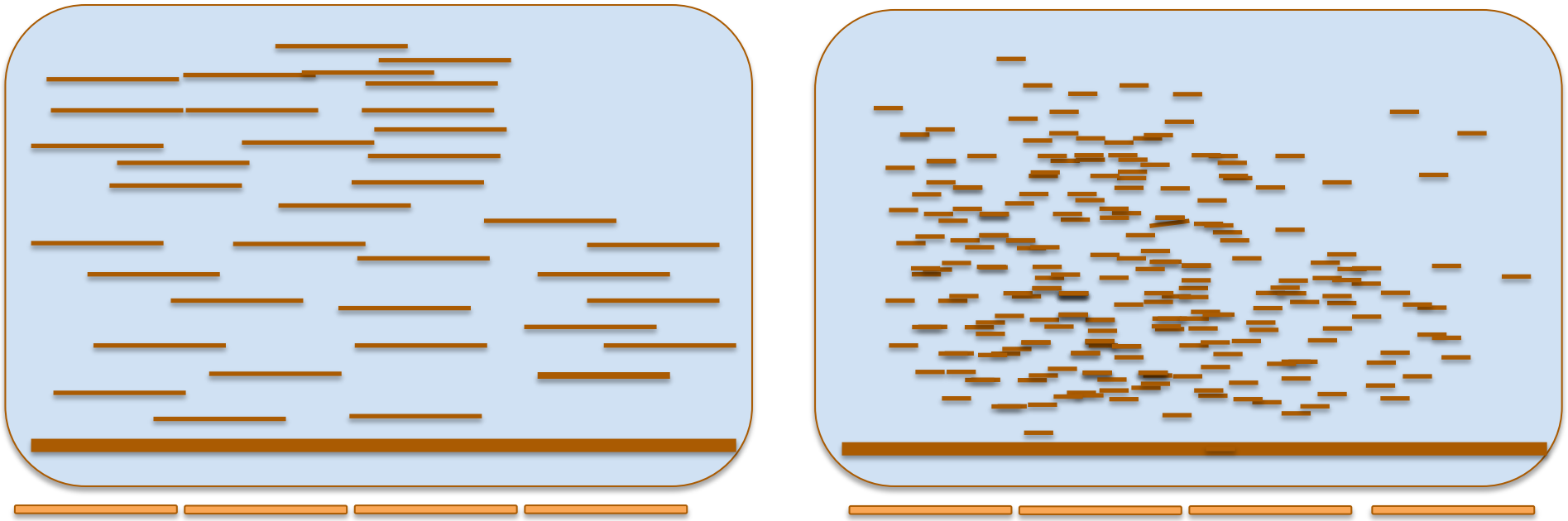
	Sample1	Sample2	Sample3	Sample4	Sample5
IGS1	3	3	4	4	0
IGS2	1	3	0	2	2
IGS3	1	2	2	2	3
IGS4	2	1	2	1	1



Challe

- **Ass** ass
- **Binning?** nally expensive)
- **Mapping?** – difficult (eg. accuracy, computationally expensive)
- **4 petabase** pairs of DNA in a gram of soil(Jack A. Gilbert,2013)

IGS(informative genomic segment) can represent the novel information of a genome



IGS (informative genomic segment):
genomic segments that do not share genomic content with the same length of read

- The more IGSs there are in a sample, the higher diversity , higher richness the sample has.
- The more IGSs two samples share, the more similar the two samples are.

(meta)Genome size = number of IGSs x IGS length (read length)

Outline

- Background and motivation
- An efficient k-mer counting approach
- Novel method to investigate microbial diversity
  - Concept of IGS (informative genomic segment)
  - Testing IGS method on simulated data sets
  - Testing IGS method on real metagenome data
- Summary

Simulated data sets with different species composition

sample ID	species composition	sequencing depth	abundance of species	size of metagenome (bp)
sample1	AAAB	10	A:30 B:10	200K
sample2	AABC	10	A:20 B:10 C:10	300K
sample3	ABCD	10	A:10 B:10 C:10 D:10	400K
sample4	ABCE	10	A:10 B:10 C:10 E:10	400K
sample5	AFGH	10	A:10 F:10 G:10 H:10	400K
sample6	IFGH	10	I:10 F:10 G:10 H:10	400K

No sequencing error is introduced in this simulated data

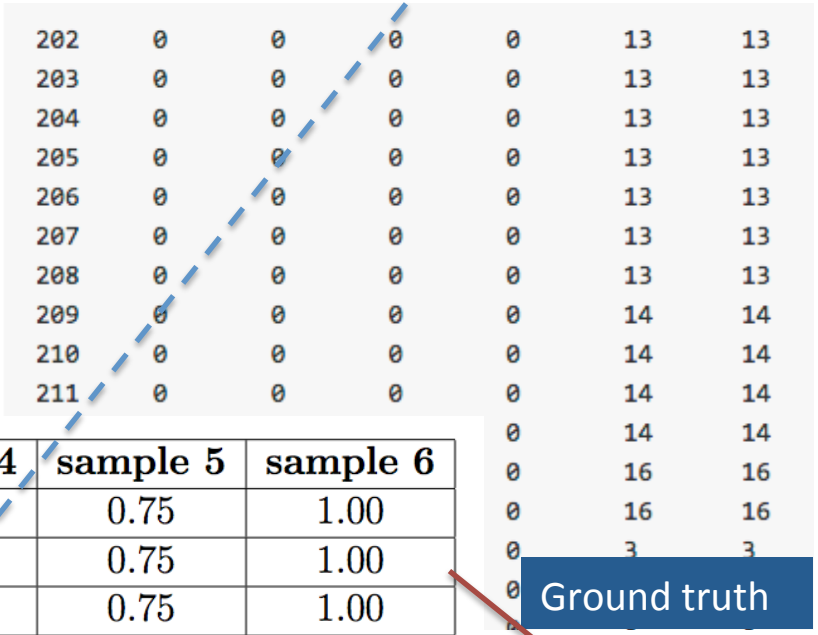
	observed IGS	ACE	simpson evenness	estimated genome size (Kbp)	real genome size (Kbp)
sample1	2002	2002.0	0.76	200.2	200
sample2	3038	3038.0	0.83	303.8	300
sample3	4076	4076.0	0.91	407.6	400
sample4	4078	4078.0	0.91	407.8	400
sample5	4069	4069.0	0.91	406.9	400
sample6	4087	4087.0	0.91	408.7	400

1. Estimated (meta)genome size matches real size perfectly.
2. Evenness metric can represent species distribution correctly

Samples-by-species table

	sample1	sample2	sample3	sample4	sample5	sample6
genomeA	30	20	10	10	10	
genomeB	10	10	10	10		
genomeC		10	10	10		
genomeD			10			
genomeE				10		
genomeF					10	10
genomeG					10	10
genomeH					10	10
genomeI						10

Samples-by-IGSs table



202	0	0	0	0	13	13
203	0	0	0	0	13	13
204	0	0	0	0	13	13
205	0	0	0	0	13	13
206	0	0	0	0	13	13
207	0	0	0	0	13	13
208	0	0	0	0	13	13
209	0	0	0	0	14	14
210	0	0	0	0	14	14
211	0	0	0	0	14	14
				0	14	14
				0	16	16
				0	16	16
				0	3	3

	sample 1	sample2	sample 3	sample 4	sample 5	sample 6
sample 1	0.00	0.25	0.50	0.50	0.75	1.00
sample 2	0.25	0.00	0.25	0.25	0.75	1.00
sample 3	0.50	0.25	0.00	0.25	0.75	1.00
sample 4	0.50	0.25	0.25	0.00	0.75	1.00
sample 5	0.75	0.75	0.75	0.75	0.00	0.25
sample 6	1.00	1.00	1.00	1.00	0.25	0.00

Ground truth

Mantel Correlation: 0.9714

	sample 1	sample2	sample 3	sample 4	sample 5	sample 6
sample 1	0.00	0.35	0.60	0.66	0.80	1.00
sample 2	0.35	0.00	0.42	0.51	0.84	1.00
sample 3	0.60	0.42	0.00	0.56	0.89	1.00
sample 4	0.66	0.51	0.56	0.00	0.89	1.00
sample 5	0.80	0.84	0.89	0.89	0.00	0.42
sample 6	1.00	1.00	1.00	1.00	0.25	0.00

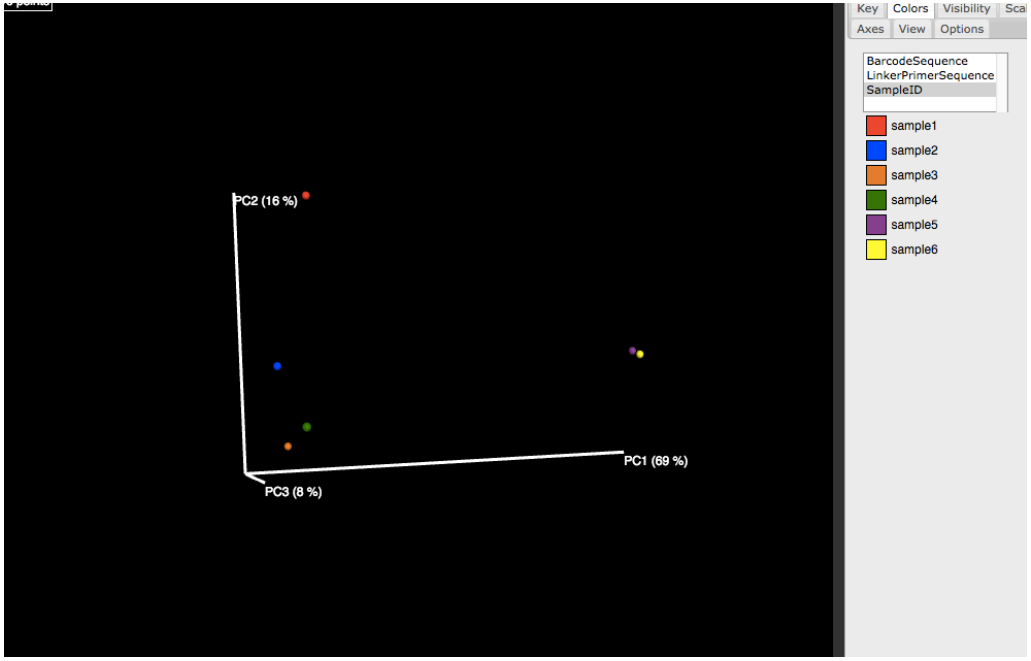
Using IGS method

Calculated dissimilarity matrix matches the matrix calculated from species composition.

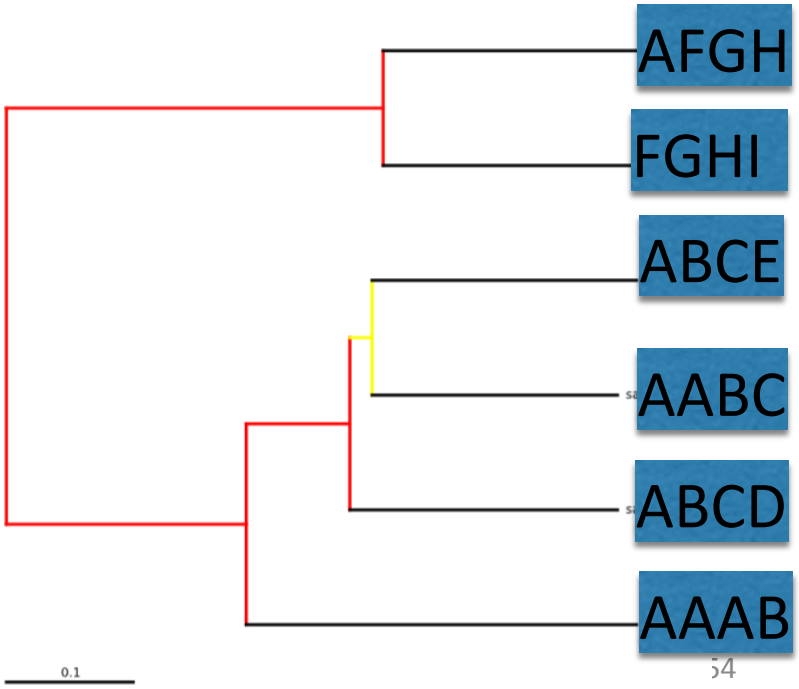
With an accurate dissimilarity matrix, it will be straightforward to perform PCoA or clustering analysis to demonstrate the relationship between samples.

	sample 1	sample2	sample 3	sample 4	sample 5	sample 6
sample 1	0.00	0.35	0.60	0.66	0.80	1.00
sample 2	0.35	0.00	0.42	0.51	0.84	1.00
sample 3	0.60	0.42	0.00	0.56	0.89	1.00
sample 4	0.66	0.51	0.56	0.00	0.89	1.00
sample 5	0.80	0.84	0.89	0.89	0.00	0.42
sample 6	1.00	1.00	1.00	1.00	0.25	0.00

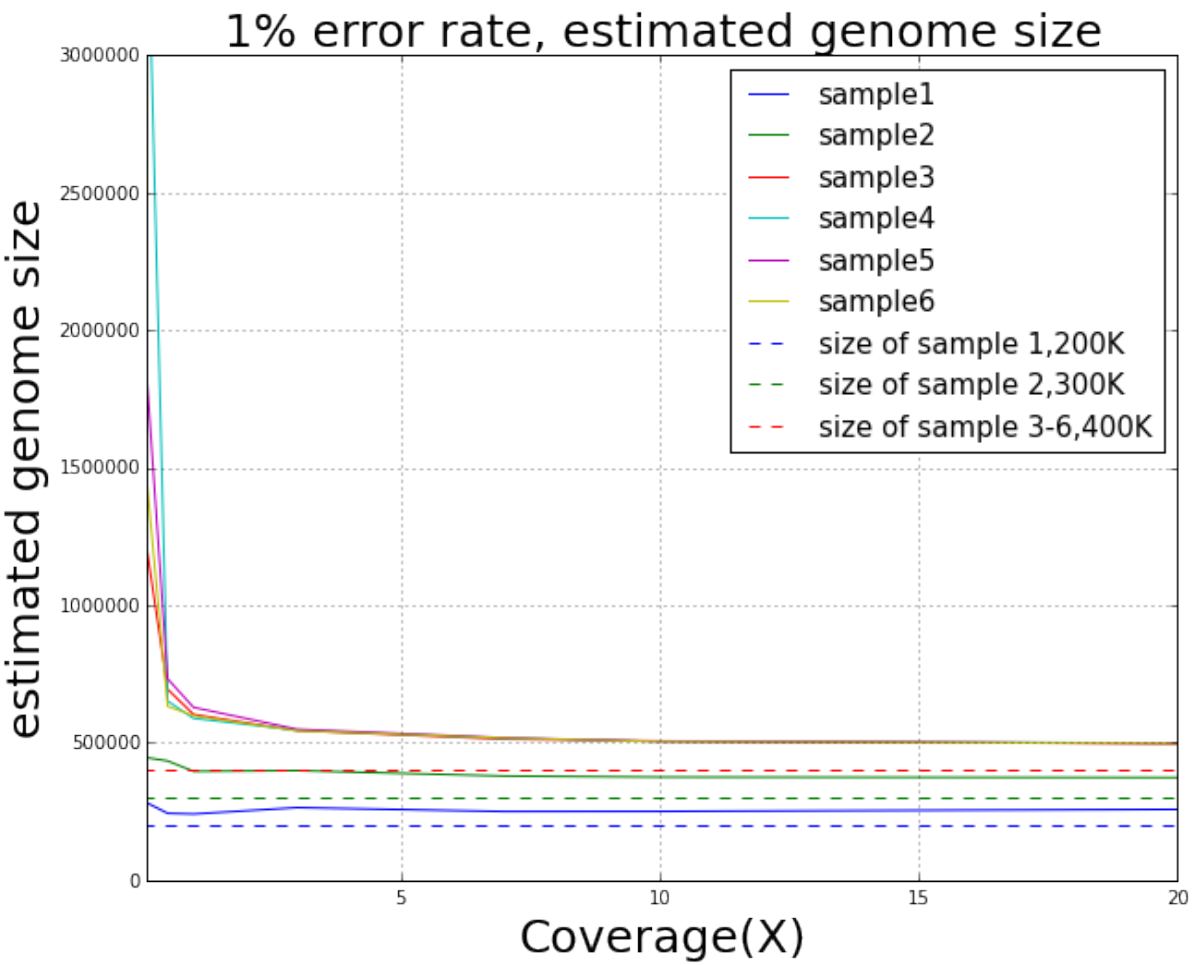
Ordination (PCoA)



Clustering

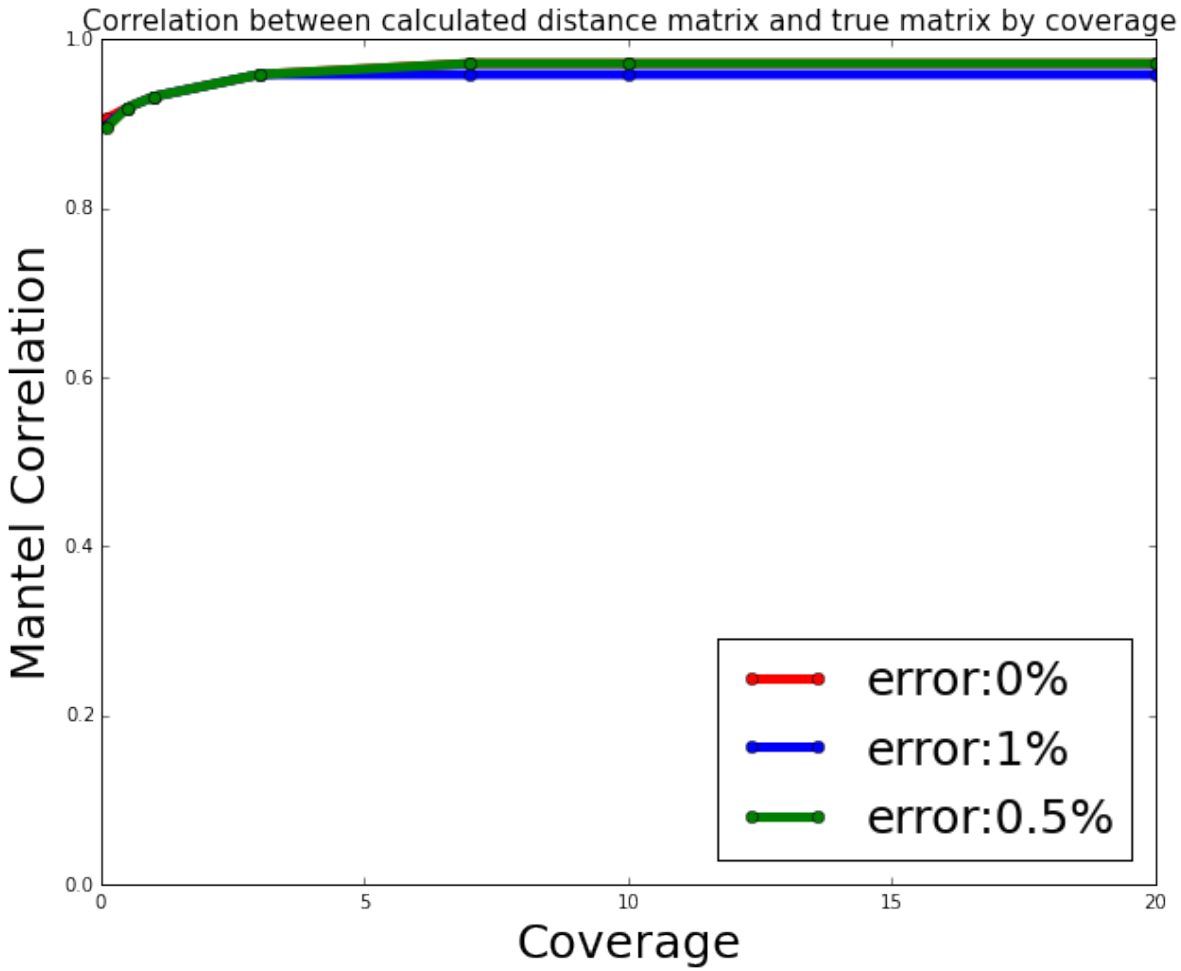


Influence of error rate and coverage to alpha diversity analysis



Metagenome size is more accurate with increasing coverage and lower error rate

Influence of error rate and coverage to beta diversity analysis



Beta diversity is less prone to error rate and low coverage

Outline

- Background and motivation
- An efficient k-mer counting approach
- Novel method to investigate microbial diversity
  - Concept of IGS (informative genomic segment)
  - Testing IGS method on simulated data sets
  - Testing IGS method on real metagenome data
- Summary

Sorcerer II Global Ocean Sampling Expedition

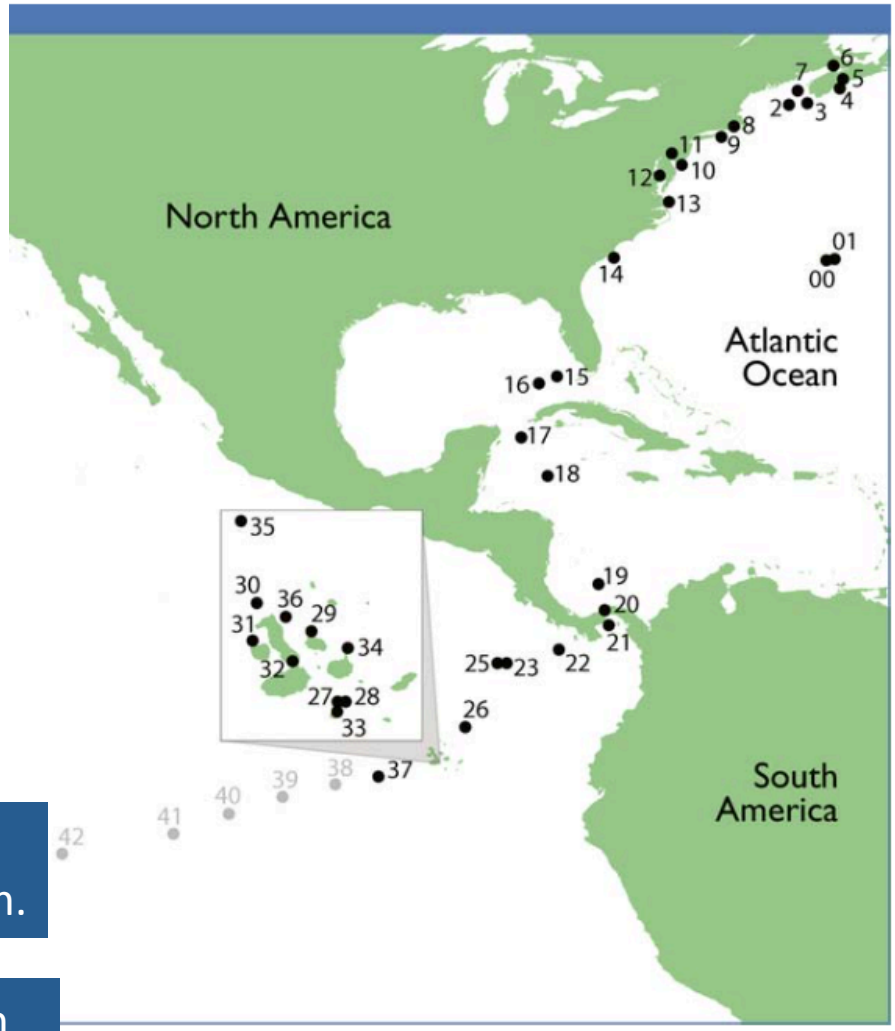
OPEN ACCESS Freely available online PLoS BIOLOGY

The *Sorcerer II* Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific

Douglas B. Rusch[1*], Aaron L. Halpern[1], Granger Sutton[1], Karla B. Heidelberg[1,2], Shannon Williamson[1], Shibu Yooseph[1], Dongying Wu[1,3], Jonathan A. Eisen[1,3], Jeff M. Hoffman[1], Karin Remington[1,4], Karen Beeson[1], Bao Tran[1], Hamilton Smith[1], Holly Baden-Tillson[1], Clare Stewart[1], Joyce Thorpe[1], Jason Freeman[1], Cynthia Andrews-Pfannkoch[1], Joseph E. Venter[1], Kelvin Li[1], Saul Kravitz[1], John F. Heidelberg[1,2], Terry Utterback[1], Yu-Hui Rogers[1], Luisa I. Falcón[5], Valeria Souza[5], Germán Bonilla-Rosso[5], Luis E. Eguiarte[5], David M. Karl[6], Shubha Sathyendranath[7], Trevor Platt[7], Eldredge Bermingham[8], Victor Gallardo[9], Giselle Tamayo-Castillo[10], Michael R. Ferrari[11], Robert L. Strausberg[1], Kenneth Nealson[1,12], Robert Friedman[1], Marvin Frazier[1], J. Craig Venter[1]

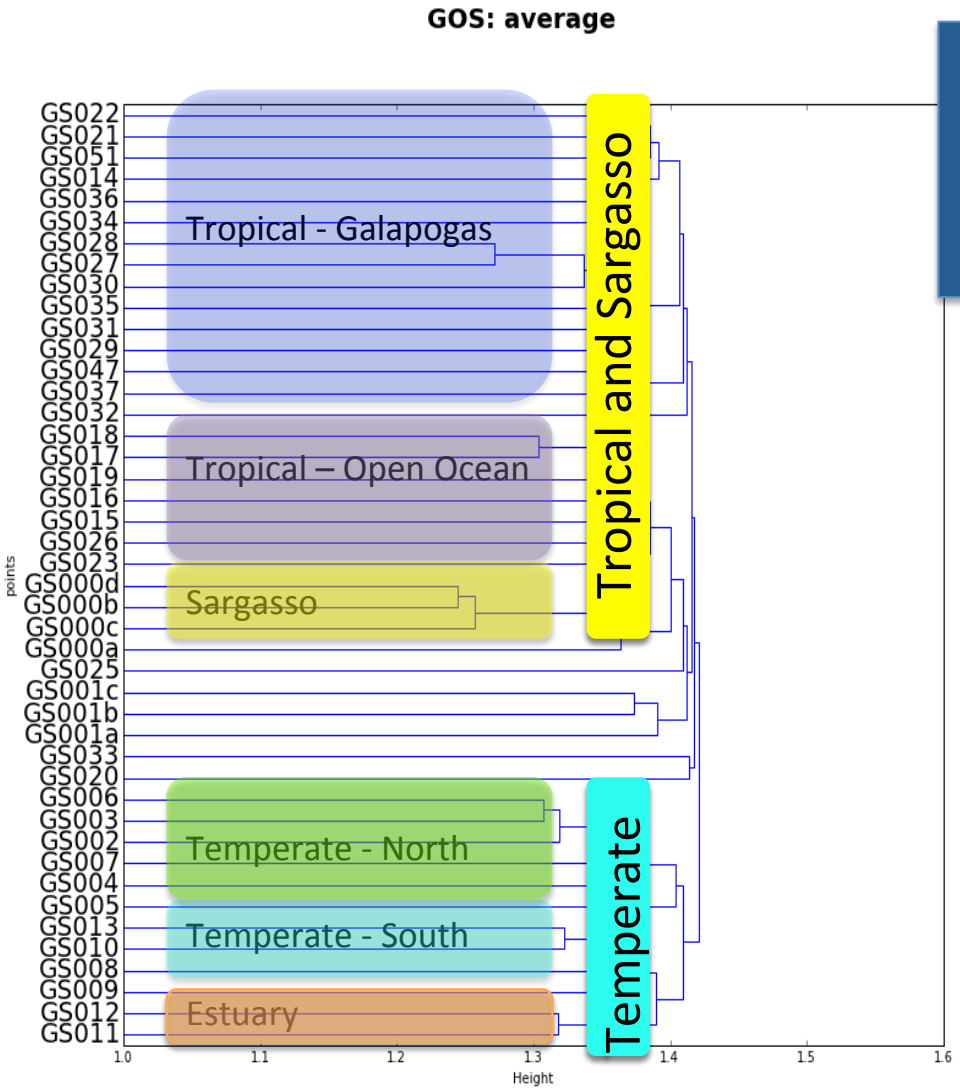
1 J. Craig Venter Institute, Rockville, Maryland, United States of America, **2** Department of Biological Sciences, University of Southern California, Avalon, California, United States of America, **3** Genome Center, University of California Davis, Davis, California, United States of America, **4** Your Genome, Your World, Rockville, Maryland, United States of America, **5** Departmento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, Mexico City, Mexico, **6** Department of Oceanography, University of Hawaii, Honolulu, Hawaii, United States of America, **7** Bedford Institute of Oceanography, Dartmouth, Nova Scotia, Canada, **8** Smithsonian Tropical Research Institute, Balboa, Ancon, Republic of Panama, **9** Departamento de Oceanografía, Universidad de Concepción, Concepción, Chile, **10** Escuela de Química, Universidad de Costa Rica, San Pedro, Costa Rica, **11** Department of Environmental Sciences, Rutgers University, New Brunswick, New Jersey, United States of America, **12** Department of Earth Sciences, University of Southern California, Los Angles, California, United States of America

The world's oceans contain a complex mixture of micro-organisms that are for the most part, uncharacterized both genetically and biochemically. We report here a metagenomic study of the marine planktonic microbiota in which surface (mostly marine) water samples were analyzed as part of the *Sorcerer II* Global Ocean Sampling expedition. These samples, collected across a several-thousand km transect from the North Atlantic through the Panama Canal and ending in the South Pacific yielded an extensive dataset consisting of 7.7 million sequencing reads (6.3 billion bp). Though a few major microbial clades dominate the planktonic marine niche, the dataset contains great diversity with 85% of the assembled sequence and 57% of the unassembled data being unique at a 98% sequence identity cutoff. Using the metadata associated with each sample and sequencing library, we developed new comparative genomic and assembly methods. One comparative genomic method, termed "fragment recruitment," addressed questions of genome structure, evolution, and taxonomic or phylogenetic diversity, as well as the biochemical diversity of genes and gene families. A second method, termed "extreme assembly," made possible the assembly and reconstruction of large segments of abundant but clearly nonclonal organisms. Within all abundant populations analyzed, we found

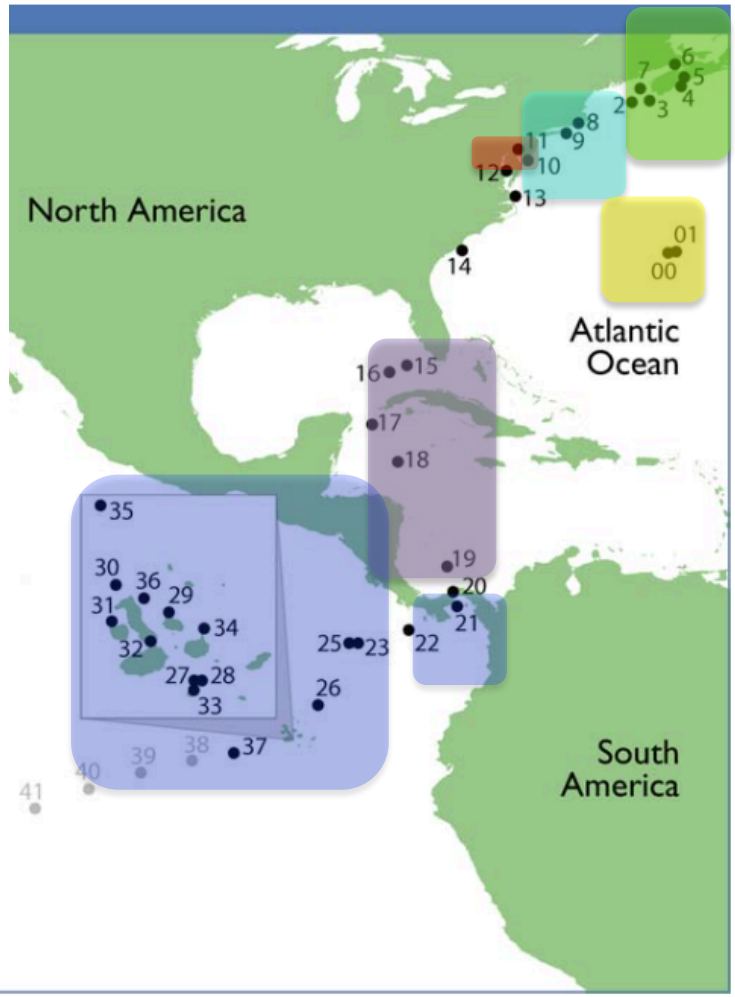


It has been calculated that microbes in ocean account for about half of the biomass on Earth.

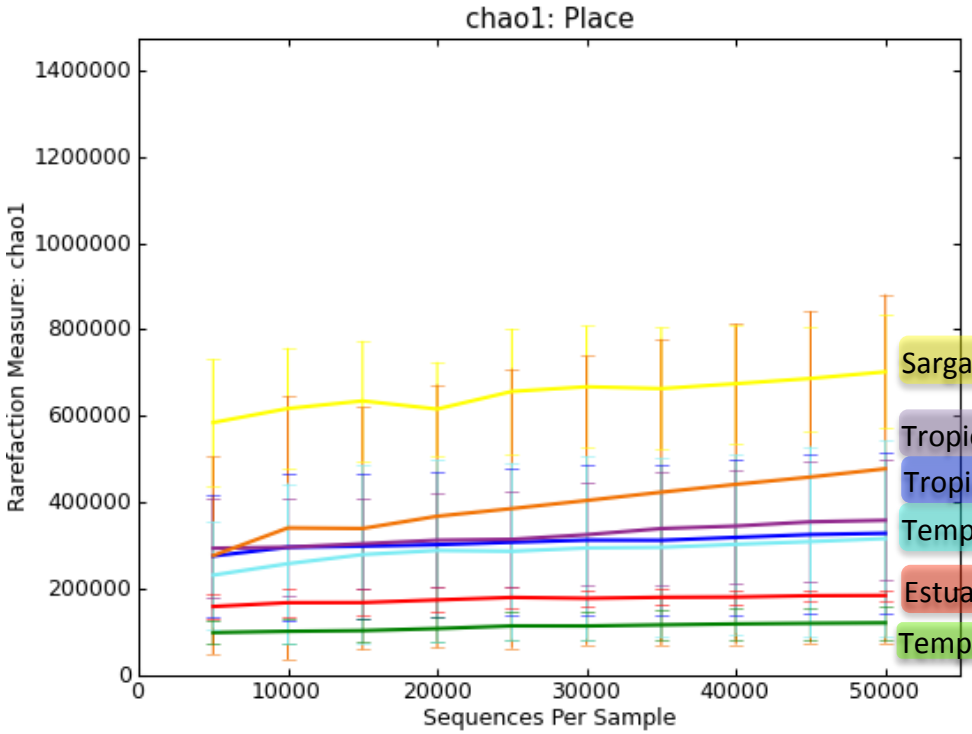
In a drop (one millilitre) of seawater, one can find 10 million viruses, one million bacteria

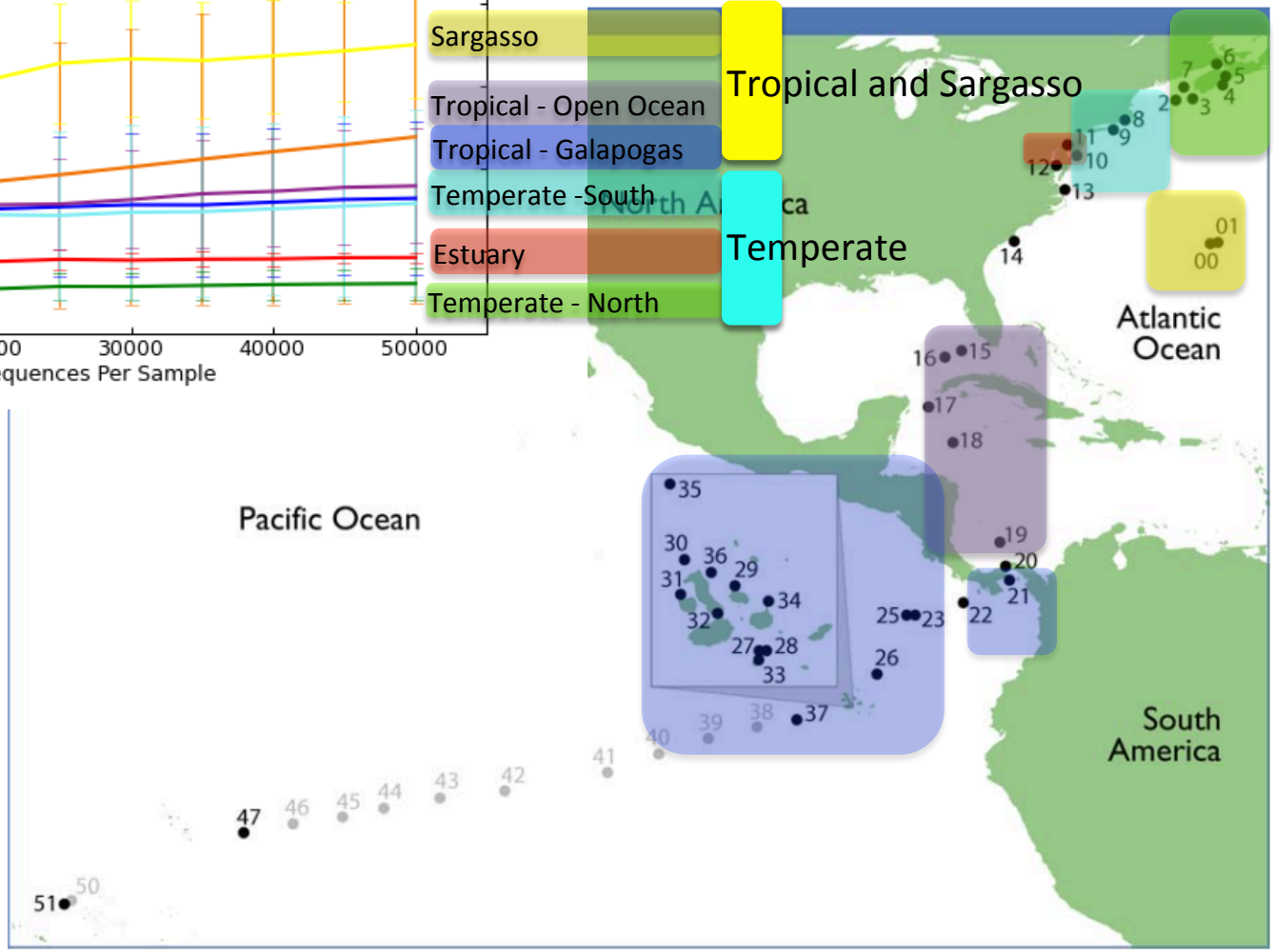


Samples can be separated by locations using IGS-based method – as good as the result in original research based on alignment-based method or better.

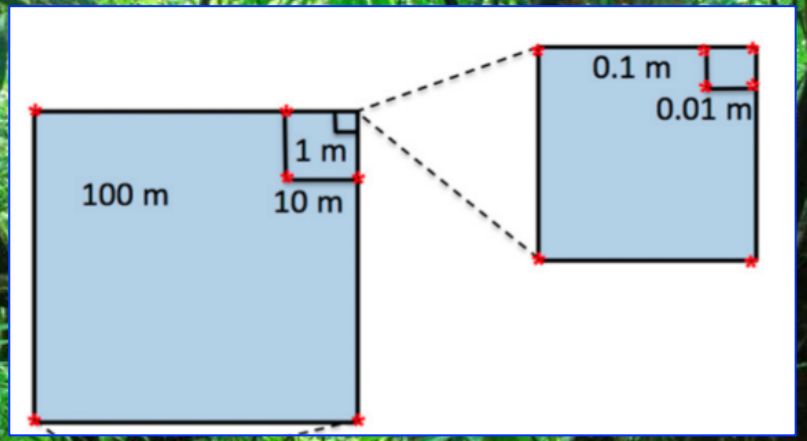


Alpha diversity analysis shows samples from tropical area have higher richness.



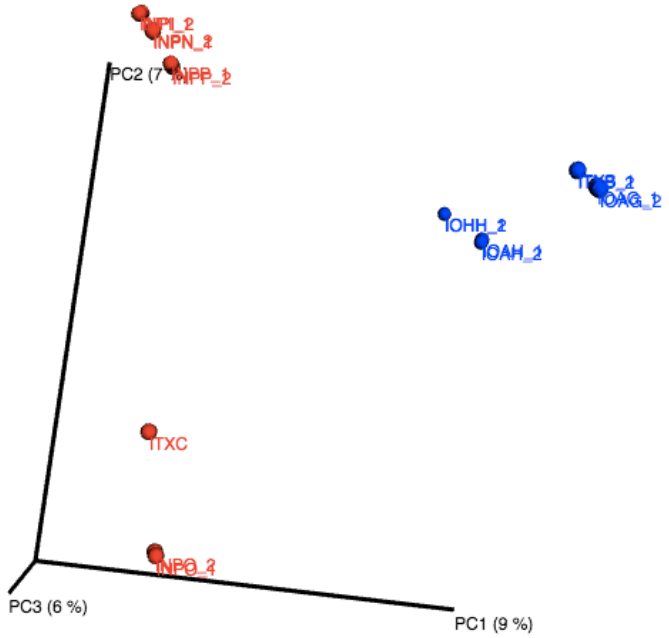


ARMO (Amazon Rain Forest Microbial Observatory) project



Soil metagenome sample	Number of reads	Number of k-mers k=29	Number of unique k-mers
IOHH.2301.7.1859	273,021,397	32,762,567,744	22,333,571,795

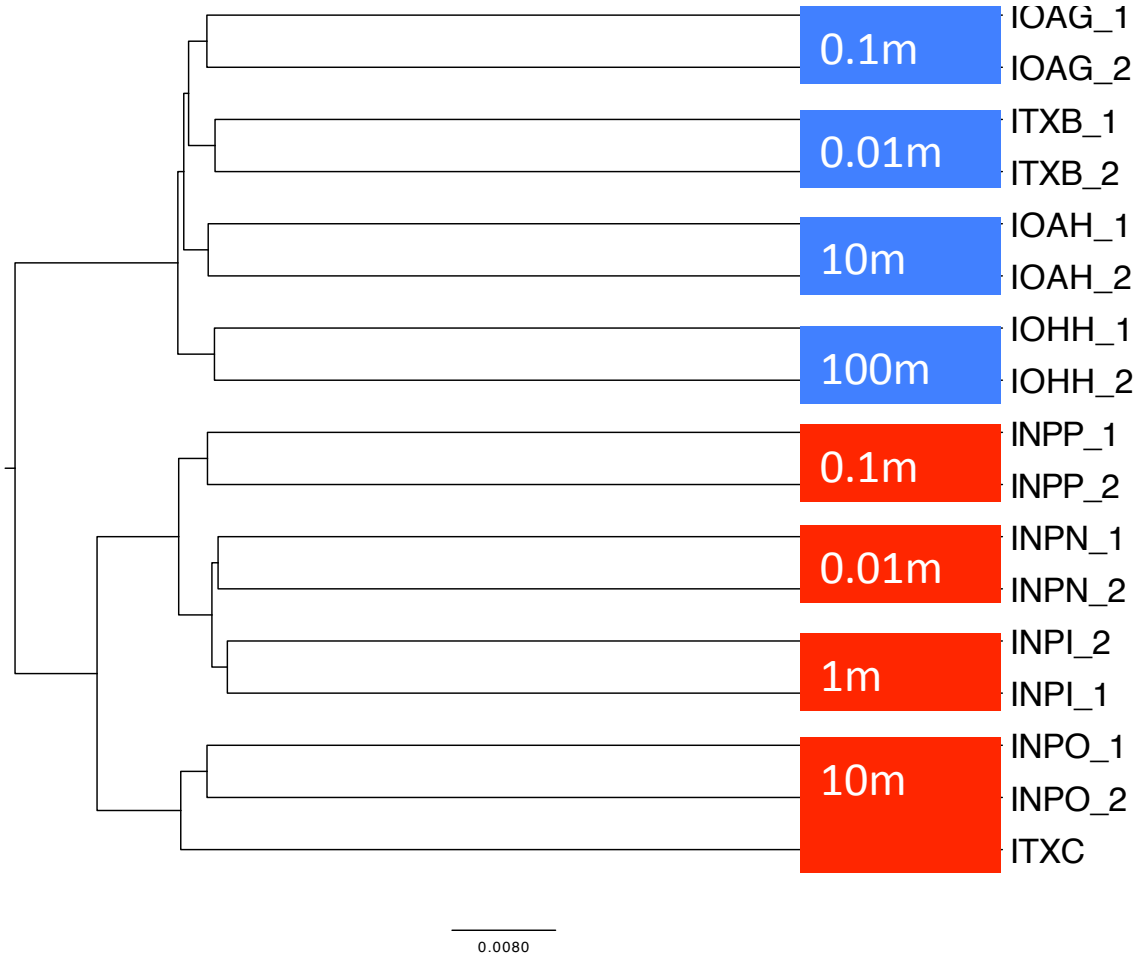
21 samples, 1T size of data, 5 billion reads.

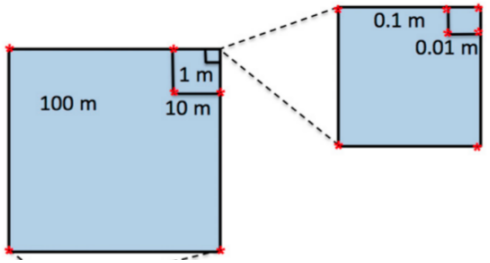


Blue: Prairie
Red: Forest

1. Conversion (forest -> pasture) affects composition
2. Prairie samples more similar in space
3. Subsamples with 8M reads are used.

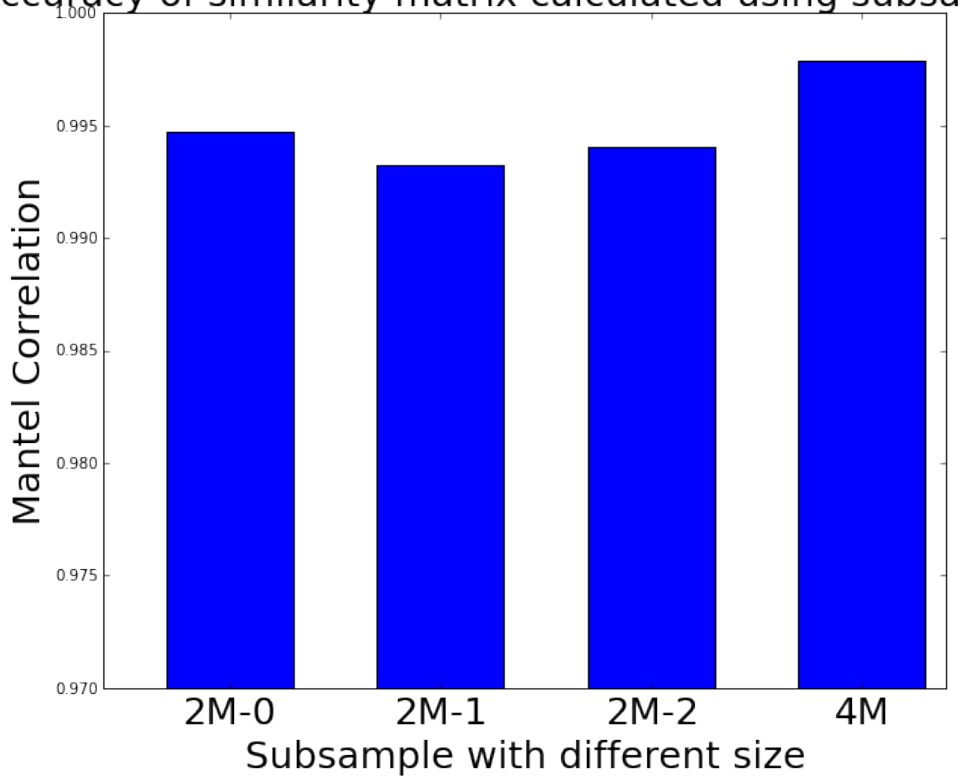
Samples can be clustered by different treatments and distances using IGS method.

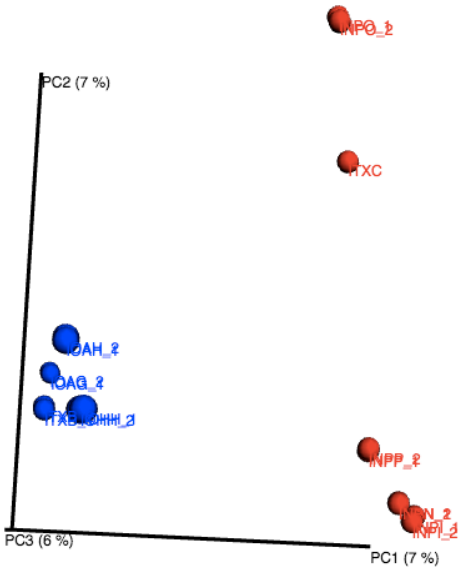




Blue: Prairie
Red: Forest

Accuracy of similarity matrix calculated using subsamples

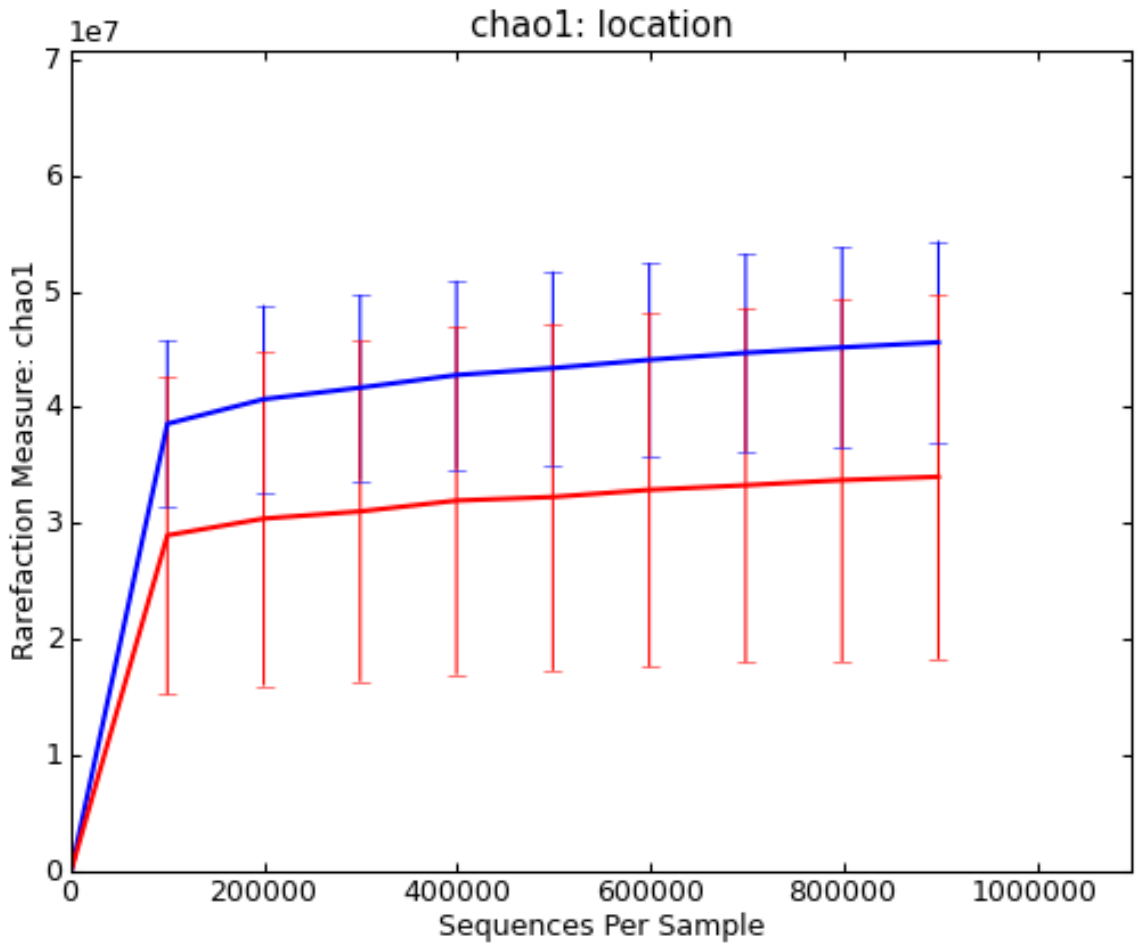




Blue: Prairie
Red: Forest

Mantel correlation to similarity matrix calculated from subsample with 8M reads

Using subsamples can get decent result of beta diversity.



Blue: Prairie
Red: Forest

Alpha diversity analysis shows samples from prairie area have higher richness than forest area.

Outline

- Background and motivation
- An efficient k-mer counting approach
- Novel method to investigate microbial diversity
  - Concept of IGS (informative genomic segment)
  - Testing IGS method on simulated data sets
  - Testing IGS method on real metagenome data
- Summary

Summary

- A whole framework with potentials:
  - Take all information into account, rare species
  - Alpha diversity (richness, evenness) (Chao1, ACE estimators, etc.)
  - Beta diversity (abundance-based, incidence-based)
  - Integrate with existing packages (khmer, mothur, QIIME, scikit-bio,etc.)
- Scalable, efficient:
  - Based on efficient k-mer counting
  - Using subsamples may be enough for specific task

What's next

- Extend applications beyond diversity analysis
  - Extracts interesting reads by coverage profile
  - Reads binning/classification based on reads coverage profile
  - Genome size estimation/sequencing effort evaluation
- Refine pipeline, optimize performance
  - Adjust IGS resolution
  - Improve methods to handle large number of IGSs more efficiently
  - Iterative beta diversity analysis

Publications

- IGS method **Using the concept of informative genomic segment to investigate microbial diversity of metagenomic sample** (in preparation)
- error analysis **Crossing the streams: a framework for streaming analysis of short DNA sequencing reads**. Qingpeng Zhang, Sharon Awad, C. Titus Brown. (submitted) https://peerj.com/preprints/890/
- khmer **These are not the k-mers you are looking for: efficient online k-mer counting using a probabilistic data structure.** Qingpeng Zhang, Jason Pell, Rosangela Canino-Koning, Adina Chuang Howe, C. Titus Brown. PLOS ONE, July 25, 2014, DOI: 10.1371/journal.pone.0101271 http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0101271
- digital normalization **A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data**. C. Titus Brown, Adina Howe, Qingpeng Zhang, Alexis B. Pyrkosz, Timothy H. Brom. Submitted. http://arxiv.org/abs/1203.4802
- seaworm symbiont **Genomic versatility and functional variation between two dominant heterotrophic symbionts of deep-sea Osedax worms**. Shana K Goffredi, Hana Yi, Qingpeng Zhang, Jane E Klann,Isabelle A Struve, Robert C Vrijenhoek and C Titus Brown. The ISME Journal doi:10.1038/ismej.2013.201

Acknowledgement

- **Dr. Titus Brown**
- **Lab members of GED**
  - Dr. Sherine Awad
  - Michael Crusoe
  - Jiarong Guo
  - Luiz Irber
  - Camille Scott
- **Collaborators**
  - Dr. James Tiedje
  - Dr. Joan Ross
  - Dr. Shana K Goffredi
  - Yiseul KIm
- **Committee members**
  - Jaimes Cole
  - Richard Enbody
  - Yanni Sun
  - Eric Torng
- **Former members of GED**
  - Dr. Adina Howe
  - Eric McDonald
  - Dr. Jason Pell
  - Dr. Likit Preeyanon
  - Dr. Elijah Lowe
- **khmer developers**