



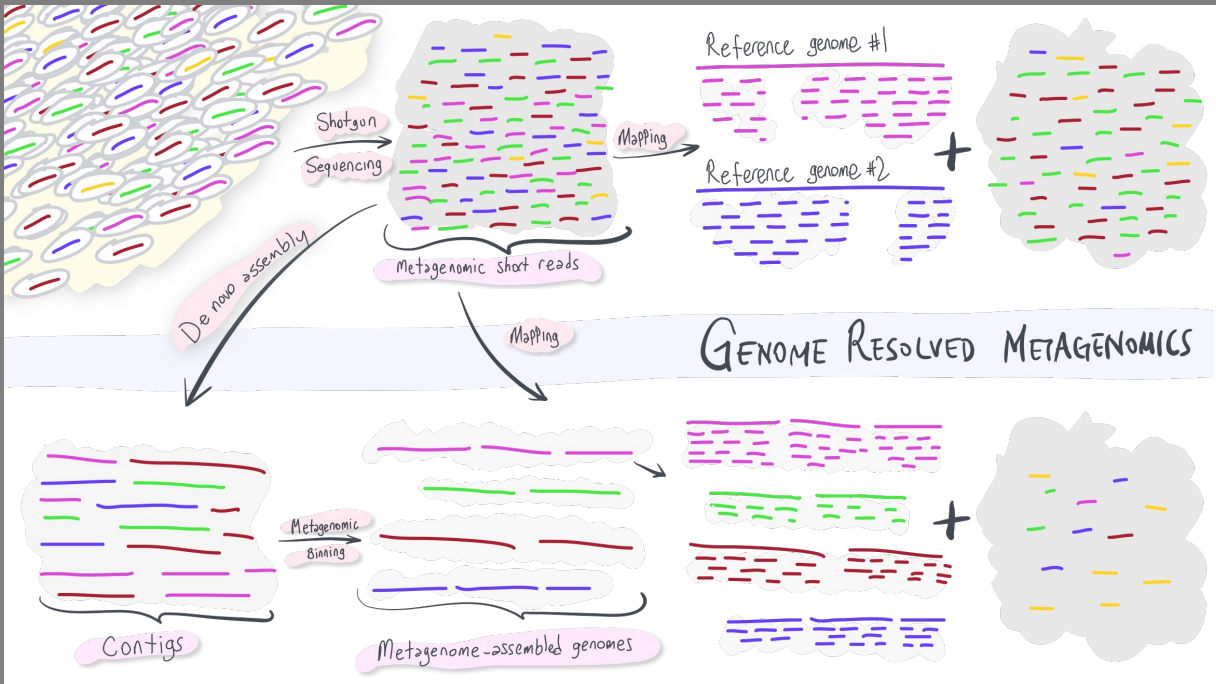
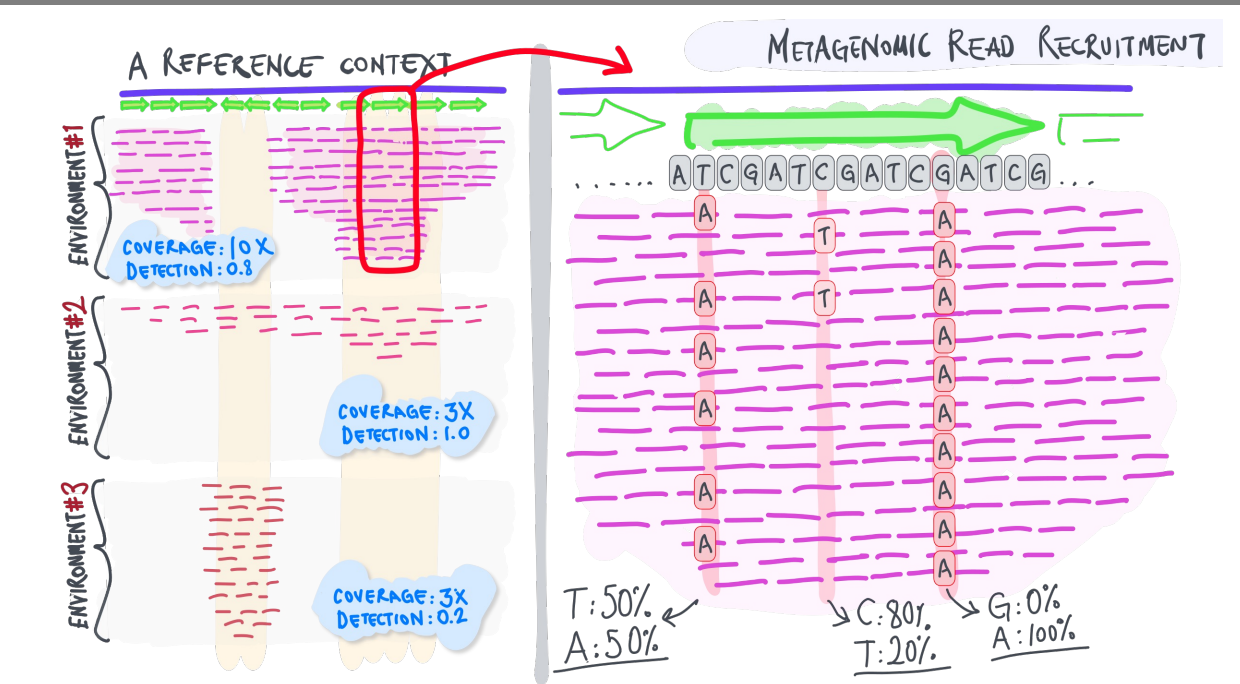
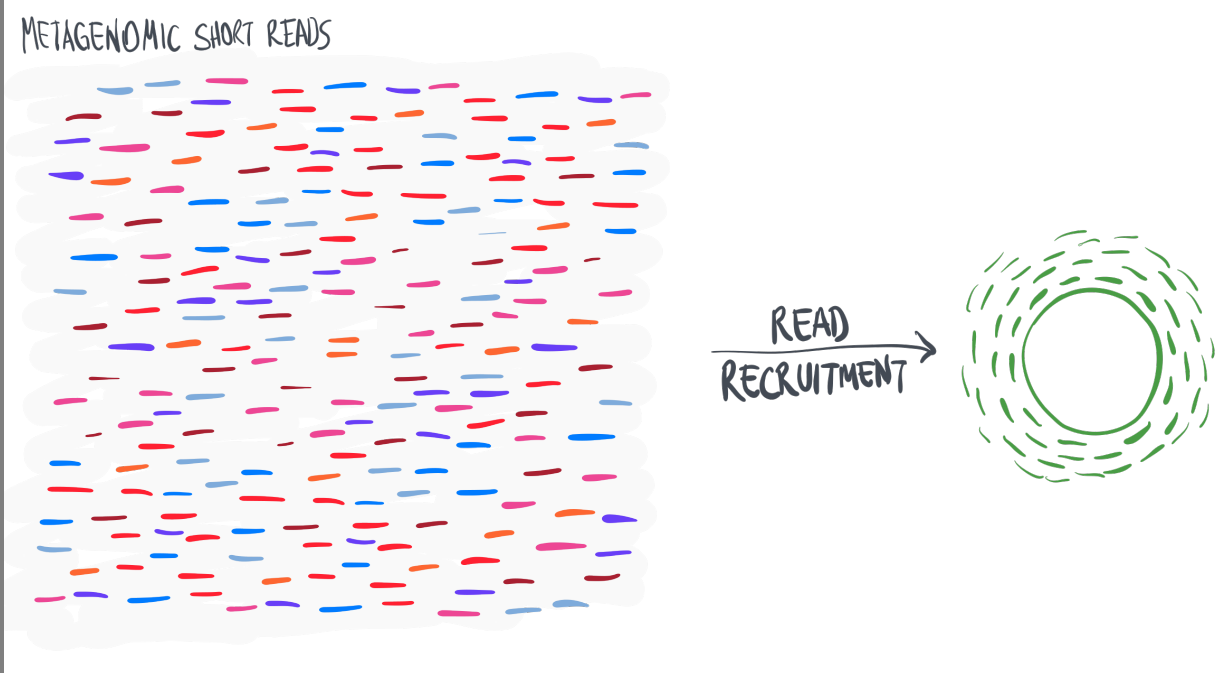
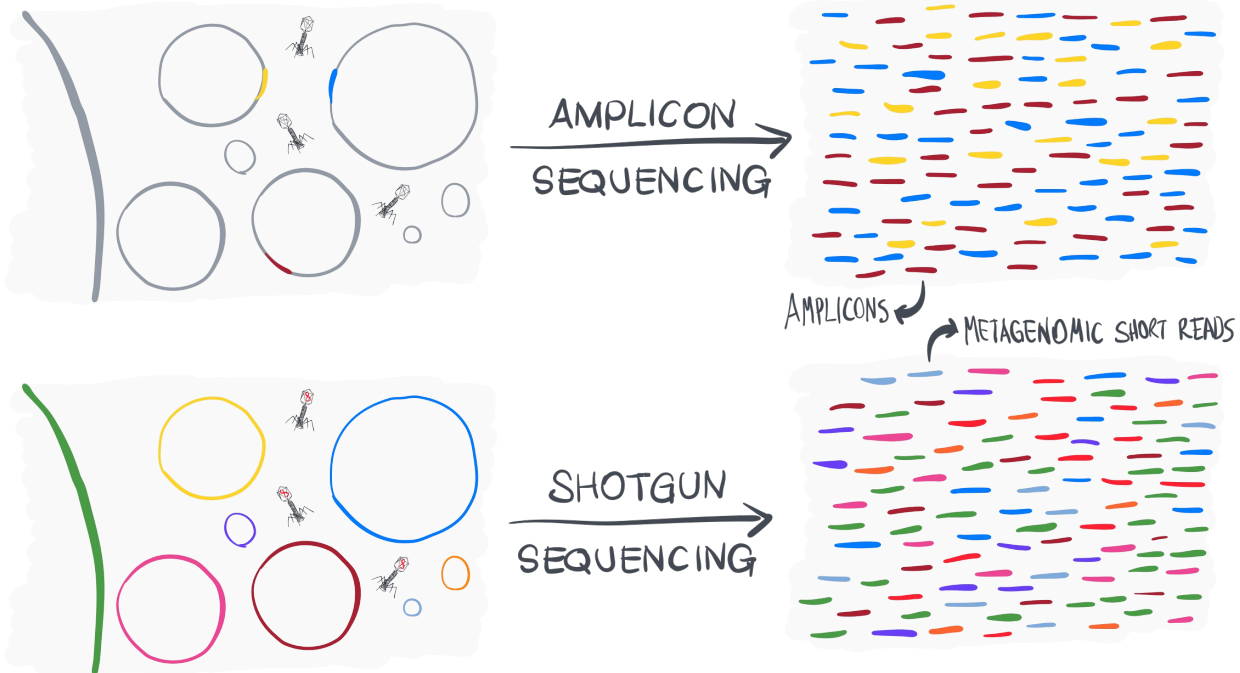
Previously in STAMPS

Introduction to

shotgun

metagenomics and

read recruitment



The background features the word "OMICS" in a large, stylized font. The letters "O", "M", "I", and "S" are rendered in a dark red, glossy, brush-stroke style. The letters "O", "M", and "S" are also overlaid with thick, black, calligraphic strokes. A vertical white line is positioned to the left of the text.

Now
Genome-resolved
metagenomics: key concepts
in **reconstructing genomes**
from metagenomes



Introduction to genome binning

Sequence composition

Computing k-mer frequencies

Differential coverage

Completion and contamination



> Introduction to genome binning

Sequence composition

Computing k-mer frequencies

Differential coverage

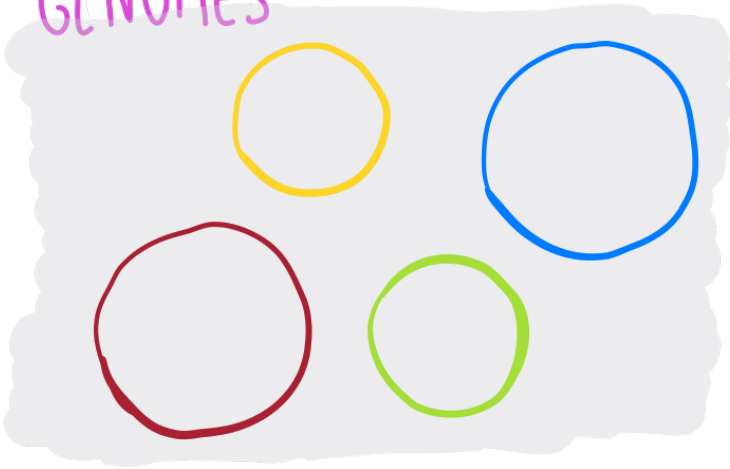
Completion and contamination

A bright sun with rays shining through a cloudy sky. The sun is the central focus, with its rays extending outwards. The clouds are soft and white, contrasting with the blue sky. The overall scene is bright and clear.

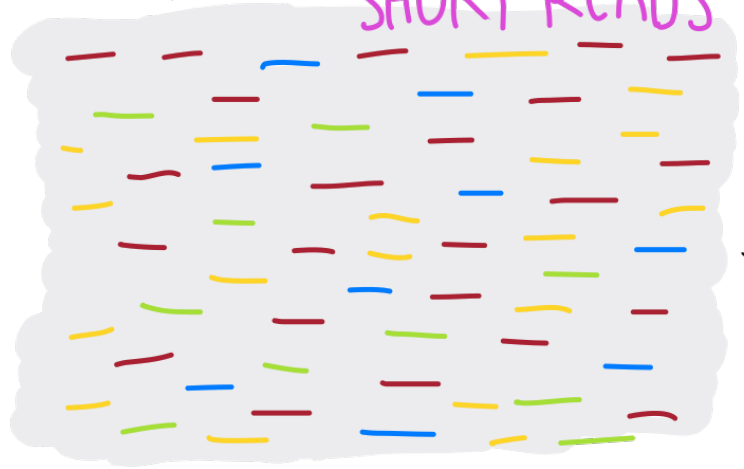
**Accessing genomes of
microbes we have not
yet cultivated**

SHOTGUN SEQUENCING

GENOMES

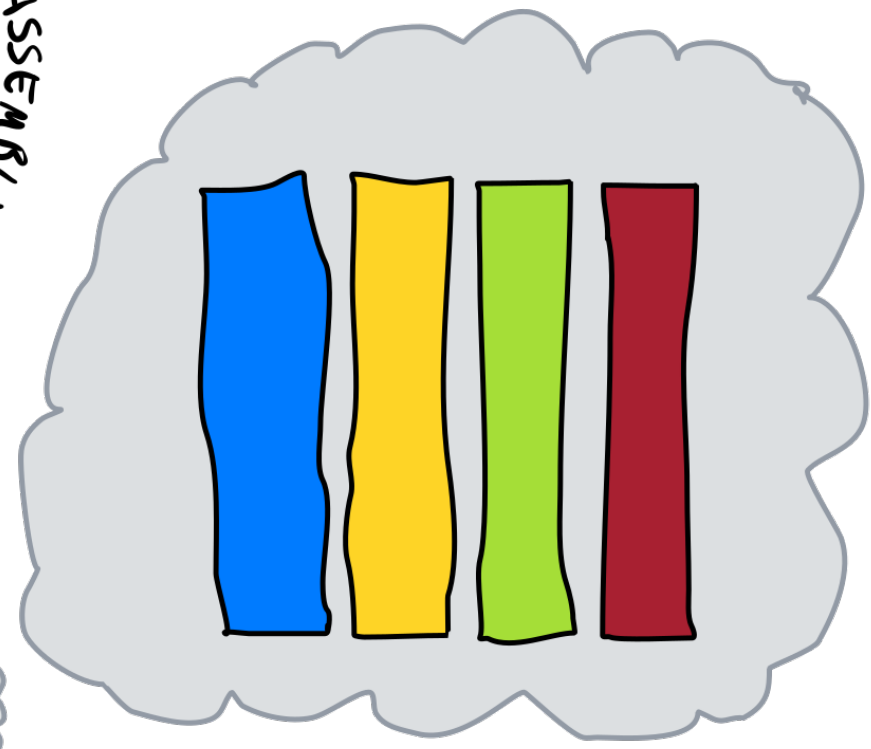
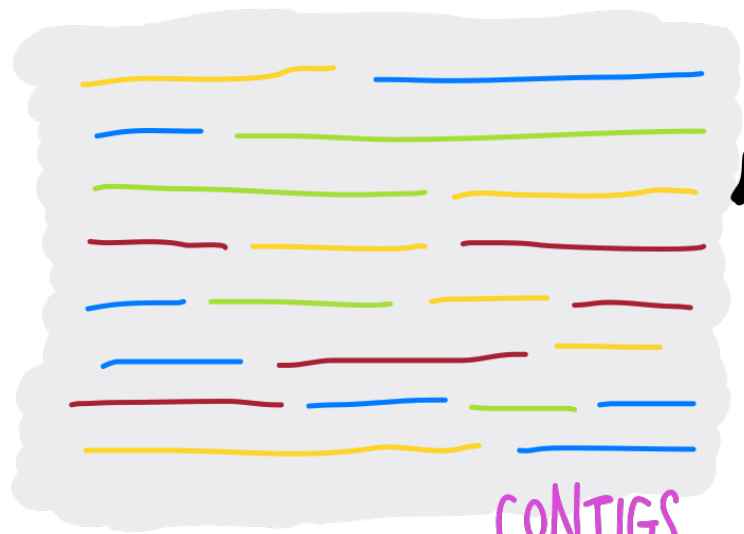


SHORT READS

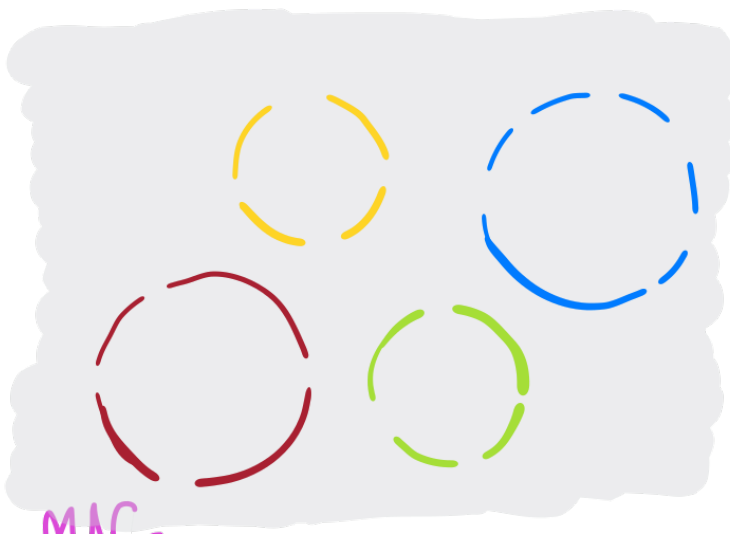


DE-NOVO ASSEMBLY

CONTIGS



MAGs



METAGENOMIC BINNING



SHOTGUN SEQUENCING

GENOMES

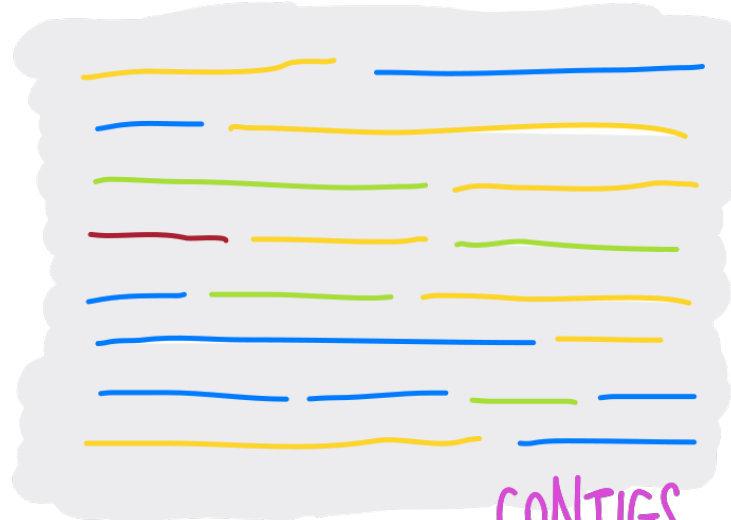
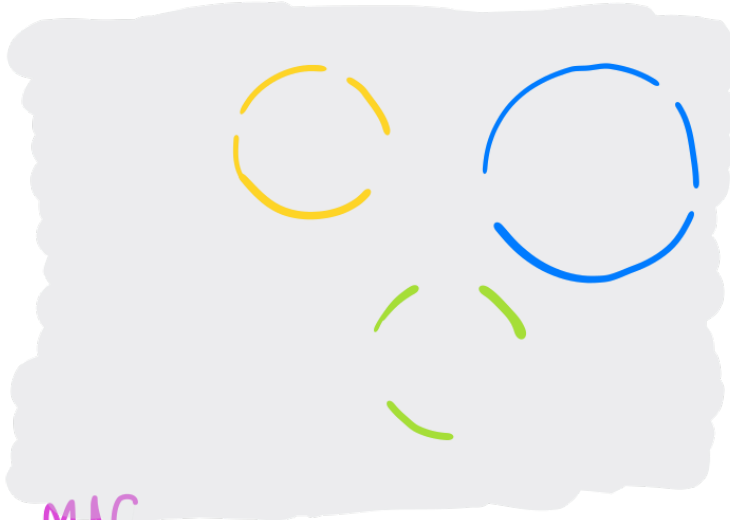
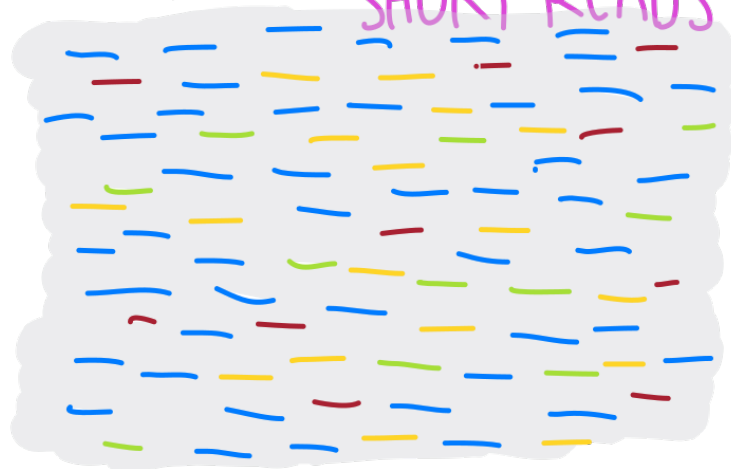
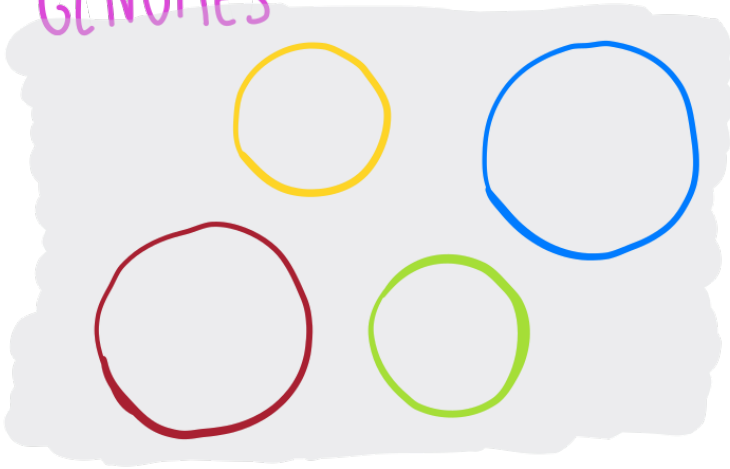
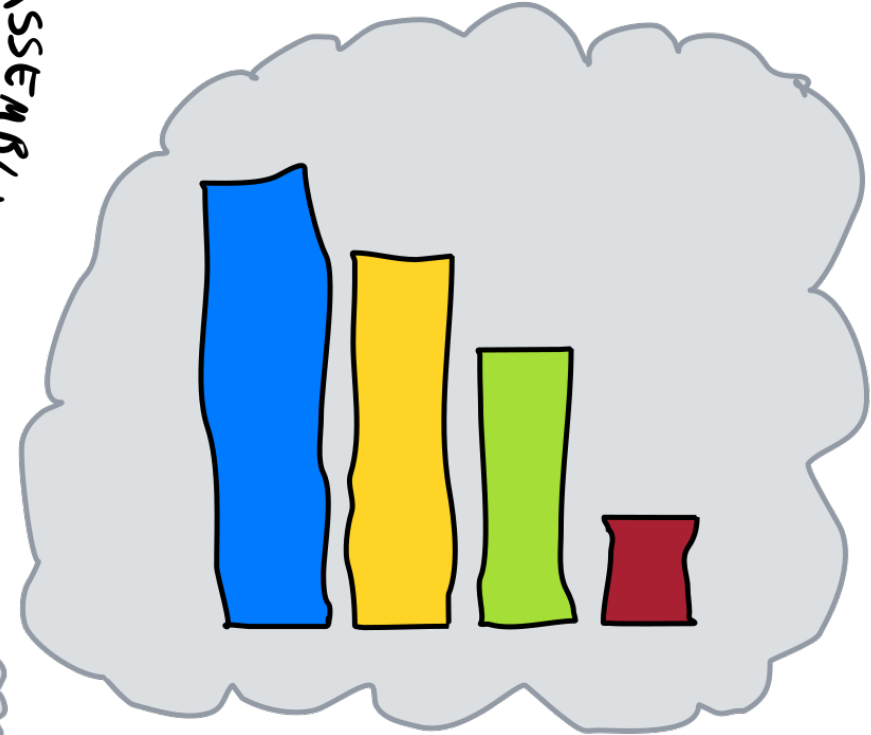
SHORT READS

DE-NOVO ASSEMBLY

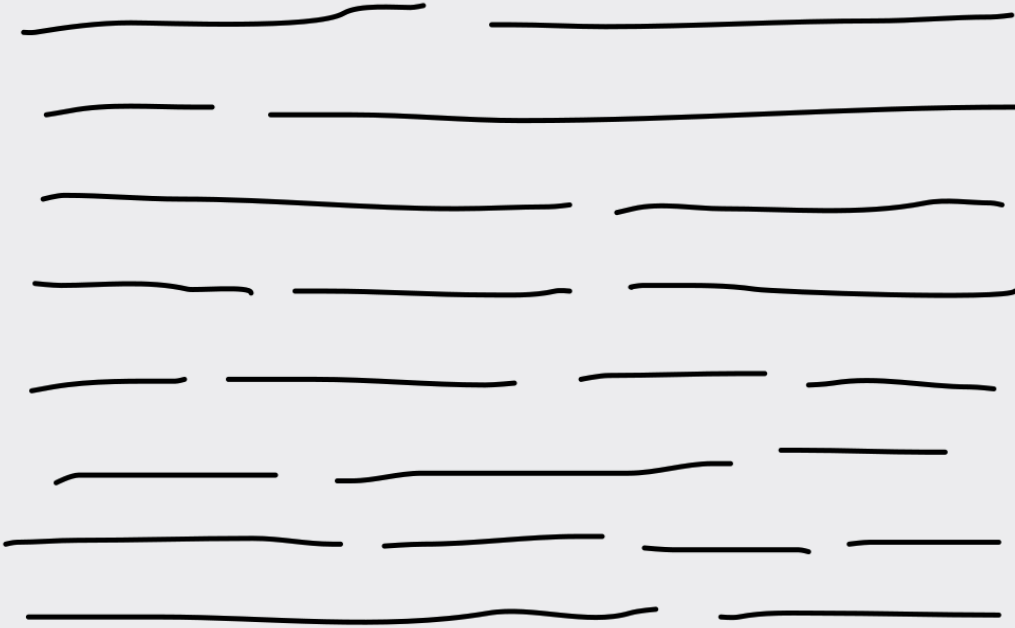
MAGs

CONTIGS

METAGENOMIC BINNING



CONTIGS




MAGs







Community structure and metabolism through reconstruction of microbial genomes from the environment

Gene W. Tyson, Jarrod Chapman, Philip Hugenholtz, Eric E. Allen, Rachna J. Ram, Paul M. Richardson, Victor V. Solovyev, Edward M. Rubin, Daniel S. Rokhsar & Jillian F. **Banfield** 

A new view of the tree of life

Laura A. Hug, Brett J. Baker, Karthik Anantharaman, Christopher T. Brown, Alexander J. Probst, Cindy J. Castelle, Cristina N. Butterfield, Alex W. Hermsdorf, Yuki Amano, Kotaro Ise, Yohey Suzuki, Natasha Dudek, David A. Relman, Kari M. Finstad, Ronald Amundson, Brian C. Thomas & Jillian F. **Banfield** 

Unusual biology across a group comprising more than 15% of domain Bacteria

Christopher T. Brown, Laura A. Hug, Brian C. Thomas, Itai Sharon, Cindy J. Castelle, Andrea Singh, Michael J. Wilkins, Kelly C. Wrighton, Kenneth H. Williams & Jillian F. **Banfield** 

Population Genomic Analysis of Strain Variation in *Leptospirillum* Group II Bacteria Involved in Acid Mine Drainage Formation

Sheri L Simmons , Genevieve DiBartolo , Vincent J Deneff, Daniela S. Aliaga Goltsman, Michael P Thelen, Jillian F **Banfield** 



Introduction to genome binning

> **Sequence composition**

Computing k-mer frequencies

Differential coverage

Completion and contamination

Enzymatic Synthesis of Deoxyribonucleic Acid

VIII. FREQUENCIES OF NEAREST NEIGHBOR BASE SEQUENCES IN DEOXYRIBONUCLEIC ACID

JOHN JOSSE,* A. D. KAISER, AND ARTHUR KORNBERG

From the Department of Biochemistry, Stanford University School of Medicine, Palo Alto, California

(Received for publication, October 4, 1960)

Determination of deoxyribonucleotide sequence in a deoxyribonucleic acid molecule is important from both the chemical and genetic points of view. It is also essential for answering the question of whether DNA synthesized *in vitro* by polymerase (1, 2) is a faithful copy of the nucleotide¹ sequence of the primer DNA. Although enzymatically synthesized DNA has the same over-all nucleotide composition as the particular primer DNA (3), it could not be inferred that this synthesis is a replication of the nucleotide sequences of the primer.

Because of the limitations of present methods, complete sequence studies have never been made. Sinheimer (4) has

(8). All of the labeled substrates contained P³² in the phosphate esterified to the sugar; they were prepared as described previously (1). The DNA-synthesizing enzyme was prepared from the polymerase, Fraction VII, described elsewhere (1); this enzyme was refractionated with diethylaminoethyl cellulose, yielding a preparation with a specific activity of 500 units per mg of protein. Micrococcal DNase was prepared according to Cunningham *et al.* (9); the final fraction had a specific activity of 7500 units per mg of protein.² Calf spleen phosphodiesterase was isolated by Hilmeo's procedure (10); the purified preparation had a specific activity of 82 units per mg of protein.

In the studies to be reported here, we have derived the frequencies of the 16 possible nearest neighbor pairs in a variety of DNA's by the technique of enzymatic incorporation of 5'-P³²-labeled nucleotides into DNA and then degradation of the DNA into 3'-nucleotides. Briefly, we have found that: (a) each DNA directs the synthesis of a product which has a unique and non-random pattern of the 16 nearest neighbor frequencies; (b) the DNA synthesized has the same nearest neighbor frequencies whether the primer is native DNA or enzymatically prepared DNA containing only traces of the original native DNA; and (c) the pattern of nearest neighbor frequencies in every case involves both base-pairing of adenine to thymine and of guanine to cytosine between sister strands of DNA, and opposite "polarity" of the two strands as proposed in the Watson and Crick model (7).

Genes from Nine Genomes Are Separated into Their Organisms in the Dinucleotide Composition Space

Hiroshi NAKASHIMA,^{1,*} Motonori OTA,² Ken NISHIKAWA,² and Tatsuo OOI³

School of Health Sciences, Faculty of Medicine, Kanazawa University, 5-11-80 Kodatsuno, Kanazawa 920-0942, Japan,¹ Center for Information Biology, National Institute of Genetics, Yata 1111, Mishima, Shizuoka 411-8540, Japan,² and Kyoto Women's University, Kitahiyoshi-cho 35, Higashiyama-ku, Kyoto 605, Japan³

(Received 2 September 1998)

Abstract

A set of 16 kinds of dinucleotide compositions was used to analyze the protein-encoding nucleotide sequences in nine complete genomes: *Escherichia coli*, *Haemophilus influenzae*, *Helicobacter pylori*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Synechocystis* sp., *Methanococcus jannaschii*, *Archaeoglobus fulgidus*, and *Saccharomyces cerevisiae*. The dinucleotide composition was significantly different between the organisms. The distribution of genes from an organism was clustered around its center in the dinucleotide composition space. The genes from closely related organisms such as Gram-negative bacteria, mycoplasma species and eukaryotes showed some overlap in the space. The genes from nine complete genomes together with those from human were discriminated into respective clusters with 80% accuracy using the dinucleotide composition alone. The composition data estimated from a whole genome was close to that obtained from genes, indicating that the characteristic feature of dinucleotides holds not only for protein coding regions but also noncoding regions. When a dendrogram was constructed from the disposition of the clusters in the dinucleotide space, it resembled the real phylogenetic tree. Thus, the distinct feature observed in the dinucleotide composition may reflect the phylogenetic relationship of organisms.

Key words: separation of genes; dinucleotide frequency; phylogenetic tree



Introduction to genome binning

Sequence composition

> Computing k-mer frequencies

Differential coverage

Completion and contamination

(a simple example with $k=2$)

GTTTTGGCATGATTAAGGAGTTTCTTTGTGCTTC

GTTTTGGCATGATTAAGGAGTTTTCTTTTGTGCTTC

k=2

GTTTTGGCATGATTAAGGAGTTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

k=2

GT TTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0

k=2

CTT TGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1

k=2

G **TT** GGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2

k=2

GTT **TT** GGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	3

k=2

GTTT **TG**GCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	3

k=2

GTTT **GG** CATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	3

k=2

GTTTTGGCCTGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	0	0	0	0	0	1	1	1	0	0	1	3

k=2

GTTTTG(CA)GATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	1	0	0	0	0	1	1	1	0	0	1	3

k=2

GTTTTGGG**AT**GATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	1	1	0	0	0	0	1	1	1	0	0	1	3

k=2

GTTTTGGCA**TG**ATTAAGGAGTTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	1	1	0	0	0	0	1	1	1	0	0	2	3

k=2

GTTTTGGCAT **GAT** TAAGGAGTTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	1	1	0	0	0	1	1	1	1	0	0	2	3

k=2

GTTTTGGCATCATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	2	1	0	0	0	1	1	1	1	0	0	2	3

k=2

GTTTTGGCATGATTAGGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	2	1	0	0	0	1	1	1	1	0	0	2	4

k=2

GTTTTGGCATGATTAAGGAGTTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
1	0	2	2	1	0	0	2	2	2	2	3	1	2	4	10

k=2

GTTTTGGCATGATTAAGGAGTTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
1	0	2	2	1	0	0	2	2	2	2	3	1	2	4	10



GAAGCACAAAAGAAACTCCTTAATCATGCCAAAAC

k=2

GTTTTGGCATGATTAAGGAGTTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
1	0	2	2	1	0	0	2	2	2	2	3	1	2	4	10

GAAGCACAAAAGAAACTCCTTAATCATGCCAAAAC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
10	3	2	2	4	2	0	2	2	2	0	0	1	2	1	1

k=2

GTTTTGGCATGATTAAGGAGTTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
1	0	2	2	1	0	0	2	2	2	2	3	1	2	4	10



GAAGCACAAAAGAAACTCCTTAATCATGCCAAAAC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
10	3	2	2	4	2	0	2	2	2	0	0	1	2	1	1

GTTTTGGCATGATTAAGGAGTTTTCTTTTGTGCTTC
GAAGCACAAAAGAAACTCCTTAATCATGCCAAAAC

k=2

GTTTTGGCATGATTAAGGAGTTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
1	0	2	2	1	0	0	2	2	2	2	3	1	2	4	10



GAAGCACAAAAGAAACTCCTTAATCATGCCAAAAC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
10	3	2	2	4	2	0	2	2	2	0	0	1	2	1	1

GTTTTGGCATGATTAAGGAGTTTTCTTTTGTGCTTC
GAAGCAGAAAAGAAACTCCTTAATCATGCCAAAAC

AA	AC	AG	GA	CA	CC	CG	GC	AT	TA

k=2

GTTTTGGCATGATTAAGGAGTTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
1	0	2	2	1	0	0	2	2	2	2	3	1	2	4	10

GAAGCACAAAAGAAACTCCTTAATCATGCCAAAAC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
10	3	2	2	4	2	0	2	2	2	0	0	1	2	1	1

GTTTTGGCATGATTAAGGAGTTTTCTTTTGTGCTTC
 GAAGCAGAAAAGAAACTCCTTAATCATGCCAAAAC

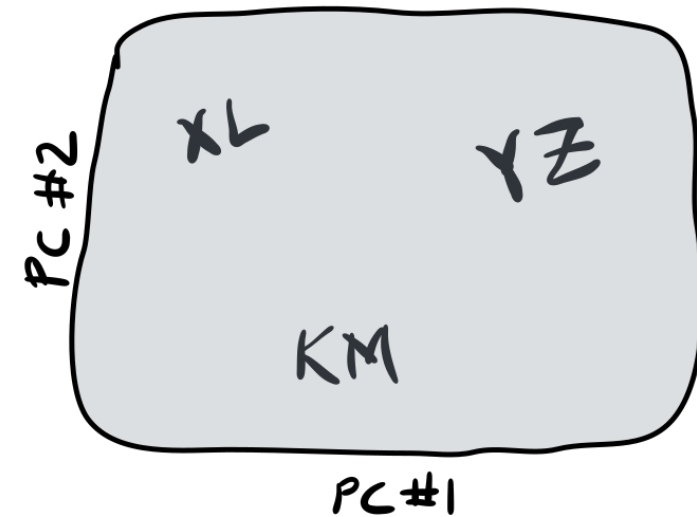
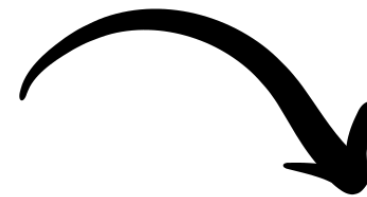
AA	AC	AG	GA	CA	CC	CG	GC	AT	TA
11	3	4	4	5	2	0	2	2	1

k=2

	AA	AC	AG	GA	CA	CC	CG	GC	AT	TA
X	11	3	4	4	5	2	0	2	2	1
Y	4	5	2	4	5	4	4	3	2	1
Z	4	5	3	2	4	1	5	5	2	3
L	11	6	3	2	2	3	2	1	1	4
K	1	1	2	2	1	8	9	10	0	0
M	0	4	4	3	4	10	4	5	0	0

k=2

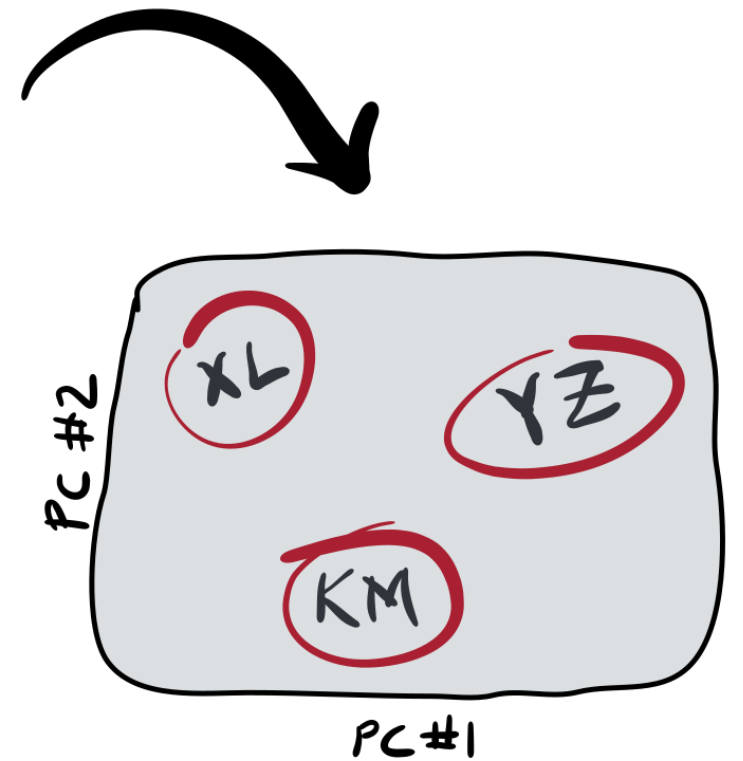
	AA	AC	AG	GA	CA	CC	CG	GC	AT	TA
X	11	3	4	4	5	2	0	2	2	1
Y	4	5	2	4	5	4	4	3	2	1
Z	4	5	3	2	4	1	5	5	2	3
L	11	6	3	2	2	3	2	1	1	4
K	1	1	2	2	1	8	9	10	0	0
M	0	4	4	3	4	10	4	5	0	0



k=2

	AA	AC	AG	GA	CA	CC	CG	GC	AT	TA
X	11	3	4	4	5	2	0	2	2	1
Y	4	5	2	4	5	4	4	3	2	1
Z	4	5	3	2	4	1	5	5	2	3
L	11	6	3	2	2	3	2	1	1	4
K	1	1	2	2	1	8	9	10	0	0
M	0	4	4	3	4	10	4	5	0	0

k=2



Genes from Nine Genomes Are Separated into Their Organisms in the Dinucleotide Composition Space

Hiroshi NAKASHIMA,^{1,*} Motonori OTA,² Ken NISHIKAWA,² and Tatsuo OOI³

School of Health Sciences, Faculty of Medicine, Kanazawa University, 5-11-80 Kodatsuno, Kanazawa 920-0942, Japan,¹ Center for Information Biology, National Institute of Genetics, Yata 1111, Mishima, Shizuoka 411-8540, Japan,² and Kyoto Women's University, Kitahiyoshi-cho 35, Higashiyama-ku, Kyoto 605, Japan³

(Received 2 September 1998)

Abstract

A set of 16 kinds of dinucleotide compositions was used to analyze the protein-encoding nucleotide sequences in nine complete genomes: *Escherichia coli*, *Haemophilus influenzae*, *Helicobacter pylori*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Synechocystis* sp., *Methanococcus jannaschii*, *Archaeoglobus fulgidus*, and *Saccharomyces cerevisiae*. The dinucleotide composition was significantly different between the organisms. The distribution of genes from an organism was clustered around its center in the dinucleotide composition space. The genes from closely related organisms such as Gram-negative bacteria, mycoplasma species and eukaryotes showed some overlap in the space. The genes from nine complete genomes together with those from human were discriminated into respective clusters with 80% accuracy using the dinucleotide composition alone. The composition data estimated from a whole genome was close to that obtained from genes, indicating that the characteristic feature of dinucleotides holds not only for protein coding regions but also noncoding regions. When a dendrogram was constructed from the disposition of the clusters in the dinucleotide space, it resembled the real phylogenetic tree. Thus, the distinct feature observed in the dinucleotide composition may reflect the phylogenetic relationship of organisms.

Key words: separation of genes; dinucleotide frequency; phylogenetic tree

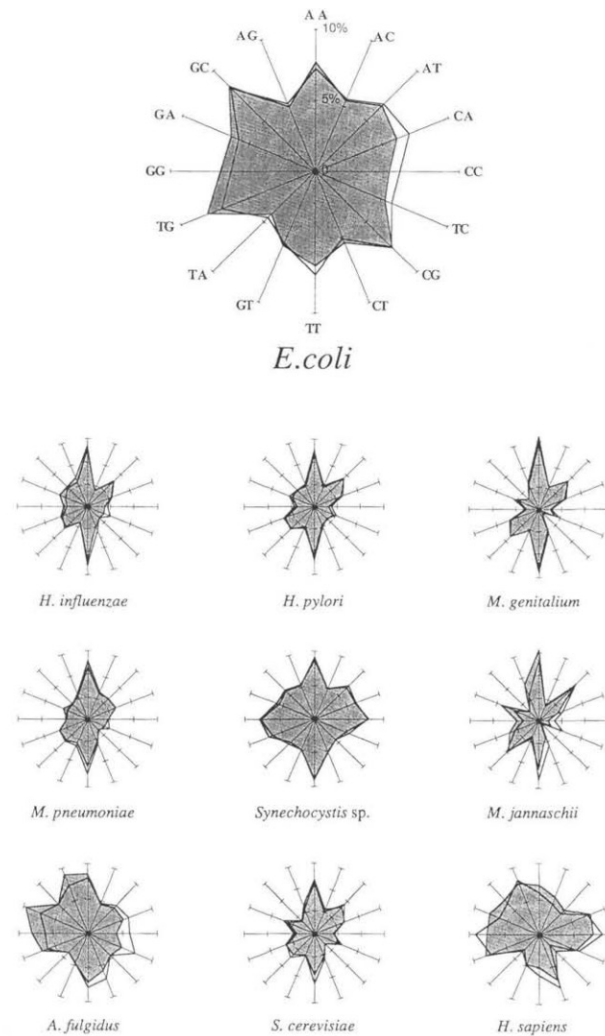


Figure 1. Star-diagrams presenting the dinucleotide composition. The mean compositions (%) over all genes (shaded) and the entire genome (non-shaded) are plotted. The data for human are exceptions (see Sequence data section), depicted here as references. The radial axes of 16 dinucleotides are allotted so that the complementary dinucleotides, AA/TT, AC/GT, etc., occupy counter positions along the circle. Complementary pairs should have equivalent amounts in the total composition over a genome. Note that the scale is different for each diagram. The innermost broken circle indicates the 5% level.

Peter A. Noble¹
Robert W. Citek²
Oladele A. Ogunseitan³

¹**Belle W. Baruch Institute for
Marine Biology and Coastal
Research, University of South
Carolina, Columbia, SC, USA**

²**Department of Soil and
Environmental Science, University
of California at Riverside,
Riverside, CA, USA**

³**Department of Environmental
Analysis and Design, University
of California at Irvine,
Irvine, CA, USA**

Tetranucleotide frequencies in microbial genomes

A computational strategy for determining the variability of long DNA sequences in microbial genomes is described. Composite portraits of bacterial genomes were obtained by computing tetranucleotide frequencies of sections of genomic DNA, converting the frequencies to color images and arranging the images according to their genetic position. The resulting images revealed that the tetranucleotide frequencies of genomic DNA sequences are highly conserved. Sections that were visibly different from those of the rest of the genome contained ribosomal RNA, bacteriophage, or undefined coding regions and had corresponding differences in the variances of tetranucleotide frequencies and GC content. Comparison of nine completely sequenced bacterial genomes showed that there was a nonlinear relationship between variances of the tetranucleotide frequencies and GC content, with the highest variances occurring in DNA sequences with low GC contents (less than 0.30 mol). High variances were also observed in DNA sequences having high GC contents (greater than 0.60 mol), but to a much lesser extent than DNA sequences having low GC contents. Differences in the tetranucleotide frequencies may be due to the mechanisms of intercellular genetic exchange and/or processes involved in maintaining intracellular genetic stability. Identification of sections that were different from those of the rest of the genome may provide information on the evolution and plasticity of bacterial genomes.

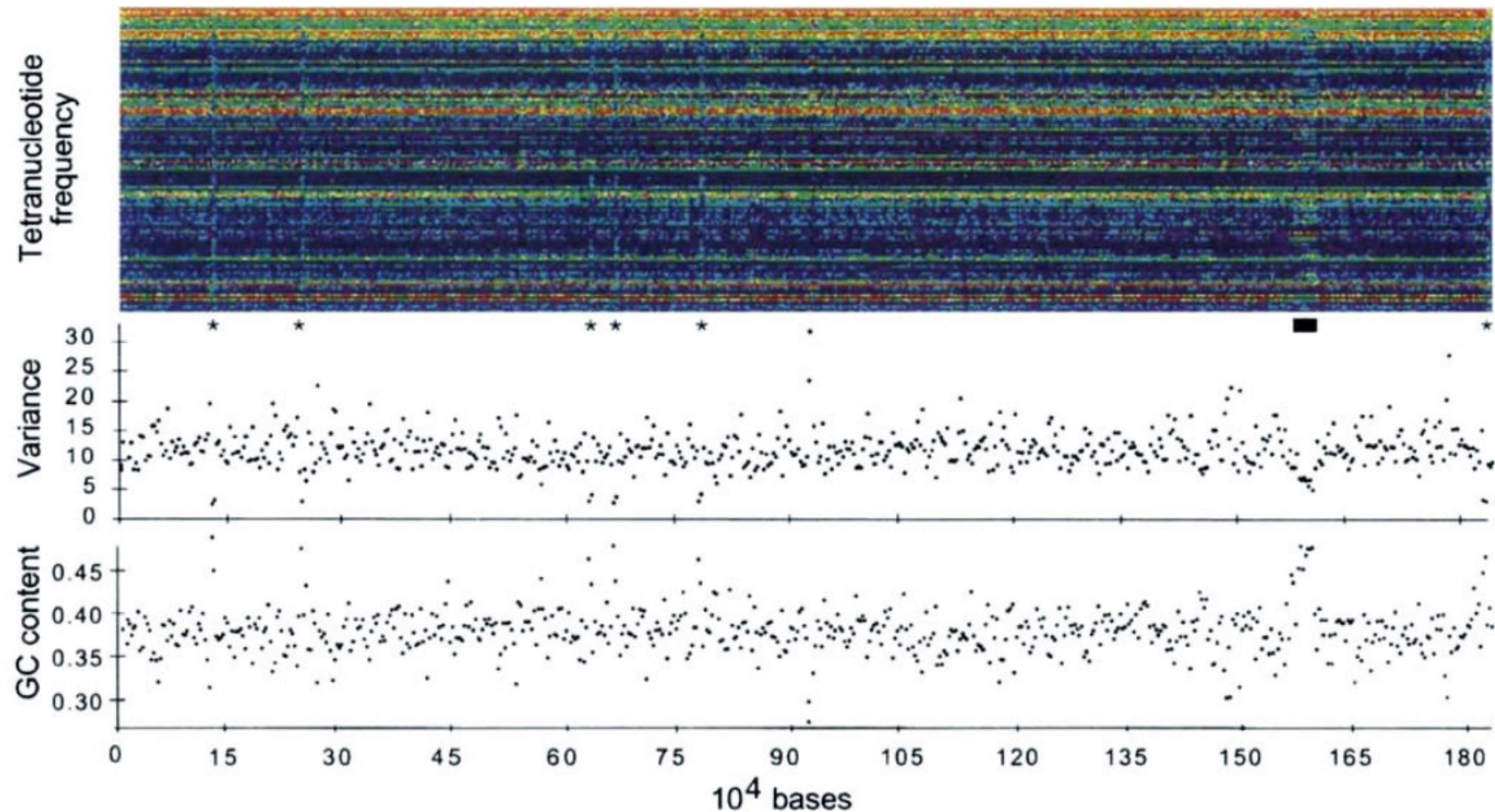


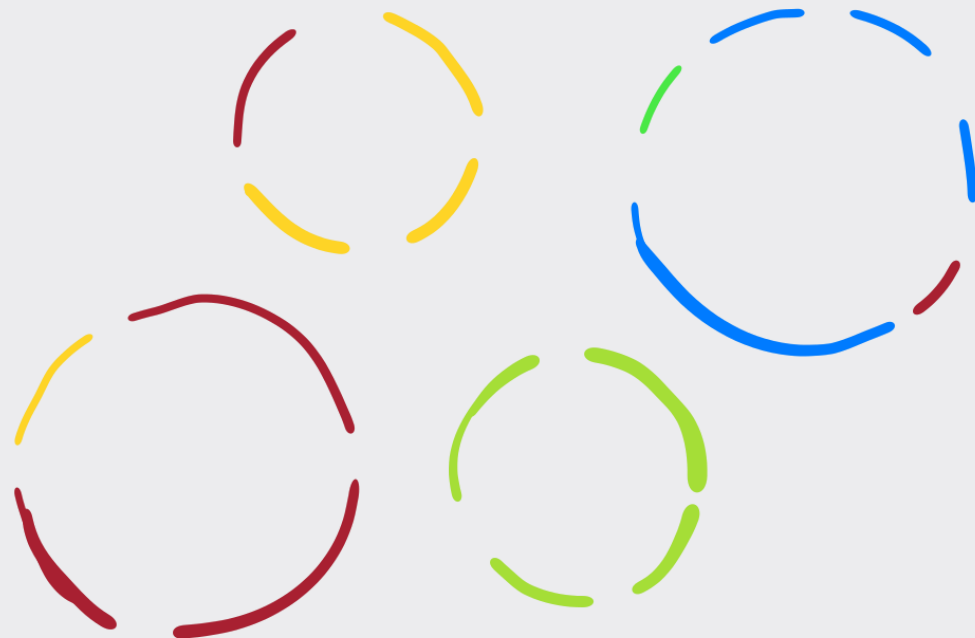
Figure 1. Fingerprints, variances of tetranucleotide frequencies, and GC values of sections of the *Haemophilus influenzae* Rd genome are consecutively ordered from the *NotI* restriction site [9]. Each column of the color image represents the fingerprint obtained from the analysis of one DNA sequence (*i.e.*, a 3000 bp section). Each row represents the frequency of a specific tetranucleotide and its complement. Tetranucleotides are arranged alphabetically on the *y*-axis. Each tetranucleotide is represented by a box, whose color is determined by its frequency, ranging from purple (low) to red (high). A star (*) identifies sections containing ribosomal RNA. The black bar identifies the location of the cryptic Mu-like bacteriophage. The variance and GC values were computed from the analysis of one section.

SEQUENCE COMPOSITION

CONTIGS



MAGs





Introduction to genome binning

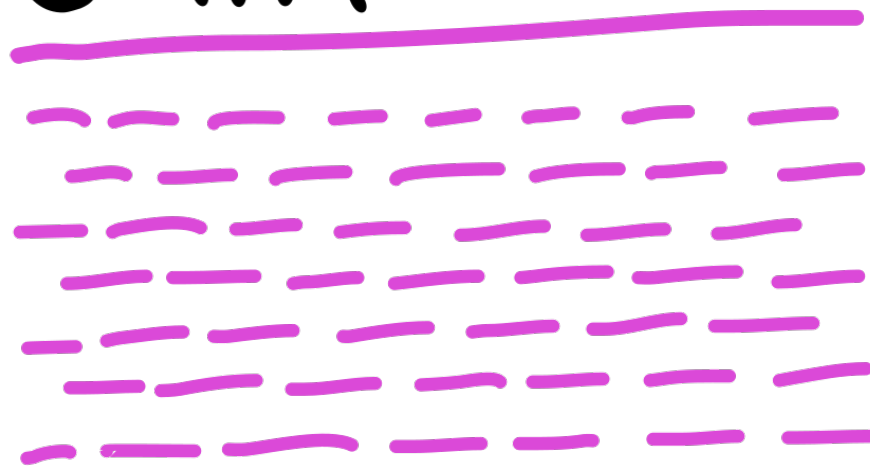
Sequence composition

Computing k-mer frequencies

> Differential coverage

Completion and contamination

CONTIG #1

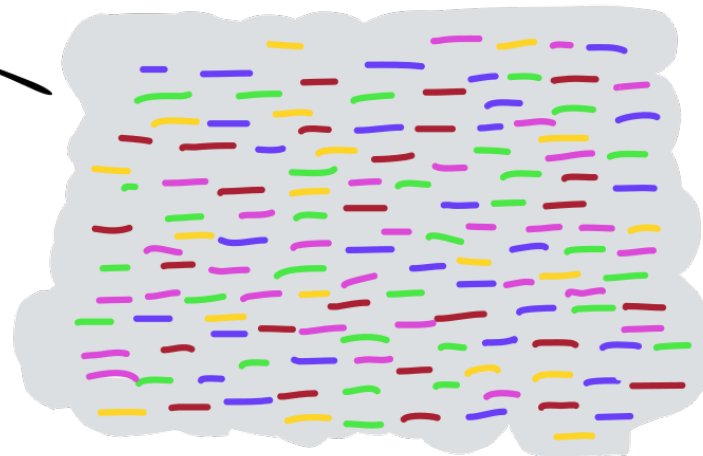


COVERAGE: ~7X

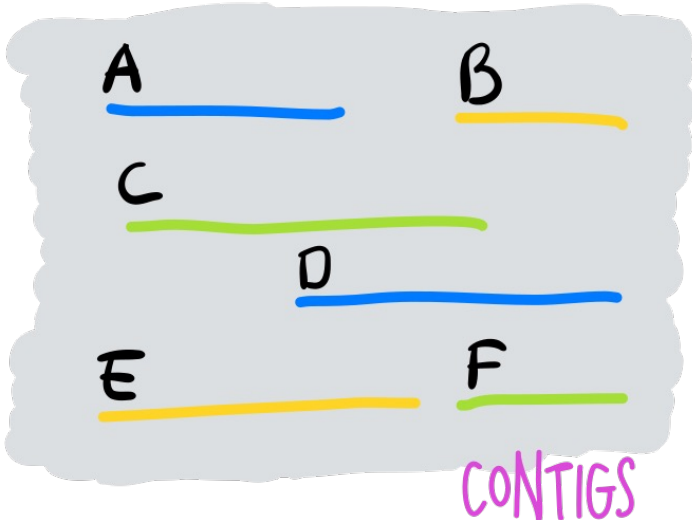
CONTIG #2



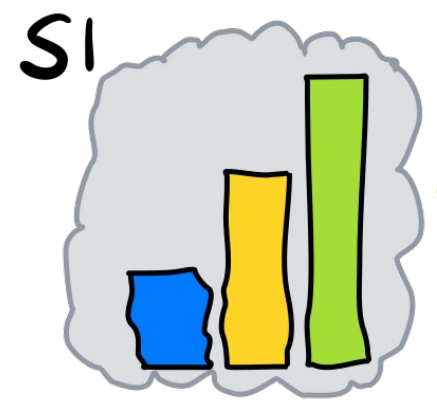
COVERAGE: ~4X



METAGENOMIC READS



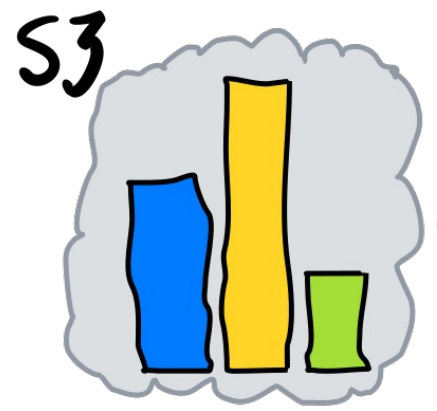
CONTIGS



- A - 1X
- B - 3X
- C - 5X
- D - 1X
- E - 3X
- F - 5X

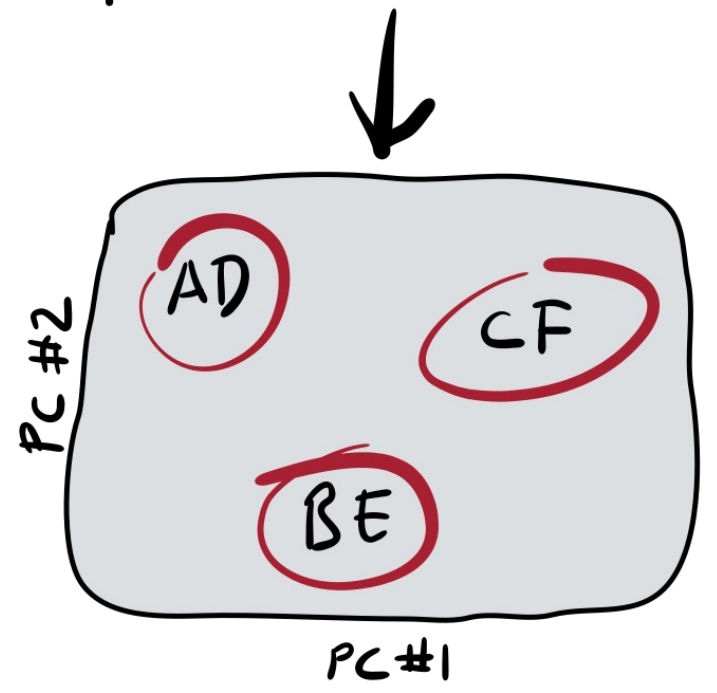


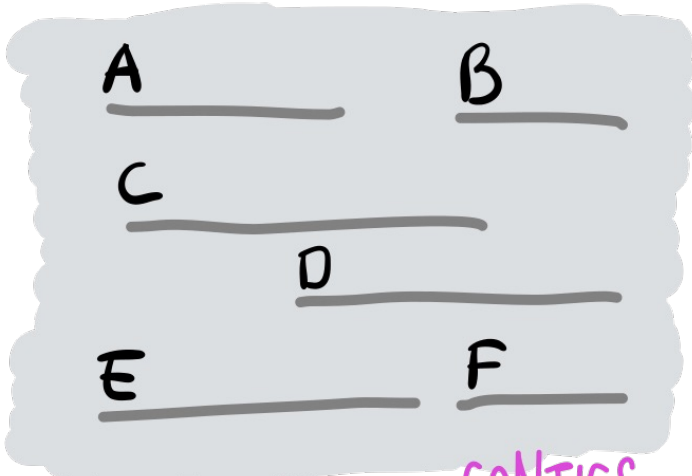
- A - 5X
- B - 1X
- C - 3X
- D - 5X
- E - 1X
- F - 3X



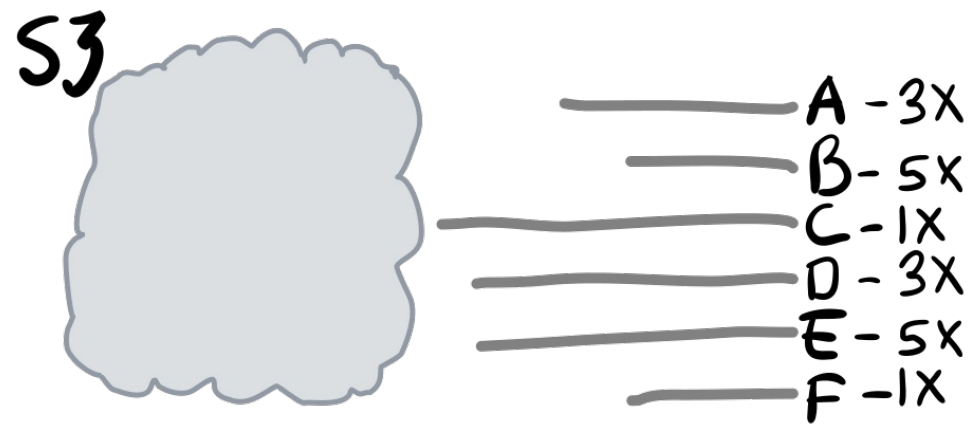
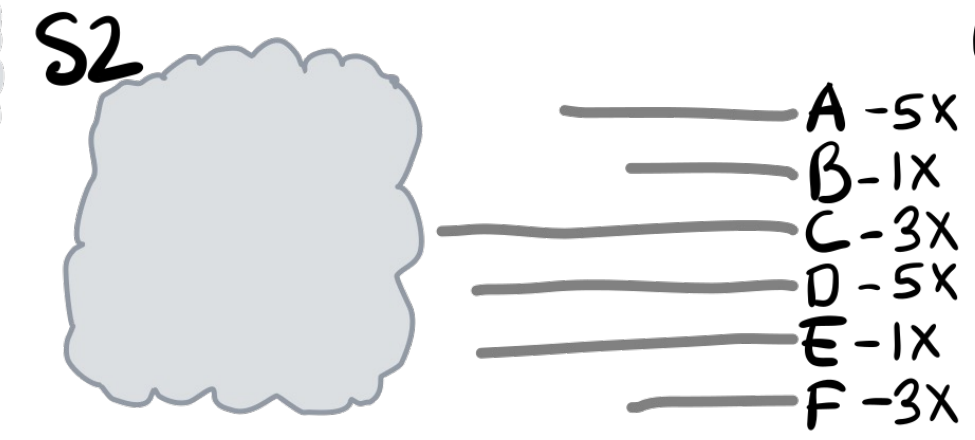
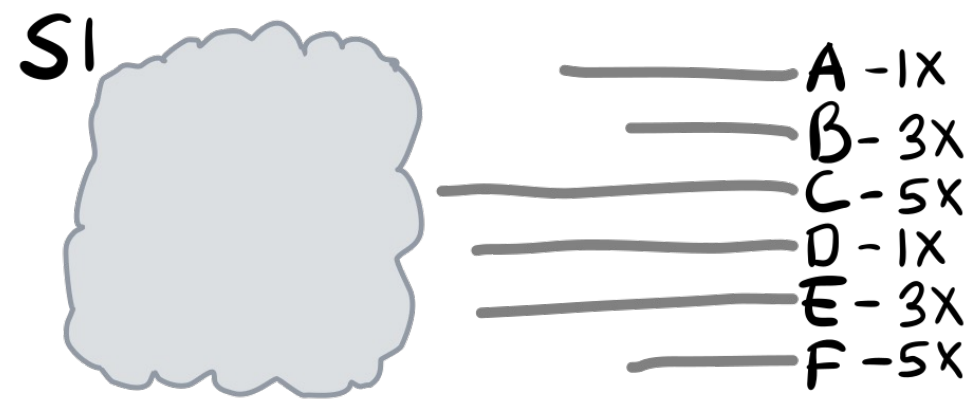
- A - 3X
- B - 5X
- C - 1X
- D - 3X
- E - 5X
- F - 1X

	A	B	C	D	E	F
S1	1	3	5	1	3	5
S2	5	1	3	5	1	3
S3	3	5	1	3	5	1

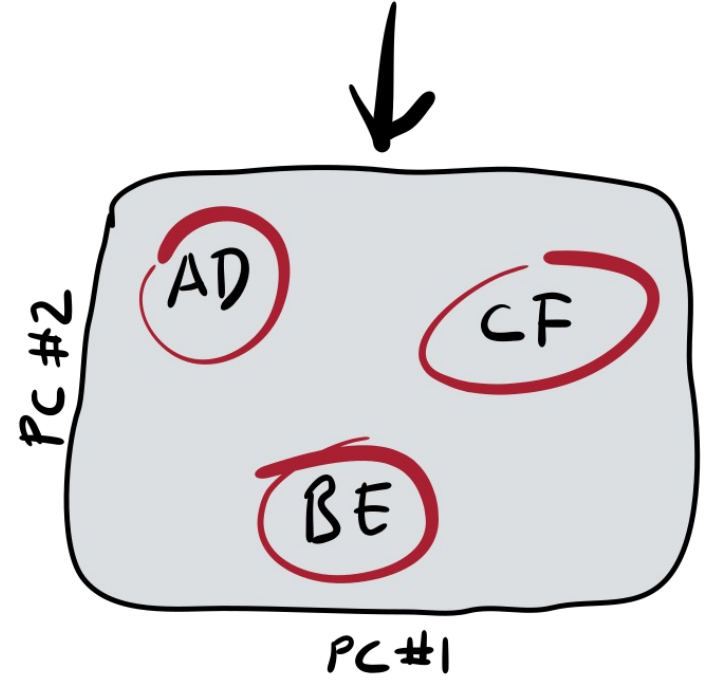




CONTIGS

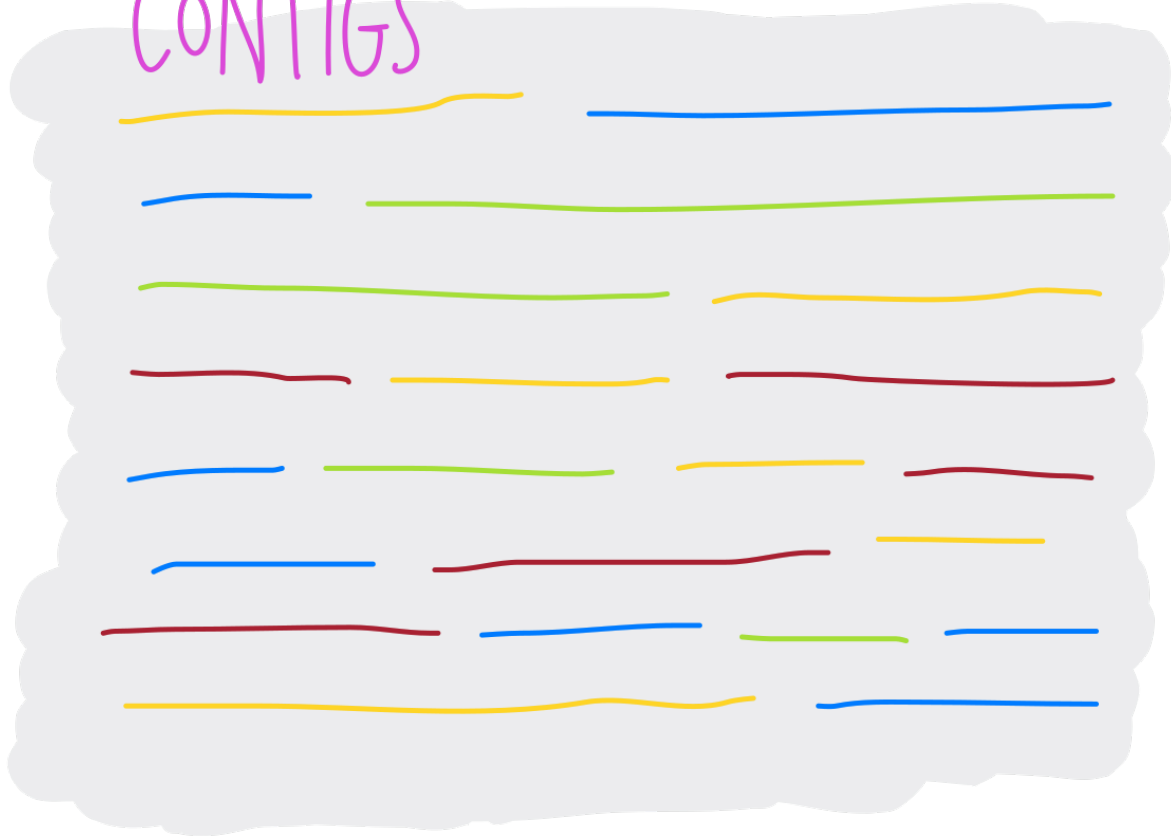


	A	B	C	D	E	F
S1	1	3	5	1	3	5
S2	5	1	3	5	1	3
S3	3	5	1	3	5	1

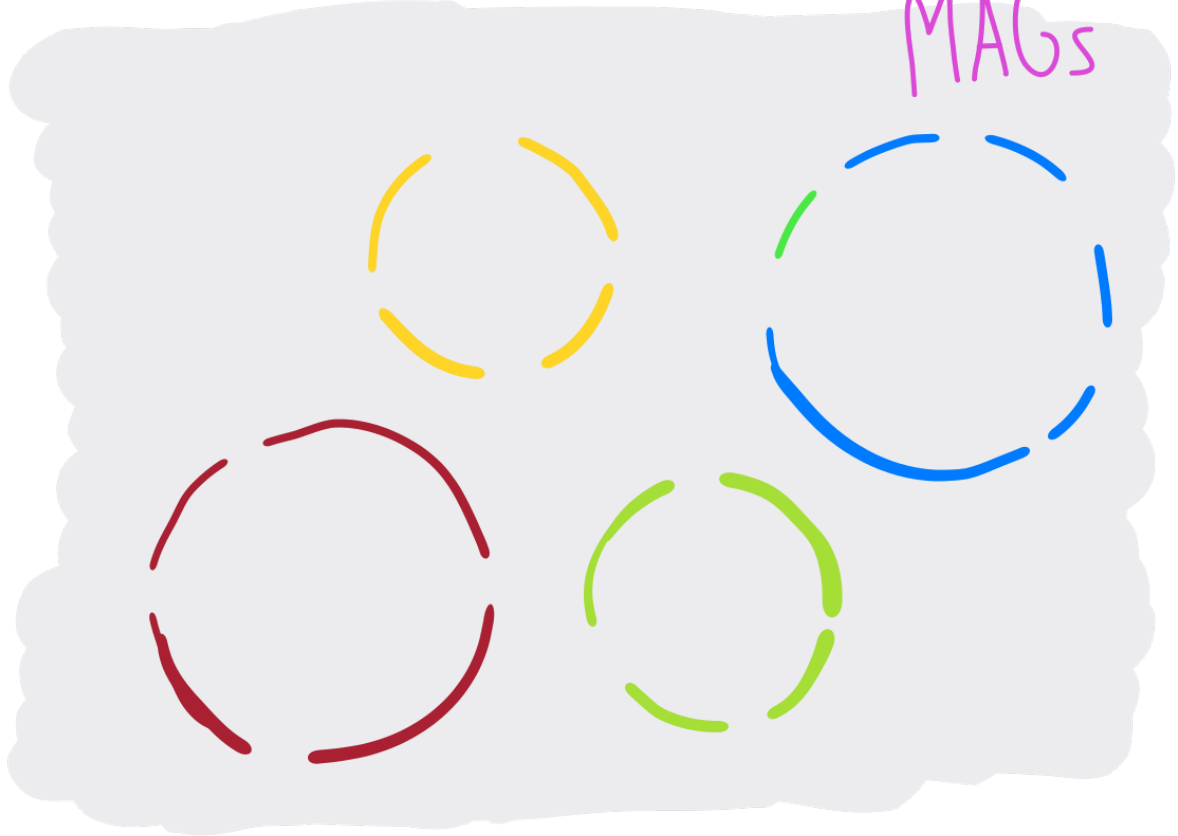


SEQUENCE COMPOSITION

CONTIGS



MAGs



DIFFERENTIAL COVERAGE



Introduction to genome binning

Sequence composition

Computing k-mer frequencies

Differential coverage

> Completion and contamination

Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea

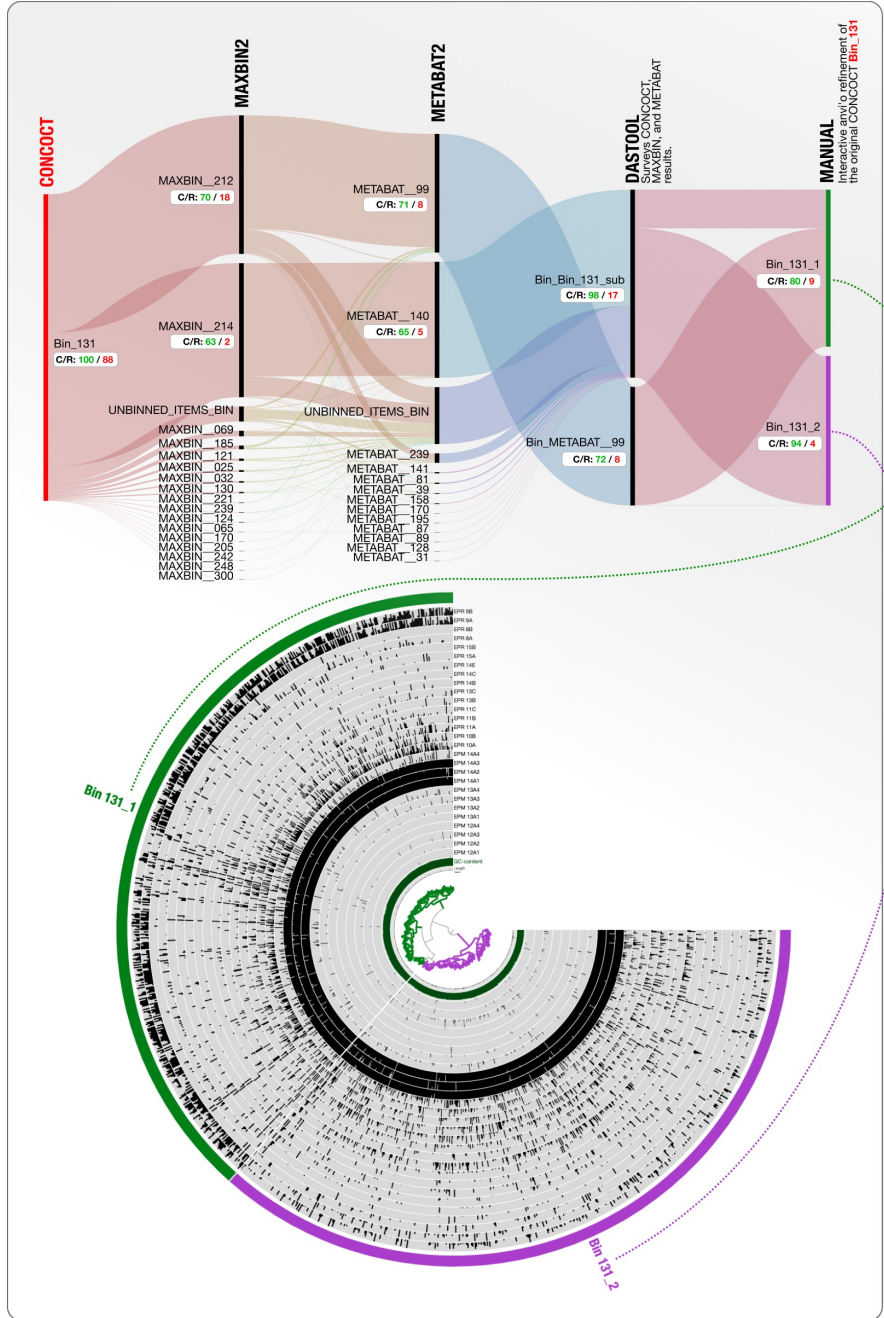
Robert M Bowers , Nikos C Kyrpides, Ramunas Stepanauskas, Miranda Harmon-Smith, Devin Doud, T B K Reddy, Frederik Schulz, Jessica Jarett, Adam R Rivers, Emiley A Eloë-Fadrosh, Susannah G Tringe, Natalia N Ivanova, Alex Copeland, Alicia Clum, Eric D Becraft, Rex R Malmstrom, Bruce Birren, Mircea Podar, Peer Bork, George M Weinstock, George M Garrity, Jeremy A Dodsworth, Shibu Yooseph, Granger Sutton, Frank O Glöckner, Jack A Gilbert, William C Nelson, Steven J Hallam, Sean P Jungbluth, Thijs J G Ettema, Scott Tighe, Konstantinos T Konstantinidis, Wen-Tso Liu, Brett J Baker, Thomas Rattei, Jonathan A Eisen, Brian Hedlund, Katherine D McMahon, Noah Fierer, Rob Knight, Rob Finn, Guy Cochrane, Ilene Karsch-Mizrachi, Gene W Tyson, Christian Rinke, The Genome Standards Consortium, Alla Lapidus, Folker Meyer, Pelin Yilmaz, Donovan H Parks, A Murat Eren, Lynn Schriml, Jillian F Banfield, Philip Hugenholtz & Tanja Woyke 

Accurate and complete genomes from metagenomes

Lin-Xing Chen,¹ Karthik Anantharaman,^{1,7} Alon Shaiber,^{2,3} A. Murat Eren,^{3,4}
and Jillian F. Banfield^{1,5,6}

¹Department of Earth and Planetary Sciences, University of California, Berkeley, California 94720, USA; ²Graduate Program in Biophysical Sciences, University of Chicago, Chicago, Illinois 60637, USA; ³Department of Medicine, University of Chicago, Chicago, Illinois 60637, USA; ⁴Bay Paul Center, Marine Biological Laboratory, Woods Hole, Massachusetts 02543, USA; ⁵Department of Environmental Science, Policy, and Management, University of California, Berkeley, California 94720, USA; ⁶Earth and Environmental Sciences, Lawrence Berkeley National Laboratory, University of California, Berkeley, California 94720, USA

Genomes are an integral component of the biological information about an organism; thus, the more complete the genome, the more informative it is. Historically, bacterial and archaeal genomes were reconstructed from pure (monoclonal) cultures, and the first reported sequences were manually curated to completion. However, the bottleneck imposed by the requirement for isolates precluded genomic insights for the vast majority of microbial life. Shotgun sequencing of microbial communities, referred to initially as community genomics and subsequently as genome-resolved metagenomics, can circumvent this limitation by obtaining metagenome-assembled genomes (MAGs); but gaps, local assembly errors, chimeras, and contamination by fragments from other genomes limit the value of these genomes. Here, we discuss genome curation to improve and, in some cases, achieve complete (circularized, no gaps) MAGs (CMAGs). To date, few CMAGs have been



Visualizing the fate of contigs across metagenomic binning algorithms

a post by **A. Murat Eren (Meren)**

[Web](#)
[Email](#)
[Twitter](#)
[LinkedIn](#)
[Github](#)

[ORCID](#)

and **Jarrold J. Scott**

[Web](#)
[Email](#)

Visualizing contig coverages to better understand microbial population structures

a post by **Emily Fogarty**

[Email](#)
[Github](#)
[ORCID](#)

and **Ryan Moore**

[Twitter](#)
[LinkedIn](#)
[Github](#)

