Metagenomics 101

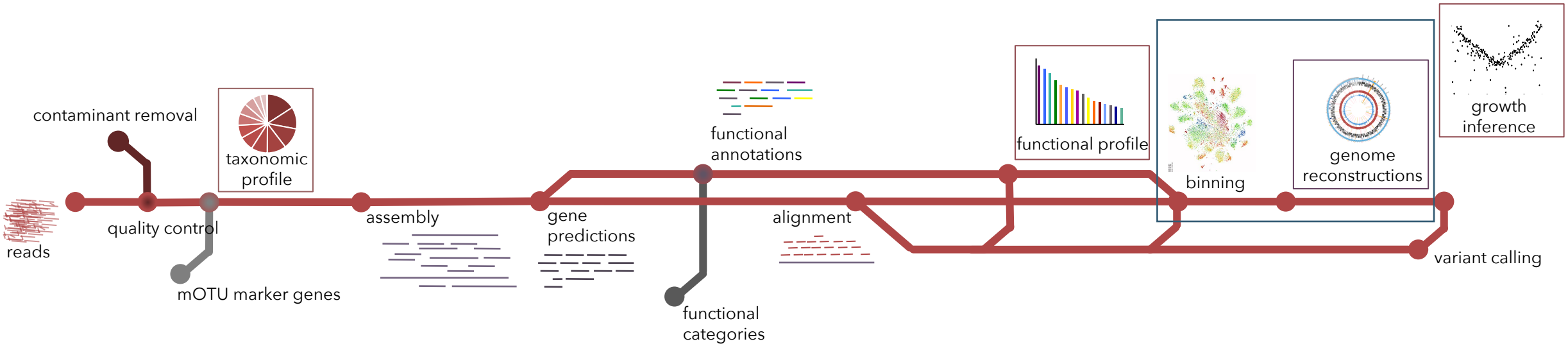
Session 8: Genome reconstruction

Anna Heintz-Buschart

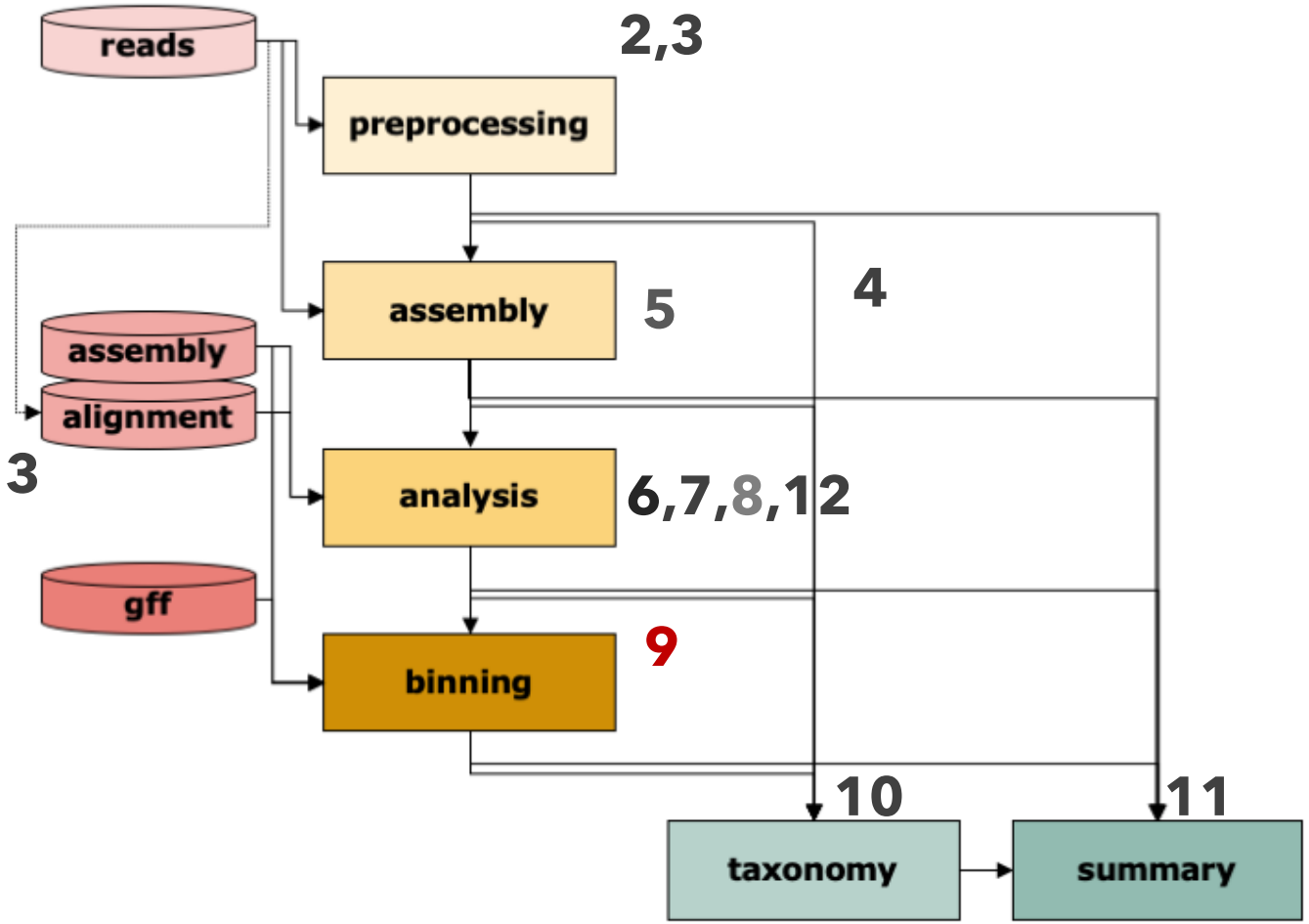
May 2022

University of Amsterdam
Life Sciences

Metagenomics (+ other omics) pipeline



Metagenomics (+ other omics) pipeline



Today

- what's the aim and why is this a problem?
- common features of genomes
- algorithms / approaches
- refinement and quality control

Aim

- point of reference in the absence of (suitable) cultured isolates:

- characterisation of un-cultured microbial taxa
- resource for short-read annotation
- pangenomic potential
- anchor for the integration of functional meta-omics

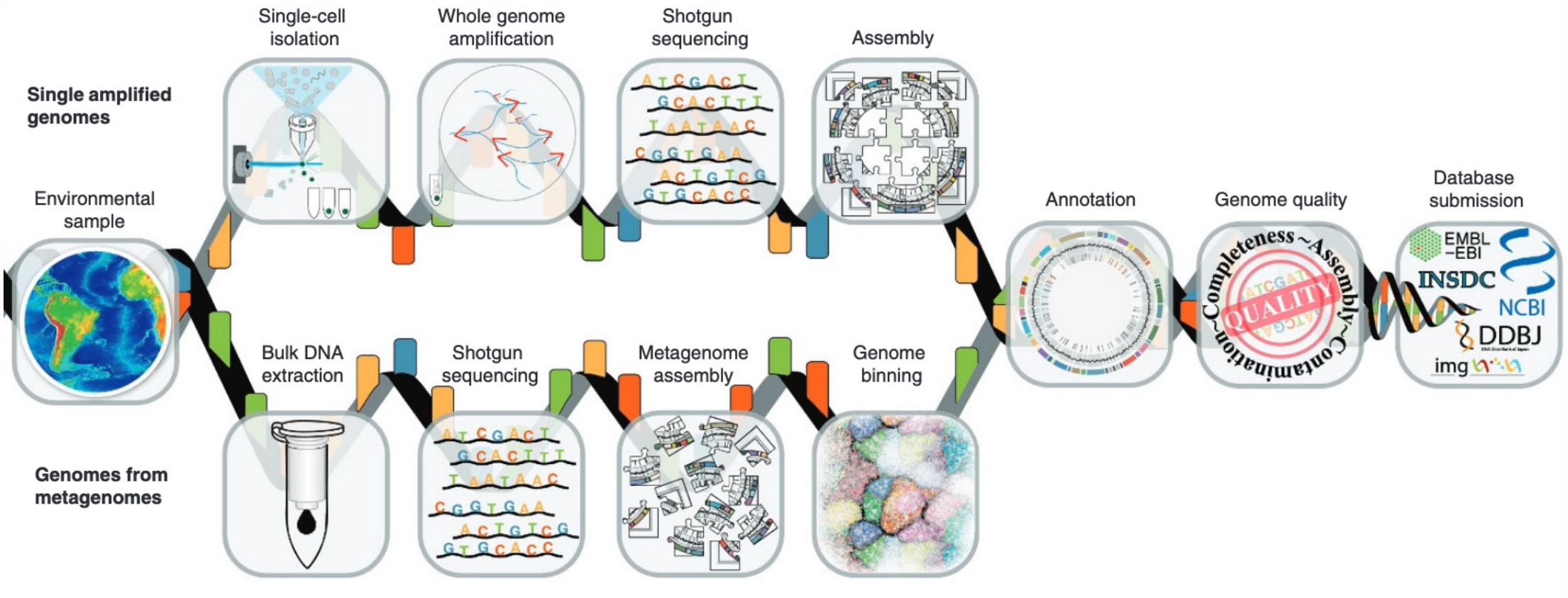
Terminology

- MAG: metagenome-assembled genome
- as opposed to SAGs (single-cell sequencing based genomes) and classical isolate-based genomes

- genome reconstructions

- bins
- binning: putting similar items in a group
  - also used in the context of taxonomic profiling

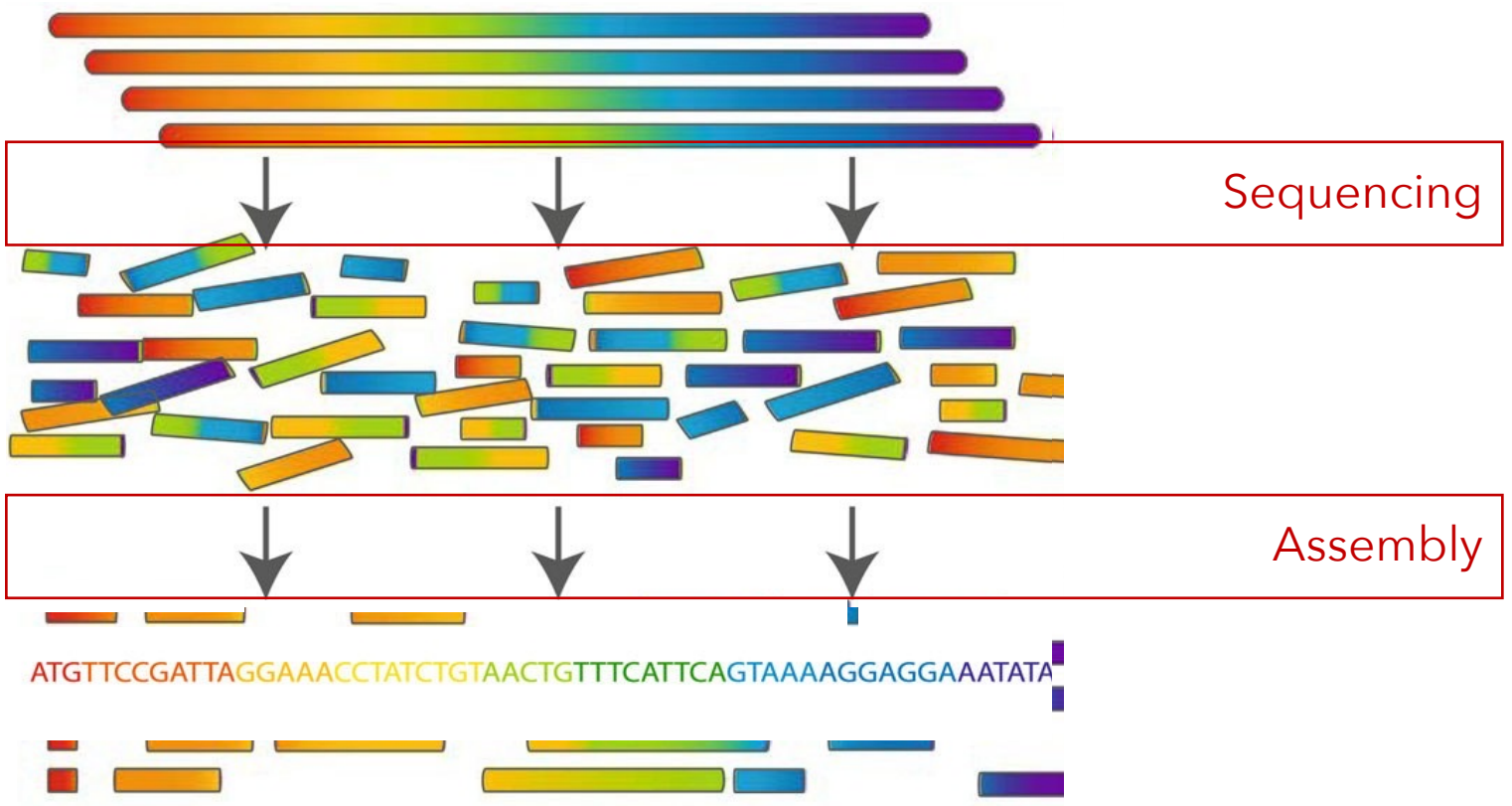
Workflow



Bowers et al. (2017) Nat Biotechnol 35, 725–731. https://doi.org/10.1038/nbt.3893

Recap assembly

- puzzling sequencing reads back together



adapted from Commins et al. 2009 Biol Proced Online 11: 52

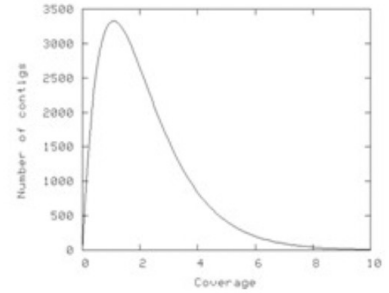
Recap assembly

- number of contigs representing a genome depends on:
- number of reads derived from this genome

Lander-Waterman statistics

L = read length
T = minimum detectable overlap
G = genome size
N = number of reads
c = coverage (NL / G)
$\sigma = 1 - T/L$

$E(\#islands) = Ne^{-c\sigma}$
$E(island\ size) = L((e^{c\sigma} - 1) / c + 1 - \sigma)$
contig = island with 2 or more reads



14

Pop, Delcher & Salzberg https://slideplayer.com/slide/4410317/

Recap assembly

- number of contigs representing a genome depends on:

- number of reads derived from this genome:
  - sequencing depth
  - diversity

- presence of difficult sequences:
  - high/low GC
  - repeats / low-complexity regions
  - high similarity to other genomes (incl. phages)

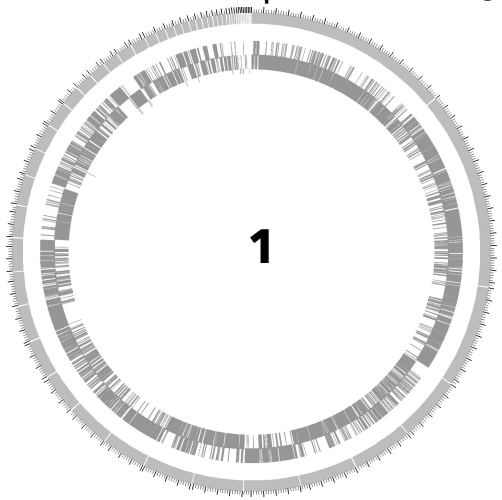
Group contigs, but based on what?

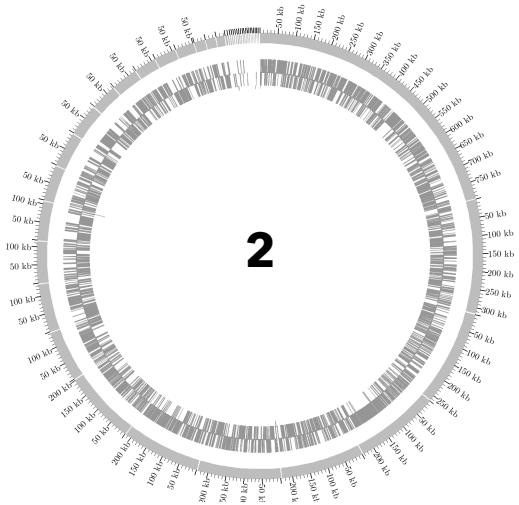
Group contigs, but based on what?

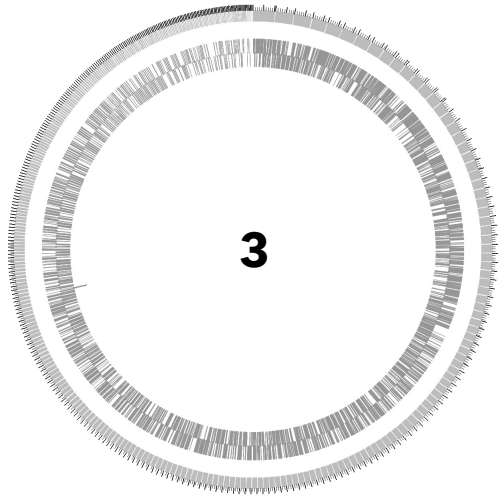
AA AC AG AT CA CC CG GA GC TA

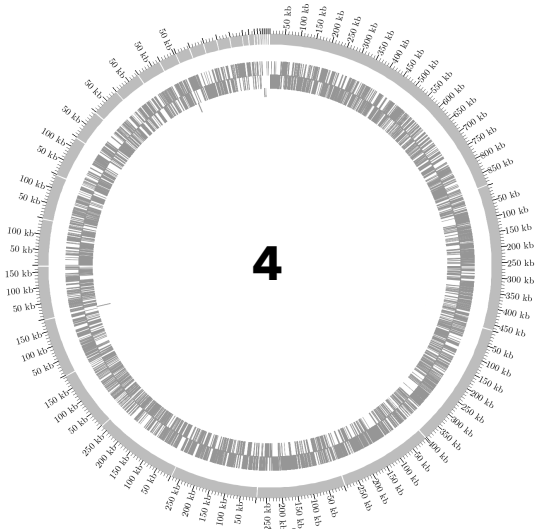
Paenibacillus sp.

Brevibacterium casei



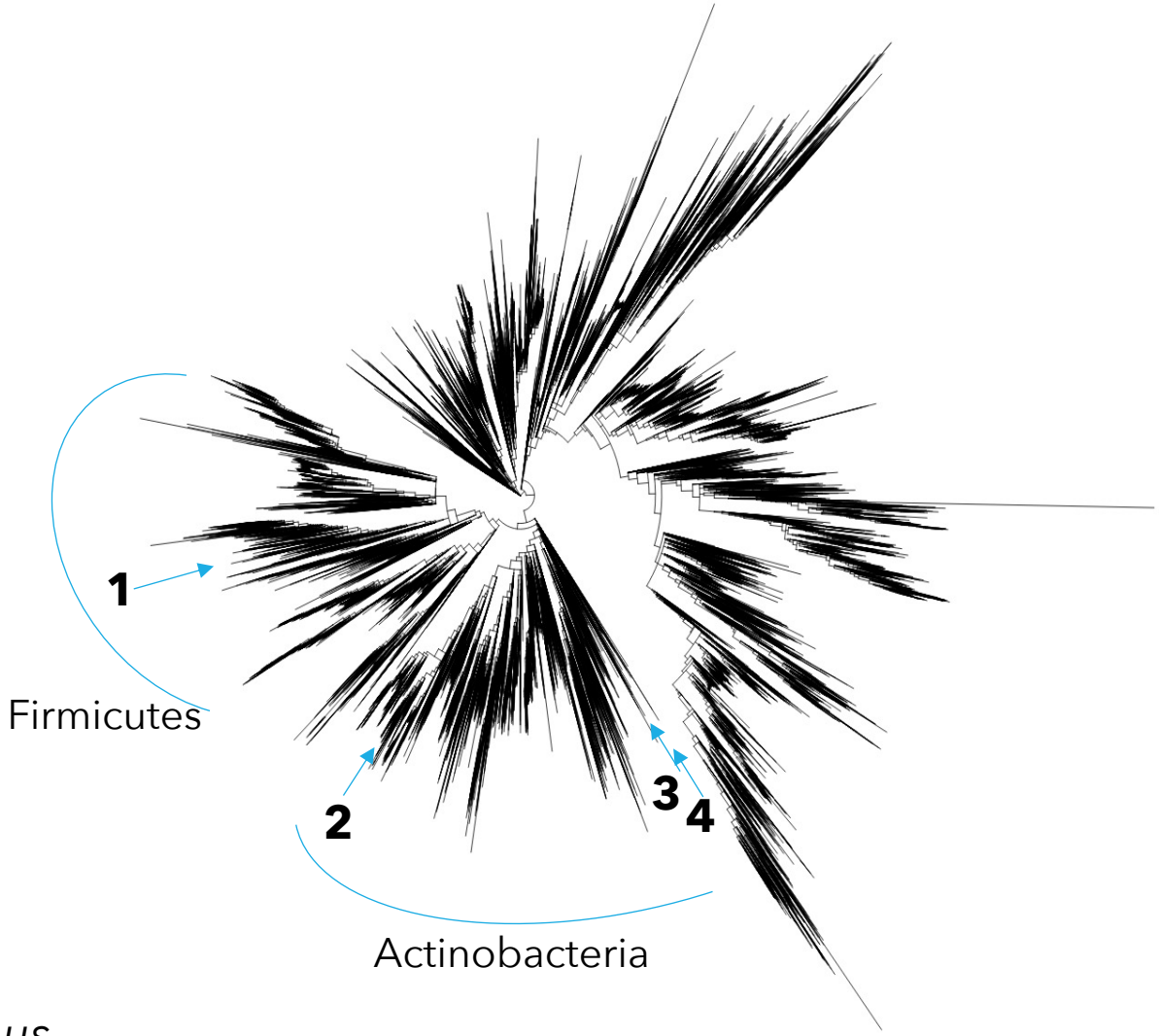






Patulibacter medicamentivorans

Patulibacter americanus



*k*Mer profile

	AA	AC	AG	AT	CA	CC	CG	GA	GC	TA
17233	692	669	823	336	821	938	773	1010	756	167

[['M1_2_V1_contig_17233',
'TACCGAGGCATATCAAGCCGCTATTGCAGCCTGATGAATGGCGACCAGGAAGGGGTTCGAACCC
TCGACCTCCGGCGTGACAGGCCGGCGTTCTAACCAGCTGAACTACCTGGCCGAATTTATGGTGGG
AACAACAGGGCTCGAACCTGTGACCCTCTGCTTGTAAGGCAGATGCTCTCCCAGCTGAGCTATGC
TCCCCACTCGAAATAATCGCTTCGGCGAACGACAAGGGTTATTATACAGAATAACCCCTGCCGTG
TCAACCTTATTTTTTGATTTTTTCAAGAAATTTTTTGAAAAGGGAAATGCATTACTTTCCCTGCT
TTTTCAGATCGGCGATCATGGCGGTCAGATCCTCTTTGGTAAAGTTCTGCACCCTGTCACAGAAC
TGGCAGGTGAGCTCAGCAGAGCCCTGCTCGTCCACGATCTTTTCAAGCTCCTTTGACCCCAGCGA
CAGCAGCGCGCGCTCCGTGCGCTCGCGCGAGCAGTAGCAGCGGTATTCGATCGGATCGACGGAGA
GGATCTCCATGTCAAAATCAGACAGCACCGTTTTGAGCAGCACCGCAGGATCAGGATTCTCCTTT
AAGAGATTCGTCACGCTGGGAGCTGCGTAGATGCCGCCCTCAACCTTGGTGATGACATCCTCACC
CGCGCCGGGGAGGAGCTGGATGAGATAGCCGCCCGCGGTGAGCACGCTGCGGTCGCGGTCGATGA

*k*Mer profile

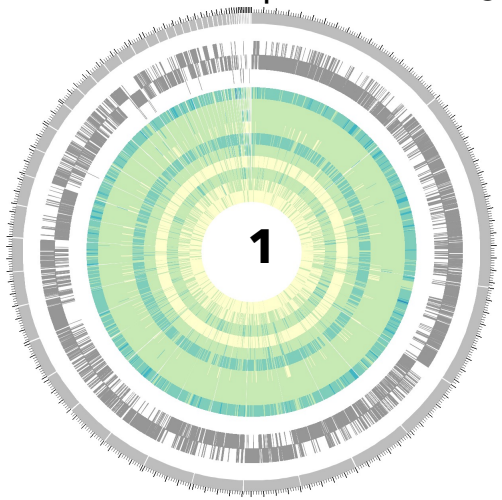
	AA	AC	AG	AT	CA	CC	CG	GA	GC	TA
17233	692	669	823	336	821	938	773	1010	756	167
35980	931	592	796	411	734	839	765	991	739	243

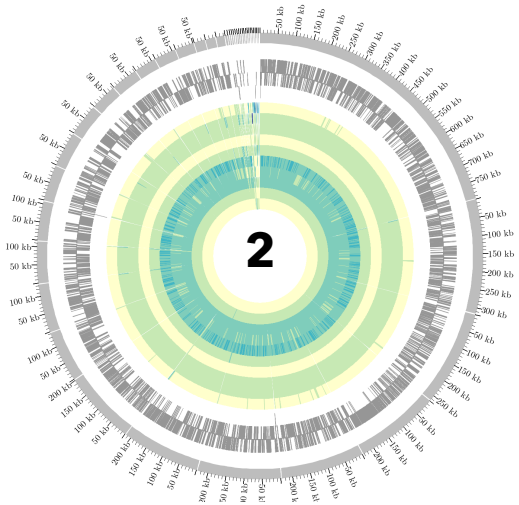
['M1_2_V1_contig_35980',
'GCCTTATCTTCATAAATAATATAATCTCTCACATCTTTGATCCACATAAAAAAACTCTC
CTTTTATGGAGAGTATAACGTAATCTCAAACTTCCTGCAACGAAAGTTTTCCAGATTATG
AGGACGTCAAAGACGGTCATCACTTTTTACTTTTCCAGATTTCAAAAATGATGACCGTCT
CAATGTTCTTCGTTATTCAATTTATTTTCCTCTTAGTATTCTTTCTATCTTATTATAGAT
TACTTTATAGAGTTCTGTATTCAATTTTTAATATAAAGATAAATTTTATGATTCTCTGAT
TCCCTGCTCATAATCCATATGATAATACTATCACTGGTTTTACTTAGAAAGTTTTATAGA
TTTAAATTATAATTTCACGGATTATAATTTAGATTTTATTTCGAAATATCGGATACACTT
TTTCTCTCTATTCGTGATAAGCAATCATAAACCTAACTTCTTAGATTCCCAACTGTTTAT
TTATCCATTGTACTTTAACAGTTTCCAGAACACAAATGGCAGATGTTCCAATCCTCTTTG
TAAAGTATCATTTGAAAAAGTACCTTAATATTTCTTATGTAATATCCCCGTACCTTTATG

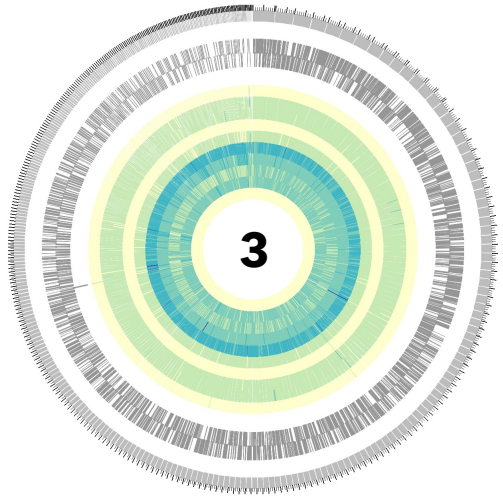
*k*Mer profiles of different bacteria

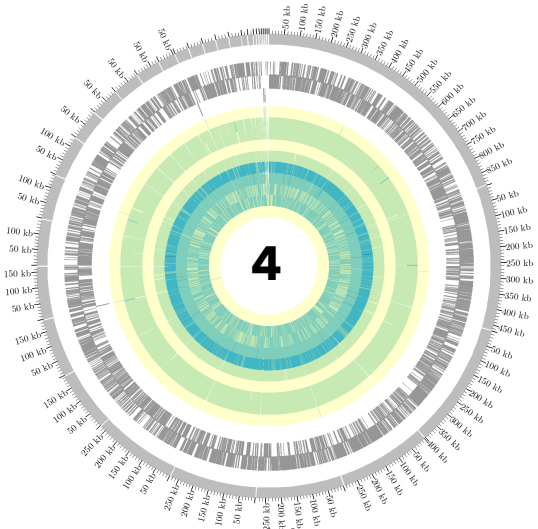
Paenibacillus sp.

Brevibacterium casei





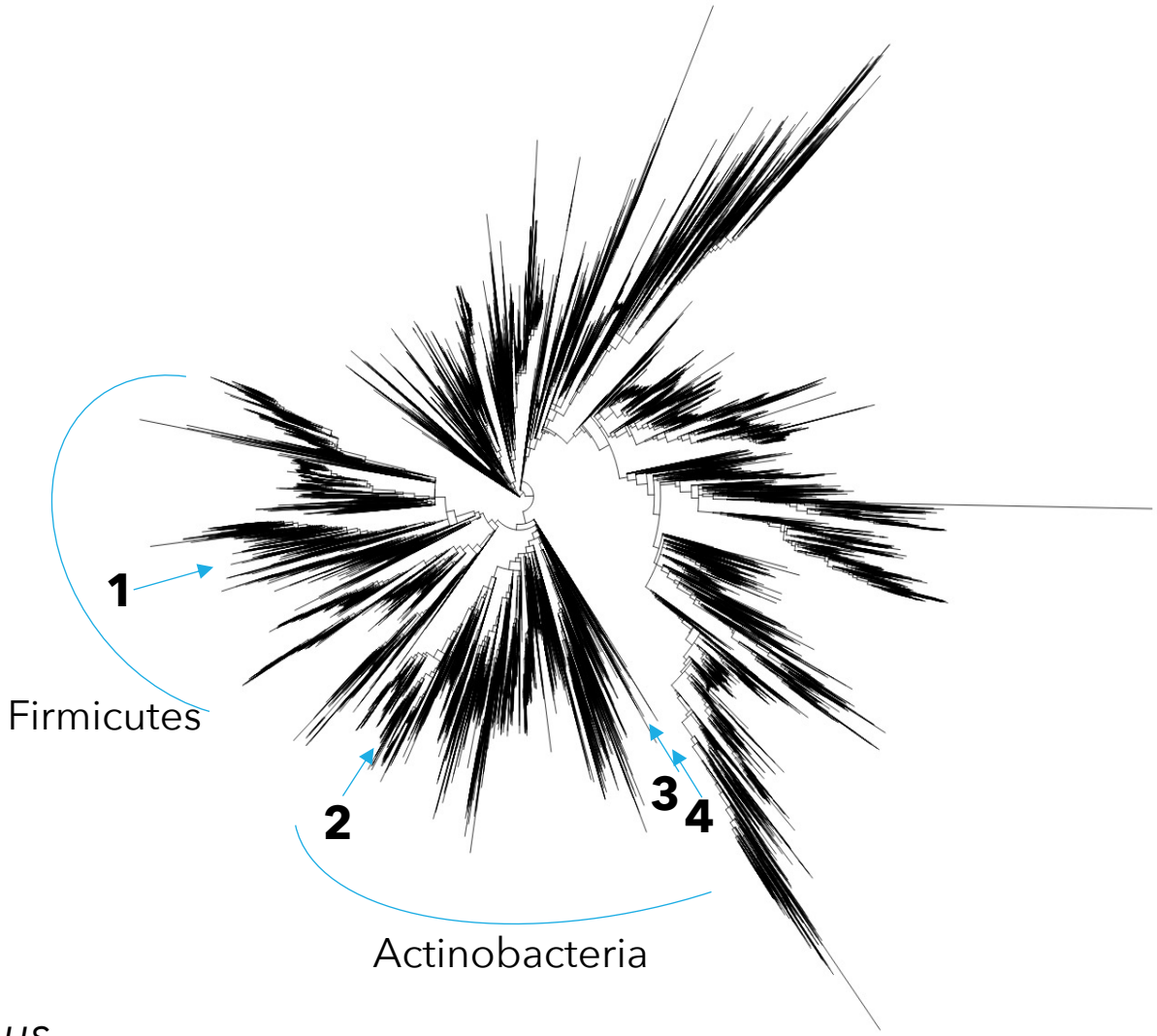




Patulibacter medicamentivorans

Patulibacter americanus

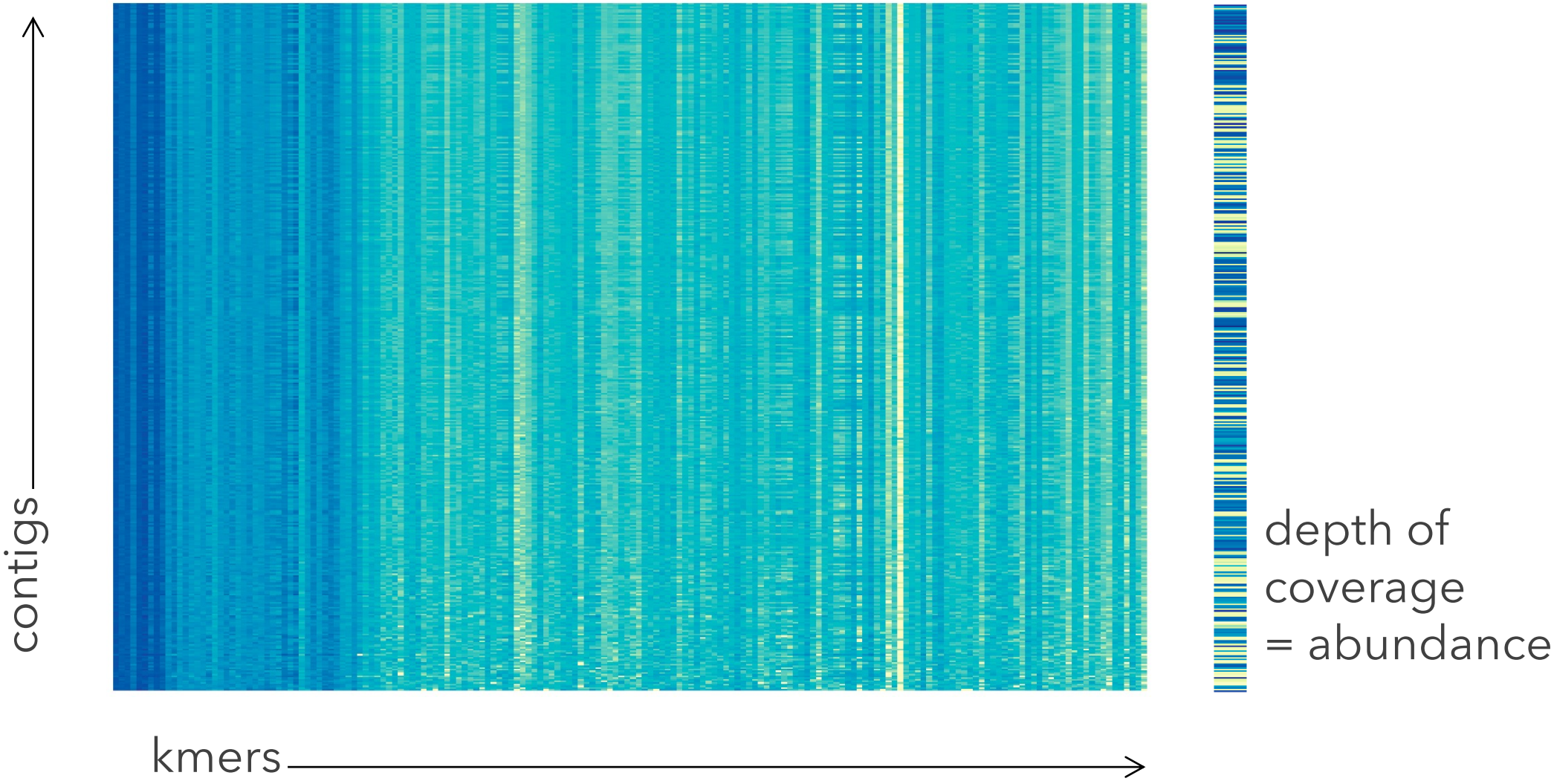
AA AC AG AT CA CC CG GA GC TA



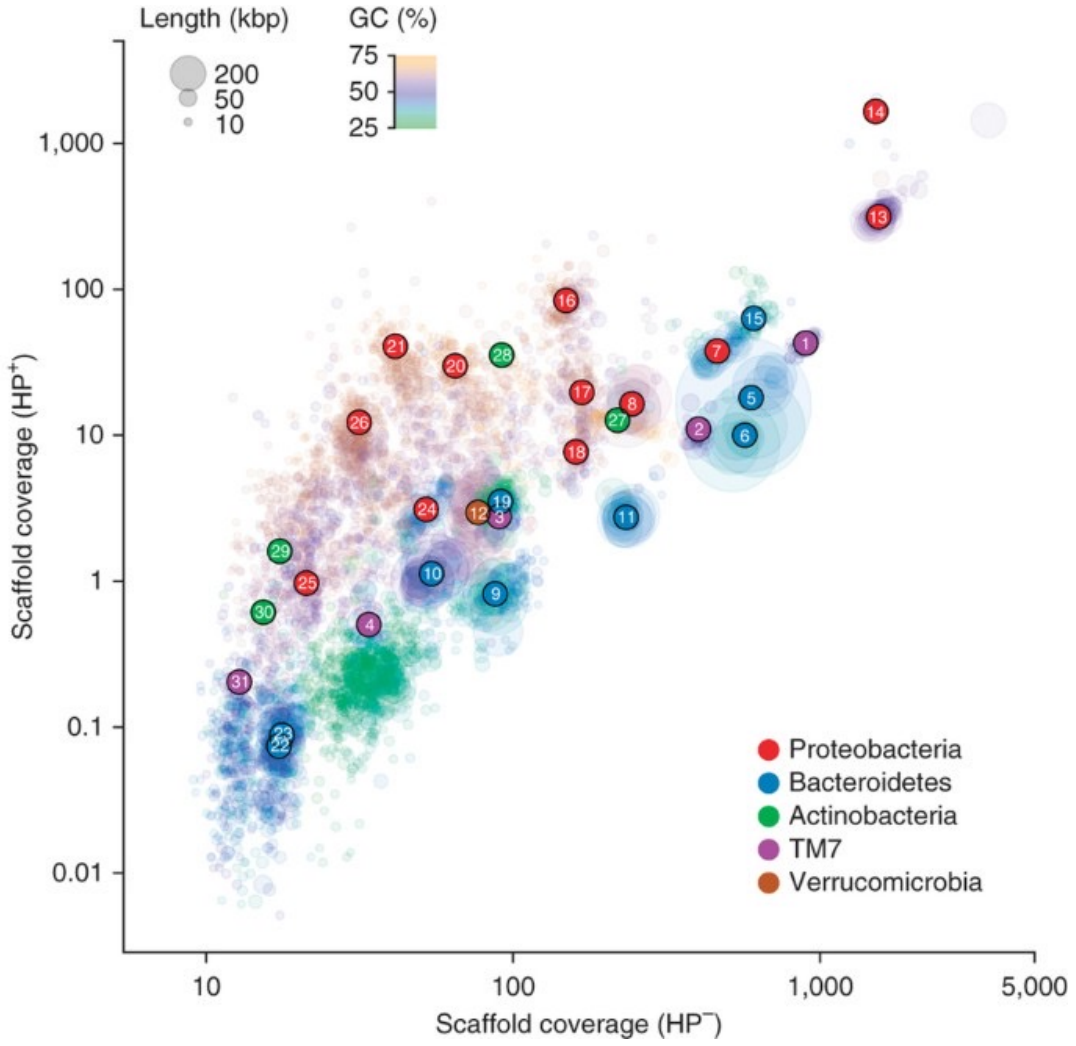
*k*mers, again

- word sizes for alignment seeding: BLAST default 11, BWA default 19
- *k*mers for taxonomy: kraken1/2 default 31/33
- *k*mers for diversity: nonpareil 24
- *k*mers for assembly: metaSPAdes between 25 and 127
- *k*mers for binning: 3-6

*k*Mer profiles in a metagenome

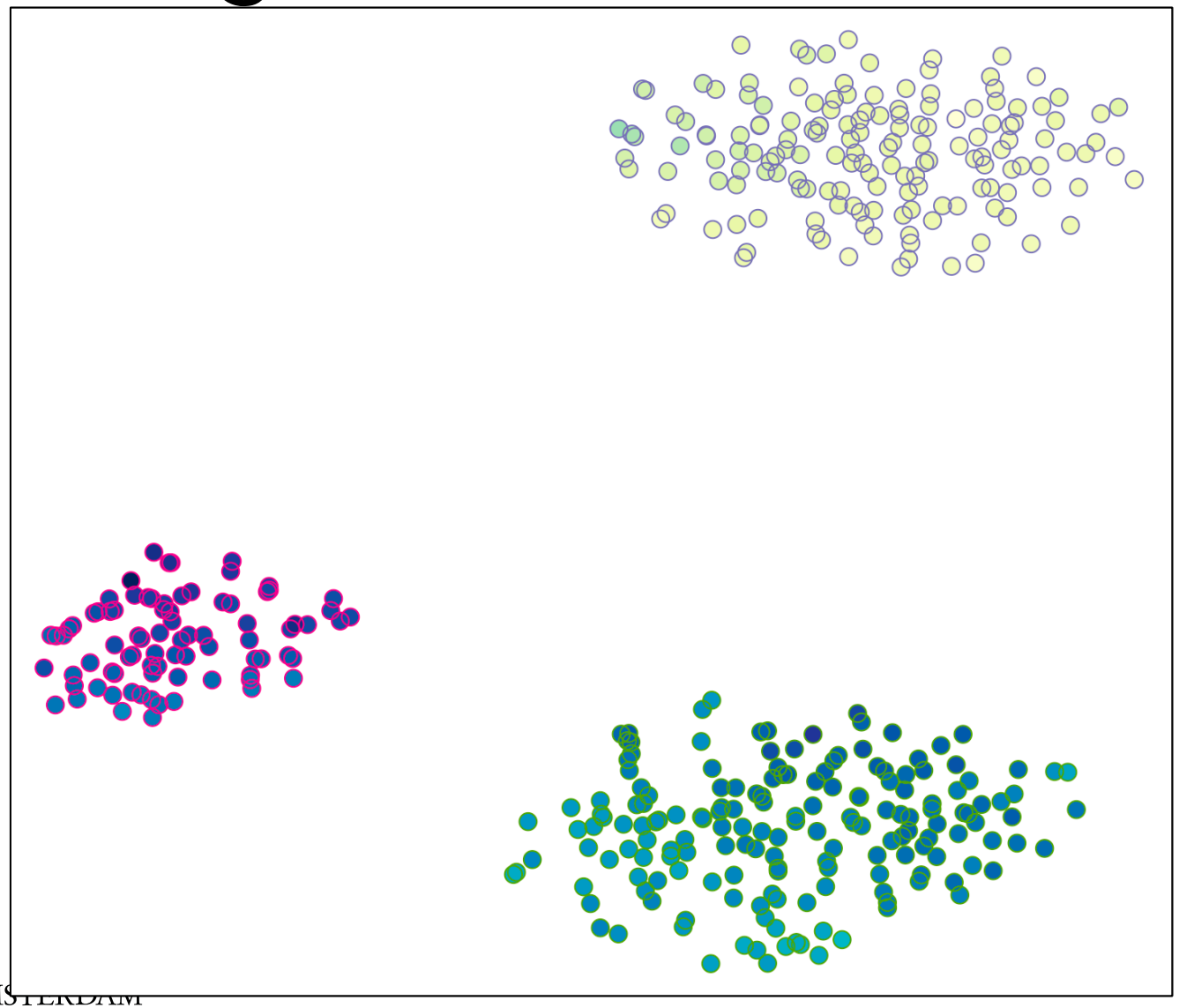


Coverage, pure and simple



Albertsen *et al*. (2013) Nat Biotechnol 31, 533–538. https://doi.org/10.1038/nbt.2579

Clustering



Group contigs, but how?

PeerJ

MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies

Dongwan D. Kang[1], Feng Li[2], Edward Kirton[1], Ashleigh Thomas[1], Rob Egan[1], Hong An[2] and Zhong Wang[1,3,4]

Binning metagenomic contigs by coverage and composition

Johannes Alneberg[1,8], Brynjar Smári Bjarnason[1,8], Ino de Bruijn[1,2], Melanie Schirmer[3], Joshua Quick[4,5], Umer Z Ijaz[3], Leo Lahti[6,7], Nicholas J Loman[4], Anders F Andersson[1,9] & Christopher Quince[3,9]

available as **Supplementary Software** and at https://github.com/BinPro/CONCOCT.

To validate CONCOCT, we constructed two synthetic mock metagenome data sets. The species mock was designed to test the ability of CONCOCT to resolve species-level variation in a complex community. It consists of 96 samples, each comprising random paired-end reads from the same 101 species but with different relative frequencies (see Online Methods). The strain mock contains only 20 genomes across 64 samples, but 5 genomes are from different strains of *Escherichia coli*; it was constructed to investigate the impact of strain-level variation on clustering (see **Supplementary Tables 1** and **2** for lists of genome sequences

nature biotechnology

LETTERS

https://doi.org/10.1038/s41587-020-00777-4

Check for updates

Improved metagenome binning and assembly using deep variational autoencoders

Jakob Nybo Nissen[1,2], Joachim Johansen[2], Rosa Lundbye Allesøe[2], Casper Kaae Sønderby[3], Jose Juan Almagro Armenteros[1], Christopher Heje Grønbech[3,4], Lars Juhl Jensen[2], Henrik Bjørn Nielsen[5], Thomas Nordahl Petersen[6], Ole Winther[3,4,7] and Simon Rasmussen[2]✉

Bioinformatics, 32(4), 2016, 605–607
doi: 10.1093/bioinformatics/btv638
Advance Access Publication Date: 29 October 2015
Applications Note

OXFORD

Sequence analysis

MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets

Yu-Wei Wu[1,2,*], Blake A. Simmons[1,2,3] and Steven W. Singer[1,2]

nature COMMUNICATIONS

ARTICLE

Check for updates

https://doi.org/10.1038/s41467-022-29843-y OPEN

A deep siamese neural network improves metagenome-assembled genomes in microbiome datasets across different environments

Shaojun Pan[1,2], Chengkai Zhu[1,2,3], Xing-Ming Zhao[1,2,4,5]✉ & Luis Pedro Coelho[1,2]✉

CONCOCT

- 1 assembly: mapping of multiple samples

- coverage vectors: average depth of coverage per contig per sample
- *k*mer frequency: 4-mers

- pseudo-count
- normalization of coverage vector to sequencing depth
- concatenation of both matrices
- log-transformation

- PCA (components explaining 90% of variation)
- clustering by Gaussian mixture model

Alneberg *et al*. (2014) Nat Methods 11, 1144–1146. https://doi.org/10.1038/nmeth.3103

MetaBAT2

- 1 assembly: mapping of multiple samples

- coverage matrix: average depth of coverage per contig per sample
- *k*mer frequency: 4-mers
- Pearson correlation of abundances

- quantile normalization of coverage, (correlation,) and *k*mer data
- geometric mean of all data's similarities as score per contig pair

- graph-based clustering (contigs as nodes, scores as edge weight):
- graph-building: incorparate set of contigs with highest similarity
- graph-partitioning: accelerated label propagation

Kang *et al*. (2019) PeerJ 7:e7359. https://doi.org/10.7717/peerj.7359

VAMB

- 1 or more assemblies: mapping of many samples

- coverage matrix: reads per kilobase per million mapped reads
- *k*mer frequency: 4-mers

- variational autoencoder to get a latent representation matrix, using a reconstruction error made up of cross-entropy as abundance error and sum of squares for *k*mer error
- clustering by adaptive iterative medoid, with cluster boundaries determined by cosine distance density
- repeat clustering step until everything is clustered

- optionally split by assembly of origin

Nissen *et al*. (2021) Nat Methods 39, 555-560. https://doi.org/10.1038/s41587-020-00777-4

MaxBin2

- 1 assembly: mapping of multiple samples

- coverage matrix: reads/contig length per contig per sample
- *k*mer frequency: 4-mers

- Expectation-Maximization algorithm (probability that contig S belongs to a genome based on the *k*mer frequency and coverage matrix):
- Gaussian distribution estimate of Euclidean distance between *k*mers
- Poisson distribution for coverage distance
- combination by multiplication
- initialize number of genomes based on the average number of present essential, single-copy genes
- iterate up to 50x

Wu *et al*. (2015) Bioinformatics 32, 605–607. https://doi.org/10.1093/bioinformatics/btv638

COCACOLA

- 1 assembly: mapping of multiple samples

- coverage vectors: average depth of coverage per contig per sample
- *k*mer frequency: 4-mers
- linkage by paired ends in multiple samples
- alignment to the same taxonomy

- pseudo-count
- normalization of coverage vector to sequencing depth
- concatenation of both matrices
- weight matrix (normalized Laplace) for linkage/taxonomy information

- initialize clustering by K-means on L_1 norm, then solve optimization (minimization) of genome assignment of each contig using alternating non-negative least squares

Lu *et al*. (2017) Bioinformatics 33:791-798. https://doi.org/10.1093/bioinformatics/btw290

SemiBin

- 1 assembly: mapping of multiple samples

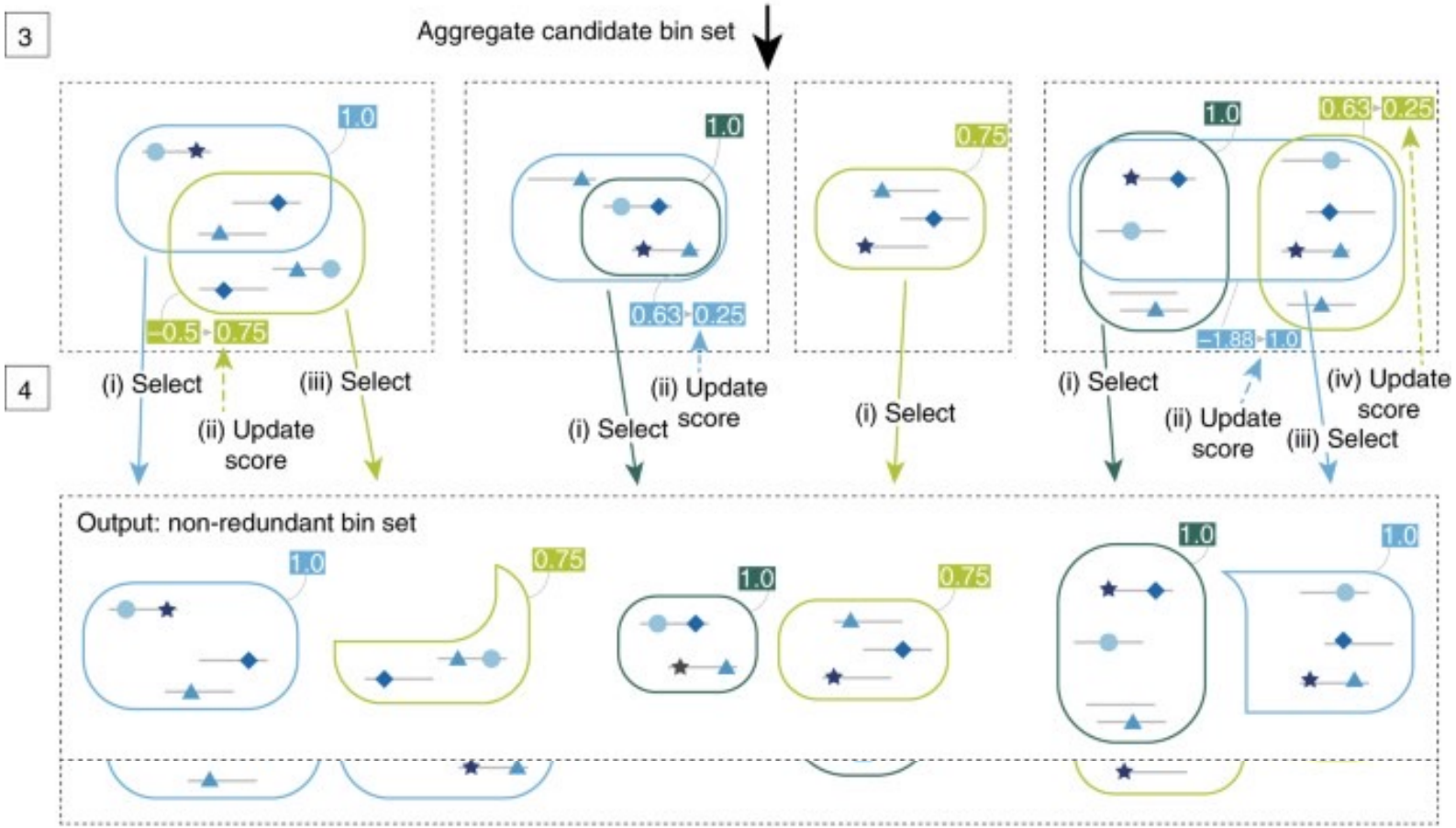
- coverage vectors: average depth of coverage per contig per sample
- *k*mers: 4-mers
- taxonomic annotation based on sequence clustering

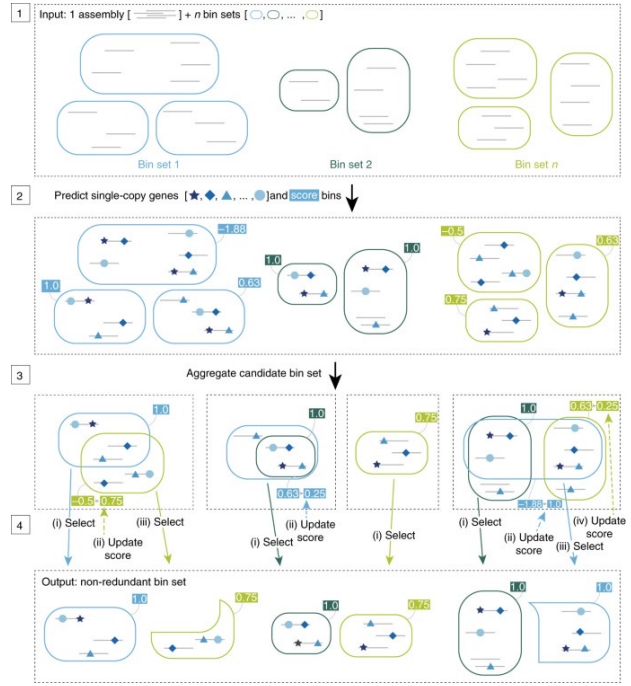
- pseudo-count, normalization of *k*mer vector to contig length
- scale coverage vectors to similar order of magnitude as *k*mers
- establish must-links and can't-links based on taxonomic annotation

- embedding based on deep siamese neural network in 100 dimensions
- use euclidean distances as edges in graph
- partition the graph into communities using Infomap

Meyer *et al*. (2022) Nat Methods 19, 429–440. https://doi.org/10.1038/s41592-022-01431-4

Refinement

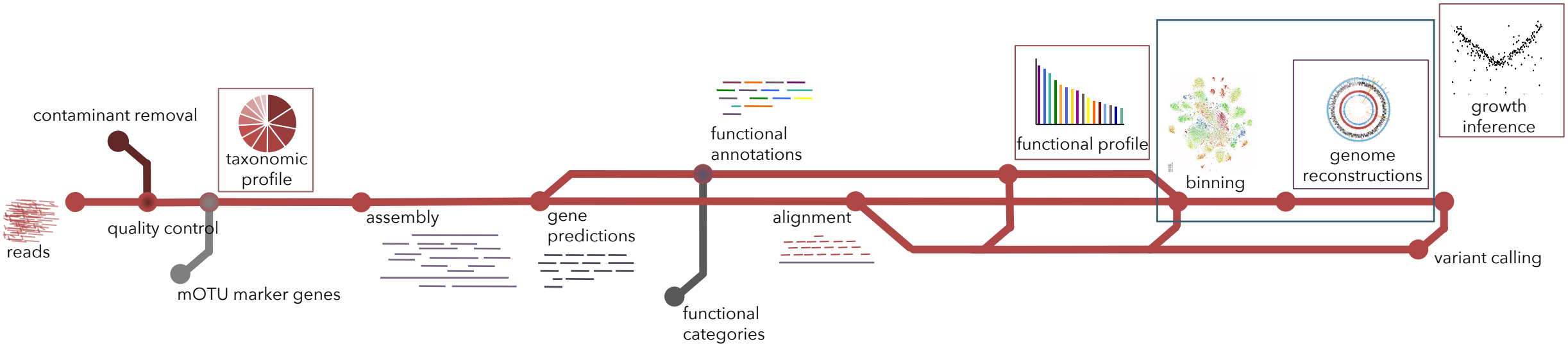




DAS Tool

Sieber *et al.* (2018) Nat Microbiol 3, 836–843. https://doi.org/10.1038/s41564-018-0171-1

Metagenomics (+ other omics) pipeline



MetaBAT2, MaxBin2, binny$_1$, DAS tool

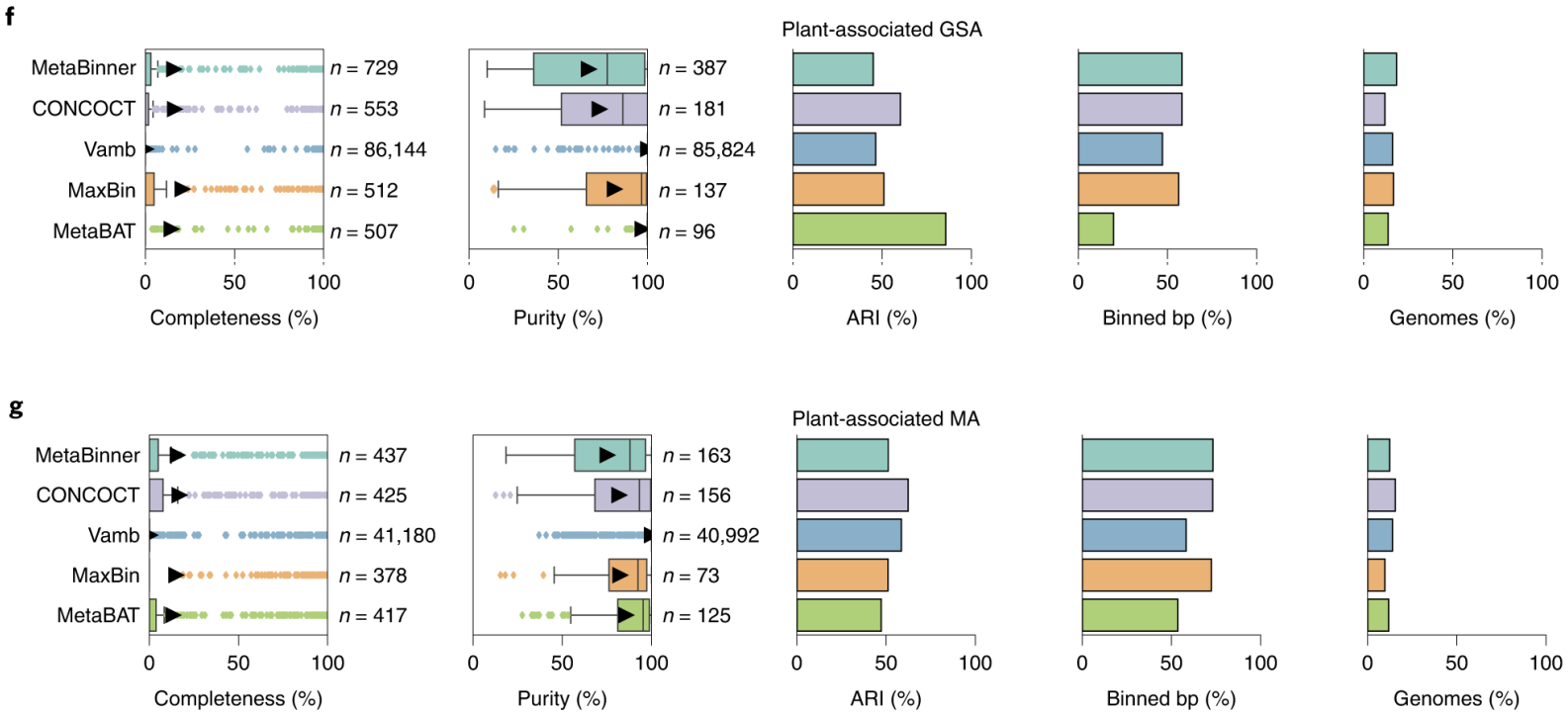
- binny$_1$:
- 1 assembly, 1 alignment

- *k*mers: 4-mer frequencies
- pseudo-count, centre-log ration scaling

- iterate over:
- t-SNE for embedding
- clustering by DB-SCAN
- assess completeness based on essential, single-copy genes
- split based on coverage depth

Benchmarking

- based on known genomes



nature methods
ANALYSIS
https://doi.org/10.1038/s41592-022-01431-4
Check for updates
OPEN

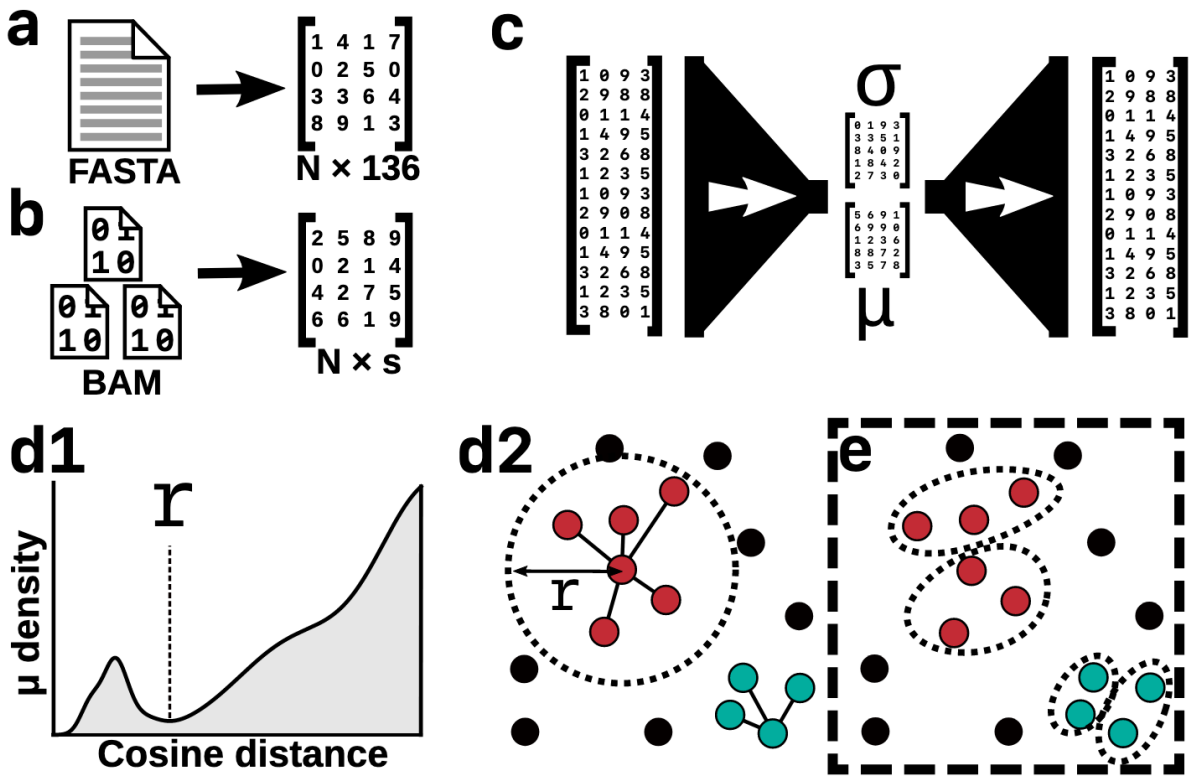
Critical Assessment of Metagenome Interpretation: the second round of challenges

Fernando Meyer[1,2,76], Adrian Fritz[1,2,3,76], Zhi-Luo Deng[1,2,4], David Koslicki[5], Till Robin Lesker[3,6], Alexey Gurevich[7], Gary Robertson[1,2], Mohammed Alser[8], Dmitry Antipov[9], Francesco Beghini[10], Denis Bertrand[11], Jaqueline J. Brito[12], C. Titus Brown[13], Jan Buchmann[14], Aydin Buluç[15,16], Bo Chen[15,16], Rayan Chikhi[17], Philip T. L. C. Clausen[18], Alexandru Cristian[19,20], Piotr Wojciech Dabrowski[21,22], Aaron E. Darling[23], Rob Egan[24,25], Eleazar Eskin[26], Evangelos Georganas[27], Eugene Goltsman[24,25], Melissa A. Gray[19,28], Lars Hestbjerg Hansen[29], Steven Hofmeyr[15,16], Pingqin Huang[30], Luiz Irber[13], Huijue Jia[31,32], Tue Sparholt Jørgensen[33,34], Silas D. Kieser[35,36], Terje Klemetsen[37], Axel Kola[38], Mikhail Kolmogorov[39], Anton Korobeynikov[9,40], Jason Kwan[41], Nathan LaPierre[26], Claire Lemaitre[42], Chenhao Li[11], Antoine Limasset[43], Fabio Malcher-Miranda[44], Serghei Mangul[12], Vanessa R. Marcelino[45,46], Camille Marchet[43], Pierre Marijon[47], Dmitry Meleshko[9], Daniel R. Mende[48], Alessio Milanese[49,50], Niranjan Nagarajan[51,52], Jakob Nissen[53], Sergey Nurk[54], Leonid Oliker[15,16], Lucas Paoli[49], Pierre Peterlongo[42], Vitor C. Piro[44], Jacob S. Porter[55], Simon Rasmussen[56], Evan R. Rees[41], Knut Reinert[57], Bernhard Renard[44,58], Espen Mikal Robertsen[37], Gail L. Rosen[19,28,59], Hans-Joachim Ruscheweyh[49], Varuni Sarwal[26], Nicola Segata[10], Enrico Seiler[57], Lizhen Shi[60], Fengzhu Sun[61], Shinichi Sunagawa[49], Søren Johannes Sørensen[62], Ashleigh Thomas[24,63], Chengxuan Tong[11], Mirko Trajkovski[35,64], Julien Tremblay[65], Gherman Uritskiy[66], Riccaro Vicedomini[17], Zhengyang Wang[30], Ziye Wang[67], Zhong Wang[68,69,70], Andrew Warren[55], Nils Peder Willassen[37], Katherine Yelick[15,16], Ronghui You[30], Georg Zeller[50], Zhengqiao Zhao[19], Shanfeng Zhu[71,72], Jie Zhu[31,32], Ruben Garrido-Oter[73], Petra Gastmeier[38], Stephane Hacquard[73], Susanne Häußler[6], Ariane Khaledi[6], Friederike Maechler[38], Fantin Mesny[73], Simona Radutoiu[74], Paul Schulze-Lefert[73], Nathiana Smit[6], Till Strowig[6], Andreas Bremges[1,3], Alexander Sczyrba[75] and Alice Carolyn McHardy[1,2,3,4]✉

Meyer *et al*. (2022) Nat Methods 19, 429–440. https://doi.org/10.1038/s41592-022-01431-4

Side note: why are some tools used more than others?

Improved metagenome binning and assembly using deep variational autoencoders



Side note: why are some tools used more than others?

Y.Y.Lu et al.

coverage across multiple samples for binning. Compared with recent approaches such as CONCOCT, GroopM, MaxBin and MetaBAT, COCACOLA performs better in three aspects. First, COCACOLA reveals superiority with respect to precision, recall and Adjusted Rand Index (ARI). Second, COCACOLA shows better robustness in the case of varying number of samples. COCACOLA is scalable and faster than CONCOCT, GroopM, MaxBin and MetaBAT.

In addition, the COCACOLA framework seamlessly embraces customized knowledge to facilitate binning accuracy. In our study, we have investigated two types of knowledge, in particular, the co-alignment to reference genomes and linkage between contigs provided by paired-end reads. We find that both co-alignment and linkage information facilitate better binning performance in the majority of the cases.

2 Materials and methods

2.1 Problem formulation

A microbial community is composed of a abundance levels, and our objective is to omic OTU bins from which they were ori expected to be disentangled based on con discriminative abundance or dissimilarity of l-mer composition. The rationale of bi relies on the underlying assumption that the same OTU share similar relative abun composition.

Formally, we encode the abundance an OTU by a $(M+V)$ dimensional feature ve where M is the number of samples, V is the and K is the total OTU number. Specifi abundance of the k-th OTU in the m-th sa spectively. And $W_{M+v,k}$ stands for the l-mer relative frequency composition of the k-th OTU, $v = 1, 2, \cdots, V$. Similarly, the feature vector of the n-th contig is denoted as X_n. Let $\mathbb{H}_{kn}$ be the indicator function describing whether the n-th contig belongs to the k-th OTU, i.e. $\mathbb{H}_{kn} = 1$ means the n-th contig originating from the k-th OTU and $\mathbb{H}_{kn} = 0$ otherwise. Therefore, X_n can be represented as:

$$X_n = \mathbb{H}_{1n} W_1 + \mathbb{H}_{2n} W_2 + \cdots + \mathbb{H}_{kn} W_K, \quad n = 1, 2, \cdots, N \quad (1)$$

where N is the number of contigs. Equation (1) can be further written into the matrix form:

$$X \approx W\mathbb{H} \qquad s.t. \quad W \geq 0, \quad \mathbb{H} \in \{0,1\}^{K \times N}, \|\mathbb{H}_n\|_0 = 1 \quad (2)$$

where $W = (W_1, W_2, \cdots, W_K)$ is a $(M+V) \times K$ non-negative matrix with each column encoding the feature vector of the corresponding OTU. And $\mathbb{H} = (\mathbb{H}_1, \mathbb{H}_2, \cdots, \mathbb{H}_N)$ is a $K \times N$ binary matrix with each column encoding the indicator function of the corresponding contig. $\|\mathbb{H}_n\|_0 = \sum_{k=1}^{K} \mathbb{H}_{kn} = 1$ ensures the n-th contig belongs exclusively to only one particular OTU.

The matrices W and $\mathbb{H}$ are obtained by minimizing a certain objective function. In this article we use Frobenius norm, commonly known as the sum of squared error:

$$\arg\min_{W,\mathbb{H}\geq 0} \|X - W\mathbb{H}\|_F^2 \qquad s.t. \quad \mathbb{H} \in \{0,1\}^{K \times N}, \|\mathbb{H}_n\|_0 = 1 \quad (3)$$

Note that Equation (3) is NP-hard by formulation as an integer programming problem with an exponential number of feasible solutions (Jiang *et al.*, 2014). A common procedure to tackle Equation (3) relaxes binary constraint of $\mathbb{H}$ with numerical values. Hence

COCACOLA 793

Equation (3) is reformulated as the following minimization problem:

$$\arg\min_{W,H} \|X - WH\|_F^2 \qquad s.t. \quad W, H \geq 0 \quad (4)$$

where H serves as a coefficient matrix instead of an indicator matrix. In the scenario of Equation (4), $W_{\cdot k}$, the feature vector of the k-th OTU represents the centroid of the k-th cluster. Meanwhile, each contig $X_{\cdot n}$ is approximated by a weighted mixture of clusters, where the weights are encoded in $H_{\cdot n}$. In other words, relaxation of binary constraint makes the interpretation from hard clustering to soft clustering, where hard clustering means that a contig can be assigned to one OTU only, while soft clustering allows a contig to be assigned to multiple OTUs. It has been observed that by imposing sparsity on each column of H, the hard clustering behavior can be facilitated (Kim and Park, 2008). Therefore, Equation (4) is further modified through the Sparse Non-negative Matrix Factorization form (Kim and Park, 2008):

The feature matrix of contigs is denoted as $X = [p\ q]^T$, as the combination of coverage profile p and composition profile q. To be specific, X is a $(M+V) \times N$ non-negative matrix of which each column represents the feature vector of a particular contig.

2.3 Incorporating additional knowledge into binning

We consider two types of additional knowledge that may enhance the binning accuracy (Basu *et al.*, 2008). One option is paired-end reads linkage. Specifically, a high number of links connecting two contigs imply high possibility that they belong to the same OTU. Because the linkage may be erroneous owing to the existence of chimeric sequences, we keep linkages that are reported through multiple samples. The other option is co-alignment to reference genomes. That is, two contigs mapped to the same reference genome support the evidence that they belong to the same OTU.

We encode additional knowledge by an undirected network in

the tetra-mer composition denotes the tetra-mer frequency for the contig itself plus its reverse complement. Owing to palindromic tetra-mers, $V = 136$.

Adopting the notation of CONCOCT (Alneberg *et al.*, 2014), the coverage of all the N contigs is represented by an $N \times M$ matrix Y, where N is the number of contigs of interest and Y_{nm} indicates the coverage of the n-th contig from the m-th sample. Whereas the tetra-mer composition of the N contigs are represented by an $N \times V$ matrix Z where Z_{nv} indicates the count of v-th tetra-mer found in the n-th contig. Before normalization, a pseudo-count is added to each entry of the coverage matrix Y and composition matrix Z, respectively. As for the coverage, a small value is added, i.e. $Y'_{nm} = Y_{nm} + 100/L_n$, analogous to a single read aligned to each contig as prior, where L_n is the length of the n-th contig. As for the composition, a single count is simply added, i.e. $Z'_{nv} = Z_{nv} + 1$.

The coverage matrix Y is first column-wise normalized (i.e. normalization within each individual sample), followed by row-wise normalization (i.e. normalization across M samples) to obtain coverage profile p. The row-wise normalization aims to mitigate sequencing efficiency heterogeneity among contigs.

$$Y''_{nm} = \frac{Y'_{nm}}{\sum_{n=1}^{N} Y'_{nm}} \qquad p_{nm} = \frac{Y''_{nm}}{\sum_{m=1}^{M} Y''_{nm}} \quad (6)$$

The composition matrix Z is row-wise normalized for each contig (i.e. normalization across M tetra-mer count) to obtain composition profile q:

$$q_{nv} = \frac{Z'_{nv}}{\sum_{v=1}^{V} Z'_{nv}} \quad (7)$$

Equation (3) changes to the following form:

$$\arg\min_{W,H\geq 0} \|X - WH\|_F^2 + \alpha \sum_{n=1}^{N} \left\|H_{\cdot n}\right\|_1^2 + \beta\, Tr(H\mathcal{L}H^T) \quad (9)$$

where the parameter $\beta > 0$ controls the trade-off of belief between unsupervised binning and additional knowledge. Namely, large β indicates strong confidence on the additional knowledge. Conversely, small β puts more weight on the data.

To use multiple additional knowledge sources together, a combined *Laplacian matrix* is constructed as a weighted average of individual *Laplacian matrices* $\tilde{\mathcal{L}} = \left(\sum_d \alpha_d \mathcal{L}_d\right) / \left(\sum_d \alpha_d\right)$ where each positive weight α_d reflects the contribution of the corresponding information. For simplicity, weights are treated equally in the article.

2.4 Optimization by alternating non-negative least squares

Among comprehensive algorithms to solve Equation (9), the multiplicative updating approach (Lee and Seung, 1999) is most widely used. Despite its simplicity in implementation, slow convergence is of high concern. This article adopts a more efficient algorithm with provable convergence called alternating non-negative least squares (ANLS) (Kim and Park, 2008). ANLS iteratively handles two non-negative least square subproblems in Equation (10) until convergence. The ANLS algorithm is summarized in Algorithm 1.

$$H \leftarrow \arg\min_{H\geq 0} \|X - WH\|_F^2 + \alpha \sum_{n=1}^{N} \left\|H_{\cdot n}\right\|_1^2 + \beta\, Tr(H\mathcal{L}H^T) \quad (10a)$$

794 Y.Y.Lu et al.

$$W \leftarrow \arg\min_{W\geq 0} \left\|X^T - H^T W^T\right\|_F^2 \quad (10b)$$

We solve Equation (10a) by block coordinate descent, that is, we divide Equation (10a) into N subproblems and minimize the objective function with respect to each subproblem at a time while keeping the rest fixed:

$$\arg\min_{H_{\cdot n}\geq 0} \|X_{\cdot n} - WH_{\cdot n}\|_2^2 + \alpha\|H_{\cdot n}\|_1^2 + \beta H_{\cdot n}^T \mathcal{L} H_{\cdot n}, \quad n = 1, \cdots, N$$

$$= \arg\min_{H_{\cdot n}\geq 0} \|X_{\cdot n} - WH_{\cdot n}\|_2^2 + \alpha\|H_{\cdot n}\|_1^2 + \beta H_{\cdot n}^T (H_{\cdot n} - 2\sum_{n'=1}^{N} \mathcal{A}_{nn'} H_{\cdot n'}^{old})$$

$$= \arg\min_{H_{\cdot n}\geq 0} \|X_{\cdot n} - WH_{\cdot n}\|_2^2 + \alpha\|H_{\cdot n}\|_1^2 + \beta \|H_{\cdot n} - \sum_{n'=1}^{N} \mathcal{A}_{nn'} H_{\cdot n'}^{old}\|_2^2 \quad (11)$$

where the matrix H^{old} denotes the value of H obtained from the pre-

assignment.

The distance measurement contributes crucially to the success of binning. Ideally, a proper distance measurement exhibits more distinguishable taxonomic difference. The traditional K-means approach takes Euclidean distance as default measurement to quantify closeness. However, as for the coverage profile, Su *et al.* (2012) shows L_1 distance produces more reasonable binning results than Euclidean and correlation-based distances. As for the composition profile, L_1 distance also reveals superiority over Euclidean and cosine distances (Liao *et al.*, 2014). Therefore, our method adopts K-means clustering with L_1 distance. Once preliminary K-means clustering is achieved, we eliminate suspicious clusters with few contigs using the bottom-up L Method (Salvador and Chan, 2004). Performance comparisons with respect to L_1 and Euclidean distance are given in the supplementary material.

2.6 Parameter tuning

We have two parameters (α, β) to be tuned in our algorithm. Traditional cross-validation-like strategy demands searching through a two dimensional grid of candidate values, which is computationally unaffordable in the case of large datasets. Instead, we first search a good marginal α value by fixing $\beta = 0$. After that, a one-dimensional search is performed on a range of candidate β values while keeping α fixed.

In our implementation, when $\beta = 0$, α is approximated by the regression of the corresponding Lagrange Multipliers from N constrained problems $\text{argmin}_{H_{\cdot n}\geq 0} \|X - WH_{\cdot n}\|_F^2$ with constraint $(\|H_{\cdot n}\|_1 - 1)^2 = 0$, where $n = 1, \cdots, N$. The resulting α is denoted

Algorithm 1. Optimization by ANLS

Input: feature matrix $X \in \mathbb{R}^{(M+V)\times N}$, initial basis matrix $W \in \mathbb{R}^{(M+V)\times K}$ and coefficient matrix $H \in \mathbb{R}^{K\times N}$, tolerance threshold ε, maximum iteration threshold T

1: **repeat**
2: Obtain optimal H of Equation (10a) by fixing W
3: Obtain optimal W of Equation (10b) by fixing H
4: **until** A particular stopping criterion involving ε is satisfied or iteration number exceeds T

Output: W, H

by α^*. Then we run the algorithm with respect to each candidate β and fixed $\alpha = \alpha^*$, resulting in corresponding binning results with cluster number. Notice that traditional internal cluster validices are only applicable on the basis of fixed cluster number o (Wiwie *et al.*, 2015), such as Sum of Square Error and -Bouldin index (Davies and Bouldin, 1979). To be specific, ices have the tendency toward monotonically increase or de- as the cluster number increases (Liu *et al.*, 2013). We tackle act of monotonicity by adopting TSS (Tang-Sun-Sun) minion index (Tang *et al.*, 2005), that is, we choose the candidate minimum TSS value, recorded as β^*. Then we can solve on (9) by using (α^*, β^*) as selected regularization parameters.

COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge

st-processing

sulting binning obtained from Algorithm 1 may contain clust are closely mixed to each other. Therefore, we define *separonductance* as an effective measurement to diagnose the g closeness of pairwise clusters, so as to determine whether to merge them. Namely, we consider each cluster as having a spherical scope centered at its centroid. To be robust against outliers, the radius is chosen as the third quartile among the intra-cluster distances. The *separable conductance* between the c_1-th cluster and the c_2-th cluster, $sep(c_1, c_2)$, is defined as the number of contigs from the c_1-th cluster also included in the spherical scope of the c_2-th cluster, divided by the smaller cluster size of two. Intuitively, the *separable conductance* exploits the overlap between two clusters. The procedure of post-processing works as follows: we keep picking the pair of clusters with maximum *separable conductance* and merge them until it fails to exceed a certain threshold. The threshold is set to be 1 in this study.

2.8 Datasets

Alneberg *et al.* (2014) simulated a 'species' dataset and another 'strain' dataset. Both simulated datasets were constructed based on 16S rRNA samples originated from the Human Microbiome Project (HMP) (Consortium *et al.*, 2012). The relative abundance profiles of the different species/strains for the simulation were based on the HMP samples as well.

The simulated 'species' dataset consisted of 101 different species across 96 samples. It aimed to test the ability of CONCOCT to cluster contigs in complex populations (Alneberg *et al.*, 2014). The species were approximated by the OTUs from HMP with >3% sequence differences. Each species was guaranteed to appear in at least 20 samples. A total of 37 628 contigs remain for binning after co-assembly and filtering.

The simulated 'strain' dataset aimed to test the ability of CONCOCT to cluster contigs at different levels of taxonomic

Limitations

- binning is just as good as the assembly

- longer contigs perform better
- -> most binners use contigs >1,000 bp, some >2,500 bp

- more samples perform better than fewer samples

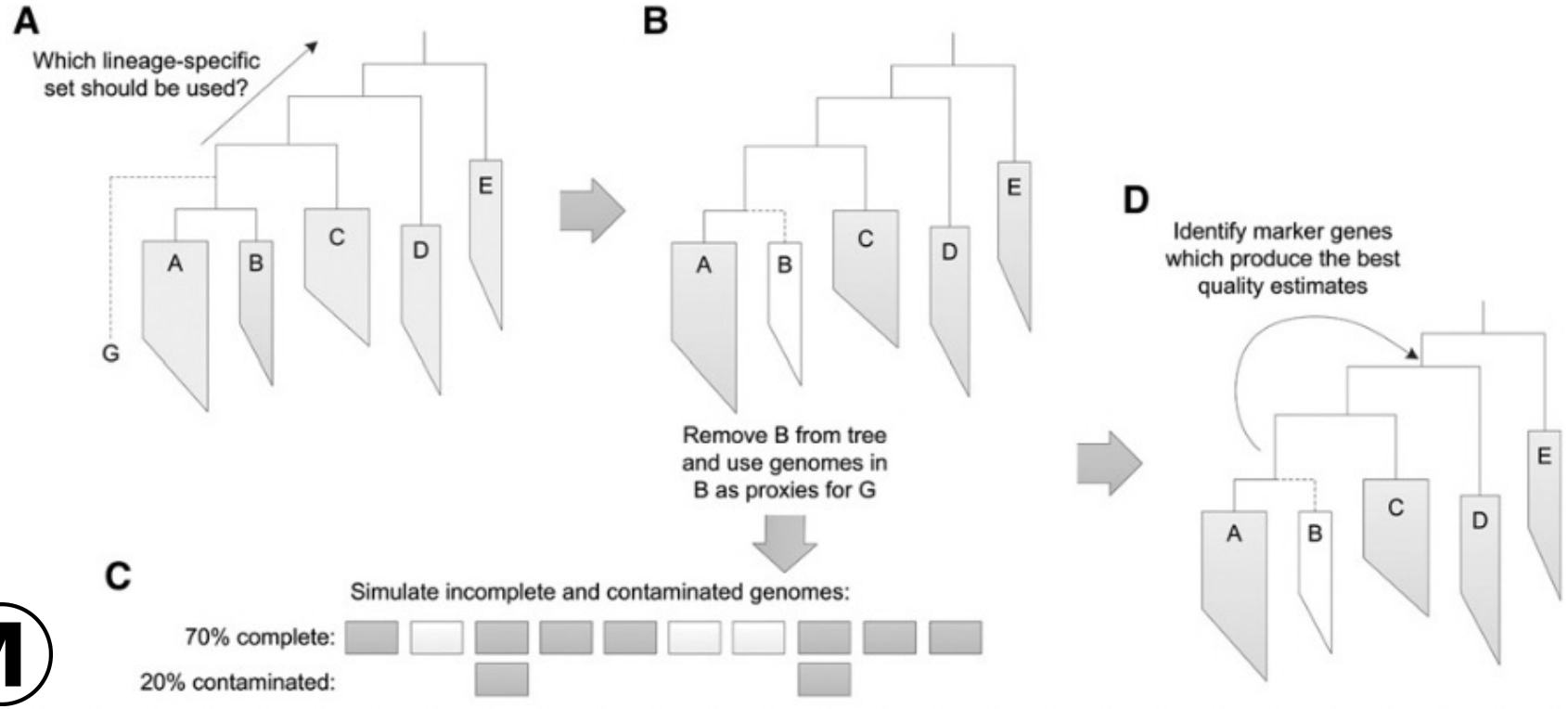
- the more knowledge is used, the less likely new organisms are found

- how to deal with partial genomes?

Quality assessment of MAGs

- completeness
- contamination

- based on gene content



CheckM

Parks *et al*. (2015) Genome Res. 25(7):1043-55. doi: 10.1101/gr.186072.114.

Quality measures and reporting

assembly quaility
***Finished**: Single, validated, contiguous sequence per replicon without gaps or ambiguities with a consensus error rate equivalent to Q50 or better. Assembly statistics*.*
***High Quality Draft**:Multiple fragments where gaps span repetitive regions. Assembly statistics*. Presence of the 23S, 16S and 5S rRNA genes and at least 18 tRNAs.*
***Medium Quality Draft**:Many fragments with little to no review of assembly other than reporting of standard assembly statistics*.*
***Low Quality Draft**:Many fragments with little to no review of assembly other than reporting of standard assembly statistics*.*
completeness score
***High Quality Draft**: >90%*
***Medium Quality Draft**: >50%*
***Low Quality Draft**: < 50%*
contamination score
***High Quality Draft:** < 5%*
***Medium Quality Draft:** < 10%*
***Low Quality Draft:** < 10%*
completeness software
Checkm, anvi'o, BUSCO or other

Bowers et al. (2017) Nat Biotechnol 35, 725–731. https://doi.org/10.1038/nbt.3893

Alternatives?

- long reads

  higher DNA quality demands
  more expensive -> lower depth
  more computational effort

- HiC

  lab/computational protocols not mature

- single-cell genomics

  lower throughput / depth
  technical challenges

- taxonomic annotation at gene/contig level

  only works for well-described organisms
  HGT events can’t be observed

Further reading

Review

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]

Chen *et al*. 2020, Genome Res. 30(3):315-333 https://doi.org/10.1101/gr.258640.119

METHOD

binny: an automated binning algorithm to recover high-quality genomes from complex metagenomic datasets

Oskar Hickl [1], Pedro Queirós [2] Paul Wilmes [3], Patrick May [1], and Anna Heintz-Buschart [4,

Hickl *et al*. bioRχiv https://doi.org/10.1101/2021.12.22.473795

Thanks for your attention!

a.u.s.heintzbuschart@uva.nl

SP C2.205

github.com/a-h-b

twitter.com/_a_h_b_

University of Amsterdam
Life Sciences