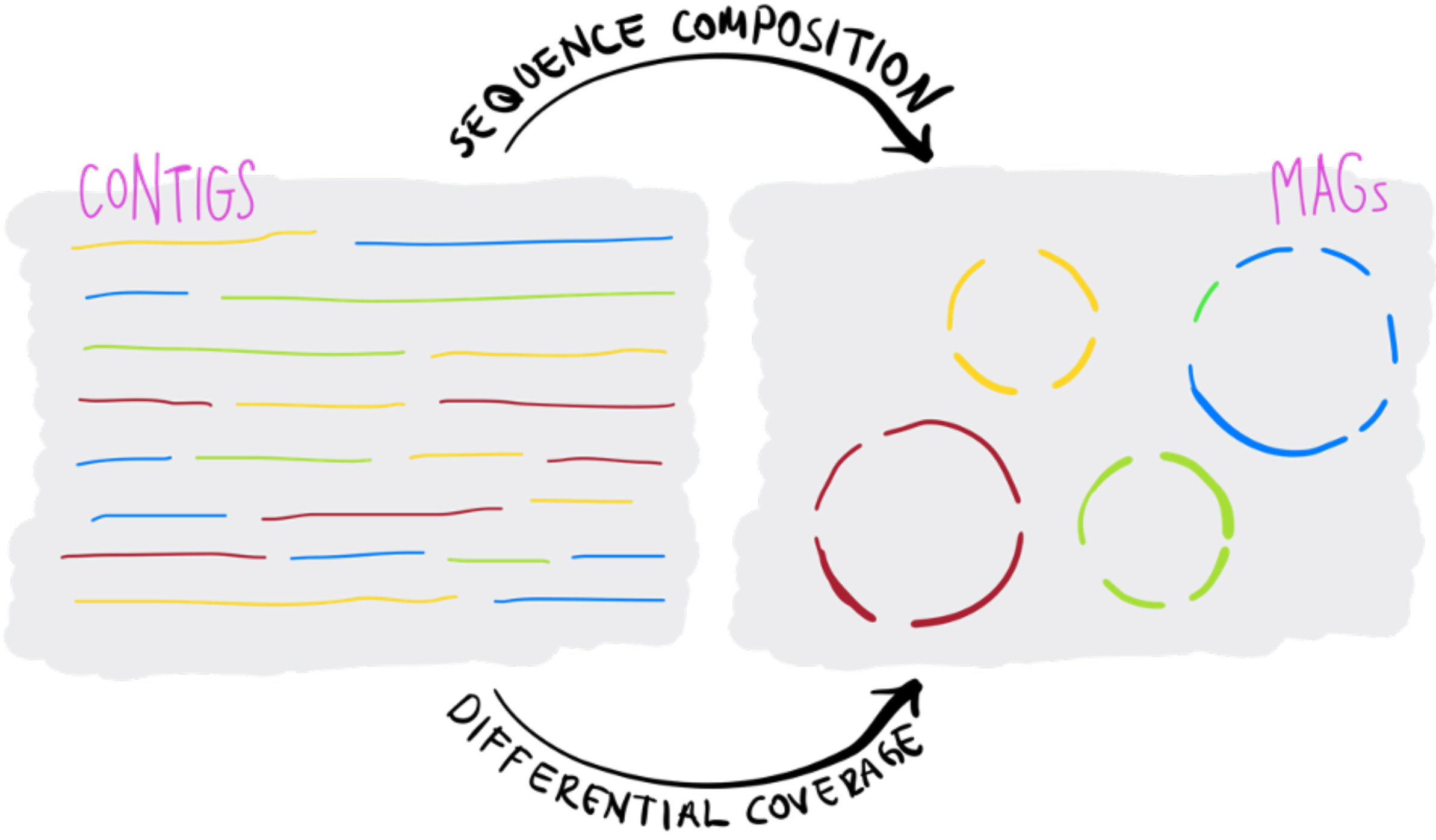
Metagenomics workshop

Module 2: from bins to genomes

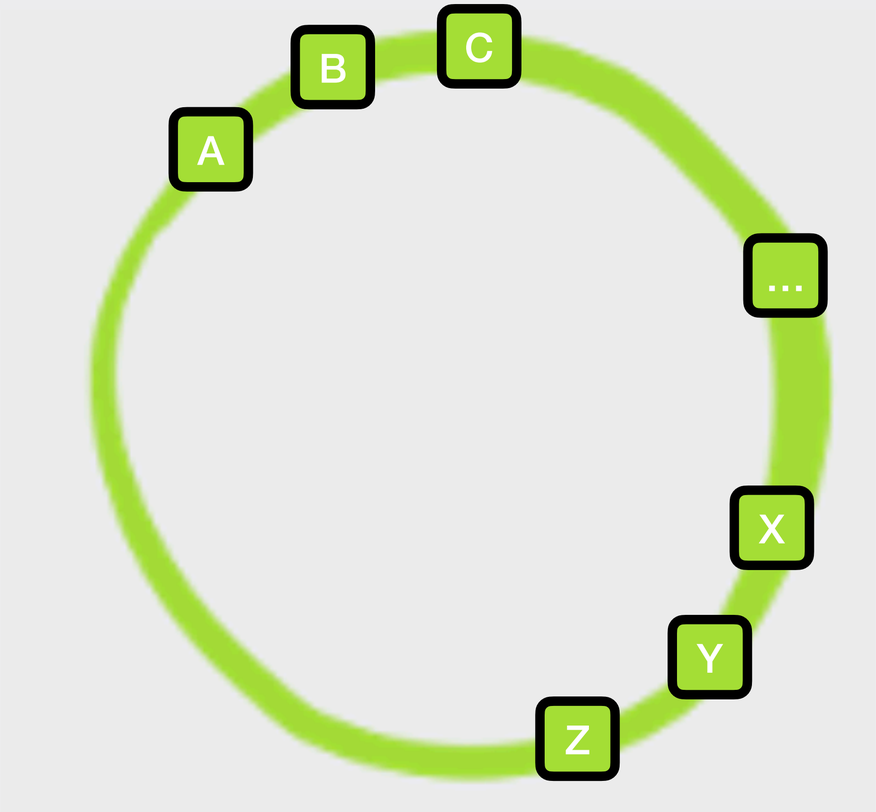
Lucas Paoli (paolil@ethz.ch), Sunagawa Lab
Institute of Microbiology, ETH Zurich

Serina Robinson (serina.robinson@eawag.ch)
Dept. of Environmental Microbiology, EAWAG

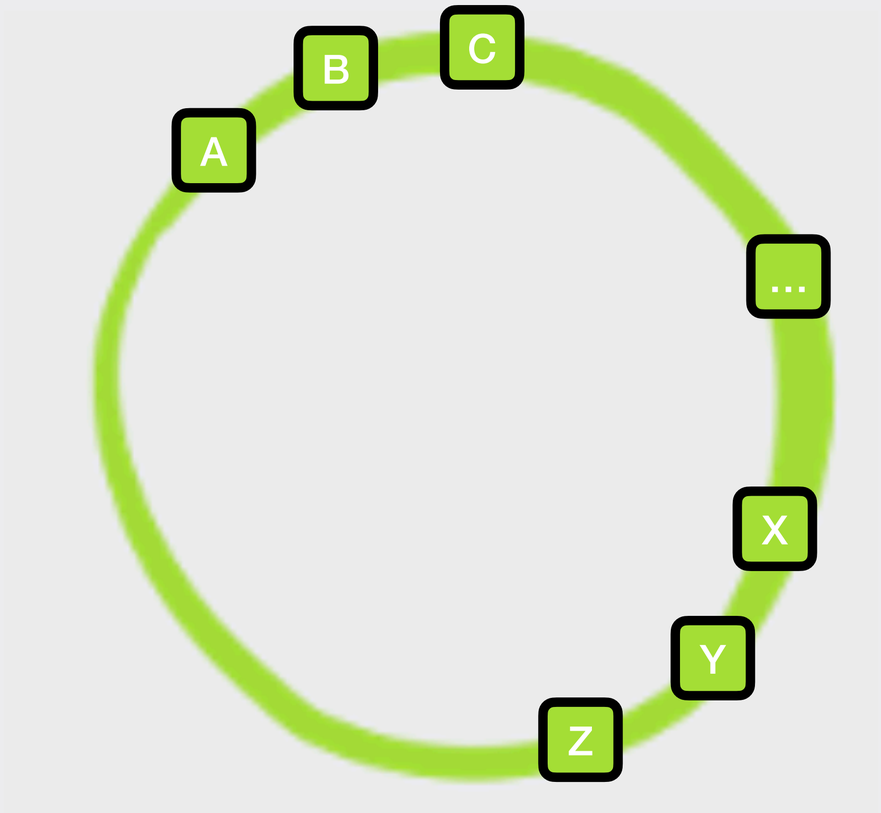

SEQUENCE COMPOSITION
CONTIGS
MAGs
DIFFERENTIAL COVERAGE

Distinguishing spurious groups of contigs from draft genomes: completeness and contamination

Universal single-copy marker genes

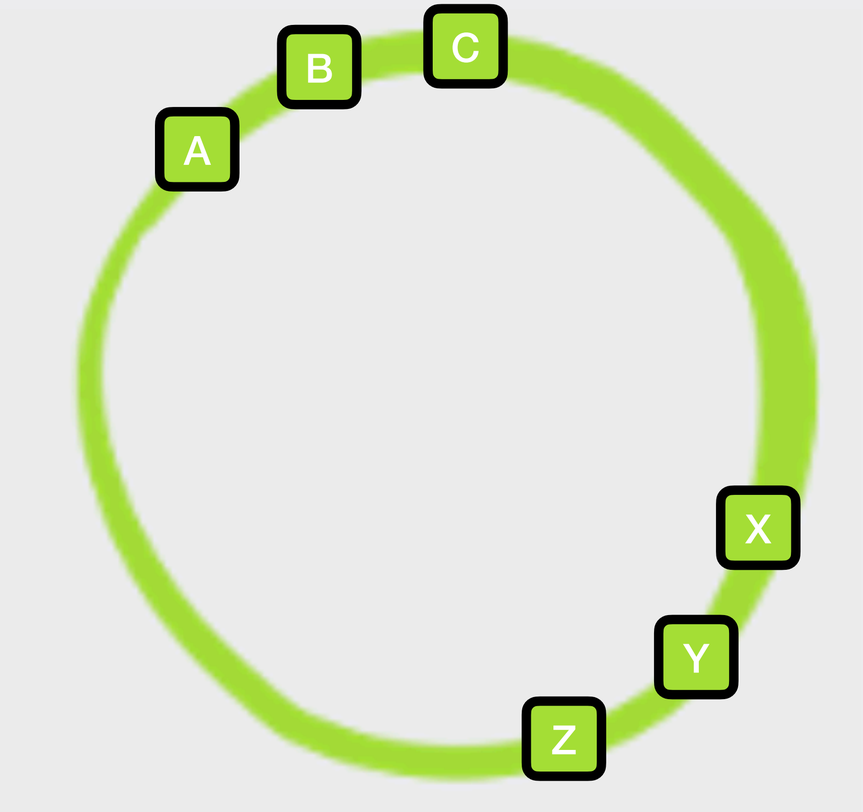


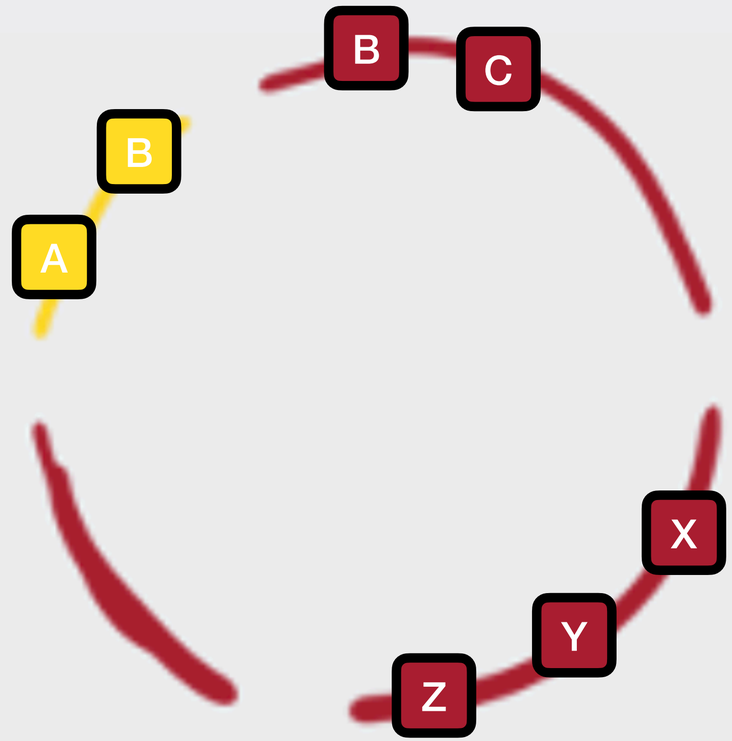
Universal single-copy marker genes

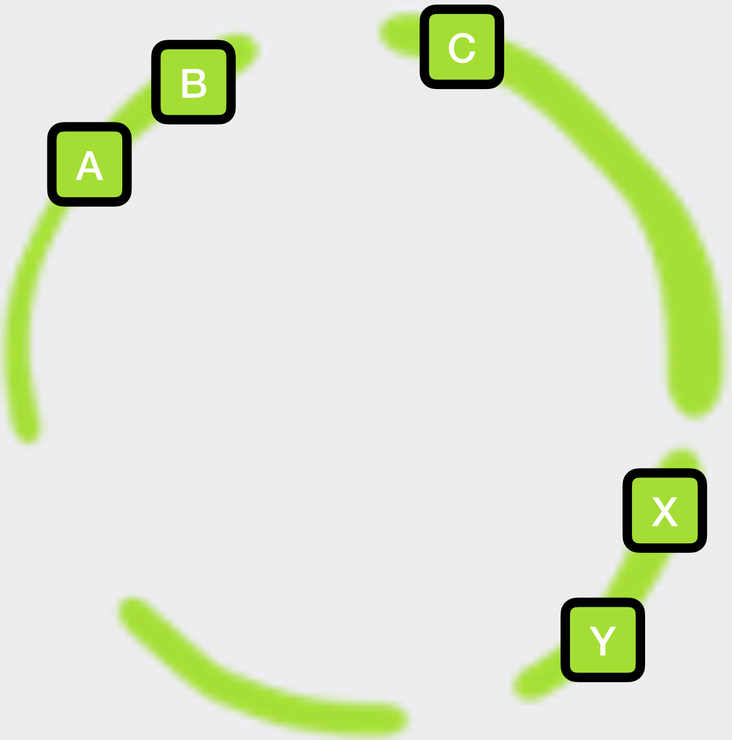


Between 40 and 120 for Bacteria/Archaea depending on cutoffs

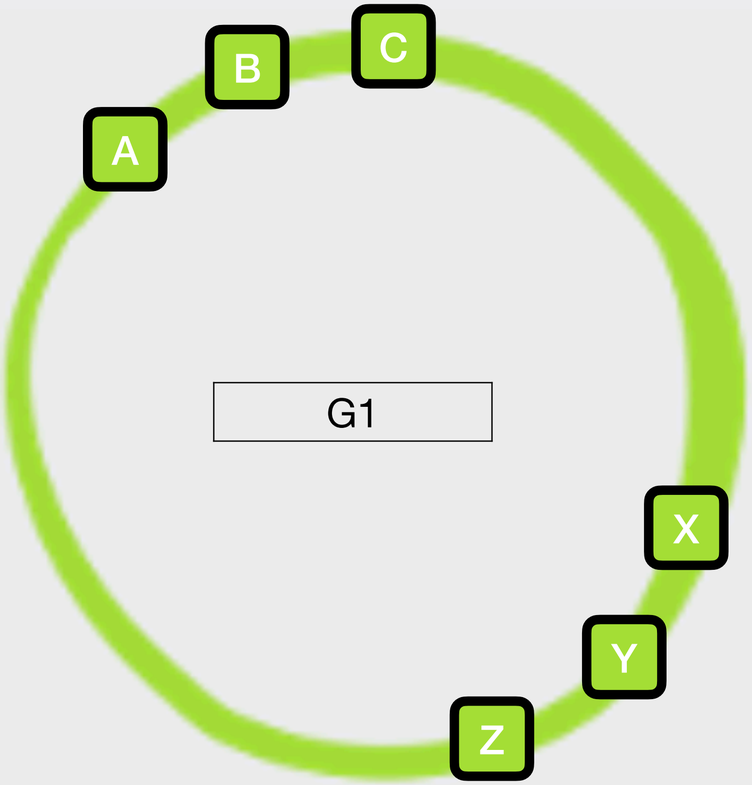
Universal single-copy marker genes

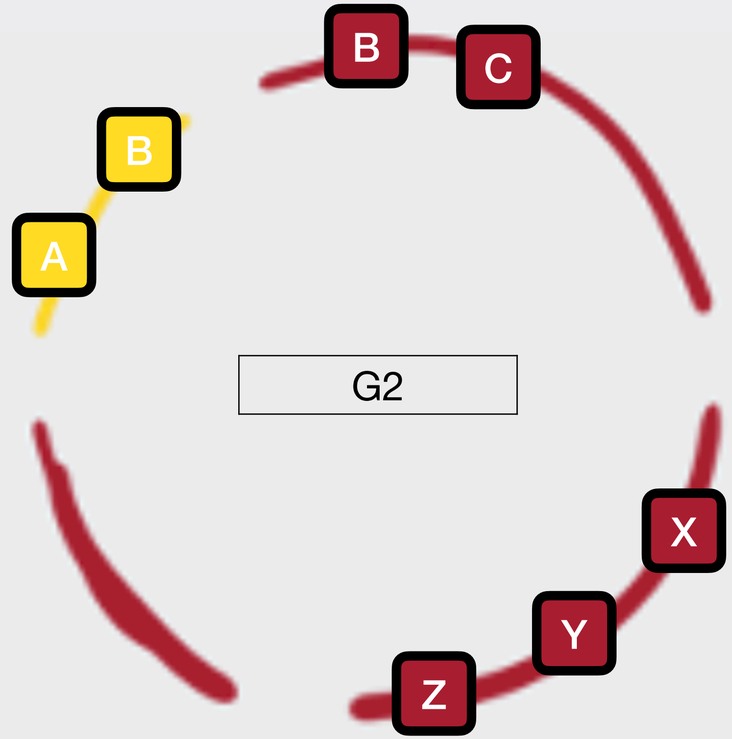


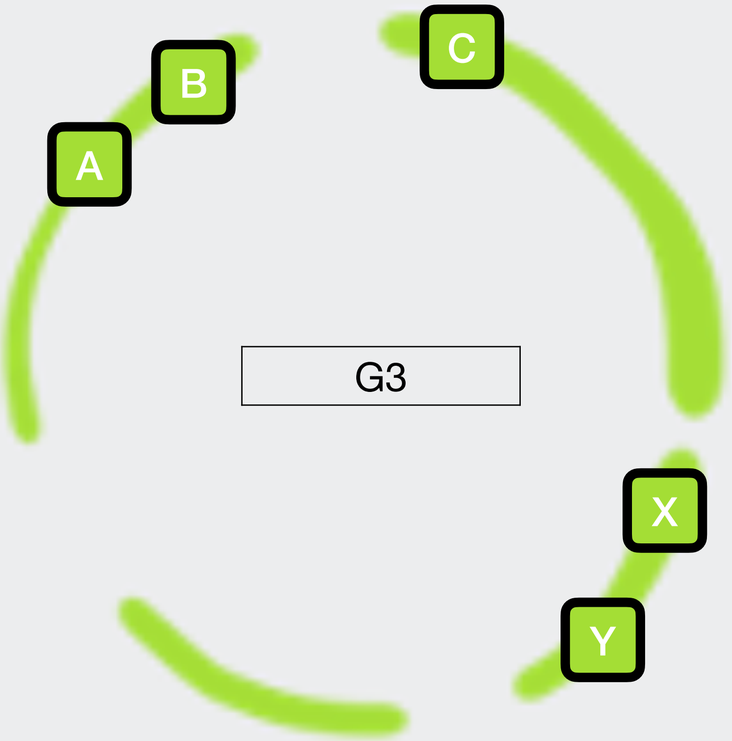




Universal single-copy marker genes



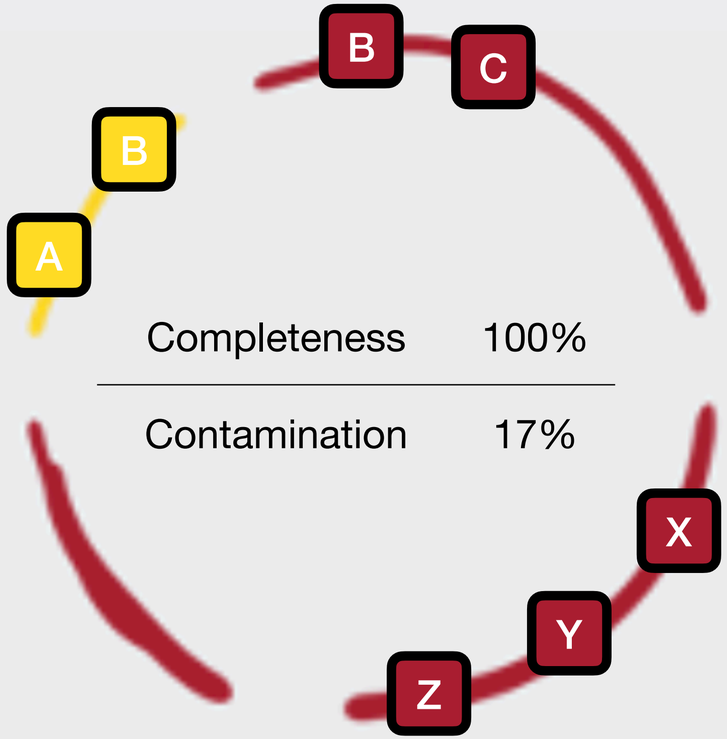




	G1	G2	G3
A	1	1	1
B	1	2	1
C	1	1	1
X	1	1	1
Y	1	1	1
Z	1	1	0

Universal single-copy marker genes

Completeness	100%
Contamination	0%



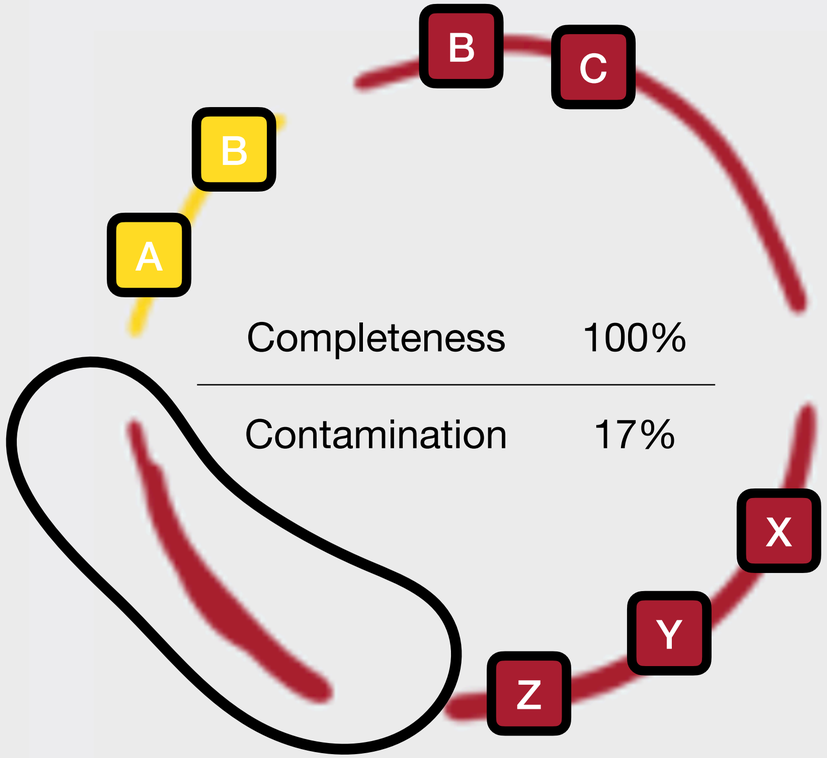
Completeness	100%
Contamination	17%

Completeness	83%
Contamination	0%

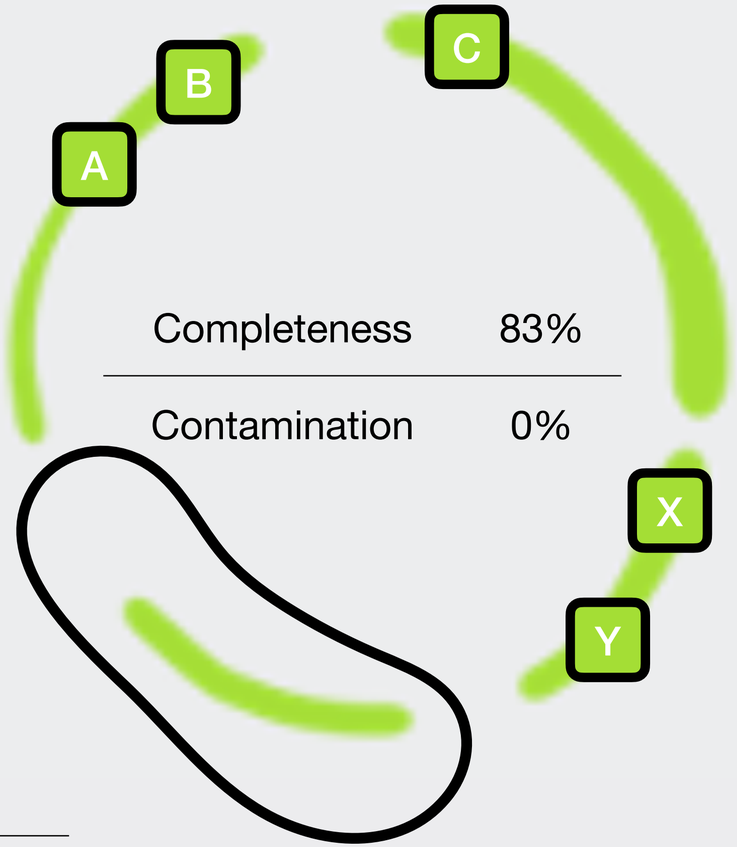
	G1	G2	G3
A	1	1	1
B	1	2	1
C	1	1	1
X	1	1	1
Y	1	1	1
Z	1	1	0

Universal single-copy marker genes

Completeness	100%
Contamination	0%



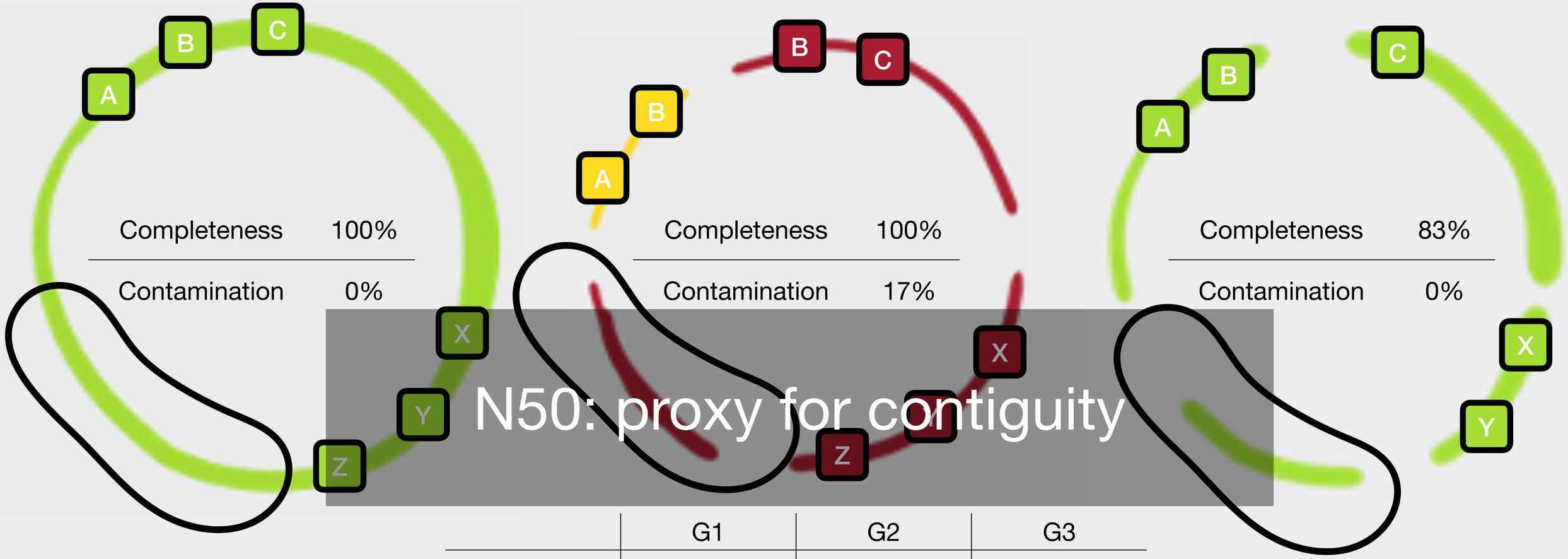
Completeness	100%
Contamination	17%



Completeness	83%
Contamination	0%

	G1	G2	G3
A	1	1	1
B	1	2	1
C	1	1	1
X	1	1	1
Y	1	1	1
Z	1	1	0

Universal single-copy marker genes



N50: proxy for contiguity

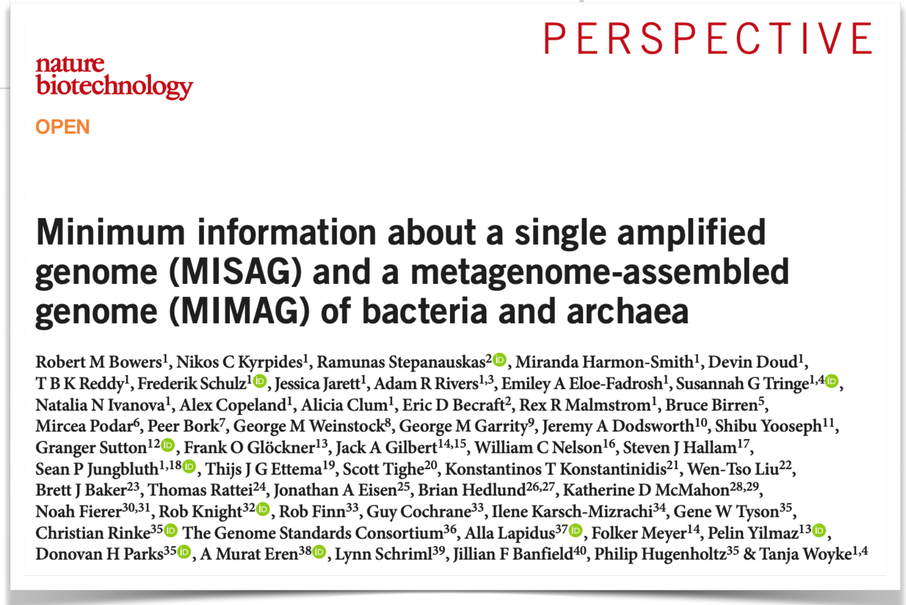
	G1	G2	G3
A	1	1	1
B	1	2	1
C	1	1	1
X	1	1	1
Y	1	1	1
Z	1	1	0

In practice: evaluation software

- **CheckM**
- **Anvi’o**
- **BUSCO**
- **…**

Critically thinking about community standards

High-quality draft (SAG/MAG)	
Assembly quality[a]	Multiple fragments where gaps span repetitive regions. Presence of the 23S, 16S, and 5S rRNA genes and at least 18 tRNAs.
Completion[b]	>90%
Contamination[c]	<5%
Medium-quality draft (SAG/MAG)	
Assembly quality[a]	Many fragments with little to no review of assembly other than reporting of standard assembly statistics.
Completion[b]	≥50%
Contamination[c]	<10%



PERSPECTIVE

nature biotechnology

OPEN

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

Robert M Bowers[1], Nikos C Kyrpides[1], Ramunas Stepanauskas[2], Miranda Harmon-Smith[1], Devin Doud[1], T B K Reddy[1], Frederik Schulz[1], Jessica Jarett[1], Adam R Rivers[1,3], Emiley A Eloe-Fadrosh[1], Susannah G Tringe[1,4], Natalia N Ivanova[1], Alex Copeland[1], Alicia Clum[1], Eric D Becraft[2], Rex R Malmstrom[1], Bruce Birren[5], Mircea Podar[6], Peer Bork[7], George M Weinstock[8], George M Garrity[9], Jeremy A Dodsworth[10], Shibu Yooseph[11], Granger Sutton[12], Frank O Glöckner[13], Jack A Gilbert[14,15], William C Nelson[16], Steven J Hallam[17], Sean P Jungbluth[1,18], Thijs J G Ettema[19], Scott Tighe[20], Konstantinos T Konstantinidis[21], Wen-Tso Liu[22], Brett J Baker[23], Thomas Rattei[24], Jonathan A Eisen[25], Brian Hedlund[26,27], Katherine D McMahon[28,29], Noah Fierer[30,31], Rob Knight[32], Rob Finn[33], Guy Cochrane[33], Ilene Karsch-Mizrachi[34], Gene W Tyson[35], Christian Rinke[35] The Genome Standards Consortium[36], Alla Lapidus[37], Folker Meyer[14], Pelin Yilmaz[13], Donovan H Parks[35], A Murat Eren[38], Lynn Schriml[39], Jillian F Banfield[40], Philip Hugenholtz[35] & Tanja Woyke[1,4]

Helena Klein

Genome-based classification and taxonomy

RESOURCE

nature biotechnology

A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life

Donovan H Parks, Maria Chuvochina, David W Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil & Philip Hugenholtz

Taxonomy is an organizing principle of biology and is ideally based on evolutionary relationships among organisms. Development of a robust bacterial taxonomy has been hindered by an inability to obtain most bacteria in pure culture and, to a lesser extent, by the historical use of phenotypes to guide classification. Culture-independent sequencing technologies have matured sufficiently that a comprehensive genome-based taxonomy is now possible. We used a concatenated protein phylogeny as the basis for a bacterial taxonomy that conservatively removes polyphyletic groups and normalizes taxonomic ranks on the basis of relative evolutionary divergence. Under this approach, 58% of the 94,759 genomes comprising the Genome Taxonomy Database had changes to their existing taxonomy. This result includes the description of 99 phyla, including six major monophyletic units from the subdivision of the Proteobacteria, and amalgamation of the Candidate Phyla Radiation into a single phylum. Our taxonomy should enable improved classification of uncultured bacteria and provide a sound basis for ecological and evolutionary studies.

Bioinformatics, 36(6), 2020, 1925–1927
doi: 10.1093/bioinformatics/btz848
Advance Access Publication Date: 15 November 2019
Applications Note

OXFORD

Genome analysis

GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database

Pierre-Alain Chaumeil*, Aaron J. Mussig, Philip Hugenholtz and Donovan H. Parks*

Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, St Lucia, QLD 4072, Australia

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on July 12, 2019; revised on October 15, 2019; editorial decision on November 11, 2019; accepted on November 13, 2019

RESOURCE

https://doi.org/10.1038/s41564-022-01214-9

nature microbiology

Check for updates

OPEN

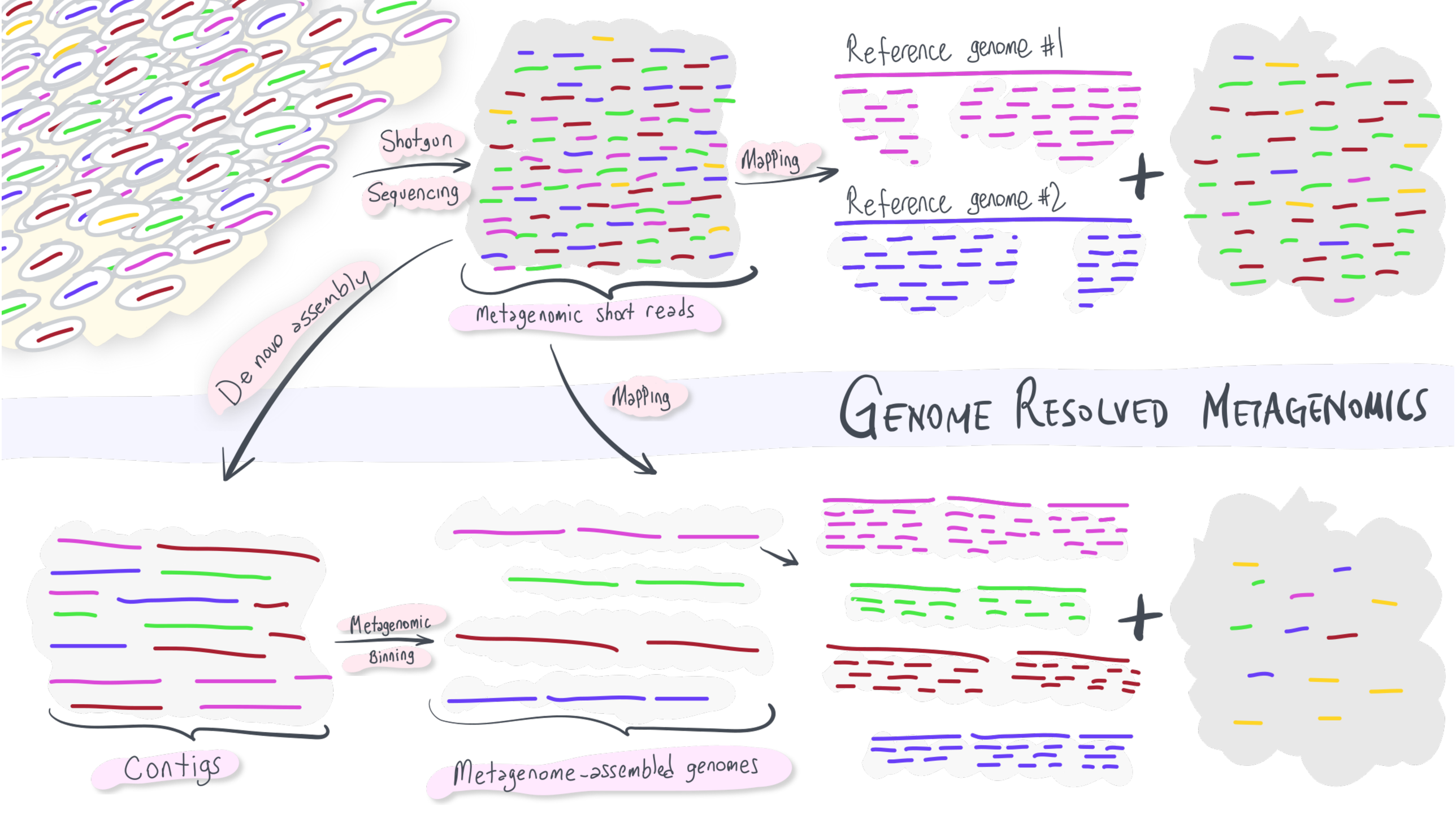
SeqCode: a nomenclatural code for prokaryotes described from sequence data

Brian P. Hedlund[1], Maria Chuvochina[2], Philip Hugenholtz[2], Konstantinos T. Konstantinidis[3], Alison E. Murray[4], Marike Palmer[1], Donovan H. Parks[2], Alexander J. Probst[5], Anna-Louise Reysenbach[6], Luis M. Rodriguez-R[7], Ramon Rossello-Mora[8], Iain C. Sutcliffe[9], Stephanus N. Venter[10] and William B. Whitman[11]

Most prokaryotes are not available as pure cultures and therefore ineligible for naming under the rules and recommendations of the International Code of Nomenclature of Prokaryotes (ICNP). Here we summarize the development of the SeqCode, a code of nomenclature under which genome sequences serve as nomenclatural types. This code enables valid publication of names of prokaryotes based upon isolate genome, metagenome-assembled genome or single-amplified genome sequences. Otherwise, it is similar to the ICNP with regard to the formation of names and rules of priority. It operates through the SeqCode Registry (https://seqco.de/), a registration portal through which names and nomenclatural types are registered, validated and linked to metadata. We describe the two paths currently available within SeqCode to register and validate names, including *Candidatus* names, and provide examples for both. Recommendations on minimal standards for DNA sequences are provided. Thus, the SeqCode provides a reproducible and objective framework for the nomenclature of all prokaryotes regardless of cultivability and facilitates communication across microbiological disciplines.

Cultivation-independent genome-resolved metagenomics: a summary

Helena Klein


Shotgun
Sequencing
Metagenomic short reads
Mapping
Reference genome #1
Reference genome #2
De novo assembly
Mapping
GENOME RESOLVED METAGENOMICS
Contigs
Metagenomic
Binning
Metagenome-assembled genomes

Material

This course uses a lot of material from https://merenlab.org/momics/, I invite you to have a look.

If you want details on the bioinformatics behind you can start by having a look here: https://astrobiomike.github.io/genomics/metagen_anvio

A. Murat Eren (Meren) (PI)

☐ Web ☐ Email ☐ Twitter ☐ LinkedIn ☐ Github ☐ ORCiD

☐ Address: Knapp Center for Biomedical Discovery, 900 E. 57th St., MB 9, RM 9118, Chicago, IL 60637 USA

☐ Phone: +1-773-702-5935 ☐ Fax: +1-773-702-2281

I am a computer scientist with a deep appreciation for the complexity of life. I design algorithms and experiments to better understand microbes and their ecology. [photos: 1, 2, 3].

» MBL Fellow, Marine Biological Laboratory.
» Assistant Professor, The Department of Medicine at the University of Chicago.
» Committee on Microbiology, The Biomedical Sciences Cluster at the University of Chicago.

Mike Lee

☐ Web ☐ Email ☐ Twitter ☐ LinkedIn ☐ Github

» NASA Space Biology Fellow, NASA Ames Research Center.
» JCVI Research Fellow, J. Craig Venter Institute.

✍ Combining reference genome annotations with your own in pangenomes (Sat, Dec 01, 2018)
✍ Anvi'o 'views' demystified (Mon, May 08, 2017)
✍ Making anvi'o use your own HMM collection (Sat, May 21, 2016)

Questions?

Image: François Aurat