Encapsulated Bacteria Session 6: *De novo* Genome Assembly

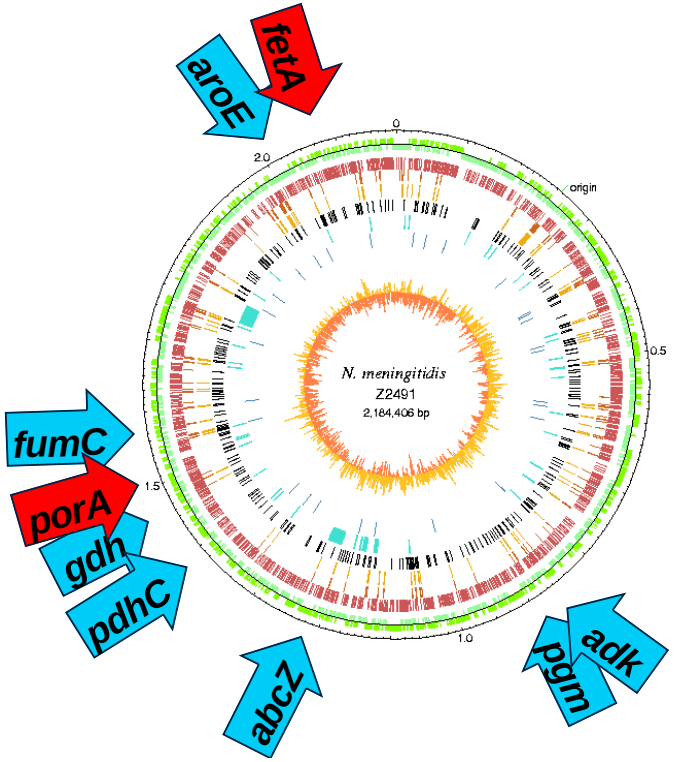
Genomics and Clinical Microbiology 2024

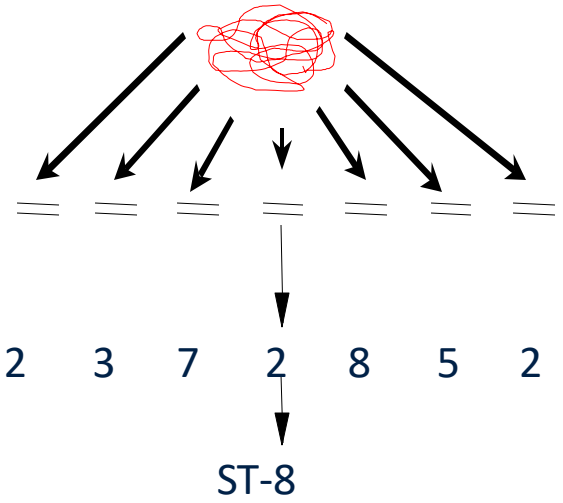
Martin Maiden, Made Krisna, Kasia Parfitt, Keith Jolley

Department of Biology

UNIVERSITY OF
OXFORD

First generation genomic typing: single locus and MLST





B: P1.7,16: F5-1: ST-33 (cc32)

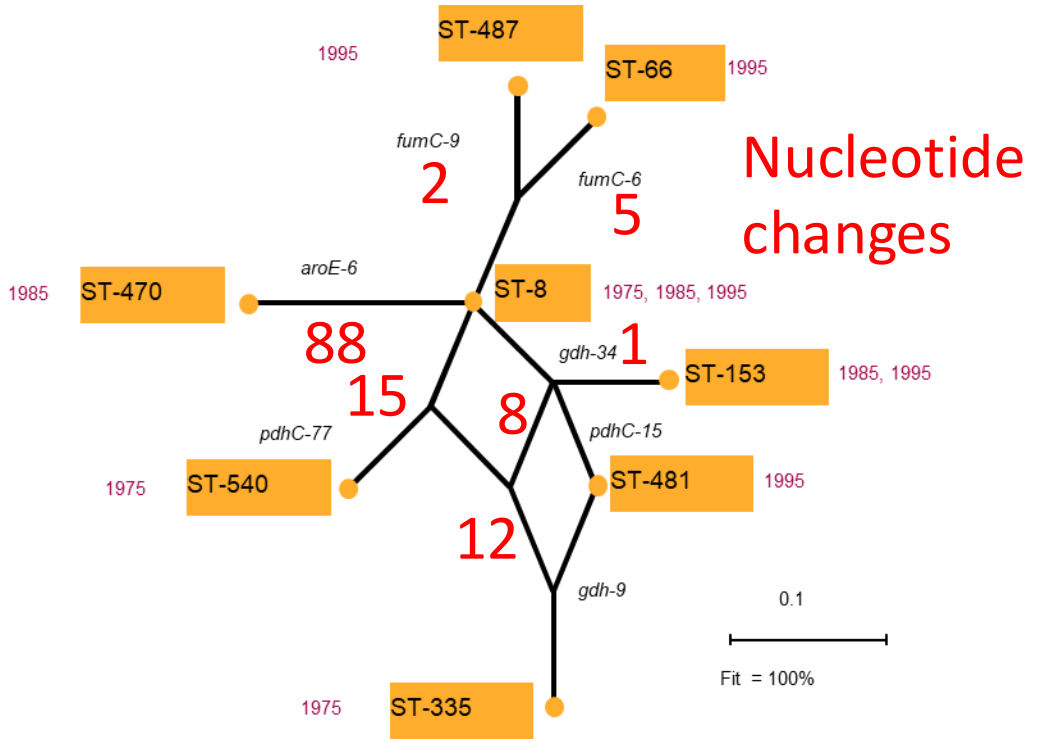
Antigen type (fine type)

Sequence type & clonal complex

Maiden, MCJ, Bygraves, JA, Feil, E, Morelli, G, Russell, JE, Urwin, R, Zhang, Q, Zhou, J, Zurth, K, Caugant, DA, Feavers, IM, Achtman, M & Spratt, BG. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci USA* 95, 3140-3145.

MLST: allele-based analyses, sequence types (STs) and clonal complexes (ccs)

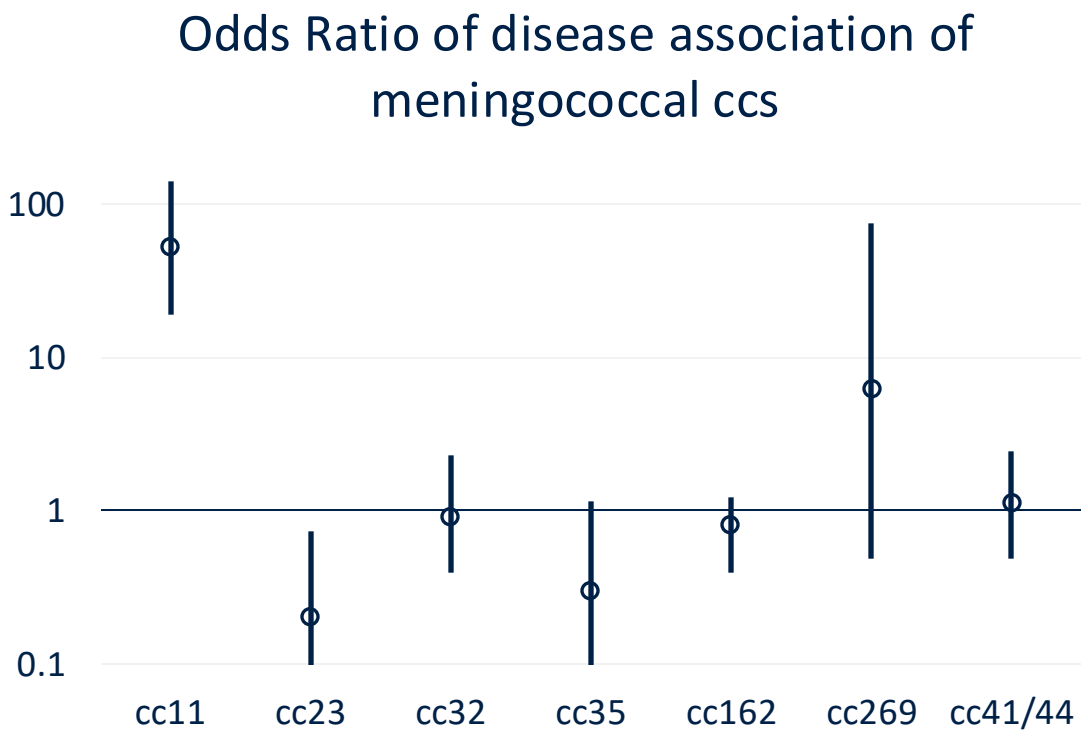
ST	*adk*	*abcZ*	*aroE*	*fumC*	*gdh*	*pdhC*	*pgm*
8	2	3	7	2	8	5	2
66	2	3	7	6	8	5	2
153	2	3	7	2	34	5	2
335	2	3	7	2	9	15	2
470	2	3	6	2	8	5	2
481	2	3	7	2	9	5	2
487	2	3	7	9	8	5	2
540	2	3	7	2	8	77	2



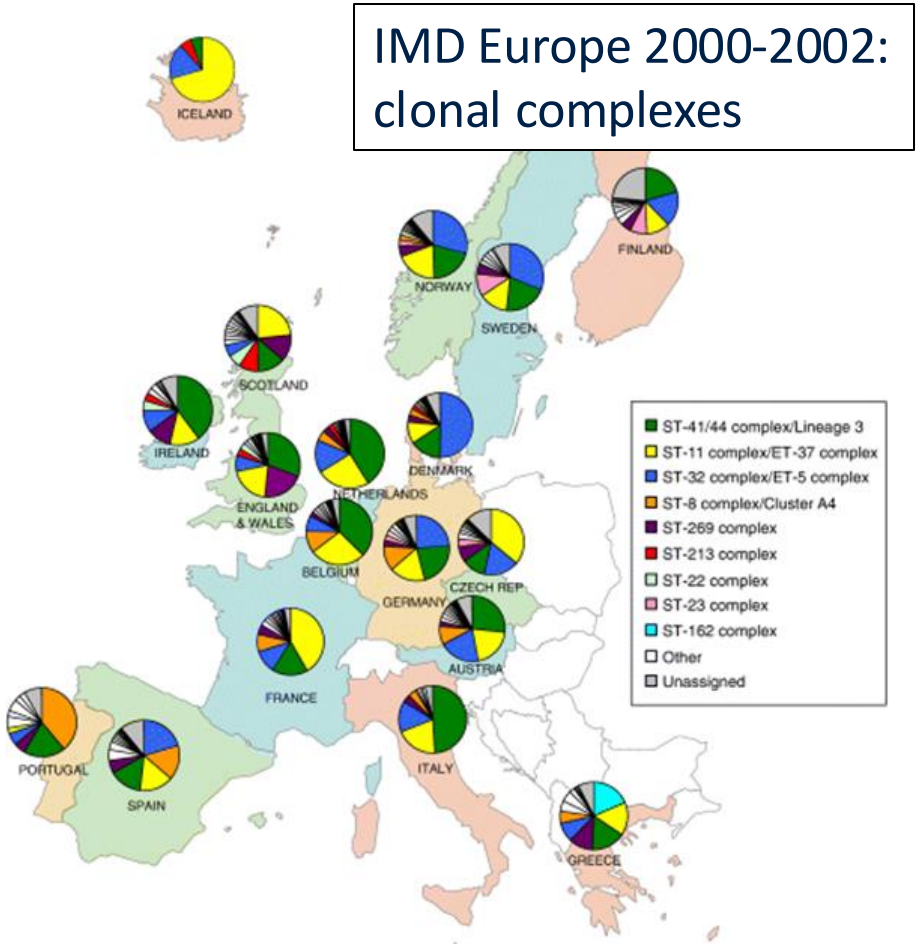
ST-8 Clonal Complex: cc8

Russell, J. E., Urwin, R., Gray, S. J., Fox, A. J., Feavers, I. M. & Maiden, M. C. (2008). Molecular epidemiology of meningococcal disease in England and Wales 1975-1995, before the introduction of serogroup C conjugate vaccines. *Microbiology* 154, 1170-1177.

Meningococcal hyperinvasive clonal complexes

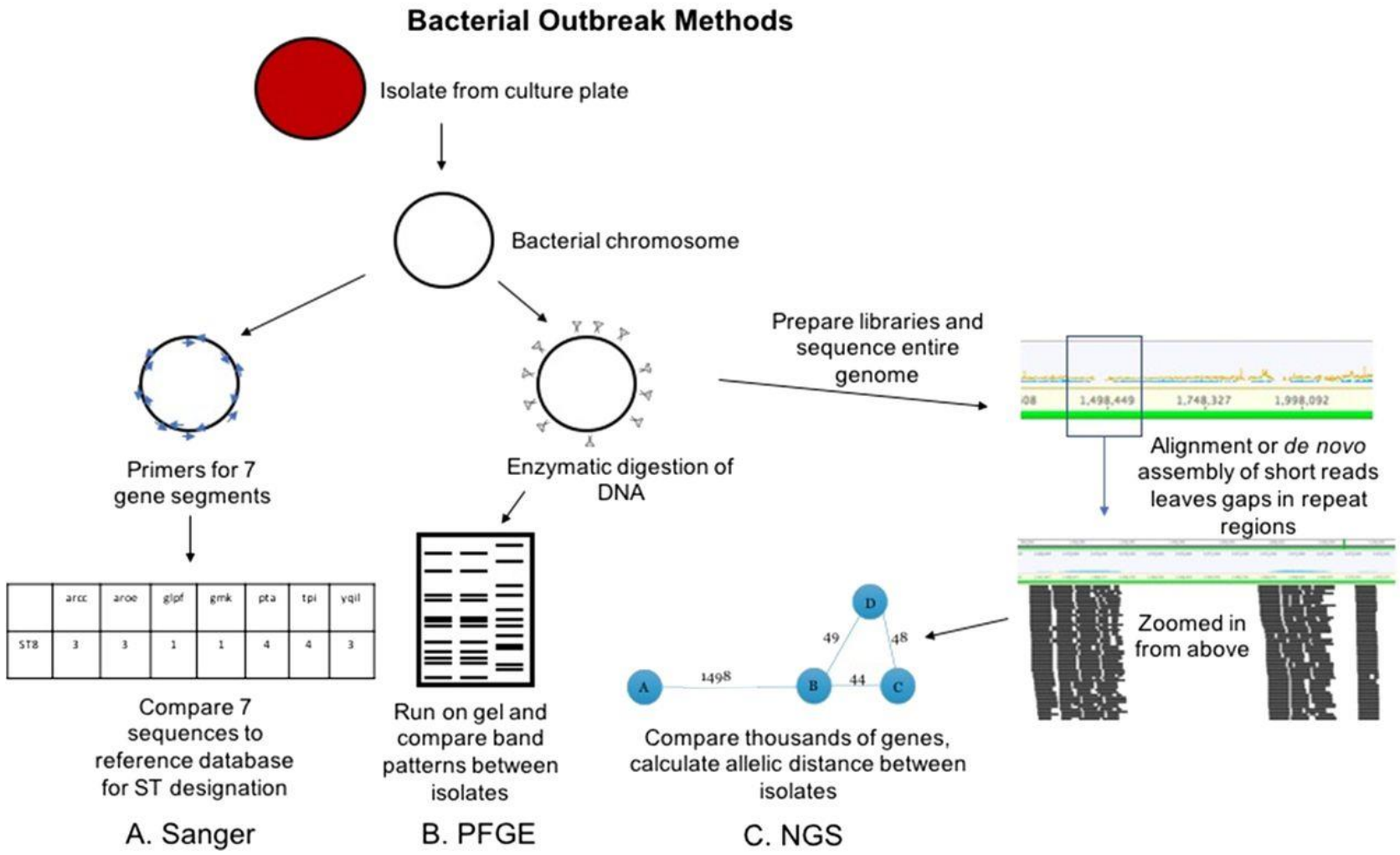


Yazdankhah, S. P., Kriz, P., Tzanakaki, G., Kremastinou, J., Kalmusova, J., Musilek, M., Alvestad, T., Jolley, K. A., Wilson, D. J., McCarthy, N. D., Caugant, D. A. & Maiden, M. C. (2004). Distribution of serogroups and genotypes among disease-associated and carried isolates of *Neisseria meningitidis* from the Czech Republic, Greece, and Norway. *J Clin Microbiol* **42**, 5146-5153.



Brehony, C., Jolley, K. A. & Maiden, M. C. (2007). Multilocus sequence typing for global surveillance of meningococcal disease. *FEMS Microbiol Rev* **31**, 15-26.

Conventional and NGS methods for bacterial characterisation.



Dougherty, C.E. & Graf, E. (2019). Conventional and NGS methods for bacterial outbreak investigation. *Clin. Lab. Sci.* **32**, 70-77.

Whole genome *de novo* sequence pipeline

Isolate growth

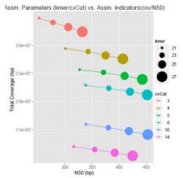
Bacterial cells/clinical specimen

DNA Extraction

Purified DNA

Sequencing (Illumina)

Short-read sequences



de novo assembly (VELVET, SPADES)

Assembled contiguous sequences ('High Quality Draft' Genome)

Database deposition (BIGSDB)

Phenotype & provenance linkage and annotation

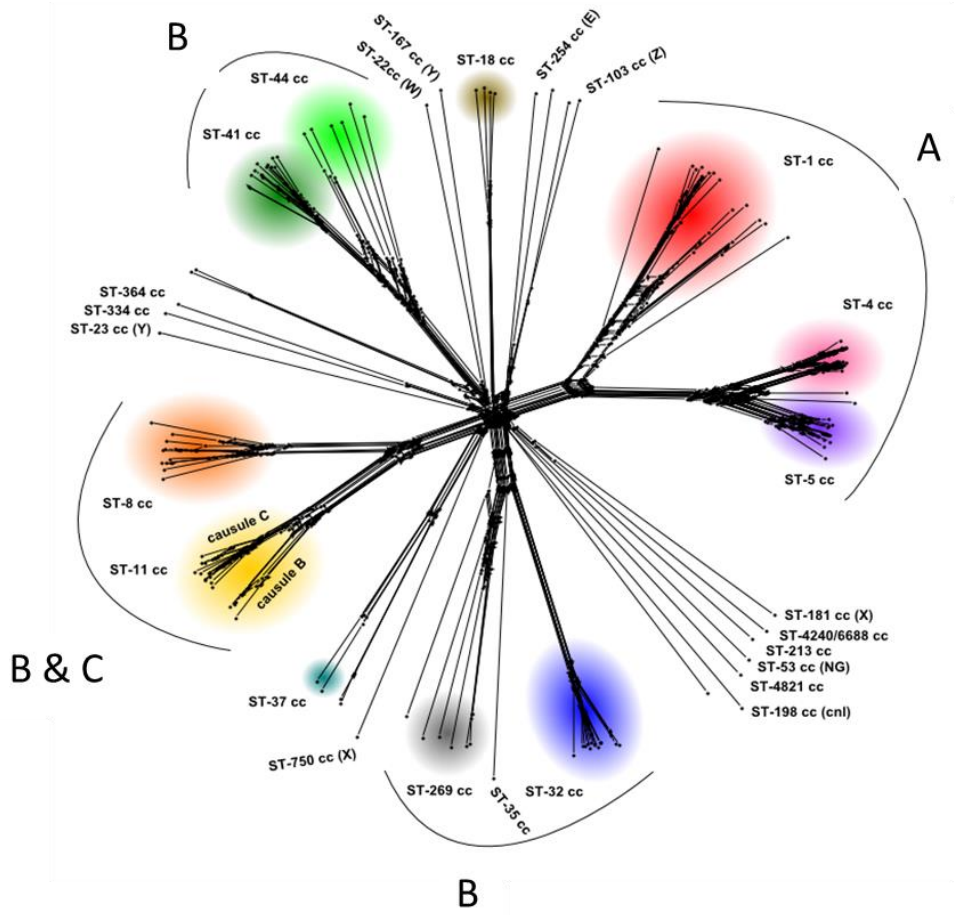
Annotated, curated, web-accessible WGS sequences

'Plain language' data

Bratcher, H. B., Bennett, J. S. & Maiden, M. C. J. (2012). Evolutionary and genomic insights into meningococcal biology. *Future Microbiology* **7**, 873-885.

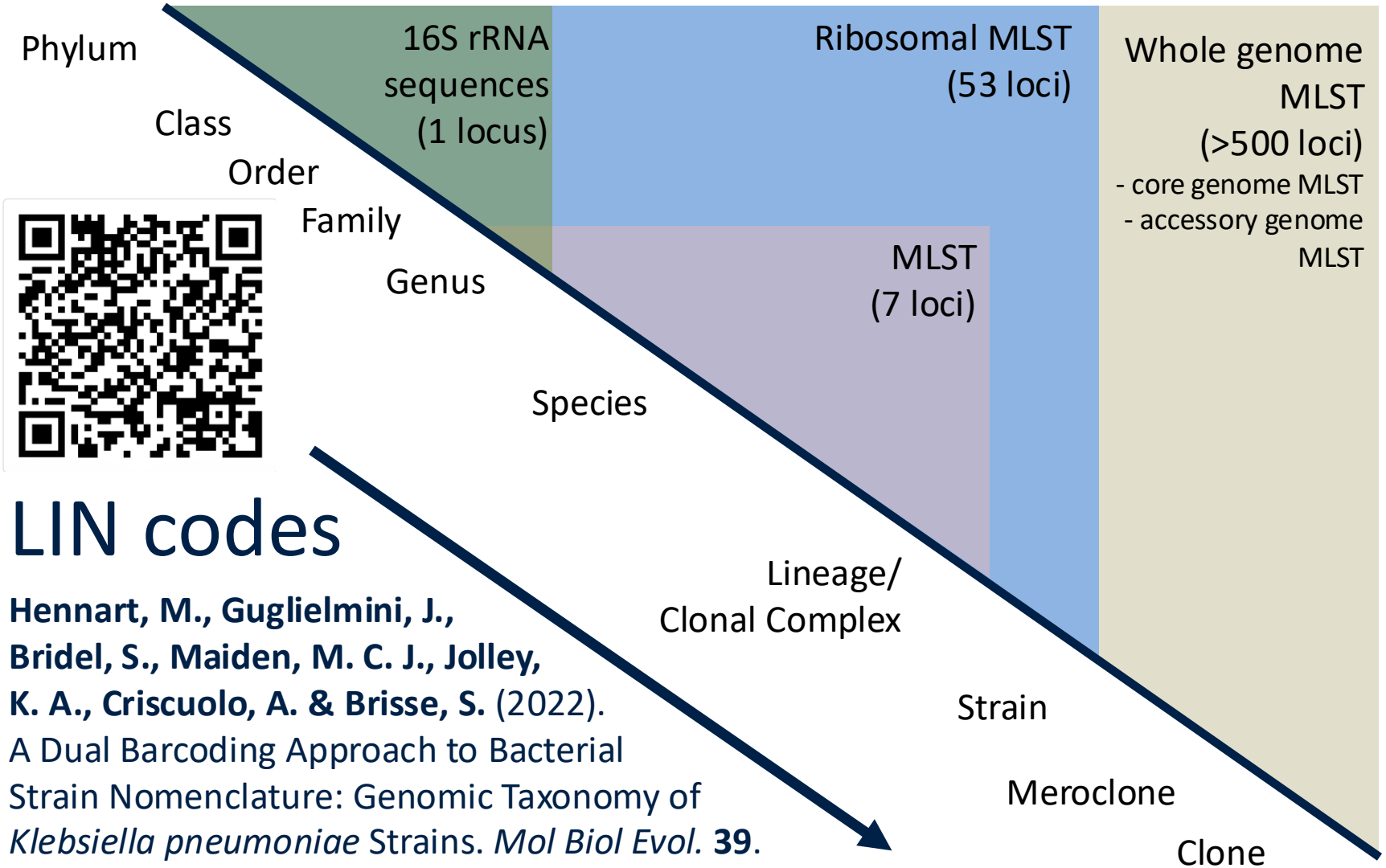
Meningococcal population Structure revealed by genome sequencing

- Meningococcal population structure first studied by Multilocus Enzyme Electrophoresis (MLEE)
  - the concept of hyperinvasive meningococci.
- Replaced by seven locus MLST analyses in 1998, which identifies sequence types (STs) and groups of STs ‘clonal complexes’ (ccs).
- Whole genome analysis identifies lineages, which closely correspond to ccs and hyperinvasive lineages.
- Lineages are associated with vaccine antigens, including capsules (serogroups).



Bratcher, H. B., Corton, C., Jolley, K. A., Parkhill, J., and Maiden, M. C. (2014). A gene-by-gene population genomics platform: *de novo* assembly, annotation and genealogical analysis of 108 representative *Neisseria meningitidis* genomes. *BMC Genomics* **15**, 1138

Conceptual framework: sequence data, nomenclature, phenotype.

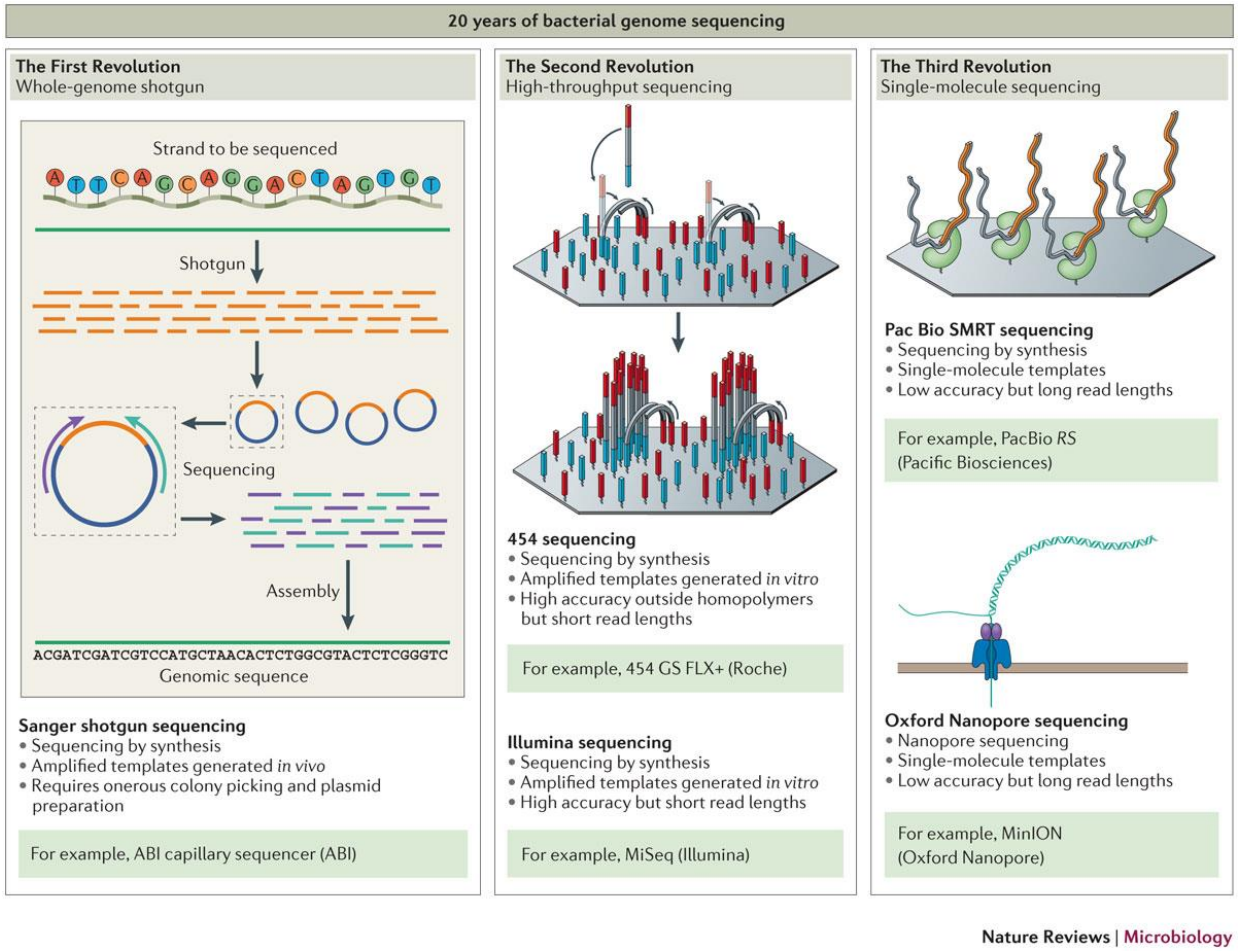


LIN codes

Hennart, M., Guglielmini, J., Bridel, S., Maiden, M. C. J., Jolley, K. A., Criscuolo, A. & Brisse, S. (2022). A Dual Barcoding Approach to Bacterial Strain Nomenclature: Genomic Taxonomy of *Klebsiella pneumoniae* Strains. *Mol Biol Evol.* **39**.

Maiden, M. C. J., Jansen van Rensburg, M. J., Bray, J. E., Earle, S. G., Ford, S. A., Jolley, K. A. & McCarthy, N. D. (2013). MLST revisited: the gene-by-gene approach to bacterial genomics. *Nature Reviews Microbiology* **11**, 728-736.

Whole genome sequencing technologies



Loman, N. J. & Pallen, M. J. (2015). Twenty years of bacterial genome sequencing. *Nat Rev Microbiol.* **13**, 787-794.

Approaches to Genome Assembly

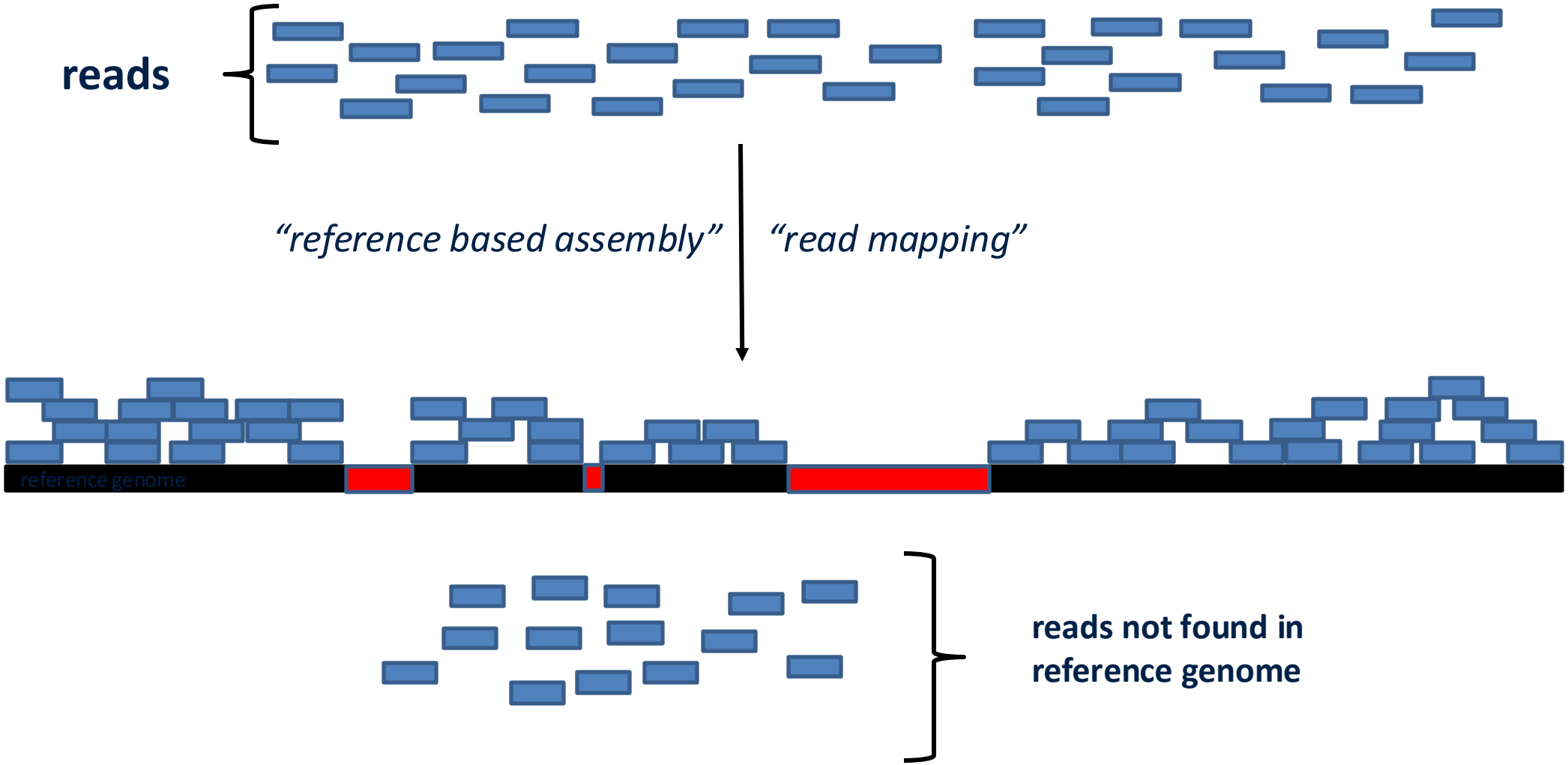
Comparative, reference-based assembly:

- using a reference genome from the same organism or a closely related sample is used to guide the assembly process by aligning the reads;
- primary use is for resequencing applications.

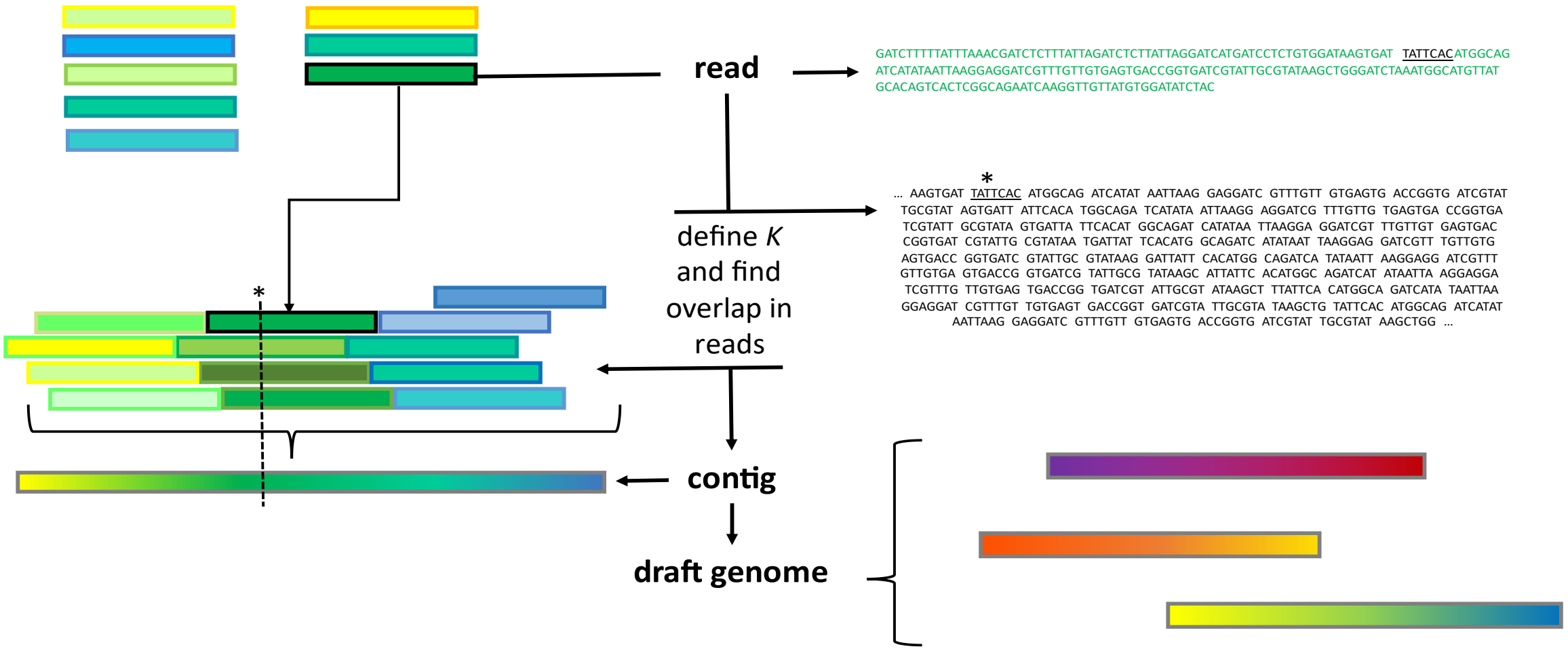
***de novo* assembly:**

- assembly in the strictest sense, no map or guidance is used for assembling the genome;
- used to assemble genomes that have not been previously sequenced.

Read mapping, reference-based assembly



De novo assembly of a draft genome



Assembly gaps

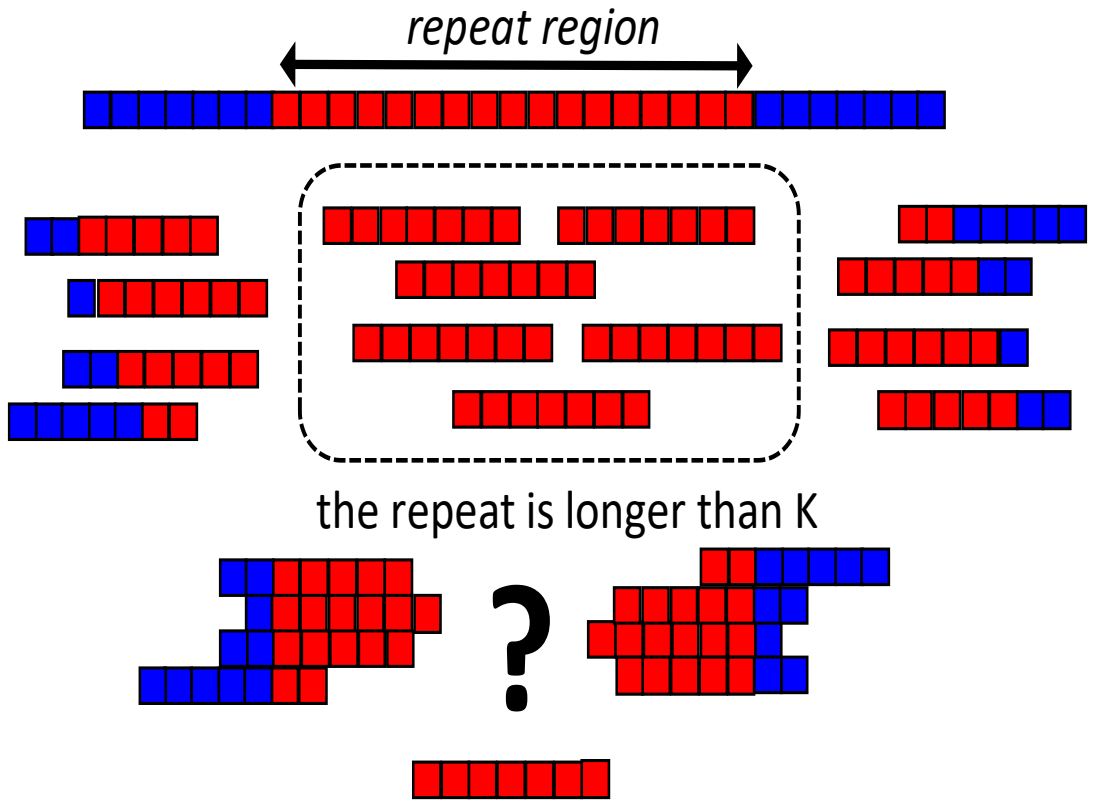
Biased base content:

- AT and GC rich regions are under-represented;
- other chemistry quirks.

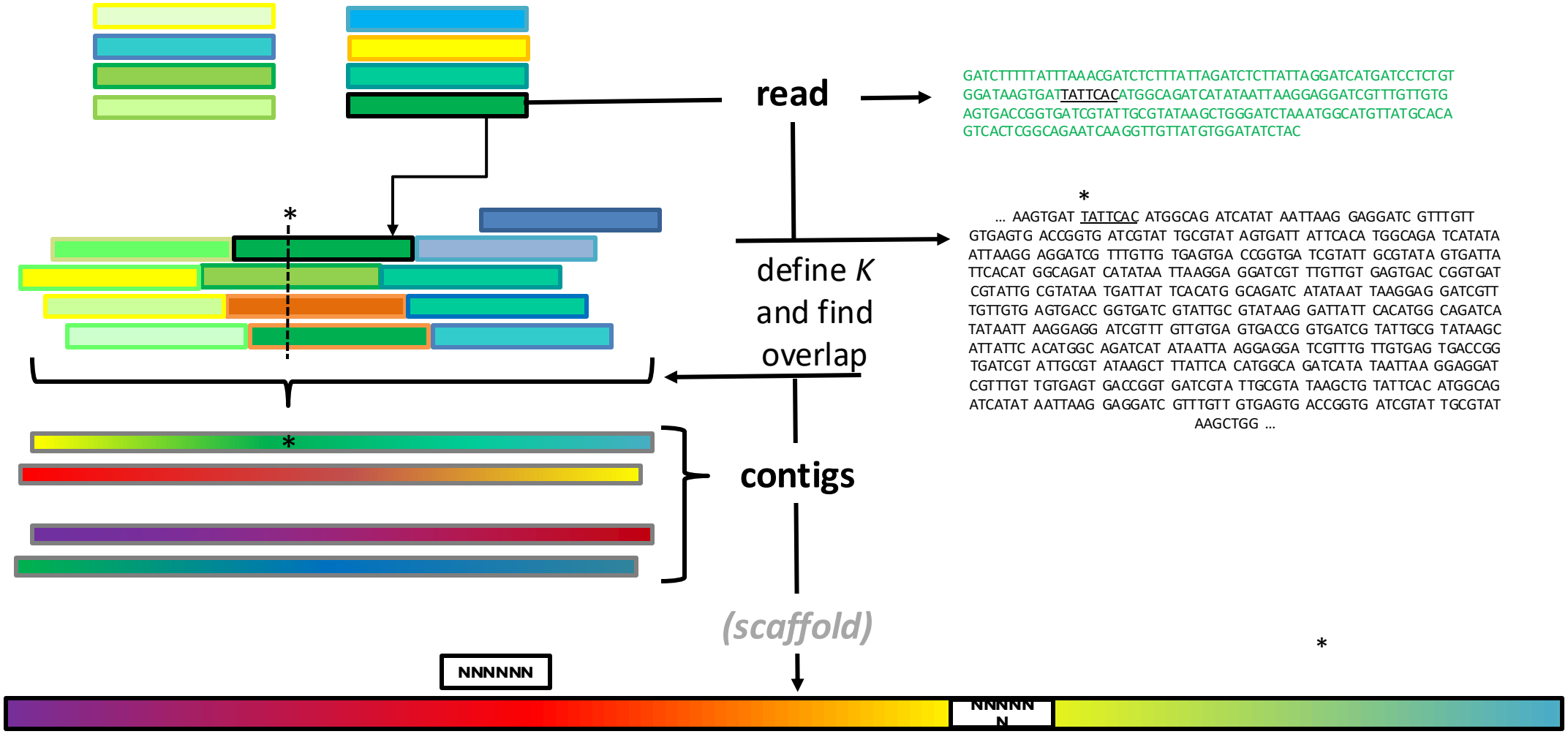
Sequence compression due to repetitive sequences.

➢Increased depth needed:

- ✓ sequencing errors,
- ✓ polymorphic sites.

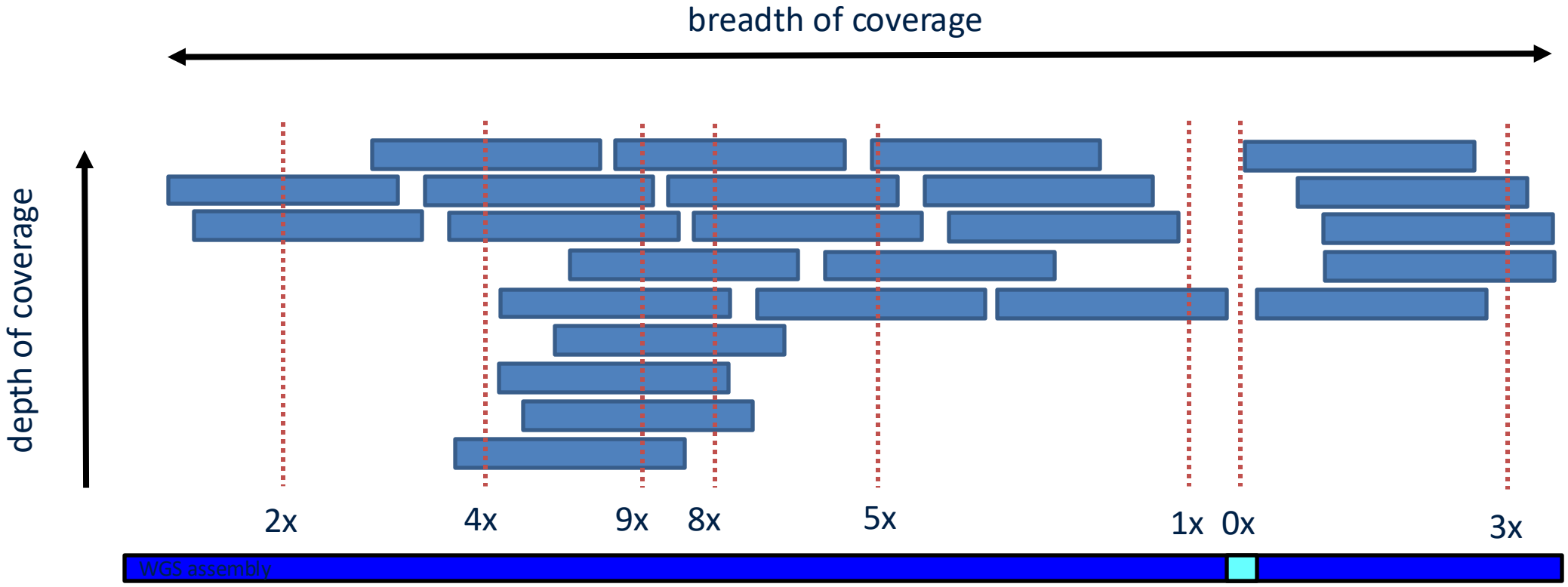


Scaffolding of a draft genome



groups contigs in order but this is a hypothesis that the scaffold is correct

Uniformity of genome coverage



COVERAGE: fraction of the genome sequenced by at least one read
DEPTH: average number of reads that cover any given region
UNIFORMITY: measures the evenness of the coverage depth across the genome

Some quality metrics

Metric	Definition
Number of contigs	Total number of contigs in an assembly
Total length	Combined length of all contigs
Maximum length	Largest contig
Mean length	Average contig size
Coverage	Fraction of the genome sequenced by at least one read
Uniformity	Evenness of the coverage depth across the genome
Depth	Average number of reads that cover any given region
N50	Length of the shortest contig for which longer and equal length contigs cover at least 50 % of the assembly
L50	Number of sequence contigs that are longer than, or equal to, the N50 length and therefore include half the bases of the assembly
N90	Length of the shortest contig for which longer and equal length contigs cover at least 90 % of the assembly