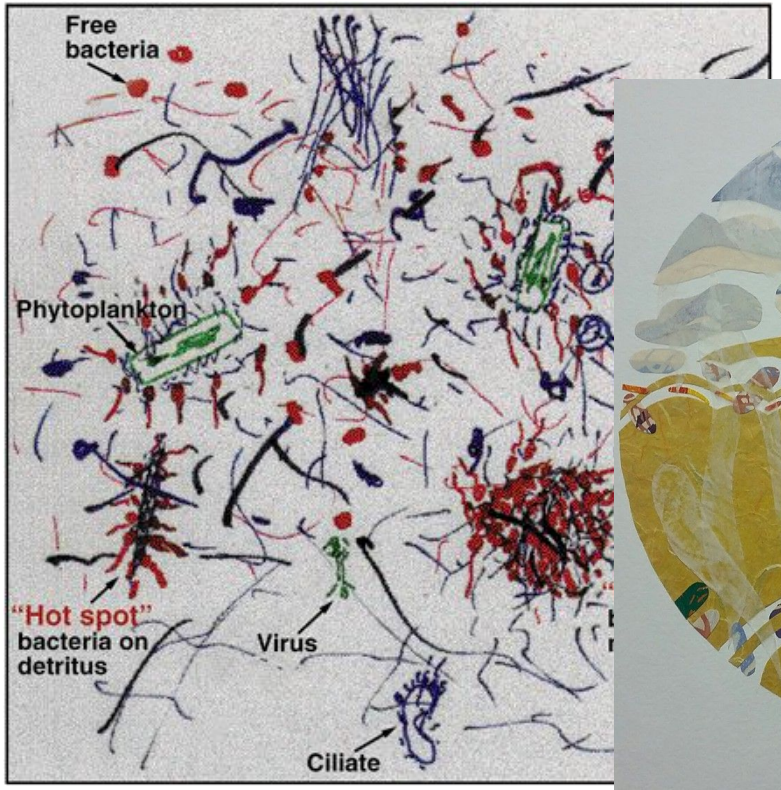
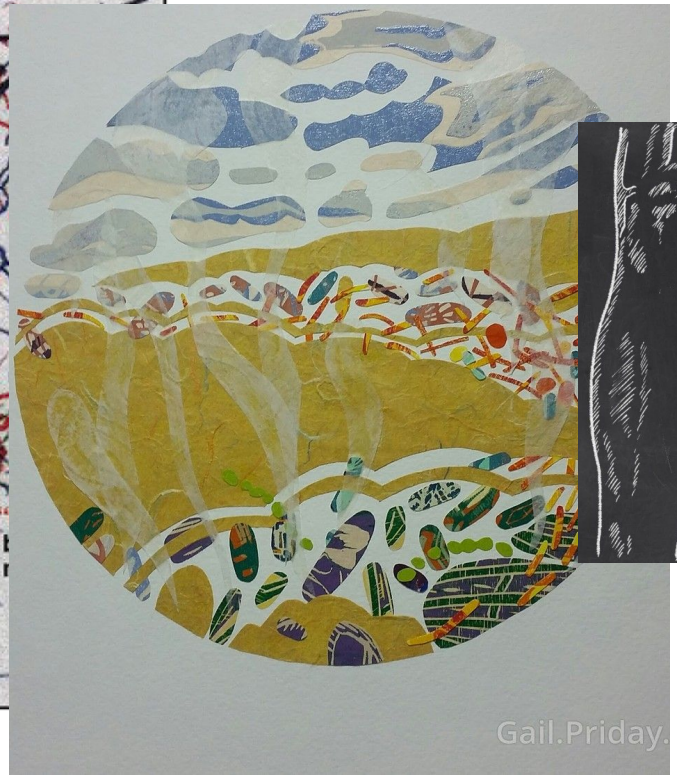
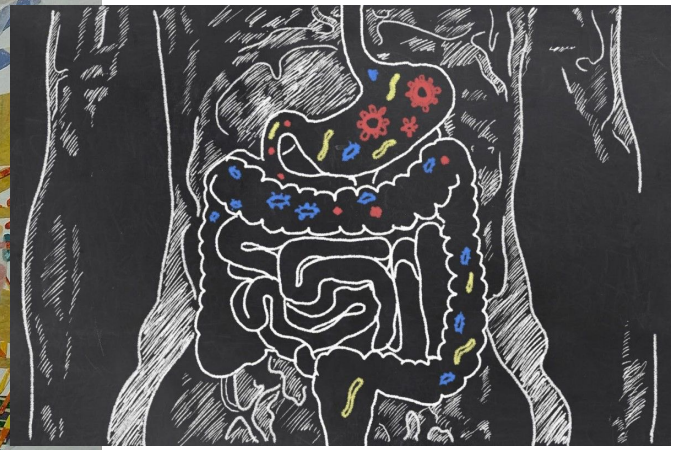
Metagenomics Lesson 2







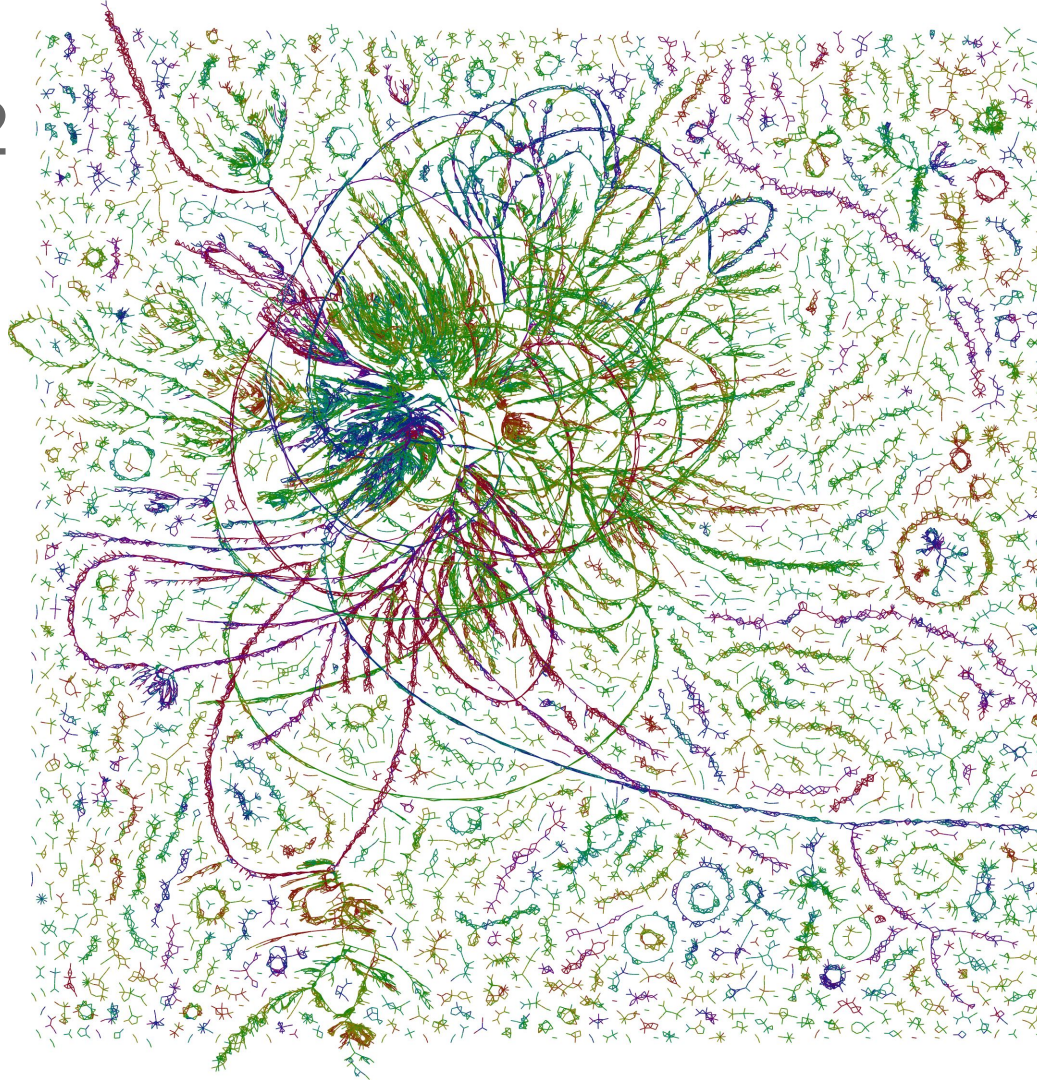
BVCN

biovcnet.github.io

Metagenomics Lesson 2

Taxonomic Classification using k-mers

1. The power of k-mers
2. Choosing a classifier
3. Demo: Adapter trimming
4. Demo: Kraken2



A “k-mer” is a word of DNA that is *k* bases long

4^1 1-mers: **A, T, C, G**

4^2 2-mers: **AA, AT, AC, AG, TA, TT, TC, TG, CA, CT, CC, CG, GA, GT, GC, GG**

4^3 3-mers:

All codons are 3-mers but not all 3-mers are codons

UUU, UUC: Phe UUA, UUG: Leu	UCU, UCC, UCA, UCG: Ser	UAU, UAC: Tyr UAA, UAG: Stop	UGU, UGC: Cys UGA: Stop UGG: Trp
CUU, CUC, CUA, CUG: Leu	CCU, CCC, CCA, CCG: Pro	CAU, CAC: His CAA, CAG: Gln	CGU, CGC, CGA, CGG: Arg
AUU, AUC, AUA: Ile AUG: Met	ACU, ACC, ACA, ACG: Thr	AAU, AAC: Asn AAA, AAG: Lys	AGU, AGC: Ser AGA, AGG: Arg
GUU, GUC, GUA, GUG: Val	GCU, GCC, GCA, GCG: Ala	GAU, GAC: Asp GAA, GAG: Glu	GGU, GGC, GGA, GGG: Gly

Exponential growth of nucleotide and amino acid k-mers

$4^1 = 4$

$4^2 = 16$

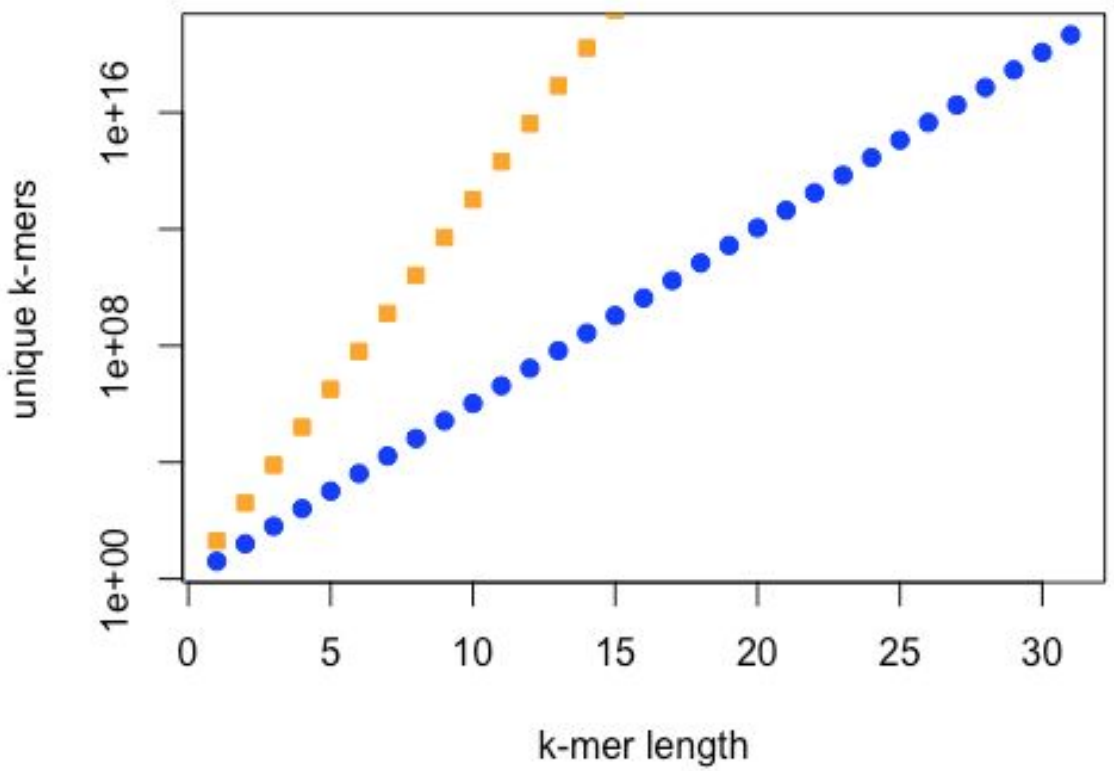
$4^3 = 64$

$4^4 = 256$

$4^8 = 65536$

$4^{16} = 4294967296$

$4^{32} = 18446744073709551616$



Extracting kmers

Typically k-mers are extracted by running a k-length window across all of the reads and sequences.

Given a sequence of length 16, you could extract 11 k-mers of length 6 from it like so:

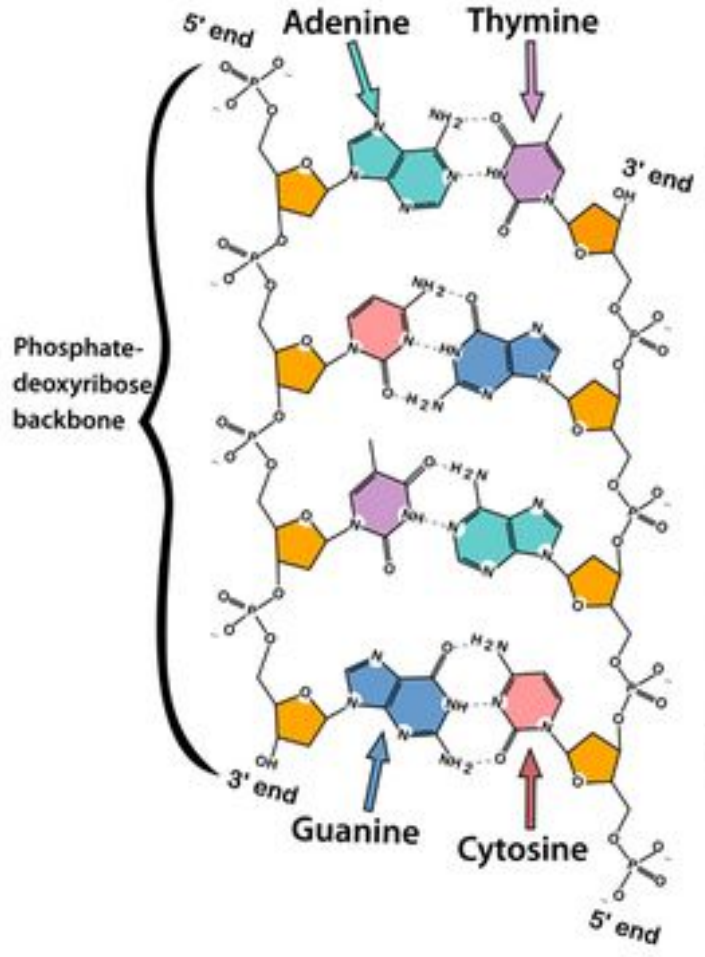
```
AGGATGAGACAGATAG

AGGATG
 GGATGA
  GATGAG
   ATGAGA
    TGAGAC
     GAGACA
      AGACAG
       GACAGA
        ACAGAT
         CAGATA
          AGATAG
```

Complementarity

k-mer analysis is often simplified by storing (“hashing” or “indexing”) only the lexicographically lower reverse complement (the “canonical” string), i.e. we can equally choose to store **`ACTG`** or **`CAGT`** in the picture to the right (reading 5’ to 3’).

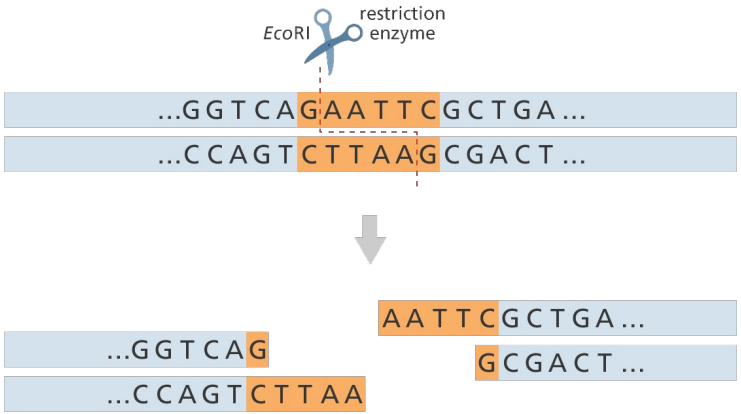
`TTTT → AAAA`
`TATA → TATA`
`CGAT → ATCG`
`TGAC → GTCA`



Palindromes

In molecular biology, a palindrome is “a DNA sequence that is its own reverse complement”.

Restriction enzyme cut sites are common examples of palindromes, e.g. *EcoRI:*



Some tools (particularly assemblers) require odd-length k-mers to avoid palindromes, which induce self-loops in de Bruijn graphs.

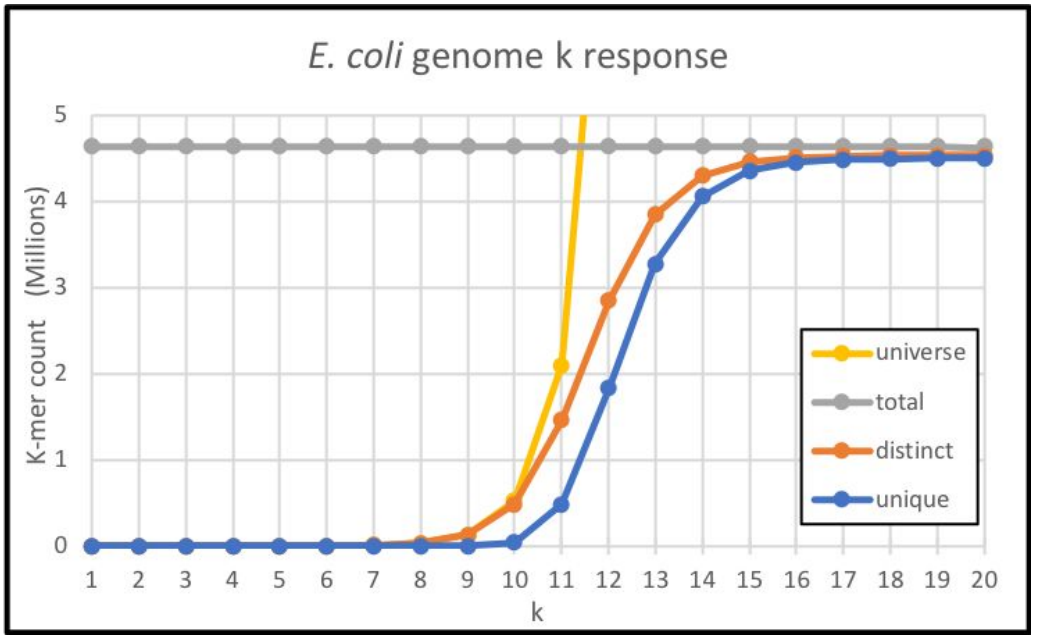
An odd-length k-mer can never be its own reverse complement:

5'- **G A T T C** -3'
3'- **C T A A G** -5'

Choosing a k-mer length

The optimal k-mer length will be determined by the goal of your analysis.

1-mers: GC-content analysis
3-mers: Codon usage
4 or 5-mers: Metagenomic binning
7 to 35-mers: Taxonomic Classification
21 to 127-mers: Genome assembly



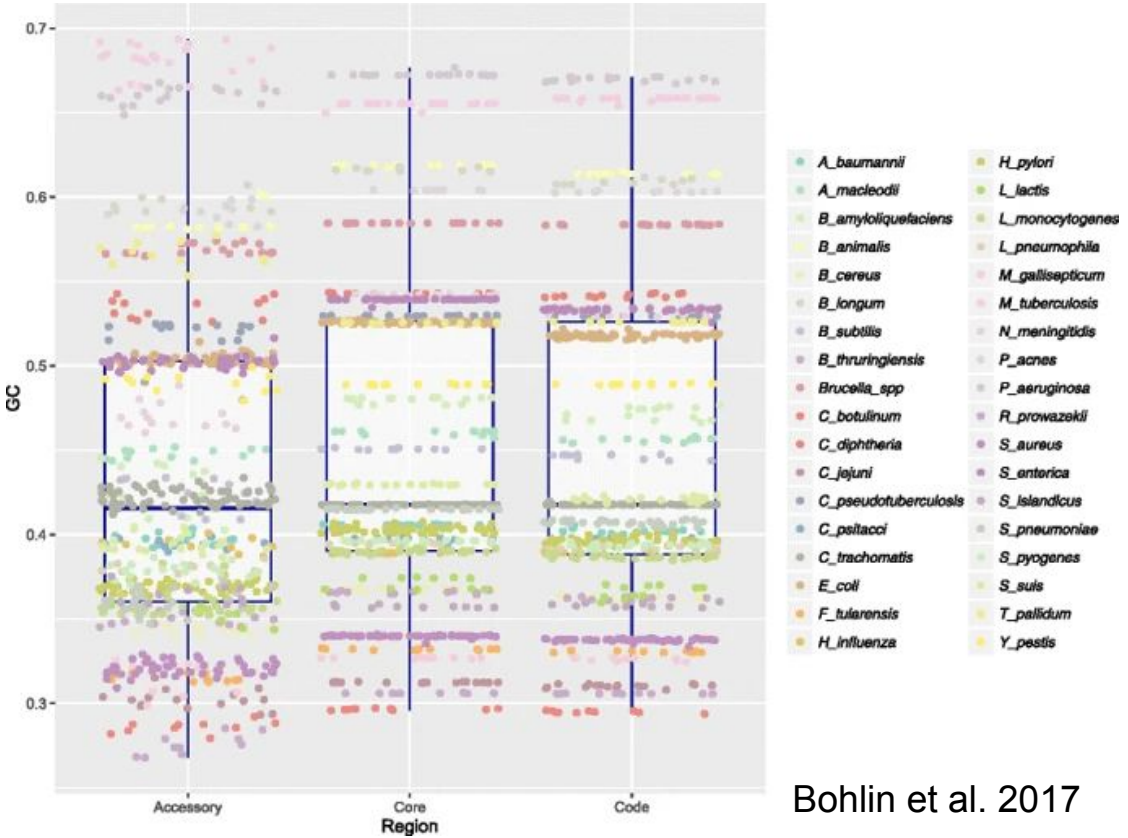
Bernardo J. Clavijo

G+C content

The distribution of k-mers in genomes is not uniform.

Even at the level of 1-mers (e.g. G+C, A+T) bacterial genomes have different, but highly conserved proportions.

This figure shows how GC content is distributed in the core, accessory and whole genome regions of some bacteria. The standard deviations are generally a couple percent.

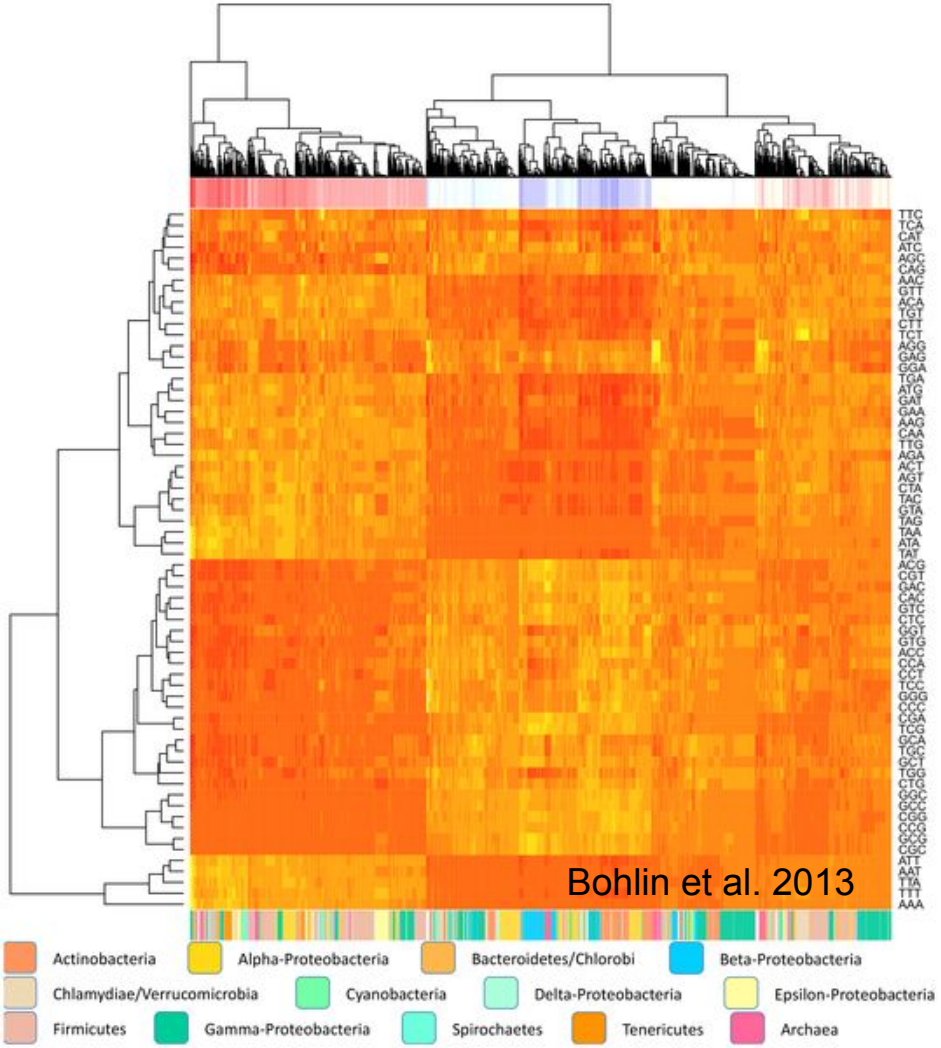


Bohlin et al. 2017

Codon usage

At the level of 3-mers, bacterial genomes have unique, but highly conserved usage of codons. This extends also to untranslated regions and may be primarily related to GC content.

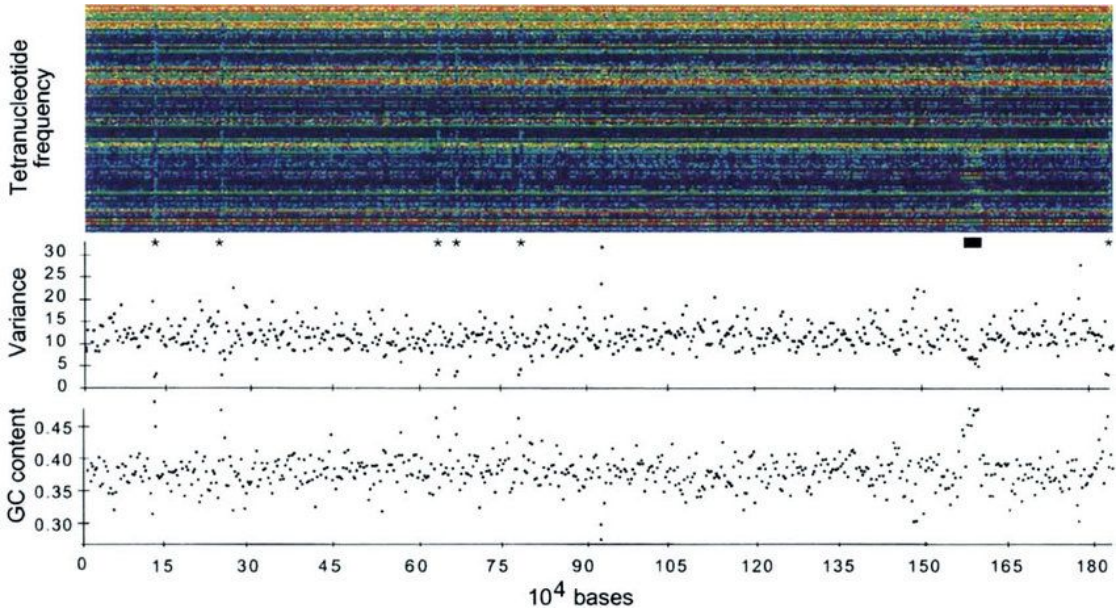
This heatmap shows more (yellow) and less (red) frequently used codons (rows) in some bacterial genomes (columns). AT-rich (red) and GC-rich (blue) genomes are shown in a row at the top.



Tetranucleotide Frequency

At the level of 4-mers (aka “tetranucleotide frequency”), bacterial genomes have conserved “fingerprints” which are used in metagenomic binning tools and to detect horizontal gene transfer events.

This figure shows an *H. influenzae* genome, where each column is the 4-mer fingerprint of a 3000 bp region and each row is a 4-mer colored from low abundance (purple) to high (red).



Starred (*) sections contain ribosomal RNA. The black bar identifies a bacteriophage.

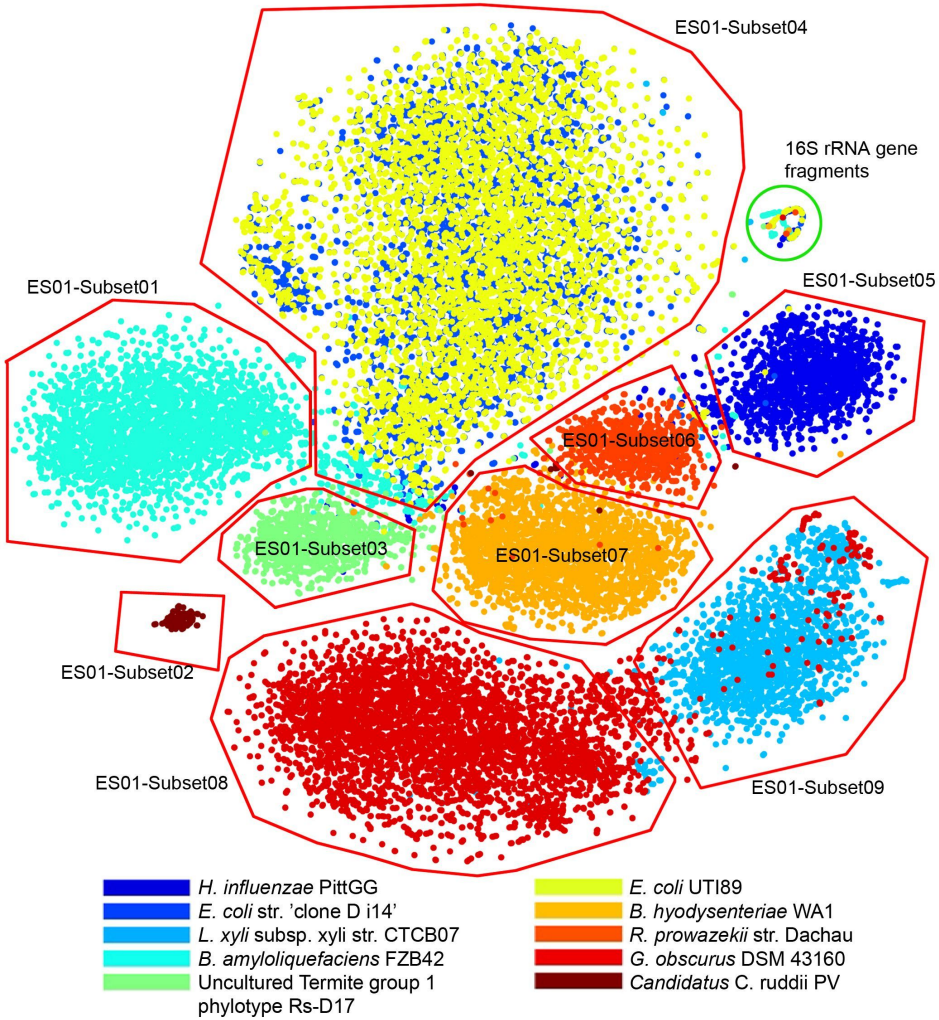
Noble et al. (1998).

Dimensional reduction

As the number of dimensions increases rapidly with the length of the k-mer, dimensional reduction algorithms like PCA, t-SNE, DBSCAN, and UMAP can be used to project (or “embed”) them back into 2D for visualization.

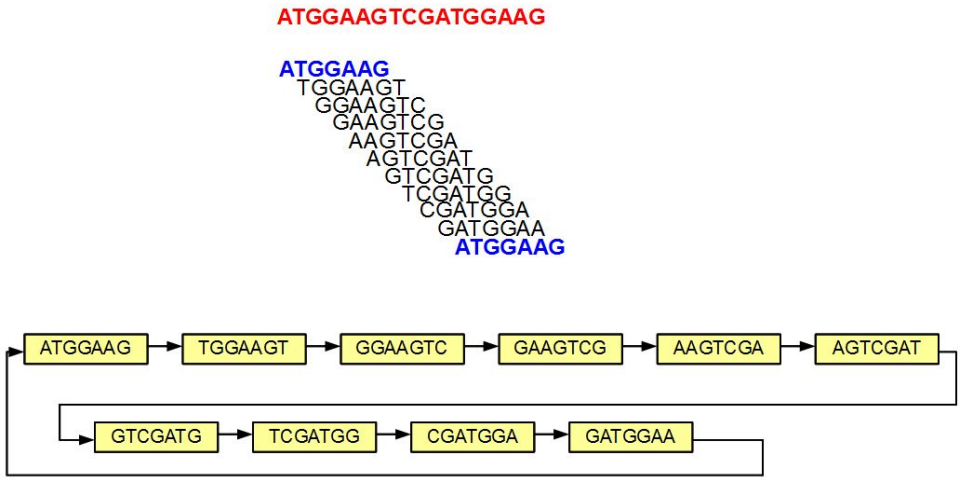
This figure shows a 4-mer t-SNE-based visualization and human-augmented binning of an evenly distributed simulated microbial community.

Laczny et al. (2014)



Genome Assembly

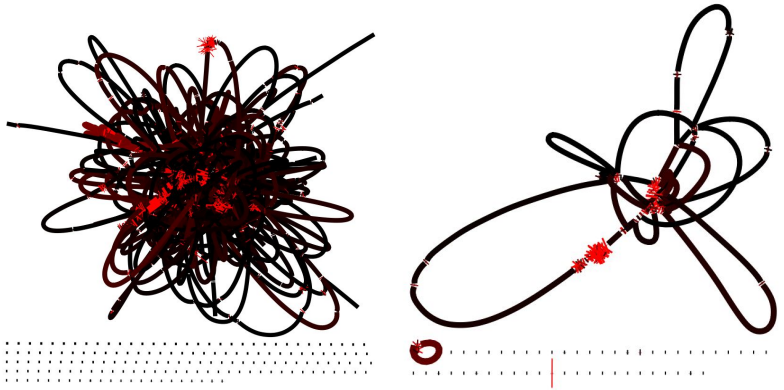
Many genome assemblers use de Bruijn graphs based on k-mer overlap.



https://homolog.us/Tutorials/book4/p2.1.html

If the k-mer length is too short, many unrelated regions are assembled together.

But k-mers cannot be longer than sequence lengths, thus the benefit of long-read technologies.



k = 21
3190 contigs

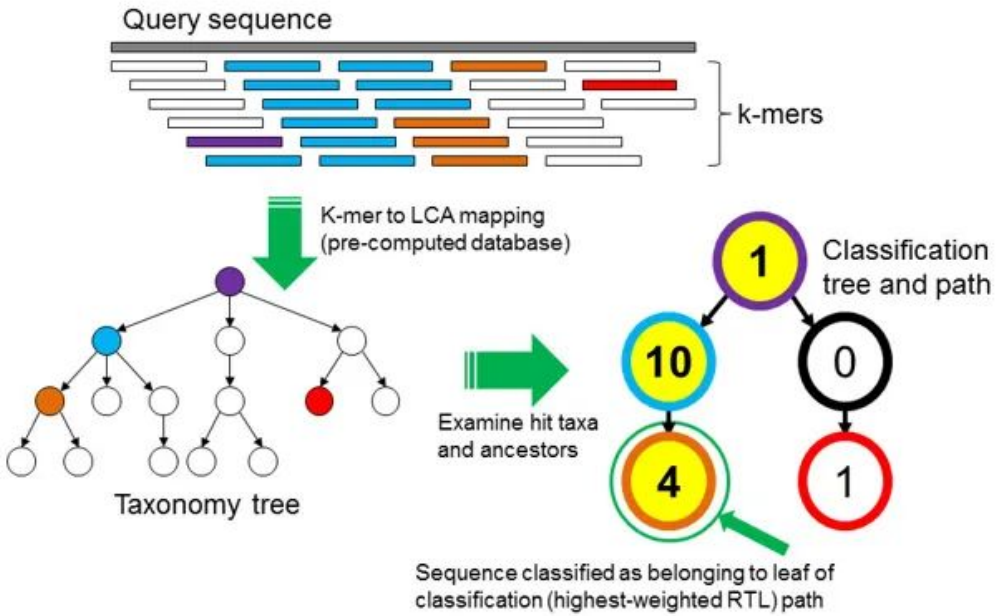
k = 127
288 contigs

Taxonomic Classification

As k-mer length increases, the probability of a random match between two distantly related taxa decreases exponentially.

This can be used to classify DNA sequences based on their k-mer profiles by comparing the k-mers to a pre-computed database of taxonomically-known k-mers.

This technique is also used for identifying adapters and other contaminants in raw reads.



Wood & Salzberg (2014)

Choosing a Metagenomic Taxonomic Classifier

Sequence format

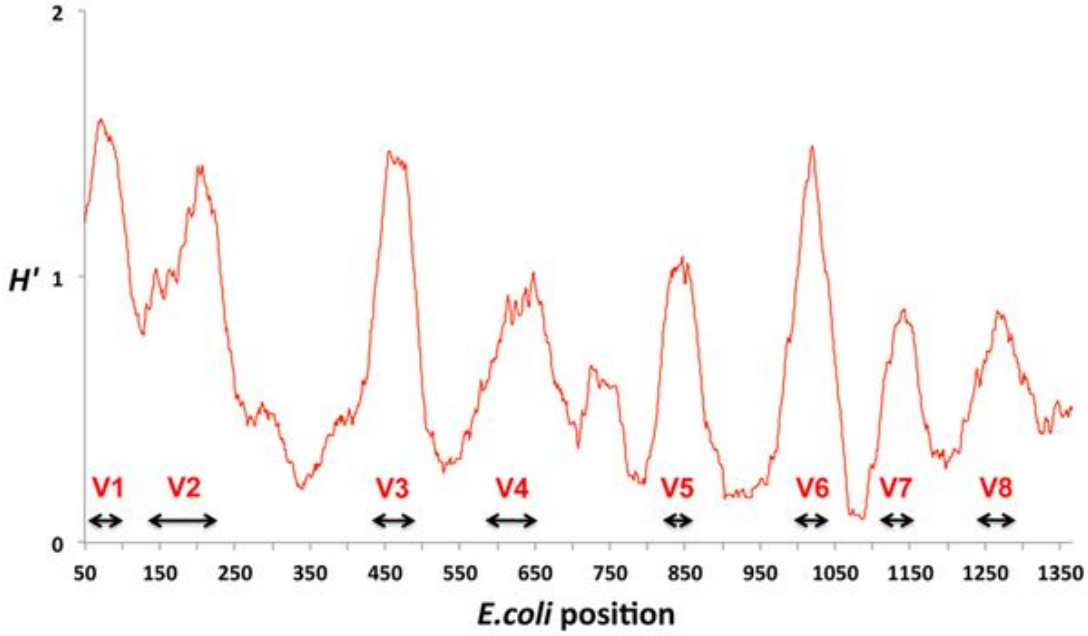
- Nucleotide k-mers grow at 4^x
- Amino acid k-mers grow at 20^x
- So a nucleotide 31-mer = Amino Acid 14-mer

Database

- Whole genomes
- Marker genes
- Ribosomal RNA genes

Search algorithm

- k-mer indexing
- MinHash (Lesson 3)
- BWT (Lesson 4)
- BLAST-like (#FunctionalAnnotation)



Variance in nucleotide composition in 16S ribosomal RNA genes. Vasileiadis et al. (2012)

Choosing a Metagenomic Taxonomic Classifier

k-mers + genomes

- Kraken (don't use anymore!)
- Kraken2 (nuc or aa)
- KrakenUniq (nuc)
- CLARK (nuc)
- MOTHUR classify.seqs (16S rRNA)

BWT + genomes

- Centrifuge (nuc)
- Kaiju (aa)

BWT + marker genes

- MetaPhlAn2
- mOTU
- AMPHORAnet

A review of methods and databases for metagenomic classification and assembly

Florian P. Breitwieser, Jennifer Lu and Steven L. Salzberg

Table 3. Metagenomic classifiers, aligners and profilers

Tool	Synopsis	Reference	Web site
Kraken	Fast taxonomic classifier using in-memory k-mer search of metagenomics reads against a database built from multiple genomes	[64]	https://ccb.jhu.edu/software/kraken/
Kraken-HLL	Extension of Kraken counting unique k-mers for taxa and allowing multiple databases		https://github.com/fbreitwieser/kraken-hll
CLARK(-S)	Fast taxonomic classifier using in-memory k-mer search of metagenomics reads against a database built from completed genomes. S extension uses spaced k-mer seeds for better classification	[65, 66]	http://clark.cs.ucr.edu
Kallisto	Taxonomic profiler using pseudo-alignment with k-mers using techniques based on transcript (RNA-seq) quantification	[67]	https://github.com/pachterlab/kallisto
k-SLAM	Taxonomic classifier using database of nonoverlapping k-mers in genomes. Reads are split into k-mers, and overlaps found by lexicographical ordering are pseudo-assembled	[68]	https://github.com/aindj/k-SLAM
Kaiju	Fast taxonomic classifier against protein sequences using FM-index with reduced amino acid alphabet	[69]	https://github.com/bioinformatics-centre/kaiju
DIAMOND	Protein homology search using spaced seeds with a reduced amino acid alphabet, 2000–20 000 times faster than BLASTX	[70]	https://github.com/bbuchfink/diamond
BLAST+	Highly sensitive nucleotide and translated-nucleotide protein alignment	[61, 71]	https://blast.ncbi.nlm.nih.gov
MEGAN6/CE	Desktop and Web metagenomics analysis suite. Uses BLAST or diamond to match sequences and assigns LCA of matches	[72, 73]	http://ab.inf.uni-tuebingen.de/software/megan6/
DUDes	Top-down assignment of metagenomics reads	[74]	https://sourceforge.net/projects/dudes/
Taxonomer	Web-based metagenomics classifier including binning and visualization	[75]	http://taxonomer.iobio.io/
GOTTCHA	Taxonomic profiler that maps reads against short unique subsequences ('signature') at multiple taxonomic ranks	[76]	http://lanl-bioinformatics.github.io/GOTTCHA/
LMAT(-ML)	K-mer-based taxonomic read classifier using extensive database including draft genomes and eukaryotes. ML (Marker Library) extension reduces RAM requirements by stringent pruning of non-informative and overlapping k-mers	[77, 78]	https://sourceforge.net/projects/lmat/
taxator-tk	Uses BLAST or LAST output for binning and taxonomic assignment via overlapping regions and pairwise distance measures	[79]	https://github.com/fungs/taxator-tk
Centrifuge	Fast taxonomic classifier using database compressed with FM-index, database and output format similar to Kraken	[80]	http://ccb.jhu.edu/software/centrifuge/
MetaPhlAn 2	Marker gene-based taxonomic profiler	[81]	https://bitbucket.org/biobakery/metaphlan2
mOTU	Taxonomic profiler based on a set of 40 prokaryotic marker genes	[82]	http://www.bork.embl.de/software/mOTU/
Mash	MinHash-based taxonomic profiler enabling superfast overlap estimations	[83]	http://mash.readthedocs.io
sourmash	Alternative implementation of MinHash algorithm using fast searches with sequence bloom trees for taxonomic profiling	[84]	https://github.com/dib-lab/sourmash
PanPhlAn	Pan-genome-based phylogenomic analysis	[2]	http://segatalab.cibio.unitn.it/tools/panphlan/

Demo #2 by Eric Collins on k-mer-based trimming (bbtools) and taxonomic classification (kraken2)

Kraken Version	Database	Build time*	Classification Time**	Database Size	MiSeq Genus			HiSeq Genus		
		(hours)	(sec/10M reads)	(GB)	Sensitivity	Precision	F1 Score	Sensitivity	Precision	F1 Score
Kraken 1	Standard 1[A]	11.7	59.2	217.3	88.4%	98.6%	93.2%	88.3%	99.2%	93.5%
	MiniKraken_v1[A] 8Gb	2.0	57.9	8.5	80.4%	99.4%	88.9%	71.1%	99.4%	82.9%
	Standard 2[B]	15.8	63.6	240.8	88.2%	98.5%	93.1%	88.0%	99.3%	93.3%
	MiniKraken_v2[B] 8Gb	2.1	58.6	8.5	79.0%	99.4%	88.0%	68.4%	99.4%	81.1%
Kraken 2	Standard 1[A]	1.1	20.2	31.2	89.4%	97.9%	93.5%	89.6%	99.1%	94.1%
	MiniKraken2_v1[A] 8Gb	0.5	16.6	8.0	85.4%	98.6%	91.6%	84.0%	99.3%	91.0%
	Standard 2[B]	3.9	26.0	34.7	89.3%	97.8%	93.3%	89.2%	99.1%	93.9%
	MiniKraken2_v2[B] 8Gb	0.7	24.3	8.0	84.8%	98.4%	91.1%	82.6%	99.3%	90.2%

*Build time calculated with 32 threads

**Classification time calculated with 16 threads

[A]Standard 1 = refseq archaea, bacteria, viral

[B]Standard 2 = refseq archaea, bacteria, viral, and human

PRE-BUILT KRAKEN 2 DATABASES

The following databases can be found at ftp://ftp.ccb.jhu.edu/pub/data/kraken2_dbs/. Information about how to download/install/use databases is provided: README.md.

Databases are pre-built, including the required hash.k2d, opts.k2d, and taxo.k2d files. Each database also includes 100mer, 150mer, and 200mer Bracken files.

- Kraken 2 16S Greengenes 13_8 DB (69.4 MB)
- Kraken 2 16S RDP 11.5 DB (164 MB)
- Kraken 2 16S Silva 132 DB (115 MB)

These additional databases have been provided by non-CCB labs:

- GTDB_r89_54k linkA collection of database files for use with Centrifuge, Kraken 1, or Kraken 2 that can be used to classify metagenomes using the GTDB_389_54k index. More information and details at: https://github.com/rrwick/Metagenomics-Index-Correction
- Maxikraken2 and Kraken2-microbial databases. These databases are maintained by LomanLab. More information at the link provided.

Additional Resources

- [C. Titus Brown] A post about k-mers - this time for taxonomy!
- [Jennifer Lu et al.] How to Choose Your Metagenomics Classification Tool
- [Bernardo J. Clavijo] k-mer counting, part I: Introduction
- De Bruijn graph of a genome
- Comparing datasets using sourmash