



Genomic DNA k-mers: basic notions and applications

Maria Nuc, Paweł Krajewski

IGR PAN

k-mer – what is it?

k-mer: a substring of a length k contained in a given string.

In biology: having a sequence (DNA, RNA, protein), k-mers are all the subsequences of a length k contained in this sequence.

Example:

AAAGAGTTGTGGTAA

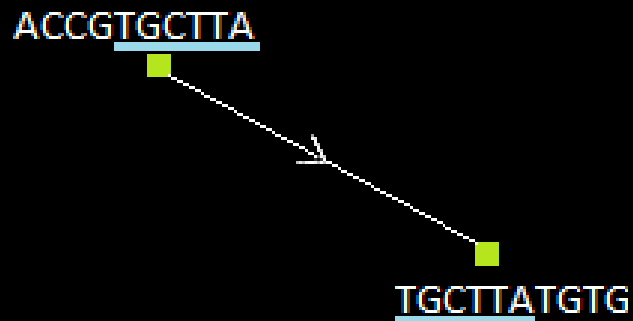
10-mers: AAAGAGTTGT, AAGAGTTGTG,
AGAGTTGTGG, GAGTTGTGGT,
AGTTGTGGTA and GTTGTGGTAA

sequence assembly

Sequence assembly – the process of aligning and merging obtained fragments of a DNA sequence in order to reconstruct the original sequence.

The idea of using k-mers in a sequence assembly is to build a special directed graph representing fragments of a DNA sequence and all their overlaps (regions of the same sequence).

The de Bruijn graph – a directed graph whose set of vertices represents the set of fixed-length strings and the edges represent suffix-to-prefix overlaps.



The k -mer graph – a form of de Bruijn graph, where vertices represent k -mers of a given sequence and edges represent overlaps of k -mers of length $k-1$.

example

$k=5$, length of overlaps = 4

CCGGTGCC CCGGT, CGGTG, GGTGC, GTGCC

CGAAGCGA CGAAG, GAAGC, AAGCG, AGCGA

GTGCCCGAAG GTGCC, TGCCC, GCCCG, CCCGA,
CCGAA, CGAAG

AACACCGGTG AACAC, ACACC, CACCG, ACCGG,
CCGGT, CGGTG

example

$k=5$, length of overlaps = 4

CCGGTGCC

CCGGT, CGGTG, GGTGC, GTGCC

CGAAGCGA

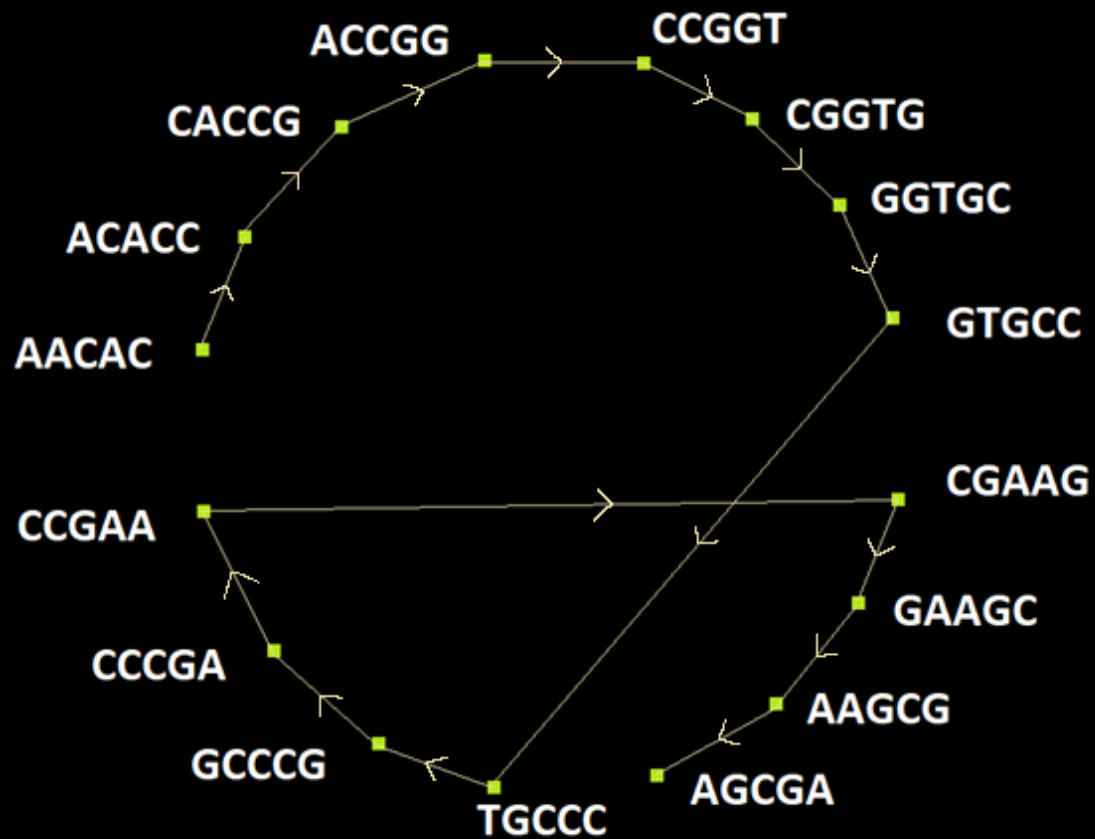
CGAAG, GAAGC, AAGCG, AGCGA

GTGCCCGAAG

~~GTGCC~~, TGCCC, GCCCG, CCCGA,
CCGAA, ~~CGAAG~~

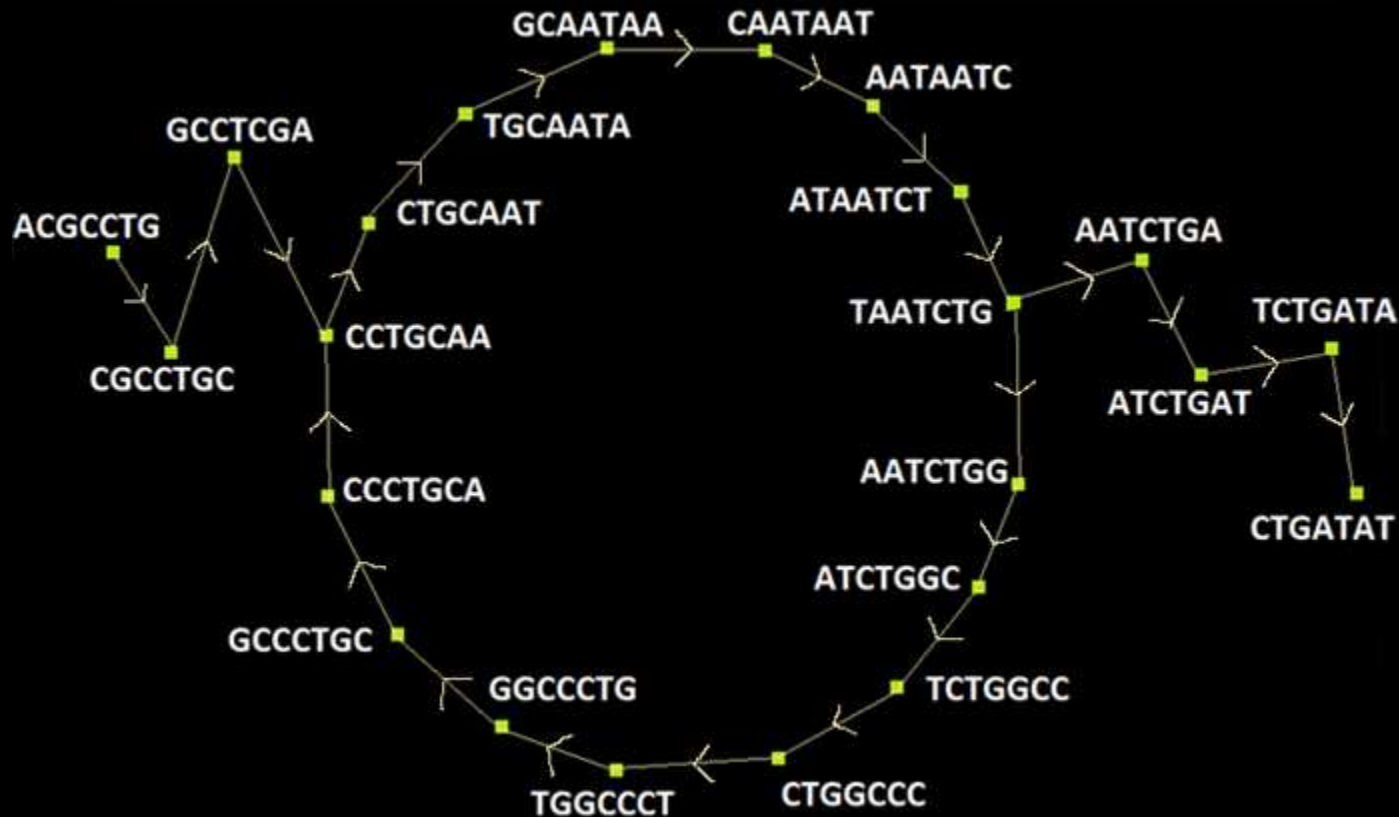
AACACCGGTG

AACAC, ACACC, CACCG, ACCGG,
~~CCGGT~~, ~~CGGTG~~



AACACCGGTGCCCGAAGCGA

some problems: e.g. repeats



Is it **ACGCCTGCAATAATCTGATAT**?

Or **ACGCCTGCAATAATCTGGCCCTGCAATAATCTGATAT**?

Or **ACGCCTGCAATAATCTGGCCCTGCAATAATCTGGCCCTGCAATAATCTGATAT...**?

...



ABYSS



SOAPdenovo

AllPaths

Velvet assembler



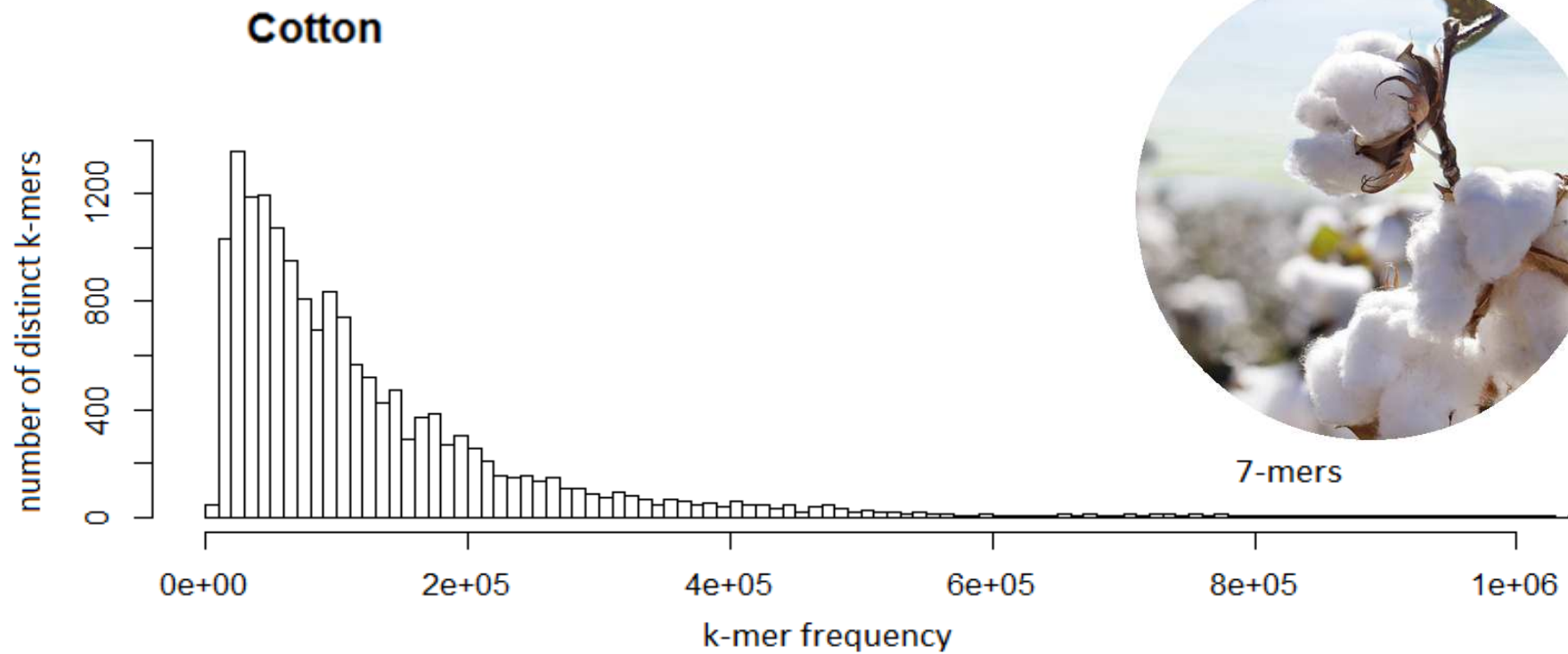
Euler assembler

k-mer spectra & genomic barcodes

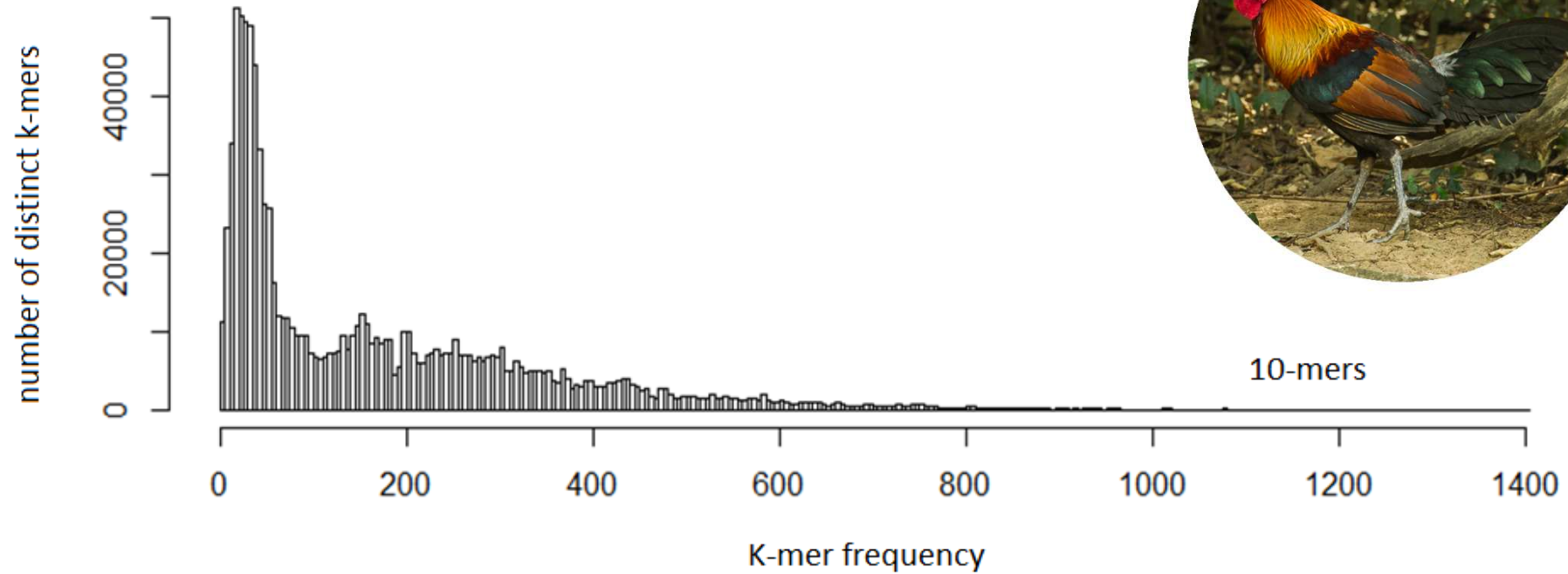
k-mer spectrum - a graphical representation of a dataset showing how many k-mers are repeated in data exactly p times, for $p=0,1,2,\dots$ (p is less or equal $L-p+1$, where L is the length of a given sequence).

Genomic barcodes – k-mer spectra of regions of a genome with a localization.

examples



Red junglefowl (*Gallus gallus*)



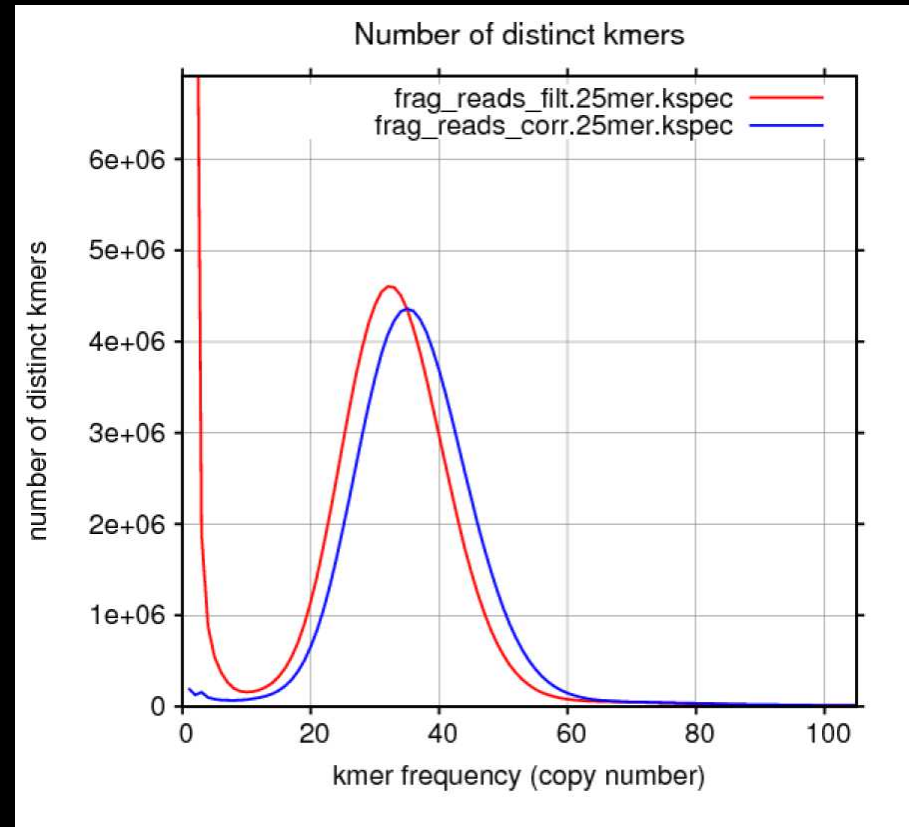
multimodality of k-mer spectra

One of the things which can be observed in k-mer spectra is its modality. Most species have unimodal spectra but all tetrapods have multimodal spectra.

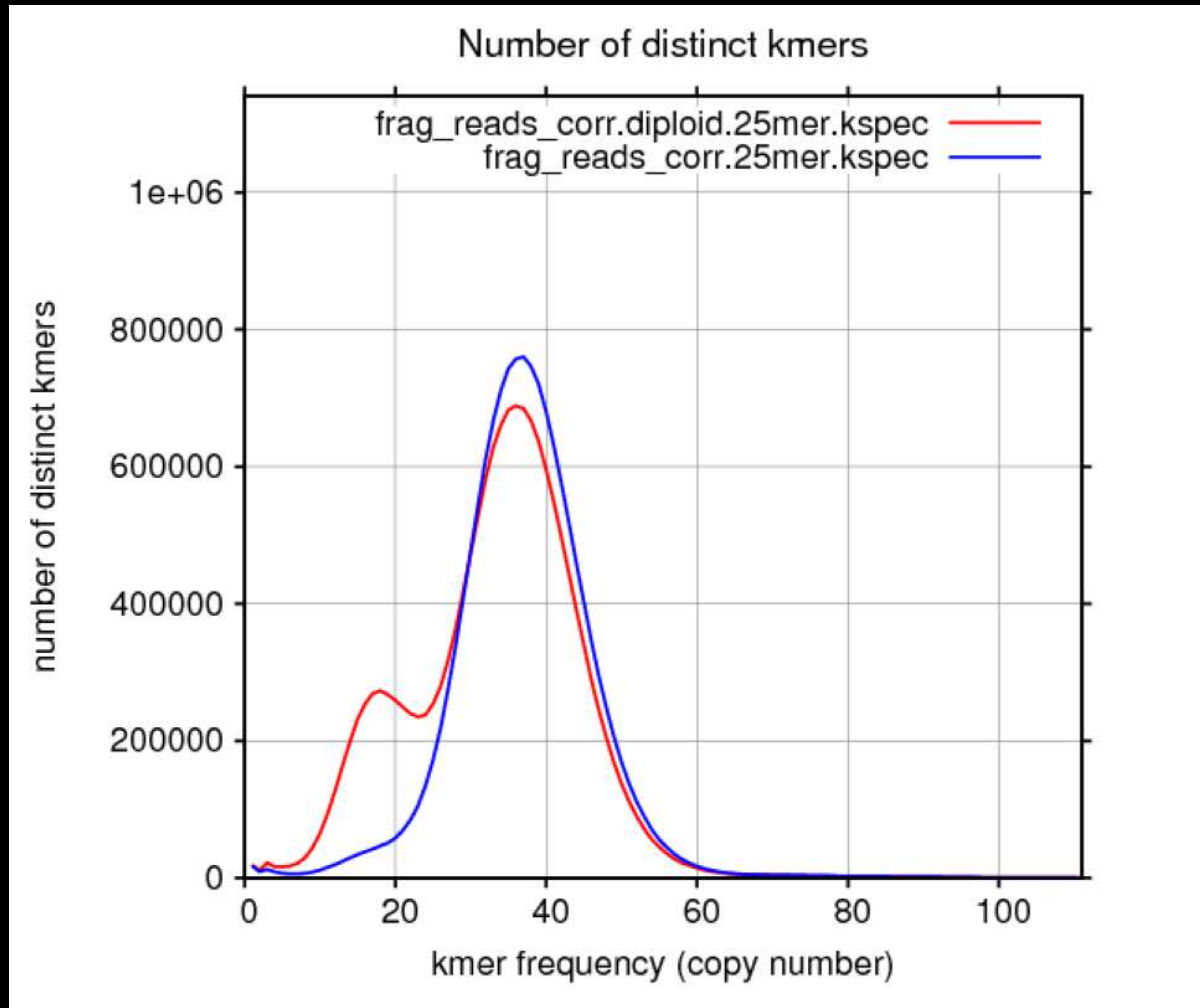
Researchers showed [1], that the multimodality of k-mer spectra depends on values of G+C content and CpG dinucleotide suppression.

quality of data

Another application of k-mer spectra is checking the quality of data.



identification of polimorphism or mutation



horizontal transfer detection

Horizontal gene transfer (HGT) – the movement of genetic material between two organisms which are not in parent-offspring relation.

Scientists observed that different microbial species have their own genomic word pattern signatures. Therefore if they are looking for fragments of genome of a given organism that appeared there via HGT, they are looking for an atypical word pattern composition in this genome. To search for these patterns they use genomic barcodes and calculating the distance between them.

genome size estimate

We can estimate the genome size using k-mer spectrum and some more informations.

$$G = T / N$$

Where:

G – genome size estimate

T – total number of bases

N – depth of read coverage

Depth of read coverage -
the number of bases of all
short reads that match a
genome divided by the
length of this genome.



$$G = T / N$$
$$N = M * L / (L - k + 1)$$

Where:

G – genome size estimate

T – total number of bases

N – depth of read coverage

M – mean k-mer coverage

k – k-mer size

L – read length

Mean k-mer coverage – we read it from k-mer spectrum plot: the peak with the largest k-mer multiplicity is the value we're looking for.

k-mers – applications

- Sequence assembly
- Checking quality of data
- Identification of polymorphism or mutation
- Genome size estimation
- Barcoding
- Horizontal transfer detection
- Analysis of k-mer spectrum
- And more...

choosing k

Choosing a proper k is crucial in usage of k -mers. Too small k can be not sufficient to get the clear information about structures we want to know, while too big k extends the calculation time and can be too „severe“ for our data.

lower k-mer sizes

- ❖ Fewer edges in the graph → less space we need to store the data.
- ❖ Bigger chance for all the k-mers to overlap during construction of the de Bruijn graph.
- ❖ Risk: many vertices in the graph can lead into a single k-mer and this will make the reconstruction of the genome more difficult.
- ❖ Information loss: e.g.

AGTCGTAGATGCTG vs. ACGT

- ❖ Problem: repetitions: e.g. **ATGTGTGTGTGTGTACG**

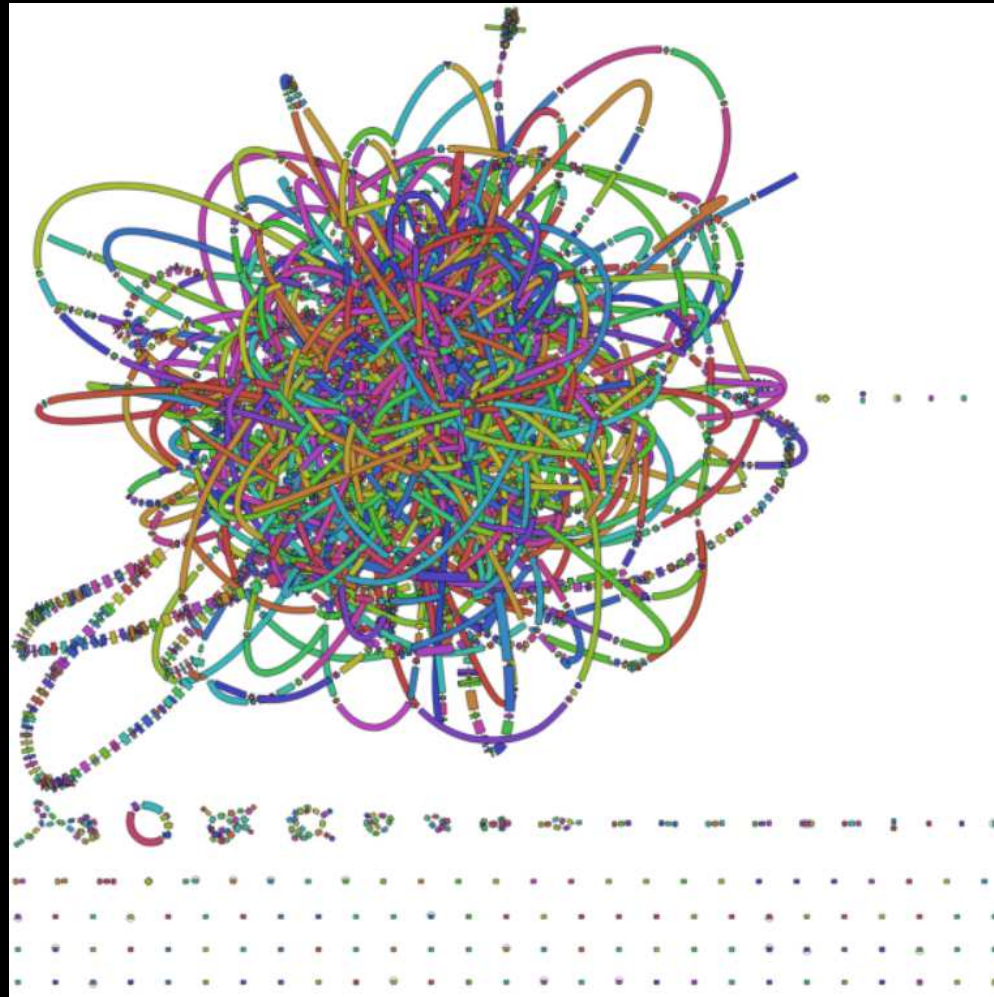
higher k-mer sizes

- ❖ More edges in the graph → More memory we need to store the data
- ❖ The number of vertices in de Bruijn graph will also decrease - there will be fewer paths to traverse in the graph.
- ❖ Risk: due to larger k-mers increasing the chance that it will not overlap with another k-mer by $k - 1$. It can lead to disjoints in the reads, and a higher amount of smaller contigs.
- ❖ It alleviates the problem of small repeat regions.

51-mer assembly

(*Salmonella* genome from 100bp Illumina reads, Velvet)

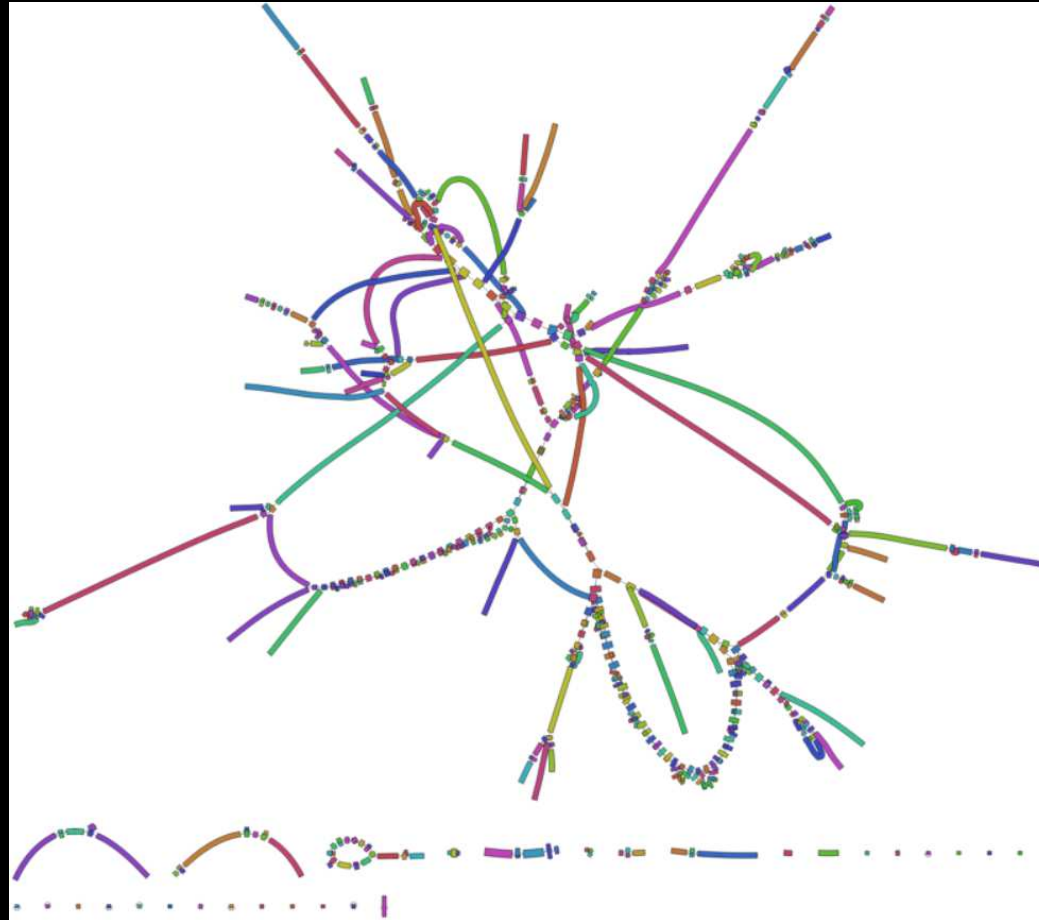
(!) Very complex, tangled graph with too big amount of edges and nodes



Source: <https://github.com/rrwick/Bandage/wiki/Effect-of-kmer-size> (Ryan Wick)

71-mer assembly

Fewer vertices and edges, more 'dead ends'.

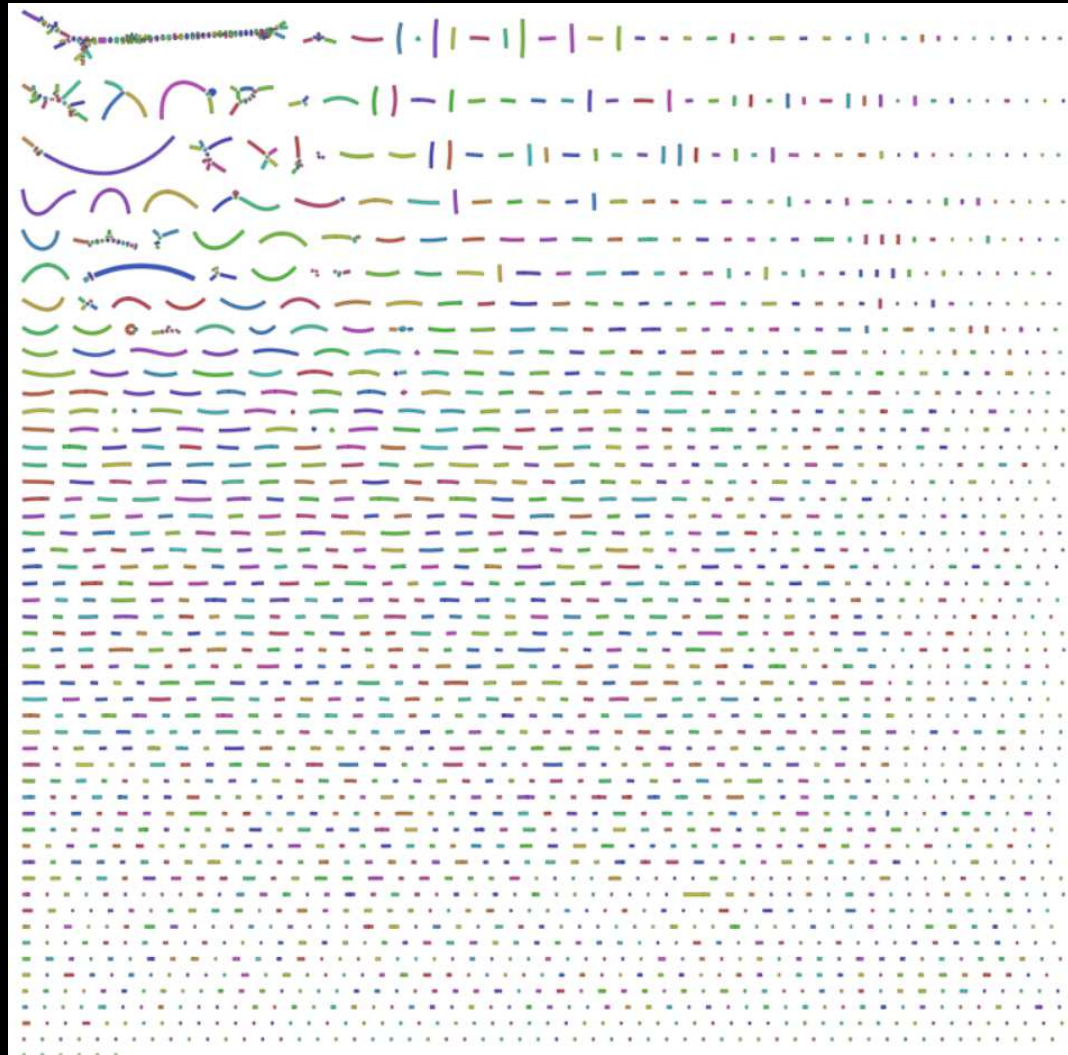


The best of these three examples, because...

Source: <https://github.com/rrwick/Bandage/wiki/Effect-of-kmer-size> (Ryan Wick)

91-mer assembly

A lot of disconnected nodes.



Source: <https://github.com/rrwick/Bandage/wiki/Effect-of-kmer-size> (Ryan Wick)

references

- Chor B., Horn D., Goldman N., et al. (2009). Genomic DNA k-mer spectra: models and modalities. *Genome Biology* 10: R108.
- Sievers A., Bosiek K., Bisch M., et al. (2017). K-mer content, correlation, and position analysis of genome DNA sequences for the identification of function and evolutionary features. *Genes* 8 (4): E122.
- Tang K., Lu Y.Y., Sun F., (2018). Background Adjusted Alignment-Free Dissimilarity Measures Improve the Detection of Horizontal Gene Transfer. *Front. Microbiol* 9:711.
- <http://software.broadinstitute.org/allpaths-lg/blog/wp-content/uploads/2014/05/KmerSpectrumPrimer.pdf>
- <https://github.com/rrwick/Bandage/wiki/Effect-of-kmer-size> (Ryan Wick)

Thank You for your attention

