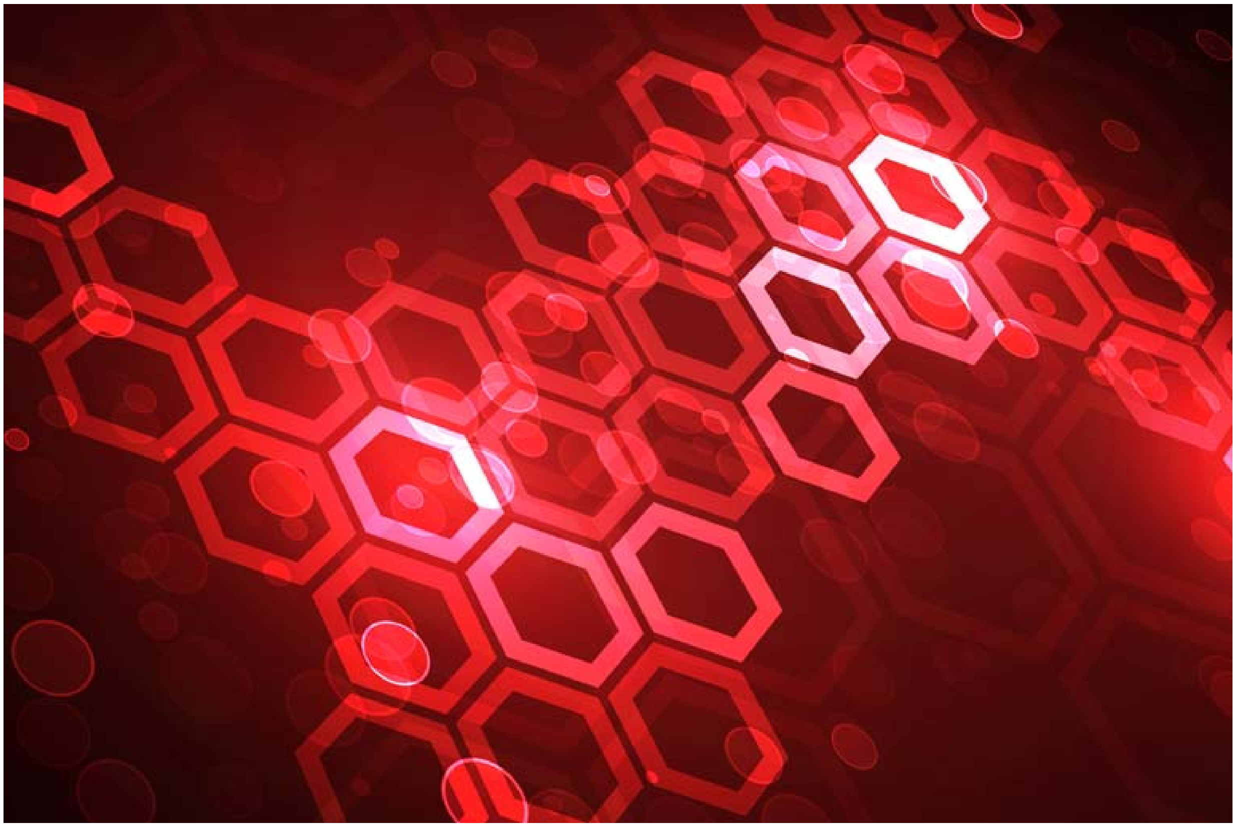
PHYS.ORG

Rapid validation for genome assemblies? Introducing KAT: K-mer Analysis Toolkit

December 5 2016



Credit: Earlham Institute

Genome assembly projects are costly in both time and money; where identifying problems with your data post-assembly can be a real setback. With the K-mer Analysis Toolkit (KAT), researchers can access and confirm their results at every stage.

Genome assembly with NGS technologies is like trying to do the hardest jigsaw puzzle you can imagine. The final jigsaw represents the full genome, and the individual pieces represent small fragments of the genome read out by the sequencer. Counterintuitively, to make the data more manageable, it is actually easier to first break these pieces into even smaller pieces called K-mers.

K-mers represent small fragments of the original genome with a fixed number (K) of DNA base pairs. A computer can efficiently work with large quantities of K-mers, then identify connections between these fragments to build-up a representation of the original genome.

K-mer-based techniques are commonly used to efficiently generate genome assemblies, KAT, however, is built to examine and compare K-mer datasets, using each distinct K-mer's underlying properties, such as frequency and nucleotide composition.

Initially, KAT can analyse sequencing data to identify error levels, biases and contamination. Information from this analysis can help researchers decide whether to proceed with downstream tasks such as genome assembly. KAT can then internally back-check your assembly to determine completeness and accuracy without any external reference data - a really useful feature when studying new organisms.

Lead Software Developer, Daniel Mapleson, said on the new tool: "Imagine genome assembly like lego. Instead of trying to piece together long, 8x2-stud pieces with 6x2-stud pieces and 5x2-stud pieces, it's more like making a staircase pattern out of the smaller 2x2-bit pieces, overlapping one stud at a time.

"However, K-mers are not only useful for assembling a genome, by counting the number of K-mers in a sequencing dataset you can learn a lot about it. By looking at the K-mer frequency profiles (K-mer spectra) we can assess the quality of the sequencing data in the first instance, such as working out if the dataset is clean, contains contaminants or is biased in some way. KAT can give answers to these questions quickly, even for non-model organisms where a reference is not available."

Project Leader and corresponding author Bernardo Clavijo commented: "The first thing many researchers do after sequencing a genome is to use-check the K-mer spectra of their data. This tells you if the information you will need to assemble the genome is there before you spend a lot of time, effort and money on doing the rest of the analysis. Now with KAT, researchers can do all kinds of validation and information comparison at this initial stage; but to also carry this forward to validation, we have included the relevant information at the end of the assembly.

"In terms of assembly validation, the tool is particularly useful with diploid genomes that can carry more than one copy of a gene, certain regions can be falsely duplicated or deleted during assembly, leading the researcher to believe there's more or less copies of a gene than there really are. KAT can help to detect these artefacts by tracking both the data generated from the sequencer and data from the assembler, ultimately leading to faster, more accurate conclusions."

The paper titled: KAT: A K-mer Analysis Toolkit to quality control NGS datasets and genome assemblies is published in *Bioinformatics*.

More information: Daniel Mapleson et al, KAT: A K-mer Analysis Toolkit to quality control NGS datasets and genome assemblies, *Bioinformatics* (2016). DOI: 10.1093/bioinformatics/btw663

Provided by Earlham Institute

Citation: Rapid validation for genome assemblies? Introducing KAT: K-mer Analysis Toolkit (2016, December 5) retrieved 27 April 2024 from https://phys.org/news/2016-12-rapid-validation-genome-kat-k-mer.html