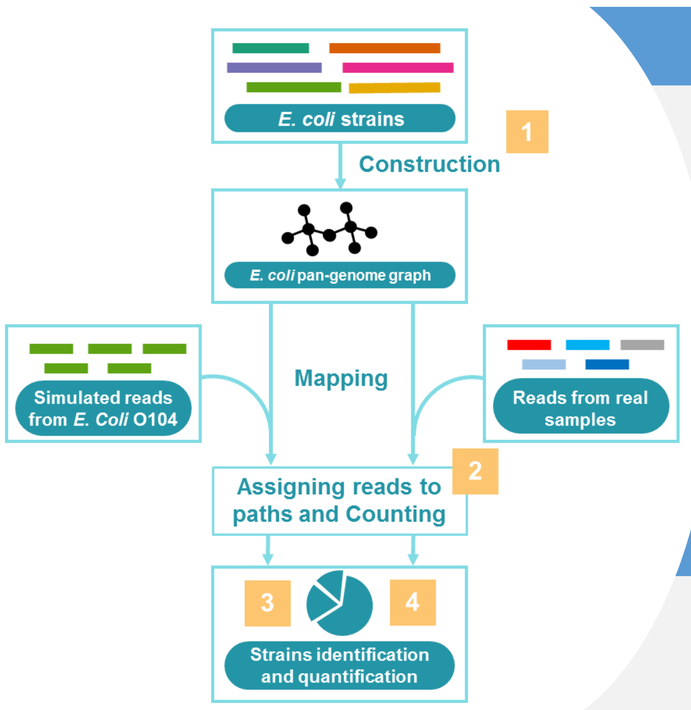
From genomics to metagenomics: benchmark of variation graphs

Kévin da Silva, Nicolas Pons, Magali Berland, Florian Plaza Oñate, Mathieu Almeida, Pierre Peterlongo

From genomics to metagenomics: benchmark of variation graphs

Kévin Da Silva[1,2], Nicolas Pons[2], Magali Berland[2], Florian Plaza Oñate[2], Mathieu Almeida[2], Pierre Peterlongo[1]



Background

In the metagenomics field, the classical approach for quantitative analysis of sequencing data consists into aligning sequence reads to a non redundant reference gene catalogue that represents a specific ecosystem [1]. However this approach lacks flexibility and exhaustiveness. To overcome those biases, the reads should be aligned to a more informative reference structure covering the variants encountered in the population and also more complete with a full genome catalogue. Recently, the pangenome concept has been increasingly used as it opens new ways to investigate multiple genomes of close individuals, as for characterizing the different strains of a species. Erik Garrison et al. have developed “vg”, a toolkit for creating variation graphs, bidirected DNA sequence graphs that represents multiple genomes, including their genetic variation [2]. With a perspective towards metagenomics, we foresee vg as a tool enabling to build a catalogue of pangenomes from metagenomic samples.

Objectives

As a proof of concept, we constructed a pangenome graph for *Escherichia coli* with three main objectives: evaluating time complexity to take it into consideration for future scaling, defining paths (successive nodes) in the graph corresponding to strains, and identifying and quantifying strains in a sample by mapping its reads on the graph.

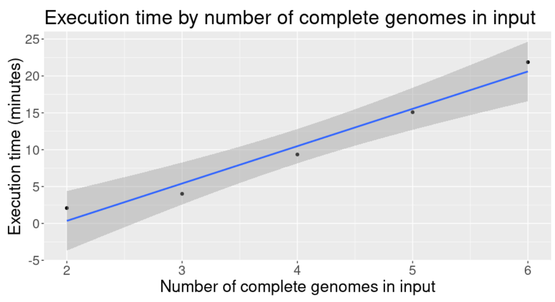
1 Graph construction and time complexity

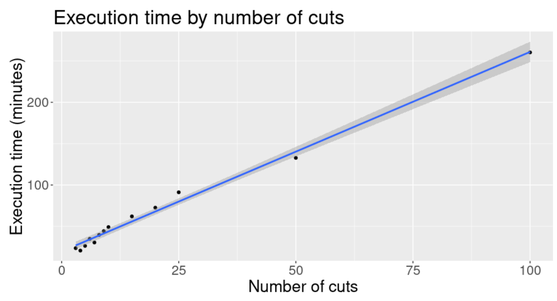
Data

- Inputs: 6 strains of *E. coli* → O157:H7, IAI39, BL21, SE15, O104:H4, K-12
- Graph construction using VG

Graph construction

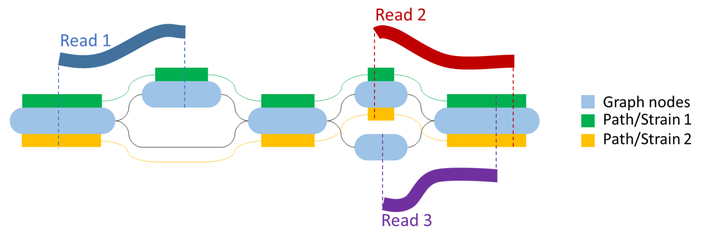
- Complete genomes as inputs:
  - Execution time with increasing number of genomes was recorded (left).
  - Time linear, 20 minutes for constructing the graph from 6 genomes.
- Contigs as inputs:
  - Complete genomes of two of the strains were cut into parts (“contigs”). Parts were overlapping on 20 kbp to ensure there was no ambiguity on the order of the parts.
  - Execution time with increasing number of cuts was recorded (right). Time linear, 4 hours using both genomes cut in 100 parts each.





2 Assigning reads to paths and counting

- A read is assigned to a path if the whole read matches a set of successive nodes belonging to the same path.
- Three different ways of counting the number of reads assigned to a path:
  - Total Count (read is counted for each path it is assigned to)
  - Unique Count (read is counted if it is assigned to a unique path)
  - Adjusted Count (Unique Count is used to define the relative abundance of the strains in the sample, this ratio is then applied to Total Count)



Total Count	Path/Strain 1	Path/Strain 2
Read 1	1	0
Read 2	1	1
Read 3	0	0

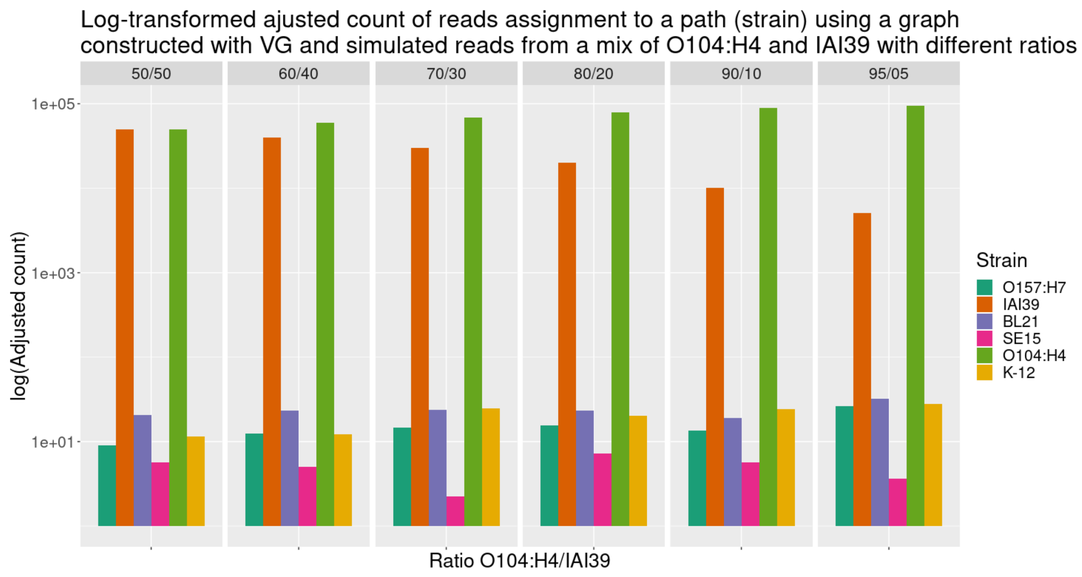
Unique Count	Path/Strain 1	Path/Strain 2
Read 1	1	0
Read 2	0	0
Read 3	0	0

Adjusted Count	Path/Strain 1	Path/Strain 2
Read 1	$1*r_1$	$0*r_2$
Read 2	$1*r_1$	$1*r_2$
Read 3	$0*r_1$	$0*r_2$

Compute relative abundance r_i for each *i* strain and apply it to the Total Count

3 Mapping and counting using simulated data

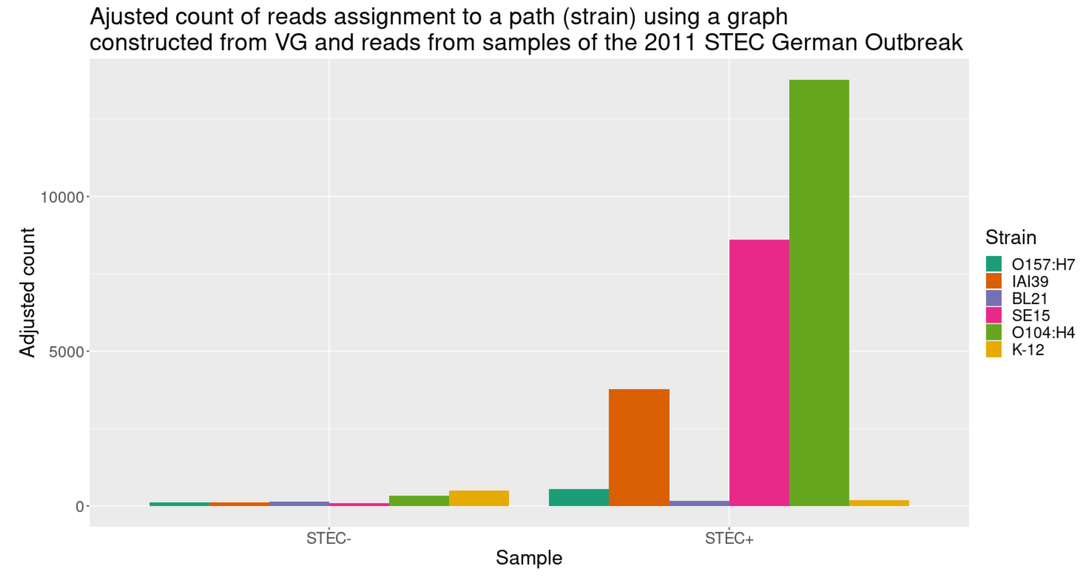
Results on simulated data: 100 000 reads of length 200 bp were generated without errors using a mix of O104:H4 and IAI39 *E. coli* strains at different ratios and then mapped on the graph.



The strains and their original abundance are correctly identified.

4 Mapping and counting using real data

Results on real data: 2 samples randomly selected from a collection after the German outbreak of Shiga-toxigenic *E. coli* (STEC) O104:H4 in 2011 mapped on the graph → 1 sample positive for STEC (STEC+) where O104:H4 is expected and 1 negative (STEC-) where O104:H4 should be absent.



In the STEC+ sample, as opposed to the STEC-, the O104:H4 strain has a higher signal, demonstrating that we can identify and quantify a strain among a whole metagenomic sample composed of several species.

Unité GenScale
team.inria.fr/genscale/

Unité MetaGenoPolis
www.mgps.eu
contact@mgps.eu

[1] Univ Rennes, CNRS, Inria, IRISA, - UMR 6074, F-35000 Rennes

[2] MetaGenoPolis, INRA, Université Paris-Saclay, 78350, Jouy-en- Josas, France

Contact: kevin.da-silva@inria.fr

References

[1] Li J, et al. An integrated catalog of reference genes in the human gut microbiome. Nat. Biotechnol. 32, 834–841, 2014.

[2] Garrison E, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol. 36(9):875–9, 2018.