Francesco Andreace

# Can I decide it or should I put the official one?

## Analysis of human pangenome graphs using k-mer-based applications

**Thesis submitted for the degree of Philosophiae Doctor**

Department of Computational Biology
Insitut Pasteur Paris, Universite' de Paris Cite

Edite doctoral school, Sorbonne Université

**2024**

**© Francesco Andreace, 2024**

*Series of dissertations submitted to the*
*Insitut Pasteur Paris, Universite' de Paris Cite, Sorbonne Universite'*
*No.* 1234

ISSN 1234-5678

All rights reserved. No part of this publication may be
reproduced or transmitted, in any form or by any means, without permission.

Cover: TBD - in PHDUIO.cls.
Print production: Pasteur Paris.

*To Sofia, my sweet old cat that lives so well without overthinking.*

# Preface

This thesis is submitted in partial fulfillment of the requirements for the degree of *Philosophiae Doctor* at Sorbonne Université. The research presented here was conducted at the Insitut Pasteur, under the supervision of Dr. Rayan Chikhi and Dr. Yoann Dufresne. This work was supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 956 229 (+Panagaia +Pasteur + INCEPTION).

The thesis is a collection of the different projects I worked on during my stay at Insitut Pasteur. I begin with an small foreword of the research output of my PhD, and a gentle introduction of the scientific concepts needed to understand the rest of the manuscript. In the first section I present the published paper I am first author of, together with other unpublished work I lead or independently developed. In the second section I present other results of my scientific production, with novel elaboration of the work that appeared in the other papers I am co-author and presentation of projects that hava or will be submitted to revision. The common theme is pangenomics and computational methods used to generate and use such models to infer relevant information. This essay ends with a chapter showcasing future perspectives and conclusions.

## Acknowledgements

Thanks for spending time reading this.

**Francesco Andreace**
Paris, September 2024

# Contents

# List of Figures

# List of Tables

## 0.1 Sequencing data

### 0.1.1 Next-generation and third generation sequencing

### 0.1.2 Reads

## 0.2 *k*-mers and how to store them

## 0.3 Pangenomics, pangenomes and pangenome graphs

### 0.3.1 The premisis for human pangenomics

#### A LINEAR REFERENCE FOR ALL GENOMIC ANALYSIS

Since the beginning of genomics, all analysis based on sequencing data depended upon the use of a single linear reference genome, i.e. the best assembled genome available for a species, to extract useful information from the DNA. We now know that this approach is suboptimal in a wide range of applications as a lot of genetic material of the species cannot be present in a single linear refererence: this is valid for eukariotes and even more for bacteria. **A SEQUENCING REVOLUTION**

Right now we are witnessing a real revolution in the sequencing. As the price is significantly lowering, also thanks to competition of new companies entering the market, new scientific discoveries and technological advances are leading to a remarkable increase of quality, in term of per-base error rate, and throughput. This means than right now we dispose of a rich wealth of high quality sequencing information to produce hundreds or thousands of new first grade assemblies.

**A QUALITY REVOLUTION** This limitation at the beginning was not solvable due to the scarcity of high quality assembled genomes as the technologies of sequencing and computational tools were not mature enough. For example, the Human Genome Project took 13 years to produce its result [**humangenomeproject**] and the absence of long reads with decent error rate made it impossible to automatically resolve repetitive regions like telomers and centromeres [**human-pangenomics-era**], producing a reference only 92% complete [**t2t**]. This problem was only solved in 2022 [**t2t**]. At the same time, many consortia are producing increasingly more genomes to a level comparable to the T2T consortium. For example, the HPRC, i.e. the Human Pangenome Reference Consortium released 47 new human genomes (92 haplotypes) in 2021 and has recently released other 153 genomes to a total of 400 haplotypes. The ability to produce such high quality data for human genomes is the main driver of the

Right now we are witnessing a real revolution in the sequencing. As the price is significantly lowering, also thanks to competition of new companies

entering the market, new scientific discoveries and technological advances are leading to a remarkable increase of quality, in term of per-base error rate, and throughput. This means than right now we dispose of a rich wealth of high quality sequencing information to produce hundreds or thousands of new first grade assemblies. This progress lead to a shift in paradigm with increasing effort from the scientific community to propose new methods to analyse one or multiple genomes: not anymore by comparing it against a single reference sequence but against a comprehensive representation of the species.

This novel way to overcome the limits of "linear genomic" and consider all the variation in a single species is called pangenomics.

Various efforts are being made on producing reference pangenomes of yeasts, bacterias, plants and animals, including humans. In order to do so, new tools to construct and then analyse and use such representations are being developed. It is important here to notice, as it will be stressed in the next sections and chapters, that construction is just the first step and that is very important to understand and work on which are the operations that can be

succesfully performed by these representations.

## 0.4 Graphs

### 0.4.1 De Bruijn Graphs

#### 0.4.1.1 Colored and Compacted De Bruijn Graphs

### 0.4.2 Variation Graphs

# Papers

# Appendices