Francesco Andreace

k-mers to rule them all

Analysis of human pangenome graphs and other 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

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

ISSN 1234-5678

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

To Sofia

Preface

refsection:1

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 Doc. Rayan Chikhi and Doc. 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 (+Pasteur + INCEPTION).

The thesis is a collection of the published paper I am first author of, novel elaboration of the work I did for other papers where I am not first author and presentation of work that has or will be submitted to revision and is then not yet published. The common theme to them is the development k-mer-based methods and the use of such for (meta/pan/-)genomics. The paper is preceded by an introductory chapter that provides background information and motivation for the work and is succeded by considerations and perspectives.

Acknowledgements

Thanks for your consideration.

Francesco Andreace
Paris, August 2024

Contents

List of Figures

List of Tables

Chapter 1

Introduction

r esfesce:citnitorro: 2

As sequencing is every day more accessible in terms of cost and better in terms of accuracy, increasingly more high quality assemblies are being produced for organisms of the same species. This is also coming at lower processing time: the Human Genome Project took 13 years to produce its result in 2003, while now genomes of similar quality are being produced by the hundreds in a few years time. In addition to this, Project like the UK BioBank, which sequenced the genome of 100 thousands people, pose the challenge of how to efficiently analyze jointly all the genetic data of a relatively similar population. Genomics software has been developed under the assumption that only one or few top grade sequences of a single species were available as reference to help analyzing relatively limited amount of new data. This means the new publicly available information is currently not used to improve the quality of present studies. Moreover, the velocity and heterogeneity of new sequences depoisted in public databases, like ENA or SRA, makes current algorithms unfit to jointly analyze rapidly the population data that can be inferred from them. For this reason, a new transformative approach is emerging beyond the single reference genome: pan-genomics. Its aim is to reduce the observational bias of current genomic analyses by capturing the entire genetic diversity within a single population, species or group of similar ones, into a complex representation called pan-genome. To do so, novel computational methods are therefore needed to process up to thousands large and complex genomes. As the research field of computational pan-genomics is relatively new, at the present moment there is no one-fit-all solution that satisfies the requirements and enable straightforward downstream analysis for all genomics applications. In fact, different models are being used to tackle the various problem of representing, indexing, compressing and The model that is currently more researched is an improvement of a sequence graph, called variation graph, in which relationship between shared parts and differences in genomes are modeled by nodes and edges. Alternately, one known way of efficiently represent genomic data is to decompose sequences of variable length, known as reads, into tokens of fixed size, called k-mers. This name comes from the use of an arbitrary value, namely k, that is the fixed length of such tokens. The k-mer model has proven its effectiveness in many bioinformatics techniques and particularly for assembling genomes from reads of a particular species. More recently they are gaining success by enabling superior processing of complex data such as ancient DNA or metagenomes from marine samples compared to other methods. Many data structures have been proposed to represent sequencing data into sets of k-mers, each of them with its strengths and limitations. In this dissertation I present my work on analyzing, developing and applying computational methods, mostly k-mer based, for pan-genomics.

this needs more iterations

1.1 Sequencing data

1.1.1 Next-generation and third generation sequencing

1.1.2 Reads

1.2 *k*-mers and how to store them

1.3 Graphs

1.4 De Bruijn Graphs

1.4.1 Colored and Compacted De Bruijn Graphs

1.5 Variation Graphs

Chapter 2

Pushing the limit of pan-genome construction methods

rseefcs:epcutsihoim:g3

In this first chapter we present and discuss two situations in which we have pushed the limit of pan-genome constructing methods. In the first one we analyzed the current state of the art methods available at the moment and stress-tested them to generate what was, to the best of our knowledge, the largest human pan-genome produced at the time. The second one was the generation of a yeast pan-genome reference for the species *Lodderomyces elongisporus*. In order to best capture the suspected rearrangement events between 3 chromosomes, we had to modify one of the best known pangenome construction pipeline: this lead to challenges and discussion about how to achieve balance between biological correctness and genome variation resolution of the pan-genome.

2.1 Pangenomics and pangenome graphs

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. 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 telomeres and centromeres [**human-pangenomics-era**], producing a reference only 92% complete[**t2t**]. This problem was only solved in 2022 [**t2t**].

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.

2.2 Motivation

The following paper originates from a discussion early in my PhD journey, on which are the best suited tools for large cohort pangenomes of species with large genomes, like animal or pants. As pointed out in the introduction, there is no one-fits-all solution and most of the tools, at the time of the analysis, were freshly released or distributed under development. It was therefore
In order to evaluate the current state of the art of pangeome builing tools, we decided to perform a thorough assessment of the best available methods by giving as input the largest dataset we could produce to mimic the conditions that they could be required to be used in the near future. We decided to test on human data because of the significance and usefulness of pangenomes of our species.
There are multiple ways of representing a group of genomes to be analyzed or used jointly. One that took traction in the last few years has been graphs. Graphs can represent the sequences as labels of nodes, relationship between them (adjacency or overlap) as edges and infer difference in the genomes as different collections of nodes in the graph. We specifically focused our attention on the most used ones, variation graphs and De Bruijn Graphs, In variation graph edges represent adjacencies, i.e the genome is spelled by a walk on nodes connected by an edge. In De Bruijn Graphs they represent overlaps i.e. the suffix of a node is the prefix of the next node connected to it: this implies that edges can exists between nodes that are not adjecent in the genome. As discussed in the next sessions, this distinction implies several differences in how these graphs can be used for downstream analysis.
In this article, we surveyed the the methods and tools that build such graphs, then tested them on different datasets and finally analyzed their features. The result is a small guide on which are the best applications for each of these tools and which are the weaknesses they suffer.
The work we performed was intended for publication, but, to the best of my knowledge, the manuscript has never been put in production.

2.3 Current limitations: a *Lodderomyces elongisporus* pangenome reference

Here I present another example of pushing the boundaries of variation graph pangenome building tools, precisely in building graph that represent inter-chromosomal events like rearrangements. This was done when working with the consortium I belong to, the EU Commission founded Marie Curie ITN Alpaca consortium, on builindg a pangenome reference for the medically interesting yeast strain of *Lodderomyces elongisporus*.

We sequenced, in collaboration with a lab at the Comenius University in Bratislava, Slovakia, 11 different samples using ONT. The computational work I present here was done in collaboration with members of the consortium, in large part by Simon Heumos, whom I would like to thank for the time spent discussing and working together. This work was another example of how difficult it is to produce biologically significant pangenome graphs with th current state-of-the-art tools and offers insights on other areas that should be improved.

2.3.1 *Lodderomyces elongisporus*: genetic characteristichs and interest

Lodderomyces elongisporus is a diploid yeast that has been isolated from, among many sources, humans and it is recently emerging as pathogenic. It is phylogenetically placed in the Candida clade and the side of its genome is usually between 15 and 16 Mb, 2 orders of magnitude smaller than a human genome [**Lodderomyces**]. Its DNA is organized into 8 chromosomes, named from A to H, of different length from around 3.5Mbp of chr A to 800 Kbp of chr H and a 35 Kbp mitochondrial DNA. Our analysis shows that it has a stable core genome of 13Mbp. Increasing reports of (mostly bloodstream) infection in mainly immunosuppressed adults makes it an increasingly important subject of studies: an outbreak war reported occurring in a neonatal ICU in Dheli, India from September 2021 to February 2022 with 1 death[**lodelo_india**].

2.3.2 Building ad hoc pangenome reference using variation graphs

Given high quality assemblies generated by the sequencing information we got of these 11 samples, we decided to build a pangenome graph using different tools. As noted in the previous section, dBGs are very easy and simple to generate, but their usefulness is limited for visual analysis for complex biological events interpretation and study. To this end we also decided to produce variation graphs using first `pggb` and then `Minigraph-Cactus`, in a similar way to what has been done with the Human Draft Pangenome Reference [**hdpr**], to then inspect the graph and gain knowledge of the kind of variations present between the samples.

It is important to notice that in order to produce the best biological correct result, several rounds of parameter tuning and manual curation are needed, with

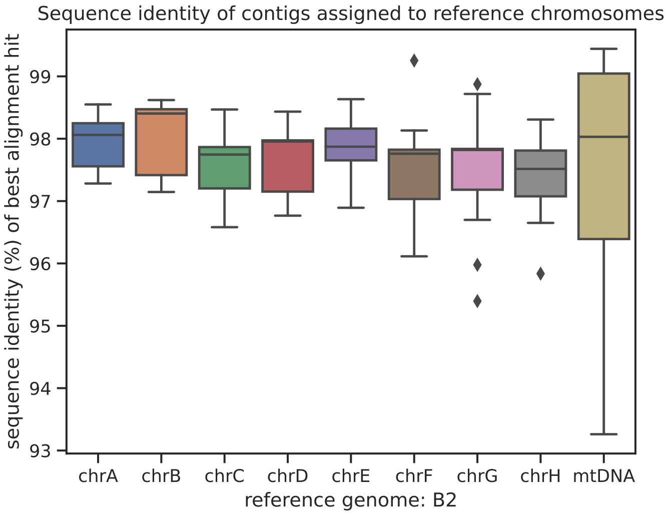

fig:lodelo_alignment_scores

Figure 2.1: Sequence Identity of contigs assigned to reference chromosomes.

knowledge far superior of the one of a normal user: to this end, Simon help was paramount. Samples were called in alphabetical order from A to K, followed with a number greater or equal than 0 that flagged the version of the assembly. Moreover, one sample, B2, for which a member of the consortium produced a high quality assembly after manual curation, was used as relative reference. Another sample, J, was fully resolved into chromosomes while the others were divided into smaller contigs.

2.3.2.1 Determining chromosomal communities

As variation graphs pangenome construction pipelines use mapping or alignment to infer graphs, the first step consists in grouping together the sequences of all the genomes by chromosome, in order to peform computation separately for any chromosome: each group of sequence will be in processed in a isolated way compared to the one of other groups. The final output graph will have (at least) a separate connected component for every group give as input: this means that without any pre-processing, no inter-chromosomal event can be detected. As the rest of the assemblies, with the exception of the J2 sample, were not resolved into single chromosomes, each of them was aligned to the reference B2 using `wfmash` alignment segment size of 10k and 95% sequence identity. The identity scores of the alignment of the genomes to the reference B2 assemby is shown in picture **??**. We detected chromosome-crossing syntenies for multiple contigs in alignment hits of chromosomes C, G, and H, indicating that they correspond to a singular same recombination group: this was consistent with the genomic

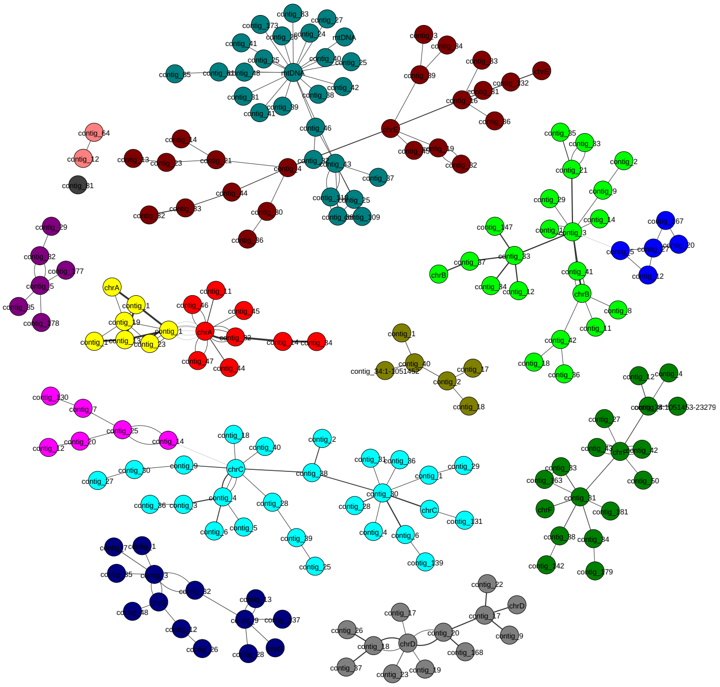

fig:lodelo_communities

Figure 2.2: Community partition of the contigs based on alignment scores.

study, based on SNPs, performed on the isolates related to the Dheli outbreak. In their finding, this is more frequent in hospital/patient populations than in fruit ones [**lodelo_india**]. This event was not straightforwardly identifiable from simple community separation, like the one suggested in the manual of `pggb`. We therefore decided to run the pipelines with just one single community for chromosomes C,G and H. In figure **??** it is displayed the community partition of the assemblies based on Louvain algorithm. The latter does not show any condensation of the 3 chromosomes into a single community: this reinforces the fact that "manual" inspection of the alignments is still required in absence of high quality assemblies in order to produce biologically valuable graphs.

2.3.2.2 Customizing the pipelines

As `pggb` uses all-vs-all alignment of a collection of sequence as first step to infer the graph, it enables the representation of recombination among chromosomes placed inside the same community, as seen also for human acrocentric chromosomes [**Guarracino2023**].

This is not the case for the other well known pipeline for variation graphs construction: `Minigraph-Cactus`. As the first step is based on Minigraph, it

must have a single reference sequence for any collection of sequence that has to be considered together, be it a single chromosome or a group of them. This means that there is no feature to have chromosome C,G and H considered together in input.
To try to overcome this limitation of the approach we tried to produce a graph that respected the condition of having the three chromosome inside the same connected component of the graph, at the cost of biological correctness. We therefore produced a chimeric contig consisting of the concatenation of the three chromosomes assemblies of the B2 sample. The rationale was to provide it as a backbone so that there is a single reference for the `Minigraph` construction step. The expectancy was to therefore produce a graph that showed the recombination from the mapping of the contigs of the other genomes.
By building a graph using `Minigraph` with the chimeric chromosome CGH and all the contigs of the other genomes assigned to chromsome C, G and H does not represent any recombination event, as can be seen in figure [**fig:cgh__mingraph**]. This was somewhat expected, as it is known that `Minigraph` does not also consider inversion between genomes. When the complete modified pipeline of `Minigraph-Cactus` is run, it is possible to see the tangle between the chormosomes, as shown in figures **??**

2.3.3 Focus on the tangle

2.4 Conclusion and Perspectives

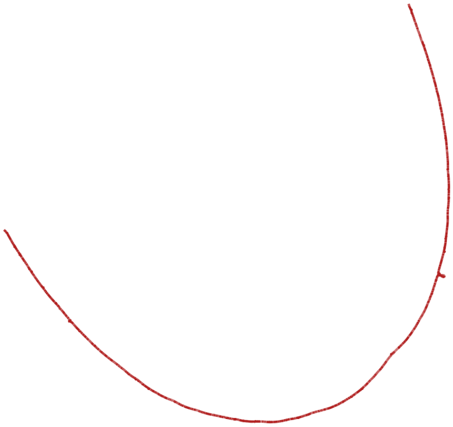

Figure 2.3: Graph of chimeric chromosome CGH from sample B2 and all the contigs of the other genomes aligning to it produced with `Minigraph`. The graph is linear and no inter-chromosomal event is visible.

fig: cgh_mingraph

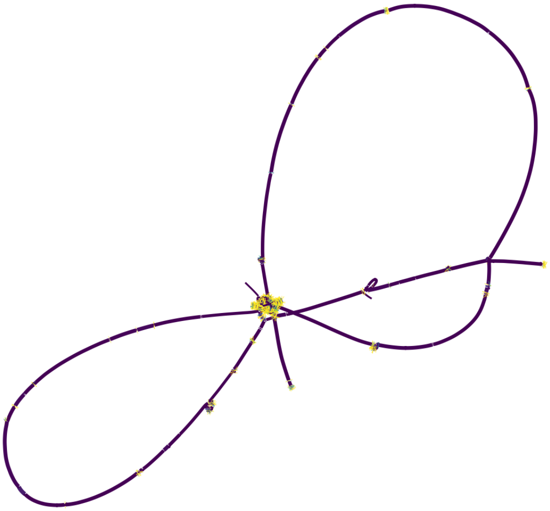

Figure 2.4: The graph after all the other steps of the `Minigraph-Cactus` pipeline, coloured by depth, after simplification of variants < 1kbp using the command `gfatools asm -b 1000 -u`. The large recombination event is now visible.

fig: cgh_mcactus

Figure 2.5: Difference in output between `Minigraph` and `Minigraph-Cactus` of the chimeric graph produced to visualize the inter-chromosomal event between C,G and H.

fig: chromosome_ cgh_minigraph

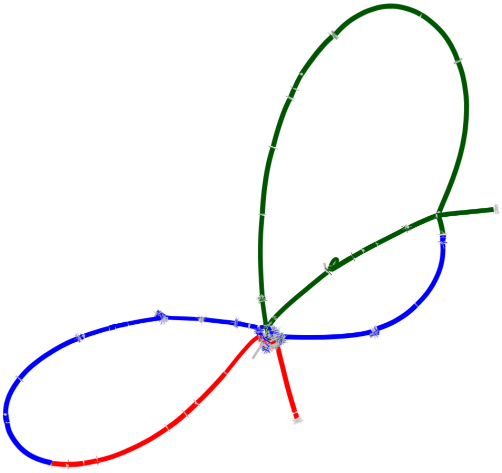

fig:lodelo_graph_tangle

Figure 2.6: The tangle of chromosomes C, G and H in the `Minigraph-Cactus` variation graph. Nodes are colored based on `Minigraph` alignment of chromosome C (dark green), G (blue) and H (red) of the reference assembly B2. The three chromsomes are bound together because of the construction.

Chapter 3

Reducing the complexity of De Bruijn Graphs for Pangenomics

secfcomplexity

3.1 Too big, too complex, too difficult to use

3.2 DBG-based tools for pangenomics

3.2.1 Requirements

3.2.1.1 Bacterial pangenomics

3.2.1.2 Human pangenomics

3.2.2 Representing sets of *k*-mer sets

3.3 Linearize a masked human reference pangenome graph for visualization purposes

3.4 Converting ccdBGs into unitig matrices for downstream analyses: muset

3.5 Quotient filters to represent kmer sets: optimizin existing methods

3.6 Perspectives and future work

Chapter 4

Perspectives and future work

refsection:5
perspectives

4.1 On human pangenomics: graphs and beyond

The result of the analysis I conducted in the first phase of my PhD, presented in chapter **??** serves as basis to understand what are the features, the limitations and the usefulness of the software that is currently used or developed to build pangenome graphs. These are based upon the latest developments in terms of computer science algorithms to provide the best computational performance now possible and represents a huge leap compared to the currently standard software used for genomic analysis. Here I will present a few considerations and perspectives that stem from this as well as from 2 more years of thoughts and discussions with peers of my doctoral program, my supervisors and other colleagues and experts in the field.

As we are possibly at the beginning of a change of paradigm between linear reference sequences and genomics analysis to pangenome references and pangenomics analysis, there are a few things that need to be adressed as soon as possible.

Reproducibility and stability of computation has to be the main focus of the next years for pangenome reference software developing.

In the case of leading general-purpose pangenome reference building tools, like `pggb` and `Minigraph-Cactus`, that produce variation graphs, it is of upmost importance that the graph generated from a set of sequences is exactly the same when the same data is fed as input. This means that the heuristics used to generate the variation graph are independent of the ordering of the input sequence and do not contain any stochastic process that might alter the structure of the graph. If a tool can produce two variation graph that can spell the same input genomes but that do not have the same internal order, downstream analysis, like read mapping, loci visualization and other application become biased toward the graph, making it less desirable to genomic analysts that rely on the stability of a linear reference. Current liftover and graph-mapping solutions, in my view, can only be a temporary solution if pangenomics is to be adopted and fully accepted in the genomics and genetics field.

There should be guarantees or estimates on the overall biological correctness of pangenome graphs.

While `Minigraph-Cactus` omits centromeric variation, `pggb` does at the cost of producing more complex graphs. The trade-off is not trivial as gaining on "variation resolution" leads, also, to graphs that are more difficult to interpret, especially as the number of input genomes increases. Moreover, De Bruijn Graphs are difficult to untangle and understand already at small case. A very useful and interesting future development would be to design a method to evaluate thoroughly the (computational and biological) quality of the pangenomic data structure produced. This tool would be a necessary Quality Contro (QC) step in all custom-pangenome based analysis. For human pangenomics, this would be useful for application where a different reference compared to the HPRC precomputed one is needed, for other cases, like bacterias, virus or fungi, where only specific strains and not the entire species is to be considered.

De Bruijn Graph methods need a common color file format or interface to push the development of application-specific tools. Mathematically clear, computationally efficient and output-stable de Bruijn Graph methods, like the ones that use colored-compacted dBGs, won't succeed in being real alternatives of alignment based software to perform pangenomic analysis if there won't be a consensus between the main developers on at least a minimum common interface that let users write tools to exploit the information they contain. Standardized file formats [**kff**] and interfaces for (colored) queries would help other researchers commit into developing tools for k-mer based approaches, independently of the latest tool in the scene. At the current moment, the landscape is quite diverse and new tool are constantly being developed, discouraging, in my view, the needed investment of resources to develop tools for dbg-based downstream analysis. Writing software for genomics application of k-mer based pangenome representations is crucial to make this representation useful to the end users. As the representation of references is better suited to variation graphs, applications of k-mer based tools could provide added value on genomic studies of specific (sub-)populations.

Graphs are not the only *k*-mer based pangenome representation. For this specific use-case, other data structure based on k-mers can be used to extract valuable insights. As already described, unitig matrices are a new powerful example of k-mer based data structures that can represent the genetic content of a population and its diversity. k-mer matrices [**kmtricks_2022**] can be seen as a pangenome where rows are the k-mers present in the whole populations and columns are the specific individuals. Then the value in the matrix could be an absence/presence binary value, defining a *de facto* euqivalent representation of a colored dBG. Unitig matrices then contain the same information of a cdBG.
In my opinion there is great value in being able to demonstrate equivalence between such representation and to develop tools to change of representation such that, depending on the specific application, each user can decide which is of more interest and not be limited by specific tools building specifically formatted data structures. This is another interesting path forward in the field.

4.2 Indexing data structures and metagenomics

Chapter 5

Conclusions

refsection:6
conclusions

This manuscript is the result of 40 months of doctoral journey. The common theme is the use of k-mer based algorithms and tools to tackle complex genomic problems that consider multiple different samples at the same time. This effort was always driven by the desire to produce software or analyses that would help answer biological questions.

I hasn't been a straightforward path as many projects I embarked in did not get to a stage of meaningful contribution to the scientific community. Nevertheless such efforts taught me important lessons about working alone or with other colleagues and strengthened my understanding of other genomics and bioinformatics fields.

I reckon that the scope of this doctoral thesis might seem quite broad, starting with an analysis of computational pangenomics methods for human genomes representations and ending with data structures to represent kmers in a cache efficient way. In my view this is the result of constant curiosity about the whole sequence bioinformatics problems tackled by the team where I was and the reflection of a comprehensive and multifaceted approach to the challenges of the field.

We proposed an analysis on the construction and representation of pangenome graph from high quality human haploid assemblies was well received by the community as it shed light on the characteristics of both their internal representation and the methods to generate them. It stresses the importance of selecting the kind of graph that best fits the particular application, specifically in the way it represents variations in the DNA sequence of the individuals. Variation graphs are better to perform specific downstream analysis and are more intuitive to understand, manipulate, visualise and analyse while de Bruijn Graphs are more efficient to generate, scale better and give guarantees on the preservation of the input sequences. Finally, this work stresses the difficulties of proposing a one-fits-all solution and points out areas of research that would make kmer based approaches more attractive to the genomic community. I believe pangenomics is the key of solving many issues and deficiencies present in current genomics approaches: it is a novel area that is evolving now and will need much more effort to produce viable solutions to all the genomics tasks biased by the use of a single reference sequence.

My work on the Backpack Quotient Filter has been focused on coding the underlying data structure, the Rank Select Quotient filter, in a way that could be easily used and manipulated for different high level interfaces that would implement different filters. The final method proposed on top of this data structure a particular encoding of the counting information together with the integration of the Fimpera scheme for kmer storing and retrieving. My work lied on the demonstration of the efficacy of the implementation I wrote by recreating on top of it the Counting Quotient Filter as originally conceived together with the Fimpera scheme. This demonstrated that implementation differences affect the magnitude of the data such data structures can analyse, even if the information provided as output is the same. This work served as confirmation that improvements in the representation ok kmer sets can drastically change the way in which large data collections can be interrogated and explored, even if very complex. In order to improve in the future such analysis power, new tools will require enhanced methods and refined coding techniques to exploit the maximum out of computational power to analyse exponentially increasing amounts of data.

As the understanding of public metagenomic repositories is important when developing software that is meant to encode their informations, I also helped in the writing of a section about such repositories in a method primer review. My contribution to this large collaborative effort was to provide to potential users a first, concise and expert overview on which are the main efforts to store, categorise and analyse public metagenomics data.

Finally, I gave my contribution to other projects, like MUSET, that is a pipeline for the construction of uniting matrices (both abundance and presence/absence). This is the first effort to produce such data structure and is an advancement from kmer matrices, that contained almost the same information while being more complex to analyse. This data structure can serve different applications, from pangenomics to metagenomics data as well as cancer transcriptomics, as it provides a very clear representation of the genetic content of multiple samples and their difference that is easily accessible by data analysis and machine learning tools.

The tools, pipelines and data structures I analysed or coded are still not the definitive answer to any of the problems that pangenomic and metagenomics still face but are a clear demonstration of what are good steps in the right direction. In the future there is plenty of work to do. In the pangenomic field, methods to demonstrate the biological correctness of pangenome graphs, like it is now done with assemblies, as well to scale pangenomic representation to handle large collections of eukaryotes. In metagenomics new methods will need to follow the pace of the rapid increase of data that is being sequence everywhere in the world. Finally, all the methods that will need to analyse and compare multiple individuals or samples will need to use kmers as part or as bedrock of their development, as I think this thesis proved, are a simple, versatile and effective way of representing DNA sequences.

Papers

Appendices