Population genomics analyses on pangenome graphs

Flavia Villani[1], Francesco Porto[2], Andrea Guarracino[3], Robert W. Williams[4], Pjotr Prins[4], Gianluca Della Vedova[2], Erik Garrison[5], Vincenza Colonna[1]

[1]National Research Council, Institute of Genetics and Biophysics Adriano Buzzati-Traverso, Napoli, Italy; [2]Department of Informatics, Systems, and Communication, University of Milano-Bicocca, Italy; [3]Centre for Molecular Bioinformatics, Department of Biology, University Of Rome Tor Vergata, Rome, Italy; [4]Department of Genetics, Genomics and Informatics, College of Medicine, UTHSC [5]University of California, Santa Cruz, US

Abstract

Introduction

Population genomics is the study of the causes and the consequences of genetic variability within and among populations. Population genomics is based on the study of variable sites. The accuracy of the inferences made by population genomics analyses is strictly correlated to the amount of information on genetic variation. For this reason, the field of population genomics has been particularly active in the last ten years, due to the unprecedented availability of genomic sequences that made possible the identification of millions of novel genetic variants [1, 2, 3].

Nevertheless, most of the population genomics studies are based on genomic variants which are simple to detect like single nucleotide variants and and a very few studies have taken in consideration complex structural variants so far. This is mostly due to the inability to have reliable data set of complex structural variants, a limitation that is now being tackled by the use of long-reads sequence technology and pangenomes.

Standard approaches in sequence analysis relate sequences to a single linear reference genome. Sequence fragments produced by NGS technologies are mapped and assembled against a reference genome, and genetic variants are identified through comparison with it. While this is an efficient way of processing sequence information, the approach has a fundamental problem, i.e. substantial difference from the reference sequence are hard to observe and describe. As pangenome we refer to the entire set of genomic elements in a given species or clade. Pangenomic methods allow us to overcome limitations of the use of the reference genomes, relating all genomes directly to each other; sequence and variation are combined [4]. In pangenome variation graphs, genetic variants appear as bubbles. These sites have a common starting context (a single inbound node), a common exit point (a single outbound node), and a diversity of possible paths that connect the two, each of which represents an allele [5]. We consider these bubbles in the context of a data model developed to represent the basic components of variation graphs, the handlegraph abstraction [6]. This data structure breaks down the elements of a variation graph and proposes a programming interface based on them. We use this data model as the basis for algorithms to find bubbles, their alleles, and the frequencies of these alleles among the genomes embedded in the graph.

Pangenomes are large, and unwieldy to work with as raw collections of sequences. One possible approach to processing them considers the collection of genomes and their mutual alignment in a compact, graphical model. As a lossless representation of the pangenome and its embedded sequence variation, these variation graphs should in principle support any kind of population genetic analysis that would be completed on simpler representations of the genomes and their variation. But, because this pangenomic approach is quite recent, the software for

population genetic analyses currently available are still mostly based on genomic data in the linear format.

Here we respond to this need by implementing a VGPOP, a set of tools for population genetics on genome graphs. At a high level, our work has two parts. We first uncover genomic variation embedded in pangenomic variation graphs by developing and implementing straight forward algorithms for bubble detection on variation graphs. We then demonstrate the calculation of basic population genetic parameters over variation graphs.

Methods

Implementation of the VGPOPlibrary We developed a library named VGPOP to conduct standard population genetics analyses using pangenomic data models. Typically represented in the Graphical Fragment Assembly (GFA) format [7], these models can represent whole genome alignments in a compact graphical structure. The library is written in the Python programming language under MIT license; the code is publicly available on GitHub (`https://github.com/Flavia95/VGpop`). Currently VGPOP has three sets of functions detailed in the following paragraphs.

1. Functions for the identification of variable sites The first mandatory step for any further population genetics analysis is to extract from the graphs the information about variable sites, i.e. the regions where more that one type of sequence is present. Any population genetic analysis is indeed based on the information contained in the variable segments of the sequence and their occurrence in the population under investigation. Because of their appearance in the pangenome graph, variable sites are referred to as bubbles. We implemented two main functions for bubble detection, namely BUBBLEPOP and BUBBLECALL.

The **BUBBLEPOP** function takes as input a GFA file and gives as output a dictionary, i.e. a table of correspondences between region of the graph and sequence variants. It explores the graph using the two recursive algorithms, the Depth First Search (DFS) [8, 9] and the Breadth First Search (BFS) [10, 11].

In BUBBLEPOP, we run BFS on the tree obtained by DFS. Starting from the tree root, the DFS explores the tree until it finds a bubble, that is a pair of nodes whose distance from the root is the same. When this happens it calculates the distance from the root of all the nodes in the bubble.

At the beginning of the bubbles all paths share the same identical node, and this is true also at the end of the bubble.

Once the pangenome has been decomposed with BUBBLEPOP in a tree whose information on the node distance from the root is stored in dictionary, **BUBBLECALL** explicits the content of the bubbles and its position in relation to a chosen reference sequence in three steps:

1. Choosing the reference path - We consider all the possible paths that connect the initial node and the final node of each bubble, the first path in the GFA file is chosen as reference (REF)

2. Variant identification - In this step BUBBLECALL iterates over all available nodes to analyze paths within the node and compare them with the reference node. BUBBLECALL considers all the possible paths pairs(x,y) in which x is the REF and y is any other path. A node is called as a variant if:

   - (i) it is supported by at least one path;
   - (ii) the node sequence is different from the sequence of the corresponding reference node;
   - (iii) if its distance from the root is the same as the one of the reference node, than the variant is classified as a Single Nucleotide Variants (SNV);
   - (iv) if its distance from the root is smaller that the one of the reference node then the variant is classified as a Deletion;
   - (v) if its distance from the root is greater that the one of the reference node then the variant is classified as an Insertion.

3. Variant positioning - This step defines the position of a variants with respect to the reference sequence. When the two paths were used to call the variants, the length of the sequences was taken into account in order to map the variants on the individual paths.

Re-implementation in Rust Population genomics analysis requires the study of a large number of individuals of any species, therefore the pangenomic approach has to be implemented in a way that is applicable to graphs of any complexity. For this reason, we decided to re-implement the core functions of our library in Rust; this project is publicly available on GitHub at `https://github.com/HopedWall/rs-gfatovcf`. Rust is a programming language which allows us to build reliable and efficient programs when compared to other languages, such as Python, which was used for our original implementation.

In order to achieve scalability, the following changes were made:

1. we employed a non-recursive strategy for building the spanning tree, since the original procedure required an excessive amount of memory on large graphs. The Rust implementation, instead, uses a queue-based approach, which prevents this type of problem.

2. we introduced as a parameter the maximum amount of edges to traverse during the BUBBLECALL step. This is required since finding all paths between two given nodes is a problem which is known to be NP-hard, hence it may take exponential time. This change limits the running time but might result in missing some paths.

3. introduced the ability to set only specific paths as references, avoiding the variant identification with respect to all the paths in the graph. This should increase performances when simpler analyses are required.

2. Functions for format conversion

The **GFA2VCF** function of VGPOP takes as input a graph in the GFA format and outputs a corresponding linear representation in the VCF format, i.e. the file format that is currently used to store sequence information on variable sites. To do this *gfa2vcf* uses first BUBBLEPOP to decompose the pangenome in a tree and then BUBBLECALL to identify the variable sites. Finally the dictionary of the variable site is formatted according to the vcf specifications.

3. Functions for population genetics

GFA2ALLELEFREQ - The frequency of an allele is an indication of how common the allele is in a population. It is calculated by counting how many times the allele appears in the population, divided by the total number of copies of the gene. The code we developed for the GFA2ALLELEFREQ function of VGPOPtakes as input a graph in GFA format and a metadata file (with information on paths, individuals, and populations), and outputs a file that contains the allele frequencies for variable loci per each population. In GFA2ALLELEFREQ the allele frequency corresponds to the number of paths that support a node (i.e. a variant) divided by the total number of paths actually realized. The frequencies of monomorphic nodes (i.e. frequency = 1) are not reported. GFA2ALLELEFREQ first uses bubblepop and bubblecall to read the graph, and than applies calculation of frequencies.

GFA2FST - The Wright's fixation index (Fst) is a measure of population differentiation due to genetic structure [12]. It is estimated from genetic polymorphism data, such as SNV or microsatellites. Several formulae exists for its calculation among which the one that estimates it as the standardized variance of allele frequencies among sub-populations. The code we developed for the GFA2FST function of VGPOP, takes as input an allele frequencies file, and as output a file that contains the calculation of Fst. GFA2FST first uses BUBBLECALL and BUBBLEPOP to read the graph, and than applies GFA2ALLELEFREQ to calculate allele frequencies and than calculates F_{st} as the standardized variance of allele frequencies among subpopulations: F_{st}= s_2/p (1-p) with s_2 and p being the variance and mean, respectively, of the allele frequencies.

GFA2TAJIMASD - The test statistic developed by Tajima [13] allows to identify non-random evolution of DNA sequences and consists in the ratio between two estimate of the effective population size (i.e. a measure of genetic diversity [14]): the number of segregating sites and the nucleotide diversity. The code we developed for GFA2TAJIMASD takes as input a GFA and outputs the corresponding value of the test statistic.

Results

***Calculation of F_{st} on simulated data using* GFA2FST**. To test if the calculation made by VGPOPare accurate, we applied VGPOPfunctions to data for which we can predict ranges of expectations for the parameters calculated by VGPOP. In particular we used sequence data produced by simulation under a known demographic scenario of two populations separating from a common ancestral population to measure the degree of separation calculated as F_{st}.

As simulation scenario we considered a model adapted from [15] with two diploid populations separating without subsequent migration. The first population is bigger in size compared to the second, and through time develops maintaining constant size until 5k generations ago when it starts to exponentially expand. The second population develops through time maintaining constant size. We considered three possible scenarios for separation time: 5k (T1), 10k (T2), and 15k (T3) generations ago. The expectation is that the longer the separation time, the higher will be the F_{st}, with scenario T3 having the higher F_{st} compared to T2 and T1.

We used the software ms [15] to produce 100 replicates of simulated variable sites in a 10kb region for eighty individuals under each of the three scenario. The variable sites were transformed in sequences that include also the invariable part (using Seq-Gen [16]) and the sequences were used to reconstruct the pangenome of the simulated data that was then processed with the GFA2FST function of the VGPOPlibrary. F_{st} calculation was validated using vcftools REF, that uses a different F_{st} formula.

We found that the F_{st} trend vary according to expectation of the three simulated scenario, i.e. the lowest value is found at T1 and the highest at T3. We observe the same trend when calculating F_{st} with a different formula as a control. Nevertheless, the absolute values of F_{st} obtained from VGPOPare lower that those obtained from vcftools, suggesting that a further comparison would be required to fully clarify the discordance and improve the VGPOPlibrary.

Allele frequencies at variable loci of the human HLA region using GFA2ALLELEFREQ The HLA region is located on the short arm of chromosome 6 from 6p21.1 to p21.3 in a region spanning 7Mb. The class II region includes genes for the α and β chains of the MHC class II molecules HLA-DR, HLA-DP and HLA-DQ. In addition, the genes encoding the DMα and DM β chains, as well as the genes encoding the α and β chains of the DO molecule (DO α and DO β, respectively), are also located in the MHC class II region [17].

In the latest version of the human reference genome (GRCh38), there are alternate loci highly polymorphic where the sequence variation is too complex to be represented with a single sequence [18]. These loci are known to co-segregate with disease and are therefore of great interest in population genetics. Sequence reads alignment in the HLA region, is known to be particularly difficult, particularly in regions originating from highly polymorphic regions and regions absent from the reference genome.

We considered three genes of the HLA region, *HLA-E*, and *HLA-DMA*, and *HLA-C*. For these three genes we started from eleven (*HLA-DMA*), nine (*HLA-E*), and ten (HLA-C) sequences downloaded from GenBank. We used the sequences to reconstruct the pangenomes. The pangenomes of *HLA-E* and *HLA-DMA* are less complex compared to the pangenome of *HLA-C*, suggesting less diversity in these two genes compared to *HLA-C*. We used the pangenomes to detect of variable sites (bubbles) with BUBBLEPOP, and then the allele frequencies are calculated as the number of paths supporting the variant node divided by total number of paths using GFA2ALELLEFREQ.

Variant identification in Sars-CoV-2 using rs-gfatovcf

Since the main motivation for re-implementing GFA2VCF in Rust was the ability to use it on larger graphs, we considered the Sars-CoV-2 pangenome available at `http://covid19.genenetwork.org/` in GFA format. This pangenome is composed of sequences of approximately 1.2 GBytes and with 78571 fragments, obtained from 15127 genomes. rs-gfatovcf is capable of obtaining a VCF file from it in 16 minutes on a machine with 256GB RAM, we found 294626 variants. While this result is satisfactory, we want to exploit concurrent and parallel computing to reduce its running time.

Conclusions

We have presented the results of our project to develop VGPOP, a library for population genetic analyses based on pangenome graphs.

The use of pangenomes and variation graphs is one of the major changes in genomics. Because this approach is quite recent, there has been little focus on developing software for population genetic analyses from pangenomes, and in fact almost all the already available software is based on genomic data in the linear format. With our project we contributed to fill this gap by writing software for population genetic analyses able to deal with pangenomes.

Two function of VGPOP, BUBBLEPOP and BUBBLECALL, have the primary function to parse the pangenome and identify the variable sites (bubbles). These two functions are exploited by some of the others, like GFA2ALLELEFREQ and GFA2FST that instead produce population genetics summary statistics. Finally, other functions, e.g. GFA2VCF are utility to convert file formats. This set of functions does not cover all possible needs for population genetic analyses, but it shows that several types of function are required to cover all possible tasks.

We first tested VGPOPon simulated data where we could rely on known expectations. We demonstrated that with VGPOPwe can reliably estimates genetic distance between a pair of populations in three scenarios of increasing genetic diversity, using as a measure of diversity F_{st}, one of the basic summary statistics in population genetics. We also demonstrated that VGPOPcan calculate allele frequencies in regions of the genome with complex genetic variability, such as the HLA region, a complex variable region due to the high degree of similarity and polymorphism of its genes. The range of complexity in the variability of the HLA region made it also possible to test the limitations of VGPOP. Finally, we focused on the Sars-CoV-2 pangenome, which we chose for its current international relevance. This pangenome also acts as a benchmark for what the Rust version of VGPOPcan do, as it targets bigger graphs, which would cause memory problems in the original Python implementation.

Overall, with our project we demonstrated that VGPOPcan calculate the basic statistics for population genomics inference directly from pangenomes. VGPOPis able to process pangenomic data, therefore putatively access complex variants scantily considered so far in population genomics. Even if in its current form VGPOPis only effective with simple variants, it has the potential to be adapted also for more complex ones. To our knowledge, this is the first such exploration that has been undertaken in the scope of this representation. Our work suggests a series of follow-up studies to extend related population genetic metrics to pangenome models. We hope to explore the development of haplotype-based scans for genetic selection (e.g nSL[19], iHS[20], and xp-ehh[21]) to pangenome graphs, as well as other measures of frequency differentiation between populations could be applied to alleles in bubbles in the graph (e.g. PBS [22]).

We are also aware of the current limitations of VGPOP, namely (1) the inability to detect complex bubbles and (2) its overall running time on larger graphs. In order to address (1), we are looking into a new bubble detection algorithm [5]. In order to address (2), we plan on exploiting parallel computing, which we hope will drastically improve the the running time of our functions.

References

[1] . G. P. Consortium *et al.*, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, pp. 68–74, 2015.

[2] N. A. Rosenberg, "Standardized subsets of the hgdp-ceph human genome diversity cell line panel, accounting for atypical and duplicated samples and pairs of close relatives," *Annals of human genetics*, vol. 70, no. 6, pp. 841–847, 2006.

[3] K. Karczewski and L. Francioli, "The genome aggregation database (gnomad)," *MacArthur Lab*, 2017.

[4] J. M. Eizenga, A. M. Novak, J. A. Sibbesen, S. Heumos, A. Ghaffaari, G. Hickey, X. Chang, J. D. Seaman, R. Rounthwaite, J. Ebler, *et al.*, "Pangenome graphs," *Annual Review of Genomics and Human Genetics*, vol. 21, 2020.

[5] B. Paten, J. M. Eizenga, Y. M. Rosen, A. M. Novak, E. Garrison, and G. Hickey, "Superbubbles, Ultrabubbles, and Cacti," *J. Comput. Biol.*, vol. 25, pp. 649–663, 07 2018.

[6] J. M. Eizenga, A. M. Novak, E. Kobayashi, F. Villani, C. Cisar, S. Heumos, G. Hickey, V. Colonna, B. Paten, and E. Garrison, "Efficient dynamic variation graphs," *Bioinformatics*, 2020.

[7] "Gfaformat."

[8] R. E. Korf, "Depth-first iterative-deepening: An optimal admissible tree search," *Artificial intelligence*, vol. 27, no. 1, pp. 97–109, 1985.

[9] Wikipedia contributors, "Depth-first search — Wikipedia, the free encyclopedia," 2020. [Online; accessed 12-June-2020].

[10] S. Beamer, K. Asanovic, and D. Patterson, "Direction-optimizing breadth-first search," in *SC'12: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, pp. 1–10, IEEE, 2012.

[11] "Dfs, bfs."

[12] R. R. Hudson, "Generating samples under a wright–fisher neutral model of genetic variation," *Bioinformatics*, vol. 18, no. 2, pp. 337–338, 2002.

[13] F. Tajima, "Statistical method for testing the neutral mutation hypothesis by dna polymorphism.," *Genetics*, vol. 123, no. 3, pp. 585–595, 1989.

[14] D. L. Hartl and A. G. Clark, *Principles of population genetics*, vol. 116.

[15] R. R. Hudson, "ms a program for generating samples under neutral models," *Bioinformatics*, 2004.

[16] A. Rambaut and N. C. Grass, "Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic trees," *Bioinformatics*, vol. 13, no. 3, pp. 235–238, 1997.

[17] R. D. Campbell and J. Trowsdale, "Map of the human mhc," *Immunology today*, vol. 14, no. 7, pp. 349–352, 1993.

[18] H. P. Eggertsson, H. Jonsson, S. Kristmundsdottir, E. Hjartarson, B. Kehr, G. Masson, F. Zink, K. E. Hjorleifsson, A. Jonasdottir, A. Jonasdottir, *et al.*, "Graphtyper enables population-scale genotyping using pangenome graphs," *Nature genetics*, vol. 49, no. 11, p. 1654, 2017.

[19] A. Ferrer-Admetlla, M. Liang, T. Korneliussen, and R. Nielsen, "On detecting incomplete soft or hard selective sweeps using haplotype structure," *Molecular biology and evolution*, vol. 31, no. 5, pp. 1275–1291, 2014.

[20] B. F. Voight, S. Kudaravalli, X. Wen, and J. K. Pritchard, "A map of recent positive selection in the human genome," *PLoS Biol*, vol. 4, no. 3, p. e72, 2006.

[21] P. C. Sabeti, P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, X. Xie, E. H. Byrne, S. A. McCarroll, R. Gaudet, *et al.*, "Genome-wide detection and characterization of positive selection in human populations," *Nature*, vol. 449, no. 7164, pp. 913–918, 2007.

[22] X. Yi, Y. Liang, E. Huerta-Sanchez, X. Jin, Z. X. P. Cuo, J. E. Pool, X. Xu, H. Jiang, N. Vinckenbosch, T. S. Korneliussen, *et al.*, "Sequencing of 50 human exomes reveals adaptation to high altitude," *Science*, vol. 329, no. 5987, pp. 75–78, 2010.