Extended abstract

Enabling multiscale variation analysis with genome graphs

Brice Letcher[1]*, Martin Hunt[1] and Zamin Iqbal[1]

[1] *EMBL-EBI, Hinxton, United Kingdom*
***Corresponding author**: bletcher@ebi.ac.uk

Abstract

Standard approaches to characterising genetic variation revolve around mapping reads to a reference genome, but high genetic diversity leads to biases in mapping and variation detection. Genome graphs have been proposed as a means to address this and alleviate mapping bias. However when genotyping genome graphs, we need to define which variant sites are in the graph and what reference to express them against. Notably, with enough samples or in highly diverse genomic regions "nested variation" naturally occurs- a long deletion which is an alternate allele to multiple SNPs, or diverged haplotypes with small variants on top of each. There is currently no tool that models these relationships and meaningfully outputs variation at multiple scales.

We demonstrate our software `gramtools` can accurately genotype dense variation at multiple scales, outperforming reference-based variant callers and state of the art genome graph tools on two datasets of microbial pathogens. Many species and genes of great interest harbour high levels of genetic diversity where multiscale variation naturally occurs and requires consideration. We provide a new output format for accessing all variation in directed acyclic genome graphs allowing straightforward genotyping of sample cohorts, finer resolution of genetic variation and the definition of alternate references.

Keywords
Genome graphs— Variant calling — P. falciparum — M. tuberculosis

1. Introduction

Genome graphs are graph structures extending single, linear reference genomes with known population genetic variation or candidate variants. They are used as objects that remove reference bias [1] and as objects that enable genotyping across samples at the same variant sites [2].

In genome graphs built from enough samples or in highly diverse genomic regions, defining which variant sites are present and what reference to express them against becomes non-trivial. In particular in such graphs variation starts to appear at multiple scales, with two naturally occurring cases. First, when analysing structural variants and small variants together, SNPs can occur under long deletions. Second, in genes with divergent forms or in long insertions, SNPs can occur on top of alternate haplotypes.

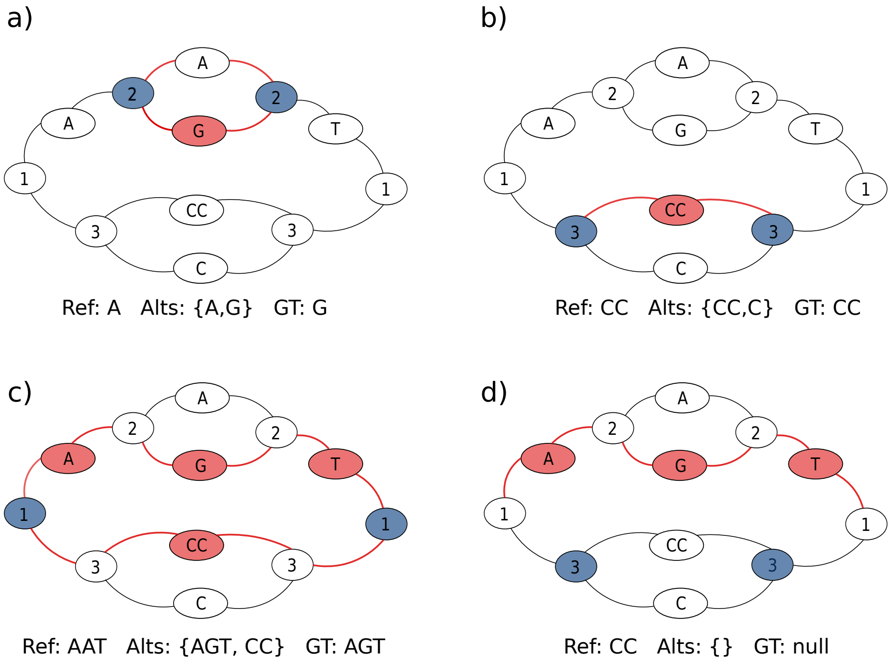
There is currently no tool that models these relationships and meaningfully outputs variation at multiple scales. Here we present a framework to identify, call and output all identified variation in directed acyclic genome graphs using the opensource software `gramtools` (https://github.com/iqbal-lab-org/gramtools). We give applications in two microbial datasets illustrating genotyping performance compared to the state of the art and a new analysis in a previously inaccessible genomic region.

2. Methods

`gramtools` implements a workflow for building, genotyping and augmenting genome graphs. To genotype, we map reads from whole-genome sequencing experiments to a unique data structure developed for `gramtools` [3] and record coverage with awareness of horizontal (genomic repeats) and vertical (allelic repeats) mapping uncertainty.

Genotyping produces three main outputs: a personalised reference genome for the sample, a VCF of called variants expressed against the standard reference genome, and a JSON of calls at each variant site in the graph. The latter includes variant sites which are ”nested” in others and sites which occur on different sequence backgrounds or references.

The algorithm for nested genotyping is illustrated in Fig. 1. We refer to each outgoing branch from a parent site as a **haplogroup**, for group of related haplotypes.



a) Ref: A Alts: {A,G} GT: G

b) Ref: CC Alts: {CC,C} GT: CC

c) Ref: AAT Alts: {AGT, CC} GT: AGT

d) Ref: CC Alts: {} GT: null

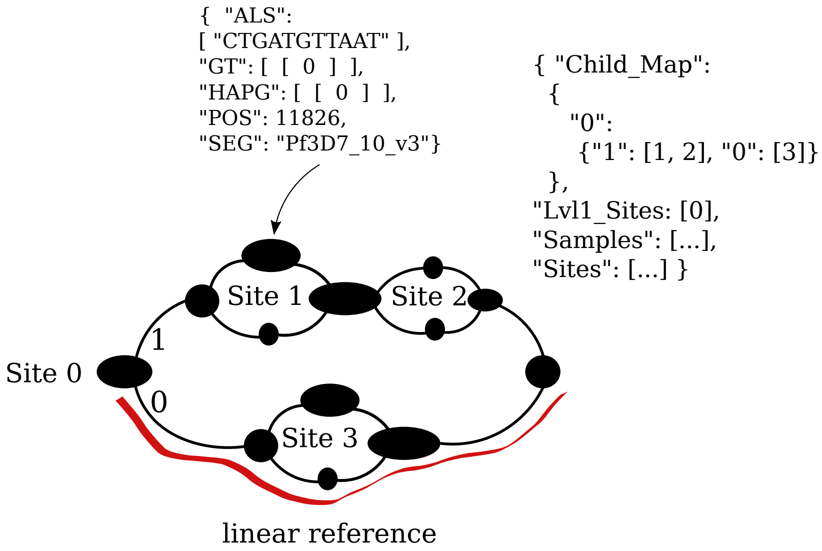
Figure 1. Nested genotyping procedure. Nodes with numbers mark variant sites. In each panel, blue-filled nodes mark which site is being processed, red-filled nodes mark called alleles, and red paths mark alleles considered for genotyping. The example shows haploid genotyping. a. Genotyping of child site 2. b. Genotyping of child site 3. c. Genotyping of parent site 1. d. Invalidation (null calling) of site 3.

3. Results and Discussion

3.1 Multiscale-aware variant call format

We developed a JSON-based output format providing one entry per identified site in a directed acyclic genome graph and storing parent/child relationships between sites. This enables two features. First, it makes incompatibilities between sites explicit allowing genotyping to enforce consistency (Fig.1). Second, it enables defining alternate references based on haplogroups, and which ones variants fall on.

An example is given in Fig. 2. In contrast to VCF, the format also records graph topology allowing queries such as extracting all variant records under a given haplogroup.



```
{  "ALS":
[ "CTGATGTTAAT" ],
"GT": [  [  0  ]  ],
"HAPG": [  [  0  ]  ],
"POS": 11826,
"SEG": "Pf3D7_10_v3"}
```

```
{ "Child_Map":
  {
    "0":
      {"1": [1, 2], "0": [3]}
  },
"Lvl1_Sites: [0],
"Samples": [...],
"Sites": [...] }
```

Figure 2. JSON variant call format introduced in `gramtools`. A graph with nested variation is shown; `gramtools` gives each identified site a number ID. Black nodes contain sequence. Haplogroups, groups of related haplotypes in the graph, are labeled on the edges leaving the first node of Site 0. The red path shows the embedded linear reference genome, and Site 1 and 2 occur on a non-reference sequence background. Top-right text shows part of the top-level of the call format. ”Child_Map” associates a site ID to sites occurring under it: here we record that site 0 contains Site 1 and Site 2 under haplogroup 1, and Site 3 under haplogroup 0. ”Lvl1_Sites” gives site IDs which are not children of any other sites, allowing recursive exploration of the child map. ”Sites” is an array indexed by each site ID: each entry is a JSON containing the same information as a VCF line, shown here above Site 1.

3.2 Genotyping performance

3.2.1 Comparison with reference-based variant callers

We performed an experiment on a genome graph of variation from 2,500 samples in four clinically relevant surface antigens of the malaria parasite *P. falciparum.* Using 14 validation samples with long-read assemblies we show `gramtools` genotype calls outperform variant callers `samtools` and `cortex` run against the reference genome alone. We further show the `gramtools` inferred personalised reference genome allows those tools to discover previously inaccessible variation, and that gramtools finds recombinants between input haplotypes in the graph.

3.2.2 Comparison with state of the art genome graph tools

We built graphs containing 45 distinct deletions between 100 and 13,000 bp found in 17 samples and all variation overlapping the deletions in a further 1,000 samples of *M. tuberculosis*. Using long-read assemblies for the 17 samples we show gramtools is better able to resolve these regions compared to state of the art tools vg [1] and graphtyper2 [4]. Our nesting-aware genotyping process guarantees mutually exclusively calling deletions and the small variants overlapping them.

3.3 Analysis of variation on top of locally defined references

We genotyped 700 *P. falciparum* samples at the surface antigen DBLMSP2 in which two diverged forms are known to segregate, likely due to balancing selection [5]. We show how gramtools recovers the two forms and is able to output variation on top of each diverged form allowing the study of variation on different references.

3.4 Discussion

We have presented a method for identifying, calling and outputting multiscale variation in gramtools. Analogous to the recently proposed rGFA format for describing sequences in genome graphs [6], we provide a format for describing variant calls in genome graphs. We believe such formats are required to better study and express variation in genome graphs.

To be useful, genome graphs should support three concepts: compatibility, consistency and interpretability. Compatibility is maintaining support for linear references. Consistency is outputting a fixed set of variants for a given genome graph. Interpretability is providing a simple way of analysing variation at multiple scales or on different references. In gramtools we propose a framework and format implementing each of these concepts.

Acknowledgments

The authors thank Rachel Colquhoun for the tool used to construct graphs, Sorina Maciuca for algorithms in gramtools and Robyn Ffrancon for software engineering in gramtools.

Brice Letcher is a predoctorate fellow at EMBL-EBI, funded by EMBL and registered at the University of Cambridge.

References

[1] Erik Garrison, Jouni Sirén, Adam M Novak, Glenn Hickey, Jordan M Eizenga, Eric T Dawson, William Jones, Shilpa Garg, Charles Markello, Michael F Lin, Benedict Paten, and Richard Durbin. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, 36(9):875–879, August 2018.

[2] Jonas Andreas Sibbesen, Lasse Maretty, and Anders Krogh. Accurate genotyping across variant classes and lengths using variant graphs. *Nature Genetics*, 50(7):1054, July 2018.

[3] Sorina Maciuca, Carlos del Ojo Elias, Gil McVean, and Zamin Iqbal. A natural encoding of genetic variation in a Burrows-Wheeler Transform to enable mapping and genome inference. In Springer, editor, *Proceedings of the 16th International Workshop on Algorithms in Bioinformatics, Volume 9838 of Lecture Notes in Computer Science*, pages 222–233, 2016.

[4] Hannes P. Eggertsson, Snaedis Kristmundsdottir, Doruk Beyter, Hakon Jonsson, Astros Skuladottir, Marteinn T. Hardarson, Daniel F. Gudbjartsson, Kari Stefansson, Bjarni V. Halldorsson, and Pall Melsted. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nature Communications*, 10(1):5402, November 2019. Number: 1 Publisher: Nature Publishing Group.

[5] Alfred Amambua-Ngwa, Kevin K. A. Tetteh, Magnus Manske, Natalia Gomez-Escobar, Lindsay B. Stewart, M. Elizabeth Deerhake, Ian H. Cheeseman, Christopher I. Newbold, Anthony A. Holder, Ellen Knuepfer, Omar Janha, Muminatou Jallow, Susana Campino, Bronwyn MacInnis, Dominic P. Kwiatkowski, and David J. Conway. Population Genomic Scan for Candidate Signatures of Balancing Selection to Guide Antigen Characterization in Malaria Parasites. *PLOS Genetics*, 8(11):e1002992, November 2012.

[6] Heng Li, Xiaowen Feng, and Chong Chu. The design and construction of reference pangenome graphs. *arXiv:2003.06079 [q-bio]*, March 2020. arXiv: 2003.06079.