Strain-Aware Assembly of Genomes from Mixed Samples Using Flow Variation Graphs

Jasmijn A. Baaijens[1,2], Leen Stougie[1,3,4], and Alexander Schönhuth[1,4,5(✉)]

[1] Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
alexander.schoenhuth@cwi.nl
[2] Harvard Medical School, Boston, MA, USA
[3] Vrije Universiteit, Amsterdam, The Netherlands
[4] INRIA-Erable, Lyon, France
[5] Utrecht University, Utrecht, The Netherlands

Extended Abstract

Introduction. The goal of strain-aware genome assembly is to reconstruct all individual haplotypes from a mixed sample at the strain level and to provide abundance estimates for the strains. Given that the use of a reference genome can introduce significant biases, de novo approaches are most suitable for this task. So far, reference-genome-independent assemblers have been shown to reconstruct haplotypes for mixed samples of limited complexity and genomes not exceeding 10000 bp in length. This renders such approaches applicable to viral quasispecies, but one cannot use them for bacterial sized genomes. In experiments presented here, we notice that even reference-dependent approaches tend to struggle with bacterial sized genomes.

We present VG-Flow, a de novo approach that enables full-length haplotype reconstruction from pre-assembled contigs of complex mixed samples. Our method increases contiguity of the input assembly and, at the same time, it performs haplotype abundance estimation. VG-Flow is the first approach to require polynomial, and not exponential runtime in terms of the underlying graphs. Since runtime increases only linearly in the length of the genomes in practice, it enables the reconstruction also of genomes that are longer by orders of magnitude, thereby establishing the first de novo solution to strain-aware full-length genome assembly applicable to bacterial sized genomes.

Methods. The methodical novelty that underlies VG-Flow's advances is to derive *flow variation graphs* from the (common) variation graphs that one constructs from the input contigs. General variation graphs [2,3] derived from input contigs had been presented in earlier work as a means for overcoming linear reference induced biases and aiming at the reconstruction of full-length strain-level haplotypes [1]. We introduce the concept of a flow variation graph and cast the relevant computational problem in terms of this graph, which renders the problem polynomial-time solvable for the first time.

R. Schwartz (Ed.): RECOMB 2020, LNBI 12074, pp. 221–222, 2020.
https://doi.org/10.1007/978-3-030-45257-5_14

Our approach consists of five steps, depicted in Fig. 1. As input it takes a data set of next-generation sequencing reads and a collection of strain-specific contigs assembled from the data. The final output is presented as a *genome variation graph* capturing all haplotypes present, along with the estimated relative abundances. While the already efficient or practially feasible steps (1) and (5) correspond to prior work [1], steps (2), (3) and (4) are novel, and replace the exponential-runtime procedure presented earlier.

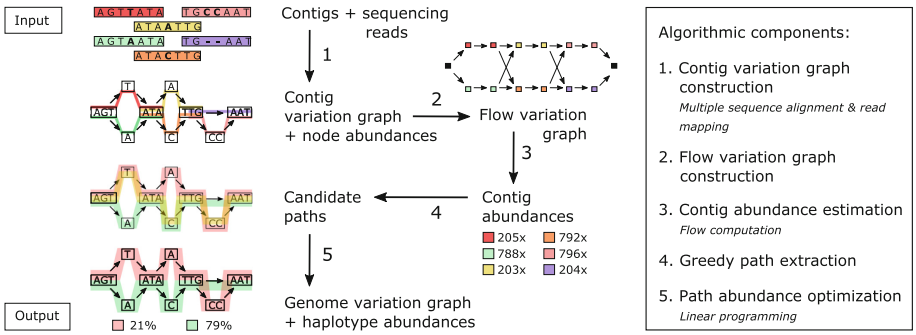

Fig. 1. Algorithm overview

Results. We demonstrate that VG-flow scales approximately linearly in genome size in practice, which allows to process mixtures of genomes that are longer on orders of magnitude. In this, VG-Flow presents the first comprehensive solution to assembling haplotypes from mixed samples at the strain level, also for small bacterial genomes and samples of considerably increased complexity. Benchmarking experiments show that our method outperforms state-of-the-art approaches on mixed samples from viral genomes in terms of assembly accuracy as well as abundance estimation. Experiments on longer, bacterial sized genomes demonstrate that VG-Flow is the only current approach that can reconstruct full-length haplotypes from mixed samples at the strain level in human-affordable runtime.

A full version of this paper is available at https://doi.org/10.1101/645721 and the software can be downloaded from https://bitbucket.org/jbaaijens/vg-flow.

References

1. Baaijens, J., Van der Roest, B., Köster, J., Stougie, L., Schönhuth, A.: Full-length de novo viral quasispecies assembly through variation graph construction. Bioinformatics **35**(24), 5086–5094 (2019). https://doi.org/10.1093/bioinformatics/btz443
2. Garrison, E., et al.: Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat. Biotechnol. **36**, 875–879 (2018). https://doi.org/10.1038/nbt.4227
3. Paten, B., Novak, A., Eizenga, J., Garrison, E.: Genome graphs and the evolution of genome inference. Genome Res. **27**(5), 665–676 (2017). https://doi.org/10.1101/gr.214155.116