Critique: "De novo assembly and genotyping of variants using colored de Bruijn graphs"

Zamin Iqbal, Mario Caccamo, Isaac Turner, Paul Flicek, Gil McVean

1. Method outline

The paper presents a novel method of variant detection based on de novo assembly of sample genomes. It extends well-known de Bruijn graph framework of genome assembly by introducing additional labeling: edges that are induced from a specific genome are colored with the color corresponding the genome. The method is implemented in tool called Cortex.

The key observation that lies within their approach is that variants in the graph generate specific structures known as bulges. By constructing the graph from a set of reads and adding reference genome to the graph, the method finds these structures in the graph.

Variations are found by two algorithms: Bubble Caller (BC) and Path Divergence Caller (PD). The Bubble Caller algorithm locates clean bulges in the graph, i.e. bulges that are formed by two branches that do not interfere with any other part of the graph. The Path Divergence Caller is more sophisticated, it allows the branch that belongs to the reference to interfere with the other parts of the graph.

Since bulges in the graph could be induced not only by variations, but by sequencing errors and intergenomic repeats, the authors developed a probabilistic approach of classifying graph structures.

2. Critique

The paper is clearly written and the supplementary note presents detailed description of the method developed by the authors. Source code as well as documentation is freely available on the project's sourceforge page, it is open-source and distributed under GNU GPL v.3 license. The documentation is very detailed.

The introduction presents rational motivation of the work – it is clear that the traditional method of variant calling (mapping of short reads to a reference) has it's limitations. Additionally, authors criticize currently existing assemblers for being "variation-unaware" since they tend to produce "consensus" sequences in case of ambiguity and thus lose information. However, this statement is not supported by any quantative evidence — it would be very interesting to know how much real information that spoils variant detection current state-of-the-art assemblers lose comparing to the "careful" assembler developed by authors. It is worth to note Cortex performs graph cleaning operations and thus may lose some information by itself.

The authors assess capabilities of Cortex on both simulated and real data. Runs on the simulated data showed quite high power of detecting short ($1 - 100$ BP) and moderately sized variants ($100 - 1000$ BP), from 90% to 50%. However, for large indels Cortex was unable to detect only homozygous sites with power 35%. It shows limited ability of Cortex to detect large indels.

For the real data, they took an individual from 1000 Genomes Project and compared Cortex variant calling capabilities with traditional mapping methods used in 1K project. Cortex detected 87% of homozygous variant sites and 67% of heterozygous sites. However, it was able to detect larger structural variants, indels of size higher than 20 BP that cannot be detected using traditional methods.

The authors also note that their approach has some limitations. For example, Cortex doesn't use any correction tools and doesn't utilize paired information.

3. Conclusion

The paper presents a novel approach for variant calling. It is based on de novo assembly of raw reads by using colored de Bruijn graphs framework. Cortex has high power of detecting SNVs, short and moderate size variants. The method was shown to be able to detect short indels that are unlikely to be detected by traditional mapping methods. It also doesn't require a complete reference and can call variants using genomes represented as a set of raw reads which is a great advantage over conventional approaches.