A graph method for population genotyping of structural variants

P Krusche[1], E Dolzhenko[2], S Chen[2], N Johnson[2], MA Bekritsky[1], A Gross[2], BR Lajoie[2], V Rajan[2], Z Kingsbury[1], SJ Humphray[1], SS Ajay[2], RJ Taft[2], DR Bentley[1], MA Eberle[2]

[1]Illumina Cambridge Ltd, Saffron Walden, Essex, CB10 1XL, UK [2]Illumina Inc, 5200 Illumina Way, San Diego, CA 92122, USA

Contact Peter Krusche: pkrusche@illumina.com and Michael A. Eberle: meberle@illumina.com

Introduction

Accurate **detection and annotation of structural variants (SVs)** is a critical component of clinical variant calling pipelines. Here we present a method for **genotyping deletions, insertions and substitutions jointly across large sample cohorts**.

A significant challenge with genotyping SVs across multiple samples is that the same variant may have different representations in different samples. Nearby variants and sequence instability around the breakpoints of the SVs are some of the causes of these issues. As a result, SVs can be difficult to aggregate across samples which leads to underestimation of variant frequencies and other errors.

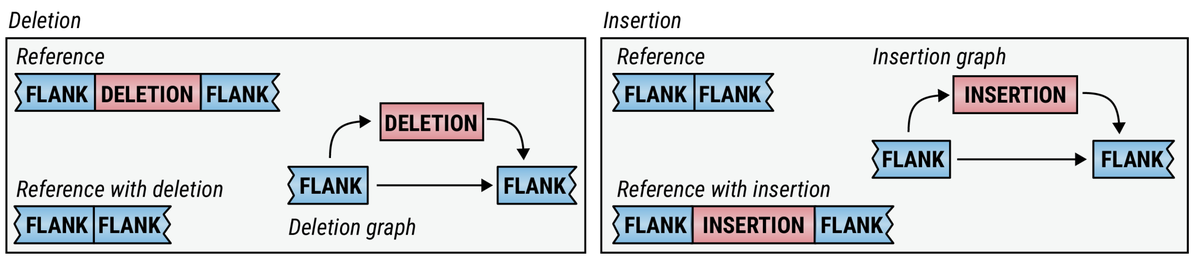
We have developed a joint genotyping method for SVs based on re-aligning sequence reads to breakpoint graphs. Using this method, we are able to evaluate breakpoints uniformly across many samples and genotype SVs jointly in a population.

To demonstrate the utility of our method we show that it **achieves superior genotyping performance compared to the leading single-sample variant caller** [1] by testing haplotype concordance across the Platinum Genomes pedigree and cohort-level Hardy-Weinberg consistency.

Graph-based Realignment

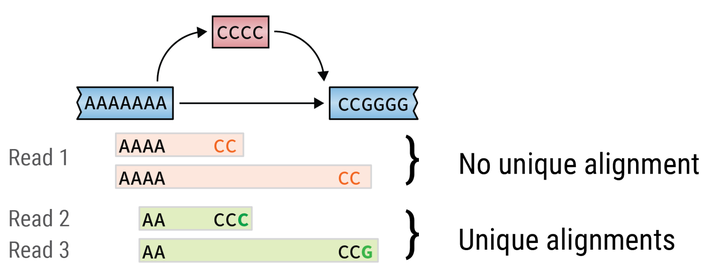
Template Graph Construction

We construct a template sequence graph for each event by adding nodes for all distinct allele sequences and then connecting the nodes with edges according to REF and ALT paths.



Read Recovery

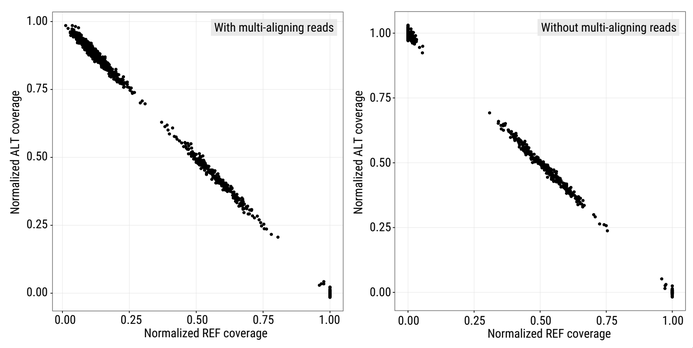
We extract reads from each BAM file that map to target regions around the breakpoints. We also extract their mates if they map outside of the target regions.



Read Alignment

We use an exact matching heuristic and gssw [3] to align reads to a template graph for each sample.

We analyze the dynamic programming matrix to detect multi-mapping reads.



Variant Genotyping

Candidate Variants

- We used published long insertions (233 based on short + long reads [2]; 483 based on short read assemblies [3])
- >100k reconciled Manta cohort and Platinum Genomes calls

Validation Datasets

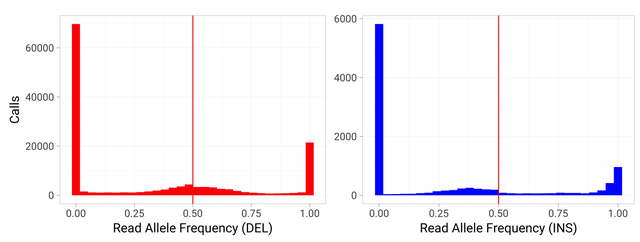
- 17 Platinum Genomes (PG) samples (HiSeq 2000 2x100 WGS)
- 220 samples from Polaris (HiSeqX 2x150 WGS)
- Variants validate when they pass PG inheritance or have AF > 0.05 in Polaris and their genotypes in HWE (p-value > 0.05) [4]

Genotyping Model

We assume that read counts supporting each allele are Poisson-distributed and infer model parameters using EM.

Although the read allele frequency becomes skewed when not all ALT reads can be retrieved, the events can still be genotyped.

$$freq = \frac{\#ALT}{\#ALT + \#REF}$$

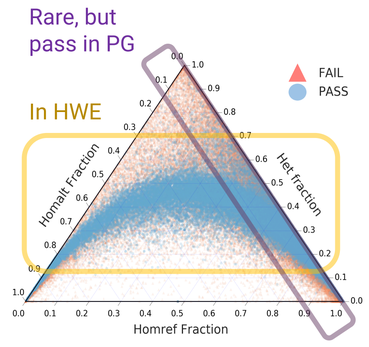


Genotyping Results

Validating Candidate Variants

Candidate list of **185,395 indels with length > 10**
- ➜ 183,098 bi-allelic in >1 Platinum Genomes or Polaris samples.
- ➜ 37,878 pass in PG ; 100,154 monomorphic in PG
- ➜ **70,009 variant calls were validated.**

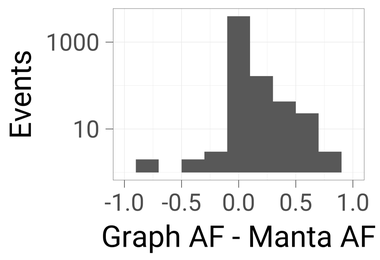
Variant calls on Polaris and PG are available at https://github.com/Illumina/Polaris



Comparing Graph-based and Manta Genotypes in the Polaris Cohort

- The comparison is based on 15,325 candidate deletions longer than 50bp (which is the min. recommended size for variants Manta can genotype).
- On Polaris, we see 9716 in one or both callers with minimum AF >= 0.1; **Graph method calls 98%** of these with >5% AF, while Manta calls only 80%.

	Graph no HWE	Graph in HWE
Manta no HWE	8102	**2104**
Manta in HWE	**922**	4182



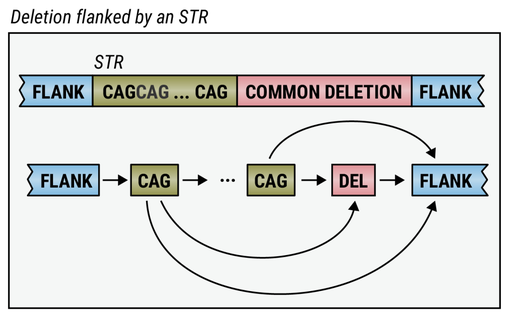
Conclusion

Population-based joint genotyping can better distinguish between heterozygous and homozygous variants because it produces more variant calls that are in HWE.

Ongoing Work

STR Breakpoint Graphs

We can model repeats with graphs by
- representing repeat sequence in a graph as loops around a repeat unit node,
- unrolling the loops into multiple repeat-unit nodes to enable DAG alignment,
- genotyping the repeat using these alignments as described previously [6].



Using the STR graphs **we can genotype both the STR and the deletion**!

Resources and Links

Platinum Genomes
https://github.com/illumina/platinumgenomes

The Polaris Study
https://github.com/illumina/polaris

Join us!
illumina.wd1.myworkdayjobs.com/illumina-careers

References

1. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32,** 1220–1222 (2016).
2. English, A. C. *et al.* Assessing structural variation in a personal genome-towards a human reference diploid genome. *BMC Genomics* **16,** 286 (2015).
3. Kehr, B. *et al.* Diversity in non-repetitive human sequences not found in the reference genome. *Nat Genet* **49,** 588–593 (2017).
4. Hardy-Weinberg Principle on Wikipedia: https://en.wikipedia.org/wiki/Hardy%E2%80%93Weinberg_principle
5. Graph-striped Smith Waterman implementation: https://github.com/vgteam/gssw
6. Dolzhenko, E. *et al.* Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Research* (2017). doi:10.1101/gr.225672.117