# A graph-based pipeline to evaluate common structural variations (SVs) based on haplotypes and reassembly

Sun-Gou Ji, Lizao Li, Gülfem Demir, Péter Kómár and Deniz Kural

Seven Bridges Inc., 1 Main St, Cambridge MA, USA

# CHECK OUT OUR OTHER POSTER @ PgmNr 1263/F (Kallberg et al.)

# I. Background

# 

#### SV Breakpoints

- Not consistent across studies
  - Due to errors in current technology (biochemical, computational, or both)
  - Multiple SV events cluster together in complex regions of the genome (microhomology/repeats/etc.)
  - Overlapping SVs in multiple individuals can be a consequence of convergent evolution caused by different evolutionary events
- Goal: Identify SVs common in the population with consistent breakpoints (evolutionarily conserved SVs fixed in the population)
- → Filter for SVs in HWE that also have well-defined haplotypes

# II. Methods

all individuals

#### - 1. Conserved SVs from 1000 Genomes

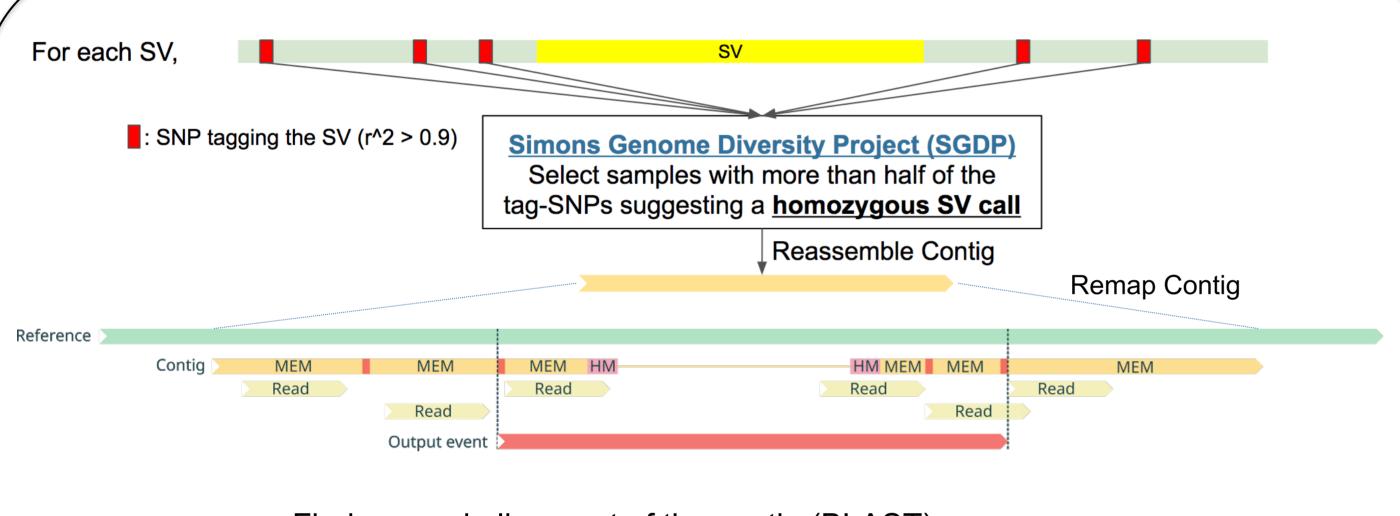
- 1000 Genomes SV (Sudmant *et al*, 2015)
  HWE P > 0.05 (bonferroni-corrected) in all five super population
  At least one tag-SNP (r² > 0.9) across
  - DG\
    - > 2 technologies
      > 5% frequency
      > 250 unique samples

> 2 studies

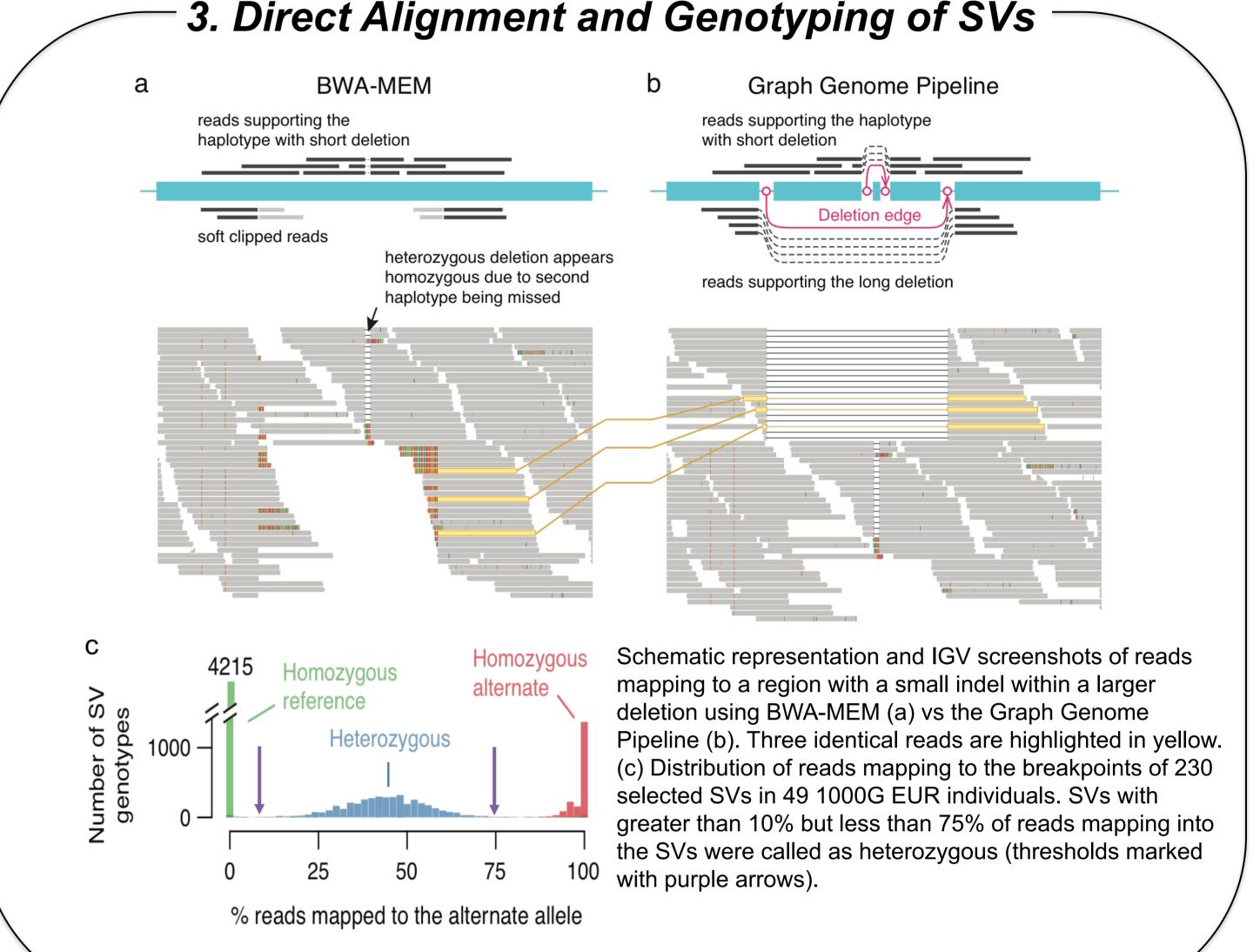
> 2 algorithms

> 80% of selected SVs were present in at least one high-confidence human genome (CHM1, CHM13, AK1, Huref)

### 2. Reassembly and Remapping of Alternate Contig



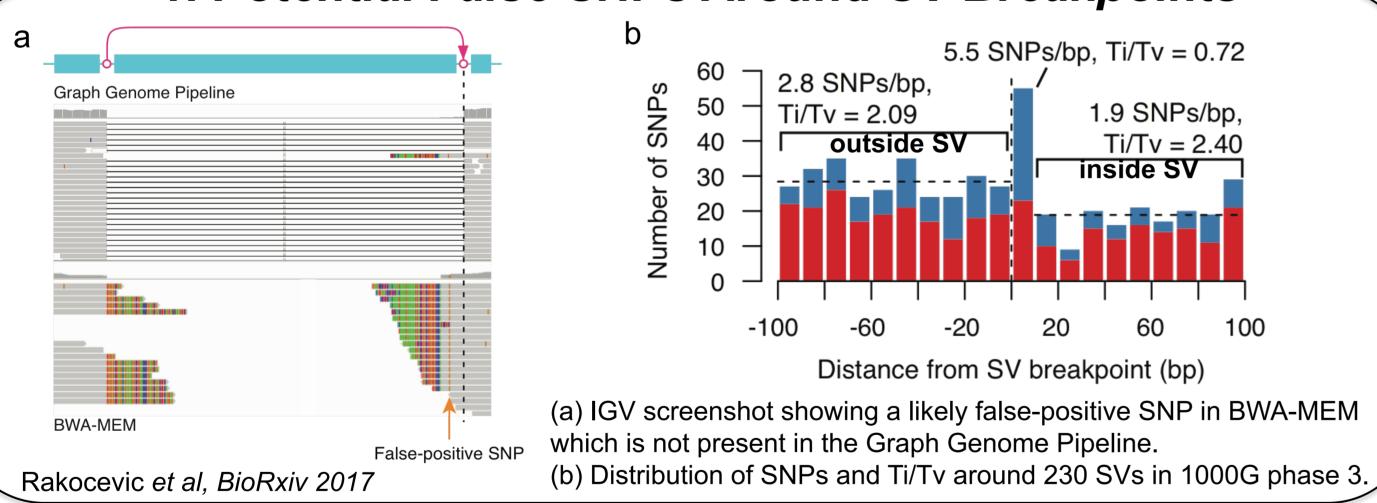
- Find gapped alignment of the contig (BLAST)
- Identify maximal exact matches (MEM) longer than read length
- → A total of 1,118 SVs included in the pan-genome graph



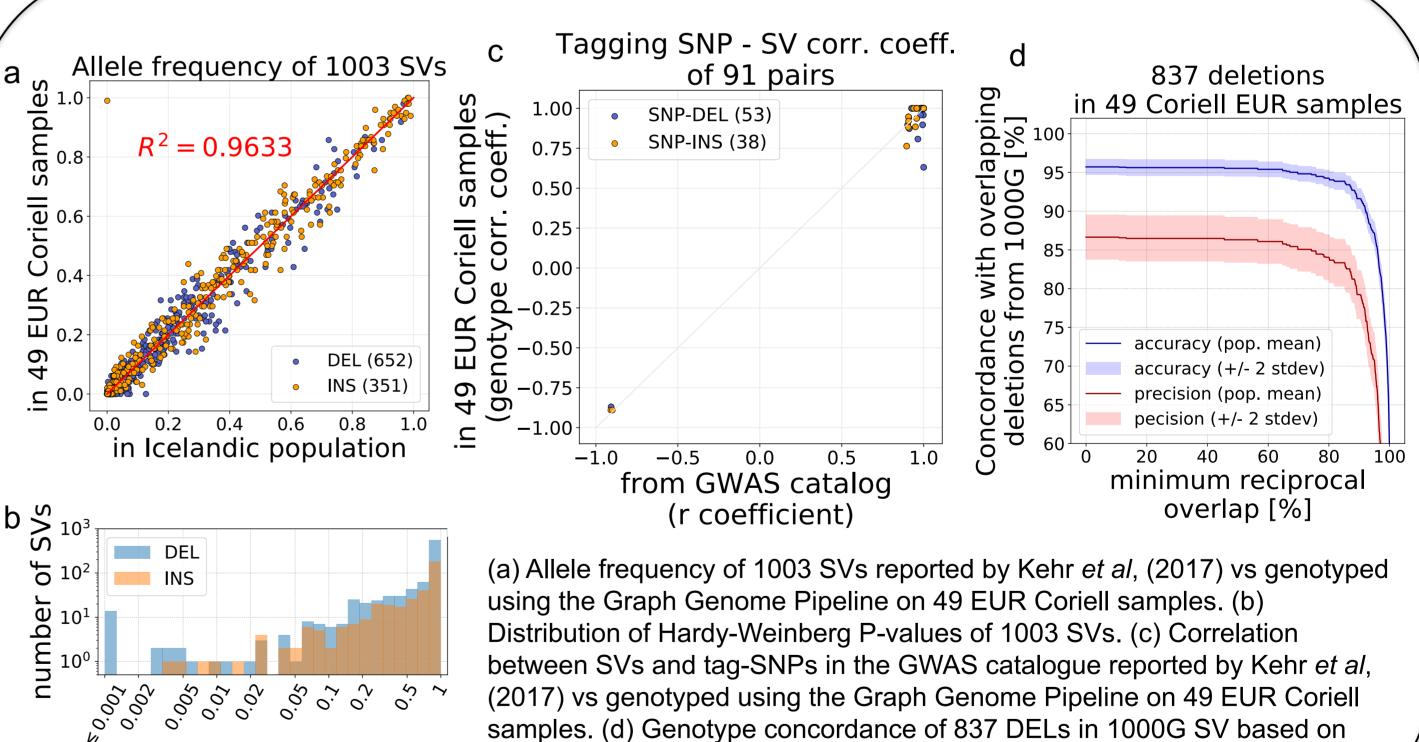
Rakocevic et al, BioRxiv 2017

# III. Results

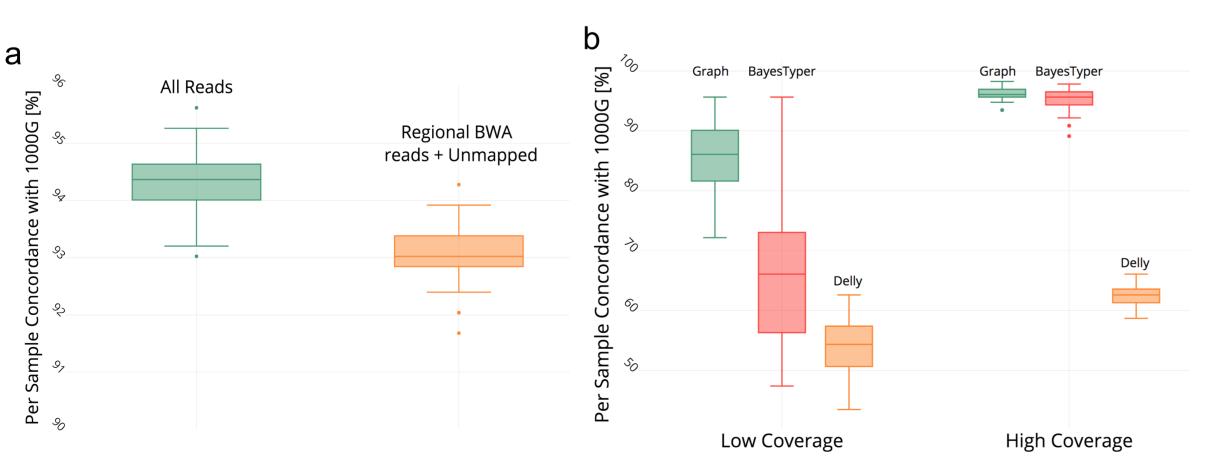
# - 1. Potential False SNPs Around SV Breakpoints



# 3. Genotyping of deCODE SVs in 1000G EUR



## 2. SV Genotype Concordance with 1000 Genomes SV -



(a) Distribution of per-sample concordance of 1118 SVs using the Graph Genome Pipeline on all raw reads vs re-alignment and genotyping of BWA-aligned reads from SV regions and unmapped reads. (b) Comparison of Graph Genome Pipeline vs BayesTyper (https://github.com/bioinformatics-centre/Bayes Typer) vs Delly (Rausch *et al*, 2012) on low coverage (5-7X) 1000 Genomes phase 3 raw reads and high coverage (30X) 49 EUR individuals sequenced as part of the Coriell cohort.

# 4. SV Confidence Based on Genotype Concordance

minimum reciprocal overlap.

#### - Pre-alignment features

HWE exact test p-value

- Population evidence (DGV, 1000G, GoNL, Icelandic, etc.)
- Personal genome evidence (CHM1, HuRef, 1000G trio, etc.)
  Consistency of breakpoints in assembled contigs

#### Graph-alignment features

- Microhomology length/content around flanking regions
- K-mer content of new pathsSimulation of reads and alignment

SV genotype confidence score

# IV. Discussion

- We can directly align reads to an SV-augmented pan-genome graph using the SBG graph pipeline
- No need for realignment
- Straightforward genotyping of SVs
- Rescue of potential FP SNPs/indels near SVs
- Current limitations
- Limited to common SVs with consistent breakpoints in regions of high complexity
- SVs with true wobbly breakpoints will lead to FP SNPs
- General information regarding the SBG graph pipeline can be found @ poster PgmNr 1263/F (Kallberg et al.)