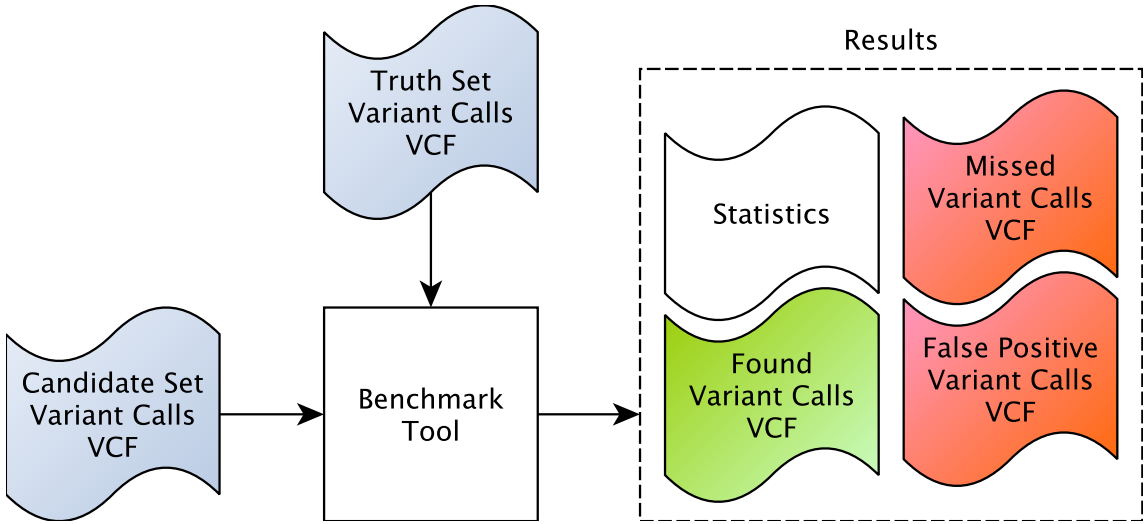
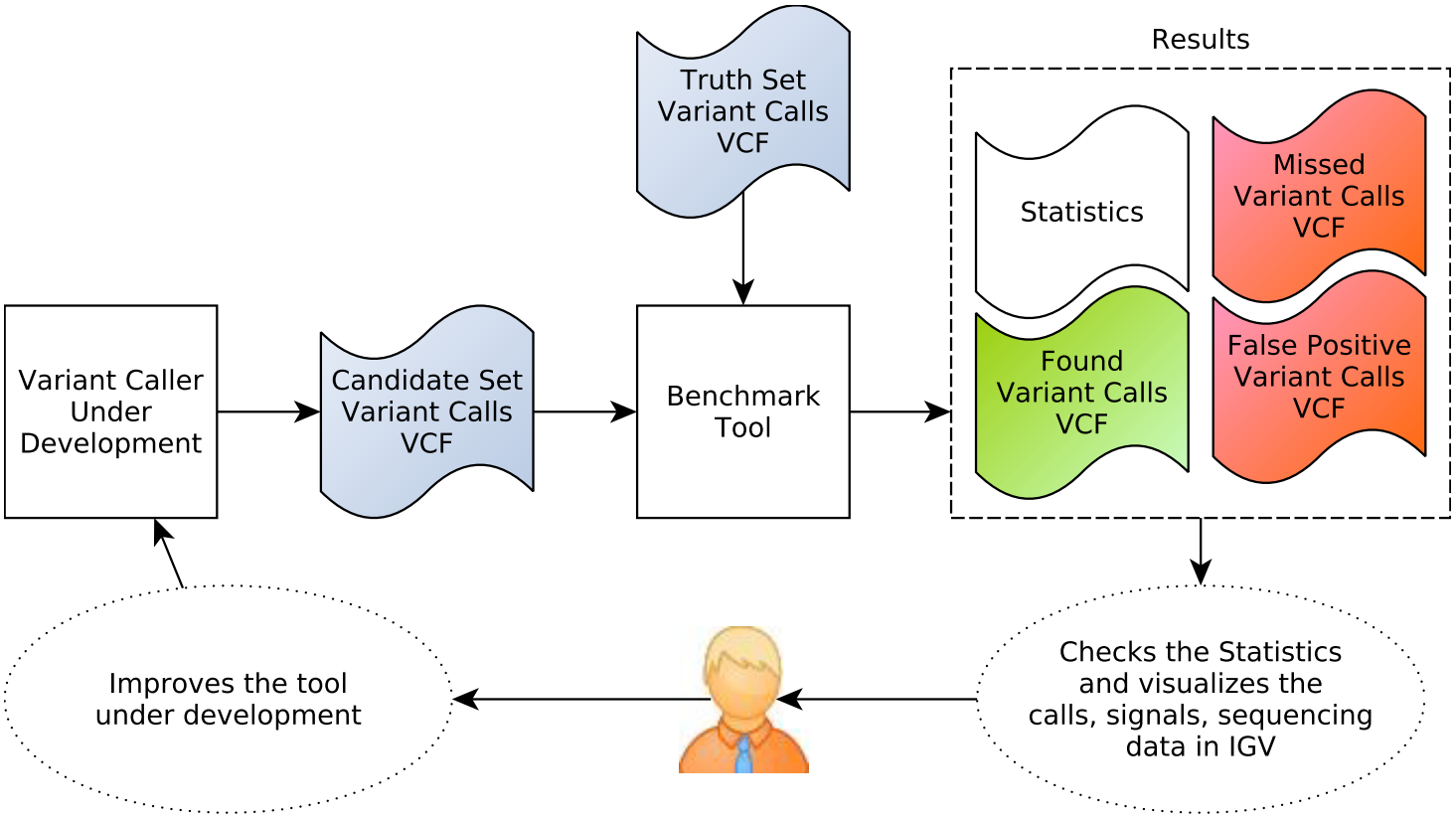
Benchmarking Variant Callers

- Compare variant calls to a reference truth set
- Split results for individual analysis



Benchmarking Variant Callers



RESULTS

And comparison to the state of the art

Evaluation

Kosugi *et al. Genome Biology* (2019) 20:117
https://doi.org/10.1186/s13059-019-1720-5

Genome Biology

- Datasets and results for other software from :

RESEARCH Open Access

Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing

Shunichi Kosugi[1,2], Yukihide Momozawa[3], Xiaoxi Liu[3], Chikashi Terao[1,2], Michiaki Kubo[4] and Yoichiro Kamatani[1,2*]

- Evaluated results from 69 software
  - Simulated Datasets
  - Real Datasets

Deletion Caller Results

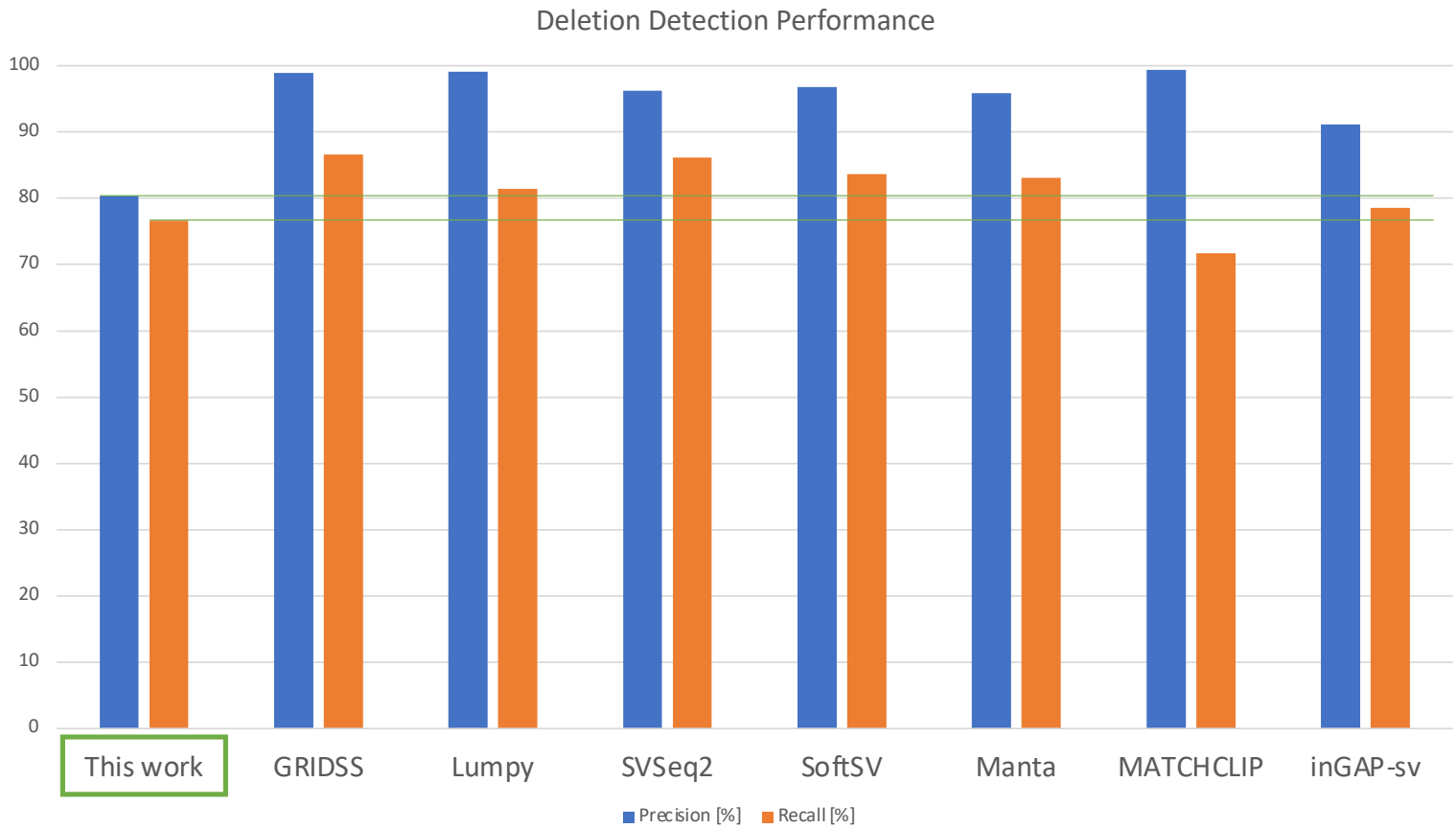
Deletion Caller – Simulated Data

Precision :

$$\frac{\#\ Deletions\ found}{\#\ Deletions\ predicted}$$

Sensitivity (Recall) :

$$\frac{\#\ Deletions\ found}{\#\ Existing\ deletions}$$



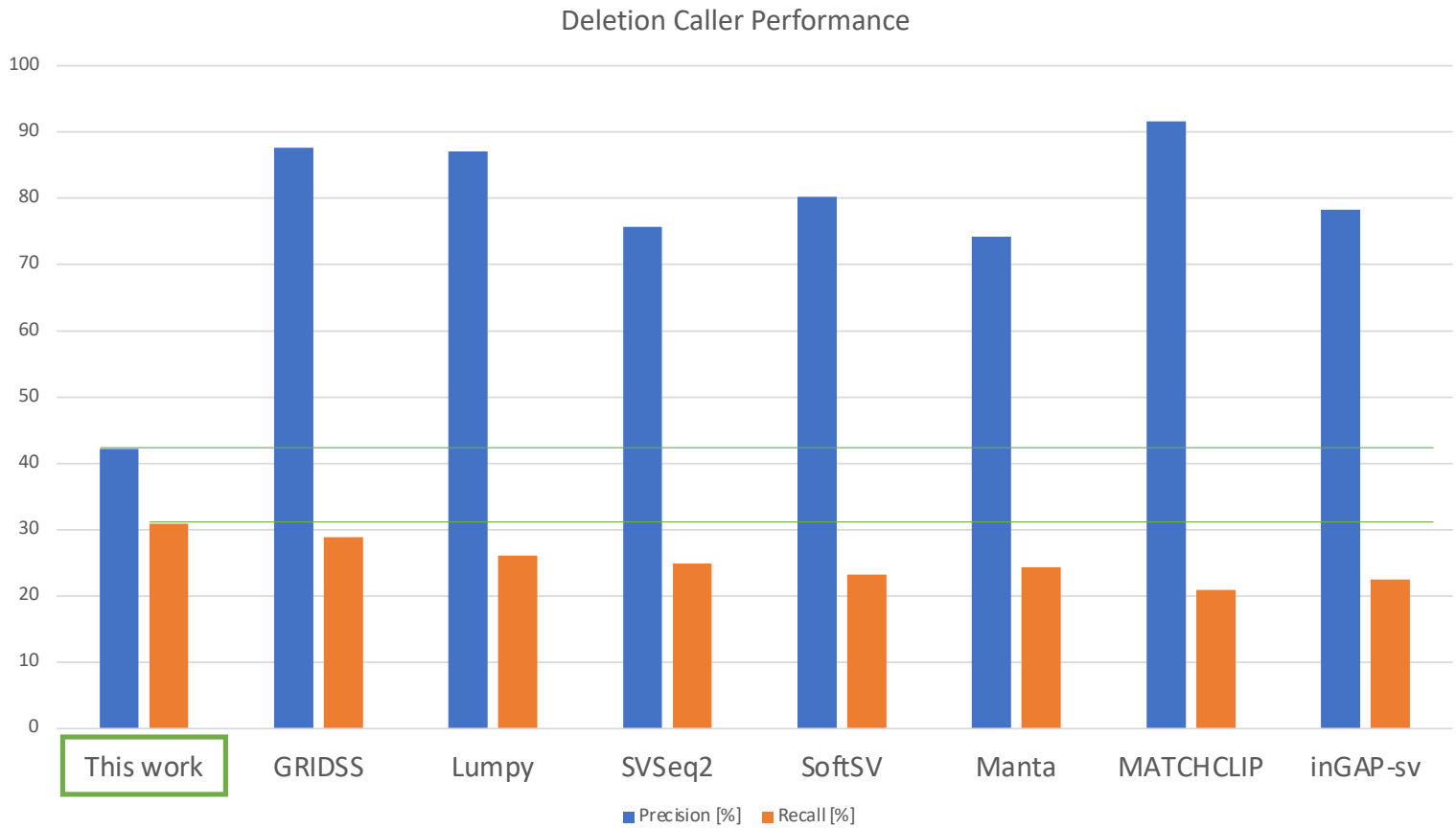
Deletion Caller – Real Data – HG001 (NA12878)

Precision :

$$\frac{\#\ Deletions\ found}{\#\ Deletions\ predicted}$$

Sensitivity (Recall) :

$$\frac{\#\ Deletions\ found}{\#\ Existing\ deletions}$$



Note : new results evaluated on HG001 (NA12878) with DGV known variants + long read calls. In report only HG002 with GIAB data set, (46.4%, 29.2%)

Deletion Caller – Real Data – HG001 (NA12878)

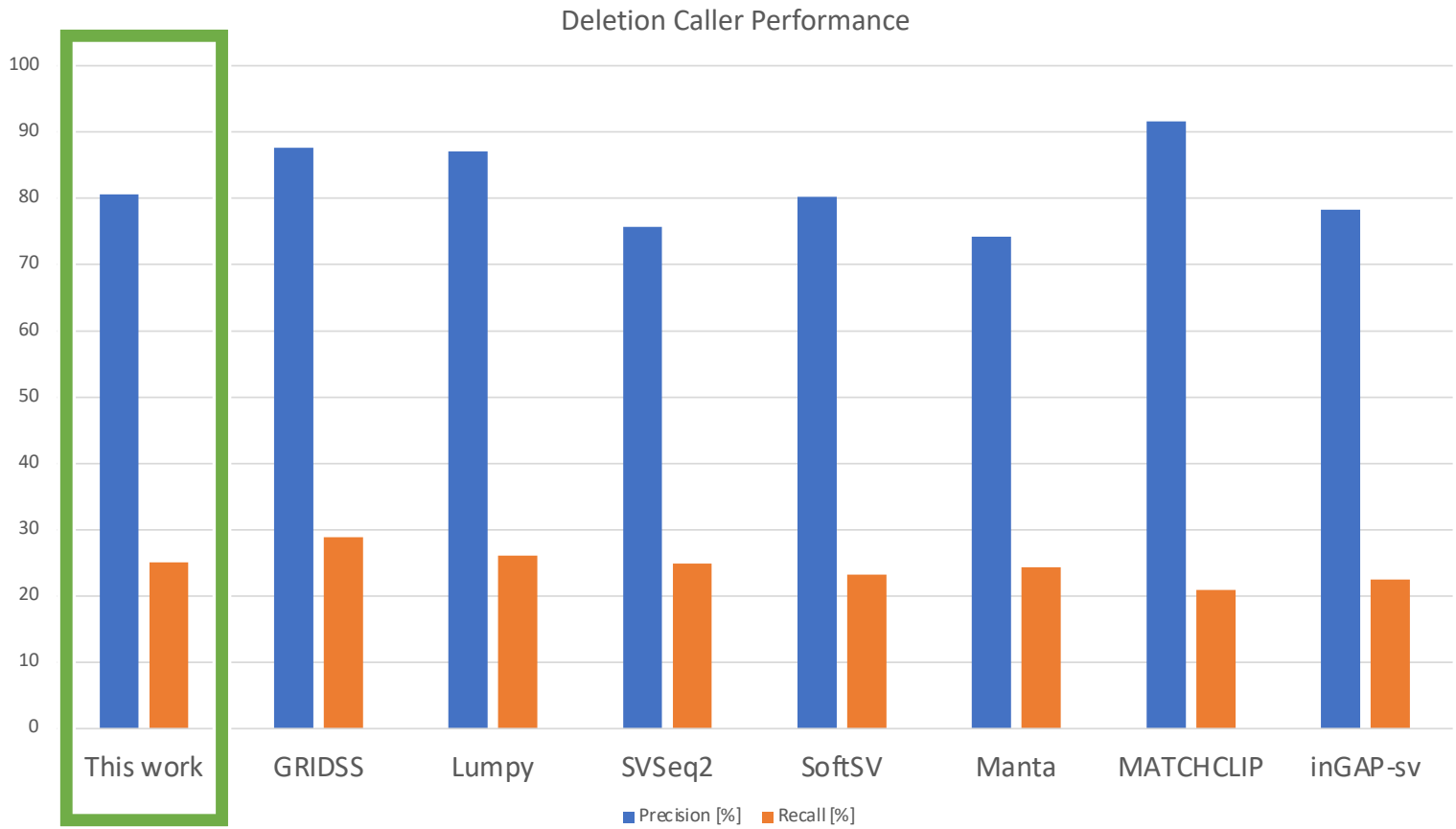
Precision :

$$\frac{\#\ Deletions\ found}{\#\ Deletions\ predicted}$$

Sensitivity (Recall) :

$$\frac{\#\ Deletions\ found}{\#\ Existing\ deletions}$$

Threshold for number of pairs with reads mapped too far away changed from 10% to 20% of coverage !



Duplication Caller Results

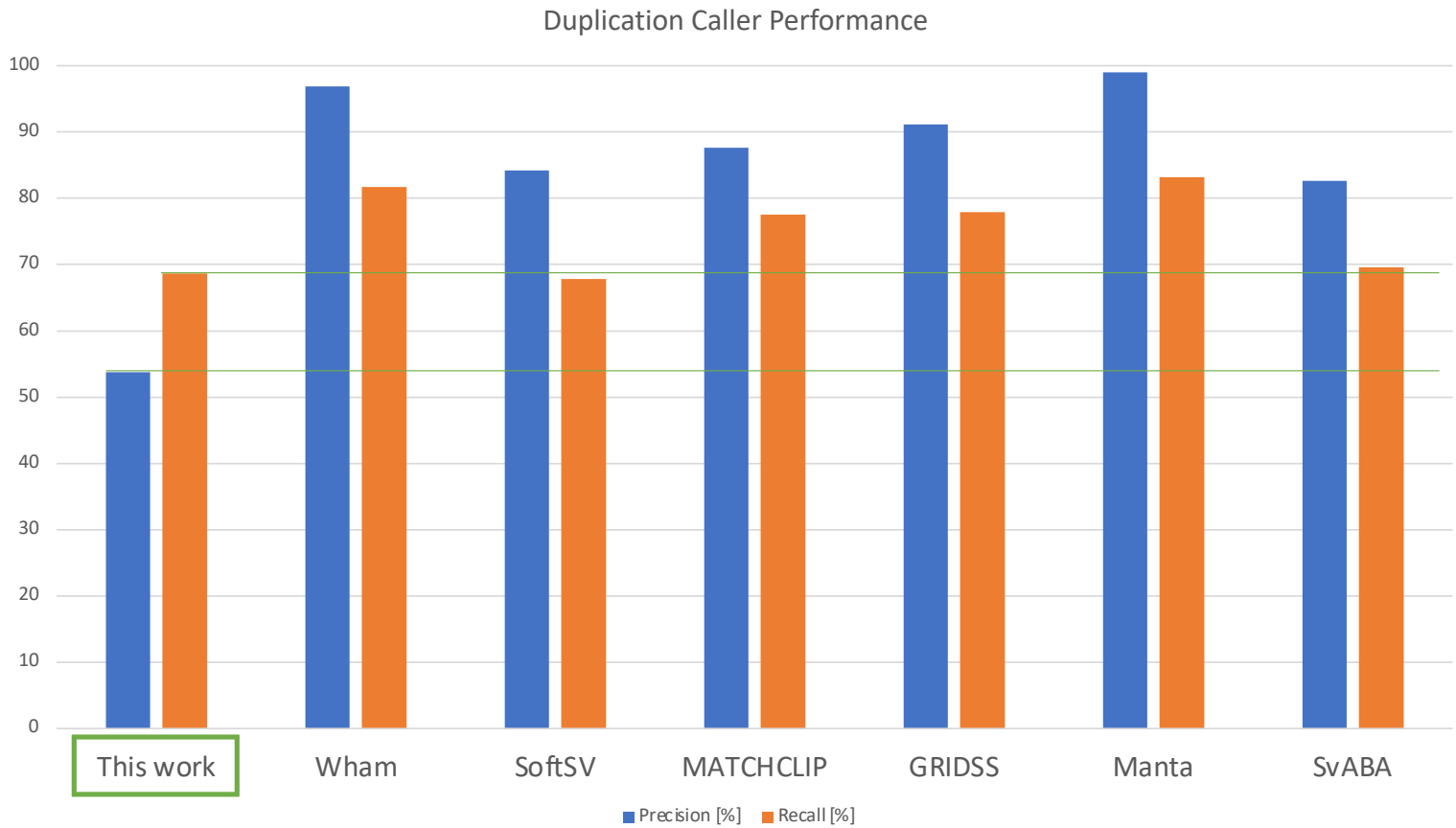
Duplication Caller – Simulated Data

Precision :

$$\frac{\#\ Duplications\ found}{\#\ Duplications\ predicted}$$

Sensitivity (Recall) :

$$\frac{\#\ Duplications\ found}{\#\ Existing\ duplications}$$



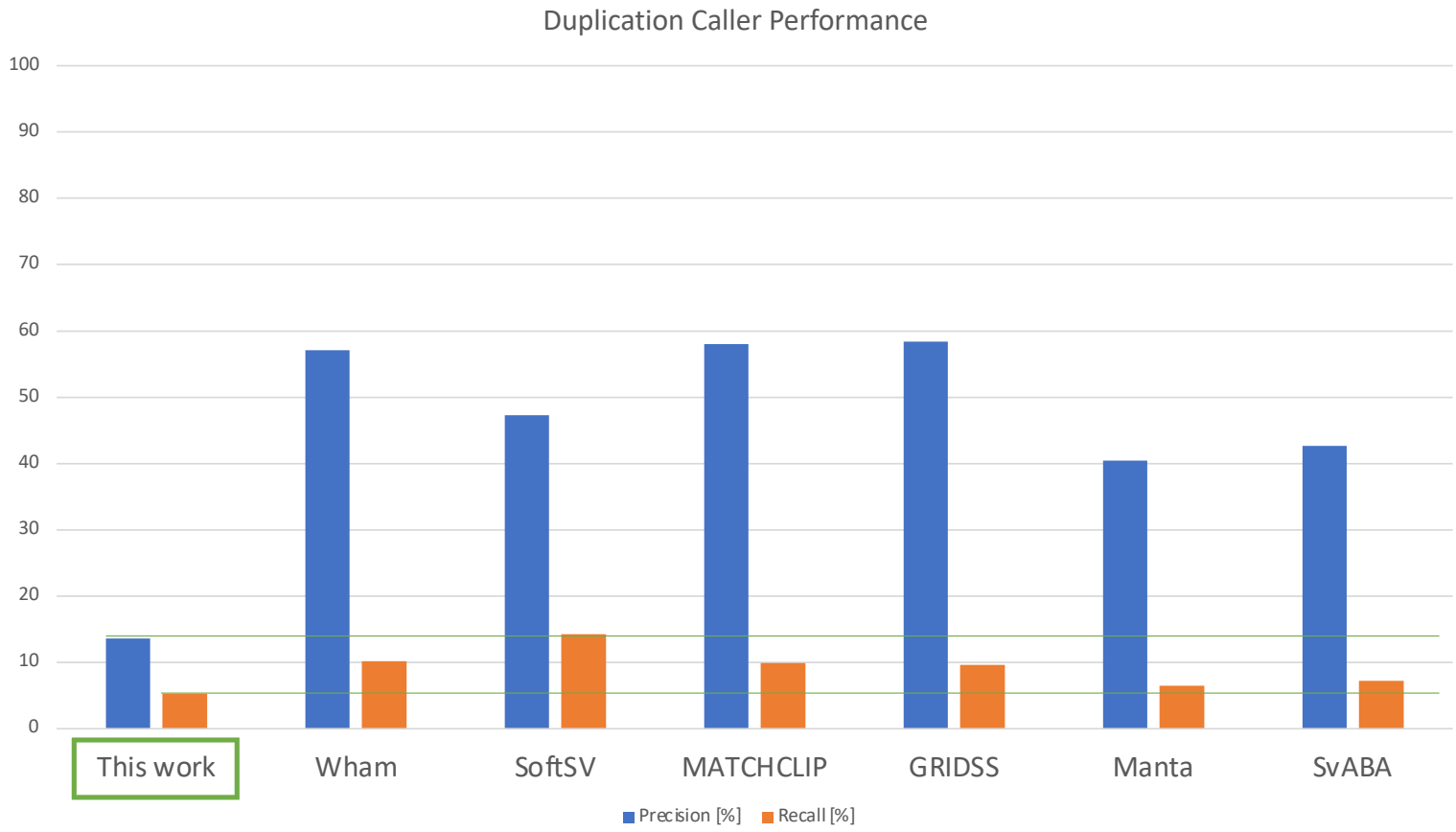
Duplication Caller – Real Data (NA12878)

Precision :

$$\frac{\#\ Duplications\ found}{\#\ Duplications\ predicted}$$

Sensitivity (Recall) :

$$\frac{\#\ Duplications\ found}{\#\ Existing\ duplications}$$



Inversion Caller Results

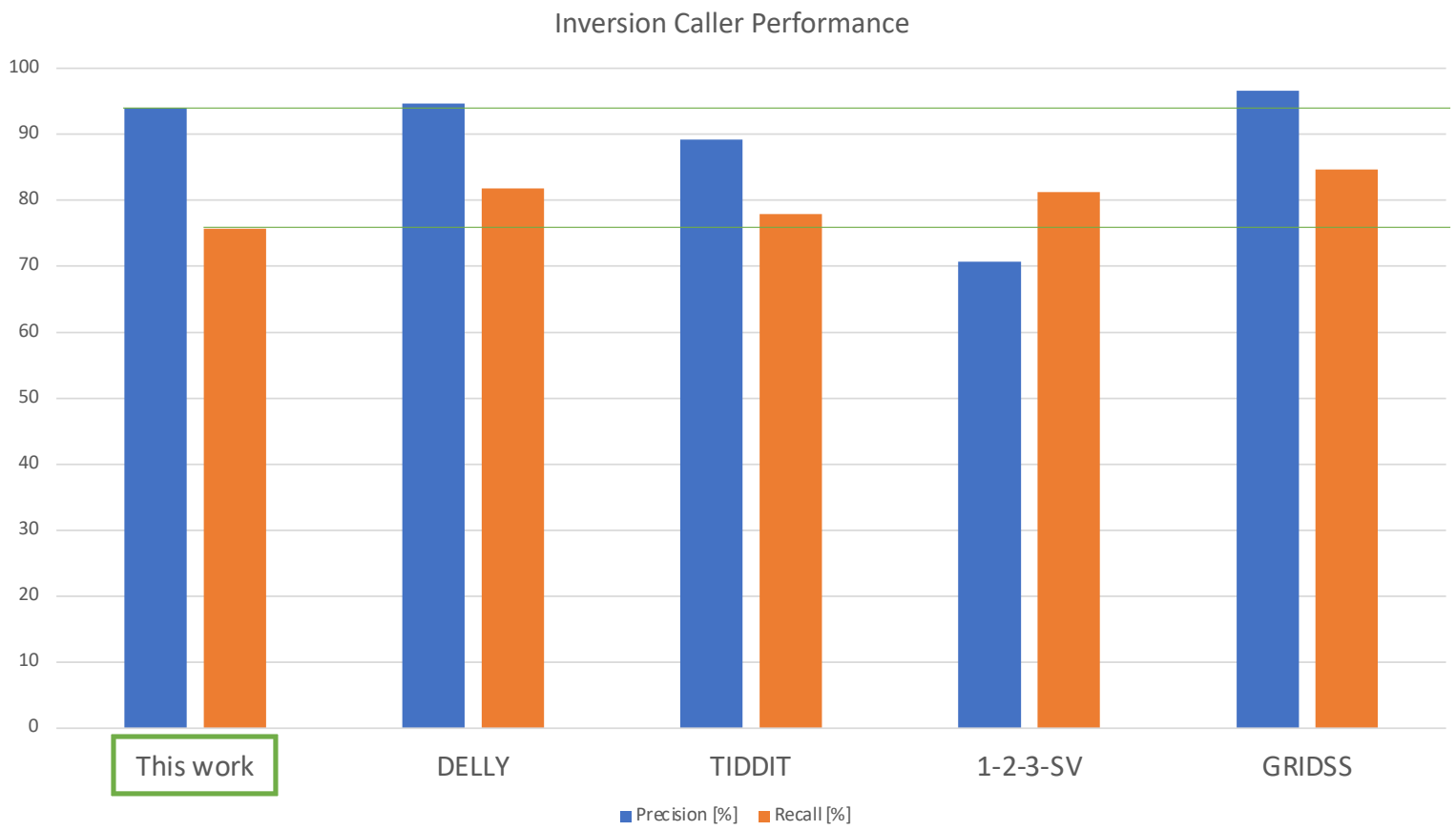
Inversion Caller – Simulated Data

Precision :

$$\frac{\#\ Inversions\ found}{\#\ Inversions\ predicted}$$

Sensitivity (Recall) :

$$\frac{\#\ Inversions\ found}{\#\ Existing\ Inversions}$$



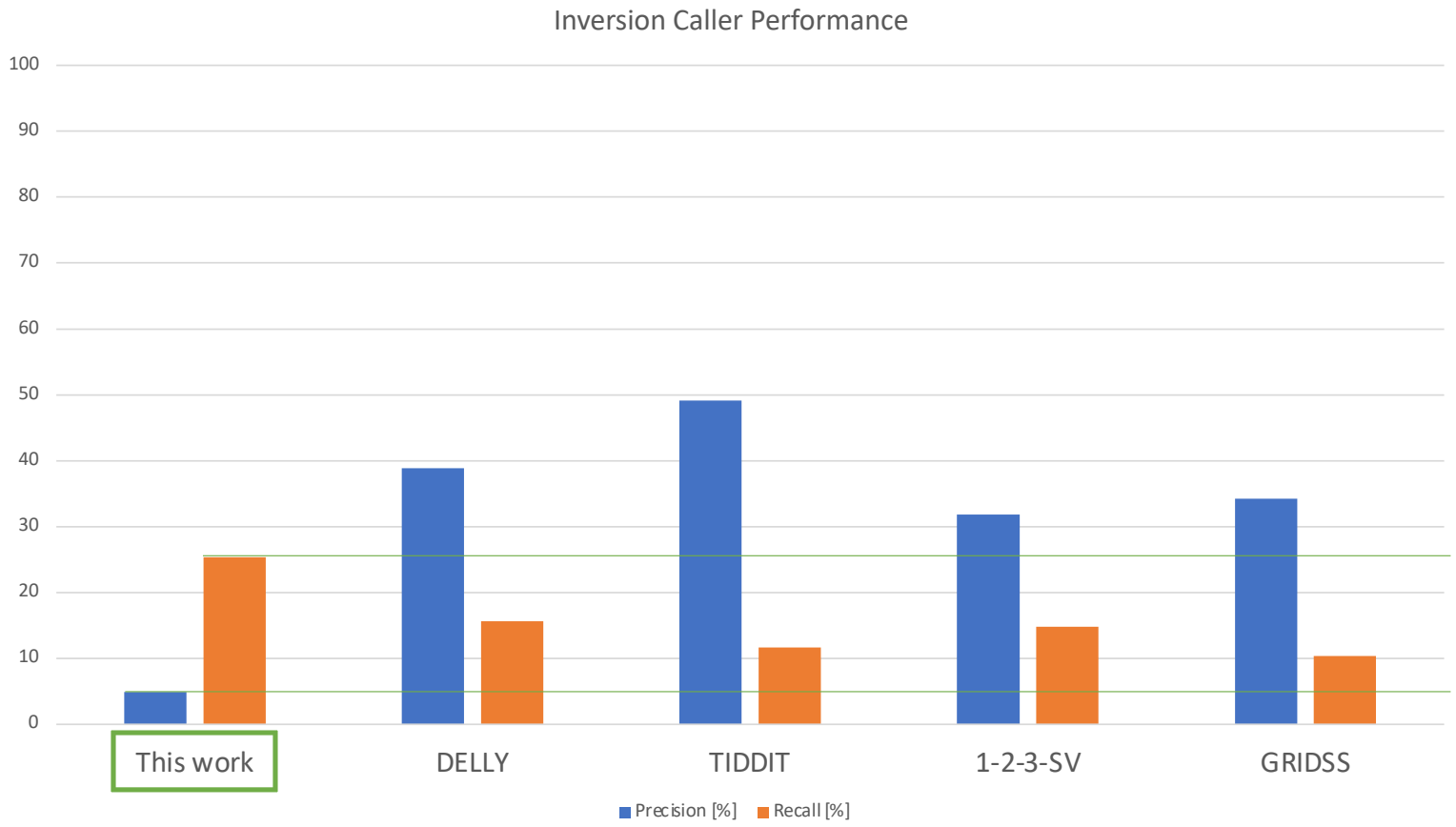
Inversion Caller – Real Data (NA12878)

Precision :

$$\frac{\#\ Inversions\ found}{\#\ Inversions\ predicted}$$

Sensitivity (Recall) :

$$\frac{\#\ Inversions\ found}{\#\ Existing\ Inversions}$$



Insertion Caller

- Not benchmarked because unfinished
- Predicted insertion sites were assessed with long reads
  - PacBio long reads from Real Data (HG001)

- Assembly results were manually explored for several regions

Results – Runtime

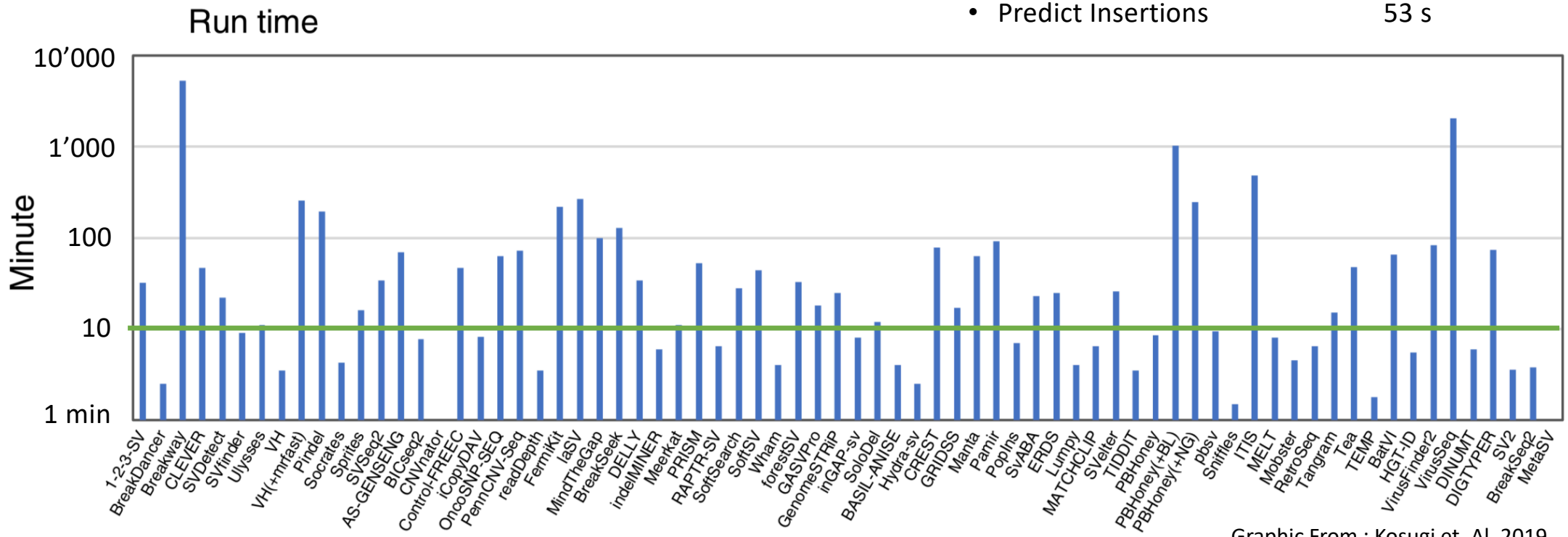
- A Whole-Genome can be analyzed in about one hour on a normal computer (30x Coverage, 100-200 GB alignment file)
  - Extracting all signals ~ 30 minutes
  - Running Deletion, Duplication, and Inversion calling ~ 10-20 minutes
- Parallelized at the chromosome level running on 8 threads

Results – Runtime

- Comparison to the state of the art
  - Single Core
  - Chromosome 8

Results – Runtime (single core)

- Chromosome 8 — **Total : 10 min 8 s**
  - Signal extraction — 5 min 46 s
  - Call Deletions — 48 s
  - Call Duplications — 1 min 20 s
  - Call Inversions — 1 min 18 s
  - Predict Insertions — 53 s



Log Scale

Graphic From : Kosugi et. Al. 2019
Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing

Results Summary

- Collection of variant callers
  - Detection capabilities close to the state of the art
  - Require tuning to get better precision (filter false positives)

- Fast runtime
  - Short development loop (~1 min to analyze a whole chromosome)
  - Can be used for cohort or population studies (~1 hour per whole-genome)

- Analysis of the missed variants and false positives
  - Insights on the caller performance