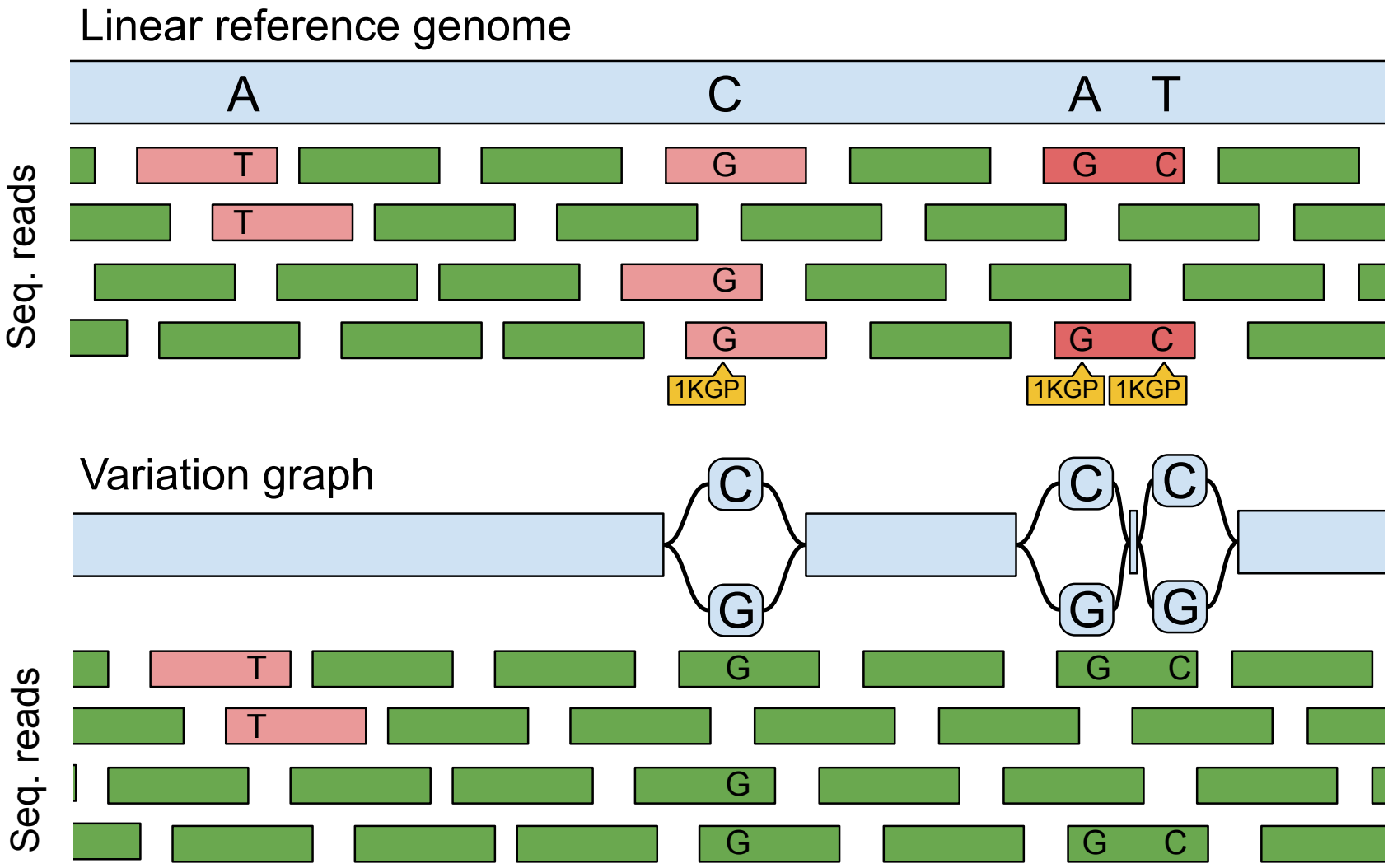
Genotyping structural variants in pangenome graphs using the vg toolkit

UNIVERSITY OF CALIFORNIA
SANTA CRUZ
Genomics Institute

Jean Monlong
November 7, 2019

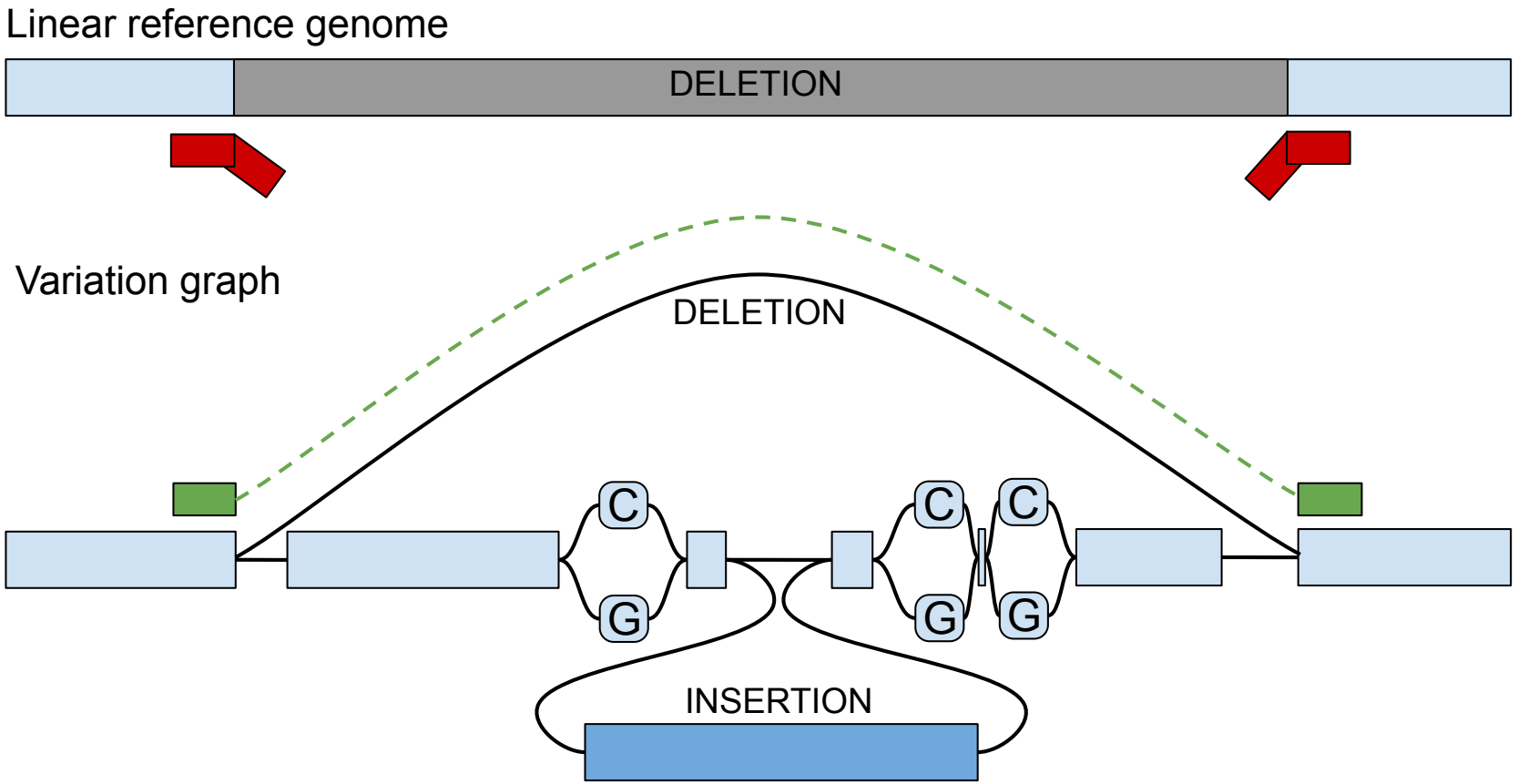
Genome Informatics

Pangenome graphs and variant-aware read mapping



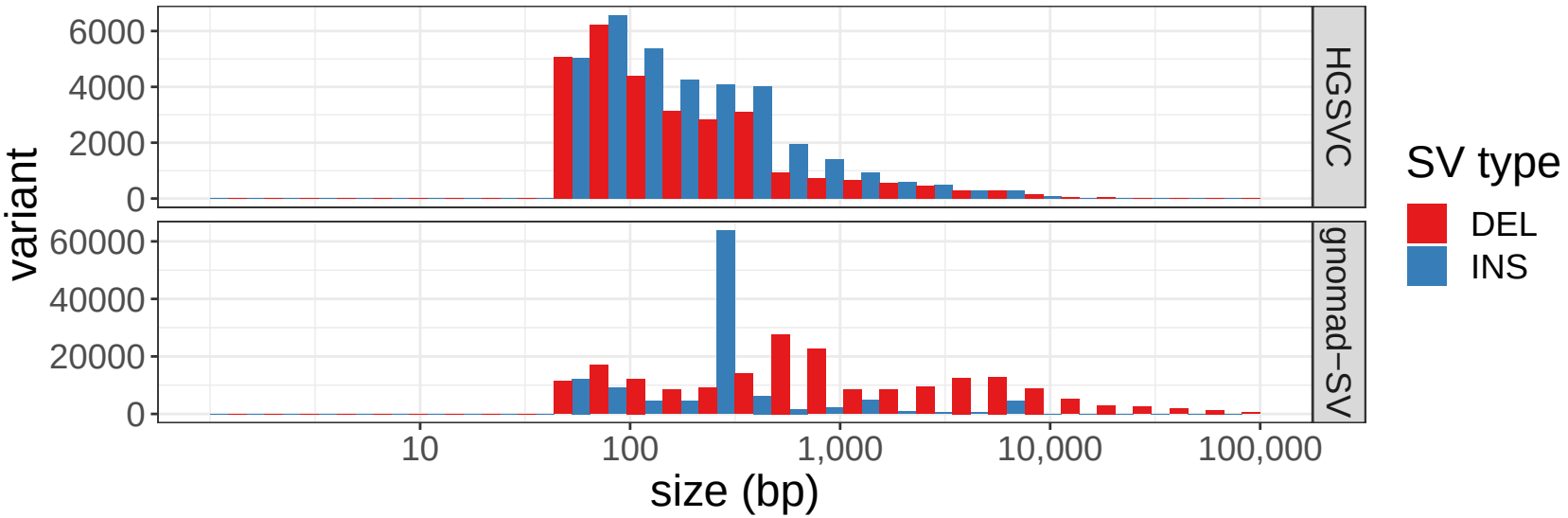
Mapping reads across structural variants

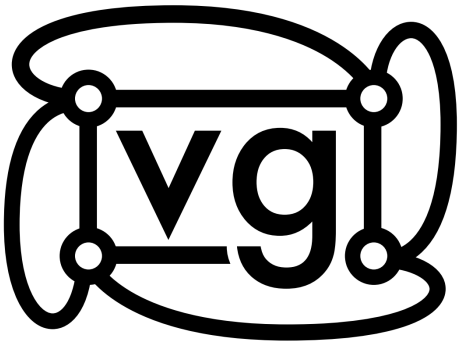
Structural variants are genomic variants larger than 50 bp, e.g. insertions, deletions, inversions translocations.



SV catalogs from long-read sequencing studies

Ref.	Project	Samples
Chaisson et al. 2019	Human Genome Structural Variation Consortium (**HGSVC**)	3
Audano et al. 2019	**SVPOP**	15
Zook et al. 2019	Genome in a Bottle (**GIAB**)	1





The **vg toolkit** is a complete, open source solution for **graph construction**, **read mapping**, and **variant calling**.

`https://github.com/vgteam/vg`

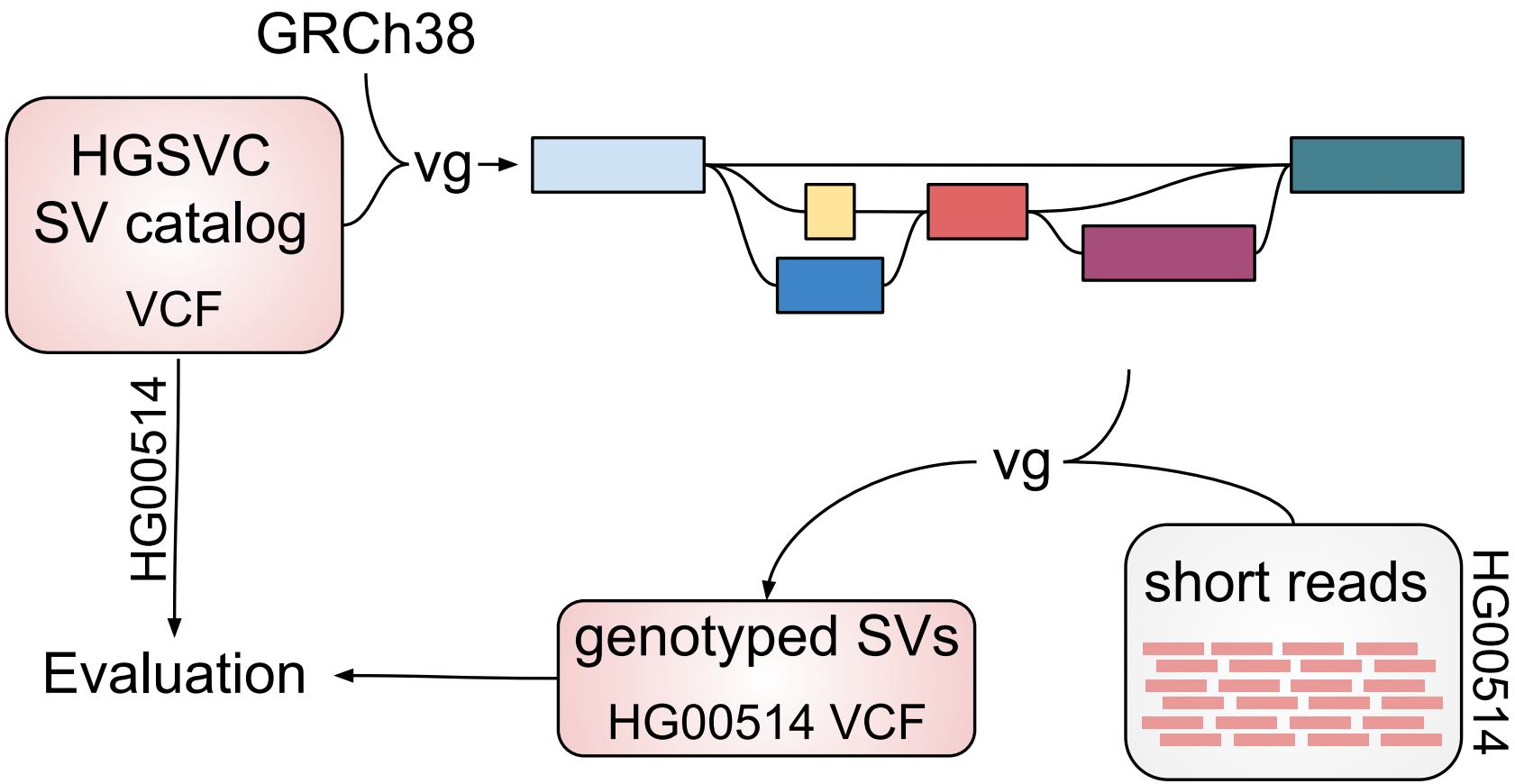
Garrison et al. Nature Biotech 2018

Can we genotype SVs from short-read sequencing datasets with the vg toolkit?

Starting from public SV catalogs or *de novo* assemblies.

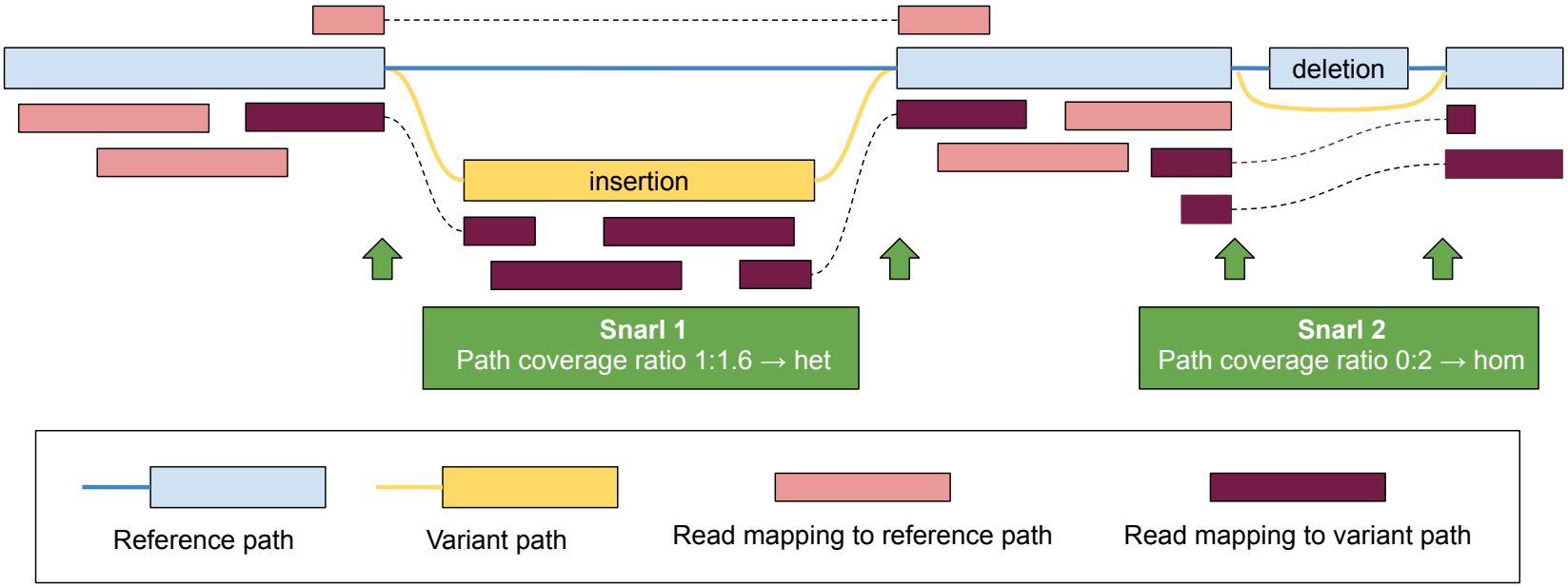
Hickey et al. bioRxiv 2019

Genotyping public SV catalogs in human



Evaluate genotype predictions for a sample from the truth set (e.g. HG00514).

Genotyping variants in vg

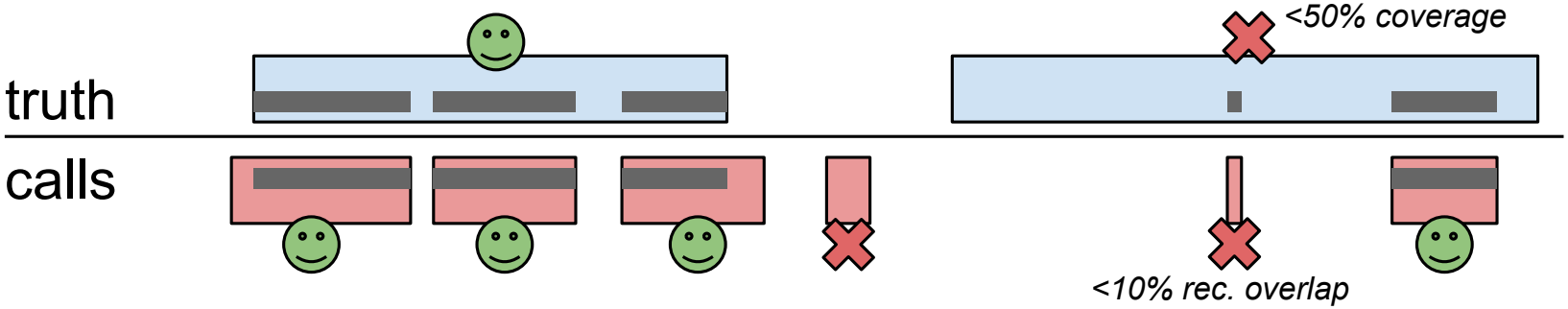


- Genotyping is based on the path coverage.
- A snarl is a variant site in the graph, a “bubble”.

Evaluating SV genotypes with a truth set

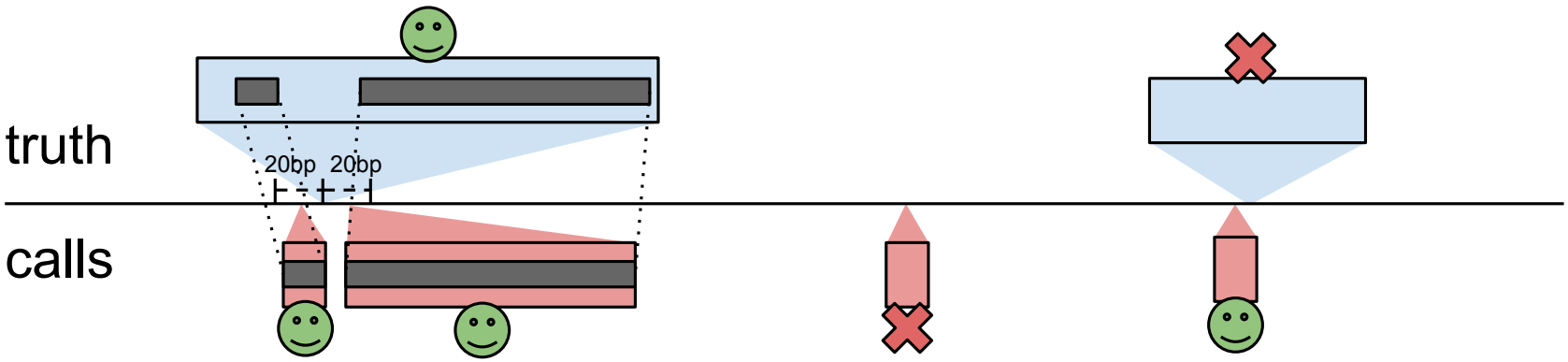
Deletions/Inversions

At least 50% coverage and 10% reciprocal overlap



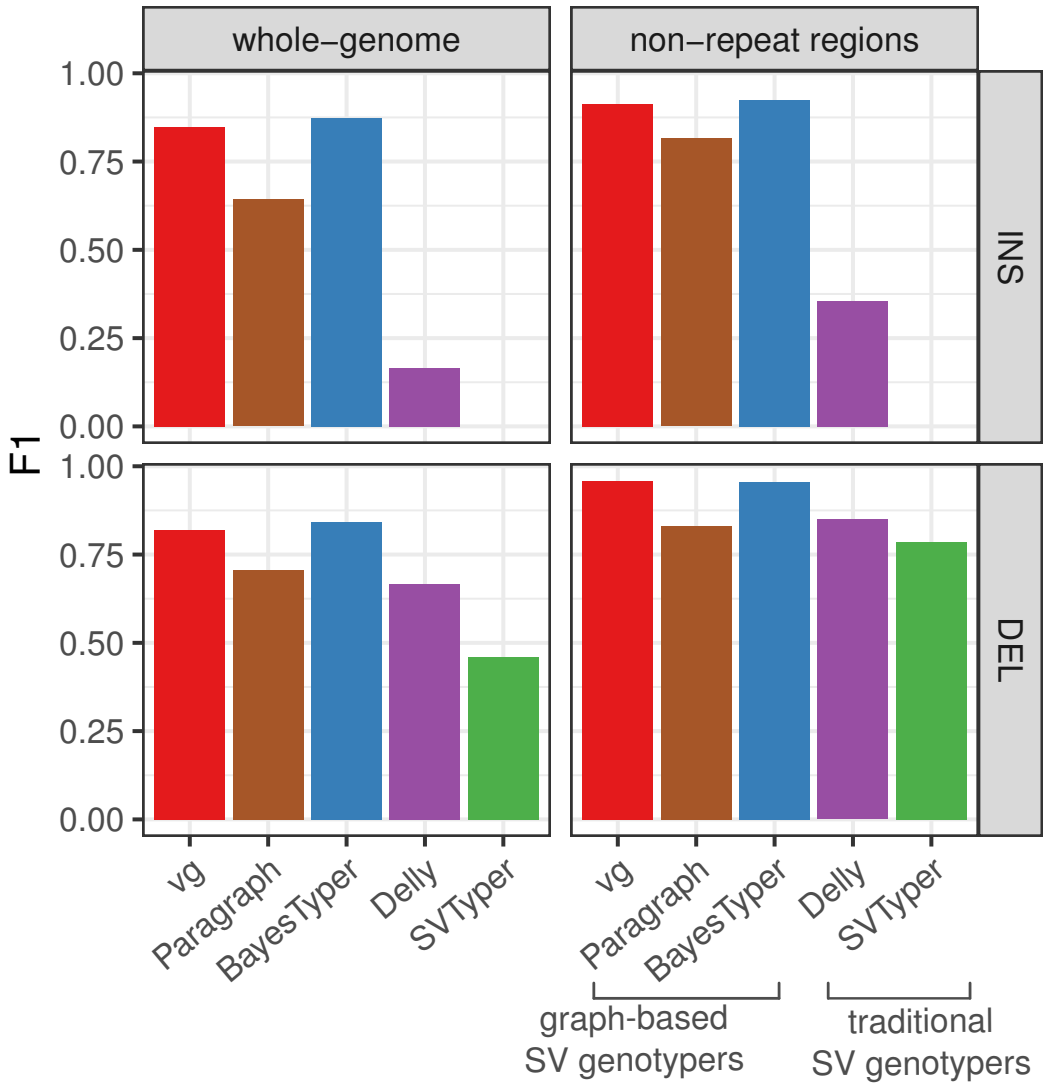
Insertions

At least 50% of inserted sequence matching nearby insertions



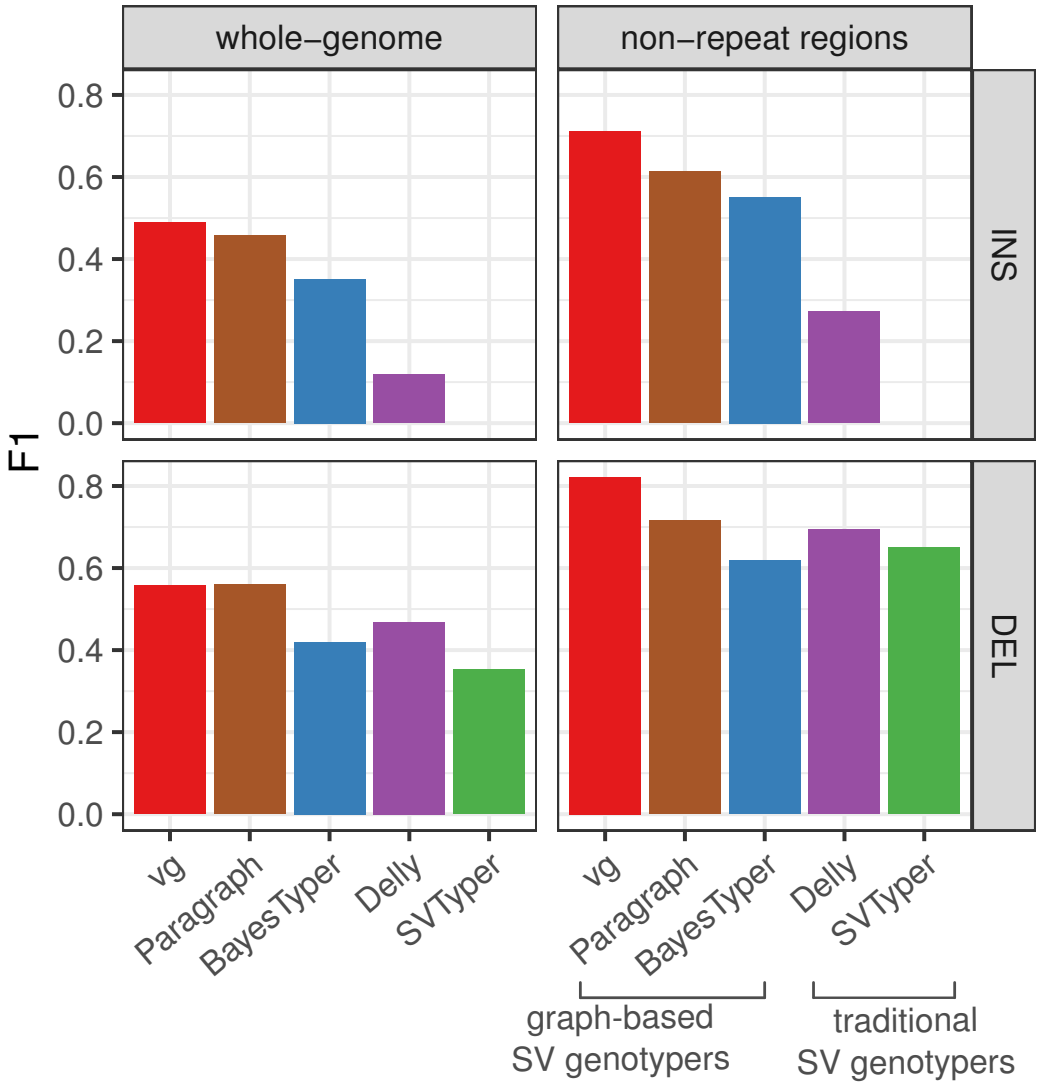
R package: `https://github.com/jmonlong/sveval`

Results on HGSVC - Simulated reads



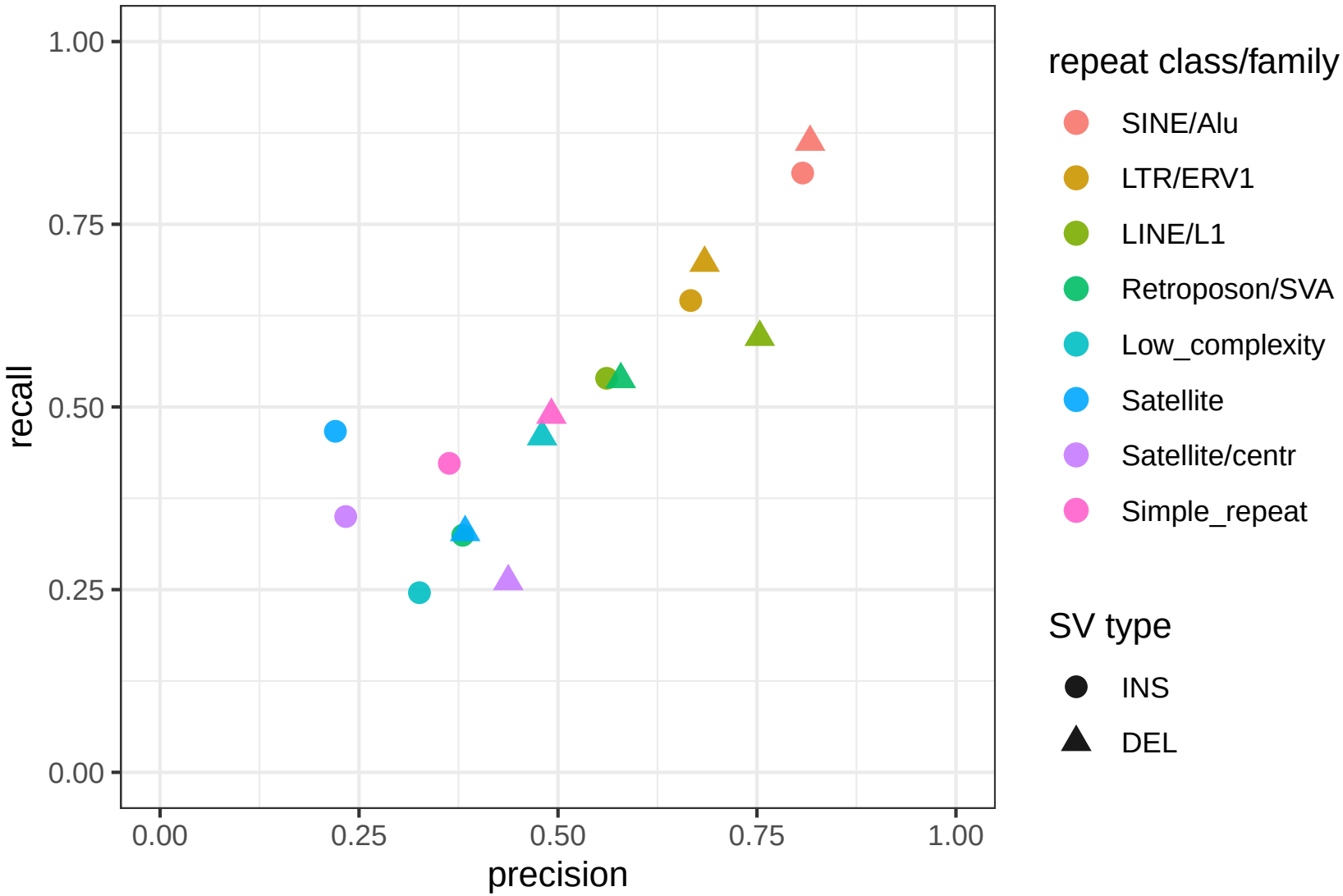
Non-repeat regions: regions not overlapping segmental duplications or simple repeats

Results on HGSVC - Real reads



Non-repeat regions: regions not overlapping segmental duplications or simple repeats

Simple repeat/low complexity regions are challenging



SV sequence annotated with RepeatMasker. Class assigned if covered ≥80% by a repeat element.

Challenges with the VCF format

Multiple equivalent representations, over-simplification, impractical.

Figure 8: Inversion

can be described equivalently in two ways. Either one uses the short hand notation described previously (recommended for simple cases):

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
2	321682	INV0	T	<INV>	6	PASS	SVTYPE=INV;END=421681

or one describes the breakends:

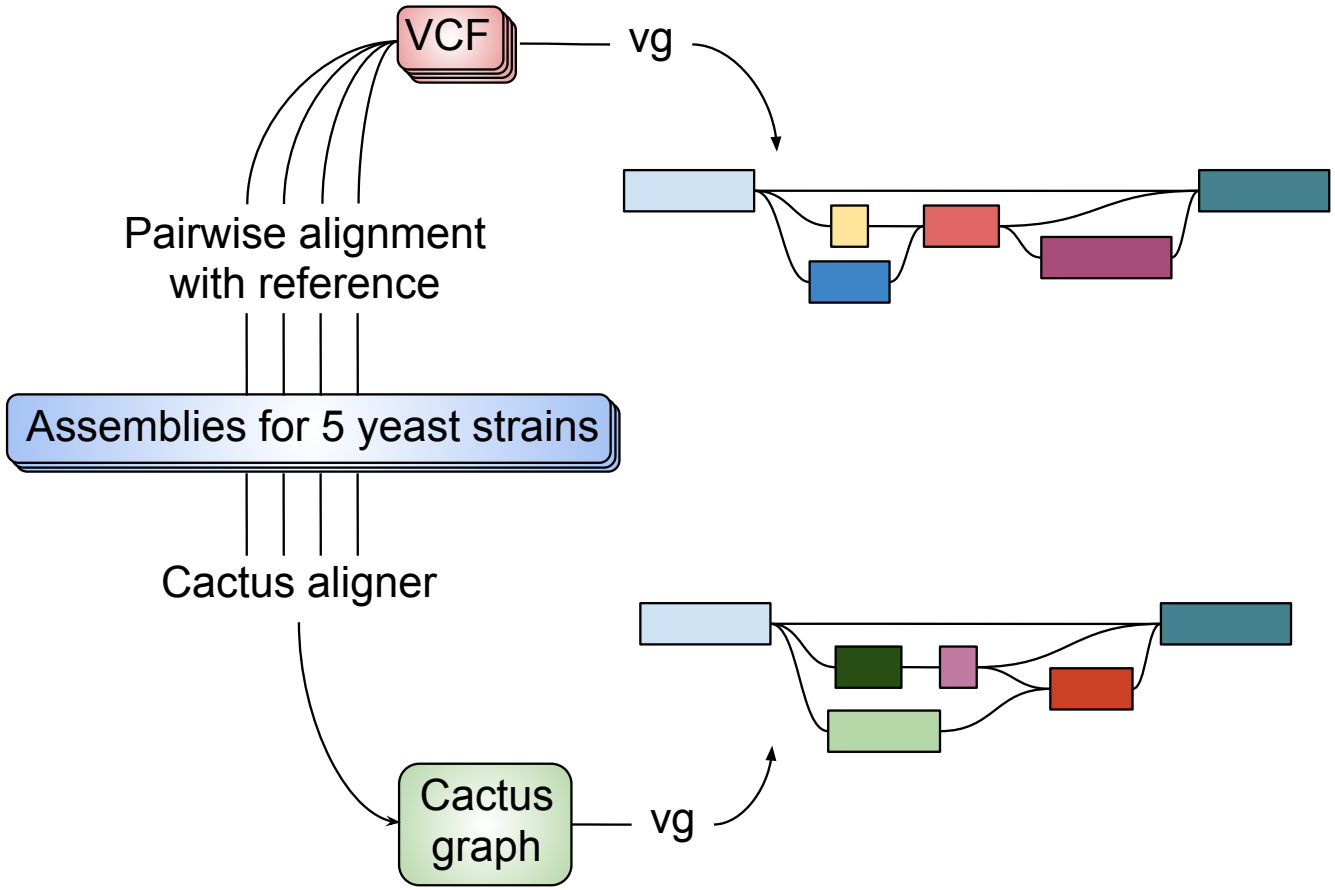
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
2	321681	bnd_W	G	G]2 : 421681]	6	PASS	SVTYPE=BND;MATEID=bnd_U;EVENT=INV0
2	321682	bnd_V	T	[2 : 421682[T	6	PASS	SVTYPE=BND;MATEID=bnd_X;EVENT=INV0
2	421681	bnd_U	A	A]2 : 321681]	6	PASS	SVTYPE=BND;MATEID=bnd_W;EVENT=INV0
2	421682	bnd_X	C	[2 : 321682[C	6	PASS	SVTYPE=BND;MATEID=bnd_V;EVENT=INV0

VCF v4.2 specs

Why not start directly from de novo assemblies?

Analysis of 12 yeast strains from 2 clades

Selected 5 strains to build graph: one reference + 2 per clade.

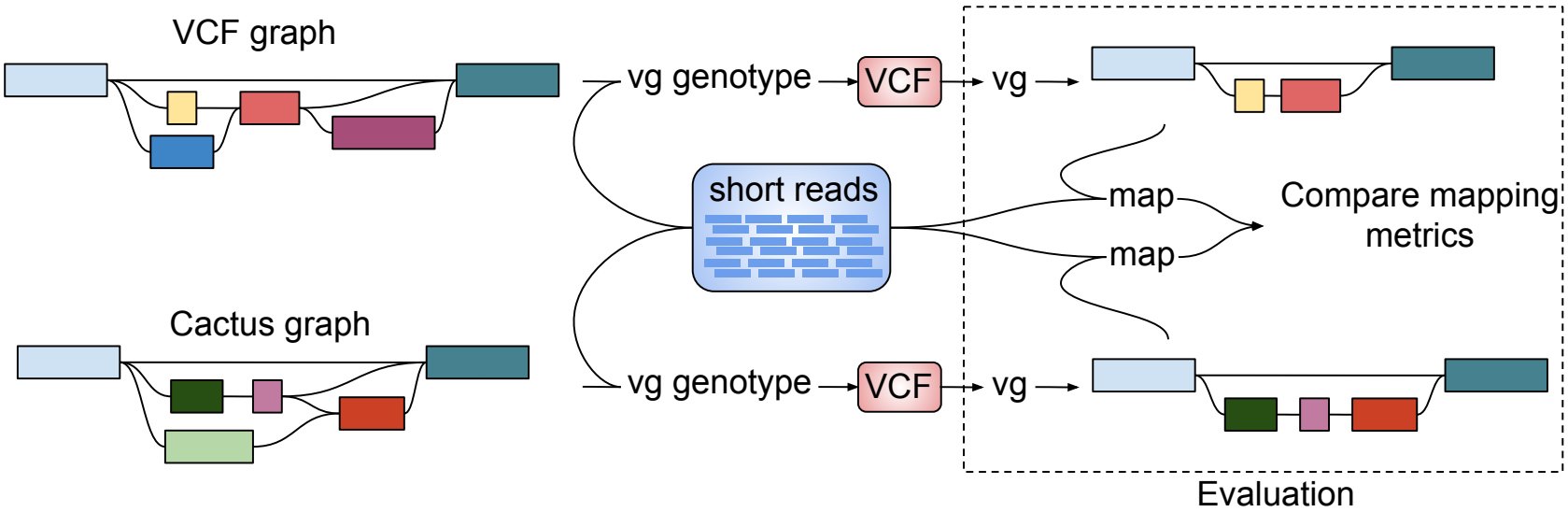


Evaluating SV genotyping using mapping statistics

- No gold-standard to compare with.

Evaluating SV genotyping using mapping statistics

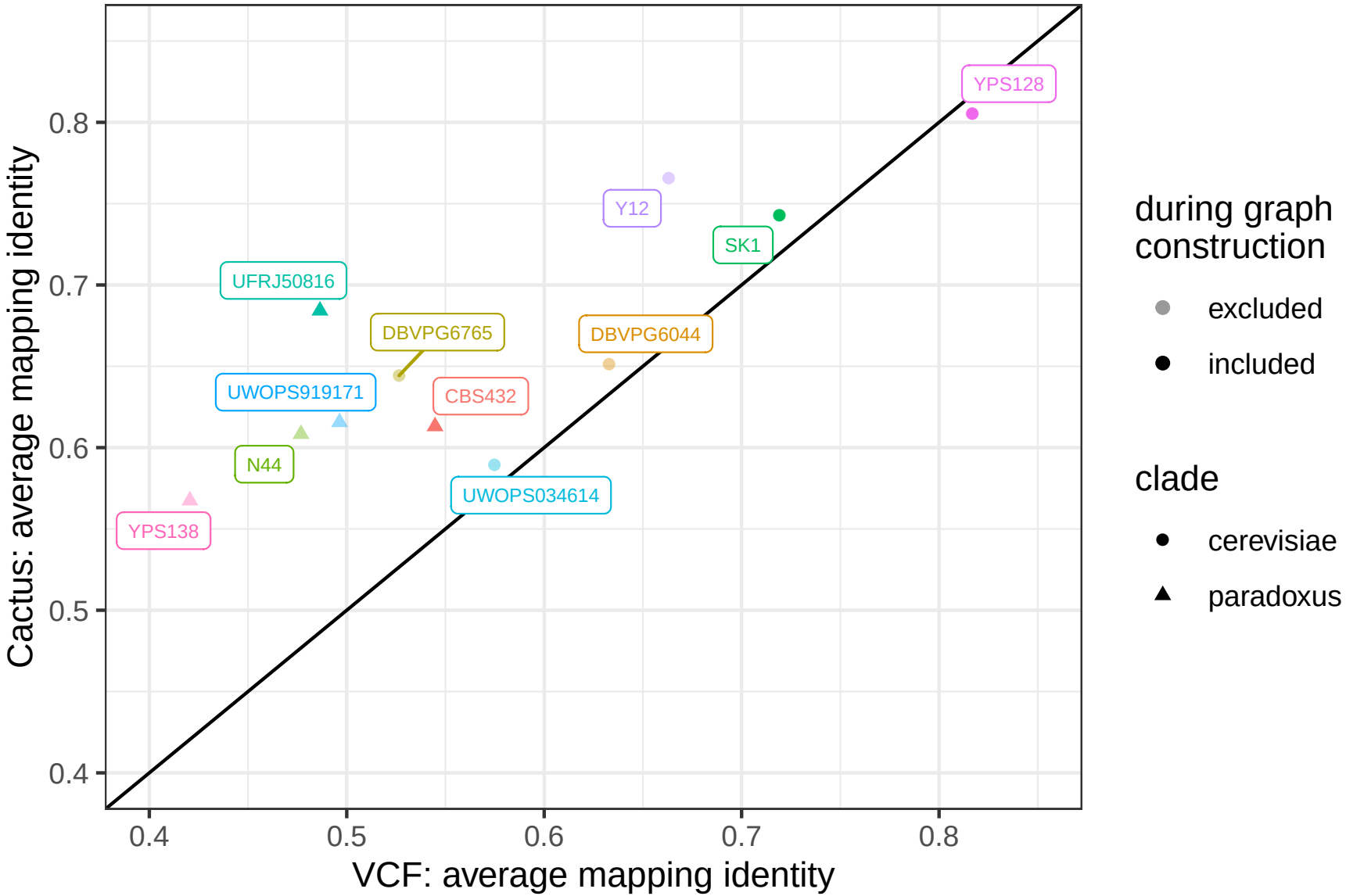
- No gold-standard to compare with.
- Map reads to a sample graph built from the SV calls:



- Mapping quality $\sim$ Sample graph quality $\sim$ SV calls quality.

Better mapping for SVs called in the cactus graph

Analysis restricted to reads at variation sites.



Conclusions

- The vg toolkit can integrate and genotype SVs.
- Graphs from *de novo* assemblies alignment performs better.

Hickey et al. bioRxiv 2019 `https://jmonlong.github.io/manu-vgsv/`

Conclusions

- The vg toolkit can integrate and genotype SVs.
- Graphs from *de novo* assemblies alignment performs better.

Hickey et al. bioRxiv 2019 `https://jmonlong.github.io/manu-vgsv/`

Future directions

- Experiment with **high-quality human de novo assemblies** (e.g. the Human PanGenome Project).
- **Combine** public SV catalogs and **genotype** SVs in a large and diverse cohort.

Acknowledgment

Benedict Paten
Glenn Hickey
David Heller
Adam Novak
Erik Garrison
Jouni Siren
Jordan Eizenga
Charles Markello
Xian Chang
Robin Rounthwaite
Jonas Sibbesen
Eric T. Dawson

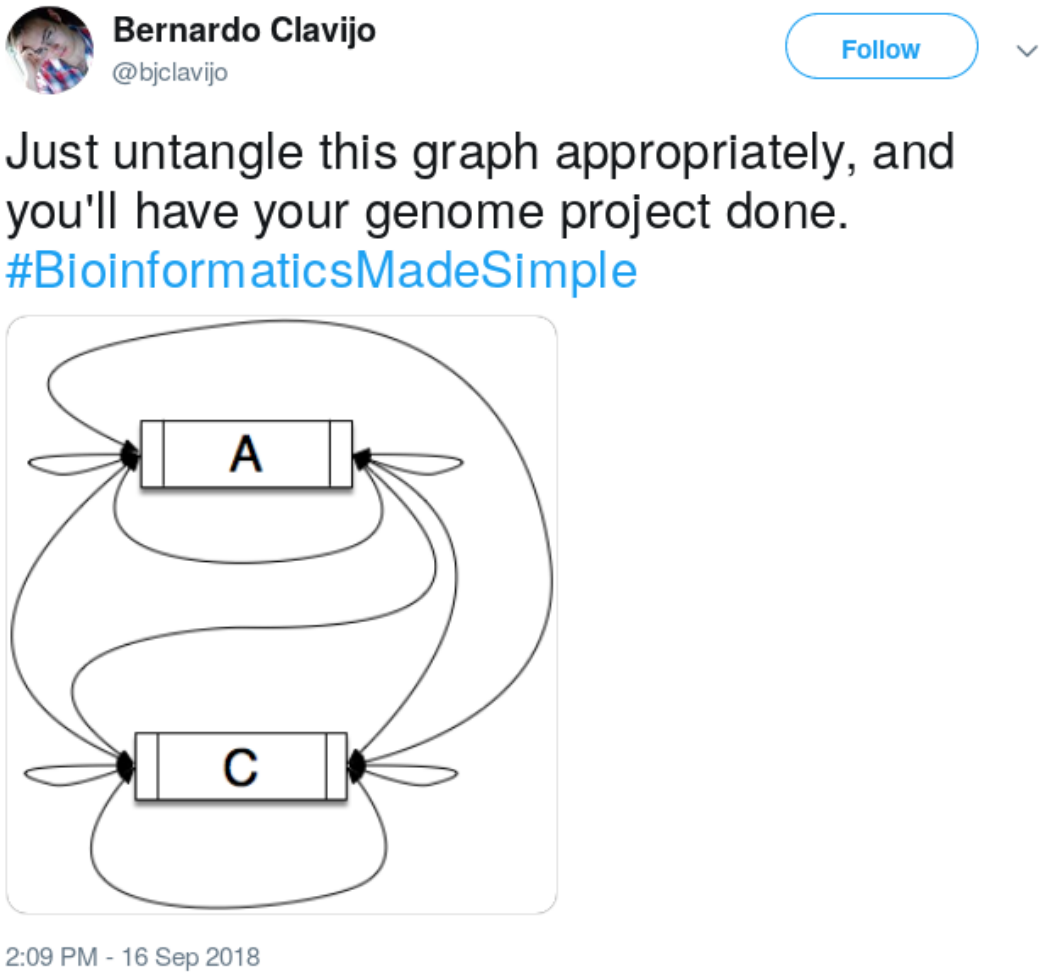
UNIVERSITY OF CALIFORNIA
SANTA CRUZ
Genomics Institute

EMBL
European Molecular Biology Laboratory

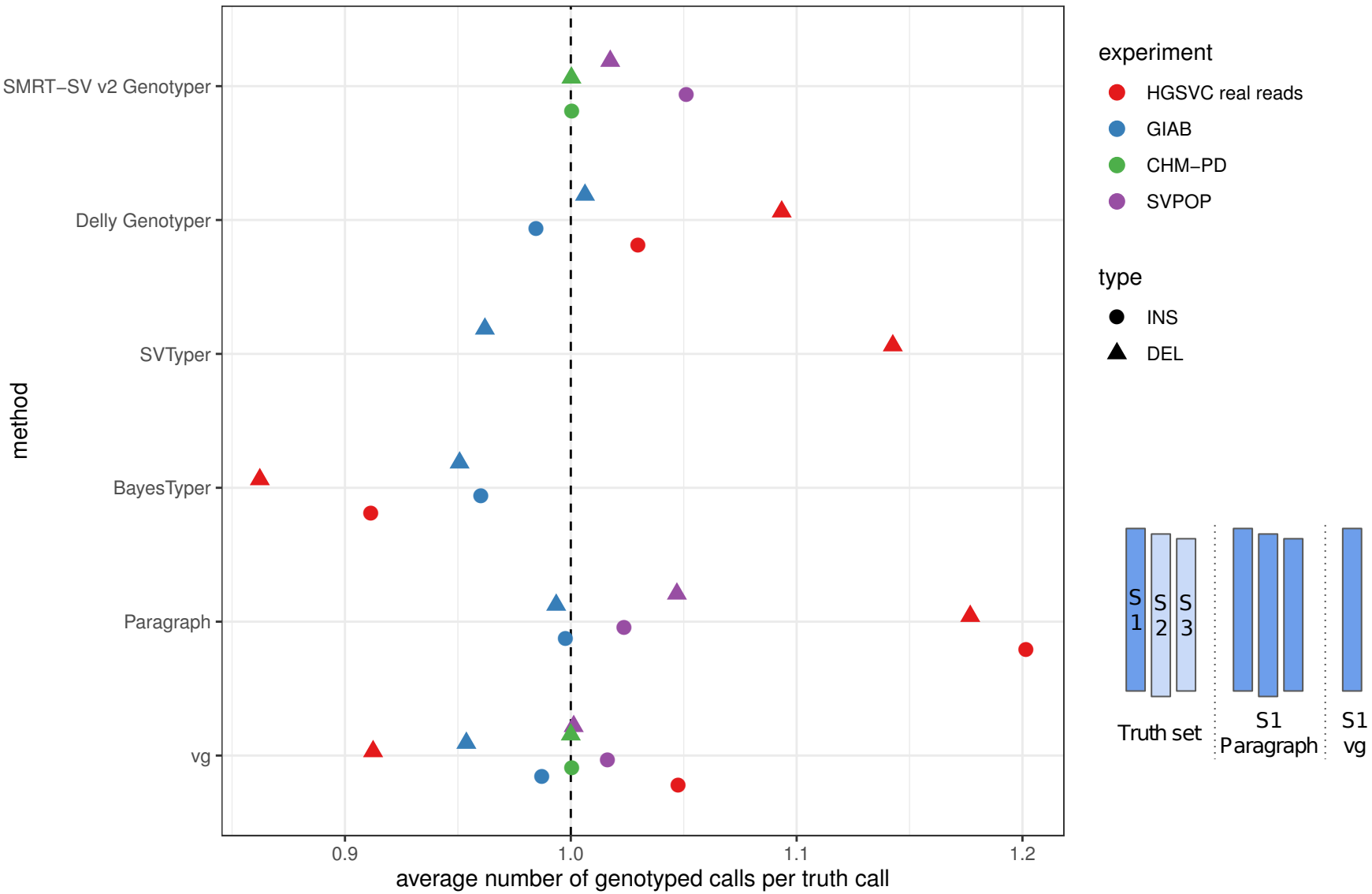
W.M. KECK FOUNDATION

NIH
NHGRI

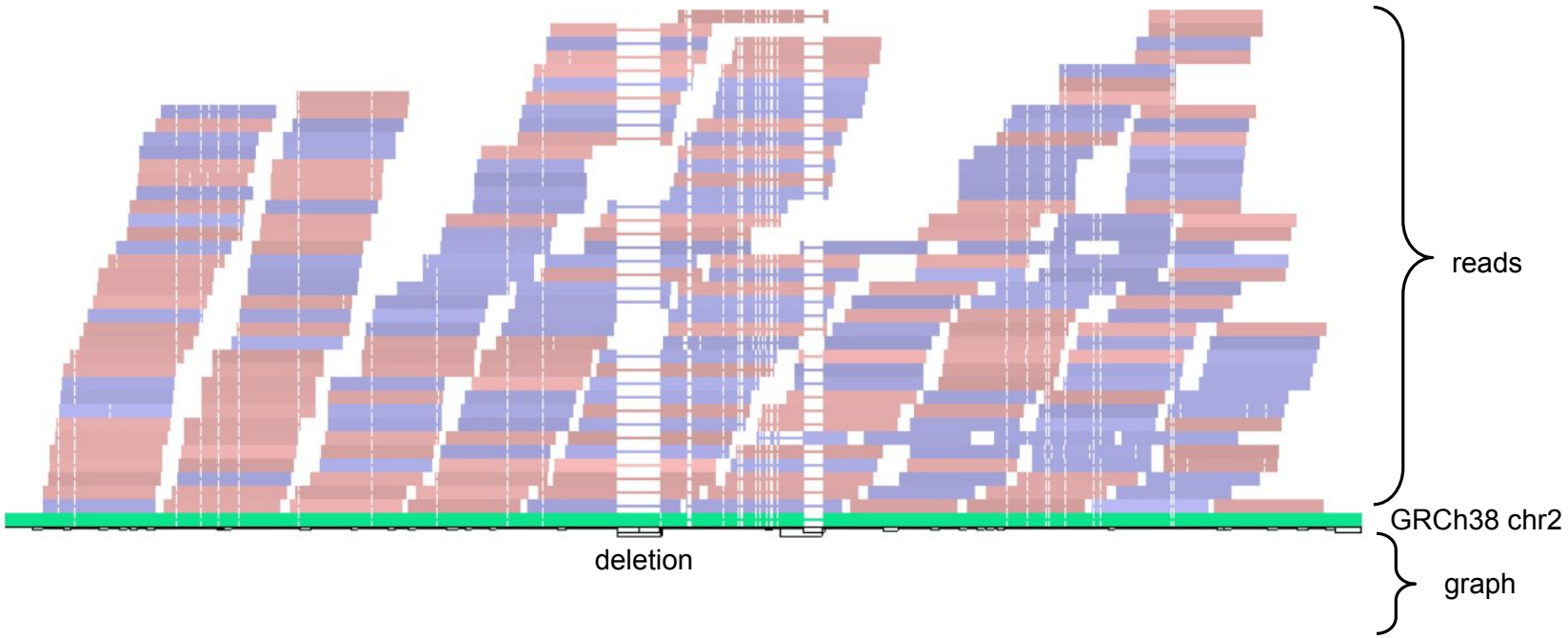
Universal genome graph



Some methods "over-genotype" similar variants

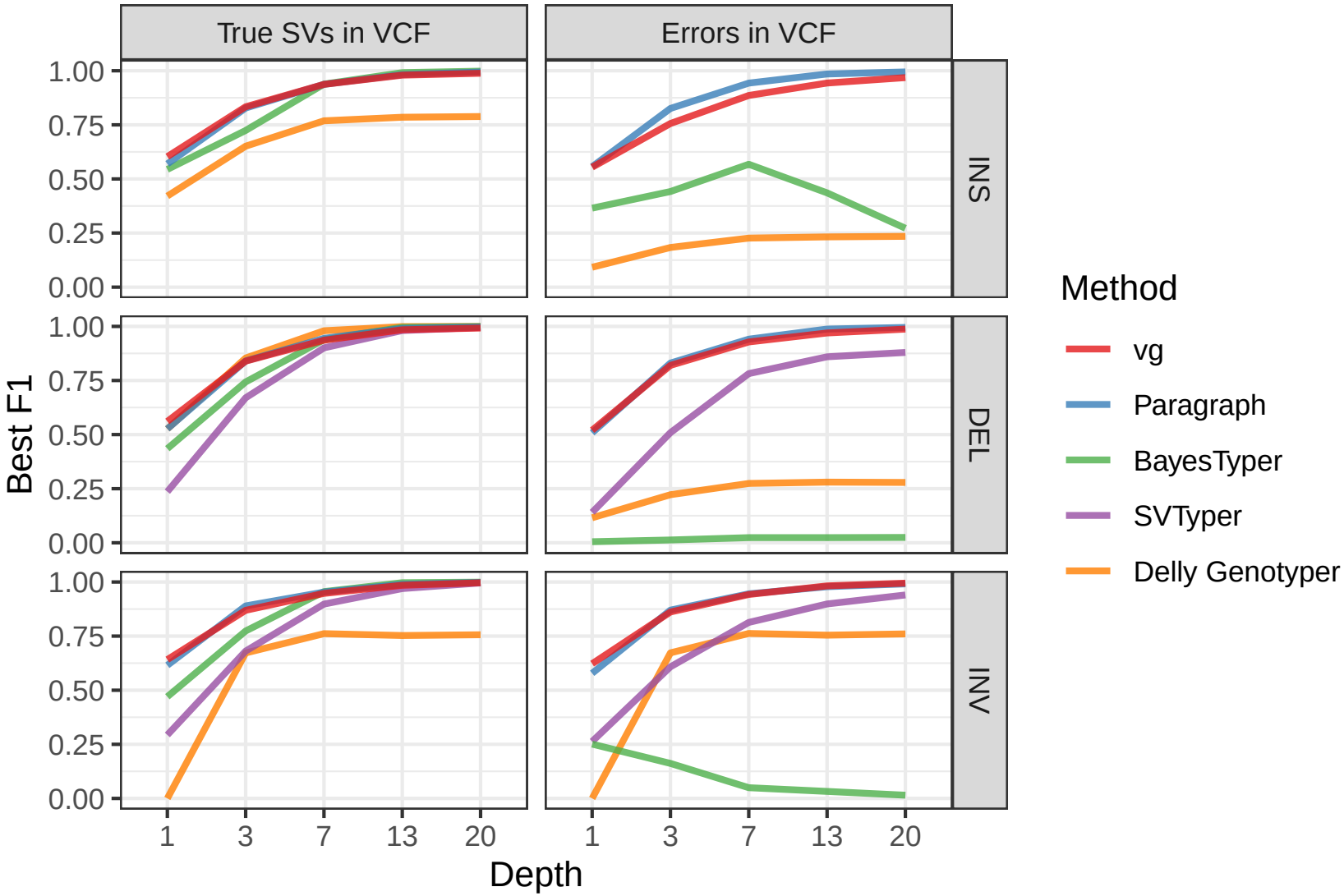


Deletion correctly genotyped by vg

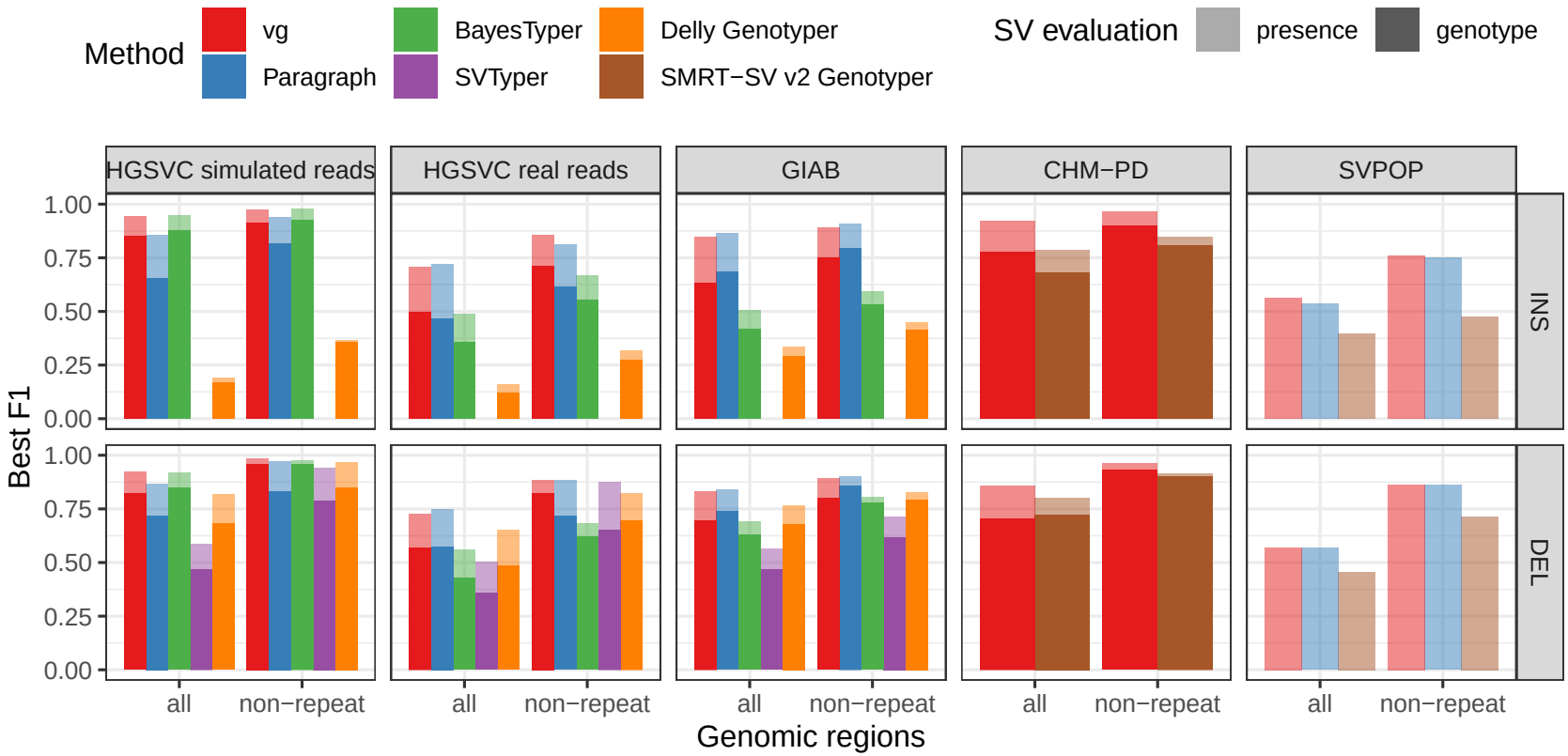


51 bp homozygous deletion in the 3' UTR of the LONRF2 gene.

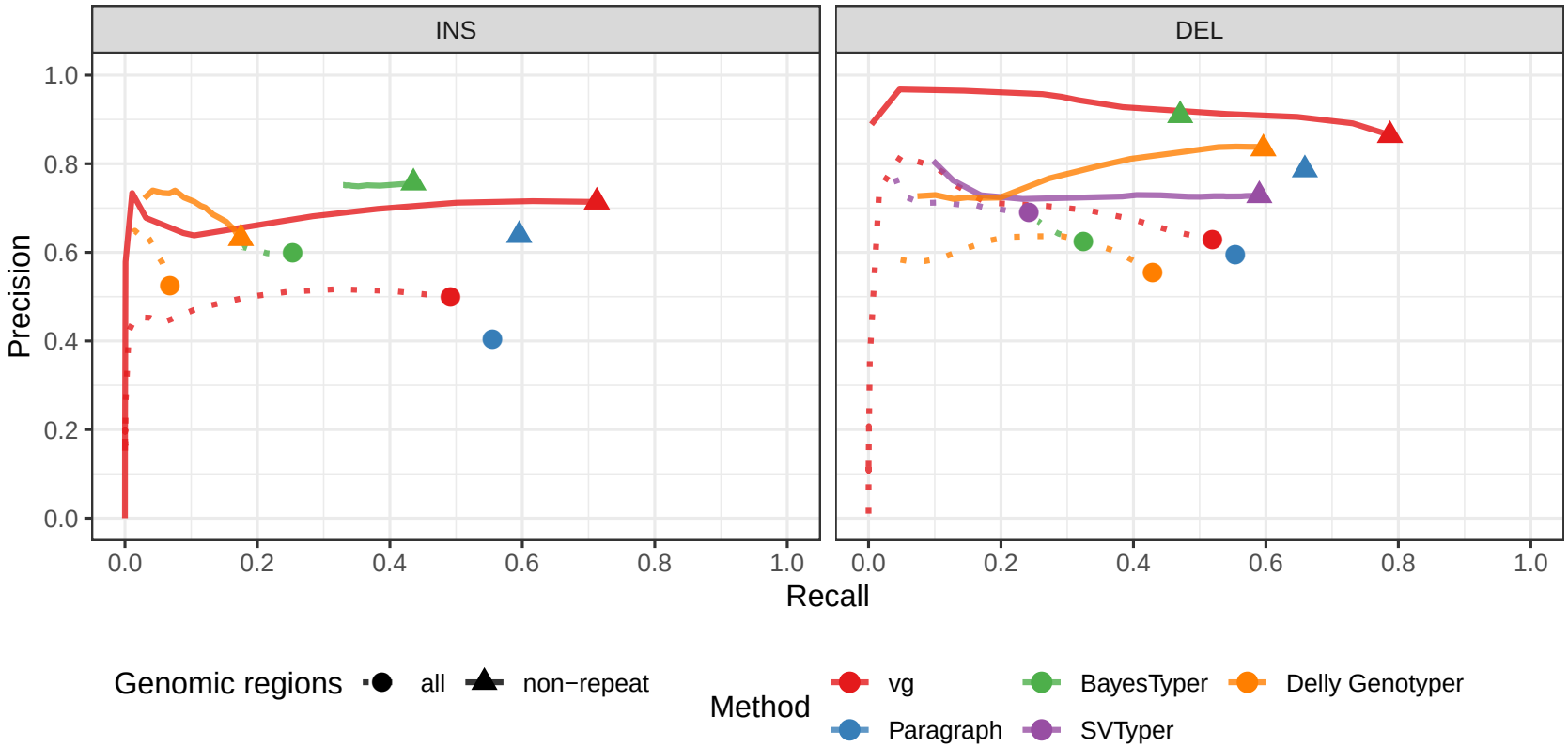
Simulation experiment



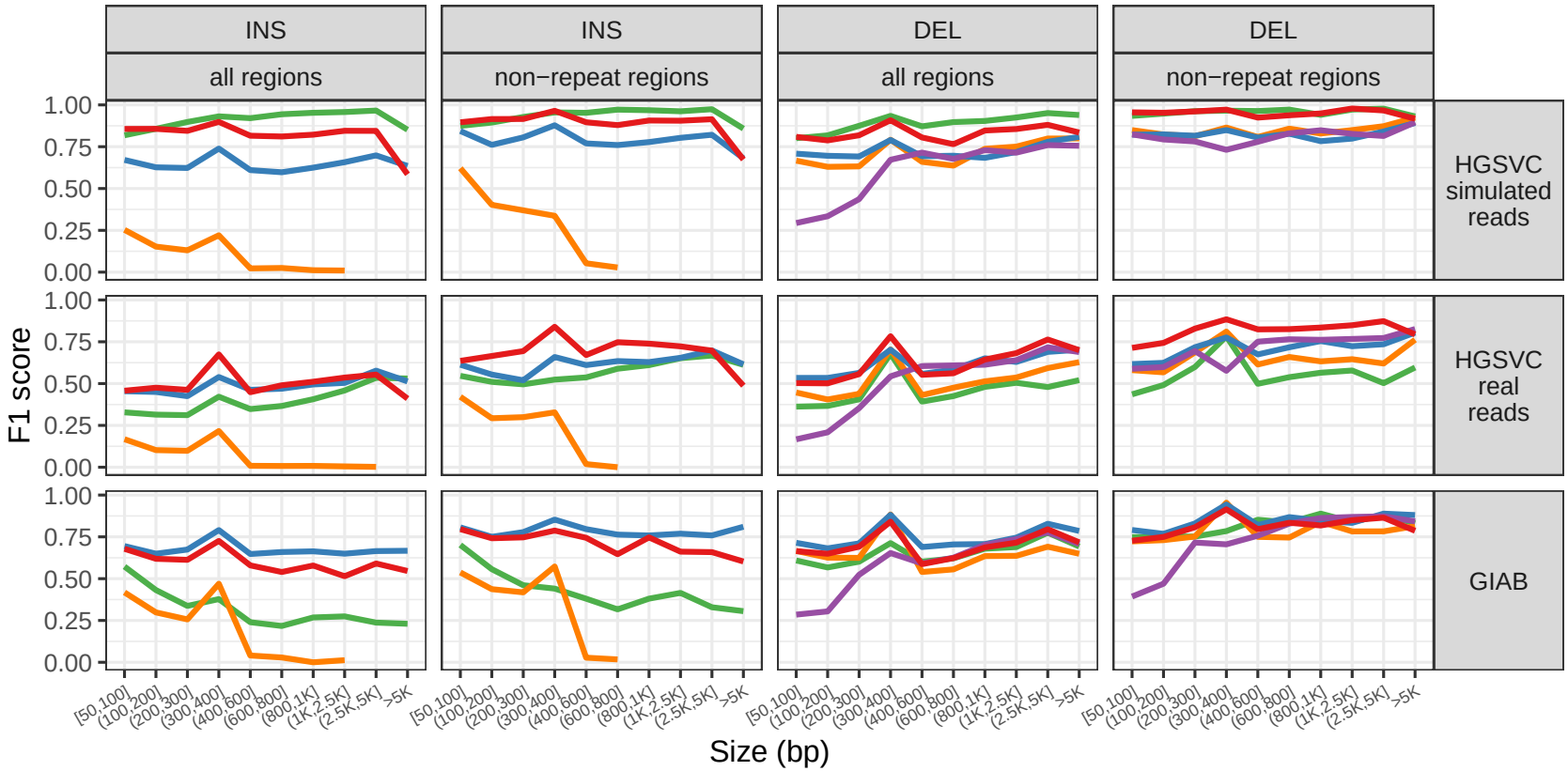
SV catalog summary results



Precision-recall curve



Evaluation per SV size



Better mapping for SVs called in the cactus graph

Analysis restricted to reads at variation sites.