









Genotyping structural variation in variation graphs with the vg toolkit

This manuscript ([permalink](#)) was automatically generated from [jmonlong/manu-vgs@73fe12](#) on March 6, 2019.

Authors

 Glenn Hickey^{1, },  David Heller^{1, },  Jean Monlong^{1, },  Adam Novak¹,  Benedict Paten^{1, †}

 — These authors contributed equally to this work

[†] — To whom correspondence should be addressed: bpaten@ucsc.edu

1. UC Santa Cruz Genomics Institute, University of California, Santa Cruz, California, USA

Abstract

Introduction

Structural variation (SV) represents genomic mutation involving 50 bp or more and can take several forms, such as for example deletions, insertions, inversions, or translocations. Although whole-genome sequencing (WGS) made it possible to assess virtually any type of structural variation, many challenges remain. In particular, SV-supporting reads are difficult to map to reference genomes. Multi-mapping, caused by widespread repeated sequences in the genome, is another issue because it often resembles SV-supporting signal. As a result, many SV detection algorithms have been developed and multiple methods must usually be combined to minimize false positives. Several large-scale projects used this ensemble approach, cataloging tens of thousands of SV in humans[1,2]. SV detection from short-read sequencing remains laborious and of lower accuracy, explaining why these variants and their impact have been under-studied as compared to single-nucleotide variants (SNVs) and small insertions/deletions (indels).

Over the last few years, exciting developments in sequencing technologies and library preparation made it possible to produce long reads or retrieve long-range information over kilobases of sequence. These approaches are maturing to the point where it is feasible to analyze the human genome. This multi-kbp information is particularly useful for SV detection and de novo assembly. In the last few years, several studies using long-read or linked-read sequencing have produced large catalogs of structural variation, the majority of which were novel and sequence-resolved[3,4,5,6,7]. These technologies are also enabling high-quality de novo genome assemblies to be produced[3, 8], as well as large blocks of haplotype-resolved sequences[9]. These technological advances promise to expand the amount of known genomic variation in humans in the near future.

In parallel, the reference genome is evolving from a linear reference to a graph-based reference that contains known genomic variation[10,11,12]. By having variants in the graph, mapping rates are increased and variants are more uniformly covered, including indels and variants in complex regions[11]. Both the mapping and variant calling become variant-aware and benefit in terms of accuracy and sensitivity. In addition, different variant types are called simultaneously by a unified framework. Graphs have also been used locally, i.e. to call variants at the region level. GraphTyper[13] and BayesTyper[14] both construct variation graphs of small regions and use them for variant genotyping. Here again, the graph-approach showed clear advantages over standard approaches that use the linear reference. Other SV genotyping approaches compare read mapping in the reference genome and a sequence modified with the SV. For example SMRT-SV was designed to genotype SVs identified on PacBio reads[4], SVTyper uses paired-end mapping and split-read mapping information[15], and Delly provides a genotyping feature in addition to its discovery mode[16].

Results

Structural variation in vg

In addition to SNV and short indels, vg can handle large deletions and insertions (and inversion?) (Figure 1a). As a proof-of-concept we simulated genomes and SVs with a size distribution matching real SVs[17]. Some errors were added at the breakpoints to investigate their effect on genotyping. In all simulations, vg performed as well or better than SVtyper[15] and Delly[16] (Figure 1b). Overall, vg was more robust to errors around the breakpoints. For insertions, vg performed almost as well as in the presence of errors in contrast to Delly and BayesTyper whose performance dropped significantly. For deletions, the F1 dropped for all methods but remained high for vg and SVtyper.

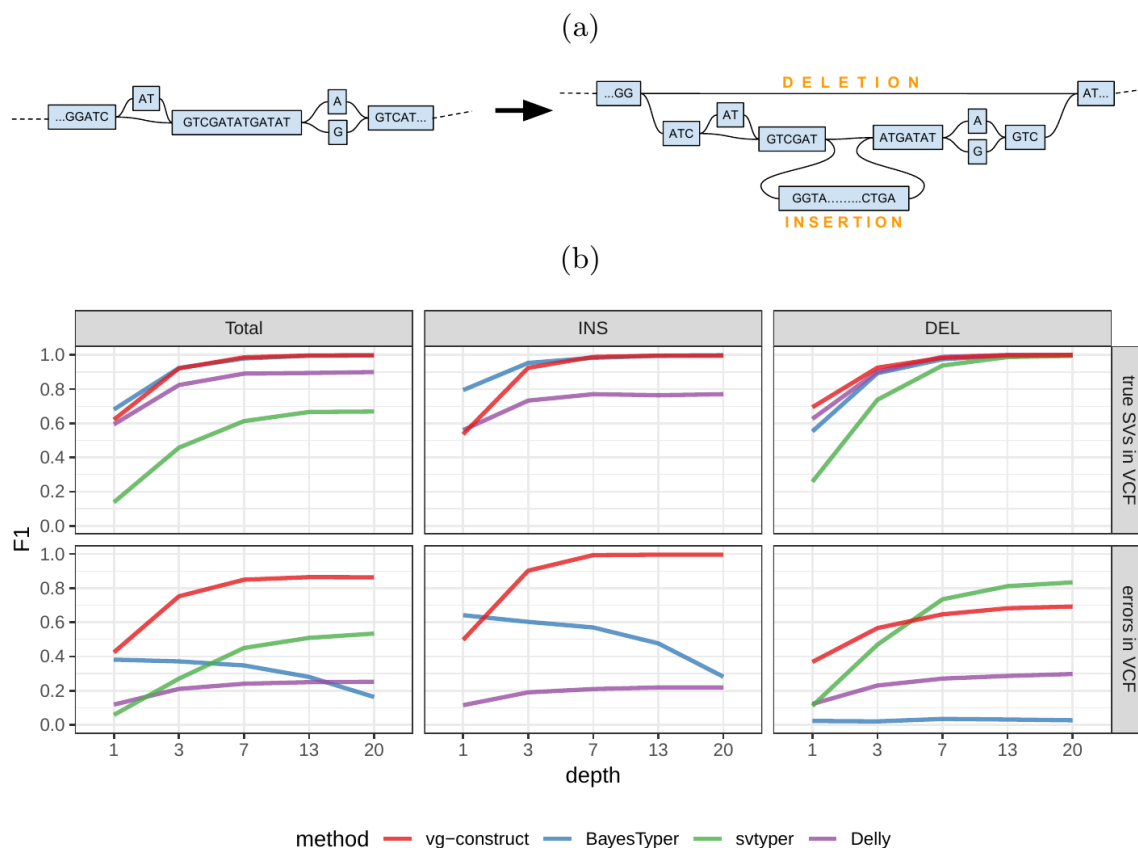


Figure 1: **Structural variation in vg.** a) Adding large deletions and insertions in a variation graph. b) Simulation experiment. For each experiment (method, depth and input VCF with/without errors), the deciles of the call qualities were used as threshold and the maximum F1 is reported on the y-axis.

HGSVC

Chaisson et al.[\[17\]](#) provide a high-quality SV catalog of three samples, obtained using a consensus from different sequencing, phasing and variant calling technologies.

(Whole-genome) Simulation

The phasing information in the HGSVC VCF was used to extract two haplotypes for sample HG00514, and 30X paired-end reads were simulated using vg sim. The reads were used to call VCFs then compared back to the original HGSVC calls (Figure [2a](#) and Table [1](#)).

When restricting the comparisons to regions not identified as tandem repeats or segmental duplications in the Genome Browser (Table [2](#)).

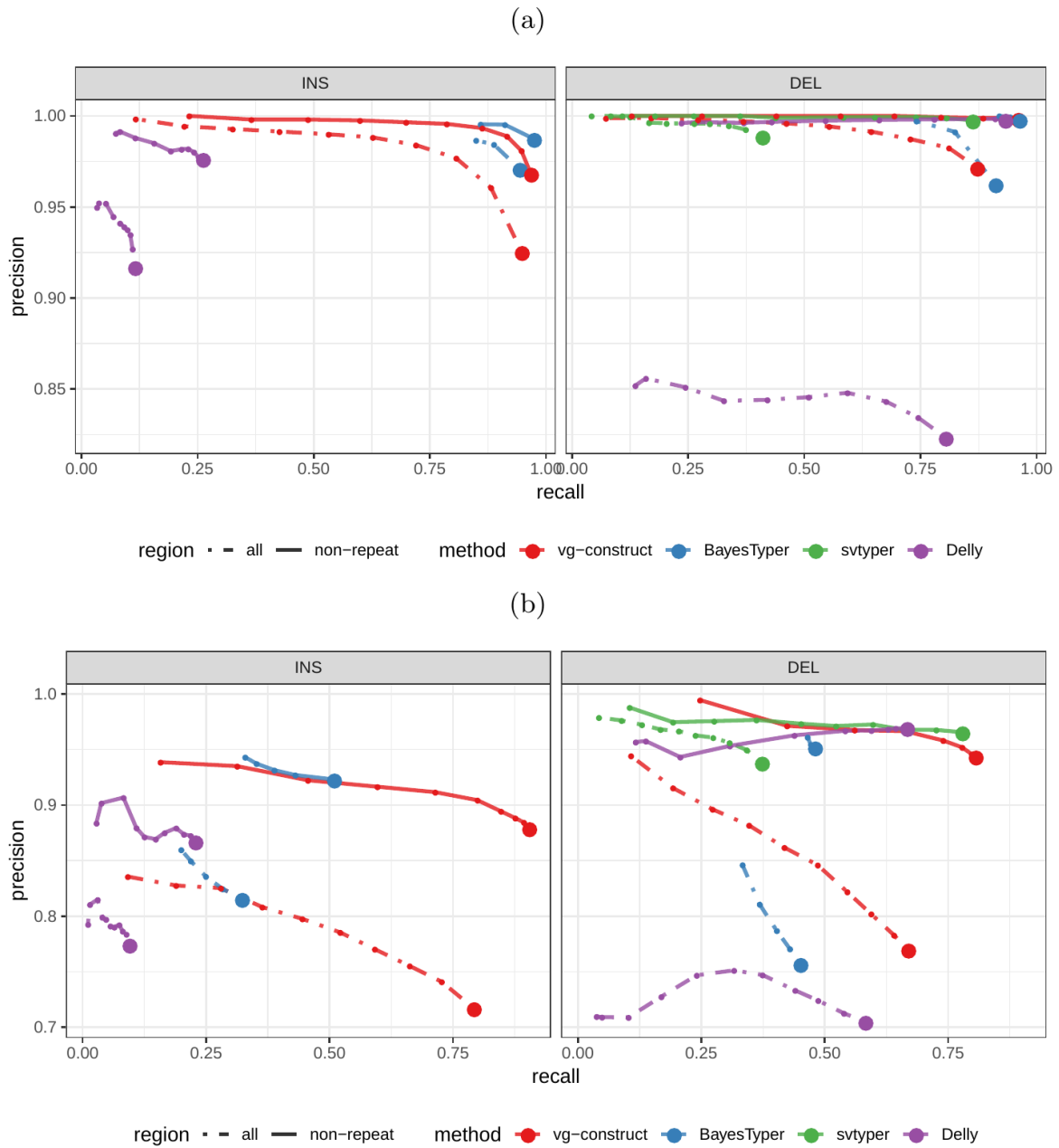


Figure 2: **Structural variants from the HGSVC dataset.** a) Simulated reads. b) Real Illumina reads.

Genotyping evaluation results in Fig 5.

(Whole-genome) Real reads

Figure 2b. Tables 3 and 4 for results over the genome or when restricting the comparisons to regions not identified as tandem repeats or segmental duplications in the Genome Browser.

The input SVs could also be derived from short-reads. For example, SVs were first discovered by Manta(REF) from Illumina reads and then genotyped. For BayesTyper, the precision drops for insertions while the recall increases for deletions (Figure [???]).

Genotyping evaluation results in Fig 6.

Genotyping SV using vg and de novo assemblies

We investigated whether genome graphs derived from genome-genome alignments yield advantages for SV genotyping. To this end, we analyzed public sequencing datasets for 12 yeast strains from two clades (*S. cerevisiae* and *S. paradoxus*) [18]. From these datasets, we generated two different types of genome graphs. The first graph type (in the following called *construct graph*) was created from a linear reference genome of the *S.c.* S288C strain and a set of SVs relative to this reference strain in VCF format. We compiled the SV set using the output of three methods for SV detection from genome assemblies: Assemblytics [19], AsmVar [20] and paftools [21]. All three methods were run to detect SVs between the reference strain *S.c.* S288C and each of the other 11 strains. Merging the results from the three methods and the 11 strains provided us with a high-sensitivity set of SVs occurring in the two yeast clades. We used this set to construct the *construct graph*. The second graph (in the following called *cactus graph*) was derived from a multiple genome alignment of all 12 strains using our Cactus tool [22]. While the *construct graph* is still mainly linear and highly dependent on the reference genome, the cactus graph is completely unbiased in that regard.

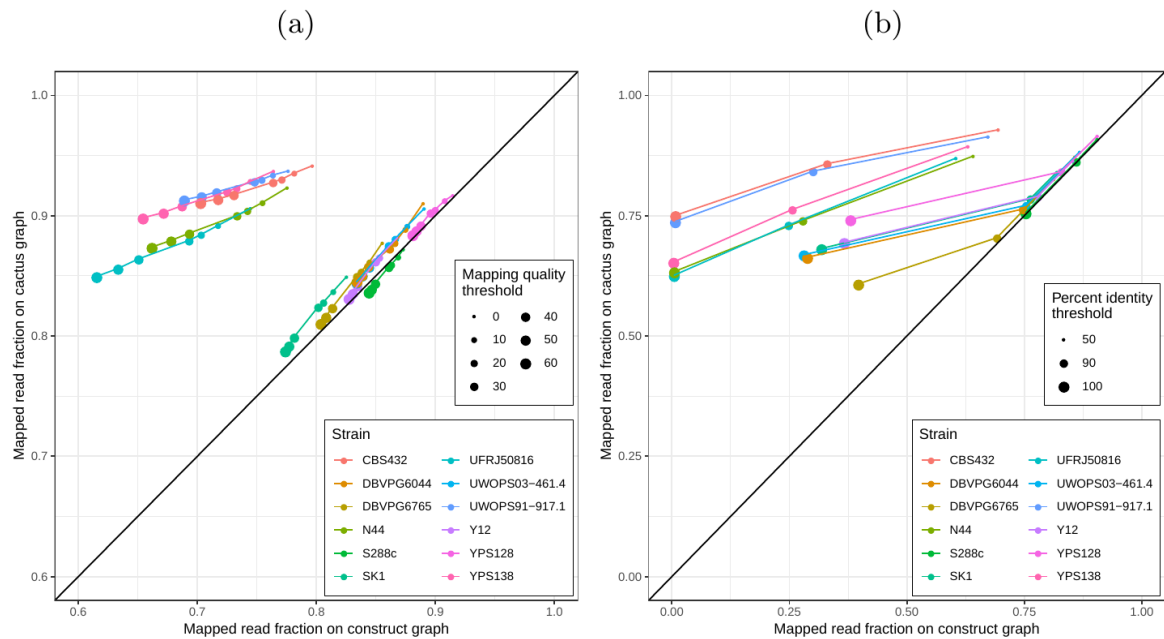


Figure 3: **Mapping comparison.** The fraction of reads mapped to the cactus graph (y-axis) and the construct graph (x-axis) are compared. a) Stratified by mapping quality threshold. b) Stratified by percent identity threshold.

In a first step, we tested our hypothesis that the *cactus graph* has higher mappability due to its better representation of sequence diversity among the yeast strains. Fig. 3a shows the fraction of Illumina reads from the 12 strains that was mapped with a mapping quality above a certain threshold to the *cactus graph* and to the *construct graph*. Generally, more reads were mapped to the *cactus graph* than to the *construct graph* regardless of the chosen mapping quality threshold. Only for the reference strain S.c. S288C, both graphs exhibited similar mappability. This suggests that not the higher sequence content in the *cactus graph* alone (XX Mb compared to XX Mb in the *construct graph*) drives the improvement in mappability. Instead, our measurements suggest that genetic distance to the reference strain increases the advantage of the *cactus graph* over the *construct graph*. Consequently, the benefit of the *cactus graph* is largest for strains in the *S. paradoxus* clade and smaller for reads from strains in the *S. cerevisiae* clade.

When we explored the mapping identity of the short reads on the graphs, we observed a similar trend (see Fig. 3b). For strains in the *S. paradoxus* clade, the *cactus graph* enabled substantially more mappings with high percent identity than the *construct graph*. With strains in the *S. cerevisiae* clade, the difference was smaller, at least for a percent identity threshold up to 90%. When comparing read fractions with perfect identity (i.e. percent identity threshold = 100%), the *cactus graph* clearly outperforms the *construct graph* on 11 out of 12 samples. The only exception again is the reference strain S288C.

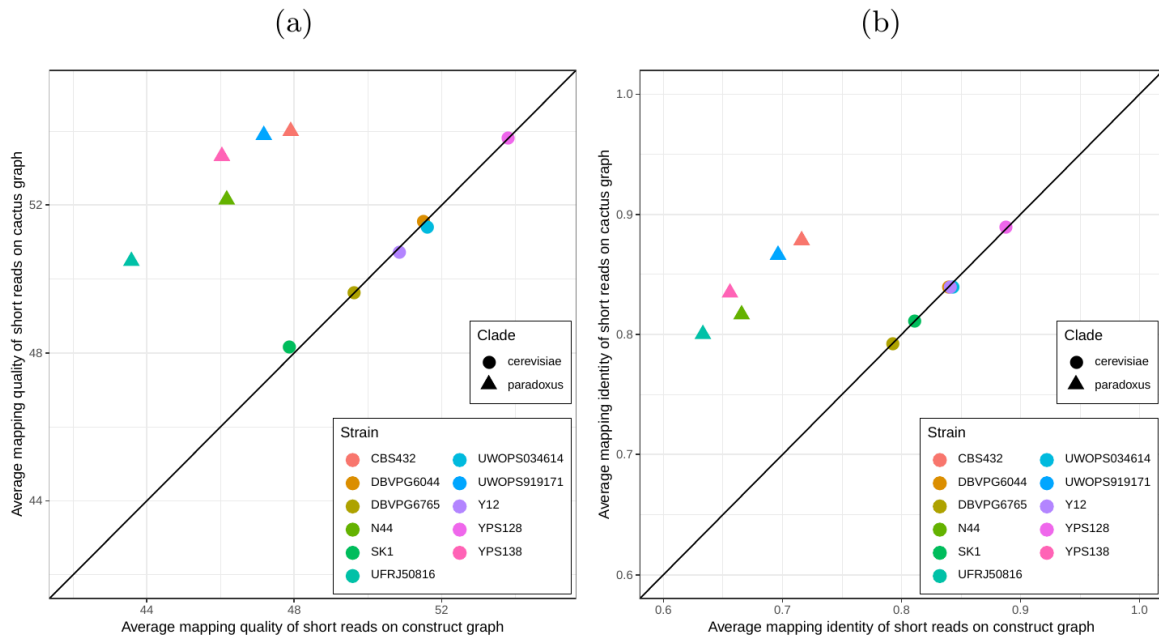


Figure 4: SV genotyping comparison. a) Average mapping quality of short reads mapped to the *cactus graph* (y-axis) and *construct graph* (x-axis). b) Average mapping identity of short reads mapped to the *cactus graph* (y-axis) and *construct graph* (x-axis). Colors and shapes represent the 11 non-reference strains and two clades, respectively

Next, we compared the SV genotype performance of both graphs. We mapped short reads from the 11 non-reference strains to both graphs and called variants using *vg*'s variant calling module. To compare the callsets from both graphs, we generated a sample graph for each callset using the reference genome and the callset. Each sample graph is a graph representation of the respective callset. If a given callset is correct, we would expect that reads from the same sample can be mapped confidently and with high identity to the corresponding sample graph. Therefore, we compared the average mapping identity of the short reads on both types of sample graphs (see Figure 4b). Similar to the results of our mapping analysis above, the *cactus graph* clearly outperformed the *construct graph* for strains in the *S. paradoxus* clade. With strains in the *S. cerevisiae* clade, both graphs were on a par.

This trend was confirmed when we looked at two our measures, average mapping quality and average alignment score (see Figures 4a and S2).

Methods

Discussion

Supplementary Material

Table 3: HGSVC experiment using real reads.

Graph	type	TP	TP.baseline	FP	FN	precision	recall	F1
HGSVC-Construct	Total	18436	18500	6575	8206	0.7378	0.6927	0.7145
	INS	10984	10600	3542	5085	0.7495	0.6758	0.7107
	DEL	7452	7900	3033	3121	0.7226	0.7168	0.7197
HGSVC-1KG-Construct	Total	17802	17946	6221	8760	0.7426	0.672	0.7055
	INS	10647	10262	3304	5423	0.7564	0.6543	0.7017
	DEL	7155	7684	2917	3337	0.7248	0.6972	0.7107
HGSVC-Bayestyper	Total	4342	4840	1048	21866	0.822	0.1812	0.2969
	INS	1786	1883	309	13802	0.859	0.1201	0.2107
	DEL	2556	2957	739	8064	0.8001	0.2683	0.4018
SVPOP-Construct	Total	9091	9931	10235	16775	0.4925	0.3719	0.4238
	INS	6972	7420	6706	8265	0.5253	0.4731	0.4978
	DEL	2119	2511	3529	8510	0.4157	0.2278	0.2943

////////////////////////////////////.

Table 4: HGSVC experiment using real reads and restricting the comparisons to non-repeat regions.

Graph	type	TP	TP.baseline	FP	FN	precision	recall	F1
HGSVC-Construct	Total	5197	5244	854	831	0.86	0.8632	0.8616
	INS	3708	3626	459	498	0.8876	0.8792	0.8834
	DEL	1489	1618	395	333	0.8038	0.8293	0.8164
HGSVC-1KG-Construct	Total	5103	5155	865	920	0.8563	0.8486	0.8524
	INS	3642	3555	464	569	0.8845	0.862	0.8731
	DEL	1461	1600	401	351	0.7996	0.8201	0.8097
HGSVC-Bayestyper	Total	1560	1731	274	4344	0.8633	0.2849	0.4284
	INS	883	901	69	3223	0.9289	0.2185	0.3538
	DEL	677	830	205	1121	0.8019	0.4254	0.5559
SVPOP-Construct	Total	3251	3480	941	2595	0.7872	0.5728	0.6631
	INS	2859	3009	780	1115	0.7941	0.7296	0.7605
	DEL	392	471	161	1480	0.7453	0.2414	0.3647

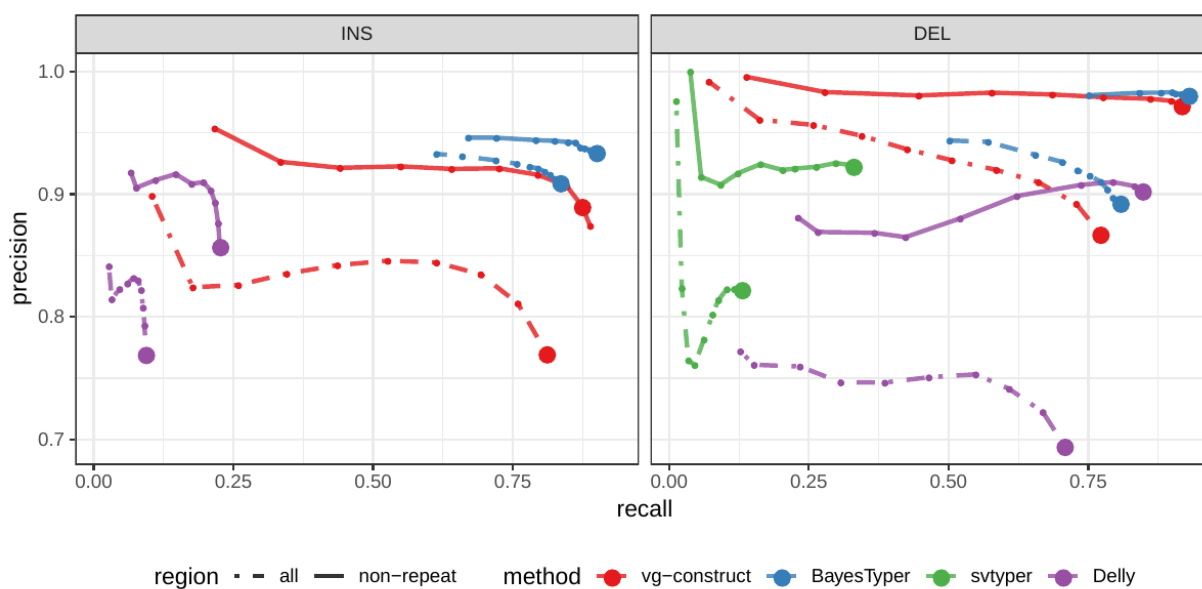


Figure 5: **Structural variants from the HGSVC dataset.** Genotyping evaluation for simulated reads.

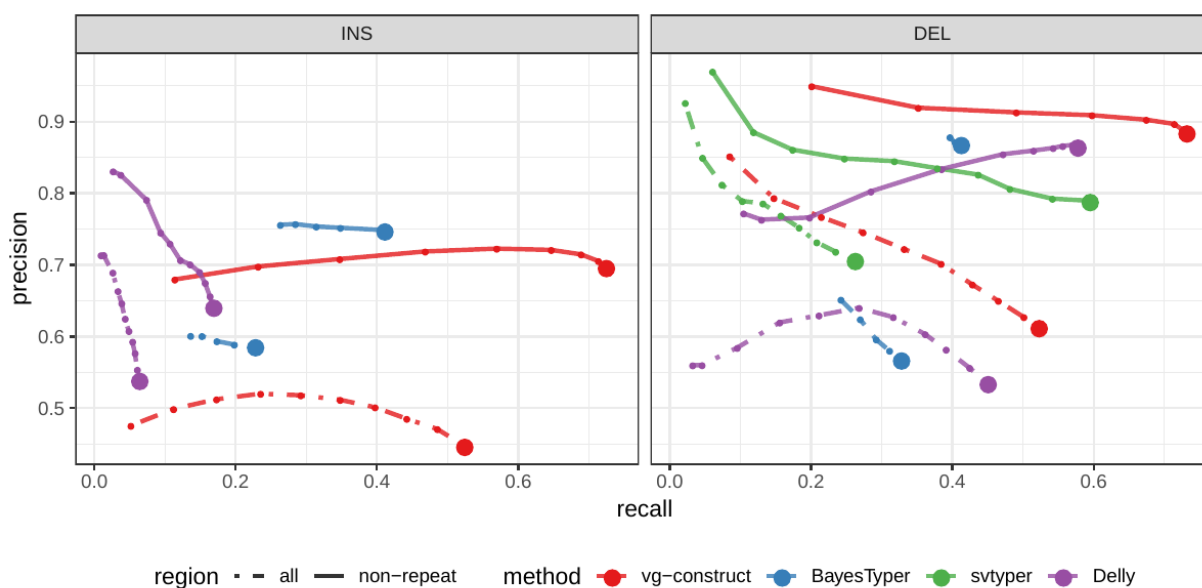


Figure 6: **Structural variants from the HGSVC dataset.** Genotyping evaluation for real reads.

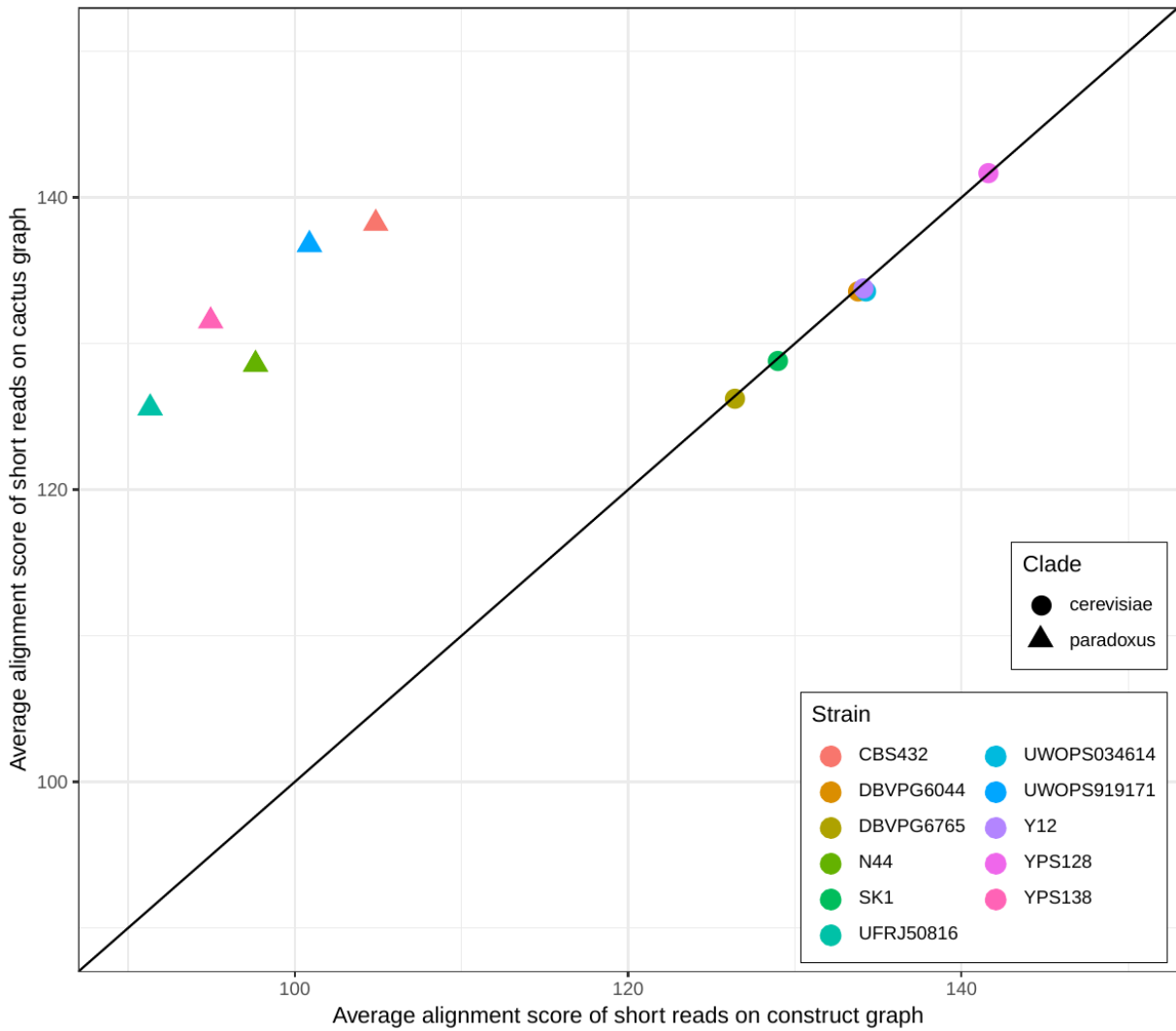


Figure S2: **SV genotyping comparison.** Average alignment score of short reads mapped to the *cactus graph* (y-axis) and *construct graph* (x-axis) is compared. Colors and shapes represent the 11 non-reference strains and two clades, respectively

References

1. An integrated map of structural variation in 2,504 human genomes

Peter H. Sudmant, Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, Kai Ye, Goo Jun, ... Jan O. Korbel

Nature (2015-10) <https://doi.org/73c>

DOI: [10.1038/nature15394](https://doi.org/10.1038/nature15394) · PMID: [26432246](https://pubmed.ncbi.nlm.nih.gov/26432246/) · PMCID: [PMC4617611](https://pubmed.ncbi.nlm.nih.gov/PMC4617611/)

2. Whole-genome sequence variation, population structure and demographic history of the Dutch population

Laurent C Francioli, Androniki Menelaou, Sara L Pulit, Freerk van Dijk, Pier Francesco Palamara, Clara C Elbers, Pieter BT Neerincx, Kai Ye, Victor Guryev, ... Cisca Wijmenga

Nature Genetics (2014-06-29) <https://doi.org/f6bxm8>

DOI: [10.1038/ng.3021](https://doi.org/10.1038/ng.3021) · PMID: [24974849](https://pubmed.ncbi.nlm.nih.gov/24974849/)

3. Resolving the complexity of the human genome using single-molecule sequencing

Mark J. P. Chaisson, John Huddleston, Megan Y. Dennis, Peter H. Sudmant, Maika Malig, Fereydoon Hormozdiari, Francesca Antonacci, Urvashi Surti, Richard Sandstrom, Matthew Boitano, ... Evan E. Eichler

Nature (2014-11-10) <https://doi.org/w69>

DOI: [10.1038/nature13907](https://doi.org/10.1038/nature13907) · PMID: [25383537](https://pubmed.ncbi.nlm.nih.gov/25383537/) · PMCID: [PMC4317254](https://pubmed.ncbi.nlm.nih.gov/PMC4317254/)

4. Discovery and genotyping of structural variation from long-read haploid genome sequence data

John Huddleston, Mark J.P. Chaisson, Karyn Meltz Steinberg, Wes Warren, Kendra Hoekzema, David Gordon, Tina A. Graves-Lindsay, Katherine M. Munson, Zev N. Kronenberg, Laura Vives, ... Evan E. Eichler

Genome Research (2016-11-28) <https://doi.org/f9x79h>

DOI: [10.1101/gr.214007.116](https://doi.org/10.1101/gr.214007.116) · PMID: [27895111](https://pubmed.ncbi.nlm.nih.gov/27895111/) · PMCID: [PMC5411763](https://pubmed.ncbi.nlm.nih.gov/PMC5411763/)

5. Mapping and phasing of structural variation in patient genomes using nanopore sequencing

Mircea Cretu Stancu, Markus J. van Roosmalen, Ivo Renkens, Marleen M. Nieboer, Sjors Middelkamp, Joep de Ligt, Giulia Pregno, Daniela Giachino, Giorgia Mandrile, Jose Espejo Valle-Inclan, ... Wigard P. Kloosterman

Nature Communications (2017-11-06) <https://doi.org/gftpt9>

DOI: [10.1038/s41467-017-01343-4](https://doi.org/10.1038/s41467-017-01343-4) · PMID: [29109544](https://pubmed.ncbi.nlm.nih.gov/29109544/) · PMCID: [PMC5673902](https://pubmed.ncbi.nlm.nih.gov/PMC5673902/)

6. Genome-wide reconstruction of complex structural variants using read clouds

Noah Spies, Ziming Weng, Alex Bishara, Jennifer McDaniel, David Catoe, Justin M Zook, Marc Salit, Robert B West, Serafim Batzoglou, Arend Sidow

Nature Methods (2017-07-17) <https://doi.org/gbnhww>

DOI: [10.1038/nmeth.4366](https://doi.org/10.1038/nmeth.4366) · PMID: [28714986](https://pubmed.ncbi.nlm.nih.gov/28714986/) · PMCID: [PMC5578891](https://pubmed.ncbi.nlm.nih.gov/PMC5578891/)

7. Characterizing the Major Structural Variant Alleles of the Human Genome

Peter A. Audano, Arvis Sulovari, Tina A. Graves-Lindsay, Stuart Cantsilieris, Melanie Sorensen, AnneMarie E. Welch, Max L. Dougherty, Bradley J. Nelson, Ankeeta Shah, Susan K. Dutcher, ... Evan E. Eichler

Cell (2019-01) <https://doi.org/gfthvz>

DOI: [10.1016/j.cell.2018.12.019](https://doi.org/10.1016/j.cell.2018.12.019) · PMID: [30661756](https://pubmed.ncbi.nlm.nih.gov/30661756/)

8. Nanopore sequencing and assembly of a human genome with ultra-long reads

Miten Jain, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, ... Matthew Loose

Nature Biotechnology (2018-01-29) <https://doi.org/gczffw>

DOI: [10.1038/nbt.4060](https://doi.org/10.1038/nbt.4060) · PMID: [29431738](https://pubmed.ncbi.nlm.nih.gov/29431738/) · PMCID: [PMC5889714](https://pubmed.ncbi.nlm.nih.gov/PMC5889714/)

9. Phased diploid genome assembly with single-molecule real-time sequencing

Chen-Shan Chin, Paul Peluso, Fritz J Sedlazeck, Maria Nattestad, Gregory T Concepcion, Alicia Clum, Christopher Dunn, Ronan O'Malley, Rosa Figueroa-Balderas, Abraham Morales-Cruz, ... Michael C Schatz

Nature Methods (2016-10-17) <https://doi.org/f9fv4w>

DOI: [10.1038/nmeth.4035](https://doi.org/10.1038/nmeth.4035) · PMID: [27749838](https://pubmed.ncbi.nlm.nih.gov/27749838/) · PMCID: [PMC5503144](https://pubmed.ncbi.nlm.nih.gov/PMC5503144/)

10. Genome graphs and the evolution of genome inference

Benedict Paten, Adam M. Novak, Jordan M. Eizenga, Erik Garrison

Genome Research (2017-03-30) <https://doi.org/f95nhd>

DOI: [10.1101/gr.214155.116](https://doi.org/10.1101/gr.214155.116) · PMID: [28360232](https://pubmed.ncbi.nlm.nih.gov/28360232/) · PMCID: [PMC5411762](https://pubmed.ncbi.nlm.nih.gov/PMC5411762/)

11. Variation graph toolkit improves read mapping by representing genetic variation in the reference

Erik Garrison, Jouni Sirén, Adam M Novak, Glenn Hickey, Jordan M Eizenga, Eric T Dawson, William Jones, Shilpa Garg, Charles Markello, Michael F Lin, ... Richard Durbin

Nature Biotechnology (2018-08-20) <https://doi.org/gd2zqs>

DOI: [10.1038/nbt.4227](https://doi.org/10.1038/nbt.4227) · PMID: [30125266](https://pubmed.ncbi.nlm.nih.gov/30125266/) · PMCID: [PMC6126949](https://pubmed.ncbi.nlm.nih.gov/PMC6126949/)

12. Fast and accurate genomic analyses using genome graphs

Goran Rakocovic, Vladimir Semenyuk, Wan-Ping Lee, James Spencer, John Browning, Ivan J. Johnson, Vladan Arsenijevic, Jelena Nadj, Kaushik Ghose, Maria C. Suci, ... Deniz Kural

Nature Genetics (2019-01-14) <https://doi.org/gftd46>

DOI: [10.1038/s41588-018-0316-4](https://doi.org/10.1038/s41588-018-0316-4) · PMID: [30643257](https://pubmed.ncbi.nlm.nih.gov/30643257/)

13. GraphTyper enables population-scale genotyping using pangenome graphs

Hannes P Eggertsson, Hakon Jonsson, Snaedis Kristmundsdottir, Eiríkur Hjartarson, Birte Kehr,

Gisli Masson, Florian Zink, Kristjan E Hjorleifsson, Aslaug Jonasdottir, Adalbjorg Jonasdottir, ...
Bjarni V Halldorsson
Nature Genetics (2017-09-25) <https://doi.org/gbx7v6>
DOI: [10.1038/ng.3964](https://doi.org/10.1038/ng.3964) · PMID: [28945251](https://pubmed.ncbi.nlm.nih.gov/28945251/)

14. Accurate genotyping across variant classes and lengths using variant graphs

Jonas Andreas SibbesenLasse Maretty, Anders Krogh
Nature Genetics (2018-06-18) <https://doi.org/gdndnz>
DOI: [10.1038/s41588-018-0145-5](https://doi.org/10.1038/s41588-018-0145-5) · PMID: [29915429](https://pubmed.ncbi.nlm.nih.gov/29915429/)

15. SpeedSeq: ultra-fast personal genome analysis and interpretation

Colby Chiang, Ryan M Layer, Gregory G Faust, Michael R Lindberg, David B Rose, Erik P Garrison, Gabor T Marth, Aaron R Quinlan, Ira M Hall
Nature Methods (2015-08-10) <https://doi.org/gcpgfh>
DOI: [10.1038/nmeth.3505](https://doi.org/10.1038/nmeth.3505) · PMID: [26258291](https://pubmed.ncbi.nlm.nih.gov/26258291/) · PMCID: [PMC4589466](https://pubmed.ncbi.nlm.nih.gov/PMC4589466/)

16. DELLY: structural variant discovery by integrated paired-end and split-read analysis

T. Rausch, T. Zichner, A. Schlattl, A. M. Stutz, V. Benes, J. O. Korbel
Bioinformatics (2012-09-07) <https://doi.org/f38r2c>
DOI: [10.1093/bioinformatics/bts378](https://doi.org/10.1093/bioinformatics/bts378) · PMID: [22962449](https://pubmed.ncbi.nlm.nih.gov/22962449/) · PMCID: [PMC3436805](https://pubmed.ncbi.nlm.nih.gov/PMC3436805/)

17. Multi-platform discovery of haplotype-resolved structural variation in human genomes

Mark J.P. Chaisson, Ashley D. Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J. Gardner, Oscar Rodriguez, Li Guo, Ryan L. Collins, ... Charles Lee
Cold Spring Harbor Laboratory (2017-09-23) <https://doi.org/gftxhc>
DOI: [10.1101/193144](https://doi.org/10.1101/193144)

18. Contrasting evolutionary genome dynamics between domesticated and wild yeasts

Jia-Xing Yue, Jing Li, Louise Aigrain, Johan Hallin, Karl Persson, Karen Oliver, Anders Bergström, Paul Coupland, Jonas Warringer, Marco Cosentino Lagomarsino, ... Gianni Liti
Nature Genetics (2017-04-17) <https://doi.org/f93kpp>
DOI: [10.1038/ng.3847](https://doi.org/10.1038/ng.3847) · PMID: [28416820](https://pubmed.ncbi.nlm.nih.gov/28416820/) · PMCID: [PMC5446901](https://pubmed.ncbi.nlm.nih.gov/PMC5446901/)

19. Assemblytics: a web analytics tool for the detection of variants from an assembly

Maria Nattestad, Michael C. Schatz
Bioinformatics (2016-06-17) <https://doi.org/f9c485>
DOI: [10.1093/bioinformatics/btw369](https://doi.org/10.1093/bioinformatics/btw369) · PMID: [27318204](https://pubmed.ncbi.nlm.nih.gov/27318204/) · PMCID: [PMC6191160](https://pubmed.ncbi.nlm.nih.gov/PMC6191160/)

20. Discovery, genotyping and characterization of structural variation and novel sequence at single nucleotide resolution from de novo genome assemblies on a population scale

Siyang LiuShujia Huang, Junhua Rao, Weijian Ye, Anders Krogh, Jun Wang
GigaScience (2015-12) <https://doi.org/f75r4n>
DOI: [10.1186/s13742-015-0103-4](https://doi.org/10.1186/s13742-015-0103-4) · PMID: [26705468](https://pubmed.ncbi.nlm.nih.gov/26705468/) · PMCID: [PMC4690232](https://pubmed.ncbi.nlm.nih.gov/PMC4690232/)

21. Minimap2: pairwise alignment for nucleotide sequences

Heng Li

Bioinformatics (2018-05-10) <https://doi.org/gdhibqt>

DOI: [10.1093/bioinformatics/bty191](https://doi.org/10.1093/bioinformatics/bty191) · PMID: [29750242](https://pubmed.ncbi.nlm.nih.gov/29750242/) · PMCID: [PMC6137996](https://pubmed.ncbi.nlm.nih.gov/PMC6137996/)

22. Cactus: Algorithms for genome multiple sequence alignment

B. Paten, D. Earl, N. Nguyen, M. Diekhans, D. Zerbino, D. Haussler

Genome Research (2011-06-10) <https://doi.org/bk4697>

DOI: [10.1101/gr.123356.111](https://doi.org/10.1101/gr.123356.111) · PMID: [21665927](https://pubmed.ncbi.nlm.nih.gov/21665927/) · PMCID: [PMC3166836](https://pubmed.ncbi.nlm.nih.gov/PMC3166836/)