









# Genotyping structural variation in variation graphs with the vg toolkit

This manuscript ([permalink](#)) was automatically generated from [jmonlong/manu-vgs@4cf1e74](#) on March 14, 2019.

## Authors

---

 Glenn Hickey<sup>1, </sup>,  David Heller<sup>1, </sup>,  Jean Monlong<sup>1, </sup>,  Adam Novak<sup>1</sup>,  Benedict Paten<sup>1, †</sup>

 — These authors contributed equally to this work

<sup>†</sup> — To whom correspondence should be addressed: [bpaten@ucsc.edu](mailto:bpaten@ucsc.edu)

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

# Abstract

---

Variation graphs have been proposed to represent human genomes and improve on different aspects of genomics analysis, such as read mapping and variant calling. Structural variants (SVs) have been associated with diseases but they remain under-studied due to their complexity and certain technological challenges. Still, thousands of SVs have been characterized and better catalogs are now being generated thanks to new technologies. There is now an opportunity to integrate previously neglected SVs into the unified framework of variation graphs. We first show that the vg toolkit is capable of genotyping insertions, deletions and inversions, even in the presence of small errors in the variant definition. We then benchmarked vg across three high-quality sequence-resolved SV catalogs generated by recent studies. vg was compared to state-of-the-art SV genotypers using simulated and real Illumina short reads. On real data, vg produced the best genotype predictions systematically in all datasets. Finally, we found that better graphs could be constructed directly from de novo assembly alignment. We experimented with assemblies from 12 yeast strains and showed that SV genotyping was improved compared to graphs built from intermediate SV catalogs in the VCF format. Our results demonstrate the power of variation graphs for SV genotyping. Beyond single nucleotide variants and short insertions/deletions, the vg toolkit now includes SV in its unified variant calling framework and provides a natural solution to integrate high-quality SV catalogs and assemblies.

# 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. SVs have long been associated with developmental disorders, cancer and other complex diseases and phenotypes[1]. However, SVs have been under-studied for technological reasons and due to their complexity as compared to other types of genomic variation. Although whole-genome sequencing (WGS) made it possible to assess virtually any type of structural variation, many challenges remain. SV-supporting reads are generally difficult to map to reference genomes, in part because most SVs are larger than the sequencing reads. Repeated sequences in the genome often confuse read mapping algorithms, which can produce mappings that seem to support an SV. In practice, large-scale projects had to combine several methods to achieve better accuracy. This methodology has been used to compile catalogs with tens of thousands of SVs in humans[2,3]. Overall, SV detection from short-read sequencing remains laborious and of lower accuracy than small variant detection. This explains 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 have made it possible to produce long reads or retrieve long-range information over kilobases of sequence. This is particularly useful for SV detection and de novo assembly. Several recent studies using long-read or linked-read sequencing have produced large catalogs of structural variation, the majority of which was novel and sequence-resolved[4,5,6,7,8]. These technologies are also enabling the production of high-quality de novo genome assemblies[4,9], and large blocks of haplotype-resolved sequences[10]. Such technical advances promise to expand the amount of known genomic variation in humans in the near future. However, their cost prohibits their use in large-scale studies that require hundreds or thousands of samples, such as disease association studies.

At the same time, the reference genome is evolving from a linear reference to a graph-based reference that incorporates known genomic variation[11,12,13]. By including variants in the graph, both read mapping and variant calling become variant-aware and benefit in term of accuracy and sensitivity[12]. In addition, different variant types are called simultaneously by a unified framework. vg was the first openly available tool that scaled to multi-gigabase genomes and provides read mapping, variant calling and haplotype modeling[12]. In vg, graphs can be built from both variant catalogs in the VCF format or assembly alignment. Other genome graph implementations have also been used specifically to genotype variants. Using a sliding-window approach, GraphTyper realigns reads to a graph build from known SNVs and short indels[14]. BayesTyper build graphs of both short variants and SVs, and genotypes variants based on the khmer distribution of sequencing reads[15]. Here again, the graph-based approaches showed clear advantages over standard methods that use the linear reference.

Other SV genotyping approaches typically compare read mapping to the reference genome and to a sequence modified with the SV. For example SMRT-SV was designed to genotype SVs identified on PacBio reads[5]. SVTyper uses paired-end mapping and split-read mapping information and can genotype deletions, duplications, inversions and translocations[16]. Delly provides a genotyping feature in addition to its discovery mode and can genotype all types of SVs although the VCF needs special formatting for some[17]. SMRT-SV2 is a machine learning tool that was trained to genotype SVs from the alignment of read to the reference genome augmented with SV-containing sequences as alternate contigs[8].

We show that the unified variant calling framework implemented in vg is capable of genotyping deletions, insertions and inversions. We compare vg with state-of-the-art SV genotypers: SVTyper[16], Delly[17], BayesTyper[15] and SMRT-SV2[8]. On simulation, vg is robust to small errors in the breakpoint location and outperforms most other methods on shallow sequencing experiments. Starting from SVs discovered in recent long-read sequencing studies[18,19,20,8], we evaluated the genotyping accuracy when using simulated or real Illumina reads. Across all three datasets that we tested, vg is the best performing SV genotyper on real short-read data for all SV types and sizes. Going further, we show that building graphs from the alignment of de novo assemblies leads to better genotyping performance.

## Results

---

### Structural variation in vg

In addition to SNV and short indels, vg can handle large deletions, insertions and inversions (Figure 1a). As a proof-of-concept we simulated genomes and different types of SVs with a size distribution matching real SVs[18]. We compared vg against SVTyper, Delly and BayesTyper across different level of sequencing depth. Some errors were also added at the breakpoints to investigate their effect on genotyping (see Methods). The results are shown in Figure 1b. When using the correct breakpoints, vg tied with Delly as the best genotyper for deletions, and with BayesTyper as the best genotyper for insertions. For inversions, vg was the second best genotyper after BayesTyper. The differences between the methods were the most visible at lower sequencing depth. In the presence of 1-10 bp errors in the breakpoint location, the performance of Delly and BayesTyper dropped significantly. The dramatic drop for BayesTyper can be explained by its khmer-based approach that requires exact SV definition. In contrast, vg was only slightly affected by the presence of errors in the input VCF (Figure 1b). For vg, the F1 scores for all SV types decreased no more than of 0.07 point. Overall, these results show that vg is capable of genotyping SVs and is robust to errors in the input VCF.

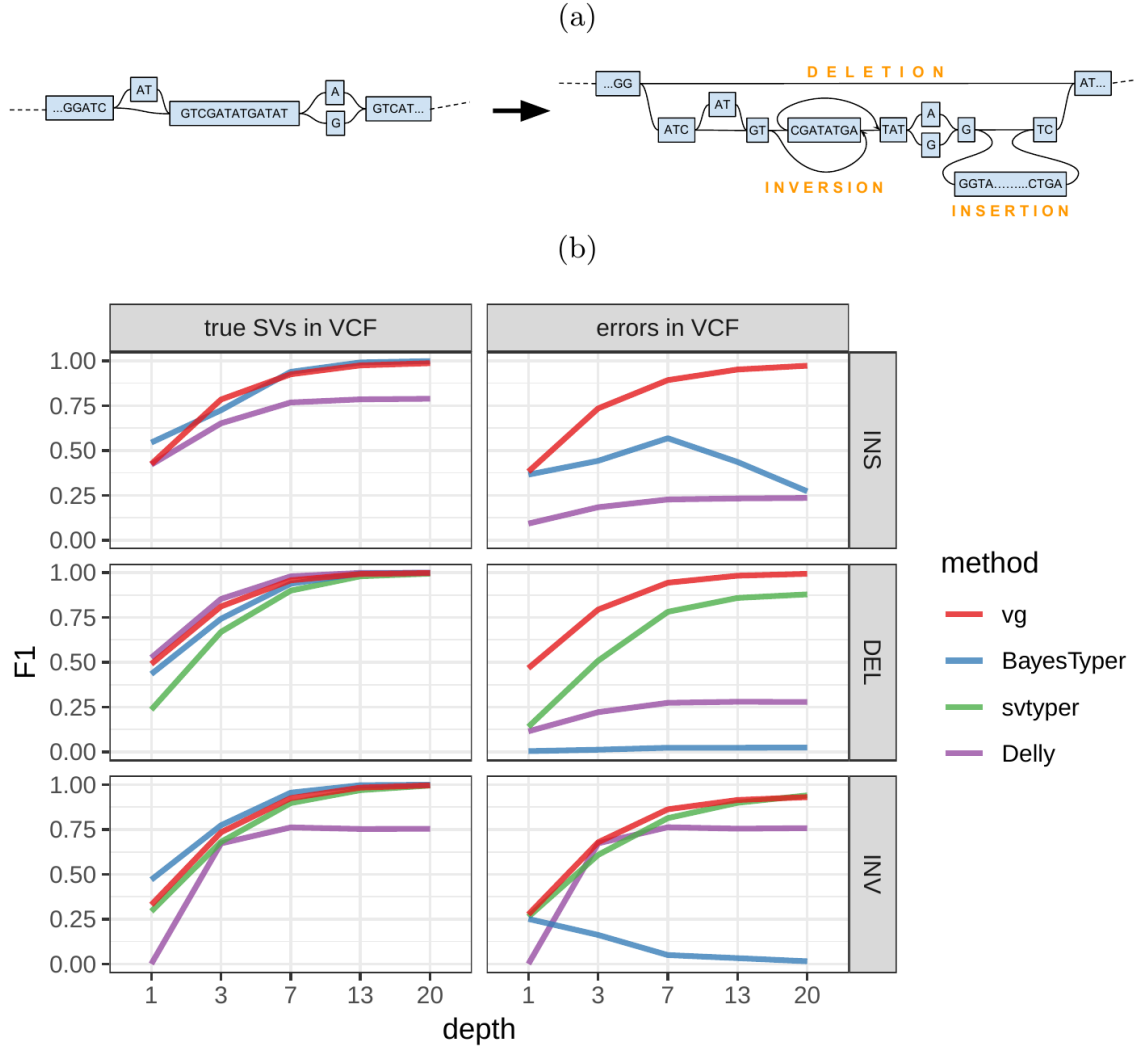


Figure 1: **Structural variation in vg.** a) Adding large deletions and insertions in a variation graph. b) Simulation experiment. For each experiment (method, depth and input VCF with/without errors), the maximum F1 was picked when using different quality thresholds, and is reported on the y-axis.

## HGSVC dataset

The Human Genome Structural Variation Consortium (HGSVC) generated a high-quality SV catalog of three samples, obtained using a consensus from different sequencing, phasing and variant calling technologies[18]. The three samples come from different human populations: a han Chinese individual (HG00514), a Puerto-Rican individual (HG00733), and a Yoruban Nigerian individual (NA19240). These SVs were used to construct a graph with vg and as input for the other genotypers. SVs were genotyped from short reads and compared with the original catalog (see [Methods](#)).

First, by simulating reads for HG00514, we compared the different methods in the ideal situation where the SV catalog is correct and matches exactly the SVs supported by the reads. While vg outperformed Delly and SVTyper, BayesTyper showed the best F1 score and precision-recall trade-

off (Figures 2 and S1, Table S1). When restricting the comparisons to regions not identified as tandem repeats or segmental duplications, the genotyping predictions were significantly better for all methods, with vg almost as good as BayesTyper on deletions (F1 of 0.944 vs 0.955). We observed similar results when evaluating the absence/presence of a SV instead of the exact genotype (Figures 2 and S2). Overall, both graph-based methods, vg and BayesTyper, outperformed the two other methods tested.

We then repeated the analysis using real Illumina reads from HG00514, to benchmark the methods on a more realistic experiment. Here vg clearly outperformed other approach, most likely because of its graph-based strategy and robustness to errors in the SV catalog (Figures 2 and S3). In non-repeat regions and across the whole genome, the F1 scores and precision-recall curves were higher for vg compared to other methods. For example, for deletions in non-repeat regions, the F1 score for vg was 0.801 while the second best method, Delly, had a F1 score of 0.692. We observed similar results when evaluating the absence/presence of a SV instead of the exact genotype (Figures 2 and S4).

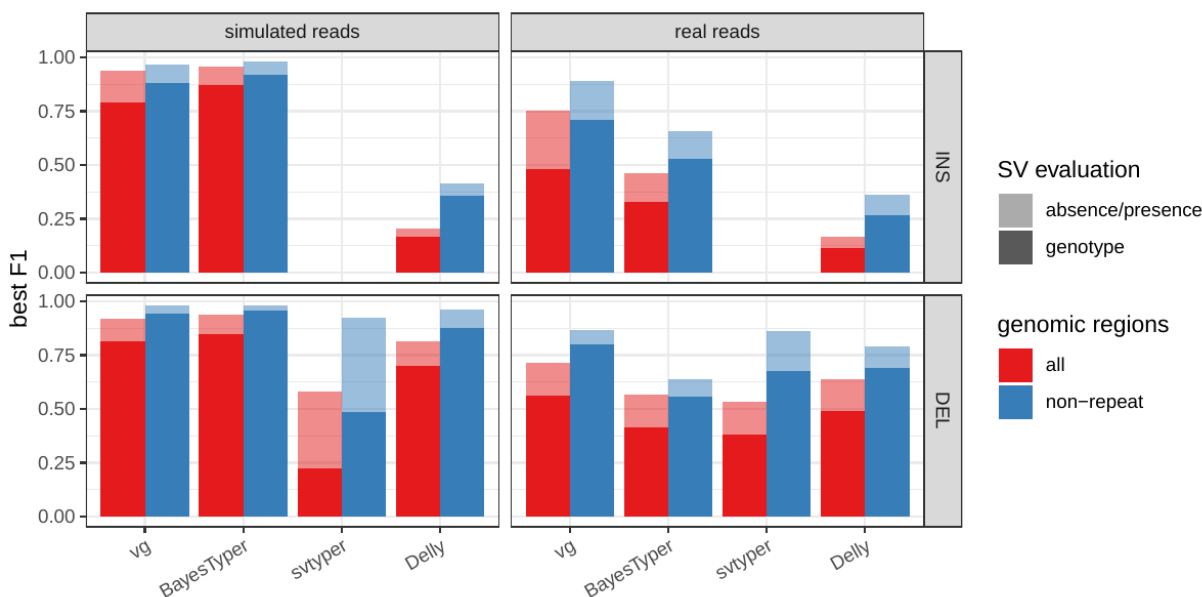


Figure 2: **Structural variants from the HGSVC dataset.** Simulated and real reads from HG00514 were used to genotype SVs and compared with the high-quality calls from Chaisson et al.[18]. Maximum F1 score for each method, across the whole genome (red) or focusing on non-repeat regions (blue). The calling and genotyping evaluation are shown with different shapes.

## Other long-read datasets

The Genome in a Bottle (GiaB) consortium is currently producing a high-quality SV catalog for a Ashkenazim individual (HG002)[19,20]. Dozens of SV callers and datasets from short, long and linked reads were used to produce this set of SVs. vg performed similarly on this dataset than in the HGSVC dataset, with a F1 score of XX and XX for insertions and deletions respectively (FIG). As before, other methods produced lower F1 scores and precision-recall curves (FIG).

A recent study by Audano et al. generated a SV catalog using long-read sequencing across 15 individuals [8]. These variants were then genotyped from short reads across 440 individuals using SMRT-SV2, a machine-learning genotyper implemented for this study. We first called SVs from the pseudo-diploid genome and reads used to train SMRT-SV2 and constructed by merging datasets from two haploid cell lines[8]. Although no false-positives were predicted by SMRT-SV2, the higher recall for vg resulted in a higher F1 score (e.g. 0.809 vs 0.726 across the whole genome, see Table S2). Using publicly available Illumina reads, we then genotyped SVs in one of the 15 individuals that was used for discovery in Audano et al.[8]. Compared to SMRT-SV2, vg had a better precision-recall curve and a higher F1 for both insertions and deletions (FIG and Table S3). The overall F1 score for vg was 0.574 versus XX for SMRT-SV2. Of note, Audano et al. had identified 27 sequence-resolved inversions, 7 of which were predicted correctly by vg. Inversions are often complex, harboring additional variation that makes their characterization and genotyping challenging.

## 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*) [21]. 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 [22], AsmVar [23] and paftools [24]. 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 [25]. While the *construct graph* is still mainly linear and highly dependent on the reference genome, the *cactus graph* is completely unbiased in that regard.

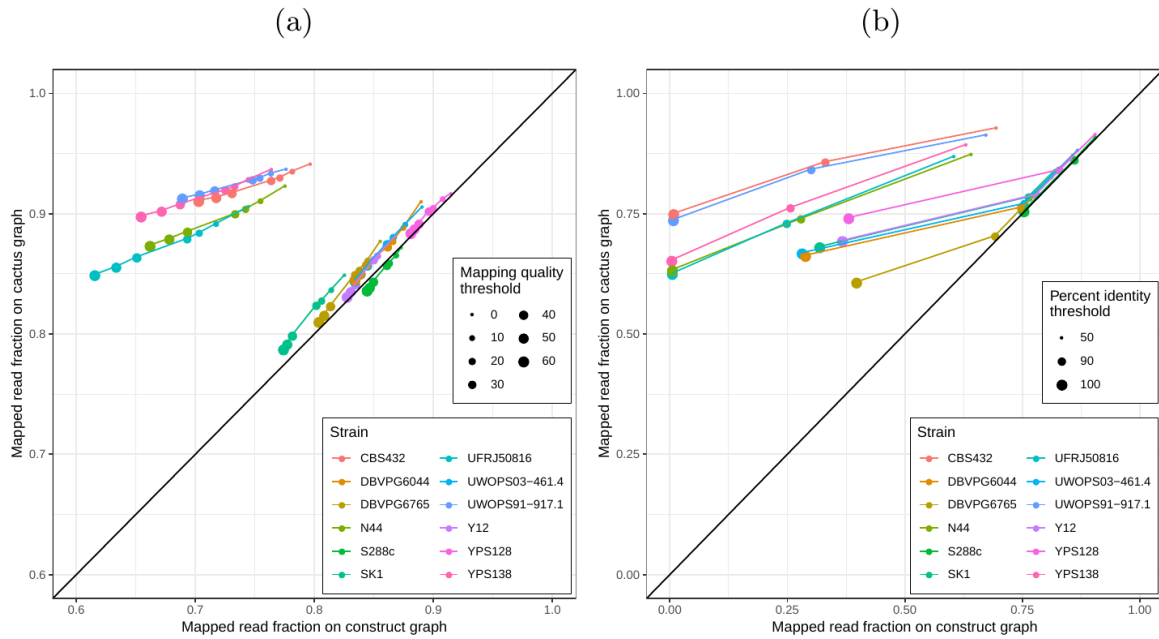
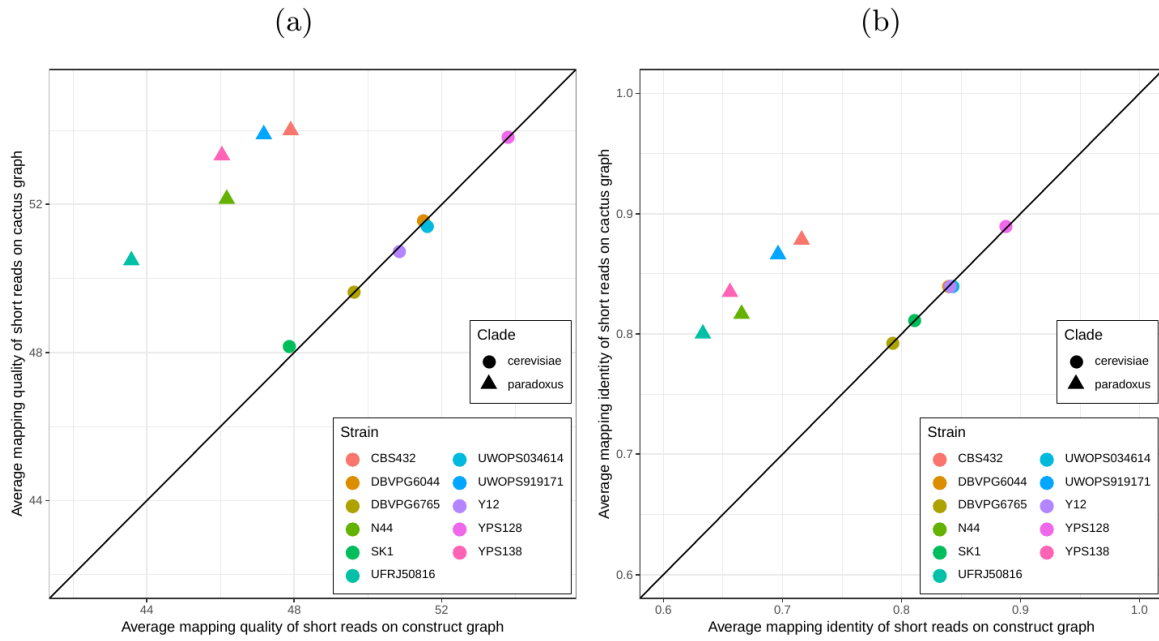


Figure 3: **Mapping comparison.** The fraction of reads mapped to the cactus graph (y-axis) and the construct graph (x-axis) are compared. a) Stratified by mapping quality threshold. b) Stratified by percent identity threshold.

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. Figure 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 Figure 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 S6).

# Methods

---

## Simulation experiment

### HGSVC

#### (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.

When restricting the comparisons to regions not identified as tandem repeats or segmental duplications in the Genome Browser (Table [S1](#)).

#### (Whole-genome) Real reads

Illumina reads were downloaded from ...

## De novo assembly alignment from 12 yeast strains

# Discussion

---

*Potential topics for the discussion.*

### Providing a resource to be used by large-scale sequencing project

As a result of this study we provide a variation graph containing XX millions of SNVs and indels from the 1000 Genomes Project as well as XX thousands of SVs derived from long-read sequencing. This variation graph could serve as a richer reference for large scale projects that use short-read sequencing. For instance, more and more large-scale projects are sequencing the genomes of thousands or hundreds of thousands of individuals, e.g the Pancancer Analysis of Whole Genomes, the Genomics England initiative, and the TOPMed consortium(REFS). These large WGS studies will provide a deeper look into the mechanism of common diseases and, in some cases, will be used directly in a clinical setting. Clinicians and researchers are eager to use these growing WGS resources to interrogate the importance of SVs in disease at a scale never achieved before, either to get a more complete picture of the genetic factors of a disease or to produce a more comprehensive clinical report. As sequencing reaches the clinic, whole-genome sequencing will become routine for many patients. Clinicians will rely on variant calling and

interpretation for diagnosis and treatment. For variant interpretation in particular, a comprehensive and unified characterization of the genomic variation will be extremely valuable.

## **Easier to use**

Some methods require additional information or special VCF formatting [26]. SVTyper was designed to use VCFs created by Lumpy. The genotyping module from Delly was implemented for variants found by its discovery module. SMRT-SV requires a VCF with information about XXX. Nebula, a new khmer-based genotyper, requires reads from a sample containing the SV during khmer selection[27]. In contrast, vg can take as input either explicit or symbolic VCFs, as well as assembly alignment.

## **Assemblies are the future**

Our results suggest that constructing a graph from de novo assembly alignment is more representative of the sequencing reads and leads to better SV genotyping. De novo assemblies for human are becoming more and more common, for example from optimized mate-pair libraries[28] or long-read sequencing[28]. For an optimal representation of the genomic variation, we expect the future graphs to include information from the alignment of numerous de novo assemblies. Aligning assembled contigs to existing variation graphs, like to ones created from SVs catalogs, is still experimental but could generate a genome graph augmented with both existing variant databases and new high-quality assemblies.

## **Future improvements in vg**

The vg toolkit is in active development. Read mapping is an area of constant improvement, both in term of computational efficiency and accuracy. For example, haplotype information can be modeled in variation graph and, in the future, assist read mapping and variant calling. These upcoming developments will directly benefit SV genotyping with vg.

## **Limitations**

Copy number variants (CNVs) are currently represented as deletions or insertions. For this reason duplications are represented as additional sequence rather than encoded as a loop in the graph. While this is sufficient to represent single copy changes, such as deletions or single tandem duplications, CNVs with multi-copies states are not addressed by the current implementation. The genotyping algorithm would need to be extended to model copy number in order to assess these variants.

Near-breakpoint resolution is necessary. Simulations have shown that SV genotyping with vg is robust to errors up to 10 bp in breakpoint location.

The genotyping evaluation of inversions is limited by the lack of existing gold-standards. We showed that vg is capable of genotyping simple inversions from simulation or the few discovered in the SV catalog from Audano et al.[8]. However most inversions are complex and involve small insertions/deletions around their breakpoints(REF). While these complex variants are difficult to represent in the VCF format, they would naturally be represented through the alignment of de novo assemblies. For example, in our experiment with yeast assemblies, we identified XX variants that can be considered complex inversion as they contain at least XX inverted bases.

## Supplementary Material

---

Table S1: HGSVC experiment. Precision, recall and F1 score for the call set with the best F1 score.

Experiment	Method	Region	Type	Precision	Recall	F1
Simulated reads	vg-construct	all	INS	0.770	0.810	0.790
			DEL	0.868	0.771	0.817
		non-repeat	INS	0.891	0.873	0.882
			DEL	0.973	0.917	0.944
	BayesTyper	all	INS	0.910	0.835	0.871
			DEL	0.893	0.807	0.848
		non-repeat	INS	0.935	0.899	0.917
			DEL	0.981	0.930	0.955
	svtyper	all	DEL	0.823	0.130	0.224
		non-repeat	DEL	0.923	0.329	0.486
	Delly	all	INS	0.770	0.093	0.166
			DEL	0.695	0.707	0.701
		non-repeat	INS	0.858	0.226	0.357
			DEL	0.903	0.847	0.874
Real reads	vg-construct	all	INS	0.448	0.523	0.482
			DEL	0.614	0.522	0.564
		non-repeat	INS	0.697	0.724	0.710
			DEL	0.885	0.731	0.801
	BayesTyper	all	INS	0.587	0.227	0.327
			DEL	0.568	0.327	0.415
		non-repeat	INS	0.748	0.410	0.530
			DEL	0.869	0.412	0.559
	svtyper	all	DEL	0.707	0.262	0.382

Experiment	Method	Region	Type	Precision	Recall	F1
		non-repeat	DEL	0.790	0.594	0.678
	Delly	all	INS	0.540	0.063	0.113
			DEL	0.535	0.450	0.489
		non-repeat	INS	0.642	0.168	0.266
			DEL	0.866	0.577	0.692

////////////////////////////////////

Table S2: SVPOP experiment. Calling evaluation for the pseudo-diploid genome built from CHM cell lines in Audano et al.[8].

Method	Region	Type	TP	FP	FN	Precision	Recall	F1
vg-construct	all	Total	9,415	1,369	3,085	0.873	0.753	0.809
		INS	5,789	1,192	1,467	0.829	0.798	0.813
		DEL	3,626	177	1,618	0.954	0.692	0.802
	non-repeat	Total	2,966	273	620	0.919	0.827	0.871
		INS	2,148	270	501	0.894	0.811	0.851
		DEL	818	3	119	0.996	0.873	0.931
SMRT-SV2	all	Total	9,224	0	6,956	1.000	0.570	0.726
		INS	5,322	0	3,631	1.000	0.594	0.746
		DEL	3,902	0	3,325	1.000	0.540	0.701
	non-repeat	Total	2,816	0	1,914	1.000	0.595	0.746
		INS	2,020	0	1,306	1.000	0.607	0.756
		DEL	796	0	608	1.000	0.567	0.724

Table S3: SVPOP experiment. Calling evaluation in the HG5014 individual, one of the 15 individuals used to generate the high-quality SV catalog in Audano et al.[8].

Method	Region	Type	TP	FP	FN	Precision	Recall	F1
vg-construct	all	Total	13,630	9,897	10,376	0.580	0.568	0.574
		INS	8,867	7,850	5,172	0.533	0.632	0.578
		DEL	4,752	2,039	5,150	0.697	0.480	0.568
		INV	11	8	54	0.579	0.169	0.262
	non-repeat	Total	3,864	1,165	855	0.768	0.819	0.792
		INS	2,677	985	514	0.731	0.839	0.781
		DEL	1,180	176	321	0.869	0.786	0.825
		INV	7	4	20	0.636	0.259	0.368

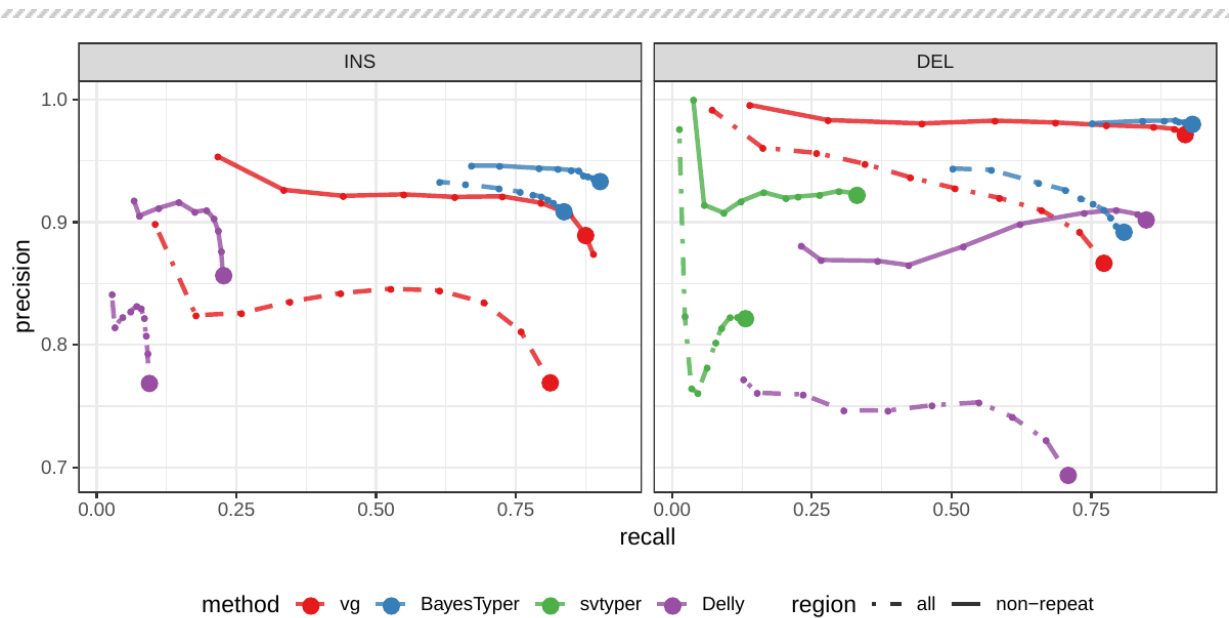


Figure S1: **Structural variants from the HGSVC dataset.** Genotyping evaluation for simulated reads.

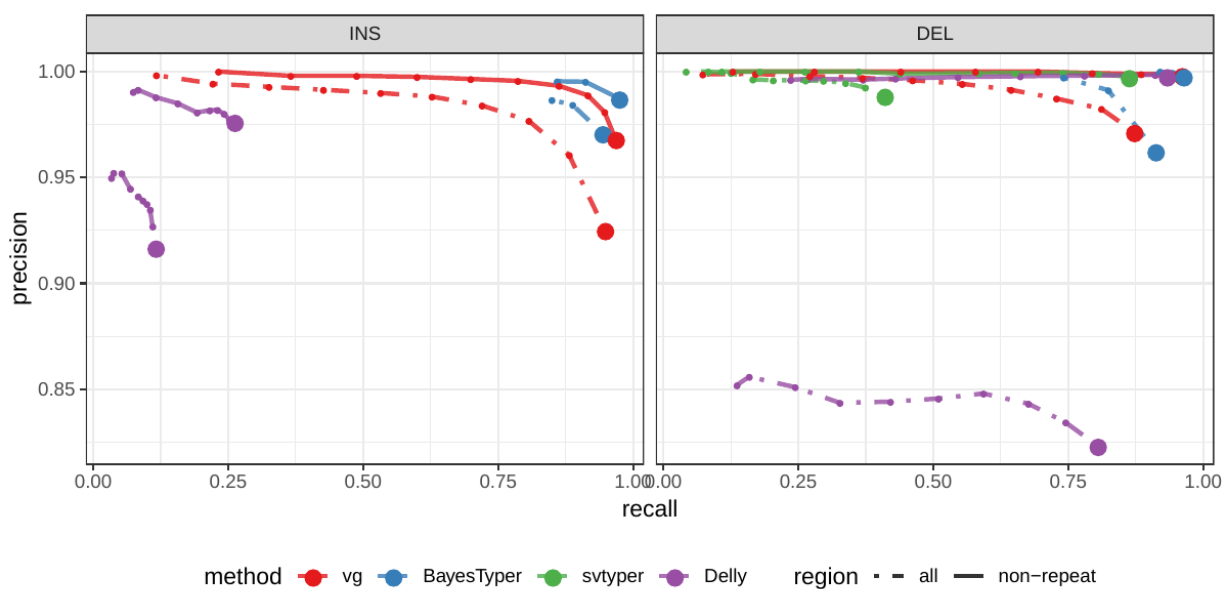


Figure S2: **Structural variants from the HGSVC dataset.** Calling evaluation for simulated reads.

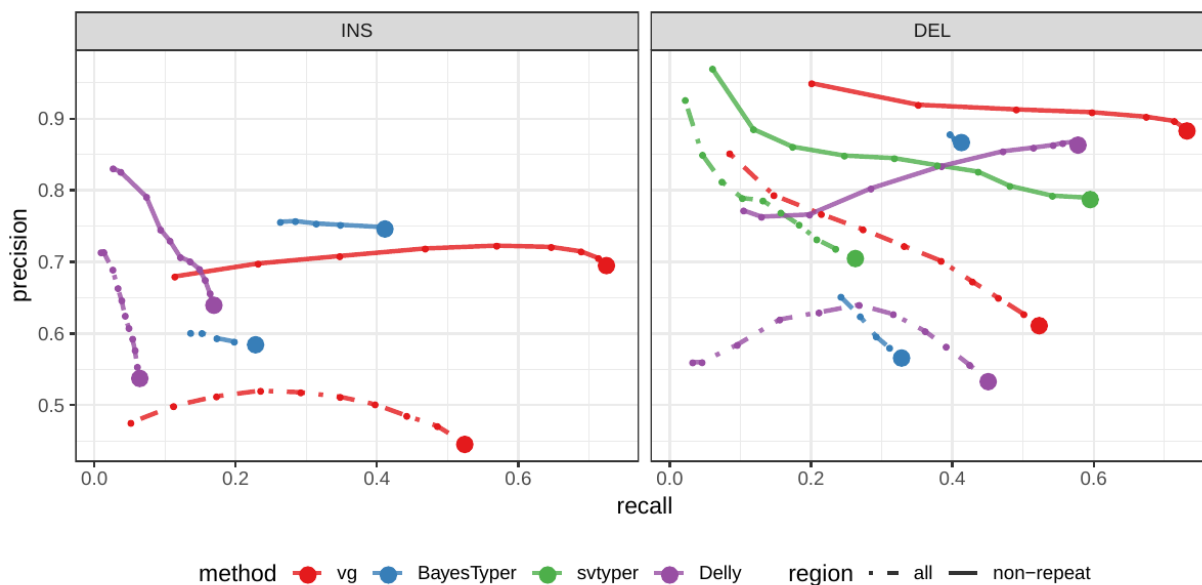


Figure S3: **Structural variants from the HGSVC dataset.** Genotyping evaluation for real reads.



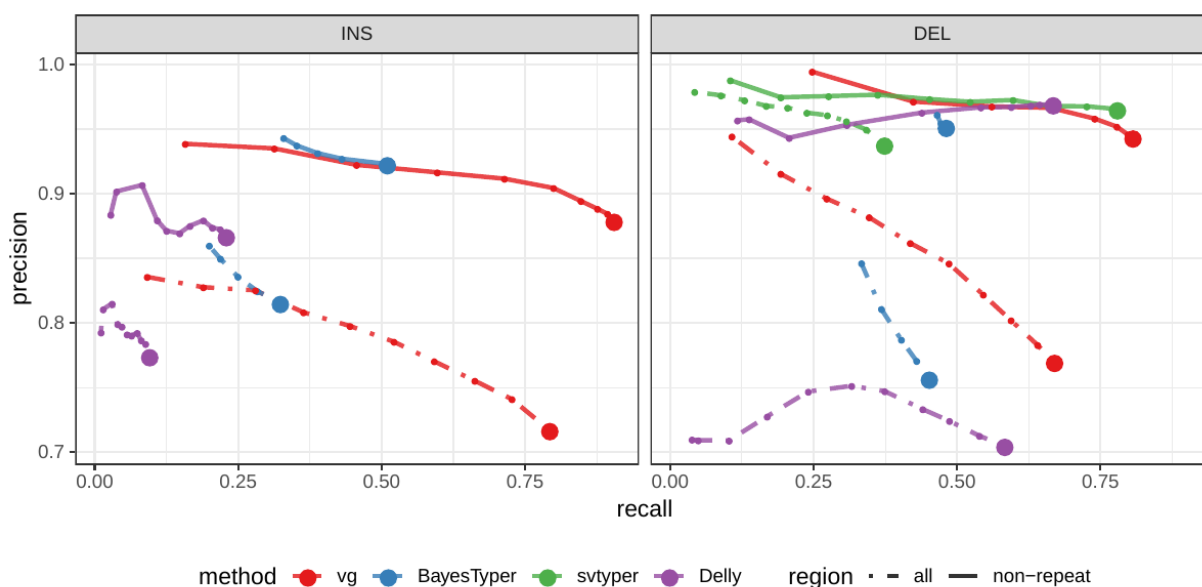


Figure S4: **Structural variants from the HGSVC dataset.** Calling evaluation for real reads.

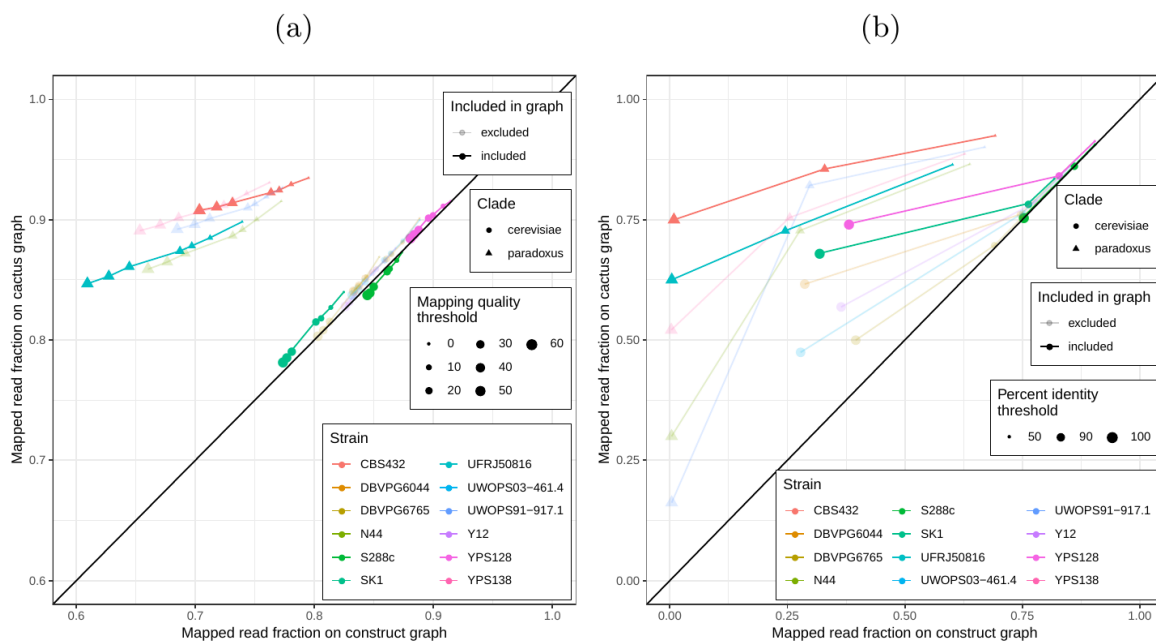


Figure S5: **Mapping comparison on graphs of five strains.** The fraction of reads mapped to the cactus graph (y-axis) and the construct graph (x-axis) are compared. a) Stratified by mapping quality threshold. b) Stratified by percent identity threshold.

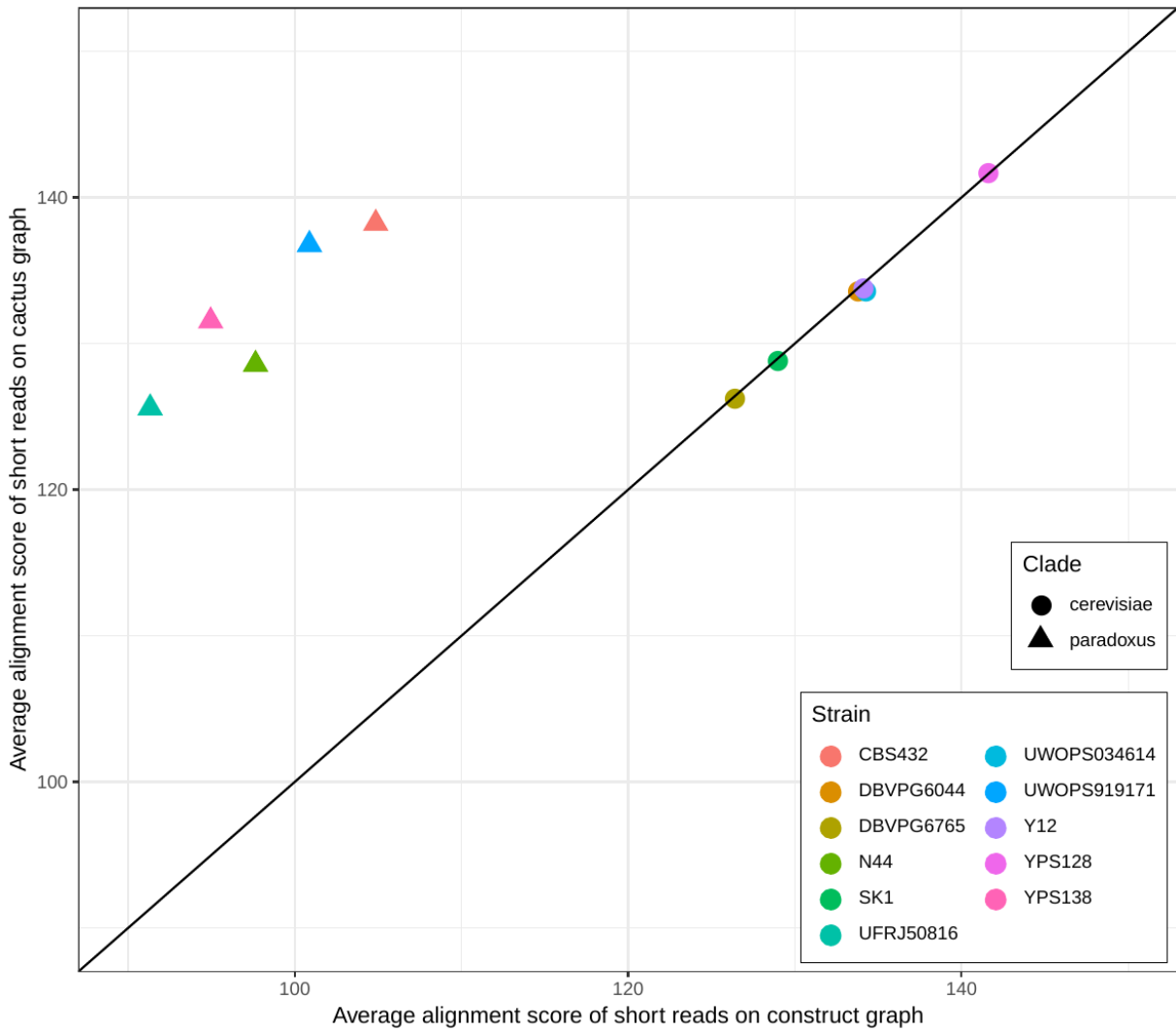


Figure S6: **SV genotyping comparison.** Average alignment score of short reads mapped to the *cactus graph* (y-axis) and *construct graph* (x-axis) is compared. Colors and shapes represent the 11 non-reference strains and two clades, respectively

# References

---

## 1. Phenotypic impact of genomic structural variation: insights from and for human disease

Joachim Weischenfeldt, Orsolya Symmons, François Spitz, Jan O. Korbel

*Nature Reviews Genetics* (2013-02) <https://doi.org/f4nhxh>

DOI: [10.1038/nrg3373](https://doi.org/10.1038/nrg3373) · PMID: [23329113](https://pubmed.ncbi.nlm.nih.gov/23329113/)

## 2. 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/)

## 3. 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/)

## 4. 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, Fereydoun 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/)

## 5. 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/)

## 6. 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/)

#### **7. 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/gbnhkw>

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/)

#### **8. 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/)

#### **9. 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/)

#### **10. 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/)

#### **11. 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/)

#### **12. 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/)

#### **13. 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/)

**14. 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/)

**15. 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/)

**16. 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/)

**17. 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/)

**18. 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)

**19. Extensive sequencing of seven human genomes to characterize benchmark reference materials**

Justin M. Zook, David Catoe, Jennifer McDaniel, Lindsay Vang, Noah Spies, Arend Sidow, Ziming Weng, Yuling Liu, Christopher E. Mason, Noah Alexander, ... Marc Salit  
*Scientific Data* (2016-06-07) <https://doi.org/f84nqc>  
DOI: [10.1038/sdata.2016.25](https://doi.org/10.1038/sdata.2016.25) · PMID: [27271295](https://pubmed.ncbi.nlm.nih.gov/27271295/) · PMCID: [PMC4896128](https://pubmed.ncbi.nlm.nih.gov/PMC4896128/)

**20. Reproducible integration of multiple sequencing datasets to form high-confidence SNP, indel, and reference calls for five human genome reference materials**

Justin Zook, Jennifer McDaniel, Hemang Parikh, Haynes Heaton, Sean A Irvine, Len Trigg,

Rebecca Truty, Cory Y McLean, Francisco M De La Vega, Chunlin Xiao, ...  
*Cold Spring Harbor Laboratory* (2018-03-13) <https://doi.org/gfwsmj>  
DOI: [10.1101/281006](https://doi.org/10.1101/281006)

**21. 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/)

**22. 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/)

**23. 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/)

**24. Minimap2: pairwise alignment for nucleotide sequences**

Heng Li  
*Bioinformatics* (2018-05-10) <https://doi.org/gdhbqt>  
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/)

**25. 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/)

**26. Evaluation of computational genotyping of Structural Variations for clinical diagnoses.**

Varuna Chander, Richard A Gibbs, Fritz J Sedlazeck  
*Cold Spring Harbor Laboratory* (2019-02-22) <https://doi.org/gfwf66>  
DOI: [10.1101/558247](https://doi.org/10.1101/558247)

**27. Nebula: Ultra-efficient mapping-free structural variant genotyper**

Parsoa Khorsand, Fereydown Hormozdiari  
*Cold Spring Harbor Laboratory* (2019-03-04) <https://doi.org/gfwf67>  
DOI: [10.1101/566620](https://doi.org/10.1101/566620)

**28. Sequencing and de novo assembly of 150 genomes from Denmark as a population reference**

Lasse Maretty, Jacob Malte Jensen, Bent Petersen, Jonas Andreas Sibbesen, Siyang Liu, Palle Villesen, Laurits Skov, Kirstine Belling, Christian Theil Have, Jose M. G. Izarzugaza, ... Mikkel Heide Schierup

*Nature* (2017-07-26) <https://doi.org/gbpnnx>

DOI: [10.1038/nature23264](https://doi.org/10.1038/nature23264) · PMID: [28746312](https://pubmed.ncbi.nlm.nih.gov/28746312/)