Genotyping structural variation in variation graphs with the vg toolkit

This manuscript (permalink) was automatically generated from jmonlong/manu-vgsv@f2b6e54 on January 28, 2019.

Authors

Glenn Hickey[1,☯], David Heller[1,☯], Jean Monlong[1,☯], 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 were 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] (*REF_PETER_SOON*). These technologies are also enabling high-quality de novo genome assemblies to be produced[3,7], as well as large blocks of haplotype-resolved sequences[8]. 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[10,11,9]. By having variants in the graph, mapping rates are increased and variants are more uniformly covered, including indels and variants in complex regions[10]. Both the mapping and variant calling become variant-aware and benefit in term 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[12] and BayesTyper[13] 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[14], and Delly provides a genotyping feature in addition to its discovery mode[15].

Results

HGSVC

[3] provides a high-quality SV catalog of three samples, obtained using a consensus from different sequencing, phasing and variant caling technologies.

(Whole-genome) Simulation

The phasing information in the HGSVC VCF was used to extract two haplotypes for sample HG00514, and 30X pairend-end reads were simulated using vg sim. The reads were used to call VCFs then compared back to the original HGSVC calls.

Graph	type	TP	TP.baseline	FP	FN	precision	recall
HGSVC-Construct	Total	24451	24089	3119	2617	0.8854	0.902
INS	14596	14264	775	1421	0.9485	0.9094	0.9285
DEL	9855	9825	2344	1196	0.8074	0.8915	0.8474
HGSVC-1KG-Construct	Total	24172	23815	3236	2891	0.8804	0.8917
INS	14540	14111	836	1574	0.9441	0.8996	0.9213
DEL	9632	9704	2400	1317	0.8017	0.8805	0.8393
SVPOP-Construct	Total	10548	11559	5990	15147	0.6587	0.4328
INS	7733	8223	2266	7462	0.784	0.5243	0.6284
DEL	2815	3336	3724	7685	0.4725	0.3027	0.369
SVPOP-1KG-Construct	Total	10403	11369	6750	15337	0.6275	0.4257
INS	7497	7934	2198	7751	0.7831	0.5058	0.6146
DEL	2906	3435	4552	7586	0.4301	0.3117	0.3615

When restricting the comparisons to regions not identified as tandem repeats or segmental duplications in the Genome Browser:

Graph	type	TP	TP.baseline	FP	FN	precision	recall
HGSVC-Construct	Total	5901	5822	452	253	0.928	0.9584
INS	4026	3952	98	172	0.9758	0.9583	0.967
DEL	1875	1870	354	81	0.8408	0.9585	0.8958
HGSVC-1KG-Construct	Total	5880	5785	486	290	0.9225	0.9523
INS	4024	3922	123	202	0.9696	0.951	0.9602
DEL	1856	1863	363	88	0.8369	0.9549	0.892
SVPOP-Construct	Total	3565	3856	390	2219	0.9081	0.6347
INS	3091	3246	239	878	0.9314	0.7871	0.8532
DEL	474	610	151	1341	0.8016	0.3127	0.4499
SVPOP-1KG-Construct	Total	3574	3817	562	2258	0.8717	0.6283
INS	3066	3180	253	944	0.9263	0.7711	0.8416
DEL	508	637	309	1314	0.6734	0.3265	0.4398

(Whole-genome) Real reads

Graph	type	TP	TP.baseline	FP	FN	precision	recall
HGSVC-Construct	Total	18436	18500	6575	8206	0.7378	0.6927
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
INS	10647	10262	3304	5423	0.7564	0.6543	0.7017
DEL	7155	7684	2917	3337	0.7248	0.6972	0.7107
SVPOP-Construct	Total	9091	9931	10235	16775	0.4925	0.3719
INS	6972	7420	6706	8265	0.5253	0.4731	0.4978
DEL	2119	2511	3529	8510	0.4157	0.2278	0.2943

When restricting the comparisons to regions not identified as tandem repeats or segmental duplications in the Genome Browser:

Graph	type	TP	TP.baseline	FP	FN	precision	recall
HGSVC-Construct	Total	5197	5244	854	831	0.86	0.8632
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
INS	3642	3555	464	569	0.8845	0.862	0.8731
DEL	1461	1600	401	351	0.7996	0.8201	0.8097
SVPOP-Construct	Total	3251	3480	941	2595	0.7872	0.5728
INS	2859	3009	780	1115	0.7941	0.7296	0.7605
DEL	392	471	161	1480	0.7453	0.2414	0.3647

Methods

Discussion

References

1. **An integrated map of structural variation in 2,504 human genomes**
Peter H. SudmantTobias 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 · PMID: 26432246 · PMCID: PMC4617611

2. **Whole-genome sequence variation, population structure and demographic history of the Dutch population**
Laurent C FrancioliAndroniki 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 · PMID: 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, 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 · PMID: 25383537 · PMCID: 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 · PMID: 27895111 · PMCID: 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 · PMID: 29109544 · PMCID: 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/gbnhwk
DOI: 10.1038/nmeth.4366 · PMID: 28714986 · PMCID: PMC5578891

7. **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 · PMID: 29431738 · PMCID: PMC5889714

8. **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 · PMID: 27749838 · PMCID: PMC5503144

9. **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 · PMID: 28360232 · PMCID: PMC5411762

10. **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 · PMID: 30125266 · PMCID: PMC6126949

11. **Fast and accurate genomic analyses using genome graphs**
Goran Rakocevic, Vladimir Semenyuk, Wan-Ping Lee, James Spencer, John Browning, Ivan J. Johnson, Vladan Arsenijevic, Jelena Nadj, Kaushik Ghose, Maria C. Suciu, … Deniz Kural
Nature Genetics (2019-01-14) https://doi.org/gftd46
DOI: 10.1038/s41588-018-0316-4 · PMID: 30643257

12. **Graphtyper enables population-scale genotyping using pangenome graphs**
Hannes P Eggertsson, Hakon Jonsson, Snaedis Kristmundsdottir, Eirikur 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 · PMID: 28945251

13. **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 · PMID: 29915429

14. **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 · PMID: 26258291 · PMCID: PMC4589466

15. **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 · PMID: 22962449 · PMCID: PMC3436805