nature reviews genetics

https://doi.org/10.1038/s41576-023-00590-0

Review article

Check for updates

Variant calling and benchmarking in an era of complete human genome sequences

In the format provided by the authors and unedited

Supplementary Information

Supplementary Box 1: Classes and sources of errors in variant calls

Errors in variant calls can arise at each stage of genome analysis, from library preparation to sequencing to mapping to variant calling. Errors and variability introduced in each of these stages are often inter-related, but we outline different sources of errors below, summarized in the figure in ref [1].

Library preparation can include extracting DNA from cells, purifying DNA, size-selecting DNA, adding adapters for the sequencing technology, adding barcodes, making copies of DNA molecules, and enriching DNA from certain regions such as medically-relevant exons. If amplification methods like PCR are used to make many copies of small quantities of DNA or enrich particular genomic regions of interest, variants may be missed due to lower coverage for certain genomic contexts (e.g., extreme GC content), and there may be a higher variant error rate due to noisier sequences at homopolymers and tandem repeats. Targeted sequencing and linked reads typically require amplification steps that introduce these variant calling errors, though barcoding methods can enable downstream analyses to detect very low allele fractions. Targeted sequencing methods also typically have less uniform coverage over the regions of interest, which can cause lower sensitivity to variants near the edges of the targeted regions, but their higher coverage of most of the targeted regions can enable more robust variant calling, especially for somatic and mosaic variants. Some new approaches to targeted enrichment do not require amplification and/or maintain long DNA fragments that can be used for long-read sequencing, including long tandem repeats that cannot be called with other targeted methods. Finally, while most sequencing methods require shorter DNA fragments that cannot be mapped easily to repetitive genome regions, new approaches to isolate very long DNA can enable exceptionally long reads of hundreds of kilobase-pairs up to several million base-pairs.

The sequencing process itself can also introduce errors in variant calls due to systematic errors from the sequencing chemistry or limitations in the achievable read length. For example, certain sequence motifs may cause errors in one direction and not the other (e.g., GGT is sometimes read as GGG by Illumina).[2] Similarly, some highly repetitive satellite DNA sequences like HSat2 and HSat3 arrays were found to have strand bias and shorter read lengths for ONT and higher than normal coverage for HiFi. Coverage was lower than normal for both HiFi and ONT in AT-rich HSat1 arrays.[3] Also, homopolymers and tandem repeats can cause inaccurate base calls in surrounding sequences in short reads, or cause indel errors in long reads from single molecule sequencing methods. Variants may also be missed because the technology does sequence certain DNA motifs well, such as the GA-rich simple repeats in current PacBio HiFi data.[4] Certain technologies have lower error rates and/or produce much longer reads, and lower error rates and longer reads can enable variant calls in more repetitive regions.

Even short reads can be mapped confidently to most of the human genome, but certain regions such as segmental duplications, long tandem repeats, and satellite DNA can be challenging to map. Variant calls are most frequently missed or filtered when reads cannot be confidently mapped. In some cases, false variant calls are made due to mis-mapped reads, particularly if variants are not filtered stringently or events like gene conversions occurred in the sample.

Inaccurate variant calls also result from large differences between the sample and the reference genome, both due to errors in the reference and structural variability in the population. For example, some segmental duplications are missing copies or have extra copies in GRCh37 and/or GRCh38 and are not representative of the population. For example, 74 genes (22 protein-coding), including the known medically-relevant genes CBS, CRYAA, KCNE1, and H19 are falsely duplicated in GRCh38 (and similarly MRC1 on GRCh37).[5] These extra copies of genes cause ambiguously and incorrectly mapped reads and missed variant calls and mis-genotyped calls, unless the false copies are masked.[6,7] These false duplications cause many of the differences in variant calls seen in a recent comparison between exome sequencing reads aligned to GRCh37 vs. GRCh38.[8] Similarly, variant calls can be missed or mis-genotyped if the reference genome has more copies of a region than some individuals because the region is copy number variant in the population.

Conversely, GRCh37 and GRCh38 are missing copies of some segmental duplications that are in some or all individuals. Reads from the missing sequence often map incorrectly to the reference, causing several classes of variant calling errors. GRCh37 and GRCh38 are missing the gene KCNJ17, which causes variant calling errors in its medically relevant paralog KCNJ18 and KCNJ12, and a recent publication

about the new T2T-CHM13 reference explores specific examples of the different types of errors.[9] For example, false positives can result from paralog-specific variants (PSVs) coming from the missing sequence or from variants being called on the reference when in fact they should be variants on the missing sequence. PSVs are often filtered out by large population-level studies (e.g., with the GATK filter "Inbreedingcoeff") because they are heterozygous in most individuals, which is not expected from Hardy Weinberg Equilibrium. However, these can sometimes hide true variants that occur in some individuals at the same location, causing false negatives. False negatives can also result from extra reads aligning that cause the fraction of reads containing the variant to be lower than expected. Finally, genotype errors can occur from extra reference-matching reads, where homozygous variants appear to be heterozygous, or variants are detected in fewer individuals in the population. Some approaches intentionally remove extra copies of highly similar genes from the reference so that variants can be called with short reads, but these approaches do not permit assigning variant calls to a particular gene copy without downstream analysis. Many segmental duplications are copy number variable in the population, so the reference may match the copy number of some individuals but have fewer copies than other individuals, resulting in errors in those individuals.

Mapping of a phased de novo assembly of an individual to the reference alleviates many of the mapping challenges with individual short and long reads, but approaches to do this are less mature and often give different results in the most repetitive genomic regions, particularly regions with large structural variants.[10,11] The T2T Consortium recently generated the first accurately assembled centromere/satellite regions for all chromosomes in an effectively haploid human,[12–14] but quality assessment, alignment, and variant calling in these regions in many individuals is an area of active research, as discussed in the Advances in Variant Calling section in the main text.

Supplementary Box 2: Human non-germline variant calling

Since the focus of this review is on germline variant calling, we refer the readers to a recent review of somatic variant calling for a comprehensive examination of challenges.[15] Important considerations from germline variant benchmarking that are likely useful to somatic variant benchmarking are stratification to identify "clustering of errors" that correlate with sequencing technology or variant calling bias. Given the extensive filtering that is done with germline variants to remove possibly erroneous somatic variants, the stratifications may be different than those currently produced by GIAB/GA4GH. Another important item for somatic variant calling is proper handling representation differences especially when benchmarking of complex variants. As illustrated in Figure 5, comparing variants is a vital step though likely additionally challenging for low VAF somatic variants.

De novo variants are variants in a child that are not inherited from either parent. These variants can be important for diagnosing diseases like autism.[16] However, they are challenging to detect accurately because only ~100 exist in most individuals, so even a small number of false positives results in low precision.[17] In addition, variants that are represented differently between the parents and child can falsely appear to be de novo variants.[6] Therefore, several methods have been developed specifically to improve identification of de novo mutations. Typically, these methods call variants jointly between the parents and child to minimize representation differences or variants that are missed in the parents due to low coverage or low number of reads supporting the variant.[18] When calling variants in immortalized cell lines, it is useful to delineate between de novo variants that arise in the cell lines as compared to those likely present in blood cells. These are termed 'cell-line de novo' or 'non-germline' (somatic or arising in cell culture) DNM variants, which typically are several hundred to a couple thousand per cell line.[17]

While most variant calling is performed on DNA extracted from cells, specialized methods have been developed to call variants from other data types. For example, variants can be called from bulk RNA-sequencing data, but biases can be introduced by the library preparation steps, spliced read alignments, insufficient expression, and allele-specific expression.[19,20] Single cell RNA-seq[21] and targeted DNA sequencing[22] can be used to detect variants at the level of an individual cell, though RNA variant detection can be affected by expression levels and allele-specific expression, and DNA variant detection can suffer from allele dropout. Sequencing of cell free DNA in the blood is frequently used for prenatal screening of chromosomal abnormalities in the fetus,[23] and for monitoring cancer patients for minimal residual disease by doing deep targeted sequencing of particular mutations in the tumor.[24]

Supplementary References

1. Poplin, R., Zook, J. M. & DePristo, M. Challenges of Accuracy in Germline Clinical Sequencing Data. *JAMA* **326**, 268–269 (2021).
2. Meacham, F. *et al.* Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics* **12**, 451 (2011).
3. Mc Cartney, A. M. *et al.* Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *bioRxiv* 2021.07.02.450803 (2021) doi:10.1101/2021.07.02.450803.
4. Nurk, S. *et al.* HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020).
5. Wagner, J. *et al.* Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat. Biotechnol.* 1–9 (2022).
6. Wagner, J. *et al.* Benchmarking challenging small variants with linked and long reads. *Cell Genomics* **2**, (2022).
7. Behera, S. *et al.* Fixing reference errors efficiently improves sequencing results. *bioRxiv* 2022.07.18.500506 (2022) doi:10.1101/2022.07.18.500506.
8. Li, H. *et al.* Exome variant discrepancies due to reference-genome differences. *Am. J. Hum. Genet.* **108**, 1239–1250 (2021).
9. Aganezov, S. *et al.* A complete reference genome improves analysis of human genetic variation. *bioRxiv* 2021.07.12.452063 (2021) doi:10.1101/2021.07.12.452063.
10. Li, H. *et al.* A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat. Methods* **15**, 595–597 (2018).
11. Audano, P. A. *et al.* Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**, 663–675.e19 (2019).
12. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
13. Logsdon, G. A. *et al.* The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021).

14. Altemose, N. *et al.* Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022).
15. Cortés-Ciriano, I., Gulhan, D. C., Lee, J. J.-K., Melloni, G. E. M. & Park, P. J. Computational analysis of cancer genome sequencing data. *Nat. Rev. Genet.* **23**, 298–314 (2022).
16. Sanders, S. J. *et al.* Multiple Recurrent De Novo CNVs, Including Duplications of the 7q11.23 Williams Syndrome Region, Are Strongly Associated with Autism. *Neuron* **70**, 863–885 (2011).
17. Conrad, D. F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* **43**, 712–714 (2011).
18. Cleary, J. G. *et al.* Joint Variant and *De Novo* Mutation Identification on Pedigrees from High-Throughput Sequencing Data. *J. Comput. Biol.* **21**, 405–419 (2014).
19. Piskol, R., Ramaswami, G. & Li, J. B. Reliable identification of genomic variants from RNA-seq data. *Am. J. Hum. Genet.* **93**, 641–651 (2013).
20. Brouard, J.-S., Schenkel, F., Marete, A. & Bissonnette, N. The GATK joint genotyping workflow is appropriate for calling variants in RNA-seq experiments. *J. Anim. Sci. Biotechnol.* **10**, 44 (2019).
21. Schnepp, P. M., Chen, M., Keller, E. T. & Zhou, X. SNV identification from single-cell RNA sequencing data. *Hum. Mol. Genet.* **28**, 3569–3583 (2019).
22. Pellegrino, M. *et al.* High-throughput single-cell DNA sequencing of acute myeloid leukemia tumors with droplet microfluidics. *Genome Res.* **28**, 1345–1352 (2018).
23. Barrett, A. N., Zimmermann, B. G., Wang, D., Holloway, A. & Chitty, L. S. Implementing prenatal diagnosis based on cell-free fetal DNA: accurate identification of factors affecting fetal DNA yield. *PLoS One* **6**, e25202 (2011).
24. Tie, J. *et al.* Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer. *Sci. Transl. Med.* **8**, 346ra92 (2016).