TYPE Opinion
PUBLISHED 19 March 2024
DOI 10.3389/fbinf.2024.1384497

OPEN ACCESS

EDITED BY
Leena Salmela,
University of Helsinki, Finland

REVIEWED BY
Giacomo Baruzzo,
University of Padua, Italy

*CORRESPONDENCE
Inken Wohlers,
✉ iwohlers@fz-borstel.de

RECEIVED 09 February 2024
ACCEPTED 07 March 2024
PUBLISHED 19 March 2024

CITATION
Olbrich M, Bartels L and Wohlers I (2024), Sequencing technologies and hardware-accelerated parallel computing transform computational genomics research. *Front. Bioinform.* 4:1384497. doi: 10.3389/fbinf.2024.1384497

Sequencing technologies and hardware-accelerated parallel computing transform computational genomics research

Michael Olbrich[1], Lennart Bartels[2] and Inken Wohlers[2,3]*

[1]Center for Biotechnology, Khalifa University for Science and Technology, Abu Dhabi, United Arab Emirates, [2]Biomolecular Data Science in Pneumology, Research Center Borstel, Borstel, Germany, [3]University of Lübeck, Lübeck, Germany

KEYWORDS

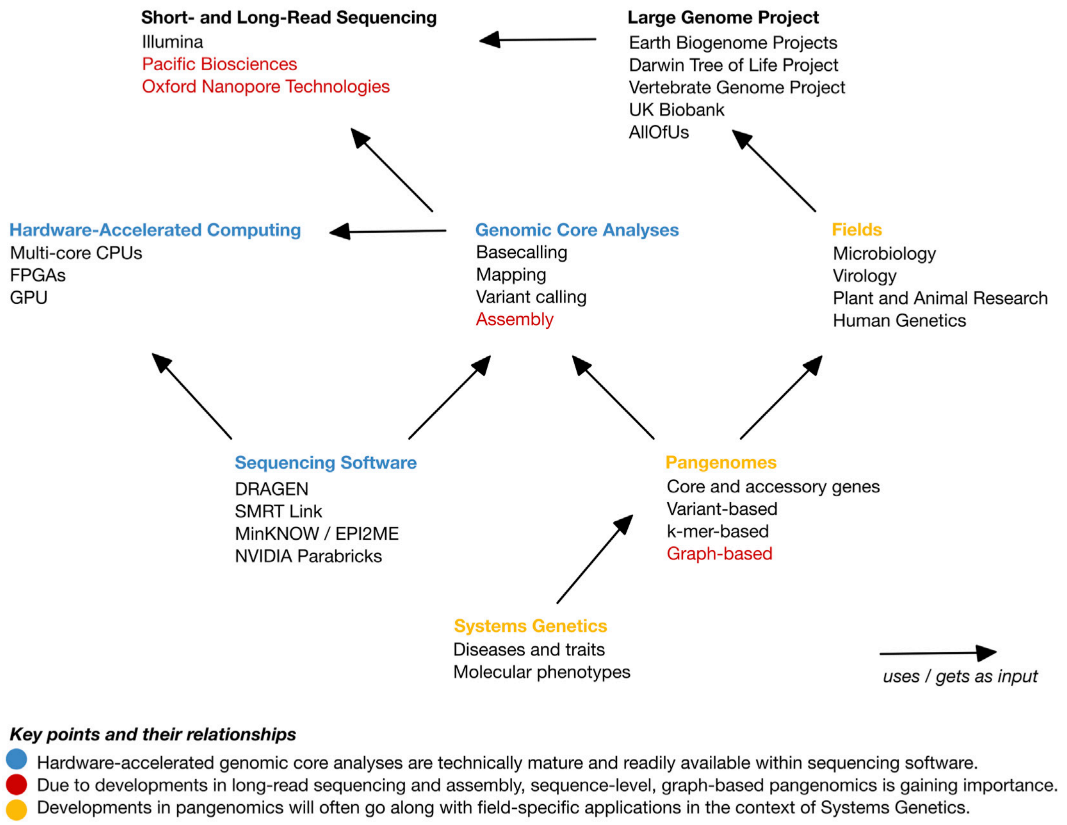
genome sequencing, parallel computing, computational genomics, sequencing software, pangenomics, long-read sequencing, genome references, systems genetics

Introduction

High-throughput sequencing and hardware-accelerated computing have both developed tremendously within the last decade. In genomics, progress is fueled by new technologies that allow the sequencing of increasingly longer reads with continuously improving accuracy at steadily decreasing costs. Shifts in parallel computing arise from the observation that specific computations can be efficiently parallelized on certain hardware. Developments in the field then gained momentum when deep learning became omnipresent, which is one application that can be very efficiently parallelized on general-purpose Graphics Processing Units (GPUs).

This article discusses the impact these sequencing and hardware-accelerated computing developments have on genomic data analysis. Based on the presented observations, we provide our view on how different ongoing efforts and lines of research will come together and transform computational genomics research.

An overview of this opinion piece is depicted in Figure 1. We commence by providing a background of high-throughput sequencing and current state and endeavors in genome research. Subsequently, we briefly introduce parallel computing. We then postulate that there are four core genome sequencing data analyses: basecalling, read mapping, variant identification, and assembly. Furthermore, we elucidate the extent of their integration and the utilization of hardware-based acceleration in software suites predominantly offered by sequencing providers. Finally, we observe that currently, and across fields, a shift towards improved genomic representations, typically referred to as pangenomes, is ongoing. A pangenome represents a set of genomes or genomic sequences and is thus devised for a specific purpose. This renders approaches much more application-field specific than current general-purpose core analyses. In the Discussion, we provide our view on a forthcoming shift from representing genome data towards linking genetic data with molecular and phenotypic information, i.e., towards Systems Genetics.



- ● Hardware-accelerated genomic core analyses are technically mature and readily available within sequencing software.
- ● Due to developments in long-read sequencing and assembly, sequence-level, graph-based pangenomics is gaining importance.
- ● Developments in pangenomics will often go along with field-specific applications in the context of Systems Genetics.

FIGURE 1
Overview of the topics of this opinion article and their relationships. The main points are highlighted in color. *Blue color:* Hardware-accelerated genomic core analyses are technically mature and readily available within sequencing software. *Red color:* Due to developments in long-read sequencing and assembly, sequence-level, graph-based pangenomics is gaining importance. *Yellow color:* Developments in pangenomics will often go along with field-specific applications in the context of Systems Genetics.

Genome sequencing developments and increase of genomic reference data

High-throughput sequencing, typically referred to as next-generation sequencing, has been around for more than two decades. Second-generation sequencing technology, dominated by Illumina, uses a sequencing-by-synthesis approach, restricting the length of reads obtained to typically 100–200 bases. Third-generation sequencing technologies have been around for the last decade, but have undergone considerable technological improvements within the last five years, resulting in gradually improved base accuracy and increased read lengths. The two different, major technologies are single-molecule real-time (SMRT) sequencing represented by Pacific Biosciences (PacBio) and nanopore sequencing, represented by Oxford Nanopore Technologies (ONT). Both technologies can generate reads with lengths up to megabases and base accuracies varying between protocols but up to more than 99%, e.g., for the PacBio high-fidelity (HiFi) protocol (Logsdon et al., 2020). These recent technological improvements made long-read sequencing a proclaimed method of the year 2022 (Marx, 2023).

Technological advancements and reduction in sequencing costs have resulted in up to a doubling of sequencing data within major sequence read archives (Arita et al., 2021) in the last five years (Katz et al., 2022), exceeding 50 petabytes (Yuan et al., 2024). Third-generation sequencing is increasingly applied, with currently more than 760,000 raw read files at the European Nucleotide Archive attributed to PacBio and Oxford Nanopore, respectively, of which about 700,000 PacBio and nearly all ONT files were submitted since 2019 (retrieved via https://www.ebi.ac.uk/ena/browser/advanced-search on 07/02/2024; search terms ‘instrument_platform = "ILLUMINA/PACBIO_SMRT/OXFORD_NANOPORE” AND first_created>2019-01-01’). In the same five-year period, about 21.3 million submitted read files were attributed to Illumina sequencing; thus, long-read sequencing files amounted to ~7% of submissions.

Developments have also facilitated the initiation of increasingly expansive genome sequencing projects in terms of scale and scope. Examples are the Earth BioGenome Project (Lewin et al., 2022), the Darwin Tree of Life Project (Darwin Tree of Life Project Consortium, 2022), the Vertebrate Genomes Project (VGP) (Rhie et al., 2021), as well as a large number of human genome projects, among them the UK Biobank (Halldorsson et al., 2022) and AllOfUs (All of Us Research Program Investigators et al., 2019; Ramirez et al., 2022). The number of sequenced genomes also increased in plant research (Sun et al., 2022) and Microbiology (Anani et al., 2020).

Parallel and GPU-accelerated computing speeds up computation by orders of magnitude

Parallel computing refers to utilizing multiple computing units to address a specific problem, wherein computations are executed concurrently, significantly improving computational speed. Suitable hardware resources are required on which the parallel calculations can be performed. Modern multi-core central processing units (CPUs) are capable of running many execution threads simultaneously. A special form of parallelization can be achieved by using graphics processing units (GPUs), initially designed for rapid parallel execution of mathematical calculations in computer graphics applications. With a vast quantity of processing units, GPUs surpass even the largest CPUs in terms of parallelization. GPUs have since been repurposed across various domains to accelerate tasks in which the same operation is applied to different subsets of data. Manufacturers have subsequently developed specialized GPUs and application programming interfaces to promote the development of GPU-accelerated applications. This gave rise to general-purpose computing on GPUs, the most well-known example of which is deep learning.

Acceleration by yet another order of magnitude can be achieved by concurrent computation on multiple machines. An example of such distributed computing is parallel processing on compute clusters using scientific workflow management systems such as snakemake (Köster and Rahmann, 2012) or nextflow (Di Tommaso et al., 2017), which also facilitate deployment in the cloud. Distributed computing for individual core genome analyses is a topic of interest with various tools available (Zou et al., 2021).

Hardware-accelerated primary and secondary sequencing data analysis is technically mature and readily available

Genome sequencing data processing can be divided into primary, secondary, and tertiary analyses. Primary analysis can be considered the generation of sequence data from the sequencing devices' raw measurements, typically in conjunction with PHRED scores that estimate individual base accuracy. This so-called base calling is typically performed during sequencing. Secondary analyses are read mapping, i.e., providing for each read the reference sequence position to which it matches, and variant calling, i.e., identifying differences from a specific reference sequence. Finally, genome assembly, i.e., reconstructing fully haplotype-phased genomes from sequencing data, will likely become an integral genomic secondary analysis of long-read data. We examined these core analyses, as well as their implementation status regarding hardware acceleration in the software solutions provided by the major sequencing companies Illumina, PacBio, and ONT.

Illumina's DRAGEN is commercial software that can be used directly on specific Illumina machines (NovaSeq X Series, NextSeq 1000/2000), on servers on-premise, or in the cloud. It utilizes the sequencer's onboard Field Programmable Gate Arrays (FPGAs) to accelerate primary analysis, i.e., the generation of FASTQ files from the binary base call (BCL) files. Concerning secondary analysis, DRAGEN provides the Genome Analysis Toolkit (GATK) (McKenna et al., 2010) for variant calling, which comes with a proprietary hardware acceleration that speeds up analyses significantly (Betschart et al., 2022). Besides secondary analyses, an extensive range of common tertiary analyses can be performed with DRAGEN, incurring additional licensing costs.

PacBio's SMRT Link software includes the SMRT analysis module for secondary analysis, which provides various types of analyses of HiFi sequencing data. The variant calling workflow uses deep learning-based DeepVariant for small variant detection (Poplin et al., 2018), which can be run with GPU acceleration (Yun et al., 2021).

ONT's MinKNOW software controls the sequencing device and offers GPU-accelerated base calling. For the secondary analysis, ONT provides the EPI2ME platform, a collection of open-source workflows that can be run free of charge.

Besides easy-to-use graphical software provided by all three major sequencing providers, major industry players, including NVIDIA, have recognized the genomic analysis market's potential and have introduced frameworks like Parabricks that leverage GPUs to improve processing speed (Clara Parabricks 4.0.0, 2024). The acceleration of established genomic alignment and variant calling pipelines by factors of 10–100 exemplifies the impact of these technological and algorithmic developments (O'Connell et al., 2023). A notable aspect of this framework is its full accessibility to academics, with charges applied solely for commercial usage.

Deep learning is the key enabling technology for long-read base calling, with ONT using PyTorch (Paszke et al., 2019) and PacBio using TensorFlow (Developers, 2024). Accordingly, the respective latest sequencing devices are equipped with on-board GPUs. Specific deep learning achievements are improved PacBio HiFi read generation with DeepConsensus (Baid et al., 2023) and ONT models that allow detection of various base modification types (Ahsan et al., 2024) currently with basecaller Dorado.

Pangenome approaches are increasingly applied across all research fields

The concept of pangenomes emerged in the area of microbial genomics (Tettelin et al., 2005) and has specific relevance in clinical microbiology (Anani et al., 2020). The pangenome was defined as the set of genes that occur within a bacterial phylogenetic clade, distinguishing between a core genome of shared genes, and an accessory genome of remaining genes. Besides gene-level approaches, k-mer-based approaches are used in microbiology, representing the pangenome as a presence-absence matrix of unique k-mers within contributing genomes.

Recently, pangenomics has been extended to plants (Li et al., 2022), animals (Golicz et al., 2020), and humans (Liao et al., 2023). Eukaryotic genomes are typically assessed on sequence- and not gene-level, since they contain more non-coding sequences and a larger core genome. Thus, the representation of sequence-level diversity is the primary scope of pangenomics in eukaryotes. The major benefit of sequence-level pangenomic approaches is the improvement of variant calling since variants can be identified more thoroughly with respect to a preferably diverse and comprehensive set of reference sequences.

To form a multi-genome reference, so-called pangenome graphs (Eizenga et al., 2020) are created—typically from high-quality assembled genomes. In these graphs, nodes represent subsequences and edges the adjacency of the sequences within an observed genome. Since graphs represent multiple genomes, the genetic variation within a population is captured better than with existing linear genome references (Ballouz et al., 2019; Liao et al., 2023).

Human genetics stands out as the driving force that propelled computational genomics toward harnessing the potential of long reads for high-quality assembly and subsequent pangenomic representations. This journey began with the achievement of a telomere-to-telomere assembly, encompassing all previously uncharted regions of the human genome (Miga et al., 2020; Nurk et al., 2022; Rhie et al., 2023) via diploid human assembly (Ebert et al., 2021; Porubsky et al., 2021) in conjunction with assembly algorithmic developments (e.g., Hifiasm (Cheng et al., 2021) and Verkko (Rautiainen et al., 2023)), achieving a first draft human pangenome in 2023 (Liao et al., 2023).

Discussion

In the forthcoming decade, *de novo* assembly from long-read sequencing data stands to ascend as the coveted gold standard in many applications. Further, high-quality assembly in conjunction with pangenome representations will, for the first time, provide a complete picture of genomes and genomic variation that is not reference-biased. Such improved genomic resolution will provide opportunities to discover genotype-phenotype relationships that so far have been overlooked. This applies universally across genomes of viruses, microbes, plants, animals, and humans, all of which are currently the focus of extensive, discipline-specific genome projects. These ambitious projects, in return, catalyze the development of high-performance, hardware-accelerated implementations of largely open-source core analysis tools and corresponding graphical software for easy use with sequencing devices. In contrast, analyses that require the precise resolution afforded by long-read data, notably structural variant calling and assembly, necessitate further research and development in computational genomics. Given that these are at an earlier research stage, hardware acceleration options beyond CPU utilization have yet to be fully explored.

With the availability of longer assembled sequences up to entire genomes, pangenomic approaches will gain importance. Pangenomes, however, will often have application-specific requirements. Although pangenomic representations of genomes from high-quality assemblies are, in theory, superior to single reference-based variant lists, computational tools are still needed to construct, process, and annotate such representations for specific applications. We believe that this is a major focus of computational genomics research in the next decade and that field- and application-specific characteristics will play an important role, possibly resulting in a multitude of pangenome-centered tools. Specifically, we think that developments in pangenome representations will go along and often align with the development and application of methods that link sequence representations with phenotypes. This interplay of pangenome approaches with methods from statistical genetics and machine learning will help unlock the potential of Systems Genetics, eventually providing a holistic understanding of biological systems from the genomic viewpoint.

Author contributions

MO: Writing–original draft, Writing–review and editing. LB: Writing–original draft, Writing–review and editing. IW: Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ahsan, M. U., Gouru, A., Chan, J., Zhou, W., and Wang, K. (2024). A signal processing and deep learning framework for methylation detection using Oxford Nanopore sequencing. *Nat. Commun.* 15, 1448. doi:10.1038/s41467-024-45778-y

All of Us Research Program Investigators; Denny, J. C., Rutter, J. L., Goldstein, D. B., Philippakis, A., Smoller, J. W., et al. (2019). The "all of us" research Program. *N. Engl. J. Med.* 381, 668–676. doi:10.1056/NEJMsr1809937

Anani, H., Zgheib, R., Hasni, I., Raoult, D., and Fournier, P.-E. (2020). Interest of bacterial pangenome analyses in clinical microbiology. *Microb. Pathog.* 149, 104275. doi:10.1016/j.micpath.2020.104275

Arita, M., Karsch-Mizrachi, I., and Cochrane, G. (2021). The international nucleotide sequence database collaboration. *Nucleic Acids Res.* 49, D121–D124. doi:10.1093/nar/gkaa967

Baid, G., Cook, D. E., Shafin, K., Yun, T., Llinares-López, F., Berthet, Q., et al. (2023). DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer. *Nat. Biotechnol.* 41, 232–238. doi:10.1038/s41587-022-01435-7

Ballouz, S., Dobin, A., and Gillis, J. A. (2019). Is it time to change the reference genome? *Genome Biol.* 20, 159. doi:10.1186/s13059-019-1774-4

Betschart, R. O., Thiéry, A., Aguilera-Garcia, D., Zoche, M., Moch, H., Twerenbold, R., et al. (2022). Comparison of calling pipelines for whole genome sequencing: an empirical study demonstrating the importance of mapping and alignment. *Sci. Rep.* 12, 21502. doi:10.1038/s41598-022-26181-3

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat. Methods* 18, 170–175. doi:10.1038/s41592-020-01056-5

Clara Parabricks 4.0.0 (2024). NVIDIA docs. Available at: https://docs.nvidia.com/clara/parabricks/4.0.0/index.html (Accessed February 2, 2024).

Darwin Tree of Life Project Consortium: Blaxter, M., Mieszkowska, N., Di Palma, F., Holland, P., Durbin, R., et al. (2022). Sequence locally, think globally: the Darwin tree of Life project. *Proc. Natl. Acad. Sci. U. S. A.* 119, e2115642118. doi:10.1073/pnas.2115642118

Developers, T. (2024). *TensorFlow*. doi:10.5281/zenodo.10713739

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* 35, 316–319. doi:10.1038/nbt.3820

Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M. J., et al. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* 372, eabf7117. doi:10.1126/science.abf7117

Eizenga, J. M., Novak, A. M., Sibbesen, J. A., Heumos, S., Ghaffaari, A., Hickey, G., et al. (2020). Pangenome graphs. *Annu. Rev. Genomics Hum. Genet.* 21, 139–162. doi:10.1146/annurev-genom-120219-080406

Golicz, A. A., Bayer, P. E., Bhalla, P. L., Batley, J., and Edwards, D. (2020). Pangenomics comes of age: from bacteria to plant and animal applications. *Trends Genet.* 36, 132–145. doi:10.1016/j.tig.2019.11.006

Halldorsson, B. V., Eggertsson, H. P., Moore, K. H. S., Hauswedell, H., Eiriksson, O., Ulfarsson, M. O., et al. (2022). The sequences of 150,119 genomes in the UK Biobank. *Nature* 607, 732–740. doi:10.1038/s41586-022-04965-x

Katz, K., Shutov, O., Lapoint, R., Kimelman, M., Brister, J. R., and O'Sullivan, C. (2022). The Sequence Read Archive: a decade more of explosive growth. *Nucleic Acids Res.* 50, D387–D390. doi:10.1093/nar/gkab1053

Köster, J., and Rahmann, S. (2012). Snakemake--a scalable bioinformatics workflow engine. *Bioinforma. Oxf. Engl.* 28, 2520–2522. doi:10.1093/bioinformatics/bts480

Lewin, H. A., Richards, S., Lieberman Aiden, E., Allende, M. L., Archibald, J. M., Bálint, M., et al. (2022). The Earth BioGenome project 2020: starting the clock. *Proc. Natl. Acad. Sci. U. S. A.* 119, e2115635118. doi:10.1073/pnas.2115635118

Li, W., Liu, J., Zhang, H., Liu, Z., Wang, Y., Xing, L., et al. (2022). Plant pan-genomics: recent advances, new challenges, and roads ahead. *J. Genet. Genomics* 49, 833–846. doi:10.1016/j.jgg.2022.06.004

Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., et al. (2023). A draft human pangenome reference. *Nature* 617, 312–324. doi:10.1038/s41586-023-05896-x

Logsdon, G. A., Vollger, M. R., and Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* 21, 597–614. doi:10.1038/s41576-020-0236-x

Marx, V. (2023). Method of the year: long-read sequencing. *Nat. Methods* 20, 6–11. doi:10.1038/s41592-022-01730-w

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi:10.1101/gr.107524.110

Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., et al. (2020). Telomere-to-telomere assembly of a complete human X chromosome. *Nature* 585, 79–84. doi:10.1038/s41586-020-2547-7

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., et al. (2022). The complete sequence of a human genome. *Science* 376, 44–53. doi:10.1126/science.abj6987

O'Connell, K. A., Yosufzai, Z. B., Campbell, R. A., Lobb, C. J., Engelken, H. T., Gorrell, L. M., et al. (2023). Accelerating genomic workflows using NVIDIA Parabricks. *BMC Bioinforma.* 24, 221. doi:10.1186/s12859-023-05292-2

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "PyTorch: an imperative style, high-performance deep learning library," in *Advances in neural information processing systems 32* (New York, NY: Curran Associates, Inc.), 8024–8035. Available at: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* 36, 983–987. doi:10.1038/nbt.4235

Porubsky, D., Ebert, P., Audano, P. A., Vollger, M. R., Harvey, W. T., Marijon, P., et al. (2021). Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat. Biotechnol.* 39, 302–308. doi:10.1038/s41587-020-0719-5

Ramirez, A. H., Sulieman, L., Schlueter, D. J., Halvorson, A., Qian, J., Ratsimbazafy, F., et al. (2022). The all of us research Program: data quality, utility, and diversity. *Patterns N. Y. N.* 3, 100570. doi:10.1016/j.patter.2022.100570

Rautiainen, M., Nurk, S., Walenz, B. P., Logsdon, G. A., Porubsky, D., Rhie, A., et al. (2023). Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat. Biotechnol.* 41, 1474–1482. doi:10.1038/s41587-023-01662-6

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., et al. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 592, 737–746. doi:10.1038/s41586-021-03451-0

Rhie, A., Nurk, S., Cechova, M., Hoyt, S. J., Taylor, D. J., Altemose, N., et al. (2023). The complete sequence of a human Y chromosome. *Nature* 621, 344–354. doi:10.1038/s41586-023-06457-y

Sun, Y., Shang, L., Zhu, Q.-H., Fan, L., and Guo, L. (2022). Twenty years of plant genome sequencing: achievements and challenges. *Trends Plant Sci.* 27, 391–401. doi:10.1016/j.tplants.2021.10.006

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., et al. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome.". *Proc. Natl. Acad. Sci. U. S. A.* 102, 13950–13955. doi:10.1073/pnas.0506758102

Yuan, D., Ahamed, A., Burgin, J., Cummins, C., Devraj, R., Gueye, K., et al. (2024). The European nucleotide archive in 2023. *Nucleic Acids Res.* 52, D92–D97. doi:10.1093/nar/gkad1067

Yun, T., Li, H., Chang, P.-C., Lin, M. F., Carroll, A., and McLean, C. Y. (2021). Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinforma. Oxf. Engl.* 36, 5582–5589. doi:10.1093/bioinformatics/btaa1081

Zou, Y., Zhu, Y., Li, Y., Wu, F.-X., and Wang, J. (2021). Parallel computing for genome sequence processing. *Brief. Bioinform.* 22, bbab070. bbab070. doi:10.1093/bib/bbab070