References

[1] The 1000 Genomes Project Participants. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[2] Mohamed Ibrahim Abouelhoda, Stefan Kurtz, and Enno Ohlebusch. Replacing suffix trees with enhanced suffix arrays. *Journal of Discrete Algorithms*, 2(1):53–86, 2004.

[3] JM Adams, PGN Jeppesen, F Sanger, and BG Barrell. Nucleotide sequence from the coat protein cistron of r17 bacteriophage RNA. *Nature*, 223(5210):1009–1014, 1969.

[4] Cornelis A Albers, Gerton Lunter, Daniel G MacArthur, Gilean McVean, Willem H Ouwehand, and Richard Durbin. Dindel: accurate indel calls from short-read data. *Genome Research*, 21(6):961–973, 2011.

[5] Morten E Allentoft, Martin Sikora, Karl-Göran Sjögren, Simon Rasmussen, Morten Rasmussen, Jesper Stenderup, Peter B Damgaard, Hannes Schroeder, Torbjörn Ahlström, Lasse Vinner, et al. Population genomics of bronze age Eurasia. *Nature*, 522(7555):167–172, 2015.

[6] Manuel Allhoff, Alexander Schönhuth, Marcel Martin, Ivan G Costa, Sven Rahmann, and Tobias Marschall. Discovering motifs that induce sequencing errors. *BMC Bioinformatics*, 14(5):S1, 2013.

[7] Carlos Alonso-Blanco, Jorge Andrade, Claude Becker, Felix Bemm, Joy Bergelson, Karsten M Borgwardt, Jun Cao, Eunyoung Chae, Todd M Dezwaan, Wei Ding, et al. 1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. *Cell*, 166(2):481–491, 2016.

[8] Stephen F Altschul and Bruce W Erickson. Optimal sequence alignment using affine gap costs. *Bulletin of Mathematical Biology*, 48(5-6):603–616, 1986.

[9] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3): 403–410, 1990.

[10] Alberto Apostolico. The myriad virtues of subword trees. In *Combinatorial algorithms on words*, pages 85–96. Springer, 1985.

[11] Oswald T Avery, Colin M MacLeod, and Maclyn McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *Journal of Experimental Medicine*, 79(2):137–158, 1944.

[12] Shankar Balasubramanian, David Klenerman, and David Bentley. Arrayed biomolecules and their use in sequencing, 2004. US Patent 6,787,308.

[13] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, Sergey I Nikolenko, Son Pham, Andrey D Prjibelski, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19(5): 455–477, 2012.

[14] Markus J Bauer, Anthony J Cox, and Giovanna Rosone. Lightweight algorithms for constructing and inverting the BWT of string collections. *Theoretical Computer Science*, 483:134–148, 2013.

[15] Richard Bellman. On the theory of dynamic programming. *Proceedings of the National Academy of Sciences*, 38(8):716–719, 1952.

[16] David R Bentley, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, John Milton, Clive G Brown, Kevin P Hall, Dirk J Evers, Colin L Barnes, Helen R Bignell, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008.

[17] Anders Bergström, Jared T Simpson, Francisco Salinas, Benjamin Barré, Leopold Parts, Amin Zia, Alex N Nguyen Ba, Alan M Moses, Edward J Louis, Ville Mustonen, et al. A high-definition view of functional genetic variation from natural yeast genomes. *Molecular Biology and Evolution*, 31(4):872–888, 2014.

[18] Derek M Bickhart, Benjamin D Rosen, Sergey Koren, Brian L Sayre, Alex R Hastie, Saki Chan, Joyce Lee, Ernest T Lam, Ivan Liachko, Shawn T Sullivan, et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature Genetics*, 49(4):643–650, 2017.

[19] Etienne Birmelé, Pierluigi Crescenzi, Rui Ferreira, Roberto Grossi, Vincent Lacroix, Andrea Marino, Nadia Pisanti, Gustavo Sacomoto, and Marie-France Sagot. Efficient bubble enumeration in directed graphs. In *International Symposium on String Processing and Information Retrieval*, pages 118–129. Springer, 2012.

[20] Inanç Birol, Shaun D Jackman, Cydney B Nielsen, Jenny Q Qian, Richard Varhol, Greg Stazyk, Ryan D Morin, Yongjun Zhao, Martin Hirst, Jacqueline E Schein, et al. De novo transcriptome assembly with ABySS. *Bioinformatics*, 25(21):2872–2877, 2009.

[21] Mathieu Blanchette, W James Kent, Cathy Riemer, Laura Elnitski, Arian FA Smit, Krishna M Roskin, Robert Baertsch, Kate Rosenbloom, Hiram Clawson, Eric D Green, et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research*, 14(4):708–715, 2004.

[22] Nathan Blow. Decoding the unsequenceable, 2015. URL https://www.future-science.com/doi/pdf/10.2144/000114252.

[23] Alexander Bowe, Taku Onodera, Kunihiko Sadakane, and Tetsuo Shibuya. Succinct de Bruijn graphs. In *International Workshop on Algorithms in Bioinformatics*, pages 225–235. Springer, 2012.

[24] Ljiljana Brankovic, Costas S Iliopoulos, Ritu Kundu, Manal Mohamed, Solon P Pissis, and Fatima Vayani. Linear-time superbubble identification algorithm for genome assembly. *Theoretical Computer Science*, 609:374–383, 2016.

[25] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527, 2016.

[26] Adrian W Briggs, Udo Stenzel, Matthias Meyer, Johannes Krause, Martin Kircher, and Svante Pääbo. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Research*, 38(6):e87–e87, 2009.

[27] S. R. Browning and B. L. Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genetics*, 81(5):1084–1097, 2007.

[28] Michael Burrows and David J Wheeler. A block-sorting lossless data compression algorithm. *Digital Equipment Corporation technical reports*, 124, 1994.

[29] Jonathan Butler, Iain MacCallum, Michael Kleber, Ilya A Shlyakhter, Matthew K Belmonte, Eric S Lander, Chad Nusbaum, and David B Jaffe. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Research*, 18(5):810–820, 2008.

[30] Bruno Canard and Robert S Sarfati. DNA polymerase fluorescent substrates with reversible 3'-tags. *Gene*, 148(1):1–6, 1994.

[31] Jun Cao, Korbinian Schneeberger, Stephan Ossowski, Torsten Günther, Sebastian Bender, Joffrey Fitz, Daniel Koenig, Christa Lanz, Oliver Stegle, Christoph Lippert, et al. Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature Genetics*, 43(10):956–963, 2011.

[32] Humberto Carrillo and David Lipman. The multiple sequence alignment problem in biology. *SIAM Journal on Applied Mathematics*, 48(5):1073–1082, 1988.

[33] John Castiblanco. *A primer on current and common sequencing technologies.* El Rosario University Press, 2013. URL https://www.ncbi.nlm.nih.gov/books/NBK459463/.

[34] Mark JP Chaisson, Ashley D Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J Gardner, Oscar Rodriguez, Li Guo, Ryan L Collins, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *bioRxiv:193144*, 2018.

[35] Mahul Chakraborty, Nicholas W VanKuren, Roy Zhao, Xinwen Zhang, Shannon Kalsow, and JJ Emerson. Hidden genetic variation shapes the structure of functional elements in drosophila. *Nature Genetics*, 50(1):20–25, 2018.

[36] Danny Challis, Lilian Antunes, Erik Garrison, Eric Banks, Uday S Evani, Donna Muzny, Ryan Poplin, Richard A Gibbs, Gabor Marth, and Fuli Yu. The distribution and mutagenesis of short coding indels from 1,128 whole exomes. *BMC Genomics*, 16(1):143, 2015.

[37] Zheng Chang, Guojun Li, Juntao Liu, Yu Zhang, Cody Ashby, Deli Liu, Carole L Cramer, and Xiuzhen Huang. Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. *Genome Biology*, 16(1):30, 2015.

[38] J Michael Cherry, Caroline Adler, Catherine Ball, Stephen A Chervitz, Selina S Dwight, Erich T Hester, Yankai Jia, Gail Juvik, TaiYun Roe, Mark Schroeder, et al. SGD: Saccharomyces genome database. *Nucleic Acids Research*, 26(1):73–79, 1998.

[39] Rayan Chikhi and Guillaume Rizk. Space-efficient and exact de Bruijn graph representation based on a bloom filter. *Algorithms for Molecular Biology*, 8(1):22, 2013.

[40] Rayan Chikhi, Antoine Limasset, and Paul Medvedev. Compacting de Bruijn graphs from sequencing data quickly and in low memory. *Bioinformatics*, 32(12): i201–i208, 2016.

[41] 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, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, 13(12):1050–1054, 2016.

[42] Deanna M Church. Genomes for all. *Nature Biotechnology*, 36(9):815–816, 2018.

[43] Richard M Clark, Gabriele Schweikert, Christopher Toomajian, Stephan Ossowski, Georg Zeller, Paul Shinn, Norman Warthmann, Tina T Hu, Glenn Fu, David A Hinds, et al. Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. *Science*, 317(5836):338–342, 2007.

[44] Jonathan Cohen. Graph twiddling in a mapreduce world. *Computing in Science & Engineering*, 11(4):29–41, 2009.

[45] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[46] Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics*, 19(1):118–135, 2018.

[47] ENCODE Project Consortium et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[48] International Human Genome Sequencing Consortium et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[49] UK10K Consortium et al. The UK10K project identifies rare variants in health and disease. *Nature*, 526(7571):82–90, 2015.

[50] Francis HC Crick. On protein synthesis. *Symposia of the Society for Experimental Biology*, 12:138–163, 1958.

[51] Francis HC Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[52] Jesse Dabney, Matthias Meyer, and Svante Pääbo. Ancient DNA damage. *Cold Spring Harbor Perspectives in Biology*, page a012567, 2013.

[53] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and VCFtools. *Bioinformatics*, 27(15): 2156–2158, 2011.

[54] Matei David, Lewis Jonathan Dursi, Delia Yao, Paul C Boutros, and Jared T Simpson. Nanocall: an open source basecaller for oxford nanopore sequencing data. *Bioinformatics*, 33(1):49–55, 2016.

[55] Peter de Barros Damgaard, Rui Martiniano, Jack Kamm, J Víctor Moreno-Mayar, Guus Kroonen, Michaël Peyrot, Gojko Barjamovic, Simon Rasmussen, Claus Zacho, Nurbol Baimukhanov, et al. The first horse herders and the impact of early Bronze Age steppe expansions into Asia. *Science*, 360(6396):1422–1442, 2018.

[56] Nicolaas Govert De Bruijn. A combinatorial problem. *Koninklijke Nederlandse Akademie v. Wetenschappen*, 49(49):758–764, 1946.

[57] Daniel Aguirre de Cárcer, Alberto López-Bueno, David A Pearce, and Antonio Alcamí. Biodiversity and distribution of polar freshwater DNA viruses. *Science Advances*, 1(5):e1400127, 2015.

[58] David Deamer, Mark Akeson, and Daniel Branton. Three decades of nanopore sequencing. *Nature Biotechnology*, 34(5):518–524, 2016.

[59] Jeffrey Dean and Sanjay Ghemawat. MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[60] O. Delaneau, J. Marchini, and J. F. Zagury. A linear complexity phasing method for thousands of genomes. *Nature Methods*, 9(2):179–181, 2012.

[61] Arthur L Delcher, Simon Kasif, Robert D Fleischmann, Jeremy Peterson, Owen White, and Steven L Salzberg. Alignment of whole genomes. *Nucleic Acids Research*, 27(11):2369–2376, 1999.

[62] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo Del Angel, Manuel A Rivas, Matt Hanna, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–498, 2011.

[63] Alexander Dilthey, Charles Cox, Zamin Iqbal, Matthew R Nelson, and Gil McVean. Improved genome inference in the MHC using a population reference graph. *Nature Genetics*, 47(6):682–688, 2015.

[64] Olga Dudchenko, Sanjit S Batra, Arina D Omer, Sarah K Nyquist, Marie Hoeger, Neva C Durand, Muhammad S Shamim, Ido Machol, Eric S Lander, Aviva Presser Aiden, et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science*, 356(6333):92–95, 2017.

[65] Richard Durbin. Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). *Bioinformatics*, 30(9):1266–1272, 2014.

[66] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, 1998.

[67] Dent A Earl, Keith Bradnam, John St John, Aaron Darling, Dawei Lin, Joseph Faas, Hung On Ken Yu, Buffalo Vince, Daniel R Zerbino, Mark Diekhans, et al. Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Research*, 21(12):2224–2241, 2011.

[68] Sarah G Earle, Chieh-Hsi Wu, Jane Charlesworth, Nicole Stoesser, N Claire Gordon, Timothy M Walker, Chris CA Spencer, Zamin Iqbal, David A Clifton, Katie L Hopkins, et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nature Microbiology*, 1(5):16041, 2016.

[69] MA Eberle, M Kallberg, HY Chuang, P Tedder, S Humphray, D Bentley, and E Margulies. Platinum genomes: A systematic assessment of variant accuracy using a large family pedigree. In *60th Annual Meeting of The American Society of Human Genetics*, pages 22–26, 2013.

[70] Robert A Edwards and Forest Rohwer. Viral metagenomics. *Nature Reviews Microbiology*, 3(6):504–510, 2005.

[71] Hannes P Eggertsson, Hakon Jonsson, Snaedis Kristmundsdottir, Eirikur Hjartarson, Birte Kehr, Gisli Masson, Florian Zink, Kristjan E Hjorleifsson, Aslaug Jonasdottir, Adalbjorg Jonasdottir, et al. Graphtyper enables population-scale genotyping using pangenome graphs. *Nature Genetics*, 49(11):1654–1660, 2017.

[72] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, et al. Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138, 2009.

[73] John Ellson, Emden Gansner, Lefteris Koutsofios, Stephen C North, and Gordon Woodhull. Graphviz—open source graph drawing tools. In *International Symposium on Graph Drawing*, pages 483–484. Springer, 2001.

[74] Ester Falconer, Mark Hills, Ulrike Naumann, Steven SS Poon, Elizabeth A Chavez, Ashley D Sanders, Yongjun Zhao, Martin Hirst, and Peter M Lansdorp. DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nature Methods*, 9(11):1107–1112, 2012.

[75] Xian Fan, Mark Chaisson, Luay Nakhleh, and Ken Chen. HySA: A Hybrid Structural variant Assembly approach using next generation and single-molecule sequencing technologies. *Genome Research*, 27(5):793–800, 2017.

[76] Michael Farrar. Striped Smith–Waterman speeds database searches six times over other SIMD implementations. *Bioinformatics*, 23(2):156–161, 2007.

[77] Paolo Ferragina and Giovanni Manzini. Opportunistic data structures with applications. In *Proceedings of the 41st Annual Symposium on the Foundations of Computer Science*, pages 390–398. IEEE, 2000.

[78] Paolo Ferragina and Giovanni Manzini. Opportunistic data structures with applications. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 390–398. IEEE, 2000.

[79] Paolo Ferragina and Giovanni Manzini. An experimental study of an opportunistic index. In *Proceedings of the twelfth annual ACM-SIAM symposium on Discrete Algorithms*, pages 269–278. Society for Industrial and Applied Mathematics, 2001.

[80] Paolo Ferragina and Giovanni Manzini. Indexing compressed text. *Journal of the ACM (JACM)*, 52(4):552–581, 2005.

[81] Paolo Ferragina, Giovanni Manzini, Veli Mäkinen, and Gonzalo Navarro. An alphabet-friendly FM-index. In *String Processing and Information Retrieval*, pages 150–160. Springer, 2004.

[82] Paolo Ferragina, Fabrizio Luccio, Giovanni Manzini, and S Muthukrishnan. Structuring labeled trees for optimal succinctness, and beyond. In *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)*, pages 184–193. IEEE, 2005.

[83] Paolo Ferragina, Fabrizio Luccio, Giovanni Manzini, and Shan Muthukrishnan. Compressing and indexing labeled trees, with applications. *Journal of the ACM (JACM)*, 57(1):4, 2009.

[84] Walter Fiers, Roland Contreras, Fred Duerinck, Guy Haegeman, Dirk Iserentant, Jozef Merregaert, W Min Jou, Francis Molemans, Alex Raeymaekers, A Van den Berghe, et al. Complete nucleotide sequence of bacteriophage ms2 RNA: primary and secondary structure of the replicase gene. *Nature*, 260(5551):500–507, 1976.

[85] Walter M Fitch. An improved method of testing for evolutionary homology. *Journal of Molecular Biology*, 16(1):9–16, 1966.

[86] Robert D Fleischmann, Mark D Adams, Owen White, Rebecca A Clayton, Ewen F Kirkness, Anthony R Kerlavage, Carol J Bult, Jean-Francois Tomb, Brian A Dougherty, Joseph M Merrick, et al. Whole-genome random sequencing and assembly of haemophilus influenzae Rd. *Science*, 269(5223):496–512, 1995.

[87] Travis Gagie, Giovanni Manzini, and Jouni Sirén. Wheeler graphs: A framework for BWT-based data structures. *Theoretical Computer Science*, 698:67–78, 2017.

[88] Emden R Gansner and Stephen C North. An open graph visualization system and its applications to software engineering. *Software: practice and experience*, 30(11): 1203–1233, 2000.

[89] Emden R Gansner, Eleftherios Koutsofios, Stephen C North, and K-P Vo. A technique for drawing directed graphs. *IEEE Transactions on Software Engineering*, 19(3):214–230, 1993.

[90] Richard C Gardner, Alan J Howarth, Peter Hahn, Marianne Brown-Luedi, Robert J Shepherd, and Joachim Messing. The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing. *Nucleic Acids Research*, 9(12):2871–2888, 1981.

[91] Erik Garrison and Gabor Marth. Haplotype-based variant detection from short-read sequencing. *arXiv:1207.3907*, 2012.

[92] Erik Garrison, Jouni Sirén, Adam M Novak, Glenn Hickey, Jordan M Eizenga, Eric T Dawson, William Jones, Shilpa Garg, Charles Markello, Michael F Lin, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, 36(9):875–879, 2018.

[93] Jay Ghurye, Arang Rhie, Brian P Walenz, Anthony Schmitt, Siddarth Selvaraj, Mihai Pop, Adam M Phillippy, and Sergey Koren. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *bioRxiv:261149*, 2018.

[94] André Goffeau, Bart G Barrell, Howard Bussey, RW Davis, Bernard Dujon, Heinz Feldmann, Francis Galibert, JD Hoheisel, Cr Jacq, Michael Johnston, et al. Life with 6000 genes. *Science*, 274(5287):546–567, 1996.

[95] Simon Gog, Timo Beller, Alistair Moffat, and Matthias Petri. From theory to practice: Plug and play with succinct data structures. In *13th International Symposium on Experimental Algorithms, (SEA 2014)*, pages 326–337, 2014.

[96] David Gordon, Chris Abajian, and Phil Green. Consed: a graphical tool for sequence finishing. *Genome Research*, 8(3):195–202, 1998.

[97] Osamu Gotoh. An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 162(3):705–708, 1982.

[98] Osamu Gotoh. Optimal sequence alignment allowing for long gaps. *Bulletin of Mathematical Biology*, 52(3):359–373, 1990.

[99] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology*, 29(7):644–652, 2011.

[100] Catherine Grasso and Christopher Lee. Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. *Bioinformatics*, 20(10):1546–1556, 2004.

[101] Richard E Green, Johannes Krause, Adrian W Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, Heng Li, Weiwei Zhai, Markus Hsi-Yang Fritz, et al. A draft sequence of the Neandertal genome. *Science*, 328(5979):710–722, 2010.

[102] Stuart J. Green, Reigh P. Monreal, Alan T. White, Thomas G. Bayer, Stuart J. Green, Reigh P. Monreal, Alan T. White, Thomas G. Bayer, Yasmin D. Arquiza, Alan T. White, Stuart J. Green, R. Buenaflor, and Jr. Nd Y. D. Arquiza. Phrap documentation, 1999.

[103] Roberto Grossi and Jeffrey Scott Vitter. Compressed suffix arrays and suffix trees with applications to text indexing and string matching. *SIAM Journal on Computing*, 35(2):378–407, 2005.

[104] Roberto Grossi, Ankur Gupta, and Jeffrey Scott Vitter. High-order entropy-compressed text indexes. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 841–850. Society for Industrial and Applied Mathematics, 2003.

[105] Sarah Guthrie, Abram Connelly, Peter Amstutz, Adam F Berrey, Nicolas Cesar, Jiahua Chen, Radhika Chippada, Tom Clegg, Bryan Cosca, Jiayong Li, et al. Tiling the genome into consistently named subsequences enables precision medicine and machine learning with millions of complex individual data-sets. Technical report, PeerJ PrePrints, 2015. URL https://peerj.com/preprints/1426/.

[106] Ronald Haentjens Dekker, Dirk Van Hulle, Gregor Middell, Vincent Neyt, and Joris Van Zundert. Computer-supported collation of modern manuscripts: Collatex and the beckett digital manuscript project. *Digital Scholarship in the Humanities*, 30(3):452–470, 2014.

[107] Robert S Harris. *Improved Pairwise Alignment of Genomic DNA*. PhD thesis, The Pennsylvania State University, 2007.

[108] Timothy D Harris, Phillip R Buzby, Hazen Babcock, Eric Beer, Jayson Bowers, Ido Braslavsky, Marie Causey, Jennifer Colonell, James DiMeo, J William Efcavitch, et al. Single-molecule DNA sequencing of a viral genome. *Science*, 320(5872): 106–109, 2008.

[109] Steffen Heber, Max Alekseyev, Sing-Hoi Sze, Haixu Tang, and Pavel A Pevzner. Splicing graphs and EST assembly problem. *Bioinformatics*, 18(suppl_1):S181–S188, 2002.

[110] Desmond G Higgins and Paul M Sharp. Clustal: a package for performing multiple sequence alignment on a microcomputer. *Gene*, 73(1):237–244, 1988.

[111] Jonathon T Hill, Bradley L Demarest, Brent W Bisgrove, Yi-Chu Su, Megan Smith, and H Joseph Yost. Poly peak parser: Method and software for identification of unknown indels using sanger sequencing of polymerase chain reaction products. *Developmental Dynamics*, 243(12):1632–1636, 2014.

[112] B. Howie, J. Marchini, and M. Stephens. Genotype imputation with thousands of genomes. *G3 (Bethesda)*, 1(6):457–470, 2011.

[113] Lin Huang, Victoria Popic, and Serafim Batzoglou. Short read alignment with populations of genomes. *Bioinformatics*, 29(13):i361–i370, 2013.

[114] Xiaoqiu Huang. A contig assembly program based on sensitive detection of fragment overlaps. *Genomics*, 14(1):18–25, 1992.

[115] Yao-Ting Huang and Chen-Fu Liao. Integration of string and de Bruijn graphs for genome assembly. *Bioinformatics*, 32(9):1301–1307, 2016.

[116] Julian Huxley. *Evolution: the modern synthesis*. George Allen and Unwin, 1942.

[117] Ramana M Idury and Michael S Waterman. A new algorithm for DNA sequence assembly. *Journal of Computational Biology*, 2(2):291–306, 1995.

[118] Z. Iqbal, M. Caccamo, I. Turner, P. Flicek, and G. McVean. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics*, 44(2): 226–232, 2012.

[119] Z. Iqbal, I. Turner, and G. McVean. High-throughput microbial population genomics using the Cortex variation assembler. *Bioinformatics*, 29(2):275–276, 2013.

[120] Zamin Iqbal, Mario Caccamo, Isaac Turner, Paul Flicek, and Gil McVean. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics*, 44(2):226–232, 2012.

[121] Marten Jäger, Max Schubach, Tomasz Zemojtel, Knut Reinert, Deanna M Church, and Peter N Robinson. Alternate-locus aware variant calling in whole genome sequencing. *Genome Medicine*, 8(1):130, 2016.

[122] 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, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, 36(4):338–345, 2018.

[123] Christine Jandrasits, Piotr W Dabrowski, Stephan Fuchs, and Bernhard Y Renard. seq-seq-pan: Building a computational pan-genome data structure on whole genome alignment. *BMC Genomics*, 19(1):47, 2018.

[124] Stefan Jänicke, Annette Geßner, Greta Franzini, Melissa Terras, Simon Mahony, and Gerik Scheuermann. Traviz: A visualization for variant graphs. *Digital Scholarship in the Humanities*, 30(suppl_1):i83–i99, 2015.

[125] W Min Jou, G Haegeman, M Ysebaert, and W Fiers. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature*, 237(5350):82–88, 1972.

[126] Arthur B Kahn. Topological sorting of large networks. *Communications of the ACM*, 5(11):558–562, 1962.

[127] Rolf G Karlsson and Patricio V Poblete. An $O(m \log \log D)$ algorithm for shortest paths. *Discrete Applied Mathematics*, 6(1):91–93, 1983.

[128] John J Kasianowicz, Eric Brandin, Daniel Branton, and David W Deamer. Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences*, 93(24):13770–13773, 1996.

[129] John Kececioglu. The maximum weight trace problem in multiple sequence alignment. In *Annual Symposium on Combinatorial Pattern Matching*, pages 106–119. Springer, 1993.

[130] John D Kececioglu and Eugene W Myers. Combinatorial algorithms for DNA sequence assembly. *Algorithmica*, 13(1-2):7–51, 1995.

[131] Birte Kehr, Kathrin Trappe, Manuel Holtgrewe, and Knut Reinert. Genome alignment with graph data structures: a comparison. *BMC Bioinformatics*, 15(1): 99, 2014.

[132] W James Kent. BLAT—the BLAST-like alignment tool. *Genome Research*, 12(4): 656–664, 2002.

[133] W James Kent and David Haussler. Assembly of the working draft of the human genome with GigAssembler. *Genome Research*, 11(9):1541–1548, 2001.

[134] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at UCSC. *Genome Research*, 12(6):996–1006, 2002.

[135] Paul Julian Kersey, James E Allen, Irina Armean, Sanjay Boddu, Bruce J Bolt, Denise Carvalho-Silva, Mikkel Christensen, Paul Davis, Lee J Falin, Christoph Grabmueller, et al. Ensembl genomes 2016: more genomes, more complexity. *Nucleic Acids Research*, 44(D1):D574–D580, 2015.

[136] Daehwan Kim, Ben Langmead, and Steven L Salzberg. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, 12(4):357–360, 2015.

[137] Daehwan Kim, B Langmead, and S Salzberg. HISAT2: graph-based alignment of next-generation sequencing reads to a population of genomes. https://github.com/infphilo/hisat2, 2017.

[138] Daniel C Koboldt, Ken Chen, Todd Wylie, David E Larson, Michael D McLellan, Elaine R Mardis, George M Weinstock, Richard K Wilson, and Li Ding. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, 25(17):2283–2285, 2009.

[139] Yuichi Kodama, Martin Shumway, and Rasko Leinonen. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Research*, 40(D1):D54–D56, 2011.

[140] Klaus-Peter Koepfli, Benedict Paten, Genome 10K Community of Scientists, and Stephen J O'Brien. The Genome 10K Project: a way forward. *Annual Review of Animal Biosciences*, 3(1):57–111, 2015.

[141] Sergey Koren, Brian P Walenz, Konstantin Berlin, Jason R Miller, Nicholas H Bergman, and Adam M Phillippy. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, 27(5): 722–736, 2017.

[142] Jonas Korlach, Patrick J Marks, Ronald L Cicero, Jeremy J Gray, Devon L Murphy, Daniel B Roitman, Thang T Pham, Geoff A Otto, Mathieu Foquet, and Stephen W Turner. Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proceedings of the National Academy of Sciences*, 105(4):1176–1181, 2008.

[143] Anna Kuosmanen, Topi Paavilainen, Travis Gagie, Rayan Chikhi, Alexandru Tomescu, and Veli Mäkinen. Using minimum path cover to boost dynamic programming on DAGs: co-linear chaining extended. In *International Conference on Research in Computational Molecular Biology*, pages 105–121. Springer, 2018.

[144] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

[145] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

[146] Iosif Lazaridis, Nick Patterson, Alissa Mittnik, Gabriel Renaud, Swapan Mallick, Karola Kirsanow, Peter H Sudmant, Joshua G Schraiber, Sergi Castellano, Mark Lipson, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513(7518):409–413, 2014.

[147] Christopher Lee, Catherine Grasso, and Mark F Sharlow. Multiple sequence alignment using partial order graphs. *Bioinformatics*, 18(3):452–464, 2002.

[148] Wan-Ping Lee, Michael P Stromberg, Alistair Ward, Chip Stewart, Erik P Garrison, and Gabor T Marth. Mosaik: A hash-based algorithm for accurate next-generation sequencing short-read mapping. *PLoS ONE*, 9(3):e90581, 2014.

[149] Rasko Leinonen, Hideaki Sugawara, Martin Shumway, and International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Research*, 39(suppl_1):D19–D21, 2010.

[150] Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 31(10):1674–1676, 2015.

[151] Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, 2011.

[152] Heng Li. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics*, 28(14):1838–1844, 2012.

[153] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997*, 2013.

[154] Heng Li. Fast construction of FM-index for long sequence reads. *Bioinformatics*, 30(22):3274–3275, 2014.

[155] Heng Li. Fermikit: assembly-based variant calling for Illumina resequencing data. *Bioinformatics*, 31(22):3694–3696, 2015.

[156] Heng Li. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, 32(14):2103–2110, 2016.

[157] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.

[158] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[159] Heng Li and Richard Durbin. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, 26(5):589–595, 2010.

[160] Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11):1851–1858, 2008.

[161] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[162] Ruiqiang Li, Yingrui Li, Karsten Kristiansen, and Jun Wang. SOAP: short oligonucleotide alignment program. *Bioinformatics*, 24(5):713–714, 2008.

[163] Ruiqiang Li, Yingrui Li, Hancheng Zheng, Ruibang Luo, Hongmei Zhu, Qibin Li, Wubin Qian, Yuanyuan Ren, Geng Tian, Jinxiang Li, et al. Building the sequence map of the human pan-genome. *Nature Biotechnology*, 28(1):57–63, 2010.

[164] Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

[165] DJ Lightfoot, David Erwin Jarvis, T Ramaraj, R Lee, EN Jellen, and PJ Maughan. Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (Amaranthus hypochondriacus) chromosomes provide insights into genome evolution. *BMC Biology*, 15(1):74, 2017.

[166] Henry W Lin and Max Tegmark. Critical behavior in physics and probabilistic formal languages. *Entropy*, 19(7):299, 2017.

[167] Vivian Link, Athanasios Kousathanas, Krishna Veeramah, Christian Sell, Amelie Scheu, and Daniel Wegmann. ATLAS: analysis tools for low-depth and ancient samples. *bioRxiv:105346*, 2017.

[168] David J Lipman, Stephen F Altschul, and John D Kececioglu. A tool for multiple sequence alignment. *Proceedings of the National Academy of Sciences*, 86(12): 4412–4415, 1989.

[169] Gianni Liti, David M Carter, Alan M Moses, Jonas Warringer, Leopold Parts, Stephen A James, Robert P Davey, Ian N Roberts, Austin Burt, Vassiliki Koufopanou, et al. Population genomics of domestic and wild yeasts. *Nature*, 458 (7236):337–341, 2009.

[170] Bo Liu, Hongzhe Guo, Michael Brudno, and Yadong Wang. deBGA: read alignment with de Bruijn graph-based seed and extension. *Bioinformatics*, 32(21):3224–3232, 2016.

[171] Nicholas J Loman, Joshua Quick, and Jared T Simpson. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods*, 12(8):733–735, 2015.

[172] Sorina Maciuca, Carlos del Ojo Elias, Gil McVean, and Zamin Iqbal. A natural encoding of genetic variation in a Burrows-Wheeler Transform to enable mapping and genome inference. In *International Workshop on Algorithms in Bioinformatics*, pages 222–233. Springer, 2016.

[173] Tanja Magoč and Steven L Salzberg. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21):2957–2963, 2011.

[174] Michael A Mahowald, Federico E Rey, Henning Seedorf, Peter J Turnbaugh, Robert S Fulton, Aye Wollam, Neha Shah, Chunyan Wang, Vincent Magrini, Richard K Wilson, et al. Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla. *Proceedings of the National Academy of Sciences*, 106(14):5859–5864, 2009.

[175] BWJ Mahy, JJ Esposito, and JC Venter. Sequencing the smallpox virus genome: prelude to destruction of a virus species. *ASM News*, 57:577–580, 1991.

[176] Udi Manber and Gene Myers. Suffix arrays: a new method for on-line string searches. *SIAM Journal on Computing*, 22(5):935–948, 1993.

[177] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, Michael S Braverman, Yi-Ju Chen, Zhoutao Chen, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.

[178] Gabor T Marth, Ian Korf, Mark D Yandell, Raymond T Yeh, Zhijie Gu, Hamideh Zakeri, Nathan O Stitziel, LaDeana Hillier, Pui-Yan Kwok, and Warren R Gish. A general approach to single-nucleotide polymorphism discovery. *Nature Genetics*, 23 (4):452–456, 1999.

[179] Jeffrey A Martin and Zhong Wang. Next-generation transcriptome assembly. *Nature Reviews Genetics*, 12(10):671–682, 2011.

[180] Rui Martiniano, Anwen Caffell, Malin Holst, Kurt Hunter-Mann, Janet Montgomery, Gundula Müldner, Russell L McLaughlin, Matthew D Teasdale, Wouter Van Rheenen, Jan H Veldink, et al. Genomic signals of migration and continuity in Britain before the Anglo-Saxons. *Nature Communications*, 7:10326, 2016.

[181] Ryan McDaniell, Bum-Kyu Lee, Lingyun Song, Zheng Liu, Alan P Boyle, Michael R Erdos, Laura J Scott, Mario A Morken, Katerina S Kucera, Anna Battenhouse, et al. Heritable individual-specific and allele-specific chromatin signatures in humans. *Science*, 328(5975):235–239, 2010.

[182] Duccio Medini, Claudio Donati, Herve Tettelin, Vega Masignani, and Rino Rappuoli. The microbial pan-genome. *Current Opinion in Genetics & Development*, 15(6): 589–594, 2005.

[183] Gregor Mendel. Versuche über Pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brunn*, 44:1–47, 1866.

[184] Androniki Menelaou and Jonathan Marchini. Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. *Bioinformatics*, 29(1): 84–91, 2012.

[185] Alexander S Mikheyev and Mandy MY Tin. A first look at the oxford nanopore minion sequencer. *Molecular Ecology Resources*, 14(6):1097–1102, 2014.

[186] Jason R Miller, Arthur L Delcher, Sergey Koren, Eli Venter, Brian P Walenz, Anushka Brownley, Justin Johnson, Kelvin Li, Clark Mobarry, and Granger Sutton. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*, 24(24): 2818–2824, 2008.

[187] Ilia Minkin, Son Pham, and Paul Medvedev. TwoPaCo: An efficient algorithm to build the compacted de Bruijn graph from many complete genomes. *Bioinformatics*, 33(24):4024–4032, 2016.

[188] Robi D Mitra, Jay Shendure, Jerzy Olejnik, George M Church, et al. Fluorescent in situ sequencing on polymerase colonies. *Analytical Biochemistry*, 320(1):55–65, 2003.

[189] Anthony P Monaco and Zoia Larin. YACs, BACs, PACs and MACs: artificial chromosomes as research tools. *Trends in Biotechnology*, 12(7):280–286, 1994.

[190] Burkhard Morgenstern, Andreas Dress, and Thomas Werner. Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proceedings of the National Academy of Sciences*, 93(22):12098–12103, 1996.

[191] Yulia Mostovoy, Michal Levy-Sakin, Jessica Lam, Ernest T Lam, Alex R Hastie, Patrick Marks, Joyce Lee, Catherine Chu, Chin Lin, Željko Džakula, et al. A hybrid approach for de novo human genome sequence assembly and phasing. *Nature Methods*, 13(7):587–590, 2016.

[192] Martin D Muggli, Alexander Bowe, Noelle R Noyes, Paul S Morley, Keith E Belk, Robert Raymond, Travis Gagie, Simon J Puglisi, and Christina Boucher. Succinct colored de Bruijn graphs. *Bioinformatics*, 33(20):3181–3187, 2017.

[193] Eugene W Myers. Toward simplifying and accurately formulating fragment assembly. *Journal of Computational Biology*, 2(2):275–290, 1995.

[194] Eugene W Myers. The fragment assembly string graph. *Bioinformatics*, 21 (suppl_2):ii79–ii85, 2005.

[195] Eugene W Myers. Efficient local alignment discovery amongst noisy long reads. In *International Workshop on Algorithms in Bioinformatics*, pages 52–67. Springer, 2014.

[196] Eugene W Myers and Webb Miller. Optimal alignments in linear space. *Bioinformatics*, 4(1):11–17, 1988.

[197] Eugene W Myers and Webb Miller. Approximate matching of regular expressions. *Bulletin of Mathematical Biology*, 51(1):5–37, 1989.

[198] Eugene W Myers, Granger G Sutton, Art L Delcher, Ian M Dew, Dan P Fasulo, Michael J Flanigan, Saul A Kravitz, Clark M Mobarry, Knut HJ Reinert, Karin A Remington, et al. A whole-genome assembly of drosophila. *Science*, 287(5461): 2196–2204, 2000.

[199] Eugene W Myers, Granger G Sutton, Hamilton O Smith, Mark D Adams, and J Craig Venter. On the sequencing and assembly of the human genome. *Proceedings of the National Academy of Sciences*, 99(7):4145–4146, 2002.

[200] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.

[201] Cédric Notredame, Desmond G Higgins, and Jaap Heringa. T-coffee: a novel method for fast and accurate multiple sequence alignment1. *Journal of Molecular Biology*, 302(1):205–217, 2000.

[202] Adam M Novak, Erik Garrison, and Benedict Paten. A graph extension of the positional Burrows–Wheeler transform and its applications. *Algorithms for Molecular Biology*, 12:18, 2017.

[203] Adam M Novak, Glenn Hickey, Erik Garrison, Sean Blum, Abram Connelly, Alexander Dilthey, Jordan Eizenga, MA Saleh Elmohamed, Sally Guthrie, André Kahles, et al. Genome graphs. *bioRxiv:101378*, 2017.

[204] Sergey Nurk, Dmitry Meleshko, Anton Korobeynikov, and Pavel A Pevzner. metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, 27(5):824–834, 2017.

[205] Genome 10K Community of Scientists. Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. *Journal of Heredity*, 100(6):659–674, 2009.

[206] Yukiteru Ono, Kiyoshi Asai, and Michiaki Hamada. PBSIM: PacBio reads simulator—toward accurate genome assembly. *Bioinformatics*, 29(1):119–121, 2012.

[207] Taku Onodera, Kunihiko Sadakane, and Tetsuo Shibuya. Detecting superbubbles in assembly graphs. In *International Workshop on Algorithms in Bioinformatics*, pages 338–348. Springer, 2013.

[208] R Padmanabhan, Ernest Jay, and Ray Wu. Chemical synthesis of a primer and its use in the sequence analysis of the lysozyme gene of bacteriophage t4. *Proceedings of the National Academy of Sciences*, 71(6):2510–2514, 1974.

[209] Andrew J Page, Carla A Cummins, Martin Hunt, Vanessa K Wong, Sandra Reuter, Matthew TG Holden, Maria Fookes, Daniel Falush, Jacqueline A Keane, and Julian Parkhill. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31(22):3691–3693, 2015.

[210] Benedict Paten, Javier Herrero, Kathryn Beal, Stephen Fitzgerald, and Ewan Birney. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Research*, 18(11):1814–1828, 2008.

[211] Benedict Paten, Mark Diekhans, Dent Earl, John St John, Jian Ma, Bernard Suh, and David Haussler. Cactus graphs for genome comparisons. *Journal of Computational Biology*, 18(3):469–481, 2011.

[212] Benedict Paten, Dent Earl, Ngan Nguyen, Mark Diekhans, Daniel Zerbino, and David Haussler. Cactus: Algorithms for genome multiple sequence alignment. *Genome Research*, 21(9):1512–1528, 2011.

[213] Benedict Paten, Jordan M Eizenga, Yohei M Rosen, Adam M Novak, Erik Garrison, and Glenn Hickey. Superbubbles, ultrabubbles, and cacti. *Journal of Computational Biology*, 25(7):649–663, 2018.

[214] N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich. Ancient admixture in human history. *Genetics*, 192(3): 1065–1093, 2012.

[215] William R Pearson and David J Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85(8):2444–2448, 1988.

[216] Jane Peterson, Susan Garges, Maria Giovanni, Pamela McInnes, Lu Wang, Jeffery A Schloss, Vivien Bonazzi, Jean E McEwen, Kris A Wetterstrand, Carolyn Deal, et al. The NIH human microbiome project. *Genome Research*, 19(12):2317–2323, 2009.

[217] Pavel A Pevzner, Haixu Tang, and Michael S Waterman. An eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences*, 98 (17):9748–9753, 2001.

[218] Ryan Poplin, Dan Newburger, Jojo Dijamco, Nam Nguyen, Dion Loy, Sam S Gross, Cory Y McLean, and Mark A DePristo. Creating a universal snp and small indel variant caller with deep neural networks. *bioRxiv:092890*, 2017.

[219] David Porubskỳ, Ashley D Sanders, Niek Van Wietmarschen, Ester Falconer, Mark Hills, Diana CJ Spierings, Marianna R Bevova, Victor Guryev, and Peter M Lansdorp. Direct chromosome-length haplotyping by single-cell sequencing. *Genome Research*, 26(11):1565–1574, 2016.

[220] Kay Prüfer. snpAD: an ancient DNA genotype caller. *Bioinformatics*, 2018. doi: 10.1093/bioinformatics/bty507.

[221] Robert F Purnell, Kunal K Mehta, and Jacob J Schmidt. Nucleotide identification and orientation discrimination of DNA homopolymers immobilized in a protein nanopore. *Nano Letters*, 8(9):3029–3034, 2008.

[222] Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[223] Goran Rakocevic, Vladimir Semenyuk, James Spencer, John Browning, Ivan Johnson, Vladan Arsenijevic, Jelena Nadj, Kaushik Ghose, Maria C Suciu, Sun-Gou Ji, et al. Fast and accurate genomic analyses using genome graphs. *bioRxiv:194530*, 2018.

[224] Rajeev Raman, Venkatesh Raman, and S Srinivasa Rao. Succinct indexable dictionaries with applications to encoding k-ary trees and multisets. In *Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 233–242. Society for Industrial and Applied Mathematics, 2002.

[225] Benjamin Raphael, Degui Zhi, Haixu Tang, and Pavel Pevzner. A novel method for multiple alignment of sequences with repeated and shuffled elements. *Genome Research*, 14(11):2336–2346, 2004.

[226] Tobias Rausch, Anne-Katrin Emde, David Weese, Andreas Döring, Cedric Notredame, and Knut Reinert. Segment-based multiple sequence alignment. *Bioinformatics*, 24(16):i187–i192, 2008.

[227] Mikko Rautiainen, Veli Mäkinen, and Tobias Marschall. Bit-parallel sequence-to-graph alignment. *bioRxiv:323063*, 2018.

[228] Gabriel Renaud, Kristian Hanghøj, Eske Willerslev, and Ludovic Orlando. gargammel: a sequence simulator for ancient DNA. *Bioinformatics*, 33(4):577–579, 2016.

[229] Anthony Rhoads and Kin Fai Au. PacBio sequencing and its applications. *Genomics, Proteomics & Bioinformatics*, 13(5):278–289, 2015.

[230] Andy Rimmer, Hang Phan, Iain Mathieson, Zamin Iqbal, Stephen RF Twigg, Andrew OM Wilkie, Gil McVean, Gerton Lunter, WGS500 Consortium, et al. Integrating mapping-, assembly-and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*, 46(8):912–918, 2014.

[231] Gordon Robertson, Jacqueline Schein, Readman Chiu, Richard Corbett, Matthew Field, Shaun D Jackman, Karen Mungall, Sam Lee, Hisanaga Mark Okada, Jenny Q Qian, et al. De novo assembly and analysis of RNA-seq data. *Nature Methods*, 7 (11):909–912, 2010.

[232] Michael G Ross, Carsten Russ, Maura Costello, Andrew Hollinger, Niall J Lennon, Ryan Hegarty, Chad Nusbaum, and David B Jaffe. Characterizing and measuring bias in sequence data. *Genome Biology*, 14(5):R51, 2013.

[233] Nicole Rusk. Torrents of sequence. *Nature Methods*, 8(1):44–44, 2010.

[234] F Sanger, GG Brownlee, and BG Barrell. A two-dimensional fractionation procedure for radioactive nucleotides. *Journal of Molecular Biology*, 13(2):373–398, 1965.

[235] F Sanger, Ar R Coulson, GF Hong, DF Hill, and GB d Petersen. Nucleotide sequence of bacteriophage λ DNA. *Journal of Molecular Biology*, 162(4):729–773, 1982.

[236] Frederick Sanger and Hans Tuppy. The amino-acid sequence in the phenylalanyl chain of insulin. 1. the identification of lower peptides from partial hydrolysates. *Biochemical Journal*, 49(4):463–481, 1951.

[237] Frederick Sanger, Gilian M Air, Bart G Barrell, Nigel L Brown, Alan R Coulson, John C Fiddes, Clyde A Hutchison III, Patrick M Slocombe, and Mo Smith. Nucleotide sequence of bacteriophage φx174 DNA. *Nature*, 265(5596):687–695, 1977.

[238] Frederick Sanger, Steven Nicklen, and Alan R Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74 (12):5463–5467, 1977.

[239] Stephan Schiffels, Wolfgang Haak, Pirita Paajanen, Bastien Llamas, Elizabeth Popescu, Louise Loe, Rachel Clarke, Alice Lyons, Richard Mortimer, Duncan Sayer, et al. Iron age and Anglo-Saxon genomes from East England reveal British migration history. *Nature Communications*, 7:10408, 2016.

[240] Michael Schmid, Daniel Frei, Andrea Patrignani, Ralph Schlapbach, Juerg E Frey, Mitja NP Remus-Emsermann, and Christian H Ahrens. Pushing the limits of de novo genome assembly for complex prokaryotic genomes harboring very long, near identical repeats. *bioRxiv:300186*, 2018.

[241] Desmond Schmidt and Robert Colomb. A data structure for representing multi-version texts online. *International Journal of Human-Computer Studies*, 67(6): 497–514, 2009.

[242] Holger Schmitt, Ung-Jin Kim, Tatiana Slepak, Nikolaus Blin, Melvin I Simon, and Hiroaki Shizuya. Framework for a physical map of the human 22q13 region using bacterial artificial chromosomes (BACs). *Genomics*, 33(1):9–20, 1996.

[243] Korbinian Schneeberger, Jörg Hagmann, Stephan Ossowski, Norman Warthmann, Sandra Gesing, Oliver Kohlbacher, and Detlef Weigel. Simultaneous alignment of short reads against multiple genomes. *Genome Biology*, 10(9):R98, 2009.

[244] Valerie A Schneider, Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A Kitts, Terence D Murphy, Kim D Pruitt, Françoise Thibaud-Nissen, Derek Albracht, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, 27 (5):849–864, 2017.

[245] Marcel H Schulz, Daniel R Zerbino, Martin Vingron, and Ewan Birney. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28(8):1086–1092, 2012.

[246] Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145, 2008.

[247] Jay Shendure, Gregory J Porreca, Nikos B Reppas, Xiaoxia Lin, John P McCutcheon, Abraham M Rosenbaum, Michael D Wang, Kun Zhang, Robi D Mitra, and George M Church. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309(5741):1728–1732, 2005.

[248] Stephen T Sherry, M-H Ward, M Kholodov, J Baker, Lon Phan, Elizabeth M Smigielski, and Karl Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311, 2001.

[249] Jonas Andreas Sibbesen, Lasse Maretty, The Danish Pan-Genome Consortium, and Anders Krogh. Accurate genotyping across variant classes and lengths using variant graphs. *Nature Genetics*, 50:1054–1059, 2018.

[250] F Sigaux. Cancer genome or the development of molecular portraits of tumors. *Bulletin de L'Academie Nationale de Medecine*, 184(7):1441–7, 2000.

[251] Birgit Sikkema-Raddatz, Lennart F Johansson, Eddy N de Boer, Rowida Almomani, Ludolf G Boven, Maarten P van den Berg, Karin Y van Spaendonck-Zwarts, J Peter van Tintelen, Rolf H Sijmons, Jan DH Jongbloed, et al. Targeted next-generation sequencing can replace Sanger sequencing in clinical diagnostics. *Human Mutation*, 34(7):1035–1042, 2013.

[252] Jared T Simpson. Exploring genome characteristics and sequence quality without a reference. *Bioinformatics*, 30(9):1228–1235, 2014.

[253] Jared T Simpson and Richard Durbin. Efficient construction of an assembly string graph using the FM-index. *Bioinformatics*, 26(12):i367–373, 2010.

[254] Jared T Simpson and Richard Durbin. Efficient de novo assembly of large genomes using compressed data structures. *Genome Research*, 22(3):549–556, 2012.

[255] Jared T Simpson, Kim Wong, Shaun D Jackman, Jacqueline E Schein, Steven JM Jones, and Inanç Birol. ABySS: a parallel assembler for short read sequence data. *Genome Research*, 19(6):1117–1123, 2009.

[256] Jouni Sirén. Indexing variation graphs. In *2017 Proceedings of the ninteenth workshop on algorithm engineering and experiments (ALENEX)*, pages 13–27. SIAM, 2017.

[257] Jouni Sirén, Niko Välimäki, and Veli Mäkinen. Indexing finite language representation of population genotypes. In *International Workshop on Algorithms in Bioinformatics*, pages 270–281. Springer, 2011.

[258] Jouni Sirén, Niko Välimäki, and Veli Mäkinen. Indexing graphs for path queries with applications in genome research. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 11(2):375–388, 2014.

[259] Jouni Sirén, Erik Garrison, Adam M Novak, Benedict Paten, and Richard Durbin. Haplotype-aware graph indexes. *arXiv:1805.03834*, 2018.

[260] Temple F Smith. Functional genomics—bioinformatics is ready for the challenge. *Trends in Genetics*, 14(7):291–293, 1998.

[261] Temple F Smith and Michael S Waterman. Comparison of biosequences. *Advances in Applied Mathematics*, 2(4):482–489, 1981.

[262] Rodger Staden. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research*, 6(7):2601–2610, 1979.

[263] Kraig R Stevenson, Joseph D Coolon, and Patricia J Wittkopp. Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. *BMC Genomics*, 14(1):536, 2013.

[264] Erich C Strauss, Joan A Kobori, Gerald Siu, and Leroy E Hood. Specific-primer-directed DNA sequencing. *Analytical Biochemistry*, 154(1):353–360, 1986.

[265] P. H. Sudmant, J. O. Kitzman, F. Antonacci, C. Alkan, M. Malig, A. Tsalenko, N. Sampas, L. Bruhn, J. Shendure, 1000 Genomes Project, and E. E. Eichler. Diversity of human copy number variation and multicopy genes. *Science*, 330(6004): 641–646, 2010.

[266] Peter H Sudmant, Tobias Rausch, Eugene J Gardner, Robert E Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, Kai Ye, Goo Jun, Markus Hsi-Yang Fritz, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81, 2015.

[267] Granger G Sutton, Owen White, Mark D Adams, and Anthony R Kerlavage. TIGR Assembler: A new tool for assembling large shotgun sequencing projects. *Genome Science and Technology*, 1(1):9–19, 1995.

[268] Hajime Suzuki and Masahiro Kasahara. Acceleration of nucleotide semi-global alignment with adaptive banded dynamic programming. *bioRxiv:130633*, 2017.

[269] Artem Tarasov, Albert J Vilella, Edwin Cuppen, Isaac J Nijman, and Pjotr Prins. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, 31(12): 2032–2034, 2015.

[270] Robert Tarjan. Depth-first search and linear graph algorithms. *SIAM Journal on Computing*, 1(2):146–160, 1972.

[271] Aaron E Tenney, Jia Qian Wu, Laura Langton, Paul Klueh, Ralph Quatrano, and Michael R Brent. A tale of two templates: Automatically resolving double traces has many applications, including efficient pcr-based elucidation of alternative splices. *Genome Research*, 17(2):212–218, 2007.

[272] Hervé Tettelin, Vega Masignani, Michael J Cieslewicz, Claudio Donati, Duccio Medini, Naomi L Ward, Samuel V Angiuoli, Jonathan Crabtree, Amanda L Jones, A Scott Durkin, et al. Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*, 102(39):13950–13955, 2005.

[273] Julie D Thompson, Desmond G Higgins, and Toby J Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–4680, 1994.

[274] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with tophat and cufflinks. *Nature Protocols*, 7(3):562–578, 2012.

[275] Susannah Green Tringe, Christian Von Mering, Arthur Kobayashi, Asaf A Salamov, Kevin Chen, Hwai W Chang, Mircea Podar, Jay M Short, Eric J Mathur, John C Detter, et al. Comparative metagenomics of microbial communities. *Science*, 308 (5721):554–557, 2005.

[276] Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire M Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. The human microbiome project. *Nature*, 449(7164): 804–810, 2007.

[277] Isaac Turner, Kiran V Garimella, Zamin Iqbal, and Gil McVean. Integrating long-range connectivity information into de Bruijn graphs. *Bioinformatics*, 34(15): 2556–2565, 2018.

[278] Esko Ukkonen. On-line construction of suffix trees. *Algorithmica*, 14(3):249–260, 1995.

[279] Daniel Valenzuela and Veli Mäkinen. CHIC: a short read aligner for pan-genomic references. *bioRxiv:178129*, 2017.

[280] Robert Vaser, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, 27(5): 737–746, 2017.

[281] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

[282] George Vernikos, Duccio Medini, David R Riley, and Herve Tettelin. Ten years of pan-genome analyses. *Current Opinion in Microbiology*, 23:148–154, 2015.

[283] Y. Wang, J. Lu, J. Yu, R. A. Gibbs, and F. Yu. An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. *Genome Research*, 23(5):833–842, 2013.

[284] Robert H Waterston, Eric S Lander, and John E Sulston. On the sequencing of the human genome. *Proceedings of the National Academy of Sciences*, 99(6):3712–3716, 2002.

[285] James D Watson, Francis HC Crick, et al. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

[286] Detlef Weigel and Richard Mott. The 1001 genomes project for Arabidopsis thaliana. *Genome Biology*, 10(5):107, 2009.

[287] Peter Weiner. Linear pattern matching algorithms. In *Switching and Automata Theory, 1973. SWAT'08. IEEE Conference Record of 14th Annual Symposium on*, pages 1–11. IEEE, 1973.

[288] David A Wheeler, Maithreyan Srinivasan, Michael Egholm, Yufeng Shen, Lei Chen, Amy McGuire, Wen He, Yi-Ju Chen, Vinod Makhijani, G Thomas Roth, et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452(7189):872–876, 2008.

[289] Ryan R Wick, Mark B Schultz, Justin Zobel, and Kathryn E Holt. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*, 31(20): 3350–3352, 2015.

[290] Ray Wu. Nucleotide sequence analysis of DNA. *Nature New Biology*, 236(68): 198–200, 1972.

[291] Sun Wu, Udi Manber, and Eugene Myers. A subquadratic algorithm for approximate regular expression matching. *Journal of Algorithms*, 19(3):346–360, 1995.

[292] Jia-Xing Yue, Jing Li, Louise Aigrain, Johan Hallin, Karl Persson, Karen Oliver, Anders Bergström, Paul Coupland, Jonas Warringer, Marco Cosentino Lagomarsino, et al. Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nature Genetics*, 49(6):913–924, 2017.

[293] Georg Zeller, Richard M Clark, Korbinian Schneeberger, Anja Bohlen, Detlef Weigel, and Gunnar Rätsch. Detecting polymorphic regions in Arabidopsis thaliana with resequencing microarrays. *Genome Research*, 18(6):918–929, 2008.

[294] Daniel Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5):821–829, 2008.

[295] Mengyao Zhao, Wan-Ping Lee, Erik P Garrison, and Gabor T Marth. SSW library: An SIMD Smith-Waterman C/C++ library for use in genomic applications. *PLoS ONE*, 8(12):e82138, 2013.

[296] Grace XY Zheng, Billy T Lau, Michael Schnall-Levin, Mirna Jarosz, John M Bell, Christopher M Hindson, Sofia Kyriazopoulou-Panagiotopoulou, Donald A Masquelier, Landon Merrill, Jessica M Terry, et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology*, 34 (3):303–311, 2016.

[297] Boyan Zhou, Shaoqing Wen, Lingxiang Wang, Li Jin, Hui Li, and Hong Zhang. Antcaller: an accurate variant caller incorporating ancient DNA damage. *Molecular Genetics and Genomics*, 292(6):1419–1430, 2017.

[298] Jacob Ziv and Abraham Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, 1977.

[299] Justin M Zook, Brad Chapman, Jason Wang, David Mittelman, Oliver Hofmann, Winston Hide, and Marc Salit. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature Biotechnology*, 32(3): 246–251, 2014.

[300] Justin M Zook, David Catoe, Jennifer McDaniel, Lindsay Vang, Noah Spies, Arend Sidow, Ziming Weng, Yuling Liu, Christopher E Mason, Noah Alexander, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data*, 3:160025, 2016.

Related publications

During my graduate studies I have been an author on a number of publications which are related to this work. These include the following:

- Garrison, Erik, Jouni Sirén, Adam M. Novak, Glenn Hickey, Jordan M. Eizenga, Eric T. Dawson, William Jones et al. "Variation graph toolkit improves read mapping by representing genetic variation in the reference." *Nature Biotechnology*, 36(9):875–879, (2018).
- Paten, Benedict, Jordan M. Eizenga, Yohei M. Rosen, Adam M. Novak, Erik Garrison, and Glenn Hickey. "Superbubbles, ultrabubbles, and cacti." *Journal of Computational Biology*, 25(7):649–663, (2018).
- Garg, Shilpa, Mikko Rautiainen, Adam M. Novak, Erik Garrison, Richard Durbin, and Tobias Marschall. "A graph-based approach to diploid genome assembly." *Bioinformatics*, 34(13):i105–i114, (2018).
- Sirén, Jouni, Erik Garrison, Adam M. Novak, Benedict Paten, and Richard Durbin. "Haplotype-aware graph indexes." *arXiv:1805.03834* (2018).
- Paten, Benedict, Adam M. Novak, Jordan M. Eizenga, and Erik Garrison. "Genome graphs and the evolution of genome inference." *Genome Research*, 27(5):665–676, (2017).
- Novak, Adam M., Glenn Hickey, Erik Garrison, Sean Blum, Abram Connelly, Alexander Dilthey, Jordan Eizenga et al. "Genome graphs." *bioRxiv:101378* (2017).
- Computational pan-genomics consortium. "Computational pan-genomics: status, promises and challenges." *Briefings in Bioinformatics*, 19(1):118–135, (2016).
- Novak, Adam M., Erik Garrison, and Benedict Paten. "A graph extension of the positional Burrows–Wheeler transform and its applications." *Algorithms for Molecular Biology*, 12:18, (2017).