References

1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010. ISSN 0028-0836. doi:10.1038/nature09534. URL http://dx.doi.org/10.1038/nature09534. 5, 15, 115

Abouelhoda, Mohamed I., Kurtz, Stefan, and Ohlebusch, Enno. The enhanced suffix array and its applications to genome analysis algorithms in bioinformatics. In Guigó, Roderic and Gusfield, Dan, editors, *Algorithms in Bioinformatics*, volume 2452 of *Lecture Notes in Computer Science*, chapter 35, pages 449–463. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2002. ISBN 978-3-540-44211-0. doi:10.1007/3-540-45784-4_35. URL http://dx.doi.org/10.1007/3-540-45784-4_35. 16

Abouelhoda, Mohamed I., Kurtz, Stefan, and Ohlebusch, Enno. Replacing suffix trees with enhanced suffix arrays. *Journal of Discrete Algorithms*, 2(1):53–86, 2004. ISSN 15708667. doi:10.1016/S1570-8667(03)00065-0. URL http://dx.doi.org/10.1016/S1570-8667(03)00065-0. 22

Adams, Mark D., Sutton, Granger G., Smith, Hamilton O., Myers, Eugene W., and Venter, J. Craig. The independence of our genome assemblies. *Proceedings of the National Academy of Sciences*, 100(6):3025–3026, 2003. ISSN 1091-6490. doi:10.1073/pnas.0637478100. URL http://dx.doi.org/10.1073/pnas.0637478100. 4

Albers, Cornelis A., Lunter, Gerton, MacArthur, Daniel G., McVean, Gilean, Ouwehand, Willem H., and Durbin, Richard. Dindel: Accurate indel calls from short-read data. *Genome Research*, 21(6):961–973, 2011. ISSN 1549-5469. doi:10.1101/gr.112326.110. URL http://dx.doi.org/10.1101/gr.112326.110. 88, 105

Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, 408(6814):796–815, 2000. ISSN 0028-0836. doi:10.1038/35048692. URL http://dx.doi.org/10.1038/35048692. 3

Barnett, Derek W., Garrison, Erik K., Quinlan, Aaron R., Strömberg, Michael P., and Marth, Gabor T. BamTools: a c++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, 27(12):1691–1692, 2011. ISSN 1460-2059. doi:10.1093/bioinformatics/btr174. URL http://dx.doi.org/10.1093/bioinformatics/btr174. 58

Bauer, Markus J., Cox, Anthony J., and Rosone, Giovanna. Lightweight BWT construction for very large string collections combinatorial pattern matching. volume 6661 of *Lecture Notes in Computer Science*, chapter 20, pages 219–231. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-21457-8. doi:10.1007/978-3-642-21458-5_20. URL http://dx.doi.org/10.1007/978-3-642-21458-5_20. 32, 46

Bentley, David R., Balasubramanian, Shankar, Swerdlow, Harold P., Smith, Geoffrey P., Milton, John, Brown, Clive G., Hall, Kevin P., Evers, Dirk J., Barnes, Colin L., Bignell, Helen R., et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008. ISSN 1476-4687. doi:10.1038/nature07517. URL http://dx.doi.org/10.1038/nature07517. 5, 47

Bentley, Jon L. and Sedgewick, Robert. Fast algorithms for sorting and searching strings. In *SODA '97: Proceedings of the eighth annual ACM-SIAM symposium on Discrete algorithms*, pages 360–369. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1997. ISBN 0-89871-390-0. URL http://portal.acm.org/citation.cfm?id=314321. 32

Berger, Emery D., McKinley, Kathryn S., Blumofe, Robert D., and Wilson, Paul R. Hoard: a scalable memory allocator for multithreaded applications. *SIGPLAN Not.*, 35(11):117–128, 2000. ISSN 0362-1340. doi:10.1145/356989.357000. URL http://dx.doi.org/10.1145/356989.357000. 58

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., et al. The complete genome sequence of escherichia coli k-12. *Science (New York, N.Y.)*, 277(5331):1453–1462, 1997. ISSN 0036-8075. doi:10.1126/science.277.5331.1453. URL http://dx.doi.org/10.1126/science.277.5331.1453. 3

Burrows, M. and Wheeler, D. J. A block-sorting lossless data compression algorithm. Technical Report 124, 1994. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.37.6774. 16, 28

Butler, Jonathan, MacCallum, Iain, Kleber, Michael, Shlyakhter, Ilya A., Belmonte, Matthew K., Lander, Eric S., Nusbaum, Chad, and Jaffe, David B. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome research*, 18(5):810–820, 2008. ISSN 1088-9051. doi:10.1101/gr.7337908. URL http://dx.doi.org/10.1101/gr.7337908. 13

C. elegans Sequencing Consortium. Genome sequence of the nematode c. elegans: a platform for investigating biology. *Science (New York, N.Y.)*, 282(5396):2012–2018, 1998. ISSN 0036-8075. doi:10.1126/science.282.5396.2012. URL http://dx.doi.org/10.1126/science.282.5396.2012. 3, 61

Catchen, Julian M., Amores, Angel, Hohenlohe, Paul, Cresko, William, and Postlethwait, John H. Stacks: Building and genotyping loci de novo from Short-Read sequences. *G3: Genes, Genomes, Genetics*, 1(3):171–182, 2011. ISSN 2160-1836. doi:10.1534/g3.111.000240. URL http://dx.doi.org/10.1534/g3.111.000240. 74

Chaisson, Mark J. and Pevzner, Pavel A. [duplicate] short read fragment assembly of bacterial genomes. *Genome Research*, 18(2):324–330, 2008. ISSN 1549-5469. doi:10.1101/gr.7088808. URL http://dx.doi.org/10.1101/gr.7088808. 8, 54

Chikhi, Rayan and Rizk, Guillaume. Space-Efficient and exact de bruijn graph representation based on a bloom filter algorithms in bioinformatics. volume 7534 of *Lecture Notes in Computer Science*, chapter 19, pages 236–248. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-33121-3. doi:10.1007/978-3-642-33122-0_19. URL http://dx.doi.org/10.1007/978-3-642-33122-0_19. 8, 120

Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87, 2005. ISSN 1476-4687. doi:10.1038/nature04072. URL http://dx.doi.org/10.1038/nature04072. 4

Conrad, Donald F., Keebler, Jonathan E., DePristo, Mark A., Lindsay, Sarah J., Zhang, Yujun, Casals, Ferran, Idaghdour, Youssef, Hartl, Chris L., Torroja, Carlos, Garimella, Kiran V., et al. Variation in genome-wide mutation rates within and between human families. *Nature genetics*, 43(7):712–714, 2011. ISSN 1546-1718. doi:10.1038/ng.862. URL http://dx.doi.org/10.1038/ng.862. 102, 106, 107, 109

Conway, Thomas, Wazny, Jeremy, Bromage, Andrew, Zobel, Justin, and Beresford-Smith, Bryan. Gossamer a resource-efficient de novo assembler. *Bioinformatics*, 28(14):1937–1938, 2012. ISSN 1460-2059. doi:10.1093/bioinformatics/bts297. URL http://dx.doi.org/10.1093/bioinformatics/bts297. 14

Conway, Thomas C. and Bromage, Andrew J. Succinct data structures for assembling large genomes. *Bioinformatics (Oxford, England)*, 27(4):479–486, 2011. ISSN 1367-4811. doi:10.1093/bioinformatics/btq697. URL http://dx.doi.org/10.1093/bioinformatics/btq697. 8, 14

de Bruijn, N. G. A combinatorial problem. *Koninklijke Nederlandse Akademie v. Wetenschappen*, 49:758–764, 1946. 8

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society.*

Series B (Methodological), 39(1):1–38, 1977. ISSN 00359246. doi:10.2307/2984875. URL http://dx.doi.org/10.2307/2984875. 90

DePristo, Mark A., Banks, Eric, Poplin, Ryan, Garimella, Kiran V., Maguire, Jared R., Hartl, Christopher, Philippakis, Anthony A., del Angel, Guillermo, Rivas, Manuel A., Hanna, Matt, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*, 43(5):491–498, 2011. ISSN 1546-1718. doi:10.1038/ng.806. URL http://dx.doi.org/10.1038/ng.806. 68, 97, 102

Dinh, Hieu and Rajasekaran, Sanguthevar. A memory-efficient data structure representing exact-match overlap graphs with application for next-generation DNA assembly. *Bioinformatics*, 27(14):1901–1907, 2011. ISSN 1460-2059. doi:10.1093/bioinformatics/btr321. URL http://dx.doi.org/10.1093/bioinformatics/btr321. 119

Drmanac, Radoje, Sparks, Andrew B., Callow, Matthew J., Halpern, Aaron L., Burns, Norman L., Kermani, Bahram G., Carnevali, Paolo, Nazarenko, Igor, Nilsen, Geoffrey B., Yeung, George, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science (New York, N.Y.)*, 327(5961):78–81, 2010. ISSN 1095-9203. doi:10.1126/science.1181498. URL http://dx.doi.org/10.1126/science.1181498. 5

Earl, Dent, Bradnam, Keith, St. John, John, Darling, Aaron, Lin, Dawei, Fass, Joseph, Yu, Hung On Ken, Buffalo, Vince, Zerbino, Daniel R., Diekhans, Mark, et al. Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Research*, 21(12):2224–2241, 2011. ISSN 1549-5469. doi:10.1101/gr.126599.111. URL http://dx.doi.org/10.1101/gr.126599.111. 10, 14, 70, 71

Eid, John, Fehr, Adrian, Gray, Jeremy, Luong, Khai, Lyle, John, Otto, Geoff, Peluso, Paul, Rank, David, Baybayan, Primo, Bettman, Brad, et al. Real-Time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138, 2009. ISSN 1095-9203. doi:10.1126/science.1162986. URL http://dx.doi.org/10.1126/science.1162986. 5

Euler, Leonard. Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, 8:128–140, 1741. URL http://www.math.dartmouth.edu/~{}euler/pages/E053.html. 23

Ferragina, P. and Manzini, G. Opportunistic data structures with applications. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, volume 0, pages 390–398. IEEE Comput. Soc, Los Alamitos, CA, USA, 2000. ISBN 0-7695-0850-2. ISSN 0272-5428. doi:10.1109/SFCS.2000.892127. URL http://dx.doi.org/10.1109/SFCS.2000.892127. 16, 28

Ferragina, Paolo, Gagie, Travis, and Manzini, Giovanni. Lightweight data indexing and compression in external memory. In *Proceedings of the Latin American Theoretical Informatics Symposium*. 2010. URL http://arxiv.org/abs/0909.4341. 46

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., et al. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*, 260(5551):500–507, 1976. ISSN 0028-0836. URL http://view.ncbi.nlm.nih.gov/pubmed/1264203. 2

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., et al. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, 269(5223):496–512, 1995. ISSN 1095-9203. doi:10.1126/science.7542800. URL http://dx.doi.org/10.1126/science.7542800. 3

Fleury, M. Deux problemes de geometrie de situation. *Journal de mathematiques elementaires*, pages 257–261, 1883. 24

Flicek, Paul, Amode, M. Ridwan, Barrell, Daniel, Beal, Kathryn, Brent, Simon, Carvalho-Silva, Denise, Clapham, Peter, Coates, Guy, Fairley, Susan, Fitzgerald, Stephen, et al. Ensembl 2012. *Nucleic Acids Research*, 40(D1):D84–D90, 2012. ISSN 1362-4962. doi:10.1093/nar/gkr991. URL http://dx.doi.org/10.1093/nar/gkr991. 114

Genome 10K Community of Scientists. Genome 10K: A proposal to obtain Whole-Genome sequence for 10000 vertebrate species. *Journal of Heredity*, 100(6):659–674, 2009. ISSN 1465-7333. doi:10.1093/jhered/esp086. URL http://dx.doi.org/10.1093/jhered/esp086. 5

Girard, Simon L., Gauthier, Julie, Noreau, Anne, Xiong, Lan, Zhou, Sirui, Jouan, Loubna, Dionne-Laporte, Alexandre, Spiegelman, Dan, Henrion, Edouard, Diallo, Ousmane, et al. Increased exonic de novo mutation rate in individuals with schizophrenia. *Nat Genet*, 43(9):860–863, 2011. ISSN 1061-4036. doi:10.1038/ng.886. URL http://dx.doi.org/10.1038/ng.886. 106

Gnerre, Sante, MacCallum, Iain, Przybylski, Dariusz, Ribeiro, Filipe J., Burton, Joshua N., Walker, Bruce J., Sharpe, Ted, Hall, Giles, Shea, Terrance P., Sykes, Sean, et al. [duplicate] high-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences*, 108(4):1513–1518, 2011. ISSN 1091-6490. doi:10.1073/pnas.1017351108. URL http://dx.doi.org/10.1073/pnas.1017351108. 13, 14, 71

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., et al. Life with 6000 genes. *Science*, 274(5287):546–567, 1996. ISSN 1095-9203. doi:10.1126/science.274.5287.546. URL http://dx.doi.org/10.1126/science.274.5287.546. 3

Gonnella, Giorgio and Kurtz, Stefan. Readjoiner: a fast and memory efficient string graph-based sequence assembler. *BMC Bioinformatics*, 13(1):82+, 2012. ISSN 1471-2105. doi:10.1186/1471-2105-13-82. URL http://dx.doi.org/10.1186/1471-2105-13-82. 119

Green, Phil. Whole-genome disassembly. *Proceedings of the National Academy of Sciences*, 99(7):4143–4144, 2002. ISSN 1091-6490. doi:10.1073/pnas.082095999. URL http://dx.doi.org/10.1073/pnas.082095999. 4

Grossi, Roberto and Vitter, Jeffrey S. Compressed suffix arrays and suffix trees with applications to text indexing and string matching (extended abstract). In

Proceedings of the thirty-second annual ACM symposium on Theory of computing, STOC '00, pages 397–406. ACM, New York, NY, USA, 2000. ISBN 1-58113-184-4. doi:10.1145/335305.335351. URL http://dx.doi.org/10.1145/335305.335351. 17

Gusfield, Dan. *Algorithms on strings, trees, and sequences : computer science and computational biology.* Cambridge Univ. Press, 1997. ISBN 0521585198. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/0521585198. 16, 22

Healy, John, Thomas, Elizabeth E., Schwartz, Jacob T., and Wigler, Michael. Annotating large genomes with exact word matches. *Genome research*, 13(10):2306–2315, 2003. ISSN 1088-9051. doi:10.1101/gr.1350803. URL http://dx.doi.org/10.1101/gr.1350803. 16, 57

Idury, R. M. and Waterman, M. S. A new algorithm for DNA sequence assembly. *Journal of computational biology*, 2(2):291–306, 1995. ISSN 1066-5277. URL http://view.ncbi.nlm.nih.gov/pubmed/7497130. 8

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001. ISSN 0028-0836. doi:10.1038/35057062. URL http://dx.doi.org/10.1038/35057062. 4

Iqbal, Zamin, Caccamo, Mario, Turner, Isaac, Flicek, Paul, and McVean, Gil. De novo assembly and genotyping of variants using colored de bruijn graphs. *Nature genetics*, 44(2):226–232, 2012. ISSN 1546-1718. doi:10.1038/ng.1028. URL http://dx.doi.org/10.1038/ng.1028. 14, 15, 74, 75, 93, 118

Jou, W. M., Haegeman, G., Ysebaert, M., and Fiers, W. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature*, 237:82–88, 1972. doi:10.1038/237082a0. URL http://dx.doi.org/10.1038/237082a0. 2

Kececioglu, John D. and Myers, Eugene W. Combinatorial algorithms for DNA sequence assembly. In *Algorithmica*, volume 13, pages 7–51. 1993. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.44.5469. 7

Kelley, David R., Schatz, Michael C., and Salzberg, Steven L. Quake: quality-aware detection and correction of sequencing errors. *Genome biology*, 11(11):R116+, 2010. ISSN 1465-6914. doi:10.1186/gb-2010-11-11-r116. URL http://dx.doi.org/10.1186/gb-2010-11-11-r116. 47, 48

Kent, W. James. BLATthe BLAST-like alignment tool. *Genome Research*, 12(4):656–664, 2002. ISSN 1549-5469. doi:10.1101/gr.229202. \%20Article\%20published\%20online\%20before\%20March\%202002. URL http://dx.doi.org/10.1101/gr.229202.%20Article%20published%20online%20before%20March%202002. 16

Ko, P. and Aluru, S. Space efficient linear time construction of suffix arrays. *Journal of Discrete Algorithms*, 3(2-4):143–156, 2005. ISSN 15708667. doi:10.1016/j.jda.2004.08.002. URL http://dx.doi.org/10.1016/j.jda.2004.08.002. 32

Kurtz, Stefan, Narechania, Apurva, Stein, Joshua C., and Ware, Doreen. A new method to compute k-mer frequencies and its application to annotate large repetitive plant genomes. *BMC genomics*, 9(1):517+, 2008. ISSN 1471-2164. doi:10.1186/1471-2164-9-517. URL http://dx.doi.org/10.1186/1471-2164-9-517. 16

Lam, T. W., Li, Ruiqiang, Tam, Alan, Wong, Simon, Wu, Edward, and Yiu, S. M. High throughput short read alignment via bi-directional BWT. In *2009 IEEE International Conference on Bioinformatics and Biomedicine*, volume 0, pages 31–36. IEEE, Los Alamitos, CA, USA, 2009. ISBN 978-0-7695-3885-3. doi:10.1109/BIBM.2009.42. URL http://dx.doi.org/10.1109/BIBM.2009.42. 36

Lam, T. W., Sung, W. K., Tam, S. L., Wong, C. K., and Yiu, S. M. Compressed indexing and local alignment of DNA. *Bioinformatics*, 24(6):791–797, 2008. ISSN 1460-2059. doi:10.1093/bioinformatics/btn032. URL http://dx.doi.org/10.1093/bioinformatics/btn032. 16

Langmead, Ben, Trapnell, Cole, Pop, Mihai, and Salzberg, Steven. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25–10, 2009. ISSN 1465-6906. doi:10.1186/gb-2009-10-3-r25. URL http://dx.doi.org/10.1186/gb-2009-10-3-r25. 16

Li, Fugen and Stormo, Gary D. Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics*, 17(11):1067–1076, 2001. ISSN 1460-2059. doi:10.1093/bioinformatics/17.11.1067. URL http://dx.doi.org/10.1093/bioinformatics/17.11.1067. 16

Li, Heng. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)*, 27(21):2987–2993, 2011. ISSN 1367-4811. doi:10.1093/bioinformatics/btr509. URL http://dx.doi.org/10.1093/bioinformatics/btr509. 15

Li, Heng. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics*, 28(14):1838–1844, 2012. ISSN 1460-2059. doi:10.1093/bioinformatics/bts280. URL http://dx.doi.org/10.1093/bioinformatics/bts280. 14, 15, 93, 105, 118, 119

Li, Heng and Durbin, Richard. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–1760, 2009. ISSN 1367-4811. doi:10.1093/bioinformatics/btp324. URL http://dx.doi.org/10.1093/bioinformatics/btp324. 16, 55, 66

Li, Heng and Durbin, Richard. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 26(5):589–595, 2010. ISSN 1367-4811. doi:10.1093/bioinformatics/btp698. URL http://dx.doi.org/10.1093/bioinformatics/btp698. 62, 68

Li, Heng and Homer, Nils. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, 11(5):473–483, 2010. ISSN 1477-4054. doi:10.1093/bib/bbq015. URL http://dx.doi.org/10.1093/bib/bbq015. 15, 74

Li, Ruiqiang, Fan, Wei, Tian, Geng, Zhu, Hongmei, He, Lin, Cai, Jing, Huang, Quanfei, Cai, Qingle, Li, Bo, Bai, Yinqi, et al. The sequence and de novo assembly of the giant panda genome. *Nature*, 463(7279):311–317, 2010a. ISSN 1476-4687. doi:10.1038/nature08696. URL http://dx.doi.org/10.1038/nature08696. 4, 13

Li, Ruiqiang, Li, Yingrui, Zheng, Hancheng, Luo, Ruibang, Zhu, Hongmei, Li, Qibin, Qian, Wubin, Ren, Yuanyuan, Tian, Geng, Li, Jinxiang, et al. Building the sequence map of the human pan-genome. *Nat Biotech*, 28(1):57–63, 2010b. ISSN 1546-1696. doi:10.1038/nbt.1596. URL http://dx.doi.org/10.1038/nbt.1596. 61

Li, Ruiqiang, Yu, Chang, Li, Yingrui, Lam, Tak-Wah, Yiu, Siu-Ming, Kristiansen, Karsten, and Wang, Jun. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–1967, 2009. ISSN 1460-2059. doi:10.1093/bioinformatics/btp336. URL http://dx.doi.org/10.1093/bioinformatics/btp336. 16

Li, Ruiqiang, Zhu, Hongmei, Ruan, Jue, Qian, Wubin, Fang, Xiaodong, Shi, Zhongbin, Li, Yingrui, Li, Shengting, Shan, Gao, Kristiansen, Karsten, et al. [duplicate] de novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, 20(2):265–272, 2010c. ISSN 1549-5469. doi:10.1101/gr.097261.109. URL http://dx.doi.org/10.1101/gr.097261.109. 8, 13, 47, 48, 54, 61, 67, 71

Maccallum, Iain, Przybylski, Dariusz, Gnerre, Sante, Burton, Joshua, Shlyakhter, Ilya, Gnirke, Andreas, Malek, Joel, McKernan, Kevin, Ranade, Swati, Shea, Terrance P., et al. ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. *Genome biology*, 10(10):R103+, 2009. ISSN 1465-6914. doi:10.1186/gb-2009-10-10-r103. URL http://dx.doi.org/10.1186/gb-2009-10-10-r103. 13

Malde, Ketil, Coward, Eivind, and Jonassen, Inge. Fast sequence clustering using a suffix array algorithm. *Bioinformatics*, 19(10):1221–1226, 2003. ISSN 1460-2059. doi:10.1093/bioinformatics/btg138. URL http://dx.doi.org/10.1093/bioinformatics/btg138. 16

Manber, Udi and Myers, Gene. Suffix arrays: a new method for on-line string searches. In *SODA '90: Proceedings of the first annual ACM-SIAM symposium on Discrete algorithms*, pages 319–327. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1990. ISBN 0-89871-251-3. URL http://portal.acm.org/citation.cfm?id=320176.320218. 16, 28

Maxam, A. M. and Gilbert, W. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2):560–564, 1977. ISSN 0027-8424. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC392330/. 2

McLaren, William, Pritchard, Bethan, Rios, Daniel, Chen, Yuan, Flicek, Paul, and Cunningham, Fiona. Deriving the consequences of genomic variants with the ensembl API and SNP effect predictor. *Bioinformatics (Oxford, England)*, 26(16):2069–2070, 2010. ISSN 1367-4811. doi:10.1093/bioinformatics/btq330. URL http://dx.doi.org/10.1093/bioinformatics/btq330. 114

Meacham, Frazer, Boffelli, Dario, Dhahbi, Joseph, Martin, David, Singer, Meromit, and Pachter, Lior. Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics*, 12(1):451+, 2011. ISSN 1471-2105. doi:10.1186/1471-2105-12-451. URL http://dx.doi.org/10.1186/1471-2105-12-451. 79

Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002. ISSN 0028-0836. doi:10.1038/nature01262. URL http://dx.doi.org/10.1038/nature01262. 4

Myers, Eugene W. The fragment assembly string graph. *Bioinformatics*, 21(suppl 2):ii79–ii85, 2005. ISSN 1460-2059. doi:10.1093/bioinformatics/bti1114. URL http://dx.doi.org/10.1093/bioinformatics/bti1114. 7, 22, 24, 28, 44, 55, 83

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

1091-6490. doi:10.1073/pnas.092136699. URL http://dx.doi.org/10.1073/pnas.092136699. 4

Nagarajan, Niranjan and Pop, Mihai. Parametric complexity of sequence assembly: theory and applications to next generation sequencing. *Journal of computational biology : a journal of computational molecular cell biology*, 16(7):897–908, 2009. ISSN 1557-8666. doi:10.1089/cmb.2009.0005. URL http://dx.doi.org/10.1089/cmb.2009.0005. 24

Nik-Zainal, Serena, Alexandrov, Ludmil B., Wedge, David C., Van Loo, Peter, Greenman, Christopher D., Raine, Keiran, Jones, David, Hinton, Jonathan, Marshall, John, Stebbings, Lucy A., et al. Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–993, 2012a. ISSN 00928674. doi:10.1016/j.cell.2012.04.024. URL http://dx.doi.org/10.1016/j.cell.2012.04.024. 109, 110

Nik-Zainal, Serena, Van Loo, Peter, Wedge, David C., Alexandrov, Ludmil B., Greenman, Christopher D., Lau, King Wai W., Raine, Keiran, Jones, David, Marshall, John, Ramakrishna, Manasa, et al. The life history of 21 breast cancers. *Cell*, 149(5):994–1007, 2012b. ISSN 1097-4172. doi:10.1016/j.cell.2012.04.023. URL http://dx.doi.org/10.1016/j.cell.2012.04.023. 109

Ning, Z., Cox, A. J., and Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome research*, 11(10):1725–1729, 2001. ISSN 1088-9051. doi:10.1101/gr.194201. URL http://dx.doi.org/10.1101/gr.194201. 16

Nong, Ge, Zhang, Sen, and Chan, Wai H. Linear suffix array construction by almost pure Induced-Sorting. *Data Compression Conference*, 0:193–202, 2009. ISSN 1068-0314. doi:10.1109/DCC.2009.42. URL http://dx.doi.org/10.1109/DCC.2009.42. 32, 46

Pell, Jason, Hintze, Arend, Canino-Koning, Rosangela, Howe, Adina, Tiedje, James M., and Brown, C. Titus. Scaling metagenome sequence assembly with probabilistic de bruijn graphs. *Proceedings of the National Academy of Sciences*, 109(33):13272–13277, 2012. ISSN 1091-6490. doi:10.1073/pnas.1121464109. URL http://dx.doi.org/10.1073/pnas.1121464109. 8, 120

Pevzner, P. A. 1-Tuple DNA sequencing: computer analysis. *Journal of biomolecular structure & dynamics*, 7(1):63–73, 1989. ISSN 0739-1102. URL http://view.ncbi.nlm.nih.gov/pubmed/2684223. 8

Pevzner, Pavel A., Tang, Haixu, and Waterman, Michael S. An eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences*, 98(17):9748–9753, 2001. ISSN 1091-6490. doi:10.1073/pnas.171285098. URL http://dx.doi.org/10.1073/pnas.171285098. 8, 23, 47, 48, 78

Pop, Mihai, Kosack, Daniel S., and Salzberg, Steven L. Hierarchical scaffolding with bambus. *Genome research*, 14(1):149–159, 2004. ISSN 1088-9051. doi:10.1101/gr.1536204. URL http://dx.doi.org/10.1101/gr.1536204. 55

Prufer, Kay, Munch, Kasper, Hellmann, Ines, Akagi, Keiko, Miller, Jason R., Walenz, Brian, Koren, Sergey, Sutton, Granger, Kodira, Chinnappa, Winer, Roger, et al. The bonobo genome compared with the chimpanzee and human genomes. *Nature*, 486(7404):527–531, 2012. ISSN 0028-0836. doi:10.1038/nature11128. URL http://dx.doi.org/10.1038/nature11128. 4

Puglisi, Simon J., Smyth, W. F., and Turpin, Andrew H. A taxonomy of suffix array construction algorithms. *ACM Comput. Surv.*, 39(2):4+, 2007. ISSN 0360-0300. doi:10.1145/1242471.1242472. URL http://dx.doi.org/10.1145/1242471.1242472. 31

Rasmussen, Kim R., Stoye, Jens, and Myers, Eugene W. Efficient q-gram filters for finding all epsilon-matches over a given length. *Journal of computational biology : a journal of computational molecular cell biology*, 13(2):296–308, 2006. ISSN 1066-5277. doi:10.1089/cmb.2006.13.296. URL http://dx.doi.org/10.1089/cmb.2006.13.296. 7, 22

Sanders, Stephan J., Murtha, Michael T., Gupta, Abha R., Murdoch, John D., Raubeson, Melanie J., Willsey, A. Jeremy, Ercan-Sencicek, A. Gulhan, DiLullo, Nicholas M., Parikshak, Neelroop N., Stein, Jason L., et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, 485(7397):237–241, 2012. ISSN 1476-4687. doi:10.1038/nature10945. URL http://dx.doi.org/10.1038/nature10945. 106

Sanger, F., Brownlee, G. G., and Barrell, B. G. A two-dimensional fractionation procedure for radioactive nucleotides. *Journal of Molecular Biology*, 14(1):303+, 1965. ISSN 00222836. doi:10.1016/S0022-2836(65)80253-4. URL http://dx.doi.org/10.1016/S0022-2836(65)80253-4. 2

Sanger, F., Nicklen, S., and Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, 1977. ISSN 1091-6490. doi:10.1073/pnas.74.12.5463. URL http://dx.doi.org/10.1073/pnas.74.12.5463. 3

Scally, Aylwyn, Dutheil, Julien Y., Hillier, LaDeana W., Jordan, Gregory E., Goodhead, Ian, Herrero, Javier, Hobolth, Asger, Lappalainen, Tuuli, Mailund, Thomas, Marques-Bonet, Tomas, et al. Insights into hominid evolution from the gorilla genome sequence. *Nature*, 483(7388):169–175, 2012. ISSN 1476-4687. doi:10.1038/nature10842. URL http://dx.doi.org/10.1038/nature10842. 4, 13

Schmid, C. W. and Deininger, P. L. Sequence organization of the human genome. *Cell*, 6(3):345–358, 1975. ISSN 0092-8674. URL http://view.ncbi.nlm.nih.gov/pubmed/1052772. 3

Schneider, Gregory F. and Dekker, Cees. DNA sequencing with nanopores. *Nat Biotech*, 30(4):326–328, 2012. ISSN 1087-0156. doi:10.1038/nbt.2181. URL http://dx.doi.org/10.1038/nbt.2181. 120

Simpson, Jared T. and Durbin, Richard. Efficient construction of an assembly string graph using the FM-index. *Bioinformatics*, 26(12):i367–i373, 2010. ISSN 1460-2059. doi:10.1093/bioinformatics/btq217. URL http://dx.doi.org/10.1093/bioinformatics/btq217. ii, 19, 59, 119

Simpson, Jared T. and Durbin, Richard. [duplicate] efficient de novo assembly of large genomes using compressed data structures. *Genome research*, 22(3):549–556, 2012. ISSN 1549-5469. doi:10.1101/gr.126953.111. URL http://dx.doi.org/10.1101/gr.126953.111. ii, 44, 102, 119

Simpson, Jared T., Wong, Kim, Jackman, Shaun D., Schein, Jacqueline E., Jones, Steven J. M., and Birol, İnanç. ABySS: A parallel assembler for short read sequence data. *Genome Research*, 19(6):1117–1123, 2009. ISSN 1549-5469. doi: 10.1101/gr.089532.108. URL http://dx.doi.org/10.1101/gr.089532.108. 8, 13, 54, 55, 61

Sirén, Jouni. Compressed suffix arrays for massive data. In *String Processing and Information Retrieval*, pages 63–74. 2009. doi:10.1007/978-3-642-03784-9_7. URL http://dx.doi.org/10.1007/978-3-642-03784-9_7. 53

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner, C., Kent, S. B. H., and Hood, L. E. Fluorescence detection in automated DNA sequence analysis. *Nature*, 321(6071):674–679, 1986. ISSN 0028-0836. doi:10.1038/321674a0. URL http://dx.doi.org/10.1038/321674a0. 3

Staden, R. A strategy of DNA sequencing employing computer programs. *Nucleic acids research*, 6(7):2601–2610, 1979. ISSN 0305-1048. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC327874/. 6

The International Cancer Genome Consortium. International network of cancer genome projects. *Nature*, 464(7291):993–998, 2010. ISSN 1476-4687. doi: 10.1038/nature08987. URL http://dx.doi.org/10.1038/nature08987. 5

Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485(7400):635–641, 2012. ISSN 1476-4687. doi:10.1038/nature11119. URL http://dx.doi.org/10.1038/nature11119. 13

Välimäki, Niko, Ladra, Susana, and Mäkinen, Veli. Approximate All-Pairs Suffix/Prefix overlaps. In Amir, Amihood and Parida, Laxmi, editors, *Combinatorial Pattern Matching*, volume 6129 of *Lecture Notes in Computer Science*, pages 76–87. Springer Berlin / Heidelberg, 2010. doi:10.1007/978-3-642-13509-5_8. URL http://dx.doi.org/10.1007/978-3-642-13509-5_8. 52

Valouev, Anton, Ichikawa, Jeffrey, Tonthat, Thaisan, Stuart, Jeremy, Ranade, Swati, Peckham, Heather, Zeng, Kathy, Malek, Joel A., Costa, Gina, McKernan, Kevin, et al. A high-resolution, nucleosome position map of c. elegans reveals a lack of universal sequence-dictated positioning. *Genome research*, 18(7):1051–1063, 2008. ISSN 1088-9051. doi:10.1101/gr.076463.108. URL http://dx.doi.org/10.1101/gr.076463.108. 5

Venter, J. Craig, Adams, Mark D., Myers, Eugene W., Li, Peter W., Mural, Richard J., Sutton, Granger G., Smith, Hamilton O., Yandell, Mark, Evans, Cheryl A., Holt, Robert A., et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001. ISSN 1095-9203. doi:10.1126/science.1058040. URL http://dx.doi.org/10.1126/science.1058040. 4, 7, 22

Wakeley, J. The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. *Trends in ecology & evolution*, 11(4):158–162, 1996. ISSN 0169-5347. URL http://view.ncbi.nlm.nih.gov/pubmed/21237791. 115

Warren, René L., Sutton, Granger G., Jones, Steven J. M., and Holt, Robert A. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, 23(4):500–501, 2007. ISSN 1460-2059. doi:10.1093/bioinformatics/btl629. URL http://dx.doi.org/10.1093/bioinformatics/btl629. 13

Waterston, Robert H., Lander, Eric S., and Sulston, John E. On the sequencing of the human genome. *Proceedings of the National Academy of Sciences*, 99(6):3712–3716, 2002. ISSN 1091-6490. doi:10.1073/pnas.042692499. URL http://dx.doi.org/10.1073/pnas.042692499. 4

Waterston, Robert H., Lander, Eric S., and Sulston, John E. More on the sequencing of the human genome. *Proceedings of the National Academy of Sciences*, 100(6):3022–3024, 2003. ISSN 1091-6490. doi:10.1073/pnas.0634129100. URL http://dx.doi.org/10.1073/pnas.0634129100. 4

Wheeler, David A., Srinivasan, Maithreyan, Egholm, Michael, Shen, Yufeng, Chen, Lei, McGuire, Amy, He, Wen, Chen, Yi-Ju, Makhijani, Vinod, Roth, G. Thomas, et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452(7189):872–876, 2008. ISSN 0028-0836. doi:10.1038/nature06884. URL http://dx.doi.org/10.1038/nature06884. 5

Ye, Kai, Schulz, Marcel H., Long, Quan, Apweiler, Rolf, and Ning, Zemin. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21):2865–2871, 2009. ISSN 1460-2059. doi:10.1093/bioinformatics/btp394. URL `http://dx.doi.org/10.1093/bioinformatics/btp394`. 112

Zerbino, Daniel R. and Birney, Ewan. Velvet: Algorithms for de novo short read assembly using de bruijn graphs. *Genome Research*, 18(5):821–829, 2008. ISSN 1549-5469. doi:10.1101/gr.074492.107. URL `http://dx.doi.org/10.1101/gr.074492.107`. 8, 13, 54, 61