References

Albers, C., Lunter, G., MacArthur, D., McVean, G., et al. Dindel: Accurate indel calls from short-read data. *Genome Research*, 21(6):961, 2011. 4

Ananda, G., Chiaromonte, F., and Makova, K. A genome-wide view of mutation rate co-variation using multivariate analyses. *Genome Biology*, 12(3):R27, 2011. 87

Anderson, S. Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Research*, 9(13):3015, 1981. 37

Ball, E., Stenson, P., Abeysinghe, S., Krawczak, M., et al. Microdeletions and microinsertions causing human genetic disease: common mechanisms of mutagenesis and the role of local DNA sequence complexity. *Human muttaion*, 26(3):205–213, 2005. 10

Bansal, V. and Libiger, O. A probabilistic method for the detection and genotyping of small indels from population-scale sequence data. *Bioinformatics*, 2011. 25

Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research*, 27(2):573, 1999. 6, 29

Bentley, D., Balasubramanian, S., Swerdlow, H., Smith, G., et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008. 4, 16, 28, 31

Bhangale, T., Rieder, M., Livingston, R., and Nickerson, D. Comprehensive identification and characterization of diallelic insertion–deletion polymorphisms in 330 human candidate genes. *Human molecular genetics*, 14(1):59, 2005. 25

Brinkmann, B., Klintschar, M., Neuhuber, F., Huhne, J., et al. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *The American Journal of Human Genetics*, 62(6):1408–1415, 1998. 87

Brownlee, G., Sanger, F., and Barrell, B. Nucleotide sequence of 5S-ribosomal RNA from Escherichia coli. *Nature*, 1967. 3

Bugaut, A. and Balasubramanian, S. A sequence-independent study of the influence of short loop lengths on the stability and topology of intramolecular DNA G-quadruplexes. *Biochemistry*, 47(2):689–697, 2008. 115

Calafell, F., Shuster, A., Speed, W., Kidd, J., et al. Short tandem repeat polymorphism evolution in humans. *European Journal of Human Genetics*, 6(1):38–49, 1998. 87

Carrilho, E. DNA sequencing by capillary array electrophoresis and microfabricated array systems. *Electrophoresis*, 21(1):55–65, 2000. 3

Chen, K., McLellan, M., Ding, L., Wendl, M., et al. PolyScan: an automatic indel and SNP detection approach to the analysis of human resequencing data. *Genome research*, 17(5):659, 2007. 5

Cochrane, G., Akhtar, R., Bonfield, J., Bower, L., et al. Petabyte-scale innovations at the European Nucleotide Archive. *Nucleic acids research*, 37(suppl 1):D19, 2009. 28

Cohen, J. Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. *PLoS Biology*, 2(12):e439, 2004. 2

Consortium, T..G.P. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073, 2010. 31, 72, 88, 119

Di Rienzo, A., Peterson, A., Garza, J., Valdes, A., et al. Mutational processes of simple-sequence repeat loci in human populations. *Proceedings of the National Academy of Sciences of the United States of America*, 91(8):3166, 1994. 10

Dohm, J., Lottaz, C., Borodina, T., and Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic acids research*, 36(16):e105, 2008. 101

Drake, J., Charlesworth, B., Charlesworth, D., and Crow, J. Rates of spontaneous mutation. *Genetics*, 148(4):1667, 1998. 98

Durbin, R. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge Univ Pr, 1998. 57

Ellegren, H. Microsatellite mutations in the germline::: implications for evolutionary inference. *Trends in Genetics*, 16(12):551–558, 2000. 11

Ellegren, H. Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics*, 5(6):435–445, 2004. 11, 55

Ewing, B. and Green, P. Base-calling of automated sequencer traces usingPhred. II. error probabilities. *Genome research*, 8(3):186, 1998. 34

Ewing, B., Hillier, L., Wendl, M., and Green, P. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome research*, 8(3):175, 1998. 36

Feuk, L., Carson, A., and Scherer, S. Structural variation in the human genome. *Nature Reviews Genetics*, 7(2):85–97, 2006. 4

Gatchel, J. and Zoghbi, H. Diseases of unstable repeat expansion: mechanisms and common principles. *Nature Reviews Genetics*, 6(10):743–755, 2005. 11

Green, E., Guyer, M., et al. Charting a course for genomic medicine from base pairs to bedside. *Nature*, 470(7333):204–213, 2011. 2

Hamosh, A., Scott, A., Amberger, J., Bocchini, C., et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(suppl 1):D514, 2005. 10

Hanawalt, P. Transcription-coupled repair and human disease. *Science*, 266(5193):1957, 1994. 116

Hazel, P., Huppert, J., Balasubramanian, S., and Neidle, S. Loop-length-dependent folding of G-quadruplexes. *Journal of the American Chemical Society*, 126(50):16405–16415, 2004. 115

Hoeijmakers, J. et al. Genome maintenance mechanisms for preventing cancer. *Nature*, 411(6835):366–374, 2001. 116

Homer, N., Merriman, B., and Nelson, S. BFAST: an alignment tool for large scale genome resequencing. *PLoS One*, 4(11):e7767, 2009. 4

Hormozdiari, F., Alkan, C., Eichler, E., and Sahinalp, S. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome research*, 19(7):1270, 2009. 16

Huppert, J. and Balasubramanian, S. Prevalence of quadruplexes in the human genome. *Nucleic Acids Research*, 33(9):2908, 2005. 115

Huse, S., Huber, J., Morrison, H., Sogin, M., et al. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome biology*, 8(7):R143, 2007. 74

International Human Genome Sequencing Consortium, B. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001. 2

Kasai, K., Nakamura, Y., and White, R. Amplification of a variable number of tandem repeats (VNTR) locus (pMCT118) by the polymerase chain reaction (PCR) and its application to forensic science. *Journal of forensic sciences*, 35(5):1196, 1990. 87

Kashi, Y., King, D., and Soller, M. Simple sequence repeats as a source of quantitative genetic variation. *Trends in Genetics*, 13(2):74–78, 1997. 11

Kashi, Y. and King, D. Simple sequence repeats as advantageous mutators in evolution. *TRENDS in Genetics*, 22(5):253–259, 2006. 11

Kelkar, Y., Tyekucheva, S., Chiaromonte, F., and Makova, K. The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome research*, 18(1):30, 2008. 118

Koboldt, D., Chen, K., Wylie, T., Larson, D., et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, 25(17):2283, 2009. 26

Korbel, J., Abyzov, A., Mu, X., Carriero, N., et al. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biology*, 10(2):R23, 2009. 15

Korbel, J., Urban, A., Affourtit, J., Godwin, B., et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318(5849):420, 2007. 13, 32, 34

Kovtun, I. and McMurray, C. Features of trinucleotide repeat instability in vivo. *Cell research*, 18(1):198–213, 2008. 11

Krawitz, P., Rodelsperger, C., Jager, M., Jostins, L., et al. Microindel detection in short-read sequence data. *Bioinformatics*, 26(6):722, 2010. 4

Kuhn, R., Karolchik, D., Zweig, A., Trumbower, H., et al. The UCSC genome browser database: update 2007. *Nucleic acids research*, 35(suppl 1):D668, 2006. 29

Lai, Y. and Sun, F. The relationship between microsatellite slippage mutation rate and the number of repeat units. *Molecular biology and evolution*, 20(12):2123, 2003. 87

Lee, S., Hormozdiari, F., Alkan, C., and Brudno, M. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nature methods*, 6(7):473–474, 2009. 82, 83

Legendre, M., Pochet, N., Pak, T., and Verstrepen, K. Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome research*, 17(12):1787, 2007. 99

Lenzmeier, B. and Freudenreich, C. Trinucleotide repeat instability: a hairpin curve at the crossroads of replication, recombination, and repair. *Cytogenetic and genome research*, 100(1-4):7–24, 2000. 11

Levy, S., Sutton, G., Ng, P., Feuk, L., et al. The diploid genome sequence of an individual human. *PLoS biology*, 5(10):e254, 2007. 2

Li, H. and Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754, 2009. 4, 32

Li, H., Handsaker, B., Wysoker, A., Fennell, T., et al. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078, 2009. 73

Li, H., Ruan, J., and Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research*, 18(11):1851, 2008. 4, 32

Lygo, J., Johnson, P., Holdaway, D., Woodroffe, S., et al. The validation of short tandem repeat (STR) loci for use in forensic casework. *International Journal of Legal Medicine*, 107(2):77–89, 1994. 87

Madsen, B., Villesen, P., and Wiuf, C. Short tandem repeats in human exons: a target for disease mutations. *BMC genomics*, 9(1):410, 2008. 10

Mahtani, M. and Willard, H. A polymorphic X-linked tetranucleotide repeat locus displaying a high rate of new mutation: implications for mechanisms of mutation at short tandem repeat loci. *Human molecular genetics*, 2(4):431, 1993. 11

Mardis, E. The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24(3):133–141, 2008. 13

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9):1297, 2010. 4

McKernan, K., Peckham, H., Costa, G., McLaughlin, S., et al. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome research*, 19(9):1527, 2009. 4

Meacham, F., Boffelli, D., Dhahbi, J., Martin, D., et al. Identification and correction of systematic error in high-throughput sequence data. 2011. 101

Medvedev, P., Stanciu, M., and Brudno, M. Computational methods for discovering structural variation with next-generation sequencing. *nature methods*, 6:S13–S20, 2009. 5

Metzker, M. Sequencing technologiesthe next generation. *Nature Reviews Genetics*, 11(1):31–46, 2009. 3

Mills, R., Luttig, C., Larkins, C., Beauchamp, A., et al. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome research*, 16(9):1182, 2006. 25

Moore, G. et al. Cramming more components onto integrated circuits. *Proceedings of the IEEE*, 86(1):82–85, 1998. 1

Myers, E. The fragment assembly string graph. *Bioinformatics*, 21(suppl 2):ii79, 2005. 28

Nachman, M. and Crowell, S. Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156(1):297, 2000. 98

Nature Jobs. New opportunities in the genomic era. 2011. 2

Ning, Z., Cox, A., and Mullikin, J. SSAHA: a fast search method for large DNA databases. *Genome Research*, 11(10):1725, 2001. 36

Ning, Z., Spooner, W., Spargo, A., Leonard, S., et al. The SSAHA trace server. 2004. 36

Pearson, C., Edamura, K., and Cleary, J. Repeat instability: mechanisms of dynamic mutations. *Nature Reviews Genetics*, 6(10):729–742, 2005. 10, 11

Pearson, W. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics*, 11(3):635–650, 1991. 36

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL `http://www.R-project.org`. ISBN 3-900051-07-0. 104

Ruitberg, C., Reeder, D., and Butler, J. STRBase: a short tandem repeat DNA database for the human identity testing community. *Nucleic Acids Research*, 29(1):320, 2001. 87

Rumble, S., Lacroute, P., Dalca, A., Fiume, M., et al. SHRiMP: accurate mapping of short color-space reads. *PLoS computational biology*, 5(5):e1000386, 2009. 4

Sanger, F., Coulson, A., Hong, G., Hill, D., et al. Nucleotide sequence of bacteriophage [lambda] DNA. *Journal of Molecular Biology*, 162(4):729–773, 1982. 3

Sindi, S., Helman, E., Bashir, A., and Raphael, B. A geometric approach for classification and comparison of structural variants. *Bioinformatics*, 25(12):i222, 2009. 15

Sprecher, C., Puers, C., Lins, A., and Schumm, J. General approach to analysis of polymorphic short tandem repeat loci. *BioTechniques*, 20(2):266–277, 1996. 88

The Economist. Data, data everywhere. 2010. 1

Tuzun, E., Sharp, A., Bailey, J., Kaul, R., et al. Fine-scale structural variation of the human genome. *Nature genetics*, 37(7):727–732, 2005. 13, 15

Urquhart, A., Kimpton, C., Downes, T., and Gill, P. Variation in short tandem repeat sequencesa survey of twelve microsatellite loci for use as forensic identification markers. *International Journal of Legal Medicine*, 107(1):13–20, 1994. 87

Volik, S., Raphael, B., Huang, G., Stratton, M., et al. Decoding the fine-scale structure of a breast cancer genome and transcriptome. *Genome research*, 16(3):394, 2006. 13

Wang, J., Wang, W., Li, R., Li, Y., et al. The diploid genome sequence of an Asian individual. *Nature*, 456(7218):60–65, 2008. 4

Weber, J. and Wong, C. Mutation of human short tandem repeats. *Human molecular genetics*, 2(8):1123, 1993. 87

Wheeler, D., Srinivasan, M., Egholm, M., Shen, Y., et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452(7189):872–876, 2008. 4

Whittaker, J., Harbord, R., Boxall, N., Mackay, I., et al. Likelihood-based estimation of microsatellite mutation rates. *Genetics*, 164(2):781, 2003. 87

Xu, H., Chakraborty, R., and Fu, Y. Mutation rate variation at human dinucleotide microsatellites. *Genetics*, 170(1):305, 2005. 88

Zeitouni, B., Boeva, V., Janoueix-Lerosey, I., Loeillet, S., et al. SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics*, 26(15):1895, 2010. 16

Zhivotovsky, L., Rosenberg, N., and Feldman, M. Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *The American Journal of Human Genetics*, 72(5):1171–1186, 2003. 11