Bibliography

[1] G. Achaz, P. Netter, and E. Coissac. Study of intrachromosomal duplications among the eukaryote genomes. *Mol Biol Evol*, 18(12):2280–2288, 2001.

[2] M. M. Alba and R. Guigo. Comparative analysis of amino acid repeats in rodents and humans. *Genome Res*, 14(4):549–554, 2004.

[3] S. F. Altschul. Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol*, 219(3):555–65, 1991.

[4] S. F. Altschul and W. Gish. Local alignment statistics. *Methods Enzymol*, 266:460–80, 1996.

[5] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990.

[6] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, 1997.

[7] W. Amos, S. J. Sawcer, R. W. Feakes, and D. C. Rubinsztein. Microsatellites show mutational bias and heterozygote instability. *Nat Genet*, 13(4):390–1, 1996.

[8] P. Andolfatto. Adaptive evolution of non-coding DNA in Drosophila. *Nature*, 437(7062):1149–52, 2005.

[9] P. F. Arndt, C. B. Burge, and T. Hwa. DNA sequence evolution with neighbor-dependent mutation. *J Comput Biol*, 10(3-4):313–322, 2003.

[10] P. F. Arndt and T. Hwa. Regional and time-resolved mutation patterns of the human genome. *Bioinformatics*, 20(10):1482–1485, 2004.

[11] P. F. Arndt and T. Hwa. Identification and measurement of neighbor-dependent nucleotide substitution processes. *Bioinformatics*, 21(10):2322–2328, 2005.

[12] P. F. Arndt, D. A. Petrov, and T. Hwa. Distinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. *Mol Biol Evol*, 20(11):1887–1896, 2003.

[13] A. Arneodo, E. Bacry, P. V. Graves, and J. F. Muzy. Characterizing long-range correlations in DNA sequences from wavelet analysis. *Phys Rev Lett*, 74(16):3293–3296, 1995.

[14] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig *et al.* Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, 2000.

[15] J. A. Bailey and E. E. Eichler. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet*, 7(7):552–564, 2006.

[16] P.-F. Baisnee, S. Hampson, and P. Baldi. Why are complementary DNA strands symmetric? *Bioinformatics*, 18(8):1021–33, 2002.

[17] E. V. Ball, P. D. Stenson, S. S. Abeysinghe, M. Krawczak, D. N. Cooper, and N. A. Chuzhanova. Microdeletions and microinsertions causing human genetic disease: common mechanisms of mutagenesis and the role of local DNA sequence complexity. *Hum Mutat*, 26(3):205–13, 2005.

[18] M. Barluenga, K. N. Stolting, W. Salzburger, M. Muschick, and A. Meyer. Sympatric speciation in Nicaraguan crater lake cichlid fish. *Nature*, 439(7077):719–23, 2006.

[19] M. R. Barnes and I. C. Gray. Amino acid properties and consequences of subsitutions. In *Bioinformatics for Geneticists*. Wiley, 2003.

[20] M. A. Batzer, G. E. Kilroy, P. E. Richard, T. H. Shaikh, T. D. Desselle, C. L. Hoppens, and P. L. Deininger. Structure and variability of recently inserted Alu family members. *Nucleic Acids Res*, 18(23):6793–8, 1990.

[21] M. J. Benton. Diversification and extinction in the history of life. *Science*, 268(5207):52–8, 1995.

[22] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res*, 28(1):235–242, 2000.

[23] P. Bernaola-Galvan, P. Carpena, R. Roman-Roldan, and J. L. Oliver. Study of statistical correlations in DNA sequences. *Gene*, 300(1-2):105–15, 2002.

[24] R. J. Britten, L. Rowen, J. Williams, and R. A. Cameron. Majority of divergence between closely related DNA samples is due to indels. *Proc Natl Acad Sci USA*, 100(8):4661–4665, 2003.

[25] M. Brudno, C. B. Do, G. M. Cooper, M. F. Kim, E. Davydov, NISC Comparative Sequencing Program, E. D. Green, A. Sidow, and S. Batzoglou. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res*, 13(4):721–731, 2003.

[26] S. V. Buldyrev. Power law correlations in DNA sequences. In E. V. Koonin, editor, *Power Laws, Scale-Free Networks and Genome Biology*, pages 126–164. Landes Bioscience, 2006.

[27] R. Bundschuh. An analytic approach to significance assessment in local sequence alignment with gaps. In *Proceedings of the 4th Annual International Conference, RECOMB 2000*, pages 86–95. ACM, New York, 2000.

[28] R. Bundschuh. Asymmetric exclusion process and extremal statistics of random sequences. *Phys Rev E*, 65(3):031911, 2002.

[29] C. D. Bustamante, A. Fledel-Alon, S. Williamson, R. Nielsen, M. Todd Hubisz, S. Glanowski, D. M. Tanenbaum, T. J. White, J. J. Sninsky, R. D. Hernandez *et al.* Natural selection on protein-coding genes in the human genome. *Nature*, 437(7062):1153–1157, 2005.

[30] C. T. Caskey, A. Pizzuti, Y. H. Fu, R. G. Fenwick, and D. L. Nelson. Triplet repeat mutations in human disease. *Science*, 256(5058):784–9, 1992.

[31] F.-C. Chen, C.-J. Chen, W.-H. Li, and T.-J. Chuang. Human-specific insertions and deletions inferred from mammalian genome sequences. *Genome Res*, 17(1):16–22, 2007.

[32] F.-C. Chen and W.-H. Li. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am J Hum Genet*, 68(2):444–456, 2001.

[33] Z. Cheng, M. Ventura, X. She, P. Khaitovich, T. Graves, K. Osoegawa, D. Church, P. DeJong, R. K. Wilson, S. Pääbo, M. Rocchi, and E. E. Eichler. A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature*, 437(7055):88–93, 2005.

[34] N. Chia and R. Bundschuh. A practical approach to significance assessment in alignment with gaps. In *Proceedings of the 9th Annual International Conference, RECOMB 2005*, Lecture Notes in Computer Science, pages 474–488. Springer, Berlin, 2005.

[35] Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87, 2005.

[36] N. A. Chuzhanova, E. J. Anassis, E. V. Ball, M. Krawczak, and D. N. Cooper. Meta-analysis of indels causing human genetic disease: mechanisms of mutagenesis and the role of local DNA sequence complexity. *Hum Mutat*, 21(1):28–44, 2003.

[37] N. A. Chuzhanova, M. Krawczak, L. A. Nemytikova, V. D. Gusev, and D. N. Cooper. Promoter shuffling has occurred during the evolution of the vertebrate growth hormone gene. *Gene*, 254(1-2):9–18, 2000.

[38] O. Clay and G. Bernardi. Compositional heterogeneity within and among isochores in mammalian genomes. II. Some general comments. *Gene*, 276(1-2):25–31, 2001.

[39] R. G. Clegg and M. Dodson. Markov chain-based method for generating long-range dependence. *Phys Rev E*, 72(2 Pt 2):026118, 2005.

[40] F. S. Collins, M. L. Drumm, J. L. Cole, W. K. Lockwood, G. F. V. Woude, and M. C. Iannuzzi. Construction of a general human chromosome jumping library, with application to cystic fibrosis. *Science*, 235(4792):1046–9, 1987.

[41] H. J. Cooke, W. R. Brown, and G. A. Rappold. Hypervariable telomeric sequences from the human sex chromosomes are pseudoautosomal. *Nature*, 317(6039):687–92, 1985.

[42] C. Coulondre, J. H. Miller, P. J. Farabaugh, and W. Gilbert. Molecular basis of base substitution hotspots in Escherichia coli. *Nature*, 274(5673):775–780, 1978.

[43] F. H. Crick. On protein synthesis. *Symp Soc Exp Biol*, 12:138–63, 1958.

[44] F. H. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–3, 1970.

[45] C. Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London, UK, 1859.

[46] N. de la Chaux, P. W. Messer, and P. F. Arndt. DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage. *BMC Evol Biol*, 7(1):191, 2007.

[47] Drosophila 12 Genomes Consortium. Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, 450(7167):203–18, 2007.

[48] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis*. Cambridge University Press, Cambridge, UK, 1998.

[49] H. Ellegren. Characteristics, causes and evolutionary consequences of male-biased mutation. *Proc Biol Sci*, 274(1606):1–10, 2007.

[50] EMBL-EBI. Mapping of go terms to go slim terms. Website, 2007. `ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/goslim/goaslim.map`.

[51] A. Eyre-Walker. The genomic rate of adaptive evolution. *Trends Ecol Evol*, 21(10):569–75, 2006.

[52] J. C. Fay, G. J. Wyckoff, and C.-I. Wu. Testing the neutral theory of molecular evolution with genomic data from Drosophila. *Nature*, 415(6875):1024–6, 2002.

[53] W. Feller. *An Introduction to Probability Theory and Its Applications*. Wiley, New York, 1966.

[54] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, 2004.

[55] W. M. Fitch. Homology a personal view on some of the problems. *Trends Genet*, 16(5):227–31, 2000.

[56] D. Freije, C. Helms, M. S. Watson, and H. Donis-Keller. Identification of a second pseudoautosomal region near the Xq and Yq telomeres. *Science*, 258(5089):1784–7, 1992.

[57] M. B. Gerstein, C. Bruce, J. S. Rozowsky, D. Zheng, J. Du, J. O. Korbel, O. Emanuelsson, Z. D. Zhang, S. Weissman, and M. Snyder. What is a gene, post-ENCODE? History and updated definition. *Genome Res*, 17(6):669–81, 2007.

[58] D. Graur and W.-H. Li. *Fundamentals of molecular evolution*. Sinauer Associates, 2000.

[59] J. A. M. Graves. Sex chromosome specialization and degeneration in mammals. *Cell*, 124(5):901–14, 2006.

[60] T. R. Gregory, editor. *The Evolution of the Genome*. Elsevier Academic Press, 2005.

[61] A. J. F. Griffiths, W. M. Gelbart, J. H. Miller, and R. C. Lewontin. *Modern Genetic Analysis*. W. H. Freeman and Company, 1999.

[62] S. Grossmann and B. Yakir. Large deviations for global maxima of independent superadditive processes with negative drift and an application to optimal sequence alignment. *Bernoulli*, 10:829–845, 2004.

[63] R. C. Hardison. Comparative genomics. *PLoS Biol*, 1(2):E58, 2003.

[64] M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, 22(2):160–174, 1985.

[65] S. T. Hess, J. D. Blake, and R. D. Blake. Wide variations in neighbor-dependent substitution rates. *J Mol Biol*, 236(4):1022–33, 1994.

[66] D. Holste, I. Grosse, S. Beirer, P. Schieg, and H. Herzel. Repeats and correlations in human DNA sequences. *Phys Rev E*, 67(6):061913, 2003.

[67] T. J. P. Hubbard, B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts *et al.* Ensembl 2007. *Nucleic Acids Res*, 35:D610–D617, 2006.

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

[69] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

[70] Z. Jiang, H. Tang, M. Ventura, M. F. Cardone, T. Marques-Bonet, X. She, P. A. Pevzner, and E. E. Eichler. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat Genet*, 39(11):1361–8, 2007.

[71] M. A. Jobling and C. Tyler-Smith. The human Y chromosome: an evolutionary marker comes of age. *Nat Rev Genet*, 4(8):598–612, 2003.

[72] T. H. Jukes and C. R. Cantor. Evolution of protein molecules. In H. N. Munro, editor, *Mammalian protein metabolism*, pages 21–132. Academic Press, New York, 1969.

[73] S. Karlin and S. F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci USA*, 87(6):2264–8, 1990.

[74] S. Karlin and S. F. Altschul. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc Natl Acad Sci USA*, 90(12):5873–7, 1993.

[75] S. Karlin and V. Brendel. Patchiness and correlations in DNA sequences. *Science*, 259(5095):677–80, 1993.

[76] S. Karlin and A. Dembo. Limit distributions of the maximal segmental score among Markov-dependent partial sums. *Adv Appl Prob*, 24:113–140, 1992.

[77] W. J. Kent, R. Baertsch, A. Hinrichs, W. Miller, and D. Haussler. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci USA*, 100(20):11484–9, 2003.

[78] P. Khaitovich, I. Hellmann, W. Enard, K. Nowick, M. Leinweber, H. Franz, G. Weiss, M. Lachmann, and S. Pääbo. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science*, 309(5742):1850–1854, 2005.

[79] M. Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, 1983.

[80] M. Kimura and T. Ohta. The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, 61(3):763–771, 1969.

[81] M. Kimura and T. Ohta. Protein polymorphism as a phase of molecular evolution. *Nature*, 229:467–469, 1971.

[82] J. L. King and T. H. Jukes. Non-Darwinian evolution. *Science*, 164(881):788–98, 1969.

[83] M. C. King and A. C. Wilson. Evolution at two levels in humans and chimpanzees. *Science*, 188(4184):107–116, 1975.

[84] A. S. Kondrashov and I. B. Rogozin. Context of deletions and insertions in human coding sequences. *Hum Mutat*, 23(2):177–85, 2004.

[85] T. G. Krontiris. Minisatellites and human disease. *Science*, 269(5231):1682–3, 1995.

[86] E. M. Kvikstad, S. Tyekucheva, F. Chiaromonte, and K. D. Makova. A macaque's-eye view of human insertions and deletions: Differences in mechanisms. *PLoS Comput Biol*, 3(9):e176, 2007.

[87] B. T. Lahn, N. M. Pearson, and K. Jegalian. The human Y chromosome, in the light of evolution. *Nat Rev Genet*, 2(3):207–16, 2001.

[88] G. Levinson and G. A. Gutman. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol*, 4(3):203–221, 1987.

[89] B. Lewin. *Genes VII*. Oxford University Press, Oxford, 2000.

[90] W. Li. Spatial 1/f spectra in open dynamical-systems. *Europhys Lett*, 10(5):395–400, 1989.

[91] W. Li. Expansion-modification systems: A model for spatial 1/f spectra. *Phys Rev A*, 43(10):5240–5260, 1991.

[92] W. Li and D. Holste. Spectral analysis of guanine and cytosine fluctuations of mouse genomic DNA. *Fluct Noise Lett*, 4(3):453–464, 2004.

[93] W. Li and D. Holste. Universal 1/f noise, crossovers of scaling exponents, and chromosome-specific patterns of guanine-cytosine content in DNA sequences of the human genome. *Phys Rev E*, 71(4):041910, 2005.

[94] W. Li and K. Kaneko. Long-range correlation and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence. *Europhys Lett*, 17:655–660, 1992.

[95] W.-H. Li. *Molecular Evolution*. Sinauer Associates, 1997.

[96] W.-H. Li, S. Yi, and K. Makova. Male-driven evolution. *Curr Opin Genet Dev*, 12(6):650–6, 2002.

[97] M. R. Lieber, Y. Ma, U. Pannicke, and K. Schwarz. Mechanism and regulation of human non-homologous DNA end-joining. *Nat Rev Mol Cell Biol*, 4(9):712–720, 2003.

[98] P. Lio and N. Goldman. Models of molecular evolution and phylogeny. *Genome Res*, 8(12):1233–1244, 1998.

[99] C. D. Livingstone and G. J. Barton. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput Appl Biosci*, 9(6):745–756, 1993.

[100] G. Lunter and J. Hein. A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics*, 20 Suppl 1:I216–I223, 2004.

[101] M. Lynch and J. S. Conery. The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155, 2000.

[102] K. D. Makova, S. Yang, and F. Chiaromonte. Insertions and deletions are male biased too: a whole-genome analysis in rodents. *Genome Res*, 14(4):567–73, 2004.

[103] H. A. Makse, S. Havlin, M. Schwartz, and H. E. Stanley. Method for generating long-range correlations for large systems. *Phys Rev E*, 53(5):5445–5449, 1996.

[104] J. Maynard-Smith. *Evolutionary Genetics*. Oxford University Press, 1989.

[105] J. H. McDonald and M. Kreitman. Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, 351(6328):652–4, 1991.

[106] P. W. Messer. Solvable sequence evolution models and genomic correlations. Master's thesis, Institute For Theoretical Physics, University of Cologne, Germany, 2005.

[107] P. W. Messer. Table of all identified indels. Website, 2007. `http://evogen.molgen.mpg.de/data/indels.txt.gz`.

[108] P. W. Messer. Table of identified indels in coding regions. Website, 2007. `http://evogen.molgen.mpg.de/data/coding_indels41.txt`.

[109] P. W. Messer and P. F. Arndt. CorGen–measuring and generating long-range correlations for DNA sequence analysis. *Nucleic Acids Res*, 34:W692–5, 2006.

[110] P. W. Messer and P. F. Arndt. The majority of recent short DNA insertions in the human genome are tandem duplications. *Mol Biol Evol*, 24(5):1190–7, 2007.

[111] P. W. Messer, P. F. Arndt, and M. Lassig. Solvable sequence evolution models and genomic correlations. *Phys Rev Lett*, 94(13):138103, 2005.

[112] P. W. Messer, R. Bundschuh, M. Vingron, and P. F. Arndt. Alignment statistics for long-range correlated genomic sequences. In *Proceedings of the 10th Annual International Conference, RECOMB 2006*, Lecture Notes in Computer Science. Springer, Berlin, 2006.

[113] P. W. Messer, R. Bundschuh, M. Vingron, and P. F. Arndt. Effects of long-range correlations in DNA on sequence alignment score statistics. *J Comput Biol*, 14(5):655–68, 2007.

[114] P. W. Messer, M. Lassig, and P. F. Arndt. Universality of long-range correlations in expansion-randomization systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(10):P10004, 2005.

[115] W. Miller, K. D. Makova, A. Nekrutenko, and R. C. Hardison. Comparative genomics. *Annu Rev Genomics Hum Genet*, 5:15–56, 2004.

[116] R. E. Mills, C. T. Luttig, C. E. Larkins, A. Beauchamp, C. Tsui, W. S. Pittard, and S. E. Devine. An initial map of insertion and deletion (indel) variation in the human genome. *Genome Res*, 16(9):1182–1190, 2006.

[117] T. H. Morgan. *A critique of the theory of evolution*. Princeton University Press, 1916.

[118] R. Mott. Maximum likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores. *Bull Math Biol*, 54:59–75, 1999.

[119] S. Mumm, B. Molini, J. Terrell, A. Srivastava, and D. Schlessinger. Evolutionary features of the 4-Mb Xq21.3 XY homology region revealed by a map at 60-kb resolution. *Genome Res*, 7(4):307–14, 1997.

[120] NCBI. Basic Local Alignment Search Tool. Website, 2007. `http://www.ncbi.nlm.nih.gov/BLAST`.

[121] NCBI. The Reference Sequence Collection. Website, 2007. `http://www.ncbi.nlm.nih.gov/RefSeq`.

[122] R. Nielsen, editor. *Statistical Methods in Molecular Evolution*. Springer, Berlin, 2005.

[123] C. Notredame. Recent evolutions of multiple sequence alignment algorithms. *PLoS Comput Biol*, 3(8):e123, 2007.

[124] S. Ohno, U. Wolf, and N. B. Atkin. Evolution from fish to mammals by gene duplication. *Hereditas*, 59(1):169–187, 1968.

[125] E. M. Ostertag and H. H. Kazazian. Biology of mammalian L1 retrotransposons. *Annu Rev Genet*, 35:501–38, 2001.

[126] J. Overington, D. Donnelly, M. S. Johnson, A. Sali, and T. L. Blundell. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci*, 1(2):216–226, 1992.

[127] D. C. Page, M. E. Harper, J. Love, and D. Botstein. Occurrence of a transposition from the X-chromosome long arm to the Y-chromosome short arm during human evolution. *Nature*, 311(5982):119–23, 1984.

[128] Y. Park, S. Sheetlin, and J. L. Spouge. Accelerated convergence and robust asymptotic regression of the Gumbel scale parameter for gapped sequence alignment. *J Physics A*, 38:97–108, 2005.

[129] S. Pascarella and P. Argos. Analysis of insertions/deletions in protein structures. *J Mol Biol*, 224(2):461–471, 1992.

[130] H. Pearson. Genetics: what is a gene? *Nature*, 441(7092):398–401, 2006.

[131] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley. Long-range correlations in nucleotide sequences. *Nature*, 356(6365):168–70, 1992.

[132] P. Pfeiffer, S. Thode, J. Hancke, and W. Vielmetter. Mechanisms of overlap formation in nonhomologous DNA end joining. *Mol Cell Biol*, 14(2):888–895, 1994.

[133] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C*. Cambridge University Press, 1997.

[134] A. Razin and A. D. Riggs. DNA methylation and gene function. *Science*, 210(4470):604–10, 1980.

[135] R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen *et al.* Global variation in copy number in the human genome. *Nature*, 444(7118):444–54, 2006.

[136] Rhesus Macaque Genome Sequencing and Analysis Consortium. Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, 316(5822):222–34, 2007.

[137] D. B. Roth, T. N. Porter, and J. H. Wilson. Mechanisms of nonhomologous recombination in mammalian cells. *Mol Cell Biol*, 5(10):2599–2607, 1985.

[138] P. C. Sabeti, S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. S. Mikkelsen, D. Altshuler, and E. S. Lander. Positive natural selection in the human lineage. *Science*, 312(5780):1614–1620, 2006.

[139] S. A. Sawyer, J. Parsch, Z. Zhang, and D. L. Hartl. Prevalence of positive selection among nearly neutral amino acid replacements in Drosophila. *Proc Natl Acad Sci USA*, 104(16):6504–10, 2007.

[140] S. Saxonov, P. Berg, and D. L. Brutlag. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci USA*, 103(5):1412–7, 2006.

[141] J. Schmutz, J. Wheeler, J. Grimwood, M. Dickson, J. Yang, C. Caoile, E. Bajorek, S. Black, Y. M. Chan, M. Denys *et al.* Quality assessment of the human genome sequence. *Nature*, 429(6990):365–8, 2004.

[142] A. Schwartz, D. Chan, L. Brown, R. Alagappan, D. Pettay, C. Disteche, B. McGillivray, A. de la Chapelle, and D. Page. Reconstructing hominid Y evolution: X-homologous block, created by X-Y transposition, was disrupted by Yp inversion through LINE-LINE recombination. *Hum Mol Genet*, 7(1):1–11, 1998.

[143] M. C. Simmler, F. Rouyer, G. Vergnaud, M. Nystrom-Lahti, K. Y. Ngo, A. de la Chapelle, and J. Weissenbach. Pseudoautosomal DNA sequences in the pairing region of the human sex chromosomes. *Nature*, 317(6039):692–7, 1985.

[144] R. R. Sinden and R. D. Wells. DNA structure, mutations, and human genetic disease. *Curr Opin Biotechnol*, 3(6):612–22, 1992.

[145] S. Sinha and E. D. Siggia. Sequence turnover and tandem repeats in cis-regulatory modules in Drosophila. *Mol Biol Evol*, 22(4):874–885, 2005.

[146] H. Skaletsky, T. Kuroda-Kawaguchi, P. J. Minx, H. S. Cordum, L. Hillier, L. G. Brown, S. Repping, T. Pyntikova, J. Ali, T. Bieri *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, 423(6942):825–37, 2003.

[147] N. G. Smith and A. Eyre-Walker. Adaptive protein evolution in Drosophila. *Nature*, 415(6875):1022–4, 2002.

[148] S. F. Smith and M. S. Waterman. Comparison of biosequences. *Adv Appl Math*, 2:482–489, 1981.

[149] T. F. Smith, M. S. Waterman, and C. Burks. The statistical distribution of nucleic acid similarities. *Nucleic Acids Res*, 13(2):645–56, 1985.

[150] D. Sornette. *Critical Phenomena in Natural Sciences*. Springer, Berlin, 2004.

[151] H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, Z. D. Goldberger, S. Havlin, R. N. Mantegna, S. M. Ossadnik, C. K. Peng, and M. Simons. Statistical mechanics in biology: how ubiquitous are long-range correlations? *Physica A*, 205(1-3):214–53, 1994.

[152] G. R. Sutherland and R. I. Richards. Simple tandem DNA repeats and human genetic disease. *Proc Natl Acad Sci USA*, 92(9):3636–41, 1995.

[153] M. S. Taylor, C. P. Ponting, and R. R. Copley. Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes. *Genome Res*, 14(4):555–566, 2004.

[154] E. E. Thomas. Short, local duplications in eukaryotic genomes. *Curr Opin Genet Dev*, 15(6):640–4, 2005.

[155] E. E. Thomas, N. Srebro, J. Sebat, N. Navin, J. Healy, B. Mishra, and M. Wigler. Distribution of short paired duplications in mammalian genomes. *Proc Natl Acad Sci USA*, 101(28):10349–10354, 2004.

[156] J. W. Thomas, J. W. Touchman, R. W. Blakesley, G. G. Bouffard, S. M. Beckstrom-Sternberg, E. H. Margulies, M. Blanchette, A. C. Siepel, P. J. Thomas, J. C. McDowell *et al* . Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, 424(6950):788–793, 2003.

[157] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. D. Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent *et al*. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, 23(1):137–44, 2005.

[158] G. Toth, Z. Gaspari, and J. Jurka. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res*, 10(7):967–981, 2000.

[159] Y. Y. Tseng and J. Liang. Estimation of amino acid residue substitution rates at local spatial regions and application in protein function inference: a Bayesian Monte Carlo approach. *Mol Biol Evol*, 23(2):421–436, 2006.

[160] D. C. van Gent, J. H. Hoeijmakers, and R. Kanaar. Chromosomal stability and the DNA double-stranded break connection. *Nat Rev Genet*, 2(3):196–206, 2001.

[161] E. Viguera, D. Canceill, and S. D. Ehrlich. Replication slippage involves DNA polymerase pausing and dissociation. *EMBO J*, 20(10):2587–95, 2001.

[162] B. F. Voight, S. Kudaravalli, X. Wen, and J. K. Pritchard. A map of recent positive selection in the human genome. *PLoS Biol*, 4(3):e72, 2006.

[163] R. F. Voss. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Phys Rev Lett*, 68(25):3805–3808, 1992.

[164] X. J. Wang. Statistical physics of temporal intermittency. *Phys Rev A*, 40(11):6647–6661, 1989.

[165] S. T. Warren, F. Zhang, G. R. Licameli, and J. F. Peters. The fragile X site in somatic cell hybrids: an approach for molecular cloning of fragile sites. *Science*, 237(4813):420–3, 1987.

[166] H. Watanabe, A. Fujiyama, M. Hattori, T. D. Taylor, A. Toyoda, Y. Kuroki, H. Noguchi, A. BenKahla, H. Lehrach, R. Sudbrak *et al.* DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature*, 429(6990):382–388, 2004.

[167] M. S. Waterman. *Introduction to computational biology: maps, sequences, and genomes*. CRC Press, 1995.

[168] M. S. Waterman and M. Vingron. Rapid and accurate estimates of statistical significance for sequence data base searches. *Proc Natl Acad Sci USA*, 91(11):4625–8, 1994.

[169] R. H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002.

[170] J. D. Watson. *The Double Helix*. Weidenfeld and Nichols, 1968.

[171] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–8, 1953.

[172] M. T. Webster, N. G. C. Smith, and H. Ellegren. Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. *Proc Natl Acad Sci USA*, 99(13):8748–8753, 2002.

[173] O. Weiss and H. Herzel. Correlations in protein sequences and property codes. *J Theor Biol*, 190(4):341–53, 1998.

[174] K. H. Wolfe, P. M. Sharp, and W.-H. Li. Mutation rates differ among regions of the mammalian genome. *Nature*, 337(6204):283–5, 1989.

[175] J. O. Wrabl and N. V. Grishin. Gaps in structurally similar proteins: towards improvement of multiple sequence alignment. *Proteins*, 54(1):71–87, 2004.

[176] Y. K. Yu, R. Bundschuh, and T. Hwa. Statistical significance and extremal ensemble of gapped local hybrid alignment. In M. Lassig and A. Valleriani, editors, *Biological Evolution and Statistical Physics*, Lecture Notes in Physics. Springer, Berlin, 2002.

[177] Z. Zhang and M. Gerstein. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res*, 31(18):5338–5348, 2003.

[178] E. Zuckerkandl and L. Pauling. Molecular disease, evolution, and genetic heterogeneity. In M. Kasha and B. Pullman, editors, *Horizons in Biochemistry*, pages 189–225. Academic Press, New York, 1962.

[179] E. Zuckerkandl and L. Pauling. Molecules as documents of evolutionary history. *J Theor Biol*, 8(2):357–366, 1965.

Notation and abbreviations

Chapter 1

DNA.....	Deoxyribonucleic acid
RNA.....	Ribonucleic acid
RS.....	replication slippage
UCO.....	unequal crossing over
$\vec{s}$.....	nucleotide sequence $\vec{s} = (s_1, \ldots, s_N)$
$\mathbf{Q}$.....	instantaneous rate matrix
$\vec{\rho}$.....	nucleotide frequencies $\vec{\rho} = (\rho_A, \rho_C, \rho_G, \rho_T)$
$\vec{\pi}$.....	stationary nucleotide distribution $\vec{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$
$P_{ij}(r)$.....	joint probability to find nucleotides $s_x = i$ and $s_{x+r} = j$
$C(r)$.....	correlation function $C(r) = \sum_i [P_{ii}(r) - \rho_i^2]$

Chapter 2

SSR.....	simple sequence repeat
NHEJ.....	nonhomologous end joining
l.....	length of an indel
d.....	trace extension of an indel
I, D.....	insertion rate, deletion rate (bp/Mbp)
p.....	p-value
z.....	z-score, measures difference from mean in standard deviations
PDB.....	Protein Data Bank
GO.....	Gene Ontology

Chapter 3

$\vec{s}$.....	binary sequence with $s_i = \pm 1$ (in section 3.7, $s_i \in \{A, C, G, T\}$)
$\mu, \delta_\ell, \gamma_\ell^+, \gamma_\ell^-$..	rates of mutation, duplication, random insertion, deletion
ℓ, $\ell_{\max}$.....	segment length, maximal segment length
k, r, t.....	sequence position, distance between positions, time
$\lambda, \mu_{\text{eff}}$.....	growth rate, effective mutation rate
$\langle\cdot\rangle$.....	ensemble average
$P_k^{\pm}(t)$.....	probability to find $s_k = \pm 1$ at time t
$P_{\text{eq(op)}}(k, r)$..	joint probability $\Pr[s_k = (\neq) s_{k+r}]$
$C(r)$.....	correlation function $C(r) = \langle s_k s_{k+r} \rangle$ in the binary model, in the four-letter model $C(r) = \sum_i [P_{ii}(r) - \rho_i^2]$

χ, α.....	scaling exponents $\alpha = 2\chi$
$m(L)$.....	composition bias in a sequence segment of finite length L
$\mathcal{S}(x), \mathcal{P}_\chi(x)$..	scaling functions
η_ℓ, ν_ℓ.....	biased insertion rate, bias parameter
ϵ.....	Monte Carlo step-size parameter
$\mathbf{q}$.....	mutation rate matrix of the four-letter model
μ, g.....	mutation parameter, GC-content
π_x.....	stationary frequency of nucleotide $x \in \{A, C, G, T\}$
$\vec{P}(r)$.....	16-dimensional vector of all $P_{ij}(r) = \Pr[s_x = i \wedge s_{x+r} = j]$
$\vec{p}, \vec{P}_0$.....	$\vec{P}(r) = \vec{p}\, r^{-\alpha} + \vec{P}_0$
$\mathbf{Q}$.....	16 by 16 matrix $\mathbf{Q} = \mathbf{I}_4 \otimes \mathbf{q} + \mathbf{q} \otimes \mathbf{I}_4$

Chapter 4

$\vec{a}, \vec{b}$.....	nucleotide sequences $\vec{a} = (a_1, \ldots, a_N),\ \vec{b} = (b_1, \ldots, b_M)$
$\mathcal{A}$.....	alignment
$s(a, b)$.....	match-mismatch scoring matrix
μ.....	mismatch penalty
S.....	alignment score
$\gamma, \gamma_i, \gamma_e$.....	gap cost, gap initiation cost, gap extension cost
K.....	parameter determining the mean of a Gumbel distribution
λ.....	decay parameter of a Gumbel distribution
ρ_x.....	frequency of nucleotide x
h.....	score of a global alignment
$Z_N(\lambda)$.....	generating function
$\vec{s}$.....	diagonal vector of scores in the alignment lattice
pdf(X).....	probability density function of random variable X
$P(X)$.....	probability of X
$\boldsymbol{\sigma}$.....	covariance matrix
$C(r)$.....	correlation function $C(r) = \sum_i [P_{ii}(r) - \rho_i^2]$
α.....	scaling exponent of long-range correlations
β.....	decay parameter of short-range correlations
Γ.....	Euler-Mascheroni constant $\Gamma \approx 0.5772$