



Universiteit  
Leiden  
The Netherlands

## **Metagenomics : beyond the horizon of current implementations and methods**

Khachatryan, L.

### **Citation**

Khachatryan, L. (2020, April 28). *Metagenomics : beyond the horizon of current implementations and methods*. Retrieved from <https://hdl.handle.net/1887/87513>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/87513>

**Note:** To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/87513> holds various files of this Leiden University dissertation.

**Author:** Khachatryan, L.

**Title:** Metagenomics : beyond the horizon of current implementations and methods

**Issue Date:** 2020-04-28

---

---

## Bibliography

- [1] WB Whitman, DC Coleman, and WJ Wiebe. Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences*, 95(12):6578–6583, 1998.
- [2] AA Juwarkar, SK Yadav, PR Thawale, P Kumar, SK Singh, and T Chakrabarti. Developmental strategies for sustainable ecosystem on mine spoil dumps: a case of study. *Environmental monitoring and assessment*, 157(1-4):471–481, 2009.
- [3] C Pedros-Alio. Dipping into the rare biosphere. *Science*, 5809:192, 2007.
- [4] C Pedros-Alio. The rare bacterial biosphere. *Annual review of marine science*, 4:449–466, 2012.
- [5] GT Pecl, MB Araujo, JD Bell, J Blanchard, TC Bonebrake, I-C Chen, TD Clark, RK Colwell, F Danielsen, B Evengard, et al. Biodiversity redistribution under climate change: Impacts on ecosystems and human well-being. *Science*, 355(6332):eaai9214, 2017.
- [6] K Todar. Bacteria and archaea and the cycles of elements in the environment. Retrieved June, 7:2014, 2012.
- [7] M Paumann, G Regelsberger, C Obinger, and GA Peschek. The bioenergetic role of dioxygen and the terminal oxidase(s) in cyanobacteria. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1707(2-3):231–253, 2005.
- [8] PG Falkowski and LV Godfrey. Electrons, life and the evolution of earth's oxygen cycle. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 363(1504):2705–2716, 2008.
- [9] C Gougoulas, JM Clark, and LJ Shaw. The role of soil microbes in the global carbon cycle: tracking the below-ground microbial processing of plant-derived carbon for manipulating carbon dynamics in agricultural systems. *Journal of the Science of Food and Agriculture*, 94(12):2362–2371, 2014.
- [10] MG Klotz, DA Bryant, and TE Hanson. The microbial sulfur cycle. *Frontiers in microbiology*, 2:241, 2011.
- [11] J Chen, Y Li, Y Tian, C Huang, D Li, Q Zhong, and X Ma. Interaction between microbes and host intestinal health: modulation by dietary nutrients and gut-brain-endocrine-immune axis. *Current Protein and Peptide Science*, 16(7):592–603, 2015.
- [12] SE Erdman and T Poutahidis. Microbes and oxytocin: benefits for host physiology and behavior. In *International review of neurobiology*, volume 131, pages 91–126. Elsevier, 2016.
- [13] E Patterson, JF Cryan, GF Fitzgerald, RP Ross, TG Dinan, and C Stanton. Gut

- microbiota, the pharmabiotics they produce and host health. *Proceedings of the Nutrition Society*, 73(4):477–489, 2014.
- [14] LK Ursell, W van Treuren, JL Metcalf, M Pirrung, A Gewirtz, and R Knight. Replenishing our defensive microbes. *Bioessays*, 35(9):810–817, 2013.
  - [15] TA Van der Meulen, HJM Harmesen, H Bootsma, FKL Spijkervet, FGM Kroese, and A Vissink. The microbiome–systemic diseases connection. *Oral diseases*, 22(8):719–734, 2016.
  - [16] GJM Christensen and H Bruggemann. Bacterial skin commensals and their role as host guardians. *Beneficial microbes*, 5(2):201–215, 2013.
  - [17] Y Belkaid and S Tamoutounour. The influence of skin microorganisms on cutaneous immunity. *Nature Reviews Immunology*, 16(6):353, 2016.
  - [18] PC Calder. Feeding the immune system. *Proceedings of the Nutrition Society*, 72(3):299–309, 2013.
  - [19] DA Cowan, J-B Ramond, TP Makhalanyane, and P de Maayer. Metagenomics of extreme environments. *Current opinion in microbiology*, 25:97–102, 2015.
  - [20] P Blum. *Archaea: Ancient Microbes, Extreme Environments, and the Origin of Life*, volume 50. Gulf Professional Publishing, 2001.
  - [21] L Tazi, DP Breakwell, AR Harker, and KA Crandall. Life in extreme environments: microbial diversity in great salt lake, utah. *Extremophiles*, 18(3):525–535, 2014.
  - [22] C Bang, T Dagan, P Deines, N Dubilier, WJ Duschl, S Fraune, U Hentschel, H Hirt, N Hulter, T Lachnit, et al. Metaorganisms in extreme environments: do microbes play a role in organismal adaptation? *Zoology*, 2018.
  - [23] J Rifkin. *The biotech century*. Sonoma County Earth First/Biotech Last, 1998.
  - [24] S Farmer. Probiotic, lactic acid-producing bacteria and uses thereof, October 8 2002. US Patent 6,461,607.
  - [25] JR Postgate. Economic importance of sulphur bacteria. *Phil. Trans. R. Soc. Lond. B*, 298(1093):583–600, 1982.
  - [26] M Fernandez, B Del Rio, DM Linares, MC Martin, and MA Alvarez. Real-time polymerase chain reaction for quantitative detection of histamine-producing bacteria: use in cheese production. *Journal of dairy science*, 89(10):3763–3769, 2006.
  - [27] A Mayra-Makinen and M Bigret. Industrial use and production of lactic acid bacteria. *Food science and Technology*, 139:175–198, 2004.
  - [28] BS Dien, MA Cotta, and TW Jeffries. Bacteria engineered for fuel ethanol production: current status. *Applied microbiology and biotechnology*, 63(3):258–266, 2003.
  - [29] SF Bender, C Wagg, and MGA van der Heijden. An underground revolution: biodiversity and soil ecological engineering for agricultural sustainability. *Trends in Ecology & Evolution*, 31(6):440–452, 2016.
  - [30] M-N Xing, X-Z Zhang, and H Huang. Application of metagenomic techniques in mining enzymes from microbial communities for biofuel synthesis. *Biotechnology advances*, 30(4):920–929, 2012.
  - [31] M Wainwright, J Lederberg, and J Lederberg. History of microbiology. *Encyclopedia of microbiology*, 2:419–437, 1992.
  - [32] Y Stanier, M Doudoroff, EA Adelberg, et al. General microbiology. *General microbiology*, 1958.

- [33] N Hall. Advanced sequencing technologies and their wider impact in microbiology. *Journal of Experimental Biology*, 210(9):1518–1525, 2007.
- [34] SE Hasnain. Impact of human genome sequencing on microbiology. *Indian journal of medical microbiology*, 19(3):114, 2001.
- [35] JT Staley and A Konopka. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual Reviews in Microbiology*, 39(1):321–346, 1985.
- [36] A Escobar-Zepeda, A Vera-Ponce de Leon, and A Sanchez-Flores. The road to metagenomics: from microbiology to dna sequencing technologies and bioinformatics. *Frontiers in genetics*, 6:348, 2015.
- [37] National Research Council et al. *The new science of metagenomics: revealing the secrets of our microbial planet*. National Academies Press, 2007.
- [38] C Simon and R Daniel. Achievements and new knowledge unraveled by metagenomic approaches. *Applied microbiology and biotechnology*, 85(2):265–276, 2009.
- [39] R Knight, A Vrbanc, B C Taylor, A Aksenov, C Callewaert, J Debelius, A Gonzalez, T Kosciulek, L-I McCall, D McDonald, et al. Best practices for analysing microbiomes. *Nature Reviews Microbiology*, page 1, 2018.
- [40] S Junemann, N Kleinbolting, S Jaenicke, C Henke, J Hassa, J Nelkner, Y Stolze, S P Albaum, A Schluter, A Goesmann, et al. Bioinformatics for ngs-based metagenomics and the application to biogas research. *Journal of biotechnology*, 261:10–23, 2017.
- [41] DJ Lane, B Pace, GJ Olsen, DA Stahl, ML Sogin, and NR Pace. Rapid determination of 16s ribosomal rna sequences for phylogenetic analyses. *Proceedings of the National Academy of Sciences*, 82(20):6955–6959, 1985.
- [42] TM Schmidt, EF DeLong, and NR Pace. Analysis of a marine picoplankton community by 16s rna gene cloning and sequencing. *Journal of bacteriology*, 173(14):4371–4378, 1991.
- [43] PJ Turnbaugh, RE Ley, M Hamady, CM Fraser-Liggett, R Knight, and JI Gordon. The human microbiome project. *Nature*, 449(7164):804, 2007.
- [44] HMP Integrative. The integrative human microbiome project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell host & microbe*, 16(3):276, 2014.
- [45] V Robles-Alonso and F Guarner. From basic to applied research: lessons from the human microbiome projects. *Journal of clinical gastroenterology*, 48:S3–S4, 2014.
- [46] SM Bakhtiar, JG LeBlanc, E Salvucci, A Ali, R Martin, P Langella, J-M Chatel, A Miyoshi, LG Bermudez-Humaran, and V Azevedo. Implications of the human microbiome in inflammatory bowel diseases. *FEMS microbiology letters*, 342(1):10–17, 2013.
- [47] J Lloyd-Price, G Abu-Ali, and C Huttenhower. The healthy human microbiome. *Genome medicine*, 8(1):51, 2016.
- [48] XC Morgan, N Segata, and C Huttenhower. Biodiversity and functional genomics in the human microbiome. *Trends in genetics*, 29(1):51–58, 2013.
- [49] X C Morgan, T L Tickle, H Sokol, D Gevers, K L Devaney, D V Ward, J A Reyes, S A Shah, N LeLeiko, S B Snapper, et al. Dysfunction of the intestinal

- microbiome in inflammatory bowel disease and treatment. *Genome biology*, 13(9):R79, 2012.
- [50] D Gevers, S Kugathasan, L A Denson, Y Vazquez-Baeza, W Van Treuren, B Ren, E Schwager, D Knights, S J Song, M Yassour, et al. The treatment-naïve microbiome in new-onset crohn's disease. *Cell host & microbe*, 15(3):382–392, 2014.
- [51] C Pedros-Alio. Genomics and marine microbial ecology. *International Microbiology*, 9(3):191–197, 2006.
- [52] PG Falkowski, RT Barber, and V Smetacek. Biogeochemical controls and feedbacks on ocean primary production. *Science*, 281(5374):200–206, 1998.
- [53] CA Suttle. Marine viruses—major players in the global ecosystem. *Nature Reviews Microbiology*, 5(10):801, 2007.
- [54] NA Bokulich, ZT Lewis, K Boundy-Mills, and DA Mills. A new perspective on microbial landscapes within food production. *Current opinion in biotechnology*, 37:182–189, 2016.
- [55] M Trindade, LJ van Zyl, J Navarro-Fernández, and A Abd Elrazak. Targeted metagenomics as a tool to tap into marine natural product diversity for the discovery and production of drug candidates. *Frontiers in microbiology*, 6:890, 2015.
- [56] MM Zhang, Y Qiao, EL Ang, and H Zhao. Using natural products for drug discovery: the impact of the genomics era. *Expert opinion on drug discovery*, 12(5):475–487, 2017.
- [57] SM Techtmann and TC Hazen. Metagenomic applications in environmental monitoring and bioremediation. *Journal of industrial microbiology & biotechnology*, 43(10):1345–1354, 2016.
- [58] JL Metcalf, ZZ Xu, A Bouslimani, P Dorrestein, DO Carter, and R Knight. Microbiome tools for forensic science. *Trends in biotechnology*, 35(9):814–823, 2017.
- [59] SE Schmedes, AE Woerner, NMM Novroski, FR Wendt, JL King, KM Stephens, and B Budowle. Targeted sequencing of clade-specific markers from skin microbiomes for forensic human identification. *Forensic Science International: Genetics*, 32:50–61, 2018.
- [60] EN Hanssen, E Avershina, K Rudi, P Gill, and L Snipen. Body fluid prediction from microbial patterns for forensic application. *Forensic Science International: Genetics*, 30:10–17, 2017.
- [61] L Brinkac, TH Clarke, H Singh, C Greco, A Gomez, MG Torralba, B Frank, and KE Nelson. Spatial and environmental variation of the human hair microbiota. *Scientific reports*, 8(1):9017, 2018.
- [62] N Fierer, CL Lauber, N Zhou, D McDonald, EK Costello, and R Knight. Forensic identification using skin bacterial communities. *Proceedings of the National Academy of Sciences*, 107(14):6477–6481, 2010.
- [63] S Lax, JT Hampton-Marcell, SM Gibbons, GB Colares, D Smith, JA Eisen, and JA Gilbert. Forensic analysis of the microbiome of phones and shoes. *Microbiome*, 3(1):21, 2015.
- [64] RR Dunn, N Fierer, JB Henley, JW Leff, and HL Menninger. Home life: factors structuring the bacterial diversity found within and between homes. *PloS one*, 8(5):e64133, 2013.
- [65] GE Flores, ST Bates, JG Caporaso, CL Lauber, JW Leff, R Knight, and N Fierer. Diversity, distribution and sources of bacteria in residential

- kitchens. *Environmental microbiology*, 15(2):588–596, 2013.
- [66] GE Flores, ST Bates, D Knights, CL Lauber, J Stombaugh, R Knight, and N Fierer. Microbial biogeography of public restroom surfaces. *PloS one*, 6(11):e28132, 2011.
- [67] S Lax, DP Smith, J Hampton-Marcell, SM Owens, KM Handley, NM Scott, SM Gibbons, P Larsen, BD Shogan, S Weiss, et al. Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science*, 345(6200):1048–1052, 2014.
- [68] SE Schmedes, AE Woerner, and B Budowle. Forensic human identification using skin microbiomes. *Applied and environmental microbiology*, pages AEM-01672, 2017.
- [69] F Schlutzen, A Tocilj, R Zarivach, J Harms, M Gluehmann, D Janell, A Bashan, H Bartels, I Agmon, F Franceschi, et al. Structure of functionally activated small ribosomal subunit at 3.3 Å resolution. *Cell*, 102(5):615–623, 2000.
- [70] CR Woese, O Kandler, and ML Wheelis. Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proceedings of the National Academy of Sciences*, 87(12):4576–4579, 1990.
- [71] CR Woese and GE Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090, 1977.
- [72] WG Weisburg, SM Barns, DA Pelletier, and DJ Lane. 16s ribosomal dna amplification for phylogenetic study. *Journal of bacteriology*, 173(2):697–703, 1991.
- [73] J Wagner, P Coupland, H P Browne, T D Lawley, S C Francis, and J Parkhill. Evaluation of pacbio sequencing for full-length bacterial 16s rRNA gene classification. *BMC microbiology*, 16(1):274, 2016.
- [74] GJ Olsen, R Overbeek, N Larsen, TL Marsh, MJ McCaughey, MA Maciukenas, W-M Kuan, TJ Macke, Y Xing, and CR Woese. The ribosomal database project. *Nucleic Acids Research*, 20(suppl):2199–2200, 1992.
- [75] JR Cole, Q Wang, JA Fish, B Chai, DM McGarrell, Y Sun, CT Brown, A Porras-Alfaro, CR Kuske, and JM Tiedje. Ribosomal database project: data and tools for high throughput rRNA analysis. *Nucleic acids research*, 42(D1):D633–D642, 2013.
- [76] P Yilmaz, LW Parfrey, P Yarza, J Gerken, E Pruesse, C Quast, T Schweer, J Peplies, W Ludwig, and FO Glockner. The SILVA and “all-species living tree project (ltp)” taxonomic frameworks. *Nucleic acids research*, 42(D1):D643–D648, 2013.
- [77] D McDonald, MN Price, J Goodrich, EP Nawrocki, TZ DeSantis, A Probst, GL Andersen, R Knight, and P Hugenholtz. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME journal*, 6(3):610, 2012.
- [78] B Yang, Y Wang, and P-Y Qian. Sensitivity and correlation of hypervariable regions in 16s rRNA genes in phylogenetic analysis. *BMC bioinformatics*, 17(1):135, 2016.
- [79] CM Burke and AE Darling. A method for high precision sequencing of near full-length 16s rRNA genes on an illumina miseq. *PeerJ*, 4:e2492, 2016.
- [80] J Shin, S Lee, M-J Go, SY Lee, SC Kim, C-H Lee, and B-K Cho. Analysis of

- the mouse gut microbiome using full-length 16s rRNA amplicon sequencing. *Scientific reports*, 6:29681, 2016.
- [81] S Ceuppens, D De Coninck, N Botteldoorn, F Van Nieuwerburgh, and M Uyttendaele. Microbial community profiling of fresh basil and pitfalls in taxonomic assignment of enterobacterial pathogenic species based upon 16S rRNA amplicon sequencing. *International journal of food microbiology*, 257:148–156, 2017.
- [82] FE Dewhirst, Z Shen, MS Scimeca, LN Stokes, T Boumenna, T Chen, BJ Paster, and JG Fox. Discordant 16S and 23S rRNA gene phylogenies for the genus *helicobacter*: implications for phylogenetic inference and systematics. *Journal of bacteriology*, 187(17):6106–6118, 2005.
- [83] S Hong, J Bunge, C Leslin, S Jeon, and SS Epstein. Polymerase chain reaction primers miss half of rRNA microbial diversity. *The ISME Journal*, 3(12):1365, 2009.
- [84] PHA Timmers, HCA Widjaja-Greefkes, CM Plugge, and AJM Stams. Evaluation and optimization of PCR primers for selective and quantitative detection of marine anme subclusters involved in sulfate-dependent anaerobic methane oxidation. *Applied microbiology and biotechnology*, 101(14):5847–5859, 2017.
- [85] RJ Case, Y Boucher, I Dahllof, C Holmstrom, WF Doolittle, and S Kjelleberg. Use of 16s rRNA and rpoB genes as molecular markers for microbial ecology studies. *Applied and environmental microbiology*, 73(1):278–288, 2007.
- [86] FE Angly, PG Dennis, A Skarshewski, I Vanwonterghem, P Hugenholtz, and GW Tyson. Copyrighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. *Microbiome*, 2(1):11, 2014.
- [87] M Perisin, M Vetter, JA Gilbert, and J Bergelson. 16stimator: statistical estimation of ribosomal gene copy numbers from draft genome assemblies. *The ISME journal*, 10(4):1020, 2016.
- [88] S Louca, M Doebeli, and L W Parfrey. Correcting for 16s rRNA gene copy numbers in microbiome surveys remains an unsolved problem. *Microbiome*, 6(1):41, 2018.
- [89] MGI Langille, J Zaneveld, JG Caporaso, D McDonald, D Knights, JA Reyes, JC Clemente, DE Burkepille, RLV Thurber, R Knight, et al. Predictive functional profiling of microbial communities using 16s rRNA marker gene sequences. *Nature biotechnology*, 31(9):814, 2013.
- [90] R Ranjan, A Rani, A Metwally, HS McGee, and DL Perkins. Analysis of the microbiome: advantages of whole genome shotgun versus 16s amplicon sequencing. *Biochemical and biophysical research communications*, 469(4):967–977, 2016.
- [91] M Tessler, J S Neumann, E Afshinnekoo, M Pineda, R Hersch, L F M Velho, B T Segovia, F A Lansac-Toha, M Lemke, R DeSalle, et al. Large-scale differences in microbial biodiversity discovery between 16s amplicon and shotgun sequencing. *Scientific reports*, 7(1):6589, 2017.
- [92] CB Driscoll, TG Otten, NM Brown, and TW Dreher. Towards long-read metagenomics: complete assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a freshwater lake co-culture. *Standards in genomic sciences*, 12(1):9, 2017.



- [93] BL Brown, M Watson, SS Minot, MC Rivera, and RB Franklin. Minion nanopore sequencing of environmental metagenomes: a synthetic approach. *GigaScience*, 6(3):1–10, 2017.
- [94] Simon Andrews et al. Fastqc: a quality control tool for high throughput sequence data. 2010.
- [95] M Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):pp–10, 2011.
- [96] AM Bolger, M Lohse, and B Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.
- [97] S Lindgreen, KL Adair, and PP Gardner. An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific reports*, 6:19233, 2016.
- [98] PHA Sneath, RR Sokal, et al. *Numerical taxonomy. The principles and practice of numerical classification*. 1973.
- [99] PD Schloss, SL Westcott, T Ryabin, JR Hall, M Hartmann, EB Hollister, RA Lesniewski, BB Oakley, DH Parks, CJ Robinson, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, 75(23):7537–7541, 2009.
- [100] JG Caporaso, J Kuczynski, J Stombaugh, K Bittinger, FD Bushman, EK Costello, N Fierer, AG Pena, JK Goodrich, JI Gordon, et al. QIIME allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5):335, 2010.
- [101] RC Edgar. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461, 2010.
- [102] Stephen F Altschul, W Gish, W Miller, EW Myers, and DJ Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.
- [103] B Buchfink, C Xie, and DH Huson. Fast and sensitive protein alignment using diamond. *Nature methods*, 12(1):59, 2014.
- [104] M Hamada, Y Ono, K Asai, and MC Frith. Training alignment parameters for arbitrary sequencers with last-train. *Bioinformatics*, 33(6):926–928, 2016.
- [105] H Li and R Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics*, 25(14):1754–1760, 2009.
- [106] B Langmead and SL Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357, 2012.
- [107] WJ Kent. BLAT - the blast-like alignment tool. *Genome research*, 12(4):656–664, 2002.
- [108] AV Aho, JE Hopcroft, and JD Ullman. On finding lowest common ancestors in trees. *SIAM Journal on computing*, 5(1):115–132, 1976.
- [109] DH Huson, AF Auch, J Qi, and SC Schuster. Megan analysis of metagenomic data. *Genome research*, 17(3):000–000, 2007.
- [110] DE Wood and SL Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, 15(3):R46, 2014.
- [111] D Kim, L Song, FP Breitwieser, and SL Salzberg. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome research*, 2016.
- [112] M Burrows and DJ Wheeler. A block-sorting lossless data compression algorithm. 1994.

- [113] P Ferragina and G Manzini. Indexing compressed text. *Journal of the ACM (JACM)*, 52(4):552–581, 2005.
- [114] AL Delcher, S Kasif, RD Fleischmann, J Peterson, O White, and SL Salzberg. Alignment of whole genomes. *Nucleic acids research*, 27(11):2369–2376, 1999.
- [115] R Ounit, S Wanamaker, TJ Close, and S Lonardi. Clark: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC genomics*, 16(1):236, 2015.
- [116] TAK Freitas, P-E Li, MB Scholz, and PSG Chain. Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. *Nucleic acids research*, 43(10):e69–e69, 2015.
- [117] N Segata, L Waldron, A Ballarini, V Narasimhan, O Jousson, and C Huttenhower. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*, 9(8):811, 2012.
- [118] F Meyer, D Paarmann, M D’Souza, R Olson, EM Glass, M Kubal, T Paczian, A Rodriguez, R Stevens, A Wilke, et al. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics*, 9(1):386, 2008.
- [119] M Rho, H Tang, and Y Ye. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic acids research*, 38(20):e191–e191, 2010.
- [120] J Droge, I Gregor, and AC McHardy. Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods. *Bioinformatics*, 31(6):817–824, 2014.
- [121] B Liu, T Gibbons, M Ghodsi, T Treangen, and M Pop. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *Genome biology*, 12(1):P11, 2011.
- [122] S Sunagawa, DR Mende, G Zeller, F Izquierdo-Carrasco, SA Berger, JR Kultima, LP Coelho, M Arumugam, J Tap, HB Nielsen, et al. Metagenomic species profiling using universal phylogenetic marker genes. *Nature methods*, 10(12):1196, 2013.
- [123] D Koslicki, S Foucart, and G Rosen. Quikr: a method for rapid reconstruction of bacterial communities via compressive sensing. *Bioinformatics*, 29(17):2096–2102, 2013.
- [124] D Koslicki, S Foucart, and G Rosen. Wgsquikr: fast whole-genome shotgun metagenomic classification. *PloS one*, 9(3):e91784, 2014.
- [125] P Menzel, KL Ng, and A Krogh. Fast and sensitive taxonomic classification for metagenomics with kaiju. *Nature communications*, 7:11257, 2016.
- [126] N-P Nguyen, S Mirarab, B Liu, M Pop, and T Warnow. Tapp: taxonomic identification and phylogenetic profiling. *Bioinformatics*, 30(24):3548–3555, 2014.
- [127] GGZ Silva, DA Cuevas, BE Dutilh, and RA Edwards. Focus: an alignment-free model to identify organisms in metagenomes using non-negative least squares. *PeerJ*, 2:e425, 2014.
- [128] S Hunter, M Corbett, H Denise, M Fraser, A Gonzalez-Beltran, C Hunter, P Jones, R Leinonen, C McAnulla, E Maguire, et al. Ebi metagenomics—a new resource for the analysis and archiving of metagenomic data. *Nucleic acids research*, 42(D1):D600–D606, 2013.
- [129] E Quevillon, V Silventoinen, S Pillai, N Harte, N Mulder, R Apweiler, and R Lopez. Interproscan: protein domains identifier. *Nucleic acids research*, 33(suppl\_2):W116–W120, 2005.

- [130] J-H Lee, H Yi, and J Chun. rrnaselector: a computer program for selecting ribosomal rna encoding sequences from metagenomic and metatranscriptomic shotgun libraries. *The Journal of Microbiology*, 49(4):689, 2011.
- [131] A Bateman, L Coin, R Durbin, RD Finn, V Hollich, S Griffiths-Jones, A Khanna, M Marshall, S Moxon, ELL Sonnhammer, et al. The pfam protein families database. *Nucleic acids research*, 32(suppl\_1):D138–D141, 2004.
- [132] DH Haft, JD Selengut, and O White. The tigrfams database of protein families. *Nucleic acids research*, 31(1):371–373, 2003.
- [133] TK Attwood and ME Beck. Prints—a protein motif fingerprint database. *Protein Engineering, Design and Selection*, 7(7):841–848, 1994.
- [134] N Hulo, A Bairoch, V Bulliard, L Cerutti, E De Castro, PS Langendijk-Genevaux, M Pagni, and CJA Sigrist. The prosite database. *Nucleic acids research*, 34(suppl\_1):D227–D230, 2006.
- [135] DWA Buchan, SCG Rison, JE Bray, D Lee, F Pearl, JM Thornton, and CA Orengo. Gene3d: structural assignments for the biologist and bioinformaticist alike. *Nucleic acids research*, 31(1):469–473, 2003.
- [136] IF Spellerberg and PJ Fedor. A tribute to claude shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the ‘shannon–wiener’ index. *Global ecology and biogeography*, 12(3):177–179, 2003.
- [137] EH Simpson. Measurement of diversity. *nature*, 1949.
- [138] JR Bray and JT Curtis. An ordination of the upland forest communities of southern wisconsin. *Ecological monographs*, 27(4):325–349, 1957.
- [139] C Lozupone, M Hamady, and R Knight. Unifrac—an online tool for comparing microbial community diversity in a phylogenetic context. *BMC bioinformatics*, 7(1):371, 2006.
- [140] A Kislyuk, S Bhatnagar, J Dushoff, and JS Weitz. Unsupervised statistical clustering of environmental shotgun sequences. *BMC bioinformatics*, 10(1):316, 2009.
- [141] DD Roumpeka, RJ Wallace, F Escalettes, I Fotheringham, and M Watson. A review of bioinformatics tools for bio-prospecting from metagenomic sequence data. *Frontiers in genetics*, 8:23, 2017.
- [142] Y-W Wu and Y Ye. A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *Journal of Computational Biology*, 18(3):523–534, 2011.
- [143] Y Wang, HCM Leung, S-M Yiu, and FYL Chin. Metacluster 4.0: a novel binning algorithm for NGS reads and huge number of species. *Journal of Computational Biology*, 19(2):241–249, 2012.
- [144] T Van Lang, T Van Hoai, et al. A two-phase binning algorithm using l-mer frequency on groups of non-overlapping reads. *Algorithms for Molecular Biology*, 10(1):2, 2015.
- [145] K Song, J Ren, G Reinert, M Deng, MS Waterman, and F Sun. New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. *Briefings in bioinformatics*, 15(3):343–353, 2013.
- [146] S Giroto, C Pizzi, and M Comin. Metaprob: accurate metagenomic reads binning based on probabilistic sequence signatures. *Bioinformatics*, 32(17):i567–i575, 2016.

- [147] X Ding, F Cheng, C Cao, and X Sun. Dectico: an alignment-free supervised metagenomic classification method based on feature extraction and dynamic selection. *BMC bioinformatics*, 16(1):323, 2015.
- [148] H Cui and X Zhang. Alignment-free supervised classification of metagenomes by recursive svm. *BMC genomics*, 14(1):641, 2013.
- [149] W Liao, J Ren, K Wang, S Wang, F Zeng, Y Wang, and F Sun. Alignment-free transcriptomic and metatranscriptomic comparison using sequencing signatures with variable length markov chains. *Scientific reports*, 6:37243, 2016.
- [150] CC Laczny, C Kiefer, V Galata, T Fehlmann, C Backes, and A Keller. Busybee web: metagenomic data analysis by bootstrapped supervised binning and annotation. *Nucleic acids research*, 45(W1):W171–W179, 2017.
- [151] Y Wang, H Hu, and X Li. Mbmc: An effective markov chain approach for binning metagenomic reads from environmental shotgun sequencing projects. *Omics: a journal of integrative biology*, 20(8):470–479, 2016.
- [152] RM Kotamarti, M Hahsler, D Raiford, M McGee, and MaH Dunham. Analyzing taxonomic classification using extensible markov models. *Bioinformatics*, 26(18):2235–2241, 2010.
- [153] H-S Seok, W Hong, and J Kim. Estimating the composition of species in metagenomes by clustering of next-generation read sequences. *Methods*, 69(3):213–219, 2014.
- [154] VI Ulyantsev, SV Kazakov, VB Dubinkina, AV Tyakht, and DG Alexeev. Metafast: fast reference-free graph-based comparison of shotgun metagenomic data. *Bioinformatics*, 32(18):2760–2767, 2016.
- [155] VB Dubinkina, DS Ischenko, VI Ulyantsev, AV Tyakht, and DG Alexeev. Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis. *BMC bioinformatics*, 17(1):38, 2016.
- [156] S Chatterji, I Yamazaki, Z Bai, and JA Eisen. Compostbin: A dna composition-based algorithm for binning environmental shotgun reads. In *Annual International Conference on Research in Computational Molecular Biology*, pages 17–28. Springer, 2008.
- [157] M Comin and M Schimd. Fast comparison of genomic and meta-genomic reads with alignment-free measures based on quality values. *BMC medical genomics*, 9(1):36, 2016.
- [158] G Benoit, P Peterlongo, M Mariadasou, E Drezen, S Schbath, D Lavenier, and C Lemaitre. Multiple comparative metagenomics using multiset k-mer counting. *PeerJ Computer Science*, 2:e94, 2016.
- [159] SV Thankachan, SP Chockalingam, Y Liu, A Apostolico, and S Aluru. Alfred: a practical method for alignment-free distance computation. *Journal of Computational Biology*, 23(6):452–460, 2016.
- [160] Y Wang, X Lei, S Wang, Z Wang, N Song, F Zeng, and T Chen. Effect of k-tuple length on sample-comparison with high-throughput sequencing data. *Biochemical and biophysical research communications*, 469(4):1021–1027, 2016.
- [161] C Rinke, P Schwientek, A Sczyrba, NN Ivanova, IJ Anderson, J-F Cheng, A Darling, S Malfatti, BK Swan, EA Gies, et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499(7459):431, 2013.

- [162] CE Mason and S Tighe. Focus on metagenomics. *Journal of Biomolecular Techniques: JBT*, 28(1):1, 2017.
- [163] S Kumar, KK Krishnani, B Bhushan, and MP Brahmane. Metagenomics: retrospect and prospects in high throughput age. *Biotechnology research international*, 2015, 2015.
- [164] DB Roszak and RR Colwell. Survival strategies of bacteria in the natural environment. *Microbiological reviews*, 51(3):365, 1987.
- [165] EJ Stewart. Growing unculturable bacteria. *Journal of bacteriology*, pages JB-00345, 2012.
- [166] KI Mohr. Diversity of myxobacteria-we only see the tip of the iceberg. *Microorganisms*, 6(3), 2018.
- [167] M Hamady, C Fraser-Liggett, R Knight, et al. The human microbiome project: Exploring the microbial part of ourselves in a changing world. *Nature*, 449(7164):804–10, 2007.
- [168] W-L Wang, S-Y Xu, Z-G Ren, L Tao, J-W Jiang, and S-S Zheng. Application of metagenomics in the human gut microbiome. *World journal of gastroenterology: WJG*, 21(3):803, 2015.
- [169] S Al Khodor, B Reichert, and IF Shatat. The microbiome and blood pressure: can microbes regulate our blood pressure? *Frontiers in pediatrics*, 5:138, 2017.
- [170] R Kolde, EA Franzosa, G Rahnavard, AB Hall, H Vlamakis, C Stevens, MJ Daly, RJ Xavier, and C Huttenhower. Host genetic variation and its microbiome interactions within the human microbiome project. *Genome medicine*, 10(1):6, 2018.
- [171] M Hattori and T D Taylor. The human intestinal microbiome: a new frontier of human biology. *DNA research*, 16(1):1–12, 2009.
- [172] Atsushi Kouzuma, Shunichi Ishii, and Kazuya Watanabe. Metagenomic insights into the ecology and physiology of microbes in bioelectrochemical systems. *Bioresource technology*, 2018.
- [173] FH Coutinho, GB Gregoracci, JM Walter, CC Thompson, and FL Thompson. Metagenomics sheds light on the ecology of marine microbes and their viruses. *Trends in Microbiology*, 2018.
- [174] RJ Boissy, DJ Romberger, WA Roughead, L Weissenburger-Moser, JA Poole, and TD LeVan. Shotgun pyrosequencing metagenomic analyses of dusts from swine confinement and grain facilities. *PloS one*, 9(4):e95578, 2014.
- [175] B Carbonetto, N Rascovan, R Alvarez, A Mentaberry, and MP Vazquez. Structure, composition and metagenomic profile of soil microbiomes associated to agricultural land use and tillage systems in argentine pampas. *PloS one*, 9(6):e99949, 2014.
- [176] JQ Su, B Wei, CY Xu, M Qiao, and YG Zhu. Functional metagenomic characterization of antibiotic resistance genes in agricultural soils from China. *Environment international*, 65:9–15, 2014.
- [177] SJ Finley, ME Benbow, and GT Javan. Microbial communities associated with human decomposition and their potential use as postmortem clocks. *International journal of legal medicine*, 129(3):623–632, 2015.
- [178] A Fornaciari. Environmental microbial forensics and archaeology of past pandemics. *Microbiology spectrum*, 5(1), 2017.
- [179] Q Peng, X Wang, M Shang, J Huang, G Guan, Y Li, and B Shi. Isolation of a novel alkaline-stable lipase from a

- metagenomic library and its specific application for milkfat flavor production. *Microbial cell factories*, 13(1):1, 2014.
- [180] M Drancourt, C Bollet, A Carlzio, R Martelin, J-P Gayral, and D Raoult. 16S ribosomal DNA sequence analysis of a large collection of environmental and clinical unidentifiable bacterial isolates. *Journal of clinical microbiology*, 38(10):3623–3630, 2000.
- [181] G Muyzer, EC De Waal, and AG Uitterlinden. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Applied and environmental microbiology*, 59(3):695–700, 1993.
- [182] M Balvociute and DH Huson. SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? *BMC genomics*, 18(2):114, 2017.
- [183] JM Janda and SL Abbott. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*, 45(9):2761–2764, 2007.
- [184] J-H Ahn, B-Y Kim, J Song, and H-Y Weon. Effects of PCR cycle number and DNA polymerase type on the 16S rRNA gene pyrosequencing analysis of bacterial communities. *Journal of Microbiology*, 50(6):1071–1074, 2012.
- [185] JP Brooks, DJ Edwards, MD Harwich, MC Rivera, JM Fettweis, MG Serrano, RA Reris, NU Sheth, B Huang, P Girerd, et al. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC microbiology*, 15(1):66, 2015.
- [186] AJ Pinto and L Raskin. PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PloS one*, 7(8):e43093, 2012.
- [187] SW Kembel, M Wu, JA Eisen, and JL Green. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS computational biology*, 8(10):e1002743, 2012.
- [188] EN Hanssen, KH Liland, P Gill, and L Snipen. Optimizing body fluid recognition from microbial taxonomic profiles. *Forensic Science International: Genetics*, 37:13–20, 2018.
- [189] GJ Olsen and CR Woese. Ribosomal rna: a key to phylogeny. *The FASEB journal*, 7(1):113–123, 1993.
- [190] P Yilmaz, R Kottmann, E Pruesse, C Quast, and FO Glockner. Analysis of 23s rna genes in metagenomes—a case study from the global ocean sampling expedition. *Systematic and applied microbiology*, 34(6):462–469, 2011.
- [191] L Yang, Z Tan, D Wang, L Xue, M-X Guan, T Huang, and R Li. Species identification through mitochondrial rna genetic analysis. *Scientific reports*, 4:4089, 2014.
- [192] AE Budding, M Hoogewerf, CMJE Vandenbroucke-Grauls, and PHM Savelkoul. Automated broad-range molecular detection of bacteria in clinical samples. *Journal of clinical microbiology*, 54(4):934–943, 2016.
- [193] FCA Quaak, T van Duijn, J Hoogenboom, AD Kloosterman, and I Kuiper. Human-associated microbial populations as evidence in forensic casework. *Forensic Science International: Genetics*, 36:176–185, 2018.
- [194] AE Woerner, NMM Novroski, FR Wendt, A Ambers, R Wiley, SE Schmedes, and B Budowle. Forensic human identification with targeted microbiome markers using nearest neighbor classification. *Forensic Science International: Genetics*, 38:130–139, 2019.

- [195] J Jovel, J Patterson, W Wang, N Hotte, S O'Keefe, T Mitchel, T Perry, D Kao, AL Mason, KL Madsen, et al. Characterization of the gut microbiome using 16S or shotgun metagenomics. *Frontiers in microbiology*, 7:459, 2016.
- [196] N Shah, H Tang, TG Doak, and Y Ye. Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics. In *Biocomputing 2011*, pages 165–176. World Scientific, 2011.
- [197] B Steven, LV Gallegos-Graves, SR Starkenburg, PS Chain, and CR Kuske. Targeted and shotgun metagenomic approaches provide different descriptions of dryland soil microbial communities in a manipulated field study. *Environmental microbiology reports*, 4(2):248–256, 2012.
- [198] A Almeida, AL Mitchell, A Tarkowska, and RD Finn. Benchmarking taxonomic assignments based on 16s rRNA gene profiling of the microbiota from commonly sampled environments. *BMC Genomics*, 17(1), 2016.
- [199] T Sijen. Molecular approaches for forensic cell type identification: on mRNA, miRNA, DNA methylation and microbial markers. *Forensic Science International: Genetics*, 18:21–32, 2015.
- [200] TH Clarke, A Gomez, H Singh, KE Nelson, and LM Brinkac. Integrating the microbiome as a resource in the forensics toolkit. *Forensic Science International: Genetics*, 30:141–147, 2017.
- [201] R D'Amore, U Z Ijaz, M Schirmer, JG Kenny, R Gregory, AC Darby, M Shakya, M Podar, C Quince, and N Hall. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC genomics*, 17(1):55, 2016.
- [202] A Sczyrba, P Hofmann, P Belmann, D Koslicki, S Janssen, J Droge, I Gregor, S Majda, J Fiedler, E Dahms, and et al. Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nature methods*, 14(11):1063, 2017.
- [203] MA Peabody, T Van Rossum, R Lo, and FSL Brinkman. Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. *BMC bioinformatics*, 16(1):362, 2015.
- [204] K Mavromatis, N Ivanova, K Barry, H Shapiro, E Goltsman, AC McHardy, I Rigoutsos, A Salamov, F Korzeniewski, M Land, and et al. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature methods*, 4(6):495, 2007.
- [205] ABR McIntyre, R Ounit, E Afshinnekoo, RJ Prill, E Henaff, N Alexander, SS Minot, D Danko, J Foox, S Ahsanuddin, and et al. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome biology*, 18(1):182, 2017.
- [206] AM Walsh, F Crispie, O O'Sullivan, L Finnegan, MJ Claesson, and PD Cotter. Species classifier choice is a key consideration when analysing low-complexity food microbiome data. *Microbiome*, 6(1):50, 2018.
- [207] VC Piro, M Matschkowski, and BY Renard. Metameta: integrating metagenome analysis tools to improve taxonomic profiling. *Microbiome*, 5(1):101, 2017.
- [208] A Escobar-Zepeda, EE Godoy-Lozano, L Raggi, L Segovia, E Merino, RM Gutierrez-Rios, K Juarez, AF Licea-Navarro, L Pardo-Lopez, and A Sanchez-Flores. Analysis of sequencing strategies and tools for

- taxonomic annotation: Defining standards for progressive metagenomics. *Scientific reports*, 8(1):12034, 2018.
- [209] E Plummer, J Twin, D M Bulach, SM Garland, and SN Tabrizi. A comparison of three bioinformatics pipelines for the analysis of preterm gut microbiota using 16S rRNA gene sequencing data. *Journal of Proteomics & Bioinformatics*, 8(12):283, 2015.
- [210] H Hasman, D Saputra, T Sicheritz-Ponten, O Lund, CA Svendsen, N Frimodt-Moller, and FM Aarestrup. Rapid whole genome sequencing for the detection and characterization of microorganisms directly from clinical samples. *Journal of clinical microbiology*, pages JCM-02452, 2013.
- [211] A Klindworth, E Pruesse, T Schweer, J Peplies, C Quast, M Horn, and FO Glockner. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic acids research*, 41(1):e1–e1, 2013.
- [212] A Bankevich, S Nurk, D Antipov, AA Gurevich, M Dvorkin, AS Kulikov, VM Lesin, SI Nikolenko, S Pham, AD 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.
- [213] SY Anvar, L Khachatryan, M Vermaat, M van Galen, I Pulyakhina, Y Ariyurek, K Kraaijeveld, JT den Dunnen, P de Knijff, P Ac't Hoen, et al. Determining the quality and complexity of next-generation sequencing data without a reference genome. *Genome biology*, 15(12):555, 2014.
- [214] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.
- [215] T Magoc and SL Salzberg. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21):2957–2963, 2011.
- [216] NA O'Leary, MW Wright, JR Brister, S Ciufu, D Haddad, R McVeigh, B Rajput, B Robbertse, B Smith-White, D Ako-Adjei, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–D745, 2015.
- [217] FP Breitwieser and SL Salzberg. Pavian: Interactive analysis of metagenomics data for microbiomics and pathogen identification. *BioRxiv*, page 084715, 2016.
- [218] A Wilke, T Harrison, J Wilkening, D Field, EM Glass, N Kyrpides, K Mavrommatis, and F Meyer. The m5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. *BMC bioinformatics*, 13(1):141, 2012.
- [219] HB Mann and DR Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.
- [220] Y Benjamini, Yand Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.
- [221] C Van Rijsbergen. Information retrieval: theory and practice. In *Proceedings of the Joint IBM/University of Newcastle*



- upon Tyne Seminar on Data Base Systems, pages 1–14, 1979.
- [222] N Nagarajan, C Cook, MP Di Bonaventura, H Ge, A Richards, KA Bishop-Lilly, R DeSalle, TD Read, and M Pop. Finishing genomes with limited resources: lessons from an ensemble of microbial genomes. *BMC genomics*, 11(1):242, 2010.
- [223] DJ Edwards and KE Holt. Beginner’s guide to comparative bacterial genome analysis using next-generation sequence data. *Microbial informatics and experimentation*, 3(1):2, 2013.
- [224] EA Grice and JA Segre. The skin microbiome. *Nature Reviews Microbiology*, 9(4):244, 2011.
- [225] S Bikel, A Valdez-Lara, F Cornejo-Granados, K Rico, S Canizales-Quinteros, X Soberon, L Del Pozo-Yauner, and A Ochoa-Leyva. Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: towards a systems-level understanding of human microbiome. *Computational and structural biotechnology journal*, 13:390–401, 2015.
- [226] MJ Gosalbes, JJ Abellan, A Durban, AE Perez-Cobas, A Latorre, and A Moya. Metagenomics of human microbiome: beyond 16S rDNA. *Clinical Microbiology and Infection*, 18:47–49, 2012.
- [227] S Maccaferri, E Biagi, and P Brigidi. Metagenomics: key to human gut microbiota. *Digestive diseases*, 29(6):525–530, 2011.
- [228] R Martin, S Miquel, P Langella, and LG Bermudez-Humaran. The role of metagenomics in understanding the human microbiome in health and disease. *Virulence*, 5(3):413–423, 2014.
- [229] SL Edmonds-Wilson, NI Nurinova, CA Zapka, N Fierer, and M Wilson. Review of human hand microbiome research. *Journal of dermatological science*, 80(1):3–12, 2015.
- [230] HE Blum. The human microbiome. *Advances in medical sciences*, 62(2):414–420, 2017.
- [231] E Holmes, JV Li, JR Marchesi, and JK Nicholson. Gut microbiota composition and activity in relation to host metabolic phenotype and disease risk. *Cell metabolism*, 16(5):559–564, 2012.
- [232] AP Bhatt, MR Redinbo, and SJ Bultman. The role of the microbiome in cancer development and therapy. *CA: a cancer journal for clinicians*, 67(4):326–344, 2017.
- [233] I Cho and MJ Blaser. The human microbiome: at the interface of health and disease. *Nature Reviews Genetics*, 13(4):260, 2012.
- [234] JL Sonnenburg and F Backhed. Diet-microbiota interactions as moderators of human metabolism. *Nature*, 535(7610):56, 2016.
- [235] BH Mullish, JR Marchesi, MR Thursz, and HRT Williams. Microbiome manipulation with faecal microbiome transplantation as a therapeutic strategy in clostridium difficile infection. *QJM: An International Journal of Medicine*, 108(5):355–359, 2014.
- [236] RD Moloney, L Desbonnet, G Clarke, TG Dinan, and JF Cryan. The microbiome: stress, health and disease. *Mammalian Genome*, 25(1-2):49–74, 2014.
- [237] AV Contreras, B Cocom-Chan, G Hernandez-Montes, T Portillo-Bobadilla, and O Resendis-Antonio. Host-microbiome interaction and cancer: Potential application in precision medicine. *Frontiers in physiology*, 7:606, 2016.

- [238] C He, Y Shan, and W Song. Targeting gut microbiota as a possible therapy for diabetes. *Nutrition Research*, 35(5):361–367, 2015.
- [239] CJ Marx. Can you sequence ecology? metagenomics of adaptive diversification. *PLoS biology*, 11(2):e1001487, 2013.
- [240] S Hiraoka, C-C Yang, and W Iwasaki. Metagenomics and bioinformatics in microbial ecology: current status and beyond. *Microbes and environments*, 31(3):204–212, 2016.
- [241] R Tiwari, L Nain, N E Labrou, and P Shukla. Bioprospecting of functional cellulases from metagenome for second generation biofuel production: a review. *Critical reviews in microbiology*, 44(2):244–257, 2018.
- [242] MOA Sommer, GM Church, and G Dantas. A functional metagenomic approach for expanding the synthetic biology toolbox for biomass conversion. *Molecular systems biology*, 6(1):360, 2010.
- [243] V Kunin, A Copeland, A Lapidus, K Mavromatis, and P Hugenholtz. A bioinformatician’s guide to metagenomics. *Microbiology and molecular biology reviews*, 72(4):557–578, 2008.
- [244] SS Mande, MH Mohammed, and TS Ghosh. Classification of metagenomic sequences: methods and challenges. *Briefings in bioinformatics*, 13(6):669–681, 2012.
- [245] Leandro N Lemos, Roberta R Fulthorpe, Eric W Triplett, and Luiz FW Roesch. Rethinking microbial diversity analysis in the high throughput sequencing era. *Journal of microbiological methods*, 86(1):42–51, 2011.
- [246] P Janssen, L Goldovsky, V Kunin, N Darzentas, and CA Ouzounis. Genome coverage, literally speaking: The challenge of annotating 200 genomes with 4 million publications. *EMBO reports*, 6(5):397–399, 2005.
- [247] KB Akondi and VV Lakshmi. Emerging trends in genomic approaches for microbial bioprospecting. *Omics: a journal of integrative biology*, 17(2):61–70, 2013.
- [248] JC Hunter-Cevera. The value of microbial diversity. *Current Opinion in Microbiology*, 1(3):278–285, 1998.
- [249] NR Pace. Mapping the tree of life: progress and prospects. *Microbiology and molecular biology reviews*, 73(4):565–576, 2009.
- [250] J-D Grattepanche, LF Santoferrara, GB McManus, and LA Katz. Diversity of diversity: conceptual and methodological differences in biodiversity estimates of eukaryotic microbes as compared to bacteria. *Trends in microbiology*, 22(8):432–437, 2014.
- [251] L Zinger, A Gobet, and T Pommier. Two decades of describing the unseen majority of aquatic microbial diversity. *Molecular Ecology*, 21(8):1878–1896, 2012.
- [252] B Szalkai, I Scheer, K Nagy, BG Vertessy, and V Grolmusz. The metagenomic telescope. *PloS one*, 9(7):e101605, 2014.
- [253] GL Rosen, R Polikar, DA Caseiro, SD Essinger, and BA Sokhansanj. Discovering the unknown: improving detection of novel species and genera from short reads. *BioMed Research International*, 2011, 2011.
- [254] Y Ono, K Asai, and M Hamada. Pbsim: Pacbio reads simulator—toward accurate genome assembly. *Bioinformatics*, 29(1):119–121, 2012.
- [255] E Van Eijk, SY Anvar, HP Browne, WY Leung, J Frank, AM Schmitz, AP Roberts, and WK Smits. Complete

- genome sequence of the *Clostridium difficile* laboratory strain 630 $\delta$  erm reveals differences from strain 630, including translocation of the mobile element ctn 5. *BMC genomics*, 16(1):31, 2015.
- [256] MF Haroon, S Hu, Y Shi, M Imelfort, J Keller, P Hugenholtz, Z Yuan, and GW Tyson. Anaerobic oxidation of methane coupled to nitrate reduction in a novel archaeal lineage. *Nature*, 500(7464):567, 2013.
- [257] MJ Chaisson and G Tesler. Mapping single molecule sequencing reads using basic local alignment with successive refinement (blasr): application and theory. *BMC bioinformatics*, 13(1):238, 2012.
- [258] C-S Chin, P Peluso, FJ Sedlazeck, M Nattestad, GT Concepcion, A Clum, C Dunn, R O'Malley, R Figueroa-Balderas, A Morales-Cruz, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature methods*, 13(12):1050, 2016.
- [259] L van der Maaten and G Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.
- [260] M Ester, H-P Kriegel, J Sander, X Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.
- [261] J Frank, S Lucker, RHAM Vossen, MSM Jetten, RJ Hall, HJM Op den Camp, and SY Anvar. Resolving the complete genome of *kuenenia stuttgartiensis* from a membrane bioreactor enrichment using single-molecule real-time sequencing. *Scientific reports*, 8(1):4580, 2018.
- [262] DB Goldstein, A Allen, J Keebler, EH Margulies, S Petrou, S Petrovski, and S Sunyaev. Sequencing studies in human genetics: design and interpretation. *Nature Reviews Genetics*, 14(7):460, 2013.
- [263] A Nekrutenko and J Taylor. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics*, 13(9):667, 2012.
- [264] M Costello, TJ Pugh, TJ Fennell, C Stewart, L Lichtenstein, JC Meldrim, JL Foster, DC Friedrich, D Perrin, D Dionne, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic acids research*, 41(6):e67–e67, 2013.
- [265] C Alkan, BP Coe, and EE Eichler. Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12(5):363, 2011.
- [266] JM Kidd, N Sampas, F Antonacci, T Graves, R Fulton, HS Hayden, C Alkan, M Malig, M Ventura, G Gianuzzi, et al. Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nature methods*, 7(5):365, 2010.
- [267] H Li and N Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in bioinformatics*, 11(5):473–483, 2010.
- [268] J Kuczynski, CL Lauber, WA Walters, LW Parfrey, JC Clemente, D Gevers, and R Knight. Experimental and analytical tools for studying the human microbiome. *Nature Reviews Genetics*, 13(1):47, 2012.
- [269] S Subramanian and S Kumar. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. *Genome research*, 13(5):838–844, 2003.

- [270] J Sved and A Bird. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proceedings of the National Academy of Sciences*, 87(12):4692–4696, 1990.
- [271] Ms Csuros, L Noe, and G Kucherov. Reconsidering the significance of genomic word frequencies. *Trends in Genetics*, 23(11):543–546, 2007.
- [272] C Acquisti, G Poste, D Curtiss, and S Kumar. Nullomers: really a matter of natural selection? *PloS one*, 2(10):e1022, 2007.
- [273] J Josse, AD Kaiser, and A Kornberg. Enzymatic synthesis of deoxyribonucleic acid VIII. frequencies of nearest neighbor base sequences in deoxyribonucleic acid. *Journal of Biological Chemistry*, 236(3):864–875, 1961.
- [274] B Chor, D Horn, N Goldman, Y Levy, and T Massingham. Genomic DNA k-mer spectra: models and modalities. *Genome biology*, 10(10):R108, 2009.
- [275] R Hariharan, R Simon, MR Pillai, and TD Taylor. Comparative analysis of DNA word abundances in four yeast genomes using a novel statistical background model. *PloS one*, 8(3):e58038, 2013.
- [276] B Jiang, JS Liu, and ML Bulyk. Bayesian hierarchical model of protein-binding microarray k-mer data reduces noise and identifies transcription factor subclasses and preferred k-mers. *Bioinformatics*, 29(11):1390–1398, 2013.
- [277] Y Liu, J Schroder, and B Schmidt. Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data. *Bioinformatics*, 29(3):308–315, 2012.
- [278] H Chae, J Park, S-W Lee, KP Nephew, and S Kim. Comparative analysis using k-mer and k-flank patterns provides evidence for CpG island sequence evolution in mammalian genomes. *Nucleic acids research*, 41(9):4783–4791, 2013.
- [279] DR Kelley, MC Schatz, and SL Salzberg. Quake: quality-aware detection and correction of sequencing errors. *Genome biology*, 11(11):R116, 2010.
- [280] R Chikhi and P Medvedev. Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, 30(1):31–37, 2013.
- [281] A Brazma, I Jonassen, J Vilo, and E Ukkonen. Predicting gene regulatory elements in silico on a genomic scale. *Genome research*, 8(11):1202–1215, 1998.
- [282] GE Sims, S-R Jun, GA Wu, and S-H Kim. Alignment-free genome comparison with feature frequency profiles (ffp) and optimal resolutions. *Proceedings of the National Academy of Sciences*, pages pnas-0813249106, 2009.
- [283] JT Simpson. Exploring genome characteristics and sequence quality without a reference. *Bioinformatics*, 30(9):1228–1235, 2014.
- [284] T Lappalainen, M Sammeth, MR Friedlander, P AC’t Hoen, J Monlong, MA Rivas, M Gonzalez-Porta, N Kurbatova, T Griebel, PG Ferreira, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506, 2013.
- [285] P AC’t Hoen, MR Friedlander, J Almlof, M Sammeth, I Pulyakhina, SY Anvar, JFJ Laros, HPJ Buermans, O Karlberg, M Brannvall, et al. Reproducibility of high-throughput mrna and small rna sequencing across laboratories. *Nature biotechnology*, 31(11):1015, 2013.
- [286] WA Kusters and JFJ Laros. Metrics for mining multisets. In *Research and Development in Intelligent systems XXIV*, pages 293–303. Springer, 2008.

- [287] PJ Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.
- [288] J Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.
- [289] G Lunter and M Goodson. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome research*, 21(6):936–939, 2011.
- [290] AR Quinlan and IM Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.
- [291] HH Li, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, and R Durbin. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.
- [292] JG Caporaso, CL Lauber, EK Costello, D Berg-Lyons, A Gonzalez, J Stombaugh, D Knights, P Gajer, J Ravel, N Fierer, et al. Moving pictures of the human microbiome. *Genome biology*, 12(5):R50, 2011.
- [293] KJ Stacey, GR Young, F Clark, DP Sester, TL Roberts, S Naik, MJ Sweet, and DA Hume. The molecular basis for the lack of immunostimulatory activity of vertebrate DNA. *The Journal of Immunology*, 170(7):3614–3620, 2003.
- [294] P Kaufmann, A Pfefferkorn, M Teuber, and L Meile. Identification and quantification of bifidobacterium species isolated from food with genus-specific 16S rRNA-targeted probes by colony hybridization and PCR. *Applied and Environmental Microbiology*, 63(4):1268–1273, 1997.
- [295] KJV Nordstrom, MC Albani, GVe James, C Gutjahr, B Hartwig, F Turck, U Paszkowski, G Coupland, and K Schneeberger. Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using k-mers. *Nature biotechnology*, 31(4):325, 2013.
- [296] SN Gardner and BG Hall. When whole-genome alignments just won't work: kSNP v2 software for alignment-free snp discovery and phylogenetics of hundreds of microbial genomes. *PloS one*, 8(12):e81760, 2013.
- [297] G Marcais and C Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770, 2011.
- [298] M Crusoe, G Edverson, J Fish, A Howe, E McDonald, J Nahum, K Nanlohy, H Ortiz-Zuazaga, J Pell, J Simpson, et al. The khmer software package: enabling efficient sequence analysis. URL <http://dx.doi.org/10.6084/m9.figshare.979190>, 2014.
- [299] DS DeLuca, JZ Levin, A Sivachenko, T Fennell, M-D Nazaire, C Williams, M Reich, W Winckler, and G Getz. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*, 28(11):1530–1532, 2012.
- [300] N Segata, D Boernigen, TL Tickle, XC Morgan, WS Garrett, and C Huttenhower. Computational metaomics for microbial community studies. *Molecular systems biology*, 9(1):666, 2013.
- [301] KT Konstantinidis, A Ramette, and JM Tiedje. The bacterial species definition in the genomic era. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 361(1475):1929–1940, 2006.

- [302] M Schlöter, M Lebuhn, T Heulin, and A Hartmann. Ecology and evolution of bacterial microdiversity. *FEMS microbiology reviews*, 24(5):647–660, 2000.
- [303] DL Hartl and DE Dykhuizen. The population genetics of *Escherichia coli*. *Annual review of genetics*, 18(1):31–68, 1984.
- [304] PA Cotter and VJ DiRita. Bacterial virulence gene regulation: an evolutionary perspective. *Annual Reviews in Microbiology*, 54(1):519–565, 2000.
- [305] RW Jackson, E Athanassopoulos, G Tsiamis, JW Mansfield, A Sesma, DL Arnold, MJ Gibbon, J Murillo, JD Taylor, and A Vivian. Identification of a pathogenicity island, which contains genes for virulence and avirulence, on a large native plasmid in the bean pathogen *Pseudomonas syringae* pathovar *phaseolicola*. *Proceedings of the National Academy of Sciences*, 96(19):10875–10880, 1999.
- [306] Antimicrobial Resistance WHO. Global report on surveillance. *Antimicrobial Resistance, Global Report on Surveillance*, 2014.
- [307] PM Bennett. Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria. *British journal of pharmacology*, 153(S1):S347–S357, 2008.
- [308] T Foster. *Staphylococcus*. 1996.
- [309] S Jarraud, C Mougél, J Thioulouse, G Lina, H Meugnier, F Forey, X Nesme, J Etienne, and F Vandenesch. Relationships between *Staphylococcus aureus* genetic background, virulence factors, agr groups (alleles), and human disease. *Infection and immunity*, 70(2):631–641, 2002.
- [310] R Urwin and MCJ Maiden. Multilocus sequence typing: a tool for global epidemiology. *Trends in microbiology*, 11(10):479–487, 2003.
- [311] MCJ Maiden, JA Bygraves, E Feil, G Morelli, J E Russell, R Urwin, Q Zhang, J Zhou, K Zurth, DA Cautant, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences*, 95(6):3140–3145, 1998.
- [312] M Dreyer, L Aguilar-Bultet, S Rupp, C Guldemann, R Stephan, Al Schöck, A Otter, Ge Schupbach, S Brisse, M Lecuit, et al. *Listeria monocytogenes* sequence type 1 is predominant in ruminant rhombencephalitis. *Scientific reports*, 6:36419, 2016.
- [313] MD Ismail, I Ali, S Hatt, EA Salzman, AW Cronenwett, CF Marrs, AH Rickard, and B Foxman. Association of *Escherichia coli* ST131 lineage with risk of urinary tract infection recurrence among young women. *Journal of global antimicrobial resistance*, 13:81–84, 2018.
- [314] S Jena, S Panda, KC Nayak, and DV Singh. Identification of major sequence types among multidrug-resistant *Staphylococcus epidermidis* strains isolated from infected eyes and healthy conjunctiva. *Frontiers in Microbiology*, 8:1430, 2017.
- [315] C-R Usein, AS Ciontea, CM Militaru, M Condei, S Dinu, M Oprea, D Cristea, V Michelacci, G Scavia, LC Zota, et al. Molecular characterisation of human shiga toxin-producing *Escherichia coli* OMLST26 strains: results of an outbreak investigation, Romania, February to August 2016. *Eurosurveillance*, 22(47), 2017.
- [316] MH Antwerpen, K Prior, A Mellmann, S Höppner, WD Splettstoesser, and D Harmsen. Rapid high resolution genotyping of *Francisella tularensis* by

- whole genome sequence comparison of annotated genes ("MLST+"). *PLoS One*, 10(4):e0123298, 2015.
- [317] VI Siarkou, F Vorimore, N Vicari, S Magnino, A Rodolakis, Y Pannekoek, K Sachse, D Longbottom, and K Laroucau. Diversification and distribution of ruminant *Chlamydia abortus* clones assessed by MLST and mlva. *PLoS One*, 10(5):e0126433, 2015.
- [318] B Heym, M Le Moal, L Armand-Lefevre, and M-H Nicolas-Chanoine. Multilocus sequence typing (MLST) shows that the 'Iberian' clone of methicillin-resistant *Staphylococcus aureus* has spread to france and acquired reduced susceptibility to teicoplanin. *Journal of Antimicrobial Chemotherapy*, 50(3):323–329, 2002.
- [319] Y Yamaoka. *Helicobacter pylori* typing as a tool for tracking human migration. *Clinical Microbiology and Infection*, 15(9):829–834, 2009.
- [320] Z Qi, Y Cui, Q Zhang, and R Yang. Taxonomy of *Yersinia pestis*. In *Yersinia pestis: Retrospective and Perspective*, pages 35–78. Springer, 2016.
- [321] W Wade. Unculturable bacteria—the uncharacterized organisms that cause oral infections. *Journal of the Royal Society of Medicine*, 95(2):81–83, 2002.
- [322] S Bhattacharya, N Vijayalakshmi, and SC Parija. Uncultivable bacteria: Implications and recent trends towards identification. *Indian journal of medical microbiology*, 20(4):174, 2002.
- [323] M Pinto, V Borges, M Antelo, M Pinheiro, A Nunes, J Azevedo, MJ Borrego, J Mendonca, D Carpinteiro, L Vieira, and et al. Genome-scale analysis of the non-cultivable treponema pallidum reveals extensive within-patient genetic variation. *Nature microbiology*, 2(1):16190, 2017.
- [324] D Smajs, M Strouhal, and S Knauf. Genetics of human and animal uncultivable treponemal pathogens. *Infection, Genetics and Evolution*, 61:92–107, 2018.
- [325] KW Larssen, A Nor, and K Bergh. Rapid discrimination of *Staphylococcus epidermidis* genotypes in a routine clinical microbiological laboratory using single nucleotide polymorphisms in housekeeping genes. *Journal of medical microbiology*, 67(2):169–182, 2018.
- [326] SA Nachappa, SM Neelambike, C Amruthavalli, and NB Ramachandra. Detection of first-line drug resistance mutations and drug–protein interaction dynamics from tuberculosis patients in south india. *Microbial Drug Resistance*, 24(4):377–385, 2018.
- [327] SS Khoramrooz, SA Dolatabad, FM Dolatabad, M Marashifard, M Mirzaii, H Dabiri, A Haddadi, SM Rabani, HRG Shirazi, and D Darban-Sarokhalil. Detection of tetracycline resistance genes, aminoglycoside modifying enzymes, and coagulase gene typing of clinical isolates of *Staphylococcus aureus* in the southwest of iran. *Iranian journal of basic medical sciences*, 20(8):912, 2017.
- [328] J Sarvari, A Bazargani, MR Kandekar-Ghahraman, A Nazari-Alam, M Motamedifar, et al. Molecular typing of methicillin-resistant and methicillin-susceptible *Staphylococcus aureus* isolates from shiraz teaching hospitals by PCR-rflp of coagulase gene. *Iranian journal of microbiology*, 6(4):246, 2014.
- [329] RA Viau, LM Kiedrowski, BN Kreiswirth, M Adams, F Perez, D Marchaim, DM Guerrero, KS Kaye, LK Logan, MV Villegas, et al. A comparison of molecular typing methods applied to *Enterobacter cloacae* complex:

- hsp60 sequencing, rep-PCR, and MLST. *Pathogens & immunity*, 2(1):23, 2017.
- [330] KA Jolley and MCJ Maiden. BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC bioinformatics*, 11(1):595, 2010.
- [331] M Inouye, H Dashnow, L-A Raven, MB Schultz, BJ Pope, T Tomita, J Zobel, and KE Holt. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome medicine*, 6(11):90, 2014.
- [332] R Tewolde, T Dallman, U Schaefer, CL Sheppard, P Ashton, B Pichon, M Ellington, C Swift, J Green, and A Underwood. MOST: a modified MLST typing tool based on short read sequencing. *PeerJ*, 4:e2308, 2016.
- [333] MV Larsen, S Cosentino, S Rasmussen, C Friis, H Hasman, R Marvig, L Jelsbak, TS Pontén, DW Ussery, FM Aarestrup, et al. Multilocus sequence typing of total genome sequenced bacteria. *Journal of clinical microbiology*, pages JCM-06094, 2012.
- [334] N-F Alikhan, Z Zhou, MJ Sergeant, and M Achtman. A genomic overview of the population structure of salmonella. *PLoS genetics*, 14(4):e1007261, 2018.
- [335] A Gupta, IK Jordan, and L Rishishwar. stringMLST: a fast k-mer based tool for multilocus sequence typing. *Bioinformatics*, 33(1):119–121, 2016.
- [336] AJ Page, N-F Alikhan, HA Carleton, T Seemann, JA Keane, and LS Katz. Comparison of classical multi-locus sequence typing software for next-generation sequencing data. *Microbial genomics*, 3(8), 2017.
- [337] H 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.
- [338] P Danecek, A Auton, G Abecasis, CA Albers, E Banks, MA DePristo, RE Handsaker, G Lunter, GT Marth, ST Sherry, et al. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, 2011.
- [339] VI Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8):707–710, 1966.
- [340] T Wirth, D Falush, R Lan, F Colles, P Mensa, LH Wieler, H Karch, PR Reeves, MCJ Maiden, H Ochman, and et al. Sex and virulence in escherichia coli: an evolutionary perspective. *Molecular microbiology*, 60(5):1136–1151, 2006.
- [341] A Koehler, H Karch, T Beikler, TF Flemmig, S Suerbaum, and H Schmidt. Multilocus sequence analysis of porphyromonas gingivalis indicates frequent recombination. *Microbiology*, 149(9):2407–2415, 2003.
- [342] R Leinonen, H Sugawara, M Shumway, and International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic acids research*, 39(suppl\_1):D19–D21, 2010.
- [343] M Enersen, I Olsen, AJ van Winkelhoff, and DA Caugant. Multilocus sequence typing of porphyromonas gingivalis strains from different geographic origins. *Journal of clinical microbiology*, 44(1):35–41, 2006.
- [344] F Alhashash, X Wang, K Paszkiewicz, M Diggle, Z Zong, and A McNally. Increase in bacteraemia cases in the east midlands region of the uk due to mdr escherichia coli ST73: high levels of genomic and plasmid diversity in causative isolates. *Journal of Antimicrobial Chemotherapy*, 71(2):339–343, 2015.



- [345] T Cohen, PD van Helden, D Wilson, C Colijn, MM McLaughlin, I Abubakar, and RM Warren. Mixed-strain mycobacterium tuberculosis infections and the implications for tuberculosis treatment and control. *Clinical microbiology reviews*, 25(4):708–719, 2012.
- [346] M Dzunkova, A Moya, X Chen, C Kelly, and G D’Auria. Detection of mixed-strain infections by facs and ultra-low input genome sequencing. *Gut microbes*, pages 1–5, 2018.
- [347] KE Raven, T Gouliouris, J Parkhill, and SJ Peacock. Genome-based analysis of enterococcus faecium bacteremia associated with recurrent and mixed-strain infection. *Journal of clinical microbiology*, 56(3):e01520–17, 2018.
- [348] G Yu, D Fadrosch, JJ Goedert, J Ravel, and AM Goldstein. Nested pcr biases in interpreting microbial community structure in 16s rna gene sequence datasets. *PLoS One*, 10(7):e0132253, 2015.
- [349] M Schirmer, UZ Ijaz, R D’Amore, N Hall, WT Sloan, and C Quince. Insight into biases and sequencing errors for amplicon sequencing with the illumina miseq platform. *Nucleic acids research*, 43(6):e37–e37, 2015.
- [350] D-L Sun, X Jiang, QL Wu, and N-Y Zhou. Intragenomic heterogeneity in 16s rna genes causes overestimation of prokaryotic diversity. *Applied and environmental microbiology*, pages AEM–01282, 2013.
- [351] K Kennedy, MW Hall, MDJ Lynch, G Moreno-Hagelsieb, and JD Neufeld. Evaluating bias of illumina-based bacterial 16s rna gene profiles. *Applied and environmental microbiology*, pages AEM–01451, 2014.
- [352] R Poretsky, LM Rodriguez, C Luo, D Tsementzi, and KT Konstantinidis. Strengths and limitations of 16s rna gene amplicon sequencing in revealing temporal microbial community dynamics. *PLoS One*, 9(4):e93827, 2014.

