rooted in graph theory in some form. The effort of the Data Working Group of the GA4GH is a prominent example for this. We have also discussed how the transition in terms of data structures will affect operations such as read mapping, variant discovery, genotyping and phasing, all of which are at the core of modern genomics research. Last but not least, we have analyzed the issues that arise in visualizing pan-genomes, and we have briefly discussed future issues in uncertain data handling, recently an ever recurring theme in genome data analysis, often arising from the repetitive structure of many genomes.

We have concentrated on computational challenges of pan-genomics in this survey. We are aware that there are also political challenges that have to be addressed that concern data sharing and privacy. Clearly, the usefulness of any pan-genomic representation will increase with the number of genomes it represents, strengthening its expressive and statistical power. Unfortunately, however, only a fraction of the sequenced data is currently publicly available. This is partly owing to the confidential nature of human genetic data, but also, to a large extent, by missing policies and incentives to make genomic data open access or to prevent intentional withholding of data. Funding agencies like the National Institutes of Health in the USA have started to address these issues [158] (see also http://www.nih.gov/news-events/news-releases/nih-issues-finalized-policy-genomic-data-sharing).

Future directions

Overall, we have provided a broad overview of computational pan-genomics issues, which we hope will serve as a reference for future research proposals and projects. However, so far, we have mostly been addressing how to deal with genomes as sequences, that is, from a 'one-dimensional' point of view, and so we have been focusing on storing and analyzing sequences and the mutual relations of particular subsequence patches, like variant alleles and their interlinkage, genes and/or transcriptomes. We have done this because we believe that at this point in genomics history, only the consistent exploration and annotation of exhaustive amounts of sequence information can lay the solid foundation for additional 'pan-genomics oriented' steps.

Yet, even after having resolved the corresponding issues—and we are hopeful that, at this point, our summary has helped to consistently structure these—there is more to follow. New approaches have already appeared on the horizon that will benefit from the cornerstone provided by primarily sequence-driven pan-genomics. For example, it can be expected that one can lift pan-genomes into three dimensions in the mid-term future, thanks to rapidly developing techonolgies that allow to infer their three-dimensional conformation. This will mean that future, three-dimensional pan-genomes will not only represent all sequence variation applying for species or populations, but also encode their spatial organization as well as their mutual relationships in that respect.

Epigenomics topics have not been exhaustively addressed here either, but will need to be addressed as soon as the first 'primary' pan-genomes stand. Technologies that do not only map sequential and three-dimensional arrangement, but also additional biochemical modifications have likewise been on a steep rise recently. Most importantly, we will be in position to link sequential pan-genomes to maps that indicate hypo- and hypermethylated regions relatively soon. Likely, the integration of such basic biochemical modifications will serve as template for further, often more complex elements of biochemical genomic maps.

In summary, the emergence of computational pan-genomics as a field is a major advance in contemporary genomics research. We have entered an era that holds the promise to close large gaps in global maps of genomes and to draw the full picture of their variability. We therefore believe that we can expect to witness amazing, encompassing insights about extent, pace and nature of evolution in the mid-term future.

Key Points

- Many disciplines, from human genetics and oncology to plant breeding, microbiology and virology, commonly face the challenge of analyzing rapidly increasing numbers of genomes.
- Simply scaling up established bioinformatics pipelines will not be sufficient for leveraging the full potential of such rich genomic data sets.
- Novel, qualitatively different computational methods and paradigms are needed and we will witness the rapid extension of computational pan-genomics, a new sub-area of research in computational biology.
- The transition from the representation of reference genomes as strings to representations as graphs is a prominent example for such a computational paradigm shift.

Funding

The Netherlands Organization for Scientific Research (NWO) Vidi (639.072.309 to A.S., 864.14.004 to B.E.D.); CAPES/BRASIL (to B.E.D.); the Academy of Finland (284598 [CoECGR] to V.M. and D.V.); the Russian Scientific Foundation (14–11–00826 to L.D.); Institut de Biologie Computationnelle (ANR-11-BINF-0002 to E.R.); and the French Colib'read project (ANR–12–BS02–0008 to E.R.). NSFC 31671372 (to K. Y.); the Dutch Graduate School for Experimental Plant Sciences (054EPS15 to S.S.); the EMGO Institute for Health and Care Research (EMGO+) to K.O.; the National Human Genome Research Institute (1U54HG007990 [BD2K] to B.P. and A.M.N., 5U41HG007234 [GENCODE] to B.P.); the W. M. Keck Foundation (DT06172015 to B.P. and A.M.N.); the Simons Foundation (SFLIFE# 351901 to B.P. and A.M.N.); the ARCS Foundation (2014–15 ARCS fellowship to A.M.N.); Edward Schulak (Edward Schulak Fellowship in Genomics to A.M.N.)

Acknowledgments

We are deeply grateful to the Lorentz Center for hosting the workshop 'Future Perspectives in Computational Pan-Genomics' (8–12 June 2015), which gave rise to this article. In particular, we like to thank the Lorentz Center staff, who turned organizing and attending the workshop into a great pleasure. The workshop received additional financial support by KNAW, Bina Technologies, ERIBA, PacBio and Genalice. E.E.E. is an investigator of the Howard Hughes Medical Institute.

Competing interests

Ole Schulz-Trieglaff is an employee of Illumina Inc. and receives stocks as part of his compensation. Illumina is a

public company that develops and markets systems for genetic analysis.

The Computational Pan-Genomics Consortium

Tobias Marschall, Manja Marz, Thomas Abeel, Louis Dijkstra, Bas E. Dutilh, Ali Ghaffaari, Paul Kersey, Wigard P. Kloosterman, Veli Mäkinen, Adam M. Novak, Benedict Paten, David Porubsky, Eric Rivals, Can Alkan, Jasmijn A. Baaijens, Paul I. W. De Bakker, Valentina Boeva, Raoul J. P. Bonnal, Francesca Chiaromonte, Rayan Chikhi, Francesca D. Ciccarelli, Robin Cijvat, Erwin Datema, Cornelia M. Van Duijn, Evan E. Eichler, Corinna Ernst, Eleazar Eskin, Erik Garrison, Mohammed El-Kebir, Gunnar W. Klau, Jan O. Korbel, Eric-Wubbo Lameijer, Benjamin Langmead, Marcel Martin, Paul Medvedev, John C. Mu, Pieter Neerincx, Klaasjan Ouwens, Pierre Peterlongo, Nadia Pisanti, Sven Rahmann, Ben Raphael, Knut Reinert, Dick de Ridder, Jeroen de Ridder, Matthias Schlesner, Ole Schulz-Trieglaff, Ashley D. Sanders, Siavash Sheikhizadeh, Carl Shneider, Sandra Smit, Daniel Valenzuela, Jiayin Wang, Lodewyk Wessels, Ying Zhang, Victor Guryev, Fabio Vandin, Kai Ye and Alexander Schönhuth. All consortium members affiliations can be found in the supplementary data online.

Supplementary Data

Supplementary files are available online at http://bib.oxfordjournals.org/.

References

1. Fleischmann RD, Adams MD, White O, *et al*. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 1995;**269**(5223):496–512.
2. Goffeau A, Barrell BG, Bussey H, *et al*. Life with 6000 genes. *Science* 1996;**274**(5287):546–67.
3. Lander ES, Linton LM, Birren B, *et al*. Initial sequencing and analysis of the human genome. *Nature* 2001;**409**(6822):860–921.
4. Venter JC, Adams MD, Myers EW, *et al*. The sequence of the human genome. *Science* 2001;**291**(5507):1304–51.
5. Weigel D, Mott R. The 1001 genomes project for Arabidopsis thaliana. *Genome Biol* 2009;**10**(5):107.
6. The 100 Tomato Genome Sequencing Consortium; Aflitos S, Schijlen E, *et al*. Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. *Plant J* 2014;**80**(1):136–48.
7. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* 2014;**46**(8):818–25.
8. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* 2015;**526**(7571):68–74.
9. Tettelin H, Masignani V, Cieslewicz MJ, *et al*. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial 'pan-genome'. *Proc Natl Acad Sci USA* 2005;**102**(39):13950–5.
10. Morgante M, De Paoli E, Radovic S. Transposable elements and the plant pan-genomes. *Curr Opin Plant Biol* 2007;**10**(2):149–55.
11. Rogers J, Gibbs RA. Comparative primate genomics: emerging patterns of genome content and dynamics. *Nat Rev Genet* 2014;**15**(5):347–59.
12. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* 2012;**486**(7402):207–14.
13. Sigaux F. Cancer genome or the development of molecular portraits of tumors [in French]. *Bull Acad Natl Med* 2000;**184**(7):1441–9. discussion 1448–1449.
14. Vernikos G, Medini D, Riley DR, *et al*. Ten years of pan-genome analyses. *Curr Opin Microbiol* 2015;**23**:148–54.
15. Dilthey A, Cox C, Iqbal Z, *et al*. Improved genome inference in the MHC using a population reference graph. *Nat Genet* 2015;**47**(6):682–8.
16. Heber S, Alekseyev M, Sze SH, *et al*. Splicing graphs and EST assembly problem. *Bioinformatics* 2002;**18(Suppl 1)**:S181–8.
17. Domingo E. Quasispecies theory in virology. *J Virol* 2002;**76**(1):463–5.
18. Kent WJ, Sugnet CW, Furey TS, *et al*. The human genome browser at UCSC. *Genome Res* 2002;**12**(6):996–1006.
19. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. *Brief Bioinform* 2013;**14**(2):178–92.
20. Kafarski P. Rainbow code of biotechnology. *Chemik* 2012;**66**(8):811–6.
21. Hall RJ, Draper JL, Nielsen FGG, *et al*. Beyond research: a primer for considerations on using viral metagenomics in the field and clinic. *Front Microbiol* 2015;**6**:224.
22. Liti G, Carter DM, Moses AM, *et al*. Population genomics of domestic and wild yeasts. *Nature* 2009;**458**(7236):337–41.
23. Dutilh BE, Thompson CC, Vicente ACP, *et al*. Comparative genomics of 274 Vibrio cholerae genomes reveals mobile functions structuring three niche dimensions. *BMC Genomics* 2014;**15**(1):654.
24. Xiao J, Zhang Z, Wu J, *et al*. A brief review of software tools for Pangenomics. *Genomics Proteomics Bioinformatics* 2015;**13**(1):73–6.
25. Nguyen N, Hickey G, Raney BJ, *et al*. Comparative assembly hubs: web-accessible browsers for comparative genomics. *Bioinformatics* 2014;**30**(23):3293–301.
26. Doolittle WF. Phylogenetic classification and the universal tree. *Science* 1999;**284**(5423):2124–8.
27. Crisp A, Boschetti C, Perry M, *et al*. Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. *Genome Biol* 2015;**16**(1):50.
28. Huson DH, Scornavacca C. A survey of combinatorial methods for phylogenetic networks. *Genome Biol Evol* 2011;**3**:23–35.
29. Dutilh BE, Backus L, Edwards RA, *et al*. Explaining microbial phenotypes on a genomic scale: GWAS for microbes. *Brief Funct Genomics* 2013;**12**(4):366–80.
30. Read TD, Massey RC. Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. *Genome Med* 2014;**6**(11):109.
31. Li J, Jia H, Cai X, *et al*. An integrated catalog of reference genes in the human gut microbiome. *Nat Biotechnol* 2014;**32**(8):834–41.
32. Williamson SJ, Rusch DB, Yooseph S, *et al*. The sorcerer II global ocean sampling expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS One* 2008;**3**(1):e1456.
33. Brum JR, Ignacio-Espinoza JC, Roux S, *et al*. Patterns and ecological drivers of ocean viral communities. *Science* 2015;**348**(6237):1261498.
34. Howe AC, Jansson JK, Malfatti SA, *et al*. Tackling soil diversity with the assembly of large, complex metagenomes. *Proc Natl Acad Sci USA* 2014;**111**(13):4904–9.

35. Qin J, Li Y, Cai Z, *et al*. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 2012;**490**(7418):55–60.
36. Loman NJ, Constantinidou C, Christner M, *et al*. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of shiga-toxigenic escherichia coli o104:h4. *JAMA* 2013;**309**(14):1502–10.
37. Dutilh BE, Huynen MA, Strous M. Increasing the coverage of a metapopulation consensus genome by iterative read mapping and assembly. *Bioinformatics* 2009;**25**(21):2878–81.
38. Allen LZ, Ishoey T, Novotny MA, *et al*. Single virus genomics: a new tool for virus discovery. *PLoS One* 2011;**6**(3):e17722.
39. Malboeuf CM, Yang X, Charlebois P, *et al*. Complete viral RNA genome sequencing of ultra-low copy samples by sequence-independent amplification. *Nucleic Acids Res* 2013;**41**(1):e13.
40. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods* 2015;**12**(8):733–35.
41. Wang J, Moore NE, Deng YM, *et al*. MinION nanopore sequencing of an influenza genome. *Virology* 2015;**6**:766.
42. Beerenwinkel N, Günthard HF, Roth V, *et al*. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Virology* 2012;**3**:329.
43. Falconer E, Hills M, Naumann U, *et al*. DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat Methods* 2012;**9**(11):1107–12.
44. Lengauer T, Sing T. Bioinformatics-assisted anti-HIV therapy. *Nat Rev Microbiol* 2006;**4**(10):790–97.
45. Bartha I, Carlson JM, Brumme CJ, *et al*. A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. *eLife* 2013;**2**:e01123.
46. Carlson JM, Brumme CJ, Martin E, *et al*. Correlates of protective cellular immunity revealed by analysis of population-level immune escape pathways in HIV-1. *J Virol* 2012;**86**(24):13202–16.
47. Daugherty MD, Malik HS. Rules of engagement: molecular insights from host-virus arms races. *Annu Rev Genet* 2012;**46**(1):677–700.
48. Edwards RA, Rohwer F. Viral metagenomics. *Nat Rev Microbiol* 2005;**3**(6):504–10.
49. Mokili JL, Rohwer F, Dutilh BE. Metagenomics and future perspectives in virus discovery. *Curr Opin Virol* 2012;**2**(1):63–77.
50. Barabaschi D, Guerra D, Lacrima K, *et al*. Emerging knowledge from genome sequencing of crop species. *Mol Biotechnol* 2012;**50**(3):250–66.
51. Huang X, Kurata N, Wei X, *et al*. A map of rice genome variation reveals the origin of cultivated rice. *Nature* 2012;**490**(7421):497–501.
52. Jiao Y, Zhao H, Ren L, *et al*. Genome-wide genetic changes during modern breeding of maize. *Nat Genet* 2012;**44**(7):812–15.
53. Mace ES, Tai S, Gilding EK, *et al*. Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. *Nat Commun* 2013;**4**:2320.
54. Lek M, Karczewski K, Exome Aggregation Consortium, *et al*. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;**536**(7616):285–91.
55. MacArthur DG, Manolio TA, Dimmock DP, *et al*. Guidelines for investigating causality of sequence variants in human disease. *Nature* 2014;**508**(7497):469–76.
56. The International HapMap Consortium. A haplotype map of the human genome. *Nature* 2005;**437**(7063):1299–320.
57. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010;**11**(7):499–511.
58. van Rheenen W, Shatunov A, Dekker AM, *et al*. Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat Genet* 2016; **48**(9):1043–8.
59. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet* 2011;**12**(5):363–76.
60. Chaisson MJP, Huddleston J, Dennis MY, *et al*. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* 2015;**517**(7536):608–11.
61. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature* 2009;**458**(7239):719–24.
62. Kandoth C, McLellan MD, Vandin F, *et al*. Mutational landscape and significance across 12 major cancer types. *Nature* 2013;**502**(7471):333–9.
63. Lawrence MS, Stojanov P, Polak P, *et al*. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 2013;**499**(7457):214–18.
64. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer* 2012;**12**(5):323–34.
65. McGranahan N, Swanton C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell* 2015;**27**(1):15–26.
66. Dutilh BE, van Noort V, van der Heijden RTJM, *et al*. Assessment of phylogenomic and orthology approaches for phylogenetic inference. *Bioinformatics* 2007;**23**(7):815–24.
67. Snel B, Huynen MA, Dutilh BE. Genome trees and the nature of genome evolution. *Annu Rev Microbiol* 2005;**59**:191–209.
68. Ciccarelli FD, Doerks T, von Mering C, *et al*. Toward automatic reconstruction of a highly resolved tree of life. *Science* 2006;**311**(5765):1283–87.
69. Menconi G, Battaglia G, Grossi R, *et al*. Mobilomics in saccharomyces cerevisiae strains. *BMC Bioinformatics* 2013;**14**:102.
70. Boeckmann B, Marcet-Houben M, Rees JA, *et al*. Quest for orthologs entails quest for tree of life: in search of the gene stream. *Genome Biol Evol* 2015;**7**(7):1988–99.
71. de Been M, Lanza VF, de Toro M, *et al*. Dissemination of cephalosporin resistance genes between *Escherichia coli* strains from farm animals and humans by specific plasmid lineages. *PLoS Genet* 2014;**10**(12):e1004776.
72. Williams TA, Foster PG, Cox CJ, *et al*. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* 2013;**504**(7479):231–6.
73. Moroz LL, Kocot KM, Citarella MR, *et al*. The ctenophore genome and the evolutionary origins of neural systems. *Nature* 2014;**510**(7503):109–14.
74. Zhong B, Sun L, Penny D. The origin of land plants: a phylogenomic perspective. *Evol Bioinform Online* 2015;**11**:137–41.
75. Eppinger M, Pearson T, Koenig SSK, *et al*. Genomic epidemiology of the haitian cholera outbreak: a single introduction followed by rapid, extensive, and continued spread characterized the onset of the epidemic. *mBio* 2014;**5**(6):e01721–14.
76. Holden MTG, Hsu LY, Kurt K, *et al*. A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant staphylococcus aureus pandemic. *Genome Res* 2013;**23**(4):653–64.
77. Greenman CD, Pleasance ED, Newman S, *et al*. Estimation of rearrangement phylogeny for cancer genomes. *Genome Res* 2012;**22**(2):346–61.
78. Cooper CS, Eeles R, Wedge DC, *et al*. Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and

morphologically normal prostate tissue. *Nat Genet* 2015;**47**(4):367–72.
79. Glusman G, Cox HC, Roach JC. Whole-genome haplotyping approaches and genomic medicine. *Genome Med* 2014;**6**(9):73.
80. Allhoff M, Schönhuth A, Martin M, *et al*. Discovering motifs that induce sequencing errors. *BMC Bioinformatics* 2013;**14(Suppl 5)**:S1.
81. Ross MG, Russ C, Costello M, *et al*. Characterizing and measuring bias in sequence data. *Genome Biol* 2013;**14**(5):R51.
82. Snyder MW, Adey A, Kitzman JO, *et al*. Haplotype-resolved genome sequencing: experimental methods and applications. *Nat Rev Genet* 2015;**16**(6):344–58.
83. Schneider GF, Dekker C. DNA sequencing with nanopores. *Nat Biotechnol* 2012;**30**(4):326–8.
84. Laver T, Harrison J, O'Neill PA, *et al*. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol Detect Quantif* 2015;**3**:1–8.
85. Ashton PM, Nair S, Dallman T, *et al*. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat Biotechnol* 2015;**33**(3):296–300.
86. Madoui MA, Engelen S, Cruaud C, *et al*. Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC Genomics* 2015;**16**(1):327.
87. Kuleshov V, Xie D, Chen R, *et al*. Whole-genome haplotyping using long reads and statistical methods. *Nat Biotechnol* 2014;**32**(3):261–6.
88. Zheng GX, Lau BT, Schnall-Levin M, *et al*. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol* 2016; **34**(3):303–11.
89. Burton JN, Adey A, Patwardhan RP, *et al*. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat Biotechnol* 2013;**31**(12):1119–25.
90. Teague B, Waterman MS, Goldstein S, *et al*. High-resolution human genome structure by single-molecule analysis. *Proc Natl Acad Sci USA* 2010;**107**(24):10848–53.
91. Hastie AR, Dong L, Smith A, *et al*. Rapid genome mapping in nanochannel arrays for highly complete and accurate *de novo* sequence assembly of the complex *aegilops tauschii* genome. *PLoS One* 2013;**8**(2):e55864.
92. Mak ACY, Lai YYY, Lam ET, *et al*. Genome-wide structural variation detection by genome mapping on nanochannel arrays. *Genetics* 2016;**202**(1):351–62.
93. Kersey PJ, Allen JE, Armean I, *et al*. Ensembl genomes 2016: more genomes, more complexity. *Nucleic Acids Res* 2016;**44**(D1):D574–80.
94. Pertea M, Kim D, Pertea GM, *et al*. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc* 2016; 1650–67.
95. Edgar RC, Batzoglou S. Multiple sequence alignment. *Curr Opin Struct Biol* 2006;**16**(3):368–73.
96. Notredame C. Recent evolutions of multiple sequence alignment algorithms. *PLoS Comput Biol* 2007;**3**(8):e123.
97. Navarro G, Mäkinen V. Compressed full-text indexes. *ACM Comput Surv* 2007;**39**(1):61.
98. Rahn R, Weese D, Reinert K. Journaled string tree—a scalable data structure for analyzing thousands of similar genomes on your laptop. *Bioinformatics* 2014;**30**(24):3499–505.
99. Pevzner PA, Tang H, Tesler G. *De Novo* repeat classification and fragment assembly. *Genome Res* 2004;**14**(9):1786–96.
100. Paten B, Herrero J, Beal K, *et al*. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res* 2008;**18**(11):1814–28.
101. Paten B, Diekhans M, Earl D, *et al*. Cactus graphs for genome comparisons. *J Comput Biol* 2011;**18**(3):469–81.
102. Paten B, Earl D, Nguyen N, *et al*. Cactus: algorithms for genome multiple sequence alignment. *Genome Res* 2011;**21**(9):1512–28.
103. Kehr B, Trappe K, Holtgrewe M, *et al*. Genome alignment with graph data structures: a comparison. *BMC Bioinformatics* 2014;**15**(1):99.
104. Herbig A, Jäger G, Battke F, *et al*. GenomeRing: alignment visualization based on SuperGenome coordinates. *Bioinformatics* 2012;**28**(12):i7–i15.
105. Melsted P, Pritchard JK. Efficient counting of k-mers in DNA sequences using a bloom filter. *BMC Bioinformatics* 2011;**12**(1):333.
106. Rizk G, Lavenier D, Chikhi R. DSK: k-mer counting with very low memory usage. *Bioinformatics* 2013;**29**(5):652–3.
107. Deorowicz S, Kokot M, Grabowski S, *et al*. KMC 2: fast and resource-frugal k-mer counting. *Bioinformatics* 2015;**31**(10):1569–76.
108. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. *Genomics* 2010;**95**(6):315–27.
109. Chikhi R, Limasset A, Jackman S, *et al*. On the representation of de Bruijn graphs. In: R Sharan (ed.), *Research in Computational Molecular Biology, volume 8394 of Lecture Notes in Computer Science*. Springer International Publishing, Switzerland, 2014, 35–55.
110. Iqbal Z, Caccamo M, Turner I, *et al*. *De novo* assembly and genotyping of variants using colored de Bruijn graphs. *Nat Genet* 2012;**44**(2):226–32.
111. Holley G, Wittler R, Stoye J. Bloom filter trie - a data structure for pan-genome storage. In: *Proceedings of WABI. Springer-Verlag, Berlin Heidelberg, volume 9289 of LNBI*, 2015, 217–230.
112. Minkin I, Patel A, Kolmogorov M, *et al*. Sibelia: a scalable and comprehensive synteny block generation tool for closely related microbial genomes. In: A Darling, and J Stoye (eds), *Algorithms in Bioinformatics, number 8126 in Lecture Notes in Computer Science*. Springer-Verlag, Berlin Heidelberg, 2013, 215–229.
113. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014;**15**(3):R46.
114. Drezen E, Rizk G, Chikhi R, *et al*. GATB: Genome Assembly & Analysis Tool Box. *Bioinformatics* 2014;**30**(20):2959–61.
115. Marcus S, Lee H, Schatz MC. SplitMEM: a graphical algorithm for pan-genome analysis with suffix skips. *Bioinformatics* 2014;**30**(24):3476–83.
116. Beller T, Ohlebusch E. Efficient construction of a compressed de Bruijn graph for Pan-genome analysis. In: F Cicalese, E Porat and U Vaccaro (eds), *Combinatorial Pattern Matching, number 9133 in Lecture Notes in Computer Science*. Springer International Publishing, Switzerland, 2015, 40–51.
117. Baier U, Beller T, Ohlebusch E. Graphical pan-genome analysis with compressed suffix trees and the burrows-wheeler transform. *Bioinformatics* 2016;**32**(4):497–504.
118. Sheikhizadeh S, Schranz E, Akdel M, *et al*. Pantools: representation, storage, and exploration of pan-genomic data. *Bioinformatics* 2016;**32**(17):i487–93.
119. Ernst C, Rahmann S. PanCake: a data structure for pangenomes. In: T Beißbarth, M Kollmar, A Leha, B Morgenstern, AK Schultz, S Waack, and E Wingender (eds), *German Conference on Bioinformatics 2013. volume 34 of OpenAccess Series in Informatics (OASIcs)*. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2013, 35–45.

120. Durbin R. Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). *Bioinformatics* 2014;**30**(9):1266–72.
121. Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 2003;**165**(4):2213–33.
122. Beckstein C, Böcker S, Bogdan M, *et al*. Explorative analysis of heterogeneous, unstructured, and uncertain data: a computer science perspective on biodiversity research. In: M Helfert, A Holzinger, O Belo, and C Francalanci (eds), *Proceedings of the 3rd International Conference on Data Management Technologies and Applications, DATA 2014*, Vienna, Austria. SCITEPRESS, 2014, 251–57.
123. Stein LD, Knoppers BM, Campbell P, *et al*. Data analysis: create a cloud commons. *Nature* 2015;**523**(7559):149–51.
124. Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform* 2010;**11**(5):473–83.
125. Mäkinen V, Navarro G, Sirén J, *et al*. Storage and retrieval of individual genomes. In: S Batzoglou (ed.), *Research in Computational Molecular Biology, number 5541 in Lecture Notes in Computer Science*. Springer-Verlag, Berlin Heidelberg, 2009, 121–137.
126. Mäkinen V, Navarro G, Sirén J, *et al*. Storage and retrieval of highly repetitive sequence collections. *J Comput Biol* 2010;**17**(3):281–308.
127. Gagie T, Puglisi SJ. Searching and indexing genomic databases via kernelization. *Bioinform Comput Biol* 2015;**3**:12.
128. Schneeberger K, Hagmann J, Ossowski S, *et al*. Simultaneous alignment of short reads against multiple genomes. *Genome Biol* 2009;**10**(9):R98.
129. Danek A, Deorowicz S, Grabowski S. Indexes of large genome collections on a PC. *PLoS One* 2014;**9**(10):e109384.
130. Limasset A, Cazaux B, Rivals E, *et al*. Read mapping on de Bruijn graphs. *BMC Bioinformatics* 2016;**17**:237.
131. Huang L, Popic V, Batzoglou S. Short read alignment with populations of genomes. *Bioinformatics* 2013;**29**(13):i361–70.
132. Sirén J, Välimäki N, Mäkinen V. Indexing graphs for path queries with applications in genome research. *IEEE/ACM Trans Comput Biol Bioinform* 2014;**11**(2):375–88.
133. Sirén J, Välimäki N, Mäkinen V, Indexing finite language representation of population genotypes. In: TM Przytycka and MF Sagot (eds), *Algorithms in Bioinformatics, number 6833 in Lecture Notes in Computer Science*. Springer-Verlag, Berlin Heidelberg, 2011, 270–281.
134. Tattini L, D'Aurizio R, Magi A. Detection of genomic structural variants from next-generation sequencing data. *Front Bioeng Biotechnol* 2015;**3**:92.
135. Alioto TS, Buchhalter I, Derdak S, *et al*. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat Commun* 2015;**6**:10001.
136. Layer RM, Kindlon N, Karczewski KJ, *et al*. Efficient genotype compression and analysis of large genetic-variation data sets. *Nat Methods* 2016; **13**(1):63–5.
137. Tewhey R, Bansal V, Torkamani A, *et al*. The importance of phase information for human genomics. *Nat Rev Genet* 2011;**12**(3):215–23.
138. Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. *Nat Rev Genet* 2011;**12**(10): 703–14.
139. Rimmer A, Phan H, Mathieson I, *et al*. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet* 2014;**46**(8):912–18.
140. Patterson M, Marschall T, Pisanti N, *et al*. WhatsHap: weighted haplotype assembly for future-generation sequencing reads. *J Comput Biol* 2015;**22**(6):498–509.
141. Pirola Y, Zaccaria S, Dondi R, *et al*. HapCol: accurate and memory-efficient haplotype assembly from long reads. *Bioinformatics* 2016;**32**(11):1610–7.
142. Kuleshov V. Probabilistic single-individual haplotyping. *Bioinformatics* 2014;**30**(17):i379–85.
143. Aguiar D, Istrail S. Haplotype assembly in polyploid genomes and identical by descent shared tracts. *Bioinformatics* 2013;**29**(13):i352–60.
144. Berger E, Yorukoglu D, Peng J, *et al*. Haptree: a novel Bayesian framework for single individual polyplotyping using NGS data. *PLoS Comput Biol* 2014;**10**(3):e1003502.
145. Zagordi O, Bhattacharya A, Eriksson N, *et al*. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics* 2011;**12**(1):119.
146. Töpfer A, Marschall T, Bull RA, *et al*. Viral quasispecies assembly via maximal clique enumeration. *PLoS Comput Biol* 2014;**10**(3):e1003515.
147. Hennig A, Bernhardt J, Nieselt K. Pan-Tetris: an interactive visualisation for Pan-genomes. *BMC Bioinformatics* 2015;**16(Suppl 11)**:S3.
148. Nielsen CB, Cantor M, Dubchak I, *et al*. Visualizing genomes: techniques and challenges. *Nat Methods* 2010;**7**:S5–S15.
149. Nguyen N, Hickey G, Zerbino DR, *et al*. Building a Pan-genome reference for a population. *J Comput Biol* 2015;**22**(5):387–401.
150. Darling ACE, Mau B, Blattner FR, *et al*. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* 2004;**14**(7):1394–403.
151. Waterhouse AM, Procter JB, Martin DMA, *et al*. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 2009;**25**(9):1189–91.
152. Weisenfeld NI, Yin S, Sharpe T, *et al*. Comprehensive variation discovery in single human genomes. *Nat Genet* 2014;**46**(12):1350–55.
153. Shannon P, Markiel A, Ozier O, *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**(11):2498–504.
154. Wick RR, Schultz MB, Zobel J, *et al*. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* 2015;**31**(20):3350–2.
155. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 2008;**18**(11):1851–8.
156. Nielsen R, Paul JS, Albrechtsen A, *et al*. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* 2011;**12**(6):443–51.
157. Kavak P, Yüksel B, Aksu S, *et al*. Robustness of massively parallel sequencing platforms. *PLoS One* 2015;**10**(9):e0138259.
158. Paltoo DN, Rodriguez LL, Feolo M, *et al*. Data use under the NIH GWAS Data Sharing Policy and future directions. *Nat Genet* 2014;**46**(9):934–8.