iScience

CellPress
OPEN ACCESS

Perspective

Algorithms meet sequencing technologies – 10th edition of the RECOMB-Seq workshop

Rob Patro[1] and Leena Salmela[2,*]

SUMMARY

DNA and RNA sequencing is a core technology in biological and medical research. The high throughput of these technologies and the consistent development of new experimental assays and biotechnologies demand the continuous development of methods to analyze the resulting data. The RECOMB Satellite Workshop on Massively Parallel Sequencing brings together leading researchers in computational genomics to discuss emerging frontiers in algorithm development for massively parallel sequencing data. The 10th meeting in this series, RECOMB-Seq 2020, was scheduled to be held in Padua, Italy, but due to the ongoing COVID-19 pandemic, the meeting was carried out virtually instead. The online workshop featured keynote talks by Paola Bonizzoni and Zamin Iqbal, two highlight talks, ten regular talks, and three short talks. Seven of the works presented in the workshop are featured in this edition of iScience, and many of the talks are available online in the RECOMB-Seq 2020 YouTube channel.

The development of sequencing technologies has revolutionized biological and medical research in last decades. High-thoughput sequencing by synthesis made genomic analysis cost effective bringing it to the standard toolbox of most biological and medical research groups. The growth in available sequencing data has been staggering as exemplified by the number of deposited genomes in the RefSeq database over the years (Leary et al., 2015), as shown in Figure 1A. Today, the field is dominated by two types of sequencing technologies: massively parallel sequencing by synthesis technologies such as Illumina that produce short but highly accurate sequencing reads and long reads produced by third-generation sequencing technologies such as Pacific Biosciences' single molecule, real-time (SMRT) sequencing and Oxford Nanopore's pore-based sequencing which produce long but highly erroneous sequencing reads. Short reads typically are a few hundred base pairs in length with an error rate of $\sim 0.1\%$, whereas long reads are tens of thousands of base pairs long with an error rate of up to 13% (Kozińska et al., 2019). Long reads are favored in applications where long range information is important, such as genome assembly and determining long structural variations. On the other hand, the per-base sequencing cost is lower for short reads, and thus, they are popular in applications where the depth of sequencing is crucial such as metagenomics and RNA sequencing. Some tools attempt to combine the best of both worlds and develop hybrid strategies using both short and long reads. Yet other technologies, such as optical mapping, Hi-C sequencing, or linked reads, are used to further improve the analysis (Rice and Green, 2019).

Two basic problems arise for analyzing sequencing data: Mapping localizes the sequencing reads on a reference genome or genome graph, and assembly reconstructs the sequence or sequences from which the reads originate. Both of these problems have been studied extensively, yet fundamental challenges remain. Mapping may produce incorrect alignments and assemblies may be fragmented and contain errors. Also, computational efficiency needs to be addressed. HASLR (Haghshenas et al., 2020) introduces a new method for genome assembly using a hybrid approach. Its assembly strategy is based on a backbone graph which is built on preassembled short read contigs, which are aligned to long reads to reveal connections between the contigs. CONNET (Zhang et al., 2020) addresses the high number of errors present in long read assemblies by providing a fast method based on deep learning to compute the consensus between aligned reads.

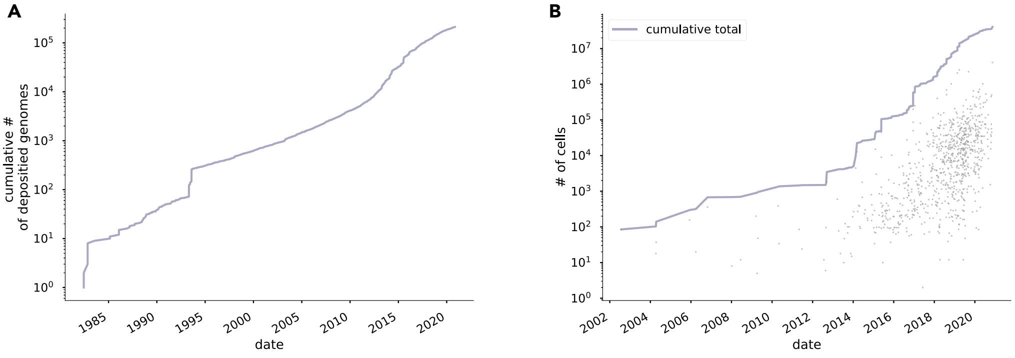
Short read aligners such as BWA (Li and Durbin, 2009) and Bowtie (Langmead and Salzberg, 2012) rely on Burrows-Wheeler transform (BWT)-based techniques for efficient mapping of short reads to a reference

[1] Department of Computer Science and Center for Bioinformatics and Computational Biology, University of Maryland, MD, USA

[2] Department of Computer Science and Helsinki Institute for Information Technology HIIT, University of Helsinki, Helsinki, Finland

*Correspondence: leena.salmela@helsinki.fi

https://doi.org/10.1016/j.isci.2020.101956



deposited genomes

single-cell RNA-seq experiments

Figure 1. The growth in available genomic data—both raw sequencing data and processed data—is staggering

The plot in Figure 1A shows the growth, over time, of the total number of genomes deposited in the RefSeq database (Leary et al., 2015); note the y axis is on a log scale. The number of available assemblies has been increasing at an exponential rate, and the availability of such a wide and growing variety of references highlights the importance of developing scalable approaches for pan-genomic representation and indexing. Likewise, the plot in Figure 1B (with data as reported in [Svensson et al., 2019]) shows the growth, over time, of the total number of reported cells in different single-cell RNA-seq sequencing experiments, with the blue line signifying the total cumulative number of reported cells. The clear trend is that more recent studies report sequencing results on more individual cells, with one recent study (Cao et al., 2020) reporting $\sim$ 4 million cells.

genome. When long reads emerged, transforming these techniques directly to long read aligners was problematic as BWT-based seed-finding techniques were less efficient in reads having higher error rates. Instead, techniques based on minimizers (Kucherov, 2019) were introduced and these form the basis of many long read aligners such as Minimap (Li, 2018) today. However, repetitive regions remain problematic for minimizer-based aligners because they are often discarded in the minimizer index for efficiency reasons and thus reads are not aligned correctly or at all to such areas. Jain et al. (2020) propose weighted minimizers as a method to mitigate these shortcomings and to produce accurate alignments in repetitive areas.

As high error rates are problematic for both assembly algorithms, as well as read aligners, methods have been developed to correct sequencing errors in the reads before further processing (Zhang et al., 2019). This can greatly improve the accuracy of assembly or alignment. Because of the high accuracy of short reads, long reads are often corrected by hybrid methods utilizing both short and long reads. Correction of whole-genome sequencing reads has received a fair amount of attention, and fewer methods have focused on other types of data. TALC (Broseus et al., 2020) studies hybrid error correction of RNA-seq data. Special attention is paid to taking into account transcript abundance and architecture in the correction process to produce more accurate results.

RNA-seq—the sequencing of RNA molecules (or their reverse-transcribed counterparts)—is one of the most popular and widely used sequencing assays. Sequencing the RNA molecules in the cell can provide a window into the dynamic processes that occur within different types of tissues, in response to different stimuli, in different disease states, or under a host of other perturbations. While RNA-seq enables the posing and answering of an array of questions that are different from those arrived at by genome sequencing, it also brings with it a distinct set of computational challenges.

One of the core challenges in processing RNA-seq data is the quantification of transcript abundance levels from the underlying set of sequencing reads and a known or derived (i.e. assembled) annotation of the transcripts in the organism being assayed. Many challenges stand in the way of accurate quantification, such as prevalent multimapping of sequencing reads among genes and transcripts (Li et al., 2010), extensive sample-specific sequence (Roberts et al., 2011; Jones et al., 2012) and fragment-level (Love et al., 2016; Patro et al., 2017) biases, and reference divergence (Munger et al., 2014). Many methods have been developed to tackle the problem of transcript abundance estimation (Trapnell et al., 2010; Li et al., 2010; Turro et al., 2011; Glaus et al., 2012; Roberts and Pachter, 2013; Patro et al., 2014, 2017) that apply a range of different probabilistic models and inference techniques to estimate transcript abundance.

The XAEM method (Deng et al., 2019) adopts a bilinear model for transcript-level quantification that aims to perform multi-sample inference, considering evidence from multiple samples within the same RNA-seq experiment jointly when performing quantification. The model can be viewed as a generalization of more common transcript quantification models where the so-called "design" matrix is fixed, and inference solves for the maximum likelihood parameters under this design matrix and the observed sequences. Instead, the XAEM method takes both the parameters and the design matrix as unknown and uses an alternating expectation maximization algorithm to solve the resulting inference problem. This provides a mechanism to share information across the samples being quantified and to account, within the model, for certain sample-specific biases even when the full causes of such biases may not be known. The authors demonstrate that the XAEM algorithm accurately quantifies transcript abundances and is particularly effective, compared to existing methods, in quantifying paralogous sets of genes. They also provide a mechanism for determining when a set of transcripts may not be reliably distinguishable based on the annotation and observed sequencing reads and offer the potential to quantify them as a group.

The TALC method (Broseus et al., 2020), mentioned briefly above, tackles another important problem in the processing of RNA-seq data—error correction for long-read RNA-seq data. Long read RNA-sequencing offers to greatly improve the quality of transcript assembly (Kovaka et al., 2019; Tung et al., 2019; Workman et al., 2019; Wyman et al., 2020). Yet, to maximize this potential, the long sequencing reads should be error corrected to improve alignment (either to the reference or among multiple long-read sequences), which, in turn, can lead to more accurate identification of transcript structure. While many methods have been introduced for "hybrid" long-read error correction (that attempt to correct error-prone long reads with low error rate short reads), TALC is specifically designed to handle long-read transcriptome data, which is distinct since the distribution of sequencing reads follows the underlying distribution of transcript abundances in the samples being assayed. TALC makes use of a weighted de Bruijn graph constructed on short reads and corrects long reads by identifying them with a coverage-consistent path through the weighted de Bruijn graph. The authors show that hybrid error correction of long reads using TALC improves base-level error rates, improves recovery of correct exon structure (which aids in accurate transcript identification), and also leads to improved quantification from long-read sequencing data. TALC generally outperforms other hybrid error correction methods that were not designed for and therefore are not particularly aware of the fact that the paired long and short reads are arising from the transcriptome.

While bulk RNA-seq remains a popular experimental assay, the prevalence of single-cell RNA-seq data has been increasing drastically over the past few years as shown in Figure 1B. These groundbreaking technologies allow measuring gene expression at the resolution of individual cells, letting scientists probe gene expression at unprecedented resolution. This, in turn, has helped contribute to our understanding of cellular diversity, treatment resistance, and developmental biology among other areas. There were three talks at RECOMB Satellite Workshop on Massively Parallel Sequencing (RECOMB-Seq) on the topic of single-cell RNA-sequencing analysis.

Do et al. (2020) presented Sphetcher, a new method for sketching large single-cell RNA-seq data. The sketch is a small representation of the larger data set that retains key properties, like the overall geometry of the transcriptional space, while also retaining information about potentially rare cell types. These sketches can act as small stand-ins for the larger data set to aid in many downstream analyses, like trajectory inference. Sphetcher works by producing a spherical sketch of the underlying data such that all cells in the data set are covered by small spheres centered around a subset of the original cells. Do et al. demonstrate that Sphetcher produces a faithful sketch of the original data set, resulting in a smaller Hausdorff distance to the original data set than alternative sketching approaches like geometric sketching (Hie et al., 2019).

Another key challenge in the analysis of single-cell RNA-seq data is the integration of multiple data sets into a unified "atlas" of cells. While different samples are assayed using the same technology, manifold technical differences can make the comparison of cells across samples difficult and make it challenging to properly integrate data from different samples together. To address this challenge, Mandric et al. (2020) introduce BATMAN. This method determines anchor cells and builds an anchor graph between the two data sets to be integrated. It then resolves which anchors to match by solving a minimum weight bipartite matching problem, while simultaneously computing batch effect correction factors for the anchor cells and extrapolating these corrections to the rest of the data set. The authors show that BATMAN outperforms existing approaches in both simulated and experimental data when integrating multiple different data sets in the presence of batch effects.

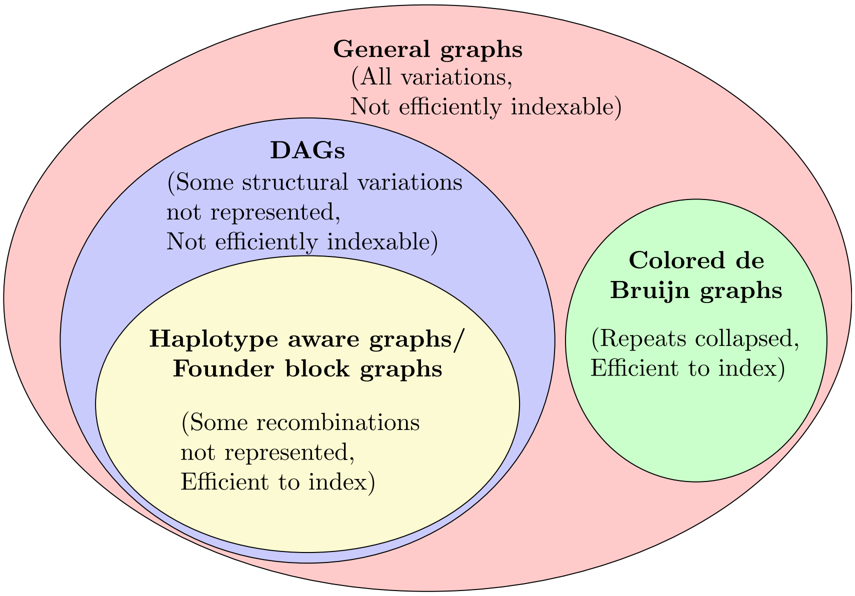


Figure 2. Graph-based representations of pangenomes and the trade-offs between what can be represented and what can be efficiently indexed

General graphs can express all variations and directed acyclic graphs (DAGs) only miss some structural variations but these representations are not efficiently indexable (Equi et al., 2020). Haplotype-aware graphs (Sirén et al., 2019) and founder block graphs (Mäkinen et al., 2020) are a restricted form of DAGs that can be efficiently indexed. Colored de Bruijn graphs (Iqbal et al., 2012) are efficiently indexable (Almodaresi et al., 2018) but collapse repeats.

The remaining single-cell work presented at RECOMB-Seq focused on bridging the analysis gap between single-cell RNA-seq and single-cell ATAC-seq and DNA methylation data. Danese et al. (2019) described EpiScanpy which brings a host of the visualization and analysis techniques available in the popular Scanpy tool (Wolf et al., 2018) to bear in single-cell epigenetic analysis using single-cell ATAC-seq and single-cell DNA methylation data. As multi-omic single-cell data become more common, such analysis frameworks will become increasingly important. To demonstrate the utility of EpiScanpy, Danese et al. apply it to single-cell atlases of mouse brain containing single-cell ATAC-seq, single-cell DNA methylation, and single-cell RNA-seq data. They show that EpiScanpy identifies differential methylation and chromatin accessibility between cell clusters that complement each other and also the single-cell gene expression data. These different data modalities can thus complement each other, for many purposes such as the discovery of cell-type markers, when analyzed by powerful and unifying frameworks such as Scanpy and EpiScanpy.

Pangenomics was an unexpectedly unifying theme of RECOMB-Seq 2020, as it was the main topic of the keynotes on both days of the workshop. Prof. Paola Bonizzoni showcased an in-depth analysis of previous and current developments in the graph-based representation of genomes, including the relationships and trade-offs of different representations as shown in Figure 2, and the primary computational challenges to be addressed as the community seeks to scale graph-based solutions to the ever-growing wealth of genomic data. Her keynote laid out the theoretical foundations of graph-based genome representation, highlighting known hardness results, but focusing on recent scalability improvements and motivating the need for continued algorithmic development to meet the promise of graph-based genome and pangenome analysis.

Dr. Zamin Iqbal, the second keynote speaker, focused strongly on bacterial pangenomics and the particular challenges and opportunities brought by bacterial genome diversity. He argued that the graphical representation of bacterial pangenomes is a powerful tool for the analysis of genetic variation among of large collections of bacterial strains and species. Dr. Iqbal discussed the design, implementation, and validation of ”pandora” (Colquhoun et al., 2020), a tool that his lab has created to discover and characterize genomic variants in bacterial data using both long-read (ONT) and short-read (Illumina) data. The results, which demonstrate high sensitivity and a low degree of reference bias for pandora, strongly motivate the continued development and use of graph-based tools for pangenomic analysis, especially in bacterial genomes.

Due to projects such as the Vertebrate Genome Project and UK10K, an increasing number of genomes are becoming available, and thus, the development of tools for the analysis of these massive data sets is becoming more and more important. Pangenomics and the graph representations of pangenomes discussed above is one example of such analysis but also to answer fundamental questions about genome evolution and function these genomes need to be compared to each other. However, comparison of these extremely long sequences is computationally difficult and as the data volumes keep growing new methods are needed. Pairwise alignment of sequences has quadratic complexity and thus quickly becomes infeasible for long sequences. Thus, homologous blocks are first identified instead. Apart from being a starting point for building multiple alignments and pangenome graphs, homologous blocks can be used for studying genome rearrangements and phylogenetics. Bubbz (Minkin and Medvedev, 2020) addresses this problem by building a compacted de Bruijn graph of the sequences and studying the paths induced by them in the graph.

To understand genome structure in a population, a first step is to compute haplotype blocks for the population. A haplotype block is a region in the genome where little or no recombinations occur, i.e., the block is inherited as a whole (Gabriel et al., 2002). Williams and Mumey (2020) identify such blocks given a set of single-nucleotide polymorphisms (SNPs) called for many individuals from the same population. Specifically, they extend the notion of haplotype blocks to include wildcards as in practice SNP sets are seldom perfect and thus some SNP calls are missing for some individuals.

Metagenomics studies microbial communities in, e.g., environmental samples from any biome (Gilbert et al., 2014) or in samples taken from the human gut or skin (Turnbaugh et al., 2007). Metagenomics attempts to answer two questions about these samples: who is there and what are they doing. Because metagenomics attempts to capture the whole biological diversity of the sample, the data sets are huge, easily surpassing the data set sizes for sequencing a single organism. Therefore, scalable and efficient algorithms are even more important for metagenomics than for sequencing single organisms. Both Utro et al. (2020) and LaPierre et al. (2020) provide scalable and efficient solutions for metagenomics. Utro et al. (2020) present a new method based on BWT for phylogenetic and functional annotation, while Metalign by LaPierre et al. (2020) introduces an efficient alignment-based method based on min hash for taxonomic profiling. Besides the two fundamental questions of taxonomic profiling and functional annotation, some special questions arise in metagenomics. Pellow et al. (2020) study the assembly of plasmids in metagenomic samples. Plasmids are important, e.g., to human health since they can transfer antibiotic resistance from one bacteria to another.

Computational biology is intrinsically tied to the technological development in molecular biology as exemplified by how the design of read aligners has developed as sequencing technologies have advanced. One of the most ground-breaking discoveries in biology in the recent years is CRISPR/Cas9 (clustered regularly interspaced short palindromic repeats) gene editing (Cong et al., 2013; Mali et al., 2013; Doudna and Charpentier, 2014) for which Emmanuelle Charpentier and Jennifer A. Doudna were awarded the Nobel prize in chemistry in 2020. CRISPR/Cas9 makes it possible to edit a living organism's genome, thus opening up new horizons in medicine and biology. Such discoveries also bring up new computational problems. In the case of CRISPR/Cas9, a central problem is detecting off-target sites, i.e., regions of the genome that gene editing could target instead of the intended site. CRISPRitz (Cancellieri et al., 2019) detects such off-target sites in a variant-aware fashion while being computationally more efficient than previous work. In the coming years, we expect many such new computational problems to arise and to be discussed in the future meetings of RECOMB-Seq.

Despite the challenges of organizing RECOMB-Seq as a virtual workshop this year, we had a program with great talks. Many of the talks are available on the RECOMB-Seq 2020 YouTube channel (RECOMB-Seq 2020, 2020). Moreover, as the virtual meeting allowed participation from anywhere in the world, we had a considerably larger audience than in previous editions of the meeting. Thus, while the meeting was certainly different than in past years, it was nonetheless quite successful. This, of course, is due to the great contributed talks, fantastic keynote addresses, and thought-provoking audience interaction.

The dynamics of a virtual meeting are quite different than those of an in-person meeting, and we note certain benefits and drawbacks. The biggest drawbacks are the most obvious ones; there is no opportunity for the chance in person interactions (e.g. during a coffee break) that can be so gratifying, especially at smaller meetings like RECOMB-Seq. Also, while remote attendance allows synchronous interaction from participants across the world, these participants still reside in their own local time zones, making attendance of certain parts of the meeting more difficult for some than others. However, there were clear benefits as well. The recording of the talks was generally well received. The video conference format also allows for a broader diversity in the way in which questions are asked and fielded by speakers. In addition to directly asking questions, audience members may also type their questions to be read by speakers directly or chosen by session moderators. This may appeal to attendees who are more comfortable formulating their question in writing than having to ask it in real time in front of the audience. Further, the video conference format chat allows for a discussion, via chat, with presenters and other audience members, which seems to add to the experience. These benefits suggest that, even when in-person meetings resume, there may be reasons to consider allowing remote attendance or to avail ourselves of some of the technical capabilities offered by large-scale videoconferencing software. These are possibilities that RECOMB-Seq may consider going forward.

The success of the conference was due to the hard work of many people. We want to thank the speakers for their excellent talks, the program committee for their hard work in reviewing and selecting the scientific program, and the steering committee of RECOMB-Seq for their support. Finally, we would especially like to thank the organizing committee at University of Padua for making RECOMB-Seq a success despite the difficult circumstances caused by the COVID-19 pandemic. As we reflect back on 10 years of RECOMB-Seq and look forward to RECOMB-Seq 2021, it is certainly clear that the need for algorithm development for massively parallel sequencing data is as pressing and important as ever and that the RECOMB-Seq community continues to rise to this challenge.

ACKNOWLEDGMENTS

L.S. is supported by the Academy of Finland (grants 308030 and 323233). R.P. is supported by NIH R01 HG009937 and by NSF CCF-1750472 and CNS-1763680.

AUTHOR CONTRIBUTIONS

Conceptualization, R.P. and L.S.; Writing - Original manuscript, R.P. and L.S.; Writing - Review & Editing, R.P. and L.S.; Funding Acquisition, R.P. and L.S.

DECLARATION OF INTERESTS

R.P. is a co-founder and CTO of Ocean Genomics Inc.

REFERENCES

Almodaresi, F., Sarkar, H., Srivastava, A., and Patro, R. (2018). A space and time-efficient index for the compacted colored de bruijn graph. Bioinformatics *34*, i169–i177.

Broseus, L., Thomas, A., Oldfield, A.J., Severac, D., Dubois, E., and Ritchie, W. (2020). TALC: Transcript-level aware long read correction. bioRxiv. https://doi.org/10.1101/2020.01.10.901728. https://www.biorxiv.org/content/early/2020/06/22/2020.01.10.901728.

Cancellieri, S., Canver, M.C., Bombieri, N., Giugno, R., and Pinello, L. (2019). CRISPRitz: rapid, high-throughput and variant-aware in silico off-target site identification for CRISPR genome editing. Bioinformatics *36*, 2001–2008, https://doi.org/10.1093/bioinformatics/btz867.

Cao, J., O'Day, D.R., Pliner, H.A., Kingsley, P.D., Deng, M., Daza, R.M., Zager, M.A., Aldinger, K.A., Blecher-Gonen, R., Zhang, F., et al. (2020). A human cell atlas of fetal gene expression. Science *370*, eaba7721, https://doi.org/10.1126/science.aba7721.

Colquhoun, R.M., Hall, M.B., Lima, L., Roberts, L.W., Malone, K.M., Hunt, M., Letcher, B., Hawkey, J., George, S., Pankhurst, L., and Iqbal, Z. (2020). Nucleotide-resolution bacterial pan-genomics with reference graphs. bioRxiv. https://doi.org/10.1101/2020.11.12.380378. https://www.biorxiv.org/content/early/2020/11/13/2020.11.12.380378.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., and Zhang, F. (2013). Multiplex genome engineering using crispr/cas systems. Science *339*, 819–823, https://doi.org/10.1126/science.1231143. https://science.sciencemag.org/content/339/6121/819.

Danese, A., Richter, M.L., Fischer, D.S., Theis, F.J., and Colomé-Tatché, M. (2019). EpiScanpy: integrated single-cell epigenomic analysis. bioRxiv. https://doi.org/10.1101/648097. https://www.biorxiv.org/content/early/2019/05/24/648097.

Deng, W., Mou, T., Kalari, K.R., Niu, N., Wang, L., Pawitan, Y., and Vu, T.N. (2019). Alternating EM algorithm for a bilinear model in isoform quantification from RNA-seq data. Bioinformatics *36*, 805–812, https://doi.org/10.1093/bioinformatics/btz640.

Do, V.H., Elbassioni, K., and Canzar, S. (2020). Sphetcher: spherical thresholding improves sketching of single-cell transcriptomic heterogeneity. iScience *23*, 101126, https://doi.org/10.1016/j.isci.2020.101126. http://www.sciencedirect.com/science/article/pii/S2589004220303114.

Doudna, J.A., and Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. Science *346*, https://doi.org/10.1126/science.1258096. https://science.sciencemag.org/content/346/6213/1258096.

Equi, M., Mäkinen, V., and Tomescu, A.I. (2020). Graphs cannot be indexed in polynomial time for sub-quadratic time string matching, unless seth fails. arXiv. https://arxiv.org/abs/2002.00629.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., et al. (2002). The structure of haplotype blocks in the human genome. Science *296*, 2225–2229, https://doi.org/10.1126/science.1069424. https://science.sciencemag.org/content/296/5576/2225.

Gilbert, J.A., Jansson, J.K., and Knight, R. (2014). The earth microbiome project: successes and aspirations. BMC Biol. *12*, 69.

Glaus, P., Honkela, A., and Rattray, M. (2012). Identifying differentially expressed transcripts from RNA-seq data with biological variation. Bioinformatics *28*, 1721–1728.

Haghshenas, E., Asghari, H., Stoye, J., Chauve, C., and Hach, F. (2020). HASLR: fast hybrid assembly of long reads. iScience, 101389, https://doi.org/10.1016/j.isci.2020.101389. http://www.

sciencedirect.com/science/article/pii/S2589004220305770.

Hie, B., Cho, H., DeMeo, B., Bryson, B., and Berger, B. (2019). Geometric sketching compactly summarizes the single-cell transcriptomic landscape. Cell Syst. *8*, 483–493.

Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., and McVean, G. (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat. Genet. *44*, 226–232.

Jain, C., Rhie, A., Zhang, H., Chu, C., Walenz, B.P., Koren, S., and Phillippy, A.M. (2020). Weighted minimizer sampling improves long read mapping. Bioinformatics *36*, i111–i118, https://doi.org/10.1093/bioinformatics/btaa435.

Jones, D.C., Ruzzo, W.L., Peng, X., and Katze, M.G. (2012). A new approach to bias correction in rna-seq. Bioinformatics *28*, 921–928.

Kovaka, S., Zimin, A.V., Pertea, G.M., Razaghi, R., Salzberg, S.L., and Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. *20*, 1–13.

Kozińska, A., Seweryn, P., and Sitkiewicz, I. (2019). A crash course in sequencing for a microbiologist. J. Appl. Genet. *60*, 103–111.

Kucherov, G. (2019). Evolution of biosequence search algorithms: a brief survey. Bioinformatics *35*, 3547–3552, https://doi.org/10.1093/bioinformatics/btz272.

Langmead, B., and Salzberg, S. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

LaPierre, N., Alser, M., Eskin, E., Koslicki, D., and Mangul, S. (2020). Metalign: efficient alignment-based metagenomic profiling via containment min hash. Genome Biol. *21*, 242.

Leary, N.A.O., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2015). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44*, D733–D745, https://doi.org/10.1093/nar/gkv1189.

Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A., and Dewey, C.N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics *26*, 493–500.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100, https://doi.org/10.1093/bioinformatics/bty191.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760, https://doi.org/10.1093/bioinformatics/btp324.

Love, M.I., Hogenesch, J.B., and Irizarry, R.A. (2016). Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. Nat. Biotechnol. *34*, 1287.

Mäkinen, V., Cazaux, B., Equi, M., Norri, T., and Tomescu, A.I. (2020). Linear time construction of indexable founder block graphs. In 20th International Workshop on Algorithms In Bioinformatics (WABI 2020), Volume 172 of Leibniz International Proceedings In Informatics (LIPIcs), C. Kingsford and N. Pisanti, eds. (Schloss Dagstuhl–Leibniz-Zentrum für Informatik), pp. 7:1–7:18, https://doi.org/10.4230/LIPIcs.WABI.2020.7. https://drops.dagstuhl.de/opus/volltexte/2020/12796.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013). RNA-guided human genome engineering via cas9. Science *339*, 823–826, https://doi.org/10.1126/science.1232033. https://science.sciencemag.org/content/339/6121/823.

Mandric, I., Hill, B.L., Freund, M.K., Thompson, M., and Halperin, E. (2020). BATMAN: fast and accurate integration of single-cell RNA-seq datasets via minimum-weight matching. iScience *23*, 101185, https://doi.org/10.1016/j.isci.2020.101185. http://www.sciencedirect.com/science/article/pii/S2589004220303709.

Minkin, I., and Medvedev, P. (2020). Scalable pairwise whole-genome homology mapping of long genomes with bubbz. iScience *23*, 101224, https://doi.org/10.1016/j.isci.2020.101224. http://www.sciencedirect.com/science/article/pii/S2589004220304090.

Munger, S.C., Raghupathy, N., Choi, K., Simons, A.K., Gatti, D.M., Hinerfeld, D.A., Svenson, K.L., Keller, M.P., Attie, A.D., Hibbs, M.A., et al. (2014). RNA-Seq alignment to individualized genomes improves transcript abundance estimates in multiparent populations. Genetics *198*, 59–73.

Patro, R., Mount, S.M., and Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat. Biotechnol. *32*, 462.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods *14*, 417.

Pellow, D., Probst, M., Furman, O., Zorea, A., Segal, A., Mizrahi, I., and Shamir, R. (2020). SCAPP: an algorithm for improved plasmid assembly in metagenomes. bioRxiv. https://doi.org/10.1101/2020.01.12.903252. https://www.biorxiv.org/content/early/2020/04/04/2020.01.12.903252.

RECOMB-Seq. (2020). RECOMB-seq 2020 YouTube Channel. https://www.youtube.com/channel/UCI_DNmSqGvJcdba_F0ZjbaA.

Rice, E.S., and Green, R.E. (2019). New approaches for genome assembly and scaffolding. Annu. Rev. Anim. Biosci. *7*, 17–40, https://doi.org/10.1146/annurev-animal-020518-115344.

Roberts, A., and Pachter, L. (2013). Streaming fragment assignment for real-time analysis of sequencing experiments. Nat. Methods *10*, 71–73.

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J.L., and Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol. *12*, 1–14.

Sirén, J., Garrison, E., Novak, A.M., Paten, B., and Durbin, R. (2019). Haplotype-aware graph indexes. Bioinformatics *36*, 400–407, https://doi.org/10.1093/bioinformatics/btz575.

Svensson, V., da Veiga Beltrame, E., and Pachter, L. (2019). A curated database reveals trends in single-cell transcriptomics. Bioinformatics *36*, 400–407.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

Tung, L.H., Shao, M., and Kingsford, C. (2019). Quantifying the benefit offered by transcript assembly with Scallop-LR on single-molecule long reads. Genome Biol. *20*, 1–18.

Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., and Gordon, J.I. (2007). The human microbiome project. Nature *449*, 804–810.

Turro, E., Su, S.-Y., Gonçalves, Â., Coin, L.J., Richardson, S., and Lewin, A. (2011). Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. Genome Biol. *12*, R13.

Utro, F., Haiminen, N., Siragusa, E., Gardiner, L.-J., Seabolt, E., Krishna, R., Kaufman, J.H., and Parida, L. (2020). Hierarchically labeled database indexing allows scalable characterization of microbiomes. iScience *23*, 100988, https://doi.org/10.1016/j.isci.2020.100988. http://www.sciencedirect.com/science/article/pii/S2589004220301723.

Williams, L., and Mumey, B. (2020). Maximal perfect haplotype blocks with wildcards. iScience *23*, 101149, https://doi.org/10.1016/j.isci.2020.101149. http://www.sciencedirect.com/science/article/pii/S2589004220303345.

Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15.

Workman, R.E., Tang, A.D., Tang, P.S., Jain, M., Tyson, J.R., Razaghi, R., Zuzarte, P.C., Gilpatrick, T., Payne, A., Quick, J., et al. (2019). Nanopore native RNA sequencing of a human poly (A) transcriptome. Nat. Methods *16*, 1297–1305.

Wyman, D., Balderrama-Gutierrez, G., Reese, F., Jiang, S., Rahmanian, S., Forner, S., Matheos, D., Zeng, W., Williams, B., Trout, D., et al. (2020). A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. bioRxiv. https://doi.org/10.1101/672931. https://www.biorxiv.org/content/early/2020/03/24/672931.

Zhang, H., Jain, C., and Aluru, S. (2019). A comprehensive evaluation of long read error correction methods. bioRxiv. https://doi.org/10.1101/519330. https://www.biorxiv.org/content/early/2019/05/29/519330.

Zhang, Y., Liu, C.-M., Leung, H.C., Luo, R., and Lam, T.-W. (2020). CONNET: accurate genome consensus in assembling nanopore sequencing data via deep learning. iScience *23*, 101128, https://doi.org/10.1016/j.isci.2020.101128. http://www.sciencedirect.com/science/article/pii/S2589004220303138.