Pseudoalignment facilitates assignment of error-prone Ultima Genomics reads

A. Sina Booeshaghi[1] and Lior Pachter[2,3,*]

[1]Department of Mechanical Engineering, California Institute of Technology, Pasadena, CA
[2]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA
[3]Department of Computing and Mathematical Sciences, Pasadena, CA
[*]Address correspondence to: lpachter@caltech.edu

Abstract

We analyze single-cell RNA-seq data sequenced with Ultima Genomics technology and find high error rates in and near homopolymers. To compensate for these errors, we explore the use of pseudoalignment for read assignment, and find that it can perform better than standard read alignment. Our pseudoalignment read assignment for Ultima Genomics data is available as part of the kallisto-bustools kb-python package available at https://github.com/pachterlab/kb_python.

Introduction

Despite numerous improvements in DNA sequencing technology and dramatic reductions in the price of sequencing over the past fifteen years (LeMieux 2019; Goodwin, McPherson, and McCombie 2016), the cost of sequencing can limit the scope of projects for biology labs (Haichao Li et al. 2019), and is a barrier to adoption of routine sequencing in the clinic (Schwarze et al. 2020). The recently unveiled Ultima Genomics sequencing technology (Almogy et al. 2022) has been advertised as providing a solution to these challenges by way of delivering high-throughput sequencing at a small fraction of the cost of current technologies ("Ultima Genomics Delivers the $100 Genome" 2022)[1].

For single-cell RNA-seq analysis, a pilot project conducted by scientists at the Broad Institute and at Ultima Genomics found that the "data show comparable results to existing technology" (Simmons et al. 2022). However, in examining the preprint we noticed that this claim is mostly based on an apples-to-oranges comparison of (on average) 176 bp long Ultima Genomics cDNA reads to 55 bp long Illumina cDNA reads (Supplementary Figure 1). Furthermore, lower quality scores for the Ultima Genomics reads than for the Illumina reads (Figure 1) motivated us to analyze the data to see whether higher error rates in Ultima Genomics reads reduce alignment rates, and consequently degrade expression estimates for genes.

[1] The "cost of sequencing" is not a well-defined concept. It can refer to only the cost of reagents for a sequencing run, or it can include other costs such as library preparation, personnel, analysis, space etc. Thus, statements such as "the $100 genome" are not meaningful; for example the oft cited NIH website for cost estimates (Kris A. Wetterstrand 2019) includes numerous production costs other than just instruments and reagents. This is because the reagent costs can be dominated by other costs (Haichao Li et al. 2019). Furthermore, there is no accepted accuracy or completeness standard for the "sequence of a human genome". For example Nebula Genomics sells a $99 genome ("Nebula Genomics - 30x Whole Genome Sequencing - DNA Testing" 2022), but at 0.4x coverage rather than the more common 30x.

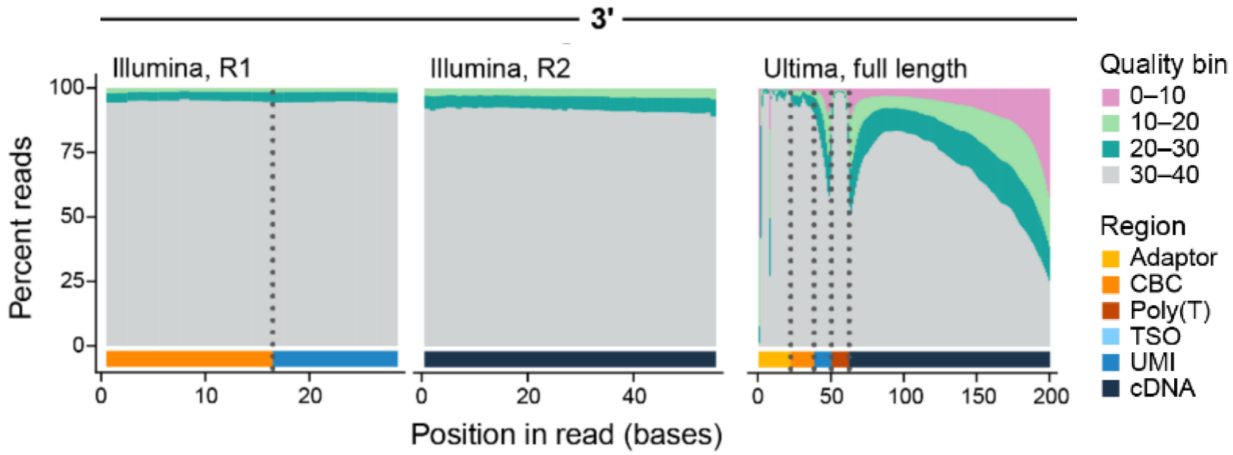

Figure 1: A reproduction of part of Extended Data Figure 1 from (Simmons et al. 2022) which is licensed under the CC-BY 4.0 license. The Illumina reads display high quality across the barcode, UMI and cDNA sequences. The Ultima Genomics reads have lower quality at and around the Poly(T) tract, and also degraded quality after 100 bp.

Results

(Simmons et al. 2022) prepared PBMC libraries (SRX14293374) that were sequenced with both Ultima Genomics (SRR18145555) and Illumina (SRR18145553) sequencers. To perform a like-to-like comparison of (SRR18145553 and SRR18145555), we trimmed the Ultima Genomics reads to a maximum length of 55bp so as to match the length of the Illumina reads. This trimming was similar to the procedure implemented by (Simmons et al. 2022) for the analysis underlying their Extended Data Figure 3H. We then pre-processed the data using kallisto-bustools (Melsted et al. 2021; Bray et al. 2016), which we modified in order to be able to parse the Ultima Genomics data (Supplementary Figure 2). This resulted in gene count matrices derived from both the Ultima Genomics and Illumina sequenced cDNA libraries (see Methods). Our results corroborated the findings of (Simmons et al. 2022) that at a high-level, the “data show comparable results” (Supplementary Figure 3). However, we noticed that not all genes had similar numbers of counts, and to understand why that may be the case we examined the nuclear gene with the highest difference, which was *TMSB4X*. This gene happens to be the 10th most highly expressed gene in the Illumina dataset. We found a 1.96x-fold difference in UMI counts depending on whether the library was sequenced with Ultima Genomics or Illumina; (Simmons et al. 2022) also identified *TMSB4X* as an outlier in a comparison of Illumina vs. Ultima Genomics (see Simmons et al. 2022 Supplementary Table 2) but did not investigate the cause for the difference.

To understand why the Ultima Genomics UMI counts were much lower than the Illumina UMI counts, we re-aligned the reads to the *TMSB4X* gene using HISAT2 (Kim et al. 2019), and examined the alignments with the IGV browser (Robinson et al. 2011); Figure 2). The number of aligned Ultima Genomics reads was more than 4 times lower than the number of aligned Illumina reads (Table 1), and we noticed a much higher incidence of errors in the Ultima

Genomics reads, specifically around homopolymer runs such as the $(T)_8$ homopolymer near the 3' end of the gene. This is consistent with the drop in quality scores near the $(T)_n$ homopolymers resulting from the poly(A) tracts at the 3' end of genes in (Simmons et al. 2022). To quantify the differences, we computed the error rates for Illumina and Ultima Genomics and found that the Ultima Genomics error rate was 10-fold higher (Table 1). We note that the errors displayed in Figure 2 and the estimates in Table 1 are only lower bounds for the error rate of Ultima Genomics because we could not align the reads with the most errors (Supplementary Figure 4). In addition to Ultima Genomics displaying a much higher mismatch error rate than Illumina for the *TMSB4X* gene, particularly concerning is the much higher rate of insertions and deletions (Figure 3a), with Ultima Genomics producing insertions and deletions up to 7 bp long. For context, liquid biopsies can benefit from accurate sequencing of 1 error per 10 million bases (Higgins et al. 2019), and the accuracy of Ultima Genomics for *TMSB4X* is worse than 80,000 indels per 10 million bases. These errors are not only present at and around the $(T)_8$ homopolymer in the *TMSB4X* gene, but are also apparent at and around much shorter homopolymers down to 3bp (Figure 2).

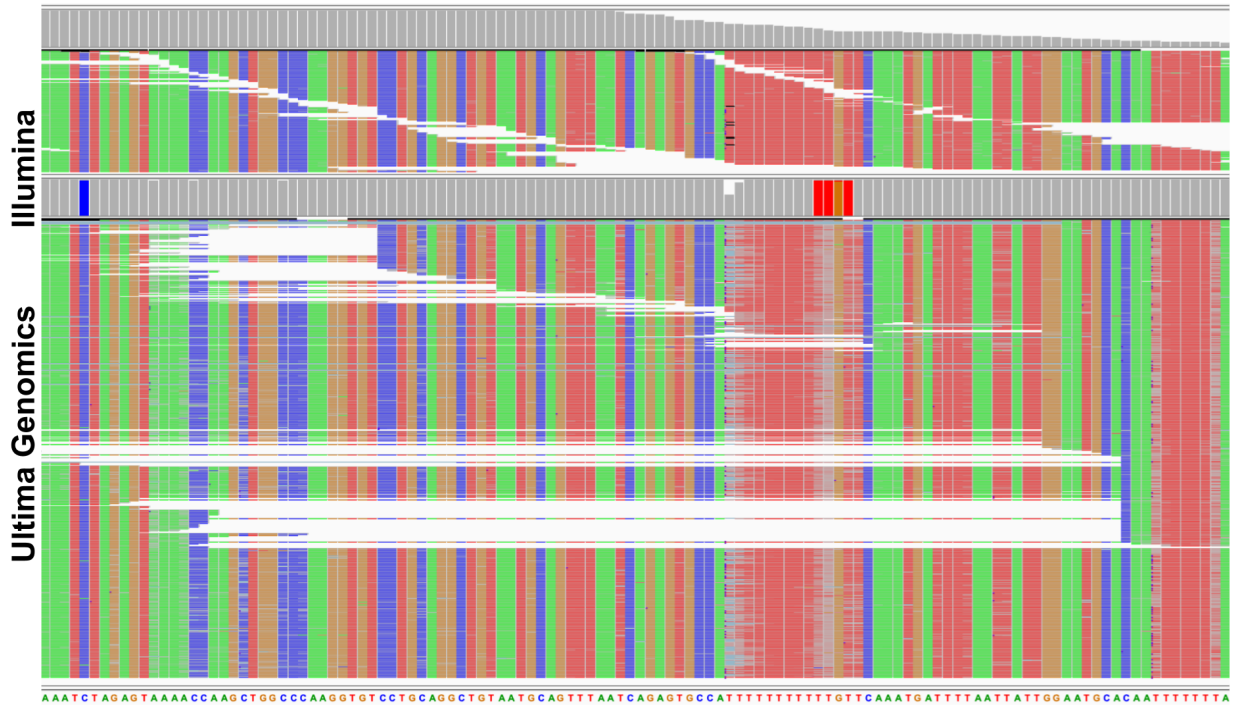

Figure 2: Illumina and Ultima Genomics reads aligning to the *TMSB4X* gene viewed on the IGV browser (Robinson et al. 2011). The gray streaks in the Ultima Genomics alignments reveal far more mismatches than in the Illumina alignments, especially around homopolymers. The coverage tracks above the alignments show that Ultima Genomics has a large number of indels at the $(T)_8$ homopolymer whereas Illumina has far fewer indels.

In light of the overall concordance in counts between Ultima Genomics and Illumina when reads were pseudoaligned (Supplementary Figure 3), we hypothesized that pseudoalignment, which is robust to errors in reads, could rescue some of the unaligned reads originating from the *TMSB4X* gene. We found that 0.1929% of Ultima Genomics reads pseudoaligned, which is 2.14

times higher than the fraction of aligned reads (0.0902%). In the case of Illumina reads, the higher base-call quality translated to little difference between the fraction of aligned (0.4118%) and pseudoaligned (0.4078%) reads. Thus, for *TMSB4X,* pseudoalignment recovered more than double the number of reads that could be assigned, demonstrating that it is a method well-suited to counting of error-prone Ultima Genomics reads for single-cell RNA-seq. Nevertheless, the Ultima Genomics error rate is so high that even pseudoalignment fails to rescue all the reads.

	Errors / 1000 bp	Avg. quality	Alignment (%)	Pseudoalignment (%)
Illumina	1.34	35.9	0.4118	0.4078
Ultima Genomics	13.4	32.6	0.0902	0.1929

Table 1: Summary *TMSB4X* gene alignment and pseudoalignment statistics for the Ultima Genomics and Illumina technologies.

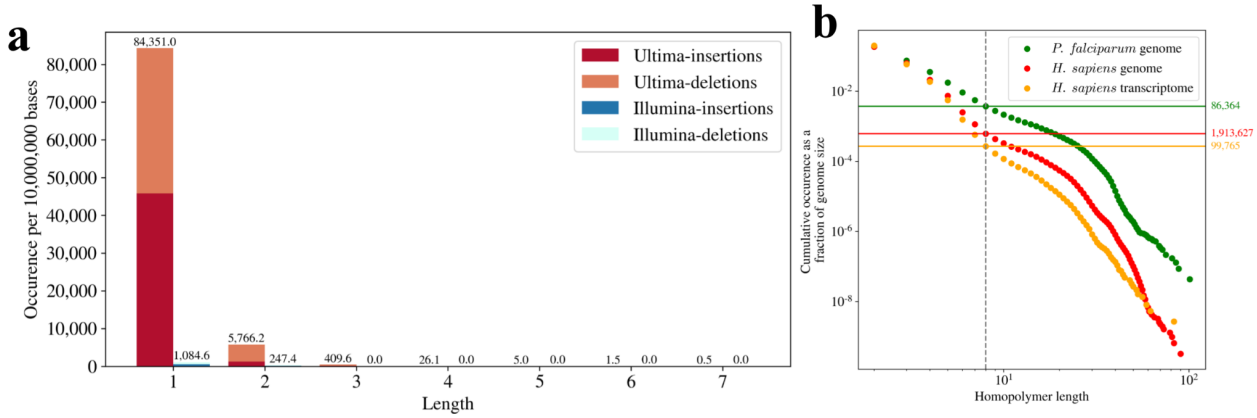

Figure 3: (a) The number of insertion and deletion errors in the *TMSB4X* gene for the Ultima Genomics and Illumina technologies. (b) The number of homopolymers (reported as a fraction of genome size) in the *H. sapiens* genome and transcriptome, as well as in the AT-rich *P. falciparum* genome.

Discussion

In the Ultima Genomics preprint (Almogy et al. 2022), the company discusses the challenges of sequencing homopolymers with their non-terminating chemistry, but touts an accuracy of 90% for homopolymers of length 8, and characterizes this as “good accuracy up to length 8-10 bases”. We find that in the Ultima Genomics single-cell RNA-seq reads, the homopolymer challenge presents as more acute than what may be imagined from summary statistics. The *TMSB4X* example demonstrates that Ultima Genomics displays poor performance not only in regions with long homopolymer runs, extreme GC content, or highly repetitive sequences. While the genome-wide accuracy of Ultima Genomics technology is likely to be better than for the *TMSB4X* gene, a comprehensive assessment of the error profiles was not possible because at the time of writing of this preprint, the data in (Almogy et al. 2022) has not been released (the

preprint states that “the data will be made available in the near future”). However, regardless of the genome-wide performance of Ultima Genomics technology, it is worth noting that improvement of sequencing technology is best guided by understanding worst-case performance.

Furthermore, the poor performance of Ultima Genomics on the *TMSB4X* gene is clinically relevant, because *TMSB4X* is possibly a biomarker for renal cancer as it has been shown to be predictive for survival (Morita and Hayashi 2018; “Expression of TMSB4X in Renal Cancer - The Human Protein Atlas” 2018; Uhlen et al. 2017). Moreover, *TMSB4X* is relevant for several clinical applications (Crockford et al. 2010).

While pseudoalignment of Ultima Genomics reads provides an improvement over standard alignment due to robustness to error, the improvement may not suffice for all genes, resulting in biases that may be difficult to compensate for during analysis. This problem may be addressable by improved alignment or pseudoalignment algorithms that are robust to homopolymers in the reference sequence. Some algorithmic ideas for this challenge have been developed in the context of other sequencing technologies that have poor performance in sequencing homopolymers; see, for example, (Feng et al. 2016) that was developed for Ion Torrent sequencing technology.

Unfortunately, in its current form, it seems that while Ultima Genomics sequencing may be promising in the long run, it is not currently a direct replacement for Illumina or BGI sequencers, both of which have much higher accuracy. The problems we found in sequencing regions with homopolymers, will be magnified in whole-genome sequencing applications. We counted the number of homopolymers in the human genome and found 1,913,627 homopolymers of at least length 8; the number of homopolymers can be even higher than that in other genomes (Figure 3b). Even for single-cell RNA-seq, the homopolymer problem has implications beyond just read assignment. (Simmons et al. 2022) observed that the poly(T) homopolymers in their reads corrupted the UMIs, and they therefore had to truncate the last 3 bases of each UMI (implemented by replacing the last 3 bases of each UMI with AAA). While this compensates for mismatches due to the poly(T) homopolymers, it has drawbacks in terms of accurate molecule counting (Melsted et al. 2021).

Hopefully Ultima Genomics will eventually find a way to reduce error rates so they can be competitive with existing technologies. The “many degrees of freedom” provided by their architecture design (Almogy et al. 2022) is possibly reason for optimism. The genomics community is already benefiting from cost reductions following the introduction of BGI sequencers (Drmanac et al. 2020; Hahn et al. 2021; LeMieux 2020), and will benefit even further if Ultima Genomics can similarly reduce sequencing costs.

Methods

The code to reproduce all the figures and results in the preprint is available at https://github.com/pachterlab/BP_2022 and provides a complete description of the methods.

Data Availability

All data used in this preprint is available on GEO under accession GSM5917802.

Software

- anndata 0.7.6 (Virshup et al. 2021)
- bustools 0.40.0 (Melsted et al. 2021; Melsted, Ntranos, and Pachter 2019)
- IGV 2.13.0 (Robinson et al. 2011)
- kallisto 0.48.0 (Bray et al. 2016)
- kb-python 0.27.2 (Melsted et al. 2021)
- ffq 0.2.1 (Gálvez-Merchán et al. 2022)
- gget 0.1.1 (Luebbert and Pachter 2022)
- HISAT2 2.2.1 (Kim et al. 2019)
- htslib 1.10 (Bonfield et al. 2021)
- Matplotlib 3.5.1 (Hunter 2007)
- Numpy 1.21.6 (Harris et al. 2020)
- Pandas 1.3.5 (Mckinney 2011)
- samtools 1.10 (Heng Li et al. 2009)
- seqkit 0.12.0 (Shen et al. 2016)

References

Almogy, Gilad, Mark Pratt, Florian Oberstrass, Linda Lee, Dan Mazur, Nate Beckett, Omer Barad, et al. 2022. “Cost-Efficient Whole Genome-Sequencing Using Novel Mostly Natural Sequencing-by-Synthesis Chemistry and Open Fluidics Platform.” *bioRxiv*. https://doi.org/10.1101/2022.05.29.493900.

Bonfield, James K., John Marshall, Petr Danecek, Heng Li, Valeriu Ohan, Andrew Whitwham, Thomas Keane, and Robert M. Davies. 2021. “HTSlib: C Library for Reading/writing High-Throughput Sequencing Data.” *GigaScience* 10 (2). https://doi.org/10.1093/gigascience/giab007.

Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. “Near-Optimal Probabilistic RNA-Seq Quantification.” *Nature Biotechnology* 34 (5): 525–27.

Crockford, David, Nabila Turjman, Christian Allan, and Janet Angel. 2010. “Thymosin beta4: Structure, Function, and Biological Properties Supporting Current and Future Clinical Applications.” *Annals of the New York Academy of Sciences* 1194 (April): 179–89.

Drmanac, Snezana, Matthew Callow, Linsu Chen, Ping Zhou, Leon Eckhardt, Chongjun Xu, Meihua Gong, et al. 2020. “CoolMPS™: Advanced Massively Parallel Sequencing Using Antibodies Specific to Each Natural Nucleobase.” *bioRxiv*. https://doi.org/10.1101/2020.02.19.953307.

“Expression of TMSB4X in Renal Cancer - The Human Protein Atlas.” 2018. February 27, 2018. https://www.proteinatlas.org/ENSG00000205542-TMSB4X/pathology/renal+cancer/KIRC.

Feng, Weixing, Sen Zhao, Dingkai Xue, Fengfei Song, Ziwei Li, Duojiao Chen, Bo He, Yangyang Hao, Yadong Wang, and Yunlong Liu. 2016. “Improving Alignment Accuracy on Homopolymer Regions for Semiconductor-Based Sequencing Technologies.” *BMC Genomics* 17 Suppl 7 (August): 521.

Gálvez-Merchán, Ángel, Kyung Hoi (joseph) Min, Lior Pachter, and A. Sina Booeshaghi. 2022. “Metadata Retrieval from Sequence Databases with Ffq.” *bioRxiv*. https://doi.org/10.1101/2022.05.18.492548.

Goodwin, Sara, John D. McPherson, and W. Richard McCombie. 2016. “Coming of Age: Ten Years of next-Generation Sequencing Technologies.” *Nature Reviews. Genetics* 17 (6): 333–51.

Hahn, Oliver, Tobias Fehlmann, Hui Zhang, Christy N. Munson, Ryan T. Vest, Adam Borcherding, Sophie Liu, et al. 2021. “CoolMPS for Robust Sequencing of Single-Nuclear RNAs Captured by Droplet-Based Method.” *Nucleic Acids Research* 49 (2): e11.

Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. “Array Programming with NumPy.” *Nature* 585 (7825): 357–62.

Higgins, Jacob, Gabriel Pratt, Charles C. Valentine, Lindsey N. Williams, and Jesse J. Salk. 2019. “Redefining ‘Gold Standard’: Ultra-Sensitive Characterization of Commercial DNA Standards with Duplex Sequencing.” *Blood* 134 (November): 2093.

Hunter, John D. 2007. “Matplotlib: A 2D Graphics Environment.” *Computing in Science & Engineering* 9 (3): 90–95.

Kim, Daehwan, Joseph M. Paggi, Chanhee Park, Christopher Bennett, and Steven L. Salzberg. 2019. “Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-Genotype.” *Nature Biotechnology* 37 (8): 907–15.

Kris A. Wetterstrand, M. S. 2019. “DNA Sequencing Costs: Data.” Genome.gov. NHGRI. March 13, 2019. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data.

LeMieux, Julianna. 2019. “All Aboard The Genome Express.” *Genetic Engineering & Biotechnology News* 39 (1): 34, 35, 38, 40–41.

———. 2020. “MGI Delivers the $100 Genome at AGBT Conference.” Genetic Engineering and

Biotechnology News. February 6, 2020. https://www.genengnews.com/news/mgi-delivers-the-100-genome-at-agbt-conference/.

Li, Haichao, Kun Wu, Chenchen Ruan, Jiao Pan, Yujin Wang, and Hongan Long. 2019. "Cost-Reduction Strategies in Massive Genomics Experiments." *Marine Life Science & Technology* 1 (1): 15–21.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79.

Luebbert, Laura, and Lior Pachter. 2022. "Efficient Querying of Genomic Reference Databases with Gget." *bioRxiv*. https://doi.org/10.1101/2022.05.17.492392.

Mckinney, Wes. 2011. "Pandas: A Foundational Python Library for Data Analysis and Statistics." 2011. https://www.dlr.de/sc/portaldata/15/resources/dokumente/pyhpc2011/submissions/pyhpc2011_submission_9.pdf.

Melsted, Páll, A. Sina Booeshaghi, Lauren Liu, Fan Gao, Lambda Lu, Kyung Hoi Joseph Min, Eduardo da Veiga Beltrame, Kristján Eldjárn Hjörleifsson, Jase Gehring, and Lior Pachter. 2021. "Modular, Efficient and Constant-Memory Single-Cell RNA-Seq Preprocessing." *Nature Biotechnology* 39 (7): 813–18.

Melsted, Páll, Vasilis Ntranos, and Lior Pachter. 2019. "The Barcode, UMI, Set Format and BUStools." *Bioinformatics* 35 (21): 4472–73.

Morita, Tsuyoshi, and Ken 'ichiro Hayashi. 2018. "Tumor Progression Is Mediated by Thymosin-β4 through a TGFβ/MRTF Signaling Axis." *Molecular Cancer Research: MCR* 16 (5): 880–93.

"Nebula Genomics - 30x Whole Genome Sequencing - DNA Testing." 2022. May 25, 2022. https://web.archive.org/web/20220525222324/https://nebula.org/whole-genome-sequencing-dna-test/.

Robinson, James T., Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. 2011. "Integrative Genomics Viewer." *Nature Biotechnology* 29 (1): 24–26.

Schwarze, Katharina, James Buchanan, Jilles M. Fermont, Helene Dreau, Mark W. Tilley, John M. Taylor, Pavlos Antoniou, et al. 2020. "The Complete Costs of Genome Sequencing: A Microcosting Study in Cancer and Rare Diseases from a Single Center in the United Kingdom." *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 22 (1): 85–94.

Shen, Wei, Shuai Le, Yan Li, and Fuquan Hu. 2016. "SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation." *PloS One* 11 (10): e0163962.

Simmons, Sean K., Gila Lithwick-Yanai, Xian Adiconis, Florian Oberstrass, Nika Iremadze, Kathryn Geiger-Schuller, Pratiksha I. Thakore, et al. 2022. "Single Cell RNA-Seq by Mostly-Natural Sequencing by Synthesis." *bioRxiv*. https://doi.org/10.1101/2022.05.29.493705.

Uhlen, Mathias, Cheng Zhang, Sunjae Lee, Evelina Sjöstedt, Linn Fagerberg, Gholamreza Bidkhori, Rui Benfeitas, et al. 2017. "A Pathology Atlas of the Human Cancer Transcriptome." *Science* 357 (6352). https://doi.org/10.1126/science.aan2507.

"Ultima Genomics Delivers the $100 Genome." 2022. May 31, 2022. https://www.ultimagenomics.com/blog/ultima-genomics-delivers-usd100-genome.

Virshup, Isaac, Sergei Rybakov, Fabian J. Theis, Philipp Angerer, and F. Alexander Wolf. 2021. "Anndata: Annotated Data." *bioRxiv*. https://doi.org/10.1101/2021.12.16.473007.