GenoMiX: Accelerated Simultaneous Analysis of Human Genomics, Microbiome Metagenomics, and Viral Sequences

Tianqi Zhang*, Antonio González†, Niema Moshiri*, Rob Knight†, Tajana Rosing*
*CSE Department, UC San Diego, La Jolla, CA 92093, USA
{tiz014,a1moshir, tajana}@ucsd.edu
†Bioengineering Department, UC San Diego, La Jolla, CA 92093, USA
antgonza@gmail.com, rknight@ucsd.edu

2023 IEEE Biomedical Circuits and Systems Conference (BioCAS) | 979-8-3503-0026-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/BioCAS58349.2023.10388531

***Abstract*—Personalized medicine tailors treatments based on the individual characteristics of each patient. Recent work targets comprehensive analysis of patient samples that contain human, viral, and bacterial genomes. Novel technologies like Simul-seq enable simultaneous analysis of such samples. However, existing workflows are slow and disjoint. In this paper, we present FPGA accelerated genomics infrastructure, GenoMiX , that supports multiple real-world analysis pipelines used in personalized medicine, including phylogenetic assignment for viral pathogens, variant calling that is key for cancer genomics, and microbiome metagenome analysis. We integrate our workflow with Qiita, an open-source framework for managing and analyzing multi-omics datasets. GenoMiX is not only up to $30\times$ faster than comparable CPU-based tools, but it also addresseNs the challenges associated with handling reference databases of varying sizes, encompassing viruses, human genomes, and microbiomes. Our optimized infrastructure was a key component enabling success of the award-winning UCSD's "Return to Learn" program during the COVID-19 public health emergency in San Diego County.**

I. Introduction

Personalized medicine tailors treatments to the individual characteristics of each patient. Protocols such as Simul-seq [1] have enabled the comprehensive collection and analysis of patient samples that contain not just human genomes but also bacterial and viral samples. This is critical to understanding how human health is affected not only by diseases such as cancer but also by native bacteria and outside pathogens that interact with our bodies and the medicines we take. Recent work in personalized medicine seeks to understand the interaction between these various aspects. However, state of the art tools such as [2]–[6] are too slow to address this need.

Many accelerators have been proposed to date for various components of the analysis pipeline, such as those based on processing in memory (PIM) [7], [8], near-data computing [9], ASICs [10], and FPGAs [11]–[15]. Most of these are not deployed in the actual analysis pipelines since they a) only address a few components of the pipeline, b) their accuracy may not be as good as CPU based tools, and c) they are more difficult to use and less compatible with the existing tools. For example, [7], [8] only accelerate filtering and pairwise alignment, respectively; [9] accelerates seeding and pre-alignment filtering but neglects computationally bounded Levenshtein distance computing; [11] runs from seeding to pairwise alignment but doesn't support backtrace to get the compact idiosyncratic gapped alignment report (CIGAR); [12] can get Bowtie2 [4]-like alignment accuracy, but cannot get the mapping quality (MAPQ) needed by downstream analysis.

In this work, we design a unified and fast tool, GenoMix, that accelerates viral sequence analysis, variant calling in human genomics, and bacterial metagenomic comparative analysis. GenoMiX uses Qiita [16], an open-sourced framework for managing and analyzing multi-omics datasets via a user-friendly webpage, as a front end. As shown in Figure 1, the three analysis pipelines share a common backbone: short read alignment and sequence trimming, which we accelerate using FPGA. In order to provide end-to-end support to medical practitioners, we integrate downstream analysis tools into GenoMiX such as DeepVariant [17] for variant calling, Pangolin [18] for viral analysis, and Unifrac [19] that is targeted at microbiome. The phylogenetic assignment generates the viral concise sequence and classifies it. Variant calling detects whether the genome's mismatches or indels are variants or sequencing machine-induced errors. In metagenome community ecology, the samples, comprising multiple bacterial species, use the unifrac distance as a metric, to discover and compare a variety of microbiome communities. Run as a plugin of Qiita [16], GenoMiX offers users the flexibility to run all three pipelines simultaneously on the same dataset or only specific pipelines they require on separate datasets. This ensures an adaptable tool suitable for a variety of use cases, from multi-omics analysis to single-dataset investigations. To summarize, our system is:

- The first multi-in-one tool that integrates and accelerates three different analysis pipelines by up to $30\times$ speedup at state of the art accuracy. Each U55c FPGA running GenoMiX provides the same computation power of up to 475 CPUs.
- Designed to efficiently handle reference databases of various sizes involving viruses, human, and microbiome, with as many as billions of base pairs;
- Integrated into state of the art Qiita [16] genomic study management platform. The platform has been used over the last three years in San Diego County as a part of UCSD's award winning "Return to Learn" program [20] during the COVID-19 public health emergency.

This work was supported in part by CRISP, center in JUMP 1.0, PRISM and CoCoSys, centers in JUMP 2.0, an SRC program sponsored by DARPA, and NSF grants number 2003279, 1911095, 1826967, 2100237, 2112167, 2052809, and 2112665.

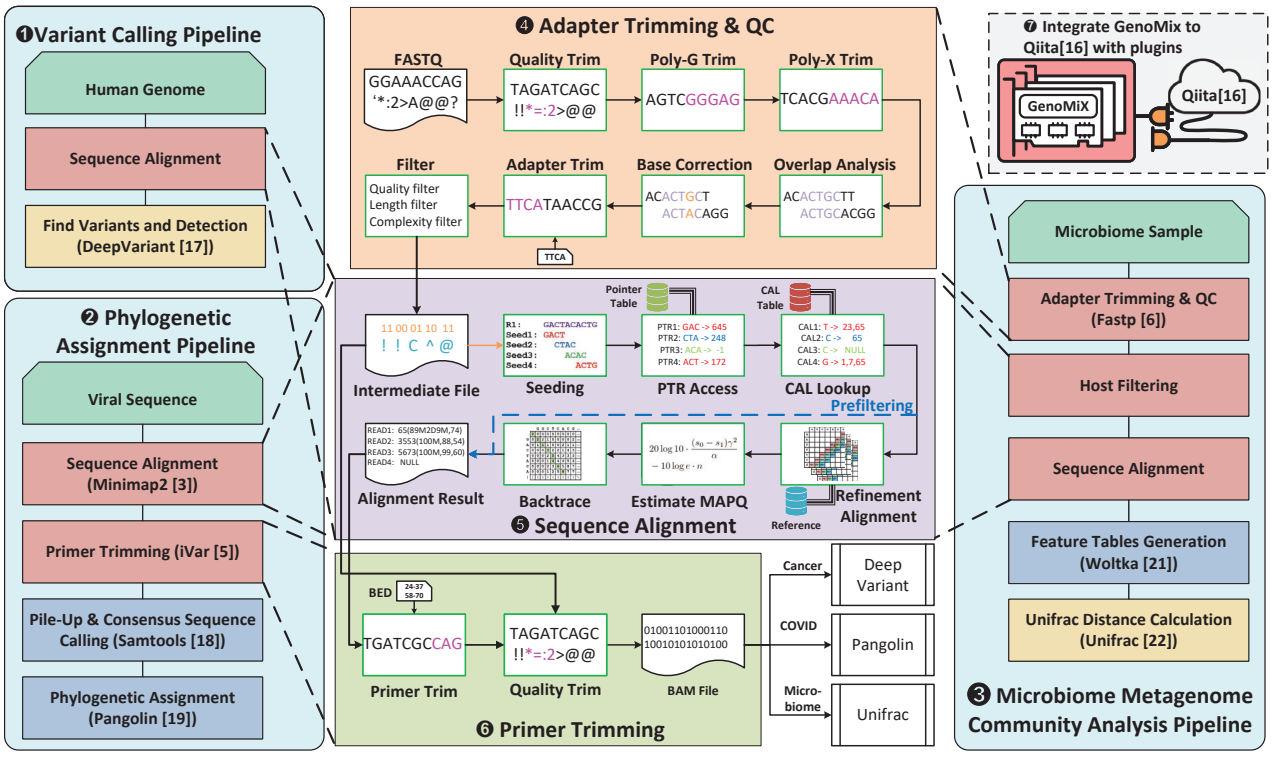

Fig. 1. Overview of GenoMiX . GenoMiX is made up with the component ❹-❼ to accelerate the pipeline ❶-❸. Standard tools chain includes aligner [2]–[4], trimmer [5], [6], and downstream modules [17]–[19], [21], [22].

II. GenoMiX Design and Implementation

Figure 1 provides an overview of GenoMiX which runs as a plugin in Qiita [16], as shown in ❼. GenoMiX accelerates three pipelines (❶-❸) with four components (❹-❼). The webpage frontend of Qiita [16] provides GenoMiX a unified and user-friendly interface for the three analysis pipelines that are represented in blue boxes: ❶ runs variant calling using DeepVariant [17] on GPU; ❷ showcases viral phylogenetic assignment pipeline utilized in *UCSD's "Return to Learn" Program* [20]; ❸ executes Unifrac [19] which computes dissimilarity between microbial communities based on their evolutionary relationships. The three pipelines have common components represented in red boxes: sequence alignment (❺), adapter trimming and quality control (❹), and primer trimming (❻). We profiled each of the three pipelines and found that trimming and alignment, when run on CPU, is by far the slowest, with $67\% - 98\%$ overhead, so GenoMiX accelerates all of these on FPGA obtaining up to $30\times$ speedup. As shown in ❹ adapter and ❻ primer trimming, which are detailed in Section II-A, remove the low-quality regions and artifact segments. The alignment component of GenoMiX , illustrated in ❺ and discussed in Section II-B, uses seeding and extension-based method with a 2-stage index table. It can not only identify mapping locations but also generates CIGAR strings and accurately estimates MAPQ values. The structure of the index table is discussed in II-C. After alignment and trimming are completed, each of the three pipelines runs their individual downstream tools: Pangolin [18] on CPU, DeepVariant [17], and Unifrac [19] on GPU, to obtain the final results that are then displayed using Qiita [16] front end.

A. GenoMiX Trimming and Quality Control (QC)

GenoMiX trimming and quality control stages enhance read alignment accuracy. This involves removing artifact DNA sequences, like adaptors and primers, and discarding low-quality reads and segments similar to Fastp [6] and iVar [5] which run on CPU. GenoMiX utilizes a sliding window to detect and discard adaptor parts when mismatches fall below a specific threshold for **adapter trimming**. It calculates moving average phred scores, which determine the sequencing quality of each base, to remove low-quality areas. This **quality trimming** can be set to one of three modes: *cut_front*, *cut_tail*, and *cut_right*. **PolyX/G trimming** eliminates consecutive identical bases at the read tail that may lack relevance. For example, polyG may occur in tails from NextSeq platforms as G indicates no signal. When the input data is paired, GenoMiX can perform **base correction** to identify overlap regions and correct mismatched base pairs if one base's phred score is much higher than the other. GenoMiX also has many **filtering** options, such as discard reads with too many undefined bases, low average phred scores, insufficient length, or low complexity. Primer sequences are known segments used to amplify the target DNA fragment. Unlike adapter trimming, which requires base-by-base matching but only involves tens of sequences, **primer trimming** deals with hundreds of artifact primer sequences introduced by multiplex amplicon sequencing. GenoMiX handles primer trimming in similar way to adapter trimming.

B. GenoMiX Alignment

In DNA sequencing, short reads are often mismatched with the reference. GenoMiX uses seeding-extension-based

method to handle mismatches or indels. The seeding stage identifies several k-mers from the sequence but may indicate a large number of less optimal candidate locations. GenoMiX prefilters those to improve the alignment quality.

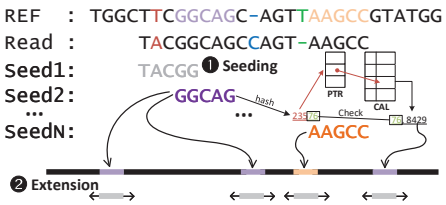

Fig. 2. Example of Seeding-and-extension based sequence alignment

Seeding: Seeds are extracted by intercepting K-mers with interval l. Since read direction isn't predetermined, GenoMiX stores entries of seeds corresponding to the original and its reverse complement in the same bucket to reduce the number of random accesses. Hash function converts a K-mer to a $2K+1$ bits integer, where the lower $2K$ bits are smaller component of the 2bits-1bp representation of the seed and its reverse complementary, and the most significant bit indicates whether the original seed or the reverse complementary seed is smaller. GenoMiX then looks up the candidate locations of seed in a 2-stage hash table. The hashed value is split into two parts: H_{ptr} and H_{cal}, which corresponds to *PTR* table and *CAL* table respectively. The PTR table is a conflict-free mapping: the H_{ptr}-th item of the PTR table points to the entry of the CAL table storing this seed. The CAL table stores H_{cal}-$position$ pair, where the key is the rest of hashing value of the seed, and the value is the location of this seed in the reference.

Prefiltering speeds up and improves the quality of sequence alignment by removing unlikely matches. GenoMiX can identify the most likely locations where the read may align by subtracting the offset of each seed from the beginning of the read from the candidate positions of the seed in the reference genome. To account for the possibility of small indels, positions that are adjacent within a certain tolerance are treated as the same. The prefiltering will discard the positions only a few seeds point to, making it can focus on more promising candidate sequences. In addition to filtering out the less-promising candidate positions, GenoMiX 's prefilter will also calculate the Hamming distance to verify if the read is able to align to the reference perfectly with only a few mismatches and no indels. The number of mismatch tolerance is a parameter making the trade-off between speed and the possibility to find the best alignment.

Extension performs alignment using the Smith-Waterman and Needleman-Wunsch algorithms and produces CIGAR as output. Depending on the downstream applications, global, local, or semi-global alignments may be used. GenoMiX estimates the mapping quality MAPQ similar to BWA-MEM [2].

C. GenoMiX Table Indexing

GenoMiX limits CAL row size to improve sequence alignment efficiency. Each CAL row has a H_{cal} cell and a seed position cell. The seed position is a 32-bit integer to fit the whole human genome. The bucket size is also limited to avoid the CAL table search bottleneck. Bucket size is limited to 32 rows, and if a bucket exceeds this limit, its H_{cal} are sorted by the number of positions they point to, and half of the positions are randomly discarded. This is repeated until the bucket size is less than 32. GenoMiX table structure is optimized based on the reference length. The default length H_{ptr} and H_{cal} is chosen for the reference as large as the human genome. When the reference genome is smaller, such as with COVID-19, there is only a small number of CAL entries that share the same H_{ptr}. In GenoMiX , the length of H_{ptr} is reduced to decrease the PTR table size, while the length of H_{CAL} is extended to maintain a constant seed length.

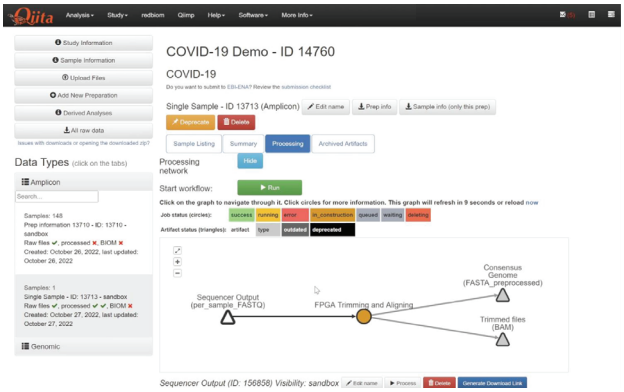

Fig. 3. Screenshot of Qiita webpage with COVID-19 analysis task

III. Evaluation and Results

Experimental Setup: Qiita [16] was used to drive the GenoMiX as a front end. Figure 3 shows the screenshot of Qiita with GenoMiX when running COVID-19 analysis, but we can run all three types of genomics data analysis - human, bacteria, and virus. The acceleration kernels are written in C++ and compiled by Xilinx Vitis HLS to run on Xilinx FPGA U55c [23] which has 16GB high bandwidth memory (HBM) split into 32 pseudo channels interconnected by 8 AXI switch boxes. GenoMiX optimizes the architecture based on the table size to leverage intra and inter box connectivity. For example, viral reference is small and thus we store the PTR table in a single HBM channel and the CAL table in internal SRAM, with multiple table copies to enhance throughput. The CAL tables for bacterial reference genome databases, in contrast, are typically so large that they cannot fit in a single FPGA, so GenoMiX divides them into multiple 2GB subtables to ensure locality within the switch boxes. Our FPGA uses AMD EPYC 7F72 CPU with 128GB RAM as the host.

Datasets: Our design was evaluated using real-world datasets, including COVID sequencing data obtained from the *UCSD's "Return to Learn" Program* [20], human whole genome sequencing data of NA12878 from the *Genome in a Bottle Project*, and microbiome metagenomic sequencing data from the *EMP500 Project* [24]. The size of the reference genome varied across the datasets, ranging from a single virus genome to a comprehensive microbiome reference database consisting of over 10,000 bacterial and archaeal genomes. Table I provides more details on them.

A. Varaint Calling

We tested the accuracy of GenoMiX on the cancer variant detection task using sequencing data of a female subject NA12878 with Platinum Genomes small variant truthset [29].

TABLE I
REFERNECE GENOME

Description	Name	Total Length	PTR Table	CAL Table	Ref Table
COVID-19	NC_045512.2 [25]	30Kbp	256MB	233KB	7.5KB
Chromosome 21	NC_000021.8 [26]	48Mbp	4GB	251MB	12MB
Whole human genome	GRCh38 [27]	3.1Gbp	4GB	16GB	786MB
Microbiome database	WoL [28]	33Gbp	40GB	195GB	8.2GB

GenoMiX has 12× higher throughput when running on a single FPGA as compared with the BWA-MEM [2] running on 16 cores. As shown in Figure 4, the difference of end-to-end results between using GenoMiX and BWA-MEM [2] is small. True positives refer to the number of variants that the downstream application, DeepVariant [17], accurately classifies. Conversely, errors refer to the number of variants that DeepVariant is able to detect the variants but misclassify them.

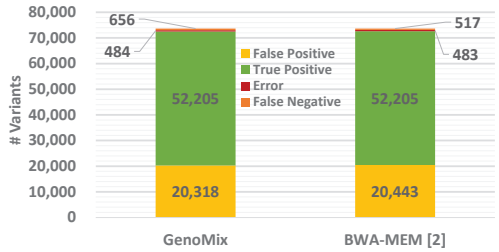

Fig. 4. Variant calling accuracy

B. Phylogenetic Assignment

We tested the COVID-19 phylogenetic assignment task with COVID sequencing data obtained from the *UCSD's "Return to Learn Program"* [20]. Figure 5 shows the runtime of GenoMiX compared to the state of the art BWA-MEM [2]. The baseline is running with 16 cores. GenoMiX is 3× faster than CPU for alignment and 30× faster for trimming. Pile-up and consensus sequence generation stages are pipeline bottlenecks after GenoMiX acceleration on FPGA. Results of the accelerated pipeline are very close in terms of accuracy to the original toolchain [30], with the difference between the consensus sequences generated of only 0.04%. Most of the difference is caused by the insufficient coverage of sequencing, and would not influence the end-to-end results. GenoMiX is able to get exactly the same phylogenetic assignment results as the baseline state of the art tool, Minimap2 [3].

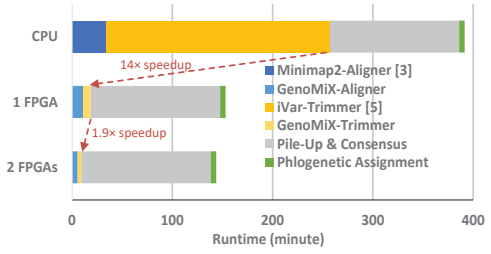

Fig. 5. End-to-end runtime of phylogenetic assignment

C. Microbiome Metagenomics Comparative Analysis

We tested the microbiome metagenomics comparative analysis workloads with common use EMP500 [24] dataset and WoL [28] reference genome database. As shown in Figure 6, alignment is the slowest component of the whole pipeline, due to the large size of the WoL [28] database. Bowtie2 [4], a state of the art tool often used for alignment of microbial data due to its high accuracy, uses 44GB of memory when running with WoL [28] database on a CPU. The challenge of GenoMiX aligner is the scale of the microbiome reference genome database: it may contain over 10K bacterial and archaeal genomes which results in over 190 GB of index tables. However, a single U55c only has 16 GB of onboard memory, which is insufficient to store the index tables. To solve this problem, we divided the original reference into 10 segments and built 10 sets of tables. At least 2 FPGAs are required to finish the alignment. With 3 FPGAs, which contain as much memory as Bowtie2 consumes, we can achieve 4× speedup with comparable accuracy. GenoMiX adapter trimmer and QC are 15× faster vs Fastp [6] on CPU. Woltka [22], which is used for feature table generation, has the highest throughput, as it is relatively simple. Even if it is run on CPU can lead the throughput. Unifrac is not shown as it runs in a few seconds on TB-size short-read datasets. We would need 679 CPU cores or only 29 FPGAs in order for alignment and trimming to match the throughput of Woltka [22].

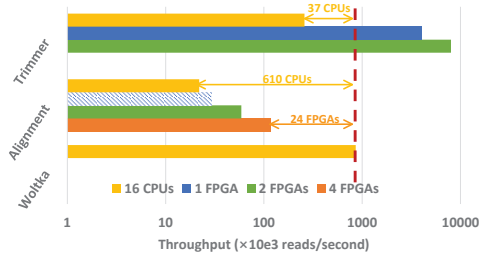

Fig. 6. Microbiome metagenomics comparative analysis throughput. GenoMiX is run with 1, 2, and 4 FPGAs while the other tools are run with 16 CPUs.

IV. CONCLUSIONS AND FUTURE WORK

In this paper, we presented an FPGA accelerated genomics infrastructure, GenoMiX , that supports multiple real-world analysis pipelines used in personalized medicine, including the phylogenetic assignment for viral pathogens, variant calling that is key for cancer genomics, and microbiome metagenome analysis. GenoMiX has been integrated with Qiita [16], an open-source framework for managing and analyzing multi-omics datasets. It is up to 30× faster than comparable CPU-based tools. The accelerated pipeline has been deployed for COVID-19 analysis in San Diego County during the last three years, as a part of UCSD's "Return to Learn" program [20]. It shows comparable end-to-end results with the ones generated by the mainstream tools.

As we look toward the future, we will investigate its applicability in other fields of biology and medicine, such as proteomics and transcriptomic. Besides, we aim to make GenoMiX even more modular, enabling researchers to integrate new pipeline components as the field of genomics advances.

V. ACKNOLEDGEMENTS

We thank Jeff DeReus for assisting with dataset migration. Our appreciation goes to Ameen Akel for valuable insights into memory profiling, and to Grant Cheng, Annie Liu, and Kai Lee for their help with Qiita platform integration.

REFERENCES

[1] J. A. Reuter, D. V. Spacek, R. K. Pai *et al.*, "Simul-seq: combined dna and rna sequencing for whole-genome and transcriptome profiling," *Nature methods*, vol. 13, no. 11, pp. 953–958, 2016.

[2] H. Li, "Aligning sequence reads, clone sequences and assembly contigs with bwa-mem," 2013.

[3] ——, "Minimap2: pairwise alignment for nucleotide sequences," *Bioinformatics*, vol. 34, no. 18, pp. 3094–3100, 2018.

[4] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with bowtie 2," *Nature methods*, vol. 9, no. 4, pp. 357–359, 2012.

[5] N. D. Grubaugh, K. Gangavarapu, J. Quick *et al.*, "An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using primalseq and ivar," *Genome biology*, vol. 20, no. 1, pp. 1–19, 2019.

[6] S. Chen, Y. Zhou, Y. Chen *et al.*, "fastp: an ultra-fast all-in-one fastq preprocessor," *Bioinformatics*, vol. 34, no. 17, pp. i884–i890, 2018.

[7] F. Hameed, A. A. Khan, and J. Castrillon, "Alpha: A novel algorithm-hardware co-design for accelerating dna seed location filtering," *IEEE Transactions on Emerging Topics in Computing*, vol. 10, no. 3, pp. 1464–1475, 2021.

[8] W. Xu, S. Gupta, N. Moshiri *et al.*, "Rapidx: High-performance reram processing in-memory accelerator for sequence alignment," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2023.

[9] W. Huangfu, K. T. Malladi, A. Chang *et al.*, "Beacon: Scalable near-data-processing accelerators for genome analysis near memory pool with the cxl support," in *2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2022, pp. 727–743.

[10] Y. Turakhia, G. Bejerano, and W. J. Dally, "Darwin: A genomics co-processor provides up to 15,000 x acceleration on long read assembly," *ACM SIGPLAN Notices*, vol. 53, no. 2, pp. 199–213, 2018.

[11] B. Khaleghi, T. Zhang, C. Martino *et al.*, "Salient: Ultra-fast fpga-based short read alignment," in *2022 International Conference on Field-Programmable Technology (ICFPT)*, 2022, pp. 1–10.

[12] H.-C. Ng, S. Liu, I. Coleman *et al.*, "Acceleration of short read alignment with runtime reconfiguration," in *2020 International Conference on Field-Programmable Technology (ICFPT)*, 2020, pp. 256–262.

[13] J. Arram, T. Kaplan, W. Luk *et al.*, "Leveraging fpgas for accelerating short read alignment," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 14, no. 3, pp. 668–677, 2017.

[14] B. Khaleghi, T. Zhang, N. Shao *et al.*, "Fast: Fpga-based acceleration of genomic sequence trimming," in *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2022, pp. 510–514.

[15] "Illumina dragen secondary analysis," https://www.illumina.com/products/by-type/informatics-products/dragen-secondary-analysis.html.

[16] "Github repo of qiita," https://github.com/qiita-spots/qiita.

[17] R. Poplin, P.-C. Chang, D. Alexander *et al.*, "A universal snp and small-indel variant caller using deep neural networks," *Nature biotechnology*, vol. 36, no. 10, pp. 983–987, 2018.

[18] Á. O'Toole, E. Scher, A. Underwood *et al.*, "Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool," *Virus evolution*, vol. 7, no. 2, p. veab064, 2021.

[19] C. Lozupone, M. E. Lladser, D. Knights *et al.*, "Unifrac: an effective distance metric for microbial community comparison," *The ISME journal*, vol. 5, no. 2, pp. 169–172, 2011.

[20] UCSD, "Return to learn program," https://returntolearn.ucsd.edu, 2020.

[21] H. Li, B. Handsaker, A. Wysoker *et al.*, "The sequence alignment/map format and samtools," *bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[22] Q. Zhu, S. Huang, A. Gonzalez *et al.*, "Phylogeny-aware analysis of metagenome community ecology based on matched reference genomes while bypassing taxonomy," *mSystems*, vol. 7, no. 2, pp. e00 167–22, 2022.

[23] "Alveo u55c high performance compute card," https://www.xilinx.com/products/boards-and-kits/alveo/u55c.html.

[24] J. P. Shaffer, L.-F. Nothias, L. R. Thompson *et al.*, "Multi-omics profiling of earth's biomes reveals that microbial and metabolite composition are shaped by the environment," *bioRxiv*, 2021. [Online]. Available: https://www.biorxiv.org/content/early/2021/06/06/2021.06.04.446988

[25] "Severe acute respiratory syndrome coronavirus 2 isolate wuhan-hu-1, complete genome," https://www.ncbi.nlm.nih.gov/nuccore/1798174254.

[26] "Homo sapiens chromosome 21, grch37.p13 primary assembly," https://www.ncbi.nlm.nih.gov/nuccore/NC_000021.8.

[27] "Genome reference consortium human build 38," https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/.

[28] Q. Zhu, U. Mai, W. Pfeiffer *et al.*, "Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains bacteria and archaea," *Nature communications*, vol. 10, no. 1, p. 5477, 2019.

[29] M. A. Eberle, E. Fritzilas, P. Krusche *et al.*, "A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree," *Genome research*, vol. 27, no. 1, pp. 157–164, 2017.

[30] N. Moshiri, K. M. Fisch, A. Birmingham *et al.*, "The vireflow pipeline enables user friendly large scale viral consensus genome reconstruction," *Scientific reports*, vol. 12, no. 1, p. 5077, 2022.