Sairam Behera

Sairam.behera@bcm.edu @srbehera11
https://srbehera.github.io/
https://github.com/srbehera
https://www.linkedin.com/in/srbehera/

Professional Summary

Currently, I am working as a post-doctoral research fellow in Dr. Sedlazeck's group at Baylor College of Medicine. My current research focuses on genetic variations i.e., structural variants (SV) and small variants (SNP and small indels), and their implications on biomedical research and clinical analysis. I graduated from the The University of Nebraska Lincoln with a doctoral degree in computer science. My research during my Ph.D. focused on bioinformatics algorithms e.g. development of tools and analysis of transcriptome assembly using short Illumina sequencing data as well as long-read sequencing data from third-generation technologies e.g. PacBio and Nanopore. I have almost 8+ years of experience in computational biology research that focuses on large-scale sequencing data analysis, transcriptome assembly, RNA-Seq analysis and algorithm development. I have also worked for 5 years as a software developer in companies such as Cognizant, Credit-Suisse, AT&T in India, Switzerland, and USA.

Research Interests

Graph Theory, Algorithm Optimization, Bioinformatics, NGS data analysis, Structural Variation

Employment History

Dec 2020 –	**Post-doc Research Fellow** at Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas.
Aug 2016 – Dec 2020	**Graduate Research Assistant** at Moriyama Lab, School of Biological Sciences, University of Nebraska-Lincoln.
Aug 2012 – May 2016	**Graduate Teaching Assistant** at Department of Computer Science and Engineering, University of Nebraska-Lincoln.
Mar 2012 – Aug 2012	**Software Developer** at Tech Mahindra, USA.
May 2010 – Aug 2011	**Graduate Research Assistant** at Department of Computer Science and Engineering, University of Texas at Dallas.
May 2005 – Jun 2009	**Software Developer** at Cognizant Technology Solutions, India.

Skills

Bioinformatics	NGS, Nanopore/PacBio, RNA-Seq, Assembly, Differential Gene Expression, Structural variations
Coding	C, C++, Java, Python, R, Perl
Databases	MySQL, Oracle
Web Dev	HTML, CSS, JavaScript

Education, Training, and Experience

2020 – present **Post-doc. Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX**
Trained in the structural and small variant analysis using short and long sequencing data for comparative genomics and biomedical research; mentored by Dr. Fritz Sedlazeck

Project 1: Fixing GRCh38 reference errors

- Identification of likely artifacts in GTEx, gnomAD, 1000 Genomes Project, and other important genomic resources leading to wrong interpretations for medically relevant genes
- A new approach together with a modified GRCh38 version that improves the subsequent analysis across these genes
- Improvements over multi-ethnic control samples across short and long-read DNA-, and RNA-sequencing

Project 2: Copy number analysis of LPA KIV-2 region using multi-ethnic cohort data

- Scaled detailed analysis of KIV-2 copy number of LPA gene that is one of the hard to assess regions of the human genome and has an impact on the cardiovascular risk
- Comparision of copy number distribution among various ethnicities (African, European, Asian, and Hispanics) using datasets from 1000 genome project (ikGP), Atherosclerosis Risk in Communities (ARIC), and Hispanic Community Health Study/Study of Latinos (HCHS/SOL)
- Analysis of LPA copy number with Lp(a) protein measurement and known small nucleotide variations that impact protein level
- Datasets: 3,202 WGS from 1kGP, 3,006 WGS from ARIC + SOL cohort; Tool: short-read based CNV tool developed by Illumina

Other Projects :

- k-mer based Structural variant genotyping using Illumina WGS data
- Analysis and comparison of single-cell WGS data of brain samples from different amplification methods (dMDA, picoplex and PTA)
- Benchmarking of structural and small variants of pancreatic cancer, medically relevant genes in X and Y chromosomes
- Comparision of assembly-based and mapping-based structural variant approaches
- Differential gene expression analysis of male reproductive tract-specific genes of mouse and human samples
- LPA diversity analysis using pan-genome approaches

Education, Training, and Experience (continued)

2012 – 2019 **Ph.D. Computer Science, University of Nebraska-Lincoln, NE**
Thesis title: *Suffix tree, minwise hashing and streaming and algorithms for big data analysis in bioinformatics*

- Identification of conserved non-coding sequences using suffix tree-based approach; used in comparative genomics study of grass species (collaboration with Dr. James Schnable at UNL)
- k-mer abundance estimation using streaming approaches
- Comparision and benchmarking of transcriptome assembly algorithms for plant species (collaboration with Dr. Etsuko Moriyama, UNL)

2009 – 2011 **M.S. Computer Science, the University of Texas at Dallas, TX**
Thesis title: *Algorithm and simulation for optimal delivery of volumetric modulated arc therapy (VMAT)*.

2001 – 2005 **B.Tech. (Hons) Computer Science, National Institute of Technology, Rourkela, India.**
First Class Honors.

Research Publications

Journal Articles

1. Behera, S., LeFaive, J., Orchard, P., Mahmoud, M., Paulin, L. F., Farek, J., Soto, D. C., Parker, S., Smith, A. V., Dennis, M. Y., Zook, J. M., & Sedlazeck, F. (2022). Fixing reference errors efficiently improves sequencing results. *bioRxiv*. https://doi.org/10.1101/2022.07.18.500506
2. Chin, C.-S., Behera, S., Metcalf, G. A., Gibbs, R. A., Boerwinkle, E., & Sedlazeck, F. J. (2022). A pan-genome approach to decipher variants in the highly complex tandem repeat of lpa. *bioRxiv*. https://doi.org/10.1101/2022.06.08.495395
3. Gan, L., Park, K., Chai, J., Updike, E. M., Kim, H., Voshall, A., Behera, S., Yu, X.-H., Cai, Y., Zhang, C., Wilson, M. A., Mower, J. P., Moriyama, E. N., Zhang, C., Kaewsuwan, S., Liu, Q., Shanklin, J., & Cahoon, E. B. (2022). Divergent evolution of extreme production of variant plant monounsaturated fatty acids. *Proceedings of the National Academy of Sciences*, *119*(30), e2201160119. https://doi.org/10.1073/pnas.2201160119
4. Smolka, M., Paulin, L. F., Grochowski, C. M., Mahmoud, M., Behera, S., Gandhi, M., Hong, K., Pehlivan, D., Scholz, S. W., Carvalho, C. M., Proukakis, C., & Sedlazeck, F. J. (2022). Comprehensive structural variant detection: From mosaic to population-level. *bioRxiv*. https://doi.org/10.1101/2022.04.04.487055
5. Voshall, A., **Behera, Sairam**, Li, X., Yu, X.-H., Kapil, K., Deogun, J. S., Shanklin, J., Cahoon, E. B., & Moriyama, E. N. (2021). A consensus-based ensemble approach to improve transcriptome assembly. *BMC Bioinformatics*, *22*(513). https://doi.org/https://doi.org/10.1186/s12859-021-04434-8
6. **Behera, Sairam**, Deogun, J. S., & Moriyama, E. N. (2020). Minisoclust: Isoform clustering using minhash and locality sensitive hashing. https://doi.org/10.1145/3388440.3412424
7. **Sairam Behera**, Deogun, J. S., & Moriyama, E. N. (2020). MinCNE: identifying conserved non-codingelements using min-wise hashing. *Advances in Computer Vision and Computational–Biology*.

8. **Behera, Sairam**, Gayen, S., Deogun, J. S., & Vinodchandran, N. V. (2018). Kmerestimate: A streaming algorithm for estimating k-mer counts with optimal space usage. *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 438–447. https://doi.org/10.1145/3233547.3233587

9. **S. Behera**, Li, X., Schnable, J., & Deogun, J. S. (2017). Dice: Discovery of conserved noncoding sequences efficiently. *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 79–82. https://doi.org/10.1109/BIBM.2017.8217628

10. **S, Behera**, Voshall, A., Deogun, J. S., & Moriyama, E. N. (2017). Performance comparison and an ensemble approach of transcriptome assembly. *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, *00*, 2226–2228. https://doi.org/10.1109/BIBM.2017.8218005

11. **Sairam Behera**, Lai, X., Liang, Z., Lu, Y., Deogun, J. S., & Schnable, J. C. (2017). Stag-cns: An order-aware conserved noncoding sequences discovery tool for arbitrary numbers of species. *Molecular Plant*, *10*(7), 990–999. https://doi.org/https://doi.org/10.1016/j.molp.2017.05.010

12. Pavlovikj, N., Begcy, K., **S. Behera**, Campbell, M., Walia, H., & Deogun, J. S. (2016). Analysis of transcriptome assembly pipelines for wheat. *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 137–140. https://doi.org/10.1109/BIBM.2016.7822507

13. Pavlovikj, N., Begcy, K., **S. Behera**, Campbell, M., Walia, H., & Deogun, J. S. (2014). A comparison of a campus cluster and open science grid platforms for protein-guided assembly using pegasus workflow management system. *2014 IEEE International Parallel Distributed Processing Symposium Workshops*, 546–555.

14. Pavlovikj, N., Begcy, K., **Behera, Sairam**, Campbell, M., Walia, H., & Deogun, J. S. (2014a). Comparing and optimizing transcriptome assembly pipeline for diploid wheat. *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, 603–604. https://doi.org/10.1145/2649387.2662450

15. Pavlovikj, N., Begcy, K., **Behera, Sairam**, Campbell, M., Walia, H., & Deogun, J. S. (2014b). Evaluating distributed platforms for protein-guided scientific workflow. *Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment*, 38:1–38:8. https://doi.org/10.1145/2616498.2616551

16. Pavlovikj, N., Begcy, K., **Behera, Sairam**, Campbell, M., Walia, H., & Singh Deogun, J. (2014). Evaluating assembly pipeline for transcriptomes. *Proceedings of the 6th International Conference on Bioinformatics and Computational Biology, BICOB 2014*.

17. **S. Behera**, Daescu, O., & Papiez, L. (2011). Optimal delivery of volumetric modulated arc therapy (vmat) for moving target. *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, 297–304. https://doi.org/10.1109/BIBMW.2011.6112390

Talks and Presentations

Oral Presentation

1. Baylor HGSC seminar talk, 2022, Title. *Towards fixing GRCh38 reference genome*
2. International Workshop on String Algorithms in Bioinformatics (StringBio), 2018, Title. *Suffix Tree Approach to Discover Conserved Non-Coding Sequences in Plants*
3. ACM BCB 2018, Title. *KmerEstimate: A Streaming Algorithm for Estimating k-mer Counts with Optimal Space Usage*

Talks and Presentations (continued)

4. IEEE BIBM 2017, Title. *DiCE: Discovery of conserved noncoding sequences efficiently*
5. UNL Center for Plant Science Innovation seminar talk, 2017, Title. *Towards Improving Transcriptome Assemblies*

Poster presentation

1. "Fixing falsely duplicated and collapsed regions of the GRCh38 reference genome", ASHG 2022 (Reviewer's choice abstract)
2. "Comprehensive genomics at scale using DRAGEN pipeline", CSHL Genome Informatics 2021
3. "Identification of allele-specific KIV-2 repeats among multi-ethnic groups and association with LP(a) measurements", ASHG 2021
4. "Performance comparison and an ensemble approach of transcriptome assembly", IEEE BIBM 2017
5. "STAG-CNS: Conserved non-coding sequence discovery tool", UNL Plant Science Retreat 2016, Nebraska City, NE

6 "Towards Improving Transcriptome Assemblies", UNL Plant Science Retreat 2016, Nebraska City, NE

7. "Transcriptome analysis of drought responses in wheat", UNL Plant Science Retreat 2014, Nebraska City, NE

Miscellaneous Experience

Teaching and Mentoring Experience

Spring 2014, 2015, 2016	**Instructor** CSE 251K: C Programming
Fall 2012 - Spring 2016	**Graduate Teaching Assistant** CSE 155E/H: Computer Science I, CSE 423/823: Design and Analysis of Algorithms, CSE 923: Development and Analysis of Efficient Algorithms and CSE 924: Graph Algorithms
Fall 2016	Mentored two masters students for their thesis on clustering and k–mer counting problems in bioinformatics.
Summer 2017	Mentored an undergraduate student on his research on NGS data analysis of corn–snake as a part of Nebraska INBRE Summer Undergraduate Research Program.

Awards and Achievements

2010	**Merit Award**, Dean's Excellence Scholarship for Graduate Students, University of Texas at Dallas.
2015	**Outstanding Teaching Assistant**, Department of Computer Science and Engineering, University Nebraska-Lincoln.
2018	**Graduate Student Conference Travel Grant**, College of Engineering, University of Nebraska-Lincoln.
	NSF Travel award, ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB).
	Travel Grant 2018 International Workshop on String Algorithms in Bioinformatics (StringBio).