PANKAJ PRADEEP

65 West 106th Street
New York, NY 10025

LinkedIn | GitHub Pages

(332) 250-4539
pp2831@columbia.edu

EDUCATION

Columbia University — New York, NY
Master of Science in Biomedical Engineering — Dec 2023
Coursework: Machine Learning for Functional Genomics, Statistical Machine Learning for Genomics, Computational Modeling of Physiological Systems, Biomedical Innovation, Applied Data Science

Vellore Institute of Technology (VIT) — Vellore, India
Bachelor of Technology in Biotechnology — July 2022
Coursework: Bioinformatics (Genetic Alignment and Protein Engineering), Molecular Biology, Immunology and Immunotechnology, Genetic Engineering, Programming (Python, R, MATLAB)

SKILLS

Bioinformatics	**Omics Data Analysis,** Samtools, GATK, **Seurat,** Bioconductor, **GEO, TCGA, CCLE**, UCSC, NCBI
Programming	**R/RStudio, Python**, Machine Learning, Bash, **Unix**, MATLAB, **HPC, Statistical Testing**
Research	scRNAseq, bulk RNAseq, Proteomics, Molecular Biology, Genomics, DNA/RNA Extraction, PCR
Packages	Scikit-learn, Pandas, NumPy, Keras, Tensorflow, Seaborn, **Monocle3, Scanpy**
Other	Trello, JIRA, Asana, Microsoft Office, LaTeX, Mendeley, Google Docs, Google Colab

EXPERIENCE

Columbia Irving Medical Centre - Gertrude H. Sergivesky Centre — Sep 2023 – Dec 2023
Student Research Worker *under Dr. Giuseppe Tosto* — New York, NY
• Performed quality control analysis of 2500 FASTQ files, leveraging fq2vcf for Bioinformatics Analysis to convert samples to VCF for further investigation of Alzheimer's Disease in underrepresented populations
• Designed a package on RStudio to evaluate 4 AD risk factors to visualize local ancestry distinguished by 3 major groups
• Co-ordinated with post-docs to troubleshoot errors in pipeline set-up and multi-tasking jobs (qsub, Unix) on HPC cluster
• Spearheaded deconvolution of RNAseq data to single cell compositions to determine biomarkers using CIBERSORTx

GlaxoSmithKline Pharmaceuticals — Jun 2023 – Aug 2023
Summer Intern – Computational Biology Oncology — Collegeville, PA
• Investigated 45 clinical signatures for gene and protein expression from public databases to determine drivers of tumors
• Implemented a correlation analysis pipeline in Python to visualize and distinguish patterns across transcriptomics and proteomics in 947 cancer cell lines from public databases, based on different subgroups (tissue, tumor, cancer type)
• Ingested 12 scRNAseq cancer datasets from GEO in Python into in-house Bioinformatics pipelines for further analysis while working individually and checking-in regularly with team sitting in GSK Germany
• Proposed an idea during GSK Science Meeting to use Earth Mover's Distance to quantify and demonstrate differences between correlation patterns across 30+ cancer groups

Columbia University - Herbert and Florence Irving Institute for Cancer Dynamics — Oct 2022 – Jun 2023
Student Research Worker — New York, NY
• Leveraged AmpliconArchitect to identify regions in genome where fragmentation or breakpoints have occurred, to generate amplicons visualizations annotated with these regions to show presence of extrachromosomal DNA (ecDNA)
• Analyzed 80 whole genome samples in the HPC (bedtools, samtools, bwa) to locate and identify regions of potential hits of extrachromosomal DNA using Integrated Genome Viewer, UCSC Browser for refining amplicon images
• Worked independently and as a team in setting up Packages for further analyzing genome data for esophageal cancer
• Organized literature surveys (PubMed) with detailed reports to further understand mechanism of ecDNA occurrence

RESEARCH PROJECTS

Columbia University: Enhancing TIGER Model for Cas13 Off-Target Prediction for Indel gRNA Data — Sep 2023 – Dec 2023
• Initiated the evaluation of various RNN and Transformer architectures for Cas13d off-target activity prediction
• Re-architected CNN model for compatibility with input data with insertions/deletions by leveraging Levenshtein (edit) distance. Enhanced model robustness by increasing tolerance of input diversity for better off-target score prediction
• Modelled utils package to assess performance of re-tuned model with indel's effect as labels to infer model compatibility

Columbia University: BRCA-SAE Multi-Omics Data Integration using Stacked Autoencoders — Jan 2023 – May 2023
• Invented a Stacked Autoencoder Model to represents 3 different data modalities in a latent space to infer driving biomarkers for cancer progression in datasets for 1000 patients ingested from TCGA for breast invasive carcinoma
• Collaborated effectively with teammate to troubleshoot errors in model set-up and training for multi-omics analysis
• Performed Feature Attribution to discover top 20 biomarkers which are drivers of cancer progression across all datasets

INTERESTS

Tennis, Soccer, Hack-a-thons, Medical Devices, Point-of-care diagnostics, Automobiles