Nikhil Mark Lakra

Personal Site | nikhilmarklakra@gmail.com | +91-9599895356 | LinkedIn | GitHub

Experience

Bioinformatics Scientist I, Elucidata — Oct 2022 – Present, New Delhi

- Collaborated with leading pharmaceutical companies, research institutions, and innovative startups, applying **bulk and single-cell RNA-Seq transcriptomics** to generate actionable insights for **early-stage drug discovery and development** across a spectrum of therapeutic areas.
- Developed an efficient, scalable, and reproducible gene similarity analysis workflow optimized for large-scale single-cell RNA-seq datasets (averaging 100k+ cells), leveraging **Pandas**, **Scikit-learn**, **Scanpy**, and **scVI-tools** within a **dockerized environment**, directly contributing to a **40% increase** in contract renewal value.
- Designed and deployed a **Python** and **R**-based meta-analysis pipeline incorporating **KNN imputation** to enhance differential gene expression and pathway enrichment analyses, enabling the **identification of key therapeutic targets** from complex multi-study datasets.
- Developed a workflow using pre-trained machine learning models (e.g., **scVI-tools, CellTypist**) to automate the annotation of cell types in scRNA-Seq data, **reducing the time to identify and accurately annotate cell types by 20%.**

Education

Indraprastha Institute of Information Technology Delhi (IIITD) — Aug 2019 – Jan 2022
M.Tech in Computational Biology — CGPA: 7.96/10.0

- **Coursework:** Machine Learning, Data Science in Genomics, Biostatistics, Computer Aided Drug Design

Projects and Thesis

clustermole_py: Python Package for Single-Cell Cluster Annotation — GitHub Link

- Developed a **Python package** inspired by **clustermole (R)** for **biological annotation of single-cell RNA-seq clusters** using gene set enrichment analysis.
- Implemented modules for **Enrichr API integration** and **Gene Set Variation Analysis (GSVA)**, enabling DE-free cluster annotation.
- Designed for **seamless integration with Scanpy AnnData** objects, facilitating common single-cell workflows.

Thesis: Predicting Selection Pressure on SNPs in Human Populations — Link to Thesis

- Developed a **machine learning framework** (Python/scikit-learn) to identify SNPs under positive selection across 17 global populations using **1.7M SNPs** from Phase III of the 1000 Genomes Project.
- **Processed and analyzed VCF files** for 2,504 individuals, filtering biallelic SNPs and calculating population genetics statistics (FST, XP-EHH, DDAF) to train a **Random Forest classifier** (MCC: 0.89, AUC: 0.94).
- Validated findings through **PCA clustering** and **eQTL analysis** (GTEx), linking selected SNPs to tissue-specific gene expression in fibroblasts and whole blood.

Technologies

Languages: Proficient in Python, R, SQL, Bash

Technologies & Frameworks: Genomic and Transcriptomic Analysis (DESeq2, edgeR, LIMMA, Gene Set Enrichment Analysis), Single-cell RNA-seq Analysis (Scanpy, Seurat, scVI-tools), Data Manipulation & Analysis (Pandas, NumPy), Machine Learning (scikit-learn, TensorFlow, PyTorch, NLTK)

Pipeline Development & Cloud Tools: Docker, Git, AWS, Snakemake, PostgreSQL, SQLite