312 Home Park Ave NW
Atlanta, GA, 30318
(470)-800-6887

Shreyash Gupta

shreyash.hanu@gmail.com
linkedin.com/in/sgupta755
github.com/Hanuphant

Curious and insightful Bioinformatics Data Scientist with expertise in statistical genomics. Seeking to develop and implement various integrative multi-omics technologies.

Education

MS Bioinformatics, Georgia Institute of Technology **Aug 2021 — Dec 2022**
BTech Biotechnology, National Institute of Technology, Warangal **Jul 2017 — May 2021**

Research Experience

Bioinformatics Analyst Geisinger **Mar 2023 — Present**

- Performed various Genome-Wide Association Studies for multiple complex traits for obesity and other anthropomorphic features extracted from MyCode EHR BioBank which has over 230k WES sequenced samples
- Executed multiple post-GWAS secondary analyzes including conditional analysis, fine mapping, colocalization, and gene enrichment analyzes
- Worked with Liver adipose tissue RNA-Seq data sets to perform cis-QTL mapping and colocalization of the major waist-hip ratio GWAS signals across 600 significant loci
- Extracted and filtered multiple phenotype and clinical data for various ICD codes and other diagnosis codes from in-house MyCode EHR BioBank
- Regularly collaborated with interdepartmental labs and multiple consortia most prominently with GIANT, CHARGE and ORAcLE working groups

Graduate Research Assistant Jordan Lab, Georgia Tech **Aug 2021 — Dec 2022**

Biological Health Score correlation with Genetic Variation in UKBB full-sibling cohort Aug 2022 — Dec 2022

- Studied the probable effect of genetic variation on the blood biomarkers among full-sibling pairs from UK BioBank sample cohort
- Computed genetic kinship estimates for UKBB sample cohort to determine full-siblings pairs using King's estimator
- Calculated the Biological Health Score (BHS) based on the blood biomarkers for each sample
- Evaluated the correlation of change in BHS with respect to genetic variation exhibited by full-sibling pair and monozygotic twin pairs using multivariate statistical analysis

Genomic Ancestry Inference on COVIRT Aug 2021 — May 2022

- Performed Genetic Ancestry Inference RNA-Seq analysis on COVIRT Data by estimating ancestry estimates while comparing gene expression levels to potentially find population ancestry effects on COVID-19 patient gene expression
- Using RNA-Seq variant calling pipelines on extracted COVIRT RNA-Seq data and 1000 Genome Project reference data, found ancestry estimates for the samples using ADMIXTURE
- Differential Gene Expression pattern was analyzed using LIMMA for each ancestral population to find significantly expressed genes
- Applied Gene Enrichment Analysis (FGSEA) for gene expression pathway analysis among the ancestral groups

Bioinformatics Research Assistant Data Science, Oerth Bio **May 2022 — Aug 2022**

- Produced small molecular compound counts from DNA-encoded chemical library experiment generated NGS data dump via deploying Levenshtein algorithm, batch-processing and multi-processing to improve the number of match counts, robustness and time efficiency of the script
- Performed further enrichment analysis of DEL counts to filter counts of small molecular compounds as well as identify top 50-100 molecules with significant binding affinity
- Developed effective Python scripts to partially automate the creation of conservation scorecards and alignment of key binding residues in E3 Ligase homologs across species
- Devised a computational E3 Ligase Discovery approach to provide insights and analysis for targeting E3 ligases binding to the protein of interest by applying hierarchical clustering techniques and HMM model-based tools for domain determination
- Leveraging REST-APIs conceptualized custom Python scripts for API calls to UniProt, PDB and muSiteDeep to specific analysis of POI

Graduate Teaching Assistant Cell and Molecular Biology, Georgia Tech **Jan 2022 — May 2022**

- Delivered several review sessions based on the course structure of Cell and Molecular Biology
- Supervised group project work among students and also provided suggestions in different phases of their project
- Evaluated submissions and provided valuable and apt feedback to students

Undergraduate Research Saxena Lab, NITW **Sept 2020 — Apr 2021**

- Performed Machine Learning based identification of breast cancer sub-types using RNA-Seq Data
- Extracted RNA-Seq Expression data from cBioPortal cancer database to further study breast cancer classification
- Applied principal component analysis on 20,400 genes and further utilized and compared ML training models like Random Forest, Artificial NNs and SVM Kernels

Projects

Exploratory Study of the Google Search Symptom Trends **Oct 2022 - Dec 2022**

- Investigated the viability of the Google Search Symptom Trends to analyze the COVID pandemic on a regional level using model transfer, correlation and clustering analysis, anomaly detection, and forecasting as complex metrics
- Conducted correlation analysis of the symptom searches to measure the association of the symptom search trends to COVID-19 pandemic trends via cross-correlation function, Granger causality test as well as recursive feature elimination strategies
- With feature engineering introduced multiple metrics based on cross-correlation function and Granger causality test to cluster the high dimensional data using UMAP and OPTICS.
- Found relevant 3 clusters of 43 symptoms of directly associated COVID-19 symptoms, early-stage and mild infection symptoms
- Assisted in the selection of forecasting models as well as identify which set of symptoms perform forecasting better
- Final insights available on https://ritusinha128.github.io/CS8803_EPI/
- Code for the study available as a tarball on the website above as well as on the GitHub repo: Hanuphant/EPI-GoogleSymptom-Trends

Human Computational Genomics **Sep 2022 - Dec 2022**

- Developed RShiny app for identifying transcription factor targets of differentially expressed genes
- Works on the integration of bulk RNA-Seq and transcription factor ChIP-Seq experiment using BETA to aid in the inference of the underlying regulatory gene networks discussed on previous workflows in studies
- Implemented REST-API calls to fetch the ChIP-Seq files from the ENCODE Repository based on the desired biological cell line
- Accessible website to the tool RanCh: https://genapp2022.biosci.gatech.edu/team5/
- Code for the tool RanCh: Hanuphant/RanCh

Computational Genomics **Jan 2022 — May 2022**

- Performed functional annotation for predicted genes from Salmonella isolates and developed a predictive webserver data pipeline to integrate genome assembly, gene prediction, functional annotation and comparative genomics pipelines.
- Utilized a mixture of homology based and ab initio tools such as InterProScan and EggNOG mapper for the functional annotation of predicted genes after genome assembly.
- Implemented Django based Nginx server as well as crontab for back-end filesystem server connect for the development of the webserver
- Code for Functional Annotation work available on the GitHub repo: Hanuphant/Functional_Annotation_Pipeline
- Code for Predictive Webserver available on the GitHub repo: Hanuphant/Webserver_Computational_Genomics

Biomedical Image Processing for COVID CT-scans **Feb 2022 — Apr 2022**

- Developed a Data Quality Control pipeline for COVID-19 CT Scans to address data imbalance issues, missing data, and anomaly detection issues as well as enhanced standardization for better prediction, segmentation, and classification of COVID-19 CT scans
- Enhanced image data quality by the application of Generative Adversarial Network (GAN) deep learning concepts
- Addressed data class imbalance by exploring loss calculation hyper-parameter options
- Report available on: Deep Learning for Disease Classification and Outcome Interpretation on COVID-19 CT Images
- Code for Biomedical Image Processing available on GitHub repo: onlyshawn/HackathonCovid19Proj

Publications

1. Kurniansyah, N. *et al.* Polygenic scores for obstructive sleep apnoea reveal pathways contributing to cardiovascular disease. en. *EBioMedicine* **117,** 105790 (July 2025).
2. Zhang, X. *et al.* Whole genome sequencing analysis of body mass index identifies novel African ancestry-specific risk allele. en. *Nat. Commun.* **16,** 3470 (Apr. 2025).
3. Pabbathi, N. P. P. *et al.* Role of metagenomics in prospecting novel endoglucanases, accentuating functional metagenomics approach in second-generation biofuel production: a review. en. *Biomass Convers. Biorefin.* **13,** 1371–1398 (2023).

Technical Skills

Computational Skills	Data Science including deep learning (Pytorch), Visualization (Rshiny and Dash), REST-APIs
Programming Languages	Expert ability with Python, intermediate in R and Bash (*nix) terminal, SQL and Octave

Certifications

- **Machine Learning:** Machine Learning by Andrew Ng (Certified)
- **Statistical Genomics:** Statistics for Genomic Data Science by Jeff Leeks (Certified)