Novel Predictive Algorithms for Metastasis-Associated CNVs via Cross-Cancer Variant Frequency Analysis.

Research Project: Design and Implementation of Novel Predictive Algorithms for Identifying Genetic Mutations and Variations via cross-cancer analysis of variant allele frequencies in Cancer Genomes (with an emphasis on common metastasis-associated CNVs).

The results of this study will be published and the developed tool will be deployed on Linux as well as on the web using frameworks such as Flask, Nextflow, or Snakemake.

Overview

This curriculum provides a structured 5-month training program on cancer bioinformatics and machine learning. It integrates biological knowledge, data science, and computational tools to analyze copy number variations (CNVs), RNA-Seq, and proteomics data from The Cancer Genome Atlas (TCGA). The focus is on Lung (LUAD, LUSC), Breast (BRCA), and Colorectal (COAD) cancers, covering data preprocessing, analysis, and machine learning applications for mutation prediction.

Each activity will be performed across all three selected cancer types.

Month 1 – Foundations in Cancer Genomics and Data Acquisition

Week	Topics & Activities	Reasoning & Discussion Focus	Content & Learning Objectives	Tutor/Speaker
1	**Introduction to Cancer Genomics** – Overview of cancer genetics, metastasis pathways, and key mutations	Explore why understanding the biological basis of cancer (driver vs. passenger mutations, CNVs, somatic mutations) is vital for	*Objectives:* • Grasp core concepts in cancer genetics and metastasis • Identify key genetic markers in LUAD, LUSC, BRCA, and COAD	Mr. Seyi

		predictive modeling		
2	**Introductory Bioinformatics** – Definitions, file formats (FASTA, VCF, BAM), and major databases (TCGA, Ensembl, UCSC)	Discuss the importance of standardized data formats and repositories for reproducibility and data sharing in research	*Objectives:* • Understand essential bioinformatics terminology • Learn about data repositories and file formats	Chioma
3	**Data Acquisition for Cancer Projects** – Methods to obtain CNV, RNA-Seq, clinical, and proteomics data	Debate the pros and cons of different data collection approaches: 1. Field research (in-house data collection) 2. Requesting controlled/private data 3. Utilizing and merging freely available data from cancer databases Discuss how integrating multiple data dimensions yields a “whole data” perspective	*Objectives:* • Learn data retrieval strategies • Understand merging heterogeneous data (clinical, CNV, RNA-Seq, proteomics)	Faruq/Chioma
4	**Introduction to RNA-Seq & Gene Expression Analysis** – Fundamentals of RNA-Seq, differential expression, and an overview of proteomics data	Explore why RNA-Seq is the gold standard for transcriptomics and how proteomics adds another dimension to understanding cancer biology	*Objectives:* • Understand RNA-Seq workflows and gene expression normalization • Connect transcriptomic and proteomic insights with genomic data	Chioma/Faruq

Note: Data acquired in Month 1 will serve as the raw input for feature extraction and analysis in later modules (notably in Months 3 and 4).

Month 2 – Data Preprocessing and Programming Foundations

Week	Topics & Activities	Reasoning & Discussion Focus	Content & Learning Objectives	Tutor/Speaker
5	**Linux for Bioinformatics** – Setting up a Linux environment, essential commands, file management, and running FastQC for CNV quality control	Discuss why Linux is favored in bioinformatics and the benefits of reproducible command-line workflows	*Objectives:* • Master Linux/Bash basics • Perform data quality control efficiently	Faruq
6	**R for Data Visualization** – Data manipulation and visualization of RNA-Seq/gene expression data	Examine the role of visual data exploration in detecting patterns and anomalies in complex genomic datasets	*Objectives:* • Clean and visualize RNA-Seq data using R • Interpret gene expression plots	Chioma
7	**Python Fundamentals** – Introduction to Python (variables, data structures, control flow, and functions)	Compare Python with other languages; emphasize Python's flexibility for data processing in bioinformatics	*Objectives:* • Write basic Python scripts for genomic data • Use loops and functions to process data	Faruq
8	**APIs, Data Merging, and Transformation** – Using APIs for data retrieval,	Discuss challenges in integrating heterogeneous data sources	*Objectives:* • Retrieve data via APIs	Ayomide

	merging clinical and omics data, normalization techniques	and the importance of normalization for downstream analysis	• Merge and normalize datasets • Prepare data for feature extraction	

Month 3 – Advanced Data Handling & Biopython Applications

Week	Topics & Activities	Reasoning & Discussion Focus	Content & Learning Objectives	Tutor/Speaker
9	**Introduction to Biopython** – Parsing FASTA/VCF files, sequence I/O, and working with annotation data	Discuss how Biopython enables reproducible and modular analysis; revisit data from Month 1 for deeper feature extraction	*Objectives:* • Parse and manipulate biological sequences • Link sequence data with earlier acquired clinical and genomic datasets	Faruq
10	**Sequence Alignment & Object-Oriented Programming (OOP) in Python** – Performing pairwise and multiple sequence alignments; implementing OOP for data encapsulation	Explore the benefits of OOP for managing complex genomic data; review why accurate alignments are critical for mutation detection	*Objectives:* • Conduct sequence alignments using Biopython • Build simple classes for genome data analysis	Ayomide/Faruq
11	**Advanced Data Handling with Pandas** – Data wrangling,	Emphasize the need for structured data manipulation; discuss different	*Objectives:* • Use Pandas to clean and merge datasets	Ayomide

	cleaning, and feature extraction from mutation and CNV data	methods for feature extraction and transformation	• Extract features for machine learning from integrated genomic data	
12	**Data Visualization for Genomic Insights** – Creating and interpreting visualizations (mutation frequencies, coverage maps) using Matplotlib and Seaborn	Discuss how effective visualization leads to better understanding and hypothesis generation in bioinformatics projects	*Objectives:* • Generate and interpret visualizations • Connect visual patterns with underlying biological processes	Gloria

Month 4 – Machine Learning for Genomic Analysis

Week	Topics & Activities	Reasoning & Discussion Focus	Content & Learning Objectives	Tutor/Speaker
13	**Introduction to Machine Learning & Feature Engineering** – Overview of supervised learning; designing features from genomic data	Discuss why ML is transformative in genomics and the criteria for effective feature selection from biological data	*Objectives:* • Understand ML basics and prepare feature vectors • Evaluate which features best represent genomic variations	Ayomide
14	**Linear Models for Mutation Prediction** – Logistic regression applied to mutation data	Analyze the trade-offs between model simplicity and predictive power; explore interpretability in linear models	*Objectives:* • Implement logistic regression • Interpret model outputs in a biological context	Ayomide

15	**Random Forests for CNV & Mutation Analysis** – Training and tuning ensemble models; evaluating feature importance	Discuss the robustness of ensemble methods and how they can uncover hidden patterns in noisy data	*Objectives:* • Build and tune Random Forest models • Assess the importance of different genomic features	Ayomide/Faruq
16	**Support Vector Machines (SVM) for Variant Classification** – Theory, kernel functions, and SVM applications in high-dimensional genomic data	Explore the rationale behind SVMs' effectiveness in complex data spaces and the concept of margin maximization	*Objectives:* • Apply SVMs to classify genomic variants • Optimize SVM parameters for accuracy	Faruq
17	**Neural Networks (CNNs) for Mutation Prediction** – Building convolutional neural networks for sequence analysis	Delve into why deep learning excels at capturing spatial patterns in DNA sequences and the challenges of overfitting	*Objectives:* • Design and train a CNN for mutation detection • Evaluate network performance on genomic data	Ayomide

Month 5 – Model Evaluation, Tool Development, and Reporting

Week	Topics & Activities	Reasoning & Discussion Focus	Content & Learning Objectives	Tutor/Speaker
18	**Comparative Model Evaluation &**	Discuss why multiple metrics are necessary	*Objectives:*	Faruq/Ayomide

	Statistical Analysis – Performance metrics (Accuracy, Precision, Recall, AUC-ROC), cross-cancer comparisons, statistical tests	for robust evaluation; interpret ROC curves and significance tests in context	• Compare ML models rigorously • Validate model differences statistically	
19	**Results Interpretation & Scientific Reporting** – Synthesizing findings, visualizing results, and report writing for publication	Examine best practices in scientific communication and how to effectively translate complex data into publishable research	*Objectives:* • Prepare detailed reports and visualizations • Develop a manuscript for publication	Mr. Seyi
20	**Tool Development & Deployment** – Building a Linux-based command-line tool, deploying on the web (using Flask, Nextflow, or Snakemake), and documentation	Discuss the importance of reproducibility and accessibility; compare deployment frameworks and their pros and cons	*Objectives:* • Develop a deployable bioinformatics tool • Document installation and usage for end users	Ayomide/Tide/Col
21	**Final Project Presentations & Workshops** – Student presentations, peer review, and discussion of future research directions	Engage in reflective discussions on project challenges, lessons learned, and potential publication strategies	*Objectives:* • Present and critique final projects • Collaborate on refining research for publication	Panel of All Tutors

Final Integrated Project

- **Objective:** Combine CNV, RNA-Seq, clinical, and proteomics data for LUAD, LUSC, BRCA, and COAD to design and implement predictive ML models that identify metastasis-associated CNVs.
- **Deliverables:** A deployable command-line and web-based tool, a comprehensive scientific report, and a manuscript prepared for publication.
- **Evaluation:** Projects are assessed on technical accuracy, model interpretability, reproducibility, and the clarity of scientific reasoning.

Conclusion

This 5-month curriculum not only covers the technical skills required in cancer genomics, data preprocessing, and machine learning but also embeds dedicated sessions for critical reasoning and scientific communication. It ensures that students are well prepared to develop a publishable research project and deploy a robust, user-friendly tool on Linux and the web using frameworks like Flask, Nextflow, or Snakemake.