GENOMAC INNOVATION HUB (G-iHUB)
...Exploring the Future of Health Tech & Broader Tech Innovations
Address: Genomac Holdings, Beside Alari Akata Filling Station, Under G, Ogbomosho, Oyo State, Nigeria.
Tel: +2348077191794 Email: contact@genomac.io
Website: www.genomac.io

Genomics Datascience with Python Internship Curriculum

Start Date: November, 2025.

Overview

This curriculum outlines a training initiative designed to:

- Empower students and early-career professionals in computational biology and bioinformatics

- Expand our network of collaborators, instructors, and ecosystem partners

Program Topic: Genomics Datascience with Python: Learn how to Deploy a Genomics Data Ml model (1-month Internship)

Delivery Format

- Duration: 8 Weeks

- Mode: Online (Zoom + Google Colab)

- Format: Weekly live training / Student support forum

Outcomes

- Build core skills in Python, data wrangling, and genomic preprocessing

- Prepare students for advanced ML pipelines in genomics

Syllabus

Module 1: Foundations of Genomics & Python for Bioinformatics

- 1.1 Welcome + Course Overview

- 1.2 Key Concepts in Molecular Biology
    - DNA, RNA, proteins, mutations, gene expression
- 1.3 Python Refresher (For Biologists or New Coders)
- 1.4 Setting Up Your Bioinformatics Environment
  (Biopython, Pandas, Jupyter, Conda, GitHub)

Module 2: Working with Genomic Sequence Data

- 2.1 Parsing FASTA & FASTQ Files with Biopython
- 2.2 Sequence Extraction, Cleaning & Slicing
- 2.3 K-mer Frequency Analysis and GC Content
- 2.4 Sequence Encoding Techniques (One-Hot, k-mer, embeddings)

Module 3: Clinical and Phenotypic Data Integration

- 3.1 Understanding Clinical Genomic Data (e.g., TCGA)
- 3.2 Parsing Metadata & EHR-style Tables (CSV, TSV)
- 3.3 Cleaning and Label Encoding Phenotypes
- 3.4 Merging Clinical and Genomic Datasets

Module 4: RNA-Seq Data Processing

- 4.1 Introduction to RNA-Seq and Transcriptomics
- 4.2 From Raw Counts to TPM/FPKM
    - (Use preprocessed data for simplicity)
- 4.3 Differential Gene Expression (DEG) Basics
- 4.4 Visualizing Gene Expression (Volcano plots, heatmaps)

Module 5: Proteomics Data Handling

- 5.1 Basics of Proteomics and Mass Spec Output
- 5.2 Parsing Protein Quantification Tables
- 5.3 Linking Proteins to Genes (Uniprot, Ensembl)
- 5.4 Normalizing and Transforming Proteomics Data

Module 6: Copy Number Variation (CNV) Data

- 6.1 What is CNV and Why It Matters?
- 6.2 Exploring CNV Datasets (e.g., SEG files, GISTIC output)
- 6.3 Extracting Regions, Chromosomes, and Gene-Level CNVs
- 6.4 Integrating CNV with Clinical or Expression Data

Module 7: Feature Engineering & ML Readiness

- 7.1 Feature Selection Across Data Types (DNA, RNA, CNV, Clinical)
- 7.2 Dimensionality Reduction: PCA, t-SNE for Omics Data
- 7.3 Label Preparation for ML (e.g., Disease vs Control)
- 7.4 Train-Test Split & Preprocessing Pipelines

Module 8: Capstone Project – Multi-Omics Integration