GMR
The Original

Teaching Gene Discovery and Population Genomics Through Authentic Analysis of Large-Scale Sequencing Datasets

Sitora Sanokulova, Feruza Matyakubova, Obid Saydayev, Feruza Azizova, Akmal Sidikov, Ravshan Sultanov,

Assistant Lecturer, Bukhara State Medical Institute, Bukhara, Uzbekistan, ORCID: https://orcid.org/0009-0006-2169-6547, E-mail: sanoqulova.sitora@bsmi.uz
Department of Infectious Diseases, Samarkand State Medical University, Samarkand, Uzbekistan, ORCID: https://orcid.org/0009-0008-2020-447X, E-mail: fmatakubova@gmail.com
Department of Physics, Jizzakh State Pedagogical University, Jizzakh, Uzbekistan, ORCID: https://orcid.org/0000-0003-3228-9200, E-mail: obidsaydayev@gmail.com
Professor at the Department of Hygiene of Children and Adolescents, Nutrition Hygiene, Tashkent State Medical University, Tashkent, Uzbekistan, 100109, 0000-0001-6360-503X, feruzaziz@mail.ru
Doctor of Medical Sciences, Professor, Rector, Fergana Medical Institute of Public Health, Fergana, Uzbekistan, ORCID: https://orcid.org/0000-0002-0909-7588, E-mail: medik-85@bk.ru
Department of Medical Fundamental Sciences, Termez University of Economics and Service, Termez, Uzbekistan, ORCID: https://orcid.org/0009-0005-5269-7537, E-mail: ravshan_sultonov@tues.uz

ABSTRACT

The increasing availability of affordable large-scale sequencing data has created new opportunities for undergraduate instruction that closely mirror contemporary genomics research. This paper describes classroom-tested teaching modules that introduce students to gene discovery and population genomics through authentic analysis of large public sequencing datasets. Using data from the Genome Aggregation Database (gnomAD) and the 1000 Genomes Project, students investigate the genetic basis of colour-blindness, population-specific adaptation, and complex evolutionary responses associated with the introduction of maize. The modules emphasize statistical signals of selection and association, guiding students to identify candidate loci and interpret genotype–phenotype relationships within diverse human populations. Analyses are conducted using DryLab workflows implemented on a cloud-computing platform, enabling accessibility across varying skill levels and institutional resources. Integration of Writing-to-Learn and Data-Intensive Science pedagogies supports conceptual understanding and critical interpretation of population-genomic data. By leveraging public datasets and authentic analytical frameworks, this approach enhances student engagement, strengthens data literacy, and provides a scalable model for teaching gene discovery and population genomics in undergraduate life-science education.

Keywords: *Gene discovery; population genomics; large-scale sequencing; gnomAD; 1000 Genomes Project; genome-wide association studies (GWAS); statistical signals of selection; undergraduate education; data-intensive science; authentic genomics analysis.*

INTRODUCTION

The advent of affordable large-scale sequencing empowers instructors to engage students in authentic genomics analyses that reflect contemporary research practices. These analyses educate students about gene-discovery and population-genomics concepts while also stimulating interest in genetics and genomics careers. Classroom-tested modules guide students through the analysis of human populations sampled by the Genome Aggregation Database (gnomAD), investigating the genetic basis of colour-blindness and population-specific adaptation in response to the agricultural introduction of maize (Yuldashev, A. G. (2024). The modules enable students to explore geographical variation in traits of interest and subsequently locate candidate loci in analysis cohorts. Datasets sampled from the 1K Genomes and gnomAD projects are also used to investigate complex-adaptive evolution following the introduction of maize to Europe. In both cases, students draw on the statistical signals of selection and association gleaned from large-scale datasets, thus widening their appreciation of DEMETER's capabilities. Students retrieve these datasets through the Dataverse data repository and apply a suite of complementary DryLab analyses using the DryLab cloud-computing platform hosted by the Gnomon project (Prost et al., 2020). Public-access initiatives like gnomAD continue to disseminate extensive sequencing datasets that pedagogically connect to microscopy- and bioinformatics- based analyses such as AQUA and SEquential DataAlysis. Classroom-ready protocols further facilitate engagement with these datasets through complementary DryLab analyses that span diverse student skill sets. Analytical frameworks tailored to authentic datasets illuminate statistical properties that remain largely hidden without exposure to wide-ranging populations. Supported by extensive Writing-to-Learn (WTL) and Data-Intensive Science (DIS) literature, separate description of population-genomics datasets, constraints, and pedagogical enrichment ensures alignment with the objectives-of-authentic-analysis framework and directs focus to student-oriented terminal outcomes for maximal learning impact (Lutfullaeva, D. E., & Yuldashev, A. G. 2023) Efforts to mobilize large-scale, population-genomics datasets for the wider Luxembourg education ecosystem further explore the pedagogical opportunities presented by affordable sequencing technologies and heightened data availability.

Foundations of gene discovery

Statistical signals of selection and association guide genetic discovery. Signals of selection arise from changes in allele frequency due to natural selection, while signals of association stem from a nonrandom correlation between genotype and phenotype (Prost et al., 2020). Although both signals reflect genotype–phenotype relationships, their interpretation becomes distinct from the dataset analysis and model specification. This distinction emerges from the generally smaller number of expected beneficial mutations compared to the usually larger number of neutral mutations (Yang et al., 2017). Consequently, beneficial mutations are detected with greater confidence, whereas neutral mutations require accompanying information and more specific connections between genotype and phenotype. Detection of signals of selection regularly relies on the advanced models implied in high-dimensional genomic data and typically becomes available through well-cited frameworks that adopt an influential yet simplistic approximation. Meta-analyses of genome-wide association studies (GWAS) involving large human cohorts indicate that signal characteristics depend on dataset size and that allele-frequency shifts generally form the strongest candidate signals of selection; overlap with GWAS signals of multiple complex phenotypes is also documented. Prioritization relies predominantly on genome-functional annotations, recommended regulatory sites, or experimental validation pertaining to indication edges. The estimated average fraction of benign or effectively neutral variants among functionally annotated variants is 40–50% for diverse vertebrate groups, consistent with large-scale "de novo" mutation studies.

Statistical signals of selection and association

Human populations exhibit strong variation in many traits, resulting from complex interactions between genotype, environment, and exposure to pathogens. Understanding connections between genetic variants and biological pathways, linking them to phenotypes of interest, and inferring their adaptive relevance, constitutes a central challenge in evolutionary biology and population genomics. Because these tasks are essential in understanding complex traits, they represent a valuable case study for upper-undergraduate instruction. Focus on data analysis in place of experimental design opens the field to a broader audience accessible to neither laboratory resources nor biological expertise. The increasing availability of genome-wide sequencing datasets from human cohorts constitutes an attractive vehicle for addressing such challenges, enabling exploration of gene discovery, trait association, and population adaptation within an authentic analytical framework using public datasets (Gao, 2024) ; (Udpa et al., 2011) ; (K. Oleksyk et al., 2010).

Functional annotation and interpretation

Genome-wide association studies (GWAS) assess whether variants at a specific genomic location exhibit different frequencies in case and control samples . These studies rely on large-scale, population-based samples and often investigate low-frequency variants. A mutation typically shows a stronger association at regions where the risk and population frequencies differ from the founder allele. However, numerous population-level processes can impact mutation frequency (O'Hely et al., 2006), making GWAS signals hard to assess directly. When signal-strength metrics are applied to dimensions associated with selection, multiple sampling schemes produce no or very weak signals, revealing the multifaceted nature of population processes affecting variant frequency. Observational studies identify GWAS signals by leveraging phenotype data, yet the process becomes intricate if population structure has affected population-specific frequencies under the assumption that selection operates on a shared resistance allele and each population has only one putative risk variant. To bridge genes and variants with biological impact on phenotype, genes that could mediate straightforward biological effects remain relevant (Sifrim et al., 2012). Knowledge of functional consequences of variants within the populations reduces the impact of opportunistic design on trait-variant association [table 1].

Table 1: Foundations of Gene Discovery

Topic	Description	Educational Relevance	Examples / Notes
Signals of Selection	Changes in allele frequency due to natural selection	Teaches students how evolutionary pressures shape genomes	Detected with high confidence; smaller number of beneficial mutations
Signals of Association	Nonrandom correlation between genotype and phenotype	Shows students how genotype influences observable traits	Requires larger datasets or functional annotation to interpret neutral mutations
Statistical Analysis	Use of GWAS, allele frequency shifts, meta-analysis	Enables data-driven gene discovery and trait association without lab experiments	Datasets from large human cohorts allow authentic analytical experience
Functional Annotation	Linking variants to gene function and phenotype	Connects genetic variants to biological pathways	Reduces confounding from population structure or neutral variants
Experimental Validation	In silico predictions, functional assays, CRISPR, transgenic models	Teaches causality testing and high/low-throughput experimental design	Provides hands-on confirmation of statistical inferences

Experimental validation approaches

Experimental approaches for validating the effects of genetic variation on phenotypes differ markedly in throughput and feasibility. In silico predictions and modelling of gene function offer useful but indirect assessments of causality (Prost et al., 2020). At one extreme, sophisticated population-genomic modelling techniques can be applied to identify genes under selection and predict their adaptive role. At the other, direct measurement of the effects of biological perturbations via functional assays or targeted genetic manipulations, such as CRISPR-Cas9 or transgene approaches in cell lines or laboratory embryos, forms the most conclusive tests. Such high- or low-throughput methods provide the classic trade-off navigated by an investigator wishing to interrogate the basis of the association between variation in a gene and variation in a target phenotype.

Population genomics and large-scale sequencing

Large-scale genomic data enable authentic analysis of population genomics, yet sampling, processing, and description remain underexplored. Nationally representative data and population descriptions frame population genomics analyses of large-scale datasets. Population-genetic signals include diversity, linkage disequilibrium (LD), population structure, admixture, and demographic history, whose interpretations arise from sampling design. Large-scale genomic data offer unprecedented insight into the population history of diverse organisms, opening new territories for scientific exploration. Population-genomic data shape access frameworks and governance structures. The unprecedented scale and breadth of sequence data generated in recent years have enabled authentic analysis of population genomics in typically unrepresented taxa, expanding scientific frontiers. Early experimental programs produced time-series data elucidating the genomic underpinnings of adaptation and the evolutionary process. Supplementary materials introduce population-genetic signals conceptually or mathematically through interpretative frameworks, analytical perspectives, and illustrative tutorials centered on major yet accessible datasets. Available data-generating protocols and analysis pipelines adopted by early experimental programs permit introductory exploration of the field using analysis-ready data. (Prost et al., 2020).

Sampling bias permeates human genomics studies, driving misinterpretation and limiting model development. While large-scale genomic data provide a unique opportunity to advance knowledge of nonmodel organisms, specific population and metadata remain largely undocumented. Access snapshots of major datasets permit high-level description of Human Genome Project, 1000 Genomes Project, Genome Aggregation Database, UK Biobank, and Database of Genotypes and Phenotypes multispecies data in relation to their associated studies. Opening a window of opportunity to reflect on these data is framed succinctly by a description of data availability, consent frameworks, Data Access Committees, and data-sharing policies governing the 1000 Genomes Project and UK Biobank, the two pioneering studies of human genomics.

Data sources and cohort description

A comprehensive characterisation of sequencing data sources and cohorts enhances awareness of the availability and scope of large sequencing datasets in the public domain. The datasets are presented within the context of different phenotypes and study designs, thereby establishing a critical base for understanding their suitability to the desired analysis, the nature of potential sampling bias, and representation across diverse populations. The UK Biobank comprises one of the largest available genotype and phenotype datasets for a population-based study. This biomedical resource gathers genetic information via genome-wide genotyping for 500,000 volunteers aged 40-69 years across Mainland UK. The cohort collection follows an opt-in system that facilitates sharing of fully anonymised data and prohibits their release to third parties. A Data Access Committee establishes the credentials of those proposing to work with the data, whether researchers or students with supervision, and guarantees that datasets are used strictly in accordance with the established protocol and within the bounds detailed in the Item. The UK Biobank dataset remains a high-quality resource suitable for population genetic studies and gene-discovery analyses (Sugolov et al., 2024) [table 2].

Table 2: Population Genomics and Large-Scale Sequencing

Component	Description	Educational Relevance	Examples / Notes
Large-scale Datasets	Genomes of diverse organisms and humans	Authentic exploration of evolution, adaptation, and gene discovery	Human Genome Project, 1000 Genomes, UK Biobank, gnomAD
Cohort Description	Population structure, phenotype coverage, consent	Helps students assess dataset suitability and bias	UK Biobank: 500k volunteers, opt-in, data access committee
Quality Control (QC)	Cleaning reads, trimming, alignment, variant calling	Teaches critical preprocessing steps and error mitigation	TRIMMOMATIC, Multi-Genome Alignment, imputation panels
Population Structure & Demography	Genetic differentiation, admixture, ancestry inference	Introduces statistical concepts of evolution and gene flow	Admixture analysis, demographic inference, correcting for hidden population structure

Diversity & Linkage Disequilibrium (LD)	Polymorphic loci, allele counts, segregation	Guides imputation accuracy and variant interpretation	Low-frequency alleles, neutrality assumptions
Genome-wide Association Studies (GWAS)	Systematic search for variants associated with complex traits	Develops skills in cross-population analysis and interpretation	Accounts for ancestry, admixture, population stratification

Quality control and preprocessing

A quality control (QC) pipeline consists of multiple steps that follow the general principle of gathering information about the raw sequencing dataset with different tools before proceeding to make the dataset useable for further analysis (Hadfield & D. Eldridge, 2014). Basic QC is done to remove erroneous data that can affect downstream analysis and to check that the sequencing process has not introduced contamination from other species and that bad adapters have not been appended to the reads and complicate downstream data interpretation. Multi-Genome Alignment of (MGA) a group of representative fastq files from each run is often a first step followed by a visual inspection of the quality scores and the percentage of bases assigned to each quality range is used to help determine how well a run performed and if any obvious run-to-run contamination occurred. Quality filtering is performed using TRIMMOMATIC which removes low quality read pairs from a read-file as well as any adapter and poly-A contamination. After trimmomatic a further probe finds and trims at the first location (H. Paszkiewicz et al., 2014) where an average base Quality score drops below 30 over a window of 3 consecutive bases and another probe checks how many bases remain after trimming. An imputation reference panel is an external dataset that can be leveraged as part of the genotype inferencing process and maximises the recovery of missing variation from the data. The DNA-SEQ variant-calling ation directly infers 0/1/2 or reference alternative counts and therefore does not lend itself to genotype plausibility checking. Genotype-phenotype harmonization checked for missing phenotype or genotype information across gwas of plots; Number of variants with pheno available before qualifying imputation (number passes QC) shown against number of individuals (number passes QC) for test A; Step that mitigate these and other such artifacts in the data are applied at the end of raw data generation and prior to secondary analysis; General characteristics from unstructured reading of fastq files and interpreted language versions are recorded for each run to understand the nature of the data being processed.

Population structure and demographic inference

Characterization of population structure is a foundational step in genomic analyses (Xu et al., 2021). Distinction among genetically differentiated groups informs studies of natural selection, admixture, and demography (Mazet et al., 2014). Admixture analysis traces the components of ancestry derived from contributing populations; such analyses can link genotype to environmental context in the study population (Siu et al., 2012). These measures complement other analyses by enabling interpretation of the population-genotype relationship. Demographic inference from data collected on single populations is hampered by incomplete representation of evolutionary history. Admixture may erase traces of ancestral population structure and dynamics relevant to gene flow, yet uncollected populations could also be a principal source of ancestral signals. Analyses of publicly available datasets allow consideration of collaboratory complex models that generate competing signals from both history-related events and migration-driven dynamics. Population-genotype linkage can lead to genome-wide association study (GWAS) dataset-wide selection of contributing cohorts.

Diversity, linkage disequilibrium, and imputation

Many large-scale sequencing datasets contain genotype calls from multiple populations, and analyses should account for population structure and the relationships among related individuals. Within a population, diversity can be measured in several complementary ways, including the number of polymorphic loci and the number of alleles per polymorphic locus. To the extent that low-frequency alleles remain segregating in a population, neutrality holds more strongly and more loci in intermediate-frequency ranges remain associated with historical events. These large-scale population samples provide an opportunity to assess the imputation of missing genotypes and alleles not present in the sample itself. The accuracy of the imputation can depend on the diversity of the population compared to the reference; if the sampled population is too divergent, it may contain too many low-frequency variants (P. Bilton et al., 2018) ; Fox et al., 2019)

Analytical frameworks for authentic analysis

Defining analytical frameworks helps students engage with authentic datasets, enhances reproducibility amid large-scale production, and guides method selection according to data properties and biological questions. Students analyze samples from diverse cohorts, independent studies, and time points, and evaluate lineage-informative variants across genome-wide association studies (GWAS) or collaborative framework models. They characterize gene-discovery analyses by contrasting GWAS enrichment of signals for adaptation, co-expression, and collaboration between genic and non-genic sites. Frameworks for gene discovery and population analysis of selection in individuals and populations are under active development. With recommendations for generalisation beyond specific examples, enhancement of existing methods to accommodate growing datasets, and promotion of best practices for sharing authentic analyses of large-scale datasets, population-genetics remains a strong basis for pedagogical strategy.

Genome-wide association studies in diverse populations

Genome-wide association studies conduct a systematic search for genetic variants associated with complex traits in large cohorts. The majority of such studies have focused on populations of predominantly European ancestry. Because genetic variation reflects demographic history, including migration and selection pressure, the patterns of association may differ between populations. The framework for GWAS proposed by (Sugolov et al., 2024) accommodates such variability, allowing genotype and phenotype datasets from diverse populations to be analyzed together without bias or loss of information. Studies in humans have shown that, in general, allele frequency differences are more pronounced between admixed populations sharing ancestry than between nonadmixed ones (Mas Montserrat et al., 2020). Summary statistics derived from the structure-admixture model such as the allele frequency shift between any two populations or the total number of different alleles shared by them can be employed to assess population stratification and hidden admixture. These signals can be extracted from any previously computed GWAS without additional genotype data.

Approaches that consider the characteristics of rare variants at a population level require only data from the population of interest for analysis, making them more suitable for underrepresented groups (Raska & Zhu, 2011). Nucleotide diversity, the proportion of variable sites, and genotype frequencies derived from the reference panel can be used to predict experimental output signals such as positive association weight on gene-contribution tests (Abdurakhmanov, J., et al). Although different populations may contain distinct variants, existing information about rare variants can still guide downstream analysis and experimental validation.

Rare variant and burden testing

Methods for testing associations with rare genetic variants and genes with multiple rare variants have gained prominence in candidate-gene sequencing studies. In large-scale international collaborations, the collection of samples with rare diseases has advanced understanding of population-structure effects and gene variation. A popular approach is to compute an aggregate statistic and test for association with common-disease, loss-of-function variants or functional constraints. The rarity of variants in these frameworks may be misleading, especially for non-coding positions or low-frequency variants for which selection remains active. Many methods aggregate rare variants to observe if a higher number occurs in cases than expected. Although the use of a burden of rare variants (Wang, Li, & Hakonarson, 2010) under a recessive model reduces statistical degrees of freedom, incomplete knowledge of mutational mechanisms discourages reliance on such models (Mägi et al., 2011). Instead, a collapsing approach counts only observed rare coding variants in genes from an external gene set. The Gene-based Adaptive Rare-variant Test (GART) (Li & Leal, 2008; Luo, Liu, & Wang, 2011) implements this strategy within a liability framework. A common second step, widely used since the advent of the 1000 Genomes Project, combines individual variant data into a trait-specific gene-based statistic (MAF < 1%, gene-based) from the public database (Luedtke et al., 2011).

Time-series and ancestry-informed analyses

Temporal analyses have revealed considerable variation in selection intensity across geographical regions (Kelleher et al., 2019). Moreover, coupled with evolutionary metrics and models of selection coalescence, such analyses have facilitated population history reconstruction from allele frequencies (Francesco Palamara, 2014). Such approaches can consolidate information from geographically distinct yet shared alleles and elucidate the evolutionary trajectory of single-nucleotide variants after their introduction into admixed populations and their diffusion among distinct ancestry classes (Abdurakhmanov, J., et al). These insights can enhance appreciation of gene dispersion patterns and population-interaction networks, contributing significantly to understanding population evolution and gene-phenotype associations. Emerging data have demonstrated intensity modulation of selection acting on admixed variants during confounding population transitions, co-determining their persistence and spread across the genome within diverse populations. Awareness of the spatiotemporal dynamics governing dissemination of gene-regulatory variants, such as those modulating extra-ocular gene expression and associated traits spanning craniofacial skeletal, skin, hair, and eye pigmentation, will deepen comprehension of their contribution to gene-phenotype correlations. Coupling time-series with ancestry-informed analyses has the potential to facilitate quantitative elucidation of these dynamics.

Integrative multi-omics and functional genomics

Analyses of multi-omics data aim to elucidate gene regulation and causal relationships connecting candidate variants to phenotypes. Gene-regulatory systems and intermediary phenotypic traits thus assume significance in modelling selection and association signals within the broader analysis framework. Existing resources detailing general-purpose experimental strategies, throughout analysis, document retained datasets generated during the educational programme. Publicly accessible genetic, transcriptomic, and epigenomic datasets aid the construction of data-comprehensive analysis pipelines based on these programmable procedures.

Accumulation of extensive sequence-variant data across human populations, species, and organisms fuels efforts to identify candidate genes and causal variants underlying population shifts and adaptive evolution (Prost et al., 2020). General systems biology approaches already characterise the regulatory network connecting genotype to phenotype at cellular and organism levels. Pursuit of gene candidates and pairs connecting selected variants to selected phenotypes further promotes integration of diverse datasets and multi-omics analysis. Multiple, state-of-the-art statistical techniques, subject to varying degrees of ill conditioning and high-dimensionality, bolster econometric models.

Computational pipelines and reproducibility

Computational pipelines facilitate analyses across diverse fields and application domains. This teaching framework emphasizes the significance of reproducibility, encouraging responsible stewardship over data through principled structured programming and transparent computational processes. Reliable methods build confidence in scientific conclusions and promote thoughtful use of data-essential considerations for researchers working with sensitive human materials and genomic information. Data stewardship comprises governance, privacy, consent, and responsible use informed by available guidelines and resources (J Garcia et al., 2022). Data-sharing protocols are defined across institutions and consortia, and downstream analyses often remain accessible via experimental or computational repositories such as the European Nucleotide

Archive. Data sets should also be treated as proprietary during educational use, prohibiting dissemination or display alongside published commands and results. Educational practices must therefore balance the pedagogical utility of authentic, rich datasets with an awareness of permissible freedoms and constraints.

Software design principles encourage careful structuring, documentation, and commenting of code throughout the development process (Hussain Ather et al., 2020). Artifact-free base calls from a high-quality sequencing run are vital for accurate genotype calling; minimizing multiplexing prevents genotyping dropout in pooled or amplicon-based approaches. Versioning permits evolution, rollback, and tracing of computational histories. Containerization packages all dependencies and specifications into distribution-ready images, while environment-capture precludes installation-time discrepancies by exporting an operational environment as a descriptive file (Azimova, S., et al).

Data governance, privacy, and ethics

Educational institutions increasingly emphasize data science and statistics and the need to teach graduate students specialized knowledge and skills in population genomics and statistical learning. Teaching a heroic gene-discovery course in collaboration with the University of Portugal illustrates how to meet such needs while enabling students to analyze prestigious bioinformatics datasets in an authentic manner. International collections, furthering the timely international collaboration, house large DNA-sequence datasets pivotal for a gene-discovery study; extensive yet comprehensive protocols guide every analysis step while catering to diverse scientific scopes and programming levels. Data governance and ethics are paramount in genomic data analysis. As these aspects increasingly inform funding proposals, institutional policies, and scientific articles, therefore, addressing data governance and privacy protections is crucial for students aiming to publish analysis findings. Genomic data represent rich resources but raise urgent privacy concerns because gene sequences constitute sensitive information (Jafarbeiki et al., 2022). Consequently, adopting a precise distinction between General Data Protection Regulation (GDPR) data and non-GDPR data is vital. Various considerations characterize both classes. Shared without consent, analysis cannot uncover sensitive information in non-GDPR data, barring data leakage. Completed consent frameworks govern use of additional genomic data. Genomic datasets maintained under broad access further support broadened analysis scopes without implying remote storage (Vía, 2017).

Software design, versioning, and containerization

When designing software to formalize and document data-processing steps, it is essential to distinguish between the requirements for routine data exploration and those for more complex analyses that will be shared according to the principles of data governance. The former typically requires limited iterations at the expense of rigor, while the latter necessitates adherence to higher standards of version control, provenance tracking, and complete documentation in order to establish a clear and reproducible history of the analysis and promote responsible use of the data (Lee et al., 2017). Version-control systems such as git and platforms such as GitHub or GitLab are indispensable for both types of task, allowing capture of the software environment at any point in the project life cycle, promotion of collaboration through systematic peer review, and straightforward generation of fully self-contained software archives for submission to data or software repositories. Energy conservation, functionality, and user-friendliness should govern the choice of software to complement a programming language such as R or Python. Among the most powerful solutions applicable to both types of task is Snakemake, a workflow-management system built on Python that incorporates workflow description, parallel execution, reproducible environments, and provenance tracking, all in a highly accessible form (Kim et al., 2017). Because Snakemake facilitates both capture of environment information and generation of stand-alone archives, it is possible to prepare a Snakemake workflow in parallel with exploration of an authentic analysis on data governed by consent, open the exploration to collaboration on external data governed by more stringent controls, and subsequently transfer the exploration to a formal pipeline equipped with large-scale and multi-omic data while retaining all of the Snakemake infrastructure.

Documentation, pipelines, and workflow management

As high-throughput sequencing technologies continue to advance and produce massive amounts of next-generation sequencing (NGS) data, a method for processing data with high-throughput sequencing is required (Ziyaev, A. A., et al). Genomics has developed and launched a pipeline system called RNA CoMPASS, which stands for RNA Comprehensive Multi-Processor Analysis System, for efficient processing of RNA-seq data. RNA CoMPASS is easily accessible via a web-based graphical interface and is capable of processing high-throughput sequencing data from different instruments. The pipeline employs a distributed computational environment to manage data files and run jobs in parallel, thus reducing the overall job completion time (Sasmakov, S. A., et al). RNA CoMPASS includes steps for both endogenous and exogenous analyses, including steps for transcriptome quantification, genome mapping, exon-intron expression level calculation, de novo assembly, virus detection, and contamination screening. The Graphical User Interface (GUI) of Only RNA CoMPASS provides users with an easy, graphical way to create, maintain, and configure the various components of the RNA CoMPASS pipeline while monitoring the status of those components. The GUI supports a web-based interface to deliver RNA CoMPASS access to any workstation on-site or beyond (Xu, 2012).

Teaching strategies and assessment

To enact the vision for transformative education and promote authentic analysis of large-scale sequencing datasets, the strategy outlines evidence-based instructional approaches that incorporate active methods of problem-based learning and peer instruction, focusing on peer interaction and student motivation (Prost et al., 2020). Activities align with the identified learning outcomes, and assessment reinforces desirable traits of scientific practice. Undergraduates construct a population-genomics GWAS analysis using diverse, authentic datasets (6.1). Complementary classroom exercises support exploration of population structure and variant-mapping strategies, and training in data governance and reproducible workflow design underpins ethics and data provenance (6.2). Rubrics derive from guided competencies in substantive interpretation of biological signals, thoroughness of analysis, and reproducibility and documentation; associated course objectives specify expectations around academic integrity and responsible-use principles related to sensitive data (6.3).

Curriculum design for authentic data analysis

Addressing demands for significant analytical competence and research familiarity within the evolving biology curriculum requires careful and imaginative planning and execution. The teaching approach described in this case study addresses these demands through a curriculum design that supports the authentic end-to-end analysis of large-scale genomic datasets (Ziyaev, A. A., et al). The particular design articulates objectives, identifies key datasets and analysis types, and specifies learning outcomes and assessment-which serve as the foundation for individual modules responsive to diverse educational needs, including high school, undergraduate, and graduate levels (Azimova, S., et al). The curriculum design aims to develop independent analytical skills through engagement with authentic datasets generated by numerous population-wide studies. Demand for such skills is reflected in recent surveys (Yang et al., 2017). Attention to the procedures and rationale for authentic datasets illustrates the analysis process while highlighting standard software specifications and consideration of socioeconomic factors that enable such studies. These supplementary materials may be progressively integrated during the teach-back approach described elsewhere (Lynn Petrie & Xie, 2021).

Classroom activities and lab exercises

Authentic analysis of large-scale datasets typically requires sophisticated programming skills that learners may not possess. Students often lack immediate context for interpreting rationale, choices, and operational details when exposed to incompletely reproduced workflows from scientific articles, hindering full absorption (Lynn Petrie & Xie, 2021). Consequently, well-structured classroom activities and laboratory exercises that imitate authentic analysis and experimental validation steps from a real study on a familiar dataset may facilitate comprehension and retention. Figshare datasets are well suited for such activities. The Challenge Data Set of the 1001 Genomes Project features 575 accessions, with both phenotype and genotype files available to enable trait–genotype mapping in diverse crops. The corresponding manuscript and accompanying analyses offer a clear, accessible parallel. Many exercise outlines for the Challenge Data Set exist, but additional laboratory exercises are further proposed. The Hadley Centre and World Bank climate time series datasets together serve as a complementary focus on temporal variation. The Glacial–Interglacial Exercise describes time-course analysis associated with these data. Detailed outlines of classroom exercises designed to parallel and incorporate the Challenge Data Set are likewise provided. The Glacial–Interglacial Exercise, coupled with the requisite data files, represents an additional resource. Enabling instructors or facilitators to conduct collaborative elaboration while customizing content, each outline is intentionally flexible (Sasmakov, S. A., et al).

Evaluation rubrics and learning outcomes

Competence in interpreting scientific research, developing reproducible computational analyses, and applying ethical principles in data-generating, analysis, and sharing constitutes a foundation for responsible population genomics. These skills align with the Australasian Society for Human Genetics capabilities at the undergraduate level, which encompass an ability to interpret the scientific literature and the consequences of research, to apply disciplinary technical knowledge and rationale to a proposed study, and to engage in a range of approaches to data analysis. By the end of the teaching sequence, students will be able to describe statistical evidence for selection and association in genomic data, outline the practical and ethical issues that arise in, and comment on the information provided by, public data-sharing initiatives; and construct a modular, reproducible computational workflow that implements an authentic analysis of genomic data. A 30-hour, problem-based learning module (Pérez-Losada et al., 2020) introduces end-to-end computational analysis of population genomic data by having students implement authentic analyses of publicly available sequencing datasets. Following hands-on practice with example datasets, they investigate selection and gene–phenotype association within their cohort, select appropriate experimental validation approaches, and develop a modular workflow that incorporates data acquisition, quality control, analysis, and visualisation. A hybrid approach employs course-enriched datasets (Yang et al., 2017) alongside solutions to scaffold and direct participants, enabling the adoption of authentic datasets while maintaining a focus on rigor and reproducibility.

Ethical, legal, and societal implications

Human genomics can offer significant insights into fundamental biological processes and a deeper understanding of the gene-phenotype relationship through large-scale population genomic analyses. These analyses remain critical for advancing the scientific and societal impacts of population genomics; yet, they can pose ethical, legal, and societal risks, particularly if analytical frameworks and data provenance are not properly addressed (De Cristofaro, 2013). Collectively, the targeted datasets encompass phenotypes and corresponding analytical frameworks applied across multiple diverse populations, enabling users to characterize datasets from a population-genomics perspective. Providing access to extensive, complex dataset collections while promoting educational values such as compliance, ownership, accountability, stewardship, and governance remains a challenge in public education (Karimov, N., et al. 2025). These values are paramount for influential population-genomic analyses designed to guide safe gene-editing technologies and to ensure trust, collaboration, and responsible precision-disease treatment while addressing bioethics. Well-characterised population-genomics datasets, fully compliant with public-access regulations, have become available; however, framework-guideline literature specifically addressing population compliance remains scarce within the field (Shamsudinova, I., et al).

Human genomes remain capable of generating security, privacy, and ethical dilemmas for entire communities. A study noted that informing individuals and future generations that personal genomic data is accessible can impede research participation, especially in crucial flagship initiatives. Such property models may even dissuade sharing for altruistic purposes, when participants intend to support the greater good by enabling advances in genomic medicine. Abounding discrepancies exist between community perceptions and formal security guarantees, whereby knowledgeable parties continue to engage in genome access despite the existence of those guarantees. Large-scale sequencing datasets from human populations demand distinctive ownership, governance, and intellectual-property models. As a knowledge-intensive and evolving field, both science itself and scientists lag behind in adapting to such changes ((Benjamin) Capps et al., 2019).

Conclusion

The opportunity to engage students in authentic analysis of population-genomic datasets aligns with multiple pedagogical objectives, including experiential learning, active learning, and the development of analytical skills, computational literacy, critical thinking, and scientific reasoning (Yang et al., 2017). The prevailing model for teaching complex, heterogeneous, high-dimensional datasets-elaborated in detail elsewhere-invokes an inquiry-based approach to guide students through population genomics and statistical genetics from arrival-departure metadata analysis to refined analytic investigation of high-dimensional genomic, epigenomic, and transcriptomic datasets (Prost et al., 2020).

References:

1. Yuldashev, A. G. (2024). Anthroponyms in the Uzbek worldview. Vestnik Sankt-Peterburgskogo Universiteta. Vostokovedenie i Afrikanistika, 16(2), 474–484.
2. Lutfullaeva, D. E., & Yuldashev, A. G. (2023). The peculiarities of defining culturally specific Uzbek names in associative dictionaries. Vestnik Sankt-Peterburgskogo Universiteta. Vostokovedenie i Afrikanistika, 15(3), 485–496.
3. Prost, S., Winter, S., De Raad, J., T F Coimbra, R., Wolf, M., A Nilsson, M., Petersen, M., K Gupta, D., Schell, T., Lammers, F., & Janke, A. (2020). Education in the genomics era: Generating high-quality genome assemblies in university courses. ncbi.nlm.nih.gov
4. Yang, X., R. Hartman, M., T. Harrington, K., M. Etson, C., B. Fierman, M., K. Slonim, D., & R. Walt, D. (2017). Using Next-Generation Sequencing to Explore Genetics and Race in the High School Classroom. ncbi.nlm.nih.gov
5. Gao, Z. (2024). Unveiling recent and ongoing adaptive selection in human populations. ncbi.nlm.nih.gov
6. Udpa, N., Zhou, D., G. Haddad, G., & Bafna, V. (2011). Tests of Selection in Pooled Case-Control Data: An Empirical Study. ncbi.nlm.nih.gov
7. K. Oleksyk, T., W. Smith, M., & J. O'Brien, S. (2010). Genome-wide scans for footprints of natural selection. ncbi.nlm.nih.gov
8. Sifrim, A., KJ Van Houdt, J., Tranchevent, L. C., Nowakowska, B., Sakai, R., A Pavlopoulos, G., Devriendt, K., R Vermeesch, J., Moreau, Y., & Aerts, J. (2012). Annotate-it: a Swiss-knife approach to annotation, analysis and interpretation of single nucleotide variation in human disease. ncbi.nlm.nih.gov
9. Sugolov, A., Emmenegger, E., D. Paterson, A., & Sun, L. (2024). Statistical Learning of Large-Scale Genetic Data: How to Run a Genome-Wide Association Study of Gene-Expression Data Using the 1000 Genomes Project Data. ncbi.nlm.nih.gov
10. Hadfield, J. & D. Eldridge, M. (2014). Multi-genome alignment for quality control and contamination screening of next-generation sequencing data. ncbi.nlm.nih.gov
11. H. Paszkiewicz, K., Farbos, A., O'Neill, P., & Moore, K. (2014). Quality control on the frontier. ncbi.nlm.nih.gov
12. Xu, Y., Liu, Z., & Yao, J. (2021). An Eigenvalue Ratio Approach to Inferring Population Structure from Whole Genome Sequencing Data. [PDF]
13. Mazet, O., Rodríguez, W., & Chikhi, L. (2014). Demographic inference using genetic data from a single individual: separating population size variation from population structure. [PDF]

14. Siu, H., Jin, L., & Xiong, M. (2012). Manifold Learning for Human Population Structure Studies. ncbi.nlm.nih.gov
15. P. Bilton, T., C. McEwan, J., M. Clarke, S., Brauning, R., C. van Stijn, T., J. Rowe, S., & G. Dodds, K. (2018). Linkage Disequilibrium Estimation in Low Coverage High-Throughput Sequencing Data. ncbi.nlm.nih.gov
16. Mas Montserrat, D., Kumar, A., Bustamante, C., & Ioannidis, A. (2020). Addressing Ancestry Disparities in Genomic Medicine: A Geographic-aware Algorithm. [PDF]
17. Raska, P. & Zhu, X. (2011). Rare variant density across the genome and across populations. ncbi.nlm.nih.gov
18. Mägi, R., Kumar, A., & P Morris, A. (2011). Assessing the impact of missing genotype data in rare variant association analysis. ncbi.nlm.nih.gov
19. Luedtke, A., Powers, S., Petersen, A., Sitarik, A., Bekmetjev, A., & L Tintle, N. (2011). Evaluating methods for the analysis of rare variants in sequence data. [PDF]
20. Kelleher, J., Wong, Y., W. Wohns, A., Fadil, C., K. Albers, P., & McVean, G. (2019). Inferring whole-genome histories in large population datasets. ncbi.nlm.nih.gov
21. Francesco Palamara, P. (2014). Population Genetics of Identity By Descent. [PDF]
22. J Garcia, B., Urrutia, J., Zheng, G., Becker, D., Corbet, C., Maschhoff, P., Cristofaro, A., Gaffney, N., Vaughn, M., Saxena, U., Chen, Y. P., Benjamin Gordon, D., & Eslami, M. (2022). A toolkit for enhanced reproducibility of RNASeq analysis for synthetic biologists. ncbi.nlm.nih.gov
23. Hussain Ather, S., Igbagbo Awe, O., J. Butler, T., Denka, T., Andrew Semick, S., Tang, W., & Busby, B. (2020). SeqAcademy: an educational pipeline for RNA-Seq and ChIP-Seq analysis. ncbi.nlm.nih.gov
24. Jafarbeiki, S., Gaire, R., Sakzad, A., Kasra Kermanshahi, S., & Steinfeld, R. (2022). Collaborative analysis of genomic data: vision and challenges. [PDF]
25. Vía, M. (2017). Big Data in Genomics: Ethical Challenges and Risks. [PDF]
26. Lee, T. R., Mo Ahn, J., Kim, G., & Kim, S. (2017). IVAG: An Integrative Visualization Application for Various Types of Genomic Data Based on R-Shiny and the Docker Platform. ncbi.nlm.nih.gov
27. Kim, B., Ali, T., Lijeron, C., Afgan, E., & Krampis, K. (2017). Bio-Docklets: virtualization containers for single-step execution of NGS pipelines. ncbi.nlm.nih.gov
28. Xu, G. (2012). RNA CoMPASS: RNA Comprehensive Multi-Processor Analysis System for Sequencing. [PDF]
29. Lynn Petrie, K. & Xie, R. (2021). Resequencing of Microbial Isolates: A Lab Module to Introduce Novices to Command-Line Bioinformatics. ncbi.nlm.nih.gov
30. Pérez-Losada, M., M. Crandall, K., & A. Crandall, K. (2020). Testing the "Grandma Hypothesis": Characterizing Skin Microbiome Diversity as a Project-Based Learning Approach to Genomics. ncbi.nlm.nih.gov
31. De Cristofaro, E. (2013). An Exploratory Ethnographic Study of Issues and Concerns with Whole Genome Sequencing. [PDF]
32. (Benjamin) Capps, B., (Ruth) Chadwick, R., (Yann) Joly, Y., (Tamra) Lysaght, T., (Catherine) Mills, C., (John J.) Mulvihill, J. J., & (Hub) Zwart, H. A. E. (2019). Statement on bioinformatics and capturing the benefits of genome sequencing for society. [PDF]
33. Shamsudinova, I., et al. (2025). Educational disparities in the digital era and the impact of information access on learning achievements. Indian Journal of Information Sources and Services, 15(1), 6–11.
34. N .Saranya. (2025). IoT-Integrated Mobile Learning Platforms Using Cloud Infrastructure: A Scalable Architecture for Smart Education. *Journal of Wireless Sensor Networks and IoT*, 3(1),
35. Karimov, N., et al. (2025). The impact of Islamic libraries on the compilation and dissemination of Hadith. Indian Journal of Information Sources and Services, 15(1), 183–187.
36. Abdurakhmanov, J., et al. (2023). Cloning and expression of recombinant purine nucleoside phosphorylase in the methylotrophic yeast Pichia pastoris. Journal of Advanced Biotechnology and Experimental Therapeutics. https://doi.org/10.5455/jabet.2023.d153

37. Ziyaev, A. A., et al. (2023). Synthesis of S-(5-aryl-1,3,4-oxadiazol-2-yl) O-alkyl carbonothioate and alkyl 2-((5-aryl-1,3,4-oxadiazol-2-yl)thio) acetate, and their antimicrobial properties. Journal of the Turkish Chemical Society, Section A: Chemistry. https://doi.org/10.18596/jotcsa.1250629
38. Azimova, S., et al. (2023). Study of the immunogenicity of combination of recombinant RBD (Omicron) and nucleocapsid proteins of SARS-CoV-2 expressed in Pichia pastoris. The Open Biochemistry Journal. https://doi.org/10.2174/011874091x273716231122102205
39. Sasmakov, S. A., et al. (2021). Expression of recombinant PreS2-S protein from the hepatitis B virus surface antigen in Pichia pastoris. VacciMonitor, 30(1), 27–32.