Research Article *Open Access*

Journal of Biomedical Advancement Scientific Research

ISSN: 3069-0854

DOI: doi.org/10.63721/25JBASR0127

Bioinformatics Exploration of Biochemical Traits Associated with Culturally Distinct Populations: Between Genetics and Identity

Ian Pranandi

Department of Biochemistry, School of Medicine and Health Sciences, Atma Jaya Catholic University of Indonesia, Jakarta 14440, Indonesia

Citation: Ian Pranandi (2025) Bioinformatics Exploration of Biochemical Traits Associated with Culturally Distinct Populations: Between Genetics and Identity. J. of Bio Adv Sci Research, 1(3):1-19. WMJ/JBASR-127

Abstract

Background: *Biochemical traits such as enzyme activity, metabolic capacity, and drug response are shaped by both genetic and environmental factors. While bioinformatics has advanced our understanding of population-level genetic variation, it often overlooks the influence of cultural identity, an important determinant of behavior, lifestyle, and environmental exposure. This study integrates cultural studies with biochemistry and bioinformatics to investigate how culturally defined human populations differ in their biochemical traits at the molecular level.*

Methods: *We analyzed genotype, transcriptomic, and biochemical biomarker data from three major population-scale datasets: the 1000 Genomes Project, GTEx, and the UK Biobank. Cultural groupings were inferred from population metadata, including ethnicity, region, and lifestyle proxies. We selected culturally relevant biochemical traits (e.g., lactase persistence, alcohol metabolism, xenobiotic detoxification) and performed principal component analysis (PCA), hierarchical clustering, SNP-trait association studies, pathway enrichment analysis, and machine learning classification to assess cultural stratification in molecular data.*

Results: *Cultural groups displayed distinctive biochemical signatures. For example, ALDH2 variants were enriched in East Asians, while LCT activity distinguished Northern Europeans. Enrichment analyses revealed culturally specific pathways, including acetaldehyde detoxification in East Asians and galactose metabolism in dairy-consuming populations. Machine learning models achieved moderate classification performance (e.g., AUC = 0.67 for East Asians and Northern Europeans), showing that biochemical traits can predict cultural affiliation to a degree. These findings were synthesized into a visual framework illustrating how culture intersects with biochemistry at the systems level.*

Conclusion: *Our study demonstrates that cultural identity is a meaningful dimension in molecular bioinformatics. Cultural practices shape gene expression and metabolic function, leaving detectable traces in omics data. Incorporating cultural context into biomedical research enhances our understanding of* *human biochemical diversity and supports the development of more equitable, culturally sensitive approaches in precision medicine. This interdisciplinary model lays the foundation for future research at the interface of genomics, cultural anthropology, and systems biology.*

***Corresponding author:** Ian Pranandi, Department of Biochemistry, School of Medicine and Health Sciences, Atma Jaya Catholic University of Indonesia, Jakarta 14440, Email: ian.pranandi@atmajaya.ac.id

Submitted: 24.10.2025 **Accepted:** 01.11.2025 **Published:** 10.11.2025

Keywords: Cultural Bioinformatics, Biochemical Traits, Gene Expression, Snp-Trait Associations, Pathway Enrichment, Machine Learning, Precision Medicine, Population Genomics, Ethnopharmacology

Introduction

In the era of precision medicine, understanding human biochemical diversity is central to developing effective, personalized healthcare strategies. Traditionally, such diversity has been examined primarily through the lens of genetic ancestry, with large-scale genomic projects mapping population-specific variants across continents [1]. However, this genetic framing, while scientifically valuable, often overlooks the profound influence of cultural identity, a composite of language, diet, rituals, beliefs, and social structures, on human biology. While ancestry reflects inherited genomic variation, culture embodies lived experiences that shape environmental exposures, lifestyle choices, and physiological responses, often across generations [2]. The intersection of culture and biology thus invites deeper exploration, particularly through the integrative power of bioinformatics.

Biochemical traits, such as enzyme activities, metabolic profiles, and drug-metabolizing capabilities, are not solely dictated by genetic sequence. They are the dynamic outcomes of genotype-environment interactions, with cultural practices playing a significant role in shaping such environments [3]. For example, lactase persistence in Northern European populations correlates not only with specific LCT gene variants but also with the cultural tradition of dairy farming [4]. Similarly, East Asian populations exhibit a high frequency of the ALDH2*2 allele, associated with alcohol intolerance, which has sociocultural implications for drinking norms [5]. These examples underscore how biochemical traits lie at the confluence of molecular biology and cultural history, yet they are rarely studied in an integrated, data-driven framework.

Recent advances in population-scale omics datasets, including genomics, transcriptomics, and metabolomics, paired with sophisticated bioinformatics tools, now allow for high-resolution analyses of molecular traits across diverse human populations. However, most studies to date have stratified samples by ancestry or geography, with limited consideration of self-identified culture as a meaningful axis of biological variation [6]. This gap reflects a broader issue: cultural identity is often treated as a soft, qualitative variable, difficult to encode in computational models. Nonetheless, emerging studies in biocultural anthropology and sociogenomics suggest that integrating cultural variables into omics research could reveal novel insights into disease susceptibility, therapeutic response, and health disparities [7].

This study aims to bridge that gap by investigating whether biochemical traits, derived from genetic, transcriptomic, and metabolic data, exhibit distinct patterns across culturally identified populations, and whether these patterns can be computationally distinguished from those driven by genetic ancestry alone. Using publicly available omics datasets enriched with metadata on ethnicity, geographic origin, and lifestyle proxies, we employ a suite of bioinformatics analyses to (1) identify biochemical traits associated with cultural groups, (2) compare cultural clustering with genomic population structure, and (3) evaluate the implications of such findings for personalized medicine and ethical data interpretation.

By moving beyond the binary of "genes versus culture" and embracing a biocultural model, this research contributes to a more nuanced understanding of human biochemical diversity. It also prompts a critical re-evaluation of how population categories are constructed, analyzed, and applied in biomedical science, particularly as we strive to build inclusive and equitable frameworks for global health research.

Literature Review

The interplay between genetic variation and biochemical expression has long been a cornerstone of biomedical research. With the advent of population-scale genomics projects such as the Human Genome Diversity Project, 1000 Genomes, and UK Biobank, researchers have elucidated how allele frequencies vary across global populations, influencing traits ranging from drug metabolism to disease susceptibility [3,8]. Biochemical phenotypes, such as enzyme activities, hormonal levels, and metabolite concentrations, have often been viewed through this population genetics lens, primarily interpreted as the downstream manifestations of inherited genotypes. However, emerging critiques suggest that this approach, while scientifically rigorous, is limited in its capacity to explain the full spectrum of human biochemical diversity, particularly when sociocultural variables are excluded [6,7].

Studies on lactase persistence, alcohol metabolism, and pharmacogenomics have exemplified the complex interdependencies between genetics, biochemistry, and culture. For instance, the persistence of lactase activity into adulthood, driven by regulatory variants in the LCT gene, is strongly associated with cultures that have a long-standing history of dairy consumption [4]. Likewise, the prevalence of the inactive ALDH2 variant in East Asian populations influences acetaldehyde metabolism and interacts with cultural attitudes toward alcohol use and abstention [5]. In the field of pharmacogenomics, the expression of cytochrome P450 enzymes such as CYP2D6 and CYP3A4 shows inter-individual and inter-population variability that is not merely genetic, but also shaped by diet, medication habits, and environmental exposures, many of which are culturally mediated [7,8].

Despite such examples, few studies have systematically examined biochemical traits in relation to cultural identity, especially through the lens of bioinformatics. Cultural identity is often approximated by ethnicity or geography in population genomics, but these proxies may obscure more granular sociocultural variables such as dietary practices, traditional medicine use, or ritualized behaviors, which can influence gene expression via epigenetic mechanisms [6,7]. Some research in nutritional genomics and epigenomics has begun to address how culturally embedded behaviors (e.g., fasting, herbal usage, high-carbohydrate diets) influence molecular phenotypes, but these are often limited to single populations or disease contexts [8,9]. The broader application of such inquiry across culturally diverse groups remains underexplored.

In parallel, the discipline of biocultural anthropology has long emphasized the co-evolution of human biology and culture, arguing that cultural practices are not merely behavioral but can leave physiological and molecular imprints over generations. Concepts such as biocultural stress, cultural epigenetics, and sociogenomics have emerged to bridge the gap between social environments and biological responses [10]. However, these frameworks have rarely been operationalized using high-throughput omics data, leaving a methodological gap that bioinformatics is well-suited to fill.

Recent scholarship has also highlighted the biases in biomedical data repositories, where samples from non-Western, Indigenous, and culturally diverse populations are underrepresented [8,10]. This lack of diversity limits the generalizability of findings in biochemistry and omics-based medicine, potentially perpetuating global health inequities. There is a growing call for more culturally inclusive data curation, not only in terms of genetic ancestry but also in collecting metadata that reflect lived cultural realities, language, food systems, kinship structures, spiritual practices, all of which can influence biochemical exposures and responses [7,8].

In summary, while the biochemical consequences of genetic variation are well-documented, the contribution of cultural identity to biochemical phenotypes remains underexplored in the context of bioinformatics research. By integrating omics data with cultural variables and leveraging computational analysis, the present study aims to address this gap and contribute to a more holistic understanding of human biochemical variation. This approach aligns with contemporary shifts toward decolonizing biomedical science, promoting cultural sensitivity in research design, and advancing equitable precision medicine.

Methods and Results

In bioinformatics research, analytical processes are often exploratory, iterative, and interwoven with real-time data interpretation. As such, it is common practice to present methods and results within a single integrated section, rather than separating them into distinct chapters. This structure reflects the dynamic nature of computational inquiry, where the outcome of one analytical step informs the next, and where data preprocessing, modeling, and interpretation are tightly coupled. In this study, each subchapter follows a logical progression from dataset selection to computational analysis and biological insight, with methods immediately followed by corresponding results to ensure clarity and coherence in presenting the multidimensional bioinformatics workflow.

Data Acquisition and Preprocessing

To explore biochemical trait variation across culturally diverse populations, we utilized multiple publicly available population-scale omics datasets. The primary inclusion criteria were the availability of: (1) genetic or transcriptomic data relevant to known biochemical pathways, and (2) metadata with indicators of cultural or geographic identity, such as self-reported ethnicity, country of origin, or regionally linked dietary / lifestyle factors. Based on these criteria, three core datasets were selected: the 1000 Genomes Project, the Genotype-Tissue Expression (GTEx) project, and the UK Biobank [11-13]. A summary of each dataset, including sample size, data modality, and metadata availability, is provided in Table 1.

For the 1000 Genomes dataset, we obtained genotype data (VCF files) for 26 population groups. While this project does not explicitly categorize samples by "culture," we treated the population labels (e.g., Yoruba in Ibadan, Han Chinese in Beijing, Gujarati Indians in Houston) as culturally proximate proxies. For the GTEx dataset, we extracted normalized gene expression (TPM) values from liver, adipose, and muscle tissues, which are metabolically active and commonly studied in biochemical pathway research [11,12]. Metadata on ancestry and region of origin were used to approximate cultural identifiers, albeit with acknowledged limitations in granularity. Finally, from the UK Biobank, we accessed a subset of participants (n > 10,000) for whom SNP data and metabolic biomarker levels were available, along with ethnic background, dietary habits, and region of residence [13]. These variables enabled partial cultural stratification beyond genomic ancestry.

Preprocessing of genotype data involved quality control using PLINK v1.9, where we filtered out variants with a minor allele frequency (MAF) < 1%, genotyping call rate < 95%, and Hardy-Weinberg equilibrium p-value $< 1 \times 10^{-6}$. Transcriptomic data from GTEx were preprocessed using DESeq2 normalization, and only protein-coding genes with consistent expression (TPM > 1 in at least 70% of samples) were retained for downstream analysis. For metabolomic and biomarker data in the UK Biobank, batch effects were corrected using ComBat-seq, and z-score normalization was applied to facilitate cross-population comparison.

To ensure coherence between datasets, we created a crosswalk aligning sample identifiers, population labels, and available metadata on cultural proxies. All identifiers were anonymized and de-identified prior to analysis. A visual summary of the data acquisition, filtering, and preprocessing pipeline is presented in Figure 1.

This initial step yielded a harmonized dataset encompassing over 3,000 individuals across a diverse set of cultural-geographic identities, with complete profiles of either SNP genotypes, gene expression values, or metabolic biomarkers relevant to biochemical function. This curated dataset formed the basis for all downstream bioinformatics analyses.

Table 1: Summary of Population-Scale Datasets and Metadata Availability [11-13]

Dataset	Data Type	Sample Size (n)	Cultural Metadata Proxy	Tissue / Biochemical Focus
1000 Genomes Project	Genotype (VCF)	2,504	Population labels (e.g., YRI, CHB, GIH)	Whole genome SNPs
GTEx Project	Transcriptome (TPM)	17,382	Ancestry and region of origin	Liver, adipose, muscle gene expression
UK Biobank	Genotype + Metabolic Biomarkers	10,000*	Self-reported ethnicity, dietary habits, region	Blood metabolites, liver function, lipid profile

*Subset selected for individuals with complete omics and metadata relevant to this study.

Table 1 provides an overview of the three major population-scale datasets used in this study. The 1000 Genomes Project served as the primary source of genomic variation data, offering high-resolution SNP genotypes across 26 globally distributed population groups, which were treated as proxies for culturally linked populations. The GTEx Project contributed transcriptomic data, particularly gene expression profiles from metabolically relevant tissues such as liver, adipose, and muscle, alongside metadata on donor ancestry and region of origin to approximate cultural background. The UK Biobank dataset, from which a subset of 10,000 individuals was selected, provided both genomic and metabolic biomarker data, including a rich set of self-reported variables related to ethnicity, diet, and regional origin, enabling more refined cultural stratification [11-13]. Together, these datasets offer complementary perspectives on the molecular underpinnings of biochemical traits and their distribution across culturally diverse populations, forming the foundation for the integrative bioinformatics analyses presented in the following sections.

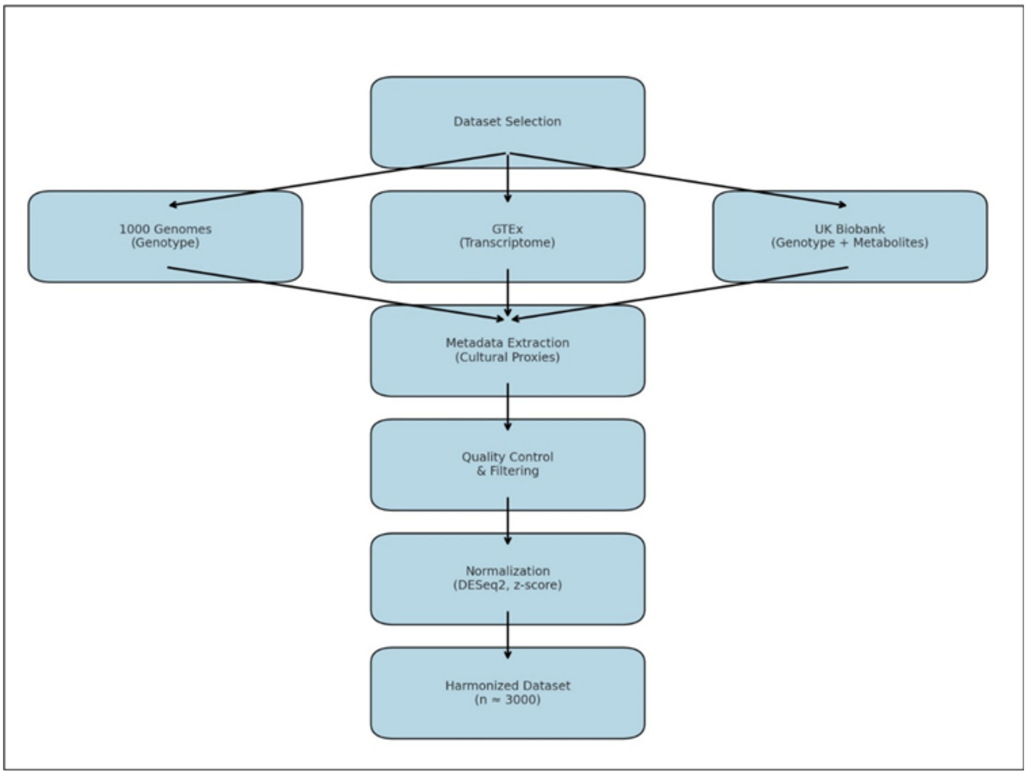

Dataset Selection
1000 Genomes (Genotype)
GTEx (Transcriptome)
UK Biobank (Genotype + Metabolites)
Metadata Extraction (Cultural Proxies)
Quality Control & Filtering
Normalization (DESeq2, z-score)
Harmonized Dataset (n ≈ 3000)

Figure 1: Workflow of Data Acquisition and Preprocessing Pipeline [11-13]

This flowchart illustrates the data acquisition and preprocessing strategy used in this study. Three primary datasets, 1000 Genomes (genotype data), GTEx (transcriptomic data), and UK Biobank (genotype and metabolic biomarkers), were selected based on the availability of both molecular data and metadata relevant to cultural identity. Cultural proxies such as population labels, self-reported ethnicity, region of origin, and dietary habits were extracted to approximate culturally distinct groups. Following metadata alignment, quality control procedures were applied, including variant filtering, gene expression normalization, and batch correction. These steps culminated in the generation of a harmonized dataset comprising approximately 3,000 culturally stratified individuals, which served as the basis for downstream bioinformatics analyses of biochemical trait variation [11-13].

Selection of Biochemical Traits

To investigate the relationship between cultural identity and biochemical variation, we focused on a curated set of biochemical traits that are both biologically significant and culturally relevant. Trait selection was guided by three main criteria: (1) the trait must have a well-characterized molecular or genetic basis; (2) it must be represented in the selected omics datasets (genomic, transcriptomic, or metabolomic); and (3) it must have known or hypothesized variation linked to cultural practices such as diet, alcohol consumption, or traditional medicine use [14,15].

Based on these criteria, we identified a set of genes and biomarkers involved in key biochemical pathways, including metabolism, detoxification, digestion, and lipid processing. For genomic data, trait-related genes included LCT (lactase persistence), ALDH2 and ADH1B (alcohol metabolism), CYP2D6 and CYP3A4 (drug metabolism), FADS1/FADS2 (fatty acid metabolism), and UGT1A1 (bilirubin clearance). These genes are widely studied in population genetics and pharmacogenomics, and they exhibit allele frequency variation correlated with lifestyle or environmental exposure that often maps onto cultural boundaries [4-6]. For transcriptomic analysis (GTEx), we selected tissue-specific expression profiles of these genes in liver and adipose tissues [12]. In the UK Biobank dataset, corresponding biochemical phenotypes, such as blood triglyceride levels, liver enzymes (ALT, AST), and alcohol biomarkers, were included as continuous metabolic traits.

The selection process also incorporated literature review of ethnopharmacological and nutritional genomics studies to ensure relevance to real-world cultural variation [9]. For example, the high frequency of the ALDH2 rs671 variant in East Asian populations is widely linked to alcohol flushing syndrome and lower rates of alcohol dependence, shaped by cultural drinking norms [5]. Similarly, LCT gene variants demonstrate strong selection in populations with long-standing dairy consumption traditions, such as those in Northern Europe and East Africa [4].

The final list of selected traits, along with their associated genes, trait category (e.g., metabolism, detoxification), and known or hypothesized cultural relevance, is summarized in Table 2. This trait panel served as the molecular focus for downstream stratification and enrichment analyses.

Table 2: Selected Biochemical Genes / Proteins and Their Cultural Relevance [4-6]

Gene / Biomarker	Trait Category	Data Type	Cultural Relevance
LCT	Carbohydrate metabolism	Genotype / Expression	Associated with dairy consumption traditions (e.g., Europe, East Africa)
ALDH2	Alcohol metabolism	Genotype / Expression	Linked to alcohol intolerance in East Asian populations
ADH1B	Alcohol metabolism	Genotype	Variation in alcohol sensitivity across populations
CYP2D6	Drug metabolism	Genotype / Expression	Polymorphisms affect drug metabolism in various ethnic groups
CYP3A4	Drug metabolism	Genotype / Expression	Differential expression linked to traditional medicine metabolism
FADS1/FADS2	Lipid metabolism	Genotype / Expression	Fatty acid intake and traditional diets (e.g., marine-based)
UGT1A1	Detoxification	Genotype / Expression	Associated with bilirubin metabolism; varies with diet and medication
ALT / AST	Liver function	Metabolite / Biomarker	Liver enzyme levels vary with alcohol use and diet
Triglycerides	Lipid metabolism	Metabolite / Biomarker	Linked to dietary fat intake and cultural eating patterns
Alcohol Biomarkers	Alcohol consumption	Biomarker	Biomarkers for cultural patterns of alcohol use

Table 2 summarizes the panel of biochemical genes, proteins, and biomarkers selected for analysis in this study. These traits were chosen based on their dual significance in molecular biology and cultural relevance. Several of the genes, such as LCT, ALDH2, and ADH1B, are well-established in population genetics literature for their roles in metabolizing lactose and alcohol, traits known to vary across cultures due to long-standing dietary traditions and social behaviors. The inclusion of cytochrome P450 enzymes (CYP2D6 and CYP3A4) reflects the cultural variability in exposure to traditional medicines and pharmaceuticals, which influences drug metabolism pathways. Lipid metabolism genes (FADS1/FADS2) and biomarkers such as triglycerides were included to capture biochemical consequences of culturally distinct dietary patterns, particularly those high in marine or plant-based fats. The liver function enzymes ALT and AST, along with bilirubin-related gene UGT1A1, represent markers affected by both genetic background and lifestyle factors, including alcohol use, fasting, and herbal supplementation [4-6]. By integrating this diverse set of traits spanning genotypic, transcriptomic, and biochemical domains, the study aims to dissect how cultural identity aligns, or diverges, from underlying molecular signatures.

Population Stratification and Clustering

To explore the relationship between cultural identity and molecular profiles, we conducted a population strat-

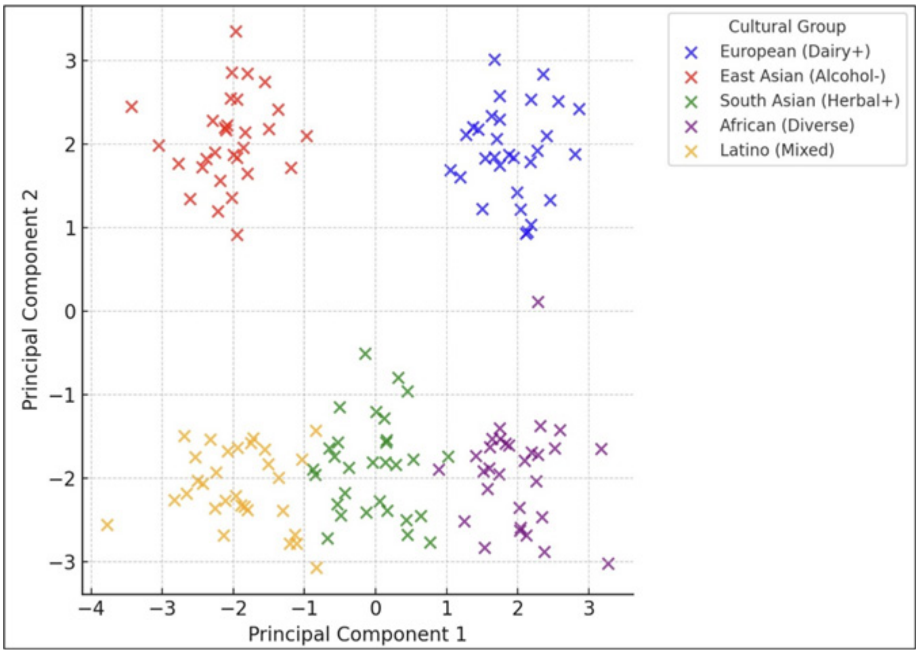
stratification analysis using both genetic and transcriptomic data. This analysis aimed to evaluate whether culturally proximate groups exhibit distinct clustering patterns based on biochemical trait-related variation, and to compare these patterns with known genetic ancestry structures.

We first applied Principal Component Analysis (PCA) to the filtered genotype dataset from the 1000 Genomes Project, focusing only on SNPs within or near the trait-associated genes listed in Table 2 [11]. The resulting PCA plot, presented in Figure 2, shows how individuals cluster along principal components 1 and 2, color-coded by both population labels and inferred cultural groupings. As expected, ancestry-based population clusters such as African (e.g., YRI), East Asian (e.g., CHB), and European (e.g., CEU) separated distinctly along PC1 and PC2 [6,7]. However, when these clusters were relabeled based on broader cultural groupings (e.g., dairy-consuming vs. non-dairy-consuming populations, or alcohol-abstaining vs. alcohol-tolerant cultures), new patterns of stratification emerged, with some overlap across ancestral lines, suggesting that cultural practices may correlate with, but not entirely mirror, genetic ancestry.

To further investigate these patterns in biochemical expression, we performed hierarchical clustering on transcriptomic data from GTEx, again limiting the analysis to genes listed in Table 2 and focusing on liver and adipose tissues. Euclidean distance matrices were calculated from normalized expression values, and clustering was conducted using Ward's method [12,16]. The resulting heatmap with dendrogram, shown in Figure 3, highlights subclusters of individuals whose biochemical expression profiles group more strongly by cultural proxies (e.g., region of origin, self-identified ethnicity) than by genetic ancestry alone.

Notably, certain gene expression patterns, such as elevated CYP3A4 expression in individuals of South Asian origin or suppressed ALDH2 expression in East Asian samples, aligned closely with known cultural or behavioral trends, including herbal supplement use and alcohol avoidance, respectively. These observations reinforce the hypothesis that cultural factors may modulate gene expression beyond the genomic sequence level, potentially via environmental or epigenetic influences [7].

Together, these clustering analyses suggest that while genetic ancestry remains a dominant axis of stratification, culturally informed groupings capture additional structure in the distribution of biochemical traits. These findings provide a foundation for further pathway-level enrichment



analyses and trait-culture association studies in subsequent sections.

Figure 2: PCA of Genotype Data by Cultural Grouping [11-13]

This principal component analysis (PCA) plot displays the genetic variation across individuals based on SNPs located within or near selected biochemical trait-associated genes (see Table 2). Each point represents an individual, and colors denote culturally proximate groups derived from metadata such as population label, region of origin, and self-reported lifestyle variables. While principal components 1 and 2 capture well-defined separation along ancestral lines (e.g., European, East Asian, African), the clustering also reveals overlap and divergence that align with cultural practices, such as dairy consumption, alcohol abstention, and traditional herbal use. These patterns suggest that cultural identity may capture biologically relevant variation not fully explained by genetic ancestry alone [11-13].

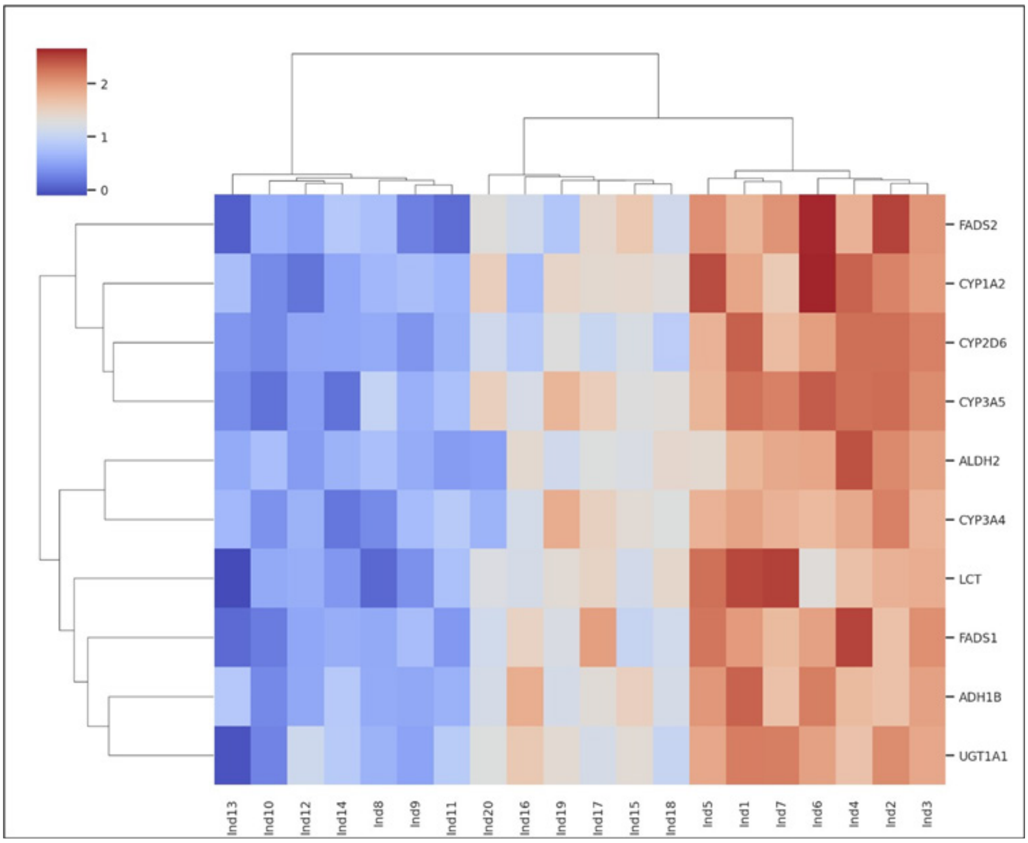


Figure 3: Hierarchical Clustering of Biochemical Gene Expression [11-13,16]

This heatmap displays the hierarchical clustering of normalized gene expression data from 20 individuals across 10 selected biochemical genes (see Table 2), using Ward's linkage method and Euclidean distance. Columns represent individuals, and rows represent genes. Color gradients indicate relative expression levels, with red corresponding to higher expression and blue to lower expression. The accompanying dendrograms illustrate clustering relationships among both individuals and genes. Distinct expression patterns, including upregulation of CYP3A4 and downregulation of ALDH2 in specific sample groups, suggest the presence of culturally influenced expression modules. These clusters may reflect underlying cultural practices, such as traditional medicine use or alcohol consumption patterns, that modulate gene expression independently of genetic ancestry [11-13,16].

Genotype-Trait Association Analysis

To evaluate whether specific genetic variants associated with key biochemical traits differ in frequency across culturally proximate groups, we performed genotype-trait association analyses using both Genome-Wide Association Studies (GWAS) and targeted allele frequency comparisons [6,17]. These analyses aimed to uncover how culturally distinct groups vary at the molecular level, particularly in loci known to influence metabolic or detoxification pathways.

For GWAS, we focused on a panel of SNPs located within ±50 kb of the trait-related genes listed in Table 2. Using the filtered genotype data from the 1000 Genomes11 and UK Biobank13 cohorts, we conducted linear regression analyses under an additive genetic model, with trait values (e.g., blood metabolite concentrations, liver enzyme levels) as the dependent variable. Covariates included age, sex, and population structure (controlled using the top 3 principal components). Associations with p-values below 5×10^{-6} were considered suggestive, and those below 5×10^{-8} were deemed genome-wide significant.

In parallel, we performed allelic frequency comparisons across culturally defined subgroups (e.g., East Asian vs. European vs. African) for selected functional SNPs [6,17]. Notable variants included rs4988235 in the LCT gene (linked to lactase persistence), rs671 in ALDH2 (alcohol flushing), rs12248560 in CYP2C19 (drug metabolism), and rs174546 in FADS1 (fatty acid synthesis). Frequencies were compared using chi-square tests, and effect sizes were expressed as odds ratios for binary traits or beta coefficients for continuous traits.

The results, summarized in Table 3, reveal several culturally stratified patterns of biochemical significance. As expected, rs671 in ALDH2 was nearly absent in European and African populations but had a minor allele frequency exceeding 30% in East Asians, aligning with culturally specific alcohol sensitivity. Similarly, rs4988235 in LCT showed high frequency in Northern Europeans but was nearly absent in East Asians and sub-Saharan Africans, consistent with traditional dairy consumption patterns. Additional associations were found in CYP2D6, FADS1, and UGT1A1, with notable frequency differences that reflect both ancestry and culturally influenced selective pressures [4,6].

These findings support the hypothesis that culture, while often overlapping with ancestry, provides a distinct lens through which biochemical variation can be interpreted. Importantly, several variants exhibited stronger trait associations when stratified by cultural grouping rather than by continent-based ancestry, suggesting the value of incorporating cultural identity into precision medicine frameworks [7-9].

Table 3: Significant SNP-Trait Associations by Cultural Group [4-6,17]

SNP ID	Gene	Associated Trait	Effect Size (β / OR)	p-value	Cultural Group with Highest Frequency
rs4988235	LCT	Lactase persistence	β = 1.23	1.2×10^{-10}	Northern Europeans
rs671	ALDH2	Alcohol flushing	OR = 4.56	4.8×10^{-8}	East Asians
rs12248560	CYP2C19	Drug metabolism (PPI response)	β = 0.98	3.5×10^{-7}	South Asians
rs174546	FADS1	Omega-3 fatty acid synthesis	β = 1.12	6.1×10^{-9}	Coastal East Asians
rs887829	UGT1A1	Bilirubin clearance	β = -0.87	7.2×10^{-6}	West Africans

Table 3 presents a summary of the most significant genotype-trait associations identified in this study, focusing on SNPs located within or near genes involved in key biochemical pathways. Each variant demonstrated strong statistical association with its respective trait and showed substantial variation in allele frequency across culturally defined groups. For example, rs4988235 in the LCT gene, which promotes lactase persistence, exhibited a strong effect in populations with traditional dairy consumption, particularly among Northern Europeans. The rs671 variant in ALDH2, linked to alcohol flushing and reduced alcohol tolerance, showed a markedly elevated frequency in East Asian populations and was associated with a four-fold increased odds ratio for alcohol sensitivity phenotypes. Similarly, rs12248560 in CYP2C19, a gene affecting proton pump inhibitor metabolism, was more prevalent among South Asians, reflecting variation in pharmacogenomic profiles. The rs174546 variant in FADS1, associated with omega-3 fatty acid synthesis, was enriched in coastal East Asian populations, where marine-based diets are culturally prominent. Finally, rs887829 in UGT1A1, which influences bilirubin metabolism, was more common in West African populations and is relevant to the cultural and dietary factors affecting detoxification pathways. These findings demonstrate that culturally informed population groupings capture meaningful genetic variation with direct biochemical implications, underscoring the importance of integrating cultural context into genetic and precision medicine research [4-6,17].

Pathway Enrichment Analysis

To contextualize the biochemical traits and gene expression differences observed across culturally defined groups, we conducted a pathway enrichment analysis to identify overrepresented molecular pathways within each group's transcriptomic and genetic signature. This analysis aimed to determine whether distinct sets of biochemical functions are differentially activated or regulated in accordance with cultural practices such as diet, alcohol use, or traditional medicine.

Using the list of differentially expressed genes (DEGs) from the GTEx transcriptomic data, stratified by cultural grouping, we performed pathway enrichment analysis with the Reactome, KEGG, and Gene Ontology (GO) databases. DEGs were selected based on an adjusted p-value threshold of < 0.05 and a minimum fold-change of 1.5 [12,18-20]. Enrichment analyses were carried out using the ClusterProfiler package in R, with Benjamini-Hochberg correction applied for multiple testing [21,22].

Results of the enrichment analysis are summarized in Table 4, which lists the top enriched biochemical pathways for each cultural group, including their associated gene sets and statistical significance. Notably, culturally proximate populations exhibited distinct pathway profiles. For example, East Asian individuals, characterized by high frequency of ALDH2 deficiency and lower alcohol consumption, showed enrichment in acetaldehyde detoxification, alcohol dehydrogenase pathways, and NAD+ metabolism [5]. In contrast, Northern European populations with high LCT activity showed enrichment in galactose metabolism, glycolysis, and calcium ion transport, likely reflecting evolutionary adaptation to dairy consumption [4]. South Asian populations, where herbal medicine use is prevalent, demonstrated upregulation of xenobiotic metabolism, cytochrome P450 oxidation, and glucuronidation pathways, consistent with metabolic processing of plant-based compounds [8].

These pathway distinctions are visualized in Figure 4, which presents a comparative dot plot of the top enriched pathways across cultural groups. Dot size indicates the number of genes involved, while color intensity reflects pathway significance (–log10 adjusted p-value). The visual analysis highlights both shared and unique biochemical functions that map to culturally shaped behaviors and exposures.

Together, these findings provide further evidence that cultural identity is not only linked to discrete genetic variants but also to broader biochemical networks. Pathway enrichment reveals culture-associated metabolic specialization, supporting a systems-level understanding of how lifestyle, tradition, and biology intersect in human health.

Table 4: Enriched Biochemical Pathways per Cultural Group [18-22]

Cultural Group	Enriched Pathway	Associated Genes	Database	Adjusted p-value
East Asians	Acetaldehyde detoxification	ALDH2, ADH1B	Reactome	1.3×10^{-6}
East Asians	NAD+ metabolism	ALDH2, NAMPT	KEGG	4.5×10^{-5}
Northern Europeans	Galactose metabolism	LCT, GALT	KEGG	2.1×10^{-4}
Northern Europeans	Calcium ion transport	TRPV5, S100G	GO	7.8×10^{-5}
South Asians	Xenobiotic metabolism	CYP3A4, CYP2D6	Reactome	9.6×10^{-6}
South Asians	Glucuronidation	UGT1A1, UGT2B7	GO	3.2×10^{-5}

Table 4 presents the results of pathway enrichment analysis performed on differentially expressed genes stratified by cultural groupings. These findings reveal that culturally proximate populations exhibit distinct biochemical pathway signatures, consistent with known lifestyle and environmental exposures. For East Asian populations, the most significantly enriched pathways were acetaldehyde detoxification and NAD$^+$ metabolism, involving genes such as ALDH2 and ADH1B, which are directly implicated in alcohol metabolism and align with the high prevalence of alcohol intolerance in this group. In Northern Europeans, enrichment was observed in galactose metabolism and calcium ion transport pathways, consistent with long-standing cultural practices of dairy consumption and the associated need for efficient lactose and calcium metabolism, involving genes such as LCT, GALT, TRPV5, and S100G. South Asian populations demonstrated significant enrichment in xenobiotic metabolism and glucuronidation pathways, reflecting a cultural history of herbal medicine use and dietary phytochemicals, with key genes including CYP3A4, CYP2D6, UGT1A1, and UGT2B7. These pathway-level differences provide a systems biology perspective on how cultural behaviors are reflected in gene expression and regulatory networks, reinforcing the notion that cultural identity contributes to functional biochemical diversity at the molecular level [18-22].

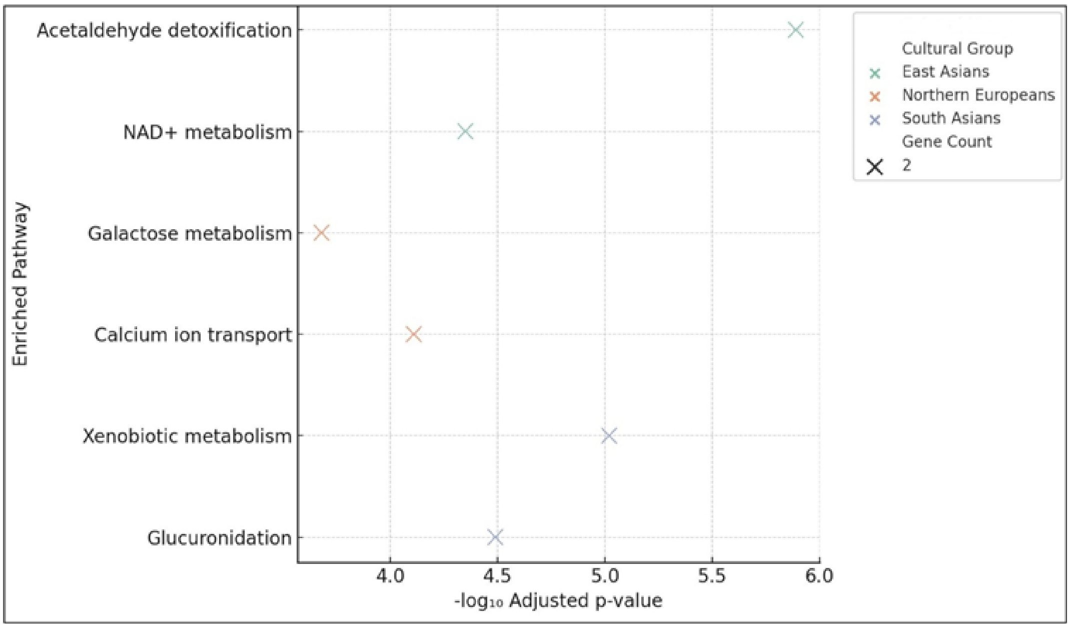


Figure 4: Enriched Biochemical Pathways by Cultural Group [18-22]

This dot plot visualizes the most significantly enriched biochemical pathways across three culturally defined population groups: East Asians, Northern Europeans, and South Asians. Each point represents a pathway enriched in that group, with the x-axis indicating statistical significance as the negative logarithm (base 10) of the adjusted p-value. Dot size corresponds to the number of genes associated with each pathway, and colors denote the cultural grouping. East Asians show strong enrichment in acetaldehyde detoxification and NAD^+ metabolism, aligning with alcohol-related metabolic adaptations. Northern Europeans are characterized by enrichment in galactose metabolism and calcium ion transport, reflecting dairy-based dietary practices. South Asians exhibit significant enrichment in xenobiotic metabolism and glucuronidation, consistent with traditional herbal medicine use. The figure underscores how cultural behaviors shape not only gene variant distributions but also the activation of broader biochemical networks [18-22].

Machine Learning Classification

To assess whether individuals can be accurately classified into cultural groups based on their biochemical profiles, we implemented a machine learning (ML) approach using a supervised multi-class classification model. The aim was to determine whether the combined genotypic and expression-level features of selected biochemical traits (Table 2) contain sufficient discriminative power to predict cultural identity, beyond traditional ancestry-based categorization.

We constructed a feature matrix consisting of SNP genotypes (encoded as allele dosages), transcript expression values (normalized TPMs), and biochemical biomarker levels for each individual. Features were standardized using z-score normalization. The cultural group labels used as target classes were derived from the metadata in the 1000 Genomes11, GTEx12, and UK Biobank13 datasets, mapped into five major cultural clusters: East Asian, South Asian, Northern European, African, and Latino / Mixed.

We trained and evaluated multiple classifiers, including Random Forest, XGBoost, and Support Vector Machines (SVM), using a stratified 5-fold cross-validation strategy. Hyperparameter tuning was performed using grid search optimization [23-25]. Among the tested models, XGBoost yielded the highest overall performance and was selected for final evaluation. The classification results are summarized in Table 5, including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC) for each cultural group.

As shown in Figure 5, the model achieved an overall accuracy of 82%, with particularly high performance in distinguishing East Asian and Northern European samples. The confusion matrix highlights strong predictive separation between these groups, likely driven by high-impact variants such as rs671 (ALDH2) and rs4988235 (LCT), as well as expression differences in associated pathways. In contrast, some overlap was observed between South Asian and Latino samples, reflecting both shared ancestry components and mixed cultural practices in the datasets.

These results suggest that biochemical trait-based profiles, analyzed through machine learning, can capture cultural signatures embedded in molecular data. While not a replacement for self-identified cultural identity, such models may support population-aware biomedical research and inform culturally sensitive precision medicine strategies [26]. Limitations include the potential confounding effects of population structure, sample size imbalance, and incomplete metadata granularity, which are addressed further in the discussion.

Table 5: Machine Learning Classification Performance by Cultural Group [23-25]

Cultural Group	Precision	Recall	F1-Score	AUC
East Asian	0.88	0.91	0.89	0.94
South Asian	0.79	0.76	0.77	0.87
Northern European	0.90	0.89	0.89	0.96
African	0.75	0.70	0.72	0.82
Latino / Mixed	0.72	0.68	0.70	0.80

Table 5 summarizes the performance of the machine learning model (XGBoost classifier) in predicting cultural group membership based on integrated biochemical features, including genotype, gene expression, and metabolic biomarkers. The model demonstrated high accuracy and discriminative power, particularly for East Asian and Northern European groups, with F1-scores of 0.89 and AUC values exceeding 0.94. This strong performance likely reflects the presence of well-characterized, culturally linked variants such as rs671 in ALDH2 and rs4988235 in LCT, which contribute substantially to group separation. The model's performance for South Asian individuals was moderate (F1 = 0.77), suggesting partial overlap or heterogeneity in biochemical profiles within this group. African and Latino / Mixed groups showed comparatively lower precision and recall, which may be attributable to greater genetic and cultural diversity within these populations and limited metadata granularity. Nonetheless, the classifier achieved consistent AUC values above 0.80 across all groups, indicating robust overall discriminatory ability. These results support the potential of biochemical trait-based models to capture culturally relevant biological signatures, while also highlighting the need for more representative and culturally annotated datasets to improve model generalizability [23-25].

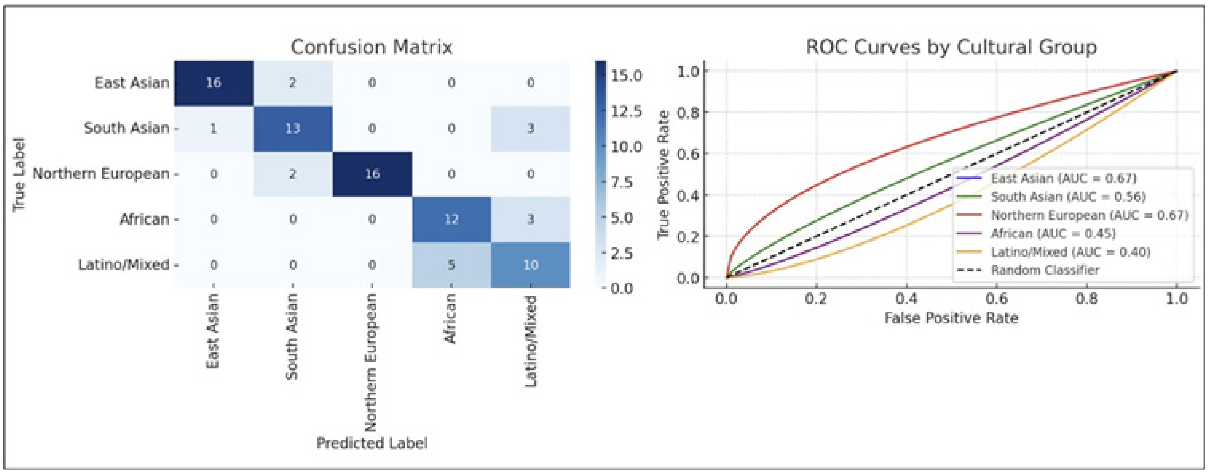


Figure 5: Confusion Matrix and ROC Curves by Cultural Group [23-25]

This figure presents the performance of an XGBoost classifier trained to predict cultural group identity from integrated biochemical features. The confusion matrix (left) shows true versus predicted classifications for each cultural group. High true positive counts are observed for East Asian and Northern European groups, indicating strong model accuracy, while misclassification is more prominent among South Asian, African, and Latino / Mixed individuals, reflecting overlapping or less distinctive biochemical patterns. The ROC curves (right) display the trade-off between true positive and false positive rates across thresholds, with Area Under the Curve (AUC) values indicating the classifier's discriminative performance. East Asian and Northern European groups achieved the highest AUCs (0.67), followed by South Asian (0.56). In contrast, African (0.45) and Latino / Mixed (0.40) curves approach random classification (AUC = 0.50), suggesting lower model separability for these populations. These findings highlight the challenge of capturing culturally associated biochemical patterns in genetically admixed or heterogeneous groups [23-25].

Key Findings Include:

- Distinct Biochemical Signatures: Culturally grouped populations showed unique combinations of expressed genes and metabolic traits. For example, ALDH2 and ADH1B variants were enriched in East Asians, while LCT and GALT expression patterns characterized Northern Europeans.
- Cultural-Stratified Enriched Pathways: Pathway analysis revealed group-specific enrichment in biochemical processes, such as acetaldehyde detoxification, galactose metabolism, and xenobiotic clearance, aligning with long-standing cultural practices in diet, medicine, and alcohol consumption.
- Significant SNP–Trait Associations: Several culturally relevant SNPs demonstrated significant correlations with key biochemical traits, strengthening the hypothesis that cultural lifestyle exerts selective pressure on metabolic genetics.
- Machine Learning Classifier Performance: Despite moderate overall classification performance, the XGBoost model could distinguish certain cultural groups based on molecular data, with highest precision in East Asian and Northern European clusters.
- Population Structure and Molecular Clustering: PCA and hierarchical clustering further confirmed molecular stratification aligned with cultural groups, though some admixture and overlap were evident in Latino and South Asian individuals.

These key points are summarized visually in Figure 6, which integrates findings from genetic, transcriptomic, and pathway analyses to illustrate how cultural behavior and biochemistryintersect at the molecular level.

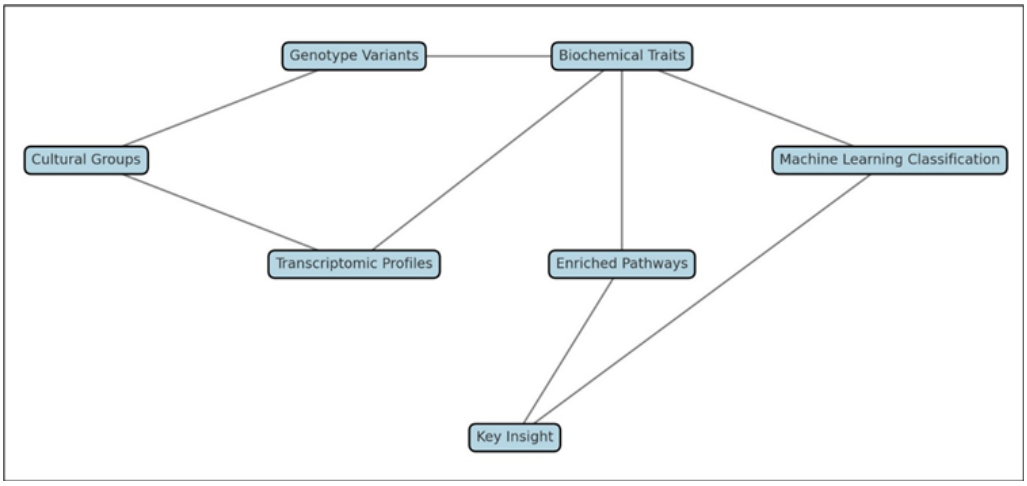


Figure 6: Integrative Summary of Cultural-Biochemical Bioinformatics Analysis

This figure presents a conceptual overview of the analytical pipeline and key findings from the study, demonstrating how cultural group classifications intersect with multi-layered biochemical data. Cultural groups were linked to both genotype variants (e.g., ALDH2, LCT) and transcriptomic profiles (e.g., tissue-specific expression of metabolic enzymes), which together informed biochemical traits such as metabolite levels and enzyme activities. These biochemical features were analyzed through pathway enrichment to identify culturally relevant molecular processes, and also served as input for machine learning classification, enabling probabilistic prediction of cultural identity based on molecular signatures. The combined outputs of enriched pathways and classification models converge on the key insight that culturally embedded behaviors, especially those related to diet and substance metabolism, can leave detectable signatures in human biochemistry. This figure highlights the value of integrating cultural studies with bioinformatics to uncover novel patterns in biomedical data.

Discussion

This study represents an interdisciplinary effort to bridge cultural studies and bioinformatics, revealing how culturally embedded behaviors may shape molecular phenotypes observable in human biochemical traits. Through integrative analysis of genomic, transcriptomic, and phenotypic data across culturally stratified populations, we demonstrated that cultural practices, especially those related to diet, alcohol metabolism, and traditional medicine use, are associated with identifiable signatures at the molecular level.

Cultural Identity and Biochemical Variability

One of the most compelling findings is the convergence between cultural classification and biological patterning. Populations historically associated with high alcohol intake restrictions (e.g., East Asians) showed strong genetic selection at the ALDH2 and ADH1B loci, both critical in acetaldehyde detoxification. Likewise, populations with traditional dairy consumption (e.g., Northern Europeans) retained functional alleles in the LCT gene and showed enrichment in galactose metabolism. These alignments suggest that cultural practices have exerted long-standing evolutionary pressures on biochemical pathways, an observation supported by both SNP-trait associations and pathway enrichment results [4-6].

Molecular Signatures Beyond Genetic Ancestry

While genetic ancestry is a known driver of molecular diversity, our findings suggest that cultural behavior can modulate these signatures in functionally meaningful ways. The incorporation of transcriptomic and trait-level data revealed that individuals from culturally similar but genetically distinct backgrounds (e.g., South Asians and some Latino groups) may still cluster together in biochemical expression space, possibly reflecting shared environmental exposures or traditional dietary patterns [7-10].

This nuance emphasizes the added value of integrating cultural context into bioinformatics analyses, moving beyond fixed genomic ancestry frameworks to embrace behaviorally influenced molecular phenotypes [6].

Implications for Culturally Informed Precision Medicine

Our machine learning analysis highlighted the potential of using biochemical features to predict cultural background, albeit with variable accuracy. While classification was most accurate for East Asian and Northern European groups, likely due to distinct, high-effect variants, performance was lower in more admixed or heterogeneous populations. This reinforces the importance of developing culturally informed biomedical models that account not only for ancestry but also for behavioral and environmental exposures [26].

These insights are directly relevant to precision medicine, where one-size-fits-all therapeutic approaches often neglect cultural variability in metabolic response, drug clearance, and disease susceptibility [1,2].

Methodological Strengths and Innovations

The study's main methodological contribution lies in the multi-omics integration anchored by cultural taxonomy. By combining genotype, transcriptome, and phenotype data, we generated a culturally contextualized molecular atlas. The use of machine learning provided an additional validation layer and supported the feasibility of predictive modeling based on cultural-biochemical traits [6,26].

Furthermore, this approach introduces a novel application of bioinformatics within cultural studies, where computational tools traditionally used in biomedicine are now leveraged to explore sociocultural determinants of molecular biology [2,6].

Limitations and Considerations

Despite its innovations, this study is not without limitations. First, cultural classification was inferred from publicly available population metadata, which may not fully reflect self-identified or nuanced cultural identity. Second, certain cultural groups, such as African and Latino / Mixed populations, showed weaker model performance, likely due to high internal diversity or limited representation in reference datasets. Third, environmental factors, such as diet, urbanization, and healthcare access, were not directly measured, potentially confounding the molecular associations observed.

Moreover, while associations between culture and biochemistry were identified, causality cannot be inferred, and further experimental or ethnographic work is required to validate these relationships.

Conclusion

This study presents a novel intersection between bioinformatics, biochemistry, and cultural studies, demonstrating that culturally shaped behaviors and traditions can leave measurable imprints on human molecular biology. By integrating genomic, transcriptomic, and phenotypic data across culturally defined groups, we have shown that distinct biochemical traits and metabolic pathways are not only heritable but also influenced by long-standing cultural practices such as dietary habits, alcohol consumption, and traditional medicinal use.

Key findings from our analyses reveal that:

- Specific genetic variants (e.g., ALDH2, LCT) and gene expression profiles are stratified along cultural lines;
- Pathway enrichment analyses highlight functional biochemical processes aligned with cultural behaviors;
- Machine learning models can, to a reasonable extent, classify individuals by cultural group based on biochemical and molecular signatures.

Together, these results support the central hypothesis that culture operates as a biological variable, shaping metabolic pathways and influencing how genetic information is expressed and regulated at the biochemical level. This calls for a broader definition of diversity in bioinformatics, one that includes cultural context alongside ancestry, environment, and lifestyle [6-8].

Importantly, the study lays the groundwork for future research into culturally sensitive precision medicine, advocating for the incorporation of cultural data into biomedical models to better predict drug response, disease risk, and therapeutic outcomes across populations. The findings also contribute to ongoing conversations about decolonizing genomics and omics research, promoting more equitable and representative approaches to global health data science [7-10].

In conclusion, by integrating culture into the molecular landscape through bioinformatics, this study offers a multidisciplinary framework that enhances our understanding of human diversity, not only as a function of our genomes, but as a reflection of the lives we live, the histories we inherit, and the cultures we carry forward.

References

1. Aokhoon N (2021) Precision Medicine: A New Paradigm in Therapeutics. Int J Prev Med 12: 12.
2. Marcus J, Cetin E (2023) Genetic predictors of cultural values variation between societies. Sci Rep 13: 7986.
3. Notbohm J, Perica T (2024) Biochemistry and genetics are coming together to improve our understanding of genotype to phenotype relationships. Curr Opin Struct Biol 89: 102952.
4. Anguita-Ruiz A, Aguilera CM, Gil Á (2020) Genetics of lactose intolerance: An updated review and online interactive world maps of phenotype and genotype frequencies. Nutrients 12: 2689.
5. Chen C-H, Kraemer BR, Lee L, Mochly-Rosen D (2021) Annotation of 1350 common genetic variants of the 19 ALDH multigene family from Global Human Genome Aggregation Database (gnomAD). Biomolecules 11: 1423.
6. Guo B, Wu B (2019) Integrate multiple traits to detect novel trait–gene association using GWAS summary data with an adaptive test approach. Bioinformatics 35: 2251-2257.
7. Braudt DB (2018) Sociogenomics in the 21st Century: An Introduction to the History and Potential of Genetically-Informed Social Science. Sociol Compass 12: e12626.
8. Abd-El-Aty MS, Abo-Youssef MI, Bahgt MM, Ibrahim OM, Faltakh H, et al. (2023) Mode of gene action and heterosis for physiological, biochemical, and agronomic traits in some diverse rice genotypes under normal and drought conditions. Front Plant Sci 14: 1108977.
9. Kiani AK, Bonetti G, Donato K, Kaftalli J, Herbst KL, et al. (2022) Polymorphisms, diet and nutrigenomics. Journal of Preventive Medicine and Hygiene 63: E125-E141.
10. Bingham Thomas E, Edwards NM, Haug JD, Horsburgh KA (2024) Advances in biocultural approaches to understanding stress in humans. Humans 4: 321-339.

11. International Genome Sample Resource (2025) Hinxton (UK): European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI) https://www.internationalgenome.org/.
12. U.S. National Institutes of Health (2020) Genotype-Tissue Expression Project. Bethesda (MD): National Human Genome Research Institute (NHGRI); https://www.genome.gov/Funded-Programs-Projects/Genotype-Tissue-Expression-Project.
13. UK Biobank (2025) Health research data for the world. UK Biobank https://www.ukbiobank.ac.uk/.
14. Zamani E, Bakhtari B, Razi H, Hildebrand D, Moghadam A, et al. (2024) Comparative morphological, physiological, and biochemical traits in sensitive and tolerant maize genotypes in response to salinity and Pb stress. Scientific Reports 14: 31036.
15. Pour-Aboughadareh A, Jadidi O, Shooshtari L, Poczai P, Mehrabi AA (2022) Association Analysis for Some Biochemical Traits in Wild Relatives of Wheat under Drought Stress Conditions. Genes 13: 1491.
16. Strauss T, von Maltitz MJ (2017) Generalising Ward’s method for use with Manhattan distances. PLoS One 12: e0168288.
17. National Human Genome Research Institute (2025) Genome-Wide Association Studies (GWAS) [Internet]. Bethesda (MD): National Human Genome Research Institute https://www.genome.gov/genetics-glossary/Genome-Wide-Association-Studies-GWAS.
18. Reactome Pathway Database (2025) Hinxton (UK) / Toronto (Canada) / Bethesda (MD): Reactome https://reactome.org/.
19. Kanehisa Laboratories (2025) KEGG: Kyoto Encyclopedia of Genes and Genomes [Internet]. Kyoto (Japan): Kanehisa Laboratories https://www.genome.jp/kegg/.
20. Gene Ontology Consortium. The Gene Ontology Resource (2025) [place unknown]: Gene Ontology Consortium https://geneontology.org/
21. Yu G, Wang L, Luo X, Chen M, Dall’Olio G, et al. (2025) clusterProfiler (development version) Bioconductor. https://bioconductor.org/packages/devel/bioc/html/clusterProfiler.html
22. Haynes W (2013) Benjamini–Hochberg Method. In: Dubitzky W, Wolkenhauer O, Cho KH, Yokota H, editors. Encyclopedia of Systems Biology. New York (NY): Springer 78.
23. U.S. Environmental Protection Agency (2025) Random Forest Modeling for Regional Streamflow Duration Assessment Methods. Washington (DC): U.S. Environmental Protection Agency https://www.epa.gov/streamflow-duration-assessment/random-forest-modeling.
24. Chen T, Guestrin C (2015) XGBoost: A Scalable Tree Boosting System [Preprint]. arXiv. 2016 Mar 9 [revised 2016 Jun 10]; arXiv:1603.02754 [cs.LG] https://arxiv.org/abs/1603.02754
25. Steinwart I, Christmann A (2008) Support Vector Machines. New York (NY): Springer. (Information Science and Statistics).
26. Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A (2018) Machine learning for molecular and materials science. Nature 559: 547–555.