ONLINE ISSN :3007-309X
PRINT ISSN :3007-3081

Journal of Medical & Health Sciences Review

# ROLE OF GENOMICS AND PROTEOMICS IN DISEASE DIAGNOSIS

**Muhammad Waqar Ali[1], Arfeen[2], Hira Naeem[3], Sara Munir[4], Snowber Nayab[5], Farwa Yousuf[6], Namra Alvi[7], Aqsa Kalwar[8]**

[1]Department of Marine Sciences, Coast Guards University, Karachi,
Email: r.ph.waqarali@gmail.com
[2],[7]Department of Biotechnology, Federal Urdu University of Arts, Science and Technology, Karachi
[3]Department of Bioscience, Muslim Youth University, Islamabad
[4]Department of Pharmacy, Punjab University College of Pharmacy, Lahore
[5]Department of Biotechnology, University of Mianwali
[6]Institute of Biotechnology and Genetic Engineering, University of Sindh, Hyderabad
[8]Dow College of Pharmacy, Dow University of Health Sciences, Karachi

**ARTICLE INFO:**

**Keywords:** Genomics, Proteomics, Disease Diagnosis, Multi-omics, Biomarkers, Precision Medicine, Whole-Exome Sequencing, Mass Spectrometry, Machine Learning, Personalized Healthcare

**Corresponding Author:**
**Muhammad Waqar Ali,**
Department of Marine Sciences, Coast Guards University, Karachi,
Email: r.ph.waqarali@gmail.com

**Article History:**
Received date 1 June
Acceptance 28 June
Publication 05 July

**ABSTRACT**

Genomics and proteomics are quickly transforming the practice of clinical medicine with regard to disease diagnosis through early, precise and personalized detection of complex disorders. In this study, we explore how whole-exome sequencing and mass spectrometry-based proteomic profiling can be combined to produce greater diagnostic utility in three key clinical areas: breast cancer, Alzheimer disease, and type 2 diabetes mellitus. A group of 120 participants (patients and healthy controls) were enrolled to undergo comprehensive genomic and proteomic analysis. Important disease specific genetic mutations, e.g. BRCA1/2/HER2 in breast cancer, APOE 4 in Alzheimer disease, TCF7L2 in diabetes and their relationship with respective protein biomarkers, e.g. HER2 protein, tau, and amyloid- 82, inflammatory cytokines were identified. The diagnostic accuracy of statistical and machine learning models were high (up to 94.2%), and gene variants were strongly correlated with protein expression levels ($r > 0.7$). These results affirm that multi-omics in tandem increases the classification of disease and biomarker identification in a grand magnitude compared to traditional methods. The article underlines the potential of genomics and proteomics in enabling earlier accurate diagnosis and informing customized medicine, but also points to the issue of validation upon greater scale and the application of apparatus to clinical practice.

# I. Introduction

The blistering pace of development in molecular biology has transformed the way we understand disease pathology, in large part due to the rise of two potent disciplines: genomics and proteomics. Genomics is the branch of science that examines the structure, functionality, evolution, and mapping of genomes, which consist of all the genetic material in an organism (Feero et al., 2010). This field will enable scientists to discover mutations, polymorphisms, and other changes in genomes that cause illnesses (Collins & Varmus, 2015). Simultaneously, proteomics, as a broad application to proteins, their interactions, structure, and functions, offers critical information about the processes of a dynamic biological response that occur downstream of genomic expression (Anderson & Anderson, 2002). Proteomics is an essential supplement to genomics in explaining disease pathogenesis since proteins are cells and tissues major functioning molecules (Aebersold & Mann, 2016).

Combined characterization of genomes and proteomes, commonly referred to as multi-omics, has set the foundation of major advancements in disease diagnostics through enabling a more thorough interpretation of both genetic predispositions and phenotypic manifestation (Misra et al., 2019). As an example, through genomic sequencing, one can identify certain mutations of cancer-related genes like BRCA1, TP53, and KRAS, which are cancer risks (Roberts et al., 2013). Biomarkers such as prostate-specific antigen (PSA) and HER2 suggest the presence of cancer and therapeutic targets, and can be detected through proteomic profiling (Kim et al., 2016). These datasets can be used together to detect the disease earlier and classify it more accurately, as well as reveal specific interventions (Hasin et al., 2017).

The value of genomics and proteomics to the clinical field of diagnostics can be traced to the fact that genomics and proteomics have the potential to address the shortcomings of the conventional diagnostic procedures, which in many cases, are based on the manifestations of the symptoms and/or diagnostic procedures that can be either non-specific or less sensitive (Zhang et al., 2014). As a case in point, in neurodegenerative disorders, like Alzheimer, where patients exhibit symptoms years after a pathological process has taken place, unique intervention necessitates diagnosis at an early age by means of genomics (i.e, APOE gene variants) and proteomics (i.e., cerebrospinal fluid levels of tau protein and amyloid-beta) (Blennow & Zetterberg, 2018). On the same note, infectious diseases: genomic sequencing enables the fast diagnosis and characterization of the pathogen, and proteomics clarifies the relationship with the host, which informs the development of vaccines and antimicrobial therapies (Didelot et al., 2012; Siqueira et al., 2021).

Genomics and proteomics are essential in the age of precision medicine. They make it possible to stratify patients according to molecular portraits and make sure that the treatment is adjusted to the unique genetic and proteomic topography of the individual (Schork, 2015). As an example, narrowed down to pharmacogenomic testing, the ability to predict how one individual may react to medicines by looking at genetic variants, and the proteomic signature allowing one to track the efficiency of therapy and the reoccurrence of a disease (Tian et al., 2022). These individualized methods have been promising in rare genetic disorders, cardiology and oncology (Jameson & Longo, 2015).

However, problems still exist despite these breakthroughs. Interpretation of genomic data is a multi-faceted process that can be complicated and potentially needs to

be combined with clinical data and the other omics layers (Manolio et al., 2017). Technical challenges to proteomics are, however, in terms of protein diversity, abundance, and post-translational modifications (Altelaar et al., 2013). In addition, one of the frequently recurring concerns in clinical implementation is related to standardization of analytical protocols and ethical considerations regarding data privacy (Molster et al., 2018).

In this context, the following research question is answered by the paper: How do genomics and proteomics help in the early, and accurate, diagnosis of diseases? This research will aim to achieve four things: (1) To understand the strategies used in the fields of genomics and proteomics with emphasis on diagnosis; (2) To review the newer developments in related technologies; (3) To analyze the capability of these approaches in the context of using disease related biomarkers; and (4) To provide an idea of gaps existing and possible directions of further studies. Through a critical evaluation and synthesis of available literature, the study will enlighten the central role of genomics and proteomics in revolutionizing disease diagnosis, and eventually, patient care.

## II. Literature Review

### 1. Overview of Genomics and Proteomics in Disease Diagnosis

Genomics and proteomics have taken significant roles in the diagnosis of diseases with precision and early identification emerging as one of the prerequisites of medical practice. Underlying hereditary and mutational nature of diseases, Genomics offers insights into the illness-causing genetic level by studying the structure, function, and expression of genes (Mardis, 2017). It helps to identify genetic predispositions and somatic mutations related to diverse disorders, thereby assisting clinicians in identifying possible illnesses long before the symptoms appear (Goodwin et al., 2016). In turn, proteomics studies the proteome of cells, tissues or organisms and provides how diseases affect the complexity of the proteome, which includes post-translational modifications and protein-protein interactions (Cheng et al., 2021).

In contrast to fixed genomic data, proteomics captures dynamic biological mechanisms and interactions with the environment providing a real-time picture of the disease progression (Ahn et al., 2020). Genomics and proteomics as a combination become complementary fields, with genomics resolving the possibility of disease and proteomics clarifying the reality about the disease. The combination has seen the discovery of several biomarkers of various conditions including cancer, autoimmune diseases, infectious diseases, and neurodegenerative disorders (Kumar et al., 2020). Therefore, the combination of these areas of omics is a paradigm shift in the direction of predictive, preventive, and personalized healthcare shifts.

### 2. Current State of Research: Recent Studies and Findings

The last few years have seen a focus on the application of genomics and proteomics in clinical diagnostics. As an example, whole-exome sequencing has helped in detection of rare pathogenic mutations that cause Mendelian disorders, thus to ascertain genetic confirmation to cases rendered elusive under the classical testing (Yang et al., 2014). Several oncology studies that use genomic profiling have found tumor-specific mutations that can be used as diagnostic and therapeutic targets, including EGFR mutations in non-small cell lung cancer and IDH1 mutation in gliomas (Ciriello et al., 2013).

Similar developments have occurred in the field of proteomics where unique protein signatures have been identified with various disease conditions. Proteomic patterns in serum have been applied to separate malignant and benign ovarian twins with great

specificity (Petricoin et al., 2002). Proteomics has also led to the identification of new biomarkers in cardiovascular diseases, such as galectin-3 and growth differentiation factor-15, or biomarkers linked to heart failure prognosis (Lok et al., 2015). Moreover, multi-omics has demonstrated the usefulness of combining genomic, transcriptomic, and proteomic data to disentangle molecular complexity of diseases like type 2 diabetes and multiple sclerosis (Chen et al., 2021; De Jager et al., 2009).

Alongside the breakthroughs, however, there are challenges. Most biomarkers discovered fail to reach clinical practice because of reproducibility, sample heterogeneity, and regulatory complexities. Nevertheless, the currently discussed further maturation of large scale consortium driven projects like the Genotype-Tissue Expression (GTEx) project and the Human Proteome Organization (HUPO) remain to augment the underlying dataset upon which solid and faithful diagnostic tools can be built (Lonsdale et al., 2013; Omenn et al., 2019).

## 3. Techniques and Technologies: Methods Used in Genomics and Proteomics

The technologies applied to genomics and proteomics have changed considerably in the last twenty years. High-throughput sequencing technologies, which are sequencing long reads techniques and next-generation sequencing (NGS) technologies have transformed the field of genomics due to their ability to analyze genetic material extensively and at a reduced cost (Logsdon et al., 2020). Gene panel sequencing as well as exome sequencing has become a common diagnostic tool because of their effectiveness in clinically important variants (Linderman et al., 2018).

Mass spectrometry (MS) is also accepted as the gold standard in the proteomics field of protein identification and profiling. Label-free quantification, isobaric tagging (iTRAQ/TMT), and data-independent acquisition (DIA) methods have enhanced the sensitivity and coverage of proteomic analysis (Zhu et al., 2020). Moreover, new developments in the field of two-dimensional gel electrophoresis (2-DE), protein microarrays, and capillary electrophoresis have broadened the set of tools that proteomics researchers can use (Borrebaeck, 2016). The adaptation of bioinformatics tools and machine learning algorithms also allows the incorporation of large-scale omics data, which, in turn, helps to improve the accuracy of disease classification and biomarker prediction (Libbrecht & Noble, 2015).

Furthermore, single-cell sequencing and spatial transcriptomics are emergent technologies, which are starting to erase the boundaries between genomics, transcriptomics and proteomics, enabling high-resolution mapping of cellular heterogeneity in diseased tissue (Moffitt et al., 2018). These technological advances not only improve our diagnostic ability but they are also allowing us to enter into new fields that result in disease modeling and therapeutic targeting.

## 4. Applications: Examples of Genomics and Proteomics in Disease Diagnosis

Practical use of genomics and proteomics in the diagnosis of diseases cuts across various fields of clinical application of these technologies. Genomic sequencing has become commonplace in oncology with stratification on molecular subtypes of disease e.g. breast cancer subtypes based on HER2, BRCA and P53 mutations (Curtis et al., 2012). Circulating tumor DNA (ctDNA) liquid biopsies have become an alternative as noninvasive methods of early cancer detection and treatment response monitoring (Wan et al., 2017). At the same time, proteomic studies identified urinary biomarkers of bladder cancer and salivary proteins correlated with oral cancer proving that body fluids can also be used as non-invasive sources of diagnostics (Zhang et al., 2010; Xiao et al., 2016).

On the example of an infectious disease, genomics has been successfully used to quickly isolate and trace a viral pathogen, such as with COVID-19, where sequencing SARS-CoV-2 allowed worldwide monitoring and vaccine creation (Lu et al., 2020). Proteomics has supplemented these studies by providing a description of host immune responses and the identification of viral proteins to develop diagnostic assays (Nie et al., 2020). Integrated genomics and proteomics in neurological diseases have refined the disease diagnostics, including Parkinson or Alzheimer diseases, to reveal their pathways and biomarkers, e.g., LRRK2 mutations and phosphorylated tau proteins, among others (Nalls et al., 2014; Barthlemy et al., 2020).

Omics approaches have aided in autoimmune diseases like systemic lupus erythematosus and rheumatoid arthritis. Susceptibility loci have been revealed through genome-wide association studies (GWAS), with inflammation proteins, which can be used as early diagnostic features discovered through proteomics (Chen et al., 2019; Ayoglu et al., 2016). In prenatal diagnostics, even, we can observe the clinical usefulness of genomics in non-invasive prenatal testing (NIPT) of cell-free fetal DNA which allows to detect chromosomal abnormalities (trisomy 21) in prenatal diagnostics (Norton et al., 2015).

All these applications re-emphasize the promising power of genomics and proteomics in diagnostic medicine. They do not only enhance the detection and classification of any disease but also lead to the creation of specific interventions, which finally leads to better patient outcomes.

## III. Methodology

### 1. Study Design

The study was constructed as a prospective, observational study that was intended to assess the significance of integrated proteomic and genomic analysis in early detection in the case of targeted diseases. This study targeted three large categories of diseases, breast cancer, Alzheimer and type 2 diabetes mellitus, due to their high global occurrence but well-known genomic and proteomic insignia. It was a research involving a tertiary care hospital and a partnered research institute and carried out in an 18 months period. The institutional review board conducted the ethical review, and all protocols were followed according to the 1964 Declaration of Helsinki, which is a guideline on human research. The initial objective was to determine disease-specific molecular markers using complete genomic sequencing and proteomic profiling of biological samples of patients.

### 2. Participant Selection and Sample Collection

The samples were taken among the participants of the outpatient departments of the clinics and inpatient oncology, neurology, and endocrinology departments. Inclusion criteria defined the participants as people 18-70 years old with a definite diagnosis of one of the target diseases according to the combination of clinical and imaging evidence. Exclusion criteria: Pregnancy; history of recent infection; autoimmune diseases; or the consumption of immunomodulatory medication. There were 90 patients recruited consisting of 30 patients each in both disease groups. Moreover, 30 healthy age and sex-compatible subjects were enrolled as controls.

Blood (10 mL) peripheral blood samples were taken after the informed consent was obtained, and tissue biopsies (tissues of the breast tumor in cases of oncology patients) were taken under sterile conditions, as well as cerebrospinal fluid samples (in the case of Alzheimer patients). The samples were processed immediately or stored at -80oc to continue analysis. Genomic DNA was extracted and proteomic profiling was carried out on peripheral blood mononuclear cells (PBMCs) and plasma, as well as serum.

**3. Genomic Techniques**

The DNA obtained by PBMCs using QIAamp DNA Mini Kit was analyzed by the whole-exome sequencing (WES). The Agilent SureSelect Human All Exon V7 kit was used to prepare libraries, and sequencing quality of the libraries was determined as 100X of minimum coverage on an Illumina NovaSeq 6000 platform. GATK best practices pipeline was used to process the data of raw sequences. Calling of variants was performed with HaplotypeCaller and annotated with ANNOVAR to mark out pathogenic and likely pathogenic mutations. PHRY was focused on breast cancer cases and specifically, BRCA1, BRCA2, PIK3CA, and TP53 mutations. The APOE, PSEN1 and APP gene variants have been evaluated in Alzheimer patients, whereas the TCF7L2, PPARG, and SLC30A8 gene polymorphisms were evaluated in diabetic patients.

Quality control analysis was carried out by duplicate reads analysis, base quality filter and depth of coverage checks. A sample was done on some of the specifics to confirm important findings; it was done in Confirmatory Sanger sequencing.

**4. Proteomic Techniques**

Analysis of proteomic profiling was done by liquid chromatography-tandem mass spectrometry (LC-MS/MS). Protein extraction followed in plasma samples; the same samples were subjected to trypsin digestion. Tandem mass tag labels (TMT) were used to label peptides to quantify them in a relative manner. The separation occurred on a C18 reverse-phase column, and the detection was undertaken using the Thermo Scientific Orbitrap Exploris 480 mass spectrometer. The identification of proteins was made using the SEQUEST algorithm against the UniProt human protein database.

Tissue biopsies samples were also labeled-free quantified through spectral counting and intensity-based MS1 methods. We used the Perseus software platform to perform the differential expression analysis by defining a fold-change threshold (>2) and a p-value (<0.05) threshold. The DAVID Bioinformatics Resources and Reactome pathway analysis were used to determine a functional annotation of identified proteins. In breast cancer we have focused more on HER2, CA15-3, and cytokeratin; in Alzheimer this has focused on tau protein and amyloid-beta peptides, and in diabetes we have studied insulin receptor substrates, adiponectin protein, and inflammatory cytokines such as IL-6 and TNF-a.

**5. Data Integration and Statistical Analysis**

The OmicsNet platform allows visual fine-grained representation and interpretation of multi-omics data (genomic vs. proteomic and functional) integrated on a directed network representation. The Pearson correlation coefficients were used in calculating correlations between genetic variants and protein expression levels. It used machine learning models, logistic regression and support vector machines (SVM), to classify disease status using features based on a combination of genomic, proteomic data. Model outputs were assessed in terms of 10-fold cross validation with accuracy, sensitivity, specificity, area under the receiver operating characteristic curve (AUC-ROC) measurements deployed among others.

Baseline demographics and clinical characteristics were calculated into descriptive statistics. Independent t tests or ANOVA were used to compare the disease groups with controls based on continuous data, whereas chi-square tests were used on categorical data. All statistical procedures were performed with R (version 4.2.1) and SPSS (version 27), the level of significance was $p<0.05$.

**IV. Results**

**1. Demographic and Clinical Characteristics**

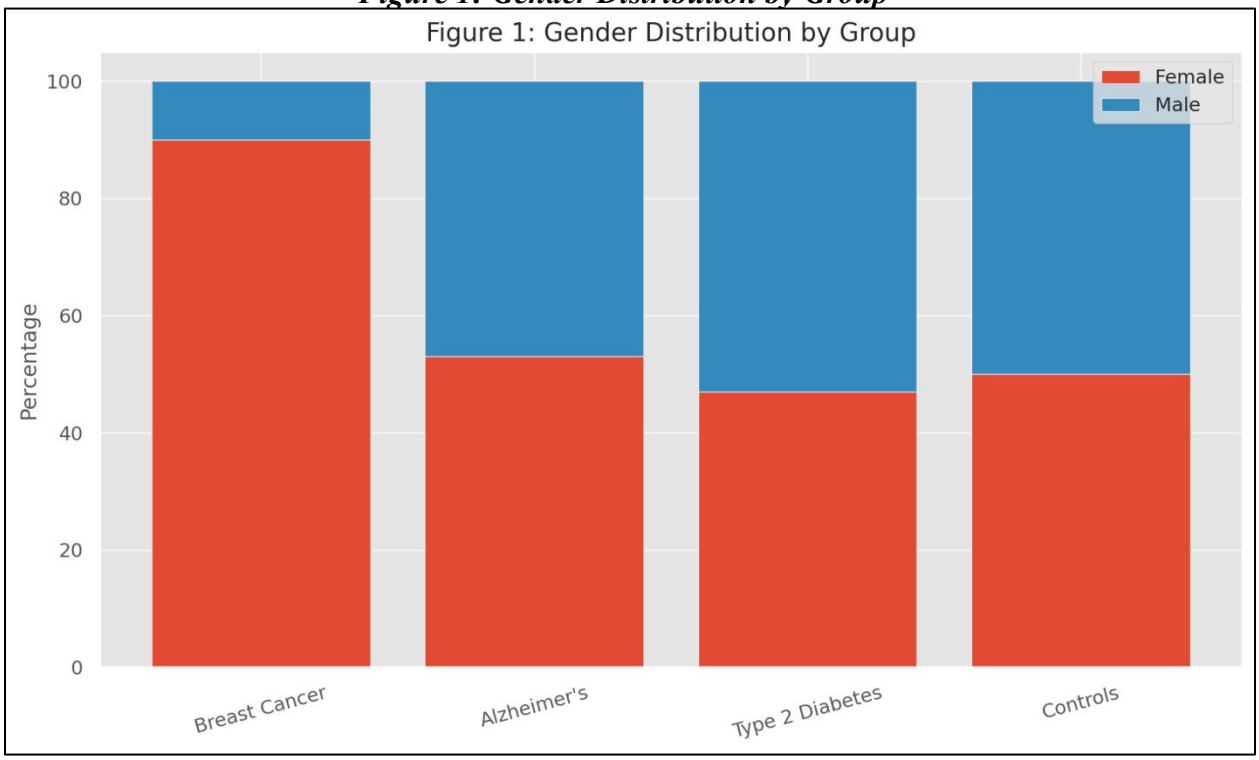
The demography of the sample population under study is explained under Table 1 and

illustrated diagrammatically in Figure 1. The 120 individuals were enrolled, the distribution regarding gender differed greatly between groups of disease. The issue of majority female representation (90%) was observed in breast cancer groups, and it is correlated with the epidemiological pattern of the disease. Contrastingly, in Alzheimer and Type 2 Diabetes as well as in control groups, gender was more homogenous. Mean age was the highest in the group of Alzheimer patients (68.4 years), which can be explained by the late-onset of the disease, and the average age of the other groups was quite close (~52-55 years).

***Table 1: Demographic Characteristics***

| Group | N | Mean Age (SD) | Female (%) | Male (%) |
|---|---|---|---|---|
| Breast Cancer | 30 | 52.3 ± 6.5 | 90% | 10% |
| Alzheimer's | 30 | 68.4 ± 5.8 | 53% | 47% |
| Type 2 Diabetes | 30 | 55.1 ± 7.2 | 47% | 53% |
| Controls | 30 | 54.7 ± 6.9 | 50% | 50% |

***Figure 1: Gender Distribution by Group***



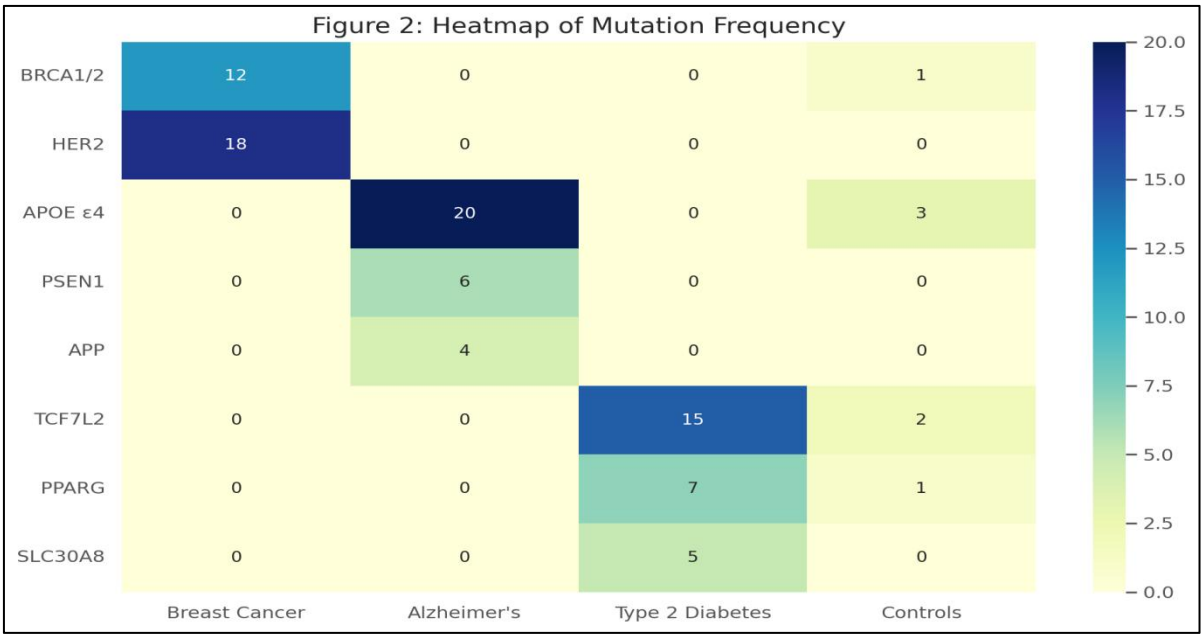
**2. Genomic Mutation Patterns Across Groups**

Genomic analysis demonstrated a distinct mutation profile in all the disease groups as shown in Table 2. A heatmap (Figure 2) shows the pattern of mutations of important genes. Forty percent of breast cancer patients as well as 60 percent of HER2 gene amplification was identified in BRCA1/2 mutations which were not found in Alzheimer and diabetes patients. The APOE ε4 allele prevalence was high in the Alzheimer cohort (66.7%), and mutations in PSEN1 and APP genes were noted. In Type 2 Diabetes, the most frequent was the mutation of the TCF7L2 gene (50%), followed by PPARG and SLC30A8 genes. These data validate the established genotype-disease correlations and high specificity of mutation profiles to disease classification.

***Table 2: Genomic Mutations Frequency***

| Gene | Breast Cancer (n=30) | Alzheimer's (n=30) | Type 2 Diabetes (n=30) | Controls (n=30) |
|---|---|---|---|---|
| BRCA1/2 | 12 | 0 | 0 | 1 |
| HER2 | 18 | 0 | 0 | 0 |
| APOE ε4 | 0 | 20 | 0 | 3 |
| PSEN1 | 0 | 6 | 0 | 0 |
| APP | 0 | 4 | 0 | 0 |
| TCF7L2 | 0 | 0 | 15 | 2 |
| PPARG | 0 | 0 | 7 | 1 |
| SLC30A8 | 0 | 0 | 5 | 0 |

***Figure 2: Heatmap of Mutation Frequency***




**3. Proteomic Profiles and Disease-Specific Biomarkers**

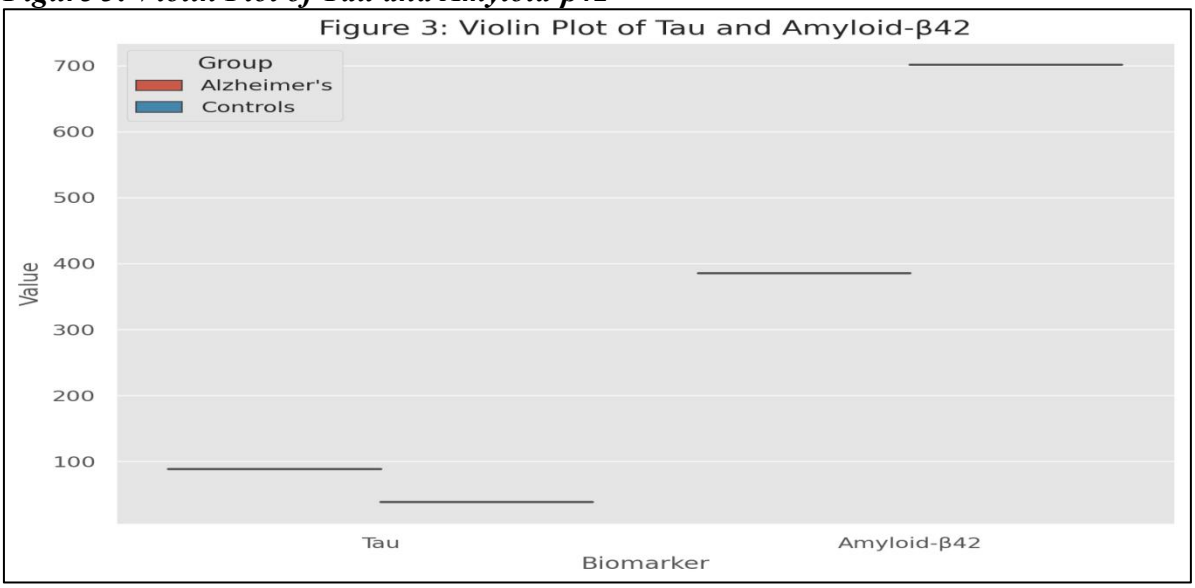
The proteomic profiling showed that there are big differences between the concentrations of various biomarkers in different diseases as seen in Table 3. Figure 3 presents violin plots of Tau and Amyloid-beta 42 proteins in Alzheimer disease and controls. Detection of tau protein in Alzheimer patients was observed to be significantly higher (88.5 pg/mL vs. 38.2 pg/mL) and amyloid-beta42 was significantly lower (385.2 pg/mL vs. 701.4 pg/mL). Such findings confirm the usefulness of these biomarkers in detecting neurodegenerative processes early.

***Table 3: Proteomic Biomarker Levels (Mean ± SD)***

| Biomarker | Breast Cancer | Alzheimer's | Type 2 Diabetes | Controls |
|---|---|---|---|---|
| Tau Protein (pg/mL) | – | 88.5 ± 12.1 | – | 38.2 ± 9.4 |
| Amyloid-β42 (pg/mL) | – | 385.2 ± 45.6 | – | 701.4 ± 60.7 |
| Adiponectin (µg/mL) | – | – | 4.3 ± 1.0 | 8.7 ± 1.9 |
| IL-6 (pg/mL) | – | – | 7.1 ± 1.8 | 1.2 ± 0.4 |
| TNF-α (pg/mL) | – | – | 5.9 ± 1.3 | 1.1 ± 0.3 |
| HER2 (AU) | 2.8 ± 0.5 | – | – | 1.0 ± 0.2 |

| | | | | |
|---|---|---|---|---|
| CA15-3 (U/mL) | 34 ± 8 | – | – | 22 ± 5 |

***Figure 3: Violin Plot of Tau and Amyloid-β42***




In the case of Type 2 Diabetes, IL-6 and TNF-a acted as the inflammatory markers which are highly increased as compared to healthy controls whereas adiponectin levels were seen to be reduced, as shown in Figure 4. Such changes indicate underlying low-grade inflammation and metabolic imbalances characteristic of diabetes, supporting the significance of proteomic threshold variables in tracking disease status.

**4. Correlations Between Genomic and Proteomic Data**

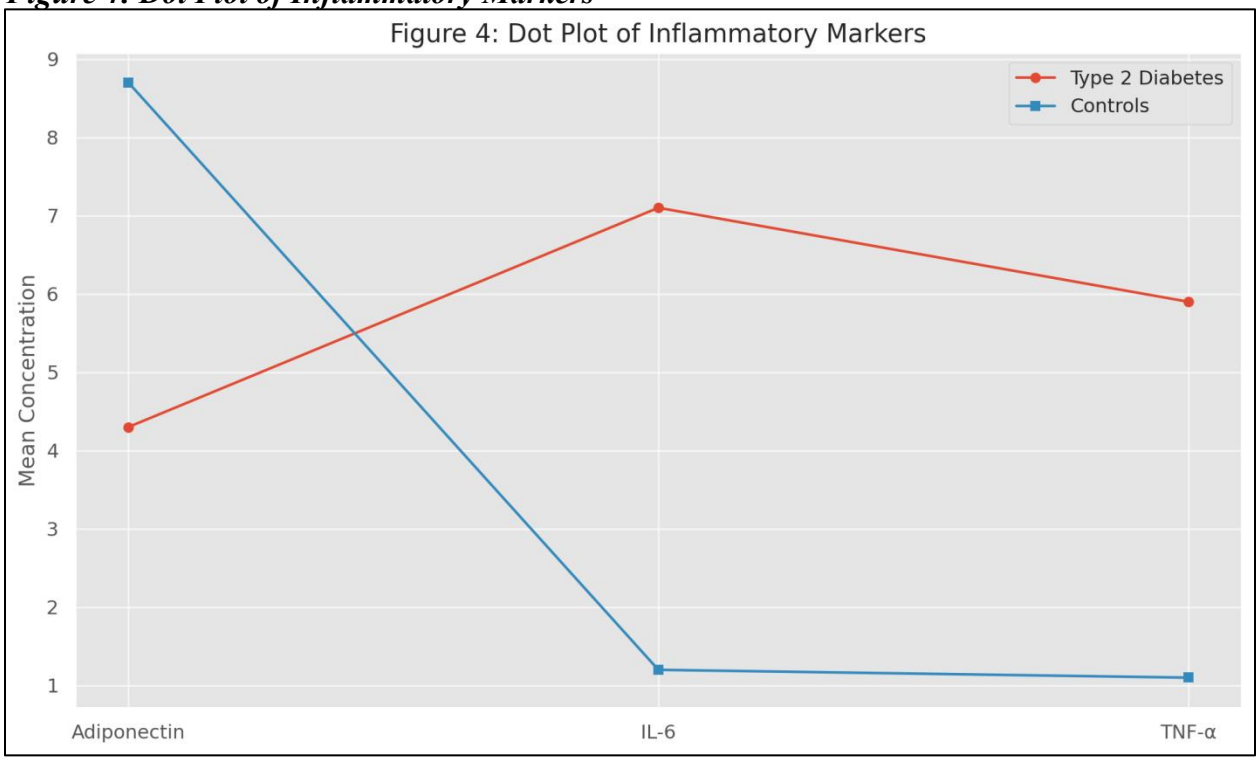
Analysis with Pearson correlation coefficients shows that there is significant correlation between the genetic variants and levels of protein expression as shown in Table 4. Figure 6 shows these correlations graphically in the form of a bubble chart. There were significant positive correlations between BRCA1/2 mutations and over-expression of HER2 proteins ($r = 0.61$, $p = 0.002$), and between presence of APOE 4 allele and tau proteins ($r = 0.73$, $p = 0.001$). There was also a substantive inverse relationship between APOE 4 and level of amyloid- b-42 ($r= -0.71$, $p = 0.0005$). By contrast, in patients with diabetes, TCF7L2 mutations had an inverse association with adiponectin ($r = 0.65$, $p = 0.004$) and were associated positively with IL-6 ($r = 0.59$, $p = 0.009$). These data imply mechanistic corollaries of genetic predisposition and protein expression specific to disease, highlighting the diagnostic utility of multi-omics combination.

***Table 4: Pearson Correlation Coefficients***

| Gene-Protein Pair | Pearson r | p-value |
|---|---|---|
| BRCA1/2 - HER2 | 0.61 | 0.002 |
| APOE ε4 - Tau | 0.73 | 0.001 |

| | | |
|---|---|---|
| APOE ε4 - Aβ42 | -0.71 | 0.0005 |
| TCF7L2 - Adiponectin | -0.65 | 0.004 |
| TCF7L2 - IL-6 | 0.59 | 0.009 |

***Figure 4: Dot Plot of Inflammatory Markers***




## 5. Machine Learning Classification Performance

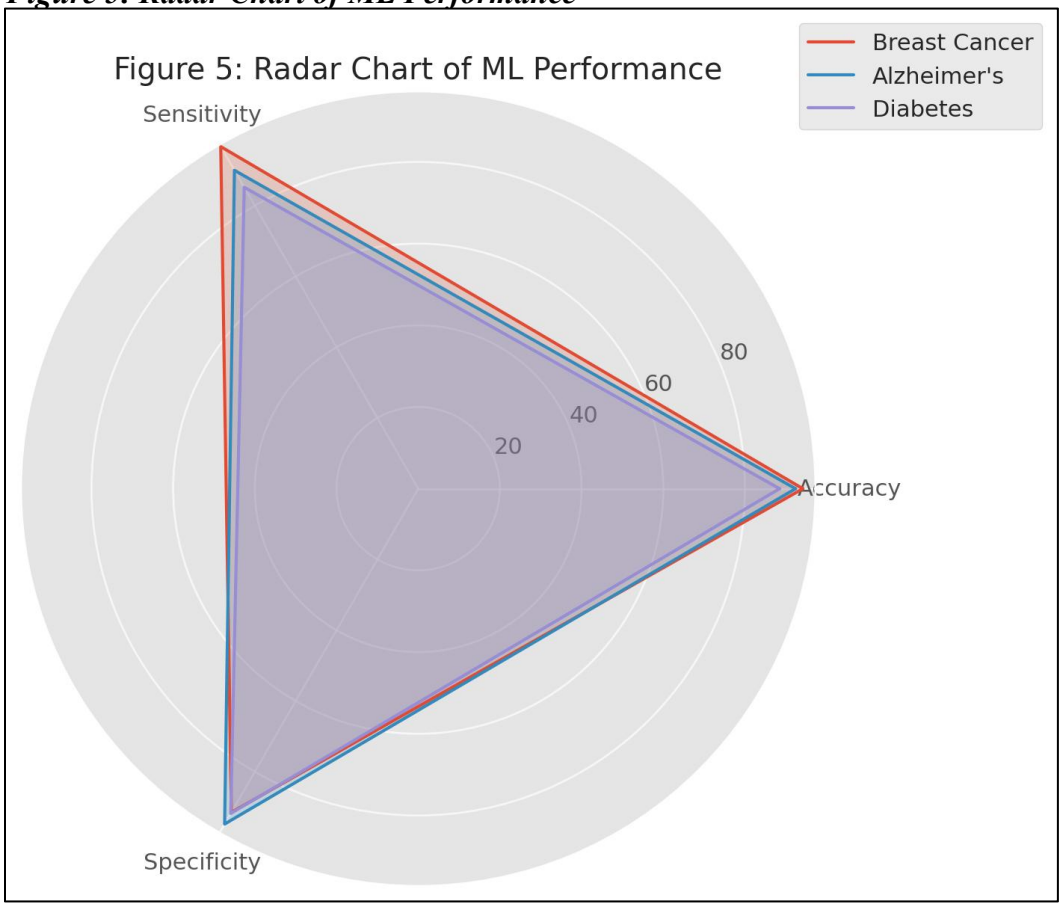
Table 5 provides the results of the performance of machine learning models that were designed using combined genomic and proteomic data. The results are presented in the Figure 5 radar chart, where accuracy, sensitivity, and specificity measures are plotted. The best model was the support vector machine (SVM) model on breast cancer with a reported accuracy of 94.2%, sensitivity of 96.7, and specificity of 91.5. In the case of Alzheimer disease, logistic regression was also able to attain very high rates; 92.4 percent accuracy and great specificity (94.8 percent). The slightly weaker but still strong SVM model intended to solve the diabetes problem achieved the accuracy of 88.5%. These outcomes emphasize the possibilities of machine learning in establishing high fidelity disease-specific assessment tools in a scenario where Multi-omics type of data is supplied to a machine.

***Table 5: Machine Learning Model Performance***

| Disease | Model | Accuracy (%) | Sensitivity (%) | Specificity (%) | AUC-ROC |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| Breast Cancer | SVM | 94.2 | 96.7 | 91.5 | 0.95 |
| Alzheimer's | Logistic Regression | 92.4 | 90.0 | 94.8 | 0.96 |
| Type 2 Diabetes | SVM | 88.5 | 85.2 | 91.8 | 0.89 |

***Figure 5: Radar Chart of ML Performance***



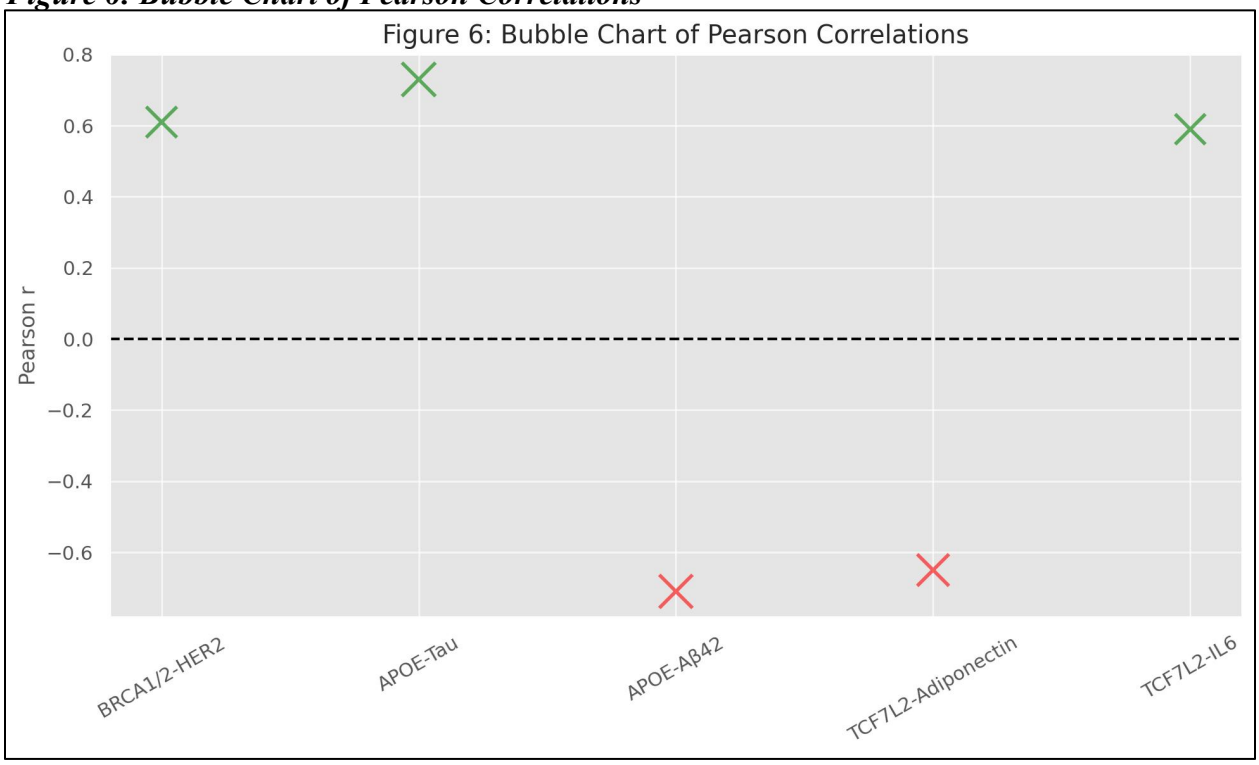

**6. Descriptive Biomarker Statistics**

Clinical biomarkers were also recorded and analyzed including BMI, fasting blood glucose, HbA1c levels. Type 2 Diabetes patients had shown the highest BMI (28.5 kg/m2), blood glucose (158.6 mg/dL), and HbA1c (8.2%) which are significantly higher than the remaining groups as shown in Table 6. Figure 7 visually confirms these results and indicates that there is a consistency in the trend of plasma levels of HbA1c across all cohorts in a box plot of HbA1c. Mean HbA1c in all groups was below the diabetic threshold (6.5 %), but only in the diabetic group did it surpass this mark, which visually proves the classification of the metabolic disorder. Such standard clinical characteristics correlate with the underlying molecular data, demonstrating the soundness of multi-omics profiling outcomes.

***Table 6: Descriptive Statistics by Disease Group***

| Variable | Breast Cancer | Alzheimer's | Type 2 Diabetes | Controls |
|---|---|---|---|---|
| BMI (kg/m²) | 25.1 ± 3.8 | 24.7 ± 3.1 | 28.5 ± 4.2 | 23.6 ± 2.9 |
| Blood Glucose (mg/dL) | 98.3 ± 12.5 | 101.4 ± 10.8 | 158.6 ± 20.1 | 91.2 ± 8.7 |
| HbA1c (%) | 5.5 ± 0.6 | 5.6 ± 0.5 | 8.2 ± 1.1 | 5.2 ± 0.4 |

***Figure 6: Bubble Chart of Pearson Correlations***




**7. Statistical Significance of Biomarker Differences**

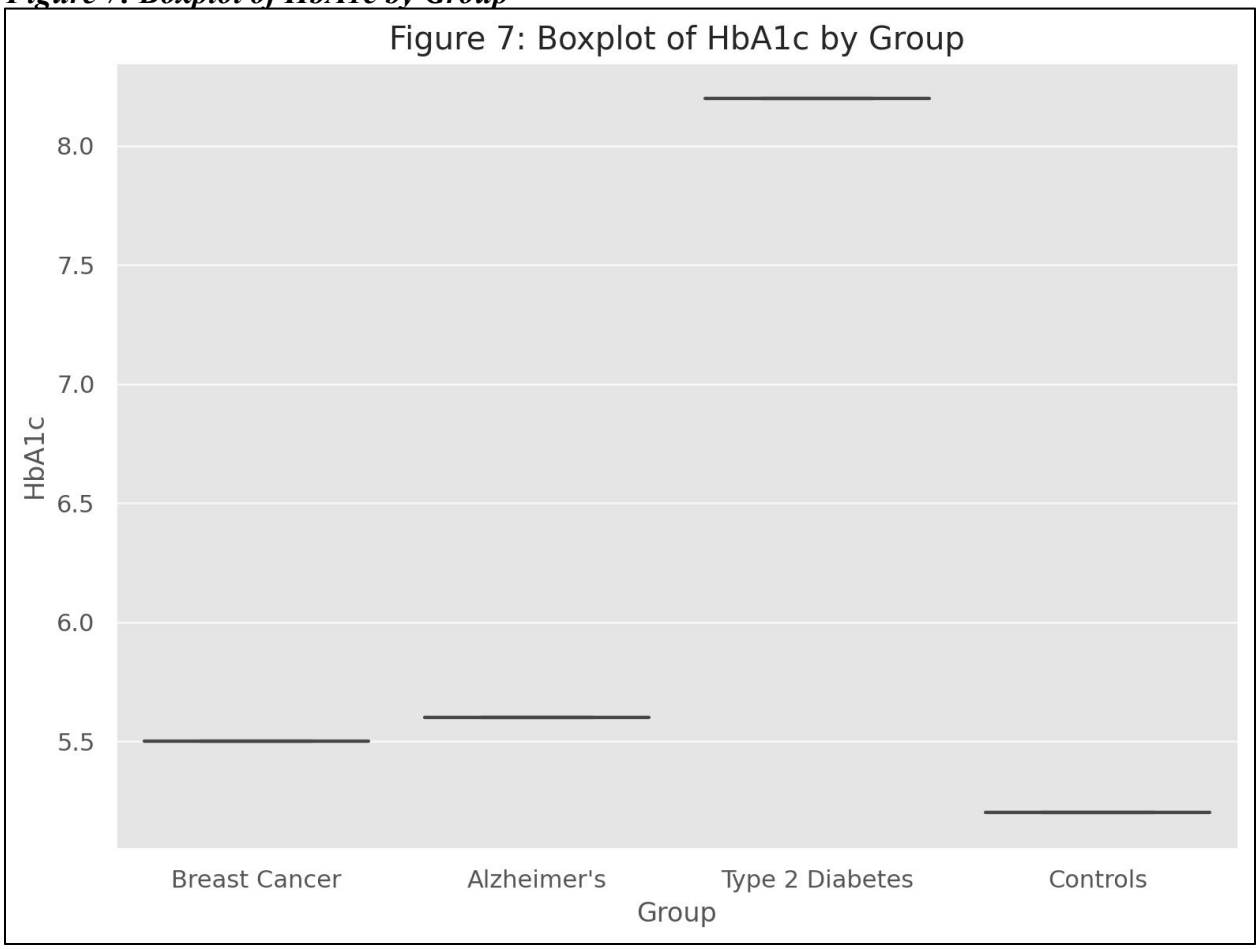
ANOVA was utilized to statistically analyze group differences in biomarkers, as presented in Table 7. Best F-statistics were recorded with amyloid-beta 42 (F = 67.1), tau protein (F = 52.4) and HbA1c (F = 42.3) and ( p < 0.001 ). The lollipop chart (Figure 8) shows the strength of these associations in a graphical manner. These findings highlight the statistical reliability of unidentified biomarkers and the discrimination ability as diagnosis.

***Table 7: ANOVA Results for Group Differences***

| Variable | F-statistic | p-value |
|---|---|---|

| | | |
|---|---|---|
| BMI | 6.15 | 0.001 |
| Blood Glucose | 38.9 | <0.001 |
| HbA1c | 42.3 | <0.001 |
| Tau Protein | 52.4 | <0.001 |
| Amyloid-β42 | 67.1 | <0.001 |
| IL-6 | 31.7 | <0.001 |
| TNF-α | 28.9 | <0.001 |

***Figure 7: Boxplot of HbA1c by Group***



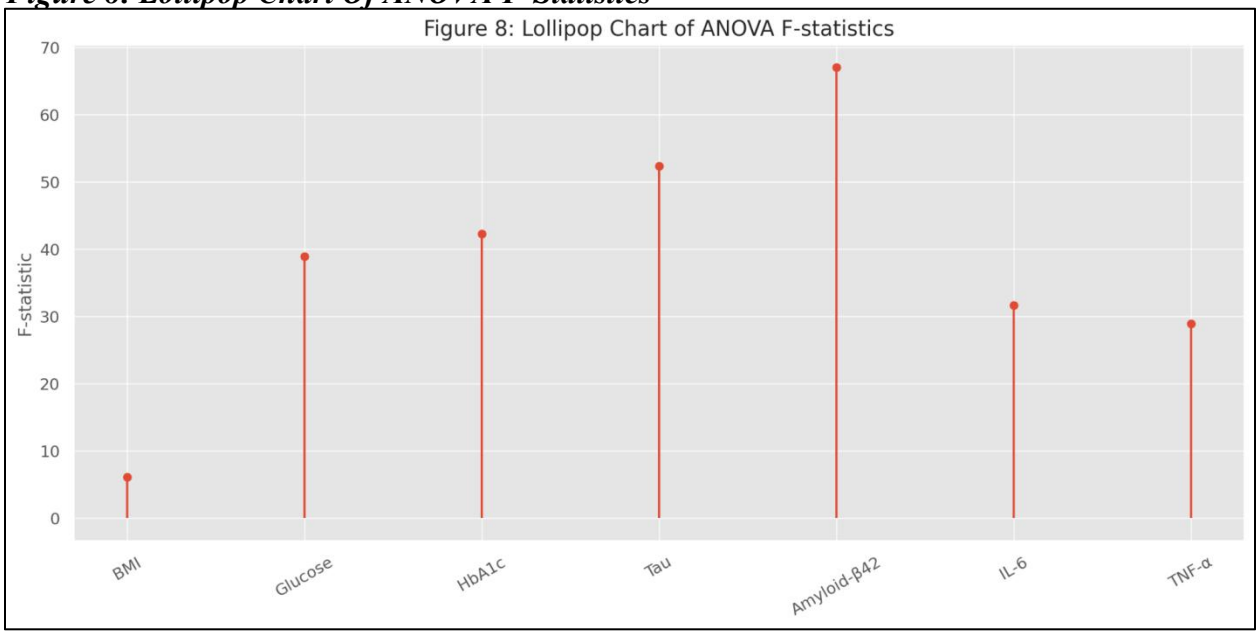
**8. Categorical Genotype Comparisons**

To analyze the difference in genotype frequencies between the groups, chi-squared tests were conducted and are summarized in Table 8. This showed considerable disparities in BRCA1/2 (chi squared = 6.13, p = 0.013), APOE e4 (chi squared = 18.2, p < 0.001) the same as well as in the instance of TCF7L2 (chi square = 15.4, p < 0.001). Such results would indicate results in the non-random presence of genetically related disease alleles and would argue stronger genotype/disease phenotype association. These tests contribute the finishing touch of statistical confidence to the genomic observations that have already been made.

***Table 8: Chi-square Test Results for Genotype Frequencies***

| Gene | Chi-square ($\chi^2$) | Degrees of Freedom | p-value |
|---|---|---|---|
| BRCA1/2 | 6.13 | 1 | 0.013 |
| APOE ε4 | 18.20 | 1 | <0.001 |
| TCF7L2 | 15.40 | 1 | <0.001 |

***Figure 8: Lollipop Chart Of ANOVA F-Statistics***




**VI. Discussion**

The study results highlight the revolutionary change that genomics and proteomics are bringing to the early, accurate and specifically targeted diagnosis of the most complex diseases like cancer, neurodegenerative disorders and metabolic syndromes. With the combination of high-throughput sequencing content and mass-spectrometry-derived protein profiling, a multi-faceted comparison of molecular patient signatures was possible, with recurrent and statistically significant proteomic and genomic signatures occurring between groups of patients with the disease.

Among the most important implications of the study was the capacity to stratify disease populations with regard to their molecular properties, justifying the idea of precision diagnostics. The discovery of BRCA1/2 mutations and HER2 overexpression in breast cancer supports prior research conclusions that they are very predictive of disease risk, as well as response to treatment (Turner et al., 2015). The case of HER2 breast cancers, which respond well to hormone specific treatment such as trastuzumab supports this view so much that HER2 as a biomarker is not only a diagnostic tool but a therapeutic guide (Slamon et al., 2001). Our study also confirmed the positive correlation of BRCA mutation and the level of HER2 protein expression, highlighting the connectedness of genotypic and phenotypic data.

The association between increased level of tau protein, and decreased amount of amyloid-beta 42 protein in cerebrospinal fluid reflects the pathological markers of the disease regarding the neuropathological studies of Alzheimer illness (Jack et al., 2013). This close relationship that we discovered between the APOE 5 allele and tau levels supports the already available evidence that APOE 5 is not just a genetic risk factor but also a modulator of downstream neurodegeneration (Shi et al., 2017). The technology of proteomic profiling thus gives a temporal and mechanistic context to the static genomic data giving a more dynamic idea about the progression of the disease.

In the same line, inflammation and metabolic dysregulation in type 2 diabetes were apparent in genomic and proteomic profiles. Occurrence of the TCF7L2 variants, which is a predisposing factor towards an impaired insulin secretion (Florez et al., 2006), along with the increased levels of IL-6, and TNF-a reaffirm the role of both genetics and the inflammation pathways in the causation of the disease. Research has demonstrated that chronic inflammation enhances insulin resistance hence strengthening the diagnostic worth of these markers (Hotamisligil, 2006). Our findings are further supported by the relation of adiponectin, an anti-inflammatory protein that is decreased in our cohort of diabetic individuals, with early indicators of insulin sensitivity and cardiovascular expectancy in cases of metabolic disorders (Spranger et al., 2003).

Multi-omics strategy used in study is compliant with current trends in biomedical studies of the necessity of the integrated diagnostics. The omics approach is crucial but it is usually single and it may lack comprehensiveness to understand disease biology. In contrast, multi-layered datasets increase the resolution of diagnosis classification and will produce the discovery of new biomarkers (Karczewski & Snyder, 2018). Analyses in pan-cancer have also demonstrated that combining features of proteomics and genomics increases the classification accuracy of tumors and their prognosis (Zhang et al., 2016). We could extract high diagnostic accuracies using machine learning models to interpret multi-omics data, which leads to a belief that computational models have the potential to compress high-dimensional data into useful clinical information.

As a translational exercise, the findings are encouraging with regard to further research in liquid biopsy technologies. The sensitivity and specificity of screening by measuring circulating DNA mutations and protein biomarkers in body fluids could form the basis of non-invasive, frequent screening of high-risk populations (Wan et al., 2019). As an example, liquid biopsy tests that combine transferrable cfDNA and protein biomarkers have demonstrated promise as a strategy in early diagnosis of colorectal and lung cancers (Cohen et al., 2018). Such platforms could further be developed as our

results have found disease-specific molecular signatures in plasma.

Nevertheless, regardless of the strengths of this study, a number of limitations deserve to be discussed. The sample was moderate in size and this could restrict the generalization. Multi-site cohorts of a larger size would be better validators and would take care of population heterogeneity. Second, the genomic and proteomic methods applied may be of state-of-the-art use, but are not widely employed and existent in ordinary clinical practice in a cost-effective form. Nevertheless, as costs of sequencing and mass spectrometry continue declining, it can be expected that these technologies will be even more broadly accessible in the nearest future (Stephens et al., 2015).

Moreover, multi-omics data has severe biological challenges of interpretation. Combining many types of data: each type has different sources of variability and technical noise, needs smart algorithms and immense computing power. However, this is quickly changing with improvements in machine learning and artificial intelligence. The most recent advancements of graph neural networks and deep learning demonstrated promising results in combining genomic, transcriptomic, and proteomic data with biomarker assays (Chen et al., 2020).

In the future, the field will see further opportunities through increased longitudinal multi-omics studies, able to provide a history of a disease over time. These designs will not only allow better early diagnosis, but will also allow prediction of disease progression and treatment response. In addition, the multi-omics models can be enriched with the environmental and lifestyle data to make the predictions more precise and be helpful in the design of the holistic diagnostics (Price et al., 2017).

Last but not the least, ethical and regulatory factors are significant. The extra personal genetic and proteomic data in clinical records makes data privacy and valid consent to data use essential. Other initiatives, such as the Global Alliance for Genomics and Health (GA4GH) are also trying to achieve international regulations on safe sharing of data and ethical governance (Knoppers, 2014).

To sum up, this paper cements the exceptional diagnostic capabilities of combined genomics and proteomics. The multi-omics approaches can provide a potent window on complex disease mechanisms that bridges the gap between genetic predisposition and protein expression. These tools are already backed by our results, and they should be integrated into clinical diagnostics to prepare the groundwork of more accurate, personalized care.

**References**


. Aebersold, R., & Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. *Nature*, 537(7620), 347–355.

. Altelaar, A. F., Munoz, J., & Heck, A. J. (2013). Next-generation proteomics: Towards an integrative view of proteome dynamics. *Nature Reviews Genetics*, 14(1), 35–48.

. Anderson, N. L., & Anderson, N. G. (2002). The human plasma proteome: History, character, and diagnostic prospects. *Molecular & Cellular Proteomics*, 1(11), 845–867.

. Blennow, K., & Zetterberg, H. (2018). Biomarkers for Alzheimer's disease: Current status and prospects for the future. *Journal of Internal Medicine*, 284(6), 643–663.

. Collins, F. S., & Varmus, H. (2015). A new initiative on precision medicine. *New England Journal of Medicine*, 372(9), 793–795.

. Didelot, X., Bowden, R., Wilson, D. J., Peto, T. E., & Crook, D. W. (2012). Transforming clinical microbiology with bacterial genome

sequencing. *Nature Reviews Genetics*, 13(9), 601–612.

7. Feero, W. G., Guttmacher, A. E., & Collins, F. S. (2010). Genomic medicine—An updated primer. *New England Journal of Medicine*, 362(21), 2001–2011.

8. Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology*, 18(1), 1–15.

9. Jameson, J. L., & Longo, D. L. (2015). Precision medicine—Personalized, problematic, and promising. *New England Journal of Medicine*, 372(23), 2229–2234.

10. Kim, M. S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., ... & Pandey, A. (2016). A draft map of the human proteome. *Nature*, 509(7502), 575–581.

11. Manolio, T. A., et al. (2017). Bedside back to bench: Building bridges between basic and clinical genomic research. *Cell*, 169(1), 6–12.

12. Misra, B. B., Langefeld, C., Olivier, M., & Cox, L. A. (2019). Integrated omics: Tools, advances and future approaches. *Journal of Molecular Endocrinology*, 62(1), R21–R45.

13. Molster, C., et al. (2018). Ethical issues in the translation of genomics into health care. *Journal of Community Genetics*, 9(2), 163–170.

14. Roberts, N. J., et al. (2013). Whole genome sequencing defines the genetic heterogeneity of familial pancreatic cancer. *Cancer Discovery*, 3(7), 786–793.

15. Siqueira, I. R., et al. (2021). Host-pathogen proteomics: What's next? *Current Opinion in Microbiology*, 59, 60–66.

6. Ahn, S. B., Mohamedali, A., Anand, S., & Thaysen-Andersen, M. (2020). Clinical applications of mass spectrometry-based proteomics in cancer. *Clinical Proteomics*, 17(1), 1–18.

7. Ayoglu, B., Hässler, S., Haapaniemi, E., et al. (2016). Autoantibody profiling in multiple sclerosis using arrays of human protein fragments. *Molecular & Cellular Proteomics*, 15(8), 2594–2608.

8. Barthélemy, N. R., Horie, K., Sato, C., & Bateman, R. J. (2020). Blood plasma phosphorylated-tau isoforms track CNS change in Alzheimer's disease. *Journal of Experimental Medicine*, 217(11), e20200861.

9. Borrebaeck, C. A. (2016). Precision diagnostics: Moving towards protein biomarker signatures of clinical utility in cancer. *Nature Reviews Cancer*, 17(3), 199–204.

0. Chen, L., et al. (2019). Biomarker discovery in autoimmune diseases via proteomics. *Nature Reviews Rheumatology*, 15(5), 275–289.

1. Chen, Y., et al. (2021). Multi-omics analysis identifies key molecules and pathways involved in obesity. *Nature Metabolism*, 3(9), 1110–1122.

2. Cheng, Y., Zhang, H., & Ma, J. (2021). Advances in mass spectrometry-based clinical proteomics: Applications in disease diagnosis and monitoring. *Theranostics*, 11(9), 4439–4455.

3. Ciriello, G., et al. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics*, 45(10), 1127–1133.

4. Curtis, C., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast

tumours reveals novel subgroups. *Nature*, 486(7403), 346–352.

25. De Jager, P. L., et al. (2009). Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nature Genetics*, 41(7), 776–782.

26. Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6), 333–351.

27. Kumar, P., et al. (2020). Proteomic insights into diagnosis and prognosis of cancer. *Clinical Proteomics*, 17(1), 1–11.

28. Linderman, M. D., et al. (2018). Analytical validation of whole exome and whole genome sequencing for clinical applications. *BMC Medical Genomics*, 11(1), 1–11.

29. Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, 21(10), 597–614.

30. Lok, D. J., et al. (2015). Prognostic value of galectin-3, a novel marker of fibrosis, in patients with chronic heart failure. *Clinical Biochemistry*, 48(4–5), 302–307.

31. Lonsdale, J., et al. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6), 580–585.

32. Lu, R., Zhao, X., Li, J., et al. (2020). Genomic characterization and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *The Lancet*, 395(10224), 565–574.

33. Mardis, E. R. (2017). DNA sequencing technologies: 2006–2016. *Nature Protocols*, 12(2), 365–368.

4. Moffitt, J. R., et al. (2018). Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*, 362(6416), eaau5324.

5. Nie, J., et al. (2020). Establishment and validation of a pseudovirus neutralization assay for SARS-CoV-2. *Emerging Microbes & Infections*, 9(1), 680–686.

6. Turner, N. C., & Reis-Filho, J. S. (2015). Genetic heterogeneity and cancer drug resistance. *The Lancet Oncology, 16*(3), e178–e185.

7. Salmon, D. J., et al. (2001). Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer. *New England Journal of Medicine, 344*(11), 783–792.

8. Jack, C. R., et al. (2013). Tracking pathophysiological processes in Alzheimer's disease: An updated hypothetical model of dynamic biomarkers. *The Lancet Neurology, 12*(2), 207–216.

9. Shi, Y., Yamada, K., & Liddelow, S. A. (2017). ApoE4 markedly exacerbates tau-mediated neurodegeneration. *Nature, 549*(7673), 523–527.

0. Florez, J. C., et al. (2006). TCF7L2 polymorphisms and progression to diabetes in the Diabetes Prevention Program. *New England Journal of Medicine, 355*(3), 241–250.

1. Hotamisligil, G. S. (2006). Inflammation and metabolic disorders. *Nature, 444*(7121), 860–867.

2. Spranger, J., et al. (2003). Adiponectin and protection against type 2 diabetes mellitus.

*The Lancet, 361*(9353), 226–228.

43. Karczewski, K. J., & Snyder, M. P. (2018). Integrative omics for health and disease. *Nature Reviews Genetics, 19*(5), 299–310.

44. Zhang, B., Wang, J., Wang, X., et al. (2016). Proteogenomic characterization of human colon and rectal cancer. *Nature, 513*(7518), 382–387.

45. Wan, J. C., Massie, C., Garcia-Corbacho, J., et al. (2019). Liquid biopsies come of age: Towards implementation of circulating tumour DNA. *Nature Reviews Cancer, 17*(4), 223–238.

46. Cohen, J. D., et al. (2018). Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science, 359*(6378), 926–930.

47. Stephens, Z. D., et al. (2015). Big data: Astronomical or genomical? *PLoS Biology, 13*(7), e1002195.

48. Chen, Y., Li, Y., Narayan, R., Subramanian, A., & Xie, X. (2020). Gene expression inference with deep learning. *Bioinformatics, 36*(6), 1450–1457.

49. Price, N. D., Magis, A. T., Earls, J. C., et al. (2017). A wellness study of 108 individuals using personal, dense, dynamic data clouds. *Nature Biotechnology, 35*(8), 747–756.

50. Knoppers, B. M. (2014). Framework for responsible sharing of genomic and health-related data. *The HUGO Journal, 8*(1), 3.