## Introduction

The advent of high-throughput transcriptomic profiling algorithms has transformed the clinical and research landscape of triple-negative breast cancer (TNBC), a heterogeneous and therapeutically challenging subtype lacking hormone receptor and HER2 amplification. Population-scale studies have established robust TNBC molecular subtypes—basal-like immune suppressed (BLIS), basal-like immune activated (BLIA/IM), luminal androgen receptor (LAR), and mesenchymal (MES)—each bearing implications for prognosis and response to emerging therapies (Burstein et al., 2015; Jiang et al., 2019; Hu et al., 2021)[1–3]. Multiple clinical trials, including FUTURE and I-SPY2, are now exploring subtype-directed stratification and therapy; yet, in practice, the translation of subtyping to individualized care remains limited, especially as clinical genomics increasingly yields "n=1" data sets of variable depth and coverage (Liu et al., 2020).

Single-patient (‘n=1’) multi-omic analysis is central to both research and precision oncology, yet introduces substantial practical and conceptual challenges. Batch effects, sequencing platform differences, and gene coverage disparities can undermine comparability with reference cohorts (Leek et al., 2010; Johnson et al., 2007; Nygaard et al., 2016)[5–7]. Furthermore, limited marker detection (relative to canonical signatures), ambiguous “borderline” expression phenotypes, and the lack of robust statistical uncertainty metrics complicate the assignment of an individual unambiguously to a single TNBC subtype (Lazar et al., 2013; Liu et al., 2020). Although recent cohort-based studies have attempted to develop transferable centroids and cross-cohort normalization strategies (Hu et al., 2021), the real-world informativity of such methods in the context of incomplete, technically heterogeneous n=1 data is unresolved.

Here, we present a comprehensive single-case molecular subtyping of TNBC. Using harmonized, batch-corrected integration of one patient’s RNA-seq with 716 reference TNBCs from the TCGA, FUSCC, and Burstein datasets, we systematically interrogate the feasibility and interpretive boundaries of both genome-wide and curated marker-set–based approaches. Importantly, we examine not only subtype assignment per se, but the degree of ambiguity and confidence underlying this designation—quantified by proximity metrics (correlation, distance to centroids), classifier discordance, and visualization—while maintaining transparency about the specific technical contingencies and limitations imposed by the available data. Our objective is to inform both clinician and patient, and to set realistic expectations for the practical use of molecular subtyping in n=1 TNBC cases.

## 

## **Data and Methods**

### **Patient Sample Processing and Gene Expression Quantification**

The patient RNA-seq data was processed from aligned BAM files using the Rsubread package. Gene-level read counts were generated using featureCounts with the following parameters: exon-level counting using GENCODE v19 annotation, minimum 1 bp overlap between reads and features, exclusion of multi-mapping reads, and unstranded library assumption. Raw read counts were filtered to retain only protein-coding genes with detectable expression.

Expression values were normalized as Fragments Per Kilobase Million (FPKM) using the formula: FPKM = (raw\_count × 10⁹) / (gene\_length × total\_mapped\_reads). Genes with FPKM ≥ 0.01 and non-missing gene symbols were retained for analysis. For duplicate gene symbols, the transcript with highest expression was selected. Quality control included verification of housekeeping gene expression (ACTB, GAPDH, B2M, HPRT1) and confirmation of low expression of hormone receptor genes (ESR1, PGR, ERBB2) consistent with triple-negative status.

### **Datasets and Sample Characteristics**

We analyzed 717 triple-negative breast cancer (TNBC) samples from four sources: The Cancer Genome Atlas (TCGA; n=158), the Fudan University Shanghai Cancer Center cohort (FUSCC; n=360), the Burstein reference dataset (n=198), and one patient case. All reference samples (n=716) had complete subtype annotations distributed as follows: BLIS (n=266, 37.2%), IM (n=184, 25.7%), LAR (n=138, 19.3%), and MES (n=128, 17.9%). The TCGA cohort showed subtype distribution of BLIS (67), IM (43), LAR (20), and MES (28). The FUSCC cohort contained BLIS (139), IM (87), LAR (81), and MES (53) samples. The Burstein dataset included BLIS (60), IM (54), LAR (37), and MES (47) samples.

### **Gene Expression Data Processing**

Expression data from all platforms were log₂-transformed (log₂(expression + 1)) prior to analysis. From the original 78 Burstein TNBC marker genes, we identified 26 genes (33.3%) that were present across all datasets after intersecting gene annotations. The final analysis utilized 4,887 genes common to all datasets for the pan-genomic approach and 26 Burstein marker genes for the targeted approach. Missing Burstein genes (n=52) included key markers such as ESR1, AR, FOXA1, and CD36, primarily due to limited gene coverage in the patient RNA-seq data (6,952 genes vs 16,090+ in reference datasets).

### **Batch Effect Correction**

ComBat batch correction (sva package) was applied to the 716 reference samples across three technical batches (TCGA, FUSCC, Burstein) using TNBC subtype as the biological covariate to preserve subtype-specific expression patterns. The patient sample was added post-correction to avoid singleton batch complications while enabling comparison against the harmonized reference space.

### **Linear Discriminant Analysis**

We performed supervised classification using Linear Discriminant Analysis with 716 reference samples as training data and the patient sample as the test case. Three discriminant functions were computed to optimally separate the four TNBC subtypes. Classification was performed using two feature sets: (1) all 4,887 common genes and (2) the 26 available Burstein marker genes. Patient classification was based on proximity to subtype centroids in the discriminant space.

### **Visualization and Pattern Analysis**

Hierarchical clustering heatmaps were generated using Ward.D2 linkage with Euclidean distance. Expression data were z-score normalized across samples for visualization purposes only. To address extreme outlier effects, we implemented adaptive robust scaling that excluded the top and bottom 10% of genes (approximately 4 genes for the 26-gene set) from color scale determination while preserving actual expression values. Subtype centroids were calculated as mean z-scores within each subtype from the batch-corrected reference data.

This methodology enabled both genome-wide and targeted marker-based classification while controlling for technical batch effects and optimizing visualization of expression patterns.

## **Results**

### **Dataset Characteristics and Quality Control**

A total of 717 TNBC samples were analyzed, comprising 158 TCGA samples, 360 FUSCC samples, 198 Burstein reference samples, and one patient case. All 716 reference samples (100%) had complete subtype annotations with the following distribution: BLIS (n=266, 37.2%), IM (n=184, 25.7%), LAR (n=138, 19.3%), and MES (n=128, 17.9%). ComBat batch correction was successfully applied to harmonize expression data across the three reference datasets while preserving subtype-specific expression signatures.

### **Gene Expression Analysis**

Of the 78 established Burstein TNBC marker genes, 26 genes (33.3%) were available across all datasets for analysis. Gene availability varied by platform: 74/78 genes in TCGA, 74/78 in FUSCC, 75/78 in the Burstein dataset, and 27/78 in the patient sample. The limited gene coverage in the patient data (6,952 genes) compared to reference datasets (≥16,090 genes) resulted in the exclusion of 52 key Burstein markers, including ESR1, AR, FOXA1, and CD36.

### **Linear Discriminant Analysis Classification**

Two parallel LDA classifications were performed using different feature sets. Using all 4,887 common genes, the patient was classified as IM subtype. However, using the 26 available Burstein marker genes, the patient was classified as BLIS subtype, representing a discordant classification between the two approaches.

**Hierarchical Clustering and Expression Pattern Analysis**

Unsupervised hierarchical clustering analysis using the 26 Burstein marker genes revealed the patient's strongest similarity to the IM subtype centroid (Pearson correlation r=0.392), followed by BLIS (r=0.308), with negative correlations to LAR (r=-0.431) and MES (r=-0.576). Euclidean distance analysis confirmed closest proximity to IM (distance=9.002) and BLIS (distance=9.087) centroids.

### **Differential Gene Expression Profile**

The patient exhibited a distinct expression profile characterized by extreme upregulation of immune-related genes. The most highly expressed genes included HERC5 (z-score=5.27), LAMP3 (z-score=4.26), and GBP5 (z-score=3.55), all associated with interferon response pathways. Additional upregulated genes included PROM1 (z-score=1.34) and ST3GAL6 (z-score=1.07). Conversely, the patient showed significant downregulation of SPOCK1 (z-score=-2.36), AGTR1 (z-score=-2.04), CD2 (z-score=-1.90), and GPX3 (z-score=-1.66).

### **Subtype-Specific Gene Signature Analysis**

Gene-by-gene analysis revealed that the patient's expression pattern matched IM-dominant signatures for key immune activation genes (HERC5, LAMP3, GBP5), where IM showed the highest expression among all subtypes. However, the patient also exhibited concordance with BLIS-dominant patterns for seven genes, including downregulation of CD2, PTGER4, and SIDT1, and upregulation of BCL11A. Only one gene (CHI3L1) showed concordance with the LAR-dominant pattern, while no genes matched MES-dominant signatures.

### **Classification Confidence and Consistency**

The correlation-based similarity analysis indicated low confidence in subtype assignment (r=0.392 for best match), suggesting mixed or atypical molecular features. The discordance between LDA methods (IM vs BLIS classification) and the relatively weak correlations to established subtype centroids indicated that this patient case exhibits features spanning multiple TNBC subtypes rather than fitting clearly into a single molecular category.

## Discussion

This single-patient analysis underscores both the potential and inherent constraints of precision molecular subtyping in TNBC. By integrating rigorous batch effect correction, cross-dataset harmonization, and both targeted and genome-wide classifier application, we sought to analogize standard population-calibrated genomics with the interpretive granularity and limitations of a real-world, individualized n=1 context.

## Data Integration and Technical Rigor:

Our methodology reflects multiple principles for analytic robustness: all reference samples were processed with uniform log2(FPKM+1) transformation, stringent gene intersection, and subtype-preserving empirical Bayes batch correction (ComBat, using subtype as a covariate) (Johnson et al., 2007). The patient sample, sequenced on a non-matched platform and with limited gene coverage (6,952 genes, 27/78 Burstein markers covered), was mapped into this harmonized reference space only after correction, thus circumventing the “singleton batch” artifact and guarding against artificial overcorrection or subtype-batch confounding (Lazar et al., 2013).

LDA was implemented with both a 4,887-gene genome-wide feature set and a 26-gene marker subset (Burstein et al., 2015), allowing direct assessment of the sensitivity of classification to gene selection and technical sparseness. The discordant subtype assignments—IM/BLIA per genome-wide LDA vs. BLIS by marker-restricted LDA—mirror published observations of classifier instability when marker panels are incomplete or platform differences dominate variance (Liu et al., 2020). Our approach to ambiguity assessment is in line with recent recommendations: use of both proximity (Euclidean, correlation) to subtype centroids and qualitative pattern visualization (heatmaps, LDA projections) to contextualize “soft” subtype assignment (Jiang et al., 2019; Hu et al., 2021).

## Biological Interpretation and Subtyping Ambiguity:

The classification ambiguity in our case is not merely technical. Over a third of cases in population cohorts likewise fail to match any single centroid with high confidence, typically reflecting either biological admixture (“borderline” or “hybrid” states) or tumor microenvironmental complexity (Hu et al., 2021; Wang et al., 2024). Our patient’s profile, marked by strong upregulation of ISGs (GBP5, HERC5, STAT1, LAMP3) alongside repression of canonical T-cell and adaptive immunity genes (CD2), typifies so-called “immune ghosting”—a phenotype increasingly recognized as reflecting transient or spatial immune exclusion, and predictive of mixed or intermediate response to immunotherapy (Savas et al., 2018; Qiu et al., 2024). This paralleled the lack of coherent assignment to BLIS or BLIA, and the moderate correlations (max r≈0.39) to centroids, as seen in recent cohort analyses (Liu et al., 2020; Hu et al., 2021).

Notably, the patient’s genomic context (TP53 and NOTCH1 mutations, MCL1/MYC amplification, CREBBP loss) may further reinforce the axis of immune suppression and stemness (Teschendorff et al., 2023), yet signature scores and IPS (Immune Profile Score) remain relatively high, sustaining some rationale for immunomodulatory strategies and highlighting potential for spatial or clonal heterogeneity in immune responsiveness (Schmid et al., 2020).

## Translational and Clinical Utility:

Our findings exemplify the tension between methodological rigor and clinical pragmatism in n=1 applications. While “hard” subtype classification is equivocal in this case, the analysis uncovers intelligible biological axes (immune activity, immune suppression, stem-like signatures) that are themselves directly relevant to therapeutic option generation, trial selection, and patient–clinician dialogue (Dalton et al., 2022; Hu et al., 2021). Recent umbrella and platform trials (FUTURE, KEYNOTE-522) have prospectively incorporated multi-modal signatures and allow for “intermediate” or “hybrid” subtype assignment in stratification schemas (Hu et al., 2021; Schmid et al., 2020).

Limitations and Future Directions:  
This analysis is subject to well-documented limitations: technical confounding from unmatched sequencing, incomplete marker coverage (particularly missing AR, ESR1, FOXA1, CD36), lack of formal statistical uncertainty quantification, and the risk of overinterpretation from modest maximum correlations or distances to centroids. Furthermore, the inability to batch-correct the patient and reference data jointly may under- or over-adjust for real biological signal (Leek et al., 2010; Nygaard et al., 2016). These limitations are intrinsic to the current clinical translation of n=1 genomics, and caution against overconfident assignment (‘quantitative uncertainty’ was not formally measured, but ambiguity is substantial as evidenced by pattern discordance and low similarity metrics).

Conclusion:  
Population-developed molecular subtyping frameworks, when applied to patient-level data with rigorous harmonization and transparent reporting of ambiguity, can yield hypothesis-generating and context-enriching insights—even when technical limitations and residual ambiguity persist. This approach maximizes both the scientific and practical value of patient-controlled omics data, supporting increasingly sophisticated and participatory models of cancer care (Dalton et al., 2022; Hu et al., 2021).

FIGURES



