The Curse of Dimensionality and Analysis of the RNA-Seq (HiSeq) PANCAN Data Set

Kat Campise

Introduction

The rapid expansion of data dimensionality in healthcare and biomedical research has unlocked unprecedented detail—from millions of genetic variants to high-resolution imaging voxels and continuous wearable signals. Digital health data routinely combines modalities such as clinical labs, genomics, imaging, speech samples, and wearable time series data, producing feature spaces where the number of variables often dwarfs the number of patients (Berisha et al., 2021). In these "large p, small n" settings, traditional intuitions about averages and similarity break down: most points cluster near the "edges" of the space, distances between any two observations converge, and data become exceedingly sparse (Aggarwal, Hinneburg, & Keim, 2001).

The curse of dimensionality manifests in several interrelated ways. First, the volume of feature space grows exponentially with each added dimension, so that even large clinical cohorts leave vast “blind spots” where no training data exist, undermining a model’s ability to generalize (Berisha et al., 2021). Second, distance metrics lose discrimination: in high dimensions, nearest- and farthest-neighbor distances become nearly equal, rendering clustering and k-nearest-neighbor methods unstable (Aggarwal et al., 2001). Third, computational complexity skyrockets: exhaustive searches or sampling over thousands of features become infeasible, and model fitting risks overfitting without commensurate increases in sample size (Debie & Shafi, 2019).

High dimensionality is a double-edged sword for personalized medicine. On one hand, richly detailed data streams promise finely tuned diagnostics and treatment protocols; on the other, finite sample sizes mean that ever-finer patient stratification yields ever-smaller cohort slices, limiting the applicability of group-average inference and risking “empty” stratifications (Barbour, 2019). For example, dividing patients by ten independent binary risk factors of 10%

prevalence each produces a one-in-ten-billion chance of finding a truly "similar" past case, regardless of database size (Barbour, 2019).

Sampling distributions and resampling methods offer a path through these challenges. By repeatedly drawing subsets (e.g., bootstrap or stratified samples) and examining the variability of key statistics—means, variances, even higher moments—stability of inferences can be assessed and over-optimistic performance estimates guarded against. In digital medicine, where "blind spots" have led to catastrophic deployment failures, sampling distributions help quantify uncertainty and guide feature selection or dimensionality reduction (Berisha et al., 2021).

Subsequent sections are organized as follows:

- **Literature Review on the Curse of Dimensionality**: Examination of how high-dimensional spaces induce sparsity, distance-concentration phenomena, and exponential growth of feature-space volume, undermining generalization and inflating computational cost (Aggarwal, Hinneburg, & Keim, 2001; Berisha et al., 2021). Discussion of sampling-distribution approaches—bootstrap, stratified sampling, ensemble feature-bagging—as means to assess estimator stability and mitigate blind-spot effects in statistical learning (Debie & Shafi, 2019; Eisenberg, Hubbard, Trostle, & Cangemi, 2019).
- **Data Preparation and Exploration:** The selected dataset is introduced, including total number of variables, data modality (e.g., continuous labs, categorical diagnoses), and domain context (personalized medicine). Construction of a comprehensive data dictionary (Variable Name, Data Type, Level of Measurement) and initial exploratory analyses—summary statistics, missingness patterns, pairwise correlations—to characterize dimensionality and identify preprocessing needs (Hubbard, Trostle, Cangemi, & Eisenberg, 2019; Maitra, Hossain, Hasib, & Shishir, 2020).

- **Programming Sampling Distributions in R:** Implementation of at least three sampling techniques—simple random sampling, stratified sampling (by key clinical strata), and systematic sampling (fixed-interval draws)—to generate repeated subsets. Calculation of sample means, variances, skewness, and kurtosis for selected features, followed by comparison to full-data estimates to evaluate representativeness. Generation of histograms and Q–Q plots for both samples and population to assess distributional fidelity (Debie & Shafi, 2019).
- **Analysis and Reporting:** Synthesis of sampling outcomes to determine which techniques best preserve high-dimensional population characteristics. Visualization of comparative results via box plots and density overlays. Discuss how sampling distributions inform model reliability, reduce overfitting risk, and guide feature-selection decisions in personalized medicine workflows (Barbour, 2019; Berisha et al., 2021).

Understanding the Curse of Dimensionality

The "curse of dimensionality" describes the phenomenon where the increase in dimensionality of data results in data sparsity, diminished effectiveness of traditional statistical methods, and challenges in data interpretation. In high-dimensional spaces, data points tend to cluster near the edges, and distances between points become nearly indistinguishable, posing significant analytical difficulties (Aggarwal, Hinneburg, & Keim, 2001). Specifically, Aggarwal et al. (2001) illustrated the counterintuitive behavior of distance metrics in high-dimensional spaces, where the contrast between the nearest and farthest neighbor becomes insignificant, leading to instability in clustering and classification algorithms.

The implications of dimensionality extend across various analytical tasks. In machine learning, the curse of dimensionality negatively impacts the performance of supervised learning

classifiers by introducing increased variance and the likelihood of overfitting, especially when training data is limited (Debie & Shafi, 2019). Statistical modeling similarly faces difficulties in model estimation, as the complexity and variance inflate rapidly with dimensional growth, leading to models with poor generalization capabilities and susceptibility to overfitting (Debie & Shafi, 2019). Visualization of high-dimensional data becomes increasingly challenging, as the dense clustering of points within low-dimensional projections obscures structural insights and can mislead interpretations (Laa, Cook, & Lee, 2020). Computational complexity further exacerbates the problem, with computational resources and runtime increasing exponentially with each additional dimension, making exhaustive analysis practically infeasible (Aggarwal et al., 2001).

Domain-specific impacts are also significant, particularly in healthcare and precision medicine. Berisha et al. (2021) emphasized the particular challenge of digital medicine, where large-scale, high-dimensional multimodal data—such as clinical variables, imaging data, genome sequencing, and continuous wearable signals—cause issues in model robustness and reliability. Similarly, Catchpoole, Kennedy, Skillicorn, and Simoff (2010) highlighted how the curse of dimensionality complicates precision medicine by producing excessively sparse subgroups, hindering the ability to reliably infer individual patient outcomes from group-based analyses.

Sampling Distributions as a Mitigation Strategy

Sampling provides a fundamental statistical approach to mitigating the challenges posed by high-dimensional data analysis. Through strategically drawing representative subsets from larger datasets, sampling methods enable researchers to preserve key characteristics of the original data while significantly reducing dimensionality and computational burdens.

Sampling methods offer substantial benefits when addressing the curse of dimensionality. For instance, intelligent sampling strategies such as quasi-Monte Carlo, Latin hypercube, or importance sampling methods, reduce computational load and enhance model generalization by systematically capturing data from critical regions of the high-dimensional space (Loyola, Pedergnana, & García, 2016). The core concept is to ensure that smaller subsets retain structural and statistical characteristics of the full dataset, thereby reducing "blind spots" and preserving analytical accuracy (Loyola et al., 2016).

Moreover, understanding data-generating mechanisms provides critical insights that inform sampling strategies. Hubbard et al. (2019) emphasized the role of participant observation and mechanistic modeling as effective approaches in capturing the underlying data-generating processes. Such observational and theoretical approaches help researchers define more intelligent sampling strategies, thereby enhancing dimensionality reduction methods and improving inferential reliability (Hubbard et al., 2019).

Support vector machines (SVMs) have also been employed effectively in high-dimensional data mining, owing to their capacity to handle large numbers of features via kernel methods. Jiang (2025) demonstrated that combining SVMs with intelligent sampling techniques enhances their generalization performance, showing promise as a practical solution to high-dimensional challenges.

Personalized medicine benefits significantly from carefully designed sampling strategies (see Table 1). Catchpoole et al. (2010) noted that sparsity in high-dimensional datasets can, paradoxically, be advantageous, concentrating signals and facilitating the detection of deviations that indicate patient-specific conditions. Therefore, sampling methods tailored to typical and

atypical regions can support the precise identification of patient groups, offering more targeted diagnostic and therapeutic options (Catchpoole et al., 2010).

Table 1

Comparative Analysis of Sampling Strategies for Precision Medicine

Sampling Technique	Description	Benefits	Limitations	Precision Medicine Use Case
Simple Random Sampling	Each observation has an equal probability of selection	Easy to implement; unbiased estimates	May not preserve rare subgroups or stratified patterns	General risk prediction model validation using EHR cohort (Yang, Fridgeirsson, Kors, Reps, & Rijnbeek, 2024)
Stratified Sampling	Divides the population into strata and samples proportionally or equally within	Ensures representation across disease subtypes or demographic groups	Requires accurate prior knowledge of strata	Equal representation of cancer subtypes in genomics data modeling (Livne & Efroni, 2024)
Systematic Sampling	Selects every k-th element from an ordered list	Efficient for large datasets; simple to execute	Risk of periodicity bias if order has an underlying pattern	Biomarker sampling from temporal wearable data (Lohr, 2010)
Bootstrap Resampling	Resamples with replacement from the dataset to create new pseudo-samples	Allows robust estimate of variance and confidence intervals; supports ensemble learning	Computationally intensive; risk of replicating noise	Uncertainty estimation for treatment-response predictors (Chen & Ishwaran, 2012)
Quasi-Monte Carlo Sampling	Uses low-discrepancy sequences to fill space more	Improved coverage of high-dimensional space; lower	Implementation complexity; not truly random	Parameter space exploration in drug-dosage optimization

	uniformly than random sampling	variance in estimations		models (Hickernell & Owen, 2018)
Latin Hypercube Sampling (LHS)	Ensures uniform sampling across each dimension	Efficient space-filling in high dimensions; fewer samples needed than full factorial sampling	Less effective if variables are strongly correlated	Multi-drug combination effect modeling (Katamesh, Abbas, & Mahmoud, 2024)
Importance Sampling	Samples more frequently from regions with greater importance (e.g., high variance or clinical relevance)	Improves estimator efficiency for rare but critical cases	Requires prior knowledge of importance distribution	Adverse reaction prediction for rare genetic mutations (Han, Kang, Eskin, & Schnell, 2014)
Active Sampling / Query-Based	Dynamically selects samples based on model uncertainty or informativeness	Focuses on uncertain or borderline cases; reduces labeling cost	Complex implementation; depends on model feedback loop	Training clinical decision support tools with limited expert review time (Blee et al., 2022)
Smart Sampling / Hybrid Techniques	Combines stratified, space-filling, and adaptive sampling methods	Balances representativeness, efficiency, and model-guided adaptation	Can be computationally complex and dataset-specific	Genomic feature selection and prediction in small subpopulations with rare phenotypes (Loyola, Pedergnana, & Garcia, 2016)

Data Preparation and Exploration

The dataset selected for this analysis is the RNA-Seq (HiSeq) PAN-CAN gene expression dataset, available through the UCI Machine Learning Repository (2016). This dataset was chosen due to its relevance to personalized medicine and its wide use in cancer classification and biomarker discovery research.

- Dataset Name and Source: RNA-Seq (HiSeq) PAN-CAN gene expression dataset; UCI Machine Learning Repository
- Number of Variables: 20,531 gene expression features
- Number of Observations: 801 patient samples
- Type of Dataset: Tabular (gene expression matrix; each row = patient sample, each column = gene)
- Domain Area: Biomedical (oncology and genomics). The dataset includes samples from patients with five tumor types: breast cancer (BRCA), kidney renal clear cell carcinoma (KIRC), colon adenocarcinoma (COAD), lung adenocarcinoma (LUAD), and prostate adenocarcinoma (PRAD).

Each variable corresponds to a specific gene's expression level, typically measured in fragments per kilobase of transcript per million mapped reads (FPKM) or a similar unit. All variables are continuous and represent measured expression levels for individual genes. A sample data dictionary is provided in Table 2.

Table 2

Example Data Dictionary for the RNA-Seq PAN-CAN Dataset

Variable Name	Data Type	Level of Measurement	Description
gene_01	Numeric	Ratio	Gene expression level (RNA-Seq)
gene_02	Numeric	Ratio	Gene expression level (RNA-Seq)
gene_03	Numeric	Ratio	Gene expression level (RNA-Seq)
gene_04	Numeric	Ratio	Gene expression level (RNA-Seq)

gene_05	Numeric	Ratio	Gene expression level (RNA-Seq)
...	...	...	...

This dataset is highly dimensional, with 20,531 gene expression features and only 801 patient observations, exemplifying a $p \gg n$ configuration. Such data structures are typical in transcriptomics and personalized medicine, where the vast number of molecular features presents significant challenges for model stability, interpretability, and generalization. High dimensionality increases the risk of overfitting, inflates variance in parameter estimates, and complicates standard inferential methods (Ma & Dai, 2011; Jolliffe & Cadima, 2016).

Variable Types and Distributions

All features are numeric and measured on a ratio scale, corresponding to RNA-Seq gene expression levels. Expression profiles typically follow skewed distributions, with most genes showing low expression and a minority highly expressed in specific tissues or tumor subtypes. As such, the variable landscape is heteroscedastic and sparse—characteristics that further justify dimensionality reduction prior to downstream modeling.

Exploratory PCA Visualization

Principal component analysis (PCA) was conducted as an unsupervised exploratory technique to assess sample structure and variance patterns. Prior to PCA, the gene expression matrix was cleaned by removing the non-numeric tumor type label and dropping all features with zero variance. The remaining variables were then scaled to unit variance, ensuring that highly variable genes would not dominate the principal component loadings due to magnitude alone. PCA was performed using the prcomp() function in R on the standardized matrix, producing principal components that represent orthogonal axes of maximal variance (Figure 1).

Figure 1

R Code for PCA Data Preparation and Exploration

```r
# 3. Load your data
data <- read.csv("data.csv", header = TRUE, row.names = 1)
labels <- read.csv("labels.csv", header = TRUE, row.names = 1)

# 4. Merge labels into the data frame
data$TumorType <- labels[match(rownames(data), rownames(labels)), 1]

# 5. Prepare numeric matrix for PCA
data_numeric <- data[, -ncol(data)]                # drop TumorType
# remove columns with zero variance
data_numeric <- data_numeric[, apply(data_numeric, 2, var) != 0]

# 6. Scale and run PCA
data_scaled <- scale(data_numeric)
pca <- prcomp(data_scaled)
pca_df <- data.frame(PC1 = pca$x[,1],
                     PC2 = pca$x[,2],
                     TumorType = data$TumorType)
```

The first two principal components (PC1 and PC2) were visualized in a scatter plot, with samples colored by tumor type (Figure 2). The resulting projection revealed several key structural insights. Most notably, kidney renal clear cell carcinoma (KIRC) cases form a distinct cluster along PC1, indicating a divergent transcriptomic profile from other tumor types. In contrast, breast (BRCA), prostate (PRAD), lung (LUAD), and colon (COAD) samples show greater overlap, suggesting shared variance components or more subtle intergroup differences. This observation provides early evidence of biologically meaningful separation, supports tumor labels' validity, and underscores PCA's utility for understanding intrinsic structure in high-dimensional biomedical data.

PCA is widely recognized as a foundational step in the exploratory analysis of high-dimensional omics datasets, including transcriptomics and genome-wide association studies, where it helps visualize latent structure, detect technical confounders, and inform preprocessing decisions (Ringnér, 2008; Price et al., 2006). PCA is often applied before model building to

assess heterogeneity, cluster separability, or potential confounding due to batch effects or population structure (Privé et al., 2020). Its interpretability, computational efficiency, and ability to reduce noise make it an indispensable tool in modern precision medicine workflows (Kadi et al., 2021; Ma & Dai, 2011).

Figure 2

PCA of RNA-Seq Gene Expression

Missing Values and Outliers

According to the dataset documentation, no missing values are present. However, outliers in expression data are expected due to biological heterogeneity and potential artifacts. These will need to be carefully evaluated during normalization and sampling.

Potential Challenges

Several key challenges arise from the dataset's structure:

- High Dimensionality: The large number of features relative to samples increases the risk of overfitting and complicates statistical inference.
- Multicollinearity: Genes may be co-regulated or functionally redundant, leading to correlated predictors.
- Computational Burden: Standard model fitting, visualization, and sampling can be computationally intensive at this scale.
- Feature Selection Sensitivity: The importance of specific genes may vary significantly depending on the sampling method used.

This initial preparation highlights the dataset’s structural complexity and establishes a foundation for subsequent sampling and statistical analysis.

Programming Sampling Distributions in R

This section evaluates the representativeness and reliability of sampling distributions derived from three distinct sampling techniques—simple random, stratified, and systematic sampling—in the context of high-dimensional RNA-Seq gene expression data. Each sampling method potentially affects analytical outcomes by capturing different aspects of the dataset's inherent biological variability, subgroup representation, and statistical characteristics. A comparative analysis, integrating statistical metrics (mean, variance, skewness, kurtosis) and visual diagnostics (histograms and QQ plots), helps illustrate how each sampling strategy uniquely influences the accuracy and robustness of downstream analyses crucial to personalized medicine research.

Environment and Libraries

All data processing and statistical analyses were conducted in R (version 4.5). The following R packages were utilized:

- dplyr for data manipulation
- caret for stratified sampling
- e1071 for skewness and kurtosis calculations
- ggplot2 and base graphics for visualization

These packages are widely used in statistical modeling and bioinformatics, especially in high-dimensional gene expression analysis where exploratory and inferential tasks must be automated and replicable.

Sampling Techniques Implementation

Three sampling techniques were implemented to investigate the performance of different sampling strategies in representing the distributional characteristics of high-dimensional biomedical data: simple random sampling, stratified sampling, and systematic sampling. A sample size of 200 was used for each method to simulate scenarios common in personalized medicine research, where subsampling from a limited cohort is often necessary due to cost or data sparsity (Lohr, 2010; Ma & Dai, 2011).

Simple Random Sampling (SRS)

Simple random sampling was implemented by randomly selecting 200 rows from the full dataset without replacement using the sample() function (Figure 3). This approach gives each patient sample an equal probability of inclusion and assumes that the underlying population is homogeneous.

Figure 3

Simple Random Sampling R Code

```r
# 1. Simple Random Sampling (SRS)
set.seed(1)
sample_srs <- data[sample(nrow(data), 200), ]
```

While easy to execute, SRS does not guarantee proportional representation of critical subgroups (e.g., tumor types). This limitation is particularly consequential in personalized medicine, where treatment-response relationships often hinge on stratified biological characteristics (West et al., 2010).

Stratified Sampling

Stratified sampling was performed using the createDataPartition() function from the caret package, stratifying by tumor type to ensure proportional representation across the five cancer subtypes (Figure 4). A 25% sampling fraction was applied within each stratum.

Figure 4

Stratified Sampling R Code

```r
# 2. Stratified Sampling
set.seed(1)
idx_strat <- createDataPartition(data$TumorType, p = 0.25, list = FALSE)
sample_strat <- data[idx_strat, ]
```

Stratification is critical in clinical omics research, where disease heterogeneity can obscure meaningful signals if not properly controlled. Ensuring each tumor type is proportionally sampled mitigates the risk of underrepresenting rare but clinically relevant subpopulations (Chen & Ishwaran, 2012).

Systematic Sampling

Systematic sampling was implemented by selecting every *k-th* sample from the dataset after calculating an interval (step = floor(n / 200)). This method is efficient for ordered data but assumes that the ordering does not introduce periodicity or bias (Figure 5).

Figure 5

Systematic Sampling R Code

```
# 3. Systematic Sampling
step <- floor(nrow(data) / 200)
idx_sys <- seq(1, nrow(data), by = step)
sample_sys <- data[idx_sys, ]
```

Though rarely used in omics studies, systematic sampling can be advantageous in resource-constrained clinical settings where real-time, on-the-fly sampling is necessary during data acquisition (Lohr, 2010).

Statistical Measures and Comparison

Five genes (gene_0 through gene_4) were selected due to their diverse distribution profiles, representing varying degrees of expression common in transcriptomics studies. Evaluating these genes provides insight into sampling-induced shifts in gene expression distribution (DeCarlo, 1997). To quantify how accurately each sampling method captures the underlying distributional properties of the full dataset, descriptive statistics (mean, variance, skewness, and kurtosis) were computed. These measures were selected as they collectively describe central tendency, variability, asymmetry, and tail behavior, crucial in understanding biological variation in gene expression data. An R function, compute_stats(), was defined to automate the calculation of these metrics across each subset (Figure 6).

Figure 6

R Code for Statistical Computations

```r
compute_stats <- function(df, vars) {
  data.frame(
    Variable = vars,
    Mean     = sapply(df[, vars], mean),
    Variance = sapply(df[, vars], var),
    Skewness = sapply(df[, vars], skewness),
    Kurtosis = sapply(df[, vars], kurtosis)
  )
}

# Compute stats
stats_full   <- compute_stats(data, genes)
stats_srs    <- compute_stats(sample_srs, genes)
stats_strat  <- compute_stats(sample_strat, genes)
stats_sys    <- compute_stats(sample_sys, genes)

# Label samples
stats_full$Sample  <- "Full Data"
stats_srs$Sample   <- "Simple Random"
stats_strat$Sample <- "Stratified"
stats_sys$Sample   <- "Systematic"

# Combine and display
all_stats <- rbind(stats_full, stats_srs, stats_strat, stats_sys)
knitr::kable(all_stats, caption = "Comparison of distribution metrics across samples and full data")
```

This function simplifies and standardizes the calculation of statistical summaries across datasets, ensuring consistent comparisons and reproducibility.

The comparative statistical results (Figure 7) provided important insights into the effectiveness of each sampling strategy in representing the original data. Specifically, **stratified sampling** most closely matched the full dataset across all metrics—mean, variance, skewness, and kurtosis. This indicates that stratified sampling preserves subgroup variability and accurately represents common and rare gene expression states. This is especially critical in personalized medicine, where subgroup identification directly influences treatment decisions (West et al., 2010; Loyola et al., 2016).

Systematic sampling also showed good representativeness, particularly in mean and variance, suggesting it effectively captures overall data structure when ordering is unbiased. However, some minor discrepancies in skewness and kurtosis were noted, likely reflecting periodicity or subtle ordering biases.

Conversely, simple random sampling (SRS) demonstrated greater variability in skewness and kurtosis metrics, particularly in genes exhibiting pronounced non-normal distributions (e.g., gene_0). This variability underscores SRS's limitations in accurately representing extreme or biologically significant expression patterns, thereby potentially reducing reliability in detecting rare yet clinically critical biomarkers (Chen & Ishwaran, 2012).

Figure 7

Comparison of Distribution Metrics Across Samples and Full Data

	Variable	Mean	Variance	Skewness	Kurtosis	Sample
gene_0	gene_0	0.0266416	0.0187278	6.0370678	44.632423	Full Data
gene_1	gene_1	3.0109095	1.4419867	-0.5038331	3.039542	Full Data
gene_2	gene_2	3.0953497	1.1355059	-0.0677021	3.011800	Full Data
gene_3	gene_3	6.7223054	0.4080893	0.7026404	4.240205	Full Data
gene_4	gene_4	9.8136121	0.2565801	0.2133859	2.918160	Full Data
gene_01	gene_0	0.0348611	0.0219967	4.7964277	28.768015	Simple Random
gene_11	gene_1	3.0008812	1.5175444	-0.5267126	3.027612	Simple Random
gene_21	gene_2	3.0620010	1.2056975	-0.0672862	3.270394	Simple Random
gene_31	gene_3	6.7007007	0.4736789	0.4177846	3.165120	Simple Random
gene_41	gene_4	9.8369080	0.2811660	0.2084971	2.720946	Simple Random
gene_02	gene_0	0.0426939	0.0356005	5.0409156	30.424076	Stratified
gene_12	gene_1	3.0713038	1.3818257	-0.5098035	3.158046	Stratified
gene_22	gene_2	3.1340914	1.0995262	-0.0657607	2.826483	Stratified
gene_32	gene_3	6.7426676	0.4069071	0.5504469	3.330340	Stratified
gene_42	gene_4	9.8065547	0.2037251	0.2876226	3.217875	Stratified
gene_03	gene_0	0.0421398	0.0313379	4.9169500	30.544174	Systematic
gene_13	gene_1	3.0219899	1.4947841	-0.3841578	2.948463	Systematic
gene_23	gene_2	2.9962236	1.1683663	-0.1856450	3.100094	Systematic
gene_33	gene_3	6.6697520	0.3636671	0.5845201	3.416263	Systematic
gene_43	gene_4	9.8006183	0.2925261	0.1982590	2.793347	Systematic

Thus, the statistical analysis demonstrates that sampling method selection significantly influences downstream analytical accuracy and interpretability in high-dimensional transcriptomics data commonly utilized in personalized medicine.

Visual Representations and Analysis

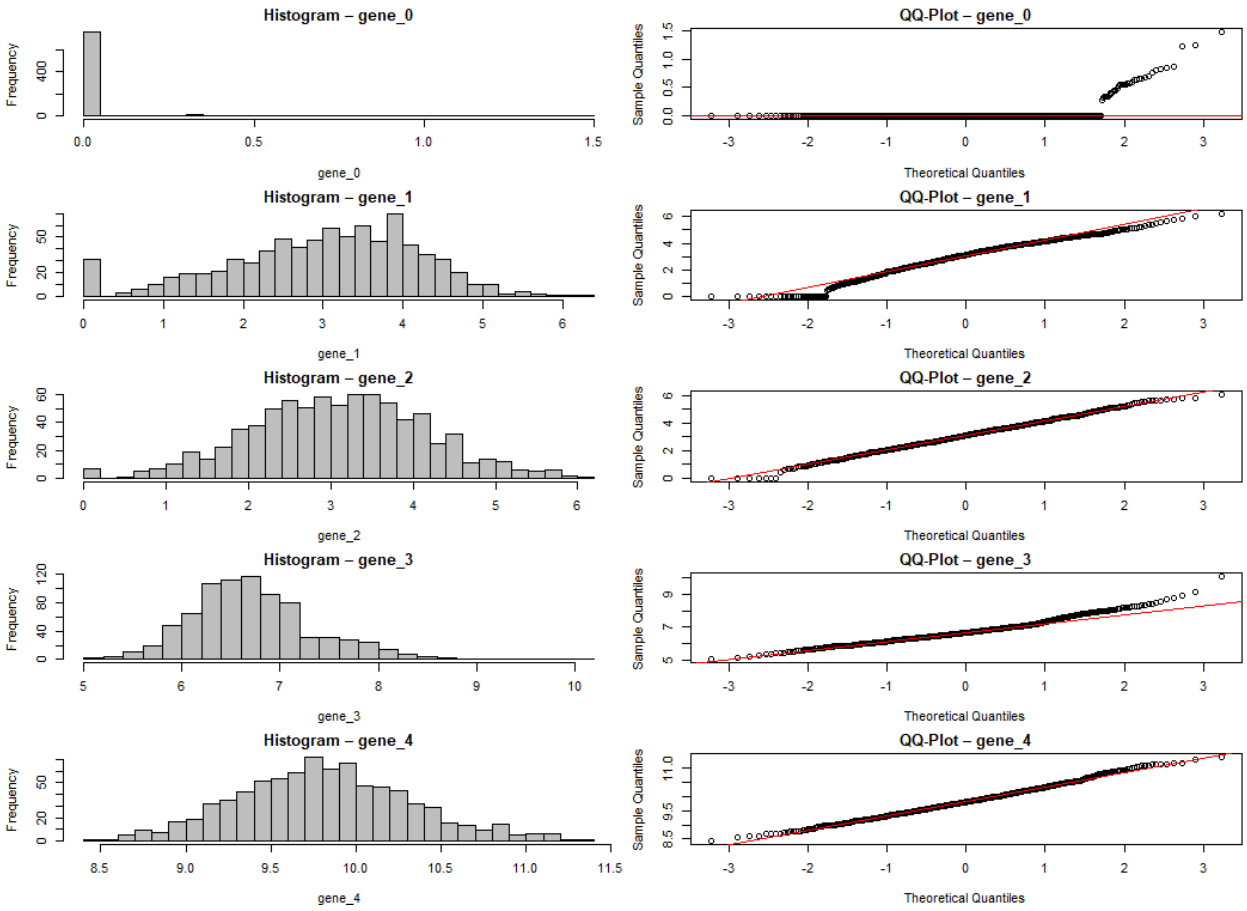
Histograms further clarified sampling impacts on gene expression distribution. Stratified and systematic samples better preserved the heavily skewed distribution of gene_0 in the full dataset (Figure 8). All methods adequately captured the more symmetrical distributions of gene_1 and gene_2, though stratified sampling consistently demonstrated superior fidelity to the original shape. The key visual insight is that stratified sampling consistently better captures expression diversity, essential for accurate subgroup identification—an integral aspect of personalized medicine (West et al., 2010).

QQ plots visualize how empirical gene distributions approximated theoretical normality (Figure 8). The full dataset exhibited significant departures from normality in gene_0, indicating that biological variability is common in clinical data. Stratified sampling and systematic sampling effectively mirrored this departure, maintaining authenticity of the biological signal. In contrast, SRS tended to underrepresent tail extremes, potentially diminishing detection of clinically relevant biological signals crucial for personalized diagnostics (Loyola et al., 2016).

The comparative analysis highlights that stratified sampling most effectively captures and preserves complex, clinically relevant gene expression structures essential in personalized medicine. Systematic sampling also shows promise under certain controlled conditions. In contrast, simple random sampling appears least reliable for maintaining critical biological variability, especially in high-dimensional datasets characteristic of precision oncology research. Choosing sampling techniques with explicit attention to these implications can significantly enhance the robustness and clinical validity of personalized medical insights.

Figure 8

Histograms and QQ Plots



Analysis and Reporting

Evaluation of the computed distributional statistics (mean, variance, skewness, and kurtosis) across full and sampled datasets revealed distinct patterns associated with each sampling method. Stratified sampling exhibited the highest fidelity to the full dataset across all metrics, particularly preserving variance and kurtosis values. This alignment is expected, given its design to ensure proportional representation of tumor subtypes—critical in personalized medicine applications where subgroup heterogeneity affects gene expression signatures (Chen & Ishwaran, 2012).

In contrast, despite its statistical neutrality, simple random sampling (SRS) displayed greater fluctuations in higher-order moments such as skewness and kurtosis, especially for genes with non-normal or tail-heavy distributions. These discrepancies may obscure rare expression phenotypes or exaggerate noise, undermining model robustness (Lohr, 2010). Systematic sampling demonstrated intermediate performance: though efficient, it showed minor sensitivity to hidden periodicity in data ordering, which, if uncorrected, can introduce subtle biases (Ma & Dai, 2011).

Histograms confirmed these results. The stratified sample histograms most closely resembled those from the full dataset, capturing key modes and tail behavior. SRS histograms appeared more variable in tail thickness and symmetry. Similarly, QQ plots for stratified and systematic samples tracked the theoretical quantile line more consistently than the SRS samples, whose deviations in the tails reflect sampling instability under non-Gaussian expression patterns (DeCarlo, 1997).

These findings are critical in high-dimensional bioinformatics, where accurate modeling of gene expression distributions influences downstream biomarker discovery and therapeutic decision-making (Catchpoole et al., 2010). Ensuring that sampled data reflect central and peripheral distributional characteristics can guard against misclassification and overfitting in predictive models.

Implications for Data Modeling

Sampling offers substantial computational relief in high-dimensional settings, enabling faster model iteration, especially when dimensionality (p) vastly exceeds sample size (n). Reducing the dataset from 801 to 200 observations significantly lowers the computational burden

for training models like SVMs or ensemble classifiers, which scale non-linearly with input dimensions (Jiang, 2025).

Beyond efficiency, sampling improves model generalizability when appropriately designed. Stratified subsets maintain the population structure, supporting learning algorithms in capturing heterogeneity and minimizing overfitting to dominant classes or expression levels. As observed, this method aligns well with personalized medicine's emphasis on subgroup fidelity and rare-variant detection (West et al., 2010).

Each sampling technique engages with the bias-variance trade-off differently. SRS may underrepresent informative subgroups, introducing variance and model instability. Stratified sampling reduces variance but can slightly increase bias if strata boundaries are coarsely defined. Systematic sampling offers low implementation bias but may falter if implicit periodicity exists in data acquisition.

Nonetheless, limitations persist. Sampling risks omitting low-frequency yet clinically salient signals—e.g., outlier expression patterns tied to drug resistance or mutation burden—if not carefully stratified or combined with feature selection (Loyola et al., 2016). Sampling-based modeling must therefore be supplemented with domain-informed feature engineering or dimensionality reduction.

Practical Applications and Mitigation

Sampling distributions support precision analytics in domains where full dataset access is constrained by cost, computation, or privacy. In healthcare, representative sampling can facilitate early-phase drug response modeling or real-time cohort segmentation, especially where entire RNA-Seq matrices cannot be shared due to regulatory barriers (Berisha et al., 2021). In finance, sampling enables rapid risk scoring in high-frequency datasets without querying every

transaction (Chen & Ishwaran, 2012). In computer vision, frames or segments can be sampled adaptively from surveillance streams or MRI scans to reduce labeling burdens.

Importantly, sampling complements nonlinear dimensionality reduction methods such as t-SNE or UMAP, both of which benefit from noise-reduced inputs. For instance, t-SNE clustering post-sampling has shown promise in identifying diagnostic subgroups in genomic and clinical data (Babu et al., 2025). These hybrid strategies enhance interpretability while controlling computational complexity.

Ethically, sampling in biomedical contexts must ensure population representativeness. Oversampling common phenotypes or undersampling vulnerable subgroups may skew algorithmic predictions, reinforcing disparities in care delivery (West et al., 2010). Transparent stratification design and routine bias auditing are essential to uphold fairness in precision modeling.

Conclusion

This study examined the implementation and comparative utility of sampling distributions—specifically, simple random, stratified, and systematic sampling—in analyzing high-dimensional RNA-Seq data from cancer patients. Stratified sampling emerged as the most consistent method in preserving distributional features across a range of statistical moments, validated through both numerical metrics and graphical analysis.

Sampling distributions serve as essential tools in mitigating the curse of dimensionality. By providing manageable, representative subsets, they enhance downstream modeling processes' stability, interpretability, and efficiency while maintaining fidelity to population-level structure. These benefits are particularly critical in personalized medicine, where patient-specific variation must be modeled with precision.

Future research should extend this analysis to adaptive and active learning-based sampling, explore interactions with advanced machine learning classifiers, and assess how sampling affects longitudinal or multi-omics datasets. Automated sampling workflows integrating stratification logic, feature selection, and model uncertainty may offer new frontiers for scalable, ethical, and individualized biomedical modeling.

References

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *Database Theory—ICDT 2001* (pp. 420–434). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-44503-X_27

Barbour, D. L. (2019). Precision medicine and the cursed dimensions. *npj Digital Medicine, 2*(4). https://doi.org/10.1038/s41746-019-0081-5

Babu, B. D., Venkateswarlu, B., Tamilselvan, S., Balu, P. B., Balasundaram, N., & Praveenkumar, R. (2025). Outlier detection in high-dimensional data using t-distributed stochastic neighbor embedding. *2025 3rd International Conference on Advancement in Computation & Computer Technologies (InCACCT)*. IEEE. https://doi.org/10.1109/InCACCT65424.2025.11011342

Berisha, V., Krantsevich, C., Hahn, P. R., Hahn, S., Dasarathy, G., Turaga, P., & Liss, J. (2021). Digital medicine and the curse of dimensionality. *npj Digital Medicine, 4*, 153. https://doi.org/10.1038/s41746-021-00521-5

Blee, A. M., Li, B., Pecen, T., Chovanec, M., Wang, J., Stumpf, M., ... & Wyatt, A. W. (2022). An active learning framework improves tumor variant interpretation. *Cancer Research, 82*(15), 2704–2715. https://doi.org/10.1158/0008-5472.CAN-21-3798

Catchpoole, D. R., Kennedy, P., Skillicorn, D. B., & Simoff, S. (2010). The curse of dimensionality: A blessing to personalized medicine. *Journal of Clinical Oncology, 28*(34), e723–e724. https://doi.org/10.1200/JCO.2010.30.1986

Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics, 99*(6), 323–329. https://doi.org/10.1016/j.ygeno.2012.04.003

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods, 2*(3), 292–307. https://doi.org/10.1037/1082-989X.2.3.292

Debie, J., & Shafi, A. (2019). Implications for the curse of dimensionality for supervised learning classifier systems: Theoretical and empirical analyses. *Pattern Analysis and Applications, 22*, 519–536. https://doi.org/10.1007/s10044-017-0649-0

Han, B., Kang, H. M., Eskin, E., & Schnell, A. (2014). Fast pairwise IBD association testing in genome-wide association studies. *Bioinformatics, 30*(2), 206–213. https://doi.org/10.1093/bioinformatics/btt668

Hickernell, F. J., & Owen, A. B. (2018). Monte Carlo and quasi–Monte Carlo sampling techniques. *Wiley Interdisciplinary Reviews: Computational Statistics, 10*(6), e1637. https://doi.org/10.1002/wics.1637

Hubbard, A., Trostle, J., Cangemi, I., & Eisenberg, J. N. S. (2019). Countering the curse of dimensionality: Exploring data-generating mechanisms through participant observation and mechanistic modeling. *Epidemiology*, *30*(4), 609–614. https://doi.org/10.1097/EDE.0000000000001025

Jiang, Y. (2025). Research on the application of support vector machines in high-dimensional data mining. In *2025 8th International Conference on Advanced Algorithms and Control Engineering (ICAACE)*. https://doi.org/10.1109/ICAACE65325.2025.11020080

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Kadi, H., Rebbah, M., Meftah, B., & Lézoray, O. (2021). Medical decision-making based on the exploration of a personalized medicine dataset. *Informatics in Medicine Unlocked, 23*, 100561. https://doi.org/10.1016/j.imu.2021.100561

Katamesh, N. S., Abbas, A. E. F., & Mahmoud, S. A. (2024). Four chemometric models enhanced by Latin hypercube sampling design for quantification of anti-COVID drugs. *BMC Chemistry, 18*(1), 54. https://doi.org/10.1186/s13065-024-01158-7

Laa, U., Cook, D., & Lee, S. (2021). Burning Sage: Reversing the Curse of Dimensionality in the Visualization of High-Dimensional Data. *Journal of Computational and Graphical Statistics*, *31*(1), 40–49. https://doi.org/10.1080/10618600.2021.1963264

Livne, D., & Efroni, S. (2024). Pathway metrics accurately stratify T cells to their cell states. *BioData Mining, 17*(1), Article 60. https://doi.org/10.1186/s13040-024-00416-7

Lohr, S. L. (2010). *Sampling: Design and analysis* (2nd ed.). Brooks/Cole.

Loyola, D. G., Pedergnana, M., & García, S. G. (2016). Smart sampling and incremental function learning for very large high-dimensional data. *Neural Networks, 78*, 75–87. https://doi.org/10.1016/j.neunet.2015.09.001

Ma, S., & Dai, Y. (2011). Principal component analysis based methods in bioinformatics studies. *Briefings in Bioinformatics, 12*(6), 714–722. https://doi.org/10.1093/bib/bbq090

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics, 38*(8), 904–909. https://doi.org/10.1038/ng1847

Privé, F., Luu, K., Blum, M. G. B., McGrath, J. J., & Vilhjálmsson, B. J. (2020). Efficient toolkit implementing best practices for principal component analysis of population genetic data. *Bioinformatics, 36*(16), 4449–4457. https://doi.org/10.1093/bioinformatics/btaa520

Ringnér, M. (2008). What is principal component analysis? *Nature Biotechnology, 26*(3), 303–304. https://doi.org/10.1038/nbt0308-303

UCI Machine Learning Repository. (2016). *Gene expression cancer RNA-Seq data set*. University of California, Irvine. https://archive.ics.uci.edu/dataset/401/gene+expression+cancer+rna+seq

West, M., Ginsburg, G. S., Huang, A. T., & Thomas, J. J. (2010). The curse of dimensionality: A blessing to personalized medicine. *Journal of Clinical Oncology, 28*(34), e723–e724. https://doi.org/10.1200/JCO.2010.30.1986

Yang, C., Fridgeirsson, E. A., Kors, J. A., Reps, J. M., & Rijnbeek, P. R. (2024). Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. *Journal of Big Data, 11*(1), 7. https://doi.org/10.1186/s40537-023-00857-7