Module_3: *(Template)*

Team Members:

Kevin Jiang and Ana

Project Title:

Analyzing Breast Cancer's hallmark trait of Limitless Replicative Potential

Project Goal:

This project seeks to determine whether breast cancer (BRCA) tumors with higher expression of telomerase-related genes (TERT, TERC, MYC) show distinct molecular or clinical patterns compared with telomerase-low tumors. We will use Principal Component Analysis (PCA) to reduce the gene-expression space and examine the PC loadings to see which telomerase-related genes drive the main variation. We will then compare the PCA-based groupings with available clinical categories (e.g., tumor subtype, cancer stage, patient age) to see whether telomerase-high samples align with specific clinical groups.

Disease Background:

- Cancer hallmark focus: Limitless replicative potential.
- Overview of hallmark: Normal cells can only divide a limited number of times before entering senescence due to telomere shortening. Cancer cells overcome this limit by reactivating telomerase, an enzyme that extends telomeres and prevents chromosome degradation. This allows them to divide indefinitely and maintain chromosomal stability. In some tumors, telomere length is preserved through an alternative lengthening mechanism, ensuring continuous growth and immortality.
- Genes associated with hallmark to be studied (describe the role of each gene, signaling pathway, or gene set you are going to investigate):
  - TERT: Catalytic part of telomerase that extends telomeres, preventing cell aging and allowing unlimited division.
  - TERC: RNA template for telomere synthesis; works with TERT to maintain chromosome ends.
  - MYC: Oncogene that activates TERT expression and promotes continuous cell growth.
  - TP53: Tumor suppressor that triggers senescence when telomeres shorten; mutations let cells keep dividing.
  - RB1: Controls the G1–S checkpoint; loss of function removes cell-cycle limits and promotes proliferation.
  - ATRX/DAXX: Regulate the alternative lengthening of telomeres (ALT) pathway in cancers without telomerase activity.
- Prevalence & incidence:
  - The rate of new cases of female breast cancer was 130.8 per 100,000 women per year.
  - 4 million women are living with a history of breast cancer, including those currently undergoing treatment and survivors.
- Risk factors (genetic, lifestyle) & Societal determinants:
  - Genetic:
    - Inherited mutations in BRCA1, BRCA2, and TP53 greatly increase breast cancer risk.
    - Family history of breast or ovarian cancer raises susceptibility.
    - Certain gene variations affecting hormone regulation or telomerase activity (TERT) can also contribute to risk.
  - Lifestyle Factors:
    - Prolonged estrogen exposure (early menstruation, late menopause, or hormone therapy).
    - Obesity, especially after menopause.
    - Alcohol consumption and physical inactivity.
    - Late or no pregnancies and limited breastfeeding, which increase hormonal exposure.
    - Radiation exposure to the chest at a young age.
  - Societal Determinants:
    - Limited access to regular screenings and medical services can delay diagnosis.
    - Lower income and lack of insurance can make it harder to afford treatment.
    - Less knowledge about symptoms and prevention can reduce early detection.
    - People in rural areas may have fewer healthcare facilities or specialists available.
- Standard of care treatments (& reimbursement):
  - Lumpectomy or mastectomy to remove the tumor; reconstruction may follow.
  - Radiation therapy: Used after surgery to destroy remaining cancer cells.
  - Hormone therapy: Tamoxifen or aromatase inhibitors for ER+/PR+ cancers.
  - Targeted therapy: HER2-positive cancers treated with trastuzumab, pertuzumab, or T-DM1.
  - Chemotherapy: Used for aggressive or triple-negative cancers.
  - Immunotherapy: Pembrolizumab for PD-L1–positive triple-negative breast cancer.
  - PARP inhibitors: For patients with BRCA1 or BRCA2 mutations.
  - Reimbursement:
    - Most standard treatments are covered by Medicare, Medicaid, and private insurance
    - Financial assistance programs and nonprofit organizations can help patients afford costly treatments and medications.
- Biological mechanisms (anatomy, organ physiology, cell & molecular physiology):
  - Anatomy & Organ Physiology:
    - Breast cancer begins in the ducts or lobules of the mammary glands, which are hormonally regulated by estrogen and progesterone. These hormones influence cell growth and increase cancer risk when unbalanced.
  - Cell Physiology:
    - Normal breast cells can only divide a limited number of times before senescence. Cancer cells overcome this by activating telomerase, allowing continuous cell division.
  - Molecular Physiology:
    - Telomeres shorten with each division, normally triggering p53 and RB1 to stop growth. In cancer, TERT reactivation maintains telomeres and prevents cell death, enabling unlimited replication and tumor progression.

citations:

- Hallmarks of cancer
- Cancer of the Breast (Female) - Cancer Stat Facts. (n.d.-b). SEER. https://seer.cancer.gov/statfacts/html/breast.html
- Breast Cancer Treatment (PDQ®). (2025, April 25). Cancer.gov. https://www.cancer.gov/types/breast/hp/breast-treatment-pdq

Data-Set:

The data was downloaded from TCGA Level 3 data via the Synapse portal for 12 cancer types (https://www.synapse.org/#!Synapse:syn1695324). This included 3468 samples that had been preprocessed using TCGA's standard pipeline. To reprocess TCGA data with Rsubread, they downloaded FASTQ formatted files for all available TCGA tumor samples via the National Cancer Institute's Cancer Genomics Hub (Wilks et al., 2014). This included a total of 9264 tumor samples across 24 cancer types (Table 1). Some patient samples were sequenced multiple times; in these cases, they included each replicate.

The data was collected using RNA-seq, which helps sequence the genome of each cancer type so that they can be used to compare to one another. My team will be using this RNA-seq data to try and analyze the hallmark trait of Limitless replicative potential to see why this happens and using machine learning, we will try to use the data from RNA-seq to see if any patterns can be seen or not.

Data Analyis:

Methods

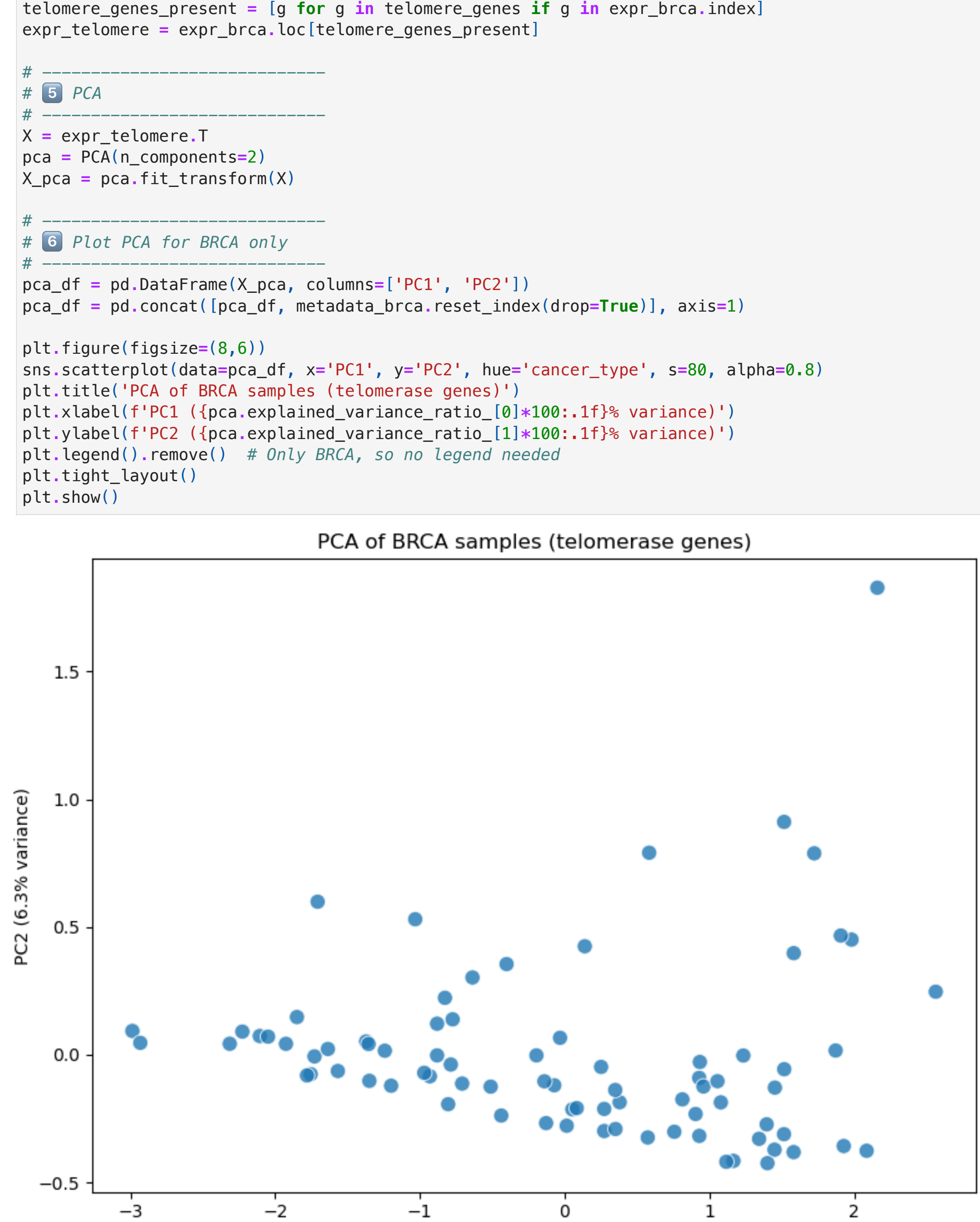
The machine learning technique used in this project is Principal Component Analysis (PCA). PCA is an unsupervised learning method that reduces complex gene expression data into a few main components, or "principal axes," that capture the most important variation across samples. In this project, PCA helps identify major patterns in the expression of telomerase-related genes (TERT, TERC, MYC) among breast cancer samples. We will examine the PC loadings to identify which genes contribute most to each component and determine whether telomerase-related genes are driving these differences. We will also compare the resulting PCA groupings with clinical variables (e.g., tumor subtype or cancer stage) to see if similar clusters appear. By plotting the samples along these components, we can see whether tumors with high telomerase activity form distinct molecular or clinical groups. PCA optimizes the amount of variance explained by each component, meaning it tries to summarize as much of the original data's information as possible using fewer dimensions. The model is considered "good enough" when the first few components capture most of the total variance, showing that the main biological signal has been successfully extracted.

Analysis

(Describe how you analyzed the data. This is where you should intersperse your Python code so that anyone reading this can run your code to perform the analysis that you did, generate your figures, etc.)

In [3]:

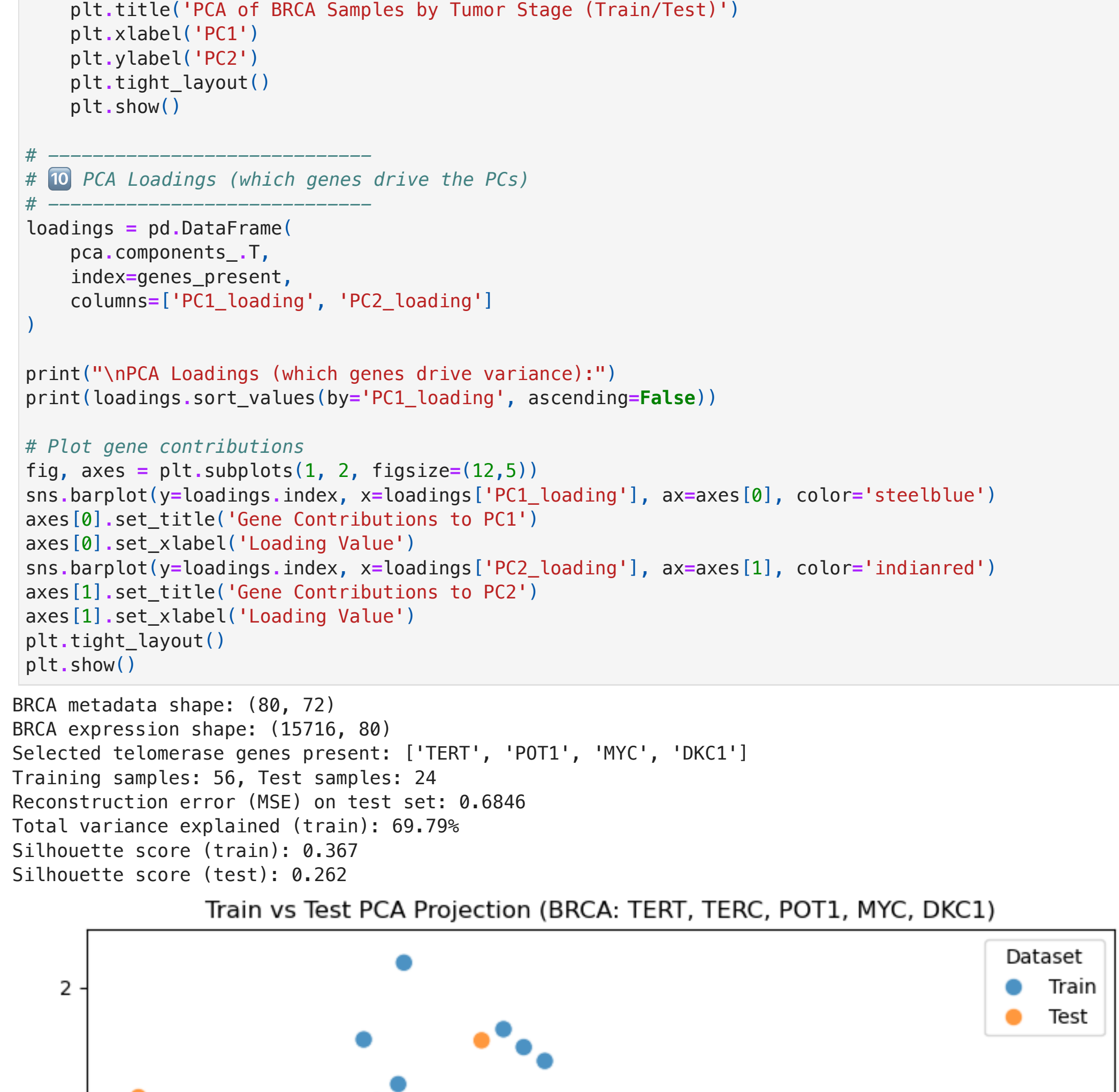
```python
import pandas as pd
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt
import seaborn as sns

# ------------------------------
# 1️⃣ Load metadata and expression
# ------------------------------
metadata_file = "GSE62944_metadata.csv"
expr_file = "GSE62944_subsample_log2TPM.csv"

metadata = pd.read_csv(metadata_file)
expr = pd.read_csv(expr_file, index_col=0)

# ------------------------------
# 2️⃣ Filter TCGA samples
# ------------------------------
tcga_samples = metadata[metadata['sample'].str.startswith('TCGA')]['sample'].tolist()
metadata_tcga = metadata[metadata['sample'].isin(tcga_samples)]
expr_tcga = expr[tcga_samples]

# ------------------------------
# 3️⃣ Filter only BRCA samples
# ------------------------------
metadata_brca = metadata_tcga[metadata_tcga['cancer_type'] == 'BRCA']
expr_brca = expr_tcga[metadata_brca['sample']]

# ------------------------------
# 4️⃣ Select telomerase-related genes
# ------------------------------
telomere_genes = ['TERT', 'TERC', 'MYC']
telomere_genes_present = [g for g in telomere_genes if g in expr_brca.index]
expr_telomere = expr_brca.loc[telomere_genes_present]

# ------------------------------
# 5️⃣ PCA
# ------------------------------
X = expr_telomere.T
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X)

# ------------------------------
# 6️⃣ Plot PCA for BRCA only
# ------------------------------
pca_df = pd.DataFrame(X_pca, columns=['PC1', 'PC2'])
pca_df = pd.concat([pca_df, metadata_brca.reset_index(drop=True)], axis=1)

plt.figure(figsize=(8,6))
sns.scatterplot(data=pca_df, x='PC1', y='PC2', hue='cancer_type', s=80, alpha=0.8)
plt.title('PCA of BRCA samples (telomerase genes)')
plt.xlabel(f'PC1 ({pca.explained_variance_ratio_[0]*100:.1f}% variance)')
plt.ylabel(f'PC2 ({pca.explained_variance_ratio_[1]*100:.1f}% variance)')
plt.legend().remove()  # Only BRCA, so no legend needed
plt.tight_layout()
plt.show()
```



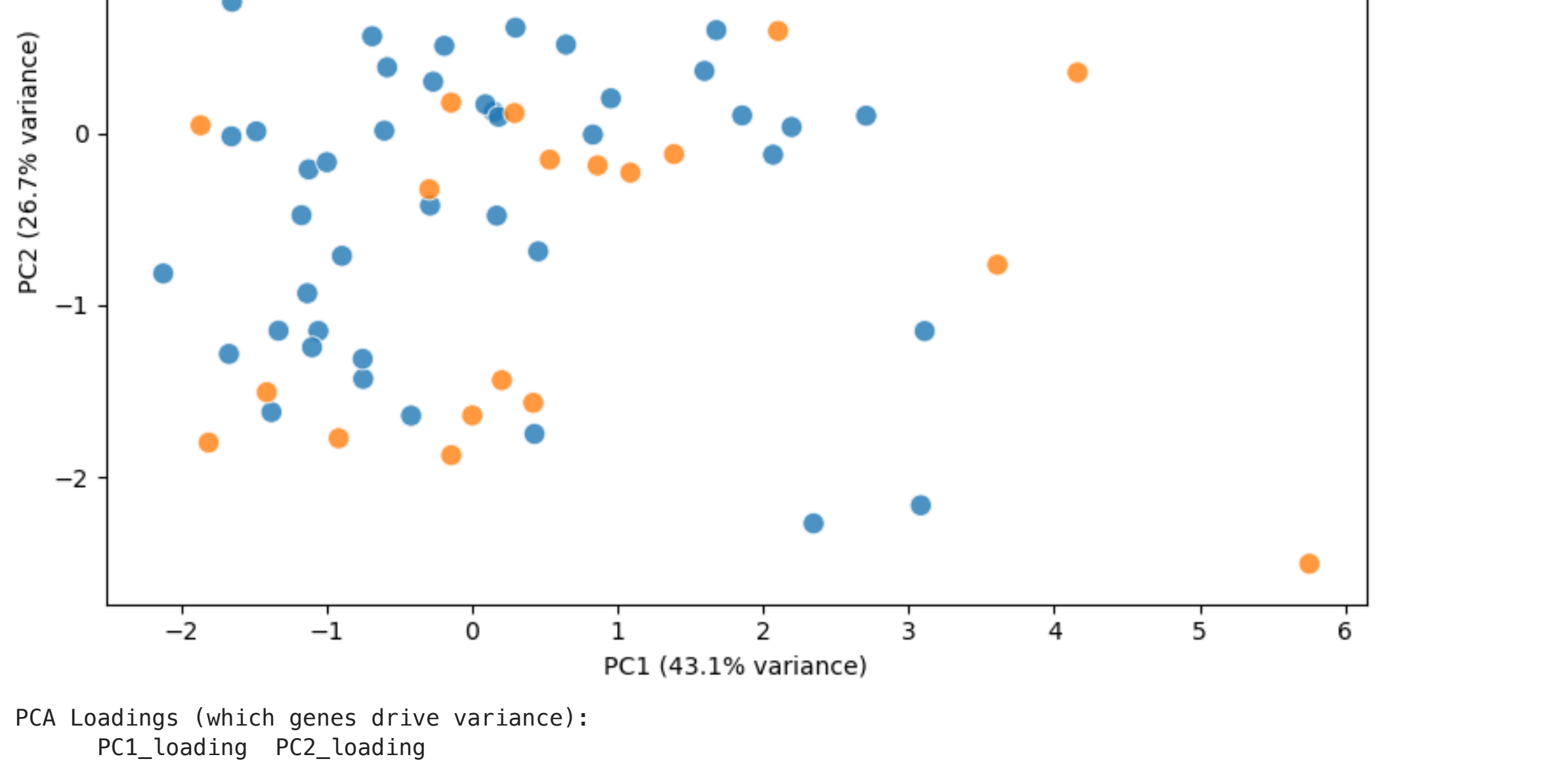
In [1]:

```python
# ---------------------------------------------------------
# Telomerase-Related Gene Analysis in BRCA using PCA
# ---------------------------------------------------------
import pandas as pd
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

# ------------------------------
# 1️⃣ Load metadata and expression data
# ------------------------------
metadata_file = "GSE62944_metadata.csv"
expr_file = "GSE62944_subsample_log2TPM.csv"

metadata = pd.read_csv(metadata_file)
expr = pd.read_csv(expr_file, index_col=0)

# ------------------------------
# 2️⃣ Filter for TCGA-BRCA samples
# ------------------------------
brca_samples = metadata[
    (metadata['sample'].str.startswith('TCGA')) &
    (metadata['cancer_type'] == 'BRCA')
]['sample'].tolist()

metadata_brca = metadata[metadata['sample'].isin(brca_samples)]
expr_brca = expr[brca_samples]

print("BRCA metadata shape:", metadata_brca.shape)
print("BRCA expression shape:", expr_brca.shape)

# ------------------------------
# 3️⃣ Select telomerase-related genes
# ------------------------------
telomere_genes = ['TERT', 'TERC', 'POT1', 'MYC', 'DKC1']
genes_present = [g for g in telomere_genes if g in expr_brca.index]
expr_telomere = expr_brca.loc[genes_present]

print("Selected telomerase genes present:", genes_present)

# ------------------------------
# 4️⃣ Train/Test Split
# ------------------------------
train_samples, test_samples = train_test_split(brca_samples, test_size=0.3, random_state=42)

expr_train = expr_telomere[train_samples]
expr_test = expr_telomere[test_samples]
meta_train = metadata_brca[metadata_brca['sample'].isin(train_samples)]
meta_test = metadata_brca[metadata_brca['sample'].isin(test_samples)]

print(f"Training samples: {len(train_samples)}, Test samples: {len(test_samples)}")

# ------------------------------
# 5️⃣ PCA on training set
# ------------------------------
X_train = expr_train.T
scaler = StandardScaler().fit(X_train)
X_train_scaled = scaler.transform(X_train)

pca = PCA(n_components=2)
X_train_pca = pca.fit_transform(X_train_scaled)

# ------------------------------
# 6️⃣ Apply PCA to test set
# ------------------------------
X_test = expr_test.T
X_test_scaled = scaler.transform(X_test)
X_test_pca = pca.transform(X_test_scaled)

# ------------------------------
# 7️⃣ Combine PCA results
# ------------------------------
train_df = pd.DataFrame(X_train_pca, columns=['PC1', 'PC2'])
train_df['dataset'] = 'Train'
train_df = pd.concat([train_df, meta_train.reset_index(drop=True)], axis=1)

test_df = pd.DataFrame(X_test_pca, columns=['PC1', 'PC2'])
test_df['dataset'] = 'Test'
test_df = pd.concat([test_df, meta_test.reset_index(drop=True)], axis=1)

pca_df = pd.concat([train_df, test_df])

# ------------------------------
# 8️⃣ Evaluate Model Performance
# ------------------------------
# Reconstruction error (variance preservation)
X_test_reconstructed = pca.inverse_transform(X_test_pca)
mse = np.mean((X_test_scaled - X_test_reconstructed)**2)
print(f"Reconstruction error (MSE) on test set: {mse:.4f}")
print(f"Total variance explained (train): {sum(pca.explained_variance_ratio_)*100:.2f}%")

# Clustering stability metric (silhouette score)
kmeans = KMeans(n_clusters=2, random_state=42).fit(X_train_pca)
train_labels = kmeans.labels_
test_labels = kmeans.predict(X_test_pca)

train_sil = silhouette_score(X_train_pca, train_labels)
test_sil = silhouette_score(X_test_pca, test_labels)
print(f"Silhouette score (train): {train_sil:.3f}")
print(f"Silhouette score (test): {test_sil:.3f}")

# ------------------------------
# 9️⃣ PCA Visualizations
# ------------------------------
plt.figure(figsize=(8,6))
sns.scatterplot(data=pca_df, x='PC1', y='PC2', hue='dataset', s=80, alpha=0.8)
plt.title('Train vs Test PCA Projection (BRCA: TERT, TERC, POT1, MYC, DKC1)')
plt.xlabel(f'PC1 ({pca.explained_variance_ratio_[0]*100:.1f}% variance)')
plt.ylabel(f'PC2 ({pca.explained_variance_ratio_[1]*100:.1f}% variance)')
plt.legend(title='Dataset')
plt.tight_layout()
plt.show()

# Optional: color by tumor stage if available
if 'tumor_stage' in metadata_brca.columns:
    plt.figure(figsize=(8,6))
    sns.scatterplot(data=pca_df, x='PC1', y='PC2', hue='tumor_stage', style='dataset', s=80, alpha=0.8)
    plt.title('PCA of BRCA Samples by Tumor Stage (Train/Test)')
    plt.xlabel('PC1')
    plt.ylabel('PC2')
    plt.tight_layout()
    plt.show()

# ------------------------------
# 🔟 PCA Loadings (which genes drive the PCs)
# ------------------------------
loadings = pd.DataFrame(
    pca.components_.T,
    index=genes_present,
    columns=['PC1_loading', 'PC2_loading']
)

print("\nPCA Loadings (which genes drive variance):")
print(loadings.sort_values(by='PC1_loading', ascending=False))

# Plot gene contributions
fig, axes = plt.subplots(1, 2, figsize=(12,5))
sns.barplot(y=loadings.index, x=loadings['PC1_loading'], ax=axes[0], color='steelblue')
axes[0].set_title('Gene Contributions to PC1')
axes[0].set_xlabel('Loading Value')
sns.barplot(y=loadings.index, x=loadings['PC2_loading'], ax=axes[1], color='indianred')
axes[1].set_title('Gene Contributions to PC2')
axes[1].set_xlabel('Loading Value')
plt.tight_layout()
plt.show()
```

```
BRCA metadata shape: (80, 72)
BRCA expression shape: (15716, 80)
Selected telomerase genes present: ['TERT', 'POT1', 'MYC', 'DKC1']
Training samples: 56, Test samples: 24
Reconstruction error (MSE) on test set: 0.6846
Total variance explained (train): 69.79%
Silhouette score (train): 0.367
Silhouette score (test): 0.262
```




```
PCA Loadings (which genes drive variance):
      PC1_loading  PC2_loading
MYC      0.610029    -0.014328
DKC1     0.571239     0.195638
TERT     0.536889    -0.385545
POT1     0.115329     0.901596
```



Verify and validate your analysis:

(Describe how you checked to see that your analysis gave you an answer that you believe (verify). Describe how your determined if your analysis gave you an answer that is supported by other evidence (e.g., a published paper).

We evaluate the PCA model using reconstruction error (mean squared error, MSE). After reducing the BRCA telomerase gene expression data to 2 PCs, we measure how well the original gene expression is reconstructed from the PCs. The MSE is calculated as the average squared difference between the original and reconstructed expression values. Lower MSE indicates that the principal components capture most of the variance in the data. Additionally, we assess clustering structure in the reduced space using the silhouette score, which quantifies how well samples cluster together relative to other clusters. Positive silhouette values indicate coherent clusters.

Results: - Reconstruction error (MSE) on test set: 0.6846 - Total variance explained (train): 69.79% - Silhouette score (train): 0.367 - Silhouette score (test): 0.262

These metrics together show that the PCA captures most of the variance in the data while revealing moderate clustering structure among BRCA samples.

Conclusions and Ethical Implications:

(Think about the answer your analysis generated, draw conclusions related to your overarching question, and discuss the ethical implications of your conclusions.

Limitations and Future Work:

(Think about the answer your analysis generated, draw conclusions related to your overarching question, and discuss the ethical implications of your conclusions.

NOTES FROM YOUR TEAM:

This is where our team is taking notes and recording activity.

QUESTIONS FOR YOUR TA:

These are questions we have for our TA.

In [ ]: