Module 3

Team Members:

Max Calcoen, Will Rousseau

Project Title:

What hallmarks and pathways are overexpressed in Prostate Adenocarcinoma?

Project Goal:

Can we accurately predict cancer type based on gene expression? Which cancer types are most conflated with random forest analysis, and why? Are immunoevasive cancer types more or less conflated than others? Can we find specific significant genes for certain cancer types?

Disease Background:

Cancer Hallmark Focus: **Evading Immune Destruction**

Overview of hallmark:

Cancer cells hijack the body's immune checkpoint systems (the "off switches" that prevent the immune system from attacking our own cells). This allows tumors to escape detection and destruction through three strategies: camouflage (hiding from detection), coercion (suppressing immune responses), or cytoprotection (protecting against immune attack).

Key mechanisms: activating checkpoint molecules (PD-1/PD-L1, CTLA-4) that shut down T cells, hiding tumor markers, releasing immunosuppressive signals (TGF-β, IL-10), and recruiting cells that help the tumor.

Genes associated with hallmark to be studied:

1. CD274 (PD-L1 gene)

- **Role:** Makes PD-L1 protein that acts like a "don't attack me" signal on tumor cells, shutting off immune responses when it connects with PD-1 on T cells
- **Pathway:** Activated by inflammatory signals and cancer-promoting pathways (PI3K/AKT, JAK-STAT)
- **Significance:** Most commonly exploited immune escape mechanism; target of many FDA-approved immunotherapies

2. PDCD1 (PD-1 gene)

- **Role:** Makes PD-1, a brake pedal for T cells. High levels indicate "exhausted" T cells that can no longer fight effectively
- **Pathway:** High PD-1 levels are a sign of exhausted T cells that have been overstimulated
- **Significance:** Blocking PD-1 can reactivate exhausted T cells and restore their cancer-fighting ability

3. CTLA4 gene

- **Role:** Makes CTLA-4 protein that shuts down T cells during early immune response in lymph nodes
- **Pathway:** Blocks CD80/CD86 molecules that would normally activate T cells, preventing them from multiplying and attacking
- **Significance:** First immune checkpoint successfully targeted in cancer therapy (drug: ipilimumab)

4. HLA genes (MHC Class I: HLA-A, HLA-B, HLA-C) and B2M

- **Role:** Create molecules that display cancer markers on cell surfaces (like ID badges for immune cells). When tumors lose these, they become invisible to T cells
- **Mechanisms of downregulation:** Loss of transporter proteins (TAP1/TAP2), B2M mutations, reduced transcription factors, or DNA methylation
- **Significance:** Very common in cancers; helps tumors resist immunotherapy

5. TGFB1 (TGF-β gene)

- **Role:** Produces powerful immunosuppressive signal that disables killer T cells and NK cells
- **Effects:** Creates immunologically inactive tumor environments where immune cells are absent or unable to function
- **Significance:** Major contributor to immunotherapy resistance

6. IDO1 (Indoleamine 2,3-dioxygenase 1 gene)

- **Role:** Makes enzyme that starves T cells by depleting tryptophan and creating toxic metabolites
- **Pathway:** Tryptophan breakdown pathway
- **Significance:** Being tested as drug target in combination immunotherapies

7. IL10 gene

- **Role:** Creates IL-10, an anti-inflammatory signal that prevents dendritic cells from activating T cells
- **Significance:** Helps create an immunosuppressive tumor environment

8. LAG3 (Lymphocyte-activation gene 3)

- **Role:** Another immune checkpoint that shuts down T cell function by interfering with how they recognize targets
- **Significance:** Early marker of T cell exhaustion; FDA-approved drugs target this (2022)

Sources:

- https://molecular-cancer.biomedcentral.com/articles/10.1186/s12943-024-02023-w
- https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2021.636568/full
- https://www.nature.com/articles/s41392-025-02280-1

Prevalence & Incidence

- Over 2 million new US cancer cases and 618,000 deaths expected in 2025
- **Most common:** Men (prostate, lung, colorectal); Women (breast, lung, colorectal)
- **Deadliest:** Lung cancer (nearly 1 in 5 cancer deaths worldwide)
- **Survival varies:** >95% for thyroid/prostate; 13% for pancreatic; 22% for liver/esophageal
- **Trends:** Six of top 10 cancers increasing; early-onset cancers (under 50) jumped 79% globally from 1990-2019

Sources:

- https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21871
- https://www.cancer.gov/about-cancer/understanding/statistics

---

Risk Factors (Genetic, Lifestyle) & Societal Determinants

Genetic:

- Only 5-10% of cancers from inherited defects; 90-95% from environment/lifestyle
- Inherited mutations (BRCA1/BRCA2) increase predisposition, making additional mutations accumulate easier
- Cancer requires multiple gene mutations working together

Lifestyle:

- Tobacco: 25-30% of cancer deaths; diet: 30-35%; infections: 15-20%
- Key factors: smoking, processed/red meat, alcohol, obesity, inactivity, sun exposure, toxins
- Healthy lifestyle reduces risk even with genetic predisposition

Societal Determinants:

- Native Americans face 2-3× higher death rates for kidney, liver, stomach, cervical cancers; Black Americans have 2× higher mortality for prostate, stomach, uterine cancers
- Neighborhood poverty shows biological changes in tumors (different gene expression)
- States without Medicaid expansion have highest cancer mortality
- Access barriers: healthcare, insurance, transportation, healthy food

Sources:

- https://pmc.ncbi.nlm.nih.gov/articles/PMC2515569/
- https://www.statnews.com/2024/08/08/cancer-genetics-neighborhood-social-determinants/

Standard of Care Treatments (& Reimbursement)

Traditional Approaches:

- Surgery: Robotic/fluorescence-guided technology
- Radiation: 3D-targeted precision
- Chemotherapy: Improved tumor-targeting
- Hormonal therapy: For hormone-dependent cancers (breast, prostate)

Immunotherapy:

- 150+ FDA approvals since 2011; 17 in 2024 alone (20-fold increase in use)
- Checkpoint inhibitors (81% of approvals) unleash cancer-fighting T cells
- 2024 breakthroughs: First TIL therapy, TCR-engineered therapy, IL-15 treatment
- Also: CAR-T cell therapy, bispecific antibodies, antibody-drug conjugates

Targeted Therapies:

- Tumor genetic testing guides treatment (especially lung cancers with EGFR, ALK, ROS1, KRAS mutations)
- Treatment tailored to tumor's molecular profile

Reimbursement:

- Annual costs: $6,000 to$100,000+ depending on type/stage
- Out-of-pocket limits (2022): $8,700 individual$,17,400 family (most patients hit within 1-3 months)
- Medicare: Pays 80% after deductible; patients pay 20%
- ACA mandates coverage for cancer diagnosis and treatment

Sources:

- https://www.cancerresearch.org/blog/introducing-the-cri-cancer-immunotherapy-insights-impact-report
- https://www.facingourrisk.org/support/insurance-paying-for-care/treatment/overview

---

Biological Mechanisms (Anatomy, Organ Physiology, Cell & Molecular Physiology)

Core Principle:

- Cancer is damaged DNA that changes how cells grow
- Starts with a single cell acquiring multiple mutations over time, transforming into a cancer cell capable of unlimited growth

Two Main Gene Classes:

Oncogenes (accelerators):

- Normally help cells grow; when mutated get stuck "on"
- One mutated copy can cause problems (dominant effect)
- Examples: RAS, MYC, EGFR, HER2

Tumor Suppressors (brakes):

- Normally stop damaged cells from dividing; when lost, cells grow unchecked
- Both copies usually need inactivation (recessive effect)
- TP53 is the "guardian of the genome"
- Examples: TP53, Rb, BRCA1/BRCA2

Molecular Mechanisms:

- **DNA mutations:** Single nucleotide substitutions (KRAS mutations in 5% of cancers), chromosomal rearrangements (BCR-ABL fusion in leukemia)
- **Epigenetic changes:** Activate/deactivate genes without changing DNA sequence
- **Gene expression alterations:** Genes acting differently without DNA damage (more common than mutations)

Multi-Step Process:

- Cancer requires multiple mutations accumulated over time
- Explains age-related cancer risk (more time = more mutations)
- Cancer cells compete; those with survival advantages take over

Common Cancer Hallmarks (Shared Across Types):

- Sustained proliferation, evading growth suppressors, resisting cell death
- Enabling immortality, inducing angiogenesis, activating invasion/metastasis
- Immune evasion, reprogramming metabolism

Cancer Type Differences:

- Same mutations cause different cancers depending on: cell type, tissue context, timing, microenvironment, cooperating mutations
- Some tumors have <5% of cancer genes mutated, suggesting alternative paths

Sources:

- https://www.cancer.gov/about-cancer/causes-prevention/genetics
- https://www.nature.com/articles/s41598-023-39608-2

AI Use: Info and sources were gathered by Will, Claude was used to format this section

Data-Set:

The Cancer Genome Atlas (TCGA) RNA-seq dataset, re-processed by Rahman et al., includes 24 cancer types, with abbreviations detailed in the original paper. The data were subsetted to include the top 3,000 most variable genes out of approximately 15,000 protein-coding genes. The subset comprises approximately 50–100 tumor samples per cancer type, totaling 1,802 out of 9,264 tumor samples. Metadata from the original dataset were retained, including around 70 of the 526 columns.

Our analysis specifically focused on the cancer type metadata.

Data Analyis:

Methods

The machine learning technique I am using is: random forest classifier. This classifier optimizes for accuracy by reducing both bias and variance through ensemble learning. Each decision tree in the forest is trained on a random bootstrapped sample of the data and features, and the final prediction is made by aggregating the votes. The model determines it is “good enough” based on metrics such as accuracy, precision, AUC, or out-of-bag error on a test set.

Analysis

We trained a random forest classifier on the RNA count data to classify the different cancer types. We used a confusion matrix to find which cancers were conflated by the random forest, suggesting similar expression levels and similar disease pathways.

In [2]:
```
# FILE: visualize.ipynb, exported as py
# %%
import pandas as pd
import os
```

```
import csv

# %%
data_path = "data"

log2tpm_path = "GSE62944_subsample_log2TPM.csv"

metadata_path = "GSE62944_metadata_percent_nonNA_by_cancer_type.csv"

log2tpm = pd.read_csv(os.path.join(data_path, log2tpm_path), index_col=0)
metadata = pd.read_csv(os.path.join(data_path, metadata_path), index_col=0)

# %%
display(log2tpm)
display(metadata.head())

# %%
print(log2tpm.columns)

# %%
metadata_path = "GSE62944_metadata.csv"

metadata = pd.read_csv(os.path.join(data_path, metadata_path), index_col=0)

# display(metadata.head())

# display(metadata["cancer_type"].T.value_counts())


cols = pd.Series(log2tpm.columns).to_frame().T
# add name to first row "sample"

cols.columns = log2tpm.columns
cols.set_index(pd.Index(["sample"]), inplace=True)

log2tpm_temp_merge = pd.concat([cols, log2tpm], ignore_index=False)  # temp for mer

# display(log2tpm_temp_merge)


merged = log2tpm_temp_merge.T.merge(metadata, left_on="sample", right_index=True).T
merged = merged.drop("sample")

# %%
display(
    pd.concat([merged.head(), merged.tail(71)], axis=0)
)  # display 5 genes, 71 metadata rows
```

	TCGA-E9-A1NI-01A-11R-A14D-07	TCGA-E2-A1LK-01A-21R-A14D-07	TCGA-BH-A0B2-01A-11R-A10J-07	TCGA-E2-A107-01A-11R-A10J-07	TCGA-LL-A5YN-01A-11R-A28M-07	TCGA-BH-A0DQ-01A-11R-A084-07	TCGA-D8-A73X-01A-11R-A32P-07	TCGA-AR-A0TP-01A-11R-A084-07	A1
A1BG	3.397369	3.466089	3.789771	3.967578	4.733007	3.011343	4.324578	3.687565	2.3
A1CF	0.008857	0.039562	0.065051	0.000000	0.014260	0.008014	0.021112	0.020984	0.0
A2M	7.575125	6.643613	9.024479	7.573842	7.459105	9.279760	7.526283	12.770294	8.2
A2ML1	0.397610	7.625124	0.428689	0.465410	1.126008	0.083225	0.274456	0.220169	0.1
A4GALT	5.277425	5.244677	4.072650	4.208381	5.249234	5.123996	5.465881	3.083101	5.3
...	...	...	...	...	...	...	...	...	
ZYG11A	1.658629	2.652582	2.507865	2.362634	2.432761	1.672325	2.001393	1.917802	2.2
ZYG11B	3.493572	3.785275	4.166826	3.468905	2.606925	4.017574	3.577492	4.326777	3.4
ZYX	7.014571	7.094611	7.705058	7.618421	7.957508	8.138775	7.582862	9.333544	7.9
ZZEF1	3.665065	3.384731	3.715974	4.325992	3.257663	4.058356	4.132838	3.294119	3.0
ZZZ3	4.253068	4.253857	4.444500	4.097372	3.636987	4.636269	4.776437	5.297698	3.9

15716 rows × 1802 columns

cancer_type	cancer_type_full	sample	cancer_type.1	bcr_patient_barcode	ajcc_pathologic_tun
ACC	Adrenocortical Carcinoma	1.0	1.0	1.0000	
BLCA	Bladder Urothelial Carcinoma	1.0	1.0	0.9500	
BRCA	Breast Invasive Carcinoma	1.0	1.0	0.9500	
CESC	Cervical Squamous Cell Carcinoma and Endocervi...	1.0	1.0	0.9625	
COAD	Colon Adenocarcinoma	1.0	1.0	0.9000	

5 rows × 99 columns

```
Index(['TCGA-E9-A1NI-01A-11R-A14D-07', 'TCGA-E2-A1LK-01A-21R-A14D-07',
       'TCGA-BH-A0B2-01A-11R-A10J-07', 'TCGA-E2-A107-01A-11R-A10J-07',
       'TCGA-LL-A5YN-01A-11R-A28M-07', 'TCGA-BH-A0DQ-01A-11R-A084-07',
       'TCGA-D8-A73X-01A-11R-A32P-07', 'TCGA-AR-A0TP-01A-11R-A084-07',
       'TCGA-E2-A1IF-01A-11R-A144-07', 'TCGA-EW-A6SD-01A-12R-A33J-07',
       ...
       'TCGA-N5-A4RF-01A-11R-A28V-07', 'TCGA-N6-A4VF-01A-31R-A28V-07',
       'TCGA-N5-A4RN-01A-12R-A28V-07', 'TCGA-QM-A5NM-01A-11R-A28V-07',
       'TCGA-N5-A4RJ-01A-11R-A28V-07', 'TCGA-N5-A4RO-01A-11R-A28V-07',
       'TCGA-N5-A4RV-01A-21R-A28V-07', 'TCGA-N6-A4VD-01A-11R-A28V-07',
       'TCGA-N5-A4RT-01A-11R-A28V-07', 'TCGA-ND-A4WC-01A-21R-A28V-07'],
      dtype='object', length=1802)
```

	TCGA-E9-A1NI-01A-11R-A14D-07	TCGA-E2-A1LK-01A-21R-A14D-07	TCGA-BH-A0B2-01A-11R-A10J-07	TCGA-E2-A107-01A-11R-A10J-07	TCGA-LL-A5YN-01A-11R-A28M-07	TCGA-BH-A0DQ-01A-11R-A084-07	
A1BG	3.397369	3.466089	3.789771	3.967578	4.733007	3.011343	4.
A1CF	0.008857	0.039562	0.065051	0.0	0.01426	0.008014	0.
A2M	7.575125	6.643613	9.024479	7.573842	7.459105	9.27976	7.
A2ML1	0.39761	7.625124	0.428689	0.46541	1.126008	0.083225	0.
A4GALT	5.277425	5.244677	4.07265	4.208381	5.249234	5.123996	5.
...	...	...	...	...	...	...	
family_history_cancer_type	NaN	NaN	NaN	NaN	NaN	NaN	
tumor_response	NaN	NaN	NaN	NaN	NaN	NaN	
ecog_score	NaN	NaN	NaN	NaN	NaN	NaN	
lymph_nodes_examined_count	YES	YES	YES	[Not Available]	YES	[Not Available]	
residual_tumor.1	NaN	NaN	NaN	NaN	NaN	NaN	

76 rows × 1802 columns

In [3]:
```
# FILE: random_forest.ipynb, exported as py
# %%
import sklearn as sk
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import os

# %%
# get data
data_path = "data"

log2tpm_path = "GSE62944_subsample_log2TPM.csv"

metadata_path = "GSE62944_metadata.csv"

log2tpm = pd.read_csv(os.path.join(data_path, log2tpm_path), index_col=0)
metadata = pd.read_csv(os.path.join(data_path, metadata_path), index_col=0)

display(log2tpm.head())
display(metadata.head())
```

```
# %%
# random forest classification using all features
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix

y = "cancer_type"


# select features and target
features = log2tpm.copy().T
target = metadata[y].T.to_frame()

display(target.head())
display(features.head())


# one-hot encode categorical variables
features = pd.get_dummies(features, drop_first=True)
# split into training and testing sets, set seed for reproducibility
X_train, X_test, y_train, y_test = train_test_split(
    features, target, test_size=0.2, random_state=42
)
# train the model
rf = RandomForestClassifier(n_estimators=100, random_state=10)
rf.fit(X_train, y_train.values.ravel())
# make predictions
y_pred = rf.predict(X_test)
# evaluate the model
accuracy = accuracy_score(y_test, y_pred)
print(f"Random Forest Classifier Accuracy: {accuracy:.4f}")
print("\nClassification Report:")
print(classification_report(y_test, y_pred))
print("\nConfusion Matrix:")
print(confusion_matrix(y_test, y_pred))
# feature importance
feature_names = features.columns
feature_importances = pd.Series(
    rf.feature_importances_, index=feature_names
).sort_values(ascending=False)
print("\nFeature importances:")
print(feature_importances.head(20))


# %%
# sort by precision

report = classification_report(y_test, y_pred, output_dict=True)
# extract precision
class_precisions = {
    k: v["precision"]
    for k, v in report.items()
    if isinstance(v, dict)
    and "precision" in v
    and k not in ["macro avg", "weighted avg"]
}
```

```
sorted_precisions = sorted(class_precisions.items(), key=lambda x: x[1], reverse=Tr

print("Cancer types sorted by precision:")
for cancer, precision in sorted_precisions:
    print(f"{cancer}: {precision:.4f}")

# %%
# plot confusion matrix to see conflated cancer types
plt.figure(figsize=(10, 7))
sns.heatmap(
    confusion_matrix(y_test, y_pred),
    annot=True,
    fmt="d",
    cmap="Blues",
    xticklabels=rf.classes_,
    yticklabels=rf.classes_,
)
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Confusion Matrix")
plt.show()

# %%
cm = confusion_matrix(y_test, y_pred, labels=rf.classes_)
labels = rf.classes_

# find the most conflated pairs
conflation = []
for i in range(len(labels)):
    for j in range(len(labels)):
        if i != j:  # diagonal is accurate prediction- no conflate
            conflation.append(((labels[i], labels[j]), cm[i, j]))

# sort by number of misclassifications
conflation_sorted = sorted(conflation, key=lambda x: x[1], reverse=True)
print("Most conflated cancer type pairs:")
for (actual, predicted), count in conflation_sorted[:10]:
    if count > 0:
        print(f"{actual} predicted as {predicted}: {count} times")
```

	TCGA-E9-A1NI-01A-11R-A14D-07	TCGA-E2-A1LK-01A-21R-A14D-07	TCGA-BH-A0B2-01A-11R-A10J-07	TCGA-E2-A107-01A-11R-A10J-07	TCGA-LL-A5YN-01A-11R-A28M-07	TCGA-BH-A0DQ-01A-11R-A084-07	TCGA-D8-A73X-01A-11R-A32P-07	TCGA-AR-A0TP-01A-11R-A084-07	A1
A1BG	3.397369	3.466089	3.789771	3.967578	4.733007	3.011343	4.324578	3.687565	2.3
A1CF	0.008857	0.039562	0.065051	0.000000	0.014260	0.008014	0.021112	0.020984	0.0
A2M	7.575125	6.643613	9.024479	7.573842	7.459105	9.279760	7.526283	12.770294	8.2
A2ML1	0.397610	7.625124	0.428689	0.465410	1.126008	0.083225	0.274456	0.220169	0.1
A4GALT	5.277425	5.244677	4.072650	4.208381	5.249234	5.123996	5.465881	3.083101	5.3

5 rows × 1802 columns

sample	cancer_type	bcr_patient_barcode	bcr_patient_uuid	patient_id	gender	race
TCGA-E9-A1NI-01A-11R-A14D-07	BRCA	TCGA-E9-A1NI	c65b835a-2d38-4250-a173-0780d2c2cf58	A1NI	FEMALE	WHITE
TCGA-E2-A1LK-01A-21R-A14D-07	BRCA	TCGA-E2-A1LK	1d27253f-b036-44e7-a04d-8da5bbf57419	A1LK	FEMALE	BLACK OR AFRICAN AMERICAN
TCGA-BH-A0B2-01A-11R-A10J-07	BRCA	TCGA-BH-A0B2	57a1604c-60b7-4b30-a75e-f70939532c5c	A0B2	FEMALE	WHITE
TCGA-E2-A107-01A-11R-A10J-07	BRCA	TCGA-E2-A107	eff2360e-399a-4167-ab2b-798e27bef739	A107	FEMALE	WHITE
TCGA-LL-A5YN-01A-11R-A28M-07	BRCA	TCGA-LL-A5YN	EF1B3332-CD7F-41BB-A2D3-2538E7BECC5C	A5YN	FEMALE	BLACK OR AFRICAN AMERICAN

5 rows × 71 columns

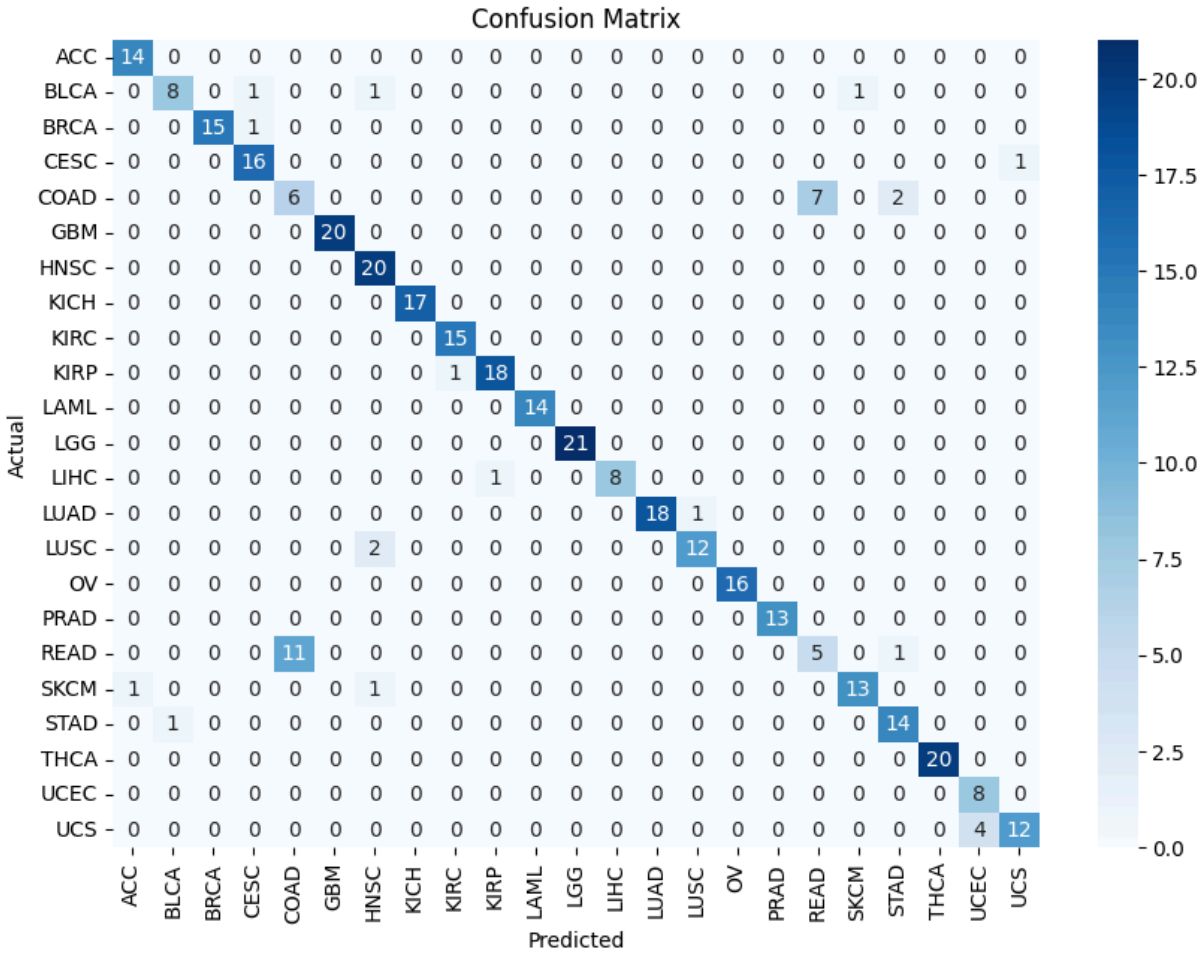
sample	cancer_type
TCGA-E9-A1NI-01A-11R-A14D-07	BRCA
TCGA-E2-A1LK-01A-21R-A14D-07	BRCA
TCGA-BH-A0B2-01A-11R-A10J-07	BRCA
TCGA-E2-A107-01A-11R-A10J-07	BRCA
TCGA-LL-A5YN-01A-11R-A28M-07	BRCA

	A1BG	A1CF	A2M	A2ML1	A4GALT	AAAS	AACS	AADAC	AAI
TCGA-E9-A1NI-01A-11R-A14D-07	3.397369	0.008857	7.575125	0.397610	5.277425	6.212456	5.245303	0.051776	1.476
TCGA-E2-A1LK-01A-21R-A14D-07	3.466089	0.039562	6.643613	7.625124	5.244677	5.326190	5.787772	0.114339	3.802
TCGA-BH-A0B2-01A-11R-A10J-07	3.789771	0.065051	9.024479	0.428689	4.072650	5.379163	4.717324	0.278396	1.656
TCGA-E2-A107-01A-11R-A10J-07	3.967578	0.000000	7.573842	0.465410	4.208381	6.475950	6.015882	0.000000	1.364
TCGA-LL-A5YN-01A-11R-A28M-07	4.733007	0.014260	7.459105	1.126008	5.249234	6.162779	5.161940	0.122220	0.900

5 rows × 15716 columns

```
Random Forest Classifier Accuracy: 0.8947

Classification Report:
              precision    recall  f1-score   support

         ACC       0.93      1.00      0.97        14
        BLCA       0.89      0.73      0.80        11
        BRCA       1.00      0.94      0.97        16
        CESC       0.89      0.94      0.91        17
        COAD       0.35      0.40      0.38        15
         GBM       1.00      1.00      1.00        20
        HNSC       0.83      1.00      0.91        20
        KICH       1.00      1.00      1.00        17
        KIRC       0.94      1.00      0.97        15
        KIRP       0.95      0.95      0.95        19
        LAML       1.00      1.00      1.00        14
         LGG       1.00      1.00      1.00        21
        LIHC       1.00      0.89      0.94         9
        LUAD       1.00      0.95      0.97        19
        LUSC       0.92      0.86      0.89        14
          OV       1.00      1.00      1.00        16
        PRAD       1.00      1.00      1.00        13
        READ       0.42      0.29      0.34        17
        SKCM       0.93      0.87      0.90        15
        STAD       0.82      0.93      0.88        15
        THCA       1.00      1.00      1.00        20
        UCEC       0.67      1.00      0.80         8
         UCS       0.92      0.75      0.83        16

    accuracy                           0.89       361
   macro avg       0.89      0.89      0.89       361
weighted avg       0.90      0.89      0.89       361


Confusion Matrix:
[[14  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  8  0  1  0  0  1  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0]
 [ 0  0 15  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0 16  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1]
 [ 0  0  0  0  6  0  0  0  0  0  0  0  0  0  0  0  0  7  0  2  0  0  0]
 [ 0  0  0  0  0 20  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0 20  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0 17  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0 15  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  1 18  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0 14  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0 21  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  1  0  0  8  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0 18  1  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  2  0  0  0  0  0  0  0 12  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 16  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 13  0  0  0  0  0  0]
 [ 0  0  0  0 11  0  0  0  0  0  0  0  0  0  0  0  0  5  0  1  0  0  0]
 [ 1  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0 13  0  0  0  0]
 [ 0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 14  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 20  0  0]
```

```
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  8  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  4 12]]

Feature importances:
TRPM8      0.002854
FEZF2      0.002825
RPS10      0.002765
STAR       0.002747
CYP11A1    0.002743
ESR1       0.002588
TRPS1      0.002549
CREB3L4    0.002351
CRYGN      0.002319
PAX8       0.002313
VTN        0.002215
C1orf61    0.002187
PMP2       0.002121
AGRP       0.002113
S100B      0.002071
BCAR1      0.002033
APOA2      0.001998
SOX17      0.001956
GALNT14    0.001934
TG         0.001928
dtype: float64
Cancer types sorted by precision:
BRCA: 1.0000
GBM: 1.0000
KICH: 1.0000
LAML: 1.0000
LGG: 1.0000
LIHC: 1.0000
LUAD: 1.0000
OV: 1.0000
PRAD: 1.0000
THCA: 1.0000
KIRP: 0.9474
KIRC: 0.9375
ACC: 0.9333
SKCM: 0.9286
LUSC: 0.9231
UCS: 0.9231
BLCA: 0.8889
CESC: 0.8889
HNSC: 0.8333
STAD: 0.8235
UCEC: 0.6667
READ: 0.4167
COAD: 0.3529
```




```
Most conflated cancer type pairs:
READ predicted as COAD: 11 times
COAD predicted as READ: 7 times
UCS predicted as UCEC: 4 times
COAD predicted as STAD: 2 times
LUSC predicted as HNSC: 2 times
BLCA predicted as CESC: 1 times
BLCA predicted as HNSC: 1 times
BLCA predicted as SKCM: 1 times
BRCA predicted as CESC: 1 times
CESC predicted as UCS: 1 times
```

The following code creates a random forest classifier for a specific type of cancer, PRAD (Prostate Adenocarcinoma). Then, it performs feature selection to find the top 20 most significant genes for PRAD according to our data.

In [ ]:
```
# random forest classification for PRAD

from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix

# split by specific cancer vs not
metadata["PRAD_pos"] = metadata["cancer_type"].apply(lambda x: "PRAD" in x)
display(metadata)
```

```
y = "PRAD_pos"


# select features and target
features = log2tpm.copy().T
target = metadata[y].T.to_frame()

display(target.head())
display(features.head())


# one-hot encode categorical variables
features = pd.get_dummies(features, drop_first=True)
# split into training and testing sets, set seed for reproducibility
X_train, X_test, y_train, y_test = train_test_split(
    features, target, test_size=0.2, random_state=42
)
# train the model
rf = RandomForestClassifier(n_estimators=100, random_state=20)
rf.fit(X_train, y_train.values.ravel())
# make predictions
y_pred = rf.predict(X_test)
# evaluate the model
accuracy = accuracy_score(y_test, y_pred)
print(f"Random Forest Classifier Accuracy: {accuracy:.4f}")
print("\nClassification Report:")
print(classification_report(y_test, y_pred))
# feature importance
feature_names = features.columns
feature_importances = pd.Series(
    rf.feature_importances_, index=feature_names
).sort_values(ascending=False)

# plot confusion matrix to see PRAD_pos conflation
plt.figure(figsize=(10, 7))
sns.heatmap(
    confusion_matrix(y_test, y_pred),
    annot=True,
    fmt="d",
    cmap="Blues",
    xticklabels=rf.classes_,
    yticklabels=rf.classes_,
)
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Confusion Matrix")
plt.show()
```

sample	cancer_type	bcr_patient_barcode	bcr_patient_uuid	patient_id	gender	race
TCGA-E9-A1NI-01A-11R-A14D-07	BRCA	TCGA-E9-A1NI	c65b835a-2d38-4250-a173-0780d2c2cf58	A1NI	FEMALE	WHITE
TCGA-E2-A1LK-01A-21R-A14D-07	BRCA	TCGA-E2-A1LK	1d27253f-b036-44e7-a04d-8da5bbf57419	A1LK	FEMALE	BLACK OR AFRICAN AMERICAN
TCGA-BH-A0B2-01A-11R-A10J-07	BRCA	TCGA-BH-A0B2	57a1604c-60b7-4b30-a75e-f70939532c5c	A0B2	FEMALE	WHITE
TCGA-E2-A107-01A-11R-A10J-07	BRCA	TCGA-E2-A107	eff2360e-399a-4167-ab2b-798e27bef739	A107	FEMALE	WHITE
TCGA-LL-A5YN-01A-11R-A28M-07	BRCA	TCGA-LL-A5YN	EF1B3332-CD7F-41BB-A2D3-2538E7BECC5C	A5YN	FEMALE	BLACK OR AFRICAN AMERICAN
...	...	...	...	...	...	...
TCGA-N5-A4RO-01A-11R-A28V-07	UCS	TCGA-N5-A4RO	4F4906DC-7EBD-47F1-A8F5-B35D3950E740	A4RO	FEMALE	BLACK OR AFRICAN AMERICAN
TCGA-N5-A4RV-01A-	UCS	TCGA-N5-A4RV	3AF5B391-E72F-463D-A086-A86C6C30A51A	A4RV	FEMALE	WHITE

sample	cancer_type	bcr_patient_barcode	bcr_patient_uuid	patient_id	gender	race
21R-A28V-07						
TCGA-N6-A4VD-01A-11R-A28V-07	UCS	TCGA-N6-A4VD	14213209-2217-4812-9A19-D9B2B6718467	A4VD	FEMALE	WHITE
TCGA-N5-A4RT-01A-11R-A28V-07	UCS	TCGA-N5-A4RT	1791E250-70AC-439C-828B-15BA811935CC	A4RT	FEMALE	WHITE
TCGA-ND-A4WC-01A-21R-A28V-07	UCS	TCGA-ND-A4WC	6385194A-D75B-4BD3-9E70-9AA36250D5B9	A4WC	FEMALE	ASIAN

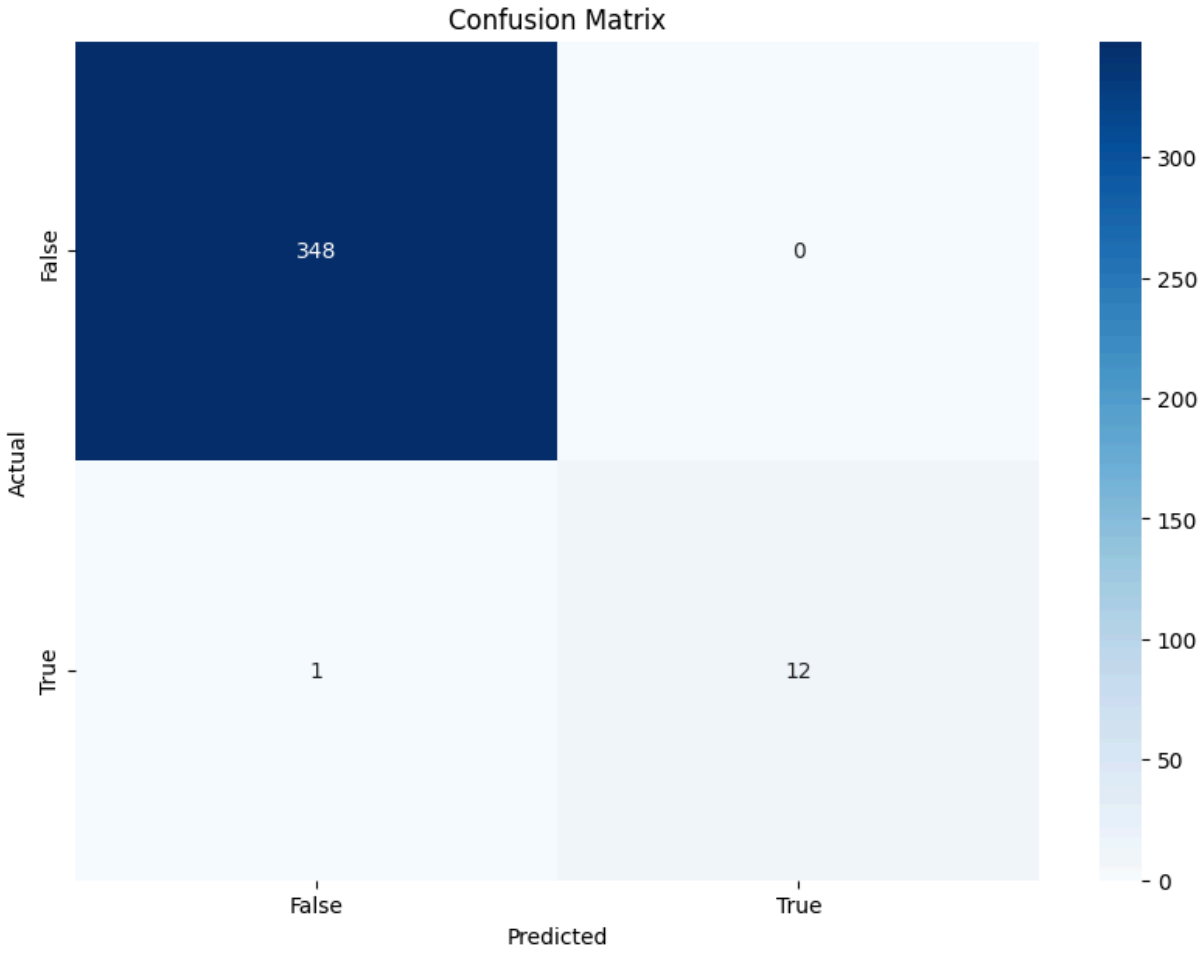
1802 rows × 72 columns

sample	PRAD_pos
TCGA-E9-A1NI-01A-11R-A14D-07	False
TCGA-E2-A1LK-01A-21R-A14D-07	False
TCGA-BH-A0B2-01A-11R-A10J-07	False
TCGA-E2-A107-01A-11R-A10J-07	False
TCGA-LL-A5YN-01A-11R-A28M-07	False

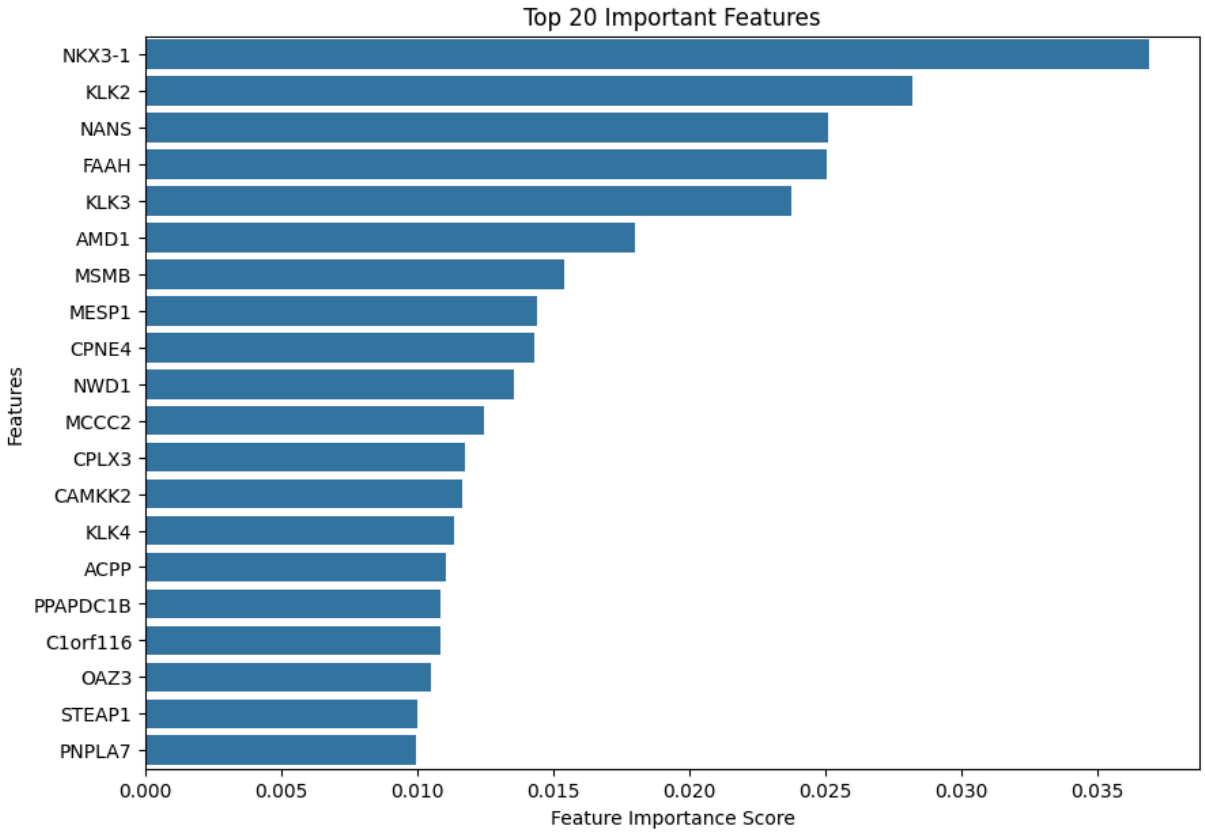
	A1BG	A1CF	A2M	A2ML1	A4GALT	AAAS	AACS	AADAC	AAI
TCGA-E9-A1NI-01A-11R-A14D-07	3.397369	0.008857	7.575125	0.397610	5.277425	6.212456	5.245303	0.051776	1.476
TCGA-E2-A1LK-01A-21R-A14D-07	3.466089	0.039562	6.643613	7.625124	5.244677	5.326190	5.787772	0.114339	3.802
TCGA-BH-A0B2-01A-11R-A10J-07	3.789771	0.065051	9.024479	0.428689	4.072650	5.379163	4.717324	0.278396	1.656
TCGA-E2-A107-01A-11R-A10J-07	3.967578	0.000000	7.573842	0.465410	4.208381	6.475950	6.015882	0.000000	1.364
TCGA-LL-A5YN-01A-11R-A28M-07	4.733007	0.014260	7.459105	1.126008	5.249234	6.162779	5.161940	0.122220	0.900

5 rows × 15716 columns

```
Random Forest Classifier Accuracy: 0.9972

Classification Report:
              precision    recall  f1-score   support

       False       1.00      1.00      1.00       348
        True       1.00      0.92      0.96        13

    accuracy                           1.00       361
   macro avg       1.00      0.96      0.98       361
weighted avg       1.00      1.00      1.00       361
```



In [9]:
```python
# find top 20 important features
top_20_features = feature_importances.head(20)
# plot top 20 important features
plt.figure(figsize=(10, 7))
sns.barplot(x=top_20_features.values, y=top_20_features.index)
plt.xlabel("Feature Importance Score")
plt.ylabel("Features")
plt.title("Top 20 Important Features")
plt.show()
```



Verify and validate your analysis:

A random forest is an ensemble learning method that builds multiple decision trees and combines their predictions. It's like asking a crowd of experts (trees) for their opinion and taking a vote, rather than relying on just one expert. A confusion matrix displays how well a model's predictions match the actual outcomes. Any predictions off the diagonal show when the model gets "confused."

Our total confusion matrix is very promising, showing minimal conflation outside of a few certain cancer types. We also used a train/test split in our model. As shown in the results above, our general (all cancer treated separately) random forest classifier accuracy is 0.8947 and the accuracy of our specific classifier for PRAD is 0.9972. According to Claude, most published cancer studies achieve an accuracy between 0.85-0.95 with a random forest classifier.

The feature selection for prostate adenocarcinoma led to a very exciting result. The NKX3-1, KLK2, and KLK3 genes have all been shown to play crucial roles in prostate function and cancer detection. 3 out of our 5 most important genes from this single data set support these findings.

Sources: https://pmc.ncbi.nlm.nih.gov/articles/PMC5844678/
https://pmc.ncbi.nlm.nih.gov/articles/PMC10126698/

https://www.ncbi.nlm.nih.gov/gene/354 https://www.ncbi.nlm.nih.gov/gene/3817 https://www.ncbi.nlm.nih.gov/gene/4824

Conclusions and Ethical Implications:

Conclusions

The most immunoevasive cancer types in this data set (PRAD, OV, BRCA) have almost zero conflation. This means these cancer types are biologically unique, compared to more conflated cancer types (READ, UCEC) which are biologically similar to others. These immunoevasive cancer types may use unique biological mechanisms and pathways because using common mechanisms would make them more detectable by the immune system.

The most over-expressed pathways for prostate adenocarcinoma that our model found which are corroborated by previous findings are NKX3-1, KLK2, and KLK3. Our promising feature selection results show that looking at genes expression could predict cancer diagnosis with enough data and a perfected ML model.

Ethical Implications

Because these immunoevasive cancer types are biologically unique, they may require more specified potentially expensive treatments. There are racial disparities in prostate cancer especially - black men have a 2x higher mortality from prostate cancer than white men. Part of this disparity is likely socioeconomic, including access to health care and specialized treatments.

Predictive testing is also a big ethical debate in medicine. First of all, overtesting is already a problem in mental and physical disorders, leading to misinformation about health issues on the rise. Also, getting a predictive test and knowing you will likely get cancer or another detrimental disease can be positive for early prevention, but it can lead to mental health issues living with the weight of that knowledge.

Sources: https://blog.dana-farber.org/insight/2018/06/enhancing-immunotherapy-race-make-cold-tumors-hot/ https://www.ncbi.nlm.nih.gov/gene/354 https://www.ncbi.nlm.nih.gov/gene/3817 https://www.ncbi.nlm.nih.gov/gene/4824 https://pmc.ncbi.nlm.nih.gov/articles/PMC7342480/ https://pmc.ncbi.nlm.nih.gov/articles/PMC9826514/

Limitations and Future Work:

A demographic analysis of TCGA (our dataset) shows that white cases were over-represented and Asian and Hispanic cases were under-represented compared to the general population. This means applying this model clinically will be less affective for Asian and Hispanic

subjects. Additionally, TCGA primarily includes data from academic medical centers in the United States, potentially limiting generalizability to other healthcare settings and countries. Some cancer types had fewer samples than others, which may have affected model performance for rarer cancer types.

Predictive testing already exists for cancer and uses blood or saliva to test for genome mutations instead of more invasive tissue samples for gene expression. Therefore, this analysis shouldn't be adapted for predictive diagnosis. However, it could improve cancer classification. Tissue biopsy is already used to identify cancer types from a tumor, but this could improve detection of immunoevasive cancers.

NOTES FROM YOUR TEAM:

10/21

- Learned optimization methods
- Initial look through data

10/23

- Will out sick
- Max Looked deeper at data, performed random forest on genetic markers
- Will filled out background asynch

10/30

- Set up virtual environment on Will's system
- Decided on question
- Presented initial random forest analysis

11/6

- Used confusion matrix, accuracy score to begin verification section
- Started conclusions and ethical implications
- NKX3-1 gene shows as important feature, verified as important to prostate cancer, may add to analysis

11/8

- Completed verification section and submitted 3rd check in

11/11

- Added feature selection for PRAD to analysis section (Very promising results!)

- Completed conclusions and limitations sections

QUESTIONS FOR YOUR TA:

None