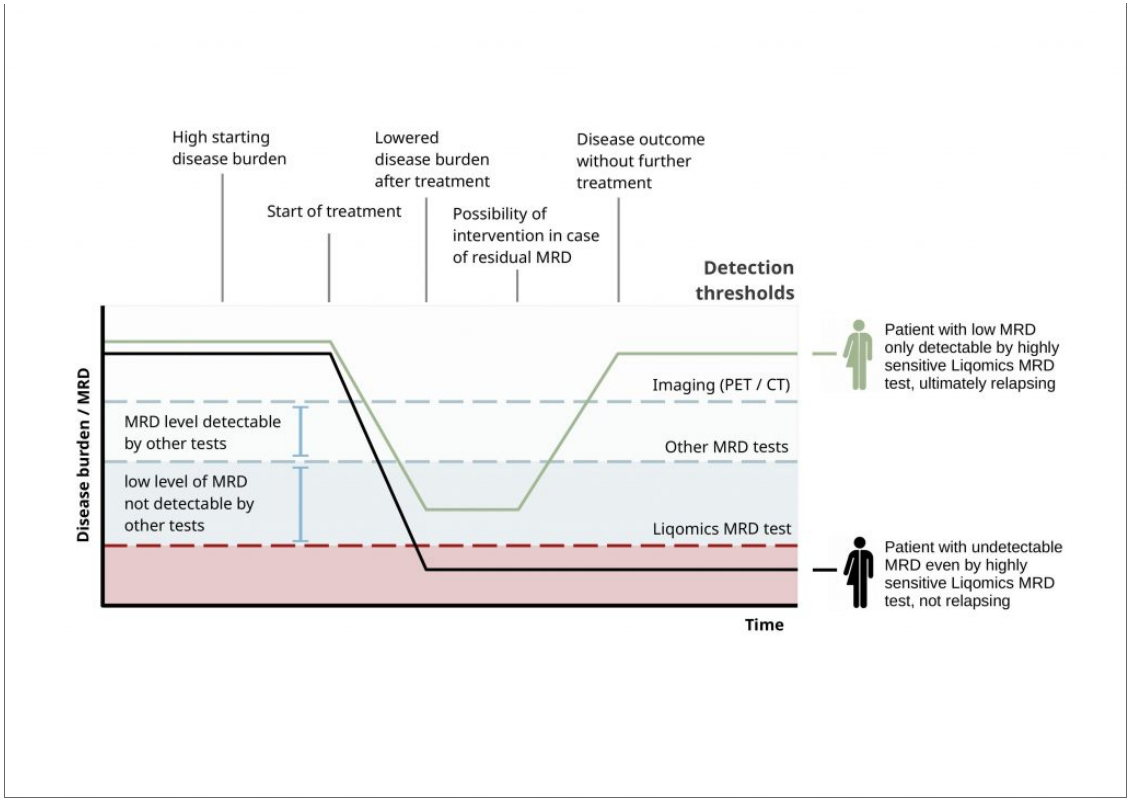
About LIQOMICS

- **Academic Spin-off from Cologne:**
  Focused on cutting-edge tumor diagnostics through liquid biopsy technology.

Key Service: LymphoVista

- **What It Does:**
  Detects and monitors **lymphomas** and **hematological cancers**.
- **How It Works:**
  - **Blood Sample Analysis:** Detects **cell-free DNA** and **circulating tumor DNA**.
  - **Disease Monitoring:** Tracks **Minimal Residual Disease (MRD)** to assess treatment response.

Sensitive MRD Testing for Cancer Monitoring



Cell-Free DNA (cfDNA)

What Is cfDNA?

- **Degraded Small DNA Fragments:**
    - Size: **50 - 200 bp**, cleaved by **nucleases**.

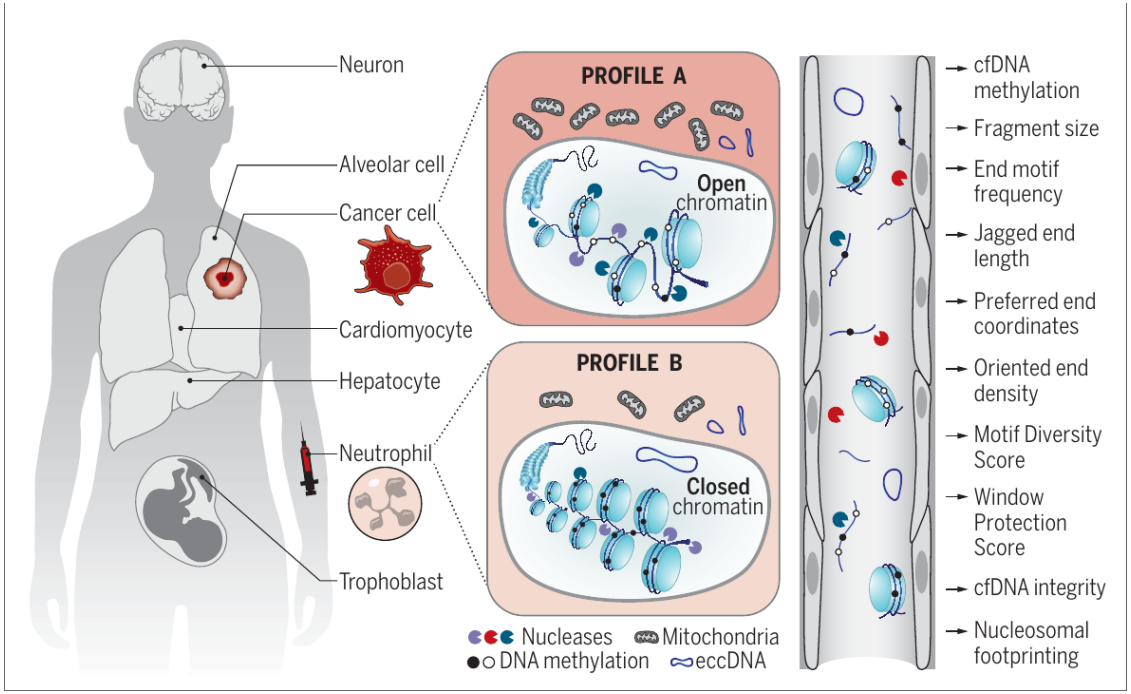
Source of cfDNA:

- **Released by Dead Cells:**
    - Occurs via **apoptosis**, **necrosis**, or **active secretion** from living cells.

Key Molecular Features:

- **Tissue-Specific Signatures:**
    - **Fragment Size:** Reflects tissue-specific fragmentation patterns.
    - **Methylation Status:** Epigenetic modifications characteristic of cell types.
    - **End Motifs:** Specific DNA ends reflecting nuclease activity.

Chromatin Organization and Nuclease Activity Define cfDNA Signatures



Projects Overview

---

1. Hodgkin Lymphoma Project

- **Data Source:**
    - Cell-free DNA sequencing from **Hodgkin Lymphoma Patients**.
    - Samples collected **before** and **after two cycles of chemotherapy**.
    - Additional **relapse information** included.
- **Goal:**
    - Build a **Machine Learning Model** to **predict relapse** after treatment.
- **Result:**
    - **Challenge:** Insufficient sample size for a reliable predictive model.

Projects Overview

---

2. Solid Tumors (Finale_DB Project)

- **Data Source:**
    - Cell-free DNA sequencing from **4 Publications:**
        - **Jiang_2015, Cristiano_2019, Snyder_2016, Sun_2019**.
    - **>800 Samples:** Includes both **healthy** and **cancer** samples.
- **Goal:**
    - Develop a **Machine Learning Model** to **distinguish cancer from non-cancer** for solid tumors.

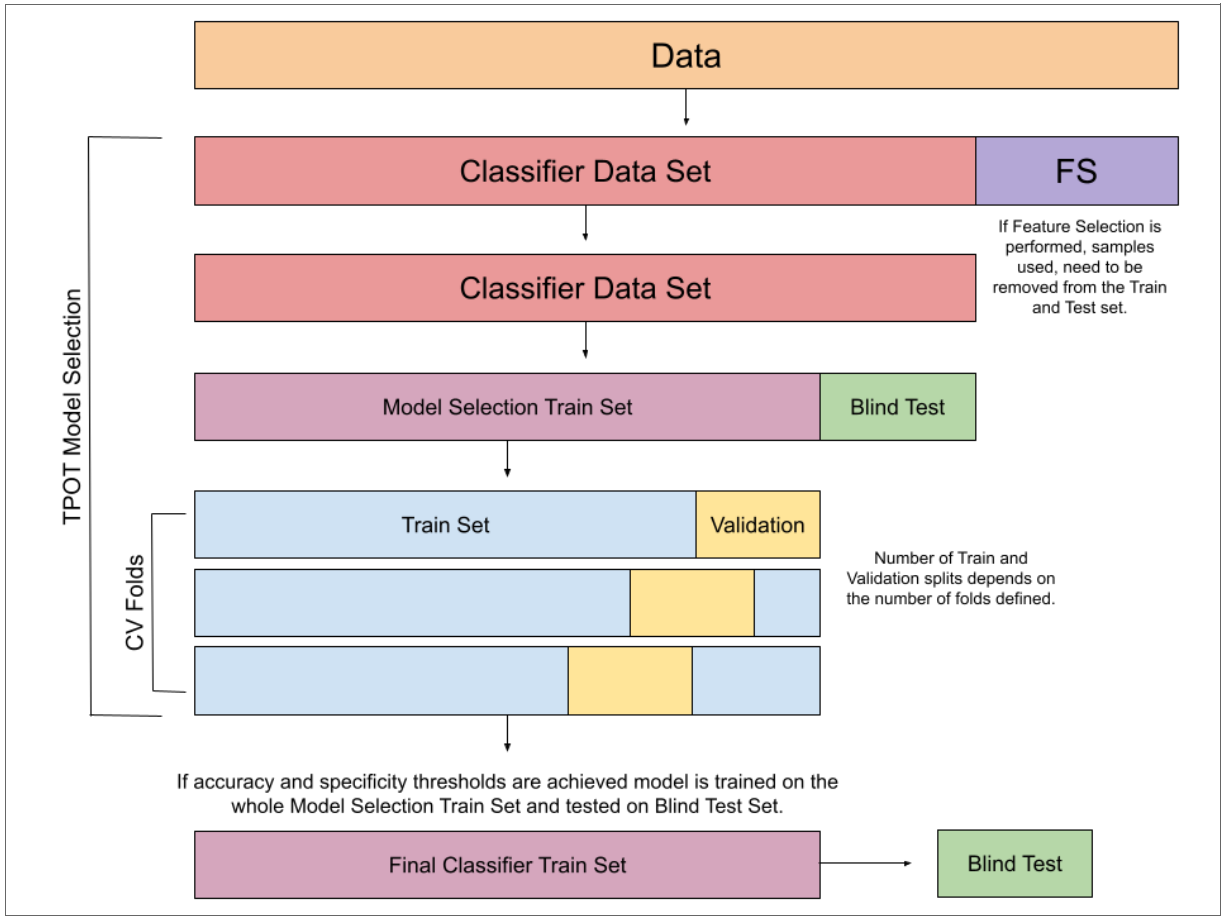
Note: Detailed results from this project will be presented on the following slides.

Summary of Data Sources and Samples

Source	Diseases Covered	Total Samples
Cristiano_2019	Bile_duct_cancer Breast_cancer Colorectal_cancer Duodenal_cancer Gastric_cancer Healthy Lung_cancer Ovarian_cancer Pancreatic_cancer	536
Jiang_2015	Cirrhosis Healthy Hepatitis_B Liver_cancer	225
Snyder_2016	Bladder_cancer Breast_cancer Colorectal_cancer Esophageal_cancer Head_and_neck_cancer Healthy Inflammatory_bowel_disease Kidney_cancer Liver_cancer Lung_cancer Ovarian_cancer Pancreatic_cancer Prostate_cancer Skin_cancer Systemic_lupus_erythematosus Testicular_cancer Uterine_cancer	58
Sun_2019	Colorectal_cancer Liver_transplant	29

Pipeline

1. Downsampling
2. Feature Normalization
3. Feature Selection
4. Machine learning (TPOT)

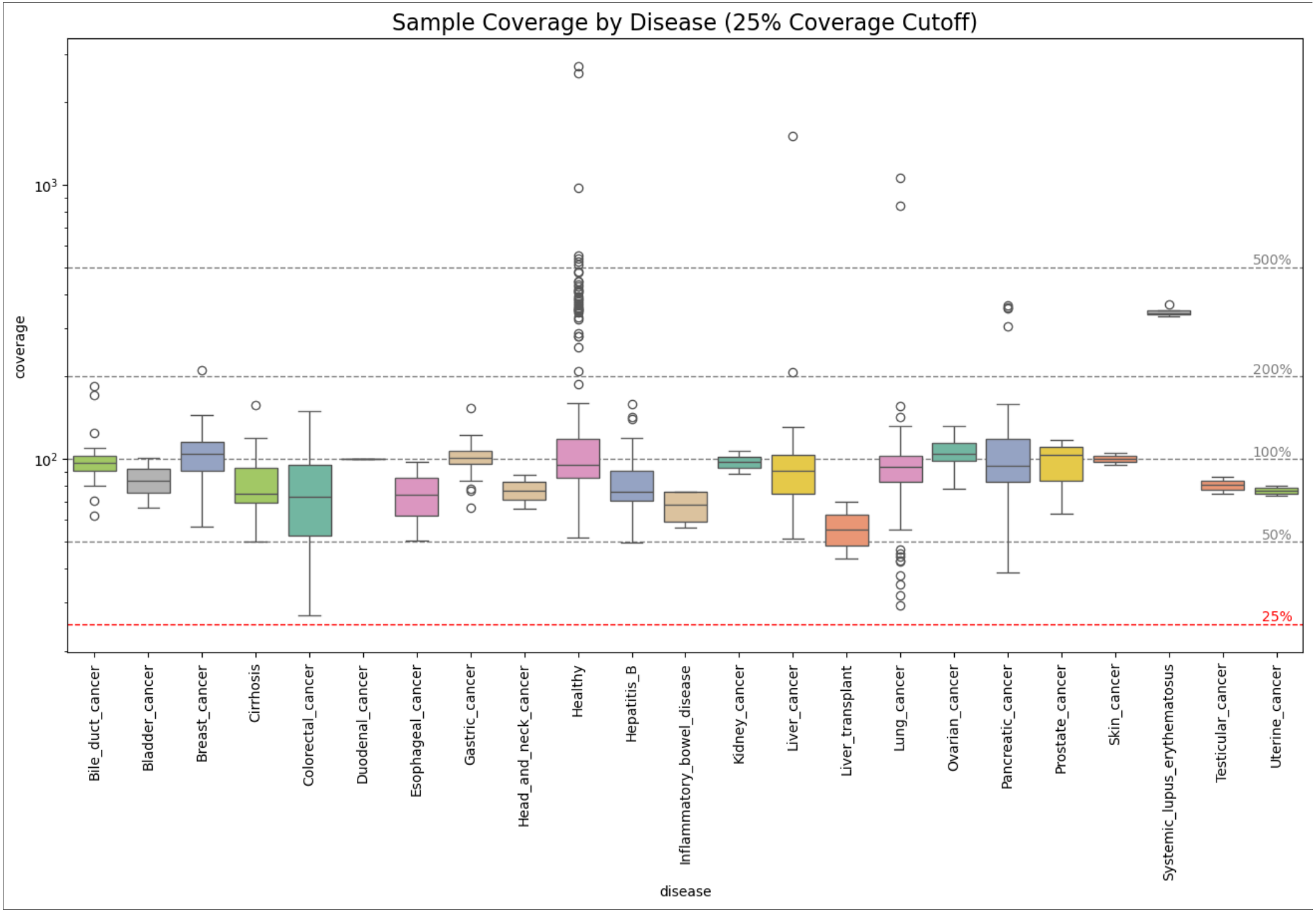


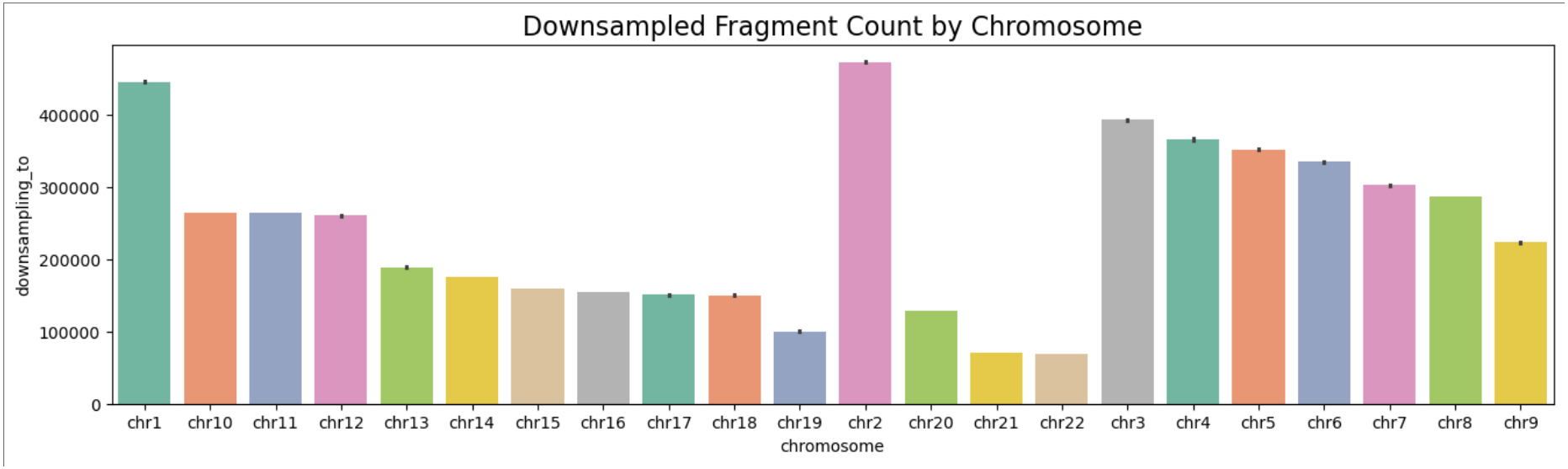
Downsampling Process

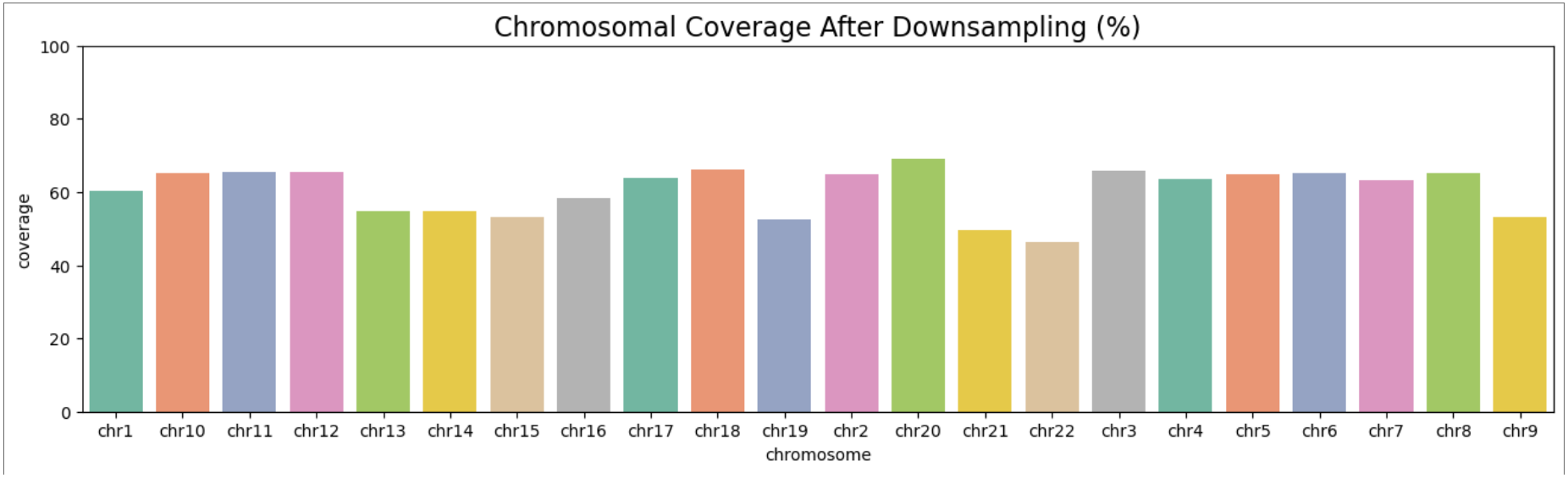
Key Steps:

1. **Handling Coverage Variability:**
    - Coverage between samples **varies significantly**, affecting comparability.
2. **Equalization:**
    - **Downsample all samples** to the same number of **DNA fragments** to ensure consistency in analysis.

```
Number of total samples after downsampling: 848
```



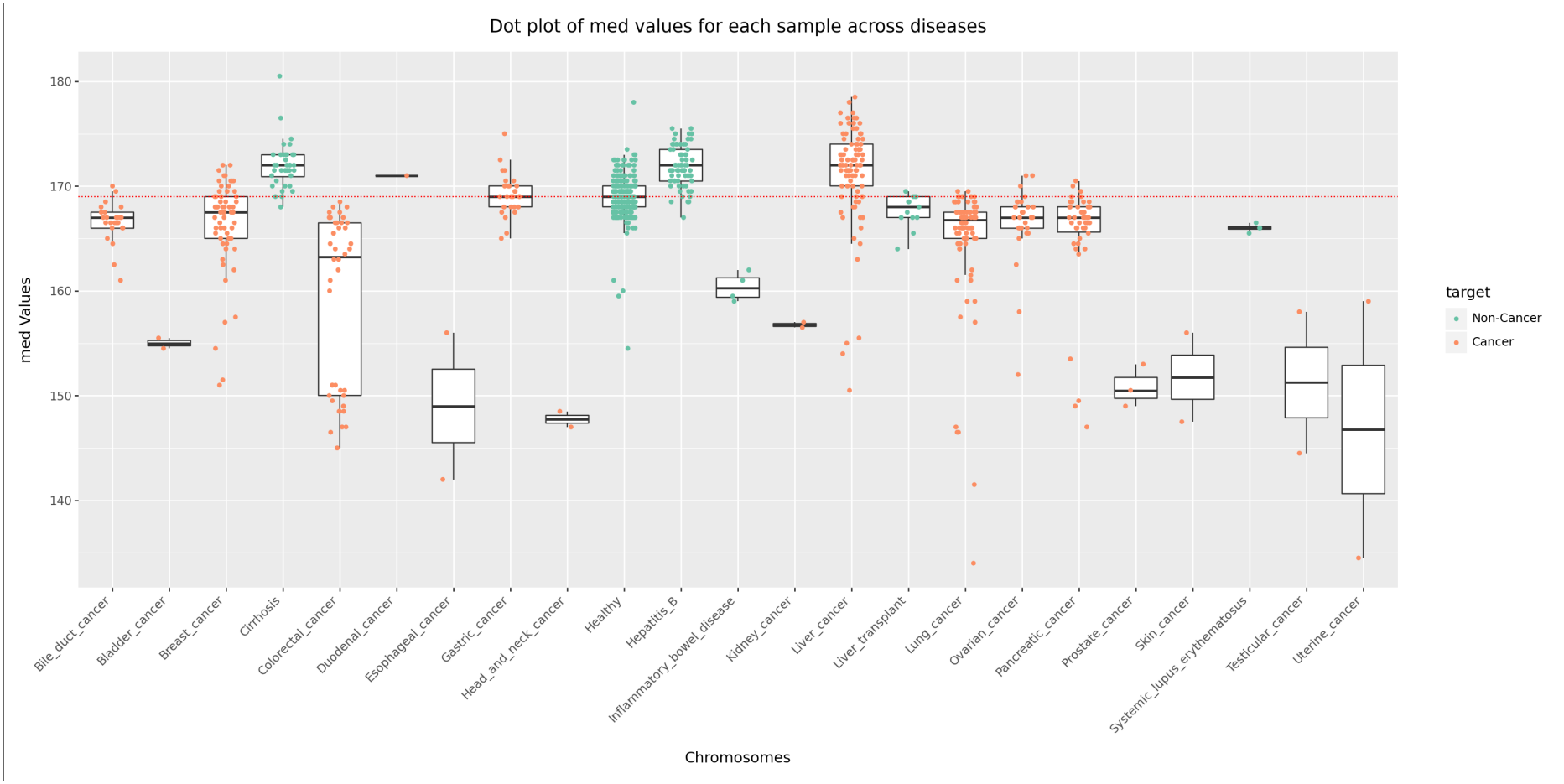

Downsampled Fragment Count by Chromosome
downsampling_to
0
100000
200000
300000
400000
chr1
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr2
chr20
chr21
chr22
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chromosome

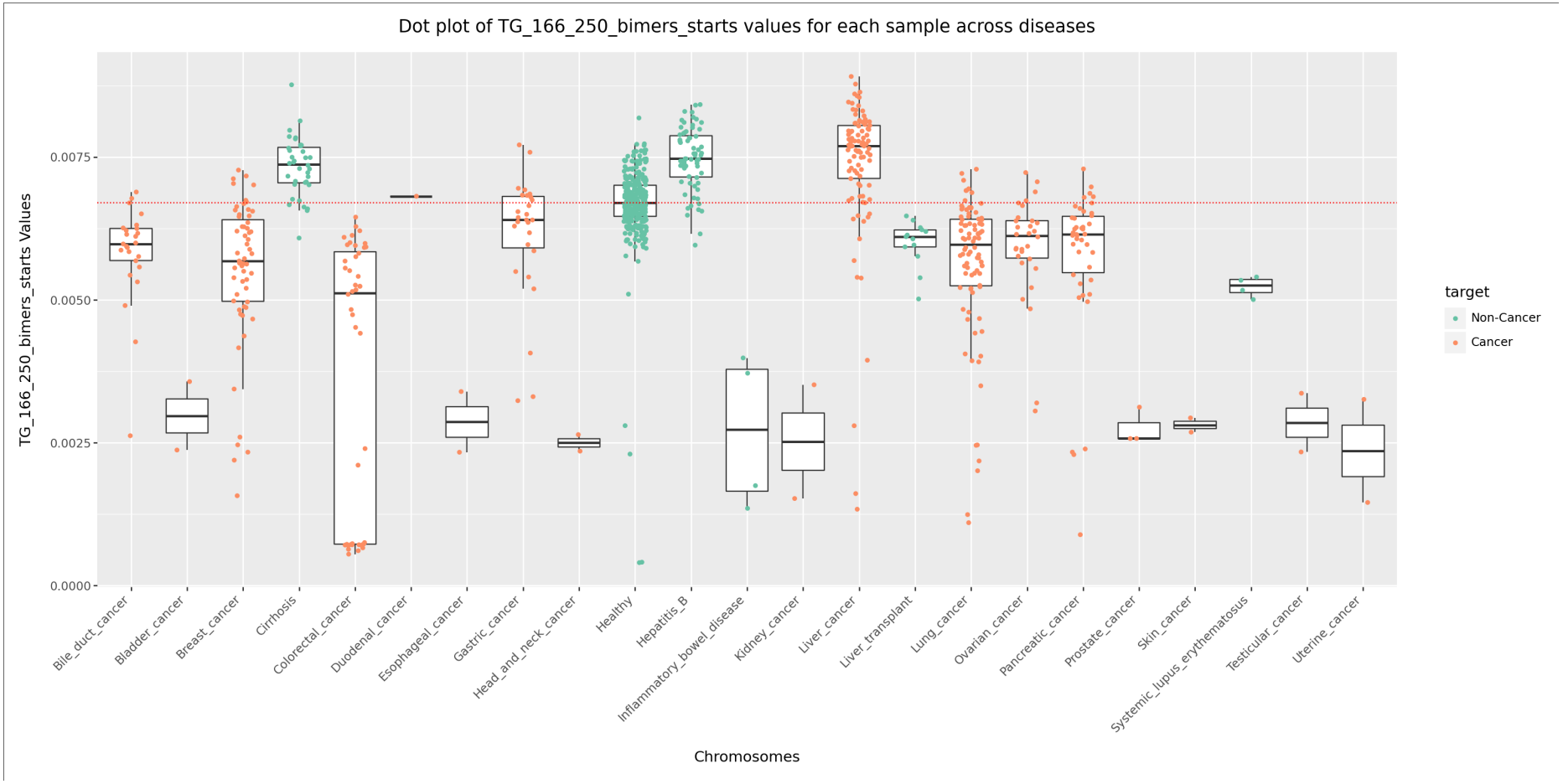

Chromosomal Coverage After Downsampling (%)
coverage
0
20
40
60
80
100
chr1
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr2
chr20
chr21
chr22
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chromosome

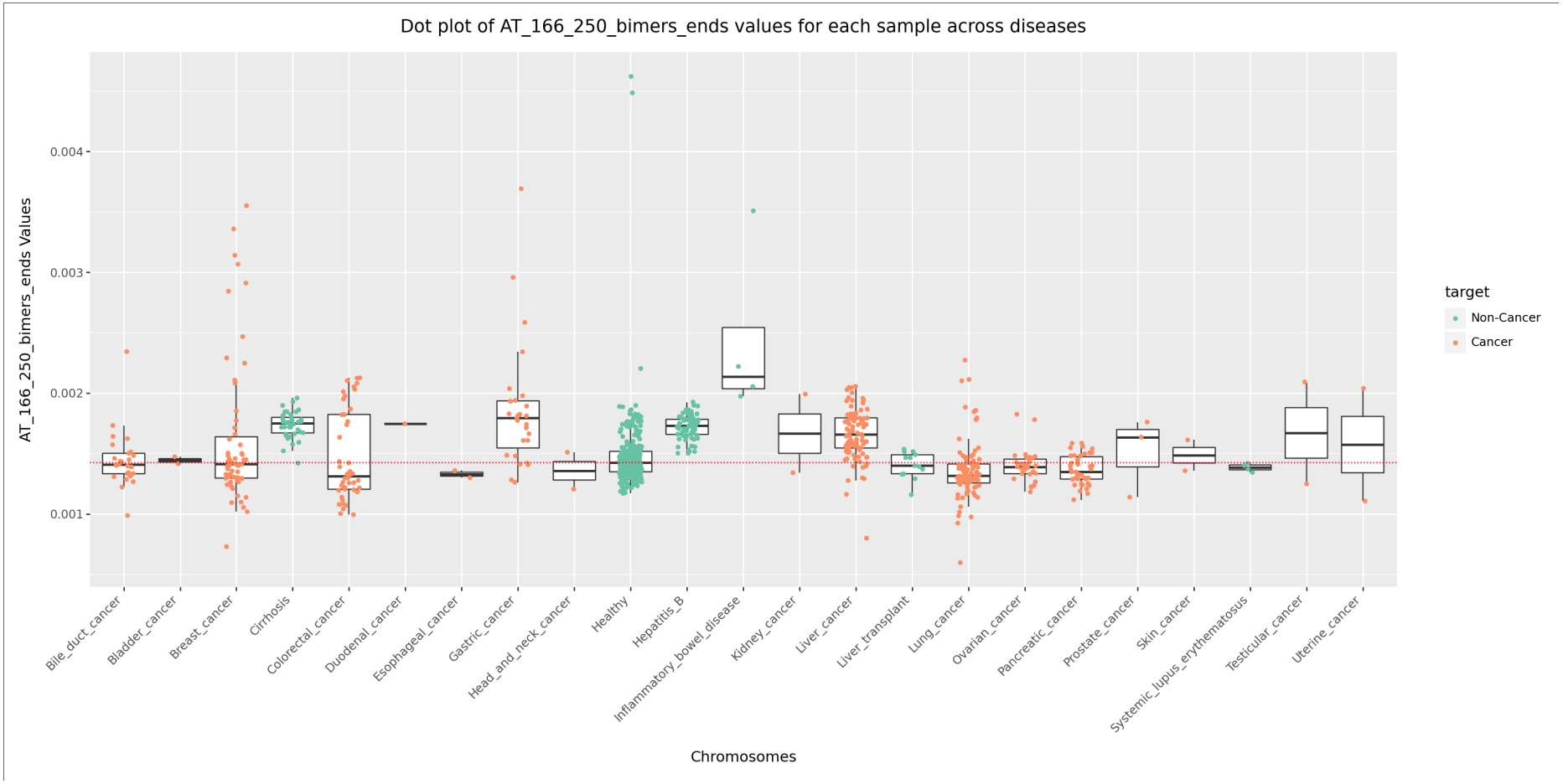
Feature Normalization Process

Key Steps:

1. **Window Sampling:**
    - Process **5 million windows** on each chromosome for accurate feature representation.
2. **Normalization:**
    - Apply **feature normalization** to prepare data for **machine learning** models.
3. **Data Readiness:**
    - Ensure features are **scaled** and **standardized** for better model performance.


Dot plot of med values for each sample across diseases
med Values
180
170
160
150
140
Bile_duct_cancer
Bladder_cancer
Breast_cancer
Cirrhosis
Colorectal_cancer
Duodenal_cancer
Esophageal_cancer
Gastric_cancer
Head_and_neck_cancer
Healthy
Hepatitis_B
Inflammatory_bowel_disease
Kidney_cancer
Liver_cancer
Liver_transplant
Lung_cancer
Ovarian_cancer
Pancreatic_cancer
Prostate_cancer
Skin_cancer
Systemic_lupus_erythematosus
Testicular_cancer
Uterine_cancer
Chromosomes
target
Non-Cancer
Cancer


Dot plot of TG_166_250_bimers_starts values for each sample across diseases
TG_166_250_bimers_starts Values
0.0075
0.0050
0.0025
0.0000
Bile_duct_cancer
Bladder_cancer
Breast_cancer
Cirrhosis
Colorectal_cancer
Duodenal_cancer
Esophageal_cancer
Gastric_cancer
Head_and_neck_cancer
Healthy
Hepatitis_B
Inflammatory_bowel_disease
Kidney_cancer
Liver_cancer
Liver_transplant
Lung_cancer
Ovarian_cancer
Pancreatic_cancer
Prostate_cancer
Skin_cancer
Systemic_lupus_erythematosus
Testicular_cancer
Uterine_cancer
Chromosomes
target
Non-Cancer
Cancer


Dot plot of AT_166_250_bimers_ends values for each sample across diseases
AT_166_250_bimers_ends Values
0.004
0.003
0.002
0.001
Bile_duct_cancer
Bladder_cancer
Breast_cancer
Cirrhosis
Colorectal_cancer
Duodenal_cancer
Esophageal_cancer
Gastric_cancer
Head_and_neck_cancer
Healthy
Hepatitis_B
Inflammatory_bowel_disease
Kidney_cancer
Liver_cancer
Liver_transplant
Lung_cancer
Ovarian_cancer
Pancreatic_cancer
Prostate_cancer
Skin_cancer
Systemic_lupus_erythematosus
Testicular_cancer
Uterine_cancer
Chromosomes
target
Non-Cancer
Cancer

Feature Selection Process

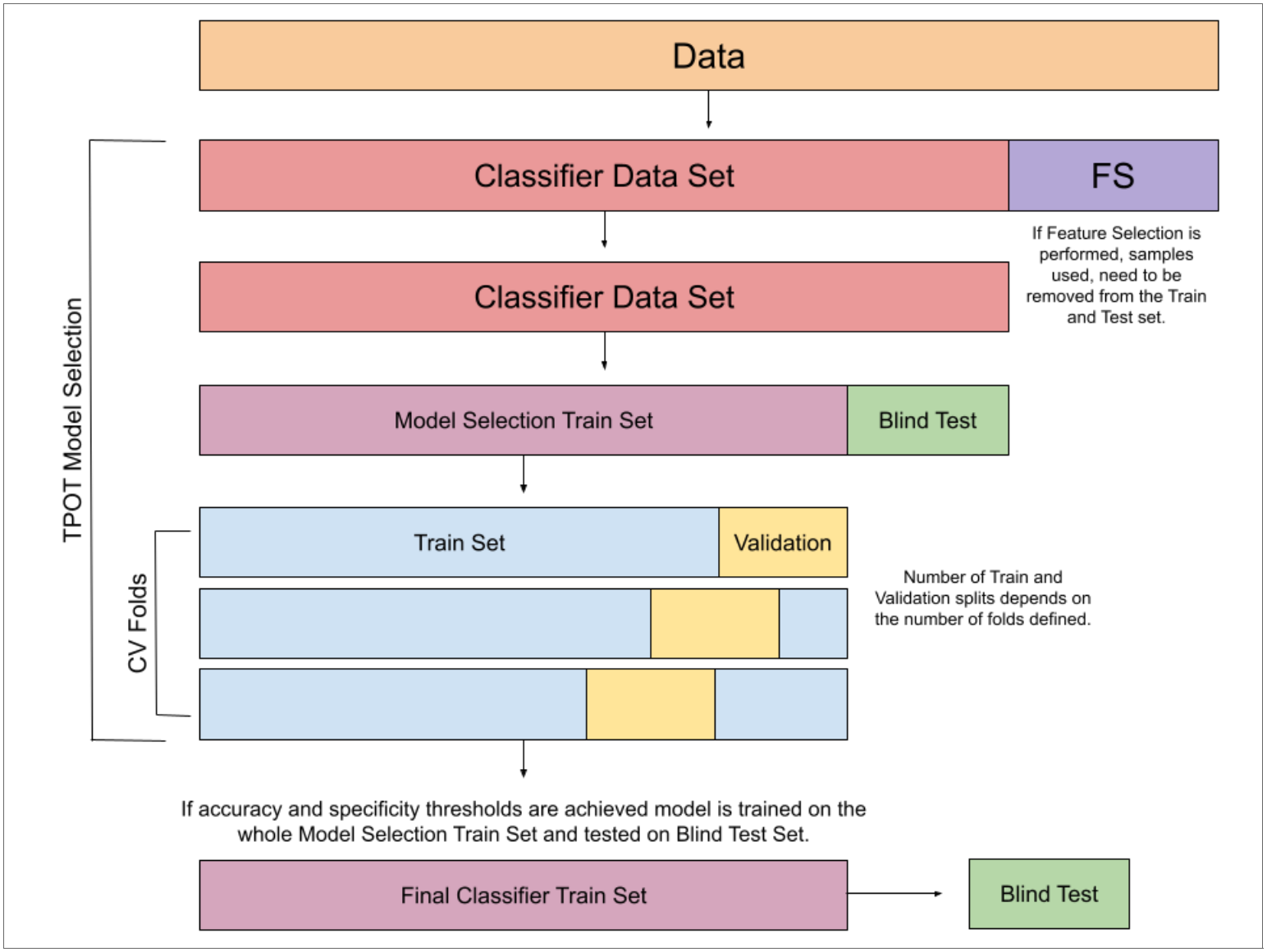
Key Steps:

1. **Sample Selection:**
    - Use **5%** of samples (42 out of 828) for feature selection to avoid overfitting.
2. **Significance Testing:**
    - Apply the **Mann-Whitney U Test** to remove **non-significant features** ($p > 0.05$).
3. **Correlation Handling:**
    - **Group correlated features** (threshold: $|r| > 0.8$) and retain the **most significant feature** from each group.

Result:

- **Selected Features: 78 Features** out of **76,801 Features** (0.1% selected)

TPOT


Data
Classifier Data Set
FS
If Feature Selection is performed, samples used, need to be removed from the Train and Test set.
Classifier Data Set
Model Selection Train Set
Blind Test
TPOT Model Selection
CV Folds
Train Set
Validation
Number of Train and Validation splits depends on the number of folds defined.
If accuracy and specificity thresholds are achieved model is trained on the whole Model Selection Train Set and tested on Blind Test Set.
Final Classifier Train Set
Blind Test

Classifier Performance on Training Data

Best Classifiers After Training

Classifier	HP_count	Accuracy	Sensitivity	Specificity	Weighted_Score
BernoulliNB	475	0.581818	0.240984	0.928333	0.652825
ExtraTreesClassifier	55	0.788430	0.660656	0.918333	0.815049
RandomForestClassifier	228	0.814876	0.762295	0.868333	0.825830
XGBClassifier	459	0.806612	0.798361	0.815000	0.808331

Selected Classifiers for Full Training

Classifier	Count
RandomForestClassifier	151
XGBClassifier	8

Key Insights:

- **Best Performers:** `RandomForestClassifier` and `XGBClassifier`
- **Next Step:** Evaluate the selected classifiers on the **blind test set** for performance validation.

Feature Importance for Training Data

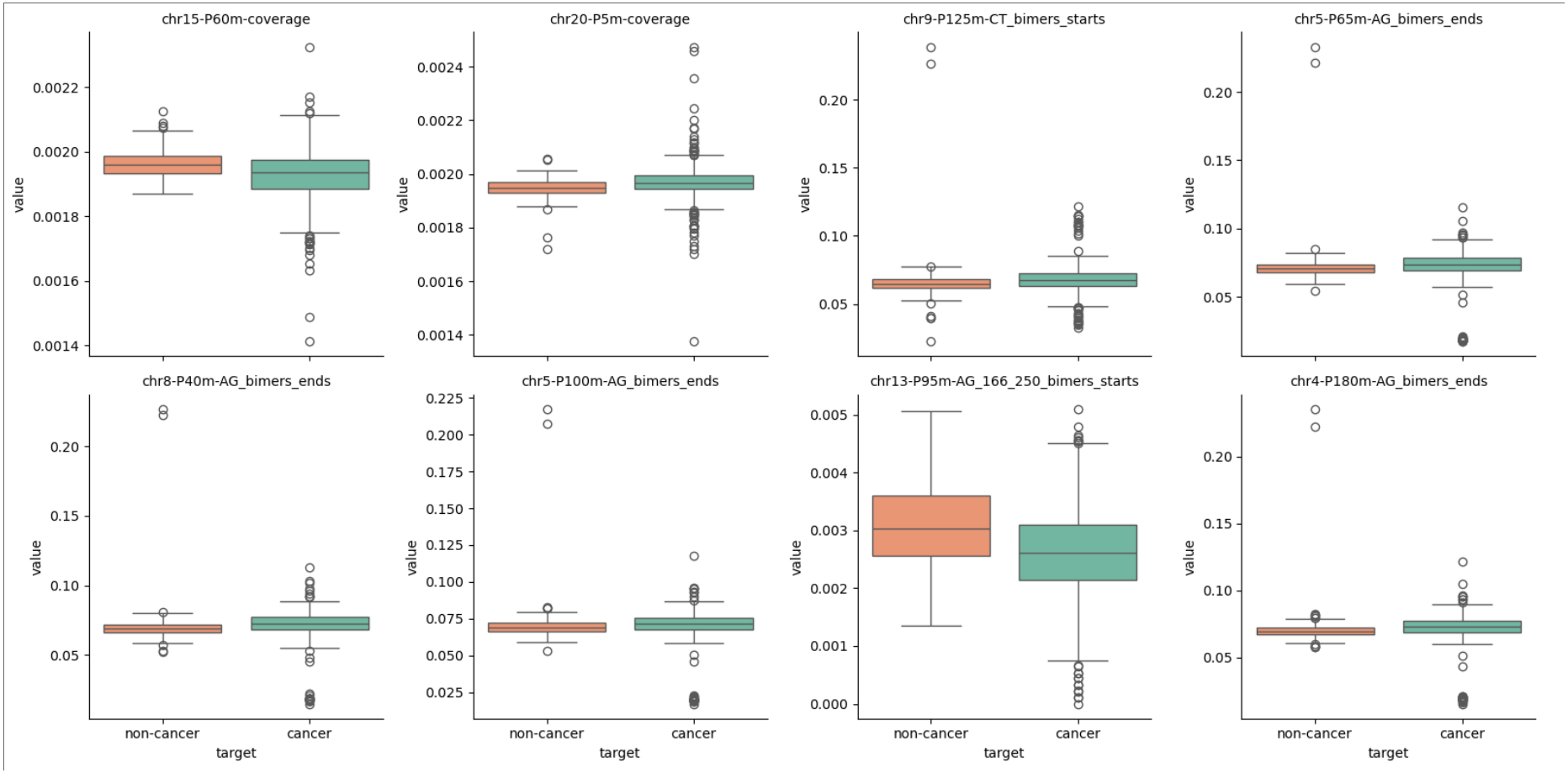
Overlap of Top 10 Important Features by Classifier

Feature	RandomForestClassifier_228	XGBClassifier_459
chr15-P60m-coverage	1	1
chr20-P5m-coverage	2	6
chr9-P125m-CT_bimers_starts	3	4
chr5-P65m-AG_bimers_ends	4	2
chr8-P40m-AG_bimers_ends	5	8
chr5-P100m-AG_bimers_ends	7	5
chr13-P95m-AG_166_250_bimers_starts	8	9
chr4-P180m-AG_bimers_ends	10	3

Key Insights:

- **Top Feature for Both Models:** `chr15-P60m-coverage`
- **Different Feature Rankings:** Each classifier relies on slightly different features.
- **Outlook:** Analyze **biological significance** of top-ranked features.

Visualizing Top 10 Important Features



Classifier Performance on Blind Test Set

Results Overview:

The following classifiers achieved perfect performance in predicting blind test set samples.

Classifier	Accuracy	Sensitivity	Specificity	Confusion_Matrix
RandomForestClassifier_228	1.000000	1.000000	1.000000	[[81 0] [0 81]]
XGBClassifier_459	1.000000	1.000000	1.000000	[[81 0] [0 81]]

Important Note:

- **Testing on a larger sample size is needed** to obtain more **realistic accuracy values**.

Internship Takeaways

What I Learned During My Internship

Technical & Analytical Skills:

- **Python Development:** Advanced proficiency in Python coding and pandas for data analysis.
- **Linux Environment:** Comfortable working in Linux-based systems for data processing.
- **NGS Pipeline Development:** Experience in next-generation sequencing (NGS) workflow design and implementation.
- **Fragmentomics & Feature Engineering:** Expertise in fragmentomics, feature extraction, and feature selection.
- **Machine Learning & Model Development:** Applied ML techniques using TPOT and scikit-learn for predictive modeling.

Internship Takeaways

What I Learned During My Internship

Business & Professional Insights:

- **Start-Up Environment:** Gained insights into start-up dynamics, project management, and business development in a biotech setting.

Takeaway:
This internship has significantly enhanced both my **technical** and **professional** skills, preparing me for future challenges in **computational biology**, **bioinformatics**, and **machine learning**.