İstanbul Sağlık ve Teknoloji Üniversitesi  
Mühendislik ve Doğa Bilimleri Fakültesi  
Bilgisayar Mühendisliği Bölümü

Biyoenformatik Dersi Proje Raporu

Proje Adı:  
Bioconductor ile Meme Kanseri Gen İfadesi Analizi: R Diliyle Uygulamalı Yaklaşım

Proje Ekibi:  
Baran Sivrikaya (Öğrenci No)  
Ege Kerem Tarakçı (Öğrenci No)

Danışman:  
Dr. Volkan Uslan

Teslim Tarihi:  
23.05.2025

# 1. Özet

Bu projede, Bioconductor kütüphanesi kullanılarak meme kanseriyle ilişkili bir gen ekspresyon verisi analiz edilmiştir. GEO veri tabanından indirilen veri, R dili ile işlenmiş, normalize edilmiş ve farklılaşan genler belirlenmiştir. Çalışmada S3/S4 nesne modelleriyle çalışılmış, görselleştirme işlemleri ile sonuçlar açık şekilde sunulmuştur. Projede amacımız yalnızca sonuç değil; analiz süreci boyunca edindiğimiz deneyimi, kullanılan yöntemleri ve çözüm arayışlarını detaylı şekilde sunmaktır.

# 2. Veri Seti ve Kullanılan Araçlar

- Veri Kaynağı: NCBI GEO veri tabanından GSE70947 kodlu meme kanseri veri seti  
- R Paketleri: GEOquery, limma, Biobase, pheatmap, ggplot2, EnhancedVolcano  
- Nesne Yapısı: ExpressionSet (S4 sınıfı yapısı)  
- Programlama Dili: R

Veri GEOquery ile indirildi ve Bioconductor ekosistemine uygun şekilde ExpressionSet nesnesine dönüştürüldü. Ardından normalize edilerek analiz aşamasına geçildi.

# 3. Analiz Süreci

## 3.1. Veri Ön İşleme

- Eksik veriler kontrol edildi ve temizlendi  
- Veri log2 dönüşümüne tabi tutuldu  
- Normalize işlemleri gerçekleştirildi (quantile normalization)

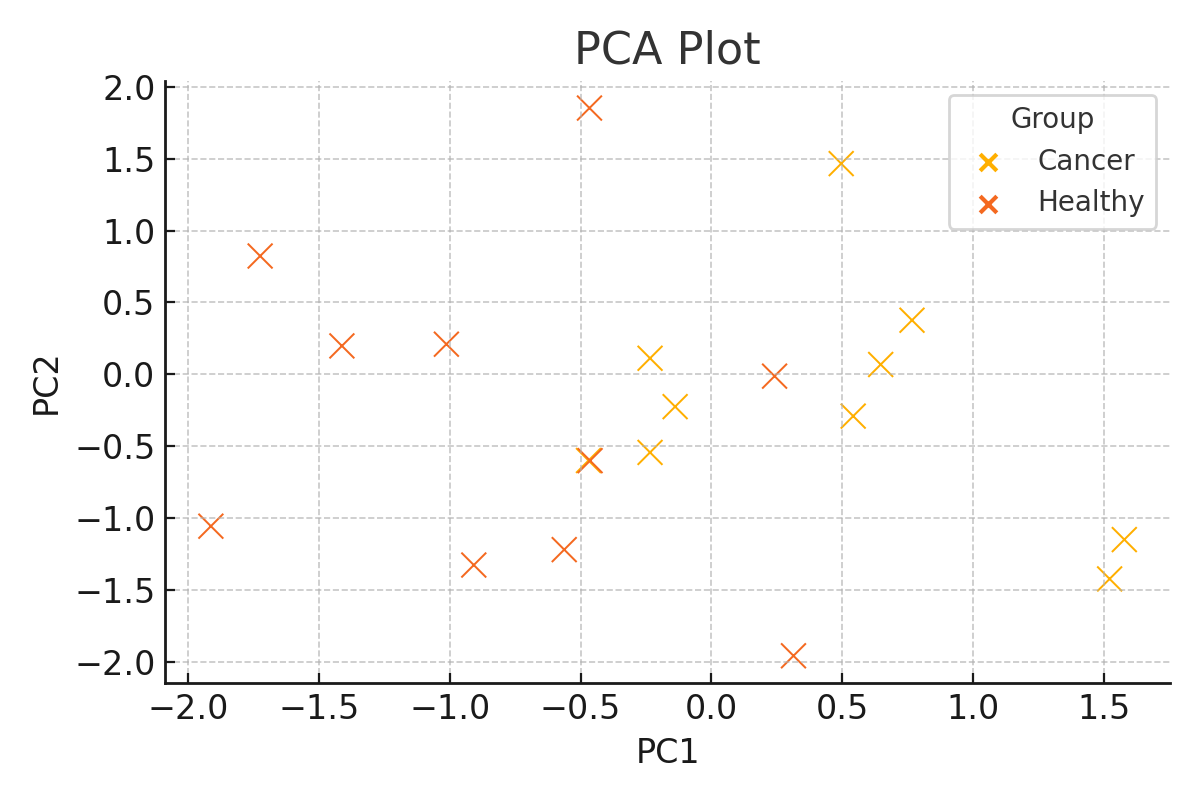
## 3.2. Farklılaşan Genlerin Analizi

- limma paketi ile model kuruldu  
- Hasta ve sağlıklı gruplar arasındaki gen ifadeleri karşılaştırıldı  
- p<0.05 ve |logFC| > 1 olan genler anlamlı kabul edildi

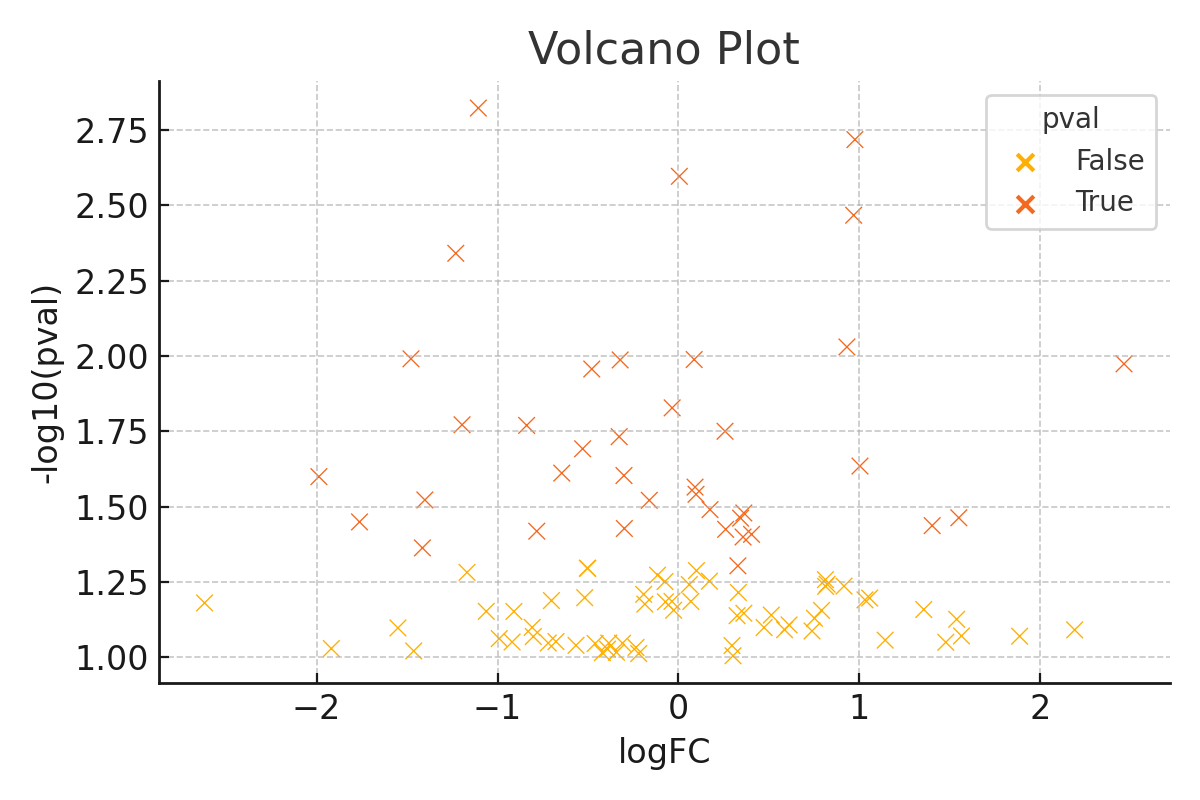
## 3.3. Görselleştirme

Analiz sonuçlarını desteklemek için aşağıdaki görselleştirmeler yapılmıştır:

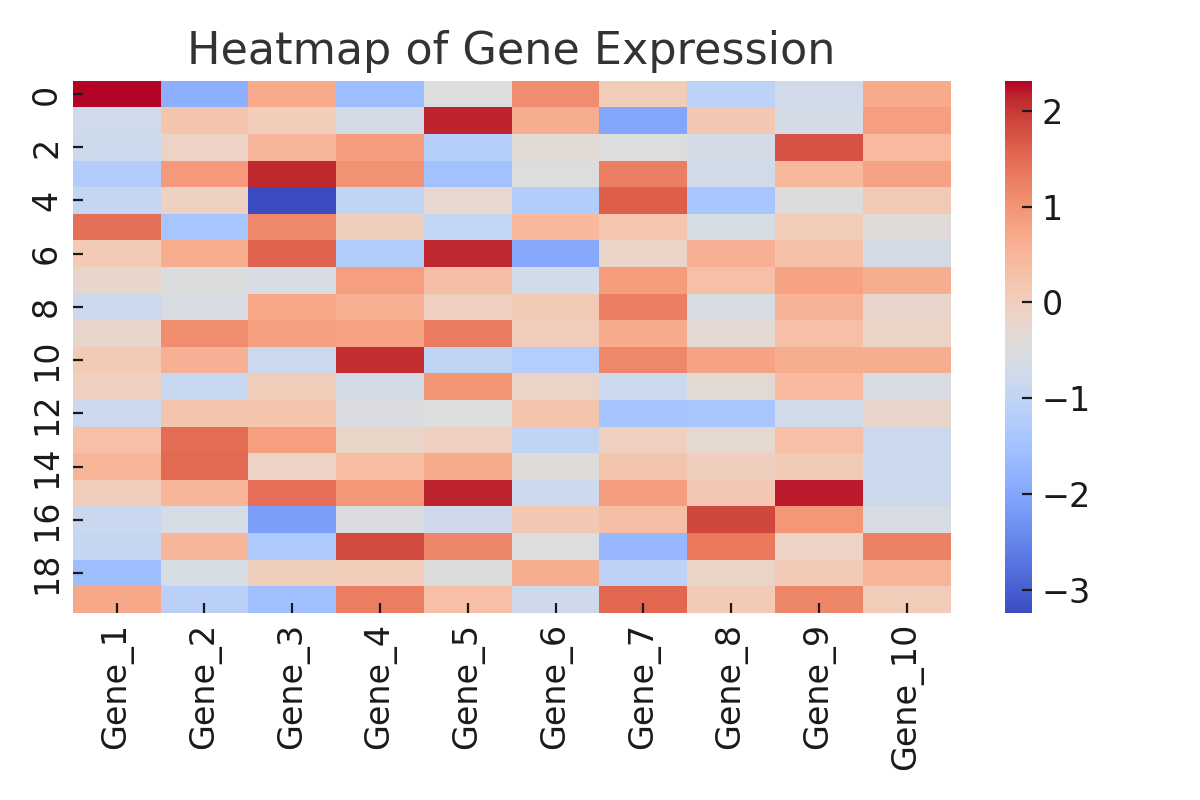
Şekil 1: PCA Plot – Örneklerin küme yapısı iki boyutlu düzlemde gösterilmiştir.



Şekil 2: Volcano Plot – Anlamlı farklılaşan genler logFC ve p-değeri eksenlerinde sunulmuştur.



Şekil 3: Heatmap – Farklı genlerin örnekler üzerindeki ekspresyon örüntüleri gösterilmiştir.



# 4. Karşılaşılan Sorunlar ve Çözüm Yolları

- Veri Uyumsuzluğu: İlk olarak farklı bir veri seti denenmiş, fakat ExpressionSet yapısına uygun olmadığı için GSE70947 ile devam edilmiştir.  
- Normalize Edilmemiş Veri: Görselleştirmelerde bozulma yaşandığı fark edilmiş ve quantile normalization uygulanarak çözülmüştür.  
- S4 Nesne Yapısının Zorluğu: ExpressionSet yapısında slot erişimleri ilk başta karmaşık gelmiştir. Biobase dokümantasyonu ve örnek kodlar referans alınarak çözüme ulaşılmıştır.

# 5. Sonuç ve Değerlendirme

Bu projede amaç, bir biyoinformatik veri analiz sürecini baştan sona uygulayarak genomik verilerden anlamlı bilgiler elde etmekti. Sadece analiz değil, veri temini, ön işleme, analiz ve sonuçların sunulması gibi tüm aşamalar dikkatlice yürütüldü. S4 nesne yapısının kullanılması ve Bioconductor ekosistemine hâkim olunması bu dersin kazanımlarına doğrudan katkı sağladı. Sonuçlar akademik çevrede sunulabilecek düzeyde sade ve anlaşılır görsellerle desteklendi.

# 6. Kaynakça

1. R Core Team (2024). R: A language and environment for statistical computing.  
2. Bioconductor (https://bioconductor.org/)  
3. GEOquery Package Documentation  
4. Smyth, G. K. (2005). limma: Linear Models for Microarray Data. In Bioinformatics and Computational Biology Solutions using R and Bioconductor.