A red text on a white background

Description automatically generated

**MÜHENDİSLİK VE DOĞA BİLİMLERİ FAKÜLTESİ**

**BYM454 - BİYOENFORMATİK DERSİ PROJE ÖDEVİ**

**R ile Uygulamalı Veri Analizi: COVID-19 ve Sağlıklı Bireyler Arasında Gen İfadesi Farklılıklarının İncelenmesi**

**HAZIRLAYANLAR**

MERT TOSUN - 210601027

BEYTULLAH ALKAN - 210601025

RESUL EKREM ÖZDEMİR – 210601019

**BİLGİSAYAR MÜHENDİSLİĞİ**

TESLİM TARİHİ

16.05.2025

**PROJENİN AMACI**

Bu projenin amacı, COVID-19 hastalarının gen ekspresyon profillerini sağlıklı bireylerle karşılaştırarak, hastalığa özgü genetik düzeydeki değişiklikleri belirlemektir. Gen ekspresyonu, bir genin hücre içinde ne kadar aktif olduğunu gösterir ve RNA-Seq yöntemi sayesinde bu aktivite sayısal olarak ölçülebilir.

Projemiz kapsamında, NCBI-GEO veri tabanında yer alan GSE196822 numaralı veri seti kullanılmıştır. Bu veri setinde COVID-19 hastalarından ve sağlıklı bireylerden alınan kan örneklerine ait RNA-Seq sayım verileri yer almaktadır. Bu veriler kullanılarak, her iki grup arasında anlamlı şekilde farklı ifade edilen (diferansiyel olarak ifade edilen) genler tespit edilmiştir.

Amaç istatistiksel olarak anlamlı düzeyde farklılık gösteren genleri belirleyerek, COVID-19'un biyolojik etkilerini gen düzeyinde daha iyi anlamak ve potansiyel biyobelirteç adaylarını ortaya koymaktır.

**YÖNTEMLER VE MATERYALLER**

**1. Kullanılan Veri Seti**

Bu çalışmada, NCBI GEO (Gene Expression Omnibus) veri tabanında yer alan GSE196822 numaralı veri seti kullanılmıştır. Veri seti, COVID-19 hastaları ile sağlıklı bireylerden alınan tam kan örneklerine ait RNA-Seq gen ekspresyon verilerini içermektedir. Veri, insan (Homo sapiens) türüne aittir.  
Veri setinde:

* 40 COVID-19 hastasına ait örnek,
* 9 sağlıklı bireye ait örnek bulunmaktadır.
* Bu örneklerden elde edilen sayım verileri analizde kullanılmıştır.

**2.  Kullanılan R Paketleri ve Kütüphaneler**

Proje boyunca kullanılan temel R paketleri şunlardır:

* **edgeR**: RNA-Seq sayım verilerinin normalize edilmesi ve voom fonksiyonunun kullanılması
* **limma**: Lineer modelleme ve diferansiyel gen analizi
* **ggplot2**: Volcano plot gibi görselleştirmeler
* **BiocManager**: Bioconductor paketlerinin kurulumu için kullanıldı.

**GİTHUB REPO LİNKİ: https://github.com/mertt1010/Covid19GeneAnalysisWithR**

**3. Veri Yapısı**

Analizde kullanılan veri, RNA-Seq teknolojisi ile elde edilen ve gen ekspresyon seviyelerini temsil eden bir ham sayım matrisi (raw count matrix) formatındadır. Veri seti, GSE196822 numaralı çalışmadan alınmıştır ve her bir hücredeki değer, bir genin ilgili örnekte kaç kez okunduğunu (read count) göstermektedir.

Veri yapısı şu şekildedir:

* Satırlar: Ensembl gen kimlikleri (örnek: ENSG00000000419), her biri bir geni temsil eder.
* Sütunlar: Bireylere ait örneklerdir. Örneğin:
  + BK\_Cov\_41 → COVID-19 hastasına ait örnek
  + BK\_HC\_5I20 → Sağlıklı bireye ait örnek
* Hücre değerleri: Her bir genin, ilgili örnekteki ifade düzeyini belirtir. Bu değerler, sayım (count) şeklindedir ve RNA-Seq analizinde genin aktifliği hakkında bilgi verir.

Örneğin aşağıdaki satırda:

| **Gen ID** | **BK\_Cov\_41** | **BK\_HC\_5I20** |
| --- | --- | --- |
| ENSG00000000419 | 164 | 200 |

“ENSG00000000419” adlı genin,

* COVID hastasında (BK\_Cov\_41) 164 kez,
* Sağlıklı bireyde (BK\_HC\_5I20) 200 kez ifade edildiği (okunduğu) anlaşılmaktadır.

Bu yapı, diferansiyel gen ekspresyon analizine doğrudan uygun olup istatistiksel analizlerin temelini oluşturur. Ek olarak, örneklerin ait olduğu gruplar (COVID vs Healthy) ayrı bir metadata çerçevesi olan coldata ile tanımlanmıştır.

**ANALİZ SÜRECİ**

Bu projede COVID-19 hastaları ile sağlıklı bireyler arasında gen ekspresyon farklılıklarını belirlemek amacıyla RNA-Seq verisi üzerinde diferansiyel gen ekspresyon analizi yapılmıştır.

1. **Sayım verisinin yüklenmesi**

GEO'dan indirilen .csv.gz formatındaki ham sayım verisi içeri aktarılmıştır. Veride satırlar genleri (Ensembl ID), sütunlar ise örnekleri (hastalar) temsil etmektedir.

1. **Örneklerin gruplandırılması**

Örnek isimlerine göre COVID ve Healthy grupları otomatik olarak sınıflandırılmıştır.

1. **Gerekli paketlerin kurulumu ve yüklenmesi**

Analiz için edgeR ve limma paketleri kullanılmıştır.

1. **Sayım verisinin DGEList formatına dönüştürülmesi**

edgeR paketi ile sayım verisi analiz formatına çevrilmiştir.

1. **Varyans stabilizasyonu ve normalize edilmesi (voom)**

voom fonksiyonu ile log dönüşüm uygulanmış ve ağırlıklandırmalı normalize veri elde edilmiştir.

1. **Model matrisi oluşturulması**

Gruplar arasında karşılaştırma yapabilmek için model matrisi tanımlanmıştır.

1. **Lineer modelleme ve eBayes uygulanması**

Limma paketi ile genler arasında istatistiksel karşılaştırma yapılmıştır.

1. **Tüm genler için sonuçların çıkarılması**

İstatistiksel analiz sonucu her gen için log2 fold change, p-value ve adjusted p-value hesaplanmıştır.

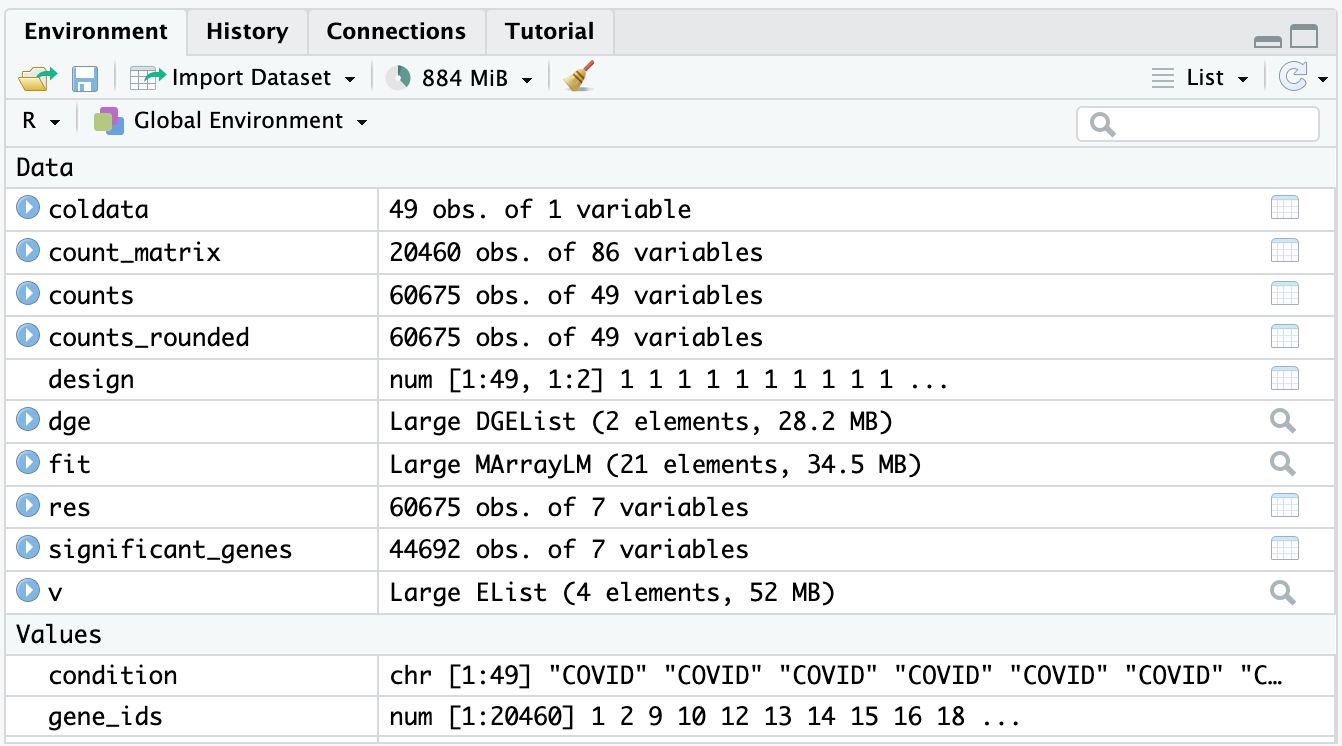
1. **Volcano plot oluşturulması**

Anlamlı genlerin görsel olarak değerlendirilmesi için volcano plot hazırlanmıştır.

1. **Anlamlı genlerin filtrelenmesi ve kaydedilmesi**

|log2FC| > 1 ve adj.P.Val < 0.05 kriterlerine göre anlamlı genler belirlenmiş ve CSV dosyasına kaydedilmiştir.

Bu adımlar sonucunda toplam 44.692 genin anlamlı şekilde farklı ifade edildiği belirlenmiştir. Bunlardan 44.264'ü COVID grubunda yukarı regüle olmuş, 428’i aşağı regüle olmuştur.

\*şekil-1

R ortamında analiz sırasında oluşturulan veri nesneleri gösterilmektedir. Sayım verisi (counts), örnek bilgileri (coldata), model matrisi (design), normalize edilmiş veri (v), lineer model sonuçları (fit) ve anlamlı bulunan genler (significant\_genes) bu ortamda kaydedilmiştir. Toplam 49 örnek ve 60.675 gen kullanılmış, sonuçta 44.692 anlamlı gen elde edilmiştir.

metin, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.\*şekil-2

Volcano plot grafiği, analiz sonucunda elde edilen genlerin log2 fold change ve -log10(p-value) değerlerine göre dağılımını göstermektedir. Kırmızı kesikli çizgiler biyolojik anlamlılık sınırını (|log2FC| > 1), mavi çizgi ise istatistiksel anlamlılık sınırını (p < 0.05) temsil etmektedir. Sağ üstteki noktalar COVID grubunda yüksek oranda artmış genleri (upregulated), sol üstteki noktalar ise baskılanmış genleri (downregulated) ifade etmektedir.

**İstatistiksel Analiz Yöntemi**

Bu projede, COVID-19 hastaları ile sağlıklı bireyler arasında gen ekspresyon düzeylerini karşılaştırmak amacıyla **diferansiyel gen ekspresyonu analizi** gerçekleştirilmiştir. Analiz sürecinde aşağıdaki istatistiksel yöntemler kullanılmıştır.

**Diferansiyel Gen Ekspresyonu Analizi Nedir?**

Bu analiz şunu anlamaya çalışır:

“Bir gen, **bir grupta (projemizde COVID-19 hastalarında)** diğer gruba (sağlıklı bireyler) göre **daha fazla mı ya da daha az mı çalışıyor?**”

RNA-Seq gibi teknolojiler sayesinde her genin hücrede ne kadar aktif olduğunu ölçebiliyoruz. Bu aktiflik, genin “ifade edilme (ekspresyon)” düzeyidir

Projemizde:

* **COVID-19 hastaları ile sağlıklı bireyler karşılaştırıldı.**
* Her gen için istatistiksel test yapıldı (limma, eBayes).
* **Log2 Fold Change**: Genin ne kadar arttığını/azaldığını gösterdi.
* **Adj. P-değeri < 0.05** olan genler **anlamlı** kabul edildi.
* Adj.P değeri, yani düzeltilmiş p-değeri, yapılan çok sayıda testin sonucunda ortaya çıkabilecek yanlış anlamlılıkları düzeltmek için kullanılan değerdir.

| **Kavram** | **Anlamı** |
| --- | --- |
| **p-değeri** | Genin gruplar arasında farklılık gösterme olasılığı |
| **adj.P değeri** | Bu p-değerinin, **çoklu test düzeltmesi** sonrası daha güvenilir hâli |
| **adj.P<0.05** | Bu genin, çok sayıda test yapılmasına rağmen hâlâ anlamlı olduğu kabul edilir |

**1. Lineer Modelleme – limma Paketi ile**

Her bir gen için, gruplar (COVID ve Healthy) arasında anlamlı bir fark olup olmadığı limma paketinin lmFit() fonksiyonu ile test edilmiştir. Bu adımda gen ifadesi ile grup bilgisi arasındaki ilişki lineer model ile modellenmiştir:

**2. Empirik Bayes Düzeltmesi – eBayes() Fonksiyonu**

Lineer model sonuçları, eBayes() fonksiyonu ile işlenmiştir. Bu yöntem sayesinde genler arası varyans farklılıkları stabilize edilmiş ve daha güvenilir p-değerleri elde edilmiştir

**3. Anlamlı Genlerin Belirlenmesi – topTable() Fonksiyonu**

Modelleme sonuçları, topTable() fonksiyonu ile özetlenmiştir. Her bir gen için aşağıdaki değerler hesaplanmıştır:

* **Log2 Fold Change (log2FC):** Genin ifade düzeyindeki değişimin büyüklüğü
* **P-value:** Farkın istatistiksel anlamlılığı
* **Adj.P.Val (FDR):** Çoklu test düzeltmesi yapılmış p-değeri

**Yani kısaca bir genin anlamlı olup olmadığını nasıl anlıyoruz?**

**Projemizde , her gen için iki şeye baktık:**

1. **log2 Fold Change (log2FC):**

Genin ifade düzeyinin iki grup (COVID vs Healthy) arasında ne kadar değiştiğini gösterir.

* Pozitif ise → COVID grubunda gen daha **aktif**.
* Negatif ise → COVID grubunda gen daha **baskılanmış**.
* |log2FC| > 1 → Değişim **biyolojik olarak anlamlı** sayılır.

1. **adj.P değeri (düzeltilmiş p-değeri):**

Genin istatistiksel olarak gerçekten farklı çalışıp çalışmadığını gösterir.

* adj.P.Val < 0.05 ise → istatistiksel olarak anlamlı kabul edilir.

Bir genin anlamlı olduğunu **hem ne kadar değiştiğine** hem de bu **değişimin istatistiksel olarak güvenilir olup olmadığına** bakarak anlıyoruz. Biz bu projede log2FC > 1 ve adj.P.Val < 0.05 olan genleri anlamlı kabul ettik. Bu iki şartı sağlayan genler hem biyolojik hem de istatistiksel olarak gerçekten farklı çalışıyor demektir.

Gen neye göre değişiyor? Ne kadar değişti dediğimizde, genin neyinde değişiklik oluyor?

Genin DNA dizisi değişmiyor.

Burada genetik mutasyon değil, o genin hücre içinde ne kadar çalıştığı (yani ne kadar aktif olduğu) ölçülüyor.

Peki Ne Ölçülüyor?

**mRNA miktarı ölçülüyor.**

RNA-Seq tekniğiyle, her genin hücrede **kaç kopya mRNA ürettiği** tespit edilir. Bu sayı, o genin **ifade düzeyini** gösterir.

**Örneğin:**

* COVID hastasının kanında bir gen **100 kez** ifade edilmiş (100 mRNA kopyası),
* Sağlıklı kişide aynı gen **25 kez** ifade edilmiş.

Bu durumda gen, COVID hastasında **4 kat daha aktif** çalışıyor demektir.  
Bu fark log2 dönüşümüyle hesaplanır → log2(100/25) = **+2**

Ne kadar değişti?” dediğimiz şey:

**İki grup arasında, genin mRNA üretim miktarındaki fark**  
Ölçülen değer: **log2 Fold Change (log2FC)**

Projemizde **log2FC (ifade artışı) değerine göre en yüksek değişim gösteren ilk 5 gen :**

**metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.**

* log2FC > 4 gibi değerler, gen ifadesinin 16 kat ve üzeri arttığını gösterir.
* Adj.P.Val < 0.001 seviyeleri, bu farkların tesadüfi olmadığını ve istatistiksel olarak çok güçlü olduğunu gösterir.
* Bu genler muhtemelen COVID-19'a karşı bağışıklık yanıtında veya hastalığın oluşturduğu stres, inflamasyon ve savunma mekanizmalarında rol alıyor olabilir.

Analiz sonucunda, COVID hastalarında ifadesi en çok artan genler arasında **ENSG00000263244** ve **ENSG00000205609** gibi genler yer aldı. Bu genler, sağlıklı bireylere göre yaklaşık 16 kata kadar daha fazla çalışıyordu. Bu fark istatistiksel olarak da anlamlıydı. Bu genler muhtemelen COVID-19’un vücutta oluşturduğu bağışıklık veya stres yanıtıyla ilişkilidir.

**SONUÇ**

Bu projede, COVID-19 hastalarının genetik düzeyde sağlıklı bireylerden nasıl farklılaştığını anlamak amacıyla gen ekspresyon analizi yapılmıştır. Analizde kullanılan veri seti, gerçek hasta ve sağlıklı bireylerden alınan kan örneklerinden elde edilen RNA-Seq verilerine dayanmaktadır.

R programında yapılan analizlerde, toplam 60.675 gen incelenmiş ve bunlardan 44.692 genin COVID-19 hastalarında anlamlı düzeyde farklı çalıştığı (farklı ifade edildiği) belirlenmiştir. Bu genlerden büyük çoğunluğu hastalarda daha fazla aktif hâle gelirken, bazıları daha az çalışır hâle gelmiştir. Bu durum, COVID-19’un vücutta gen düzeyinde önemli etkiler yarattığını göstermektedir.

Şekil 1’de (RStudio Global Environment ekranı) analiz süreci boyunca oluşturulan veriler ve işlem adımları görülmektedir. Şekil 2’de ise (Volcano Plot grafiği), hangi genlerin COVID-19 hastalarında daha fazla ya da daha az çalıştığını görsel olarak ifade eden grafik yer almaktadır. Bu grafik sayesinde genler arasındaki farklar açıkça görülebilmektedir.

Sonuç olarak, COVID-19 hastalığı vücutta genlerin çalışma düzenini değiştirmekte ve bu değişiklikler, hastalığın etkilerinin anlaşılmasında bize önemli ipuçları vermektedir. Elde edilen gen listeleri, ileride yapılacak çalışmalarda hastalığın seyrini ve etkilerini incelemek için değerli bir kaynak olabilir.