ZONGULDAK BÜLENT ECEVİT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YENİ NESİL RNA SEKANSLAMA VE MİKRODİZİN VERİLERİNİN ANALİZİ İLE KANSERDE TRANSKRİPTOMİK BİLGİLERİN ELDESİ

MOLEKULER BİYOLOJİ ANABİLİM DALI

YÜKSEK LİSANS TEZİ

EBUBEKİR AYHAN

AĞUSTOS 2020

ZONGULDAK BÜLENT ECEVİT ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

YENİ NESİL RNA SEKANSLAMA VE MİKRODİZİN VERİLERİNİN ANALİZİ İLE KANSERDE TRANSKRİPTOMİK BİLGİLERİN ELDESİ

MOLEKÜLER BİYOLOJİ ANABİLİM DALI

YÜKSEK LİSANS TEZİ

EBUBEKİR AYHAN

DANIŞMAN: DR. ÖĞR. ÜYESİ KEREM MERT ŞENSES

ZONGULDAK

Ağustos 2020

KABUL:

Ebubekir AYHAN tarafından hazırlanan "Yeni Nesil RNA Sekanslama ve Mikrodizin Verilerinin Analizi ile Kanserde Transkriptomik Bilgilerin Eldesi" başlıklı bu çalışma jürimiz tarafından değerlendirilerek Zonguldak Bülent Ecevit Üniversitesi, Fen Bilimleri Enstitüsü, Moleküler Biyoloji Anabilim Dalında Yüksek Lisans Tezi olarak oybirliğiyle kabul edilmiştir.24/08/2020

Danışman: Dr. Öğr. Üyesi Kerem Mert ŞENSES .....
Zonguldak Bülent Ecevit Üniversitesi, Fen Edebiyat Fakültesi, Moleküler Biyoloji ve Genetik Bölümü

Üye: Dr. Öğr. Üyesi Tolga ACUN .....
Zonguldak Bülent Ecevit Üniversitesi, Fen Edebiyat Fakültesi, Moleküler Biyoloji ve Genetik Bölümü

Üye: Dr. Öğr. Üyesi Can TÜRK .....
Lokman Hekim Üniversitesi, Tıp Fakültesi, Tıbbi Mikrobiyoloji ABD

ONAY:

Yukarıdaki imzaların, adı geçen öğretim üyelerine ait olduğunu onaylarım. ..../..../20....

Prof. Dr. Ahmet ÖZARSLAN
Fen Bilimleri Enstitüsü Müdürü

“Bu tezdeki tüm bilgilerin akademik kurallara ve etik ilkelere uygun olarak elde edildiğini ve sunulduğunu; ayrıca bu kuralların ve ilkelerin gerektirdiği şekilde, bu çalışmadan kaynaklanmayan bütün atıfları yaptığımı beyan ederim.”

Ebubekir AYHAN

ÖZET

Yüksek Lisans Tezi

YENİ NESİL RNA SEKANSLAMA VE MİKRODİZİN VERİLERİNİN ANALİZİ İLE KANSERDE TRANSKRİPTOMİK BİLGİLERİN ELDESİ

Ebubekir AYHAN

Zonguldak Bülent Ecevit Üniversitesi
Fen Bilimleri Enstitüsü
Moleküler Biyoloji Anabilim Dalı

Tez Danışmanı: Dr. Öğr. Üyesi Kerem Mert ŞENSES
Ağustos 2020, 66 sayfa

RNA sekanslama yöntemi son yıllarda hızla gelişen yeni nesil sekanslama teknolojilerinin bir çeşididir. RNA sekanslama (RNA-Seq) yönteminde biyolojik materyalin (örneğin kanser hücre hattının) transkribe ettiği gen bölgelerinin transkriptlerinin sekansları elde edilir. Bu sekanslar ya da diğer bir deyişle RNA-Seq veri setleri araştırmacıların kolaylıkla ulaşabileceği internet veri tabanlarında mevcuttur.

Bu çalışmanın konusu, genel çerçevede kanser biyolojisidir. Daha derine inildiğinde ise in silico yollarla ulaşılabilen yeni nesil RNA sekanslama ve gen ifadesi mikrodizin verilerinin biyoinformatik yöntem ve araçlar kullanılarak analiz edilmesiyle kanser hücre hatlarına ait transkriptomik verilerin elde edilmesi olarak tanımlanabilir. Aynı kanser hücre hattının farklı laboratuvarlarda çoğaltılması ve bu kanser hücre hattına ait kontrol gruplarının RNA sekanslama sonucu oluşan RNA-Seq veri setleri arasında gen ifadesi, alternatif ekson kullanımı gibi farklıların olup olmadığı analiz edildi.

ÖZET (devam ediyor)

İnternet veri tabanları kullanılarak çalışmak istediğimiz kanser hücre hatlarına ait RNA-Seq veri setleri indirilip, o veri setine ait sekansların kaliteleri uygun biyoinformatik araçlar kullanılarak denetlenmiştir. İyi sekans okuması kalitesine sahip veriler insan referans genomuna hizalandırıldı.

Hizalama sonucu oluşan dosyalar, görselleştirme aracı kullanılarak kanser hücre hatlarının insan genomuna hangi düzeyde hizalandığı kontrol edildi. Çalışmamızın diğer basamağında hücre hatlarındaki gen ifadesi farklılıkları, alternatif ekson kullanımı analizleri yapıldı.

Sonuç olarak, aynı hücre hattı olmasına rağmen farklı laboratuvarlarda büyütülüp, pasajlanmış olan kanser hücre hatları arasında gen ifadesi ve alternatif ekson kullanımı bakımından ciddi farklar tespit ettik.

Anahtar Kelimeler: RNA Sekanslama, Kanser Hücre Hattı, Kanser Biyolojisi
Bilim Kodu: 401.02.02

ABSTRACT

M. Sc. Thesis

TRANSCRIPTOMIC DATA RETRIEVAL FROM CANCER USING NEXT GENERATION RNA SEQUENCING AND MICROARRAY DATA ANALYSES

Ebubekir AYHAN

Zonguldak Bülent Ecevit University
Graduate School of Natural and Applied Sciences
Department Of Moleculer Biology And Genetıcs

Thesis Advisor: Assist. Prof. Dr. Kerem Mert ŞENSES
August 2020, 66 pages

RNA sequencing method is a kind of new generation sequencing technologies that have been developing rapidly in recent years. In the RNA sequencing (RNA-Seq) method, sequences of the transcripts of the gene regions transcribed by the biological material (e.g. cancer cell line) are obtained. These sequences or in other words, RNA-Seq datasets are available in internet databases that researchers can easily access.

The subject of this study is cancer biology in general framework. When digging deeper, it can be defined as obtaining transcriptomic data of cancer cell lines by analysing the next generation RNA sequencing and gene expression microarray data which can be accessed in silico ways using bioinformatics methods and tools. It was analysed whether there were any differences such as differential gene expression and alternative exon usage among the RNA-Seq datasets formed by generating same cell line in different laboratory and using the control groups of this cancer cell line in RNA sequencing.

ABSTRACT (continued)

RNA-Seq datasets of cancer cell lines that we want to work with using internet databases were downloaded and the quality of the sequences of that dataset were checked using appropriate bioinformatics tools. Data with good sequence reading quality aligned to human reference genome. The files formed as a result of the alignment were checked by using the visualization tool how the cancer cell lines are aligned to the human genome. In the other step of our study, differential gene expression and alternative exon usage analyzes were performed on cell lines.

As a result, we found significant differences in differential gene expression and alternative exon usage between cancer cell lines that were grown and passaged in different laboratories despite being the same cell line.

Keywords: RNA-Sequencing, Cell Lines, Cancer Biology

Science Code: 401.02.02

TEŞEKKÜR

Yüksek lisansımın ilk gününden son gününe kadar her koşulda desteklerini ve zamanını ayıran, bir an olsun bile yardımını esirgemeyen sevgili danışman hocam Dr.Öğr Kerem Mert Şenses'e çok teşekkür ederim.

Lisans hayatımın başlangıcından yüksek lisans hayatımın bitişine kadar her türlü desteği ve yardımlarını esirgemeyen Dr.Öğr Arzu Bahar Erol'a çok teşekkür ederim.

Her zaman güzel bir aileye sahip olduğum için ne kadar şanslı olduğumu hissettiren ve hayatım boyunca attığım her adımda, bu günlere gelmemde maddi ve manevi desteğini esirgemeyen aileme çok teşekkür ederim.

İÇİNDEKİLER

İÇİNDEKİLER (devam ediyor)

ŞEKİLLER DİZİNİ

No Sayfa

ŞEKİLLER DİZİNİ (devam ediyor)

No Sayfa

ÇİZELGELER DİZİNİ

No Sayfa

SİMGELER VE KISALTMALAR DİZİNİ

SİMGELER

P: p değeri

KISALTMALAR

DNA **:** Deoksiribo Nükleik Asit
ENA **:** European Nucleotide Archive
GEO **:** Gene Expression Omnibus
IGV : Integrative Genomics Viewer
NGS **:** Next-generation sequencing (Yeni nesil sekanslama)
RNA **:** Ribo Nükleik Asit
RNA-Seq **:** RNA sekanlama

BÖLÜM 1

GİRİŞ VE AMAÇ

Son yüzyıl içerisinde kanser hücre hattı çalışmaları kanser biyolojisinin anlaşılmasında çok önemli katkılar sağlamıştır. İn vitro deney modellerinin oluşturulmasında sıkça kullanılan bir model olan kanser hücre hatları birçok güvenilirlik tartışmasının odağında olmuştur. Bu tartışmalardan en çok bilineni HeLa kontaminasyonu vakasıdır. Stanley Gartler tarafından yayımlanan bir raporda, çeşitli kanser araştırmalarında kullanılan beyaz ırk tümörü kökenli 18 farklı hücre hattının G6PD enziminin A izozimini taşıdığı ve bu izozimin Afrikalı Amerikalı Henrietta Lacks isimli hastadan izole edilen HeLa isimli kanser hücre hattında da olduğu tespit edilmiştir [1]. Bu tespitin yapıldığı tarihten sonra bu 18 hücre hattı birçok çalışmada kullanılmaya devam edilmiş, Buehring ve arkadaşları 1969-2004 arasında yayınlanmış 220 farklı yayında HeLa kontaminanti olan bu hücre hatlarının in vitro model olarak kullanıldığını tespit etmiştir [2]. Bu ve bu gibi birçok rapor hücre hatlarının kullanılmasına karar verilmeden önce hücre hattının kimliği ile ilgili bir ön çalışma yapma gerekliliğini göstermektedir.

DNA diziliminin okunması (DNA sekanslanması) 1975'e kadar oldukça zor bir uğraştı. Bu tarihten sonra Frederick Sanger ve arkadaşlarının yapmış oldukları çalışmalar genetik biliminin çağ atlamasını sağladı. Sanger ve ark. 1977'de zincir sonlandırma, diğer bir adıyla Sanger sekanslama yöntemini tanıttılar. Sanger dizileme yöntemi, 1977'de ilk kez sunulduğundan beri biyolojik araştırmalarda en etkili yeniliklerden biri olmuştur [3].

Bu keşif modern dizilemenin temelini oluşturdu ancak ilerleyen süreçlerde İnsan Genom Projesi'nin başlamasıyla birlikte kullanılan bu tekniğin oldukça zaman alıcı ve yüksek maliyete neden olduğu anlaşılmış ve bu durum yeni bir dizileme metodunun geliştirilmesinin gerekliliğini ortaya çıkarmıştır. Bu yeni teknolojiler daha önce kullanılan Sanger'in (zincir sonlandırma) metoduna göre DNA ve RNA'yı oldukça hızlı ve daha düşük bir maliyetle dizileme imkanı sağlıyordu. Yeni Nesil Sekanslama (Next-Generation Sequencing-NGS): Büyük çaplı paralel veya derin sekanslama, genomik araştırmalarda devrim yaratan bir DNA sekanslama teknolojisini

tanımlayan ilgili terimlerdir [4]. NGS, tek bir örnekten alınan milyonlarca parçaya ayrılmış bir DNA molekülünün her bir parçasının aynı anda ve uyum içerisinde paralel olarak sekanslanmasını temel alır. NGS'de daha önceki dizileme tekniklerinden farklı olarak paralel sekanslama reaksiyonu aynı sürede yapılarak yüksek hacimli ve hızlı sonuçlar alınabilmektedir. Diğer önemli ve dikkat çekici bir özelliği ise yüksek doğrulukta dizileme kapasitesine sahip olmasıdır.

Yeni nesil sekanslama teknolojisinin gelişmesiyle birlikte sekanslama işlemi sonucu oluşan ham dataların amaca göre analiz edebilmesini mümkün kılan araçlar (İng. 'tool'lar) geliştirilmiş ve bu araçların geliştirilme süreçleri her geçen gün gelişmekte olan genomik alanında yapılan çalışmalar neticesinde devam etmektedir. Sekanslama sonucu oluşan verilerin uygun biyoinformatik ve genomik araçlarla analizi çok önemlidir.

Son yüzyıl içerisinde kanser hücre hattı çalışmaları kanser biyolojisinin anlaşılmasında çok önemli katkılar sağlamıştır. İn vitro deney modellerinin oluşturulmasında sıkça kullanılan bir model olan birçok farklı tip kanser dokusunu temsil eden kanser hücre hatlarına ait çok çeşitli yüksek çıktılı (İng. 'high throughput') moleküler profilleme yapılmıştır. Bu profillemelerden en iyi bilineni NCI-60 kanser hücre hattı paneline ait olan profillemelerdir. NCI-60 hücre hattı panelini oluşturan kanser kimliği teyit edilmiş 53 farklı hattın 'gene expression array'leri ile gen ifadesi profillemesi [5-7], SNP genotyping mikroarray'leri ile tüm genom SNP profillemesinin [8] yanı sıra metilom profillemesi [9] gibi birçok farklı moleküler profillemesi yapılmıştır. NCI-60 paneline dahil olan ve bu panelin haricindeki birçok farklı hücre hattında da daha ileri moleküler profilleme yöntemleri kullanılarak gen ifadesi ölçümleri yapılmış, bu ölçümlerin depolandığı ham veri dosyaları çeşitli biyoinformatik veri tabanlarında tüm kanser araştırmacılarının bedava olarak erişebileceği şekilde analiz edilebilir olarak kullanımlarına sunulmuştur. Bu ileri gen ifadesi profillemelerinin en gelişmişi NGS teknolojisidir ve NGS teknolojisi ile elde edilmiş ham gen ifadesi verilerine Gene Expression Omnibus ve ArrayExpress gibi veri tabanları aracılığıyla erişilmektedir.

Bu çalışmanın konusu, genel çerçevede kanser biyolojisidir. Daha derine inildiğinde ise in silico yollarla ulaşılabilen yeni nesil RNA sekanslama [ve gen ifadesi mikrodizin] verilerinin biyoinformatik yöntem ve araçlar kullanılarak analiz edilmesiyle kanser hücre hatlarına ait

transkriptomik verilerin elde edilmesi olarak tanımlanabilir. Bu çalışmada farklı kanser hücre hatlarına ait RNA sekanslama ham verileri uygun biyoinformatik yöntemler ve araçlar kullanılarak analiz edilmiştir. Kanser hücre hatları gen ifadesi farklılıklarına ve alternatif ekzon kullanımı profillerine göre değerlendirilmiş, bulgular tezin sonuç kısmında paylaşılmıştır.

BÖLÜM 2

GENEL BİLGİLER

2.1 RNA SEKANSLAMA

RNA sekanslama (RNA-Seq), bir hücrenin transkriptomuna ilişkin görüş sağlamak için yüksek verimli sekanslama yöntemlerinin yeteneklerini kullanır. Önceki Sanger sıralaması ve mikrodizi tabanlı yöntemlerle karşılaştırıldığında RNA-Seq, transkriptomun dinamik doğasının çok daha yüksek bir kapsamını ve daha yüksek çözünürlüğünü sağlar [10]. Yeni nesil dizileme yöntemlerinin ortak özelliği, kılcal elektroforez bazlı Sanger dizileme cihazlarının kapasitesini aşan dizileme reaksiyonlarını tek bir deneyde gerçekleştirebilmektir. Gen ekspresyonunu ölçmenin ötesinde RNA-Seq tarafından üretilen veriler yeni transkriptlerin keşfini, alternatif olarak eklenmiş genlerin tanımlanmasını ve alele özgü ekspresyonun saptanmasını kolaylaştırır [10]. Günümüzde değişik firmalara ait yeni nesil dizileme platformları bulunmaktadır. Her platformun kendine özgü dizileme teknolojisi vardır. Bir çalışmada kullanılacak yeni nesil sekanslama platformunun seçimi, o çalışmanın türüne göre değişiklik göstermektedir.

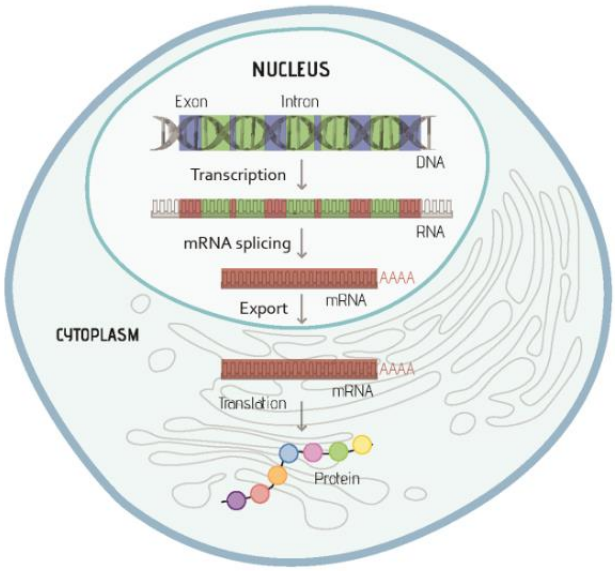
RNA-seq yöntemi son yıllarda hızla gelişen yeni nesil sekanslama teknolojilerinin bir çeşitidir. RNA-seq ile birlikte yaygın olarak kullanılan yeni nesil sekanslama yöntemleri arasında Whole Exome Sequencing (WES) olarak bilinen, genlerin kodlayan sekansları olan ekzonlarin tamamının sekanslanmasi ile uygulanan bir yöntemdir. Bir başka yaygın olarak kullanılan yeni nesil sekanslama yöntemi ise Whole Genome Sequencing (WGS) olarak bilinen tüm genomik DNA'nın sekanslanmasi esasına dayanan yöntemdir. Tüm yeni nesil sekanslama yöntemleri ortak bir prensiple çalışır. Bu prensip ise çok yüksek miktarda genetik (DNA) veya transkriptomik (RNA) sekans bilgisinin üretilmesidir. RNA-seq yönteminde biyolojik materyalin transkribe ettiği (ifade ettiği) gen bölgelerinin transkriptlerinin sekansları elde edilir.

2.2 DİFERANSİYEL GEN EKSPRESYONU

Gen ekspresyonu, genlerde kodlanmış bilginin (nükleotid sekansının) protein gibi işlevsel moleküllerin ve hücrenin yapılarını üretmek için kullandığı işlem olarak tanımlanabilir. Bazı genler bazı RNA formlarının (örneğin; Transfer tRNA, Ribozomal rRNA) üretilmesinden sorumludur.

Gen ekspresyonu iki ana adımdan oluşur. Bunlardan ilki transcription (ifade), hedef Messenger(mesajcı) RNA'nın (mRNA'nın) üretilmesidir. Diğer bir ifadeyle bir genin DNA sekansının RNA kopyasını yapmaktır. Protein kodlayan bir gen için RNA kopyası veya transkripti, bir polipeptit (protein veya protein alt birimi) oluşturmak için gereken bilgiyi taşır. İkinci adım ise translation (çeviri) olarak adlandırılır. Burada ise transkripsiyon sonucu elde edilen mRNA' daki koda göre ribozomlarda meydana gelen amino asit zinciri veya polipeptit sentezi sürecidir.

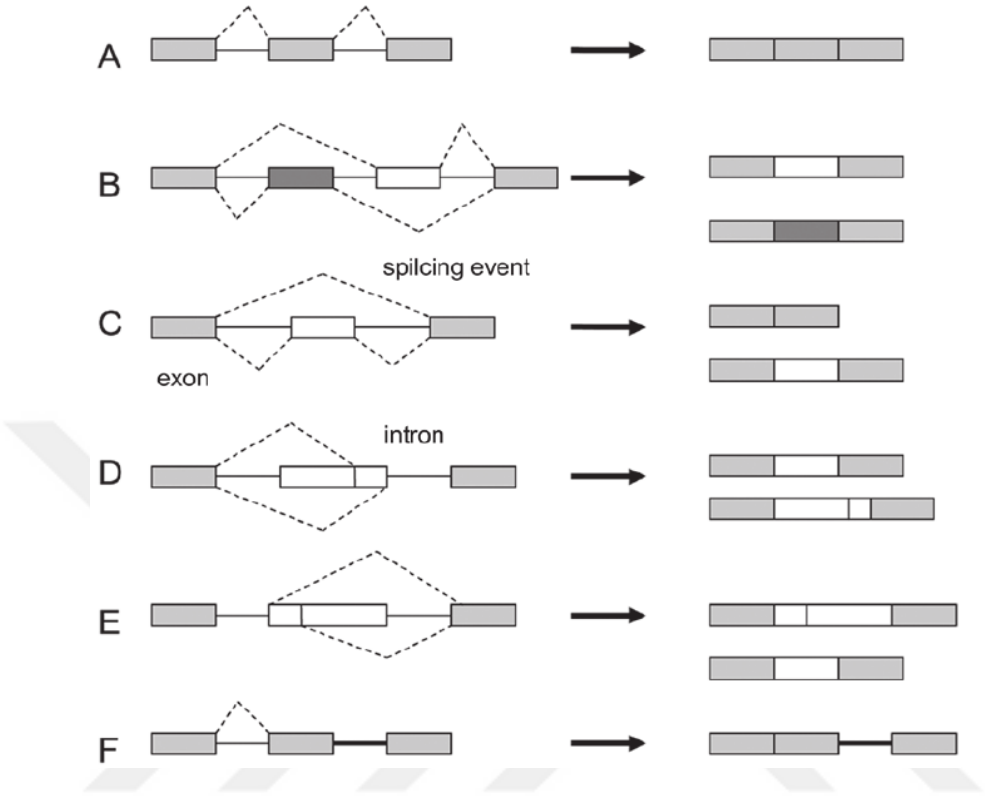
Diferansiyel gen ekspresyonu, bir tedaviye (İng. "treatment") veya herhangi bir gelişim aşamasına yanıt olarak bir genin kontrollü bir aşamaya göre ifadesindeki (mRNA miktarındaki) artmayı veya azalmayı ifade eden durumudur.



Şekil 2.1 Ökaryot bir hücrede DNA'dan proteine bilgi akışına genel bakış [11].

2.3 ALTERNATİF EKSON KULLANIMI

Alternatif splicing (ekleme), olgun mRNA üretimi sırasında gerçekleşen bir moleküler hücre biyolojisi olayıdır. Alternatif splicing farklı eksonik veya intronik segmentlerin olgun mRNA dizisine dahil edilmesi veya hariç tutulması sonucunda farklı mRNA izoformlarının oluşumu ile karakterize edilir. Bu süreç, sınırlı sayıda genden yola çıkıp çok çeşitli proteininin üretilmesinden sorumludur [12]. Beş tip alternatif splicing olayı vardır; constitutive splicing (yapıcı birleştirme), mutually exclusive exons (karşılılı özel eksonlar), cassette alternative exon (kaset alternatif eksonu), alternative 3’ splice site (alternatif 3 ‘ek yeri), alternative 5’ splice site (alternatif 5 ‘ek yeri), intron retention (intron tutma) dır [13]. Alternatif ekson kullanımı, bağımsız olarak dahil edilebilecek veya hariç tutulabilen ayrı eksonlar olan kaset eksonları ve bir grup ekson varyantından bir tanesinin seçilmesini içeren karşılıklı ayrı eklemeyi (splice edilmeyi) içerir [14]



Şekil 2.2 Beş tip alternatif splicing olayı. (A) Yapıcı birleştirme (Constitutive splicing) (B) karşılık özel eksonlar (mutually exclusive exons) (C) kaset alternatif eksonu (cassette alternative exon) (D) alternatif 3'ek yeri (alternative 3' splice site) (E) alternatif 5'ek yeri (alternative 5' splice site) (F) intron tutma (intron retention) [14].

2.4 ANALİZLERDE KULLANILAN CİHAZLAR VE BİYOİNFORMATİK ARAÇLAR

2.4.1 Cihazlar

Analizler için uygun yazımları kullanmak ve analizleri yapmak için iki bilgisayar kullanıldı.

- Dell Precision T3630 iş istasyonu(OMEGA). Intel C246 Çipset, Intel Xeon E-2136 (6 Core) işlemci, 16GB DDR4 2666MHz ECC ram, nVidia Quadro P2000 (5GB GDDR5) ekran kartı, Ubuntu/Linux işletim sistemi
- Casper Nirvana notebook, Intel core i5-4200M işlemci, 8 gigabyte ram, NVDIA Geforce 740M ekran kartı, Ubuntu/Linux işletim sistemi

2.4.2 Yazılımlar

Minion: Minion, sekanslama verilerinde 3 'adaptör sekansının varlığını çıkarmak veya test etmek için küçük bir yardımcı programdır [15].

Cutadapt: Yeni nesil sekanlama çalışmalarında, sekans parçalarının eksiksiz ve doğru okunabilmesi için bu sekans parçalarına üç üssü (3') ve beş üssü (5') uçlarından adaptör adı verilen sentetik oligonükleotitler eklenir. Cutadapat yazılımı, yeni nesil sekanslama sekans okumalarından adaptörler dizilerini uzaklaştırır. Komut satırında çalışır [16].

FASTQC: NGS platformları sekansları hatalı okuyabilmektedir. Özellikle sonlara doğru reaksiyona giren enzimlerin ve kimyasalların verimliliğinin düşmesi ve sekans okuma cihazının kirlenmesi sonucu hatalı okuma artmaktadır. Bunun yanında DNA kütüphanelerinin hazırlanmasında meydana gelen kontaminasyonlar da hataya yol açmaktadır. NGS çıktı formatı olan fastq, bir sekansın her bir nükleotiti için doğruluk değerleri bulundurur (kalite skoru). FastQC, fastq formatındaki NGS verilerini okuyarak, verilerin kalitesi hakkında bilgi verir. Grafiksel kullanıcı arayüzüne sahiptir ve komut satırında da çalışmaktadır [17].

Trimmomatic: NGS verilerinde, düşük kaliteli veya adaptörler gibi teknik sekansların varlığı analizlerin kötü sonuçlanmasına neden olur. Trimmomatic, bu düşük kaliteye sahip sekansları kesip uzaklaştırır. Komut satırında çalışır [18].

HISAT2: Yeni nesil sekans okumalarını (hem DNA hem de RNA) insan genomları popülasyonuna ve tek bir referans genomuna eşlemek için hızlı ve hassas bir hizalama programıdır. Komut satırında çalışır [19]

Samtools: Sequnce (Sekans) Alingment (Hizalama) / Map (Harita) (SAM) formatı, okuma hizalamalarını referans sekanslara karşı saklamak için genel bir hizalama formatıdır**.** SAMtools, SAM formatındaki hizalama sonrası işlemler için indeksleme, değişken arayan (İng. "variant caller") ve hizalama görüntüleyici gibi çeşitli yardımcı programlar uygular ve böylece okuma hizalamalarını işlemek için evrensel araçlar sağlar. Komut satırında çalışır [20].

StringTie: StringTie, RNA-Seq hizalamalarının potansiyel transkriptlere hızlı ve yüksek verimli bir birleştiricisidir. Her bir gen lokusu için çoklu ek varyantları temsil eden tam

uzunlukta transkriptleri birleştirmek ve nicelemek için yeni bir ağ akış algoritması ve isteğe bağlı bir de novo montaj adımı kullanır. Komut satırında çalışır [21].

Integrative Genomics Viewer (IGV): Haritalama, varyant, anotasyon gibi genomik veri setlerinin görselleştirilmesinde kullanılmaktadır. Grafiksel kullanıcı ara yüzüne sahiptir [22].

R: R, istatistiksel hesaplama ve grafikler için ücretsiz bir yazılım ortamıdır. Çok çeşitli UNIX platformları, Windows ve MacOS üzerinde derlenir ve çalışır. Komut satırında çalışır [23].

RStudio: R programlama dili için yeni bir entegre geliştirme ortamıdır. RStudio, R'nin çeşitli bileşenlerini (konsol, kaynak düzenleme, grafik, geçmiş, yardım vb.) Birleştirmeyi amaçlayan açık kaynaklı bir projedir. Komut satırında çalışır [24].

Ballgown: Transkript birleşimini gen-gen bazında görselleştirmek, eksonlar, intronlar, transkriptler veya genler için bolluk tahminleri çıkarmak ve doğrusal model tabanlı diferansiyel ekspresyon analizleri gerçekleştirmek için kullanılabilir [25]. R'de paket hali kütüphaneden yüklenerek kullanılabilir.

DEXSeq: RNA-seq verilerinde diferansiyel ekson kullanımını test etmek için istatistiksel bir yöntemdir. DEXSeq genelleştirilmiş doğrusal modeller kullanır ve biyolojik varyasyonu dikkate alarak yanlış keşiflerin güvenilir bir şekilde kontrol edilmesini sağlar. DEXSeq, yüksek hassasiyetli genler ve birçok durumda diferansiyel ekson kullanımı olan eksonları tespit eder.

DEXSeq, göreceli ekson kullanımındaki değişiklikleri yani sadece genin yukarı veya aşağı regülasyonunun sonucu olmayan ayrı ayrı eksonların ekspresyonundaki değişiklikleri bulmak için tasarlanmıştır [26].

2.4.3 Veri tabanları

Gene Expression Omnibus (GEO): GEO, yüksek verimli gen ifadesi ve genomik hibridizasyon deneylerinden heterojen veri setlerinin gönderilmesini, saklanmasını ve alınmasını kolaylaştıran esnek ve açık bir tasarım sunar [27]. Web sitesi: http://www.ncbi.nlm.nih.gov/geo/

European Nucleotide Archive (ENA): EMBL-EBI tarafından sağlanan ENA, otuz yılı aşkın bir süredir dünyanın genel sıralama verilerini arşivlemekten ve bu önemli kaynağı bilimsel araştırmalara küresel araştırma çabalarını desteklemek ve hızlandırmak için sunmaktan sorumludur [28]. Avrupa Nükleotid Arşivi (ENA), ham sekanslama verilerini, sekans montaj bilgisini ve fonksiyonel anotasyonu kapsayan dünyanın nükleotit sekanslama bilgilerinin kapsamlı bir kaydını sağlar.

ArrayExpress: Yüksek verimli sekanslama (HTS) ve mikrodizi tabanlı deneylerden elde edilen fonksiyonel genomik verileri depolar [29]. Yüksek verimli fonksiyonel genomik deneylerinden veri depolar ve bu verileri araştırma topluluğuna yeniden kullanılmak üzere sağlar.

Ensembl: Ensembl projesi, taslak insan genomunun ilk sürümlerinden bu yana genomik veri kümelerinin toplanması, işlenmesi, entegre edilmesi ve yeniden dağıtılması, genomik araştırmalarının kamuya açık verilerin hızlı bir şekilde dağıtılması yoluyla hızlandırılması amacıyla gerçekleştirilmiştir. Büyük miktarlarda ham veri böylece çok sayıda kanaldan, özellikle tarayıcımızdan (http://www.ensembl.org) kullanıma sunulan bilgiye dönüştürülür [30].

2.4.4 Dosya formatları

fastq: Dizinin ve bu sekansı oluşturan her bir nükleobazın cihaz tarafından doğru okunma (kalite) değerinin tutulduğu metin esaslı bir dosya biçimdir. Birinci satır sekans adını, ikinci

satır nükleobaz sekansını, üçüncü satır sekans hakkında açıklamayı (tercih olarak boş bırakılabilinir) ve dördüncü satır nükleobazların kalite değerini içerir. Birinci satır "@", üçüncü satır "+" karakteri ile başlamalıdır.

@Sekans adı

ATCGCTACTCGTA

+Sekans hakkında bilgi(tercihe bağlı)

'!%&/(=?!!'''(==??

fasta: Sadece sekans bilgisinin tutulduğu metin esaslı bir dosya biçimidir. Yaygın olarak kullanılan bir formattır. Dizi adı ">" karakteri ile başlamalıdır.

>Sekans adı

ATCGCTACTCGTA

SAM: Sekans Hizalama Haritası (SAM), bir referans sekansa hizalanmış biyolojik sekansların depolanması için metin tabanlı bir formattır [31]. Yeni nesil sekanslama teknolojileri tarafından üretilen nükleotit sekansları gibi verilerin depolanması için yaygın olarak kullanılır ve standart, eşlenmemiş sekansları içerecek şekilde genişletilmiştir. Referans genoma haritalanan bir sekansın; genom üzerindeki konumu, haritalama kalite değeri, var ise haritalanan bölgedeki dizi eklenme/silinme bilgisi, nükleobaz sekansı, bazların doğru okunma değeri gibi bilgileri içerir.

BAM: Sam dosya biçiminin sıkıştırılması ile elde edilen ikili dosya biçimidir.

GTF: Gen transfer formatı (GTF), gen yapısı hakkında bilgi tutmak için kullanılan bir dosya formatıdır. Anotasyon (kimliklendirme) dosyası olarak adlandırılır.

BÖLÜM 3

GEREÇ VE YÖNTEM

3.1 RNA-Seq datasetlerinin indirilmesi

Kanser hücre hatlarına ait RNA-Seq datasetleri ve bu datasetlere ait bilgiler, herkesin erişime açık olan internet veri tabanlarından ücretsiz olarak indirildi. Bu veri tabanları European Nucleotide Archive (ENA), ArrayExpress ve Gene Expression Omnibus (GEO) dur. Bu veri tabanları kullanılarak kanser hücre hatlarına ait RNA-Seq datasetleri indirildi.

Çizelge 3.1: Kullanılan kanser hücre hattı datasetlerine ait bilgiler.

Kanser hücre hattı	Kanser hücre tipi	Datasetler (GEO kodu)	Sekanlama platformu	Kütüphane düzeni
MCF7	Breast adenocarcinoma	GSE59251 GSE63189	Illumina HiSeq 2000	Single
HeLa	Cervical adenocarcinoma	GSE75410 GSE77913	Illumina HiSeq 2000	Paired
HCT116	Colon carcinoma	GSE120071 GSE131249	Illumina HiSeq 2500	Single
A549	Lung carcinoma	GSE80182 GSE136105	Illumina HiSeq 2500	Paired

Çizelge 3.2: Hücre hatları datasetlerine ait kontrol örnek (İng. “sample)” kodlarının bilgileri.

Hücre hattı	Datasetler (GEO kodu)	Kontrol örnek ENA kodu	Kontrol örnek ENA kodu	Kontrol örnek ENA kodu
MCF7	GSE59251	SRR1509730	SRR1509731	SRR1509732
	GSE63189	SRR1648590	SRR1648591	SRR1648596
HeLa	GSE75410	SRR2960983	SRR2960986	

	GSE77913	SRR3169158	SRR3169161	
HCT116	GSE120071	SRR7865855	SRR7865856	SRR7865857
	GSE131249	SRR9058970	SRR9058971	SRR9058972
A549	GSE80182	SRR10009495	SRR10009496	SRR10009497
	GSE136105	SRR3362661	SRR3362662	SRR3362663

3.2 ANALİZ ÖNCESİ ADIMLAR

3.2.1 İndirilen datsetlerin incelenmesi

İndirilen datasetlere ait örneklerin içeriğine dair bilgi almak analiz için oldukça önemlidir. Bu bilgiler; temel istatistik bilgileri, sekans kalitesi, sekans kalite skoru, sekans uzunluğu, adaptör sekansın olup olmadığıdır. İndirilen datasetler ham veri dosya formatı olan FASTQ formatındadır. Bu bilgileri elde etmek için FASTQC adlı yazılım kullanıldı.

3.2.2 Adaptör sekansın bulunması

Bir önceki adımda FASTQC ile incelenen verilerde adaptör sekans olduğu tespit edilen fastq dosyaları, minion yazılımı kullanılarak adaptör sekansın nükleotid dizilimi tespit edilir.

3.2.3. Adaptör sekansın uzaklaştırılması

Adaptör sekans içeren verilerle çalışmak analiz sonuçlarını yanlış yönde etkiler. Bu yüzden adaptör sekansa sahip verilerden bu sekanslar uzaklaştırılmalıdır. Minion yazılımı kullanılarak tespit edilen adaptör sekansı, CutAdapt yazılımı kullanılarak fastq dosyalarından kesilerek uzaklaştırıldı.

3.2.4 Düşük kaliteye sahip okumaların uzaklaştırılması

FASTQC yazılımı ile incelenen verilerde, düşük okuma kalitesine sahip sekanslar Trimmomatic yazılımı kullanılarak kesildi. Bu işlemler için Trimmomatic değerleri; LEADING:30, TRAILING:30, MINLEN:36 şeklindeydi.

LEADİNG: Eğer sekans okuma kalitesi 30'un altındaysa bazları okumanın başından başlayarak keser

TRAILING: Eğer sekans okuma kalitesi 30'un altındaysa bazları okumanın sonundan başlayarak keser.

MINLEN: Okuma uzunluğu 36'dan düşükse okumayı bırakır.

3.2.5 Okumaların referans genoma hizalanması

Önceki adımlarda okumalardan adaptör sekansı ve düşük kaliteye sahip okumalar uzaklaştırıldı. Bu sayede elimizde hizalama yapmak için uygun veriler kaldı. Okumaları referans genoma hizalama yapmak için öncelikle HISAT2 web sitesinden (https://ccb.jhu.edu/software/hisat2/manual.shtml) insan genomu önceden indekslenerek siteye yüklenmiş olan *genome_tran* dosyası indirildi. Bunun akabinde HISAT2 yazılımı kullanılarak okumalar insan genomuna hizalandı. Hizalama sonucunda sam uzantılı dosyalar elde edildi.

3.2.6 Referans genoma hizalanan okumaların görüntülenmesi

HISAT2 yazılımı kullanarak hizalan okumalar sonucu elde edilen .sam uzantılı dosyalar, Samtools yazılımı kullanılarak .bam formatına dönüştürüldü. Bam dosyaları, genom okuma dizilerinin insan referans genomunda hangi bölgelere denk geldiğinin ve bu bölgelerin kaç defa okunmuş olduğunun bilgisini içermektedir. Tek başına .bam uzantılı dosyalar, okumaların referans genoma nasıl hizalandığını görmek için yeterli değildir .bam uzantılı dosyalar indekslenmelidir. İndeksleme için, .bam uzantılı dosyalar Samtools ve IGV yazılımları kullanıldı. Daha sonra okumaların referans genoma nasıl hizalandığı IGV adlı yazılımla kontrol edildi.

3.3 Diferansiyel gen ekspresyon analizi

Örnekleri referans genoma hizaladıktan sonra, gen ekspresyon analizi için ilk olarak her bir örnek için ifade edilen genleri ve transkriptleri bir araya getirip ve ölçüm yapmamız gerekiyor. Bu ölçümü yapmak için StringTie yazlımı, Ensembl veri tabanından indirmiş olduğumuz içerisinde gen anotasyon bilgilerini içeren gtf uzantılı dosyayı ve örneğin bam uzantılı dosyasını

kullandık. StringTie yazlımı kullanarak her bir örnek için transkriptleri birleştirdik (İng. “transcript assemble”). Bu işlemin ardından her bir örnek için birleştirmiş olduğumuz transkriptleri bu sefer tek bir dosya halinde StringTie yazılımı kullanarak birleştirdik. Çıkan dosya formatı da gtf uzantılı ve bu dosya içinde her bir örneğin transkript birleşmesi mevcut.

Bu işlemin ardından Ballgown yazılımında kullanmak için, bir önceki basamakta birleştirmiş olduğumuz transkriptlerin, transkript bolluğunu tahmin etmek ve Ballgown yazılımı için okuma Çizelgelarını oluşturmak için StringTie yazılımı kullandık.

Gen ekspresyon analizini tamamlamak için R programı kullanımına geçildi. R programaında, kütüphaneden Ballgown yüklendi. Ballgown ve R kullanılarak örneklerin diferansiyel gen ekspresyonu analiz edildi ve grafikler oluşturuldu.

Yukarıdaki adımlar her bir hücre hattı için ayrı ayrı yapıldı. Bu diferensiyel gen ekspresyonu analizinde kullanılan metotlar Mihaela Pertea ve arkadaşları tarafından 2016 yılında yapılan çalışmada detaylı olarak açıklanmıştır [32].

3.4 Alternatif ekson kullanımı analizi

Bu analizin ilk adımı olarak DEXseq yazılımının kurulu olduğu dosyada mevcut olan iki python kodunu kullanmamız gerekiyor. Bu kodlar *dexseq_count.py* ve *dexseq_prepare_annotation.py*’dir. *Dexseq_prepare_annotation.py* Ensembl gtf dosyasını alır ve dosyayı daraltılmış ekson sayma bölmeleriyle bir gff dosyasına dönüştürür. İlk olarak *dexseq_prepare_annotation.py* kullanarak Ensembl gtf dosyasını gff uzantılı hale çevirdik. Çünkü bir GTF dosyasında birçok ekson, bunları içeren her transkript için bir kez olmak üzere birden çok kez görünür. Ekson sayma kutularını tanımlamak için bu bilgiyi daraltmamız gerekir.

Sonraki adımda her sam dosyası için, bir önceki adımda hazırlanan gff dosyasında tanımlanmış olan ekson sayım bölmelerinin her biri ile çakışan okuma sayısını saydık. Bunu python_count.py kullanarak yaptık.

Bu işlemin ardında analizi tamamlamak için R programına geçildi. R kütüphanesinden DEXSeq yüklendikten sonra alternatif ekson kullanımı analizi ve sonuçlara ait grafikler elde edildi.

Yukarıdaki basmaklar her bir kanser hücre hattı için ayrı ayrı yapıldı. Alternatif ekson kullanımı analiz yöntemi basamaklarına dair detaylı bilgiler Reyes ve arkaşları tarafından yapılan çalışmada bulunabilir [33].

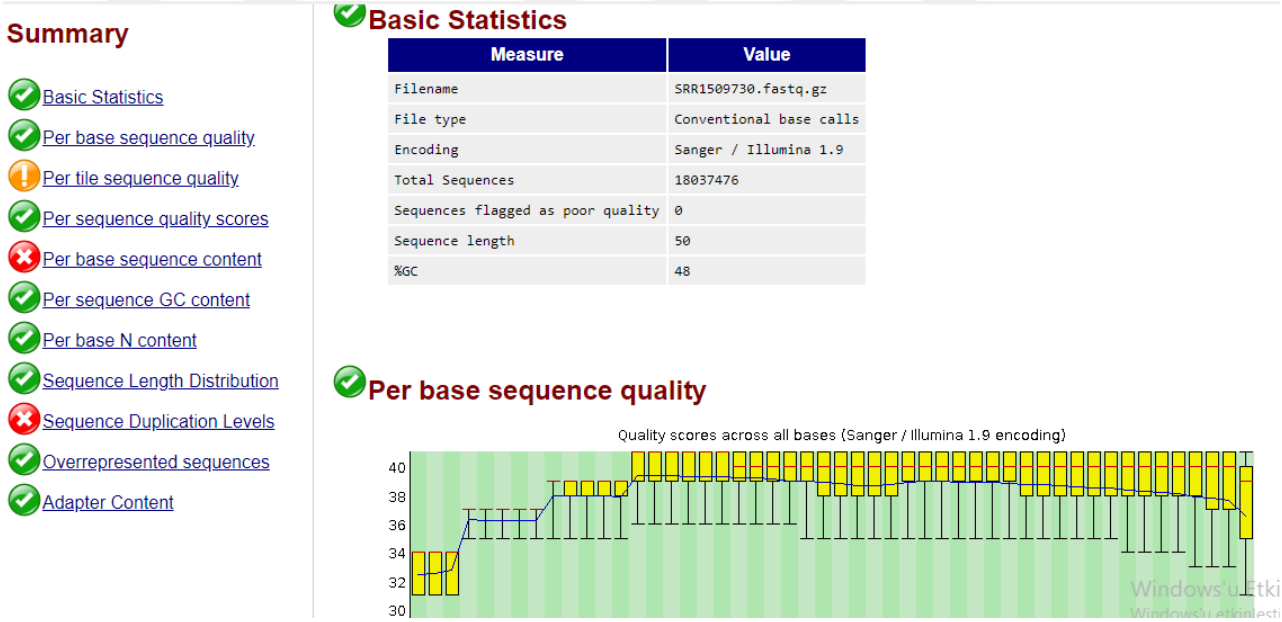
BÖLÜM 4

BULGULAR

4.1 ANALİZ ÖNCESİ BULGULAR

4.1.1 FASTQC ve Trimmomatic bulguları

İndirilen ham RNA-Seq datasetleri hakkında bilgi edinmek için FASTQC yazılımı kullanıldı.



Şekil 4.1 MCF7 kanser hücre hattı GSE59251 datasetine ait FASTQC sonucu

FASTQC ile incelenen ham verilerde sekans kalitesi düşük ise Trimmomatic yazlılımı kullanılarak düşük kaliteye sahip okumalar uzaklaştırıldı.

Çizelge 4.1: Trimmomatic öncesi ve sonrasına ait dataset örnek bilgileri.

Hücre hattı	Dataset	Veri ismi	Veri sekans uzunluğu	Trimmomatic sonrası uzunluk	Kesilen sekans uzunluğu
MCF7	GSE59251	SRR1509730	24308901	24059303	249598
		SRR1509731	21619285	21417095	202190
		SRR1509732	18037476	17872172	165304
	GSE63189	SRR1648590	11275933	11092727	183206
		SRR1648591	11110357	10856615	253742
		SRR1648596	10901186	10743675	157511

Çizelge 4.2: Trimmomatic öncesi ve sonrasına ait dataset örnek bilgileri.

Hücre hattı	Dataset	Veri ismi	Veri sekans uzunluğu	Trimmomatic sonrası uzunluk	Kesilen sekans uzunluğu
HCT116	GSE120071	SRR7865855	61273390	60946372	327018
		SRR7865856	65210417	64843600	366817
		SRR7865857	66947177	66544363	402814
	GSE131249	SRR9058970	26255001	26177517	77484
		SRR9058971	25708122	25627475	80647
		SRR9058972	27134587	27047096	157511

Çizelge 4.3: Trimmomatic öncesi ve sonrasına ait dataset örnek bilgileri.

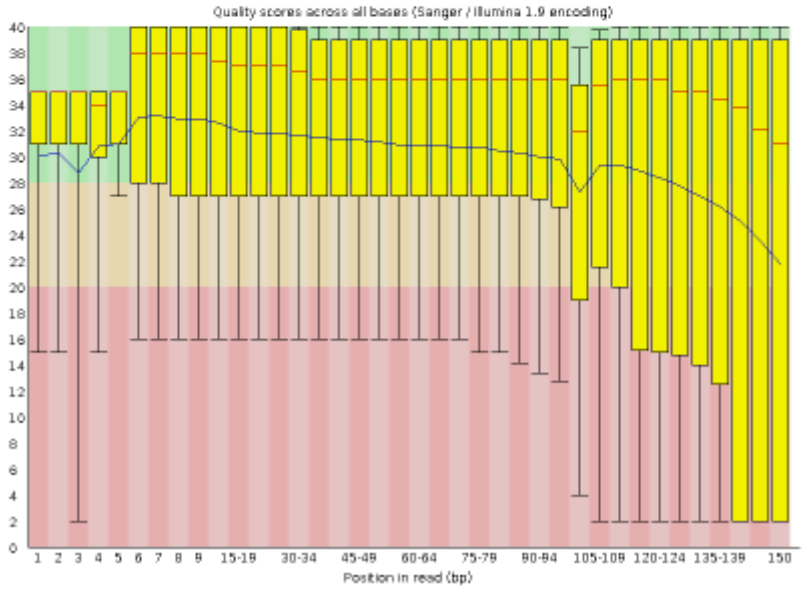
Hücre hattı	Dataset	Veri ismi	Veri sekans uzunluğu	Trimmomatic sonrası uzunluk	Kesilen sekans uzunluğu
A549	GSE80182	SRR3362661	31757382	27255878	4501504
		SRR3362662	37327954	31920621	5407333
		SRR3362663	34160910	28965798	5195112
	GSE136105	SRR10009495	50238619	49971463	267156
		SRR10009496	48755722	48442431	313291
		SRR10009497	52644489	52305940	338549

HeLa kanser hücre hattı yapılan FASTQC analizi sonucu, sekans okuma kalitesi analiz için yeterli olduğundan dolayı Trimmomatic yazılımı kullanılmamıştır.

Basic Statistics

Measure	Value
Filename	SRR3362661_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	31757382
Sequences flagged as poor quality	0
Sequence length	150
%GC	50

Per base sequence quality

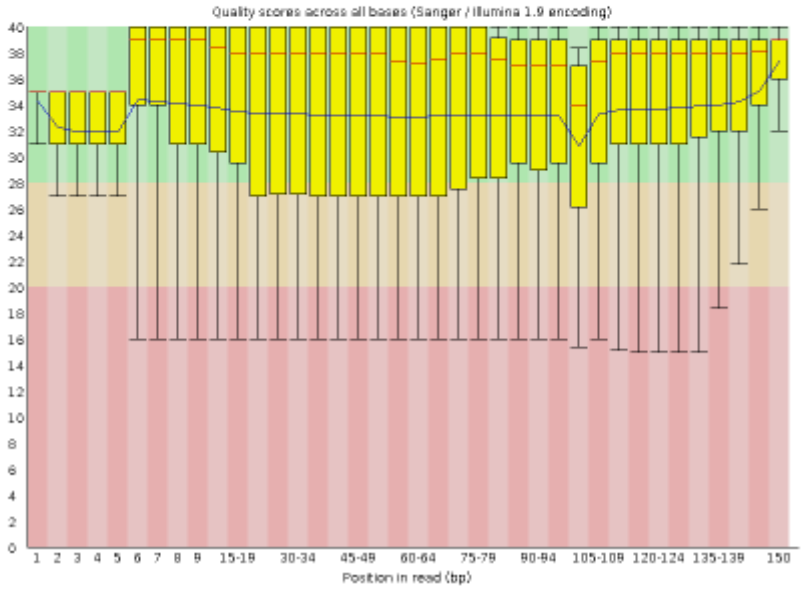


Şekil 4.2 A549 kanser hücre hattı GSE80182 datasetine ait FASTQC sonucu.

Basic Statistics

Measure	Value
Filename	SRR3362661_p1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	26992870
Sequences flagged as poor quality	0
Sequence length	15-150
%GC	49

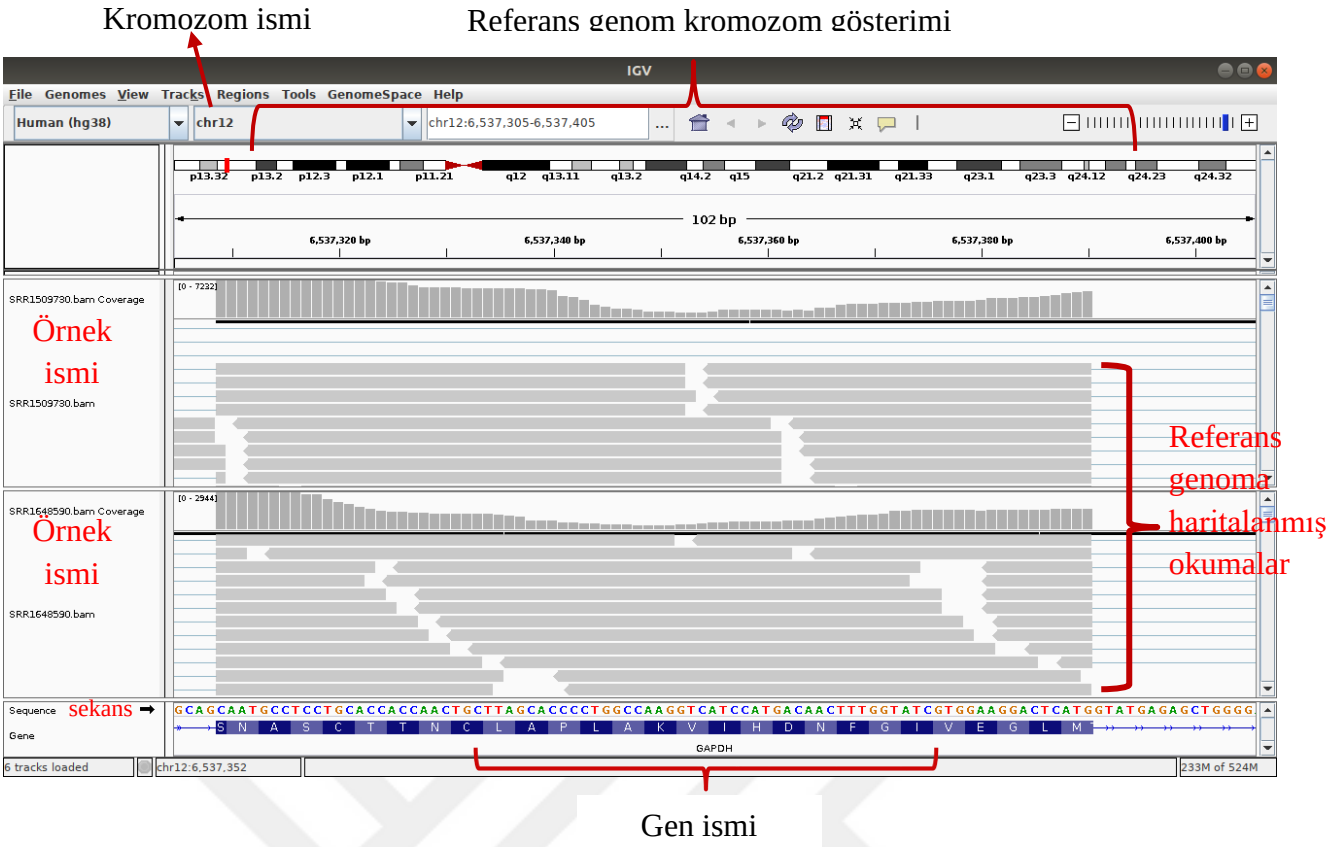
Per base sequence quality



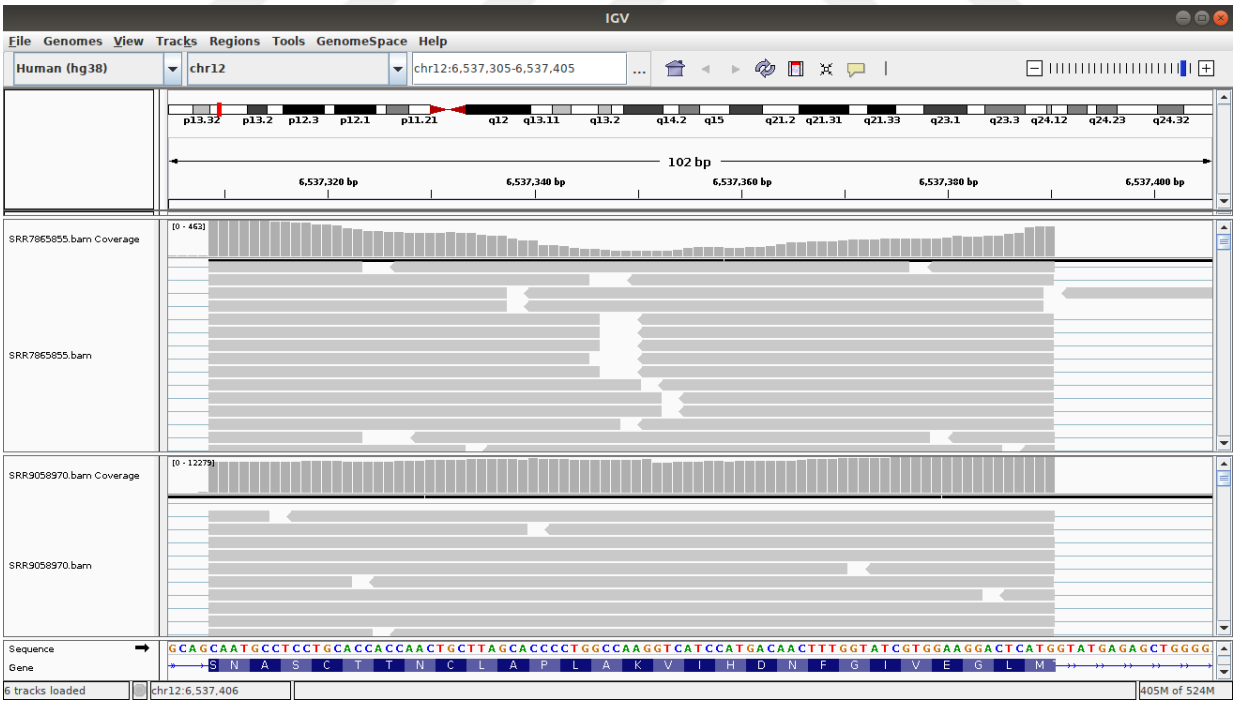
Şekil 4.3 A549 kanser hücre hattı GSE80182 datasetine ait Trimmomatic sonrası FASTQC sonucu.

4.1.2 Datasetlerin referans genoma hizalanması

Adaptör sekansı ve düşük okuma kalitesine sahip okumalar örneklerden uzaklaştırıldıktan sonra HISAT2 ile insan referans genomuna hizalandı. Hizalama sonucu çıkan sam uzantılı dosyalar bam formatına dönüştürüldükten sonra IGV ile görüntülendi.



Şekil 4.4 Referans genoma hizalanan okumaların IGV yazılımında görüntülenmesi



Şekil 4.5 HCT116 hücre hattı GSE120071 dataseti (üstte) ve GSE131249 (altta) ait örneklerin referans genoma hizalanması.

4.2 DİFERANSİYEL GEN EKSPRESYON ANALİZİ

4.2.1 MCF7 kanser hücre hattı

Ballgown ve R kullanılarak MCF7 kanser hücre hattına ait GSE59251 dataseti ve GSE63189 dataseti diferansiyel gen ekspresyon için karşılaştırıldığında, toplamda 12.380 gen analiz edildi ve bunların 6.774 tanesinde istatiksel olarak anlamlı bir şekilde ($p<0.05$) diferansiyel gen ekspresyonu tespit edildi. Toplamda 29.381 transkript arasından 7.892 tanesi datasetler arasında farklı ifade edilen transkriptlerin sayısı olarak kaydedildi.

Çizelge 4.4: MCF7 kanser hücre hattı datasetleri arasında farklı ifade edilen transkriptler ($p<0.05$).

geneNames	geneIDs	feature	id	fc	pval	qval
MKRN2	ENSG00000075975	transcript	147967	0.369765548108486	0.000100066777888541	0.00619546805112366
AC004542.6	ENSG00000276965	transcript	144419	378.380.650.676.668	0.000100161578036273	0.00619546805112366
CUTC	ENSG00000119929	transcript	25554	0.219563152858374	0.000101319460635296	0.00625392242211269
MAGOHB	ENSG00000111196	transcript	42960	0.235257204702241	0.000101575124213715	0.0062558748414124
SPESP1	ENSG00000258484	transcript	71062	0.162053879460193	0.000102015106988151	0.0062558748414124
FTLP3	ENSG00000226608	transcript	134227	198.128.481.374.223	0.000102050982322188	0.0062558748414124
PDS5B	ENSG00000083642	transcript	54625	332.546.353.826.438	0.000102529507824078	0.0062558748414124
MYOF	ENSG00000138119	transcript	25020	0.0276827421802307	0.00010261847086146	0.0062558748414124
GAPDHP65	ENSG00000235587	transcript	221218	0.521329985160521	0.00010282481629531	0.0062558748414124
DIO2	ENSG00000211448	transcript	63573	0.179599290822747	0.00010284154890583	0.0062558748414124
PSMB1	ENSG00000008018	transcript	191773	0.240448760107996	0.000103256657628337	0.0062681484664838
KCNMB2-AS1	ENSG00000237978	transcript	158801	210.624.354.647.838	0.000103756148769008	0.00628548331336538
TTC3P1	ENSG00000215105	transcript	223249	321.812.256.332.061	0.000104404508006639	0.0063117466044096
POLR3H	ENSG00000100413	transcript	145893	183.872.307.763.531	0.000106783994396675	0.0064423419699563
H3F3C	ENSG00000188375	transcript	44413	0.318011104878062	0.000109436356377834	0.00658883112036302
RNF114	ENSG00000124226	transcript	137777	0.280912641021933	0.000109759738386006	0.00659277180051612
MAP3K20	ENSG00000091436	transcript	128489	0.278663191433695	0.000109950586510088	0.00659277180051612
ZNF680	ENSG00000173041	transcript	195653	140.046.970.050.417	0.000110656850814572	0.00662160678978197
TFAP4	ENSG00000090447	transcript	76612	0.344536830239932	0.000111368213046692	0.00662550030615674
RPS27	ENSG00000177954	transcript	12462	0.00368744306678681	0.000111524245560934	0.00662550030615674
C1orf115	ENSG00000162817	transcript	17823	925.673.828.290.656	0.000111659393200236	0.00662550030615674
PEX26	ENSG00000215193	transcript	142325	0.307192173622351	0.000111838763520389	0.00662550030615674
FTO	ENSG00000140718	transcript	81031	907.787.839.942.741	0.000112187660801899	0.00662550030615674
RANP1	ENSG00000236603	transcript	183606	0.346177105588193	0.000112702070418469	0.00662550030615674
EIF2B1	ENSG00000111361	transcript	52851	0.496338340009615	0.000112838110706726	0.00662550030615674
FOLR1	ENSG00000110195	transcript	36274	0.149767180909259	0.000112940061499089	0.00662550030615674
CBSL	ENSG00000274276	transcript	139217	0.109357420975295	0.000113037184491072	0.00662550030615674

fc(fold change): GSE63189 datasetinin GSE59251 datasetine karşı aralarındaki ekspresyon oranını gösterir; dolayısıyla 1'in altındaki değerler transkriptin GSE59251'de daha düşük bir seviyede ifade edildiği anlamına gelir.

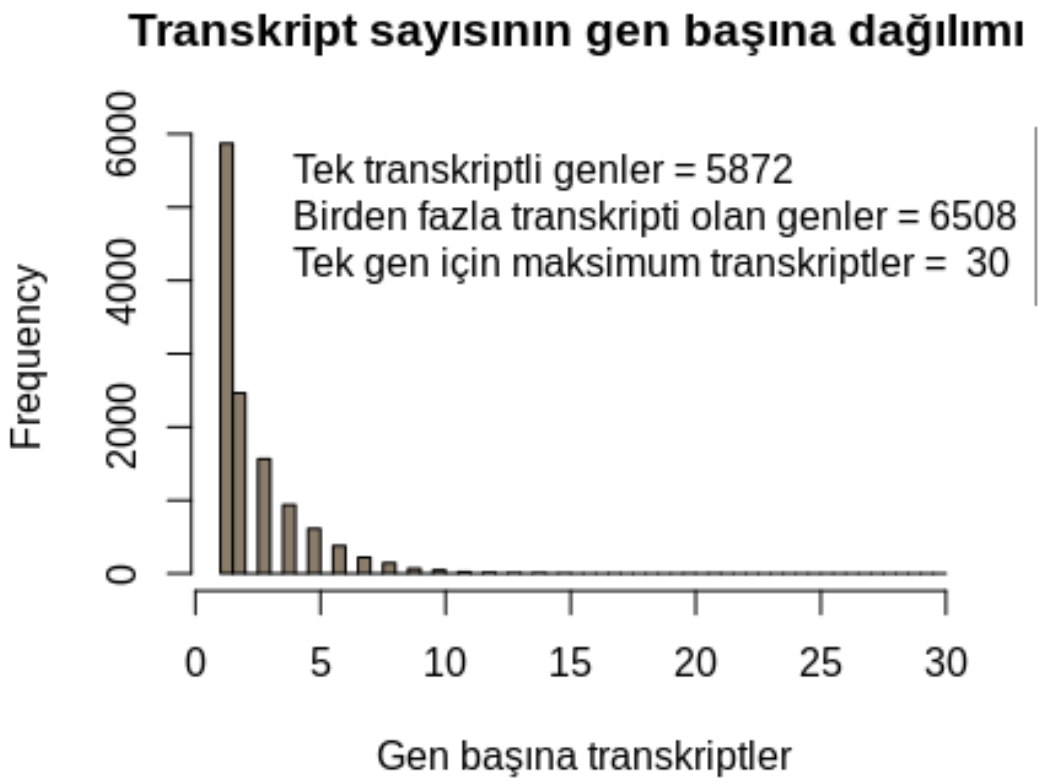
İlk kolon gen isimlerini, ikinci kolon ise gene ID'lerini göstermektedir. Altıncı kolona baktığımızda P değeri (pval) görülmektedir. P değeri, bir karşılaştırmada istatiksel anlamlı fark vardır diyebileceğimiz zaman olası hata miktarının gösterir. Bu hatanın maksimum kabul edilebilir düzeyi 0.05 olarak kabul edilir. P değeri 0,05'in altında bir değer ise karşılaştırma sonucunda anlamlı farklılık bulunduğu anlamına gelir. q değeri (qval) ise, optimize edilmiş FDR yaklaşımı kullanılarak bulunan ayarlanmış p değerlerine verilen addır. FDR yaklaşımı, q-

değerlerinin bir listesini üretmek için p-değeri dağılımının karakteristiklerini kullanarak optimize edilmiştir. 0.05 değerinde bir FDR ayarlı p-değeri (veya q-değeri), önemli testlerin % 5'inin yanlış pozitiflerle sonuçlanacağı anlamına gelir [34].

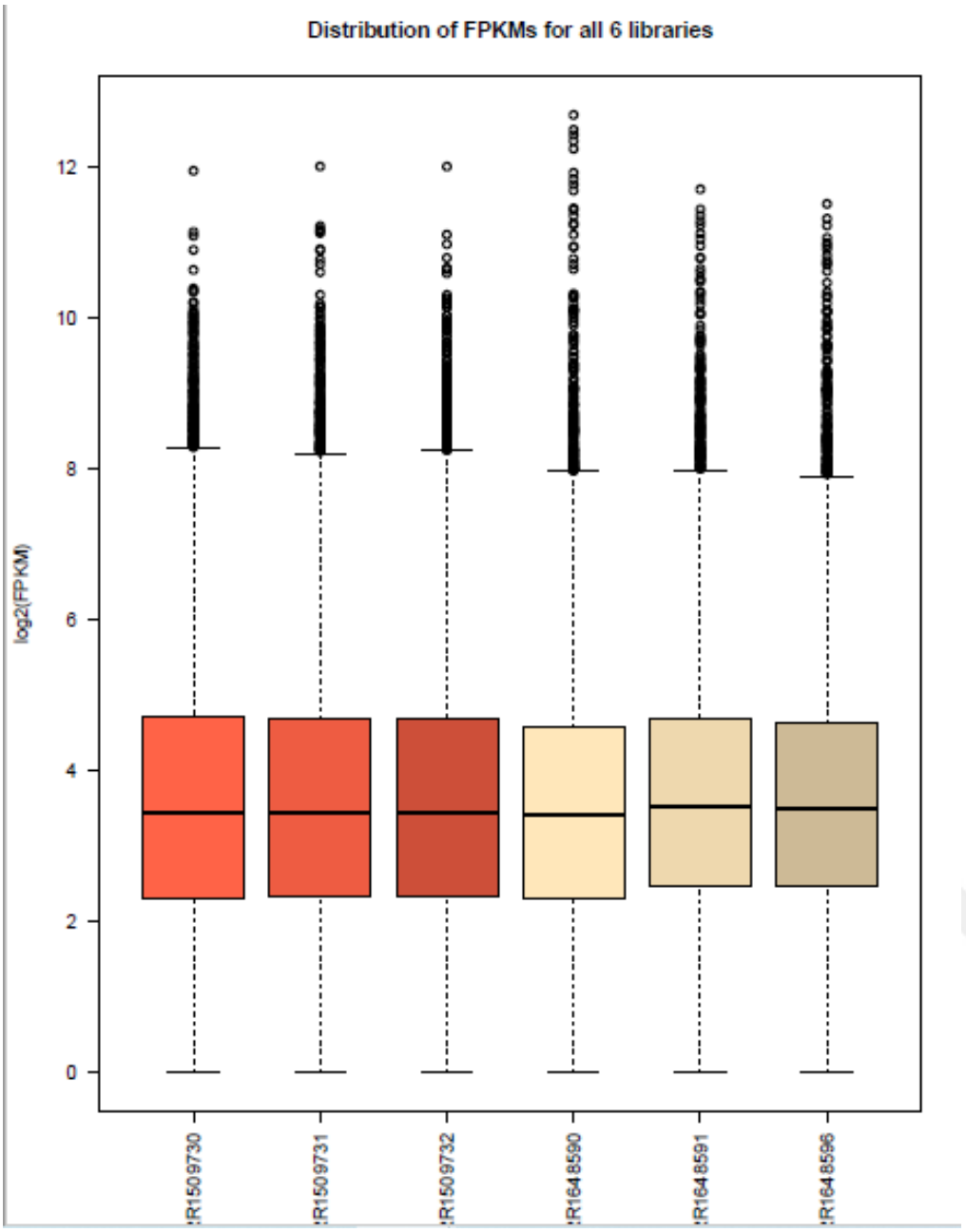
Çizelge 4.5: MCF7 kanser hücre hattı datasetleri arasında farklı ifade edilen genler (p<0.05).

featu	id	fc	pval	qval
gene	ENSG00000145687	257.985.648.607.734	0.000100240224969905	0.00352549427593019
gene	ENSG00000147454	0.268588515428856	0.000100877851096914	0.00353786911212408
gene	ENSG00000226701	791.548.356.183.832	0.000102084886537179	0.00357008727494429
gene	ENSG00000126878	243.921.055.232.531	0.000102735388998854	0.00357548820863564
gene	ENSG00000102683	0.214279902259633	0.00010283624027807	0.00357548820863564
gene	ENSG00000026508	0.298481150699877	0.000103107021041327	0.00357548820863564
gene	ENSG00000197620	0.392489897440403	0.000103394570168946	0.00357548820863564
gene	ENSG00000158604	371.381.620.756.807	0.000103688030577587	0.00357564851963934
gene	ENSG00000248783	414.893.217.090.907	0.00010444494327233	0.00359140826334357
gene	ENSG00000136933	0.39713884125101	0.000104922243379768	0.00359140826334357
gene	ENSG00000198380	233.243.789.429.323	0.000105337250372983	0.00359140826334357
gene	ENSG00000167600	0.316293446270292	0.000106189708258597	0.00359140826334357
gene	ENSG00000178035	0.543150145525879	0.000106217149324461	0.00359140826334357
gene	ENSG00000167325	0.47786783836261	0.000106393151209305	0.00359140826334357
gene	ENSG00000237289	23.833.188.204.806	0.000106396120154617	0.00359140826334357
gene	ENSG00000163082	716.951.884.301.313	0.00010646581846907	0.00359140826334357
gene	ENSG00000113209	416.838.513.650.056	0.000106912212400889	0.00359666627587774
gene	ENSG00000170448	0.576032678387902	0.000108069217193418	0.00360819103270239
gene	ENSG00000005189	0.390834353742969	0.00010809249853283	0.00360819103270239
gene	ENSG00000132294	0.285214917080699	0.000108129149687608	0.00360819103270239
gene	ENSG00000204308	230.627.540.140.916	0.000108811146464216	0.00362091741036533
gene	ENSG00000157077	277.954.446.726.521	0.00010909549225091	0.00362091741036533
gene	ENSG00000144560	193.674.512.282.727	0.000110277557244798	0.00364770047724254
gene	ENSG00000133313	173.663.202.743.425	0.000110491734973017	0.00364770047724254
gene	ENSG00000174989	123.879.258.098.584	0.000110900978008477	0.00365147369081102

İlk kolon özelliği göstermektedir, Çizelge 4.5'de gen olarak belirtilmiştir. İkinci kolon gen ismini belirmektedir. Üçüncü, dördüncü ve beşinci kolonlar sırasıyla fold change, p değerini ve q değerini göstermektedir.

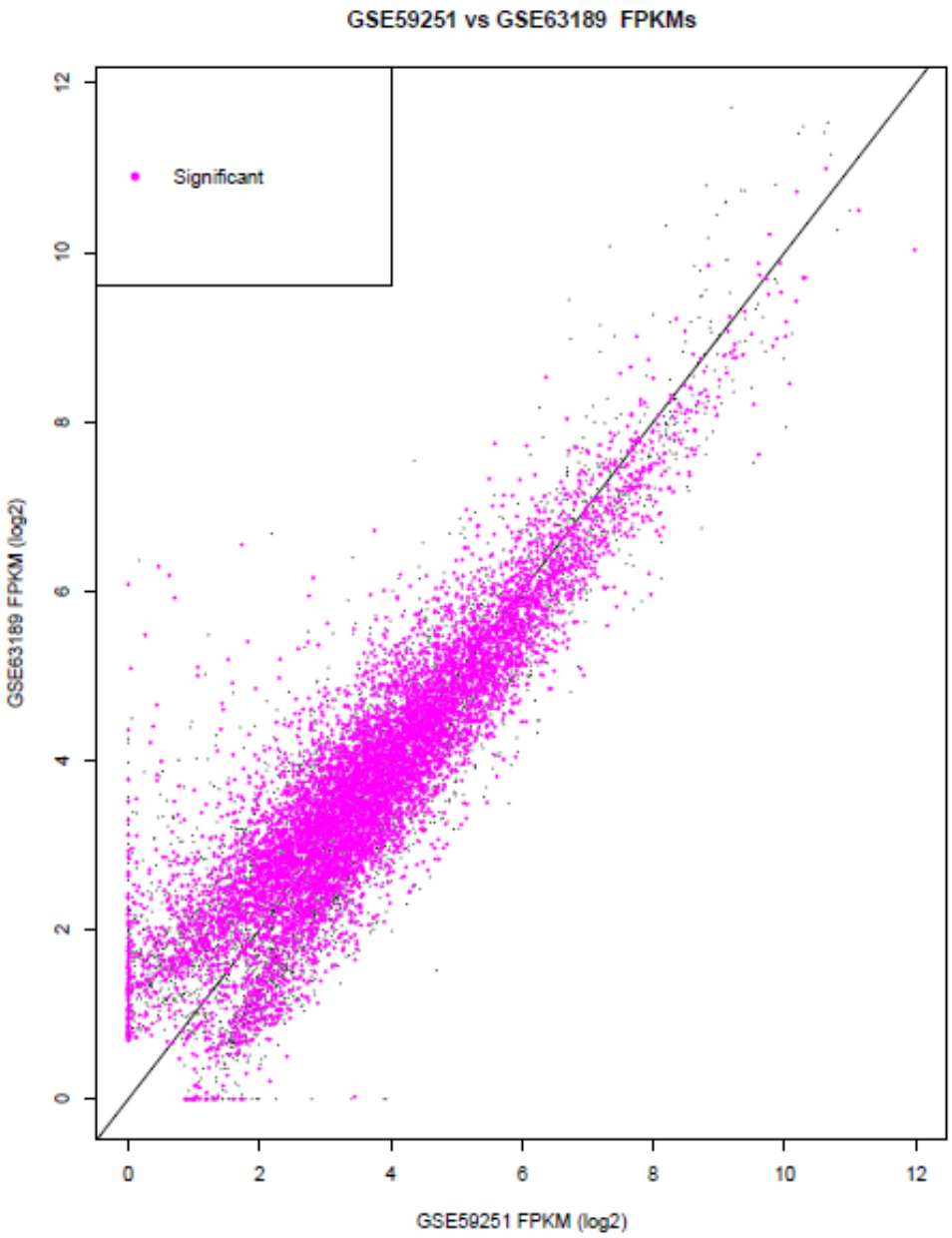


Şekil 4.6 Transkript sayısının gen başına dağılımı. Yatay eksen gen başına düşen transkript sayısını göstermekte, dikey eksen ise gen başına düşen transkript sıklığını göstermektedir.



Şekil 4.7 FPKM değerlerinin 6 örnek arasında dağılımı. Aynı renkte olanlar aynı dataseti gösterir; kırmızı olanlar GSE59251'i, gri olanlar ise GSE63189'u göstermektedir.

Yukarıda(şekil 4.6) gen bolluğunun (FPKM değerleri olarak ölçüldü), örnekler arasında dağılımı gösterilmiştir. FPKM (Fragments Per Kilobase Million): Kilobaz Milyon Başına Parça Sayısı anlamına gelir.



Şekil 4.8 GSE59251 ve GSE63189 datasetlerinin ifade değerlerinin gösterimi. Diferansiyel ifade farklılıkları mor renkte gösterilmektedir.

4.2.2 A549 kanser hücre hattı

Ballgown ve R kullanılarak A549 kanser hücre hattına ait GSE80182 dataseti ve GSE136105 dataseti diferansiyel gen ekspresyon için karşılaştırıldığında, analiz edilen 8.884 gen içinde 785 tanesinde istatiksel olarak ($p<0.05$) diferansiyel gen ifadesi tespit edildi. Toplamda 22.966

transkript arasından 2.903 tanesi datasetler arasında farklı ifade edilen transkriptlerin sayısı olarak kaydedildi.

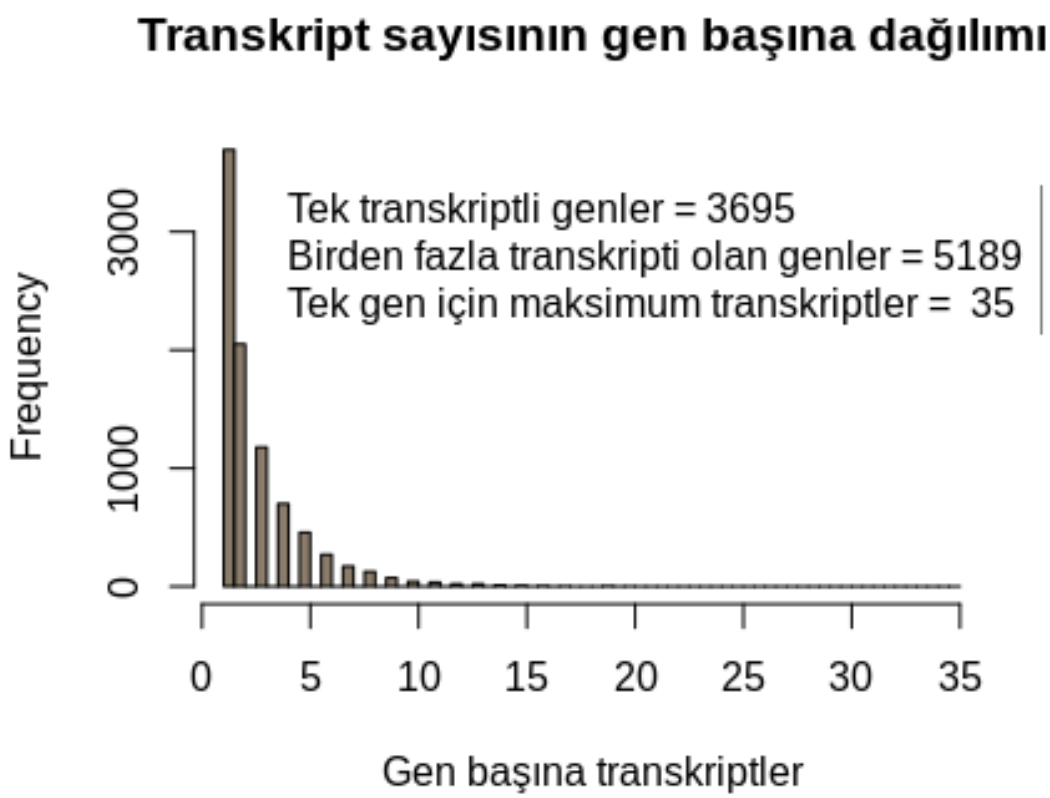
Çizelge 4.6: A549 kanser hücre hattı datasetleri arasında farklı ifade edilen transkriptler ($p<0.05$).

geneNames	geneIDs	feature	id	fc	pval	qval
RN7SK	MSTRG.24093	transcript	200370	978.371.725.103.834	0.000104279681834685	0.0353442226620906
.	MSTRG.24917	transcript	206352	210.367.376.914.259	0.00010585006613506	0.0353442226620906
KRT7	MSTRG.6245	transcript	49994	0.000897353940259361	0.000107495067337204	0.0353442226620906
HIST1H2BN	MSTRG.23650	transcript	196871	390.386.582.084.068	0.000107728624329284	0.0353442226620906
KLB	MSTRG.20970	transcript	175818	57.825.073.248.225	0.000110312678911595	0.0356822673786435
HIST1H2AB	MSTRG.23546	transcript	196563	17.114.412.876.869	0.000112799637436156	0.0359799510188717
PECR	MSTRG.17087	transcript	141426	695.888.638.874.064	0.000116929135285604	0.0364981889004694
COL6A1	MSTRG.18417	transcript	153064	0.301152365836395	0.000119002639481947	0.0364981889004694
ATF7	MSTRG.6282	transcript	50465	548.279.422.422.989	0.000119192030285431	0.0364981889004694
GPR37	MSTRG.25882	transcript	214904	612.041.038.712.215	0.000124012893497394	0.037474738316594
NID1	MSTRG.2690	transcript	20621	0.153925341345978	0.000126117129300551	0.0376156622274865
HIST1H4B	MSTRG.23544	transcript	196561	181.505.111.768.246	0.000134358805149448	0.0393836208839482
HIST2H3C	MSTRG.1697	transcript	12457	621.552.909.427.773	0.000136788178343505	0.0393836208839482
HIST2H3A	MSTRG.1702	transcript	12464	680.918.328.108.039	0.000138780509354897	0.0393836208839482
SCARNA2	MSTRG.1390	transcript	10397	93.700.469.683.439	0.00014059314559467	0.0393836208839482
HIST1H3C	MSTRG.23548	transcript	196566	172.140.305.787.968	0.000140619041734902	0.0393836208839482
HIST1H4A	MSTRG.23543	transcript	196560	164.794.736.786.965	0.000145471354427773	0.0402517485034727
TXLNGY	MSTRG.29687	transcript	244328	0.0940519264824822	0.000149564301047622	0.0408915921173773
PUDP	MSTRG.28662	transcript	236531	0.0650296067874428	0.000160413872103193	0.0431655002561247
SNORD17	MSTRG.17515	transcript	145441	863.387.061.713.268	0.000161640382392525	0.0431655002561247
CYB5A	MSTRG.13804	transcript	111428	0.210177037635849	0.000163547992392266	0.0431729102675951
DACT2	MSTRG.24872	transcript	205920	100.214.880.651.325	0.000166497421019263	0.0434520428537317
HIST2H2BF	MSTRG.1693	transcript	12450	890.447.245.971.556	0.000172434343451044	0.0444958104685021
HIST1H4I	MSTRG.23605	transcript	196778	286.235.303.172.976	0.00017536974509258	0.0447504618421798
CAVIN2	MSTRG.16914	transcript	139843	0.0848761999504641	0.000189424863879295	0.0478058398225482

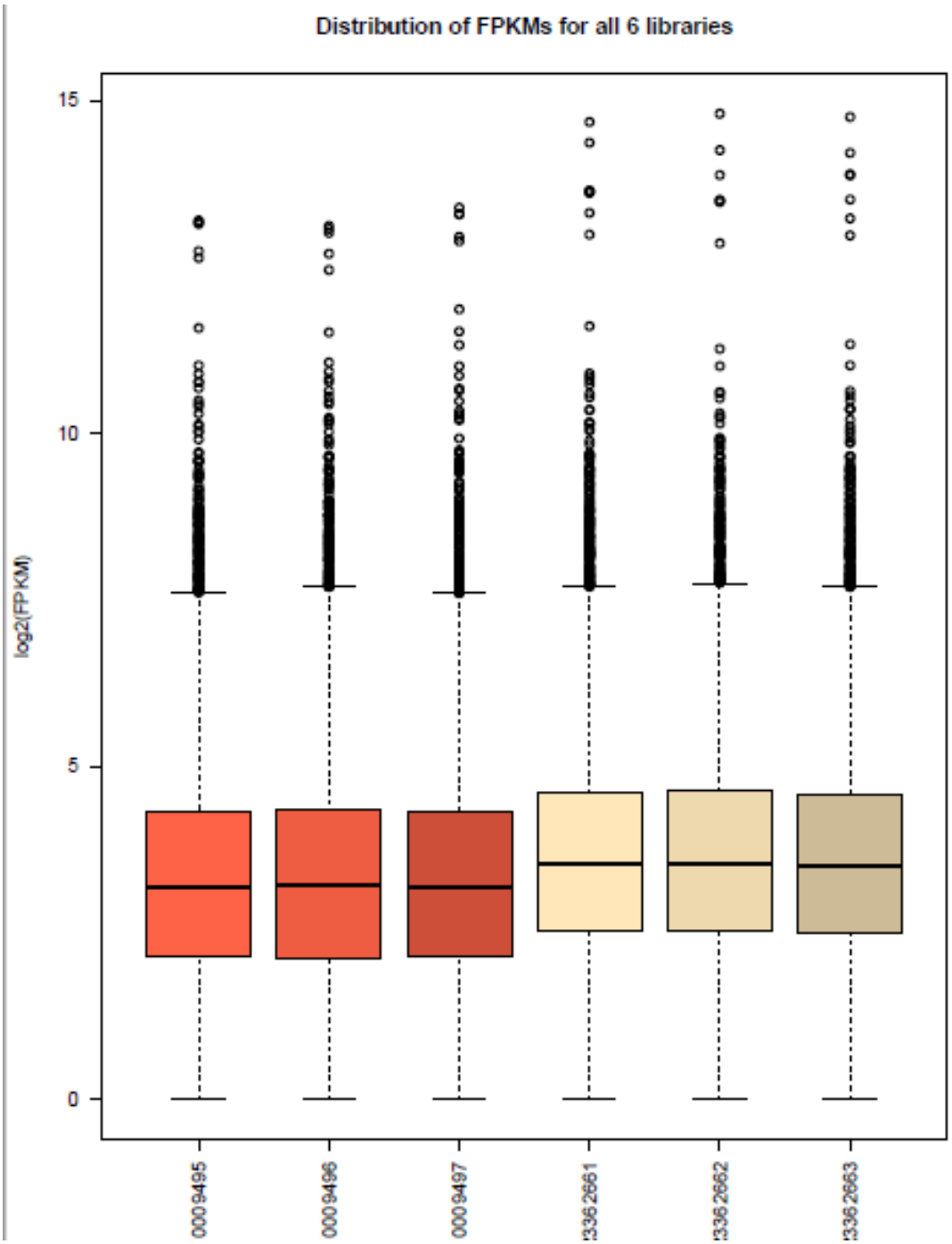
fc(fold change): GSE136105 datasetinin datasetine GSE80182 karşı aralarındaki ekspresyon oranını gösterir; dolayısıyla 1'in altındaki değerler transkriptin GSE80182'de daha düşük bir seviyede ifade edildiği anlamına gelir.

Çizelge 4.7: A549 kanser hücre hattı datasetleri arasında farklı ifade edilen genler (p<0.05).

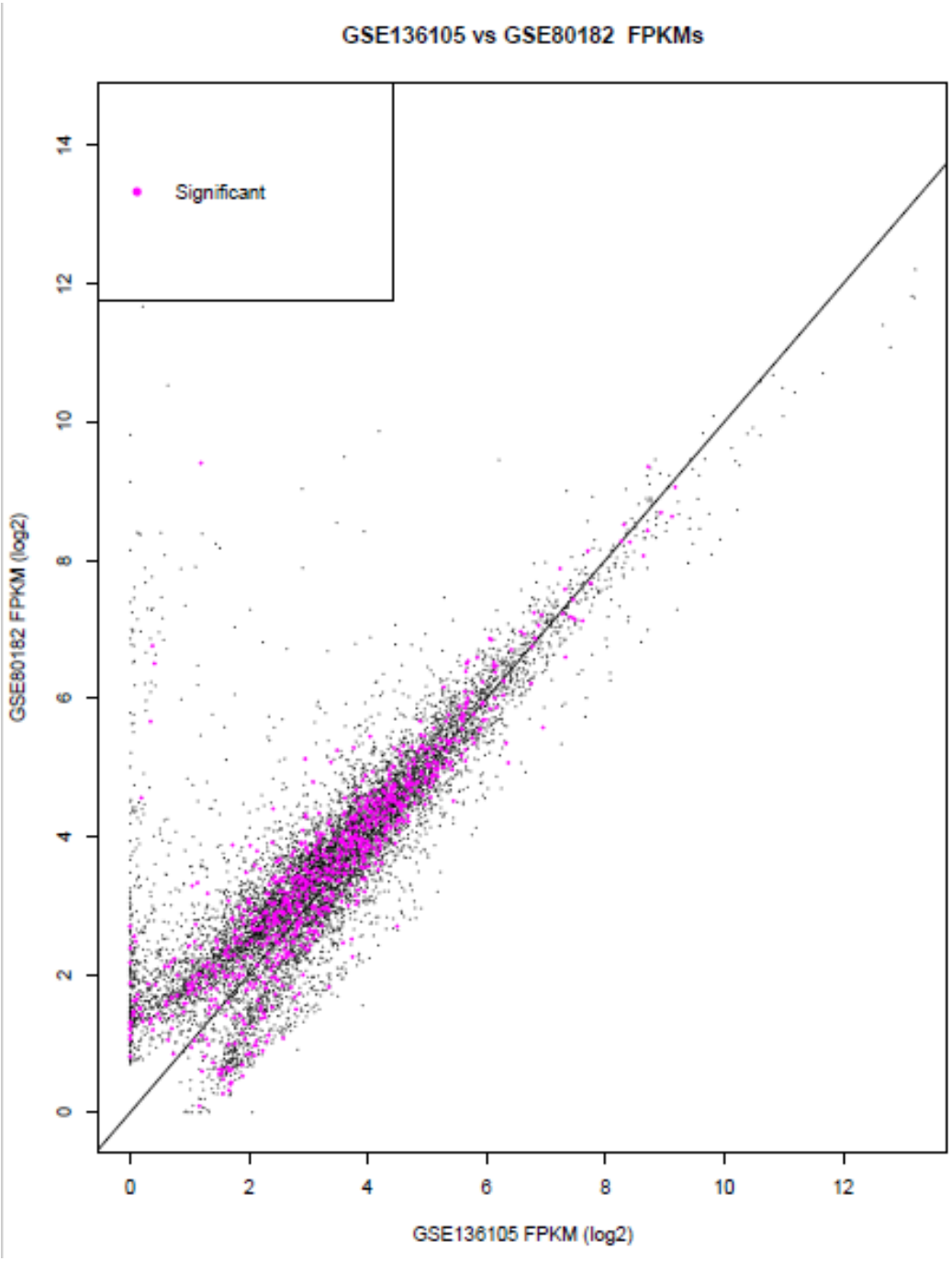
featu	id	fc	pval	qval
gene	MSTRG.17458	0.0376922772948471	0.000121120041078449	0.119558938326771
gene	MSTRG.9315	350.781.760.431.334	0.000156174534049036	0.132782001448678
gene	MSTRG.23599	227.162.339.901.135	0.000164408151275941	0.132782001448678
gene	MSTRG.9319	309.855.226.699.368	0.000186832404878046	0.138318257078046
gene	MSTRG.6631	217.841.739.143.898	0.000210987123280604	0.144185354094222
gene	MSTRG.21842	0.00971630726128401	0.000258985243792531	0.153506711748205
gene	MSTRG.19811	291.185.861.425.188	0.000292041447542468	0.153506711748205
gene	MSTRG.2565	260.696.793.713.488	0.000312671768433526	0.153506711748205
gene	MSTRG.29546	0.0679648530923465	0.00031302564433755	0.153506711748205
gene	MSTRG.25588	192.073.175.697.767	0.000319190373487821	0.153506711748205
gene	MSTRG.28546	818.816.877.516.105	0.000328301162000888	0.153506711748205
gene	MSTRG.23556	325.235.420.700.743	0.000353011519770297	0.156807717081966
gene	MSTRG.9145	287.551.487.010.545	0.000440508266628115	0.178179654345182
gene	MSTRG.16295	903.302.909.512.042	0.000441237325033095	0.178179654345182
gene	MSTRG.16869	112.934.174.780.532	0.000480226937723427	0.184777633908574
gene	MSTRG.24252	779.420.616.513.398	0.000505191893407142	0.184777633908574
gene	MSTRG.23654	947.978.486.082.097	0.000523480167512025	0.184777633908574
gene	MSTRG.1001	0.211826931690338	0.000551344506385831	0.184777633908574
gene	MSTRG.28550	739.438.168.880.741	0.000605072441583254	0.184777633908574
gene	MSTRG.23547	205.326.478.888.972	0.00060784821772697	0.184777633908574
gene	MSTRG.8623	36.285.797.632.111	0.000621161054680908	0.184777633908574
gene	MSTRG.28368	0.0926089561122997	0.00062396769667461	0.184777633908574
gene	MSTRG.1559	137.349.580.381.732	0.00069766642287028	0.189501001835205
gene	MSTRG.23568	685.470.396.876.976	0.000733851619057369	0.189501001835205
gene	MSTRG.19742	0.491841800949464	0.000755538089846519	0.189501001835205



Şekil 4.9 Transkript sayısının gen başına dağılımı. Yatay eksen gen başına düşen transkript sayısını göstermekte, dikey eksen ise gen başına düşen transkript sıklığını göstermektedir.



Şekil 4.10 FPKM değerlerinin 6 örnek arasında dağılımı. Aynı renkte olanlar aynı dataseti gösterir; kırmızı olanlar GSE80182'yi, gri olanlar ise GSE136105'i göstermektedir.



Şekil 4.11 GSE80182 ve GSE136105 datasetlerinin ifade değerlerinin gösterimi. Diferansiyel ifade farklılıkları mor renkte gösterilmektedir.

4.2.3 HCT116 kanser hücre hattı

Ballgown ve R kullanılarak HCT116 kanser hücre hattına ait GSE120071 dataseti ve GSE131249 dataseti diferansiyel gen ekspresyon için karşılaştırıldığında, analiz edilen 10.796 gen içinde 1.781 tanesinde istatiksel olarak ($p<0.05$) diferansiyel gen ifadesi tespit edildi.

Toplamda 26.608 transkript arasından 1.925 tanesi datasetler arasında farklı ifade edilen transkriptlerin sayısı olarak kaydedildi ($p<0.05$).

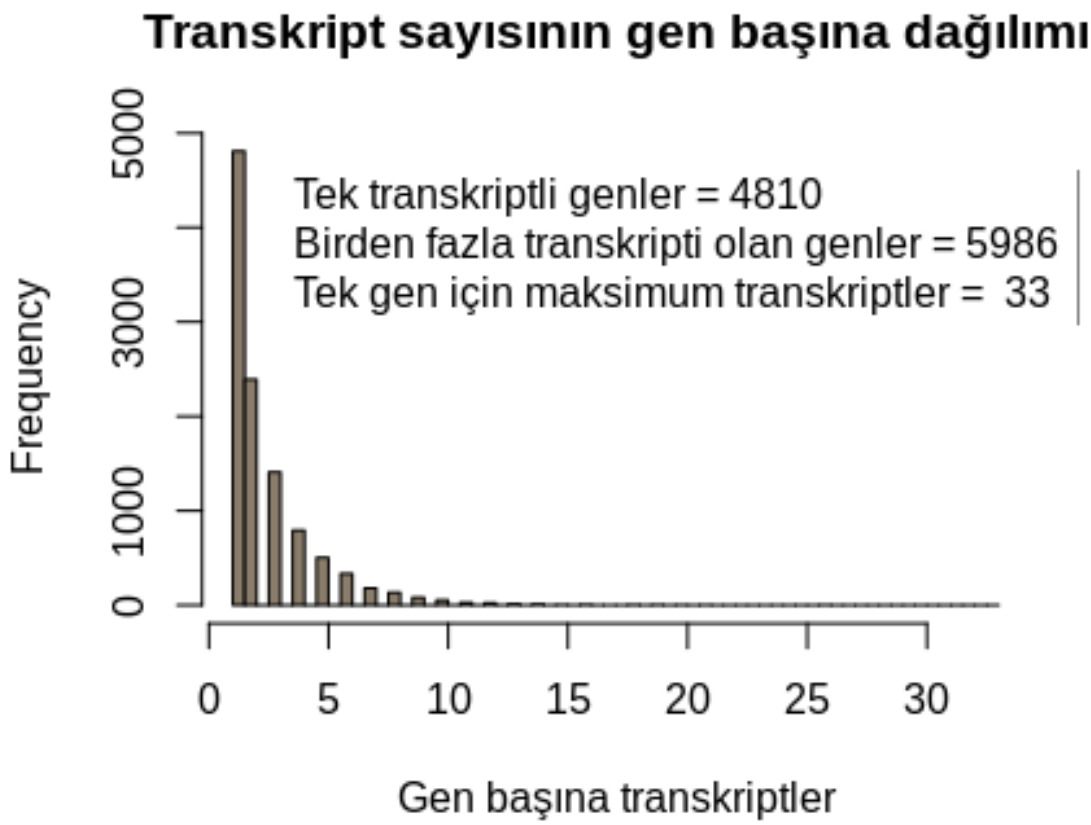
Çizelge 4.8: HCT116 kanser hücre hattı datasetleri arasında farklı ifade edilen transkriptler ($p<0.05$).

geneNames	geneIDs	feature	id	fc	pval	qval
RN7SL396P	MSTRG.26537	transcript	226789	0.000150916873010428	0.000103254313960299	0.159058313001552
.	MSTRG.10360	transcript	86469	0.00739226938149611	0.000109454434126022	0.159058313001552
MON1A	MSTRG.19187	transcript	164533	472.992.066.097.383	0.000110278319763735	0.159058313001552
CIRBP	MSTRG.13572	transcript	113043	831.984.442.901.906	0.000116995718023594	0.159058313001552
GPI	MSTRG.14309	transcript	119764	2,50E+09	0.00011956502485666	0.159058313001552
WASH5P	MSTRG.13515	transcript	112519	0.0054887711343726	0.000119916974963719	0.159058313001552
RPS27	MSTRG.1855	transcript	13832	153.732.010.801.588	0.000122767550832403	0.159058313001552
HSPA13	MSTRG.17713	transcript	151613	158.252.046.298.217	0.000124878139626694	0.159058313001552
.	MSTRG.6397	transcript	52878	8,57E+09	0.000125534597603449	0.159058313001552
RPL18	MSTRG.14732	transcript	124023	802.568.553.130.455	0.000134942552528772	0.163206883531162
MED13	MSTRG.12526	transcript	103629	0.0493911692823078	0.000152191083095277	0.170645665068268
DPM1	MSTRG.17495	transcript	149833	37.652.063.126.578	0.000166607079634451	0.170645665068268
EIF2S3	MSTRG.28174	transcript	239074	0.01036101226758	0.00016851518844696	0.170645665068268
.	MSTRG.17126	transcript	146489	0.0140868915018088	0.000172909680360211	0.170645665068268
ANXA1	MSTRG.27237	transcript	232507	344.310.920.512.454	0.000173159687193447	0.170645665068268
TMEM140	MSTRG.25391	transcript	217278	159.765.934.415.726	0.000185827429453278	0.172156259121301
MBTPS1	MSTRG.10968	transcript	91665	47.956.925.718.702	0.000187632723786746	0.172156259121301
TXNDC11	MSTRG.10028	transcript	83856	825.080.473.400.684	0.000219661502895097	0.194825108967758
TRIM44	MSTRG.4431	transcript	34509	0.403859100016661	0.000247410404785375	0.208218617073824
ATP5MC3	MSTRG.16412	transcript	139851	322.262.080.530.588	0.000252597163130197	0.208218617073824
MBOAT7	MSTRG.14921	transcript	126049	37.828.924.228.718	0.000278929527236738	0.208218617073824
TWSG1	MSTRG.13083	transcript	108637	151.798.526.864.438	0.000280127704637811	0.208218617073824
.	MSTRG.14338	transcript	120016	408.653.369.659.877	0.000286249637781566	0.208218617073824
.	MSTRG.27978	transcript	237946	0.000727377166610704	0.000290230760365717	0.208218617073824
PKM	MSTRG.9281	transcript	77433	103.956.731.199.647	0.000301311463318354	0.208218617073824
CSRP1	MSTRG.2345	transcript	17948	299.676.164.944.984	0.000306462610334401	0.208218617073824

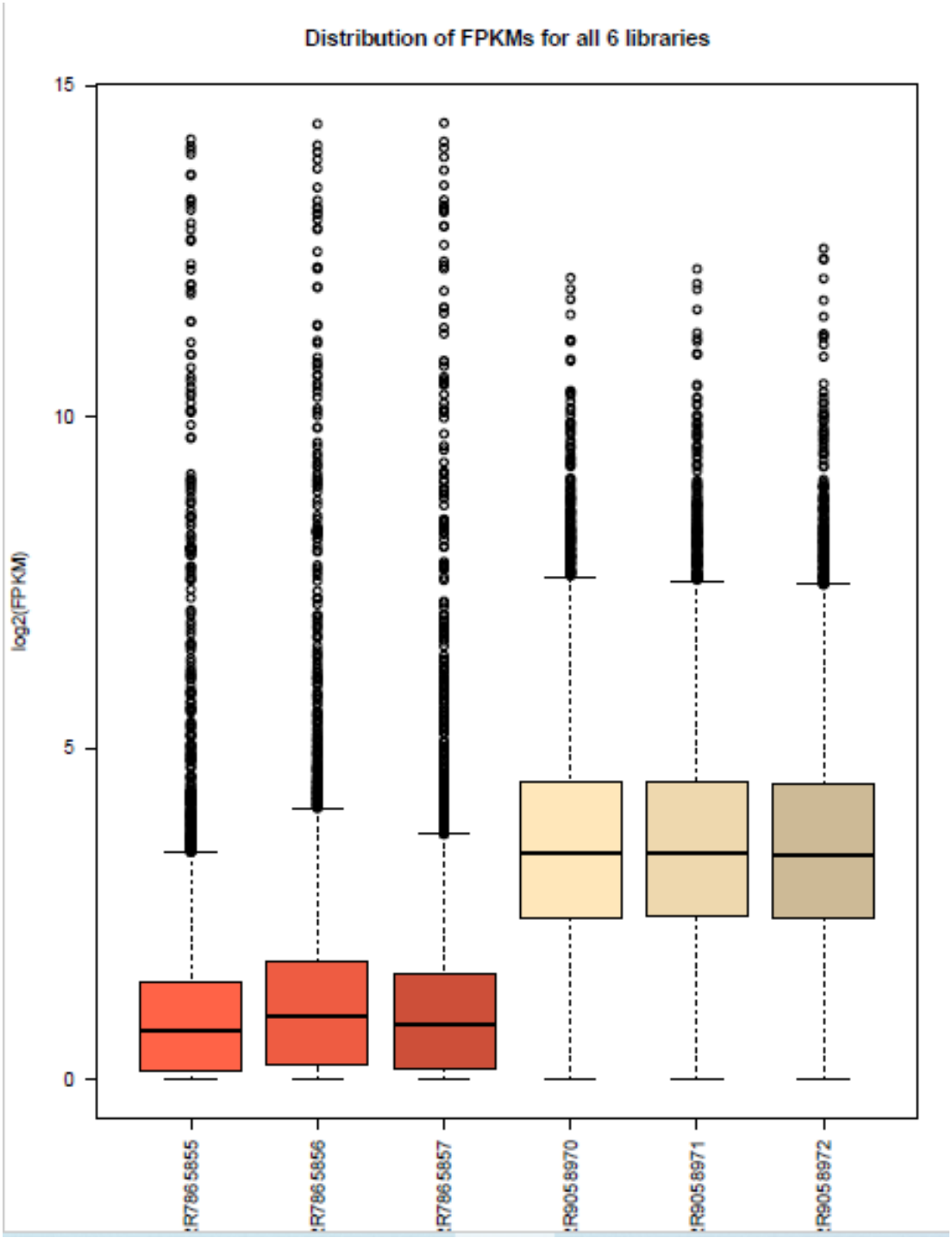
fc(fold change): GSE131249 datasetinin GSE120071 datasetine karşı aralarındaki ekspresyon oranını gösterir; dolayısıyla 1'in altındaki değerler transkriptin GSE120071'de daha düşük bir seviyede ifade edildiği anlamına gelir.

Çizelge 4.9: HCT116 kanser hücre hattı datasetleri arasında farklı ifade edilen genler ($p<0.05$).

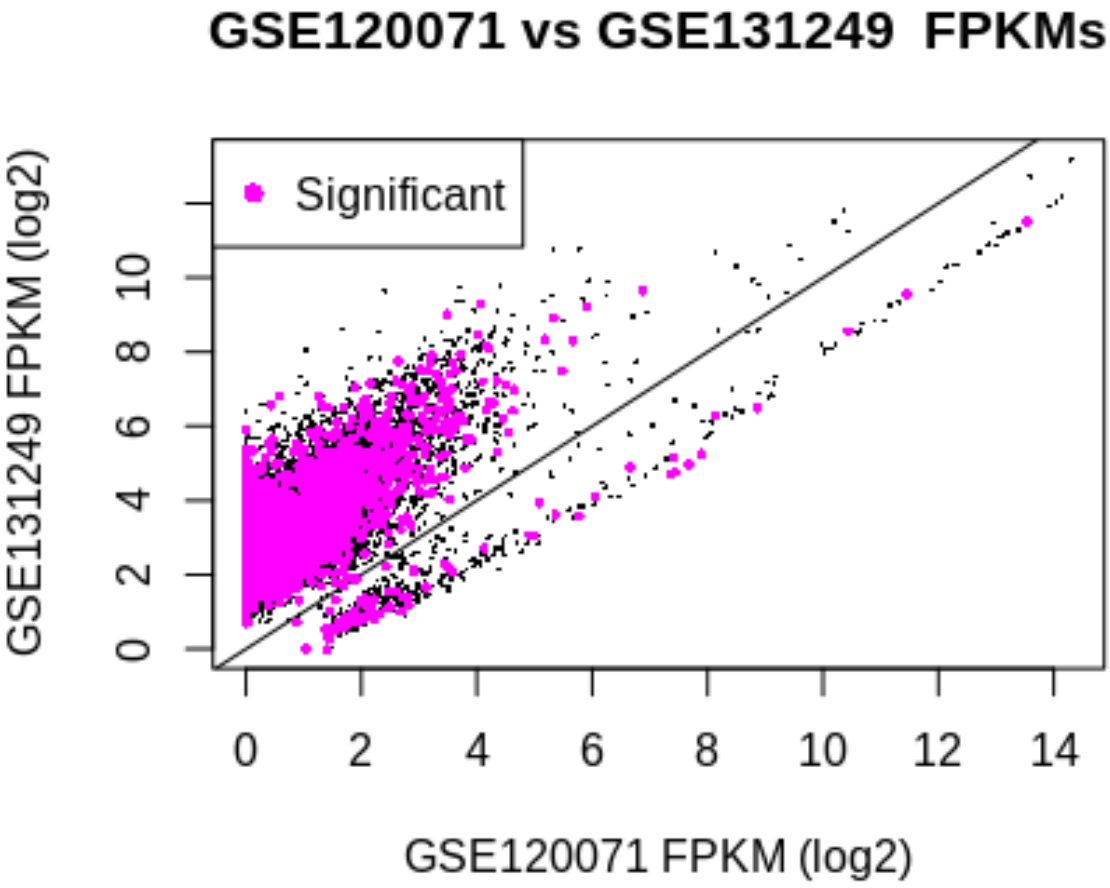
featu	id	fc	pval	qval
gene	MSTRG.9947	942.170.777.204.932	0.000111766386506673	0.0879861034685099
gene	MSTRG.10028	841.332.926.278.695	0.000118520453179283	0.0879861034685099
gene	MSTRG.27883	923.908.338.201.971	0.000122248198594632	0.0879861034685099
gene	MSTRG.15022	0.015231094114016	0.000135909997124761	0.0917052705599327
gene	MSTRG.10637	391.227.708.943.542	0.000156279511619228	0.0956343073406351
gene	MSTRG.13083	156.192.637.751.076	0.00015944956762981	0.0956343073406351
gene	MSTRG.2345	290.253.743.845.239	0.000173806185926217	0.098758504382076
gene	MSTRG.4157	0.00188016763525691	0.000199538711213032	0.107710996312795
gene	MSTRG.13740	0.0877438030426407	0.000209527386827002	0.107717031818301
gene	MSTRG.28398	0.000619610078872746	0.000278592421049551	0.119644552416924
gene	MSTRG.18285	484.978.413.627.798	0.000279533128027309	0.119644552416924
gene	MSTRG.25184	0.000561347519921785	0.000280952321955685	0.119644552416924
gene	MSTRG.11359	470.244.481.943.251	0.000312041575426525	0.119644552416924
gene	MSTRG.5959	155.107.250.307.086	0.000315651881441559	0.119644552416924
gene	MSTRG.13659	462.789.242.616.712	0.000316789519596572	0.119644552416924
gene	MSTRG.15095	194.599.785.794.028	0.000330033275434127	0.119644552416924
gene	MSTRG.4613	119.971.520.891.548	0.000342047586665006	0.119644552416924
gene	MSTRG.4150	0.00565010600620987	0.000346806744593753	0.119644552416924
gene	MSTRG.3299	102.869.149.538.029	0.000347661912576847	0.119644552416924
gene	MSTRG.24342	24.980.342.203.165	0.000354633723355091	0.119644552416924
gene	MSTRG.28958	0.0128171093912896	0.000397061721727709	0.129632828974264
gene	MSTRG.11369	154.641.462.568.689	0.000414907704113454	0.129632828974264
gene	MSTRG.5262	74.928.879.304.241	0.00050461776527233	0.129632828974264
gene	MSTRG.4567	0.185886792369327	0.000535933862877203	0.129632828974264
gene	MSTRG.16717	18.983.361.527.628	0.000541801016427867	0.129632828974264



Şekil 4.12 Transkript sayısının gen başına dağılımı. Yatay eksen gen başına düşen transkript sayısını göstermekte, dikey eksen ise gen başına düşen transkript sıklığını göstermektedir.



Şekil 4.13 FPKM değerlerinin 6 örnek arasında dağılımı. Aynı renkte olanlar aynı dataseti gösterir; kırmızı olanlar GSE120071'i, gri olanlar ise GSE131249'u göstermektedir.



Şekil 4.14 GSE12071 ve GSE131249 datasetlerinin ifade değerlerinin gösterimi. Diferansiyel ifade farklılıkları mor renkte gösterilmektedir.

4.2.4 HeLa kanser hücre hattı

Ballgown ve R kullanılarak HeLa kanser hücre hattına ait GSE75410 dataseti ve GSE77913 dataseti diferansiyel gen ekspresyon için karşılaştırıldığında, analiz edilen 16.742 gen içinde 6.031 tanesinde istatiksel olarak ($p<0.05$) diferansiyel gen ifadesi tespit edildi. Toplamda 27.119 transkript arasından 2.813 tanesi datasetler arasında farklı ifade edilen transkriptlerin sayısı olarak kaydedildi ($p<0.05$).

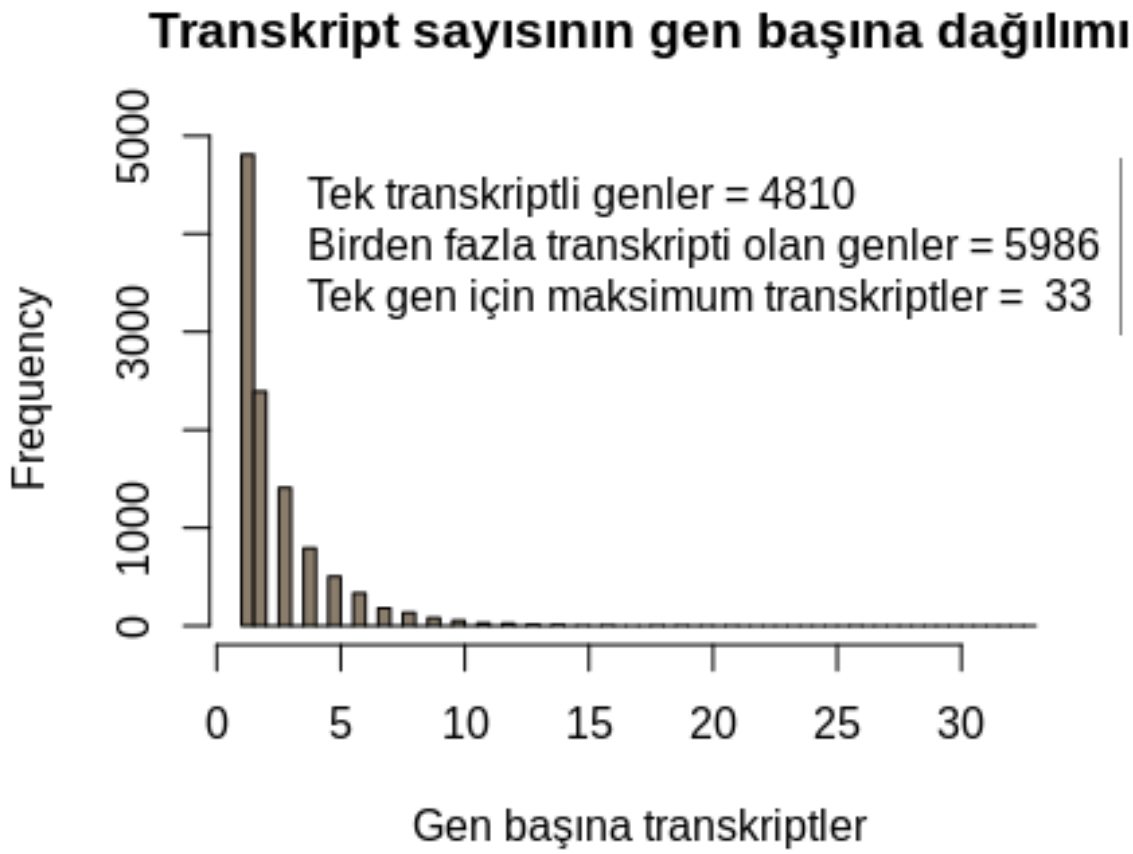
Çizelge 4.10: HeLa kanser hücre hattı datasetleri arasında farklı ifade edilen transkriptler ($p<0.05$).

geneNames	geneIDs	feature	id	fc	pval	qval
MOSPD3	MSTRG.36722	transcript	2E+05	179.056.849.803.328	0.000105610257148392	0.339192175459994
.	MSTRG.18158	transcript	1E+05	0.0922646244892218	0.000105979339048767	0.339192175459994
.	MSTRG.4574	transcript	24741	0.216125204569764	0.000108079416974549	0.339192175459994
TPGS2	MSTRG.18461	transcript	1E+05	0.133280433882609	0.00010905913159398	0.339192175459994
.	MSTRG.9667	transcript	56718	117.163.372.980.727	0.000122778617860586	0.339192175459994
RIOK1	MSTRG.33230	transcript	2E+05	154.094.499.140.559	0.000165475592063924	0.339192175459994
.	MSTRG.42006	transcript	2E+05	0.338455693096636	0.000171688421308325	0.339192175459994
HIST1H3A	MSTRG.33468	transcript	2E+05	186.885.944.100.799	0.000172456477429694	0.339192175459994
.	MSTRG.15338	transcript	90964	970.062.357.339.144	0.000189708546077805	0.339192175459994
.	MSTRG.30683	transcript	2E+05	781.606.367.816.024	0.000196580403452717	0.339192175459994
GTF3C2	MSTRG.20849	transcript	1E+05	107.385.608.918.665	0.000216675273362998	0.339192175459994
CFAP298	MSTRG.25067	transcript	2E+05	109.604.405.755.051	0.000255273340239448	0.339192175459994
TTLL12	MSTRG.26034	transcript	2E+05	787.494.664.682.714	0.000257442919036732	0.339192175459994
EEF1B2	MSTRG.23251	transcript	1E+05	245.510.701.322.293	0.000266237158770832	0.339192175459994
ANKRD33B	MSTRG.30830	transcript	2E+05	434.607.533.580.382	0.000299070749872898	0.339192175459994
CDH2	MSTRG.18373	transcript	1E+05	0.0222197203292188	0.000334138831140751	0.339192175459994
PI4K2A	MSTRG.5315	transcript	28393	0.592807903583461	0.000401025292124824	0.339192175459994
MAT2B	MSTRG.32825	transcript	2E+05	43.649.087.299.548	0.000411936198848828	0.339192175459994
TRAF7	MSTRG.14231	transcript	83215	942.380.146.466.013	0.000446070300068713	0.339192175459994
ZDHHC2	MSTRG.37768	transcript	2E+05	0.0820776236189987	0.000466066034215062	0.339192175459994
NDUFA11	MSTRG.19109	transcript	1E+05	695.591.183.294.901	0.000480495427089056	0.339192175459994
HMGB1P10	MSTRG.25601	transcript	2E+05	422.077.323.387.439	0.000484128444279253	0.339192175459994
SPIRE2	MSTRG.15747	transcript	93483	295.584.434.291.544	0.00048493908512437	0.339192175459994
MRPL14	MSTRG.34061	transcript	2E+05	296.426.756.627.644	0.000495149397181316	0.339192175459994
EFNA5	MSTRG.31953	transcript	2E+05	0.226615731592291	0.000496899670220596	0.339192175459994
.	MSTRG.23638	transcript	1E+05	507.266.987.669.484	0.00052753679475237	0.339192175459994
PPP1R3D	MSTRG.24688	transcript	2E+05	914.617.360.709.072	0.000541697977669342	0.339192175459994
.	MSTRG.11239	transcript	65744	340.344.191.528.774	0.000587266702803735	0.339192175459994
TEX2	MSTRG.17502	transcript	1E+05	644.675.707.583.053	0.000600583243000385	0.339192175459994
.	MSTRG.5071	transcript	27180	0.134517187438127	0.00064194536206863	0.339192175459994
DDX23	MSTRG.8648	transcript	49945	58.263.558.939.291	0.000660459052737594	0.339192175459994

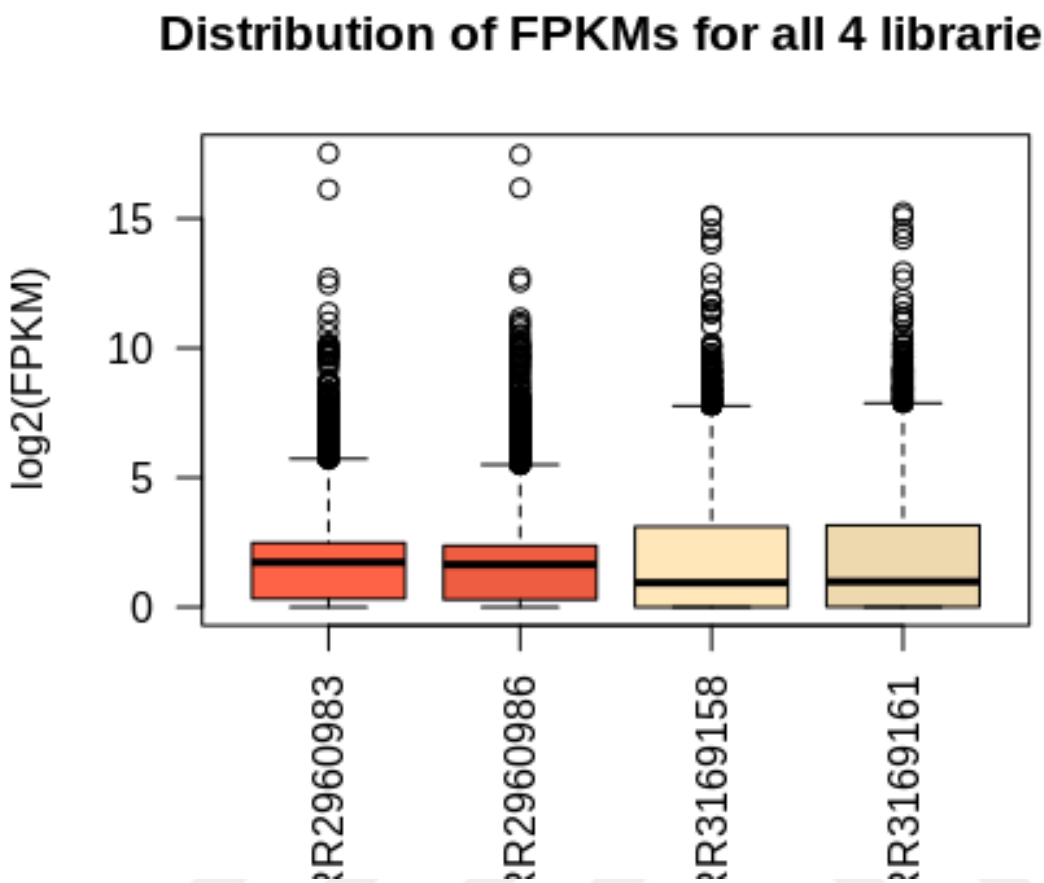
fc(fold change): GSE77913 datasetinin GSE75410 datasetine karşı aralarındaki ekspresyon oranını gösterir; dolayısıyla 1'in altındaki değerler transkriptin GSE75410'de daha düşük bir seviyede ifade edildiği anlamına gelir.

Çizelge 4.11: HeLa kanser hücre hattı datasetleri arasında farklı ifade edilen genler (p<0.05).

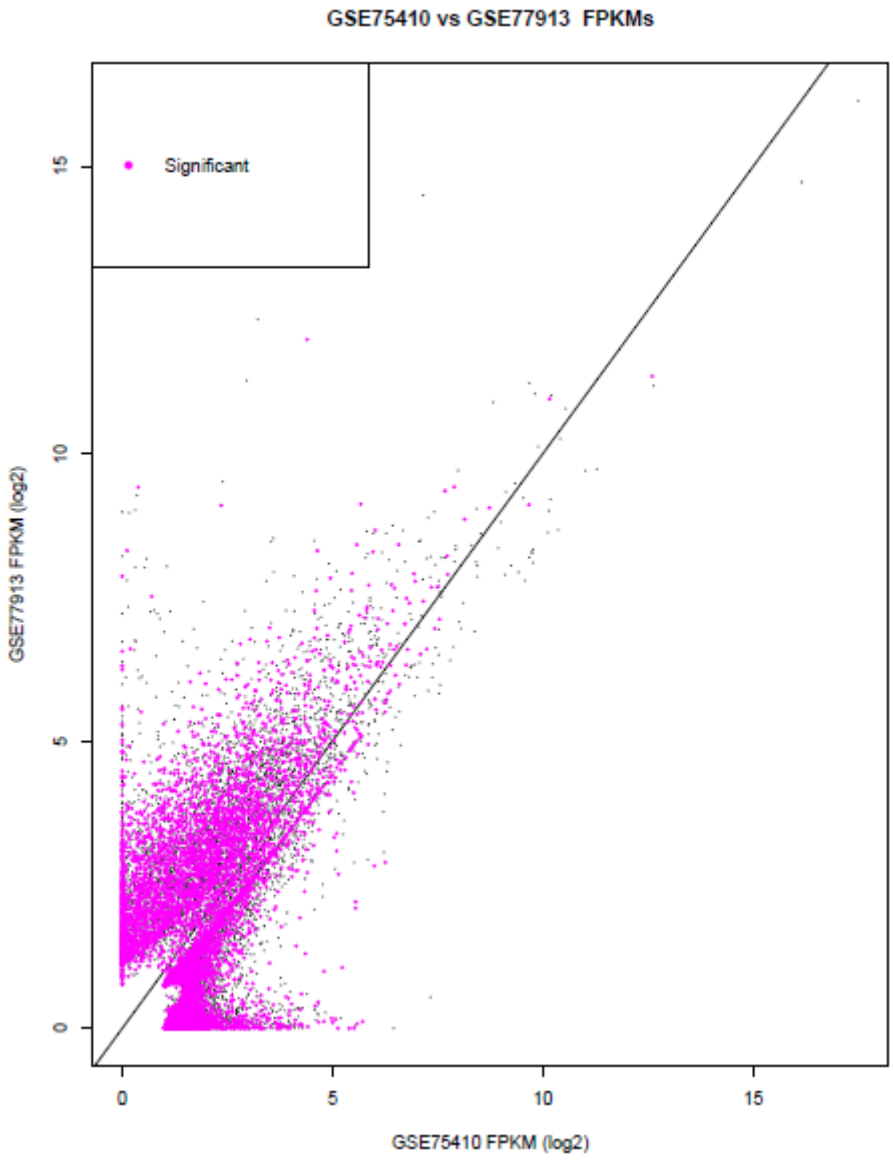
featu	id	fc	pval	qval
gene	MSTRG.30830	434.548.845.672.938	0.000110231125488358	0.114704079402141
gene	MSTRG.6769	345.072.841.666.699	0.000119832482983195	0.114704079402141
gene	MSTRG.34409	0.251925031993651	0.000124199534296987	0.114704079402141
gene	MSTRG.2690	141.037.204.573.869	0.000126012588927527	0.114704079402141
gene	MSTRG.28215	0.146695888122599	0.00013085376355737	0.114704079402141
gene	MSTRG.15574	0.340664505574435	0.000133596359361832	0.114704079402141
gene	MSTRG.25677	571.658.412.692.379	0.000143853770573577	0.114704079402141
gene	MSTRG.21294	0.231741545223821	0.000164715008257876	0.114704079402141
gene	MSTRG.15747	295.632.181.825.486	0.000178695203678858	0.114704079402141
gene	MSTRG.17502	644.453.994.401.352	0.000221382784384438	0.114704079402141
gene	MSTRG.19942	384.170.927.964.416	0.000225305676752274	0.114704079402141
gene	MSTRG.31211	0.342028176561783	0.000226942772281191	0.114704079402141
gene	MSTRG.26974	886.685.445.076.998	0.000236889367062121	0.114704079402141
gene	MSTRG.36593	0.148700826132756	0.000252291113492564	0.114704079402141
gene	MSTRG.17843	706.526.083.786.578	0.000255159347851586	0.114704079402141
gene	MSTRG.19551	437.420.754.596.268	0.000257127234131405	0.114704079402141
gene	MSTRG.9794	431.118.781.653.346	0.00025866189077195	0.114704079402141
gene	MSTRG.35648	114.156.180.238.595	0.000259820030042546	0.114704079402141
gene	MSTRG.5034	0.398519304009567	0.000261962000791427	0.114704079402141
gene	MSTRG.19509	721.637.973.718.425	0.00027608794894296	0.114704079402141
gene	MSTRG.18937	213.359.298.218.341	0.000278113634841759	0.114704079402141
gene	MSTRG.16930	349.782.434.762.233	0.000278540300380214	0.114704079402141
gene	MSTRG.36723	522.765.808.274.856	0.0002828557206771	0.114704079402141
gene	MSTRG.14243	777.144.343.216.215	0.000284410024747439	0.114704079402141
gene	MSTRG.22168	0.214511566349859	0.000301975487551398	0.114704079402141
gene	MSTRG.36298	483.615.470.528.382	0.000306841929800017	0.114704079402141



Şekil 4.15 Transkript sayısının gen başına dağılımı. Yatay eksen gen başına düşen transkript sayısını göstermekte, dikey eksen ise gen başına düşen transkript sıklığını göstermektedir.



Şekil 4.16 FPKM değerlerinin 4 örnek arasında dağılımı. Aynı renkte olanlar aynı dataseti gösterir; kırmızı olanlar GSE75410'nu, gri olanlar ise GSE77913'ü göstermektedir.



Şekil 4.17 GSE75410 ve GSE77913 datasetlerinin ifade değerlerinin gösterimi. Diferansiyel ifade farklılıkları mor renkte gösterilmektedir.

4.3 ALTERNATİF EKSON KULLANIMI

4.3.1 MCF7 kanser hücre hattı

DEXSeq ve R kullanılarak MCF7 kanser hücre hattı alternatif ekson kullanımı için analiz edildi. Analiz sonucunda, % 5'lik yanlış keşif oranı (false discovery rate) ile toplam 30.771 eksonik bölgenin 1.021 tanesinde alternatif ekson kullanımı tespit edildi. . İncelenen toplam 1.115 gen arasından 607 tanesi etkilenmiştir ($p<0.05$).

Çizelge 4.12 : MCF7 kanser hücre hattı DEXSeq deney tasarımı.

sample	condition
SRR1509730	GSE59251
SRR1509731	GSE59251
SRR1509732	GSE59251
SRR1648590	GSE63189
SRR1648591	GSE63189
SRR1648596	GSE63189

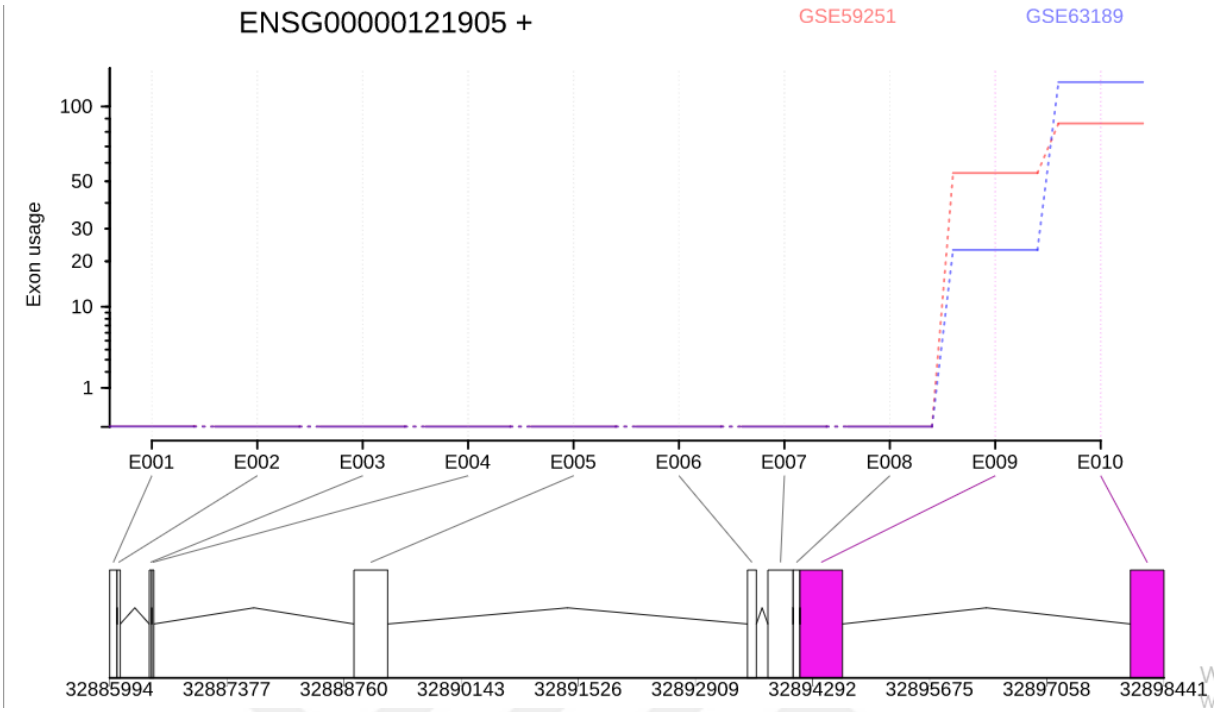
Çizelge 4.12'de, ilk kolon örnekleri, ikinci kolon ise koşulu belirtmektedir. Analizde koşul dataset olarak belirlenmiştir.

Çizelge 4.13: MCF7 kanser hücre hattı alternatif ekson kullanımı sonuç çizelgesi

testForDEU result table

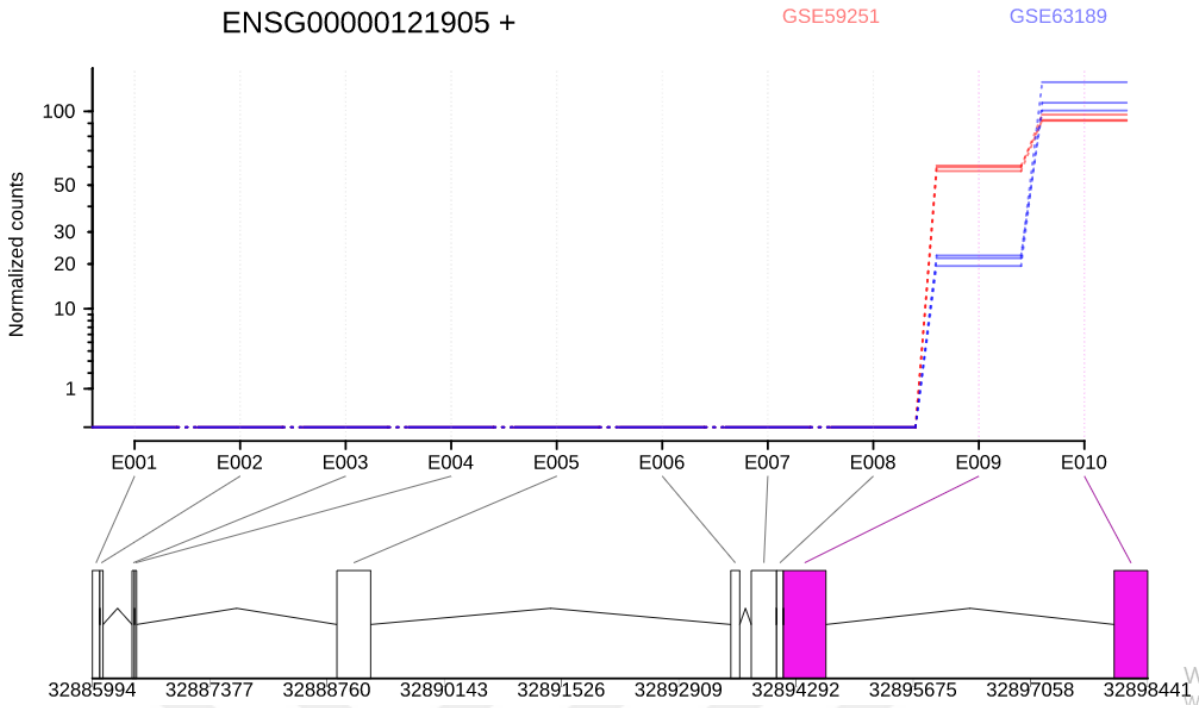
geneID	chr	start	end	total_exons	exon_changes
ENSG00000001630 ENSG00000001631 ENSG00000285772 ENSG00000240720 ENSG00000285953 ENSG00000243107	7	92112153	92246166	103	1
ENSG00000004478	12	2794970	2805423	26	1
ENSG00000004487	1	23019443	23083689	35	1
ENSG00000005187	16	20610243	20797581	47	1
ENSG00000005436	2	75652000	75710985	40	1
ENSG00000006634	7	87876216	87909553	29	1
ENSG00000006712	19	39385629	39391154	29	2
ENSG00000007341	1	112523514	112620825	66	1
ENSG00000007516	16	1333601	1349441	66	1
ENSG00000013441	2	200853009	200864744	43	1
ENSG00000014216	11	65180566	65212006	86	1
ENSG00000025770 ENSG00000272821	22	50508224	50524780	43	1
ENSG00000029364	14	69398015	69462390	28	1
ENSG00000033170	14	65410592	65744121	42	2
ENSG00000035403	10	73995193	74121363	35	1
ENSG00000041802	3	194640791	194672463	32	1
ENSG00000049768 ENSG00000286181	X	49250436	49270521	36	3
ENSG00000049860	5	74640023	74722647	38	1
ENSG00000062716 ENSG00000284190	17	59707192	59842255	49	1
ENSG00000063046	12	53006158	53042209	46	10
ENSG00000063244	19	55654146	55674715	29	2
ENSG00000064607	19	18990888	19034023	51	1
ENSG00000064652	5	122843439	123029354	33	1
ENSG00000065427	16	75627474	75648643	40	1

Yukarıdaki Çizelgeda(Çizelge 4.13); ilk kolon gen isimlerini, ikinci kolon kromozom isimlerini, üçüncü kolon transkript başlama bölgesini, dördüncü kolon transkript bitiş bölgesini, beşinci kolon genin toplamda kaç eksonu olduğunu, son kolon ise ekson değişimini göstermektedir.

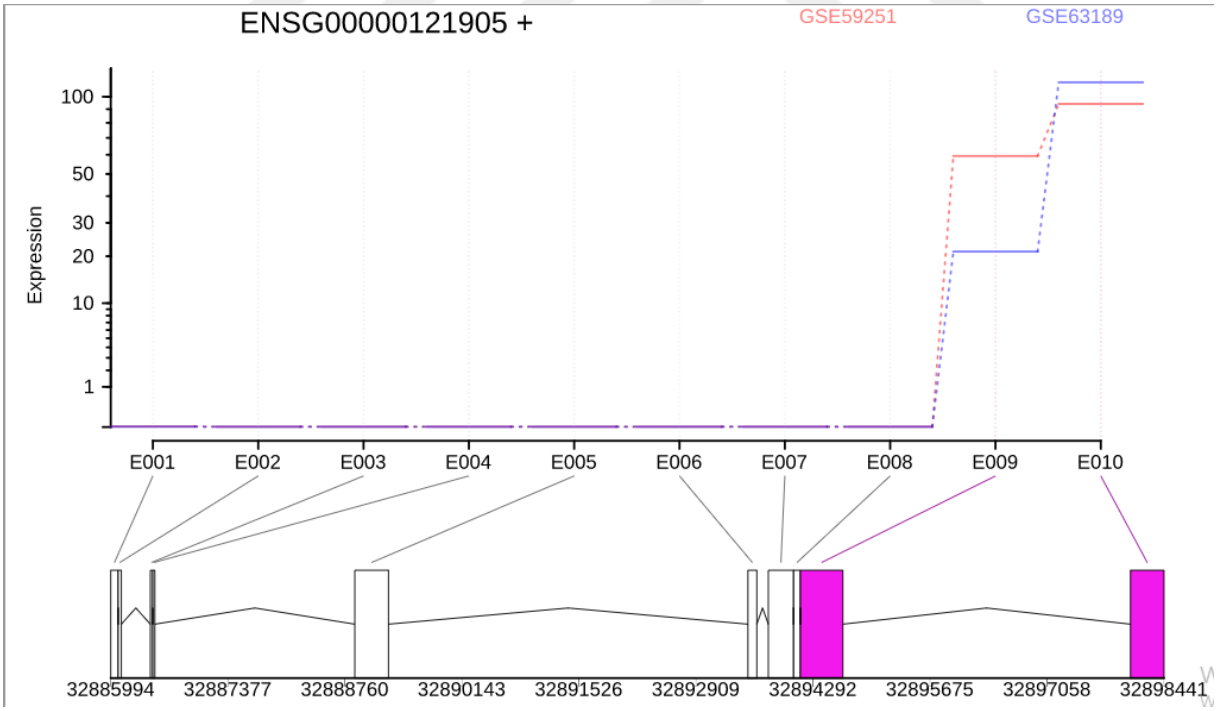


Şekil 4.18 Ekson kullanım grafiği. Mor renkte gösterilen, anlamlı bir şekilde diferansiyel ekson kullanımı gösteren ekson.

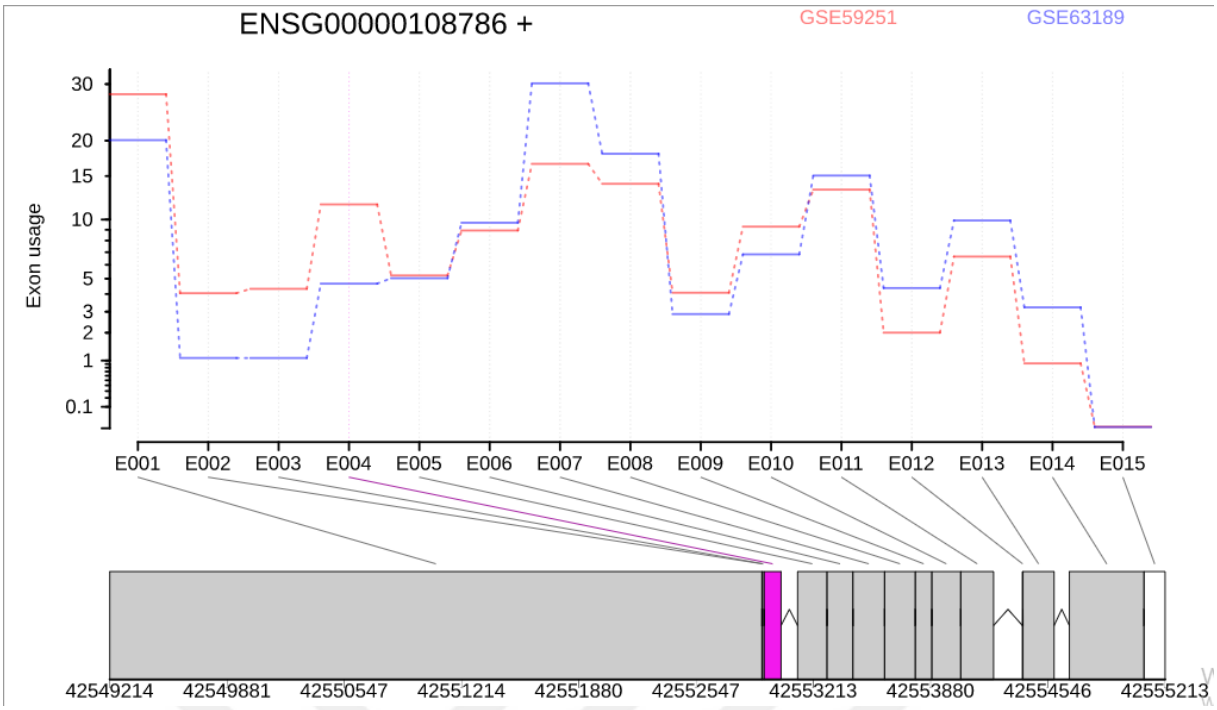
Şekil 4.19'da grafikte kırmızı renkler GSE59251 datasetini, mavi renkte gösterilen çizgiler GSE63189 datasetini göstermektedir. Yatay eksen genin eksonlarını belirtir. Toplam 10 eksonu bulunan genin, 9.ekson ve 10.eksonunda alternatif ekson kullanımı tespit edilmiştir.



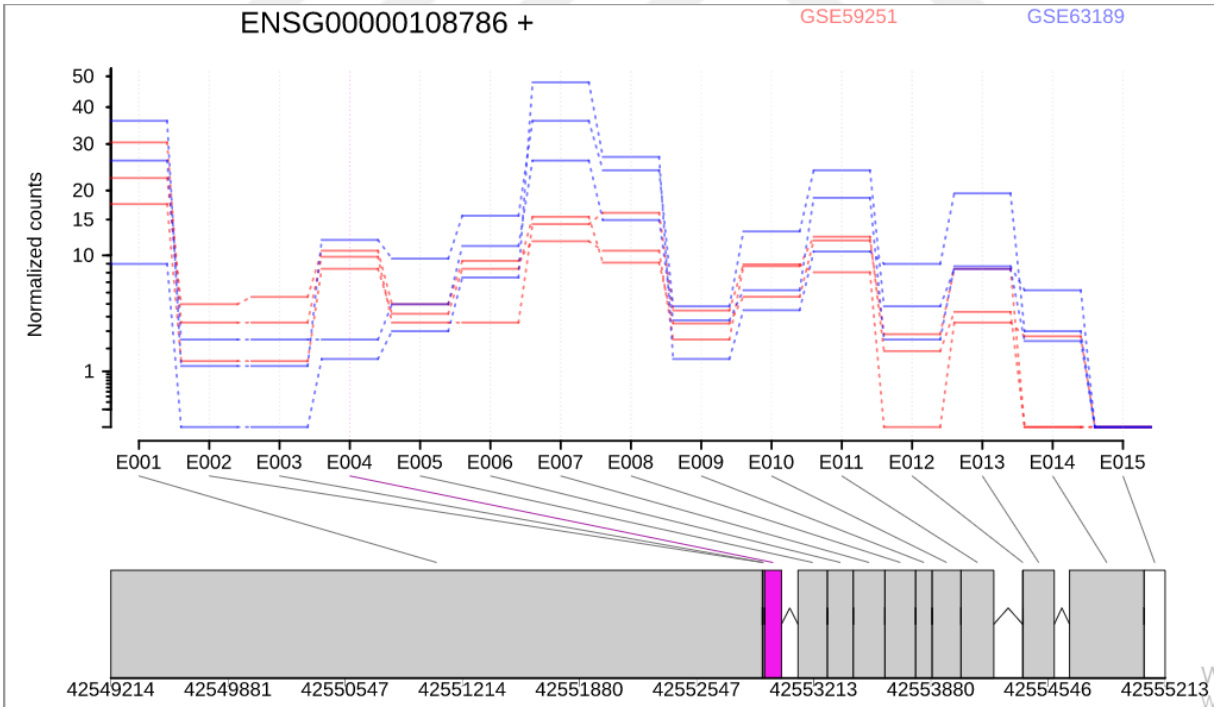
Şekil 4.19 Örneklerin her birindeki her eksonun normalize sayım değerleri. Mor renkte gösterilen, anlamlı bir şekilde diferansiyel ekson kullanımı gösteren ekson.



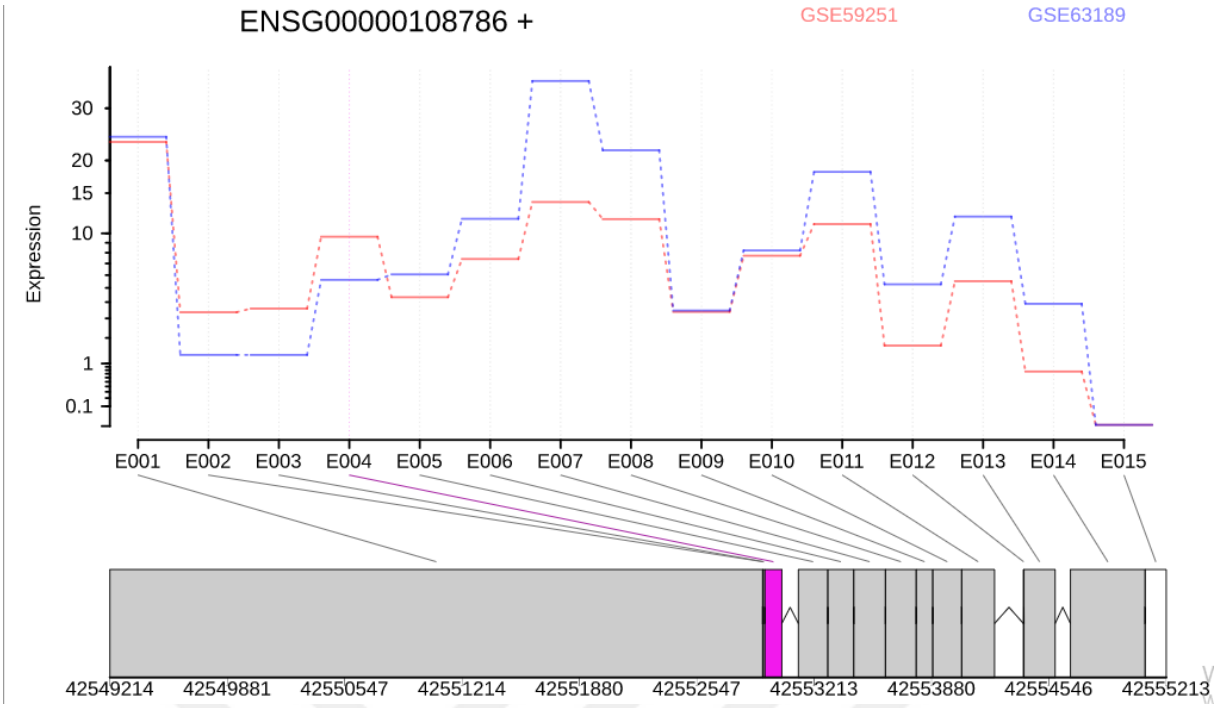
Şekil 4.20 Ekspresyon grafiği. Mor renkte gösterilen, anlamlı bir şekilde diferansiyel ekson kullanımı gösteren ekson.



Şekil 4.21 Ekson kullanım grafiği. Mor renkte gösterilen, anlamlı bir şekilde diferansiyel ekson kullanımı gösteren ekson.



Şekil 4.22 Örneklerin her birindeki her eksonun normalize sayım değerleri. Mor renkte gösterilen, anlamlı bir şekilde diferansiyel ekson kullanımı gösteren ekson.



Şekil 4.23 Ekspresyon grafiği. Mor renkte gösterilen, anlamlı bir şekilde diferansiyel ekson kullanımı gösteren ekson.

4.3.2 HeLa kanser hücre hattı

DEXSeq ve R kullanılarak HeLa kanser hücre hattı alternatif ekson kullanımı için analiz edildi. Analiz sonucunda, % 5'lik yanlış keşif oranı (false discovery rate) ile toplam 318.664 eksonik bölgenin 89.032 tanesinde alternatif ekson kullanımı tespit edildi. İncelenen toplam 15.545 gen arasından 11.479 tanesi etkilenmiştir ($p<0.05$).

Çizelge 4.14: HeLa kanser hücre hattı DEXSeq deney tasarımı.

sample	condition
SRR2960983	GSE75410
SRR2960986	GSE75410
SRR3169158	GSE77913
SRR3169161	GSE77913

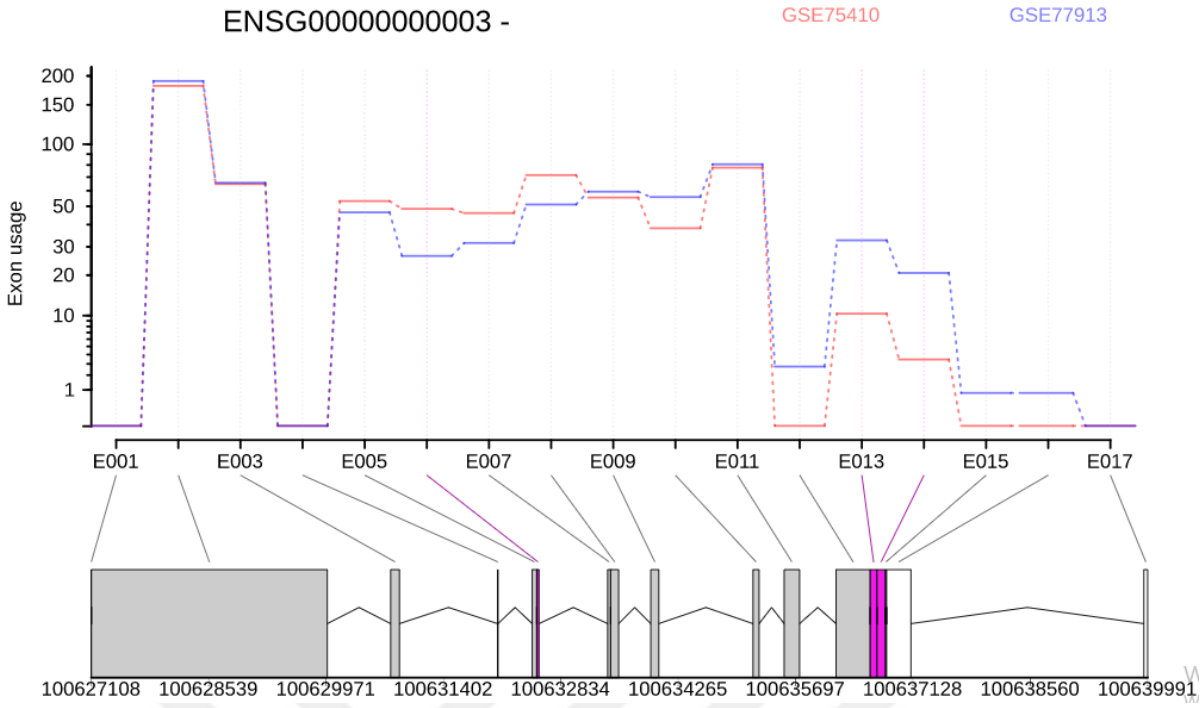
Çizelge 4.14'de, ilk kolon örnekleri, ikinci kolon ise koşulu belirtmektedir. Analizde koşul dataset olarak belirlenmiştir.

Çizelge 4.15: HeLa kanser hücre hattı alternatif ekson kullanımı sonuç çizelgesi

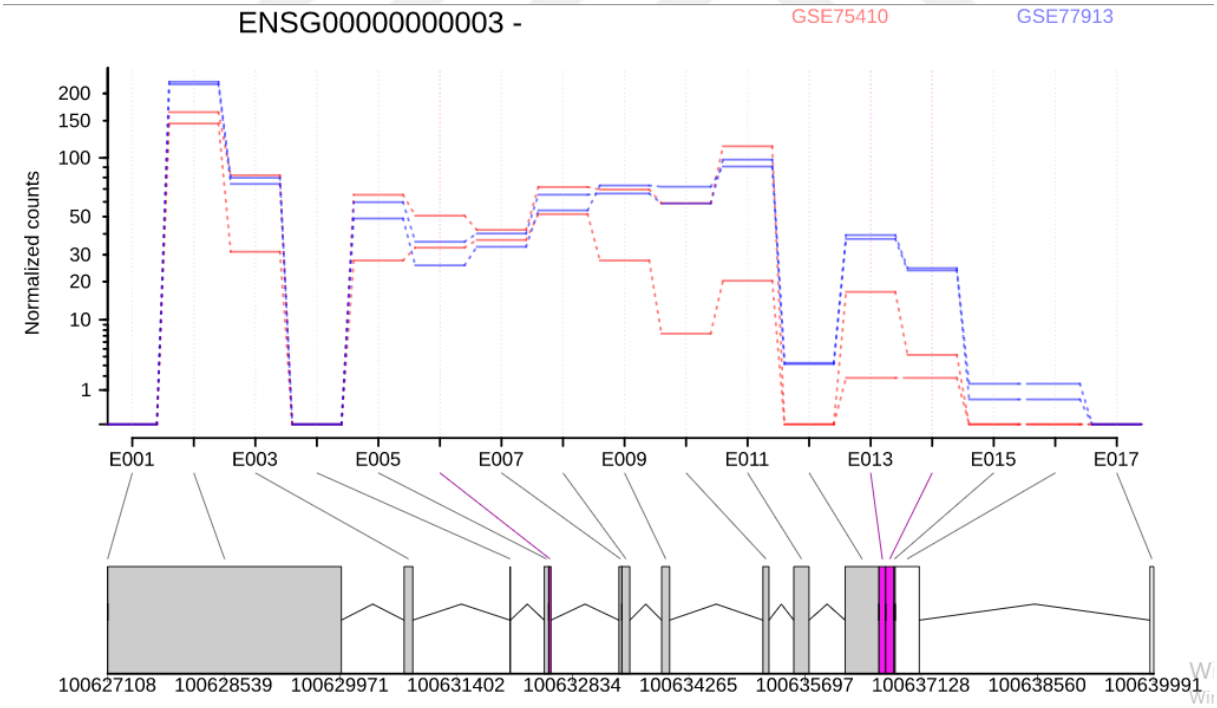
testForDEU result table

geneID	chr	start	end	total_exons	exon_changes
ENSG00000000003	X	100627108	100639991	17	3
ENSG00000000419	20	50934867	50958555	18	9
ENSG00000000457	1	169849631	169894267	22	10
ENSG00000000460	1	169662007	169854080	46	16
ENSG00000000971	1	196651878	196747504	32	3
ENSG00000001036	6	143494812	143511720	10	7
ENSG00000001084 ENSG00000231683	6	53497341	53617171	66	12
ENSG00000001167	6	41072945	41099976	12	3
ENSG00000001460	1	24356999	24416934	34	7
ENSG00000001461	1	24415802	24472976	26	5
ENSG00000001497	X	65512582	65534775	21	8
ENSG00000001617	3	50155045	50189075	39	5
ENSG00000001629	7	92245974	92401383	33	7

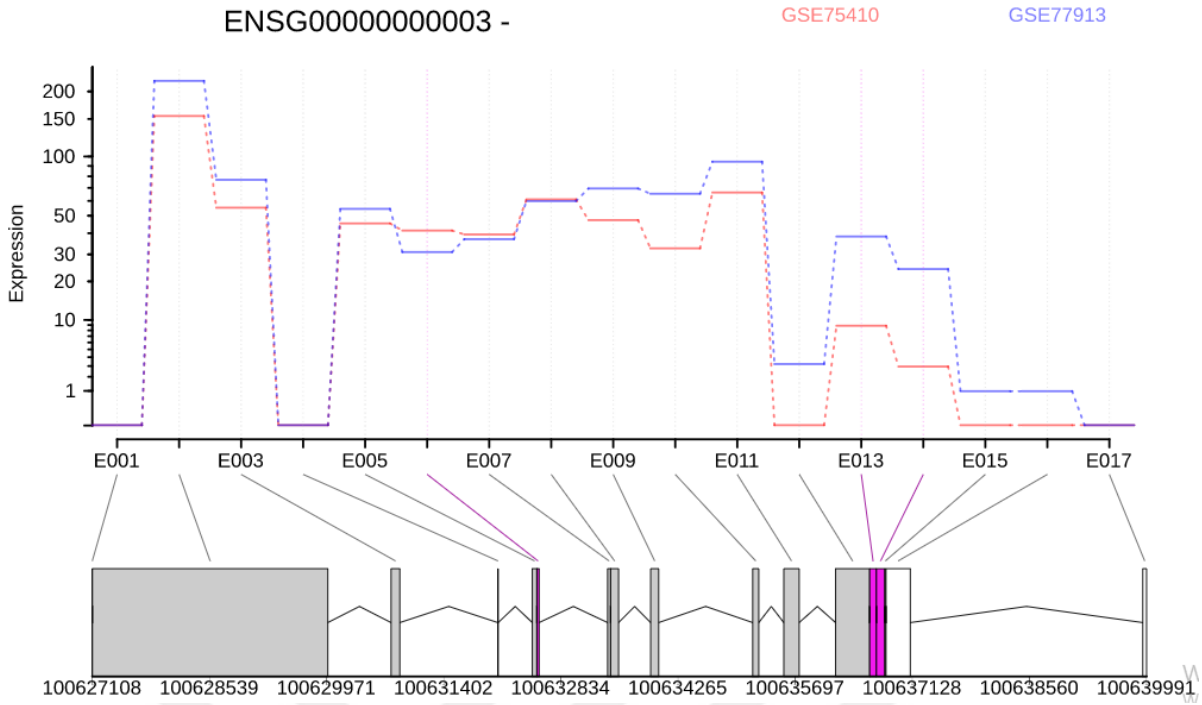
Yukarıdaki Çizelgede (Çizelge 4.15); ilk kolon gen isimlerini, ikinci kolon kromozom isimlerini, üçüncü kolon transkript başlama bölgesini, dördüncü kolon transkript bitiş bölgesini, beşinci kolon genin toplamda kaç eksonu olduğunu, son kolon ise ekson değişimini göstermektedir.



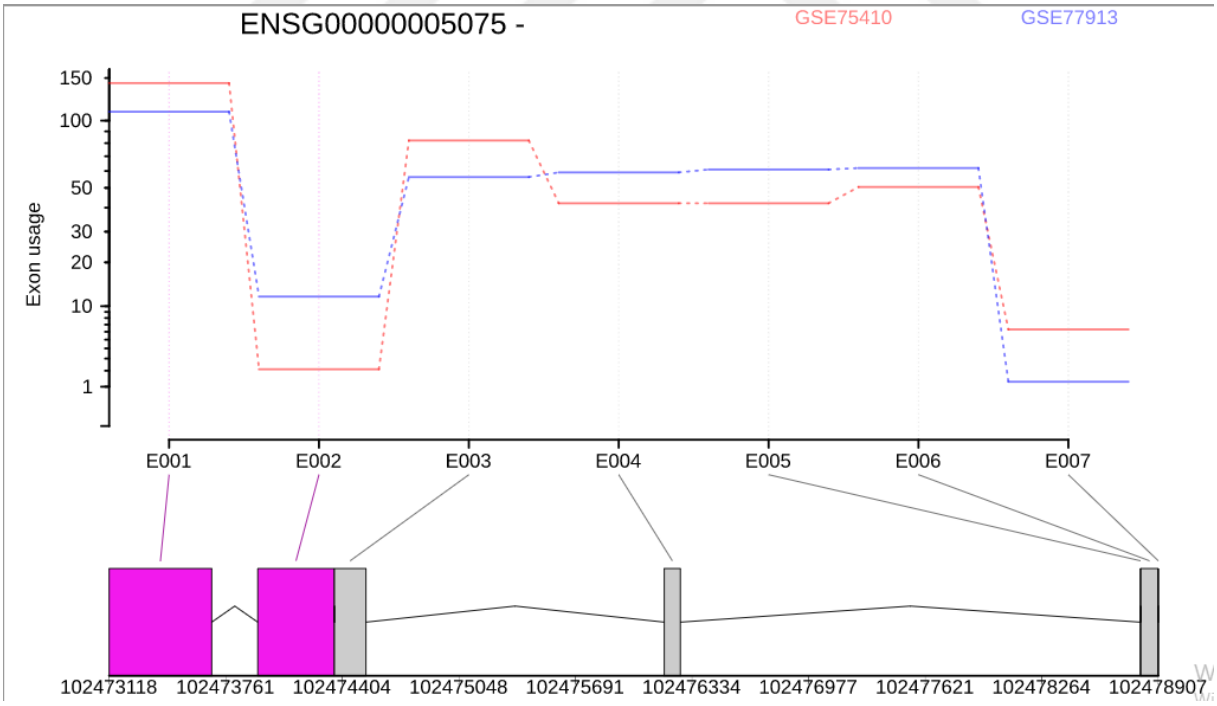
Şekil 4.24 Ekson kullanım grafiği. Mor renkte gösterilen, anlamlı bir şekilde diferansiyel ekson kullanımı gösteren ekson.



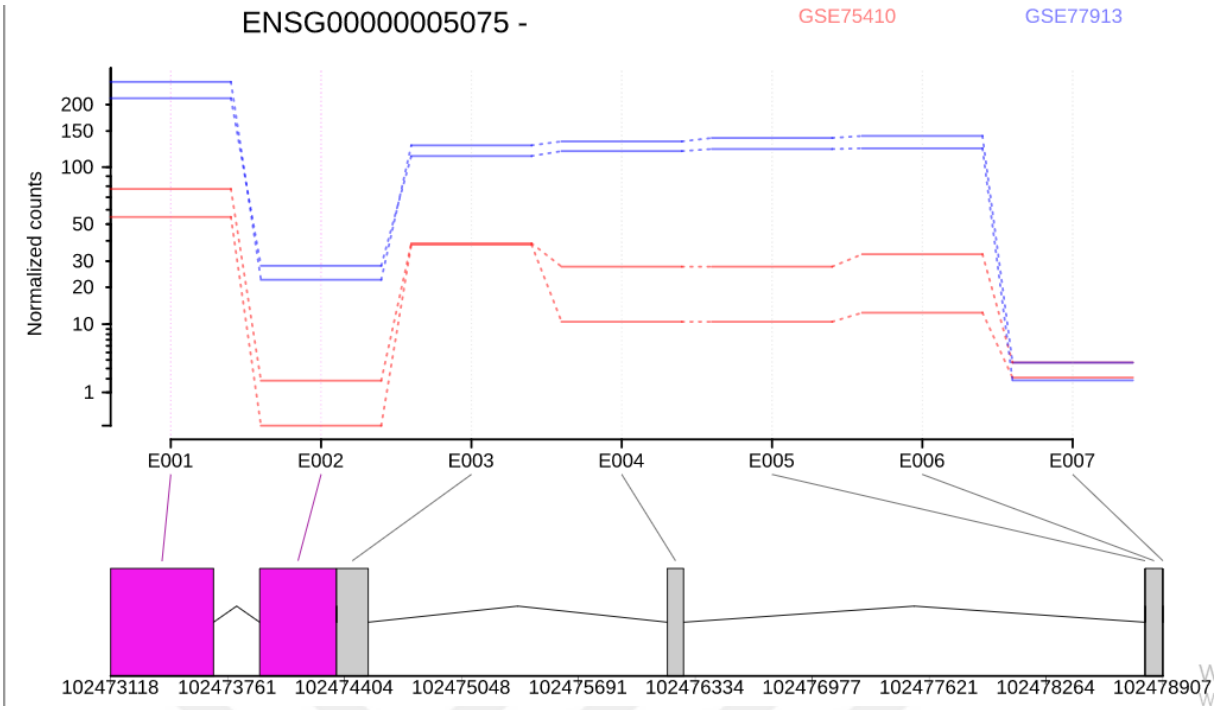
Şekil 4.25 Örneklerin her birindeki her eksonun normalize sayım değerleri. Mor renkte gösterilen, anlamlı bir şekilde diferansiyel ekson kullanımı gösteren ekson.



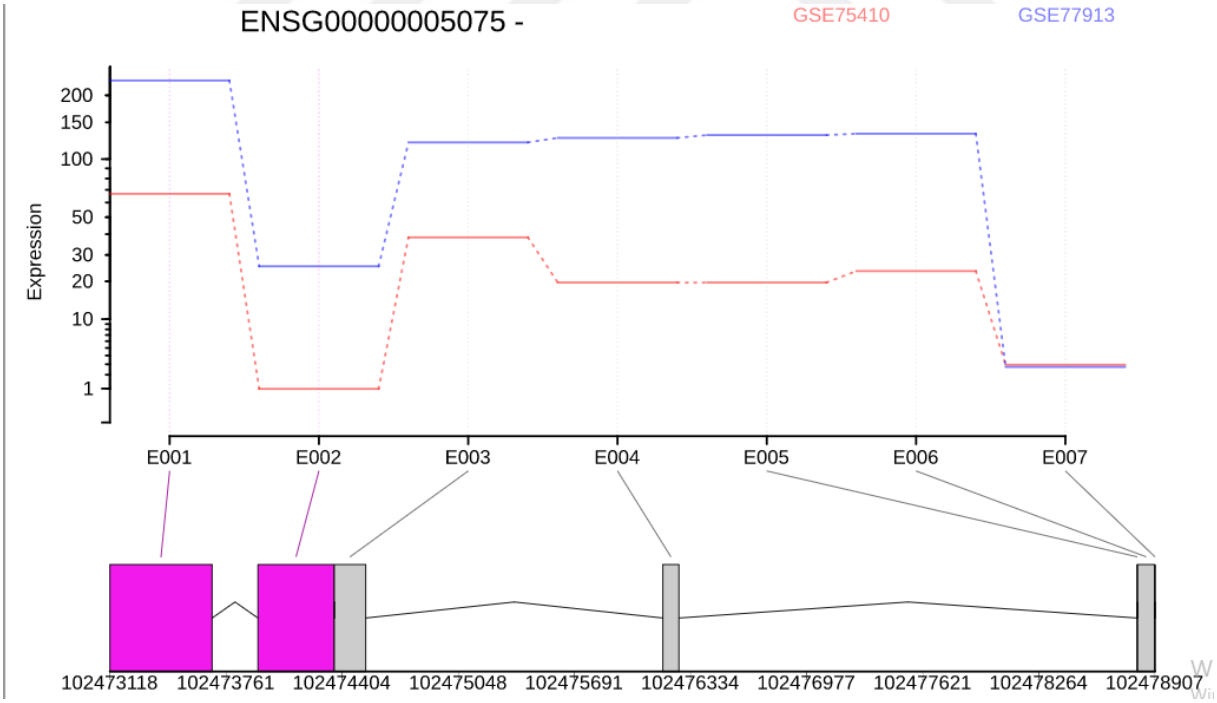
Şekil 4.26 Ekspresyon grafiği. Mor renkte gösterilen, anlamlı bir şekilde diferansiyel ekson kullanımı gösteren ekson.



Şekil 4.27 Ekson kullanım grafiği. Mor renkte gösterilen, anlamlı bir şekilde diferansiyel ekson kullanımı gösteren ekson.



Şekil 4.28 Örneklerin her birindeki her eksonun normalize sayım değerleri. Mor renkte gösterilen, anlamlı bir şekilde diferansiyel ekson kullanımı gösteren ekson.



Şekil 4.29 Ekspresyon grafiği. Mor renkte gösterilen, anlamlı bir şekilde diferansiyel ekson kullanımı gösteren ekson.

4.3.3 HCT116 kanser hücre hattı

DEXSeq ve R kullanılarak HCT116 kanser hücre hattı alternatif ekson kullanımı için analiz edildi. Analiz sonucunda, % 5'lik yanlış keşif oranı (false discovery rate) ile toplam 178.313 eksonik bölgenin 52.637 tanesinde alternatif ekson kullanımı tespit edildi. İncelenen toplam 12.229 gen arasından 10.610 tanesi etkilenmiştir ($p<0.05$).

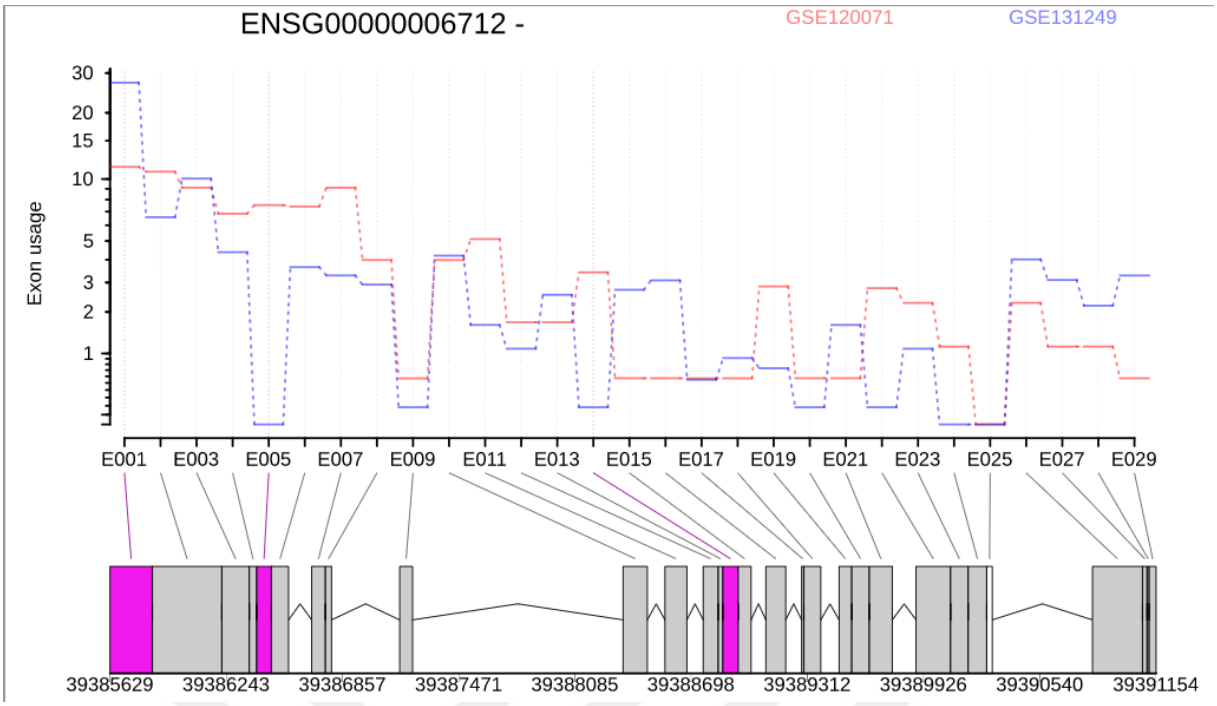
Çizelge 4.16: HCT116 kanser hücre hattı DEXSeq deney tasarımı.

sample	condition
SRR7865855	GSE120071
SRR7865856	GSE120071
SRR7865857	GSE120071
SRR9058970	GSE131249
SRR9058971	GSE131249
SRR9058972	GSE131249

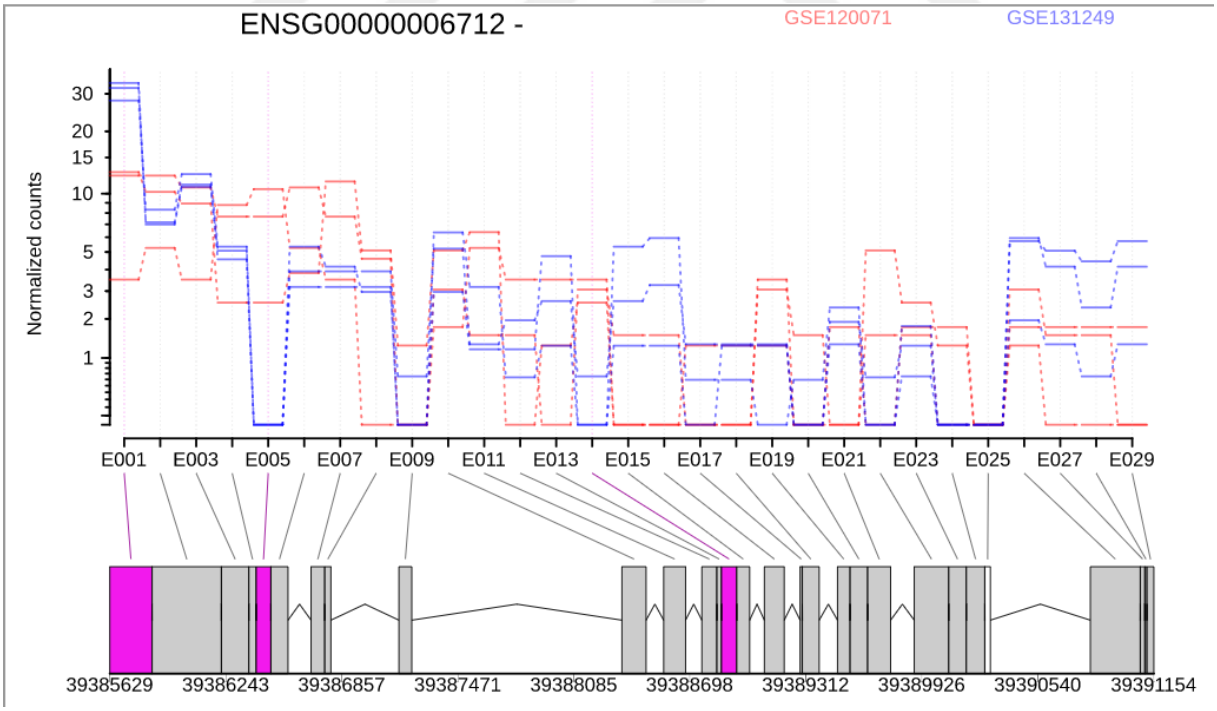
Çizelge 4.17: HCT116 kanser hücre hattı alternatif ekson kullanımı sonuç çizelgesi.

geneID	chr	start	end	total_exons	exon_changes
ENSG00000001036	6	143494812	143511720	10	2
ENSG00000001084 ENSG00000231683	6	53497341	53617171	66	4
ENSG00000001629	7	92245974	92401383	33	1
ENSG00000001630 ENSG00000001631 ENSG00000285772 ENSG00000240720 ENSG00000285953 ENSG00000243107	7	92112153	92246166	103	1
ENSG00000002330	11	64269830	64284704	18	1
ENSG00000002587	4	11393150	11429564	5	2
ENSG00000002834	17	38869859	38921770	25	4
ENSG00000003056	12	8940361	8949761	36	1
ENSG00000003147	7	8113184	8262687	54	1
ENSG00000003987	8	17296794	17413528	31	4
ENSG00000004059 ENSG00000106328	7	127588386	127602144	37	1
ENSG00000004455	1	33007940	33080996	44	1
ENSG00000004478	12	2794970	2805423	26	1
ENSG00000004487	1	23019443	23083689	35	6
ENSG00000004897	17	47117703	47189422	57	1

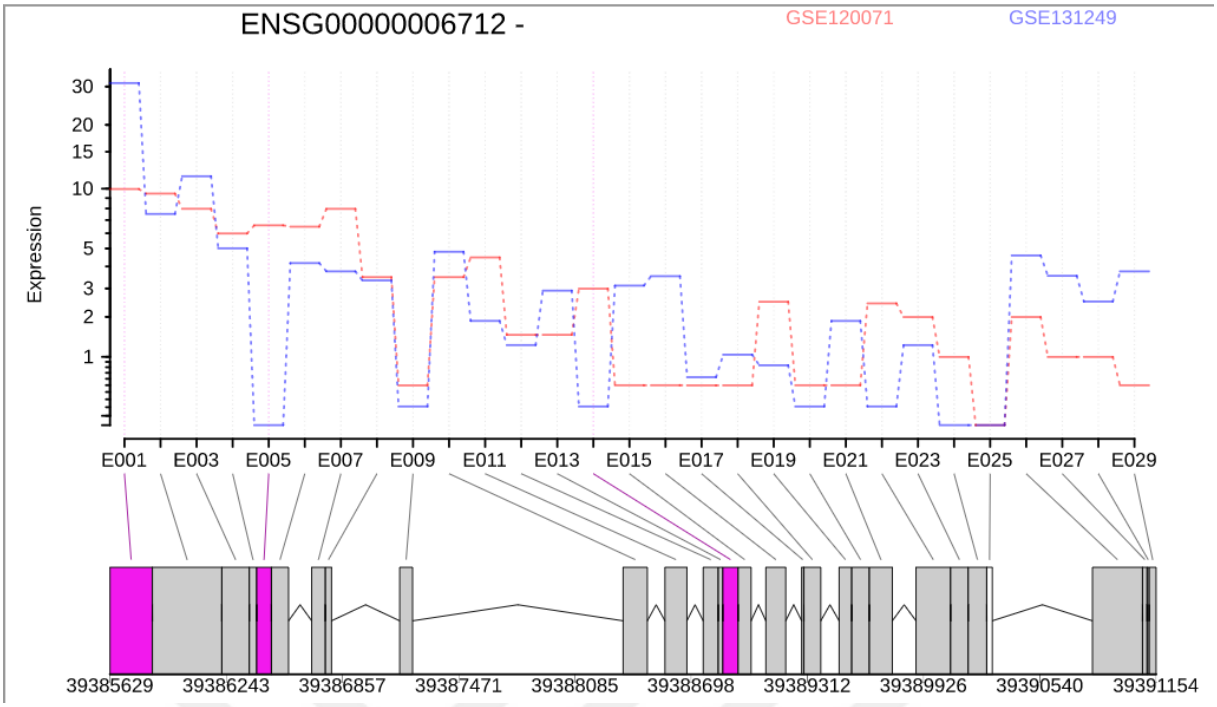
Yukarıdaki Çizelgeda (Çizelge 4.17); ilk kolon gen isimlerini, ikinci kolon kromozom isimlerini, üçüncü kolon transkript başlama bölgesini, dördüncü kolon transkript bitiş bölgesini, beşinci kolon genin toplamda kaç eksonu olduğunu, son kolon ise ekson değişimini göstermektedir.



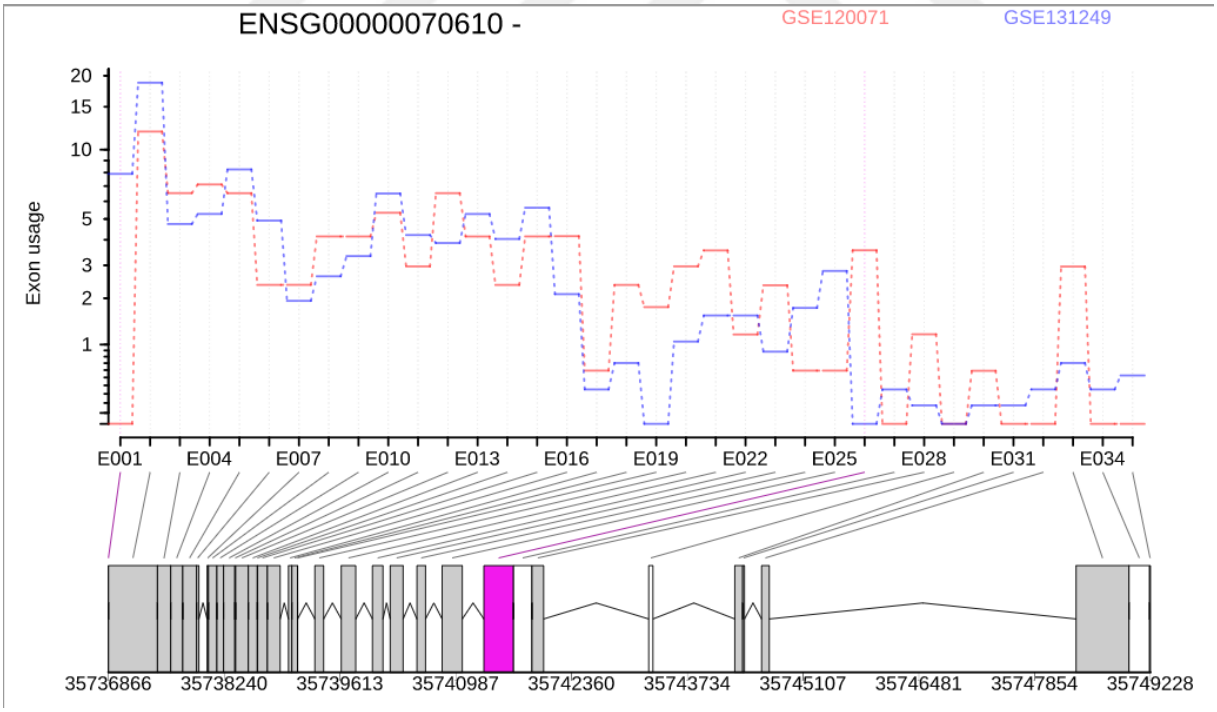
Şekil 4.30 Ekson kullanım grafiği. Mor renkte gösterilen, anlamlı bir şekilde diferansiyel ekson kullanımı gösteren ekson.



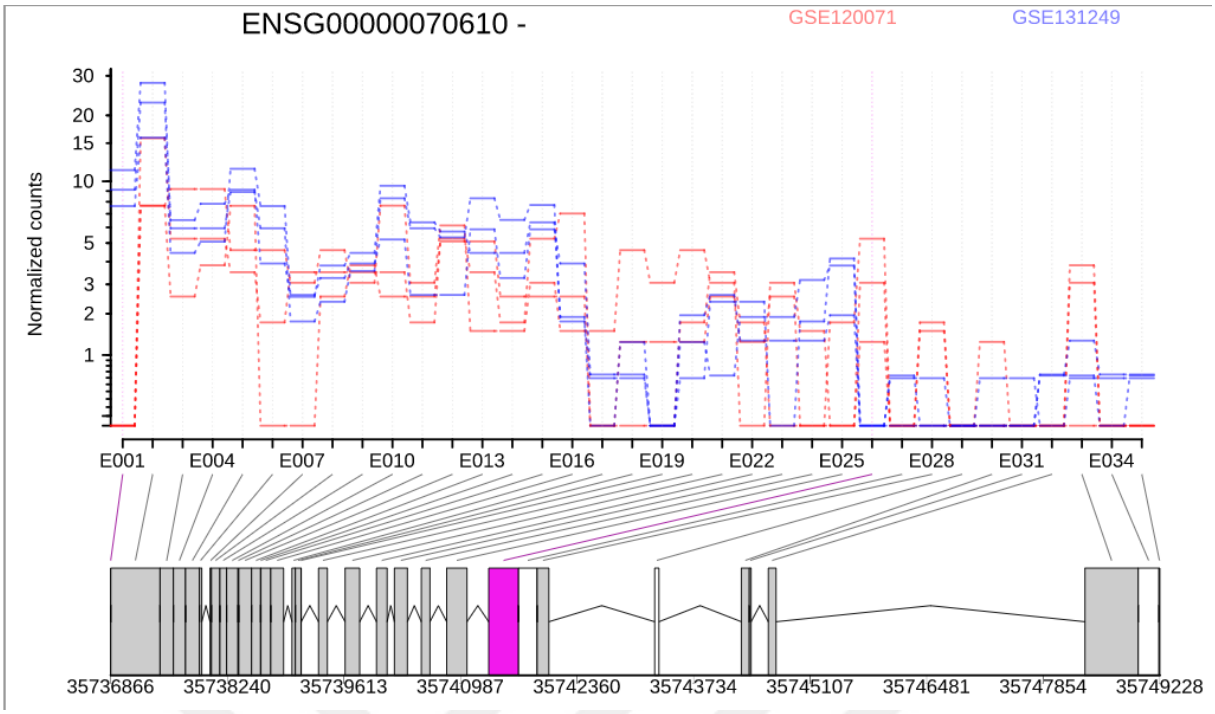
Şekil 4.31 Örneklerin her birindeki her eksonun normalize sayım değerleri. Mor renkte gösterilen, anlamlı bir şekilde diferansiyel ekson kullanımı gösteren ekson.



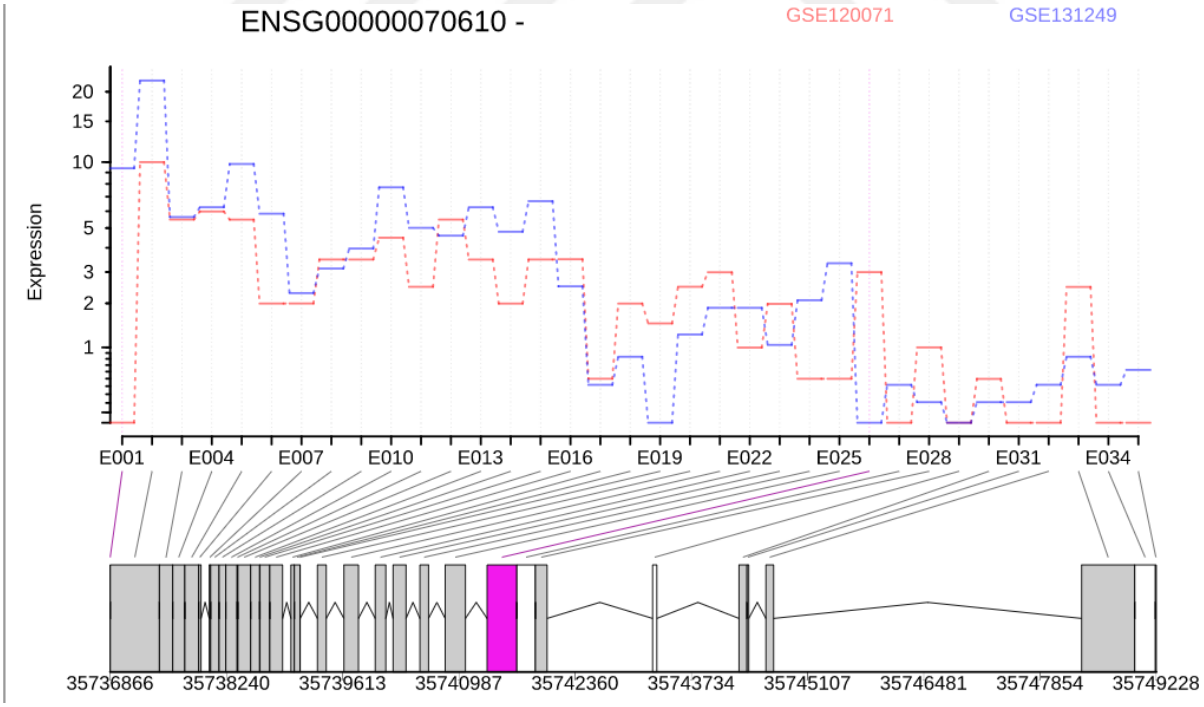
Şekil 4.32 Ekspresyon grafiği. Mor renkte gösterilen, anlamlı bir şekilde diferansiyel ekson kullanımı gösteren ekson.



Şekil 4.33 Ekson kullanım grafiği. Mor renkte gösterilen, anlamlı bir şekilde diferansiyel ekson kullanımı gösteren ekson.



Şekil 4.34 Örneklerin her birindeki her eksonun normalize sayım değerleri. Mor renkte gösterilen, anlamlı bir şekilde diferansiyel ekson kullanımı gösteren ekson.



Şekil 4.35 Ekspresyon grafiği. Mor renkte gösterilen, anlamlı bir şekilde diferansiyel ekson kullanımı gösteren ekson.

4.3.4 A549 kanser hücre hattı

R ve DEXSeq kullanılarak alternatif ekson kullanımı analizi yapılan A549 kanser hücre hattında anlamlı bir sonuç bulunamamıştır.

BÖLÜM 5

TARTIŞMA VE SONUÇ

Yeni nesil sekanslama teknolojisi önceki sekanslama yöntemlerinin (Sanger sekanslama) verimsiz kalmasına istinaden ortaya çıkmıştır. NGS kullanılarak bütün bir insan genomu tek bir gün içinde sekanslanabilir. Buna karşılık, insan genomunu deşifre etmek için kullanılan önceki Sanger sekanslama teknolojisi, son taslağı sunmak için on yıldan fazla bir süreye ihtiyaç duyuyordu [35]. Yeni nesil dizileme teknolojisi her geçen gün daha hacimli, kolay kullanılabilir ve ekonomik bir hal almaktadır. Bu durum ile ters orantılı olan tek şey elde edilen verilerin büyüklüğü ve verilerin işlenme süresidir [36]. Yeni nesil sekanslamanın gelişmesiyle bu verileri analiz edecek biyoinformatik metotlar da paralel olarak gelişmektedir.

Hücre hatları, benzer genotiplere ve fenotiplere sahip neredeyse sınırsız hücre kaynağı sağlar. Kullanımları hayvan ve insan deneyleri ile ilgili etik sorunları önler [37]. Hücre hatları, kanser araştırmalarının hız kazanması ile özellikle kanser ilaçlarının geliştirilmesinde ve etkilerinin belirlenmek istenmesinden dolayı büyük önem kazanmıştır. İn vitro deney modellerinin oluşturulmasında sıkça kullanılan bir model olan kanser hücre hatları birçok güvenilirlik tartışmasının odağında olmuştur. Fasterius ve arkadaşları çalışmalarında [38] belirli kanser hücre hatlarında RNA-seq verilerine göre yapılan varyant analizinde aynı sonucunda, aynı isimle temsil edilen hücre hattı çiftlerinde büyük ölçüde RNA ifadesi farklılıkları olduğunu tespit etmiştir.

Bu çalışmada, daha önceden RNA-Seq analiz yapılmış ve internet veri tabanlarında bulunan kanser hücre hatları verilerini indirilerek; kanser hücre hatlarındaki diferansiyel gen ekspresyonu ve alternatif ekson kullanımı analiz edildi. Bu analizler kanser hücre hatlarının ne tür moleküler değişikliklere maruz kaldığını anlamak için yapıldı.

Diferansiyel gen ekspresyonu yapılan analizlerde, dört farklı kanser hücre hattı için anlamlı bir şekilde diferansiyel gen ekspresyonu tespit edildi. MCF7 kanser hücre hattı için toplamda analiz edilen 12.380 gen arasından 6.774 tanesinde diferansiyel gen ekspresyonu tespit edildi

($p<0.05$). A549 kanser hücre hattında incelenen 8.884 gen arasından 785 tanesinde diferansiyel gen ekpresyonu tespit edildi ($p<0.05$). HCT116 kanser hücre hattı için analiz edilen 10.796 gen arasında 1781 tanesinde diferansiyel gen ekspresyonu tespit edildi ($p<0.05$). Son olarak analiz edilen HeLa kanser hücre hattında 16.742 gen arasından 6031 tanesinde diferansiyel gen ekspresyonu tespit edildi ($p<0.05$).

İncelen kanser hücre hatlarında alternatif ekson kullanımı için yapılan analizlerde, A549 hücre hattı haricinde diğer kanser hücre hatlarında anlamlı bir şekilde alternatif ekson kullanımı tespit edildi. A549 hücre hattı DEXSeq yazılımı kullanılarak analiz edildi, analiz sonucunda çıktı olarak anlamlı bir sonuç bulunamamıştır ibaresi yer almıştır. Ancak diğer kanser hücre hatlarına bakıldığı zaman; MCF7 kanser hücre hattında toplam 30.771 eksonik bölgenin 1.021 tanesinde alternatif ekson kullanımı, Hela kanser hücre hattında toplam 318.664 eksonik bölgenin 89.032 tanesinde alternatif ekson kullanımı, HCT116 kanser hücre hattında toplam 178.313 eksonik bölgenin 52.637 tanesinde alternatif ekson kullanımı tespit edildi ($p<0.05$).

Bu sonuçlar, aynı kanser hücre hattı olmasına rağmen farlı laboratuvarlarda pasajlanmış ve kültüre edilmiş hücre hatlarında moleküler farklılıklar olduğunu göstermektedir. Gen ekspresyon ve alternatif ekson kullanımı analiz sonuçlarına bakıldığı zaman, aynı genetik materyale sahip iki farklı hücrenin birbirinden farklı transkriptomik profiller oluşturduğunu söyleyebiliriz. Bunun sebebi olarak, hücre hatlarının farklı pasajlamalar ve büyütülme işlemlerinden geçirilmeleri epigenetik farklılaşmalara neden olduğu düşünülmektedir. Hücre kültürü ortamındaki değişkenlerin hücrelerde yaratabileceği epigenetik farklılaşmaların hücre hatları kullanırken göz önüne alınmalıdır.

İlerleyen çalışmalarda diferansiyel gen ekspresyonu ve alternatif ekson kullanımına ek olarak RNA düzenlenmesi (RNA editing) analizi de yapılarak, kanser hücre hatlarındaki moleküler değişiklikler aydınlatılabilir. Çalışmamız in silico analiz olduğu için in vitro deneylerler (kantitatif PCR) ile doğrulanmalıdır.

KAYNAKLAR

[1] Gartler S M (1968) Apparent Hela cell contamination of human heteroploid cell lines. *Nature*, 217(5130):750-751.

[2] Buehring G C, Eby E A, Eby M J (2004) Cell line cross-contamination: how aware are Mammalian cell culturists of the problem and how to monitor it? *In vitro cellular & developmental biology Animal*, 40(7):211-215.

[3] Pettersson E, Lundeberg J, Ahmadian A (2009) Generations of sequencing Technologies. *Genomics*, 93(2):105-111

[4] Behjati S, Tarpey P S. (2013) What is next generation sequencing? Archives of Disease in Childhood - *Education and Practice* ;98:236-238.

[5] Uma T, Shankavaram, William C, Reinhold, Satoshi N, Sylvia M, Daisaku M, Krishna K, Chary, Mark A, Reimers, Uwe S, Ari K, Douglas D, Jeffrey C, Eric P, Kaldjian, Dominic A, Scudiero, Emanuel P, Lance L, Jae K L, John N. Weinstein (2007) Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study. *Mol Cancer Ther* 6(3):820-832. doi:10.1158/1535-7163.MCT-06-0650

[6] Reinhold W C, Reimers M A, Lorenzi P, Jennifer H, Uma T S, Micah S Z, Kimberly J B, Satoshi N, Ogechi I, Yves G P, John N W (2010) Multifactorial regulation of E-cadherin expression: an integrative study. *Mol Cancer Ther*;9(1):1-16. doi:10.1158/1535-7163.MCT-09-0321

[**7] Giovinazzi S, Sirleto P, Aksenova V, Viacheslav M M, Roberto Z, William C, Alexander M I** (2014) Usp7 protects genomic stability by regulating Bub3. *Oncotarget* 5(11):3728-3742. doi:10.18632/oncotarget.1989

[8] Garraway L A, Widlund H R, Rubin M A, Gad G, Aaron J B, Sridhar R, Rameen B, Danny A M, Scott R G, Jinyan D, Charles L, Stephan N W, Cheng L, Todd R G, David L R, Matthew L M, David E F, William R S (2005) Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature*; 436(7047):117-122. doi:10.1038/nature03664

[9] Reinhold W C, Varma S, Sunshine M, Rajapakse V, Luna A, Kohn K W, Stevenson H, Wang Y, Heyn H, Nogales V , Moran S , Goldstein D J, Doroshow J H, Meltzer P S, Esteller M, Pommier Y (2017) The NCI-60 Methylome and Its Integration into CellMiner. *Cancer Research* 77(3):601-612. doi:10.1158/0008-5472.CAN-16-0655

[10] Kukurba K R, Montgomery S B. (2015) RNA Sequencing and Analysis. *Cold Spring Harb Protocol* 2015(11):951–969. Published 2015 Apr 13. doi:10.1101/pdb.top084970

KAYNAKLAR (devam ediyor)

[11] Kuang J, Yan X, Genders A J, Granata C, Bishop D J (2018) An overview of technical considerations when using quantitative real-time PCR analysis of gene expression in human exercise research. *PLoS ONE* 13(5): e0196438. https://doi.org/10.1371/journal.pone.0196438

[12] Sadeque, A, Serão, N V, Southey, Delfino K R, Rodriguez-Zas S L (2012) Identification and characterization of alternative exon usage linked glioblastoma multiforme survival. *BMC Med Genomics* 5, 59. https://doi.org/10.1186/1755-8794-5-59

[13] Wang, Xiao-Zhong. (2015). Mechanism of alternative splicing and its regulation (Review). *BIOMEDICAL REPORTS*. 3. 152-158. 10.3892/br.2014.407.

[14] Laderas T G, Walter N A, Mooney M, Vartanian K, Darakjian P, Buck K, Harrington C A, Belknap J, Hitzemann R, McWeeney S K (2011) Computational detection of alternative exon usage. *Front Neurosci*. 2011, 5: 69.

[15] Matthew P A, Davis, Stijn van Dongen, Cei Abreu-Goodger, Bartonicek N, Anton J (2013) *Enright.Methods*. 63(1):41–49

[16] Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal*, 17, pp.10–12.

[17] Andrews S (2010) FastQC: a quality control tool for high throughput sequence data.

[18] Bolger A M, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114-2120. doi:10.1093/bioinformatics/btu170

[19] Kim D, Paggi J M, Park C, Bennett C, Salzberg S L ((2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* 37, 907–915. https://doi.org/10.1038/s41587-019-0201-4

[20] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 25(16):2078-2079.

[21] Pertea M, Pertea G, Antonescu C, Chang T C, Mendell J T, Salzberg S L (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 33, 290–295 https://doi.org/10.1038/nbt.3122

[22] Robinson J, Thorvaldsdóttir H, Winckler W, Guttman M, Lander E S, Getz G, Mesirov J P (2011) Integrative genomics viewer. *Nat Biotechnol* 29, 24–26 https://doi.org/10.1038/nbt.1754

[23] Ripley B D (2001) The {R} project in statistical computing. MSOR Connections. Newsletter of the LTSN Maths, Stats & OR Network (The University of Birmingham, Edgbaston, U.K.) 1, 23–25

[24] R Studio. (2012) R Studio: integrated development environment for R, Version 0.97.390. Boston, MA: R Studio. Available at: http://www.rstudio.org.

[25] Frazee A, Pertea G, Jaffe A, Langmead B, Salzberg S L, Leek J T (2015) Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nat Biotechnol* 33, 243–246. https://doi.org/10.1038/nbt.3172

[26] Anders S, Reyes A, Huber W (2012) Detecting differential usage of exons from RNA-seq data. *Genome Res.* 22(10):2008-2017. doi:10.1101/gr.133744.111

[27] Edgar R, Domrachev M, Lash A E (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30(1):207-210. doi:10.1093/nar/30.1.207

[28] Harrison P W, Alako B, Amid C, Cerdeño-Tárraga A, Cleland I, Holt S, Hussein A, Jayathilaka S, Kay S, Keane T, Leinonen R, Liu X, Martínez-Villacorta J, Milano A, Pakseresht N, Rajan J, Reddy K, Richards E, Rosello M, Silvester N, Smirnov D, Toribio A L , Vijayaraja S, Cochrane G (2018) The European Nucleotide Archive. *Nucleic Acids Res.* 2019;47(D1):D84-D88. doi:10.1093/nar/gky1078

[29] Gabriella R, Nikolay K, Marco B, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Ison J, Keays M, Kurbatova N, Malone J, Mani R, Mupo R, Pereira R P, Pilicheva E, Rung J, Sharma A, Tang Y A, Ternent T, Tikhonov A, Welter D, Williams E, Brazma A, Parkinson H, Sarkans U (2013) ArrayExpress update—trends in database growth and links to data analysis tools, *Nucleic Acids Research*, 41(1):987–990, https://doi.org/10.1093/nar/gks1174

[30] Daniel Z, Premanand A, Wasiu A, Amode M R, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón A G, Gil, Gordon L, Haggerty L, Haskell E, Hourlier T, Izuogu O G, Janacek S H, Juettemann T, To J K, Laird M R, Lavidas I, Liu I, Loveland J E, Maurel T, McLaren W, Moore B, Mudge J, Murphy D N, Newman V, Nuhn M, Ogeh D, Ong C K, Parker A, Patricio M, Riat H S, Schuilenburg H, Sheppard D, Sparrow H, Taylor K, Thormann A, Vullo A, Walts B, Zadissa A, Frankish A, Hunt S E, Kostadima M, Langridge N, Martin F J, Muffato M, Perry E, Ruffier M, Staines D M, Trevanion S J, Aken B L, Cunningham F, Yates A, Flicek P (2018) Ensembl 2018, *Nucleic Acids Research*, 46(1):754–761, https://doi.org/10.1093/nar/gkx1098

[**31] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R** (2009) "The Sequence Alignment/Map format and SAMtools". *Bioinformatics*. 25 (16): 2078–2079. doi:10.1093/bioinformatics/btp3

[32] Pertea M, Kim D, Pertea G M, Leek J T, Salzberg S L. (2016) Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc*. 11(9):1650-1667. doi:10.1038/nprot.2016

KAYNAKLAR (devam ediyor)

[33] Reyes A, Anders S, Huber W (2013) Inferring differential exon usage in RNA-Seq data with the DEXSeq package.

[34] P-values, False Discovery Rate (FDR) and q-values (t.y)

Waters, Adres: http://www.nonlinear.com/transomics/metabolomics/v1.0/faq/pq-values.aspx

URL-1 <http://www.nonlinear.com/>, Ziyaret tarihi 29.05.2020

[35] Behjati S, Tarpey P S What is next generation sequencing? (2013) *Arch Dis Child Educ Pract Ed*. 98(6):236-238. doi:10.1136/archdischild-2013-30434

[36] Greene C S, Tan J, Ung M, Moore J H, Cheng C (2016) Big data bioinformatics [published correction appears in J Cell Physiol. Jan;231(1):257]. *J Cell Physiol*. 2014;229(12):1896-1900. doi:10.1002/jcp.24662

[37] Masters J (2000) Human cancer cell lines: fact and fantasy. *Nat Rev Mol Cell Biol 1*, 233–236 https://doi.org/10.1038/35043102

[38] Fasterius E, Al-Khalili Szigyarto C (2018) Analysis of public RNA-sequencing data reveals biological consequences of genetic heterogeneity in cell line populations. *Scientific reports*, 8(1):11226.

ÖZGEÇMİŞ

Gümüşhane'nin Şiran ilçesinde 25.07.1995 tarihinde dünyaya geldim. İlkokulu 4.sınıfa kadar Mertekli İlköğretim okulunda okuduktan sonra 5.sınıftan 8.sınıfa kadar Mithatpaşa ilköğretim okulunda okudum. İlköğretimin ardından liseyi, Şehit Tuna Teğmen Anadolu lisesi'nde okumaya başladım. 2013 yılında liseden mezun olduktan sonra aynı yıl Zonguldak Karaelmas Üniversitesi' Moleküler biyoloji ve genetik bölümünü kazandım. Lisans yıllarında eğitim alırken 2.sınıf yazında (2015), Hatay Mustafa Kemal Araştıma Ve Uygulama Merkezi genetik laboratuvarında 1 ay staj yaptım. 2017 yılı bahar döneminde ERASMUS değişim programıyla yaklaşık 5 ay Çek Cumhuriyeti/Brno Masaryk Üniversitesi'nde öğrenim gördüm. Lisans hayatımı 2018 yılında noktaladıktan sonra aynı yıl, Zonguldak Bülent Ecevit Üniversitesi Fen bilimleri enstitüsü Moleküler biyoloji anabilim dalında yüksek lisans hayatıma başladım. Yüksek lisans döneminde, I. Uşak Uluslararası Sağlık Bilimleri ve Biyoteknoloji Kongresi'nde sunum yaptım (2019). Yüksek lisans eğitimimi 2020 yılında bitirdim. Özel ilgi alanlarım (Hobiler); Türk tarihi, bilim, cosmos, sinema, basketbol, futbol, tenis.

İLETİŞİM BİLGİLERİ:

E-posta: bekirayhan.29@gmail.com Tel: 05447210465