İstanbul Üniversitesi Sağlık Bilimleri Enstitüsü Sağlık Bilimlerinde İleri Araştırmalar Dergisi
Istanbul University Institute of Health Sciences Journal of Advanced Research in Health Sciences

Sağlık Bilimlerinde İleri Araştırmalar Dergisi 2020, Cilt 3, Sayı 2
DOI: 10.26650/JARHS2020-737495 Araştırma Makalesi / Research Article

Lösemi Modelinde Tüm Genom RNA Dizileme Analiz Algoritması Geliştirilmesi

Whole Genome RNA Sequencing Analysis Algorithm in Leukemia Model

Eda Sun[1,2], **Müge Sayitoğlu**[3]

ÖZ

Amaç: RNA Dizileme teknolojisi gen anlatım farklılıkları ve kodlayan bölgedeki varyasyonlar, kodlama yapmayan küçük RNAların anlatımları ve gen füzyonlarının belirlenmesi ile bu farklılıkların nedenlerini sunabilmektedir. Ancak bu kadar enformatik bilgiler sunabilen bu teknolojinin analizlerinin yapılması ve yorumlanması oldukça zorludur. T- hücreli akut lenfoblastik lösemi (T-ALL) de prognostik öneme sahip ve hastalığın takibinde kullanılabilecek güvenilir bir genetik belirteç bulunmamakla birlikte, doğrudan tedavi protokolünü ve tedavide yararlanılacak yeni hedef proteinleri belirlemede esas olacak moleküler alt yapı ve sınıflandırma da bilinmemektedir.

Gereç ve Yöntem: Biz de bu çalışmamızda, T-ALL gibi karmaşık bir genomik arka plana sahip lösemi hücrelerinde RNA-dizileme için en uygun enformatik iş akış algoritmasını oluşturmayı amaçladık. Bu çalışmada RNA dizileme ile Jurkat ve Molt 4 hücre hatları dizilenmiştir. Doğrulama ve karşılaştırma amacıyla açık veri bankalarından elde edilen sağlıklı timosit alt grupları ve T-ALL hasta (n=12) örnekleri (GSE48173) kullanılmıştır.

Bulgular: Açık erişimli veri araçları ile gerçekleştirdiğimiz enformatik analizlerde doku spesifik alternatif kırpılma ürünlerinin kantitatif tayinini, spesifik gen varyasyonlarını ve global gen anlatım düzeylerini başarılı bir şekilde tespit ettik ve T-ALL hasta verisinde aynı yaklaşımları kullanarak doğrulama yaptık.

Sonuç: Çalışmamızın sonucunda lösemi hastalarının veri analizinde kullanılabilecek uygun araçlar ve algoritma belirlenmiştir.

Anahtar Kelimeler: RNA-Dizileme, enformatik, akut lenfoblastik lösemi

ABSTRACT

Objective: RNA Sequencing technology can offer gene expression differences and the reasons for these differences by detecting variations in the coding region, expession of non-coding RNAs and gene fusions. However, it is very difficult to analyze and interpret this technology, which can provide such valuable information. Although there is no reliable genetic marker for T-cell acute lymphoblastic leukemia (T-ALL), which can be used in the follow-up of the disease, the molecular infrastructure and classification that will be directly used in determining the treatment protocol and the new target proteins to be used in treatment are not known.

Material and Methods: In this study, we aimed to establish the most suitable workflow algorithm for RNA sequencing in cell lines belonging to a group with a complex genomic background such as T-ALL. With this study, the Jurkat and Molt4 cell lines were sequenced by RNA sequencing. In order to increase the significance of our study, the results of different thymocyte subgroups and 12 T-ALL patient samples (GSE48173) were investigated.

Results: We conducted a bioinformatics data approach by using open access data tools, and we successfully detected the tissue specific quantitative alternative splicing gene products, gene specific variations and global gene expression levels, and verified them using the same approach in T-ALL patient data.

Conclusion: Aside from these molecular findings that we have achieved, one of our goals in this study was to develop an algorithm of transcriptomic data, which is difficult to work with and to interpret, and showed the correctness of our algorithm by confirming the data described in the literature.

Keywords: RNA-Sequencing, informatics, acute lymphoblastic leukemia

[1] İstanbul Üniversitesi, Sağlık Bilimleri Enstitüsü, İstanbul, Türkiye

[2] İstinye Üniversitesi, Tıp Fakültesi, Histoloji ve Embriyoloji Anabilim Dalı, İstanbul, Türkiye

[3] İstanbul Üniversitesi Aziz Sancar Deneysel Tıp Araştırma Enstitüsü, Genetik Anabilim Dalı, İstanbul, Türkiye

ORCID: E.S. 0000-0003-0320-5784;
M.S. 0000-0002-8648-213X

Sorumlu yazar/Corresponding author:
Müge Sayitoğlu,
İstanbul Üniversitesi Aziz Sancar Deneysel Tıp Araştırma Enstitüsü, Genetik Anabilim Dalı, İstanbul, Türkiye
E-posta: mugeay@istanbul.edu.tr

Başvuru/Submitted: 14.05.2020
Kabul/Accepted: 12.06.2020

Atıf/Citation: Sun E, Sayitoglu M. Whole Genome RNA Sequencing Analysis Algorithm in Leukemia Model. Sağlık Bilimlerinde İleri Araştırmalar Dergisi 2020; 3(2): 26-34.
https://doi.org/10.26650/JARHS2020-737495

GİRİŞ

Yeni Nesil Dizileme (YND) teknolojileri, insan genom projesinin tamamlanmasıyla beraber projenin çıktısı olarak sağlık çalışmalarında önemli bir yer edinmiştir. YND teknolojileri genomik, traskriptomik, epigenetik düzenleyiciler ve genomdaki varyasyonları hakkında yüksek hassasiyette veriler sunmaktadır (1). Bu veri, örnek olarak kullanılacak nükleik asidin fragmente edilerek her bir fragmanın eş zamanlı paralel olarak çok sayıda okunmasıyla gerçekleştirilmektedir (2). YND teknolojilerinden transkriptom dizilemenin için birincil kütüphane hazırlamada kullanılan biyolojik materyal RNA'dır ve bu başlığın altındaki tüm teknolojiler aynı zamanda RNA-Dizileme olarak da adlandırılır. RNA Dizileme metodolojisinde RNA kütüphanesi hazırlandıktan sonra, ribozomlar uzaklaştırılıp, takiben cDNA sentezlenen bir örnek hazırlama protokolü ile başlamaktadır (3). Transkriptom dizilemenin en büyük avantajı, aslında bir gen anlatım çalışması olmasıdır. Bu özelliğinden dolayı, tüm transkriptom verisinin anlık bir görüntüsünü bize sunmaktadır. Tüm transkriptom verisi, hücresel transkripsiyonel profilinin kapsamlı olarak incelenmesini sağlamaktadır. Bu diğer YND teknolojilerinin bize sunamadığı alternatif kırpılma bölgelerinin, novel transkriptlerin ve gen füzyonlarının tespitini sağlamaktadır (4,5). Tüm bu sunduğu avantajların yanında RNA-dizileme teknolojisi aynı zamanda yeniden hizalama yöntemi kullanarak (6) 18-22 baz çiftinden oluşan gen anlatım sırasında düzenleyici, gen baskılayıcı yada gen susturucu olarak görev alan ve transkripsiyonel ve translasyonel düzenleyici etkisini olan küçük RNAlar, kodlanmayan RNAlar ve mikroRNAların anlatım profilleri hakkında da bilgi sunmaktadır. Tüm bu avantajlarının yanında, ham verinin referans veriye doğru hizalanmasından başlayıp, doğru iş akış algoritması kullanımından ve çıkan sonuçların doğru yorumlanması RNA dizilemenin zorlukları arasındadır (7) ve RNA dizileme analizleri hakkında belirlenmiş bir altın standart bulunmamaktadır.

Akut lenfoblastik lösemi, B ve T lenfosit gelişiminin erken safhasında meydana gelen somatik genetik değişiklikler ile ortaya çıkan ve lenfositlerin aşırı artışı ile sonuçlanan bir lösemi tipidir (8). T-ALL hastaları için prognostik öneme sahip ve hastalığın takibinde kullanılabilecek güvenilir bir genetik anlatım değişimi veya varyasyon bulunmamakla birlikte büyük bir grup hasta için, doğrudan tedavi protokolünü ve tedavide yararlanılacak yeni hedef proteinleri belirlemede esas olacak moleküler alt yapı ve sınıflandırma da bilinmemektedir (9).

Bu çalışmamızda, T-ALL hücre hatlarını kullanarak RNA dizileme için oluşturulmuş farklı analiz araçlarını karşılaştırarak genomik arka planı oldukça karışık olan bu hastalık grubu için en uygun analiz algoritma yaklaşımının belirlenmesi amaçlanmıştır.

GEREÇ VE YÖNTEM

Örneklem

Çalışma Lösemilerin tanıları için oluşturulmuş immünfenotiplendirme protokollerine göre gruplandırılmış ve ticari olarak üretilen ALL hücre hatlarından Jurkat ve Molt4 seçilmiştir. T-ALL hastalarında sıklıkla aktivasyonu görülen sinyal ileti yolları için bir model oluşturmak amacıyla seçtiğimiz hücre hatları LiCl ile aktive edilerek dizilenmiştir (10). 4 x10^6 hücre, 12 ml %10 FCS, 2mM L-glutamin, streptomisin (100mg/mL) ve penisilin (100U/mL) içeren RPMI 1649 besiyeri içerisinde 240µl 1M LiCl eklenerek bir gece inkübe edilmiştir. LiCl muamelesi kanonik WNT yolağındaki ß-katenin yıkım kompleksinde kilit rolü olan GSK3ß inhibitörü olarak görev alarak Wnt yolağının aktivasyonunu sağlamaktadır (11).

Kontrol örnekleri olarak da CD3+/CD4+/CD8- ve CD4+/CD8- sağlıklı timosit alt tiplerine ait RNA-dizileme verisi ve 12 T-ALL hastasına ait Gen Anlatım Omnibus (GEO) veri tabanından alınan RNA-dizileme verisi kullanılmıştır (GSE48173).

RNA İzolasyonu

Hücre kültüründen toplanan Jurkat ve Molt4 hücreleri, 600 µl Solüsyon D içinde homojenize edildi ve kit protokolüne uygun bir şekilde total RNA izole edilmiştir (Qiagen, Almanya). Elde edilen RNA materyallerinin bozulma miktarına dayalı prensip ile 28S/18S oranını ölçmek için RNA örneklerinin kalite kontrollerini çipli sistem olan Bioanalyzer (Agilent, ABD) ile gerçekleştirilmiştir.

Transkriptom Dizileme

Örnek grupları arasındaki transkriptom düzeyindeki farklılıklarını görebilmek için gerçekleştirdiğimiz RNA dizileme hizmet alımıyla gerçekleştirilmiştir. Örneklerimiz, Illumina HiSeq 2500 teknolojisi ile dizilenmiştir. İllumina cihazının dizileme teknolojisi sentezleyerek dizileme (Sequence by Synthesis (SBS)) teknolojisine dayandırılmıştır.

Biyoenformatik Analizler

RNA dizileme veri analizinde mevcut bir altın standart bulunmamaktadır. Bu nedenle farklı algoritma kombinasyonları test edilip en uygun algoritmanın belirlenmesi amaçlanmıştır. Çalışmada karşılaştırma için kullanılan tüm analiz araçları Tablo 1'de belirtilmiştir.

Dizileme Verilerinin Kalite Kontrolü ve Veri Temizleme

RNA dizileme verisinin kalite kontrol değerlendirmeleri ve tekrarlanan adaptörlerin tespiti, "Fasta Quality Control" (FastQC) (Babraham Bioinformatics) tespit aracı ile gerçekleştirilmiştir. FastQC aracının "fastq-mcf" alt aracı ile, dizileme sonucunda elimizdeki FastQ formatındaki ham veri, ön işleme raporlaması ile kalite değerlendirmesi yapıp okuma kalite değerleri için belirlenen eşik değere göre değerlendirilmiştir (28 baz çifti ve üzeri okuma değerleri kabul edilir). Bu eşik değer taban alınarak düşük kalitedeki okumalar temizlenip kırpılmış, örneklerin dizileme esnasında karışmamaları için eklenen işaret olan adaptörlerden de kalanlar temizlenmiştir.

Tablo 1. Çalışmamızdaki analizlerde kullanılan analiz programları ve veri tabanları

Program/Veri Tabanı Adı	**Açıklama**
fastq-dump	SRA formatındaki RNA dizileme verisini FASTQ biçimine dönüştürüp analiz edilemeye uygun hale getirir.
fast-qc (Fast Quality Control)	FASTQ dosyalarının kalite kontrol değerlendirmelerini ve tekrarlanan adaptörlerini tespit eder.
Pyton	Pyton dilinde yazılan program ile tekrarlanan adaptörleri FASTQ formatında kaydeder.
Fast-mcf	Belli okuma değeri altında kalan bölgeleri kırpıp, adaptörleri kendi bölgeleri için "false-pozitif"liği önlemek için veriden temizler.
RSEM Generator	Elde edilen referans genomu hizalama için hazırlayıp, "bowtie2" algoritmasına göre hizalamayı gerçekleştirmektedir.
STAR	Elde edilen ham veriyi referans genoma hizalamayı gerçekleştiren ve literatürde en çook tercih edilen araçtır.
The R Project for Statistical Computing	R grafik ve kapsamlı istatistiksel analizlerin yapılmasını sağlayan ücretsiz bir ara yüzdür.
Bioconductor	Yüksek çözünürlüklü verilerin analizinde R ara yüzünü kullanan bir biyoenformatik kaynaktır.
EBSeq2	RNA dizileme grupları arasında farklı anlatıma uğrayan genleri tespit eden veri analizidir.
Cufflinks	Cole Trapnell's Lab tarafından geliştirilmiş RNA dizileme grupları arasındaki anlatım farklılıklarını belirleyen araçtır.
ClustVis	İnteraktif ısı haritaları ile genom verilerinin görselleştirilmesinde ve analizinde kullanılan bir uygulama birimidir.
UCSC http://genome.ucsc.edu GRCh337/hg19	Referans sekans ve genomla ilgili bilgiler içeren, kullanıcıya çalışma alanı sağlayan bir veri tabanıdır.
DAVID (DAVID Bioinformatic Database) http://david.abcc.ncifcrf.gov	Fonksiyonel anotasyon analizlerinde kullanılan bir veri tabanıdır
GEO (Gene Expression Omnibus) http://www.ncbi.nlm.nih.gov/geo/	NCBI veri tabanının altında, ekspresyon ve varyasyon verilerinin depolandığı ve paylaşıldığı bir veri tabanıdır .

Referans Genoma Hizalanması ve Gen Anlatımlarının Profillendirilmesi

Kalite kontrol ve temizleme sonrasında değerlendirmeye uygun hale getirdiğimiz veriler, literatürde en çok kullanılan BOWTIE2 ve STAR hizalama araçları ile hg19 referans genoma hizalanmıştır. Hizalama sırasında genlerin uzunlukları ve ilgili gen bölgesi için okuma derinliği değerlendirerek "Fragment Per Kilobase Per Million" (FPKM) şeklinde ifade edilen gen anlatım değerleri normalizasyonu tamamlanmış şekilde hesaplanmıştır. Kalite kontrol verisinde yer alan okuma derinliği ve fragman uzunluğu parametrelerinin önemi Tablo 2'de hesaplama ile gösterilmiştir.

$$FPKM = \frac{ER \times 10^9}{EL \times MR \times 2}$$

ER) Gen bölgesi için okuma derinliği, *EL*) İlgili gen bölgesinin uzunluğu, *MR)* Deneydeki toplam derinlik değeri.

Gen Anlatım Analizi

Elde ettiğimiz FPKM değerlerini kullanarak tüm veri içerisindeki anlamlı olarak anlatımı değişen genleri tespit etmek için EBSeq paketi ve Cufflinks aracı kullanılmıştır. EBSeq, R programlama içerisindeki biyolojik analizler yapılabilecek platform olan Bioconductor bünyesinde bir pakettir ve gen anlatım profilleri hesaplamalarında kullanılmıştır. Cufflinks (Cole Trapnell's Lab) ise, RNA-dizileme için diferansiyel gen anlatımı hesaplama üzere ortak üç matematik ve hesaplamalı Biyoloji laboratuvarlarının geliştirdiği bir araçtır (http://cole-trapnell-lab.github.io/cufflinks/) (12).

Gen anlatım hesaplamaları sonucunda anlamlı değişiklik gösteren ($p<0,05$) genlerin logaritmik kat değişimleri tespit edilmiş ve bu genler ve değişim kat sayıları ClusVis (13) programı aracılığıyla R dilinde yazılmış program ile ısı haritasına yerleştirilmiştir. Tespit ettiğimiz genler, "The Database for Annotation, Visualization and Integrated Discovery" (DAVID) veri tabanı kullanılarak yolak ve zenginleştirme analizleri yapılmıştır (http://david.abcc.ncifcrf.gov) (14).

Gen Anlatım Analizi

Hizalanmış veri içerisinden, WNT yolağı ilişkili genlerin FPKM değerleri çekilip EBSeq ve Cufflinks ile bu genlerin anlatım profilleri hesaplanmış, anlamlı olan genlerin ($p<0.05$) logaritmik kat değişimleri tespit edilmiş ve bu genler ısı haritasında görselleştirilmiştir.

Alternatif Kırpılma Ürünlerinin Analizi

mRNA oluşurken ortaya çıkan gen ürünleri, alternatif noktalardan kırpılma sonucunda farklı ürünler oluşabilir. Bu oluşan alternatif ürünlerin anlatım miktarlarını ifade eden FPKM değerleri, GraphPad programına yerleştirilip LiCl ile uyarılmış hücre hatlarında, kontrol örneklerinde ve hastalardaki durumlarını gösteren dağılım grafiği çizilmiştir.

Varyant Analizleri

Hizaladığımız verinin çıktılarından biri de "Variant Calling File" (.vcf) olan verideki varyasyonlar

Tablo 2. Okuma derinliği ve gen uzunluğuna göre normalize edilmemiş (öncesi) ve edilmiş (sonrası) değerlerin temsili olarak gösterimi

	Gen Adı	**1. Tekrar Okuma Sayısı**	**2. Tekrar Okuma Sayısı**	**3. Tekrar Okuma Sayısı**
Öncesi	A Geni (2 kb)	10	12	30
	B Geni (4 kb)	20	25	60
	C Geni (1 kb)	5	8	15
	D Geni (10 kb)	0	0	1
Sonrası	**Gen Adı**	**1. Tekrar Okuma Sayısı**	**2. Tekrar Okuma Sayısı**	**3. Tekrar Okuma Sayısı**
	A Geni (2 kb)	1.43	1.33	1.42
	B Geni (4 kb)	1.43	1.39	1.32
	C Geni (1 kb)	1.43	1.78	1.42
	D Geni (10 kb)	0	0	0.009

Farklı sayıdaki okuma değerine sahip olan farklı uzunluktaki genlerin aslında aynı gen anlatım profilini gösterebileceğini göstermek için değerler temsili olarak gösterilmiştir

Tablo 3. İki farklı araçlarla hizalanmış ve gen anlatım profilleri hesaplanmış örneklere genel bakış

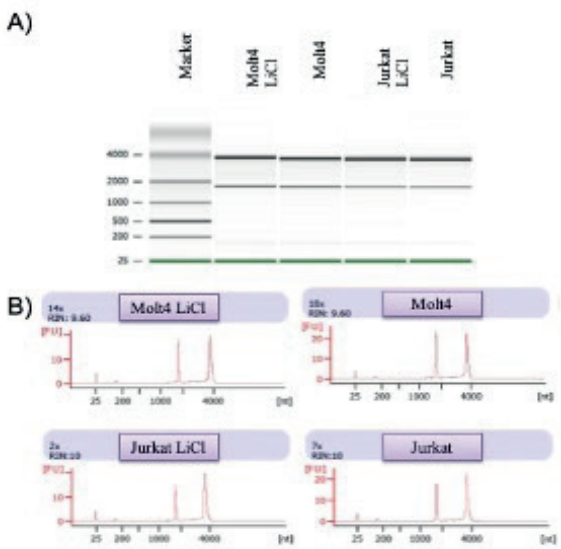
Tam Kapsamlı Analiz			
Hizalama Araçları	**LiCI ile Uyarılmış Hücre Hatları vs Timus Havuzu**	**EB-Seq**	**Cufflinks**
	Bowtie	1784	116
	STAR	1638	1677

ile ilgili bilgi içeren bir dosyadır. Hizalama sonucunda elde ettiğimiz .vcf uzantılı dosya, Illumina firmasının varyant analizleri için geliştirdiği ticari yazılım olan VariantStudio'da (v3.0.12) analiz edilmiştir.

BULGULAR

RNA Kalite Kontrolü

Yeni nesil transkriptom dizilemeden verimli sonuç alabilmek için RNA'nın 28S/18S oranını temsil eden RIN (RNA Integrity Number; RNA Bütünlük Sayısı) sayısının 7'den büyük olması gerekmektedir. Çipli sistem kullanılarak yapılan ölçümlerde (Bioanalizör, Agilent) Jurkat için RIN sayısı 10, Jurkat LiCl için 10, Molt4 için 9.6 ve Molt4 LiCl için de 9.6 olarak ölçülmüştür. Örneklerin çipli sistemdeki yürütme sonuçları Şekil 1'de gösterilmiştir.



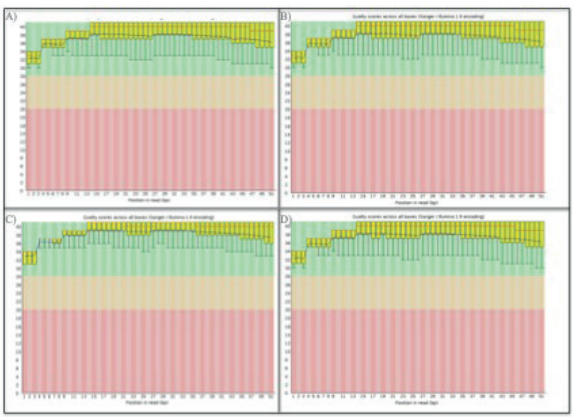
Şekil 1. RNA örneklerinin Biyoanalizör 2100 cihazındaki (Agilent) ölçüm sonuçları. A) çipli sisteme yüklenmiş örneklerin jel görüntüsü. B) 28S/18S oranının hesaplanabilmesi için gerekli okuma değerlerinin grafikleri

Enformatik Bulgular

Ön İşleme

Kalite Kontrol ve Temizleme

Dizileme ham verimizde toplam 71.018.097 okuma sayısı, 30 baz üzerindeki fragmanlarda okuma %94,81 ve 35,65 okuma derinliğine ulaşılmıştır. Dizileme sonucunda elde ettiğimiz fastq dosyalarımız, FastQC (Babraham Bioinformatics) kullanılarak dizilenen örneklerimizin kalite değerleri ve dizi içinde kalmış adaptörler belirlenmiştir. Örnekler içinde tespit edilen adaptörler kırpıldı. Şekil 2 'de örneklerin kalite kontrol grafikleri verilmiştir.



Şekil 2. RDizilenmiş RNA örneklerinin okumaları için gerçekleştirilmiş kalite kontrol analizi sonucu. Sarı barlarda gösterilen her bir fragmanın okuma kaliteleri bulundukları alanlarda belirtilmiştir. Yeşil alan; kaliteli okuma, sarı ve kırmızı alanlar ise dizilerde kalan adaptörler veya kısa okumalar olduğunu işaret etmektedir

Referans Genoma Hizalanması ve Gen Anlatımlarının Profillendirilmesi

Örnekler literatürde en çok tercih edilen Bowtie2(15) ve STAR(16) hizalama araçları kullanılarak hg19 referans genomuna hizalandı. Hizalanan veride gen okuma değerlerini referans alarak her genin mutlak anlatım değerini ifade eden FPKM değerleri hesaplanmıştır.

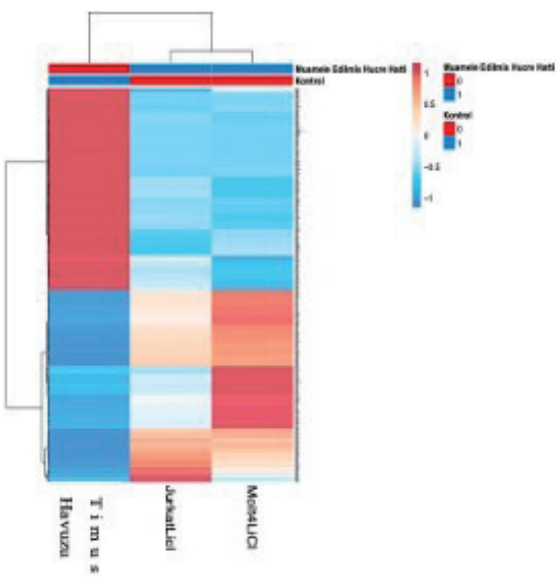
Gen Anlatım Analizi

Örnek havuzundan belirlenen grupların her biri EBSeq ve Cufflinks analiz araçları ile farklı anlatım profili gösteren genler tespit edilmiştir ($p<0.05$). Bu sonuçları dikkate aldığımızda, tam kapsamlı analiz-

de Bowtie hizalama sonucunda elde edilen verilerde EBSeq kullanılarak gen anlatımlarını hesaplandığında 1784, Cufflinks kullanıldığında ise 116 diferansiyel anlatım farklılığı gösteren gen tespit edildi. STAR hizalama aracından elde edilen veri sonucunda ise Bowtie ile 1638 ve Cufflinks ile 1677 adet anlatım farklılığı gösteren gen tespit edilmiştir. Anlatım farklılığı olduğunu tespit ettiğimiz gen sayılarında en çok veriye STAR hizalama aracı ile EBSeq hesaplama aracının en doğru yaklaşım gösterdiği ve bu araçlar kullanarak devam edilmeye karar verilmiştir.

Tüm Transkriptom Analiz

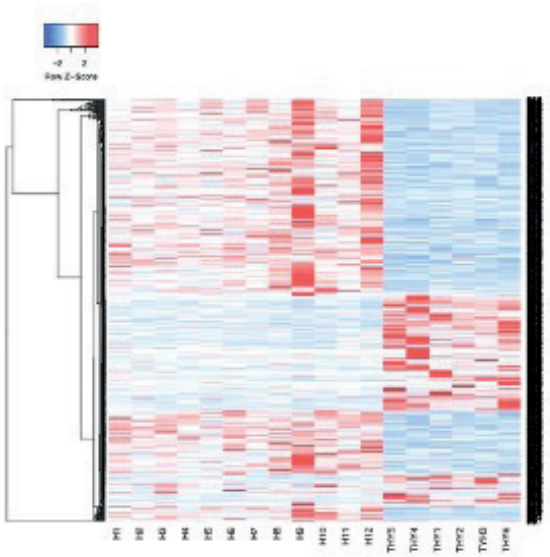
Isı haritası için oluşturduğumuz algoritmada en çok 1200 gen ile çalışılması mümkündür. Seçilen araçlardan elde edilen sonucunda elde ettiğimiz sonuçlar doğrultusunda, logaritmik değer ve p değerleri açısından değerlendirildiğinde en anlamlı sonuç veren 1200 gene ait gen anlatım profilleri, ısı haritalarına yerleştirilip gen anlatım düzeyleri görselleştirilmiştir (Şekil 3). LiCl muamelesi görmüş hücre hatlarının, kontrol örneği olan timüs havuzundan farklı bir anlatım profili gösterdiği, ısı haritasında herhangi bir etiketleme yapılmadan kümelenmesi ile gösterilmiştir.



Şekil 3. Tüm transkriptom analizi sonucu, LiCl ile muamele edilmiş hücre hattı ve hastalardaki gen anlatım profillerinin kesişim kümesi (n=426)

Hücre hatları için yaptığımız analizler, veri tabanlarından elde edilen 12 adet T-ALL hastasına ait RNA dizileme sonuçları verisinde de uygulanmış ve elde ettiğimiz sonuçlar doğrultusunda hasta grubunun, kontrol örneği olan timus havuzundan farklı bir anlatım profili gösterdiği, ısı haritasında herhangi bir etiketleme yapılmadan kümelenmesi ile gösterilmiştir.

Hem uyarılmış hücre hatlarına ait veriler hem de hastalardan elde edilen veriler birleştirildiğinde, 426 adet genin ortak profil gösterdiği belirlenmiştir (Şekil 4).



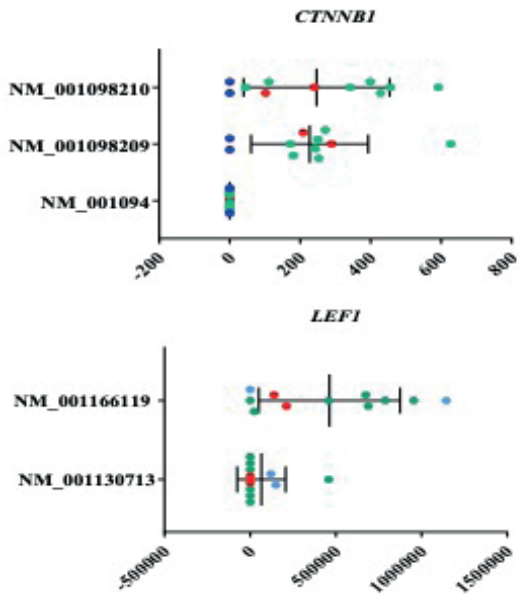
Şekil 4. Hasta verilerinde, tüm transkriptom analizinin gen anlatım ısı grafiği. "H" kodları hasta örnekleri, "THY" kodları kontrol timus havuzunu temsil etmektedir

Alternatif Kırpılma Ürünleri

T-ALL hastalarında daha önceden tanımlanmış ve WNT yolağında kilit rolü olan genler seçilerek alternatif kırpılma ürünlerinin anlatımları hesaplanmıştır. T-ALL hastaları ve sağlıklı bireylerde anlatım düzeylerinin farklılığı veritabanlarında *CTNNB1* için üç transkript ve *LEF1* için tanımlanmış 2 transkript tespit edilmiş ve bunların dokular arasındaki anlatım düzey farklılıkları Şekil 5'te gösterilmiştir.

Varyant Analizleri

Yöntemimizin doğruluğunu gösterebilmek adına seçilen hücre hatlarında daha önce tanımlanmış varyasyonlar RNA dizileme yöntemi ile da analiz edil-



Şekil 5. *CTNNB1* ve *LEF1* için tanımlanmış transkriptlerin dokular arasındaki anlatım farklılıkları. (Kırmızılar Uyarılmış hücre hatları; Yeşiller hastaları; Maviler de kontrol örneklerini temsil etmektedir.)

miştir. Jurkat hücre hattı için literatürde tanımlanan 12. kromozomda yer alan olmuş CDK4 ve 1. kromozomda lokalize olmuş *ARF1* gen varyasyonları; Molt4 hücre hattı için literatürde tanımlanan 1. kromozomda yer alan olmuş *NRAS* geninde tanımlanan varyasyon ve yine 1. kromozomda lokalize olmuş *ARF1* genindeki varyasyonlar tespit edilmiştir.

TARTIŞMA

Yeni nesil RNA dizileme, son derece dinamik olan hücre ya da dokuya özgün transkriptom repertuarını belirlemede kullanılabilecek yeni bir teknolojidir. Yeni nesil RNA dizi analizi teknolojisi uygulamaları ile kanser genom çalışmaları yeni bir boyut kazanmıştır. RNA dizileme sayesinde gen anlatım profilleri, alternatif kırpılma ürünleri ve varyasyonlar gibi genoma özgü değişiklikleri yüksek çözünürlükte tespit edilebilmenin yanında miRNA bağlanma bölgeleri gibi fonksiyonel önemi olan verinin de tespitine olanak sağlamıştır. RNA dizileme ile yüksek çözünürlükte verilerin elde edilebilmesi için ortalama 50 milyon okuma sayısı, %85 üzerinde 30 bazdan uzun okuma ve 20 kattan fazla okuma derinliği olması gerekmektedir. Bu çalışmadaki veri kalitesini değerlendirdiğimizde, anlamlılığı yüksek veriler elde edebilmek için bu eşik değerlerin oldukça üzerinde kaliteli bir dizileme verisi elde edilmiştir.

RNA dizilemenin avantajlarından biri alternatif kırpılma ürünlerinin anlatım farklılıklarını saptamıştır. Zhao ve arkadaşlarının, içerisinde bizim de kullandığımız Jurkat ve Molt4'ün e bulunduğu lösemi hücre hatları ile yaptıkları çalışmada, *IKZF2* geninin farklı transkript ürünlerinin T hücre proliferasyonuna ve apoptozuna müdahale ettiği gösterilmiştir (17). Adamia ve arkadaşlarının yayınladıkları derlemede de, kırpılma hatalarının malin dönüşümlere sebep olabileceği ve aday olarak belirlenen alternatif kırılma ürünlerinin özellikle ilaca dirençli klonların tedavisinde kullanılabileceği belirtilmiştir (18). Çalışmamızdan elde ettiğimiz alternatif kırpılma ürün sonuçlarına baktığımızda ise, *LEF1* için dört alternatif kırpılma ürününden biri (NM_001130713) tümör baskılayıcı özellikteki transkript varyant, diğeri ise (NM_001166119) onkojenik özellikli transkript varyanttır. Elimizdeki verilerde uyarılmış hücre hatlarında ve altı hastada tümör baskılayıcı özellikteki transkriptin hiç anlatıma girmediği, sadece bir hastada ve sağlıklı kontrol örneklerinde anlatımın görüldüğünü tespit edilmiştir. Onkojenik özellikte olan transkriptin ise; uyarılmış hücre hatlarında, altı hastada ve Jurkat için kontrol örneğinde anlatımını belirledik. Literatürde, tespit ettiğimiz iki alternatif kırpılma ürününün anlatımları hakkında bilgiye ulaşılamamıştır.

Varyant analizlerinde önceden bildirilen (Jurkat için *CDK4* ve *ARF1*; Molt4 için *NRAS* ve *ARF1)* genetik varyasyonların (19). RNA dizileme varyant analizi ile de hassas bir şekilde saptanabildiğini görülmüştür. Bu bulguya ek olarak, Tomov ve arkadaşlarının 2016 yılında yayınladıkları makalede, RNA Dizileme analizlerinde kullandıkları algoritma öncelikle ham verinin FASTQC aracıyla kalite kontrolünün tespitiyle başlanmıştır; ardından veriler hizalanıp bizim de kullandığımız EBSeq aracı ile gen anlatım düzeyleri hesaplanmış ve logaritma 2 tabanındaki artış miktarları gösterilmiştir (20). Bu algoritma bizim çalışmamız için belirlediğimiz algoritma ile birebir

örtüşmektedir ve iş akış protokollerimizin doğruluğu bir farklı yoldan da gösterilmiştir.

RNA dizileme yöntemi, gen anlatım analizleri için yüksek hassasiyetteki mikro-dizileme yöntemini geçersiz kılacak kadar güçlü bir transkriptom analiz yöntemi olarak karşımıza çıkmıştır. Analizlerinin göreceli olarak daha kolay olduğu mikro-dizi yöntemi, anlatımları anlamlı düzeyde değişen genlerin tespiti için çoğu çalışmada hala tercih edilen bir yöntemdir. RNA dizileme analizleri ise, araştırıcıların analiz yapmak için bir çok aracı kombine bir şekilde kullanması gereken ve zorluklarla karşı karşıya kaldığı bir yöntemdir. Ancak, mikro-dizilerde yapılan analiz sonucunda, önceden bilinen transkriptlerin anlatım düzeylerine ulaşma imkanı varken, RNA dizileme analizleri sonucunda; gen anlatım düzeyleri, kırpılma bölgelerinin tespit edilmesi ve *novel* transkriptlerin tespit edilebilmektedir (21). Literatürde, 31 T-ALL hastası ve 18 hücre hattı toplamda 49 T-ALL örneği ile yapılan bir çalışmada ekzom ve transkriptom datası karşılaştırılarak, ekzomda tespit edilmeyip transkriptom datasında tespit edilebilen yeni driver mutasyonları göstermişlerdir (22).

Sonuç olarak çalışmamızda farklı algoritmalar karşılaştırılmış ve lösemi örneklerinin analizi için RNA-Dizileme data analiz algoritması oluşturulmuş, gen anlatım düzeyindeki farklılıklar; alternatif kırpılma ürünlerinin doku spesifik anlatımları ve veri tabanında T-ALL ile ilişkilendirilmiş varyantları belirleyeceğimiz araçlar ve RNA dizileme analiz algoritma yaklaşımı belirlenmiştir. Bu çalışmada hastalığı temsil eden hücre hatları ve açık veri tabanlarından küçük bir T-ALL hasta kohortu verisi kullanılmıştır. Enformatik analizlerin güvenilirliği daha büyük veri setlerinde yapılacak analziler ve validasyon çalışmaları ile kesinlik kazanacaktır.

Hakem Değerlendirmesi: Dış bağımsız.

Peer Review: Externally peer-reviewed.

Yazar Katkıları: Çalışma Konsepti/Tasarım- E.S., M.S.; Veri Toplama- E.S., M.S.; Veri Analizi/Yorumlama- E.S., M.S.; Yazı Taslağı- E.S., M.S.; İçeriğin Eleştirel İncelemesi- M.S..; Son Onay ve Sorumluluk- E.S., M.S.

Author Contributions: Conception/Design of Study- E.S., M.S.; Data Acquisition- E.S., M.S.; Data Analysis/Interpretation- E.S., M.S.; Drafting Manuscript- E.S., M.S.; Critical Revision of Manuscript- M.S.; Final Approval and Accountability- E.S., M.S.

Çıkar Çatışması: Yazarlar çıkar çatışması beyan etmemişlerdir

Conflict of Interest: Authors declared no conflict of interest.

Finansal Destek: Bu çalışma, İstanbul Üniversitesi Bilimsel Araştırma Projeleri Birimi tarafından desteklenmiştir. (Proje No: TYL-2016-20440)

Financial Disclosure: This study was supported by Istanbul University Scientific Research Projects Unit. (Project No: TYL-2016-20440)

KAYNAKLAR/REFERENCES

1. Behjati S, Tarpey PS. What is next generation sequencing? Arch Dis Child Educ Pract Ed. 2013;98(6):236–8.
2. Johnsen JM, Nickerson DA, Reiner AP. Massively parallel sequencing: The new frontier of hematologic genomics. Blood. 2013;122(19):3268–75.
3. Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. Nature Reviews Genetics. 2009.
4. Ozsolak F, Milos PM. RNA sequencing: Advances, challenges and opportunities. Nat Rev Genet. 2011;12(2):87–98.
5. Costa V, Angelini C, De Feis I, Ciccodicola A. Uncovering the complexity of transcriptomes with RNA-Seq. J Biomed Biotechnol. 2010;2010:19.
6. David M, Dzamba M, Lister D, Ilie L, Brudno M. SHRiMP2: Sensitive yet Practical Short Read Mapping. Bioinformatics [Internet]. 2011 Apr 1 [cited 2018 Jul 13];27(7):1011–2. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr046
7. Williams AG, Thomas S, Wyman SK, Holloway AK. RNA-seq Data: Challenges in and Recommendations for Experimental Design and Analysis. Curr Protoc Hum Genet. 2014;83.

8. Terwilliger T, Abdul-Hay M. Acute lymphoblastic leukemia: a comprehensive review and 2017 update. Blood Cancer J. 2017;7(6):e577.

9. Van Vlierberghe P, Ferrando A. The molecular basis of T cell acute lymphoblastic leukemia. J Clin Invest. 2012;122(10):3398–406.

10. Galli C, Piemontese M, Lumetti S, Manfredi E, Macaluso GM, Passeri G. GSK3b-inhibitor lithium chloride enhances activation of Wnt canonical signaling and osteoblast differentiation on hydrophilic titanium surfaces. Clin Oral Implants Res. 2013 Aug;24(8):921–7.

11. Gottardi CJ, Gumbiner BM. Distinct molecular forms of β-catenin are targeted to adhesive or transcriptional complexes. J Cell Biol. 2004 Oct 25;167(2):339–49.

12. Cufflinks [Internet]. [cited 2020 Jun 23]. Available from: http://cole-trapnell-lab.github.io/cufflinks/

13. Metsalu T, Vilo J. ClustVis: A web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. Nucleic Acids Res. 2015;43(W1):W566-70.

14. DAVID Functional Annotation Bioinformatics Microarray Analysis [Internet]. [cited 2020 Jun 23]. Available from: https://david.ncifcrf.gov/

15. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Apr 4;9(4):357–9.

16. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. Sequence analysis STAR: ultrafast universal RNA-seq aligner. 2013 [cited 2020 Jun 10];29(1):15–21. Available from: http://code.google.com/p/rna-star/.

17. Zhao S, Liu W, Li Y, Liu P, Li S, Dou D, et al. Alternative splice variants modulates dominant-negative function of Helios in T-cell leukemia. PLoS One. 2016;11(9):e0163328.

18. Adamia S, Pilarski P, Bar-Natan M, Stone R, Griffin J. Alternative Splicing in Chronic Myeloid Leukemia (CML): A Novel Therapeutic Target? Curr Cancer Drug Targets. 2013;13(7):735–48.

19. Bennett JM. The Leukemia-Lymphoma Cell Line Facts Book. Leukemia Research. 2002.

20. Tomov ML, Olmsted ZT, Dogan H, Gongorurler E, Tsompana M, Otu HH, et al. Distinct and Shared Determinants of Cardiomyocyte Contractility in Multi-Lineage Competent Ethnically Diverse Human iPSCs. Sci Rep. 2016;6(37636).

21. Ramsköld D, Kavak E, Sandberg R. How to analyze gene expression using RNA-sequencing data. Methods Mol Biol. 2012;802:259–74.

22. Kalender Atak Z, Gianfelici V, Hulselmans G, De Keersmaecker K, Devasia AG, Geerdens E, et al. Comprehensive Analysis of Transcriptome Variation Uncovers Known and Novel Driver Events in T-Cell Acute Lymphoblastic Leukemia. PLoS Genet. 2013;9(12):e1003997.