Düzce Üniversitesi Bilim ve Teknoloji Dergisi

Araştırma Makalesi

Normalizasyon Yöntemlerinin RNA- Seq Verileri Üzerinde Çıkarılan Gen Birlikte İfade Edilme Ağlarının Performansına Etkisi

Mustafa Özgür CİNGİZ [a,*]

[a] *Bilgisayar Mühendisliği Bölümü, Mühendislik ve Doğa Bilimleri Fakültesi, Bursa Teknik Üniversitesi, Bursa, TÜRKİYE*

** Sorumlu yazarın e-posta adresi: mustafa.cingiz@btu.edu.tr*

DOI: 10.29130/dubited.803846

ÖZET

Protein sentezi sürecinde meydana gelen farklılaşmaların metabolik hastalıklar, kanser gibi kompleks hastalıklara neden olduğu farklı çalışmalarda belirtilmiştir. Protein sentezindeki değişimlerin anlaşılması için proteinleri oluşturan genlerin belirlenmesi ve bu genlerin diğer genlerle ilişkilerin ortaya çıkarılması gerekmektedir. Yeni nesil dizileme teknikleriyle hastalıklara neden olan moleküler düzeyde ilişkilerin doğruluklu olarak belirlenmesi kolaylaşmıştır. Gen birlikte ifade edilme (GBİE) ağları düzenleyen-düzenleyici ilişkisi içermeden benzer biyolojik süreçlere katılan genler arasındaki ilişkileri araştırmacılara göstermektedir. Çalışmamızda RNA-Seq verileri kullanılarak prostat kanseriyle ilişkili GBİE ağları elde edilmiştir. RNA- Seq verileri farklı nükleotit uzunluğundaki genlerden ve farklı sayıda okumalar içeren örneklerden oluştuğu için normalizasyon teknikleri moleküler ilişki çıkarımında önem taşımaktadır. Çalışmamızda gen birlikte ifade edilme ağları ham veri ve farklı iki normalizasyon yaklaşımı olan M- Değerinin Kırpılmış Ortalaması (MDKO), Göreceli Log İfadesi (GLİ) hesaplamalarıyla ayrı ayrı oluşturulmuş veriler üzerinde çıkartılarak örtüşme analizi ve topolojik performans değerlendirilmesi yapılmıştır. Örtüşme analizine göre normalize edilmiş RNA- Seq verileri kullanarak elde edilmiş gen birlikte ifade edilme ağlarının ham verilere göre daha fazla literatürde bulunan ilişkileri tahmin ettiği gözlemlenmiştir. İki normalizasyon yöntemiyle elde edilen GBİE'lere ait örtüşme analizi performans metrikleri değerleri ise birbirlerine yakın çıkmıştır. Topolojik değerlendirme sonuçlara göre normalize edilmiş veriler üzerinde elde edilen GBİE ağlarının ölçeksiz ağ tanımına daha yakın olduğu gözlemlenmiştir. Çalışmamızda aynı zamanda ham ve normalize edilmiş veriler üzerinde GBİE ağ çıkarım algoritmaları olan C3NET, ARACNE ve WGCNA yaklaşımlarının performansları da karşılaştırılmıştır.

Anahtar Kelimeler: *RNA- Seq, Normalizasyon, Gen Ağı Çıkarımı, Gene Birlikte İfade Ağları*

Effect of Normalization Methods on the Performance of Gene Co-expression Networks Inferred on RNA-Seq Data

ABSTRACT

Different studies prove that differentiation on protein synthesis causes different metabolic disorders such as cancer and diabetics. The inference of disease related genes and to derive their interactions enable us to understand the differentiation on protein synthesis. Next generation sequencing techniques can reveal relations of diseases more precisely at molecular level. Gene co-expression networks can reveal interactions between genes without regulator-regulatee information. We utilized RNA- Seq data to infer gene co-expression networks of prostate cancer in our study. RNA- Seq data consists of genes whose nucleotide size may be different from sample to sample. Sample sizes of RNA- Seq data also vary for each samples. RNA- Seq data normalization is an important task to infer robust and reliable gene co-expression networks. We utilized normalized RNA- Seq

Geliş: 01/10/2020, Düzeltme: 29/01/2021, Kabul: 30/01/2021

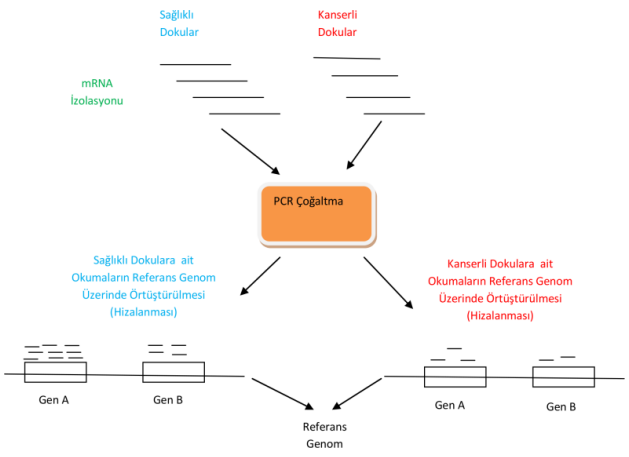
data that are obtained using two different normalization methods ,which are Trimmed Mean of M- values (TMM) and Relative Log Expression (RLE), and raw RNA- Seq data to infer gene co-expression networks for the performance comparison. We applied overlap and topological analyses to evalaute the performance of raw data based GCN, normalized data based GCNs in our study. Normalization on RNA-Seq data leads to predict more validated gene- gene relations, which are evaluated in overlap analysis, than gene- gene relations on raw dataset. Two normalization methods based gene co-expression networks present similar performance results in overlap analysis. GCNs that are derived from TMM and RLE normalizations resemble scale free topology more than raw based GCN in topological assessment. We also compare the performance results of gene network inference algorithms, which are C3NET, ARACNE and WGCNA, on raw and normalized datasets.

Keywords: *RNA- Seq, Normalization, Gene Network Inference, Gene Co-expression Networks*

I. GİRİŞ

Hastalıklara neden olan genlerin keşfi ve bu genler arasındaki ilişkiler sonucunda ortaya çıkan biyolojik süreçler hastalıkları anlamak açısından önem taşımaktadır. Yeni nesil dizileme teknikleri ile elde edilen veriler hücrelerde farklı fenotiplere neden olan proteinler ve o proteinlerle ilişkili genlerin keşfinde sıklıkla kullanılmaktadır. Yeni nesil dizileme tekniklerinde ilk ve en sık kullanılanılan veri kümeleri mRNA gen ifadesi verileridir. Farklı fenotiplere ait örneklerdeki mRNA ifade değerlerindeki değişimler kullanılarak hastalıklara ait farklı ifade edilen genler (differentially expressed genes) belirlenebilmekte ve hastalıkla ilişkili genler çıkarılmaktadır [1]. Dizileme verileriyle farklı biyolojik örneklerdeki mRNA değerleri kullanılarak gen- gen, transkripsiyon faktörü- gen, miRNA- gen gibi ilişkiler belirlenerek hastalıklara neden olan ilişkileri molekül seviyede anlaşılmasını sağlamaktadır [2].

RNA-Seq analizi transkriptoma kapsamlı bir bakış açısı sağlamasının yanı sıra RNA'yı üreten DNA alanlarının belirlenmesi ve bu alanların genlerle ilişkilendirilmesinde de sıklıkla kullanılmaktadır. RNA-Seq analizi, RNA moleküllerinden DNA dizilimi parçalarını elde etmeye yarayan deneysel adımları içerir. RNA-Seq ile genom üzerinde genlere ait daha keşfedilmemiş gen izoformlarının belirlenmesi mümkündür [3]. RNA-Seq analiziyle tek nükleotit polimorfizmi, mutasyon, yeni gen izoformların keşfi ve gen birleşimi gibi pek çok çalışma alanında kullanılmaktadır [4]. RNA- Seq analizinin temel adımları Şekil 1'de gösterilmiştir. Bu örnek aynı zamanda sağlıklı ve kanserli dokulardan alınan örneklerin fenotiplere göre okuma sayılarındaki değişimi de sunmaktadır. Kanserli ve normal dokulardan RNA izolasyonuyla hücrelerden çıkarıldıktan sonra RNA'lar, genellikle mRNA'lar, parçalanmasıyla fragman denilen küçük parçalarına ayrılır. Tersine transkripsiyonla RNA parçaları ile cDNA'lar oluşturulur ve Polimeraz Zincirleme Tepkimesiyle 10-100 milyonlarca çoğaltılarak okumalar (read) oluşturulur. Okuma 25-125 baz çift uzunluğundaki DNA parçalarıdır ve referans genomla örtüştürülerek genome üzerindeki ilişkili olduğu genler bulunur [4-7].



Şekil 1. *RNA- Seq analiz adımları*

Literatürde hastalıklara göre farklı ifade edilen genlerin keşfinde mRNA gen ifadesi verileriyle birlikte RNA- Seq verilerini de kullanan pek çok çalışma yer almaktadır. Bu çalışmalarda RNA- Seq verileriyle farklı fenotiplere neden olan genler bulunarak farklı biyolojik veri setleri kullanılarak elde edilen genlerle karşılaştırılmakta veya entegre edilerek hastalıkla ilişkili aday genler belirlenmektedir [8-10].

Fenotiplerin belirlenmesinde sadece farklı ifade edilen genlerin belirlenmesi yetersiz kalabilmektedir. Hastalığa neden olan biyolojik süreçlerde yer alan ilişkili genlerin belirlenmesiyle hastalığın biyolojik değerlendirilmesi daha doğruluklu yapılmaktadır. Birbiriyle ilişkili genlerde düzenleyici- düzenlenen bilgisi bulunuyorsa bu ağlar gen düzenleyici ağ, bulunmuyorsa gen birlikte ifade edilme ağı (GBİEA, gene co-expression network) olarak adlandırılmaktadır. mRNA gen ifadesi verilerinde olduğu gibi RNA-Seq verileri de kullanarak gen- gen ilişkilerini içeren gen ağları literatürde son yıllarda moleküler seviyede ilişkilerin incelenmesinde kullanılmaktadır [10].

RNA- Seq ham okuma verileri veri kümesindeki örneklere sayısı ve genlerin uzunluğuna göre değişmesi nedeniyle normalize edilerek çalışmalarda kullanılmaktadır. Aghababazadeh [11] kütüphane boyutu ve örneklerdeki gürültüyü gidermek için normalizasyon yöntemlerini kullanarak farklı ifade edilen genlerin analizini göğüs kanseriyle ilgili mRNA mikrodizin verileri, RNA- Seq verileri ve RNA- Seq simülasyon verileri üzerinde gerçekleştirilmiştir. Smid ve arkadaşları [12] Kanser Genom Atlas (The Cancer Genome Atlas, TCGA) verileri üzerinde farklı ifade edilen genlerin bulunmasında normalizasyon yöntemlerini karşılaştırmıştır. Robinson [13] karaciğerde farklı fenotipe neden olan RNA- Seq verileri üzerinde farklı ifade edilen genlerin çıkarımı için normalizasyon yöntemi seçiminin etkisini incelemiştir[14]. Dillies [15] ve arkadaşları insan, a. fumigatos ve m. musculus organizmalarina ait ham ve normalize edilmiş RNA- Seq verileri üzerinde farklı ifade edilen genleri bulmaya çalışmışlardır. İlgili çalışmada da normalizasyon işlemi gerçekleştirilen verilerin ham verilerden daha başarılı sonuçlar verdiği gözlemlenmiştir. Mandelboum ve arkadaşları 35 farklı RNA-Seq veri kümesinde farklı normalizasyon yöntemleri ve ham veriyi kullanarak karşılaştırma yaparak gen kümesi zenginleştirme analizinde yanlış pozitif tahmin sayılarının MDKO ve GLİ normalizasyonu kullandığında azaldığını sonucuna varmıştır [16]. Bir başka çalışmada iki farklı RNA- Seq veri kümesi üzerinde farklı ifade edilen genlerin keşfinde normalizasyon yöntemleri karşılaştırılmış ve normalizasyon yöntemlerinin farklı ifade edilen genlerin keşfine etkisinin kısıtlı olduğu sonucu ortaya konmuştur [17]. Benzer bir çalışma zebrabalığına ait RNA- Seq verileri üzerine yapılarak farklı ifade edilen genlerin normalize verilerle daha başarılı bir şekilde keşfedildiği belirtilmiştir [18]. Paşmina keçisine ait RNA- Seq verileri ile paşmina keçisine ait özel yünün diğer hayvanlardaki biyolojik süreçlerden ayıran farklı ifade edilen genleri MDKO normalizasyonu kullanılarak elde edilmiştir [19]. Bir diğer çalışmada RNA- Seq verileri üzerinde sadece MKDO normalizasyonu yapıldıktan sonra gen düzenleyi ağlar elde edilerek yeni transkripsiyon faktör- gen ilişkileri bulunmuştur [20].

RNA Seq verileri üzerinde normalizasyon gerçekleştiren çalışmalar büyük çoğunlukla farklı ifade edilen genlerin keşfi için kullanılmıştır. Belirttiğimiz çalışmalarda M- değerlerinin kırpılmış ortalaması (MDKO) ve Göreceli log ifadesi (GLİ) normalizasyon yöntemleri total okuma sayısı normalizasyonu, üst çeyrek normalizasyonu, medyan normalizasyonu, örtüşen milyon adet okuma sayısı normalizasyonu gibi farklı normalizasyon yöntemlerinden daha başarılı sonuçlar verdiği gözlemlenmiştir [12-16].

Normalizasyon yöntemlerinin RNA- Seq verileri üzerinde GBİEA çıkarımına etkisi üzerine çalışmalar yetersizdir. Çalışmamız normalizasyonun GBİEA'ların performansına etkisini incelediği için özgün bir çalışmadır. Farklı ifade edilen genlerin belirlenmesiyle ilgili tüm çalışmalarda en başarılı iki normalizasyon yöntem olan GDİ ve MDKO çalışmamızda kullanılarak GBİEA çıkarımları gerçekleştirilmiştir. Normalizasyonun GBİEA'ların performansına etkisini gözlemleyebilmek için çalışmamızda ham RNA - Seq verileri kullanılarak da GBİEA çıkarımı yapılmıştır. Tüm sonuçlar literatür verileri kullanılarak örtüşme ve topolojik analizlere göre değerlendirilmiştir.

Giriş bölümünden sonra sonra malzeme ve yöntem bölümünde çalışmamızda kullanılan veri kümesi, GBİEA çıkartma algoritmaları ve normalizasyon tekniklerine yer verilmiştir. Bölüm 3'te GBİEA'ların

örtüşme analizi ve topolojik değerlendirilmesi yapılmıştır. Bölüm 4'te ise sonuçlarımız özetlenmiş ve literatürde yer alan diğer çalışmalarla karşılaştırılarak değerlendirmemiz tamamlanmıştır.

II. MATERYAL VE METHOD

A. VERİ KÜMESİ

Çalışmamızda The Cancer Genome Atlas (TCGA) [21] projesinin prostat kanseriyle ilişkili RNA- Seq veri kümesi GBİE ağları çıkarımında kullanılmıştır. TCGA veri kümesi 495 prostat kanseri örnekten oluşmakta ve her bir örneğin 20,502 gen parçası (probe) içermektedir. TCGA veri kümesi normalizasyon uygulanmamış ham verilerden oluşmaktadır.

B. GEN BİRLİKTE İFADE EDİLME AĞI ÇIKARIM ALGORİTMALARI

GBİEA çıkarımında regresyon tabanlı, Bayes tabanlı ve bilgi teorisi tabanlı yaklaşımlar sıklıkla kullanılmaktadır. Çalışmamızda bilgi teorisi tabanlı üç farklı algoritma ile gen ağı çıkarımı gerçekleştirilmiştir. Kullanılan gen ağı çıkarım algoritmalarından ortak bilgi değeri (mutual information) ile gen ağı çıkaran ARACNE ve C3NET algoritmalarının ön adımı olan Relevance Network yaklaşımı da ayrıca açıklanmıştır. Gen ağı çıkarımında kullandığımız bir diğer algoritma ise ilişki skoru değerine göre ağ oluşturan WGCNA algoritmasıdır.

Relevance Network (RelNet): RelNet, RNA- Seq verileri üzerinde tüm gen ikilileri arasında ortak bilgi değeri hesaplayarak ilişkileri belirler. Denklem 1'de görüldüğü gibi X ve Y genlerinin arasındaki ortak bilgi değeri çıkarım formulü gösterilmiştir. Ortak bilgi değeri bir değişkenin bilgi değerinin bir başka değişkene göre değişimini entropi hesabıyla göstermektedir. RelNet'te eşik değeri belirlenmesi için veriler permütasyon test için yüzlerce/binlerce defa karıştırıldıktan (shuffle) sonra eşik ilişki değeri belirlenir [22]. Eşik değerinden büyük ortak bilgi değerine sahip gen çiftleri ile GBİEA oluşturulur. Denklem 1'e göre X ve Y değişkenlerinin tüm değerleri göz önüne alınarak X ve Y'nin birleşik olasılık (p(x,y)) değeri, X ve Y'nin olasılık değerlerini (p(x) ve p(y)) kullanılarak ortak bilgi değeri çıkartılmaktadır. Ortak bilgi değeri bir rassal değişkenin başka bir rassal değişken ile değerinin belirlenebilmesi durumunu göstermektedir ve değişkenler arası doğrusal olmayan ilişkileri de belirleyebilmektedir.

$$\mathrm{MI}(\mathrm{X},\mathrm{Y}) = \sum_{\mathrm{x}\in\mathcal{X}} \sum_{\mathrm{y}\in\mathcal{Y}} \mathrm{p}(\mathrm{x},\mathrm{y}) \log \frac{\mathrm{p}(\mathrm{x},\mathrm{y})}{\mathrm{p}(\mathrm{x})\mathrm{p}(\mathrm{y})} \quad (1)$$

Conservative Causal Core Networks(C3NET): İki adımdan oluşmakla birlikte ilk adımı RelNet yaklaşımını içermektedir. Genler arasındaki ilişkiler ortak bilgi değeriyle hesaplandıktan sonra ortak bilgi değeri için eşik değeri belirlenmesi için permütasyon test uygulanır. Belirlenen eşik değerinin altında ortak bilgi değerine sahip genler arasındaki ilişkiler GBİEA oluşumunda kullanılmaz.

İkinci adımda bir genin sadece en yüksek ortak bilgi değerine sahip gen ile ilişkisinin belirlenip diğer genlerle olan ilişkilerinin elenerek gen ağı oluşturulmaktadır. Böylece ikinci adımda en yüksek ortak bilgi değerine göre eleme işlemi gerçekleştirilmiş olur. C3NET algoritması için c3net R paketi [23] kullanılarak GBİEA çıkarımı gerçeklenmiştir. C3NET algoritmik karmaşıklık açısından basit ve başarılı bir algoritma olduğu için çalışmamızda kullanılmıştır.

Accurate Cellular Networks (ARACNE): C3NET gibi ilk adımı RelNet yaklaşımından oluşmaktadır. Permütasyon testi ile genler arası ortak bilgi değeri için bir eşik değeri hesaplanır. İkinci adımda ise dolaylı ilişkilerin elenmesiyle geride kalan gen- gen ilişkilerinin seçilmiyle gen ağı oluşturulur.

$$\mathrm{MI}(\mathrm{X},\mathrm{Z}) \leq \operatorname{argmin}\{\mathrm{MI}(\mathrm{X},\mathrm{Y}), \mathrm{MI}(\mathrm{Y},\mathrm{Z})\} \quad (2)$$

Denklem 2'ye göre X ve Z genleri arasındaki ilişkinin iki genin de Y geniyle olan ilişkisinden dolaylı olarak ortaya çıkıp çıkmadığı kontrol edilmiştir. X ile Y veya Y ile Z genleri arasındaki ortak ilişki değerlerinden en küçük olanı X ile Z genleri arasındaki ortak ilişki değerinden büyükse X ile Z arasındaki ilişki dolaylı ilişki olarak belirlenir ve ağ çıkarımında kullanılmaz. ARACNE dolaylı ilişkisi olan gen ilişkilerini eleyerek daha kuvvetli ilişkilerle bir GBİEA elde etmeye çalışmaktadır. Literatürde yapılan çalışmalarda ARACNE ile elde edilen GBİEA'ların farklı değerlendirme tekniklerine göre başarılı olduğu gözlemlenmiştir [6,24]. ARACNE algoritması ikinci adım olarak dolaylı ilişkileri elemiştir. ARACNE algoritması için minet R paketinden yararlanılmıştır. [25,26]

Weighted Gene Co-expression Network Analysis (WGCNA): WGCNA bundan önce anlatılan üç gen ağı çıkarım yaklaşımından farklı olarak ortak bilgi değeri yerine genler arasındaki ilişki skorunu Denklem 3'te gösterildiği gibi Pearson ilişki katsayısı formülüyle hesaplamaktadır. Denklemde x ve y genlerinin farklı örneklerdeki değerleri, ortalamaları ve bu değerlerin kareleri kullanılarak Pearson ilişki katsayısı hesaplanmaktadır.

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)((\sum y_i^2 - n\bar{y}^2))}} \quad (3)$$

WGCNA GBİEA çıkarımında ölçeksiz ağ (scale free network) tanımına uygun ağlar belirlemek için ilişki skoru eşik değeri belirler. Ölçeksiz ağ tanımına uyması için Pearson ilişki katsayılarının belirli sayıda üssü alınarak GBİEA oluşturulmaya çalışılır. WGCNA R paketi kapsamlı bir paket olmakla birlikte GBİEA çıkarımının dışında Pearson ilişki değerleri kullanılarak birbiriyle ilişki modüllerin bulunması ve bu modüllerin biyolojik süreçlerle incelenmesi gibi farklı çalışmalarda kullanılmaktadır [6]. WGCNA algoritması için aynı isimli R paketi [27] çalışmamızda kullanılmıştır.

C. NORMALİZASYON YÖNTEMLERİ

RNA-Seq verileri dizileme sonucunda genlerle eşleşen okuma sayılarını göstermektedir. Burada okuma sayılarını etkileyen iki önemli faktör bulunmaktadır. Bunlardan ilki RNA- Seq verilerindeki toplam okuma sayısının örnekten örneğe değişebilmesidir. İkinci önemli faktör ise uzun nükleotit diziliminden oluşan genlere ait okuma sayılarının kısa nükleotit diziliminden oluşan genlere ait okuma sayılarından fazla olabilmesidir.

Çalışmamızda ham okuma verileri, M- değerlerinin kırpılmış ortalaması (MDKO)ve Göreceli Log İfadesi (GLİ) normalizasyon yöntemleri ayrı ayrı kullanılarak elde edilmiş GBİEA'ların performansları karşılaştırılmıştır.

C. 1. M- Değerinin Kırpılmış Ortalaması (MDKO) Normalizasyonu

RNA- Seq verilerinde bir örnekteki toplam okuma sayısı, okuma sayısı fazla olan az sayıda gene ait okuma sayısından oldukça etkilenmektedir. Bu nedenle normalizasyon gerçekleştirilmeyen durumda örnekler arasında farklı ifade edilen genlerin bulunması ve genler arasındaki ilişkilerin çıkarılması problematiktir. MDKO, okuma sayısı fazla olan az sayıdaki genler ile ve az okuma sayısı olan genlerin toplam okuma sayına etkisini normalize etme varsayımına dayanmaktadır. MDKO, Log-Örnek Değişikliği (M-değeri) ve Mutlak İfade Seviyesi (A- değeri) gibi iki metriği kullanarak normalizasyon gerçekleştirmektedir. Bir genin M-değeri bir örneğin referans örneğe göre okuma sayısındaki değişimini gösterirken, A-değeri ise genin örnek ve referans değerdeki ortalamasını belirtmektedir. Referans örnek belirlenirken tüm örneklerdeki toplam okuma sayıları küçükten büyüğe sıralandıktan sonra kutu grafiğinde üçüncü çeyrek (Q3) değere karşılık gelen okuma değeri referans örnek değeri olarak belirlenir.

$$M_g(j,r) = \log_2\left(\frac{K_{gj}}{D_j}\right) - \log_2\left(\frac{K_{gr}}{D_r}\right)$$
$$A_g(j,r) = 0.5 * (\log_2\left(\frac{K_{gj}}{D_j}\right) + \log_2\left(\frac{K_{gr}}{D_r}\right)) \quad (4)$$

G genine ait M ve A değerleri Denklem 4'de gösterilmiştir. K_{gj} değeri g geninin j örneğindeki okuma sayısını, K_{gr} değeri g geninin referans örnekteki okuma sayısını, D_j j örneğindeki toplam okuma sayısını, D_r referans örnekteki toplam okuma sayısını göstermektedir. Genlerin okuma değerleri kullanıcı tarafından belirlenen M ve A değerlerinden düşük veya yüksek ise genlerin okuma değerleri sıfır olarak atanır. Örneğin M değeri %10 olarak belirlendiği taktirde tüm genlerin M değerleri belirlendikten sonra bu değerin en yüksek %10 ve en düşük %10'luk diliminden M değerine sahip olan genler veri kümesinden çıkartılır ve GBİEA oluşumunda kullanılmamaktadır.

$$w_g(j,r) = \left(\frac{D_j - K_{gj}}{D_j K_{gj}} + \frac{D_r - D_{gr}}{D_r D_{gr}}\right)$$
$$f_j = \frac{\sum_{g\,\varepsilon\,G} w_g(j,r).M_g(j,r)}{w_g(j,r)} \quad (5)$$

Denklem 5'te g tüm genleri, $M_g(j,r)$ değeri g geninin j örneğindeki M- değerini, $w_g(j,r)$ değeri g genin j örneği ve r referans değerlerine göre aldığı yeni normalizasyon katsayı değerlerini göstermektedir. Bu hesaplama g geninde olduğu gibi tüm genler için hesaplandıktan sonra f_j düzeltme değeri elde edilir ve bu değer j örneğindeki tüm genlerin okuma sayılarına uygulanarak MDKO normalizasyonu gerçekleştirilmiş olur. Çalışmamızda MDKO normalizasyonu için edgeR R paketi [28] sabit tanımlı M ve A değeri parametreleriyle birlikte kullanılmıştır. M değeri için sabit parametre değeri %30, A değeri için ise %5 olarak belirlenmiştir. Buna göre M değeri %30'luk üst ve alt sınarda olan genler ile A değeri %5'in üstünde ve altında olan genler normalizasyonda kullanılmamaktadır.

C. 2. Göreceli Log İfadesi (GLİ) Normalizasyonu

Genlerin okuma sayıları farklı örneklerdeki ifade değerine göre değişkenlik gösterebilmektedir. Farklı fenotiplere veya zamana göre belirleyici olmayan genlerin okuma sayısı örneklere ait düzeltme değerleriyle orantılıdır. GLİ okuma sayılarını, ifade seviyesi ve sıralama derinliği ile orantılı olduğu varsayımına dayanmaktadır. GLİ genlerin okuma değerlerinin aynı genlerin tüm örneklerdeki okuma değerlerinin geometrik ortalamasına oranını alarak normalizasyon işlemi gerçekleştirir. Örneklerin düzeltme faktör değeri (C_j), j örneği için düzeltme değeri, hesaplanırken örnekte geçen tüm genlerin okuma sayısı aynı genlerin tüm örneklerdeki geometrik ortalamasına bölünmesiyle elde edilen değerlerin medyan değeri olarak elde edilmektedir. Buradaki yaklaşımla farklı durumlarda sıklık değeri istatistiksel olarak değişmeyen genlerin tüm örneklerdeki gen okuma sayıları birbirine benzeyecek ve denklem 6'da ilgili genler için sonuç sıfıra yakınsayacaktır. Denklem 6'da j örneğinde örnekteki tüm genlere uygulanacak düzeltme katsayı değerinin (örneğe ait normalizasyon katsayı değeri) hesaplanması gösterilmiştir.

$$C_j = \text{medyan}_i\left(\frac{k_{ij}}{(\pi_{v=1}^{m} k_{iv})^{1/m}}\right) \quad (6)$$

Burada k_{ij}, j örneğindeki i genine ait okuma sayısını, m değeri RNA- Seq verisindeki tüm örnek sayısını göstermektedir. k_{iv}, tüm örneklerdeki i geninin okuma sayısını göstermektedir. Bir örneğin düzeltme değeri olan düzeltme faktör değeri (C_j), j örneğindeki tüm genler için okuma değerinin aynı genlerin tüm örneklerde yer alan okuma değerinin geometrik ortalamasına olan oranı olarak alınır. Bu oran tüm genler için çıkarıldıktan sonra tüm genlerin için elde edilen değerin medyan değeri alınarak j örneğinin düzeltme faktör değeri bulunur ve bu değer ile j örneğindeki genlerin okuma değeri çarpılır, GLİ normalizasyonu j örneğindeki tüm genler için tamamlanmış olur. GLİ normalizasyonunu

uygulamak için DESeq2 R paketi [29] sabit parametreleriyle birlikte çalışmamız kapsamında kullanılmıştır.

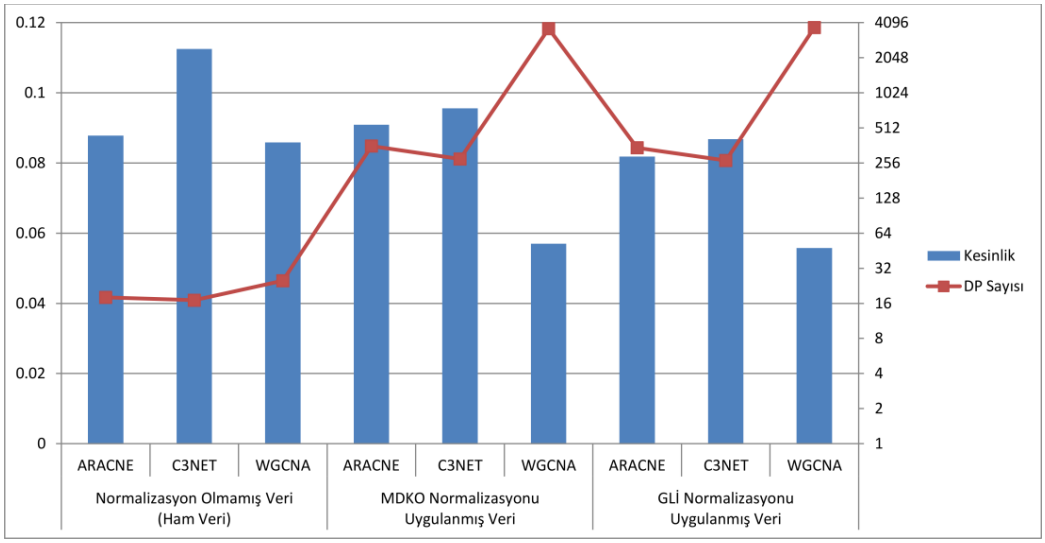
III. DENEY SONUÇLARI

Elde edilen GBİE ağlarının performansı ilk olarak örtüşme analizi ile değerlendirilmiştir. GAnet R paketindeki [30] 1,599,366 protein- protein (gen- gen) ilişkisini içeren literatür verisi örtüşme analizinde doğrulama veri kümesi olarak kullanılmıştır. GBİE ağında ve literatür verisinde yer alan ilişkiler doğru pozitif (DP), GBİE ağında yer alan literatür verisinde yer almayan gen- gen ilişkileri yanlış pozitif (YP) olarak örtüşme analizinde kullanılmıştır. Literatür verisinde yer alan GBİE ağında yer almayan ilişkiler ise yanlış negatif (YN) olarak örtüşme analizinde kullanılmaktadır. Hem GBİE ağlarında hem de literatüre verisinde olmayan ilişkiler ise doğru negatif (DN) olarak gösterilmektedir. Çalışmamızda doğrulama veri kümesinin 1.5 milyonun üzerinde ilişki içermesi nedeniyle YN değeri yüksek olacağı için örtüşme analizinde DP ve kesinlik değerleri performans metrikleri olarak kullanılmıştır. Kesinlik DP sayısının (DP+YP) sayısına bölümü olarak elde edilmektedir.

Gen ağı çıkarım algoritmaları ham veri, MDKO ve GLİ normalizasyonu uygulanmış verilere uygulanarak GBİE ağları üç farklı veri kümesi için de elde edilmiştir. GBİE ağlarının örtüşme analizindeki performans değerleri Şekil 2'de gösterilmiştir. Örtüşme analizinin istatistiksel olarak anlamlı olup olmadığı Fisher Kesinlik Testi kullanılarak gerçeklenmiştir.

$$p = \frac{\binom{DP+YN}{DP}\binom{YP+DN}{YP}}{\binom{DP+YN+YP+DN}{DP+YP}} \tag{7}$$

Denklem 7'de Fisher Kesinlik Testi'nde elde edilen sonuçların istatistiksel olarak anlamlılığının hesaplanmasında kullanılan p değerinin çıkarımı gösterilmiştir. Çalışmamızda Fisher Kesinlik Testi için GAnet R paketi kullanılmıştır [30]. Eşik değeri olan p değeri 0.05 seçilerek bu değerin üzerinde bir değer elde edildiğinde GBİE ağına ait ilişkiler rastlantısal olarak elde edildiği kabul edilmiş ve performans analizinde değerlendirilmemiştir. Çalışmamızdaki tüm GBİE ağlarına ait en yüksek p değeri ham veri üzerinde ARACNE algoritmasıyla 6.42e-17 olarak elde edilmiştir. Diğer ağlara ait p değerleri ise bu değerden daha küçüktür. Bu nedenle çalışmamızda elde edilen ve örtüşme analizine göre değerlendirilen GBİE ağlarının sonuçları Fisher Kesinlik Testine göre istatistiksel olarak anlamlıdır.



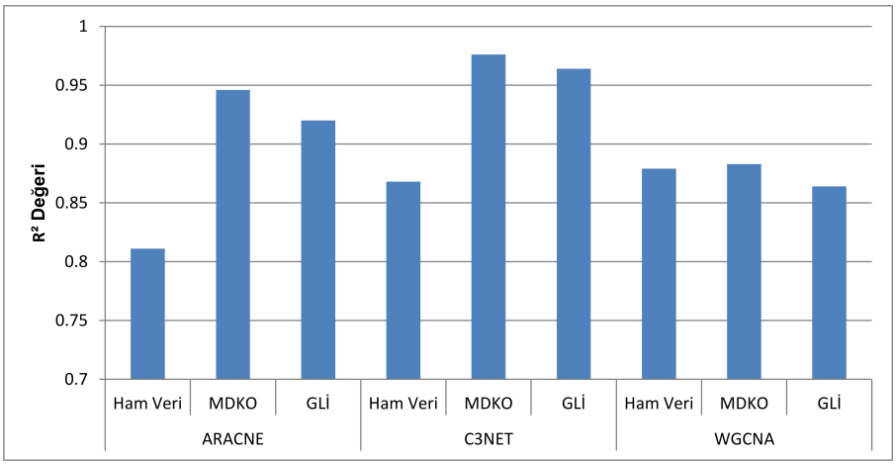
Şekil 2. *Örtüşme analizi sonuçları*

Şekil 2'de gösterilen örtüşme analizi değerlerine göre en yüksek kesinlik değeri C3NET algoritmasıyla ham veriler üzerinden elde edilen GBİEA'dan elde edilmiştir. Ham veriler üzerinden elde edilen gen GBİEA'ların kesinlik değerleri normalizasyon verileri üzerinde elde edilen GBİE'ların kesinlik değerlerinden daha yüksek çıkmıştır. Bununla birlikte MDKO ve GLİ normalizasyonu uygulandıktan sonra elde edilen veriler kullanılarak elde edilmiş GBİEA'ların DP değerleri ham veri kullanılarak elde edilen GBİEA'ların DP değerlerinden daha yüksek çıkmıştır. Ham verilerden elde edilen GBİEA'lara ait DP değerleri 16-30 aralığında iken bu değerler MDKO ve GLİ normalizasyonlarıyla elde edilen veriler üzerinde özellikle WGCNA algoritmasıyla elde edilen GBİEA'lar için 4000'nin üzerinde çıktığı gözlemlenmiştir. Ham ve 2 farklı yöntemle normalize edilmiş veriler üzerinde en yüksek kesinlik değerleri sırasıyla C3NET, ARACNE ve WGCNA gen ağı çıkarım algoritmalarıyla elde edilmiştir. WGCNA algoritması ham ve normalize verilerde en yüksek DP değerlerini veren GBİEA'ları çıkarmasına rağmen aynı zamanda tüm veri kümelerinde en düşük kesinlik değerine sahip GBİEA'ların da çıkarımını sağlamıştır. Şekil 2'de elde edilen sonuçlara göre ARACNE algoritması ham ve normalizasyon uygulanmış veri üzerinde benzer kesinlik değerleri elde ederek kesinlik sonucu değerlenmesine göre normalizasyondan en az etkilenen GBİEA çıkarım algoritması olmuştur. Örtüşme analizinde normalizasyon yöntemlerini karşılaştırdığımızda MDKO normalizasyonuyla elde edilen GBİEA'ların kesinlik değeri GLİ normalizasonu uygulanan veriler üzerinde elde edilmiş GBİEA'ların kesinlik değerinden yüksek çıkmıştır. Her iki normalizasyon yöntemiyle oluşturulmuş GBİEA'lara ait DP değerleri birbirlerine yakın olduğu gözlemlenmiştir. Çalışmamızda elde edilen kesinlik ve DP değerleri literatürde yer alan gen ağı çıkarım çalışmalarıyla örtüşmektedir [9,10,31].

Farklı normalizyon yaklaşımlarının GBİEA çıkarımına etkisini incelemek için kullandığımız ikinci yaklaşım topolojik değerlendirmedir. GBİE ağlarında yer alan düğümlerin (genler) komşuluk değerini gösteren derece değerlerinin dağılımı kuvvet yasası dağılımıdır. Kuvvet yasası dağılımına göre az sayıda genin derece değeri fazlayken geri kalan çoğu genin derece değeri düşüktür ve elde edilen bu ağlar ölçeksiz ağlardır.

$$P(\text{derece}) = \beta \, x \, \text{derece}^{-\alpha}$$
$$\log P(\text{derece}) = \log \beta - \alpha \, x \log \text{derece} \quad (8)$$

Denklem 8'de P olasılık fonksiyonu, β olasılık fonksiyonun ölçekleme katsayısı, α ise üssel parametre olup GBİE ağlarındaki genlerin düğüm derecelerinin kuvvet yasası dağılımına uyması için iki ile üç arasında bir sabit olarak alınmaktadır [32]. Denklem 8'in alt satırındaki eşitlikte denklemin her iki tarafının da doğal logaritması alınarak denklem üssel değerden kurtarılmış ve doğrusal regresyon ile düğümlerin derece değerleri modellenebilir hale gelmiştir. Burada düğümlerin derece değerlerinin olasılık değerleri bağımlı değişken, düğüm derece değerleri ise bağımsız değişken olarak alınarak en küçük kareler yöntemini kullanan regresyon modeliyle modellenmiştir. Regresyon ile elde edilen doğrusal modelin başarısı uygunluk iyiliği belirleme katsayısı (coefficient of determination), R^2, değeri ile belirlenmiştir. GBİE ağlarının genlerine ait derece değerlerinin doğrusal modellenmesi elde edildikten sonra modelin başarısı R^2 değerine göre değerlendirilmiştir. R^2 değeri bire yakınsa GBİEA ölçeksiz ağ tanımına uygun bir ağ olmakla birlikte bu değer sıfır değerine yakınsa GBİEA rassal ağ özelliği taşımaktadır. Biyolojik ağlarda R^2 değerinin bire yakın olması beklenmektedir. Şekil 3'te ham ve farklı normalizasyon yöntemleri ile oluşturulmuş RNA- Seq veri kümesi üzerinde elde edilen GBİE ağlarının R^2 değeri verilmiştir.



Şekil 3. *Topolojik değerlendirme sonuçları*

Şekil 3'teki sonuçlara göre RNA- Seq verilerinin normalizasyonu ile ARACNE ve C3NET tarafından elde edilen GBİE ağları ölçeksiz ağ tanımına daha uygun hale gelmiştir. WGCNA ölçeksiz ağ tanımına uygun GBİE ağları çıkartması nedeniyle normalizasyonun topolojik değerlendirmede etkisi daha kısıtlı olmuştur. MDKO normalizasyonuyla elde edilen GBİE ağlarının R^2 değerleri ham veri ve GLİ normalizasyonu uygulanan verilerden daha yüksek çıkmıştır. Şekil 3'e göre C3NET tabanlı GBİE ağlarının ölçeksiz ağ tanımına daha uygun olduğu gözlemlenmiştir.

IV. TARTIŞMA ve ÖNERİLER

Dizileme teknolojilerindeki ilerlemelerle birlikte farklı fenotipe neden olan moleküler seviyedeki ilişkilerin belirlenmesi daha doğruluklu olarak belirlenmeye başlamıştır. Moleküler ilişkilerin çıkarımında uzun süredir kullanılan mRNA mikrodizin verileriyle birlikte son yıllarda RNA- Seq verileri gen ağlarının çıkarımında sıklıkla kullanılmaktadır. RNA- Seq verileri farklı nükleotit sayılarına sahip genler ve veri kümesindeki örneklerin farklı sayıda okumalarından oluşması nedeniyle normalizasyon işlemleri uygulanarak kullanılmaktadır. RNA-Seq verileri üzeride normalizasyonun etkisi önemli olduğu farklı ifade edilen genlerin çıkarımıyla ilgili çalışmalarda anlaşılmıştır [4,5]. Farklı ifade edilen genlerle ilgili yapılan çalışmalarda [13,14] M- değerlerinin kırpılmış ortalaması (MDKO) ve Göreceli log ifadesi (GLİ) normalizasyon yöntemleri kullanılarak elde edilmiş farklı ifade edilen genlerin değerlendirme aşamasında total okuma sayısı normalizasyonu, üst çeyrek normalizasyonu, medyan normalizasyonu, örtüşen milyon adet okuma sayısı normalizasyonu ve ham veriler kullanılarak elde edilmiş farklı ifade edilen genlerin performansından daha iyi olduğu sonucu elde edilmiştir.

Literatürde normalizasyon yöntemlerinin karşılaştırılması farklı ifade edilen genlerin çıkarımında kullanılmakla birlikte GBİEA'ların çıkarımıyla ilgili çalışma sayısı çok kısıtlıdır. Çalışmamızda ham veri ve farklı iki normalizasyon yaklaşımlarıyla elde edilmiş GBİEA'ların performansları örtüşme analizi ve topolojik değerlendirme ile karşılaştırılmıştır.

Örtüşme analizine göre her ne kadar en yüksek kesinlik değerleri ham veri kullanarak elde edilmiş GBİEA'lar üzerinde elde edilmiş olsa da bu ağlara ait DP değerleri normalize veriler kullanılarak elde edilen GBİEA'ların DP değerinden oldukça düşük olduğu gözlemlenmiştir. Bu sonuç bize normalize edilmiş veri üzerinde GBİEA çıkarım algoritmalarının çok daha fazla sayıda gen- gen ilişkisi belirleyebildiği sonucunu vermektedir. Elde edilen sonuçlara göre normalizasyonun GBİEA çıkarımına etkisi ile ilgili çalışma olmadığı için sonuçlarımızla karşılaştırma fırsatımız olmamıştır. Örtüşme analizinde GBİEA çıkarım algoritmaları karşılaştırmasında ise C3NET algoritmasıyla elde edilen GBİEA'ların kesinlik değeri üç farklı veri üzerinde de diğer gen ağı çıkarım algoritmalarından yüksek çıkmıştır. WGCNA algoritmasıyla elde edilen GBİEA'ların ise tüm veriler üzerinde DP

değerlerinin en yüksek DP değerleri olduğu gözlemlenmiştir. MDKO ve GLİ normalizasyon yöntemlerinin performansları karşılaştırıldığında kesinlik değerlerinde MDKO yaklaşımıyla elde edilen GBİEA'ların az da olsa yüksek olduğu gözlemlenmiştir. Elde ettiğimiz sonuçlar farklı ifade edilen genlerle ilgili yapılan çalışmalardaki [11-21] normalizasyonun performansa olumlu etkisiyle paralellik göstermektedir.

Normalizasyonun GBİE ağlarının topolojik yapısına etkisi de çalışmamız kapsamında incelenmiştir. GBİE ağlarında yer alan düğümlerin komşuluk değeri olan derece değerleri kuvvet yasası dağılımına göre dağılması ve buna göre ağların ölçeksiz ağ tanımına uygun olması beklenmektedir. Çalışmamızda düğümlerin derece olasılık değerleri regresyon ile doğrusal olarak modellenmiş ve modellin başarısı R^2 değeriyle ölçülmüştür. R^2 değeri yüksek olan, bire yakın, GBİE ağları ölçeksiz tanımına uygundur. Elde ettiğimiz sonuçlara göre normalizasyon uygulanarak elde edilmiş GBİE ağları ölçeksiz ağ tanımına ham veri ile elde edilen GBİE ağlarına göre daha uygun olarak belirlenmiştir. En başarılı R^2 değerleri MDKO normalizasyonu yöntemiyle elde edilmiştir ve topolojik değerlendirmede en başarılı gen ağı çıkarım algoritmaları C3NET olarak belirlenmiştir.

Çalışmamızın sonucunda ham veri, MDKO ve GLİ normalizasyonu uygulanmış veriler örtüşme analizi ve topolojik değerlendirmelere göre karşılaştırılmıştır. Elde edilen sonuçlara göre normalize edilmiş veriler üzerinde gen çıkarım ağlarının örtüşme analizinde GBİE ağlarının DP değerlerinde yükselme sağladığı gözlemlenmiştir. Normalize edilmiş veriler üzerinde elde edilen GBİE ağları aynı zamanda ölçeksiz ağ tanımına daha uygun olarak elde edilmiştir. Bundan sonraki çalışmamızda normalizasyonun GBİE ağlarındaki biyolojik süreçleri belirlemedeki etkisinin incelenmesi hedeflenmektedir.

V. KAYNAKLAR

[1] A. Korotkov, J. D. Mills, J.A. Gorter, E.A. Van Vliet, E. Aronica, "Systematic review and meta-analysis of differentially expressed miRNAs in experimental and human temporal lobe epilepsy," *Scientific Reports*, vol. 7, no. 1, pp. 1-13, 2017.

[2] S. M. Salleh, G. Mazzoni, P. Løvendahl, H.N. Kadarmideen, "Gene co-expression networks from RNA sequencing of dairy cattle identifies genes and pathways affecting feed efficiency," *BMC Bioinformatics*, vol. 19, no. 1, pp. 513, 2018.

[3] Y. Hu et al., "Improving the diversity of captured full-length isoforms using a normalized single-molecule RNA-sequencing method," *Communications Biology*, vol. 3, no. 1, pp. 1-15, 2020.

[4] F. Ozsolak, P. M. Milos, "RNA sequencing: advances, challenges and opportunities," *Nature Reviews Genetics*, vol. 12, no. 2, pp. 87-98, 2011.

[5] M. Garber, M. G. Grabherr, M. Guttman and C. Trapnell, "Computational methods for transcriptome annotation and quantification using RNA-seq," *Nature Methods*, vol. 8, no. 6, pp. 469, 2011.

[6] M. Ö. Cingiz and B. Diri, "Two-tier combinatorial structure to integrate various gene co-expression networks of prostate cancer," *Gene*, vol. 721, pp. 144102, 2019.

[7] C. F. Xu, C. H. Yu, Y. M. Li, "Regulation of hepatic microRNA expression in response to ischemic preconditioning following ischemia/reperfusion injury in mice," *OMICS A Journal of Integrative Biology*, vol. 13, no. 8, pp. 513-520, 2009.

[8] S. Ballouz, W. Verleyen and J. Gillis, "Guidance for RNA-seq co-expression network construction and analysis: safety in numbers," *Bioinformatics*, vol. 31, no. 13, pp. 2123-2130, 2015.

[9] R. de Matos Simoes, S. Dalleau, K. E. Williamson and F. Emmert-Streib, "Urothelial cancer gene regulatory networks inferred from large-scale RNAseq, Bead and Oligo gene expression data," *BMC Systems Biology*, vol. 9, no. 1, pp. 21, 2015.

[10] O. D. Iancu, S. Kawane, D. Bottomly, R. Searles, R. Hitzemann and S. McWeeney, "Utilizing RNA-Seq data for de novo coexpression network inference," *Bioinformatics*, vol. 28, no. 12, pp. 1592-1597, 2012.

[11] F. Abbas-Aghababazadeh, Q. Li, B. L. Fridley, "Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing," *PloS One*, vol. 13, no. 10, pp. e0206312, 2018.

[12] M. Smid et al., "Gene length corrected trimmed mean of M-values (GeTMM) processing of RNA-seq data performs similarly in intersample analyses while improving intrasample comparisons," *BMC Bioinformatics*, vol. 19, no. 1, pp. 1-13, 2018.

[13] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data," *Genome Biology*, vol. 11, no. 3, pp. R25, 2010.

[14] Y. Lin et al., "Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual Drosophila melanogaster," *BMC Genomics*, vol. 17, no. 1, pp. 1-20, 2016.

[15] M. A. Dillies et al., "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis," *Briefings in Bioinformatics*, vol. 14, no. 6, pp. 671-683, 2013.

[16] S. Mandelboum, Z. Manber, O. Elroy-Stein, R. Elkon, "Recurrent functional misinterpretation of RNA-seq data caused by sample-specific gene length bias," *PLoS Biology*, vol. 17, no. 11, pp. e3000481, 2019.

[17] F. Seyednasrollah, A. Laiho, L. L. Elo, "Comparison of software packages for detecting differential expression in RNA-seq studies," *Briefings in Bioinformatics*, vol. 16, no. 1, pp. 59-70, 2015.

[18] D. Risso, J. Ngai, T. P. Speed, S. Dudoit, "Normalization of RNA-seq data using factor analysis of control genes or samples," *Nature Biotechnology*, vol. 32, no. 9, pp. 896-902, 2014.

[19] B. Bhat, M. Yaseen, A. Singh, S. M. Ahmad, N. A. Ganai, "Identification of potential key genes and pathways associated with the Pashmina fiber initiation using RNA-Seq and integrated bioinformatics analysis," *Scientific Reports*, vol. 11, no. 1, pp. 1-9, 2021.

[20] J. J. Velazquez et. al., "Gene regulatory network analysis and engineering directs development and vascularization of multilineage human liver organoids," *Cell Systems*, vol. 12, no. 1, pp. 41-55, 2020.

[21] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger and K. Ellrott, "The cancer genome atlas pan-cancer analysis project," *Nature Genetics*, vol. 45, no. 10, pp. 1113, 2013.

[22] P. E. Meyer, K. Kontos, F. Lafitte and G. Bontempi, "Information-theoretic inference of large transcriptional regulatory networks," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2007, pp. 1-9, 2007.

[23] G. Altay and F. Emmert-Streib, “Inferring the conservative causal core of gene regulatory networks,” *BMC Systems Biology*, vol. 4, no. 1, pp. 1-13, 2010.

[24] M. Ö. Cingiz, B. Diri, “Topological and biological assessment of gene networks using miRNA-target gene data,”in *Innovations in Intelligent Systems and Applications Conference*, Turkey, 2019, pp. 1-4.

[25] A. A. Margolin et al., “ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context,” *BMC Bioinformatics*, vol. 7, no. 1, pp. S7, 2006.

[26] P. E. Meyer, F. Lafitte and G. Bontempi, “minet: AR/Bioconductor package for inferring large transcriptional networks using mutual information,” *BMC Bioinformatics*, vol. 9, no. 1, pp. 461, 2008.

[27] P. Langfelder and S. Horvath, “WGCNA: an R package for weighted correlation network analysis,” *BMC Bioinformatics*, vol. 9, no. 1, pp. 559, 2008.

[28] M. D. Robinson, D. J. McCarthy and G. K. Smyth, “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data,” *Bioinformatics*, vol. 26, no. 1, pp. 139-140, 2010.

[29] M. I. Love, W. Huber and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2,” *Genome Biology*, vol. 15, no. 12, pp. 550, 2014.

[30] G. Altay, N. Altay and D. Neal, “Global assessment of network inference algorithms based on available literature of gene/protein interactions,” *Turkish Journal of Biology*, vol. 37, no. 5, pp. 547-555, 2013.

[31] M. Ö. Cingiz, G. Biricik and B. Diri, “ARNetMiT R Package: association rules based gene co-expression networks of miRNA targets,” *Cellular and Molecular Biology (Noisy-le-grand)*, vol. 63, no. 3, pp. 18-25, 2017.

[32] F. Chung, L. Lu, “Connected components in random graphs with given expected degree sequences,” *Annals of Combinatorics*, vol. 6, no.2, pp. 125-145, 2002.