

T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YENİ NESİL DİZİLEME YÖNTEMİ İLE TÜRK FINDIĞI (*CORYLUS*
COLURNA L.) VE TOMBUL FINDIĞIN (*CORYLUS AVELLANA* L.)
TRANSKRİPTOM ANALİZLERİNİN YAPILMASI VE
KARŞILAŞTIRILMASI

Salih ULU

DOKTORA TEZİ

Moleküler Biyoloji ve Genetik Anabilim Dalı

Moleküler Biyoloji ve Genetik Programı

Danışman

Prof. Dr. Nehir ÖZDEMİR ÖZGENTÜRK

Temmuz, 2022

T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

**YENİ NESİL DİZİLEME YÖNTEMİ İLE TÜRK FINDIĞI (*CORYLUS*
COLURNA L.) VE TOMBUL FINDIĞIN (*CORYLUS AVELLANA* L.)
TRANSKRİPTOM ANALİZLERİNİN YAPILMASI VE
KARŞILAŞTIRILMASI**

Salih ULU tarafından hazırlanan tez çalışması 20.07.2022 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Moleküler Biyoloji ve Genetik Anabilim Dalı, Moleküler Biyoloji ve Genetik Programı **DOKTORA TEZİ** olarak kabul edilmiştir.

Prof. Dr. Nehir ÖZDEMİR ÖZGENTÜRK
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

Prof. Dr. Nehir ÖZDEMİR ÖZGENTÜRK, Danışman

Yıldız Teknik Üniversitesi

Doç. Dr. Emel ORDU, Üye

Yıldız Teknik Üniversitesi

Prof. Dr. Cemal ÜN, Üye

Ege Üniversitesi

Prof. Dr. Semiha ERİŞEN, Üye

Yıldız Teknik Üniversitesi

Prof. Dr. Günay ANACAK, Üye

Ege Üniversitesi

Danışmanım Prof. Dr. Nehir ÖZDEMİR ÖZGENTÜRK sorumluluğunda tarafımca hazırlanan Yeni Nesil Dizileme Yöntemi ile Türk Fındığı (*Corylus colurna* L.) ve Tombul Fındığın (*Corylus avellana* L.) Transkriptom Analizlerinin Yapılması ve Karşılaştırılması başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Salih ULU

İmza



Bu çalışma, Yıldız Teknik Üniversitesi Bilimsel Araştırma Proje Koordinatörlüğü'nün FBA-2017-3134 numaralı projesi ile desteklenmiştir.



Sevgili Eşim Zehra

ve

Çevresine Işık Saçan Oğlum Melih'e

TEŞEKKÜR

Hayat boyu öğrenmenin eşsiz bir kanıtını sunan bu doktora tezinin ortaya çıkmasında sunduğu sonsuz destek, doktora öğrenimim boyunca bana karşı gösterdiği sabır ve anlayış, tez çalışmam esnasında karşılaştığım her türlü problemde akılcı yönlendirmeleriyle önümde ışık olan, öğrencisi olmaktan gurur duyduğum değerli hocam, danışmanım Prof Dr. Nehir ÖZDEMİR ÖZGENTÜRK'e teşekkürlerimi sunarım.

Tez çalışmama başlamamla birlikte Tez İzleme Komitesinde yer alan ve pozitif bakışını her toplantıda bana hissettiren, katıldığı son toplantıda dahi ciddi rahatsızlığına rağmen bilimin ışığının sönmesine müsaade etmeyen hocam Prof. Dr. Kemal Özdem ÖZTABAK'a sonsuz saygılarımı sunarım.

Hep yanımda olduğunu bildiğim annem Hanım ULU ve çok erken ayrılığıyla fiziken yanımda olamasa da her zaman varlığını hissettiğim babam Hüseyin ULU'ya teşekkürlerimi sunarım.

Bitmek bilmeyen enerjileriyle, farklı bakış açılarıyla, beni de harekete geçiren motivasyonlarıyla hayatımın en güzel zamanlarına imza atmış olan iki temel dinamiğim; ailemizin neşesi biricik oğlumuz Melih ULU ve her koşulda bana inancı eksilmeyen, beni daima destekleyen ve hayatıma yön veren sevgili eşim Zehra ÖMEROĞLU ULU'ya teşekkürü bir borç bilirim.

Salih ULU

İÇİNDEKİLER

SİMGE LİSTESİ	viii
KISALTMA LİSTESİ	ix
ŞEKİL LİSTESİ	xi
TABLO LİSTESİ	xiii
ÖZET	xiv
ABSTRACT	xvi
1 GİRİŞ	1
1.1 Literatür Özeti.....	1
1.2 Tezin Amacı	4
1.3 Hipotez	4
2 GENEL BİLGİLER	5
2.1 Fındık.....	5
2.1.1 <i>Corylus avellana</i> L. (Avrupa Fındığı)	6
2.1.1.1 Tombul.....	9
2.1.2 <i>Corylus colurna</i> L. (Türk Fındığı)	10
2.2 Ribonükleik Asit (RNA)	12
2.2.1 RNA Dizileme (RNA-seq)	12
2.2.1.1 Derleme Öncesi Kalite Kontrol ve Filtreleme	14
2.2.1.2 <i>De novo</i> Transkriptom Derlemesi	15
2.2.1.3 Derleme Sonrası Kalite Kontrol	17
2.2.1.4 Hizalama ve Bolluk Tahmini	18
2.2.1.5 Derlemenin Küçültülmesi ve Gereksiz Tekrarların Azaltılması	19
2.2.1.6 Diferansiyel İfade Analizi	20
2.2.1.7 Dizilerin Translasyonu.....	21
2.2.1.8 Transkriptoma İşlevsel Açıklama Eklenmesi	21
3 MATERYAL VE YÖNTEM	26
3.1 Materyal.....	26
3.2 Yöntem	26
3.2.1 Total RNA İzolasyonu.....	26
3.2.2 RNA'ların Kalite Kontrolü	28
3.2.3 RNA Dizileme.....	28

3.2.4 RNA-seq Analizleri	28
3.2.4.1 Hazırlık	29
3.2.4.2 Dizilemenin Kalite Kontrolü	29
3.2.4.3 Trinity Kullanılarak RNA-seq'in <i>de novo</i> Derlemesi	30
3.2.4.4 Transkript İfade Seviyelerinin Belirlenmesi	30
3.2.4.5 Derleme Kalitesinin Değerlendirilmesi	31
3.2.4.6 Örnek Veri Kalitesinin Değerlendirilmesi.....	32
3.2.4.7 Diferansiyel İfade Analizi	33
3.2.4.8 İşlevsel Açıklama Eklenmesi	34
3.2.4.9 Veritabanı Oluşturulması.....	35
3.2.4.10 Gen Ontoloji ve KEGG Yolak Analizleri	37
3.2.4.11 SSR Varlığının Araştırılması	37
3.2.5 RNA-seq Verilerinin Gerçek Zamanlı Kantitatif PCR (qPCR) ile Validasyonu	38
3.2.6 RNA-seq Verilerinin NCBI Veritabanına Yüklenmesi.....	39
4 SONUÇ VE ÖNERİLER	40
4.1 Total RNA İzolasyon Sonuçları ve Karakterizasyonu	40
4.2 RNA-seq Sonuçları	41
4.2.1 Dizilemenin Değerlendirilmesi	41
4.2.2 RNA-seq Verilerinin Kalite Kontrolleri.....	42
4.2.3 RNA-seq Verilerinin <i>de novo</i> Derlemesi.....	45
4.2.4 Transkriptlerin İfade Seviyelerinin Belirlenmesi.....	46
4.2.5 Derleme Kalitesi Analizi Sonuçları	47
4.2.6 Örneklerin Veri Kalitesi Sonuçları	50
4.2.7 Diferansiyel İfade Analizi Sonuçları.....	50
4.2.8 İşlevsel Açıklama Eklenmesi.....	54
4.2.9 Veritabanı Oluşturulması	56
4.2.10 Gen Ontoloji ve Yolak Analizi.....	57
4.2.11 SSR Analizi Sonuçları.....	62
4.3 qPCR Validasyon Sonuçları	62
4.4 RNA-seq Verilerinin NCBI SRA Veritabanına Yüklenmesi	63
4.3 Tartışma ve Öneriler	63
KAYNAKÇA	69
TEZDEN ÜRETİLMİŞ YAYINLAR	87

SİMGE LİSTESİ

bp	Baz Çifti
dk	Dakika
rpm	Dakikada dönüş sayısı
°C	Derece santigrad
g	Gram
gb	Gigabayt
km	Kilometre
m	Metre
µg	Mikrogram
mm	Milimetre
ng	Nanogram
sn	Saniye
cm	Santimetre

KISALTMA LİSTESİ

BP	Biyolojik Proses
BGI	Beijing Genomics Institute
CC	Hücresel Komponent
cDNA	Komplementer DNA
CDS	Protein Kodlayan Dizi
COG	Clusters of Orthologous Genes
DE	Diferansiyel İfade
DNA	Deoksiribonükleik asid
EST	Expressed Sequence Tags
FDA	ABD Gıda ve İlaç Dairesi
FDR	False Discovery Rate
FPKM	Fragments Per Kilobase Million
GO	Gen Ontoloji
KEGG	Kyoto Encyclopedia of Genes and Genomes
k-mer	k Uzunluğundaki Bütün Alt Diziler
LFC	Log2FoldChange Değeri
lncRNA	Uzun Kodlamayan RNA
MF	Moleküler Fonksiyon
mRNA	Messenger RNA
N	Dizideki belirsiz baz
NCBI	National Center for Biotechnology Information
ncRNA	Kodlamayan RNA
Nr	NCBI Protein Veritabanı
Nt	NCBI Nükleotid Veritabanı
ORF	Açık Okuma Çerçevesi
PATH	Trinity Sistem Yolu
PCA	Temel Bileşenler Analizi
qPCR	Gerçek Zamanlı Kantitatif PCR
R	İstatistiksel Programlama Dili
RNA	Ribonükleik asid

RNA-seq	RNA Dizileme
rRNA	Ribozomal RNA
SNP	Tek Nükleotid Farklılıkları
SRA	NCBI Dizileme Okuma Arşivi
SSR	Simple Sequence Repeats
tRNA	Transfer RNA
TPM	Milyonda Transkript Sayısı
UTR	Translate Edilemeyen Bölge
YND	Yeni Nesil Dizileme



ŞEKİL LİSTESİ

Şekil 2.1	Dünyada fındık üretimi ve ihracatı istatistikleri.....	6
Şekil 2.2	Fındık meyvesinin genel yapısı.....	8
Şekil 2.3	<i>de novo</i> transkriptom derlemesinin akış şeması	14
Şekil 2.4	Trinity transkriptom derlemesinin işleyişi	16
Şekil 2.5	Transkriptoma işlevsel açıklama eklenmesi adımları	22
Şekil 4.1	MOPS jel elektroforez yöntemi ile doğrulama	41
Şekil 4.2	FastQC programından elde edilen temel istatistikler A) <i>C. colurna</i> L. B) <i>C. avellana</i> L. Tombul	42
Şekil 4.3	<i>C. colurna</i> L.(A) ve <i>C. avellana</i> L. Tombul (B) temiz okumalarının FastQC programı ile kontrolü sonucu elde edilen grafikler	43
Şekil 4.4	Trinity.fasta içeriğinin örnek görüntüsü	46
Şekil 4.5	RSEM tarafından oluşturulan <i>C. colurna</i> L. transkriptlerinin ifade seviyeleri örnekleri.....	46
Şekil 4.6	RSEM tarafından oluşturulan <i>C. avellana</i> L. Tombul transkriptlerinin ifade seviyeleri örnekleri.....	47
Şekil 4.7	Derleme kalitesinin değerlendirilmesi (ExN50).....	49
Şekil 4.8	Veri kalitesinin değerlendirilmesi A) Derlemeye hizalanan okuma sayıları B) Transkript başına düşen okuma sayısının(log2) dağılımı ..	50
Şekil 4.9	İki fındık türü arasındaki diferansiyel olarak ifade edilen genler (EdgeR'a göre).....	51
Şekil 4.10	Diferansiyel olarak farklı ifade edilen, P değeri 0.001 ve kat değişikliği >4 olarak ifade edilen transkriptlerin sıcaklık haritası	51
Şekil 4.11	Diferansiyel olarak ifade edilen transkriptlerin $FDR \leq 0.05$ olacak şekilde filtrelenerek A) MA plot ve B) volcano plot ile gösterimi.....	54
Şekil 4.12	Transdecoder programının çıktısı olan “Trinity.fasta.transdecoder.pep” dosyasının içeriği	55
Şekil 4.13	SignalP aracının çıktısı olan “signalp.out” dosyasının içeriği	55
Şekil 4.14	Tmhmm aracının çıktısı olan “tmhmm.out” dosyasının içeriği.....	56
Şekil 4.15	Trinotate.xls dosyası içeriği.....	56
Şekil 4.16	<i>C. colurna</i> L. ve <i>C. avellana</i> L. Tombul transkriptlerinin GO kategorilerine göre sınıflandırılması A) Biyolojik Proses (BP) B) Hücresel Komponent (CC) C) Moleküler Fonksiyon (MF)	57
Şekil 4.17	<i>C. colurna</i> L. ve <i>C. avellana</i> L. Tombul transkriptlerinin KEGG Yolak analizi sonuçları.....	58

Şekil 4.18 <i>C. colurna</i> L.'nin diferansiyel olarak ifade edilen genlerinin GO terimleri ilişki ağı.....	59
Şekil 4.19 <i>C. avellana</i> L. Tombul'un diferansiyel olarak ifade edilen genlerinin GO terimleri ilişki ağı.....	60
Şekil 4.20 <i>C. colurna</i> L. ve <i>C. avellana</i> L. Tombul'un diferansiyel olarak ifade edilen genlerinin GO kategorilerindeki dağılımları	61
Şekil 4.21 Transkript başına tanımlanan toplam SSR motif sayıları dağılımı	62
Şekil 4.22 TL1, GMPM1, N, 2MMP, At1g29670, CHIB1 genlerinin qPCR ekspresyon seviyeleri	63



TABLO LİSTESİ

Tablo 2.1 <i>Corylus avellana</i> L. Sistematığı	7
Tablo 2.2 <i>C. avellana</i> L. meyvesinin besin değeri tablosu	9
Tablo 2.3 <i>Corylus colurna</i> L. Sistematığı.....	10
Tablo 3.1 qPCR için kullanılan primer dizileri	38
Tablo 4.1 Nanodrop spektrofotometre ölçüm sonuçları	40
Tablo 4.2 <i>C. colurna</i> L. (Türk Fındığı) ve <i>C. avellana</i> L. Tombul Fındık örneklerinin dizileme sonrası RNA dizileri istatistiksel sonuçları	41
Tablo 4.3 Trinity derleme kalitesi ile ilgili istatistikler	47
Tablo 4.4 Okumaların Trinity derlemesindeki temsil oranları.....	48
Tablo 4.5 Tam uzunluktaki kodlayan transkriptlerin sayısı.....	49
Tablo 4.6 İki tür arasında diferansiyel olarak ifade edilmiş gen ve transkriptler	52
Tablo 4.7 <i>C. colurna</i> L.'de <i>C. avellana</i> L. Tombul'a kıyasla artış gösteren ilk 20 gen	52
Tablo 4.8 <i>C. avellana</i> L. Tombul'da <i>C. colurna</i> L.'ye kıyasla artış gösteren ilk 20 gen	53

Yeni Nesil Dizileme Yöntemi İle Türk Fındığı (*Corylus colurna* L.) Ve Tombul Fındığın (*Corylus avellana* L.) Transkriptom Analizlerinin Yapılması Ve Karşılaştırılması

Salih ULU

Moleküler Biyoloji ve Genetik Anabilim Dalı

Doktora Tezi

Danışman: Prof. Dr. Nehir ÖZDEMİR ÖZGENTÜRK

Fındık, tarih boyunca insanlar için değerli bir gıda olmuştur. Gıda sektöründe çerez olarak, özellikle pasta ve çikolata yapımında tüketilmesi yanı sıra yağ bitkisi olarak kullanılan, ekonomik değeri oldukça yüksek bir bitkidir. Huşgiller (Betulaceae) familyasından Coryleae alt familyasının *Corylus* cinsi içerisinde tanımlanan 25'ten fazla türün ticari olarak en çok tercih edilenleri ise Anadolu kaynaklı olanlarıdır. Tüm dünyanın ihtiyacının yaklaşık %70'inin Türkiye'den karşılanması ve anavatanının Türkiye olması ayrıca Türk ekonomisine katkısı açısından da oldukça önemli bir bitkidir. Ülkemizde birçok çeşidinin ticari olarak yetiştirilmesine rağmen, en yaygın olarak yetiştirilen türleri; ticari olarak en çok tercih edilen Tombul çeşidi olarak isimlendirilen *Corylus avellana* L. Tombul ve ağaç formunda olan ve güçlü kök yapısı sebebiyle ıslah çalışmalarında anaç olarak kullanılan ve Türk Fındığı olarak bilinen *Corylus colurna* L.'dir.

Bu çalışmada Giresun Fındık Araştırma Enstitüsünün koleksiyonunda yer alan Tombul ve Türk fındık ağaçlarının yaprak, erkek ve dişi çiçek, tomurcuk ve zuruflarından total RNA izolasyonu yapıldı. Yeni Nesil Dizileme teknolojisi kullanılarak Illumina HiSeq4000 ile iki uçlu olarak 100 bp uzunluklarında

okumalardan oluřan iki farklı kütüphane oluřturuldu. Dizileme sonrası *C. colurna* L. ve *C. avellana* L. Tombul'dan yaklaşık 10,30 GB ve 10,48 GB transkriptom verisi elde edildi. *C. colurna* L. ve *C. avellana* L. Tombul okumaları *de novo* olarak 95979 transkripte derlendi. TRAPID platformu kullanılarak transkriptlerin GO ve KEGG fonksiyonel özellikleri belirlendi. *C. colurna* L. ve *C. avellana* L. Tombul için sırasıyla 27051 ve 25312 SSR tanımlandı. RNA-seq verilerinin qPCR ile validasyonu için TL1, GMPM1, N, 2MMP, At1g29670, CHIB1 genleri seçildi.

Yapılan bu çalışma ile Anadolu kaynaklı iki fındık türü; en kaliteli fındık kabul edilen Tombul çeşidi fındığın ve anaç olarak kullanılan Türk Fındığının ilk defa kapsamlı transkriptom analizi yapıldı, bu fındık türlerinin gen ekspresyon profilleri ilk defa karşılaştırılarak geniş bir bilgi elde edildi.

Bu veriler fındık yetiřtiricilięi, tıbbi ve endüstriyel arařtırmalar için açık bir kaynak oluřturacaktır.

Anahtar Kelimeler: *Corylus colurna* L., Türk Fındığı, *Corylus avellana* L., Tombul Fındık, transkriptom analizi

Transcriptome Analyzes of Turkish Filbert (*Corylus colurna* L.) and Tombul Hazelnut (*Corylus avellana* L. Tombul) with Next Generation Sequencing

Salih ULU

Department of Molecular Biology and Genetics

Doctor of Philosophy Thesis

Supervisor: Prof. Dr. Nehir ÖZDEMİR ÖZGENTÜRK

Hazelnut has become a valuable nutrient throughout the history. As it is used in cake and chocolate products as well as appetizers and also oil extraction, its economic value is high. More than 25 species have been identified in Coryleae subfamily and Betulaceae family in taxonomical order, whereas commercially most preferable ones are grown in Anatolia region. Hazelnut is an economically important plant for Turkey since 70% of world need is supplied from Turkey. A lot of hazelnut species grow in Turkey, but the most widespread species are *Corylus avellana* L. Tombul which has high economically value because of their special flavor and oil content and Turkish hazelnut *Corylus colurna* L. which is used for breeding as rootstock because of their resistance against extreme nature conditions.

In this study, Total RNA was isolated from young leaves, flowers (male and female), bud, husk shoot of *C. colurna* L. and *C. avellana* L. Tombul from the collection of Giresun Hazelnut Research Institute. Two different libraries were sequenced using Illumina HiSeq4000 with 100 bp paired-end. We generated 10,30 and 10,48 gb of transcriptome data from *C. colurna* L. and *C. avellana* L. Tombul,

respectively. The reads from *C. colurna* L. and *C. avellana* L. Tombul were *de novo* assembled into 95979 transcripts. The transcripts of *C. colurna* L. and *C. avellana* L. Tombul were GO ve KEGG functionally annotated using the TRAPID platform, respectively. We identified 25312 and 27051 Simple Sequence Repeats (SSRs) for *C. colurna* L. and *C. avellana* L. Tombul, respectively. TL1, GMPM1, N, 2MMP, At1g29670, CHIB1 genes were selected for validation with qPCR.

With this study a wide transcriptome analysis were done for Anatolia originated two hazelnuts that most quality Tombul hazelnut (*Corylus avellana* L.) and used as roodstocks Turkish hazelnut (*Corylus colurna* L.) and this was given wide information about their gene expression profile. We report the first de novo transcriptome data of *C. avellana* L. Tombul and compare *C. colurna* L. of commercial importance. These data constitute a valuable extension of the publicly available transcriptomic resource useful for breeding, medicinal and industrial research studies.

Keywords: *Corylus colurna* L., Turkish filbert, *Corylus avellana* L., Tombul hazelnut, transcriptome analysis

1.1 Literatür Özeti

Literatür arařtırmaları sonucunda doktora tezinin odaklandığı konu olan Türk Fındığı (*Corylus colurna* L.) ile Tombul Fındığın (*Corylus avellana* L. Tombul) Yeni Nesil Dizileme yöntemi ile transkriptom analizlerinin yapılması ve karşılaştırılmasının tek bir çalışma içerisinde yer aldığı bir arařtırmaya rastlanmamıştır. Fakat tez çalışmasının fikri zeminini oluşturabilecek çalışmaların özetlerine aşağıda yer vermeye çalışılacaktır.

Arařtırmalarda Tombul Fındık çeşidinin ve Türk Fındığının ekonomik ve genetik özellikleri açısından değerli bitkiler olduğu gösterilmiştir.

Türk Fındığı çok düşük ve çok yüksek sıcaklıklara dayanıklıdır, 1700 m. gibi yüksek rakımlarda yaşayabilmektedir, ağaç formunda olması ve diğer özellikleriyle diğer fındık türlerinden farklılaşmakta ve bu özelliklerinden dolayı ıslah çalışmalarında öne çıkabilmektedir. Ayrıca bu fındık türünün 600 yaşlara varan genetik özellikleri korunmuş örnekleri ülkemizin farklı bölgelerinde tespit edilmiştir [1], [2] .

Tombul Fındığın yağ ve protein içeriğinin yüksek olmasından dolayı en kaliteli ve en çok tüketilen fındık çeşidi olduğu bildirilmiştir ayrıca Tombul Fındığın yağ ve protein içeriği açısından sanayide kullanıma çok elverişli olduğu ifade edilmiştir [3], [4].

Corylus avellana L. varyeteleri ve *Corylus colurna* L. ile ayrı ayrı yapılan çalışmalarda antioksidan aktivite ve fenolik kompozisyon varlığını gösteren güçlü bulgulara rastlanmıştır [5], [6].

Fındık ekonomik değeri oldukça yüksek bir bitki olmasına rağmen, bu bitki için moleküler düzeyde çalışmalar çok az yapılmıştır.

DNA ekstraksiyonunun, fındığın içerdiği yüksek tanenler, alkaloidler, flavanoidler ve polisakkaritler nedeni ile her zaman başarılı olamadığı ve değişik ekstraksiyon yöntemlerinin kullanılması gerektiği belirlenmiştir [7].

DNA miktarının yaprak yaşı ve genotipe bağlı olarak farklılık gösterdiği fındık yaprakları ile yapılan çalışmalarda tespit edilmiş ve 0,5 - 0,7 g fındık yaprağından yaklaşık 150 ng ile 15 µg arasında DNA elde edildiği görülmüştür [8].

İtalya'da yetiştiriciliği yapılan 19 fındık çeşit ve genotipinin karakterizasyonu üzerine yapılan çalışmada 30 adet RAPD primeri kullanılmıştır. Çalışma sonucunda elde edilen bulgularla oluşturulan dendogramda iki ana çeşit grubu oluşmuş, Tombul çeşidi, 'Imperale Trebizonde', 'Fructo Rubro' ve 'Jean's' çeşitleri ile aynı grupta yer almıştır [9].

C. avellana L. türüne ait 274 genotipin genetik profillerini incelemek amacıyla 10 SSR primeri kullanılmış, incelenen genotiplerin 84'ünün aynı klonlar olduğu anlaşılmış ve kullanılan primerlerin yüksek oranda polimorfik oldukları belirlenmiştir [10].

Farklı coğrafik bölgelerde yayılım gösteren *C. avellana* L. türüne ait 270 adet klonun genetik profillerini incelemek amacıyla yapılan çalışmada 21 SSR primeri kullanılmıştır. Sonuçlara göre genotiplerin Orta Avrupa, Karadeniz, İngiltere ve İtalya-İspanya şeklinde dört ana coğrafi bölge altında toplandığı bildirilmiştir [11].

Bu çalışmada kullanılan Yeni Nesil Dizileme (YND) teknolojisi ile ilgili çalışmalara aşağıda yer verilmiştir.

Geleneksel yöntemlere kıyasla YND'nin en başta gelen avantajı, büyük miktardaki dizileme verilerinin üretilmesidir [12].

YND'nin temel kavramları, genel uygulamaları ve genomikteki potansiyel etkilerini referanslarıyla incelenmiş ve günümüzde ve gelecekte mümkün olan platformları ve biyoinformatik yaklaşımlar ortaya konulmuştur. Araştırmacılar biyoinformatik araçların YND teknolojisinin başarısında çok önemli bir yere sahip olduğunu belirtmişlerdir. Ayrıca YND teknolojilerinin genomik üzerinde

etkilerinin çok geniş kapsamlı olacağını ve büyük olasılıkla önümüzdeki yıllarda farklı alanlara da kayacağını ifade etmişlerdir [13].

Araştırmacıların hizmetine sunulan Trinity adı verilen bir dizi programla RNA-Seq verilerinden referans genom oluşturma ve analizler için bu platform kullanılarak de novo transkript dizisi kurulmasını sağlanmıştır. Bu sayede referans genomu mevcut olmayan türlerin transkriptom analizlerinin de yapılması mümkün hale gelmiştir. Ayrıca çalışmalarda karşılaştırmalı örneklerin farklı ifade edilen transkriptlerinin tanımlanması için R/Bioconductor paketleri ve protein kodlayan genlerin tanımlanması için yaklaşımlar sunulmuştur [14].

Illumina RNA-seq metodu kullanılarak *Vitis vinifera*'nın farklı durumlarındaki transkriptom analizi yapılmış ve 59 milyon okuma sonucunda 17.324 ifade olan gen elde edilmiştir. Bu sonuçlarla genlerin SNP analizleri ve splice mekanizmaları ortaya çıkarılmıştır [15].

Avrupa fıındığı "Jefferson" ile transkriptom çalışması için bitkinin farklı organlarından örnekler alınarak RNA'ları izole edilmiş ve Illumina RNA dizileme (RNA-Seq) ile 28.255 kontig elde edilmiştir [16].

Nükleik asit dizilemenin, belirli bir DNA ya da RNA molekülü içinde mevcut nükleotid tam sırasını belirlemek için iyi bir yöntem olduğu ifade edilmiştir. Ayrıca YND teknolojisi adım adım açıklanmış ve bu yöntemin kalıp hazırlanması, dizileme ve görüntüleme, dataların analizleri şeklinde üç temel adımda ilerlediği söylenmiştir [17].

Vitis vinifera cv. Corvina türünün *de novo* transkriptom karakterizasyonu için 45 farklı örnekten 1 total RNA alınarak RNA havuzu oluşturulmuş ve RNA-seq dizileme yapılmıştır [18].

Çin fıındığının (*Corylus mandshurica*) transkriptom analizi yapılmak üzere RNA-seq ile EST kütüphanesi kurulmuştur. Çalışmada 37.846 EST dizisi incelenip GO analizleri yapılmıştır. Taxol sentezinde sorumlu olan enzim ve transkripsiyon faktörlerinin 45 EST'de homoloji gösterdiğini tespit edilmiştir. Bunun yanında *C. avellana* ile karşılaştırılarak soğuk ve fungal etkilerle ilgili stres tolerans genlerine karşılaştırmalı olarak bakılmıştır [19].

Corylus heterophylla Fisch'in soğuğa dayanıklılık genlerinin transkriptomik dizilemesini yapmak için Illumina RNA-seq kullanılmıştır. Çalışmada total RNA izolasyonları için RNeasy Plant Mini Kit (Qiagen), DNA kontaminasyonunu önlemek için RNase-free DNase I (Qiagen) ve RNA konsantrasyonunu artırmak ve pürifiye etmek için RNA MinElute Kit (Qiagen) kullanılmıştır. Bu çalışmada 28 milyon okuma elde edilmiş, 82.684 kontig oluşturulmuştur. 40.941 unigen belirlenmiş ve GO kategorilerine göre tanımlanmıştır. Bu çalışma esnasında NCBI Nr, Swiss-Prot, KEGG veritabanlarını kullanılmıştır [20].

Fındıkta meyve gelişimi esnasında boş ve dolu ovüllerin transkriptom analizi ve gen ifadesi profillerini çıkarmak için bir çalışma yapılmış, çalışmalar referans genom olmadığı için *de novo* transkriptom dizileme ve RNA-seq analizleri ile yürütülmüştür. Hizalama ve kontig oluşturmak için Trinity platformu kullanılmıştır. Protein tahmini için NR, NT, SwissProt, KEGG, COG, and GO veritabanlarından faydalanılmıştır [21].

Türk Fındığının Kastamonu bölgesinde yeni bir destinasyonunu rapor edilmiş ve bu fındığın yağ ve profil içeriğine bakılmıştır [22], [23]

C. avellana L.'nin gen ifadesi analizi yapılmış [24], *C. avellana* L.'nin mahsül gelişimi hedefli olarak kromozom ölçekli bir çalışma yürütülmüştür [25]. Ayrıca Tömbul Fındık kloroplast genomunun tamamlanabilmesi için farklı araçların kullanıldığı bir çalışma ortaya konulmuştur [26].

1.2 Tezin Amacı

Ekonomik olarak önemli olan Tömbul Fındık (*C. avellana* L. Tömbul) ile ıslah çalışmalarında kullanılan ve genetik olarak önemli bir tür olan Türk Fındığının (*C. colurna* L.) YNS ile karşılaştırmalı transkriptom analizi yapılarak gen ifadesi düzeyinde farklılıkların ortaya konulması amaçlanmıştır.

1.3 Hipotez

Tömbul Fındığın (*C. avellana* L. Tömbul) dünya çapında tercih sebebi olmasını sağlayan özellikleri ve Türk Fındığının (*C. colurna* L.) güçlü genetik yapısı ve ıslah çalışmalarında kullanılmasını sağlayan özellikleri transkriptom çalışmalarıyla ortaya konulmaktadır.

2.1 Fındık

Fındık, Fagales takımının Betulaceae familyasına ait popüler bir ağaç meyvesidir. *Corylus* cinsine dahil olan fındık ağaçları kuzey yarım kürenin ılıman bölgelerinde yabani olarak yetişmektedirler. Bu yabani türler Japonya, Çin, Türkiye, Avrupa ülkeleri, ABD gibi çok geniş bir alana yayılmıştır ve bunlar içinde en fazla yabani yayılım bölgesi Avrupa'dır [27]. Bu yayılımın kökeninin Anadolu, Orta Asya ve Kafkasya olduğu, Anadolu'da da ilk defa Doğu Karadeniz Bölgesinde kültüre edildiği bilinmektedir [28], [29].

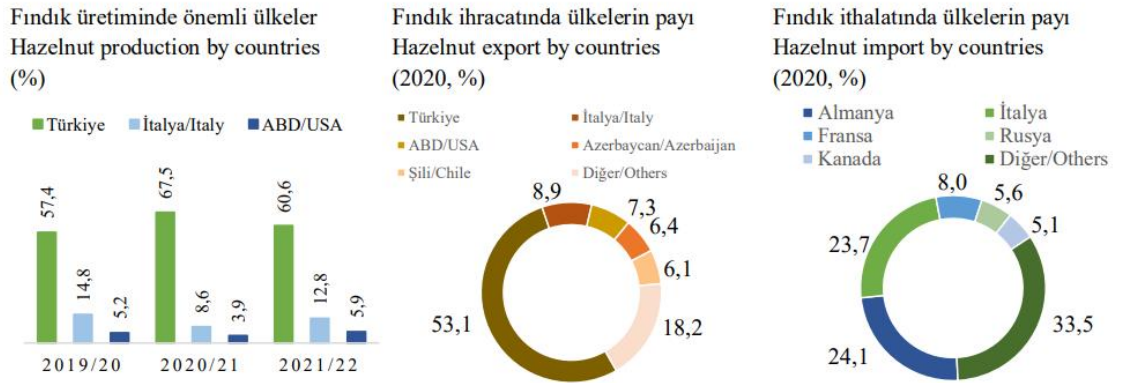
Corylus cinsinin üyesi olan türler diploiddir ve $2n = 2x = 22$ kromozoma sahiptir. Sporofitik uyumsuzluk etmenli yabancı tozlanmanın sonucu olarak *Corylus* cinsi büyük bir morfolojik varyasyona sahiptir. *Corylus* cinsi içerisinde 25'ten fazla tür tanımlanmıştır [27]. Bu türlerden ağaç formunda olanlardan başlıcaları *C. colurna* L., *C. jacquemontii* Decne, *C. chinensis* Franch, ve *C. ferox* Wall türleri iken; çalı formunda olanlar *C. avellana* L., *C. maxima* Mill, *C. americana* Marshall, *Corylus cornuta* Marshall, *C. heterophylla* Fitcher, *C. sieboldiana* Blume türleri olarak sıralanabilir [30]

Fındık türleri arasında pek çoğu yabani kalmasına rağmen bir kaç tür ticari öneme sahiptir; *C. avellana* L., *C. colurna* L. ve *C. maxima* Mill [27]. 2019 yılı verilerine göre fındık dünya kabuklu meyve üretiminde 4. sırada yer almaktadır [31]. Fındığın popüleritesi her geçen gün artmakta, bunun sebepleri arasında kendine has lezzeti, besin içeriği, sağlığa olan katkıları sayılmaktadır [32]–[35]. İçeriğindeki proteinler, lifler, yağlar ve mineraller açısından oldukça zengin olan fındık önemli bir gıda maddesi konumundadır. Bünyesinde lipitler ve fenolik bileşikler barındırmaktadır. Fındık yağı %92 oranında doymamış yağ asitlerinden oluşmaktadır [oleik (ω -9) ve linoleik asit (ω -6)]. Fındığın kendine has bu besinsel bileşimiyle kardiyovasküler sistem ve kolesterol seviyesi kontrolü gibi insan

sağlığına olumlu katkıları tespit edilmiştir [36]–[40]. ABD Gıda ve İlaç Dairesinin (FDA) fındık vb. ürünleri kalp sağlığını destekleyen ürünler kategorisine almasının da bu ilginin sebepleri arasında olduğu düşünülmektedir [35].

Bu artan ilgi ile birlikte fındık üretimi de günden güne artmaktadır. Şekil 2.1’de gösterildiği üzere, Tarım ve Orman Bakanlığı, 2022 Tarım Ürünleri Piyasa Raporu verilerine göre Türkiye dünyadaki fındık üretiminde %60’lık oran ve 665 bin tonluk üretim ile ilk sırada yer almaktadır. Sakarya, Düzce, Samsun, Ordu, Giresun ve Trabzon illerindeki üretim ile Karadeniz Bölgesi Türkiye fındık üretiminin %92’si karşılanmaktadır [41]. Türkiye’yi 120 bin tonluk fındık üretimi ile İtalya, 59 bin ton ile ABD takip etmektedir [31].

Fındık Türkiye ekonomisine katkısı oldukça büyük bir tarım ürünüdür. Bu ürün sayesinde hem tarımsal istihdam sağlanmakta hem de ihracat yolu ile ülke ekonomisine katkı sağlanmaktadır. 2022 verilerine göre fındık ihracatından elde edilen gelir 2 milyar USD’dir. Toprak Mahsulleri Ofisi Fındık Sektör 2020 Raporuna göre, fındık, ülkemizin tarımsal ihracatında %13 ile ilk sırada yer almaktadır.



Şekil 2.1 Dünyada fındık üretimi ve ihracatı istatistikleri

2.1.1 *Corylus avellana* L. (Avrupa Fındığı)

Literatürde Avrupa Fındığı veya Adi Fındık olarak anılan ve ülkemizin doğal yayılış alanlarından biri olduğu *Corylus avellana* L. (Tablo 2.1) Karadeniz Bölgesi ormanlarının 20-1500 m’lik yükseltilerinde sıklıkla, nadiren de Doğu Anadolu, Güney Anadolu ve Batı Anadolu’daki dere kenarlarında görülmektedir [42]. Karadeniz Bölgesinde özellikle 60 km’lik iç kesimlere ve 750 m yüksekliklere kadar

yetiştirilse de fındık üretiminin en yoğun yapıldığı yerler 0-500 m yükselti aralığıdır.

Fındık verimini pek çok unsur etkilemekle birlikte başlıcaları sıcaklık, yağış, don ve rüzgardır. Yağış miktarının yılın her mevsiminde dengeli olması gereklidir [43]. Fındık üretiminde en kritik koşul sıcaklıktır. En az -8, -10 °C ve en fazla 36, 38 °C sıcaklıklarda verim alınabilmektedir, ortalama sıcaklığın 13-15 °C olduğu yerler üretim için ideal koşullardır. İlkbahar mevsimi donları verimi düşürmektedir, ayrıca Karadeniz kıyı kesimlerindeki fön rüzgarları da fındık tarımını etkileyen koşullardandır [43].

Toprak koşulları açısından çok seçici olmamasına rağmen humuslu, tınlı-humuslu ve derinliği fazla olan topraklara uyumu yüksektir. Buna karşın kum oranının yüksek olduğu killi topraklarda verimli bir tarım mümkün olmamaktadır [44].

Tablo 2.1 *Corylus avellana* L. Sistematigi

Divisio (bölüm)	Spermatophyta
Subdivisio (alt bölüm)	Angiospermae
Classis (sınıf)	Magnoliatae (Dicotyledoneae)
Subclassis (alt sınıf)	Hamamelidae
Ordo (takım)	Fagales
Familia (familya)	Betulaceae
Subfamilia (alt familya)	Coryleae
Genus (cins)	<i>Corylus</i> L. (hazelnut)
Species (tür)	<i>Corylus avellana</i> L. (European Hazelnut)

5-6 yaşından itibaren meyve vermeye başlayan fındık ağaçlarının ortalama uzunlukları 6-7 m olmakta iken 20-22 m uzunluğa ulaşan örneklerine de rastlanmaktadır. *C. avellana* L. 60 cm derinliklere ulaşan saçak köke sahipken su ve besin ihtiyaçlarını çoğunlukla ilk 20 cm'lik kısımdan karşılamaktadır.

Meyveyi dışarıdan koruyan renkli tüylü brakte yaprağına zuruf denilmektedir. Meyvenin olgunlaşma sürecinde zuruf meyveyi sıkarak ve dip kısımlarından kahverengileşme başlar. Kendi zurufları içerisinde alt kısımlarından birbirlerine bitişik halde gelişen meyveler kümesine çotanak adı verilmektedir (Şekil 2.2). Çotanak ise ince bir dalla yıllık sürgüne bağlanmaktadır [45].



Şekil 2.2 Fındık meyvesinin genel yapısı [46]

Fındık tek evcikli, farklı zamanlarda olgunlaşan erkek ve dişi çiçekleri olan (dikogami) ve erkek ve dişi çiçeklerin aynı ağaç üzerinde görüldüğü bir bitkidir [47]. Çiçeklenmenin ve tozlaşmanın görüldüğü zamanlar Kasım-Mart aralığıdır. Yumurta ve tohum taslağı geç olgunlaştığı için tozlaşmadan 2-2,5 ay sonra döllenme gerçekleşir. Döllenme sonrasında 15. günden itibaren de meyve görülmeye başlar [48]. Karanfil, 4-24 tane dişi çiçeğin bir arada bulunduğu tomurcuğun adıdır; kedicik ise, 150-200 adet erkek çiçeğin bir arada bulunduğu silindirik yapının adıdır ve Temmuz ayında oluşmaya başlar.

Corylus avellana L., yüksek oranda doymamış yağ barındıran, protein ve vitamin içeren bir besindir [49]. Vitamin ve mineraller açısından bakıldığında ise, kalsiyum, potasyum, sodyum, demir, çinko, bakır, magnezyum ve manganez, E, B1, B2, B3, B6, α -tokoferol barındırdığı (Tablo 2.2) tespit edilmiştir [50].

Tablo 2.2 *C. avellana* L. meyvesinin besin değeri tablosu

	100 gr.	1 avuç
Karbonhidrat (g)	17	4
Protein (g)	14	3,6
Yağ (g)	61	14,6
Lif (g)	10	2,3
Kolesterol (mg)	0	0
Sodyum (mg)	0	0
Potasyum (mg)	680	163,2
Kalsiyum (mg)	114	27,4
Vitamin A (mg)	20	4,8
Vitamin C (mg)	6	1,5
Demir (mg)	5	1,1

2.1.1.1 Tombul

Tüm dünyada en fazla kabul gören fındık çeşidi olan Tombul, Karadeniz kıyı bölgesinde, özellikle de Giresun civarında yetişmektedir. Yüksek meyve kalitesi uluslararası pazarda aranan bir ürün olmasını sağlamıştır. Tombul Fındığın periyodisite özelliği olmasına rağmen bakımlı bahçelerde her yıl yüksek verime ulaşabilmektedir [51].

Parlak kahverengi, yaklaşık 1 mm kalınlığında, kolay kırılan bir kabuğa sahiptir. Meyve kabuğu tam doldurmaktadır. Meyve ölçüleri kabuklu şekilde; ortalama uzunluk 17.5 mm, genişlik 17 mm ve ağırlığı 1.4 g'dır. İç meyve zarı pürüzsüz, parlak, açık kahverengi, ince ve üzerinde hafif damarları vardır ve kolayca soyulabilmektedir. Beyaz, gevrek ve parlak olan meyve etinin göbek boşluğu küçüktür.

Tombul Fındığın zurufları meyve boyunun yaklaşık 2,5 katı kadar olmakta, 3'lü veya 4'lü çotanaklar oluşturmaktadır [51].

2.1.2 *Corylus colurna* L. (Türk Fındığı)

Corylus colurna L. (Tablo 2.3) literatürlerde “Turkish Hazelnut” veya “Turkish Filbert” şeklinde anılan, 15-25 m boylarında düzgün, tek gövdeye sahip ağaç formunda bir bitkidir [52]. En sık kullanılan ismi Türk Fındığı olmakla birlikte, ağaç formu nedeniyle ağaç fındığı olarak da anılmakla, Ayı Fındığı, Kaya Fındığı ve Balkan Fındığı olarak da bilinmektedir. Gökbulak Fındığı ve Budağan Fındığı gibi yayılış yerine göre isimlendirildiği de görülmüştür [53]. Bazı kaynaklar *C. arborescens* Mönch., veya *C. bizantica* Clus. sinonimlerine de yer vermiştir [54].

Tablo 2.3 *Corylus colurna* L. Sistematigi

Divisio (bölüm)	Spermatophyta
Subdivisio (alt bölüm)	Angiospermae
Classis (sınıf)	Magnoliatae (Dicotyledoneae)
Subclassis (alt sınıf)	Hamamelidae
Ordo (takım)	Fagales
Familia (familya)	Betulaceae
Subfamilia (alt familya)	Coryleae
Genus (cins)	<i>Corylus</i> L. (hazelnut)
Species (tür)	<i>Corylus colurna</i> L. (Turkish Hazelnut)

Türk Fındığı oldukça geniş aralıktaki yaşam koşullarında varlığını sürdürebilmektedir. Balçıklı toprakları sevmekle birlikte aşırı alkalileşmiş topraklarda da bulunabilmektedir, Nemli ortamları tercih etmekle birlikte [55] kuraklığa dayanımı oldukça yüksektir, aynı zamanda -30 °C soğuklara da yaşamını sürdürebilmektedir [56].

Türk Fındığı, hem dikey hem de yatay yönde kuvvetli bir kazık kök sistemi oluşturabilmesi, aynı zamanda kök ve kütük sürgünü verme yeteneği, fırtına ve kar zararlarına, taşkına ve sele karşı dirençli olması sebebiyle erozyon kontrolünde ve ağaçlandırma çalışmalarında kullanılabilir [2], [56], [57].

Türk Fındığı bu özellikleriyle, öncelikli olarak *Corylus avellana* L. olmak üzere diğer fındık türlerinin kültüre edilmesinde kök altlığı olarak da kullanılmaktadır bu da onun fındığın kültüre edilmesindeki önemini artırmaktadır [58].

Bu fındık türü, olgunluk döneminde piramit yapısında ileriki dönemlerinde tepesi geniş bir ağaç görünümüne sahip olmaktadır. Genç sürgünlerinde soluk pas rengi ve sık, basit ve yapışkan bezeli tüylerin görüldüğü gövde yapısı, yaşlı örneklerinde kalın, derin çatlakları olan mantarlı bir görünüme kavuşmaktadır [59].

Yapraklar 7-14 x 5-11cm boyutlarında yumurta, geniş yumurta ve nadiren ters yumurta biçiminde çift sıralı dişli veya dilimli dişlidir. Uçları aniden sivrileşen yapraklar küçük bir damla ucu ile sonlanır. 2,5-4 cm uzunluğunda olan yaprak sapları siğilli yapışkan tüylere sahiptir. 3-5 meyve bir arada bulunmaktadır [59].

Diğer fındık türlerinden, meyve örtüsünün kenarları ince şeritler halinde ve düzensiz olması, geriye doğru kıvrılan sivri uçları ve üzerinin yapışkan tüylerle kaplı olmasıyla kolayca ayrılmaktadır. Meyvelerin boyutları 15-20 x 10-18 mm, üstten hafif basık şekilde, geniş yumurta görünümündedir. Diğer fındık meyvelerine kıyasla Türk Fındığının perikarbi çok kalın kabukludur. Bir diğer karakteristik özelliği ise meyvenin örtüye bağlandığı mat kısmın fındığın yarı boyuna ulaşmasıdır [53]

C. colurna L.'nin Türkiye'deki diğer fındık türlerinden farklı diğer bir özelliği de epigeik çimlenme özelliğine sahip olan tohumlarının olması ve etli ve sulu bir meyve örtüsü, kupula ve zuruftan oluşan involukruma sahip olmasıdır [60]

Ülkemizde Batı Anadolu'da, Kazdağı, Karabük, Kastamonu ve Bolu; Doğu Karadeniz'de, Rize ve Trabzon'da bitki florasında bulunan Türk Fındığının düşey yayılışı ise 800-1700 m aralığıdır. Türkiye'de yoğun olarak bulunduğu bölge Kuzeybatı Anadolu Ormanlarıdır [58], [61]–[63].

Meyveli bir tür olan Türk Fındığının meyveleri “Fındık Şekeri” ve “Bolu Çikolatası” gibi yörelere özel tatlılarda kullanılmakta veya kuruyemiş olarak tüketilmektedir. Ayrıca meyve tohumları ve yaprakları tıbbi değere sahip maddeler barındırmaktadır. Yüksek miktarda oleik asit barındıran tohumlarının yanında, yaprakları da flavonoid adı verilen oldukça güçlü bir antioksidan madde içermektedir [58]. *Corylus colurna* L. meyvelerinde, %65-70 yağ, %16-19 protein ve %3-6 su bulunmaktadır. Yağın %92’si doymamış yağlar, %8’i doymuş yağlardan meydana gelmiştir [58], [64]. Mineral içeriği ise 100 g’da, 562 mg potasyum, 407 mg fosfor, 77 mg kalsiyum ve 20 mg sodyum şeklinde sıralanmaktadır [58], [65].

2.2 Ribonükleik Asit (RNA)

DNA olarak depolanan genetik bilginin RNA olarak kopyalanması ve sonrasında da proteinlere çevrilmesini ifade eden bilgi akışına santral dogma denilmektedir [66], [67]. Çevresel faktörler tarafından da değiştirilebilen genetik bilginin bu ifadesi bir organizmanın fenotipini belirlemektedir.

Ribonükleik asitler (RNA’lar), hücrelerde ve organizmalarda önemli bir biyomolekül sınıfıdır [68]. Genomun transkripsiyonu yapılan veya ifadesi gerçekleştirilen transkriptom çıktısını ifade ederler. Genin ifadesinde ve sonrasında da genotipin fenotipe bağlanmasında çok önemli roller üstlenen çok sayıda RNA tipi vardır. Ribozomal RNA’lar (rRNA) ve taşıyıcı RNA’lar (tRNA) proteinleri sentezleyen yapıları oluştururlar. Kodlamayan RNA’lar (ncRNA) ise gen ifadesinin kendisi de dahil pek çok biyolojik sürecin düzenlenmesinde görev alırlar. Mesajcı RNA’lar (mRNA) önemli bir RNA sınıfını oluşturur. mRNA dizileri ribozomlar tarafından sentezlenecek olan proteinleri kodlar [69]–[71]. Yakın zamana kadar mRNA dışındaki RNA tipleri kodlamayan RNA’lar olarak kabul edilirken, son araştırmalar uzun kodlamayan RNA’ların (lncRNA) kısa peptidleri kodlayabildiğini göstermiştir [72].

2.2.1 RNA Dizileme (RNA-seq)

Uygun fiyatlı YND platformlarının ortaya çıkmasıyla birlikte RNA-seq kullanılarak RNA’nın yüksek verimli profillemesini yapmak mümkün hale gelmiştir [73]–[76]. RNA-seq çeşitli amaçlar için kullanılabilir [77]. En popüler kullanım

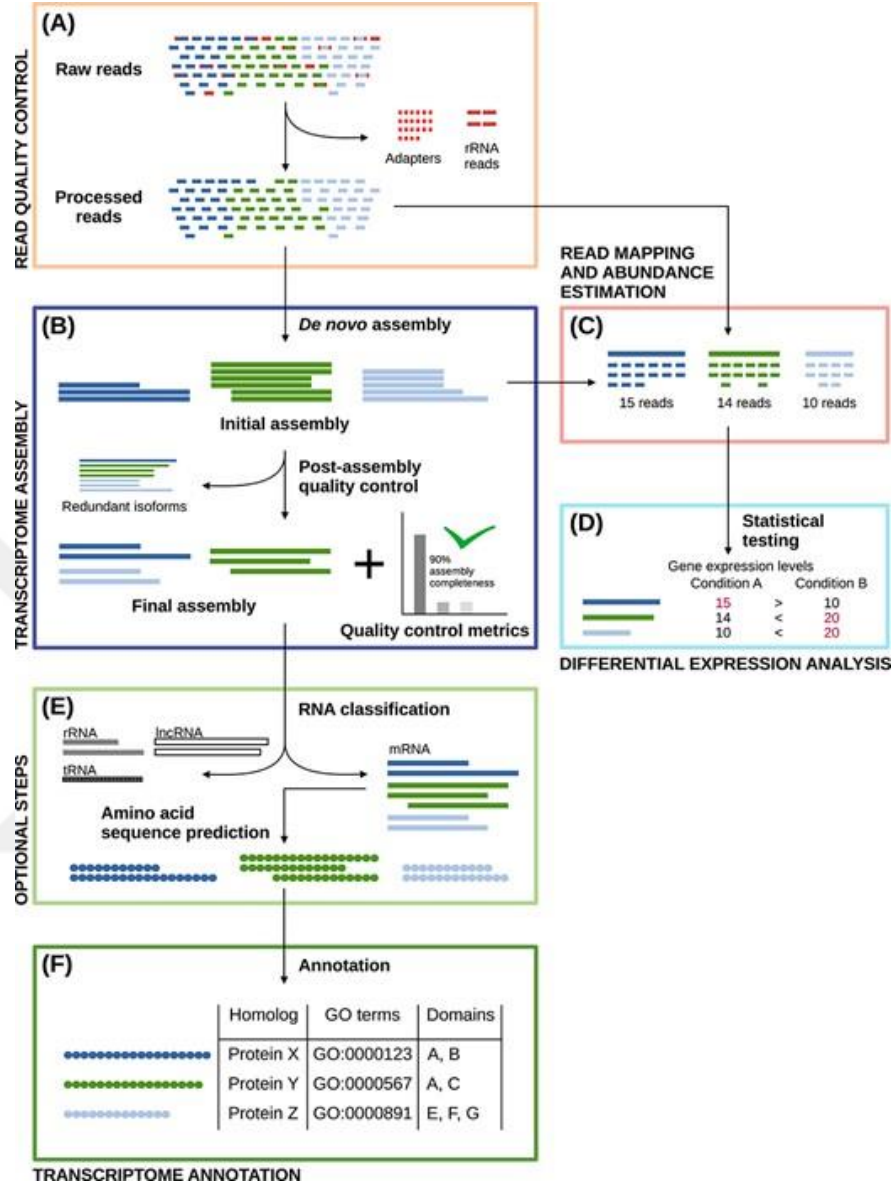
durumları transkriptomların fonksiyonel anotasyonu diye adlandırılan bir organizmanın genlerinin ve proteinlerinin bir kataloğunu oluşturmak ve diferansiyel ifade (DE) analizi adıyla anılan gen ifadesindeki değişiklikleri incelemektir. RNA-seq yaygın olarak bir hücre popülasyonundan gelen materyalin dizileme için bir araya getirildiği yaklaşımı temsil eder [78].

Günümüzde çoğu RNA dizileme çalışması kısa okuma dizilimine dayanır. RNA molekülleri izole edilir ve zenginleştirilir. Sonrasında tamamlayıcı DNA'ya (cDNA) ters kopyalanır. cDNA dizileri fragmanlara ayrılır, rastgele seçilerek PCR ile amplifiye edilir böylece dizileme için cDNA kütüphanesi elde edilmiş olur. Dizileme çıktısı, içerisinde nükleotidlerin sıralandığı milyonlarca kısa okumadan oluşmaktadır [75], [79], [80]. Bu kısa okuma dizileri 50-350 baz çifti (bp) arasında herhangi bir uzunlukta olabilir. Dizileme için kullanılan kütüphane yalnızca belli uzunluktaki parçaları tutmak için dizileri yeniden boyutlandırır yani filtreler. Bu adımdan sonra kısa okumalar tekrar birleştirilmekte ve ismine transkriptom derlemesi adı verilmektedir [80]–[82].

Diziler referans kılavuzlu veya *de novo* olarak birleştirilebilir. Referans kılavuzlu yaklaşımda organizmanın veya yakın bir türün genomu gereklidir. Daha sonra okumaların hangi genlerden kaynaklandığını belirlemek ve ardından ona karşılık gelen transkriptleri yeniden yapılandırmak için okumalar bu referans genomla eşleştirilmektedir [81].

De novo transkriptom düzeneğinde ise referans kılavuzlunun aksine referans gerekmemektedir. Akış şeması Şekil 2.3'de gösterilmiştir. Yeniden yapılandırma sürecine kaynaklık edecek herhangi harici bir bilgi bulunmadığı için süreç latince anlamıyla en baştan (sıfırdan) yani *de novo*'dur. Yalnızca okumalarda yer alan bilgiler kullanılarak gerçekleştirilmektedir. Bu yaklaşım bir genom mevcut olmadığında veya referans kılavuzlu bir derleme istenmediğinde tercih edilir [81], [83]. Ancak *de novo* birleştirilmiş diziler kendi başlarına bilgi vermezler. Onlar öncelikle anlaşılır şekilde tanımlanmalı ve sonrasında biyolojik anlamlılıklarının ortaya konulabilmesi için fonksiyonel ve evrimsel özelliklerinin karakterize edilmesi gereklidir. Bu sürece yeni açıklamalar eklenmesi veya transkriptom

anotasyonu adı verilmektedir. Bu sürecin amacı dizilerin işlevlerini açıklamak olduğundan genellikle fonksiyonel anotasyon olarak da adlandırılır [78].



Şekil 2.3 *de novo* transkriptom derlemesinin akış şeması [78]

2.2.1.1 Derleme Öncesi Kalite Kontrol ve Filtreleme

Dizileme işleminde üretilen okumalar derleme için gerekli verileri oluşturur. Modern dizileme teknolojilerinin hata oranları düşük olsa da ürettikleri veriler hatasız değildir [84]. Okumaların bolluğu, okuma uzunluğu, GC içeriği, k-mer bileşimi ve sistemsel hatalar gibi pek çok etmen derlemenin kalitesini ve sonraki tüm süreci doğrudan etkiler [85]. Bu nedenle *de novo* transkriptom derlemesindeki ilk adım okuma verilerinin kalite kontrolünü içerir. Bu aşamadaki

kalite kontrol hem verilerin kontrolünü hem de gerekli görüldüğünde düzeltmeyi veya filtrelemeyi ifade eder.

Derleme öncesi kısa okumaların kalitesi için, baz başına PHRED kalite puanı, tanımsız (N) bazların ortalama görülme sıklığı, GC içeriği, okuma uzunluğunun dağılımı, aşırı temsil edilen dizilerin tespiti ve adaptör kalıntılarının varlığının kontrol edilmesi gerekmektedir. Bu kontrollerin yapılabilmesi için hızlı sonuç veren ve sonuçları grafikler ve görsellerle destekleyen FastQC [86] birden fazla veri seti birlikte incelenmek isteniyorsa MultiQC [87] ve FastQC'ye alternatif yeni bir araç olarak Falco [88] kullanılabilir.

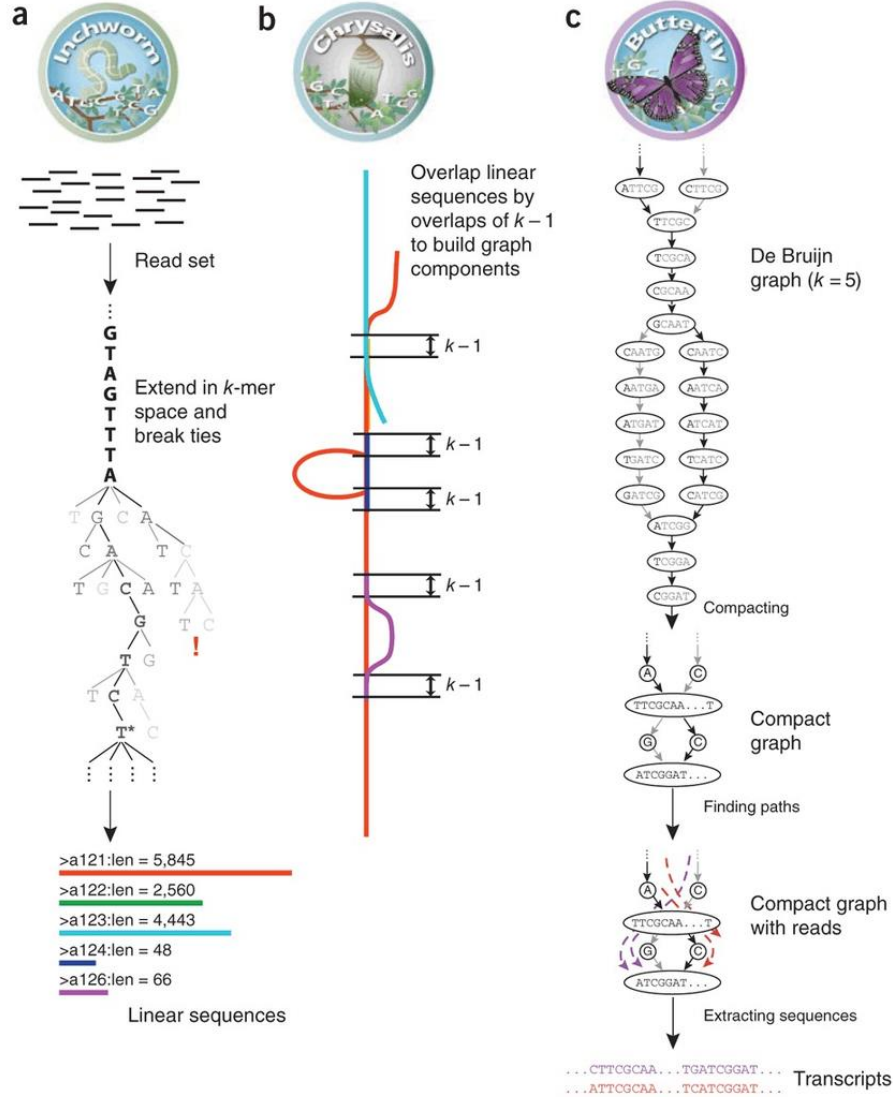
Sonraki adımda anormal okumaları düzeltmek veya hariç tutmak için çeşitli önlemler alınabilir. Uygulanacak ilk prosedür k-mer tabanlı okuma hatalarıdır. Rconnector [89] dizileme esnasında oluşturulan rastgele hataları düzeltmek için kullanılabilir. Kalite kontrol sonuçları verilerde adaptör varlığı tespit ediyorsa bunlar birleştirmeden önce kaldırılmalıdır. Bunun için cutadapt [90] kullanılabilir. Dizileme verilerinde belirsiz baz içeren (N) okumalar da bulunabilmektedir. Benzer şekilde veriler istenilen minimum kalite skorunda tutulmak üzere filtrelenebilir. Aynı zamanda aşırı kısa okumaları verilerden çıkarmak da faydalı olabilir. Otomatik adaptör algılama ve çıkarma, N içeriği tespiti ve çıkarma, düşük kaliteli bazların filtrelenmesi gibi işlemlerin tümü için Fastp [91] kullanılabilmektedir. Bu araca alternatif bir araç da Trimmomatic [92] aracıdır.

Temel temizlik işlemi gerçekleştirildikten sonra, veriler kirleticilerin varlığı açısından değerlendirilebilir. Bunlar tipik olarak ilgili organizma türünden ve/veya RNA türünden kaynaklanmayan okumalardır. İki tür kirleticiden bahsedilebilir, yabancı kirleticiler ve aynı kökenli kirleticiler. Yabancı kirleticiler için yani farklı organizma kalıntıları için Kraken2 [93] kullanılabilmektedir. Aynı kökenli kirleticiler içinse hedef RNA türü dışındaki kirleticilerdir. Bunun için SortMeRNA [94] kullanılabilmektedir.

2.2.1.2 *De novo* Transkriptom Derlemesi

RNA-seq okumaları, farklı transkriptlerin farklı bölümlerine karşılık gelen parçaların bir karışımını içerir. Transkripsiyonel gürültü [95], [96], dizileme

hataları [97] ve alternatif derleme [98], [99] kaynaklı transkript izoformları da bu verilerde temsil edilmektedir. Derlemenin amacı bu belirsizliği doğru bir şekilde ortadan kaldırmak ve kaynak dizilerin hatasız bir şekilde temsili için yeniden oluşturmaktır. Bu tipik olarak daha uzun, birleştirilmiş diziler yani kontigler elde etmek için örtüşmeleri inceleyerek elde edilir [81], [100].



Şekil 2.4 Trinity transkriptom derlemesinin işleyişi [101]

De novo derlemesi için çok sayıda araç mevcuttur, en bilindik De Bruijn tabanlı derleme programı Trinity'dir [14], [101]. 2011'de yayınlanmasından bu yana ilgili yayına 10.000'den fazla atıf yapılmıştır. Çok kapsamlı ilişkili pek çok araca sahiptir ve kullanımı nispeten kolaydır. Trinity'nin en belirgin özelliklerinden birisi biyolojik olarak birbiriyle ilişkili olabilecek kontig kümelerini tanımlamasıdır ve

bunları izoform olarak belirlemesidir [102], [103]. Trinity'nin transkriptom derlemesinin adımları Şekil 2.4'te gösterilmiştir. Bu programa alternatif olarak SOAP-denovo [104], Oases [105] ve Trans-ABYSS [106] gibi derleyici verilebilir.

2.2.1.3 Derleme Sonrası Kalite Kontrol

RNA-seq verilerinin karmaşık doğası nedeniyle *de novo* düzenekler intronik dizileri ve diğer transkripsiyonel yan ürünleri içerebilirler. Ayrıca birleştirme işleminin kendisinin de hatalara açık bir işleyişi vardır [107]. İlgisi olmayan fakat benzer transkriptler zaman zaman birleştirme esnasında yanlış bir şekilde tek kontigde birleştirilebilmektedir [108] veya k-mer uzunluk seçimleri uygun olmadığı için hatalı birleştirmeler meydana gelebilmektedir. Bu nedenle bir *de novo* transkriptom derleme düzeneğinin kalitesini değerlendirmek açıklama eklemekten ve ileri seviye analizlerden daha önemli bir adımdır. Düşük kaliteli bir derleme genlerin tanımlanması ve diferansiyel ifade analizi gibi pek çok analizde hatalı yorumlamalara sebep olabilmektedir [78].

Bir derlemenin kalitesi birkaç açıdan değerlendirilebilir. Birincisi dizi uzunluğu ve parçalanmadır. Çok sayıda kısa kontig içeren bir derleme parçalanmış olarak kabul edilebilir. Bunun kötü dizileme veya yanlış birleştirme sonucu olması mümkündür. Dizi uzunluğu istatistiklerini (N50) hesaplaması için SeqKit [109] ve benzeri araçlar kullanılabilir. İkincisi ise okumaların kullanılma oranıdır. Kaliteli bir derleme, okumaların çoğunu kullanıldığı ve ayrıca çok fazla diziyle eşleşen okumaların az miktarda olduğu derlemelerdir. Bu ölçümlerin tümü okumaları birleştirilmiş dizilere göre hizalayarak kolayca kontrol edilebilir. N50 metriği, genom düzeneğindeki nükleotitlerin yarısının bu uzunluğa eşit veya daha uzun dizilerde olduğu dizi uzunluğunu tanımlayan basit bir ölçüdür [110]. Değer tüm genom elemanlarının tam birleşimini gösterir. Bu sebeple Trinity N50 yerine ExN50 kullanılmasını tavsiye eder. Böylelikle tüm genom elemanlarının değerlendirilmesi gibi sadece tam uzunluktaki dizilere odaklanmaz aksine kısa okumaların önemini ortaya koyar. ExN50 metriği, geleneksel N50'nin transkriptom düzeneklerini değerlendirmek için uygun hale getirilmiş halidir. Bu yaklaşım Trinity tarafından kullanılmaktadır [14].

De novo transkriptom düzeneklerinin kalitesini değerlendirmek amacıyla zamanla bazı araçlar geliştirilmiştir. Bu araçlar temelde bahsedilen metrikleri ölçer ve ek istatistikler barındırır. Bu konuda en popüler araç TransRate'dir [111]. Bunun dışında DETONATE [112], rnaQUAST'da [113] kullanılabilir.

2.2.1.4 Hizalama ve Bolluk Tahmini

Okumaların hizalanması ve transkript bolluğu tahmini geniş anlamıyla RNA-seq'in DE analizi için kullanılır. Okumaların haritalanması bolluk tahmini için bir ön koşuldur [114]. Bununla birlikte, derlemenin kalite kontrolü için hizalama metrikleri de kullanılabilir. İyi kalite *de novo* derlenmiş bir transkriptom okumaların büyük çoğunluğunun derlemeye eşleşmesi/hizalanması anlamına gelir. Yani okumaların pek çoğu birleştirme esnasında kullanılmış olacaktır. Genel bir kural olarak okumaların %70'den fazla oranda derlemeye katkısı ve haritalanmayan çok az okuma iyi bir derleme anlamına gelmektedir. Yine bazı okumalar birden fazla kontig ile de eşleşebilmektedir.

Bolluk tahmini, adından da anlaşılacağı gibi, derlemedeki transkriptlerin ifade düzeyinin anlaşılması sürecini ifade eder. Bolluklar tahmin edilmektedir çünkü çoklu eşleşen okumaların kaynağını belirlemek imkansızdır ve transkriptlerin gerçek ifade seviyeleri genellikle bilinmemektedir. Bu nedenle, çoğu yöntem tipik olarak transkript bollukları için olabilecek maksimum değerleri üretir. Bu değerler transkript bazında okuma desteğini ve milyonda transkript (TPM) gibi normalleştirilmiş bir ifade metriğini içerir. Bu değerler DE analizi için çok önemlidir, ancak derlemenin kalite kontrol amaçları için de kullanılabilir [78].

Okuma hizalaması ve bolluk tahmini genellikle birlikte yapılır. Bu prosedürde iki ana yaklaşım vardır. Birincisi okumaların ilk olarak Bowtie2 [115] veya STAR [116] gibi genel amaçlı bir hizalayıcı kullanarak derlenmiş kontiglere hizalandığı iki aşamalı bir işlemdir. Çıktısı dizileri ve bunlara hizalanmış okumaları listeleyen BAM [117] dosyasıdır. Bu dosya daha sonra bolluk tahmini yapmak için RSEM [118] ve benzeri araçta kullanılır. Buna alternatif diğer yol ise tek adımlı gerçek olmayan hizalama yapan yaklaşımdır. Bu yol benzer yaklaşımlara sahip Kallisto [119] ve Salmon [120] adında iki gerçek olmayan hizalama yapan araç kullanılarak uygulanabilmektedir.

Gerçek olmayan hizalama yapan araçlar nispeten yüksek doğrulukları ve kullanım kolaylıkları ve hızları sebebiyle tercih sebebi olmaktadır.

2.2.1.5 Derlemenin Küçültülmesi ve Gereksiz Tekrarların Azaltılması

De novo derleyiciler tipik olarak genomdaki genlerin sayısına bağlı olarak beklenenden çok fazla dizi üretir [121]. Derleme sonucunda elde edilen kontigler protein kodlayan transkriptlere ek olarak transkripsiyonel oluşumlar, pre-mRNA ve ncRNA'yı içerir [107]. Ayrıca fazla dizilerin başka bir kaynağı transkript izoformları olarak ortaya çıkan alternatif birleştirmelerdir [102], [103], [122]. Tüm bu dizileri saklamak her zaman gerekli olmayabilir. Bu nedenle derlemenin küçültülmesi kontrol edilebilir boyutlarda bir dizi seti elde etme yolunda önemli bir adım olabilir.

Küçültmeye yönelik ilk yaklaşım tüm derlemeye göre temsili olarak kabul edilebilecek kontigleri manuel olarak seçmektir. Fakat derlenen kontigler arasındaki ilişki önceden bilinmeden bunun yapılması mümkün değildir. Artık çoğu popüler derleyici transkriptleri otomatik olarak izoform gruplara ayırmaktadır.

Derlemenin küçültülmesi için ikinci yaklaşım, bir kümeleme aracı kullanmaktır. Bu özellikle derlenen kontiglerin belirsizliği giderilmiş gen-izoform ilişkisine sahip olmadığı veya derlemenin gerçekten fazla olduğu yani hemen hemen aynı diziye sahip birçok kontiğin bir araya getirildiği durumlarda yararlıdır. Kümeleme araçlar CD-HIT [123], [124] ve MMSeq2 [125] dizileri kümeler halinde gruplandırmak ve temsili dizileri ortaya çıkarmak için kullanılmaktadır.

Derlemenin küçültülmesi için diğer bir yaklaşım SuperTranscripts [126] aracı ile sunulmaktadır. Araç her gen kümesi için temsili bir izoform seçmek yerine izoformlardaki benzersiz eksonları tek bir doğrusal dizi haline getirir [14].

Derlemenin küçültülmesi işlemi gerekli görüldüğü hallerde yapılmalıdır. Derlemenin küçültülmesi sezgisel bir adımdır, *de novo* olarak bir araya getirilmiş kontiglerin kökenleri bilinmemektedir ve önemli biyolojik bilgiler taşıyabilmektedirler, derlemede temsil edilmesi gereken dizilerin kaybolabileceğinin göz önünde bulundurulması ve buna göre dikkatli olunması gereklidir.

2.2.1.6 Diferansiyel İfade Analizi

Fizyolojik veya çevresel koşullardaki değişikliklere yanıt olarak gen ifadesindeki değişiklikleri değerlendirmek, RNA-seq yaklaşımının ana hedeflerinden biridir. En basit durumda bu, deney ve kontrol koşullarına maruz bırakılan örneklerden RNA'nın yakalanmasıyla elde edilir. Daha sonra, dizilenen okumalar kontrol senaryosundan farklı olarak deneysel koşullar altında ne kadar farklı ifade edildiğini değerlendirmek için organizmanın genlerine haritalanır. Bu Diferansiyel İfade Analizi olarak ifade edilir [127]. Bu tür ifade karşılaştırmaları yoluyla çeşitli koşullar altında genlerin etkinliği hakkında bir bakış elde etmek mümkündür.

Bir DE analizini gerçekleştirmek için organizmanın gen dizileme veri koleksiyonu gereklidir. Açıklamalı bir genom mevcutsa genlerin kendileri kullanılabilir. Herhangi bir genom mevcut değilse, transkriptler genlerin temsil edildiği *de novo* transkriptom derlemesi ile elde edilen referans genom ile temsil edilebilirler. DE analizi için gerekli olan bir *de novo* transkriptom düzeneğinde, tüm koşullardan gelen okumalar ve tüm kopyalar derleme için bir araya toplanır: bu, okumaların daha sonra eşlenebileceği ve nicelleştirilebileceği tek, ortak bir “referans” transkriptom üretir [14].

Daha sonra, veriler DE'nin ortaya çıkarılabilmesi için analiz edilirler. Sonuç olarak ortaya çıkan tablo, her satırın benzersiz bir diziye temsil ettiği ve her sütunun benzersiz bir örneği veya kopyayı temsil ettiği bir veri setidir. Bu tablodaki her hücre, belirli bir örnek kopyada belirli bir diziye atanan okuma sayısını göstermektedir.

DE analizi yapabilen çeşitli programlama dillerinde bir dizi paket vardır. Uyguladıkları yöntemler farklılık gösterse de [114] hepsi benzer görevleri yerine getirirler: örnekler arasındaki sıralama derinliklerindeki farklılıkları hesaba katmak için okuma sayılarını normalleştirmek [128], isteğe bağlı olarak gereksiz bilgiyi azaltmak [129], veriye okuma sayıları dağılımı eklemek, her bir genin ilgilenilen koşullar arasındaki ifade farklarını test etmek ve çoklu testlerle üretilen P değerlerini düzeltmek.

Bir genin veya transkriptin diferansiyel olarak ifade edilip edilmediği bir dizi sayısal değerle belirlenir.

Düzeltilmiş P değerleri, iki koşul arasındaki genlerin veya transkriptlerin ifadesindeki farkın istatistiksel olarak anlamlı olup olmadığını gösterir. Küçük bir P değeri, iki koşul arasındaki okuma sayılarının farklılığının tamamen şansa bağlı olma olasılığının çok düşük olduğunu gösterir, yani farkın kaynağının biyolojik bir fenomen olması kuvvetle muhtemeldir.

log2FoldChange değeri, ifadedeki farkın büyüklüğünü tanımlar: iki koşuldan biri temel alınır ve diğerinin ifadesindeki değişiklik buna göre hesaplanır. Adından da anlaşılacağı gibi, bu, karşılaştırılan iki koşulun ortalama sayılarının oranının log2 değeridir. Pozitif bir log2FoldChange (lfc) değeri, karşılaştırma için temel olarak benimsenen koşula göre yukarı düzenlemeyi, negatif bir değer ise aşağı düzenlemeyi gösterir [78].

Günümüzde DE analizi için kullanılan en popüler araçların tümü R istatistiksel programlama dili ile kullanılmak üzere geliştirilmiştir. Bu amaçla en sık kullanılan programlar DESeq2 [130], edgeR [131] ve Limma [132], [133].

2.2.1.7 Dizilerin Translasyonu

RNA-seq verilerinin ileri analizlerinde protein kodlayan dizilerin (CDS) doğru karakterizasyonu, sadece bir transkriptomun protein kodlayan kısmının profilini çıkarmak için değil, aynı zamanda UTR'lerin ve gen düzenlemesi bağlamında ilgi çekici olabilecek kodlamayan dizi ve bölgelerin sınıflandırılması için de önemlidir [134].

Kodlama bölgelerini tahmin edebilen ve ardından bunları amino asit dizilerine çevirebilen birçok araç vardır. Bu araçların prensibi, tahminleri için nükleotid bileşiminin yanı sıra açık okuma çerçevelerinin (ORF) uzunluğunu da hesaba katan olasılıklı modellere dayanmaktadır. Bu araçlardan başlıcaları TransDecoder [14], Prodigal [135] ve CadAn'dır [134].

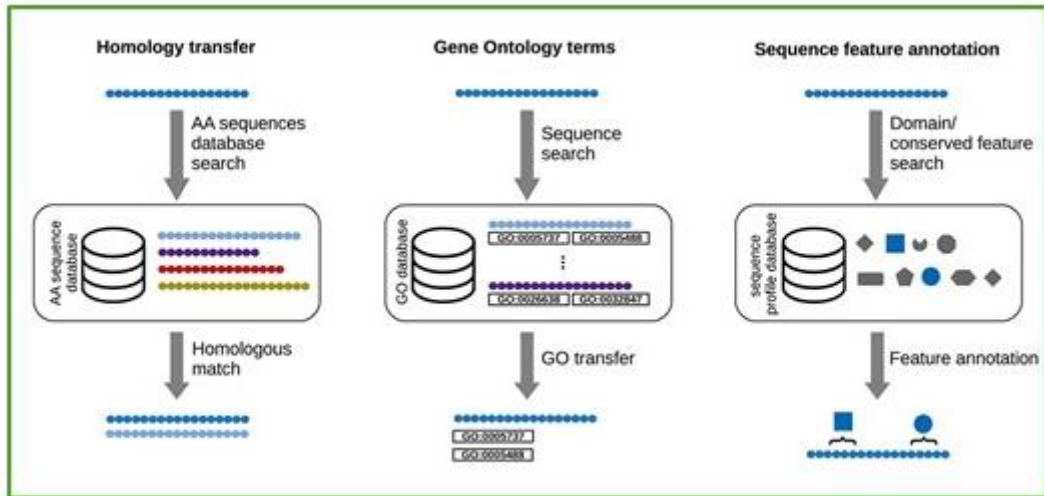
2.2.1.8 Transkriptoma İşlevsel Açıklama Eklenmesi

Bir transkriptom derlendikten ve kalite kontrolü yapıldıktan sonra, dizileri, verilerin elde edildiği koşullardaki tek tek ve toplu olarak temsil ettikleri işlevselliği aydınlatmak için incelenebilir. Örneğin derlenmiş transkriptomdaki aşırı temsil edilmiş bir transkript bir yapısal proteini kodlayabilir, bu da örneğin

alınması esnasında hücrede gerçekleşen yapısal bir modifikasyon aktivitesi varlığını ifade eder.

İşlevsel açıklama, in silico yöntemlerini kullanarak dizinin biyolojik işlevselliği ile ilgili bilgiler çıkarma ve atama işlemidir. İşlevsel açıklamanın genellikle mRNA'ların açıklamasına atıfta bulunduğu anlaşılır, çünkü bu dizilerin çevrildiği proteinler hücre içinde çeşitli aktiviteler yürütürler ve dolayısıyla hücrenin işleyişine katkıda bulunurlar. Bu nedenle işlevsel açıklama sürecinin RNA sınıflandırması ve amino asit dizisi tahmini ile başladığı söylenebilir.

Transkriptoma işlevsel açıklamalar eklenmesine standart bir yaklaşım getiren tanımlama henüz ortaya konulamamıştır. Fakat şu adımlar işlevsel açıklama prosedüründe kullanılmaktadır. Dizi arama yoluyla homoloji transferi ve kimlik ataması, dizi özelliği açıklama ve Gen ontolojisi (GO) ve biyokimyasal yol açıklama. Şekil 2.5'te gösterilen bu üç adımın tümü zorunlu değildir ve kesin çizgilerle ayrılmamıştır [78].



Şekil 2.5 Transkriptoma işlevsel açıklama eklenmesi adımları [78]

a- Homoloji Transferi Yoluyla Kimlik Ataması

Homoloji transferi, transkriptoma işlevsel açıklama eklemenin en temel biçimi olarak kabul edilebilir. Bu aşamada, bir dizi araması temelinde daha önce tanımlanmamış bir diziye tanımlayıcı bir kimlik ve işlevsel özellikler atanır. Bu yöntemle derlenmiş diziler arama araçlarına sorgu olarak girilir. İyi açıklamalı referans dizilerden oluşan bir veritabanında eşleştirilir [136]. Birden fazla eşleşen

sonuçtan e-değeri en düşük olan yanıt sorguya cevap olarak kaydedilmiş olur. E-değeri gözlemlenen dizi benzerliğinin tamamen şans eseri olarak ortaya çıkma olasılığının bir göstergesidir [137]. Başka bir deyişle yeterince düşük e-değeri korunmuş işlevi ima eden homolojinin göstergesidir ve bu sebeple bu adıma homoloji transferi ismi verilmektedir [137].

Homoloji transferi hem nükleotid dizileriyle hem de transkriptomlardan çevrilmiş protein dizileriyle gerçekleştirilebilir. Proteinler onların karşılığı olan mRNA dizilerinden daha fazla korunurlar. Protein dizi aramaları 21 amino asitlik genişletilmiş alfabesi ve nükleotid karşılıklarına kıyasla daha kısa dizi uzunlukları nedeniyle daha hassas ve hızlıdır [78].

Bu amaç için en sık kullanılan dizi arama aracı BLAST+ [138] paketidir. BLAST, farklı arama stratejileri için uzmanlaşmış birkaç alt araçtan oluşur. Blastn, nükleotid dizi hedeflerine karşı nükleotid sorgusu yapmak üzere; blastp, amino asit sorgularına karşılık amino asit hedeflidir. Paket ayrıca blastx ile nükleotid sorguları yapılarak amino asit hedeflerinin taranmasını sağlar. Yine paket yüksek hassasiyetli taramalar için iki seçenek sunmaktadır, deltablast, amino asit sorgularına karşılık amino asit veritabanlarında hassas aramalar yapar, Psiblast ise dizi profili aramalarını kullanarak amino asit sorgularını amino asit veritabanında tarayarak protein homologlarını tespit eder. Blast dizi aramalarının temel dayanağı olmasına rağmen çok yavaştır ve artan girdi boyutlarıyla hız açısından ölçeklenememektedir. *De novo* transkriptomlarına açıklama eklemek için 100 binden fazla sorgu yapmak gerekebileceğinden BLAST, özellikle büyük iş akışları için uygun olmayan bir seçenek halini almaktadır. BLAST'a alternatif olan hem hızlı hem de yüksek hassasiyette birkaç araç mevcuttur. Diamond [139], yalnızca protein veri tabanlarında arama yapmaya yönelik özel amaçlı bir araçtır. Diamond hem hızlı hem de BLAST komut satırını taklit etmesi sebebiyle tercih sebebidir. Diğer alternatif ise MMseqs2 [125], bazı yetenekleri ile BLAST'a eşdeğer bir alternatif olabilmektedir.

Homoloji aktarımı yoluyla açıklamanın kalitesi, kullanılan referans veritabanının kalitesine bağlıdır. Yakından ilişkili türlere ait referanslardan oluşan tek bir veritabanı potansiyel olarak daha az yanlış açıklamayla sonuçlanacak olsa da,

sınıflandırmadan bağımsız bir veritabanı gözden kaçabilecek yeni dizilere açıklama eklemeye değerli olabilmektedir.

Tamamı referans veritabanı olarak kullanılabilen veya manuel olarak seçilmiş bir referans dizisi seti için kaynak olarak kullanılabilecek birkaç genel amaçlı veritabanı mevcuttur. NCBI'nin NR (protein) ve NT (nükleotid) veritabanları [140] günümüzde aktif olan en büyük dizi veritabanlarıdır. Uniprot konsorsiyumunun Swiss-Prot veritabanı manuel olarak kontrol edilmiş yüksek kalitede protein dizileri barındırmaktadır. UniProt/TrEMBL veritabanı [141] daha fazla sayıda diziye sahip karşılığıdır.

b- Dizi Özelliklerinin Açıklamalara Eklenmesi

Açıklama eklenmesinin çok önemli bir yönü protein domainleri, düzensiz bölgeler, motifler, transmembran sarmallar ve benzeri işlevsel dizi özelliklerinin kesin olarak tanımlanmasıdır. Bunlara homoloji aktarımı yoluyla açıklama eklenebiliyorken bu süreç hem hataya açık hem de açıklamalar yeterli olmayabilir. Bu sebeple bu tür açıklamaların eklenebilmesi için bu özelliklerin varlığını açıkça tespit eden yaklaşımlar tercih edilmelidir.

Protein domainleri HMMER3 [142] adlı araçla tespit edilebilmekte, sinyal peptitleri SignalP [143] aracıyla tahmin edilmektedir, fLPS [144] ise protein motiflerini tahmin edebilmektedir.

InterProScan [145], birçok özellik tahmin metodunu, veritabanlarını ve analizleri tek bir kullanıcı dostu arayüzle bir araya getiren çok yönlü bir araçtır. Proteinlerin fonksiyonel ve yapısal alanlarını açıklamanın yanı sıra dizileri gen ontolojisi temelinde protein aileleri şeklinde sınıflandırmak ve transmembran bölgelerini tespit etmek için de kullanılır.

c- Gen Ontolojisi ve Yolak Açıklamasının Eklenmesi

Dizileri belirli biyolojik olaylarla tutarlı bir şekilde ilişkilendiren kontrol edilmiş bir sözlükten (ontoloji) tanımlayıcılar kullanmak oldukça yararlı olabilmektedir. Bu amaçla, derlenen diziler Gen Ontoloji (GO) terimleriyle açıklanabilir [146], [147]. GO terimleriyle dizilere açıklama eklemek için birkaç olasılık vardır.

İlk olarak, GO terimleri dizi arama yoluyla homolog dizilerden transfer edilebilirler. InterProScan [145] kullanılarak –goterm komut satırı girdisi kullanılarak terimlere açıklama eklenebilir. Buna alternatif açık kaynak bir yazılım olan eggNOG-mapper [148], [149] bağımsız bir GO açıklama aracı olarak kullanılabilir. Diğer bir alternatif, lisanslı bir araç olan BLAST2GO’nun da işlevsel açıklama paketi mevcuttur [150].

Yolak açıklamasının eklenmesi dizilerin bir veya daha fazla biyokimyasal yolak atanmasını ifade eder. İki popüler yolak açıklayan veritabanı vardır, The Kyoto Encyclopedia of Genes and Genomes (KEGG) [151], [152] ve Reactome [153]. Reactome daha insan merkezli olduğu için KEGG’i özellikle model olmayan organizmalar için daha yaygın kullanılan bir veritabanı haline getirmiştir. InterProScan, eggNOGmapper ve BLAST2GO bu veritabanından bilgiler çekebilmektedir, farklı bir araç kullanımına ihtiyaç duyulmamaktadır. Alternatif olarak açıklama ekleme paketi olan Trinotate Swiss-Prot ve BLAST sonuçları gibi pekçok GO açıklama eklenmesini destekler [14].

3.1 Materyal

RNA izolasyonu yapılacak olan *C. colurna* L. ve *C. avellana* L. Tombul bitki örneklerinin alınacağı ağaçlar için, tür tespitlerinin doğru yapıldığından ve geçmiş hikayelerinin bilindiği ve kontrollü bir şekilde büyütüldüğünden emin olunan Giresun Fındık Araştırma Enstitüsü uygulama bahçeleri tercih edilmiştir.

Örnek alma dönemi olarak fındık bitkilerinin gelişimi için aktif bir dönem olan Nisan-Mayıs ayları tercih edilmiştir. Seçilen ağaçlar meyve verme olgunluğuna erişmiş 6-10 yıllık ağaçlardır. Örnek alımı için, Tombul Fındık için 3 ağaç ve Türk Fındığı için 3 ağaç, toplam 6 ağaç kullanılmıştır. Her bir ağacın yaprak, erkek ve dişi çiçek, tomurcuk ve zuruf örnekleri ayrı ayrı toplanmış ve sıvı azotta korunarak laboratuvara taşınmış, laboratuvarında ise -80 °C'de muhafaza edilmiştir.

3.2 Yöntem

3.2.1 Total RNA İzolasyonu

Total RNA izolasyonları RNeasy Plant Mini Kit (Qiagen) kullanılarak her bir fındık ağacından temin edilen 4 dokudan ayrı ayrı yapılmıştır, ardından DNA kontaminasyonunu önlemek için RNase-free DNase I (Qiagen) kullanılmıştır.

Total RNA eldesi için şu adımlar izlenmiştir:

1. -80 °C'de muhafaza edilen 6 ağaca ait toplam 24 farklı örnek laboratuvar koşullarında sıvı azot ve havan kullanılarak toz haline getirilmiştir.
2. Her bir izolasyon için kullanılmak üzere Buffer RLC hazırlanmıştır.
3. Her bir izolasyon için kullanılmak üzere Buffer RPE hazırlanmıştır.
4. Tüplere 100'er mg örnek dokulardan konulmuş, üzerlerine 450 µg Buffer RLC eklenerek ve güçlü bir şekilde vortekslenmiştir.
5. Ardından lizat 2 ml'lik lila kolonlu koleksiyon tüpüne aktarılmış, 2 dk en yüksek devirde santrifüj edilmiştir.

6. Süpernatantlar alınıp yeni tüpe aktarılmıştır. Aktarılan kısımlar ölçülmüş, değerin yarısı kadar etanol konulmuştur.
7. Örnekler 2 ml'lik pembe kolonlu koleksiyon tüpüne aktarılmış ve 15 sn 10.000 rpm'de santrifüj edilmiştir.
8. Alt kısım atıldıktan sonra, spin kolona sırasıyla 700 µl Buffer RW1 eklenip 15 sn 10.000 rpm'de santrifüj edilip alt taraf atılarak, 500 µl Buffer RPE eklenip 15 sn 10.000 rpm'de santrifüj edilip alt taraf atılarak ve son olarak 500 Buffer µl RPE eklendi 2dk 10.000 rpm'de santrifüj edilerek işlemler tamamlanmıştır.
9. Koleksiyon tüpü boşaltılıp tekrar 1 dk en yüksek hızda santrifüj edilen kolonlar yeni koleksiyon tüplerine taşınmış ve 30 µl RNase-free water konulmuş, tekrar 1 dk 10.000 rpm'de santrifüjlenmiştir. Bu iki defa tekrarlanarak işlem sonlandırılmıştır.

RNeasy Plant Mini Kit (Qiagen) sonrası elde edilen total RNA'ların RNA konsantrasyonunu artırmak ve pürifiye etmek için RNA MinElute Kit (Qiagen) kullanılmıştır.

RNA MinElute Kit (Qiagen) ile şu işlemler yapılmıştır;

1. Örneklerin hacimlerinin µl'ye ayarlanması için nükleazsız su kullanılmıştır. 350 µl Buffer RTL eklenerek karıştırılmıştır.
2. 250 µl %100 etanol eklenip karıştırılarak RNA'nın dilüe olması sağlanmıştır.
3. Örnekler 2 ml'lik özel tüpteki RNeasy MinElute spin kolonlarına aktarılmış ve 15 sn 10.000 rpm'de santrifüjlenmiştir. Sonrasında filtreden geçen kısım atılmıştır.
4. Spin kolonlar yeni 2 ml'lik özel tüplere yerleştirilmiştir. 500 µl Buffer RPE eklenerek 2 dk 10.000 rpm'de santrifüjlenmiştir.
5. RNeasy MinElute spin kolonları yeni 2 ml'lik özel tüplere yerleştirilerek membranın kuruması için 5dk en yüksek hızda santrifüjlenmiştir ve filtreden geçen kısım atılmıştır.

6. Spin kolonlar yeni 1.5 ml'lik tüplere alınmış ve 14 µl nükleazsız su doğrudan spin kolona uygulamıştır. RNA'nın elüe olması için en yüksek hızda 1 dk santrifüjlenmiştir.

Türk Fındığının 3 ağacından elde edilen 12 farklı total RNA'dan RNA havuzu oluşturulmuştur. Bu işlem Tombul Fındık örnekleri için de tekrarlanmıştır.

3.2.2 RNA'ların Kalite Kontrolü

Elde edilen RNA'ların kalitesi Nanodrop 2000 Spectrophotometer (Thermo scientific) ile NanoDrop ölçümü ve MOPS jel elektroforez yöntemi ile teyit edilmiştir. NanoDrop spektrometresinde 230, 260 ve 280 dalga boylarında ölçüm yapılmıştır.

Formaldehit jel elektroforezinde yürütme tamponu olarak 1X MOPS buffer kullanılmıştır. İzole edilen RNA'ların bütünlüğü anlamak için, Formaldehit jele 1,25 µL örnek yüklenmiş ve 60 volt enerjiyle 1,5 saat yürütüldükten sonra görülmüştür.

İntakt RNA'ların kalitesi, YND'nin yapıldığı Beijing Genomics Institute (BGI)'de Agilent 2100 Bioanalyzer yardımıyla tekrar analiz edilip dizileme için uygunluğu onaylanmıştır.

3.2.3 RNA Dizileme

YND yapılmak üzere iki fındık türüne ait izole edilip kalite kontrolleri yapılan total RNA'lar uygun konsantrasyonlarda BGI firmasına gönderilmiştir.

BGI, Agilent 2100 ve NanoDrop yöntemi ile tekrar miktar ve kalite tayini yapmış ve sekanslama için kaliteli RNA elde edildiğinden cDNA kütüphanesi kurulmuştur.

cDNA kütüphaneleri Illumina HiSeq 4000 Dizileme Sistemi (Illumina, Inc., ABD) kullanılarak dizilenmiştir. Dizileme sonrası tüm ham veriler filtrelenmiştir. Bu filtreleme işlemi adaptörlerin, düşük kaliteli verilerin ve kontaminasyonların uzaklaştırılması şeklinde gerçekleştirilmiştir.

3.2.4 RNA-seq Analizleri

BGI tarafından yapılan dizileme sonucu elde edilen RNA-seq verileri, adaptör kalıntıları ve düşük kaliteli dizilerden arındırılmış şekilde temiz veri olarak

gönderilmiştir. *C. colurna* L. ve *C. avellana* L. fındık türüne ait iki RNA-seq verisinin biyoinformatik analizlerine; FastQC ile dizilemenin kalite kontrolünün yapılması, Trinity ile derleme ve GO analizleri şeklinde devam edilmiştir.

3.2.4.1 Hazırlık

Biyoinformatik analizler için işletim sistemi olarak temelde Linux Ubuntu versiyon 21.04 kullanılmıştır. Bazı araçların kullanımında Windows 10 ve MacOS'da kullanılmıştır. Aşağıda analizler için kullanılan bazı temel programların versiyon bilgileri yer almaktadır.

FASTQC versiyon 0.11.9

Trinity versiyon 2.9.1

BLAST+ versiyon 2.9.0

MISA versiyon 2.1

3.2.4.2 Dizilemenin Kalite Kontrolü

Dizilemenin kalite kontrolünün yapılması için FastQC programı kullanılmıştır. FastQC, Babraham Enstitüsü tarafından geliştirilen yüksek çıktılı dizileme verilerinin kalite kontrolünün yapılması için hızlı sonuç veren bir çözümdür.

FastQC aracılığıyla;

1. Dosya ismi, türü, dizi sayısı, dizilerin uzunluğu ve dizideki GC bazlarının oranı gibi temel istatistikler,
2. Dizilerdeki bazların kalite puanları,
3. Dizilerin ortalama kalite puanları,
4. Dizilerin baz içeriği oranları ve ortalama GC oranları,
5. Dizilerdeki tahmin edilemeyen bazların (N) pozisyonları,
6. Dizilerin uzunlukları,
7. Dizilerdeki duplikasyon miktarı,
8. Aşırı tekrarlanan dizilerin baz dizilimleri ve tekrar miktarları
9. Adaptör kalıntıları ve k-mer içerikleri,

tespit edilmektedir.

3.2.4.3 Trinity Kullanılarak RNA-seq'in *de novo* Derlemesi

Dizileme ve temizleme işlemi sonrasında elde edilen temiz okumaların derlemesinin gerçekleştirilmesi için Trinity programı [14] kullanılmıştır.

Trinity RNA-seq transkript verilerinin verimli ve sağlam bir şekilde *de novo* derlenmesini sağlayan bir programdır. Büyük hacimli RNA-seq okumalarını işlemek için üç bağımsız modülü biraraya getirir; Inchworm, Chrysalis ve Butterfly.

Inchworm, RNA-seq verilerini benzersiz transkript dizileri olarak birleştirir. Tam uzunlukta transkriptler üretir ve bunların benzersiz kısımlarını rapor eder.

Chrysalis, Inchworm kontiglerini kümeler haline getirir ve her küme için Bruijn grafikleri oluşturur.

Butterfly, Oluşturulan grafikleri paralel olarak işler, alternatif izoformları ve tam uzunluktaki transkriptleri rapor eder ve paralog genlere karşılık gelen transkriptleri ayırır.

Trinity ile iki RNA-seq verisi tek bir komut ile çalıştırılmış oluşturulan derleme ile diferansiyel genlerin kıyaslanması ve ileri analizler mümkün hale gelmiştir.

Örnek kod aşağıdaki gibidir.

```
$ PATH/Trinity --seqType fq --max_memory 64G \
--left A_1.fastq,C_1.fastq --right A_2.fastq,C_2.fastq \
--CPU 16 --output trinity_outdir
```

Çıktıları incelemek için aşağıdaki komut kullanılmıştır.

```
$ head trinity_outdir/Trinity.fasta
```

3.2.4.4 Transkript İfade Seviyelerinin Belirlenmesi

Trinity tarafından yeniden oluşturulan transkriptlerin ifade seviyelerini tahmin etmek için ilk olarak orjinal okumaları Trinity transkriptlerine karşı hizalama için Bowtie çalıştırılmış, ardından her bir kontig ile eşleşen okuma sayısını tahmin etmek için RSEM kullanılmıştır. Bu komut her fındık türü için ayrı ayrı çalıştırılmıştır.


```
$ PATH /util/align_and_estimate_abundance.pl --seqType fq \
--left A_1.fastq --right A_2.fastq --transcripts Trinity.fasta
output_prefix avellana --est_method RSEM --aln_method bowtie --
trinity_mode
--output_dir avellana_rsem --prep_reference
```

İzoform seviyesinde çıktıları incelemek için aşağıdaki komut kullanılmıştır.

```
$ head avellana_rsem/avellana.isoforms.results / column -t
```

Gen seviyesinde çıktıları incelemek için aşağıdaki komut kullanılmıştır.

```
$ head avellana_rsem/avellana.genes.results / column -t
```

3.2.4.5 Derleme Kalitesinin Değerlendirilmesi

Trinity derleme kalitesinin incelenmesi, Trinity özet raporuna bakılması, N50, okumaların ne kadar temsil edildiği ve exN50 istatistiklerinin değerlendirilmesi şeklinde farklı yollarla gerçekleştirilmiştir.

Derlenmiş transkript sayısı için aşağıdaki komut kullanılmıştır.

```
$ grep '>' Trinity.fasta | wc -l
```

N50 değeri, derlenen bazların en az yarısının verilen değerdeki baz sayısı kadar kontig uzunluğunda olduğunu ifade eder. Temel istatistikler içerisinde N50, N40, N30 gibi değerlere ulaşmak mümkündür.

Trinity derlemesi hakkında temel istatistikler aşağıdaki kodla incelenmiştir.

```
$ PATH /util/TrinityStats.pl Trinity.fasta
```

Genel olarak yüksek kaliteli bir derlemede, okumaların en az %70'sinin uygun çiftler olarak var olması beklenir. Okumaların temsiline tespiti için şu adımlar izlenmiştir.

Bowtie2 dizini oluşturulması için aşağıdaki komut kullanılmıştır.

```
$ bowtie2-build Trinity.fasta Trinity.fasta
```

Okuma adına göre sıralanmış BAM dosyası elde etmek için aşağıdaki komut kullanılmıştır.

```
$ bowtie2 --local --no-unal -x Trinity.fasta -q -1 A_1.fastq -2 A_2.fastq \
/ samtools view -Sb - / samtools sort -no - - > bowtie2.nameSorted.bam

$ bowtie2 --local --no-unal -x Trinity.fasta -q -1 C_1.fastq -2 C_2.fastq \
/ samtools view -Sb - / samtools sort -no - - > bowtie2.nameSorted.bam
```

ExN50 değeri N50 kontig istatistiğini göstermektedir ancak en yüksek düzeyde ifade edilen transkriptlerle sınırlıdır.

ExN50 değerinin hesaplanması için aşağıdaki komut kullanılmıştır.

```
$ PATH /util/misc/contig_ExN50_statistic.pl
Trinity_trans.TMM.EXPR.matrix Trinity.fasta > ExN50.stats
```

ExN50 değerlerini görmek ve grafik oluşturmak için aşağıdaki komutlar kullanılmıştır.

```
$ cat ExN50.stats | column -t

$ PATH /util/misc/plot_ExN50_statistic.Rscript ExN50.stats
```

Tam uzunluktaki kodlayan transkriptlerin sayılarını değerlendirmek için BLASTX komutuyla SwissProt kullanılmıştır.

En iyi eşleşen proteinleri tespit etmek üzere (E Value: 1e-20) *blastx.outfmt6* dosyası oluşturmak için aşağıdaki kod kullanılmıştır.

```
$ blastx -query Trinity.fasta \
-db sprot.pep -out blastx.outfmt6 \
-evalue 1e-20 -num_threads 2 -max_target_seqs 1 -outfmt 6
```

En iyi SwissProt eşleşmelerin uzunlukları görmek için aşağıdaki kod kullanılmıştır.

```
$ PATH /util/analyze_blastPlus_topHit_coverage.pl \
blastx.outfmt6 Trinity.fasta sprot.pep | column -t
```

3.2.4.6 Örnek Veri Kalitesinin Değerlendirilmesi

DE analizine geçmeden önce veriler incelenmelidir. Bu inceleme örneklerin aşırı benzer olup olmadığına bakılması ve data kirliliğinin varlığının incelenmesidir.

Örnekleri karşılaştırmak için aşağıdaki kodlar ayrı ayrı çalıştırılmıştır.

```

$ PATH /Analysis/DifferentialExpression/PtR \
-m Trinity_trans.counts.matrix -s samples.txt --log2 --compare_samples
$ PATH /Analysis/DifferentialExpression/PtR \
-m Trinity_trans.counts.matrix -s samples.txt --log2 --sample_cor_matrix
$ PATH /Analysis/DifferentialExpression/PtR \
-m Trinity_trans.counts.matrix -s samples.txt --log2 --CPM --prin_comp 3

```

3.2.4.7 Diferansiyel İfade Analizi

Diferansiyel olarak ifade edilen transkriptleri tanımlamak için pek çok araç mevcuttur, Bu aşamada en popüler uygulamalardan birisi olan edgeR kullanılmıştır. EdgeR, R içerisindeki Bioconductor paketinin bir parçasıdır ve kompleks istatistiksel analiz yöntemler ile çalışmaktadır.

edgeR'ı çalıştırmak için aşağıdaki örnek kodlar kullanılmıştır.

```

$ PATH /Analysis/DifferentialExpression/run_DE_analysis.pl \
--matrix Trinity_trans.counts.matrix --samples_file samples.txt \
--method edgeR --output edgeR_trans

```

Çıktı olarak verilen “*.DE_results” dosyaları örneklerin ikili karşılaştırılmalarının sonucunda diferansiyel olarak ifade edilen transkriptleri tanımlamak için oluşturulmuş dosyalardır.

“*.DE_results” dosyalarını incelemek için aşağıdaki kod kullanılmıştır.

```

$ head
edgeR_trans/Trinity_trans.counts.matrix._A_vs_C_.edgeR.DE_results \
column -t

```

İkili karşılaştırmada, P değeri 0,001 ve kat değişikliği >4 olacak şekilde ifade edilen transkriptlerin çıkarılması ve ısı haritasının elde edilmesi için aşağıdaki kodlar kullanılmıştır.

```

$ PATH /Analysis/DifferentialExpression/analyze_diff_expr.pl \
--matrix ../Trinity_trans.TMM.EXPR.matrix --samples ../samples.txt -P 1e-
3 -C 2

```

```
$ wc -l diffExpr.P1e-3_C2.matrix

$ PATH /Analysis/DifferentialExpression/PtR \

-m diffExpr.P1e-3_C2.matrix.log2.centered.dat -s ../samples.txt \

--gene_dist euclidean --sample_dist euclidean

--heatmap --heatmap_scale_limits "-4,4"
```

Diferansiyel olarak ifade edilen genleri hesaplamak için aşağıdaki kodlar çalıştırılmıştır.

```
$ PATH /Analysis/DifferentialExpression/run_DE_analysis.pl \

--matrix Trinity_genes.counts.matrix --samples_file samples.txt \

--method edgeR --output edgeR_genes

$ PATH /Analysis/DifferentialExpression/analyze_diff_expr.pl \

--matrix ../Trinity_genes.TMM.EXPR.matrix \

--samples ../samples.txt -P 1e-3 -C 2
```

3.2.4.8 İşlevsel Açıklama Eklenmesi

Diferansiyel olarak ifade edilen transkriptlerin hangileri olduklarını veya hangi biyolojik işlevleri temsil ettiğini anlamak için bir dizi programı kullanılmış bu sayede bu transkriptlere işlevsel açıklamalar ekleyerek anlamlı hale getirilmişlerdir.

Transdecoder, transkript dizileri içerisindeki olası kodlayan bölgelerin tespiti için kullanılmaktadır. Transkriptlerdeki uzun açık okuma çerçevelerini (ORF) tanımlar ve bunları dizilerin kompozisyonlarına göre puanlar.

Transdecoder iki adımlı bir işlemdir öncelikle tüm uzun ORF'ler tanımlanır, sonrasında ise hangi ORF'lerin kodlanacağını tahmin eden adım çalıştırılır. İşlem sonunda transkriptler içindeki tahmin edilen bölgelere karşılık gelen protein dizilerini içeren "Trinity.fasta.transdecoder.pep" dosyası oluşturulmuş olur.

Kod dizilimi aşağıdaki gibidir.

```
$ TransDecoder/TransDecoder.LongOrfs -t Trinity.fasta
```

```
$ TransDecoder/TransDecoder.Predict -t Trinity.fasta
```

```
$ less Trinity.fasta.transdecoder.pep
```

Olası homolog bilgilerini yakalamak için e-değeri 1e-5 olacak şekilde BLASTX çalıştırılmıştır. Sonrasında sadece tahmin edilen protein dizilerinin araştırılması için BLASTP ile dizi homolojilerine bakılmıştır.

Kod dizilimi aşağıdaki gibidir.

```
$ blastx -db sprot.pep -query Trinity.fasta -num_threads 2 \
```

```
-max_target_seqs 1 -outfmt 6 -evalue 1e-5 > swissprot.blastx.outfmt6
```

```
$ blastp -query Trinity.fasta.transdecoder.pep -db sprot.pep -num_threads 2 \
```

```
-max_target_seqs 1 -outfmt 6 -evalue 1e-5 > swissprot.blastp.outfmt6
```

Tahmin edilen protein dizileri kullanılarak Pfam veritabanına karşı HMMER araştırması yapılmıştır ve korunmuş alanlar tespit edilmeye çalışılmıştır.

```
$ hmmscan --cpu 16 --domtblout TrinotatePFAM.out Pfam-A.hmm \  
Trinity.fasta.transdecoder.pep > pfam.log
```

SignalP aracı sinyal peptidlerini ve tmhmm aracı ise transmembrane domainleri tespit etmek için kullanılmıştır.

Kod dizilimleri aşağıdaki gibidir.

```
$ signalp -f short -n signalp.out Trinity.fasta.transdecoder.pep > sigP.log
```

```
$ less signalp.out
```

```
$ tmhmm --short < Trinity.fasta.transdecoder.pep > tmhmm.out
```

```
$ less tmhmm.out
```

3.2.4.9 Veritabanı Oluşturulması

Elde edilen verilere hızlı bir erişim sağlamak için bir veritabanı oluşturmak bir gerekliliktir. Bunun için içerisine SWISSPROT kayıtlarının, Pfam domainleri hakkında genel verinin olduğu Trinotate veritabanı oluşturulmuştur. Bu

veritabanına çalışmadaki sonuçlar dahil edilerek çalışmaya özel bir veritabanı elde edilmiştir.

Aşağıdaki örnek kodlar sırasıyla çalıştırılarak veritabanı elde edilmiştir.

```
$ Trinotate Trinotate.sqlite init
--gene_trans_map ../Trinity.fasta.gene_trans_map \
--transcript_fasta ../Trinity.fasta \
--transdecoder_pep Trinity.fasta.transdecoder.pep
$ Trinotate Trinotate.sqlite LOAD_swissprot_blastx
swissprot.blastx.outfmt6
$ Trinotate Trinotate.sqlite LOAD_swissprot_blastp
swissprot.blastp.outfmt6
$ Trinotate Trinotate.sqlite LOAD_pfam TrinotatePFAM.out
$ Trinotate Trinotate.sqlite LOAD_tmhmm tmhmm.out
$ Trinotate Trinotate.sqlite LOAD_signalp signalp.out
```

Elde edilen veritabanının tüm verilerine ulaşmak için bir rapor oluşturulmuştur.

```
$ Trinotate Trinotate.sqlite report > Trinotate.xls
```

İçerisine DE verileri de dahil edilerek geniş kapsamlı bir veritabanına dönüştürülmüştür.

```
$ Trinotate/util/transcript_expression/import_expression_and_DE_results.pl \
--sqlite Trinotate.sqlite --transcript_mode --samples_file ../samples.txt \
--count_matrix ../Trinity_trans.counts.matrix \
--fpkm_matrix ../Trinity_trans.TMM.EXPR.matrix
$ Trinotate/util/transcript_expression/import_expression_and_DE_results.pl \
--sqlite Trinotate.sqlite --gene_mode --samples_file ../samples.txt \
--count_matrix ../Trinity_genes.counts.matrix \
--fpkm_matrix ../Trinity_genes.TMM.EXPR.matrix
```

```
$ Trinotate/util/transcript_expression/import_expression_and_DE_results.pl \
--sqlite Trinotate.sqlite --transcript_mode --samples_file ../samples.txt \
--DE_dir /edgeR_trans

$ Trinotate/util/transcript_expression/import_expression_and_DE_results.pl \
--sqlite Trinotate.sqlite --gene_mode --samples_file ../samples.txt \
--DE_dir /edgeR_genes
```

3.2.4.10 Gen Ontoloji ve KEGG Yolak Analizleri

GO, genlerin fonksiyonlarına göre sınıflandırıldığı, standart numaralandırmaya sahip bir sistemdir. Bu sistem genleri 3 ontoloji grubunda sınıflandırır. Moleküler Fonksiyon (MF), Biyolojik Proses (BP) ve Hücresel Komponent (CC). MF kategorisi, moleküler seviyedeki aktivitelerin sınıflandırıldığı, BP kategorisi, moleküler faaliyetlerin biyolojik hedeflerini kategorize etmektedir. CC kategorisi ise, hücrealtı yapılar ve makromoleküler bileşenlerin hücre içi lokasyonlarını ve durumlarını ifade etmek için kullanılmaktadır.

KEGG yolak analizi, metabolizma, hücresel işlemler, organizmalardaki sistemler, hastalıklar ve ilaç geliştirme amaçlı moleküler etkileşimler, reaksiyonlar ve ilişki ağları hakkında bilgi veren bir veritabanıdır.

Türk Fındığı ve Tömbül Fındık transkriptom verilerinin GO ve KEGG yolak analizleri için TRAPID online tool kullanılmış ve EggNOG 4.5.1 veritabanı ile tarama gerçekleştirilmiştir. Aynı zamanda Trinotate ve AmiGO da diferansiyel olarak ifade edilen genlerin değerlendirilmesinde kullanılmıştır.

3.2.4.11 SSR Varlığının Araştırılması

Mikrosatellitler bitki genetiği ve adli tıpta sıklıkla kullanılan bir işaretleme sistemidir. SSR'ların tanımlanması için MISA yazılımı kullanılmış, *C. colurna* L. ve *C. avellana* L. Tömbül derlemeleri ayrı ayrı çalıştırılmıştır.

Aşağıdaki kod örneği kullanılmıştır.

```
$ perl misa.pl colurna.fasta
$ perl misa.pl avellana.fasta
```

3.2.5 RNA-seq Verilerinin Gerçek Zamanlı Kantitatif PCR (qPCR) ile Validasyonu

C. colurna L. ve *C. avellana* L. Tombul RNA-seq verilerinin içerisinde bitki büyümesi, direnç gelişimi, lipid metabolizması ve savunma geliştirme ile ilgili TL1 (Thaumatococcus-like protein 1), GMPM1 (18 kDa tohum olgunlaşma proteini), N (direnç proteini N), 2MMP (Metalloproteinaz 2-MMP), At1g29670 (GDSL esteraz/lipaz) ve CHIB1 (Asidik endokitinaz) genlerinin gerçek zamanlı kantitatif PCR (qPCR) ile validasyonu yapılmıştır.

Bitki örneklerinden Plant Mini Kit (Qiagen, USA) kullanılarak üretici talimatlarına göre total RNA elde edilmiş, First stranded cDNA sentez kiti (SAMscript, GeneMark, Atlanta) kullanılarak cDNA'nın birinci ipliği sentezlenmiştir. 6 gen için Primer3 aracıyla varsayılan parametreler kullanılarak primerler sentezlenmiş ve benzersizlikleri BLAST aracıyla nr (NCBI) veritabanında ve erime eğrisi analizine göre kontrol edilmiştir. Primer dizilerinin listesi Tablo 3.1'de verilmiştir.

Tablo 3.1 qPCR için kullanılan primer dizileri

Gen	Forward/Reverse	Sequence (5' - 3')
TL1	Forward	ACTCTTCGGCCTTGTCATGG
	Reverse	GGGTTTCTGGTCCGCAGTTA
GMPM1	Forward	CATATGCAGGGAGGAAAGAA
	Reverse	CTCGAGTTAAGTACCCCCAG
N	Forward	TTATGCGAAGGACCCACACC
	Reverse	CGACTCTGCTGACTGAACCA
2MMP	Forward	TACTACATATTCCCGCCCGT
	Reverse	TTCCCTAGCATCTCTCCAGG
At1g29670	Forward	AGGTGGGCTGGACTCAATTC
	Reverse	AATGGCTCTGCATGTGTGGA
CHIB1	Forward	TTGAGAGCTTCCTTGCCACC
	Reverse	TACTGTGACCGTCCGTTTGG

Gerçek zamanlı kantitatif PCR (qPCR) Roche LightCycler® 480 Real-Time PCR System (Roche, USA) aracılığıyla gerçekleştirilmiştir. İki fındık türü arasındaki gen ifadesi değişiminin tahmini için $\Delta\Delta C_t$ karşılaştırmalı yöntem [154] kullanılmıştır.

3.2.6 RNA-seq Verilerinin NCBI Veritabanına Yüklmesi

NCBI Dizi Okuma Arşivi (SRA) veritabanı, yüksek verimli dizileme verilerinin halka açık bir şekilde tutulduğu bir arşivdir. Pek çok çalışma alanından verileri kabul etmektedir. SRA arşivi tekrarlanabilirliği artırmak ve veri analizi yoluyla yeni keşifleri kolaylaştırmak için dizileme verilerini ve hizalama bilgilerini depolamaktadır. Çalışmasının sonunda RNA-seq verileri NCBI-SRA veritabanına yüklenmiştir.



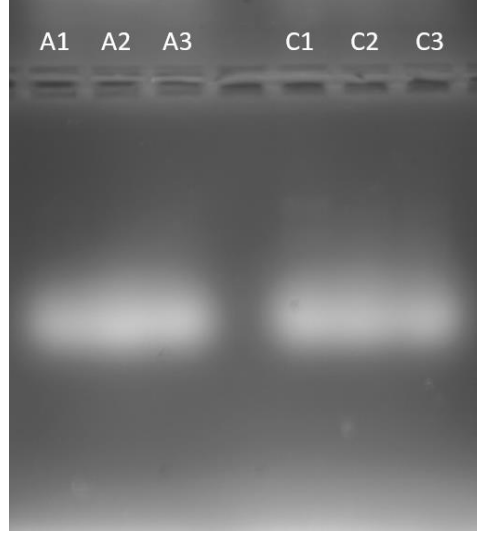
4.1 Total RNA İzolasyon Sonuçları ve Karakterizasyonu

Total RNA izolasyonları RNeasy Plant Mini Kit (Qiagen) kullanılarak her bir fındık ağacından temin edilen 4 dokudan ayrı ayrı ile yapılmış, sonrasında DNA kontaminasyonunu önlemek için RNase-free DNase I (Qiagen) ve RNA konsantrasyonunu artırmak ve pürifiye etmek için RNA MinElute Kit (Qiagen) kullanılmıştır. Total RNA'lar elde edildikten sonra spektrofotometrik yöntemle (Nanodrop) miktarı belirlenmiştir (Tablo 4.1). Buna ek olarak MOPS jel elektroforez yöntemi ile de doğrulama yapılmıştır (Şekil 4.1).

Türk Fındığının 3 ağacından elde edilen 12 farklı total RNA'dan RNA havuzu oluşturulmuştur. Bu işlem Tömbül Fındık örnekleri için de tekrarlanmıştır.

Tablo 4.1 Nanodrop spektrofotometre ölçüm sonuçları

Örnek	Nükleik asit konsantrasyonu (ng/μL)	OD 260/280	OD 260/230
A1	43,11	1,76	0,18
A2	63,16	1,58	0,27
A3	63,34	1,95	0,41
C1	44,96	1,72	0,40
C2	24,43	2,00	0,12
C3	57,01	1,87	0,33



Şekil 4.1 MOPS jel elektroforez yöntemi ile doğrulama

4.2 RNA-seq Sonuçları

4.2.1 Dizilemenin Değerlendirilmesi

İki fındık türünün yaprak, erkek ve dişi çiçek, tomurcuk ve zuruf örneklerinden elde edilen iki total RNA örneği BGI tarafından Illumina HiSeq 4000™ Sequencing System kullanılarak dizilenmiştir. YND sonrası BGI tarafından ilk temizleme ve filtreleme işlemi gerçekleştirilmiş ve temiz data gönderilmiştir.

Tablo 4.2 *C. colurna* L. (Türk Fındığı) ve *C. avellana* L. Tombul Fındık örneklerinin dizileme sonrası RNA dizileri istatistiksel sonuçları

Örnek İsmi	Temiz Okumalar	Temiz baz sayısı	Okuma Uzunluğu(bp)	Q20(%)	GC(%)
<i>C. colurna</i>	41,140,316	4,114,031,600	100	98.05%	47.42%
<i>C. avellana</i>	41,854,426	4,185,442,600	100	98.13%	47.87%

BGI tarafından gönderilen sıkıştırılmış dosya içeriğinde iki türe ait çift uçlu (pair-end) okumaların ikişer dosyası yer almıştır. Dosyalar FASTQ formatındadır. Adaptör kalıntılarından temizlenen ve kalitesiz dizilerin filtrelendiği temiz verinin toplam okuma sayısı 83 milyondur. Dizilemedeki her bir okumaların uzunluğu 100 bp'dir. Temiz veriye ait istatistikler Tablo 4.2'de yer almaktadır.

4.2.2 RNA-seq Verilerinin Kalite Kontrolleri

Dizileme sonucu BGİ'dan gelen verilerin kalite kontrol işlemleri FastQC adlı programla yapılmıştır. Temiz dataların kalite kontrolünden sonra *de novo* transkriptom analizine devam edilmiştir.

FastQC programı her bir örnek dosya isminin, dosya tipinin, dizileme sisteminin, dizi sayısının, dizi uzunluklarının ve dizideki GC baz oranının yer aldığı temel istatistiklerin verildiği bir sonuç (Şekil 4.2) elde edilmiştir.

A) *C. colurna* L.

B) *C. avellana* L. Tombul

✓ Basic Statistics

Measure	Value
Filename	C-213_1.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	20570158
Sequences flagged as poor quality	0
Sequence length	100
%GC	47

✓ Basic Statistics

Measure	Value
Filename	A-214_1.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	20927213
Sequences flagged as poor quality	0
Sequence length	100
%GC	47

Şekil 4.2 FastQC programından elde edilen temel istatistikler A) *C. colurna* L.

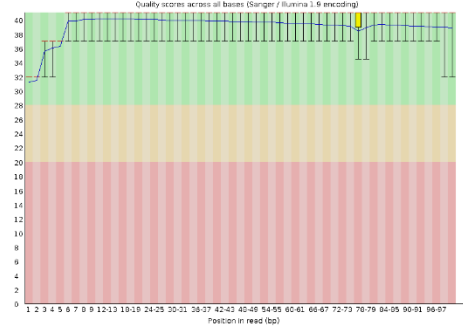
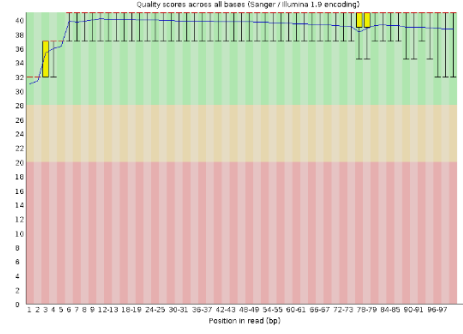
B) *C. avellana* L. Tombul

Ayrıca Şekil 4.3'de Türk Fındığı ve Tombul Fındık RNA-seq verilerinin adaptör ve düşük kaliteli dizilerden arındırılmış hali olan temiz verinin ayrıntılı FastQC sonuçları verilmiştir.

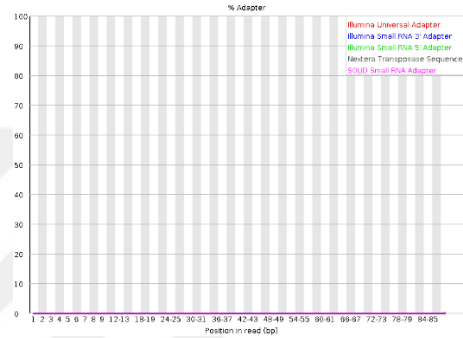
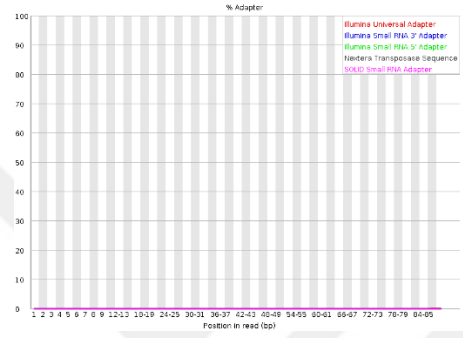
A) *C. colurna* L.

B) *C. avellana* L. Tombul

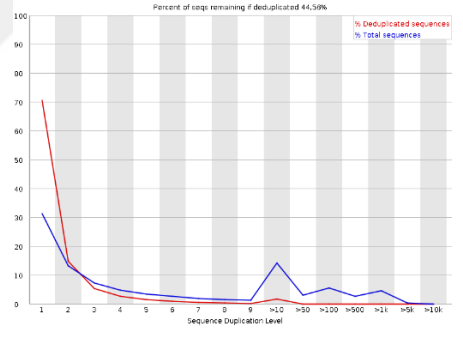
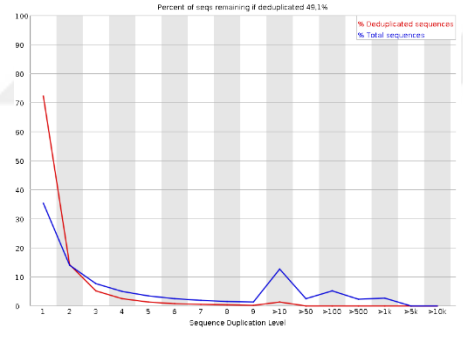
1)



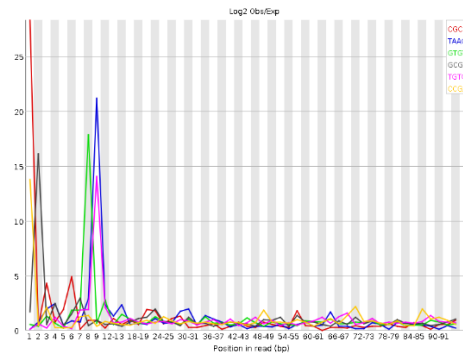
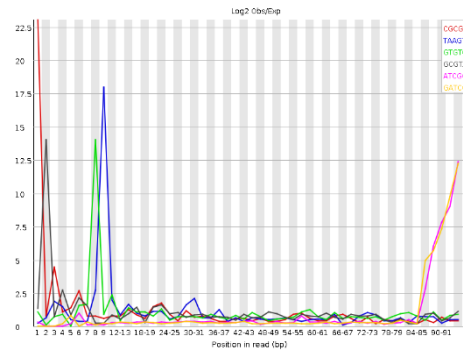
2)



3)



4)

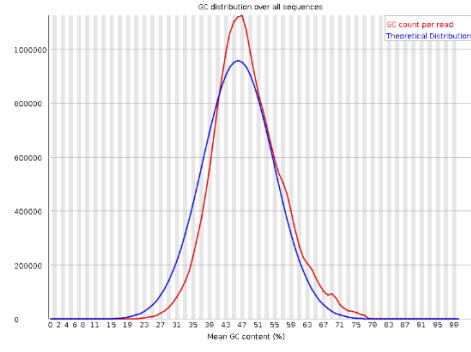
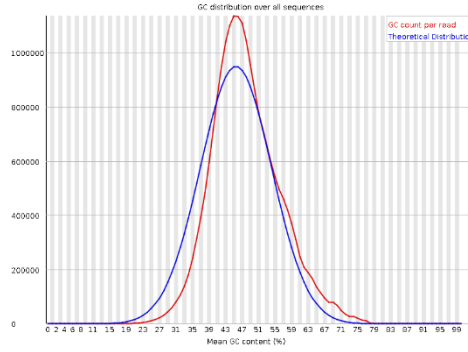


Şekil 4.3 *C. colurna* L.(A) ve *C. avellana* L. Tombul (B) temiz okumalarının FastQC programı ile kontrolü sonucu elde edilen grafikler

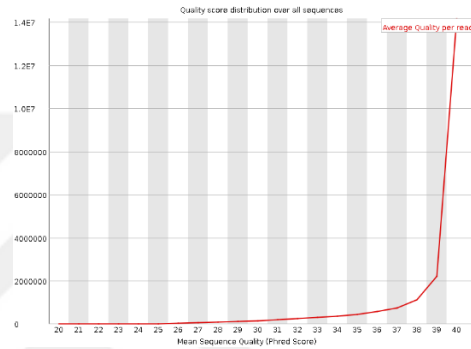
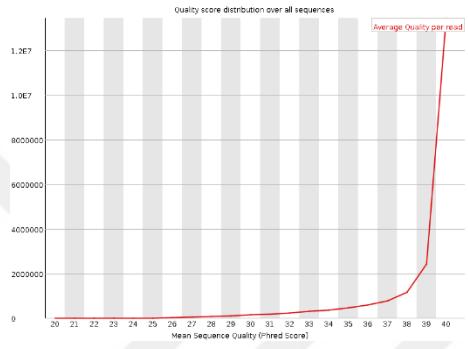
A) *C. colurna* L.

B) *C. avellana* L. Tombul

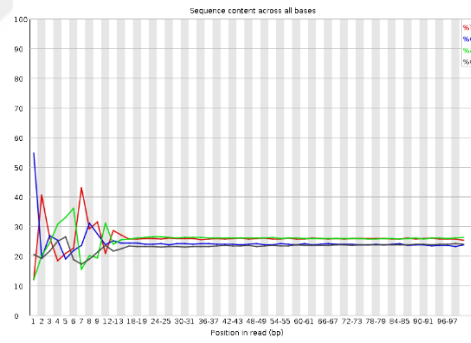
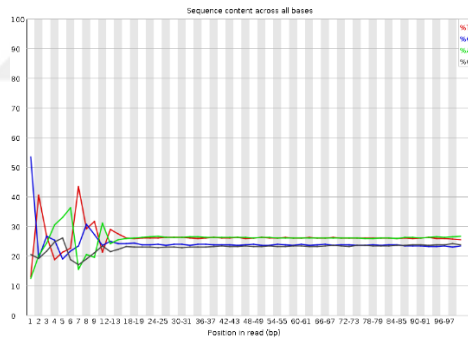
5)



6)



7)



Şekil 4.3 *C. colurna* L.(A) ve *C. avellana* L. Tombul (B) temiz okumalarının FastQC programı ile kontrolü sonucu elde edilen grafikler (devamı)

1. grafik, her bir baz konumunun dizileme kalitesini vermektedir. Burada ortalama kalite değerini ifade eden “BoxWhisker” grafiği kullanılmış, kalite değerleri üç renk ile ifade edilmiştir. Yeşil bölge, kalite değerinin en yüksek olduğu kısım, sarı bölge kalitenin kabul edilebilir seviyede olduğu kısımdır. Grafik incelendiğinde kalite seviyelerinin yüksek olduğu görülmüştür.

2. grafik adaptör içeriği ile ilgilidir. Bu grafik bilinen adaptör içeriklerinin verilerde yer alma durumunu ifade eder. Herhangi bir adaptör varlığında grafikte varlığı

yükselerek ifade edilmektedir. Grafik incelendiğinde verilerde adaptör varlığına rastlanmamıştır.

3. grafik dizilerin duplikasyon seviyelerini ifade etmektedir. Dizilerdeki duplikasyon olmadığında bu durum kırmızı çizgi ile ifade edilmektedir. Verilerden birisi aşırı duplikasyona sahip değilse temsil edilme oranını azaltmamak için müdahale edilmemesi tercih edilmektedir.

4. grafikte k-mer içeriği gösterilmekte, her pozisyonadaki k-merlerin sayısı ölçeklenmektedir. K-merdeki sapmaların listelendiği ve en en çok sapma yapan 6 k-mer içeriklerinin yer verildiği grafiklerde benzer yoğunlukta bir varlık görülmüştür.

5. grafik okumalardaki GC içeriğini belirtmekte, veri içeriğindeki dizilerin GC dağılımı ile beklenen GC oranı karşılaştırılmıştır. İki örnek de benzer davranışlar sergilemiştir.

Okumalardaki her birinin kalite skorunun ortalama değerlerinin elde edilmesiyle oluşturulan 6. grafik, dizilemedeki düşük kaliteli değerlere sahip okuma varlığını göstermesi açısından değerlidir. PHRED skorunun 20'nin altında olması kalite değerlerinin düşük olduğu anlamına gelmektedir. Grafikteki gibi sonuçların 40'a yaklaşması ise dizileme kalitesinin iyi olduğunu ifade eder.

Şekil 4.3'deki 7. grafik 4 bazın kendi aralarındaki oransal ifadesidir. Her bir bazın bağıl miktarı, transkriptomdaki veya genomdaki bazların oranlarını temsil eder.

4.2.3 RNA-seq Verilerinin *de novo* Derlemesi

Türk Fındığı ve Tombul Fındık RNA-seq verilerinin Trinity ile derlemesi esnasında yaklaşık 40 GB'lık veri topluluğu ve nihai olarak elde edilmesi amaçlanan Trinity.fasta adlı dosya oluşmuştur. Bu dosya içeriğindeki transkript sayısı sorgulandığında toplam 95.979 derlenmiş transkript elde edildiği görülmüştür. Dosya içeriğinde, okumaların derlemesi sonucunda elde edilen transkriptlerin baz dizileri, her birisi için özel olarak atanan ve işlemler boyunca kullanılacak bir isim, transkript uzunluğunu ve Bruijn grafikleri oluşturulurken üretilen yapılandırılma bilgileri yer almaktadır. Şekil 4.4'de Trinity.fasta içeriğinden örnek görsele yer verilmiştir.

transcript_id	gene_id	length	effective_length	expected_count	TPM	FPKM	IsoPct
TRINITY_DN0_c0_g1_i1	TRINITY_DN0_c0_g1	1085	909.13	15.53	1.26	0.99	6.49
TRINITY_DN0_c0_g1_i2	TRINITY_DN0_c0_g1	676	500.13	0.00	0.00	0.00	0.00
TRINITY_DN0_c0_g1_i3	TRINITY_DN0_c0_g1	1311	1135.13	279.47	18.13	14.31	93.51
TRINITY_DN0_c0_g2_i1	TRINITY_DN0_c0_g2	1440	1264.13	0.00	0.00	0.00	0.00
TRINITY_DN0_c0_g3_i1	TRINITY_DN0_c0_g3	471	295.20	72.26	18.03	14.23	100.00
TRINITY_DN0_c0_g4_i1	TRINITY_DN0_c0_g4	1017	841.13	668.74	58.55	46.22	100.00
TRINITY_DN0_c1_g1_i1	TRINITY_DN0_c1_g1	1793	1617.13	173.58	7.91	6.24	18.03
TRINITY_DN0_c1_g1_i2	TRINITY_DN0_c1_g1	1994	1818.13	0.00	0.00	0.00	0.00
TRINITY_DN0_c1_g1_i3	TRINITY_DN0_c1_g1	1912	1736.13	28.22	1.20	0.94	2.73

Şekil 4.6 RSEM tarafından oluşturulan *C. avellana* L. Tombul transkriptlerinin ifade seviyeleri örnekleri

4.2.5 Derleme Kalitesi Analizi Sonuçları

Derlemenin genel kalitesi nicel veya nitel olarak değerlendirmiştir. *grep '>' Trinity.fasta | wc -l* sorgusuna yanıt olarak 95.979 verilmiş yani derleme sonrası 95.979 Trinity transkripti derlenmiştir. Bu transkriptlerle ilgili diğer istatistikler ayrıca Tablo 4.3'de verilmiştir. Buna göre 57.698 gen üretilmiş, GC yüzdesi 43,59 olmuştur. N50 değeri 1929, ortalama kontig uzunluğu 1116,82 baz ve derlenen baz sayısı 107.190.859 olmuştur.

Tablo 4.3 Trinity derleme kalitesi ile ilgili istatistikler

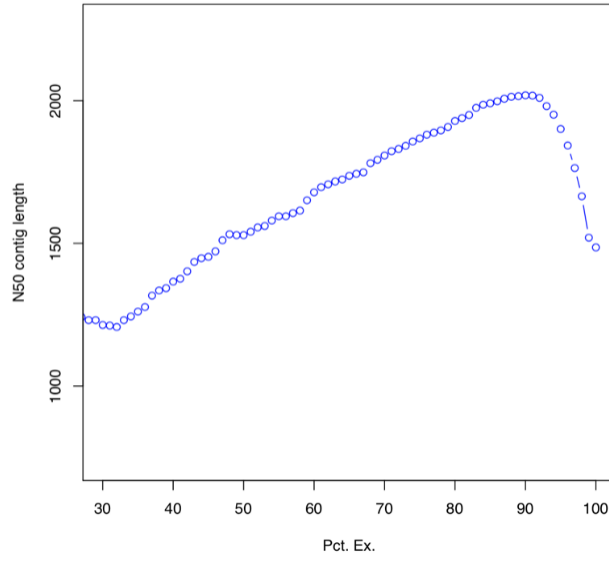
Transkript sayısı	
Toplam Gen Sayısı	57.698
Toplam Transkript Sayısı	95.979
GC Yüzdesi	43.59
Tüm Kontiglerin İstatistikleri	
Kontig N10	4283
Kontig N20	3314
Kontig N30	2720
Kontig N40	2280
Kontig N50	1929
Ortalama Kontig Uzunluğu	1116,82
Derlenen Baz Sayısı	107.190.859

Yüksek kaliteli derlemenin en önemli göstergelerinden biri olan okumaların temsil oranının belirlenmesi için Bowtie2 dizini oluşturulmuş ve BAM dosyası elde edilmiştir. Sonuçlar değerlendirildiğinde uyumsuz çift sayısının az olması, uyumluluk oranının %70'in üzerinde olması derlemenin sağlıklı olması açısından gereklidir. Tablo 4.4'deki sonuçlar incelendiğinde, genel uyum oranının *C. avellana* L. Tombul için %99,17 ve *C. colurna* L. %98,09 olduğu, ortalamasının %98.63 olduğu görülmüştür.

Tablo 4.4 Okumaların Trinity Derlemesindeki Temsil Oranları

	<i>C. avellana</i> L. Tombul	<i>C. colurna</i> L.
Eşleşen Okuma Sayısı	20.927.213	20.570.158
0 Uyumlu Hizalama Oranı	%2,45	%3,97
1 Defa Uyumlu Hizalama Oranı	%28,64	%26,19
> 1 Uyumlu Hizalama Oranı	%68,91	%69,83
Genel Uyum Oranı	%99,17	%98,09

Geleneksel bir ölçü olan N50 yerine daha modern ve kabul gören bir kalite istatistiği olan ExN50 değeri N50 Kontig istatistiğini de kullanır fakat en yüksek düzeyde ifade edilen transkriptleri ölçer. İyi kaliteli bir derlemede ExN50 tepe zirvesinin %90'larda olması beklenir. Şekil 4.7'de yer verilen grafikte derleme ExN50 değerinin %92 olduğu görülmüştür.



Şekil 4.7 Derleme kalitesinin değerlendirilmesi (ExN50)

Tam uzunluktaki kodlayan transkriptlerin değerlendirmesinde BLASTX komutuyla SWISSPROT veritabanında e-değeri $1e-20$ olacak şekilde sorgu gerçekleştirilmiş, sonuçların bir özeti Tablo 4.5’de paylaşılmıştır.

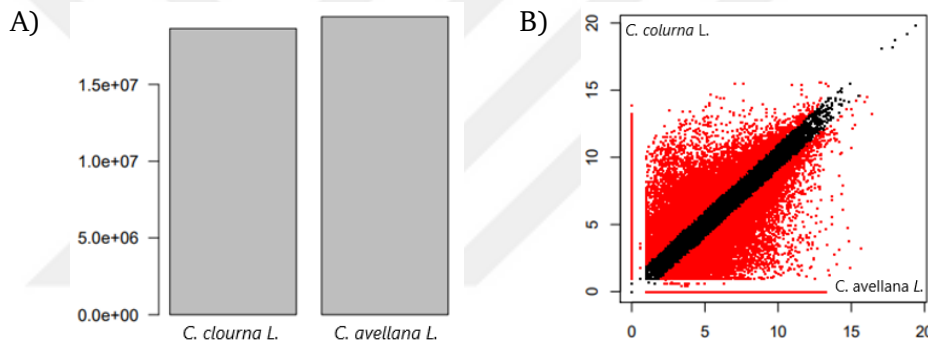
Tablo 4.5 Tam Uzunluktaki Kodlayan Transkriptlerin Sayısı

Oran (%)	Eşleşen Protein Sayısı	Toplam Protein Sayısı
100	7268	7268
90	1456	8724
80	839	9563
70	714	10277
60	735	11012
50	740	11752
40	814	12566
30	1026	13592
20	1100	14692
10	330	15022

Buna göre veritabanındaki 7268 proteinin, derlemede elde edilen transkriptlerin uzunluklarının %90 ile %100 arasında eşleştiği görülmüştür. Yine SWISSPROT veritabanı ile %80 ile %90 arasında eşleşme sağlayan transkript sayısı 1456 olmuştur.

4.2.6 Örneklerin Veri Kalitesi Sonuçları

Örneklerin veri kaliteleri diferansiyel ifade analizinin sonuçlarını da etkilemektedir. Bu sebeple veriler aşırı benzerliğin olup olmadığı ve data kirliliği açısından incelenmiştir. Şekil 4.8’ de görüldüğü gibi derlemeye hizalanan okuma sayısının iki örnekte farklılığı (A) ve transkript başına düşen okuma sayısının (log2) dağılımına göre (B) değerlendirilerek iki örneğe ait verilerin farklılıkları ortaya konulmuştur.



Şekil 4.8 Veri kalitesinin değerlendirilmesi A) Derlemeye hizalanan okuma sayıları B) Transkript başına düşen okuma sayısının (log2) dağılımı

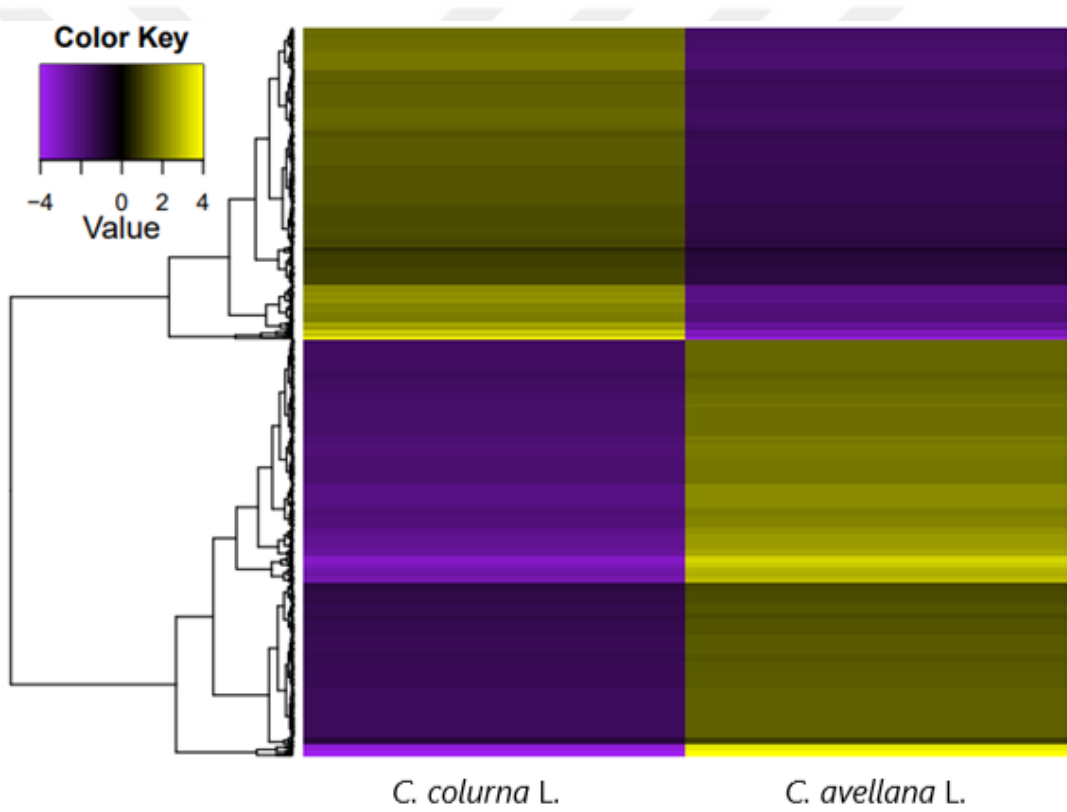
4.2.7 Diferansiyel İfade Analizi Sonuçları

İki fındık türü arasındaki diferansiyel olarak ifade edilen transkriptleri ve genleri tespit etmek için EdgeR kullanılmış, karşılaştırma sonuçları ve tespit edilen transkriptlere ve genlere ait DE_result dosyasındaki veriler Şekil 4.9 örneğindeki gibi tablolaştırılmıştır. Bu veriler her bir transkriptte özel olarak logaritmik kat değişikliği (logFC), logCPM, P-Değeri ve FDR gibi istatistiksel değerleri de barındırmaktadır.

sampleA	sampleB	logFC	logCPM	PValue	FDR				
TRINITY_DN2257_c0_g1_i5	A-214	C-213	-10.1275554058443	9.80844193200736	4.42900047984386e-25	9.32880371069513e-21			
TRINITY_DN2257_c0_g1_i3	A-214	C-213	9.26414412870827	9.6926004530355	8.46782158316222e-23	8.91788630030729e-19			
TRINITY_DN5606_c0_g1_i8	A-214	C-213	8.75970990658169	9.04165189810716	2.50776745011646e-21	1.76070352672677e-17			
TRINITY_DN3050_c0_g2_i1	A-214	C-213	8.67918989343838	9.53154996335611	3.42425768704777e-21	1.80312849155718e-17			
TRINITY_DN2486_c0_g1_i1	A-214	C-213	-8.72579028824846	8.71791568550978	5.10067235595336e-21	1.90856204384197e-17			
TRINITY_DN1008_c0_g1_i5	A-214	C-213	8.66858855528006	8.78634170718496	5.43672423826227e-21	1.90856204384197e-17			
TRINITY_DN6935_c1_g1_i1	A-214	C-213	8.64121837985558	8.52304943270974	8.29499686753895e-21	2.18852569767422e-17			
TRINITY_DN6_c0_g1_i12	A-214	C-213	-8.71272938867319	8.33747041430297	8.3123038415201e-21	2.18852569767422e-17			
TRINITY_DN872_c0_g1_i7	A-214	C-213	8.5509673528599	8.90916502122768	1.24038707610755e-20	2.90291922045037e-17			

Şekil 4.9 İki fındık türü arasındaki diferansiyel olarak ifade edilen genler (EdgeR'a göre)

Bu ikili karşılaştırma sonucunda elde edilen verilerden, P değeri 0,001 ve kat değişikliği >4 olarak ifade edilen transkriptlerle Şekil 4.10'daki sıcaklık haritası oluşturulmuştur. Isı haritasının sol tarafında kümelenen transkript ifade profilleri benzer ifade modellerine sahip büyük transkript kümelerini temsil eder.



Şekil 4.10 Diferansiyel olarak farklı ifade edilen, P değeri 0.001 ve kat değişikliği >4 olarak ifade edilen transkriptlerin sıcaklık haritası

İki tür arasında farklı ifade edilen transkript sayısının 1528 olduğu gen sayısının ise 736 olduğu bulunmuştur. Tablo 4.6'de iki tür arasında artma ve azalma gösteren diferansiyel olarak ifade edilen gen ve transkriptlerin sayısı gösterilmiştir.

Tablo 4.6 İki tür arasında diferansiyel olarak ifade edilmiş gen ve transkriptler

	İfadesi artış gösteren transkript sayısı	İfadesi artış gösteren gen sayısı
<i>C. avellana</i> L. Tombul	876	439
<i>C. colurna</i> L.	652	297

C. colurna L.'de diferansiyel olarak ifade edilen genlerin içerisinde *C. avellana* L. Tombul ile kıyaslandığında en fazla artış gösteren ilk 20 gen Tablo 4.7'de listelenmiştir.

Tablo 4.7 *C. colurna* L.'de *C. avellana* L. Tombul'a kıyasla artış gösteren ilk 20 gen

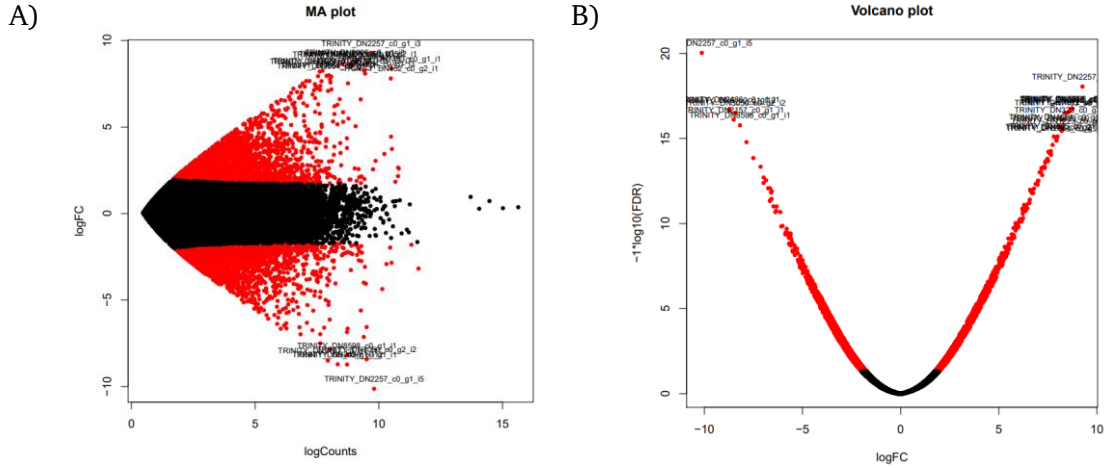
Gen İsmi	Örnek A	Örnek B	logFC	logCPM	PValue	FDR
TRINITY_DN12164_c0_g1	A-214	C-213	-10,41	9,73	6,20E-26	1,56E-21
TRINITY_DN1154_c1_g3	A-214	C-213	-8,47	9,61	1,11E-20	1,40E-16
TRINITY_DN8598_c0_g1	A-214	C-213	-8,09	7,61	4,30E-19	3,61E-15
TRINITY_DN8055_c0_g1	A-214	C-213	-7,86	7,12	3,03E-18	1,91E-14
TRINITY_DN562_c0_g1	A-214	C-213	-7,23	7,14	1,08E-16	4,53E-13
TRINITY_DN4972_c0_g1	A-214	C-213	-7,06	6,58	5,21E-16	1,46E-12
TRINITY_DN1745_c0_g1	A-214	C-213	-7,03	8,64	1,67E-16	6,03E-13
TRINITY_DN1090_c1_g1	A-214	C-213	-7,00	6,52	7,82E-16	1,81E-12
TRINITY_DN8388_c0_g3	A-214	C-213	-6,93	7,16	7,91E-16	1,81E-12
TRINITY_DN1154_c1_g1	A-214	C-213	-6,60	8,48	3,09E-15	5,56E-12
TRINITY_DN1154_c0_g2	A-214	C-213	-6,56	11,01	2,92E-15	5,56E-12
TRINITY_DN2300_c0_g2	A-214	C-213	-6,42	5,88	4,28E-14	5,39E-11
TRINITY_DN244_c0_g2	A-214	C-213	-6,42	8,09	1,10E-14	1,85E-11
TRINITY_DN7890_c0_g1	A-214	C-213	-6,39	5,97	4,63E-14	5,56E-11
TRINITY_DN8388_c0_g1	A-214	C-213	-6,28	10,11	1,97E-14	3,11E-11
TRINITY_DN9272_c0_g1	A-214	C-213	-6,15	8,19	6,54E-14	7,22E-11
TRINITY_DN817_c0_g2	A-214	C-213	-6,06	6,37	2,70E-13	2,62E-10
TRINITY_DN7829_c0_g1	A-214	C-213	-6,02	5,23	8,39E-13	6,61E-10
TRINITY_DN3694_c0_g3	A-214	C-213	-6,00	5,59	6,21E-13	5,05E-10
TRINITY_DN11919_c0_g1	A-214	C-213	-5,99	6,37	3,94E-13	3,55E-10

C. avellana L. Tombul'da diferansiyel olarak ifade edilen genlerin içerisinde *C. colurna* L. ile kıyaslandığında en fazla artış gösteren ilk 20 gen Tablo 4.8'de listelenmiştir.

Tablo 4.8 *C. avellana* L. Tombul'da *C. colurna* L.'ye kıyasla artış gösteren ilk 20 gen

Gen İsmi	Örnek A	Örnek B	logFC	logCPM	PValue	FDR
TRINITY_DN13944_c0_g1	A-214	C-213	7,63	7,11	1,34E-17	6,76E-14
TRINITY_DN8971_c0_g1	A-214	C-213	6,91	7,85	5,20E-16	1,46E-12
TRINITY_DN1024_c0_g1	A-214	C-213	6,81	8,12	8,96E-16	1,88E-12
TRINITY_DN2967_c0_g1	A-214	C-213	6,52	5,95	2,72E-14	3,61E-11
TRINITY_DN3003_c0_g1	A-214	C-213	6,38	6,68	2,60E-14	3,61E-11
TRINITY_DN26184_c0_g1	A-214	C-213	6,32	5,81	9,75E-14	1,02E-10
TRINITY_DN2323_c0_g4	A-214	C-213	6,32	7,46	2,66E-14	3,61E-11
TRINITY_DN1223_c0_g1	A-214	C-213	6,17	7,68	6,59E-14	7,22E-11
TRINITY_DN893_c0_g3	A-214	C-213	6,10	6,16	2,46E-13	2,48E-10
TRINITY_DN5743_c0_g2	A-214	C-213	5,95	6,61	3,75E-13	3,50E-10
TRINITY_DN908_c0_g5	A-214	C-213	5,90	5,66	1,19E-12	8,85E-10
TRINITY_DN786_c0_g1	A-214	C-213	5,83	5,49	2,21E-12	1,55E-09
TRINITY_DN13095_c0_g1	A-214	C-213	5,80	6,08	1,26E-12	9,06E-10
TRINITY_DN1748_c1_g3	A-214	C-213	5,78	6,93	1,00E-12	7,65E-10
TRINITY_DN497_c0_g1	A-214	C-213	5,60	7,62	2,49E-12	1,69E-09
TRINITY_DN729_c2_g1	A-214	C-213	5,57	8,50	2,54E-12	1,69E-09
TRINITY_DN13605_c0_g1	A-214	C-213	5,53	6,43	5,87E-12	3,61E-09
TRINITY_DN3551_c0_g1	A-214	C-213	5,43	7,93	6,98E-12	4,09E-09
TRINITY_DN8563_c0_g1	A-214	C-213	5,40	5,35	2,79E-11	1,30E-08
TRINITY_DN6054_c0_g1	A-214	C-213	5,39	8,48	8,60E-12	4,93E-09

Diferansiyel olarak ifade edilen transkriptlerin görselleştirilmesi için EdgeR'dan yararlanılmış ve $FDR \leq 0.05$ olacak şekilde filtrelenerek Şekil 4.11'deki MA Plot (A) ve Volcano Plot (B) oluşturulmuştur. Grafiklerdeki kırmızı noktalar verilen FDR değerine göre diferansiyel olarak ifade edilen transkriptleri ifade etmektedir.



Şekil 4.11 Diferansiyel olarak ifade edilen transkriptlerin $\text{FDR} \leq 0.05$ olacak şekilde filtrelenerek A) MA plot ve B) Volcano plot ile gösterimi

4.2.8 İşlevsel Açıklama Eklenmesi

Transkriptlerin özelliklerini belirlemek için birkaç program kullanılmış ve bir dizi işlem yapılmıştır. Elde edilen yeni bilgilerle transkriptlere özelliklerinin ve görevlerinin aydınlatılmasını sağlayacak açıklamalar eklenmiştir.

Transdecoder, transkriptlerdeki olası kodlayan bölgelerin tespitinde kullanılmaktadır. Öncelikle uzun açık okuma çerçeveleri tespit edilmiş ve sonrasında dizilerin kompozisyonuna göre puanlama yapılmıştır. Programın çalıştırılması sonucu “Trinity.fasta.transdecoder.pep” dosyası çıktı olarak sağlanmış, içeriğinin bir örneği Şekil 4.12’de verilmiştir. Bu peptit dosyasındaki başlık satırlarında orijinal transkriptlerin isimleri yer alır. Type kısmı proteinin bir başlangıç ve bir durdurma kodonu içerdiğini belirten “tamamlanmış” olup olmadığını ya da hangi kısmın eksik olduğu bilgisini verir. Ayrıca kodlama bölgesinin hangi dizide bulunduğu ve bu transkript dizisindeki ORF’nin koordinatlarını belirtmek için (+) veya (-) şeklinde ifade edilir.

Trinity.fasta.transdecoder.pep dosyasında ORF sorgusu yapıldığında transkriptler içerisinde 4523 adet kodlayan bölgelerin tespit edildiği görülmüştür.

Tahmin edilen protein dizileri kullanılarak Pfam veritabanına karşı HMMER sorgusu yapılmış ve korunmuş domainler tespit edilmiştir.


```

>TRINITY_DN1000_c0_g1_i1.p1 TRINITY_DN1000_c0_g1~TRINITY_DN1000_c0_g1_i1.p1 ORF type:complete len:829 (-),score=150.23
TRINITY_DN1000_c0_g1_i1:337-2823(-)
MDGEANLPVRPASERILIRPNTVKPQDTTSSSSSSSPDFLRNVQAFAKRRHRPIGAQSN
ILRPRRMLVPQREVSRTSDSSVGPMDAKKSQDVVSLRQGHMAKGSIVATKSAATVLGET
QDDASITPPSISGTITKTFDESFNQFDVQRDQPKFVTGCKDDRAMPLHGESQNTDGHKK
VQLSIGSNVSEQEMEWIASNQIEASTVVNDPKHRSFHNTDSERSAAGITSLAKRTTVDV
QQDLHLRLNLSQATQSSVGPSCATTTSVHSTAPMLNLTTCSHLHRDSSSHVAVEP
LVDININSQCKGQGNVQVLSNPSPKDVGEMIIDEKAIAGKASNSAIDTHVEVKQHDLSKE
QQGRVPKEIDIPENPSLHDNKSTKGGYAGDVTNVQSQGPLSKNSSSDVKLEPSKSERQE
RAASGKGASAPRKRSYDPDLFFKVGKLYQRLGKIGSGGSSEVHKVSSDCAIYALKKIK
LKGRDYATAYGFCQEIQYLKKLKGKNNIIQLIDFEVTDALLQEVNMGSMNKGGRVKDD
GYIYMWLEYGEIDLHMLSQKWKEMDGSNQITIDENWLRIFYWQQILHAVSTIHEERIVHSD
LKPANFLLVKGSLKLDIFGIAKAIMSDTTNIQRDSQVGTLSYMSPEAFMCNENDANGNII
KCGRPSDIWSLGCILYQMWYGRTPFAEYKTFWAKFKVITDPNHEITYEPVPPNWL DLMK
KCLAWDRNERWRIPQLLQHPFLVPPVPEPRLSSSQDQSKLLQLIAETCAGDREASALC
CQLEELLKDPVQKITSPLLTSRDQCKLLSQMSKLCFQLQGRLANSEN*
>TRINITY_DN1000_c0_g1_i3.p1 TRINITY_DN1000_c0_g1~TRINITY_DN1000_c0_g1_i3.p1 ORF type:complete len:829 (-),score=145.01
TRINITY_DN1000_c0_g1_i3:337-2823(-)
MDGEANLPVRPASERILIRPNTVKPQDTTSSSSSSSPDFLRNVQAFAKRRHRPIGAQSN
ILRPRRMLVPQREVSRTSDSSVGPMDAKKSQDVVSLRQGHMAKGSIVATKSAATVLGET
QDDASITPPSISGTITKTFDESFNQFDVQRDQPKFVTGCKDDRAMPLHGESQNTDGHKK
VQLSIGSNVSEQEMEWIASNQIEASTVVNDPKHRSFHNTDSERSAAGITSLAKRTTVDV
QQDLHLRLNLSQATQSSVGPSCATTTSVHSTAPMLNLTTCSHLHRDSSSHVAVEP
LVDININSQCKGQGNVQVLSNPSPKDVGEMIIDEKAIAGKASNSAIDTHVEVKQHDLSKE
QQGRVPKEIDIPENPSLHDNKSTKGGYAGDVTNVQSQGPLSKNSSSDVKLEPSKSERQE
RAASGKGASAPRKRSYDPDLFFKVGKLYQRLGKIGSGGSSEVHKVSSDCAIYALKKIK
LKGRDYATAYGFCQEIQYLKKLKGKNNIIQLIDFEVTDALLQEVNMGSMNKGGRVKDD
GYIYMWLEYGEIDLHMLSQKWKEMDGSNQITIDENWLRIFYWQQILHAVSTIHEERIVHSD
LKPANFLLVKGSLKLDIFGIAKAIMSDTTNIQRDSQVGTLSYMSPEAFMCNENDANGNII
KCGRPSDIWSLGCILYQMWYGRTPFAEYKTFWAKFKVITDPNHEITYEPVPPNWL DLMK
KCLAWDRNERWRIPQLLQHPFLVPPVPEPRLSSSQDQSKLLQLIAETCAGDREASALC
CQLEELLKDPVQKITSPLLTSRDQCKLLSQMSKLCFQLQGRLANSEN*
>TRINITY_DN1000_c0_g1_i4.p1 TRINITY_DN1000_c0_g1~TRINITY_DN1000_c0_g1_i4.p1 ORF type:complete len:852 (-),score=157.43
TRINITY_DN1000_c0_g1_i4:337-2892(-)
MDGEANLPVRPASERILIRPNTVKPQDTTSSSSSSSPDFLRNVQAFAKRRHRPIGAQSN
ILRPRRMLVPQREVSRTSDSSVGPMDAKKSQDVVSLRQGHMAKGSIVATKSAATVLGET
QDDASITPPSISGTITKTFDESFNQFDVQRDQPKFVTGCKDDRAMPLHGESQNTDGHKK
VQLSIGSNVSEQEMEWIASNQIEASTVVNDPKHRSFHNTDSERSAAGITSLAKRTTVDV
QQDLHLRLNLSQATQSSVGPSCATTTSVHSTAPMLNLTTCSHLHRDSSSHVAVEP
LVDININSQCKGQGNVQVLSNPSPKDVGEMIIDEKAIAGKASNSAIDTHVEVKQHDLSKE
QQGRVPKEIDIPENPSLHDNKSTKGGYAGDVTNVQSQGPLSKNSSSDVKLEPSKSERQE
RAASGKGASAPRKRSYDPDLFFKVGKLYQRLGKIGSGGSSEVHKVSSDCAIYALKKIK
LKGRDYATAYGFCQEIQYLKKLKGKNNIIQLIDFEVTDALLQEVNMGSMNKGGRVKDD
GYIYMWLEYGEIDLHMLSQKWKEMDGSNQITIDENWLRIFYWQQILHAVSTIHEERIVHSD
LKPANFLLVKGSLKLDIFGIAKAIMSDTTNIQRDSQVGTLSYMSPEAFMCNENDANGNII
KCGRPSDIWSLGCILYQMWYGRTPFAEYKTFWAKFKVITDPNHEITYEPVPPNWL DLMK
KCLAWDRNERWRIPQLLQHPFLVPPVPEPRLSSSQDQSKLLQLIAETCAGDREASALC
CQLEELLKDPVQKITSPLLTSRDQCKLLSQMSKLCFQLQGRLANSEN*

```

Şekil 4.12 Transdecoder programının çıktısı olan
“Trinity.fasta.transdecoder.pep” dosyasının içeriği

SignalP aracı sinyal peptitlerinin tespiti için çalıştırılmış tahmin edilen sinyal peptitleri “signalp.out” dosyasında Şekil 4.13’de görüldüğü gibi “SP(Sec/SPI)” şeklinde belirtilmiştir. Bu dosya içerisinde *grep 'sec/SPI' signalp.out | wc -l* sorgusu yapıldığında 336 sinyal peptidinin işaretlendiği görülmüştür.

TRINITY_DN2518_c0_g1_i12.p1	SP(Sec/SPI)	0.981005	0.018995	CS pos: 22-23. GFS-QN. Pr: 0.4377
TRINITY_DN2518_c0_g1_i2.p1	SP(Sec/SPI)	0.981005	0.018995	CS pos: 22-23. GFS-QN. Pr: 0.4377
TRINITY_DN2518_c0_g1_i3.p1	SP(Sec/SPI)	0.981005	0.018995	CS pos: 22-23. GFS-QN. Pr: 0.4377
TRINITY_DN2518_c0_g1_i4.p1	SP(Sec/SPI)	0.981005	0.018995	CS pos: 22-23. GFS-QN. Pr: 0.4377
TRINITY_DN2518_c0_g1_i6.p1	SP(Sec/SPI)	0.981005	0.018995	CS pos: 22-23. GFS-QN. Pr: 0.4377
TRINITY_DN2518_c0_g1_i8.p1	SP(Sec/SPI)	0.981005	0.018995	CS pos: 22-23. GFS-QN. Pr: 0.4377
TRINITY_DN2518_c0_g1_i9.p1	SP(Sec/SPI)	0.981005	0.018995	CS pos: 22-23. GFS-QN. Pr: 0.4377
TRINITY_DN2522_c0_g1_i10.p1	SP(Sec/SPI)	0.995286	0.004714	CS pos: 27-28. SHG-AD. Pr: 0.9538
TRINITY_DN2522_c0_g1_i12.p1	SP(Sec/SPI)	0.995286	0.004714	CS pos: 27-28. SHG-AD. Pr: 0.9538
TRINITY_DN2522_c0_g1_i4.p1	SP(Sec/SPI)	0.995286	0.004714	CS pos: 27-28. SHG-AD. Pr: 0.9538
TRINITY_DN2522_c0_g1_i9.p1	SP(Sec/SPI)	0.995286	0.004714	CS pos: 27-28. SHG-AD. Pr: 0.9538
TRINITY_DN2524_c0_g1_i3.p1	OTHER	0.004706	0.995294	
TRINITY_DN2525_c2_g1_i1.p1	SP(Sec/SPI)	0.730678	0.269322	CS pos: 23-24. RIC-QN. Pr: 0.1798
TRINITY_DN2525_c2_g1_i3.p1	OTHER	0.001366	0.998634	
TRINITY_DN2525_c2_g1_i4.p1	SP(Sec/SPI)	0.730678	0.269322	CS pos: 23-24. RIC-QN. Pr: 0.1798
TRINITY_DN252_c0_g2_i1.p1	OTHER	0.000798	0.999202	
TRINITY_DN2531_c0_g1_i2.p1	SP(Sec/SPI)	0.984122	0.015878	CS pos: 21-22. GRC-HR. Pr: 0.7170
TRINITY_DN2531_c0_g1_i3.p1	SP(Sec/SPI)	0.984122	0.015878	CS pos: 21-22. GRC-HR. Pr: 0.7170
TRINITY_DN2531_c0_g1_i4.p1	SP(Sec/SPI)	0.984122	0.015878	CS pos: 21-22. GRC-HR. Pr: 0.7170
TRINITY_DN2531_c0_g1_i6.p1	SP(Sec/SPI)	0.984122	0.015878	CS pos: 21-22. GRC-HR. Pr: 0.7170
TRINITY_DN2531_c0_g1_i7.p1	SP(Sec/SPI)	0.984122	0.015878	CS pos: 21-22. GRC-HR. Pr: 0.7170

Şekil 4.13 SignalP aracının çıktısı olan “signalp.out” dosyasının içeriği

Transmembran proteinlerini tahmin etmek için tmhmm aracı çalıştırılmış ve bu alanlar "tmhmm.out" dosyasında Şekil 4.14'de de görüldüğü gibi işaretlenmiştir. Çıktı olarak alınan "tmhmm.out" dosyasında yapılan sorguda en az 1 sarmala sahip transmembran proteinlerin sayısı 887 olarak tespit edilmiştir ve transmembran proteini olarak işaretlenen bu proteinlerin sarmal sayısına, boyutu ve zarın içerisinde veya dışarısında bulunduğu gibi konumuna ait bilgilerin bulunduğu topoloji bilgilerine de yer vermiştir.

TRINITY_DNS137_c0_g1_i1.p1	len=848	ExpAA=41.70	First60=17.70	PredHel=2	Topology=i7-29o446-468i
TRINITY_DNS145_c0_g1_i2.p1	len=811	ExpAA=113.41	First60=3.89	PredHel=5	Topology=o564-586i628-645o672-694i733-750o760-779i
TRINITY_DNS145_c0_g1_i7.p1	len=827	ExpAA=113.38	First60=3.87	PredHel=5	Topology=o580-602i644-661o688-710i749-766o776-795i
TRINITY_DNS145_c0_g1_i8.p1	len=827	ExpAA=113.38	First60=3.87	PredHel=5	Topology=o580-602i644-661o688-710i749-766o776-795i
TRINITY_DNS145_c0_g1_i9.p1	len=799	ExpAA=88.73	First60=3.87	PredHel=3	Topology=o644-661i688-710o749-768i
TRINITY_DNS147_c0_g1_i3.p1	len=843	ExpAA=0.00	First60=0.00	PredHel=0	Topology=o
TRINITY_DNS153_c0_g1_i2.p1	len=1410	ExpAA=0.02	First60=0.00	PredHel=0	Topology=o
TRINITY_DNS153_c0_g1_i4.p1	len=1395	ExpAA=0.02	First60=0.00	PredHel=0	Topology=o
TRINITY_DNS157_c0_g1_i2.p1	len=1055	ExpAA=0.08	First60=0.00	PredHel=0	Topology=o
TRINITY_DNS157_c0_g1_i3.p1	len=1055	ExpAA=0.08	First60=0.00	PredHel=0	Topology=o

Şekil 4.14 Tmhmm aracının çıktısı olan "tmhmm.out" dosyasının içeriği

4.2.9 Veritabanı Oluşturulması

Elde edilen tüm işlevsel açıklamalara ve diğer verilere hızlı bir erişim sağlamak için Trinotate kullanılarak veritabanı oluşturulmuş, içeriğine SWISSPROT kayıtları, Pfam bilgileri eklenmiştir. Bu dosya "Trinotate.sqlite" adı ile tutulmuştur. Ayrıca veritabanına, çalışma sonucunda elde edilen farklı ifade edilen transkriptler, sinyal peptidleri bilgileri, transmembran proteinleri verileri de dahil edilerek Şekil 4.15'de görüldüğü gibi sorgu yapılabilir bir xls sayfası elde edilmiştir.

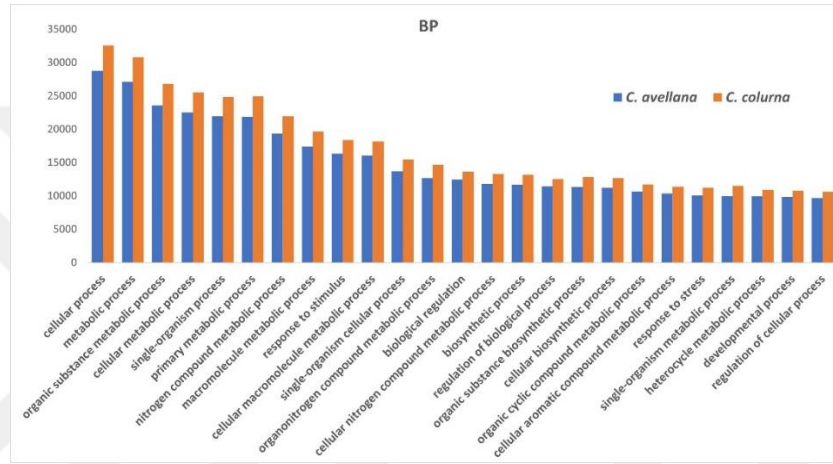
transcript_id	sprot_Top_BLASTX_hit	prot_id	prot_coords	sprot_Top_BLASTP_Pfam	SignalP	TmHMM	gene_ontology_Pfam
TRINITY_DN13348_c0_g1_i1	sp F4HR11 KN6_ARATH	TRINITY_DN13348_c0_g1_i1.p1	314-3070[-]	sp F4HR11 KN6_ARATH	PF00225.23*Tyros	ExpAA=18.97*Pre	GO:0003777*molecular_function
TRINITY_DN13399_c0_g1_i2	sp Q92396 TYRO_PODA	TRINITY_DN13399_c0_g1_i2.p1	139-2967[+]	sp P00440 TYRO_N	PF00264.20*Tyros	ExpAA=18.97*Pre	GO:0016491*molecular_function
TRINITY_DN13399_c0_g1_i1	sp Q92396 TYRO_PODA	TRINITY_DN13399_c0_g1_i1.p1	139-2937[+]	sp P00440 TYRO_N	PF00264.20*Tyros	ExpAA=18.97*Pre	GO:0016491*molecular_function
TRINITY_DN13339_c0_g1_i6	sp XSJ851 SECAB_ARAT	TRINITY_DN13339_c0_g1_i6.p1	220-2727[+]	sp XSJ851 SECAB_ARAT	PF07393.11*Sec1	ExpAA=154.31*Pre	GO:0006887*biological_process
TRINITY_DN13339_c0_g1_i1	sp XSJ851 SECAB_ARAT	TRINITY_DN13339_c0_g1_i1.p1	220-2727[+]	sp XSJ851 SECAB_ARAT	PF07393.11*Sec1	ExpAA=154.31*Pre	GO:0006887*biological_process
TRINITY_DN13385_c0_g1_i1	sp Q5N870 DCL3A_ORY	TRINITY_DN13385_c0_g1_i1.p1	222-5276[-]	sp Q5N870 DCL3A_ORY	PF00270.29*DEAD	ExpAA=154.54*Pre	GO:0003676*molecular_function
TRINITY_DN13320_c0_g1_i2	.	TRINITY_DN13320_c0_g1_i2.p1	418-3582[+]	.	PF11145.8*DUF25	ExpAA=154.31*Pre	GO:0003676*molecular_function
TRINITY_DN13320_c0_g1_i1	.	TRINITY_DN13320_c0_g1_i1.p1	25-3408[+]	.	PF11145.8*DUF25	ExpAA=154.31*Pre	GO:0003676*molecular_function
TRINITY_DN13366_c0_g1_i1	sp Q8RWZ5 SD25_ARAT	TRINITY_DN13366_c0_g1_i1.p1	1-2646[+]	sp Q8RWZ5 SD25_ARAT	PF01453.24*B	ExpAA=57.32*Pre	GO:0004672*molecular_function
TRINITY_DN13358_c0_g1_i1	sp Q93XK2 STSYN_PEA	TRINITY_DN13358_c0_g1_i1.p1	338-2584[+]	sp Q93XK2 STSYN_PEA	PF05691.12*Raffi	ExpAA=22.98*Pre	GO:0005524*molecular_function
TRINITY_DN13380_c0_g1_i1	sp Q8LBB7 LOG5_ARAT	TRINITY_DN13380_c0_g1_i1.p1	633-3536[-]	sp Q8LBB7 LOG5_ARAT	PF04576.15*Zei	ExpAA=22.98*Pre	GO:0005524*molecular_function
TRINITY_DN13380_c0_g1_i2	sp Q8LBB7 LOG5_ARAT	TRINITY_DN13380_c0_g1_i2.p1	633-3446[-]	sp Q8LBB7 LOG5_ARAT	PF04576.15*Zei	ExpAA=22.98*Pre	GO:0005524*molecular_function
TRINITY_DN13335_c0_g1_i1	sp Q65440 BAME3_ARA	TRINITY_DN13335_c0_g1_i1.p1	252-3215[+]	sp Q65440 BAME3_ARA	PF08263.12*LRRN	ExpAA=35.52*Pre	GO:0005515*molecular_function
TRINITY_DN9436_c0_g1_i12	sp Q55FG3 VP13C_DICD	TRINITY_DN9436_c0_g1_i12.p1	250-9705[+]	sp Q55FG3 VP13C_DICD	PF06650.12*SHR	ExpAA=35.52*Pre	GO:0005515*molecular_function
TRINITY_DN9436_c0_g1_i16	sp Q55FG3 VP13C_DICD	TRINITY_DN9436_c0_g1_i16.p1	250-9702[+]	sp Q55FG3 VP13C_DICD	PF06650.12*SHR	ExpAA=35.52*Pre	GO:0005515*molecular_function
TRINITY_DN9436_c0_g1_i11	sp Q55FG3 VP13C_DICD	TRINITY_DN9436_c0_g1_i11.p1	250-9675[+]	sp Q55FG3 VP13C_DICD	PF06650.12*SHR	ExpAA=35.52*Pre	GO:0005515*molecular_function
TRINITY_DN9436_c0_g1_i5	sp Q55FG3 VP13C_DICD	TRINITY_DN9436_c0_g1_i5.p1	250-9678[+]	sp Q55FG3 VP13C_DICD	PF06650.12*SHR	ExpAA=35.52*Pre	GO:0005515*molecular_function
TRINITY_DN9436_c0_g1_i15	sp O42926 VP13B_SCHP	TRINITY_DN9436_c0_g1_i15.p1	250-8814[+]	sp O42926 VP13B_SCHP	PF06650.12*SHR	ExpAA=35.52*Pre	GO:0005515*molecular_function
TRINITY_DN9436_c0_g1_i14	sp O42926 VP13B_SCHP	TRINITY_DN9436_c0_g1_i14.p1	1348-8931[+]	sp O42926 VP13B_SCHP	PF06650.12*SHR	ExpAA=35.52*Pre	GO:0005515*molecular_function
TRINITY_DN9436_c0_g1_i8	sp Q55FG3 VP13C_DICD	TRINITY_DN9436_c0_g1_i8.p1	250-9675[+]	sp Q55FG3 VP13C_DICD	PF06650.12*SHR	ExpAA=35.52*Pre	GO:0005515*molecular_function
TRINITY_DN9436_c0_g1_i1	sp QOWQJ7 KN5B_ARA	TRINITY_DN9436_c0_g1_i1.p1	175-3321[+]	sp QOWQJ7 KN5B_ARA	PF16796.5*Micro	ExpAA=35.52*Pre	GO:0008017*molecular_function
TRINITY_DN9455_c0_g1_i6	sp P42730 CLP1_ARAT	TRINITY_DN9455_c0_g1_i6.p1	247-2988[-]	sp P42730 CLP1_ARAT	PF02861.20*Clp	ExpAA=35.52*Pre	GO:0005524*molecular_function
TRINITY_DN9455_c0_g1_i5	sp P42730 CLP1_ARAT	TRINITY_DN9455_c0_g1_i5.p1	274-3015[-]	sp P42730 CLP1_ARAT	PF02861.20*Clp	ExpAA=35.52*Pre	GO:0005524*molecular_function
TRINITY_DN9455_c0_g1_i3	sp P42730 CLP1_ARAT	TRINITY_DN9455_c0_g1_i3.p1	236-2977[-]	sp P42730 CLP1_ARAT	PF02861.20*Clp	ExpAA=35.52*Pre	GO:0005524*molecular_function
TRINITY_DN9455_c0_g1_i4	sp P42730 CLP1_ARAT	TRINITY_DN9455_c0_g1_i4.p1	236-2977[-]	sp P42730 CLP1_ARAT	PF02861.20*Clp	ExpAA=35.52*Pre	GO:0005524*molecular_function
TRINITY_DN9455_c0_g1_i1	sp P42730 CLP1_ARAT	TRINITY_DN9455_c0_g1_i1.p1	163-2904[-]	sp P42730 CLP1_ARAT	PF02861.20*Clp	ExpAA=35.52*Pre	GO:0005524*molecular_function
TRINITY_DN9489_c0_g1_i7	sp Q8SF2 FBL11_ARAT	TRINITY_DN9489_c0_g1_i7.p1	237-3155[-]	sp Q8SF2 FBL11_ARAT	PF00651.31*BTB	ExpAA=35.52*Pre	GO:0005515*molecular_function

Şekil 4.15 Trinotate.xls dosyası içeriği

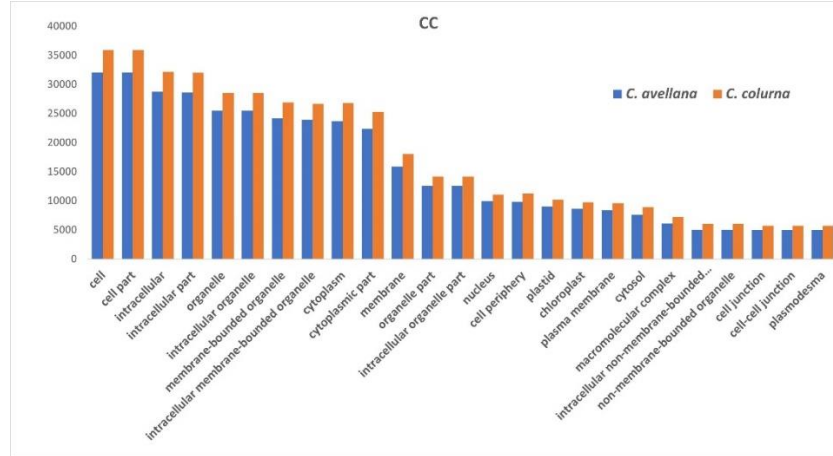
4.2.10 Gen Ontoloji ve Yolak Analizi

Türk Fındığı ve Tombul Fındığa ait transkriptlerin moleküler ve biyolojik fonksiyonlarını araştırmak için GO analizi ile biyolojik proses, moleküler fonksiyon ve hücresel komponent şeklinde 3 grupta kategorize edilmiştir. *C. colurna* L. ve *C. avellana* L. Tombul transkriptleri için GO ve KEGG zenginleştirme analizi TRAPID web yazılımıyla p değeri 0.05 olacak şekilde değerlendirilmiştir. Şekil 4.16'de her kategori için ilk 15 terime yer verilmiştir. Şekil 4.17'de ise KEGG yolak analizi sonucundaki ilk 15 terim gösterilmiştir.

A



B



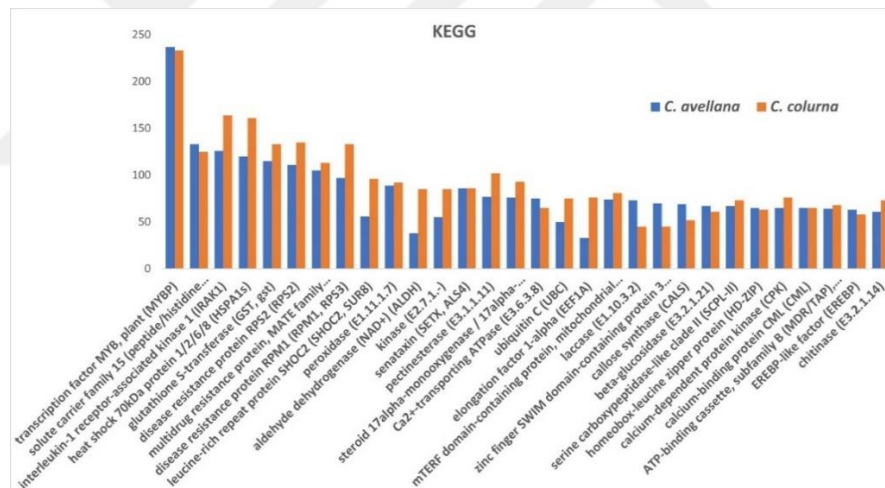
Şekil 4.16. *C. colurna* L. ve *C. avellana* L. Tombul transkriptlerinin GO kategorilerine göre sınıflandırılması A) Biyolojik Proses (BP) B) Hücresel Komponent (CC) C) Moleküler Fonksiyon (MF)

MF

■ *C. avellana* ■ *C. colurna*

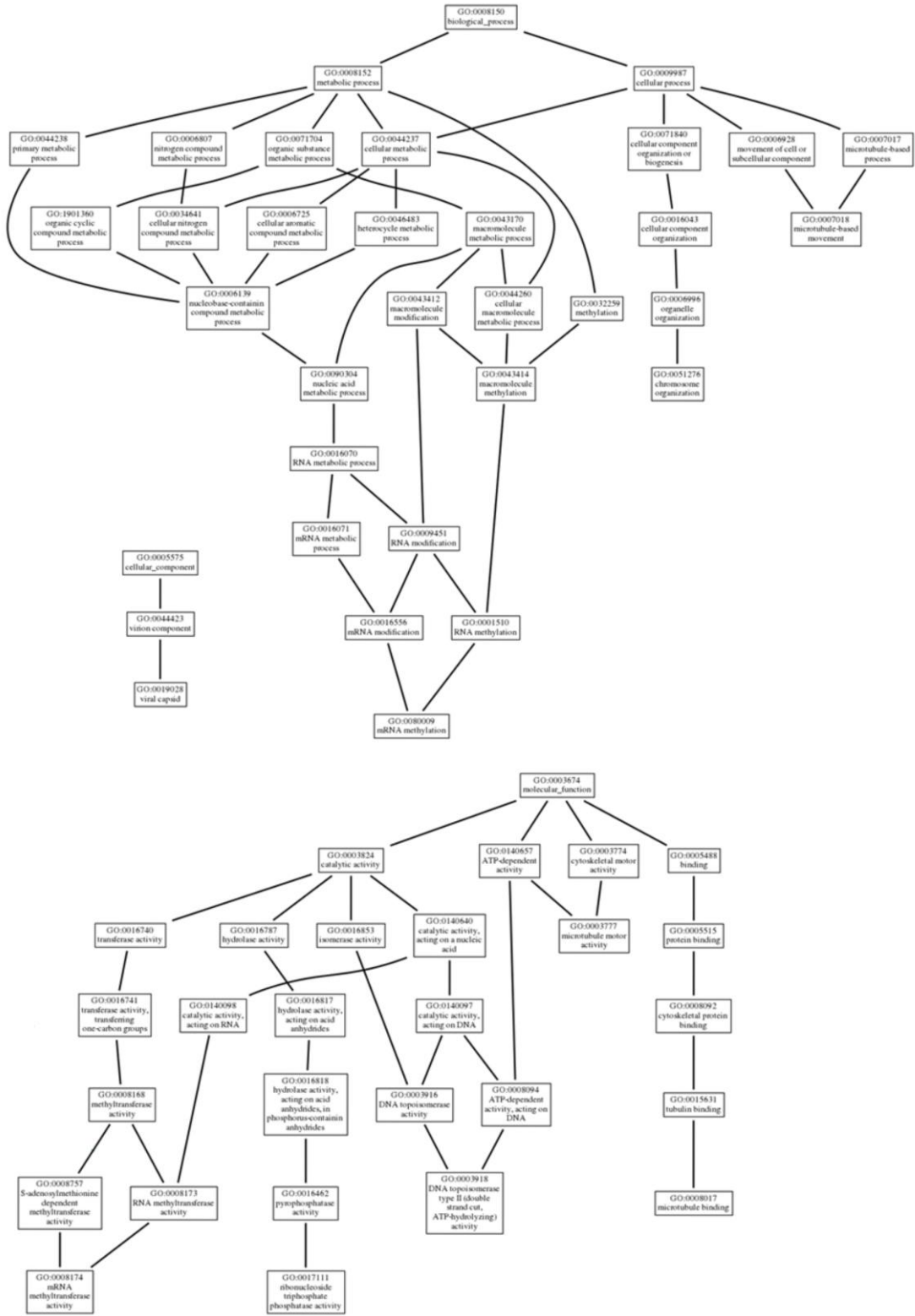
Molecular Function	<i>C. avellana</i>	<i>C. colurna</i>
binding	33000	38000
catalytic activity	21000	24000
protein binding	18000	21000
organic cyclic compound binding	18000	21000
heterocyclic compound binding	18000	21000
ion binding	15000	17000
small molecule binding	10000	12000
anion binding	10000	12000
nucleotide binding	10000	12000
transferase activity	9000	11000
nucleic acid binding	9000	11000
carbohydrate derivative binding	9000	11000
ribonucleotide binding	9000	11000
purine nucleotide binding	9000	11000
adenyl nucleotide binding	9000	11000
purine ribonucleoside binding	9000	11000
triphosphate binding	9000	11000
hydrolyase activity	8000	10000
ATP binding	8000	10000
cation binding	7000	9000
metal ion binding	7000	9000
transferase activity, transferring—	6000	8000
DNA binding	5000	7000

Komponent (CC) C) Moleküler Fonksiyon (MF) (



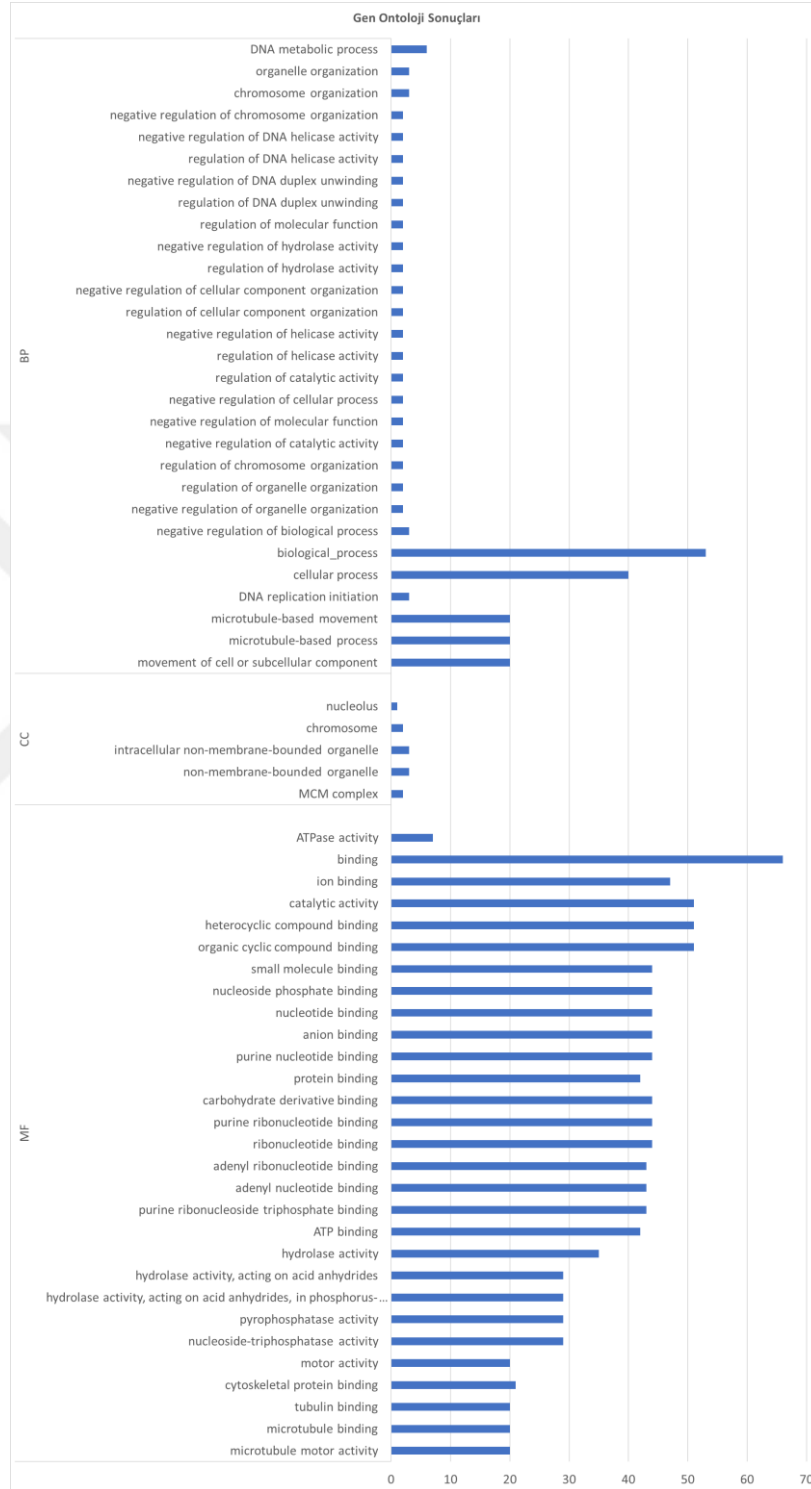
analizi sonuçları

C. colurna L.'nin diferansiyel olarak ifade edilen genlerinin biyolojik proses ve moleküler fonksiyon kategorilerindeki GO terimlerinin ilişki ağı Şekil 4.18'de verilmiştir.



Şekil 4.19 *C. avellana* L. Tombul'un diferansiyel olarak ifade edilen genlerinin GO terimleri ilişki ağı

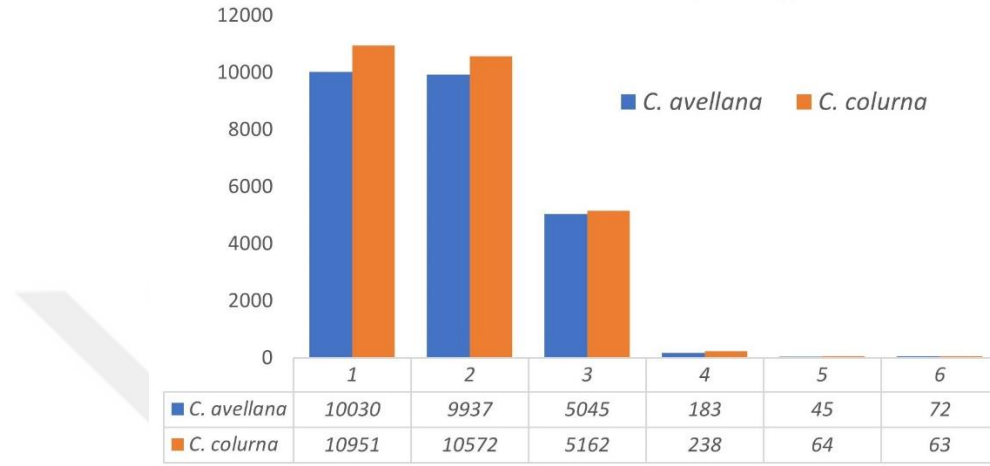
C. colurna L. ve *C. avellana* L. Tombul'un diferansiyel olarak ifade edilen genlerinin biyolojik proses, hücresel komponent ve moleküler fonksiyon kategorilerindeki dağılımları Şekil 4.20'de verilmiştir.



Şekil 4.20 *C. colurna* L. ve *C. avellana* L. Tombul'un diferansiyel olarak ifade edilen genlerinin GO kategorilerindeki dağılımları

4.2.11 SSR Analizi Sonuçları

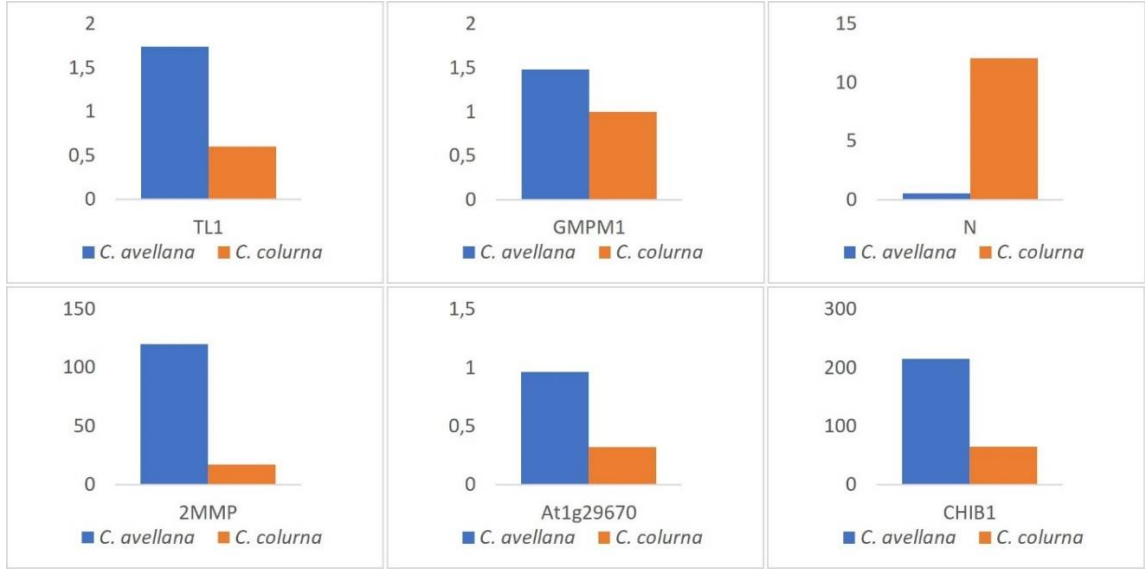
Mikrosatellit varlığını tespit etmek için kullanılan MISA yazılımı ile *C. colurna* L. ve *C. avellana* L. Tombul kütüphanelerinde ayrı ayrı gerçekleştirilen tarama sonucunda Tombul Fındık için 25.312, Türk Fındığı için 27.051 SSR motifinin varlığı belirlenmiştir. Farklı tekrar tiplerinin dağılımı Şekil 4.19'da gösterilmiştir.



Şekil 4.21 Transkript başına tanımlanan toplam SSR motifi sayıları dağılımı

4.3 qPCR Validasyon Sonuçları

RNA-seq'in verilerinin doğruluğu ve tekrarlanabilirliğini onaylamak için bitki gelişimi, direnç geliştirme, lipid metabolizması ve savunma yanıtı ile ilgili TL1 (Thaumatococcus-like protein 1), GMPM1 (18 kDa tohum olgunlaşma proteini), N (direnç proteini N), 2MMP (Metalloproteinaz 2-MMP), At1g29670 (GDSL esterase/lipaz) ve CHIB1 (Asidik endokitinaz) genleri qPCR ile validiye edilmiştir. Şekil 4.20'de gösterildiği gibi TL1, GMPM1, 2MMP, At1g29670, CHIB1 genleri *C. avellana* L. Tombul'da *C. colurna* L.'ya göre daha fazla eksprese olmuştur. N geninin ise *C. colurna* L.'da ifadesi *C. avellana* L. Tombul'a göre daha yüksek olduğu tespit edilmiştir.



Şekil 4.22 TL1, GMPM1, N, 2MMP, At1g29670, CHIB1 genlerinin qPCR ekspresyon seviyeleri

4.4 RNA-seq Verilerinin NCBI SRA Veritabanına Yüklenmesi

Farklı araştırmalara kaynaklık etmesi ve çalışmaların genişletilebilmesi için verilere erişilebilirlik önemlidir. Bu bağlamda NCBI SRA veritabanı önemli bir görev üstlenmiştir. Yüksek verimli dizileme verilerinin arşivi niteliğinde olan bu veritabanı sorgulanabilirliği ve kullanılabilirliği ile en değerli veri kaynaklarından biridir.

Bu açıdan çalışmanın verilerinin farklı araştırmalara da kaynaklık edebilmesi için *C. colurna* L. ve *C. avellana* L. Tombul'a ait RNA-seq dizileme bilgileri SRR19139693 ve SRR19139694 erişim numaralarıyla NCBI SRA veritabanına sunulmuştur.

4.5 Tartışma ve Öneriler

Halk arasında fındık olarak anılan *Corylus* cinsi bitkiler, dünyanın birçok yerinde ekonomiye katkıları, gıda olarak tüketilmeleri, peyzaj bitkisi olarak kullanımları, çikolata ve tatlı sektöründeki yerleri ve sağlık alanındaki faydalarından dolayı yeri doldurulamaz bir konumdadır. *C. avellana* L. (Avrupa Fındığı), *Corylus* türleri arasında en yaygın ve ekonomik açıdan en önemli türü olup, Türkiye'nin Karadeniz bölgesinde yaygın olarak yayılış göstermektedir [155]. *C. avellana*

L.'nin “Tombul”, “Çakıldak”, “Jefferson” ve “Palaz” gibi birçok çeşidi vardır [16], [24]. *C. avellana* L. Tombul Fındık çeşidi kalite açısından diğerlerine göre oldukça üstün olarak kabul edilmektedir [156]. Bunun yanı sıra Türkiye’de doğal olarak yetiştirilen *C. colurna* L. (Türk Fındığı) da anaç olarak kullanımı, uzun yıllar boyunca verimli olarak yaşamını sürdürmesi, dayanıklı kök yapısı, zorlu stres ve doğa koşulları dayanımıyla önemli fındık türlerinden biridir [22].

Son yıllarda bitki moleküler çalışmalarında, moleküler mekanizmaları aydınlatabilmek ve bu alanda yeni bilgilere ulaşabilmek için *de novo* dizileme teknolojisi sıklıkla kullanılmaktadır. [157]–[159]. Çoğunlukla referans genoma sahip olmayan türlerin transkriptom çalışmalarında kullanılan *de novo* derleme protokolleri yeni keşiflerin önünü açmıştır ve bu yaklaşım model olmayan organizmaları inceleyen araştırmacıların tercihi haline gelmiştir.

C. avellana L. için yapılan genomik çalışmalara bakıldığında dikkat çeken birkaç kaynak mevcuttur. *C. avellana* Jefferson, Çakıldak ve Palaz çeşitleri için çeşitli genomik çalışmalar yapılmıştır. Avrupa fındığı "Jefferson" ile transkriptom çalışması yapılması için bitkinin farklı organlarından örnekler alınarak RNA'ları izole edilmiş ve 28.255 kontig elde edilmiştir [16]. *C. avellana* L.'nin olumsuz koşullar altındaki tomurcuk patlaması sürecini incelemek üzere RNA-seq çalışması yürütülmüştür [24]. Ancak, *C. avellana* L. Tombul çalışmaları Tombul Fındığın kloroplast genomu üzerine odaklanıldığı bir çalışma [26] ve içerisinde Tombul Fındığın da bulunduğu 78 fındık çeşidinin SSR profillerinin çıkarılarak veritabanı oluşturulmaya çalışıldığı bir çalışma [160] ile sınırlıdır. Bununla birlikte, *C. colurna* L. ile ilgili henüz herhangi bir genom çalışması yayınlanmamıştır.

Bu çalışmada sırasıyla, *C. colurna* L. ve *C. avellana* L. Tombul’un genç yapraklarından, erkek ve dişi çiçeklerinden, tomurcuklarından ve zuruflarından alınan örnekler havuz haline getirilmiş, RNA-seq dizileme teknolojisi kullanılarak dizilenmiş ve *de novo* derlemesi ile kapsamlı bir RNA-seq veri seti oluşturulmuştur. Türk Fındığı ve Tombul Fındığın transkriptom verilerinin derleme sonrasında bolluk tahmini, derlemenin kalitesinin belirlenmesi, veri kalitelerinin değerlendirilmesi, diferansiyel ifade analizi, işlevsel açıklama eklenmesi, gen ontoloji ve yolak analizleri ve SSR motiflerinin belirlenmesi

işlemleri yapılmış qPCR ile validasyon adımı sonrasında RNA-seq verilerinin NCBI, SRA veritabanına yüklenmesiyle çalışma tamamlanmıştır.

Tez çalışması ile, RNA-seq dizileme ve *de novo* derlemesi ile birlikte 95.979 transkripti ve 57.698 unigene elde edilmiştir. Verilerin GC yüzdesi 43,59, N50 değeri 1929, ortalama kontig uzunluğu 1116,82 ile derleme kalitesinin iyi olduğu görülmüştür (Tablo 4.3). Okumaların temsil oranı ortalaması ise %98,63 tespit edilmiştir (Tablo 4.4). Bu sonuç ile derlemenin sağlıklı olması için gerekli olan %70 ve üzeri uyumluluk oranının çok üstü bir okuma temsil oranı yakalandığı anlaşılmıştır. İyi kalitede bir derlemenin ExN50 değerinin %90 olması beklenirken bu çalışmada tepe zirvesi %92 ile ortalamanın üstüne çıkmıştır (Şekil 4.7).

Tam uzunluktaki kodlayan transkriptlerin değerlendirilmesi için SWISSPROT veritabanında e-değeri $1e-20$ olacak şekilde sorgusunda, veritabanındaki 7268 proteinin %90'dan fazla eşleştiği, 1456 proteinin %80-90 arasında eşleşme gösterdiği belirlenmiştir (Tablo 4.5).

C. colurna L. ve *C. avellana* L. Tombul'un türe özgü veri kalitelerinin incelenmesi amacıyla hizalanan okuma sayıları ve transkript başına düşen okuma sayısı ile değerlendirmeler yapılmış ve iki türün verilerinin ayrışıkları görülmüştür (Şekil 4.8).

İki fındık türü arasında diferansiyel olarak ifade edilen transkript ve genleri tespit etmek amacıyla EdgeR kullanılmış, P değeri 0,001 ve kat değişikliği >4 olacak şekilde değerlendirme yapıldığında türler arasında farklı ifade edilen transkript sayısının 1528, gen sayısının 736 bulunmuştur. Bunlardan, *C. colurna* L.'de ifadesinde artış görülen transkript sayısının 652 ve ifadesinde artış görülen gen sayısının 297 olduğu, *C. avellana* L. Tombul'da ise ifadesinde artış görülen transkript sayısının 876 olduğu ve gen sayısının 439 olduğu belirlenmiştir (Tablo 4.6).

Transkriptlerin özelliklerinin belirlenmesi amacıyla bir dizi işlem yapılmış ve görevlerinin aydınlatılmasını sağlayacak açıklamalar eklenmiştir. Öncelikle uzun açık okuma çerçeveleri tespit edilmiş, kodlama bölgesinin hangi dizide olduğu ve bu transkript dizisindeki koordinatları belirlenmiştir. Mevcut transkriptler içerisindeki ORF sorgu sonucuna göre 4523 adet kodlayan bölgenin olduğu tespit

edilmiştir. Sinyal peptidlerinin tespiti için SignalP kullanılarak tahmin edilen 336 sinyal peptidi işaretlenmiştir. Transmembran proteinlerinin belirlenmesi amacıyla tmhmm çalıştırılmış ve en az 1 sarmala sahip transmembran proteini sayısı 887 olarak tespit edilmiş, bu proteinlerin sarmal sayısı, boyutu, zardaki konumu gibi bilgiler de not edilmiştir.

Transkript ve genlere ait SWISSPROT kayıtları, pfam bilgileri, farklı ifade edilen gen ve transkriptler, sinyal peptit bilgileri, transmembran proteinlerine ait bilgiler gibi elde edilen tüm işlevsel açıklamalar ve diğer veriler ile bir veritabanı oluşturulmuş, sonrasında xls dosyası ile sorgulanabilir hale getirilmiştir (Şekil 4.15).

C. colurna L. ve *C. avellana* L. Tombul'un kendi derlemelerinden elde genlerin biyolojik fonksiyonunu belirlemek için GO ve KEGG yolak analizleri yapılmıştır. GO analizi; biyolojik süreçler, hücresel komponentler ve moleküler fonksiyon terimleri olarak kategorize edilmiştir. GO sonuçlarına göre, *C. avellana* L. Tombul ve *C. colurna* L.'nin kendi derlemelerinden analiz edilen genleri, BP, CC ve MF'de çoğunlukla benzer GO terimlerinde anote edilmiştir (Şekil 4.16). Biyolojik proses incelendiğinde, genlerin hücresel ve metabolik proseslerin, organik substans metabolik proseslerin, hücresel metabolik ve primer metabolik prosesler başlıkları altında toplandığı görülmüştür. Hücresel komponent kategorisinde ise hücre, hücre bileşenleri, intraselüler ve intraselüler organellerde yoğunlaştığı görülmüştür. Moleküler fonksiyon kategorisinde ise, genlerin bağlanma, katalitik aktivite, protein bağlanma, organik siklik komponent bağlama ve iyon bağlama kategorilerinde anote olduğu görülmüştür.

C. colurna L. ve *C. avellana* L. Tombul'un diferansiyel olarak ifade edilen genlerinin biyolojik proses, hücresel komponent ve moleküler fonksiyon kategorilerindeki dağılımlarına bakıldığında, biyolojik süreçler kategorisinde, hücresel ve biyolojik süreçler, mikrotübül tabanlı hareket ve süreçler, hücre veya hücre altı komponentlerin taşınması, DNA metabolik süreçleri ve biyolojik sürecin negatif düzenlenmesi terimlerinin öne çıktıkları; hücresel komponentler kategorisinde, hücre içi zarla çevrili olmayan organel, zarla çevrili olmayan organel, kromozom, çekirdekçik ve MCM kompleksi terimlerinde öne çıktıkları;

moleküler fonksiyon kategorisinde ise binding, iyon binding, katalitik aktivite, heterosiklik ve organik siklik bileşikleri bağlama terimlerinde anote oldukları görülmüştür (Şekil 4.20).

KEGG yolak analizlerinde ise; *C. colurna* L. genleri transkripsiyon faktörü MYB plant, peptit/histidin taşıyıcısı, interlökin-1 reseptörü ile ilişkili kinaz 1, ısı şoku 70kDa proteini, glutatyon S-transferaz, hastalık direnç proteini RPS2, çoklu ilaca direnç proteini RPM1, pektinesteraz ve lösün bakımından zengin tekrar proteini yolaklarında yer almıştır. *C. avellana* L. genleri ise, transkripsiyon faktörü MYB Plant, peptit/histidin taşıyıcı, interlökin-1 reseptörü ile ilişkili kinaz 1, ısı şoku 70kDa proteini, glutatyon S-transferaz, hastalık direnç proteini RPS2, çoklu ilaç direnç proteini, MATE ailesi, hastalık direnç proteini RPM1, peroksidaz ve senataxin KEGG yolaklarında yer almıştır (Şekil 4.17).

Ayrıca bu çalışmada *C. colurna* L. ve *C. avellana* L. Tombul'un transkriptom verilerinden MISA programı kullanılarak sırasıyla toplam 27051 ve 25312 SSR motifi tanımlanmıştır. Bu, *C. colurna* L.'den SSR motiflerinin ilk keşfidir. *C. avellana* L. Tombul ve *C. colurna* L.'nin tüm dizileme verilerinde SSR'lar karakterize edilerek, farklı iskele/kontig bölgelerinde SSR'ların sıklıkları ve dağılımları analiz edilmiştir (Şekil 4.21). *C. colurna* L. ve *C. avellana* L. Tombul'daki SSR motiflerinin yaklaşık olarak %80'i di-nükleotid ve tri-nükleotid olduğu görülmüştür.

Transkriptom sonuçlarını deneysel olarak doğrulamak için, gelişim, savunma cevabı ve lipid metabolizması ile ilgili 6 genin qPCR ile ekspresyon seviyelerine bakılmıştır. Sonuçlara göre TL1 (Thaumatin-like protein 1), GMPM1 (18 kDa tohum olgunlaşma proteini), 2MMP (Metalloproteinaz 2-MMP), At1g29670 (GDSL esteraz/lipaz) ve CHIB1 (Asidik endokitinaz) genlerinin ekspresyonlarının *C. avellana* L. Tombul'da *C. colurna* L.'ye göre daha fazla olduğu görülmüştür. Öte yandan, N (direnç proteini N) geninin ifadesinin ise *C. colurna* L.'de daha fazla olduğu görülmüştür (Şekil 4.22). qPCR sonuçları, transkriptom sonuçlarından elde edilen verilerle ile karşılaştırıldığında sonuçların uyumlu olduğu saptanmıştır.

Bu tez çalışmasının verilerinin farklı araştırmalara da kaynaklık edebilmesi için *C. colurna* L. ve *C. avellana* L. Tombul'a ait RNA-seq dizileme bilgileri SRR19139693 ve SRR19139694 erişim numaralarıyla NCBI SRA veritabanına sunulmuştur.

C. colurna L. ve *C. avellana* L. Tombul'un *de novo* transkriptom analizinin yapıldığı bu çalışma, dünyada ekonomik açıdan önemli olan iki fındık türünün gen ekspresyon profilleri hakkında yapılmış önemli bir çalışmadır. Bu çalışma, *C. colurna* L.'nin gen ekspresyon profili çıkarılarak NCBI veritabanına gen bilgilerinin eklendiği ilk çalışmadır. Ayrıca bu çalışmanın devamında gerçekleştirilen *C. avellana* L. Tombul ve *C. colurna* L.'deki korunmuş miRNA'ların keşfine yönelik çalışma ile [161] daha ileri seviyeye taşınmıştır.

Bu çalışma ile, yeni gen keşifleri, genlerin fonksiyonel açıklaması ve bitkide biyobelirteç keşfinde önemli olan SSR'ların tespitinin bu alanda çalışan bilim insanlarına fayda sağlayacağını düşünmekteyiz.

- [1] M. Genç, Ş. T. Güner, S. Gülcü, and H. Fakir, “Afyon-Dereçine Türk Fındığı (Corylus colurna L.) Bükü,” *Orman ve Av*, vol. 74, no. 6, pp. 13–18, Jun. 1998.
- [2] S. Polat, “Türk Fındığı’nın (Corylus Colurna) Türkiye’deki Yeni Bir Yayılış Alanı,” *Marmara Coğrafya Dergisi*, vol. 0, no. 29, Dec. 2014, doi: 10.14781/mcd.51065.
- [3] M. M. Thompson, “Breeding for filbert varieties suitable for shelling,” *Proc. Nut Growers Soc. OR, WA, and BC*, vol. 67, pp. 35–42, 1982.
- [4] M. Ayfer, A. Uzun, and F. Baş, *Türk fındık çeşitleri*. Ankara: Karadeniz Bölgesi Fındık İhracatçılar Birliği Yayınları, 1986.
- [5] E. Pelvan, C. Alasalvar, and S. Uzman, “Effects of Roasting on the Antioxidant Status and Phenolic Profiles of Commercial Turkish Hazelnut Varieties (Corylus avellana L.),” *Journal of Agricultural and Food Chemistry*, vol. 60, no. 5, pp. 1218–1223, Feb. 2012, doi: 10.1021/jf204893x.
- [6] E. Riethmüller, G. Tóth, A. Alberti, M. Sonati, and A. Kéry, “Antioxidant Activity and Phenolic Composition of Corylus colurna,” *Nat Prod Commun*, vol. 9, pp. 679–682, Apr. 2014.
- [7] U. Galderisi, M. Cipollaro, G. di Bernardo, L. de Masi, G. Galano, and A. Cascino, “Identification of hazelnut (Corylus avellana) cultivars by RAPD analysis,” *Plant Cell Reports*, vol. 18, no. 7–8, pp. 652–655, Mar. 1999, doi: 10.1007/s002990050637.
- [8] N. V. Bassil and A. N. Azarenko, “Rapid Markers For Self-Incompatibility In Corylus Avellana L.,” *Acta Horticulturae*, no. 556, pp. 537–544, Jul. 2001, doi: 10.17660/ActaHortic.2001.556.78.

- [9] M. L. Miaja, R. Vallania, C. Me, A. Akkak, O. Nassi, and G. Lepori, "Varietal Characterization In Hazelnut By Rapd Markers," *Acta Horticulturae*, no. 556, pp. 247–250, Jul. 2001, doi: 10.17660/ActaHortic.2001.556.35.
- [10] T. Gökirmak, S. A. Mehlenbacher, and N. V. Bassil, "Investigation Of Genetic Diversity Among European Hazelnut (*Corylus Avellana*) Cultivars Using Ssr Markers," *Acta Horticulturae*, no. 686, pp. 141–149, Jul. 2005, doi: 10.17660/ActaHortic.2005.686.18.
- [11] T. Gökirmak, S. A. Mehlenbacher, and N. v. Bassil, "Characterization of European hazelnut (*Corylus avellana*) cultivars using SSR markers," *Genetic Resources and Crop Evolution*, vol. 56, no. 2, pp. 147–172, Mar. 2009, doi: 10.1007/s10722-008-9352-8.
- [12] M. L. Metzker, "Sequencing technologies — the next generation," *Nature Reviews Genetics*, vol. 11, no. 1, pp. 31–46, Jan. 2010, doi: 10.1038/nrg2626.
- [13] J. Zhang, R. Chiodini, A. Badr, and G. Zhang, "The impact of next-generation sequencing on genomics," *Journal of Genetics and Genomics*, vol. 38, no. 3, pp. 95–109, Mar. 2011, doi: 10.1016/j.jgg.2011.02.003.
- [14] B. J. Haas *et al.*, "De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis," *Nature Protocols*, vol. 8, no. 8, pp. 1494–1512, Aug. 2013, doi: 10.1038/nprot.2013.084.
- [15] S. Zenoni *et al.*, "Characterization of Transcriptional Complexity during Berry Development in *Vitis vinifera* Using RNA-Seq," *Plant Physiology*, vol. 152, no. 4, pp. 1787–1795, Apr. 2010, doi: 10.1104/pp.109.149716.
- [16] E. R. Rowley *et al.*, "Assembly and characterization of the European hazelnut 'Jefferson' transcriptome," *Crop Science*, vol. 52, no. 6, pp. 2679–2686, Nov. 2012, doi: 10.2135/cropsci2012.02.0065.
- [17] A. Grada and K. Weinbrecht, "Next-Generation Sequencing: Methodology and Application," *Journal of Investigative Dermatology*, vol. 133, no. 8, pp. 1–4, Aug. 2013, doi: 10.1038/jid.2013.248.

- [18] L. Venturini *et al.*, “De novotranscriptome characterization of *Vitis vinifera* cv. Corvina unveils varietal diversity,” *BMC Genomics*, vol. 14, no. 1, p. 41, Dec. 2013, doi: 10.1186/1471-2164-14-41.
- [19] H. Ma, Z. Lu, B. Liu, Q. Qiu, and J. Liu, “Transcriptome analyses of a Chinese hazelnut species *Corylus mandshurica*,” *BMC Plant Biology*, vol. 13, no. 1, p. 152, Dec. 2013, doi: 10.1186/1471-2229-13-152.
- [20] X. Chen *et al.*, “Transcriptome Sequencing and Identification of Cold Tolerance Genes in Hardy *Corylus* Species (*C. heterophylla* Fisch) Floral Buds,” *PLoS ONE*, vol. 9, no. 9, p. e108604, Sep. 2014, doi: 10.1371/journal.pone.0108604.
- [21] Y. Cheng, J. Liu, H. Zhang, J. Wang, Y. Zhao, and W. Geng, “Transcriptome Analysis and Gene Expression Profiling of Abortive and Developing Ovules during Fruit Development in Hazelnut,” *PLOS ONE*, vol. 10, no. 4, p. e0122072, Apr. 2015, doi: 10.1371/journal.pone.0122072.
- [22] S. Ayan, D. Aydınöz, E. N. Yer, and E. Ünal, “Turkish Filbert (*Corylus colurna* L.) a new distribution area in Northwestern Anatolia Forests: (Provinces of Müsellimler, Tunuslar in Ağlı, Kastamonu/Turkey),” *Biological Diversity and Conservation*, vol. 9, no. 1, pp. 128–135, 2016.
- [23] S. Ayan, E. Ünal, A. İslam, O. E. Sakıcı, and E. N. Yer, “Kastamonu yöresinde yetişen Türk fıındığının (*Corylus colurna* L.) yağ ve protein içeriği,” *Artvin Çoruh Üniversitesi Orman Fakültesi Dergisi*, pp. 1–7, Apr. 2018, doi: 10.17474/artvinofd.296580.
- [24] M. Kavas, A. Kurt Kızıldoğan, and H. İ. Balık, “Gene expression analysis of bud burst process in European hazelnut (*Corylus avellana* L.) using RNA-Seq,” *Physiology and Molecular Biology of Plants*, vol. 25, no. 1, pp. 13–29, Jan. 2019, doi: 10.1007/s12298-018-0588-2.
- [25] S. J. Lucas, K. Kahraman, B. Avşar, R. J. A. Buggs, and I. Bilge, “A chromosome-scale genome assembly of European hazel (*Corylus avellana* L.) reveals targets for crop improvement,” *The Plant Journal*, vol. 105, no. 5, pp. 1413–1430, Mar. 2021, doi: 10.1111/tpj.15099.

- [26] K. Kahraman and S. J. Lucas, "Comparison of different annotation tools for characterization of the complete chloroplast genome of *Corylus avellana* cv Tombul," *BMC Genomics*, vol. 20, no. 1, p. 874, 2019, doi: 10.1186/s12864-019-6253-5.
- [27] S. Özbek, "Fındık Yetiştiriciliği," *Çukurova Üniversitesi Ziraat Fakültesi Yayınları*, vol. 128, pp. 286–321, 1978.
- [28] S. A. Özkurt, "Fındık ekimi, bakımı, fındıklara zarar veren böcekler mücadelesi, hastalıkları, tedavisi ve fındığın ekonomideki durumu.," *Tarım Bakanlığı Neşriyat Müdürlüğü*, vol. 676, p. 80, 1950.
- [29] A. İslam, "Ordu ili merkez ilçede yetiştirilen fındık çeşitlerinde klon seleksiyonu," Çukurova Üniversitesi Fen Bilimleri Enstitüsü, Adana, 2000.
- [30] Thompson MM, Langersted HB, and Mehlenbacher AS, "Hazelnuts.," *Fruit Breeding*, vol. 3, no. 3, pp. 125–184, 1996.
- [31] FAOSTAT, "<https://www.fao.org/faostat/en/#data/QCL>," 2021.
- [32] E. Pelvan, E. Ö. Olgun, A. Karadağ, and C. Alasalvar, "Phenolic profiles and antioxidant activity of Turkish Tombul hazelnut samples (natural, roasted, and roasted hazelnut skin)," *Food Chemistry*, vol. 244, pp. 102–108, Apr. 2018, doi: 10.1016/j.foodchem.2017.10.011.
- [33] S. K. Chang, C. Alasalvar, B. W. Bolling, and F. Shahidi, "Nuts and their co-products: The impact of processing (roasting) on phenolics, bioavailability, and health benefits – A comprehensive review," *Journal of Functional Foods*, vol. 26, pp. 88–122, Oct. 2016, doi: 10.1016/j.jff.2016.06.029.
- [34] C. Alasalvar *et al.*, "Turkish Tombul Hazelnut (*Corylus avellana* L.). 2. Lipid Characteristics and Oxidative Stability," *Journal of Agricultural and Food Chemistry*, vol. 51, no. 13, pp. 3797–3805, Jun. 2003, doi: 10.1021/jf021239x.
- [35] C. Alasalvar and B. W. Bolling, "Review of nut phytochemicals, fat-soluble bioactives, antioxidant components and health effects," *British Journal of Nutrition*, vol. 113, no. S2, pp. S68–S78, Apr. 2015, doi: 10.1017/S0007114514003729.

- [36] S. Perna *et al.*, “Effects of Hazelnut Consumption on Blood Lipids and Body Weight: A Systematic Review and Bayesian Meta-Analysis,” *Nutrients*, vol. 8, no. 12, p. 747, Nov. 2016, doi: 10.3390/nu8120747.
- [37] J. Sabaté, “Nut Consumption and Blood Lipid Levels,” *Archives of Internal Medicine*, vol. 170, no. 9, p. 821, May 2010, doi: 10.1001/archinternmed.2010.79.
- [38] P. M. Kris-Etherton, F. B. Hu, E. Ros, and J. Sabaté, “The Role of Tree Nuts and Peanuts in the Prevention of Coronary Heart Disease: Multiple Potential Mechanisms,” *The Journal of Nutrition*, vol. 138, no. 9, pp. 1746S-1751S, Sep. 2008, doi: 10.1093/jn/138.9.1746S.
- [39] S. M. Mercanligil *et al.*, “Effects of hazelnut-enriched diet on plasma cholesterol and lipoprotein profiles in hypercholesterolemic adult men,” *European Journal of Clinical Nutrition*, vol. 61, no. 2, pp. 212–220, Feb. 2007, doi: 10.1038/sj.ejcn.1602518.
- [40] G. E. Fraser, “Nut consumption, lipids, and risk of a coronary event,” *Clinical Cardiology*, vol. 22, no. S3, pp. 11–15, Jul. 1999, doi: 10.1002/clc.4960221504.
- [41] Demir Taki and Beyhan Neriman, “Samsun İlinde Yetiştirilen Fındıkların Seleksiyonu Üzerine Bir Araştırma,” *Türk J. Agric For*, vol. 24, pp. 173–183, 1998.
- [42] A. Gençer and U. Özgül, “Yaygın fındık (*Corylus avellana* L.) odunundan soda yöntemi ile kâğıt hamuru üretim parametrelerinin belirlenmesi,” *Turkish Journal of Forestry / Türkiye Ormancılık Dergisi*, vol. 16, no. 2, Nov. 2015, doi: 10.18182/tjf.09112.
- [43] S. Doğanay, “Trabzon ilinde Fındık Tarımı,” *Doğu Coğrafya Dergisi*, vol. 10, no. 13, pp. 233–252, 2005.
- [44] H. Doğanay, “Türkiye Fındık Meyvacılığındaki Yeni Gelişmeler,” *Doğu Coğrafya Dergisi*, vol. 17, no. 27, pp. 1–22, 2012.
- [45] H. İ. Balık, G. Kızılcı, K. Ak, and A. Sezer, *Fındık Yetiştiriciliği*. Ankara: Gıda Tarım ve Hayvancılık Bakanlığı, 2012.

- [46] S. Marley, "Corylus avellana, lore of hazelnuts," <https://research.reading.ac.uk/herbarium/2020/12/05/adventbotany-2020-day-5-corylus-avellana-lore-of-hazelnuts/>, 2020.
- [47] N. Beyhan and F. Odabaş, "İklimsel Faktörlerin Fındıkta Verimlilik Üzerine Etkileri ve Yetiştiricilik Açısından Önemi," *Ondokuz Mayıs Üniversitesi Ziraat Fakültesi Dergisi*, vol. 11, pp. 177–188, Jan. 1996.
- [48] A. R. Demirbaş, *Fındık Tarımı*. Samsun: İl Tarım Müdürlüğü, 2010.
- [49] J.-W. Zong, T.-T. Zhao, Q.-H. Ma, L.-S. Liang, and G.-X. Wang, "Assessment of Genetic Diversity and Population Genetic Structure of Corylus mandshurica in China Using SSR Markers," *PLOS ONE*, vol. 10, no. 9, p. e0137528, Sep. 2015, doi: 10.1371/journal.pone.0137528.
- [50] F. Açıktürk, "Effects of geographical origin and variety on vitamin and mineral composition of hazelnut (Corylus avellana L.) varieties cultivated in Turkey," *Food Chemistry*, vol. 65, no. 3, pp. 309–313, May 1999, doi: 10.1016/S0308-8146(98)00201-5.
- [51] M. Şenyurt, "Corylus Colurna L. Anacına Bazı Fındık Çeşitlerinin Aşılabilirliğinin İncelenmesi," Abant İzzet Baysal Üniversitesi, Bolu, 1997.
- [52] T. H. Everett, *The New York Botanical Garden illustrated encyclopedia of horticulture*. NewYork & London: Garland Publishing, Inc. , 1988.
- [53] F. Yaltırık, *Dendroloji-Gymnospermae (Açık tohumlular)*. İstanbul: İstanbul Üniversitesi Orman Fakültesi, 1988.
- [54] H. Kayacık, *Orman ve Park Ağaçlarının Özel Sistematiği*, vol. 2. İstanbul: Orm. Fak Yay. No: 247, 1977.
- [55] S. Polat and Y. Güney, "Türk Fındığı'nın (Corylus Colurna) Türkiye'deki Yeni Bir Yayılış Alanı," *Akademik Sosyal Araştırmalar Dergisi*, vol. 3, no. 18, pp. 449–460, 2015.
- [56] S. Tosun, "Cadde (Yol) Ağac olarak Amerika'da ve Avrupa'da popülerleşen türk fındığı (Corylus colurna L.)," *Orman ve Av Dergisi, S*, vol. 3, pp. 22–25, 2012.

- [57] M. Arslan, P. Şenel, Z. Özpay Palazoğlu, and E. Çiçek, *Önemli Orman Ağacı Türlerimizden Türk Fındığı (Corylus Colurna L.) Tohumlarına Ga3 Uygulamasının Çimlenme Engelinin Giderilmesi Ve Fidan Büyümesine Etkisi*. Bolu: Orman ve Su İşleri Bakanlığı Batı Karadeniz Ormancılık Araştırma Enstitüsü, 2013.
- [58] M. Arslan, “Geçmişimizle Paylaştığımız Ve Geleceğe Miras Bırakmamız Gereken Doğal Türlerimizden Türk Fındığı (Corylus Colurna L.),” in *1 st international Non-Wood Forest Products Symposium*, 2006, pp. 302–311.
- [59] F. Yaltırık, *Dendroloji-Gymnospermae 2. Baskı*. İstanbul Üniversitesi Orman Fakültesi, 2000.
- [60] V. Erdogan, “Phylogenetic analysis of hazelnut species (Corylus, Corylaceae) based on morphology and phenology,” *Herb systematic botany*, Jan. 2002.
- [61] P. H. Davis, *Flora of Turkey and the East Aegean Islands I*. Edinburgh: Edinburgh University Press, 1982.
- [62] R. Anşın and Özkan Z.C, *Tohumlu Bitkiler. Odunsu Taksonlar*. Trabzon: Karadeniz Teknik Üniversitesi Basımevi, 1993.
- [63] F. Yaltırık, *Dendroloji Ders Kitabı II, Angiospermae*. İstanbul: İstanbul Üniversitesi Matbaası, 1993.
- [64] V. Erdogan and A. Aygün, “Fatty Acid Composition and Physical Properties of Turkish Tree Hazel Nuts,” *Chemistry of Natural Compounds*, vol. 41, pp. 378–381, Jul. 2005, doi: 10.1007/s10600-005-0156-1.
- [65] J. NinicTodorovic, “Biochemical composition of filbert seeds,” *Jugoslovensko Vocarstvo*, vol. 26, no. 3–4, pp. 23–30, Jan. 1992, [Online]. Available: <https://eurekamag.com/research/002/570/002570007.php>
- [66] F. Crick, “On protein synthesis.,” *Symp Soc Exp Biol*, vol. 12, pp. 138–63, 1958.
- [67] F. Crick, “Central Dogma of Molecular Biology,” *Nature*, vol. 227, no. 5258, pp. 561–563, Aug. 1970, doi: 10.1038/227561a0.

- [68] C. Bucciarelli and M. Selbach, “mRNAs, proteins and the emerging principles of gene expression control,” *Nature Reviews Genetics*, vol. 21, no. 10, pp. 630–644, Oct. 2020, doi: 10.1038/s41576-020-0258-4.
- [69] D. Holoch and D. Moazed, “RNA-mediated epigenetic regulation of gene expression,” *Nature Reviews Genetics*, vol. 16, no. 2, pp. 71–84, Feb. 2015, doi: 10.1038/nrg3863.
- [70] L. Statello, C.-J. Guo, L.-L. Chen, and M. Huarte, “Gene regulation by long non-coding RNAs and its biological functions,” *Nature Reviews Molecular Cell Biology*, vol. 22, no. 2, pp. 96–118, Feb. 2021, doi: 10.1038/s41580-020-00315-9.
- [71] P. Schimmel, “The emerging complexity of the tRNA world: mammalian tRNAs beyond protein synthesis,” *Nature Reviews Molecular Cell Biology*, vol. 19, no. 1, pp. 45–58, Jan. 2018, doi: 10.1038/nrm.2017.77.
- [72] J. Li and C. Liu, “Coding or Noncoding, the Converging Concepts of RNAs,” *Frontiers in Genetics*, vol. 10, May 2019, doi: 10.3389/fgene.2019.00496.
- [73] G. B. Stefano, “Comparing Bioinformatic Gene Expression Profiling Methods: Microarray and RNA-Seq,” *Medical Science Monitor Basic Research*, vol. 20, pp. 138–142, 2014, doi: 10.12659/MSMBR.892101.
- [74] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: a revolutionary tool for transcriptomics,” *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, Jan. 2009, doi: 10.1038/nrg2484.
- [75] R. Stark, M. Grzelak, and J. Hadfield, “RNA sequencing: the teenage years,” *Nature Reviews Genetics*, vol. 20, no. 11, pp. 631–656, Nov. 2019, doi: 10.1038/s41576-019-0150-2.
- [76] B. E. Slatko, A. F. Gardner, and F. M. Ausubel, “Overview of Next-Generation Sequencing Technologies,” *Current Protocols in Molecular Biology*, vol. 122, no. 1, Apr. 2018, doi: 10.1002/cpmb.59.
- [77] Y. Han, S. Gao, K. Muegge, W. Zhang, and B. Zhou, “Advanced Applications of RNA Sequencing and Challenges,” *Bioinformatics and Biology Insights*, vol. 9s1, p. BBI.S28991, Jan. 2015, doi: 10.4137/BBI.S28991.

- [78] V. Raghavan, L. Kraft, F. Mesny, and L. Rigerte, “A simple guide to *de novo* transcriptome assembly and annotation,” *Briefings in Bioinformatics*, vol. 23, no. 2, pp. 1–30, Mar. 2022, doi: 10.1093/bib/bbab563.
- [79] S. L. Salzberg, “Next-generation genome annotation: we still struggle to get it right,” *Genome Biology*, vol. 20, no. 1, p. 92, Dec. 2019, doi: 10.1186/s13059-019-1715-2.
- [80] K. R. Kukurba and S. B. Montgomery, “RNA Sequencing and Analysis,” *Cold Spring Harbor Protocols*, vol. 2015, no. 11, p. pdb.top084970, Nov. 2015, doi: 10.1101/pdb.top084970.
- [81] J. A. Martin and Z. Wang, “Next-generation transcriptome assembly,” *Nature Reviews Genetics*, vol. 12, no. 10, pp. 671–682, Oct. 2011, doi: 10.1038/nrg3068.
- [82] R. Hrdlickova, M. Toloue, and B. Tian, “<scp>RNA</scp> -Seq methods for transcriptome analysis,” *WIREs RNA*, vol. 8, no. 1, Jan. 2017, doi: 10.1002/wrna.1364.
- [83] V. Peona, M. H. Weissensteiner, and A. Suh, “How complete are ‘complete’ genome assemblies?-An avian perspective,” *Molecular Ecology Resources*, vol. 18, no. 6, pp. 1188–1195, Nov. 2018, doi: 10.1111/1755-0998.12933.
- [84] N. Stoler and A. Nekrutenko, “Sequencing error profiles of Illumina sequencing instruments,” *NAR Genomics and Bioinformatics*, vol. 3, no. 1, Jan. 2021, doi: 10.1093/nargab/lqab019.
- [85] T. I. Garcia *et al.*, “Effects of short read quality and quantity on a de novo vertebrate transcriptome assembly,” *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology*, vol. 155, no. 1, pp. 95–101, Jan. 2012, doi: 10.1016/j.cbpc.2011.05.012.
- [86] S. Andrews, “FastQC: a quality control tool for high throughput sequence data.” 2010.
- [87] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, “MultiQC: summarize analysis results for multiple tools and samples in a single report,”

- Bioinformatics*, vol. 32, no. 19, pp. 3047–3048, Oct. 2016, doi: 10.1093/bioinformatics/btw354.
- [88] G. de Sena Brandine and A. D. Smith, “Falco: high-speed FastQC emulation for quality control of sequencing data,” *F1000Res*, vol. 8, p. 1874, Jan. 2021, doi: 10.12688/f1000research.21142.2.
- [89] L. Song and L. Florea, “Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads,” *Gigascience*, vol. 4, no. 1, p. 48, Dec. 2015, doi: 10.1186/s13742-015-0089-y.
- [90] M. Martin, “Cutadapt removes adapter sequences from high-throughput sequencing reads,” *EMBnet J*, vol. 17, no. 1, p. 10, May 2011, doi: 10.14806/ej.17.1.200.
- [91] S. Chen, Y. Zhou, Y. Chen, and J. Gu, “fastp: an ultra-fast all-in-one FASTQ preprocessor,” *Bioinformatics*, vol. 34, no. 17, pp. i884–i890, Sep. 2018, doi: 10.1093/bioinformatics/bty560.
- [92] A. M. Bolger, M. Lohse, and B. Usadel, “Trimmomatic: a flexible trimmer for Illumina sequence data,” *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, Aug. 2014, doi: 10.1093/bioinformatics/btu170.
- [93] D. E. Wood, J. Lu, and B. Langmead, “Improved metagenomic analysis with Kraken 2,” *Genome Biology*, vol. 20, no. 1, p. 257, Dec. 2019, doi: 10.1186/s13059-019-1891-0.
- [94] E. Kopylova, L. Noé, and H. Touzet, “SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data,” *Bioinformatics*, vol. 28, no. 24, pp. 3211–3217, Dec. 2012, doi: 10.1093/bioinformatics/bts611.
- [95] K. Struhl, “Transcriptional noise and the fidelity of initiation by RNA polymerase II,” *Nature Structural & Molecular Biology*, vol. 14, no. 2, pp. 103–105, Feb. 2007, doi: 10.1038/nsmb0207-103.
- [96] M. Cavallaro *et al.*, “3'-5' crosstalk contributes to transcriptional bursting,” *Genome Biology*, vol. 22, no. 1, p. 56, Dec. 2021, doi: 10.1186/s13059-020-02227-5.

- [97] K. D. Hansen, S. E. Brenner, and S. Dudoit, “Biases in Illumina transcriptome sequencing caused by random hexamer priming,” *Nucleic Acids Research*, vol. 38, no. 12, pp. e131–e131, Jul. 2010, doi: 10.1093/nar/gkq224.
- [98] S. Canzar, S. Andreotti, D. Weese, K. Reinert, and G. W. Klau, “CIDANE: comprehensive isoform discovery and abundance estimation,” *Genome Biology*, vol. 17, no. 1, p. 16, Dec. 2016, doi: 10.1186/s13059-015-0865-0.
- [99] F. Ozsolak and P. M. Milos, “RNA sequencing: advances, challenges and opportunities,” *Nature Reviews Genetics*, vol. 12, no. 2, pp. 87–98, Feb. 2011, doi: 10.1038/nrg2934.
- [100] E. Bushmanova, D. Antipov, A. Lapidus, and A. D. Prjibelski, “rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data,” *Gigascience*, vol. 8, no. 9, Sep. 2019, doi: 10.1093/gigascience/giz100.
- [101] M. G. Grabherr *et al.*, “Full-length transcriptome assembly from RNA-Seq data without a reference genome,” *Nature Biotechnology*, vol. 29, no. 7, pp. 644–652, Jul. 2011, doi: 10.1038/nbt.1883.
- [102] C. J. McManus and B. R. Graveley, “RNA structure and the mechanisms of alternative splicing,” *Current Opinion in Genetics & Development*, vol. 21, no. 4, pp. 373–379, Aug. 2011, doi: 10.1016/j.gde.2011.04.001.
- [103] Y. Zhang, J. Qian, C. Gu, and Y. Yang, “Alternative splicing and cancer: a systematic review,” *Signal Transduction and Targeted Therapy*, vol. 6, no. 1, p. 78, Dec. 2021, doi: 10.1038/s41392-021-00486-7.
- [104] Y. Xie *et al.*, “SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads,” *Bioinformatics*, vol. 30, no. 12, pp. 1660–1666, Jun. 2014, doi: 10.1093/bioinformatics/btu077.
- [105] M. H. Schulz, D. R. Zerbino, M. Vingron, and E. Birney, “Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels,” *Bioinformatics*, vol. 28, no. 8, pp. 1086–1092, Apr. 2012, doi: 10.1093/bioinformatics/bts094.

- [106] G. Robertson *et al.*, “De novo assembly and analysis of RNA-seq data,” *Nature Methods*, vol. 7, no. 11, pp. 909–912, Nov. 2010, doi: 10.1038/nmeth.1517.
- [107] A. H. Freedman, M. Clamp, and T. B. Sackton, “Error, noise and bias in de novo transcriptome assemblies,” *Molecular Ecology Resources*, vol. 21, no. 1, pp. 18–29, Jan. 2021, doi: 10.1111/1755-0998.13156.
- [108] L. S. Arroyo Mühr, C. Lagheden, S. S. Hassan, S. N. Kleppe, E. Hultin, and J. Dillner, “De novo sequence assembly requires bioinformatic checking of chimeric sequences,” *PLOS ONE*, vol. 15, no. 8, p. e0237455, Aug. 2020, doi: 10.1371/journal.pone.0237455.
- [109] W. Shen, S. Le, Y. Li, and F. Hu, “SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation,” *PLOS ONE*, vol. 11, no. 10, p. e0163962, Oct. 2016, doi: 10.1371/journal.pone.0163962.
- [110] E. S. Lander *et al.*, “Initial sequencing and analysis of the human genome,” *Nature*, vol. 409, no. 6822, pp. 860–921, Feb. 2001, doi: 10.1038/35057062.
- [111] R. Smith-Unna, C. Boursnell, R. Patro, J. M. Hibberd, and S. Kelly, “TransRate: reference-free quality assessment of de novo transcriptome assemblies,” *Genome Research*, vol. 26, no. 8, pp. 1134–1144, Aug. 2016, doi: 10.1101/gr.196469.115.
- [112] B. Li *et al.*, “Evaluation of de novo transcriptome assemblies from RNA-Seq data,” *Genome Biology*, vol. 15, no. 12, p. 553, Dec. 2014, doi: 10.1186/s13059-014-0553-5.
- [113] E. Bushmanova, D. Antipov, A. Lapidus, V. Suvorov, and A. D. Prjibelski, “rnaQUAST: a quality assessment tool for *de novo* transcriptome assemblies: Table 1,” *Bioinformatics*, vol. 32, no. 14, pp. 2210–2212, Jul. 2016, doi: 10.1093/bioinformatics/btw218.
- [114] A. Conesa *et al.*, “A survey of best practices for RNA-seq data analysis,” *Genome Biology*, vol. 17, no. 1, p. 13, Dec. 2016, doi: 10.1186/s13059-016-0881-8.

- [115] B. Langmead and S. L. Salzberg, “Fast gapped-read alignment with Bowtie 2,” *Nature Methods*, vol. 9, no. 4, pp. 357–359, Apr. 2012, doi: 10.1038/nmeth.1923.
- [116] A. Dobin *et al.*, “STAR: ultrafast universal RNA-seq aligner,” *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013, doi: 10.1093/bioinformatics/bts635.
- [117] H. Li *et al.*, “The Sequence Alignment/Map format and SAMtools,” *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009, doi: 10.1093/bioinformatics/btp352.
- [118] B. Li and C. N. Dewey, “RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome,” *BMC Bioinformatics*, vol. 12, no. 1, p. 323, Dec. 2011, doi: 10.1186/1471-2105-12-323.
- [119] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, “Near-optimal probabilistic RNA-seq quantification,” *Nature Biotechnology*, vol. 34, no. 5, pp. 525–527, May 2016, doi: 10.1038/nbt.3519.
- [120] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, “Salmon provides fast and bias-aware quantification of transcript expression,” *Nature Methods*, vol. 14, no. 4, pp. 417–419, Apr. 2017, doi: 10.1038/nmeth.4197.
- [121] D. M. Bryant *et al.*, “A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors,” *Cell Reports*, vol. 18, no. 3, pp. 762–776, Jan. 2017, doi: 10.1016/j.celrep.2016.12.063.
- [122] S. Zhao, “Alternative splicing, RNA-seq and drug discovery,” *Drug Discovery Today*, vol. 24, no. 6, pp. 1258–1267, Jun. 2019, doi: 10.1016/j.drudis.2019.03.030.
- [123] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, “CD-HIT: accelerated for clustering the next-generation sequencing data,” *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, Dec. 2012, doi: 10.1093/bioinformatics/bts565.
- [124] W. Li and A. Godzik, “Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences,” *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, Jul. 2006, doi: 10.1093/bioinformatics/btl158.

- [125] M. Steinegger and J. Söding, “MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets,” *Nature Biotechnology*, vol. 35, no. 11, pp. 1026–1028, Nov. 2017, doi: 10.1038/nbt.3988.
- [126] N. M. Davidson, A. D. K. Hawkins, and A. Oshlack, “SuperTranscripts: a data driven reference for analysis and visualisation of transcriptomes,” *Genome Biology*, vol. 18, no. 1, p. 148, Dec. 2017, doi: 10.1186/s13059-017-1284-1.
- [127] A. Oshlack, M. D. Robinson, and M. D. Young, “From RNA-seq reads to differential expression results,” *Genome Biology*, vol. 11, no. 12, p. 220, 2010, doi: 10.1186/gb-2010-11-12-220.
- [128] J. Zyprych-Walczak *et al.*, “The Impact of Normalization Methods on RNA-Seq Data Analysis,” *BioMed Research International*, vol. 2015, pp. 1–10, 2015, doi: 10.1155/2015/621690.
- [129] W. W. Wilfinger, R. Miller, H. R. Eghbalnia, K. Mackey, and P. Chomczynski, “Strategies for detecting and identifying biological signals amidst the variation commonly found in RNA sequencing data,” *BMC Genomics*, vol. 22, no. 1, p. 322, Dec. 2021, doi: 10.1186/s12864-021-07563-9.
- [130] 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, p. 550, Dec. 2014, doi: 10.1186/s13059-014-0550-8.
- [131] 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, Jan. 2010, doi: 10.1093/bioinformatics/btp616.
- [132] A. McDermaid, B. Monier, J. Zhao, B. Liu, and Q. Ma, “Interpretation of differential gene expression results of RNA-seq data: review and integration,” *Briefings in Bioinformatics*, vol. 20, no. 6, pp. 2044–2054, Nov. 2019, doi: 10.1093/bib/bby067.

- [133] M. E. Ritchie *et al.*, “limma powers differential expression analyses for RNA-sequencing and microarray studies,” *Nucleic Acids Research*, vol. 43, no. 7, pp. e47–e47, Apr. 2015, doi: 10.1093/nar/gkv007.
- [134] P. G. Nachtigall, A. Y. Kashiwabara, and A. M. Durham, “CodAn: predictive models for precise identification of coding regions in eukaryotic transcripts,” *Briefings in Bioinformatics*, vol. 22, no. 3, May 2021, doi: 10.1093/bib/bbaa045.
- [135] D. Hyatt, G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, and L. J. Hauser, “Prodigal: prokaryotic gene recognition and translation initiation site identification,” *BMC Bioinformatics*, vol. 11, no. 1, p. 119, Dec. 2010, doi: 10.1186/1471-2105-11-119.
- [136] W. R. Pearson, “BLAST and FASTA Similarity Searching for Multiple Sequence Alignment,” 2014, pp. 75–101. doi: 10.1007/978-1-62703-646-7_5.
- [137] W. R. Pearson, “An Introduction to Sequence Similarity (‘Homology’) Searching,” *Current Protocols in Bioinformatics*, vol. 42, no. 1, Jun. 2013, doi: 10.1002/0471250953.bi0301s42.
- [138] C. Camacho *et al.*, “BLAST+: architecture and applications,” *BMC Bioinformatics*, vol. 10, no. 1, p. 421, Dec. 2009, doi: 10.1186/1471-2105-10-421.
- [139] B. Buchfink, K. Reuter, and H.-G. Drost, “Sensitive protein alignments at tree-of-life scale using DIAMOND,” *Nature Methods*, vol. 18, no. 4, pp. 366–368, Apr. 2021, doi: 10.1038/s41592-021-01101-x.
- [140] R. Agarwala *et al.*, “Database resources of the National Center for Biotechnology Information,” *Nucleic Acids Research*, vol. 46, no. D1, pp. D8–D13, Jan. 2018, doi: 10.1093/nar/gkx1095.
- [141] A. Bateman *et al.*, “UniProt: the universal protein knowledgebase in 2021,” *Nucleic Acids Research*, vol. 49, no. D1, pp. D480–D489, Jan. 2021, doi: 10.1093/nar/gkaa1100.

- [142] S. R. Eddy, “Accelerated Profile HMM Searches,” *PLoS Computational Biology*, vol. 7, no. 10, p. e1002195, Oct. 2011, doi: 10.1371/journal.pcbi.1002195.
- [143] J. J. Almagro Armenteros *et al.*, “SignalP 5.0 improves signal peptide predictions using deep neural networks,” *Nature Biotechnology*, vol. 37, no. 4, pp. 420–423, Apr. 2019, doi: 10.1038/s41587-019-0036-z.
- [144] P. M. Harrison, “fLPS: Fast discovery of compositional biases for the protein universe,” *BMC Bioinformatics*, vol. 18, no. 1, p. 476, Dec. 2017, doi: 10.1186/s12859-017-1906-3.
- [145] P. Jones *et al.*, “InterProScan 5: genome-scale protein function classification,” *Bioinformatics*, vol. 30, no. 9, pp. 1236–1240, May 2014, doi: 10.1093/bioinformatics/btu031.
- [146] M. Ashburner *et al.*, “Gene Ontology: tool for the unification of biology,” *Nature Genetics*, vol. 25, no. 1, pp. 25–29, May 2000, doi: 10.1038/75556.
- [147] S. Carbon *et al.*, “The Gene Ontology resource: enriching a Gold mine,” *Nucleic Acids Research*, vol. 49, no. D1, pp. D325–D334, Jan. 2021, doi: 10.1093/nar/gkaa1113.
- [148] J. Huerta-Cepas *et al.*, “Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper,” *Molecular Biology and Evolution*, vol. 34, no. 8, pp. 2115–2122, Aug. 2017, doi: 10.1093/molbev/msx148.
- [149] J. Huerta-Cepas *et al.*, “eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses,” *Nucleic Acids Research*, vol. 47, no. D1, pp. D309–D314, Jan. 2019, doi: 10.1093/nar/gky1085.
- [150] S. Gotz *et al.*, “High-throughput functional annotation and data mining with the Blast2GO suite,” *Nucleic Acids Research*, vol. 36, no. 10, pp. 3420–3435, Apr. 2008, doi: 10.1093/nar/gkn176.

- [151] M. Kanehisa, “KEGG: Kyoto Encyclopedia of Genes and Genomes,” *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, Jan. 2000, doi: 10.1093/nar/28.1.27.
- [152] M. Kanehisa, M. Furumichi, Y. Sato, M. Ishiguro-Watanabe, and M. Tanabe, “KEGG: integrating viruses and cellular organisms,” *Nucleic Acids Research*, vol. 49, no. D1, pp. D545–D551, Jan. 2021, doi: 10.1093/nar/gkaa970.
- [153] B. Jassal *et al.*, “The reactome pathway knowledgebase,” *Nucleic Acids Research*, Nov. 2019, doi: 10.1093/nar/gkz1031.
- [154] K. J. Livak and T. D. Schmittgen, “Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the $2^{-\Delta\Delta CT}$ Method,” *Methods*, vol. 25, no. 4, pp. 402–408, 2001, doi: <https://doi.org/10.1006/meth.2001.1262>.
- [155] E. R. Rowley, R. VanBuren, D. W. Bryant, H. D. Priest, S. A. Mehlenbacher, and T. C. Mockler, “A Draft Genome and High-Density Genetic Map of European Hazelnut (*Corylus avellana* L.),” *bioRxiv*, 2018, doi: 10.1101/469015.
- [156] A. İslam, “Fındık ıslahında gelişmeler,” *Akademik Ziraat Dergisi*, Jan. 2020, doi: 10.29278/azd.667662.
- [157] Y.-W. Ban *et al.*, “Comparative transcriptome analysis reveals higher expression of stress and defense responsive genes in dwarf soybeans obtained from the crossing of G. max and G. soja,” *Genes & Genomics*, vol. 41, no. 11, pp. 1315–1327, Nov. 2019, doi: 10.1007/s13258-019-00846-2.
- [158] Y. Jo, S. Lian, J. K. Cho, H. Choi, H. Chu, and W. K. Cho, “De novo transcriptome assembly of two different *Prunus salicina* cultivars,” *Genomics Data*, vol. 6, pp. 262–263, Dec. 2015, doi: 10.1016/J.GDATA.2015.10.015.
- [159] P. McGarvey *et al.*, “De novo assembly and annotation of transcriptomes from two cultivars of *Cannabis sativa* with different cannabinoid profiles,” *Gene*, vol. 762, p. 145026, Dec. 2020, doi: 10.1016/j.gene.2020.145026.

- [160] P. Boccacci, A. Akkak, and R. Botta, "DNA typing and genetic relations among European hazelnut (*Corylus avellana* L.) cultivars using microsatellite markers," *Genome*, vol. 49, no. 6, pp. 598–611, Jun. 2006, doi: 10.1139/g06-017.
- [161] B. Yirmibeş, H. N. Aydın, Z. Ömeroğlu Ulu, S. Ulu, and N. Özdemir Özgentürk, "Analysis of some conserved miRNAs in hazelnut (*Corylus avellana* L. and *Corylus colurna* L.) by real-time PCR," *Sigma Journal of Engineering and Natural Sciences – Sigma Mühendislik ve Fen Bilimleri Dergisi*, pp. 170–176, 2021, doi: 10.14744/sigma.2021.00006.



Konferans Bildirileri

1. N. Özdemir Özgentürk, S. Ulu, Z. Ömeroğlu Ulu, A. Akar, “Preparation of de novo transcriptome analysis from Turkish filbert (*Corylus colurna* L.) and Tombul hazelnut (*Corylus avellana* L. Tombul)”, 3th International Conference on Transcriptomics, Bangkok, Thailand, Transcriptomics 2017 5:3 DOI 10.4172/2329-8936-C1-015
2. N. Özdemir Özgentürk, Z. Ömeroğlu Ulu, A. Akar, S. Ulu, “Generation Of Rnaseq Data For *Corylus Avellana* L.”, International Biological, Agricultural And Life Science Congress 2019, Lviv, Ukrain, Trakya University Publisher No: 220, ISBN: 978-975-374-249-8, 249
3. N. Özdemir Özgentürk, Z. Ömeroğlu Ulu, A. Akar, S. Ulu, “Obtaining the Transcriptome Data for *Corylus colurna* L.”, Seed Science And Archaeobotanical Research in Anatolia, SARA Symposium 2019, İzmir, Türkiye