T.C.
AKDENİZ ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ
Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı

KANSER GENETİĞİ VERİTABANI OLUŞTURULMASI VE BİYOİNFORMATİK ARAÇLARINA ENTEGRASYONU İLE YENİ KANSER VERİ SETLERİNİN ANALİZİ

Mehmet Kemal SAMUR

Doktora Tezi

Antalya, 2013

T.C.
AKDENİZ ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ
Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı

KANSER GENETİĞİ VERİTABANI OLUŞTURULMASI VE BİYOİNFORMATİK ARAÇLARINA ENTEGRASYONU İLE YENİ KANSER VERİ SETLERİNİN ANALİZİ

Mehmet Kemal SAMUR

Doktora Tezi

Tez Danışmanı

Prof. Dr. Osman SAKA

Antalya, 2013

Sağlık Bilimleri Enstitüsü Müdürlüğüne;

Bu çalışma jürimiz tarafından Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı, Tıp Bilişimi programında Doktora tezi olarak kabul edilmiştir. …./…./….

Tez Danışmanı : Prof. Dr. Osman SAKA
Akdeniz Üniversitesi
Tıp Fakültesi
Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı

Üye : Prof. Dr. Ergun KARAAĞAOĞLU
Hacettepe Üniversitesi
Tıp Fakültesi
Biyoistatistik Anabilim Dalı

Üye : Prof. Dr. Gültekin SÜLEYMANLAR
Akdeniz Üniversitesi
Tıp Fakültesi
Nefroloji Bilim Dalı

Üye : Yrd. Doç. Dr. Uğur BİLGE
Akdeniz Üniversitesi
Tıp Fakültesi
Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı

Üye : Yrd. Doç. Dr. K. Hakan GÜLKESEN
Akdeniz Üniversitesi
Tıp Fakültesi
Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı

ONAY:

Bu tez, Enstitü Yönetim Kurulunca belirlenen yukarıdaki jüri üyeleri tarafından uygun görülmüş ve Enstitü Yönetim Kurulu'nun …./…./2013 tarih ve ……../……. sayılı kararıyla kabul edilmiştir.

Prof. Dr. İsmail ÜSTÜNEL

Enstitü Müdürü

Sağlık Bilimleri Enstitüsü Kurulu ve Akdeniz Üniversitesi Senato Kararı

Sağlık Bilimleri Enstitüsü'nün 22/06/2000 tarih ve 02/09 sayılı enstitü kurul kararı ve 23/05/2003 tarih ve 04/44 sayılı senato kararı gereğince "Sağlı Bilimleri Enstitülerinde lisansüstü eğitim gören doktora öğrencilerinin tez savunma sınavına girebilmeleri için doktora bilim alanında SCI tarafından taranan dergilerde en az 1 yurtdışı yayın yapması gerektiği" ilkesi gereğince yapılan yayınların listesi aşağıdadır (Orjinalleri ekte sunulmuştur).

1. Yan Z, Shah PK, Amin SB, **Samur MK**, Huang N, Wang X, Misra V, Ji H, Gabuzda D, Li C. Integrative analysis of gene and miRNA expression profiles with transcription factor-miRNA feed-forward loops identifies regulators in human cancers. Nucleic Acids Res 2012, 40(17):e135.

ÖZET

Açık veritabanlarında depolanan fonksiyonel genomik platformlar kullanılarak tümörden elde edilmiş genom çapındaki profiller, büyük projelerin ve daha küçük çaplı araştırma ekiplerinin ilgisi ile beraber astronomik boyutlara ulaşmaktadır. Bunun sonucunda, günümüzde bu verileri mevcut bilgiler ışığında entegre edebilecek ve yorumlayabilecek aynı zamanda kanser araştırmacıları ile paylaşabilecek araçlara acil ihtiyaç vardır. Bu tez projesinde amaç bu ihtiyaca cevap verebilecek bir biyoinformatik platformu oluşturmaktır.

Geliştirilen platformun fonksiyonları kanser araştırmacılarının orijinal hipotezler geliştirirken sıklıkla ihtiyaç duydukları analizleri gerçekleştirebilecek şekilde oluşturulmuştur. Platform farklı kanser türleri için gen, protein, miRNA ekspresyonlarına ilişkin profilleri, mutasyon bilgilerini, gen kopya sayısı değişimlerini ve protein – protein etkileşimi bilgilerini saklayabilmektedir. Platform araştırmacıların onkogenomik veri ile gerçekleştirilmiş temel, entegre ve ağ analizlerine ilişkin sonuçları sorgulayabilmelerini sağlamaktadır. Gen ve miRNA ekspresyonu farkı analizi, SNP verisinden elde edilen kopya sayısı değişiminin temel analiz olarak sunulduğu platformda gen ekspresyonu, miRNA ekspresyonu, mutasyon bilgisi ve kopya sayısı değişimlerini entegre eden çeşitli analiz yöntemlerini ve bunlar ile beraber genelleştirilmiş gen seti zenginleştirme analizi entegre analiz yöntemleri olarak yer almaktadır. Ağ analizi özelliği ise bir birine eşlik eden gen ekspresyonu değişimlerini, gen ekspresyonunu düzenleyici ağları ve protein – protein etkileşimlerini saklamayı ve görselleştirmeyi içermektedir. Son olarak, geliştirilen platformda gen ekspresyonu verisi ile klinik veri birleştirilerek sağ kalım analizi gerçekleştirilebilmektedir.

Geliştirilen platform mevcut hali ile 90 farklı genomik kanser araştırmasından elde edilen 20,000'den fazla hastaya ilişkin bilgiyi saklamaktadır. Farklı veri tiplerinin kullanılabilmesi, onkogenetik üzerinde etkili faktörlerin tespit edilebilmesi için orijinal entegre analiz yöntemlerini desteklemesi ve gen veya yolaklar üzerinden etkili değişimlerin sorgulanabilmesi ile öne çıkan platform literatürde var olan çalışmalara göre araştırmacılara çok zengin ve orijinal bir kaynak sunmaktadır.

Anahtar Kelimeler: Biyoinformatik, Kanser, Entegre Analiz, Genetik, Gen Ekspresyonu, miRNA Ekspresyonu, Protein Ekspresyonu, Gen Kopya Sayısı, Mutasyon

ABSTRACT

Genome-wide profiles of tumors obtained using functional genomics platforms are being deposited to the public repositories at an astronomical scale, as a result of focused efforts by individual laboratories and large projects. Consequently, there is an urgent need for reliable tools that integrate and interpret these data in light of current knowledge and disseminate results to cancer researchers in a user-friendly manner. Here we aim to develop a bioinformatics platform to meet this need.

The developed platform query functionalities are designed to fulfill most frequent analysis needs of cancer researchers with a view to generate novel hypotheses. It stores gene, protein and miRNA expression profiles, mutation information and copy number alterations for multiple cancer types, and protein-protein interaction information. The platfom allows querying of results of primary analysis, integrative analysis and network analysis of oncogenomics data. The querying for primary analysis includes differential gene and miRNA expression as well as changes in gene copy number measured with SNP microarrays. It provides results of integrative analysis of gene expression profiles with copy number alterations, mutations, protein expressions and miRNA profiles as well as generalized integrative analysis using gene set enrichment analysis. The network analysis capability includes storage and visualization of gene co-expression, inferred gene regulatory networks and protein-protein interaction information. Finally, the platform provides correlations between gene expression and clinical outcomes in terms of univariate survival analysis.

At present the platform provides different types of information extracted from 90 cancer genomics studies with information from more than 20,000 patients. The presence of multiple data types, novel integrative analysis for identifying regulators of oncogenesis, network analysis and ability to query gene lists/pathways are distinctive features of the platform. Finally, compared to current tools in the literature, the developed platform offers a novel bioinformatics platform to the community with it's rich content.

Key Words: Bioinformatics, Cancer, Integrative Analysis, Genetics, Gene Expression, Protein Expression, miRNA Expression, Gene Copy Number, Mutations

TEŞEKKÜR

Bu tez projesinin yürütülmesinde danışmanım olan Prof. Dr. Osman SAKA'ya, çalışmanın planlama aşamasından itibaren her konuda yardımcı olan Prof. Dr. Cheng LI, Prof. Dr. Nikhil C. MUNSHI ve Dr. Parantu K. SHAH'a teşekkürlerimi sunarım.

Ayrıca biyoinformatik alanında çalışmaya başlamamda yardımcı olan ve projenin başından itibaren destek veren Yrd. Doç. Dr. Özgür TOSUN'a teşekkür ederim.

Yüksek lisans eğitiminden itibaren tıp bilişimi alanında bana yol gösteren anabilim dalı öğretim üyelerimiz Yrd. Doç. Dr. Uğur BİLGE, Yrd. Doç. Dr. Kemal Hakan GÜLKESEN ve Yrd. Doç. Dr. Neşe ZAYİM'e teşekkür ederim. Ayrıca bu maratonu beraber koştuğumuz çalışma arkadaşlarıma desteklerinden dolayı teşekkür ederim.

Son olarak, böyle bir projenin var olabilmesi için en büyük desteği veren değerli eşim ve çalışma arkadaşım Anıl AKTAŞ SAMUR'a çok teşekkür ederim.

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

BWA	:	Burrows-Wheeler Aligner
CSS	:	Cascading Style Sheets
DFCI	:	Dana Farber Cancer Institute
DNA	:	Deoksiribonükleik Asit
FDR	:	False Discovery Rate
GEO	:	Gene Expression Omnibus
GSEA	:	Gene Set Enrichment Analysis
GWAS	:	Genome-wide Association Study
HTML	:	Hyper Text Markup Language
JSON	:	JavaScript Object Notation
LOH	:	Loss of Heterozygosity
MAQC	:	The MicroArray Quality Control
miRNA	:	Mikro Ribonükleik Asit
MM	:	Multiple Myeloma
MSigDB	:	The Molecular Signatures Database
NCBI	:	Amerikan Ulusal Biyoteknoloji Bilgi Merkezi
NGS	:	Next Generation Sequencing
PCR	:	Polymerase Chain Reaction
RNA	:	Ribonükleik Asit
SNP	:	Single-Nucleotide Polymorphism
TCGA	:	The Cancer Genome Atlas
TF	:	Transkripsiyon Faktörü

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

GİRİŞ

Son 30 yılda bilgi teknolojilerinde yaşanan hızlı gelişim, sağlık araştırmalarını da önemli derecede etkilemiştir. Özellikle 2000'li yıllardan sonra yaşanan değişimler genom temelli projelerin artmasına ve yeni araştırmalar planlanmasına olanak vermiştir.

"İnsan Genom" projesinin 10. yıl dönümünde çeşitli teknolojilerin getirdiği yüksek hacimli genomik veriye karşın bu veriyi işleyebilecek kaynakların aynı hızda büyümemiş olması, yeni analiz araç ve algoritmalarına duyulan ihtiyaç biyoinformatik alanında bir dar boğaz oluşturmaktadır [1].

Genomik veri hacmindeki büyümenin, özellikle de kanser ile ilgili üretilen genomik verinin bu kadar hızlı artmasındaki en önemli nedenlerinden birisi düşen maliyetler ve git gide artan araştırma faaliyetleridir. "Kanser Genom Atlası" (TCGA – The Cancer Genome Atlas)[2] ve "Kanser Genom Konsorsiyumu" (Cancer Genome Consortium)[3] gibi büyük çaplı projeler ve araştırma laboratuvarları yeni kaynaklardan sürekli yeni veri elde etmemizi sağlamaktadır.

Veri hacmindeki büyük artışla birlikte ortaya çıkan depolama, analiz yöntemi geliştirme ve kaynak kullanımı problemlerine rağmen, toplanan veriden elde edilebilecek bulguların kansere ilişkin yeni bilgiler sunmada büyük önemi vardır. Collins ve arkadaşları yaptıkları bir çalışmada bu verilerden elde edilebilecek ve insan yararına dönüştürülebilecek faydaları aşağıdaki temel başlıklar altında toplamışlardır[4]. Buna göre:

- Bu veri sayesinde hastalıkların ortaya çıkışı, gelişmesi ve çevresel faktörlerle ne şekilde etkileşim içinde olduğunu ortaya koyabilecek çalışmalar gerçekleştirilebilir
- Genom tabanlı tanı koyma modelleri geliştirilebilir ve değerlendirilebilir, hastalığın seyri ve tipi ile ilgili genom düzeyinde bilgi edinilebilir
- Hastalığa karşı kullanılacak ilaçlar belirlenebilir, genom üzerinden elde edilen bilgiler ışığında bulunan ilaç hedefleri temel alınarak yeni tedaviler geliştirilebilir
- Hastalıklar erken teşhis edilebilir ya da moleküler düzeyde daha net sınıflandırmalar yapılabilir
- Genomik bilgiyi tedavi süreçlerine katılabilecek yöntemlerin gelişmesine destek olabilir.

Bu verilerden elde edilebilecek sonuçların araştırmacılar ile paylaşılabilir ve sorgulanabilir olması kanser araştırmaları yürüten bütün bilim insanları için faydalı olacaktır.

Onkogenomik profillerden anlamlı bilgiler elde edebilmek için çeşitli araştırma soruları sorulmaktadır. Örneğin: Benim ilgilendiğim gen ya da miRNA hasta ve kontrol grubu arasında farklı ekspresyon değerlerine sahip mi? Araştırdığım genin kopya sayısı hasta bireylerde nasıl değişiyor? Gen ekspresyonundaki değişim, genin kopya sayısındaki değişim ile açıklanabilir mi? Hangi genler veya biyolojik düzenleyiciler tümör oluşumunda etkili? Belirlenen tümör tipinde hangi genlerin ekspresyon seviyesindeki değişimin sağ kalım üzerine etkisi var? Belirli kanser türlerine özel önemli değişiklikler gözlenen yolak veya modüller hangileridir? Bu kanser türünün neden olduğu değişiklikler hangi biyolojik süreçleri etkiliyor?

Bu tarz araştırma sorularına yanıtlar bulmak kanser alanında araştırmalar gerçekleştiren bilim insanlarının sürekli üzerinde durduğu konulardandır. Ancak elde edilen verilerden bu soruları yanıtlayacak bulgular elde edebilmek için büyük ölçekli analiz altyapısı ve analizlerin yürütülebilmesi için de uzmanlaşmış bir ekip gerekmektedir.

Araştırmacıların sıklıkla ilgilendikleri araştırma sorularına yanıt verebilecek, hipotezlerini oluşturmada ve çalışmalarını yürütmede gerekli yardımı sağlayabilecek çevrimiçi erişilebilen bir platform oluşturmak genomik bilginin yayılması açısından ideal bir ortam olarak düşünülmektedir. Ancak böyle bir platformun oluşturulması zorlu bir süreç gerektirmektedir. Farklı ekipler tarafından farklı platformlarda üretilen veri, verinin kalitesi, veriye eşlik eden eksik genomik lokasyon bilgileri, veriye uygun analizi seçmede ve uygulamada yaşanan nitelikli uzman eksikliği sıkıntıları bunlardan bazılarıdır. Bunlara ek olarak istatistik ve yapay zeka tekniklerini bilmek, verinin önişlenmesi sürecinden entegre analizine kadar giden yolda bir gerekliliktir. Yapılan son çalışmalar farklı veri türlerinin entegre edilebilmesi için geliştirilen araç ve yöntemlerin hala evrimleşme sürecinde olduğunu ve hala çeşitli problemlerle karşı karşıya olunduğunu göstermiştir[5]. Bütün bunlar birleştirildiğinde entegre bir platformun ortaya konulabilmesi için prosedürel ya da teorik bilgi, istatistik bilgisi ve analiz ağının oluşturulabilmesi için teknik bilginin gerekliliği ortaya çıkmaktadır.

Günümüzde genomik verinin depolanması ve araştırmacılar ile paylaşılabilmesi için GEO[6, 7] ve ArrayExpress[8, 9] gibi çeşitli büyük ölçekli ve açık veritabanları kurulmuştur. Ancak bu girişimler verinin paylaşılmasına odaklanmış, verinin analiz edilebilmesine ilişkin bir çalışma yapılmamıştır. Eldeki verinin analiz edilebilmesi içinde farklı araştırma grupları çeşitli ortamlar hazırlamışlardır. Oncomine[10, 11] ve Geneinvestigator[12] bunlara örnektir. Ancak bu çalışmalarda sadece verinin alt bir kümesine ya da belirli analizlere odaklanılmış, entegre analiz yöntemlerini ağ ve temel analiz yöntemleriyle birlikte araştırmacılara sunulmamıştır.

Bu noktadan hareketle bu tez projesinin amaçları aşağıdaki gibi sıralanabilir:

- Farklı tümör tipleri ile ilgili çeşitli platformlardan elde edilen fonksiyonel genomik (gen ve miRNA profilleri, gen kopya sayıları, protein bilgileri, yolaklara ilişkin bilgiler vb.) ve diğer büyük ölçekli verileri saklayabilen ve kullanabilen bir platform oluşturmak

- Oluşturulan platform üzerinde yöneticilerin onayı ile analizlerin otomatikleştirilmiş şekilde yapılabilmesini sağlamak ve sonuçlarını araştırmacıların sorgulayabilecekleri kaynaklara dönüştürmek

- Araştırmacıların kendi parametreleri ile ve araştırma sorularına uygun şekilde çalışma zamanında belirli temel, ağ ya da entegre genomik analizleri gerçekleştirebilmelerine imkan sağlamak

- Elde edilen sonuçları görselleştirebilen, verinin yorumlanmasında ve anlamlandırılmasında araştırmacılara yardımcı olan bir platform kurarak gerekli durumlarda kullanıcıların analizde kullandıkları veriyi indirebilmelerine imkan sağlamak

- Veri setlerinden elde edilen analiz bulgularını karşılaştırabilecek ve bu sayede farklı durumlarda genomik bilginin ne şekilde değişim gösterdiğine dair sonuçlar sunabilecek bir platform geliştirmek.

GENEL BİLGİLER

2.1. Biyoinformatik

Genetik biliminin temelleri 1866 yılında günümüzde Mendel yasası olarak bilinen ve kalıtımla ilgili teorilerin ilk olarak ortaya atıldığı zamana kadar dayanmaktadır. 1900'lerin ilk çeyreğinde öncelikle Wilhem Johannsen ilk kez gen terimini kullanmış ve ardından da 1915 yılında Morgan Hunt tarafından kromozomların üzerinde yer alan genlerin kalıtımın yapı taşları olduğu ortaya konulmuştur. 1953 yılında James Watson ve Francis Crick'in DNA'nın çift sarmallı yapısını keşfetmeleri[13] ve 1966 yılında Marshall Nirenber'in DNA'nın protein yapımındaki rolünü[14] tanımlamasından sonra genetik alanda yapılan çalışmalar git gide önem kazanmaya başlamıştır. 1977 yılında Sanger ve arkadaşları tarafından bulunan DNA dizileme modeli[15] ile günümüzde ulaşılan genetik veri üretme teknolojisinin de temelleri atılmıştır.

Genetik biliminin önem kazanmasına paralel olarak çeşitli hastalıkların genetik kalıtım ve değişimlerle ilgisi de incelenmeye başlanmıştır. Farklı hastalıklar ve türler üzerinde yapılan çalışmalar çevresel etkilerin olduğu kadar genetik faktörlerin de hastalıkların ortaya çıkmasında etken olduğunu göstemiştir [16, 17].

Bütün bu buluşlar ve araştırmaların sonucunda gelişen teknolojinin de katkısıyla genetik alanda üretilen verinin miktarı hızla artmıştır. Günümüzde çeşitli hastalıklara neden olabilecek genetik faktörlerinin incelenmesinden, kök hücre araştırmalarına kadar pek çok alanda genetik temelli araştırmalar yürütülmekte ve veriler elde edilmektedir.

1995 ve 1997 yıllarında sırasıyla mikrodizi teknolojisinin genetik araştırmalarda kullanılmaya başlanması[18] ve ökaryotik genomun yayınlanması[19] genetik çalışmaları hızlandırmıştır. İlerleyen yıllarda ise insan genomunun yayınlanmasını takiben [20-22] yaşanan genomik veri artışına paralel olarak bu veriler ile çalışabilecek bir bilim dalına da ihtiyaç duyulmuştur.

Biyoinformatik; biyolojik bilimlerden elde edilen nükleik asit (DNA/RNA), protein dizileri veya benzeri çok boyutlu bilgilerin bilgisayar kaynakları ve matematiksel ve istatistiksel modeller yardımıyla yorumlanması, yeni bilgi ve sonuçlara ulaşılmasını hedef alan çok disiplinli bir bilim dalıdır.

Amerikan Ulusal Biyoteknoloji Bilgi Merkezi (NCBI) internet sitesi üzerinden yayınladığı dokümanlarda biyoinformatiği; biyoloji, bilgisayar bilimi ve bilgi teknolojilerinin birleşiminden oluşmuş bir bilim olarak nitelendirmektedir.

Buna göre biyoinformatiğin üç önemli alt disiplini vardır. Bu disiplinler; yeni algoritmalar ve istatistiksel yöntemler geliştirerek büyük çaplı veri setlerindeki ilişkilerin ortaya çıkarılmasında, nükleotid, amino asit dizileri, protein domainleri ve protein yapıları gibi farklı tipteki verilerin analiz edilmesinde ve yorumlanmasında, farklı tipteki bilginin yönetilmesi ve erişimi için etkili araçların geliştirilmesinde önemli roller üstlenmektedirler (http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html).

2005 yılında ilk olarak Margulies ve arkadaşları tarafından duyurulan ve günümüzde git gide diğer teknolojilerin yerini alan gelecek nesil dizileme[23] teknikleri sayesinde bir bireyden elde edilebilecek genetik veri miktarı 500 Gb büyüklüğünü geçebilmektedir[24]. Onlarca veya yüzlerce bireyden elde edilen verinin klasik yaklaşımlarla analiz edilmesi mümkün olmamakta, bu alanda uzmanlaşmış araştırmacılar ve araştırma teknikleri gerektirmektedir. Bu anlamda biyoinformatiğin biyolojik bilimlere oldukça büyük katkıları olmaktadır. Biyolojik araştırmaların çok boyutlu veriler üzerinde doğrulanması veya tam tersi yönde çok boyutlu ve büyük miktardaki verilerden elde edilen bulguların biyolojik olarak doğrulanmak üzere araştırmacılara sağlanması biyoinformatiğin temel konusudur.

Biyoinformatik araştırmacıları birbirleri ile ilişkili olan çeşitli alt alanlarda araştırmalarını sürdürmekte ve bu alanların birbirleri ile etkileşimlerinden yeni bulgulara ulaşabilmektedirler. Bu alt alanlar aşağıdaki gibi üç ana grup içinde toplanabilirler.

- Genetik verinin saklanması ve gerektiğinde araştırmaların tekrar edilebilmesini sağlamak amacıyla veri saklama ve getirme odaklı biyoinformatik veritabanlarının geliştirilmesi

- Saklanan farklı türlerdeki verilerin analizlerini gerçekleştirebilecek araçların, algoritmaların ve metodolojilerin kurulmasını amaçlayan analiz odaklı araştırmalar

- Hücre içinde yer alan genetik faktörlerin sonucu oluşan farklı veri türlerini kullanmayı ve farklı genetik fonksiyon ve yapılardan yeni veriler elde etmeyi amaçlayan araştırmalar.

Bahsedilen alt alanlar ile ilgili bilgiler ilerleyen bölümlerde verilmektedir. Bu tez çalışması kapsamında oluşturulmuş olan platform bahsedilen alt alanların tamamından faydalanarak kanser araştırmaları yapmakta olan araştırmacılara araştırmalarını güçlendirecek ve hızlandıracak kolay erişilebilir ve kullanılabilir bir ortam sağlamayı amaçlamaktadır.

2.2. Biyolojik Veritabanları

Doksanlı yılların ortasından itibaren artan veri boyutuna paralel olarak bu verilerin saklanabilmesi, gerektiğinde geri getirilebilmesi ve yeniden analiz veya değerlendirilebilmesine olanak sağlayacak veri depolama ortamlarına, bir başka değişle genetik veritabanlarına ihtiyaç duyulmuştur. Bu ihtiyaç doğrultusunda farklı veri türleri, farklı biyolojik türler veya farklı araştırma konularını temel alan çeşitli veritabanları kurulmuştur. Bu veritabanlarının kurulmasında büyük çaplı projelerin veya daha küçük araştırma gruplarının etkisi büyük olmakla beraber, ticari amaçlı kurulmuş olanları da mevcuttur.

Günümüzde genetik verinin saklandığı ve paylaşıldığı veritabanları sakladıkları verinin türüne göre sınıflandırılmaktadır. Biyoinformatik araştırmacıları tarafından sıklıkla takip edilen ve yıllık olarak veritabanları ile ilgili özel bir sayı yayınlayan "Nucleic Acid Reseach" isimli bilimsel dergi son sayısında şimdiye kadar veritabanı kataloglarına girmeyi başarmış 1380 veritabanını aşağıda belirtilen 14 farklı kategori altında toplanmıştır [25].

- Nükleotid dizi veritabanları
- RNA dizi veritabanları
- Protein dizi veritabanları
- Yapısal veritabanları
- Genomik veritabanları (Omurgasız canlılar için)
- Metabolik ve sinyal yolağı veritabanları
- İnsan ve diğer omurgalı canlı veritabanları
- İnsan genleri ve hastalıkları veritabanları
- Mikrodizi verisi ve diğer gen ekspresyonu veritabanları
- Proteomik kaynaklar veritabanları
- Diğer moleküler biyoloji veritabanları
- Organel veritabanları
- Bitki veritabanları
- İmmünolojik veritabanları
- Hücre biyolojisi veritabanları

Bu tez çalışmasında yukarıda yer alan 8 ana başlığa ilişkin çeşitli kaynaklardan faydalanılmıştır.

Veritabanlarını içerdikleri verinin türüne göre sınıflandırmaya ek olarak amaca göre de sınıflandırmak mümkündür. Buna göre veritabanları 3 başlık atında toplanabilirler [26, 27].

- Veri arşivleyen
- Farklı kaynaklardan veri toplayıp paylaşan
- Farklı verilerin entegrasyonunu sağlayan

Mevcut tez çalışması sonucu ortaya çıkan platform özellik bakımından birinci ve üçüncü temel özellikleri taşımaktadır.

Günümüzde biyolojik veritabanları oluşturma konusunda sarf edilen yoğun çaba, açık olarak erişilebilen biyolojik veritabanlarının sürekli evrim geçirmesine ve bu kaynakların "NAR moleküler Biyoloji Veritabanı" veya "Database" gibi yüksek etki faktörlü akademik kataloglarda toplanmasına neden olmaktadır [26, 28, 29]. Teknolojideki gelişmenin sonucu olarak artan veri miktarı, kurulan veritabanı sayısını arttırmış olsa da pek çok araştırmacı bu artışın biyolojik araştırmalara yeteri kadar etki etmediğini düşünmektedir. Bunun en önemli nedeni ise oluşturulan ve paylaşılan kaynakların çoğunlukla bir standarttan uzak olması, dolayısıyla verinin paylaşılması, kullanılması ve yeniden analizinde çeşitli problemlerin ortaya çıkarmasıdır [26, 30-33]. Gaudet ve arkadaşları problemlerin belirli oranda giderilmesi amacıyla oluşturulacak biyolojik veritabanlarının belirli standartları taşıması gerektiğine dikkat çekerek "BioDBcore" [26] başlığı altında bir takım standartlar belirlemişlerdir. Günümüzde çeşitli bilimsel yayın organları yeni bir çalışmayı yayınlamadan önce oluşturulan biyolojik veritabanının bu standartlara uygun olmasını beklemektedirler.

Biyolojik veritabanlarının içinde bulunduğu bu durum göz önüne alınarak bu tez çalışmasında oluşturulan ve kanser verileri içeren bir veritabanı olarak da kullanılabilecek platform, belirli standartlar göz önüne alınarak hazırlanmıştır.

Bu çalışma kapsamında kullanılan kanser veri setleri de farklı tür ve araştırma konularına uygun genomik veri depolayan veritabanlarından veya kanser araştırma laboratuvarlarından elde edilmiş ve bu standartlara uygun hale getirilmiştir.

Büyük ölçekli veritabanlarının en başında NCBI tarafından kurulan veritabanları gelmektedir. NCBI bünyesinde kurulan ve günümüzde araştırmacılar tarafından farklı amaçlarla yaygın olarak kullanılan veritabanlarından bazıları BioProject, BioSample [34], dbGaP [35], RefSeq [36], dbSNP[37], SRA[38, 39] ve sağlık alanında toplanan mikrodizi verilerinin çoğunluğunu barındıran GEO'dur [6,

7, 40-45]. Bu veritabanlarına ilaveten Avrupa Biyoinformatik Enstitüsü tarafından kurulan ve bazı veri setleri için GEO ile beraber çalışan ArrayExpress [8, 9] ve Stanford Mikrodizi Veritabanı[46] gen ekspresyonu verilerine ulaşılabilen açık erişimli veritabanlarındandır.

Gen ekspresyonu bu alanda üretilen en yaygın veri türlerinden bir tanesi olsa da tek veri türü değildir. Bu nedenle diğer veri türleri için de farklı veritabanları kurulmuştur. Japon DNA Veri Bankası [47], Avrupa Biyoinformatik Enstitüsü Nükleotid arşivi[48] ve GenBank'ın [49] ortaklaşa girişimi ile oluşturulan nükleotid dizi veritabanı kendi alanlarındaki büyük veritabanlarındandır[50, 51]. Ensembl gibi genom dizilerinin tutulduğu veritabanları [52, 53], STRING gibi yüzlerce farklı türe ait milyonlarca protein bilgisinin saklandığı protein veritabanları [54-56], KEGG [57, 58] veya Reactome[59-61] gibi yolaklar için özelleşmiş veritabanları ya da HUGO[62] gibi gen sembollerini standartlaştırmayı temel alan spesifik veritabanları bu alanda yapılabilecek pek çok çalışmaya örnekler oluşturmakta, aynı zamanda biyoinformatik temelli analizlerde sıkça başvurulmaktadır.

Veritabanları açısından oldukça zengin olan biyoinformatik alanında bu zenginliğin yarattığı çeşitli problemler de vardır. Örneğin aynı görevi üstlenen pek çok kaynağın her biri farklı kullanıma, veri saklama ve sağlama özelliklerine sahiptir. Henüz yeni gelişmekte olan alan içinde tam oturmayan standartlar, bu kaynakları kullanmada deneyime sahip kişileri bile zaman zaman zorlamaktadır. Hatta bazı entegrasyon çalışmalarının farklı kaynaklar için tekrar edilmesine, aynı analizin farklı kaynaklardan alınan veriler için yeniden düzenlenerek tekrarlanmasına neden olabilmektedir.

Bu problem özellikle günümüzde bilişim kaynaklarını kullanma konusunda sıkıntı yaşayan araştırmacılar için daha da büyük bir sıkıntı haline gelmektedir. Bu nedenle araştırmacıların veritabanlarından doğrudan analiz yaptırabilir olması veya analize hazır verilere hızlı şekilde ulaşabilmeleri önemli bir gereksinim oluşturmaktadır.

2.3. Biyoinformatik Analiz Araçları

Her geçen gün daha fazla genom dizisi ortaya çıkarılmakta, bunlara ilişkin fonksiyonlar ve özellikler anlaşılmakta, protein ve genler arasındaki etkileşimler üzerine çalışılmaktadır. İçerdikleri verinin türüne göre farklı fonksiyonları olan veritabanlarının bu biyolojik verinin yönetilebilmesinde ve erişilebilir kılınmasında paha biçilemez önemleri vardır[63]. Ancak günümüzde araştırmacılar daha önce hiç olmadığı kadar çok veri toplama, saklama ve analiz etme imkânına sahiptirler [64]. Bir hastadan eş zamanlı olarak veri elde edebilen sistemler ve hastaların klinik bilgileri saklayabilen sistemler sayesinde daha önce eşi benzeri görülmemiş bir veri çağı ile karşılaşılmaktadır [27, 64, 65]. Elde edilen verinin büyüklüğü ve çeşitliliği nedeniyle pek çok analiz aracı geliştirilmeye başlanmış bu araçların kimileri spesifik veri türleri ve kaynaklar ile çalışırken kimileri ise daha genel amaçlarla kullanılabilir hale getirilmişlerdir.

Çoğu zaman pek çok farklı kaynağın ve veri türünün sonucu olarak araştırmacılar çalışmalarını yürütürken kendilerine özel geliştirdikleri betikler veya programlarla farklı kaynaklardan topladıkları verileri yeniden biçimlendirmek durumunda kalmışlardır [66].

Günümüzde bu veri yığınları içerisinden mevcut bilgileri geliştirebilecek ve insan sağlığına katkı sağlayabilecek analizleri gerçekleştiren analiz araçları geliştirmek için biyoinformatik araştırmacıları yoğun çalışmalar yürütmektedirler. On yıl öncesi ile karşılaştırıldığında geliştirilen analiz araçlarının sayısı ve türü her gün artmakta veya geliştirilen teknolojilerin tekrar kullanılabilirliği sağlanmaktadır. Bioinformatics.org, Sourceforge gibi çok amaçlı geliştirme ortamlarında yüzlerce biyoinformatik uygulamasına erişilebilmekte veya geliştiricilerin kendi kurumları üzerinden çeşitli kaynaklara erişilebilmektedir. Bunlara ek olarak BioPerl ve Biojava gibi yaygın olarak kullanılan programlama dillerinin de biyoinformatik için özel olarak geliştirdiği kütüphaneler veya eklentiler olabilmektedir [67].

Bu başlık altında biyoinformatik alanında kullanılan analiz araçları ile ilgili genel bilgiler ve bu çalışmada kullanılanlardan bazıları tanıtılmıştır.

Biyoinformatik analiz araçları da veritabanında olduğu gibi çeşitli ana başlıklar altında gruplanabilmektedir. Akademik hakemli "Nucleic Acid Reseach" dergisinin biyoinformatik analiz araçlarını ele aldığı yıllık yayınlanan özel sayısına göre bu kategoriler aşağıdaki gibi belirlenebilir [68].

- **Bilgisayar temelli araçlar:** Biyoinformatik alanında sıklıkla kullanılan programlama dilleri vb.
- **DNA araçları:** DNA dizi analizinde kullanılan, DNA dizilerinin karşılaştırılması, maniple edilmesi veya derlenmesi gibi işlemleri gerçekleştiren analiz araçları
- **Eğitim araçları:** Biyolojik teknik, materyaller ve biyoinformatik alanı ile ilgili eğitimsel içeriğe ulaşılabilen araçlar
- **Ekspresyon araçları:** Gen ekspresyonunun düzenlenmesi ve tahmini veya alternatif ekleme gibi konuları kapsayan araçlar
- **İnsan genomu:** İnsan genomuna ilişkin dizilerin, polimorfizm veya genomik bilgilerin işlendiği araçlar
- **Literatür:** Biyoinformatik alanında yapılacak literatür taramalarına veya literatürden veri madenciliği, metin madenciliği gibi yöntemlerle bilgi elde eden araçlar
- **Model Organizmalar:** Memeliden, çeşitli tek hücreli canlılara veya virüslerin genetik modellenmesinde kullanılan araçlar

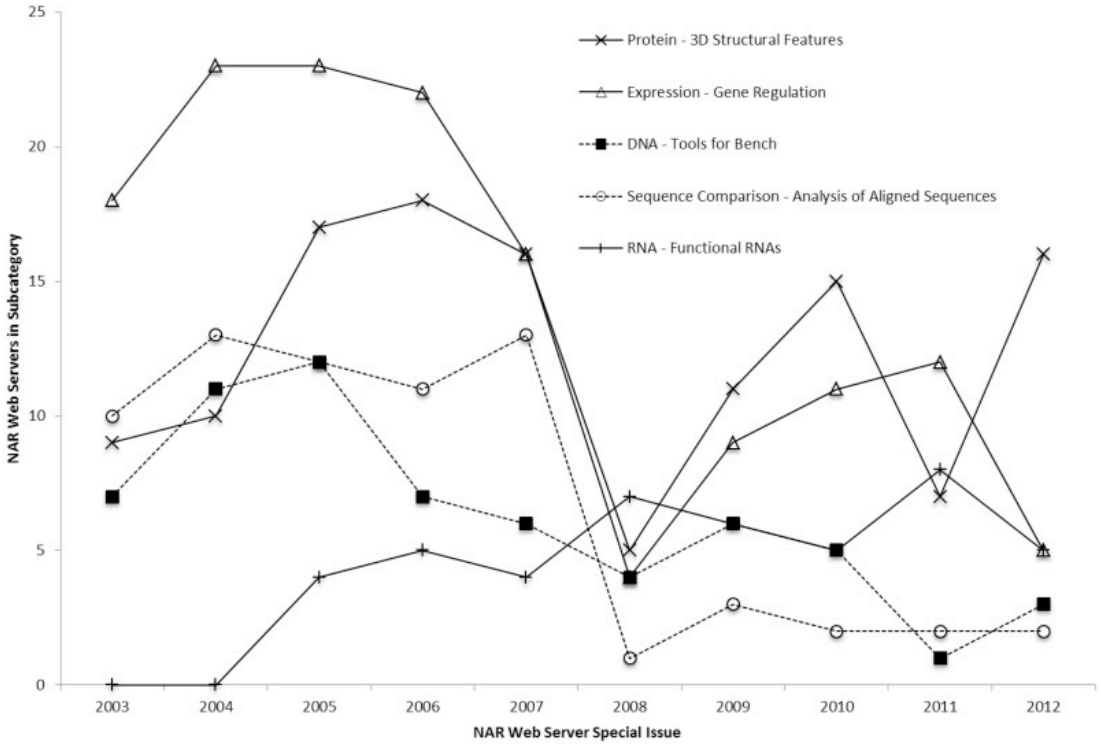
- **Diğer Moleküller:** Günümüzde yaygın olarak araştırılan genetik materyaller olan DNA, RNA ve proteinlerin dışında kalan genetik materyallere yönelik araçlar

- **Protein:** Proteinlerin dizisi ve yapısı ile ilgili analizlerde kullanılan araçlar

- **RNA:** RNA'nın fonksiyonel özellikleri, motifleri, dizileri veya görselleştirilmesinde kullanılan araçlar

- **Dizi Karşılaştırması:** Nükleik asit veya protein dizilerinin karşılaştırılması, benzerliklerinin tespit edilmesi, referans genomla ilişkilendirilmesi gibi alanlarda kullanılan araçlar.

Kısa süre içerisinde pek çok farklı kategorinin oluşmasının en büyük nedeni biyolojik verilerin çeşitliliğinde yaşanan hızlı değişimler ve teknolojinin bu değişimlerden etkilenmesi olarak gösterilebilir. Şekil 2.1 son 10 yılda bu alanda yaşanana trend değişikliklerine bir bakış açısı kazandırmaktadır.



Şekil 2.1. Son on yıl içerisinde NAR dergisi özel sayısında yayınlanan biyoinformatik araçlarının dağılımı [68]

Yukarıda bahsedilen başlıklar altında biyoinformatik çözümlemeleri için binlerce farklı uygulama geliştirilmiştir. Bu araçların bir kısmı oldukça özelleşmiş ve sadece belirli bir kaynak ile çalışabilirken, bazıları daha genel amaçlı oluşturulmuş ve farklı kaynaklardan elde edilen veri ve girdiler ile çalışabilir ve hatta birbirleri ile entegre edilebilir durumdadırlar.

Pek çok biyoinformatik araştırmacısı temelde analizleri kendileri gerçekleştirebilmek ve analizlerin gerçekleştirilmesi aşamasında biyolojik verinin getirdiği gereksinimlerden dolayı esnek davranabilmek isterler. Bunun için çoğunlukla kod veya betik temelli çalışan programlama dili veya programlama dili eklentilerini yoğun olarak kullanmaktadır. Bu eklenti ve programlama ortamları arasında en yaygın kullanılanı R istatistiksel programlama dili ve bunun bileşeni olan Bioconductor [69] paketidir. Bioconductor genetik araştırmaların herhangi bir seviyesinde (verilerin analize hazırlanması, herhangi bir seviyede analizi veya görselleştirilmesi vb.) içerdiği 600'den fazla alt paketle pek çok farklı veri türü için kullanılabilmektedir. Bu paketlere farklı örnekler verilebilir. Örneğin Affy[70] ve lumi[71] paketleri mikrodizi analizinde oldukça sık kullanılan, çip üreticilerinin ürettiği farklı formattaki verilerin ön işlemesinden son analizlerine kadar kullanılabilecek paketlerdir. Bioconductor sadece mikrodizi analizi yapabilen paketlerle sınırlandırılmamış, biyoinformatik alanında kullanılan pek çok analiz için farklı araştırmacılar ve kurumlar tarafından oluşturulmuş diğer alandaki paketleri de içerecek şekilde tasarlanmıştır. Örneğin sitometri veya PCR analizleri için kullanılabilecek flowCore[72], HTqPCR[73], qpcrNorm[74] gibi paketler bunlara örnektir. Bahsedilen biyolojik veri analizi veya mikrodizi analizinin yanında günümüzde hızla önem kazanmakta olan ve önümüzdeki zaman diliminde daha da yaygınlaşarak mevcut yöntemlerin yerine geçeceği düşünülen gelecek nesil dizileme ile ilgili pek çok paket de bu kullanışlı R birimi altında yer almaktadır. Örneğin "shortRead"[75] bu tarz verinin ön işlemesinde kullanılırken bu paketten elde edilen çıktılar ile "eastRNAseq" [76] veya "edgeR"[77] gibi paketler ileri düzey analizler gerçekleştirilebilir. Bioconductor, sunduğu pek çok farklı türdeki analiz paketine ilave olarak bu analizlerin gerçekleştirilmesinde vazgeçilmez kaynaklar olan ve eldeki biyolojik bilginin anlamlı atıflarla yorumlanabilmesini sağlayacak yardımcı kaynaklara da erişim sağlamaktadır. Örneğin "biomaRt" paketi ile BioMart'ın[78] sunduğu geniş kaynaklar analiz anında erişilebilir durumdadır.

Her ne kadar R istatistiksel programlama dili ve beraberinde sunulan paketler biyoinformatik analizlerinde çok kullanılsa da bu alanı sadece bu kaynaklarla sınırlamak ebetteki mümkün değildir. R gibi programlama dili temelli Biopython[79] veya Bioperl[80] gibi kod geliştirilerek analizlerin yapılabildiği başka ortamlar da mevcuttur. R, Python veya Perl programlama dilinin eklentileri ile beraber biyoinformatik alanında yaygın kullanımı yukarıda görülmektedir. Ancak diğer programlama dilleri de bu alanda araştırmacının ya da araştırma gruplarının deneyimlerine göre tercih edilebilirler. Örneğin analizlerde ya da araç geliştirmede kullanılan ve sistem kaynaklarını iyi kullandıkları düşünülen C/C++ veya Java gibi diller de oldukça yaygın kullanıma sahiptir.

Her ne kadar bu alanda çalışan araştırmacılar kendi araçlarını ya da esneklik kazanmak açısından kendi betiklerini yazarak programlama dillerinden faydalanmayı istiyor olsalar da yine de farklı ekipler tarafından geliştirilen yazılımlara da ihtiyaç duymaktadırlar. Farklı ekiplerin geliştirdiği iyi dokümante edilmiş araçlar, çoğu zaman araştırmacıların işlerini kolaylaştırmaktadır. Bu araçlar kimi zaman tek başına çalışan dChip [81-83], GSEA [84], IGV [85], Aracne[86] gibi araçlar olabileceği gibi kimi zaman ise web ortamından çalışan ve çeşitli programlama dillerini, farklı paketleri ya da geliştirilmiş yazılımları kullanıcılara kümelenmiş analiz sunucuları üstünden sağlamayı amaçlayan GenePattern[87] veya Galaxy[88-90] gibi ortamlar da olabilmektedir.

Görüldüğü gibi pek çok açıdan incelendiğinde biyoinformatik analizleri için araştırmacıların kullanabilecekleri özelleşmiş ya da daha genel kapsamlı ve kullanıcı tarafından özelleştirilebilecek pek çok analiz aracı mevcuttur. Ancak bu noktada en büyük handikap, veritabanlarında da olduğu gibi araştırmacıların bu kadar çok alternatifle baş edebilmelerinin imkansız olması ve farklı analiz araçlarının farklı gereksinimlerinin olması nedeniyle kimi zaman basit analizler için bile aynı işlerin tekrar tekrar yapılması gerekliliğidir.

Bu tez çalışmasında gerçekleştirilen platform, kanser ile ilgili araştırmalar yapan grupların özellikle gerekli biyoinformatik desteğine sahip olamadıkları durumlarda kolay kullanımı ile faydalanabilecekleri bir altyapı oluşturmayı amaçlamaktadır. Ayrıca çeşitli analizler için gerekli olan güçlü bilişim altyapılarına sahip olmalarına gerek kalmadan sonuca ulaşabilecekleri bir platform oluşturulması hedeflenmiştir.

2.4. Bioinformatik Araştırmalarında Kullanılan Veri Tipleri

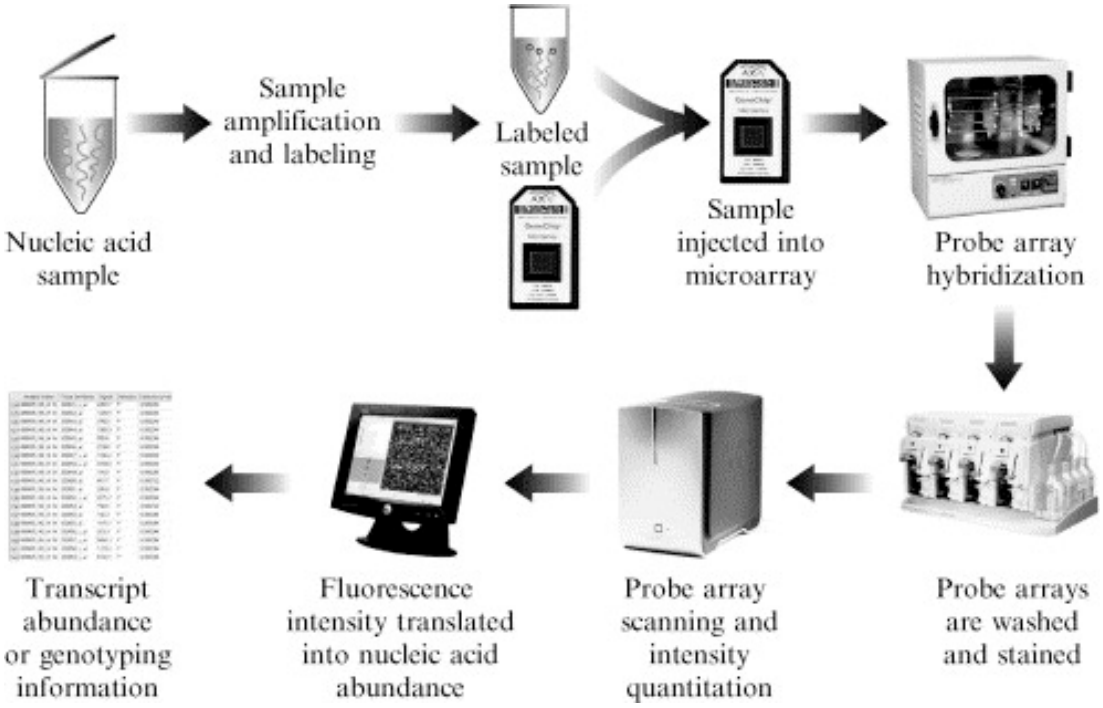
Günümüzde geniş kapsamlı genetik araştırmalardan elde edilebilecek veri kaynakları da teknolojiye paralel olarak gelişmektedir. Bu gelişim kullanılan veri kaynaklarının da çeşitlenmesine neden olmaktadır. Bu çeşitlilik aynı zamanda üretilen veriden daha çok bilgi elde edebilmeyi sağlamaktadır.

DNA'dan protein üretimine kadar geçen süreçte hatta protein üretiminden sonra bu proteinlerin birbirleri ile etkileşimi sonucu hücrede gerçekleşen aktivitelerin her birinin ölçülmesi için günümüzde çeşitli çalışmalar gerçekleştirilmektedir. Dolayısıyla hücre içerisinde genetik kodların taşındığı ve işlendiği farklı kaynaklardan veri elde edilmeye çalışılmaktadır.

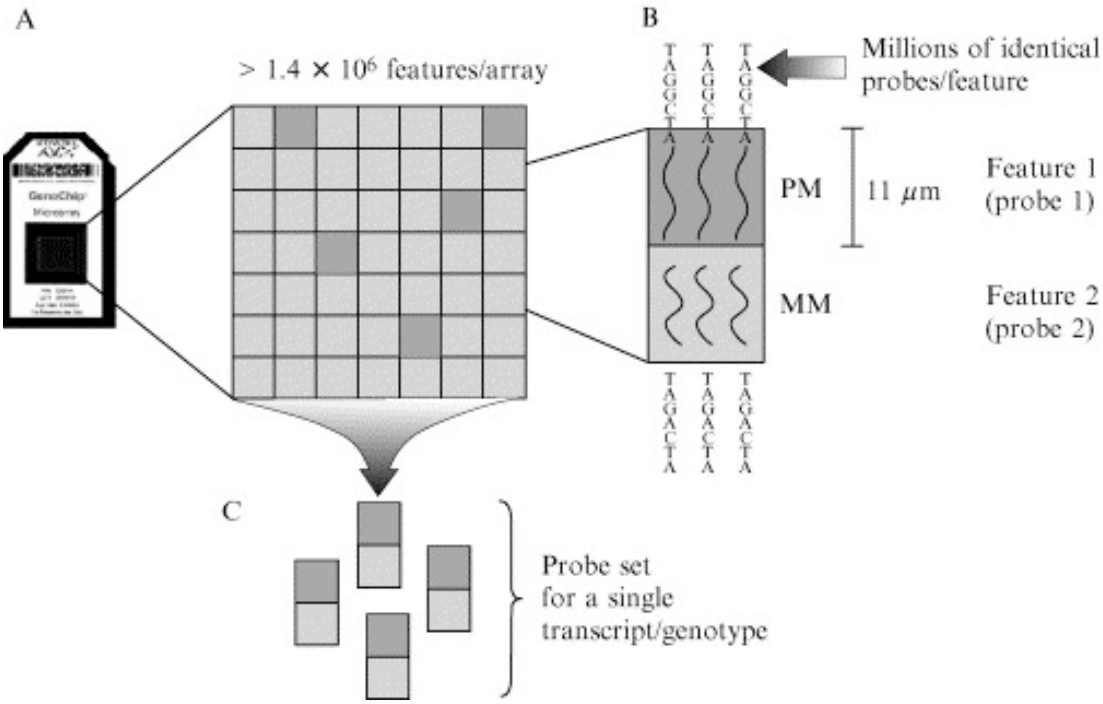
Günümüzde çeşitli firmalar genetik araştırmalarda kullanılabilecek veriler üretebilen teknolojiler üretmektedir. Genel olarak ürettikleri cihazlar benzer teknikleri kullanmakla beraber, aynı veri türünü üreten farklı yaklaşımları benimseyenler de vardır. Firmalar arasında rekabet aynı zamanda veri üretilebilen alan ve üretilen verinin kalitesi konusunda da süreli olarak olumlu gelişmeler yaşanmasını sağlamaktadır.

1997 yılında Lashkari ve arkadaşlarının [19] gen ekspresyonu verisini elde etmek için mikrodizi çiplerini kullanması ile çok boyutlu veri üretimi alanında özellikle gen ekspresyonu verisinin üretimi konusunda bir patlama yaşanmıştır. Gen regülasyonundaki farklılıkların genom üzerindeki etkileri 40 yıldan uzun zamandır araştırılan bir konudur ve gen ekspresyonu seviyesinin bununla ilişkili olduğu bilinmektedir [91]. Dolayısıyla gen ekspresyonu seviyesini gözlemlemek kompleks biyolojik olayların anlaşılması açısından oldukça önemlidir [92]. Gen ekspresyonu seviyesini ölçmek için DNA'nın sarmal yapısından elde edilen ve hücre içinde protein sentezi başlamadan proteini oluşturacak kodları taşıyan RNA'lar incelenmektedir. Bunun için günümüzde mikrodizi veya gerçek zamanlı PCR veya gelecek nesil dizileme platformları kullanılabilir [92].

Özellikle bütün genomun incelenebilmesine olanak sağladığı için mikrodizi araçlarının PCR'a göre avantajları vardır. Ancak mikrodizi çalışmaları da kendi içerisinde çeşitli dezavantajlar barındırmaktadır. Bunlardan bazılarının üstesinden gelebilmek için "The MicroArray Quality Control" (MAQC) konsorsiyumu 2006 yılında yaptığı yayınla bu çalışmalara ilişkin çeşitli kriterler belirlemiştir[93]. Günümüzde Affymetrix, Illumina, Agilent gibi firmalar farklı tekniklerle gen ekspresyonu verisini üretmektedirler. Teknikler arasındaki en temel fark üretilen verinin ölçümü sırasında kullanılan yöntemdir. Kimi firmalar tek kanallı veri üretirken (üretilen sinyal sadece ilgili dokudan ölçülen lazer sinyalinin yoğunluğuna göre ölçülmekte) kimi firmalar ise iki kanallı (ölçülen sinyal hasta ve kontrol grubu olarak nitelendirilen iki dokuda ölçülüp elde edilen sinyal değerinin logaritması alınarak hesaplanıyor) teknolojik çözümler kullanmaktadır[94]. Her ne kadar farklı teknolojiler kullanılsa da yapılan çalışmalar bu farklı üreticilerin ürettikleri verinin birbiri ile tutarlı olduğunu ortaya koymaktadır [95]. Şekil 2.2 Affymetrix firması tarafından üretilen GeneChip teknolojisi ile gen ekspresyonu verisinin nasıl ölçüldüğünü göstermektedir. Şekil 2.3'de ise kullanılan çip teknolojisinin yapısı gösterilmektedir.



Şekil 2.2. GeneChip teknolojisi ile gen ekspresyonunun ölçülmesini anlatan örnek iş akış şeması [96].



Şekil 2.3. Gen ekspresyonu ölçümünde kullanılan GeneChip teknolojisine ait kartuşların örnek yapısı [96].

Gen ekspresyonu haricinde genomdan farklı türlerde veri elde edilebilecek başka genetik yöntemler de mevcuttur. RNA üzerinden bilgi üretebilen bir diğer yöntem de gen regülasyonunda önemli etkileri olduğu bilinen [97-99] miRNA'ların ekspresyonunun ölçülmesine yardımcı olan teknolojilerdir. miRNA ekspresyon ölçümleri de RNA üzerinden ölçüldüğü için gen ekspresyonu ölçümünde kullanılan teknolojiler bu veri için de uyarlanarak kullanılabilmektedir.

Tek nükleotid polimorfizm (SNP) verileri de biyoinformatik araştırmalarında yaygın olarak kullanılan bir diğer veri türüdür. SNP'ler DNA'nın sarmal yapısı içerisinde birbiri ile eşleşmiş adenin, timin, guanin veya sitozin nükleik asit çiftlerini temsil etmektedir. İnsan genomunda 10 milyardan fazla olduğu düşünülen SNP'lerin hastalıkların kalıtsal yatkınlık veya yakalanma riski nedenlerinin incelendiği, genotip ve allel frekansı bakımından istatistiksel farkların ortaya konulmaya çalışıldığı genom çapında ilişki analizlerinde (GWAS)[100-102] yaygın olarak kullanıldığı görülmektedir. Bu analize ek olarak SNP verilerini ölçen teknolojiler gen veya SNP'lerin kopya sayılarını ölçebilmekte, bu sayede bütün genomda teorik olarak 2 kopya olması gereken genlerin kopya sayılarındaki değişimler dikkate alınarak analizler yapılabilmektedir. Bu veri türü ile heterozigotluğun incelendiği LOH analizleri de yapılabilir.

Mikrodizi teknolojisi kullanılarak üretilebilen bir diğer veri tipi ise proteinler ve etkileşimlerinin anlaşılması için oluşturulmuş ChiP-chip verileridir. Bu teknoloji ChiP ve mikrodizi tekniklerinin birbirine entegre edilmesiyle ortaya çıkmış, DNA ve proteinin etkileşimini ölçerek veri üreten bir tekniktir [103].

Günümüzde mikrodizi araştırmaları her ne kadar araştırmalarda büyük rol oynasa da kullanılan teknolojinin getirdiği dezavantajları ortadan kaldırmak için NGS (Next Generation Sequencing – Gelecek Nesil Dizileme) teknolojileri hızla yaygınlaşmaktadır. Hurd ve Nelson mikrodizi çalışmalarının kısıtlılıkları yayınladıkları bir makalede aşağıdaki gibi listelemişlerdir[104].

- Mikrodizi tasarımı daha önceden genom ve genomik özelliklerle ilgili bilgi gerektirir. Eksik, hatalı ya da güncel olmayan bilgiler hatalı sonuçlara neden olabilir

- Mikrodizi teknolojisinin en büyük engellerinden birisi çapraz-hibridizasyondur

- Yüksek sinyal ve gürültü seviyeleri hibridizasyon ile birleşerek yüksek tutarlılıkta veriler elde edilmesini engellemektedir

- DNA mikrogramlarının dizilere hibrit edilmesi gerekmektedir ve bu aşamada kullanılan PCR temelli yaklaşım örneğin düzenlenmesi aşamasında yanlılığa sebep olabilir

- Farklı teknolojilerle verilerin elde edilmesi, kullanılan farklı ön işlemeler analizlerin tekrar edilmesinde problem oluşturmaktadır

Yukarıda belirtilen mikrodizi kısıtlılıklarına NGS teknolojisi aşağıda belirtilen çözümleri önermektedir [104].

- Genomun yorumlanabilmesi için daha önceden dizinin bilinmesi gerekli değildir
- Alınan örnekler doğrudan dizilenebilir, daha önceden kullanıcı tarafından oluşturulmuş dizilerle çapraz hibridizasyon yapılmasına gerek yoktur
- Örnekten elde edilen sinyalin miktarı örneklerin birbirleri ile orantılanması ile değil bulunan dizilerin sayısı ile ifade edilmektedir. Bu da dinamik bir alanda ölçüm yapabilme olanağı sağlar
- NGS için PCR'dan kaynaklanan yanlılığı azaltmada veya yok etmede materyallerin nanogramları yeterlidir
- Veri genom çapında toplandığı için araştırmacılar RNA'dan faktör bağlanmalarına kadar bilinen ya da bilinmeyen bütün genetik bilgileri aynı veri ve platformla araştırılabilmektedir. Her bir veri türü için farklı bir veri çipi kullanılmasına gerek yoktur. Ayrıca bütün platformların aynı çıktıyı elde etmesi beklendiğinden platformlar arası tutarlılığın yüksek olması beklenmekte, bunun da çalışmaların tekrar edilebilirliğini arttıracağı tahmin edilmektedir

Son madde aynı zamanda bize NGS teknolojisi yardımı ile mikrodizi teknolojisini kullanarak elde ettiğimiz bütün verilere erişebileceğimizi de göstermektedir. NGS teknolojisinin önümüzdeki on yıl içerisinde en sık kullanılan veri üretme tekniği olması ve biyoinformatik ile beraber genetik araştırmalarında çok hızlı gelişimleri getirmesi beklenmektedir.

2.5. Genomik Verilerin Analizi

Biyoinformatik ile genomdan elde edilen verilerin istatistiksel yöntemler yardımıyla anlamlandırılması ve araştırmacılara biyolojik olarak doğrulanabilecek hipotezler sunulması amaçlanmaktadır.

Farklı yöntem ve araçlarla farklı genomik kaynaklardan elde edilen verilerin analiz edilmesinde istatistiksel yöntemler temeli oluşturmaktadır. Veri setinin içerdiği veri türüne ve verinin parametrik analiz varsayımlara uygunluğuna göre parametrik ya da parametrik olmayan yöntemler tercih edilmektedir.

Biyoinformatik alanında geliştirilen metodolojiler istatistiksel yöntemleri temel almasına karşın verinin ve biyolojik yapının getirdiği kompleks yapı kimi durumlarda aynı sonuca ulaşmak için farklı yöntemlerin denenmesini veya bilinen istatistiksel teorilerin araştırmaya göre modifiye edilmesini zorunlu kılmaktadır. Örneğin parametrik analiz varsayımlarını sağlayan iki gruba ait gen ekspresyonu

verilerini içeren bir veri setinde iki grup arasında gen ekspresyonu farkının anlamlı olup olmadığını 5'den fazla farklı test ile değerlendirmek mümkündür.

2.5.1. Entegre Analiz Yöntemleri

Genomdan elde edilen verilere uygulanan istatistiksel veya makine öğrenme temelli metotlar yardımıyla bulunan ve çeşitli anlamlılık düzeylerinde değerlendirilen sonuçların yeni bilgilere ulaşmadaki rolü açıktır. Ancak tek bir veri tipi, tek bir gen veya tek bir SNP için bulunan sonucun biyolojik olarak anlamlı olması çoğu durumda imkânsızdır. Hastalıkların yol açtığı farklılıkları anlamak ya da biyolojik olarak anlamlı hedefler bulabilmek için genomik verinin moleküler ağlarla entegre edilmesi gerekmektedir[105]. Çünkü hastalığa neden olan genlerin büyük kısmının fonksiyonel olarak ilişkili oldukları ve birbirleri ile biyolojik yolaklar üzerinde etkileşime girerek işlev gösterdikleri bilinmektedir. Böylece kompleks fenotiplerin etiyolojisine de katkıda bulundukları bilinmektedir [106-108].

Entegre edilmiş analizler veri tipinden kaynaklanan kısıtlılıkların aşılmasında da önemli getiriler sağlamaktadır. Bu nedenle farklı veri tiplerinin; örneğin gen kopya sayısı ile gen ekspresyonunun [109-113] birlikte analiz edilmesi biyolojik kompleksliği anlama açısından avantaj sağlamaktadır.

Entegre analiz yöntemlerinde de çeşitli uygulama şekilleri olabilmektedir. Kimi zaman aynı örnekten elde edilmiş farklı tipteki veriler, kimi zaman ise farklı örneklerden elde edilen veriler entegre olarak analiz edilebilir. Ancak aynı bireyden alınan örneklerden elde edilmiş veriler ile yapılan analizler daha başarılı sonuçlar verdikleri için ve yanlış pozitif oranını azalttıkları için her zaman önemlidirler [5, 114, 115].

2.6. Biyoinformatik ve Kanser

Kanser günümüzdeki en yaygın ölüm nedenlerinden birisidir ve insan vücudunda bir sistem içerisindeki pek çok organı, farklı sistemleri veya her ikisini birden etkileyebilen bir hastalıktır. Tanı koymada yaşanan zorluklar, tedaviler ve hastalığın seyri, varyasyonların çeşitliliği, süresi, lokasyonu, duyarlılığı, ilaçlara karşı direnci, hücrelerin farklılıkları ve orijinleri gibi pek çok değişken tarafından etkilenir. Gen ve proteinler arasındaki ağı ve etkileşimi gösteren ve sürekli artan kanıtlar kanserin ve moleküler mekanizmasının izlenmesinde önemli bir rol oynamaktadır. Hastalığın tanısının, tedavilerin ve sağ kalımının iyileştirilmesi için sistem biyolojisinin, klinik bilimin, genomik tabanlı teknolojilerin ve biyoinformatiğin beraber çalıştığı yeni bir konsept oluşturmak gerekmektedir [116]. Kanserin genetik değişimlerin sonucunda ortaya çıktığı günümüzde açık olarak bilinmektedir. Bu noktada klinik biyoinformatik; klinik bilgiyi, biyoinformatiği, tıp bilişimini, bilgi teknolojilerini, matematiği ve genomik bilgileri birleştiren; erken tanı, kanser hastalarına etkili tedavi ve öngörülebilir bir sağ kalım sunmayı amaçlayan kritik bir elementtir [116, 117].

Kanserin ortaya çıkasında, ilerlemesinde RNA ve proteinlerin ekspresyon seviyeleri ile bunların düzenlenmelerinin etkisi olduğu bilinmektedir [118-122]. Transkripsiyondan, protein sentezine kadar pek çok farklı değişkenin etkilediği normal yaşam döngüsü içinde yaşanan en ufak değişiklik bile büyük sonuçlar doğurmakta, kanserin gelişmesine neden olabilmektedir. Kanser, içinde barındırdığı pek çok farklı türle ve hatta türlerin kendi içinde farklı sınıflandırmalarıyla başlı başına araştırmacılar için güçlü bir odak teşkil etmektedir. Örneğin multiple myeloma (MM) bir lenfosit kanseri iken bu hastalığın bilinen alt türlerinin birbirinden çok farklı genetik değişimlerin sonucunda olduğu [123] ve bu farklıkların hastalığa karşı mücadelede büyük önem taşıdığı bilinmektedir.

Günümüzde DNA'da oluşan binlerce varyasyonun kanserin farklı türleri ile olan ilişkileri bilinmektedir [124, 125]. Ayrıca bu genetik değişimlerin fenotip ve kullanılan ilaca yanıt gibi bilgilerle birleştirilmesi hem genel anlamda hem de bireysel anlamda hastalıkla mücadele için büyük önem taşımaktadır [124]. Kişiselleştirilmiş tıp uygulamaları gelişen teknolojiye paralel olarak günümüzde pek çok alanda hızla uygulanmaktadır. İnsanın biyolojik yapısının kompleks olması bazı durumlarda tanı ve tedavi yöntemlerinin genellenememesine neden olmakta ya da aynı yöntemin farklı bireyler üstünde farklı sonuçlar üretmesine neden olmaktadır. Bu bireysel farklılıkları göz önüne alarak insanlara uygulanacak tanı ve tedavi yöntemleri, genellenmiş yöntemlere göre daha başarılı sonuçlar üretebilmektedir. Aynı zamanda fenotip özelliklerine ya da çevresel faktörlere bağlı farklılıklar da bu metotlar sayesinde dikkate alınabilmekte, tanı ve tedavide başarının artması sağlanabilmektedir.

Kişiselleştirilmiş tıptan beklenen avantajları kullanmak için günümüzde kanser tedavisinde de biyoinformatik araçlar ve genomik verinin yardımı ile hastalara kişiselleştirilmiş tedavi uygulamaları yapılmaktadır. Her ne kadar bütün genom dizisi pratikte kullanılamasa da [124, 126] günümüzde kanser tedavisinde genetik bilgiden faydalanarak kişiselleştirilmiş tıp uygulamaları ile hastalıkla mücadele eden örnekler mevcuttur [127, 128]. Yakın zamanda tüm genomu dizilemenin hasta başına maliyetlerinin 1000 doların altına düşeceği tahmin edilmektedir [129]. Hızla düşen maliyetler genomdan elde edilen bilgi ile daha çok klinik araştırma yapılabileceğini ve kişiselleştirilmiş tıp alanında daha sık başvurulacak bir kaynak olduğunu göstermektedir. Ayrıca kanser gibi çok fazla genetik değişimlerin neden olduğu hastalıklarda biyoinformatik kaynaklarına daha fazla ihtiyaç duyulacağına da bir işarettir.

GEREÇLER VE YÖNTEMLER

3.1. Veri Setleri

Bu çalışma kapsamında kullanılan veri setlerinin büyük çoğunluğu "Genel Bilgiler" bölümünde bahsi geçen, araştırmacıların yaptıkları çalışmaları yayınlamak için verilerini depolamak ve erişime açmak zorunda oldukları veritabanlarından toplanmıştır. Veritabanlarından veriler alınırken verinin sahibi tarafından gerekli izinlerin sağlanıp sağlanmadığı göz önünde bulundurulmuştur.

Bu çalışma kapsamında gerekli izinleri olan farklı kanser türlerine ait verilerin büyük bir bölümü, bu alanda en çok başvurulan ve araştırmacıların çalışmalarına ait bir yayın yapmak istediklerinde genellikle verilerini depolamaları gereken GEO[6, 7] ve ArrayExpress[8, 9] veritabanlarından toplanmıştır. Bahsedilen veritabanlarına ek olarak, Ulusal Kanser Enstitüsü tarafından kurulan "Kanser Genom Atlası Projesi" (TCGA) [2] kapsamında toplanmış ve araştırmacılarla farklı formlarda paylaşılan verilerden faydalanılmıştır. Son olarak üniversite ve araştırma kurumlarının kendi alanlarında depoladıkları, literatürde var olan ve paylaşılan veri kaynaklarına da başvurulmuştur.

Oluşturulan çalışmada depolanan veri setleri toplam örnek sayısına ve ilişkili oldukları kanser türüne göre aşağıdaki tabloda sıralanmıştır.

Tablo 3.1. Çalışmada kullanılan veri setleri ve ilişkili oldukları kanser türleri

Erişim Numarası	Örnek Sayısı	Kanser Türü
Broad-Prostate	102	Prostat
GSE10387	12	Lösemi
GSE11036	12	Lösemi
GSE11121	200	Meme
GSE11417	188	Kolon
GSE11522	20	MM
GSE12093	136	Meme
GSE12417	405	Lösemi
GSE12702	80	Prostat
GSE12896	384	MM
GSE12945	62	Kolon
GSE13557	40	Lösemi
GSE13591	158	MM

GSE13989	27	Lenfoma
GSE14230	15	MM
GSE14680	10	MM
GSE14804	21	Beyin
GSE14814	90	Akciğer
GSE14860	291	Uterus
GSE14960	23	Beyin
GSE14994	229	Böbrek
GSE15127	167	Lenfoma
GSE15526	264	Özofagus
GSE15842	67	Lenfoma
GSE15852	86	Meme
GSE16122	203	MM
GSE16125	84	Kolon
GSE16131	368	Lenfoma
GSE16406	203	Lösemi
GSE16441	68	Böbrek
GSE16558	130	MM
GSE16619	203	Meme
GSE17306	106	MM
GSE17385	6	MM
GSE17498	102	MM
GSE17536	177	Kolon
GSE18155	94	Germ Hücresi
GSE18333	82	Prostat
GSE18797	12	Akciğer
GSE18805	82	Akciğer
GSE18828	18	Beyin
GSE19388	13	Akciğer
GSE19539	140	Yumurtalık
GSE21032	743	Prostat
GSE21036	142	Prostat
GSE2113	52	MM
GSE21349	492	MM
GSE22058	397	Karaciğer
GSE23720	370	Meme
GSE2658	559	MM
GSE26863	558	MM
GSE28976	16	Meme
GSE30563	6	Beyin
GSE3141	111	Akciğer

GSE32688	96	Pankreas
GSE4475	221	Lenfoma
GSE5364	341	Meme / Kolon / Karaciğer / Akciğer / Özofagus / Tiroid
GSE5900	78	MM
GSE6344	40	Böbrek
GSE6477	162	MM
GSE6691	56	MM
GSE6980	16	MM
GSE7068	23	Akciğer
GSE7116	26	MM
GSE7390	198	Meme
GSE7425	61	Lenfoma
GSE7545	102	Meme
GSE7696	84	Beyin
GSE8894	138	Akciğer
GSE9154	42	Meme
GSE9829	288	Karaciğer
GSE9845	197	Karaciğer
LadanyiLab	128	Akciğer
TCGA-BRCA	599	Meme
TCGA-GBM	558	Beyin
TCGA-OV	594	Yumurtalık
TCGA-READ	69	Rektum
TCGA - BLCA	28	Mesane
TCGA-LUAD	229	Akciğer
TCGA-CESC	36	Serviks
TCGA-COAD	179	Kolon
TCGA-KIRC	403	Böbrek
TCGA-LAML	199	Lösemi
TCGA-LUSC	178	Akciğer
TCGA-PRAD	83	Prostat
TCGA-STAD	133	Mide
TCGA-THCA	323	Tiroid

3.2. Verilerin Önişlemesi

Farklı araştırmalardan ve kanser türlerinden elde edilen veri setlerinin biyoinformatik analizlerde kullanılabilmesi için öncelikle ön bir işleme tabii tutulmaları gerekmektedir.

Önişleme aşamasında mikrodizi veya gelecek nesil dizileme yöntemlerinden elde edilen veri, renk veya nükleik asit dizi analize uygun olarak sayısal ifadelere veya gürültüsü azaltılmış dizilere dönüştürülmektedirler.

Mikrodizide verilerin ön işlemesi özellikle büyük önem taşımaktadır. Çünkü mikrodizi teknolojisi kullanılarak üretilen veri yüksek gürültü barındırmakta, bu verinin kalitesini etkilemekte ve aynı zamanda dağılımı normal dağılımdan uzaklaştırmaktadır. Veri önişleme araçları, platformların oluşturduğu ham verileri girdi olarak kullanıp literatürde var olan algoritmalar yardımı ile gürültü seviyesini azaltmayı ve veriyi normal dağılıma yaklaştırmayı hedeflemektedirler. Takip eden bölümde öncelikle mikrodizi verileri için önişleme adımları anlatılmakta daha sonra ise gelecek nesil dizileme teknolojilerinde önişlemenin nasıl yapıldığına değinilmektedir. Teknolojik olarak yaklaşımları farklı olan bu iki yöntemin önişlemeleri veri analizinde oldukça büyük öneme sahiptir.

3.2.1. Verilerin Okunması

Mikrodizi platformları genel olarak iki formda ham veri oluşturmaktadırlar. Bunlar iki kanallı ya da iki renkli olarak düşünülen mikrodiziler (bu tür platformlar kırmızı ve yeşil renklerde iki farklı görüntü üretmekte ve analizler bunun üzerinden başlamaktadır) ve veriyi aynı renkte ışık kullanarak farklı tonlarda çıktılarla ifade tek renkli platformlardır.

Kullanılan platformun türüne göre önişleme araçları değişebilmektedir. Bazı araçlar her iki veri türünü kullanabilirken bazı araçlar sadece bir versiyon ile çalışabilmektedir.

Bu çalışmada kullanılan veri setlerinin çoğunluğu tek renkli platformlardan elde edilmiş verileri kullanmaktadır. Bununla beraber iki renkli platformlardan elde edilmiş veri setleri de mevcuttur.

Platformların genellikle görüntü dosyası şeklinde ürettikleri ham çıktılar uygun yazılımlar ile bilgisayar ortamında okunarak bir sonraki adıma hazır hale getirilmektedir. Bu aşamada genellikle mikrodizi platformları yoğunluk bilgisinin saklandığı görüntü dosyası ile beraber okunan yoğunluk bilgisinin genomun hangi bölgesine ait olduğunu da benzersiz tanımlayıcı numaraları ile işaretlemektedir. Bu benzersiz numaralar ile daha sonra üreticilerden elde edilebilecek atıf dosyaları kullanılarak ölçülen bilginin genom üzerindeki hangi lokasyona karşılık geldiği bulunmaktadır.

3.2.2. Arka Plan Düzenlenmesi

Arka plan düzenlenmesi, mikrodizi verilerde normalizasyon ve sonrası adımların gerçekleştirilebilmesi için büyük önem taşımaktadır. Mikrodizi daha önceki bölümlerde de değinildiği gibi bir lazer ışık kaynağının içinde sadece DNA'nın belirli bölgelerinin eşleşebileceği nükleik asit dizileri taşıyan binlerce küçük gözeneğe tutularak elde edilen ışının yoğunluğuna göre verinin üretildiği bir tekniktir. Ancak böyle bir teknik ile elde edilen ve mavi parmak izleri olarak nitelendirilen genetik veri içerisinde büyük gürültüleri de barındırmaktadır.

Teknolojiden kaynaklı spesifik olmayan bağlanma veya uzamsal heterojenlik gibi problemlerin ortadan kaldırılması için arkaplan düzenleme işleminin uygulanması gerekmektedir[130]. Bu nedenle farklı mikrodizi temel teknolojileri için uygulanabilecek farklı arkaplan düzenleme algoritmaları mevcuttur.

Örneğin iki kanallı veri için günümüzde yaygın olarak kullanılan ve 2007 yılında Ritchie ve arkadaşları tarafından geliştirilen daha sonra ise Silver ve arkadaşları tarafından modifiye edilen metot, gözlenen piksel yoğunluklarını sırasıyla birisi normal dağılmış diğeri ise üstel dağılmış arkaplan gürültüsü ve sinyali temsil eden rastgele 2 değişkenin toplamını temel alarak işlemektedir [130, 131]. Diğer taraftan yine iki kanallı bir başka platform için ise 2002 yılında Kooperberg ve arkadaşları ön ve arka plan yoğunluklarını kullanan bununla beraber standart sapma, hesaplamada kullanılan piksel sayısı gibi değişkenleri göz önüne alan bir başka metot önermiştir[132].

Tek kanallı teknolojilerde ise negatif kontrol probları yardımıyla arka plan düzenleme işlemi gerçekleştirilirken, negatif ve pozitif kontrol probları yardımı ile normalizasyon işlemi gerçekleştirilebilmektedir. Shi ve arkadaşları tarafından uyarlanan metodoloji, tek kanallı platformlarda arka plan gürültüsünün düzenlenmesi açısından yaygın kullanılan bir örnektir[133].

3.2.3. Normalizasyon

Normalizasyonun amacı elde edilen veriyi normal dağılıma yaklaştırmanın yanında çeşitli teknolojilerin getirdiği farklılıkları da ortadan kaldırarak biyolojik farklılıklara odaklanmaktır[134].

Normalizasyonda da prosedürler kullanılan teknolojiye göre farklılık göstermektedir. Genel olarak kullanılabilecek metotlar şu şekilde sıralanabilir [135].

- Cyclic loess
- Zıtlık Tabanlı
- Nicelik Normalizasyonu
- Ölçeklendirme

- Non-lineer metot
- RMA

Her bir yöntemin kendisine göre çeşitli avantajları olmakla beraber farklı dezavantajlar da barındırmaktadırlar. Ancak RMA ve lineer olmayan yöntemler diğerlerine göre daha sık tercih edilmektedir.

3.2.4. Gelecek Nesil Dizileme Verisinin Önişleme Süreçleri

Gelecek nesil dizileme, mikrodizi teknolojisinin giriş bölümünde anlatılan çeşitli teknik problemlerinin üstesinden gelmek için geliştirilen ve DNA'nın bütün dizisinin okunduğu, araştırmacılara sadece belirli SNP'ler ya da DNA üzerinde belirli bir bölgeye dair bilgi vermek yerine tüm diziyi sağlayan bir teknolojidir. İlk olarak 2005 yılında Roche firması tarafından yayınlanmış bir makale[23] ile araştırmalarda kullanılmaya başlanan teknoloji günümüzde sağladığı avantajlarla göreceli olarak eski sayılabilecek mikrodizileme gibi tekniklerin yakın zamanda önüne geçecek gibi görünmektedir.

Gelecek nesil dizileme teknolojisi ilk yayınlandığı zamandan günümüze kadar hızla gelişmeye devam etmiştir. Çeşitli firmalar tarafından geliştirilen ve biyolojik materyallerden genetik verinin elde edilmesini sağlayan cihazlarla ilgili en büyük eleştiri, okuyabildikleri nükleik asit dizilerinin uzunlukları olmuştur. Ancak hızla gelişen teknoloji sayesinde günümüzde oldukça uzun diziler okunabilir hale gelmiş, hatta teknolojilerin ürettikleri veri inanılmaz boyutlara ulaşmıştır [136].

Gelecek nesil dizileme teknolojileri tarafından üretilen verinin de biyoinformatik alanında ileri düzey analizlere alınmadan önce çeşitli önişlemelerden geçirilmesi gerekmektedir. Ancak üretilen verinin yapısı ve boyutları gereği çok güçlü bilgisayar kaynaklarına, depolama alanına ve analiz yazılımlarına gerek duyulmaktadır[1].

Farklı üreticiler tarafından geliştiriliyor olsalar da dizileme cihazları genel olarak aynı formatta ham veri üretmekte ve yaygın olarak kabul görmüş formatlarda bu verileri önişleme için saklamaktadırlar. Okuma verisi olarak adlandırılan bu veriler için yaygın kabul gören formatlardan bir tanesi "fastq" formatıdır. Bu formatta okunan DNA dizisi 4 satırlık metin şeklinde saklanmaktadır. Örnek bir fastq formatı şu şekilde olmalıdır:

```
@
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGT
+
!"*((((***+))%%%++)(%%%%).1***-+*"))**55CCF>>>>>>CCCCCCC65
```

Bu formata göre ilk ve opsiyonel olan üçüncü satır tanımlayıcı numaralarını ve açıklamalarını içerirken ikinci satır elde edilen dizi verisini son satır ise veriye ait çeşitli kalite göstergelerini içermektedir.

Fastq veya benzeri yaygın kabul görmüş formatta alınan gelecek nesil dizileme verilerinde ilk yapılması gereken işlem bu verilerin eğer mevcutta var ise bir genom ile ilişkilendirilmesidir. İnsan genomu başta olmak üzere araştırmalarda yaygın olarak kullanılan çeşitli türlere ait genom bilgileri bu amaçla oluşturulmuş çeşitli veritabanlanlarından elde edilebilmektedir.

Referans genomla eldeki verilerin eşleştirilmesi işlemi için farklı platformlarda geliştirilmiş yazılımlar kullanılabilir. Temel görevleri kısa parçalar halinde okunmuş nükleik asit dizilerini referans genoma eşleştirmek olan bu yazılımlar eldeki verinin büyüklüğüne, okumadan elde edilen dizinin uzunluğuna, referans genomun türüne göre farklı çalışma sürelerine sahiptirler. Günümüzde Burrows-Wheeler Aligner (BWA)[137, 138] veya Bowtie[139] gibi genellikle sistem düzeyinde işlem yapabileceğiniz araçlar yaygın olarak kullanılmakla beraber platform üreticileri de çeşitli araçlar sağlamaktadırlar.

Ham şekilde saklanan veri yukarıda bahsedilen araçlarla veya benzeri görevi üstlenen diğer yazılımlar yardımıyla referans genomda ilişkili oldukları bölgeler ile eşleştirilmektedir. Takip eden ileri analiz işlemleri ise araştırma sorularına göre farklılık göstermektedir. Kimi analizlerde belirli gen veya bölgelere ait diziler karşılaştırılabilecekken kimi analizlerde bu verilerden elde edilen diziler ek araçlarla sayısal verilere dönüştürülmekte ve analizlere devam edilmektedir.

Elde edilen dizi verisinin sayısal işlemlere dönüştürülmesi genellikle sayım araçları ile gerçekleştirilmektedir. Bu araçlar nükleik asit dizilerinin sayılması ve normalize edilmesi işlemlerini gerçekleştirmektedirler. Veriler normalize edildikten sonra ileri analizlerin gerçekleştirilebileceği formata dönüştürülebilmektedirler. Bu aşamada da "HTseq" ve benzeri araçlara ihtiyaç duyulmaktadır.

Veriler genom üzerindeki bölgelerle eşleştirilip sayım araçları ile sayılarak ve normalize edilerek farklı ortamlarda gerçekleştirilebilecek analizlere hazır hale getirilmiştir. Bu aşamadan sonra araştırmacılar ilgilendikleri analiz türlerine göre çeşitli araçlar veya geliştirdikleri betiklerden yardım alarak analizlerini gerçekleştirebilirler.

3.3. Kullanılan Biyoinformatik Analiz Yöntemleri

Bu tez çalışması kapsamında oluşturulan entegre analiz aracı 11 farklı analizi gerçekleştirebilmekte, bununla beraber ilave olarak eklenen modül ile sağ kalım ve gen ekspresyonu farkı analiz sonuçlarından elde edilen sonuçları veri setleri için karşılaştırabilmektedir. Çalışma kapsamında kullanılan analizler genel olarak 5 ana grup altında toplanmıştır. Bunlar:

- Temel analiz yöntemleri
    - Gen/miRNA Ekspresyon Farkı Analizi

    - Gen Kopya Sayısı Analizi
- Gen Ağı Analizleri
    - ARACNE Ağ Analizi
    - WGCNA Ağ Analizi
- Entegre Analizler
    - Mutasyon Analizi
    - Gen Seti Zenginleştirme Analizi
    - GemiNi Analizi
    - Gen Kopya Sayısı ve Ekspresyonu Entegre Analizi
    - Protein-Protein Etkileşimi
- Sağ Kalım Analizi
- Gen Ekspresyonu/ Sağ Kalım Temelli Karşılaştırma

3.3.1. Gen/miRNA Ekspresyon Farkı Analizi

Gen ekspresyonu basitçe genlerin protein yapılarına dönüşümleri sürecinde aktiflik düzeylerinin ölçüldüğü bir genomik veri türüdür. mRNA'lardan elde edilen mikrodizi veya gelecek nesil dizileme verileri gerekli ön işlemeler yapıldıktan sonra analize hazır hale dönüştürülürler. Bu dönüşümden sonra birden fazla fenotip grupları arasında genin aktivasyonun nasıl değiştiğini anlamak için gen ekspresyonu farkı analizi yapılması gerekmektedir.

Analiz aşamasında farklı analiz yöntemleri önerilmiş ve kullanılmaktadır. Bunlar genel olarak aşağıdaki gibi gruplanabilir.

- **t testi ile analiz:** t testi normallik varsayımlarının sağlandığı durumlarda bağımsız iki grup arasındaki ortalamaların farklı olup olmadığını test etmede kullanılan temel istatistiksel yöntemdir. Ancak düşük ekspresyon seviyelerinde (insan genomunda var olan bütün genler aynı anda aktive olmayabilirler bu da ekspresyon seviyelerinin düşük kalmasına sebep olmaktadır) t test iki grup arasında anlamlı fark bulabilmektedir. Bu problemi ortadan kaldırmak için t testini temel alarak düzenlenmiş diğer yöntemler geliştirilmiş veya başka istatistiksel yaklaşımlar temel alınmıştır[140].

- **B İstatistiği:** Bu yöntem ilk olarak Lonnstedt[141] ve arkadaşları tarafından deneysel Bayes tahmini üzerine kurularak önerilmiş daha

sonra ise Smyth[142] tarafından geliştirilmiştir. Gerçekte aktive olmamış ama t testi ile ayırt edilemeyen genlerin ayırt edilmesinde başarılı olmuş bir yöntemdir[140].

- **Bayes T istatistiği:** Baldi[143] ve arkadaşları tarafından geliştirilen ve Bayesien olasılık kuramını temel alan bu yöntem düşük ekspresyon seviyelerindeki varyans problemlerini çözmede başarılı olmuştur. Örneklem sayısının az olduğu durumlarda başarılı bir yöntemdir[140].

- **SAM:** Bu yöntem t testinde karşılaşılan küçük varyans probleminin üstesinden gelmek için t testine benzer bir analiz yöntemi izleyip daha sonra permütasyon analizi ile FDR tahminlemesi üzerine kurulmuştur[140, 144].

- **Samroc:** Bu metot iki grup arasında genlerin farklı ekspresyona sahip olma ihtimaline göre genleri sıralar. Metodun genel amacı hatalı negatif ve hatalı pozitif değerlerinin tahminlenmesidir. Genel olarak SAM ile benzerlik göstermesine rağmen paydaya eklenen değişken açısından farklılık vardır[140, 145].

Bu 5 başlık altında toplanan yöntemeler genel olarak mikrodizi çiplerinden elde edilen gen ve miRNA ekspresyonlarının analiz edilmesinde başvurulan yöntemlerdir. Geliştirilen sistemde ön işleme sonrası analizlerin gerçekleştirilmesi için R istatistiksel programlama dili içerisinde yer alan "limma"[146] paketinden faydalanılmıştır.

Bu yöntemlere ek olarak gelecek nesil dizileme verisi için de çeşitli yöntemler ve yaklaşımlar geliştirilmiştir. Bunlardan bir tanesi yine R ortamında kullanılabilecek olan ve negatif binomial dağılımı temel alarak geliştirilmiş "DESeq" paketidir[147]. RNAseq verisi için çalışmada kullanılan bir diğer analiz paketi ise "edgeR" paketidir. Bu yöntemde de bir öncekinden farklı olarak deneysel Bayes yaklaşım temel alınmıştır[77].

Gen ekspresyonu analizinde kullanılan bu yöntemler miRNA ekspresyonu analizlerinde de aynen kullanılabilmektedirler. miRNA'ler genlerin regülasyonunu düzenlemede önemli role sahip küçük nükleik asit dizileridir ve aynı genler gibi ekspresyon değerleri taşımaktadırlar. Bu ekspresyon değerlerinin değişimi genlerin aktivasyon düzeylerini etkilemekte, dolayısıyla temel birim olan proteinlerin üretilmesine etki etmektedirler. İki farklı fenotip arasında miRNA'lerin aktivasyonunu değerlendirmek bu nedenle oldukça önemlidir. Gen ekspresyonu ile benzer yöntemler ve paketler kullanıldığı için miRNA ekspresyonu fark analizi ayrıca anlatılmamıştır.

3.3.2. Gen Kopya Sayısı Analizi

Kanser, başlangıcında ve ilerlemesinde genomik değişikliklerin sıklıkla görüldüğü kompleks bir hastalıktır. Bütün insan genomuna bakıldığında kanser

vakalarında büyük miktarda gen kopya sayısı değişikliklerinin yaşandığı da bilinmektedir. Kanser türleri için genellikle genomun farklı bölgelerinde kopya sayısı değişiklikleri yaşansa da bu değişiklikler onko genleri veya tümör baskılayıcı genleri etkileyerek hastalığın başlangıcında ya da ilerlemesinde etkili olmaktadırlar[148-153].

Gen kopya sayısı değişiklikleri bu kadar sık yaşandığı ve hastalığın seyrinde büyük önem taşıdığı için, farklı kanser türlerinden elde edilen gen kopya sayısı verilerinin analiz edilmesi ve değişikliklerin sorgulanabilir olması araştırmacılar açısından oldukça önemlidir.

Bu tez çalışmasında gen kopya sayısı verilerinin elde edilebilmesi için farklı veritabanlarından toplanan Affymetrix SNP platformlarına ait veriler dChip-SNP[82] yazılımı yardımı ile analiz edilerek SNP düzeyinde kopya sayıları elde edilmiştir.

SNP'ler için elde edilen kopya sayılarının gen kopya sayısına dönüştürülmesi sırasında "UCSC hg19" genom atıflarından faydalanılmıştır. Kopya sayısının gen düzeyine dönüştürülmesi sorgu sırasında sistem tarafından gerçekleştirilmekte olup, sırasıyla öncelikle sorgulanmak istenen genin genom üzerindeki lokasyonu tespit edilmekte daha sonra bu lokasyon içinde yer alan SNP'ler matris olarak saklanan ön analizi yapılmış dosyalardan çıkarılmakta ve ilgili lokasyon içindeki SNP'lerin ortalama kopya sayısı gen kopya sayısı olarak bildirilmektedir.

3.3.3. ARACNE Ağ Analizi

Basit olarak düşünüldüğünde gen ekspresyonu genlerin aktivasyonunu gösteren bir bilgidir. Ancak bir genin ekspresyonunu ve ekspresyon seviyesinde yaşanan değişikliği tek başına ele almak yetersiz kalmaktadır. Çünkü insan genomu gibi kompleks yapılarda gen ekspresyonunun düzenlenmesinde pek çok değişken rol oynayabilmektedir. Temel olarak gen veya genler hücre içerisinde çeşitli proteinlerin kodlanmasından sorumludurlar ve genlerin kodlanması sonucu ortaya çıkan proteinler birbirleri ile etkileşerek ya da kodlanmış bir protein, protein kodlamak için var olan bir geni etkileyerek bu süreçte karmaşık görevler üstlenirler[154]. Bununla beraber gen düzeyinde yaşanan çeşitli mutasyon veya değişiklikler de geni ve dolayısıyla o genin etkileşimde olduğu diğer gen ağlarını etkileyebilmektedir.

Böylesine çok değişkenli ve kompleks bir yapı içinde genlerin birbirleri ile olan etkileşimlerini anlamak için çeşitli ağ tabanlı yöntemler geliştirilmiştir. Bu yöntemlerden bir tanesi Aracne[86, 154] metodu olarak isimlendirilmektedir. Bu yöntem mikrodizi verisinden elde edilen mRNA'lara ait ekspresyon değerlerini kullanarak gen ağları oluşturmaktadır. Ancak bunu yaparken temel istatistiksel analizlerden doğrudan faydalanmak yerine genlerin birbirleri ile olan doğrudan veya dolaylı ilişkilerini göz önüne almakta, dolayısıyla birbirleri ile ilgili olmayan ancak doğrudan istatistiksel metotlarla incelendiğinde aralarında anlamlı ilişki olabileceği düşünülen varyasyonları eleyebilmektedir. Bu işlem için genlerin proteinlerle, proteinlerin birbirleri ile veya gen transkripsiyonunu düzenleyen transkripsiyon

faktörleri ile olan ilişkilerini göz önüne alarak tersine mühendislik yöntemleri ile gen ağlarını oluşturmak hedeflemektedir [86, 154].

Aracne algoritmasını kullanmak için Califano lab tarafından geliştirilmiş C++ veya Java tabanlı uygulamalar kullanılmaktadır. Uygulamalar temel olarak aynı algoritmayı işlettiklerinden hangi versiyonun kullanılacağı tamamen araştırmacılara bırakılmıştır. Bu tez çalışmasında C++ kaynak kodları alınarak analiz sunucularında yeniden derlenmiş ve analizlerde kullanılmıştır. Bununla beraber araştırmacılar tarafından oluşturulan 2000'den fazla transkripsiyon faktörü içeren liste analizlerde girdi parametresi olarak sisteme verilmiş ve oluşturulan ağların bu TF'ler çerçevesinde görselleştirilmesine olanak verilmiştir. Görselleştirme işlemi için ise Flash tabanlı olarak geliştirilen Cytoscape Web[155] projesinden faydalanılmış oluşturulan gen ağları çalışma anında JSON formatına dönüştürülerek web eklentisi yardımı ile görselleştirilmiştir.

3.3.4. WGCNA Ağ Analizi

Bir önceki başlık altında belirtildiği gibi genlerin birbirleri ile etkileşim içinde oldukları bilinmektedir. Genlerin bu etkileşimlerini ortaya çıkarmak için geliştirilen bir diğer yöntem ise WGCNA'dir[156].

Bu yöntem gen ekspresyonu verilerini kullanarak genler arasında kurduğu ağırlıklandırılmış korelasyon analizi ile gen ağlarını belirlemede ve bu ağların fenotiplerle ilişkilendirilmesinde kullanılan bir yöntemdir[156].

Geliştiriciler tarafından R paketi olarak sunulan analiz yönteminin kullanılması için çeşitli parametrelerin veri setine göre optimize edilmesi gerekmektedir. Aynı zamanda bellek kullanımı konusunda da kısıtlılıkları bulunan yöntemin kullanılabilmesi için farklı fenotipik özelliklere ait verileri içeren gen ekspresyonu veri setleri, her gen için ortanca değeri kullanılarak genler en yüksek ekspresyon değerinden en düşük ekspresyon değerine olacak şekilde sıralanmış, ilk 5000 gen analiz için kullanılmıştır. Yöntem kendi görselleştirmelerini ve genlerden elde ettiği modüllerin fenotipler ile olan ilişkilerini çıktı olarak üretmektedir.

Yapılan çalışmada geliştirilen analiz modülü gen ekspresyonu verisi içerisinden en yüksek ekspresyon değerine sahip 5000 geni seçme, parametrelerin ayarlanması ve verinin fenotiplere göre hazırlanması işlemlerini gerçekleştirdikten sonra WGCNA analiz paketini R ortamından çağırarak gerekli işlemleri gerçekleştirmektedir.

3.3.5. Gen Seti Zenginleştirme Analizi

Her ne kadar genetik bilgiden çeşitli istatistiksel yöntemler yardımıyla genlere ve ağlarına ilişkin bilgiler elde edilebilir olsa da bunların biyolojik olarak anlamlı ve bilinen yolaklarla ilişkilendirilmesi araştırmacılar açısından büyük önem taşımaktadır.

Gen seti zenginleştirme analizi biyolojik olarak bilinen ya da çeşitli kaynaklardan derlenmiş yolaklar ile eldeki verinin kullanılarak hangi yolakların aktif olduğunu ya da hangi yolaklarda değişimlerin yaşandığını tespit etmeye yönelik analizleri içermektedir.

Klasik olarak kullanılan analizlerden farklı olarak bu çalışmada kullanılan GSEA[84] aracı genleri iki veya daha fazla sınıf arasında gruplayarak sıralamadaki yerlerini temel alan analizler gerçekleştirmektedir. Bu sayede aşağıda sayılan klasik analiz yöntemleri ile yaşanabilecek problemlerin aşılması öngörülmektedir[84].

- GSEA ile iki sınıf arasında sınırlı istatistiksel farklılık var olsa bile genler analiz içinde kalabilmekte
- Analiz sonucunda elde edilen anlamlı gen listeleri çok uzun olsa bile araştırma için biyolojik olarak bir anlam içermiyorlarsa analizin dışında bırakılabilmekte
- Bir biyolojik yolaktaki genlerin tamamını incelemek yerine bir kısmında yaşanan değişiklikleri de göz önüne alarak analizleri gerçekleştirmekte (örn: bir yolaktaki genlerin %20'sinde yaşanan değişiklik o yolağın farklı çalışmasına ya da diğer genleri etkilemesine neden olabilir)

Analiz için aynı algoritmayı temel alan farklı platformlarda çalışabilen sistemler geliştirilmiştir. Bu çalışmada Java tabanlı olarak geliştirilen GSEA aracı kullanılmaktadır. Analizlerde gen ekspresyonu verisi analiz modülü tarafından fenotip bilgileri ile beraber hazır hale getirildikten sonra analizde kullanılacak yolak veritabanı tercihlerine göre analiz yürütülmektedir. Broad Enstitüsü tarafından oluşturulan ve yönetilen moleküler yolak veri tabanları 6 temel kategoriye ayrılmıştır. Bunlar:

- **Pozisyonel gen setleri:** Genlerin kromozomlar üstündeki pozisyonlarına göre yolaklar oluşturulmuştur
- **Oluşturulmuş gen setleri:** Bu setler literatür ve online veritabanları yardımı ile araştırmacılar tarafından oluşturulmuş gen setlerini içermektedir
- **Motif gen setleri:** Bu setler gen regülasyonunda biyolojik olarak çalışan miRNA, transkripsiyon faktörü gibi çeşitli motifleri içeren gen setleridir
- **Hesaplamalı gen setleri:** Bu setler büyük kanser veri setlerinden yapılan analizler sonucu oluşturulmuş ve kanser ile ilgisi olabilecek gen setlerini içermektedir

- **Gen Ontolojisi (GO) setleri:** Biyolojik olarak hücre içindeki ve yaşamsal döngüdeki rolleri bilinen gen setlerini içermektedir

- **Onkogen setleri:** Bu setler ise onkogen olarak bilinen ve kanserin başlangıcında ve ilerlemesinde etkisi olduğu biyolojik olarak ortaya konulmuş genlere ait veri setleri içermektedir.

3.3.6. GemiNi (Gen ve miRNA Entegre Analizi)

GemiNi metodu gen ve miRNA ekspresyon verilerini kullanarak ileri beslemeli modellemeyi temel alan bir yöntemdir.

miRNA'ların gen regülasyonundaki rolünü göz önüne alarak, transkripsiyon faktörlerini de analize dâhil eden yöntem ileri beslemeli ağlar kurarak mRNA ve miRNA verileri arasındaki ilişkiyi transkripsiyon faktörleri üzerinden ortaya koymayı amaçlamaktadır[157].

Temel olarak iki farklı fenotipe ve aynı kişiden elde edilmiş örneklere ait mRNA ve miRNA verilerini girdi olarak kullanan metot, daha önceden oluşturulmuş gen-TF, miRNA-TF, gen-miRNA ilişkilerini kullanarak ileri beslemeli ağlar oluşturmakta ve bu ağların mevcut fenotipler üzerindeki değişimlerine göre ileri beslemeli ağların anlamlılıklarını ortaya koymaktadır[157].

Permutasyon temelli analiz gerçekleştiren GemiNi algoritması R platformu üzerinden analizi gerçekleştirmekte, analizde daha önceden oluşturulmuş ileri beslemeli ağ bilgilerini yine R ortamında saklanan veri yapılarından elde etmektedir. Geliştirilen proje içerisinde kurulan analiz katmanı uygun veriler sisteme aktarıldıktan sonra tetiklenerek analizi gerçekleştirebilmekte ve analiz sonucunda üretilen grafik ve ileri beslemeli ağ çıktılarını dosya depolama alanlarında saklayarak kullanıcılarla paylaşılabilir kılmaktadır.

Ayrıca geliştirilen projeye ek olarak eklenen modül ile araştırmacıların kendi verilerini çalışma zamanında analiz edebilmelerine de olanak sağlanmıştır. Uygun formattaki veri sisteme yüklendikten sonra analiz parametreleri ayarlanarak analizler sistem üzerinden R ortamında deneyimsiz araştırmacılar için gerçekleştirilebilir.

3.3.7. Kopya Sayısı ve Gen Ekspresyonu Entegre Analizi

Gen kopya sayısı analizinin anlatıldığı bölümde kanserin farklı türlerinde genlerin kopya sayısında sıkça değişimler yaşandığından, bu değişimlerin hastalığın ortaya çıkmasında ve ilerlemesindeki rolünden bahsedilmiştir. Genlerin kopya sayısında yaşanan değişimler aynı zamanda genlerin ekspresyon seviyelerini de etkileyebilmektedir. Bu etkileşimin ortaya konulabilmesi için iki farklı veri türünü analiz edebilen bir yöntemin kullanılması gerekmektedir. Ancak bu değişimlerin incelenmesi aşamasında gen ekspresyonunun farklı etmenler tarafından da kontrol edildiğinin bilinmesi önemlidir.

Bu çalışma kapsamında genlerin kopya sayısı ve ekspresyonuna ait verileri entegre edebilmek için DRI[158] isimli algoritma kullanılmıştır. Yöntemin temeli korelasyon analizine dayanmakla beraber eş örneklerden elde edilmiş kopya sayısı ve ekspresyon verilerini genlere göre değerlendirmektedir. Yöntem analiz sırasında rastgele permutasyon analizi gerçekleştirerek genler için korelasyonun yanında q değerlerini de vermekte ve düzeltilmiş olan bu değerler ile hatalı olabilecek ya da düşük anlamlılık değerine sahip olan genler elenmektedir.

R istatistiksel programlama dili ortamından erişilen analiz paketine tez çalışması kapsamında oluşturulan analiz katmanından erişim sağlanmakta, eş gen ekspresyonu ve kopya sayısı verileri ile 1000 permutasyon üzerinden analizler gerçekleştirilmektedir. Analiz sonuçları veri setleri için veritabanına aktarılmakta ve sorgulanabilir hale getirilmektedir.

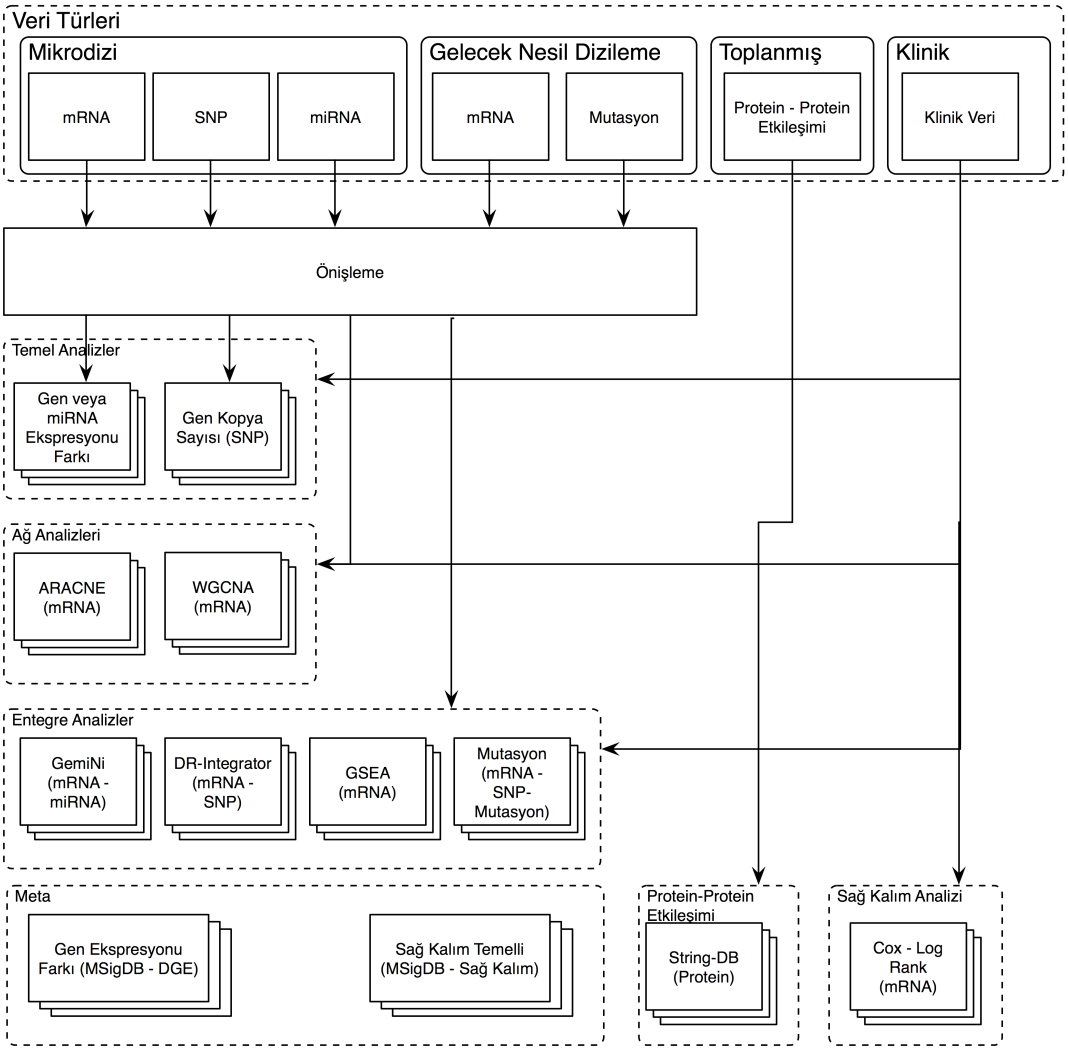
3.3.8. Mutasyon Analizi

Genlerin aktivasyon seviyelerindeki veya kopya sayılarındaki değişimi anlamanın ve bunu biyolojik olarak anlamlandırmanın yanında, geni oluşturan nükleik asit dizilerinde nasıl değişikliklerin gerçekleştiğini bilmek de araştırmalar açısından büyük öneme sahiptir.

Gen üzerinde yaşanan mutasyonlar genlerin fonksiyonlarını etkilemekte dolayısıyla aynı sistem içerisinde çalışan diğer genlere ve sistemin bütününe etki etmektedirler.

Gen mutasyon bilgileri SNP analizlerinden elde edilebileceği gibi gelecek nesil dizileme teknolojileri ile de tespit edilebilmektedir.

Bu çalışma kapsamında oluşturulan mutasyon bilgileri TCGA projesi için toplanan ve araştırmacılar ile paylaşılan verilerden elde edilmiştir. TCGA araştırmacılara ham veriden gen ile ilişkilendirilmiş 4. seviye veriye kadar farklı aşamalarda veriler sunmaktadır. TCGA üzerinden farklı kanser türleri için toplanmış mutasyon verileri RNAseq ve kategorize edilmiş gen kopya sayısı verisi ile Broad Enstitüsü tarafından oluşturulan genom veri analiz merkezine ait “Firehose” portalından elde edilmiştir. Veritabanında depolanan mutasyon bilgileri çalışma zamanında R ortamından erişilerek RNAseq ve kopya sayısı bilgileri ile birleştirilmekte, araştırmacılara kopya sayısına bağlı gen ekspresyonu değişiminin görselleştirilmesi ve mutasyon türünün etkileri aynı grafik üzerinde aktarılmaktadır.



Şekil 3.1. Veri türüne göre analiz akışlarının genel görüntüsü

3.4. Web Tabanlı Biyoinformatik Analiz ve Sorgu Aracının Geliştirilmesi

Çalışma kapsamında bahsedilen analizlerin yapılabilmesi ve özellikle araştırmacılar için görsel olarak sunulabilmesi için birden çok sunucu üzerinden çalışan çok katmanlı mimari ile çalışan bir analiz ve web portalı geliştirilmiştir.

Geliştirilen platformun genel görüntüsü şekil 3.2'de gösterilmektedir.

3.4.1. Kullanıcı Arayüzleri

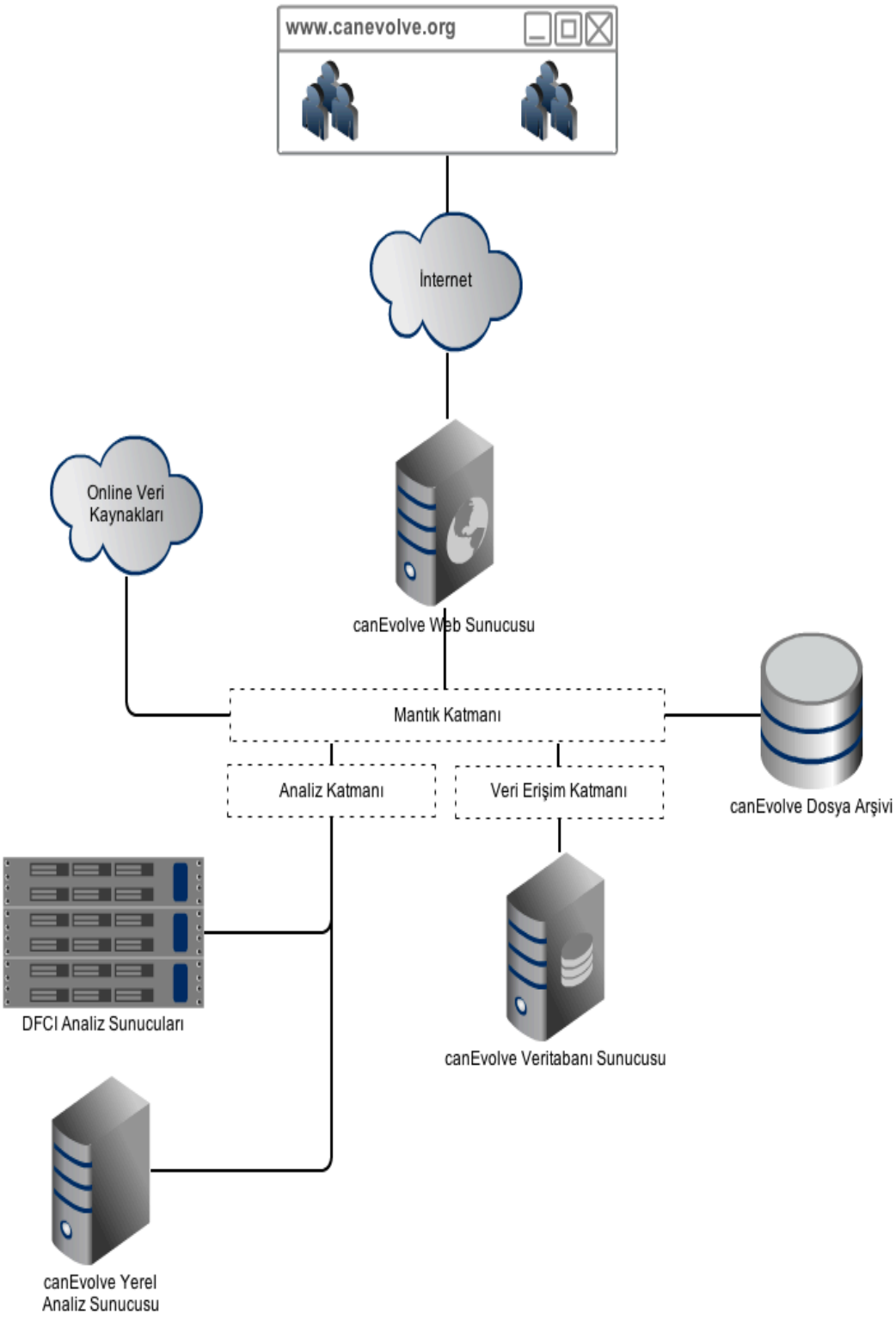
Kolay kullanımı ve dağıtımı desteklemek amacıyla web tabanlı olarak geliştirilen kullanıcı arayüzleri, kullanıcıların kanser veri setleri ile ilgili aradıkları çeşitli bilgilerine tek ortamdan erişebilmelerini sağlamak amacıyla analiz tiplerine göre gruplanmıştır.

Kullanıcılar platforma ait domain adresi ile erişim sağladıklarında veritabanında depolanan analizleri sorgulayabilecekleri arayüz ile karşılaşmaktadırlar.

Kullanıcı arayüzleri kullanıcıların sorgulama anında kesinti yaşamalarını engellemek amacıyla AJAX(Asynchronous JavaScript and XML – Asenkron JavaScript ve XML) araçlarıyla donatılmış ve asenkron çağrılar sunucu tarafına göndererek analizlerin ve sonuçlarının getirilmesi sağlanmıştır.

AJAX çağrıları esnek web tasarımı sağlaması nedeniyle HMTL(Hyper Text Markup Language – Zengin Metin Dili) ve CSS(Cascading Style Sheets - Basamaklı Stil Şablonları) kullanılarak zenginleştirilmiştir. Bununla beraber sistem PHP programlama dili kullanılarak hazırlanmış arayüzler de içermektedir. PHP programlama dili ile geliştirilen ve asenkron olarak çağrılan modüller ayrıca web sunucusu üzerinden kullanıcıya ait istekleri almakta ve yönetmekte de kullanılmaktadır.

Kullanıcı arayüz deneyiminin geliştirilmesi için ek olarak DOJO JavaScript kütüphanesinden de faydalanılmış, çeşitli arayüz fonksiyonları bu kütüphane yardımıyla zenginleştirilmiştir.



Şekil 3.2. Geliştirilen platformunun genel görüntüsü

3.4.2. Web Sunucusu ve Mantık Katmanı

Çalışmada kullanılan web sunucusu projenin başlangıcından itibaren yürütülen iş birliği kapsamında Harvard Üniversitesi / Dana Farber Kanser Enstitüsü tarafından desteklenen altyapı çalışmaları nedeniyle Amerika Birleşik Devletlerin'de tutulmaktadır.

canEvolve web sunucusunun temel görevi internet üzerinden kullanıcılar ile iletişimi sağlamak ve kullanıcıların isteklerine göre iş akışını yönetmektir. Bu işlemlerin gerçekleştirilebilmesi amacıyla çok katmanlı mimari temel alınarak canEvolve omurga sistemi oluşturulmuştur. Sanal olarak katmanlandırılmış bu sistemin tepesinde mantık katmanı yer almaktadır.

Mantık katmanının temel görevleri aşağıdaki gibi tasarlanmıştır:

- Sistem üzerindeki bütün iş akışlarını kontrol etmek
- Kaynak kullanımını optimize etmek
- Analiz katmanına analiz için gerekli parametre ve depolama kaynaklarını sağlamak
- Veri iletişim katmanı ile beraber çalışarak veritabanı kayıtlarını yönetmek
- Online veri kaynaklarına erişimleri yönetmek, dış kaynaklardan otomatik olarak getirilen verileri arşivlemek
- Web sunucu üzerinden kullanıcı sorgularına ait çıktıları kullanıcı ekranlarına göndermek

Mantık katmanı, geliştirilen platformun merkezinde olup bütün işlemlerin doğru şekilde yönetilmesinden sorumludur. Kullanıcılar bir kanser türüne ait sorgu gönderdiklerinde sorgunun türüne göre diğer katmanlar ile iletişime geçerek sonuçları hazırlayan mantık katmanı, aynı zamanda sistem yöneticileri tarafından sisteme eklenecek yeni veri setleri ve bu setlere uygulanacak analizlerin de iş akışı içerisinde yönetilmesine olanak sağlayacak şekilde tasarlanmıştır.

3.4.3. Analiz ve Veri Erişim Katmanları

Analiz ve veri erişim katmanlarının temel görevleri sırasıyla biyoinformatik analizlerini yönetmek ve veritabanına erişim sağlayarak veritabanı kayıtlarını yönetmektir.

Analiz katmanı, mantık katmanının analize hazır hale getirdiği verileri yine mantık katmanı sayesinde elde ettiği parametreler ile analiz etme görevini üstlenecek şekilde tasarlanmıştır. Platform analizlerinin büyük bölümünü R tabanlı kodlar

üzerinden yürütülüyor olsa da bunlara ilave olarak C++, Java ve Python tabanlı analiz modüllerini de çalışmaktadır.

Analiz modülü iki farklı analiz türünü gerçekleştirebilmektedir. Bunlardan birincisi kullanıcılardan gelen sorgu türüne göre ön analizleri yapılmış verileri kullanarak ileri analizleri tamamlamak ve ilgili çıktıları geçici depolama alanında hazırlamaktır. Çalışma zamanında karşılaşılan ileri analizlere ek olarak bu modül aynı zamanda belirli analizler için kullanıcıların kendi veri setlerini kullanarak analiz başlatmalarına olanak verecek şekilde oluşturulmuştur. Ancak kullanıcılar platformun bütün analiz özelliklerini kullanamamaktadır. Bunun nedeni analiz seviyesinde kullanılan sistem kaynaklarının iş birliği kapsamında belirli ölçüde kullanılabilir olması ve sistemin gereksiz analiz yüklerinden uzak tutulmak istenmesidir.

Analiz modülü aynı zamanda sadece canEvolve yöneticileri tarafından kullanılabilen ve platformun desteklediği bütün analizlerin yürütülebildiği, sonuçlarının yönetici onayı ile analiz sonunda kullanıcılar ile paylaşılabildiği yönetim analiz modülünden gelen istekleri de karşılamakla sorumludur.

Bu modül, kendisine mantık katmanı tarafından gönderilen analizleri gerçekleştirmek için sistem seviyesinde erişim hakkına sahip olduğu bir başka yerel analiz sunucusunu veya DFCI'da görev alan bütün araştırmacıların ortak olarak erişebildikleri 512 işlemcilik analiz kapasitesine sahip kümelenmiş analiz sunucusuna da kullanabilmektedir.

Analizlere ait iş akışları ve analiz kod şablonları daha önceden sisteme yöneticiler tarafından tanıtılmakta ve testler tamamlandıktan sonra analizin çalışacağı yer ile beraber analiz katmanının yerel konfigürasyonu içerisine aktarılmaktadır. Yeni bir analiz geldiğinde katman konfigürasyonda belirtilen sistem kaynaklarını kullanarak ilgili analiz sunucusuna uzak bağlantı gerçekleştirip analizleri sonuna kadar devam ettirmekte ve oluşan çıktıları dosya arşiv alanı ve/veya veritabanına mantık katmanı üzerinden taşımaktadır.

Veri erişim katmanı ise veritabanında gerçekleştirilen işlemlerin nesne temelli olarak yürütülmesini ve efektif kod kullanımını sağlamak, veritabanı bağlantı yönetimini tek bir noktadan yapmak adına kurulmuştur. Veritabanı üzerinde yapılacak bütün işlemler bu katman yardımıyla yapılmaktadır.

BULGULAR

4.1. Geliştirilen Platformun İçeriği

Geliştirilen kanser genomik portalı "www.canevolve.org" domain adresi üzerinden araştırmacıların kullanımına sunulmuştur. Kanser araştırmacılarının basit veya entegre analiz yöntemleri ile analiz edilmiş veri setleri üzerinden araştırdıkları sonuca kolayca ulaşmalarını sağlayacak şekilde tasarlanan sistem, standart gen sembollerini kullanarak pek çok analizin sonucunu araştırmacılar ile paylaşabilmektedir.

Mevcut haliyle platform üzerinde 98 veri setine ait 20000'den fazla hastadan toplanmış genomik bilginin farklı yöntemler ile analiz edilmiş sonuçlarına ulaşılabilmektedir. 10'dan fazla entegre ve basit analiz yöntemini destekleyen platformun içeriği tablo 4.1 de özetlenmiştir.

Tablo 4.1. Kullanılan analizler ve içerdikleri veri seti sayıları

Analiz Grubu	Analiz	Kullanılan Yöntem	Toplam Veri seti
Temel	Gen Ekspresyonu Farkı	LIMMA	68
Temel	miRNA Ekspresyonu Farkı	LIMMA	19
Temel	Kopya Sayısı Değişikliği	dChipSNP	32
Ağ	Düzenleyici Ağ	ARACNE	13
Ağ	Korele Ekspresyon	WGCNA	16
Entegre	Gen Seti Zenginleştirme	GSEA	16
Entegre	Gen Ekspresyonu ve miRNA Eksrepsyonu	GemiNI	6
Entegre	Gen Ekspresyonu ve Kopya Sayısı Değişimi	DR-Integrator	6
Entegre	Mutasyon	TCGA (RNAseq / Gistic2)	13
Sağ Kalım	Sağ kalım analizi	Survival	22

4.2. Kullanıcı Arayüzleri

Analiz sonuçlarının sorgulanabilmesi veya sistem yöneticileri tarafından ön işlemesi yapılmış veri setleri üzerinde analizlerin gerçekleştirilebilmesi için oluşturulmuş "Sorgu" arayüzü, basit ve aynı mantığın farklı analizlerde kullanılabilmesine olanak verecek şekilde tasarlanmıştır.

Gen ekspresyonu, kopya sayısı, ARACNE, sağ kalım analizi, kopya sayısı ve gen ekspresyonu entegre analizi gibi analizler kullanıcılar tarafından belirlenen veri setleri ve gen sembolleri kullanılarak gerçekleştirilmektedir. miRNA ekspresyonu analizinde ise gen sembolü yerine miRanda sembollerinin kullanılması gerekmektedir.

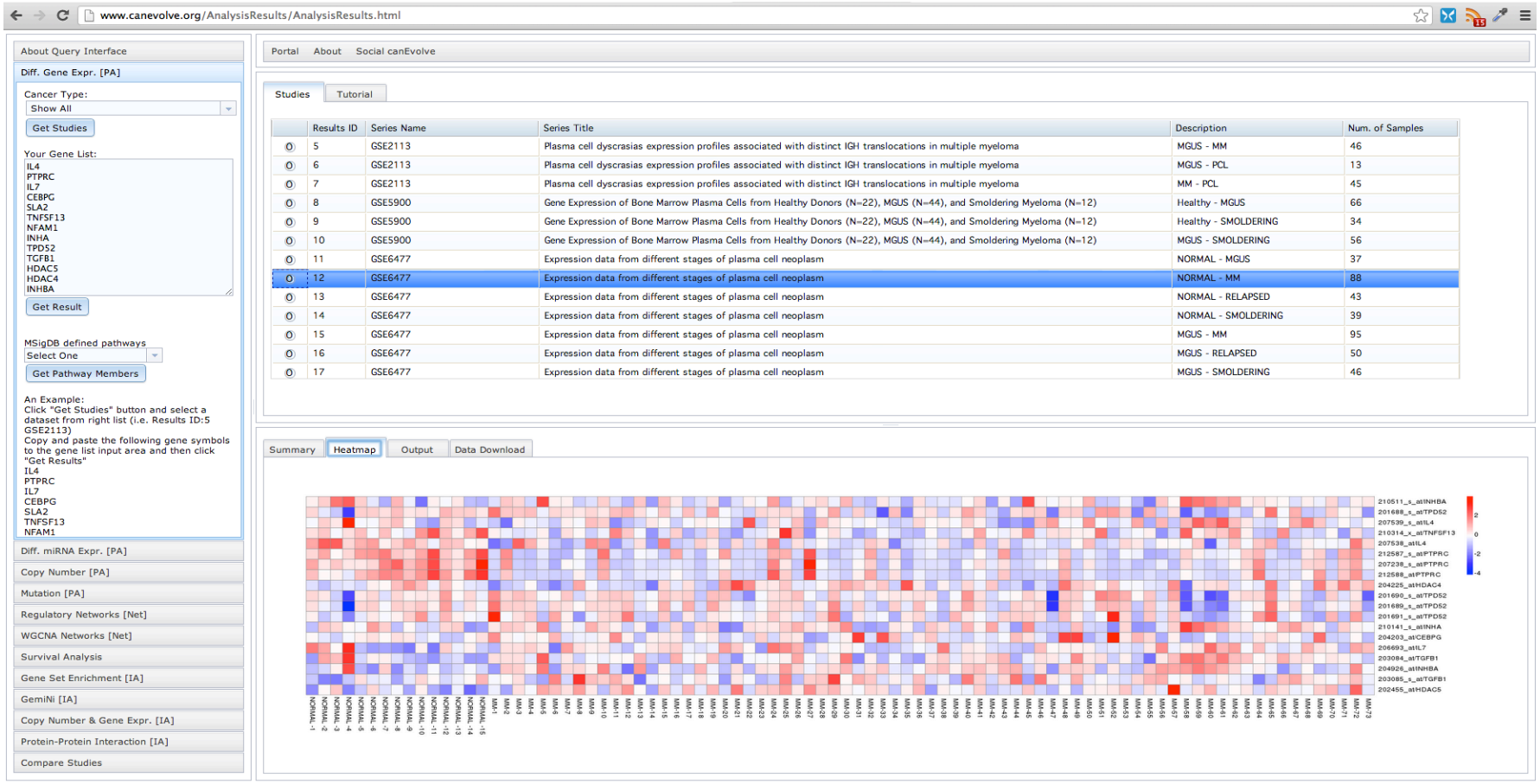
Bunlara ek olarak sadece çalışma seçilerek WGCNA, gen seti zenginleştirme ve GemiNi ile analiz edilmiş verilerin sonuçlarına ulaşılabilmektedir. Bu analizler diğer analizlerden farklı olarak çalışma anında gerçekleştirilmemekte, platform tarafından daha önce yapılmış analizler üzerinden sorgulanarak kullanıcılara aktarılmaktadır.

Ayrıca protein-protein etkileşimi de yayınlanmış biyolojik kaynaklardan sadece insan genomunu kapsayacak şekilde toplanmış ve birleştirilerek saklanmıştır. Çalışma anında sadece veritabanı sorgusu ile elde edilen veri üzerinden görselleştirme sağlanmaktadır.

Sorgu arayüzünün diğer analizlere göre çalışma anında farklı tutum izlediği tek modül gen ekspresyonu veya sağ kalım analizlerinin karşılaştırıldığı modülüdür. Bu modül araştırmacı tarafından belirlenen yolaklar için aynı veri türünü kullanan farklı veri setlerindeki sonuçların karşılaştırılabilmesi için geliştirilmiştir. Analiz çalışma zamanında gerçekleştirilmekte olup kullanıcılara analiz sonuçlarını tablo ve görseller halinde sunabilmektedir.

Sorgu arayüzünde gen sembolleri kullanılarak analiz ve sorgulama gerçekleştirilen modüllerde yolaklarla da işlem yapmak mümkündür. Bunun için MSigDB tarafından oluşturulmuş yolaklardan seçim yapıldıktan sonra üye genlerin analiz için kullanılabilir olması sağlanmıştır.

Sorgulama arayüzüne ek olarak kullanıcılara açılan ve kendi veri setlerini veya online kaynaklarda depolanmış veri setlerini analiz edebilmelerini sağlayan arayüzler de ilgili analizin parametrelerine göre hazırlanmış ve erişime açılmıştır. Hali hazırda GemiNi ve gen ekspresspresyonu farkı analizlerini gerçekleştiren arayüzlere platform üzerinden erişim mümkündür. Sisteme ait bir ekran görüntüsü örneği Şekil 4.1'de verilmiştir.



Şekil 4.1. Kullanıcı arayüz örneği

4.3. Platformun Kullanılması ve Örnek Senaryolar ile İçeriğin Doğrulanması

4.3.1. Gen Ekspresyonu Farkı Analizi

Gen ekspresyonu farkı analizi çalışma zamanında kullanıcı tarafından belirlenen gen sembolleri veya seçilen yolağa üye gen sembollerini kullanarak seçilen veri seti üzerinde analizi gerçekleştirmektedir.

Analiz kullanıcı tarafından başlatıldıktan sonra sunucu tarafında ilgili katmanlarda çalıştırılmaktadır. Analiz sonucunda işlem özeti, grafik, sonuç tablosu, kullanılan veri seti ile oluşturulan ısı haritasının indirilebilmesini sağlayan indirme seçenekleri arayüze çıktı olarak aktarılmaktadır.

Mevcut versiyon 12 farklı kanser türü için 68 farklı karşılaştırmaya olanak vermekle beraber içerdiği 4000'e yakın yolak bilgisi ile biyolojik olarak anlamlandırılabilecek sorgular gerçekleştirmeyi olanaklı kılmaktadır.

Geliştirilen platform yardımı ile yapılan ve literatür yardımı ile doğrulanan bir kaç örnek senaryo aşağıdaki gibidir.

GSE6477 veri setini kullanarak Tiedemann ve arkadaşları tarafından yapılmış bir çalışmada multiple myeloma ile ilgili olarak 15 kinaz geni üzerinde değişiklikler rapor edilmiştir[159]. Bu genler geliştirilen platform yardımı ile aynı veri seti üzerinde analiz edildiğinde benzer sonuçlara ulaşıldığı aşağıda görülmektedir.

Bu çalışmaya ait analiz sonuçları Tablo 4.2'de verilmiş ve Şekil 4.2'de ekspresyon değerleri z skoruna dönüştürülerek ısı haritası ile görselleştirilmiştir.

Literatür incelendiğinde platformun depoladığı böbrek kanserine ait bir veri seti ile yapılan çalışmada "sodyum iyon taşıma" yolağına ait üye genlerin normal bireyler ile kanserli vakalar karşılaştırıldığında farklılık gösterdiği bildirilmiştir[160].

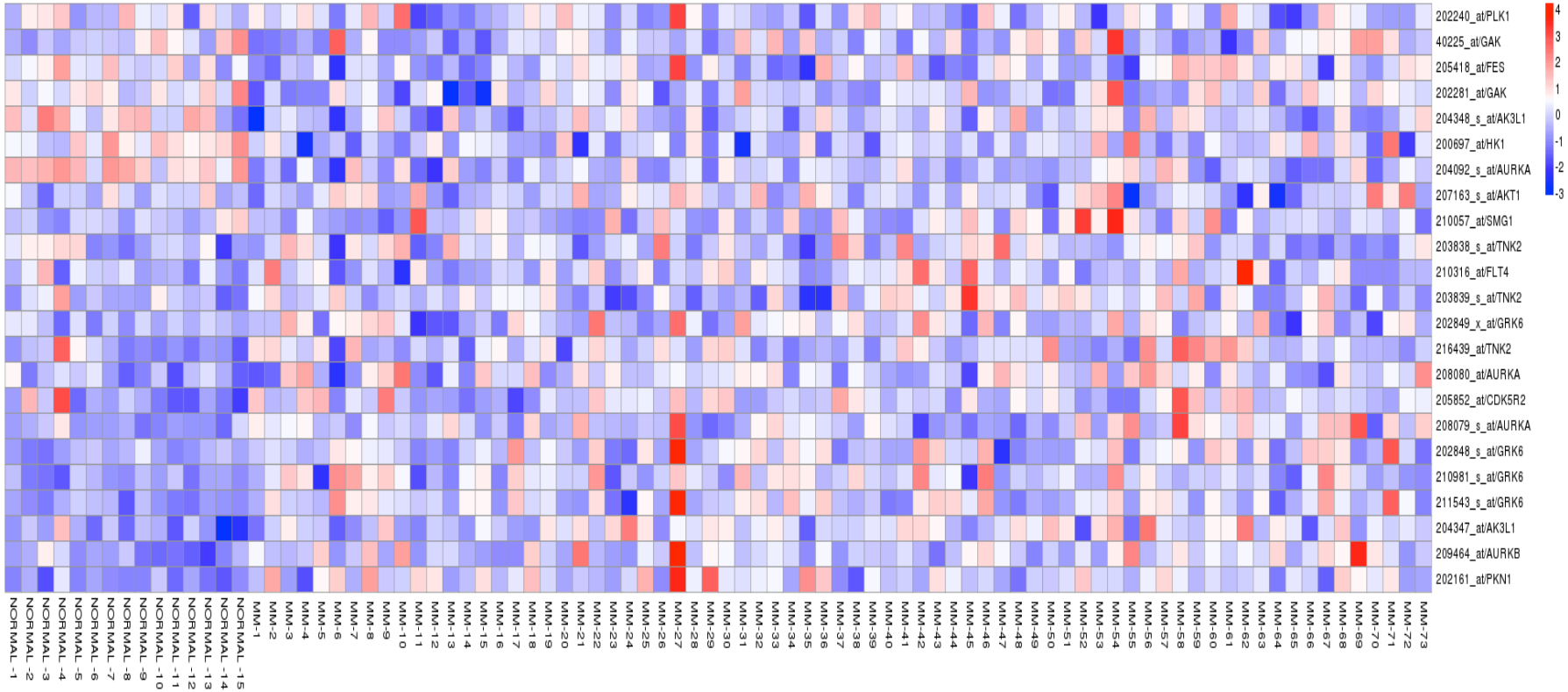
Bu literatür bilgisi ışığında aynı analiz, geliştirilen platform kullanarak tekrarlandığında elde edilen sonucun literatürle uyumlu olduğu görülmektedir. Tablo 4.3 ve Şekil 4.3 yapılan analize ait sonuçları özetlemekte ve ekspresyon değerlerinin z skoruna dönüştürülerek ısı haritası yardımı ile görselleştirilmesini içermektedir.

Tablo 4.2. GSE6477 Multiple Myeloma veri seti gen ekspresyonu farkı analizi sonuçları

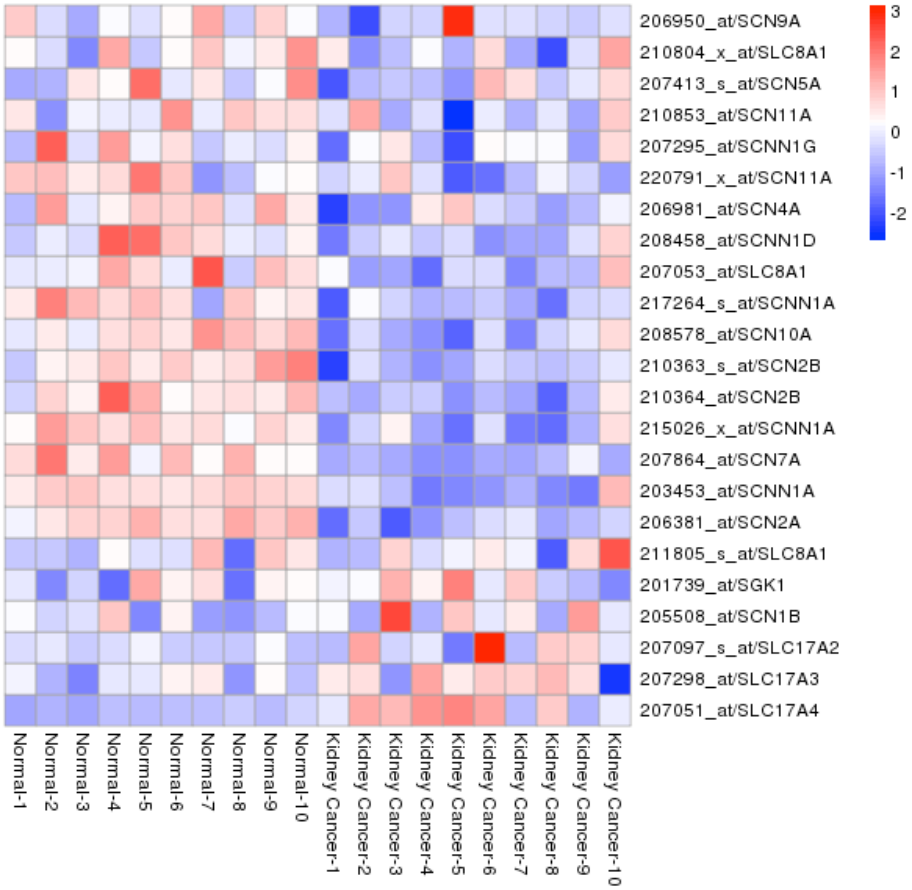
Probeset ID	Gen Sembolü	log FC	Ort. Eksp.	t	p Değeri	Düzeltilmiş p Değeri	B
204348_s_at	AK3L1	-0.184136	5.59685	-3.2882	0.00144488	0.00817369	-1.85367
204347_at	AK3L1	0.138557	4.17509	3.58026	0.000558692	0.00379322	-0.967117
207163_s_at	AKT1	0.0243588	7.16791	0.327749	0.743871	0.83796	-6.91159
208080_at	AURKA	0.102609	5.00822	2.2447	0.0272664	0.0810751	-4.50951
208079_s_at	AURKA	0.180398	4.97232	2.40354	0.0183119	0.0595606	-4.16083
204092_s_at	AURKA	-0.389457	5.90424	-5.76627	1.15E-07	2.98E-06	7.17222
209464_at	AURKB	0.308226	5.93409	3.65693	0.000431668	0.00308001	-0.724988
205852_at	CDK5R2	0.160384	4.92554	2.33528	0.0217778	0.0680677	-4.31326
205418_at	FES	-0.131347	6.46518	-2.12427	0.0364208	0.100703	-4.75967
210316_at	FLT4	0.050953	4.33573	0.985365	0.327116	0.488296	-6.48142
40225_at	GAK	-0.062294	7.98072	-0.77062	0.44297	0.597971	-6.66875
202281_at	GAK	-0.139042	7.40196	-2.22784	0.0284122	0.0836562	-4.54527
202848_s_at	GRK6	0.288554	8.61024	2.72688	0.00770082	0.0304791	-3.38735
202849_x_at	GRK6	0.114096	9.07631	1.66456	0.0995183	0.212183	-5.59806
211543_s_at	GRK6	0.44916	7.0541	3.57656	0.000565639	0.00383221	-0.978702
210981_s_at	GRK6	0.399348	7.79403	3.507	0.000712519	0.00458992	-1.19488
200697_at	HK1	-0.358182	7.54063	-3.53316	0.00065349	0.00428034	-1.11396
202161_at	PKN1	0.358517	6.85745	4.283	4.65E-05	0.000487121	1.38675
202240_at	PLK1	-0.051248	5.88554	-0.74187	0.460115	0.613532	-6.6904
210057_at	SMG1	0.0212705	3.45662	0.565389	0.573232	0.710775	-6.80547
203838_s_at	TNK2	0.0429811	6.35721	0.590297	0.556488	0.696931	-6.7911
216439_at	TNK2	0.125912	6.82642	1.85867	0.0663797	0.158146	-5.26697
203839_s_at	TNK2	0.0731892	7.95433	1.16333	0.247804	0.402641	-6.29225

Tablo 4.3. GSE6344 Böbrek kanseri veri seti için gen ekspresyonu farkı analiz sonuçları

Probeset ID	Gen Sembolü	log FC	Ort. Eksp.	t	p Değeri	Düzeltilmiş p Değeri	B
208578_at	SCN10A	-0.445053	6.72333	-4.78758	0.00010054	0.000542059	0.775989
210853_at	SCN11A	-0.0742793	4.30542	-1.34703	0.192407	0.264817	-6.36554
220791_x_at	SCN11A	-0.0962056	4.06044	-2.01255	0.0572372	0.0964325	-5.3431
205508_at	SCN1B	0.122619	6.07139	1.41709	0.171204	0.240568	-6.27306
206381_at	SCN2A	-0.818599	5.83592	-7.63732	1.78E-07	3.33E-06	7.17077
210363_s_at	SCN2B	-0.354509	5.20515	-4.82397	9.22E-05	0.000505402	0.862139
210364_at	SCN2B	-0.330401	5.55166	-4.87101	8.25E-05	0.000463607	0.973447
206981_at	SCN4A	-0.23346	6.20967	-2.90225	0.00855353	0.0197327	-3.58591
207413_s_at	SCN5A	-0.0853812	6.14843	-1.34688	0.192456	0.264853	-6.36574
207864_at	SCN7A	-0.266542	3.6281	-5.53525	1.75E-05	0.00013419	2.53263
206950_at	SCN9A	-0.0572934	3.52375	-0.892063	0.382524	0.466979	-6.86546
217264_s_at	SCNN1A	-0.448779	5.31053	-4.68293	0.000128848	0.00066079	0.528029
203453_at	SCNN1A	-2.85948	8.26508	-5.96511	6.55E-06	6.13E-05	3.52293
215026_x_at	SCNN1A	-0.529559	6.78003	-5.02814	5.70E-05	0.000346273	1.34463
208458_at	SCNN1D	-0.281913	5.57924	-2.95138	0.00765208	0.0179731	-3.47977
207295_at	SCNN1G	-0.127798	4.54573	-1.57931	0.129295	0.18992	-6.04437
201739_at	SGK1	0.217258	9.73361	1.01211	0.323071	0.406263	-6.75126
207097_s_at	SLC17A2	0.23806	5.99792	1.43967	0.164783	0.232883	-6.24243
207298_at	SLC17A3	0.622128	9.47357	1.4945	0.150003	0.215382	-6.16642
207051_at	SLC17A4	1.271	6.44794	4.6708	0.000132612	0.000674349	0.499273
207053_at	SLC8A1	-0.235471	4.11489	-3.07092	0.00582416	0.0143832	-3.21857
210804_x_at	SLC8A1	-0.0808608	4.41182	-1.25745	0.222473	0.298672	-6.47802



Şekil 4.2. GSE6477 Multiple Myeloma veri seti için 15 kinaz genine ait ısı haritası



Şekil 4.3. GSE6344 Böbrek kanseri veri setine ait "Sodyum iyon taşıma" yolağı gen ekspresyonu ısı haritası

4.3.2. miRNA Ekspresyonu Farkı Analizi

İki farklı fenotip arasında gen ekspresyonu analizine benzer şekilde çalışan miRNA ekspresyonu farklı analizi, bir önceki analiz ile aynı algoritma kullanılarak tasarlanmış bir modüldür. miRNA ekspresyon verilerini kullanarak analizleri gerçekleştiren modüle gen ekspresyonu modülünden farklı olarak miRNA isimlerini parametre olarak girmek gerekmektedir.

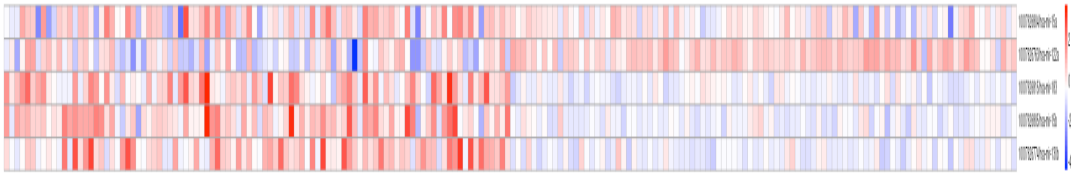
Çıktı olarak analizde kullanılan veri setinin özeti, ısı haritası, sonuç tablosu ve üretilen çıktıların indirilebilmesini sağlayan alanları üreten modül, kullanım açısından çalışma anında analiz gerçekleştiren diğer modüllere göre miRNA verisinin küçük boyutlu olmasından dolayı daha hızlı çalışmaktadır.

Literatürden bulunan karaciğer kanserine ilişkin bir analiz GSE22058 erişim numaralı veri setini kullanarak HSA-MIR-130B, HSA-MIR-183, HSA-MIR-122A, HSA-MIR-15B, HSA-MIR-15A gibi miRNA'ların hastalar ve kontrol grubu arasında farklı ekspresyon seviyelerine sahip olduğunu ortaya koymuştur[161, 162]. Bu analiz aynı veri setini depolayan platform üzerinde tekrarlandığında bahsedilen

miRNA'ların hasta ve kontrol grubunda farklı ekspresyona sahip oldukları Tablo 4.4 ve Şekil 4.4'de görülmektedir.

Tablo 4.4. GSE22058 Karaciğer tümörü veri seti miRNA ekspresyon analizi sonuçları

Probeset ID	miRNA Sembol	log FC	t	p Değeri	Düzeltilmiş p Değeri	B
10007626774	HSA-MIR-130B	-0.387235	-10.309	3.75E-20	4.13E-19	34.57588449
10007626805	HSA-MIR-15B	-0.27273	-10.0751	1.80E-19	1.72E-18	33.01793592
10007626815	HSA-MIR-183	-0.427575	-10.0378	2.31E-19	2.12E-18	32.76981652
10007626763	HSA-MIR-122A	0.286372	7.72111	5.97E-13	2.63E-12	18.13551354
10007626804	HSA-MIR-15A	-0.0659583	-2.58399	0.0105011	0.0171128	-4.66088921



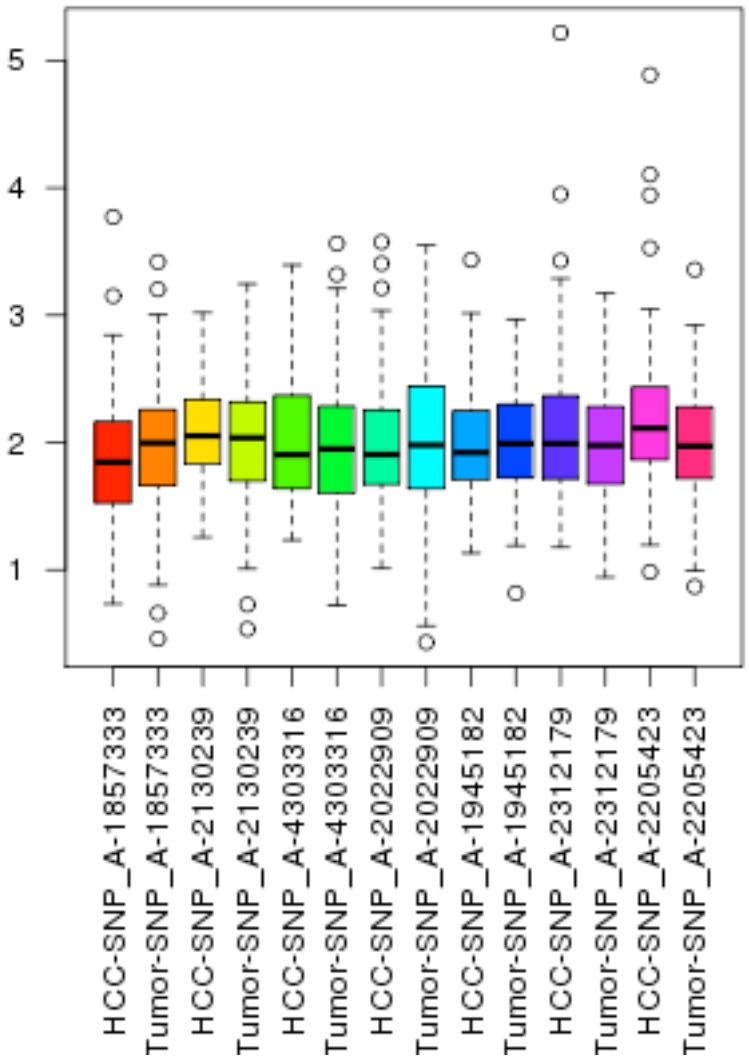
Şekil 4.4. GSE22058 Karaciğer tümörü veri seti miRNA analizi ısı haritası

4.3.3. Gen Kopya Sayısı Analizi

Gen kopya sayısı değişimlerinin sorgulanabildiği modül için önişlemesi yapılmış ve tamamı Affymetrix platformundan alınmış SNP verileri dChip-SNP yazılımı ile matris formatına dönüştürülerek sorgulamaya hazır hale getirilmiştir. Analizler platformdan bağımsız bir ortamda gerçekleştirilmekte olup yönetici modülü ile sonuçlar platforma aktarılmaktadır. Platform üzerinde önişlemesi ve matris formatına dönüştürmesi gerçekleştirilmeyen tek analiz kopya sayısı analizidir.

Kopya sayısı sorgulaması için kullanıcılar ilgilendikleri veri setini belirlemek ve araştırdıkları gen sembollerini sisteme girmek zorundadırlar. Platform üzerinde kurulan eşleştirme algoritması kullanıcıların girdikleri gen sembollerine göre genlerin lokasyonlarını bulmakta ve o lokasyon içerisinde yer alan SNP bilgilerini kayıtlardan getirmektedir. Birden fazla fenotipe ait verinin olduğu veri setleri için fenotiplere göre örnekler gruplanmakta ve her gen için ortalama kopya sayısı hesaplanmaktadır. Hesaplanan ortalama değerler bir tablo halinde araştırmacılara sunulmakta ve kutu çizgi grafiği şeklinde görselleştirilmektedir. Ayrıca sonuç ekranından kullanıcılar grafiği ve analizde kullanılan ön işlemesi tamamlanmış veriyi indirebilmektedirler.

Aşağıda Şekil 4.5 ve Tablo 4.5'de GSE9845 veri setine ait hepatoselüler karsinom ve sirozlu fenotiplere ait dokulardan elde edilmiş veriler ile oluşturulan kutu çizgi grafiği ve genlerin kopya sayısındaki değişimler görülebilmektedir. Chiang ve arkadaşları yayınladıkları bir makalede bu veri setini kullanarak yaptıkları analizlerde VEGFA ve CCND1 genlerine ait değişimlerin olduğundan bahsetmişlerdir[163]. Bu genlere ait platformun oluşturduğu çıktı örneği aşağıda gösterilmiştir.



Şekil 4.5. GSE9845 veri setinden elde edilen VEGFA, CCND1, CTNNB1 genleri ile ilişkili SNP'lere ait kopya sayıları

Tablo 4.5. GSE9845 veri setinden elde edilen VEGFA, CCND1, CTNNB1 genleri ile ilişkili SNP'lere ait kopya sayıları

Fenotip	Gen Sembolü	SNP ID	dbSNP ID	Ortalama Kopya Sayısı
Karaciğer Tümörü	CTNNB1	SNP_A-1857333	rs9824212	1.85607767
Siroz	CTNNB1	SNP_A-1857333	rs9824212	2.000010638
Karaciğer Tümörü	CTNNB1	SNP_A-2130239	---	2.06631068
Siroz	CTNNB1	SNP_A-2130239	---	2.000021277
Karaciğer Tümörü	CTNNB1	SNP_A-4303316	rs2293301	2.049165049
Siroz	CTNNB1	SNP_A-4303316	rs2293301	1.999989362
Karaciğer Tümörü	CTNNB1	SNP_A-2022909	rs2293300	1.994417476
Siroz	CTNNB1	SNP_A-2022909	rs2293300	2
Karaciğer Tümörü	CTNNB1	SNP_A-1945182	rs9838277	2.009067961
Siroz	CTNNB1	SNP_A-1945182	rs9838277	2.000021277
Karaciğer Tümörü	VEGFA	SNP_A-2312179	rs752662	2.101330097
Siroz	VEGFA	SNP_A-2312179	rs752662	2
Karaciğer Tümörü	VEGFA	SNP_A-2205423	rs16896629	2.184145631
Siroz	VEGFA	SNP_A-2205423	rs16896629	2.000010638

4.3.4. ARACNE Ağ Analizi

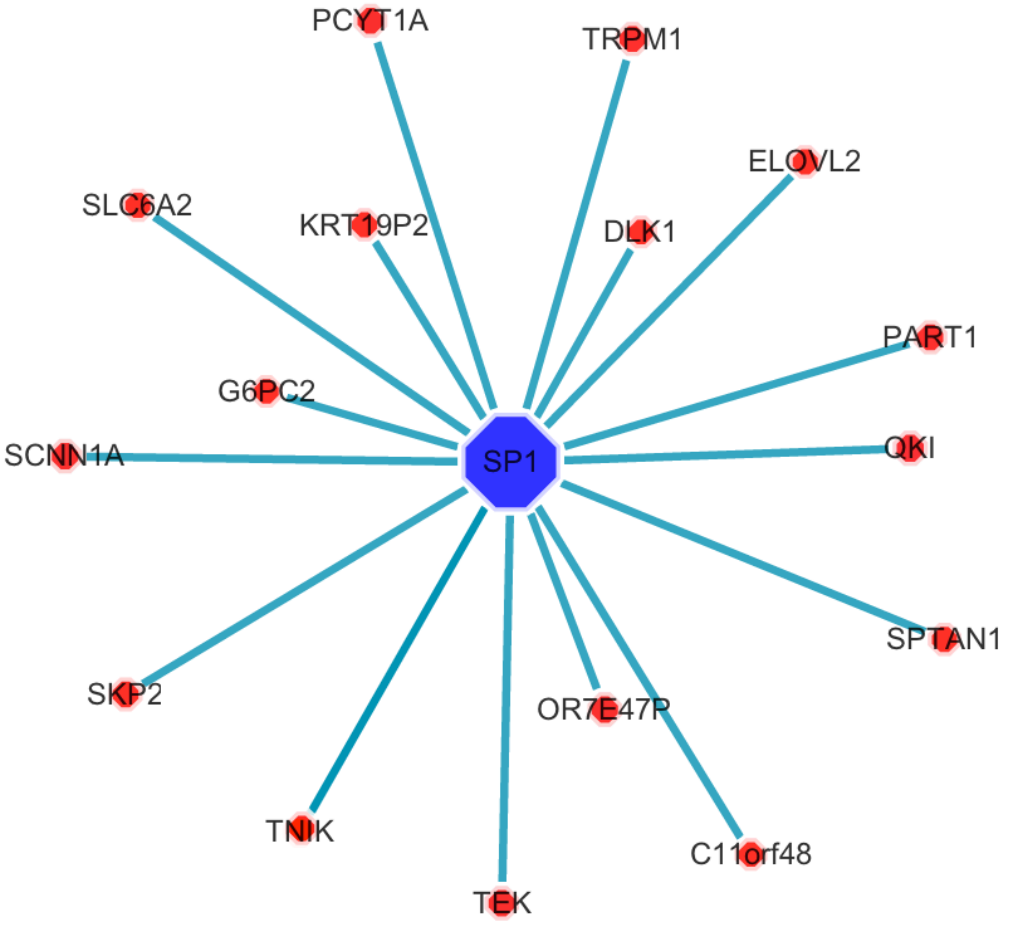
Aracne algoritması gen ekspresyonu verisi üzerinden transkripsiyon faktörlerini kullanarak tersine mühendislik yöntemi ile ilişkili olan genleri belirlemek için geliştirilmiş bir araçtır. Kullanıcılar öncelikle incelemek istedikleri veri setini seçerek ilgilendikleri gene ait ağı oluşturmak için gen sembolünü platforma parametre olarak bildirmelidirler.

Kullanıcının bildirdiği gen sembolü ve seçtiği veri setine göre sistem seviyesinde çalıştırılan analiz, kullanıcılara analiz sonucunda oluşan ağı görselleştirerek geri getirmektedir.

Bu fonksiyonun oluşturulması aşamasında analizlerde kullanılmak üzere araştırmacılar tarafından 2000'den fazla transkripsiyon faktörünün listesi çıkarılmıştır. Mevcut versiyonda analizler bu TF'ler üzerinden yapılmaktadır.

Sistem, analiz için öncelikle gen ve ilişkili olduğu diğer genleri belirlemekte ve algoritmanın ağırlıklandırma yöntemine göre onları ilişkilendirdikten sonra Flash tabanlı görselleştirme aracı ile ekrana çalışma zamanında ağı oluşturmaktadır. Araştırmacılar isterlerse yakınlaştırma ya da uzaklaştırma gibi fonksiyonlarla oluşturulan ağa ait elemanları inceleyebilmektedir.

Aşağıda şekil 4.6'da Multiple Myelomada etkili olduğu bilinen SP1[164] transkripsiyon faktörü için oluşturulan ağ şeması görülebilir.



Şekil 4.6. SP1 transkripsiyon faktörü için Aracne algoritması ile GSE6477 veri setinden üretilmiş ağ görüntüsü

4.3.5. WGCNA Ağ Analizi

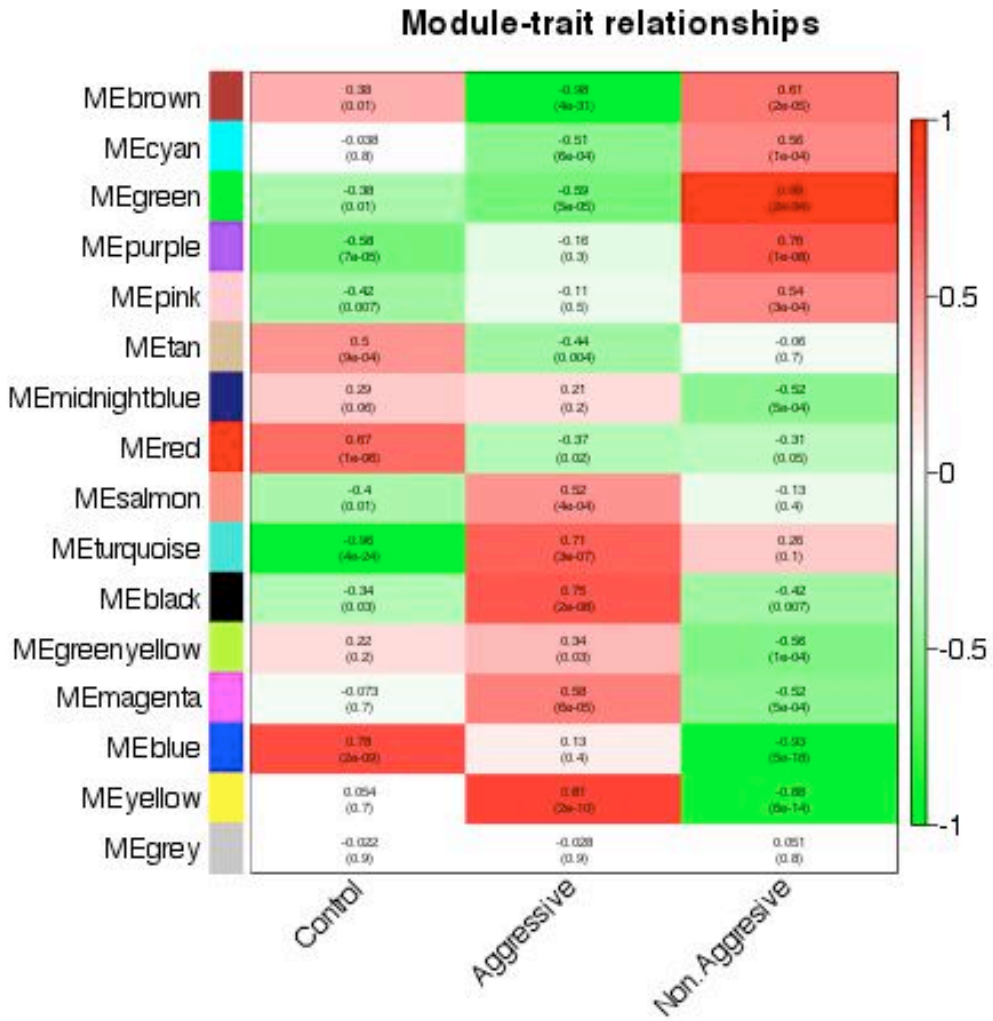
Bu algoritma ağırlıklandırılmış korelasyon temelli analiz gerçekleştirerek gen ağlarını oluşturmakta ve oluşturduğu ağları fenotiplere göre analiz ederek farklı gruplar için anlamlılık seviyelerini bulmaktadır.

Daha önce bahsedilen analizlerden farklı olarak sadece veri seti seçimi ile çalışan ve platformda depolanan analiz sonuçlarını sorgulayan sistemde analizler veri setine ait fenotip bilgileri ve ekspresyon verileri kullanılarak sistem yöneticileri onayı ile platform tarafından daha önce gerçekleştirilmektedir.

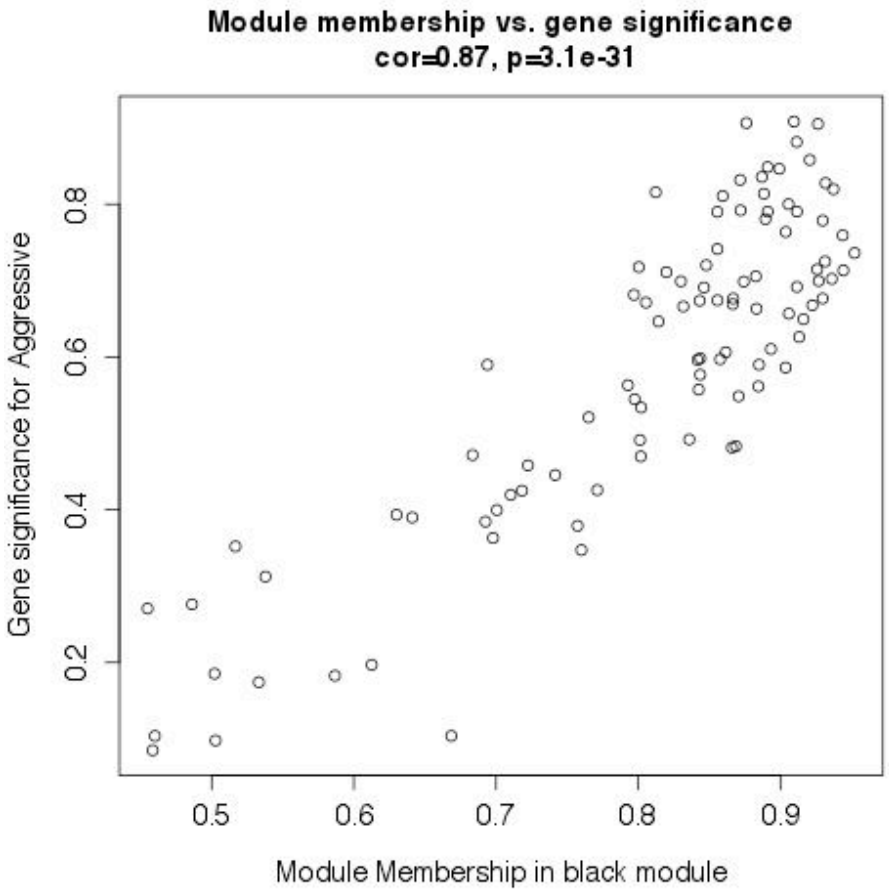
Analiz sonucu olarak kullanılan parametreleri, genlerin kümelenmesini, oluşturulan ağlar ile fenotipler arasındaki ilişkiyi, ağlar ile fenotipler arasındaki

korelasyon grafiklerini ve son olarak ağların kullanıcılar tarafından yerel bilgisayarlarında görselleştirebilmesine olanak vermek için VisANT çıktılarını platform kullanıcılar ile paylaşmaktadır.

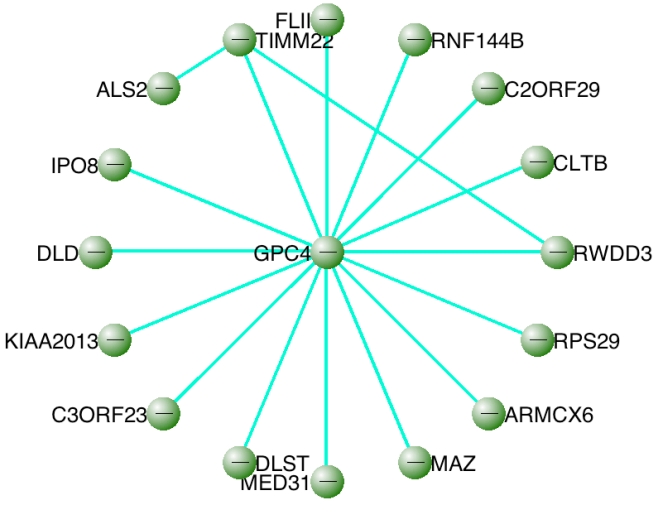
Aşağıda Şekil 4.7'de meme kanseri veri setine uygulanmış olan WGCNA algoritmasına göre 3 farklı grup için ağların anlamlılık düzeylerini ve altında Şekil 4.8'de "Siyah modül" için hastalığın agresif olduğu gruba ait modül üyeliği ve gen anlamlılığı saçılım grafiği görülebilir. Son olarak ise Şekil 4.9'da Siyah modül için aralarında ilişki düzeyi yüksek olan genlerin VisANT ile görselleştirilmiş şekli yer almaktadır.



Şekil 4.7. GSE28976 meme kanseri veri setinden WGCNA için elde edilmiş gen ağları ve gruplar ile ilişkileri



Şekil 4.8. GSE28976 meme kanseri veri setinden WGCNA algoritması ile elde edilmiş "Siyah modül" için ağ üyesi genlerin anlamlılıklarını gösteren saçılım grafiği



Şekil 4.9. "Siyah modül" üyesi genlerden ilişki düzeyleri yüksek olanların ağ grafiği

4.3.6. Gen Seti Zenginleştirme Analizi

Gen seti zenginleştirme analizleri eldeki genomik bilginin biyolojik olarak anlamlandırılmasındaki önemli adımlardan birisidir. Bu nedenle kanser araştırmaları gibi kompleksliğin yüksek olduğu çalışmalarda genomik bilginin tek başına yorumlanması veya anlamlandırılması yerine entegre analiz yöntemleri ile biyolojik olarak da anlamlandırılabilmesi önemlidir.

Platform gen seti zenginleştirme modülünde Broad Enstitüsü tarafından geliştirilen GSEA aracını kullanmaktadır. Analiz katmanı sistem seviyesinde platform yöneticileri onayı ile veri setlerini bu araca taşımakta ve analizleri gerçekleştirmektedir.

Analizleri tamamlanan veri setleri yine yönetici onayı ile sonuç arşivine aktarılmakta ve çalışma zamanında kullanıcılara veri setinden elde edilmiş analizlere yönelik sonuçlar getirilmektedir.

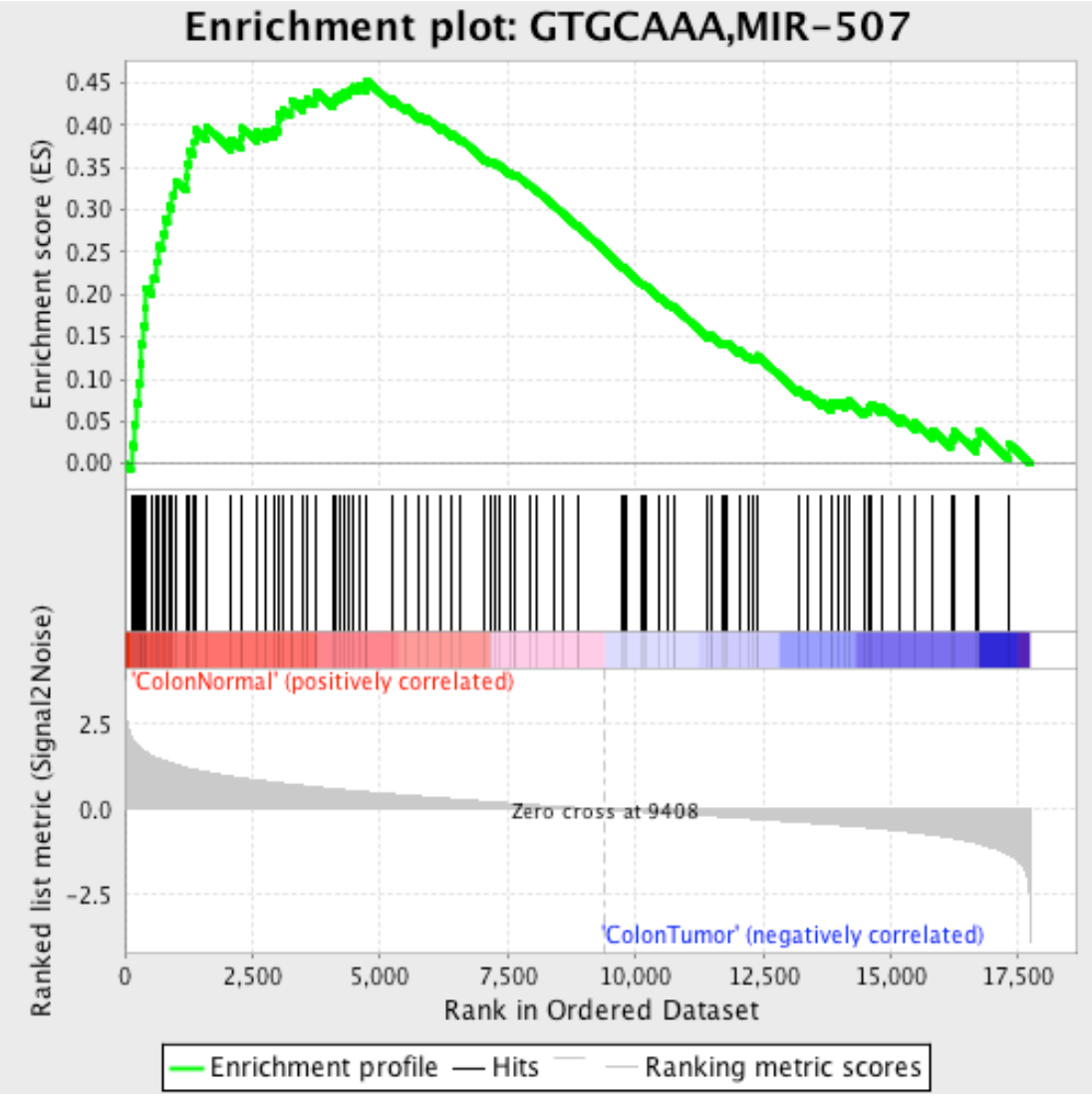
Platform analiz aşamasında GSEA aracının sunduğu bütün yolak veritabanlarını kullanmak yerine kanser ile diğerlerine göre daha yüksek ilişki içinde olan ve sonuçlarının önemli olduğu düşünülen 6 farklı hedef dosyasını kullanmaktadır. Bunlar aşağıdaki gibidir.

- Gen ontolojisi
- miRNA hedefleri
- Transkripsiyon faktör hedefleri
- BioCarta gen setleri
- Kanonik yolaklar
- KEGG yolakları

Kullanıcılar WGCNA analizinde olduğu gibi sadece veri setini seçerek sonuçlara ulaşabilmektedirler. Platform analizlerin çıktısı olarak kullanıcılara yukarıda belirtilen 6 başlık için sonuçları GSEA aracının ürettiği formatta sunmaktadır. Bu arayüz üzerinden fenotiplere göre farklı anlamlılık düzeylerinde gruplanmış yolakların listesine ve analiz sonucunda üretilen grafiklere erişilebileceği gibi yolaklarla ilgili detay bilgilere de Broad Enstitüsüne ait sayfalar üzerinden erişim sağlanmıştır.

Sonuç ekranında aynı zamanda araştırmacıların gerek duymaları halinde inceleyebilecekleri analiz parametreleri ve analiz setinde kullanılan genlere ait işaretçileri görebilecekleri seçenekler yer almaktadır.

Geliştirilen modülden elde edilen analiz sonuçları literatür ile karşılaştırıldığında, modülün literatür ile uyumlu bulgular verdiği ortaya koyulmuştur. Örneğin kolon kanseri için TCGA üzerinden elde edilen veri seti ile miRNA hedeflerinin gen seti zenginleştirme analizi miR-507, miR-99, miR-7 gibi miRNA'ların anlamlı olduğunu göstermektedir. Literatüre bakıldığında bu miRNA'lar çeşitli genleri hedef alarak kolon kanserinin gelişmesinde etkili olmaktadır[165, 166]. Aşağıda Şekil 4.10'da bu veri seti için platform tarafından araştırmacılara sağlanan miR-507'ye ait zenginleştirme grafiği örnek olarak verilmiştir.



Şekil 4.10. TCGA Kolon kanseri veri seti (eş örnekler) ile GSEA modülü kullanılarak elde edilen miR-507 için zenginleştirme grafiği

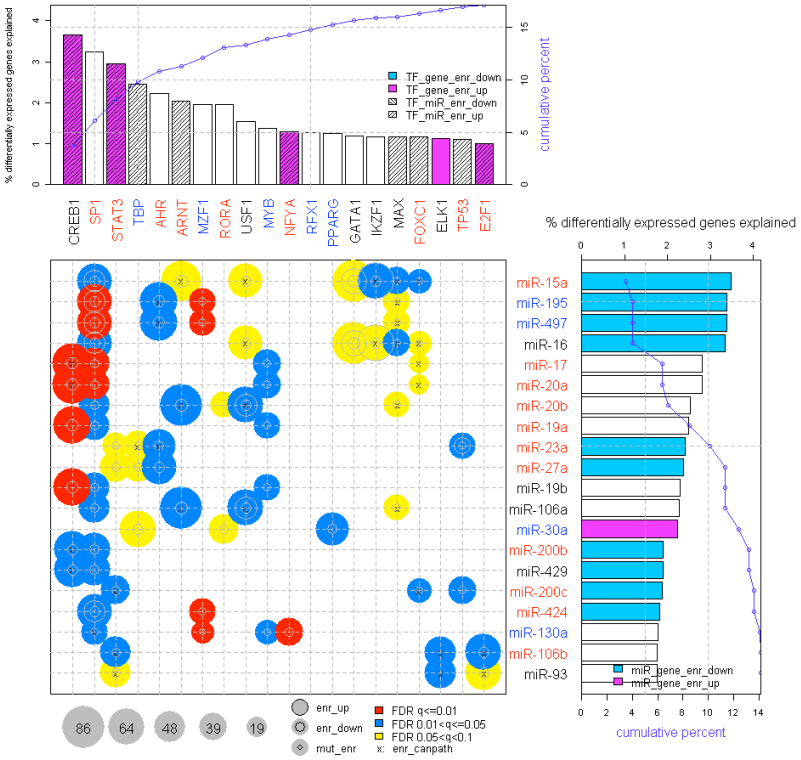
4.3.7. GemiNi (Gen Ekspresyonu ve miRNA Ekspresyonu Entegre Analizi)

GemiNi bu tez projesinde yer alan araştırmacılar ve iş birliği yapılan araştırma ekibi tarafından ortak bir çalışma sonucunda geliştirilmiş bir yöntemdir. Gen ekspresyonu ve miRNA ekspresyonu verisini entegre ederek analiz etmek için geliştirilmiş bu metot transkripsiyon faktörleri, miRNA'lar ve onların ortak hedefi olan genleri kullanarak ileri beslemeli ağlar oluşturmaktadır[157].

Gen seti zenginleştirme analizinde olduğu gibi platform yöneticilerinin onayı ile çalışan bu fonksiyon için de sistemin oluşturduğu ön analizler depolama alanında tutulmakta ve kullanıcılara veri seti seçimine bağlı olarak sonuçlar aktarılmaktadır.

Analiz sonucu olarak araştırmacılarla TF, miRNA ve ortak hedeflerini gösteren grafikleri, oluşturulan ileri beslemeli ağa ilişkin ilişki grafiklerini ve sonuçları liste seklinde döndüren platformdan veri seti seçilerek analiz sonuçlarına ulaşılabilmektedir.

Bu analiz modülü ile GSE18805 erişim numaralı akciğer kanserine ilişkin veri seti analiz edildiğinde Şekil 4.11'deki gibi TF, miRNa ve hedef genlerin olduğu grafik analiz sonucu olarak oluşturulmuştur.



Şekil 4.11. GSE18805 akciğer kanseri veri setinin GemiNi metodu ile analiz edilmesinden elde edilen ilk 20 TF, miRNA ve hedef genleri

Bu veri seti üzerinde yapılan analize göre yöntem ilk 3 transkripsiyon faktörü CREB1[167], SP1[168], STAT3[168], ilk 3 miRNA'yı da miR-15a[169], miR-195 [170]ve miR-497[170] şeklinde belirleyen algoritmanın sonuçları literatür ile uyumludur.

4.3.8. Gen Kopya Sayısı ve Gen Ekspresyonu Entegre Analizi

Platform gen kopya sayısı ve gen ekspresyonu emtegrasyonunda literatürde yer alan yöntemleri baz alarak R istatistiksel programlama dili üzerinden analizleri gerçekleştirmektedir. Kanserde kopya sayısı değişimlerinin (kopya sayısındaki artış veya gen kopyalarının silinmesi) sıkça rastlanan bir durum olduğu daha önceki bölümlerde de gösterildiği gibi bilinmektedir. Bu kopya sayısı değişimlerinin ekspresyon seviyesi üzerine etki etmesi de araştırmacılar açısından beklenen bir durumdur. Bu nedenle bu etkileşimi ortaya koymak için geliştirilen modül, sorgulama için gen sembolü ve veri setini girdi bilgisi olarak kullanmakta ve araştırmacıların ilgilendikleri genlere ilişkin analiz sonuçlarını ve veri setinin kapsadığı bütün genlere ilişkin sonuçları iki farklı tablo halinde araştırmacılara sunmaktadır.

Bu modül analizlerin gerçekleştirilmesi için korelasyon tabanlı analizler uygulayan DR-Integrator[158] aracını kullanmaktadır. Analiz modülüne entegre edilmiş R istatistiksel programla dili yardımı ile gerekli paketler kullanılarak analizler gerçekleştirilmektedir.

Aşağıda bu modül yardımıyla GSE26863 MM veri seti için elde edilmiş ilk 10 gene ilişkin korelasyon değerleri ve FDR değerleri verilmiştir. Bu ilk 10 gen içerisinde yer alan BIRC2 ve FAF1 gibi genler onko gen olarak tanımlanmış olup kanserin oluşmasında ve gelişmesinde etkileri bilinmektedir. Aynı zamanda başka bir çalışmada bu genlerin homozigotluğun kaybolması ve sağ kalım üzerine etkileri olduğu da gösterilmiştir[171].

Tablo 4.6. GSE26863 MM veri setinden DR-Integrator algoritması ile elde edilen ilk 10 gen

Gene Sembolü	Sıra	Korelasyon	FDR
BIRC2	1	0.8666	0
PSMD4	2	0.7784	0
SDHC	3	0.7614	0
UBAP2L	4	0.75	0
MRPL9	5	0.7386	0
JTB	6	0.736	0
FAF1	7	0.7358	0
GPR89A	8	0.7352	0
WHSC1L1	9	0.7351	0
GSTT1	10	0.7346	0

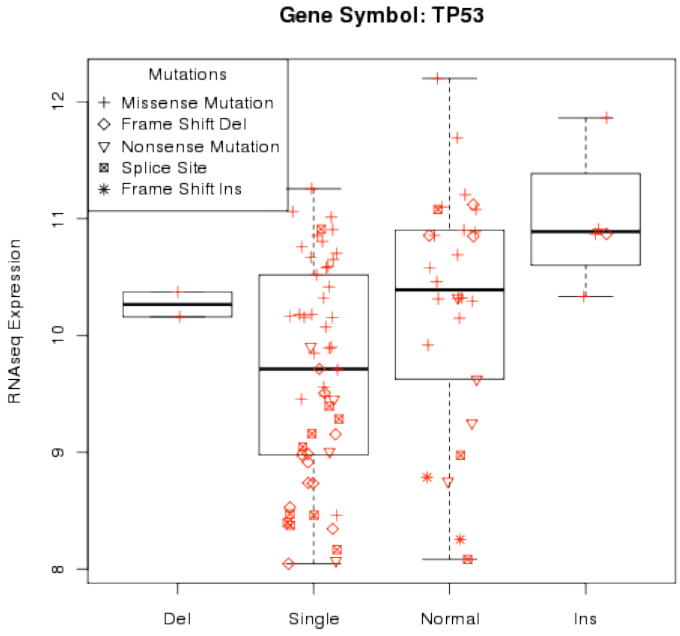
4.3.9. Mutasyon Analizi

Her ne kadar diğer analiz yöntemleri ile genom üzerindeki farklı elementlerin ne şekilde değiştiğini ortaya koyuyor olsak da bu değişikliğin asıl sebebini bilmek onarılmasında ya da olası etkilerinin değiştirilmesinde, sebeplerin anlaşılmasında büyük önem taşımaktadır. Mutasyon verisinin elde edilmesi işlemi için de farklı analiz yöntemleri kullanılabilir.

Bu proje kapsamında mutasyon bilgileri kanser için oluşturulmuş en büyük kaynaklarından bir tanesi olan TCGA üzerindeki verilerden TCGA'in takip ettiği analiz yöntemleri takip edilerek matrisler halinde alınmış ve sorgulanabilir hale dönüştürülmüştür.

Aynı zamanda mutasyon bilgisine ulaşılan örnekler için yine TCGA üzerinde depolanan veriler incelenerek ilgili RNAseq (gelecek nesil dizileme platformu ile elde edilen gen ekspresyonu verisi) ve kopya sayısı verileri de elde edilmiş, kopya sayısına göre ekspresyon değişimleri ve ilişkili mutasyonların aynı kutu çizgi grafikleri üzerinde gösterilmesi sağlanmıştır.

Gen sembollerinin ve veri setinin seçiminden sonra analize başlayan modül detaylı olarak örneklerden elde edilen mutasyon bilgilerini ve grafikleri üreterek araştırmacılara sağlamaktadır. Şekil 4.12 tümör baskılayıcı gen olarak bilinen ve akciğer kanserindeki etkileri daha önce ortaya konulmuş TP53[172] geni için TCGA'den elde edilmiş akciğer kanseri veri setine ilişkin analiz sonuçlarının bir kısmını göstermektedir.



Şekil 4.12. TCGA akciğer kanseri veri setinden elde edilen TP53 geninin kopya sayısına göre ekspresyon değişimini ve mutasyonları gösteren kutu çizgi grafiği

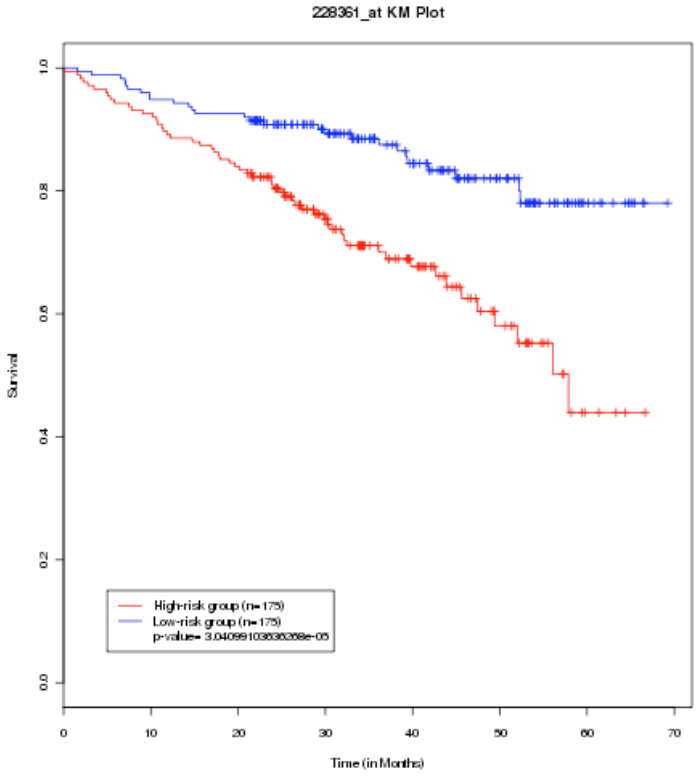
4.3.10. Sağ kalım analizleri

Sağ kalım analizi çeşitli yöntemler ile bulunan gen sembollerinin kanser türleri üzerindeki sağ kalım etkisini ortaya koymak amacıyla platforma entegre edilmiş bir modüldür.

Bu modül yardımıyla araştırmacılar araştırdıkları gen veya genlerin araştırma yaptıkları kanser türünde sağ kalımı ne şekilde etkilediğini inceleyebileceklerdir. Modül "Cox regresyon" ve "Log-Rank" sağ kalım testlerini entegre R istatistiksel programlama dili ortamından çağırarak ön işlemesi yapılmış ve sağ kalım bilgileri platform yöneticileri tarafından onaylanmış veri setlerini kullanarak analizi gerçekleştirmektedir. K-M grafiklerinin oluşturulmasında sistem eldeki örnekleri ekspresyon değerlerinin ortancasına göre 2 gruba ayırarak karşılaştırmayı yapmaktadır.

Araştırmacılara tek gen üzerinden analiz gerçekleştirdikleri durumda K-M grafiklerini de çıktı olarak sağlayan modül, çoklu gen sembolleri ile analiz başlatıldığında ise analiz sonuçlarını tablo şeklinde sağlamaktadır.

Aşağıda çeşitli kanser türleri üzerinde etkisi olduğu bilinen E2F2 geni için GSE2658 MM veri seti üzerinde yapılan sağ kalım analizi için platformun ürettiği KM grafiklerinden bir tanesi ve sonuç tablosu yer almaktadır.



Şekil 4.13. E2F2 geni için GSE2658 MM veri seti üzerinde yapılan sağ kalım analizi KM grafiği

Tablo 4.7. E2F2 geni için GSE2658 MM veri setinden elde edilen sağ kalım analizi sonuçları

Probeset ID	Cox p değeri	Cox düzeltilmiş p değeri	Logrank p değeri	Logrank düzeltilmiş p değeri
207042_at	0.017782558	0.364000215	0.269536811	0.826227715
228361_at	2.25E-06	0.001539186	3.04E-05	0.152602806
235582_at	0.008302797	0.266851014	0.22329531	0.802804024

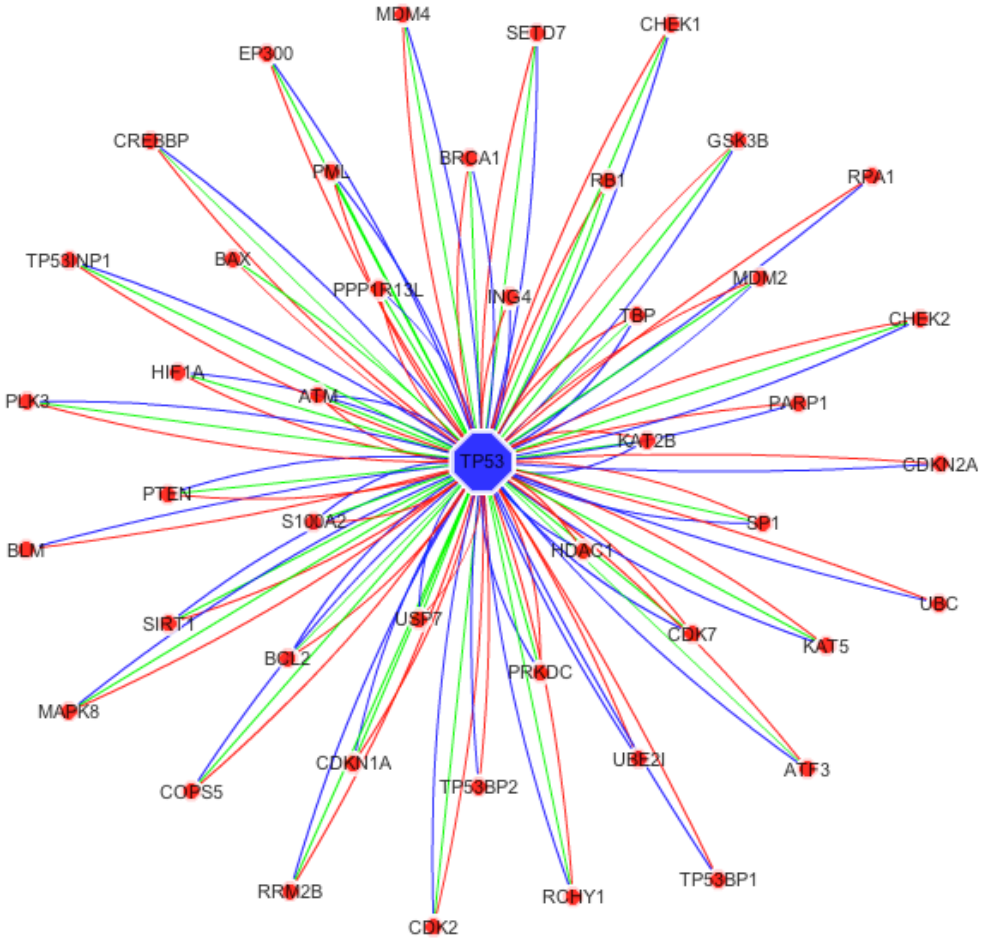
4.3.11. Protein-Protein Etkileşimleri

Protein-protein etkileşimlerini bilmek biyolojik olarak büyük önem taşımaktadır. Hücre içerisinde kendilerine genler tarafından kodlanan görevleri yerine getiren bu yapıların birbirleri üzerinde ve genler üzerindeki etkileri bilinmektedir. Kendi aralarındaki etkileşimi anlamak için günümüzde çeşitli araştırmalar kapsamında kurulmuş pek çok veritabanı yer almaktadır.

Geliştirilen platformun kanser araştırmalarında kullanılacak kapsamlı bir kaynak oluşturması amacıyla protein-protein etkileşimleri de mevcut veritabanlarından bilişim kaynakları kullanılarak insan genomu için çıkarılmış ve sorgulanıp görselleştirilebilir şekilde platformun veritabanına eklenmiştir. Bu işlem için STRING[55, 56] projesine ait açık erişime sunulmuş veri kaynakları ve oluşturulan analiz platformu kullanılmıştır.

Mevcut versiyon 4 milyon satıra yakın protein etkileşimi, bağlanma türü, kaynağı ve gen bilgisini kullanarak gen sembolleri üzerinden araştırmacıların bu bilgilere ulaşabilmelerini sağlamaktadır.

STRING projesi farklı kaynaklardan elde ettiği bulgulara göre proteinler arasındaki ilişki ağırlıklandırılabildiğinden sorgulama anında araştırmacılar 1 ile 999 arasındaki bir skor ile istedikleri şekilde esnek veya detaylı bir görselleştirmeye sahip olabilmektedirler. Aşağıda şekil 4.14'de TP53 ile ilişkileri yüksek derecede kanıtlanmış diğer proteinler gösterilmektedir.



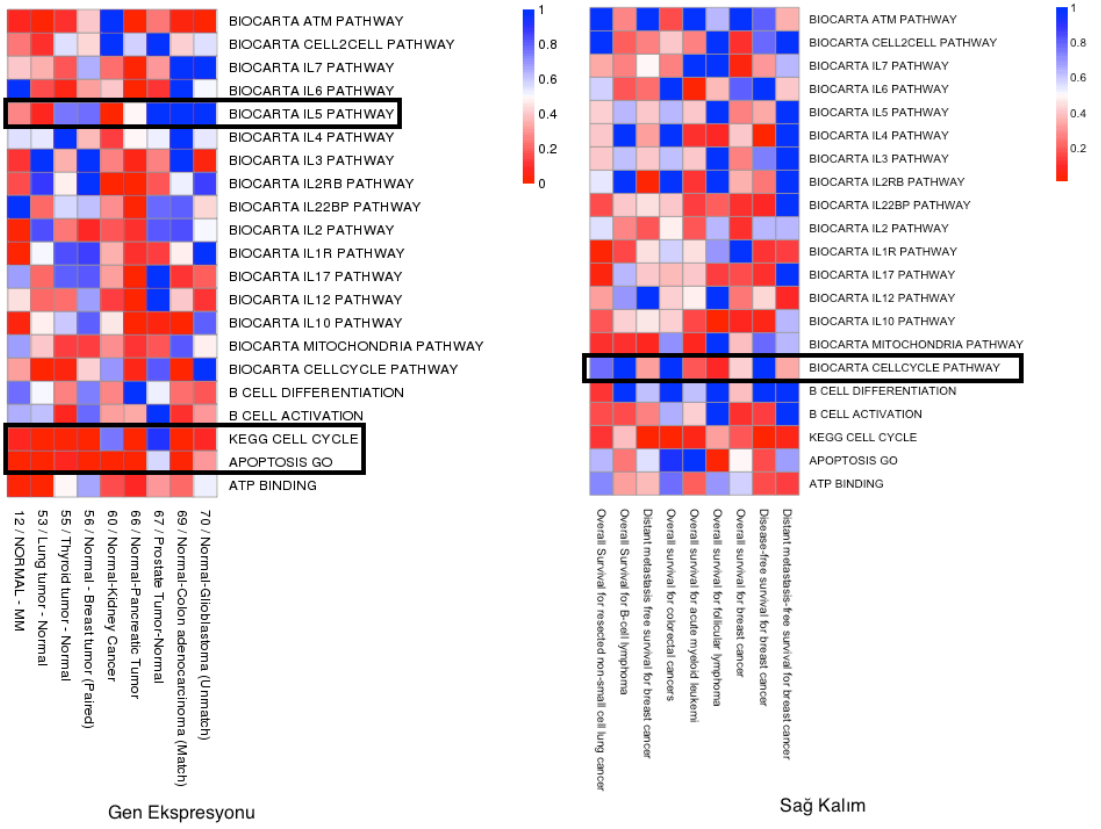
Şekil 4.14. TP53 ile ilişkili proteinler

4.3.12. Meta Analiz Modülleri ve Kanserin Evriminin Modellenmesi

Geliştirilen platformun "Çalışmaları Karşılaştır" fonksiyonu MSigDB tarafından sağlanan yolakların farklı veri setlerinde analiz edilerek sonuçlarının karşılaştırılmasına olanak sağlamak amacıyla geliştirilmiştir. Gen ekspresyonu farkı analizi ve sağ kalım analizi için ayrı ayrı kullanılabilen fonksiyon sayesinde kullanıcı tarafından belirlenen yolaklar farklı veri setlerinde analiz edilerek veri setlerinin birbirleri ile karşılaştırılmalarına imkân sağlamaktadır.

Bu modüller yardımıyla araştırmacıların seçtikleri veri setleri ve yolaklara göre analizler çalışma zamanında gerçekleştirilip analiz sonuçları kullanıcılara ısı haritaları ve sonuç tabloları şeklinde ulaştırılmaktadır.

Bu modül ile yapılan ve 9 farklı kanser tipi için 21 farklı yolağın karşılaştırıldığı analize ilişkin ısı haritası şekil 4.15'de verilmiştir.

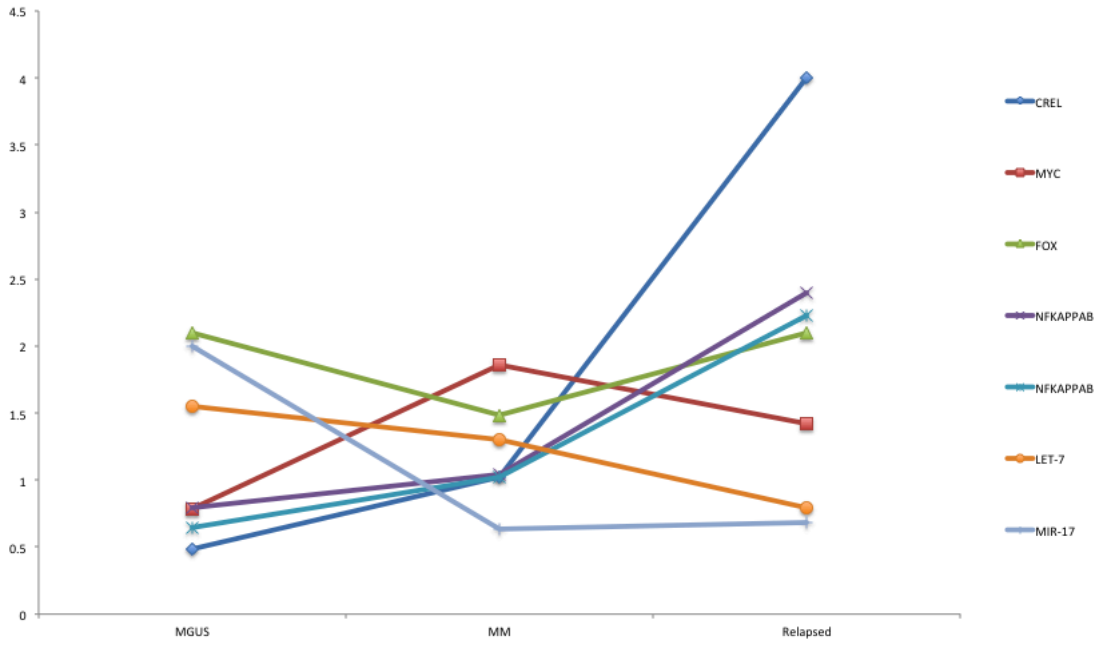


Şekil 4.15. 9 farklı kanser türü ve 21 farklı yolak için platformun oluşturduğu analiz sonuçları

Gen ekspresyonu veri setlerinden elde edilen analiz sonuçlarına göre hücrenin yaşamsal döngüsü ile ilgili olan "hücre siklusu" ve "apoptosis" gibi yolakların neredeyse bütün kanser türlerine ait karşılaştırmalarda etkilendiği görülürken, "IL5" gibi yolakların kanser tipine göre değişik anlamlılıklar gösteriyor olması kanserli hücrelerin türe göre farklı davranışlar sergilediğini ortaya koymakta ve araştırmacılara farklı kanser türlerindeki değişimi anlamada yardımcı olmaktadır. Benzer şekilde örüntülerin sağ kalım analizinden elde edilen bulgularda da olduğu görülmektedir. "hücre siklusu" yolağının meme kanserinde etkisinin[173] diğerlerine göre daha belirgin olduğu örnek bunun bir örneğidir.

Farklı çalışmalardan elde edilen meta analiz bulguları bize sadece çalışmaları gen ekspresyonu değişimi ya da sağ kalım açısından karşılaştırma avantajı sağlamamış, bununla beraber kanserin farklı türleri için ve aynı kanser türünde farklı alt gruplar için değişimlerin ya da normal hücre yapısından kanserli hücre yapısına doğru değişimin ne şekilde modelleneceğine dair fikirler sağlamıştır.

Bu analizlerin bütün türler için yapılması şimdilik veri imkânları ile kısıtlı olsa da iş birliği çerçevesinde MM ile ilgili çeşitli analizler gerçekleştirilmiştir. Aşağıda Şekil 4.16'da farklı yolakların normal örneklerden tekrarlayan MM gelişimine doğru nasıl değiştiği modellenmiştir.



Şekil 4.16. Normal-MGUS-MM-Tekrarlayan MM geçişlerinde etkilenen yolaklar.

Bu analizler üye veya hedef genlerdeki anlamlı ekspresyon seviyesi değişimine bağlı olarak transkripsiyon faktörleri, miRNA hedefleri, metabolik ve sinyal yolaklarında yaşanan değişimi ortaya koymuştur. MYC, FOXO ve NF-Kappa B'nin rolü MGUS'dan MM'e geçişte daha önce Davies ve arkadaşları tarafından[174], miR-17 Chen ve arkadaşları[175] tarafından, let-7 ise Lionetti ve arkadaşları[176] tarafından gösterilmiştir. Benzer şekilde MM'in tekrar ettiği hastalarda miR484 ve CREL'de yaşanan anlamlı değişimler ise Tian ve arkadaşları[177] tarafından gösterilmiştir.

Henüz geliştirme aşamasında yapılan bu analizler ile bulunan sonuçların doğrulanması çalışmalarına iş birliği yapılan ekiple devam edilmektedir.

TARTIŞMA

Bu tez projesi kapsamında geliştirilen platformun temel amacı, kanser alanında çalışmalar sürdüren araştırmacılara tek bir kaynaktan biyoinformatik temelli analiz sonuçlarına ulaşabilecekleri ve ilerleyen aşamalarda platformun desteklediği yöntemleri kullanarak kendi verilerini analiz edebilecekleri bir ortam sağlamaktır.

Projenin geliştirilebilmesindeki en önemli bileşenlerden bir tanesi analizleri gerçekleştirilecek verinin toplanmasıdır. Mevcut versiyonda 100'e yakın farklı çalışmadan toplanmış çeşitli kanser türlerine ilişkin veri setleri kullanılmıştır. Ancak bu verilerin standart bir formda olmayışı ve klinik bilgilerin paylaşılmıyor olması ya da klinik bilgilerin de farklı formatlarda paylaşılması ön işleme ve verilerin analize hazırlanması sürecinde çeşitli zorluklar yaşatmıştır. Özellikle klinik verilerin saklanma şekillerinde bir standart olmayışı analizlerle ilgili süreci zorlaştırmıştır. MIAME[178] gibi çeşitli standartlar genomik veri için belirlenmiş olsa da farklı kaynakların farklı formatta veri talepleri, standartların her türlü veri için belirlenmemiş olması ve klinik verinin ne şekilde eşlik edeceğinin tam olarak belirlenmemesi farklı kaynaklardan toplanan verinin analiz edilebilir olmasını zorlaştırmaktadır. Gelecek nesil dizileme teknolojileri ile daha da artması beklenen veri hacminin bir an önce standart bir saklama formatına dönüştürülmesi bundan sonraki araştırmalar için kolaylık sağlayacaktır.

Geliştirilen platform ile genomik ve klinik verinin paylaşılması amaçlanmıştır. Bu nedenle sistem üzerinde depolanan veri, yaygın kullanımı göz önüne alınarak R istatistiksel programlama dilinin kullanabileceği veri paketlerine dönüştürülmüş ve araştırmacılarla paylaşılabilir hale getirilmiştir. Bu, araştırmacılara analizleri farklı parametrelerle kendilerine ait yerel sunucularda tekrar etme ya da platformun desteklemediği analizleri kendi ortamlarında uygulayabilme açısından avantaj sağlayacaktır.

Geliştirilen platform üzerinde çeşitli kanser türlerinden elde edilmiş genomik veri ve eğer var ise beraberindeki klinik veri kullanılarak 11 farklı türde analiz yapılabilmekte ve bu analizlerden elde edilen bulgular için 2 farklı meta analiz yöntemi uygulanabilmektedir. Kanser araştırmalarında biyolojik olarak anlamlı hipotezler üretip araştırma sorularına cevap vermek amacıyla geliştirilmiş platformda kullanılan analizlerin tamamı işbirliği yapılan kanser araştırmacıları ile beraber seçilmiş ve araştırmacıların deneyimleri bu konuda yol gösterici etmenlerden bir tanesi olmuştur. Mevcut versiyonda çoğunluğu MM araştırmacıları olmak üzere 9 farklı ülkeden araştırmacın verdiği geri bildirimlerle platformun geliştirilmesine ve yeni fonksiyonlar eklenmesine devam edilmektedir.

Analizlerin belirlenmesi ve verinin hazırlanması kadar analizleri gerçekleştirecek altyapının da kurulması önemlidir. Proje kapsamında analizlerin

gerçekleştirilebilmesi, verilerin yönetilebilmesi ve saklanabilmesi için çok katmanlı mimaride birden fazla sunucu üzerinde çalışan bir platform geliştirilmiştir. Geliştirilen analiz altyapısı platformun desteklediği bütün analizleri uygulayabilecek şekilde kurulmasına karşın biyoinformatik analiz araçlarındaki çeşitlilik nedeni ile kopya sayısı analizinin önişleme işlemleri platformun dışında gerçekleştirilmiştir. Birbirinden farklı platformlardaki ve farklı analiz araçlarını gerektiren veriyi işleyebilecek bir altyapı oluşturmak avantaj sağlamanın yanında çeşitli yönetim zorluklarını da beraberinde getirmiştir.

Analizlerin gerçekleştirilmesinde verilen öncelikli karar, analizlerde kullanılacak verinin düzeyi ve analizlerin çalışma zamanında mı yoksa sistem yöneticisi onayı ile daha önce mi analiz edileceğini belirlemektir.

Genomik verilerin çok boyutlu ve yüksek hacimli olması nedeniyle analizlerin gerçekleştirilebilmesi için yüksek işgücüne sahip bilgisayar kaynaklarının kullanılması gerekmektedir. Bu nedenle geliştirilen platform analizleri iki farklı ortamda gerçekleştirmektedir. Bir tanesi kümelenmiş bilgisayarlardan oluşan analiz sunucusu diğeri ise yerel ve küçük çaplı analizlerin gerçekleştirilmesinde kullanılan 16 işlemcili yerel analiz sunucusudur.

İki farklı ortamda analizlerin yönetilmesi karmaşık bir süreçtir ve bu nedenle analiz katmanı kuyruk mantığını uygulamaktadır. Eğer analiz yükü fazla ise yeni gelen analiz istemleri kuyruğa alınarak bekletilmektedir. Bu şekilde sunucular üzerinde oluşabilecek hatalara karşı önlem alınmış ve analizlerin sağlıklı şekilde yürütülmesi hedeflenmiştir. Bu aşamada sadece sistem yöneticisi onayı ile gerekli analizlerin DFCI üzerindeki diğer araştırmacılar ile beraber kullanılan kümelenmiş analiz ortamına taşınmasına olanak tanınmış, bunun haricinde kalan ve platformun anonim kullanıcıları olan araştırmacılar tarafından gerçekleştirilen analizlerin lokal sunucu üzerinde işleme alınması hedeflenmiştir.

Günümüzde araştırmacılara onko genetik veri üzerinden sorgulama ya da analiz yapma olanağı sağlayan çeşitli veritabanları veya web portalları bulunmaktadır. Ancak bu çalışmaların pek çoğu tek bir analiz türüne veya veri türüne odaklanmış çalışmalardır (caSNP[148] , GCOD[179], PrognoScan[180]). Geliştirilen platform mevcutlardan farklı olarak pek çok analiz türünü, farklı veri tipleri ve kanser türleri için tek bir kaynaktan sunabilmektedir. Ayrıca sağladığı görselleştirme seçenekleri ile verinin ve analiz sonuçlarının anlaşılmanı kolaylaştırmaktadır. Geliştirilmiş olan platformun benzer öncüllerine göre en önemli avantajlarından biri, araştırmacılar için daha kapsamlı analiz ve görselleştirme olanakları sunabilmesidir.

Kısıtlı sayıda analizle ve sadece üçüncü seviye TCGA verisini kullanacak şekilde geliştirilen cBio[181] projesinden farklı olarak açık olarak erişilebilen bütün verileri kullanma kapasitesine sahip olan platform, ayrıca sunduğu farklı analiz seçenekleri ile de kanser araştırmacıları için büyük bir kanyak oluşturmaktadır.

Bir başka benzer çalışma ise Oncomine[10, 11] projesidir. Ancak bu proje pek çok özelliği araştırmacılara ücretli olarak sunmakla beraber ne cBio ne de Oncomine projesi ağ analizlerini ve entegre analiz yöntemlerini desteklememektedir. Tek bir tür veriye ya da kanser türüne odaklanmadan geliştirilen projede, sorgulama ve analiz modülleri de genel amaçlı oluşturulmuştur. Bu sayede geliştirilen platform transkripsiyon faktörleri ya da miRNA gibi gen ekspresyonunu düzenleyici faktörlerin tespit edilmesinde ya da yolaklar seviyesinde biyolojik olarak anlamlı anormal gen ekspresyonu ya da kopya sayısına sahip gen veya gen gruplarının tespit edilmesinde ve bunların sağ kalım üzerine olan etkilerinin ortaya konulmasında kolay kullanılır bir araç olmuştur. Mevcutta var olan kaynakların hiçbirisi bu kadar kapsamlı fonksiyonlar sunmamaktadırlar.

Geliştirilen platformun mevcut durumu daha çok depolanan veri kaynakları üzerinden araştırmacılara ışık tutuyor olsa da kullanıcıların kendi verilerini de analiz edebilmelerini sağlayan çeşitli fonksiyonlar ile donatılmıştır. Araştırma ekibi tarafından geliştirilen "GemiNi" analiz modülü, gen kopya sayısı ve ekspresyonunu entegre etmek için oluşturulmuş "DosageEffect" paketi ve GEO gen ekspresyonu verisini analiz etmede kullanılan "GEO-Miner" bunlardan bazılarıdır. Ancak henüz depolanmış veri de olduğu gibi bütün analizlerin kullanıcılar tarafından geliştirilen platformun kaynaklarını kullanarak yürütülmesi mümkün değildir.

Araştırmacılar araştırmanın planlanmasında, yürütülmesinde, bulgularının farklı veri ve yöntemlerle incelenmesi gerektiğinde bir çok kaynağa gereksinim duymaktadırlar. Bu çalışmada oluşturulan platform tek bir çatı altında kanser araştırması yürüten bir araştırmacıya gereksinim duyabileceği pek çok kaynağı sağlamayı hedeflemektedir.

Bir araştırmanın yürütülmesinde en temel unsur veridir. Kimi durumlarda genomik araştırmalarda araştırma ekipleri kendi ürettikleri verileri kullanırken kimi durumlarda da literatürde tanımlanmış veriye ihtiyaç duyarlar. Ancak mevcut kaynaklardan sadece kanser araştırmalarına özgü veriye erişim çok kolay değildir. Bu nedenle bu kanser araştırmalarını temel alarak veriyi depolayan ve farklı kanser türlerine göre gruplayabilen bir platform araştırmacıların aradıkları veriye erişebilmeleri bakımından büyük bir avantajdır. Genomik verinin yanında klinik verinin de araştırmacılar ile paylaşılabilir ve bütün veri setleri için aynı standartlar gözetilerek veriyle ilişkilendirilmiş olması, paylaşılan verinin önemini arttırmaktadır. Geliştirilen platform getirdiği bu avantajlar ile kanser araştırması yürüten ekipler için önemli bir başvuru kaynağı olacağı öngörülmektedir.

Bununla beraber veriye erişim bu çalışmalarda en kritik sorun değildir. Araştırmacılar aradıkları veriye literatür veya farklı genomik veritabanları sayesinde ulaşsalar bile bu verilerin analizlere hazır hale getirilmesi için gereken uzmanlık bilgisi, zaman ve teknik altyapı gereksinimleri veriye erişim kadar ciddi bir sorundur. Günümüzde genomik verinin üretilmesinde farklı platformlar kullanılmaktadır. Hepsi temelde hücre içindeki aynı genomik fonksiyonu ölçmek için kullanılıyor olsalar da aslında farklı platformlardan elde edilen verinin analizlere hazırlanma süreçleri de farklıdır. Her bir veri türü ve platform için uygun ön işleme araçlarının,

algoritmalarının ve teknik altyapının hazırlanması gerekmektedir. Ayrıca aynı veri türü için literatürde var olan pek çok yöntemden uygun olanının seçilebilmesi nitelikli uzman gerektirmektedir. Bu çalışmanın yürütülmesi sırasında beraber çalışılan farklı uzmanların bilgi ve deneyimleri uygun yöntemlerin belirlenmesinde önemli bir yol gösterici olmuştur. Ayrıca farklı platform ve veri türleri için gerekli olan onlarca aracın ve güçlü teknik altyapının tek bir platform altında toplanması, bu platfomu mevcut araçlardan farklı olarak güçlü bir veri hazırlama aracına dönüştürmüştür. Mevcut hali ile farklı platformlardan elde edilen mikrodizi ya da gelecek nesil dizileme verilerinin ön işleme gereksinimlerini verinin türüne ve elde edildiği platfoma uygun şekilde gerçekleştirebilen canEvolve projesi ,farklı platformları bir arada destekleyebilen ender biyoinformatik araçlarındadır.

Genomik temelli araştırmalarda en önemli aşama veri analizinin hipoteze göre en uygun biçimde gerçekleştirilebilmesidir. Farklı genomik fonksiyonlara ait verilerin analiz edilmesi çeşitli biyoinformatik analiz araçları gerekmektedir. Ancak bu araçların kullanılabilmesi için araştırmacıların veri türüne ve platformuna uygun pek çok yardımcı parametreyi de yönetmesi gerekir. Bu durum üstesinden gelmek araştırmacılar için bir handikaptır. Çünkü çoğunluğu sistem seviyesinde çalışan ve çok da kullanıcı dostu olmayan biyoinfomatik araçlarının kullanımı özellikle teknik anlamda yeterli desteği bulmayan araştırma ekipleri için problemler oluşturmaktadır. Ayrıca farklı kaynaklar için farklı gereksinimler olacağından analizlerin belirli bir sistematik oturtulmadan yürütülmesi zordur. Bilimsel çalışmaların en başında planlanan araştırma sorularına ek olarak araştırma ilerledikçe üretilebilen yeni hipotezlere hızlı yanıt üretmek de zordur. Bunun zorluğun temel nedeni, analizlerin yeni hipotezlere göre tekrarlanması ya da ek analizlerin yapılma gerekliliğidir. Bütün bunlar dikkate alındığında geliştirilen platform sağladığı temel, entegre ve ağ analizleri ile güçlü bir biyoinformatik analiz aracıdır. Oluşturulan altyapı pek çok analizin kullanıcı tarafında belirlenen veri setleri ve gen sembolleri ile hızlı ve kolay bir şekilde yapılabilmesine olanak vermekle birlikte araştırmacının yeni bulgu ve bilgilere göre analizleri tekrarlamasını ya da farklı veri setleri kullanarak bulgularını doğrulamasını sağlamaktadır. Bunu sağlamak için oluşturmuş analiz altyapısı tamamen kanser araştırması yürüten ekiplerin iş birliği ile oluşturulduğundan araştırmalarda sıklıkla karşılaşılan pek çok soruya cevap sağlayabilir niteliktedir. Ayrıca geliştirilen platform desteklediği 11 farklı analiz türü ile araştırmacılara büyük getiriler sağlayabilecek düzeydedir.

Genomik temelli analizlerin seçimi araştırma konularına göre farklılık göstermektedir. Genomik değişimlerden kaynaklanan hastalıklar arasında da farklılıklar vardır. Bu nedenle doğru veri türünü doğru yöntemlerle analiz etmek oldukça önemlidir. Bu projede kullanılan analizlerin seçilmesi aşamasında farklı kanser türleri üzerinde uzmanlaşmış araştırmacıların ortak görüşü ile belirlenmiş ve literatürde sıklıkla başvurulan analizler dikkate alınmıştır. Temel analiz özelliklerine bakıldığında gen ekspresyonu, miRNA ekspresyonu ve gen kopya sayısı değişimini temel alan analizleri sunan platform, bizim bildiğimiz kadarı ile bu 3 farklı veri türü için temel analizleri tek bir ortamdan sağlayan tek platformdur. Temel analizler tek başlarına belirli gen veya SNP'ler için güçlü sonuçlar bulsalar da kanser gibi

kompleks hastalıklarda sadece gen veya SNP bazındaki değişiklik yerine biyolojik olarak anlamlı ve iş birliği içinde çalışan gen veya SNP gruplarının belirlenmesi önemlidir. Bunun için iki farklı sağ kalım analizi yöntemini destekleyen platform, ayrıca bu ağların görselleştirilebilir olmasını da sağlayarak elde edilen bulguların araştırmacılar tarafından yorumlanabilmesine kolaylık sağlamaktadır.

Hücre içinde yaşanan genomik fonksiyonların hücrede bir döngü ile ilerlediği biyolojik olarak bilinmektedir. Bu döngü dikkate alındığında farklı genomik değişimlerin birbirini nasıl etkilediğini ortaya koyabilmek kompleks değişikliklerin anlaşılması bakımından biyolojik olarak oldukça önemlidir. Bunu temel alarak geliştirilen 5 farklı entegre analiz yöntemi ile güçlendirilen platform, kanserin kompleks yapısı göz önüne alındığında araştırmalar için önemli bulgulara hızlı erişim sağlayabilecek kapasitededir.

Kanser, neden olduğu sonuçlar göz önüne alındığında araştırmacılar tarafından titizlikle değerlendirilmesi gereken bir hastalıktır. Yaşanan bir genomik değişimin sağ kalım süresine etkisi ya da tedaviden sonra hastalığın tekrarlamasına kadar geçen süreye etkisini bilmek hastalıkla mücadelede öneli bir adımdır. Bu nedenle sağ kalım süresine ilişkin klinik veriyi bulunduran veri setleri ile sağ kalım analizlerini de yapabilmeyi sağlayan sistem, araştırmacıların yaşanan genomik değişimle ilişkili olarak hastalığın seyrinin ne şekilde değiştiğini görebilmelerini sağlamaktadır. Sistem, içinde barındırdığı meta analiz altyapısı ile farklı genomik veri setlerinden elde edilen bulguların bir biri ile karşılaştırılabilir ve birleştirilebilir olmasını sağlamaktadır. Kullanıcılar, elde ettikleri bulguları aynı türe ait farklı veri setlerinde test edebilmekte veya türler arasındaki uyumunu platform yardımı ile karşılaştırabilmektedirler.

Mevcut platformlar incelendiğinde bu kadar farklı veri türünü ve analizi tek bir ortamdan sunan başka bir çalışma bulunmamaktadır. Her ne kadar kullanıcı tarafında işlemlerin kolaylaştırılması benimsenmiş olsa da analiz tarafında aslında durum farklıdır. Farklı platformlardan elde edilen ve farklı genomik fonksiyonlara ait veri türlerinin uygun şekilde analiz edilebilmesi için 100'e yakın farklı biyoinformatik analiz aracını ve programlama paketini birleştiren platform araştırmanın planlanmasından, sonuçlandırılmasına kadar geçen süreçte biyoinformatik yöntemler ile cevaplanabilecek sorular için araştırmacılara büyük bir kanser araştırma kütüphanesi sunmaktadır.

Literatürdeki verilerin analiz edilebilir olmasının yanında kullanıcılar kendi verilerini de analiz edebilecek kolay kullanılan ve çeşitli tipte analizleri gerçekleştiren araçlara ihtiyaç duymaktadırlar. Platformda desteklenen farklı analiz yöntemlerinin bir kısmı kullanıcıların kendi verilerini de analiz edebilecekleri web tabanlı analiz aracına dönüştürülmüştür. Bu sayede araştırmacılara literatürdeki verilerin analizinden ve kendi verilerinden elde edilen bulguları birleştirme imkanı sağlanmaktadır. Mevcut hali ile çeşitli teknik kısıtlılıkları bulunan analiz araçları için ilerleyen versiyonlarda bu sınırların giderilmesi hedeflenmektedir.

Çalışmanın ilerleyen aşamalarında bütün analizlerin bulut bilişim teknolojisi kullanılarak Hadoop gibi bir altyapıya taşınması için çalışmalar devam etmektedir. Ayrıca analizlerin tamamının R ortamından kontrol edilebilmesini sağlayacak analiz alt yapasını geliştirme çalışmaları da devam etmektedir. Bu işlemlerde de sona ulaşıldığında, ilerleyen versiyonlarda, özellikle karmaşık analiz araçlarını kullanamayan, analizlerin gerçekleştirilmesi için yeterli bilişim altyapısı olmayan araştırmacılara kendi verilerini analiz edebilecekleri kolay kullanılabilir online bir platformda sağlanmış olacaktır. Ayrıca platform üzerindeki mevcut veriler ile kendi verilerinden elde ettikleri bulguları birleştirebilmelerine olanak sağlanmış olacaktır.

SONUÇLAR

1. Farklı tümör tiplerinden ve farklı genomik veri platformlarından elde edilen verilerin tek bir biyoinformatik kaynağında toplanması oldukça karmaşık bir süreçtir. Ancak genomik veri alanında standartların belirlenmesi ve veri üreten platformların bu standartlara uygun veriler üretmeye başlamasıyla bu sorunun üstesinden gelmede avantaj sağlayacaktır.

2. Kanser kompleks bir hastalıktır ve aynı orijinli türlerin bile moleküler seviyede ciddi farkları olabilmektedir. Bu kompleks yapıyı analiz edebilecek güçlü araçlara ve analiz metotlarına ihtiyaç vardır. Ayrıca hastalığa yol açan genetik değişikliklerin zaman içerisinde ne şekilde farklılaştığını anlayabilecek araç ve yöntemler kullanmak kanserde tanı, tedavi ve yeni ilaç hedeflerinin geliştirilmesi açısından büyük önem taşımaktadır

3. Genomik araştırmaların gerçekleştirilmesinde genomdan elde edilen veri kadar eşlik eden klinik verinin de önemi büyüktür. Genomik veri ile beraber klinik veri paylaşımının artması bu kompleks hastalığı bilim insanlarının daha iyi anlamasını sağlayacaktır.

4. Farklı analiz yöntemlerini barındıran ve araştırmacılara kolay kullanılabilir şekilde sunan veri ve analiz kaynaklarının araştırmaların planlanması ve yürütülmesi açısından büyük avantajları vardır

5. Aalizlerden elde edilen bulguların birleştirilebilmesi hastalığın anlaşılmasında, yeni teşhis, tedavi yöntemlerinin planlanmasında büyük rol oynamaktadır

6. Genomik veri için temel, entegre ve ağ analizlerini bir arada sunabilen ve görselleştirebilen araçlar büyük önem taşımaktadır. Biyolojik olarak karmaşık olan süreçlerin tam olarak ortaya konulabilmesi, güçlü bulgular ile biyolojik doğrulamaların gerçekleştirilmesi bu tarz araçların başarısına bağlıdır

7. Genomik verinin analiz edilebilmesi için oluşturulması gereken analiz altyapısı oldukça karmaşıktır ve farklı uzmanlık alanlarının bir araya gelmesi ile başarılı bir şekilde kurulabilmektedir.

8. Geliştirilen platform ile standartlara uygun veri sağlayan çevrimiçi veritabanlarından sistem yöneticilerinin verileri otomatik olarak sisteme alabilmeleri sağlanmıştır

9. Platform ön işleme kapasitesi bakımından farklı veri türleri ve platformlar için işler şekilde tasarlanmış bu sayede farklı araştırmalarda toplanan ve paylaşılan verinin analizlere uygun hale getirilmesi sağlanmıştır.

10. Platforma eklenen temel, ağ entegre analiz yöntemleri ile kanser çalışmalarında araştırmacıların kullanabilecekleri büyük bir araştırma alt yapısı kurulmuştur. Ayrıca genomik analiz yöntemleri sağ kalım gibi klinik açıdan önemli yöntemlerle desteklenmiştir.

11. Geliştirilen platform ile farklı veri setlerinden elde edilen bulgular birbiri ile karşılaştırılabilir hale getirilmiş ve bu sayede kanser tipleri arasında ya da aynı kanser tipine ilişkin farklı çalışmalardaki sonuçların incelenmesi sağlanmıştır

12. Biyoinformatik alanında kullanılan 100'den fazla araç ve kütüphanenin birbiri ile entegrasyonu sayesinde güçlü bir biyoinformatik analiz platformu kurulmuştur

Geliştirilen platform mevcut araçlarla karşılaştırıldığında desteklediği analiz yöntemi ve içerdiği veri büyüklüğü bakımında çok daha kapsamlı bir araç olarak görülmektedir

KAYNAKLAR

1. Pennisi E. Human genome 10th anniversary. Will computers crash genomics? Science 2011, 331(6018):666-668.

2. Cancer Genome Atlas Research N. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008, 455(7216):1061-1068.

3. International Cancer Genome C, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabe RR, Bhan MK, Calvo F, Eerola I, Gerhard DS, Guttmacher A, Guyer M, Hemsley FM, Jennings JL, Kerr D, Klatt P, Kolar P, Kusada J, Lane DP *et al*. International network of cancer genome projects. Nature 2010, 464(7291):993-998.

4. Collins FS, Green ED, Guttmacher AE, Guyer MS, Institute USNHGR. A vision for the future of genomics research. Nature 2003, 422(6934):835-847.

5. Huang N, Shah PK, Li C. Lessons from a decade of integrating cancer copy number alterations with gene expression profiles. Brief Bioinform 2012, 13(3):305-316.

6. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Holko M, Ayanbule O, Yefanov A, Soboleva A. NCBI GEO: archive for functional genomics data sets--10 years on. Nucleic Acids Res 2011, 39(Database issue):D1005-1010.

7. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002, 30(1):207-210.

8. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, Sansone SA. ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 2003, 31(1):68-71.

9. Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Holloway E, Kurbatova N, Lukk M, Malone J, Mani R, Pilicheva E, Rustici G, Sharma A, Williams E, Adamusiak T, Brandizi M *et al*. ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments. Nucleic Acids Res 2011, 39(Database issue):D1002-1004.

10. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, Barrette TR, Anstet MJ, Kincead-Beal C, Kulkarni P, Varambally S, Ghosh D, Chinnaiyan AM. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. Neoplasia 2007, 9(2):166-180.

11. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM. ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia 2004, 6(1):1-6.

12. Hruz T, Laule O, Szabo G, Wessendorp F, Bleuler S, Oertle L, Widmayer P, Gruissem W, Zimmermann P. Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. Adv Bioinformatics 2008, 2008:420747.

13. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 1953, 171(4356):737-738.

14. Leder P. Retrospective. Marshall Warren Nirenberg (1927-2010). Science 2010, 327(5968):972.

15. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 1977, 74(12):5463-5467.

16. Lander ES, Schork NJ. Genetic dissection of complex traits. Science 1994, 265(5181):2037-2048.

17. Bloss CS, Jeste DV, Schork NJ. Genomics for disease treatment and prevention. Psychiatr Clin North Am 2011, 34(1):147-166.

18. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995, 270(5235):467-470.

19. Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY, Brown PO, Davis RW. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci U S A 1997, 94(24):13057-13062.

20. Venter JC. A part of the human genome sequence. Science 2003, 299(5610):1183-1184.

21. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K *et al*. Initial sequencing and analysis of the human genome. Nature 2001, 409(6822):860-921.

22. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M *et al*. The sequence of the human genome. Science 2001, 291(5507):1304-1351.

23. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC *et al*. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005, 437(7057):376-380.

24. Shokralla S, Spall JL, Gibson JF, Hajibabaei M. Next-generation sequencing technologies for environmental DNA research. Mol Ecol 2012, 21(8):1794-1805.

25. Galperin MY, Fernandez-Suarez XM. The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. Nucleic Acids Res 2012, 40(Database issue):D1-8.

26. Gaudet P, Bairoch A, Field D, Sansone SA, Taylor C, Attwood TK, Bateman A, Blake JA, Bult CJ, Cherry JM, Chisholm RL, Cochrane G, Cook CE, Eppig JT, Galperin MY, Gentleman R, Goble CA, Gojobori T, Hancock JM, Howe DG *et al*. Towards BioDBcore: a community-defined information specification for biological databases. Nucleic Acids Res 2011, 39(Database issue):D7-10.

27. Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, Hill DP, Kania R, Schaeffer M, St Pierre S, Twigger S, White O, Rhee SY. Big data: The future of biocuration. Nature 2008, 455(7209):47-50.

28. Cochrane GR, Galperin MY. The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources. Nucleic Acids Res 2010, 38(Database issue):D1-4.

29. Landsman D, Gentleman R, Kelso J, Francis Ouellette BF. DATABASE: A new forum for biological databases and curation. Database (Oxford) 2009, 2009:bap002.

30. Attwood TK, Kell DB, McDermott P, Marsh J, Pettifer SR, Thorne D. Calling International Rescue: knowledge lost in literature and data landslide! Biochem J 2009, 424(3):317-333.

31. Seringhaus MR, Gerstein MB. Publishing perishing? Towards tomorrow's information architecture. BMC Bioinformatics 2007, 8:17.

32. Philippi S, Kohler J. Addressing the problems with life-science databases for traditional uses and systems biology. Nat Rev Genet 2006, 7(6):482-488.

33. Goble C, Stevens R. State of the nation in data integration for bioinformatics. J Biomed Inform 2008, 41(5):687-693.

34. Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I, Kimelman M, Pruitt KD, Resenchuk S, Tatusova T, Yaschenko E, Ostell J. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. Nucleic Acids Res 2012, 40(Database issue):D57-63.

35. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, Popova N, Pretel S, Ziyabari L, Lee M, Shao Y, Wang ZY, Sirotkin K, Ward M, Kholodov M, Zbicz K *et al*. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 2007, 39(10):1181-1186.

36. Pruitt KD, Tatusova T, Klimke W, Maglott DR. NCBI Reference Sequences: current status, policy and new initiatives. Nucleic Acids Res 2009, 37(Database issue):D32-36.

37. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001, 29(1):308-311.

38. Kodama Y, Shumway M, Leinonen R, International Nucleotide Sequence Database C. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res 2012, 40(Database issue):D54-56.

39. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database C. The sequence read archive. Nucleic Acids Res 2011, 39(Database issue):D19-21.

40. Barrett T, Edgar R. Mining microarray data at NCBI's Gene Expression Omnibus (GEO)*. Methods Mol Biol 2006, 338:175-190.

41. Barrett T, Edgar R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. Methods Enzymol 2006, 411:352-369.

42. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R. NCBI GEO: mining millions of expression profiles--database and tools. Nucleic Acids Res 2005, 33(Database issue):D562-566.

43. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R. NCBI GEO: mining tens of millions of expression profiles--database and tools update. Nucleic Acids Res 2007, 35(Database issue):D760-765.

44. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Edgar R. NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res 2009, 37(Database issue):D885-890.

45. Edgar R, Barrett T. NCBI GEO standards and services for microarray data. Nat Biotechnol 2006, 24(12):1471-1472.

46. Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, Kaloper M, Weng S, Jin H, Ball CA, Eisen MB, Spellman PT, Brown PO, Botstein D, Cherry JM. The Stanford Microarray Database. Nucleic Acids Res 2001, 29(1):152-155.

47. Tateno Y, Imanishi T, Miyazaki S, Fukami-Kobayashi K, Saitou N, Sugawara H, Gojobori T. DNA Data Bank of Japan (DDBJ) for genome scale research in life science. Nucleic Acids Res 2002, 30(1):27-30.

48. Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tarraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R, Hoad G, Jang M, Pakseresht N, Plaister S, Radhakrishnan R, Reddy K, Sobhany S, Ten Hoopen P, Vaughan R, Zalunin V *et al*. The European Nucleotide Archive. Nucleic Acids Res 2011, 39(Database issue):D28-31.

49. Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res 2012, 40(Database issue):D48-53.

50. Cochrane G, Karsch-Mizrachi I, Nakamura Y, International Nucleotide Sequence Database C. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res 2011, 39(Database issue):D15-18.

51. Karsch-Mizrachi I, Nakamura Y, Cochrane G, International Nucleotide Sequence Database C. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res 2012, 40(Database issue):D33-37.

52. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kahari AK, Keefe D, Keenan S, Kinsella R, Komorowska M *et al*. Ensembl 2012. Nucleic Acids Res 2012, 40(Database issue):D84-90.

53. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E *et al*. The Ensembl genome database project. Nucleic Acids Res 2002, 30(1):38-41.

54. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C. STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 2009, 37(Database issue):D412-416.

55. Snel B, Lehmann G, Bork P, Huynen MA. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res 2000, 28(18):3442-3444.

56. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 2011, 39(Database issue):D561-568.

57. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000, 28(1):27-30.

58. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 2012, 40(Database issue):D109-114.

59. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res 2005, 33(Database issue):D428-432.

60. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B, Kanapin A, Lewis S, Mahajan S, May B, Schmidt E, Vastrik I, Wu G, Birney E, Stein L, D'Eustachio P. Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res 2009, 37(Database issue):D619-622.

61. Joshi-Tope G, Vastrik I, Gopinath GR, Matthews L, Schmidt E, Gillespie M, D'Eustachio P, Jassal B, Lewis S, Wu G, Birney E, Stein L. The Genome Knowledgebase: a resource for biologists and bioinformaticists. Cold Spring Harb Symp Quant Biol 2003, 68:237-243.

62. Seal RL, Gordon SM, Lush MJ, Wright MW, Bruford EA. genenames.org: the HGNC resources in 2011. Nucleic Acids Res 2011, 39(Database issue):D514-519.

63. Stein LD. Integrating biological databases. Nat Rev Genet 2003, 4(5):337-345.

64. Forman MR, Greene SM, Avis NE, Taplin SH, Courtney P, Schad PA, Hesse BW, Winn DM. Bioinformatics: Tools to accelerate population science and disease control research. Am J Prev Med 2010, 38(6):646-651.

65. Doctorow C. Big data: Welcome to the petacentre. Nature 2008, 455(7209):16-21.

66. Stein L. Creating a bioinformatics nation. Nature 2002, 417(6885):119-120.

67. Stajich JE, Lapp H. Open source tools and toolkits for bioinformatics: significance, and where are we? Brief Bioinform 2006, 7(3):287-296.

68. Brazas MD, Yim D, Yeung W, Ouellette BF. A decade of Web Server updates at the Bioinformatics Links Directory: 2003-2012. Nucleic Acids Res 2012, 40(Web Server issue):W3-W12.

69. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G *et al*. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004, 5(10):R80.

70. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 2004, 20(3):307-315.

71. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. Bioinformatics 2008, 24(13):1547-1548.

72. Hahne F, LeMeur N, Brinkman RR, Ellis B, Haaland P, Sarkar D, Spidlen J, Strain E, Gentleman R. flowCore: a Bioconductor package for high throughput flow cytometry. BMC Bioinformatics 2009, 10:106.

73. Dvinge H, Bertone P. HTqPCR: high-throughput analysis and visualization of quantitative real-time PCR data in R. Bioinformatics 2009, 25(24):3325-3326.

74. Mar JC, Kimura Y, Schroder K, Irvine KM, Hayashizaki Y, Suzuki H, Hume D, Quackenbush J. Data-driven normalization strategies for high-throughput quantitative RT-PCR. BMC Bioinformatics 2009, 10:110.

75. Morgan M, Anders S, Lawrence M, Aboyoun P, Pages H, Gentleman R. ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. Bioinformatics 2009, 25(19):2607-2608.

76. Delhomme N, Padioleau I, Furlong EE, Steinmetz LM. easyRNASeq: a bioconductor package for processing RNA-Seq data. Bioinformatics 2012, 28(19):2532-2533.

77. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010, 26(1):139-140.

78. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A. BioMart--biological queries made easy. BMC Genomics 2009, 10:22.

79. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJ. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009, 25(11):1422-1423.

80. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD *et al*. The Bioperl toolkit: Perl modules for the life sciences. Genome Res 2002, 12(10):1611-1618.

81. Li C, Wong WH. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A 2001, 98(1):31-36.

82. Lin M, Wei LJ, Sellers WR, Lieberfarb M, Wong WH, Li C. dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. Bioinformatics 2004, 20(8):1233-1240.

83. Zhong S, Li C, Wong WH. ChipInfo: Software for extracting gene annotation and gene ontology information for microarray analysis. Nucleic Acids Res 2003, 31(13):3483-3486.

84. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005, 102(43):15545-15550.

85. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. Nat Biotechnol 2011, 29(1):24-26.

86. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 2006, 7 Suppl 1:S7.

87. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. Nat Genet 2006, 38(5):500-501.

88. Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J. Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol 2010, Chapter 19:Unit 19 10 11-21.

89. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A. Galaxy: a platform for interactive large-scale genome analysis. Genome Res 2005, 15(10):1451-1455.

90. Goecks J, Nekrutenko A, Taylor J, Galaxy T. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 2010, 11(8):R86.

91. Romero IG, Ruvinsky I, Gilad Y. Comparative studies of gene expression and the evolution of gene regulation. Nat Rev Genet 2012, 13(7):505-516.

92. Bates S. The role of gene expression profiling in drug discovery. Curr Opin Pharmacol 2011, 11(5):549-556.

93. Consortium M, Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, Lee KY, Luo Y, Sun YA, Willey JC, Setterquist RA, Fischer GM, Tong W, Dragan YP, Dix DJ, Frueh FW *et al*. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 2006, 24(9):1151-1161.

94. Yang YH, Speed T. Design issues for cDNA microarray experiments. Nat Rev Genet 2002, 3(8):579-588.

95. Barnes M, Freudenberg J, Thompson S, Aronow B, Pavlidis P. Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. Nucleic Acids Res 2005, 33(18):5914-5923.

96. Dalma-Weiszhausz DD, Warrington J, Tanimoto EY, Miyada CG. The affymetrix GeneChip platform: an overview. Methods Enzymol 2006, 410:3-28.

97. Bi CL, Chng WJ. miRNA deregulation in multiple myeloma. Chin Med J (Engl) 2011, 124(19):3164-3169.

98. Gao W, Xu J, Shu YQ. miRNA expression and its clinical implications for the prevention and diagnosis of non-small-cell lung cancer. Expert Rev Respir Med 2011, 5(5):699-709.

99. Gommans WM, Berezikov E. Controlling miRNA regulation in disease. Methods Mol Biol 2012, 822:1-18.

100. Baranzini SE, Galwey NW, Wang J, Khankhanian P, Lindberg R, Pelletier D, Wu W, Uitdehaag BM, Kappos L, Gene MSAC, Polman CH, Matthews PM, Hauser SL, Gibson RA, Oksenberg JR, Barnes MR. Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. Hum Mol Genet 2009, 18(11):2078-2090.

101. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science 2008, 322(5903):881-888.

102. Holmans P, Green EK, Pahwa JS, Ferreira MA, Purcell SM, Sklar P, Wellcome Trust Case-Control C, Owen MJ, O'Donovan MC, Craddock N. Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. Am J Hum Genet 2009, 85(1):13-24.

103. Wu J, Smith LT, Plass C, Huang TH. ChIP-chip comes of age for genome-wide functional analysis. Cancer Res 2006, 66(14):6899-6902.

104. Hurd PJ, Nelson CJ. Advantages of next-generation sequencing versus the microarray in epigenetic research. Brief Funct Genomic Proteomic 2009, 8(3):174-183.

105. Yang Y, Adelstein SJ, Kassis AI. Integrated bioinformatics analysis for cancer target identification. Methods Mol Biol 2011, 719:527-545.

106. Zhang M, Liang L, Morar N, Dixon AL, Lathrop GM, Ding J, Moffatt MF, Cookson WO, Kraft P, Qureshi AA, Han J. Integrating pathway analysis and genetics of gene expression for genome-wide association study of basal cell carcinoma. Hum Genet 2012, 131(4):615-623.

107. Pedroso I. Gaining a pathway insight into genetic association data. Methods Mol Biol 2010, 628:373-382.

108. Carlborg O, Haley CS. Epistasis: too often neglected in complex trait studies? Nat Rev Genet 2004, 5(8):618-625.

109. Louhimo R, Hautaniemi S. CNAmet: an R package for integrating copy number, methylation and expression data. Bioinformatics 2011, 27(6):887-888.

110. Schafer M, Schwender H, Merk S, Haferlach C, Ickstadt K, Dugas M. Integrated analysis of copy number alterations and gene expression: a bivariate assessment of equally directed abnormalities. Bioinformatics 2009, 25(24):3228-3235.

111. Bicciato S, Spinelli R, Zampieri M, Mangano E, Ferrari F, Beltrame L, Cifola I, Peano C, Solari A, Battaglia C. A computational procedure to identify significant overlap of differentially expressed and genomic imbalanced regions in cancer datasets. Nucleic Acids Res 2009, 37(15):5057-5070.

112. Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, Pochanard P, Mozes E, Garraway LA, Pe'er D. An integrated approach to uncover drivers of cancer. Cell 2010, 143(6):1005-1017.

113. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics 2009, 25(22):2906-2912.

114. Lee H, Kong SW, Park PJ. Integrative analysis reveals the direct and indirect interactions between DNA copy number aberrations and gene expression changes. Bioinformatics 2008, 24(7):889-896.

115. Monni O, Barlund M, Mousses S, Kononen J, Sauter G, Heiskanen M, Paavola P, Avela K, Chen Y, Bittner ML, Kallioniemi A. Comprehensive copy number and gene expression profiling of the 17q23 amplicon in human breast cancer. Proc Natl Acad Sci U S A 2001, 98(10):5711-5716.

116. Wu D, Rice CM, Wang X. Cancer bioinformatics: a new approach to systems clinical medicine. BMC Bioinformatics 2012, 13:71.

117. Wang X, Liotta L. Clinical bioinformatics: a new emerging science. J Clin Bioinforma 2011, 1(1):1.

118. Blekherman G, Laubenbacher R, Cortes DF, Mendes P, Torti FM, Akman S, Torti SV, Shulaev V. Bioinformatics tools for cancer metabolomics. Metabolomics 2011, 7(3):329-343.

119. Larkin SE, Zeidan B, Taylor MG, Bickers B, Al-Ruwaili J, Aukim-Hastie C, Townsend PA. Proteomics in prostate cancer biomarker discovery. Expert Rev Proteomics 2010, 7(1):93-102.

120. Korkola J, Gray JW. Breast cancer genomes--form and function. Curr Opin Genet Dev 2010, 20(1):4-14.

121. Byrum S, Montgomery CO, Nicholas RW, Suva LJ. The promise of bone cancer proteomics. Ann N Y Acad Sci 2010, 1192:222-229.

122. Chari R, Thu KL, Wilson IM, Lockwood WW, Lonergan KM, Coe BP, Malloff CA, Gazdar AF, Lam S, Garnis C, MacAulay CE, Alvarez CE, Lam WL. Integrating the multiple dimensions of genomic and epigenomic landscapes of cancer. Cancer Metastasis Rev 2010, 29(1):73-93.

123. Avet-Loiseau H, Li C, Magrangeas F, Gouraud W, Charbonnel C, Harousseau JL, Attal M, Marit G, Mathiot C, Facon T, Moreau P, Anderson KC, Campion L, Munshi NC, Minvielle S. Prognostic significance of copy-number alterations in multiple myeloma. J Clin Oncol 2009, 27(27):4585-4590.

124. Fernald GH, Capriotti E, Daneshjou R, Karczewski KJ, Altman RB. Bioinformatics challenges for personalized medicine. Bioinformatics 2011, 27(13):1741-1748.

125. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 2009, 106(23):9362-9367.

126. McGuire AL, Burke W. An unwelcome side effect of direct-to-consumer personal genome testing: raiding the medical commons. JAMA 2008, 300(22):2669-2671.

127. Hudis CA. Trastuzumab--mechanism of action and use in clinical practice. N Engl J Med 2007, 357(1):39-51.

128. Gambacorti-Passerini C. Part I: Milestones in personalised medicine--imatinib. Lancet Oncol 2008, 9(6):600.

129. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, Dudley JT, Ormond KE, Pavlovic A, Morgan AA, Pushkarev D, Neff NF, Hudgins L, Gong L, Hodges LM, Berlin DS, Thorn CF, Sangkuhl K, Hebert JM, Woon M *et al*. Clinical assessment incorporating a personal genome. Lancet 2010, 375(9725):1525-1535.

130. Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, Holloway A, Smyth GK. A comparison of background correction methods for two-colour microarrays. Bioinformatics 2007, 23(20):2700-2707.

131. Silver JD, Ritchie ME, Smyth GK. Microarray background correction: maximum likelihood estimation for the normal-exponential convolution. Biostatistics 2009, 10(2):352-363.

132. Kooperberg C, Fazzio TG, Delrow JJ, Tsukiyama T. Improved background correction for spotted DNA microarrays. J Comput Biol 2002, 9(1):55-66.

133. Shi W, Oshlack A, Smyth GK. Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. Nucleic Acids Res 2010, 38(22):e204.

134. Smyth GK, Speed T. Normalization of cDNA microarray data. Methods 2003, 31(4):265-273.

135. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003, 19(2):185-193.

136. Schuster SC. Next-generation sequencing transforms today's biology. Nat Methods 2008, 5(1):16-18.

137. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009, 25(14):1754-1760.

138. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010, 26(5):589-595.

139. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009, 10(3):R25.

140. Kim SY, Lee JW, Sohn IS. Comparison of various statistical methods for identifying differential gene expression in replicated microarray data. Stat Methods Med Res 2006, 15(1):3-20.

141. Lonnstedt I, Speed T. Replicated microarray data. Stat Sinica 2002, 12(1):31-46.

142. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 2004, 3:Article3.

143. Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. Bioinformatics 2001, 17(6):509-519.

144. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001, 98(9):5116-5121.

145. Broberg P. Ranking genes with respect to differential expression. Genome Biol 2002, 3(9):preprint0007.

146. Smyth GK. Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor. Edited by Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W. New York: Springer; 2005: 397-420.

147. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol 2010, 11(10):R106.

148. Cao Q, Zhou M, Wang X, Meyer CA, Zhang Y, Chen Z, Li C, Liu XS. CaSNP: a database for interrogating copy number alterations of cancer genome from SNP array data. Nucleic Acids Res 2011, 39(Database issue):D968-974.

149. Myllykangas S, Bohling T, Knuutila S. Specificity, selection and significance of gene amplifications in cancer. Semin Cancer Biol 2007, 17(1):42-55.

150. Stark M, Hayward N. Genome-wide loss of heterozygosity and copy number analysis in melanoma using high-density single-nucleotide polymorphism arrays. Cancer Res 2007, 67(6):2632-2642.

151. Weir BA, Woo MS, Getz G, Perner S, Ding L, Beroukhim R, Lin WM, Province MA, Kraja A, Johnson LA, Shah K, Sato M, Thomas RK, Barletta JA, Borecki IB, Broderick S, Chang AC, Chiang DY, Chirieac LR, Cho J *et al*. Characterizing the cancer genome in lung adenocarcinoma. Nature 2007, 450(7171):893-898.

152. Hu X, Stern HM, Ge L, O'Brien C, Haydu L, Honchell CD, Haverty PM, Peters BA, Wu TD, Amler LC, Chant J, Stokoe D, Lackner MR, Cavet G. Genetic alterations and oncogenic pathways associated with breast cancer subtypes. Mol Cancer Res 2009, 7(4):511-522.

153. Tchatchou S, Burwinkel B. Chromosome copy number variation and breast cancer risk. Cytogenet Genome Res 2008, 123(1-4):183-187.

154. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human B cells. Nat Genet 2005, 37(4):382-390.

155. Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD. Cytoscape Web: an interactive web-based network browser. Bioinformatics 2010, 26(18):2347-2348.

156. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008, 9:559.

157. Yan Z, Shah PK, Amin SB, Samur MK, Huang N, Wang X, Misra V, Ji H, Gabuzda D, Li C. Integrative analysis of gene and miRNA expression profiles with transcription factor-miRNA feed-forward loops identifies regulators in human cancers. Nucleic Acids Res 2012, 40(17):e135.

158. Salari K, Tibshirani R, Pollack JR. DR-Integrator: a new analytic tool for integrating DNA copy number and gene expression data. Bioinformatics 2010, 26(3):414-416.

159. Tiedemann RE, Zhu YX, Schmidt J, Yin H, Shi CX, Que Q, Basu G, Azorsa D, Perkins LM, Braggio E, Fonseca R, Bergsagel PL, Mousses S, Stewart AK. Kinome-wide RNAi studies in human multiple myeloma identify vulnerable kinase targets, including a lymphoid-restricted kinase, GRK6. Blood 2010, 115(8):1594-1604.

160. Tun HW, Marlow LA, von Roemeling CA, Cooper SJ, Kreinest P, Wu K, Luxon BA, Sinha M, Anastasiadis PZ, Copland JA. Pathway signature and cellular differentiation in clear cell renal cell carcinoma. PLoS One 2010, 5(5):e10696.

161. Burchard J, Zhang C, Liu AM, Poon RT, Lee NP, Wong KF, Sham PC, Lam BY, Ferguson MD, Tokiwa G, Smith R, Leeson B, Beard R, Lamb JR, Lim L, Mao M, Dai H, Luk JM. microRNA-122 as a regulator of mitochondrial metabolic gene network in hepatocellular carcinoma. Mol Syst Biol 2010, 6:402.

162. Liu AM, Yao TJ, Wang W, Wong KF, Lee NP, Fan ST, Poon RT, Gao C, Luk JM. Circulating miR-15b and miR-130b in serum as potential markers for detecting hepatocellular carcinoma: a retrospective cohort study. BMJ Open 2012, 2(2):e000825.

163. Chiang DY, Villanueva A, Hoshida Y, Peix J, Newell P, Minguez B, LeBlanc AC, Donovan DJ, Thung SN, Sole M, Tovar V, Alsinet C, Ramos AH, Barretina J, Roayaie S, Schwartz M, Waxman S, Bruix J, Mazzaferro V, Ligon AH *et al*. Focal gains of VEGFA and molecular classification of hepatocellular carcinoma. Cancer Res 2008, 68(16):6779-6788.

164. Fulciniti M, Amin S, Nanjappa P, Rodig S, Prabhala R, Li C, Minvielle S, Tai YT, Tassone P, Avet-Loiseau H, Hideshima T, Anderson KC, Munshi NC. Significant biological role of sp1 transactivation in multiple myeloma. Clin Cancer Res 2011, 17(20):6500-6509.

165. Yang L, Belaguli N, Berger DH. MicroRNA and colorectal cancer. World J Surg 2009, 33(4):638-646.

166. Catto JW, Miah S, Owen HC, Bryant H, Myers K, Dudziec E, Larre S, Milo M, Rehman I, Rosario DJ, Di Martino E, Knowles MA, Meuth M, Harris AL, Hamdy FC. Distinct microRNA alterations characterize high- and low-grade bladder cancer. Cancer Res 2009, 69(21):8472-8481.

167. Park JK, Park SH, So K, Bae IH, Yoo YD, Um HD. ICAM-3 enhances the migratory and invasive potential of human non-small cell lung cancer cells by inducing MMP-2 and MMP-9 via Akt and CREB. Int J Oncol 2010, 36(1):181-192.

168. Blaine SA, Wick M, Dessev C, Nemenoff RA. Induction of cPLA2 in lung epithelial cells and non-small cell lung cancer is mediated by Sp1 and c-Jun. J Biol Chem 2001, 276(46):42737-42743.

169. Bandi N, Zbinden S, Gugger M, Arnold M, Kocher V, Hasan L, Kappeler A, Brunner T, Vassella E. miR-15a and miR-16 are implicated in cell cycle regulation in a Rb-dependent manner and are frequently deleted or down-regulated in non-small cell lung cancer. Cancer Res 2009, 69(13):5553-5559.

170. Li D, Zhao Y, Liu C, Chen X, Qi Y, Jiang Y, Zou C, Zhang X, Liu S, Wang X, Zhao D, Sun Q, Zeng Z, Dress A, Lin MC, Kung HF, Rui H, Liu LZ, Mao F, Jiang BH *et al*. Analysis of MiR-195 and MiR-497 expression, regulation and role in breast cancer. Clin Cancer Res 2011, 17(7):1722-1730.

171. Dickens NJ, Walker BA, Leone PE, Johnson DC, Brito JL, Zeisig A, Jenner MW, Boyd KD, Gonzalez D, Gregory WM, Ross FM, Davies FE, Morgan GJ. Homozygous deletion mapping in myeloma samples identifies genes and an expression signature relevant to pathogenesis and outcome. Clin Cancer Res 2010, 16(6):1856-1864.

172. Imielinski M, Berger AH, Hammerman PS, Hernandez B, Pugh TJ, Hodis E, Cho J, Suh J, Capelletti M, Sivachenko A, Sougnez C, Auclair D, Lawrence MS, Stojanov P, Cibulskis K, Choi K, de Waal L, Sharifnia T, Brooks A, Greulich H *et al*. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell 2012, 150(6):1107-1120.

173. Miecznikowski JC, Wang D, Liu S, Sucheston L, Gold D. Comparative survival analysis of breast cancer microarray studies identifies important prognostic genetic pathways. BMC Cancer 2010, 10:573.

174. Davies FE, Dring AM, Li C, Rawstron AC, Shammas MA, O'Connor SM, Fenton JA, Hideshima T, Chauhan D, Tai IT, Robinson E, Auclair D, Rees K, Gonzalez D, Ashcroft AJ, Dasgupta R, Mitsiades C, Mitsiades N, Chen LB, Wong WH *et al*. Insights into the multistep transformation of MGUS to myeloma using microarray expression analysis. Blood 2003, 102(13):4504-4511.

175. Chen L, Li C, Zhang R, Gao X, Qu X, Zhao M, Qiao C, Xu J, Li J. miR-17-92 cluster microRNAs confers tumorigenicity in multiple myeloma. Cancer Lett 2011, 309(1):62-70.

176. Lionetti M, Biasiolo M, Agnelli L, Todoerti K, Mosca L, Fabris S, Sales G, Deliliers GL, Bicciato S, Lombardi L, Bortoluzzi S, Neri A. Identification of microRNA expression patterns and definition of a microRNA/mRNA regulatory network in distinct molecular groups of multiple myeloma. Blood 2009, 114(25):e20-26.

177. Tian W, Liou HC. RNAi-mediated c-Rel silencing leads to apoptosis of B cell tumor cells and suppresses antigenic immune response in vivo. PLoS One 2009, 4(4):e5028.

178. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A,

Sarkans U, Schulze-Kremer S *et al*. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 2001, 29(4):365-371.

179. Liu F, White JA, Antonescu C, Gusenleitner D, Quackenbush J. GCOD - GeneChip Oncology Database. BMC Bioinformatics 2011, 12:46.

180. Mizuno H, Kitada K, Nakai K, Sarai A. PrognoScan: a new database for meta-analysis of the prognostic value of genes. BMC Med Genomics 2009, 2:18.

181. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C, Schultz N. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2012, 2(5):401-404.

ÖZGEÇMİŞ

Mehmet Kemal Samur, 1983 yılında Konya'da doğdu. İlk ve orta öğrenimini Antalya'da tamamlayarak 2001 yılında Gazi Üniversitesi Teknik Eğitim Fakültesi Bilgisayar Sistemleri Bölümünde üniversite eğitimine başladı. 2005 yılında lisans eğitimini tamamladı ve 2006 yılında Akdeniz Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik ve Tıp Bilişimi Anabilim Dalında yüksek lisans eğitimine başladı. 2008 yılında yüksek lisans eğitimini tamamlayarak aynı anabilim dalında doktora eğitimine başladı. Halen Akdeniz Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik ve Tıp Bilişimi Anabilim Dalında araştırma görevlisi olarak çalışmaktadır ve yabancı dili İngilizce'dir.

E K L E R

Published online 29 May 2012 *Nucleic Acids Research, 2012, Vol. 40, No. 17* ***e135***
doi:10.1093/nar/gks395

Integrative analysis of gene and miRNA expression profiles with transcription factor–miRNA feed-forward loops identifies regulators in human cancers

Zhenyu Yan[1], Parantu K. Shah[1], Samir B. Amin[1], Mehmet K. Samur[1,2], Norman Huang[1], Xujun Wang[3], Vikas Misra[4], Hongbin Ji[5], Dana Gabuzda[4] and Cheng Li[1,*]

[1]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, MA 02215, USA, [2]Department of Biostatistics and Medical Informatics, Akdeniz University, Antalya, Turkey, [3]Department of Bioinformatics, Tongji University, Shanghai, China, [4]Department of Cancer Immunology and AIDS, Dana-Farber Cancer Institute, Boston, MA 02215, USA and [5]State Key Laboratory of Cell Biology, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China

Received September 29, 2011; Revised April 8, 2012; Accepted April 15, 2012

ABSTRACT

We describe here a novel method for integrating gene and miRNA expression profiles in cancer using feed-forward loops (FFLs) consisting of transcription factors (TFs), miRNAs and their common target genes. The dChip-GemiNI (Gene and miRNA Network-based Integration) method statistically ranks computationally predicted FFLs by their explanatory power to account for differential gene and miRNA expression between two biological conditions such as normal and cancer. GemiNI integrates not only gene and miRNA expression data but also computationally derived information about TF–target gene and miRNA–mRNA interactions. Literature validation shows that the integrated modeling of expression data and FFLs better identifies cancer-related TFs and miRNAs compared to existing approaches. We have utilized GemiNI for analyzing six data sets of solid cancers (liver, kidney, prostate, lung and germ cell) and found that top-ranked FFLs account for ~20% of transcriptome changes between normal and cancer. We have identified common FFL regulators across multiple cancer types, such as known FFLs consisting of MYC and miR-15/miR-17 families, and novel FFLs consisting of ARNT, CREB1 and their miRNA partners. The results and analysis web server are available at http://www.canevolve.org/dChip-GemiNi.

INTRODUCTION

A fundamental challenge in cancer systems biology is to identify the regulators of transcriptomic changes during disease progression. These transcriptomic changes are regulated by many different mechanisms including genetic and epigenetic modifications (1). Transcription factors (TFs) and microRNAs (miRNAs) are important regulators at the transcriptional and post-transcriptional levels that modulate transcriptome changes and therefore gene expression in response to cellular environment and signals. Both TFs and miRNAs are known to act as oncogenes or tumor suppressors in human cancers (2–4). Therefore, understanding and utilizing regulatory network information for TFs and miRNAs and their target genes could shed light on altered regulatory genes and pathways in human cancers and suggest novel therapeutic targets. Integrative analysis of both data types is underscored by a recent study showing that destabilization of target mRNAs by miRNA is the predominant mechanism to reduce gene expression, highlighting an essential role of miRNAs in gene regulation (5).

The miRNA-mediated feed-forward loops (FFLs) consisting of TFs and miRNAs are recurrent and important network motifs that form functional modules in the larger regulatory network (6,7). These FFL network motifs consist of a TF, a miRNA and their common target genes (defined as FFL target genes), where the TF regulates the transcription of the miRNA and both the TF and the miRNA regulate a common set of target genes (6–10).

*To whom correspondence should be addressed. Tel: +1 617 632 3498; Fax: +1 617 632 5444; Email: cli@hsph.harvard.edu

The FFLs govern many aspects of normal cell functions and diseases: creating bistable switches of gene expression in developing tissues for spatial avoidance; controlling the time sequence of gene expression to create temporal avoidance; and minimizing expression fluctuation against noise (11). For example, the FFL consisting of c-Myc, miR-17 cluster and E2F1 modulates cellular proliferation in cancer (3,8); the FFL formed by p53 and miR-34a-c promotes cell cycle procession (9); and several FFLs involving miR-7 buffer gene expression against environmental fluctuation during development (10). There are several databases of FFLs for development and cancer (12,13). However, large-scale experimental identification of FFLs and their roles in cancer has not been carried out.

A large amount of genome-wide gene expression and miRNA expression profiles for the same set of samples and covering multiple cancer types are now available from focused efforts of individual laboratories as well as large projects, such as TCGA (14) and ICGC (15). A common theme among currently available integrative analysis approaches is to first identify differentially expressed genes and miRNAs and then look for enriched gene ontology (GO) groups and pathways or miRNA–target gene pairs that are negatively correlated in expression level (16,17). While these methods can generate biological hypotheses that involve single genes or pathways, they do not fully use the genetic network architecture implied by the TF and miRNA regulation. Although researchers have studied FFLs in specific diseases or computationally discover them using genome scans (18–20), the network structures of TF, miRNA and regulated genes have not been used in integrative analysis of gene and miRNA expression data in a systematic manner.

We hypothesized that dysregulation of TF–miRNA FFLs could account for a large proportion of transcriptome changes between normal and disease states such as cancer. Therefore, we investigated the transcriptome changes by looking at gene, TF and miRNA expression profiles in the context of FFL networks. We developed a novel integrative analysis method dChip-GemiNI (Gene and miRNA Network-based Integration), which not only combines gene and miRNA expression profiles available for a disease process, but also incorporates regulatory network structure in the form of computationally identified TF–miRNA FFLs. The utilization of FFLs also provides a principled way to integrate these expression profiles. GemiNI statistically ranks predicted FFLs by their explanatory power to account for differential gene and miRNA expression between two biological conditions such as normal and cancer and assesses their significance using permutation.

We applied dChip-GemiNI to six paired gene and miRNA data sets of human cancers. We identified common miRNAs, TFs and FFLs across cancer types and quantified the proportion of transcriptome changes in cancer, which can be explained by top-ranking FFLs. Validation with systematic literature mining suggested that integrative analysis of expression and FFLs can better predict cancer-related TFs and miRNAs compared with using gene expression data alone, modeling FFLs better identifies cancer-related regulators and FFL-based integrative analysis is more robust. We identified well-known as well as novel FFLs that are common across multiple cancer types or cancer-specific. These top ranked, novel FFLs form experimentally testable hypotheses that regulatory interactions of the involved TFs, miRNAs and their target genes are driving regulators and effectors in one or multiple cancer types.

METHODS

We developed the dChip-GemiNI analysis method to integrate the regulatory network structure with the combined gene and miRNA expression profiling data generated for two biological states, such as normal and cancer. At present, dChip-GemiNI focuses on TF–miRNA FFLs where a miRNA mediates the effect of a TF on their common downstream target genes. The multistep process leading to data integration is outlined in Figure 1.

The first step identifies candidate FFLs from the regulatory relationships between TFs, miRNAs and target genes based on computational prediction of gene targets. In the second step, appropriate cancer data sets with both miRNA and mRNA expression profiles are identified and pre-processed. In the third step, we compute the network motif score (NMS) and the false discovery rate (FDR) for each candidate TF–miRNA FFL. The NMS is a function of multiple scores, including TF and miRNA binding scores to their target sequences, differential expression *P*-values of the FFL components between normal and cancer tissues, and TF and miRNA's target enrichment in differentially expressed genes and miRNAs. The significance of NMS is assessed by their *P*-values based on null distributions obtained by permuting sample group labels and by FDR based on network permutation. Each step is described in detail below.

Construction of candidate TF–miRNA FFL networks

Using the TF–target and the miRNA-target information from motif scanning and prediction databases, we curated candidate TF–gene, TF–miRNA and miRNA–gene regulatory pairs. See Supplementary Figure S1 for the data set statistics. Based on these pairs, we further constructed TF–miRNA–gene networks using matrix representation of regulation targets. The scores of binding affinity were also used in advanced analysis with matrices containing these scores.

TF–target gene relation

We used the tfbsConsSites and tfbsConsFactors data tables from the UCSC Genome Browser that contain the location and score of TF binding sites (TFBSs) conserved in the human/mouse/rat whole genome alignments. The tfbsConsFactors table contains position frequency matrix (PFM) of TFBS motifs from TRANSFAC. We searched for TFBS in the 5-kb promoter region upstream of the transcription start site (TSS) of RefSeq genes. In addition, we performed a human-only TFBS search using TRANSFAC matrices v7.0 and UCSC hg18 genome assembly with the Perl TFBS module (21).

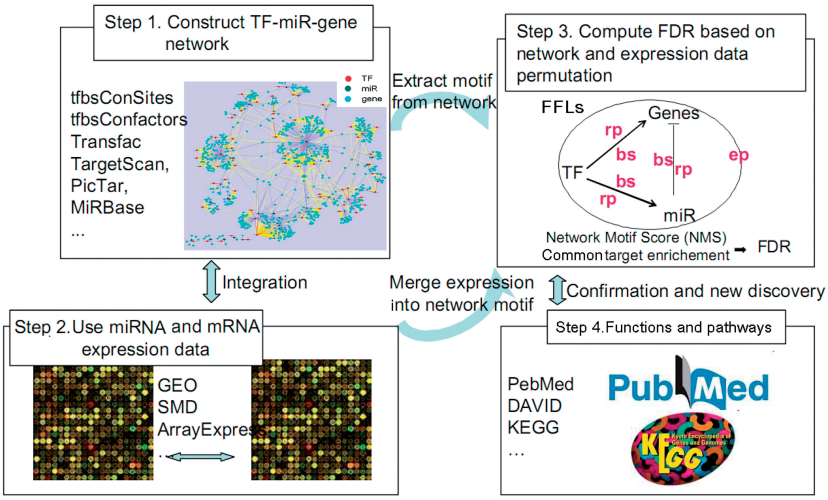


Figure 1. dChip-GemiNI workflow for integrated network and expression analysis. We construct candidate TF–miRNA–gene network using different databases (Step 1), obtain gene and miRNA expression profiles from different studies (Step 2), select significant FFL motifs through integration of network motifs and expression data (Step 3) and further validate the results (Step 4). See Materials and Methods section for details. Step 3 shows a typical TF–miRNA–gene FFL. bs: binding score; rp: enrichment *P*-value; ep: differential expression *P*-value.

The TF matrix accessions were linked to SWISS-PROT IDs and mapped to NCBI gene IDs using BioMart.

TF–target miRNA relation

We downloaded the miRBase (22) pre-miRNA genomic locations from UCSC table browser to locate miRNA promoter regions on the genome. We searched for TFBS in 5-kb upstream of pre-miRNAs, assuming TFs regulate miRNAs in a similar way as they regulate genes (23,24). We considered miRNA clusters and miRNA host genes when making TF–target miRNA predictions. miRNAs tend to form clusters as polycistrons on chromosome and are transcribed together. Moreover, many miRNAs are located within the introns of other genes and are often transcribed together with the host genes. To predict TFBS in these cases, we used the promoter of the first miRNA in 5′ for the miRNAs in a cluster and the promoter of the host gene for the miRNAs in genes.

miR–target gene relation

There are several major miRNA–target gene prediction databases including TargetScan, PicTar and miRanda (25). These predictions have different strengths and weaknesses in terms of specificity and sensitivity (26,27). Analysis of protein-level changes after miRNA knockout and transfection shows that TargetScan and PicTar are more accurate (28,29). We used TargetScan Release 5.1 as the primary database for miRNA target gene prediction, since TargetScan considers sequence conservation and is constantly updated, and PicTar as the supporting database. This is also comparable to the use of tfbsConsSites as the primary database for TF–target gene and target miRNA prediction. We then used matrices to store binary regulatory relationships between TF, miRNA and genes and used matrix operations for efficient search of candidate FFLs and calculation of other statistics such as the number of TF binding sites and total binding affinity scores.

Gene and miRNA expression data sets

We downloaded data sets with both mRNA and miRNA expression profiles with normal and cancer samples for the same cancer type from GEO (30) and ArrayExpress (31). We utilized mRNA and miRNA expression profiles that were either from the same samples (paired) or different samples (unpaired). The data set include liver cancer (hepatocellular carcinoma), kidney cancer (renal cell carcinoma), prostate cancer, testicular cancer (germ cell tumors) and two independent non-small cell lung cancer (NSCLC) cohorts (32–38). See Supplementary Table S1 for details on the data set, including sources, platforms and sample sizes.

Computing differential expression and gene set enrichment

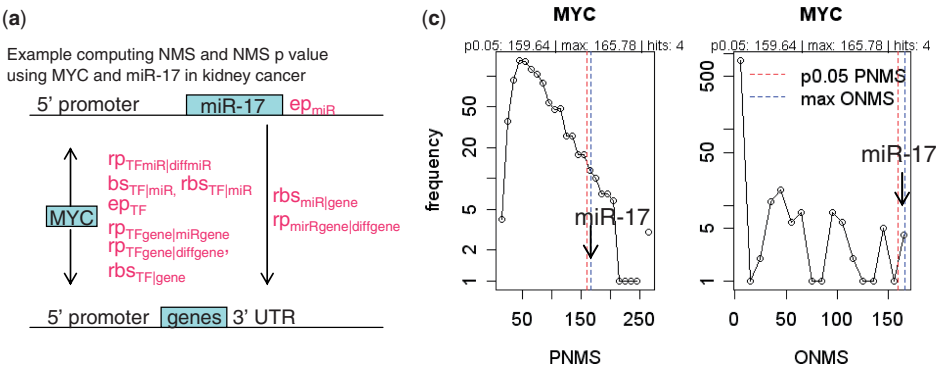
We wrote modules in R-programming language to handle both miRNA and gene expression data with control and disease samples. We first normalized expression data and filtered out low-expression gene or miRNA probes by a specific low-expression cutoff value, so that >50% of the probes were kept, whose expression values are larger than the cutoff value in >50% samples. We followed standard analysis methods to identify differentially expressed miRNAs and genes as well as regulating TFs or miRNAs whose target miRNAs or genes are enriched in differentially expressed gene lists with $P < 0.05$. The *t*-tests or paired *t*-tests were used to identify a gene or miRNA's differential expression between two sample groups

(e.g. normal versus cancer), and the nominal *P*-values were stored as NMS factors for subsequent permutation analysis. We used Fisher's exact test on the frequency tables of TF or miRNA target genes versus up/down/no change genes to compute *P*-values for the enrichment of target genes in differentially expressed genes or miRNAs. Our methods are applicable to any type of miRNA and gene expression data with appropriately formatted data files, sample information and gene information files.

Defining the significance of FFLs altered in cancer

NMS and FDR

We integrated the statistics from expression data analysis with network structure information into a NMS for candidate FFLs. For each candidate, FFL consisting of a TF, a miRNA (miR below) regulated by the TF and their common target genes, we computed its NMS as a product of bs, the binding affinity score from target prediction; ep, the differential *e*xpression *P*-value from the *t*-test of two group of samples; and rp, the en*r*ichment *P*-value from the Fisher's exact test to assess a target gene set's enrichment in differentially expressed genes or another target gene set (Figure 2). The TF–gene, miR-gene and TF–miR links in an FFL are represented by both the binding score bs and target enrichment score rp in the NMS calculation. We used subscripts to indicate the node, link or other information for each scoring factor (Figure 2), e.g. $rp_{TFgene|diffgene}$ refers to the enrichment *P*-value of TF-targeted genes (TFgene) in differentially expressed genes (diffgene) ($ep_{gene} < 0.05$). Other relations, such as TF to miR (TFmiR) and miR to gene (miRgene) are represented in similar way as shown in Figure 2. A variation of bs (binding score), rbs (relative binding score), was defined as the ratio between the median binding score of the differentially expressed targets of a TF or miRNA and the median binding scores of all the targets of the TF or miRNA, and it weights more in the total score if differentially



(b)

Factor	Description	value
$rp_{TFmiR\|diffmiR}$	Enrichment p between MYC target miRs and differential miRs	0.24
$bs_{TF\|miR}$	MYC TFBS score on miR-17 promoter	2.97
$rbs_{TF\|miR}$	Relative MYC TFBS score on miR promoter (differential miRs vs all miRs)	1.00
ep_{TF}	MYC differential p value	+0.015
ep_{miR}	miR-17 differential p value	+0.001
$rp_{TFgene\|miRgene}$	Enrichment between MYC target genes and miR-17 target genes	2.64e-5
$rp_{TFgene\|diffgene}$	Enrichment between MYC target genes and differential genes	+0.042
$rbs_{TF\|gene}$	Relative MYC TFBS score on gene promoter (differential genes vs all genes)	1.00
$rbs_{miR\|gene}$	Relative miR binding score on gene 3' UTR (differential genes vs all genes)	1.00
$rp_{miRgene\|diffgene}$	Enrichment between miR-17 target genes and differential genes	-0.0057
NMS	Network motif score = f(bs, rbs, ep, rp)	
N_P	Number of permutations	1000
PNMS	1000 maxT permuted MYC/miRs NMS using permuted gene/miR expression	
ONMS	Observed MYC/miR-17 NMS using observed gene/miR expression	161.97
$N_{PNMS>ONMS}$	Number of PNMS larger than ONMS	45
NMS p	NMS p value = $(N_{PNMS>ONMS}) / N_P$	0.045
S	Max(NMS p, $rp_{TFgene\|miRgene}$)	0.045
FDR	Compute FDR by using network and expression permutation	0.028

Figure 2. Computation of NMS and FDR with multiple scoring factors. (**a**) Computing NMS for the FFL of MYC and miR-17 using different scoring factors. (**b**) The list of scoring factors and example values and results. The ± signs indicate up/down-regulation of single or enriched genes. (**c**) Left: the distribution of N_P (1000) maximal permutated NMS statistics (PNMS); right: the distribution of the observed NMS from all FFLs involving MYC (ONMS). With 1000 permutations, there are 45 PNMS larger than the observed NMS of MYC-miR-17 FFL, resulting in a permutation $P = 0.045$. The *P*-values for all TF–miRNA FFLs can be computed similarly. The NMS *P*-values are combined with common target enrichment *P*-values during network permutation to further compute FDR. See Materials and Methods section for details.

expressed targets also have higher sequence binding score. The *P*-value components ep and rp are converted from *P*-values linearly as 2 − *P*-value, which ranges between 1 and 2 and gives linear weights to TF and miRNA enrichment or differential expression.

Finally, the NMS of a FFL is defined as the product of two components:

$$\mathrm{NMS} = \mathrm{f(bs,ep,rp)} = \prod(\mathrm{bs_{TF|miR}}, \mathrm{rp_{TFgene|miRgene}}) * \prod(\mathrm{ep_{TF}}, \mathrm{ep_{miR}}, \mathrm{rp_{TFmiR|diffmiR}}, \mathrm{rp_{TFgene|diffgene}}, \mathrm{rp_{miRgene|diffgene}}, \mathrm{rbs_{TF|miR}}, \mathrm{rbs_{TF|gene}}, \mathrm{rbs_{miR|gene}})$$

The first component is based on the candidate FFL network and does not change for different expression data set. It is computed using the TF–miRNA binding score ($\mathrm{bs_{TF|miR}}$), as well as the enrichment *P*-value for TF and miRNA common target genes ($\mathrm{rp_{TFgene|miRgene}}$), which checks whether the TF and miRNA share significant number of target genes by Fisher's exact test. Such enrichment indicates that they are more likely to interact to regulate a set of common genes in specific pathways. The second component of NMS is based on both expression data and the candidate FFL network. It is computed using ep, rp, bs and rbs associated with nodes (genes and miRNAs) and edges (regulatory relationship) in the FFL network.

Assessing the NMS significance by permutation

With the NMS statistics for all candidate FFLs, we then estimated their significance by comparing them with the null distributions formed by randomly permuting the sample group labels of expression data set and the regulatory network structure. We first computed NMS permutation *P*-values by permuting sample group labels. For each TF (MYC as an example in Figure 2c), we computed the distribution of the maximal NMS statistic among all the FFLs involving the TF using the permuted data sets and used this distribution as the MaxT null distribution. The permutation *P*-value for the NMS of a FFL involving this TF and a miRNA is defined as the proportion of permutated maximal NMS statistics (PMNS) for this TF larger than the observed NMS (ONMS): NMS *P*-value = $(\mathrm{N_{PNMS > ONMS}})/\mathrm{Np}$, where Np is the number of permutations. We then computed empirical median FDR values at different *P*-value cutoffs by comparing the number of filtered FFLs in the original and in the permuted regulatory network. Supplementary Figure S2b shows the relationship between FDR and cutoff of the *S* score, defined as Max(TF–miRNA common target enrichment *P*-value, NMS *P*-value) to filter for significant FFLs. The figure shows that FDR decreases as S cutoff decreases. We then converted *S* scores to FDR for all FFLs using data set-specific relation between *S* and FDR.

RESULTS

dChip-GemiNI identifies significant TF–miRNA FFLs altered in human cancers

A detailed example of computing NMS and FDR for a well-known FFL consisting of MYC and miR-17 (8), in kidney cancer, is shown in Figure 2. It obtains a significant NMS $P < 0.05$ and FDR < 0.05 for the kidney cancer data set. We visualize significant FFLs identified by the GemiNI analysis in a summary plot (Figure 3). The plot highlights top TFs STAT1, MYC and USF1 (Figure 3a) and top miRNAs miR-15a, miR-16 and miR-20b (Figure 3d) as the regulators forming FFLs that account for a large proportion of transcriptome changes in kidney cancer. The plot also allows for a quick identification of other significant FFLs as well as involved TFs and miRNAs (Figure 3c), helping construct specific biological hypotheses on their expression changes and roles in a cancer type. Other examples of significant FFLs are available in the summary plots of Supplementary File S1.

Both NMS *P*-value and TF–miRNA common target enrichment *P*-value are essential in selecting significant FFLs altered between normal and cancer. Venn diagrams (Supplementary Figure 7a and b) show that thresholding TF–miRNA common target enrichment *P*-value alone gives rise to more than 4000 significant FFLs (enrichment $P < 0.05$, FDR < 0.1 compared to network permutation). In contrast, thresholding NMS *P*-value gives much focused FFLs (approximately 200). Using NMS *P*-value and TF–miRNA common enrichment *P*-value together (as *S* score = max of the two *P*-values) will call significant FFLs as their intersection: around two-third of the FFLs from the NMS *P*-value cutoff and a very small portion of FFLs from the enrichment *P*-value cutoff. Although FFLs with significant enrichment of TF–miRNA common targets are independent of expression data set, the S-filtered FFLs depend on expression data and cancer-specific alteration of FFLs (Table 1).

We applied dChip-GemiNI to six cancer data sets with both gene and miRNA expression data in normal and cancer samples and identified significant FFLs. They include data sets for liver, kidney, prostate, germ cell tumors and two independent data sets of non-small cell lung cancer (Supplementary Table S1). We also included tables of significant FFLs and their common target genes in Supplementary File S2 for the kidney cancer data. The FFLs with significant FDR were used in the downstream analysis and literature and computational validation.

Validation of significant FFLs for their roles in cancer

Our literature search identified only six experimentally validated miRNA-mediated FFLs in human cancers (Supplementary Table S6), of which dChip-GemiNI identifies one FFL (E2F1 with miRNA-106a/93/25) as significant (FDR < 0.1) in lung and liver cancers. Our procedure missed some validated FFLs due to the lack of known TF motifs in our data set. Since very few experimentally validated FFLs are known in cancer, we validated the significant FFLs in several innovative ways in the absence of any gold standard set of FFLs in cancer.

We utilized the TransmiR database (39) that curates TF -miRNA regulation information from the literature as partial evidence for FFLs. We found evidence supporting multiple significant FFLs (FDR < 0.1) formed by TFs MYC, MYCN, E2F1, E2F3, SP1, EGR1 and STAT3 with several miRNAs (see Supplementary Table S2).

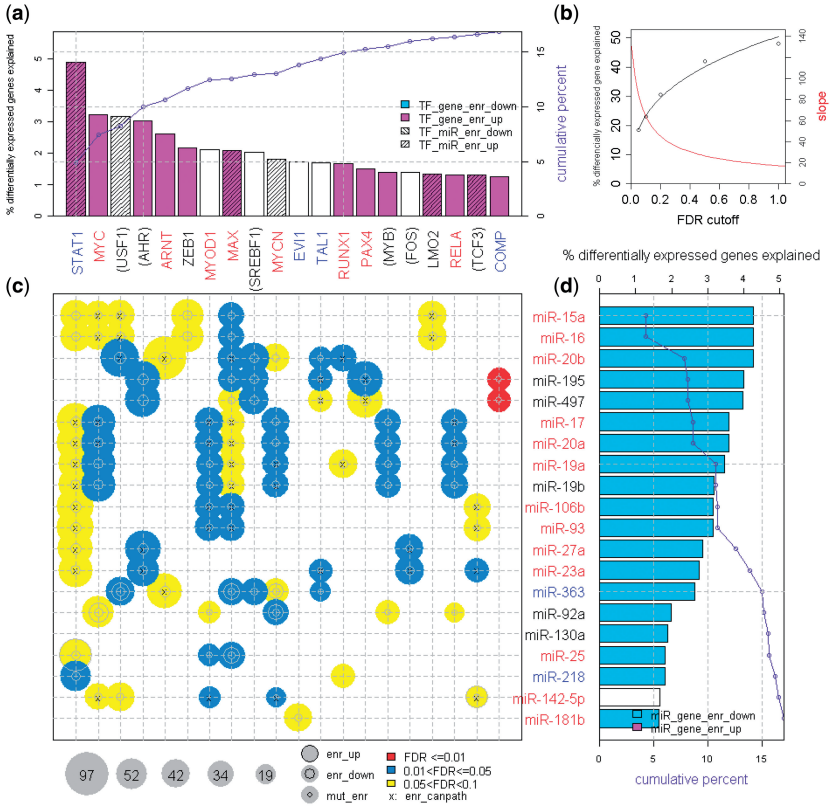


Figure 3. The summary bubble-bar plot from dChip-GemiNI analysis of the kidney cancer data, highlighting the amount of differentially expressed genes explained by FFLs, TFs and miRNAs. TFs on the top (**a**) and miRNAs on the right (**d**) are ranked by the percentage of differentially expressed genes explained by all the significant FFLs involving a TF or miRNA (the height of bars). The top 20 TFs and miRNAs are displayed (for complete plots, see Supplementary File S1). The bubbles in the lower left panel (**c**) correspond to TF–miRNA FFLs with FDR <0.1 from the kidney cancer data. The bubble size indicates the number of differentially expressed FFL target genes, and color indicates the FFL significance. The cumulative percentage of differentially expressed genes explained by TFs and miRNAs is also shown in the barplots (**a** and **d**). The parenthesis around a TF or miRNA name indicates that its differential expression is not used in NMS due to missing data or low-expression filtering (other scoring factors are still used). TF and miRNA name color: red, expression is upregulated in cancer; blue, downregulated; black, no change. The following target enrichment analyses are called at Fisher's exact test $P < 0.05$. enr_up (enr_down): TF and miRNA common target genes are enriched in upregulated (downregulated) genes; TF_gene_enr_up (TF_gene_enr_down): TF target genes are enriched in upregulated (downregulated) genes; TF_miR_enr_up (TF_miR_enr_down): TF target miRNAs are enriched in upregulated (downregulated) miRNAs; miR_gene_enr_up (miR_gene_enr_down): miRNA target genes are enriched in upregulated (downregulated) genes; Mut_enr: TF target genes and miRNA target genes are enriched in each other; enr_canpath: TF and miRNA common and differentially expressed target genes are enriched in KEGG cancer pathways.

Next, we hypothesized that even when the functional role of a particular FFL is not known in a given cancer type, both the involved TF and miRNA could have been studied together, or functional roles of the constituent TF and miRNA may have been characterized extensively, if the involved TFs and miRNAs play important roles in multiple cancer types. Therefore, we performed automatic literature mining of PubMed abstracts for combined keywords of a TF or miRNA gene name or aliases and cancer-related terms ('neoplasms', 'cancer', 'tumor', 'carcinoma', 'oncogenes'). The general cancer terms instead of specific cancer type names were used in the search to relate the literature results to top-ranked TFs and miRNAs across multiple cancer types. We found that 3.2% of all significant FFLs (FDR < 0.1) from kidney cancer co-mentioned in at least one publication, compared to only 1.2% of all TF–miRNA pairs co-mentioned. Thus, there is a 2.67-fold enrichment in literature co-occurrence for significant FFLs. We also found that there is a 1.51-fold enrichment in literature co-occurrence for significant FFLs identified by TF–miRNA common target enrichment *P*-value only without using expression data, compared to

Table 1. The top-ranked TFs and miRNAs involved in TF–miRNA FFLs across six cancer data sets

	liver_cancer GSE22058	kidney_cancer GSE16441	prostate_cancer GSE21032	germ_cell_tumor GSE18155	lung_cancer GSE18805	lung_cancer GSE2088_ E-TABM-22	Average rank	Number of presence
TFs								
USF1	2	3	5	9	8	2	4.83	6
ARNT	3	5	6	9	6	1	5.00	6
MYC	4	2	3	9	21	7	7.67	5
MAX	18	8	8	9	14	6	10.50	6
AHR	6	4	14	9	5	41	13.17	5
CREB1	5	71	2	1	1	5	14.17	5
IKZF1	9	29	21	9	14	3	14.17	4
MYCN	41	10	11	9	20	19	18.33	4
TFAP2C	29	25	11	9	20	20	19.00	2
SREBF1	16	9	52	9	26	8	20.00	4
MYCN	39	8	15	43	32	8	24.17	3
FOSB	17	39	27	42	18	21	27.33	2
TFAP2C	43	38	16	7	29	38	28.50	2
NR3C1	24	26	51	5	39	26	28.50	1
SREBF1	15	17	62	47	20	12	28.83	3
TFAP2B	26	35	17	10	40	48	29.33	2
PPARG	19	37	62	42	12	6	29.67	3
MYB	12	21	58	47	10	32	30.00	2
MZF1	70	43	13	29	10	15	30.00	3
JUND	17	33	50	42	18	21	30.17	2
miRNAs								
hsa-miR-17	3	5	5	1	4	3	3.50	6
hsa-miR-16	3	1	5	9	3	1	3.67	6
hsa-miR-195	2	3	2	9	2	4	3.67	6
hsa-miR-15a	1	1	3	9	1	8	3.83	6
hsa-miR-20a	3	5	5	9	4	6	5.33	6
hsa-miR-130a	10	13	14	3	15	7	10.33	6
hsa-miR-27a	5	9	9	9	8	26	11.00	5
hsa-miR-19a	11	6	6	9	6	33	11.83	5
hsa-miR-497	24	4	1	9	2	33	12.17	4
hsa-miR-106b	6	8	11	9	16	33	13.83	5
hsa-miR-106a	4	35	6	38	7	1	15.17	4
hsa-miR-19a	8	7	4	38	6	28	15.17	4
hsa-miR-20b	36	2	6	38	7	8	16.17	4
hsa-miR-23a	33	8	9	26	10	29	19.17	3
hsa-miR-106b	7	9	12	38	14	51	21.83	4
hsa-miR-93	7	9	12	38	14	51	21.83	4
hsa-miR-124	12	18	31	38	23	10	22.00	3
hsa-miR-424	6	19	52	17	11	31	22.67	4
hsa-miR-92a	10	14	7	4	68	34	22.83	4
hsa-miR-15b	19	15	11	38	13	51	24.50	4

In each cancer data set, TFs and miRNAs are ranked by the percentage of differentially expressed (DE) genes explained by all the significant FFLs (FDR < 0.1) involving a TF or miRNA (similar to Figure 3). The tables order TFs and miRNAs by their average ranks in the six cancer data sets. In a data set, the TFs or miRNAs with the same percentage of FFL-explained DE genes are ranked the same. Analysis of germ cell tumor results in only a few TFs and miRNAs associated with significant FFLs, and we rank the rest TFs and miRNAs at the same low rank order. Number of presence: the number of times that a TF or miRNA is present in the top 20 of individual data sets. See Supplementary File S4 for more information.

a 2.17-fold enrichment for FFLs identified using NMS *P*-value only. This supports the value of using expression data through NMS in selecting cancer-specific FFLs.

We also counted the number of cancer-related papers for each TF and miRNA (Figure 4a and Supplementary File S3). We found several TFs and miRNAs which have been studied in a large number of cancer publications due to their important and diverse roles in cancer: TP53 (more than 50 000 papers), NFKB1 (more than 30 000), ESR1 and MYC (approximately 20 000), miR-21 (more than 300 papers), miR-17 (approximately 200) and miR-16 (more than 100). The median value of the cancer-related papers is 233 for all the TFs and 9 for all the miRNAs.

We expected that the involvement of miRNAs and TFs in cancer would be reflected by both the number of cancer studies in the literature and their ranking in expression data analysis. We first assessed the correlation between TFs and miRNAs' differential expression ranking and their occurrence in cancer-related PubMed abstracts. We ranked miRNAs and TFs by their differential expression *P*-values in each cancer data set and computed their average ranks in the six cancer data sets (Figure 4b and Supplementary File 4). We found a strong correlation between the number of cancer-related papers and the average rank of each miRNA (correlation = −0.72, Figure 4c; higher rank is associated with more papers)

(a)

miRNA	number of papers	oncomiRs
miR-21	322	*#
miR-17	195	*#
miR-155	178	*#
miR-221	127	*#
miR-34a	126	*
miR-16	112	*#
miR-145	96	*#
miR-146a	89	*#
miR-125b	84	*#
let-7a	83	*#

(b)

miRNA	Average Rank	OncomiRs
miR-21	32.50	*#
miR-142-5p	51.83	*
miR-130a	53.00	
miR-96	53.83	*
miR-146a	58.17	*#
miR-145	60.67	*#
miR-17	61.17	*#
miR-30c	62.83	#
miR-106b	66.83	#
miR-25	68.67	#

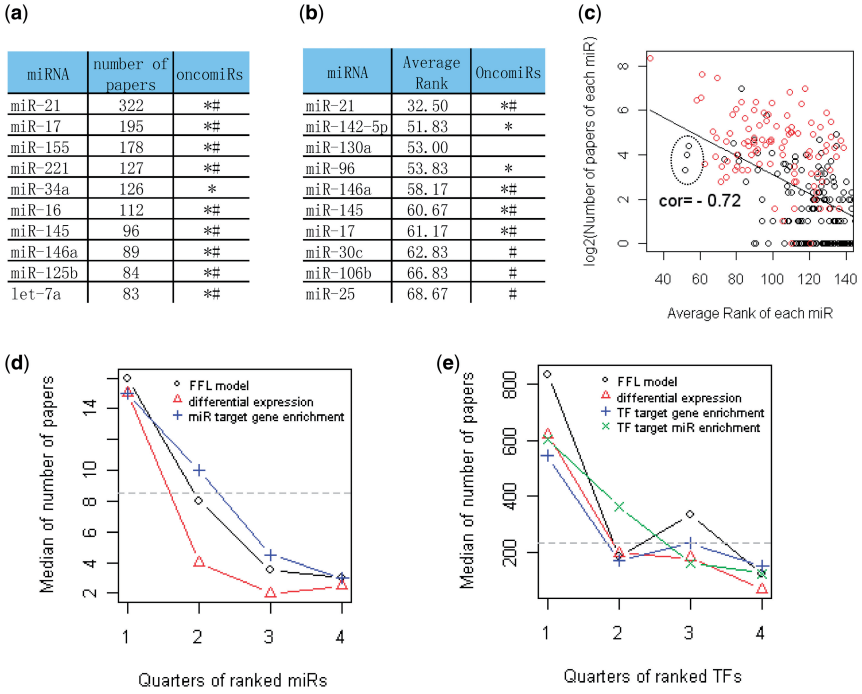


Figure 4. Validation of top TFs and miRNAs ranked by different analysis methods. (**a**) The top 10 miRNAs based on the number of cancer-related papers from literature mining. Asterisks indicate oncomiRs summarized in reviews (2,3,40,41); Hash indicate oncomiRs from the SBI OncomiR collection (http://www.systembio.com/services/microrna/oncomir-collection). See Supplementary Files S3 and S7 for details. (**b**) The top 10 miRNAs by the average rank of differential expression *P*-values from the six cancer data sets. See Supplementary File S4 for the complete list. (**c**) The scatter plot between the number of cancer-related papers and the average rank of miRNAs by differential expression. Higher rank (lower rank number) is correlated with more papers, so the correlation is negative. The red circles are known oncomiRs from SBI. The top-ranked miRNAs could also suggest novel oncomiR candidates, such as the three leftmost black points corresponding to miR-142-5p, miR-130a and miR-96 (indicated in the large circle). (**d**) The miRNAs are first ordered by their average ranks in the six cancer data sets using each analysis method and then divided into four quarters (*X*-axis). Then, the median number of cancer-related papers of the miRNAs in each quarter is plotted (*Y*-axis) for different methods. The gray horizontal line is the median number of papers for all miRNAs. (**e**) The similar plot as (d) for TFs.

and a moderate correlation for TFs (−0.30). The correlations are smaller when using the ranks in individual cancer data sets compared to using the average ranks across the six cancer data sets. The high correlation for miRNAs suggests that the miRNA's differential expression ranks averaged across cancer types can identify miRNAs important in cancer.

Next, we compared different analysis methods of ranking cancer-related TFs and miRNAs based on expression data, including differential expression analysis, target enrichment of differentially expressed genes and miRNAs and the FFL model. For differential expression and target enrichment analysis, we ranked TFs and miRNAs by *P*-values. For the FFL model, TFs and miRNAs were ranked by the number of differentially expressed FFL target genes in their associated FFLs with FDR < 0.1. We divided the miRNAs and TFs into four quarters based on their average ranks in each method (lower quarters for high-ranked miRNAs and TFs) and then plotted the median number of papers in each quarter. All the methods are able to associate high-ranked miRNAs and TFs with more cancer-related papers (Figure 4d and e). Although the miRNA differential expression method (triangles) better distinguishes the first quarter from the other quarters, the FFL model performs slightly better in selecting cancer-related miRNAs in the first quarter of ranked miRNAs (Figure 4d). Moreover, the FFL model performs significantly better than the other methods in ranking cancer-related TFs in the first quarter (Figure 4e). In summary, the FFL model performs comparably or better in identifying top cancer-related miRNAs or TFs compared to existing approaches.

FFL-based integrative analysis is more robust

We found that for the same biological process or disease, results from the FFL model are more comparable between data sets than those from traditional methods such as

differential expression or TF/miRNA target enrichment. For example, the correlation between TF or miRNA ranks of the two lung cancer data sets is the highest using the FFL model (Supplementary Table S3). In addition, the FFL analysis identifies the involvement of the miR-15/miR-17 families in both lung cancer data sets (Supplementary Figure S3), which is less prominent from traditional analysis. The results described here and in the previous section confirm the robustness of integrative analysis of gene and miRNA expression data with network motifs.

This is important as the number of differentially expressed genes, target-enriched TFs or miRNAs vary across the cancer expression data sets, possibly due to their cancer type, sample size and platform characteristics (Supplementary Table S4 and Supplementary File S4). The different data sets for the same cancer type can also result in differences. For instance, the gene and miRNA expression data are paired in lung cancer data set 1 (GSE18805), while they are from two studies (GSE2088 and E-TABM-22) in lung cancer data set 2. The lung cancer data set 2 has more samples, fewer differentially expressed genes/TFs and miRNAs, but more miRNAs with enriched target genes than the lung cancer data set 1 (Supplementary Tables S1 and S4). For some genes or miRNAs, such as the TF ARNT, the direction of expression change is opposite between the two lung cancer data sets (Supplementary File S4). In general, more differentially expressed or target-enriched genes and miRNAs lead to more TFs and miRNAs identified by significant FFLs (Supplementary Table S4 and Supplementary Figure S4).

Significant TF–miRNA FFLs explain ~20% of transcriptome changes in human cancers

To estimate transcriptome changes between normal and cancer that are likely caused by significant FFLs, we identified the target genes of all the significant FFLs in a data set and computed the ratio of differentially expressed FFL target genes to all differentially expressed genes. In the kidney cancer data, the FDR cutoffs of 0.05 and 0.1, respectively, identify approximately 230 and approximately 430 significant FFLs, which likely cause 18 and 23% of expression changes (Figure 3b). In addition, the FFL target genes of the top 20 TFs and top 20 miRNAs, ranked by the percentage of expression changes explained by all the significant FFLs associated with a TF or miRNA, account for ~17% of the differentially expressed genes in the kidney cancer data (the cumulative blue curves in Figure 3a and d).

For all the cancer data sets, there are on average 130 and 270 significant FFLs at the FDR cutoffs of 0.05 and 0.1, respectively, which explain on average 13 and 19% of expression changes in a data set. The FFL target genes of the top 20 TFs and top 20 miRNAs account for ~15% of the differentially expressed genes on average across the data sets. Interestingly, FFLs with FDR <0.1 can explain 20–23% of the expression changes for most of the cancers, but only 3% of expression changes for germ cell tumor, possibly because the small sample size of this data set leads to much fewer significant FFLs (Supplementary Table S1).

In comparison, we found that on average significant TFs or miRNAs explain 50–70% of the expression changes when using TF–target gene or miRNA–target gene enrichment analysis. This can be due to the fact that each TF or miRNA usually has hundreds of target genes, but the number of target genes of an FFL (defined as the common targets of the involved TF and miRNA) is usually less than 100. In effect, the FFL model narrows down the list of potential target genes of candidate regulators for further validation.

Meta-analysis of multiple cancer types identifies common and cancer-specific FFLs

We identified significant FFLs and involved TFs and miRNAs that are common and distinct across multiple cancer types at the FDR threshold of 0.1. Top-ranked FFLs that are common across the six cancer data sets consist of TFs USF1, ARNT, MYC, MAX, AHR, CREB1 and miRNAs mir-15a, miR-16, miR-17, miR-20a and miR-195 (Table1; see Supplementary File S4 for the complete list). Most of the top miRNAs belong to two families of miRNA clusters, the miR-15/16/195/424/497 family and miR-17-5p/20/93.mr/106/519.d family, located on chromosomes 3, 7, 13, 17 and X (see Supplementary Table S5 for details). These miRNA clusters are potentially co-regulated by dozens of TFs including ARNT, MYC and CREB1. Supplementary Figure S5 shows the regulatory relationships among them.

The miR-15 and miR-17 families have been experimentally verified to be oncomiRs and have tumor suppressor or oncogene functions in cancer [Supplementary Figure S6 (2,3,40,41)]. In our analysis, the expression levels of most miR-15 and miR-17 family members increase in most cancer types relative to normal, and there is clear expression correlation in the miRNA clusters (Supplementary File S5). miRNA clusters and miRNA families are widely distributed in the genome: there are over 200 miRNAs forming over 60 physical clusters on chromosomes and over 200 miRNAs belonging to approximately 100 miRNA families. Therefore, in our view, the identification of miR-15/17 cluster family is likely not due to the bias introduced by computational methods for target identification (e.g. the entire miRNA family sharing largely overlapping target genes).

The target gene functions of top-ranked FFLs are diverse and often enriched of cancer-related GO terms and pathways. For example, significant GO terms in the differentially expressed target genes of the MYC/miR-17 FFL in kidney cancer include transcriptional regulation, apoptosis, cell cycle arrest, RAS signaling and cytokine production (Table 2). Significant GO terms for other selected FFLs involving MYC, CREB1, SP1 and mir-15/17 in several cancer types are listed in Supplementary File S6.

We also identified cancer type-specific FFLs and associated top TFs and miRNAs (Table 1 and Supplementary File S4). For example, OCT1/POU2F1 ranks as number 4 in lung cancer data 1 but ranks lower in other cancers; STAT1 ranks as number 1 in kidney cancer, while

its average rank across the data sets is not high. For miRNAs, miR-548d-3p ranks within top three in germ cell tumor but lower in other cancers. These results generate novel hypotheses to study cancer type-specific functions of top-ranked FFLs, TFs and miRNAs and can be queried at http://www.canevolve.org/dChip-GemiNi.

DISCUSSION

Transcriptome changes are an important mediator in complex diseases such as cancers, which have multiple biological hallmarks (42). Among many genetic regulators, TFs and miRNAs frequently form FFLs and other network motifs to regulate gene expression in a combinatorial manner. Such regulation may be disrupted in cancer due to mutations and chromosome abnormalities. Success in treating cancer patients calls for better understanding of common and cancer type-specific regulatory processes in cancer cells.

Here, we have developed dChip-GemiNI, a novel analysis approach integrating FFL network structures with gene and miRNA expression data, and identified significant TF–miRNA FFLs that likely cause ~20% of the transcriptome changes. Conventional analysis methods focus on differentially expressed genes and miRNAs between biological processes or disease states, but expression changes may not be detectable for many TFs and miRNAs, whose expression level can be low and their effects can be exerted by other means such as protein phosphorylation or localization. However, these effects may be probed through changes in the network of the regulated genes. Gene set and pathway enrichment analyses method partially utilize such information via TF or miRNA target gene sets. dChip-GemiNI overcomes limitations from traditional analysis methods by combining multiple data sources to rank network motifs such as FFLs by their explanatory power to account for transcriptomic changes. We have shown that this FFL-based modeling can identify TF and miRNA regulatory pathways commonly altered across cancer types and do so more consistently than existing analysis approaches. Our results confirm the roles of miRNA-mediated pathways in human cancers and generate novel biological hypotheses on specific TF–miRNA FFLs as altered regulators.

TF–miRNA FFLs may be common drivers of pathogenesis across multiple cancer types

We have identified common TFs and miRNAs associated with significant FFLs across cancer types. For example, MYC, ARNT and USF1 are ranked within the top 20 TFs in all the cancer types except the germ cell tumor (Table1). MYC has multiple roles in cancer and is known to involve in miRNA regulation (3,41,43–45). MYC mutation or dysregulation can cause a cascade of expression changes of miRNAs and genes and lead to abnormal cell cycle and apoptosis functions (Supplementary Figure S6). Our study

Table 2. Significantly enriched GO terms in the differentially expressed target genes of the MYC and miR-17 FFL in kidney cancer

Category	Term	Count	*P*-value	FDR %
GOTERM_BP_FAT	GO:0006350~transcription	20	6.75E-05	0.098527
GOTERM_BP_FAT	GO:0045449~regulation of transcription	20	0.0011513	1.667907
GOTERM_BP_FAT	GO:0016265~death	9	0.0041554	5.898623
GOTERM_BP_FAT	GO:0006917~induction of apoptosis	6	0.0058396	8.195772
GOTERM_BP_FAT	GO:0012502~induction of programmed cell death	6	0.005916	8.298779
GOTERM_BP_FAT	GO:0007050~cell cycle arrest	4	0.0061271	8.582609
GOTERM_BP_FAT	GO:0046578~regulation of Ras protein signal transduction	5	0.0069286	9.653135
GOTERM_BP_FAT	GO:0001816~cytokine production	3	0.0120185	16.18359
GOTERM_BP_FAT	GO:0051056~regulation of small GTPase-mediated signal transduction	5	0.0129273	17.30217
GOTERM_BP_FAT	GO:0008219~cell death	8	0.0142681	18.92731
GOTERM_BP_FAT	GO:0006357~regulation of transcription from RNA polymerase II promoter	8	0.0150876	19.90583
GOTERM_MF_FAT	GO:0005083~small GTPase regulator activity	6	0.0036058	4.2172
GOTERM_MF_FAT	GO:0030695~GTPase regulator activity	7	0.0039461	4.606721
GOTERM_MF_FAT	GO:0060589~nucleoside–triphosphatase regulator activity	7	0.0043937	5.116735
GOTERM_MF_FAT	GO:0030528~transcription regulator activity	13	0.0090264	10.25106
GOTERM_MF_FAT	GO:0003677~DNA binding	17	0.0094193	10.67461
GOTERM_MF_FAT	GO:0003700~transcription factor activity	10	0.0098228	11.10768
INTERPRO	IPR003070:Orphan nuclear receptor	2	0.0096936	11.07411
INTERPRO	IPR001331:Guanine–nucleotide dissociation stimulator, CDC24, conserved site	3	0.0177529	19.4124
KEGG_PATHWAY	hsa05214:Glioma	3	0.0142992	12.54349
PIR_SUPERFAMILY	PIRSF002524:nerve growth factor IB-like nuclear receptor	2	0.0052646	3.816163
SP_PIR_KEYWORDS	Transcription regulation	20	9.40E-06	0.010708
SP_PIR_KEYWORDS	Transcription	20	1.29E-05	0.014701
SP_PIR_KEYWORDS	Phosphoprotein	37	0.0006004	0.681797
SP_PIR_KEYWORDS	Nucleus	26	0.0007915	0.897858
SP_PIR_KEYWORDS	DNA binding	15	0.0017858	2.015405
SP_PIR_KEYWORDS	Guanine–nucleotide releasing factor	4	0.0067022	7.373974
SP_PIR_KEYWORDS	transcription factor	3	0.0181031	18.78729
SP_PIR_KEYWORDS	Zinc finger	12	0.0181405	18.82247
UP_SEQ_FEATURE	Compositionally biased region:Ser-rich	8	0.0003965	0.50712

GO terms are filtered by FDR < 0.2 and GO analysis are performed by the DAVID functional annotation Web server.

confirms that FFL networks involving MYC and the miR-15 and miR-17 families are potential common drivers of malignant progression across cancer types. We show that MYC regulates multiple top-ranked miRNAs associated with significant FFLs, including the miRNA clusters in the miR-15/16/195/424/497 and miR-17-5p/20/93.mr/106/519.d families (Supplementary Figure S5).

In addition, our analysis ranks ARNT (aryl hydrocarbon receptor nuclear translocator) as number 2 in prostate cancer and number 4 in liver cancer (Table 1). ARNT has recently been shown to associate with tumor growth and progression of liver cancer (46). Our analysis ranks CREB1 (cAMP responsive element binding protein 1) as one of the top TFs by average rank and number 1 in the lung cancer data set 1 and germ cell tumor. Recent studies have shown that CREB is involved in tumor initiation, progression and metastasis (47), and it has growth suppression or cancer inhibition function in lung and prostate cancers (48,49). However, CREB's regulation of miRNAs in cancer has not been well studied and our results provide a new hypothesis of CREB-miRNA network's alteration in cancer and potential miRNA partners.

Modeling FFLs better identifies potential cancer regulatory network

We have shown that modeling TF–miRNA FFLs helps to better identify cancer-related TFs, as suggested by the higher number of cancer-related papers in top-ranked TFs compared with the methods of differential expression and target enrichment. TFs' mRNA expression level is usually low, their functions are frequently altered at the protein interaction or modification level and they often form regulatory modules. These factors hinder simpler approaches to study TF functions in cancer using gene expression data. In evaluating TFs, the FFL model integrates information from both mRNA and miRNA expression through the network structure of target genes and improves the cancer relevance of the TF ranking. For example, MYC and CREB1 are not among the top TFs according to differential expression or target enrichment analysis (Supplementary File S4), but through FFLs they are connected to oncomiRs in the miR-15 and miR-17 families, many of which are top-ranked miRNAs by differential expression or miRNA target gene enrichment across cancer data sets. Therefore, in the FFL model, MYC and CREB1 are ranked high as common cancer-related TFs because their associated significant FFLs can explain a high percentage of expression changes.

Future improvement of network and expression data integration

Although the analysis results from dChip-GemiNI shows consistency and specificity across multiple data sets, there are areas of improvement in follow-up work. First, we can improve the target prediction and candidate regulatory network by including gene targets identified from genome-wide profiling of TFs using chromatin immunoprecipitation followed by microarray hybridization or next-generation sequencing. Second, the proposed methodology is not limited to one type of network motifs, such as FFL. We can study other common network motifs, such as feedback loops consisting of one miRNA and one TF, which often form bistable switches in development and differentiation. Third, we may consider correlations between TF, miRNA and gene expression when each sample has both miRNA and gene expression data. If such correlations are available for both control and disease samples, correlation changes between control and disease can be incorporated into the scoring of network motif models to identify altered regulatory network. As genomics studies of cancer generate more large data sets of multiple data types, integrative analysis methods will play ever-growing roles in understanding the causes of cancer and revealing novel drug targets.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–6, Supplementary Figures 1–7 and Supplementary Files 1–8.

ACKNOWLEDGEMENTS

We thank Curtis Huttenhower, Nikhil C. Munshi, Tao Lu, Bruce Yankner, Yu Zhang, Andrea Richardson, Yong Zhang, Meng Wang, Zhen Zhou and members of Cheng Li Lab for constructive discussion. Finally, we are grateful to the anonymous reviewers for their critical insights that have improved the quality of this work. Z.Y., P.K.S. and C.L. designed the study and wrote the manuscript. Z.Y. carried out the data collection and analysis and prepared the figures. M.K.S. constructed the web server. S.B.A., N.H. and X.W. contributed to data collection. V.M., D.G. and H.J. contributed to the biological interpretation. All authors discussed the methods and results and approved the manuscript.

FUNDING

National Institutes of Health [R01 GM077122 to Z.Y., P.K.S. and C.L., DP1 DA28994 to D.G. and V.M.]; Dana Foundation (to Z.Y.). Claudia Adams Barr Program in Innovative Basic Cancer Research and the Multiple Myeloma Career Development award (to P.K.S.); Claudia Adams Barr Program in Innovative Basic Cancer Research (to D.G. and V.M.); H.J. is a scholar of the Hundred Talents Program of the Chinese Academy of Sciences. Funding for open access charge: NIH [R01 GM077122].

Conflict of interest statement. None declared.

REFERENCES

1. Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.
2. Croce,C.M. (2009) Causes and consequences of microRNA dysregulation in cancer. *Nat. Rev. Genet.*, **10**, 704–714.

3. Esquela-Kerscher,A. and Slack,F.J. (2006) Oncomirs—microRNAs with a role in cancer. *Nat. Rev. Cancer*, **6**, 259–269.
4. Calin,G.A., Sevignani,C., Dumitru,C.D., Hyslop,T., Noch,E., Yendamuri,S., Shimizu,M., Rattan,S., Bullrich,F., Negrini,M. *et al.* (2004) Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc. Natl Acad. Sci. USA.*, **101**, 2999–3004.
5. Guo,H., Ingolia,N.T., Weissman,J.S. and Bartel,D.P. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, **466**, 835–840.
6. Mangan,S. and Alon,U. (2003) Structure and function of the feed-forward loop network motif. *Proc. Natl Acad. Sci. USA.*, **100**, 11980–11985.
7. Tsang,J., Zhu,J. and van Oudenaarden,A. (2007) MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. *Mol. Cell*, **26**, 753–767.
8. O'Donnell,K.A., Wentzel,E.A., Zeller,K.I., Dang,C.V. and Mendell,J.T. (2005) c-Myc-regulated microRNAs modulate E2F1 expression. *Nature*, **435**, 839–843.
9. He,L., He,X., Lim,L.P., de Stanchina,E., Xuan,Z., Liang,Y., Xue,W., Zender,L., Magnus,J., Ridzon,D. *et al.* (2007) A microRNA component of the p53 tumour suppressor network. *Nature*, **447**, 1130–1134.
10. Li,X., Cassidy,J.J., Reinke,C.A., Fischboeck,S. and Carthew,R.W. (2009) A microRNA imparts robustness against environmental fluctuation during development. *Cell*, **137**, 273–282.
11. Shalgi,R., Brosh,R., Oren,M., Pilpel,Y. and Rotter,V. (2009) Coupling transcriptional and post-transcriptional miRNA regulation in the control of cell fate. *Aging*, **1**, 762–770.
12. El Baroudi,M., Cora,D., Bosia,C., Osella,M. and Caselle,M. (2011) A curated database of miRNA mediated feed-forward loops involving MYC as master regulator. *PLoS One*, **6**, e14742.
13. Friard,O., Re,A., Taverna,D., De Bortoli,M. and Cora,D. (2010) CircuitsDB: a database of mixed microRNA/transcription factor feed-forward regulatory circuits in human and mouse. *BMC Bioinformatics*, **11**, 435.
14. McLendon,R., Friedman,A., Bigner,D., Van Meir,E.G., Brat,D.J., Mastrogianakis,G.M., Olson,J.J., Mikkelsen,T., Lehman,N., Aldape,K. *et al.* (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.
15. Hudson,T.J., Anderson,W., Artez,A., Barker,A.D., Bell,C., Bernabe,R.R., Bhan,M.K., Calvo,F., Eerola,I., Gerhard,D.S. *et al.* (2010) International network of cancer genome projects. *Nature*, **464**, 993–998.
16. Fulci,V., Colombo,T., Chiaretti,S., Messina,M., Citarella,F., Tavolaro,S., Guarini,A., Foa,R. and Macino,G. (2009) Characterization of B- and T-lineage acute lymphoblastic leukemia by integrated analysis of MicroRNA and mRNA expression profiles. *Genes Chromosomes Cancer*, **48**, 1069–1082.
17. Qin,L.X. (2008) An integrative analysis of microRNA and mRNA expression—a case study. *Cancer Inform.*, **6**, 369–379.
18. Guo,A.Y., Sun,J., Jia,P. and Zhao,Z. A novel microRNA and transcription factor mediated regulatory network in schizophrenia. *BMC Syst. Biol.*, **4**, 10.
19. Shalgi,R., Lieber,D., Oren,M. and Pilpel,Y. (2007) Global and local architecture of the mammalian microRNA-transcription factor regulatory network. *PLoS Comput. Biol.*, **3**, e131.
20. Re,A., Cora,D., Taverna,D. and Caselle,M. (2009) Genome-wide survey of microRNA-transcription factor feed-forward regulatory circuits in human. *Mol. Biosyst.*, **5**, 854–867.
21. Lenhard,B. and Wasserman,W.W. (2002) TFBS: computational framework for transcription factor binding site analysis. *Bioinformatics*, **18**, 1135–1136.
22. Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.
23. Kim,Y.K. and Kim,V.N. (2007) Processing of intronic microRNAs. *EMBO J.*, **26**, 775–783.
24. Lagos-Quintana,M., Rauhut,R., Lendeckel,W. and Tuschl,T. (2001) Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853–858.
25. Rajewsky,N. (2006) microRNA target predictions in animals. *Nat. Genet.*, **38(Suppl.)**, S8–S13.
26. Kertesz,M., Iovino,N., Unnerstall,U., Gaul,U. and Segal,E. (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet.*, **39**, 1278–1284.
27. Sethupathy,P., Megraw,M. and Hatzigeorgiou,A.G. (2006) A guide through present computational approaches for the identification of mammalian microRNA targets. *Nat. Methods*, **3**, 881–886.
28. Selbach,M., Schwanhausser,B., Thierfelder,N., Fang,Z., Khanin,R. and Rajewsky,N. (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature*, **455**, 58–63.
29. Baek,D., Villen,J., Shin,C., Camargo,F.D., Gygi,S.P. and Bartel,D.P. (2008) The impact of microRNAs on protein output. *Nature*, **455**, 64–71.
30. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. *et al.* (2011) NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.
31. Parkinson,H., Sarkans,U., Kolesnikov,N., Abeygunawardena,N., Burdett,T., Dylag,M., Emam,I., Farne,A., Hastings,E., Holloway,E. *et al.* (2011) ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **39**, D1002–D1004.
32. Burchard,J., Zhang,C., Liu,A.M., Poon,R.T., Lee,N.P., Wong,K.F., Sham,P.C., Lam,B.Y., Ferguson,M.D., Tokiwa,G. *et al.* (2010) microRNA-122 as a regulator of mitochondrial metabolic gene network in hepatocellular carcinoma. *Mol. Syst. Biol.*, **6**, 402.
33. Liu,H., Brannon,A.R., Reddy,A.R., Alexe,G., Seiler,M.W., Arreola,A., Oza,J.H., Yao,M., Juan,D., Liou,L.S. *et al.* (2010) Identifying mRNA targets of microRNA dysregulated in cancer: with application to clear cell Renal Cell Carcinoma. *BMC Syst. Biol.*, **4**, 51.
34. Taylor,B.S., Schultz,N., Hieronymus,H., Gopalan,A., Xiao,Y., Carver,B.S., Arora,V.K., Kaushik,P., Cerami,E., Reva,B. *et al.* (2010) Integrative genomic profiling of human prostate cancer. *Cancer Cell*, **18**, 11–22.
35. Palmer,R.D., Murray,M.J., Saini,H.K., van Dongen,S., Abreu-Goodger,C., Muralidhar,B., Pett,M.R., Thornton,C.M., Nicholson,J.C., Enright,A.J. *et al.* (2010) Malignant germ cell tumors display common microRNA profiles resulting in global changes in expression of messenger RNA targets. *Cancer Res.*, **70**, 2911–2923.
36. Puissegur,M.P., Mazure,N.M., Bertero,T., Pradelli,L., Grosso,S., Robbe-Sermesant,K., Maurin,T., Lebrigand,K., Cardinaud,B., Hofman,V. *et al.* (2011) miR-210 is overexpressed in late stages of lung cancer and mediates mitochondrial alterations associated with modulation of HIF-1 activity. *Cell Death Differ.*, **18**, 465–478.
37. Yanaihara,N., Caplen,N., Bowman,E., Seike,M., Kumamoto,K., Yi,M., Stephens,R.M., Okamoto,A., Yokota,J., Tanaka,T. *et al.* (2006) Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. *Cancer Cell*, **9**, 189–198.
38. Fujiwara,T., Hiramatsu,M., Isagawa,T., Ninomiya,H., Inamura,K., Ishikawa,S., Ushijima,M., Matsuura,M., Jones,M.H., Shimane,M. *et al.* (2012) ASCL1-coexpression profiling but not single gene expression profiling defines lung adenocarcinomas of neuroendocrine nature with poor prognosis. *Lung Cancer*, **75**, 119–125.
39. Wang,J., Lu,M., Qiu,C. and Cui,Q. (2010) TransmiR: a transcription factor-microRNA regulation database. *Nucleic Acids Res.*, **38**, D119–D122.
40. Cho,W.C. (2007) OncomiRs: the discovery and progress of microRNAs in cancers. *Mol. Cancer*, **6**, 60.
41. Garzon,R., Marcucci,G. and Croce,C.M. Targeting microRNAs in cancer: rationale, strategies and challenges. *Nat. Rev. Drug Discov.*, **9**, 775–789.
42. Hanahan,D. and Weinberg,R.A. Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.
43. Meyer,N. and Penn,L.Z. (2008) Reflecting on 25 years with MYC. *Nat. Rev. Cancer*, **8**, 976–990.

44. Dews,M., Homayouni,A., Yu,D., Murphy,D., Sevignani,C., Wentzel,E., Furth,E.E., Lee,W.M., Enders,G.H., Mendell,J.T. *et al*. (2006) Augmentation of tumor angiogenesis by a Myc-activated microRNA cluster. *Nat. Genet.*, **38**, 1060–1065.
45. Mestdagh,P., Fredlund,E., Pattyn,F., Schulte,J.H., Muth,D., Vermeulen,J., Kumps,C., Schlierf,S., De Preter,K., Van Roy,N. *et al*. MYCN/c-MYC-induced microRNAs repress coding gene networks associated with poor outcome in MYCN/c-MYC-activated tumors. *Oncogene*, **29**, 1394–1404.
46. Liang,Y., Li,W.W., Yang,B.W., Tao,Z.H., Sun,H.C., Wang,L., Xia,J.L., Qin,L.X., Tang,Z.Y., Fan,J. *et al*. (2012) Aryl hydrocarbon receptor nuclear translocator is associated with tumor growth and progression of hepatocellular carcinoma. *Int. J. Cancer*, **130**, 1745–1754.
47. Xiao,X., Li,B.X., Mitton,B., Ikeda,A. and Sakamoto,K.M. Targeting CREB for cancer therapy: friend or foe. *Curr. Cancer Drug Targets*, **10**, 384–391.
48. Aggarwal,S., Kim,S.W., Ryu,S.H., Chung,W.C. and Koo,J.S. (2008) Growth suppression of lung cancer cells by targeting cyclic AMP response element-binding protein. *Cancer Res.*, **68**, 981–988.
49. Kumar,A.P., Bhaskaran,S., Ganapathy,M., Crosby,K., Davis,M.D., Kochunov,P., Schoolfield,J., Yeh,I.T., Troyer,D.A. and Ghosh,R. (2007) Akt/cAMP-responsive element binding protein/cyclin D1 network: a novel target for prostate cancer inhibition in transgenic adenocarcinoma of mouse prostate model mediated by Nexrutine, a Phellodendron amurense bark extract. *Clin. Cancer Res.*, **13**, 2784–2794.