

ANKARA YILDIRIM BEYAZIT ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ



**RETİNOBLASTOM HASTALIĞINDA YENİ NESİL DİZİLEME
VERİ ANALİZİ İLE BİR ARDIŞIK DÜZENİN GELİŞTİRİLMESİ**

Doktora Tezi

Gülistan ÖZDEMİR ÖZDOĞAN

Elektrik ve Bilgisayar Mühendisliği

Ekim, 2020

ANKARA

**RETİNOBLASTOM HASTALIĞINDA YENİ NESİL
DİZİLEME VERİ ANALİZİ İLE BİR ARDIŞIK
DÜZENİN GELİŞTİRİLMESİ**

Ankara Yıldırım Beyazıt Üniversitesi

Fen Bilimleri Enstitüsü

Elektrik ve Bilgisayar Mühendisliği Doktora Tezi

Gülistan ÖZDEMİR ÖZDOĞAN

Ekim, 2020

ANKARA

DOKTORA TEZİ SONUÇ FORMU

GÜLİSTAN ÖZDEMİR ÖZDOĞAN tarafından **Dr. Öğretim Üyesi Hilal KAYA** danışmanlığında tamamlanan **“RETİNOBLASTOM HASTALIĞINDA YENİ NESİL DİZİLEME VERİ ANALİZİ İLE BİR ARDIŞIK DÜZENİN GELİŞTİRİLMESİ”** başlıklı tezi okuduk ve bir doktora tezi için kapsam ve içeriğinin tam olarak yeterli olduğunu onaylıyoruz.

Dr. Öğr. Üyesi Hilal KAYA

Danışman

Prof. Dr. İlyas ÇANKAYA

Jüri Üyesi

Doç. Dr. Baha ŞEN

Jüri Üyesi

Dr. Öğr. Üyesi Ayhan AKBAŞ

Jüri Üyesi

Dr. Öğr. Üyesi Cevat RAHEBİ

Jüri Üyesi

Prof. Dr. Ergün ERASLAN

Fen Bilimleri Enstitüsü Müdürü

ETİK BEYAN

Fen Bilimleri Enstitüsü'nün Tez Yazım Kılavuzuna uygun olarak hazırlanan bu tezde,

- Tüm veri, bilgi ve belgelerin akademik ve etik kurallar çerçevesinde elde edildiğini,
- Tüm bilgi, belge ve değerlendirmelerin bilimsel etik ve ahlak kurallarına uygun olarak sunulduğunu,
- Kullanılan tüm materyallerin tam olarak alıntılanmış olduğunu ve referans alındığını,
- Kullanılan materyallerde herhangi bir değişikliğin yapılmadığını,
- Sunulan tüm çalışmaların orijinal olduğunu,

beyan eder, yukarıdaki ifadelerin aksi bir durumda, yasal haklarımdan feragat etmeyi kabul edeceğimi bildiririm.

Tarih: 26.10.2020

İmza:

Adı Soyadı: Gülistan ÖZDEMİR ÖZDOĞAN

TEŞEKKÜR

Tez çalışmam boyunca verdiği destek, motivasyon ve yardımlardan dolayı danışman hocam Sayın Dr. Öğretim Üyesi Hilal KAYA'ya, tez süresince tez izleme komitesi jüri üyesi olarak benim için vakitlerini ayırıp, kıymetli öneri ve değerlendirmeleriyle çalışmama yön veren Sayın Prof. Dr. İlyas ÇANKAYA'ya ve Sayın Doç. Dr. Baha ŞEN'e teşekkürlerimi sunarım. Hakem olarak çalışmamı değerlendiren ve görüşleriyle çalışmamın şekillenmesini sağlayan uluslararası akademisyenlere teşekkürü bir borç bilirim.

Çalışmam boyunca hesaplamalarımı gerçekleştirdiğim TÜBİTAK TRUBA alt yapısına ve bu yapının oluşmasında ve sürdürülmesinde emeği geçenlere teşekkürlerimi sunmak isterim.

Hayatımın her anında olduğu gibi, bu yolculukta da en büyük gücüm, desteğim, umudum, nefesim olan ANNEM'e ve beni her zaman olduğu gibi doktora eğitim sürecinde de destekleyip, hayattaki tecrübeleriyle önümü açan ablalarım Hayriye, Derya ve Feride'ye ve ailelerine teşekkür ederim.

Akademik hayatım boyunca bana hep destek olan, aktardığı değerli tecrübeleri ve paylaşımlarıyla bu yolda daha emin ilerlememi sağlayan ve bu yolculukta sabırla benimle beraber yürüyen eşim Prof. Dr. Cem ÖZDOĞAN'a teşekkür ederim.

Son olarak, Covid-19 virüsüyle mücadelede yer alan tüm doktorlara ve sağlık çalışanlarına, tüm bilim insanlarına ve pandemi boyunca emekleriyle bu mücadeleye bir şekilde ortak olan dünya üzerindeki tüm insanlara teşekkürü bir borç bilirim.

Ekim 2020

Gülistan ÖZDEMİR ÖZDOĞAN

DEVELOPMENT OF A PIPELINE WITH NEXT-GENERATION SEQUENCING DATA ANALYSIS ON RETINOBLASTOMA DISEASE

ABSTRACT

Next-generation sequencing (NGS), revolutionized genomic researches, is related to massively parallel deoxyribonucleic acid (DNA) sequencing technology. Although the cost of generating NGS data was decreased compared to initially emerging stages of this technology, its cost might still be somewhat a problem according to studied data. New strategies such as pool-seq and low-coverage data have been developed to overcome this cost problem. Despite decreasing cost, it is important to elucidate whether they are efficient in NGS studies.

Within the scope of this thesis, a pipeline has been developed for pool-seq and low-coverage sequencing data obtained from tumors on retinoblastoma. Retinoblastoma is an eye malignancy in childhood that is initiated by RB1 mutation or MYCN amplification and can cause to the loss of vision of eye(s), and even sometimes life. In order to evaluate the effectiveness of the developed pipeline, obtained results on both the disease data with the required features and some other non-disease data exhibiting similar characteristics as much as possible were compared by working in conjunction with a standard counterpart. It has been observed that the developed pipeline is able to call larger number of variants and achieves to higher sensitivity and F-score values. Furthermore, results related to variants, variants called in disease-associated genes and variant types in retinoblastoma data are also presented. In order to evaluate the effectiveness of the developed pipeline more precisely, it is suggested to use cancer data with higher mutation rates and larger pools.

Since the alignment step in NGS data analysis is highly time-consuming and also inherently compatible with the GPU, some versions of the alignment algorithms running on the CPU have been developed for GPU executions. BWA which is the alignment algorithm utilized within the developed pipeline and BarraCUDA which is GPU adaptation developed for running under the CUDA environment were examined on different data sets and their performance was evaluated in detail. Accordingly, it

has been observed that BarraCUDA significantly reduces the computation time even on a single GPU and has a similar alignment rate to BWA as stated in all the data studied. In order to fully understand degree of the contribution of the GPU, BarraCUDA's performance on data with having a high number of reads and also the effect of using more than one GPU should be examined.

Keywords: NGS data analysis, Retinoblastoma, Low-coverage sequencing, Pool-seq, Alignment, CUDA, GPU

RETİNOBLASTOM HASTALIĞINDA YENİ NESİL DİZİLEME VERİ ANALİZİ İLE BİR ARDIŞIK DÜZENİN GELİŞTİRİLMESİ

ÖZ

Yeni nesil dizileme (YND), genomik araştırmalarda devrim yaratan büyük ölçüde paralel deoksiribonükleik asit (DNA) dizilemeye dayalı bir teknolojidir. YND verisi oluşturma maliyeti, bu teknolojinin ortaya çıktığı zamandaki ile karşılaştırıldığında azalmış olsa da, çalışılan veriye göre yine de bir problem oluşturabilir. Bu maliyet sorununun üstesinden gelebilmek için havuz dizileme ve düşük kapsamalı dizileme verileri gibi yeni stratejiler geliştirilmiştir. Düşen maliyete rağmen, bu stratejilerin YND çalışmalarında etkili olup olmadıklarını değerlendirebilmek önemlidir.

Bu tez kapsamında, tümör verilerinden elde edilen havuz dizileme ve düşük kapsama ile dizilenmiş retinoblastom verisi için bir ardışık düzen geliştirilmiştir. Retinoblastom, çocukluk çağında RB1 mutasyonu veya MYCN amplifikasyonu ile başlayan ve göz (ler) in görme kaybına ve hatta bazen ölüme yol açabilen bir göz kanseridir. Geliştirilen ardışık düzenin etkinliğini değerlendirebilmek için, hem bu özellikteki hastalık verisi hem de mümkün olabildiğince benzer özellikteki diğer veriler üzerinde standart bir ardışık düzenle birlikte çalışılarak sonuçlar karşılaştırılmıştır. Geliştirilen ardışık düzenin, daha fazla sayıda varyant çağırabildiği ve daha yüksek duyarlılık ve F-skor değerleri elde ettiği gözlemlenmiştir. Ek olarak, retinoblastom verisinde varyantlar, hastalık ile ilişkili genlerde çağrılan varyantlar ve varyant türleri ile ilgili sonuçlar sunulmuştur. Geliştirilen ardışık düzenin etkinliğini daha net değerlendirebilmek adına, daha yüksek mutasyon oranlarına ve daha büyük havuzlara sahip kanser verilerinin kullanılması önerilmektedir.

Öte yandan, YND veri analizinde hizalama adımı hem zamana gereksinim duyan hem de bu adımın temel karakteristiği sebebiyle GPU'ya uyumlu olduğundan, CPU'da çalışan hizalama algoritmalarının bazılarının GPU'da çalışabilen versiyonları geliştirilmiştir. Geliştirilen ardışık düzen kapsamında kullanılan hizalama algoritması olan BWA ve GPU için CUDA ortamında geliştirilen versiyonu olan BarraCUDA farklı veri setleri üzerinde incelenerek, performansları değerlendirilmiştir. Buna göre, BarraCUDA'nın tek GPU üzerinde bile ciddi anlamda çalışma süresini azalttığı ve

alıřılan tm verilerde belirtildiđi gibi BWA ile benzer hizalama oranına sahip olduđu gzlemlenmiřtir. GPU'nun katkısının tam olarak anlařılabilmesi iin, BarraCUDA'nın okuma sayısı fazla olan veriler zerindeki yaklařımı ve birden fazla GPU kullanımının etkisi incelenmelidir.

Anahtar Kelimeler: YND veri analizi, Retinoblastom, Dřk kapsamalı dizileme, Havuz dizileme, Hizalama, CUDA, GPU

İÇİNDEKİLER

DOKTORA TEZİ SONUÇ FORMU.....	ii
ETİK BEYAN.....	iii
TEŞEKKÜR.....	iv
ABSTRACT.....	v
ÖZ	vii
İÇİNDEKİLER	ix
TERMİNOLOJİ	xi
TABLolar LİSTESİ.....	xiii
ŞEKİLLER LİSTESİ	xiv
BÖLÜM 1 - GİRİŞ.....	1
1.1 Literatür Özeti	3
1.1.1 Retinoblastom	3
1.1.2 Yeni Nesil Dizileme (YND)	5
1.1.3 Yeni Nesil Dizileme Veri Analizi	8
1.1.4 YND Veri Analizi Araçları	16
1.1.5 Havuz Dizileme (Pool-seq)	18
1.1.6 Düşük Kapsamalı (Low-coverage) Dizileme.....	20
1.2 Tezin Amacı	23
1.3 Orijinal Katkı.....	24
1.4 Tez Organizasyonu	25
BÖLÜM 2 - ÖN BİLGİ.....	26
BÖLÜM 3 - GELİŞTİRİLEN ARDIŞIK DÜZEN	32
3.1 Hastalık Gen Listesinin Oluşturulması.....	32
3.2 Gereksinimler	32
3.3 Kullanılan Veriler.....	35
3.4 Ardışık Düzen.....	38
3.4.1 Referans Genom.....	38
3.4.2 Kalite Kontrolü.....	39
3.4.3 Hizalama	43
3.4.4 Hizalama Sonrası İşlemler	45
3.4.5 Alt-Örnekleme Süreci	46
3.4.6 Varyant Çağırma	48

3.4.7	Varyant Filtreleme	52
3.4.8	Yüksek Güvenilirliğe Sahip Varyantların Bulunması.....	59
3.4.9	Varyant Anotasyonu.....	60
3.4.10	Anotasyon Çıktılarının İşlenmesi.....	61
BÖLÜM 4 - BULGULAR.....		62
4.1	Karşılaştırmalı Sonuçlar	62
4.2	Hastalık Verisi Sonuçları.....	64
4.3	CUDA Çalışmaları	68
BÖLÜM 5 - SONUÇ VE ÖNERİLER		76
KAYNAKLAR		78
ÖZGEÇMİŞ.....		90

TERMİNOLOJİ

Kısaltmalar

ASCII	American Standard Code for Information Interchange
BAM	Binary Alignment Map
bç	Baz Çifti (bp-Base Pair)
BQSR	Baz Kalite Skoru Yeniden Kalibrasyonu (Base Quality Score Recalibration)
BWA	Burrows-Wheeler Alignment
BWT	Burrows-Wheeler transform
CNV	Kopya Sayısı Değişiklikleri (Copy Number Variations)
CUDA	Compute Unified Device Architecture
ÇU	Çift Uçlu (PE - Paired-end)
DNA	Deoksiribonükleik asit (Deoxyribonucleic acid)
GATK	The Genome Analysis Toolkit
GIAB	The Genome in a Bottle Konsorsiyumu
GP	Gerçek Pozitif (TP - True Positive)
GPGPU	Genel Amaçlı GPU programlama (General Purpose Computing on GPU)
GPU	Grafik İşlem Birimi (Graphics Processing Unit)
indel	İnsersiyon - Delesyon (indel, Insertion-Deletion)
İP	İş parçacığı
NCBI	The National Center for Biotechnology Information
RefSeq	NCBI Reference Sequence Database
RNA	Ribonükleik asit (Ribonucleic acid)
SAM	Sequence Alignment Map
SRA	Sequence Read Archive
TED	Tüm Ekzom Dizileme (WES - Whole Exome Sequencing)
TGD	Tüm Genom Dizileme (WGS - Whole Genome Sequencing)
TNP	Tek Nükleotid Polimorfizmi (SNP - Single Nucleotide Polymorphism)
TNV	Tek Nükleotid Varyant (SNV - Single Nucleotide Variant)
TRUBA	Türk Ulusal e-Bilim e-Altyapısı
TU	Tek Uçlu (SE - Single-end)
VCF	Variant Call Format

VQSR	Varyant Kalite Skoru Yeniden Kalibrasyonu (Variant Quality Score Recalibration)
YN	Yanlış Negatif (FN - False Negative)
YND	Yeni Nesil Dizileme (NGS - Next Generation Sequencing)
YP	Yanlış Pozitif (FP - False-Positive)

Terimler

Ardışık Düzen	Pipeline
Çift Uçlu	Paired-end
Derinlik	Depth
Duyarlılık	Sensitivity (Recall)
Düşük Kapsamalı Dizileme	Low-coverage Sequencing
Havuz Dizileme	Pool-seq
İş parçacığı	Thread
Kapsam	Coverage
Katı Filtreleme	Hard-filtering
Kesinlik	Precision
Okuma	Read
Okuma Grubu	Read Group
Sonek Ağacı	Suffix Tree
Tanımlama	Annotation
Tek Uçlu	Single-end

TABLÖLER LİSTESİ

Tablo 1. 1: DNA dizilemede kullanılan araçlar/yazılımlar	18
Tablo 3. 1: Geliştirilen ardışık düzende kullanılan araç/programlama dili/geliştirme ortamı listesi	35
Tablo 3. 2: Retinoblastom veri araştırması sonuçları.....	36
Tablo 3. 3: Hastalık verisinin özeti	37
Tablo 3. 4: Kalite skorları ve karşılık gelen doğruluk değerleri [106].....	41
Tablo 3. 5: Illumina fastq dosya bilgileri [111]	44
Tablo 3. 6: Kullanılan veriler ve bu verilerin alt-örneklemesinden elde edilen çalışma verileri	47
Tablo 3. 7: VQSR için kullanılan varyant veritabanları ve kaynak seti listesi	54
Tablo 3. 8: Katı filtrelemede kullanılması önerilen tanımlamalar ve eşik değerleri. 59	
Tablo 4. 1: Standart ve Geliştirilen ardışık düzenin karşılaştırılması	63
Tablo 4. 2: RB1 geninde bulunan varyantlar	66
Tablo 4. 3: CUDA ile yapılan analizler için kullanılan tüm veriler ve çeşitli özellikleri.....	70
Tablo 4. 4: BWA ve BarraCUDA'nın sonuçları	71

ŞEKİLLER LİSTESİ

Şekil 1. 1: Kanser araştırmalarında mevcut ve sonraya yönelik yaklaşımların durumu [2].	2
Şekil 1. 2: Sağlıklı gözün retinoblastomlu bir gözle karşılaştırılması.	4
Şekil 1. 3: A) Sağ gözdeki lökokori belirtisi B) retinoblastom tümörü [9]	5
Şekil 1. 4: DNA dizileme, [20].	7
Şekil 1. 5: Yeni nesil dizilemenin kanser araştırmaları ve klinik uygulamalara entegrasyonu [26].	10
Şekil 1. 6: TGD ve TED uygulamalarının iş akışı [27].	11
Şekil 1. 7: TED veri analizinin genel akış şeması [29].	12
Şekil 1. 8: Kullanılan biyoinformatik iş akışı [33].	13
Şekil 1. 9: Takip edilen TED iş akışı [37].	14
Şekil 1. 10: Tüm genom dizileme yaklaşımında havuz dizileme verisinin oluşturulması [52].	19
Şekil 1. 11: Dizilemede kapsam ve derinlik kavramları [61].	21
Şekil 2. 1: Canlıların genomlarını inceleyen genomik disiplinin temel birimleri [73].	26
Şekil 2. 2: DNA'nın çift sarmal yapısının gösterimi [74]	27
Şekil 2. 3: TNP örnekleme [75].	28
Şekil 2. 4: indel örnekleme [75].	29
Şekil 2. 5: CNV örneği [75].	29
Şekil 2. 6: Varyant örnekleme [75].	29
Şekil 2. 7: (a) TNP'ler, (b) haplotipler, (c) etiket (tag) TNP'ler [77].	30
Şekil 2. 8: VCF dosyası örneği [75]	31
Şekil 3. 1: Geliştirilen ardışık düzenin özeti.	33
Şekil 3. 2: YND veri analizi için takip edilen iş akışı	34
Şekil 3. 3: Örnek fastq formatı ve açıklamaları	40
Şekil 3. 4: (Soldan sağa olmak üzere) RB-H1 ve RB-H2'ye ait kalite skorları grafiği.	42
Şekil 3. 5: Kalite skor grafiğinin oluşma süreci	43
Şekil 3. 6: RB-H1 verisi için okuma grubu oluşturulması	45
Şekil 3. 7: RB-H1 (üst) ve RB-H2 (alt) hastalık verilerine dair edilen deneysel- atanan kalite skorları (empirical-reported quality score) grafiği.	50
Şekil 3. 8: GATK HaplotypeCaller'ın temel çalışma adımları [116].	51
Şekil 3. 9: VQSR sürecinin iş akışı [120].	53

Şekil 3. 10: VariantRecalibrator adımı sonunda üretilen model grafikleri örneği [120].	55
Şekil 3. 11: Hastalık verisi ile VariantRecalibrator adımı sonunda üretilen model grafikleri.	57
Şekil 3. 12: Seçilen veri ile VariantRecalibrator adımı sonunda üretilen tanımlama çifti dağılım grafiği örnekleri.	58
Şekil 4. 1: Kromozomlardaki varyant dağılımı	64
Şekil 4. 2: Retinoblastom gen listesindeki varyant dağılımı	65
Şekil 4. 3: Varyant çeşitleri (RefSeq veritabanı)	65
Şekil 4. 4: Zaman – Veri seti grafiği.	72
Şekil 4. 5: Küçük verilerin çalışma zamanı grafiği.	73
Şekil 4. 6: Daha büyük verilerin çalışma zamanı grafiği.	74
Şekil 4. 7: Veri seti içindeki en büyük verinin (SRR622457) çalışma zamanı grafiği.	75

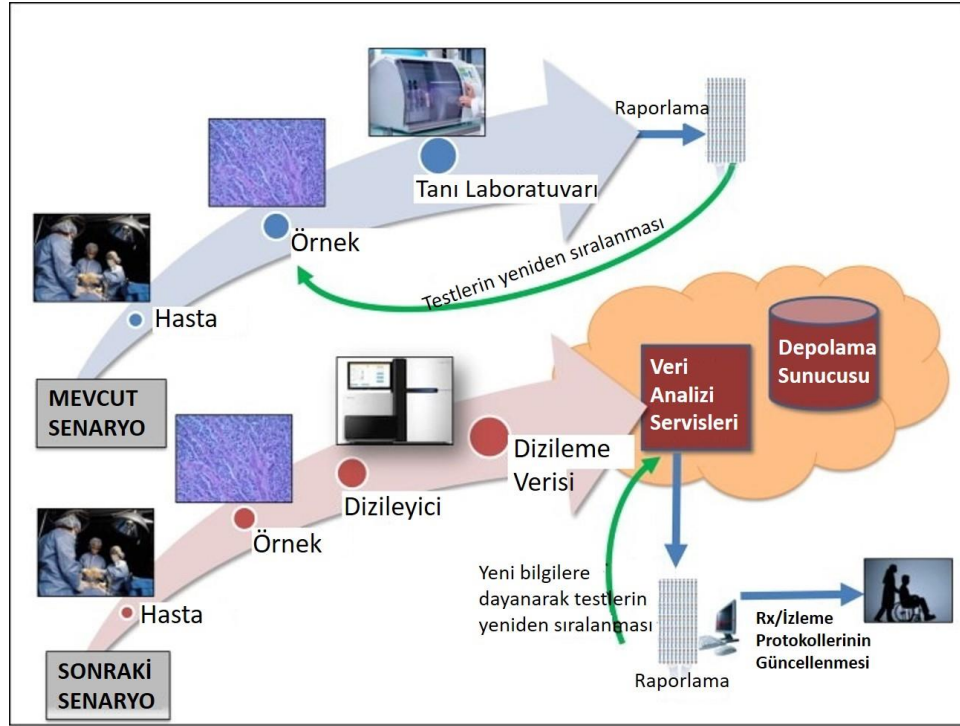
BÖLÜM 1

GİRİŞ

Kanser, insanların hayatını olumsuz bir şekilde etkileyen ciddi bir rahatsızlıktır. Genel olarak, hücrelerin kontrol dışı çoğalması ve büyümesi olarak tanımlanabilir. İnsan hücrelerinin normal bir biyolojik döngüsünde hücreler büyür, bölünür ve ölür. Ancak, kanserde hücrenin ölümü gerçekleşmez. Bunun yerine, gerekmediği halde hücre bölünmeye devam eder ve tümör denilen yapılar oluşur. Tümörler arasında, kansere sebep olan ve yayılma eğiliminde olan ‘kötü huylu’ olanlar olabildiği gibi, hücrelerin aşırı büyümesi sonucu oluşan ve çoğunlukla zararsız yapıda olan ‘iyi huylu’ tümörler de olabilir.

100’den fazla çeşidi olan kanser, ortaya çıktığı organ ya da dokunun ismi ile çağrılır [1]. Örneğin, gözün retina bölümünde oluşan kötü huylu tümörlerin sebep olduğu kanser retinoblastom olarak isimlendirilir. Bazen, kanser bir bölgede ortaya çıkıp, buradan kan ya da lenf sistemi ile vücudun diğer bölgelerine yayılır. Bu duruma ‘metastaz’, bu durumu yaratan kanserlere de ‘metastatik kanserler’ denir. Metastatik kanserler, insan hayatını daha ciddi şekilde etkileyen ve tedavinin yapılamadığı durumlarda ölüme sebebiyet veren kanserler olduğundan daha tehlikelidirler.

Günümüzde insan ölümlerine sebep olan faktörlerin arasında üst sıralarda yer alan kanser hastalığı bu özelliği ile birçok araştırmanın da odağı haline gelmiştir. Eskiden belli bir disiplinden olan kişiler tarafından yapılan bu bilimsel çalışmalar, zamanla farklı disiplinlerin bir araya gelerek daha etkin çözümlerin bulunabildiği çalışmalara dönüşmeye başlamıştır. Klinik çalışmalarda belli bir hasta için geliştirilen çözümler anlık fayda sağlarken, hastalardan alınan örneklerin moleküler olarak incelenmesiyle elde edilen bilgiler diğer hastaların da iyileşme sürecine katkı sağlayarak, depolanıp gelecekteki tedavi süreçleri için de bir kaynak oluşturmaktadır (Şekil 1.1).



Şekil 1. 1: Kanser araştırmalarında mevcut ve sonraya yönelik yaklaşımların durumu [2]. (Şekil, Kaynak [2]'den alıntılanarak Türkçeleştirilmiştir.)

Kanser hastalığı ile ilgili en önemli dönüm noktası İnsan Genom Projesidir. Genom, herhangi bir organizmanın sahip olduğu genetik bilgilerin tümüdür. 1990 yılında başlayan İnsan Genom Projesi, insan genomundaki yaklaşık üç milyar baz (A, T, G, C) çiftinin diziliminin bulunmasını, genlerin tanımlanmasını ve kalıtsal özelliklerin nesiller boyunca incelenebileceği haritalar çıkarmayı amaçlayan [3] uluslararası bir bilimsel projedir. 2003 yılında tamamlanan proje, kanserin sebepleri, genetik rahatsızlıkların tanı ve tedavisi, yeni ilaçların üretilmesi, genlerin fonksiyonelliğinin araştırılması ve biyoinformatik alanının gelişmesi gibi konularda önemli keşiflerin de önünü açmıştır [4].

Biyoinformatik adı, ilk olarak Paulien Hogeweg ve Ben Hesper tarafından 1970'li yılların başında 'biyotik sistemlerdeki bilişim süreçlerinin incelenmesi' tanımı ile kullanılmıştır [5]. Bilgisayar teknolojisindeki gelişmeler biyolojik verilerin üretilme kapasitesini arttırmıştır. Bu artış ile üretilen büyük miktardaki verinin etkin ve verimli bir şekilde işlenmesi ihtiyacının doğması biyoinformatik biliminin

gelişmesini sağlamıştır. Biyoinformatik, genel olarak, biyoloji, bilgisayar bilimleri ve istatistik dallarından oluşan disiplinlerarası bir bilim dalıdır.

Biyoinformatiğin temel olarak üç amacı vardır [6]. Bunlar, çeşitli araştırma grupları tarafından üretilen verilerin depolanarak diğer çalışmalarda kullanılabilmesi, bu verileri etkin bir şekilde analiz etmeyi sağlayan araçların ve kaynakların geliştirilmesi ve son olarak geliştirilen bu araçlarla verilerin analiz edip biyolojik açıdan anlamlı sonuçlar elde edilebilmesini sağlamaktır. Bu kapsamda, kanser gibi genetik hastalıkların sebeplerinin ve bu hastalıklarda etken olan genlerin tespitinin araştırılması ve ilaç keşifleri biyoinformatiğin uygulama alanlarından bazılarıdır.

Biyoinformatiğin temel amaçlarından olan biyolojik verilerin depolanması ve analizi yüksek başarımlı hesaplama gücüne ihtiyaç duyar. Ayrıca, veri analizi için gerekli olan araçların geliştirilmesi için yazılım geliştirme süreçleri kullanılır. Bunlar dışında, dinamik programlama, makine öğrenme, veri madenciliği, sinir ağları, genetik algoritmalar, veritabanı tasarımı ve geliştirilmesi, hesaplamalı geometri, doğal dil işleme, grafik ve grafik arayüzleri konuları [7] bilgisayar bilimi ile ilgilenen insanların biyoinformatikteki çalışma alanlarıdır.

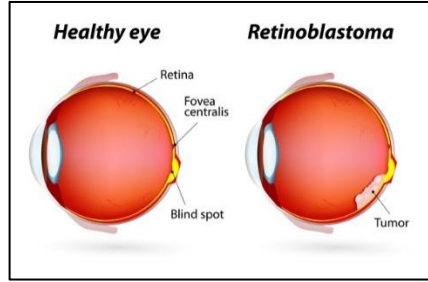
1.1 Literatür Özeti

1.1.1 Retinoblastom

Retinoblastom, çocukluk çağında retinada oluşan ve nadir görülen bir göz tümörü olmakla beraber, çocukluk döneminde en sık görülen göz içi kanseridir [8]. Çoğunlukla beş yaş altı çocuklarda görülen retinoblastom, tek gözde oluşabildiği gibi her iki gözde de gelişebilir. Erken tanı birçok kanser türünde olduğu gibi bu hastalık için de önemlidir. Retinoblastomun tedavi edilmediği durumlarda, tümör vücuttaki diğer organlara ve merkezi sinir sistemine yayılım göstererek hayati sonuçlar doğurabilir [9].

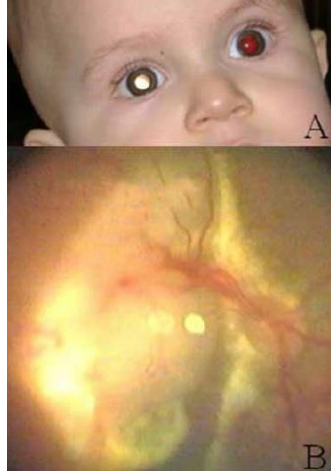
Retina, göz küresinin arka iç duvarını kaplayan ve ışığa duyarlı sinir hücrelerinden oluşan bir tabakadır. Bünyesindeki hücrelerle ışığı algılayıp optik sinir ile beyne ileterek görme işlevinin gerçekleşmesini sağlar. Genel olarak, bir hücrenin kontrol dışı büyümesi olarak tanımlanan kanser hastalığındaki durum retinoblastom için de

geçerlidir. Gözün ilk gelişim aşamasında retinayı doldurmak için çoğalan ‘retinoblast’ ismindeki göz hücrelerinin çoğalma süreci olgun retina hücrelerinin oluşmasıyla belli bir noktada durur, ancak bazı durumlarda, bu hücrelerin kontrolden çıkmasıyla retinoblastom kanserinin oluştuğu görülür [10]. Sağlıklı bir gözün retinoblastom tümörüne sahip bir gözle karşılaştırıldığı durum Şekil 1.2’de sunulmuştur.



Şekil 1. 2: Sağlıklı gözün retinoblastomlu bir gözle karşılaştırılması. (Designua/Shutterstock.com dan alınmıştır.) (Healthy eye: Sağlıklı göz, Retinoblastoma: Retinoblastom hastalığına sahip göz, Blind spot: Kör nokta, Tumor: Tümör)

Retinoblastomun en yaygın iki klinik belirtisi lökokori ve şaşılıktır [8]. Göz bebeğinden gelen beyaz ışık yansıması olan ve ‘beyaz pupilla refleksi’ olarak isimlendirilen lökokori, klinik testlerde doktorlar tarafından bulunabildiği gibi, çocuğun flaş ile çekilen fotoğrafları sırasında da fark edilebilir [9]. Şekil 1.3 (A)’da retinoblastomlu çocuğun gözlerine ışık verildiğinde her iki göz bebeğinin verdiği tepki görülmektedir. Buna göre, sağ gözdeki beyazlık lökokori olup, buna sebep olan göz tümörü ise Şekil 1.3 (B)’de görülmektedir Diğer belirti ve semptomlar ise, gözdeki ağrı veya kızarıklık, göz çevresi enfeksiyonu, normalden daha büyük olan göz küresi, gözün renkli kısmı irisin renginde değişiklik olarak sayılabilir [11, 12].



Şekil 1. 3: A) Sağ gözdeki lökokori belirtisi B) retinoblastom tümörü [9]

Retinoblastom geni olarak da bilinen ve hastalığın oluşmasında önemli bir yere sahip olan RB1 geni, 13. kromozomun 14q bandında yer alır. RB1 geni, tümör baskılayıcı gen olarak bilinen, hücrelerin büyümelerini düzenleyen ve kontrollü bir şekilde bölünmesinden sorumlu bir gendir [12]. Retinoblastom oluşması için, her biri bir ebeveynden gelmek üzere iki alele (bkznz. Bölüm 2) sahip olan RB1 geninin, retinoblast hücrelerindeki her iki kopyasında da mutasyon ya da kayıp olması gerekir [9].

Retinoblastom, genel olarak kalıtsal ve kalıtsal olmayan olmak üzere iki farklı şekilde oluşur. Kalıtsal retinoblastomda, vücuttaki hücrelerin tümünde RB1 geninde mutasyon varken, kalıtsal olmayan retinoblastom ise sadece tek bir retina hücresinde başlayan mutasyon ile gerçekleşir [9]. Retinoblastom, tek gözde ya da her iki gözde oluşmasına bağlı olarak “unilateral” ya da “bilateral” retinoblastom olarak isimlendirilir. Kalıtsal retinoblastoma sahip hastalarda aynı zamanda bir beyin tümörü gelişme durumu olursa “trilateral” retinoblastom oluşmuş olur [11]. Kalıtsal retinoblastoma sahip hastalarda beyinde oluşabilecek kötü huylu tümörler dışında, kemik kanseri, yumuşak doku kanserleri, melanom gibi kanserlerin gelişme durumu da söz konusudur [9].

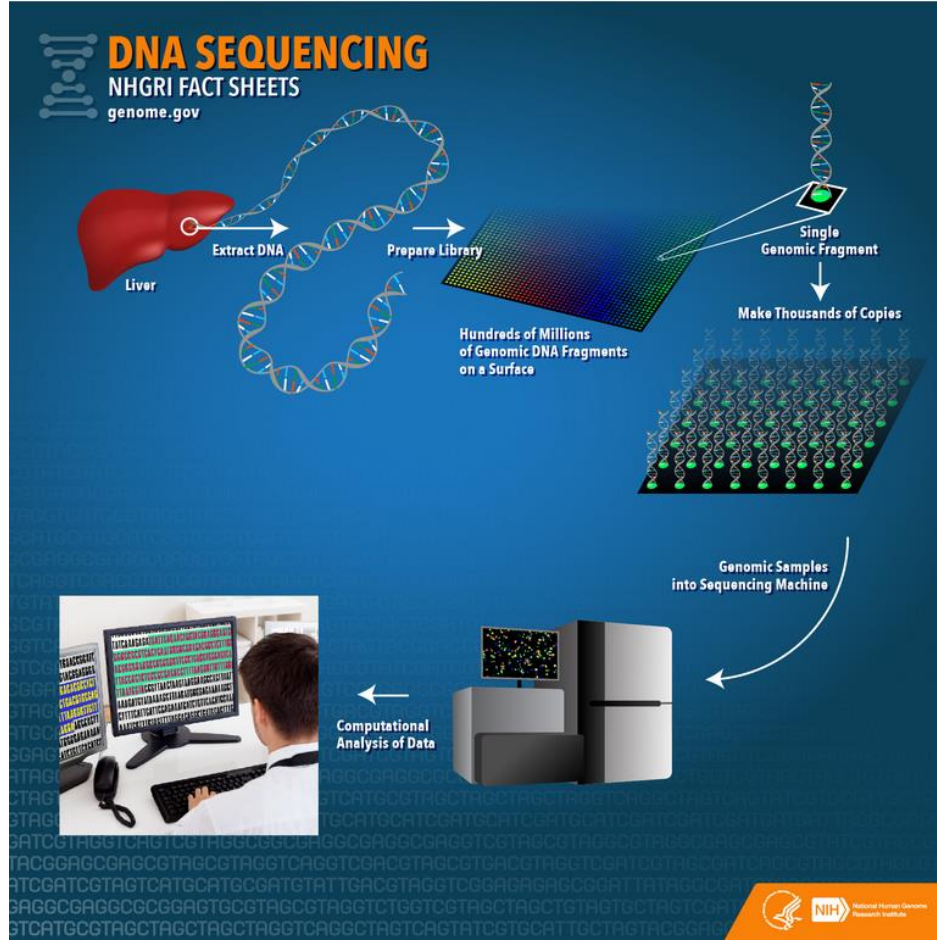
1.1.2 Yeni Nesil Dizileme (YND)

Dizileme, canlıların genetik bilgilerinin tutulduğu DNA üzerindeki nükleotitlerin sıralamasının bulunması sürecidir ve genom çalışmalarında önemli bir yere sahiptir.

Birçok farklı genom ve hastalık üzerinde yapılan araştırmalar ve sonucunda elde edilen keşifler dizileme ve dizileme teknolojisindeki gelişmelerin sonucudur.

DNA dizileme için, 1975'te Frederick Sanger tarafından zincir sonlandırma yöntemine dayanarak geliştirilen metot, 1977 yılında Sanger ve arkadaşları tarafından yayımlanarak [13] literatüre sunulmuştur. Aynı yıl, Maxam ve Gilbert tarafından da kimyasal parçalama yöntemine dayalı başka bir DNA dizileme metotuna dair çalışma [14] yayımlanmıştır. 1980 yılında, Frederick Sanger ve Walter Gilbert'e DNA dizilemeye yaptıkları bu katkılardan dolayı kimya dalında Nobel ödülü verilmiştir [15]. Bu ödül, Frederick Sanger'in 1958 yılında proteinlerin, özellikle de insülinin yapısına dair yaptığı çalışmalardan dolayı aldığı ilk Nobel ödülünden sonraki ikinci ödüdür [16]. Sanger tarafından geliştirilen metot, yüksek doğruluğa sahip olması [17] sebebi ile DNA dizilemenin altın standardı olarak bilinir [18]. Sanger dizileme yaklaşımı otomatikleştirilerek, DNA dizileme alanında "Birinci Nesil Dizileme" teknolojisi olarak kabul görmüş olup, İnsan Genom Projesi de otomatik Sanger dizileme ile gerçekleştirilmiştir [17]. İnsan Genom Projesi, hem harcanan milyarlarca dolar hem de on üç yıllık tamamlanma süresi [18] ile tarihin maliyetli projeleri arasındadır. Her ne kadar Sanger dizileme ile bu süreç başarılı bir şekilde tamamlansa da, hem onun kadar güvenilir hem de daha az maliyetli olan yeni bir teknolojiye ihtiyaç duyulduğu görülmüştür [17]. Bunun için, 2004 yılında Ulusal İnsan Genomu Araştırma Enstitüsü (The National Human Genome Research Institute, NHGRI) tarafından zaman ve maliyet problemini ortadan kaldırabilmek amacıyla yeni bir proje başlatılmıştır. Proje, DNA'nın paralel olarak dizilenmesi esasına dayalı yaklaşımlar üzerine kurulu yeni bir teknolojinin doğmasına sebep olmuştur ve bu yönde geliştirilen yaklaşımların tümü "Yeni Nesil Dizileme" ya da "Masif Paralel Dizileme" olarak isimlendirilmiştir [17]. YND, temel olarak kütüphanenin hazırlanması, dizileme ve veri analizi adımlarından oluşur. Şekil 1.4'de YND tabanlı DNA dizilemenin ana hatları görülmektedir. Buna göre, herhangi bir canlıdan alınan DNA örneğinden elde edilen yüz milyonlarca DNA parçası ile kütüphane oluşturulur. Hazırlanan her bir DNA parçası için binlerce kopya oluşturulur. Her bir DNA parçasının paralel bir şekilde sıralayıcılarla dizilenmesiyle, her bir baz birden fazla okunarak doğruluğu sağlanmış olur [19]. Son olarak, dizileme sonucu oluşan ham verinin analiz edilerek farklılıkların bulunması, doğrulanması ve hastalıklarla ilişkili olanların tespiti gerekir.

YND sürecinde, biyoinformatik alanında çalışan insanların odaklandığı kısım veri analizi kısmı olup, tez çalışması kapsamında da bu kısma odaklanılacaktır.



Şekil 1. 4: DNA dizileme, [20]. (Liver: Karaciğer, Extract DNA: DNA'nın Çıkarılması, Prepare Library: Kütüphane Hazırlanması, Hundreds of Millions of Genomic DNA Fragments on a Surface: Yüzey Üzerinde Yüz Milyonlarca Genomik DNA Parçası, Single Genomic Fragment: Tek Genomik Parça, Make Thousands of Copies: Binlerce Kopyanın Oluşturulması, Genomic Samples into Sequencing Machine: Genomik Örneklerin Dizileme Cihazlarına Gönderilmesi, Computational Analysis of Data: Verilerin Hesaplamalı Analizi)

YND ile tüm bir insan genomu ilk kez 2008 yılında, DNA'nın çift sarmal yapısını bulanlardan biri olan James D. Watson'a ait genomun dizilenmesiyle gerçekleşmiştir [21]. Burada, YND'nin hem birkaç aylık bir süre zarfında tamamlanması ile zaman maliyeti, hem de bir milyon dolardan daha aza düşen maliyeti ile daha avantajlı olduğu görülmüştür. YND teknolojisinin kanser genomları üzerindeki uygulamalarına yönelik yapılan bir derleme çalışmasında, YND ile ilk kez bir kanser genomunun

dizilenmesinin Ley ve arkadaşları tarafından [22] 2008 yılında Akut Miyeloid Lösemi (AML) hastalığı ile yapıldığı, 2008 ve 2012 yılları arasında en az 25 farklı kansere ait yaklaşık 800 genomun dizilendiği belirtilmiştir [23]. Tez kapsamında da, YND ile çalışılan kanserler üzerine yapılan literatür araştırmasında [23-26] birçok farklı kanser türünde YND teknolojisinin çalışıldığı görülmüştür.

YND teknolojisi, zaman ve maliyet açısından avantaj sağlasa da, sahip olduğu birtakım özellikler sebebiyle kullanımda bazı zorluklar da ortaya çıkarmıştır. YND her bir DNA parçasının paralel dizilenmesi esasına dayandığından, dizileme sırasında oluşan büyük veri [17, 18] bu teknolojinin beraberinde getirmiş olduğu zorluklardan biridir. Bunun dışında, YND’de kısa okuma uzunlukları sebebi ile hizalama ve birleştirme (assembly) süreçlerinin daha zor olması ve tek bir okumada Sanger dizilemeye göre daha yüksek hata oranları olması da karşılaşılan diğer zorluklardır [18]. Yeni nesil dizileme verisi üretebilen farklı platformlar mevcut olup, bunlardan Illumina ve Ion Torrent sıklıkla kullanılanlar arasındadır.

1.1.3 Yeni Nesil Dizileme Veri Analizi

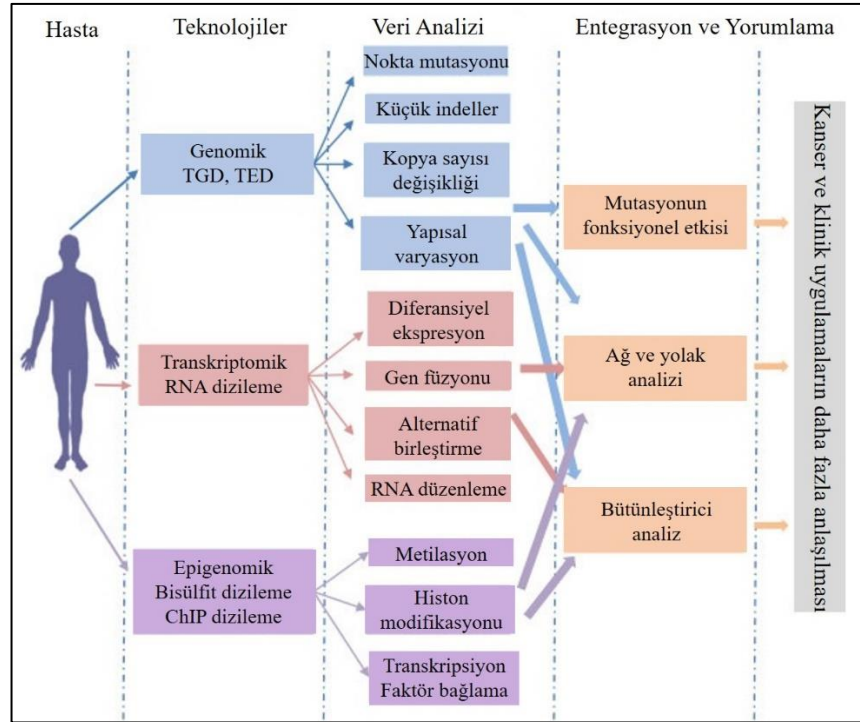
Yeni nesil dizileme sürecinin laboratuvar kısmının tamamlanmasının ardından oluşan büyük ham veri üzerinde veri analizi süreci gerçekleşir ve bu süreç temel olarak beş adımdan oluşur [27]:

- Kalite değerlendirme (kontrolü): Yeni nesil dizileme platformlarından elde edilen ham verinin kalitesinin değerlendirilerek, tanımlanmış standartlara uymayan okumaların düzeltilmesi, yok edilmesi ya da doğrulanması adımıdır.
- Hizalama: Bu adım, üst üste binen çok sayıdaki kısa okuma parçalarının birleştirilmesiyle orijinal dizinin yeniden elde edilmesi sürecidir ve iki şekilde gerçekleşir [28]: Birincisi, bakteri gibi küçük genomlar için kullanılan ve herhangi bir referans genoma ihtiyaç duymayan birleştirme (de novo) yaklaşımıdır. İkincisi ise, insan genomu gibi karmaşık genomların hizalanması sırasında kullanılan ve aynı türden ya da o türle yakından ilişkili bir referans genoma ihtiyaç duyan yaklaşımdır. Hizalama, temelde, bilgisayar bilimindeki dizgi eşleştirme problemi olmakla birlikte, dizgiler için tam bir eşleştirme yapmak hedeflenirken, hizalamada tam eşleşmenin olmadığı yerlerde

olabilecek varyasyonları yakalayabilmek ve eşleşmeyi hızlı bir şekilde gerçekleştirebilmek hedeflenir [29]. Önemli olan, okuma parçalarını referans genoma yüksek verim ve doğrulukla hizalamaktır. Karşılaşılabilecek problemler, okuma parçalarının kısa oluşu, tekrarlı dizilerin olması ve büyük ölçekli polimorfizmlerdir [18, 30].

- Varyant Tanımlama (Çağırma): Bu adım, hizalanmış dizi ile çalışılan genoma ait referans genomun karşılaştırılmasıyla dizi üzerindeki farklılıkların bulunması sürecidir.
- Varyant Anotasyonu: Bir önceki adımda çağrılan çok sayıdaki varyant arasından odaklanılması gerekenlerin seçilebilmesi için bu varyantların anlamlandırılması gerekir. Bunun için, daha önceden oluşturulmuş ve nette yayımlanmış çeşitli varyant veritabanları kullanılır [28].
- Görselleştirme: YND veri analizi ile elde edilen verilerin yorumlanmasında görsellikten yararlanılabilir. Örneğin, varyant tanımlama adımı ile elde edilen farklılıklar görsel bir araçla değerlendirilebilir.

Yeni nesil dizilemenin kanser araştırmaları ve klinik uygulamalardaki çözümleri Şekil 1.5’de görüldüğü gibi genomik, transkriptomik ve epigenomik olmak üzere üç farklı kategoride sınıflandırılabilir [26]. Bu tez kapsamında genomik altındaki teknolojilere odaklanılacak olup, genomik altındaki metotlar Tüm Genom Dizileme (TGD) ve Tüm Ekzom Dizileme (TED)’dir.



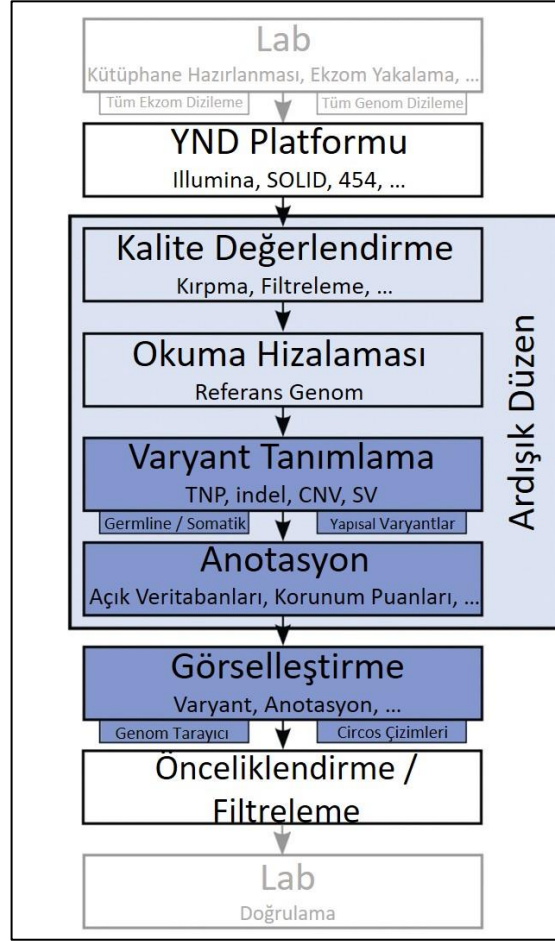
Şekil 1. 5: Yeni nesil dizilemenin kanser araştırmaları ve klinik uygulamalara entegrasyonu [26].

(Şekil, Kaynak [26]'den alıntılanarak Türkçeleştirilmiştir.) (Alternatif birleştirme: Alternative splicing, RNA düzenleme: RNA editing, Faktör bağlama: Factor binding)

YND teknolojileri kapsamındaki TGD, TED ve hedefli dizileme metotları, karmaşık hastalıklara sebep olan genlerin ve varyasyonların tanımlanmasını sağlayan metotlardır [27]. TGD, bütün bir genomu dizilemeye dayanan bir yaklaşım iken, TED sadece genom üzerindeki protein kodlayan ‘ekzon’ adı verilen bölgelerin birleşiminden oluşan ekzomun dizilenmesine dayalı bir yaklaşımdır [31]. TED’in daha düşük bir maliyetle küçük bir veri için büyük kapsama değerleri ile bir dizileme verisi üretebilme potansiyeli olduğundan [24], zaman ve maliyet açısından TGD’e göre daha avantajlı olduğu söylenebilir. Hedefli dizileme ise, seçilen bir bölgenin dizilenmesi esasına dayanır.

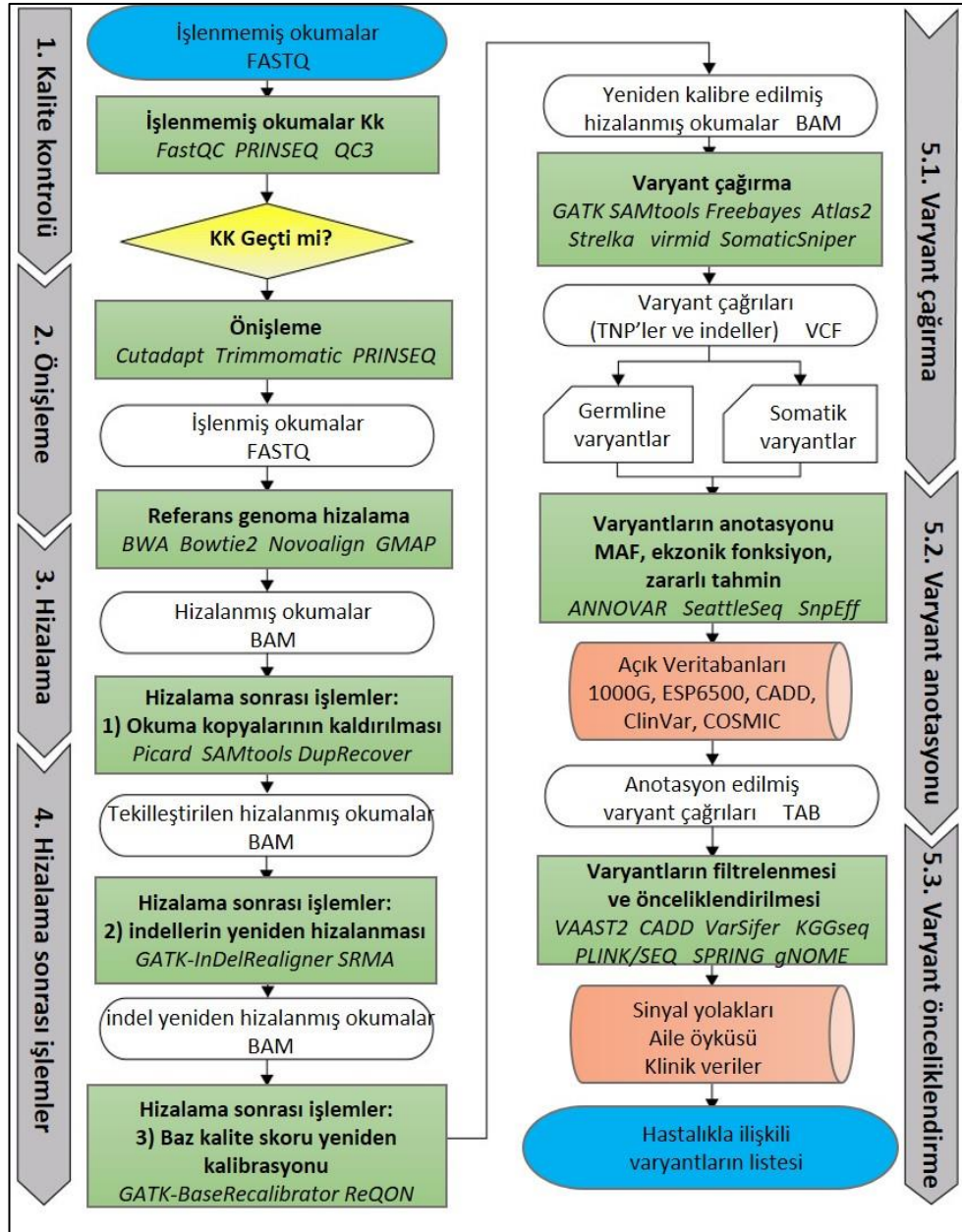
TGD ve TED uygulamalarının temel iş akışı Şekil 1.6’da sunulmaktadır. Burada, laboratuvar sürecinin ardından, seçilen YND platformu ile gerçekleştirilen dizileme ve bu sürecin sonunda oluşan ham veri üzerinde çalışılan veri analizi adımları gösterilmektedir. Buna göre, Kalite Değerlendirme, Hizalama, Varyant Tanımlama ve

Anotasyon olarak dört adımdan oluşan ardışık düzeni Görselleştirme, Önceliklendirme/Filtreleme ve Doğrulama adımları takip eder.



Şekil 1. 6: TGD ve TED uygulamalarının iş akışı [27]. (Şekil, Kaynak [27]'den alıntılanarak Türkçeleştirilmiştir.) (Korunum Puanları: Conservation Scores, Circos Çizimleri: Circos Plots)

Daha düşük maliyetli ve büyük kapsama değerleriyle dizilenmiş olması sebebi ile hastalıklara sebep olan varyasyonları belirlemede tercih edilen bir yaklaşım olan TED'in biyoinformatik analizine yönelik yapılan bir çalışmada, veri analizinin genel adımları Şekil 1.7'deki gibi sunulmuştur [29]. Burada, farklı adımlarda kullanılabilecek analiz araçları ve veritabanları da belirtilmiştir. Sunulan adımlar, çeşitli veri setleri üzerinde de denenerek, hassas bir varyant çağırma süreci için farklı araç ve kaynak kullanımları önerilmiştir.



Şekil 1. 7: TED veri analizinin genel akış şeması [29]. (Şekil, Kaynak [29]'den alıntılanarak Türkçeleştirilmiştir.) (Kk, KK: Kalite Kontrolü, Tekilleştirilen hizalanmış okumalar: Deduplicated read alignment, zararlı tahmin: deleterious prediction, Sinyal yolları: Signaling pathways)

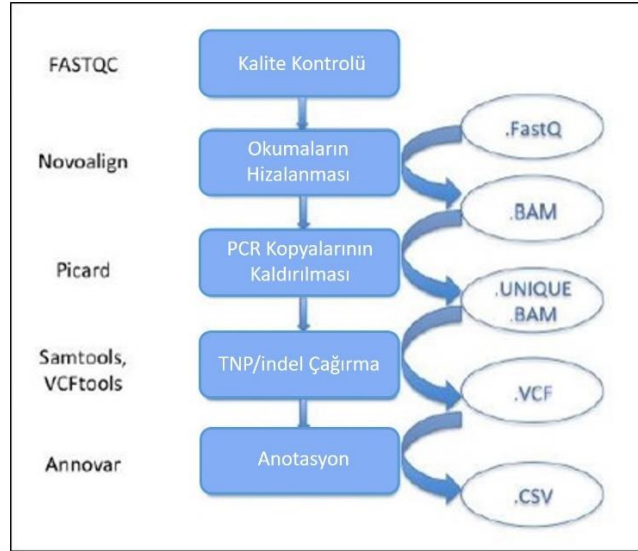
Yeni nesil dizileme ve veri analizine yönelik yapılan literatür araştırmasında, YND'nin başta kanser olmak üzere farklı hastalıklarda çalışıldığı görülmüştür. Ağır kombine immün yetmezliği hastası olduğu düşünülen hasta grubu üzerinde yapılan bir çalışmada TED ile hastalık ile ilişkili genlerde oluşan genetik değişikliklerin belirlenebilmesi hedeflenmiştir [32]. Nörodejeneratif hastalıklara yönelik yapılan bir

çalışmada ise seçilmiş ailelerden alınan ekzom dizileme verileri üzerinde Şekil 1.8’de sunulan biyoinformatik analizi çalışılarak hastalık ile ilişkili varyantlar bulunmaya çalışılmıştır [33]. Hastalığa sebep olan değişimlerin belirlenmesinde geleneksel yöntemlerin başarılı olamaması sebebi ile TED’e dayalı YND teknolojisinden yararlanıldığı ve YND teknolojisinin bu hastalıkların altında yatan genetik sebepleri ortaya çıkarabileceği konusundaki beklentileri belirtilmiştir.



Şekil 1. 8: Kullanılan biyoinformatik iş akışı [33]. (Şekil, Kaynak [33]’den alıntılanarak Türkçeleştirilmiştir.) (Nedensel Varyantlar: Causal Variants)

TED ile yapılan başka bir çalışmada, tüm ekzom dizileme ile dizilenmiş kolorektal kanser verisinde kopya sayısı değişiklikleri (Copy Number Variations, CNV) bölgelerini tespit edebilmek amaçlanmıştır [34]. TED metodu ile hastalıklarda etkin olan genler, L1 sendromu [35] ve ailesel meme ve yumurtalık kanseri [36] olmak üzere iki farklı hastalıkta çalışılmıştır. TED yaklaşımı kullanan diğer bir çalışmada Parkinson ve Distoni hastalığı üzerinde çalışılarak, takip edilen TED iş akışı, her adımda kullanılan veri analizi araçları ve her adım sonunda elde edilen dosya formatları Şekil 1.9’da sunulmuştur [37].



Şekil 1. 9: Takip edilen TED iş akışı [37]. (Şekil, Kaynak [37]'den alıntılanarak Türkçeleştirilmiştir.)

YND ile hastalıkların gelişiminde etken olan genlerin belirlenmesi, mesane kanseri [38] ve prostat kanserinde [39] YND veri analizine ek olarak yolak analizi (pathway analysis), zenginleştirme analizi (enrichment analysis) gibi farklı biyoinformatik yaklaşımlarla beraber de çalışılmıştır.

TED ve TGD metotlarını sayısal olarak karşılaştırabilmek adına, 2013 yılında yayımlanmış bir çalışmada verilen bilgiler kullanılabilir [27]. Buna göre, ortalama olarak TED'nin yaklaşık %90'ı daha önceden bulunmuş ve veritabanlarından erişilebilir olan 12000 varyant, TGD'nin ise 144000 yeni olmak üzere yaklaşık 5 milyon varyant çağırabildiği belirtilmiştir. Gelişen teknolojiler ve/veya geliştirilen araçlarla bu sayılar farklılaşabilse de, verilen rakamlar ile karşılaştırmalı bir bakış sunulmak istenmiştir.

TGD ve TED metotları karşılaştırıldığında, TED'nin sadece genom üzerindeki ekzon bölgelerine odaklandığı ve bu nedenle sahip olduğu bazı avantajlardan (düşük maliyet, daha kısa zaman) bahsedilmiştir. Öte yandan, TGD ve TED ile bulunabilen varyant sayıları karşılaştırıldığında TGD'nin bulabildiği varyant sayısının çokluğu ve protein kodlamayan bölgelerdeki bu varyantların ne anlama geldiği kanser genomu üzerine yapılan bir çalışmada sorgulanmıştır [25]. Bu çalışmada, kimi araştırmacıların protein kodlamayan bölgelerdeki farklılıkların kanser çalışmalarında etkili olabileceği ve yeni

keşiflerin önünü açabileceğini savunurken, kimi araştırmacıların ise protein kodlamayan bölgelerdeki farklılıkları anlamaya çalışmanın kanser çalışmalarındaki ilerlemeye sekte vurabileceği görüşünü savundukları belirtilmiştir. Hem dizileme teknolojisindeki gelişmelerle maliyetin düşmesi hem de birbirine zıt bu iki görüş sebebi ile kanser hastalığında TGD ile de çalışmalar yapılmaktadır.

Yapılan literatür araştırmasında, YND veri analizi odaklı çalışmalar da yapıldığı görülmüştür. Örneğin, bu yönde yapılan bir çalışmada, DNA ve RNA dizileme veri analizi incelenerek bu süreçler farklı kanser verilerinde uygulanmıştır [40]. Başka bir çalışmada, yeni bir yazılım kullanarak DNA ve RNA dizileme ile Farklılaşmamış Pleomorfik Sarkom (Undifferentiated Pleomorphic Sarcoma, UPS) kanseri üzerinde veri analizi çalışılarak mutasyonların belirlenmesi hedeflenmiştir [41]. CLC-bio adındaki bu yazılımın beta sürümü kullanılarak, çalışılan kanser türünde tanımlı olan mutasyonların belirlenebildiği belirtilmiştir. Ayrıca, bu yazılımın analiz süreçlerini kodlamaya gerek duymadan, arayüzle gerçekleştirebilen ilk güçlü yazılım olduğu da vurgulanmıştır. Diğer bir çalışmada, farklı veri türlerini (YND, biyolojik imaj, klinik veri) bütünleştirmeye yönelik ‘Personalized Oncology Suite’ (POS) adında Java tabanlı bir web uygulaması geliştirilmiştir [42]. Çalışma kolorektal kanseri ile gerçekleşse de, geliştirilen uygulamanın hastalığa bağlı parametreleri değiştirebilmeye olanak tanınması sebebi ile diğer hastalıkların da bu uygulamadan yararlanabileceği belirtilmiştir.

YND veri analizi çalışmalarının bir kısmı ise, bir ardışık düzen (pipeline) oluşturma esasına dayanır. Ekzom dizileme veri analizi için yüksek başarımlı hesaplama (High Performance Computing, HPC) yapan sistemlerde çalışabilen ardışık düzenler geliştirilmiştir [43-45]. Bunlardan biri, etkili bir şekilde veriyi işleyebilmek ve doğru sonuçlar alabilmek adına farklı tasarım prensipleri ve paralelleştirme tekniklerinin ifade edildiği, otomatik hale getirilmiş bir iş akışının geliştirildiği çalışmadır [43]. Bu iş akışı içerisinde, farklı açık kaynak kodlu araçların yanı sıra, bazı kısımlar için yazar tarafından geliştirilen araçların da kullanıldığı belirtilmiştir. Diğer çalışmada, veri analizi adımları için geliştirilen araçların bir ardışık düzen içerisinde birleştirilmesiyle varyantların belirlenmesi hedeflenmiştir [44]. Ardışık düzen ile çalışmanın YND veri analizi sonuçlarını daha kolay anlamayı sağladığı belirtilmiş olup, geliştirilen ardışık

düzenin klinik veri üzerinde denenmesi sonucu ekzom dizileme analizi için elverişli olduğu sonucuna varılmıştır. Diğer bir çalışmada ise, mevcut ekzom dizileme veri analizi ardışık düzeni sistem gereksinimlerinin ve çalışma zamanının azaltıldığı, daha esnek ve genişletilebilir bir formda olacak şekilde değiştirilmiştir [45]. Burada amaç, büyük miktardaki ekzom dizileme verisinden hastalık yapıcı varyantların ve genlerin belirlenmesini sağlayan bir süreç oluşturmaktır. Bu büyük veri, 87 farklı projeden alınan 4567 ekzom verisi ile oluşturulmuştur. Öte yandan, farklı varyant çağırma araçları ile varyant çağırma süreci çalışılarak, sonuçlar karşılaştırılmış ve etkili bir süreç için değerlendirmelerde bulunulmuştur. Bir ardışık düzen oluşturmaya amaçlayan çalışmalara başka bir örnek ise, mikrodizin ve dizileme teknolojileri kullanılarak kanserle ilişkili biyolojik belirteçlerin analizine yönelik bir çalışma olup, bu kapsamda istatistiksel ve biyoinformatik araçların uygulanması ya da geliştirilmesi hedeflenmiştir [46]. Çalışma içerisinde, genomik ve transkriptomik dizileme verileri için bir ardışık düzen geliştirilmiştir. Bunun için, hem varolan biyoinformatik araçları hem de ağ zenginleştirme analizine dayalı önerilen yeni yaklaşım kullanılmıştır.

1.1.4 YND Veri Analizi Araçları

YND veri analizi için pratikte birçok farklı araç ve yazılım geliştirilirken, teoride farklı platformlardan alınan veriler üzerinde çalışabilen, varyantların belirlenmesi sırasında geniş ve güncel bilgi sağlayabilen ve kullanıcıya etkin bir arayüz sunabilen bir yazılım geliştirilmesi hedeflenir [28]. YND veri analizi yapabilen birçok farklı araç/yazılım olmasına karşın, hangi aracın/yazılımın en iyi olduğu konusunda belli bir uzlaşımın olmadığı görülmektedir [30]. Bunun nedenleri, araçların sürekli yenilenmesinden dolayı doğru bir karşılaştırma yapabilme olanağının olmaması, araçların farklı platformlardan gelen veriler için geliştirilmeleri, birçok aracın tek bazlık varyasyonlar dışındaki mutasyonları bulmakta zayıf oldukları ve araçların performanslarının direkt karşılaştırılabilecekleri verinin sınırlı olmasıdır.

Varyantların belirlenmesinde en temel kriter, doğru ve etkili bir hizalama yapılmış olmasıdır. Bir dizi hizalama aracı ya da yazılımının performansı, hizalama hızı, hafıza gereksinimi, duyarlılık (sensitivity; hizalama oranı) ve doğruluk (doğru hizalama oranı) ölçütleriyle değerlendirilir [47]. YND veri analizi sürecinin hızlı gelişimi sebebiyle bu yönde yapılan değerlendirmeler güncel kalamasa da, varyant çağırma

süreci için çalışılan veri kümesi içindeki tüm verilerin eş zamanlı olarak bir arada analizini sağlayan yaklaşımlar önerilmektedir [48]. Bunun yanı sıra, bir varyant çağırma aracının varyant olmadığı halde varyant olarak belirlenen sayıyı (Yanlış Pozitif - YP) azaltması da varyant çağırma süreci için yapılacak öneriler arasındadır [49]. Varyant anotasyonu için geliştirilen araçlar farklı yaklaşımlar kullansa da, birçoğu yapısal farklılıklar yerine, tek ya da birkaç bazın farklılığına dayalı varyantları bulmayı hedefler [27].

Varyant analizi üzerine yapılan bir çalışmada, 205 tane YND veri analizi aracı incelenmiş ve bunlardan 32 tanesi seçilerek dört veri seti üzerinde test edilmiştir [27]. Çalışmanın sonuçlarına göre, her bir veri analizi adımı için aşağıda listelenen öneriler ve tespitlerde bulunulmuştur:

- Kalite kontrolü için filtreleme ve kırpma işlevi olan araçların kullanılması tavsiye edilmektedir.
- Hizalama adımı için farklı araçların geliştirildiği ve zamanla bu araçların iyileştirildiği ve bazı araçların, *BWA*, *Bowtie* ve *SOAP* gibi, hizalamada belli bir standarda ulaşarak, sık kullanılan araçlar arasına girdiği belirtilmiştir.
- Varyant çağırmada tüm varyantların tek bir yaklaşım ile belirlenemeyebileceği ve bu sebeple de aynı çalışmada birkaç aracın birlikte kullanılması önerilmiştir.
- Varyant anotasyonu araçları için, kendi içlerindeki varyant veritabanlarını sürekli güncel tutabilen araçların seçilmesi ve örün (web) tabanlı uygulamalar tercih edilecekse güvenlik konusunun göz ardı edilmemesi gerektiği belirtilmiştir.
- Görselleştirme araçları için de hem veri güvenliği hem de yasal konular konusunda uyarıda bulunulmuştur ve veri analizinin farklı adımlarında oluşan veri formatlarını tanıyan ve kullanabilen araçlar seçilmesinin daha anlamlı sonuçlar üretebileceği belirtilmiştir.

Daha önce bahsedildiği gibi YND veri analizi için birçok farklı araç ya da yazılım geliştirildiğinden, bunların tümünü burada sunabilmek mümkün değildir. Ancak, tez kapsamında DNA dizilemeye yönelik yapılan literatür araştırması [24, 26, 27, 29, 30,

49, 50] sırasında ulaşılan araç/yazılımlar çalışıldığı analiz adımları ile birlikte Tablo 1.1’de sunulmuştur.

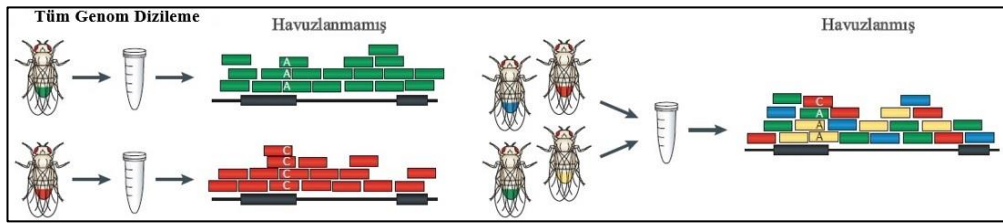
Tablo 1. 1: DNA dizilemede kullanılan araçlar/yazılımlar

Veri Analizi Adımı		Araç / Yazılım
Kalite Kontrolü		FastQC, FastQ Screen, FASTX-Toolkit, NGSQC Toolkit, PRINSEQ, QC-Chain, QC3, Cutadapt, Trimmomatic, ContEST, htSeqTools, PIQA, SolexaQA, TagCleaner, TileQC
Hizalama		MAQ, BWA, ELAND, SSAHA2, Bowtie2/Bowtie, SOAP3/SOAP2, SHRIMP2/SHRIMP, Corona Lite, BFAST, Novoalign/NovoalignCS, Stampy, MOSAIK, mrFAST, mrsFAST, YOABS, GMAP
Varyant Çağırma	Mutasyon	CRISP, SNVer, FreeBayes, Atlas2, GATK, SAMtools, VarScan2/VarScan, SomaticSniper, JointSNVMix, SolSNP, Shimmer, Seurat, QuadGT, MuTect, MutationSeq, Strelka, deepSNV, Virmid, SNVMix, CASAVA, SOAPSnp
	CNV	CNVnator, RDXplorer, CONTRA, ExomeCNV, CBS, SegSeq, CNaseg
	indel ve Yapısal Varyasyon	BreakDancer, Dindel, GenomeSTRiP, Pindel, VariationHunter, PEMer, SVDetect, Breakpointer, CLEVER, GASVPro, SVMerge
Varyant Anotasyonu		ANNOVAR, AnnTools, NGS-SNP, SeattleSeq, snpEff, SVA, VARIANT, VEP
Görselleştirme		Ensemble Genome Browser, UCSC Genome Browser, VEGA Genome Browser, Artemis, Integrative Genomics Viewer (IGV), Savant, Bambino, Tablet, MagicViewer, Geneious, Circos
Tüm Veri Analizi		Araç / Yazılım
Analitik Ardışık Düzenler		HugeSeq, SIMPLEX, TREAT
İş Akışı Sistemleri		Galaxy, LONI, Taverna, Synapse

1.1.5 Havuz Dizileme (Pool-seq)

Yeni nesil dizileme teknolojilerindeki gelişmeler, bu metodolojinin giderek yaygınlaşmasını ve maliyetinin düşmesine yol açtı. Ancak, yine de çok sayıda örneğin ayrı ayrı dizilenmesi belli bir maliyet yükü doğurmaktadır [51]. Bu nedenle yeni çözümler üretilmeye çalışılmıştır. Bunlardan biri, bir grup genomdan alınan DNA’ların bir karışım halinde dizilenmesidir [52]. Bu yaklaşıma havuz dizileme ve bu şekilde oluşan veriye de havuz dizileme verisi denir. Şekil 1.10’da tüm genom dizilemede havuzlanmamış (unpooled) ve havuzlanmış (pooled) veri yaklaşımı görülmektedir. Bu yaklaşımın amacı, maliyeti biraz daha azaltmak, süreci

hızlandırmak ve varyantları belirleyebilmek olup, buradaki en önemli nokta, varyantların doğru bir şekilde belirlenebilmesi ve alel frekanslarının tam olarak tahmin edilebilmesidir. Çünkü, dizileme sırasında oluşan hatalar, düşük frekansa sahip alellerle karıştırılarak, YP varyant sayısını yükseltebilmektedir [53]. Havuz dizileme, genom çapında ilişkilendirme çalışmaları (Genome Wide Association Studies, GWAS), polimorfizm keşfi ve alel frekansı tahmini, popülasyon yeniden dizileme ve genom evrimi gibi farklı araştırma alanlarında kullanılmaktadır [54].



Şekil 1. 10: Tüm genom dizileme yaklaşımında havuz dizileme verisinin oluşturulması [52]. (Şekil, Kaynak [52]'den alıntılanarak Türkçeleştirilmiştir.)

YND için geliştirilen araçların bazıları havuz dizileme verileri için de çalışabilmekle beraber, havuz dizileme verileri için geliştirilen araçlar da mevcuttur. Bunlar arasında en bilineni olan *CRISP*, birden fazla havuz dizileme verisi içerisinde hem nadir hem de yaygın olarak görülen varyantları bulmayı sağlayan bir araçtır [55]. *PoPoolation2*, havuz dizileme verisiyle popülasyonların karşılaştırılması amacıyla tasarlanmış bir araçtır [56]. *vipR* ise R ve Java programlama diliyle havuz dizileme verisi içinden hızlı ve etkili bir şekilde varyantları belirleyebilmek için geliştirilmiş bir araçtır [57]. Havuz dizileme verisi kullanılırken gerçek varyantların dizileme hataları ile karıştırılması probleminin üstesinden gelebilmek için Bayes yaklaşımı ile *snape* isminde bir araç geliştirilmiştir [58].

Havuz dizileme verisinde kullanılabilen araçların performansı ile ilgili yapılan bir çalışmada [51] 5 aracın, GATK, CRISP, LoFreq, VarScan, ve SNVer, doğruluk, duyarlılık ve özgüllük (specificity) gibi metriklerin yanı sıra, çalışma süreleri ve hafıza kullanımı bakımından da karşılaştırılmaları yapılmıştır. Çalışmada, sentetik olarak Illumina havuz dizileme verisi üretilmiştir. Doğruluk değerleri açısından, GATK, CRISP ve LoFreq'in farklı havuz dizileme verilerinde VarScan ve SNVer'e göre daha

yüksek doğruluk verdiği görülmüştür. Öte yandan, CRISP ve LoFreq'in hafıza kullanımı, çalışma zamanı, doğruluk ve kullanım kolaylığı açısından GATK'e göre daha iyi olduğu belirtilse de, GATK'in küçük havuzlar kullanan çalışmalarda en uygun duyarlılığın arandığı durumlarda maliyetin bilinerek tercih edilebileceği belirtilmiştir.

Literatürde havuz dizileme verisi bazı hastalıklar üzerinde de çalışılmıştır. Bunlardan birinde, 387 Parkinson hastasına ait DNA örnekleri 39 havuz verisi olarak dizilenmiştir [59]. Burada, Parkinson hastalığı ile ilgili olan 71 gene ait 997 ekzom bölgesi incelenmiştir. Veri analizi için Trimmomatic, BWA, Picard, GATK ve ANNOVAR araçları kullanılmıştır. Analiz sonucunda doğrulanmış 17 adet varyantın 6 tanesinin yeni olduğu ve havuz dizileme verisi ile yeni varyantların belirlenebildiği belirtilmiştir. Başka bir çalışmada ise, nörogelişimsel bozukluklar ekzom havuz dizileme verisi ile çalışılmıştır [60]. Bunun için, 96 bireye ait veri 8 havuz verisi olarak dizilenmiş olup, varyantlar belirlenerek hastalıkla ilişkileri incelenmiştir. Ekzom tabanlı havuz dizilemeye dayalı çalışılan bu yaklaşımın uygun maliyetli olduğu ve hastalık ile ilgili hızlı tarama yapabildiği belirtilmiştir.

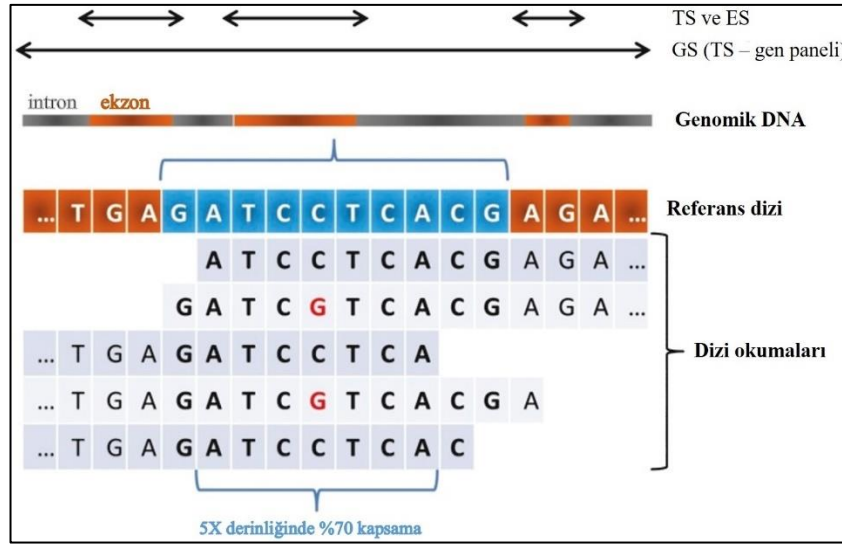
Havuz dizileme ile yapılan büyük kapsamlı bir çalışmada 996 bireye ait veri 83 havuz olarak dizilenmiştir [53]. Bu çalışmada, yanlış çağrılan varyantların kaldırılması amacıyla *Kolmogorov-Smirnov* istatistiksel testine dayanan bir filtreleme önerilmiştir. Önerilen yaklaşımın doğrulanması için havuz dizileme verisinden çağrılan varyantlarla, bir havuz içindeki verinin ayrı ayrı dizilenmesinden belirlenen varyantlar karşılaştırılmış ve önerilen filtreleme yaklaşımı ile gerçek varyantlar büyük oranda korunurken yanlış varyantlarınsa birçoğunun elendiği görülmüştür. Ayrıca, önerilen yaklaşımın diğer havuz dizileme verisi çalışmalarına da uygulanabileceği belirtilmiştir.

1.1.6 Düşük Kapsamlı (Low-coverage) Dizileme

Yeni nesil dizilemede, *okuma (read)* adı verilen parçalara ayrılan genom birçok kez dizilenir. Şekil 1.11'de farklı dizileme yaklaşımlarında *kapsam (coverage)* ve *derinlik (depth)* kavramları açıklanmıştır [61]. Seçilen yaklaşıma göre, tüm genom (GS), ekzom bölgeleri (ES) ya da sadece belli gen bölgeleri (TS) kullanılır. Şekil 1.11'de ekzon üzerinde alınan bir bölge ve bu bölgenin okumalarla (grili bölgeler) farklı

sayılarda dizilendiği görülmektedir. Şekilde, 10 bazlık bir alandaki bazların 7'sinin 5 farklı okumada kapsandığı görülmektedir. Buna göre, bu bölgenin *kapsamı* %70, *derinliği* ise 5X olur.

Kapsam, derinlik ya da kapsama derinliği olarak çağrılmakta olup, okuma derinliğini ifade etmekte ve dolayısıyla kapsam ve derinlik kavramları birbirlerinin yerine kullanılabilir [62]. Tüm genom dizileme için 30X ortalama bir kapsama derinliği iken, 10X değerinden daha aşağı değerler düşük kapsama olarak kabul edilir. 1X'in altındaki derinlik değerleri ise, çalışılan verideki okumalar arasında boşluklar bulunabileceğini gösterir.



Şekil 1. 11: Dizilemede kapsam ve derinlik kavramları [61]. (Şekil, Kaynak [61]'den alıntılanarak Türkçeleştirilmiştir.)

Tüm genom dizilemenin etkinliğinin tam olarak görülebilmesi için ihtiyaç duyulan çok sayıda örnek, dizileme maliyetlerinin düşmüş olmasına karşın yine de bir maliyet yükü doğurur ve bu durumun üstesinden gelebilmek için iki farklı yaklaşım tercih edilir [63]. İlki, tüm genom yerine sadece genomun ekzom bölgelerinin yüksek kapsama ile dizilenmesi, ikincisi ise tüm genomun düşük kapsama ile dizilenmesidir. Kompleks özellik genetiği üzerine yapılan bir çalışmada, çok sayıda bireyin düşük kapsamayla dizilenmesinin güçlü ve uygun maliyetli olduğu görülmesi sebebi ile az sayıda örneğin yüksek kapsamayla dizilenmesine alternatif olabileceği belirtilmiştir [64]. Düşük kapsamalı dizileme verisinin çok sayıda genoma gereksinim duyulan deneylerde

maliyeti etkili bir şekilde azalttığı belirtilen başka bir çalışmada, bu çeşit veri ile nadir varyantların hastalıklarla ilişkisinin tespiti için yeni yaklaşımlar önerilmiştir [65]. Ayrıca, hastalıklarla ilişkili varyant tespitinde düşük kapsamlı dizileme ve havuz dizileme stratejilerinin kullanıldığı durumlarda mümkün olduğu kadar çok sayıda bireyin deneylere dahil edilmesi gerektiği de belirtilmiştir. Düşük kapsamlı dizileme verisi ile çalışmanın maliyet avantajının yanı sıra dezavantajı da vardır. O da, doğru bir varyant tanımlama süreci için her bir lokusta yeterli sayıda okuma olmayabilir [66].

Düşük kapsamlı dizileme verisi için önerilen bir çalışmada, delesyonların belirlenmesini hedefleyen konvolüsyonel sinir ağları (CNN) yaklaşımına dayanan *CNNDel* isminde bir metot geliştirilmiştir [67]. Bunun için, farklı araçlardan yapısal varyasyonlar toplanarak özellikleri çıkartılmıştır. Daha sonra, geliştirilen *CNNDel* metodu oluşturulan varyasyon seti ile eğitilerek YP sonuçları elemesi sağlanmıştır. Çalışmada, 1000 Genom Projesi içinden alınan 26 tane düşük kapsamlı dizileme verisi çalıştırılarak, sonuçları değerlendirilmiştir ve konvolüsyonel sinir ağlarının etkili bir şekilde YP sonuçları azalttığı belirtilmiştir. Varyant çağırma için karar ağacı tabanlı *Fuwa* adında bir metodun önerildiği çalışmada TGD ve TED verilerinin yanı sıra düşük kapsamlı dizileme verisi de kullanılmıştır [66]. *Fuwa*'nın, Platypus, GATK-UnifiedGenotyper, GATK-HaplotypeCaller ve SAMtools araçları ile karşılaştırıldığında, yüksek doğruluk ve duyarlılıkta çalıştığı ve onlardan daha hızlı olduğu görülmüştür. Ayrıca, varyant çağırma araçlarının kullandığı çok sayıda parametre ve bu parametreler için kabul gören eşik değerlerinin düşük kapsamlı dizileme verileri için uygun olmayabileceği belirtilmiştir. *Fuwa*'nın makine öğrenme tabanlı geliştirilmiş olması ve farklı veri setlerine uyarlanabilmesinden dolayı, düşük kapsamlı dizileme verileri için iyi bir seçenek olabileceği sonucuna varılmıştır.

Havuz dizileme verisinde olduğu gibi düşük kapsamlı dizileme verisi için geliştirilen araçlar da mevcuttur. Geniş çaplı TNP keşfi yapabilmek için Reveel aracı [68], uzun okuma dizilerinden oluşan düşük kapsamlı verilerde yapısal varyasyonların belirlenmesini sağlayan NextSV aracı [69] ve yine yapısal varyasyonların belirlenmesi için SVSeq [70] ve SVSeq2 [71] araçları geliştirilmiştir.

Düşük kapsamlı dizileme verileri üzerinde çalışabilen araçların performansına dair yapılan bir çalışmada dört farklı TNP çağırma aracının, SOAPsnp, Atlas-SNP2,

SAMtools ve GATK, tek örnekli veriler üzerindeki sonuçları incelenmiştir [72]. GATK ve Atlas-SNP2 diğerlerine göre daha yüksek pozitif çağırma oranı ve duyarlılığa sahipken, GATK'in daha çok TNV (Tek Nükleotid Varyant, SNV) çağırdığı görülmüştür. Çalışmanın sonuçlarına göre, araçların çıktıları arasında benzerlik oranının düşük olması sebebi ile, araç seçimi, parametrelerin belirlenmesi ve doğrulama süreçlerinde dikkatli olunması gerektiği vurgulanmıştır. Ayrıca, tek bir aracın kullanılacağı durumlarda GATK'in tercih edilmesi, ancak doğruluğu arttırabilmek açısından da birden fazla aracın kullanımı önerilmektedir.

1.2 Tezin Amacı

YND, DNA parçalarının paralel olarak dizilenmesine dayanan bir teknolojidir. Ortaya çıktığı günden beri, başta kanser olmak üzere birçok farklı hastalık verisi ile farklı amaçlarla çalışılmıştır. Tez kapsamında, çocukluk çağı göz içi kanseri olan retinoblastom hastalığının YND ile çalışılması hedeflenmiştir. YND'nin deneysel süreci sonunda elde edilen veri, çeşitli analiz adımlarıyla hastaların DNA dizilimlerindeki farklılıkların ve bunların hastalıklar ile ilişkisinin tespitinde kullanılır. Bu tez kapsamında, varolan analiz araçları ile retinoblastom hastalığına ait YND verisi üzerinde veri analizi çalışılarak nokta mutasyonların (TNP) ve indellerin çağırılması ve yorumlanması hedeflenmektedir.

YND, Sanger dizilemeye göre maliyeti azaltan bir süreç olsa da, çalışma kapsamında gerekli veri setine bağlı olarak yine bir maliyet yükü doğurur. Bu yüzden, maliyeti azaltmaya yönelik farklı stratejiler geliştirilmiştir. Havuz dizileme ve düşük kapsamlı dizileme bu amaçla geliştirilen stratejilerdendir. Yapılan literatür araştırmaları sırasında, her iki strateji de ayrı ayrı olmak üzere farklı çalışmalarda çalışılsa da, her ikisinin bir hastalık verisi üzerinde bir arada kullanıldığı bir çalışmaya rastlanılmamıştır. Bu tez kapsamında, iki stratejinin bir arada kullanılarak üretildiği bir retinoblastom verisi üzerinde YND veri analizinin çalışılması hedeflenmektedir. Bu amaçla, havuz dizileme ve düşük kapsamlı dizileme verilerinin analizine uygun bir ardışık düzen geliştirilecektir. Bunun için, hem havuz dizilemede hem de düşük kapsamlı dizileme verilerinde çalışabilir araçlar seçilecektir. Bu ardışık düzenin performansının değerlendirilebilmesi için, standart bir ardışık düzenle karşılaştırılarak,

sonuçların performans metrikleri açısından değerlendirilmesi yapılacaktır. Bu değerlendirme mümkün olduğunca tez kapsamında kullanılan hastalık verisine benzer özelliklere sahip veriler üzerinde gerçekleştirilecektir.

YND veri analizinde en maliyetli adımlardan biri olan hizalama adımı için geliştirilen araçlarda bu maliyeti azaltmak adına birden fazla iş parçacığı (thread) kullanımına dayanan yaklaşım benimsenmiştir. Bu maliyeti azaltmak adına, Grafik İşlem Birimi (Graphics Processing Unit, GPU) üzerinde çalışabilen hizalama araçları da geliştirilmiştir. Bunlar, hizalamada yaygın kullanılan bazı araçların GPU'ya uyarlanmış halleridir. Bu çalışmada, geliştirilen ardışık düzende kullanılan hizalama algoritmasının GPU'da çalışır versiyonu ile hizalama için gerek duyulan hız ve istenen doğruluğun sağlanıp sağlanamayacağı farklı veri setleri üzerinde çalışılarak karşılaştırılacak ve sonuçlar değerlendirilecektir. Bunun için, tez kapsamında kullanılan veriler dışında daha yüksek kapsama ile dizilenmiş, farklı okuma uzunluğuna ve okuma sayısına sahip verilerden de yararlanılacaktır.

1.3 Orijinal Katkı

YND teknolojisindeki gelişmeler, bu teknolojinin popüler olmasını ve eskiye göre maliyetlerinin azalmasına sebep olmuştur. Öte yandan, hâlâ belli bir maliyet yükü olan YND çalışmalarındaki bu maliyeti azaltmak için havuz dizileme ve düşük kapsamalı dizileme gibi farklı stratejiler üzerinde çalışılmaya başlanmıştır. Bu iki stratejiye yönelik yapılan literatür araştırmaları sonucunda bunların farklı çalışmalarda ayrı ayrı kullanılmasına karşın, bu iki teknikle üretilen gerçek bir hastalık verisi üzerinde YND veri analizi adımlarının çalışmadığı görülmüştür. Bu iki stratejiyle üretilen retinoblastom verisi üzerinde gerçekleştirilen bu çalışma bildiğimiz kadarıyla bu yönde gerçekleşen ilk çalışmadır. Tez kapsamında, havuz dizileme ve düşük kapsamalı dizileme verileri için YND veri adımlarının çalışıldığı bir ardışık düzen geliştirilmiştir.

Retinoblastom üzerinde yapılan veri araştırmaları sonucunda, hastalığın verisine ulaşmanın çok kolay olmadığı görülmüştür. Bunun sebeplerinden birinin, retinoblastomun çocukluk çağı kanseri olması ve bu nedenle de verinin bir erişkin kanser verisi kadar sık olmadığı söylenebilir. Öte yandan, retinoblastom nadir görülen

bir kanser olmakla beraber, çocuklarda en sık görülen göz içi tümörüne bağlı oluşan bir rahatsızlıktır. Son olarak, her verinin herkese açık bir şekilde yayımlanmaması da veriye kolay ulaşılamama sebepleri arasındadır. Tez kapsamında geliştirilen ardışık düzenle, maliyeti azaltmaya yönelik stratejilerle üretilen veriler üzerinde etkin veri analizi yöntemleri çalışılırsa, klinik çalışmalarda bu stratejilerle veri üretimi için de bir motivasyon oluşturulabileceği düşünülmektedir. Dolayısıyla bu durumun retinoblastom gibi veri kısıtlı olan kanser türlerindeki veri üretimi ve paylaşımı konusunda olumlu bir katkı sağlayabileceği düşünülmektedir.

1.4 Tez Organizasyonu

Tez kapsamında retinoblastom hastalığına, YND'ye, YND veri analizi ve araçlarına, havuz dizilemeye ve düşük kapsamlı dizilemeye dair literatür araştırmasının, tezin amacının ve literatüre katkısının sunulduğu Birinci bölümün ardından, tez aşağıdaki bölümlerden oluşmaktadır:

İkinci Bölüm, tez içerisinde geçen bazı teknik terimlere dair tanımlamalar ve görseller içermektedir.

Üçüncü bölümde, hastalık gen listesinin oluşturulmasına dayalı yapılan araştırma, geliştirilen ardışık düzenin gereksinimleri ve tüm adımlarının detayları sunulmaktadır. Ayrıca, çalışma içerisinde kullanılan verilere dair bilgiler de bu bölümde yer almaktadır.

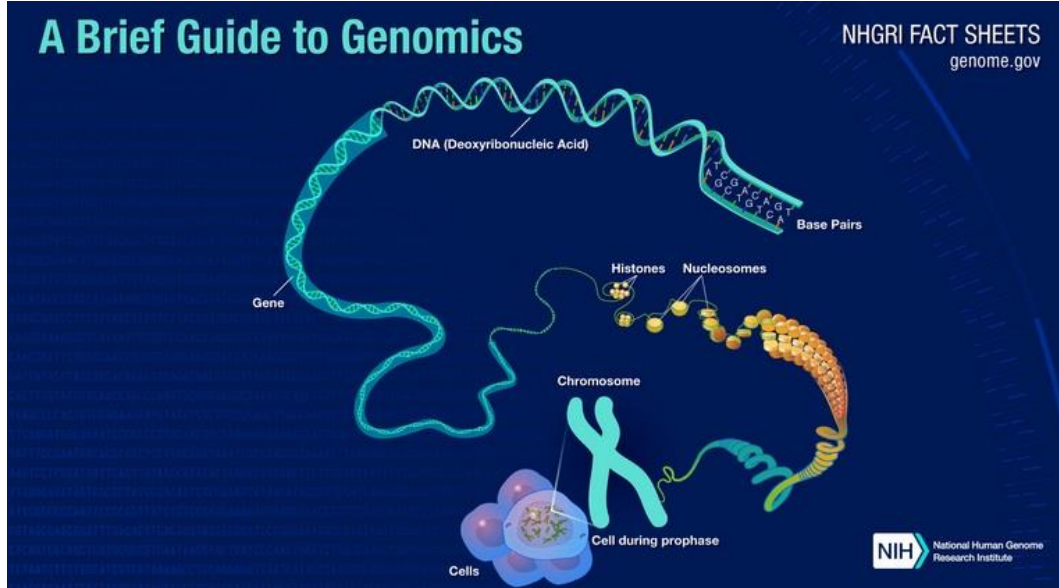
Dördüncü bölüm, geliştirilen ardışık düzen kapsamında elde edilen tüm sonuçları içermektedir. Ayrıca, CUDA ile yapılan çalışmanın detayları ve GPU üzerindeki sonuçları da bu bölümde sunulmaktadır.

Beşinci bölüm, çalışmanın sonuçlarının kısa özetini ve bu kapsamda yapılabilecek çalışmalar hakkında öneriler içermektedir.

BÖLÜM 2

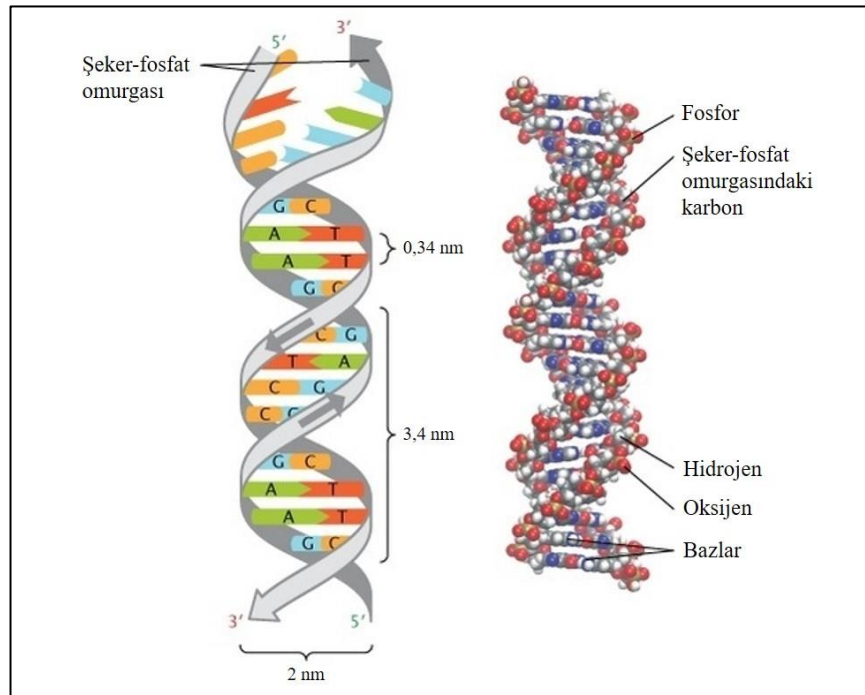
ÖN BİLGİ

- **Genom:** Bir canlının genetik bilgisinin tümüdür.
- **Deoksiribonükleik asit (Deoxyribonucleic acid, DNA):** Canlının genetik talimatlarını içeren bir moleküldür (Şekil 2.1).
- **Kromozom:** DNA'nın histon (histone) denen proteinlerin etrafına birçok kez sarılmasıyla oluşan yapılardır (Şekil 2.1). İnsan genomunda 46 çift kromozom bulunmaktadır.
- **Gen:** Kalıtımın temel fiziksel birimi olup, DNA'dan oluşur (Şekil 2.1). Protein oluşturmak için gerekli bilgileri içerir. İnsan genomunda yaklaşık 20.000 ile 25.000 arasında gen olduğu tahmin edilmektedir [73].
- **Lokus:** Bir genin veya DNA dizisinin kromozom üzerindeki fiziksel yerini ifade eder.



Şekil 2. 1: Canlıların genomlarını inceleyen genomik disiplinin temel birimleri [73]. (Base Pairs: Baz Çiftleri, Gene: Gen, Histones: Histonlar, Nucleosomes: Nükleozomlar, Chromosome: Kromozom, Cell during prophase: Ön faz esnasındaki hücre, Cells: Hücreler)

- İnsan DNA'sı birbiri üzerine sarılı çift sarmal bir yapıdan oluşur. Bu yapıyı oluşturan iplikler ise nükleotidlerden oluşur. Her bir nükleotid, azotlu baz, beş karbonlu şeker ve fosfat grubundan oluşur. Şeker ve fosfat grubu DNA'nın omurgasını (backbone) oluşturur. Adenin (Adenine, A), Guanin (Guanine, G), Timin (Thymine, T) ve Sitozin (Cytosine, C) olmak üzere dört baz vardır. Adenin ve Timin bir baz çifti oluştururken, Sitozin ve Guanin de başka bir baz çifti oluşturur. Sarmalı oluşturan iplikler birbirine zıt yönde uzanırlar ve her ipliğin fosfat ucu 5', şeker ucu ise 3' dir. DNA'nın çift sarmal yapısı Şekil 2.2'de verilmiştir. Şeklin sol tarafındaki gri kısımlar şeker-fosfat omurgasını, renkli parçalar ise bazları (G, C, A, T) göstermektedir. Sağ tarafta, DNA sarmalının moleküler yapısı gösterilmiş olup, fosfor, karbon, hidrojen, oksijen ve azot atomları sırasıyla altın, gri, beyaz, kırmızı ve mavi kürelerle temsil edilmektedir. İki baz çifti arasındaki uzaklık 0,34 nanometre (nm), çift sarmalın bir dönüşünün uzunluğu 3,4 nm, DNA molekülünün genişliği ise, 2 nm'dir.



Şekil 2. 2: DNA'nın çift sarmal yapısının gösterimi [74]. (Şekil, Kaynak [74]'den alıntılanarak Türkçeleştirilmiştir.)

- İnsan Genom Projesi: 1990 yılında başlayıp 2003 yılında tamamlanan, DNA'daki bazların sırasının belirlenmesi ve gen haritasının çıkarılması amacıyla yürütülmüş olan uluslararası bilimsel bir projedir. İnsan Genom Projesi ile insanların DNA'sının %99'undan da fazlasının aynı olduğu ortaya çıkmıştır. İnsan genomu yaklaşık olarak üç milyar baz çiftinden oluşmaktadır.
- Ekzon: Gen üzerinde protein kodlayan bölümlere verilen isimdir. Genom üzerindeki ekzonların tümü 'ekzom' olarak isimlendirilir.
- İntron: Gen üzerinde protein kodlamayan bölümlere verilen isimdir.
- Genetik varyasyon: Bir popülasyon içinde bulunan bireylerin DNA dizilerindeki farklılıklardır ve aşağıdaki çeşitleri vardır [75]:
 - Tek Nükleotid Polimorfizmi – TNP (Single Nucleotide Polymorphism – SNP): DNA dizisinde tek bir nükleotidin değişmesi sonucu oluşur. Şekil 2.3'de bir TNP görülmekte olup, referans genomda G'nin olduğu pozisyonda incelenen genom için T bulunmaktadır.

Referans DNA	ACTGACGCATGCATCATGCATGC
TNP'li DNA	ACTGACGCATGCATCATTCATGC

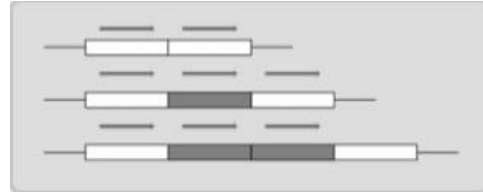
Şekil 2. 3: TNP örnekleme [75]. (Şekil, Kaynak [75]'den alıntılanarak Türkçeleştirilmiştir.)

- (indel, İnsersiyon – Delesyon) (indel, Insertion-Deletion): DNA dizisi üzerinde iki ile yüzlerce baz çifti uzunluğundaki bir bölgenin eklenmesi ya da silinmesidir [75]. Şekil 2.4'de referans genom ve insersiyon ve delesyon örneklerinin mevcut olduğu iki farklı DNA dizisi görülmektedir. İnsersiyon için, G ve C bazları arasına 3 bazlık bir eklenme, delesyon için ise, CA bazlarının olmadığı yerde iki bazlık bir silinme örneği gösterilmektedir.

Referans	ACTGACGCATGCATCATGCATGC	} indel
İnsersiyon	ACTGACGCATG GTAC ATCATGCATGC	
Delesyon	ACTGACG -- TGCATCATGCATGC	

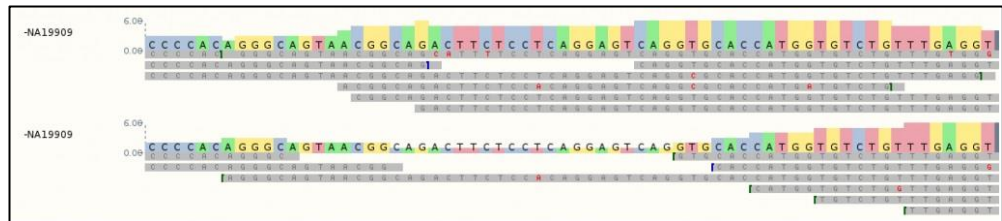
Şekil 2. 4: indel örnekleme [75]. (Şekil, Kaynak [75]'den alıntılanarak Türkçeleştirilmiştir.)

- Yapısal varyasyon: Daha büyük bir DNA dizisi üzerinde oluşan genetik varyasyonlar için kullanılır [75]. Belli bir genin kopya sayısının bireyden bireye değişmesi olarak tanımlanan kopya sayısı değişiklikleri (CNV) [76] yapısal varyasyon içine girer. Şekil 2.5'de CNV örneği görülmektedir.



Şekil 2. 5: CNV örneği [75]

- Varyant: DNA dizileri arasındaki farklılıkları, örneğin, TNP ve indel, ifade etmek için kullanılır. Şekil 2.6'da kırmızı ile belirtilenler varyantları gösterir.

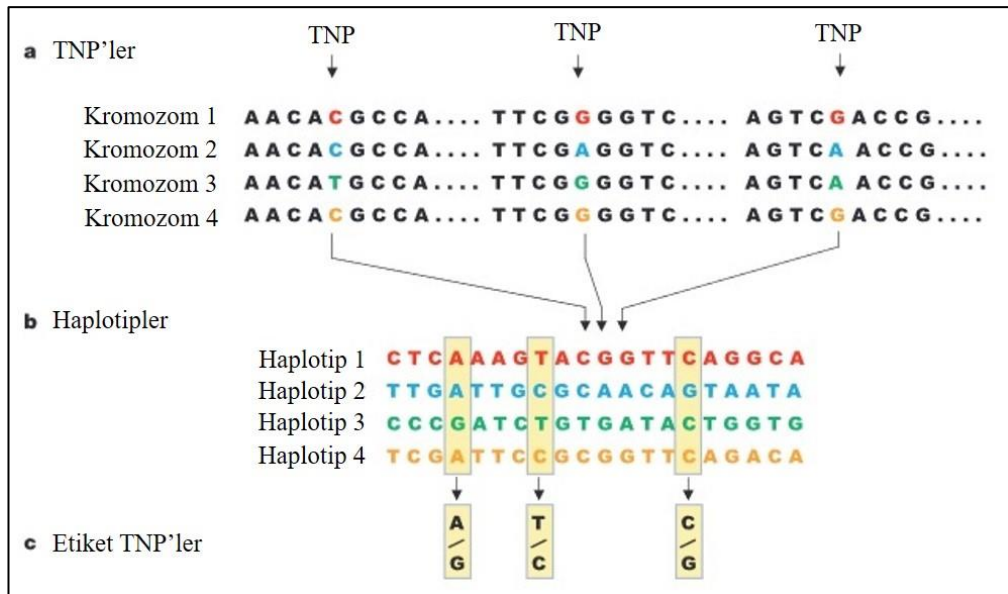


Şekil 2. 6: Varyant örnekleme [75]

- Allel: Aynı varyantın farklı versiyonları için kullanılır [75]. Referans allel varyantın tanımlandığı bölgede referans genom üzerinde bulunan bazı,

alternatif (alternative) alel ise referans genom dışındaki DNA dizisinde varyantın tanımlandığı bölgede bulunan bazı ifade eder.

- Haplotip (Haplotype): Birlikte kalıtılan DNA varyasyonu ya da polimorfizm grubudur [76]. Şekil 2.7 (a)'da dört farklı insandaki aynı kromozom bölgesi üzerinde bulunan üç tane TNP görülmektedir. Şekil 2.7 (b)'de yanyana TNP'lerdeki alellerin kombinasyonu ile elde edilen haplotip bloğu ve Şekil 2.7 (a)'da verilen 3 TNP görülmektedir. Burada, 20 tane TNP'den oluşan bir blok örnek olarak verilmiştir. Şekil 2.7 (c)'de ise 20 TNP'den oluşan bloğun 3 tanesinin 4 haplotipin tanımlanabilmesi için yeterli olduğu belirtilmektedir.



Şekil 2. 7: (a) TNP'ler, (b) haplotipler, (c) etiket (tag) TNP'ler [77]. (Şekil, Kaynak [77]'den alıntılanarak Türkçeleştirilmiştir.)

- YND veri analizi kapsamında kullanılan ve standart hale gelmiş dosya formatları şu şekildedir:
 - FASTA: Nükleotid dizileri için yaygın olarak kullanılan metin tabanlı bir dosya formatıdır.
 - FASTQ: YND ile üretilen ham veriyi gösteren metin tabanlı bir dosya formatıdır. Detayları Bölüm 3'de verilecektir.

- Sequence Alignment Map (SAM): Biyolojik dizilerin bir referans genomla hizalanması sonucu elde edilen metin tabanlı bir dosya formatıdır.
- Binary Alignment Map (BAM): SAM dosyasının ikili dosya formatıdır.
- Variant Call Format (VCF): Varyant verilerini saklamak için kullanılan metin tabanlı bir dosya formatıdır. Şekil 2.8’de örnek bir VCF dosyası verilmiştir. Dosya içerisinde, her varyant için bir satır bilgi tutulur. Her bir satır içerisinde, sırasıyla olmak üzere, varyantın bulunduğu kromozom (#CHROM), kromozom üzerindeki yerinin başlangıç koordinatı (POS), varyant tanımlayıcı (ID), referans alel (REF), alternatif alel (ALT), kalite skoru (QUAL), filtrelemeyi geçip geçmediği bilgisi (FILTER), varyant ile ilgili çeşitli bilgilerin tutulduğu bilgi kısmı (INFO), sonraki sütunların formatı (FORMAT), örneklere dair tanımlayıcı bilgiler tutulur [75].

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA19909
11	5248232	rs334	T	A	100	PASS	AA=T ;AC=1;AF=0.0273562;AFR_AF=0.0998;AMR_AF=0.0072;AN=2;DP=22876;EAS_AF=0;EUR_AF=0;EX_TARGET;NS=2504;SAS_AF=0;VT=SNP	GT	0 1

Şekil 2. 8: VCF dosyası örneği [75]

BÖLÜM 3

GELİŞTİRİLEN ARDIŞIK DÜZEN

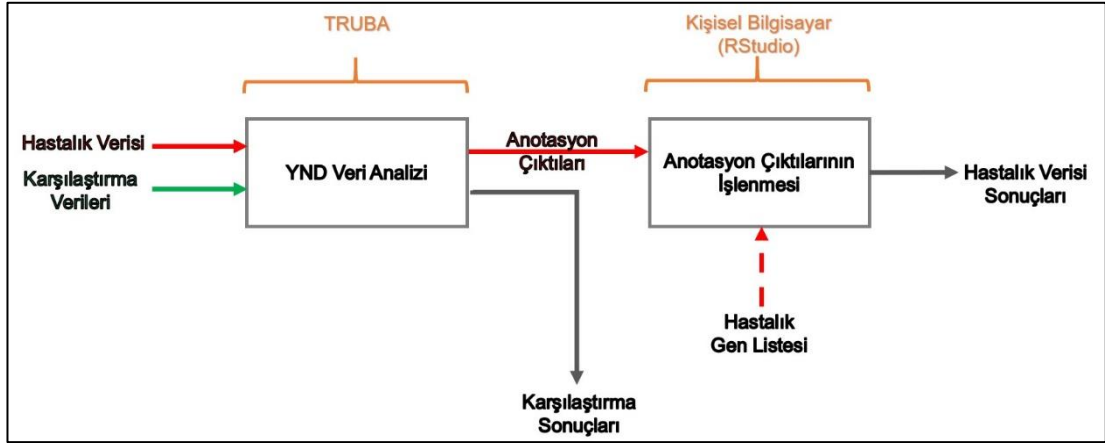
3.1 Hastalık Gen Listesinin Oluşturulması

Retinoblastom ile ilişkili gen listesini belirleyebilmek için retinoblastomda YND çalışmalarına yönelik bir literatür araştırması yapılmıştır. Araştırma sonuçlarına göre, bazı çalışmalar [78-80] YND ile varyantların tespitine yönelik iken, bazıları da genlere yönelik çalışmalardır. Buna göre, dört retinoblastom hastasının TGD ile incelendiği çalışmada retinoblastomun nispeten kararlı bir genom ve çok düşük mutasyon oranına sahip olduğu belirtilmiştir [81]. Ayrıca, SYK geninin retinoblastom için önemli bir gen olduğu da vurgulanmıştır. Öte yandan, retinoblastom hastalığındaki somatik kopya sayısı değişikliklerine yönelik bir çalışmada, iyi bilinen hastalık genleri olan RB1 ve MYCN'e ek olarak CRB1, NEK7, SOX4, NUP205, MIR181 ve DEK genleri yeni aday genler olarak tanımlanmıştır [82]. Retinoblastom üzerine hazırlanan bir derleme makalesinde KIF14, MDM4, MYCN, DEK, E2F3, CDH11 ve SYK genlerinin hastalığın ilerlemesinde önemli oldukları vurgulanmıştır [83]. Somatik varyantlarla ilgili bir çalışmada ise, RB1 genine ek olarak, BCOR ve CREBBP genlerinde tekrarlı olarak (en az iki farklı hastada olan) bulunan patojenik (hastalığa neden olan) mutasyonlar tespit edilmiştir [84]. Bu çalışmada da, diğer çalışmada [81] olduğu gibi retinoblastomun çok düşük mutasyon oranına sahip olduğu vurgulanmıştır. Bu bölümde adı geçen genler tez çalışmasında kullanılacak olup, sonraki bölümlerde 'hastalık gen listesi' olarak çağrılacaktır.

3.2 Gereksinimler

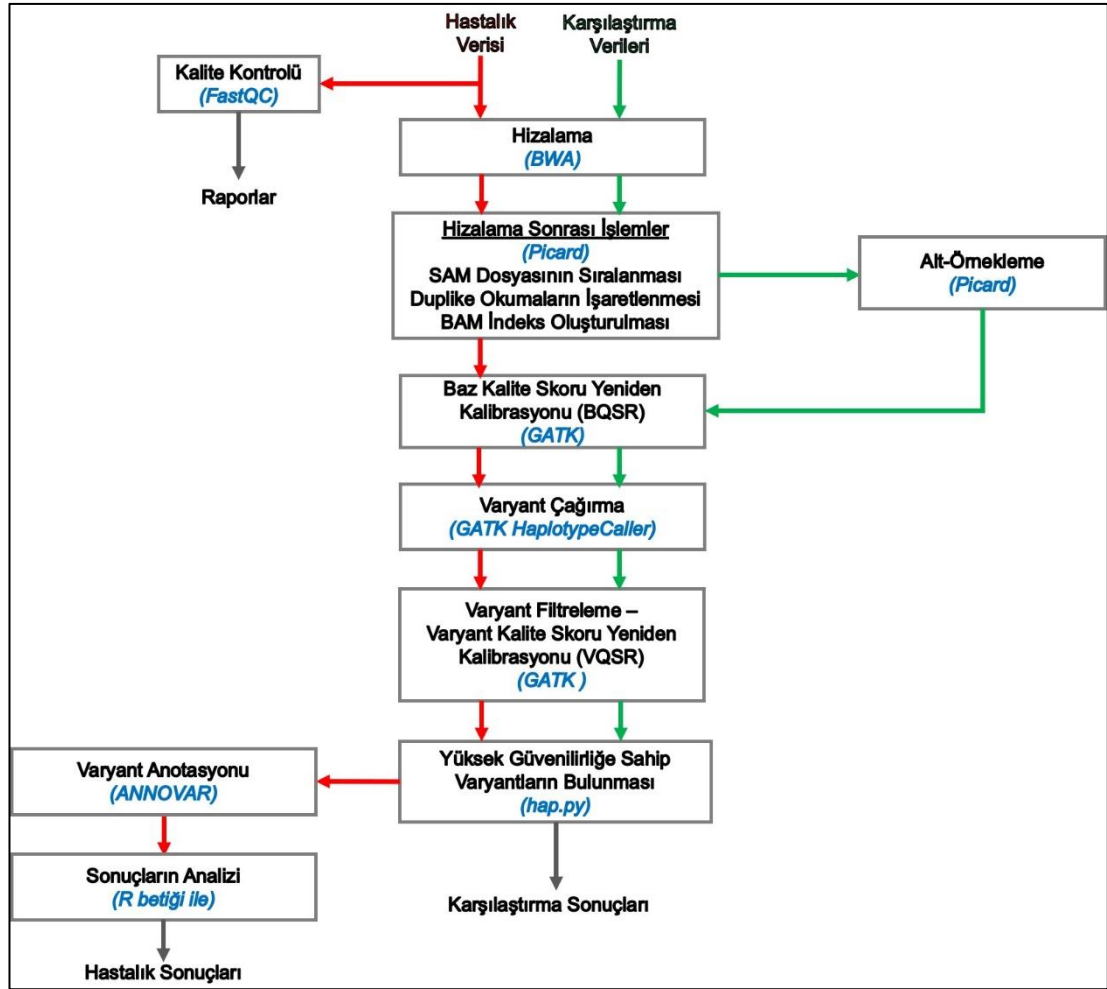
Tez kapsamında, havuz dizileme ve düşük kapsamalı dizileme verilerinin analizi için bir ardışık düzen geliştirilmiştir. Bu ardışık düzen retinoblastom hastalık verisi üzerinde çalışılmıştır. Geliştirilen ardışık düzenin, varsayılan parametrelerle çalıştırıldığı standart bir ardışık düzenle karşılaştırılarak performansının değerlendirilebilmesi için, bu iki stratejiyle dizilenmiş karşılaştırma verilerine ihtiyaç duyulmaktadır. Yapılan araştırma sonrasında, bu iki stratejinin bir arada kullanıldığı

gerçek veriye rastlanılmadığından, sadece düşük kapsamalı dizileme ile dizilenen veriler ardışık düzenlerin karşılaştırılması için kullanılmıştır. Buna göre, geliştirilen ardışık düzen, YND veri analizi ve Anotasyon çıktıların işlenmesi olmak üzere iki aşamadan oluşur (Şekil 3.1).



Şekil 3. 1: Geliştirilen ardışık düzenin özeti

İlk aşama, hastalık ve karşılaştırma verileri üzerinde YND adımlarının çalışıldığı veri analizi kısmıdır. Bu kısım için takip edilen iş akışı Şekil 3.2’de sunulmaktadır. YND veri analizi, büyük miktarda veri içerdiğinden bu verilerin analizini yapabilmek için etkin bir şekilde hesaplama yapabilen, güvenli sistemlere ihtiyaç vardır. Bu tezde de, sürecin yüksek hafıza gereksinimi nedeniyle veri analizi adımları TRUBA altyapısında gerçekleştirilmiştir. TRUBA, ülkemizdeki akademisyenlerin, öğrencilerin, araştırma gruplarının yüksek başarılı hesaplama, veri yoğun hesaplama, bilimsel veri ambarları ve bulut hesaplama konularındaki sistem gereksinimlerini karşılamak amacıyla TÜBİTAK ULAKBİM Yüksek Başarılı Hesaplama Grubu tarafından işletilen bir ulusal e-altyapıdır [85]. Şu an itibarıyla, sistemde ~19000 işlemci çekirdeği, 36 adet GPU bulunmaktadır [85]. Ardışık düzenin YND veri analizi kısmı, Şekil 3.2’de adı geçen araçların TRUBA ortamına kurulmasından sonra, yazılan kabuk betikleriyle yine TRUBA’da gerçekleştirilmektedir.



Şekil 3. 2: YND veri analizi için takip edilen iş akışı

Veri analizi süreci tüm veriler ile çalışılarak (Şekil 3.2), elde edilen varyantlar yüksek güvenilirliğe sahip varyant veritabanları ile karşılaştırılarak, ardışık düzenin performansı çeşitli metriklerle değerlendirilmektedir. Öte yandan, hastalık verisinden elde edilen ve yüksek güvenilirliğe sahip varyant veritabanlarında yer alan varyantlar ile anotasyon adımı çalışılarak, elde edilen çıktılar TRUBA ortamından kişisel bilgisayara aktarılmaktadır. Hastalık verisinden gelen bu çıktıların değerlendirilme süreci R dilinde yazılan kod ile gerçekleştirilmektedir.

Geliştirilen ardışık düzen kapsamında kullanılan tüm araçlar/programlama dili/geliştirme ortamı, sürümleri ve ilgili referanslar Tablo 3.1’de sunulmaktadır. Araçlar ve kullanıldıkları veri analizi adımları Bölüm 3.4’te detaylı olarak anlatılacaktır.

Tablo 3. 1: Geliştirilen ardışık düzende kullanılan araç/programlama dili/geliştirme ortamı listesi

Araç/Programlama Dili/Geliştirme Ortamı	Sürüm	Referans
FastQC	0.11.7	[86]
BWA	0.7.17	[87, 88]
SAMtools	1.8	[89]
Picard	2.18.7	[90]
GATK	4.0.5.1	[91]
ANNOVAR	(2018Apr16)	[92]
hap.py		[93]
R	3.6.0	[94]
RStudio	1.2.1335	[95]

3.3 Kullanılan Veriler

Ardışık düzenin test edilebilmesi için retinoblastom hastalığına ait yayımlanmış çalışma verisine/verilerine ihtiyaç vardır. Bu yönde yapılan araştırma sonucunda ulaşılan veriler Tablo 3.2’de listelenmiştir. Tez kapsamında kullanılanlar yeşil olarak belirtilmiş olup, verilere ulaşım için şu adresler kullanılabilir: (ENA: <https://www.ebi.ac.uk/ena/browser/home> GEO: <https://www.ncbi.nlm.nih.gov/gds>, ArrayExpress: <https://www.ebi.ac.uk/arrayexpress/>, dbGaP: <https://www.ncbi.nlm.nih.gov/gap/>)

Tablo 3. 2: Retinoblastom veri araştırması sonuçları. Tez kapsamında kullanılan veriler yeşil ile belirtilmiştir.

Erişim Numarası	Türü
ERR550406 ERR550407 (ENA-PRJEB6630)	Retinoblastoma tumors NGS data
GSE11488 (GEO)	Expression profiling by array
E-MTAB-3515 (ArrayExpress)	DNA-seq, genotyping design
E-MTAB-3492 (ArrayExpress)	comparative genomic hybridization by array, case control design
GSE84747 (GEO)	Non-coding RNA profiling by array
E-MTAB-4977 (ArrayExpress)	microRNA profiling by array
GSE7072 (GEO)	Non-coding RNA profiling by array
GSE41321 (GEO)	Non-coding RNA profiling by array
GSE29683 (GEO)	Expression profiling by array
GSE29684 (GEO)	Expression profiling by array
GSE33048 (GEO)	Expression profiling by array
GSE24673 (GEO)	Expression profiling by array
GSE7185 (GEO)	Expression profiling by array
phs000352.v1.p1 (dbGaP)	Case Set, Tumor vs. Matched-Normal, Whole Genome Sequencing
GSE5222 (GEO)	Expression profiling by array

Tablodaki veriler içinde, DNA dizileme verisi olan ve açık bir şekilde ulaşılabilir olan verilerden belirtilen iki veri bu tez kapsamında kullanılmıştır. Bunlardan ilki, Meksika'daki Hospital de Pediatría ve Hospital Infantil de México hastanesinden alınan [96] *PRJEB6630* erişim numaralı [97] veridir. *PRJEB6630* çalışması kapsamında retinoblastom haricinde medulloblastom (kötü huylu beyin tümörü) verisi

de mevcut olup, bu tezde sadece retinoblastom verisi kullanılmaktadır. Retinoblastom verisinin temel özellikleri aşağıda listelenmiştir:

- Illumina platformu tarafından üretilen tüm genom dizileme verisidir.
- Sadece retinoblastom tümörlerinden elde edilmiştir.
- Havuz dizileme verisidir.
 - Her bir havuz içinde 2'si kız, 2'si erkek çocuk olmak üzere 8 çocuğa ait veri iki havuz verisi (RB-H1, RB-H2) olarak dizilenmiştir.
- Düşük kapsamalı dizileme verisidir.
 - Bu veri ile yapılan çalışma kapsamında *kapsama derinliğinin* 1X'den küçük olduğu belirtilmiş olup, her bir havuz için *kapsama derinliği* (c) Denklem 3.1'e göre hesaplanmıştır [98]:

$$c = (m \times n)/q \quad (3.1)$$

Burada, m her örnek için okumaların sayısı, n okumaların ortalama uzunluğu, q ise genomun uzunluğudur. Genomun ortalama uzunluğu, 2.988.355.349 olarak alınmıştır [98]. Okumaların ortalama uzunluğu 36, RB-H1 ve RB-H2'deki okumaların sayısı ise sırayla 31.019.035 ve 33.895.793 olduğundan, Denklem 3.1'e göre, RB-H1'in c değeri 0,37; RB-H2'nin c değeri ise 0,41 olarak hesaplanır.

Bahsedilen bu veri, ilerleyen bölümlerde 'hastalık verisi' olarak çağrılacak olup veri ile ilgili özet bilgiler Tablo 3.3'de sunulmaktadır. Geliştirilen ardışık düzeni test edebilmek adına, genel olarak ulaşılabilir hastalık verilerinden bir diğeri olan hedefli yeniden dizileme verisi E-MTAB-3515 de bu tez kapsamında kullanılmış olup, bu veri ise ilerleyen bölümlerde 'hastalık test verisi' olarak çağrılacaktır.

Tablo 3. 3: Hastalık verisinin özeti

	RB-H1	RB-H2
Veri Adı	ERR550406.fastq	ERR550407.fastq
Büyüküğü	4.0 GB	4.4 GB
Okuma Sayısı	31019035	33895793
Okuma Uzunluğu	36	36

Tez kapsamında kullanılan hastalık verisi havuz dizileme ve düşük kapsamlı dizileme stratejileriyle dizilenmiştir. Retinoblastom veri araştırması sürecinde bu iki stratejiyle dizilenmiş başka bir veriye rastlanılmamıştır (Tablo 3.2). Her iki stratejinin bir arada kullanılmasıyla üretilen başka hastalık verisi sorgusu ise NCBI – SRA [99] sayfasından yapılmıştır. Buna göre, insan genomu için bu iki stratejinin bir arada kullanıldığı bir hastalık verisiyle karşılaşılmamıştır. Ancak, geliştirilen ardışık düzenin etkinliğinin ve performansının araştırılabilmesi için ek verilere ihtiyaç vardır. Bunun için, 1000 Genom Projesi’nden [100] düşük kapsamlı dizileme stratejisi ile üretilen veriler üzerine yoğunlaşmıştır. 1000 Genom Projesi, çok sayıda insandan alınan dizileme verileri ile düşük frekansa sahip varyantların belirlenmesini hedefleyen bir proje olup, bu proje ile genetik varyasyonlar için geniş kapsamlı bir kaynak oluşturulmuştur [101]. Seçilen ilk veri, NA12878 genomuna ait dizileme verisinin *kapsam* değeri 5X olacak şekilde alt örneklemeyle elde edilen veridir [102]. Diğer veri içinse, NCBI SRA veritabanında ‘low-coverage child’ arama parametreleri ile arama yapıp, *fastq* formatına sahip veriler filtrelendiğinde üç farklı genoma ait verinin olduğu görülmüştür. Bu veriler içinden NA20355 [103] no’lu genom seçilmiştir. Seçilen bu iki veri ilerleyen bölümlerde ‘karşılaştırma verileri’ olarak çağrılacaktır.

3.4 Ardışık Düzen

3.4.1 Referans Genom

YND veri analizi sürecinde kullanılan veri dışında çalışılan canlı türü için referans genomun mevcut olduğu durumlarda, referans genomun dizisine ihtiyaç duyulmaktadır. Tez kapsamında insan genomu için yayımlanmış referans genom kullanılmaktadır. İnsan genomu için yayımlanan son iki referans genom, Şubat 2009’da yayımlanmış hg19 (GRCh37) genomu ve Aralık 2013’te yayımlanmış hg38 (GRCh38) genomu olduğundan, tezde daha yeni olan hg38 ile çalışılmaktadır. hg38 için farklı sürümler ve versiyonlar bulunmakla beraber, analizler için uygun olan

genomun ‘GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz’ dosya isimli genom olduğu belirtildiğinden [104], tezde bu versiyon kullanılmaktadır.

hg38 genomu kapsamında verinin içeriği şu şekildedir [105]:

- *chr1-chr22* → 1-22 kromozomlar
- *chrX* → X kromozomu
- *chrY* → Y kromozomu
- *chrM* → Mitokondriyal DNA
- *_random* → Belli bir kromozom üzerinde olup, sırası tam belli olmayanlar
- *chrU_* → Bulunduğu kromozom bilinmeyenler

3.4.2 Kalite Kontrolü

YND veri analizinde, verinin çalışmayı yapan grup tarafından üretildiği durumlarda veri analizinin ilk adımı kalite kontrolüdür. Eğer, analiz çalışmalarında yayımlanmış veri kullanılırsa, bu veri yayımlanmadan önce kalite kontrolünden geçirildiğinden ve kontrolü geçmesi gerektiğinden bu adım atlanabilir. Tez çalışması içinde, sadece hastalık verisi ile kalite kontrolü çalışılmıştır.

Dizileme verileri, metin tabanlı *fastq* formatındaki dosyalarda, ‘okuma’ adı verilen ve kullanılan dizileme cihazına bağlı olarak farklı uzunlukta parçalar halinde tutulurlar. Bir *fastq* dosyasında her bir okuma için dört satır tutulur. Şekil 3.3’de RB-H1 verisindeki ilk beş okuma ve her okuma için tutulan bilgiler görülmektedir. Buna göre, her bir okuma için ilk satırda ‘@’ karakterinin ardından dizi tanımlayıcı, ikinci satırda okunan bazlar, üçüncü satırda ‘+’ karakteri (bazı verilerde devamında dizi tanımlayıcı görülebilir), dördüncü satırda ise kalite skorları tutulmaktadır. Kalite skorlarının etkin bir şekilde tutulabilmesi için *ASCII* (American Standard Code for Information Interchange) karakterleri kullanılır.

	-bash-4.2\$ head -20 ERR550406.fastq	
okuma_1	@ERR550406.1 HWUSI-EAS636_0014_FC:7:1:2597:1101#0/1	Dizi tanımlayıcı
	TCACGCCCGTAATCCAGCACTTTGGGAGGTGGAGG	Okunan bazlar
	+	
	GHHHHHHHHDGEGGFDE<EGGGGEHGH;9DCCCC	Kalite skorları
okuma_2	@ERR550406.2 HWUSI-EAS636_0014_FC:7:1:2739:1101#0/1	
	ATTCCATTCCATTCCATTCCGGATGATTCCATTCCA	
	+	
	BE:BEDEEEEEEGGEBG?BGG@GGDGDGGDBGGGB	
okuma_3	@ERR550406.3 HWUSI-EAS636_0014_FC:7:1:2893:1104#0/1	
	TCTCCAATAAAATACAAAATTAGGTGGGCGTGGT	
	+	
	FHHHDHGDHHDGDHGEHGGDGBGG@EGGD@@?F<?	
okuma_4	@ERR550406.4 HWUSI-EAS636_0014_FC:7:1:2926:1102#0/1	
	TCTAATAGGAAACCCTGGTTTTTAAGTTGTTTTCAA	
	+	
	=9?,04-?-@;:.3.7A?A::C>CD:FG?GC@@9C>	
okuma_5	@ERR550406.5 HWUSI-EAS636_0014_FC:7:1:2945:1106#0/1	
	GAGGAGATTGTGCCCAACTTGGCTGTGCAAGGGGC	
	+	
	8==8>?AA@@<BE?B;BBBBG?4<:EB:E<@=AB##	

Şekil 3. 3: Örnek fastq formatı ve açıklamaları

Okunan her bir nükleotidin kalitesini değerlendirebilmek için, dizileme cihazları tarafından o nükleotidin yanlış okunma olasılığı üzerinden elde edilen ve Phred kalite skoru olarak isimlendirilen bir değer atanır. Phred kalite skoru Denklem 3.2 ile tanımlanır [106]:

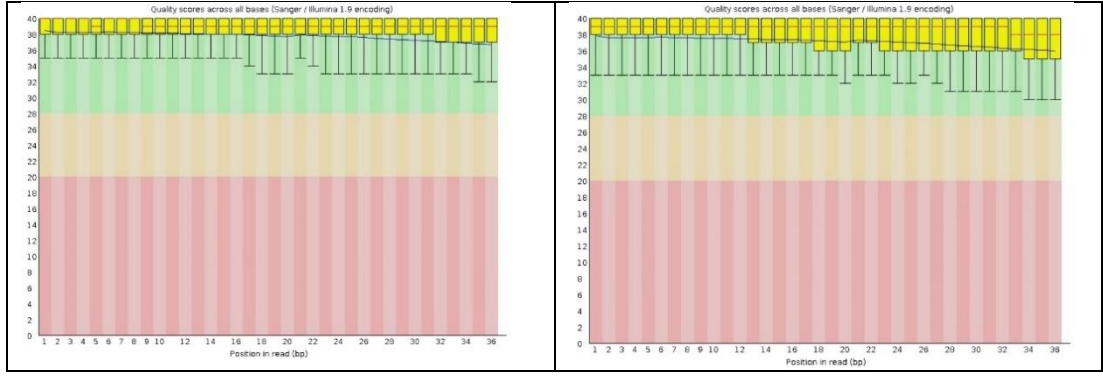
$$Q = -10 \log_{10} P \quad (3.2)$$

Burada, P nükleotidin yanlış okunma olasılığını, Q ise *Phred* kalite skorunu göstermektedir. Buna göre, P değeri 0,001 olan bir nükleotidin *Phred* kalite skoru 30 çıkar. Dizileme cihazının kullandığı kodlama sistemine göre bulunan değere belli bir değer eklenir. Örneğin, Sanger ya da Illumina 1.8+ *Phred*+33 kodlama sistemini kullandığından, bulunan değere 33 eklenir. Böylece, 63 elde edilir. *ASCII* kod tablosunda [107] 63 değeri ? karakterine karşılık geldiğinden, okunan nükleotidin kalite skoru için 0,001 kullanmak yerine ? kullanılır. *Phred* kalite skorları ve karşılık gelen doğruluk değerlerine göre (Tablo 3.4), *Phred* kalite skoru 30 olan bir nükleotidin %99,9 doğruluk ile okunduğu söylenebilir.

Tablo 3. 4: Kalite skorları ve karşılık gelen doğruluk değerleri [106]. (Tablo, Kaynak [106]'den alıntılanarak Türkçeleştirilmiştir.)

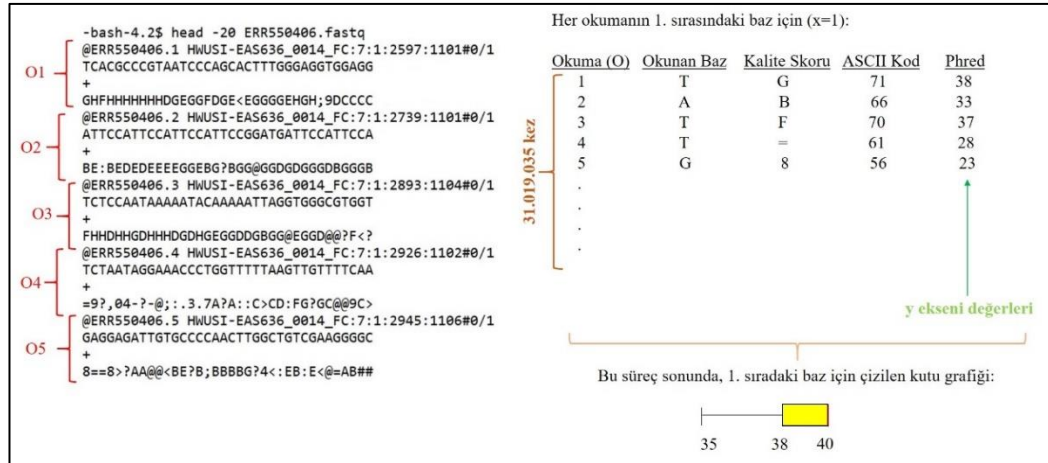
Phred Kalite Skoru (Q)	Nükleotidin yanlış okunma olasılığı (P)	Nükleotidin doğruluğu
10	0,1	%90
20	0,01	%99
30	0,001	%99,9
40	0,0001	%99,99
50	0,00001	%99,999

Çalışmada kullanılan hastalık verisi (RB-H1, RB-H2), FastQC aracı ile çalıştırılarak kalite kontrolü sonuçları incelenmiştir. FastQC, birçok YND verisiyle uyumlu çalışabilen ve verileri değerlendirebilmek için çeşitli çıktılar sunan bir kalite değerlendirme aracıdır [27]. Hastalık verisi ile çalışılan kalite kontrolü adımı sonunda, FastQC aracı tarafından farklı modüllere ait grafikler, bu modüllerin kalite kontrolünden geçip geçmediği ve ayrıca veriye dair özet bilgilerin mevcut olduğu bir rapor üretilmiştir. Buna göre, "Per base sequence content" ve "Per sequence GC content" modülleri hariç diğer tüm modüllerin kalite kontrolünü başarılı olarak geçtikleri görülmektedir. "Per base sequence quality" modülüne göre, hastalık verisi için üretilmiş grafikler Şekil 3.4'de sunulmaktadır. Burada, x eksenini okumadaki pozisyonları, y eksenini ise *Phred* kalite skorlarını göstermektedir. Hastalık verisinde okuma uzunluğunun 36 olması (Tablo 3.3) sebebi ile x eksenindeki değerler 1'den 36'ya kadardır. Grafikler üzerinde görülen üç farklı renkteki bölge kalite derecelerini gösterir. Buna göre, yeşil bölge çok iyi, sarı kabul edilebilir, kırmızı ise düşük kalite değerlerine karşılık gelmektedir. Ayrıca, grafik üzerinde görülen sarı kutu çeyrekler arasını, kırmızı çizgi medyan skorlarını, mavi çizgi ise ortalama skorları göstermektedir.



Şekil 3. 4: (Soldan sağa olmak üzere) RB-H1 ve RB-H2’ye ait kalite skorları grafiği. (FastQC aracı tarafından üretilmiştir)

Kalite kontrolü ve FastQC hakkında verilen genel bilgilerden sonra hastalık verisi için kalite skor grafiğinin nasıl oluştuğundan bahsedilecektir. Şekil 3.3’de RB-H1 verisindeki ilk 5 okumaya dair veri sunulmuştu. Her bir okuma uzunluğu 36 olduğundan Şekil 3.4’de 36 tane kutu grafiği görülmektedir. Yani, okumadaki her bir pozisyon için okunan bazların kalite skoru üzerinden bir kutu grafiği oluşturulmaktadır. Şekil 3.5’de okumanın birinci sırasındaki kutu grafiğinin oluşturulma süreci anlatılmaktadır ($x=1$). Bunun için, her okumanın birinci pozisyonundaki bazların *Phred* kalite skorları bulunmalıdır. Hastalık verisi Sanger/Illumina 1.9 (*Phred*+33) ile kodlandığından, ASCII tablosuna göre karakterlere karşılık gelen değerlerden 33 çıkarılarak *Phred* kalite skorları bulunur. Bu süreç, RB-H1 verisi için okumaların sayısı 31.019.035 olduğundan 31.019.035 kez tekrarlanır (Tablo 3.3). Bulunan *Phred* kalite skor değerlerine göre, birinci çeyrek (%25), üçüncü çeyrek (%75), alt bıyık (%10), üst bıyık (%90) ve medyan değerleri bulunarak kutu grafiği oluşturulur. Birinci sıradaki baz için yapılan bu süreç, okuma içerisindeki her baz için de ayrı ayrı yapılarak 36 farklı kutu grafiği elde edilir.



Şekil 3. 5: Kalite skor grafiğinin oluşma süreci

3.4.3 Hizalama

Hizalama için geliştirilen araçlardan bazıları Bölüm 1.1.4'teki Tablo 1.1'de sunulmuştu. Etkin bir hizalama yapabilmek için verinin özellikleri incelenerek, veriye uygun aracın seçilmesi gerekir. Tez kapsamında kullanılan hastalık verisi kısa okuma verisi (her bir okuma uzunluğu 36) olduğundan, kısa okumalar için geliştirilmiş hizalama araçlarına odaklanılmıştır. Kısa okumalar için geliştirilen araçların en bilinenleri, BWA, Bowtie2/Bowtie, SOAP, MAQ, Mosaik, Novoalign, Bfast'dir [27, 29, 108].

Hizalama algoritmalarının birçoğu referans genom ya da okunan okumalar için indeks adı verilen veri yapıları oluşturur [108]. İndeksin oluşma şekline göre, hizalama araçları temel olarak, hash tablosu kullananlar ve son ek ağacı (suffix tree) kullananlar olmak üzere ikiye ayrılırlar [108, 109]. Bunlardan hash tablosu kullananlar, daha yavaş ve daha hassasken, son ek ağacı kullananlar, hafızayı daha etkin kullanıp, daha hızlı çalışırlar [109]. BWA ve Bowtie/Bowtie2 araçları, son ek dizisini Burrows–Wheeler transform (BWT) ile birleştiren ve bu yönüyle daha etkili olan FM-index algoritmasını kullanır.

Geliştirilen ardışık düzen kapsamında, hem hızlı hem de etkili bir hizalama aracı olarak bilinen BWA (Burrows–Wheeler Alignment) [87] kullanılmıştır. BWA, *backtrack*, *SW* ve *mem* olmak üzere üç algorithmadan oluşur [110]. Bunlardan,

backtrack 100 baz çifti (bç) uzunluğuna kadar olan Illumina okumaları için, diğer ikisi ise daha uzun okumalar (70 bç ile 1Mbç arası) için kullanılır. *mem* algoritması, 70-100 bç uzunluğundaki okumalar için *backtrack*'e göre daha iyi performans vermektedir. *backtrack* algoritması iki adımda gerçekleşir: *aln* ile okumaların sonek dizisindeki koordinatları belirlenir, *samse/sampe* ile de bu koordinatların referans genom üzerindeki dönüşümleri yapılarak SAM dosyası üretilir. *samse* tek uçlu (single-end), *sampe* ise çift uçlu (paired-end) veriler için kullanılır. *mem* ise maksimum tam eşleşmeleri bulmakla başlayıp daha sonra Smith-Waterman algoritması ile bunları genişleterek hizalama sürecini gerçekleştirir ve *mem* alt komutu ile kullanılır. *SW* için *bwasw* kullanılır.

Hizalama için öncelikle 'bwa index' komutu ile referans genomun indekslenmesi gerekir. Hastalık verisi, tek uçlu ve okuma uzunluğu 36 olduğundan, hizalama için 'bwa aln' ve 'bwa samse' komutları kullanılmıştır. Karşılaştırma verileri ise, çift uçlu olup, okuma uzunlukları 100 ve 101 olduğundan, onlar için 'bwa mem' algoritması kullanılmıştır. BWA *aln* ve *mem*, birden fazla iş parçacığı ile çalışabilen bir araç olduğundan analizler sırasında 8 ve 16 adet iş parçacığı kullanılmıştır.

Hizalama adımı sırasında, varyant çağırma için gerekli olan okuma grubu (read group) ataması da gerçekleştirilmiştir. Bölüm 3.4.2'de bahsedildiği gibi, YND verileri *fastq* dosyası içerisinde dörtlü bloklar halinde tutulur ve bu blokların ilk sırasında dizi tanımlayıcıya dair bilgiler tutulmaktadır. Tablo 3.5'de, eski ve yeni Illumina *fastq* dosyasının ilk satırındaki bilgiler gösterilmektedir.

Tablo 3. 5: Illumina fastq dosya bilgileri [111]

Yeni	@<instrument>:<run number>:<flowcell ID>:<lane>:<tile>:<xpos>: <y-pos> <read>:<is filtered>:<control number>:<index>
Eski	@<machine_id>:<lane>:<tile>:<x_coord>:<y_coord>#<index>/<read>

Seçilen ve Bölüm 3.4.6'da bahsedilecek olan varyant çağırma aracı için gerekli olan ve tanımlanması gereken okuma grubu alanları şu şekildedir [112]:

- ID – Okuma grubu tanımlayıcı: (Illumina için) {FLOWCELL}.{LANE}
- PU - Platform Birimi: {FLOWCELL }.{LANE}.{SAMPLE}
- SM - Örnek: Her bir örnek için atanan isim.

- PL - Platform/Teknoloji: ILLUMINA, SOLID, HELICOS gibi.
- LB - DNA hazırlama kütüphane tanımlayıcısı: Duplike okumaları tanımlamak için bu alana ihtiyaç vardır.

Bu bilgiler doğrultusunda, uygun okuma grubu atamaları yapılmıştır. Şekil 3.6'da RB-H1 verisi için okuma grubu oluşturulma süreci görülmektedir. LB alanı için, verinin sayfasında [97] kütüphane adı 'UNSPECIFIED' olarak belirtildiği için *null* kullanılmıştır.



Şekil 3. 6: RB-H1 verisi için okuma grubu oluşturulması

Hizalama adımı sonunda, SAMtools aracı ile RB-H1 ve RB-H2 hastalık verilerinin hizalanma oranlarının sırasıyla % 97,64 ve %98,99 olduğu hesaplanmıştır.

3.4.4 Hizalama Sonrası İşlemler

Hizalama adımı ile oluşan dosyalar, yüksek verimli dizileme verilerini işleyen Java tabanlı *picard* aracı [90] ile aşağıda listelenen işlemlerin gerçekleştirilmesiyle varyant çağırma için hazır edilir:

- *SAM dosyasının sıralanması*; Hizalama adımı sonunda oluşan *SAM* dosyaları sıralanarak, ikili dosya formatı olan *BAM* formatında saklanır.
- *Duplike okumaların işaretlenmesi*; Bu adımda, duplike okumalar tanımlanarak varyant çağırma adımı göz ardı edilmeleri sağlanır. Bu amaçla, duplike okumalar onaltılık tabanda 0400 (0x0400) değeri ile işaretlenir [90].
- *BAM indeks oluşturulması*; Önceki adım sonunda oluşan *BAM* dosyaları indekslenir.

3.4.5 Alt-Örnekleme Süreci

Kullanılan her iki karşılaştırma verisi de düşük kapsamalı dizileme verisi olmakla beraber, etkin bir karşılaştırma yapabilmek adına tüm verilerin okuma sayıları mümkün olduğunca eş kılınmaya çalışılmıştır. Bunun için, hizalama ve sonrasındaki işlemlerin gerçekleştirildiği *BAM* dosyaları kullanılarak toplam okuma sayıları üzerinden *picard* aracındaki *DownsampleSam* ile alt-örnekleme yapılmıştır (Tablo 3.6). Tabloda görüldüğü gibi hastalık verisinin okuma uzunluğu 36 iken, karşılaştırma verilerinin okuma uzunluğu 100 ve 101'dir. Hizalama adımında, hastalık verisi için *BWA* aracının *samse* alt aracı kullanılırken, karşılaştırma verileri için *mem* alt aracı kullanılmıştır. NA12878 genomu için hastalık verisindeki ortalama okuma sayısı üzerinden, NA20355 genomu için ise hastalığa ait her iki verinin okuma sayıları üzerinden alt-örnekleme yapılmıştır. Ayrıca, hastalık verisinin tek uçlu, karşılaştırma verilerinin ise çift uçlu dizileme verisi olması bu süreçte göz önünde bulundurulmuştur.

Tablo 3. 6: Kullanılan veriler ve bu verilerin alt-örneklemesinden elde edilen çalışma verileri (Ok. Uzu: Okuma Uzunluğu, Alt-Örn: Alt-Örnekleme, TU: tek uçlu veri, ÇU: çift uçlu veri)

	Ok. Uzu.	Alt-Örn. Öncesi		Alt-Örn. Sonrası	
		Dosya Adı	Toplam Okuma Sayısı	Dosya Adı	Toplam Okuma Sayısı
Hastalık Verisi	36	S1_mdup.bam	31019035 (TU)	S1_mdup.bam	Alt-Örn. yapılmadı
	36	S2_mdup.bam	33895793 (TU)	S2_mdup.bam	Alt-Örn. yapılmadı
Karşılaştırma Verileri	NA12878	101	NA12878_5x_mdup. bam	92459454 x 2 (ÇU)	5XS1.bam 32420314 5XS2.bam 32424365 5XS3.bam 32422462
		100	NA20355_1_mdup. bam	48361599 x 2 (ÇU)	NA20355_down1. bam 30988327
			NA20355_2_mdup. bam	49047287 x 2 (ÇU)	NA20355_down2. bam 33910734
	NA20355	100	NA20355_1_mdup. bam	48361599 x 2 (ÇU)	NA20355_down1. bam 30988327
			NA20355_2_mdup. bam	49047287 x 2 (ÇU)	NA20355_down2. bam 33910734

NA12878 genomu için kullanılan veri SRR622461 erişim numaralı veridir. Hizalanan ve hizalama sonrası süreçlerin tamamlandığı veri dosyası üzerinde (NA12878_5x_mdup.bam) alt-örnekleme yapılmıştır. Hastalık verisinde her havuz için bir veri olmak üzere iki farklı veri dosyası mevcuttur. Bu nedenle, üç farklı alt-örnekleme (5XS1, 5XS2, 5XS3) yaratılarak, farklı alt-örneklemler ikili küme oluşturacak şekilde birlikte analiz edilerek, her kümeden elde edilen sonuçların ortalaması alınmaktadır. Buna göre, 5XS1 ve 5XS2 bir küme, 5XS1 ve 5XS3 diğer bir küme, 5XS2 ve 5XS3 de diğer bir küme olmak üzere, üç farklı veri kombinasyonu elde edilmiştir. *Picard* aracı, toplam okuma sayısı üzerinden alt-örnekleme gerçekleştirir. Buna göre, hastalık verisindeki iki verinin okuma sayılarının ortalaması

alınarak istenen okuma sayısı bulunmuştur (32457414). NA12878 genomuna ait verinin toplam okuma sayısı da (92459454×2) 'dir. İstenilen okuma sayısı, toplam okuma sayısına oranlandığında 0,175 değeri elde edilmektedir. Bu değer, “olasılık (P değeri)” olarak *picard* aracına gönderilmiştir. Öte yandan, üretilen her verinin birbirinden farklı olması için “random_seed (R değeri)” *null* olarak atanır. Buna göre, Tablo 3.6’da belirtilen okuma sayılarına sahip üç farklı alt-örneklem (5XS1, 5XS2, 5XS3) üretilmiştir.

NA20355 genomu içinse, çift uçlu olarak dizilenen iki veri seti (ERR251661 ve ERR251662) kullanılmaktadır. Yine, hizalanan ve hizalama sonrası işlemlerin gerçekleştirildiği veri üzerinden alt-örnekleme yapılmaktadır. Bu genom için iki veri olduğundan, hastalık verisindeki ortalama okuma sayısını almak yerine, birebir her hastalık verisi ile eş okuma sayıları elde edilecek şekilde alt-örnekleme yapılması hedeflenmiştir. Buna göre, birinci veri S1’in okuma sayısı olan 31019035’e eş, ikinci veri de S2’nin okuma sayısı olan 33895793’e eş olacak şekilde iki alt-örneklem (NA20355_down1, NA20355_down2) oluşturulmuştur. Bu yaklaşım ile elde edilen alt-örneklemelerin okuma sayıları 30988327 ve 33910734 olmuştur.

3.4.6 Varyant Çağırma

Literatürde farklı varyant çağırma araçları mevcut olup, bunlara dair liste Bölüm 1.1.4’de sunulmuştur. Öte yandan, geliştirilen ardışık düzen havuz dizileme ve düşük kapsamalı dizileme verilerine yönelik olduğundan, bu veriler özelinde geliştirilen araçlar olduğu ve bu araçlara dair literatür araştırması da Bölüm 1.1.5 ve Bölüm 1.1.6’da sunulmuştur. Tüm bu araştırmalar çerçevesinde, her iki stratejiyle üretilen veriler ile etkin çalışabilen varyant çağırma aracı arayışına girilmiştir. Bu bağlamda yapılan araştırmada, havuz dizileme araçlarının performansı ile ilgili yapılan çalışmada havuz dizileme verisinde kullanılabilen beş aracın, GATK, CRISP, LoFreq, VarScan ve SNVer, değerlendirilmesi yapılmıştır [51]. Buna göre, doğruluk değerleri açısından GATK, CRISP ve LoFreq’in diğerlerine göre daha iyi olduğu, CRISP ve LoFreq’in ise GATK’ya göre hafıza kullanımı, çalışma süreleri ve doğruluk değerleri açısından daha iyi olduğu ifade edilmiştir. Öte yandan, GATK’nin veri olarak küçük havuzlar kullanıldığı ve en iyi hassasiyetin istendiği durumlarda maliyetin göz ardı edilerek tercih edildiği belirtilmiştir. Düşük kapsamalı dizileme verisi üzerinde dört

farklı TNP çağırma aracının, SOAPsnp, Atlas-SNP2, SAMtools, GATK, karşılaştırıldığı çalışmada ise, GATK ve Atlas-SNP2'nin diğerlerine göre daha iyi performans gösterdiği ve GATK'nin, Atlas-SNP2'ye göre daha fazla TNV çağırabildiği ifade edilmiştir [72]. Ayrıca, yazarlar tek bir varyant çağırma aracının kullanılacağı durumlarda GATK aracının tercih edilmesini de önermişlerdir. Bu sonuçlar doğrultusunda, GATK'nin hem havuz dizileme hem de düşük kapsamlı dizileme verilerinde etkin çalıştığı anlaşılmıştır. Ayrıca, hastalık verisinde her bir havuzda dört örnek olmak üzere iki havuz olduğundan belirtilen küçük havuz kriteri [51] de sağlanmaktadır. Son olarak, GATK'nın popüler bir varyant çağırma aracı olması sebebi ile karşılaşılan problemlere daha kapsamlı ve daha ulaşılabilir çözümler sunabileceği de düşünüldüğünden, varyant çağırma aracı olarak GATK tercih edilmiştir.

3.4.6.1 The Genome Analysis Toolkit (GATK)

GATK, YND için verimli analiz araçlarının geliştirilmesini kolaylaştırmak amacıyla tasarlanan bir yapısal programlama çatısıdır [91]. GATK, Haritalama-indirgeme (MapReduce) tekniğine dayanan ve Java ile yazılan açık kaynak geliştirme ortamıdır. Haritalama-indirgeme yaklaşımı, verinin dağıtık olarak işlenmesi esasına dayanan bir programlama modelidir.

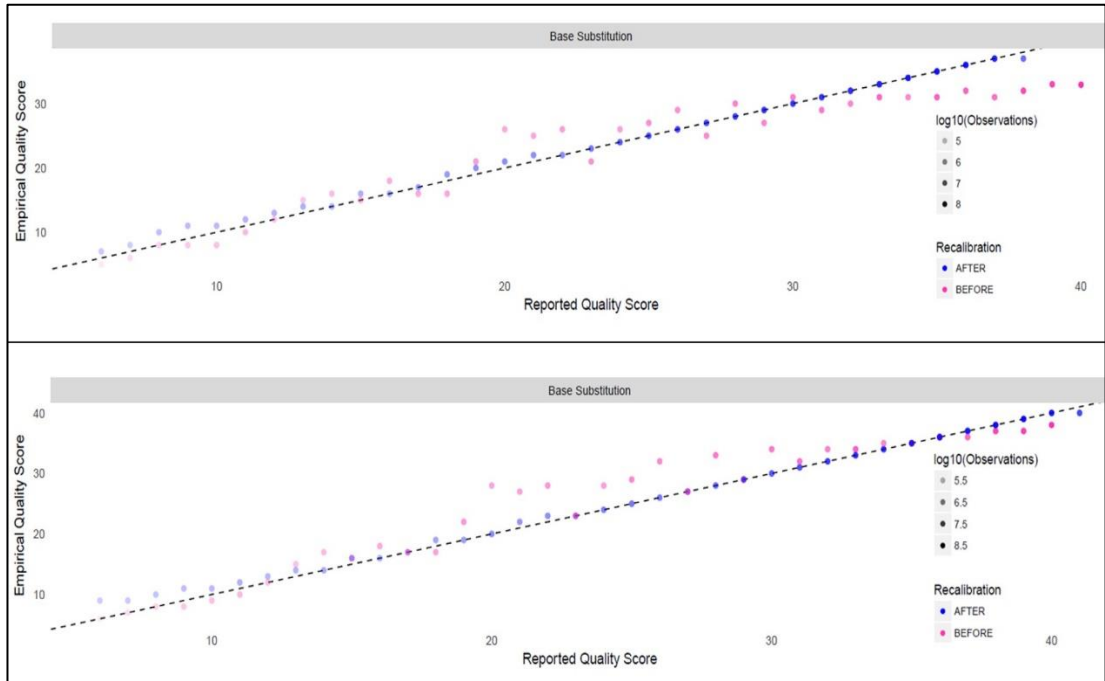
3.4.6.2 Baz Kalite Skoru Yeniden Kalibrasyonu (BQSR)

YND sürecinde, dizileme cihazları tarafından okunan her bir nükleotid için o nükleotidin ne doğrulukta okunduğunu gösteren bir kalite skoru atanır. Ancak, bazı sebeplerden dolayı bu skorlarda hatalar oluşabilir. GATK ekibi tarafından, oluşan bu hataları deneysel olarak modelleyip, bu modele dayanarak atanan skorları ayarlayabilen makine öğrenme tabanlı *BQSR* adında bir algoritma geliştirilmiştir [113]. *BQSR* iki adımda çalışır: İlk olarak, *BaseRecalibrator* ile çalışılan veri ve bilinen varyantlarla bir model oluşturulur ve kalibrasyon dosyası üretilir. Daha sonra, *ApplyBQSR* ile bu model kullanılarak nükleotidlerin kalite skorları yeniden ayarlanarak yeni bir veri dosyası (BAM) üretilir. GATK ekibi, algoritmanın etkisini tam olarak görebilmek adına, zorunlu olmamakla beraber, *BaseRecalibrator* adımının

ikinci kez çalıştırılarak ikinci bir modelle sürecin görsel olarak karşılaştırılmasını önermektedir.

Geliştirilen ardışık düzen kapsamında *BQSR* sürecinde kullanılan bilinen varyant veri tabanlarına (dbsnp_146.hg38.vcf.gz, 1000G_phase1.snps.high_confidence.hg38.vcf.gz, Mills_and_1000G_gold_standard.indels.hg38.vcf.gz) GATK Resource Bundle [114] sayfasındaki dosya aktarım protokolü (FTP) alanından ulaşılmıştır.

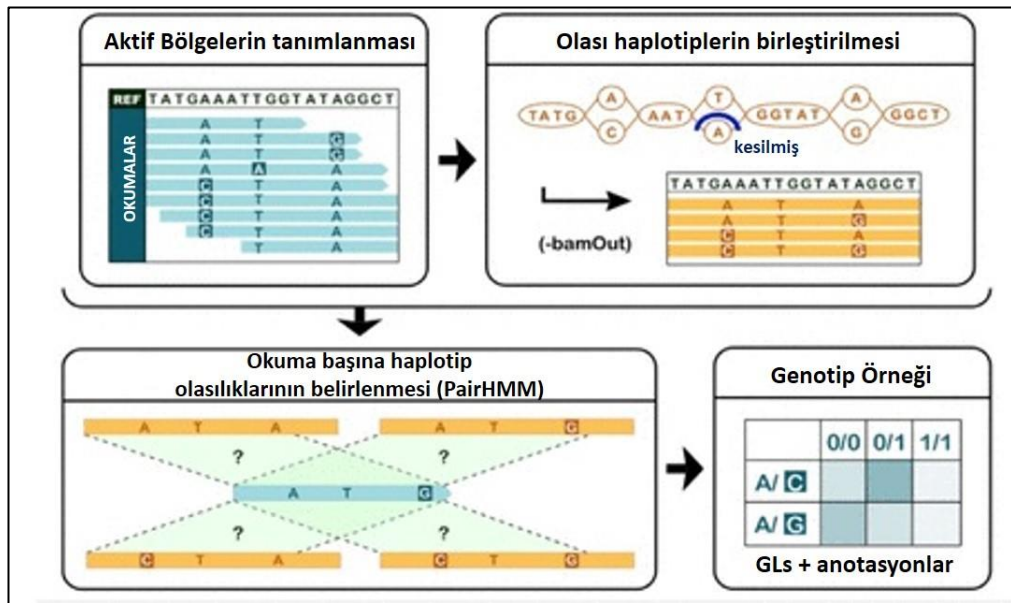
BQSR süreci sonunda, araç tarafından bazı verilerin ve grafiklerin sunulduğu bir rapor oluşturulur. Bunun için, önerildiği gibi *BaseRecalibrator* adımının iki kez çalıştırılması gerekir. Elde edilen grafiklerden Deneysel-Atanan Kalite Skoru grafiği (Şekil 3.7) incelendiğinde, yeniden kalibrasyon sonrasında önceye göre atanan kalite skorlarının deneysel skor değerlerine daha yakın olduğu görüldüğünden, kalite skorları üzerinde bir iyileşme gerçekleştiği anlaşılmaktadır.



Şekil 3. 7: RB-H1 (üst) ve RB-H2 (alt) hastalık verilerine dair edilen deneysel-atanan kalite skorları (empirical-reported quality score) grafiği. (Base Substitution: Baz İkamesi, log10 (Observations): log10(Gözlemler), Recalibration: Yeniden kalibrasyon) (Grafikler, GATK BQSR aracı tarafından üretilmiştir.)

3.4.6.3 GATK HaplotypeCaller

Geliştirilen ardışık düzen kapsamında varyant çağırma için GATK'nın *HaplotypeCaller* aracı [115, 116] kullanılmıştır. *HaplotypeCaller*, Şekil 3.8'de görüldüğü gibi dört temel adımda gerçekleşir [116, 117]: Öncelikle, program aktif bölgeleri tanımlar. Aktif bölge, genom üzerinde varyasyonların olabileceği bölgeler olup, program bu bölgeler üzerinde çalışır. Daha sonra, her bir aktif bölge için 'de Bruijn' benzeri bir çizge oluşturularak verideki olası haplotipler belirlenir. Sonra, belirlenen haplotipler çift Saklı Markov Model (pair-Hidden Markov Model, pair-HMM) ile genotip olasılıklarını hesaplamak için kullanılır. Son olarak, örnekler genotipler atanır. Her aktif bölge için, Bayes kuralı ile okumalardaki alellerin olasılığı kullanılarak her genotipin olasılığı hesaplanır. Daha sonra, en yüksek olasılığa sahip olan genotip o örneğin genotipi olarak atanır.



Şekil 3. 8: GATK HaplotypeCaller'ın temel çalışma adımları [116]. (Şekil, Kaynak [116]'den alıntılanarak Türkçeleştirilmiştir.)

Geliştirilen ardışık düzen kapsamında, *GATK HaplotypeCaller* havuz dizileme ve düşük kapsamlı dizileme için kullanılan parametreler dışında, varsayılan değerlerle ve veriler bir arada analiz edilecek şekilde (joint calling) çağrılmıştır. Havuz dizileme için *ploidy* parametresi kullanılmıştır. Diploit genom ile çalışıldığından ve her havuzda

dört örnek olduğundan *ploidy* parametresine 8 değeri atanmıştır. Düşük kapsamalı dizileme için ise, *min-pruning* ve *min-dangling-branch-length* parametrelerinin ikisine de 1 değeri atanmıştır. Bu adım sonunda VCF uzantılı bir dosya elde edilir.

3.4.7 Varyant Filtreleme

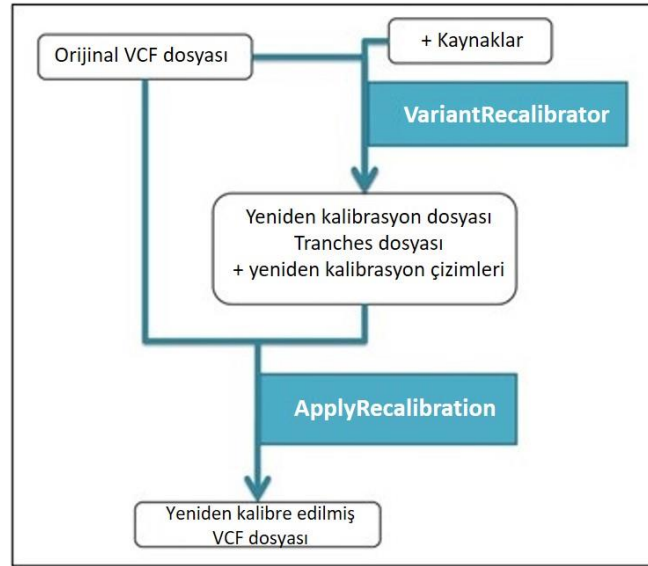
GATK aracı ile varyantlar filtrelenirken, çalışılan genom ve çalışılan verinin özelliklerinin değerlendirilmesi sonrasında, makine öğrenme tabanlı bir yaklaşım olan Varyant Kalite Skoru Yeniden Kalibrasyonu (VQSR) ya da katı filtreleme (hard-filtering) tercih edilir.

3.4.7.1 Varyant Kalite Skoru Yeniden Kalibrasyonu (VQSR)

GATK geliştiricileri, varyant filtreleme için öncelikle VQSR yaklaşımının kullanılmasını önermektedir [118]. Varyant çağırma sürecinde her ne kadar öncelikle varyantları doğru ve etkili bir şekilde çağırarak hedeflense de, gerçek varyantların (GP) yanı sıra gerçekte varyant olmadığı halde varyant olarak çağrılanlar (YP) da bulunabilir. VQSR, geliştirilen makine öğrenme tabanlı bir yaklaşımla varyantlar için tanımlanan yeni bir skor ile mümkün olduğunca GP varyantları bulup, YP olanları elemeyi hedefler [119]. VCF dosyasında, tanımlama (annotation) olarak ifade edilen, varyantlara dair bağlamsal istatistik (context statistics) metrikler tutulur. Örneğin; MQ okumaların hizalama kalitesinin karesel ortalamasının karekökü (root mean square of the mapping quality of reads), DP ise genel kapsama derinliği (overall depth of coverage) için kullanılan tanımlamalardır. VQSR varyant filtreleme için herhangi bir tanımlama değerinin belli bir eşik değeri ile karşılaştırılması yaklaşımı yerine, Gaussian karma dağılım (mixture model) ile gerçekleştirilen kümeleme tekniği ile varyantları filtreler. VQSR, tanımlama değerlerinin Gaussian dağılım göstermesi varsayımı (her ne kadar geçerli olmayan tanımlamalar olsa da, örneğin, MQ gibi.) üzerine kuruludur [119]. Literatürdeki yüksek doğruluğa sahip varyant veritabanları eğitim ve doğruluk seti için kullanılır. Böylece, bu varyantların çalışılan verideki tanımlama değerleri kullanılarak pozitif varyantları tanımlamak için bazı kurallar öğrenilir. Benzer bir durum negatif varyantlar için de geçerli olup, onları tanımlamak için de ayrıca bazı kurallar öğrenilir. Bu durum, set içindeki tüm kaynaklara uygulandıktan sonra, çalışılan veride çağrılan her bir varyant için olasılıksal bir değer

(VQSLOD) atanır. Bu değer, her bir varyantın pozitif olma olasılığının negatif olma olasılığına oranının logaritmik değeridir.

VQSR süreci Şekil 3.9’da görüldüğü gibi iki adımda gerçekleşir. (*GATK 4* ile *ApplyRecalibration* metodunun ismi *ApplyVQSR* olarak değişmiştir.)



Şekil 3. 9: VQSR sürecinin iş akışı [120]. (Şekil, Kaynak [120]’den alıntılanarak Türkçeleştirilmiştir.)

İlk adım, *VariantRecalibrator* ile Gaussian modellerinin oluşturularak, her bir varyant için VQSLOD skorunun hesaplanmasıdır. Bunun için, seçilen tanımlama değerleri ve yüksek doğruluğa sahip varyant veritabanlarından oluşan kaynaklar kullanılır. *VariantRecalibrator* adımı için tanımlanan kaynak setleri şunlardır [120]:

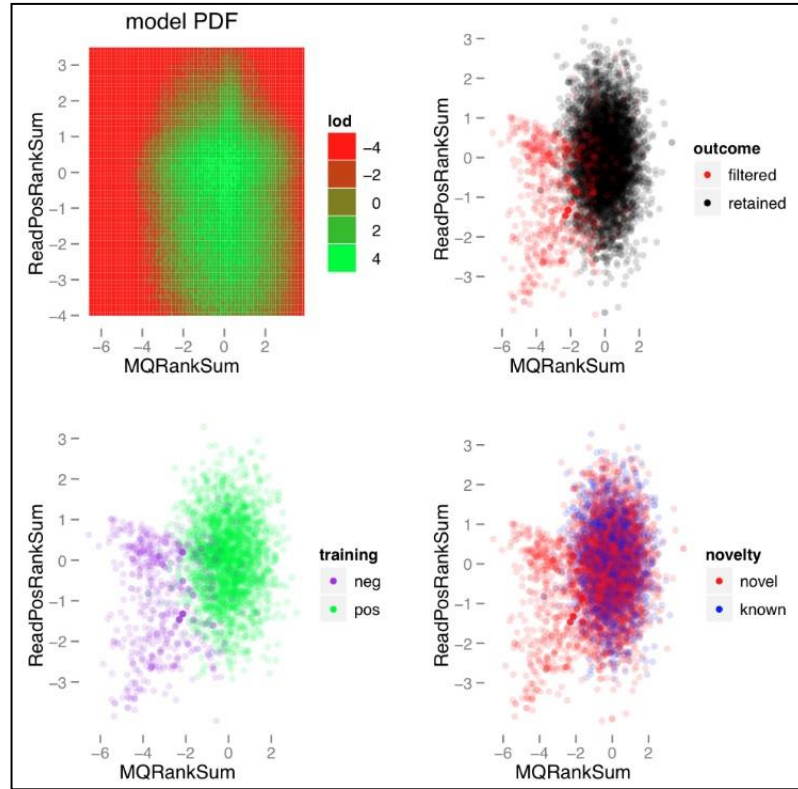
- Eğitim seti; Gaussian modellerinin oluşumu için kullanılır.
- Doğruluk seti; *VQSLOD* değerleri için uygun duyarlılık eşliğinin bulunması için kullanılır.
- Bilinen seti; Sadece raporlama için kullanılır.

Bu adımda kullanılan yüksek doğruluğa sahip varyant veritabanları GATK Resource Bundle sayfasından [114] indirilerek, Tablo 3.7’de belirtilen kaynak setlerine [115] dahil edilerek kullanılmıştır.

Tablo 3. 7: VQSR için kullanılan varyant veritabanları ve kaynak seti listesi

	Veritabanı	Eğitim	Doğruluk	Bilinen
TNP	Hapmap (hapmap_3.3.hg38.vcf.gz)	T	T	F
	Omni (1000G_omni2.5.hg38.vcf.gz)	T	T	F
	1000Genome (1000G_phase1.snps.high_confidence.hg38.vcf.gz)	T	F	F
	dbSNP (dbSNP_146.hg38.vcf.gz)	F	F	T
indel	Mills (Mills_and_1000G_gold_standard.indels.hg38.vcf.gz)	T	T	F
	dbSNP (dbSNP_146.hg38.vcf.gz)	F	F	T

VariantRecalibrator adımı sonunda tanımlamalara ait dağılım grafikleri oluşturulur. Gaussian karma dağılıma göre oluşturulan grafikler çok boyutlu olmasına karşın, anlaşılır olabilmesi açısından her tanımlama çifti için iki boyutlu grafikler raporlanır [119]. Şekil 3.10'da, *ReadPosRankSum* ve *MQRankSum* tanımlama çifti için raporlanan model grafiklerine ait bir örnek [120] görülmektedir. Buna göre, sol üstteki grafikte çalışılan verideki varyantların bu tanımlama değerlerindeki dağılım grafiği gösterilmektedir [119]. Burada, yeşil bölgedeki varyantlar yüksek kaliteli varyantları, kırmızı bölgedekiler ise düşük olasılığa sahip varyantları göstermektedir. Diğer üç grafikteki veri aynı olmakla beraber, her biri farklı özellikleri vurgulamaktadır. Örneğin; sağ üstteki grafik, varyantların filtrelenme durumlarını, sol alttaki grafik eğitim seti içindeki varyantların dağılımını, sağ alttaki grafik ise varyantların bilinen ve yeni olma durumlarını göstermektedir.



Şekil 3. 10: VariantRecalibrator adımı sonunda üretilen model grafikleri örneği [120]. (outcome: sonuç, filtered: filtrelenenler, retained: tutulanlar, training: eğitim, neg: negatif, pos: pozitif, novelty: yenilik, novel: yeni, known: bilinen)

VQSR'ın ikinci adımında ise, *ApplyVQSR* ile varyantlar filtrelenerek yeni bir VCF dosyası üretilir. Bu adımda, seçilen belli bir duyarlılık değerine göre varyantlar filtrelenir [119]. Seçilen duyarlılık değerinden daha yüksek değere sahip olan varyantlara 'PASS' değeri atanır. Bu değerden daha düşük olanlar ise filtrelenerek denk geldiği dilim belirtilir ya da elendiğini gösteren filtre tanımlamaları kullanılır.

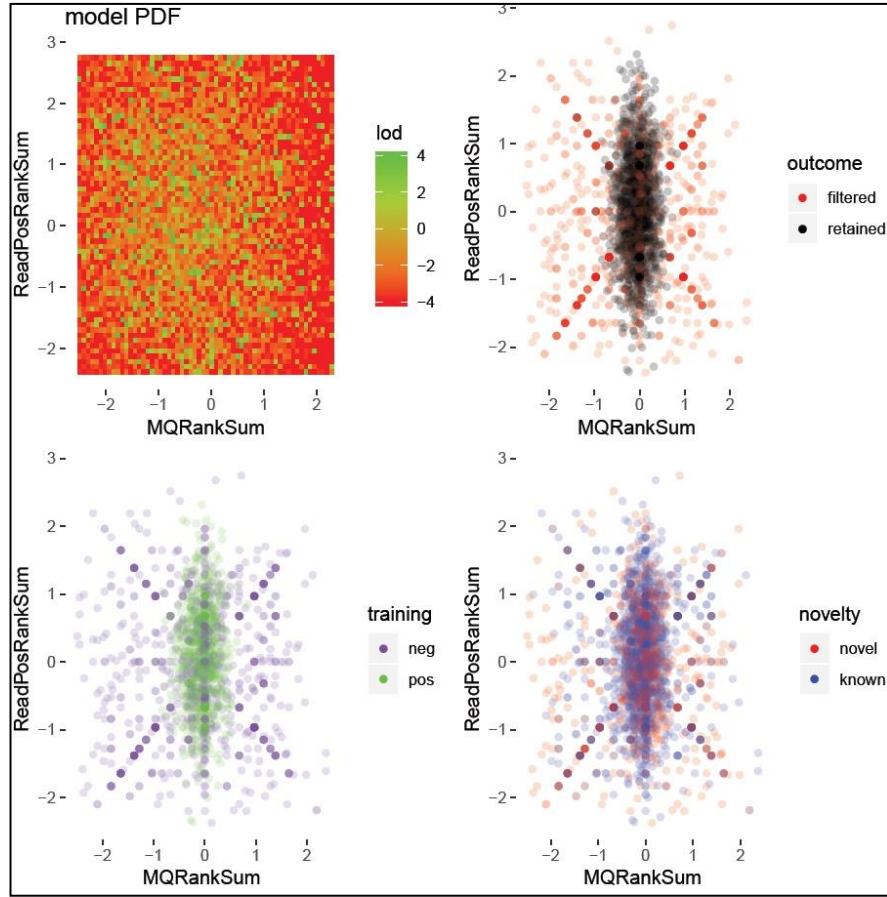
VQSR süreci için göz önünde bulundurulması gereken hususlar şunlardır [119]:

- VQSR süreci, TNP ve indeller için ayrı ayrı çalıştırılmalıdır.
- GATK ekibi, insan genomu ile yapılan bir çalışmada VQSR'ın iyi çalışabilmesi için, en az 1 tüm genom örneği ya da 30 tane ekzom örneğine ihtiyaç olduğunu tespit etmişlerdir. Eğer bu büyüklükte bir veri yoksa, ulaşılabilir veri kaynaklarından alınan örneklerle sayının çoğaltılması önerilmektedir. Bunun için de, çalışmadaki örneklerin ve alınan örneklerin

BAM dosyaları beraber kullanılarak varyant çağırma yapılması gerektiği belirtilmiştir.

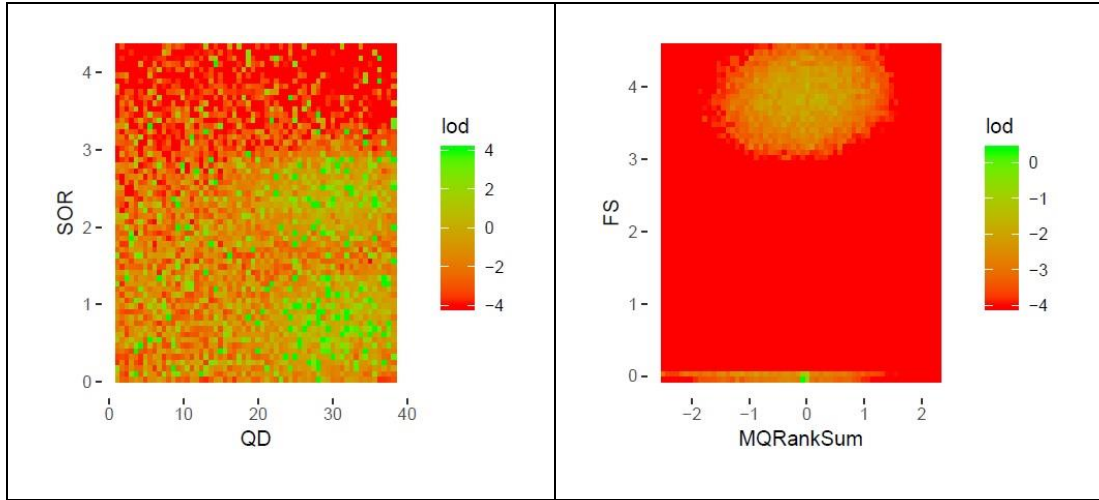
- Yukarıda bahsedilen şartları sağlayan, ama yine de küçük bir veri seti olduğundan uygun sonuçlar alınamadıysa, eğitim seti için oluşturulacak Gaussian model sayısının *maxGaussians* parametresi ile azaltılması da öneriler arasındadır.

Düşük kapsamlı dizileme verisinde de varyant filtreleme için VQSR kullanılması gerektiği belirtilmiştir [118]. Bu nedenle, geliştirilen ardışık düzen kapsamında VQSR kullanılmıştır. TNP ve indeller için ayrı ayrı çalışılmıştır. GATK geliştiricileri, VQSR’da kullanılan parametrelerin (seçilen tanımlamalar, duyarlılık değeri gibi) kullanıcıların kendi verileri ve çalışmalarının özelliklerine göre belirlenmesini önerirler. Bu kapsamda, geliştirilen ardışık düzende çalışılan veri için farklı tanımlama setleri ve Gaussian model sayısı (4, 6, 8) kullanılarak VQSR çalışılıp uygun set bulunmaya çalışılmıştır. Sonuçlara göre, *ReadPosRankSum – MQRankSum* tanımlama çifti için hastalık verisi üzerinde elde edilen model grafikleri Şekil 3.11’de sunulmaktadır.



Şekil 3. 11: Hastalık verisi ile VariantRecalibrator adımı sonunda üretilen model grafikleri. (GATK VQSR aracı tarafından üretilmiştir.) (outcome: sonuç, filtered: filtrelenenler, retained: tutulanlar, training: eğitim, neg: negatif, pos: pozitif, novelty: yenilik, novel: yeni, known: bilinen)

Seçilen veriyle çalışılan VQSR sürecinde, diğer tanımlama çiftleri için oluşan dağılım grafiklerinden iki örnek Şekil 3.12’de verilmiştir.



Şekil 3. 12: Seçilen veri ile VariantRecalibrator adımı sonunda üretilen tanımlama çifti dağılım grafiği örnekleri. (GATK VQSR aracı tarafından üretilmiştir.)

3.4.7.2 Katı Filtreleme

Varyant filtreleme için öncelikli olarak VQSR ile çalışılması önerilse de, VQSR ile çalışılmayacak bazı durumlarda tanımlamalar için belirlenen eşik değerlerine göre filtreleme yapılır. VQSR’ın kullanılamayacağı aşağıda listelenen durumlarda, Tablo 3.8’deki tanımlamalar ve eşik değerleri önerilmiştir [118].

- Küçük ölçekli deneyler (Hedef gen panelleri- targeted gene panels, 30’dan daha az ekzom verisinden oluşan ekzom dizileme çalışmaları)
- VQSR için kullanılacak uygun eğitim/doğruluk kaynakları bulunmayan genomlarla yapılan çalışmalar

Tablo 3. 8: Katı filtrelemede kullanılması önerilen tanımlamalar ve eşik değerleri

Tanımlama	TNP	indel
QualByDepth (QD)	2,0	2,0
FisherStrand (FS)	60,0	200,0
RMSMappingQuality (MQ)	40,0	-
MappingQualityRankSumTest (MQRankSum)	-12,5	-
ReadPosRankSumTest (ReadPosRankSum)	-8,0	-20,0
StrandOddsRatio (SOR)	3,0	10,0
InbreedingCoeff	-	-0,8

Katı filtreleme için göz önünde bulundurulması gereken noktalar ise şunlardır [118]:

- 10X'den küçük kapsama değerine sahip olan dizileme verilerinde (düşük kapsamalı dizileme verileri) katı filtrelemenin mümkün olamaması sebebi ile VQSR tercih edilmesi gerektiği,
- Tablo 3.8'de verilen *InbreedingCoeff* popülasyon bazlı bir metrik olduğundan çalışılan veride 10 ya da daha fazla örnek bulunması durumunda kullanılması,
- Tabloda önerilen değerlerin çalışılan veriye, organizmaya göre değişebileceği, her durum için en iyi sonuç veremeyebileceği, bu nedenle kullanılan veri ve genom için bu değerler üzerinde çalışılarak uygun değerlerin araştırılması gerekliliği.

3.4.8 Yüksek Güvenilirliğe Sahip Varyantların Bulunması

Geliştirilen ardışık düzenlerin performanslarının değerlendirilebilmesi için çağrılan varyantların doğrulanması gerekir. Bu amaçla yapılan bir çalışmada, doğrulama için The Genome in a Bottle (GIAB) Konsorsiyumu listesindeki varyantlar kullanılmıştır [121]. Bu çalışmada, ekzom veri analizi için farklı hizalama ve varyant çağırma araçlarıyla oluşturulan ardışık düzenlerin çağırdığı varyantlar GIAB listesindeki varyantlarla karşılaştırılarak performansları değerlendirilmiştir.

Bu tez çalışmasında da, yüksek güvenilirliğe sahip varyantların bulunması için GIAB kıyaslama listesi [122] kullanılmaktadır. 'Standart' ya da 'Geliştirilen' ardışık düzen tarafından çağrılan ve varyant filtreleme adımını başarılı bir şekilde geçen her varyant,

GIAB varyant listesiyle karşılaştırılarak Gerçek Pozitif (GP), Yanlış Pozitif (YP) ve Yanlış Negatif (YN) olarak değerlendirilmektedir. Buna göre;

- GP – Standart / Geliştirilen ardışık düzende çağrılmış olup GIAB listesinde bulunanlar,
- YP – Standart / Geliştirilen ardışık düzende çağrılmış olup, GIAB listesinde bulunmayanlar,
- YN – Standart / Geliştirilen ardışık düzende çağrılmayıp, GIAB listesinde bulunan varyantlar.

Bu değerlendirme, varyantların kromozom ve kromozom üzerindeki yerine göre yapılmaktadır. Bunun için, *hap.py* aracının *som* metodu kullanılmıştır. Bu değerlendirme sonrasında, her veri için her iki ardışık düzenin de kesinlik (precision), duyarlılık ve F-skorları hesaplanmaktadır (Denklem 3.3-3.5).

$$kesinlik = \frac{GP}{GP+YP} \quad (3.3)$$

$$duyarlılık = \frac{GP}{GP+YN} \quad (3.4)$$

$$F - skor = 2 \times \frac{kesinlik \times duyarlılık}{kesinlik+duyarlılık} \quad (3.5)$$

3.4.9 Varyant Anotasyonu

Geliştirilen ardışık düzenin bu aşamasında, belirlenen varyantlar *ANNOVAR* aracı ile anlamlandırılmıştır. Bunun için, yüksek güvenilirliğe sahip varyantların olduğu VCF dosyası *ANNOVAR* dosya formatına dönüştürülmüştür. *ANNOVAR*, genel olarak, gen-temelli (gene-based), bölge-temelli (region-based) ve filtre-temelli (filter-based) olmak üzere üç farklı şekilde çalışır [123]. Gen-temelli anotasyon, çağrılan varyantların protein kodlama değişikliklerine yol açıp açmadığını araştırır. Bunun için, NCBI Reference Sequence Database (RefSeq) anotasyon veritabanı (*refGene*) kullanılmıştır. RefSeq, canlılara ait genomik, transkript ve protein dizileri de dahil olmak üzere geniş kapsamlı bir dizileme veritabanıdır [124]. Bu kapsamda, tüm varyantların türlerine dair (örneğin; ekzonik, intronik gibi) bilginin tutulduğu

“_variant_function” ve ekzonik varyantların aminoasit deęişikliklerine dair bilginin tutulduęu “_exonic_variant_function” uzantılı iki dosya üretilir [123]. Bölge-temelli anotasyona göre, *cytoBand* anotasyon veritabanı (*hg38_cytoBand*) ile çalışılmıştır. Elde edilen sonuç dosyasında, her bir varyantın sitogenetik bant bilgisi (kromozom üzerindeki bölgeler) tutulmaktadır. Filtre-temelli anotasyon ile çağrılan varyantların çeşitli varyant veritabanlarında mevcut olup olmadığı araştırılırken [123], bu amaçla çalışılan her veritabanı için iki farklı dosya üretilir. Buna göre, çalışılan veritabanında olan varyantlar “_dropped” uzantılı dosyaya, çalışılan veritabanında bulunmayan varyantlar ise “_filtered” uzantılı dosyaya eklenir. Filtre-temelli yaklaşıma göre, dbSNP (*hg38_avsnp150*) ve 1000 Genomes (*hg38_ALL.sites.2015_08*) anotasyon veritabanları ile çalışılmıştır. dbSNP, küçük ölçekli insersiyon ve delesyonlar da dahil olmak üzere, insan genomundaki yaygın varyasyonlar ya da klinik çalışmalarda bulunan mutasyonların ve bunlara dair bilgilerin tutulduęu açık bir veritabanıdır [125]. 1000 Genom Projesi ise, çok sayıda insan genomunun dizilenmesine dayalı ilk proje olup, sonucunda insan genomuna ait varyasyonlar için geniş kapsamlı açık bir veritabanı oluşturulmuştur [101].

3.4.10 Anotasyon Çıktılarının İşlenmesi

Ardışık düzenin ikinci kısmı, yazılan bir kod ile “Anotasyon Çıktılarının İşlenmesi” sürecidir. Bunun için, biyoinformatik alanında sıklıkla kullanılan *R* programlama dili tercih edilmiştir. *R*, istatistiksel hesaplama ve grafikler için çözümler sunan bir dil ve yazılım ortamıdır [94]. *R* kodu, etkili bir görsel ortam sunan *RStudio* [95] ortamında yazılmıştır. Varyant anotasyonu adımıında, her anotasyon türü ve kullanılan veritabanına göre çeşitli metin dosyaları üretilir. Yazılan kod ile, üretilen dosyaların çeşitli istatistikler ve grafikler bazında deęerlendirilebilmesi ve anlamlandırılabilmesi sağlanmaktadır. Bu kısımda elde edilen sonuçlar Bölüm 4.2’de sunulacaktır.

BÖLÜM 4

BULGULAR

4.1 Karşılaştırmalı Sonuçlar

Geliştirilen ardışık düzenin etkinliğini değerlendirebilmek için, standart bir ardışık düzen tanımlanmıştır. Bu, varyant çağırma adımıyla kullanılan GATK aracının varsayılan parametreleri kullanması ile elde edilir ve sonuçlar içerisinde ‘Standart Ardışık Düzen’ ya da ‘Standart’ olarak ifade edilmektedir. Bunun yanı sıra, düşük kapsamlı dizileme stratejilerinde varyant çağırma oranını arttırdığı iddia edilen parametrelerin kullanıldığı ardışık düzen mevcuttur. Bu ise, ‘Geliştirilen Ardışık Düzen’ ya da ‘Geliştirilen’ olarak ifade edilmektedir. Hem hastalık verisi hem de karşılaştırma verileri (NA12878 ve NA20355 genomları) için Standart ve Geliştirilen ardışık düzen ayrı ayrı çalıştırılıp, sonuçları karşılaştırılarak performans değerlendirmeleri yapılmıştır (Tablo 4.1). Buna göre;

- Her üç veri için de geliştirilen ardışık düzen, standart bir ardışık düzenden ciddi oranda daha fazla TNV ve indel çağırabilmektedir. Örneğin, hastalık verisi üzerinde standart ardışık düzen 2804 indel ve 76456 TNV çağırırken, geliştirilen ardışık düzen 4652 indel ve 542035 TNV çağırmıştır.
- Öte yandan, geliştirilen ardışık düzende daha yüksek duyarlılık değerleri elde edilirken, daha düşük kesinlik değerleri elde edilmiştir.
- Kesinlik değerleri her veri için ayrı ayrı incelendiğinde, NA12878 ve NA20355 genomları için geliştirilen ardışık düzenin sonuçlarının standardın ürettiğine çok yakın olduğu gözlemlenirken, en ciddi fark hastalık verisindeki TNV sonuçlarında görülmüştür.
- Kesinlik ve duyarlılık değerlerini etkili bir şekilde değerlendirebilmek için F-skor değeri hesaplanmıştır. F-skor, kesinlik ve duyarlılık değerlerinin harmonik ortalamasıdır. Tablo 4.1’de görüldüğü gibi her üç veri için de geliştirilen ardışık düzende daha yüksek F-skor değerleri elde edilmiştir.

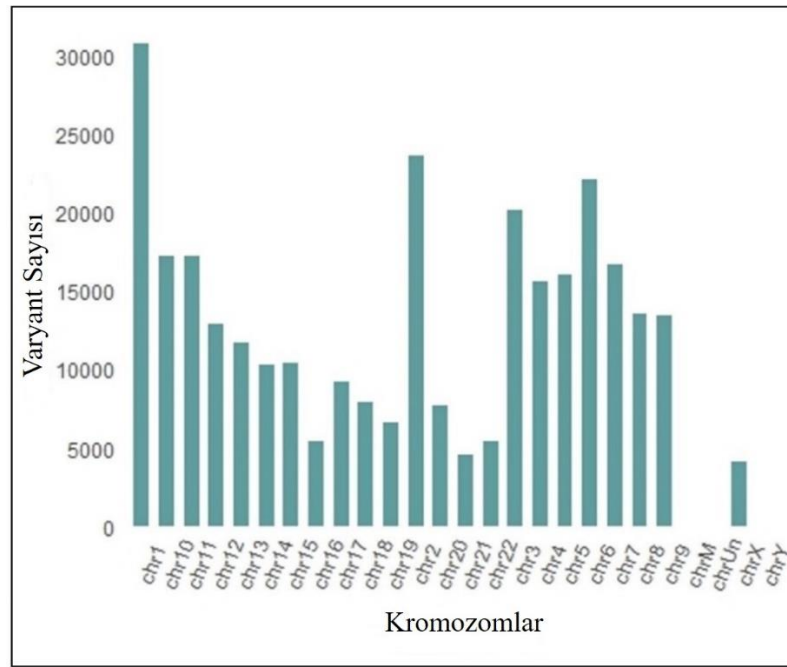
- Bu sonuçlar çerçevesinde, çağrılan indel ve TNV sayıları açısından, duyarlılık ve F-skor değerleri açısından geliştirilen ardışık düzenin standart ardışık düzenden daha iyi sonuç verdiği görülmektedir.
- Öte yandan, geliştirilen ardışık düzenin havuz dizileme ve düşük kapsama ile dizilen veri üzerinde tam anlamıyla etkili olduğunu gösterebilmek adına bu stratejilerle dizilen başka yeni nesil dizileme verilerine de ihtiyaç vardır. Bu veriler, mutasyon oranı yüksek olan, erişkin kanser verileri olabilir.

Tablo 4. 1: Standart ve Geliştirilen ardışık düzenin karşılaştırılması (STD.: STANDART, GEL.: GELİŞTİRİLEN, TNV: Tek Nükleotid Varyant, indel: insersiyon-delesyon, GP: Gerçek Pozitif, YP: Yanlış Pozitif, YN: Yanlış Negatif)

			Sayı	GP	YP	YN	DUYARLILIK (%)	KESİNLİK (%)	F-SKOR (%)
HASTALIK VERİSİ	STD.	indel	2804	1822	982	531365	0,3	65	0,7
		TNV	76456	51049	25407	3036374	1,7	66,8	3,2
	GEL.	indel	4652	2957	1695	530230	0,6	63,6	1,1
		TNV	542035	298817	243218	2788606	9,7	55,1	16,5
NA12878	STD.	indel	141095	96190	44904	436997	18	68,2	28,5
		TNV	1120217	889888	230329	2197535	28,8	79,4	42,3
	GEL.	indel	151758	101272	50486	431915	19	66,7	29,6
		TNV	1696079	1317629	378450	1769794	42,7	77,7	55,1
NA20355	STD.	indel	135781	58874	76907	474313	11	43,4	17,6
		TNV	1295076	643604	651472	2443819	20,8	49,7	29,4
	GEL.	indel	153432	64913	88519	468274	12,2	42,3	18,9
		TNV	1816074	857767	958307	2229656	27,8	47,2	35

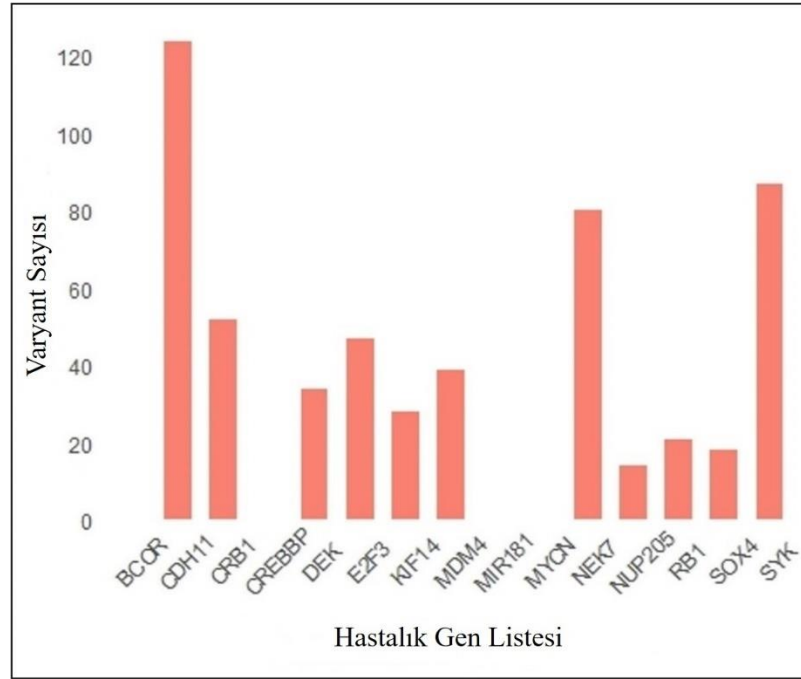
4.2 Hastalık Verisi Sonuçları

Varyant çağırma adımı bulanan varyantlardan yüksek güvenilirliğe sahip olanların belirlenmesinden sonra, sadece bu varyantlar üzerinde anotasyon adımı gerçekleştirilerek, sonuçlar elde edilmiştir. Buna göre, ANNOVAR VCF dosyasında 301775 tane TNP ve indel bulunmaktadır. Bunların kromozomlara göre dağılımına baktığımızda (Şekil 4.1), hastalık ile ilişkili RB1 geninin bulunduğu 13. kromozom üzerinde 11674 tane varyant çağırılmıştır.



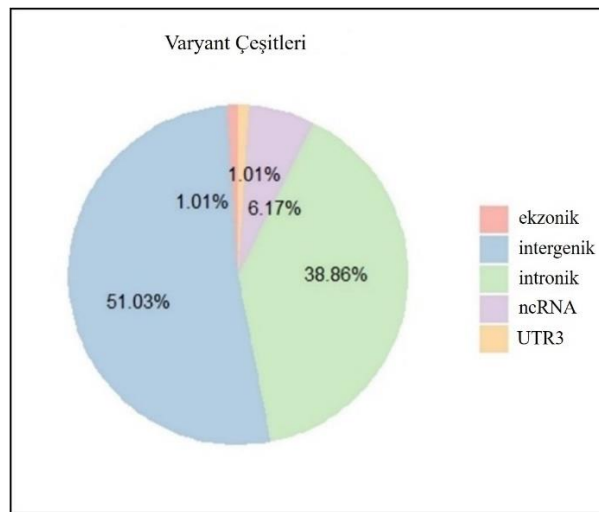
Şekil 4. 1: Kromozomlardaki varyant dağılımı

YND ile retinoblastom hastalığı kapsamında yapılan literatür araştırmasına (Bölüm 3.1) göre, hastalıkta etkin olan genlerde bulunan varyant sayılarına baktığımızda (Şekil 4.2) BCOR, CREBBP, MIR181, MYCN, dışındaki genlerde varyant çağırıldığı, ancak bunların hiçbirinin patojenik olmadığı görülmüştür.



Şekil 4. 2: Retinoblastom gen listesindeki varyant dağılımı

Çağrılan varyantların 301673 tanesi dbSNP (ANNOVAR; hg38_avsnp150) veritabanında, 298008 tanesi ise 1000 Genome veritabanında (ANNOVAR; hg38_ALL.sites.2015_08) bulunmaktadır. Öte yandan, RefSeq veritabanına göre 3057 tane ekzonik varyanttan 1342 tanesi sinonim olmayan (nonsynonymous) TNV'dir (Şekil 4.3). (Sadece % 1'in üstünde olan varyant çeşitleri gösterilmiştir.)



Şekil 4. 3: Varyant çeşitleri (RefSeq veritabanı)

RB1 geni üzerinde ekzonik varyant bulunmazken, 21 tane bilinen intronik varyant çağrılmıştır (Tablo 4.2).

Tablo 4. 2: RB1 geninde bulunan varyantlar

Gen	cytoBand	dbSNP No
LPAR6;RB1	13q14.2	rs12430215
LPAR6;RB1	13q14.2	rs1951775
LPAR6;RB1	13q14.2	rs198564
LPAR6;RB1	13q14.2	rs198567
RB1	13q14.2	rs11839271
RB1	13q14.2	rs174857
RB1	13q14.2	rs1981434
RB1	13q14.2	rs198576
RB1	13q14.2	rs198582
RB1	13q14.2	rs198583
RB1	13q14.2	rs2245534
RB1	13q14.2	rs2515668
RB1	13q14.2	rs2804090
RB1	13q14.2	rs2804094
RB1	13q14.2	rs2854342
RB1	13q14.2	rs2854348
RB1	13q14.2	rs2854352
RB1	13q14.2	rs2854353
RB1	13q14.2	rs2854363
RB1	13q14.2	rs443211
RB1	13q14.2	rs9568042

YND ile retinoblastom hastalığında varyantların belirlenmesi süreci iki ayrı çalışmada iki ayrı hedefli dizileme verisi üzerinde çalışılmıştır [78, 79]. Devarajan ve arkadaşları tarafından yapılan çalışmada FastQC ile kalite kontrolü yapılan, BWA ile hizalanan, GATK ile TNP ve indelleri, ExomeCNV and Cn. MOPS araçları ile de kopya sayısı varyasyonlarını çağırarak, bu varyantların ANNOVAR ile anlamlandırıldığı benzer bir ardışık düzen çalışılmıştır [78]. Ancak, burada çalışılan veri hem kandan hem tümörden alınmış olup, sadece RB1 gen bölgesine odaklı hedefli yeniden dizileme verisidir. Tez kapsamında çalışılan veri ise, sadece tümörden alınmış olup tüm genom dizileme verisidir ve havuz dizileme ve düşük kapsamalı dizileme stratejileriyle dizilendiğinden bu veriye uygun bir ardışık düzen çalışılmıştır. Grotta ve arkadaşları tarafından yapılan diğer çalışmada ise YND ve RB1 custom array-Comparative Genomic Hybridization (aCGH) teknikleri bir arada çalışılarak sonuçlar sunulmuştur

[79]. Bu çalışmada da, hedefli yeniden dizileme verisi kullanılmış olup, veri analizi süreci için TruSeq Amplicon iş akışı takip edilmiş ve varyant çağırma için GATK kullanılmıştır. Bu çalışmada kullanılan veri yayımlandığından, geliştirilen ardışık düzenin temel adımlarını test edebilmek adına bahsi geçen hedefli yeniden dizileme verisi (hastalık test verisi) E-MTAB-3515 erişim numarası [126] ile indirilerek ardışık düzenin bir kısmı bu veri üzerinde çalıştırılmıştır. Çalışma içerisinde, YND ya da Sanger dizileme ile 24 tane farklı varyant çağrıldığı belirtilmiştir. Çalışmada 53 hastaya ait YND verisi kullanıldığı söylenmiş olup, yayımlanan halinde sadece 23 hastaya ait veri bulunmaktadır. Dolayısıyla, test sürecinde 23 hastaya ait veri çalışılabilmiştir. Kullanılan verinin hedefli yeniden dizileme verisi ve dolayısıyla küçük bir veri olması sebebi ile BQSR kullanmadan ve filtreleme stratejisi olarak katı filtreleme kullanılarak ardışık düzen uygulanmıştır ve rapor edilen 24 varyanttan 10'u çağırılabilmiştir. Çalışılan veri büyüklüğü, makaledeki veri ile birebir uyumlu olmadığından sadece varyant bazlı karşılaştırma yapılabilmiş olup, diğer değerlerle karşılaştırma yapılamamıştır. Daha sonra, veri büyüklüğü azaltılarak yeniden ardışık düzen çalıştırılmış ve sonuçlar incelenmiştir. Hastalık verisinde 8 örnek olması sebebi ile 23 hastanın ilk 8'i kullanılarak yeni bir veri seti oluşturulduğunda, test edilen ardışık düzen rapor edilen varyantlardan sadece 4'ünü çağırabilmiştir. Sonuç olarak, geliştirilen ardışık düzen kısmi olarak bu veri üzerinde test edildiğinde benzer sonuçlar alınabildiği ve veri büyüklüğü azaltıldığında daha az varyant çağırabildiği görülmüştür.

Bu çalışma ile, havuz dizileme ve düşük kapsama stratejileriyle dizilenmiş gerçek bir hastalık verisi üzerinde YND veri analizi çalışılıp sonuçların sunulması hedeflenmişti. Burada hedef, yeni varyantların bulunması değildi. Nitekim, düşük kapsamlı dizileme ile ilgili yapılan bir çalışmada, bu stratejilerle dizilenen YND verisinde belli bir hastalıkla ilgili varyantlar bulmak için mümkün olduğu kadar çok sayıda genomun kullanılması gerektiği belirtilmiştir [65]. Üzerinde çalışılan hastalık ile ilgili dikkat çeken bir nokta, retinoblastom genomunun çok düşük mutasyon oranına sahip olmasıdır [81, 84]. Hastalık ile ilgili dikkat çeken başka bir nokta ise, retinoblastomun ilgili olduğu RB1 geni üzerinde geleneksel tekniklerle çok sayıda varyant bulunabildiği ve retinoblastomdaki varyantların karakteristiği sebebi ile bu yaklaşımların hâlâ etkin olduğudur [127]. Havuz dizileme ve düşük kapsamlı

dizileme ile dizilenmiş gerçek veriler üzerinde hastalık ile ilişkili sonuçlar alabilmek adına, mutasyon oranı daha yüksek kanserler, özellikle erişkin kanserleri ve daha geniş havuz verilerinde çalışılması gerektiği düşünülmektedir.

4.3 CUDA Çalışmaları

Grafik İşlem Birimi (GPU), aynı anda binlerce iş parçacığını çalıştırabilme yeteneğine sahip işlemci birimleridir. GPU ile merkezi işlem birimi (Central Processing Unit, CPU) arasında bazı temel farklılıklar vardır [128]: CPU, birkaç çekirdeğe sahip, seri işlemler yapmaya uygun, aynı anda az sayıda işi bir arada yürütebilen işlemci birimidir. GPU ise, çok sayıda çekirdeğe sahip, paralel işlemler için uygun olan, aynı anda binlerce işi bir anda yürütebilen işlemci birimidir. GPU, öncelikle grafikler ve oyunlar için geliştirilse de, sonradan Genel Amaçlı GPU programlama (General Purpose Computing on GPU, GPGPU) anlayışıyla yoğun hesaplama gerektiren problemlerde de kullanılabilir hale gelmiştir. GPU hesaplamada, CPU ve GPU'nun birlikte kullanımına dayanan hibrid bir model kullanılır. Programın seri kısmı CPU üzerinde çalıştırılırken, yüksek hesaplama gerektiren kısmı GPU üzerinde çalıştırılır. Compute Unified Device Architecture (CUDA), NVIDIA tarafından geliştirilen ve GPU programlamada kullanılan bir paralel hesaplama platformudur.

Yeni nesil dizileme veri analizinde hizalama adımı zamana ve kaynağa ihtiyaç duyan bir adımdır. Bu nedenle, hizalama algoritmalarının bazıları CUDA platformunda çalışılarak yoğun hesaplama için GPU'nun paralel işlem özelliğinden faydalanılmıştır. BarraCUDA [129] BWA'nın, nvBowtie [130] Bowtie'in, Cushaw [131] BWT'nin ve Cushaw2-GPU [132] ise kendi geliştirdikleri Cushaw2 algoritmasının CUDA platformunda GPU için geliştirilmiş versiyonlarıdır.

Geliştirilen ardışık düzende BWA kullanılması sebebi ile, bu bölümde de BarraCUDA ile çalışılmıştır. BarraCUDA, BWA hizalama algoritmasının GPGPU ile geliştirilmiş versiyonudur [129]. CPU'ya göre, daha yüksek bir performansla çalıştığı ve BWA ile benzer hizalama oranına sahip olduğu belirtilmiştir. BWA-backtrack algoritmasının hizalama sürecini *aln* ve *samse/sampe* adımlarıyla gerçekleştirdiğinden Bölüm 3.4.3'de bahsedilmişti. BarraCUDA aracında, en yoğun hesaplama sürecinin olduğu *aln* adımı GPU üzerinde çalışılmıştır. Ayrıca, çift uçlu (ÇU) veriler için de *sampe*

adımında iş parçacığı kullanımı getirilmiştir. BarraCUDA, NVIDIA tarafından sunulan paralel hesaplama platformu olan CUDA’da geliştirilmiştir ve şu şekilde çalışır [129]: İlk olarak, BWT algoritmasıyla indekslenen referans genom ve dizi okumaları diskten GPU’nun hafızasına yüklenir. Daha sonra, her bir okumanın hizalanması GPU’daki yüzlerce işlemciye dağıtılarak, hesaplamanın paralel olarak yapıldığı bir GPU çekirdeği (kernel) oluşturulur. Çekirdek işini tamamladığında, hizalama sonuçları GPU’dan diske geri aktarılır.

CUDA platformunda geliştirilen GPU tabanlı hizalama araçlarının performansına yönelik yapılan bir çalışmada [133], BarraCUDA ve nvBowtie araçlarının CPU ve GPU üzerindeki performansları incelenmiştir. GPU’nun çok sayıda okumadan oluşan ve okuma uzunluğu da uzun olan veri setlerinde CPU’ya oranla daha iyi performans gösterdiği belirtilmiştir. Ayrıca, çalışmada en yeni bir GPU’nun sisteme eklenmesinin fiyat/performans açısından orantısız bir artış sağlamadığı, bu nedenle de daha eski bir GPU ile çalışmanın maliyet açısından daha etkili olabileceği görülmüştür.

CUDA ile yapılan analizler için tez kapsamında kullanılan veriler dışında bahsedilen araçların test edildiği bazı yeni veriler de kullanılmıştır. Buna göre, analizler sırasında kullanılan tüm veriler ve çeşitli özellikleri Tablo 4.3’de sunulmuştur. Retinoblastom hastalık verisi (ERR550406 ve ERR550407) tek uçlu (TU), diğer tüm veriler çift uçlu(ÇU)’dur. Veri seçiminde, farklı okuma uzunluklarına ve farklı okuma sayılarına sahip veriler kullanımına özen gösterilmiştir. Veri seti içerisinde, bir tane de diğerlerine göre çok daha büyük olan bir veri (SRR622457) seçilerek GPU’nun bu veriye katkısı gözlemlenmek istenmiştir. ERR003014, SRR032215, ERR161544, SRR211279 no’lu veriler, BarraCUDA’nın ve CUDA platformunda geliştirilen diğer hizalama araçlarının test edildiği verilerden bazılarıdır.

Tablo 4. 3: CUDA ile yapılan analizler için kullanılan tüm veriler ve çeşitli özellikleri

Veri	TU/ÇU	Okuma Uzunluğu	Toplam Okuma Sayısı	İndirme Yeri
ERR550406	TU	36	31019035	[97]
ERR550407	TU	36	33895793	[97]
ERR003014	ÇU	37	22673582	[134]
ERR000589	ÇU	51	24279572	[99]
SRR032215	ÇU	76	28291390	[134]
SRR211279	ÇU	100	50937050	[134]
ERR251661	ÇU	100	96723198	[103]
ERR161544	ÇU	100	148223280	[134]
SRR622461	ÇU	101	184918908	[99]
SRR622457	ÇU	101	2873647546	[134]

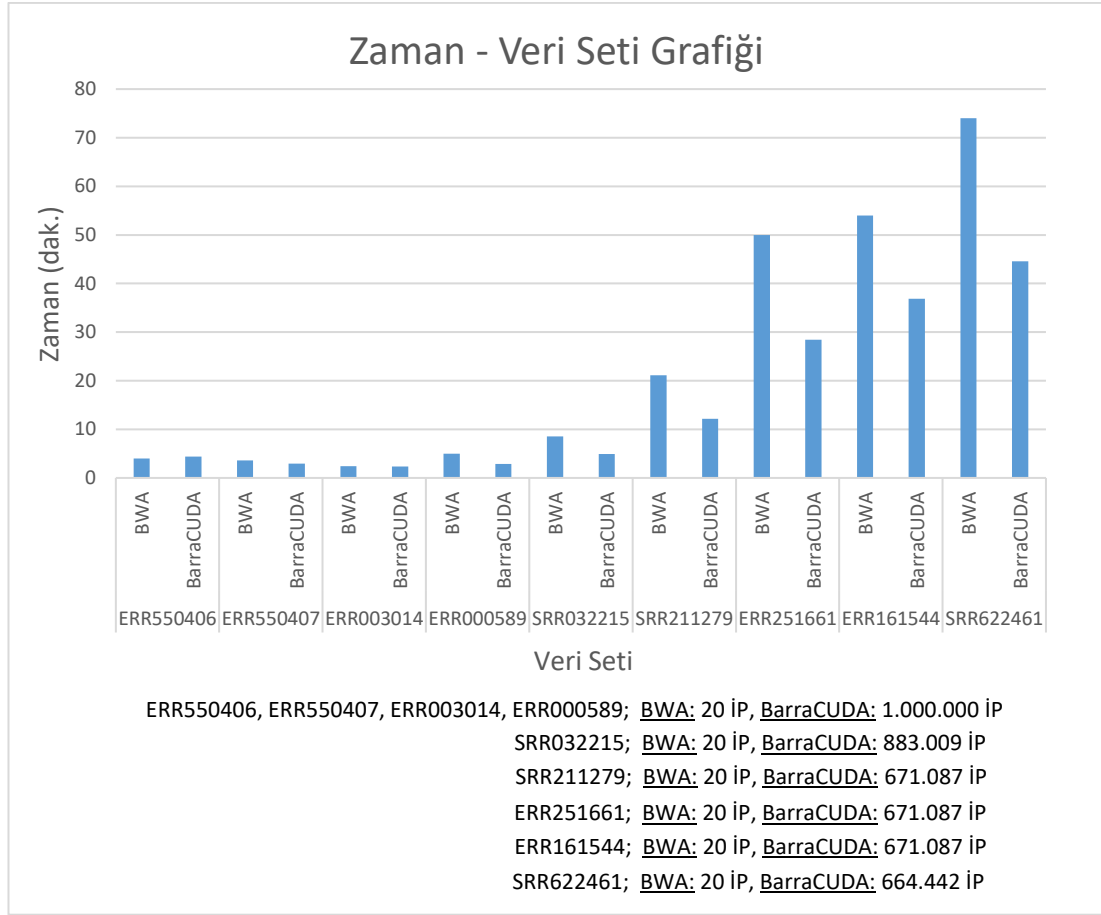
Her veri seti BWA ile CPU’da, BarraCUDA ile de GPU’da çalıştırılarak, zaman ve hizalama oranları karşılaştırılmıştır. NVIDIA P100 GPU kullanılmıştır. BWA, *aln* adımıyla 1, 8 ve 20 iş parçacığı (İP), *samse/sampe* adımıyla tek İP ile; BarraCUDA ise, *aln* adımıyla 1 GPU, *sampe* adımıyla ise 20 İP kullanılarak çalıştırılmıştır. Her yürütme üç kez tekrar edilip, süre ve hizalama oranlarının ortalamaları alınmıştır. Kullanılan verinin tek uçlu ya da çift uçlu olmasına bağlı olarak BWA ve BarraCUDA’nın *samse* ya da *sampe* alt araçları kullanılmıştır. BarraCUDA’da, BWA’dan farklı olarak, *sampe* alt aracı iş parçacıkları ile çalışılabilir hale getirilmiştir. Kullanılan iş parçacıkları sayısı dışında, araçlar tanımlı varsayılan değerlerle çalışılmıştır. Hizalama adımı referans genomun indekslenmesi ile başlar. Referans genom için oluşturulan indeks BWA ve BarraCUDA ile bir kez çalıştırılıp, tüm yürütmelerde kullanılmıştır. Buna göre, BWA ve BarraCUDA ile yapılan yürütmelerde elde edilen sonuçlar Tablo 4.4’de sunulmaktadır. Verilen çalışma zamanları *aln* ve *samse/sampe* adımlarının toplam değerleridir. BarraCUDA’nın çalışmasında [129] belirtildiği gibi tüm veri setlerinde benzer hizalama oranları elde edilmiştir. Öte yandan, BarraCUDA’nın BWA’ya göre oldukça hızlı çalıştığı görülmektedir. BarraCUDA ile, özellikle okuma uzunluğu uzun ve okuma sayısı fazla

olan verilerde, BWA kullanıldığındaki (20 İP) toplam çalışma zamanına göre neredeyse %50-60 oranında bir düşüş olduğu görülmektedir.

Tablo 4. 4: BWA ve BarraCUDA'nın sonuçları (Hiz. Oranı: Hizalama Oranı)

Veri Seti	BWA (1 İP)		BWA (8 İP)		BWA (20 İP)		BarraCUDA (1 GPU)	
	Hiz. Oranı (%)	Zaman (dak.)	Hiz. Oranı (%)	Zaman (dak.)	Hiz. Oranı (%)	Zaman (dak.)	Hiz. Oranı (%)	Zaman (dak.)
ERR550406	97,64	60,26	97,64	14,48	97,64	11,89	97,38	11,37
ERR550407	98,99	80,42	98,99	18,01	98,99	12,41	98,78	10,7
ERR003014	94,93	79,63	94,93	50,61	94,93	45,11	96,17	6,22
ERR000589	97,83	105,97	97,83	48,23	97,83	42,11	97,94	6,64
SRR032215	88,94	203,55	88,94	45,96	88,94	27,58	89,05	8,96
SRR211279	97,67	333,37	97,67	86,91	97,67	49,43	97,61	20,15
ERR251661	96,01	824,56	96,01	168,21	96,01	103,3	96,04	45,6
ERR161544	98	830,81	98	216,43	98	133,48	97,89	61,98
SRR622461	89,66	1102,82	89,66	240,31	89,66	151,49	89,7	75,66
SRR622457	-	-	-	-	89,75	2205,97	89,81	1250,53

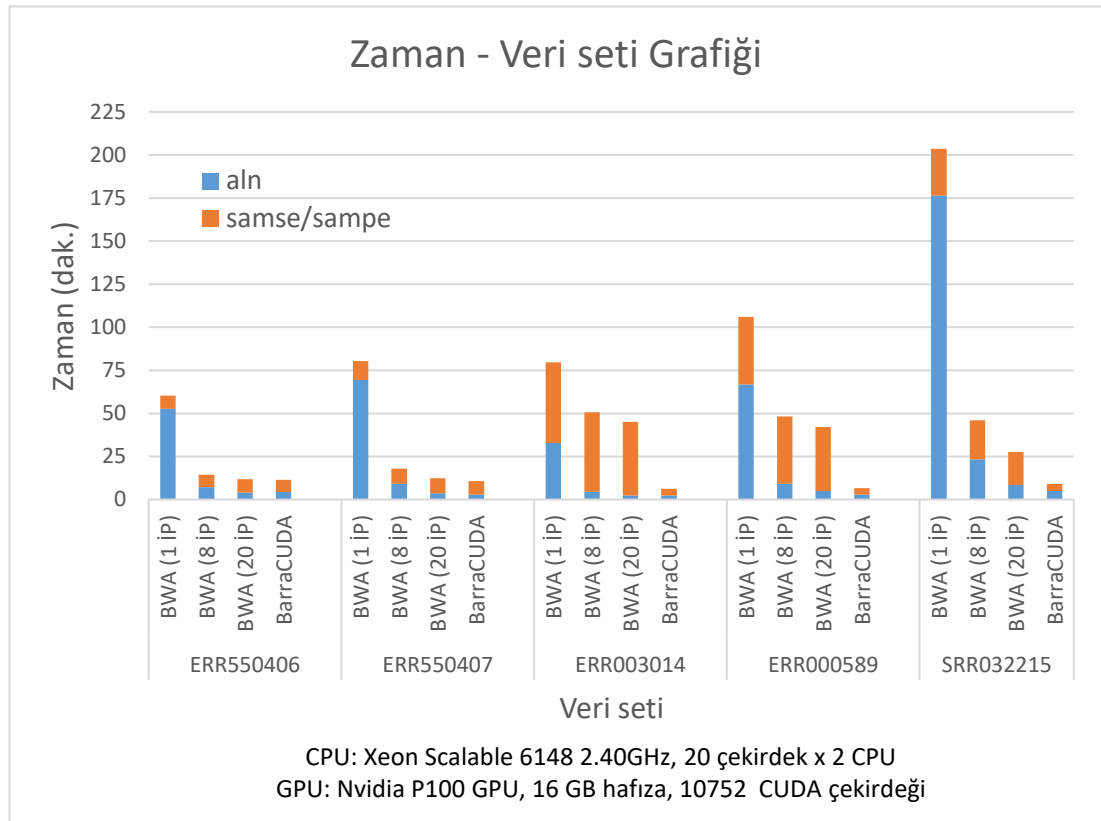
BarraCUDA, *aln* adımı GPU, *samse/sampe* adımı ise CPU kullanmaktadır. Dolayısıyla, GPU'nun etkisinin tam olarak görülebilmesi için hem BWA hem de BarraCUDA'nın *aln* adımının farklı veriler üzerindeki çalışma zamanı grafiği (Şekil 4.4) verilmiştir. Her bir veri seti için GPU'da kullanılan İP sayısı şeklin altında belirtildiği gibidir. GPU'da, çalışılan ÇU verilerinde okuma sayıları arttıkça zamanda belirgin bir azalış olduğu görülmektedir. TU verileri (ERR550406 ve ERR550407), tez kapsamında hastalık verisi olarak kullanılan düşük kapsamlı dizileme verileridir. TU verilerinde GPU'nun etkisinin değerlendirilebilmesi için yüksek kapsama ile dizilenen YND verileri kullanılabilir.



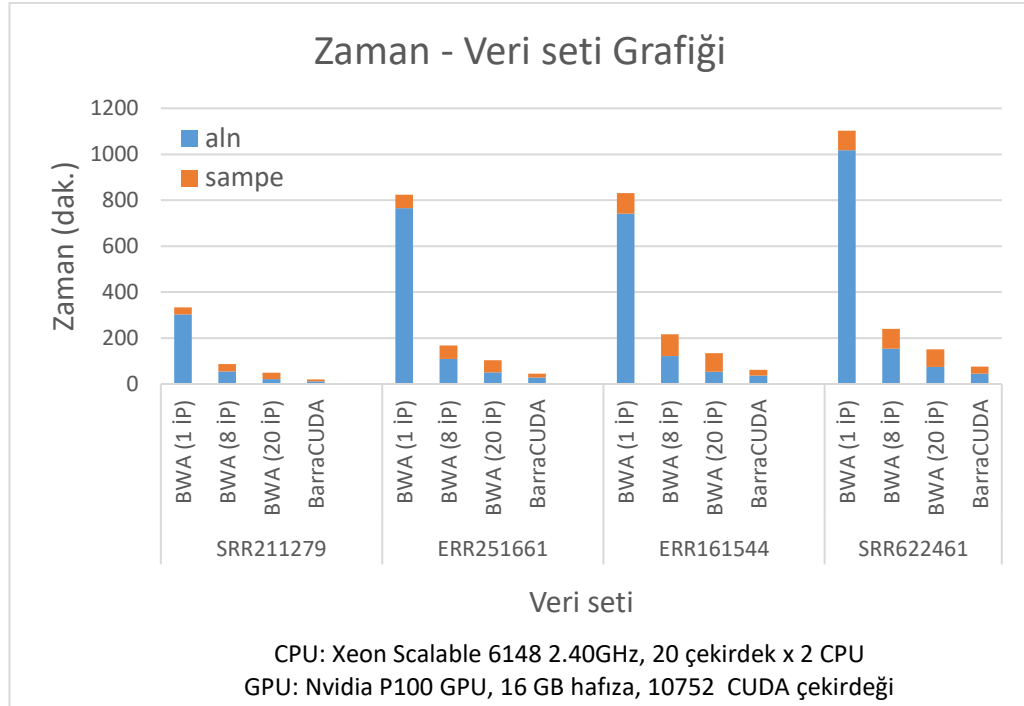
Şekil 4. 4: Zaman – Veri seti grafiği. BWA (CPU: 20 İP) ve BarraCUDA(GPU). Verilen zaman değerleri yalnızca *aln* adımı içindir.

Hizalama sürecinin, *aln* ve *samse/sampe* adımlarından oluştuğundan daha önce bahsedilmişti. Bu adımların, farklı sayıda İP ve GPU kullanımında toplam çalışma zamanına katkısı verilerin büyüklüğüne göre bölünerek iki farklı grafikte incelenmiştir (Şekil 4.5 ve Şekil 4.6). Şekillerde 1 İP, 8 İP ve 20 İP BWA ile çalışıldığını, BarraCUDA ise 1 GPU ile çalışıldığını göstermektedir. BarraCUDA’da, TU verileri için *aln* adımı GPU üzerinde, *samse* adımı ise İP kullanılmadan çalıştırılır. Buna göre, farklı sayılarda İP kullanımında *samse* adımının süresi aynı kalmakla beraber, *aln* için tek bir İP kullanıldığında alınan uzun çalışma zamanı farklı sayılarda İP ve GPU kullanımında azalmıştır (Şekil 4.5). BarraCUDA’da, ÇU verileri için *aln* adımı GPU üzerinde, *sampe* adımı ise 20 İP kullanılarak çalıştırılmıştır. Benzer şekilde, *aln* adımı için tek bir İP üzerinde alınan uzun çalışma zamanı farklı sayılarda İP ve GPU kullanımında ciddi oranda bir azalış göstermiştir (Şekil 4.5 ve Şekil 4.6). *sampe* adımı

için, BWA’da İP kullanımı desteklenmediğinden çalışma zamanları farklı yürütmelerde aynı kalırken, BarraCUDA’da İP kullanılabildiğinden çalışma zamanının azaldığı görülebilmektedir. Ayrıca, *aln* adımı açısından değerlendirildiğinde küçük verilerde GPU’nun performansı BWA ile 20 CPU İP ile sağlanabilirken (Şekil 4.5), veriler büyüdükçe GPU’nun performansı da artmaktadır (Şekil 4.6).

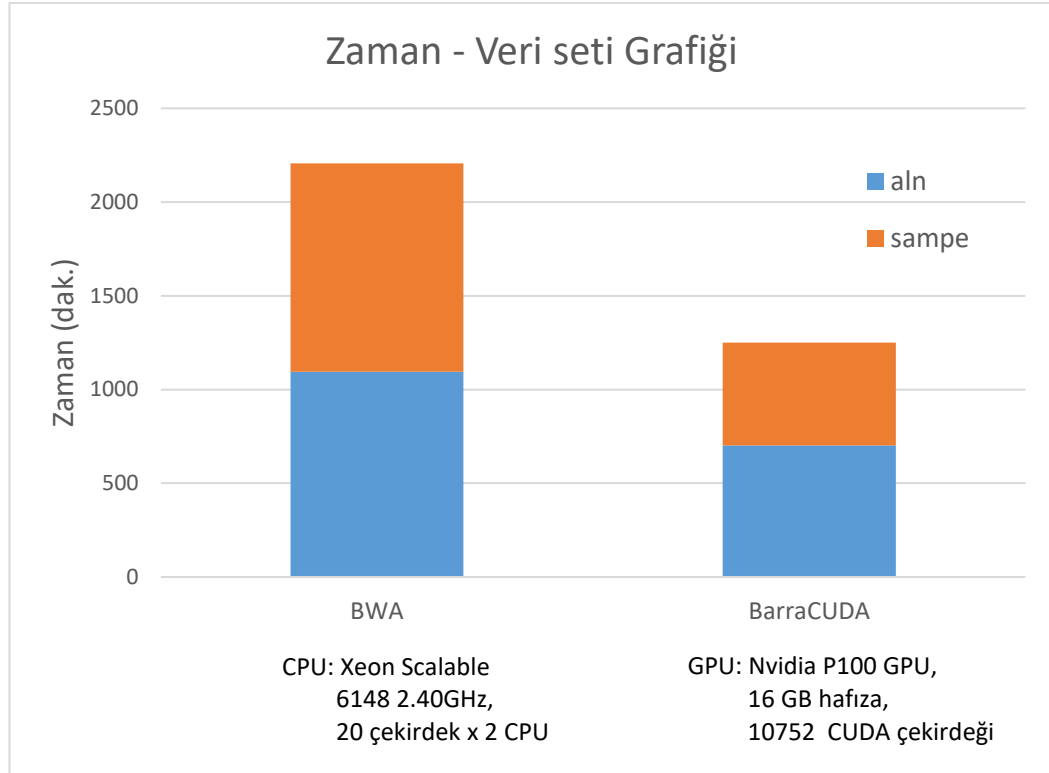


Şekil 4. 5: Küçük verilerin çalışma zamanı grafiği. BWA’de, *aln* adımı belirtilen sayıda İP, *samse/sampe* adımı tek İP ile; BarraCUDA’da, *aln* adımı 1 GPU, *sampe* adımı ise 20 CPU İP ile çalıştırılmıştır.



Şekil 4. 6: Daha büyük verilerin çalışma zamanı grafiği. BWA’de, *aln* adımı belirtilen sayıda İP, *sampe* adımı tek İP ile; BarraCUDA’da, *aln* adımı 1 GPU, *sampe* adımı ise 20 CPU İP ile çalıştırılmıştır.

Veriler içindeki toplam okuma sayısı en fazla olan verinin (SRR622457) çalışma zamanı kendi içinde değerlendirilmiştir. Bu veride 1, 4 ve 8 İP kullanılarak BWA sonuçları alınmaya çalışılmıştır. Ancak, çalışılan sistemin zaman kısıtı sebebiyle süre yeterli olamamıştır. Bu nedenle, bahsedilen veriyle, 20 İP kullanımı ile BWA ve 1 GPU kullanımı ile BarraCUDA olmak üzere, sadece iki yürütme yapılabilmektedir (Şekil 4.7). BarraCUDA, beklenildiği gibi her iki adımda da (*aln* ve *sampe*) çalışma zamanında ciddi bir düşüş sağlamıştır.



Şekil 4. 7: Veri seti içindeki en büyük verinin (SRR622457) çalışma zamanı grafiği. BWA’de, *aln* adımı 20 İP, *sampe* adımı tek İP ile; BarraCUDA’da, *aln* adımı 1 GPU, *sampe* adımı ise 20 CPU İP ile çalıştırılmıştır.

BWA algoritmasının ve GPGPU ile geliştirilen versiyonu olan BarraCUDA’nın farklı verilerle çalıştırılması sonucunda, tek bir GPU’nun çalışma zamanına ciddi anlamda olumlu katkı yaptığı görülmektedir. Ayrıca, BarraCUDA’nın belirtildiği gibi BWA ile benzer bir doğruluk oranına sahip olduğu da gözlemlenmiştir. Öte yandan, GPU’nun katkısının tam olarak anlaşılabilmesi adına okuma sayısı fazla olan (okuma uzunluğu kısa olan) başka veriler üzerinde de çalışılması gerektiği düşünülmektedir. Bu kapsamdaki verilerin gereksinim duyabileceği hafızanın hesaplamalarda bir sorun teşkil edip etmeyeceğine ise farklı veri setleri üzerinde yapılan analizler sonrasında karar verilebilir. Ayrıca, BarraCUDA’nın birden fazla GPU’yu desteklediği de belirtilmiştir. Özellikle okuma sayısı fazla olan verilerde birden çok GPU kullanımının nasıl bir etki yaratabileceği bu yönde yapılacak yeni çalışma içerisinde araştırılacaktır.

BÖLÜM 5

SONUÇ VE ÖNERİLER

Her bir DNA parçasının paralel dizilenmesine dayalı YND her ne kadar gelişimindeki ilerlemeler sayesinde maliyeti azalsa da, çalışılan veri setinin büyüklüğüne bağlı bir maliyet doğurabilir. Bunun için geliştirilen stratejilerden iki tanesi, havuz dizileme ve düşük kapsamlı dizilemedir. Bu iki strateji ile üretilen YND verisi ayrı ayrı çalışılsa da, bu iki verinin bir arada gerçek bir hastalık üzerinde çalışılmadığı görülmüştür. Çocukluk çağı kanseri olan retinoblastom hastalığında havuz dizileme ve düşük kapsamlı dizileme verisi olarak üretilen veri ile bu özellikteki veriler için bir ardışık düzen geliştirilmeye çalışılmıştır. Geliştirilen ardışık düzen, mümkün olabilen benzer verilerle standart bir ardışık düzenle karşılaştırılarak performansı incelenmiştir. Buna göre, geliştirilen ardışık düzenin standarta göre ciddi anlamda daha fazla varyant çağırabildiği görülmüştür. Geliştirilen ardışık düzende daha yüksek duyarlılık değerleri elde edilirken, daha düşük kesinlik değerleri elde edilmiş olup, en düşük kesinlik değeri hastalık verisindeki TNV çağırması sırasında görülmüştür. Her iki düzenin F-skor değerleri karşılaştırıldığında da, geliştirilen ardışık düzenin daha yüksek F-skor değerlerine sahip olduğu görülmüştür. Buna göre, geliştirilen ardışık düzenin standart olana göre daha iyi sonuçlar verdiği gözlemlenmiştir. Ayrıca, retinoblastom verisi ile anotasyon adımı da çalışılarak çağrılan varyantların anlamlandırılması sağlanmıştır. Buna göre, hastalık verisi üzerinde kromozomlara göre çağrılan varyant sayısı, retinoblastomla ilişkili genler üzerindeki varyant sayısı ve varyant çeşitleri raporlanmıştır. Öte yandan, retinoblastomla ilişkili RB1 geni üzerinde çağrılan varyantlar incelendiğinde intronik varyantların çağrıldığı görülmüştür ve patojenik varyanta rastlanmamıştır. Literatürün belirttiği gibi, retinoblastom genomunun düşük mutasyon oranına sahip olduğu desteklenmektedir. Geliştirilen ardışık düzende, hem havuz dizilemede hem de düşük kapsamlı dizileme verilerinde etkin çalıştığı belirtilen varyant çağırma aracı GATK kullanılmıştır. Ancak, bu stratejiler ile çalışabilen farklı araçların da mevcut olduğu görülmüştür ve bu araçların bu kapsamdaki veri üzerinde çalışıp çalışmayacağı ve etkin olup olmayacağı incelenebilir.

Veri analizi içinde göreceli olarak en çok zamana gereksinim duyan hizalama adımının GPU tabanlı CUDA platformunda geliştirilen bir araçla çalışılması sonucunda, tek bir GPU üzerinde dahi ciddi oranda bir zaman kazancı sağlanırken, CPU üzerinde çalışan versiyonu ile benzer bir hizalama doğruluğu elde edilmiştir. Ayrıca, hizalama sürecini oluşturan iki adımın farklı veri setleri üzerinde çalışma zamanı incelenerek bu aracın katkısı araştırılmıştır. Buna göre, hem GPU üzerinde çalışan adım hem de CPU üzerinde çalışan adımın çift uçlu veriler için iş parçacıklarıyla yürütülebilmesi sayesinde GPU versiyonu olarak geliştirilen araç ile ciddi bir zaman kazancı sağlanmıştır. Öte yandan, okuma sayısı büyük olan verilerde birden fazla GPU kullanımının etkisinin de incelenmesi gerektiği belirtilmiş olup, bu yönde tarafımızdan yapılan çalışmalar devam etmektedir. Ayrıca, GPU kullanımının performansa etkisi değerlendirilirken, her ne kadar hizalama oranları da incelenmiş ve benzer olduğu görülse de, hizalama adımı sonrası varyant çağırma araçları ile tüm bir veri analizi süreci ile varyant çağırma performansı da değerlendirilebilir. Son olarak, veri analizi süreci içinde GPU'nun çalışma zamanının azaltılmasında katkısının olabileceği başka adımlar (örneğin, varyant çağırma) araştırılabilir.

Geliştirilen ardışık düzen, büyük havuz verilerinde ve daha fazla sayıda havuz verisi içeren düşük kapsamlı farklı kanser dizileme verilerinde çalıştırılarak ardışık düzenin etkinliği daha net değerlendirilebilir. Ayrıca, mutasyon oranı daha yüksek kanserler ve erişkin verileri üzerinde geliştirilen ardışık düzeni çalışmanın daha etkili olabileceği düşünülmektedir. Bu kapsamda ulaşılabilir veriler üzerinde test edildiğinde de etkin çalışmaya devam ederse, daha az maliyetli olan bu stratejilerle veri kısıtı olan kanser türlerinde veri üretiminin ve paylaşımının desteklenebileceği düşünülmektedir.

KAYNAKLAR

- [1] What is Cancer? - National Cancer Institute. 10 Ekim 2020, <https://www.cancer.gov/about-cancer/understanding/what-is-cancer>.
- [2] Kamalakaran, S., Varadan, V., Janevski, A., Banerjee, N., Tuck D., McCombie, W. R., ve diğeri, "Translating next generation sequencing to practice: opportunities and necessary steps," *Mol Oncol*, 7(4), 743-55, 2013.
- [3] What is Human Genome Project? 15 Ekim 2020, <https://www.genome.gov/human-genome-project/What>.
- [4] Robinson, T.R. *Genetics For Dummies®*, Indianapolis, Indiana: Wiley Publishing, Inc. 2005.
- [5] Hogeweg, P., "The Roots of Bioinformatics in Theoretical Biology", *PLoS Comput Biol.*, 7(3), e1002021, 2011.
- [6] Luscombe, N. M., Greenbaum D. ve Gerstein, M., "What is bioinformatics? An introduction and overview", *Yearb Med Inform*, 1, 83-99, 2001.
- [7] Cohen, J., "Bioinformatics—an introduction for computer scientists", *ACM Computing Surveys*, 36(2), 122-158, 2004.
- [8] Aerts, I., Lumbroso-Le Rouic, L., Gauthier-Villars, M., Brisse, H., Doz F. ve Desjardins, L., "Retinoblastoma", *Orphanet J Rare Dis*, 1(31), 2006.
- [9] Tuncer, S., "Retinoblastom", *Klinik Gelişim*, 25, 56-65, 2012.
- [10] What is Retinoblastoma? - Retinoblastoma Information. 5 Ekim 2020, <https://www.cancer.org/cancer/retinoblastoma/about/what-is-retinoblastoma.html>.
- [11] Retinoblastoma Treatment (PDQ) - Patient Version - National Cancer Institute. 10 Ekim 2020, <https://www.cancer.gov/types/retinoblastoma/patient/retinoblastoma-treatment-pdq>.
- [12] Retinoblastoma: MedlinePlus Genetics. 15 Ekim 2020, <https://medlineplus.gov/genetics/condition/retinoblastoma/>.
- [13] Sanger, F., Nicklen S. ve Coulson, A. R., "DNA sequencing with chain-terminating inhibitors", *Proc Natl Acad Sci U S A*, 74(12), 5463-7, 1977.

- [14] Maxam A. M. ve Gilbert, W., "A new method for sequencing DNA", *Proc Natl Acad Sci U S A*, 74(2), 560-4, 1977.
- [15] Frederick Sanger - Facts - NobelPrice.org. 20 Eylül 2020, <https://www.nobelprize.org/prizes/chemistry/1980/sanger/facts/>.
- [16] Frederick Sanger - Facts - NobelPrice.org. 20 Eylül 2020, <https://www.nobelprize.org/prizes/chemistry/1958/sanger/facts/>.
- [17] Rizzo J. M. ve Buck, M. J., "Key principles and clinical applications of "next-generation" DNA sequencing", *Cancer Prev Res (Phila)*, 5(7), 887-900, 2012.
- [18] Voelkerding, K. V., Dames, S. A. ve Durtschi, J. D., "Next-generation sequencing: from basic research to diagnostics", *Clin Chem*, 55(4), 641-58, 2009.
- [19] Doğan, M., Eröz, R., Yüce, H. ve Özmerdivenli, R., "Yeni Nesil Dizileme (YND) Hakkında Bilinenler (Literatür Taraması)", *Düzce Tıp Fakültesi Dergisi / Duzce Medical Journal*, 19(1), 27-30, 2017.
- [20] DNA Sequencing Fact Sheet. 10 Ekim 2020, <https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Fact-Sheet>.
- [21] Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., ve diğer, "The complete genome of an individual by massively parallel DNA sequencing", *Nature*, 452, 872-6, 2008.
- [22] Ley, T. J., Mardis, E. R., Ding, L., Fulton, B., McLellan, M. D., Chen, K., ve diğer, "DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome", *Nature*, 456, 66–72, 2008.
- [23] Mwenifumbo, J. C. ve Marra, M. A., "Cancer genome-sequencing study design", *Nature Reviews Genetics*, 14, 321–332, 2013.
- [24] Meyerson, M., Gabriel, S. ve Getz, G., "Advances in understanding cancer genomes through second-generation sequencing", *Nat Rev Genet*, 11(10), 685-96, 2010.
- [25] Tran, B., Dancey, J. E., Kamel-Reid, S., McPherson, J. D., Bedard, P. L., Brown, A. M. K., ve diğer, "Cancer genomics: technology, discovery, and translation", *J Clin Oncol*, 30(6), 647-60, 2012.
- [26] Shyr, D. ve Liu, Q., "Next generation sequencing in cancer research and clinical application", *Biol Proced Online*, 15(4), 2013.

- [27] Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., ve diğeri, "A survey of tools for variant analysis of next-generation genome sequencing data", *Brief Bioinform*, 15(2), 256-78, 2014.
- [28] Cunha, M. L. R., Meijers, J. C. M. ve Middeldorp, S., "Introduction to the analysis of next generation sequencing data and its application to venous thromboembolism", *Thromb Haemost*, 114(5), 920-32, 2015.
- [29] Bao, R., Huang, L., Andrade, J., Tan, W., Kibbe, W. A., Jiang, H. ve Feng, G., "Review of Current Methods, Applications, and Data Management for the Bioinformatics Analysis of Whole Exome Sequencing", *Cancer Inform*, 13(Suppl 2), 67–82, 2014.
- [30] Ulahannan, D., Kovac, M. B., Mulholland, P. J., Cazier, J.-B. ve Tomlinson, I., "Technical and implementation issues in using next-generation sequencing of cancers in clinical practice", *Br J Cancer*, 109(4), 827–835, 2013.
- [31] Serrati, S., Summa, S. D., Pilato, B., Petriella, D., Lacalamita, R., Tommasi, S. ve Pinto, R., "Next-generation sequencing: advances and applications in cancer diagnosis", *Onco Targets Ther*, 9, 7355–7365, 2016.
- [32] Erman, B., "Ağır Kombine İmmün Yetmezlikli Hastalarda, Hastalığa Neden Olan Genetik Defektlerin Yeni Nesil Dizileme Yöntemiyle Araştırılması", Doktora Tezi, Hacettepe Üniversitesi, Sağlık Bilimleri Enstitüsü, Ankara-Türkiye, 2015.
- [33] Kartal, E., "Identification of Pathogenic Mutations in Neurodegenerative Disorders: Bioinformatic Analysis of Next Generation Sequencing Data", Master of Science, Boğaziçi University, Institute for Graduate Studies in Science and Engineering, İstanbul-Turkey, 2015.
- [34] Corraliza Marquez, A. M., "Copy number variations of colorectal cancer by whole exome sequencing data", Master Thesis, University of Vic-Central University of Catalonia, Department of Systems Biology, Barcelona-Spain, 2014.
- [35] Jazayeri, O., "Unravelling the genetic basis of hereditary disorders by high-throughput exome sequencing strategies", [Groningen]: University of Groningen, 2016.
- [36] Carrot-Zhang, J., "Unraveling the genetics of cancer using whole-exome sequencing", Doctor of Philosophy, McGill University, Department of Human Genetics, Montreal, Quebec, Canada, 2016.
- [37] Sheerin, U.-M., "The Use of Next Generation Sequencing Technologies to Dissect the Aetiologies of Parkinson's disease and Dystonia", Doctor of

Philosophy, University College of London, Institute of Neurology, London-England, 2014.

- [38] Martinez, C. B., "Whole exome sequencing of Urothelial Bladder Cancer: identification of ARID1A and STAG2 as new, important, players", PhD Thesis, Universidad Autonoma De Madrid, Department of Molecular Biology, Madrid-Spain, 2014.
- [39] Rantaperi, T., "Bioinformatic analysis of next-generation sequencing data", Master's Thesis, University of Tampere, Institute of Biomedical Technology, Finland, 2012.
- [40] Sigurgeirsson, B., "Analysis of RNA and DNA sequencing data: Improved bioinformatics applications", Doctoral Thesis, Royal Institute of Technology, School of Biotechnology, Stockholm-Sweden, 2016.
- [41] Alsaadi, A., "Identification and Validation of Mutated Signalling Pathways in Cancer", Doctor of Philosophy, University of Edinburgh, Edinburgh-England, 2016.
- [42] Dander, A., "Integration of Next-Generation Sequencing Data and Whole-Slide Bioimages for Personalized Oncology", Doctor of Philosophy, Innsbruck Medical University, Division of Bioinformatics Biocenter, Innsbruck-Austria, 2014.
- [43] Kawalia, A., "Addressing NGS Data Challenges: Efficient High Throughput Processing and Sequencing Error Detection", PhD Thesis, University of Cologne, Faculty of Mathematics and Natural Sciences, 2016.
- [44] Fischer, M., "Exome Analysis Using Next-Generation Sequencing Data", Doctoral Thesis, Graz University of Technology, Institute for Genomics and Bioinformatics, Graz-Austria, 2010.
- [45] Wieland, T., "Next-Generation Sequencing Data Analysis", Doctor of Science, Technical University of Munich, Faculty for Information Technology, Munich-Germany, 2015.
- [46] Suo, C., "Statistical Methods for the Detection, Analyses, and Integration of Biomarkers in the Human Genome and Transcriptome", PhD Thesis, Karolinska Institutet, Department of Medical Epidemiology and Biostatistics, Stockholm-Sweden, 2014.
- [47] Bao, S., Jiang, R., Kwan, W., Wang, B., Ma, X. ve Song, Y.-Q., "Evaluation of next-generation sequencing software in mapping and assembly", J Hum Genet, 56(6), 406-14, 2011.

- [48] Nielsen, R., Paul, J. S., Albrechtsen, A. ve Song, Y. S., "Genotype and SNP calling from next-generation sequencing data", *Nature Reviews Genetics*, 12, 443-51, 2011.
- [49] Chaitankar, V., Karakulah, G., Ratnapriya, R., Giuste, F. O., Brooks, M. J. ve Swaroop, A., "Next generation sequencing technology and genomewide data analysis: Perspectives for retinal research", *Prog Retin Eye Res*, 55, 1-31, 2016.
- [50] Lee, H. C., Lai, K., Lorenc, M. T., Imelfort, M., Duran, C. ve Edwards, D., "Bioinformatics tools and databases for analysis of next-generation sequence data", *Brief Funct Genomics*, 11(1), 12-24, 2012.
- [51] Huang, H. W., NISC Comparative Sequencing Program, Mullikin, J. C. ve Hansen, N. F., "Evaluation of variant detection software for pooled next-generation sequence data", *BMC Bioinformatics*, 16, 235, 2015.
- [52] Schlötterer, C., Tobler, R., Kofler, R. ve Nolte, V., "Sequencing pools of individuals - mining genome-wide polymorphism data without big funding", *Nat Rev Genet*, 15(11), 749-63, 2014.
- [53] Anand, S., Mangano, E., Barizzzone, N., Bordoni, R., Sorosina, M., Clarelli, F., ve diğer, "Next Generation Sequencing of Pooled Samples: Guideline for Variants' Filtering", *Sci Rep*, 6:33735, 2016.
- [54] Weldatsadik, R. G., Wang, J., Puhakainen, K., Jiao, H., Jalava, J., Räisänen, K., ve diğer, "Sequence analysis of pooled bacterial samples enables identification of strain variation in group A streptococcus", *Sci Rep*, 7, 45771, 2017.
- [55] Bansal, V., "A statistical method for the detection of variants from next-generation resequencing of DNA pools", *Bioinformatics*, 26(12), i318-i324, 2010.
- [56] Kofler, R., Pandey, R. V. ve Schlötterer, C., "PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq)", *Bioinformatics*, 27(24), 3435-3436, 2011.
- [57] Altmann, A., Weber, P., Quast, C., Rex-Haffner, M., Binder, E. B. ve Müller-Myhsok, B., "vipR: variant identification in pooled DNA using R", *Bioinformatics*, 27(13), i77-i84, 2011.
- [58] Raineri, E., Ferretti, L., Esteve-Codina, A., Nevado, B., Heath, S. ve Pérez-Enciso, M., "SNP calling by sequencing pooled samples", *BMC Bioinformatics*, 13, 239, 2012.

- [59] Pihlstrøm, L., Rengmark, A., Bjørnarå, K. A. ve Toft, M., "Effective variant detection by targeted deep sequencing of DNA pools: an example from Parkinson's disease", *Ann Hum Genet*, 78(3), 243-52, 2014.
- [60] Popp, B., Ekici, A. B., Thiel, C. T., Hoyer, J., Wiesener, A., Kraus, C., ve diğer, "Exome Pool-Seq in neurodevelopmental disorders", *Eur J Hum Genet*, 25(12), 1364-1376, 2017.
- [61] Biancalana, V. ve Laporte, J. "Diagnostic use of massively parallel sequencing in neuromuscular diseases: towards an integrated diagnosis", *J Neuromuscul Dis*, 2(3), 193-203, 2015.
- [62] Sims, D., Sudbery, I., Illott, N. E., Heger, A. ve Ponting, C. P., "Sequencing depth and coverage: key considerations in genomic analyses", *Nat Rev Genet*, 15(2), 121-32, 2014.
- [63] Bizon, C., Spiegel, M., Chasse, S. A., Gizer, I. R., Li, Y., Malc, E. P., ve diğer, "Variant calling in low-coverage whole genome sequencing of a Native American population sample", *BMC Genomics*, 15(1), 85, 2014.
- [64] Li, Y., Sidore, C., Kang, H. M., Boehnke, M. ve Abecasis, G. R., "Low-coverage sequencing: implications for design of complex trait association studies", *Genome Res*, 21(6), 940-51, 2011.
- [65] Navon, O., Sul, J. H., Han, B., Conde, L., Bracci, P. M., Riby, J., ve diğer, "Rare variant association testing under low-coverage sequencing", *Genetics*, 194(3), 769-79, 2013.
- [66] Li, Z., Wang, Y. ve Wang, F., "A study on fast calling variants from next-generation sequencing data using decision tree", *BMC Bioinformatics*, 19(1), 145, 2018.
- [67] Wang, J., Ling, C. ve Gao, J., "CNNDel: Calling structural variations on low coverage data based on convolutional neural networks", *BioMed Research International*, 2017, 6375059, 2017.
- [68] Huang, L., Wang, B., Chen, R., Bercovici, S. ve Batzoglou, S., "Reveel: large-scale population genotyping using low-coverage sequencing data", *Bioinformatics*, 32(11), 1686-96, 2016.
- [69] Fang, L., Hu, J., Wang, D. ve Wang, K., "NextSV: a meta-caller for structural variants from low-coverage long-read sequencing data", *BMC Bioinformatics*, 19, 180, 2018.

- [70] Zhang, J. ve Wu, Y., "SVseq: an approach for detecting exact breakpoints of deletions with low-coverage sequence data", *Bioinformatics*, 27(23), 3228-34, 2011.
- [71] Zhang, J., Wang, J. ve Wu, Y., "An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data", *BMC Bioinformatics*, 13(Suppl 6), S6, 2012.
- [72] Yu, X. ve Sun, S., "Comparing a few SNP calling algorithms using low-coverage sequencing data", *BMC Bioinformatics*, 14, 274, 2013.
- [73] A Brief Guide to Genomics. 15 Ekim 2020, <https://www.genome.gov/about-genomics/fact-sheets/A-Brief-Guide-to-Genomics>.
- [74] Pray, L., "Discovery of DNA structure and function: Watson and Crick", *Nature Education*, 1(1), 100, 2008.
- [75] Human genetic variation | Training Content Hub. 15 Ekim 2020, <https://www.ebi.ac.uk/training-beta/online/courses/human-genetic-variation-introduction/>.
- [76] Talking Glossary of Genetic Terms | NHGRI. 15 Ekim 2020, <https://www.genome.gov/genetics-glossary>.
- [77] T. I. H. Consortium, "The International HapMap Project", *Nature*, 426, 789–796, 2003.
- [78] Devarajan, B., Prakash, L., Kannan, T. R., Abraham, A. A., Kim, U., Muthukkaruppan, V. ve Vanniarajan, A., "Targeted next generation sequencing of RB1 gene for the molecular diagnosis of Retinoblastoma", *BMC Cancer*, 15, 320, 2015.
- [79] Grotta, S., D'Elia, G., Scavelli, R., Genovese, S., Surace, C., Sirleto, P., ve diğ er, "Advantages of a next generation sequencing targeted approach for the molecular diagnosis of retinoblastoma", *BMC Cancer*, 15, 841, 2015.
- [80] Li, W. L., Buckley, J., Sanchez-Lara, P. A., Maglinte, D. T., Viduetsky, L., Tatarinova, T. V., ve diğ er, "A rapid and sensitive next-generation sequencing method to detect rb1 mutations improves care for retinoblastoma patients and their families", *J Mol Diagn*, 18(4), 480–493, 2016.
- [81] Zhang, J., Benavente, C. A., McEvoy, J., Flores-Otero, J., Ding, L., Chen, X., ve diğ er, "A novel retinoblastoma therapy from genomic and epigenetic analyses", *Nature*, 481(7381), 329-34, 2012.

- [82] Kooi, I. E., Mol, B. M., Massink, M. P. G., de Jong, M. C., de Graaf, P., van der Valk, P., ve diğeri, "A meta-analysis of retinoblastoma copy numbers refines the list of possible driver genes involved in tumor progression", *PLoS One*, 11(4), e0153323, 2016.
- [83] Thériault, B. L., Dimaras, H., Gallie, B. L. ve Corson, T. W., "The genomic landscape of retinoblastoma: a review", *Clin Exp Ophthalmol*, 42(1), 33-52, 2014.
- [84] Kooi, I. E., Mol, B. M., Massink, M. P. G., Ameziane, N., Meijers-Heijboer, H., Dommering, C. J., ve diğeri, "Somatic genomic alterations in retinoblastoma beyond *rb1* are rare and limited to copy number changes", *Sci Rep*, 6, 25264, 2016.
- [85] TRUBA Wiki Sayfası. 1 Ekim 2020, http://wiki.truba.gov.tr/index.php/Ana_sayfa.
- [86] Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. 15 Ekim 2020, <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
- [87] Li, H. ve Durbin, R. "Fast and accurate short read alignment with burrows-wheeler transform", *Bioinformatics*, 25(14), 1754-60, 2009.
- [88] Li, H. ve Durbin, R., "Fast and accurate long-read alignment with burrows-wheeler transform", *Bioinformatics*, 26(5), 589-95, 2010.
- [89] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ve diğeri, "The Sequence Alignment / Map format and SAMtools", *Bioinformatics*, 25(16), 2078-9, 2009.
- [90] Picard Tools – By Broad Institute. 10 Ekim 2020, <https://broadinstitute.github.io/picard/>.
- [91] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ve diğeri, "The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data", *Genome Res*, 20(9), 1297–1303, 2010.
- [92] Wang, K., Li, M. ve Hakonarson, H., "Annovar: functional annotation of genetic variants from high-throughput sequencing data", *Nucleic Acids Res*, 38(16), e164, 2010.
- [93] Illumina/hap.py: Haplotype vcf comparison tools. 15 Ekim 2020, <https://github.com/Illumina/hap.py>.

- [94] R: the R Project for Statistical Computing. 15 Ekim 2020, <https://www.r-project.org/>.
- [95] RStudio | Open source & professional software for data science - RStudio. 15 Ekim 2020, <https://rstudio.com/>.
- [96] García-Chequer, A. J., Méndez-Tenorio, A., Olguín-López, G., Sánchez-Vallejo, C., Isa, P., Arias, C. F., ve diğer, "Illumina next generation sequencing data and expression microarrays data from retinoblastoma and medulloblastoma tissues", *Data in Brief*, 6, 908-916, 2016.
- [97] ENA Browser. 1 Ekim 2020, <https://www.ebi.ac.uk/ena/data/view/PRJEB6630>.
- [98] García-Chequer, A. J., Méndez-Tenorio, A., Olguín-Ruiz, G., Sánchez-Vallejo, C., Isa, P., Arias, C. F., ve diğer, "Overview of recurrent chromosomal losses in retinoblastoma detected by low coverage next generation sequencing", *Cancer Genet*, 209(3), 57–69, 2016.
- [99] Home - SRA - NCBI. 1 Eylül 2020, <https://www.ncbi.nlm.nih.gov/sra>.
- [100] The 1000 Genomes Project Consortium, "A global reference for human genetic variation", *Nature*, 526(7571), 68-74, 2015.
- [101] About | 1000 Genomes. 10 Ekim 2020, <https://www.internationalgenome.org/about>.
- [102] Illumina sequencing of HapMap individual NA12878 aligned by BI downsampled to 5x coverage - SRA - NCBI. 10 Ekim 2020, <https://www.ncbi.nlm.nih.gov/sra/?term=SRR622461>.
- [103] NA20355 | IGSR Sample (ERR251661 ve ERR251662). 1 Ekim 2020, <https://www.internationalgenome.org/data-portal/sample/NA20355>.
- [104] Which human reference genome to use? 1 Ekim 2020, <https://lh3.github.io/2017/11/13/which-human-reference-genome-to-use>.
- [105] Human genome reference builds - GRCh38 or hg38 - b37 - hg19 - GATK. 1 Ekim 2020, <https://gatk.broadinstitute.org/hc/en-us/articles/360035890951-Human-genome-reference-builds-GRCh38-or-hg38-b37-hg19>.
- [106] "Quality Scores for Next-Generation Sequencing", Technical Note: Sequencing, Illumina, 2011.
- [107] Ascii Table - ASCII character codes and html, octal, hex and decimal chart conversion. 15 Ekim 2020, <http://www.asciitable.com/>.

- [108] Li, H. ve Homer, N. "A survey of sequence alignment algorithms for next-generation sequencing", *Brief Bioinform*, 11(5), 473-83, 2010.
- [109] Mielczarek, M. ve Szyda, J., "Review of alignment and SNP calling algorithms for next-generation sequencing data", *J Appl Genet*, 57(1), 71-9, 2016.
- [110] bwa.1. 10 Ekim 2020, <http://bio-bwa.sourceforge.net/bwa.shtml>.
- [111] File Format Guide. 15 Ekim 2020, <https://www.ncbi.nlm.nih.gov/sra/docs/submitformats/>.
- [112] Read groups - GATK. 15 Ekim 2020, <https://software.broadinstitute.org/gatk/documentation/article.php?id=6472>.
- [113] Base Quality Score Recalibration (BQSR) - GATK. 5 Eylül 2020, <https://gatk.broadinstitute.org/hc/en-us/articles/360035890531-Base-Quality-Score-Recalibration-BQSR->.
- [114] Resource bundle - GATK. 5 Ekim 2020, <https://gatk.broadinstitute.org/hc/en-us/articles/360035890811-Resource-bundle>.
- [115] Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., ve diğer, "From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline", *Curr Protoc Bioinformatics*, 43, 11.10.1–11.10.33, 2013.
- [116] Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., ve diğer, "Scaling accurate genetic variant discovery to tens of thousands of samples", *bioRxiv*, 2017.
- [117] HaplotypeCaller - GATK. 5 Eylül 2020, <https://gatk.broadinstitute.org/hc/en-us/articles/360042913231-HaplotypeCaller>.
- [118] I am unable to use VQSR (recalibration) to filter variants - GATK. 9 Eylül 2020, <https://gatk.broadinstitute.org/hc/en-us/articles/360037499012-I-am-unable-to-use-VQSR-recalibration-to-filter-variants>.
- [119] Variant Quality Score Recalibration (VQSR) - GATK. 9 Eylül 2020, <https://gatk.broadinstitute.org/hc/en-us/articles/360035531612-Variant-Quality-Score-Recalibration-VQSR->.
- [120] GATKwr12-6-Variant_filtering.pdf. 15 Ekim 2020, https://qcb.ucla.edu/wp-content/uploads/sites/14/2016/03/GATKwr12-6-Variant_filtering.pdf.

- [121] Cornish, A. ve Guda, C., "A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference", BioMed Research International, 2015, 456479, 2015.
- [122] Genome in a Bottle | NIST. 15 Ekim 2020, <https://www.nist.gov/programs-projects/genome-bottle>.
- [123] ANNOVAR Documentation. 12 Eylül 2020, <https://doc-openbio.readthedocs.io/projects/annovar/en/latest/>.
- [124] RefSeq: NCBI Reference Sequence Database. 15 Ekim 2020, <https://www.ncbi.nlm.nih.gov/refseq/>.
- [125] Home - SNP - NCBI. 15 Ekim 2020, <https://www.ncbi.nlm.nih.gov/snp/>.
- [126] E-MTAB-3515 < Browse < ArrayExpress < EMBL-EBI. 1 Ekim 2020, <https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-3515/>.
- [127] Tomar, S., Sethi, R., Sundar, G., Quah, T. C., Quah, B. L. ve Lai, P. S., "Mutation spectrum of *rb1* mutations in retinoblastoma cases from singapore with implications for genetic management and counselling", PLoS One, 12(6), e0178776, 2017.
- [128] What's the Difference Between a CPU vs. a GPU | NVIDIA BLOG. 5 Ekim 2020, <https://blogs.nvidia.com/blog/2009/12/16/whats-the-difference-between-a-cpu-and-a-gpu/>.
- [129] Klus, P., Lam, S., Lyberg, D., Cheung, M. S., Pullan, G., McFarlane, I., ve diğeri, "BarraCUDA - a fast short read sequence aligner using graphics processing units", BMC Research Notes, 5, 27, 2012.
- [130] NVBIO: nvBowtie. 20 Eylül 2020, https://nvlabs.github.io/nvbio/nvbowtie_page.html.
- [131] Liu, Y., Schmidt, B. ve Maskell, D. L., "CUSHAW: a CUDA compatible short read aligner to large genomes based on the Burrows–Wheeler transform", Bioinformatics, 28(14), 1830-7, 2012.
- [132] Liu, Y. ve Schmidt, B., "CUSHAW2-GPU: Empowering Faster Gapped Short-Read Alignment Using GPU Computing", IEEE Design & Test, 31(1), 31 - 39, 2013.

[133] Buntara, F., Lee, B.-S., Purbojati, R. W. ve Zhou, C. X., "Is GPUs Ready to Boost Genomic Alignment Computation", 2019 International Conference on Innovative Trends in Computer Engineering (ITCE), Egypt , 2019.

[134] <ftp://ftp.sra.ebi.ac.uk/> dizini. 5 Eylül 2020, <ftp://ftp.sra.ebi.ac.uk/>.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Gülistan ÖZDEMİR ÖZDOĞAN

Doğum Tarihi : 05.03.1983

E-mail : guldemiroz@gmail.com

EĞİTİM BİLGİLERİ

Üniversite : Çankaya Üniversitesi, Bilgisayar Mühendisliği, 2006, Ankara

Yüksek Lisans : TOBB Ekonomi ve Teknoloji Üniversitesi, Bilgisayar Mühendisliği, 2010, Ankara

İŞ DENEYİMİ

Yazılım Uzmanı : Altay Grup, Ankara, (2007)

Burslu Yüksek Lisans Öğrencisi : TOBB Ekonomi ve Teknoloji Üniversitesi, Ankara, (2007-2010)

Öğretim Görevlisi : Çankaya Üniversitesi, Ankara, (2011-2015)

Şirket Müdürü, Kurucu Ortak : VARATAH TEKNOLOJİ Çözümleri Bilgisayar Sistemleri Yazılım ve Bilişim Limited Şirketi, Ankara, (2013-2017)

İLGİ ALANLARI

- Biyoinformatik
- Veri Madenciliği
- Paralel Hesaplama
- Makine Öğrenme

TEZDEN ÜRETİLEN YAYINLAR / BİLDİRİLER

1. Özdemir Özdoğan G., Kaya H., Şen B., Çankaya İ. “A Survey on Predicting Survivability of Retinoblastoma on SEER Data”. International Conference on Advanced Technologies, Computer Engineering and Science (ICATCES’18), 257-261, 2018.
2. Özdemir Özdoğan G. ve Kaya H. “Next-Generation Sequencing Data Analysis on Pool-Seq and Low-Coverage Retinoblastoma Data”. Interdisciplinary Sciences: Computational Life Sciences, 12(3), 302–310, 2020.
3. Comparison of Alignment Tools BWA and BarraCUDA on set of NGS Datasets (hazırlık aşamasında)

DİĞER YAYINLAR

1. Özdemir Özdoğan G., Abul O. ve Yazıcı A. “Paralel Veri Madenciliği Algoritmaları”. Başarım 09 - 1. Ulusal Yüksek Başarım ve Grid Konferansı, 131-137, Ankara, 2009.
2. Özdemir Özdoğan G. ve Abul O. “Task-Parallel FP-Growth on Cluster Computers”. 25th International Symposium on Computer and Information Sciences, 383-388, İngiltere, 2010.