Hastalık Tanısı İçin Yeni Nesil Dizileme Verisi Analizi: Gereksinimler ve Bir Çözüm Önerisi

Orçun Taşar[1], Esra Çınar[1] ve Hüseyin Onay[2]

[1] İdea Teknoloji Çözümleri, İstanbul, Türkiye
[2] Ege Üniversitesi Tıp Fakültesi, İzmir, Türkiye
orcun.tasar@ideateknoloji.com.tr

Özet. DNA dizileme teknolojilerinin ucuzlayarak yaygınlaşması genetik ve tıp araştırmalarının yörüngesini değiştirmiştir. Artık bu çalışmaları, Yeni Nesil Genom Dizileme sistemlerinden elde edilen büyük çaptaki veriyi anlamlandıran bilgi ve iletişim teknolojilerini kullanmadan gerçekleştirmek imkansız hale gelmiştir. Bu çalışmada, hızla çoğalan genomik verinin güvenli saklanması, erişilmesi ve analizi için gereksinimler tanımlanmış ve devam etmekte olan bir TEYDEB projesi kapsamında geliştirilmek üzere önerdiğimiz çözüm detaylandırılmıştır. "Sık Gözlenen Yenidoğan Metabolik Hastalıklarının Hızlı Tanısı için Yeni Nesil Dizi Analizi Temelli Kit Geliştirilmesi" başlıklı TEYDEB projesi kapsamında, genomik tabanlı belirteçlerin kullanılmasıyla sık gözlenen 10 metabolik hastalığın tanısı için genetik tabanlı testleri içeren bir kit geliştirilecektir. Bu çalışmada, Oxford Nanopore MinION cihazı için geliştirilecek kitin çıktısı olan verinin tanı amacıyla analizinin otomatize edilmesi; bu verilerin saklanması ve veriye erişilmesine dair önerdiğimiz çözümler sunulmaktadır.

Anahtar Kelimeler: Yeni Nesil Dizileme Verisi Analizi, Yeni Doğan Tarama Programı, Nanopor, Genetik Varyant

Diagnostic Next Generation Sequencing Data Analysis for Variant: Requirements and a Proposition

Abstract. The rapid advance in genome sequencing technology has led to the decrease of sequencing costs and hence, to the production of vast amount of genomic data. It is not possible to process the large amount of Next-Generation Sequencing Data without the use of information and communication technologies. In this study, we discuss the requirements with respect to the secure storage, access and analysis of rapidly expanding genomic data, and we present the solution we propose as part of an ongoing TEYDEB project. Within the scope of the TEYDEB project titled "Development of a fast kit for the diagnosis of frequent newborn metabolic diseases with Next-Generation Sequencing data analysis", a DNA testing kit for the diagnosis of 10 metabolic diseases will be developed using genomic markers. In this study, we propose a solution for the storage and automatization of data analysis of the data generated by Oxford Nanopore MinIon.

Keywords: Next Generation Sequencing Data Analysis, Newborn Screening Tests, Nanopore, Genetic Variants

1 Giriş

Bu çalışmada, TEYDEB destekli "Sık Gözlenen Yenidoğan Metabolik Hastalıklarının Hızlı Tanısı için Yeni Nesil Dizi Analizi Temelli Kit Geliştirilmesi" projesi kapsamında, genomik tabanlı belirteçlerin kullanılmasıyla hastalık tanısı için, veri analizinin otomatize edilmesi; verilerin aktarılması; tutulması ve sonuçların kolay anlaşılmasına dair konular tartışılmakta ve bir çözüm önerisi sunulmaktadır. Proje kapsamında MinION platformunda kullanılmak üzere on metabolik hastalık için bir yenidoğan tarama kiti geliştirilecektir. Oxford Nanopore Technologies (ONT) şirketinin geliştirdiği MinION, nanopor aracılığı ile direkt tek molekül dizilimi yapan ticarileşmiş ilk üründür. DNA örneği cihaza yüklendikten sonra çıkan ham verinin analiz edilerek hastalığa sebep olan mutasyonların varlığına dair bir sonuca ulaşılması hedeflenmektedir. Literatürde yenidoğan taramada ve doğuştan metabolik hastalıkların tanısında yeni nesil dizi analizi kullanımına yönelik az sayıda çalışma bulunsa da, MinION ile hızlı bir şekilde sonuç veren bir uygulamaya örnek bir çalışma bulunmamaktadır. Ayrıca, henüz nanopor teknolojisi ile elde edilen verilerin analizi için standart hale gelmiş bir uygulama yoktur. Kitin klinik alanda pratik uygulamaya geçmesi için kullanımının kolay olması ve yeterince hızlı olması gerekmektedir. Bu amaçla, MinION ile elde edilen genetik verinin analizi için otomatik bir çözüm geliştirilmesine; kullanım kolaylığı sağlamak için farklı kullanıcılar düşünülerek arayüzler tasarlanmasına ve saptanan varyantların raporlandığı bir arayüzün sisteme entegre edilmesine ihtiyaç duyulmaktadır.

Varolan yenidoğan tarama programlarında topuktan alınan kan örneği ile test yapılmaktadır. Bu testlerin sonuçları bazı hastalıklar için çocuğun beslenmesine bağlıdır. Örneğin Fenilketonüri taraması için bebeğin en az 48 saat beslenmiş olması gerekir. Hastaneden çıkış zamanı, bebeğin yoğun bakımda olması ve parenteral beslenmesi durumlarında bu testi yaptırmak için beklemek veya tekrardan sağlık merkezine başvurmak gerekecektir. Biyotinidaz enzim eksikliği taramasında ise 48 saatlik numunenin taranması sonucu şüpheli olabilecek bir durum varsa ikinci kere kan örneği alınması gerekebilmektedir. Geliştirilecek olan bu kit, minimal invaziv girişimle tanıya imkan sağlayacak, klinisyeni laboratuvar şartlarının sağlanması mecburiyetinden kurtaracaktır. Doğuştan metabolik hastalıkların taramasında kullanılan testlere göre çok daha duyarlı bir test oluşturulurken, bu amaçla kullanımında kit yüksek doğruluk oranıyla, doğrulama testi ihtiyacını ortadan kaldıracaktır. Doğuştan metabolik hastalıkların tanısında kullanılan çok sayıda testin yerine aynı anda hızlı bir değerlendirme imkanı sunacak, bu sayede hem zaman kaybından klinisyeni korurken, hem de maliyet etkin bir çözüm sağlayacaktır. Bu kit sayesinde hayati tehtid altındaki yenidoğanın hızlı tanısı ve sonunda hızlı tedavisine imkan verecektir.

Hastalara doğru ve hızlı genetik tanı koyulabilmesi amacıyla genom üzerindeki varyantların doğru bir şekilde saptanması gerekmektedir. Çalışmamızda bu amaca yönelik olan, DNA dizilemesi; dizileme verilerinin analiz edilmesi ile varyantların saptanması ve raporlanması; ve elde edilen verilerin güvenli saklanması konularına

yer verilmiştir. Bölüm 1.1'de genom dizileme teknolojisinde yaşanan gelişmeler ve projemiz için seçilen nanopor teknolojisine dair detaylar anlatılmaktadır. Dizileme verilerinin analiz sürecinin, rutin tanı amacıyla standart hale getirilmesi için önerilerimiz, ilgili veri formatları ve veri analizindeki önemleri çerçevesinde Bölüm 2'de anlatılmaktadır. MinION verileri için optimize edeceğimiz analiz sürecinin, bulut tabanlı bir platform üzerinde hızlı ve güvenli bir şekilde yürütülmesi amacıyla tasarladığımız web uygulamasının detayları Bölüm 3'te verilmiştir.

1.1 Genom Dizileme Teknolojisindeki Gelişmeler

Genom üzerindeki varyantların doğru ve sağlıklı bir şekilde saptanabilmesi, dolayısıyla da hastalara doğru ve hızlı bir tanı koyulabilmesi için, DNA dizileme aşaması çok önemlidir. Bu bölümde, dizileme teknolojileri alanında yaşanan teknolojik gelişmeler anlatılacaktır ve projemiz için seçtiğimiz MinION dizileme platformuna dair detaylar verilecektir.

İnsan Genom Projesi hakkındaki ilk bulgular 2001 yılında paylaşılmış ve proje 2003 yılında başarılı bir şekilde tamamlanmıştır [1, 2]. Projenin üzerinden geçen son 15 yılda genomik bilimi devrimsel bir sürece girmiş ve uygulama alanlarında kullanılan teknolojilerde büyük bir gelişme yaşanmış olup, bu gelişmeler genetik alanında yapılan çalışmaların hızlanmasına ve genetik tanının daha hızlı ve daha az maliyetle koyulabilmesine olanak sağlamıştır.

İnsan Genom Projesi, 1977 yılında Frederick Sanger tarafından yayınlanan ve birden fazla insanın Nobel ödülü almasını sağlayan bir DNA dizileme yöntemi kullanılarak gerçekleştirilmiştir [3, 4]. Bu yöntem günümüzde Sanger Dizileme olarak anılmaktadır. 1987 yılında Applied Biosystems tarafından bu dizileme tekniğini otomatize hale getiren ilk cihaz üretilmiş (ABI 370) ve bu gelişmeyle birlikte 1990 yılında Amerika Enerji Kurumu ve Ulusal Sağlık Enstitüsü (National Institutes of Health – NIH) tarafından İnsan Genom Projesi başlatılmıştır. Proje yaklaşık 13 yılda tamamlanmış ve 2.7 milyar dolara mal olmuştur [5]. Sanger dizileme yöntemi her ne kadar farklı boyutlardaki DNA fragmanlarının dizilenmesi konusunda esneklik sağlasa da yüksek maliyetler karşılığında uzun sürelerde düşük çıktılar vermesi genomik alanını yeni nesil dizileme yöntemlerine doğru itmiştir. Sanger dizileme yöntemi günümüzde hala doğrulama amacıyla kullanılmaktadır ama DNA dizileme alanındaki yarış artık yeni nesil teknikler kullanan cihazlar arasında gerçekleşmektedir.

Yeni nesil DNA dizileme yöntemleri biyokimyasal mekanizmalar bazında değişkenlik gösterse de temelde hepsi aynı işlemi gerçekleştirmeye çalışmaktadır: kitlesel bir dizilemenin aynı anda paralel bir şekilde gerçekleştirilmesi. Bu fikirle 2000 yılında piyasaya sürülen ilk yöntemlerden birinin Lynx Therapeutics tarafından geliştirilen 'kitlesel imza dizileme' yöntemi olduğunu görüyoruz. Bu süreci 2005 yılında piyasaya sürülmüş ve birçok çalışmada yeni nesil DNA dizilemenin başlangıcı olarak kabul edilen Roche 454 'pirodizileme', 2007 yılında Illumina'nın 'sentez ile dizileme' ile Applied Biosystem tarafından geliştirilen 'ligasyon ile dizileme' ve 2010 yılında Life Technologies tarafından geliştirilen 'iyon yarı iletken' sistemleri takip ettiler [6]. Şu an yeni nesil DNA dizileme pazarının en büyük üreticisi konumunda olan Illumina, MiniSeq, MiSeq, NextSeq 500/550, HiSeq serisi ve NovaSeq 6000 gibi

‘sentez ile dizileme’ sistemini kullanan dizileme platformu modelleriyle hedeflenmiş gen panellerinden tüm ekzom ya da tüm genom çalışmalarına ideal olacak şekilde farklı çözümler sunmaktadır. 1987 yılında üretilen ilk Sanger dizileme cihazı olan ABI 370’in günlük DNA dizisi okuma kapasitesi yaklaşık 400.000 nükleotit (400Kb) iken [7], bugün Illumina NovaSeq 6000 cihazı tek bir yürütmede 3 trilyon nükleotit (3000Gb) okuma kapasitesi ile aynı anda 30x derinlikte 48 insan genomunu 44 saat içinde rahatlıkla okuyabilmektedir [8].

Yeni nesil DNA dizileme yöntemleri Sanger dizilemeye göre nükleotit başına okuma maliyetini önemli ölçüde düşürüp, okuma hızını ve çıktı boyutunu arttırmış olsalar da 75 – 150 baz gibi kısa okuma uzunluklarından dolayı, DNA’daki kopya sayısı değişikliklerini (copy number variations – CNV), büyük insersiyon veya delesyon gibi yapısal varyantları saptama konusunda zayıf kalabilmektedirler. Yeni nesil DNA dizileme yöntemlerinin genetik alanında açtığı çığır ve hala daha duyulan farklı ihtiyaçlar DNA dizileme üzerine yeni yaklaşımların oluşmasını sağlamaya devam etmektedir.

Oxford Nanopore Teknolojisi. Günümüzde Oxford Nanopore firmasının ürettiği cihazlar polimeraz zincir reaksiyonu ile DNA kalıplarının çoğaltılmasına ihtiyaç duymadan tek molekül üzerinden 150 bin baza kadar başarılı bir şekilde uzun okumalar yapabilmektedir. Çalışma kapsamında tarafımızca kullanılacak olan Oxford Nanopore dizileme platformu, MinION, platform üzerinde dizileme işlemi devam ederken aynı anda veri analizine de olanak sağlamaktadır. Fiziksel boyut, okuma hızı ve stratejisi, eş zamanlı analiz süreçlerinde on beş dakikada tanı koyabilme, DNA dizileme çalışmasından metilasyon paterni gibi epigenom bilgilerinin elde edilebilmesi gibi özelliklerinden dolayı bu platformların kullandığı sistemler üçüncü nesil dizileme platformları olarak da adlandırılmaktadır [9, 10].

MinION dizileme platformu10x2x3 cm boyutlarında, 90 gr ağırlığında küçük bir cihaz olup, okuma için gereksinim duyduğu enerjiyi USB 3.0 bağlantısından sağlayabilmektedir. DNA dizisi okuma stratejisi, yaklaşık 1 nm çapındaki porlardan geçen DNA fragmanlarının iletken ortamda yarattığı pikoamper (pA) boyutundaki elektriksel değişikliklere dayanmaktadır. Her bir nükleotit elektrik akımında farklı paternlerde değişikliğe neden olduğundan dolayı, nükleotitler porlardan geçerken tanımlanabilmektedir. Bu değişiklikler por etrafında bulunan sensörler tarafından okunup eş zamanlı olarak kaydedilmektedir. Buna benzer şekilde, küçük porlardan geçirilen DNA fragmanlarının herhangi bir senteze veya ligasyona ihtiyaç duymadan direkt olarak dizilenebilmesi üzerine teoriler 1990’ların başından beri tartışılmakta olan bir konuydu, Oxford Nanopore firması ise projeye 2007 yılında başlamış ve 2014 yılında MinION Kabul Programı (MinION Access Program – MAP) ile platformun araştırma gruplarınca kullanılmasını sağlamıştır. MinION şu an ticari olarak satışta olan bir dizileme platformudur [11, 12].

Çalışmamız kapsamında MinION dizileme platformu ve verisi ile çoğunlukla çocukluk çağında başlangıç gösteren ve büyük kısmı tek gen defekti kaynaklı yenidoğan metabolik hastalıklarının tanısının hızlı, kolay ve başarıyla gerçekleştirilebilmesi için TEYDEB-1511 proje desteği ile bir çözüm geliştirilmesi hedeflenmektedir. Tanının olabilecek en hızlı şekilde yapılabilmesi projenin öncelikleri arasındadır. Bu sebeple ıslak laboratuvarda kullanılacak hızlı bir dizileme protokolü için, yaklaşık on dakikalık bir ön hazırlığa ihtiyaç duyan MinION dizileme

platformu seçilmiştir. Dizileme platformundan elde edilecek verinin sağlıklı ve tutarlı analizinin hızlı ve performanslı bir ortamda gerçekleştirilmesi için önerdiğimiz çözümler Bölüm 2 ve 3'te anlatılmıştır.

Çalışma kapsamında MinION platformu ile analizi gerçekleştirilecek sık gözlenen yenidoğan metabolik hastalıklarının listesi ve ilgili genler Tablo 1'de verilmiştir. Mendeliyen tipinde olan bu hastalıkların referans genlerine dair bilgiler Online Mendelian Inheritance in Man (OMIM) veritabanında yer almaktadır.

Tablo 1. Çalışma kapsamında MinION platformu ile analizi gerçekleştirilecek sık gözlenen on yenidoğan metabolik hastalıkları.

Yenidoğan Metabolik Hastalıkları	**İlgili Genler**
İzovalerik asidemi	IVD
Metil malonik asidemi	MUT
Fenilketonüri	PAH
Maple şurup idrar hastalığı	DBT, BCKDHB, BCKDHA
Tirozinemi tip 1	FAH
Biotinidaz eksikliği	BTD
Galaktozemi	GALT
Glukojen depo hastalığı tip 2	GAA
Mukopolisakkaridoz tip 1	IDUA
Nieman-Pick Hastalığı	SMPD1, NPC1, NPC2

2 Veri Formatları ve Analiz Süreci

MinION verisini kullanan ve günümüzde rutin tanı için standart hale gelmiş bir analiz süreci mevcut değildir. Bunun sebeplerinden biri, her ne kadar kullanımı yaygınlaşmaya başlamış da olsa, MinION dizileme platformunun uygulanabilirlik açısından yeni bir teknoloji olmasıdır. Elde edilen verinin yorumlanmasına dair günümüzde farklı bakış açıları ve farklı algoritmalarla tasarlanmış biyoenformatik araçlar mevcuttur fakat daha önce yapılan çalışmalarda çoğunlukla bu araçların verimli bir şekilde bir araya getirilmesinden ziyade MinION dizileme platformunun hangi amaçlarla kullanılabileceğine, verinin nasıl üretildiğine ve bu verinin nasıl işlenebileceğine dair kısıtlı gözlemlere odaklanılmıştır. Biz çalışmamızın ilk kısmında tamamıyla MinION verisinin nasıl daha verimli ve tutarlı bir şekilde ele alınabileceği üzerinde duracak ve rutin tanı için bir çözüm sunacağız.

Dizileme sonrası elde edilen ham verilerin varyant bilgisine dönüştürülmesi için biyoenformatik analizlerin yapılması gerekmektedir. Bu analizler, baz çağırma; kalite tayini; adaptör kırpılması; hizalama; varyant çağırma; anotasyon; görselleştirme gibi aşamaları içermektedir. Bu bölümde, çeşitli veri formatları, veri analizi sürecindeki önemleri açısından anlatılmaktadır.

Dizileme cihazlarından alınan ilk ham veriler çoğunlukla fluoresan rengi veya şiddeti, ph veya elektrik akımı değişimi gibi analitik verilerdir ve genetik veri analizine başlanabilmesi için elde edilmiş olan bu ham verilerin nükleotit bilgilerine çevrilmesi gerekmektedir (baz çağırma – base calling). Bu aşama dizileme sırasında platform içerisinde gerçekleşebileceği gibi, çevrimdışı bir şekilde ilgili

biyoenformatik araçları ile de düzenlenebilir. MinION platformunun verdiği ham veri porlar etrafına konumlanmış olan sensörlerin, DNA fragmanlarının por içinden geçişi sırasında detekte ettikleri elektriksel değişimlerdir ve bu bilgiler bir HDF5 (hiyerarşik veri formatı 5) [13] standardına dayanan FAST5 dosya formatında kaydedilir. MinION akış hücresi üzerindeki porlardan geçen her bir DNA fragmanı için sadece o okumaya özgü benzersiz bir FAST5 dosyası yaratılır. Bu dosya içerisinde, DNA fragmanının okunması sırasında por içindeki iletken ortamda yer alan elektrik akımının zamana karşı pA cinsinden değerleri ve okumaya dair meta veriler hiyerarşik bir biçimde yer alabilir. FAST5 dosyaları dizileme işlemi boyunca oluşturuldukça USB 3.0 bağlantısı üzerinden kullanıcı bilgisayarına iletilmekte, bu da dizileme devam ederken eş zamanlı analiz imkanı tanımaktadır.

FAST5 dosya formatı içerisinden genetik bilgilere ulaşabilmek için öncelikle bu verilerin nükleotit bilgilerine çevrilmesi gerekmektedir. Bunun için Oxford Nanopore'un, bulut tabanlı ve çevrimiçi kullanılabilen Metrichor ve yine çevrimdışı kullanılmak üzere Albacore adında iki aracı mevcuttur. Bundan sonraki her aşamada da olacağı gibi baz çağırma aşamasında da üçüncü parti araçlar mevcuttur. Çalışmamız kapsamında bizim ilk önceliğimiz çevrimdışı araçları kullanmak olacaktır.

Baz çağırma aşamasının ardından elde edilen veri formatı FASTQ olacaktır. FASTQ günümüzde biyoenformatik araçların çoğunlukla kullandığı ve oldukça kabul görmüş bir ilk girdi formatıdır. FAST5 formatının aksine bir FASTQ dosyasında binlerce DNA fragmanı okuması okuma kalitesi skorları, platform bilgileri, okumanın akış hücresi üzerinde hangi koordinattan geldiği gibi bilgiler ile beraber saklanabilir [14].

DNA fragmanlarının uçlarına MinION cihazına yüklenmeden önce laboratuvar ortamında adaptörler eklenmekte ve sonra tüm örneklerden gelen bütün DNA fragmanları aynı ortamda homojen bir şekilde karıştırılarak MinION üzerindeki akış hücresine aktarılmaktadır. Bu adaptörler hangi fragmanın hangi örnekten geldiğini belirten bir etiket görevini görür. Baz çağırma işlemi sırasında elde edilen ilk FASTQ dosyalarında da bu adaptör dizilerinin bilgileri bulunmaktadır. Referans genom hizalama aşamasında adaptör dizileri yanlış skorlamaya ve dolayısıyla hizalanmama durumuna sebebiyet vermemeleri için okuma uçlarından kırpılacaktır.

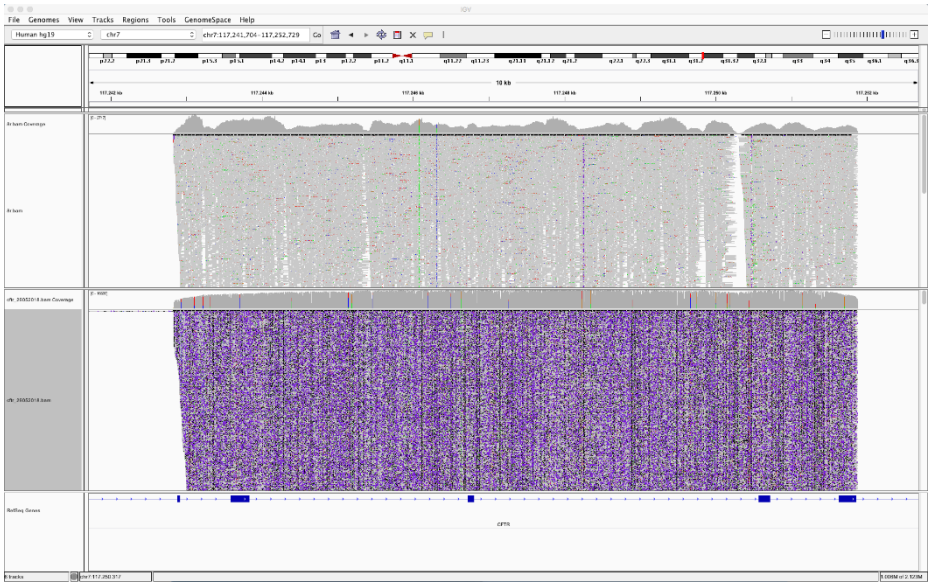
FASTQ dosyalarında elde edilmiş olan DNA okuma bilgilerinin referans genoma hizalanması için günümüzde en yaygın kabul görmüş hizalama aracı olan, Burrows-Wheeler Transform sıkıştırması ve Smith-Waterman algoritmasını kullanan Burrows-Wheeler Aligner (BWA) hizalayıcısının [15], her ne kadar 100 – 150 bazlık kısa okumalar için ideal olsa da daha uzun olan Oxford Nanopore okumaları üzerinde de yüksek performansla çalıştığı daha önce gösterilmiştir. Özellikle PacBio ve Oxford Nanopore gibi uzun okumalar için tasarlanmış olan Minimap2 hizalayıcısının doğruluk oranının BWA ile eşdeğer olduğu fakat dört kat daha hızlı çalıştığı da yapılan çalışmalarda belirtilmiştir. Çalışmamızda FASTQ dosyalarındaki Oxford Nanopore okumalarının referans genomla eşleşmesinin özellikle bu iki hizalayıcı ile gerçekleştirilmesini planlanmaktayız. Ayrıca, GraphMap, MarginAlign gibi Oxford Nanopore komünitesi içerisinde geliştirilen diğer hizalayacılar ile de performans ve doğruluk oranı açısından karşılaştırmalar yapılacaktır.

MinION verisinin en büyük handikapı şüphesiz ki hata oranının ikinci nesil dizileme platformlarına göre daha yüksek olmasıdır [9]. *CFTR* geni 17. ve 20.

ekzonlar arasında kalan bölge üzerinde Illumina NextSeq500 ve MinION tarafından gerçekleştirilen dizileme verileri arasındaki fark Şekil 1'de gösterilmiştir. Referans genoma hizalanmış DNA okumaları üzerinde bir hata doğrulaması (error-correction) yapılması tutarlı analiz sonuçlarına ulaşabilmek için oldukça önemlidir. Bunun için Saklı Markov Modeli (Hidden Markov Model) ve Yinelenen Sinir Ağı (Recurrent Neural Network) algoritmaları kullanılarak okumalar üzerinde düzeltme çalışmaları yapılacaktır. [9, 11].

FASTQ dosyasında saklanan DNA okumaları referans genoma hizalandıktan sonra, tüm hizalanan okumalar, referans genomun hangi bölgesiyle eşleştikleri bilgisiyle birlikte SAM (Sequence Alignment Mapping) formatına kaydedilecektir ama SAM dosyalarının boyutları yüksek olduğu için bu boyutu küçültmek ve daha sonraki analiz aşamalarının daha hızlı gerçekleşebilmesi için bu dosyalar BAM (Binary Alignment Mapping) formatına dönüştürülecektir [16]. Bu dönüşüm işlemi için kabul görmüş araçlardan ikisi Samtools ve Picard paketlerinde mevcuttur [17, 18]. BAM formatına sıkıştırılmış DNA okumalarının görsel olarak incelenmesinin Integrative Genomics Viewer (IGV) programı aracılığı ile yapılması planlanmaktadır [19].

Varyant çağırma (variant calling) aşaması, referans genoma hizalanmış okumalardan varyantların çıkarılması için yapılan bir hesaplama işlemi olup işlem sonucunda varyant bilgileri VCF (Variant Calling Format) [20] dosyalarına kaydedilecektir. Analiz sürecindeki bu basamak için Oxford Nanopore komünitesi içerisinde Nanopolish, Poreseq, MarginCaller gibi araçlar geliştirilmiştir. Bu araçların farklı veri setleri için tek tek denenerek doğruluk ve performans açısından karşılaştırmalarının yapılacaktır.



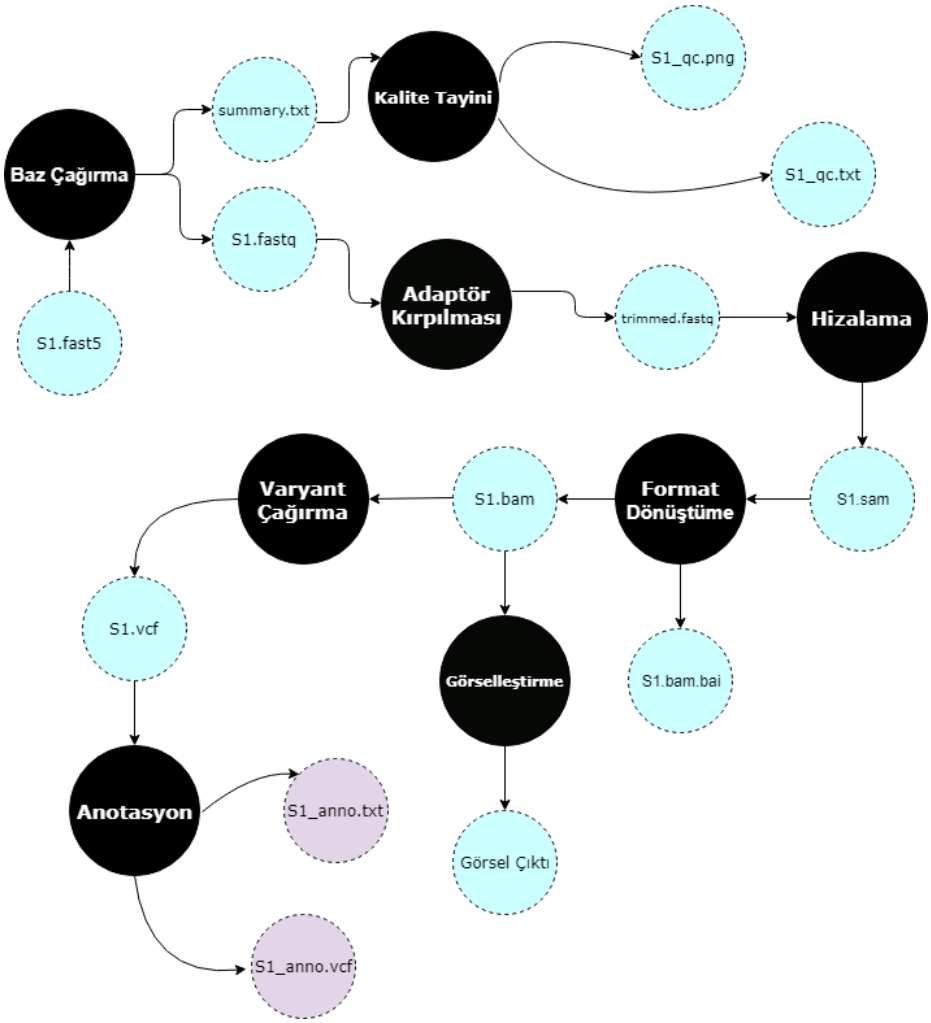
Şekil 1. *CFTR* geninin 17. ve 20. ekzonları arasındaki bölgesi için Illumina ve MinION verilerinin karşılaştırılması. Üstte yer alana 8r.bam dosyasında Illumina okumaları, alttaki cftr_29052018.bam dosyasında MinION okumaları yer almaktadır. MinION okumalarında çok fazla sayıda artefakt mevcuttur.

Tablo 2. Oxford Nanopore verilerinin analizi için kullanılabilecek biyoenformatik araçlardan bazıları [9].

Biyoenformatik Araç	**Uygulama Alanı**
BWA	Hizalama
Minimap2	Hizalama. Özellikle Nanopore ve PacBio için geliştirildi.
GraphMap	Hizalama. Uzun DNA okumalar için geliştirildi.
MarginAligner	Hizalama. Özellikle Nanopore için geliştirildi.
LAST	Hizalama. Uzun DNA ve RNA okumaları için geliştirildi.
Albacore	Baz çağırma.
DeepNano	Baz çağırma.
Poretools	Baz çağırma, format dönüştürme, görselleştirme.
Porechop	Adaptör kırpma.
ALEC	Hata düzeltme.
NanoCORR	Hata düzeltme.
Nanocorrect	Hata düzeltme.
NanoOK	Hata düzeltme, kalite tayini
minION_QC	Kalite tayini
PoreSeq	Hata düzeltme, varyant çağırma.
Nanopolish	Varyant çağırma.
MarginCaller	Varyant çağırma.

Elde edilen varyant setlerinin klinik olarak yorumlanması için gerçekleştirilecek anotasyon aşaması oldukça önemlidir. Bu aşama için biyoenformatik komünitesi tarafından geliştirilmiş Variant Effect Predictor (VEP), Annovar, SnpEff gibi araçlar mevcuttur. Bu araçlar Polyphen, SIFT, MutationTaster gibi tahmini patojenite skoru veren veri tabanlarından skor bilgisini çekerken, ExAC, 1000 Genome gibi veri tabanlarından da ilgili varyantın hangi populasyonlarda ne sıklıkla gözlendiğinin bilgisini alıp VCF dosyalarına işleyebilmektedir. Ayrıca kanıta dayalı varyant yorumlamasını da VCF dosyalarına ekleyebilmek için ClinVar veya Human Genome Mutation Database (HGMD) gibi veri tabanlarının analiz sürecimize entegre edilmesi de önceliklerimiz arasındadır.

MinION platformundan elde edilecek verinin analizi için kullanılabilecek araçlar Tablo 2'de, bu araçlar kullanılarak oluşturulacak analiz süreci süreç hakkında daha önce anlattıklarımız doğrultusunda Şekil 2'de özetlenmiştir.

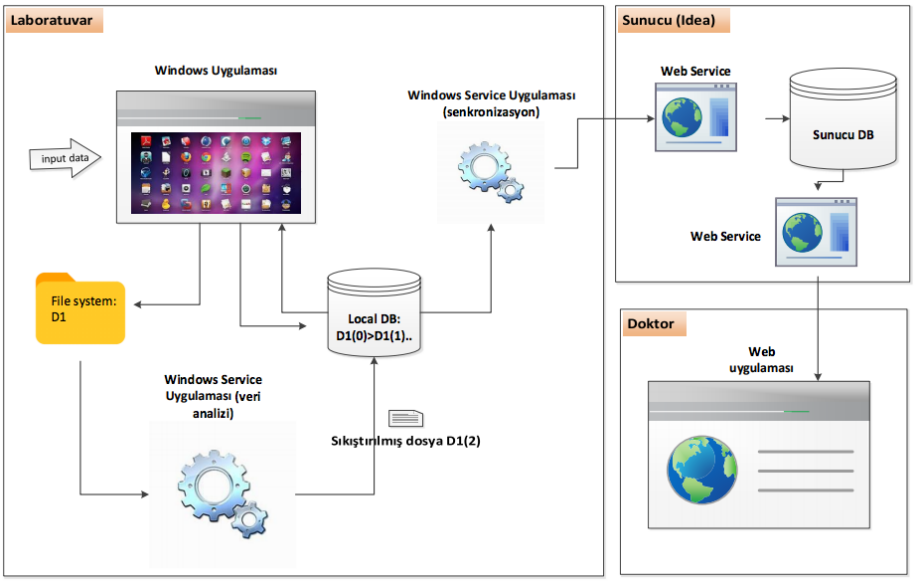


Şekil 2. MinION platform çıktıları için dizayn edilen analiz akış şeması.

3 Analiz Süreci İçin Bir Web Uygulaması

Analiz sürecinde, MinION dizileme platformundan elde edilen verinin çeşitli hesaplama aşamalarından geçerken farklı formatlara dönüştürülmesi gerekmektedir. Bu işlemler karmaşık olduğundan dolayı yüksek bilgisayar gücüne ihtiyaç duyulmaktadır. Farklı algoritmalarla tasarlanmış biyoenformatik araçları GPU veya CPU temelli çalışabildiklerinden dolayı farklı konfigürasyonlardaki bilgisayarlarda farklı performanslarla sonuç verebilmektedirler. Tüm bunlar analiz sürecini yerel bir bilgisayar üzerinde yavaşlatabilen etkenlerdir. Çalışmamızda, bu sebeplerden dolayı analiz sürecinin bulut tabanlı bir platform üzerinde hızlı ve güvenli bir şekilde yürütülmesi hedeflenmektedir.

MinION verileri için optimize edeceğimiz analiz sürecinin otomatize hale getirilmesi ve sadece tek tuşla uzaktan tüm bu sürecin bulut üzerinde çalıştırılıp hızlı bir şekilde sonuç alınabilmesi hedeflerimiz arasındadır. Bunun için veri analizi süreci bir bulut sistemi üzerinde yürütülecek, analiz sürecinin kontrolü ve son anote edilmiş varyant listesinin filtrelenmesi işlemi bu uygulama üzerinde gerçekleştirilecektir. Kişiye özel genetik verinin korunaklı bir şekilde saklanması oldukça önemli bir konu olduğundan dolayı verilerin bulutta güvenli bir şekilde saklanması sağlanacaktır. İlgili verilere erişim hakkı olan kullanıcılar dışında kimse herhangi bir veriyi gözlemleme veya yerel bir bilgisayar ağına indirme gibi eylemlerde bulunamayacaktır. Bunun sağlanabilmesi için her kullanıcı hesabı üzerinde farklı kullanıcı seviyeleri oluşturulacak ve sızma testleri ile analiz platformunun güvenliği test edilecektir.



Şekil 3. MinION Veri Analizi Uygulaması için Sistem mimarisi

Geliştirmeyi amaçladığımız sistem, MinION ham verisinin laboratuvardan web uygulamasına yüklenmesi, bu ham verinin otomatik bir şekilde analiz sürecine sokulup son varyant listesinin elde edilmesi ve bu listenin daha sonra tıbbi genetik uzmanları veya patologlar tarafından bir grafik kullanıcı arayüzü aracılığı ile incelenmesini kapsamaktadır. Bu sistemde kullanıcının yaşayabileceği en önemli problemlerden birisi verinin analiz platformuna yüklenmesi olarak karşımıza çıkmaktadır. MinION verisi her ne kadar bir tüm ekzom veya tüm genom verisi kadar büyük olmasa da, 10 GB'lık bir verinin bir sunucuya yüklenmesi yerel internet bağlantısının performansına göre kullanıcı açısından zorluklar yaşatabilmektedir. Bunun için farklı opsiyonlarda çözüm geliştirilecek fakat en önemlisi masaüstü bir yükleme uygulaması tasarlamak olacaktır.

Analiz sürecindeki diğer bir zorluk, hesaplamaların ne kadar sürede tamamlanacağıdır. Yoğun hasta sayısına sahip merkezlerde yavaş süren analiz süreçleri raporlamalarda birikmeye sebep olabilmektedir. Bu da verinin analiz platformuna yüklendikten sonra işlemin olabilecek en kısa sürede tamamlanmasını gerektirmektedir. Bulut tabanlı platform üzerinde her biyoenformatik araç kendi ihtiyacı olan konfigürasyonlardaki bilgisayar birimlerinde çalıştırabilmektedir. Bu da analiz aşamasının daha hızlı, daha verimli ve daha az maliyetle tamamlanabilmesine olanak sağlamaktadır. Verinin analiz platformuna yüklenmesinden sonuçların kullanıcıya kadar erişmesine kadar geçen süreç bir iş akışı şeklinde Şekil 3'te verilmiştir.

Analiz sürecinin kolaylıkla kontrol edilebilmesi için kullanıcı dostu, basit ama etkili bir grafik kullanıcı arayüzü oluşturulması amacıyla tasarımlara başlanacaktır. Kullanıcı ilk girdi datasını FAST5 veya FASTQ olarak web uygulamasına yükledikten sonra ilgili veriyi analiz bitiminde bir varyant listesi olarak analiz platformu üzerinde görüntüleyebilecek ve çeşitli kriterlere göre filtreleyebilecektir.

4 Sonuç

Yeni nesil dizileme sistemleri genomik alanda yapılan çalışmalara hız katmış ve nükleotit başına dizileme maliyetini önemli ölçüde düşürmüştür. Artık çok daha kısa sürede çok daha fazla genom bölgesi dizilenebilmekte ve çeşitli hastalıklara çok daha hızlı bir şekilde tanı koyulabilmektedir. Bu sistemler her ne kadar büyük bir avantaj sağlamış olsalar da eksik yönleri hala mevcuttur ve bu eksiklikler sürekli yeni yaklaşımlarla silinmeye çalışılmaktadır. Çalışmamız kapsamında da bu şekilde yeni yaklaşımlarla üretilmiş bir sistem kullanılarak veri analizi süreci optimize ve otomatize hale getirilecektir. Oxford Nanopore platformlarının dünyada kullanım oranı artmaya başlamış olmasına rağmen bu cihazlardan çıkan verinin analizi için henüz bir standart oluşmamıştır. Bizim oluşturmayı hedeflediğimiz sistem, hem ülkemizde hem de dünyada Oxford Nanopore verisini bulut üzerinde FAST5 formatından klinik rapora kadar götürmesiyle bir ilk olacak ve ülkemizde de tek molekül dizilemesi verisi üzerine bir uzmanlık alanı yaratacaktır.

Analiz sürecinde, MinION dizileme platformundan elde edilen verinin çeşitli hesaplama aşamalarından geçerken farklı formatlara dönüştürülmesi gerekmektedir. Bunun için kullanılan farklı algoritmalarla tasarlanmış biyoenformatik araçları GPU veya CPU temelli çalışabildiklerinden dolayı farklı konfigürasyonlardaki bilgisayarlarda farklı performanslarla sonuç verebilmektedirler. Tüm bunlar analiz sürecini yerel bir bilgisayar üzerinde yavaşlatabilen etkenlerdir. Çalışmamızda, bu sebeplerden dolayı analiz sürecinin bulut tabanlı bir platform üzerinde hızlı ve güvenli bir şekilde yürütülmesi hedeflenmektedir. Önerdiğimiz çözümün bulut tabanlı bir sistem olmasının diğer önemli sebebi de yapılacak analizlerin ölçeklendirilebilir olmasıdır. Ayrıca, her analiz ve her analiz sürecine dahil olan her bir biyoenformatik aracın buluttaki farklı makineler üzerinde (kendi ihtiyacı olan konfigürasyonlardaki bilgisayar birimlerinde) çalıştırılabilecek olması analiz sürecinin ölçeklendirilmesini düşük maliyetle sağlayabilecektir.

Kaynakça

1. Lander E. S., et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921 (2001).
2. Venter J. C., et al. The sequence of the human genome. Science 291:1304–1351 (2001).
3. Sanger F., Coulson A.R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975;94:441–8 (1975).
4. Sanger F., Nicklen S., Coulson A.R. DNA sequencing with chainterminating inhibitors. Proc Natl Acad Sci USA. 74:5463–7 (1977).
5. National Human Genome Research Institute. The Human Genome Project completion: Frequently Asked Questions. https://www.genome.gov/11006943, last accessed: 2018/06/13.
6. van Dijk E.L., Auger H., Jaszczyszyn Y., Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 30:418–426 (2014).
7. Robinson P. N., Piro R. M., Jager M. Computational Exome and Genome Analysis. Chapman & Hall/CRC, Oxfordshire, UK (2018).
8. Illumina. Scalability for sequencing like never before: NovaSeq System Specifications. https://www.illumina.com/systems/sequencing-platforms/novaseq/specifications.html, last accessed: 2018/06/13.
9. Minervini C.F., et al. TP53 gene mutation analysis in chronic lymphocytic leukemia by nanopore MinION sequencing. Diagn Pathol. 11:96 (2016).
10. Lu H., Giordano F., Ning Z. Oxford nanopore MinION sequencing and genome assembly. Genom. Proteom. Bioinform. 14:265–279 (2016).
11. Jain M., Olsen H.E., Paten B., Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 17(1):239 (2016).
12. Richard M. L. and Matthew D. C. A world of opportunities with nanopore sequencing. Journal of Experimental Botany, Vol. 68, No. 20 pp. 5419–5429, (2017).
13. HDF5 File Format Specification Version 3.0. https://support.hdfgroup.org/HDF5/doc/H5.format.html, last accessed: 2018/06/13
14. Cock P.J., et al. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. Apr;38(6):1767-71 (2010).
15. Li H. and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60 (2009).
16. The SAM/BAM Format Specification Working Group. Sequence Alignment/Map Format Specification. https://github.com/samtools/hts-specs/blob/master/SAMv1.pdf, last revised: 2018/05/22
17. Li H., et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25(16):2078-9 (2009).
18. Broad Institute. “Picard Tools.” Broad Institute, GitHub repository. http://broadinstitute.github.io/picard/, last accessed: 2018/06/13; version 2.18.7.
19. Helga T., James T. R., Jill P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 14, 178-192 (2013).
20. The Variant Call Format (VCF) Version 4.2 Specification. https://github.com/samtools/hts-specs/blob/master/VCFv4.2.pdf, last revised: 2017/09/25.