BULUT BİLİŞİMDE GENOMİK VERİLERİN GİZLİLİĞİ

Işıl KARABEY AKSAKALLI[1], Ahmet Fırat YELKUVAN[2], Fuat AKAL[3]
[1]Erzurum Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümü, Erzurum, Türkiye,
[2]Cumhuriyet Üniversitesi, Bilgisayar Mühendisliği Bölümü, Sivas, Türkiye
[3]Hacettepe Üniversitesi, Bilgisayar Mühendisliği Bölümü, Ankara, Türkiye
isil.karabey@erzurum.edu.tr, aftyelkuvan@cumhuriyet.edu.tr, akal@cs.hacettepe.edu.tr

ÖZET

Genomik araştırmalardan elde edilen, tıbbi veritabanları ve biyobankalar gibi elektronik ortamda saklanan veriler, araştırmacılar tarafından sorgulanmakta ve işlenmektedir. Son yıllarda genomik verilere ortak bir erişim alanı sunmak, erişimi kolaylaştırmak ve depolama birimlerini etkin bir biçimde kullanmak için bulut bilişimden yararlanılmaktadır. Genomik verilerin bulut üzerinde işlenmesi için dış kaynaklardan temin edilmesi bazı gizlilik ve güvenlik tehditlerini de beraberinde getirmektedir. Bulut bilişim sistemlerinin kullanımı bireylerin mahremiyetini tehlikeye atmadan etkin bir şekilde sağlanmalıdır. Dışa aktarılan veriler, iletim sırasında taşıma katmanı güvenliği sayesinde kolaylıkla korunurken, verilerin işlenmesi sırasında bir takım problemler meydana gelmektedir. Verinin bulutta açık bir şekilde depolanması veri sızıntısına yol açabilmektedir. Özel Bilgi Erişim (Private Information Retrieval-PIR) şemaları kullanılarak verinin şifreli bir şekilde saklanması da sorgulama sırasında sistemin yavaşlamasına sebep olmakta ve böylece PIR şemaları, gerçek dünya kullanımında etkinliğini yitirmektedir. Bu çalışmada genomik verilerin bulut ortamlarında saklanması ve sorgulanması sırasında karşılaşılacak gizlilik ve güvenlik problemleri ve bu problemlere yönelik literatürde bulunan çözümler derlenmiştir. Genomik verilere yapılan saldırılar ve bu saldırılara yönelik çözüm önerileri de bu çalışma kapsamında bir araya getirilmiştir. Aynı zamanda önerilen çözümlere rağmen hala devam eden bazı açık sorunlar da okuyucuya aktarılmıştır.

***Anahtar Kelimeler*—**Genomik verilerin gizliliği, genomik verilerin işlenmesinde güvenlik tehditleri, Genomik verileri koruma yöntemleri, Özel bilgi erişim şemaları ile şifreli veri sorgulama

Privacy of Genomic Data in Cloud Computing

ABSTRACT

Data obtained from genomic research is stored in electronic form, on medical databases and bio-data-storage, where they are made available for queries and processing by many researchers. Recently, cloud computing is becoming a popular choice given its ability to offer and facilitate shared access to genomic data, and to make effective use of storage capabilities. Outsourcing the processing of genomic data over the cloud, especially that of patients' sensitive data, entails certain confidentiality and security concerns. Various computing services must be used without endangering the individual's privacy. Although the exported data is easily protected by the transport layer security during transmission, a number of problems can still arise during the processing of the data. Storing unencrypted data in the cloud can lead to data leakage, while theoretically storing the data in encrypted form using the Private Information Retrieval (PIR) schemes causes the system to slow down during queries, limiting real-world effectiveness of PIR schemes. The present study aims to provide a more comprehensible compilation of the confidentiality of genomic data, by categorizing the solutions proposed in the literature, for confidentiality and security problems concerning the storage and query of genomic data in cloud environments. Unlike other studies, the attacks on genomic data and the proposal for this aggressive solution have been put together in this study. Furthermore, readers are informed about some open issues despite the proposed solutions.

***Keywords*—**Privacy of genomic data, Security threats in genomic data processing, Protection methods of genomic data, Encrypted data searching using Private Information Retrieval (PIR)

I. GİRİŞ (INTRODUCTION)

"Kullandığın kadar öde" prensibini uygulayarak kullanıcılarına hesaplama ve depolama kaynakları sağlayan bulut bilişim, günümüzde en popüler hesaplama modellerinden biridir. Bulut teknolojisi sayesinde kullanıcılar uygun fiyatlı, verimli hesaplama ve depolama hizmetlerine ulaşabilmektedir. Bu avantajlarının yanında bulut bilişim denildiğinde kullanıcıların aklını karıştıracak konular da gündeme gelmektedir. Bunlar arasında uzak depolama biriminin güvenilirliği, verilerin erişilebilirliği, bulut sistemi yöneticilerinin depolanmış verilere erişim hakkı ve verilerin bilgisayar korsanlarına karşı ne kadar korumalı olduğu sayılabilir. Çünkü hassas ve kişisel verilerin bulut ortamına taşınması, veri sızıntısı riskini ortaya çıkarmaktadır. Bu riske karşılık son yıllarda internet üzerinden uzak sunuculara sıklıkla yapılan aktif saldırılardan dolayı kullanıcıların veri gizliliği farkındalığı giderek artmaktadır. Veri gizliliğini sağlamak için şifreleme tekniklerini kullanmak, bir kişinin güvensiz bir ortamda dahi verilerinin korumasını sağlayabilmekte ve riskin en aza indirilmesini olası kılmaktadır [2]. Bu yüzden bulut ortamını kullanan birçok kullanıcı, veri şifreleme yolu ile bulutta sakladığı verileri için ön etkin (proactive) koruma yöntemini seçmeye başlamıştır [3]. Ön etkin koruma yaklaşımı sayesinde yeni bir tehdit ile karşı karşıya kalmadan zararlı bir eylemin gerçekleşme olasılığına bakılarak kullanıcı bilgilendirilir ve buna karşı önlemler alınabilir. Veriye yetkisiz erişebilen saldırganlara, bulut hesaplamalarını doğru yaparken sorgulanan içerikle ilgilenen dürüst-ama-meraklı servis sağlayıcılarına ve kötücül bulut sunucularına karşı veriyi korumak için bulut üzerinden yapılan hesaplamaların gizli olması gerekmektedir. Dolayısıyla, istenilen işlem bulut üzerinde şifreli veriler üzerinden gerçekleştirilmelidir. Fakat şifreli veri sorgulamanın en önemli problemlerinden biri hesaplamaların çoğunun doğrudan şifreli veriye uygulanamaması ve mevcut çözümlerin de yüksek işlemci gücü gerektirmesidir. Dolayısıyla, kullanıcı verilerinin gizliliği korunurken şifreli veri üzerinden etkili veri kullanımını başarmak, bulut bilişimin önemli araştırma konularından biri haline gelmiştir.

Eğlence, tıp, askeriye, özel sektör, kişisel depolama ve finans gibi birçok alanda kullanılan bulut bilişim, paralel işleme araçları sayesinde büyük miktarda veri ile mücadele etmede yaygın bir şekilde benimsenmiştir. Özellikle yaşam bilimlerinde sınırlı hesaplama kaynakları üzerinde büyük ölçekte genomik veri ile başa çıkmak kaçınılmaz bir zorluk haline gelmiştir. Genom verileri sayesinde bir organizmanın tüm genom dizilimi ile bireyin soy ağacını çıkarabilme ve tüm kalıtım bilgilerini çözebilme imkânı sağlanmaktadır. Modern sağlık hizmetlerini iyileştirmek ve insan genomunu daha iyi anlamak, özellikle hastalıkların teşhisinde ve tedavilere yanıt vermede genom dizilemeye ihtiyaç olduğu için ucuz ve çok sayıda genom verisi üretme gereği kaçınılmaz olmuştur.

Gelecek nesil dizileme (Next Generation Sequencing - NGS) teknolojilerinin ortaya çıkmasıyla birlikte genom dizilimi maliyetinde azalmalar meydana gelmektedir. Bu yüzden genomik verilerden sağlık hizmetleri (örn. kişiselleştirilmiş ilaçlar), biyomedikal araştırmalar (örn. yeni genotip-fenotip ilişkilerinin keşfi), doğrudan tüketiciye hizmetler (örn. hastalık riski testleri) ve adli tıp (örn. ceza soruşturmaları) gibi birçok alanda artan bir şekilde yararlanılmaktadır [4]. Dolayısıyla, üretilen genomik verinin miktarı ve hesaplama gereksinimleri ancak bulut sistemlerinin kullanılmasıyla başa çıkılabilecek hale gelmeye başlamıştır [5] [6]. Öte yandan bulut bilişim, depolama ve hesaplama gereksinimlerini karşılamakla birlikte, doğası gereği kişisel olan genomik verilerin saklanması, işlenmesi, transfer edilmesi, paylaşılması ve kullanılması gibi durumlar söz konusu olduğunda önemli güvenlik endişelerine neden olmaktadır [7]. Verilerin üçüncü kişilere açıklanmasını yasaklayan gizlilik ve veri koruma gereksinimleri bulut bilişimin genomik veri bağlamında kullanılmasının önünde engeller oluşturmaktadır.

Genomik verilerin neden özel veriler olduğu ve gizli tutulması gerektiği DNA özelliklerinden kaynaklanmaktadır [4]. DNA, her şeyden önce kişiye özgüdür ve iki bireyin DNA'sı birbirinden kolayca ayırt edilebilmektedir. Ayrıca DNA, bireyin soyağacı, sağlığı ve davranışları hakkında bilgi içermektedir. Dahası, DNA bir bireyde zamanla çok fazla değişmemekte ve insanlar tarafından gizemli olarak algılanmaktadır. Sonuç olarak, DNA verilerinin şifreli ve anonim bir şekilde

depolanması, işlenmesi, paylaşılması ve yönetilmesi gerekmektedir. Bu gereklilik hem toplumsal kaygılara hem de kamu politikasına dayanmaktadır. Verileri araştırma projelerinde kullanılan birçok kişi kendilerine ait genetik kayıtlardan elde edilen mahrem bilgilerin istismar edilmesinden çekinmektedir [8]. Bu çekinceleri hafifletmek için kişisel olarak tanımlanabilir bir biçimde bir deneğin genomik bilgisinin paylaşımını sınırlayan politikalar hazırlanmıştır. ABD'de bulunan Ulusal Sağlık Enstitüsü (National Institutes of Health-NIH)'nün genomik veri paylaşım politikası buna örnektir. Ülkemizde de 2016 yılında yürürlüğe giren 6698 sayılı Kişisel Verilerin Korunması Kanunuyla DNA gibi hassas verilerin güvenliğiyle ilgili yasal çerçeve çizilmiştir [8].

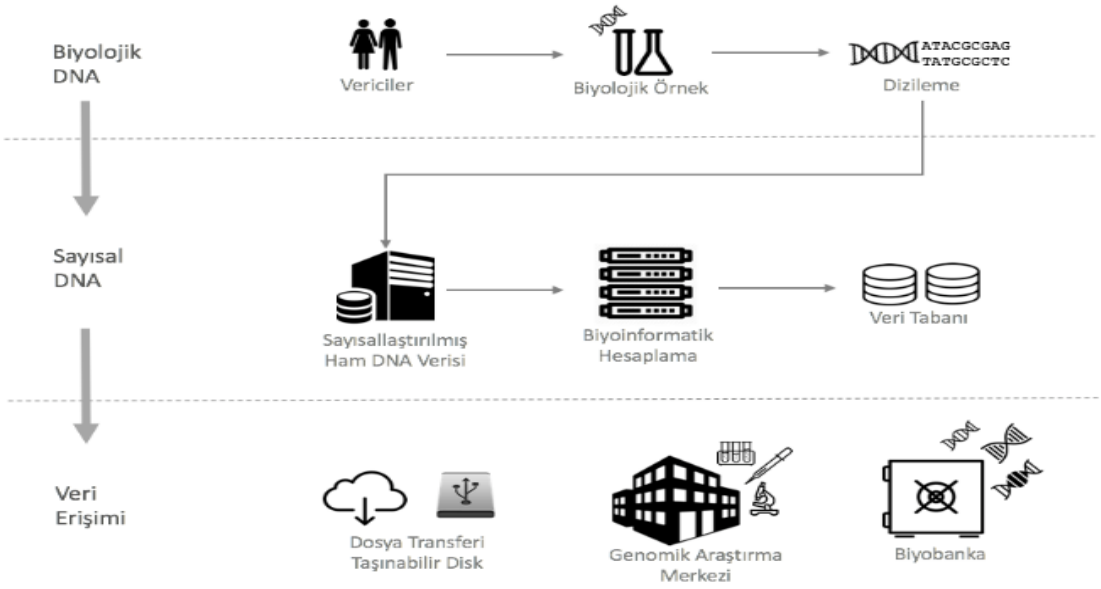
Kişiye özgü genomik verilerin güvenliği sağlanırken, aynı zamanda sağlık hizmetlerinin ve araştırmacıların gizliliği ihlal etmeden bu verilerden yararlanabilmesi de gerekmektedir. Güvenlik alanında yaygın olarak kullanılan Alice ve Bob karakterleri [9] yardımıyla örnek bir senaryo şu şekilde tanımlanabilir. Alice'in Mikrogen Genetik Tıp merkezinde, Bob'un ise Intergen Genetik Hastalıklar Tanı merkezinde bulunan başlıca araştırmacılar olduğu varsayılsın [8]. Hem Alice hem de Bob hasta verisi toplamak ve Alzheimer hastalığı üzerine genom haritalama ile ilgili araştırmalar yapmak için NIH tarafından finanse edilsinler. Alice ve Bob, NIH politikasına [10] uymak için çalışmalarının bitiminde verilerini merkezi bir depolama birimi ile paylaşmak zorundadır. Böylece başka bir enstitüde görev alan Charlie gibi üçüncü kişiler bu veriler üzerinde "Jüvenil Alzheimer ve X gen varyantı tanısı içeren kaç kayıt var?" gibi sorguları işletebilirler [8]. Dolayısıyla, Alice ve Bob'un deneklerinin kimliklerini ortaya çıkarmadan, aynı zamanda da Charlie'nin araştırmalarına engel olmadan genomik verilerin nasıl paylaşılacağı problemi ele alınmalıdır. Yani hem verinin bilimsel araştırmalar için kullanılabilirliği hem de veri gizliliğinin etkin bir şekilde sağlanması gerekmektedir.

Bu makalenin diğer bölümleri şu şekilde düzenlenmiştir. Bölüm II'de genomik veri kavramı detaylı bir şekilde anlatılmaktadır. Genomik veriler elde edilirken karşılaşılabilecek gizlilik ihlalleri ve bu ihlallere engel olmak önerilen çözümler Bölüm III' te verilmektedir. Genomik verilerin gizliliğini sağlamaya yönelik ve henüz açık olan sorunlar Bölüm IV' te yer almaktadır. Makale, sonuçların özetlendiği Bölüm V ile sonlanmaktadır.

II. GENOMİK VERİ (GENOMIC DATA)

İnsan genomu iki tamamlayıcı polimer zincirinden oluşan çift sarmallı DNA moleküllerinde kodlanmaktadır. DNA'yı oluşturan temel yapısal bileşenler nükleotit olarak adlandırılır. Bunlar, Adenine (A), Thymine (T), Guanine (G) ve Cytosine (C)'dir. Dolayısıyla insan genomu, A, T, G ve C harflerinden oluşan uzun bir karakter dizisiyle (ya da dizilimiyle) ifade edilebilmektedir. Genetik bilgi, kromozomların içerisinde bulunmakta ve her kromozom, insanın karakteristik özelliklerini oluşturan çeşitli işlevlerden sorumlu genleri içermektedir [11]. Bir kişinin DNA'sı saç, cilt, kan ve tükürük gibi çeşitli örneklerden edinilmektedir. Örnek alındıktan sonra genetik materyal, DNA Çıkarma Aracı (DNA Extraction Kit) ile çıkarılmaktadır. Daha sonra bir dizileme platformu kullanılarak ham DNA dizilimi elde edilmektedir. Yalnız, bu işlem genellikle kısa okumalar biçimindedir ve okumaların her biri genomun rastgele bir kısmından bir kaç yüz nükleotid içermektedir. DNA diziliminden elde edilen ham okumalar, dizilen genomun tüm halini elde edebilmek için referans bir genoma hizalanır [12]. Tüm genomdan kısa okumaların referans genoma hizalanma süreci "dizilim hizalama" olarak adlandırılmakta ve bu prosedürün sonucu olarak bireyin DNA dizilimindeki yaklaşık 3.2 milyar harf, ileri analizler için sunucularda eşleştirilmektedir [9]. Bu eşleme prosedürü, yüksek hesaplama maliyetini beraberinde getirmekte ve bu yüzden bulut bilişimin sağladığı paralel hesaplama ile bu verilerin işlenmesi için gereken süre önemli ölçüde azalmaktadır.

Genom dizilimi ve analiz sürecinde örnek bir senaryo Şekil 1'de gösterilen bir örnek üzerinde açıklanmaktadır. DNA diziliminden genomik verinin elde edilmesi ve bu verinin analizi sürecinin iş akışıyla bu sürecin sonunda üretilen veriye nasıl erişilebileceği Şekil 1' de görülmektedir. Süreç, vericinin bir sağlık



Şekil 1. DNA verilerinin sanal ortamda saklanması

kuruluşunda, örneğin bir klinikte kan, tükürük vb. biyolojik örnek vermesiyle başlamaktadır. DNA'nın alınan bu biyolojik örnekten çıkarılmasından sonra dizilenmesi gerçekleştirilir.

Günümüzde dizileme işlemleri daha çok Yeni Nesil Dizileme platformları ile gerçekleştirilmektedir. Vericiden örnek alınmasından dizileme işlemlerine kadar elde edilen DNA kimyasal biçimdedir ve güvenliği rutin fiziksel önlemler ile sağlanabilir. Dizilendikten sonra ise DNA sayısallaşır. Sayısallaşan bu veri biyoinformatik hesaplamaya tabii tutulur. Bu hesaplamalar genellikle yüksek başarımlı bilgisayarlar yardımıyla gerçekleştirilir. Üretilen genomik veriler hem ham veri olarak hem de FASTQ, SAM, BAM ve VCF gibi çeşitli formatlara uygun biçimde veri tabanlarında saklanabilir. Verilere uzak ya da yakından erişim verinin sahibi tarafından belirlenebilmelidir. Veri erişimi açısından çeşitli seçenekler bulunmaktadır. Örneğin, kullanıcılar kendilerine izin verilen verileri taşınabilir disklerle edinebilir ya da İnternet üzerinden indirebilir. Ya da üretilen bu veriler, bir genomik araştırma merkezine verilip ilgili konuda yapılacak araştırmalara destek sağlanabilir. Ya da üretilen veriler, kurumsallaşmış biyo-bankalar aracılığıyla araştırmacıların hizmetine sunulabilir. Bu şekilde birçok seçenek veriye uzaktan erişimi gerektirmektedir. Dolayısıyla bulut bilişim depoları üzerinden sayısallaşmış DNA verisinin gizliliğini korumak ve güvenliğini sağlamak için fiziksel önlemler artık yeterli olmayacaktır. Bunun için geliştirilen erişim kontrol yöntemleri, yalnızca yetkili kullanıcıların hassas verilere erişmesine izin vermektedir [13].

Sayısallaştırılarak kayıt altına alınan ve daha sonra kullanılmak üzere paylaşılan genomik verilerde bireye özgü mahrem bilgiler yer almaktadır [12]. Genomik veriler kullanılarak örneğin, bireyin Alzheimer, kanser ve şizofreni gibi spesifik hastalıklara yatkınlığı çıkarılıp gelecekte karşılaşabileceği hastalıklar hakkında tahmin yapılabilmekte ya da bireyin soy ağacı hakkında bilgi edinilmektedir. Dolayısıyla insanlar, genomik verilerinin gizli kalmasını istemektedir. DNA dizisinin çoğu tüm insan popülasyonunda aynı dizilime sahip olsa da her kişinin DNA'sının yaklaşık olarak sadece %0.5'i ki bu sayı milyonlarca nükleotit anlamına gelir, genetik varyasyonlardan dolayı referans genomdan farklıdır [12]. İnsan popülasyonunda en yaygın DNA varyasyonu, tek nükleotitteki farkı temsil eden Tek Nükleotid Polimorfizm (Single Nükleotid Polimorphism - SNP)'dir. Mevcut durumda insan popülasyonunda yaklaşık 54 milyon doğrulanmış SNP bulunmaktadır ve bu sayı

giderek artmaktadır [14]. Yapılan araştırmalara göre bir bireyin çeşitli hastalıklara duyarlılığı, bireyin SNP' leri incelenerek hesaplanabilmektedir. Örneğin Apolipoprotein E (ApoE) geni üzerindeki iki belirli SNP' nin (rs7412 ve rs429358) Alzheimer hastalığı için artan bir risk belirttiği bildirilmiştir. Bir başka örnekte Alzheimer hastalığına yatkınlığı analiz etmek için toplam on belirli SNP taşıyan üç adet genin bulunduğu bildirilmiştir [12]. Buradan hareketle genomik veri analizinin yaratabileceği hassasiyete dair hipotetik bir örnek şöyle verilebilir. Alice genomunu MyGenome adlı bir merkezde diziletsin ve verilerinin bu merkezin deposunda tutulmasına izin versin [4]. Alice soy ağacının bir kısmını keşfetmek, kendisi ile ilgili bazı gerçekleri öğrenmek ve aynı zamanda genetik araştırmalara katkıda bulunmak için MyGenome'un hizmetlerinden yararlanmış olsun. MyGenome'un dizilemeyi gerçekleştirmesinden yıllar sonra Alice bazı sağlık problemleri yaşasın. Özelleştirilmiş tedaviler uygulamak için genomik verileri kullanan bir doktoru ziyareti sonucunda Alice, geçmişte MyGenome'a sunduğu genomik verilerin kullanılarak şimdiki hastalığının teşhisinin ve tedavisinin yapılabileceğini öğrensin. Doğal olarak Alice gelecekte karşılaşabileceği olası sağlık sorunlarını da bilmek isteyecektir. Yapacağı araştırmalar sonunda onun genomik profilinde olan insanlar için örneğin bunama riski olduğunu öğrenmiş olabilir. Bu olaydan sonra Alice, verilerini paylaştığı MyGenome hakkında şüphe duymaya başlar. Bilgilerinin kendisinden habersiz kullanılabileceği ya da başkalarıyla paylaşılabileceği ve bu durumun da kendisi hakkında beklenmedik sonuçlara yol açabileceği endişesini yaşamaya başlar. Bu hipotetik örnekte SNP'lerin bireylerin sağlığı hakkında gizlilik hassasiyeti bilgisi taşıdığı ve SNP sızıntısının kişisel verilerin gizliliğini büyük bir tehlikeye attığı görülmektedir.

III. BULUT ÜZERİNDEN GENOMİK VERİLERE ERİŞİMDE KARŞILAŞILAN GİZLİLİK PROBLEMLERİ (PRIVACY PROBLEMS ENCOUNTERED IN ACCESSING GENOMIC DATA ON CLOUD)

Genomik verilerin paylaşımı biyomedikal keşiflerin hızını artırmak için yaşamsal önem taşımaktadır. Öte yandan, sayısallaştırılmış genomik veriler, kişiler hakkında gizli bilgilerin açığa çıkma riskini de beraberinde getirmektedir. Bu yüzden son yıllarda, genomik verilere istenmeyen erişimle ilgili gizlilik tehditleri ve olası çözümler tartışılmakta ve genomik verilerin gizliliği üzerine birçok öneri sunulmaktadır. Ancak güvenlik tanımlarının karmaşık doğası nedeniyle gerçek yaşam problemleri için önerilen çözümleri uygulamak zor olmaktadır.

Bu bölümde insan genomik verisine yönelik tehditlere karşı literatürde önerilen çözümlere ve gizlilik önlemlerine rağmen devam eden ve bulut bilişimi de ilgilendiren problemlere yer verilmiştir.

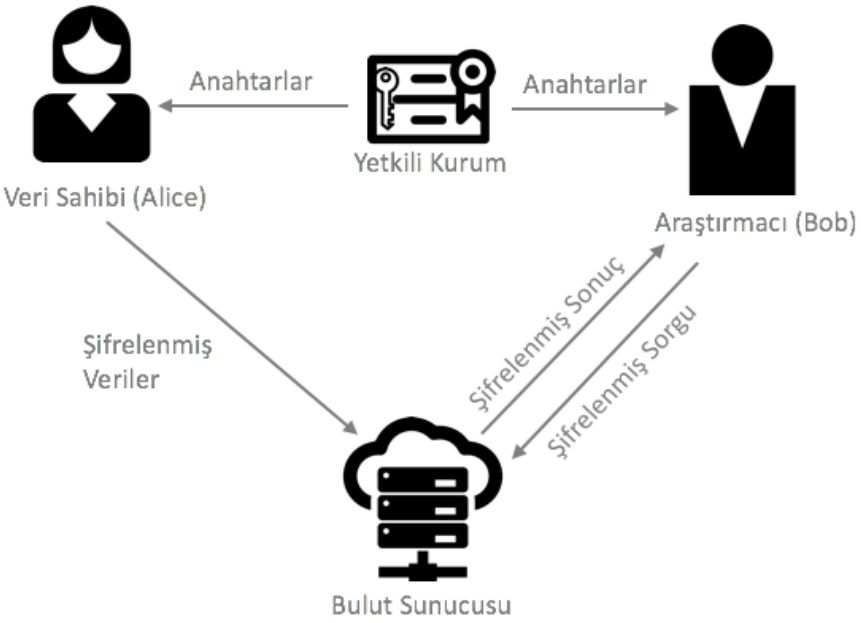
3.1. DNA dizisi hizalama ve karşılaştırma

DNA dizileri arasındaki benzerlikleri keşfetmek için yapılan dizi hizalama işlemi, yüksek ve pahalı hesaplama gereksinimlerinden dolayı çoğu zaman genel bulutlarda (public clouds) yapılmaktadır. Okuma haritalama (read mapping) işleminin buluta bırakılması hesaplama maliyetinin azalmasını sağlamaktadır [9]. Bununla beraber, genomik verilerin üçüncü kişiler tarafından kontrol edilen genel bulutlara gönderilmesi ve herkese açık bir ortamda verinin işlenmesi, gizli genom verisinin açığa çıkma riskini doğurmaktadır. Bu gizlilik ihlalini önlemek için güvenli dizi hizalama mekanizmaları tanımlanmalıdır. Örnek bir senaryo olarak Alice'in Bob tarafından kontrol edilen bir genel bulut üzerinde tüm genomu için dizi hizalama yapmak istediği varsayılsın [11]. Bu senaryoda Alice ne bulutun diğer kullanıcılarına ne de Bob'a güvenmektedir. Verilerini gönderdiği bulut altyapısında ham ya da işlenmiş herhangi bir verisinin başkaları tarafından elde edilme ihtimalini düşünmektedir. Bu yüzden hizalama sürecinde okumalar ile ilgili tüm değerli bilgiyi

korumak gerekmektedir. Bir başka örnek senaryoda Alice ve Bob karakterleri iki farklı biyo-mühendislik şirketini temsil etsin. Alice'in, belli bir DNA dizisine benzer bir dizinin, Bob'un elinde olup olmadığını kontrol etmek istediği varsayılsın [15]. Böyle bir durumda Alice'in neyi aradığı konusunda Bob'un fikrinin olmaması ve Bob'un da bu işlem yapılırken kendine ait dizi kataloglarını gereksiz yere ifşa etmemesi beklenmektedir. Çünkü bu tür karşılaştırmalar sırasında kendi dizileri hakkında herhangi bir bilgi ifşa etmemeleri veri sahipleri için büyük önem taşımaktadır.

Yapılan literatür araştırmaları sonucunda güvensiz ortamda DNA dizisi hizalama ve karşılaştırma problemi ile ilgili bir çok çözüm yolu önerilmiştir [15-17]. Chen vd. [16] tarafından yapılan bir çalışmada bir şirket ile birlikte bünyesinde özel ve genel bulutu birleştiren bir hibrit bulut üzerinde ölçeklenebilir ve güvenli bir okuma haritalama işlemi yapılmıştır. "Tohumla ve genişlet (seed-and-extend)" yönteminden esinlenen yaklaşımda haritalama görevi iki farklı buluta paylaştırılmaktadır. Bu yaklaşımda Genel Bulut, kısa okuma alt dizileri ile (tohumlar) uyumlu anahtar (keyed hash) değerlerinden kesin eşleşenleri bulmada ve gende kabaca konumlandırmak için referans dizilerinin uyumlu anahtar değerlerini okumada görev almaktadır. Özel bulut ise doğru hizalamayı bulmak için bu konumlardaki tohumları genişletmeden sorumludur. Tohumla-ve-genişlet stratejisi olarak adlandırılan yöntemde temelde iki dizinin düzenleme mesafesine bakılıp bazı alt diziler arasındaki eşleşmeler değerlendirilir. Yazarların yaklaşımında açık buluttaki uyumlu anahtar değerleri dışında hiç bir bilgi açıklanmadığı için saldırganlar doğrudan okumalara erişememektedir. Tohumla ve genişlet yaklaşımını kullanan bir diğer çalışmada, Zhao vd. [18] az sayıda insan genom dizisinin bile genomik verilerin sahibi hakkında bilgi içerdiğinden bahsederek mevcut güvenlik ve şifreleme yöntemlerinin yetersiz olduğunu savunmuşlardır. Önerdikleri algoritmada "tohumla ve genişlet" yaklaşımını kullanarak okuma eşleme işlemini güvenli bir şekilde gerçekleştirmek için hesaplamaların çoğunu genel buluta devrederken, yalnızca özel bulutta şifreleme ve şifre çözme işlemlerini gerçekleştirmektedir. Genel bulut, okumaların varsayılan konumlarını belirlemek için tohumların anahtar (hash) değerleri ile referans genomundaki ilgili alt dizeler arasındaki tam eşleşmeleri ararken, özel bulut bu tohum eşleşmelerini optimum hizalamalara genişletir ve okumalar ve referans genom arasındaki farklılıkları tespit etmektedir.

Ileri vd. [17], DNA hizalama probleminde blok zinciri (blockchain) mantığının kullanılması fikrinden yola çıkarak okuma haritalama (read mapping) iş yükünü referans genomlara dağıtmak için Coin-Application Mediator Interface (Coinami) isimli bir arayüz önermiştir. Coinami ağı üç seviyeli bir yapıdan oluşup tek olan kök otorite, alt otoritelere sertifika dağıtmaktadır. Alt otoriteler çoklu madencilere hizalama problemleri gibi atamaları yapmakta ve aldığı sonuçları doğrulamaktadır. Eğer hizalamalar geçerli ise alt otoriteler blok zincirleri imzalayıp madencilere geri göndermektedir. Bir madenci, birden fazla alt otoriteden gelen atamalar üzerinde çalışabilmektedir. Coinami iş akışında otorite, N adet genom örneğini FASTQ dosya formatında ve bilinen haritalama konumları ile birlikte tuzak (decoy) okumaların bulunduğu bir veritabanında barındırmaktadır. Önerilen arayüzde tuzak veritabanı, her referans genomu ve okuma haritalama kombinasyonu için bir kez üretilmelidir. Atamaları oluşturmak için otorite, çoklu örneklerden okumaları karıştırmakta ve çoklamakta, ayrıca atamanın %5'ini oluşturan tuzakları içermektedir. Daha sonra madenciler atamaları indirmekte; okumaları, otorite ve madenciler arasındaki değişim protokolü ile belirtilen referans genoma eşleştirmekte ve BAM dosyalarını otoriteye geri göndermektedir. Son olarak otorite, hizalamaları N adet örneğe yeniden atamakta ve sonuçları doğrulamak için tuzakları kontrol etmektedir. Eğer tuzaklar onların önceden belirlenmiş konumlarına hizalanırsa, BAM dosyasının geçerli olduğu düşünülmekte ve otorite bloğu imzalayıp ilgili madenciye geri göndermektedir.



Şekil 2. Güvenli genom verisi sorgulama

Önerilen bu sistemdeki tuzak okuma adları, veritabanı aramasından korunmak için madenci tarafından raporlanan BAM dosyaları ile doğrudan kontrol edilebilen haritalama bilgilerini içermektedir. İsimleri okumak için bir ön ek olarak eklenen iş ID'si sayesinde farklı atamalarda aynı tuzak kullanılmış olsa bile farklı şifreli okuma adı almakta ve bu da potansiyel saldırıları önlemektedir.

Atallah vd. [15] yaptıkları bir çalışmada dizi karşılaştırma için dinamik programlama üzerinden düzenleme uzaklığını kullanmışlardır. Yazarların önerdiği protokolde homomorfik şifreleme kullanılarak anlamsal olarak güvenli açık anahtar sistemi kullanılmaktadır. Karşılaştırılacak girdi değerleri, girdi sahibinin (Alice) açık anahtarı ile şifrelenip karşı tarafa gönderilmektedir. Karşı taraf (Bob) pozitif ve negatif olmak üzere rastgele bir vektör üretip bu vektörü şifreli girdiler ile toplamaktadır. Daha sonra rastgele seçilen bir permütasyona göre elde edilen bu toplamların sırası değiştirilip veri Alice'e gönderilmekte ve Alice'in özel anahtarı ile şifre çözülerek toplamlar elde edilmektedir. Son aşamada Bob tarafından, Alice'in dizisi ile karşılaştırılacak dizi ve önceki rastgele üretilen vektör değerlerin farkı alınarak DNA nükleotidleri hakkında herhangi bir bilgi ifşa edilmeden dizi karşılaştırması yapılmaktadır. Bu süreçte rastgele bir vektör üretilip girdi değerlerine eklendiği için iki taraf da verinin içeriğinden habersizdir.

3.2. Genomik verilerin sorgusu

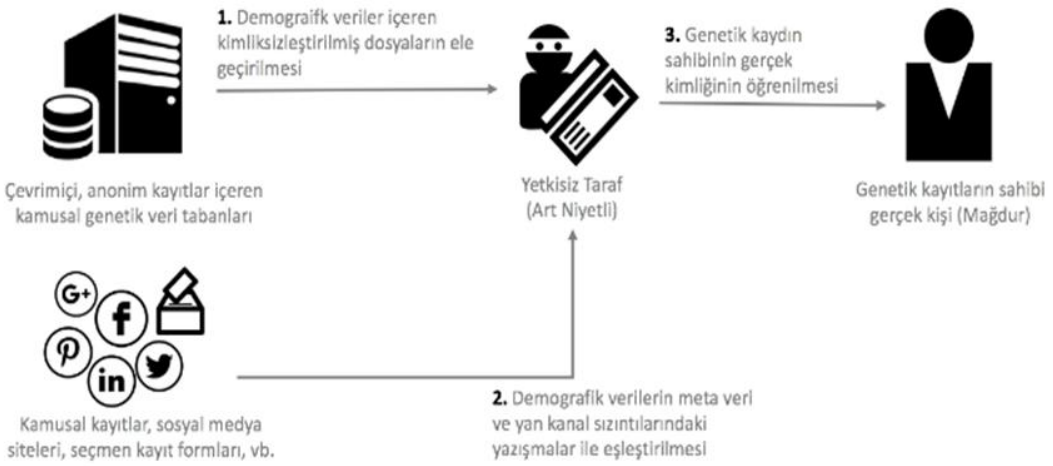
Bulut sisteminde depolanan genomlar ve bu genomlara erişmek için yapılan sorgular, sağlayıcıya ait ticari hakları korumak, saldırganlar tarafından erişimi engellemek için gizli tutulmalıdır. Ayrıca, bulut sistemi de sakladığı verinin içeriğinden ve işlettiği sorguların sonuçlardan habersiz olmalıdır. Örnek bir senaryo olarak hasta olan Alice'in bir genetik test yaptırmak istediği varsayılsın [11]. Bu hizmet Bob tarafından veriliyor olsun. Bob genetik testi gerçekleştirmek için Alice'in genetik verileri üzerinde kendi sorgusunu işletmek zorundadır. Alice genlerinin gizli kalmasını istemekte, Bob'a ve Bob'un alt yapısına güvenmemektedir. Benzer şekilde Bob da sorgusunu gizli tutmak istemekte, Alice'e ve Alice'in genomik veriler içeren hiçbir cihazına (akıllı kart, taşınabilir bellek vs.) güvenmemektedir. Bu senaryoda hem Alice hem de Bob verilerinin gizliliği konusunda endişe duymaktadır. Alice DNA profilinin açığa çıkmasını istemiyor iken Bob da çalıştırdığı sorguların gizli tutulmasını istemektedir. Bu senaryoda genom verisinin

güvenli bir biçimde nasıl sorgulanabileceği Şekil 2'de görülmektedir. Bu senaryoda verinin sahibi yetkili kurumdan edineceği anahtar yardımıyla verilerini şifreleyerek bulut üzerinde depolamaktadır. Araştırmacı bu verilere, elinde ancak geçerli bir anahtar varsa ulaşabilmektedir. Yetkilendirilmiş araştırmacı sorgusunu şifrelenmiş olarak buluta yönlendirir ve sonuçları da şifrelenmiş olarak alır. Bulut sunucusu gelen sorgu ve giden sonuçların içeriği hakkında herhangi bir bilgiye sahip değildir.

Blanton ve Aliasgari [19] tarafından yapılan bir çalışmada bir istemcinin DNA dizisinin ve bir sağlayıcının genetik teste karşılık gelen bir deseninin değerlendirildiği sonlu bir otomata aracılığı ile hataya dayanıklı (error-resilient) DNA sorgulama problemi ele alınmıştır. Hesaplama sunucularına herhangi bir bilgi ifşa etmeden güvenli kaynak aktarımı için herhangi bir sonlu otomata tipine uygulanabilen hataya dayanıklı aramada desen, bir sonlu otomata olarak temsil edilmekte ve DNA dizisi üzerinde değerlendirilmektedir. Çalışmada uygulanan sonlu otomata formülleri ile iletişim karmaşıklığı azaltılmış ve şebekeler gibi dinamik ya da durağan olmayan ortamlarda çalışmaya uygun çözüm üreten çoklu taraf dış kaynak kullanım protokolünün bir eşik değer versiyonu geliştirilmiştir. Uygulanan homomorfik şifreleme yöntemi ile şifreli metin gözlemlenerek açık metin hakkında hiçbir bilginin elde edilemediği garanti edilmiş, eşik şifreleme ile de tüm taraflar ile birlikte önceden tanımlanmış tarafların katılımının deşifreleme için gerekli olduğu bir sistem tasarlanmıştır. Böylece bir saldırganın sadece bir tarafa erişmesi ile veriyi elde edemeyeceği vurgulanmıştır.

Canim vd. [20] tarafından yapılan bir çalışmada büyük miktarlarda klinik ve genomik veriye bağlı araştırmalarda bireylerin gizliliğini sağlamak ve sistem güvenliğini korumak amacıyla şirketlerin kimliksizleştirme yöntemleri üzerinden oluşturdukları güvenliğin yeterli olmadığı, kimliksizleştirilen verilerin bir takım yöntemler kullanılarak bireylerin kimliklerini ortaya çıkarabildiği sonucuna varılmıştır. Bu yüzden kimliksizleştirme yöntemlerine alternatif olarak herhangi bir gizli kayıt ortaya çıkarmadan pratik kriptografik protokoller üzerinden biyomedikal verilerin paylaşılabileceği, yönetileceği ve analiz edilebileceğinden bahsedilmiştir. Yapılan çalışmada kriptografik donanım üzerinden hassas biyomedikal verileri işlemek ve depolamak için çoklu üçüncü taraf ihtiyacını ortadan kaldıran bir çerçeve geliştirilmiştir. Bu çerçeve ile genomik verileri işlemek için güvenli bir protokol tanımlanmış ve tipik medikal araştırmalar için böyle bir yaklaşımın verimli bir şekilde çalıştırılabileceğini göstermek için çeşitli deneyler yapılmıştır. Tasarlanan mimarinin pahalı açık anahtar şifreleme protokollerine nazaran daha hızlı ve daha etkili olduğu deneylerle kanıtlanmış olup çoklu üçüncü taraf yerine tek bir üçüncü taraf ile güven gereksinimi azaltılmıştır.

Perl vd. [21] güvenli sorgulama sırasında şifreli veri üzerinde yapılan işlem adımlarının genellikle yavaşlamaya ve karmaşıklığa yol açtığından dolayı veri korumanın zor olduğunu belirtmişlerdir. Yaptıkları çalışmada homomorfik şifrelemenin gerçek yaşam uygulamalarına uygun olmadığından yola çıkarak bu tarz uygulamalar için gerekli performansı sürdürebilen Bloom Filtre tabanlı yeni ve güvenli bir terim arama algoritması geliştirilmiştir. Çalışmada aynı zamanda Gizlenmiş Bloom Filtreleri (OBF) kullanılarak veri gizliliği sağlanmıştır. Burada Bloom filtreler, ikili arama yolu ile büyük bir veri kümesini, olası eşleşmelerin olduğu küçük bir veri kümesine indirgeme amacıyla bir ikili ağaç oluşturmak için kullanılmaktadır. Bloom filtresi uygulandıktan sonra küme üyeliğinde yanlış pozitifler üretmek için filtreler genişletilmiştir. OBF yöntemi ile açık Bloom Filtre'ye erişme ihtimalinin yeterince az olabildiği ve gerçek sonuçların yanlış pozitif kümesi sayesinde gizlenmiş hale getirilebildiği bir arama algoritması önerilmiştir. Ek olarak, veri kümesi genişletilse de elde edilen sonuçların yine de homomorfik şifreli metinde işlenecek kadar büyük olmadığı da gösterilmiştir. Geliştirilen Bloom Filtre arama algoritmasında kullanıcı kendi açık anahtarı ile OBF'den geçirilmiş şifrelenmiş arama terimini sunucuda aramaktadır. Sunucu, sorguyu işletirken veritabanını daha küçük olası eşleşmelere indirgemek için OBF'yi kullanmakta ve elde



Şekil 3. Üst veri ve yan kanal sızıntıları ile kimlik izi sürme

ettiği sonucu şifreli olarak kullanıcıya iletmektedir. Son olarak da kullanıcı kendi gizli anahtarını kullanarak bu şifreyi çözmektedir. Perl vd. [22] başka bir çalışmalarında Bloom filtrelerin PIR (Private Information Retrieval) eşdeğerlerinden 2000 kat daha hızlı işlem yaptığını deneyler ile kanıtlamıştır. Yapılan performans testlerinde 50-nükleotit uzunluğundaki insan genom dizisinde bir aramanın 2.8 GHz Intel Core i7 işlemcili bir cihaz üzerinde 0.1 saniyeden daha az bir sürede gerçekleştiği gözlenmiştir.

Franz vd. [23] genomik dizilerin mahremiyetini Gizli Markov Modeli (Hidden Markov Model - HMM) kullanarak sağlamışlardır. Yazarlar uyguladıkları modeli bir senaryo üzerinden açıklamışlardır. Senaryoda hastaların genom verilerini depolayan A isimli bir sağlık hizmeti sağlayıcısının hastalık testi yaptığı ve özel bir hastalık için HMM kullanılarak kodlanan bir modele karşı genom bölümlerini test eden başka bir B sağlayıcısı ile iletişime geçmek istediği belirtilmektedir. A hastalık yatkınlığı olasılığını belirlerken modelin genoma nasıl oturtulduğunu görmek istemekte, B ise kendi hizmetini diğer sağlayıcılardan ayırt etmeye yarayan HMM modelinin ifşa olmasını istememektedir. Problemi çözmek için çok sayıda aritmetik işlem olmasına rağmen medikal laboratuvarlarda bulunan hesaplama kaynaklarını kullanarak HMM' lerin uygulanabileceği gelişmiş protokoller sunulmuştur. Bu durumda karşılıklı olarak birbirlerine güvenmeyen iki taraftan biri HMM'yi bilirken diğer taraf genomik metni bilmekte fakat ikisi de algoritmayı çalıştırabilmektedir. Çalışmada taraflar modelin genoma uyup uymadığını öğrenirken, hem modelin parametrelerini hem de genomik diziyi birbirlerine açıklamak zorunda değildirler.

Lei vd. [24] tarafından yapılan bir diğer çalışmada mobil cihazlar üzerinde genom verilerinin gizliliği üzerinde durularak dinamik simetrik aranabilir şifreleme yöntemi (DSSE), genom verisinin özel olarak depolanması ve işlenmesi için bir çerçeve olarak sunulmuştur. Çalışmada genom verisinin bulut sağlayıcı üzerinde depolanması ve hesaplamaların kullanıcının kendi mobil cihazı üzerinde yapılması ile gizlilik ihlallerinden ve pahalı homomorfik şifrelemeden kaçınmak amaçlanmıştır. Önerilen mimaride bir sertifika kurumu, kullanıcı DNA' sının dizilenmesinden ve onun Bulut Depolama Sağlayıcı (Cloud Storage Provider-CSP)'sına güvenilir bir şekilde ulaştırılmasından sorumludur. Aynı şekilde kullanıcı da istediği verileri DSSE yöntemi ile güvenli bir şekilde CSP üzerinden almaktadır. Oluşturulan sistemin, gizlice dinleyen kişiler, güvenilir olmayan CSP ve kötü niyetli taraflar olmak üzere üç saldırı şekline karşı güvenliği test edilmiştir. Gizlice dinleme saldırılarından korunmak için basit kriptografik araçların yeterli olduğu görülmüştür. Güvenilir olmayan CSP'den ve kötü niyetli taraflardan verileri korumak için ise CSP'de saklanan tüm genetik veriler kullanıcıların simetrik gizli anahtarları ile şifrelenmekte ve böylece üçüncü

Şekil 4. DNA aracılığı ile kişi özelliklerini açığa çıkarma [1]

bir taraf ve CSP, sunucuda saklanan genomik verilerin yalın içeriğine erişememektedir.

Zhu vd. [25] benzer hasta arama problemine etkili ve gizliliği koruyan bir çözüm üretmişlerdir. Benzer hasta arama, sistemde bulunan tüm hastanelerde belirli bir hastanın genomik verisine benzer tüm hastaların araştırılmasıdır. Bu arama gerçekleştirilirken her hastanenin kendi veri kümesi tekil bir anahtar ile şifrelenmekte ve veri kümesinin herhangi bir bulut ortamında şifreli tutulması sağlanmaktadır. Yazarlar düşük hafıza gereksinimi ile her hastanenin veri kümesini indekslemek için hiyerarşik bir dizin yapısı önermişlerdir. Ayrıca, arama verimliliğini önemli ölçüde artırmak için her hastanenin bireysel indekslerinden ortak bir arama indeksi oluşturan yeni bir gizlilik koruma mekanizması geliştirmişlerdir. Şifreleme yöntemi olarak ise standart simetrik şifreleme ve özellik tabanlı şifrelemenin bir kombinasyonunu geliştirerek genomik verilere güvenli erişim kontrolü sağlamışlardır.

3.3. Üst veri ve yan kanal sızıntıları ile kimlik izi sürme

İnsan genom gizliliğini tehlikeye atan bu saldırı türünde bilgisayar korsanı ya da meraklı taraf, sahibinin kimliği gizli olan çevrimiçi insan genom verisine ve veri sahibiyle ilgili olabilecek temel demografik ayrıntılar ve sağlık koşulları gibi ek üst verilere ihtiyaç duyar [26]. Şekil 3'te art niyetli bir yetkisiz kişinin çevrimiçi anonim kayıtlara erişerek elde ettiği demografik verileri üst veri ve yan kanal sızıntılarındaki yazışmalar ile eşleştirerek genetik kaydın sahibini nasıl ortaya çıkarabileceği gösterilmektedir. Böyle bir saldırı başarılı olursa genetik ayrımcılık, mali kayıp ve genom sahibine şantaj yapma gibi ciddi sonuçlara sebep olmaktadır.

Kimlik izi sürme yöntemi kullanılarak Sweeney tarafından yapılan çalışmada [27] dönemin Massachusetts valisinin herkese açık olan seçmen kayıt formları ve hastane kayıtlarında görülen demografik verileri (doğum tarihi, cinsiyet, 5 haneli posta kodu vb.) kullanılarak valinin sağlık durumu tespit edilmiştir. Sweeney [28] başka bir çalışmada ise ABD'deki popülasyonun neredeyse %87'sinde doğum tarihi, cinsiyet ve posta kodunun benzersiz olduğu ve bu nedenle bu üç özelliği içeren herhangi bir verinin anonimleştirilmemesi gerektiğinden, aksi halde bu durumun yeniden tanımlama saldırılarına sebebiyet verebileceğinden bahsetmektedir. Savage [29], yeniden tanımlamanın verilere erişme, kimlik bilgilerinin belirlenmesi ve verilerin bilinen tanımlayıcılara bağlanması şeklinde üç adımda gerçekleşebileceğini belirterek veritabanlarına yetkisiz erişen

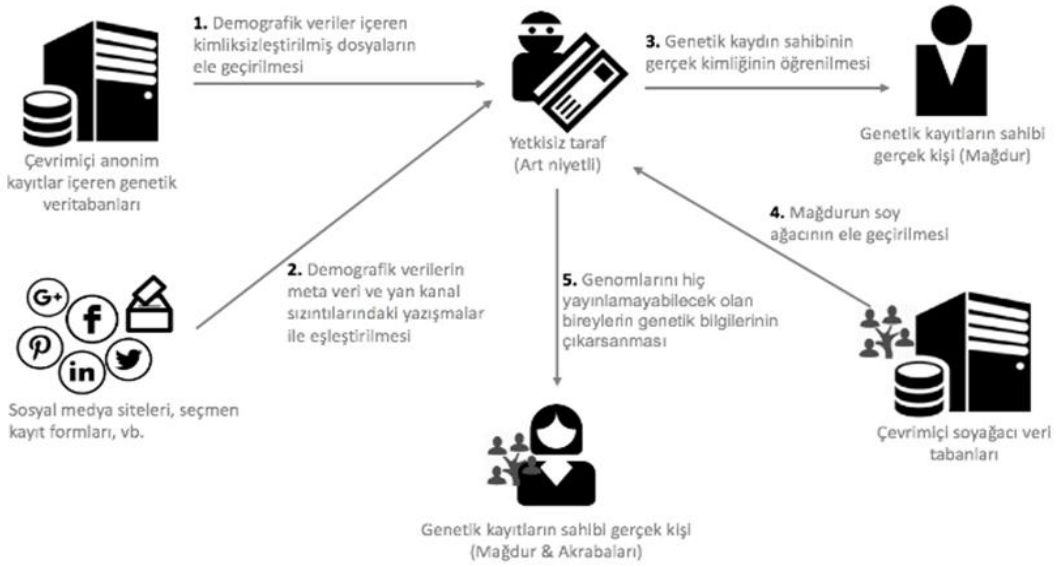
kişilerin bu adımları uygulamak için hem zamana hem de tanımlanabilir bağlantı kayıtlarına erişmek için paraya ihtiyaç duyduklarını vurgulamaktadır. Durum çalışması olarak Personal Genome Project (PGP) ve Sequence Read Archive olmak üzere iki farklı veritabanı üzerinden bireylerin yeniden tanımlanabileceği senaryolardan bahsetmiştir. Kayıtlı profillerin çoğunda yer alan doğum tarihi, posta kodu ve cinsiyet gibi üst veriler aracılığı ile saldırı gerçekleştirilebilmektedir. Ayrıca dosya isminden kişinin isim ve soy ismine ulaşılabilmektedir. Sweeney vd. [26] bir başka çalışmasında PGP web sitelerinden kopyalanan 1130 adet açık profilin yaklaşık %50'sinde açık olarak doğum tarihi, cinsiyet ve posta kodunun görüldüğünü rapor etmiştir. PGP web sitesinden sıkıştırılmış olarak indirilebilen genomik veri içeren dosyalar incelenmiş ve anonim PGP profilleri ile bunlar arasında bu dosyaların adında bulunan üst veri yardımıyla ilişkiler kurulabilmiştir. Bazı katılımcılar için indirilen sıkıştırılmış dosya açıldığı zaman elde edilen dosyanın katılımcının gerçek ismini içeren bir dosya adına sahip olduğu görülmüştür. Örneğin hx0157A_8659862.zip isimli indirilebilir bir DNA dosyasına sahip olan bir hx0157A profili açıldığında genome\Elanie\Smith\629562.txt isimli genom sahibinin adını içeren bir dosya ortaya çıkmaktadır. Ayrıca doğum tarihi ve posta kodu gibi ek veriler de kullanılarak arama sonuçları az sayıda bireye kadar daraltılabilmektedir.

Alser vd. [1] üst verilerin ilgili kişilerin kimliklerini ortaya çıkarmadaki etkisi büyük olduğu için bu verilerin veritabanlarından kaldırılması ya da kesin bir şekilde 2002 Sağlık Sigorta Taşınabilirliği ve Hesap Verebilirlik Yasası (HIPAA) Gizlilik Kuralı'nın takip edilmesi gerektiğini savunmaktadır. HIPAA altında kapsanan verilerin bazı formatlara kesin bir şekilde uyması gerektiğini (örneğin; tarihlerin sadece yıl bilgisi içermesi, posta kodlarının nüfusun 20.000 kişiden az olduğu yerlerde sadece ilk iki haneyle temsil edilmesi ve isim, sosyal güvenlik numarası, sokak adresleri gibi hiçbir tanımlayıcının açık olmayacağı vb.) belirtmişlerdir.

3.4. DNA üzerinden kişi özelliklerinin ifşa edilmesi

ADAD'da (Attribute Disclosure Attack via DNA) saldırgan, kişinin kimliği ile mahrem bilgilerini (örn. ilaç bağımlılığı) ilişkilendirmek için DNA verilerini kullanan istatistiksel bir köprü oluşturmaktadır [30]. Şekil 4'te DNA verileri kullanılarak kişilerin özelliklerinin nasıl çıkarılabileceğinin bir örneği görülmektedir. Örnekte, belirli popülasyonlardaki genotiplerin dağılımı ve istatistiklerine dayanarak yayınlanmış çalışmalara katılan bireylerin benzer özelliklerine bakılarak belli hastalıklara sahip olma olasılıkları hesaplanabilmektedir. Böylece art niyetli kişi belirli bir kişinin yayınlanmamış bilgilerini ortaya çıkarma şansını elde edebilmektedir. Yapılan deneyler sonucunda dikkatli bir şekilde seçilen 45 adet SNP verisinin dünya çapında belli popülasyonların çoğunda en fazla 10-15 hata ile eşleşme sağlamaya yeterli olduğu tespit edilmiştir [31]. Ayrıca yaklaşık 300 yaygın SNP'nin rastgele alt kümelerinin herhangi bir kişiyi benzersiz bir şekilde eşleştirmek için yeterli bilgiyi üretebileceği hesaplanmıştır [31]. Düşük sayıda SNP verisinin eşleşme için yeterli olması ile GWAS (Genome Wide Association Studies - Genom Çapında İlişkilendirme Çalışmaları) çalışmalarındaki birey seviyesinde genotip-fenotip kayıtları ADAD'a karşı savunmasız kalmaktadır. Bu problem ile başa çıkmak adına NIH de dahil olmak üzere birçok kuruluş GWAS veri setleri için iki katmanlı bir erişim sistemini (sınırlı erişim-açık erişim) benimsemiştir. Bunlardan bireysel seviyedeki genotip ve fenotipleri depolamak için sınırlı erişim alanı kullanılırken, Alel frekanslarının üst düzey veri özetleme istatistikleri, açık erişim alanında depolanmaktadır [31].

Erlich ve Narayanan [31] tarafından ADAD'a çözüm olarak önerilen bir yöntem diferansiyel gizliliktir. Diferansiyel gizlilik, özet istatistiklerin gizliliğini korumak için yayınlanmadan önce sonuçlara gürültü ekleyen bir yaklaşım sunar. Bu yöntem bir kişinin kaydına göre değişen iki veri setinin özet istatistiklerinin birbirlerine çok yakın olmasını



Şekil 5. Tamamlama saldırısı

garanti eder ve böylece aradığı kişinin veritabanında olup olmadığını bilmeyen saldırgan, kişinin mahrem özelliklerini öğrenemez.

Özellik çıkarımını azaltmak ve genomik verilerin gizliliğini korumak için Humbert vd. [32] tarafından yapılan bir diğer çalışmada bir şaşırtma mekanizmasına dayalı genomik gizliliği koruma mekanizması (GPPM) önerilmiştir. Pratikte SNP değerlerini gizleyerek, hassaslığı azaltarak, sahte SNP değerleri ekleyerek ya da SNP değerlerine gürültü ekleyerek şaşırtma işlemi uygulanabilmektedir. Bu çalışmada SNP gizleme yöntemi seçilmiştir. Sahte veri ya da gürültü ekleme işlemi, yüksek maliyetinden ötürü genomik araştırma topluluğunca tercih edilmemektedir.

3.5. Tamamlama saldırıları

Genotip isnat (Genotype imputation) olarak adlandırılan bu yöntem ile kısmi verilerden yola çıkarak genetik bilginin tamamlanması amaçlanır [31]. Bu saldırı yöntemi, ilgili verideki eksik genotip değerlerini tamamlamak için kullanıcı referans panelleri ve işaretçiler arasındaki "Bağlantı Dengesizliği" (Linkage Disequilibrium - LD)'nin avantajından yararlanır. SNP'lerin karakteristik bir özelliği olan LD, popülasyonun genetik geçmişinden dolayı tüm genomdaki herhangi bir SNP konum çifti arasında görülen bir korelasyondur [12]. LD, SNP'ler birbirlerinden bağımsız olmadığında gözlemlenir. Bu yüzden bir SNP konumu ve LD ilişkisi kullanılarak diğer SNP konumlarının içeriğinden çıkarılabilir [12, 32]. Bahsedilen tehdidin en bilinen örneği, Alzheimer hastalığıyla ilişkilendirilen ApoE genini çıkararak kendi genomunu yayınlayan Dr. Jim Watson'un ApoE risk durumudur. Çünkü Watson'ın ApoE geni üzerindeki SNP'lerin, yayınlanan SNP'ler ve LD arasındaki ilişkiler kullanılarak çıkarılabileceği gösterilmiştir [12, 32].

Şekil 5'te tamamlama yöntemi ile genetik bilgi çıkarımının nasıl yapılabileceği örneklenmiştir. Art niyetli yetkisiz bir kişi çevrimiçi anonim kayıtlara erişerek elde ettiği demografik verileri üst veri ve yan kanal sızıntılarındaki yazışmalar ile eşleştirerek genetik kaydın sahibini ortaya çıkarabilir. Daha sonra bu verileri kullanarak çevrimiçi kamusal genetik veritabanları üzerinden genetik kaydın sahibinin soy ağacını elde edebilir. Böylece genetik verilerinin yayınlanması düşünülmeyen soy ağacı üyelerinin de bilgileri yetkisiz taraflar tarafından ortaya çıkarılabilmektedir. Humbert vd. [33] tarafından yapılan çalışmada OpenSNP.org dan bulunan bir bireyin profili Facebook üzerindeki akrabaları aranarak bir

tamamlama saldırısı için kullanılmıştır. Bireyin akrabalarının genotipleri tahmin edilerek Alzheimer hastalığına olan genetik yatkınlıkları değerlendirilmiştir.

Kısmi veriler kullanılarak genetik bilgilerin yeniden yapılandırılmasına dayanan bu saldırıya çözüm olarak Alser vd. [1] her bireyin (bireyin kendisi, ailesi veya genomik araştırmacılar) kamuya açık olarak paylaştığı tüm verilere dikkat etmesi gerektiğini savunmuşlardır. Mevcut tamamlama yöntemleri ile eksik genomik bilgi tamamlanabiliyorsa genomik verilerin belirli kısımlarının veri setlerinden çıkarılması gerektiğinden bahsedilmiştir. Bir başka çözüm olarak genomun sadece bazı kısımlarına araştırmacıların erişimine izin veren adanmış kriptografik yöntemlerin kullanımı önerilmiştir.

Bu çalışmada açıklanan gizlilik problemlerinin adı, problem sonucunda oluşan saldırılar, bu saldırılara sebep olan istismar edilebilir zafiyetler, saldırı sonuçları ve çözüm önerilerinin genel bir özeti Tablo 1'de verilmiştir.

IV. GENOMİK VERİLERİN GİZLİLİĞİ İÇİN AÇIK SORUNLAR (OPEN ISSUES FOR PRIVACY OF GENOMIC DATA)

Genomik verilerin gizliliğini sağlama hususunda çoğunlukla kriptografik yöntemler kullanılmaktadır. Bu kriptografik yöntemler sadece verinin saklanması için değil, bulut tarafında üçüncü kişilerden ve bulut altyapısından veriyi gizlemek için şifreli veriler üzerinden güvenli hesaplamaların yapılması için kullanılmaktadır. Güvenli hesaplama için yaygın olarak kullanılan Tamamen Homomorfik Şifreleme (Fully Homomorphic Encryption) yöntemleri veriyi şifrelemeyi ve ardından bu veri üzerinden hesaplamaların yapılmasını desteklese de talepte bulunan kişiye ulaştığında verinin deşifrelenmesi gerekmektedir. Bu da istemci tarafında istenmeyen bir durumdur. Genomik verilerin gizliliği için kullanılabilecek bir diğer yöntem fonksiyonel şifreleme yöntemidir. Bu yöntem şifreli veri üzerinde hesaplamalara izin vermekte fakat doğrudan açık metin üretmektedir. Bu yöntemlerin ikisi de gerçek dünya uygulamalarında etkili çalışamadığı için homomorfik şifreleme kullanmadan genomik verilerin istismarını önlemek için daha özelleştirilmiş ve etkili çözümlere ihtiyaç duyulmaktadır. Veri paylaşımından ayrı olarak bulut bilişim ortamında Bölüm 3'te bahsedildiği gibi insan genomunu hizalama ve karşılaştırma, SNP değerlerinin çıkarımı, kişi özelliklerinin DNA verileri aracılığıyla çıkarımı gibi birçok güvenlik ve gizlilik tehdidi bulunmaktadır. Bu tehditlere yönelik literatürde yer alan çözümler (SNP değerlerine gürültü ekleme vb.) genom mekanizmasının doğruluğunu ve verinin etkin kullanımını olumsuz etkileyebilmekte ve yüksek hesaplama maliyeti getirmektedir. Bu yüzden araştırmacılar ve veri sahiplerinin en büyük endişelerden biri olan üçüncü taraf ortamlarından verileri korumak için güvenli ve etkili bir çözüm bulmak açık bir problem haline gelmiştir. Ayrıca verilerin anonimliği sağlandığı düşünülse bile geliştirilen bazı algoritmalar [34] anonim olan verileri halka açık kayıtlarda yer alan genomik veriler ile karşılaştırarak yeniden tanımlamaya sebebiyet vermektedir. Bu yüzden verilerin anonimliğini sağlayan gizlilik koruyan sistemler için test mekanizmalarına ihtiyaç duyulmaktadır.

V. SONUÇ (CONCLUSION)

Yüksek verimli dizileme yapılarak üretilen genomik veriler bulut bilişimin ucuz depolama ve paralel hesaplama özelliği sayesinde gün geçtikçe daha kolay ve hızlı bir şekilde toplanmakta, depolanmakta ve işlenmektedir. Soy ağacı çıkarma, yeni doğan tarama, ilaç bağımlılığı, alerji ve belirli bir hastalığa eğilim gibi artan sayıdaki DNA testleri kamusal veya özel bulut sağlayıcılara aktarılmakta ve çeşitli nedenler için bireylerin DNA'ları sağlayıcılar tarafından dizilenmekte, araştırmacılar tarafından ise bu bilgilere dolaylı ya da doğrudan erişilmektedir. DNA verilerinin güvenlik ve gizlilik problemleri henüz yeterli bir şekilde tanımlanıp ele alınmadığında kullanıcı izni olmadan genomik bilgilere istenmeyen erişimler, ciddi gizlilik ihlallerine neden olabilmektedir. Dizileme teknolojilerinin kullanımıyla birlikte genomik araştırmalar her ne kadar gelişim gösterse de mahremiyetin korunması eksikliğinden dolayı genomik bilgilerin dış ortama aktarılması faydadan çok zarara sebebiyet verebilmektedir.

Tablo 1. Genomik verilere erişimde karşılaşılan gizlilik problemleri

Problem	Saldırı	Zafiyet	Saldırı Sonuçları	Çözüm Önerleri
DNA dizisi hizalama ve karşılaştırma	○ Gizli genom verisinin açığa çıkarılması	▪ Genomik verilerin genel bulutlarda depolanması ve işlenmesi	❖ Veri gizliliği ihlali	➢ "Tohumla ve genişlet" yöntemi ile güvenli okuma haritalama işlemi [16] [18] ➢ Blockchain zinciri ile DNA hizalama işlemi [17] ➢ Homomorfik şifreleme ve dinamik programlama ile dizi karşılaştırma [15]
Genomik verilerin sorgusu	○ Gizli genom verisinin açığa çıkarılması	▪ Genomik verilerin herkese açık ortamlarda şifresiz tutulması	❖ Veri gizliliği ihlali ❖ Sağlayıcıya ait ticari hakların hilali	➢ Sonlu otomata ve homomorfik şifreleme ile hataya dayanıklı (error-resilient) DNA sorgulama [19] ➢ Verileri işlemek için çoklu üçüncü taraf ihtyacını ortadan kaldıran kriptografik donanım [20] ➢ Bloom Filtre tabanlı terim arama algoritması [21] ➢ Gizli Markov Modeli kullanarak genomik dizilerin mahremiyetini sağlama [23] ➢ Dinamik simetrik aranabilir şifreleme yöntemi [24]
Üst veri ve yan kanal sızıntıları ile kimlik izi sürme	○ Çevrimiçi anonim kayıtlara erişilmesi ○ Demografik verilerin meta veri ve yan kanal sızıntılarındaki yazışmalar ile eşleştirilmesi ○ Gerçek kaydın sahibinin kimliğinin elde edilmesi	▪ Benzersiz verilerin anonim olarak saklanması ▪ Dosya isimlerinin kayıt sahibinin gerçek ismini içermesi	❖ Veri gizliliği ihlali ❖ Genetik ayrımcılık ❖ Mali kayıp ❖ Genom sahibine şantaj yapma	➢ Doğum tarihi, cinsiyet ve posta kodu gibi benzersiz verilerin anonim olarak saklanmasından kaçınılması [26] [28] ➢ 2002 Sağlık Sigorta Taşınabilirliği ve Hesap Verebilirlik Yasası (HIPAA) Gizlilik Kuralı'nın takip edilmesi [1]

DNA üzerinden kişi özelliklerinin ifşa edilmesi	○ Bazı hastalıkların veya genotiplerin dağılımı ile ilgili çalışmalardan ve istatistiklerden bilgi edinilmesi ○ Yayınlanmış çalışmalara katılan bireylerin belli bir hastalığa sahip olma olasılıklarının hesaplanması ○ Mağdur hakkında yayınlanmış bilgilerin ortaya çıkarılması	▪ Düşük sayıda SNP verisinin eşleşme için yeterli olması ▪ Alel frekansların üst düzey veri özetleme istatistiklerinin açık erişim alanında depolanması	❖ Veri gizliliği ihlali ❖ Genetik ayrımcılık ❖ Mali kayıp ❖ Genom sahibine şantaj yapma	➢ Diferansiyel gizlilik yöntemi [31] ➢ Şaşırtma mekanizmasına dayalı genomik gizliliği koruma mekanizması (GPPM) [32]
Tamamlama saldırıları	○ Çevrimiçi anonim kayıtlara erişilmesi ○ Demografik verilerin meta veri ve yan kanal sızıntılarındaki yazışmalar ile eşleştirilmesi ○ Gerçek kaydın sahibinin kimliğinin elde edilmesi ○ Mağdurun soy ağacının çıkarılması ○ Genomları hiç yayınlanmayacak bireylerin genetik bilgilerinin elde edilmesi	▪ Benzersiz verilerin anonim olarak saklanması ▪ Kamuya açık verilerin özelliklerine dikkat edilmemesi	❖ Genetik ayrımcılık, ❖ Mali kayıp ❖ Genom sahibine şantaj yapma	➢ Genomun sadece bazı kısımlarına araştırmacıların erişimine izin veren adanmış kriptografik yöntemlerin kullanılması [1] ➢ Kamuya açık olarak paylaşılanı tüm verilere dikkat edilmesi [1] ➢ Genomik verilerin belirli kısımlarının veri setlerinden çıkarılması [1]

Bu çalışmada genom verilerinin depolanması, aktarılması ve işlenmesi sırasında meydana gelen problemler ele alınmıştır. Ayrıca genom sistemlerinin etkili çalışmasını ve araştırmacıların bu sistemlerden yararlanmasını etkilemeden verilerin gizliliği ve güvenliğini olabildiğince sağlayan yeni yöntemlerin geliştirilmesi adına literatürde ele alınan mevcut güvenlik problemleri ve çözümleri kategorilere ayrılarak özetlenmiştir. Bununla birlikte genom verilerin gizliliği için hâlihazırda var olan açık problemlere de yer verilmiştir.

KAYNAKLAR (REFERENCES)

[1]. M. Alser, N. Almadhoun, A. Nouri, C. Alkan, and E. Ayday, "Can you really anonymize the donors of genomic data in today's digital world?," in *Data Privacy Management, and Security Assurance*: Springer, 2015, pp. 237-244.

[2]. M. M. A. Aziz *et al.*, "Privacy-preserving techniques of genomic data—a survey," *Briefings in bioinformatics*, vol. 20, no. 3, pp. 887-895, 2019.

[3]. B. Wang, "Search over Encrypted Data in Cloud Computing," PhD., Virginia Polytechnic Institute and State University, 2016.

[4]. M. Naveed *et al.*, "Privacy in the genomic era," *ACM Computing Surveys (CSUR)*, vol. 48, no. 1, pp. 1-44, 2015.

[5]. X. Qiu *et al.*, "Cloud technologies for bioinformatics applications," in *Proceedings of the 2nd Workshop on Many-Task Computing on Grids and Supercomputers*, 2009, pp. 1-10.

[6]. W.J. Lu, Y. Yamada, and J. Sakuma, "Privacy-preserving genome-wide association studies on cloud environment using fully homomorphic encryption," in *BMC medical informatics and decision making*, 2015, vol. 15, no. S5: Springer, p. S1.

[7]. M. Beck *et al.*, "Genecloud: Secure cloud computing for biomedical research," in *Trusted Cloud Computing*: Springer, 2014, pp. 3-14.

[8]. M. Kantarcioglu, W. Jiang, Y. Liu, and B. Malin, "A cryptographic approach to securely share and query genomic sequences," *IEEE Transactions on information technology in biomedicine*, vol. 12, no. 5, pp. 606-617, 2008.

[9]. B. Schneier, *Applied cryptography: protocols, algorithms, and source code in C*. john wiley & sons, 2007.

[10]. NIH. "Guidance for Institutions Submitting Grant Applications and Contract Proposals under the NIH Genomic Data Sharing Policy for Human and Non-Human Data." https://gds.nih.gov/pdf/GDS_Policy_Guidance_Grant_App_Contract_Proposals.pdf (accessed April 28, 2019.

[11]. M. Akgün, A. O. Bayrak, B. Ozer, and M. Ş. Sağıroğlu, "Privacy preserving processing of genomic data: A survey," *Journal of biomedical informatics*, vol. 56, pp. 103-111, 2015.

[12]. E. Ayday, E. De Cristofaro, J.-P. Hubaux, and G. Tsudik, "The chills and thrills of whole genome sequencing," *Computer*, 2013.

[13]. L. Bonomi, Y. Huang, and L. Ohno-Machado, "Privacy challenges and research opportunities for genomic data sharing," *Nature Genetics*, pp. 1-9, 2020.

[14]. NCBI. https://www.ncbi.nlm.nih.gov/projects/SNP/ (accessed February 02, 2020, 2019).

[15]. M. J. Atallah, F. Kerschbaum, and W. Du, "Secure and private sequence comparisons," in *Proceedings of the 2003 ACM workshop on Privacy in the electronic society*, 2003, pp. 39-44.

[16]. Y. Chen, B. Peng, X. Wang, and H. Tang, "Large-Scale Privacy-Preserving Mapping of Human Genomic Sequences on Hybrid Clouds," in *NDSS*, 2012.

[17]. A. M. Ileri, H. I. Ozercan, A. Gundogdu, A. K. Senol, M. Y. Ozkaya, and C. Alkan, "Coinami: a cryptocurrency with DNA sequence alignment as proof-of-work," *arXiv preprint arXiv:1602.03031*, 2016.

[18]. Y. Zhao, X. Wang, and H. Tang, "A Secure Alignment Algorithm for Mapping Short Reads to Human Genome," *Journal of Computational Biology*, vol. 25, no. 6, pp. 529-540, 2018.

[19]. M. Blanton and M. Aliasgari, "Secure outsourcing of DNA searching via finite automata," in *IFIP Annual Conference on Data and Applications Security and Privacy*, 2010: Springer, pp. 49-64.

[20]. M. Canim, M. Kantarcioglu, and B. Malin, "Secure management of biomedical data with cryptographic hardware," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 1, pp. 166-175, 2011.

[21]. H. Perl, Y. Mohammed, M. Brenner, and M. Smith, "Fast confidential search for bio-medical data using bloom filters and homomorphic cryptography," in *2012 IEEE 8th International Conference on E-Science*, 2012: IEEE, pp. 1-8.

[22]. H. Perl, Y. Mohammed, M. Brenner, and M. Smith, "Privacy/performance trade-off in private search on bio-medical data," *Future Generation Computer Systems*, vol. 36, pp. 441-452, 2014.

[23]. M. Franz, B. Deiseroth, K. Hamacher, S. Jha, S. Katzenbeisser, and H. Schröder,

"Towards secure bioinformatics services (short paper)," in *International Conference on Financial Cryptography and Data Security*, 2011: Springer, pp. 276-283.

[24]. X. Lei, X. Zhu, H. Chi, and S. Jiang, "Privacy-preserving use of genomic data on mobile devices," in *2015 IEEE/CIC International Conference on Communications in China (ICCC)*, 2015: IEEE, pp. 1-6.

[25]. X. Zhu, E. Ayday, R. Vitenberg, and N. R. Veeraragavan, "Privacy-Preserving Search for a Similar Genomic Makeup in the Cloud," *arXiv preprint arXiv:1912.02045*, 2019.

[26]. L. Sweeney, A. Abu, and J. Winn, "Identifying participants in the personal genome project by name (a re-identification experiment)," *arXiv preprint arXiv:1304.7605*, 2013.

[27]. L. Sweeney, "Weaving technology and policy together to maintain confidentiality," *The Journal of Law, Medicine & Ethics*, vol. 25, no. 2-3, pp. 98-110, 1997.

[28]. L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557-570, 2002.

[29]. S. R. Savage, "Characterizing the risks and harms of linking genomic information to individuals," *IEEE Security & Privacy*, vol. 15, no. 5, pp. 14-19, 2017.

[30]. A. Das, "Approaches in Genomic Privacy," Computer Science & Center for Computational Molecular Biology (CCMB), Brown University, 2018.

[31]. Y. Erlich and A. Narayanan, "Routes for breaching and protecting genetic privacy," *Nature Reviews Genetics*, vol. 15, no. 6, pp. 409-421, 2014.

[32]. M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti, "Reconciling utility with privacy in genomics," in *Proceedings of the 13th Workshop on Privacy in the Electronic Society*, 2014, pp. 11-20.

[33]. M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti, "Addressing the concerns of the lacks family: quantification of kin genomic privacy," in *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, 2013, pp. 1141-1152.

[34]. B. Malin and L. Sweeney, "How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems," *Journal of biomedical informatics*, vol. 37, no. 3, pp. 179-192, 2004.