T.C.
HACETTEPE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

EKZOM VERİ SETİNDEN HASTALIĞA ÖZGÜ VARYANT VERİ TABANI OLUŞTURULMASI

Yavuz ADABALI

Biyoinformatik Programı

YÜKSEK LİSANS TEZİ

Ankara

2019

T.C.
HACETTEPE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

EKZOM VERİ SETİNDEN HASTALIĞA ÖZGÜ VARYANT VERİ TABANI OLUŞTURULMASI

Yavuz ADABALI

Biyoinformatik Programı

YÜKSEK LİSANS TEZİ

TEZ DANIŞMANI

Prof. Dr. Ayşe Nurten AKARSU

İKİNCİ DANIŞMAN

Dr. Öğr. Ü. İdil Yet

Ankara

2019

T.C. HACETTEPE ÜNİVERSİTESİ

SAĞLIK BİLİMLERİ ENSTİTÜSÜ

EKZOM VERİ SETİNDEN HASTALIĞA ÖZGÜ VARYANT

VERİ TABANI OLUŞTURULMASI

Yavuz Adabalı

Danışman: Prof. Dr. Ayşe Nurten AKARSU

İkinci Danışman: Dr. Öğr. Ü. İdil Yet

Bu tez çalışması 29/08/2019 tarihinde jürimiz tarafından "Biyoinformatik Programı"nda yüksek lisans tezi olarak kabul edilmiştir.

Jüri Başkanı:	**Doç. Dr. Yeşim AYDIN SON** (*Ortadoğu Teknik Üniversitesi*)
Tez Danışmanı:	**Prof. Dr. Ayşe Nurten AKARSU** (*Hacettepe Üniversitesi*)
Üye:	**Doç. Dr. Tunca DOĞAN** (*Hacettepe Üniversitesi*)

Bu tez Hacettepe Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca yukarıdaki jüri tarafından uygun bulunmuştur.

19 Eylül 2019

Prof. Dr. Dicle[illegible] ORHAN

Enstitü Müdürü

YAYIMLAMA VE FİKRİ MÜLKİYET BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe Üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alınarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayınlanan ***"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"*** kapsamında tezim aşağıda belirtilen koşullar haricince YÖK Ulusal Tez Merkezi / H.Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

o Enstitü / Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 2 yıl ertelenmiştir. (1)

o Enstitü / Fakülte yönetim kurulunun gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren ... ay ertelenmiştir. (2)

o Tezimle ilgili gizlilik kararı verilmiştir.

24../09./2019

Yavuz ADABALI

---

[1] *"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"*

(1) Madde 6. 1. Lisansüstü tezle ilgili patent başvurusu yapılması veya patent alma sürecinin devam etmesi durumunda, tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulu iki yıl süre ile tezin erişime açılmasının ertelenmesine karar verebilir.

(2) Madde 6. 2. Yeni teknik, materyal ve metotların kullanıldığı, henüz makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış ve internetten paylaşılması durumunda 3. şahıslara veya kurumlara haksız kazanç imkanı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulunun gerekçeli kararı ile altı ayı aşmamak üzere tezin erişime açılması engellenebilir.

*(3) Madde 7. 1. Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı, tezin yapıldığı kurum tarafından verilir *. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise, ilgili kurum ve kuruluşun önerisi ile enstitü veya fakültenin uygun görüşü üzerine üniversite yönetim kurulu tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir. Madde 7.2. Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir*

** Tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulu tarafından karar verilir.*

ETİK BEYAN

Bu çalışmadaki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi, görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu, kullandığım verilerde herhangi bir tahrifat yapmadığımı, yararlandığım kaynaklara bilimsel normlara uygun olarak atıfta bulunduğumu, tezimin kaynak gösterilen durumlar dışında özgün olduğunu, Tez Danışmanı Prof. Dr. Ayşe Nurten Akarsu ve eş danışman Dr. Öğr. Ü. İdil Yet'in danışmanlığında tarafımdan üretildiğini ve Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü Tez Yazım Yönergesine göre yazıldığını beyan ederim.

24.09.2019

Yavuz Adabalı

TEŞEKKÜR

Eğitim hayatım boyunca her zaman yanımda olan aileme, tez ve uygulama geliştirme süreçlerinde emeklerini ve desteklerini hiçbir zaman esirgemeyen benimle birlikte bu süreçte yoğun mesai harcayan başta danışmanım Prof. Dr. Ayşe Nurten AKARSU olmak üzere ikinci danışmanım Dr. Öğr. Ü. İdil YET'e , tezin neredeyse her bir satırında katkısı bulunan, sorguların hazırlanmasında emek veren, yazılım testlerini gerçekleştiren ve aynı zamanda tanıdığım en iyi "uygulama kullanıcısı" olan Dr. Öğr. Ü. Arda ÇETİNKAYA'ya, teze büyük katkılar sunan ve sorguların oluşturmasında yardımlarını esirgemeyen Ar. Gör. Can KOŞUKCU'ya; kurumsal proje ve alt yapı geliştirme süreçlerini paylaşan Kaan ÖZDEMİR'e, Yüksek Lisans eğitimim süresince kolaylıklar sağlayan Aile, Çalışma ve Sosyal Hizmetler Bakanlığı Bilgi İşlem Dairesi Başkanlığı (AÇSHB-BİDB) Daire Başkanları Mustafa ÖZAŞIK, Ömer Abdullah KARAGÖZOĞLU, Abdullah Sait BOZKURT ,Ömer ÖĞREDİCİ'ye ve AÇSHB BİDB Sosyal İstihdam Birim Yöneticisi Muhammed NASSER'e teşekkürlerimi borç bilirim.

Ayrıca, Biyoinformatik Yüksek Lisans eğitimimde bilgilerini sunan HÜ Biyoinformatik Anabilim Dalı üyesi değerli hocalarıma teşekkür ederim.

Son olarak uygulama verilerinin elde edilmesinde emeği geçen HÜ Genetik Anabilim Dalı çalışanlarına teşekkür ederim.

ÖZET

Adabalı, Y., Ekzom Veri Setinden Hastalığa Özgü Varyant Veri Tabanı Oluşturulması, Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü Biyoinformatik Programı Yüksek Lisans Tezi, Ankara, 2019. İnsan Genom Projesi ile insan referans genom dizisinin oluşturulması ardından ileri nesil dizileme teknolojileri ile katlanarak artan sayıda bireyin genomik dizileri ortaya çıkarılmaya devam etmiştir. Böylece, insanlar arasında genomik dizide milyonlarca farklılığın (varyantın) olduğunu görülmüştür. Bu genetik varyantların bir kısmının genetik hastalıklara sebep olduğu bilinmektedir; ancak bir dizileme analizinde ortaya çıkan pek çok varyanttan hangisinin hastalıkla ilişkili olduğunu saptamak zordur. Bu tezin amacı, genetik hastalık şüphesi olan kişilerden "Ekzom" dizilemesi yapılarak elde edilen genetik varyant verilerinin Türk populasyonuna özgü verileri içeren bir varyant veri tabanı oluşturmakta kullanılması ve karşılaştırmalı varyant analizi için bir platform sağlanmasıdır. Bu amaçla, veri tabanı yönetim sistemi olarak MS-SQL; uygulama arka yüz teknolojisi olarak ASP, .NET, MVC; nesne ile ilişkisel haritalama/eşleme aracı olarak Entity Framework; ön yüz teknolojisi olarak ise HTML5, CSS teknolojileriyle Bootstrap, Javascript ve Jquery kütüphaneleri kullanılmıştır. Kurulan sistem, *Ion Reporter* yazılımı aracılığıyla anote edilen tabuler dosya formatlarını (.tsv, .csv gibi) kullanan genişletilebilir bir veri tabanıdır. Veri tabanında çeşitli varyant özelliklerine göre sorgu yapılabilmektedir. Ayrıca, veri tabanındaki hastaya özgü varyantlar, hastalık tipine ve kalıtım modeline göre 3 farklı modda, kurumsal ve uluslararası varyant veri tabanlarındaki varyant sıklıklarına ve varyantların pozisyonu/etkisine göre karşılaştırmalı varyant filtrelenebilmektedir. Sonuçta, oluşturulan veri tabanı hastalıkla ilişkili olabilecek genetik varyantların Türk toplumuna özgü veriler kullanılarak hızlı ve etkin analizini sağlamakta ve büyük genetik verilerin analizini kolaylaştırmaktadır.

Anahtar kelimeler: Tüm Ekzom dizileme, Genetik veri tabanı, Genetik varyasyon, Genotip

ABSTRACT

Adabalı, Y., Building a Disease-specific Variant Database from Exome Datasets, Hacettepe University Graduate School Health Sciences Department of Bioinformatics Master's Thesis, Ankara, 2019. After the completion of human reference genome sequence with the Human Genome Project, an increasing nıumber of individuals have been sequenced with next-generation sequencing technologies. This has shown that millions of genetic differences, called genetic variants, exists between individuals. Some of these genetic variants are known to cause genetic disorders. However, it is difficult to pinpoint a disease-causing variant among the many variants present in an individual. The aim of this thesis is to collect genetic variant data from individuals with suspected genetic disorders to establish a variant database. This database will provide analysis of specific variants for Turkish population and providing a platform that allows comparative analysis of individual data. For this purpose, MS-SQL as the database mangement system; ASP, .NET, MVC in the back-end; Entity Framework as object-relational mapping tool; HTML5 and CSS technologies and Bootsrap, Javascript ve Jquery libraries in the front-end were used. The built system establishes an expandable database by incorporating tabular file formats such as .tsv, .csv which are annotated by the Ion Reporter software. The database allows query options with respect to several variant properties. In addition, variants for an individual in the database can be compared against the other variants which can be filtered by 3 modes of disease type and inheritance pattern, frequency of variants in in-house and international variant databases and the effect or position of variants. In conclusion, the established database provides a quick and effective analysis of genetic variants that can be related to several diseases by using specific data for Turkish population and facilitates the analysis of this big genetic data.

Key Words: Whole exome sequencing, Genetic database, Genetic variation, Genotype

İÇİNDEKİLER

SİMGELER ve KISALTMALAR

1KGP	1000 Genom Projesi (*1000 Genomes Project*)
3' UTR	3’ Translayona Uğramayan Bölge (*3' Untranslated Region*)
5' UTR	5’ translayona Uğramayan Bölge (*5' Untranslated Region*)
AAF	Afrika Allel Frekansı (*African Allelic Frequency*)
ABD	Amerika Birleşik Devletleri
ADA2	Adenozin Deaminaz 2 (*Adenosine Deamiınase 2*)
ADO	ActiveX Veri Objesi (*ActiveX Data Object*)
AJAX	Asenkron JavaScript ve XML (*Asynchronous JavaScript and XML*)
AMAF	Afrikalı Amerikan Minör Allel Frekansı (*African American Minor Allele Frequency*)
ANOVA	Varyans Analizi (*Analysis of Variance*)
ANSI	Amerikan Ulusal Standartlar Enstitüsü (*American National Standards Institute*)
B Tree	Dengeli Ağaç (*Balanced Tree*)
bam	İkili Hizalama / Harita Dosyası (*Binary alignment/map file*)
BL	İş Katmanı (*Business Layer*)
cDNA	Kodlayan DeoksiriboNükleik Asit (coding DeoxyriboNucleic Acid)
CERN	Avrupa Nükleer Araştırma Konseyi (*European Council for Nuclear Research*)
CMDB	Çin Milyonom Veri Tabanı *(Chinese Millionome Database)*
CODASYL	Veri Sistemleri Dilleri Konferansı (*Conference on Data Systems Languages*)
COSMIC	Kanser Somatik Mutasyonlar Kataloğu (*Catalogue of Somatic Mutations in Cancer*)
CRUD	Oluştur, Oku, Sil, Güncelle (*Create, Read, Update, Delete*)
CSS	Basamaklı Biçim Sayfaları (*Cascading Style Sheets*)
csv	Virgülle Ayrılmış Değer (*Comma Separated Value*)
DAL	Veri Erişim Katmanı (*Data Access Layer*)
DBA	Diamond-Blackfan Anemisi

dbSNP	Tek Nükleotid Değişimi Veri Tabanı (*database of Single Nucleotide Polymorphisms*)
DGV	Genomik Varyant Veri Tabanı (*Database of Genomic Variants*)
DNA	DeoksiriboNükleik Asit
DRA	Hastalık Araştırma Alanı (*Disease Research Area*)
EAAF	Doğu Asya Allel Frekansı *(East Asian Allelic Frequency*)
EFAF	Avrupa Fin Allel Frekansı (*European Finnish Allelic Frequency*)
EMAF	Avrupalı Amerikalı Minör Allel Frekansı (*European American Minor Allele Frequency*)
emPCR	Emülsiyon Temelli Polimeraz Zincir Reaksiyonu
ENFAF	Avrupa Fin Olmayan Allel Frekansı (*European Non-Finish Allelic Frequency*)
ESP	Ekzom Dizileme Projesi (*Exome Sequencing Project*)
ExAC	Ekzom Toplama Konsorsiyumu *(The Exome Aggregation Consortium)*
FAA	Fanconi Aplastik Anemisi
FASTQ	Hızlı adaptif Büzüşme Eşiği algoritması – Kalite skoru (*Fast Adaptive Shrinkage Threshold algorithm – Quality score*)
FATHMM	Gizli Markov Modelleriyle Fonksiyonel Analiz (*Functional Analysis through Hidden Markov Models*)
FBLIM1	Filamin Bağlayan LIM (lin-11, isl-1, mec-3) proteini 1 (*Filamin-Binding LIM (lin-11, isl-1, mec-3) protein 1*)
FK	İkincil Anahtar (*Foreign Key*)
GAF	Global Allel Frekansı (*Global Allelic Frequency*)
GATK	Genom Analizi Araç Seti (*Genome Analysis Toolkit*)
GMAF	Global Minör Allel Frekansı (*Global Minor Allele Frequency*)
GnomAD	Genom Toplama Veritabanı (*Genome Aggregation Database*)
GO	Gen Ontolojisi (*Gene Ontology*)
GoNL	Hollanda Genomu (*Genome of the Netherlands*)
hg16	İnsan genom tümleşkesi verisyon 16 (*Human genome assembly version 16*)
HTML	Hiper Metin İşaretleme Dili (*Hypertext Markup Language*)

http	Hiper Metin Aktarım Protokolü (*HyperText Transfer Protocol*)
HUMAF	Hacettepe Üniversitesi Minör Allel Frekansı
IBM	Uluslararası İş Makineleri Şirketi (*International Business Machines Corporation*)
ID	Tanımlayıcı (*IDentifier*)
IGV	Entegre Genomik Görüntüleyici (*Integrated Genomics Viewer*)
IGVdb	Hindistan Genom Varyasyon Veri Tabanı (*The Indian Genome Variation Database*)
IMS	Bilgi Yönetim Sistemi (*Information Management System*)
INGRES	İnteraktif Grafik ve Geri Çağırma Sistemi (*Interactive Graphics and Retrieval System*)
indel	İnsersiyon ve/veya delesyon
ISO	Uluslararası Standardizasyon Organizasyonu (*International Organization for Standardization*)
JSNP	Japon Tek Nükleotid Polimorfizm (*Japanese Single Nucleotide Polymorphism*)
LAF	Latin Allel Frekansı (*Latino Allelic Frequency*)
LINQ	Dile Entegre Sorgu (*Language INtegrated Query)*
MAF	Minör Allel Frekansı
MAM	Marmara Araştırma Merkezi
MaxMAF	Maksimun Minör Allel Frekansı
MinHomopolimer	Minimum Homopolimer Uzunluk
MinMAF	Minimum Minör Allel Frekansı
MNV	Çoklu Nükleotid Değişikliği (*Multiple Nucleotide Variation*)
mRNA	Mesajcı RNA
MS-SQL Server	Microsoft Yapılandırılmış Sorgu Dili Sunucusu (*Microsoft Structured Query Language Server*)
MVC	Model, Görüntü, Kontrolcü (*Model View Controller*)
NCBI	Ulusal Biyoteknoloji Bilgi Merkezi (*National Center for Biotechnology Information*)
NCBI34	Ulusal Biyoteknoloji Bilgi Merkezi İnsan Genom Tümleşkesi versiyon 34 (*National Center for Biotechnology Information Human Genome Assembly version 34*)

NGS	Yeni Nesil Dizileme (*Next Generation Sequencing*)
NoSQL	İlişkisel olmayan Yapısal Sorgu Dili (*Non Relational Structred Query Language*)
OAF	Diğer Allel Frekansı (*Other Allelic Frequency*)
OCA2	Okulokutanöz Albinizm Proteini 2 (*OculoCutaneous Albinism Protein 2*)
OMIM	İnsanlarda Mendelyan Kalıtım - Çevrimiçi (*Online Mendelian Inheritance in Man*)
ORM	Nesne ile İlişkisel Haritalama/Eşleme (*Object Relational Mapping*)
PanSNPdb	Pan-Asya SNP Genotipleme Veri Tabanı (*The Pan-Asian SNP Genotyping Database*)
PFAM	Protein Aileleri Veri Tabanı (*Protein Families Database*)
pH	Potansiyel Hidrojen
Php	Hiper Metin Ön İşlemcisi (*Hypertext Preprocessor*)
PK	Birincil Anahtar (*Primary Key*)
QUEL	Sorgu Dili (*Query Language*)
RNA	RiboNükleik Asit
RPS26	Ribozomal Protein Küçük 26 (*Ribozomal Protein Small 26*)
SAAF	Güney Asya Allel Frekansı (*South Asian Allelic Frequency*)
SDS	Shwachman-Diamond Sendromu
SGVP	Singapur Genom Varyasyon Projesi
SIFT	Tolere edilebiliri tolere edilemezden ayırma (*Sorting Intolerant from Tolerant*)
SLC25A12	Çözünen Madde Taşıyıcı Ailesi Proteini 25 A2 (*Solute Carrier family protein 25 A2*)
SNV	Tek Nükleotid Değişikliği (*Single Nucleotide Variation*)
SOAP	Basit Nesne Erişim Protokolü (*Simple Object Access Protocol*)
SQL	Yapılandırılmış Sorgu Dili (*Structured Query Language*)
SSMS	SQL Sunucusu Yönetim Sistemi (*SQL Server Management Studio*)
TEX11	(*Testis-Expressed gene 11*)

T-SQL	Transakt Yapılandırılmış Sorgu Dili (*Transact Structured Query Language*)
tsv	Sekmeyle Ayrılmış Değer (*Tab Seperated Value*)
TÜBİTAK	Türkiye Bilimsel ve Teknolojik Araştırma Kurumu
TÜSEB	Türkiye Sağlık Enstitüleri Başkanlığı
UCSC	Kaliforniya Üniversitesi Santa Cruz (*University of California, Santa Cruz*)
vcf	varyant belirleme formatı (*variant caller format*)
W3C	Dünya Çapında Ağ Birliği (*World Wide Web Consortium*)
XML	Genişletilebilir İşaretleme Dili (*eXtensible Markup Language*)
XSS	Siteler Arası Betik Çalıştırma Zaafiyeti (*Cross-site scripting*)

ŞEKİLLER

TABLOLAR

1. GİRİŞ

Son yüzyılın en önemli projelerinden biri olan ve başta insan olmak üzere farklı türlerin referans genomlarını dizilemeyi hedefleyen "İnsan Genom Projesi" ABD, Fransa, İngiltere, Japonya ve Çin işbirliğinde tamamlanarak 2003 yılında insan genomuna ait referans dizi belirlenmiştir. Buna göre insan genomu 3,2 milyar bazdan oluşmakta, bunun ancak %1,5'luk kısmı proteinlere kodlanan fonksiyonel gen bölgelerini içermektedir (1). Geri kalan %98,5'luk kısım ise kodlanmayan genom dizisinden oluşmakla birlikte bu bölgelerde yer alan regülatör elemanlar gen fonksiyonlarını düzenlemede kritik rol üstlenmektedirler (2). İnsan genom projesinin tamamlanması pek çok açıdan genetik alanında çığır açacak yeniliklerle doludur. Bunların önemli bir sonucu nedeni bilinmeyen genetik hastalıkların nedenlerinin tespit edilmesinde hızlı ve göreceli olarak kolay uygulanabilen yöntemlerin geliştirilmiş olmasıdır. Gelişen yüksek ölçekli dizileme yöntemleri ile birlikte insan genomunu tek koşumda dizileyebilmek mümkün hale gelmiştir. Eş zamanlı olarak kurumsal alt yapılarda yüksek ölçekli genom dizileme yöntemleri uygulanmaya başlamış ve hastalıklara neden olan genlerin saptanmasında belirgin başarı elde edilmesine neden olmuştur.

Tüm genomda yerleşik olan yaklaşık 22.000 genin kodlanan bölgelerinin hedefli dizilemesi tüm ekzom dizileme olarak adlandırılmış ve özellikle hemen tamamı Mendel kalıtımı özellikleri gösteren nadir hastalıklara neden olan gen değişikliklerinin saptanması için etkin olarak kullanılmaya başlanmıştır. Buna karşın tüm dünyada tamamlanan geniş kapasiteli ekzom dizileme sonuçları nadir hastalıkların ancak %30'unda ekzom dizileme ile hedefe ulaşılabildiğini göstermiştir (3). Bunun farklı sebepleri olmakla birlikte hastalıktan sorumlu olduğu düşünülen genom değişikliklerinin (varyant) tespit edilmesinde populasyonlara özgü varyant bilgilerinin yetersiz oluşu ön plana çıkan nedenler arasındadır. Bu nedenle, tüm ekzom ve tüm genom çalışmalarında varyant bilgilerinin veri tabanlarında toplanması ve populasyon temelli veri tabanlarının etkin veri yönetim sistemleri ile idare edilmesi kaçınılmaz bir gereklilik olarak karşımıza

çıkmıştır. Buna yönelik olarak farklı toplumlardan yüz binlerce bireyin varyant bilgilerini içeren ExAC, gnomAD, *Greater Middle East Genome Project* (Büyük Orta Doğu Genom Projesi) gibi veri tabanları ortaya çıkmış ve bu veri tabanlarının *web* uygulamaları ile erişimi ve kullanımı olanaklı hale gelmiştir (4-6). Bunlara ek olarak, tüm toplumlar kendi toplumlarına özgü genom projelerini tamamlamakta ve veri tabanlarını oluşturmaktadır. Araştırıcıların büyük bir çoğunluğu ise, kurumsal alt yapılarda elde edilen tüm ekzom ve tüm genom verilerini kurumsal veri tabanlarında (*in-house database*) toplamakta ve bu veri tabalarındaki genomik bilgileri birbirleri ile karşılaştırarak bireylerde saptanan varyant sayısını azaltma yolunu seçmektedir.

Ülkemizde populasyona özgü bir varyant veri tabanı bulunmamaktadır. Bu alanda TÜBİTAK, TÜSEB gibi farklı kurumların çabaları olmakla birlikte henüz tüm araştırıcıların kullanımına açılan bir veri tabanı oluşturulamamıştır (7,8). Bu açıdan, filtrelemede öncelik belirleme açısından kaçınılmaz bir gereklilik olan kurumsal veri tabanı oluşturulması ve bunun etkin bir veri yönetim sistemi ile idare edilmesi ülkemizde henüz mümkün değildir. Veri tabanları yapısal bilgi ya da verilerin elektronik ortamda organize şekilde toplandığı yapılardır. Veri tabanı sistemlerinin kontrolleri ise veri tabanı yönetim sistemleri olarak adlandırılan yazılımlarla gerçekleştirilmektedir. Veri tabanı yönetim sistemleri informasyon teknolojilerinin biyolojik bilgilerin yorumlanmasına doğrudan katkı yaptığı güncel ve sürekli değişime açık bir alan olarak karşımıza çıkmaktadır.

Sunulan tez çalışmasında, Hacettepe Üniversitesi Tıbbi Genetik Anabilim Dalı bünyesinde elde edilen ekzom verilerinin elektronik ortamda toplanması, farklı kalıtım kalıplarına özgü modeller geliştirilerek hızlı ve etkin filtrelemeye yönelik yazılım geliştirilmesi planlanmıştır.

2. GENEL BİLGİLER

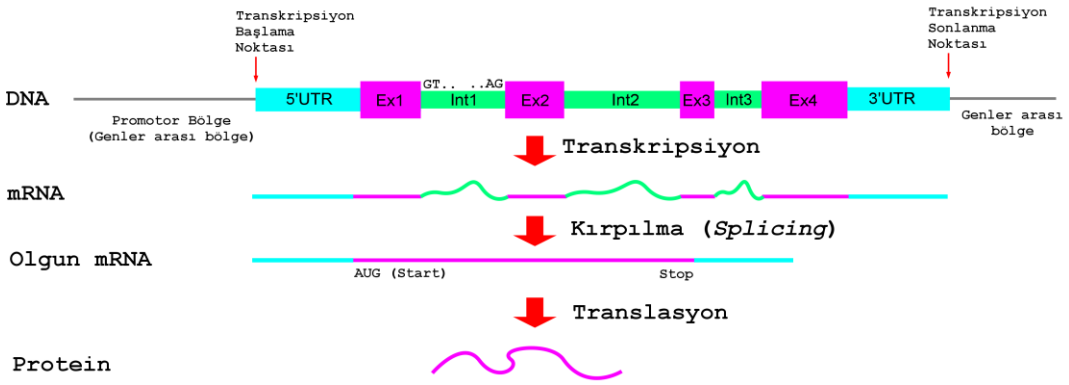
2.1. İnsan Genomu ve Genom Mimarisi

İnsan genomu, hücre çekirdeğinde bulunan 22 çift otozomal kromozom, 1 çift eşey kromozomu ve mitokondriyal DNA'dan meydana gelen yaklaşık 3,2 milyar baz çiftinin oluşturduğu genetik dizinin bütünüdür. İnsanları oluşturan esas hücreler olan somatik hücreler" diploiddir — genomları 2 adet birbirinden ufak farklılıklar içeren kopya şeklinde olup 6,4 milyar baz çiftinden oluşmaktadır (9). Eşey hücreleri olan gametlerde ise genom haploid olup DNA'larında 3,2 milyar baz çifti bulunmaktadır.

DNA molekülünün yapısındaki işlevsel birimlere gen adı verilir. Bu genlerin çoğunluğunu 20.000-21.000 adet olan protein kodlayan genler oluşturmaktadır (10). Genler, kendileri doğrudan işlev gören RNA'ların veya protein kodlanmasına aracılık edecek mRNA'ların dizisine kalıp oluştururlar. Genler dışındaki alanlara ise genler arası bölge adı verilmektedir. Bu bölgelerde çok sayıdaki tekrar dizisinin yanında genlerin başı/sonunu belirleyen diziler ve gen aktivitesinin kontrolünü sağlayan diziler de mevcuttur. Protein kodlayan bir gen, dizisini kodladığı mRNA'ların üzerindeki protein kodlamasına katılan ekzon; kodlamaya katılmayarak mRNA sentezi aşamasında kırpılma (*splicing*) mekanizması ile uzaklaştırılan intron; mRNA'nın yapısına katılan ancak protein kodlayan dizinin öncesindeki 5' UTR (*5' untranslated region*, 5' translayona uğramayan bölge) ve sonrasındaki 3' UTR (*3' untranslated region*, 3' translayona uğramayan bölge) bölgelerinden oluşur (Şekil 1.1). Tüm ekzonlar (ekzom), tüm genomun %1-2'lik bir alanını oluşturmaktadır (1,11).

İntronların kırpılmasının gerçekleşmesi için ekzonların başlangıç ve bitişinde bulunan diziler kırpılma mekanizması tarafından tanınmalıdır. Bunun için intronun ilk ve son 2 nükleotid uzunluğundaki kısmı kritik önem taşımaktadır. Bu dizi çoğunlukla ekzonun 5' verici (*donor*) ucunda konumlanmış GT baz dizisi ile 3' alıcı (*acceptor*) ucunda yer alan AG baz dizisini içermektedir. İnsan genomunun kaba taslak yapısı 80'li yılların

başından beri bilinse de genom dizinin detaylı şekilde ortaya çıkarılması 2001 yılında İnsan Genom Projesi'nin ilk sonuçları yayınlandıktan sonra gerçekleşmiştir (1).



Şekil 2.1. İnsanlarda protein kodlayan bir genin bölümleri. Protein kodlayan bir gen protein dizisine katılan ekzon; mRNA'ya kodlandıktan sonra kırpılan (*splicing*); bu dizilerin öncesindeki 5' UTR (*5' untranslated region*, 5' translayona uğramayan bölge) ve sonrasındaki 3' UTR (*3' untranslated region*, 3' translayona uğramayan bölge) bölgelerinden oluşur. Transkripsiyonun başladığı ve bittiği noktalar genin sınırını belirlerken bunlar dışında kalan bölgelere genler arası bölge adı verilir. 5'UTR öncesindeki genler arası bölge transkripsiyonun başlayacağı noktayı belirleyen promotor dizilerini de içerir. mRNA olgunlaştıktan sonra sadece ekzonlardan kaynaklanan dizilerin protein dizisini kodladığına dikkat ediniz.

2.1.1. İnsan Genom Projesi

İnsanoğlu gözlem yapıp sorgulamaya başladığı andan itibaren "Ben kimim?" sorusunun yanıtını aramıştır. 1990 yılında resmi olarak başladığında bu soruyu cevaplamak için ilk adımların atılmasını sağlayan İnsan Genom Projesi, 6 farklı ülkedeki 20 merkezde gerçekleştirilen kapsamlı bir çalışma sonucunda 14 Nisan 2003 tarihinde tamamlanmıştır. Bu merkezlerin 12'si ABD, 3'ü Almanya, 2'si Japonya ve birer tanesi Çin, Fransa ve İngiltere'de bulunmaktadır (12). İnsan haploid genom dizisinin belirlenmesini sağlayan bu proje sonrasında yaklaşık 3 milyar nükleotidden oluşan referans dizi ortaya

çıkarılmış ve özellikle nadir genetik hastalıkların etiyolojisinin aydınlatılması için dev bir adım atılmıştır. Genetik alanında yapılan çalışmaların sayısı arttıkça bireyin genomundan ziyade popülasyonun gen havuzunun karakterize edilmesinin hastalıkların etiyolojisinin saptanmasında önemi fark edilmiştir. Bazı toplumlarda mutasyon olarak kabul edilen varyantların yeniden kategorize edilmesi ve genişletilmiş popülasyon verileri ile karşılaştırılması sonucunda bu değişikliklerin polimorfizm olarak sınıflandırıldığı ve hastalığa yol açmadığı görülmüştür (13). Böylelikle, insan genom projesinin amacı olan referans dizinin oluşturulmasının yeterli olmadığı, aslında cevaplanması gereken sorunun ise "Biz kimiz?" olduğu ortaya çıkmıştır. Bununla uyumlu olarak pek çok ülke-bölge, kendi lokal popülasyon veri tabanlarını oluşturmaya başlamıştır. Genom dizilemeleri günlük hayata girmiş, çok sayıda veri kurumsal altyapılarda üretilir hale gelmiştir. Araştırıcılar, popülasyon veri tabanlarını kullanırken bunun yanı sıra kendi kurumsal verilerini birbirleri ile karşılaştırmayı da etkin bir strateji olarak benimsemişlerdir.

2.1.2. Topluma ve Bireye Özgü Genetik Değişiklikler (Varyantlar)

Belirli bir popülasyonun kendi içinde veya farklı toplumlar arasındaki yaygın genetik çeşitliliklere polimorfizm adı verilmektedir. Hastalık ile ilişkilendirilebilecek nadir değişiklikler ise "mutasyon" olarak adlandırılmaktadır (14). Çok sayıda verinin bir arada kullanılması sonucu, genetik değişikliklerin farklı toplumlarda görülme sıklıklarının değiştiği anlaşılmıştır. Genom dizisindeki çeşitlilik hem çalışılan hastalığın türüne hem de bireylerin genom arka planlarındaki değişikliklere göre şekillenmektedir. İster hastalığa sebep olsun, isterse polimorfizm olarak değerlendirilsin, referans genom dizisine göre farklılık gösteren genetik değişikliklerin tamamına "varyant" adı verilmektedir.

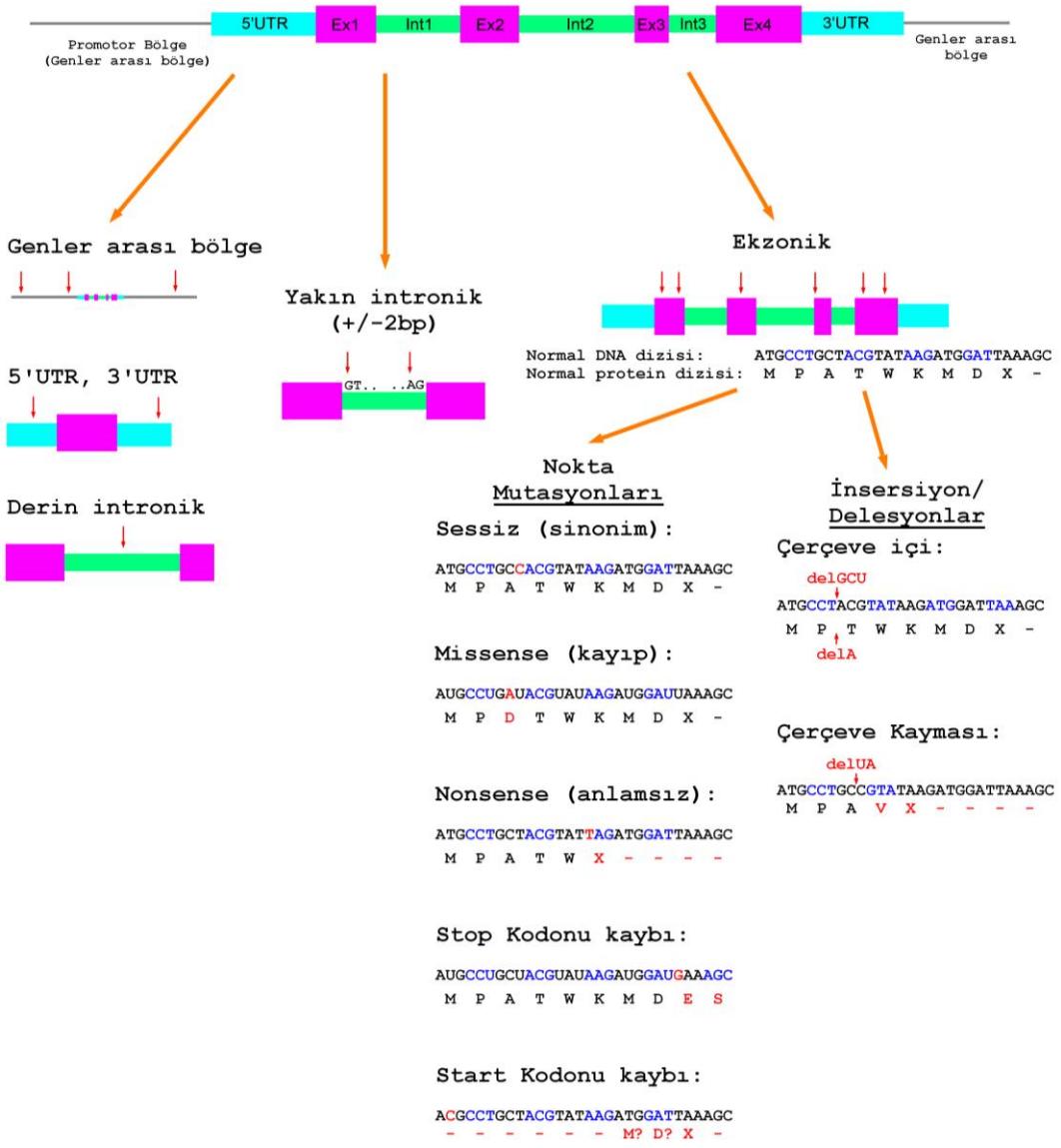
Genomda bulunan bütün varyantlar farklı toplumlar için farklı frekansta ortaya çıkmaktadır. Özellikle nadir hastalık çalışmalarında hastalık etkeni bir genetik değişikliğin mutasyon olarak adlandırılabilmesi için toplumda görülme sıklığı kritik önem taşımaktadır. Referans allelden sonra en sık rastlanan allelin toplumda görülme sıklığı

%1'in altındaysa bu genetik değişiklik nadir hastalıklar için mutasyon adayı olarak değerlendirilebilir (7). Minör allel frekansı (MAF) olarak adlandırılan ve bir toplumda belirli bir varyant için ikinci sık rastlanan allel, nadir hastalıkların genetik etiyolojisinin aydınlatılması için en önemli değerlendirme kriteridir. Özellikle artan genetik popülasyon verileri ve karakterize edilen mutasyon tipleri ile birlikte nadir hastalıklar için %1 olarak belirlenen eşik değerinin güncellenmesi önerilmiştir. Shearer ve ark., 2014 yılında yaptıkları çalışmanın sonucunda mutasyon olarak bildirilmiş MAF değerlerini gözden geçirip varyantları tekrar sınıflandırmıştır. Elde edilen sonuçlar doğrultusunda otozomal resesif geçişli nadir hastalıklar için MAF eşik değeri 0,005, otozomal dominant geçişli nadir hastalılar için ise 0,0005 olarak önerilmiştir (13).

Genetik değişiklikler, protein kodlayan ve kodlamayan genleri, hatta genler arası bölgeleri etkileyebilir. Bu genetik değişiklikler arasında özellikle protein kodlayan genlerdeki genetik değişikliklerin protein üzerine etkisini öngörmek mümkündür. Bir mutasyon, protein kodlayan gendeki yerine göre ekzonik, intronik, 5'UTR, 3'UTR ve yakın genler arası bölge mutasyonları olarak sınıflandırılabilir. Bu mutasyonlardan ekzonik mutasyonlar doğrudan amino asit yapısına etki edebilirken diğer mutasyonların etkileri ancak dolaylı olabilir. Dolaylı etkileri öngörmek zor olmakla beraber özellikle intronun ilk 2 bazı ve son 2 bazındaki değişikliklerin kırpılma düzenini değiştirerek amino asit dizisinde büyük ekleme (insersiyon), çıkarma (delesyon) ve çerçeve kaymalarına sebep olabildiği bilinmektedir (15). Ekzonik dizilerdeki değişiklikler ise nükleotid sayısını değiştirip değiştirmediğine göre ikiye ayrılabilir: Nokta mutasyonları ve insersiyon/delesyonlar. Nokta mutasyonları, sessiz (sinonim), *missense* (yanlış anlamlı mutasyon), *nonsense* (anlamsız mutasyon) ve daha nadir birkaç farklı şekilde olabilir. İnsersiyon/delesyonlar için ise eklenen/eksilen nükleotid sayısı mutasyonun etkisini belirlemede önemlidir. Her üç nükleotidin protein üzerinde bir amino asidi belirleyen kodon dizisini oluşturduğu düşünüldüğünde üçün katı sayıda olan insersiyon/delesyonlar sadece amino asit insersiyon/delesyonlarına neden olan çerçeve içi (*in-frame*) değişikliklerken; üçün katı olmayan sayıdaki değişiklikler çerçeve kaymasına (*frameshift*) sebep olmaktadır. Bu

değişiklikler ve etkileri Şekil 2.2'de özetlenmiştir. Özellikle proteinlerin erken sonlanmasına sebep olabilen çerçeve kayması ve *nonsense* mutasyonlarının protein işlevi üzerinde daha büyük kayba neden oldukları bilinmektedir. Ayrıca, bu mutasyonlarda *nonsense* aracılı RNA parçalanması mekanizması ile protein üretiminin de kısıtlanarak mutasyonların zararlı etkisini artırmaktadır (16).

Bir genetik değişikliğin mutasyon olarak adlandırılması için MAF, mutasyonun protein yapısı üzerine etkisi ve bunun gibi birçok kriter bir arada değerlendirilmelidir. Genomun tamamını tek seferde dizilemeye olanak kılan yeni nesil DNA dizileme teknolojileri ile pek çok genetik değişikliğin geniş toplumlardaki sıklığı ve dağılımını da belirlemek mümkün hale gelmiştir.



Şekil 2.2. Protein kodlayan genlerdeki farklı mutasyonların etkileri. Protein kodlayan gen mutasyonlarının gen üzerindeki yerine bakıldığında genler arası bölge, 5'UTR, 3'UTR ve derin intronik varyantların etkilerini öngörmek zordur. Öte yandan, intronun ekzona komşu 1-2 nükleotidindeki varyantlar çoğunlukla *splicing* mekanizmasını bozduğu için protein yapısında ciddi sorunlara sebep olurlar. Bunun yanında, doğrudan protein kodlayan diziyi oluşturduğu için ekzonik varyantların protein üzerindeki etkilerini öngörmek daha kolaydır. Ekzonik varyantların etkileri şeklin sağ bölümünde gösterilmiştir. Her bir varyant türünün etkisi

altında yer alan DNA dizisi ve protein dizisi ile örneklenmiştir. Bu örneklerde, kırmızı ile belirtilen değişiklikler DNA nükleotid dizisindeki değişiklikleri ve bunların protein amino asit dizisindeki sonuçlarını göstermektedir.

2.1.3. Genom ve Varyant Veri Tabanları

İnsan genom projesi tamamlanıp referans dizi aydınlatıldıktan sonra özellikle genetik hastalıklar üzerine çalışan araştırıcıların yararlanmasına yönelik veri tabanlarının oluşturulması için temeller atılmaya başlanmıştır. Resmî olarak Nisan 2003'te projenin ilk sonuçlarının yayınlanmasını takiben yaklaşık üç ay sonra ABD kaynaklı "*National Center for Biotechnology Information*" kurumunun NCBI34/hg16 genom versiyonunu yayınlaması ile ilk insan genom veri tabanı da ortaya çıkmıştır (17). Yeni nesil DNA dizileme teknolojilerinin yaygın olarak kullanılmaya başlanması ile Ekzom Dizileme Projesi (*Exome Sequencing Project*/ESP) ve 1000 Genom Projesi (*1000 Genomes Project*/1KGP) gibi geniş popülasyonlardaki yüksek ölçekli DNA dizileme çalışmaları sonucunda gerek genom gerekse varyant düzeyinde bilgi içeren veri tabanları oluşturulmaya başlanmış bugün varyant filtreleme aşamalarında etkin olarak kullanılmaktadır (18,19).

NCBI (National Center for Biotechnology Information) – GenBank Veri Tabanı

NCBI'ın (https://www.ncbi.nlm.nih.gov/) DNA dizi verilerini içeren "GenBank" arşivi, 1982 yılının aralık ayında üçüncü çıktılarını açık erişimli bir veri tabanı şekilde yayınladığında farklı organizmalara ait 606 DNA dizisi ve 680.338 nükleotidden oluşan bir veri bankası halinde idi. Haziran 2019'da 232. çıktının verilerine göre ise bu veri tabanı yine farklı organizmalara ait 213,4 milyon DNA dizisi ve 329,8 milyar nükleotidden meydana gelen geniş bir referans kaynak niteliği taşımaktadır (20). Bu kaynak, özellikle tüm genom verilerinin yüklenmesiyle yaklaşık 1 milyar genom DNA dizisi ve 4,8 trilyon nükleotid içeren devasa bir veri tabanı olarak kullanılmaktadır (20). NCBI bünyesinde

referans dizilerin bulunduğu GenBank, *web* servis hizmeti sağlayıcılarından *eutils* Python paketi ve *SOAP* XML temelli protokolü kullanmaktadır.

UCSC Genome Browser

UCSC (https://genome.ucsc.edu/); Kaliforniya Üniversitesi, *Santa Cruz* araştırıcıları tarafından oluşturulan ve idamesi sağlanan, hem kullanıcı dostu bir arayüz sunması hem de ihtiyaç duyulan verinin esnek bir şekilde indirilebilmesi gibi olanakları sayesinde sıklıkla başvurulan bir genom tarayıcısıdır. Birçok model organizma dahil olmak üzere çok sayıda omurgalı ve omurgasız canlı türünün referans genom dizilerini içerir. Bu referans dizilerin interaktif şekilde kullanılmasına olanak sağlayan bir *web* sitesi tasarımına sahiptir. UCSC bünyesindeki tarayıcı, hızlı etkileşimi destekleyecek şekilde optimize edilmiş bir arayüze sahiptir ve verilerin hızlı bir şekilde görselleştirilmesi, incelenmesi ve sorgulanması için MySQL veri tabanının üzerine kurulmuştur.

UCSC genom tarayıcısı, 2001 yılında İnsan Genom Projesi'nin ilk meyvelerinin dağıtımı için bir kaynak olarak kullanılmaya başlamıştır. Günümüzde 100'den fazla türe ait 180'in üzerinde genom dizi içermektedir (21).

The Exome Aggregation Consortium (ExAC) Browser

ExAC veri tabanı (http://exac.broadinstitute.org/), UCSC'den farklı olarak bütün bir insan referans genom verisini içermez. Bunun yerine farklı etnik kökenlerden gelen ve konjenital bir hastalığı olmadığı bilinen 60.706 bireyin tüm ekzom dizileme yöntemi ile elde edilen varyant bilgilerini barındırır (4). ExAC varyant veri tabanı, nadir genetik hastalıklar üzerine çalışan araştırma grupları için bir kontrol niteliğindedir. Bu veri tabanında 60.706 birey için tüm ekzom dizileme sonucu elde edilen varyantların tamamı, genotip sayıları, frekansları ve okuma derinliği gibi kritler kullanılarak sınıflandırmıştır. Ağustos 2016 tarihinden itibaren veri tabanına kontrol gruplarındaki kopya sayısı değişiklikleri de eklenmiştir (4).

Genome Aggregation Database (gnomAD)

gnomAD veri tabanı (https://gnomad.broadinstitute.org/) ise farklı etnik kökenlerden gelen ve konjenital bir hastalığı olmadığı bilinen 125.748 bireyin tüm ekzom dizileme, 15.708 bireyin ise genom dizileme yöntemi ile elde edilen varyant bilgilerini barındırır. gnomAD veri tabanında çalışmaya katılan bireyler için toplamda 17,2 milyon ekzonik, 261,9 milyon genomik varyant tespit edilmiştir. Bu varyantlar çalışmaya özgü bir rastgele orman (*random forest*) veri madenciliği yöntemi ile işlenmiş ve filtreleme sonucunda 14,9 milyon ekzonik, 229,9 milyon filtrelenmiş yüksek kalitede varyant elde edilmiştir (5).

Geniş kapsamlı bir referans niteliği olmasına karşın içerisinde genetik bir hastalığa sahip bireylerin de bulunabileceği göz önüne alınarak, bu veri tabanının nadir genetik hastalıklar için varyant filtreleme basamaklarında kullanılması tasarımcılar tarafından önerilmemektedir.

Popülasyonlara Özgü Veri Tabanları

Özellikle nadir hastalık çalışmalarında yüksek ölçekli DNA dizileme sonuçlarının filtrelenmesi için varyantların toplumda görülme sıklığı kritik öneme sahiptir. Her ne kadar ExAC ve gnomAD gibi farklı popülasyonlardan bireylerin oluşturduğu veri tabanları bu filtreleme basamakları için kullanışlı olsa da bazı genetik değişikliklerin toplumlara özgü olduğu düşünüldüğünde birçok ülke kendi lokal veri tabanlarını oluşturmak için çalışmalara başlamıştır. Dünyada bunun örnekleri arasında Çin Milyonom Veri tabanı (CMDB), İngiltere Genom Projesi, Hollanda Genom Projesi (GoNL), Japon Tek Nükleotid Polimorfizm Projesi (JSNP), Pan-Asya SNP Genotipleme Veri Tabanı (PanSNPdb), Hindistan Genom Varyasyon Veri Tabanı (IGVdb), Singapur Genom Varyasyon Projesi (SGVP) ve Estonya Genom Projesi olarak sıralanabilir (22-29). Ülkemizde de yüksek ölçekli DNA dizileme yöntemleri ile elde edilen veriler kullanılarak yapılan çalışmalar mevcuttur. Alkan ve ark., 2014 yılında yaptıkları genom dizileme çalışmasında 16 Türk bireye tüm genom dizileme gerçekleştirmiş ve bu bireylerin genetik yapılarını ortaya koymuştur (30). Toplamda 1.111 Orta Doğulu bireyden elde edilen ekzom dizileme verileri ile gerçekleşen

Greater Middle East Genome Project (Büyük Orta Doğu Genom Projesi) içerisinde ise toplamda 140 Türk ekzom dizileme verisi kullanılmıştır (6). Bu çalışmalar haricinde, henüz yayınlanmamış TÜBİTAK Marmara Araştırma Merkezi'nde (TÜBİTAK-MAM) gerçekleştirilmiş çok sayıda ekzom ve tüm genom dizileme verisi bulunmakta, ayrıca Türkiye Sağlık Enstitüleri Başkanlığı'nın da (TÜSEB) 100.000 Türk genom projesi kapsamında ilk aşamada gerçekleştirdiği 100 Türk bireyin genom dizileme verisi bulunmaktadır (7,8). Ancak bunlara erişim henüz mümkün değildir. Bu nedenle, ekzom/genom analizlerinde hastalıklarla ilişkili değişikliklerin saptanmasında Türkiye'ye özgü araştırıcıların genel erişimine açık bir veri tabanı olmadığı için öncelik belirleme aşamalarında büyük bir zorlukla karşılaşılmaktadır.

2.2. Yeni Nesil DNA Dizileme Teknolojileri (NGS)

İnsan genom projesi sonrasında büyük genom parçalarını kurumsal alt yapılar içinde kısa bir zamanda dizilemek ve veri elde etmek mümkün hale gelmiştir. Gene yönelik hedefli dizilemeler, tüm ekzom dizileme ve tüm genom dizileme yöntemlerinin bütünü yeni nesil dizileme (*next generation sequencing*, NGS) olarak adlandırılmaktadır. 3,2 milyar baz çiftinden meydana gelen insan genomunun yarısı tekrar dizilerinden, %25'i genler arası (intergenik) bölgelerden, %23'ü intronik bölgelerden ve yaklaşık %1,5'luk kısmı protein kodlamasına katılan ve ekzon adı verilen DNA dizilerinden meydana gelmektedir (31). Nadir genetik hastalıkların büyük bir kısmı da bu %1,5'luk protein kodlamasına katılan dizilerde meydana gelen hatalardan kaynaklanmaktadır. Bu gerçekten yola çıkarak geliştirilen, özellikle genetik hastalıkların tanı ve tedavisine yönelik tasarlanmış hedefli DNA dizileme yöntemine ekzom dizileme adı verilmektedir. Ekzom dizileme yönteminde yaklaşık 30 milyon nükleotid ve 180.000 ekzonik gen bölgesi dizilenebilmektedir. Yeni nesil dizileme yöntemlerinin pratikte en sık kullanılan türü olan tüm ekzom dizileme metodu ile konjenital genetik hastalıklara sahip bireylerin yaklaşık

%30'una kesin tanı konabilmektedir (32). Yeni nesil dizilemenin en yaygın kullanıldığı alan olan nadir hastalık çalışmaları da bu teknolojinin gelişmesi ile beraber hızlanmıştır.

2.2.1. Farklı Platformlara Göre Ekzom Veri Eldesi

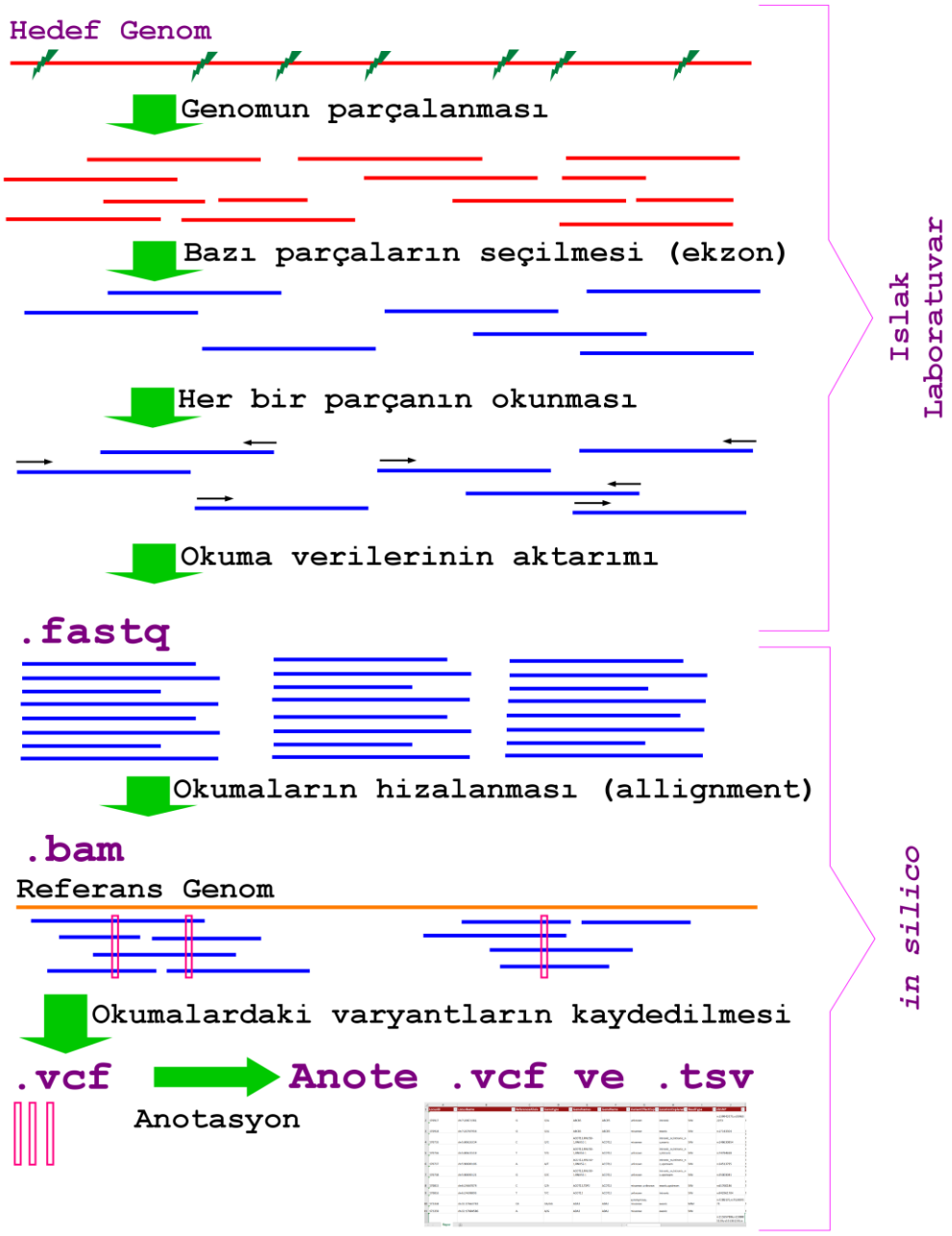
Yeni nesil DNA dizileme teknolojisi ortaya çıktığından beri birçok farklı üreticinin geliştirmiş olduğu çok sayıda platform kullanılmıştır. Bunlar arasından günümüzde en yaygın olarak tercih edilen ve kabul görenleri ise Illumina ve *Thermo Fisher Scientific* altyapılarıdır. İki platform arasındaki en belirgin fark Illumina tarafından geliştirilen yüksek okuma kapasiteli modellerin tüm genom dizileme için de elverişli olması; ancak *Thermo Fisher Scientific* altyapısındaki cihazlar ile gerçekleştirilebilen en yüksek hacimli dizileme protokolünün ekzom dizileme olmasıdır.

Illumina platformu ile gerçekleştirilen DNA dizilemede kütüphaneler, oligo nükleotidlerle kaplanmış bir yüzeye (*flowcell*) bağlanır. Amplifikasyon bu yüzey üzerinde gerçekleşir ve bu şekilde DNA kümeleri oluşturulur. Ekzonik gen bölgelerinin polimerizasyonu için nükleotidler eş zamanlı olarak tepkimenin gerçekleşeceği ortama salınır ve yüzey üzerindeki oligo nükleotidler ile eşleşemeyen moleküller ortamdan uzaklaştırılır. Oligo nükleotidler ile eşleşme gerçekleştiğinde ortaya çıkan enerji enzim tepkimeleri ile ışığa dönüştürülür ve özel kamera sistemleri ile algılanır (33).

Thermo Fisher Scientific (*Ion Torrent/Ion Proton*) sistemlerinde ise DNA parçaları, ekzonik gen bölgelerini hedefleyen primerler ile çoğaltılır. Ardından emülsiyon temelli polimeraz zincir reaksiyonu (*emPCR*) metodu kullanılarak nano boyuttaki küreler etrafında ve mikroyağ reaktörleri içerisinde bir amplifikasyon daha meydana gelir. İkinci amplifikasyondan elde edilen ürünler manyetik küreler ile yakalanır, saflaştırılır ve dizileme aşamasında nükleotid bağlanması ile değişen ortam pH'ının algılanması ile dizilenir (34). İleri nesil dizilemenin her iki sistemde de ortak kullandığı basamakların görsel bir özeti Şekil 2.3'te sunulmuştur. Her iki sistemde de büyük kopya sayısı değişikliklerinin tespit edilmesinde sorunlar yaşanmaktadır (35). Özellikle, amplifikasyon

temelli NGS çalışmalarında tek allel kaybolduğu zaman ortamda kalıp bulunduğu için amplifikasyon olacağından ilgili kaybı tespit etmek hibridizasyon bazlı tekniklere göre daha zordur. Buna ek olarak, pH değişimine duyarlı *Ion Torrent/Ion Proton* sistemleri ile yapılan dizilemelerde art arda aynı nükleotidin tekrar ettiği durumlarda (homopolimer) tekrar sayısının yanlış tespitine dayalı dizileme hataları değerlendirmede önemli bir sorun yaratmaktadır (36).

Bu tez çalışmasında üretilen veri tabanı, Hacettepe Üniversitesi Tıbbi Genetik Anabilim Dalı bünyesinde *Thermo Fisher Scientific - Ion Proton* sistemi ile üretilen ekzom dizileme verilerine göre tasarlanmıştır. Geliştirilen yazılımın aynı zamanda Illumina platformu ile üretilen ekzom dizileme verilerinin de kullanılmasına olanak sağlaması hedeflenmiştir.



Şekil 2.3. İleri nesil dizilemenin yapılışı ve elde edilen veri formatları. İleri nesil dizilemedeki ıslak laboratuar uygulamaları üst kısımda özetlenmiştir. Alt kısımda ise *in silico* olarak NGS verisinin .fastq'dan başlayarak, .bam, .vcf, anote .vcf ve .tsv formatlarında hangi özelliklerinin bulunduğu gösterilmektedir. .bam ve .vcf formatlarında görülen mor kutucuklar varyantları ifade etmektedir.

2.2.2. Ekzom Verisinden Hastalığa Özgü Varyantların Tespit Edilmesi

Tüm ekzom dizileme yönteminde ıslak laboratuvar uygulamalarını takiben yapılan dizileme işleminden sonra çıktı olarak .fastq formatında bir metin dosyası elde edilmektedir. Tüm ekzom dizileme ile elde edilen .fastq verisi, referans genoma göre hizalanmamış okumaların nükleotid dizilimlerini içermektedir. Bu okumalar referans genoma göre hizalandıktan sonra .bam uzantılı (*binary alignment/map file*) dosyalar elde edilmektedir. .bam uzantılı hizalanmış veriler kullanılarak referans genoma göre ortaya çıkan farklılıklar (varyantlar) .vcf uzantılı (*variant caller format*) dosyalar şeklinde kaydedilir (37). *Ion Proton* sisteminden elde edilen sinyaller yerel sunucuya aktarılarak .bam ve .vcf uzantılı dosyalar oluşturulur. .vcf uzantılı dosyaların *Ion Reporter* bulut sistemine yüklenmesi ile anotasyon yapılır (**Şekil 2.3**) (38,39). Bu şekilde elde edilen varyant tablolarında ilgili genetik değişikliğin karakterize edilebilmesi için varyant ile ilgili detayların yer aldığı yaklaşık 50 kolon bulunur. Kolon sayıları, farklı *Ion Reporter* versiyonlarına göre farklılık gösterebilmektedir. Varyant detaylarını içeren kolonlar ve bu kolonların açıklaması **Tablo 2.1**'de verilmiştir (38).

Tablo 2.1. *Ion Reporter* v5.10 anote .vcf Varyant tablosunda yer alan kolon başlıkları ve bu başlıkların açıklamaları (38).

Kolon Başlığı	Açıklama
Locus	Varyantın kromozomal lokasyon bilgisini içerir.
Genotype	Varyantın genotip bilgisini içerir. Örn: G/A gösterimi; ilgili örnekte belirli bir kromozom bölgesinde "Guanin" ve "Adenin" olmak üzere iki allelin varlığını gösterir.
Filter	Varyantın okuma kalitesi kriterlerini geçip geçmediğini belirtir.
Ref	İlgili genomik lokalizasyondaki referans nükleotid bilgisini içerir.
Observed Allele	Referans allele göre değişiklik gösteren nükleotid bilgisini içerir.

Tablo 2.1. (devamı) *Ion Reporter* anote .vcf Varyant tablosunda yer alan kolon başlıkları ve bu başlıkların açıklamaları.

Kolon Başlığı	Açıklama
Type	Nükleotid değişikliğinin türünü belirtir. Örn: SNV: *Single Nucleotide Variation*, Tek nükleotid değişikliği, INDEL: Küçük insersiyon/delesyon, MNV: *Multiple nucleotide variation*, Çoklu nükleotid değişikliği anlamına gelmektedir.
Genes	Varyantın bulunduğu gen/genleri belirtir.
Location	Genetik değişikliğin gen bölgesindeki karşılığını gösterir. *unknown* (bilinmiyor), *intergenic* (Genler arası), *intronic* (intronik), *exonic* (ekzonik), *utr_5* (5' UTR), *utr_3* (3' UTR), *splicesite_5* (*5' splice* bölgesi), *splicesite_3* (*3' splice* bölgesi), *upstream* (Genin 5' ucundan önce), *downstream* (Genin 3' ucundan sonra), *exonic_nc* (protein kodlamasına katılmayan genin ekzonik bölgesi), *intronic_nc* (protein kodlamasına katılmayan genin intronik bölgesi), *ncRNA* (kodlamayan RNA), *nonCoding* (kodlamayan gen) şeklinde veriler barındırmaktadır.
Length	Varyantın kaç nükleotidi kapsadığını belirtir.
% Frequency	Varyant allel okumasının toplam okuma sayısına olan oranını verir.
Strand	Varyantın bulunduğu gen/genlerin DNA molekülündeki yerleşim yönünü gösterir. "+" simgesi ilgili varyantın bulunduğu genin 5' ucundan 3' ucuna doğru genomik lokasyon olarak artan nükleotid sayısı şeklinde yerleştiğini, "-" simgesi ise azalan şekilde yerleştiğini göstermektedir.
Exon	İlgili varyantın gendeki kaçıncı ekzonda bulunduğunu belirtir.
Transcript	Varyantın bulunduğu genin NCBI transkript kodunu içerir.
Coding	Varyantın cDNA düzeyindeki adlandırmasını belirtir.
Amino Acid Change	Varyantın protein düzeyindeki adlandırmasını belirtir.

Tablo 2.1. (devamı) *Ion Reporter* anote .vcf Varyant tablosunda yer alan kolon başlıkları ve bu başlıkların açıklamaları.

Kolon Başlığı	Açıklama
Variant Effect	İlgili genetik değişikliğin protein üzerindeki etkisini gösterir. *refAllele* (referans alleli ile aynı nükleotid), *unknown* (bilinmiyor), *synonymous* (sinonim, sessiz), *missense* (yanlış anlamlı), *nonframeshiftInsertion* (çerçeve kaymasına sebep olmayan insersiyon), *nonframeshiftDeletion* (çerçeve kaymasına sebep olmayan delesyon), *nonframeshiftBlockSubstitution* (çerçeve kaymasına sebep olmayan hem insersiyon hem delesyon), *nonsense* (anlamsız), *stoploss* (Stop kodonu kaybı), *frameshiftInsertion* (çerçeve kaymasına sebep olan insersiyon), *frameshiftDeletion* (çerçeve kaymasına sebep olan delesyon), *frameshiftBlockSubstitution* (çerçeve kaymasına sebep olan hem insersiyon hem delesyon) şeklinde veriler barındırmaktadır.
PhyloP	UCSC veri tabanındaki "*Table Browser*" sekmesinden elde edilir. Varyantın bulunduğu genomik lokalizasyonun -14 ile 3 arasında değişen PhyloP evrimsel korunmuşluk değerini belirtir (21).
SIFT	Varyantın hastalığa neden olma potansiyelinin SIFT yazılımı ile tahminlenmesi sonucu 0 ile 1 arasında bir değer elde edilir. Değer 0'a yaklaştığında olası patojenik varyant, 1'e yaklaştığında ise polimorfizm anlamına gelmektedir (40).
Grantham	*Missense* varyantlar için referans ve gözlenen amino asit arasındaki farklılığı, fizikokimyasal perspektifte hesaplar. Sınır değerleri 5 ile 215 arasındadır (41). (Örn: İzolösin-Lösin değişikliğinden elde edilen skor 5'tir ve varyant patojenik olarak değerlendirilmez. Buna karşın Sistein-Triptofan değişikliğinden elde edilen skor 215'tir ve varyant olası patojenik olarak değerlendirilir)
PolyPhen	*Missense* varyantların hastalığa neden olma potansiyelinin PolyPhen-2 yazılımı ile tahminlenmesi sonucu 0 ile 1 arasında bir değer elde edilir. Değer 1'e yaklaştığında olası patojenik varyant, 0'a yaklaştığında ise polimorfizm anlamına gelmektedir (42).

Tablo 2.1. (devamı) *Ion Reporter* anote .vcf Varyant tablosunda yer alan kolon başlıkları ve bu başlıkların açıklamaları.

Kolon Başlığı	Açıklama
FATHMM	Gizli Markov Modeli aracılığıyla protein değişikliğine etki eden ve protein kodlamasına etki etmeyen varyantların 0 ile 1 değeri arasındaki FATHMM (*Functional Analysis through Hidden Markov Models*) skorunu verir (43).
PFAM	Pfam veri tabanı içerisinde varyantın bulunduğu gen bölgesinin protein domain adlandırmasını belirtir (44).
dbSNP	dbSNP veri tabanında bulunan varyantların "rs" ile başlayan kodunu belirtir (45).
DGV	Genomik yapısal yeniden düzenlenmeler için DGV (*Database of Genomic Variants*) veri tabanındaki karşılığı belirtir (46).
MAF	Varyantın 1000 Genom Projesi içerisindeki genotip oranını belirtir (17).
EMAF	Varyantın 1000 Genom Projesi kapsamında dizilenen "*European American*" popülasyonu içerisindeki genotip oranını belirtir (17).
AMAF	Varyantın 1000 Genom Projesi kapsamında dizilenen "*African American*" popülasyonu içerisindeki genotip oranını belirtir (17).
GMAF	Varyantın 1000 Genom Projesi'nin faz 1 aşamasında dizilenen 1.904 birey içerisindeki genotip oranını belirtir (*Global Minor Allele Frequency*) (17).
UCSC Common SNPs	Varyantın UCSC veri tabanındaki "*UCSC Common SNPs*" sekmesindeki karşılığını belirtir (21).
ExAC LAF	ExAC (*Exome Aggregation Consortium*) varyant veri tabanında 5.789 kontrolden oluşan "*Latin American*" popülasyonu için ilgili varyantın genotip oranını belirtir (4).
ExAC EAAF	ExAC (*Exome Aggregation Consortium*) varyant veri tabanında 4.327 kontrolden oluşan "*East Asian*" popülasyonu için ilgili varyantın genotip oranını belirtir (4).

Tablo 2.1. (devamı) *Ion Reporter* anote .vcf Varyant tablosunda yer alan kolon başlıkları ve bu başlıkların açıklamaları.

Kolon Başlığı	Açıklama
ExAC OAF	ExAC (*Exome Aggregation Consortium*) varyant veri tabanında belirli bir popülasyon içinde değerlendirilemeyen etnik kökendeki (*Other*) 454 kontrol için ilgili varyantın genotip oranını belirtir (4).
ExAC EFAF	ExAC (*Exome Aggregation Consortium*) varyant veri tabanında 3.307 kontrolden oluşan "*European Finnish*" popülasyonu için ilgili varyantın genotip oranını belirtir (4).
ExAC SAAF	ExAC (*Exome Aggregation Consortium*) varyant veri tabanında 8.256 kontrolden oluşan "*South Asian*" popülasyonu için ilgili varyantın genotip oranını belirtir (4).
ExAC ENFAF	ExAC (*Exome Aggregation Consortium*) varyant veri tabanında 33.370 kontrolden oluşan "*European Non-Finnish*" popülasyonu için ilgili varyantın genotip oranını belirtir (4).
ExAC AAF	ExAC (*Exome Aggregation Consortium*) varyant veri tabanında 5.203 kontrolden oluşan "*African American*" popülasyonu için ilgili varyantın genotip oranını belirtir (4).
ExAC GAF	ExAC (*Exome Aggregation Consortium*) varyant veri tabanında 60.706 kontrol için ilgili varyantın genotip oranını belirtir (4).
COSMIC	Varyantın bulunduğu genin COSMIC (*Catalogue of Somatic Mutations in Cancer*) veri tabanındaki adlandırmasını gösterir (47).
OMIM	Varyantın bulunduğu genin OMIM (*Online Mendelian Inheritance in Man*) veri tabanındaki karşılığını gösterir (48).
Gene Ontology	Varyantın bulunduğu genin "*The Gene Ontology*" (GO) veri tabanındaki karşılığını gösterir (49).
DRA	Varyantın bulunduğu genin "*Disease Research Area*" (DRA) veri tabanındaki karşılığını gösterir.
DrugBank	Varyantın bulunduğu genin DrugBank veri tabanındaki karşılığını gösterir (50).

Tablo 2.1. (devamı) *Ion Reporter* anote .vcf Varyant tablosunda yer alan kolon başlıkları ve bu başlıkların açıklamaları.

Kolon Başlığı	Açıklama
ClinVar	ClinVar veri tabanına girmiş varyantlar için kullanılan etiketi belirtir (51). *Uncertain significance* (Önemi belirsiz), *Conflicting interpretations of pathogenicity* (Patojenite konusunda çelişen yorumlar), *Likely benign* (Olası selim), *Benign* (Selim), *Pathogenic* (Patojenik), *Benign/Likely benign* (Selim/Olası selim), *not provided* (varyant kayıtlı değil), *Likely pathogenic* (Olası patojenik) şeklinde veriler barındırmaktadır.
Allele Coverage	Gözlenen her allel için ayrı ayrı okuma derinliğinin sayısal değerini verir.
Allele Ratio	Gözlenen her allel için ayrı ayrı okuma derinliğinin yüzdesini verir.
p-value	Varyantın kalite skoruna bağlı olan p anlamlılık değerini belirtir.
Phred QUAL Score	Varyantın *Phred* kalite skorunun sayısal değerini verir (52).
Coverage	Toplam okuma derinliğinin sayısal değerini verir.
Ref+/Ref-/Var+/Var-	Gözlenen her allel için ayrı ayrı artı ve eksi yönlü okuma derinliklerinin sayısal değerini verir.
Homopolymer Length	Varyantın bulunduğu pozisyonun öncesindeki homopolimer uzunluğunu sayısal değer şeklinde belirtir.

Ion Reporter bulut sisteminin çıktılarından .tsv (*tab seperated value*) formatında veri elde edilebilir. Elde edilen bu sekmeyle ayrılmış metin formatı *Microsoft Office Excel* yazılımı ile açılarak bireylere özgü işlenmiş varyant tabloları izlenebilir; ancak bireyler arası varyant karşılaştırması için farklı bireylerden veri içeren varyant havuzunun bir veri tabanı şeklinde birleştirilmesi gerekmektedir. Bu şekilde, bireyler arası varyantların birbirlerine göre karşılaştırılması ve filtreleme yapılması aday varyant sayısını azaltmakta ve hastalıktan sorumlu genlerin bulunmasını kolaylaştırmaktadır.

2.3. Veri Tabanları

Veri tabanları yapısal bilgi ya da verilerin elektronik ortamda organize şekilde toplandığı yapılardır. Veri tabanı sistemlerinin kontrolleri ise veri tabanı yönetim sistemleri olarak adlandırılan yazılımlarla gerçekleştirilmektedir. Günümüz modern veri tabanlarında veriler tablo olarak adlandırılan veri tabanı nesneleri yardımıyla tutulmaktadır. Tabloların her bir satırı ilgili tabloya ait farklı değerleri taşımaktadır. Bu yöntemle veriler daha kolay erişilebilir, yönetilebilir, güncellenebilir ve kontrol edilebilir hale gelmektedir. Günümüzde ilişkisel veri tabanları, bulut veri tabanları, dağıtık veri tabanları ve NoSQL veri tabanları gibi pek çok farklı amaçla kullanılan veri tabanları mevcuttur (53).

2.3.1. Veri Tabanı Mimarileri

İlişkisel Veri Tabanları

1950'lerin sonunda bilgisayarların ticari olarak erişilebilir olmasının ardından elde edilen devasa verilerin nasıl depolanacağı sorunu ortaya çıkmıştır. Ve buna istinaden 1960'larda Charles W. Bachman tarafından *Integrated Database System* adı verilen ilk veri tabanı yönetim sistemi oluşturulmuştur (54). Birkaç yıl sonra ise IBM, *Information Management System* (IMS) adını verdiği kendi veri tabanı yönetim sistemini geliştirmiştir (55,56). Sonraki yıllarda ortaya çıkan pek çok veri tabanı yönetim sisteminin ardından veri yönetim işleminin standartlaştırılması gerekliliği ortaya çıkmıştır. 1971 yılında ise *The Database Task Group* (57) adlı organizasyon "CODASYL yaklaşımı" olarak adlandırılan standartları açıklamıştır (58).

IBM'de sabit disk sistemlerinde geliştirici olarak yer alan Edgar Codd mevcut CODASYL yaklaşımlarından ve performansından memnun olmadığı için yeni veri tabanı mimarileri öneren makaleler yayınlamıştır (59). Bunların sonuncusu olarak ise büyük veri tabanlarında verilerin nasıl depolanacağı ve işleneceği hakkındaki görüşlerini kaleme

almıştır (60). Bu makalede verilerin CODASYL'in alt yapısını oluşturan navigasyon modelindeki gibi bağlı listeler şeklinde değil ilişkisel tablolar şeklinde tutulmasından bahsetmiştir. Böylece, ilişkisel veri tabanlarının temelleri atılmıştır.

Bu makalenin ardından 1973 yılında Micheal Stonebraker ve Eugene Wong ilişkisel veri tabanı sistemleri üzerinde çalışmaya başlamışlardır (61). Projelerine INGRES (*Interactive Graphics and Retrieval System*) adını vermişlerdir (62). INGRES ile ilişkisel veri tabanlarının veri işlemleri üzerinde etkili ve pratik bir model olduğunu göstermişlerdir (61). INGRES için daha sonra QUEL (61,63) olarak adlandırılan bir sorgu dili oluşturulmuştur ve buna paralel olarak IBM'de kendi ilişkisel veri tabanını oluşturarak "Yapılandırılmış Sorgu Dili" (SQL; *Structured Query Language*) sorgulama dilini geliştirmiştir (64). 1986 yılında SQL ANSI standartlarını 1987'de de ISO standartlarını oluşturulmuştur (65-67).

İlişkisel Veri Tabanı Yönetim Sistemleri

İlişkisel Veri Tabanı Yönetim Sistemleri kullanıcıların veritabanlarına erişimini sağlayan, yeni veri tabanları oluşturmasına olanak sunan ve bu veritabanlarına erişim kontrolünü gerçekleştiren yazılım sistemleri olarak tanımlanmaktadır (68). MS-SQL *Server*, IBM DB2, *Oracle*, MySQL, MongoDB gibi tüm modern veri tabanı sistemlerinin temelini oluşturmaktadır (69). İlişkisel veri tabanlarında veriler tablo olarak isimlendirilen veri tabanı nesneleri üzerinde tutulmaktadırlar. Bir tablo basitçe ilişkisel veri girdilerinin olduğu belirli sayıda satır ve sütunlardan meydana gelmektedir.

MS-SQL Server ve Transact-SQL (T-SQL)

MS-SQL *Server*, *Microsoft* tarafından kullanıma sunulmuş bir ilişkisel veri tabanıdır. MS-SQL *Management Studio* aracılığıyla veri tabanlarına bağlanılabilir, yeni veri tabanları oluşturulabilir ya da hali hazırda kurulu bulunan veri tabanı tabloları üzerinde T-SQL sorgulama diliyle okuma, güncelleme, silme ya da yeni kayıt işlemleri gerçekleştirilebilir (70).

T-SQL, SQL sorgu dili üzerine inşa edilen ve *Microsoft* tarafından geliştirilmiş daha kapsamlı sorgulama dilidir. MS-SQL veri tabanı üzerindeki sorgulama işlemleri T-SQL aracılığı ile gerçekleştirilmektedir (71). MS-SQL veri tabanı üzerinde yer alan tablolarda kullanılan veri tipleri tam ve yaklaşık nümerikler, tarih ve zaman, karakter dizileri gibi veri tipleridir (72).

Veri Tabanı Tabloları Arasındaki İlişkiler

İlişkisel veri tabanlarında oluşturulan tablolar arasında üç tip ilişki bulunmaktadır. Bunlar; "Bire-Bir", "Bire-Çok", "Çoka-Çok" ilişkilerdir (73). Veri tabanı tabloları arasındaki bu ilişki türleri, iki tür anahtar sayesinde sağlanmaktadır. Bu anahtarlar ise;

i) Birincil Anahtar (*Primary Key*) (PK): Bir veri tabanı tablosu üzerinde özgünlüğü sağlayan anahtar türüdür. Aynı tablo içerisinde aynı değere sahip başka bir PK alanı bulunamaz.
ii) İkincil Anahtar (Foreign Key) (FK): Bir tabloda PK olarak tanımlanan alanın diğer bir tabloda yer alarak tablolar arası ilişkinin kurulmasını sağlayan anahtardır.

Veri tabanları arasındaki 3 farkı ilişki türü ise aşağıda belirtildiği şekildedir:

i) Bire-Bir İlişki: İki tablo arasında yer alan ilişkiyi sağlayan anahtar, her iki tabloda da sadece bir kez tekrar eder ve aynıdır. Başka bir deyişle, ana tabloda yer alan PK sorgusu ile ulaşılan ikincil tablodaki FK, ikincil tablo için bir PK'dır (FK=PK). Genelde bir tabloda çok fazla boş değere sahip alanlar varsa performans nedenlerinden dolayı optimizasyon amaçlı olarak bire-bir ilişki kullanılır.
ii) Bire-Çok İlişki: Bir tablodaki PK alanlı bir kayıt, diğer tabloda FK alanları aracılığıyla birden çok kayda bağlanır.
iii) Çoka-Çok İlişki: Çoka-çok ilişki bağlantılı iki tablo arasında yeni bir bağlantı tablosu olarak üçüncü bir tablonun oluşturulması ile sağlanır.

2.3.2. Veri Tabanlarında Optimizasyon İşlemleri

İndeksler, veri tabanında aramaların daha hızlı gerçekleştirilmesi için Dengeli Ağaç (*B Tree*) yapısına göre çalışan yapılardır. Veriler kümelenmiş İndekslerde (clustered index) (74) yapraklar içerisinde (*leaf*) tutulurken, kümelenmemiş(non-clustered) (75) indekslerde verinin yerine adresi yapraklarda tutulmaktadır. Veri tabanlarında belirtilen PK'lar kümelenmiş indekslerdir. Kümelenmemiş indeksler ise en çok aranan sütunlara göre veri tabanı yöneticileri tarafından belirlenmektedir.

2.3.3. Dinamik *Web* Uygulamaları

İlişkisel veri tabanlarının kullanımı çok fazla yazılım bilgisi ve programlama dili becerisi gerektirmektedir. Son kullanıcıların veri tabanı üzerindeki verilere erişmek için SQL söz dizimini bilmesini beklemek pratik değildir. Bu yüzden, veri tabanında yer alan verilere kullanıcıların rahat bir şekilde erişebileceği kullanıcı dostu arayüzlere sahip dinamik yazılımlar gerekmektedir.

Dinamik *web* uygulamaları, bir uygulama sunucusu ve veri tabanı aracılığıyla kullanıcı tipine ya da ekranda yer alan filtre gibi girdilere göre içeriğin değiştiği ve kullanıcıya sunulduğu *web* uygulamalarıdır. Dinamik *web* sayfaları temelde üç ana bileşene ayrılmaktadır. Bunlardan ilki verilerin tutulduğu veri tabanı; ikincisi veri tabanı ile iletişimde bulunan ve ilgili veriler üzerinde değişiklik yapmamıza aracılık sağlayan sunucu kısmı (*server*-side, arka yüz); üçüncüsü ve sonuncusu ise son kullanıcının gördüğü *web* tarayıcıları tarafından yorumlanan arayüzlerden oluşan son kullanıcı (ön yüz) kısmıdır (76).

Dinamik *web* uygulamalarında son kullanıcı ile sunucu arasındaki etkileşim internet aracılı veri transfer protokolü (http protokolleri) aracılığıyla sağlanmaktadır (77). Yine benzer şekilde veri tabanı sunucusu ve uygulama sunucusu arasındaki iletişim benzer protokollerle yürütülümektedir.

Dinamik *web* uygulaması oluşturulurken arka yüz ve ön yüz kısımlarında farklı mimariler, programlama dilleri ve teknolojiler kullanılmaktadır. Örneğin, arka yüz kısmında kullanılan programlama dillerine Java, Aspx .Net, Phyton, Perl, Php gibi programlama dilleri örnek olarak verilebilir (78).

Ön yüz kısmında kullanılan popüler kütüphaneler ve diller Bölüm 2.3.4'te sunulmaktadır.

2.3.4. Ön yüz mimarileri

HTML (Hyper Text Markup Language): 1980 yılında CERN'de görev yapan Tim Berners Lee tarafından temelleri atılan günümüz internet dünyasının temelini oluşturan metin tabanlı veri paylaşım dilidir (79). Html ile *web* sayfaları oluşturulurken HTML etiketleri (80) yardımıyla tarayıcılara etiketler içerisinde yer alan verilerin nasıl görüntülenmesi gerektiği bilgileri verilir. Örnek olarak, bir makale içeriğinde italik olarak gösterilmek istenen metin öğeler .html uzantılı bir dosyada etiketleri arasında yer almalıdır.

HTML ortaya çıktığı ilk günden günümüze kadar beş farklı versiyona ulaşmıştır. HTML 1.0 versiyonunda sadece metinler ve resimler için basit etiketler yer alırken, günümüz versiyonu HTML 5.0 (81) da yeni etiketler ile herhangi üçüncü parti bir uygulamaya gerek kalmadan video ya da ses ögesi oynatılabilmektedir. Ayrıca, günümüzde tablet ve cep telefonu gibi mobil cihazların popüler olmasıyla birlikte bu cihazlarda da uygun görüntülenmeyi sağlayan responsive tasarımlar HTML 5.0 ile desteklenmektedir (81).

CSS (Cascading Style Sheets): CSS, *web* sayfalarının daha kullanıcı dostu hale gelmesini sağlayan ve *web* sayfaları içeriğinde renk, genişlik, derinlik, font tipi, font boyutu, sayfa boyutu, arka plan resmi gibi pek çok öğe üzerinde değişiklik yapılmasına olanak sağlayan dildir (82). HTML 3.2 ile birlikte gelen, her bir sayfa için font, renk ve benzeri özelliklerin ayrı ayrı girilmesi ile ortaya çıkan sorunlar "*World Wide Web*

Consortium (W3C)"un 1996da CSS Tavsiyelerini "*W3C CSS Recommendation (CSS1)*" yayınlamasıyla çözüme kavuşmuştur (83). Böylece tek bir css dosyası üzerinde değişiklik yapılarak tüm sayfalar üzerinde değişiklik yapılabilir hale gelinmiştir.

Bootstrap: Orijinal adıyla *"Twitter Blueprint*" olan ve "*Bootstrap*" adıyla bilinen kütüphane, Twitter'da geliştirici olarak çalışan Mark Otto ve Jacob Thornton tarafından *web* uygulamalarının ve sitelerinin ön yüz geliştirmelerini kolaylaştırmak için tasarlanmış açık kaynak ön yüz geliştirme platformudur. Bu platformun kullanımındaki en temel farklılık duyarlı (*responsive*) bir tasarım oluşturma fırsatı sunmasıdır (84). Duyarlı tasarımlar, *web* sitelerinin mobil ve tablet cihazlardan girildiğinde site içindeki resim, yazı gibi içeriklerinin ekran genişliğine göre yeniden şekillenip cihaza göre değişiklik göstermesini sağlayan tasarımlardır.

Bootstrap, ayrıca özel tasarlanmış "*jquery*" eklentileri duyarlı tasarım desteğinin yanı sıra, tipografi, simgeler, formlar, tuşlar, tablolar ve duyarlı tasarıma temel oluşturan CSS kütüphaneleri ve HTML içerir. Bu platform *Safari*, *Google Chrome*, *Firefox*, *Microsoft Edge* ve *Internet Explorer 10+* son sürümleri gibi modern tarayıcıların tümüyle uyumlu halde çalışmaktadır (85).

JavaScript*: JavaScript* aracılığıyla bir *web* sayfası üzerinde yer alan tüm HTML öğelerine (tuşlar, dinamik ve statik metin alanları) erişilebilmekte ve bu değerlerin içeriği güncellenebilmektedir. Bu sayede statik html sayfaları ile kullanıcılar arasında interaktif ilişki kurulması sağlanmaktadır (86).

Jquery: 2006 yılında John Resig tarafından geliştirilen ve açık kaynak kodlu ücretsiz, küçük ve zengin özellikli bir *javascript* kütüphanesidir (87,88). Bu kütüphane sayesinde HTML belgesi üzerinde bulunan öğelere erişim, olayların yakalanması, animasyon ve *ajax* gibi işlemler daha kolay hale gelmektedir. Bunun nedeni, *jquery*'nin *javascript* kütüphanesine göre çok daha basit bir sözdizimine sahip olmasıdır. Ayrıca, *jquery* farklı tarayıcılar için farklı *javascript* kodlarının yazılması gerekliliğini ortadan kaldırmaktadır.

2.3.5. Arka yüz mimarileri

Son kullanıcı tarafından gelen isteklerin belirli iş kuralları dahilinde sunucu üzerinde gerçekleştirildiği, aynı zamanda veri tabanı işlemlerinin gerçekleştirilmesinde kullanılan mimarilerdir. Farklı mimariler bulunmasına rağmen tez kapsamında ".Net *Framework"* mimarisi kullanıldığı için sadece bu mimari ve içeriği hakkında detaylı bilgi verilmiştir.

.NET *Framework*, *Window*s üzerinde çalışan uygulamalarda çeşitli hizmetleri yönetme ve yürütme ortamıdır (89). .Net *framework*'un çalışan uygulamalar için sunduğu hizmetler arasında, bellek yönetimi ve düşük seviyeli programlama işlemlerini yürütmek için programcıların işini kolaylaştıran pek çok kütüphane yer almaktadır. Bu mimari içerisinde yer alan diller ve kütüphanelere aşağıda değinilmiştir.

C# Programlama Dili: *Microsoft* tarafından *Java* programlama diline karşılık olarak .Net *Framework* için tasarlanmış nesne yönelimli yazılım geliştirme için kullanılan programlama dilidir (90). Bu dil geliştirilirken, söz diziminde prosedürel programlama dili olan C ve nesne yönelimli programlama dili olan C++ dillerinin söz dizimleri referans alınmıştır. Öğrenilmesi en kolay yazılım dilleri arasında yer almasından ve neredeyse tüm platformlarda yazılım geliştirme desteği sunmasından ötürü günümüz dünyasında en çok kullanılan programlama dilleri arasında yer almaktadır.

ADO .Net: Ado .Net, .Net *Framework* ile oluşturulan uygulamaların veri tabanı ile iletişim kurmasını sağlayan komut ve nesnelerden oluşan bir veri erişim kütüphanesidir (91). ADO .Net kütüphanesi içerisinde yer alan "*connection*" nesnesi ile veri tabanlarına bağlantı sağlanmakta, "*command*" nesneleri ile veri tabanı üzerinde silme, güncelleme, yeni kayıt oluşturma, veri getirme (*Create Read Update Delete*; CRUD) (92) gibi işlemler gerçekleştirilebilmektedir. "*dataset*" nesneleriyle ise veri tabanından gelen nesnelerin listesi tutulmaktadır. ADO.NET provider nesnesi yardımıyla ise geliştirilen uygulamaların

mevcut tüm veri tabanlarına (MS-SQL, MySQL, DB2, vs.) bağlanılarak çalışma imkanı sunulmuştur (93).

Dile Entegre Edilmiş Sorgu (LINQ): LINQ (*Language INtegrated Query*), **.**NET *Framework* 3.5 ile birlikte gelen C# ile SQL benzeri kod yazılmasını sağlayan yenilikçi kütüphanelerden biridir (94). LINQ mimarisi ile veriye erişim ve veriler üzerinde değişiklik yapmak çok daha kolay ve okunaklı hale gelmiştir.

Extension Yöntemleri: .NET Framework 3.5 ile birlikte gelen herhangi bir yeni nesne örneği oluşturmadan, sadece üzerinde işlem yapılacak olan nesne ile çağrılan statik tanımlı metotlardır (95). Kod okunurluğu artırmak ve kod tekrarını önlemek amacıyla kullanılmaktadır.

Jenerik Mimari: Veri tabanlarında her bir tabloya karşılık gelen nesneler bulunmaktadır. Jenerik mimari sayesinde her bir nesne için ayrı ayrı veri tabanı işlemleri yapmak yerine, tek bir tip nesne ile işlemlerin bir kez yapılması sağlanmakta ve uzun kod yazmak yerine kodlar kısaltılabilmektedir (96).

Entity Framework ORM (Objection Relational Mapping): ORM, ilişkisel veri tabanı ile nesne yönelimli programlama tabanlı yazılım mimarileriyle entegrasyonu sağlayan bir programlama tekniğidir (97). Örnek olarak, .Net *Framework* için *Entity Framework* ve *Dapper* verilebilir (98,99). ORM araçları yardımıyla kod geliştirme ortamlarında veri tabanındaki tablolara karşılık gelen yeni birer sınıf oluşturulmaktadır. Veri tabanı işlemleri (CRUD), bu sınıflardan türetilen nesneler yardımıyla gerçekleştirilmektedir. ORM araçları aynı zamanda herhangi bir veri tabanına ya da sorgu diline bağımlığı tamamen ortadan kaldırmaktadır. Örneğin, veri tabanı olarak MS-SQL kullanan bir uygulamada veri tabanı farklı bir platforma taşınmak istenirse (MySQL, MongoDB vs.) ORM kullanılarak oluşturulan kodlarda herhangi bir değişiklik yapmaya ihtiyaç duyulmadan taşınabilir. ORM kullanmanın dezavantajları ise oluşturulan SQL sorgularının daha kompleks olması ve bu nedenle SQL sorgularına kıyasla daha yavaş çalışmasıdır.

2.3.6. Yazılım Geliştirme Mimarileri

Katmanlı Mimari: Gelişmiş yazılımların, gerek bakımının kolaylaşması gerekse kod okunurluğunun artırılması amacıyla yazılımda yürütülen kod blokları genel olarak 3 katmana ayrılarak tasarlanmaktadır (100). Bu katmanlar sırasıyla aşağıdaki gibidir.

i) Veri Erişim Katmanı (Data Access Layer (DAL)): Veri tabanına erişim amaçlı kullanılan katmandır. Bu katmanda veri tabanında CRUD işlemlerine ilişkin kodlar bulunur. Bu katman iş katmanı ile iletişim halindedir.

ii) İş Katmanı (*Business Layer* (BL)): Bu katmanda veri erişim katmanında gelen verilerin hangi iş kurallarına göre bir sonraki katman olan sunum katmanına nasıl gönderileceği ya da yeni kayıt ekleme, güncelleme ya da silme işlemlerinin hangi iş kuralları ile yapılacağının belirlenmektedir. Bu katman diğer iki katmanla da iletişim halindedir.

iii) Sunum Katmanı (*Presentation* Layer): İş katmanından gelen işlenmiş verilerin son kullanıcıya gösterildiği ve benzer şekilde iş katmanına işlenmek üzere verilerin gönderildiği katmandır.

3. GEREÇ, YÖNTEM VE BİREYLER

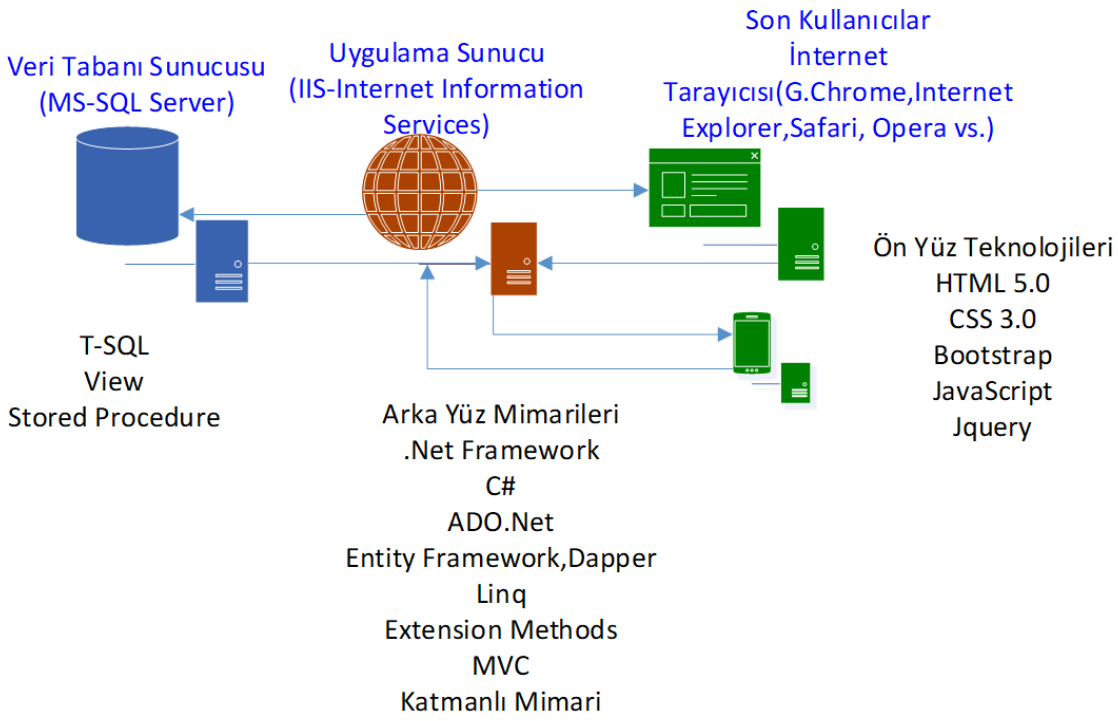
3.1. Bireyler

Bu çalışmada Hacettepe Üniversitesi Tıbbi Genetik Anabilim Dalında yerleşik olan *Ion Proton* ileri nesil dizileme sisteminden elde edilen ekzom verileri kullanılmıştır. Hastalık grupları olarak Diamond-Blackfan Anemisi (DBA, n=39), Shwachman-Diamond Sendromu (SDS, n=9) Melkersson-Rosenthal Sendromu (n=4), Fanconia Aplastik Anemisi (FAA, n=14), Mikrooti (n=1) olmak üzere farklı hastalıklara sahip toplam 67 bireyin ekzom verisi kullanılmıştır. Veri tabanında bir genetik hastalıktan şüphe duyulmayan sağlıklı bireylere ait veri bulunmamaktadır.

Çalışma protokolü Hacettepe Üniversitesi Girişimsel Olmayan Klinik Araştırmalar Etik Kurulu tarafından onaylanmıştır (Etik Kurul 31 Ocak 2018 Tarih, GO18/117-32 numaralı oluru, Bkz. EK-1).

3.2. Çalışmada Kullanılan Yöntemler

Geliştirilen uygulama, 3 katmanlı mimari olarak tasarlanmıştır. Uygulama geliştirilirken ön yüz, arka yüz ve veri tabanı geliştirmek amacıyla kullanılan teknolojiler Şekil 3.1'de özetlenmiştir.



Şekil 3.1. Çalışmada kullanılanılan 3 katmanlı mimari. Sol tarafta veri tabanı sunucusu ve veri tabanın oluşturulmasında kullanılan teknolojiler; ortada veri tabanı ile ön yüzün birbiriyle iletişimi sağlayan arka yüzde bulunan uygulama sunucusu ve bunun oluşturulmasında kullanılan teknolojiler; sağ tarafta ise ön yüzde son kullanıcının oluşturulan sistemi kullanmasını sağlayan ve kolaylaştıran teknolojiler yer almaktadır.

3.2.1. Çalışmada Kullanılan Uygulama Geliştirme Ortamları

Kod geliştirme ortamı olarak *Microsoft Visual Studio* 2017 *Enterprise*, SQL yönetim sistemi olarak ise *Microsoft* SQL *Server Management Studio* kişisel amaçlı olarak kullanılmıştır.

3.2.2. Veri tabanı İşlemleri

Veri Tabanının Oluşturulması

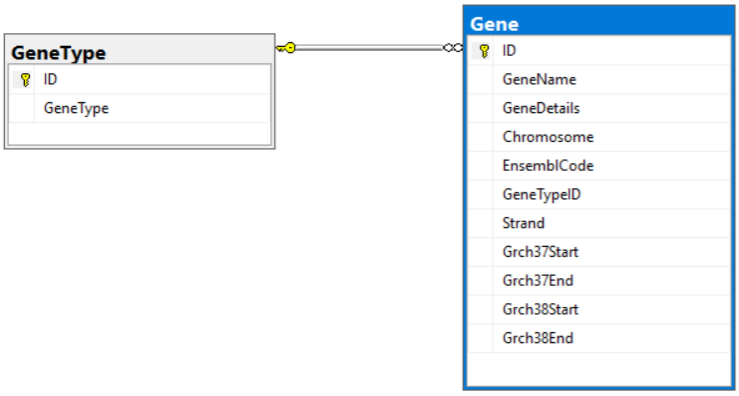
Geliştirilen olan uygulama için MS-SQL veri tabanı üzerinde "HU VariantsDB" adlı bir veri tabanı oluşturulmuştur. Bunun için, MS-SQL üzerinde *Object Explorer->Databases* menüsünden sağ tıklanarak *New Database* seçilerek açılan pencerede veri tabanının adı, dosya tiplerinin ne şekilde tutulacağı, veri tabanının nasıl genişleyeceği gibi bilgiler girilmiştir.

Tabloların Oluşturulması

Veri tabanında excel dosyalarından gelecek verilerin tutulacağı 32 ayrı tablo oluşturulmuş ve tablolar arası ilişkiler birincil anahtar (PK) ve ikincil anahatar (FK) aracılığıyla sağlanmıştır. İlişkilerin kurulmasına örnek olarak Person ve PersonLocusGenotype tablolarında yer alan PK ve FK'lar Şekil 3.2 gösterilmiştir.

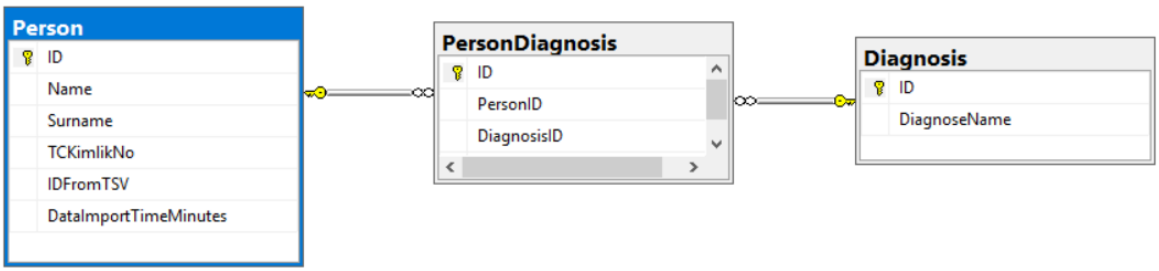
Şekil 3.2. Birincil ve ikincil anahatarlara bir örnek. **"**Person" tablosu PK alanı (Sol) ve "PersonLocusGenotype" tablosu FK alanı (Sağ)

Veri tabanında bire-çok ve çoka-çok ilişkiler sırası ile Şekil 3.3 ve Şekil 3.4'de verilmiştir. Böylelikle, bire-çok ilişki türünde bir gen sadece bir gen tipine sahip olabilirken, aynı gen tipine sahip birden fazla genin yer alabileceği gösterilmiştir. Örneğin, bir gen hem *protein coding* hem *miRNA* gen tipinde olamaz iken *protein coding* tipinde yüzlerce gen bulunmaktadır.



Şekil 3.3. **"**Gene" ve "GeneType" tabloları arasındaki bire-çok ilişki.

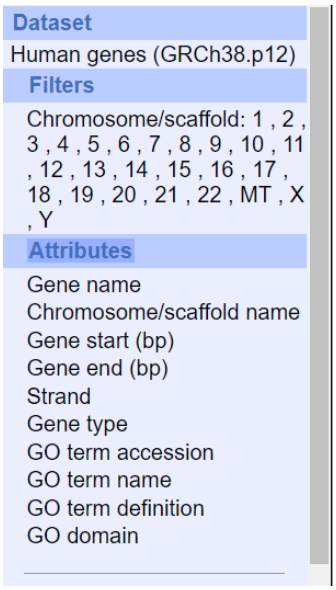
Tasarlanan veritabanında yer alan bir diğer ilişki türü ise çoka-çok ilişkidir. Bu ilişki türüne örnek olarak "Person" ve "Diagnosis" tabloları arasındaki ilişki verilebilir. Çünkü bir birey birden fazla hastalığa sahip olabilirken, aynı hastalığa sahip birden fazla birey de bulunabilmektedir (Şekil 3.3).



Şekil 3.4 "Person" ve "Diagnosis" tabloları arasındaki çoka-çok ilişki.

Gen ve Genle İlişkili Ontoloji Bilgilerinin Tablolara Aktarılması

Gen ve gene ilişkin gen ontolojisi bilgileri ensembl veri tabnından indirilmiş ve sisteme eklenmiştir (Şekil 3.5) (101).



Şekil 3.5. Ensembl filtreleme seçenekleri.

Gen Ontolojisi ve ilişkili gen bilgileri öncelikle GRCh38.p12 formatında indirilmiş ve SQL *Server Management Studio* 17 (SSMS) üzerinden "HU VariantsDB" veritabanına "*comma separated value* (csv)" formatında eklenmiştir (Bkz. EK-3). Bu işlemin ardından .csv formatındaki veriler "tmpGrch38" adında bir tabloya aktarılmıştır. "tmpGrch38" tablosundan ilişkili tablolara ("GeneType", "GODomain", "GeneOntology", "Gene" tabloları) veri aktarımını sağlayan T-SQL sorguları EK-4'de verilmiştir.

Parametre Tablolarına Verilerin Aktarılması

Ion Reporter aracılığıyla oluşturulan .tsv dosyalarında yer alan *Location*, *Type*, *VariantEffect* parametrelerinin tümü eksiksiz olarak sırasıyla "PRMLocationExplain", "PRMType", "PRMVariantEffect" tablosuna aktarılmıştır (102-104).

3.2.3. Veri aktarım uygulaması

Veri aktarım uygulaması geliştirme aşamasında IDE olarak *Visual Studio* 2017 *Enterprise* sürümü kullanılmıştır. Uygulama, *Windows Forms* uygulaması olarak

geliştirilmiştir ve C# programlama dili kullanılmıştır. Mevcut durumda gerek görülmese de, gelecekte gereği halinde uygulama *web* katmanına taşınabilmektedir.

Uygulama Katmanlarının Oluşturulması

Uygulama geliştirme ortamında boş bir çözüm (*solution*) oluşturulmuş ve ”ExcelReading” olarak adlandırılmıştır. Ardından arayüzün yer alacağı *Windows Forms* Uygulaması “ExcelReading” olarak eklenmiştir. Daha sonra ise, uygulama üzerinde veri tabanındaki tablolara karşılık gelecek sınıfların yer aldığı “Model” katmanı *ClassLibrary* projesi eklenmiştir. Yine kod tekrarını önlemek amacıyla ADO .Net sorguları için oluşturulan *Repository*, ORM adı ile *ClassLibrary* projesi olarak sisteme eklenmiştir.

Entity Framework DatabaseFirst İle Veri Tabanı Tablolarından Sınıfların Oluşturulması

EntityFramework Nuget Package Manager aracılığı ile *Solution*’da yer alan “Model” ve “ExcelReading” projelerine eklenmiştir. Ardından, *DatabaseFirst* yaklaşımıyla veri tabanındaki tablolara karşılık gelen sınıflar otomatik olarak oluşturulmuştur.

Uygulama Arayüzünün Tasarlanması

Veri aktarım uygulaması üzerinde .tsv dosyasındaki varyant bilgilerinin ait olduğu hastaya ilişkin bilgiler uygulama arayüzünde sol menüde kişi bilgileri bölümü (kişi ad, soyad ve kod bilgileri *textbox*’ları) aracılığıyla yer almaktadır. Teşhis bilgileri ise, veri tabanına daha önceden eklenmiş hastalıkların listelendiği bir *combobox* aracılığıyla kullanıcı seçimli olarak tasarlanmıştır. Hastaya ilişkin bilgiler girildikten sonra “Dosya Seç” tuşu aracılığı ile hastaya ilişkin varyantların yer aldığı .tsv uzantılı dosya seçilerek sisteme veri aktarımı başlamaktadır. Arayüzün sağ tarafında ise aktarım sonunda aktarılamayan varyant bilgileri “Eklenemeyen Alan Bilgileri” bölümünde listelenmektedir. Tasarlanan arayüz ekran görüntüsü Bölüm 4.1.3’te yer almaktadır.

Aktarım Kodlarının Geliştirilmesi

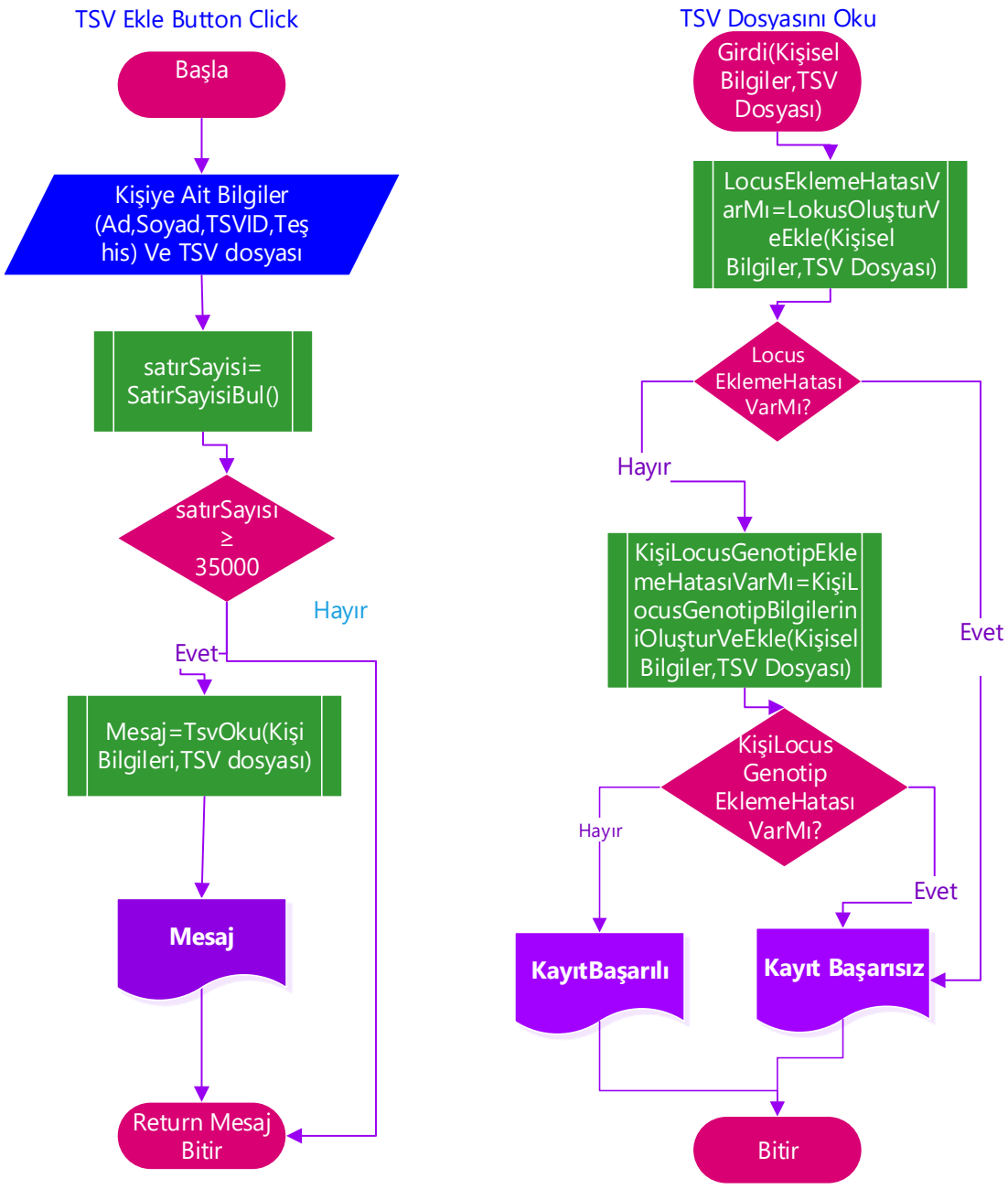
Aktarım Uygulaması geliştirilirken ORM aracı olarak *Entity Framework* ve *Dapper* kullanılmıştır. Ayrıca, bazı veri tabanı işlemleri ADO. Net aracılığıyla gerçekleştirilmiştir.

Aktarım uygulaması *solution*'ı içerisinde yer alan projeler ve bu projeler içerisinde gerçekleştirilen işlemler aşağıda sunulmuş ve tüm akış diyagramı Şekil 3.6 ve Şekil 3.7'de görselleştirilmiştir.

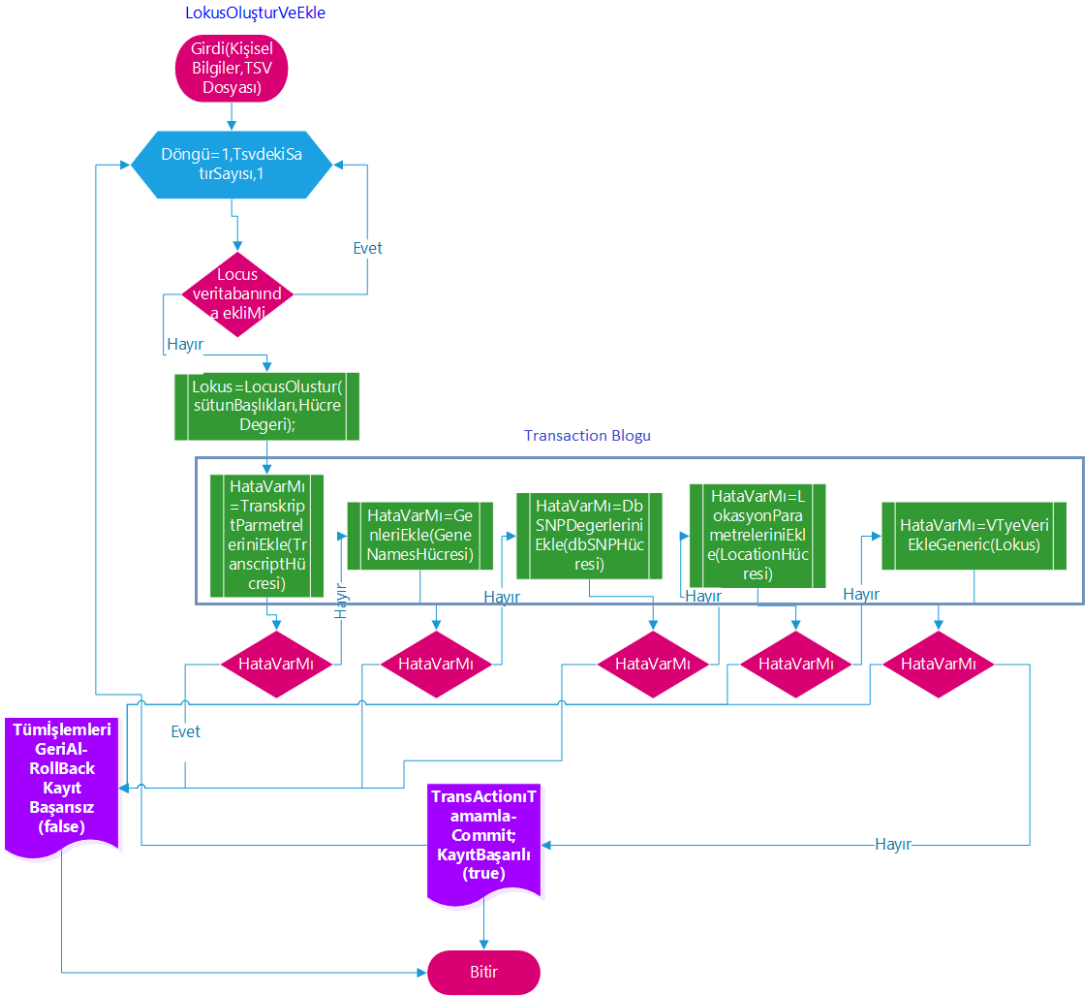
i) Model: Veri tabanı tablolarına karşılık gelen sınıfların *Entity Framework Database First* yaklaşımıyla *Class Library* projesi olarak oluşturulmuştur.

ii) ORM: *Reflection*, Jenerik mimari ve ADO.Net mimarisi kullanılarak CRUD işlemlerinin gerçekleştirildiği *Class Library* projesi olarak oluşturulmuştur.

iii) "Excel Reading": Uygulama arayüzünün bulunduğu ve bu arayüzdeki tuşlar ve diğer araçlarla kullanıcı iletişimin sağlandığı, bu sayede veri tabanına aktarım yapılan kodların yer aldığı *Windows Forms* uygulaması oluşturulmuştur.



Şekil 3.6. Veri aktarım uygulaması akış diyagramı. Uygulama arayüzünde yer alan "Dosya Seç" tuşuna tıklandığında "ExcelReading" *ClassLibrary* projesinde gerçekleştirilen işlemlere ilişkin akış diyagramı solda, diyagram içerisinde çağrılan "TSVDosyasınıOku" fonksiyonun gerçekleştirdiği işlemlere ilişkin akış diyagramı ise sağda gösterilmiştir.



Şekil 3.7. **"**TSVDosyasınıOku" fonksiyonu akış diyagramı. "TSVDosyasınıOku" fonksiyonu içerisinde kullanılan ve .tsv dosyası içerisinde yer alan *Locus*'a ait bilgilerin veri tabanında yer alan "Locus" ve ilişkili tablolara bir *transaction* bloğu içerisinde aktarılmasını sağlayan "LokusOluşturveEkle" fonksiyonuna ilişkin akış diyagramı. *Transaction* bloğu olarak gösterilen kısımda yer alan metodların herhangi birinde hata oluşursa *transaction* gereği tüm işlemler geri alınacak ve uygulamadan çıkılacaktır. Tam aksi şekilde lokusa ait tüm parametrelerin ve lokusun kendisinin aktarılması sırasında herhangi bir hata oluşmamışsa tüm veriler ilişkileriyle birlikte veri tabanına aktarılmış olacaklardır. Veri tabanında kişiye özgü varyant bilgilerinin tutulduğu "PersonLocusGenotype" tablosu ve ilşikili olduğu tablolara .tsv'deki hücrelerinden aktarım yapan "KişiLocusGenotipBilgileriniOluşturveEkle" isimli fonksiyonda benzer şekilde bir akış diyagramına sahiptir.

Veri Aktarımında Kullanılan Diğer Parametreler

Veri tabanına kaynak oluşturan .tsv'de bulunmayan; ancak veri tabanına aktarma yaparken veya yaptıktan sonra hesaplanan bazı parametreler bulunmaktadır. Bu parametreler şu şekildedir:

i) MinMAF: Her bir varyant için *Ion Reporter* yazılımındaki MAF, AMAF, EMAF, GMAF alanlarında en küçük olan MAF değerini belirtir. Eğer burada karşılığında MAF değeri olmayan bir değer varsa MinMAF değeri 0'dır.

ii) MaxMAF: Her bir varyant için *Ion Reporter* yazılımındaki MAF, AMAF, EMAF, GMAF alanlarında en büyük olan MAF değerini belirtir. Eğer burada her 4 alanın karşılığında da bir MAF değeri yoksa MaxMAF değeri 0'dır.

iii) HUMAF: Her bir varyant için veri tabanındaki varyantların veri tabanı içindeki MAF değerini belirtmek amaçlı olarak tez kapsamında oluşturulmuştur. Aşağıda görüldüğü şekilde hesaplanmaktadır (Formül 3.1):

$$\frac{\textit{(Varyantın homozigot olduğu birey sayısı)+(Varyantın heterozigot olduğu birey sayısı/2)}}{\textit{(Veri tabanındaki toplam birey sayısı)}} \qquad \textit{(Formül 3.1.)}$$

iv) Tek Yönlü Okuma: Her bir varyant için *Ion Reporter* yazılımındaki "Ref+/Ref-/Var+/Var-" sütununda yer alan okumalardan varyant verisine ait, hem artan nükleotid sayıları hem de azalan nükleotid sayıları yönünde çift yönlü okuma olup olmadığını kontrol etmek amaçlı olarak tez kapsamında oluşturulmuştur. Eğer Var+ veya Var- değerlerinden biri 0 ise tek yönlü okuma değeri veri tabanında *false* (0) olarak belirtilmiştir.

v) Genotip≠Gözlenen Hatası: Her bir varyant için veri tabanında "Genotype" (Genotip) sütunundaki sonuç "Observed" (Gözlenen) sütunundaki sonucu karşılamalıdır. Örneğin, "Genotype" sütununda G/GC şeklinde bir sonuç varsa bu referans değerin "G" olduğunu ve varyantın "GC" olduğunu göstermektedir. Bu durumda, veri tabanında "Observed" sütununda "GC" yer almaması durumunda "Genotip≠Gözlenen Hatası" vardır ve bu durum *Ion*

Reporter yazılımı tarafından yapılan bir yanlış raporlamayı göstermektedir. Her bir varyant için "Genotype" ve "Observed" alanlarına uyumluluğuna bakılarak "IsGenotypeEqGenotype" alanı tabanında *true* (1) ya da *false* (0) olarak belirlenmektedir.

vi) MinHomopolimer: Her bir varyantın bir homopolimer içinde veya komşuluğunda yer alıp almadığı *Ion Reporter* yazılımı tarafından bir veya birden fazla sayı değeri ile ifade edilmektedir. Minhomopolimer değeri Homopolimer sütununda birden fazla değer yer aldığında bunlardan en küçük olanını götermektedir.

3.2.4. *Web* Uygulamasının Geliştirilmesi

Web uygulaması geliştirilirken .Net MVC *Framework*, *Entity Framework*, *Bootstrap*, HTML 5.0, *JavaScript*, *Jquery* kütüphaneleri kullanılmıştır. Ayrıca, veri tabanı işlemleri için *Entity Framework* kullanılarak yazılmış *generic repository*, arayüz için ise geliştirilen HTML *Helper*'lar kullanılmıştır. *Web* Uygulaması *Solution*'ı içerisinde yer alan projeler ve projeler içerisinde gerçekleştirilen işlemler aşağıdaki gibidir.

i) "HU.Bioinformatics.Model": Veri tabanı tablolarına karşılık gelen sınıfların *Entity Framework* aracılığıyla oluşturulduğu *ClassLibrary* projesidir.
ii) "HU.Bioinformatics.Service": Veri tabanına bağlantı kodlarının yer aldığı *generic repository* aracılığıyla veri tabanı CRUD işlemlerinin gerçekleştirildiği *ClassLibrary* projesidir.
iii) "HU.Bioinformatics.Type": Arayüzde gösterilecek "ViewModel" nesnelerine ait sınıfların ve "Enum"ların (Enumaratör) yer aldığı *ClassLibrary* projesidir.
iv) "HU.Bioinformatics.Web": Arayüzün yer aldığı .Net MVC projesidir.

Varyant Filtreleme İşlemleri İçin İlgili View'ın Oluşturulması

İlişkili tablolar üzerinde yer alan verilerin toplu halde son kullanıcıya gösterilmesi için SQL *join* komutları ile gerekli tüm tablolar birbirine bağlanmıştır. Ardından bu "*join* sorgusu" "V_PersonLocusDetailExceptPerson" ismi ile *view* olarak veri tabanına kaydedilmiştir. Bu *view* .tsv uzantılı dosyalarda yer alan tüm verileri içermektedir. Bundan sonraki süreçte tüm varyant filtrelemeleri bu *view* üzerinde gerçekleştirilecek olan SQL sorguları aracılığı ile sağlanmıştır.

Varyant Filtreleme Modları İçin SQL Sorgularının Oluşturulması

Uygulama içerisinde "Gevşek Homozigot", "Gevşek Heterozigot", "Gevşek Bileşik Heterozigot", "Orta Homozigot", "Orta Heterozigot", "Orta Bileşik Heterozigot", "Sıkı Homozigot", "Sıkı Heterozigot", "Sıkı Bileşik Heterozigot" olmak üzere 9 farklı varyant filtreleme modu oluşturulmuştur. Her bir filtreleme için farklı parametreler kullanılmıştır. Bir filtreleme modu sırasıyla şu işlemleri yapmaktadır:

i) Filtre uygulanacak bireye ilişkin ID belirlenir.
ii) Hastaya ilişkin hastalık ID'si belirlenir.
iii) Filtreleme moduna göre eleme süreci başlar.
iv) Bileşik heterozigot filtresi durumunda aynı genden iki varyantın filtrelemeyi geçmesi durumu aranır. Bu sırada, varyant filtrelemesi için girilen *Location* ve *Variant Effect* dahil parametrelere uyan bireyin heterozigot varyantları seçilir. Bu aşama homozigot ve heterozigot varyant filtrelemede yoktur.
v) İlgili hastada ve kurumsal veri tabanında yer alan diğer hastalık ID'lerinde en az bir kez yer alan varyantlar kalıtım modeline göre elenir.
    - Homozigot: Filtrelemeye uyan başka teşhisli diğer bireylerdeki homozigot varyantlar elenir.
    - Heterozigot: Filtrelemeye uyan başka teşhisli diğer bireylerdeki heterozigot ve homozigot varyantlar elenir.

- Bileşik Heterozigot: Filtrelemeye uyan başka teşhisli diğer bireylerdeki homozigot varyantlar elenir. Bu aşama bileşik heterozigot filtresi için bir önceki aşama ile beraber yapılır.

vi) Kalan varyantlar arasından Location ve Variant Effect bilgilerine bakılarak 3 farklı sıkılıkta filtre modu çalıştırılır:

- Gevşek: Bu modun amacı varyantın etkisine göre bir filtreleme yapmadan sadece veri tabanındaki diğer varyantlarla kıyaslayarak filtreleme yapmaktadır. Gen içi ve dışı tüm varyantları dahil eder. *Location* ve *Variant Effect* bilgilerine göre filtreleme yapılmaz.
- Orta: Bu modun amacı varyantın zarar verici olma olasılığı olan 3' *splice*, 5' *splice* ve sessiz olmayan protein kodlayan ekzonik varyantlar için filtreleme yapmaktadır. *Location* bilgisine göre "*splicesite_5*, splicesite_3" ya da *Variant Effect* bilgisine göre "*missense*, *nonfameshiftInsertion*, *nonframeshiftDeletion*, *nonframeshiftBlockSubstitution*, *nonsense*, *stoploss*, *frameshiftInsertion*, *frameshiftDeletion*, *frameshiftBlockSubstitution*" olan varyantlar listelenir.
- Sıkı: Bu modun amacı varyantın zarar verici olma olasılığı yüksek olan 3' *splice*, 5' *splice* ve protein dizisinde birden fazla amino asit değişimine sebep olan varyantlar için filtreleme yapmaktadır. *Location* bilgisine göre "*splicesite_5*, splicesite_3" ya da *Variant Effect* bilgisine göre "*nonsense*, *stoploss*, *frameshiftInsertion*, *frameshiftDeletion*, *frameshiftBlockSubstitution*" olan varyantlar listelenir.

vii) Kalan varyantlar arasından sadece o kalıtım modeline uygun olan homozigot veya heterozigot varyantlar seçilir.

viii) Varyant filtrelemesi için girilen diğer parametrelere uyan varyantlar seçilir.

Web Arayüzü İle Varyant Filtrelemelerinin Gerçekleştirilmesi

Filtre Sayfası İçin *View*'ın Oluşturulması: Son kullanıcının göreceği ögeleri içerecek olan view sayfasında HTML etiketleri, MVC *web helper*'ları, *jquery* ve *bootstrap* kütüphanelerinden yararlanılmıştır.

Tasarlanan sayfada mobil telefonlar gibi küçük cihazlarda filtre ve tablo ekranı tam olarak kaplarken, büyük cihazlarda ekranın sağ kısmında filtreler (ekranın 1/4'ü) solda ise filtre sonucu dönen verilerin yer alacağı tablo (ekranın 3/4'ü) yer almıştır. Filtre menüsü içerisinde teşhis bilgileri, ilgili teşhise ilişkin hasta bilgileri, MinMAF ve MaxMAF değerlerinin girilebileceği alanlar yer almıştır. Aynı zamanda, filtre sonucu dönen değerlerin *MS-Excel* formatında çıktısının alınması sağlanmıştır. *View* içerisinde detay tuşu tıklanıldığında *Jquery* kütüphanesi ile yazılan *script* kod ile detay sayfasına yönlendirilme yapılmıştır.

Filtre Sayfası İçin İlgili *Controller*'ın Oluşturulması: *Web* arayüzünde filtreleme seçenekleri menüsünden filtreler seçilerek *controller* sınıfı içerisindeki kod bloğunda ilgili servis katmanındaki ilgili metod yardımıyla dönen değerler liste aracılığıyla *view*'a aktarılmıştır. "Varyant Filtresi" sayfası da benzer şekilde oluşturulmuştur.

Stored Procedure'lere Servis Katmanından *Dapper* ORM'i ile Erişim: Filtreleme için oluşturulan *Stored Procedure*'lere servis katmanı üzerinde EK-5'te gösterilen fonksiyon aracılığı ile erişilmiştir.

Uygulamada Kullanılan Bazı Extension Metotlar

Veri tabanına aktarılan kromozom bilgileri sayısal olarak tutulmuştur. Bu yüzden X kromozomuna sayısal değer olarak 23, Y kromozomuna ise 24 değeri verilmiştir. Ancak son kullanıcıların görmek isteyeceği ve filtreleme aşamasında ilgili değer olarak X ya da Y değerini kullanacağı düşünüldüğünden X ve Y değerlerinin nümerik olarak 23 ve 24'e dönüştürülmesini sağlayan bir *extension* metodu geliştirilmiştir.

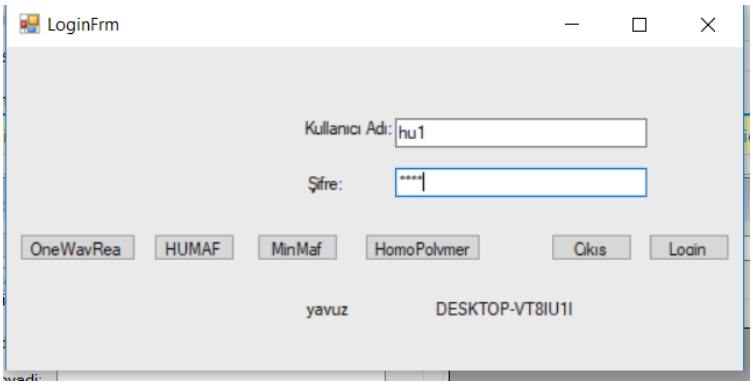
4. BULGULAR

Bu çalışmada anote edilmiş ekzom verisinden filtreleme yapmanın her türlü kullanıcı tarafından kolayca sağlanabilmesi için bir veri tabanı; veri tabanına yeni bir kişinin ekzom verisini eklemek için bir masaüstü uygulaması ve veri tabanının kolay kullanımı için bir *web* arayüzü olmak üzere iki farklı yazılım geliştirilmiştir. Bulgular, bu yazılımların genel özellikleri, performansı ve farklı senaryolarda varyant filtrelemeye yönelik uygulamalar olarak sunulmaktadır.

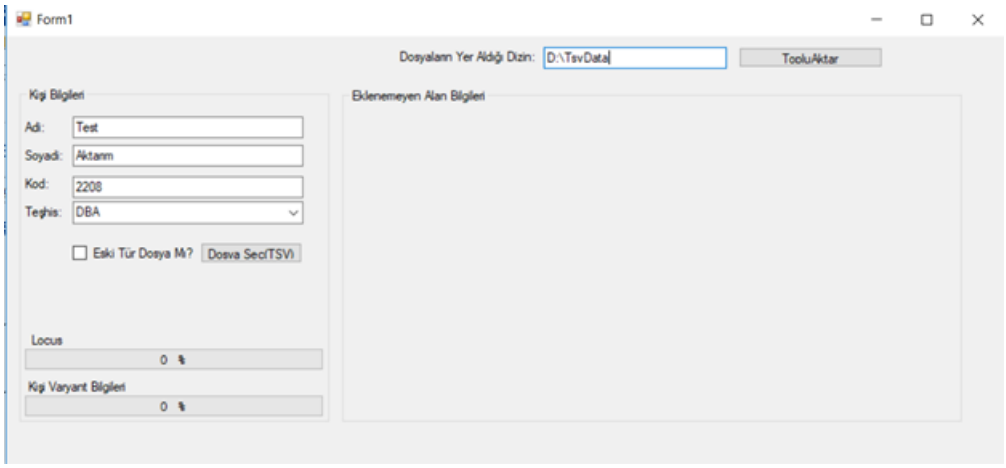
4.1. Yazılımların Genel Özellikleri

4.1.1 Veri Yüklenmesi için Geliştirilen Masaüstü Uygulaması

Veri yüklenmesi için geliştirilen masaüstü uygulaması verilerin teker teker ya da bir klasör içindeki verilerin toplu olarak yüklenmesini sağlamaktadır. Uygulama, *Ion Reporter* yazılımının 5.10 versiyonundan elde edilen anote edilmiş .tsv formatındaki ekzom verilerini bir veri tabanına entegre etmeyi sağlamaktadır. Masaüstü uygulaması, bu entegrasyon sırasında .tsv dosyası içerisinde bulunmayan; ancak veri tabanına eklenirken hesaplanan MinMAF, MaxMAF, HUMAF, Tek Yönlü Okuma, Genotip≠Gözlenen Hatası, MinHomopolimer verilerini de veri tabanına eklemektedir. Uygulamanın ekzom verisi eklemesi sırasında oluşturmadığı; ancak sonradan hesapladığı veriler de mevcuttur (HUMAF). Uygulama, bu verileri bir düğme aracılığı ile hesaplama ve güncelleme olanağı sunmaktadır. Masaüstü uygulamasının görüntüsü Şekil 4.1 ve Şekil 4.2 ’de görülmektedir.



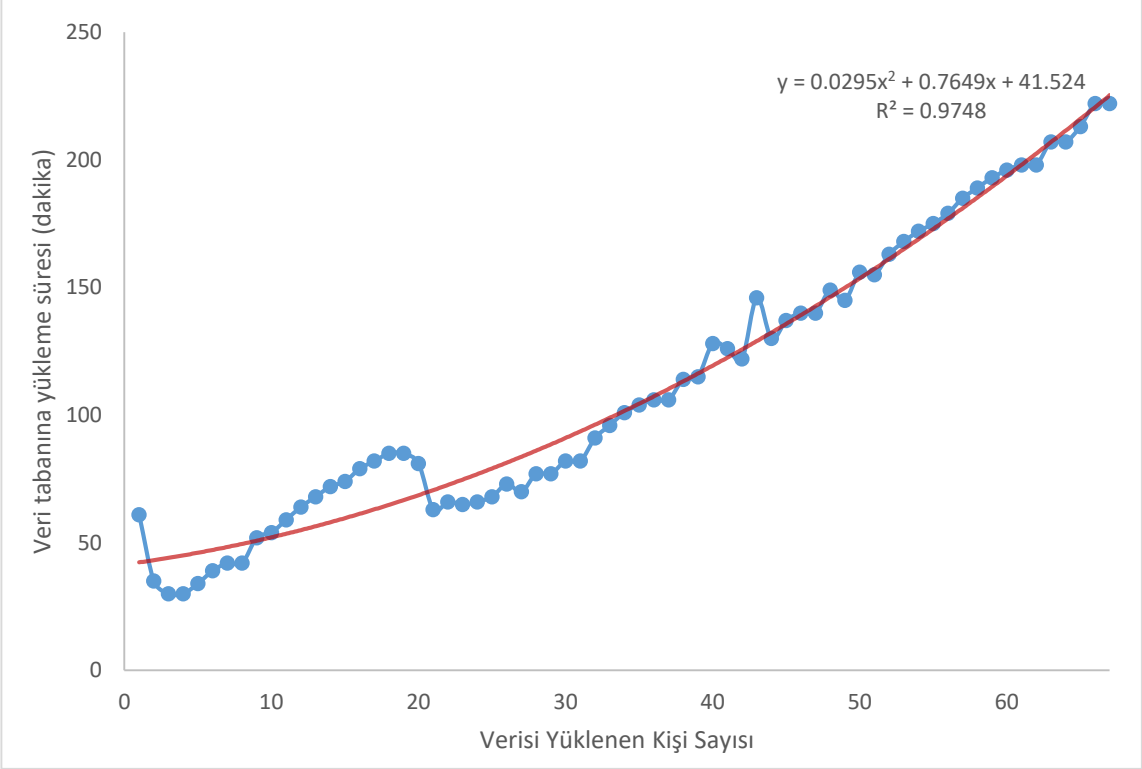
Şekil 4.1. Masaüstü uygulaması giriş ekranı. Uygulama biri giriş diğeri aktarım formu olmak üzere 2 ayrı formdan oluşmaktadır. Uygulama ilk kez çalıştırıldığında giriş ekranı ile karşılaşılmakta, kullanıcı bilgilerinin dışında herhangi bir yetkiye sahip olmadan veritabanında HUMAF, Tek Yön Okuma, MinMAF, MinHomopolimer gibi değişkenler üzerinde veri güncelleme yapılmasına olanak sağlayan tuşlar bulunmaktadır. Kullanıcı adı ve şifre alanları kullanılarak yapılan girişte kayıtlı kullanıcı ve bilgisayar aktarım yetkisine sahipse veri aktarım formu ekranı aktif hale gelecektir (Şekil 4.2).



Şekil 4.2. Masaüstü uygulama aktarım ekranı. Dosyalar (.tsv) “toplu aktar” ve “dosya seç” tuşları aracılığı ile toplu ya da tek tek aktarılabilmektedir. Aktarımın sağlıklı bir şekilde sürüp sürmediği “Lokusa ait aktarım yüzdesi” ve “Kişi Varyant aktarım yüzdesi” çubuklarından anlaşılabilmektedir. Aktarım başarılı bir şekilde gerçekleşirse “Başarılı Aktarım Mesajı” uyarı ekranda yer almakta aksi takdirde eklenemeyen alan bilgileri bölümünde aktarım sırasında karşılaşılan hatalar liste şeklinde sunulmakta ve uyarı mesajı verilmektedir.

Masaüstü uygulamasının ekzom verisi yükleme hızı ekzom verileri yüklendikçe değişmektedir. Uygulamanın veri yükleme hızını gösteren grafik Şekil 4.3'de sunulmuştur. Buna göre, veri yükleme hızı bir kişiye ait ekzom verisindeki varyantların daha önce veri tabanına kayıtlı varyant olması, toplam varyant sayısı, bilgisayarın eş zamanlı yaptığı diğer işlemler gibi bazı durumlara göre farklılık göstermektedir. Bununla beraber, uygulamanın veri yükleme hızında 2. derece bir polinomla uyumlu pozitif eğimli bir trend dikkat çekmektedir. Bu durum veri yükleme hızının hasta sayısı arttıkça giderek yavaşlayacağına işaret etmektedir (Formül 4.1).

$$(y=0{,}0295x^2+0{,}7649x+41{,}524\ ;\ r^2=0{,}9748) \qquad \text{(Formül 4.1.)}$$

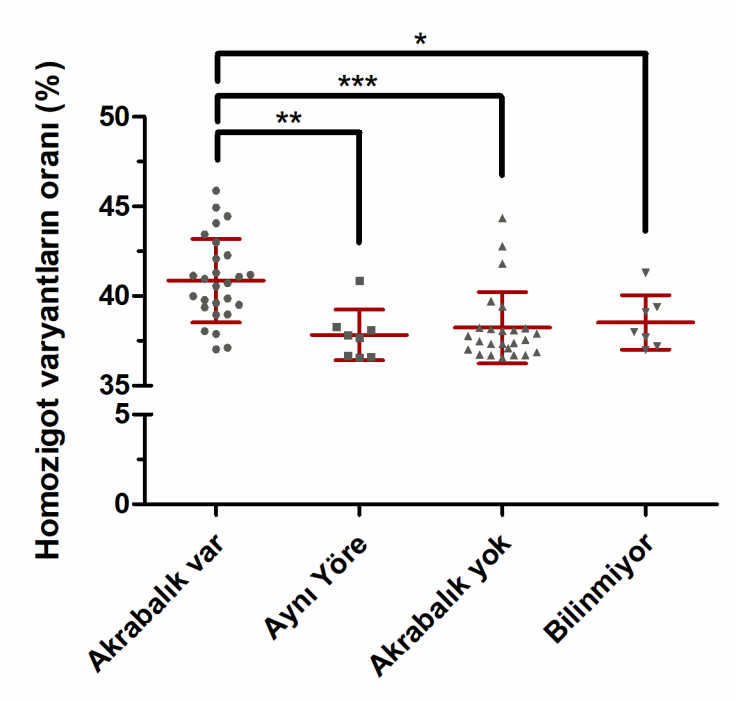


Şekil 4.3. Masaüstü uygulamasının veri yükleme hızı. Mavi işaretli noktalar her bir hasta verisinin yüklenme hızını göstermekte olup kırmızı çizgi bu veriye uyan 2. derece polinomu göstermektedir.

4.1.2 Veri tabanı

Veri tabanındaki Ekzom Verisinin Genel Özellikleri

Veri tabanı, 67 bireye ait (38 kadın, 29 erkek) *Ion Proton* sisteminden elde edilmiş ekzom verisi içermektedir. Her bireye ait ortalama 51.255,58 (45.513 – 53.195) varyant bulunmaktadır. Veri tabanındaki toplam özgün varyant sayısı 351.698 olup birey başına ortalama varyant sayısı 5.249,22'dir. Bu varyantlar toplam 257.263 lokusta yer almakta olup lokus başına 1,37 varyant bulunmaktadır. Veri tabanındaki varyantların birey başına ortalama 20.108,57 tane (17.773-22.971) homozigot; birey başına ortalama 31.147,01 tane (26.918-33.649) heterozigot varyant bulunmaktadır. Veri tabanındaki kayıtlı olan bireylerin 27'sinde ebeveynleri arası tarif edilebilen bir akrabalık bulunması sebebi ile homozigot varyant oranında bir fazlalık görülmektedir. Bu durum, Şekil 4.4'de görüldüğü gibi, ebeveynleri arasında akrabalık olan bireylerde istatistiksel olarak anlamlı ölçüde daha fazla homozigot varyant görülmesi ile de anlaşılmaktadır.



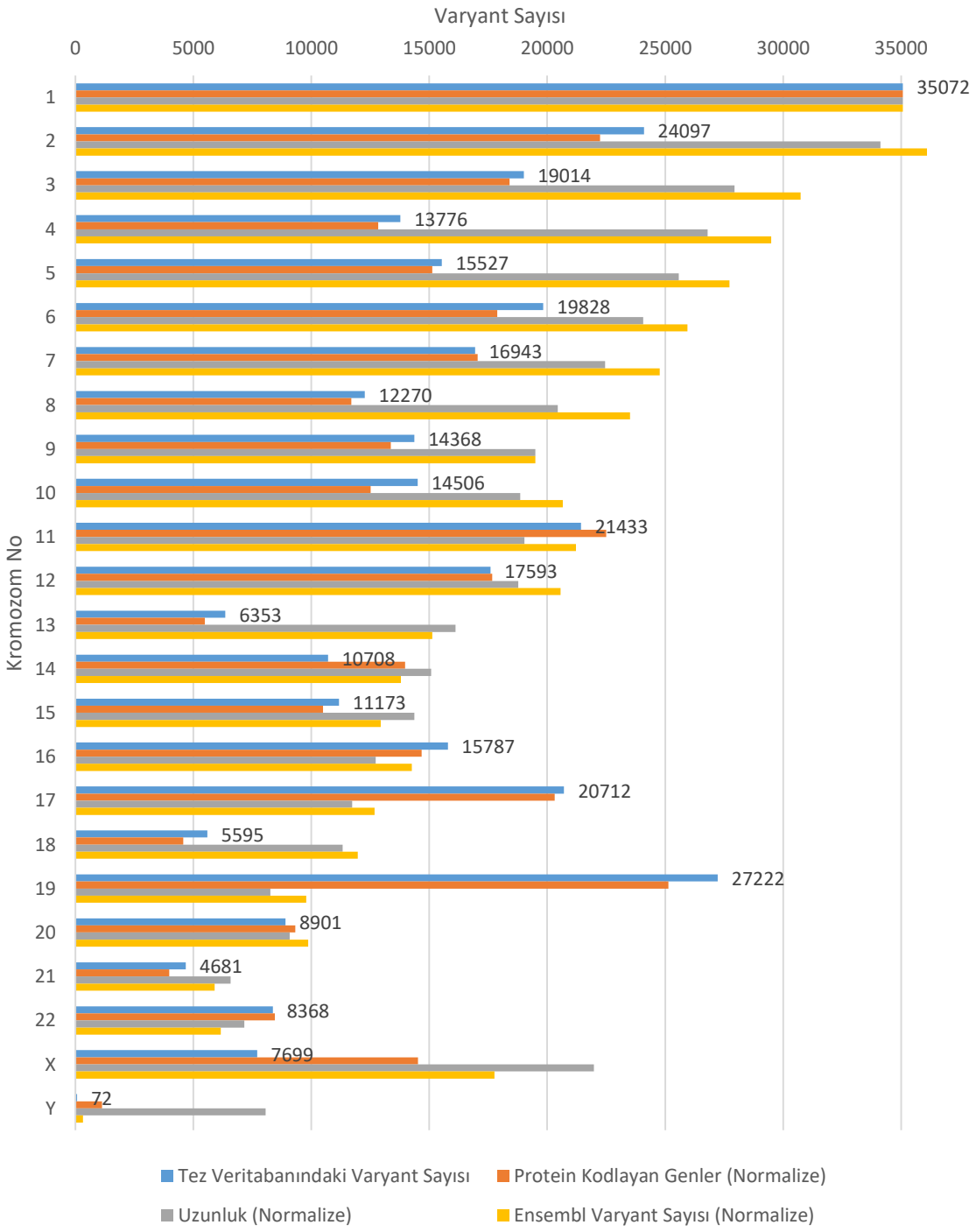
Şekil 4.4. Veri tabanında bireylerdeki homozigot varyantların oranı ile ebeveynler arasında akrabalığın ilişkisi. Ebeveynleri arasında akrabalık olan bireylerde tüm diğer gruplara kıyasla istatistiksel olarak anlamlı oranda yüksek oranda homozigot varyant bulunmaktadır. Her bir nokta bir bireyin homozigot varyant oranını göstermekte olup kırmızı çizgiler o gruptaki varyantların aritmetik ortalaması ve standart sapmasını göstermektedir. ANOVA testinin sonuçları *p<0,05; **p<0,005; ***p<0,001 olarak belirtilmiştir.

Varyantların gen içindeki yerlerine ve etkilerine bakıldığında ise, 147.140 tanesinin ekzonik ve ilginç bir şekilde çoğunluğunun ekzon dışı, özellikle intronik, bölgelere denk geldiği görülmektedir (Tablo 4.1). Nadir durumlarda bir varyant bir genin farklı *splice* varyantları sebebi ile farklı etkilere sahip olabileceğinden tablodaki toplam varyant sayıları veri tabanındaki toplam varyant sayısının üzerinde görünmektedir.

Tablo 4.1. Veri Tabanındaki Varyantların Gen Bölgelerine göre Dağılımı.

Gen bölgesi	Varyant Sayısı	%
EKZONİK	147140	39.41
Nokta Mutasyonları:		
Sinonim	66978	
Missense	75705	
Nonsense	1098	
Stop kaybı	128	
İnsersiyon/Delesyonlar:		
Çerçeve içi	1340	
Çerçeve kayması	3408	
Etkisi bilinmeyen	6396	
KODLAMAYAN EKZONİK	2808	0.75
UTR	18173	4.87
İNTRONİK		
3'-5' *splice*	1075	0.29
Diğer İntronik	189235	50.69
GENLER ARASI BÖLGE		
Gen öncesi	8019	2.15
Gen sonrası	6880	1.84

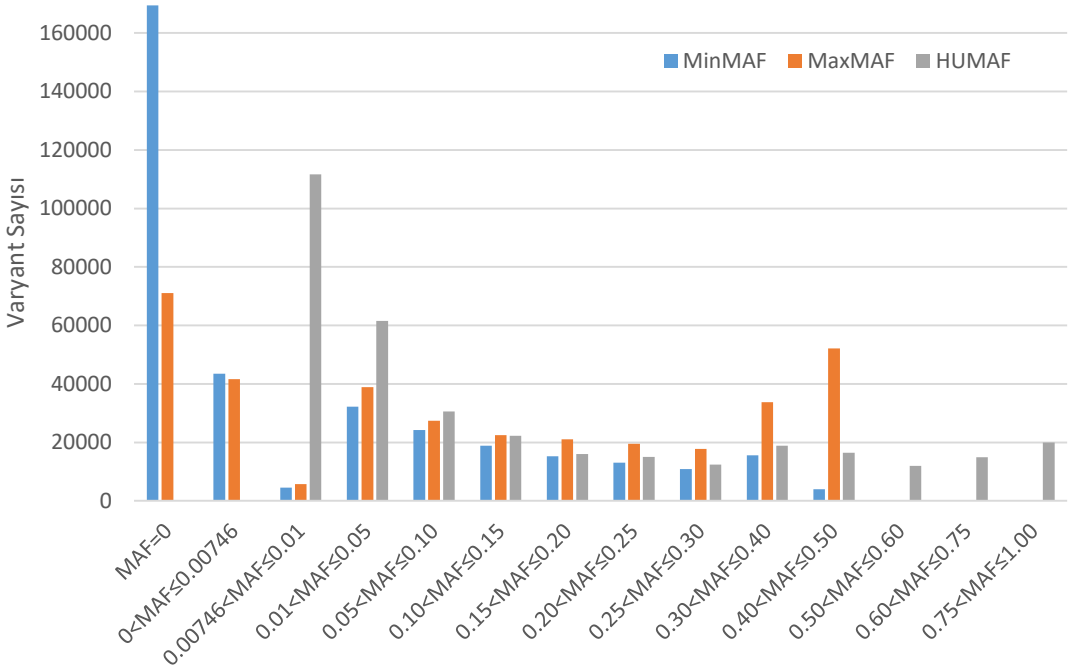
Veri tabanındaki varyantların kromozomlara göre dağılımına bakıldığında her bir kromozomdaki varyantın o kromozomda bulunan protein kodlayan gen sayıları ile yüksek korelasyon bulunduğu görülmektedir (r^2=0,943) (Şekil 4.5). Ekzom dizilemesinde özellikle protein kodlayan genler hedeflendiği için bu beklenen bir durumdur ve veri tabanındaki varyantların beklenen oranlarda dağıldığına işaret etmektedir. Buna karşın, kromozom uzunluğu ve Ensembl'da bulunan kısa varyant sayıları ile veri tabanındaki varyant sayıları arasındaki korelasyon düşüktür. Bu durum, kromozomların uzunluklarına göre farklı oranda gen içermeleri ile açıklanabilir. Gen yoğunluğu düşük 13. ve 18. Kromozomda veri tabanında az sayıda varyant bulunurken; gen yoğunluğu yüksek 17. ve 19. kromozomda veri tabanında çok sayıda varyant bulunduğu dikkati çekmektedir.



Şekil 4.5. Veri tabanındaki varyant sayılarının kromozomlara göre dağılımı. 1-22. kromozomlar ve X-Y kromozomlarında veri tabanında bulunan varyant sayılarının her bir kromozomun karşılığında belirtilmiştir. Ensembl veritabanında bulunan protein kodlayan gen, kromozom uzunluğu ve

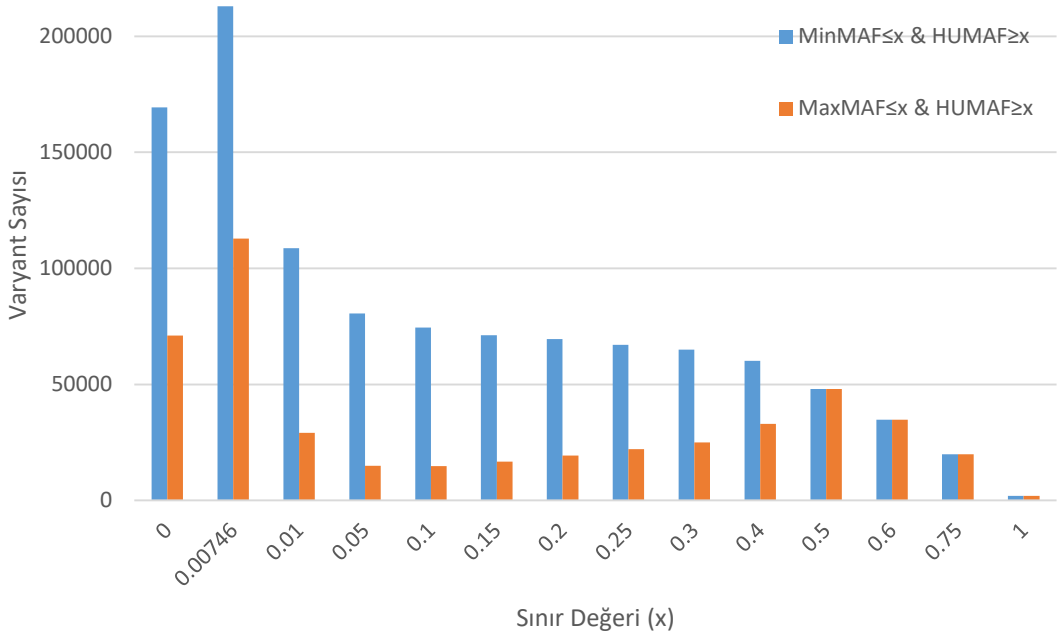
Ensembl kısa varyant sayıları, en uzun kromozom olan 1. kromozomdaki değerlere normalize edilerek belirtilmiştir (105).

Varyatların veri tabanı oluşumu sırasında hesaplanan üç farklı MAF değeri olan MinMAF, MaxMAF ve HUMAF değerlerine göre varyantların dağılımı Şekil 4.6'da sunulmuştur. Bu değerlerin nasıl oluşturulduğu Yöntem kısmında açıklanmıştır (Bölüm 3.2.3). Buna göre, MinMAF'a göre yapılan sınıflandırmada küçük değerlerde MaxMAF değerine göre daha fazla varyant bulunduğu görülmüştür. Veri tabanında MinMAF değeri 0 olan 169.391 varyant bulunurken MaxMAF değeri 0 olan 71.090 varyant mevcuttur. HUMAF değerlerine göre yapılan hesaplamada ise mümkün olan en küçük değer bir varyantın sadece bir bireyde heterozigot olarak göründüğü durumdur. Bu değer 1/134, ondalık sayı olarak yaklaşık 0,007463'tür. Veri tabanında bu değere sahip 111.675 varyant vardır. Veri tabanında MAF değeri yüksek olan varyantlara bakıldığında ise, MinMAF değeri 0,50'nin üzerinde varyant bulunmazken; MaxMAF değerine göre 145 varyantın, HUMAF değerine göre ise 46.784 varyantın MAF değeri 0,50'i aşmaktadır. Bu durum veri tabanındaki bireyler arasında sık görülen ve bu veritabanının köken aldığı popülasyona özel çok sayıda varyant bulunduğunu ortaya koymaktadır.



Şekil 4.6. Veri tabanındaki varyantların çeşitli MAF parametrelerine göre dağılımı.

Dünya genelini yansıtan MinMAF ve MaxMAF değerleri kullanıcı tarafından belirlenen sınır değerlerden küçük olan; ancak veri tabanı popülasyonda daha sık görülen varyantlar, ek varyant filtrelemesi sağlayabildikleri için kurumsal veri işlevselliği açısından oldukça önemlidir. Buradan yola çıkarak, HUMAF değerleri sınır değerlerin üzerinde olan; ancak MinMAF ve MaxMAF değerleri altında kalan varyant sayıları Şekil 4.7'de gösterilmiştir. Buna göre, HUMAF her sınır değerde ek varyant filtrelemesi sağlamaktadır. Bunun yanında, HUMAF'ın MinMAF için sağladığı varyant filtrelemesi MaxMAF için sağladığı filtrelemenin 0,4 ve altındaki her sınır değerinde 1,82 ile 5,37 kat üzerinde olmaktadır. Bu durum, HUMAF kullanıldığında MinMAF ile filtrelenen varyant sayısının MaxMAF ile elenen varyant sayısına yaklaşmasına katkı sağlamaktadır.

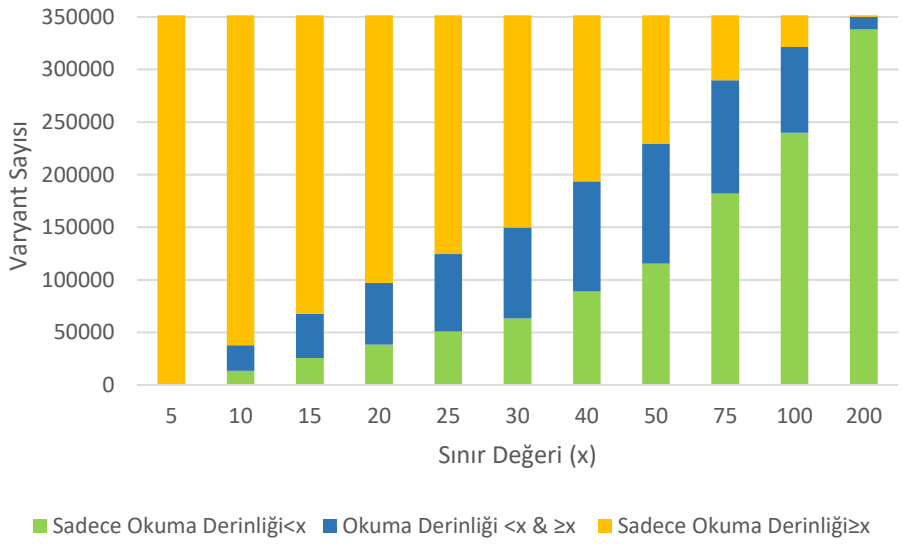


Şekil 4.7. Veri tabanında HUMAF'ın varyant filtrelemesine katkısı. Veri tabanında belirlenen bir sınır değerin altında kalan MinMAF ve MaxMAF değerine sahip varyantlar için HUMAF'ta da aynı değer sınır seçildiğinde bu değerin üzerinde kalarak filtreleme aşamalarında elenecek ek varyant sayıları.

Veri Tabanındaki Ekzom Verisinde Sistematik Olarak Taranabilecek Bazı Hatalar

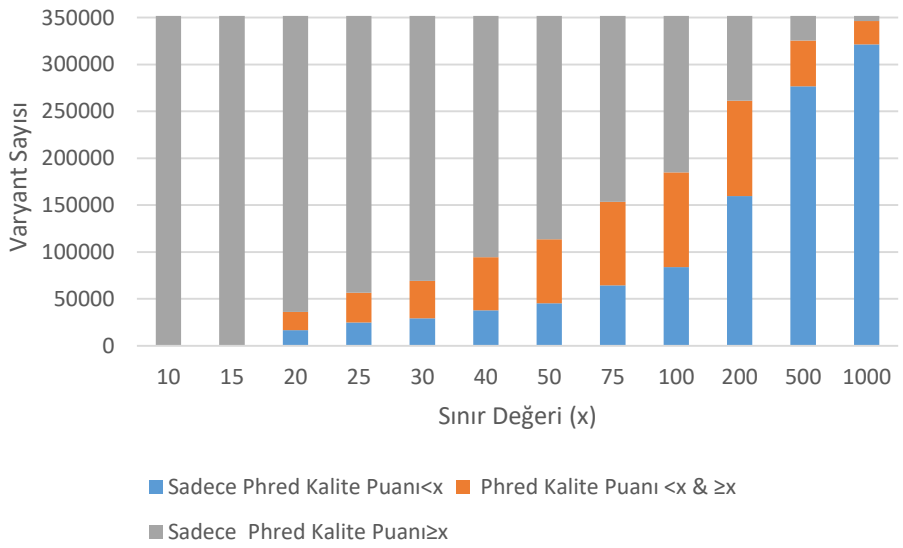
Veri tabanına kaynak oluşturan *Ion Reporter*'dan elde edilen .tsv dosyalarında bazı sistematik hatalar dikkati çekmiştir. Bu hatalar aşağıda özetlenmiş, varyant filtrelemelerinde elenebilmesi için bazı filtreleme seçenekleri oluşturulmuştur:

i) Okuma Derinliği (*Coverage*): Okuma derinliği bir varyantın kaç okumada görüldüğünü göstermektedir. Okuma derinliği düştükçe o varyantın yanlış olma olasılığı artmakta, heterozigot/homozigot ayrımı yapmak zorlaşmaktadır. Şekil 4.8'de veri tabanında yer alan varyantların okuma derinliğine göre dağılımı görülmektedir. Buna göre, veri tabanında bulunan varyantların sadece 2 tanesi 5'ten küçük okuma derinliğine sahiptir. Okuma derinliği sınırı 30 olarak kabul edildiğinde dahi varyantların %57,45'i tüm bireylerde ≥30 okuma derinliğine sahiptir.



Şekil 4.8. Veri tabanındaki varyantların okuma derinliğine göre dağılımı. Veri tabanındaki varyantların belirtilen sınır değere (x) göre dağılımlarında sarı renkli sütunlar tüm bireylerde ≥x; yeşil renkli sütunlar tüm bireylerde <x; mavi renkli sütunlar ise bazı bireylerde ≥x, bazı bireylerde <x okuma derinliğine sahip varyantları göstermektedir.

ii) *Phred* Kalite Puanı: *Phred* Kalite Puanı bir varyantın okuma verisinin güvenilirliğinin bir göstergesidir. Şekil 4.9'da veri tabanında yer alan varyantların *Phred* Kalite Puanına göre dağılımı görülmektedir. Buna göre, veri tabanında bulunan varyantların tamamı 10'dan büyük puana sahiptir. *Phred* Kalite Puanı sınırı 30 olarak kabul edildiğinde dahi varyantların %80,33'ü tüm bireylerde ≥30 *Phred* Kalite Puanı'na sahiptir.



Şekil 4.9. Veri tabanındaki varyantların *Phred* Kalite Puanı'na göre dağılımı. Veri tabanındaki varyantların belirtilen sınır değere (x) göre dağılımlarında gri renkli sütunlar tüm bireylerde ≥x; mavi renkli sütunlar tüm bireylerde <x; turuncu renkli sütunlar ise bazı bireylerde ≥x, bazı bireylerde <x *Phred* Kalite Puanı'na sahip varyantları göstermektedir.

iii) Genotip≠Gözlenen Hatası: Veri tabanında genotip ve gözlenen sütunlarında tutarsızlık olan varyantlar görülmüştür ve bunlar veri tabanından ayıklanabilmesi için işaretlenebilmesi mümkün kılınmıştır. Veri tabanında bu şekilde 3.206 varyant bulunmakta olup bunların çoğunluğu insersiyon/delesyon varyantlarıdır (Tablo 4.2).

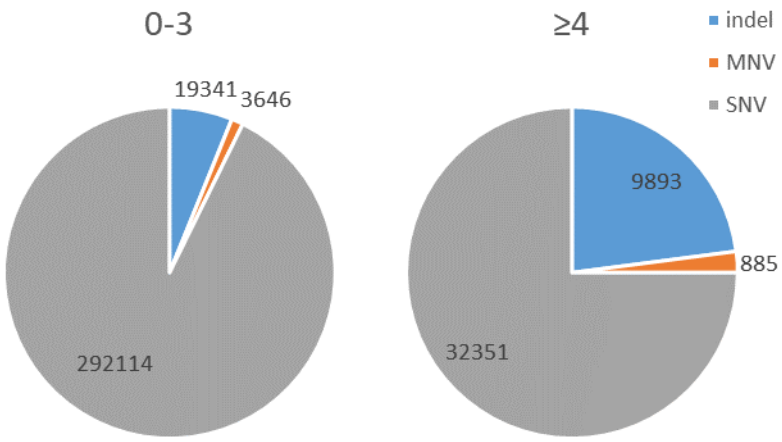
iv) Tek/Çift Yön Okumalar: Veri tabanında yer alan varyantların 28.221 tanesi sadece tek yönden okunmuş varyantlar iken; 33.934 varyant bazı bireylerde tek yönden, bazılarında çift yönden okunmuştur (Tablo 4.2). Bu varyantların dağılımına bakıldığında %86,96'sının nokta mutasyonlardan oluştuğu görülmektedir.

Tablo 4.2. Veri tabanındaki Genotip≠Gözlenen hatası olan ve Tek/Çift yönlü okunan varyantların dağılımı.

Genotip≠Gözlenen Hatası	
Hata Durumu	Varyant Sayısı
Hata yok	348492
Hata var	3206
Hatalı varyantların dağılımı	
indel:	3146
MNV:	6
SNV:	54

Tek/Çift Yön Okumalar	
Okuma yönü	Varyant Sayısı
Sadece çift yön okumalar	289543
Sadece tek yön okumalar	33934
Tek ve çift yön okumalar var	28221
Tek yön okunan varyantların dağılımı	
indel:	6527
MNV:	2198
SNV:	58174

v) Homopolimer Uzunluğu: Veri tabanında varyantların homopolimer bölgeleri ile ilşkisine bakıldığında 40.695 varyantın ≥4 homopolimer bölgesi içinde veya komşuluğunda yer aldığı görülmektedir. Bu varyantların varyant tipleri açısından dağılımı, homopolimer yakınında yer almayan varyantlar ile kıyaslandığında ≥4 homopolimer ilişkili varyantların istatistiksel olarak anlamlı ölçüde insersiyon/delesyon varyantlarından zengin olduğu görülmektedir (Ki-kare testi: p<0,0001) (Şekil 4.10). Bu da bir hata kaynağı olabilecek *Ion Proton* homopolimer hataları ile ≥4 homopolimer grubundaki insersiyon/delesyon sıklığının ilişkisini öngörmektedir.



Şekil 4.10. Homopolimer uzunluğu – varyant türü ilişkisi.

4.1.3 *Web* Arayüzü Uygulaması

Uygulama Kullanıcı Tipine Göre içerik oluşturulan dinamik sayfalardan oluşmaktadır. Uygulamaya *User* (standart kullanıcı) yetkisi ile giriş yapıldığında doğrudan Varyant Listesi sayfası ile karşılaşılacaktır. Bu sayfaya aynı zamanda *Admin* (Yönetici) yetkisiyle erişim sağlanabildiği için sayfa detaylarına aşağıda değinilecektir.

Ana sayfa

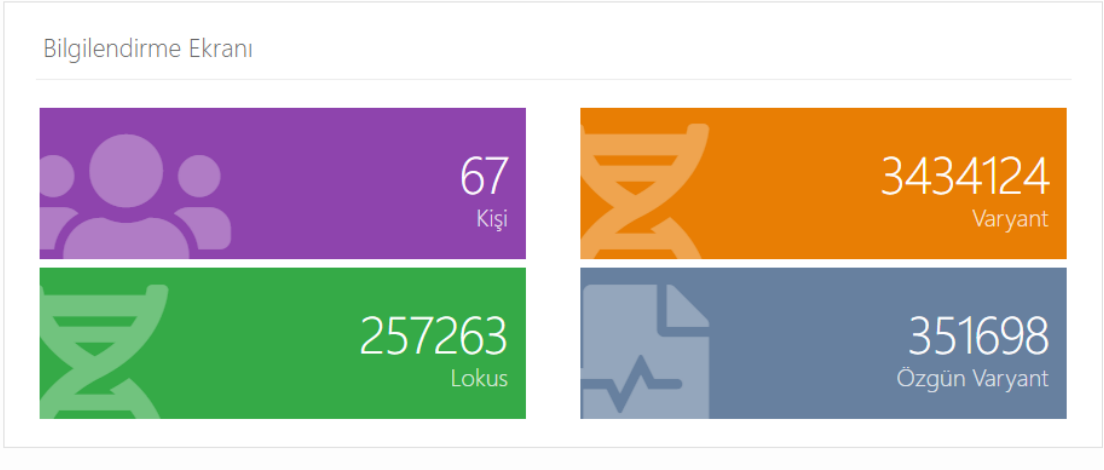
Yönetici yetkisi ile giriş yapıldığında uygulamanın karşılama ekranı üç ayrı pencereden oluşmaktadır:

i) İşlemler Ekranı: Verilere ilişkin ekleme, güncelleme, silme ve listeleme işlemlerinin yer aldığı menüler bulunmaktadır (Şekil 4.11).

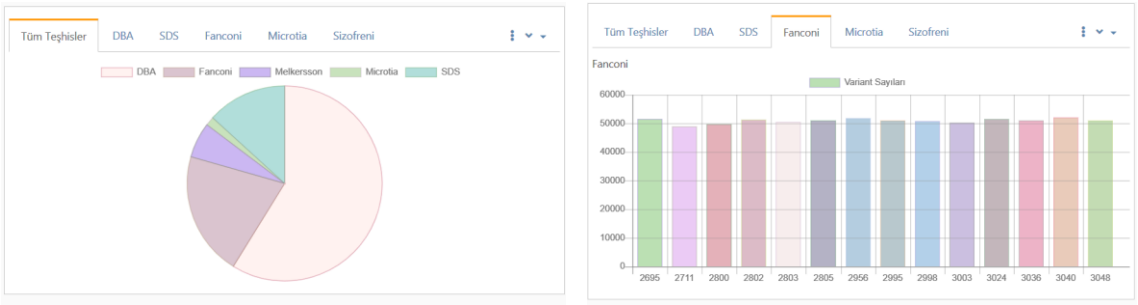
ii) Bilgilendirme Ekranı: Veri tabanında yer alan Kişi, Varyant, Lokus, Özgün Varyant bilgilerine ilişkin sayısal değerler bulunmaktadır (Şekil 4.12).

iii) İstatistik Ekranı: Veri tabanında yer alan hasta-teşhis istatistikleri pasta grafik şeklinde ve hastalıklara sahip hastaların varyant sayısı bilgileri ise çubuk grafikler şeklinde sunulmaktadır (Şekil 4.13).

Şekil 4.11. İşlemler Ekranı.



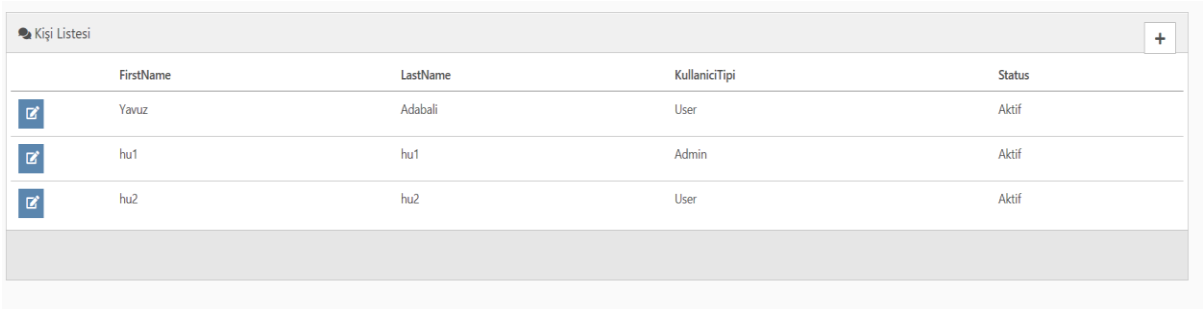
Şekil 4.12. Bilgilendirme Ekranı.



Şekil 4.13. Hastalığa özgü varyant bilgileri istatistikleri.

Kullanıcı İşlemleri

Kullanıcı işlemleri ana sayfasında veri tabanında yer alan kullanıcılara ilişkin bilgilerin listelendiği tablo bulunmaktadır. Şekil 4.14'te yer aldığı üzere, veri tabanında "Yönetici" ve "Standart Kullanıcı" olmak üzere iki tip kullanıcı bulunmaktadır. Kullanıcılar kullanıcı tiplerine göre uygulama içerisinde kısıtlamalar ile karşılaşmaktadır. Standart kullanıcı sadece varyant listesi ekranını görebilmektedir. Kullanıcı listesi üzerinde yer alan tuş ile sistemde kayıtlı kullanıcılara ilişkin güncelleme işlemleri gerçekleştirilebilmektedir. Uygulama içerisinde kullanılan listeler üzerinde yer alan güncelleme (ya da detay görüntüleme), silme ya da yeni kayıt ekleme işlemleri mümkündür.

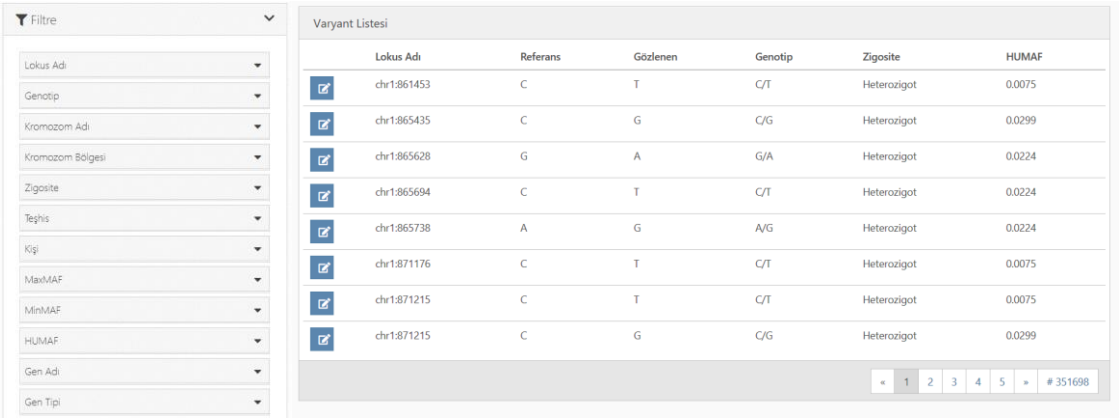


Şekil 4.14. Kullanıcı Listesi.

Varyant Listesi

Varyant Listesi karşılama sayfası veri tabanında yer alan özgün varyantların kromozom adı ve kromozom lokasyon bölgesine göre sıralanmış şekilde geldiği tablo ve ilgili parametrelerle sorgulama yapılmasına olanak sunan filtre penceresinden oluşmaktadır (Şekil 4.15). Varyant listesinin yer aldığı tablonun alt kısmında özgün varyant sayısı bilgisi yer almaktadır.

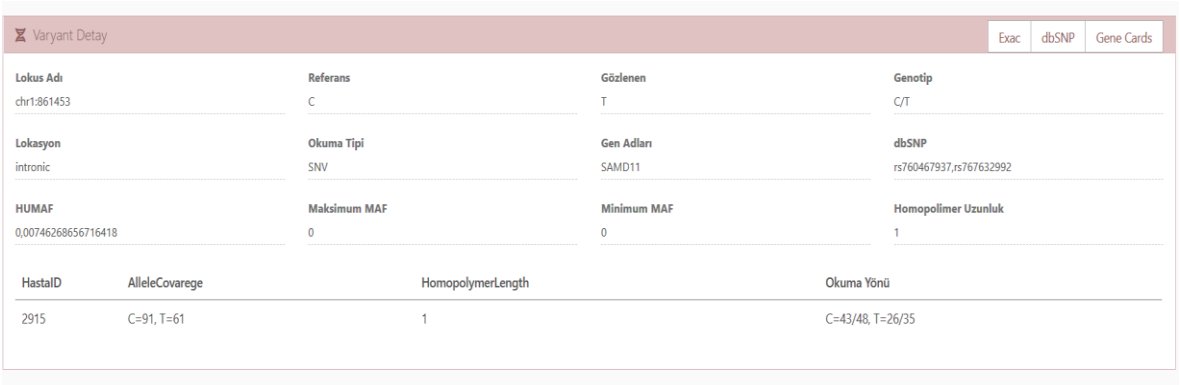
Varyant Listesi karşılama sayfasında yer alan filtre ekranı parametreleri kullanıcı tipine göre değişiklik göstermektedir. Yönetici yetkisine sahip bir kullanıcı Lokus Adı, Genotip, Kromozom Adı, Kromozom Bölgesi, Zigosite, Teşhis, Kişi, MaxMAF değeri, MinMAF değeri, HUMAF değeri, Gen Adı, Gen Tipi, Gene Ontology Domain Adı, Varyant Tipi, Varyant Etkisi, Ensemble ID, dbSNP, MinHomopolimer, Okuma Derinliği, *Phred* Kalite Puanı, Tek Yön Okuma Bilgisi, Genotip≠Gözlenen hatalarına ilişkin filtreleme yapabilmektedir.



Şekil 4.15. Filtre ekranı ve varyant listesi tablosu.

Varyant Detay

Varyant Listesi tablosu üzerinde bulunan tuş ile ilgili varyanta ilişkin detay bilgilerinin yer aldığı varyant detay sayfasına yönlendirilmektedir (Şekil 4.16). Varyant detay sayfası seçilen varyanta ilişkin detaylı bilgileri içermektedir. Ayrıca, sayfa içerisinde yer alan listede ilgili varyantın veri tabanında kaç kez görüldüğü bilgisi yer almaktadır. Eğer giriş yapılan kullanıcı tipi yönetici ise, hastaya ilişkin ID bilgisi de liste içerisinde bulunmaktadır. Varyant Detay sayfası üzerinde yer alan dış kaynak linkleri ile ExAC, dbSNP ve GeneCards sayfalarına ilgili varyanta ilişkin bilgilerle doğrudan erişim sağlanabilmektedir.



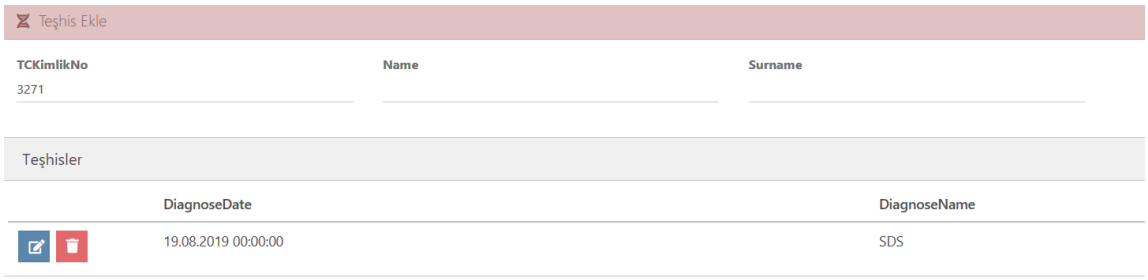
Şekil 4.16. Varyant Detay Ekranı.

Varyant Filtreleme

Karşılama sayfası seçilen filtre sonucunda dönen değerlerin listeleneceği varyant listesi tablosu ve ilgili parametrelerle sorgulama yapılmasına olanak sunan filtre penceresinden oluşmaktadır. Filtreleme penceresi yardımıyla 9 farklı filtreleme modunun (Gevşek Homozigot, Gevşek Heterozigot, Gevşek Bileşik Heterozigot, Orta Homozigot, Orta Heterozigot, Orta Bileşik Heterozigot, Sıkı Homozigot, Sıkı Heterozigot, Sıkı Bileşik Heterozigot) yer aldığı varyant filtre parametreleri ile Teşhis ve bu teşhisin konulduğu bireylerin birlikte ilişkili olarak listelendiği Teşhis – Kişi parametreleri filtreleme yapılabilmektedir. Ayrıca, MaxMAF, MinMAF, MinHomopolimer, Okuma derinliği, *Phred* Kalite Skoru, Tek Yön Okuma Bilgisi, Genotip≠Gözlenen hatalarına ilişkin seçimli filtrelemeler de yapılabilmektedir. Girilen parametrelerle yapılan sorgu sonucu dönen değerler Varyant Listesi tablosunda gösterilmektedir. Tablonun sağ üst köşesinde yer alan "Excele Gönder" tuşu aracılığıyla bu değerler excel dosyasına aktarılabilmektedir. Ayrıca, Varyant Listesi üzerinde yer alan Güncelle/Detay Görüntüle tuşu yardımıyla ilgili varyantın detaylı bilgilerine erişilebilmekte ve bu varyantın veri tabanında görüntülenme sayısına ilişkin bilgilere ulaşılabilmektedir.

Kişi Listesi

Karşılama sayfası veri tabanında yer alan hastaların ve bu hastalara ilişkin kimlik bilgilerinin yer aldığı tablodan oluşmaktadır. Tablo üzerinde yer alan güncelleme tuşu ile hastaya ilişkin verilerin güncellenebileceği "Hasta Düzenle" sayfasına yönlendirilme yapılmaktadır. "Hasta Düzenle" sayfası yardımıyla hastaya ilişkin kisişel bilgiler güncellenebilmekte, ayrıca hastaya yeni teşhis ekleme ve teşhis silme işlemleri gerçekleştirilebilmektedir (Şekil 4.17).



Şekil 4.17. Hasta düzenleme ekranı.

Teşhis Listesi

Karşılama sayfası veri tabanında yer alan hastalıkların listelendiği "Teşhis Listesi" tablosundan oluşmaktadır. Tablo üzerinde yer alan güncelleme tuşu ile ilgili satırdaki teşhis bilgileri güncellenebilmektedir. Aynı tablo üzerinde yer alan silme tuşu ile ise seçilen hastalık herhangi bir hastaya eklenmemişse silinme işlemi gerçekleştirilebilir aksi halde silme işlemi başarısız olacaktır. Yine Listenin sağ üst köşesinde yer alan ekleme tuşu ile veri tabanına yeni bir hastalık eklenebilmektedir.

4.2. Filtreleme Modlarının Değerlendirilmesi

4.2.1. Veri tabanının Varyant Filtreleme Becerisinin Değerlendirilmesi

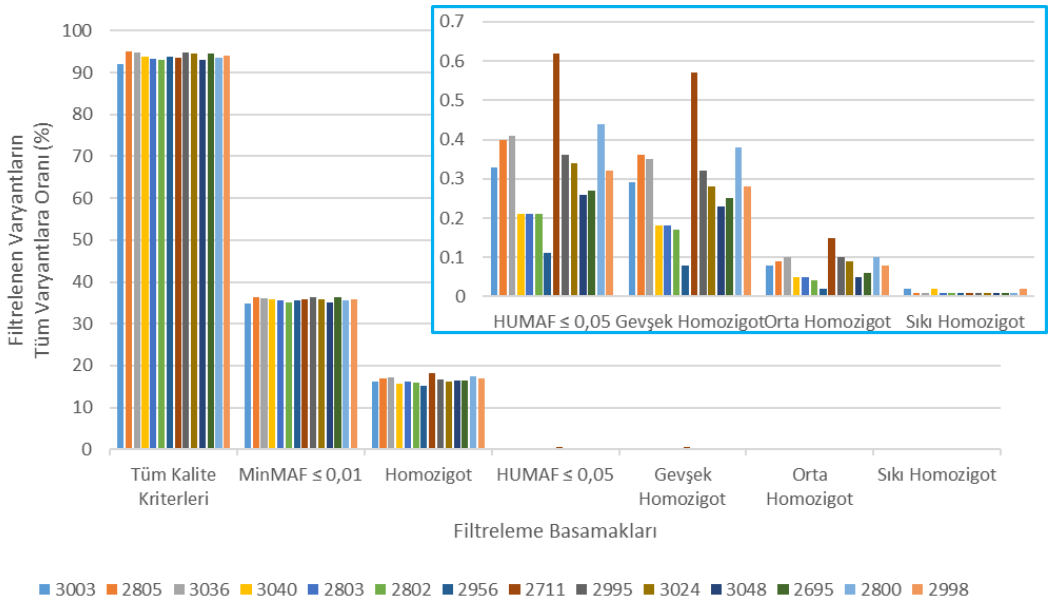
Kurulan veri tabanının varyant filtrelemedeki etkinliğini göstermek amacı ile 14 adet otozomal resesif Fanconi Aplastik Anemisi (FAA) tanılı bireye ait veri incelenmiştir (Tablo 4.3). Bu bireylerde bulunan ortalama 50.914,93 varyanttan (48.887-52143) ortalama %6,18'i çeşitli varyant kalitesi filtreleri ile elenmektedir. Bu filtrelerden "Okuma derinliği ≥ 5" filtresi en etkisiz filtre olurken, "*Phred* kalite puanı ≥ 30" filtresi ortalama %3,66 ile en büyük filtrelemeyi sağlamaktadır. Bu basamaktaki filtreler sayıca çok fazla varyant elemesi sağlamasa da elenen varyantların çoğunun gerçek olmaması sebebi ile nadirdirler ve diğer basamaklarda elenememektedirler. Bu nedenle, kalite filtreleri ile eleme basamağı önemli bir eleme aşamasıdır.

Genel olarak bakıldığında MinMAF filtrelemesinde MaxMAF filtrelemesine göre 4,99 kat daha fazla varyant ilk aşamada filtrelemeyi geçmektedir. MinMAF ile MaxMAF arasındaki fark sonraki filtreleme basamaklarında da devam etmektedir; ancak özellikle "HUMAF≤0,05" filtresi ile fark her filtreleme modunda 2 katın altına düşmektedir (Şekil 4.18 ile Şekil 4.19; Şekil 4.20 ile Şekil 4.21; Şekil 4.22 ile Şekil 4.23). Bunun nedeni, HUMAF filtresi ile MinMAF ile elenmeyen varyantların, MaxMAF ile elenmeyen varyantlara kıyasla daha çok elemeye maruz kalmalarıdır (Tablo 4.3). Tüm filtreleme basamaklarına bakıldığında kendinden bir önceki basamağa göre en etkin filtrelemeyi her filtreleme modunda veri tabanının kendi verisinden yola çıkan "HUMAF≤0,05" filtresi sağlamaktadır (%38,39-%99,28). Bu filtreleme özellikle homozigot varyantlarda (%88,43-%99,28), heterozigot varyantlara (%38,39-%77,26) göre daha etkilidir (Tablo 4.3).

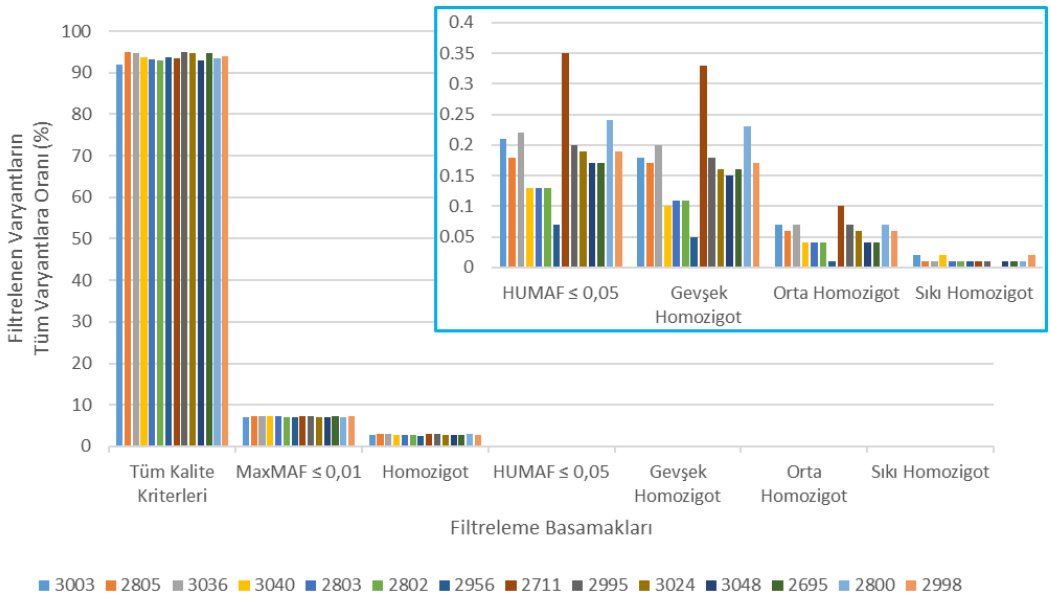
Veri tabanının kendi verisinden kaynak alan diğer filtreler de Gevşek/Orta/Sıkı ve Homozigot/Heterozigot/Bileşik Heterozigot modları olan varyant filtreleridir. Bu filtreler varyantların o tanıya sahip olmayan bireylerde görüldüğü durumda varyantın patojenik olmadığı sonucuna vararak varyantı elemektedir. Sadece bu özelliğe göre filtreleme yapan Gevşek filtreler "HUMAF≤0,05" filtresi üzerine her modda ek bir eleme sağlamaktadır. Bu durum en çok bileşik heterozigot modda; en az homozigot modda belirgindir (Şekil 4.18, Şekil 4.19 ile Şekil 4.22, Şekil 4.23). Bu durumlarda bile filtreleme modlarının ek eleme sağlaması HUMAF ile elenemeyen varyantları azaltmakta ve aday varyant listesini düşürmektedir. Üç farklı sıklıktaki filtreleme modu artan oranda varyant elemesi sağlamaktadır; ancak birey, sebebi aranan genetik hastalığın kalıtım modeli ve sık görülen mutasyon tiplerine uygun filtreleme kullanılmalıdır. Sıkı Bileşik Heterozigot filtreleme ile 14 FAA hastasından sadece 4 tanesinde 2'şer varyant saptanması dikkat çekmektedir. Bu durum, protein yapısında birden fazla amino asidin yapısını bozan 2 farklı heterozigot mutasyonun aynı gene denk gelme olasılığının düşük olması nedeniyle beklenen bir durumdur. Homopolimer filtreleri ise orta filtreler için %0-%71.43 aralığında, sıkı filtreler için ise %0-%100 aralığında bireye göre oldukça değişken miktarlarda filtreleme imkanı sunmaktadır.

Tablo 4.3. Fanconi Aplastik Anemisi teşhisli 14 bireyin varyant filtreleme basamaklarında elenen varyantları. Her bir değerin altında ilk sırada toplam varyantlardan eleme yüzdesi, ikinci sırada bir önceki filtreleme aşamasına göre eleme yüzdesi yer almaktadır. Parantez içindeki değerler ilgili değişkenin aralığını göstermektedir.

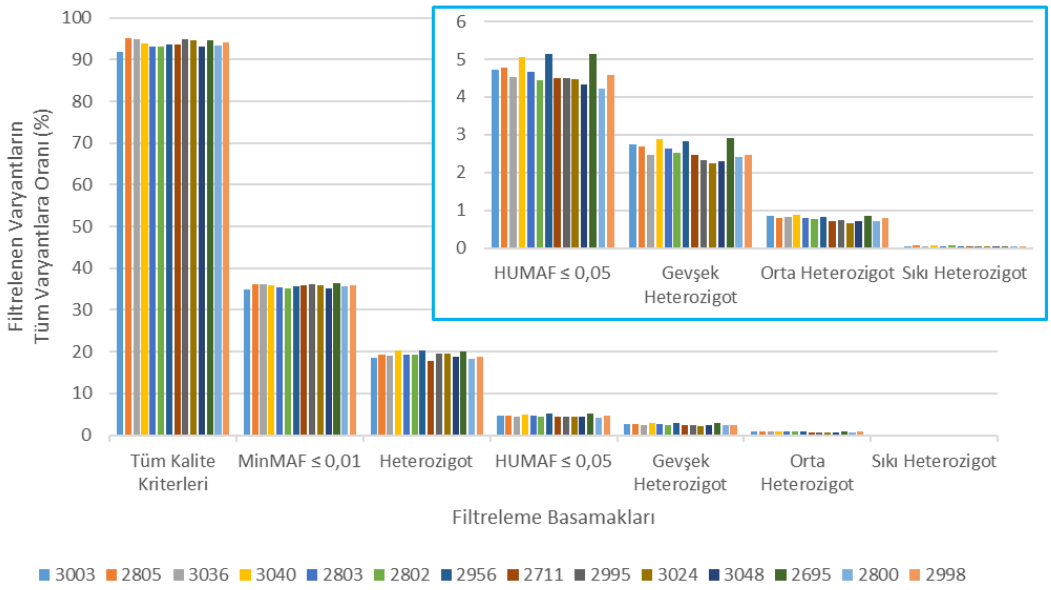
Tüm Varyantlar	**50914.93 (48877-52143)**											
Kalite Kriterleri:												
Genotip≠Gözlenen Hatası	**50667.5 (48739-51990)**											
	% 0.48 (0.24-3.12)											
Okuma derinliği ≥ 5	**50914.86 (48877-52143)**											
	% 0 (0-0)											
Phred Kalite Puanı ≥ 30	**49051.5 (46966-50410)**											
	% 3.66 (2.78-4.73)											
Çift Yönlü Okumalar	**49138.36 (47072-50183)**											
	% 3.49 (2.87-4.87)											
Tüm Kalite Kriterleri	**47768.71 (45706-48877)**											
	% 6.18 (4.98-8.07)											
MAF	MinMAF ≤ 0,01			**18203.57 (17571-18726)**				MaxMAF ≤ 0,01			**3644.57 (3463-3803)**	
				% 64.25 (63.7-65.08)							% 92.84 (92.6-93.07)	
				% 61.89 (61.56-62.21)							% 92.37 (92.09-92.6)	
Varyant Zigositesi	Hom.	**8416.79 (7910-8874)**		Het.	**9786.79 (8697-10601)**			Hom.	**1441.93 (1356-1507)**		Het.	**2224.07 (2019-2468)**
		% 83.46 (81.84-84.76)			% 80.79 (79.58-82.21)				% 97.17 (96.94-97.39)			% 95.63 (95.09-95.93)
		% 53.74 (49.5-57.27)			% 46.27 (42.73-50.5)				% 60.42 (58.3-62.64)			% 38.99 (31.22-41.7)
HUMAF ≤ 0,05		**162.57 (57-304)**			**2366.71 (2093-2665)**				**94.36 (37-173)**			**1299.36 (1155-1426)**
		% 99.68 (99.38-99.89)			% 95.36 (94.87-95.79)				% 99.82 (99.65-99.93)			% 97.45 (97.27-97.71)
		% 98.09 (96.57-99.28)			% 75.82 (74.32-77.26)				% 93.5 (88.43-97.27)			% 41.56 (38.39-47.61)
Kalıtım Modeli ve Farklı Tanıya Sahip Bireylerde Bulunmayan Varyantlar (Gevşek Filtre)	Hom.	**141.86 (42-277)**	Het.	**1304.21 (1158-1503)**	Bil. Het.	**871.57 (741-1138)**	Hom.	**83.21 (27-162)**	Het.	**1019.21 (898-1161)**	Bil. Het.	**279.57 (214-332)**
		% 99.72 (99.43-99.92)		% 97.44 (97.09-97.76)		% 98.29 (97.79-98.55)		% 99.84 (99.67-99.95)		% 98 (97.77-98.26)		% 99.45 (99.35-99.59)
		% 13.78 (7.09-26.32)		% 44.94 (41.63-49.85)		% 63.27 (56.99-66.49)		% 13.17 (5-27.03)		% 21.6 (14.85-26.4)		% 78.56 (75.79-81.91)
Nonsinonim ekzonik ve yakın *splice* varyantları (Orta Filtre)		**38.36 (10-73)**		**397.43 (343-462)**		**97.43 (73-148)**		**27.57 (7-47)**		**334.21 (274-400)**		**35.14 (23-65)**
		% 99.92 (99.85-99.98)		% 99.22 (99.11-99.34)		% 99.81 (99.72-99.86)		% 99.95 (99.9-99.99)		% 99.34 (99.23-99.47)		% 99.93 (99.88-99.95)
		% 73.24 (67.35-78.26)		% 69.5 (66.75-71.55)		% 88.82 (85.7-91.25)		% 66.97 (59.26-74.07)		% 67.24 (65.24-69.49)		% 87.49 (80.42-91.61)
Homopolimer Uzunluğu ≤ 3		**33.43 (4-64)**		**364.57 (309-421)**		**82.36 (51-126)**		**23.36 (2-40)**		**306.71 (249-368)**		**30.14 (17-60)**
		% 99.93 (99.87-99.99)		% 99.28 (99.19-99.4)		% 99.84 (99.76-99.9)		% 99.96 (99.92-100)		% 99.4 (99.29-99.52)		% 99.94 (99.88-99.97)
		% 15.76 (2.63-60)		% 8.29 (6.87-9.91)		% 15.92 (6.14-30.14)		% 18.89 (0-71.43)		% 8.26 (6.35-10.03)		% 14.96 (0-37.04)
Çerçeve kayması, *nonsense* ve yakın *splice* varyantları (Sıkı Filtre)		**5.71 (3-9)**		**26.07 (19-39)**		**0.57 (0-2)**		**5.36 (2-9)**		**23.14 (17-32)**		**0 (0-0)**
		% 99.99 (99.98-99.99)		% 99.95 (99.93-99.96)		% 100 (100-100)		% 99.99 (99.98-100)		% 99.96 (99.94-99.97)		% 100 (100-100)
		% 80.37 (30-93.75)		% 93.46 (91.56-94.95)		% 99.33 (97.26-100)		% 75.5 (14.29-93.94)		% 93.09 (90.69-94.29)		% 100 (100-100)
Homopolimer Uzunluğu ≤ 3		**2.86 (1-8)**		**19.14 (11-26)**		**0.14 (0-2)**		**2.71 (0-8)**		**17 (9-23)**		**0 (0-0)**
		% 99.99 (99.98-100)		% 99.96 (99.95-99.98)		% 100 (100-100)		% 100 (99.98-100)		% 99.97 (99.96-99.98)		% 100 (100-100)
		% 46.06 (0-85.71)		% 26.72 (14.81-42.86)		% 21.43 (0-100)		% 47.56 (0-100)		% 26.64 (5.88-47.06)		% 0 (0-0)



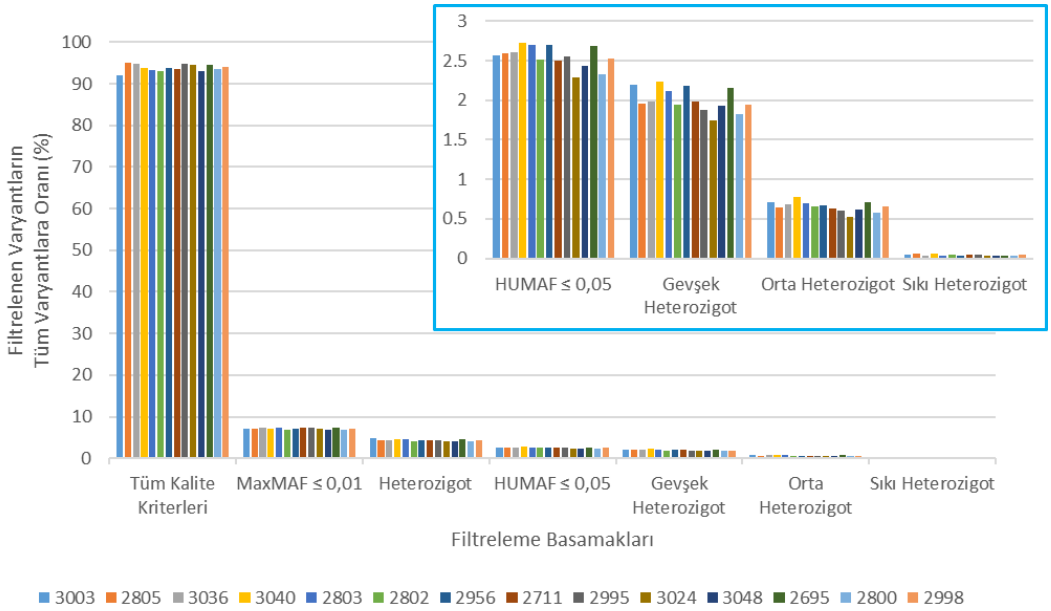
Şekil 4.18. MinMAF ve Homozigot filtrelerinde FAA tanılı bireylerde varyant filtreleme akışı. "HUMAF≤0,05"dan sonraki filtreleme basamakları sağ-üst panelde y-ekseninde daha yüksek çözünürlükle gösterilmiştir.



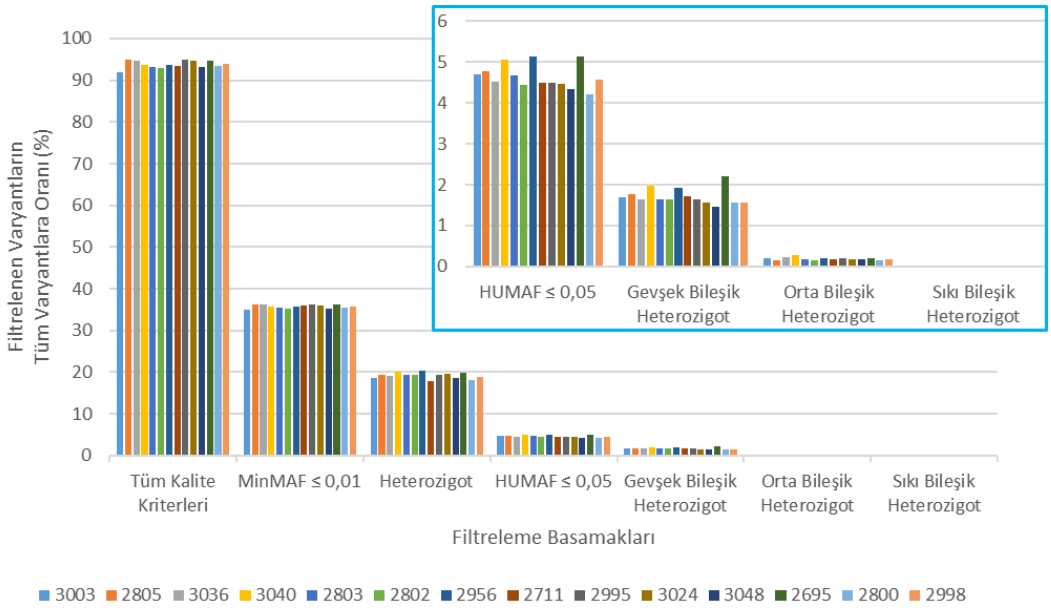
Şekil 4.19. MaxMAF ve Homozigot filtrelerinde FAA tanılı bireylerde varyant filtreleme akışı. "HUMAF≤0,05"dan sonraki filtreleme basamakları sağ-üst panelde y-ekseninde daha yüksek çözünürlükle gösterilmiştir.



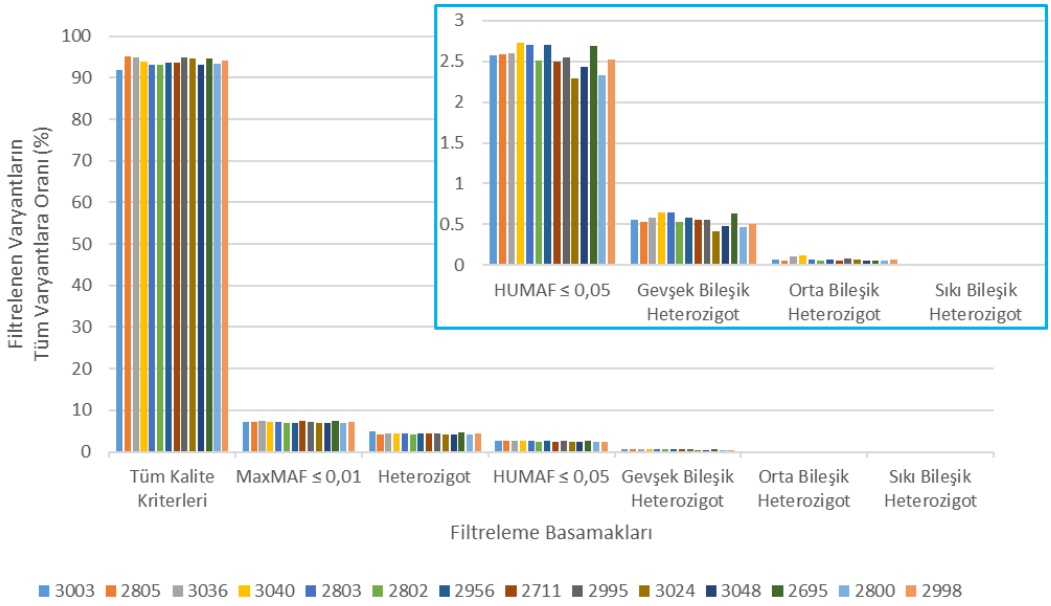
Şekil 4.20. MinMAF ve Heterozigot filtrelerinde FAA tanılı bireylerde varyant filtreleme akışı. "HUMAF≤0,05"dan sonraki filtreleme basamakları sağ-üst panelde y-ekseninde daha yüksek çözünürlükle gösterilmiştir.



Şekil 4.21. MaxMAF ve Heterozigot filtrelerinde FAA tanılı bireylerde varyant filtreleme akışı. "HUMAF≤0,05"dan sonraki filtreleme basamakları sağ-üst panelde y-ekseninde daha yüksek çözünürlükle gösterilmiştir.



Şekil 4.22. MinMAF ve Bileşik Heterozigot filtrelerinde FAA tanılı bireylerde varyant filtreleme akışı. "HUMAF≤0,05"dan sonraki filtreleme basamakları sağ-üst panelde y-ekseninde daha yüksek çözünürlükle gösterilmiştir.



Şekil 4.23. MaxMAF ve Bileşik Heterozigot filtrelerinde FAA tanılı bireylerde varyant filtreleme akışı. "HUMAF≤0,05"dan sonraki filtreleme basamakları sağ-üst panelde y-ekseninde daha yüksek çözünürlükle gösterilmiştir.

4.2.2. Farklı Filtreleme Modlarına Göre Patojenik Varyant Saptanması

Filtreleme modlarının kullanım başarısını göstermek amacı ile daha önceden patojenik varyantlara sahip olduğu belirlenmiş 4 bireyin ekzom dizileme verileri kullanılmıştır. Bu sonuçlar aşağıda sunulmuştur:

i) **2912 nolu birey:** Bu bireyin *RPS26* geninde heterozigot *nonsense* bir mutasyon bulunmaktadır (*RPS26*:c.124C>T, p.Arg42Ter). Bu mutasyonun tezde kurulan mevcut sistemde hangi filtreleme basamaklarını geçtiği Tablo 4.4'te görülmektedir. Bu mutasyon, beklendiği üzere sadece heterozigot modunda görülmüş; homozigot ve bileşik heterozigot modlarında görülmemiştir. Mutasyon, en zarar verici varyantların filtrelendiği "Sıkı" filtreleme modunda hem MinMAF hem MaxMAF filtrelemelerde filtreleme içinde kalmıştır. Bu varyant, ekzom verisindeki toplam 51.561 varyanttan başlandığında, en dar filtreleme seçeneği olarak MaxMAF, Sıkı Heterozigot ve Homopolimer filtrelerini geçen 19 varyanttan biridir. DBA tanısı ile ilişkili bir gende bulunması sebebi ile hastalık sebebi olarak bu 19 varyant arasında ön plana çıkmaktadır.

Tablo 4.4. 2912 nolu bireyin ekzom verilerinin filtrelemesi. Patojenik *RPS26*:c.124C>T (p.Arg42Ter) varyantının bulunduğu basamaklar kırmızı ile işaretlenmiştir.

Birey No:	**2912**											
Tüm Varyantlar	**51561**											
Kalite Kriterleri:												
Genotip≠Gözlenen Hatası	**51427**											
	% 0.26											
Okuma derinliği ≥ 5	**51561**											
	% 0											
Phred Kalite Puanı ≥ 30	**49420**											
	% 4.15											
Çift Yönlü Okumalar	**49704**											
	% 3.6											
Tüm Kalite Kriterleri	**48169**											
	% 6.58											
MAF		MinMAF ≤ 0,01	**18358**					MaxMAF ≤ 0,01	**3708**			
			% 64.4						% 92.81			
			% 61.89						% 92.3			
Varyant Zigositesi	Hom.	**8102**		Het.	**10256**			Hom.	**1366**		Het.	**2342**
		% 84.29			% 80.11				% 97.35			% 95.46
		% 55.87			% 44.13				% 63.16			% 36.84
HUMAF ≤ 0,05		**72**			**2433**				**46**			**1350**
		% 99.86			% 95.28				% 99.91			% 97.38
		% 99.11			% 76.28				% 96.63			% 42.36
Kalıtım Modeli ve Farklı Tanıya Sahip Bireylerde Bulunmayan Varyantlar (Gevşek Filtre)	Hom.	**56**	Het.	**1763**	Bil. Het.	**861**		Hom.	**39**	Het.	**1258**	Bil. Het. **294**
		% 99.89		% 96.58		% 98.33			% 99.92		% 97.56	% 99.43
		% 22.22		% 27.54		% 64.61			% 15.22		% 6.81	% 78.22
Nonsinonim ekzonik ve yakın *splice* varyantları (Orta Filtre)		**20**		**504**		**98**			**16**		**394**	**45**
		% 99.96		% 99.02		% 99.81			% 99.97		% 99.24	% 99.91
		% 64.29		% 71.41		% 88.62			% 58.97		% 68.68	% 84.69
Homopolimer Uzunluğu ≤ 3		**13**		**454**		**85**			**10**		**352**	**39**
		% 99.97		% 99.12		% 99.84			% 99.98		% 99.32	% 99.92
		% 35		% 9.92		% 13.27			% 37.5		% 10.66	% 13.33
Çerçeve kayması, *nonsense* ve yakın *splice* varyantları (Sıkı Filtre)		**8**		**46**		**2**			**8**		**33**	**0**
		% 99.98		% 99.91		% 100			% 99.98		% 99.94	% 100
		% 60		% 90.87		% 97.96			% 50		% 91.62	% 100
Homopolimer Uzunluğu ≤ 3		**2**		**31**		**0**			**2**		**19**	**0**
		% 100		% 99.94		% 100			% 100		% 99.96	% 100
		% 75		% 32.61		% 100			% 75		% 42.42	% 0

ii) **2683 nolu birey:** Bu bireyin *ADA2* geninde bileşik heterozigot *missense* birer mutasyonu bulunmaktadır (*ADA2*:c.1359_1360delTGinsCC, p.Asp454His ve *ADA2*:c.620T>C, p.Phe207Ser). Bu mutasyonların tezde kurulan mevcut sistemde hangi filtreleme basamaklarını geçtiği Tablo 4.5'de görülmektedir. Bu mutasyonlar, beklendiği üzere hem heterozigot hem de bileşik heterozigot modlarında görülmüş; homozigot modunda görülmemiştir. Mutasyonlar, *missense* varyantların da dahil edildiği "Orta" filtreleme modunda görülmüş ve beklendiği üzere "Sıkı" filtreleme modunda görülmemiştir. *ADA2*:c.1359_1360delTGinsCC mutasyonunun etkilediği 1359. pozisyondaki nükleotid c.1359T>C (rs752497071) sinonim bir değişikliğe sebep olmakta ve bu değişiklik için MAF>0,01'dir. *Ion Reporter* yazılımı bu MAF değerini *ADA2*:c.1359_1360delTGinsCC mutasyonuna atfettiği için MaxMAF filtrelemesinde bu mutasyon kaybolmaktadır. Ancak MinMAF filtresinde her iki mutasyon da görülmektedir. Bu varyantların her ikisi, ekzom verisindeki toplam 50.674 varyanttan başlandığında, en dar filtreleme seçeneği olarak MinMAF, Orta Bileşik Heterozigot ve Homopolimer filtrelerini geçen 71 varyant arasındadır. DBA tanısı ile ilişkili bir gende bulunması sebebi ile hastalık sebebi olarak bu 71 varyant arasında ön plana çıkmaktadırlar.

Tablo 4.5. 2683 nolu bireyin ekzom verilerinin filtrelemesi. Patojenik *ADA2*:c.1359_1360delTGinsCC (p.Asp454His) ve *ADA2*:c.620T>C (p.Phe207Ser) varyantlarının bulunduğu basamaklar kırmızı ile işaretlenmiştir. Sadece *ADA2*:c.1359_1360delTGinsCC (p.Asp454His) varyantının bulunduğu basamaklar ise sarı ile işaretlenmiştir.

Birey No:	**2683**											
Tüm Varyantlar	**50674**											
Kalite Kriterleri:												
Genotip≠Gözlenen Hatası	**50548**											
	% 0.25											
Okuma derinliği ≥ 5	**50674**											
	% 0											
Phred Kalite Puanı ≥ 30	**48344**											
	% 4.6											
Çift Yönlü Okumalar	**48798**											
	% 3.7											
Tüm Kalite Kriterleri	**47095**											
	% 7.06											
MAF	MinMAF ≤ 0,01	**17710**				MaxMAF ≤ 0,01	**3564**					
		% 65.05					% 92.97					
		% 62.4					% 92.43					
Varyant Zigositesi	Hom.	**7763**	Het.	**9947**		Hom.	**1323**	Het.	**2241**			
		% 84.68		% 80.37			% 97.39		% 95.58			
		% 56.17		% 43.83			% 62.88		% 37.12			
HUMAF ≤ 0,05		**45**		**2419**			**21**		**1341**			
		% 99.91		% 95.23			% 99.96		% 97.35			
		% 99.42		% 75.68			% 98.41		% 40.16			
Kalıtım Modeli ve Farklı Tanıya Sahip Bireylerde Bulunmayan Varyantlar (Gevşek Filtre)	Hom.	**40**	Het.	**1710**	Bil. Het. **851**	Hom.	**19**	Het.	**1203**	Bil. Het. **263**		
		% 99.92		% 96.63	% 98.32		% 99.96		% 97.63	% 99.48		
		% 11.11		% 29.31	% 64.82		% 9.52		% 10.29	% 80.39		
Nonsinonim ekzonik ve yakın *splice* varyantları (Orta Filtre)		**7**		**485**	**86**		**4**		**374**	**36**		
		% 99.99		% 99.04	% 99.83		% 99.99		% 99.26	% 99.93		
		% 82.5		% 71.64	% 89.89		% 78.95		% 68.91	% 86.31		
Homopolimer Uzunluğu ≤ 3		**5**		**432**	**71**		**3**		**331**	**27**		
		% 99.99		% 99.15	% 99.86		% 99.99		% 99.35	% 99.95		
		% 28.57		% 10.93	% 17.44		% 25		% 11.5	% 25		
Çerçeve kayması, *nonsense* ve yakın *splice* varyantları (Sıkı Filtre)		**2**		**37**	**0**		**2**		**28**	**0**		
		% 100		% 99.93	% 100		% 100		% 99.94	% 100		
		% 71.43		% 92.37	% 100		% 50		% 92.51	% 100		
Homopolimer Uzunluğu ≤ 3		**1**		**15**	**0**		**1**		**11**	**0**		
		% 100		% 99.97	% 100		% 100		% 99.98	% 100		
		% 50		% 59.46	% 0		% 50		% 60.71	% 0		

iii) **3045 nolu birey:** Bu bireyin de *ADA2* geninde homozigot çerçeve kaymasına sebep olan 2 nükleotid delesyon mutasyonu bulunmaktadır (*ADA2*:c.680_681delAT, p.Tyr227fs). Bu mutasyonun tezde kurulan mevcut sistemde hangi filtreleme basamaklarını geçtiği Tablo 4.6'da görülmektedir. Bu mutasyon, beklendiği üzere sadece homozigot modunda görülmüş; heterozigot ve bileşik heterozigot modlarında görülmemiştir. Mutasyon, en zarar verici varyantların filtrelendiği "Sıkı" filtreleme modunda hem MinMAF hem MaxMAF filtrelemelerde filtreleme içinde kalmıştır. Bu varyant, ekzom verisindeki toplam 59.631 varyanttan başlandığında, en dar filtreleme seçeneği olarak MaxMAF, Sıkı Homozigot ve Homopolimer filtrelerini geçen 3 varyanttan biridir. DBA tanısı ile ilişkili bir gende bulunması sebebi ile hastalık sebebi olarak bu 3 varyant arasında ön plana çıkmaktadır.

Tablo 4.6. 3045 nolu bireyin ekzom verilerinin filtrelemesi. Patojenik *ADA2*:c.680_681delAT (p.Tyr227fs) varyantının bulunduğu basamaklar kırmızı ile işaretlenmiştir.

Birey No:	**3045**											
Tüm Varyantlar	**49631**											
Kalite Kriterleri:												
Genotip≠Gözlenen Hatası	**49477**											
	% 0.31											
Okuma derinliği ≥ 5	**49631**											
	% 0											
Phred Kalite Puanı ≥ 30	**47859**											
	% 3.57											
Çift Yönlü Okumalar	**47856**											
	% 3.58											
Tüm Kalite Kriterleri	**46581**											
	% 6.15											
MAF		MinMAF ≤ 0,01	**17759**					MaxMAF ≤ 0,01	**3527**			
			% 64.22						% 92.89			
			% 61.88						% 92.43			
Varyant Zigositesi	Hom.	**8879**		Het.	**8880**			Hom.	**1513**		Het.	**2014**
		% 82.11			% 82.11				% 96.95			% 95.94
		% 50			% 50				% 57.1			% 42.9
HUMAF ≤ 0,05		**276**			**2194**				**154**			**1182**
		% 99.44			% 95.58				% 99.69			% 97.62
		% 96.89			% 75.29				% 89.82			% 41.31
Kalıtım Modeli ve Farklı Tanıya Sahip Bireylerde Bulunmayan Varyantlar (Gevşek Filtre)	Hom.	**273**	Het.	**1600**	Bil. Het.	**842**		Hom.	**152**	Het.	**1113**	Bil. Het. **265**
		% 99.45		% 96.78		% 98.3			% 99.69		% 97.76	% 99.47
		% 1.09		% 27.07		% 61.62			% 1.3		% 5.84	% 77.58
Nonsinonim ekzonik ve yakın *splice* varyantları (Orta Filtre)		**67**		**460**		**108**			**41**		**351**	**39**
		% 99.87		% 99.07		% 99.78			% 99.92		% 99.29	% 99.92
		% 75.46		% 71.25		% 87.17			% 73.03		% 68.46	% 85.28
Homopolimer Uzunluğu ≤ 3		**57**		**424**		**98**			**35**		**320**	**33**
		% 99.89		% 99.15		% 99.8			% 99.93		% 99.36	% 99.93
		% 14.93		% 7.83		% 9.26			% 14.63		% 8.83	% 15.38
Çerçeve kayması, *nonsense* ve yakın *splice* varyantları (Sıkı Filtre)		**10**		**41**		**0**			**8**		**34**	**0**
		% 99.98		% 99.92		% 100			% 99.98		% 99.93	% 100
		% 85.07		% 91.09		% 100			% 80.49		% 90.31	% 100
Homopolimer Uzunluğu ≤ 3		**4**		**28**		**0**			**3**		**23**	**0**
		% 99.99		% 99.94		% 100			% 99.99		% 99.95	% 100
		% 60		% 31.71		% 0			% 62.5		% 32.35	% 0

iv) **3149 nolu birey:** Bu birey, 3 farklı hastalıkla ilgili 3 farklı homozigot *missense* patojenik varyantı bir arada barındırmaktadır ve çok sayıda patojenik varyantın filtrelenmesinde sunulan filtreleme sisteminin kullanılması için iyi bir örnek teşkil etmektedir. Bu varyantlar, *SLC25A12* genindeki c.728G>A (p.Arg243Lys); *C15orf41* genindeki c.58C>A (p.Pro20Thr); ve *OCA2* genindeki c.1441G>A (p.Ala481Thr) mutasyonlarıdır. Bu mutasyonların tezde kurulan mevcut sistemde hangi filtreleme basamaklarını geçtiği Tablo 4.7'de görülmektedir. Bu mutasyonlar, beklendiği üzere sadece homozigot modunda görülmüş; heterozigot ve bileşik heterozigot modlarında görülmemiştir. Mutasyonlar, *missense* varyantların filtrelendiği "Orta" filtreleme modunda görülmüştür. *C15orf41* ve *OCA2* genlerindeki mutasyonlar hem MinMAF hem MaxMAF filtrelemelerde filtreleme içinde kalırken, *SLC25A12* genindeki mutasyon sadece MinMAF filtrelemesinde görülmüştür. *SLC25A12*'deki mutasyonun bazı toplumlardaki MAF değeri (Avrupa) 0,0166'ya kadar çıkmaktadır; ancak bu mutasyon homozigot olduğunda hastalık sebebi olduğu için Hardy-Weinberg dengesindeki bir popülasyonda homozigot görülme sıklığı 2,75/10.000'dır ve nadirdir. Bu nedenle, bazı popülasyonlarda sık görülen varyantlara hassas MaxMAF filtresinde elenmektedir. Bu varyantlar, ekzom verisindeki toplam 49.644 varyanttan başlandığında, en dar filtreleme seçeneği olarak MinMAF, Orta Homozigot ve Homopolimer filtrelerini geçen 51 varyanttan 3'üdür. DBA öntanısının yanında hastanın diğer temel bulguları olan infatil başlangıçlı persistan epilepsisi ve albinizmi açıklaması sebebi ile 3 gende bulunan mutasyonlar hastalık sebebi olarak bu 51 varyant arasında ön plana çıkmaktadır.

Tablo 4.7. 3149 nolu bireyin ekzom verilerinin filtrelemesi. Patojenik *SLC25A12*:c.728G>A (p.Arg243Lys), *C15orf41*:c.58C>A (p.Pro20Thr) ve *OCA2*:c.1441G>A (p.Ala481Thr) varyantlarının bulunduğu basamaklar kırmızı ile işaretlenmiştir. Sadece *C15orf41*:c.58C>A (p.Pro20Thr) ve *OCA2*:c.1441G>A (p.Ala481Thr) varyantlarının bulunduğu basamaklar ise sarı ile işaretlenmiştir.

Birey No:	**3149**											
Tüm Varyantlar	**49644**											
Kalite Kriterleri:												
Genotip≠Gözlenen Hatası	**49530**											
	% 0.23											
Okuma derinliği ≥ 5	**49644**											
	% 0											
Phred Kalite Puanı ≥ 30	**47500**											
	% 4.32											
Çift Yönlü Okumalar	**49611**											
	% 0.07											
Tüm Kalite Kriterleri	**46366**											
	% 6.6											
MAF	MinMAF ≤ 0,01	**17607**					MaxMAF ≤ 0,01	**3479**				
		% 64.53						% 92.99				
		% 62.03						% 92.5				
Varyant Zigositesi	Hom.	**8578**		Het.	**9029**		Hom.	**1402**		Het.	**2077**	
		% 82.72			% 81.81			% 97.18			% 95.82	
		% 51.28			% 48.72			% 59.7			% 40.3	
HUMAF ≤ 0,05		**210**			**2141**			**101**			**1184**	
		% 99.58			% 95.69			% 99.8			% 97.62	
		% 97.55			% 76.29			% 92.8			% 42.99	
Kalıtım Modeli ve Farklı Tanıya Sahip Bireylerde Bulunmayan Varyantlar (Gevşek Filtre)	Hom.	**198**	Het.	**1544**	Bil. Het.	**752**	Hom.	**100**	Het.	**1075**	Bil. Het.	**237**
		% 99.6		% 96.89		% 98.49		% 99.8		% 97.83		% 99.52
		% 5.71		% 27.88		% 64.88		% 0.99		% 9.21		% 79.98
Nonsinonim ekzonik ve yakın *splice* varyantları (Orta Filtre)		**61**		**431**		**64**		**40**		**333**		**28**
		% 99.88		% 99.13		% 99.87		% 99.92		% 99.33		% 99.94
		% 69.19		% 72.09		% 91.49		% 60		% 69.02		% 88.19
Homopolimer Uzunluğu ≤ 3		**51**		**394**		**59**		**32**		**305**		**28**
		% 99.9		% 99.21		% 99.88		% 99.94		% 99.39		% 99.94
		% 16.39		% 8.58		% 7.81		% 20		% 8.41		% 0
Çerçeve kayması, *nonsense* ve yakın *splice* varyantları (Sıkı Filtre)		**10**		**27**		**0**		**9**		**23**		**0**
		% 99.98		% 99.95		% 100		% 99.98		% 99.95		% 100
		% 83.61		% 93.74		% 100		% 77.5		% 93.09		% 100
Homopolimer Uzunluğu ≤ 3		**8**		**22**		**0**		**7**		**19**		**0**
		% 99.98		% 99.96		% 100		% 99.99		% 99.96		% 100
		% 20		% 18.52		% 0		% 22.22		% 17.39		% 0

Veri tabanında yer alan ve farklı patojenik varyantları daha önceden saptanmış bireylerde uygulanan varyant filtreleme modları ile gözden kaçan bir patojenik varyant olmamıştır, diğer bir deyişle yanlış negatif sonuç yoktur.

5.TARTIŞMA

DNA dizileme teknolojileri yıllar içerisinde büyük bir gelişme kaydederek günümüzde kısa bir sürede bir seferde bir bireyin bütün genetik yapısının dizilenmesi mümkün hale gelmiştir. Bu işlemlerin kolaylığı sebebi ile insan genetik hastalıklarının hem tanısı hem de bu hastalıklarla ilgili bilimsel araştırmalar için ileri nesil dizileme teknolojileri pek çok laboratuarda günümüzde kullanılmaktadır. Bu kapsamda geliştirilen tüm genom dizileme teknolojilerinin maliyetini düşürmek ve genomun özellikle hastalıklarla ilişkili olan genleri içeren kısmına odaklanmak için tüm ekzom dizileme teknolojileri geliştirilmiş ve sık kullanılır hale gelmiştir. Tüm ekzom dizileme teknolojileri, bu tez kapsamında da görüldüğü gibi bir bireyde bulunan çok sayıda varyantı göstermektedir; ancak bir bireyde hangi varyantın insanları birbirinden farklı kılan ve normal kabul edilebilecek bir değişiklik, hangi varyantın hastalıkla ilişkili değişiklik olduğunu anlamak, ileri nesil dizileme süreçlerini işletmekten daha çok efor gerektiren bir süreç haline gelmiştir. Bu nedenle, çeşitli algoritmalar kullanarak hastalıkla ilişkili aday varyant sayısını azaltmak ve insan eforunu daha etkili kullanmak amacıyla varyant filtreleme yazılımları geliştirilmiştir.

Bu çalışma kapsamında, *Windows* işletim sisteminde kullanılabilen bir kurumsal veri tabanı ve veri tabanından varyant filtrelemesi yapılması için proglamlama dili bilmeyen bir araştırmacıların da rahatlıkla kullanabileceği bir *web* arayüzü oluşturulmuştur. Oluşturulan veri tabanı çeşitli genetik hastalıklardan etkilendiği düşünülen 67 bireye ait ekzom dizileme verisi içermektedir. Tez kapsamında, öncelikle oluşturulan veri tabanındaki verinin niteliği incelenmiştir.

Veri tabanındaki varyantlarda ilk dikkati çeken özellik pek çok varyantın bireylerde tekrar tekrar görülmesidir. Bu durum, veri tabanında birey başına 51.255,58 varyant bulunurken birey başına ortalama 5.249,22 varyant düşmesi ile ortaya konmaktadır. Veri tabanındaki ekzonik varyantların ensembl veri tabanındaki protein kodlayan genlerin sayıları ile örtüşmesi de veri tabanının sağlıklı bir ekzom verisi ile

oluşturulduğu ve beklentileri karşıladığının göstergesidir. Buna karşın, özellikle cinsiyet kromozomlarında diğer kromozomlara kıyasla belirgin olarak beklenenden az varyant bulunmaktadır ve veri tabanının cinsiyet kromozomlarının değerlendirilmesinde yetersiz kalabileceğine bir işarettir.

Veri tabanındaki verinin dikkat çeken bir özelliği de ekzom verisinden köken almasına karşın ekzonik varyatların veri tabanındaki varyantların sadece %39,41'ini oluşturmasıdır. Bu durumun temel sebebi ekzom dizilemesi için tasarlanan sistemlerin ekzonların 5' ucu ve 3' ucundaki intronik/genler arası bölgeleri içermesidir. Bu gen bölgelerinin ekzonik bölgelerden daha fazla varyant içermesinin birkaç sebebi olabilir:

1) İntronik bölgelerdeki varyantların zarar verici etkisi, protein kodlamasına doğrudan katılan ekzonik bölgelere göre daha az olduğu için bu bölgedeki varyantlar daha tolere edilebilir ve daha çok sayıda olmaktadır;

2) intronik bölgeler, dizilenen genetik bölgelerin uç bölgelerine denk gelmekte ve bu bölgelerde okuma hataları daha sık olabilmektedir [tek yönlü okuma (intronik: %18,65; ekzonik: %15,83); Genotip≠Gözlenen hatası (intronik: %1,26; ekzonik: %0,40); *Phred* Kalite Puanı≤30 (intronik: %20,66; ekzonik: %17,90)]; 3). İntronik bölgeler daha çok homopolimer dizileri içermektedir intronik: %25,19; ekzonik: %16,55). Bu da hatalı okuma oranlarını artırmaktadır.

Veri tabanının başka bir önemli özelliği de veri tabanında akraba evliliği yapmış ebeveynlere sahip bireylerin sık görülmesidir (27/67). Bu durum, homozigot varyant sayısını artırarak homozigot varyantların filtrelenmesinde diğer uluslararası veri tabanlarında görülmeyen önemli bir katkı sağlayabilir. Duruma tersten bakıldığında ise homozigot varyant oranı bir bireyin ebeveynleri arasındaki akrabalık durumu hakkında fikir yürütmek için kullanılabilir. Buna göre, bireylerin aile hikayesinde ebeveynleri arasında akrabalık rapor etmemesine karşın, Şekil 4.4'de "akrabalık yok" grubundaki homozigot varyant oranı en yüksek olan üç bireyde ebeveynler arasında akrabalık şüphesi oluşmaktadır. Bu bireylere geri dönülüp akrabalığın yeniden sorgulanmasının

gerekliliği bu veri tabanında toplanan ekzom verisinin bu açıdan analizi ile ortaya çıkabilecek bir öngörüdür, ve bireylerin tekrar değerlendirilmesi için ipucu olabilir.

Web arayüzü kullanılarak yapılan denemelerde "HU VariantsDB"de yer alan dört bireyin ekzom verieri analiz edildiğinde bu bireylerde belirlenen analiz kriterlerine göre hastalıktan sorumlu varyantların kaçırılmadığı görülmüştür. Sıkı/Orta filtre grubuna düşen varyantlar, Homozigot/Heterozigot/Bileşik Heterozigot filtreleme grubuna düşen filtreleme seçenekleri, homopolimer bölgelerinin içinde/komşuluğunda varyantların elenmesi durumu ve hatta bir bireyde birden fazla patojenik varyant bulunması gibi çeşitli durumlarda filtreleme şemaları başarı ile patojenik varyantları kapsamıştır. Burada sunulan yöntem ile 3-71 varyant bulunduran listeler içerisinde bu patojenik varyantların yer alması eleme filtrelerinin yaklaşık 1:1000 varyantlık elemeyi yanlış negatif sonuca yol açmadan sağlayabildiğini örneklemiştir. Bu filtrelemeleri sağlayan filtreleme şeması, üç grup araçtan faydalanmaktadır: 1) Nadir olmayan varyantların elenmesi, 2) Kalıtım modeline göre eleme sağlanması, 3) Hatalı varyantların tespiti ve elenmesi.

Nadir olmayan varyantların saptanması bir varyantın MAF değeri ile mümkün olmaktadır. Veri tabanı çeşitli global veri tabanlarındaki MAF değerlerini *Ion Reporter* yazılımı aracılığı ile oluşturulan .tsv dosyasından almakta; ancak kendisi de üç adet MAF oluşturmaktadır. Bunlardan ikisi, 1000 Genom projesinden elde edilen MAF'ların özeti niteliğinde olan MinMAF ve MaxMAF değerleridir. Varyant filtremenin aslında hiçbir toplumda sık olmayan varyantları seçmesi için MaxMAF değerini kullanarak filtreleme ile her durumda başarılı olması beklenmektedir; ancak yapılan ve sunulan denemelerde iki senaryoda bunun başarısız olduğu görülmüştür (Tablo 4.5 ve Tablo 4.7). Bu durumlardan birinde *Ion Reporter*'ın aslında bulunan varyant ile aynı pozisyonda görülen fakat farklı bir varyant için atfetmesi durumu; diğerinde ise bir popülasyonda sıklığı 0,01'in hafif üzerinde olan bir varyantın elenmesi durumu vardır. Bu nedenle, oluşturulan MinMAF filtresi, final varyant listelerinde çok sayıda varyant görülmesine sebep olsa da bazen patojenik varyantların kaçırılmasına engel olduğu için mutlaka dikkate alınmalıdır.

Veri tabanında oluşturulan diğer bir filtre ise HUMAF'tır. Bu filtre veri tabanının kendi verisinden ürettiği MAF değeri olup Türkiye'ye özgüdür ve dünya için nadir olabilecek; ancak Türkiye'de sık bulunabilen varyantları da içerir. Eleme şemalarında en çok varyant elemesi sağlayan basamağın bu basamak olması (homozigot varyantlarda %88,43-%99,28; heterozigot varyantlarda %38,39-%77,26) global MAF değerlerine göre bu filtrenin belirgin bir katkı sağladığını göstermektedir. HUMAF, veri tabanındaki birey sayısı küçükken; tesadüfen bir teşhise sahip bireylerde sık görüldüğü, diğer teşhislerdeki bireylerde hiç görülmediği için filtreleme modlarına atlanan varyantların filtrelemesini kolaylaştırırken; veri tabanı boyutu büyüdükçe daha çok varyantın filtrelenmesini sağlayacaktır.

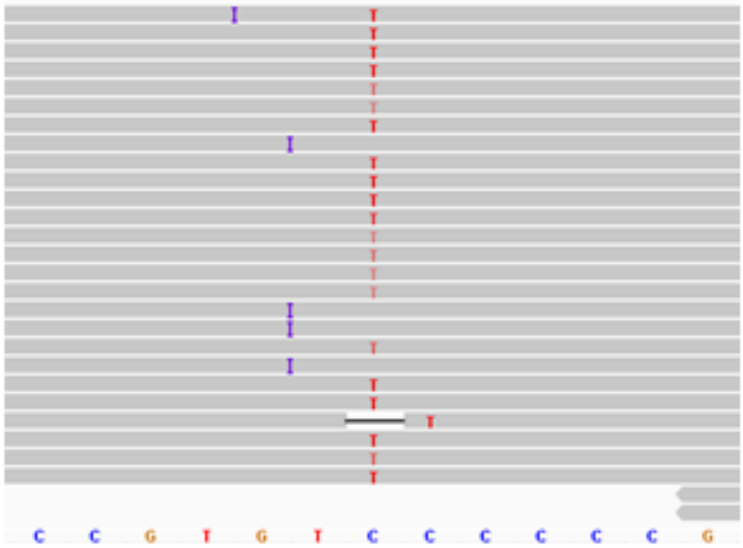
Varyant filtrelemesinde önemi büyük olan HUMAF'ın hesabı ise zordur. Veri tabanı HUMAF değerini hesaplarken .vcf kaynaklı .tsv'deki varyant veisini baz aldığı için o varyantın bulunduğu lokusa ait varyant bilgisi olmayan bireylerde homozigot referans dizisi görüldüğünü varsaymaktadır. Ancak gerçekte, o bölgenin yeterli kalitede okunamaması nedeniyle varyant olarak kaydedilmemesi de söz konusu olabilir. Bu nedenle, ilgili varyantın olduğu bölgenin veri tabanına giren tüm ekzom verilerinde yeterli okuma derinliği ve kalitesinde okunduğu varsayılmaktadır. Bu durum, aslında bazı bireylerde okunamayan çeşitli varyantların her bireyde okunmuş gibi hesaba dahil olmasına neden olarak aslında sık görülen bir varyantın nadir gibi görünmesine sebep olabilir. Bu konuda bir geliştirme sağlamanın yolu, bir genomik pozisyondaki varyantın kaç bireyde kaliteli olarak okunduğunu hesaplamak ve her HUMAF hesaplaması için o varyanta özgü bir payda kullanmaktır. Bu durum, varyant olsun olmasın tüm okuma verilerini içeren .bam dosyalarından yola çıkarak bir hesaplama yapılmasını gerektirir ki, bu hem hesaplama sürecini uzatan hem de .bam verilerine ihtiyaç doğurduğu için bu çalışmanın kapsamı dışına taşan bir durumdur. Gelecekte, her ekzom analizine özgü .bam dosyaları ile bu veri tabanının oluşturulması/desteklenmesi HUMAF hesaplarının daha doğru yapılmasını sağlayacaktır.

Bu çalışmada kalıtım modeline göre filtreleme için üç farklı mod sunulmuştur: Homozigot, Heterozigot ve Bileşik heterozigot. Bu filtreleme modları o bireydeki varyantların zigositesini dikkate almanın yanında veri tabanında farklı teşhisteki bireylerde aynı varyantın kalıtım modeli ile uygun zigositede görünüp görünmemesini de dikkate almaktadır. Bu filtreleme stratejileri açısından önemli, GATK gibi diğer filtreleme programlarında olmayan bir yeniliktir ve ek varyant elemesi sağlamaktadır (106). Ancak bu modların kullanılmasında, yanlış negatif sonuçlara sebep olabilecek ve dikkat gerektiren bir durum vardır. Veri tabanındaki bireylerin yanlış bir teşhis ile kaydedilmiş olmalarıdır. Bu durumda veri tabanında bireyin gerçek teşhisi ile uyan ve aynı patojenik varyanta sahip bir birey varsa o patojenik varyant elenecektir. Özellikle veri tabanı büyüdükçe bu sorun yaratabilecek bir durumdur. Bunun çözümü olarak, aynı teşhisin farklı teşhisteki bireylerde görülmesi için kullanıcının belirleyeceği bir tolerans sağlanabilir. Örneğin, diğer teşhislerden 1 bireyde aynı varyantın bulunmasına izin verildiğinde veri tabanına yanlış teşhis konulmuş 1 birey bulunması yanlış negatif sonuca neden olmayacaktır.

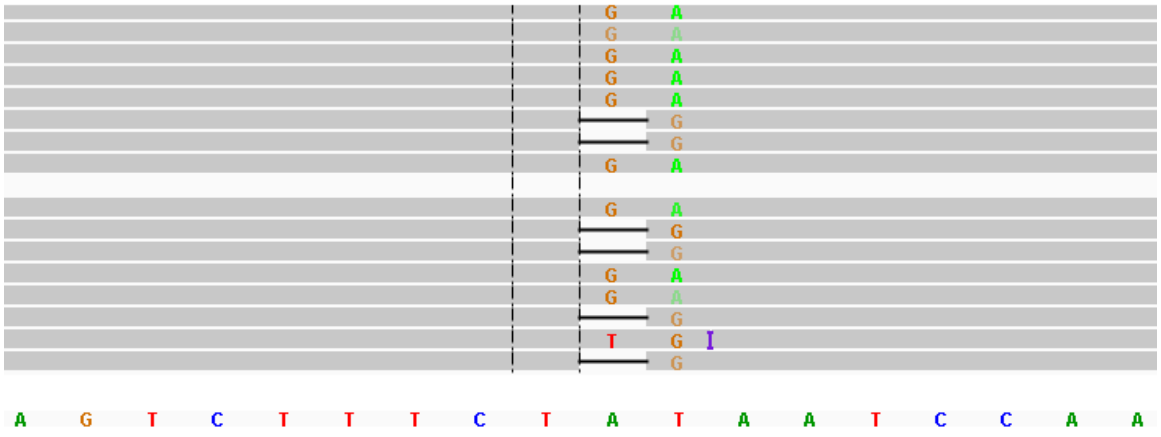
Filtreleme modlarından, bileşik heterozigot modu resesif hastalıklarda görülebilen bileşik heterozigot varyantları aramak için tasarlanmıştır. Bu filtreleme modu diğer filtreleme parametrelerinden hem önce hem sonra çalışarak filtreleme yaptığı için diğer filtreleme parametrelerinde elenen; ama aynı gende bulunan varyantların bir daha elemesini sağlamaktadır. Örneğin, bir gene ait üç heterozigot varyant varsa ve bunlardan ikisi MinMAF filtresi ile eleniyorsa geriye kalan bir varyant artık bileşik heterozigot durumda bulunmadığı için ikinci filtreleme ile tekrar elenmektedir. Bileşik Heterozigot mod, mevcut durumu ile aynı genomik pozisyona denk gelen 2 farklı varyantı öyle olmasına karşın bileşik heterozigot olarak değerlendirmemektedir. Bu sorun, çok nadir karşılaşılabilecek bir durum olması nedeniyle henüz bir çözüme gidilmemiştir; ancak gelecekte ihtiyaç duyulursa bireylerin varyantlarını veri tabanına eklerken oluşturulabilecek ekstra basamaklarla önüne geçilebilir. Bileşik heterozigot modunun avantajı bu durumun düşünüldüğü bireylerde gevşek modda bile heterozigot varyantlar

arasında %56,99-%81,91'lik bir eleme sağlaması ile ortaya çıkmaktadır. Bileşik heterozigot modunun dezavatajı ise, saptadığı varyantların cis veya trans olduğunu öngörememesidir. Bunun için ebeveyn-çocuk çalışmaları gerekmektedir.

Tez kapsamında hatalı varyantların tespiti ve elenmesi için filtreleme seçeneği üretmek üzere iki farklı yöntem izlenmştir. Birinci yöntemde, *Ion Reporter*'ın kendi sunduğu anotasyon parametreleri üzerinden filtreleme seçeneği sağlanmıştır. Bu yöntemle okuma derinliği, *Phred* Kalite Puanı, tek/çift yön okuma ve MinHomopolimer filtreleri oluşturulmuş, okuma derinliği ve *Phred* Kalite Puanı doğrudan *Ion Reporter*'dan alınırken MinHomopolimer ve tek/çift yön okuma filtreleri *Ion Reporter*'dan hesaplanmıştır. İkinci hatalı varyant filtreleme yöntemi olarak, *Ion Reporter* yazılımının sunmadığı, hatta *Ion Reporter*'ın sebep olduğu hataların saptanmasına ve bu hataların elenmesine çaba sarf edilmiştir. Bu hatalardan Genotip≠Gözlenen hatası, bir varyant için "Genotype" ve "Observed" alanlarını farklı metodlarla belirleyen *Ion Reporter* yazılımının bir hatasıdır ve daha çok homopolimer bölgelerindeki indelleri işaret etmektedir. Bu hata dışında iki çeşit daha hata saptanmıştır: 1) "Homopolimer yakını *missense* polimorfizm" (Şekil 5.1), 2) "Aynı varyantın farklı isimlendirilmesi" (Şekil 5.2). Bu hataların çözümü için sadece .tsv verisinden kaynak alacak bir yöntem mevcut olmayıp ya o bölgenin genomik dizisinin incelenmesi ile hatanın öngörülmesi ya da .bam dosyasından okuma hizalaması ve .vcf dosyasından varyant anotasyonunun baştan yapılması gerekmektedir. Bu şekilde, beklenmedik sistematik hatalarla karşılaşılması *Ion Reporter*'ın başka sistematik hatalar da yapabileceğini düşündürmekle beraber bu çalışma kapsamında başka bir hata türü ile karşılaşılmamıştır.



Şekil 5.1. "Homopolimer yakını *missense* polimorfizm" hatası. *Ion Reporter* tarafından 3149 nolu bireyde *FBLIM1*:c.571_572insT olarak anote edilen varyant, yukarıdaki gibi .bam dosyası IGV (*Integrated Genomics Viewer*) programı ile incelendiğinde 2 nükleotid uzağındaki c.573C>T polimorfizmi ve bu polimorfizmi içine alan 5bp uzunluğundaki homopolimer bölgesinin yanlış konumlandırılmasından kaynaklanmaktadır. Bu hata, homopolimer bölgesine doğrudan komşu olmadığı için MinHomopolimer filtresi ile filtrelenmemektedir. Bu türden hatalar orta homozigot filtrelemede her bireye ait ekzom versinde birkaç kez tekrarlanmaktadır.



Şekil 5.2. "Aynı varyantın farklı isimlendirilmesi" hatası. *Ion Reporter* tarafından *TEX11*:c.2566-3AT>TC olarak anote edilen varyant için 3010 nolu bireyde referans değişkeni "TAT"; "gözlenen" değişkeni "TGA" olarak kaydedilmiştir. Yukarıdaki gibi .bam dosyası IGV programı ile

> incelendiğinde aynı varyant olarak başka bir bireylerde de görünen bu varyant için tüm diğer bireylerde referans değişkeni “AT”; “gözlenen” değişkeni “GA” olarak kaydedilmiştir. Bu nedenle 3010 nolu bireyde bu varyant isimlendirme hatası nedeni ile çok nadir bir varyant gibi görünen aslında yaygın bir varyanttır.

Oluşturulan yazılım paketleri, 67 ekzom verisi ile etkin bir filtreleme sağlasa dahi veri tabanın genişlemesi hem özgün popülasyon verisinin artması hem de aynı laboratuvardan kaynaklanan sistematik hataların veri tabanında yer bulması ile filtrelemeleri daha etkin kılacaktır. Bu amaçla uygulanabilecek bir yöntem de başka kaynaklardan elde edilen (örn: İllumina) ekzom verilerinin veri tabanına yüklenmesidir. Bu şekilde veri tabanının genişleme hızı artırılabilir.

Öte yandan, veri tabanına yeni ekzom verisi kaydı, veri tabanı için giderek uzayan sürelerde gerçekleşmektedir. Mevcut performansta en uygun polinoma göre veri tabanına kayıt süresi ekponansiyel olarak artmaktadır. Bu sürenin değişim hızı ekzom verisi arttıkça yavaşlayabilir; ancak 100. ekzom verisinin veri tabanına yüklenmesi için bile öngörülen süre 6:53 saattir. Bunun önlenmesi açısından tez kapsamında tamamlanmamış olmakla birlikte HUMAF hesaplamalarının veri aktarımı sırasında arka yüz kodlaması ile değil aktarım bittikten sonra bir *stored procedure* ile gerçekleşmesi sağlanmalıdır. Bu sayede, milyonlarca varyant bilgisine ait HUMAF değeri ortalama nitelikli bir veri tabanı üzerinde bile 40-60 saniye üzerinde hesaplanacaktır. Böylece, bir hastada bulunan ortalama 51.255 varyant veri tabanına 36 dakikada (veri tabanına ilk kez aktarılan hastaya ait varyantın aktarılma süresi ve *stored procedure* ile HUMAF hesabı süresi toplamı) aktarılmasının mümkün olacağı öngörülmektedir.

Veri tabanına bazı özelliklerin eklenmesi, daha da farklı analiz seçeneklerinin kapısını aralayacaktır. Önemli olabilecek eklemelerden birisi popülasyon genetiği analizleri için daha kapsamlı bilgi sunacak olan bireylere ait cinsiyet, akrabalık ilişkileri, doğum yeri gibi bazı özelliklerin veri tabanına eklenmesidir. Bir diğer durum ise, ebeveyn verisinden faydalanarak analizlerin genişletilmesidir. *Trio* ekzom verisinin bulunduğu durumlarda kalıtım şemasının daha etkin kullanımı ve *de novo* varyantların tespiti için

trio analiz seçeneği eklenebilir. Bu varyant filtrelenmesini kolaylaştıracak aday patojenik varyant sayısını azaltacaktır. Oluşturulan veri tabanı mevcut durumda aday varyantlar için OMIM, ExAC ve GeneCards veri tabanlarına bağlantı olanağı sunması ile kullanım kolaylığı da sağlamaktadır. Bu veri tabanlarının sayısı arttırılabilir (107).

Bu veri tabanının nihai amacı verilerin ilgili araştırmacıların erişimine açılmasıdır; bu amaçla Hacettepe Üniversitesi Tıp Fakültesi Tıbbi Genetik Anabilim Dalı'ndaki kullanıcılar oluşturulan veri tabanı ve yazılımları kullanmışlar, test etmişler ve geribildirimlerini sunmuşlardır. Böylece, veri tabanının oluşturulması, arka yüz ve ön yüz mimarileri, yazılımın hedef kullanıcısı olan insan genetiği ile ilgilenen araştırmacıların kullanım kolaylıkları gözetilerek tasarlanmıştır.

Veri tabanının araştırmacılara açılmasında veri güvenliği de önemli bir konu haline gelmektedir. Veri güvenliğinin sağlanması için uygulama üzerinde her ne kadar SQL *injection*, XSS (*Cross Site Scripting*) gibi popüler veri çalma ataklarına karşı önlemler yazılım geliştirme aşamasında özellikle dikkate alınmış olsa da, veri tabanına yetkisiz erişimlerin *network* ağı üzerinden de engellemesi gerekecektir. Bunun için gerek SQL veri tabanı sunucusu üzerinde gerekse uygulama sunucusu üzerinde atakları önleyecek önlemlerin profesyonel bir ağ (*network*) ve bilgi güvenliği ekibi tarafından sağlanması ve idamesi gerekecektir. Ayrıca, uygulama intranet dışında yayınlanmaya başladığı andan itibaren kullanıcı sayısında meydana gelecek büyük artış her ne kadar uygulama geliştirme aşamasında oturum (*session*) yönetimi ile minimize edilmeye çalışılmışsa da ağ üzerindeki trafiğin artmasıyla SQL sorgularının yavaşlamasına ya da uygulamanın yanıt verme süresinin düşmesine sebep olacaktır. Bu gibi sorunların çözümü, gerek uygulama sunucusu gerekse veri tabanı sunucusu için yeni kaynaklar gerektirecektir. Uygulama sunucusu sayısının artması ve sunucular arasındaki trafiğin yönetilmesi için ise nitelikli sistem yöneticilerine ve ağ ekibine ihtiyaç duyulacaktır.

Uygulamanın ilerleyen süreçlerinde ortaya çıkması öngörülen bir diğer sorun ise filtre sorgularının ya da aramaların yavaşlaması olacaktır. Çünkü her yeni kayıt eklendiğinde daha önce oluşturulan kümelenmiş *index* yapısı bozulacak, yapılan arama

sorguları da bu bozulmayla doğru orantılı olarak yavaşlayacaktır. Bu yüzden, belirli periyotlarda veri tabanı üzerinde bu indekslerin güncellenmesi gerekecektir (108).

Bu çalışmada, kurumsal bir veri tabanından da faydalanarak ekzom verisinden varyant filtrelemesinin farklı senaryolar için çeşitlendirilmesini, yazılım dili bilmeyen kişilerce de ekzom filtrelemelerinin yapılabilmesi, sık ekzom dizileme hatalarının saptanması ve elenmesi için bir veri tabanı ve *web* arayüzü oluşturulmuştur. Bu sistem, Hacettepe Üniversitesi bünyesinde oluşturulmuş olup gelişmeye, daha çok bireye ait ekzom verisi dahil etmeye uygun şekilde tasarlanmıştır. Bu şekilde, büyüyerek araştırıcıların ekzom verilerinin analizlerini kolaylaştırması, hızlandırması ve aynı zamanda bir popülasyon veri tabanı haline gelmesi hedeflenmiştir.

6. SONUÇ VE ÖNERİLER

1. Tez kapsamında kurumsal ekzom verilerinin analizlerine yönelik kulanıcı dostu özgün bir veri tabanı, “HU VariantsDB” ve varyant filtrelemesine yönelik bir web arayüzü oluşturulmuştur.

2. Farklı kalıtım kalıpları varsayımları üzerinden veri filtreleme olanaklı hale gelmiştir.

3. Kurumsal veri filtreleme aşaması (HUMAF) homozigotlarda %80-90 civarında varyant elemesi yapabilirken bu oran heterozigotlar için %70 civarındadır. Populasyona yönelik veri eksikliği göz önüne alındığında HUMAF filtrelemesi varyant sayısını azaltmada etkin çözüm üretebilmektedir.

4. Indeks vakalarda birleşik heterozigotlara yönelik kademeli eleme stratejisi geliştirilmiştir.

5. Oluşturulan yazılımlar, Ion Reporter yazılımın tanıttığı bazı hataları tespit edebilmekte ve etkin filtreleme yapabilmektedir.

6. Varyant havuzunun genişletilmesi ve farklı alt yapılardan (Illumina) elde edilen verinin veri tabanına yüklenmesi programın filtreleme kapasitesini artıracaktır. Oluşturulan program, buna uyumlu olarak hazırlanmıştır.

7. Veri tabanına her yeni bireyin eklenmesi işlem süresini artırmaktadır. Bunun önüne geçebilmek için HUMAF hesaplamalarının veri aktarımı sırasında arka yüz kodlaması ile değil aktarım bittikten sonra bir stored procedure ile gerçekleşmesi sağlanmalıdır.

8. Bu veri tabanının nihai amacı verilerin ilgili araştırmacıların erişimine açılmasıdır. Hacettepe Üniversitesi bünyesinde oluşturulmuş olan bu sistem genişletilmeye açık olarak planlanmıştır. Bu aşama için veri güvenliği, veri tabanı güncellemeleri ve ağ trafiği yönetimi için gereklilikler göz önüne alınmalıdır.

7. KAYNAKLAR

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001; 409 (6822): 860–921.

2. Pennisi E, Genomics. ENCODE project writes eulogy for junk DNA. Science. 2012; 337 (6099): 1159–61.

3. Frésard L, Montgomery SB. Diagnosing rare diseases after the exome. Cold Spring Harb Mol Case Stud. 2018; 4(6): a003392.

4. Lek M, Karczewski KJ, Minikel Eric V, Samocha KE, Banks E, Fennell T, ve ark. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016; 536 (7616): 285–91.

5. Karczewski KJ, Francioli LC, Grace Tiao G, Cummings BB, Alföldi J, Wang Q. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. BioRxiv. 2019; 531210. https://doi.org/10.1101/531210

6. Scott EM, Halees A, Itan Y, Spencer EG, He Y, Azab MA ve ark. Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. Nat Genet. 2016; 48(9): 1071-76.

7. Cetinkaya A, Xiong JR, Vargel İ, Kösemehmetoğlu K, Canter H, Gerdan ÖF ve ark. Loss-of-Function Mutations in ELMO2 Cause Intraosseous Vascular Malformation by Impeding RAC1 Signaling. Am J Hum Genet. 2016; 99(2): 299–317.

8. Türkiye Genom Projesi Başladı [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://www.tuseb.gov.tr/haberler/turkiye-genom-projesi-basladi

9. Chial, H. DNA sequencing technologies key to the Human Genome Project. Nature Education. 2008; 1(1): 219.

10. Gloss BS, Dinger ME. Realizing the significance of noncoding functionality in clinical genomics. Exp Mol Med. 2018; 50(8): 97.

11. Venter CJ, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, ve ark. The Sequence of the Human Genome. Science. 2001: 291(5507): 1304-51.

12. Fidanoğlu P, Belder N, Erdoğan B, İlk Ö, Rarajbli F, Özdağ H. Genom projeleri 5N1H: ne, nerede, ne zaman, nasıl, neden ve hangi popülasyonda? Türk Hij Den Biyol Derg. 2014; 71: 45-60.

13. Shearer AE, Eppsteiner RW, Booth KT, Ephraim SS, Gurrola J, Simpson A, ve ark. Utilizing ethnic-specific differences in minor allele frequency to recategorize reported pathogenic deafness variants. Am J Hum Genet. 2014; 95(4): 445-53.

14. Karki R, Pandya D, Elston RC, Ferlini C. Defining "mutation" and "polymorphism" in the era of personal genomics. BMC Med Genomics. 2015; 15: 8-37.

15. Anna A, Monika G. Splicing mutations in human genetic disorders: examples, detection, and confirmation. J Appl Genet. 2018; 59(3): 253–68.

16. He F, Jacobson A. Nonsense-Mediated mRNA Decay: Degradation of defective transcripts is only part of the story. Annu Rev Genet. 2015; 49: 339–66.

17. List of UCSC genome releases [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://genome.ucsc.edu/FAQ/FAQreleases.html#release1

18. 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015; 526(7571): 68–74.

19. NHLBI GO Exome Sequencing Project (ESP) [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: http://evs.gs.washington.edu/EVS/

20. NCBI. GenBank release notes [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt

21. Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, ve ark. The UCSC Genome Browser database: 2019 update. Nucleic Acids Res. 2019; 47 (1): 853–58.

22. Liu S, Huang S, Chen F, Zhao L, Yuan Y, Francis SS, ve ark. Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History. Cell. 2018; 175(2): 347-59.

23. Turnbull C, Scott RH, Thomas E, Jones L, Murugaesu N, Pretty FB, ve ark. The 100 000 Genomes Project: Bringing whole genome sequencing to the NHS. BMJ. 2018; 361: k1687.

24. Boomsma DI, Wijmenga C, Slagboom EP, Swertz MA, Karssen LC, Abdellaoui A, ve ark. The Genome of the Netherlands: design, and project goals. Eur J Hum Genet. 2014; 22(2): 221–7.

25. Yamaguchi-Kabata Y, Nakazono K, Takahashi A, Saito S, Hosono N, Kubo M, ve ark. Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. Am J Hum Genet. 2008; 83(4): 445-56.

26. Ngamphiw C, Assawamakin A, Xu S, Shaw PJ, Yang JO, Ghang H, ve ark. PanSNPdb: the Pan-Asian SNP genotyping database. PLoS One. 2011; 6(6): e21451.

27. Indian Genome Variation Consortium. The Indian Genome Variation database (IGVdb): a project overview. Hum Genet. 2005; 118(1): 1-11.

28. Teo YY, Sim X, Ong RT, Tan AK, Chen J, Tantoso E, ve ark. Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. Genome Res. 2009; 19(11): 2154-62.

29. Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. Int J Epidemiol. 2015; 44(4): 1137-47.

30. Alkan C, Kavak P, Somel M, Gokcumen O, Ugurlu S, Saygi C, ve ark. Whole genome sequencing of Turkish genomes reveals functional private alleles and impact of genetic interactions with Europe, Asia and Africa. Genomics. 2014; 15: 963-75.

31. Makalowski W. The human genome structure and organization. Acta Biochim Pol. 2001; 48(3): 587-98.

32. Tarailo-Graovac M, Shyr C, Ross CJ, Horvath GA, Salvarinova R, Ye XC, ve ark. Exome Sequencing and the management of neurometabolic disorders. N Engl J Med. 2016; 374(23): 2246-55.

33. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, ve ark. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008; 456(7218): 53–9.

34. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, ve ark. An integrated semiconductor device enabling non-optical genome sequencing. Nature. 2011; 475(7356): 348-52.

35. Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics. 2012; 28(21): 2711–8.

36. Ivády G, Madar L, Dzsudzsák E, Koczok K, Kappelmayer J, Krulisova V, ve ark. Analytical parameters and validation of homopolymer detection in a pyrosequencing-based next generation sequencing system. BMC Genomics. 2018; 19(1): 158.

37. Zhang H. Overview of Sequence Data Formats. Methods in Molecular Biology. 2016; 1418: 3-17.

38. Annotations available in Ion Reporter™ Software [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi:

https://ionreporter.thermofisher.com/ionreporter/help/GUID-8CF8A299-8BA1-4AFC-B6FE-837F59A7E74D.html

39. Annotation sources [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://assets.thermofisher.com/TFS-Assets/LSG/manuals/MAN0017605_IonReporter5_10_UG.pdf 304-9

40. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. 2012; 40 (Web Server issue): W452-7.

41. Grantham, R. Amino acid difference formula to help explain protein evolution. Science. 1974; 185 (4154): 862–86.

42. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, ve ark. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7(4): 248–9.

43. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, ve ark. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat. 2013; 34(1): 57–65.

44. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, ve ark. The Pfam protein families database in 2019. Nucleic Acids Res. 2019; 47(D1): D427–32.

45. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, ve ark. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001; 29(1): 308–11.

46. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. 2014; 42: D986–92.

47. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N ve ark. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2019; 47(D1): D941–47.

48. Online Mendelian Inheritance in Man, OMIM. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://www.omim.org/

49. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2019; 47(D1): D330-8.

50. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, ve ark. A major update to the DrugBank database for 2018. Nucleic Acids Res. 2018; 46(D1): D1074-82.

51. Landrum MJ, Kattman BL. ClinVar at five years: Delivering on the promise. Human Mutation. 2018; 39(11): 1623-30.

52. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 1998; 8(3): 186-94.

53. Sam S. Types of databases. [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://www.tutorialspoint.com/Types-of-databases

54. Sam S. Types of databases [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://www.computerhistory.org/collections/catalog/102695039

55. IMS Then and Today, IBM's Information Management System (IMS) [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: http://idcp.marist.edu/pdfs/ztidbitz/22%20zTidBits%20(IMS_Then&ToDay).pdf

56. History of IMS: Beginnings at NASA [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi:https://www.ibm.com/support/knowledgecenter/zosbasics/com.ibm.imsintro.doc.intro/ip0ind0011003710.htm

57. Olle, T. An assessment of how the CODASYL Data Base Task Group Proposal meets the GUIDE-SHARE requirements, Conference on Data Systems Languages records, . Charles Babbage Institute Archives. [Internet]. 1972 [Erişim Tarihi 16.09.2019]. https://archives.lib.umn.edu/repositories/3/archival_objects/13960

58. Olle TW. The Codasyl Approach to Data Base Management. New York, NY, USA: John Wiley & Sons; 1978.

59. Codd EF. Derivability, Redundancy and Consistency of Relations Stored in Large Data Banks. New York, NY, USA: IBM Thomas J. Watson Research Center; 1969.

60. Codd EF. A Relational Model of Data for Large Shared Data Banks. Communications of the ACM. 1970; 13(6): 377-387.

61. Stonebraker M, Held G, Wong E, Kreps P. The design and implementation of INGRES. Journal ACM Transections on Database Systems. 1976; 1(3): 189-222.

62. Rawe LA. History of the Ingres Corporation. IEEE Ann Hist Comput. 2012; 34(4): 58-70.

63. Stonebraker M, Rawe LA. The design of POSTGRES. SIGMOD '86 (Proceedings of the 1986 ACM SIGMOD international conference on Management of data); 28-30.05.1986; Washington DC, USA. 340-55.

64. Date CJ. A Guide to the SQL Standard: a User's Guide to the Standard Relational Language SQL. Boston, MA, USA: Addison-Wesley Longman Publishing Co Inc; 1987.

65. ISO 9075:1987 Information processing systems -- Database language – SQL [Internet]. [Erişim Tarihi: 16.09.2019]. Erişim Adresi https://www.iso.org/standard/16661.html

66. ANSI/ISO/IEC International Standard (IS) Database Language SQL — Part 2: Foundation (SQL/Foundation) [Internet]. 1999 [Erişim Tarihi 16.09.2019]. Erişim Adresi: http://web.cecs.pdx.edu/~len/sql1999.pdf

67. Chamberlin DD. Early History of SQL. IEEE Ann Hist Comput. 2012; 34(4): 78–82.

68. Connolly T, Begg C. Database Systems – A Practical Approach to Design Implementation and Management. 6th ed. Essex, UK: Pearson; 2014. Chapter2, Database Environment; p.64.

69. DB-Engines Ranking of Relational DBMS [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://db-engines.com/en/ranking/relational+dbms

70. What is SQL Server Management Studio (SSMS)? [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://docs.microsoft.com/en-us/sql/ssms/sql-server-management-studio-ssms?redirectedfrom=MSDN&view=sql-server-2017

71. Transact-SQL Reference (Database Engine) [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://docs.microsoft.com/en-us/sql/t-sql/language-reference?view=sql-server-2017

72. Data types (Transact-SQL) [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://docs.microsoft.com/en-us/sql/t-sql/data-types/data-types-transact-sql?view=sql-server-2017

73. Pin P, Shan C. The Entity-Relationship Model - Toward a Unified View of Data. ACM Transactions on Database Systems. International conference on very large data bases; 22–24.09.1975; Framingham, MA, USA. 1(1): 9–36

74. Nonclustered Index Structures [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://docs.microsoft.com/en-us/previous-versions/sql/sql-server-2008-r2/ms177484(v=sql.105)

75. Clustered Index Structures [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://docs.microsoft.com/en-us/previous-versions/sql/sql-server-2008-r2/ms177443(v=sql.105)?redirectedfrom=MSDN

76. Eckerson WW. Three Tier Client/Server Architecture: Achieving Scalability, Performance, and Efficiency in Client Server Applications. Open Information Systems. 1995; 3(20): 10-1.

77. Fielding RT, Gettys J, Mogul JC, Nielsen HF, Masinter L, Leach PJ, ve ark. Hypertext Transfer Protocol – HTTP/1.1 [Internet] 1999 (Erişim Tarihi 16.09.2019]. Erişim Adresi: http://delivery.acm.org/10.1145/rfc_fulltext/RFC2616/rfc2616.txt?ip=193.140.237.30&id=RFC2616

78. What is the best programming language to learn for backend developers? [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://www.slant.co/topics/7812/~programming-language-to-learn-for-backend-developers

79. Berners-Lee, T. Information Management: A Proposal. CERN [Internet]. 1989 [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://www.w3.org/History/1989/proposal.html.

80. World Wide Web Consortium, Tags used in HTML [Internet]. 1992. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://www.w3.org/History/19921103hypertext/hypertext/WWW/MarkUp/Tags.html

81. World Wide Web Consortium, HTML 5 [Internet]. 2008. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://www.w3.org/TR/2008/WD-html5-20080122/

82. World Wide Web Consortium, What is CSS? [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://www.w3.org/standards/webdesign/htmlcss#whatcss

83. World Wide Web Consortium, Cascading Style Sheets, level 1 [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://www.w3.org/TR/CSS1/

84. Otto, M. Say hello to Bootstrap 2.0 [Internet]. 2012 [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://blog.twitter.com/developer/en_us/a/2012/say-hello-to-bootstrap-2.html

85. Browsers and devices [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://getbootstrap.com/docs/4.3/getting-started/browsers-devices/

86. Flanagan D. JavaScript - The definitive guide. 6th ed. Sebastopol, CA, USA: O'Reilly Media; 2011. Chapter 1, Introduction to JavaScript; p.1-8.

87. York R. Beginning JavaScript and CSS Development with jQuery. Indianapolis, IN, USA: Wiley; 2009. Chapter2, Selecting And Filtering; p.28.

88. Resig, J. History of jQuery [Internet]. 2017 [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://www.slideshare.net/jeresig/history-of-jquery

89. .NET Framework Guide [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://docs.microsoft.com/en-us/dotnet/framework/

90. A Tour of the C# Language Guide [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://docs.microsoft.com/en-us/dotnet/csharp/tour-of-csharp/

91. ADO.NET Overview [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://docs.microsoft.com/en-us/dotnet/framework/data/adonet/ado-net-overview

92. Hendrickson E. Explore It!: Reduce Risk and Increase Confidence With Exploratory Testing. USA: The Pragmatic Programmers; 2013. Chapter 7, Explore Entities And Their Relationships; p.70-3.

93. .NET Framework Data Providers [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://docs.microsoft.com/en-us/dotnet/framework/data/adonet/data-providers

94. LINQ [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://docs.microsoft.com/en-us/previous-versions/cc299380(v%3dmsdn.10)

95. Extension Methods (C# Programming Guide) [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://docs.microsoft.com/en-us/dotnet/csharp/programming-guide/classes-and-structs/extension-methods

96. Generics (C# Programming Guide) [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://docs.microsoft.com/en-us/dotnet/csharp/programming-guide/generics/

97. Mehta VP. Pro LINQ Object Relational Mapping with C# 2008. Berkeley, CA, USA: Apress; 2008. Chapter1, Object-Relational Mapping Concepts; p.3-17.

98. Microsoft. Entity Framework Overview. [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://docs.microsoft.com/en-us/dotnet/framework/data/adonet/ef/overview

99. What's Dapper? [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://dapper-tutorial.net/dapper

100. Fowler M. Patterns of Enterprise Application Architecture. Boston, MA, USA: Addison-Wesley Longman Publishing Co Inc; 2003.

101. Biomart – Ensembl [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://www.ensembl.org/biomart/martview/9a5493c1d8fdc471dd5208f023e7f4e0

102. Ion Reporter Software 5.12 Help – Location Filter [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://ionreporter.thermofisher.com/ionreporter/help/GUID-7E23CDA5-56A1-41BE-9BA7-582FA9B6716B.html

103. Ion Reporter Software 5.12 Help – Variant Type Filter [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://ionreporter.thermofisher.com/ionreporter/help/GUID-C5359815-CC5B-458D-93B6-C06622410168.html

104. Ion Reporter Software 5.12 Help – Variant Effect Filter [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://ionreporter.thermofisher.com/ionreporter/help/GUID-15F823D3-7D3A-4CB8-9DEC-FEBE10C89618.html

105. Ensembl genome browser 97 [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: http://grch37.ensembl.org/Homo_sapiens/Info/Index

106. DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, ve ark. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Gen. 2011; 43: 491-8.

107. Stelzer G, Rosen R, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, ve ark. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analysis. Curr Protoc Bioinformatics. 2016; 54: 1.30.1-1.30.33.

108. Reorganize and rebuild indexes [Internet]. [Erişim Tarihi 16.09.2019]. Erişim Adresi: https://docs.microsoft.com/en-us/sql/relational-databases/indexes/reorganize-and-rebuild-indexes?view=sql-server-2017

8. EKLER

EK-1: Tez Çalışması ile İlgili Etik Kurul İzni

T.C.
HACETTEPE ÜNİVERSİTESİ
Girişimsel Olmayan Klinik Araştırmalar Etik Kurulu

Sayı : 16969557-1621
Konu : ARAŞTIRMA PROJESİ DEĞERLENDİRME RAPORU

Toplantı Tarihi : 03 EYLÜL 2019 SALI
Toplantı No : 2019/20
Proje No : GO 18/117 (Onay Tarihi: 31.01.2018)
Karar No : GO 18/117-03

Kurulumuzun 31.01.2018 tarihli toplantısında GO 18/117 kayıt numarası ile onaylanmış olan, Üniversitemiz Tıp Fakültesi Tıbbi Genetik Anabilim Dalı öğretim üyelerinden Prof. Dr. Ayşe Nurten AKARSU'nun sorumlu araştırmacı olduğu, Dr. Öğr. Üyesi İdil YET, Prof. Dr. Şule ÜNAL, Uzm. Bio. Can KOŞUKCU ile birlikte çalışacakları ve Yavuz ADABALI'nın yüksek lisans tezi olan, GO 18/117 kayıt numaralı, ***"Exom Veri Setinden Hastalığa Özgü Varyant Veri Tabanı Oluşturulması"*** başlıklı proje için vermiş olduğunuz 26.08.2019 tarihli başlık değişikliği dilekçesi talebi Kurulumuzun 03.09.2019 tarihli toplantısında görüşülmüş ve **uygun bulunmuştur**. Çalışmanın başlığı ***"Ekzom Veri Setinden Hastalığa Özgü Varyant Veri Tabanı Oluşturulması"*** olarak değiştirilmiş ve kayıtlarımıza eklenmiştir. Çalışma tamamlandığında sonuçlarını içeren bir rapor örneğinin Etik Kurulumuza gönderilmesi gerekmektedir.

1.Prof. Dr. Ayşe Lale DOĞAN (Başkan)
2. Prof. Dr. Sevda F. MÜFTÜOĞLU (Üye)
3. Prof. Dr. M. Yıldırım SARA (Üye)
4. Prof. Dr. Nedret SAĞLAM (Üye)
5. Prof. Dr. Mintaze Kerem GÜNEL (Üye)
İZİNLİ
6. Prof. Dr. Oya Nuran EMİROĞLU (Üye)
7. Prof. Dr. M. Özgür UYANIK (Üye)
İZİNLİ
8. Doç. Dr. Gözde GİRGİN (Üye)
9. Doç. Dr. Fatma Visal OKUR (Üye)
İZİNLİ
10. Doç. Dr. Can Ebru KURT (Üye)
11. Doç. Dr. H. Hüsrev TURNAGÖL (Üye)
12. Dr. Öğr. Üyesi Özay GÖKÖZ (Üye)
13. Dr. Öğr. Üyesi Müge DEMİR (Üye)
İZİNLİ
14. Öğr. Gör. Dr. Meltem ŞENGELEN (Üye)
15. Av. Meltem ONURLU (Üye)

Hacettepe Üniversitesi Girişimsel Olmayan Klinik Araştırmalar Etik Kurulu
06100 Sıhhiye-Ankara
Telefon: 0 (312) 305 1082 • Faks: 0 (312) 310 0580 • E-posta: goetik@hacettepe.edu.tr
Ayrıntılı Bilgi için:

EK-2: Tez Çalışması Orijinallik Raporu

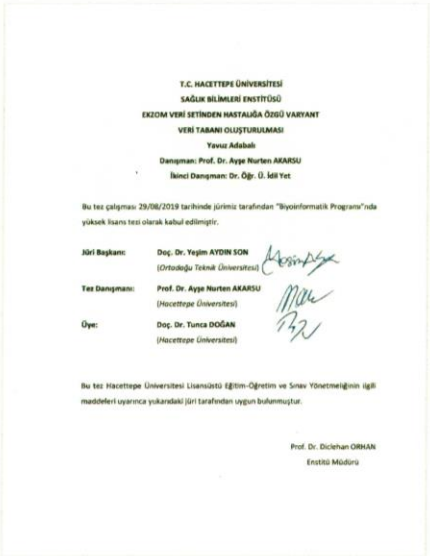
turnitin

Dijital Makbuz

Bu makbuz ödevinizin Turnitin'e ulaştığını bildirmektedir. Gönderiminize dair bilgiler şöyledir:

Gönderinizin ilk sayfası aşağıda gönderilmektedir.

Gönderen:	Yavuz Adabalı
Ödev başlığı:	Ekzom Veri Setinden Hastalığa Özg…
Gönderi Başlığı:	Ekzom Veri Setinden Hastalığa Özg…
Dosya adı:	yavuz_adabali_yl_tez.pdf
Dosya boyutu:	3.82M
Sayfa sayısı:	120
Kelime sayısı:	22,360
Karakter sayısı:	144,597
Gönderim Tarihi:	18-Eyl-2019 03:58PM (UTC+0300)
Gönderim Numarası:	1175079056



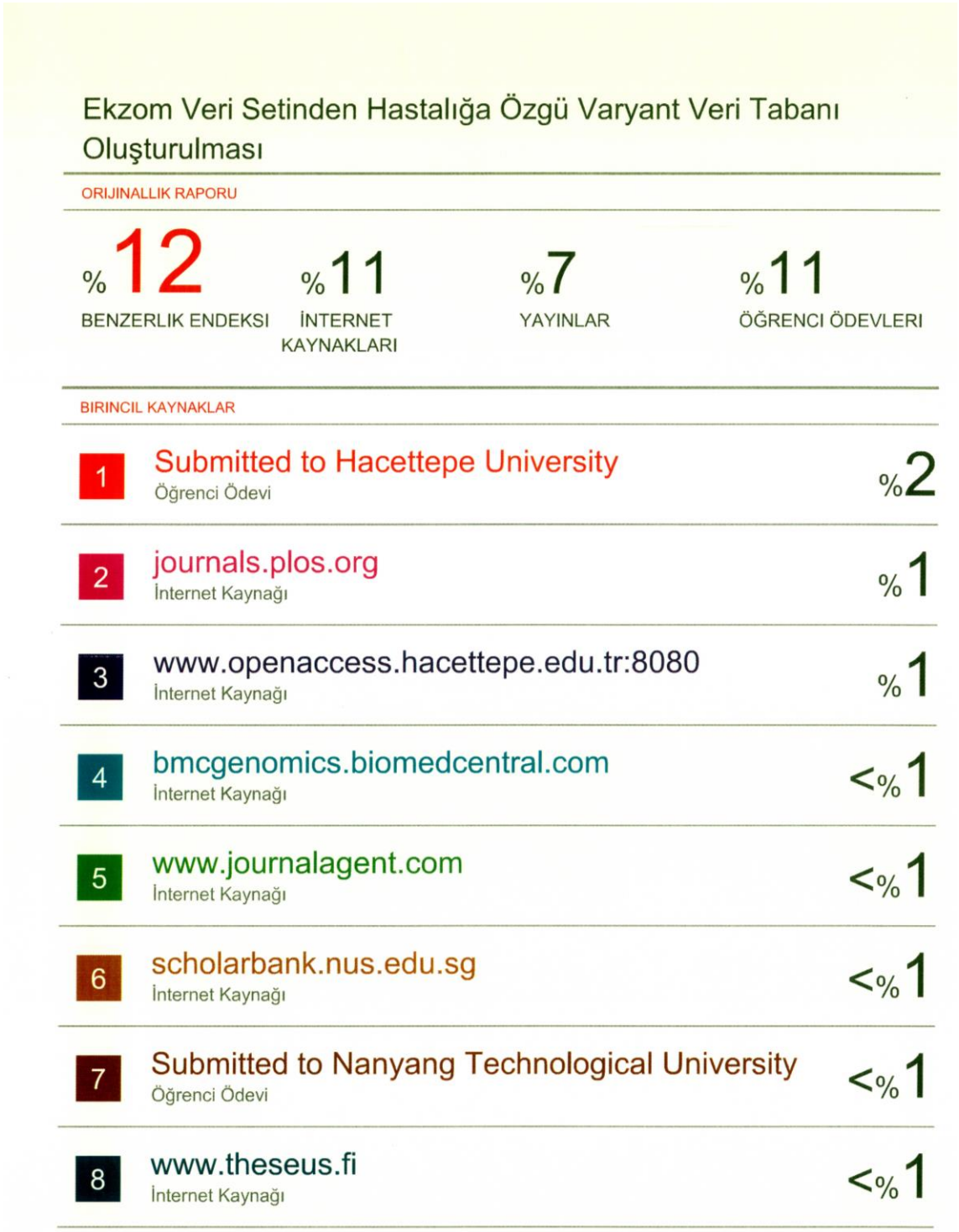
T.C. HACETTEPE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ
EKZOM VERİ SETİNDEN HASTALIĞA ÖZGÜ VARYANT
VERİ TABANI OLUŞTURULMASI
Yavuz Adabalı
Danışman: Prof. Dr. Ayşe Nurten AKARSU
İkinci Danışman: Dr. Öğr. Ü. İdil Yet

Bu tez çalışması 29/08/2019 tarihinde jürimiz tarafından "Biyoinformatik Programı"nda yüksek lisans tezi olarak kabul edilmiştir.

Jüri Başkanı: Doç. Dr. Yeşim AYDIN SON (Ortadoğu Teknik Üniversitesi)
Tez Danışmanı: Prof. Dr. Ayşe Nurten AKARSU (Hacettepe Üniversitesi)
Üye: Doç. Dr. Tunca DOĞAN (Hacettepe Üniversitesi)

Bu tez Hacettepe Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca yukardaki jüri tarafından uygun bulunmuştur.

Prof. Dr. Diclehan ORHAN
Enstitü Müdürü



Ekzom Veri Setinden Hastalığa Özgü Varyant Veri Tabanı Oluşturulması

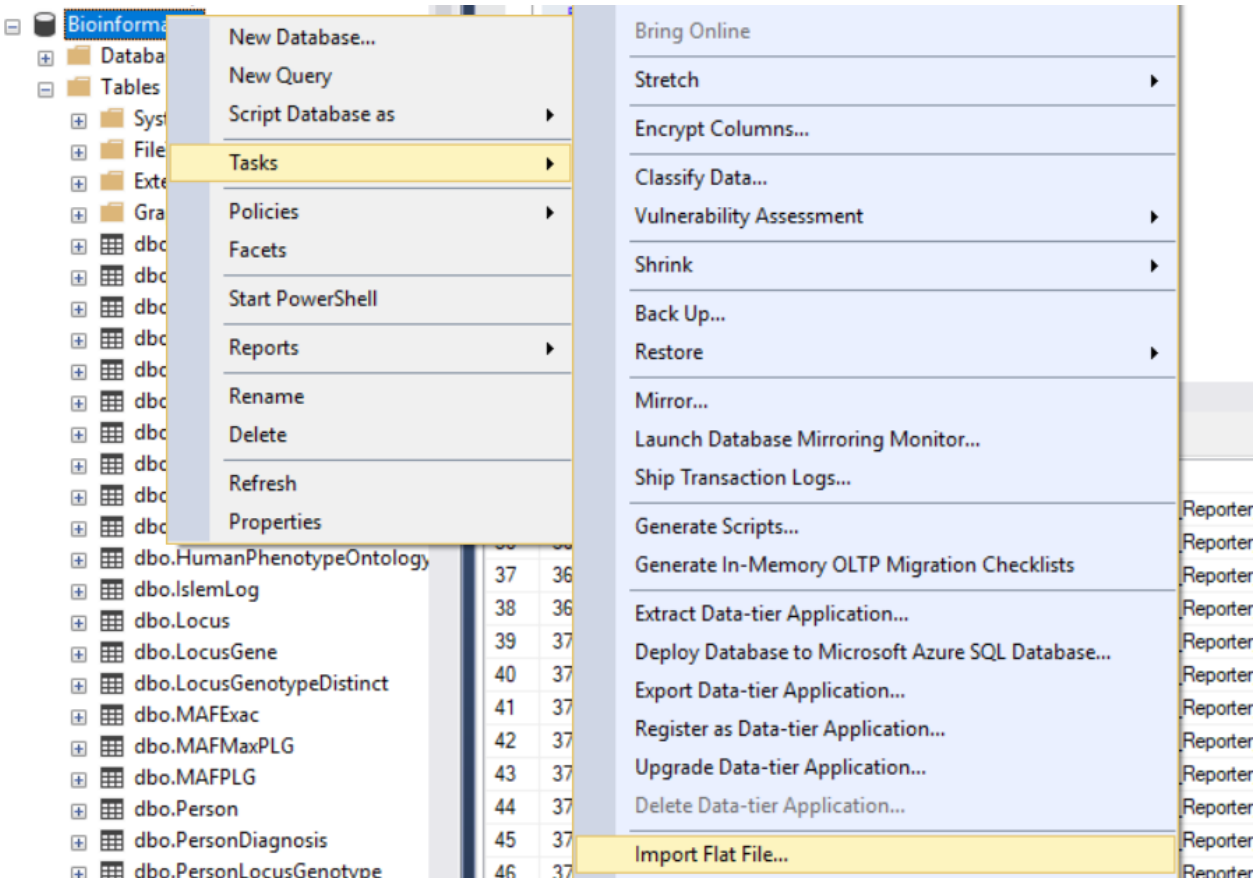
ORIJINALLIK RAPORU

%12	%11	%7	%11
BENZERLIK ENDEKSI	İNTERNET KAYNAKLARI	YAYINLAR	ÖĞRENCI ÖDEVLERI

BIRINCIL KAYNAKLAR

1	Submitted to Hacettepe University Öğrenci Ödevi	%2
2	journals.plos.org İnternet Kaynağı	%1
3	www.openaccess.hacettepe.edu.tr:8080 İnternet Kaynağı	%1
4	bmcgenomics.biomedcentral.com İnternet Kaynağı	<%1
5	www.journalagent.com İnternet Kaynağı	<%1
6	scholarbank.nus.edu.sg İnternet Kaynağı	<%1
7	Submitted to Nanyang Technological University Öğrenci Ödevi	<%1
8	www.theseus.fi İnternet Kaynağı	<%1

EK-3: MS-SQL Management Studio üzerinden csv dosyası aktarımı



EK-4: "tmpGrch38" tablosundan gen ile ilişkili tablolara aktarım yapan T-SQL sorguları

Söz konusu sorgular aşağıda "aktaran tablo -> aktarım yapılan tablo" şeklinde belirtilmiştir, güncelleme sorgularında ise tek bir alan ismi belirtilmiştir:

tmpGrch38 -> GeneType:

```
insert into GeneType(GeneType)
select trim(Gene_Type) from tmpGenesFromEnsemble group by Gene_type
```

tmpGrch38 -> GODomain:

```
insert into GODomain(GODomain)
select trim(GO_domain) from tmpGenesFromEnsemble group by trim(GO_domain)
```

GODomainID (Güncelleme):

```
UPDATE
    [tmpGrch38]
SET
    [tmpGrch38].GODomainID = god.ID
FROM
    [tmpGrch38] grch38
INNER JOIN
    GODomain god
ON
    grch38.GO_domain = god.GODomain;
```

tmpGrch38 -> GeneOntology:

```
insert into GeneOntology(GOTermDefinition, GOAccessID, GOAccessIDNumeric, GOTermName, GoDomainID)
select  Trim(SUBSTRING(go_term_definition, 1, CHARINDEX('"', go_term_definition) - 1)),
        Trim(GO_term_accession),
        Trim(SUBSTRING(GO_term_accession, CHARINDEX(':', GO_term_accession)  + 1,  LEN(GO_term_accession))) AS [Second],
        trim(GO_term_name),
        tmp.GODomainID
from [tmpGrch38] tmp where go_term_definition is not null group by   go_term_definition,GO_term_accession,GO_term_name,GODomainID
order by GO_term_accession
```

GeneType (Güncelleme):

```
----genetype
UPDATE tmpGrch38 SET tmpGenesFromEnsemble.GeneTypeID = gt.ID
FROM tmpGrch38 tmp
INNER JOIN GeneType gt ON tmp.Gene type = gt.GeneType;
```

tmpGrch38 -> Gene:

```
insert into Gene(GeneName, Chromosome,Strand,GeneTypeID, Grch37Start,Grch37End)
select
    Trim(gene_name),
        case Chromosome_scaffold_name
        when 'X' then 23
        when 'Y' then 24
        when 'MT' then 25
        ELSE Chromosome_scaffold_name
    end,
        case Strand
        when 1 then 1
        when -1 then 0
        ELSE Strand
    end,
    GeneTypeID
    ,Gene_start__bp_
    ,Gene_end__bp_
from tmpGenesFromEnsemble  group by   Gene_name,Chromosome_scaffold_name,
                                    Strand,GeneTypeID,Gene_start__bp_,Gene_end__bp_
having COUNT(*)=1
order by Gene_name
```

EK-5: “GetVariants” isimli fonksiyona çağrı yapan kod bloğu

Kullanıcının seçtiği parametreye göre filtre ile ilgili “*stored procedure”* sonucu dönen varyantları çağıran fonksiyon:

```
maxMaf = Convert.ToDouble(filters.Where(w => w.ID == "MaxMAF").Select(s => s.Value).FirstOrDefault());
var sql = "get" + variantFilter + "VariantsMaxMaf";
using (var connection = new SqlConnection(connStr))
{
    connection.Open();
    // commandType: CommandType.StoredProcedure
    return connection.Query<PersonLocusDetailListViewModel>(sql,
    new { PersonID = personID, DiagnosisID = diagnosisID, MaxMAF = maxMaf },
    commandType: CommandType.StoredProcedure, commandTimeout: 0).AsQueryable().ToPagedList(currentPage, pageSize, null);
}
```

9. ÖZGEÇMİŞ

KİŞİSEL BİLGİLER:

Adı ve Soyadı : YAVUZ ADABALI
Doğum Yeri ve Tarihi : Konya 15/03/1988
Uyruğu : T.C.
İletişim Adresi ve Telefonu : T.C. Aile, Çalışma ve Sosyal Hizmetler Bakanlığı B.İ.D.B.
Emek Mahallesi 17. Cadde No:13 Pk: 06520 Emek / ANKARA
0544 232 22 54

ÖĞRENİM DURUMU:

İlköğretim : Konya Namık Kemal İ.Ö.O. - 2002
Lise : Konya Karatay Lisesi (YDAL) Fen Bilimleri - 2006
Lisans : Karabük Üniversitesi Bilgisayar Müh. (%30 İngilizce) - 2014
Yüksek Lisans : Hacettepe Üniversitesi Biyoinformatik Yüksek Lisans Programı

MESLEKİ DENEYİMLERİ:

2010 Karabük Üniversitesi Uzaktan Eğitim Merkezi Kısmi Zamanlı Öğrenci
2011 Karabük Üniversitesi Uluslararası İlişkiler Ofisi Kısmi Zamanlı Öğrenci
2011 Centro Superior de Hostelería y Turismo de Valencia Stajyer
2012 Karabük Üniversitesi Bilgi İşlem Daire Başkanlığı Stajyer/ Kısmi Zamanlı Öğrenci
2015-Halen T.C. Aile, Çalışma ve Sosyal Hizmetler Bakanlığı Programcı

GÖREV ALDIĞI PROJELER:

1. Php ve MySQL İle Karayolu Seyahat Firmaları İçin e-Rezervasyon Sistemi
2. XML Parsing ile Mobil Güncel Döviz Kurları Uygulaması
3. Bütünleşik Sosyal Yardım Bilgi Sistemi
4. Aile Bilgi Sistemi
5. Aile Sosyal Destek Programı (ASDEP)
6. Şehit Gazi Bilgi Sistemi
7. Doğum Yardımı Bilgi Sistemi
8. Mevsimlik Gezici Tarım İşçilerinin Yaşam Şartlarının İyileştirilmesi Projesi (METİP)