

T.C.
KASTAMONU ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
GENETİK VE BİYOMÜHENDİSLİK ANA BİLİM DALI



GENETİK ANALİZLERDE KULLANILAN İSTATİSTİK
YÖNTEMLER

HÜSEYİN YILMAZ

YÜKSEK LİSANS TEZİ

PROF. DR. ORHAN KAVUNCU

MAYIS - 2024

KASTAMONU

TEZ ONAYI

Hüseyin YILMAZ tarafından hazırlanan “**GENETİK ANALİZLERDE KULLANILAN İSTATİSTİK YÖNTEMLER**” adlı tez çalışmasının savunma sınavı **27.05.2024** tarihinde yapılmış olup aşağıda verilen jüri tarafından oy birliği ile Kastamonu Üniversitesi Fen Bilimleri Enstitüsü **Genetik ve Biyomühendislik Ana Bilim Dalı Yüksek Lisans Tezi** olarak kabul edilmiştir.

Danışman	Prof. Dr. Orhan KAVUNCU Kastamonu Üniversitesi	
Jüri Üyesi	Prof. Dr. Mustafa Muhip ÖZKAN Ankara Üniversitesi	
Jüri Üyesi	Dr. Öğr. Üyesi Ferhat ULU Kastamonu Üniversitesi	

Jüri üyeleri tarafından kabul edilmiş olan bu tez Kastamonu Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulunca onanmıştır.

Enstitü Müdürü	Doç. Dr. Selçuk MEMİŞ	
----------------	-----------------------	-------

TAAHHÜTNAME

Bu tezin tasarımı, hazırlanması, yürütülmesi, araştırmalarının yapılması ve bulgularının analizlerinde bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu; ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını, bilimsel etiğe uygun olarak kaynak gösterildiğini bildirir ve taahhüt ederim.

Hüseyin YILMAZ

ÖZET

YÜKSEK LİSANS TEZİ

GENETİK ANALİZLERDE KULLANILAN İSTATİSTİK YÖNTEMLER

HÜSEYİN YILMAZ

KASTAMONU ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ

GENETİK VE BİYOMÜHENDİSLİK ANA BİLİM DALI

DANIŞMAN: PROF. DR. ORHAN KAVUNCU

Bilimin hemen her alanında araştırma sonuçlarının değerlendirilmesinde İstatistik yöntemlerin kullanılmasına ihtiyaç duyulur. Çünkü bilimsel araştırmaların çoğunda elde mevcut araştırma materyalinin, hipotetik bir popülasyonu temsil etmesi, o popülasyondan bizim araştırma materyalimiz gibi bir tesadüf örneği çekme ihtimali ne kadar büyükse o kadar mümkündür. Genetik araştırmalarda da, elde mevcut materyalin belirli bir genetik kurala uygunluğunun test edilmesi, farklı popülasyonların genetik yapılarının karşılaştırılması gibi istatistik yöntemlere başvurulmasını gerektiren analizler söz konusudur.

Bu tezin amacı, genetik çalışmalarda kullanılan istatistik yöntemleri bir araya getirerek sunmaktır. Bu amaca uygun olarak, genetik analizlerde kullanılan istatistik yöntemlerle ilgili kaynaklar taranmış ve belirli bir düzene sokularak sunulmuş, özellikle belirli DNA dizileri bakımından popülasyon içinde ve popülasyonlar arasında farklılıkları analiz etmek üzere geliştirilmiş olan testler üzerinde durulmuştur.

ANAHTAR KELİMELE: Genetik analiz, İstatistik yöntemler, Hudson–Aguade–Kreitman (HKA) testi, Mac Donald- Kreitman (MK) Testi, Tajima’nın D testi.

Mayıs 2024, 48 Sayfa

ABSTRACT

MSC THESIS

STATISTICAL METHODS USED IN GENETIC ANALYSIS

HÜSEYİN YILMAZ

**KASTAMONU UNIVERSITY INSTITUTE OF SCIENCE
DEPARTMENT OF GENETICS AND BIOENGINEERING
SUPERVISOR: PROF. DR. ORHAN KAVUNCU**

Statistical methods are needed to evaluate research results in almost every field of science. Because in most of the scientific research, the greater the probability of drawing a random sample from a hypothetical population, the possibility that our research material represent that population is greater. In genetic research also, there are analyzes that require the use of statistical methods, such as testing whether the existing research material fits a certain genetic rule and comparing the genetic structures of different populations.

The aim of this thesis is to gather and present the statistical methods used in genetic studies. According with this purpose, literatures related to statistical methods used in genetic analyzes were reviewed and presented in a certain order, especially the tests developed to analyze differences within and between populations in terms of certain DNA sequences were emphasized.

KEYWORDS: Genetic analysis, statistical methods, Hudson–Aguade–Kreitman (HKA) test, Mac Donald-Kreitman (MK) Test, Tajima’s D test.

May 2024, 48 Page

TEŞEKKÜR

Yüksek lisans eğitimi ve bu tezin yazım sürecinde bilgisini, desteğini esirgemeyen, fikirleri ve hayat görüşüyle bana çalışkan, adaletli, başarılı, erdemli biri olmayı öğreten değerli danışmanım Prof. Dr. Orhan KAVUNCU'ya,

Eğitim hayatım boyunca verdikleri desteği esirgemeyen annem, babam, büyükannem ve büyükbabama,

Manevi desteğini, sabrını, başarıma olan inancını hiç eksik etmeyen biricik eşim Sezen'e ve aile hayatımıza değer, önem ve mutluluk katan sevgili oğlum Halit Emin'e sonsuz teşekkürlerimi sunarım.

HÜSEYİN YILMAZ

Kastamonu, 2024

İÇİNDEKİLER

	<u>Sayfa</u>
TEZ ONAYI	ii
TAAHHÜTNAME	iii
ÖZET.....	iv
ABSTRACT	v
TEŞEKKÜR	vi
İÇİNDEKİLER	vii
TABLolar DİZİNİ	viii
SİMGELEr VE KISALTMALAR DİZİNİ	ix
1. GİRİŞ.....	1
2. KURAMSAL ÇERÇEVE	13
3. YÖNTEM.....	19
3.1 Student'in <i>t</i> Testi.....	28
3.2 Hudson – Kreitman – Aguade (HKA) Testi	29
3.3 Mac Donald – Kreitman (MK) Testi	31
3.4 Tajima'nın D Testi	33
4. SONUÇ VE TARTIŞMA.....	38
KAYNAKLAR	43
EKLER.....	46
EK A. Farklı Serbestlik Derecelerindeki χ^2 Değerleri ve İlişkili Olasılıklar.....	47
ÖZGEÇMİŞ.....	48

TABLÖLAR DİZİNİ

Sayfa

Tablo 3.1 MN kan gruplarının beklenen ve gözlemlenen değerleri.....	27
Tablo 3.2 A ve B gametlerine ait olasılıklar	28
Tablo 3.3 HKA testi veri tablosu	30
Tablo 3.4 Dmd lokusundaki 2 introna ait HKA testi verileri.....	30
Tablo 3.5 MK testi veri tablosu 1.....	31
Tablo 3.6 MK testi veri tablosu 2.....	32
Tablo 3.7 MK test örneği - 1.....	32
Tablo 3.8 MK test örneği - 2.....	33
Tablo 3.9 4 Allelli DNA dizisi örneği.....	36
Tablo A 1 Farklı serbestlik derecelerindeki χ^2 değerleri ve ilişkili olasılıklar	47

SİMGELER VE KISALTMALAR DİZİNİ

Simgeler

Δp	: Göçten önceki ve sonraki frekanslar farkı
$^{\circ}\text{C}$	: Celsius derece
μ	: Tüm nükleotid dizisi boyunca mutasyon oranı
Π	: Birçok dizi arasındaki ikili uyumsuzlukların ortalama miktarı

Kısaltmalar

Adh	: Antidiüretik hormonu
AIDS	: Acquired Immune Deficiency Syndrome
cm	: Santimetre
CRISPR	: Clustered Regularly Interspaced Palindromic Repeats
CTAB	: Setiltrimetilamonyum bromür
Dmd	: Duchenne muscular dystrophy
DNA	: Deoksiribo Nükleik Asit
HIV	: Human Immunodeficiency Virus
HKY	: Hasegawa – Kishino – Yano
HWE	: Hardy – Weinberg Dengesi
İng.	: İngilizce
kg	: Kilogram
LCT	: Laktaz lokusu
LD	: Linkage Disequilibrium
M.Ö.	: Milattan Önce
MAS	: Belirteç temelli seleksiyon
MK	: MacDonald- Kreitman
MS	: Multiple Skleroz
NCBI	: National Center for Biotechnology Information
NS	: Önemli değil
PCR	: Polimeraz Zincir Reaksiyonu
RNA	: Ribo Nükleik Asit
SNP	: Single Nucleotide Polymorphism
USA	: United States of America
vd.	: Ve diğerleri
WES	: Tüm Eksom Dizileme
yl	: Yüksek lisans

1. GİRİŞ

Günümüzde yaşayan ya da daha önce yaşamış tekhücreli canlılardan gelişmiş canlılara kadar var olan tüm canlılar arasında benzerlik ve farklılıklar bulunmaktadır. Canlılar âleminde gözlenen çeşitliliğin sebepleri araştırılırken bunların birbirlerinden farklılıkları kadar benzerlikleri de dikkat çekmektedir. Canlıların birbirinden neden farklı olduğu ya da hangi bakımlardan birbirine benzediği, genetik bilimi bünyesinde geliştirilmiş yöntemlerle araştırılmaktadır. Aralarında akrabalık olan canlılar birbirlerine daha fazla benzer. Bu benzerliğe kalıtım denilmektedir. Bu benzerlik akrabaların müşterek cedde sahip olmalarından kaynaklanmaktadır. Ancak benzerlik birebir aynı değildir. Tek yumurta ikizlerinin bile birbirine benzemeyen yanları vardır. Bu farklılıklara da, genel olarak varyasyon denilmektedir (Kavuncu, 2020).

Genetik bilimi, Evrim bilimi ile interaktif olarak canlılar âlemindeki bu benzerlik ve farklılıkları konu aldığı zaman, çalışma birimi popülasyondur. Popülasyon, konuya göre, bir tür olabilir, belirli bir tür içindeki ırklar olabilir, aynı ırk içinde farklı çevrelere yerleşmiş topluluklar olabilir. Ancak genetikte daha rafine bir kavram olarak popülasyon, belirli bir bölgede yaşayan belirli bir türden canlı topluluğudur. Bir popülasyona mensup bireyler arasındaki genetik farklılıklara polimorfizm, popülasyonlar arası farklılıklara ise çeşitlilik (İng. diversity) denir (Hartl ve Clark, 2007).

Fenotip ve genotip kavramları genetikte sık kullanılan kavramlar olup kısaca anlamlarını vermek gerekirse, Fenotip; üzerinde durulan özellikler bakımından canlının tespit edilen halidir (mor çiçek, sarı saç, ela göz, boy 182 cm, ağırlık 75 kg gibi). Fenotip, genlerin ve çevre etkilerinin bir fonksiyonu olarak belirlenir. Genotip; üzerinde durulan özellikler bakımından fenotipin nasıl olacağını belirleyen genler topluluğudur. Günümüzde yetişkin bir insanda etkisi belirlenen gen sayısı 35.000 civarında tahmin edilmektedir. Bu sayı tahmin edileceği üzere her geçen gün artmaktadır.

Biraz önce tanımlandığı gibi, belirli bir türden belirli bir zamanda belirli bir bölgede yaşayan canlı topluluğuna popülasyon denilmektedir. Ilgaz Dağı'ndaki kardelenler,

Van Gölü'ndeki sazan balıkları birer populasyon örneğidir. Populasyonlar arası ya da populasyonun farklı generasyonları arasındaki benzerlik ve farklılıkları Populasyon Genetiği inceler. Türler veya daha büyük canlı grupları arasındaki benzerlik ve farklılıkları çalışan, bunların nedenlerini araştıran bilim ise Evolüsyon (Evrim) bilimidir (Kavuncu, 2020).

Populasyon genetiği, başta Mendel Kanunları olmak üzere genetik ilkeleri populasyon bazında uygulayan bir alandır. Türlerde ve bir tür içindeki populasyonlarda zamanla oluşan değişiklikleri ve bu değişikliklere neden olan faktörleri inceler. Populasyon genetiğinin başlıca uygulama alanları;

- İnsanlar ve diğer canlılar arasındaki genetik çeşitliliği açıklama, araştırma, bu bilgileri analiz etme,
- Nesli giderek azalan canlıların çiftleşmelerini kontrol ederek bu türlerin neslinin korunmasına yardım etme,
- Tarımı yapılan bitkilerin, evcil hayvanların yaşama şartlarında iyileştirme yapma,
- Türler arası nükleotid ve protein akrabalık yakınlık derecesini tespit etme,
- Türlerle ait genlerde zamanla oluşan değişimi araştırma ve açıklama olarak sayılabilir (Hartl ve Clark, 2007).

Populasyon genetiği populasyonları gen düzeyinde analiz ederek populasyonları birbirleri ile karşılaştıran bir disiplindir. Bir populyasyona ait tüm bireylerin eksiksiz bir genetik analizini yapmak, her bireyin tüm genomunun nükleotid dizisini tespit etmek gibi masraflı ve karmaşık bir işi gerektirse de, bu analizler günümüzde yapılabilmektedir (Kavuncu, 2022). Bu analizleri yapmak teknolojinin ilerlemesi ve yeni yöntemlerin geliştirilmesi ile her geçen gün kolaylaşmaktadır.

Elektroforez yöntemi populasyon genetikçileri tarafından populasyonlara ait çeşitlilik ve polimorfizmi araştırma çalışmalarında kullanılır. Protein ve nükleik asitlerin ayırımında kullanılan bu yöntem elektrik alanına maruz kalan makro moleküllerin

hareketine dayanır. Jel içinde birkaç saat elektrik akımını uygulanan protein ve DNA (Deoksiribo Nükleik Asit) molekülleri elektrik alanına tepki göstermek suretiyle molekül ağırlığına bağlı olarak farklı süratlerde jel boyunca ilerler. İşlem bittiğinde moleküllerin yerleri tespit edilir. Bu yöntem gerek protein, gerekse DNA farklılıklarını tespit etmede kullanılmaktadır (Hartl ve Clark, 2007).

Canlılar arasındaki farklılıklara genetik biliminin ortaya çıkmasından çok daha önceleri de ilgi duyulmuştur. Bitki ve hayvan yetiştiriciliğinde bu farklılıklardan yararlanmak, genetiğin bir bilim dalı olarak ortaya çıkmasından belki de 10 bin yıl önce başlamıştır. İnsanlar önemli gördüğü, beğendiği, verimi yüksek bitki ve hayvanları daha çok tercih etmiş, damızlık olarak kullanmıştır. Böylelikle damızlıkların tercih edilen fenotiplerinin döllere aktarılacağı varsayılmıştır. Bu işleme sun'i seleksiyon (yapay seçim) denir (Kavuncu, 2020). Bu sayede üstün özellikli, verimi yüksek, insanlar tarafından daha çok tüketilen ya da alışverişte daha çok tercih edilen genotipler neslini devam ettirmiştir. İlgi görmeyen, verimi az, tüketimde tercih edilmeyen genotiplerin nesli ise giderek azalmış ve hatta yok olmuştur.

Tabii seleksiyon (doğal seçim) ise ortam şartlarına uyum sağlayan canlıların yaşamaya devam etmesi, uyum sağlayamayanların ise giderek yok olmasıdır. Örneğin sanayi devrimiyle oluşan endüstriyel kirlilik sonucu, siyah güveler ortamda kamufle olup yaşamaya devam ederken açık renkteki güveler, güveyle beslenen (avcı) organizmalar tarafından avlanarak sayılarında azalma ortaya çıkmıştır. Almanya, İngiltere ve Amerika Birleşik Devletleri'nde yoğun sanayileşme sonucu yakılan kömürün oluşturduğu is ve atıklar çevrenin kirlenmesine neden olmuştur. Daha önceleri ortamda kolay ayırt edilen siyah güveler değişen çevre şartları yüzünden, avcı organizmalar tarafından daha az görülür hale gelmiştir. Açık renkteki güvelerde ise tam tersi bir durum söz konusu olmuştur. Açık renkteki güveler sanayi devriminden önce ortamda güveyle beslenen (avcı) organizmalar tarafından siyahlara isteğe bağlı olarak tercih edilebiliyorken çevrenin kirlenmesinden sonra siyahlar görülemediği için avcılar tarafından mecburen daha çok tüketilmeye başlanmış ve sayılarında önemli ölçüde azalma meydana gelmiştir. Bu durumu açıklayan mekanizma endüstriyel melanizm olarak bilinmektedir (Hartl ve Clark, 2007).

Genetik bir bilim dalı olarak, 1865'te Gregor Mendel'in bezelyelerle (*Pisum sativum*) yaptığı deneme sonuçlarını yayınlaması ile doğmuştur. Mendel, genetik bilgilerin nesilden nesile kalıtım faktörleriyle aktarıldığını açıklamıştır (Hartl ve Clark, 2007).

Canlılar arasında var olan genetik çeşitlilik kalıtım deneyleri yapan Mendel'in yanı sıra Francis Galton'un da çalışma alanıydı. Galton, istatistiği genetik çalışmalarda ilk kullanan araştırmacılardan biridir. İnsanlarda saç rengi, göz rengi gibi fiziksel çeşitliliklerin yanı sıra davranışsal değişiklikleri de incelemiştir. Bu çalışmalarıyla Biyometri'nin babası kabul edilir. Canlılar arasında var olan bu genetik çeşitlilik cümlesinden olmak üzere;

- Allel frekanslarının tespiti ile türlerin alt grupları arasındaki benzerlik ve farklılıklar belirlenebilir.
- Genetik çeşitlilik tüm canlılar için iyi bir araştırma alanıdır.
- Çoğu canlı türünde üzerinde durulan özellikler bakımından genotip çoklu allele sahiptir. Diploit canlılarda her genotip iki allelden oluştuğu için tek yumurta ikizleri hariç (tek yumurta ikizlerinde bile bazı farklılıklar vardır.) iki birey aynı genotipte olamaz. Bu da DNA parmak izi gibi analizlerde araştırmacıya kolaylık sağlar (Hartl ve Clark, 2007).

1953 yılında James Watson ve Francis Crick tarafından DNA'nın üç boyutlu yapısı ve kendini kopyalaması moleküler düzeyde açıklanmasıyla Moleküler Genetik diye adlandırılan bir alt disiplin dalı ortaya çıkmıştır (Kavuncu, 2020). Moleküler genetik DNA'nın ve RNA (Ribo Nükleik Asit)'nin yapısını, işlevlerini, çeşitlerini ve görevlerini araştırır. DNA'nın kendini kopyalamasının nasıl gerçekleştiğini, protein sentezi sırasında bilginin nasıl işlendiği ve aminoasitlerin diziliminin nasıl yapıldığını da moleküler genetik inceler. Bu gibi incelemelerde de istatistik yöntemlerine ihtiyaç duyulur.

İnsanlarda boy uzunluğu, kilo, ten rengi gibi özellikler birçok genden etkilenir. Bazı ölçek ve ölçme araçları kullanılarak bu özellikler ölçülebilir. Nicel karakterler denilen bu özelliklerin kalıtımını Kantitatif Genetik inceler. Bitki ve hayvan ıslahı

alışmalarında sıklıkla kullanılan kantitatif genetik uygulamaları lokusların nicel karakterler üzerindeki etkisini açıklamaya alışmaktadır (Nielsen ve Slatkin, 2013).

Populasyon Genetiğinin önemli konularından biri de Hardy – Weinberg Dengesi (HWE)’dir. Bu denge halinde generasyonlar boyunca süreklilik gözlenen geniş populasyonlarda, seçim, mutasyon ve göçün olmadığı, ayrıca çiftleşmelerin rastgele gerçekleştiği kabul edilir. Bir populasyon bu denge halindeyken populasyonun genetik yapısı, nesilden nesile değişmeden kalır.

Bir populasyonda daha önce rastgele çiftleşme olup olmadığına bakılmaksızın rastgele çiftleşmenin olduğu ilk generasyonla birlikte genotip frekansları sabit hale gelir. Rastgele çiftleşmenin bu ilk generasyonunda ve devam eden generasyonlarda, A geninin frekansı p , a allelininki de $q = 1 - p$ olmak üzere, genotip frekansları;

- AA için p^2 ,
- Aa için $2pq$,
- aa için ise q^2 olur (Kavuncu, 2022).

Bu denge durumuna 1908 yılında birbirinden bağımsız araştırma yaparak bu sonuca ulaşan iki araştırmacının adı verilmiş ve Hardy – Weinberg Dengesi olarak literatürdeki yerini almıştır. Bu denge halini yine diğerlerinden bağımsız olarak Rus genetikçi Chetverikov da bulmuştur (Kavuncu, 2022).

Aslında HW dengesinde iki ayrı dengeden bahsedilmektedir:

- Rastgele çiftleşme sonucu gen veya genotip frekanslarında generasyonlar boyunca oluşan dengedir. Populasyonlarda rastgele çiftleşme olmadığı zaman gen frekansları değişmezken, genotip frekansları değişir. Bu durumda rastgele çiftleşme genotip frekanslarının nesiller boyunca değişmemesini sağlar.

- HW dengesinde diğer denge durumu gen frekansları ve genotip frekansları arasındaki yukarıda bahsedilen ilişki vardır. Bu denge halinde homozigotların frekansıyla heterozigotların frekansı arasında,

$$4f_1f_3 = (f_2)^2 \quad (1.1)$$

eşitliğiyle gösterilebilecek bir ilişki vardır. Bu eşitlikte $f_1 = p^2$, $f_2 = 2pq$ ve $f_3 = q^2$ yi temsil etmektedir. Bu eşitlik göç, mutasyon ve seleksiyonun etkili olmadığı, sonsuz sayıda bireyden oluşan popülasyonlarda rastgele çiftleşmenin söz konusu olduğu, her lokus için geçerlidir (Kavuncu, 2022).

Rastgele çiftleşmeden sapmalar olması popülasyonun HW dengesinin bozulmasına sebep olur. Çiftleşmelerin rastgele olmadığı popülasyonlarda çiftleşmeler fenotip olarak birbirine benzeyenlerin ve birbirine benzemeyenlerin çiftleşmesi şeklinde gruplanabilir. Assortative çiftleşme denilen ve çiftleşmenin fenotipe göre olduğu bu çiftleşmelerde birbirine benzeyen fenotipteki bireylerin çiftleşmesi pozitif assortative çiftleşme, birbirine benzemeyen fenotipteki bireylerin çiftleşmesine ise negatif assortative çiftleşme denir. Benzeyenlerin çiftleşmesi, homozigotluğu artırırken heterozigotluğu azaltır. Benzemeyenlerin çiftleşmesi, heterozigotluğu artırırken, homozigotluğu azaltır. Rastgele çiftleşmede de fenotip olarak birbirine benzeyen ve benzemeyenler arasında tamamen şansa bağlı rastgele çiftleşmeler olur; ancak bunun ihtimali assortative çiftleşmede şansa göre olandan daha fazladır. Rastgele çiftleşmeden sapmanın etkisi, popülasyonun gen frekansında değişme olması şeklinde değil de genotip frekanslarında değişme olması şeklinde olur (Kavuncu, 2022).

Popülasyonlar arasında da çiftleşmeler olur. Gerçekte coğrafi olarak yakın olan popülasyonlardan bireylerin çiftleşme olasılığı coğrafi olarak uzak olan bireylerin çiftleşme olasılığından daha fazladır. Coğrafi yapıdan kaynaklı rastgele çiftleşme olasılığının azalması, bir şekilde alt grupların birbirinden izolasyonuna, popülasyonun alt bölümlere ayrılmasına neden olur. Rastgele çiftleşen bir popülasyondan coğrafi izolasyonla birbirinden ayrılan bu alt popülasyonlarda, heterozigotluk azalır, homozigotluk artar. Buna Wahlund Etkisi denir. Böylelikle heterozigotluk beklenen HW frekanslarının altına düşer (Nielsen ve Slatkin, 2013).

Bir populasyonda gen frekansları ise evölüsyoner amiller diye adlandırılan göç, mutasyon, seleksiyon ve şans sonucu değişir. Göç, mutasyon ve seleksiyon, etkili oldukları populasyonda yaptıkları değişim miktarı ve bu değişikliğin yönü belirlenebildiği için sistematik etkili amiller olarak adlandırılır. Şans ise dispersif etkili amil olarak bilinir çünkü etkili olduğu populasyonda yaptığı değişimin miktarı belirlenebilirken bu değişimin yönü belirlenemez (Kavuncu, 2022). Tüm bu etkilerin tahmin edilmesinde de istatistik bilgilere ihtiyaç vardır. Evölüsyoner amilleri kısaca açıklayacak olursak;

➤ Göç: Bir populasyondan ayrılan alt populasyonlar, şansın ve coğrafi izolasyonun etkisiyle farklı genetik yapılara evrilirler. Bu alt populasyonlara deme (kapalı grup) denir. Bu alt populasyonlarda göçün etkisiyle gen frekansları göç edilen grupta değişir. Göçün yaptığı etki m sayısı ile hesaplanır. m ; herhangi bir genin muhacir (göç eden) olma ihtimalidir.

p_m ; muhacir olan genler arasından herhangi birinin nispi frekansı olsun. Göçün o genin frekansı p olan bir populasyondaki etkisi göçten sonraki ve önceki frekanslar farkı Δp olarak;

$$\Delta p = p_{(1)} - p = m \cdot (p_m - p) \quad (1.2)$$

(1.2) eşitliği ile bulunur (Kavuncu, 2022).

➤ Mutasyon: Genetik materyalde meydana gelen kalıtsal değişikliklerdir. Mutasyon, genetik çeşitliliğe sebep olan birinci etmendir. Genetik varyasyon, kromozomların sayısı ve yapısında meydana gelen değişikliklerin yanında aslında nokta mutasyonu veya gen mutasyonlarından kaynaklanmaktadır. Moleküler seviyedeki bu değişiklikler, DNA nükleotid dizilerinde bazı dizilerin yer değiştirmesi, silinmesi veya başka dizilerin yerine geçmesi şeklinde olabilir. Kromozom düzeyinde ise translokasyon denilen homolog olmayan kromozomlar arasında gerçekleşen parça değişimi, parça azalması (delesyon), parça çoğalması (duplikasyon) veya parçanın ters dönmesi (inversiyon) şeklinde olabilmektedir. Gen mutasyonları oranı bir gen düzeyinde her generasyonda 10^{-4} ile 10^{-6} arasında olup, bu oran çok düşük miktarda olduğu için allel frekansını değiştirme ihtimali birkaç generasyonluk küçük bir

zaman diliminde yok sayılabilir, ancak uzun bir zaman diliminde bu düşük değişiklikler birikerek etkili olur. Mutasyon, yabani tip ve mutant tip arasında tek yönlü ve çift yönlü olma durumuna göre çalışılmıştır (Kavuncu, 2022).

- **Seleksiyon (Seçilim):** Bazı genotiplerin diğer genotiplere nazaran daha çok döl oluşturma imkânına sahip olmasıdır. Bu durum doğal faktörlerin etkisinde insan müdahalesi olmadan gerçekleşiyorsa buna tabii seleksiyon denir. İnsanların kültüre aldığı türlerde renk, verim veya ekonomik üstünlüğü olan bireyleri çoğaltmasına suni seleksiyon denir. Bir popülasyondaki bireylerden herhangi birinin bir sonraki generasyona olan genetik katkısına fitness ya da Türkçe'deki anlamıyla selektif adaptif değeri denir. Özetle; seleksiyon tek yönlü ise popülasyon homozigotlaşır; heterozigotlar lehine ise polimorfik bir dengeye ulaşır (Kavuncu, 2022).
- **Şans:** Popülasyonlar model kurma çalışmalarında sonsuz büyüklükte varsayımlarına rağmen aslında sonsuz büyüklükte değildir. Popülasyonlar var olan büyüklüklerine göre ertesi nesilleri oluşturmak için döl verecek bireylerin örneklenmesinden kaynaklı bir şans faktörüne tabidir. Şansın gen frekansı değişikliğine olan etkisi, allel frekansı ve popülasyonu oluşturan birey sayısına bağlıdır. Az sayıda bireyden oluşan popülasyonda şans etkisi daha çoktur. Bir allelin frekansı ne kadar küçükse bulunduğu popülasyonda tamamen yok olma ihtimali o kadar yüksektir (Kavuncu, 2020).

Hardy - Weinberg dengesine göre bir popülasyonda allel frekanslarının bir nesilden sonraki nesile aynı kalabilmesi için popülasyonun sonsuz büyüklükte olması gerekir. Ancak gerçekte doğadaki popülasyonların genişliği (popülasyondaki birey sayısı) sonsuz değildir. Popülasyon genişliği sonsuz olmayan bu popülasyonlarda sonraki nesli oluşturacak gametler popülasyonun gamet havuzundan rastgele seçildiğinde örnekleme hatası nedeniyle allel frekanslarının nesiller arasında şans eseri farklı olmasına rastgele genetik sürüklenme (İng. random genetic drift) ya da kısaca sürüklenme (İng. drift) denir. Bu sürüklenme yüzünden bir allelin popülasyondan tamamen kaybolması veya sabitlenmesi mümkündür (Griffiths vd. 2012). Allel frekanslarındaki değişim ile ilgili bir model olan Wright – Fisher Modeli bakterilerin çoğu için geçerlidir. Çünkü eşeysiz üreyen, haploit (genetik materyalin bir kopyasını

içeren) canlılarda uygun bir modeldir. Bu modelden yararlanarak allel frekanslarındaki değişim (1.3) eşitliği ile belirlenebilir (Nielsen ve Slatkin, 2013).

$$E[f_A(t+1)] = 2Nf_A / 2N = f_A(t) \quad (1.3)$$

N; populasyonun birey sayısı,

2N; gen kopya sayısıdır.

t + 1 neslinde A allelinin frekansı, t neslindeki allel frekansına eşit olsun beklenir. Ancak bu ortalama bir değerdir; populasyondan her seferinde 2N kadar gametlik örnekler çekilse, her örnekte allel frekansı eşitlik (1.3) ile verilen değer etrafında değişik değerler alacaktır; tesadüfi sürüklenme sebebiyle oluşan değişiklik A alleli lehine de a alleli lehine de olabilir. Ancak doğal seçimde bir allel lehine olma durumuna daha çok rastlanır (Nielsen ve Slatkin, 2013).

Canlılarda meydana gelen mutasyonlar populasyonlardaki heterozigotluk oranını ve polimorfizmi artırır. Bu artış miktarını ilgili lokusta bulunan allel sayısı ve bu allellerin frekansı belirler. DNA dizilerinde yer alan her nükleotidin mutasyona uğrama olasılığı vardır. Örneğin; herhangi bir gen 500 aminoasitlik protein dizisi kopyalıyorsa 1.500 nükleotidlik DNA dizisine sahip demektir. Nükleotid dizileri 4 çeşit bazdan (Adenin, Timin, Guanin, Sitozin) oluştuğuna göre bu durumda mümkün olan allel sayısı $4^{1.500} = 2^{3.000} \approx 10^{903}$ tanedir. Canlılardaki oluşan her yeni mutasyonun o populasyonda bulunmayan bir allel oluşturma ihtimaline dayanan Kimura'nın önerdiği bu modele mutasyonun sonsuz sayıda allel modeli (İng. Infinite allele model) denir. Bu modelde sadece müşterek ceden dolayı özdeş olan iki allel benzer olabilir. Bu modele göre protein sentezinde temel birim gendir (Kavuncu, 2022). Birçok genin populasyondaki bireylerde bulunan ikiden fazla alleli vardır. Mutasyon etkisi göz önüne alındığında genetik çeşitlilik heterozigot genotip oranına bakılarak belirlenir. Bir lokusta mutasyon etkisinin oluşturduğu heterozigotluktan farklı miktarda heterozigotluk varsa diğer güçlerin özellikle doğal seçilimin etkili olduğu düşünülür. Genlerin heterozigot olma durumu allel sayısına bağlıdır. Genlerin mümkün olan allel sayısı populasyondaki birey sayısından genellikle çok olduğu için oluşan yeni

mutasyonların daha öncekilerden farklı bir allel oluşturduğu söylenebilir (Hartl ve Clark, 2007).

Elektroforez tekniğinin protein polimorfizmi belirleyebildiği dönemde sonsuz sayıda allel modeline ait istatistikler kullanılmaktayken DNA polimorfizmini belirleyebilecek düzeye gelmesiyle sonsuz sayıda site modeline ait istatistikler kullanılmaya başlanmıştır. Sonsuz sayıda site modelinde protein sentezinde temel birim nükleotidlerdir. Nükleotid polimorfizmi çalışma alanıdır. Sonsuz sayıda allel modeli ile arasındaki fark, elektroforez yönteminin gelişim kat etmesiyle olmuştur (Kavuncu, 2022). DNA dizi incelemeleri için ilk defa Kimura ve Waterson'un oluşturduğu bu modelde DNA molekülündeki uzunca bir nükleotid dizisi üzerinde çalışılır. Meydana gelen her bir mutasyonun sadece bir nükleotid bölgesinde değişiklik oluşturduğu kabul edilir. Bu model moleküler genetik alanında çalışma yapan araştırmacıların elde ettiği verilerin analizinde kullanılır (Hartl ve Clark, 2007).

Bu tezin konusu genetik analizlerde kullanılan istatistik yöntemler olduğu için, giriş bölümünde istatistikle ilgili temel bilgilere de kısaca temas etmek yararlı görülmüştür. İstatistik kelimesi, maksada göre 3 ayrı anlamda kullanılır. 1- Bir bilim dalı olarak meselâ, bilimsel araştırmalardan elde edilen verilerin toplanması, düzenlenmesi, sınıflandırılması ve bu verilerden gerekli değerlendirmeler yapılarak belirli sonuçlara ulaşılmasını sağlayan metotların tümüne istatistik denir. İstatistik sayesinde bilimsel araştırmalardaki veriler hem düzenli bir şekilde sunulabilir, hem de grafik ve tablolar sayesinde veriler arasındaki ilişkiler daha net görülebilir. İstatistik metotlar ucuz, daha hızlı ve güvenilir olması sebebiyle günümüzde yaygın bir şekilde kullanılmaktadır. Ekonomide, eğitimde, sağlıkta, askeri alanda ve genetik çalışmalarda istatistik yöntemlere başvurulur. 2- Geçmiş ve şimdiki olayları inceleyerek gelecek hakkında tahmin yapmayı sağlayan bir bilim dalı olduğu için çeşitli olaylardan toplanan veriler istatistiğin çalışma materyalini oluşturur. İstatistikte elde edilen verilere de istatistik denilmektedir; futbol maçındaki atılan gol sayısı, bir ildeki ya da ülkedeki kişi sayısı, bir toplumda aşı olan bebek sayısı, bir hastanede yatmakta olan hasta sayısı, vb. gibi... 3- İstatistik yöntemler, elde mevcut ünitelerden (örnekten) elde edilen verilerden, o ünitelerin temsil ettiği daha büyük ünite topluluğuna (populasyona) ait çıkarımlar yapmaya çalışır; bunun için örnekten hesapladığı değerleri populasyon değerlerine

genellemeler yapar. Örnekten hesaplanan bu değerlere de istatistik denilir (Kavuncu, 2021).

İstatistik yöntemler, kullanım alanlarına göre iki gruba ayrılır:

- **Deskriptif İstatistik:** Günlük hayattaki olay ya da durumlardan elde edilen verilerin sınıflandırılması, grafik, tablolar yardımıyla sunulması, ortalama, varyans ve standart sapmalarının hesaplanarak ifade edilmesi bu istatistik alanının konusudur (Düzgüneş ve Kavuncu, 1978).
- **Karar Verme İstatistiği (Inferential Statistics):** İstatistik için, belirsizlik olan durumlarda karar verme ilmi diyebiliriz. Bir populasyon hakkında çıkarımda bulunabilmek için o populasyondan tesadüfi olarak yeterli sayıda ünitelerden oluşan örnek seçilir. Her ünitenin seçilme ihtimali eşit ve seçilen bir ünitenin diğer ünitelerin seçimini etkilemediği kabul edilir. Bu seçilen tesadüfi örnekten elde edilen bilgilere göre populasyon hakkında karar verilir (Düzgüneş ve Kavuncu, 1978).

İstatistiğin tarihsel gelişimine baktığımızda istatistiğin bir bilim olarak 19. Yüzyılda ortaya çıktığını görmekteyiz. Devletlerin sahip olduğu asker sayısı, ülkelerin nüfus sayımları, tahsil edilecek vergi miktarı istatistik kapsamındaki ilk çalışmalar olmuştur. M.Ö. 3.000'ler de Mısır'da piramitlerin inşaatında çalışacak işçi sayısını tespit edebilmek için ilk nüfus sayımı yapılmıştır. Yine Osmanlı'da da belirli aralıklarla nüfus sayımı yapılmıştır. İstatistiğin hızlı gelişimi 17. Yüzyılda Fransa'daki mali ve ticari gelişmelerde istatistiğin önem kazanmasıyla olmuştur. Alman üniversitelerinde yer alan 'Devletlerin Özellikleri' adlı derste ülkelerin mali, askeri yapıları işleniyordu. İstatistik (statistik) kelimesi de buradan yani devlet (statüs) kelimesinden gelmektedir (Terzi, 2018). Devlet adına bu bilgileri toplayan kişilere ABD'de statist deniliyordu (Düzgüneş ve Kavuncu, 1978).

Belirli bir olay aynı şartlar altında tekrarlanan denemelerde aynı sonuçları veriyorsa bu olay kesin olaydır. Aynı şartlar altında bu olay defalarca gerçekleştiğinde elde edilen sonuçla olay öncesi tahmin edilen sonuç arasında benzerlik vardır. Modelin durumunu bozmayacak kadar küçük farklar da oluşabilir. Örneğin; saf su deniz

seviyesinde 1 atmosfer basınçta 100 °C’de kaynar. Aynı şartlar altında defalarca ölçüm yapılırsa küçük farklar oluşabilmekle birlikte aynı sonuca ulaşılır. Belirli şartlarda tekrarlanan bazı olaylarda ise olayın sonucu önceden tam net belirlenemiyorsa bu olaylara tesadüf olaylar denir: bir toplumdan seçilen yüz kişiden kaçının kanser olduğu, haziran ayında esecek en şiddetli rüzgârın şiddeti, bir hastaneye aralık ayında başvuru yapacak acil vaka sayısı gibi... İstatistiğin konusu işte bu tesadüfî olaylardır. Yani, aynı şartlar altında yapılan gözlemlerde farklı sonuçlar veren durumlar istatistiğin konusudur. İstatistik bu olaylardan elde edilen verileri frekans tabloları ya da grafiklere yerleştirerek verilerin düzenlenmesini sağlar ve bu verilerden yapılacak çıkarımların genellenebilme ihtimalini hesaplar (Kavuncu, 2021).

İstatistik yöntemler sayesinde daha hızlı, güvenilir, maliyeti daha az sonuçlar elde edilir. Sadece genetik alanında değil, sağlık, siyaset, fizik, kimya, ekonomi alanlarında da istatistik çalışmalara gereksinim vardır. Yapılan çalışmalardan toplanan verilerin düzenlenmesi, aritmetik ortalama, standart sapma gibi tanımlayıcı istatistiklerin hesaplanması, verilerin frekans tabloları ve grafik halinde gösterimi yanında hasta-kontrol grubu gibi karşılaştırmalar için istatistik bilimine ihtiyaç vardır. Günümüzde ise bu işlemlerin daha hızlı yapımına olanak tanıyan bilgisayar istatistik paket programları geliştirilmiştir (Kavuncu, 2020).

Bu tez, bir derleme çalışmasıdır. Genetik analizlerde kullanılan istatistik yöntemlerle ilgili kaynaklar taranmış ve belirli bir düzene sokularak sunulmuştur. Dolayısıyla bu tezin amacı, bu yöntemlerin derli toplu bir sunumunu yaparak literatürdeki bir boşluğu doldurmaktır.

2. KURAMSAL ÇERÇEVE

Günümüzde genetik analizler başlığı altında yapılan incelemeler sonucunda Türkçe ve yabancı dillerde yazılmış birçok kaynak mevcut olduğu görülmektedir. Mendel ile başlayan bu araştırma ve çalışmalar üzerinde durulan kalıtım faktörlerinin (genlerin) nesilden nesile nasıl aktarıldığına yönelmiş, kalıtım faktörlerinin (genlerin) fiziki yapısının DNA olduğu anlaşıldıktan sonra, özellikle 1953 yılında James D. Watson ve Francis Crick tarafından DNA'nın moleküler yapısını ve nesilden nesile geçişini sağlayan replikasyon modeli ortaya konduktan sonra, genin yapısı ve fonksiyonu ile ilgili çalışmalar hız kazanmıştır. 1977 yılında Frederick Sanger tarafından geliştirilen yöntemle DNA Dizileme keşfedilmiştir. 1976-1977 yıllarında ise Allan Maxam ve Walter Gilbert tarafından farklı bir DNA Dizileme yöntemi geliştirilmiştir. Yine 1980'lerde geliştirilen Floresan In Situ Hibridizasyon (FISH) tekniği de genetik çalışmalarda sıklıkla kullanılan bir yöntemdir. Genetik alanında Polimeraz Zincir Reaksiyonu (PCR) tekniği de oldukça önemli bir gelişmedir. Bu teknik DNA'nın bir bölümünün belirle koşullarda çoğaltılması prensibine dayanır. Bu keşiflere paralel olarak gelişen genetik analizler daha da önem kazanmıştır. 1990 yılında başlanıp 2003 yılında tamamlanan İnsan Genom Projesi ile insanın gen haritası çıkarılmıştır. Bu proje ile anemi, otizm, Alzheimer, kanser gibi hastalıkların nasıl oluştuğu, hangi genlerdeki farklılıkların bu hastalıklara yol açtığı ve bu hastalıkların tanı ve tedavi süreci daha iyi anlaşılmaya başlanmıştır. Yapılan genetik analizlerin sonucunda bunun gibi birçok hastalığa ilaç geliştirilmesinin önü açılmıştır (URL-1, 2023).

Genetik araştırmalarda hastalık için tanı koyma ve hastalığı tedavi etmek için zamanlamanın çok büyük önemi vardır. Kaymaz (2022) yaptığı çalışmada Sanger Sekanslama, Tüm Ekzon Dizileme (WES) gibi genetik yöntemlerin Primer İmmün Yetmezliği hastalığının tanı ve tedavisindeki önemine değinmiştir. Bu hastalıkta erken tanının çok önemli olduğunu, erken tanı için de genetik araştırmanın gerekliliğine vurgu yapmıştır. 1.407 hasta ile yapılan tanı ve tedavi sürecinde genetik testler yüksek başarı sağlamıştır (Kaymaz, 2022).

Paketli gıdalardaki genler analiz edilerek bir ürünün doğallığı hakkında bilgi sahibi olunabilir. Demirhan (2017)'ın yaptığı çalışmada yemeklik patates ve paketli gıdalardaki işlenmiş patates ürünleri içeriğindeki genler analiz edilmiştir. Çalışmada PCR tekniği kullanılmıştır. Çalışmada kullanılan ürünlerdeki DNA izolasyonu için CTAB metodu uygulanmıştır. Gıdalarda genetiği değiştirilmiş organizma tespitinde kalitatif PCR yönteminin kullanışlı olduğu vurgulanmıştır. Ancak ürün içeriğinde yabancı genin olup olmadığını tam olarak anlayabilmek için sekans analizi yapılması ve Real Time PCR tekniği kullanımının gerekliliğine işaret edilmiştir (Demirhan, 2017).

DNA dizi analizi yöntemi genetik çalışmalarda sıklıkla kullanılır. Özen (2020) Osteogenezis İmparfekta tanısı koyulan hastalarda dizi analizi yöntemi ile bu hastalığa neden olan günümüzde henüz bilinmeyen genleri tespit etmeye çalışmıştır. Hastalığa tanı koymada genetik analizlerin önemli olduğunu aktarmıştır. Fransa, Kore gibi diğer ülkelerde de benzer çalışma yapılmıştır. Akraba evliliği ve ailedeki bireylerde bu hastalığın olma ihtimali hastalığındaki etkenler arasında ilk sıralarda gösterilmiştir. Bilinmeyen gen tespiti için tüm genomun dizilenmesine ihtiyaç duyulduğu vurgulanmıştır (Özen, 2020). Bu da bize genetik çalışmaların ne kadar önemli olduğunu göstermektedir.

Yılmaz (2010) doktora tezinde mikroarray tabanlı DNA yeniden dizileme çipi ve direkt DNA dizi analizi yönteminin Wilson hastalarının genetik analizinde başarılı sonuçlar verdiğini aktarmıştır. Bu çalışmada bir gen üzerinde Wilson hastalığına sebep olan çok sayıda mutasyon taraması yapılmıştır. Yılmaz (2010) mikroarray tabanlı yeniden dizileme yönteminin genetik çalışmalarda hızlı ve kullanışlı bir teknik olduğunu vurgulamıştır.

Canlı organizmaların melezleri arasındaki benzerlikler genetik çalışmalarla tespit edilebilmektedir. Esmeray (2016) mısır melezlerinin genetik özelliklerini moleküler düzeyde analiz etmiştir. Mısır melezlerinin genetik açıdan benzerlik ve uzaklıklarını tespit ederek verim ve diğer özellikler açısından incelemelerde ve tespitlerde bulunmuştur (Esmeray, 2016).

Gen tedavisi ya da gen terapisi hastalıkların tedavi sürecinde, hastalıklara karşı ilaç geliştirmede giderek önem kazanan ve genetik çalışmaların bu alana yöneldiği bir tekniktir. İleride birçok hastalığın tedavisinde standart bir uygulama olması beklenmektedir. Hedef hücrelere DNA molekülü viral ve non-viral vektörler aracılığıyla aktarılır. Bu tedavideki en önemli aşama da bu aktarım işlemidir (Samsunlu, 2021).

Somatik hücrelerde (vücut hücresi) yapılan gen tedavisi sadece tedavi edici genin aktarıldığı somatik hücreleri etkiler ve dolayısıyla kalıtsal olmadığı için gelecek nesillere aktarılmaz. Uygulama şekillerine göre dört çeşittir;

- Hastalıklı gene dokunmadan tedavi edici genin aktarılması,
- Mutant genin normal gen ile değişimi,
- Hastalık oluşturan genin aktivitesinin engellenmesi,
- Hastalık oluşturan genlerin ön ilaç verilerek tedavi edilmesidir (Samsunlu, 2021).

Üreme hücrelerine (sperm, yumurta) hastalığı tedavi eden genin aktarılmasıyla yapılan tedavi ilgili bireyi ve kalıtsal olduğu için ondan sonraki nesilleri de etkiler. Çoğu ülkede üreme hücrelerine gen aktarımı etik olmadığı için kabul görmez. Bu nedenle günümüzde gen tedavisinin somatik hücrelerde uygulanması yasaldır. Gen terapisi alanında en çok çalışma yapan ülke Amerika'dır (Samsunlu, 2021).

DNA dizileme tekniklerinden sadece insan sağlığı alanında değil, tarımı yapılan hayvan ve bitkilerde de yararlanılmaktadır. Gedik ve Kavuncu (2021), keçilerde yaptıkları bir çalışmada, Türkiye'de yetiştirilen bazı keçi ırklarında Mucopolysaccharidosis IIID (MPS IIID) veya Sanfilippo D sendromu olarak bilinen kalıtsal hastalığa yol açan mutasyonun olup olmadığını çalışmışlardır. Buğdayda yapılan birçok SNP (Single Nucleotide Polymorphism) belirleme çalışması, Can (2023) tarafından hazırlanan yüksek lisans tezinde özetlenmiştir. Gündüz (2023) tarafından hazırlanan başka bir yıl tezinde, pamukta yaprak solgunluğu (Verticillium wilt) hastalığı ile ilgili bir çalışmada (Bardak vd. 2021), belirteç temelli seleksiyon

(MAS) için kullanımı mümkün olan ve yaprak solgunluğu da dâhil, birçok hastalığa dayanıklılık/hassaslık veya tolerans bakımından varyasyonla bağlantı halinde olabilecek birçok SNP belirteç belirlenmiştir.

2020 yılında Nobel Kimya Ödülü kazandıran bir çalışma olan CRISPR-CAS 9 tekniği DNA'nın istenen bölümünden kesilip yeniden yapıştırılmasına olanak tanır. Canlı genomlarında düzenleme yapılmasına imkân veren bir tekniktir. Kanser, Orak hücreli anemi, Multiple Skleroz (MS) gibi genetik hastalıklarla ilgili araştırmalarda ve tedavilerinde kullanılmaktadır. Sadece insan sağlığı alanında değil, biyoteknoloji, tarım, hayvancılık ve biyomedikal alanında da uygulanabilir bir tekniktir (Samsunlu, 2021).

Sporcuların spor müsabakalarındaki performanslarını arttırmak için sporcuların kullanımı yasak olan kimyasal madde kullanıp kullanmadığını belirlemek için genetik analizlere gerek vardır. Laboratuvarda yapılan incelemelerde doping yaptığı saptanan sporcu müsabakalardan men edilir (Weir, 1996). Böylece sporcuların hak etmediği bir başarıyı elde etmesi denetlenebilir.

DNA parmak izi yöntemi, adli suçlarda suçlunun kimliğini tespit etmede, babalık tayini testlerinde, kalıtsal hastalıkların tanı ve tedavisinde kullanılır. DNA parmak izi yönteminde kişiye ait DNA dizilimi ile vücut parçasının (saç, deri, kan, ter gibi) DNA dizilimi karşılaştırılarak kişinin kimliği tespit edilebilir. Bir kişinin gerçek babasının tespit edilebilmesi için çocuk babanın alleleline sahip olmalıdır. Eğer çocukta babaya ait allel tespit edilemezse ya da yoksa babasının bu kişi olmadığı anlaşılır. Weir (1996) yaptığı çalışmada babalık tayini testinin genetik alanında önemli bir bilgi sağladığını aktarmıştır. Burada dışlama olasılığı belirlenir. Bir de babalık indeksi vardır. Anne, sözde baba ve çocuğun gerçek genotiplerine bakılır. Bu çalışmalar için ek bilgiye Weir (1996)'den ulaşılabilir. DNA dizilerini karşılaştırma metodu Florida'da bir dış hekiminin hastalarında AIDS (Acquired Immune Deficiency Syndrome) virüsü olup olmadığını anlamak için HIV (Human Immunodeficiency Virus) dizilerini karşılaştırmasıyla başlamıştır (Weir, 1996). Bu genetik analizler sonucu verilere ait olasılık hesaplamaları genetik alanındaki istatistik yöntemlere örnektir.

Adli testlerde ise iki olasılık hesaplanır. Bunlar; şüpheli kişinin suçlu ya da suçsuz olduğunu gösteren hesaplamalardır. Olay yerinden temin edilen materyalden yararlanılır. Bazı suçlarda coğrafi bölge şüpheli kişi sayısını daraltırken bazen de şüphelinin herhangi bir etnik gruptan olması o suçu işleme ihtimaliyle ilişkili olmayabilir. Delil olarak olay yerinden alınan numune ile şüphelinin aynı allellere sahip olduğunu tayin edebilmek için laboratuvarında klinik testler yapılarak DNA analizi yapılmasına gerek vardır. Bu analizlerde de hatalar oluşabilir. Bunun için gerekli prosedürlere uyulmalıdır (Weir, 1996). Hatanın nereden kaynaklandığının tespiti için de istatistik testlere ihtiyaç vardır.

Bütün bu genetik analiz çalışmalarında kullanılan istatistik yöntemler araştırıldığında birçok kaynağa ulaşılmaktadır. Genetik alanında elde edilen verilerin işlenmesi, kaydedilmesi, yorumlanması, benzerlik ve farklılıklarının tespit edilmesi için istatistik yöntemlere ihtiyaç vardır. İstatistik yöntemlerde dört aşama vardır;

- Araştırmanın konu ve kapsamı belirlenir. Uygun sayıda ünitelerden bilgiler (veriler) toplanır.
- Toplanan bilgiler bu aşamada derlenir, sınıflandırılır. Grafik ve tablo olarak derlenen veriler sunulur.
- Verilerden gerekli olan çeşitli istatistikler hesaplanır ve sunulacak hale getirilir.
- Bu verilerden sonuç çıkarılarak genelleme-değerlendirme yapılır (Terzi, 2018).

Elde edilen verilerin sunum aşamasında grafik ve tablo kullanımı sıklıkla başvurulmuş bir yöntemdir. Verilerin gruplanmasını ve birlikte görünümünü sağlar. İstatistik çalışmalarda görsel olarak tercih edilebilecek histogram, çubuklu diyagram gibi birçok grafik türü vardır. Bu grafikler hakkında detaylı bilgi için Terzi (2018) kaynağı incelenebilir.

Araştırmalarda, araştırmacının çalışmanın sonucu ile ilgili bir beklentisi olur. İstatistik yöntemlerin uygulanabilmesi için, araştırmalardan elde edilecek sonucun bu beklentiyi destekleyip desteklemeyeceğinin, araştırma yapmadan önce belirlenemeyecek olması

gerekir. İstatistik yöntemler tercih edilip uygulanırken arařtırmanın bazı konulara dikkat etmesi gerekir. Öncelikle arařtırmanın sadece sonuç kısmında deęil yapılan çalıřmanın bařından sonuna kadar bir istatistikçiye ihtiya vardır. Arařtırmadaki örnek geniřlięi çoęaltılarak istatistik gü oranı yükseltilmiř olur. Arařtırmalarda kullanılacak tüm uygulamalarda oluřabilecek hatanın mümkün olduęu kadar küçük olması gerekir. Arařtırmanın yapıldıęı örnek geniřlięinin en uygun sayıda olması gerekir. Çünkü örnek geniřlięinin az olması, yapılacak çıkarımların hatalı olma ihtimalini artırır. Gereęinden fazla olduęunda da hata kaynakları çoęalır ve daha fazla maliyet ve zaman gerekir. İstatistik yöntemler uygulanırken arařtırmacı bu konulara hassasiyetle dikkat etmelidir (Kavuncu, 2018).

İstatistik arařtırmalarda etik olmak da bilimsel olarak çok önemlidir. Veriler oluřturulurken ve kaydedilirken objektif olunmalıdır. İstatistiksel arařtırmalarda arařtırmanın bařından sonuna kadar her ařamada belirli hatalar oluřabilir. Arařtırmanın düzensiz ilerlemesi, yapısının karmařık oluřu, istatistięin sadece deęerlendirme kısmında dâhil edilmesi gibi faktörler hatalara sebep olur. Önemli olan bu hataları bařtan fark edebilmek ve arařtırmayı bu hataları en aza indirerek sürdürüp tamamlamaktır (Bayram, 2015).

Genetik analizlerdeki istatistik yöntemler hakkında yapılan çalıřmalar incelendięinde Masatoshi Nei, Jukes ve Cantor, Sudhir Kumar, Anthony J.F. Griffiths ve Tajima'nın çalıřmaları olduęu görölmektedir. Hudson – Kreitman – Aguade (HKA) Testi, Mc Donald -Kreitman (MK) Testi ve Tajima'nın D Testi bu alanda sıklıkla bařvurulan testlerdir. Esasen tezin amacı bu testleri tanıtmak olduęundan bu testler yöntem bölümünde geniř olarak ele alınmıřtır.

3. YÖNTEM

Giriş bölümünde de belirtildiği gibi, bu tez bir derleme çalışmasıdır. Bu çalışmada genetik analizlerde kullanılan istatistik yöntemlerle ilgili kaynaklar incelenmiştir. Tezin Kaynakça kısmında belirtilen bu kaynaklar sistemli bir şekilde taranarak bu tez çalışması oluşturulmuştur.

Genetik çalışmalardan elde edilen verileri gruplandırmak, karşılaştırmak, değerlendirmek ve analiz edebilmek için istatistik yöntemler oldukça çeşitli olup çoğu farklı kaynaklarda yer almaktadır. Bu tez çalışması ile bu yöntemler araştırılıp derlenerek sistemli bir şekilde bu tezin içerisinde bir araya getirildi.

İki tür arasında başlangıçta aynı olan nükleotid dizisinde T zaman sonra d adet nükleotid farklılaşmışsa, birim zamanda nükleotid sitesi başına ortalama ikame oranı r 'yi, (3.1) denklemi ile bulabiliriz;

$$r = \frac{d}{2LT} \quad (3.1)$$

(3.1) denklemindeki d ; L uzunluğunda bir DNA dizisinde oluşan farklılıkların sayısını, T ise 2 türün ayrılması için geçen süreyi ifade eder. (3.1) denklemi, ikame oranı olan r ile DNA dizi karşılaştırmalarında tespit edilen farklılıkların sayısı d arasındaki ilişkiyi gösterir. Eğer birden fazla değişiklik gerçekleşirse, d tam tespit edilemeyebilir. Oluşan ikinci değişiklik birinci değişikliğin oluşturduğu etkiyi yok edebilir. Örneğin; ilk ikamede bir Adenin nükleotidi yerine bir Timin nükleotidi geçebilir. İkinci ikamede ise bu Timin nükleotidi yerine tekrar bir Adenin nükleotidi gelirse aslında 2 değişim olmasına rağmen (3.1) numaralı eşitlikteki d sayısı bunları göstermez. Gerçekte kaç tane ikame olduğunu belirleyebilmek için birçok model geliştirilmiştir. Bu sorunun çözümünde genel olarak Jukes – Cantor Modeli kullanılır. Bu model bir sitedeki nükleotidin diğer 3 nükleotid ile eşit ihtimalle değişebileceği varsayımına dayanır. Oluşan ikinci değişikliğin birinci değişikliği eski haline getirme olasılığı $1/3$, eski haline getirmeme olasılığı ise $2/3$ 'tür (Nielsen ve Slatkin, 2013).

α ; bir nükleotidden başka bir nükleotide mutasyonun gerçekleşme hızı olarak kabul edildiğinde Adenin (A) nükleotidinin; α olasılığı ile Sitozin (C) nükleotidine, α olasılığıyla Timin (T) nükleotidine, α olasılığıyla Guanin (G) nükleotidine mutasyona uğraması beklenir. A nükleotidinin bu süre zarfında değişmeme ihtimali ise $1 - 3\alpha$ 'dır. Buna göre DNA üzerindeki herhangi bir sitede $t + 1$ anında A olma ihtimali (3.2) denklemi ile bulunur (Hartl ve Clark, 2007).

$$P_{A(t+1)} = (1 - 3\alpha)P_{A(t)} + \alpha (1 - P_{A(t)}) \quad (3.2)$$

(3.2) denkleminin birinci kısmı t anında $P_{A(t)}$ oranındaki Adeninlerin değişmeme, yani Adenin olarak kalma ihtimalini, ikinci kısmı ise $1 - P_{A(t)}$ kadar Adenin dışı nükleotidlerin Adenine değişme ihtimalini verir. (3.2) denkleminde (3.3) denklemi elde edilir (Hartl ve Clark, 2007):

$$P_{A(t+1)} - P_{A(t)} = dP_{A(t)} / dt = - 4\alpha P_{A(t)} + \alpha \quad (3.3)$$

Bu diferansiyel denklemin çözümü

$$P_{A(t)} = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t} \quad (3.4)$$

(3.4) denklemindeki $P_{A(t)}$, ilk durumda Adenin nükleotidi olduğu kabul edilen bir sitede t anında yine Adenin olma ihtimalini gösterir. t süresi boyunca iki dizideki herhangi bir sitede aynı nükleotidin (N) olma ihtimali ise

$$P_{NN} = \frac{1}{4} + \frac{3}{4} e^{-8\alpha t} \quad (3.5)$$

(3.5) denkleminde iki dizi boyunca olacak değişikliklerin toplam oranı, bir dizidekinin 2 katıdır. İki dizi arasında aynı olmayan nükleotid sitelerinin oranı aşağıdaki (3.6) denkleminde görüldüğü gibi bunun 1'den farkına eşittir (Hartl ve Clark, 2007).

$$d = 1 - P_{NN} \quad (3.6)$$

Başlangıç durumunda akrabalıktan dolayı aynı olan 2 DNA dizisinin t sürede farklılaşma ihtimalinin nükleotid ikame hızıyla olan bağlantısı, denklem 3.6'dan denklem (3.7)' deki gibi bulunur (Kavuncu, 2022):

$$\hat{d} = \frac{3}{4}(1 - e^{-8\alpha t}) \quad (3.7)$$

λ' y1; var olan bir nükleotidden başka bir nükleotide mutasyon oranı kabul edersek $\lambda=3\alpha$ olur. Bu durumda t zamanında iki DNA dizisi arasındaki benzer olmayan nükleotid sitelerinin beklenen oranı k ise $k = 2\lambda t = 2(3\alpha)t = 6\alpha t$ olur. (3.7) denkleminde (3.8) denklemi elde edilir (Hartl ve Clark, 2007).

$$8\alpha t = -\ln\left(1 - \frac{4}{3}\hat{d}\right) \quad (3.8)$$

$k = \frac{3}{4}(8\alpha t)$ olduğundan k'nın ve d'nin beklenen değerlerinden (3.9) denklemi bulunur (Kavuncu, 2022).

$$\hat{k} = -\frac{3}{4}\ln\left(1 - \frac{4}{3}\hat{d}\right) \quad (3.9)$$

Tahminin varyansı da aşağıdaki (3.10) denklemi ile verilmektedir (Kavuncu, 2022):

$$\text{var}(\hat{k}) = \hat{d} (1 - \hat{d}) / [L(1 - \frac{4}{3}\hat{d})^2] \quad (3.10)$$

Denklem (3.10)'da yer alan L; karşılaştırması yapılan DNA dizilerindeki nükleotid sitelerinin sayısıdır. Jukes – Cantor Modeli, mutasyonun gerçekleşme oranının sabit ve tüm nükleotidler için eşit miktarda olduğunu varsayar. Bu alanda Kimura veya Hasegawa-Kishino-Yano (HKY) Modeli gibi, eşitliğin olmayabildiğini varsayan başka modeller olsa da sıklıkla Jukes – Cantor Modeli tercih edilir (Kavuncu, 2022).

Örnek:3.1- (Can, 2023'ten değiştirilerek alınmıştır). 303 aminoasit dizisine karşılık gelen, stop kodonuyla birlikte 912 nükleotid vardır. Bunlardan 20 nükleotidin değişmesi durumunda;

$d = 20/912 = 0,0219$ olur. Çoklu deęişmeleri de önemseyen bir düzeltme ile gerçek deęişme oranı;

$\hat{k} = -\frac{3}{4} \ln(1 - \frac{4}{3} 0,0219) \cong 0,0222$ olur. 100 birim zaman önce iki dizinin müşterek cedden dolayı özdeş olduğunu varsayalım. Buradan birim zamanda nükleotid ortalama ikame hızı α ;

$\hat{k} = 6\alpha t$ denkleminde $\alpha = \frac{\hat{k}}{6t} = 0,0222/600 = 0,000037$ olarak bulunur.

Populasyonların çoęu allel frekansları farklı alt populasyonlara ayrılmaktadır. Mutasyon, göç, genetik sürüklenme etkisi ile alt populasyonlardaki bu allel frekanslarındaki farklılıklar seçilimle daha da artabilir. Bir populasyonda allel frekanslarındaki yüksek miktardaki farklılıklar bu populasyonun çevreye uyumunu gösteren önemli bir kriterdir (Nielsen ve Slatkin, 2013).

İnsanlarda laktaz (LCT) lokusu laktaz enzimini kodlar. Bu enzim sütün laktozunu, glikoz ve galaktoza ayırır. Bebekler anne sütünü sindirebilmek için bu enzimi salgılar. Bazı yetişkinlerde ise bu genin ifadesi azalmakta ve bu kişilerin laktoza olan toleransları zayıflamaktadır. Araştırmacılar laktoz intoleransının genellikle laktaz genindeki bir mutasyondan kaynaklandığını belirtmektedir. Bu genin aktivitesi devam eden yetişkin bireylerin laktaz kalıcılığı vardır. Laktoz intoleransı oranı Danimarka ve İsveç'teki yetişkin bireylerde %1 civarında iken Tayland'daki yetişkin bireylerde bu oran %97 civarındadır. Laktoz kalıcılığına neden olan allelin tahmini on bin yıl önce doğal seçilim etkisiyle Kuzey Avrupa'daki bireylerde laktoz intoleransı bakımından olumlu etki gösterdiği tahmin edilmektedir (Nielsen ve Slatkin, 2013).

Populasyonlar arası genetik farklılaşmaya dayalı testlerin kullanıldığı çalışmalara bir diğer örnek EPAS1 lokusuyla ilgili yapılan araştırmadır. Han Çin toplumundaki bireyler ile Tibet toplumundaki bireylerin allel frekansları benzerliğinden dolayı bu populasyonlar birbiriyle bağlantılıdır. Han Çin toplumundaki bireylerde düşük frekans, Tibet toplumundaki bireylerde yüksek frekans gösteren EPAS1 geninde iki SNP bulunur. EPAS1 geni hipoksi yani oksijen oranının az olduğu durumlara organizmanın verdiği yanıtı organize eder. Tibet toplumunun büyük bir kısmı rakımı

4.000 metreden yüksek olan platoda yaşamaktadır. Bu platoda oksijen basıncı deniz seviyesindeki oksijen basıncından %40 daha azdır. Bu iki populasyon arasındaki allel frekansı farklılığı Tibetli bireylerin rakımı yüksek yerlerde yaşamaya uyum sağlamalarını kolaylaştıracak bir seçim geçirdiğini düşündürür. Tibetli bireylerdeki EPAS1 lokusu ile hemoglobin seviyeleri arasındaki ilişki de EPAS1 lokusu ile rakımı yüksek yerlerdeki yaşama uyum gösterme arasında bağlantı olabileceğini gösterir. Canlılar üzerinde etkili olan doğal seçilimin tanımlanması populasyon genetiğinin uygulama alanlarından biridir. Tibetli bireylerdeki EPAS1 lokusu üzerinde etkili olan seçilimin araştırılması canlıların geçirdiği evrimi açıklamayanın yanında canlılar arasındaki işlevsel bağlantıları da açıklamaya yarar (Nielsen ve Slatkin, 2013).

Tibetli bireylerin rakımı yüksek, oksijeni düşük seviyelerde olan yerlerde yaşayabilmelerine benzer örnek bazı kazlarda da vardır. *Anser indicus* kazı rakımı 4.000 - 6.000 metre olan Tibet göllerinde yaşar ve burada ürerler. Everest Dağı'nın zirvesinden geçerek kış aylarını ovalarda geçirir. Bir diğer örnek ise Güney Amerika'da deniz seviyesinden 5.000 - 6.000 metre yükseklerde yaşayabilen *Chloephaga melanoptera* kazıdır (Hartl ve Clark, 2007).

Genetik analizlerde kullanılan istatistik yöntemler; özelliklerine, içeriklerine ve sağladığı avantajlara göre farklı gruplara ayrılabilir. Tezin bu bölümünde genetik araştırmalarda sıklıkla kullanılan istatistik yöntemlere ait bilgiler belli bir düzende verildi. Bu yöntemlerden başka farklı yöntemler elbette olabilir ancak bu yöntemlerin bir arada yer aldığı Türkçe bir kaynak oluşturabilmek daha önemlidir. Bu da tezin önemli bir amacıdır. Tezin bu bölümünde bu testler gerekli açıklamalarıyla listelenmiştir.

İstatistik alanındaki araştırmalarda sıklıkla karşılaşılan kesikli dağılım türleri; Binomiyal dağılım ve Poisson dağılımıdır. Her ikisinin de kendine ait özellikleri vardır. Bu özelliklere bakacak olursak;

- Binomiyal Dağılım: Sayma yoluyla ulaşılan verilerden yararlanılır. Bir olayın gerçekleşme olasılığıyla gerçekleşmeme olasılığı sabit olan olayları konu alan dağılım türüdür. Tüm verilere ait olasılıkların toplamı bire eşit olur. Yeni doğan

bebeğin cinsiyetinin kız veya erkek olması, madeni paranın tuğra ya da yazı gelmesi gibi özellikler, Bernoulli Dağılımı gösterir. N denemede istenenin 0, 1, 2, ..., N kere gelmesi belirli ihtimallerle mümkündür; parayı beş kere atsak 0, 1, 2, 3, 4 veya 5 kere tuğra gelebilir, 4 çocuklu ailelerde kız çocuk sayısı 0, 1, 2, 3 veya 4 tane olabilir. İşte istenen sayısına ait bu ihtimal dağılımına Binomiyal Dağılım denir. Deneme sayısı (N) 1 olduğunda dağılım, Bernoulli dağılımıdır.

- Poisson Dağılımı: Belirli bir zamanda rastgele oluşan ender gerçekleşen olayların dağılımıdır. Bu dağılım da sayma yoluyla ulaşılan verilere ait dağılımdır. Bir bölgede bir ay içinde olan deprem sayısı, bir basketbol maçında bir oyuncunun atamadığı serbest atış sayısı, bir futbol maçında kırmızı kart gören futbolcu sayısı gibi özelliklerin dağılımı, bu dağılıma örnektir (Ercan, 1997).

Poisson Dağılımı genetikte J.B.S. Haldane tarafından geliştirilen yaklaşımda kullanılmıştır (Griffiths vd. 2012). Genetik analizlerde istatistik yöntemlerin kullanımı Galton ile başlamakla birlikte, daha sonraki yıllarda Haldane, Wright ve Fisher tarafından yapılan çalışmalarda, populasyonların genetik yapılarını incelemek üzere istatistik yöntemler çokça kullanıldı. Haldane iki lokus arasındaki yeni kombinasyonların oranını hesaplarken, iki lokus arasındaki parça değiş tokuşlarının sayısı ile yeni kombinasyonlar arasında bir ilişki kurdu. Parça değiş tokuşlarının sayısı çift ise yeni kombinasyon ortaya çıkmıyor, tek ise yeni kombinasyon ortaya çıkıyordu. İki lokus arasındaki parça değiş tokuşlarının sayısı, eğer parça değiş tokuşu ihtimali çok küçük değilse Binomiyal bir dağılım gösterir. Ancak iki kromatid arasında herhangi bir yerde parça değiş tokuşu olma ihtimali oldukça küçük bir ihtimal olduğu için parça değiş tokuşu sayısı Poisson dağılımına daha uygundur.

Parça değiş tokuşu sayısı, daha küçük segment içinde düşünülünce, Poisson dağılımı söz konusu olur. Her bir nükleotid yerine başka bir nükleotid ikamesi de çok küçük ihtimallerle söz konusu olduğu için, bir DNA dizisi boyunca nükleotid ikamelerinin sayısı da Poisson dağılımı gösterir.

LD (Linkage Disequilibrium) ve Haplotip Yapısı: Genlerin olası allelleri, gametlere rastgele dağılıyorsa herhangi bir gamet frekansı allellerin frekanslarının çarpımına eşittir. Bu denge halindeki gametik frekansları;

$$P(AB) = p_A \times p_B$$

$$P(Ab) = p_A \times q_b$$

$$P(aB) = q_a \times p_B$$

$$P(ab) = q_a \times q_b$$

şeklinde. Bir gamette bir araya gelme ihtimali, ihtimallerinin çarpımına eşit (rastgele) olan genler bağlantı dengesinde (Linkage equilibrium) iken bir gamette bir araya gelme ihtimali, ihtimallerinin çarpımına eşit (rastgele) olmayan genler ise bağlantı dengesizliği (Linkage Disequilibrium: LD) durumundadır. Birey sayısının yeterince çok olduğu rastgele çiftleşen bir popülasyonda, mutasyon, göç ve doğal seçim olmaması halinde lokusların çoğu, bağlantı dengesindedir (Hartl ve Clark, 2007).

İkişer alleli olan birbirine bağlı iki lokusun bir lokusunda A ve a allelleri, diğerinde B ve b allelleri olduğunu varsayalım. Gametler, bu allellerin 4 farklı birleşme durumunda oluşur (AB, aB, Ab, ab) ki bunlara haplotip denir (Nielsen ve Slatkin, 2013).

Herhangi bir popülasyonda ayrılan mutasyon üzerinde etkisi devam eden seçilimi saptamaya yönelik Tajima'nın D Testi gibi testler de vardır. Avantajlı allele sahip kromozomlar, bu alleller ortaya çıkmadan önceki süreçte benzer bir haplotip yapısı gösterirler. Bunun tespitinde kullanılan birkaç farklı istatistik test mevcuttur. Bu testlerin temel prensibi devam etmekte olan seçici taramadan etkilenen haplotiplerin hemen hemen aynı olması gerektiğini esas almalarıdır (Nielsen ve Slatkin, 2013). Bağlantı dengesizliğini belirlemek için Khi-Kare Bağımsızlık Testi kullanılmaktadır (Hartl ve Clark, 2007; Nielsen ve Slatkin, 2013). Bu bakımdan Khi – Kare Bağımsızlık Testine kısaca göz atmakta fayda görülmüştür.

Khi – Kare Bağımsızlık Testi: Birbirinden farklı gözlem türlerine ait kategorik verilere göre uyum iyiliğini test etmede kullanılır. Her analiz için beklenen frekansları tahmin eden sıfır hipotezini kullanır. Gözlemlenen değer, beklenen değerden şansa bağlı olmayacak kadar farklıysa sıfır hipotezi reddedilebilir. Bu durumda modelin doğruluğu kabul edilmez (Nielsen ve Slatkin, 2013).

k gözlem kategorisinde, gözlemlenen değerler; $O_1, O_2, \dots, O_k$, beklenen değerler ise $E_1, E_2, \dots, E_k$ olsun. Bu durumda Khi - Kare Testi (3.11) denklemi ile hesaplanır.

$$\chi^2 = \sum_{i=1}^k \frac{(E_i - O_i)^2}{E_i} \quad (3.11)$$

(3.11) denklemi ile hesaplanan χ^2 değerinin çok büyük çıkması gözlenen ve beklenen değerlerin birbirinden çok farklı olduğu anlamına geleceği için bu modeli reddedebiliriz.

Khi- Kare Testi uygulama basamakları;

- χ^2 ' nin hesaplanması,
- p (serbestlik derecesindeki azalma)' nin hesaplanması,
- χ^2 değerinin k – p serbestlik derecesine sahip Khi – Kare dağılımıyla karşılaştırılmasıdır (Nielsen ve Slatkin, 2013).

Khi – Kare dağılımı özel bir Gama dağılımıdır. Gama dağılımı genellikle spesifik bir olay q defa olana kadar beklenen süreye göre hesaplanır. Gama dağılımı parametreleri q ve μ olan 2 parametrelili bir dağılımdır. Gama dağılımı $\mu = 1/2$ ve $q = n/2$ (n, pozitif bir tam sayı) parametre değerlerinde;

$$f(x) = \frac{\frac{1}{2} e^{-\frac{x}{2}} \left(\frac{x}{2}\right)^{\frac{n}{2}-1}}{\Gamma\left(\frac{n}{2}\right)} \quad (3.12)$$

(3.12) denklemiyle verilen dağılım bir gösterir. Bu dağılıma Khi - Kare dağılımı da denir. Gama dağılımı hakkında detaylı bilgi için (Kavuncu, 2021) kaynağını incelemekte yarar vardır.

Örnek:3.2- (Hamilton, 2009'dan alınmıştır). Doğu Sibirya'nın yerli halkı olan 1066 Chukchi bireyinden alınan bir örnekleme ait MN kan grubu genotiplerinin beklenen ve gözlemlenen değerleri Tablo 3.1'de verilmiştir. Bu örneğin Hardy – Weinberg dengesine uygunluğunu Khi – Kare Testi'ni kullanarak açıklayalım.

Tablo 3.1 MN kan gruplarının beklenen ve gözlemlenen değerleri (Hamilton, 2009)

Genotip	Gözlemlenen	Beklenen Genotip Sayısı	Gözlemlenen - Beklenen
MM	165	$N \times \hat{p}^2 = 1.066 \times (0,4184)^2 = 186,61$	-21,6
MN	562	$N \times 2\hat{p}\hat{q} = 1.066 \times 2(0,4184)(0,5816) = 518,80$	43,2
NN	339	$N \times \hat{q}^2 = 1.066 \times (0,5816)^2 = 360,58$	-21,6

$$M \text{ kan grubunun allel frekansı } \hat{p} = 2 \times F_{MM} + F_{MN} / 2N = \frac{2 \times 165 + 562}{2 \times 1.066} = \frac{892}{2.132} = 0,4184$$

$$N \text{ kan grubu allel frekansı } \hat{q} = 1 - 0,4184 = 0,5816 \text{ olur.}$$

$$\text{Buradan } \chi^2 = (-21,6)^2 / 186,61 + (43,2)^2 / 518,80 + (-21,6)^2 / 360,58 = 7,39 \text{ bulunur.}$$

Bu örnekte 3 genotip ve 1 tahmini allel frekansı için serbestlik derecesi $3 - 1 - 1 = 1$ 'dir. Ek A'da verilen tablodan 1 serbestlik derecesi ile 7,39 χ^2 değerinin 0,01 ile 0,001 arasında bir olasılığa ait olduğu görülmektedir. Bu da bize HW dengesine göre beklenen genotip frekanslarının bu örnekteki popülasyonda olmadığını gösterir.

Örnek:3.3- A ve B allelinden oluşan bir örnekte AB gametinin gözlemlenen değeri 150, Ab gametinin gözlemlenen değeri 300, aB gametinin gözlemlenen değeri 200, ab gametinin gözlemlenen değeri ise 350'dir. Bu 4 gametin beklenen değerleri ise Tablo 3.2' de verilen bilgilere göre;

$$P(AB) = P(A) \times P(B) = 0,45 \times 0,35 = 0,1575,$$

$$P(Ab)= P(A) \times P(b) = 0,45 \times 0,65 = 0,2925,$$

$$P(aB)= P(a) \times P(B) = 0,55 \times 0,35 = 0,1925,$$

$P(ab)= P(a) \times P(b) = 0,55 \times 0,65 = 0,3575$ olduğuna göre bu gametlerin rastgele farklılaşma ihtimalini Khi - Kare testini kullanarak açıklayalım?

Tablo 3.2 A ve B gametlerine ait olasılıklar

	A	a	Toplam
B	0,15	0,20	0,35
b	0,30	0,35	0,65
Toplam	0,45	0,55	1,00

$\chi^2 = (-8)^2 / 158 + 8^2 / 292 + 8^2 / 192 + (-8)^2 / 358 = 0,405 + 0,219 + 0,333 + 0,178 = 1,135$ bulunur. Serbestlik derecesi, Tablo 3.2'ye göre satır sayısının 1 eksiği ile sütun sayısının 1 eksiğinin çarpımına eşittir. Serbestlik derecesi, $1 \times 1 = 1$ 'dir. Ek A'da verilen tablodan 1 serbestlik derecesi ile 1,135 χ^2 değerinin 0,50 ile 0,25 arasında bir olasılığa ait olduğu görülmektedir. Böylelikle iki lokustaki allellerin gametlere dağılımının birbirinden bağımsız olduğu, yani bu örneğin temsil ettiği populasyonun bağlantı dengesinde olduğu hipotezini kabul edebiliriz.

3.1 Student'in t Testi

t istatistiğinin genel özelliği, ortalamasının sıfır olmasıdır. t dağılımı serbestlik derecesine bağlı bir normal dağılım gösterir. Populasyonun varyansını bilmediğimiz örneklerde bu örneğe ait farklı ihtimalleri hesaplayabilmek için standart normal dağılım yerine t dağılımı olarak bilinen ve (3.13) denklemiyle verilen test istatistiği tercih edilir.

$$t = \frac{\bar{x} - \mu}{S_{\bar{x}}} \quad (3.13)$$

t – dağılımının Varyansı 1'den büyüktür. Dağılımın serbestlik derecesi büyürse Varyans 1'e doğru azalan bir değer alır (Kavuncu, 2021).

t – dağılımının 3 farklı kullanımı olup bunlar aşağıda kısaca özetlenmiştir.

- Tek Örneklem t – Testi (One Sample t Test): Bir araştırma hakkında çalışma yapmadan önce yaptığımız tahminlerin doğruluğunu tespit etmemizde işe yarayan bir testtir. Örneğin bir A hastalığı için ilaç geliştirilecek olsun. Ancak tedavi edilecek kişilerin kilolarıyla bu araştırmanın ilişkili olduğunu bilgisini bu test ile analiz edebiliriz. Karadeniz Bölgesi’ndeki bu A hastalarının kilo ortalamalarını 55 olarak tahmin ettiniz diyelim. Araştırma için seçilen 50 kişinin kilo ortalaması 57 çıkarken başka bir 50 kişinin ise kilo ortalaması 52 çıkabilir. Bu bizim çalışmamızda hata yaptığımızı göstermez. Bu hastalığa sahip bu bölgedeki tüm kişilere ulaşmak hem zor hem de fazla maliyet ve zaman gerektirdiği için bu istatistiksel teste ihtiyaç vardır (Eymen, 2007).
- Bağımsız Örneklem t –Testi (Independent Sample t Test): Araştırma yapılacak 2 farklı grup için ortalamaları karşılaştıran bir testtir. Örneğin: 30 yaşındaki kadın ve erkeklerden oluşan bir çalışmada bu kişilerin düzenli spor yapmalarını karşılaştırmak için, ya da bu kişilerin günlük yürüme mesafelerini karşılaştırmak için bu test kullanılır (Eymen, 2007).
- Eşleştirilmiş Örneklem t – Testi (Paired Sample t Test): Deneysel araştırmalarda deney öncesi ve sonrası değerlerin analizi için kullanılır. Ayrıca bir çalışmaya katılan kişilerin 2 ayrı konu hakkında düşüncelerini karşılaştırmak için de kullanılan bir testtir. Örneğin: bir hastanede tedavi gören kanser hastalarının kanlarında belirli bir maddenin tedaviden önce ve sonraki konsantrasyonlarını bu yöntemle karşılaştırabiliriz.

3.2 Hudson – Kreitman – Aguade (HKA) Testi

İki popülasyonun birbirinden farklı ile bu popülasyonların kendi içindeki değişiklikleri karşılaştıran testtir. Bu test ilerleyen zamanda MacDonald ve Kreitman tarafından yeniden düzenlenmiş, genişletilmiştir (Kavuncu, 2022). HKA testi; DNA dizi analizlerinden yola çıkarak doğal seçilimi belirleyebilmek için kullanılır. Gerek tür içi gerekse de türler arasındaki farklılıkları karşılaştırmaya yarayan bir testtir (Nielsen ve Slatkin, 2013).

Tür içinde ve türler arasında tahmin edilen mutasyon sayısı; popülasyondaki birey sayısına, örnek büyüklüğüne ve ayrışma zamanına göre değişir. Aynı türden aynı örneklem büyüklüğündeki çok sayıda lokus verisinde, türler arasında oluşan sabit farklılığın miktarının tür içinde var olan ayrışma bölgeleri miktarına oranı, eğer doğal seçim yoksa eşit olması gerektiği var sayılır (Nielsen ve Slatkin, 2013).

Tablo 3.3 HKA testi veri tablosu (Nielsen ve Slatkin, 2013)

	Lokus 1	Lokus 2
Türler İçi Açılan Siteler	S_1	S_2
Türler Arası Sabit Farklılıklar	F_1	F_2

İki lokus için Tablo 3.3’deki verilere göre nötr durumda;

$$E[S_1] / E[F_1] = E[S_2] / E[F_2] \quad (3.14)$$

(3.14) eşitliği olması gerekir. Bu eşitliğin sağlanamaması halinde seçilimin iki lokustan biri üzerinde etkisini gösterdiği düşünülmektedir. S_2/F_2 oranının S_1/F_1 oranından küçük çıkmasının nedeni seçici taramanın Lokus 2’yi etkilemesinden dolayı bu lokusta ayrılan bölge sayısında azalma olması olabilir. Tabii bu çıkarımdan farklı çıkarımlar da olabilir (Nielsen ve Slatkin, 2013).

Örnek:3.4-

Tablo 3.4 Dmd lokusundaki 2 introna ait HKA testi verileri (Nielsen ve Slatkin, 2013)

Coğrafi Bölge	Lokus	S	F	p değeri
Afrika	İntron 7	6	39	NS
	İntron 44	15	27	
Avrupa	İntron 7	1	39	< 0,05
	İntron 44	10	27	
Asya	İntron 7	0	39	< 0,05
	İntron 44	10	27	
Amerika	İntron 7	3	39	NS
	İntron 44	9	27	

Tablo 3.4'deki S; ayrılan bölge sayısını, F; insan ve şempanzeler arasındaki sabit farkların sayısını ifade ederken NS; önemli değil anlamındadır. Nachman ve Crowell HKA testini kullanarak insanlardaki Duchenne muscular dystrophy (Dmd) geninde yer alan intron 7 ve intron 44'teki genetik varyasyonun doğal seçilimden etkilenme durumunu araştırmışlardır. Bunun için sabit farklılığı belirleyebilmek için insan DNA dizisi ve şempanze DNA dizisi analiz edilmiş ve bu iki canlı türünün DNA dizileri karşılaştırılmıştır. Yapılan çalışmada araştırma yapılan popülasyonlarda intron 44' e göre intron 7' de değişkenlikte yüksek miktarda azalma olduğu anlaşılmıştır. Çalışmanın bu aşamasından sonra HKA testi kullanılarak genetik veriler istatistik olarak yorumlanmıştır. Araştırmacılar Asya ve Avrupa' daki popülasyonlarda intron 7 bölgesindeki farklılık miktarının doğal seçilimden dolayı azalmış olabileceğini aktarmışlardır (Nielsen ve Slatkin, 2013).

3.3 Mac Donald – Kreitman (MK) Testi

K ve L diye adlandırdığımız iki farklı popülasyon düşünelim. K popülasyonunun protein kodlayan bölgesinin nükleotid dizisi d_K , L popülasyonunda bu nükleotid dizisi d_L olsun. Bu koşullarda d_K ve d_L dizilerinde aynı olmayan nükleotidler önem arz eder. K ve L popülasyonlarında farklı ancak K ve L popülasyonlarının kendi içlerinde aynı olan siteye sabit (fixed) denir. Sabit site hariç diğer siteler polimorfik diye isimlendirilir. Daha sonra bu siteler kendi içinde sinonim(S) ve sinonim olmayan(N) şeklinde sınıflandırılır. MK testi, sinonim olmayan değişikliklerin sinonim değişikliklere oranını kullanan bir test türüdür (Kavuncu, 2022).

Tablo 3.5 MK testi veri tablosu 1 (Kavuncu, 2022)

Sinonim ve Sinonim olmayan Siteler	Sabit ve Polimorfik Siteler		
	Sabit (F)	Polimorfik (P)	Toplam (F+P)
Sinonim (S)	s_f	s_p	$S = s_f + s_p$
Sinonim Olmayan (N)	n_f	n_p	$N = n_f + n_p$
Toplam (S+ N)	$F = s_f + n_f$	$P = s_p + n_p$	$T = S+N = F +P$

Tablo 3.5'deki n_f/s_f ve n_p/s_p oranlarının mutant alleller açısından bir seçilim yok ise aynı olması beklenir. K ve L popülasyonlarında sinonim olmayanların oranı sinonimlerin oranına göre çok azdır. Çünkü sinonim mutasyonlar seçilime ve

aminoasit farklılığına neden olmaz. Sinonim olmayanların sinonimlere oranının sabit ve polimorfik sitelerde farklı olmaması K ve L populasyonlarının seçim açısından farklı olmadığını gösterir. Bu da bize bu durumda mutasyon ve şansın çeşitliliğe etkisinin araştırılmasına ihtiyaç duyulmadığını gösterir. Bunu Pearson Khi - Kare veya Fisher'in exact testi ile kontrol edebiliriz (Kavuncu, 2022).

MK testi, tür içi polimorfizmi ve türler arası çeşitliliği karşılaştırma yapması yönünden HKA testine benzer bir testtir. Bu karşılaştırmayı yaparken kodlama bölgelerinde oluşan mutasyonları eş anlamlı ve eş anlamlı olmayan mutasyonlar şeklinde ayırır (Nielsen ve Slatkin, 2013).

Tablo 3.6 MK testi veri tablosu 2 (Nielsen ve Slatkin, 2013).

	Sinonim Olmayan	Sinonim
Türler İçi Açılan Siteler	S_N	S_S
Türler Arası Sabit Farklılıklar	F_N	F_S

Tablo 3.6'daki verilerden $E[S_N] / E[S_S] = E[F_N] / E[F_S]$ veya $E[S_N] / E[F_N] = E[S_S] / E[F_S]$ eşitliğinin olması beklenir. Eşitliklerin sağlanamaması durumunda ise tarafsızlık ve seleksiyonun etkilerinden bahsedilebilir. Analiz yapılan populasyonlar bu süreçte sabit büyüklükteyse;

- $F_N / F_S > S_N / S_S$ olması pozitif seleksiyona,
- $F_N / F_S < S_N / S_S$ olması ise negatif veya dengeleyici seleksiyona kanıt oluşturur (Nielsen ve Slatkin, 2013).

Örnek:3.5-

Tablo 3.7 MK test örneği - 1 (Kavuncu, 2022)

A-Çekimser Durum	Sabit	Polimorfik
Sinonim Değil	4	15
Sinonim	12	45
Oran	0,3333	0,3333

Örnek:3.6-

Tablo 3.8 MK test örneği - 2 (Kavuncu, 2022)

B- D melanogaster için Adh lokusundan ampirik veri	Sabit	Polimorfik
Sinonim değil	2	7
Sinonim	42	17
Oran	0,0476	0,4118

3.4 Tajima'nın D Testi

MK Testi sinonim olmayan faktörlerin sinonim olanlara oranını kullanırken Tajima'nın D istatistiği bunların arasındaki farka göre işlem yapan bir testtir. Herhangi bir DNA bölgesinin a tane nükleotid içerdiğini varsayalım. Bu popülasyondan herhangi bir b genişliğinde bir örnek alalım. Eğer bu nükleotid dizisi ile örnek arasında iki veya daha fazla farklılık varsa bu site açılma gösteren site (İng. segregating site) adını alır. Nükleotid dizisi ile popülasyondan alınan örneğin dizisi ikiye karşılaştırıldığında ortaya çıkan farklılık da uyumsuzluk gösteren sitelerde (İng. mismatch or divergent site) olur (Nielsen ve Slatkin, 2013).

Standart Nötr Model etkisindeki allel frekanslarının dağılımındaki sapmaları belirlemeye yarayan bir test türüdür. Tajima D istatistiği aşağıdaki (3.15) eşitliğiyle verilmektedir:

$$D = \frac{\hat{\theta}_T - \hat{\theta}_W}{\sqrt{\hat{V}(\hat{\theta}_T - \hat{\theta}_W)}} \quad (3.15)$$

$\hat{\theta}_T$ ve $\hat{\theta}_W$ değerleri ise (3.16) ve (3.17) numaralı eşitliklerden hesaplanır (Nielsen ve Slatkin, 2013):

$$\hat{\theta}_W = S / \sum_{i=1}^{n-1} \frac{1}{i} \quad (3.16)$$

$$\hat{\theta}_T = \frac{\sum_{(i,j): i < j} d_{ij}}{\frac{n(n-1)}{2}} \quad (3.17)$$

n; örneklenen kromozom sayısı,

S; açılma gösteren site sayısı,

d; i ve j dizileri arasındaki nükleotid farklılıklarının sayısıdır (Nielsen ve Slatkin, 2013).

μ ; tüm nükleotid dizisi boyunca mutasyon oranı,

N_e ; etkin popülasyon büyüklüğü,

N ; incelenen popülasyon ile homozigotluk olarak aynı artış oranı gösteren ideal popülasyon büyüklüğü olmak üzere (3.18) eşitliği genetikte sıklıkla kullanılan bir eşitliktir (Hartl ve Clark, 2007).

$$\theta = 4N_e\mu \quad (3.18)$$

Popülasyon genetiğinde $4N_e\mu$ yerine θ sembolü kullanılır. (3.15) numaralı denklemde görüldüğü gibi, Tajima D istatistiği, θ ' nın iki ayrı tahmini arasındaki farkın istatistik olarak önemli olup olmadığını belirlemeye yönelik bir testtir.

n büyüklüğündeki dizilerin incelendiği bir çalışmada oluşabilecek ayrılan bölge miktarı eşitlik (3.19)'de verilmiştir.

$$E(S) = \theta \sum_{i=1}^{n-1} \frac{1}{i} \quad (3.19)$$

(3.19) eşitliği $\theta = 4N_e\mu$ denkleğini tahmin etmemizi sağlar. Ayrıca;

$$\alpha = \sum_{i=1}^{n-1} \frac{1}{i} = 1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n-1} \quad (3.20)$$

olmak üzere,

$$\hat{\theta} = S / \alpha \quad (3.21)$$

denkliğini tahmin etmemizi sağlar (Hartl ve Clark, 2007).

Birçok dizi arasında oluşan ikili uyumsuzlukların ortalama miktarı Π ve kararlı yapıdaki n büyüklüğündeki bir örnekte;

$$E(I\bar{I}) = \theta \quad (3.22)$$

eşitliğinden yararlanılır. (3.22) eşitliği göz önüne alındığında oluşacak tahmin;

$$\hat{\theta} = I\bar{I} \quad (3.23)$$

şeklindedir (Hartl ve Clark, 2007).

Tajima Testi (3.21) ve (3.23) denklemlerindeki θ değerleri arasındaki farka dayanır. Yani $I\bar{I} - S / \alpha$ 'nın değerine odaklanır. Eğer herhangi bir faktörün sonsuz sayıda allel modeline bağlı olarak etkisi yoksa $I\bar{I} - S / \alpha$ 'nın değeri sıfırdır. Ayrıca $I\bar{I} - S / \alpha$ değeri pozitif ve negatif değerler alabilir (Hartl ve Clark, 2007).

Tajima' nın testindeki $I\bar{I} - S / \alpha$ farkının büyüklüğü farkın standart sapması şeklinde düzenlenirse eşitlik (3.24)'te verilen Tajima' nın D İstatistiği ortaya çıkar (Hartl ve Clark, 2007).

$$D = \frac{I\bar{I} - S / \alpha}{\sqrt{V(I\bar{I} - S / \alpha)}} \quad (3.24)$$

(3.24) eşitliğinin istatistik olarak kullanılma nedeni $I\bar{I} - S / \alpha$ değerinin ortalama ve varyans değerleri içermesidir (Hartl ve Clark, 2007).

Tablo 3.9 4 Allelli DNA dizisi örneği (Hartl ve Clark, 2007)

Alleller	DNA Dizisindeki Nükleotid Siteleri															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
a	A	A	A	A	T	T	T	T	G	G	G	G	C	C	C	C
b	A	A	A	A	T	T	T	T	G	G	G	G	C	C	C	C
c	<u>G</u>	A	A	A	<u>C</u>	T	T	T	<u>A</u>	G	G	G	<u>T</u>	C	C	C
d	A	<u>G</u>	A	A	T	<u>C</u>	T	T	G	<u>A</u>	G	G	C	<u>T</u>	C	C

Tablo 3.9’da 4 allelli bir DNA dizisi yer almaktadır. Bu dizi gerçekteki DNA dizilerine göre çok kısa olup gerçekteki DNA dizilerinden daha fazla çeşitlilik göstermektedir. Her sütunda çoğunluk kuralına aykırı nükleotidlerin altı çizilmiştir.

Örnek:3.7- (Hartl ve Clark, 2007). Tablo 3.9’daki dört tane DNA dizisi verilerinden denklem (3.21) ve denklem (3.23)’deki $\hat{\theta}$ tahminlerini bulalım?

Sonsuz sayıda allel modeline göre;

S, ayırma bölgelerinin sayısı 8’dir. Bu bölgeler; 1, 2, 5, 6, 9, 10, 13 ve 14. bölgelerdir.

II , nükleotid uyumsuzluk miktarı; 4’tür. a,b,c,d dizileri arasındaki ikili karşılaştırma sayısı 6’dır. Yani dördün ikili kombinasyonu olup bu ikili karşılaştırmalar; a-b, a-c, a-d, c-d, b-c, b-d şeklindedir. Bu ikili karşılaştırmalardaki nükleotid uyumsuzluğu olan bölgeler; a-b arasında 0, a-c arasında 4, a-d arasında 4, b-c arasında 4, b-d arasında 4, c-d arasında ise 8 tanedir. Bunların toplamı ise $0 + 4 + 4 + 4 + 4 + 8 = 24$ ‘tür. Buradan $II = 24/6 = 4$ olarak bulunur (Hartl ve Clark, 2007).

Bu durumda $S = 8$, $II = 4$ ve buradan $n = 4$ olur. Denklem (3.20)’den α değeri ise;

$$\alpha = 1 + \frac{1}{2} + \frac{1}{3} = 1,833 \text{ olur.}$$

Denklem (3.21)’den $\hat{\theta} = \frac{8}{1,833} = 4,36$ iken; Denklem (3.23)’ten $\hat{\theta} = II = 4$ olur. Bu durumda $\hat{\theta}$ değerleri arasındaki fark yani $II - S/\alpha = 4 - 4,36 = -0,36$ olur. Örneklem büyüklüğü çok az olduğu için hesaplanan değer (-0,36) sıfırdan önemli ölçüde farklı olup olmadığı hakkında tahmin yürütmek için elimizde çok az veri vardır (Hartl ve

Clark, 2007). Tajima D istatistiđi, Student's t dađılımlı gösterir. Tajima's D istatistiđi gibi istatistiklerin rnekleme dađılımlı normal dađılım varsayımına uymuyorsa, Mann Whitney U testi gibi, sıra istatistiklerini kullanan (parametrik olmayan) testler kullanılabilir.



4. SONUÇ VE TARTIŞMA

Bu tez çalışmasında genetik alanında sıklıkla kullanılan özellikle belirli DNA dizileri bakımından populasyon içinde ve populasyonlar arasındaki farklılıkları analiz etmeye yarayan istatistik testler örnekleriyle birlikte tezin Yöntem bölümünde açıklanmıştır. HKa Testi bu alandaki ilk testlerdendir. Bu testlerden en çok tercih edileni Tajima D testidir. Bu testte hesaplanan Tajima D değeri, türler içinde ve türler arasında doğal seçilim olmadığı durumlarda sıfır olmalıdır. FOXP2 geni üzerinde yapılan bir araştırma Tajima D testinin kullanıldığı genetik çalışmalara örnek olarak verilebilir. İnsanlarda bulunan FOXP2 geni mutasyona uğradığında bu mutasyonun görüldüğü bireyler dilbilgisi kurallarına göre cümle kurmakta zorlanır. Bunun dışındaki zihinsel ve fiziksel aktiviteleri yapabilir. Araştırmacılar yaptıkları çalışmada çok sayıda memeliden alınan farklı türlere ait FOXP2 genini incelemişlerdir. Bu genin canlılarda yüksek oranda korunduğu ve incelenen canlılarda aminoasit dizilimlerinin ise neredeyse aynı olduğu sonucuna varmışlardır. İnsanlardaki FOXP2 geninde ise iki mutasyon tespit edilmiştir. Bunlardan birisi, 553. kodonda arginin yerine histidin ikamesi olan tek bir nükleotid mutasyonudur. Diğer de 328. kodonda arginin yerine stop kodonu ikamesi olan bir nonsense (anlamsız) mutasyondur (URL-2, 2024). Her iki mutasyon insanlarda konuşma zorluğuna yol açmaktadır. Araştırmacılar FOXP2 geninin doğal seçilim geçirme olasılığını Tajima D testini kullanarak araştırmıştır. 20 kişiden alınan genin nükleotid dizilerinin analiz edilmesi sonucu 47 SNP, $\hat{\theta}_w = 7,9 \times 10^{-4}$, $\hat{\theta}_T = 3,0 \times 10^{-3}$ tahminleri ve Tajima D değeri -2,20 bulunmuştur. Sonuçlar arasındaki farklılık, %1 düzeyinde önem arz etmektedir. Böylece bu geni yakın zamanda doğal seçilimin etkilediği sonucuna varılabilir (Nielsen ve Slatkin, 2013). Ancak farklı populasyonlarda belirli bir DNA dizisi bakımından gözlenen farklılığın nötr mutasyonlardan mı, yoksa doğal seçilim nedeniyle mi meydana geldiğini söylemek için mevcut testler yeterli görülmemekte, başka testler geliştirmek için bu alanda yeni çalışmaların yapılması gerekmektedir.

Mac Donald –Kreitman (MK) Testi, kodlama bölgelerindeki sinonim olmayan değişikliklerin sinonim değişikliklere oranıyla işlem yapılan bir testtir. Örnek 3.5 MK testi örneğinde sinonim olmayanların sinonimlere oranı sabit ve polimorfik sitelerde

0,333 olarak bulunmuş olup bu oran MK test özelliğiyle uyumludur. Yani bu örnekteki lokusta mutasyon ve şans faktörleri dışında varyasyona sebep olacak başka bir faktörün araştırılması gerekmemektedir. Ancak gerçek verilerden alınan Örnek 3.6 'da ise sabit sitede sinonim olmayanların sinonimlere oranı 0,0476 iken polimorfik sitede bu oran 0,4118 olarak hesaplanmıştır. Bu iki oran arasında önemli bir fark görülme de bu örneğe Fisher' in exact testi ya da Khi – Kare bağımsızlık testi uygulandığında aslında 0,0476 ve 0,4118 arasında istatistik olarak önemli bir fark olduğu görülecektir. Bu da bize istatistik testlerin önemini göstermektedir.

Khi – Kare bağımsızlık testine ait örnek, Örnek 3.2 ve Örnek 3.3'te verilmiş olup Fisher' in exact testi konu bütünlüğü açısından bu teze dâhil edilmemiştir. Örnek 3.3'te Khi – Kare testi kullanılarak gametlerin rastgele farklılaşma ihtimali araştırılmıştır. 1 serbestlik dereceli 1,135 χ^2 değerinin Ek A'daki tablodan 0,50 ile 0,25 arasında bir olasılığa ait olduğu bulunduğundan bu örnekteki populasyonun bağlantı dengesinde olduğu hipotezi kabul edilmiştir. Örnek 3.2 de ise 1 serbestlik dereceli 7,39 χ^2 değerinin Ek A'daki tablodan 0,01 ile 0,001 arasında bir olasılığa ait olduğu bulunduğundan bu örnekteki populasyonun HW dengesine göre beklenen genotip frekanslarında olmadığı kabul edilmiştir.

Genetik analizlerde kullanılan istatistik testlerden elde edilen genetik veriler şu şekilde gruplanabilir:

- Fenotipik Veriler: En tipik örnek Mendel'in bezelyelerle (*Pisum sativum*) yaptığı deneylerle ortaya koyduğu özelliklerdir. Mendel yaptığı çaprazlamalarda yedi adet karakter çeşidi gözlemlemiştir. Bu karakterlerde dominant ve çekinik olanların ayrışma oranlarını da açıklamış ve 3:1 olduğunu ifade etmiştir (Weir, 1996).
- Allozim Veriler: Allozim, proteinlerin farklı allel formlarıdır. Çözünmüş proteinler arasındaki yük farkını kullanarak elektroforez yoluyla elde edilen verilerdir. Bu yöntem ile, proteinlerin %30'unun çözünür olmasına ve DNA dizilerindeki farklılıkların %25'inin elektroforez ile tespit edilebiliyor olmasına rağmen oldukça hızlı ve ucuz veri elde edilir (Weir, 1996).

- Protein Sıra Verileri: Proteinler aminoasit adı verilen yapılardan oluşur. Elektroforez yöntemiyle farklı proteinler ayırt edilebilir. Proteinlerin yapısındaki bu amino asitlerin dizisinin belirlenmesiyle elde edilen veriler günümüzde ulusal biyoteknoloji bilgi merkezi olan NCBI’ da (National Center for Biotechnology Information) saklanmaktadır (Weir, 1996).
- Restriksiyon Parçaları Verileri: DNA dizilerinin restriksiyon enzimleriyle belirli uzunluklarda kesilmesi sonucu oluşan verilerdir. Restriksiyon endonükleazların kullanılması DNA dizileri hakkında ayrıntılı bilgiler elde edilmesini sağlamıştır (Weir, 1996).
- DNA Dizi Verileri: DNA nükleotid adı verilen yapı birimlerinden oluşur. 4 çeşit nükleotid vardır. Bunlar; Adenin, Timin, Guanin ve Sitozindir. Adenin ile Timin, Guanin ile de Sitozin karşılıklı dizilerek çift sarmal DNA molekülünü oluşturur. Bu nükleotid diziliminden elde edilen veriler genetikte önemli bir yere sahiptir. Günümüzde DNA dizi verileri de ulusal biyoteknoloji bilgi merkezi NCBI’ da saklanmaktadır (Weir, 1996).

Tez çalışmamızdan çıkardığımız sonuçlardan biri genetik analizlerin günümüzde ne kadar önemli olduğudur. Genetik analizlerin önemine ait bir örnek Hartl ve Clark (2007)’de verilmiştir. Bu örnekte günümüzde böceklerin pestisit direncine karşı bilim insanlarının yeni pestisit üretme gerekliliğinden bahsedilmiştir. Pestisit direnci böceklerde kimyasal pestisitlere karşı geliştirilmiş bir adaptasyondur. İlk defa 1940 ‘larda uygulanan kimyasal pestisitler böcek ve haşerelerin yaşamı için ciddi bir tehdit oluşturmuştur. Bu durumun sonucunda da bu canlıların tarım ürünlerine verdikleri zarar önemli ölçüde azalmıştır. Ancak 5 ile 50 nesil sonra 400’ den fazla böcek türü bir veya daha çok sayıda pestisite karşı direnç geliştirmiştir. Bu da genetik çalışmaları var olan bu durumun nasıl gerçekleştiğini açıklamaya yöneltmiştir. Ziraat mühendisleri de bu canlılarla etkin mücadele için yeni pestisit geliştirmek zorunda kalmıştır (Hartl ve Clark, 2007). Tüm bu süreçlerde genetik analizlere ihtiyaç vardır. Genetik analizler günümüzde;

- Genlerdeki hastalıklara neden olan yapının araştırılmasında,

- Ekonomik değeri olan bitki ve hayvanların belirli genotiplerinin çoğaltılarak gelecek nesillere aktarılmasında,
- Tüp bebek tedavisinde,
- Bir canlıdan alınan genin başka canlıya aktarılmasında,
- Canlılarda uygulanan klonlama çalışmalarında,
- DNA dizi analizlerinde,
- Hastalıklara karşı aşı ve ilaç geliştirilmesinde,
- Salgın hastalıklarla etkin mücadele amaçlarıyla yapılmaktadır. Bu genetik analizlerden elde edilen verilerin yorumlanmasında ise istatistik testlere ihtiyaç vardır.

Bu tez çalışmasından çıkarılan bir başka sonuç ise: yapay seçilimin türlerin bazı genotiplerinin neslinin devamı üzerine olan etkisinin önemli olduğudur. Geçmişten günümüze bitki ve hayvan yetiştiriciliğinde insanlar verimi yüksek, üstün özellikli genotipleri daha çok tercih ederek yetiştirmiştir. Yapay seçim denilen bu durumla gelecek nesillere tercihen beğenilen genotipler aktarılmıştır. İnsanlar tarafından tercih edilmeyen, ilgi görmeyen, alış verişte alınıp satılmayan türler ya da türlere ait belirli genotiplerin nesli ise giderek azalmış ve hatta yok olmuştur. Burada önemli bir durum da ata tohum diye adlandırılan, genetiği değiştirilmemiş bitki türleri tohumlarının giderek azalıyor olmasıdır. Çünkü insanlar günümüzde üstün özelliklerinden dolayı daha çok biyoteknoloji çalışmaları sonucu üretilen hibrit tohumları tercih etmektedir. Diğer tohumlara göre daha doğal olan bazı ata tohumları giderek azalırken bazı genotiplerin ata tohumu ise günümüzde yok olmuştur. Bu duruma hayvan yetiştiriciliğinde de rastlanmaktadır. Et ve süt verimi diğerlerine göre az olan hayvan genotiplerinin sayısı azalmış, hatta bazılarının nesli tükenmiştir. Özbek (2019) yaptığı çalışmada tarımda sürdürülebilirlik için ata tohumların kıymetinin üretici, tüketici ve yöneticiler tarafından iyi bilinmesi gerektiğini vurgulamış olup bu da bizim sonucumuzla uyumludur. Üreticilerin tohum seçiminde ata tohumu tercih etmelerinin

yanında tüketicilerinde bu alanda bilinçlendirilmesi gerektiğini ve yöneticilerin ise ata tohumlara talebi attıracak sistem geliştirmeleri gerektiğini aktarmıştır.

İncelenen kaynaklarda Poisson dağılımına pek fazla rastlanmamaktadır. Oysa mesela, bir DNA dizisi boyunca nükleotid ikamelerinin sayısının Poisson dağılımı gösterdiği düşünülürse Poisson dağılımı genetik çalışmalarda daha fazla önemsenmelidir.

Bu tez çalışması sonunda vardığımız nihai sonuç, genetik analizlerde kullanılan istatistik yöntemlerin, bir taraftan kullanımını yaygınlaştırmak, diğer taraftan da daha net değerlendirmelere yarayacak yönde geliştirilmesini ve çoğaltılmasını sağlamak için çalışmalara ihtiyaç olduğudur.

KAYNAKLAR

- Bardak, A., Çelik, S., Erdoğan, O., Ekinci, R., & Dumlupınar, Z. (2021). Association Mapping of Verticillium Wilt Disease in a Worldwide Collection of Cotton (*Gossypium hirsutum* L.). *Plants*, 10(2), 306.
- Bayram, N. (2015). Araştırmalarda istatistik ve etik. *Endüstri İlişkileri ve İnsan Kaynakları Dergisi*, 17(3), 20-29. doi:10.4026/1303-2860.2015.0286.x.
- Can, S. (2023). Belirli bir DNA dizisi bakımından popülasyonlar arası farklılık üzerine bir simülasyon çalışması. Yüksek Lisans Tezi, *Kastamonu Üniversitesi Fen Bilimleri Enstitüsü*. Kastamonu.
- Demirhan, B. (2017). Patates ve işlenmiş patates ürünlerinde genetiği değiştirilmiş organizmalarla ilgili genetik analizler. Yüksek Lisans Tezi, *Ondokuz Mayıs Üniversitesi Fen Bilimleri Enstitüsü*. Samsun.
- Düzgüneş, O., & Kavuncu, O. (1977-1978). *İstatistik sınıf I*. Yaygın Yükseköğretim Kurumu.
- Ercan, M. (1997). *Bilimsel araştırmalarda istatistik*. (2. Baskı), Kavak ve Hızlı Gelişen Tür Orman Ağaçları Araştırma Enstitüsü Müdürlüğü.
- Esmeray, M. (2016). Mısır heterotik gruplarında genetik analizler. Doktora Tezi, *Namık Kemal Üniversitesi Fen Bilimleri Enstitüsü*. Tekirdağ.
- Eymen, U. E. (2007). *SPSS 15.0 veri analiz yöntemleri*. https://yunus.hacettepe.edu.tr/~tonta/courses/spring2009/bby606/SPSS_15.0_ile_Veri_Analizi.pdf, Erişim Tarihi: 07/05/2023.
- Gedik, Y., & Kavuncu, O. (2021). Determination of mucopolysaccharidosis IIID in some goat breeds. *Turkish Journal of Veterinary & Animal Sciences*, 45(5), 964-968. doi:10.3906/vet-2101- 41.
- Griffiths, A.J.F., Wessler, S.R., Carroll, S.B., & Doebley, J. (2012). *Introduction to Genetic Analysis*. Onuncu Baskı, New York: W. H. Freeman and Company.
- Gündüz, S. (2023). QTL haritalama yöntemlerinin geçerliliği üzerine bir simülasyon çalışması. Yüksek Lisans Tezi, *Kastamonu Üniversitesi Fen Bilimleri Enstitüsü*. Kastamonu.
- Hamilton, M. B. (2009). *Population Genetics*. (1. Baskı), Wiley–Blackwell.

- Hartl, D. L., & Clark, A.G. (2007). *Principles of population genetics*. (4. Baskı), Sunderland, Massachusetts: Sinauer Associates, Inc. Publishers.
- Kavuncu, O. (2018). İstatistik uygulamalarda dikkat edilmesi gereken hususlar. *Kastamonu University Journal of Engineering and Sciences*, 4(2), 22-26.
- Kavuncu, O. (2020). *Genetik*. (1. Baskı), Kamer Yayınları.
- Kavuncu, O. (2021). *Temel ve uygulamalı istatistik*. (1. Baskı), Nobel Yayınları.
- Kavuncu, O. (2022). *Populasyon genetiği ve kantitatif genetik*. (1. Baskı), Nobel Yayınları.
- Kaymaz, D. (2022). Primer immün yetmezlik tanısında genetik yöntemlerin yeri ve önemi. Tıpta Uzmanlık Tezi, *Ankara Üniversitesi Tıp Fakültesi*. Ankara.
- Nielsen, R., & Slatkin, M. (2013). *An introduction to population genetics*. (1. Baskı), Sunderland, Massachusetts U.S.A.: Sinauer Associates, Inc. Publishers.
- Özbek, H. (2019). Tarımsal biyoçeşitliliğin korunmasında yerel tohum kullanımı ve üretici farkındalığı (Ankara ili Ayaş ilçesi örneği). Yüksek Lisans Tezi, *Tokat Gaziosmanpaşa Üniversitesi Fen Bilimleri Enstitüsü*. Tokat.
- Özen, S. (2020). Osteogenezis imperfektali hastalarda yeni nesil dizi analizi yöntemi ile hedeflenmiş moleküler genetik tanı ve sorumlu yeni genlerin araştırılması. Doktora Tezi, *Ege Üniversitesi Sağlık Bilimleri Enstitüsü*. İzmir.
- Samsunlu, E.T. (2021). Gen terapisinde CRISPR-Cas-9. *Manisa Celal Bayar Üniversitesi Sağlık Bilimleri Enstitüsü Dergisi*, 8(3), 574-580.
- Terzi, Y. (2018). *Temel istatistik I ders notları*. <https://ist-fen.omu.edu.tr/tr/hakkimizda/ders-notlari/TEMEL%20%C4%B0ST%20I%202018.pdf>, Erişim tarihi: 07/05/2023.
- URL-1. CRISPR- CAS9 NEDİR?. <https://uskudar.edu.tr/tr/crispr-cas9>, Erişim tarihi: 06/05/2023.
- URL-2. Takahashi, H., Takahashi, K., & Liu, F.C. *FOXP Genes, Neural Development, Speech and Language Disorders*, <https://www.ncbi.nlm.nih.gov/books/NBK7023/#:~:text=FOXP2%20gene%20mutation%20is%20so,involvement%20in%20speech%20and%20language>, Erişim tarihi: 22/04/2024.
- Weir, B. S. (1996). *Genetic data analysis II*. Sunderland, MA: Sinauer Assoc, Inc.

Yılmaz, A. (2010). Mikroarray tabanlı yeniden dizileme sistemi ile Wilson hastalarında ATP7B geninde moleküler genetik analizler. Doktora Tezi, *Hacettepe Üniversitesi Biyoloji Anabilim Dalı*. Ankara.





EKLER

EK A. Farklı Serbestlik Derecelerindeki χ^2 Değerleri ve İlişkili Olasılıklar

Tablo A 1 Farklı serbestlik derecelerindeki χ^2 değerleri ve ilişkili olasılıklar (Hamilton, 2009)

Serbestlik derecesi	0,5	0,25	OLASILIKLAR 0,1	0,05	0,01	0,001
1	0,4549	1,3233	2,7055	3,8415	6,6349	10,8276
2	1,3863	2,7726	4,6052	5,9915	9,2103	13,8155
3	2,3660	4,1083	6,2514	7,8147	11,3449	16,2662
4	3,3567	5,3853	7,7794	9,4877	13,2767	18,4668
5	4,3515	6,6257	9,2364	11,0705	15,0863	20,5150