YYÜ TAR BİL DERG
(YYU J AGR SCI)
2017, 27(4): 531-542

Geliş Tarihi (Received): 22.08.2017
Kabul Tarihi (Accepted): 01.11.2017
doi: 10.29133/yyutbd.335745

Araştırma Makalesi/*Research Article (Original Paper)*

Beagle Genotip Atama Yönteminin Değerlendirilmesi ve Bir Uygulama

Sinem İZDEŞ BARANSEL[1], Gazel SER[2*]

[1]Van Yüzüncü Yıl Üniversitesi Fen Bilimleri Enstitüsü, Zootekni ABD, Van, Türkiye
[2]Van Yüzüncü Yıl Üniversitesi, Ziraat Fakültesi, Zootekni Bölümü, 65100 Van, Türkiye
[*]e-posta: gazelser@yyu.edu.tr

Özet: Bu çalışma, genotip atama yöntemlerinden olan Beagle programının tanıtılarak, oluşturulan üç farklı senaryoda atama performansının değerlendirilmesi amacıyla gerçekleştirilmiştir. Populasyon tabanlı atama yöntemleri içerisinde yer alan Beagle programı eksik tek nükleotit polimorfizmler (Single Nucleotid Polimorfizm, SNP) ve gözlemlenmiş SNP'ler arasındaki bağlantı dengesizliği (Linkage Disequilibrium, LD) bilgileri ile lokal haplotip küme (Browning) modelini kullanarak, eksik genotipleri tahmin etmede yüksek oranda atama doğruluğu sunmaktadır. Bu amaçla, çalışmada 1000 Genom Projesi'nde dizilenmiş 191 bireyin 22. kromozomu üzerinde bulunan 1356 SNP bilgileri kullanılarak, veri seti %20, %50 ve %70 oranlarında şansa bağlı olarak kesilerek üç farklı senaryo oluşturulmuştur. Üç senaryo sonucunda elde edilen Alelik-R^2 değerleri, %90'dan büyük olarak elde edilmiş ve yüksek doğruluk derecesine sahip atamalar yapılmıştır. Çalışmada tanıtımı amaçlanan Beagle programında referans veri setlerinin büyüklükleri arasındaki farklılıkların, atama doğruluğu üzerine belirgin bir etkisi saptanamamıştır. Sonuç olarak, Beagle programında farklı büyüklüklere sahip örneklerden elde edilen sonuçlar doğrultusunda, yüksek doğruluk derecesinde atamalar yapabildiğini göstermiştir.

Anahtar kelimeler: Beagle, Genotip atama yöntemleri, Hedef veri seti, Referans veri seti

Evaluation of Beagle Genotype Imputation Method and An Application

Abstract: This study describes the genotype imputation method using the Beagle program and was completed with the aim of evaluating the imputation performance in three different scenarios. A population-based imputation method, the Beagle program uses linkage disequilibrium (LD) information between missing single nucleotid polimorfizm (SNP) and observed SNPs, the local haplotype cluster (Browning) model to offer high rates of imputation accuracy for estimation of deficient genotypes. With this aim, the study used 1356 SNP data from the 22nd chromosome of 191 individuals in the 1000 Genome project, cutting the dataset at random rates of 20%, 50% and 70% to create three different scenarios. The Allelic-R^2 values obtained as a result of the three scenarios were larger than 90% and imputations with high degree of accuracy were made. In this study the differences in the size of the reference datasets in the Beagle program were not identified to have a clear effect on imputation accuracy. In conclusion, in terms of results obtained from samples with different sizes using the Beagle program, imputations with high accuracy could be made.

Keywords: Beagle, Genotip imputation methods, Target data set, Reference data set

Giriş

Genetik alanında eksik gözlem tahmin yöntemleri, eksik Tek Nükleotid Polimorfizm (Single Nucleotid Polimorfizm, SNP) verilerinin tahmin edilmesine dayanan istatistiki yöntemlerdir. Bu yöntemler, genetik ilişkilendirme çalışmalarının gücünü ve kararlılığını artırmak için genom boyu ilişkilendirme (Genome Wide Association Study, GWAS) çalışmalarında yoğun olarak kullanılmaktadır. Eksik SNP verisine sahip bireylerin, örneklem kümesinden çıkarılması örneklem büyüklüğünde önemli bir azalışa sebep olmaktadır. Genotiplendirme çalışmalarının yüksek maliyetleri ve bazı durumlarda çalışmanın tekrarının mümkün olmaması gibi nedenler, genotip atama yöntemlerine olan ilgiyi arttırmaktadır (Marchini ve Howie 2010).

Genotip atama yöntemleri, eksik gözlemlerin tahminine ilişkin bir süreç olduğundan yapılan atamanın doğruluğu önemlidir. Atama işleminin doğruluğu; çalışılan populasyonun yapısına, referans populasyonun büyüklüğüne, eksik genotiplerin oranına, yüksek yoğunlukta genotiplenmiş bireylerin ve yakınlarının sayısına, bağlantı dengesizliği (linkage disequilibrium, LD) panellerindeki SNP sayısına, doğru ve atanmış genotipler arasındaki korelasyonun önemliliğine, referans ve hedef populasyon arasındaki ilişkiye bağlıdır. Atama yöntemleri, genellikle daha az yoğun genotiplenmiş çalışma örneğinin içinde bulunan eksik genotipleri tahmin

etmek amacıyla, daha yoğun genotiplerin oluşturduğu referans panelini kullanmaktadır. Dolayısıyla referans paneli ve atama yönteminin seçimi, atama doğruluğunu etkileyen en önemli faktördür (Morris ve Ramsay 2010; Pausch ve ark. 2013; Chud ve ark. 2015).

Genotip atama yöntemleri, aile tabanlı (family-based) ve populasyon tabanlı (population-based) atama yöntemleri olarak iki ana gruba ayrılsa da, bazı durumlarda her iki yöntemin kombinasyonu kullanılabilmektedir. Her iki yöntemdeki atama işlemi, atanacak genotiplerle referans populasyondaki SNP'ler arasındaki LD ilişkisine göre yapılmaktadır. Aile tabanlı ve populasyon tabanlı olarak kullanılan yöntemler için farklı yazılım programları geliştirilmiştir. Bu programlar, aynı zamanda genotip atama yöntemleridir. Findhap, AlphaImpute, FImpute şeklinde olan aile tabanlı atama yöntemleri; Mendelyan segresyon kurallarını ve pedigri bilgisini kullanarak atama yapmaktadır. Bu üç yöntem içerisinde atama doğruluğu en yüksek olan ve dolayısıyla araştırıcılar tarafından en çok tercih edilen FImpute programıdır. Beagle, Mach, Impute, fastPHASE atama yöntemleri populasyon tabanlı atama yöntemleridir ve eksik SNP'ler ile gözlemlenmiş SNP'ler arasındaki bağlantı dengesizliği (LD) bilgilerini kullanmaktadırlar. Bu yöntemler arasında en çok tercih edilen, Beagle programıdır (Sun ve ark. 2012; Chud ve ark. 2015).

Bu çalışmada, Beagle programının tanıtımı ve oluşturulan üç farklı senaryoda atama performansının değerlendirilmesi amaçlanmıştır.

Materyal ve Yöntem

Bu çalışmada ki SNP verileri, Browning (2016) tarafından oluşturulan Beagle programının (versiyon 4.1) internet sayfasında yer alan örnek veri setinden elde edilmiştir. Veri seti, 1000 Genom Projesi'nde dizilenmiş bireylerin 37. kromozomlarının genotip bilgilerinden yararlanarak, rastgele seçilen 191 bireyin 22. kromozomu üzerinde 20000086. ve 20099941. pozisyonları arasında bulunan 1356 SNP bilgileri kullanılmıştır. Çalışmamızda, rastgele seçilen üç SNP belirtecine (rs138720731, rs55902548 ve rs187930998) ilişkin atamalar değerlendirilmiştir.

Çalışmada Kullanılacak Referans ve Hedef Veri Setlerinin Oluşturulması

Hedef ve referans veri setleri Beagle program kodları kullanılarak düzenlenmiştir. Çalışmada hedef ve referans veri setlerinin belirlenmesinde üç farklı senaryo oluşturulmuştur. Senaryolar, hedef ve referans veri setlerinin farklı büyüklüklerde ya da eşit olması durumunda, atama doğruluğunun nasıl etkilediği üzerine kurgulanmıştır.

Birinci senaryo: Veri setinde yer alan 191 bireyden, şansa bağlı olarak seçilen %20'lik kısım (38 birey) hedef veri seti olarak belirlenirken, geri kalan kısım (153 birey) referans veri seti olarak belirlenmiştir

- Referans dosyasını oluşturma kodu: *zcat test.16Jun16.7e4.vcf.gz | cut -f1-162 | tr '/' '|' | gzip > ref.16Jun16.7e4.vcf.gz*
- Hedef dosyasını oluşturma kodu: *zcat test.16Jun16.7e4.vcf.gz | cut -f1-9,163-200 | gzip > target.16Jun16.7e4.vcf.gz*

İkinci senaryo: Veri setinde yer alan bireylerden, şansa bağlı olarak seçilen %50'lik kısım (96 birey) hedef veri seti olarak belirlenirken, geri kalan kısım (95 birey) referans veri seti olarak belirlenmiştir.

- Referans dosyasını oluşturma kodu: *zcat test.16Jun16.7e4.vcf.gz | cut -f1-104 | tr '/' '|' | gzip > ref.16Jun16.7e4.vcf.gz*
- Hedef dosyasını oluşturma kodu: *zcat test.16Jun16.7e4.vcf.gz | cut -f1-9,105-200 | gzip > target.16Jun16.7e4.vcf.gz*

Üçüncü senaryo: Veri setinde yer alan bireylerden şansa bağlı olarak seçilen %70'lik kısım (134 birey) hedef veri seti olarak belirlenirken, geri kalan kısım (57 birey) referans veri seti olarak belirlenmiştir.

- Referans dosyasını oluşturma kodu: *zcat test.16Jun16.7e4.vcf.gz | cut -f1-66 | tr '/' '|' | gzip > ref.16Jun16.7e4.vcf.gz*
- Hedef dosyasını oluşturma kodu: *zcat test.16Jun16.7e4.vcf.gz | cut -f1-9,67-200 | gzip > target.16Jun16.7e4.vcf.gz*

Populasyona Tabanlı Atama Yöntemi-Beagle

Beagle, haplotip faz çıkarsama ve genetik ilişki analizi çalışmalarının yapılması için geliştirilen bir yazılım programıdır. Binlerce örneğin, yüzbinlerce genotiplenmiş belirteçleri ile büyük ölçekli veri setlerini analiz etmek için tasarlanmış olan Beagle;

- Birbiriyle ilişkisi olmayan bireyler, ebeveyn-çocuk çifti ve ebeveyn-çocuk üçlüsü için genotip veriyi aşamalı olarak uygular.
- Sporadik eksik genotip verilerini çıkarsar.
- Bir referans panelinde genotiplenmiş olan, genotiplenmemiş belirteçleri atar.
- Tek belirteç ve haplotipik ilişki analizini gerçekleştirir.
- Homozygous-by-descent (HBD) veya identical-by-descent (IBD) olarak paylaşılan genetik bölgeleri bulur.

Fazı bilinmeyen genotip verisi, fazı bilinen genotip verisi, ebeveyn-çocuk ikilisi ve ebeveyn-çocuk üçlüsü verileri, Beagle'da karıştırılıp birleştirilmektedir. Beagle, her veri türü için eksik genotipleri ve genotiplenmemiş belirteçleri atayarak haplotipleri tahmin etmektedir. Genotipleri, fazlarına ayırma ve atama işlemleri yapılırken birbiriyle ilişkisiz veriler için Beagle genotip olabilirliklerin yerine çağrılmış genotipleri kullanılmaktadır (Browning 2011).

Yöntem, eksik genotip belirteçlerinin atama işleminde "lokalize haplotip kümeleme" modelini kullanmaktadır. Birbirleriyle yakın bağlantı dengesizliğindeki (linkage disequilibrium, LD) belirteçleri yakalamak için sadece lokal haplotip verilerini kullanmaktadır. Beagle, ortalama haplotip sayılarına dayanan bir düğümden, diğerine geçişlerdeki olasılıkların belirlenmesinde saklı markov modelini (Hidden Markov Model, HMM) kullanmaktadır. Bu durumda, düğüm belirli bir lokustaki aynı alellere sahip haplotip topluluğudur. Dolayısıyla, verilen düğüm, bir birini izleyen alellerin olasılıklarının belirlenmesinde kullanılmaktadır. Beagle, her bireyin haplotiplerini belirlemek için aşamalı bir algoritma kullanmaktadır. Böylece, lokal haplotip kümelerini elde edebilmektedir. Her birey için HMM kullanılarak, haplotip kümesi oluşturulur. Her birey için kümelenen haplotipler, lokal haplotip kümeleri için tekrar yapılandırılarak kullanılmaktadır. Beagle'da Markov zinciri, yüksek doğruluk düzeyinde atama yapılabilmesi amacıyla 10 iterasyonun üzerinde çalıştırılmaktadır (Larmer ve ark. 2010; Chud ve ark. 2015).

Beagle tarafından kullanılan lokal haplotip küme modeli, Browning modeli olarak da isimlendirilmektedir. Bu modelde, gözlenen haplotipler yerel çerçevede belirteçlerin haplotip benzerliğine dayalı olarak, her belirtecin konumunda kümeler halinde gruplandırılmıştır. Bir belirteçten, diğer belirtece model boyunca hareket ederken, küme üyeliği tarihsel rekombinasyon veya mutasyon olayları nedeniyle bazı değişikliklerde, kararlı kalma eğilimindedir. Browning modelinde, rekombinasyon fraksiyonları gibi herhangi bir açık parametre yoktur. Bunun yerine model, kümeler ve kümeler arası geçişlerde gözlenen frekanslar ile birlikte, bunların arasındaki olası geçişlerle temsil edilmektedir. Mach, Impute, fastPHASE gibi diğer populasyon tabanlı atama yapan yöntemlerin kullandığı, Li ve Stephens çerçevesi ile Beagle tarafından kullanılan Browning modeli arasında bazı önemli farklılıklar vardır. Bu farklılıkların ilki, her bir belirteçte konumların sayısı değişebilir. Bu Beagle'a hesaplama yükünü en aza indirme avantajı sağlarken, farklı yerlerdeki karmaşıklığın farklı düzeylerde modellemesine izin vermektedir. Diğer önemli farklılık ise Browning modelinde, gizli konumlar yani lokalize haplotip kümeleri sadece alelin tek tipini belirtmektedir. Böylece, modelin konumları herhangi bir gözlenen mutasyon içerecek olmasına rağmen, mutasyon açıkça modellenmiş olmaz. Browning modeli, Li ve Stephens çerçevesine göre daha tutucu bir modeldir. Böylece, Browning modelinde daha az parametre ile daha hızlı hesaplama süresi sonuçları elde edilmektedir (Browning 2008).

Beagle'da Atama Doğruluğunun Belirlenmesi

Beagle, atama doğruluğunun test edilmesi için çıktı dosyasında iki önemli parametre vermektedir. Bunlardan ilki, atama doğruluğunun değerlendirilmesinde kullanılan en basit ve güçlü araç olan Alelik R^2 (AR^2) istatistiğidir. Bu istatistik aynı zamanda, Beagle-R^2 olarak da bilinmektedir. Eksik SNP verilerinin doğru tahmini için, gözlenmeyen doğru genotipler ve atanmış genotiplerin alel dozları arasındaki korelasyonun karesi alınarak hesaplanmaktır. Alelik-R^2 değeri, $0 \leq R^2 \leq 1$ arasında değişim göstermektedir. Alelik-R^2 değerinin büyük (1'e yakın olması) olması genotip atama doğruluğunun en önemli göstergesidir. Diğeri ise Doz-R^2 (DR^2) değeridir. Doz-R^2, atanan genotipler, 0 ve 2 aralığında bir değere sahip olan B-alel dozu tarafından değerlendirilir ve Beagle tarafından tahmin edilen posterior genotip olasılıkları ($([P(RA)+2*P(RR)])$)

kullanılarak hesaplanmaktadır. Buna göre, Alelik R^2 (AR^2) istatistiği aşağıda verilen eşitlik kullanılarak hesaplanır.

$$AR^2 = \frac{\left[\left(\sum g_n e_n - (1/N)\right)\left(\sum g_n \sum e_n\right)\right]^2}{\left[\sum f_n - (1/N)\left(\sum e_n\right)^2\right]\left[\sum z_n - (1/N)\left(\sum z_n\right)^2\right]} \quad (1)$$

Eşitlik 1'de g_n doğru alel dozu, e_n atanan alel dozunu gösterirken, z_n ise 0,1, ya da 2 şeklinde kodlanan alellerin kopyaların sayısına karşılık gelen, genotiplerin atamadaki en yüksek posterior olasıklarını gösterir. $f_n = p_{n1} + 4p_{n2}$ şeklinde hesaplanır ve p_{nk}, n'nci örneğe karşılık gelen k sınıftaki (0,1,2) genotiplerin atanmış olasıklarını ifade etmektedir (Browning ve Browning 2009; Ramnarine 2016).

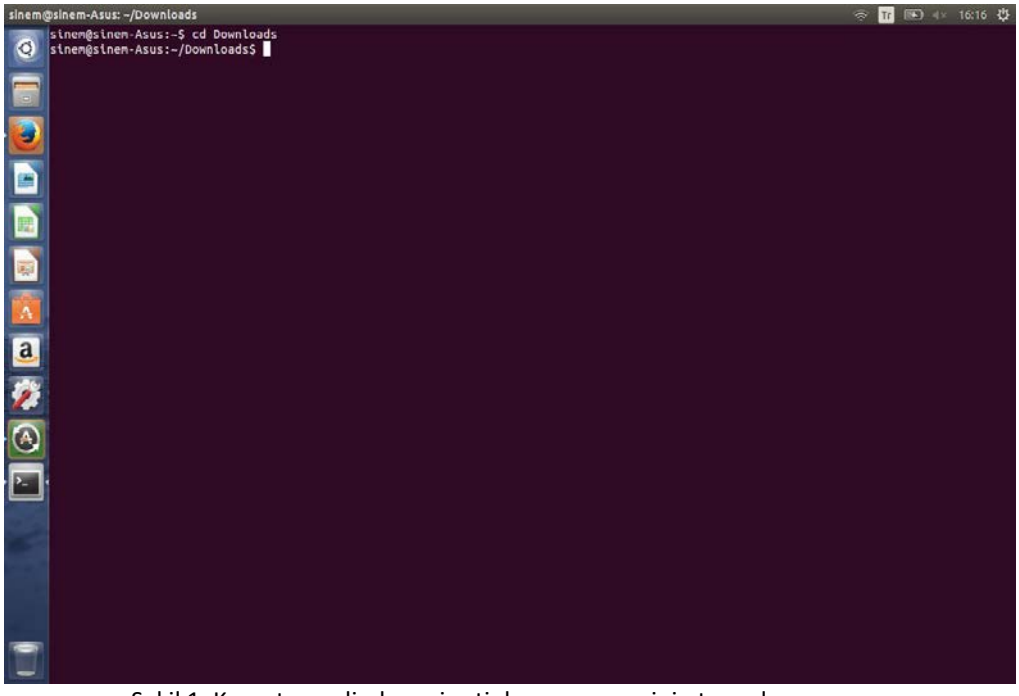
Bulgular ve Tartışma

Beagle'ın Çalışma Prensibi

Beagle, Java programlama dili kullanan bir genotip atama programıdır. Program Windows, Unix, Linux, Solaris ve Mac gibi işletim sistemlerinde çalışabilmektedir. Beagle, Java Standart Edition (SE) ve Runtime Enviroment (JRE) platformlarını kullanmaktadır. Beagle programının bir ara yüzü yoktur, bundan dolayı komut paneli üzerinde çalışmaktadır. BEAGLE programı, varyant çağırma formatındaki (Variant Call Format, VCF) verileri kullanmaktadır. Varyant çağırma formatı (Variant Call Format, VCF), sıkıştırılırmış bir şekilde depolanan bir metin formatıdır. Genomdaki konumlar hakkında meta-bilgi hatlarını, bir başlık satırını ve veri hatlarının her birini içerir. Aynı zamanda, VCF her pozisyon için örneklerde genotip bilgilerini içeren bir özelliğe sahiptir. Beagle programının, uygulama basamakları aşağıda sırası ile verilmiştir.

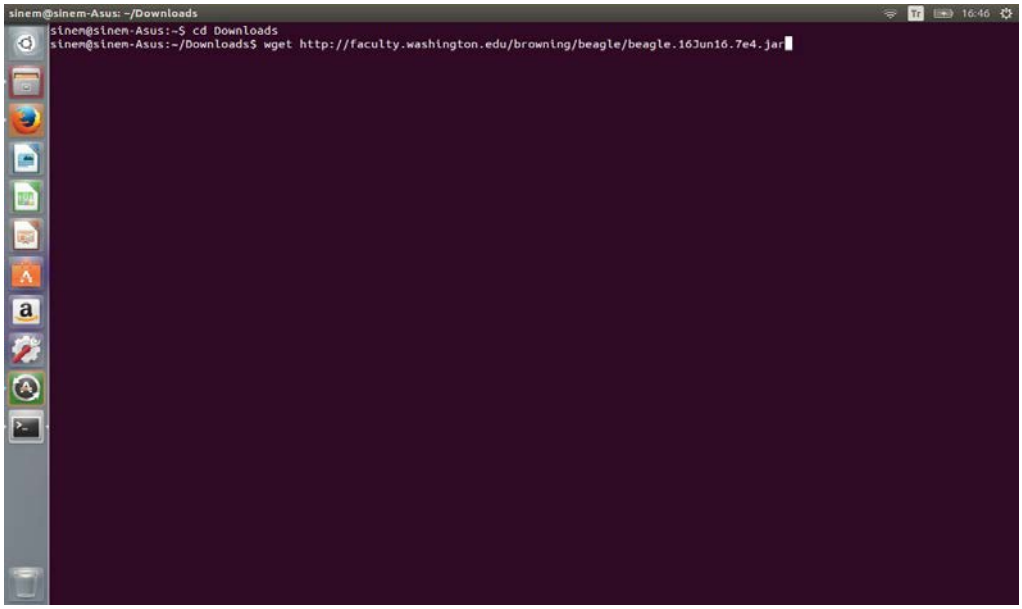
Veri Setinin Tanımlanması

Beagle'da, öncelikli olarak komut panelinde, veri setinin bulunduğu dosya tanıtılır. Bu tanıtma işlemiyle, sonraki adımlarda elde edilecek çıktılar bu dosya içerisinde yer alması sağlanmaktadır (Şekil 1).

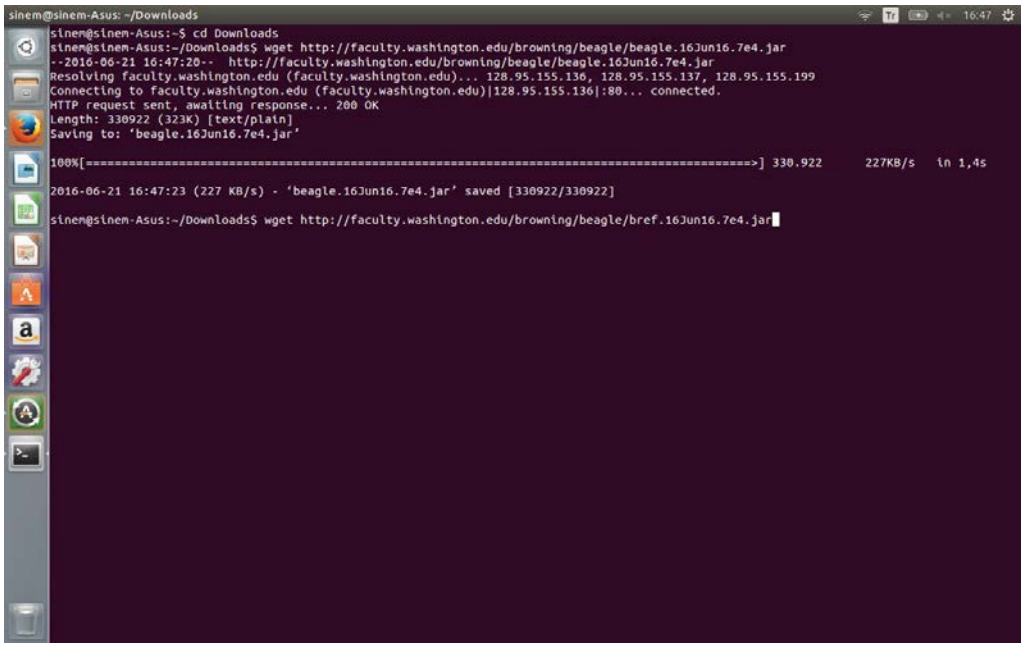


Şekil 1. Komut panelinde veri seti dosyasının yerinin tanımlanması.

İkinci adımda ise Şekil 2'de ".jar" uzantılı beagle ve Şekil 3'de de bref dosyaları tanımlanır.



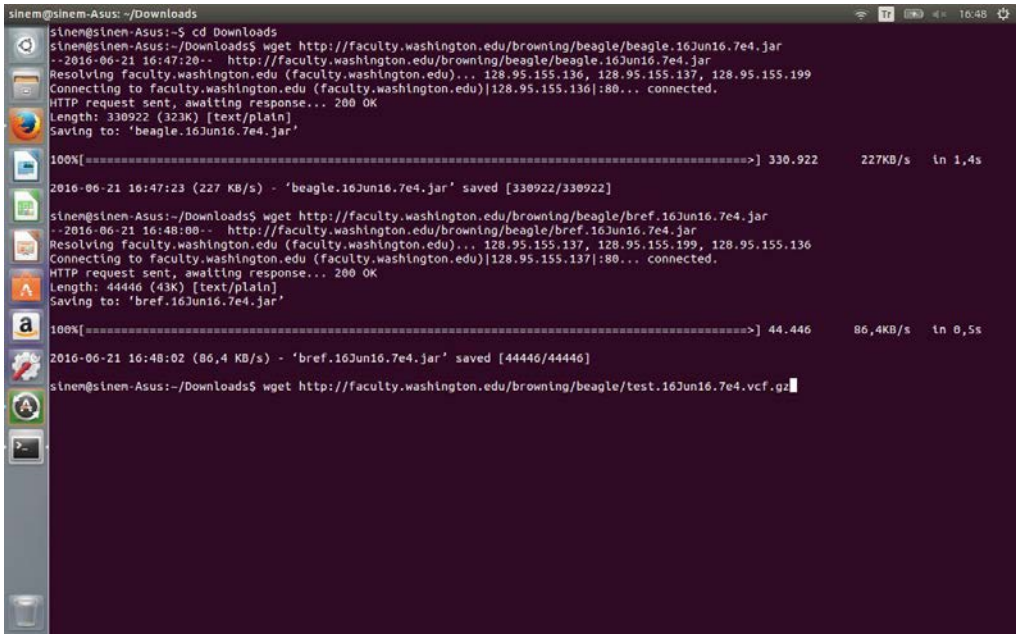
Şekil 2. ".jar" uzantılı Beagle dosyasının tanımlanması.



Şekil 3. ".jar" uzantılı bref dosyasının tanımlanması.

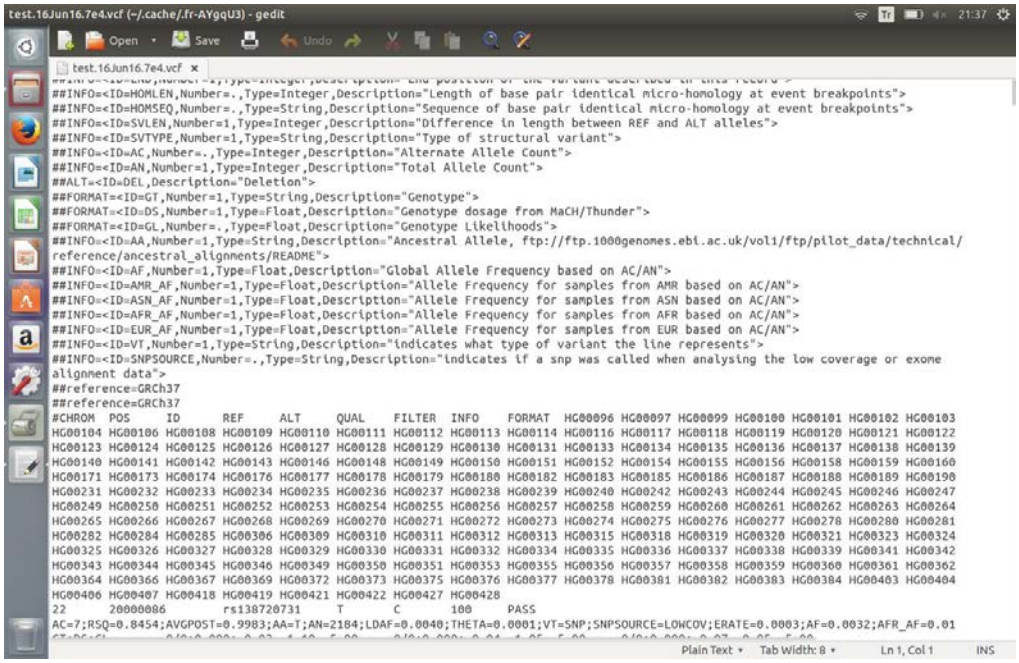
Veri Setinin Oluşturulması

Komut paneline veri setinin bulunduğu, internet adresi yapıştırılarak örnek veri setinin bilgisayara çağrılması ve ilk adımda tanıtılan dosya içerisine kaydedilmesi sağlanır. Verinin çağrılmasına ilişkin gösterim Şekil 4'de verilmiştir.

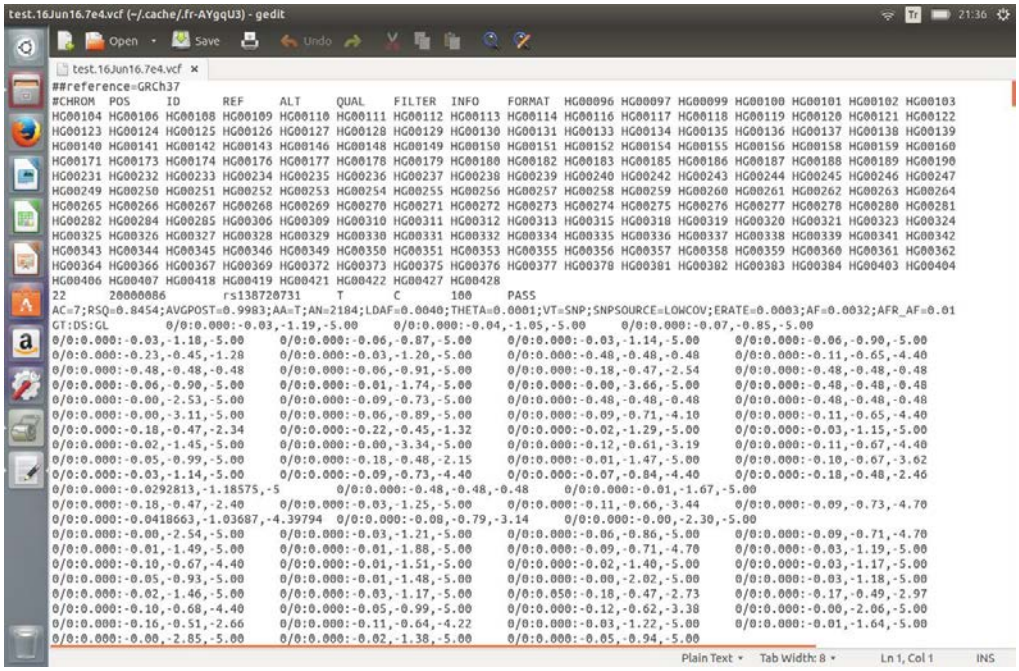


Şekil 4. Komut panelinde veri setinin çağırılması.

Klasör içerisine indirilen veri dosyası ".vcf" uzantısı olarak yer almaktadır. Dosya açıldıktan sonra, veri setine ilişkin bilgiler ve veri seti görülür. Aşağıda Şekil 5'de veri setini tanıtan özellikler ve Şekil 6'da ise veri setinin görünümü verilmiştir.



Şekil 5. Veri setinin tanımlama bilgilerinin olduğu görünüm.



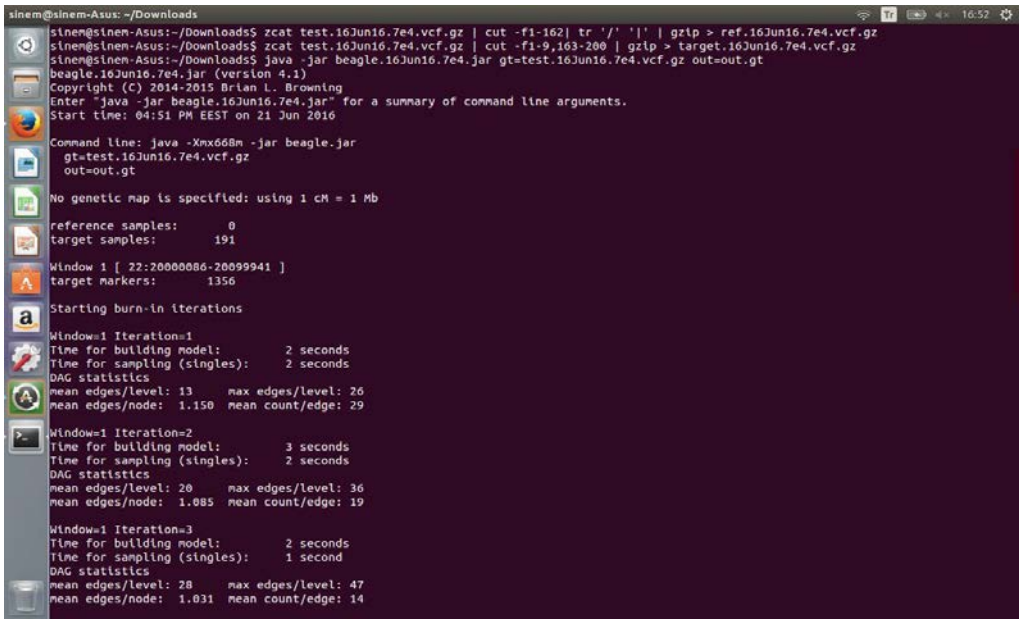
Şekil 6. Veri setinin görünümü.

Genotip (GT) Komut Satırı Argümanına İlişkin Sonuçlar

Genotip (GT) argümanında, çalışma veri setindeki 191 örneğin 15 iterasyonla genotiplerin alellerini belirler. GT formatındaki her belirteç için bir VCF çıktı dosyası bulunur. Bir genotipin aşamalı alel ayırıcısını "|" içeriyor olması, analizin devam ettiği süre boyunca genotip fazını koruduğunu belirtir. Çıktı dosyasında, GT argümanı ile elde edilen tüm genotipler hem aşamalı hem de eksiksizdir. Çıktı dosyasında aleller, referans alel ile aynıysa "0", referans alelin alternatif alellerinden biri ise "1", diğeri ise "2" olarak kodlanır ve aşağıdaki kod ile çalıştırılır.

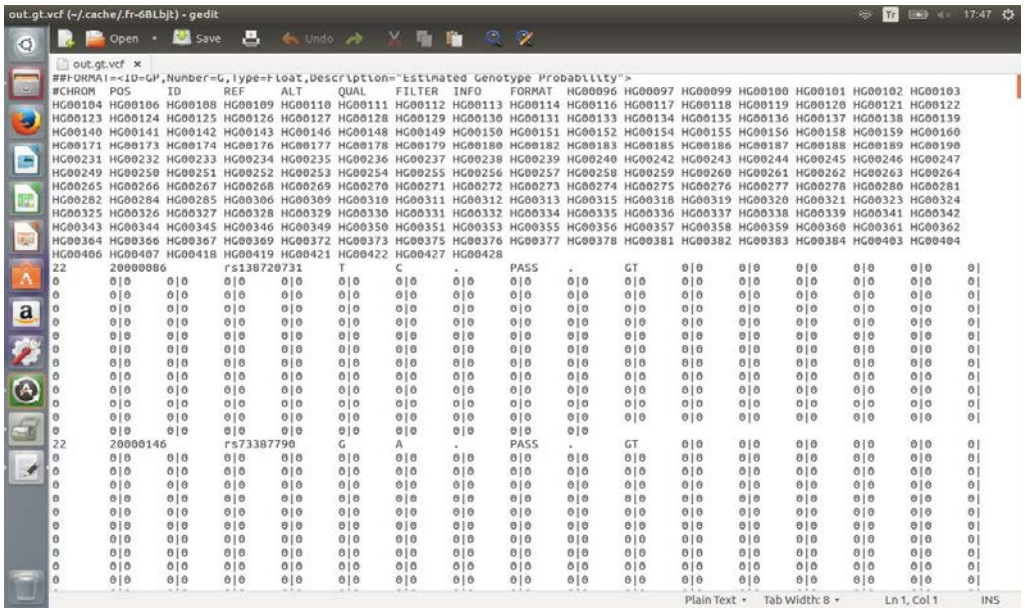
- *java -jar Beagle.16Jun16.862.jar gt=test.16Jun16.862.vcf.gz out=out.gt*

Genotip (GT) argümanının çalıştırılması sonucunda iki çıktı elde edilmektedir. İlk çıktı, iterasyon basamakları, kodların çalışma süresi gibi bilgileri içeren dosya Şekil 7'de verilmiştir.

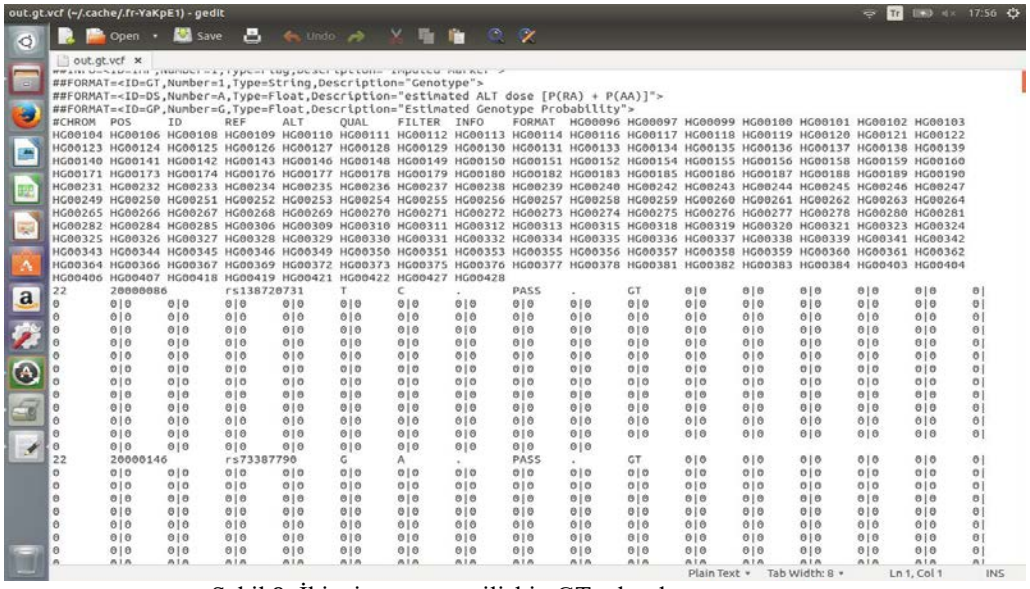


Şekil 7. GT argümanının çalıştırılmasından elde edilen örnek bir görünüm.

Her üç senaryo için elde edilen elde edilen GT çıktı sonuçlarına ilişkin örnek görünümler sırasıyla Şekil 8-10'da verilmiştir.

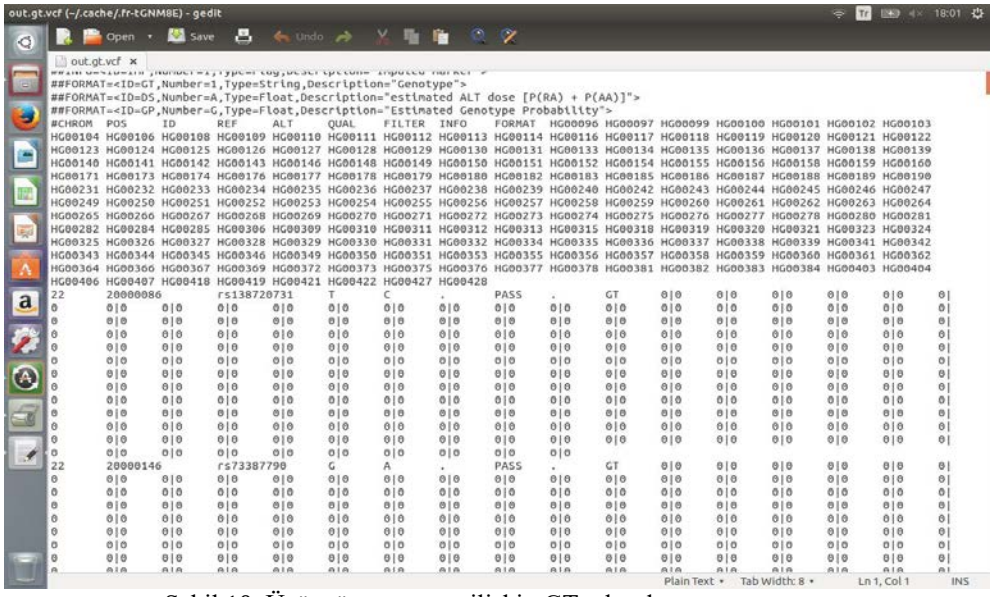


Şekil 8. Birinci senaryoya ilişkin GT çıktı dosyası.



Şekil 9. İkinci senaryoya ilişkin GT çıktı dosyası.

Şekil 8-10 incelendiğinde, birinci, ikinci ve üçüncü senaryolar için rastgele seçilmiş ve Pürin (Adenin (A), Guanin (G)) ve Primidin (Timin(T), Sitozin(C)) nükleotid ilişkisi göz önüne alınarak 22. kromozom üzerinde 20000086'ncı pozisyonda bulunan rs138720731 numaralı SNP verisinde referans alel T (timin), alternatif alel ise C (sitozin) olarak belirtilmiştir. Qual değeri default olarak verilmiştir ve Phred ölçeğine göre ALT çağrısı doğru olarak yapılmıştır. Herhangi bir filtreleme ve ek bilgi yoktur. GT değerlerinde 0|0 gösterimi bireylerin, referans alel ile aynı alele (T) sahip olduğunu göstermiştir. Aynı şekilde rastgele seçilen 22. kromozom üzerinde bulunan bir diğer rs55902548 numaralı SNP verisinde referans alel G (guanin), alternatif alel ise T (timin) olarak belirtilmiştir. GT değerlerinde 0|1 gösterimi bireylerin, ilk alelinin referans alel ile aynı alele (G), ikinci alelin ise ALT'de listelenen ilk alele yani T aleline sahip olduğunu göstermiştir.



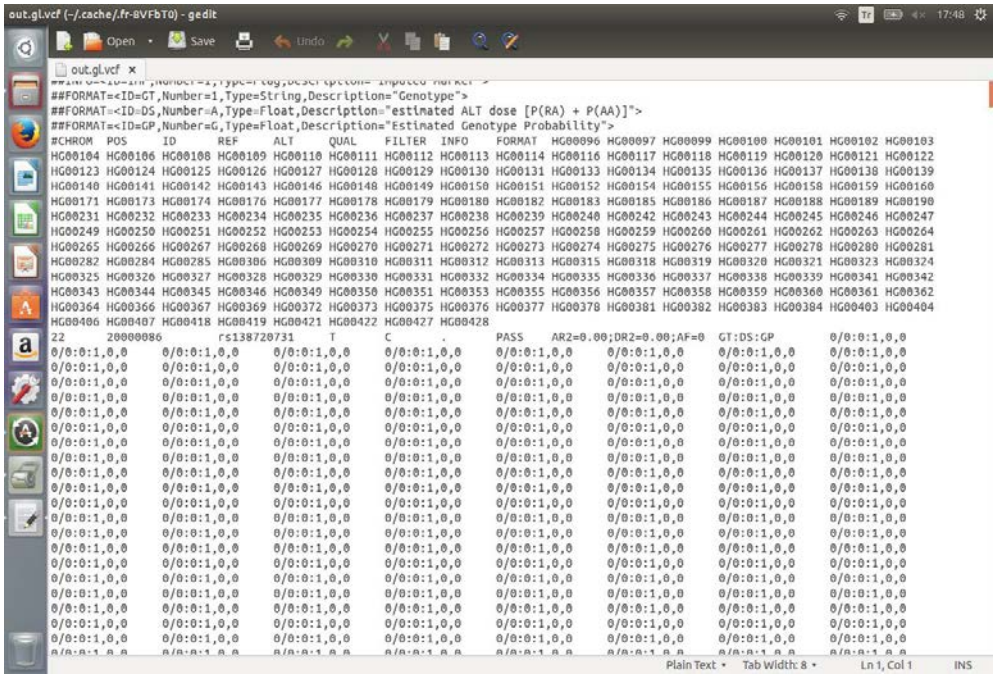
Şekil 10. Üçüncü senaryoya ilişkin GT çıktı dosyası.

Genotip Olabilirlik (GL) Komut Satırı Argümanına İlişkin Sonuçlar

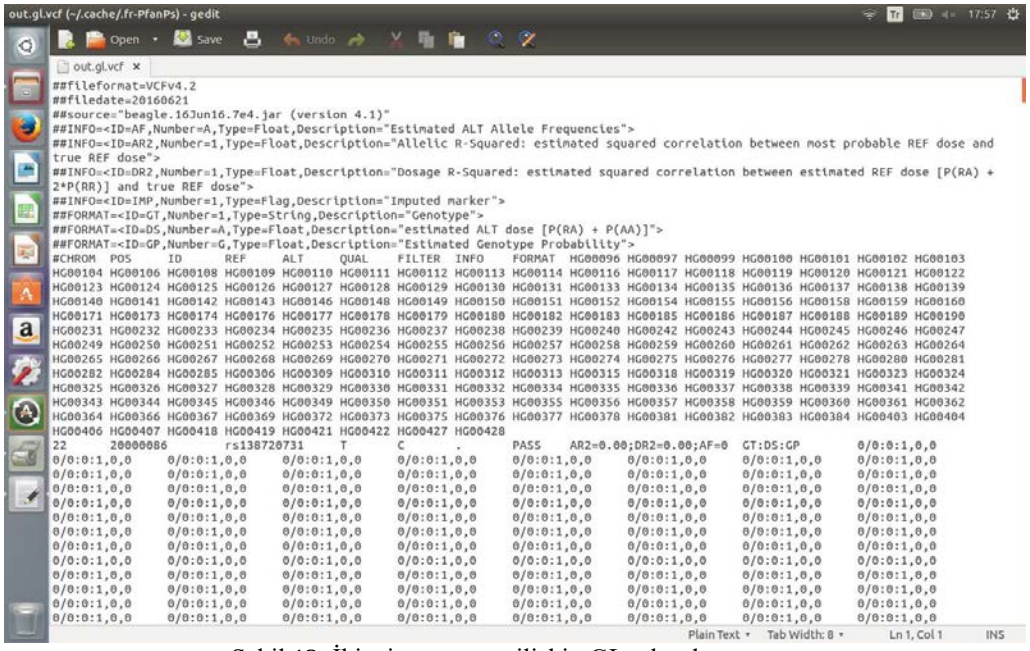
Genotip (GT) argümanı ile çalışılan veri setinden elde edilen genotip alelleri için daha sonra referans alel (REF) ve alternatif alel (ALT) alanlarında tanımlanan alel kümesindeki tüm olası genotiplerin, genotip olabilirlikleri (GL) hesaplanır. GL formatındaki, her belirtecin bir VCF çıktı dosyası bulunmaktadır. GT alanında herhangi bir verinin olmadığı durumda, GL alanı göz ardı edilir. GL argümanı aşağıdaki kod kullanılarak çalıştırılır.

- *java -jar Beagle.16Jun16.862.jar gl=test.16Jun16.862.vcf.gz out=out.gl*

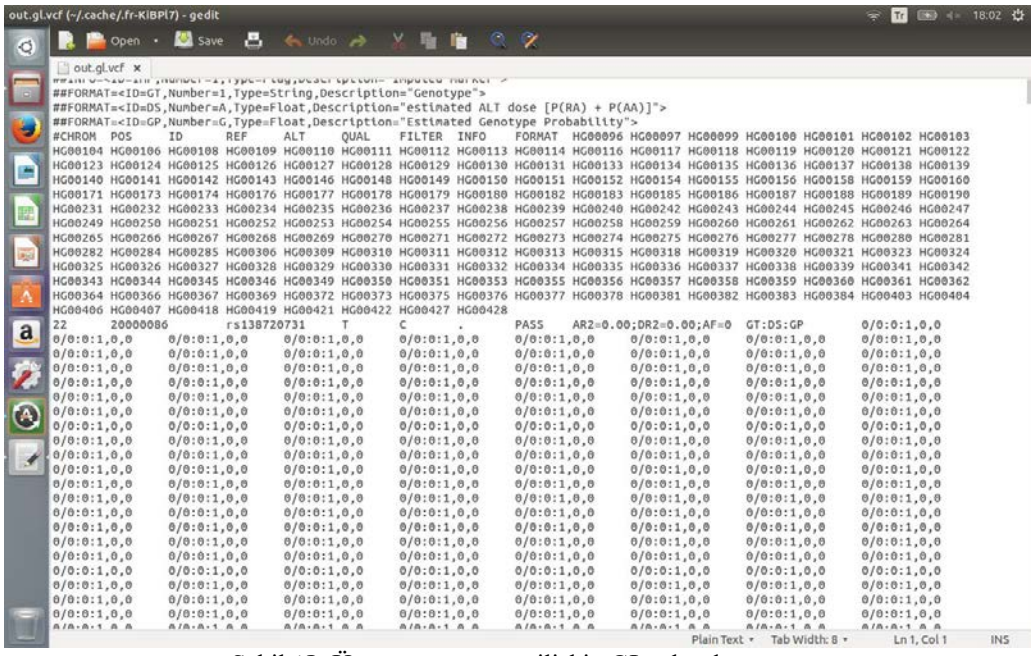
Genotip olabilirlik (GL) argümanı da GT argümana benzer şekilde iki çıktı elde edilmektedir. İlk çıktı, iterasyon basamakları, kodların çalışma süresi gibi bilgileri içeren dosya ve ikinci dosyada ise genotip atamanın doğruluğunun değerlendirilmesinde kullanılan alelik-R^2 (AR^2), doz-R^2 (DR^2) ve alel frekanslarına (AF) ilişkin çıktı elde edilmektedir. Genotip atamanın doğruluğuna ilişkin bilgilerin bulunduğu dosya GL'dir. Beagle'ın en önemli çıktısıdır. Beagle'ın en önemli çıktısıdır. Her üç senaryoda, rs138720731 numaralı SNP'den elde edilen GL çıktılarına ilişkin örnek görünümler Şekil 11-13'de verilmiştir.



Şekil 11. Birinci senaryoya ilişkin GL çıktı dosyası.



Şekil 12. İkinci senaryoya ilişkin GL çıktı dosyası.



Şekil 13. Üçüncü senaryoya ilişkin GL çıktı dosyası.

Her üç senaryodan elde edilen SNP'ler için genotip atama doğruluğunun belirlenmesinde kullanılan Alelik-R^2 (AR^2), Doz-R^2 (DR^2) ve alel frekanslarına (AF) ilişkin sonuçlar Çizelge 1'de verilmiştir.

Çizelge 1'de birinci senaryo için rs55902548 numaralı SNP için hesaplanan AR^2=0.97, DR^2=0.98 ve AF=0.14 değerlerine göre populasyon içindeki referans ve alternatif aleller içinden, en uygun alternatif alelin atamasının yapıldığı söylemek mümkündür. Benzer şekilde, rs138720731 numaralı SNP için hesaplanan AR^2=0, DR^2=0 ve AF=0 değerlerine göre populasyondaki referans alelle, atanan alellerin aynı olduğu göstermektedir. İkinci senaryo için rs187930998 numaralı SNP için hesaplanan AR^2=0.97, DR^2=0.97 ve AF=0.0026 değerlerine göre populasyon içindeki referans ve alternatif aleller içinden, hem referans allelerin hem de referans allerden farklı olarak en uygun alternatif alelin atamasının yapıldığı söylenebilir. Üçüncü senaryoda ise rs55902548 numaralı SNP için hesaplanan AR^2=0.96, DR^2=0.97 ve AF=0.14 değerlerine göre populasyon içindeki referans ve alternatif aleller içinden, hem referans allelerin hem de referans allerden farklı olarak en uygun alternatif alelin atamasının yapıldığı söylenebilir. Çizelge 1'de verinin farklı oranlarda kesilmesiyle oluşturulan üç senaryodan elde edilen atamalarda hem alelik-R^2 değerleri hem de doz-R^2 değerlerinden elde edilen atamaların

doğruluk derecesi >90%'ın üzerinde ve birbirine yakın değerler elde edilmiştir. Bununla beraber, referans veri setleri ya da hedef veri setlerinin büyüklükleri arasındaki farklılıklar, atama doğruluğu üzerine belirgin bir etkisi saptanamamıştır.

Çizelge 1. Her üç senaryodan elde edilen SNP'lerin tahmin doğruluğuna ilişkin sonuçlar

SNP'ler	Senaryolar[1]	Alelik-R^2	Doz-R^2	Alelik Frekans (AF)
rs138720731	S1(ref=153;hedef=38)	0	0	0
	S2(ref=95;hedef=96)	0	0	0
	S3(ref=57;hedef=134)	0	0	0
rs55902548	S1(ref=153;hedef=38)	0.97	0.98	0.14
	S2(ref=95;hedef=96)	0.96	0.97	0.14
	S3(ref=57;hedef=134)	0.96	0.97	0.14
rs187930998	S1(ref=153;hedef=38)	0.97	0.97	0.0026
	S2(ref=95;hedef=96)	0.97	0.97	0.0026
	S3(ref=57;hedef=134)	0.96	0.96	0.0025

[1]S1: Birinci senaryo; S2: İkinci senaryo; S3: Üçüncü senaryo; ref: referans veri seti; hedef: hedef veri seti

Beagle'ın GL komut argümanlarının çalıştırılmasıyla her üç senaryo için 15 iterasyonda ve 1Mb'lık bölgede, kodların çalışma süreleri ilişkin sonuçlar elde edilmiştir. Her üç senaryodan elde edilen çalışma süreleri Çizelge 2'de verilmiştir.

Çizelge 2. Üç senaryo için GL argümanından elde edilen çalışma süreleri

	Birinci senaryo	İkinci senaryo	Üçüncü senaryo
Çalışılan SNP sayısı	1356	1356	1356
Modelin kurulması için toplam süre	28 saniye	27 saniye	27 saniye
Örneklendirme için toplam süre	3 dakika 15 saniye	3 dakika 9 saniye	3 dakika 20 saniye
Kodun toplam çalışma süresi	3 dakika 47 saniye	3 dakika 39 saniye	3 dakika 50 saniye

Özellikle, genotip atama programlarının performans karşılaştırılmasında kullanılan bir başka yöntemde hesaplama süreleridir. Çizelge 2'de her üç senaryodan elde edilen sonuçlar incelendiğinde, modelin kurulması için programın kullanım süreleri aynı, örneklendirme süresi ve toplam çalışma süresi ise oldukça kısa olduğu belirlenmiştir.

Tartışma ve Sonuç

Bu çalışmada, referans veri setlerinin büyüklüğünün atama doğruluğuna etkisinin araştırılmasına yönelik olarak, oluşturulan üç farklı senaryodan elde edilen alelik-R^2 değerleri birbirine yakın ve atamalar >90% doğruluk derecesinde belirlenmiştir (Çizelge 1). Bu sonuçlarla paralel olarak Weng ve ark. (2013) tarafından yapılan çalışmada, farklı oranlarda oluşturulan (%20, %40, %80 ve %95) hedef veri seti senaryolarında, fastPHASE ve Beagle programlarından en doğru atamanın (>90%) Beagle'dan elde edildiğini bildirmişlerdir. Aynı zamanda, çalışmada referans populasyonun büyüklüğü arttıkça, Beagle ve fastPHASE'den daha doğru sonuçların elde edildiği belirtilmiştir. Martin ve ark. (2014) ile Browning ve Browning (2009) tarafından yapılan çalışmalarda, Beagle'da referans veri setinin büyüklüğü arttıkça daha büyük Alelik-R^2 değerlerinin elde edildiğini, dolayısıyla daha doğru atamaların yapıldığı bildirilmiştir. Bununla beraber, Beagle'ın genotip atama doğruluğunda, Alelik-R^2'nin oldukça güvenilir bir ölçü olduğu belirtilmiştir.

Genotip atamada kullanılan programlarda hesaplama zamanı ve iterasyon sayısı önemli bir kriterdir. Genotip atamada, referans veri setinde belirteçlerin sayısı ve referans örneklerin sayısı arttıkça hesaplama zamanı da artmaktadır (Browning ve Browning 2016). Çizelge 2'de 15 iterasyondan ve 1Mb uzunluğundaki bir bölgede Beagle, tüm senaryolar için hemen hemen aynı hesaplama zamanını kullanmıştır. Browning ve Browning (2016) tarafından yapılan çalışmada, hesaplama zamanlarının karşılaştırmak amacıyla 10Mb uzunluğundaki bir bölgede Minimac3, Impute2 ve Beagle (v4.1) karşılaştırmış ve 5 milyonluk bir referans veri seti için Beagle'ın daha kısa sürede sonuç verdiğini ifade etmişlerdir. Aynı zamanda Browning ve Browning (2009) tarafından yapılan çalışmada Beagle'ın HMM model kullanımıyla, diğer atama programlarına göre daha kısa sürede

sonuçların elde edildiğini ve küçük örnekler için iyi sonuçlar verirken, büyük örneklerde mükemmel sonuçlar verdiği bildirilmiştir.

Sonuç olarak, Beagle programı büyük ya da küçük örneklerde yüksek doğruluk derecesinde atamalar yapabilmektedir. Genotip atama yöntemleri arasında programın serbest olması ve kullanım kolaylığı açısından araştırıcıya birçok avantaj sağlamaktadır. Ayrıca programa ilişkin olarak ulusal literatürde bir çalışmaya rastlanmamıştır. Bu anlamda da çalışmanın, literatüre katkı sağlayacağı düşünülmektedir.

Teşekkür

Bu çalışma, ilk yazarın "Beagle genotip atama yönteminin değerlendirilmesi ve bir uygulama" isimli Yüksek lisans tezinden özetlenmiştir.

Kaynaklar

Browning BL (2011). Beagle 3.3.2**.** Department of Medicine Division of Medical Genetics University of Washington, Seattle, USA.

Browning BL, Browning SR (2009). A unified approach to genotype ımputation and haplotype-phase inference for large data sets of trios and unrelated individuals. The American Journal of Human Gen. 84: 201-223.

Browning BL, Browning SR (2016). Genotype imputation with millions of reference samples. The American Journal of Human Genetics. 98: 116–126.

Browning BL (2016). Beagle 4.1. https://faculty.washington.edu/browning/beagle/beagle.html#introduction. (run.beagle.16Jun16.d8b.example) (Erişim Tarihi: 16 Haziran, 2016.)

Browning SR (2008). Missing data imputation and haplotype phase inference for genome-wide association studies. Hum. Genet. 124 (5): 439–450.

Chud TCS, Ventura RV, Schenkel FS, Carvalheiro R, Buzanskas ME, Rosa JO, de Alvarenga Mudadu M, da Silva MVGB, Mokry FB, Marcondes CR, Luciana CA, Regitano LCA, Munari DP (2015). Strategies for genotype imputation in composite beef cattle. BMC Genetics. 16: 99.

Larmer S, Sargolzaei M, Ventura R, Schenkel F (2010). Imputation Accuracy from Low to High Density Using Within and Across Breed Reference Populations in Holstein, Guernsey and Ayrshire cattle. http://www.cdn.ca/Articles/GEBMAR2012/Imputation%20accuracy%20from%20low%20to%20high%20density%20-%20Larmer.pdf. (Erişim tarihi: 03.Ağustos, 2013).

Marchini J, Howie BN (2010). Genotype imputation for genome wide association studies. Nat. Rev. Genet. 11 (7): 499-511.

Martin AR, Tse G, Bustamante CD, Kenny EE (2014). Imputation-based assessment of next generation rare exome variant arrays. Pac. Symp. Biocomput. 241–252.

Morris DL, Ramsay PP (2010). https://immport.niaid.nih.gov/docs/standards/SNP_Imputation_Methods_Manual.pdf. (Erişim Tarihi: 14 Eylül, 2015).

Pausch H, Aiger B, Emmerling R, Edel C, Götz KU, Friesve R (2013). Imputation of high-density genotypes in the Fleckvieh cattle population. Genetics Selection Evolution. 45: 3.

Ramnarine SRS (2016). Genetic Imputation: Accuracy to Application. PhD Thesis. Washington University, St. Louis, Missouri, USA, pp:97.

Sun C, Wu XL, Weigel KA, Weigel KA, Rosa GJ, Bauck S, Woodward BW, Schnabel R D, Taylor JF, Gianola D (2012). An ensemble-based approach to ımputation of moderate-density genotypes for genomic selection with application to Angus cattle**.** Genet. Res. Camb. 94: 133-150.

Weng Z, Zhang Z, Zhang Q, Fu W, He S, Ding X (2013). Comparison of different imputation methods from low- to high-density panels using chinese Holstein cattle. Animal. 7(5): 729–735.