Biyoinformatik · **Uygulama notları**

Genomik Varyant Analizi

Dizileme Okumalarından SNP ve INDEL Saptama,
Anotasyon, Görselleştirme ve Yorumlamaya

Bu not, farklı organizmalarda referansa dayalı kısa varyant analizini; çalışma tasarımı ve ploididen okuma kanıtlarına, varyant saptamaya, filtrelemeye, anotasyona ve yorumlamaya uzanan bir bütün olarak açıklar. Ortak temel, biyolojik soruyu her hesaplama aşamasının varsayımlarıyla ilişkilendirir. İnsan germ hattı verisi, haploit izolatlar, model dışı organizmalar, poliploit ve havuzlanmış örnekler, uzun okumalar ve somatik varyasyon ayrı analiz yollarında ele alınır. Açıklamalı kayıtlar ve eşlik eden aşamalı betikler, kararları, yürütme koşullarını ve hata kaynaklarını örnekler; tamamlanmış bir gerçek veri yeniden üretimi veya doğrulanmış bir tüm genom analiz hattı oluşturmaz. İngilizce ve Türkçe sürümlerde aynı örnekler, bilimsel koşullar ve kaynakça korunur.

Eşlik eden betikler:
main_scripts/genomic_variants/

Tuna Birgün
github.com/tunabirgun

Orijinal Not: 22 Ağustos 2026 · Türkçe: 31 Ekim 2026

İçindekiler

1 Biyolojik sorudan varyant çalışmasına **2**
1.1 Yazılım seçmeden önce gözlemi tanımlayın . . . 2
1.2 Analizin sonradan oluşturamayacağı tasarım bilgisi . . . 2
1.3 İleri okuma . . . 3
2 Aleller, genotipler ve referans genom **3**
2.1 Ploidi, okumalara sorulan soruyu değiştirir . . . 3
2.2 Güvenin dört düzeyi . . . 3
3 Dosyaları bir kanıt zinciri olarak okuyun **4**
3.1 Sentetik bir hizalama . . . 4
3.2 VCF, alelleri ve örneklerdeki durumlarını tanımlar . . . 4
3.3 Koordinatlar açık dönüşüm gerektirir . . . 5
3.4 Referans güven dosyası ne ekler? . . . 5
4 Uyumlu bir referans ve örnek kaydı hazırlayın **5**
4.1 Örnek, kütüphane ve koşu kimliklerini ayırın . . . 6
4.2 Maliyetli aşamalardan önce inceleyin . . . 6
4.3 Referans seçimi bir analiz kararıdır . . . 6
5 Kısa okuma kanıtını yorumlanabilir hâle getirin **7**
5.1 Hizalama, genomik köken önerir . . . 7
5.2 Duplikeler, gözlemler arasındaki bağımlılığı gösterir . . . 7
5.3 Baz kalitesi yeniden kalibrasyonu koşulludur . . . 8
6 Alelleri keşfedin ve kohortu genotipleyin **8**
6.1 Sınırları belirli bir GATK örneği . . . 9
6.2 Ortak genotipleme neden ayrı bir işlemdir? . . . 9
6.3 Biyolojik anlam yüklemeden önce çağrıları inceleyin . . . 9
7 Filtreleme: Hangi kanıtın sonucu destekleyebileceğine karar vermek **10**
7.1 Üç filtreleme düzeyi . . . 10
7.2 Kuralları deneyden hareketle seçmek . . . 11
7.3 Güvenilir çağrı yapılabilen bölgeler paydayı belirler . . . 11
8 Normalizasyon: Karşılaştırılabilir varyant gösterimleri **11**
8.1 Referansa dayalı normalizasyon . . . 11
8.2 Çok alelli lokuslar ilişkiler taşır . . . 12
8.3 Haplotipler ve dizi etkilerinin yorumlanması . . . 12
9 Anotasyon: Dizi farkını biyolojik modelle ilişkilendirmek **12**
9.1 Uyumlu anotasyon kaynaklarını seçmek . . . 13
9.2 Örnek anotasyon komutları . . . 13
9.3 Bir alelin transkripte bağlı etkileri . . . 14
9.4 Kanıtı bağlamını kaybetmeden eklemek . . . 14
10 Doğrulama: İddiayı açıkça tanımlanmış bir alanda ölçmek **14**
10.1 Doğruluk kümesi varyantlarla birlikte bölgeleri de içerir . . . 14
10.2 Haplotipleri karşılaştırmak ve paydayı raporlamak . . . 15
10.3 Uyuşmazlıkları ve seçilmiş adayları araştırmak . . . 15
11 Özel yollar: Biyoloji değiştiğinde modeli değiştirmek **16**
11.1 İnsan germ hattı örnekleri . . . 16
11.2 Diğer diploit organizmalar: Kaynak koşullarını yeniden belirlemek . . . 16
11.3 Haploit izolatlar, mikroorganizmalar ve mantarlar . . . 16
11.4 Poliploitler ve havuzlanmış örnekler . . . 17
11.5 Uzun okumalar ve somatik varyasyon . . . 17
12 Çağrı kümesini ve lokus kanıtını görselleştirmek **18**
12.1 Soruyu yanıtlayan gösterimi seçmek . . . 18
12.2 Yerel IGV ile aleli incelemek . . . 18
12.3 Gösterilen özeti yeniden üretmek . . . 19
13 Yorumlama ve raporlama: Kanıt zincirini korumak **19**
13.1 Sentetik aday üzerinden örnek . . . 19
13.2 Tekrarlanabilirlik . . . 20
13.3 Gerçekte neyin doğrulandığını belirtmek . . . 20
14 Okumalardan değerlendirilmiş varyant kümesine aşamalı iş akışı **20**
14.1 Ortak hazırlık ve açık biyolojik dallanma . . . 21
14.2 Hizalayıcıları ve somatik çağırıcıları adil karşılaştırmak . . . 21
14.3 Yürütme kaydı olarak tekrarlanabilirlik . . . 22
14.4 Aynı protokolü yerel Galaxy kurulumuna eşlemek . . . 22
Kaynakça **23**

1 Biyolojik sorudan varyant çalışmasına

Genomik varyant çalışması, bir örneğin DNA dizisinin tanımlanmış bir referanstan nasıl ayrıldığını ve bu farkların ne anlama gelebileceğini araştırır. Dizileme, kromozomun tamamını doğrudan gözlemlemez. DNA moleküllerinden okumalar üretir; hizalama bu okumaların genomdaki olası kökenlerini belirler, varyant çağırıcı ise alternatif dizi durumlarını destekleyen kanıtı değerlendirir. Anotasyon, bildirilen aleli genlerle, transkriptlerle ve diğer özelliklerle ilişkilendirir. Bu işlemler farklı soruları yanıtladığı için analizin başında ortaya çıkan belirsizlikler son tablo yorumlanırken de önemini korur.

Bu not, genomik DNA'da tek nükleotit varyantları ile kısa ekleme ve silmelerin referansa dayalı keşfini ele alır. Açıklamaların ana ekseni, eşlenik uçlu kısa okumalarla yapılan germ hattı dizilemesidir. Aynı kavramlar başka organizma ve teknolojilerde de kullanılır; ancak insan için hazırlanmış diploit bir iş akışı tüm türlere uygulanabilecek bir reçete değildir. Sonraki bölümlerde haploit organizmalar, poliploitler, havuzlanmış örnekler, uzun okumalar, organeller ve somatik örnekler ayrı ele alınır. Yapısal varyantlar, kopya sayısı değişiklikleri, tekrar genişlemeleri ve montaja dayalı keşif ek yöntemler gerektirir; küçük varyantları içeren bir VCF, bunların bulunmadığını gösteremez.

1.1 Yazılım seçmeden önce gözlemi tanımlayın

Biyolojik birimden başlayın. Tek bir organizma, doku örneği, izolat ve birden fazla bireyin oluşturduğu havuz farklı örnekleme tasarımlarıdır. Tek bir organizmadan geldiği düşünülen DNA ekstraktı da organel DNA'sı, simbiyontlar veya kontaminantlar içerebilir. Örneğin neyi temsil etmesinin beklendiğini ve deneyin hangi varyasyon kaynaklarını birbirinden ayırabildiğini kaydedin. Örneğin havuzdaki düşük frekanslı bir alel, onu taşıyan bireyi belirlemez. Karma bir enfeksiyonda da tek bir genom içindeki heterozigotluğu temsil etmek zorunda değildir.

Araştırma sorusu, incelenecek bölgeyi de belirler. Tüm genom dizilemesi kodlayan ve kodlamayan dizileri örneklerken yakalama ve amplikon deneyleri seçilmiş bölgeleri özellikle zenginleştirir. Yakalama hedefi tasarımla belirlenmiş bölgedir; her bazda kullanılabilir kanıt bulunduğunun garantisi değildir. Kaç varyantın bulunmadığını bildirmeden önce hedefin dizilenmiş ve güvenilir çağrı yapılabilir bölümünü değerlendirin. Örnekler arası karşılaştırmalar, hem biyolojik birimin hem de incelenen bölgelerin uyumlu tanımlarını gerektirir.

Germ hattı varyant keşfi, çoğunlukla bireyin kalıtsal genomunda bulunan alelleri arar. Somatik varyant keşfi, döllenmeden sonra germ hattı dışındaki hücrelerde edinilen değişiklikleri arar. Bu değişiklikler, örneklenen klonun hücrelerinin bir bölümünde veya tamamında bulunabilir; saflık, kopya sayısı ve varsa eşlenmiş normal örnek hesaba katılmalıdır. Bu iki görev farklı kanıt modelleri kullanır. Resmî GATK germ hattı iş akışı, hazırlanmış hizalamalardan örnek başına referans güven dosyalarına ve ardından kohort genotiplemesine uzanan tutarlı bir yol sunar; bu yolun tanımlanmış olması her deney ve tür için doğrulandığı anlamına gelmez [6].

Temel parametreler

Veriyi indirmeden önce kısa bir analiz tanımı yazın: Biyolojik birim, deney ve okuma teknolojisi, referans montajı, bölgelere göre beklenen ploidi, germ hattı/somatik/karışım modeli, keşif bölgeleri, hedeflenen anotasyon ve kullanılabilecek bağımsız kontroller. Hangi seçimlerin varsayım olduğunu kaydedin. Genotip modelini değiştiren kararları çağrıları yorumlamadan önce çözün.

1.2 Analizin sonradan oluşturamayacağı tasarım bilgisi

Biyolojik tekrarlar, dizileme derinliği ve teknik tekrarlar farklı belirsizlikleri ele alır. Aynı kütüphaneden daha fazla okuma elde etmek, kapsanan lokuslardaki kanıtı güçlendirebilir; bağımsız birey sayısını artırmaz. Bir kütüphanenin yeniden dizilenmesi, koşuya özgü bir sorunu ortaya çıkarabilir; ekstraksiyon veya kütüphane hazırlama sırasında oluşan hatalar ise korunabilir. Bağımsız hazırlanmış bir kütüphane ya da farklı bir ölçüm ilkesine dayanan deney, kanıt zincirinin başka bir bölümünü sınar. Tüm tekrarlı ölçümleri eşdeğer saymak yerine kontrolleri desteklemeleri gereken iddiaya göre seçin.

Deney izin verdiğinde teknik partileri biyolojik gruplar arasında dengeleyin. Bir popülasyonun tüm üyeleri bir kütüphane protokolüyle, diğer popülasyonun üyeleri başka bir protokolle dizilenmişse varyant frekansı farkları teknik farklılıkları da yansıtabilir. Parti etiketlerini korumak incelemeyi mümkün kılar; tasarımda

eksik olan bilgiyi her zaman geri getirmez. Akrabalık ilişkileri, popülasyon yapısı ve ilişkili izolatlar da sonraki çıkarımları etkiler; çalışma üst verisinde korunmalıdır.

Bu nedenle yararlı bir çalışma çıktısı, farkların listesinden fazlasını içerir. Referans ve örnek tanımları, fiilen analiz edilen bölgeler, kanıt ve filtre durumları, yazılım ve kaynak sürümleri ile yorum sınırları da çıktının parçasıdır. Sonraki bölümler, ara dosyaların bu çıktıya nasıl katkıda bulunduğunu açıklar. Örneklerde kurmaca kontig ve örnek adları kullanılır. Bunlar biçimleri ve kararları göstermek içindir; gerçek bir veri kümesi, tamamlanmış bir iş akışı çalıştırması veya performans sonucu sunulmaz.

1.3 İleri okuma

Kavramsal temeli genişletmek için Brown'ın *Genomes 5* kitabı genetik ve genomik alanında bir ders kaynağı; Van der Auwera ve O'Connor'ın *Genomics in the Cloud* kitabı ise genomik veri analizine odaklanan hesaplamalı bir yardımcı kaynak sunar [1, 2]. Bu kitaplar kavramsal temel sağlar; yazılımın güncel davranışını tanımlayan teknik belgeler değildir. Buradaki sürüme bağlı komutlar resmî araç belgeleriyle kontrol edilmiştir.

2 Aleller, genotipler ve referans genom

Referans genom, koordinatları ve karşılaştırılacak diziyi sağlar. Referanstaki alellerin atasal, yaygın veya biyolojik açıdan üstün olduğunu söylemez. Bir örnek bir konumda referansla eşleşirken başka bir konumda alternatif alel taşıyabilir. Montajın değişmesi koordinatları, yerel dizi bağlamını ve varyantın temsilini değiştirebilir. Bu nedenle bir varyantı tanımlamak için konum ve iki harf yeterli değildir; montaj ve kontig bilgisi de gerekir.

Tek nükleotit varyantında (SNV) bir nükleotit diğerinin yerini alır. SNP terimi çağırma yazılımlarında sıklıkla işlemsel anlamda kullanılır; polimorfizm ise tarihsel olarak popülasyon içindeki varyasyonu ifade eder. Ekleme, referansa göre dizi ekler; silme, referansta bulunan diziyi çıkarır. Bunlar göreli tanımlardır: Bir referansa göre ekleme olarak adlandırılan alel, evrimsel olarak sonradan ortaya çıkmış olay olmayabilir. Burada kısa indel, seçilen küçük varyant yönteminin ele aldığı küçük ekleme/silme alellerini ifade eder; evrensel bir uzunluk sınırı koyulmaz.

2.1 Ploidi, okumalara sorulan soruyu değiştirir

Ploidi, bir lokustaki genotip modelinin temsil ettiği kromozom kopyası sayısıdır. Diploit ve iki alelli bir modelde fazlanmamış olasılıklar referans/referans, referans/alternatif ve alternatif/alternatiftir. Haploit modelde tek alel kopyası bulunur; tetraploit modelde alternatif alelin birden fazla dozajı olabilir. Tetraploit genotipi kodlayabilen bir çağırıcı, bu dozajları ayırt etmek için yine yeterli kanıta ve uygun modele ihtiyaç duyar. Ploidi parametresini ayarlamak, benzer homeolog kromozomlar arasındaki hizalama sorununu çözmez.

Ploidi aynı örneğin farklı bölgelerinde değişebilir. Eşey kromozomları, organeller, anöploit bölgeler ve kopya sayısı değişiklikleri, genom genelinde tek değer kullanma varsayımını bozabilir. Her modelin uygulanacağı biyolojik ve referansa özgü bölgeleri belirleyin. Bir poliploidin referansında birden fazla homeolog tek bölgede birleşmişse olağandışı derinlik ve görünürde heterozigotluk, birkaç genomik kopyanın aynı konuma hizalanmasını yansıtabilir. Bu gözlemler doğrudan biyolojik yoruma değil, montajın ve hizalamanın incelenmesine gerekçe oluşturur.

VCF genotip indisleri alel sırasını gösterir. 0 referans aleli, 1 ilk alternatif aleli, 2 ise ikinci alternatif aleli belirtir. Bu yüzden diploit `0/1` ile haploit `1` farklı modelleri ifade eder. Eğik çizgi fazlanmamış alelleri, dikey çizgi fazlanmış alelleri ayırır. Faz, alellerin belirtilen fazlama bağlamında kromozom kopyalarına atanmış olduğunu gösterir; yalnızca varyantlar birbirine yakın diye varsayılmamalıdır [3].

2.2 Güvenin dört düzeyi

Phred dönüşümü bir olasılığı logaritmik ölçekte ifade eder:

$$Q = -10 \log_{10}(p).$$

Hata olasılığı 10^{-3} ise karşılık gelen puan 30'dur. Bu bir matematik örneğidir; filtreleme önerisi değildir. p ile hangi olayın ifade edildiği önemlidir. Baz kalitesi bir nükleotit gözlemiyle, hizalama kalitesi okumanın

yerleştirilmesiyle, genotip kalitesi model altında yapılan genotip atamasıyla, lokus kalitesi ise varyant kaydıyla ilişkili kanıtla ilgilidir. Aynı sayısal ölçeği kullanmaları, bu farklı düzeylerdeki puanların birbirinin yerine geçebileceği anlamına gelmez [4, 3].

Kohorttaki en az bir örnekte alternatif aleli güçlü kanıtla desteklenen bir lokus düşünün. Derinliği düşük bir örneğin o lokustaki genotipi yine de belirsiz olabilir. Başka örneklerin alternatif alel taşıdığı bir konumda, bir örnek güvenle referans olarak da çağrılabilir. Bu nedenle lokusun kabulü ile genotipin kabulü ayrı kararlardır. Kohort düzeyindeki kalite değeri, her örneğin genotipini güvenilir hâle getirmez.

Dikkat edilmesi gerekenler

Eksik genotip, referans genotip değildir. Diploit gösterimde `./.`, genotipin verilmediğini; `0/0`, referans/referans çağrısını belirtir. Eksik değerleri referans çağrılarıyla değiştirmek alel sayılarını, eksiklik oranlarını ve sonraki biyolojik sonuçları değiştirir. Yalnızca varyant içeren bir VCF'de kaydın bulunmaması da o konumda yeterli referans kanıtı olduğunu göstermez.

Okuma desteği, oy sayımı değil kanıttır. Güvenilir biçimde yerleştirilmiş farklı moleküllerden gelen on gözlem ile tek bir çoğaltılmış parçadan gelen on duplike okuma aynı değildir. Zincir yönü, okuma içindeki konum, yerel dizi bağlamı ve yakın eşleşmezlikler görünürdeki bir aleli açıklayabilir. Çağırıcı bu bilginin bir bölümünü kendi modeline göre birleştirir; kalite kontrolü ve sonraki inceleme, modelin örneği hedeflenen çıkarım için yeterince iyi temsil edip etmediğini değerlendirir.

3 Dosyaları bir kanıt zinciri olarak okuyun

FASTQ, okuma dizilerini ve baz kalitelerini saklar. Yaygın dört satırlı kayıtta tanımlayıcı `@` ile başlar; ikinci satır diziyi içerir, üçüncü satır `+` ile başlar, dördüncü satır baz başına kaliteyi kodlar. Dizi ve kalite dizgelerinin uzunlukları eşit olmalıdır. Eşlenik uçlu dosyalar, DNA parçalarının dizilenen iki ucunu tanımlar; filtreleme boyunca eşleşmeleri korunmalıdır. Okuma adı biyolojik tekrarı değil, bir gözlemi tanımlar.

SAM metin tabanlı bir hizalama biçimidir; BAM aynı hizalama bilgisini ikili biçimde saklar. CRAM, referans ve çözümleme gereklilikleri ayrıca değerlendirilmesi gereken başka bir hizalama biçimidir. Hizalama, bayraklar, hizalama kalitesi ve CIGAR adı verilen işlem dizgesi aracılığıyla okumanın nereye ve nasıl yerleştirildiğini açıklar. Sıralanmış ve indekslenmiş BAM, bölgesel erişimi destekler; ancak sıralama ve indeksleme doğru hizalamayı ya da örnek kimliğinin korunduğunu kanıtlamaz [4, 5].

3.1 Sentetik bir hizalama

Aşağıdaki kayıt şematiktir; gerçek SAM'deki sekmeler okunabilirlik için boşluklarla gösterilmiştir. `toy1` kontiginin 101–106 konumlarındaki bazların `ACGTAC` olduğunu varsayın.

Örnek kayıtlar

```
r001  0  toy1  101  60  6M  *  0  0  ACGTTC  IIIIII
```

Okuma, referansın bir tabanlı 101 konumunda başlar. Hizalanan altı baz 101–106 konumlarını kapsar; okumanın beşinci bazı T, varsayılan referansın 105 konumundaki A bazından farklıdır. `6M` CIGAR ifadesi altı hizalanmış eşleşme veya eşleşmezlik anlamına gelir; altı nükleotidin özdeş olduğunun doğrulandığını göstermez. Bu örnekte eşleşmezlik, `M` işleminin içinde yer alabilir. Kalite dizgesindeki `I` harfleri kodlanmış kalitelerdir; ekleme işlemi değildir. Böyle tek bir gözlem, örneğin genotipini belirlemeye yetmez.

3.2 VCF, alelleri ve örneklerdeki durumlarını tanımlar

VCF başlığı dosya sürümünü, kontigleri ve anotasyonların anlamlarını tanımlar. Her veri kaydı kromozom, konum, tanımlayıcı, referans ve alternatif alel, kalite, filtre durumu ve bilgi alanlarını içerir. Örnek sütunları FORMAT alanına göre yorumlanır. BCF, varyant verisinin ikili temsilidir. Başlık bilimsel tanımın bir parçasıdır: Aynı görünen etiketlerin dosyalar birleştirilmeden önce uyumlu anlamlara sahip olması gerekir [3].

Örnek kayıtlar

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT sample_A
 toy1 105 .  A   T   60   PASS   .    GT:DP:GQ 0/1:18:35
 toy1 110 .  AT  A   45   PASS   .    GT:DP:GQ ./.:3:.
```

Bu kurmaca satırlar, okunabilir bir alıntıdır; tanım satırları ve sekme ayraçları çıkarıldığı için tam ve geçerli bir VCF değildir. İlk satır, 105 konumunda A'dan T'ye varyasyon önerir ve `sample_A` için diploit heterozigot genotip verir. İkinci satır, 110 konumundaki çapa A bazından sonra gelen T'nin silinmesini temsil eder; ancak bu örnek için genotip vermez. `PASS` değeri, kaydın tanımlanmış lokus filtreleme şemasıyla ilgilidir. Eksik genotipi geçersiz kılamaz. Derinlik ve kalite sayıları önerilen eşikler değil, alanları açıklayan örnek değerlerdir.

3.3 Koordinatlar açık dönüşüm gerektirir

SAM metnindeki konumlar ve VCF konumları bir tabanlıdır. BED aralıklarında yaygın olarak sıfır tabanlı başlangıç ve dışlanan bitiş kullanılır. Bu nedenle referansın 101'den 106'ya kadar olan bazları BED'de `toy1 100 106` olarak gösterilir. Uzunlukları $106 - 100 = 6$ bazdır. VCF'nin 105 konumundaki tek baz, BED'de `toy1 104 105` aralığına karşılık gelir. Komut satırındaki bölge sözdizimi ile ikili dosyanın iç koordinatları aynı olmak zorunda değildir; aralığı aktarmadan önce ilgili aracın arayüzünü kontrol edin [4].

Yukarıdaki çapalı silmede VCF referans aleli 110–111 konumlarını kapsar; BED karşılığı `toy1 109 111` olur. Yalnızca silinen T ise `toy1 110 111` aralığındadır. Bu aralıklar farklı soruları yanıtlar. Varyantlarla genomik özelliklerin kesişimi hesaplanırken çapa, değişen bazlar veya aracın tanımladığı varyant kapsamı kullanılabilir. Özellikle ekzon sınırlarındaki ekleme ve silmeler için hangi tanımın kullanıldığını belirtin. Şüpheli bir koordinat sorununu, bilinen sentetik bir örnekle sınamadan tüm konumları kaydırarak çözmeye çalışmayın.

3.4 Referans güven dosyası ne ekler?

Yalnızca varyant içeren VCF, ağırlıklı olarak bildirilen varyant lokuslarını kaydeder. gVCF ise incelenen bölgelerdeki referans ve gözlenmemiş alternatif durumların kanıtına ilişkin bilgiyi de taşır. GATK yolunda benzer referans güvenine sahip bölümler bloklar hâlinde sıkıştırılır; `<NON_REF>` gibi sembolik aleller modelde yer alır. Bunlar gözlenmiş sıradan eklemeler veya gerçek DNA dizileri değildir. gVCF, uyumlu bir sonraki genotipleme aşamasının ara çıktısıdır; yalnızca daha büyük bir son VCF değildir [7].

Referans güven bilgisi, başka bir örneğin alternatif alel sunduğu konumda referansı destekleyen kanıtı olan örnek ile kullanılabilir kanıtı olmayan örneği ayırmaya yardımcı olur. Yorum yine kapsama, çağırıcının kurallarına ve işlenen aralıklara bağlıdır. Dosya uzantısı, farklı çağırıcılar arasında uyumluluğu tek başına göstermez. Üreten yöntemi ve sürümü koruyun; seçilen ortak genotipleme aracının söz konusu gVCF modelini kabul ettiğini doğrulayın.

4 Uyumlu bir referans ve örnek kaydı hazırlayın

Referans paketi, aynı koordinat sistemini tanımlaması gereken birden fazla kaynağı bir araya getirir. FASTA diziyi sağlar; indeksi diziye erişimi mümkün kılar; dizi sözlüğü ise gerektiren yazılımlar için kontigleri tanımlar. Gen anotasyonları, hedef aralıklar ve bilinen varyant kaynakları farklı bilgiler ekler. Aynı tür adını taşıyan dosyalar; montaj sürümü, kontig adlandırması, organel dizileri, alternatif iskeleler veya anotasyon sürümü bakımından ayrılabilir. Uyumluluk, içerik ve kaynak bilgisi üzerinden gösterilmelidir.

Çalışmaya ve mevcut kaynaklara uygun montajı seçerek başlayın. Sağlayıcıyı, erişim numarasını veya sürümünü, içerdiği dizi kümesini ve yapılan değişiklikleri kaydedin. Alternatif diziler içeren kromozom düzeyindeki montaj, yalnızca birincil dizileri içeren alt kümeyle aynı değildir. Dizi eklemek veya çıkarmak, hizalayıcının değerlendirebildiği seçenekleri değiştirir. Kullanılan FASTA'nın tam kopyasına ait bütünlük kaydını yerel olarak saklayın ve indeksleri bu dosyadan üretin. Alttaki dizi değişmişse yalnızca dosya adı eşleşiyor diye bir indeksi yeniden kullanmak güvenli değildir.

Kaynak	Uyumluluk sorusu
Referans FASTA	Koordinatları hangi montaj ve dizi alt kümesi tanımlıyor?
Gen anotasyonu	Kontig adları, koordinatlar ve sürümün kaynak bilgisi bu montajla eşleşiyor mu?
Hedef aralıklar	Hangi koordinat kuralı ve deney tasarımıyla üretildiler?
Bilinen lokuslar	Aleller ve konumlar bu referansa göre geçerli mi?
Örnek çizelgesi	Tanımlayıcılar bireyleri, kütüphaneleri ve dizileme birimlerini ayırıyor mu?

Kontig adını değiştirmek sınırlı bir işlemdir. `1` adını `chr1` yapmak, ancak diziler ve koordinat sistemleri zaten uyuşuyorsa adları uzlaştırabilir. Bir montajı diğerine dönüştüremez. Montajlar arasında koordinat dönüşümü, uygun bir eşleme ve dönüştürülen alellerin doğrulanmasını gerektirir; bazı lokuslar eşlenemez veya farklı biçimde temsil edilir. Yeni bir analizde, doğrudan seçilen montaj için hazırlanmış kaynakları edinmek çoğu zaman önlenebilir belirsizlikleri azaltır.

4.1 Örnek, kütüphane ve koşu kimliklerini ayırın

Örnek çizelgesi; biyolojik örnek, ekstraksiyon, kütüphane ve dizileme birimi arasındaki ilişkiyi korumalıdır. Aynı kütüphanenin farklı dizileme şeritlerinden gelen okumalar aynı biyolojik örneğe katkı sağlar. Aynı bireyin farklı kütüphaneleri ise okumaları daha sonra birlikte analiz edilse bile ayrı kütüphane etiketleri gerektirebilir. Okuma grupları, hizalama dosyalarında örnek ve kütüphane bilgisi dâhil tanımlayıcıları taşır. Bu alanlar sonraki araçların gruplama davranışını etkiler; yalnızca görsel etiket olarak ele alınamaz [4].

Hizalamaları birleştirmeden önce tüm girdilerin beklenen örneğe ait olduğunu, aynı referansı kullandığını ve doğru okuma grubu ilişkilerini koruduğunu doğrulayın. Teknik olarak başarılı bir birleştirme, yanlışlıkla iki bireyin verisini bir araya getirebilir. Kayıtlı kimliği, organizmaya uygun deneysel veya genetik kontrollerle karşılaştırın. Üst veri herkese açık paylaşılamıyorsa hassas bilgileri okuma adlarına ya da dışa aktarılan dosya adlarına koymak yerine denetimli bir yerel eşleme tutun.

İş akışı kaydı, her sonucun girdilerini, referans kaynaklarını, yazılım ortamını ve parametre seçimlerini birbirine bağlamalıdır. Okuma kırpma, duplike politikası, aralık kısıtlaması, ploidi ve filtreleme gibi kanıtı etkileyen değişiklikleri kaydedin. İlk alınan veriyi türetilmiş dosyalardan ayrı tutun. Böylece değişen bir çağrı kümesinin yeni okumalardan mı, farklı modelden mi yoksa farklı referanstan mı kaynaklandığı açıklanabilir; her dosya yeni ve bağımsız bir sonuç sayılmaz.

4.2 Maliyetli aşamalardan önce inceleyin

Küçük yapısal kontroller, hizalama veya çağırma ciddi kaynak tüketmeden önce uyumsuz girdileri yakalayabilir. Beklenen dosya adlarını ve türlerini doğrulayın; başlıkları inceleyin, örnek adlarını kontrol edin ve kontig sözlüklerini karşılaştırın. Sıkıştırılmış varyant dosyalarının sonraki aracın kabul ettiği sıkıştırma ve indeksi kullandığını kontrol edin. Dosyanın boş olmaması ve indekslemenin başarılı olması yalnızca sınırlı özellikleri gösterir; biyolojik geçerliliği kanıtlamaz.

Komut satırı

```
# Illustrative checks on an existing prepared alignment.
samtools quickcheck sample.ready.bam
samtools view -H sample.ready.bam
samtools flagstat sample.ready.bam
```

Bu komutlar resmî SAMtools arayüzünü izler [5]. `quickcheck`, her hizalamayı bozulma açısından okumak yerine başlık ve dosya sonu özelliklerini kontrol eder. Başlık incelemesi referans ve okuma grubu tanımlarını açığa çıkarır; `flagstat` hizalama bayraklarını özetler. Çıktıları, beklenen örnekler ve deneyin özellikleriyle birlikte yorumlayın. Bu notta komutlar belgelerle karşılaştırılmıştır; bir çalışma veri kümesinde çalıştırılmamıştır.

4.3 Referans seçimi bir analiz kararıdır

Daha eksiksiz veya çeşitliliği daha iyi temsil eden referans, eski doğrusal montajın yetersiz temsil ettiği dizileri açığa çıkarabilir. Taslak insan pangenomu çalışması, değerlendirdiği graf tabanlı iş akışlarında varyant saptamanın değiştiğini göstermiştir; doğrusal hattın yalnızca FASTA dosyasını değiştirmek bu

yöntemi yeniden üretmez [37]. Yeni montaj; koordinatları, hizalamaları ve kaynak uyumunu değiştirir. İndeksleri yeniden oluşturun, eşleşen doğruluk ve anotasyon kaynaklarını kullanın; montajları karşılaştırıyorsanız koordinat dönüşümünü denetleyip kaydedin. Olson ve arkadaşlarının derlemesi bu seçimleri eksiksiz insan genomları ve karşılaştırmalı değerlendirme bağlamında ele alır [36]. Derleme önerileri ve insanda ölçülen performans, model dışı bir organizmanın doğruluğunu kanıtlamaz.

5 Kısa okuma kanıtını yorumlanabilir hâle getirin

Ön işleme, okumaları belirli bir çağırma modeline hazırlar. Amaç her kalite grafiğini ideal bir örneğe benzetmek değil, teknik sorunları belirlemek ve kanıtı tutarlı biçimde temsil etmektir. Okuma kalite profillerini, adaptör dizilerini, baz bileşimini, okuma uzunluklarını ve fazla temsil edilen dizileri deney bağlamında inceleyin. Yanlı baz bileşimi organizmadan veya kütüphane tasarımından kaynaklanabilir; kalite uyarısı verinin atılması gerektiği anlamına değil, incelenmesi gereken bir gözleme karşılık gelir.

İstenmeyen adaptör dizileri veya düşük kaliteli uçlar analizi etkiliyorsa kırpma yararlıdır; ancak hizalayıcıya verilen okumaları değiştirir. Eşleşmeleri koruyun, kalan okuma uzunluklarını ve sayılarını kaydedin. Aşırı kırpma gerçek ve bilgi taşıyan diziyi çıkarabilir, benzersiz yerleşimi azaltabilir. Kırpma kuralını kütüphane ve çağırma yolu için seçin, ardından sonuçlarını inceleyin. Mevcut diziyi ve hedeflenen analizi değerlendirmeden evrensel bir asgari uzunluk ya da kalite eşiği benimsemeyin.

İsteğe bağlı eşlikçi betik fastp 1.4.0 kullanır. `--qualified_quality_phred 15`, okuma düzeyindeki filtre için yeterli baz kalitesini tanımlar; Q15 altındaki her bazı kırpmaz veya ortalama kalitenin 15 olmasını istemez. Varsayılan `--unqualified_percent_limit 40`, yetersiz kaliteli baz oranının en fazla %40 olmasına izin verir. Kurmaca 100 bazlık bir okumada 30 bazın kalitesinin Q15'in altında olması bu kuralı ihlal etmez. Kalitesi Q15'in altında olan baz sayısı 41'e çıkarsa okuma elenir. Diğer filtreler her iki okumayı da reddedebilir. `--length_required 15`, işlemeden sonra 15 bazdan kısa kalan okumaları eler. Adaptör kırpma açıktır; tanınan NextSeq/NovaSeq verilerinde poli-G uç kırpma kendiliğinden etkinleşir. Bu betik, kayan pencereyle kalite kırpmayı veya poli-X kırpmayı açmaz [42].

Yeterli baz kalitesi eşiğini yükseltmek, okumayı kısaltmadan yetersiz sayılan bazların oranını artırabilir. Kaliteye göre kırpma ise dizi çıkarır ve sonrasında uzunluk filtresini devreye sokabilir. fastp raporlarında korunan çiftleri, elenme nedenlerini ve uzunluk dağılımlarını; ardından hizalanmayı ve güvenilir çağrı yapılabilen kapsamı karşılaştırın. Daha katı ayarlar son çağrıları her zaman iyileştirmez.

5.1 Hizalama, genomik köken önerir

Hizalayıcı, her okuma için seçilen referansta olası yerleşimler arar. Eşlenik uç bilgisi, parça uçları arasındaki ilişki yoluyla yerleşimleri kısıtlar. Tekrarlar ve çok benzer genler birden fazla kökeni destekleyebilir; referanstan uzaklaşma ve dizileme hataları ek belirsizlik getirir. Hizalama kalite puanı, hizalayıcının kendi modeli altında yerleşime duyduğu güveni özetler. Seçilen referansın doğru genomik kopyayı içerdiğini kanıtlamaz.

Hizalamadan sonra deneyin gerektirdiği ölçüde hizalanma oranlarını, kapsama dağılımını, parça uzunluğu davranışını, duplike ölçütlerini ve uyumsuz veya tamamlayıcı hizalamaları inceleyin. Tek toplu değere dayanmak yerine örnekleri ve dizileme birimlerini karşılaştırın. Ortalama kapsamı yeterli bir kütüphanede bile GC bakımından zengin bölgeler, tekrarlı lokuslar veya yakalama hedefleri yetersiz temsil edilebilir. Eksik çağrılar yorumlanırken hedeflenen analiz bölgesindeki kapsama, genom genelindeki tek ortalamadan daha açıklayıcıdır.

Koordinata göre sıralama, hizalamaları bölge temelli işlemeye hazırlar; indeksleme verimli erişim sağlar. Okuma gruplarını koruyun ve beklenen kontigleri kontrol edin. Her ikisi de BAM üretiyor diye genomik DNA için RNA-seq hizalama reçetesini kullanmayın. Ekzon birleşimlerine duyarlı hizalama ve transkript kaynaklı kapsama, genomik varyant keşfinden farklı varsayımlara dayanır. Uzun okuma çağırıcısı da burada açıklanan kısa okuma yolundan farklı hizalayıcı ve platforma özgü model gerektirebilir.

5.2 Duplikeler, gözlemler arasındaki bağımlılığı gösterir

Kütüphane çoğaltımı, tek bir özgün parçanın kopyalarından birden fazla okuma üretebilir. Bu kopyaları bağımsız moleküller saymak bir alelin desteğini abartabilir. Duplike işaretleme, sonraki yazılımın belgelenmiş politikasını uygulayabilmesi için bayrak atar. Aynı diziye sahip tüm okumaları silmekten farklıdır.

Doğal olarak özdeş parçalar, optik duplikeler ve çoğaltılmış kopyalar; kütüphane ve kullanılan yöntem bağlamında yorumlanmalıdır [4, 7].

Sıradan koordinat temelli duplike politikasını amplikon deneyine doğrudan uygulamayın: Tasarım, uçları aynı olan çok sayıda molekülü bilerek üretir. Benzersiz moleküler tanımlayıcı içeren kütüphaneler, bu tanımlayıcıları ve hata modelini yorumlayan yöntem gerektirir. Duplike ölçütleri girdi miktarı, dizileme derinliği ve kütüphane karmaşıklığına göre de değişebilir. Yüksek duplike oranı inceleme gerektirir; tek başına nedeni belirlemez veya evrensel ret eşiği oluşturmaz.

İşlem sırası önemlidir. Duplike aracı, koordinata göre sıralanmış çıktı üretmeden önce eş bilgisi, belirli bir sıralama düzeni veya kütüphane etiketi gerektirebilir. Seçilen uygulamanın tüm hazırlık gerekliliklerini izleyin. Dizinin içinden alınmış kısa bir komut çalışabilir; yine de amaçlanandan farklı okumaları işaretleyebilir. Eşlik eden hazırlık betikleri, ara sıralama düzenleri dâhil bu işlem sırasını açıkça gösterir.

5.3 Baz kalitesi yeniden kalibrasyonu koşulludur

Baz kalitesi yeniden kalibrasyonu (BQSR), bildirilen baz kaliteleri ile ampirik eşleşmezlik örüntüleri arasındaki sistematik farkları tahmin eder. Nükleotit dizisini onarmak yerine güven değerlerini ayarlar. GATK yaklaşımı, dizi bağlamı ve okuma döngüsü gibi ortak değişkenleri modeller [10]. Gerçek biyolojik farkların otomatik olarak dizileme hatası sayılmaması için bilinen varyant lokusları maskelenir [8].

Bu gereklilik, kaynakları zengin ve sınırlı organizmalar arasında pratik bir ayrım oluşturur. Bilinen lokus kümesi referansla eşleşmeli ve maskeleme amacı için yeterince güvenilir olmalıdır; indirilebiliyor olması yeterli değildir. Başlangıç kaynaklarının yinelemeli oluşturulduğu bir yolda, ön çağrı kümesinden aday maskeleme lokusları türetilebilir; ancak aynı okumalara ve modele bağımlılık sürer. Yineleme, son çağrıların doğruluğuna ilişkin bağımsız doğrulama değildir. Başlangıç lokuslarının nasıl elde edildiğini kaydedin ve mümkün olduğunda süreci ayrı kanıtla değerlendirin.

BQSR, her dizileme platformu veya çağırıcı için zorunlu bir işlem değildir. Seçilen iş akışının bunu bekleyip beklemediğini, verinin kalite modelinin uygunluğunu ve gerekli kaynakların bulunup bulunmadığını belirleyin. Ek bir ön işleme aşamasının mutlaka iyileşme sağlayacağını varsaymak yerine tanısal davranışı karşılaştırın. BQSR ayrıca, iş akışının ilerleyen kısmında çağrılmış varyantların anotasyonlarını modelleyen varyant kalite puanı yeniden kalibrasyonundan farklıdır. Benzer adlar farklı nesnelere ve kararlara karşılık gelir.

Dikkat edilmesi gerekenler

Ön işleme çıktısı, yalnızca tanımlanmış bir çağırıcının gerekliliklerine göre analize hazır hizalamadır. Sıralama, duplike işaretleme ve yeniden kalibrasyon; örneğin kontaminasyonsuz olduğunu, ploidinin doğru seçildiğini veya her lokusta çağrı yapılabileceğini göstermez. Genotip yorumuna geçmeden önce çözülmemiş örnek ve referans sorunlarını kayıtta koruyun.

6 Alelleri keşfedin ve kohortu genotipleyin

Varyant çağırma, okuma kanıtının hangi dizi durumuyla daha iyi açıklandığını değerlendirir. Basit bir kavramsal modelde çağırıcı, aday genotip G altında gözlenen okumaların D olabilirliğini, $P(D \mid G)$, hesaplar. Genotipin sonsal olasılığı, önsel varsayımlara da bağlıdır. Bu ayrım, aynı okuma sayılarının farklı model, ploidi veya parametreler altında neden aynı çağrıları üretmeyebileceğini açıklar. Genotip yorumlanırken model ve varsayımları da dikkate alınmalıdır.

Çağırıcı hangi aday alelleri değerlendireceğine de karar vermelidir. Bir indelin yakınında, farklı okuma hizalamaları aynı yerel haplotipi farklı biçimde tanımlayabilir. Montaja dayalı yöntemler, her eşleşmezliği bağımsız ele almak yerine aday yerel dizileri yeniden kurabilir ve okuma kanıtını bu dizilerle karşılaştırabilir. HaplotypeCaller, yerel montaj kullanan bir çağırıcıdır; FreeBayes başka bir haplotip temelli çerçeve sunar. VCF üreten tüm araçları eşdeğer saymak yerine, destekledikleri varsayımlar ve çıktılar ayrı değerlendirilmelidir [7, 14].

6.1 Sınırları belirli bir GATK örneği

Aşağıdaki sözdizimi, tek örnek için referans güven çağrısını gösterir. Uyumlu, indekslenmiş ve sözlüğü bulunan referans; geçerli örnek/okuma grubu bilgilerini taşıyan, hazırlanmış, koordinata göre sıralanmış ve indekslenmiş BAM; ayrıca gerçekten diploit bir analiz bölgesi varsayılır. Dosya adları kurmacadır. Sağlanmış verileri tanımlamaz veya doğrulanmış bir çalıştırma ortamı oluşturmaz.

Komut satırı

```
gatk HaplotypeCaller \
  -R reference.fa \
  -I sample.ready.bam \
  -O sample.g.vcf.gz \
  -ERC GVCF \
  --sample-ploidy 2
```

Çıktı, uyumlu bir genotipleme aşaması için örnek başına gVCF'dir. 2 değeri, öğretim örneğindeki diploit modeli açıkça belirtir; türler arası genel öneri değildir. HaplotypeCaller diploit olmayan durumları destekler; ancak olası genotip sayısı ve hesaplama yükü, ploidi ve alel sayısıyla artabilir. Güncel belgelerde bölgeye bağlı ploidi seçeneği de bulunur; etkin bölge sınırlarındaki davranışı dikkat gerektirir. Bu tür seçenekleri benimsemeden önce kurulu sürümü doğrulayın [7].

Bu kısa örnek; çalışmaya özgü aralıkları, kaynak sınırlarını ve kütüphane hazırlığına bağlı seçenekleri içermez. Ham FASTQ'dan başlayan çalıştırılabilir bir iş akışı değildir. PCR'siz hazırlandığı belgelenmiş kütüphane için GATK 4.6.2.0, `--pcr-indel-model NONE` ayarını belirtir. Bu ayar, okumaları kırpmak veya duplikeleri kaldırmak yerine PCR'ye özgü indel hata modelini kapatır. Varsayılan `CONSERVATIVE` değerini yalnızca daha fazla indel elde etmek için değiştirmeyin. Gerekçeli bir değişiklikten sonra indel kanıtını inceleyin; seçilen çağırıcının yeniden kalibre edilmiş kalite bekleyip beklemediğini belirleyin. Kurulu yardım metnini inceleyin ve kullanılan parametreleri koruyun. Sözdiziminin başarılı olması, biyolojik modelin örneğe uygunluğunu göstermez.

6.2 Ortak genotipleme neden ayrı bir işlemdir?

Bağımsız ve yalnızca varyant içeren VCF'ler, tüm lokuslarda eşdeğer kanıt taşımaz. A örneği bir aleli bildirirken B örneğinde o konum yoksa B referans olabilir, yetersiz kapsanmış olabilir veya analiz bölgesinin dışında kalabilir. Satırları birleştirmek bu durumları ayıramaz. Referans güven iş akışında örnek dosyaları ortak aday lokusların yeniden değerlendirilmesi için gereken bilgiyi korur; ortak genotipleme, örneğe özgü genotipler içeren bir kohort temsili üretir [6].

GATK yolunda, uyumlu tek örnek gVCF'leri genotiplemeden önce GenomicsDB'ye aktarılabilir. Aktarma bir depolama ve veri erişimi aşamasıdır; son kohort genotiplerini üreten istatistiksel karar değildir. Tutarlı referanslar, uyumlu başlıklar, benzersiz ve doğru örnek kimlikleri ile belirtilmiş analiz aralıkları gerekir. Belgelenmiş örnek eşleme optimizasyonu, başlık birleştirmesinin bazı kontrollerini gevşettiğinden önceden uyumluluk kontrolü özellikle önem kazanır [9].

Kohort üyeliği, sonucun kaynak kaydının parçası olur. Örnek eklemek, keşfedilen alel kümesini ve sonraki anotasyon veya filtreleme özetlerini değiştirebilir; kohort tanımını ve sürümünü koruyun. Akrabalık ve popülasyon yapısı da sonraki analizlerde ve kullanılan önsel olasılıklarda önem taşır. Ortak genotipleme kontaminasyonu gidermez, tek bölgede birleşmiş referansı onarmaz veya düşük güvenli kanıtı bağımsız gözleme dönüştürmez.

6.3 Biyolojik anlam yüklemeden önce çağrıları inceleyin

İlk çağrı kümesi, modele dayalı adaylardan oluşur. Örnek düzeyindeki eksiklik, derinlik ve kalite dağılımlarını; varyant sınıflarını ve genomik dağılımı deneyle ilişkili olarak inceleyin. Teknik partiyle ilişkili değişiklikleri ve haploit olması beklenen bölgelerde beklenmedik derecede çok heterozigot çağrıyı araştırın. Bu işaretler kontaminasyondan, karışmış örneklerden, duplike bölgelerden, yanlış ploididen veya gerçek biyolojiden kaynaklanabilir. Tek özet grafikten neden belirlemek yerine açıklamaları bağımsız kanıtla ayırın.

Yerel kanıtın önemli olduğu durumlarda temsilî lokusları bir hizalama görüntüleyicisinde inceleyin. Yerleşim belirsizliğine, okuma başlangıçlarına, zincir ve okuma konumu örüntülerine, yakındaki indellere ve

gerçek referans bağlamına bakın. Görsel inceleme, bildirilen kayıt ile onu destekleyen okumalar arasındaki uyuşmazlığı ortaya çıkarabilir; aynı zamanda seçicidir ve gözlemciye bağlıdır. Sistematik performans değerlendirmesinin yerini tutmaz veya hiç çağrılmamış bölgelerin duyarlılığını belirlemez.

Uygulama notu

Üç çıktıyı kavramsal olarak ayırın: Aday lokus, örnek genotipi ve güvenilir çağrı yapılabilen bölge. Sonraki filtreleme lokusu kabul ederken bir örneğin genotipini maskeleyebilir; yakın varyantlar filtreleri geçse bile bir bölgede güvenilir çağrı yapılamayabilir. Bu ayrımları anotasyona ve sonraki alel frekansı veya ilişkilendirme analizlerine taşıyın.

Sonraki aşamalarda alel temsili normalleştirilir, gerekçelendirilmiş lokus ve genotip filtreleri uygulanır, performans değerlendirilir ve işlevsel anotasyonlar eklenir. Her aşama, çağrı kümesinden güvenle çıkarılabilecek sonuçları değiştirir. Biyolojik açıdan ilginç bir alelin kanıtına ve modeline kadar izlenebilmesi için özgün aday çağrıları ve dönüşüm kaydını koruyun.

7 Filtreleme: Hangi kanıtın sonucu destekleyebileceğine karar vermek

Varyant çağırıcı, istatistiksel bir model altında genotip atar; filtreleme ise bu atamalardan hangilerinin amaçlanan analizde kullanılabileceğini belirler. Bunlar farklı kararlardır. Bir lokusta alternatif alelin bulunduğuna dair güçlü kanıt varken, örneklerden birinin genotipini güvenilir biçimde belirleyecek veri yetersiz olabilir. Tersine, tek bir örnekteki ikna edici genotip, kohort genelinde yanlış hizalamadan kaynaklanan bir lokusu güvenilir yapmaz. Bu nedenle bütün VCF dosyasına tek bir kalite eşiği uygulamak yerine lokus, genotip, örnek ve bölge düzeyindeki kararları ayırın.

7.1 Üç filtreleme düzeyi

Lokus filtresi, bir kaydı veya alternatif aleli hizalama kalitesi, zincir dengesizliği, derinlik ve okuma içi konum yanlılığı gibi kanıtlarla değerlendirir. *Genotip filtresi*, tek bir örneğin atamasını genotip kalitesi, bilgi sağlayan derinlik ve alel desteği üzerinden değerlendirir. *Örnek filtresi* ise kontaminasyon, beklenmeyen örnek kimliği veya genom genelinde yetersiz kapsama gibi sorunları ele alır. Her kuralın hangi düzeyde uygulandığını kaydedin. Bir kaydın `PASS` olması, o kayıttaki bütün genotiplerin kullanılabilir olduğu anlamına gelmez.

Ek açıklamaların anlamını ve kapsamını VCF başlığı tanımlar. `INFO/DP` ile `FORMAT/DP` aynı niceliği göstermeyebilir; çağırıcıya özgü derinlik, iç filtrelerin reddettiği okumaları dışarıda bırakabilir. `QUAL` lokus düzeyindeki bir hipotezle, `GQ` ise belirli bir model altında genotiple ilgilidir. Bu değerler, farklı çağırıcılar arasında evrensel olarak kalibre edilmiş olasılıklar gibi kullanılmamalıdır. Daha sonraki kararları yeniden değerlendirebilmek için filtre etiketlerini ve özgün çağrı kümesini koruyun. Nokta ile gösterilen eksik filtre durumu, değerlendirilip geçtiği açıkça belirtilen durumdan farklıdır [3].

Alel dengesi grafiği, genotipi onu destekleyen okumalarla ilişkilendirdiği için yararlıdır. Sıradan bir diploit heterozigotta iki alelin benzer destek alması yararlı bir beklentidir; kesin bir koşul değildir. Örnekleme değişkenliği, referans yanlılığı, yerel kopya sayısı ve hizalama hataları oranı değiştirebilir. Beklenti, ploidi ve biyolojik bağlamla birlikte değişir. Haploit izolatlara, poliploitlere, havuzlanmış örneklere veya düşük frekanslı somatik varyantlara diploit alel dengesi kuralı uygulamayın.

Kurmaca, iki alelli diploit bir lokusta üç örneğin genotipleri `0/0`, `0/1` ve `1/1`; FORMAT/DP değerleri sırasıyla 20, 5 ve 30 olsun. Başlangıçta `AN=6`, `AC=3`, `AF=3/6 =0.5` olur. AN, çağrılmış alelleri; AC ise alternatif alelin kopyalarını sayar. Öğretim amacıyla seçilmiş `FORMAT/DP<10` kuralı ikinci örneği maskelerse genotipler `0/0`, `./.` ve `1/1` olur; yeniden hesaplama `AN=4`, `AC=2`, `AF=2/4 =0.5` verir. Frekans aynı kalsa da payda küçülmüştür. Bu eşik önerilen filtre değildir. Maskeleme, genotipi `0/0` yapmaz; eşlikçi maskeleme aşaması INFO sayımlarını kendiliğinden yenilemez. Frekans analizinden önce sayımları korunan genotiplerden yeniden hesaplayın; örneğin `bcftools +fill-tags` ile `-t AN,AC,AF` kullanılabilir [17].

7.2 Kuralları deneyden hareketle seçmek

Ek açıklama dağılımlarını SNP ve indeller, ilgili kapsama katmanları ve dizileme partileri için ayrı ayrı inceleyin. Böyle bir kanıt varsa, gerçek olması olası varyantları bağımsız biçimde desteklenen hatalarla karşılaştırın. Çok yüksek derinlik, referansta tek bölgeye indirgenmiş bir tekrarı gösterebilir; düşük derinlik ise genotipi belirsizleştirebilir. Aynı sayısal eşik, hedef zenginleştirme deneyleri, genomlar ve organizmalar arasında bu durumların tümünü temsil edemez. Önerilen sınırlara yakın varyantları inceleyin; her kuralın kaç lokusu ve genotipi değiştirdiğini kaydedin.

GATK, sabit eşiklerle filtreleme ile varyant kalite skoru yeniden kalibrasyonunu (VQSR) ayırır. Sabit eşiklerle filtreleme, belirli ek açıklamalara açık eşikler uygular. VQSR ise uygun kaynaklar ve yeterli sayıda bilgi sağlayan varyant kullanarak bir ek açıklama modeli öğrenir; her veri kümesinde kendiliğinden iyileşme sağlamaz. Yayımlanmış insan örnekleri, ek açıklamaların davranışını açıklar; farklı bir tür, küçük bir hedef paneli veya değiştirilmiş laboratuvar protokolü için eşik belirlemez. İşleme sırasında kullanılan bir bilinen varyant kaynağı da kendiliğinden bağımsız doğruluk kümesi sayılmaz [11].

Popülasyon kontrollerinin kendilerine özgü varsayımları vardır. Beklenmeyen heterozigotluk, eksiklik, akrabalık veya alel frekansı örüntüleri bir sorunu gösterebilir; gerçek biyolojiyi de yansıtabilir. Hardy–Weinberg testi, uygun bir popülasyon modeli varsayar. Seçilim, popülasyon yapısı, akraba çiftleşmesi, klonalite, cinsiyete bağlılık ve örnek seçimi, basit bir dışlama kuralını yanıltıcı kılabilir. Bir popülasyon istatistiğini hata belirteci olarak kullanmadan önce deney tasarımını inceleyin. Olgu-kontrol analizinde gruplar arasında farklı eksiklik özellikle dikkate alınmalıdır; teknik nedenlerle gözlenemeyen veriler, biyolojik bir farkı taklit edebilir [26].

7.3 Güvenilir çağrı yapılabilen bölgeler paydayı belirler

Bir varyant kaydının bulunmaması, referans genotipini kanıtlamaz. Bir konumda kapsama olmayabilir, hizalama ölçütleri sağlanmayabilir, konum hedeflerin dışında kalabilir veya çağırıcı tarafından dışlanmış olabilir. Sorunun gerektirdiği kanıta göre güvenilir çağrı yapılabilen bölgeyi tanımlayın, bölge maskesini saklayın ve amaçlanan analiz alanıyla ilişkisini raporlayın. Örneğe özgü maske, o örneğin nerelerde değerlendirilebildiğini gösterir; kohort maskesi ise başka bir soruyu yanıtlar ve örneklerin nasıl birleştirildiğini belirtmelidir.

Örneğin iki örneğin varyant yükünü karşılaştırmak, örneklerden birinde değerlendirilebilir dizi çok daha azsa güçleşir. Yalnızca korunan varyantları saymak bu farkı gizler. Benzer biçimde, bir gendeki negatif bulgu, güvenle değerlendirilen kodlayan bazlarla kapsanmayan ekzonları ve zor paralog bölgeleri ayırmalıdır. BED dosyası genellikle sıfır tabanlı, sağ ucu dışlayan aralıklar kullanırken VCF konumları bir tabanlıdır; kesişim almadan önce bu kuralları doğrulayın. Özellikle indel sınırlarında dikkat gerekir; dayanak baz içeren kayıt, biyolojik olaydan farklı biçimde bir bölgeyle örtüşebilir.

Dikkat edilmesi gerekenler

Filtreleme, verilerin destekleyebileceği iddiaları değiştirir. Korunan kayıt sayısını, eksik duruma getirilen genotipleri, dışlanan örnekleri ve güvenilir çağrı alanının dışında kalan baz veya aralıkları ayrı ayrı raporlayın. Düzenli görünen bir VCF, dışlanan dizinin başarıyla incelendiğini göstermez.

8 Normalizasyon: Karşılaştırılabilir varyant gösterimleri

VCF kaydı, dizi farkının bir gösterimidir; farkın kendisi değildir. Tekrarlı bir dizide yinelenen bir bazın silinmesi, aynı alternatif haplotipi üretmesine rağmen birkaç farklı konumda yazılabilir. Bitişik baz değişimleri, tek bir çok nükleotitli varyant veya birkaç SNP olarak gösterilebilir. Veri tabanlarıyla eşleştirmeden, kayıtları birleştirmeden veya çağırıcıları karşılaştırmadan önce, bu işlemde eşdeğerliğin ne anlama geldiğine karar verin.

8.1 Referansa dayalı normalizasyon

Küçük varyantların normalizasyonunda genellikle geçerli VCF alelleri korunarak gereksiz ortak dizi çıkarılır ve indeller seçilen referansa göre sola hizalanır. Bu nedenle referans FASTA dosyası işlemin bir parçasıdır. Kontig adları ve dizileri VCF ile uyuşmalıdır; koordinatları dönüştürmeden yalnızca montaj

etiketini değiştirmek uyumsuzluğu düzeltmez. Normalizasyon aracı, farklı bir biyolojik iddiayı sessizce üretmek yerine tutarsız referans alellerini ortaya çıkarmalıdır [16, 3].

Sentetik `CAAAAG` referans parçasını düşünün. C bazını 1. konum sayarsak VCF'deki `POS=4, REF=AA, ALT =A` gösterimi, referansa göre sola hizalamadan sonra `POS=1, REF=CA, ALT=C` olur. İkisi de `CAAAG` dizisini üretir; korunan dayanak baz, silinmenin geçerli bir VCF kaydıyla yazılmasını sağlar. Sola hizalama, tam olarak bu diziye göre tutarlı bir gösterim seçer. Hangi fiziksel A'nın silindiğini belirlemez, okuma kanıtını güçlendirmez veya yakın iki çağrının aynı kromozomda oluştuğunu kanıtlamaz. Ters zincirdeki transkripte göre yazılan bir varyantın da genomik referans gösterimiyle uzlaştırılması gerekir.

Aşağıdaki komut, belgelenmiş bir `bcftools norm` işlemini örnekler. Uyumlu ve indekslenmiş bir referans ile kayıtları incelenmiş bir girdi VCF'si varsayar. Dosya adları yer tutucudur; komut, bu notta herhangi bir veri kümesinin işlendiğini göstermez.

Komut satırı

```
bcftools norm -f reference.fa -m -any \
  -Oz -o normalized.vcf.gz calls.vcf.gz
bcftools index normalized.vcf.gz
```

Burada `-f`, referans denetimleri ve normalizasyon için kullanılacak diziyi sağlar; `-m -any` çok alelli kayıtları ayırır, `-Oz` ise sıkıştırılmış VCF yazar. Kayıt ayırma, sonraki analize yönelik bilinçli bir tercihtir; normalizasyonun zorunlu tanımı değildir. Dönüşümden sonra çıktı başlığını ve örnek genotipleri inceleyin. Düzenli görünen bir dosyanın anlamı, sonraki analiz için yine de uygun olmayabilir [16, 17].

8.2 Çok alelli lokuslar ilişkiler taşır

İki alternatif aleli olan diploit bir lokusta `GT=1/2`, örneğin birinci ve ikinci alternatif aleli taşıdığını gösterir. Referansa göre iki bağımsız heterozigot anlamına gelmez. Bir araç bu kaydı ayırdığında, alel indeksli ek açıklamalar ve genotip olabilirlikleri uygun biçimde yeniden eşlenmelidir. Başlıktaki `Number=A`, `Number=R` ve `Number=G` bildirimleri sırasıyla alternatif alel başına, referans dâhil alel başına ve genotip başına değerleri tanımlar. Seçilen dönüşümün dosyadaki alanları doğru işlediğini denetleyin [3].

Alel ilişkileri önemliyse özgün kaydı veya ona izlenebilir bir bağlantıyı koruyun. Genotiplerin maskelenmesi ya da örnek alt kümesinin seçilmesi sonrasında alel sayıları ve frekanslarının yeniden hesaplanması gerekebilir. Ayrılan her kayda kopyalanan lokus düzeyindeki ek açıklama, alele özgü kanıt olmak zorunda değildir. Yalnızca iki alelli kayıt kabul eden bir program, girdiye bir kısıt koyar; bu kısıt lokusun biyolojik modeliyle karıştırılmamalıdır.

8.3 Haplotipler ve dizi etkilerinin yorumlanması

Aynı kodondaki iki baz değişimi birlikte yorumlandığında, ayrı ayrı anotasyonlarından farklı bir amino asit etkisi öngörülebilir. Dolayısıyla değişimlerin *cis* mi yoksa *trans* mı olduğu önemlidir. Fazı belirlenmemiş bir genotip bu soruyu yanıtlamaz. Benzer biçimde, birkaç indel ile gösterilen karmaşık bir değişim, VCF satırları tek tek eşleşmese de doğruluk haplotipiyle eşleşebilir. Kayıt düzeyinde eşitlik bazı karşılaştırmalar için fazla katıdır; genotip uyumu göz ardı edilirse başka karşılaştırmalarda fazla gevşek kalabilir.

Performans değerlendirmesinde haplotipleri dikkate alan bir karşılaştırma yöntemi kullanın; karşılaştırmanın alelleri, genotipleri veya fazı mı değerlendirdiğini belirtin. Her olayı yalnızca satırları tek biçime sokmak için parçalamak yerine, anotasyona uygun gösterimi koruyun. Normalizasyon, parçalama, faz belirleme ve referans montajı dönüştürme farklı sorunları çözer; birinin yapılması diğerlerinin doğruluğunu göstermez. Gösterim değişikliğinden sonra varyant sınıfına göre sayıları karşılaştırın ve değişen karmaşık lokusları ilerlemeden önce inceleyin [23].

9 Anotasyon: Dizi farkını biyolojik modelle ilişkilendirmek

Varyant anotasyonu, dış modelleri ve kanıtları kullanarak varyant hakkında açıklamalar ekler. Dizi etkisi tahmin aracı, bir alelin kodlayan diziyi değiştirip değiştirmediğini, bir kesip birleştirme bölgesiyle örtüşüp örtüşmediğini veya açıklanmış bir düzenleyici aralığa düşüp düşmediğini değerlendirebilir. Popülasyon veri

tabanı, gözlenen frekansı sağlayabilir. Deneysel çalışma ise işlevsel etkiyi destekleyebilir. Bu açıklamalar farklı soruları yanıtlar ve anotasyonlu çıktıda ayırt edilebilir kalmalıdır.

9.1 Uyumlu anotasyon kaynaklarını seçmek

Geçerli bir anotasyon kaynak kümesi; doğru referans montajını, gen modellerini, transkript tanımlayıcılarını ve kaynak sürümlerini bir araya getirir. Kontig adlandırmasını, koordinat kurallarını, ilgili durumlarda genetik kodu ve her veri tabanının kökenini doğrulayın. Tür adı tek başına yeterli değildir; aynı türün iki montajında koordinatlar ve gen modelleri farklı olabilir. Sürümü belirtilmeyen transkript erişim numarası da kodlayan dizideki bir değişikliği gizleyebilir. Genomik aleli ve öngörülen dizi etkisinin tanımlandığı transkripti birlikte kaydedin.

Ensembl VEP, türe özgü bir önbellek, desteklenen GFF/GTF transkript modelleri ve özel anotasyon kaynakları kullanabilir. Önbellek sürümü, yazılımın belgelenmiş uyumluluk koşullarını karşılamalı; referans montajı girdiyle eşleşmelidir. Çevrim dışı çalışma, seçilen kipin gerektirdiği yerel kaynaklara ihtiyaç duyar. Önbellek dizininin bulunması, referans ve yapılandırma kontrolünün yerini tutmaz. SnpEff kendi genom veri tabanlarını kullanır; uyumlu referans dizileri ve gen anotasyonlarından veri tabanı da oluşturabilir. Bunlar alternatif anotasyon yollarıdır; tahmin edilen etkiyi birbirinden bağımsız deneylerle doğrulamazlar [18, 19, 21, 22].

Model olmayan bir organizmada önce gen anotasyonunun kendisini inceleyin. Eksik ekzonlar, tamamlanmamış kodlayan diziler veya yanlış çeviri tablosu, varyant doğru çağrılsa bile yanıltıcı sonuçlar üretebilir. Özel SnpEff veri tabanı, mevcut kodlayan diziler ve protein dizileriyle kontrol edilmeli; uyuşmazlıklar araştırılmalıdır. Özel gen modelleriyle VEP kullanırken desteklenen biçimi ve gerekli indekslemeyi sürüme özgü kılavuzdan doğrulayın. Uygun gen modeli yoksa örtüşme ve dizi kanıtını gerçekten desteklenen düzeyde raporlayın; ilgisiz bir türün koordinatlarından güvenilir bir protein etkisi atamayın.

9.2 Örnek anotasyon komutları

Aşağıdaki VEP örneği; yazılımın, uyumlu yerel önbelleğin ve eşleşen indekslenmiş FASTA dosyasının hazır olduğunu varsayar. Kabuk değişkenlerini deney için doğrulanmış tür adı, montaj etiketi ve önbellek konumuna göre ayarlayın. Bu yer tutucular özellikle insan referansını dayatmaz. Yerel önbellek sürümü, kurulu VEP sürümünün uyumluluk kurallarını karşılamalıdır.

Komut satırı

```
vep --input_file normalized.vcf.gz \
  --output_file annotated.vep.vcf --vcf \
  --offline --cache --dir_cache "$cache_dir" \
  --species "$species" --assembly "$assembly" \
  --fasta reference.fa
```

Komut, VCF çıktısı ve yerel anotasyon kaynakları ister. VEP alanlarını ayrıştırmadan önce VCF başlığındaki `INFO/CSQ` bildirimini inceleyin. Bu bildirimin `Format:` açıklaması alan sırasını verir; sıra eski bir örnekten tahmin edilmemelidir. Uyarıları ve örnek anotasyonlu alelleri değerlendirin. İşlemin başarıyla bitmesi, bütün transkript modellerinin referansla uyuştuğunu kanıtlamaz [19, 27].

Alternatif SnpEff örneği, girdinin montajı ve anotasyon sürümüyle eşleşen, önceden kontrol edilmiş yerel bir veri tabanını varsayar. `database_id` değişkenini bu veri tabanının tam tanımlayıcısına ayarlayın; yalnızca tür adı yeterli değildir.

Komut satırı

```
java -jar snpEff.jar "$database_id" \
  normalized.vcf.gz > annotated.snpeff.vcf
```

Öngörülen bütün dizi etkilerini eşit derecede güvenilir saymak yerine `ANN` başlığını ve alel düzeyindeki uyarıları inceleyin. Uygun hazır veri tabanı yoksa bu adımdan önce eşleşen FASTA ve gen modellerinden özel veri tabanı oluşturup kontrol edin. Veri tabanı oluşturma yolu, ayrıca dizi ve çeviri kontrolleri gerektirir. Bu komutlar birbirinin alternatifidir; şemalarını uzlaştırmadan iki anotasyon çıktısını birleştirme talimatı değildir [28, 22].

9.3 Bir alelin transkripte bağlı etkileri

Bir varyant, bir transkriptte kodlayan bölgede, diğerinde intronda, üçüncüsünün ise dışında bulunabilir. Özet seçmeden önce transkript düzeyindeki dizi etkilerini koruyun. En ağır öngörülen etki, ön elemede yararlıdır; ancak kanıt desteği zayıf bir transkripti öne çıkarabilir. Ana veya kanonik transkript, yeniden üretilebilir bir seçim kuralı sağlar; incelenen doku ya da durum için biyolojik olarak ilgili transkript olduğunu garanti etmez. Kuralı belirtin ve önemli adayları incelemek için gereken alternatifleri saklayın [18].

Standart çekirdek genetik kodunu kullanan, artı zincirdeki kurmaca bir örnekte genomik `G>A` değişimi, T1 transkriptinin `GAA` kodonunu `AAA` yapar. Glutamat yerine lizin öngörülür; bu, yanlış anlamlı bir dizi etkisidir. Aynı genomik baz T2 transkriptinin intronundaysa bu transkript için intronik anotasyon verilir. Bunlar tek alelin transkripte bağlı iki açıklamasıdır; iki ayrı varyant veya transkript ifadesi ölçümü değildir. Transkript tanımlayıcılarını ve sürümlerini saklamak farkın incelenmesini sağlar.

Eş anlamlı varyant, belirli bir transkriptte kodlanan amino asidi korur; bu terim bütün moleküler süreçlerin etkilenmediğini göstermez. Yanlış anlamlı varyant amino asidi değiştirir, ancak tek başına işlev bozukluğunu kanıtlamaz. Tahmin edilen çerçeve kayması veya dur kodonu oluşumu, transkript yapısına, okuma çerçevesine ve dizi bağlamına bağlıdır. Öngörülen dizi etkisi, alelin modele dayalı tanımıdır. Biyolojik önemi için soruya uygun kanıt gerekir.

Dikkat edilmesi gerekenler

`HIGH` veya `MODERATE` gibi etiketler, öngörülen dizi etkisi kategorilerini özetler. Bunlar patojenite olasılığı, uyum başarısı ölçümü veya fenotipe varyantın neden olduğunun kanıtı değildir. Öngörülen dizi etkisini, işlevsel kanıtı, popülasyon kanıtını ve son yorumu ayrı alanlarda tutun [20].

9.4 Kanıtı bağlamını kaybetmeden eklemek

Özel veri tabanları popülasyon frekanslarını, korunmuşluk skorlarını, küratörlü fenotipleri veya projeye özgü aralıkları ekleyebilir. Eşleştirmeden önce referans montajını ve alel gösterimini uzlaştırın. Bir aralıkla örtüşme, normalize edilmiş alelin tam eşleşmesinden farklı bir açıklamayı destekler. Veri tabanında eşleşme bulunmaması; alelin veri tabanında olmamasından, kapsama eksikliğinden, gösterim uyuşmazlığından veya kaynağın organizmayı kapsamamasından kaynaklanabilir. Bu durum alelin kendiliğinden nadir olduğu anlamına gelmez.

Frekans için, mevcutsa popülasyon tanımını ve paydayı saklayın. Bir popülasyondaki frekans tahmini değişmeden başka popülasyona atanamaz; veri tabanında gözlenmiş alel de bu nedenle zararsız sayılmaz. İşlev tahmini skorlarında yöntem ve sürümü koruyun. Tahmin araçlarını birleştirirken ortak eğitim verilerini veya özellikleri dikkate alın. İlişkili tahminleri ayrı kanıt çizgileri gibi saymak desteği olduğundan güçlü gösterebilir.

Anotasyonu, sentetik veya iyi tanımlanmış az sayıda örnekle denetleyerek tamamlayın: Bir kodlayan baz değişimi, bir kodlamayan alel, bir indel ve ilgili bir sınırı olan transkript. Alel yönelimini, transkript kimliğini ve beklenen kategoriyi kontrol edin. Bu denetim yapılandırmayı ve yorumlamayı sınar; deneyin tümündeki duyarlılığı veya özgüllüğü ölçmez.

10 Doğrulama: İddiayı açıkça tanımlanmış bir alanda ölçmek

Makul görünen bir iş akışı, amaçlanan görevi yerine getirdiğine dair kanıt gerektirir. İyi tanımlanmış bir örnek üzerinden performans değerlendirmesi, belirli karşılaştırma kuralları altında referans çağrı kümesiyle uyumu ölçer. Okuma hizalamalarının incelenmesi, tekil adayları araştırır. Bağımsız laboratuvar deneyi ise seçilmiş aleli sınayabilir. Bunların hiçbiri tek başına bütün doğrulama sorularını yanıtlamaz. Kanıtı iddiaya göre seçin; iş akışının değerlendirilmesi ile belirli bir sonucun doğrulanmasını ayırın.

10.1 Doğruluk kümesi varyantlarla birlikte bölgeleri de içerir

Küçük varyant performans değerlendirmesinde genellikle bir doğruluk VCF'si ve güvenilir bölge BED dosyası bulunur. BED, referans çağrı kümesinin karşılaştırmayı desteklediği alanı tanımlar. Bunun dışındaki diziler güvenle referans kabul edilmemelidir. Tam olarak eşleşen montajı ve değerlendirme kaynağı

sürümünü kullanın. İnsan Genome in a Bottle (GIAB) kaynakları, tanımlanmış insan örnekleri ve değerlendirme bölgeleri sağlar. İnsan iş akışı geliştirmedeki yararları; mantar, bitki, havuzlanmış popülasyon veya tümördeki performansı kanıtlamaz [24, 23].

Hedef zenginleştirme deneyinde karşılaştırma ayrıca amaçlanan hedef alanıyla sınırlandırılabilir. Oluşan kesişimi ve sınırları aşan olayların nasıl ele alındığını belirtin. Değerlendirme alanının dışındaki doğruluk varyantı, sıradan bir yanlış negatif değildir. Benzer biçimde, güvenilir doğruluk alanının dışındaki ek bir çağrı, yalnızca doğruluk VCF'sinde bulunmadığı için yanlış pozitif sayılamaz. Dışlanan ve değerlendirilmeyen sonuçları görünür tutun.

Doğruluk kanıtı, iddianın gerektirdiği ölçüde bağımsız olmalıdır. Filtreleri tek bir değerlendirme örneğiyle eşleşene kadar ayarlayıp aynı örneği tarafsız performans tahmini olarak sunmak, değerlendirme verisinin yöntem geliştirmeye sızmasına yol açar. Kaynaklar elveriyorsa geliştirme kararlarını son değerlendirmeden ayırın. Aynı okumaları kullanan iki çağırıcının uyumu yararlı bir tanısal kanıttır; ancak ortak hizalama hataları ve ilişkili modeller, bu uyumun tam bağımsız doğruluk olarak kullanılmasını engeller.

10.2 Haplotipleri karşılaştırmak ve paydayı raporlamak

`hap.py` gibi haplotipleri dikkate alan araçlar, doğrudan satır eşleştirmesini yanıltıcı yapan gösterim farklarını ele alır. Karşılaştırma ayrıca genotip eşleştirmesini, filtrelemeyi, referans dizisini, bölgeleri ve varyant kategorilerini tanımlamalıdır. Sayıları yorumlamadan önce aracın çıktı kurallarını inceleyin; gösterime duyarlı eşleştirme sonrasında doğruluk ve sorgu tarafındaki olay sayımları farklı olabilir [23].

Eşleştirme tanımı üzerinde anlaşılmış basit bir olay sayımı örneğinde, pozitif öngörü değeri (precision) ve duyarlılık (recall) şöyledir:

$$\text{precision} = \frac{TP}{TP + FP}, \qquad \text{recall} = \frac{TP}{TP + FN}.$$

Pozitif öngörü değeri, değerlendirilen çağrıların ne kadarının doğruluk kümesiyle desteklendiğini; duyarlılık ise değerlendirilen doğruluk varyantlarının ne kadarının bulunduğunu sorar. Payda sıfırsa ilgili ölçüt kendiliğinden kusursuz değil, tanımsızdır. Ölçütlerle birlikte temel sayıları ve değerlendirilen alanı raporlayın. Performans tahmini yaparken örnekleme ve bağımlılık yapısına uygun bir aralıkla belirsizliği nicelleştirin. Az sayıda varyantın değerlendirilmesi, bütün bağlamlarda hassas bir performans tahmini sağlamaz [23]. Sentetik bir örnekte 90 eşleşen çağrı, desteklenmeyen 10 çağrı ve kaçırılan 20 doğruluk varyantı varsa pozitif öngörü değeri $90/100 = 0{,}90$, duyarlılık ise $90/110 \approx 0{,}818$ olur. Bunlar aritmetik örneklerdir; gözlenmiş iş akışı sonuçları değildir.

Genom genelindeki ortalama, deney açısından önemli hataları gizleyebilir. Değerlendirme kaynağı bu kategorileri destekliyorsa SNP ve indelleri, indel uzunluklarını, kapsamayı ve zor dizi bağlamlarını ayrı değerlendirin. Tekrarlar ve düşük karmaşıklıktaki bölgeler, daha kolay dizilerden farklı davranabilir. Kaynağın değerlendirmediği varyant sınıfları dâhil kapsama sınırlarını raporlayın. Küçük varyant değerlendirmesi; yapısal varyantları, kopya sayısı değişikliklerini veya tıbbi açıdan ilgili her lokusu doğrulamaz [23].

10.3 Uyuşmazlıkları ve seçilmiş adayları araştırmak

Okumaların görselleştirilmesi; desteğin tek zincirle sınırlı kalmasını, soft clipping durumunu, yakın indelleri, düşük hizalama kalitesini veya desteğin okuma uçlarında yoğunlaşmasını gösterebilir. Yalnızca işaretli baza bakmak yerine alternatif ve referans okumalarını karşılaştırın, çevredeki diziyi inceleyin. Bu gözlemler, çağrının neden araştırılması gerektiğini açıklar; görsel güven tek başına kalibre edilmiş bir doğruluk ölçütü değildir. Okuyucu, gözlenen okuma örüntüsünü onun için önerilen açıklamadan ayırabilmelidir.

Bağımsız deneylerin kendilerine özgü tasarımları ve saptama sınırları vardır. Yüksek frekanslı germ hattı aleli için seçilen deney, düşük frekanslı mozaik alele veya tekrarlı lokusa uygun olmayabilir. Hangi adayların sınandığını, neden seçildiklerini ve deneyin neleri saptayabildiğini raporlayın. Seçim sırasında belirsiz çağrılar dışlanmışsa, güçlü birkaç adayın doğrulanması bütün çağrı kümesinin pozitif öngörü değerini tahmin etmez. Uygun değerlendirme kaynağı veya bağımsız kanıt yoksa evrensel bir performans değeri üretmek yerine tamamlanan kontrolleri ve kalan belirsizliği açıklayın.

11 Özel yollar: Biyoloji değiştiğinde modeli değiştirmek

Ortak iş akışı; referans uyumunu, okuma kalitesini, varyant kanıtını, gösterimi ve anotasyonu ele alır. Bu akışın içinde kullanılan model yine de örneğe uygun olmalıdır. Çağırıcı seçmeden önce organizmayı, bölgeye göre beklenen ploidiyi, materyalin tek birey mi karışım mı olduğunu, dizileme teknolojisini ve hedeflenen alel frekansı aralığını yazın. Aracın bir girdi türünü desteklemesi yalnızca başlangıçtır; ilgili koşullardaki performans ayrıca değerlendirilmelidir.

11.1 İnsan germ hattı örnekleri

İnsan germ hattı analizi; gelişmiş referans montajlarından, popülasyon kaynaklarından ve değerlendirme örneklerinden yararlanır. Bu kaynaklar montaja özgüdür; farklı popülasyonları ve genomik bölgeleri eşit ölçüde kapsamaz. Cinsiyet kromozomları ve diğer bölgeye özgü ploidileri bilinçli seçin. Örneğin tümüne uygulanan tek bir diploit varsayımı, her genomik bölme için doğru değildir. Heteroplazmi incelendiğinde mitokondri genomu da ayrı bir analiz sorunudur.

11.2 Diğer diploit organizmalar: Kaynak koşullarını yeniden belirlemek

Diploit bir model organizmada işlemlerin genel sırası yararlı kalabilir; ancak insana ait bilinen varyantlar, frekans eşikleri, transkript öncelikleri ve doğruluk kaynakları kendiliğinden aktarılamaz. Montaj kalitesini ve suş farklılığını kontrol edin. "Alternatif" olarak çağrılan alel, seçilen referanstan farklıdır; zorunlu olarak yeni oluşmuş, nadir, zararlı veya türde bulunmayan bir alel değildir. Referans seçimi hem hizalamayı hem yorumu etkiler.

DeepVariant belgeleri, ploidi ve insan dışı uygulamalarla ilgili sınırlamalar dâhil desteklenen modelleri ve kullanım alanlarını ayırır. Bir koşulda eğitilmiş model, yazılım çalışıyor diye başka koşul için doğrulanmış sayılmamalıdır. Teknolojiye ve deney türüne uygun model seçin; yayımlanmış doğruluk değerini aktarmak yerine sonucu ilgili kanıtla değerlendirin [12, 13].

Bu yolda önce organizmaya ait referansı ve gen modellerini hazırlayın; materyal ve analiz bölgeleri için diploidiyi doğrulayın. Ardından okuma kalite kontrolünü, hizalamayı ve kütüphaneyi gözeten duplikasyon hazırlığını yürütün. Sunulan diploit çağrı yolunu ancak çağırıcının önsel varsayımlarını ve deney uyumunu inceledikten sonra kullanın. Uyumlu bilinen varyantlar ve gerekçeli bir protokol yoksa isteğe bağlı BQSR'yi atlayın; insan kaynağı kullanmayın. Filtreleri organizmaya ait kanıtla seçip değerlendirin. Aynı FASTA ile normalizasyon yapın, eşleşen modellerle anotasyon ekleyin ve kapsama, genotip ve lokus görünümlerini inceleyin. Uygun doğruluk kümesi yoksa tanısal kontrolleri ve seçilmiş bağımsız doğrulamaları raporlayın; çağırıcı uyumunu ölçülmüş doğruluk olarak sunmayın.

11.3 Haploit izolatlar, mikroorganizmalar ve mantarlar

Klonal bir haploit izolatta haploit genotip modeli uygun olabilir. Birden fazla alel görünümü araştırılmalıdır. Olası açıklamalar arasında kontaminasyon, karışık enfeksiyon veya kültür, örnek içi çeşitlilik, çoğalmış diziler ve hizalama hataları vardır. Diploit tipte heterozigotluk, tarafsız bir varsayılan açıklama değildir. Bakteri genomlarında farklı kopya sayılarında plazmitler bulunabilir; bu nedenle kapsama ve alel oranları doğru genomik bölme düzeyinde yorumlanmalıdır.

Mantarlara tek bir evrensel ploidi atanamaz. Tür ve materyale bağlı olarak haploit, diploit, poliploit ve anöploidi içeren yapılar görülebilir. *Saccharomyces cerevisiae* izolatlarını inceleyen bir çalışma, tek bir maya türü içindeki ploidi ve kromozom kopya sayısı çeşitliliğini örnekler [29]. Genotip ve alel dengesi beklentilerini seçmeden önce ilgili biyolojik durumu belirleyin. Yakın bir referans yoksa farklılaşmış veya referansta bulunmayan diziler hizalanamayabilir. Böyle bir durumda referansa dayalı küçük varyant kümesi, bütün genom farklarını değil örneğin hizalanabilen bölümünü tanımlar. Gen varlığı/yokluğu, yeniden düzenlenmeler ve aksesuar diziler ek analiz yolları gerektirir.

Anotasyonda da organizmaya özgü seçimler gerekir. İlgili durumlarda organel veya soya özgü istisnalar dâhil gen modellerini ve genetik kodu doğrulayın. Uyumlu mikrobiyal veya fungal anotasyon yerine insan sonuç veri tabanı kullanmayın. Araştırma sorusuyla ilgiliyse dirençle ilişkili varyant kataloğunun da organizma, alel ve kanıt kapsamı değerlendirilmelidir. Genel bir protein etki etiketi, direnç fenotipini kanıtlamaz.

Sunulan haploit yol, ortak hazırlıktan sonra BCFtools'u açıkça bir ploidisiyle kullanır. Yorumdan önce kontaminasyonu ve beklenmeyen karışık alel kanıtını inceleyin. Çağrıları normalize edin, organizmaya özgü filtreleri seçin, eşleşen genlerle anotasyon ekleyin; derinliği ve aday lokusları inceleyin. Bu yol tek izolata yöneliktir; metagenomik topluluk, havuzlanmış kültür veya karışık enfeksiyon için karışım modeli gerekir.

BCFtools 1.25 `mpileup` varsayılanlarında, girdi dosyası başına ham derinlik üst sınırı 250'dir (`--max-depth`); asgari hizalama kalitesi 0 (`--min-MQ`), asgari baz/BAQ kalitesi ise 1'dir (`--min-BQ`). `--max-depth` değerini yükseltmek daha fazla okuma kanıtının korunmasına izin verebilir; bellek veya işlem yükünü de artırabilir. Düşürmek ise genotiplemeden önce kanıtı azaltabilir. Bu tek örnekli yolda `--max-idepth 250` farklı bir kural uygular: Konumdaki pileup derinliği 250 veya üzerindeyse indel çağırma atlanır. Bu lokuslardan 250 okuma seçilerek devam edilmez. Eşiği yükseltmek daha derin konumlarda indellerin değerlendirilmesine izin verir; düşürmek daha fazla yüksek derinlikli konumu bu değerlendirmeden dışlar. Kalite alt sınırlarını yükseltmek daha fazla hizalamayı veya bazları dışlar. Bunların hiçbiri daha iyi çıkarımı garanti etmez; BAM derinliği, gerçekte kullanılan kanıttan yüksek olabilir. Eşlikçi betik bu varsayılanları korur ve `-a` ile `FORMAT/AD,FORMAT/DP` alanlarını açıkça ister. Bu sürümde AD zaten varsayılan anotasyondur; DP isteğe bağlıdır. AD önce referans, ardından alternatif alellerin derinliğini listeler; FORMAT/DP ise çağırıcının kurallarına göre yüksek kaliteli bazları sayar. Payda seçmeden önce başlık tanımlarını ve BAM kanıtını inceleyin. Diğer çağrı kümelerinde bulunmayan alanlar sıfır değil, eksik kalır [17].

11.4 Poliploitler ve havuzlanmış örnekler

Poliploit bireyde genotip çıkarımı, alel dozajını ve genom kopya sayısını dikkate almalıdır. İki diploit genotipi ayırmaya yeterli kapsama, birbirine yakın dozaj durumlarını ayırmayabilir. Allopoliploitlerde ilişkili homeolog diziler ayrıca hizalama belirsizliği yaratabilir. Ploidi parametresini artırmak, farklı kopyaları tek bölgeye indirgemiş bir referansı düzeltmez. Dozaj belirsizliğini ve referans yapısını birlikte yorumlayın.

FreeBayes, açık ploidi ve havuzlama kiplerini belgeler; bu nedenle modeli bu koşullara göre ayarlanabilen araçlara bir örnektir. Böyle bir kipin seçilmesi yine de uygun kapsama, parametre incelemesi ve doğrulama gerektirir. Havuzlanmış deney, birden çok genomun sağladığı kanıtı ölçer. Okuma derinliği, bağımsız örneklenmiş organizma sayısı değildir. Eşit olmayan DNA katkısı ve hem organizma hem okuma düzeyindeki örnekleme, alel frekansı tahminlerini etkiler. Havuz frekanslarını birey düzeyinde çözülmüş genotipler olarak raporlamayın [14, 15].

Sunulan poliploit yol, ikiden büyük sabit tam sayı ploidisi için FreeBayes kullanır. Hizalamadan önce referansın çökmüş, haplotipleri çözülmüş veya alt genomları ayrılmış yapısının hedeflenen kopya modelini destekleyip desteklemediğini belirleyin. Derinliği, alel dozu belirsizliğini ve homeolog hizalamalarını inceleyin; normalizasyon ve filtrelemeden sonra uyumlu özel anotasyonlar ve dozu gözeten özetler kullanın. Anöploidi veya bölgesel kopya sayısı değişimi bu sabit ploidi betiğinin dışındadır. Paket havuz frekansı çıkarımını uygulamaz.

11.5 Uzun okumalar ve somatik varyasyon

Uzun okumalar; hizalama, faz belirleme ve zor bölgelerdeki kanıtı değiştirebilir. Hata profilleri ise platforma, kimyaya ve işlemeye bağlıdır. Gerçek okuma türü için tasarlanmış hizalayıcı ve küçük varyant modeli kullanın; ilgili model sürümünü kaydedin. Kısa okuma yapılandırması, uzun okuma BAM dosyasını kabul ediyor diye uygun hâle gelmez. Küçük varyantlar ve yapısal varyantlar, farklı değerlendirme koşulları olan ayrı çıktı sınıflarıdır [13].

Somatik analiz, örneklenen dokuda klonal veya subklonal olabilen, germ hattı dışında edinilmiş değişiklikleri arar. Gözlenen alel oranları germ hattı varyasyonuna, saflığa, kopya sayısına ve kontaminasyona da bağlıdır. Diploit germ hattı genotip modeli bu soruyu yanıtlamaz. Eşleştirilmiş normal örnek, germ hattı alellerini ve bazı hataları ayırmaya yardımcı olabilir; ancak normal örneğin kontaminasyonu, normal örneğe tümör karışması ve saptama sınırları yine önemlidir. Yalnızca tümör kullanılan analizde belirsizlikler ve kaynak gereksinimleri farklıdır. Mutect2, özel bir somatik yola örnektir; çağrıları germ hattı `PASS` kuralı yerine kendisine karşılık gelen filtreleme iş akışını gerektirir [25].

Bu yollar, öğretici iş akışının nerede dallanması gerektiğini tanımlar. Hiçbiri evrensel eşik listesi sağlamaz. Biyolojik durum veya teknoloji doğrulanmış modelin dışında kalıyorsa, görünür yokluğu ya da düşük frekanslı varyasyonu yorumlamadan önce uyumsuzluğu belgeleyin ve uygun kanıtı edinin.

İnsanda eşleşmiş somatik yol için tümör ve normali tutarlı hazırlayın, kimliklerini ve eşleşmelerini doğrulayın; ardından somatik çağırıcı seçin. Mutect2 çıktısına, koşula göre kontaminasyonu ve yönelim yanlılığını gözeten kendi filtrelemesi uygulanır. Eşlik eden README gerekli kaynakları ve filtre aşamalarını belirtir. Filtrelenen çağrıları ve alel kanıtını koruyun; seçilen çıktıyı normalize edin, aynı insan montajına ait kaynaklarla anotasyon ekleyin ve tümör/normal görünümlerini inceleyin. Germ hattı popülasyon frekansı destekleyici bağlamdır; eşleşmiş normalin yerini almaz ve somatik kökeni kanıtlamaz. Eşleşmiş çift betikleri yalnızca tümörden çıkarım yapmaz.

Zheng ve arkadaşlarının 2026 derlemesi; derin öğrenme, montaj ve pangenom yaklaşımları dâhil uzun okuma kısa varyant yöntemlerini ele alır. Özeti, farklı hata profillerini ve yeni erişilebilen bölgelerin değerlendirilmesini vurgular. Bu kaynak güncel okuma yönü sunar; çalıştırılmış uzun okuma iş akışı veya bu kısa okuma betiklerine ait doğruluk güvencesi sağlamaz [41].

12 Çağrı kümesini ve lokus kanıtını görselleştirmek

Görselleştirme iki tamamlayıcı soruyu ele alır: Çağrı kümesi örnekler arasında beklenen özellikleri gösteriyor mu ve okumalar belirli bir aleli destekliyor mu? Soruyu önceden tanımlayın; çizilen veriyi, seçim kuralını ve paydayı saklayın. Bir genom görünümü doğruluk ölçümü değildir; ikna edici bir hizalama ekran görüntüsü de bağımsız doğrulama sağlamaz.

12.1 Soruyu yanıtlayan gösterimi seçmek

Biyolojik grupları özetlemeden önce derinlik, eksik genotip oranı ve genotip kalitesini örnek ve parti bazında karşılaştırın. Yalnızca ortalamaları değil dağılımları gösterin; eksik değerleri sıfırdan ayırın. SNP/INDEL sayılarında ve dizi etkisi kategorilerinde birimin VCF kaydı, normalize edilmiş alel, örnek genotipi veya transkripte özgü dizi etkisi olduğunu belirtin. Çok alelli kaydın bölünmesi ya da birden fazla transkriptin anotasyonu, yeni biyolojik gözlem oluşmadan sayıları değiştirir. Varyant yoğunluğunu yalnızca karşılaştırılabilir çağrılabilir bölgelerde karşılaştırın; kapsama farklıyken tüm referans uzunluğunu payda yapmak yapay fark üretebilir.

Çağırıcı yorumlanabilir alel sayıları veriyorsa alel derinliğini toplam alel derinliğine veya alel oranını derinliğe karşı gösterin. Paydayı tanımlayın ve çok alelli alanların başlıkta belirtilen sırasını koruyun. Diploit heterozigot örneğinde yarıma yakın oran modelle uyumlu olabilir; örnekleme, hizalama yanlılığı ve kopya sayısı dağılımı etkiler. Sabit tetraploit bölgelerde doz durumlarının beklentileri farklıdır; somatik örneklerde saflık ve kopya sayısı nedeniyle alel oranı doğrudan hücre oranına çevrilemez. Tüm organizmalara diploit alel dengesi aralığı uygulamayın.

Doğruluk karşılaştırmasında aynı doğruluk kümesini, bölge maskesini ve eşleştirme kuralını kullanarak SNP ve INDEL için pozitif öngörü değeri ile duyarlılığı ayrı gösterin. Kaynak elveriyorsa zor bölgeleri ayrıca değerlendirin. Montaja dayalı X/Y karşılaştırması, güçlü bir genel sonucun her genomik bölmedeki performansı göstermediğini örnekler [38]. Anotasyon kategorilerinin sütun grafiği, uygun arka plan ve istatistiksel model olmadan işlevsel zenginleşme kanıtı sağlamaz.

12.2 Yerel IGV ile aleli incelemek

IGV'ye kullanılan referans FASTA'yı, indeksli BAM/CRAM'ı, indeksli VCF'yi ve uyumlu gen modellerini yükleyin. Lokusa yakınlaşın; geniş ölçekte boş görünen iz, görünürlük penceresinden kaynaklanabilir. INDEL'leri ve okuma ayrıntılarını inceleyin, gerektiğinde okumaları ipliğe göre gruplayın ve hizalama kalitesini değerlendirin. Görüntüleyici sürümünü ve ayarlarını saklayın; kapsama renklendirmesi ile baz kalitesine göre saydamlık görünen kanıtı etkiler [39, 40].

Önceden belirlenen geçen, filtrelenen ve uyumsuz lokusları inceleyin. Somatik çiftlerde tümör ile normali karşılaştırılabilir kapsama ölçekleriyle birlikte gösterin. Genomik aleli, seçim nedenini, gözlenen kanıtı, inceleyeni ve çözülemeyen açıklamaları kaydedin. Çelişen okumaları kırparak gizlemek yerine yakın tekrarları ve kırpılmış hizalamaları inceleyin. Korunan bilgi içeren ekran görüntülerini ve oturumları yerelde tutun; oturumlar dosya konumlarını saklayabilir. Görsel inceleme araştırmayı destekler; gözlenmeyen aleller için duyarlılık tahmini sağlamaz.

12.3 Gösterilen özeti yeniden üretmek

Eşlik eden `09_prepare_review.sh` aşaması, açıkça seçilen VCF'den BCFtools istatistiklerini ve uzun biçimli alel/örnek tablosunu kaydeder. Kendiliğinden yalnızca `PASS` seçmez, genotipleri maskelemez, eksik etiketleri türetmez ve grafik üretmez. Sonraki `10_plot_record_counts.py` aşaması, sınıf/FILTER sayılarının yığılmış gösterimini PNG ve PDF olarak, kaynak TSV ile birlikte üretir. Her VCF kaydını bir kez sayar; çok alelli ve sembolik kayıtları ayırır. Kayıt sayısı alel sayısı veya doğruluk tahmini değildir. Karşılaştırmayı tanımlayın ve betiği koruyun. Yerel inceleme için eşleşen referans ve hizalamaları ayrıca yükleyin; özet VCF tek başına okuma kanıtını yeniden oluşturamaz.

Gösterimi paylaşmadan önce çizilen toplamları kaynak tabloyla karşılaştırın ve küçük, bilinen bir örneği bağımsız sayın. Eksenleri, birimleri, etiketleri, eksik değerleri ve kullanılan örnek/bölge seçimini kontrol edin. Alelin veya filtre durumunun bilinçli değiştirilmesi ilgili özeti değiştirmelidir. Bilimsel şekillerin raster ve vektör çıktıları ayrı incelenmelidir; bu öğretim paketi gerçek veri şekli sunduğunu ileri sürmez, çalıştırılabilir çizim komutlarını sağlar.

13 Yorumlama ve raporlama: Kanıt zincirini korumak

Yararlı bir rapor, okuyucunun sonucu deney sorusundan başlayarak okumalar, genotip, anotasyon ve dış kanıtlar üzerinden izlemesini sağlar. Son aday tablosu bu nedenle bir gen simgesiyle etki etiketinden fazlasını içermelidir. Referans montajımı, normalize edilmiş genomik aleli, örnek genotipini veya alel frekansı tahminini, filtre durumunu, ilgili derinliği ve desteği, transkript tanımlayıcısını, sonucu ve izlem kararını gerekçelendiren kanıtı saklayın. Araştırılmak üzere seçilen aday ile gösterilmiş biyolojik açıklamayı ayırın.

13.1 Sentetik aday üzerinden örnek

Uyumlu referansı ve gen anotasyonu bulunan diploit bir organizmada eğitim amaçlı deney varsayın. Sentetik aday, kodlayan bölgede iki alelli bir baz değişimidir ve çağırıcı, fazı belirlenmemiş heterozigot genotip atamıştır. Korunan gözlemlerin 24 referans destekleyici ve 18 alternatif destekleyici okuma olduğunu varsayalım. Sayılan bu okumalar içindeki gözlenen alternatif alel oranı şöyledir:

$$\frac{18}{24+18} = \frac{18}{42} \approx 0{,}429.$$

Sayılar yorumlamayı göstermek için oluşturulmuştur. Önerilen derinlik koşulu, hata modeli veya gerçek örnek sonucu değildir. Diğer okumalar, çağırıcıya özgü kurallarla dışlanmış olabilir; bu nedenle bu oran, çağırıcının kendi ek açıklamasının yerine sessizce geçirilmemelidir.

İlk soru, genotipin teknik olarak desteklenip desteklenmediğidir. Hizalama kalitesini, okumaların yerleşimini, zincir dağılımını, yakın varyasyonları ve bölgede güvenilir çağrı yapılabilirliğini inceleyin. Orta düzeyde bir alternatif alel oranı tek başına heterozigotu ne doğrular ne de reddeder. Lokus çoğalmış bir bölgeyle örtüşüyorsa, aynı sayılar yanlış kopyaya atanmış okumalardan kaynaklanabilir. Ploidi belirsizse biyolojik yoruma geçmeden önce varsayılan genotip modeli gözden geçirilmelidir.

İkinci soru, alelin seçilen gen modeli altında neyi değiştirdiğidir. Anotasyonun bir transkriptte yanlış anlamlı, diğerinde intronik sonuç bildirdiğini varsayalım. İkisini de koruyun ve kodlayan transkriptin soruyla neden ilgili olduğunu belirtin. Amino asit değişimi, tahmin edilen dizi sonucudur; protein etkinliğini ölçmez. Varyant uyumlu bir popülasyon kaynağında bulunmuyorsa, aleli bütün popülasyonlarda nadir ilan etmek yerine bulunmama durumunu kaynağın kapsamıyla birlikte raporlayın.

Üçüncü soru, biyolojik sonucun hangi kanıtla gerekçelendirileceğidir. Birlikte kalıtım, ilişkilendirme, bağımsız işlevsel deney veya başka kanıtlar, deneye göre ilgili olabilir. Her biri uygun tasarım ve kontroller gerektirir. VEP ile SnpEff'in kodon değişimi konusunda anlaşması, esas olarak kendi girdileri altında anotasyon tutarlılığını destekler; fenotip nedeninin iki bağımsız gösterimi değildir. Bu aşamada makul sonuç, "hedefli izlem gerektiren, teknik olarak olası aday" olabilir.

Gözlemleri yorumlardan ayırmak

Katman	Sentetik örnekte desteklenen açıklama	Kanıtlanmayan açıklama
Okuma kanıtı	Sayılan okumalar iki aleli de destekler.	Destekleyen her okuma hedeflenen lokustan gelir.
Genotip modeli	Çağırıcı diploit heterozigot atar.	Ploidi ve kopya sayısı yalnızca genotip etiketinden bilinir.
Anotasyon	Bir transkript amino asit değişimi öngörür.	Protein işlevi veya organizma fenotipi değişmiştir.
Veri tabanı sorgusu	Belirli kaynakta eşleşen alel bulunmamıştır.	Alel türde veya popülasyonda yoktur.
İzlem	Aday, belirtilen araştırma kuralını karşılar.	Aday, nedensel açıklamadır.

13.2 Tekrarlanabilirlik

Referans ve anotasyon sürümlerini, yazılım sürümlerini, model tanımlayıcılarını, komut argümanlarını, aralık tanımlarını ve filtreleme kurallarını tam olarak saklayın. Günlükleri ve makinece okunabilir ara özetleri, projenin veri yönetişimi kuralları altında yerel olarak koruyun. İş akışının adı tek başına yeterli değildir; sürümleri, varsayılanları ve kaynak kümeleri değişebilir. Stokastik yöntemlerde, uygulama bunları sunuyorsa ilgili rastgelelik tohumunu ve çalışma zamanı varsayımlarını saklayın. Planlanan analizden sapmaları ve gerekçelerini kaydedin.

Uygulanabilir bir rapor, analiz edilmesi amaçlanan alanı, güvenilir çağrı yapılabilen alanı ve performans değerlendirme alanını ayırır. Örnek dışlamalarını, genotip maskelemeyi, eksikliği ve varyant sınıfına göre sayıları bildirin. Hangi sayıların kayıtlara, hangilerinin alellere ait olduğunu açıklayın. Anotasyondan sonra veri tabanı değişirse özgün sonucu, sürümü açıkça belirtilen yeniden anotasyonla birlikte koruyun. Negatif rapor, değerlendirilen bölgeleri ve varyant sınıflarını, ayrıca çözülemeyen önemli boşlukları belirtmelidir.

13.3 Gerçekte neyin doğrulandığını belirtmek

Bu nottaki komutlar ve eşlik eden betikler, belgelenmiş arayüzleri ve çevrelerindeki akıl yürütmeyi örnekler. Tamamlanmış bir gerçek veri iş akışı oluşturmazlar; sentetik örnekler de performans tahmini değildir. Yürütülen bir projede hangi adımların çalıştırıldığını, hangi çıktıların incelendiğini, hangi değerlendirme kaynağı veya bağımsız doğrulamanın kullanıldığını ve nelerin sınanmadığını ayrı ayrı raporlayın. Dosya oluşturma, indeksleme ve derlemenin başarılı olması teknik işlemleri gösterir; biyolojik doğruluğu kanıtlamaz.

Araştırma analizinin çıktısı, uygun sınırlar taşıyan ve denetlenebilir bir açıklamadır. Küçük varyant kümesi, değerlendirilmiş ve referansa hizalanmış diziyle ilgili belirli soruları destekleyebilir. Tek başına fenotipin bütün genomik nedenlerini dışlayamaz, değerlendirilmemiş yapısal varyasyonu doğrulayamaz veya klinik kullanıma hazır olunduğunu gösteremez. Son ifadeyi deneye, doğrulanmış alana ve her aday için mevcut kanıta göre kurun.

14 Okumalardan değerlendirilmiş varyant kümesine aşamalı iş akışı

Eşlik eden `main_scripts/genomic_variants` dizini, aşamalara ayrılmış Bash betiklerini, örnek tablosunu ve ayrı yazılım ortamlarını içerir. Böylece bu nottaki kavramlar incelenebilir komutlarla ilişkilendirilir. Her aşamayı ayrı çalıştırın ve ilerlemeden önce çıktısını inceleyin; paket, tamamlanmış bir gerçek veri yeniden üretimi veya klinik olarak doğrulanmış analiz hattı iddiası taşımaz. İki dildeki README dosyaları güncel komutları, argümanları ve yürütme sınırlarını açıklar.

Bölüm sırası, filtreleme kararlarını gösterimden ayrı açıklar; eşlikçi betikler yürütülürken önce normalizasyon yapılır, ardından incelenmiş filtre politikası uygulanır (Bölüm 7 ve 8).

14.1 Ortak hazırlık ve açık biyolojik dallanma

Okumaları ve türetilmiş verileri, `GENOMICS_PROJECT` ile seçilen harici analiz dizininde tutun. Örnek tablosunun her satırı bir çift uçlu dizileme birimini tanımlar; örnek, fiziksel kütüphane ve platform birimi ayrı alanlardır. Aynı kütüphanenin farklı şeritleri kütüphane tanımlayıcısını paylaşır; her birimin okuma grubu tanımlayıcısı benzersizdir. Illumina okumalarında `PL:ILLUMINA` platformu, `SM` örneği, `LB` kütüphaneyi ve `PU` platform birimini belirtir. Bütün okuma gruplarına aynı tanımlayıcıyı vermek bu ayrımları kaybettirir [4].

Örnek kayıtlar

```
main_scripts/genomic_variants/
  scripts/00_setup/       # validate manifest and inputs
  scripts/01_reference/   # index the selected reference
  scripts/02_reads/       # optional fastp
  scripts/03_alignment/   # BWA or Bowtie2, per sequencing unit
  scripts/04_preparation/ # library duplicates; optional BQSR
  routes/                # biological calling models
```

Hizalayıcı indekslerini gerçekten seçilen FASTA'dan oluşturun. BWA-MEM veya Bowtie2 seçimini açıkça yapın; argümanları ve hizalama tanılarını saklayın. Yerleştirme ve kalite modelleri aynı kanıtı üretmek zorunda değildir [30, 31]. İsteğe bağlı kırpma gerekçelendirilmeli, korunan okumalar incelenmelidir. Duplike işaretleme için aynı kütüphanenin birimlerini birleştirin; ardından örneğin kütüphanelerini bir araya getirin. Sunulan SAMtools yolu, `fixmate -m` için ada göre sıralama, ardından koordinata göre sıralama ve duplike işaretleme uygular. Son çağırma girdisi koordinata göre sıralanmış ve indekslenmiştir; geçici ada göre sıralanmış BAM, son indeksli girdi değildir [5]. İsteğe bağlı BQSR, uyumlu bilinen varyant konumları ve gerekçelendirilmiş çağırıcı protokolü gerektirir. Karşılaştırmalar boyunca işaretlenmiş veya yeniden kalibre edilmiş BAM seçimini koruyun.

Sabit ploidili betiği çalıştırmadan önce yalnızca bu modelle uyumlu bölgeleri içeren BED dosyasını inceleyin. Tam FASTA ve hizalama referansı değişmeden kalır. Diploit GATK, haploit BCFtools ve poliploit FreeBayes betiklerinin her biri, incelenmiş dosyayı isteğe bağlı ikinci argüman olarak kabul eder; koordinatları hazırlanmış referansla denetler ve çağrıların yanına bir kopyasını kaydeder. Örneğin:

Komut satırı

```
bash routes/diploid_germline/05_haplotypecaller.sh \
  sample1 reviewed-diploid.bed
```

BED koordinatları sıfır tabanlıdır ve bitişi içermez. `toy1 100 200`, bir tabanlı 101–200. bazları belirtir; gerçek dosyada alanlar sekmeyle ayrılır. Betikler GATK için `-L`, BCFtools mpileup için `--regions-file`, FreeBayes için `--targets` kullanır. BED verilmezse modelin bütün referans boyunca uygunluğu gerekçelendirilmelidir; tablodaki ploidi değeri her kontigi denetlemez. Referanstan kontig çıkarmak, hizalamada olası yerleşimleri değiştirir ve aynı kısıtlamayı sağlamaz. Çağırma alanı ile sonraki değerlendirme maskesi farklı soruları yanıtlar; değerlendirme maskesi yanlış çağırma modelini düzeltemez. Yerel montaj bağlamını, dayanak bazlı indelleri ve referans güven bloklarını sınırlarda inceleyin. Ek aralık genişletme istenmez; genişletme uygulanırsa eklenen alanın da modele uygun olması gerekir.

Çağırma dalını biyolojik soru belirler. Diploit germ hattı yolu, örnek başına gVCF üretir ve uyumlu gVCF'lerden ortak genotipleme yapar. Haploit yol açıkça haploit model kullanır; sabit ploidili poliploit yol ise beyan edilmiş kopya sayısını gerektirir. Bu örnekler ploidiyi çıkarsamaz veya çelişkili girdi bilgilerini düzeltmez. Karışık ploidi, anöploidi, organeller, havuzlar ve uzun okumalar, ilgili yolun sınırlarında açıklanan ek kararları gerektirir. İnsan eşleştirilmiş somatik analizi ise normal ve tümör örneklerini kullanır; mutasyon modeli Bölüm 11'de açıklanan belirsizlikleri ele almalıdır.

14.2 Hizalayıcıları ve somatik çağırıcıları adil karşılaştırmak

Sınırları belirlenmiş bir insan somatik karşılaştırması, BWA ile Bowtie2'yi ve Mutect2, Strelka2 veya VarScan2'yi değerlendirebilir; ancak dosya adlarından oluşan matris deney tasarımı değildir. Karşılaştırmanın gerektirdiği yerlerde okuma çiftlerini, örnek eşleşmesini, referansı, kırpma kararını, duplike politikasını,

BQSR koşulunu ve değerlendirme alanını sabit tutun. Her çağırıcının belgelenmiş girdi ve filtreleme koşullarını koruyun. Mutect2, somatik filtreleme aşamalarını gerektirir; VarScan'in durum ve yüksek güven etiketleri bağımsız doğruluk kanıtı değil, çağırıcı sınıflarıdır. Strelka2, eski yöntemlerle karşılaştırma için korunmuştur: Deposu arşivlenmiştir ve Python 2 çalışma zamanı, modern araçlardan ayrı ve uyumlu bir ortamda tutulmalıdır [25, 32, 33].

Önişleme seçiminin kendisi araştırılıyorsa bunu bir etken olarak tanımlayın ve karşılaştırmayı buna göre yorumlayın. Yakalama örneklerinde, değerlendirme kaynağının yüksek güven bölgeleri ile ekzom hedef maskesinin gerçek kesişimini kullanın; sınır politikasını belirtin. İki maskeden söz etmek, kesişimlerini oluşturup kullanmadan değerlendirme alanını kısıtlamaz. Germ hattı doğruluk kümesi germ hattı alellerini tanımlar; düşük frekanslı somatik olayları kendiliğinden etiketlemez. Olay sınıfına ve alel frekansı aralığına uygun kanıt kullanın; değerlendirilmeyen çağrıları ayrı tutun [23, 24].

TP, FP ve FN sayılarını açık eşleştirme kurallarıyla, Bölüm 10'da tanımlanan pozitif öngörü değeri ve duyarlılıkla birlikte raporlayın. Doğruluk (accuracy), ayrıca tanımlanmış bir gerçek negatif evreni ve payda gerektirir; yalnızca varyant içeren sıradan VCF sayımları bunları sağlamaz. Çağırıcıları ham çağrı sayısına veya dayanağı olmayan doğruluk yüzdesine göre sıralamayın. SNP/indel pozitif öngörü değeri–duyarlılık, kapsama veya uyuşmazlık sınıfları gibi soruyu yanıtlayan özetleri seçin; anlamlı karşılaştırma için PCA veya sabit sayıda VCF dosyası zorunlu değildir.

14.3 Yürütme kaydı olarak tekrarlanabilirlik

Her aşamada girdi kaynağını ve sürümünü, tam komutu, etkin argümanları, yazılım ve ortam sürümlerini, referans ve kaynak kimliğini, çıktı konumlarını ve çıkış durumunu saklayın. Çalışma süresi raporlanıyorsa geçen süreyi, ayrılan iş parçacığı sayısını ve belleği donanım bağlamıyla kaydedin. İşlem stokastik davranış sunuyorsa rastgelelik tohumunu koruyun; her komut için tohum uydurmak yerine uygulanabilir olmadığını belirtin. Sürümleri sabitlenmiş ortam dosyası araçları yeniden kurmaya yardımcı olur; çözümlenmiş ortam dışa aktarımı ise bir platformda gerçekten kullanılan paketleri kaydeder.

Çıktı denetimlerini çalıştırmadan önce tanımlayın. Geçerli okuma çiftleri, uyumlu referans kontigleri, beklenen BAM örnek ve okuma grupları, koordinat sırası ve indeks, beklenen VCF örnekleri ve ploidi, referansla tutarlı aleller ve boş olmayan değerlendirme alanı bu denetimlere örnektir. Kasıtlı olarak bozulmuş sentetik girdilerle denetimleri sınayın ve hata çıkışı gerektirin. Dosya bütünlüğü denetimleri tutarlılığı sınar; bağımsız değerlendirme kaynakları veya deneyler biyolojik geçerliliği ele alır. Sözdizimi denetimlerini, küçük sentetik yürütme sınamalarını ve gerçek veriyle yeniden üretimi ayrı raporlayın; çalıştırılmamış aşamaları belirtin. Değişen sonucun nedenini izleyebilmek için ham çağrıları ve normalizasyon, filtreleme, anotasyon ve karşılaştırma kayıtlarını koruyun.

14.4 Aynı protokolü yerel Galaxy kurulumuna eşlemek

Yetkili bir yerel Galaxy kurulumunda aynı mantıksal aşamaları mevcut araçlarla eşleyin ve veri kümelerinin ilişkilerini geçmişte koruyun. Kurulu araç tanımlayıcısını ve sürümünü, parametreleri, referans seçimini ve işin sonucunu kaydedin; geçmişi ve yeniden kullanılabilir iş akışını tamamlayıcı kayıtlar olarak yerel dışa aktarın. Geçmiş neyin çalıştığını, iş akışı ise yeniden kullanılabilir bağlantıları açıklar. Gerekli yeniden kalibrasyon araçları dâhil olmak üzere araçların varlığını ve sürümlerini kullanılan kurulumda denetleyin; bütün Galaxy kurulumlarının aynı olanakları sunduğunu varsaymayın [34, 35]. Bu kavramsal eşleme, komut satırı sonuçlarıyla eşdeğerliği kanıtlamaz; etkin parametreleri ve çıktıları aynı biyolojik koşullar altında karşılaştırın.

Kaynaklar

[1] Brown TA. *Genomes 5*. 5th ed. CRC Press; 2023. ISBN: 9780367674076.

[2] Van der Auwera GA, O'Connor BD. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media; 2020. ISBN: 9781491975190.

[3] Global Alliance for Genomics and Health File Formats Team. *The Variant Call Format Specification: VCFv4.5 and BCFv2.2*. 25 February 2026. https://samtools.github.io/hts-specs/VCFv4.5.pdf

[4] SAM/BAM Format Specification Working Group. *Sequence Alignment/Map Format Specification*. Version 1.6; 12 August 2025. https://samtools.github.io/hts-specs/SAMv1.pdf

[5] SAMtools developers. *samtools(1) manual page*. Version 1.25; 27 October 2026. https://www.htslib.org/doc/samtools.html

[6] GATK Team. *Germline short variant discovery (SNPs + Indels)*. Updated 24 July 2024. https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels

[7] GATK Team. *HaplotypeCaller*. Updated 20 August 2026. https://gatk.broadinstitute.org/hc/en-us/articles/54698694431131-HaplotypeCaller

[8] GATK Team. *Known variants—Training resources—Truth sets*. Updated 25 June 2024. https://gatk.broadinstitute.org/hc/en-us/articles/360035890831-Known-variants-Training-resources-Truth-sets

[9] GATK Team. *GenomicsDBImport*. Updated 6 October 2022. https://gatk.broadinstitute.org/hc/en-us/articles/360057439331-GenomicsDBImport

[10] GATK Team. *Base Quality Score Recalibration (BQSR)*. Updated 29 August 2024. https://gatk.broadinstitute.org/hc/en-us/articles/360035890531-Base-Quality-Score-Recalibration-BQSR

[11] GATK Team. *Hard-filtering germline short variants*. Updated 6 November 2024. https://gatk.broadinstitute.org/hc/en-us/articles/360035890471-Hard-filtering-germline-short-variants

[12] Poplin R, Chang PC, Alexander D, et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol*. 2018;36(10):983–987. https://doi.org/10.1038/nbt.4235

[13] DeepVariant developers. *DeepVariant FAQ*. Release r1.10. https://github.com/google/deepvariant/blob/r1.10/docs/FAQ.md

[14] Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. *arXiv* [preprint]. 2012;1207.3907. https://doi.org/10.48550/arXiv.1207.3907

[15] FreeBayes developers. *FreeBayes: a haplotype-based variant detector*. Software documentation. https://github.com/freebayes/freebayes

[16] Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *GigaScience*. 2021;10(2):giab008. https://doi.org/10.1093/gigascience/giab008

[17] BCFtools developers. *bcftools(1) manual page*. https://samtools.github.io/bcftools/bcftools

[18] McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17:122. https://doi.org/10.1186/s13059-016-0974-4

[19] Ensembl. *Variant Effect Predictor: cache documentation and software guide*. https://www.ensembl.org/info/docs/tools/vep/script/vep_cache.html; https://github.com/Ensembl/ensembl-vep/tree/release/115

[20] Ensembl. *Calculated variant consequences*. https://static.ensembl.org/info/genome/variation/prediction/predicted_data.html

[21] Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*. 2012;6(2):80–92. https://doi.org/10.4161/fly.19695

[22] SnpEff developers. *Building databases*. https://pcingola.github.io/SnpEff/snpeff/build_db/

[23] Krusche P, Trigg L, Boutros PC, et al. Best practices for benchmarking germline small-variant calls in human genomes. *Nat Biotechnol*. 2019;37(5):555–560. https://doi.org/10.1038/s41587-019-0054-x

[24] National Institute of Standards and Technology. *Genome in a Bottle*. https://www.nist.gov/programs-projects/genome-bottle

[25] GATK Team. *Mutect2*. https://gatk.broadinstitute.org/hc/en-us/articles/54698777535387-Mutect2

[26] PLINK developers. *PLINK 1.9: Input filtering*. https://www.cog-genomics.org/plink/1.9/filter#hwe

[27] Ensembl. *VEP command-line configuration*. Release 115. https://github.com/Ensembl/ensembl-vep/blob/release/115/modules/Bio/EnsEMBL/VEP/Config.pm

[28] SnpEff developers. *Running SnpEff*. https://pcingola.github.io/SnpEff/snpeff/running/

[29] Peter J, De Chiara M, Friedrich A, et al. Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature*. 2018;556:339–344. https://doi.org/10.1038/s41586-018-0030-5

[30] BWA developers. *BWA manual*. https://bio-bwa.sourceforge.net/bwa.shtml

[31] Bowtie 2 developers. *Bowtie 2 manual*. https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml

[32] Illumina. *Strelka2 germline and somatic small variant caller*. Archived software repository and user guide. https://github.com/Illumina/strelka

[33] VarScan developers. *Somatic Mutation Calling*. https://varscan.sourceforge.net/somatic-calling.html

[34] Galaxy Training Network. *Understanding Galaxy history system*. https://training.galaxyproject.org/training-material/topics/galaxy-interface/tutorials/history/tutorial.html

[35] Galaxy Training Network. *Creating, Editing and Importing Galaxy Workflows*. https://training.galaxyproject.org/training-material/topics/galaxy-interface/tutorials/workflow-editor/tutorial.html

[36] Olson ND, Wagner J, Dwarshuis N, et al. Variant calling and benchmarking in an era of complete human genome sequences. *Nat Rev Genet*. 2023;24:464–483. https://doi.org/10.1038/s41576-023-00590-0

[37] Liao WW, Asri M, Ebler J, et al. A draft human pangenome reference. *Nature*. 2023;617:312–324. https://doi.org/10.1038/s41586-023-05896-x

[38] Wagner J, Olson ND, McDaniel J, et al. Small variant benchmark from a complete assembly of X and Y chromosomes. *Nat Commun*. 2025;16:497. https://doi.org/10.1038/s41467-024-55710-z

[39] IGV developers. *IGV Desktop: Alignments basics*. https://igv.org/doc/desktop/UserGuide/tracks/alignments/viewing_alignments_basics/

[40] IGV developers. *IGV Desktop: Variants (VCF)*. https://igv.org/doc/desktop/UserGuide/tracks/vcf/

[41] Zheng Z, Sedlazeck FJ, Luo R. Small variant calling in the era of long-read sequencing. *Nat Rev Genet*. Published online 20 October 2026. https://doi.org/10.1038/s41576-026-01020-7

[42] OpenGene. *fastp documentation*. https://github.com/OpenGene/fastp