GenomBilim Kisokulu Pratigi

# Dizileri Hizalama ve Varyant Belirleme

Tugce Bilgin Sonay

Mikrosatelit calismalarinin en buyuk problem veri eksikligiyken yeni nesil dizileme (NGS) teknikleri bunun tam tersi bir sorunla karsi karsiya. NGS metotlari o kadar buyuk bir veri uretiyor ki bunu analiz etmek projeler icin onemli bir zorluk teskil ediyor. Arastirmacilar bu buyuklukteki verilerle basa cikabilmek icin oldukca guclu bilgisayarlara ve isletim sistemlerine ihtiyac duyuyorlar. Biyoinformatik bu ihtiyaclari karsilamak amaciyla cikmis bir araclar ve yaklasimlar butunudur. Biyoinformatikci, biyoloji ve bilgisayar bilimi arasinda kopru gorevi gorur. Gunumuzde pek cok universite bu konuda lisans veya yuksek lisans programlari sunsa da biyoinformatik aslen kendi cabamizla da ogrenebilecegimiz bir branstir.

Bu kis okulunda son yillarda biyoinformatigin en ihtiyac duyuldugu alan olan dizilemedeki kullanimlarini ogrenecegiz. Bu amacla herkes kendisine verilen erisim numarasindaki bireyin kisa mitogenom dizilerini insan referans genomuna gore hizalayacak, mutasyonlari bulup her dizinin son halini hazirlayacak. Sonrasinda bu dizileri kullanarak filogenomik haritalarini cikaracagiz.

Bu islemleri yapmak icin hal-i hazirda pek cok biyoinformatik araci zaten internette mevcut. Bu araclarin onemli bir kismi terminal dedigimiz komut satirlarindan calistirilabiliyor. Bu da ancak linux/unix bazli isletim sistemlerinde mumkun. Dolayisiyla kis okulumuz surecinde linux isletim sisteminde calisacagiz.

# Kisa Dizi Kutuphanesinden mtDNA’yi indir

## Calisma klasorunu olustur:

Ortak kumeye

ssh -Y [egitim1@levrek1.ulakbim.gov.tr](mailto:egitim1@levrek1.ulakbim.gov.tr)

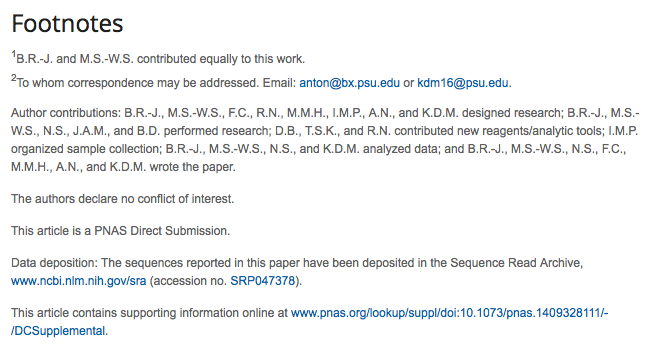
komutuyla gir. –Y sayesinde grafiklerini goruntuleyebilirsin.

Masaustunde “Genomik” adi altinda bir klasor olustur. Proje boyunca bu klasoru kullanacaksin.

## Kisa dizilerini kullanacagin makaleyi bul:

“Maternal age effect and severe germ-line bottleneck in the inheritance of human mitochondrial DNA” adli makaleyi asagidaki linkten bul:

<http://www.pnas.org/content/111/43/15474.full>



Figur . Makalenin Footnotes kismi

Makalede kac mitogenom dizisi uretilmis?

Nasil bir teknoloji kullanarak uretilmisler?

Vucudun hangi bolgesinden alinan ornekler kullanilmis?

Footnotes kisminda (Figur 1), erisim numarasi linkine (accession no.) tikla. Bu link seni arastirmanin uretip yayinladigi kisa dizilerin listesine goturecek (Figur 2).

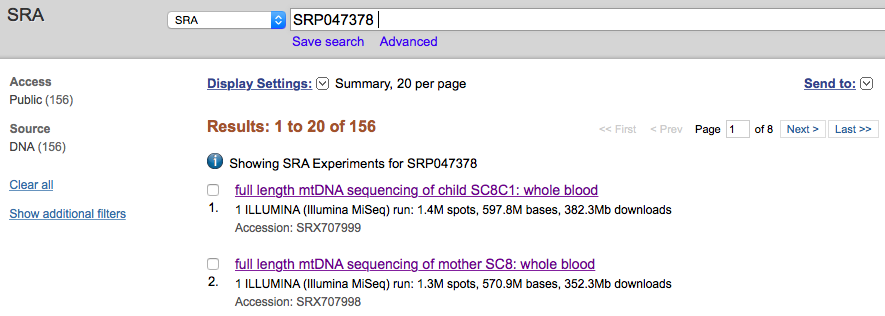
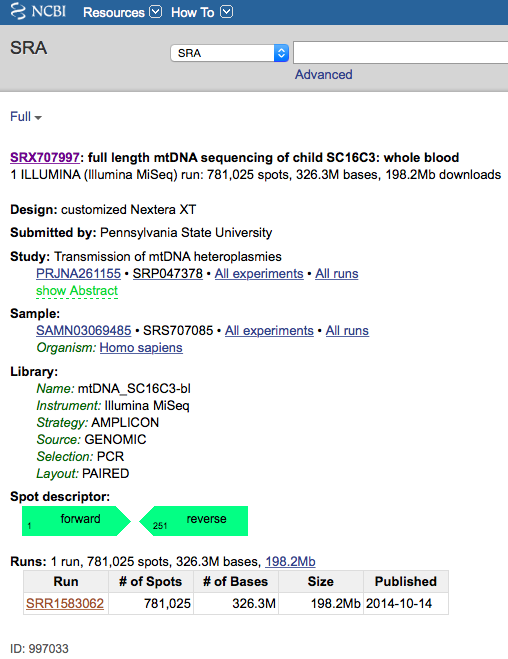


Figure . SRA sayfasi

SRA yani kisa dizi arsivi ozellikle yeni dizileme teknolojileri kullanilarak uretilmis genelde 1000 bazdan kisa dizilerin depolandigi bir veri tabanidir. SRA arastirmacilara kisa dizilerini kullanima acik olarak depolayabilecekleri bir yer gorevi gormekle kalmayip ayni zamanda aranilan diziye ve diziyle ilgili detayli deneysel bilgiye hizli bir erisime olanak verir.

## Calisacagin kisa mitogenom dizilerini indir:

SRA icinde sana verilen erisim numarasindaki mtDNAyi bul ve dizileri indir. Bunun icin ‘Run’ basliginin altindaki linkteki ‘SRR’ ile baslayan ismi not al.



Figur . Kisa dizileri indirmek

Kisa dizi dosyasini indirmek icin, SRA toolkit’i kullanacagiz.

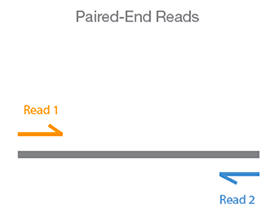
Once terminali ac ve genomik klasorune gitmek icin asagidaki komutu gir:

**cd Desktop/Genomik/**

Dosyayi indirmek icin asagidaki komutu kullanacagiz, SRRin ardindan gelen uc noktayi indirmek istedigin dizinin numarasiyla degistirmeyi unutma:

**fastq-dump –I --split-files SRR…**

Bu komut iki tane dosya indirecek: one dogru diziler icin SRR…\_1.fastq ve arkaya dogru diziler icin SRR…\_2.fastq. Iki tane dosya olmasinin sebebi dizilerin paired-end teknolojisiyle uretilmis olmasi. Paired-end, Figur 4’de de gordugun gibi bir fragmanin iki ucundan dizilenmesini saglar, bu sayede cok daha yuksek kalitede bir dizi elde edilir. Biz kolaylik acisindan bu derste sadece one dogru uzanan dizilerle calisacagiz. O yuzden SRR…\_1.fastq dosyasinin adini infile\_1.fastq olarak degistirip SRR…\_2.fastq’yu dosyandan silebilirsin.



Figur 4. paired-end okuma

Eger vaktin varsa fastq-dump komutunun nasil isledigini anlamaya calisabilirsin: <http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=toolkit_doc&f=fastq-dump>

# Filtreleme ve Kesme

## Kisa dizilerini kesfe cik:

Terminalde asagidaki komutu yazalim:

**head infile\_1.fastq**

Bu komut infile\_1.fastq dosyasini acmadan sadece ilk 10 satirina goz atmamizi saglayacak. Dosyamiz cok buyuk oldugu icin **head** (ve son satirlari gosteren **tail**) cok ise yarayan bir komut. fastq dosyasinda bir dizi dort satirdan olusur ve asagidaki gibi gozukur:

@SRR1583062.1.1 1 length=183

AAATAATAGGATGAGGCAGGAATCAAAGACAGATACTGCGACATAGGGTGCTCCGGCTCCAGCGTCTCGCAATGCTATCGCGTGCACACCCCCCAGACGAAAATACCAAATGCATGAAGAGCTCCCGTGAGTGGTTAATAGGGTGATAGACCTGTGATCCATCGTGATGTCTTATTTAAGGGG

+SRR1583062.1.1 1 length=183

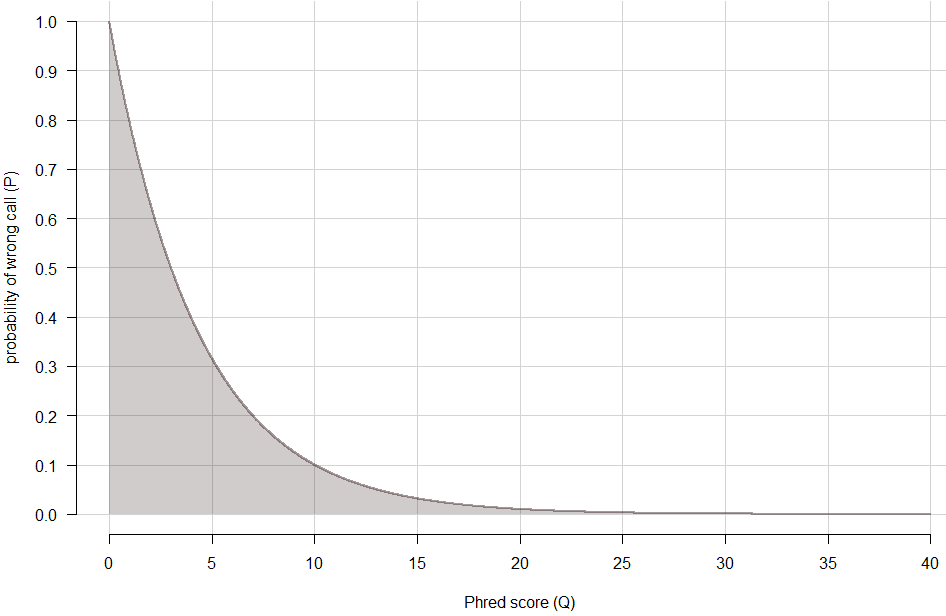
??A??BBBEEDDDDDDGGGGGGIIIIIIIIIIIIIIIIIHHHHHIIIIFHHIIIHHHHHIIIIHHHHIHIHHHIIIIIIHHHHEHHHHHHHHGGGGGGGGGHGGGGGGGGGGGGGGHGGGGGGGGGGCGEGEEGGGGGGGGGGCEGGGGGGGGGGGGGGGGGGGGGEGGEGGGGGGGGGGGGG

Ilk ve ucuncu satir dizinin adini ve uzunlugunu, ikinci satir dizinin kendisini gosterir. Son satirsa Phred kalite skorlarini gosterir. Bu satirdaki her karakter bir bazin kalite skoruna karsilik gelir.

Bir bazin Phred kalite skoru, , o bazin hatali okunma ihtimali *P*’nin logaritmasidir (Figur 5).

or

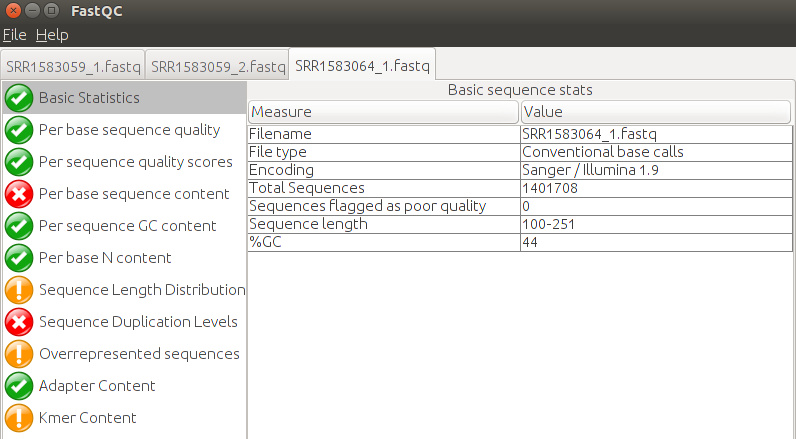
Mesela, Phred kalitesi 20 olan bir bazin yanlis okunmus olma ihtimali 0.01’dir (. Yani yuz okumadan biri bu bazi yanlis teshis edecektir. Bilimsel calismalar genelde kalite skoru en az 20 hatta mumkunse 30 olan bazlari dikkate aliyorlar.



Figur 5. Phred Kalite Verileri

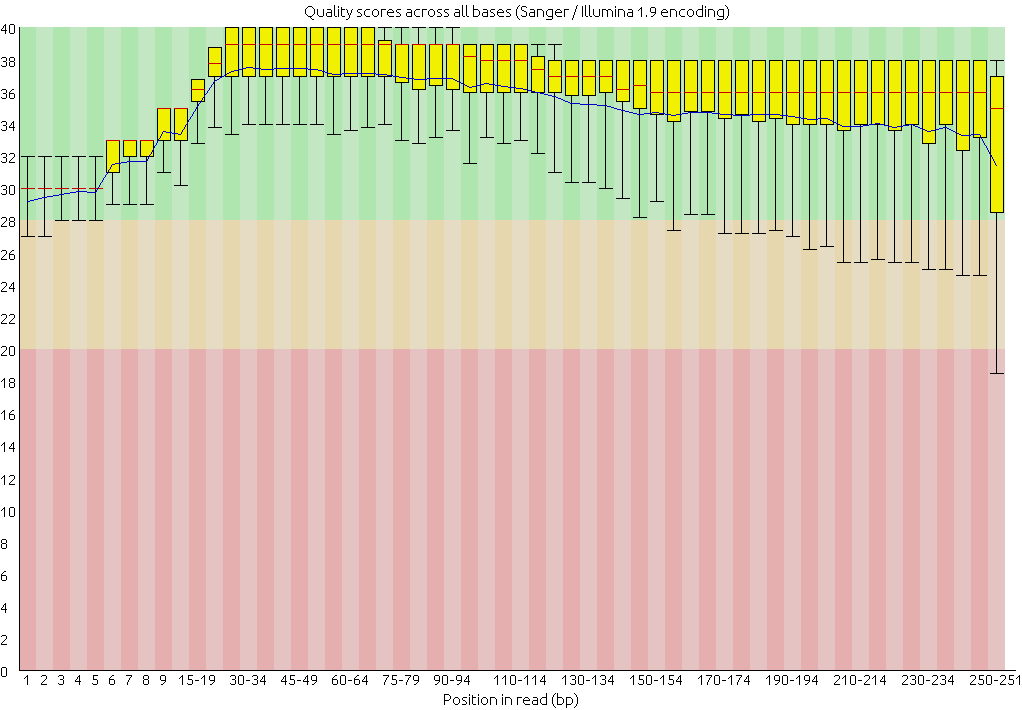
## Dizinin kalite dagilimini kontrol et:

* FastQC programini ac.
* Menu kisminda File’da fastq dosyani bul ve ac.
* Sen dosyani acar acmaz program analize baslayacak ve bir kalite raporu olusturacak. Bu islem biraz zaman alabilir.



Figur 6. FastQC programi

* Butun dizilerin analizi bitince kalite raporu otomatik olarak acilir ve bazi temel istatistikleri gosterir (Figur 6). Bizim icin en onemli kisimlar dizi sayisi (Total sequences) ve dizi uzunlugu. Dizi sayisinin coklugu genel olarak iyi bir isarettir.
* Yukaridaki menunun icindeki -> save report komutunu kullanarak Genomik klasorune raporu kaydet. Boylece istedigin zaman kalite verilerine geri donebilirsin.
* Sol taraftan daha detayli kalite raporlarina da bakabilirsin. Ornegin “Per base sequencing quality”ye tikla. Bu grafik (ornegin Figur 7) dizi boyunca uzanan baz gruplari icin kalite skorlarini gosterir.
* Grafigin nasil bir sekli var?



Figur 7. Dizi kalite grafigi

* Dusuk kaliteli bazlar genom analizi acisindan nasil bir risk olusturabilir? Bir dizinin kalitesini arttirmak icin ne yapmak gerekir?

Bu dizileri analiz etmeden once dizilerimizi dusuk kaliteli okumalardan ve bazlardan temizlemeliyiz. Bunu iki asamada yapacagiz. Ilk once kalitenin genel olarak dusuk oldugu bas ve sonlardaki bazlari kesecegiz. Sonrasinda elimizeki dizileri filtreleyerek belli bir kalite skorunun altinda bazi cok olanlari atacagiz.

* Neden once kesip sonra filtreliyoruz? Tersini yapsaydik bir sey degisir miydi?

Not: FastQC programini henuz kapatmayin.

## Kotu kalite bazlari kes:

Bas ve sonlardaki kotu kalite bazlari kesmek icin fastx fastx\_trimmer diye bir program kullanacagiz. Bunun icin:

* Terminali ac.
* Genomik klasorune git:

**cd Desktop/Genomik**

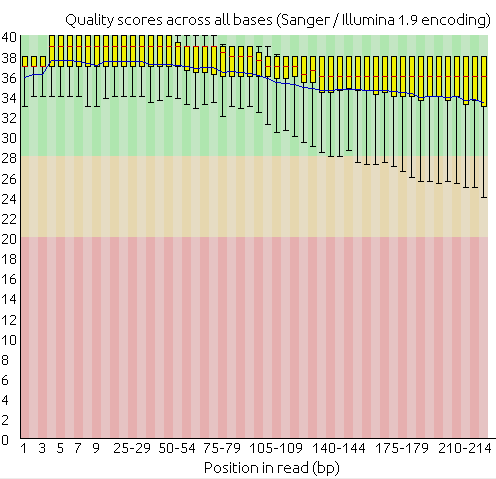
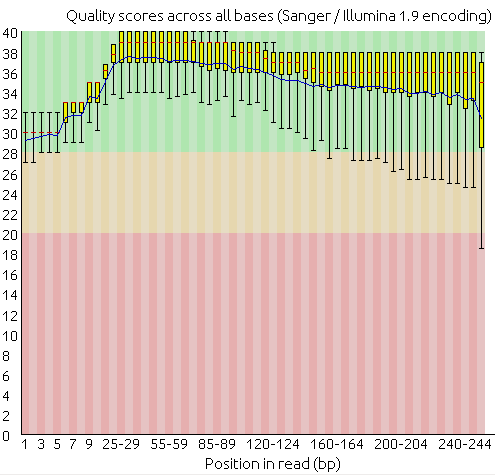
* Asagidaki komutu gir:

**fastx\_trimmer -f 20 -l 240 -i infile\_1.fastq –o infile.fastq**

* Bu komutun bitmesini beklerken asagidaki linkten ne yaptigina bakabilirsin: <http://hannonlab.cshl.edu/fastx_toolkit/commandline.html>

Linki kullanarak kullandigimiz her bir parametrenin ne ise yaradigini not al:

* + **-f**
  + **-l**
  + **-i**
  + **-o**
* Dizininde infile.fastq isimli yeni bir dosya olusacak. Icerigine bir goz at, komut yapmasini bekledigin islemi yapmis mi diye bir kontrol et.
* fastQC’ye geri don ve bu yeni dosyanin kalite raporunu cikart, kalite grafigine bak. Bu grafigi eskisiyle karsilastir. Ne goruyorsun? (mesela Fugur 8)
* Her ornegin kendine ozgu bir kalite dagilimi olacaktir. Sence senin ornegin icin bu kadar kesmek iyi mi?



Figur 8. Bas ve sondaki bazlari kesmeden once ve kestikten sonraki kalite grafigi

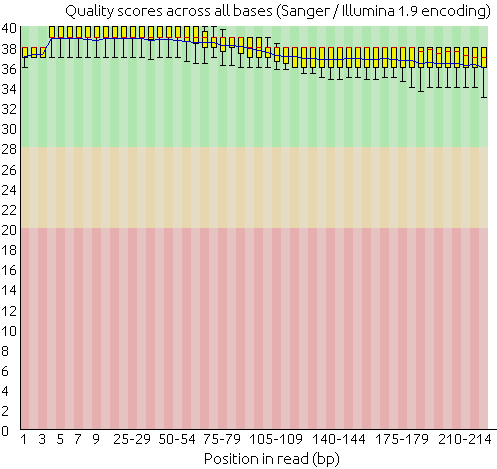
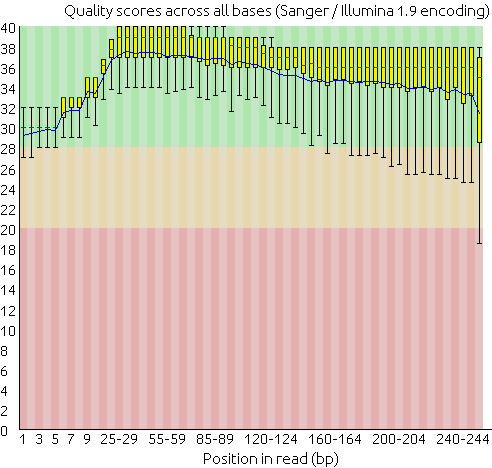
## Kotu kalite dizileri filtrele:

Dizilerin genel olarak kalitesi iyilesti. Fakat hala aralarda kotu kalite bazlarimiz mevcut. Bunun icin, simdi “fastq\_quality\_filter” isimli programi kullanarak kotu kalite bazi fazla olan dizileri filtreleyecegiz.

* Terminalde genomik klasorune git ve asagidaki komutu gir:

**fastq\_quality\_filter -q 30 -p 95 -i infile.fastq -o outfile.fastq**

* Filtreleme gerceklesirken bu komutun tam olarak ne yaptigina yandaki linkten bir goz at: <http://hannonlab.cshl.edu/fastx_toolkit/commandline.html>, ve asagiaki parametrelerin ne ise yaradigini not al:
  + **-q**
  + **-p**
  + **-i**
  + **-o**
* Filtreleme bitince, fastQC’de olusan dizinin kalite grafigini uret. Bu sefer ne degisti? (ornegin Figur 9)



Figur 9. Filtreleme oncesi ve sonrasi kalite grafigi

* Ilk dosyanin temel istatistikleriyle bu dosyaninkileri karsilastir. Orjinal dosyadaki kac diziyi (yuzde kacini) filtreleyerek atmisiz?

Kesme ve filtreleme islemlerini bitirdigimize gore artik dizilerimizi referans genomuna yerlestirmeye baslayabiliriz. Bu kisimda p 95/q 30 parametreleriyle filtrelenmis dizileri kullanacagiz.

# Insan mitogenomu

Yerlestirme islemlerine baslamadan once referans dizimizi yani insan mitogenomuna ihtiyacimiz var. Ortak kumeden ref.fasta dosyasini Genomik dizinine kopyala.

**cp ../../ref.fasta .**

# Dizileri referans genomuna yerlestir

Dizileri yerlestirmek eslestirmece oyunu gibi dusunulebilir. Amac kisa dizilerin aynilarini (veya benzerlerini) buyuk dizide yani genomda bulmaktir. Bu islemi kolaylastirmak icin biyoinformatik araclari indexleme denilen bir teknik kullanirlar. Bu islem sayesinde cok hizli bir sekilde kisa diziler genom uzerinde yerlerini bulurlar. Biz yerlestirme icin bowtie2 programini kullanacagiz.

## Genomu indexle

Hizli yerlestirme icin ilk adim referans genomumuzu indexlemek. Bunun icin terminalden Genomik klasorune git ve asagidaki komutu gir:

**bowtie2-build ref.fasta ref**

## Dizileri yerlestir

Artik kisa dizileri genoma yerlestirmeye haziriz. Bunun icin asagidaki komutu gir:

**bowtie2 -x ref -q outfile.fastq –S outfile.sam**

Bu islem 5-10 dakika icinde bir SAM (**S**equence **A**lignment/**M**ap) dosyasi uretecek. SAM dosyasi insan gozunun okuyabilecegi text formatinda bir dosyadir ve icinde her bir dizi icin bir baslik bir de yerlestirmeyle ilgili detaylarin yer aldigin bir kisim vardir. Dosyanin ilk satirlarini acip icine bir goz atabilirsin. Yandaki link dosya icerigiyle ilgili daha detayli bilgi verir: <https://samtools.github.io/hts-specs/SAMv1.pdf>

Yerlestirme bitince ekranda kisa bir rapor belirecek. Kisa dizilerinden yuzde kaci yerlestirilebilmis?

SAM dosyasini analiz etmek icin SAMTools programini kullanacagiz. Her programin kendi indexleme sekli oldugundan referans genomumuzu asagidaki komutu kullanarak tekrar indexlememiz gerek:

**samtools faidx ref.fasta**

Bu ref.fasta.fai. isimli bir dosya uretecek.

Simdi SAM dosyamizi BAM (**B**inary **A**lignment/**M**ap)’e cevirecegeiz. BAM SAM’in binary halidir. Bu islem icin asagidaki komutu kullanalim:

**samtools import ref.fasta.fai outfile.sam outfile.bam**

Asagidaki komutlarla bam dosyamizi da indexleyecegiz:

**samtools sort outfile.bam outfile.sorted.bam**

**samtools index outfile.sorted.bam**

# Varyant Tespiti

Dizimizin son halini uretmeden onceki son adimda referans genom ve elimizdeki dizi arasindaki mutasyonlari bulacagiz. Bunun icin yine SAMTools’u kullanacagiz.

## Referans genomu kullanarak mutasyonlari bul:

Kullandigimiz yerlestirme teknikleri kisa dizileri dogru yerlerine her zaman cok yuksek kesinlikte yerlestiremeyebilir. Ornegin, referans genomda uzun bir tekrar dizisinin icinde yer alan kisa bir dizi pekala tekrar dizisinin baska yerlerine de yerlesebilir. Bu durumda kullandigimiz program uygun yerlerden herhangi birini rastgele secmek durumunda kalir. Bu da yerlestirme guvenilirligini azaltir.

Yerlestirme programlari yerlestirme guvenilirliklerini tipki dizileme kaliteleri gibi skorlarlar. Yerlestirme skoru 10 olan bir dizinin yanlis yere yerlestirilmis olma ihtimali yuzde ondur. Dolayisiyla dizileme kalitesine gore degil yerlestirme kalitesine gore de filtreleme yapmamiz gerekir. Bunun icin yerlestirme kalitesi 10’dan dusuk olan dizileri terminalde, asagidaki komutla filtreleyecegiz:

**samtools mpileup outfile.sorted.bam -Q 30 -q 10 -uf ref.fasta > outfile.bcf**

Bu komutla yaklasik 10 dakika icinde outfile.bcf olarak adlandirdigin bir bcf dosyasi uretiyorsun. Bcf dosya formati tum mtDNA’n icin varyant bilgisini tutan binary bir formattir. Bu formatin icerigiyle ilgili daha fazla bilgi icin yandaki linke goz atabilirsin: <http://samtools.sourceforge.net/samtools.shtml>

# Son diziyi olustur

## Varyant dosyasini okunabilir formata cevir:

mtDNA dizinin son halini uretebilmek icin once, variant dosyasini bcf’ten, insan tarafindan okunabilir bir format vcf (Variant Call Format)’e cevirmen gerekir. Bunun icin BCFtools araclarini kullanacagiz:

**bcftools call –cV indels outfile.bcf > outfile.vcf**

Eger vaktin varsa, vcf dosyani acip yandaki link yardimiyla icerigini anlamaya calisabilirsin: <http://samtools.github.io/hts-specs/VCFv4.2.pdf>

## Varyant dosyasini indexle:

Bcftools’ta son dosya halini yaratmak icin varyantlarin indexlenmesi gerekir. Bunun icin asagidaki konutlari girecegiz:

**bcftools view outfile.vcf -O z –o outfile.vcf.gz**

**bcftools index outfile.vcf.gz**

## Varyantlari kullanarak dizinin son halini uret:

En son olarak buldugun varyantlari kullanarak elindeki dizinin mtDNA halini uret. Bunun icin terminalde asagidaki komutu gir, <myname> kismini kendi isminle degistirmeyi unutma:

**bcftools consensus outfile.vcf.gz -f ref.fasta –o <myname>.fasta**

Bu komut mtDNA’ni iceren bir fasta dosyasi uretecek (ornek: tugce.fasta).

# Filogenetik Agac

Artik siniftaki herkesin urettigi mtDNA’larla bir filogenetik agac uretebiliriz.

## Genomlari topla:

Oncelikle kendi genom dosyani ac ve ilk satirinda > isaretinden hemen sonra ismini yaz:

>NC\_012920.1 Homo sapiens mitochondrion, complete genome

yerine

>tugce

Boylece herkesin dosyalarini bir araya topladigimizda, herkesin dosyasinin basinda kendi ismi olacak. Artik dosyani ortak kumeye kaydedebilirsin.

**cp tugce.fasta ../../genomlar/**

Sonrasinda ortak alandaki tum genomlari kendi dizininde tek bir dosyaya kaydetmek icin asagidaki komutu terminalde gir:

**cat ../../genomlar/\* > genomes.fasta**

## Genomlari hizala:

Agaci olusturmadan once, elimizdeki genomlarin hepsini bir hizaya getirmemiz gerekir. Bunun icin Clustal Omega adinda bir arac kullanacagiz. Clustalo’nun nasil calistigini anlamak icin asagidaki komutu terminalinde yazabilirsin:

**clustalo --help**

Daha onceden olusturdugumuz genomes.fasta dosyasinin icindeki genomlari toplu olarak hizalamak icin asagidaki komutu terminalinde gir:

**clustalo -i genomes.fasta -o alignedgenomes.phy --outfmt=phy> alignedgenomes.phy**

Bu komut yaklasik 20 dkda icinde hizalanmis genomlarin bulundugu aligned\_genomes.fasta adinda bir dosya uretecek.

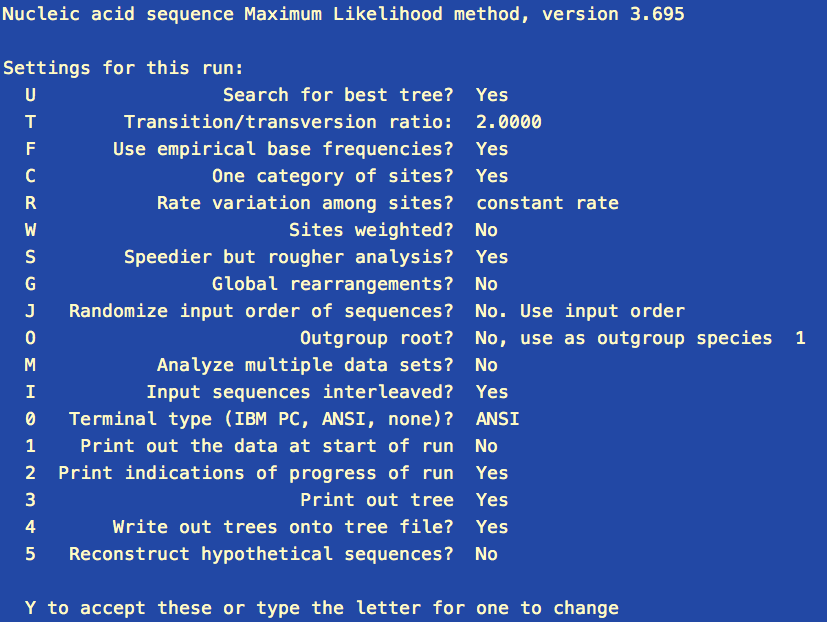
## Agaci uret:

Phylip adli programi kullanarak once DNA dizilerini maximum likelihood denilen istatiksel yontemle inceleyip dizilerimiz arasindaki uzakliklari hesaplatacagiz. Bunun icin asagidaki komutu gir:

**dnaml alignedgenomes.phy**

Bu komut bizden asagidaki gibi input dosyasini isteyecek.

Sonra da asagidaki ekranda parametrelerini goreceksin. Parametrelere bakip hepsini anladigina emin olduktan sonra kabul etmek icin y harfine bas.



Y harfine basinca ekraninda tum dizilerin listelendigini goreceksin. Bu islem uzun surebilir. Sonucta outfile ve outtree diye iki dosya uretecek. Fikir edinmek icin ikisinin de iclerini acip bak.

## Agaci ciz:

Simdi artik agacimizi cizebiliriz. Bunun icin FigTree programini kullanacagiz. Bunun icin FigTree programini acip ‘File’ menusunde ‘Open’ kismina girip oradan outtree adiyla kaydolmus agaci sec.

Bir kac saniye icerisinde agaclarimiz hazir. Asagida nasil gorunmesi gerektigiyle ilgili bir ornek bulabilirsin.

Nasil bir agac goruyorsun?

Senin dizin agacin neresinde?